RTO-MP-049

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION

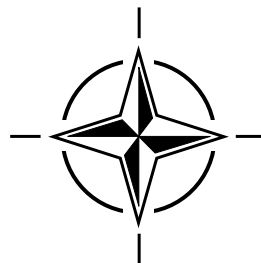BP 25, 7 RUE ANCELLE, F-92201 NEUILLY-SUR-SEINE CEDEX, FRANCE

*Note: There is an Addendum to this publication
containing the PowerPoint presentations.*

RTO MEETING PROCEEDINGS 49

New Information Processing Techniques for Military Systems

(les Nouvelles techniques de traitement de l'information
pour les systèmes militaires)

*Papers presented at the Information Systems Technology Panel (IST) Symposium held in Istanbul,
Turkey, 9-11 October 2000.*

Published April 2001

Distribution and Availability on Back Cover

Form SF298 Citation Data

Report Date ("DD MON YYYY") 00042001	Report Type N/A	Dates Covered (from... to) ("DD MON YYYY")

Title and Subtitle New Information Processing Techniques for Military Systems	Contract or Grant Number
	Program Element Number

Authors	Project Number
	Task Number
	Work Unit Number

Performing Organization Name(s) and Address(es) Research and Technology Organization North Atlantic Treaty Organization BP 25, 7 rue Ancelle F92201 Neuilly-sur-Seine Cedex France	Performing Organization Number(s)

Sponsoring/Monitoring Agency Name(s) and Address(es)	Monitoring Agency Acronym
	Monitoring Agency Report Number(s)

Distribution/Availability Statement Approved for public release, distribution unlimited

Supplementary Notes

Abstract

Subject Terms

Document Classification unclassified	Classification of SF298 unclassified
Classification of Abstract unclassified	Limitation of Abstract unlimited

Number of Pages 300	

This page has been deliberately left blank

————————————

Page intentionnellement blanche

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION

BP 25, 7 RUE ANCELLE, F-92201 NEUILLY-SUR-SEINE CEDEX, FRANCE

*Note: There is an Addendum to this publication
containing the PowerPoint presentations.*

RTO MEETING PROCEEDINGS 49

New Information Processing Techniques for Military Systems

(les Nouvelles techniques de traitement de l'information
pour les systèmes militaires)

*Papers presented at the Information Systems Technology Panel (IST) Symposium held in Istanbul,
Turkey, 9-11 October 2000.*

The Research and Technology Organization (RTO) of NATO

RTO is the single focus in NATO for Defence Research and Technology activities. Its mission is to conduct and promote cooperative research and information exchange. The objective is to support the development and effective use of national defence research and technology and to meet the military needs of the Alliance, to maintain a technological lead, and to provide advice to NATO and national decision makers. The RTO performs its mission with the support of an extensive network of national experts. It also ensures effective coordination with other NATO bodies involved in R&T activities.

RTO reports both to the Military Committee of NATO and to the Conference of National Armament Directors. It comprises a Research and Technology Board (RTB) as the highest level of national representation and the Research and Technology Agency (RTA), a dedicated staff with its headquarters in Neuilly, near Paris, France. In order to facilitate contacts with the military users and other NATO activities, a small part of the RTA staff is located in NATO Headquarters in Brussels. The Brussels staff also coordinates RTO's cooperation with nations in Middle and Eastern Europe, to which RTO attaches particular importance especially as working together in the field of research is one of the more promising areas of initial cooperation.

The total spectrum of R&T activities is covered by the following 7 bodies:

- AVT   Applied Vehicle Technology Panel
- HFM   Human Factors and Medicine Panel
- IST   Information Systems Technology Panel
- NMSG   NATO Modelling and Simulation Group
- SAS   Studies, Analysis and Simulation Panel
- SCI   Systems Concepts and Integration Panel
- SET   Sensors and Electronics Technology Panel

These bodies are made up of national representatives as well as generally recognised 'world class' scientists. They also provide a communication link to military users and other NATO bodies. RTO's scientific and technological work is carried out by Technical Teams, created for specific activities and with a specific duration. Such Technical Teams can organise workshops, symposia, field trials, lecture series and training courses. An important function of these Technical Teams is to ensure the continuity of the expert networks.

RTO builds upon earlier cooperation in defence research and technology as set-up under the Advisory Group for Aerospace Research and Development (AGARD) and the Defence Research Group (DRG). AGARD and the DRG share common roots in that they were both established at the initiative of Dr Theodore von Kármán, a leading aerospace scientist, who early on recognised the importance of scientific support for the Allied Armed Forces. RTO is capitalising on these common roots in order to provide the Alliance and the NATO nations with a strong scientific and technological basis that will guarantee a solid base for the future.

<div align="center">

The content of this publication has been reproduced
directly from material supplied by RTO or the authors.

</div>

New Information Processing Techniques
for Military Systems

(RTO MP-049 / IST-017)

Executive Summary

Information processing is a key factor for many military systems. Recent operations in Gulf-war, in Bosnia, and in Kosovo has made this requirement even more obvious. Advances in sensing and information processing/distribution technologies will enable highly innovative system concepts for achieving improvements in military mission capabilities. This Symposium essentially dealt with applications of new, promising, and unprecedented information processing techniques for military systems, where special emphasis was given to technology transfer from the commercial area. Military effectiveness requires the ability to acquire and process information in real time and to communicate this effectively on a wide front. With increasing NATO responsibilities in joint military operations involving many different national Communications and Information System (CIS) environments, the need for a unified approach to support information / data transfer services becomes more crucial. Topics addressed by the symposium were:

- Information filtering and information fusion
- Applications of soft computing (neural networks, fuzzy logic, genetic algorithms)
- Expert systems (including real-time aspects) and knowledge-based decision support
- Machine intelligence applied to future autonomous systems
- Techniques for efficient information management (including data dissemination)
- Situation analysis (incl. real-time interpretation of large amounts of battlefield information)
- Processing demands in information security (monitoring of systems and networks, incl. firewalls and intrusion detection)
- Modelling and simulation, visualisation and virtual environments
- Innovative architectures in the field of c2 systems
- Design methods for information systems and command centres
- Battlefield digitisation concepts

THE SYMPOSIUM
The three days of the symposium included 29 papers and two keynote addresses, which provide a good basis for further development. The first keynote address entitled "Autonomous Systems" (Dr. B. C. Williams, US) covered information processing onboard unmanned spacecraft. The second keynote address entitled "Model-Based Design of Information-Rich Command Organisations" (Dr. D. Serfaty, US) covered a description of team structure design based upon the team's mission. The feed back from 140 delegates was, in general, positive with the vast majority considering the contents relevant. The purpose of the symposium was to exchange information on state-of-the-art and state-of-the-practice in information processing techniques as applied to military systems. The scope of the symposium was intentionally very broad and was organised into multiple sessions.

SESSION I – INFORMATION SYSTEMS AND TECHNIQUES I
The initial two sessions consisted of eleven papers with various aspects of information systems and techniques that covered the spectrum from future revolutionary technology such as the quantum computer to fielded systems.

SESSION II – INFORMATION SYSTEMS AND TECHNIQUES II
A paper in this continuing session entitled "Battlefield Digitisation" (Gibson and White) addressed the broader range of requirements to achieve information security which, in addition to technology development and exploitation, also must consider changes to doctrine, command processes, organisations, user-requirements specification, architecture definition, procurement, training, and operational use.

SESSION Ill – SECURITY AND RELIABILITY
There were four papers in this session. The first paper (Serb and Patriciu) addressed the reliability of command and control systems in terms of fault tolerance capabilities provided by a cluster of networked nodes capable of fault detection and automatic reconfiguration such that the systems continue operation subsequent to the fault.

SESSION IV – COMMUNICATIONS
The communications session consisted of five papers. The first two papers (Berni and Mozzone; Rice) addressed wireless networks in the undersea environment. The undersea environment is particularly difficult for networks and typically involves one or more battery powered buoys to provide a gateway to the terrestrial node. Bandwidth limitations of military ships have resulted in much research into effective data compression algorithms, which enable de-compression by the receiver without loss of data quality.

SESSION V – DETECTION, FUSION, DECISION SUPPORT
Six papers were presented during this session beginning with an overview of information fusion (Whitaker). The major challenge to future Command and Control systems was highlighted. The Command and Control process is described by the Observe Orient Decide Act (OODA) Loop. Processing and fusion of electro-optical (EO) data was the subject of the next paper (Davies).

SESSION VI – VIRTUAL REALITY AND HUMAN-COMPUTER INTERFACE
This session included four papers. The first paper (Varga, McQueen, and Rossi) described the United Kingdom Master Battle Planner, which is an Air Tasking Order planning tool providing for visualisation of the scenario including showing the mission in motion.

CONCLUSIONS
The defense community has the opportunity to contribute to information technology through establishment of metrics and a theoretical/mathematical basis for information systems. This is an area where the Information Systems Technology Panel might consider an activity that would deal with military information performance requirements, modelling, simulation, and analysis.

les Nouvelles techniques de traitement de l'information pour les systèmes militaires

(RTO MP-049 / IST-017)

Synthèse

Le traitement de l'information est un élément clé pour de nombreux systèmes militaires. Les récentes opérations, guerre du Golfe, Bosnie et Kosovo n'ont fait que souligner son utilité. Les progrès réalisés dans les domaines des technologies du traitement et de la diffusion de l'information ainsi que des capteurs permettront d'élaborer des concepts de systèmes très novateurs dans le but d'améliorer l'efficacité des missions militaires. Ce symposium a traité essentiellement de la mise en œuvre de nouvelles techniques de traitement de l'information pour systèmes militaires , prometteuses et sans précédent,; avec une attention particulière pour les transferts de technologie avec le secteur commercial. L'efficacité militaire dépend de la capacité d'acquérir et de traiter des informations en temps réel et de les communiquer avec efficacité à une grande partie de la chaîne de commandement. Avec une limplication toujours plus grande de l'OTAN dans des opérations interarmées mettant en jeu différents environnements nationaux de systèmes de communications et d'information (CIS), le besoin d'une approche unifiée pour le soutien des services de transfert de données et de renseignements se fait de plus en plus sentir. Les sujets abordés lors du symposium étaient les suivants :

- Le filtrage et le fusionnement des données
- Les applications de l'ingénierie des logiciels (les réseaux neuraux, la logique floue, les algorithmes génétiques)
- Les systèmes experts (y compris les aspects temps réel) et les aides à la décision basées sur l'expérience
- L'intelligence machine appliquée aux futurs systèmes autonomes
- Les techniques de gestion efficace des informations (y compris la diffusion des données)
- L'analyse de la situation (y compris l'interprétation en temps réel de grands volumes de données du champ de bataille)
- Les exigences en matière de traitement pour la sécurité de l'information (le contrôle des systèmes et réseaux, y compris les pare feu et les systèmes de détection d'intrus)
- La modélisation et la simulation, la visualisation et les environnements virtuels
- Les architectures novatrices pour systèmes C2
- Les méthodes de conception pour systèmes d'information et centres de commandement
- Les concepts de numérisation du champ de bataille

LE SYMPOSIUM

En tout, 29 communications et 2 discours d'ouverture ont été présentés lors des 3 journées de la conférence, ce qui constitue une bonne base pour de futurs travaux. Le premier discours d'ouverture, intitulé "Les systèmes autonomes" (par le Dr.B.C.Williams, US) concernait le traitement de l'information à bord de véhicules spatiaux non habités. Le deuxième, intitulé "La conception, à base de modèles, d'organismes de commandement riches en information" (par le Dr. D.Serfaty, US) était une description de la conception d'une structure pour une collaboration optimale basée sur la mission de l'équipe. Les commentaires des 140 délégués ont été, dans l'ensemble, positifs une grande majorité considérant les communications dignes d'intérêt. Le symposium a eu pour objectif d'échanger des informations sur l'état actuel des connaissances et des pratiques dans le domaine des techniques de

traitement de l'information au sein des systèmes militaires. Le portée du symposium était délibérément très large et la réunion a été divisée en un certain nombre de sessions.

SESSION I – SYSTEMES ET TECHNIQUES DE TRAITEMENT DE L'INFORMATION (I)
Les deux premières sessions étaient composées de 11 communications, portant sur différents aspects des systèmes et des techniques de traitement de l'information, allant des futures technologies révolutionnaires telles que l'ordinateur quantique aux systèmes déployés.

SESSION II – SYSTEMES ET TECHNIQUES DE TRAITEMENT DE L'INFORMATION (II)
L'une des communications présentées lors de cette session, "La numérisation du champ de bataille" (par Ms. Gibson & White) a examiné l'éventail plus large des spécifications en matière de sécurité de l'information, lequel, en plus du développement et de l'exploitation des technologies, devra aussi incorporer des changements concernant la doctrine, les processus de commandement, les organisations, la spécification des besoins des utilisateurs, la définition des architectures, l'approvisionnement, la formation et e la mise en œuvre opérationnelle.

SESSION III – SECURITE ET FIABILITE
Quatre communications ont été présentées. La première communication (par Ms. Serb et Patriciu) a examiné la fiabilité des systèmes de commandement et de contrôle du point de vue de la tolérance aux pannes offerte par une grappe de nœuds en réseau capables de détecter les erreurs et de reconfigurer les systèmes automatiquement afin que ceux-ci puissent reprendre leur activité rapidement après l'erreur.

SESSION IV – COMMUNICATIONS
La session sur les communications a permis la présentation de 5 communications. Les deux premières (par Ms. Berni et Mozzone ; puis M. Rice) concernaient les réseaux radio en environnement sous-marin. L'environnement sous-marin pose des problèmes particuliers pour les réseaux et généralement nécessite le déploiement d'une ou de plusieurs bouées fonctionnant sur piles pour fournir une passerelle au nœud terrestre. Les conditions limitant la bande passante utilisable par les bâtiments de guerre ont conduit à de nombreux travaux de recherche sur les algorithmes de compression de données, destinés à permettre la décompression par le récepteur sans perte de qualité.

SESSION V – DETECTION, FUSIONNEMENT, AIDES A LA PRISE DE DECISIONS
Six communications ont été présentées lors de cette session, commençant par un tour d'horizon du fusionnement des données (par M. Whitaker). La session a permis de mettre en relief les enjeux pour les futurs systèmes de commandement et contrôle. Le processus de commandement et contrôle est décrit par la boucle OODA (observer, orienter, décider, agir). Le traitement et le fusionnement des données électro-optiques (EO) a fait l'objet de la communication suivante (par M. Davies).

SESSION VI – LA REALITE VIRTUELLE ET L'INTERFACE HOMME - MACHINE
Cette session a été composée de 4 communications. La première, (par Ms. Varga, McQueen et Rossi) a présenté le Système Central de Planification de l'Engagement du Royaume-Uni, qui est un outil de planification des ordres de mission aérienne incorporant la visualisation du scénario, et montrant des images de la mission en cours d'exécution.

CONCLUSIONS
Les spécialistes de la défense ont la possibilité de contribuer aux technologies de l'information en établissant des paramètres et en jetant les bases théoriques/mathématiques de futurs systèmes d'information. Il s'agit d'un domaine pour lequel la commission sur la technologie des systèmes d'information pourrait envisager la création d'une activité portant sur les spécifications des performances, la modélisation, la simulation et l'analyse de l'information militaire.

Contents

---

† Paper not available at time of production.

SESSION III: SECURITY AND RELIABILITY
Chairman: Dr W. STEIN (GE)

SESSION IV: COMMUNICATIONS
Chairman: Dr E. ANGELIDIS (GR)

SESSION V: DETECTION, FUSION, DECISION SUPPORT
Chairman: Dr F. INCE (TU)

SESSION VI: VIRTUAL REALITY AND HUMAN-COMPUTER INTERFACE
Chairman: Dr. R. JACQUART (FR)

Theme

Information processing is a key factor for many military systems. Recent operations in Gulf-war, in Bosnia, and in Kosovo have made this requirement even more obvious. Advances in hardware and software technologies for sensing, integrating, processing, and distributing information will enable highly innovative system concepts for achieving improvements in military mission capabilities. This Symposium will essentially deal with applications of new, promising, and unprecedented information processing techniques for military systems challenges of the next 5 to 10 years, where special emphasis is given to technology transfer from and to the commercial area.

Today's and future military effectiveness requires the ability to acquire and process information in real time as well as to distribute and communicate this effectively on a wide front. This is also a very important requirement of operations as demonstrated by the recent conflicts. With increasing NATO responsibilities in joint military operations involving many different national Communications and Information System (CIS) environments, the need for a unified approach to support information / data transfer services becomes more crucial. This symposium is intended to cover a large but important operational area by viewing relevant aspects of military systems: processing performance; scalability, flexibility, and adaptivity of processing; system characteristics like robustness, reliability, and security; and other relevant factors.

TOPICS TO BE COVERED:

1. The various Processing Demands of the Command and Control Cycle (e.g., sensing, integrating/fusing, and distributing information on extended/global battlefields)

2. Situation Analysis (incl. real-time interpretation of large amounts of battlefield information)

3. Architectures and Processing Demands in the field of Command and Control Systems (especially Distributed Systems)

4. Applications of Soft Computing (neural networks, fuzzy logic, genetic algorithms)

5. Machine Reasoning (including real-time aspects) and Knowledge-Based Decision Support

6. Information Architectures and Processing Demands of Future Mobile Autonomous Systems

7. New Techniques for Efficient Information Management (including Data Dissemination)

8. Processing Demands of Information Assurance (Cryptography; Monitoring Systems and Networks, incl. Firewalls and Intrusion Detection Systems)

9. Information Processing in Enabling Technologies (e.g., Intelligent Collaboration, Modeling and Simulation, Visualization, and Virtual Environments)

10. Processing Demands of Image Understanding and Speech Technology / Human Language Systems

Thème

Le traitement de l'information est un élément clé de bon nombre de systèmes militaires. Les récentes opérations pendant la guerre du Golfe et en Bosnie n'ont fait que souligner l'utilité de ce moyen. Les progrès réalisés dans les domaines des technologies du matériel et des logiciels pour la détection, l'intégration, le traitement et la diffusion des données permettront prochainement d'élaborer des concepts de systèmes très novateurs dans le but d'améliorer l'efficacité des missions militaires. Ce symposium traitera essentiellement de la mise en oeuvre de nouvelles techniques, prometteuses et sans précédent, de traitement de l'information pour systèmes militaires, au cours des prochaines cinq à dix années; l'accent étant mis sur les transferts de technologie avec le monde du commerce.

Désormais, l'efficacité militaire dépendra de sa capacité à acquérir et à assimiler les informations en temps réel et à les communiquer avec efficacité à tous les niveaux de commandement. Cette capacité est aussi l'une des principales conditions requises pour la conduite des opérations, comme en témoignent les conflits récents. Avec la multiplication des responsabilités de l'OTAN dans des opérations interarmées mettant en jeu plusieurs environnements de communications et d'information (CIS), le besoin d'une approche unifiée du soutien des services de transfert de données et de renseignements se fait de plus en plus sentir. Ce symposium couvrira un grand domaine opérationnel qui est d'une grande importance, en traitant des aspects suivants des systèmes militaires : les performances en traitement, la variabilité d'échelle, la souplesse d'emploi, l'adaptativité du traitement, les caractéristiques des systèmes telles que la robustesse, la fiabilité, la sécurité et tout autre facteur approprié.

SUJETS A TRAITER :

1. Les différentes exigences du cycle de commandement et contrôle (par exemple, la détection, l'intégration/fusion et la dissémination des informations à travers des champs de bataille étendus/globaux)

2. L'analyse de la situation (y compris l'interprétation en temps réel de grands volumes de données du champ de bataille)

3. Les exigences en matière d'architectures et de traitement dans le domaine des systèmes de commandement et contrôle (en particulier les systèmes répartis)

4. Les applications de l'ingénierie des logiciels (les réseaux neuronaux, la logique floue, les algorithmes génétiques)

5. Le raisonnement machine (y compris les aspects temps réel) et les aides à la décision basées sur les connaissances

6. Les exigences en matière d'architectures d'information et de traitement des futurs systèmes mobiles autonomes

7. Les nouvelles techniques de gestion efficace des informations (y compris la diffusion des données)

8. Les exigences en matière de sécurité de l'information (la cryptographie, les systèmes et réseaux de surveillance, y compris les pare-feux et les systèmes de détection d'intrus)

9. Le traitement de l'information dans les technologies habilitantes (par exemple, la collaboration intelligente, la modélisation et la simulation, la visualisation et les environnements virtuels)

10. Les exigences en matière de traitement des technologies vocales et de compréhension d'image et de systèmes de traitement du langage naturel

Information Systems Technology Panel

CHAIRMAN

Dr M VANT
Deputy Director General
Defence Research Establishment Ottawa
Dept of National Defence
3701 Carling Ave
OTTAWA, ON, K1A 0K2, CANADA

DEPUTY CHAIRMAN

Dr R JACQUART
Directeur du DTIM
ONERA/CERT
BP 4025
31055 TOULOUSE CEDEX 4, FRANCE

TECHNICAL PROGRAMME COMMITTEE

CHAIRMAN:	Dr J GROSCHE	GE
MEMBERS:	Dr R JACQUART	FR
	Dr W STEIN	GE
	Dr E ANGELIDIS	GR
	Dr F INCE	TU
	Prof A MILLER	US

PANEL EXECUTIVE

From Europe:

RTA-OTAN
Lt-Col A GOUAY, FAF
IST Executive
BP 25, 7 Rue Ancelle
F-92201 NEUILLY SUR SEINE CEDEX, FRANCE

Telephone: 33-1-5561 2280/82 - Telefax: 33-1-5561 2298/99

From the USA or CANADA:

RTA-NATO
Attention: IST Executive
PSC 116
APO AE 09777

HOST NATION LOCAL COORDINATOR

Dr F INCE
ISIK University
80670 MASLAK, ISTANBUL

Tel: (90) 212 286 2960 Ext 2251
FAX: (90) 212 285 2875

ACKNOWLEDGEMENTS/REMERCIEMENTS

Technical Evaluation Report

by

Judy L. Miller
Charles Stark Draper Laboratory, Inc.
555 Technology Square
Cambridge, MA 02139
UNITED STATES
jmiller@draper.com

INTRODUCTION

The 5th symposium of the Information Systems Technology Panel was entitled "New Information Processing Techniques for Military Systems." The Program Chairman was Dr. J. GROSCHE (GE). The Program Committee also included:

Dr. R. JACQUART (FR)
Dr. W. STEIN (GE)
Dr. A. ANGELIDIS (GR)
Dr. F. INCE (TU)
Prof. A. MILLER (US)

The welcoming address was presented by Brig Gen Cemal Alagoz, Director of Technical Services at the Turkish MoD, who presented three eras of war: (1) the land acquisition era, (2) the industrial era ending with weapons of mass destruction, and (3) the current information technology era. He addressed the need for information technology to compensate for the changing nature of recent threats, which have been low scale, operations other than war (OOTW), and counter terrorism. The nature of these threats effectively precludes the collection of threat data over a long period of time prior to engagement and thus demands the rapid, collection, exploitation, and dissemination of tactical information. Brig Gen Cemal Alagoz also described the changing military operations with demands for increased deployability, flexibility, interoperability, low collateral damage, and precision targeting. His welcoming address effectively established the challenges that information technologies and processes need to overcome prior to meeting military users' expectations.

There were two keynote addresses during the symposium. The first keynote address entitled "Autonomous Systems" was presented by Dr. B. C. Williams, Space Systems Laboratory & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge MA, US. Dr. Williams addressed information processing onboard unmanned spacecraft. This is important due to long communication delays from the manned ground stations to planetary spacecraft that can preclude effective response to unplanned situations. He described new search methods using onboard intelligence to enable deductive, automated reasoning to achieve autonomous control of unmanned spacecraft.

The second keynote address entitled "Model-Based Design of Information-Rich Command Organizations" was presented by Dr. D. Serfaty, Founder/President of Aptima, Woburn MA, US. Dr. Serfaty presented a very interesting description of team structure design based upon the team's mission (goals and tasks) to improve performance. Moreover, the design effort is based upon the application of a mathematical representation of the mission tasks and the application of optimization algorithms. Since information technologies in general and network-centric warfare in particular are human-centric and collaborative in nature, the idea of designing a team based upon the mission, assigning tasks, and then determining the information flows and communication paths required seems most logical. The concept of team design is especially interesting within the context of increasing coalition military operations, where new teams are formed without the benefit of prior evolution.

Additionally, the team design methodology is based upon formal algorithms driven by empirical data from subject matter experts. As presented, the research and development is currently in progress to develop and test a model-driven computer-based tool to represent and analyze information flow within the context of a complex theater warfare system having equal emphasis upon an accurate representation of the human element. There are three orthogonal axes considered: (a) organization, (b) technology, and (c) processes. For military information systems, the consideration of technology alone is necessary but not sufficient. The goal of this work is to develop a tool soundly based upon theoretical constructs, which can provide the analysis and design of human-system organizational architectures. This work acknowledges that the core process is human-centered decision making and the role of information as provided by software and hardware elements is to serve as support to the decision process. Dr. Serfaty indicated such organizational engineering has evolved from cybernetics research conducted in the 1980's at Massachusetts Institute of Technology (US). There is a definite need for development of information performance metrics, models, analysis tools for both performance and sensitivity assessment, and perhaps even the development of information modes correlated with military operations. The work presented by Dr. Serfaty represents a good step in this direction.

THEME

Information processing is a key factor for many military systems. Recent operations in the Gulf War, in Bosnia, and in Kosovo have made this requirement even more obvious. Advances in hardware and software technologies for sensing, integrating, processing, and distributing information will enable highly innovative system concepts for achieving improvements in military mission capabilities. This symposium was intended to address applications of new, promising, and unprecedented information processing techniques for military systems challenges of the next 5 to 10 years, with special emphasis upon technology transfer to and from the commercial area.

Today's and future military effectiveness requires the ability to acquire and process information in real time as well as to distribute and communicate this effectively on a wide front. This is also a very important requirement of operations as demonstrated by the recent conflicts. With increasing NATO responsibilities in joint military operations involving many different national Communications and Information System (CIS) environments, the need for a unified approach to support information/data transfer services becomes more crucial. This symposium covered a large but important operational area by examining relevant aspects of military systems: processing performance; scalability, flexibility, and adaptivity of processing; system characteristics like robustness, reliability, and security; and other relevant factors.

PURPOSE AND SCOPE

The purpose of the symposium was to exchange information on state-of-the-art and state-of-the-practice in information processing techniques as applied to military systems. The scope of the symposium was intentionally very broad and was organized into multiple sessions addressing information systems and techniques; security and reliability; communications; detection, fusion, decision support; and virtual reality and human-computer interface. There were twenty-nine papers presented over the two and a half day symposium.

EVALUATION

The topic of information processing for military operations, especially for NATO coalition operations, is an extremely important and timely topic. Information superiority is perceived as a countermeasure to asymmetric threats, low intensity conflicts, and military operations other than war. The transition from platform-centric warfare to network-centric warfare requires the exchange of information across computers communicating with wireless communications networks such as aircraft flying local area networks (FLAN). In fact, the concept of network-centric warfare is a derivative of network-centric computing and attempts to exploit information technology that has evolved

through private sector advancement of internet technology and standards such as Hyper Text Markup Language (HTML), Web Browsers (e.g., Netscape Navigator and Microsoft Internet Explorer), TCP/IP, search engines, and Java. Such developments have facilitated the interaction of computers with different operating systems. The exploitation of network-centric computing by the commercial sector to establish a competitive edge is often viewed by the military leadership as directly transferable to gaining a competitive edge in warfare. As a result, military research and development often is focussed upon the system integration of commercial-of-the-shelf (COTS) software and hardware into existing military Command and Control systems. In some cases, this has modfied the traditional development process of requirements definition followed by technology trades studies. With the use of COTS products, there is a greater emphasis upon the integration requirements since the COTS product capabilities are already pre-defined by the vendor.

A major development observed was the increasing use of internet-related techniques and client-server computer architectures. A second major trend observed was increased use of computer graphics and multimedia displays for visualization to the user.

Several papers recognized that information systems are inherently "human-centric" and involve processes such as the Command and Control Observe Orient Decide Act (OODA) loop that must be taken into account in addition to inserting new technology and information techniques.

This symposium has provided the NATO technical community the opportunity to exchange information about many on-going development projects. In general, current military systems are still attempting to catch-up to the commercial world. In the future, military systems can be expected to focus more upon their unique requirements, such as precision weapons targeting, that will require specialized military technological solutions that will not be provided by COTS products. As these requirements emerge, future symposia can be expected to include these topics.

The symposium achieved its objective of bringing together in a timely fashion, many leading engineers in this particular field. Both the Program Committee and National Hosts should be congratulated for their outstanding efforts in arranging the symposium. In particular, the symposium website was a great success.

SESSION I– INFORMATION SYSTEMS AND TECHNIQUES I

The initial two sessions consisted of eleven papers that covered various aspects of information systems and techniques. This set of papers was extremely diverse covering the spectrum from future revolutionary technology such as the quantum computer (Krutar) to fielded information systems for logistics (Burival and Reha).

The first paper (Driesenaar) discussed message standards for interoperability. In particular, the Allied Data Publications 3 (AdatP-3) documentation of the NATO Message Text Formatting System was discussed. Limitations with this system, especially with regard to automatic message distribution, were reported. It was recommended to combine AdatP-3 with the Army Tactical Command & Control Information System (ATCCIS) to provide the best of both standards in a unified approach.

The next two papers (Choenni and Bruggeman; Krusche and Tolk) dealt with the potentials of multi-media databases, data mining, and common shared data models to provide information that can support the command and control human decision making process. The need for advanced query handling techniques to provide timely response with the information of interest was identified and several techniques were described.

The fourth paper (Wunder) discussed trends in computers and wireless technology. A particular architecture comprised of COTS products was described. A client-server architecture was proposed with mobile clients including laptop computers, palmtops, and mobile telephones accessing the server.

The fifth paper (Krutar) explored the possibilities of quantum computers, which are devices that exploit qubits. In the case of digital computers, each bit is represented as a 0 or 1. In the case of quantum computers, the superposition of these states is also possible so a qubit can be a 0 and a 1 simultaneously. The quantum computer offers outstanding performance improvement for certain classes of problems. For example, Peter Shor of AT&T Laboratories in Florham Park developed a factoring algorithm in 1994 that, if executing on a quantum computer, would enable decryption of the commonly used computer security algorithms (public key cryptography) which relies upon the long time periods required by current digital computers to factor large integers.

The papers in this session presented standards for exchanging data, databases and query handling techniques, information architecture, and a thought-provoking revolutionary computer technology.

SESSION II – INFORMATION SYSTEMS AND TECHNIQUES II

The first paper in this continuing session (Gibson and White) addressed the broader range of requirements to achieve information security which, in addition to technology development and exploitation, also must consider changes to doctrine, command processes, organizations, user-requirements specification, architecture definition, procurement, training, and operational use. The potential for information and communications systems to actually exacerbate command and control problems is noted and therefore the emphasis upon inclusion of cognitive science and technology as a means to achieving cognitive dominance is considered.

The next set of three papers described various information-related systems that were developed during the 1990's. The first of the three papers (Burival and Reha) described an information system which was developed for integrated logistics management. This system supports the logistics life cycle from material cataloguing and logistics requirements through recordkeeping and inventory management to distribution and equipment maintenance. The second of the three papers (Hofmann and Hofmann) described the

main features for combat simulation models based upon robust and highly efficient rule sets. Both spatial and procedural templates were employed. Possible applications include training and a decision aid during actual operations. The third of the three papers, (Snajder and Horak) describes the systems engineering and system architecture design and development of a command and control system for ground forces.

The next paper (Reinhardt) addressed Geographic Information Systems (GIS) access via the World Wide Web. A comparison of different data transfer techniques for web-browser access to GIS data is made. Three dimensional (3D) visualization of GIS databases is described.

The last paper of the session (Voas) described an approach to assessing software survivability by injecting software faults and observing whether the software continues to operate in the presence of these faults. The motivation for this approach is the absence of metrics for quantifying the extent to which software in the information system is trustworthy for the users. Of particular relevance is the Interface Propagation Analysis, which can determine the effect of third party (e.g., COTS) software component failures or anomalous behavior on the remainder of the system. With the increasing introduction of COTS and modified COTS software in military information systems, lack of insight into proprietary software precludes many previously employed methods of system level testing.

The papers in this session covered a wide variety of systems and techniques. Many of the papers described an existing or proposed system. Several papers surveyed techniques currently being used on the internet, which have applicability to military operations.

SESSION III – SECURITY AND RELIABILITY

There were four papers in this session following the second keynote address. The first paper (Serb and Patriciu) addressed the reliability of command and control systems in terms of fault tolerance capabilities provided by a cluster of networked nodes capable of fault detection and automatic reconfiguration such that the systems

continue operation subsequent to the fault. The point was made that reliability applies to simulation and modeling systems employed for training and military exercises as well as for tactical command and control systems.

The next paper (Bieber and Siron) experimented with the application of COTS and Security Components Off The Shelf (SCOTS) augmented with a role-based access control component. One of the impacts that was particularly interesting was the need for a security officer to administer the roles. This suggests that organization structure and task loadings need to be considered in the design of secure and reliable information systems. This is an example of where the team design described by Dr. Serfaty would apply to development of new team members and roles.

The third paper (Patriciu and Serb) described the main design issues associated with the application of the worldwide Public Key Infrastructure (PKI) currently used in the business world. Such a system could enable secure information exchange using the Internet and obviating the need for special networks.

The final paper in this session (Pretschner, Slotsch, and Stauner) addressed reliability through testing and the development of automated tools to improve the testing process.

This is a most important topic for military applications. Three of the papers addressed reliability while only one paper considered information security. Although the need for information security is widely recognized, much remains to be done in the development of techniques and technologies specific to military applications.

SESSION IV – COMMUNICATIONS

The communications session consisted of five papers. The first two papers (Berni and Mozzone; Rice) addressed wireless networks in the undersea environment. The undersea environment is particularly difficult for networks and typically involves one or more battery powered buoys to provide a gateway to the terrestrial node. Bandwidth limitations of military ships have resulted in much research

into effective data compression algorithms, which enable de-compression by the receiver without loss of data quality. The third paper (Kondoz and Palaz) addressed a similar coding effort applied to speech coding algorithms. The final paper in this session (Trangeled, Vink, and Berni) described a series of naval experiments for exchanging and fusing data from multiple maritime vessels collecting data to support NATO maritime operational requirements for anti-submarine warfare, mine counter measures, and military oceanography.

In summary, this session provided a good reminder that there will always be information network users with substantially less bandwidth than others. Thus, techniques to improve the efficiency of available bandwidth will remain important.

SESSION V – DETECTION, FUSION, DECISION SUPPORT

Six papers were presented during this session beginning with an overview of information fusion (Whitaker). The major challenge to future Command and Control systems was highlighted. The Command and Control process is described by the Observe Orient Decide Act (OODA) Loop. Since in military operations the adversary will have its own OODA Loop, the challenge is to have a better and faster OODA Loop than your opponent.

Processing and fusion of electro-optical (EO) data was the subject of the next paper (Davies). An EO model was developed to generate EO data for analysis of the performance of fusion algorithms. The model was employed to determine the effects of update rate, field of regard, false alarm rate, and other parameters on the quality of the fused tactical picture.

The next paper (Harrison) describes the application of heuristic-based optimization techniques to military decision making processes such as planning for convoy movement. Advantages of this approach are faster planning and the ability to provide the planner with an estimate in the measure of confidence of the plan. This technique can be applied to other applications of a similar planning class such as Sensor Collection Management and Radio Frequency Allocation.

The next paper (Maas, Wynia, Stavnem, and Houtsma) considers the factors contributing to information overload with shipborne Command Information Centers (CIC) and especially within the context of future reduced ship manning plans. Information-related factors include: uncertainty, quantity, ambiguity, novelty, and the level of abstraction. Additionally, three human-related factors include: time constraints, workload, and multi-channel information flows. What is meant by multi-channel information flow is the operator confrontation of not only information displayed on a tactical screen, but also voice information provided on the headphone, and text information provided by e-mail and paper. To solve the operator information overload problem, a concept based upon predefined templates for operator information needs and presentation formats for each task followed by information filtering and adaptive information control using workload prediction and measurement of task progress was developed.

The next paper (Stroscher and Schneider) describes a Bayesian approach to multi-sensor data fusion for the purpose of identification of friend, foe, or neutral. A prototype was developed and results reported. The last paper in this session (Khashman) considers the detection of military targets from images through combination of the three fields of scale space analysis, edge detection, and neural networks.

In summary, various techniques and algorithms are being considered in the broad area of sensor detection, multi-sensor data fusion, and decision aids, which do not create information overloading of the user.

SESSION VI – VIRTUAL REALITY AND HUMAN-COMPUTER INTERFACE

This session included four papers. The first paper (Varga, McQueen, and Rossi) described the United Kingdom Master Battle Planner, which is an Air Tasking Order planning tool intended to provide an adaptive, decision oriented visualization environment for UK Joint Force Commanders. The Master Battle Planner is a prototype with map-based visualization in contrast with spreadsheet displays employed by other systems such as the United States Theater Battle Management Core System. The Master Battle Planner provides for visualization of the scenario including showing the mission in motion.

The next paper (Julier, Baillot, Lanzaforta, Brown, and Rosenblum) described the Battlefield Augmented Reality System to provide mobile augmented reality for dismounted warfighters. Using a wearable computer, a wireless communications network, and a tracked see-through head mounted display, the actual user's field of view would be augmented with information received from a remote transmitting location. This information would be aligned with the true information and might include augmenting a building with a wireframe plan of the interior. Clearly, precision registration of the real view with the augmented information is a formidable challenge.

The next paper (Matzke and Schmorrow) provided a tutorial on virtual environments and described the architecture of the various components. Examples of virtual environments were given. Future virtual environments are expected to integrate visual, audio, and haptic (the sense of touch) for creation of increasing realism in the simulation environment.

The final paper (Guleyupoglu and Ng) took the virtual environment one step further to define virtual engineering as the engineering process performed primarily within the virtual environment. Virtual engineering can be employed throughout the design and development life cycle of a system and can support testing, evaluation, and training, as well. Virtual engineering was applied to virtual prototyping of a ship Command Information Center which obviated the development of a costly physical mockup and facilitated collaborative development among geographically dispersed team members.

In summary, there are enormous benefits to using virtual environments for training especially in combination with constructive environments and live play. The virtual environment enables the generation of correlated sensor data of threats that does not exist in the real world during non-combat operations. These synthetic environments enable military Command and Control processes to be stimulated and provide realistic training environments for the Commander and Staff.

REACTION OF THE SYMPOSIUM PARTICIPANTS

Standard questionnaires were provided to the symposium participants and this TER author read about 30 responses. The major criticism, which is shared by this author, was that the topics did not provide enough technical material and details. Perhaps given the symposium title, it is not surprising that many, if not most, of the papers described systems that had been developed, were in development, or were recommended for development. The fundamental technologies could be the theme of a future symposium. However, as pointed out during several of the papers, information technologies are fundamentally a human-centric endeavor!

The author also observed the unusually high attendance throughout the duration of the symposium. It can only be indicative of the high level of interest on the part of the participants that the meeting room remained so full throughout the entirety of the symposium.

CONCLUSIONS AND RECOMMENDATIONS

As stated in the Purpose and Scope section, the symposium was intentionally very broad and many papers described tools and techniques available in the commercial arena that can be applied to military computer systems. These tools include web browsers, search engines, data mining techniques, network protocols and the like. However, the emphasis to date has been on information access and display. There is a strong need for military applications to address information content.

In the United States, the National Institute of Standards and Technology (NIST) is considering how to define metrics, which can be applied to information. For military applications such as targeting, the target data accuracy, precision, and senescence are just as important as the access speed. Metrics are also important for the end-to-end test and evaluation process. Another area of research and development opportunity is the simulation and modeling of information. Most information systems are developed and, if performance is established at all, it is established after the fact. What if we could predict the performance of a particular design, compare designs to determine differences in performance and even analyze sensitivities to parameters?

In conclusion, as Brig Gen Cemal Alagoz stated in his welcoming address, we have truly entered the information technology era of war. While it is important to apply existing commercial technology to military applications, it is equally important for the defense community to recognize the need for specialized information requirements imposed by warfare. The defense community has the opportunity to regain the lead in information technology through establishment of metrics and a theoretical/mathematical basis for information systems. This is an area where the Information Systems Technology Panel might consider a symposium, or task group, that would deal exclusively with military information performance requirements, metrics, modeling, simulation, and analysis. Finally, information systems and technologies will continue to remain an important topic for NATO countries especially as the nature of military operations continues with the establishment of *ad hoc* coalitions to respond to low intensity conflicts.

This page has been deliberately left blank
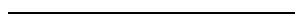
Page intentionnellement blanche

Information Exchange
in support of
C2-Interoperability

Freek N. Driesenaar, MSc
Scientist
TNO Phyisics and Electronics Laboratory
P.O. Box 96864
2509 JG The Hague
The Netherlands
Driesenaar@fel.tno.nl

1   Introduction

Large organisations, such as NATO and the armed forces of its member countries, cannot function without the availability of accurate, timely, complete and consistent information. The quality of every decision that is made depends largely on the quality of the information on which the decision is based. This makes information an essential resource for any organisation that must be managed carefully.

Due to the intensified level of co-operation between NATO countries, it has become crucial that information can also be shared *between* armed forces. National forces are deployed ever more often in crisis management situations and (disaster-)relief operations throughout the world, requiring them to work together closely with forces of other countries. Fast and effective collaboration requires a method for information dissemination that is flexible and open.

The need to share information between countries translates directly to the requirement that information can be exchanged between their command & control (C2) systems. For this to be possible, the systems must agree to exchange and interpret information in a standardised (unambiguous) way. In other words: the systems must be interoperable.

This paper focuses on two existing information exchange standards: ADatP-3 (based on formatted messages) and ATCCIS (based on database replication). After describing and analysing both AdatP-3 and ATCCIS separately, the paper compares the two information exchange standards. Ideas are set forward for a unified approach which tries to capture the best of the two worlds and the paper ends with suggestions for future work.

2   Interoperability

Interoperability is defined here as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged" [1]. To explain this concept and to identify which elements are necessary for interoperability, we will examine information exchange as it occurs in different domains (between people and between systems) and we will describe its status quo and how this came to be.

2.1   Interoperability between people

For one person (the provider) to successfully transfer information to another (the receiver), agreements must be made at various levels. First, they must agree upon a medium of communication. If the provider uses writing but the receiver is illiterate, the exchange will fail. If the provider uses speech but the receiver is deaf, again the exchange will fail (although in this case, having the receiver lip-read may solve the problem).

Second, they must agree upon a language. If the chosen medium is speech but the provider speaks in a language unknown to the receiver, there will still be no exchange; that which is spoken may be heard, but it is not understood. The root of the problem lies in the fact that different languages have different vocabularies: they use different words to express the same ideas. This can also occur within a single language, when a speaker uses a jargon that is unknown to the listener. Agreeing upon a language not only entails agreeing upon a vocabulary, but also agreeing upon a common meaning for the words. Even if the provider does speak in a language which is known by both, if both parties attach different ideas to the same words (e.g., what is their definition of "entity"?) they may think they understand one another, while in fact they disagree.

Finally, they must agree upon a common communication procedure. It is no use standardising the format of a request, for example, when in practice the receiver fails to respond to requests because they are not going through the proper channels.

The extent to which these agreements can be made determines the level of understanding that can be achieved between the provider and receiver, and as such, the potential level of interaction between them.

2.2   Interoperability between systems

The agreements that must be made between people are the same agreements that must be made between C2-systems that wish to exchange information. First, they must agree upon a medium, i.e. the type of connection

will be used to communicate: what type of cable or frequency will physically connect the systems, and what protocol will be used to transport the messages that are sent.

Second, they must agree upon a language that is to be 'spoken' by the systems, i.e. the messages that will be exchanged. Each system has its own native language, which is contained in the structure of the information that is used by that system. For example, the structure may specify that there are clients; that clients have an address and a city of residence; and that clients can place one or more orders. Different systems will generally speak different languages: a 'client' in an order-processing system can be a 'debtor' in a financial administration package and can be a 'lead' in a sales-support system. Therefore, in order to exchange information between systems, it is necessary to create a common frame of reference for the concepts which exist in the individual information structures. In other words, an exchange language must be defined, which describes the messages in terms of syntax (what do they look like) and semantics (what do they mean).

Finally, they must agree upon a set of procedures which regulates the exchange of information: what is the (higher-level) protocol for message exchange between systems, which security considerations must be taken into account, which priorities will be supported, etc.

2.3 Past to present

In the last decade interoperability has become one of the most important issues in system design. This contrasts sharply with the early years, during which there was little need for interoperability. Initially, systems were designed to operate as stand-alone, autonomous units, dedicated towards supporting the work in a particular area or department. Each system had its own form of internal data storage that provided little if any access for external parties. In the few cases that information exchange between systems was required, a dedicated coupling (in the form of a translator) was custom-built.

As technology progressed and the number of systems grew, the need for information exchange increased. It proved infeasible to continue to develop and maintain the increasing number of system-specific couplings. The focus shifted towards finding ways in which couplings could be re-used or could be used to connect multiple systems together. Hardware standards were developed concerning cables and connectors; software standards were developed concerning protocols and services. From the bottom up, the various levels of the OSI-model were filled in.

Now that many technical problems have been solved and boundaries have been pushed back, it is becoming clear that to achieve true interoperability we need some crucial standards. There are different mechanisms available today which allow systems to connect to others, but these do not tell a system how to format and interpret messages that can be sent over the connection. In other words, the medium has been taken care of; the language and the procedures have yet to be worked out.

This setting formed the point of departure for NATO, which was seeing a growing need to interconnect the C2-systems of its member nations. NATO identified the absence of a standardised military language and message exchange protocol that would help its forces to communicate and interact more effectively. To solve this problem, different projects have been initiated over the years to devise a solution. These projects have taken different approaches towards designing an information exchange mechanism, but the two most successful approaches have been the use of formatted messages by ADatP-3 and the database replication approach taken by ATCCIS. Both approaches will be examined in later sections.

3 C2 Information

In order to judge the merit of ADatP-3 and ATCCIS as approaches towards achieving C2 interoperability, we must be clear on what type of information is exchanged between C2 systems. It then becomes possible to indicate to which degree each approach succeeds in supporting specific types of information exchange.

Here we wish to consider two types of C2 information: the actual content and transfer information. Content information is the information that is to be conveyed to a receiver; it is what would normally be written in a letter. Transfer information is the information that determines how the content is to be transferred; it is what would normally be provided on the envelope that contains the letter. Both will be examined in the following subsections.

3.1 Content

As indicated above, content is the information that is being exchanged. As such, this is the information that an exchange language must be able to express. Content comes in three flavours: descriptions, events, and reporting data (for simplicity, we do make a distinction between data and information).

Descriptive data describes the static C2 world; it refers to information that does not change (often) over time. For example, the name and nickname of a unit; the maximum cross-country speed of a Leopard-2 main battle tank; and the location of a town. This type of information can generally be provided ahead of use, in the form of a database or document, but it is sometimes necessary to be able to request it as the need arises.

Event data describes the dynamics of the C2 world; it refers to information that can change (often) over time. For example, the location and status of a unit; the identity of an as yet unidentified person; the sighting of an aircraft; and the available capacity of a field hospital. This type of information can not be provided ahead of

time, but will be reported on a regular basis or as soon as the event occurs.

Finally, *reporting data* is meta-data that provides a context for interpreting description- or event data. For example, the source of the information; the reliability of the source; the credibility of the data; and the time period of validity. This type of information will generally be reported together with the data it refers to.

3.2   Transfer information

Transfer information describes how the content is to be exchanged. As such, this is the information that must be used by the exchange medium: it determines how the information is communicated. For example, the identities of the sender and the intended recipient; the priority; the classification; and the type of encryption.

4   Approach 1: ADatP-3 (Formatted messages)

4.1   Introduction

ADatP-3 (Allied Data Publication 3) is the name of the publication which documents the NATO Message Text Formatting System (FORMETS); the abbreviation is also widely used to denote that same system. FORMETS specifies the message formats that are to be used in the construction of character-oriented messages that are exchanged between national and NATO authorities and systems. The use of ADatP-3 by all NATO countries has been ratified in STANAG 5500.

The goal of ADatP-3 is to serve as a standard for information exchange in general; not to specifically support exchange between systems. For this reason, ADatP-3 focuses on defining a message standard in which messages are concise, accurate and can be quickly processed by both human operators and automated systems. ADatP-3 specifies only the permitted message formats; it does not make any assumptions concerning the communication medium (although one of the most popular exchange mechanisms for ADatP-3 messages has been ACP127).

Figure 1 - Information flow between ADatP-3 systems. The shaded area identifies the scope of ADatP-3 work.

The use of ADatP-3 is very straightforward (see Figure 1). A user can transfer information to another user by either writing a message manually, or by generating the message using an automated system. The message can then be sent over any acceptable data transfer mechanism, and after receipt can be processed manually or automatically by the receiver.

4.2   Exchange language

ADatP-3 is in fact nothing more than an exchange language. It comprises an artificial, character-based language in which:

- the vocabulary is limited to a collection of codes and words, called fields, which have an unambiguous meaning;
- sentences are limited to certain sequences of fields, which are called sets, in which the position of a field is used to determine its meaning;
- messages are limited to certain sequences of sets, called message text formats (MTFs), in which the position of a set is used to determine its meaning.

The MTF definitions in ADatP-3 are independent of one another; however, MTFs can make use of the same sets, and sets can make use of the same fields. To illustrate the structure of an ADatP-3 message, here is an example:

```
MSGID/ENEMY SITREP/RPVGS/004//
EFDT/040849Z/JUL//
EGROUP/U0004/ORC//
LOCATION/REAL/-/-/-/POINT/32UPC9307//
SOURCE/-/RPV//
TIME/AT/040840ZJUL//
```

In the example, each line is a set, and each set consists of a set identifier (the first word) followed by one or more fields. The first set identifies the MTF that was used; in this case the message is of type ENEMY SITREP.

ADatP-3 messages support primarily the exchange of event data and reporting data. If necessary, description data can be provided in the form of free text, but this has no formal structure and cannot easily be used by automated systems. Transfer data that is supported by ADatP-3 are sender, message type and SIC codes; these can all be contained in the message itself. The format makes no assumptions concerning additional transfer information that may be used by the message transfer mechanism.

4.3   Advantages

The ADatP-3 approach has a number of advantages.

First, messages can be *processed independently*. ADatP-3 messages are designed to be self-supporting; they can contain only few references to external sources. As such, an ADatP-3 system does not require messages to arrive in any particular order because it can generally interpret each message in isolation.

Second, ADatP-3 messages are indeed quite *concise*. The formatting allows a lot of information to be provided in a small space.

Third, the message formats are *man-readable*. In part, this is due in part to the choice for an entirely character-oriented format. However, because message- and set headers in the messages provide helpful context information, and because the field-codes adhere to widely used abbreviations, most messages can be read and understood without requiring detailed knowledge of the ADatP-3 format. In fact, even messages that become damaged during transfer may still provide valuable information to a human operator.

Finally, ADatP-3 is a *mature standard* in that a large amount of user-feedback has been obtained with which the format has been improved in iterative steps.

4.4   Disadvantages

The ADatP-3 approach also has a number of disadvantages.

First, ADatP-3 *defines only the syntax* of the exchange language, not the semantics. Field codes are defined in terms of what they abbreviate, but their meaning within a set or the meaning of a set within a MTF are not specified. Although the meaning can often be inferred from the context (see also the first advantage noted above), different interpretations can exist.

Second, ADatP-3 is not always elegantly designed for use in automated systems because of some *minor design flaws*: Some fields permit the use of multiple units of measure; e.g., liquid amounts can be specified in liters or in gallons. Fields are sometimes ambiguous; e.g., a date can be specified either as DDMMYY or YYMMDD. Combinations of fields permit the same information to be specified in different ways; e.g., an armoured infantry unit can be identified by `/ARMD/INF/-/-/` or by `/-/INF/-/ARMD/` or variations thereof. All of these aspects make the development of an ADatP-3 system more complex. Of course, this point relates to the first point.

Finally, ADatP-3 is *not one standard* but a set of standards. The large number of improvements made to the MTFs has resulted in a large number of different versions of ADatP-3, often incompatible with earlier versions. In some cases individual countries have made their own version by adding nation-specific codes and formats, thus adding to the problem.

5   Approach 2: ATCCIS (Database replication)

5.1   Introduction

ATCCIS (Army Tactical Command & Control Information System) is an international study aimed at achieving interoperability between the C2 systems of the participating nations. Thirteen countries are currently active within ATCCIS, and several of these countries are already developing national systems based on the ATCCIS principles.

ATCCIS aims to achieve interoperability by using distributed databases that are synchronised through database replication. The idea is to share information between users by allowing them to write to and read from the same database. However, as a single, centralised database is infeasible in practice, ATCCIS provides multiple nodes in the network with a copy of the shared database, called the replication database, and ensures that changes made to the database at any node are replicated to all other nodes. The ATCCIS solution comprises the following elements:

- an exchange language in the form of a model called the LC2IEDM (Land C2 Information Exchange Data Model), which defines the structure of the shared database;
- an exchange mechanism based on the principles of database replication called the ARM (ATCCIS Replication Mechanism), which allows changes to the shared database to be communicated between nodes; and
- a transfer protocol which is used to transfer the replication messages between the ARMs at the various nodes (this is chosen rather than built; TCP/IP is currently being used).

To illustrate the working of ATCCIS we will examine a simple information flow between two systems. Consider a situation in which two ATCCIS nodes, each comprising of a single application, a geographical information system (GIS), and a copy of the shared database, are connected through a network (see Figure 2). In this example, one user records the movement of a unit using his GIS. This information is translated by the GIS into table updates (creates, updates and deletes) and applied to the replication database (RDB). These database updates are automatically replicated by grouping them in transactions and distributing them using the ARM. On the other end of the line, the transactions are received and applied to the database, and the GIS then translates the updates into information which can be displayed to the second user.

We will now look into the exchange language and the exchange mechanism in more detail.
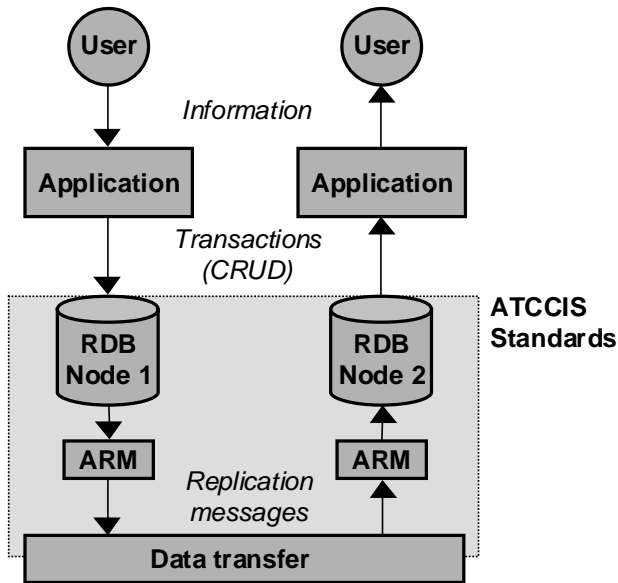
Figure 2 - Information flow between two ATCCIS nodes. The shaded area identifies the scope of ATCCIS work.

5.2   Exchange language

The ATCCIS exchange language is defined in a relational datamodel called the LC2IEDM. This model captures the structure of the information that is shared between ATCCIS users. The model is a conceptual model, meaning that it identifies the information concepts that are exchanged without stating how these are to be exchanged.

The scope of the LC2IEDM is the core army C2 information that is exchanged at an international level. Core C2 information refers to the general information concepts that are shared by virtually every unit and cell within the army. For example, the model recognises concepts such as battlefield objects (e.g., units, facilities, terrain features, control features), object characteristics (e.g., location, status, activity), object capabilities, and reports, plans and orders. The model defines these concepts at an international level, meaning that country-specific concepts (such as special naming conventions) are not supported.

However, the LC2IEDM has been developed to be extensible. For example, it is possible to locally add information concepts that are specific to a functional area (e.g., logistics, communications, and engineering) or that are used only by a single nation. In this way, the model acts as the hub of a wheel to which spokes can be added (which is why the model was initially called the ATCCIS Generic Hub Datamodel).

The LC2IEDM provides explicit support for both description data and event data, including the corresponding reporting data. The ARM handles all transfer data.

There are two reasons why the LC2IEDM is indeed the ATCCIS exchange language. First, the replication

database contains a direct implementation of this model. This means that the entities and attributes in the relational model have been translated directly to tables and columns in the database. As such, an application that wishes to access or modify the information must do so according to the structures defined in the model, i.e. the access-language is the LC2IEDM. Second, the model directly determines the format of the replication messages that are used by the ARM to exchange information; this is explained in more detail below.

5.3   Exchange mechanism

As explained earlier, ATCCIS exchanges information through the use of replication databases. This exchange can be local or remote: local exchange occurs when different users/applications access the same replication database (each replication database can serve multiple clients); remote exchange occurs when information is shared between users on different nodes through replication.

Exchange between replication databases (i.e. remote exchange) is performed by the ARM that must be present on each node. If a modification is made to the local replication database, the associated additions, changes and deletions are sent by the ARM to the ARMs on other nodes in the form of a replication message. If a replication message is received from another ARM, the ARM simply carries out the modifications contained in that message in the local database. In this way, the databases throughout the network are synchronised after each modification. Because modifications are transmitted as changes to tables and records that are identified in the LC2IEDM, the latter model directly describes the structure of replication messages also, again performing in its role as exchange language.

The ARM implements automatic replication based on contracts. A replication contract is an agreement between two users on the information that they will exchange, which is described by four main parts: a data provider, a data receiver, a contract type, and a filter. The use of contracts allows the ARM to work autonomously; modifications are communicated automatically to all nodes that have indicated an interest in the information via a contract (there is no manual trigger required from the user). Furthermore, contracts can be added and removed dynamically as the information requirements change. Note that if there is no contract, then there is no replication.

The ARM permits selective replication through the use of contract types and filters. Given that most tactical networks have bandwidth limitations, it is necessary to reduce the levels of communication as far as possible. Furthermore, due to security considerations it may be desirable to give parties only selected access to information contained in the database. This can be achieved first by using contract types to indicate the type of information that is required: for example, the user may only wish to have information concerning the common

operational picture or information about plans and orders. Next, the information content can be further refined using pre-defined filters that can be parameterised to suit individual preferences: for example, the user may wish to receive only information concerning units in a particular area. As contracts must always be accepted by both provider and receiver, security can be enforced.

All information needed by the ARM to implement automatic, selective replication is stored in the replication database. For this purpose an ARM Management Model (AMM) resides in the database next to the LC2IEDM. The AMM stores information such as the users, the topology of the network (e.g., where are the users located), which pre-defined types of contracts and filters are available, and which contracts and filters which have indeed been defined. The ARM management protocol allows nodes, users, contracts, and flow control to be managed dynamically.

5.4   Advantages

The approach taken by ATCCIS has a number of advantages.

First, the ATCCIS exchange language is *highly consistent*. Because all concepts are contained in a single model that is highly normalised, structures are only defined once. For example, there is only one standard for defining locations or date-time-groups. As another example, the identification of a unit is defined only once in the model and can be re-used wherever necessary.

Second, the ATCCIS exchange language supports *referencing*. So, instead of including for example all information about a unit, one can include a reference to the unit. Of course, this can greatly reduce the size of replication messages.

Third, ATCCIS supports *automatic distribution* of information, as explained in the previous section.

Finally, ATCCIS supports *selective distribution* of information.

5.5   Disadvantages

The approach taken by ATCCIS also has a number of disadvantages.

First, the ATCCIS Exchange Language is *too expressive* to ensure interoperable applications. On the basis of LC2IEDM it is possible to represent, and thus convey, rather complicated information constructs. As a simple example, the LC2IEDM supports report data on event data reported by someone else. The possible constructs are virtually endless and it is certainly possible that applications do not support the same ones.

Second, *event preservation is not explicitly supported*. ATCCIS subdivides events into small segments (e.g., a unit movement is subdivided into a unit segment, a point location segment, a time segment, and the relations between the segments) according to the structure of the LC2IEDM. These segments are then replicated either together or individually, possibly mixed together with segments of other events, and must be regrouped by the application on the receiving end before they can be presented to the user as the initial events. As such, there is no correspondence between events and replication messages; the application must constantly decide whether the latest replicated change will allow it to generate an event or whether it should wait for additional information. This impacts the design of ATCCIS-based applications as well as that of translators that must translate between ATCCIS and other formats (e.g., ADatP-3). It also makes it difficult to implement the filters that can be used in contracts, because a user will generally wish to filter on events rather than on table updates.

Third, *data completeness* is not signalled. It is not always possible to determine whether all database changes relating to a specific event have been received. For example, it is not possible to identify whether all unit locations in a particular plan have been collected. This adds to the problem described above concerning the translation of database updates to user events.

Fourth, ATCCIS replication messages can not be *processed independently*. One reason, of course, is the fact that a replication message can contain data relating to different events. The other reason is that ATCCIS enforces strict referential integrity – meaning that information referenced to should be passed prior to its reference.

Fifth, ATCCIS *replication messages are relatively large*. Replication message syntax does not allow the updating of an individual column in a table record; the entire record must be sent. Next, ATCCIS makes use of technical database keys, which can become very long (e.g., each unit is identified by a unique number of 18 characters). Finally, the structure of the exchange language can cause a small event (e.g., the movement of a unit) to result in many changes to the database, each of which can result in an individual replication message. As such, ATCCIS is not designed to minimise network load, even though it provides support for contracts and filters which reduce the load.

Sixth, there is no support for *varying the quality of service*. All information that is replicated is currently processed with the same level of service: it is sent intact, complete, in order and secure. However, because it is not possible to identify the battlefield event to which a replication message corresponds, it is difficult to assign other service characteristics to messages, such as priority, classification, or time-to-live qualification.

Finally, we observe that ATCCIS is still very much a *standard-to-be*. Little experience has been gained in the practical use of the products, other than what was learned during the few demonstrations held by ATCCIS itself. It

is expected that many lessons learned have yet to be fed back to the standard in order to improve it.

6 ADatP-3 versus ATCCIS

Within the NATO community, ADatP-3 and ATCCIS are viewed as being two completely different approaches towards achieving interoperability between C2 systems. This has resulted in a debate over which of the approaches will best serve for the future. In Table 1 we summarise the results of our analysis of ATCCIS and ADatP-3. Each aspect will be discussed individually below.

Table 1. Comparison between ADatP-3 and ATCCIS.
(- : poor, -/+: reasonable, +: good, NS: not supported)

Aspect	ADatP-3	ATCCIS
Consistent	-/+	+
Expressive	-/+	+
Event preservation	+	-
Data completeness	+	-
Independent processing	+	-
Message size	-/+	-/+
Referencing	-	+
Automatic distribution	NS	+
Selective distribution	NS	+
Man-readable	+	-

Consistent: AdatP-3 is not as strict as ATCCIS concerning message syntax and semantics. This is mainly due to the fact that ATCCIS uses a model to derive the syntax and define semantics.

Expressive: ATCCIS is a more expressive approach than ADatP-3, however, ATCCIS is too expressive to enforce interoperability. In ADatP-3, information constructs are constrained to that which can be formulated using the pre-defined MTFs. In ATCCIS, many information constructs are possible and C2-systems are almost bound to differ in the constructs they support, causing (possibly invisible) breaches in or even breaking of interoperability.

Event preservation: ADatP-3 preserves events; ATCCIS does not. ADatP-3 messages contain complete events and can be interpreted in isolation. ATCCIS can replicate events either in a single replication message or using multiple messages, leaving it up to the receiving application to recreate the event for the user.

Data completeness: ADatP-3 signals data completeness, ATCCIS does not. Note that data completeness relates to event preservation: ATCCIS will

implicitly signal data completeness, as soon as it preserves events.

Independent processing: In general, AdatP-3 messages can be processed independently, while ATCCIS replication messages can not be processed independently.

Message size and referencing : The amount of data that is physically transferred during information exchange will on average be the same. ADatP-3 messages are concise in comparison with the data that must be replicated when the same information is exchanged within ATCCIS. However, ATCCIS is able to refer to information that has already been sent and only has to send it once, while an ADatP-3 message must always contain all relevant information. In practice, therefore, the amount of information that must be transferred will be comparable (and can be reduced in both cases using compression techniques).

Automatic and selective distribution : ADatP-3 does not support these mechanisms, ATCCIS does. Within ADatP-3, information exchange is initiated by the sender (information-push). ATCCIS, however, allows the receiver to selectively indicate what information he wishes to receive automatically (information-pull).

Man-readable : ADatP-3 messages can be read by human operators; ATCCIS replication messages cannot. ADatP-3 makes use of standard field-codes and uses set identifiers, thus making messages fairly easy to read (although certain message types will require knowledge of the format). ATCCIS replication messages contain table identifiers, numerical database keys (which refer to entities defined in the database) and cryptic mnemonics; their contents cannot be determined without access to the database.

7 Conclusion

We take the view that ADatP-3 and ATCCIS are not completely different approaches, but rather are variations on a common theme. Both can be considered message-oriented solutions: ADatP-3 makes use of ADatP-3 messages, and ATCCIS makes use of replication messages. The main difference between the two is how the messages are generated and how they are processed.

The comparison in the previous section indicates that while neither approach is superior, they complement each other's strengths and weaknesses. This would suggest that a combination might be able to capture the best of both. We therefore come to the following recommendations concerning a unified approach.

7.1 Recommendations for a unified approach

The analysis presented in this paper gives raise to the following recommendations:

- Use a single, unified conceptual model to define the messages of the exchange language (as done in ATCCIS). Allow information structures to be re-

used (e.g., use the same form of unit identification throughout the model). This will result in *consistency* and elegance. Both ADatP-3's MTFs and ATCCIS's LC2IEDM contain many information concepts that can act as starting point for the model.

- Distinguish between description-, event- and reporting data in the model. These can even become separate models. This will keep the model simple and understandable.

- Focus on event data, as this is the most important information that is exchanged between C2 systems. Specify the individual *events* and specify how these are to be mapped to the model and back; leave no room for alternative interpretations. This will limit the *expressiveness* of ATCCIS and ensure and facilitate building interoperable C2-systems.

- Make sure that messages *preserve events* and that messages can be *processed independently.* This will simplify the development of message processing systems.

- Do not require messages to be *man-readable.* Although this was desirable in the past, expect messages to be exchanged between C2 systems only.

From experience obtained in dealing with C2-interoperability matters we would also like to add the following recommendations for the advanced reader:

- Make the conceptual model concrete: do not hide information concepts in abstractions or generic structures. These can be added later when the physical implementation is developed.

- Do not strive to develop a model that can fit on a single page. Allow the model to be multi-dimensional that can be viewed from different angles.

- Consider carefully if the proposed use of the messages (e.g., how will they be filtered and distributed) should affect the structure of the conceptual model. Try to focus only on what will be exchanged at the conceptual level, and include aspects of use at the logical- and implementation levels.

7.2   Future work

In this paper we have looked only briefly at data distribution mechanisms. Although ADatP-3 does not prescribe the use of a particular mechanism, it is primarily suited for point-to-point protocols such as telex and email. ATCCIS bases its own mechanism on automatic and selective replication. When further developing a unified approach it may be worth to consider the following:

- support for point-to-multi-point data distribution – supporting this can result in more efficient data distribution, and may even be essential for use of combat net radios;

- support for different data distribution mechanisms – this may enable a more flexible way to implement

automatic information exchange and it may also reduce network load (such as request/reply and publish/subscribe);

- support for various quality of service aspects (such as: priority, assured delivery, confirmed delivery, encryption, compression) – this is especially important in communication critical environments, where the required transmission capacities are close to or even exceed the available transmission capacity, or in cases where the required quality of service exceeds the supported quality of service (for example when an unencrypted classified message is transferred over an insecure data link).

- use of commercially available message oriented middle-ware products, such as: IBM MQSeries, TIBCO TIB/Rendezvous, Talarian MQExpress;

- support other existing interoperability related standards – on the basis of the concrete unified conceptual model and the list of 'events' it may be possible to achieve interoperability using existing standards such as CORBA, COM/DCOM, HLA, and XML. This could enhance the scope of the standard, enable the use of more COTS products, and facilitate the development of applications.

In the end, the unified approach may evolve into an information bus architecture, where C2-systems and/or C2-applications can connect to a C2-network in a 'plug-and-play' fashion.

8   References

[1] IEEE, *IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries*, New York, 1990

[2] D. Alberts, J. Garstka, F. Stein, *Network centric warfare: developing and leveraging information superiority*, 2nd Edition (Revised), August 1999

Potentials of Advanced Database Technology for Military Information Systems

Sunil Choenni[a]   Ben Bruggeman[b]

[a] *National Aerospace Laboratory, NLR, P.O. Box 90502, 1006 BM Amsterdam, The Netherlands*
[b] *Royal Netherlands Naval College, KIM, P.O. Box 10000, 1780 CA Den Helder, The Netherlands*

Abstract

Research and development in database technology evolves in several directions, which are not necessarily divergent. A number of these directions might be promising for military information systems as well. In this paper, we discuss the potentials of multi-media databases and data mining. Both directions focuss on the handling of a vague information need of a user. In general, data mining systems allow a higher degree of vagueness than multi-media systems. Information systems that are able to handle vague information needs adequately will improve the decision making process of militaries.

1 Introduction

Both the military and civil communities have come to the realisation that military information systems fundamentally differ from civil ones in some respects [6]. The difference started with the views that military and civil communities had on information systems. The military community was and is looking forward to information systems that support the *human decision making process* in a time pressured dynamic environment with multiple, often conflicting, goals, especially for command and control tasks. Databases in these information systems should be able to deal with high volume and of uncertain data at lower levels of the decision chain, and with aggregated information at higher decision making levels [12]. Furthermore, databases should be able to handle vague information needs. For a long time, the civil community had a less ambitious view on information systems. Civil information systems were developed to help employees in performing a number of tasks, often administrative, in an efficient way. This was realised by handing over routine actions to computers. Databases in civil information systems consisted of well-structured and true data. To retrieve data from these systems one should exactly specify what data is to be retrieved. In the sixties it was even worse, relevant data could only be retrieved if the system was exactly told how to navigate towards the data. As soon as the automation of routine jobs was well understood, research and de-

velopment in civil information systems became more ambitious. Currently, research in databases evolves in several directions, which are not necessarily divergent [22]. A number of these directions appears to be promising for military information systems as well. In this paper, we discuss the potentials of multi-media databases and data mining for military information systems. Both directions focuss on the handling of a vague information need of the user. In general, data mining systems allow a higher degree of vagueness than multi-media systems.

A multi-media database management system aims to reply to the (vague) information need by combining different (advanced) types of data, such as audio, video, images, etc. We present an architecture that facilitates the storage and access to different types of media servers in an integrated manner. Main properties of the architecture are modularity and extensibility. Depending on the imposed requirements and characteristics of an application, this architecture can be elaborated such that a tailored operational multi-media database system can be built, including a military multi-media database system. We report on the issues that play a role in building a military multi-media system from a database perspective.

As noted before, a second direction that is promising for military information systems is data mining. This direction has seen a recent surge in commercial development. Data mining has as goal to extract implicit, previously unknown, and potentially useful knowledge from large data sets. Extracted knowledge may support or be used in strategic decision making. In the military world, data mining can be applied to search for correlation between doctrines, to predict how an enemy will act based on large volumes of data collected about the enemy (in the past), to improve internal logistics, etc. To realise this task, data mining combines techniques from the fields of machine learning, statistics, artificial intelligence, and database technology. We discuss the requirements to mine operational databases successfully. Furthermore, we report on the results that we have obtained by mining two aircraft incident databases.

The remainder of this paper is organised as follows. In Section 2, we motivate why advanced database technology might be interesting for military applica-

tions. In Section 3, we discuss the challenges and potential solutions for multi-media database systems. In Section 4, we give a brief state of the art in the field of data mining. Furthermore, we report on an application concerning aircraft incidents. Finally, the paper is concluded in Section 5.

2 Military benefit

As stated in the introduction, military applications put complex demands on information systems, especially when information systems have to operate in non co-operative environments. Taking the developments in the operational field into account, we expect that the demands that a military information system will have to meet may become even more complex. We note that these demands are partly yielded by technological achievements.

To meet the military demands on information systems, the application of advanced information technology, including database technology, as underpinning is inevitably. For example, the manoeuvrability of aerial targets have been increased. An impact of this increase is that the performance of classical methods for trajectory tracking and prediction —which, in general, are based on simple dynamic models that do not exploit knowledge about a "situation"— suffers [7]. So, the development of alternative trajectory and predicting systems that include data and knowledge bases, which are able to handle fuzziness and uncertainty, is becoming interesting.

Today, a number of trends may be distinguished in the world in which militaries are operating in. To mention some of these trends:

- There is a growing need for joint and combined operations. In order to perform these operations successfully information from different and heterogeneous sources should be processed and integrated in a consistent manner [12].

- Due to budget cuttings, military organizations (in many countries) should work more efficiently [26]. One way to achieve this is to automate a large number of tasks that is currently performed by human-beings.

- Shorter reaction times are demanded. This means that we require high performance data processing tools [8].

- Soldiers are or will be equipped with advanced technologies and means, such as microphones, videos, digital maps, etc.

- A wide variety of methods and techniques are applied for intelligence gathering and data processing [8].

Each of these trends has in common that they strongly rely on information processing and integration techniques. For example, applications of a wide variety of

methods and techniques for intelligence gathering will result in a (wide) diversity of information, each with their own characteristics, e.g., with regard to uncertainty measures, data types, etc. The challenge is to combine all information in order to select high quality information and to present this information in a suitable way to the user. For the other above-mentioned trends a similar reasoning holds.

We feel that both multi-media database management systems and data mining systems are useful to successfully implement these trends. We note that well-defined questions to well-defined databases can be handled by standard database technology. Multi-media database management systems are intended to reply on information needs that are vague and incomplete. An example of such a question is: give me all information that is known about servo XY, in connection with a defect. Data mining systems are aimed to answer more strategic kind of questions, such as: should we consider an enemy that is able to perform actions X and Y as dangerous or not?

In the two successive sections, we discuss multi-media databases and data mining in more detail.

3 Multi-media databases

Many advanced applications, including military applications, will profit from the integration of information from different media servers. This integration generally leads to a better assessment of a situation; moreover user interactivity will further improve the assessment. Integration and the support of user interactivity are major goals of multi-media databases. Within the scope of soldier modernization programs multi-media databases are promising, since soldiers are equipped with a wide variety of advanced technology and means, such as microphones, videos, maps, etc. Multi-media databases might also be useful in traditional situations as well, e.g., for off-site support. Suppose that in a state of war a ship at sea has problems with its engine and several media servers (including handbooks) contain information about this type of engine, e.g., images of the engine, video fragments how to disconnect/connect several parts of the engine, audio fragments containing the functionality of the several parts of the engine, etc. A multi-media database system may help the crew in solving the problems at the ship by replying adequately to the information need of the crew. Suppose that the first guess of the crew is that something might be wrong with compression and formulates its information need as "give me all information about compression in a combustion chamber". As a reply the system should integrate all information available on different media servers to explain the functions and the use of the different parts of, for example, the combustion chamber in a suitable order. Furthermore, it might also be useful to demonstrate how the compression can be measured and what tools are proper for this purpose. We note that in a

state of war it is not always possible to let an expert come from land.

In order to support above-mentioned applications, advanced database architectures and query handling techniques are required. In Section 3.1, we discuss such an architecture and in Section 3.2, we report on the challenges that this architecture entails for query handling.

3.1 Architecture

In order to support multi-media application, we feel that advanced database architectures, should provide developers the following components.

- Suitable user interfaces, in which users can express their information need in a human friendly manner.

- A multi-media storage server which provides the capability to store and to access different types of data, e.g., images, video, audio, etc.

- Feature extractors. These extractors might be specialized algorithms that are able to extract relevant features from a piece of data.

- A meta database which contains features about the data stored on the multi-media storage server, such as color histograms, texture, annotated text, etc.

- A data dictionary which contains relevant information with regard to the whole system.

As depicted in Figure 1, all these components are loosely connected to each other through a network (e.g., Internet), supporting the concept of modularity and extensibility. For example, if a novel feature algorithm has been made available extracting previously unknown features, it can be plugged in into the feature extraction module and the meta database should be extended with a number of attributes or relations to store these features. The remainder of the system can be left untouched. This architecture is suitable for on-site as well as for off-site support for militaries. A more detailed description can be found in [25].

To be viable from a database perspective, the architecture has to be supported by advanced query handling techniques that exploit the meta database. The challenge in the field of query handling is to devise techniques that analyse the information need of the user and maps the result to relevant features in the meta database. These features are used to identify the data that meet the information need and to locate the data at the multi-media server. Suppose that a user provides an image and is interested in similar images stored on the server. To answer this information need, a mapping might be on color histograms. The color of the provided image can be computed and compared with the color histograms of the stored images in the meta database. Images with similar color
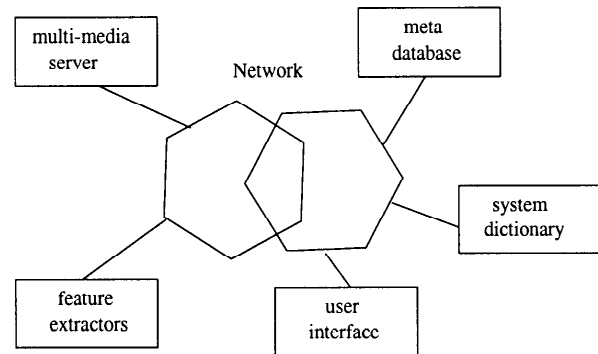
Figure 1: Architecture of a multi-media database system

histograms are retrieved from the multi-media server and presented to the user.

3.2 Query handling

In order to handle user queries, traditional optimizers[1] produce an evaluation plan for each query. This evaluation plan specifies the actual (basic) operations (i.e., joins, selections, etc.) and the order, in which these operations should be performed. The goal of the optimizer is to choose a cheap evaluation plan, primarily in terms of disk accesses. We note that, in general, it is infeasible for an optimizer to search for the cheapest plan, since the problem of query optimization is hard. However, query optimizers are performing quite well in finding a cheap evaluation plan for a single query [14, 24].

For multi-media applications, traditional query optimization techniques are inadequate for the following reasons. First, information needs formulated by a user are not exact as in traditional applications, but rather vague and incomplete. So, an optimizer should be able to handle vagueness and incompleteness.

Second, an information need is often decomposed into a sequence of queries, and current database management systems do not exploit multi query optimization techniques [5, 10, 20] but are focussed on the optimization of a single query in isolation. Multi-query optimization techniques, which search for a plan of a sequence of queries instead of a single query, may considerably speed up the processing of information needs. For example, a user is searching for pictures of cheerful ladies in uniform (e.g., to promote an airline company). Let us assume that this need is decomposed into two queries: 1) select all cheerful ladies and 2) select all cheerful uniforms. If we have a small set of uniforms, it is better to search for a set of cheerful uniforms first, and then to select pictures of cheerful ladies wearing these uniforms. So, an optimizer should be able to determine the best order in which a query of sequence may be performed.

---

[1] An optimizer is a software module in a database management system that has as goal to speed up the processing of a query. This is mainly achieved by reducing the number of disk accesses.
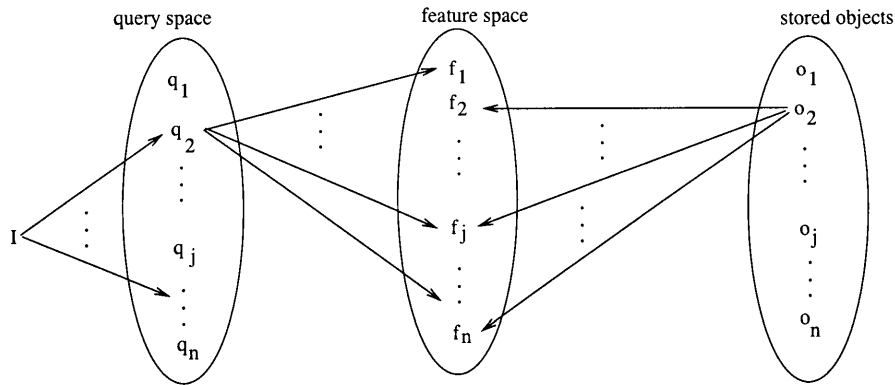
Figure 2: Mapping of information needs and objects on features

Third, the search for a proper answer to an information need is an interactive process. Suppose that a multi-media database management system comes up with a set of cheerful ladies in uniform as a result on the above-mentioned information need. By means of this output, the user specifies more precisely what in her/his opinion should be understood by cheerful ladies, e.g., by assigning a rating to each picture of the selected set. On the basis of a subset of the pictures (e.g. those with high ratings), the system searches for pictures that better meets the information need of the user. This does not automatically mean that the originally selected set can be thrown away. It is very well possible that a user refers lateron to some pictures that were not interesting on the first glance. Therefore, it might be sensible to store some of the originally selected pictures. The challenge for novel optimizers is to decide what pictures should be stored and for how long. Current database management systems do not take the interactive behaviour of users into account. Next generations database management systems may speed up the processing of information needs by exploiting the interactive behaviour of users.

Although multi-query optimization and the exploitation of interactive behaviour of users are important techniques to speed up the processing of information needs, we feel that an adequate handling of vagueness and incompleteness captured into a specified information need is of crucial importance. If an information need is wrongly interpreted by a database management system —due to inadequate techniques to handle vagueness and incompleteness— compared to what a user has meant, the system will come up with answers that are not relevant. Therefore, we discuss this issue in more detail, and we argue that solutions for this issue are application dependent.

In Section 3.1, we have noted that the meta database (see Figure 1) is exploited in processing an information need. This meta database contains features of the stored objects. In order to process an information need, this is split into a number of queries. Each query is mapped into a set of features, which is used to search for objects that (hopefully) partly meet the information need. The results of all queries

are combined to produce the final result to an information need. In Figure 2, we have sketched this process in a simplified form. The main challenge is to map a query into a set of relevant and useful features. Since this is a tough task, we know beforehand that a map on a feature for a query will be partly true and useful. Therefore, proposed solutions assign an uncertainty measure to each map. Different solutions differ in the uncertainty formalism (e.g., Bayesian theory, Dempster-Shafer theory, etc.) they choose [18, 23]. So, we can distinguish two main problems in the mapping of a query into a set of features.

- How to translate a concept that appears in a query into a set of features (often defined beforehand) that might be quite abstract, such as color histograms, texture, etc. The usefulness of solutions to this problem is application dependent. The better we understand a concept, the better we can translate this concept into a useful set of features. For example, the term data fusion in the military world stresses on the combination of data coming from different sources, while in the database world the same term stresses on the correctness of data to be stored. This also implies that the usefulness of a multi-media database system is determined by how well an application domain is understood by a developer.

- How to measure the "goodness" of a mapping of a concept to a feature, and how to propagate these measurements. We feel that the better we understand the first problem, the better we will be able to handle this problem.

We note that research and development in both above-mentioned fields are still in their childhood.

Another mapping that plays a role in Figure 2 is the mapping of stored objects into features. In the context of information retrieval, a considerable amount of research is devoted to this problem. In the field of information retrieval, one tries to find the relevant documents —given an information need— from a collection of stored documents. A widely accepted model to handle this problem is the following. For each document, a representative set of terms is extracted and

stored, e.g. in a meta database. To each term, a term and a document frequency are assigned. The term frequency expresses the number of times that a term t appears in a document d, denoted as $tf(t,d)$. The document frequency expresses the number of documents in which a term t appears, denoted as $df(t)$. The values for $tf(t,d)$ and $df(t)$ are used to retrieve relevant documents.

To determine the probability that a document d is relevant for a given set of terms T we have to compute $Pr(d|T)$. According to Bayes rule $Pr(d|T) \propto Pr(T|d)$ [16]. To determine $Pr(T = t|d)$, the following formula is widely used

$$Pr(T = t|d) = \alpha_1 \frac{df(t)}{\sum_{t \in T} df(t)} + \alpha_2 \frac{tf(t,d)}{\sum_{t \in T} tf(t,d)}$$

Setting the proper values for α_1 and α_2 is often a matter of trial and error.

Extension of the above-mentioned strategy for mapping objects to features is a promising direction in the field of multi-media databases. We note that capturing all mappings of Figure 2, into a consistent framework will be a significant progress in this field.

Since the success of multi-media databases depends on the understanding of the concepts used in an application, it is important that militaries actively participate in the development of a new generation of military information systems.

4 Data Mining

The large amount of data stored in databases may serve two purposes. First, it may help in understanding a phenomenon, and second it may help to predict the outcome of similar phenomena. In our view, data mining is a powerful tool that contributes to the realization of these purposes.

Data mining has as goal to extract potentially useful information from large databases by using a wide variety of methods and techniques, including statistical ones, that is able to explore large data sets efficiently. In practice it is seldom the case that a proper data set is available that can be directly mined. Therefore, we feel that the following four steps are equally important for effectively data mining.

1. A so-called mining question should be formulated. This question should specify the kind of information one is looking for. In general, a mining question is formulated together with domain experts.

2. Then, the data that may be used in order to answer the mining question should be selected, enriched, cleaned, and integrated, i.e., constructing a *data warehouse*. Since intelligence gathering is an important activity in many military applications, data enrichment is a key factor in building data warehouses successfully.

3. A mining algorithm has to be selected/developed that will search the data warehouse for appropriate answers to the mining question.

4. Finally, the answers of the search process should be presented in a way such that domain experts are able to understand and evaluate these answers.

We note that these four steps are defined as *Knowledge Discovery in Databases (KDD)* [15] and should be iteratively applied. In general, data mining is a highly interactive process. In practice, users start with a rough idea of the information that might be interesting, and during the mining session the user more explicitly specifies, based on, among others, the mining results obtained so far, which information should be searched for.

In Section 4.1, we describe a comprehensive and extensible architecture that supports the above-mentioned steps. Since the mining step provides us arguments to assess the performance of data mining in military applications, we focus on a number of mining techniques in Section 4.2. Then, in Section 4.3, we briefly describe a data mining application that we have performed at our laboratory.

4.1 KDD Architecture

To be viable, a data mining tool should support all the steps distinguished for knowledge discovery. In Figure 3, the architecture of a data mining tool is presented that supports all these steps. This architecture consists of a number of modules. The input consists of a mining question, the databases that should be mined, the algorithm that is selected for mining, and additional requirements that may be posed by the user. For example, a user may demand that an attribute in a database should not be involved in the mining process for reasons of privacy (e.g., income of pilots). The user has the possibility to request intermediate mining results, and if desired, the user is able to modify the input. An advanced graphical user interface is required to facilitate the presentation of the input and the interaction between tool and user. Mining results should be presented in a way that can be easily interpreted by domain experts.

Once the tool has received its input, it should extract and pass proper information to a module that contains a suite of tools to set up a data warehouse on the one hand, and on the other hand it should formalize the mining question such that it is understood by the selected mining algorithm. In general, the generation of a data warehouse may be a complex task. Therefore, tools are required to support this task. For example, there is a practical need for tools that detect conflicting data, filter noisy data, etc.

Once a data warehouse has been established, a mining algorithm will heavily interrogate the database to expose useful knowledge hidden in the database. Each
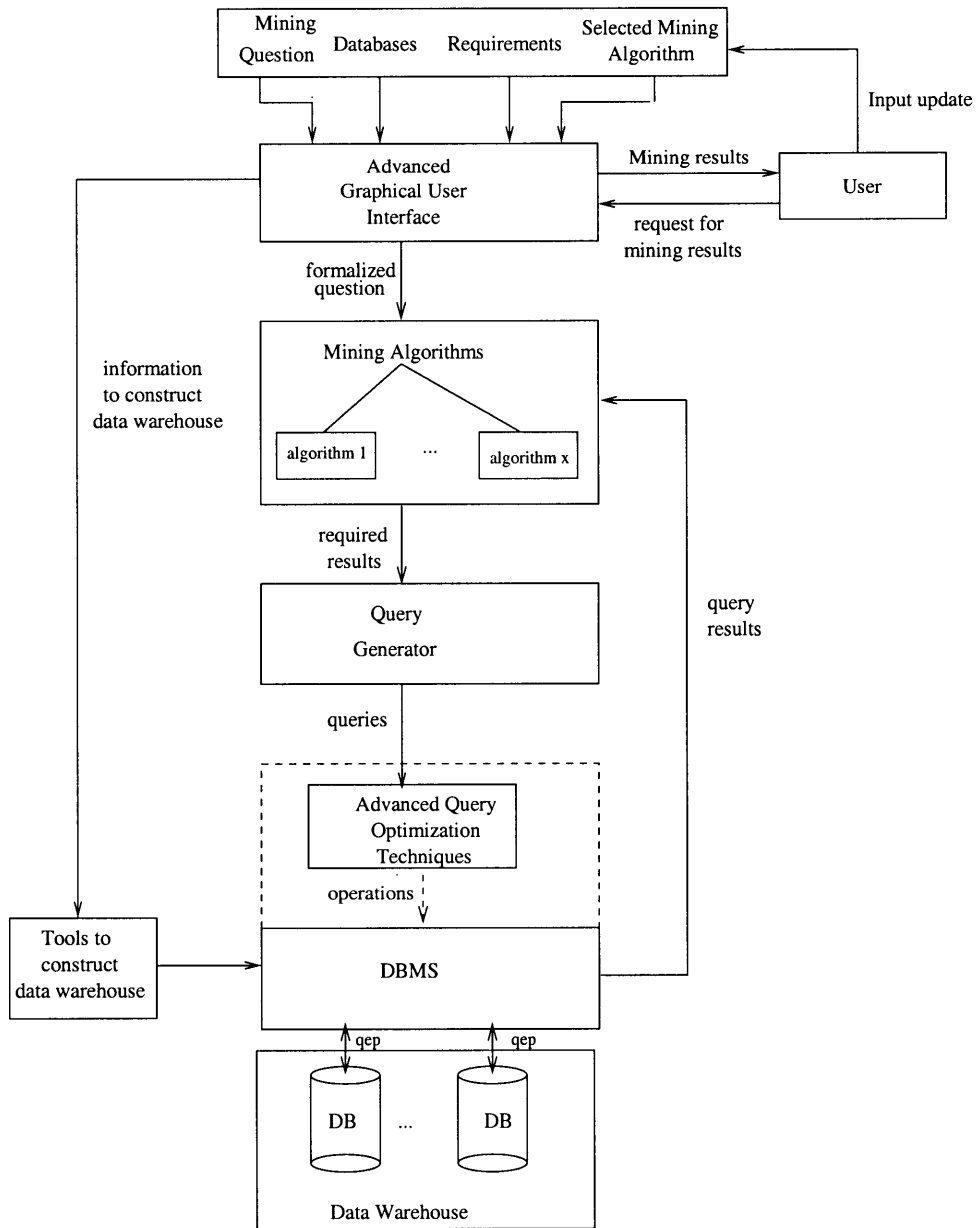
Figure 3: Architecture of a data mining tool

interrogation should be translated into a query that is understood by the underlying database management system (dbms). Therefore, a query generator should be part of a data mining tool. If a dbms receives a query, it generates an efficient query execution plan (qep), which describes step by step the operations to be performed in order to retrieve the result of a query from the data warehouse. Then, the query execution plan is performed, and the result is passed to the dbms, which passes it, in turn, to the mining algorithms module.

Since the retrieval of data from databases is still a bottleneck [9, 14], the performance of a data mining tool depends on the query processing capabilities of a database management system (dbms). We have observed that mining algorithms may pose simultaneously sequences of interdependent queries to a dbms and that exploitation of dependencies speed up query processing [5, 10, 20]. Since database management systems on the market place do not take advantage of this fact, we suggest to implement advanced query optimization techniques on top of commercial database systems in order to gain performance.

Our motivation to equip the mining algorithm module with a wide variety of algorithms is that some algorithms may very well be suited for some applications, while less suited for others.

A wide variety of techniques are available for mining purposes (step 3) varying from classical techniques, such as regression, to more novel ones, such as evolutionary computing techniques. In the following, we discuss how to set up a mining algorithm module. We briefly discuss a number of mining techniques which might be included in such a module.

4.2 Mining algorithm module

During the past years many algorithms have been developed and implemented for data mining tasks by the research as well as the commercial community. A proper question that raises up: "What mining algorithm do I need?". Although a rough categorization is possible for the mining algorithms, the question remains tough to answer. Currently, we may distinguish three categories of algorithms: classification algorithms, algorithms for association rules, and algorithms to mine time series databases.

Classification has as goal to distribute objects/tuples on the basis of common properties into a number of classes [1]. Algorithms for association rules are focussed towards the search of frequently occurring patterns in a database [2]. Mining algorithms for time series databases search for common patterns embedded in a database of sequences of events [4].

As motivated before, a mining algorithm module should be equipped with a wide variety of algorithms. In order to select the proper algorithm, a characterization of each type of algorithms is necessary. Then, a system or an expert may choose an algorithm for the problem at hand.

In Figure 4, we illustrate for classification problems how such a module may be set up. On the basis of a number of characteristics that may be relevant in choosing a mining technique, a tree is constructed. Suppose we have a mining problem and we know that attributes $x_1, x_2, x_3, ..., x_n$ are linear dependent on an attribute y. If we are searching for a function to describe this dependency, then a linear regression technique might be a good choice for this task. For a more detailed characterization of a number of mining techniques, we refer to Section 4.2.2

Today, the majority of the mining algorithms are focussed towards association rules and classification. In the following we give a brief overview of the algorithms in these fields.

4.2.1 Association rules

The field of association rules has been inspired by databases that store items purchased by a customer as a transaction [2, 3, 15]. A planning department may be interested in finding associations between sets of items. An example of such an association may be that 90% of the transactions that purchase diapers also purchase beer. This might be a good reason to place these two items close to each other to provide the customer a better service.

For reasons of convenience, we model a supermarket database as a relation $basket(i_1, i_2, i_3, ..., i_k)$, in which $i_j, 1 \leq j \leq k$, is a binary attribute that records whether an item has been sold or not. Note that a tuple in the database corresponds to a shopping basket at the counter.

Let I be the sets of all items, $X, Y \subseteq I$, and $s(X)$ the percentage of baskets containing all items in X. Then, $X \rightarrow Y$ is a association rule with regard to t_1 and t_2 iff $s(XY) \geq t_1$ and $\frac{s(XY)}{s(X)} \geq t_2$, in which $Y \not\subseteq X$. The latter equation expresses the confidence in a rule, while the former equation guarantees that itemsets should have a minimum size, and therefore a minimum support.

The problem is to find all rules that have minimum support and confidence greater than the defined threshold values t_1 and t_2. In general, two steps are distinguished in solving this problem. In step 1, all itemsets that have minimum support are selected. And in step 2, for each itemset X found in step 1 and any $Z \subset X$, all rules that have minimum confidence are generated.

Suppose that we have the following four transactions: $T_1 = \{a, b, c\}$, $T_2 = \{a, b, d\}$, $T_3 = \{a, d, e\}$ and $T_4 = \{a, b, d\}$. Let the minimum support and minimum confidence be 0.7 and 0.9 respectively. Then, the following itemsets meet the minimum support. $s(\{a\}) = 1$, $s(\{b\}) = 0.75$, $s(\{d\}) = 0.75$, $s(\{ab\}) = 0.75$, and $s(\{ad\}) = 0.75$. The valid association rules are[2]: $b \rightarrow a$ and $d \rightarrow a$, both with

---

[2]Note that $a \rightarrow b$ is not a valid rule, since $\frac{s(ab)}{s(a)} = \frac{0.75}{1} < 0.9$.
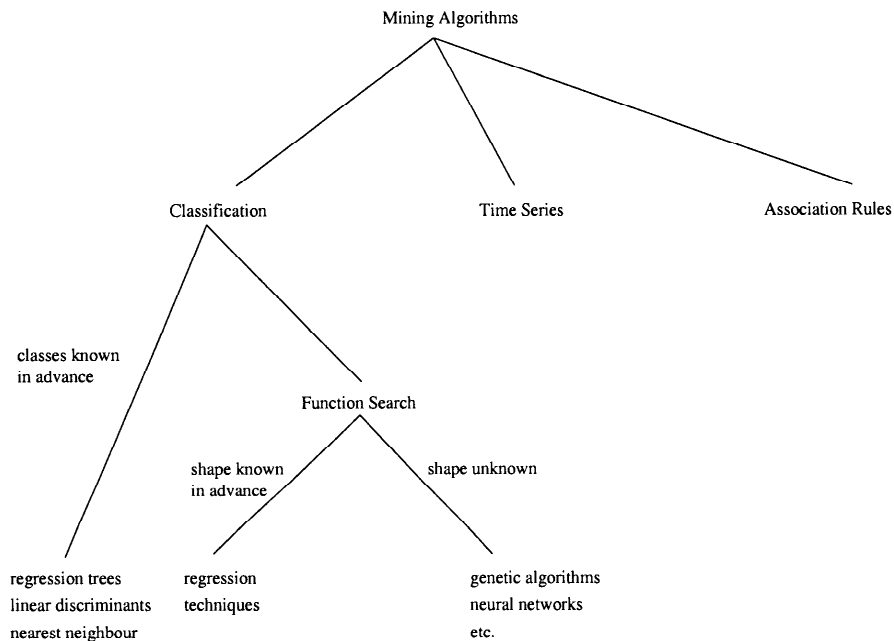
Mining Algorithms

Classification   Time Series   Association Rules

classes known
in advance

Function Search

shape known   shape unknown
in advance

regression trees   regression   genetic algorithms
linear discriminants techniques   neural networks
nearest neighbour   etc.

Figure 4: Module of mining algorithms

confidence 1.0.

Having obtained the itemsets that meet the minimum support, step 2 is straightforward. The solution for step 1 is harder. A simple solution for step 1 is to form all itemsets and obtain their support in one pass over the data. However, this solution is computationally infeasible, since the number of itemsets grows exponentially with the number of items. The challenge is to minimize the complexity and the number of passes over the database. A large number of papers in the literature report on various type of solutions for this problem, exploiting mathematical properties as well as domain knowledge.

4.2.2 Classification

As noted before, classification has as goal to distribute the tuples of a database into a number of, whether or not pre-defined, classes. In general, if the classes are not pre-defined, the term clustering is used. In the remainder of this section, we restrict ourselves to the case that the classes are pre-defined.

Today, the most popular classification techniques are based on decision trees, while evolutionary techniques are gaining considerable attention. Classical techniques, such as regression, kernel density methods, etc. are still appropriate for many data mining tasks. Due to space limits, we briefly characterize decision trees and two evolutionary techniques namely, genetic algorithms and neural networks[3]. Our characterization is based on the following terms: the assumptions on which a technique is based, the quality of the solution produced by a technique, and the "complexity" of a technique. Such a characterization can be

made for classical techniques as well.

Decision trees[21] This type of algorithms picks an attribute as root, and splits it on all possible values, resulting in a tree with depth 2. If all corresponding tuples to a leaf are in a same class C, label that leaf with C. As long as leafs are unlabeled: choose a new attribute and expand the tree by splitting it on all possible values again. This process terminates until all leafs have been labelled. Algorithms based on decision trees differ in the way a choice is made for an attribute that will be split.

Assumption: None.

Quality of solution: The results of decision trees are easily interpreted and are useful as long as the trees are not too large. In general, large trees have a higher misclassification rate than small ones. There are techniques to determine the right size of a tree.

Complexity: The most expensive operation is the splitting of an attribute on attribute values and the classification of all tuples according to these values. Furthermore, a tree grows exponentially with the number of attributes.

Genetic algorithms[19] Genetic algorithms start with an initial population. Traditionally, an individual/object in the population is represented as a string of bits. The quality of each individual is computed, called the fitness of an individual. On the basis of these qualities, a selection of individuals is made. Some of the selected individuals undergo a minor modification, called mutation. For some pairs of selected individuals a random point is selected, and the substrings behind this random point are exchanged, called cross-over. The selected individuals, whether or not modified, form a new generation and the same proce-

---

[3]For an overview of classical techniques in the context of data mining, we refer to [13], which is available upon request.

dure is repeated with the new generation until some defined criterion is met.

Assumption: A representation of a population and a quality measure should be defined.

Quality of solution: Genetic algorithms have been successfully applied in a wide variety of applications. In some cases, it is proven that it converges to a local optimum. In other cases, experiments show that convergence occurs.

Complexity: The most expensive operation is the computation of the fitness function. A genetic algorithm searches different (small) parts of a search space. The complexity is linear with the number of individuals in a population and the number of generations that should be investigated.

Neural networks[17]: A neural network is a function that maps input patterns to output patterns. It consists of nodes and connections between nodes. Nodes are organised in layers, one input layer, a number of hidden layers, and an output layer. A node in layer i is connected with all nodes in layer $i + 1$. The connections are labelled with a weight. The input nodes receive binary values from their environment. The other nodes compute a function from their weighted input and propagate the result. The function looks as follows:

$$V_j = g(\sum_k w_{j,k} * V_k)$$

in which V_j denotes the value of node j, $w_{j,k}$ the weight of the connection between node j and node k, and g is a function. The output of the network strongly depends on the function g. Often g is defined as follows:

$$g = \begin{cases} 1 & \text{if } x - \theta_x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

in which θ_x is a threshold value.

To make the network learn the correct function, we let it adjust the weights using a set of input-output pairs. A simple idea for an adjustment scheme is: if a network gives a wrong answer, the weights are adjusted proportionally to their contribution to the wrong answers.

Assumption: At least one hidden layer is necessary to approximate continuous functions.

Quality of solution: Depends on an appropriate choice of the parameters, such as number of nodes, hidden layers, etc. In general, this is a tough task.

Complexity: The computation of the values that should be propagated is the most expensive operation. Furthermore, once a network is established, the complexity is linear with the input. The number of connections grows also linear with the addition of a node in a layer.

4.3 An Application

At NLR, we have developed and implemented a retargetable data mining tool for classification tasks that is running in a Microsoft environment [11]. Retargetable means that the tool can be integrated with other database management systems, such as ORACLE, without much effort. Our mining tool is currently equipped with a mining algorithm module, a query generator module, and a simple user interface. The mining algorithm module consists of a genetic-based mining algorithm. The architecture of the tool, called SHARVIND, is depicted in Figure 5. The tool takes as input a mining question, which actually selects the part of the database where interesting knowledge should be searched, and possibly requirements (e.g., not to use certain attributes in the mining process) posed by a user. Then, a random number of expressions, called initial population, is selected. We note that an expression is a conjunction of predicates defined over a number of database attributes. The initial population is manipulated by applying the crossover and mutation operators. Since we require the number of tuples that satisfy an individual/expression to compute the fitness of an individual, individuals are translated into corresponding SQL queries which are passed to the MS Access dbms. The fittest individuals are selected to form the next generation and the manipulation process is repeated until no significant improvement of the population can be observed. As output, the tool delivers expressions whose corresponding number of tuples falls in a user-defined interval.

We have mined two real-life databases with the tool. Both databases contain aircraft incident data, one is set up by the Federal Aviation Administration (FAA) in the USA, referred as FAA database, and the other is set up by the Joint Research Centre (JRC) in Italy, referred as ECCAIRS database. We note that the FAA database is obtained from the Internet and is mined at our laboratory, while the JRC database is mined on location. In the FAA database aircraft incident are recorded from 1978 to 1995, while ECCAIRS is recently in production.

The ECCAIRS database consists of 36 relations and about 300 attributes. Two major relations of the database are the *ACS* and the *OCCS* relations. The *ACS* relation contains information with regard to aircraft, such as manufacturer, motor, speed of the aircraft, etc., and information about the environment in which the aircraft is involved, such as weather conditions. The *OCCS* relation describes in general terms an occurrence (incident or accident) and contains general information with regard to an occurrence, for example, time and location of an occurrence, etc. The relation *ACCS* contains 5202 tuples and 186 attributes and *OCCS* contains 5202 tuples and 27 attributes. Although the number of tuples is not large, mining might be interesting due to the large number of attributes.

However, currently 17 relations do not contain any

Figure 5: Architecture of SHARVIND

tuples, while other relations consist of tuples having many NULL values. So, in order to make the database suitable for mining, we have cleaned the database. We have removed attributes that have less than 2000 entries filled in, attributes whose values consist of natural language, attributes that are fully functional dependent on another attribute, and attributes with high and low selectivity factors.

After performing the removals, we have joined the relations ACS and OCCS and 64 attributes were left for mining.

At NLR, the FAA database is implemented as a single table that is sorted on an attribute, called report number, which served as primary key. In the following, we mean by the FAA database, the database as it is implemented and filled at our laboratory.

The FAA database consists of more than 70 attributes and about 60.000 tuples. As in the case of ECCAIRS, this database contains also NULL values, redundant data, and attributes with very high and low selectivity factors. Therefore, we have cleaned this database in the same way as ECCAIRS, in order to make it suitable for mining. After cleaning 30 attributes were left and 60.000 tuples.

We have mined both databases by posing several questions concerning safety aspects to our tool. We start with the question "What are the profiles of risky flights?" We have posed to the FAA databases some additional mining questions (concerning safety aspects), such as "Given the fact an incident was due to operational defects not inflicted by the pilot, what is the profile of this type of incident", etc. We have presented the mining results to safety experts at our laboratory and the overall conclusion was that the an-

swers to the mining questions were correct and promising. The mining results helped safety experts to gain insight in the databases and hopefully also knowledge in future. For example, an unexpected result from the FAA database was the following association:

aircraft_damage is ('minor') ∧ *primary_flight_type* is ('personal') ∧ *type_of_operation* is ('general operating rules') ∧ *flight_plan* is ('none') ∧ *pilot_rating* is ('no rating') → pilot_induced.

This association means that pilots without flight certificates and flight plans who are flying in private aircraft are causing more incidents than other groups. This result was on the first glance a bit strange for our safety expert, since pilots without flight certificates are not allowed to fly. After a while it appeared that the association was correct and our safety expert was able to explain the association. The pilots without certificates appeared to be students whose incidents were recorded in the FAA database as well.

5   Conclusions

We have discussed a number of trends that may be distinguished in the military community. To implement these trends successfully, an adequate processing of various types of information with acceptable performance is required. In this paper, we have briefly introduced the field of multi-media databases and data mining. We have touched on the potentials of these fields for the next generation of military information systems as well as the challenges they entail.

search possible.

References

[1] Agrawal, R., Ghosh, S., Imielinski, T., Iyer, B., Swami, A., An Interval Classiefier for Database Mining Applications, in Proc. of the 18th Very Large Data Base, 1992, pp. 560-573.

[2] Agrawal, R., Imielinski, T., Swami, A., Mining Association Rules between Sets of Items in Large Databases, in Proc. ACM SIGMOD '93 Int. Conf. on Management of Data, 1993, pp. 207-216.

[3] Agrawal, R., Srikant, R., Fast Algorithms for Mining Association Rules, in Proc. Int. Conf. on Very Large Databases, 1994, pp 487-499.

[4] Agrawal, R., Srikant, R., Mining Sequential Patterns, in Proc. 11th Int. Conf. on Data Engineering, 1995, pp. 3-14.

[5] Alsabbagh, J.R., Raghavan, V.V., *Analysis of Common Subexpression Exploitation Models in Multiple Query Processing*, in Proc. 10th Int. Conf. on Data Engineering, IEEE Press, pp. 488-497, 1994.

[6] Boyes, J., Andriole, S., (Eds.) Principles of Command and Control, AFCEA International Press, Washington, D.C., USA, 1987.

[7] Bruggeman, B., N de Reus, Tracking and Prediction: A View to the Future, Proc. AFDRs/WG15 Conf on Maritime Combat Systems Engineering, Portsdown, 1998.

[8] Bruggeman, B., Command & Control- Advanced Reasoning Methods, Proc. C4I Symposium, Den Helder, the Netherlands, 1999.

[9] Choenni, R., *On the Automation of Physical Database Design*, Ph.D. thesis, University of Twente, 1995.

[10] Choenni, R., Kersten, M., Saad, A., van den Akker, J., A Framework for Multi-Query Optimization, in Proc. COMAD '97 8th Int. Conference on Management of Data, 1997, pp. 165-182.

[11] Choenni, R., On the Suitability of Genetic-Based Algorithms for Data Mining, Advances in Database Technologies, ER '98 Workshops on Data Warehousing and Data Mining, Mobile Data Access, and Collaborative Work Support and Spatio-Temporal Data Management, Y. Kambayashi, D.L. Lee, E-P. Lim, M. Mohania, Y. Masunaga (Eds), LNCS 1552, Springer Verlag, Germany, 55-67, 1998.

[12] Choenni, R., Leijnse, K., A Framework for the Automation of Air Defence Systems, To appear in: RTA SCI Panel Symposium on Warfare Automation: Procedures and Techniques for Unmanned Vehicles, NATO RTA, Neuilly-Sur-Seine Cedex, France, 1999.

[13] Choenni, R., de Laat, R., Data Mining: A Brief Introduction, NLR memorandum ID-97-004, 1997, Amsterdam.

[14] Elmasri, R., Navathe, S.B., *Fundamentals of Database systems*, The Benjamin/Cummings Publishing Company, California, USA, 1988.

[15] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., (Eds), Advances in Knowledge Discovery and Data Mining, AAAI/The MIT Press, 1996.

[16] G. Grimmett, D. Stirzaker, Probability and Random Processes, Oxford Science Publications, Oxford University Press, USA, 1989.

[17] B. Kosko, Neural networks and Fuzzy Systems, Prentice Hall Inc., 1992.

[18] M. Lalmas and I. Ruthven. Representing and Retrieving Structured Documents using the Dempster-Shafer Theory of Evidence: Modelling and Evaluation. Journal of Documentation, 54(5):529-565, December 1998.

[19] Michalewicz, Z., Genetic Algorithms + Data Structures = Evolution Programs. Springer-Verlag, New York, USA.

[20] Sellis, T.K., *Multiple-Query Optimization*, in ACM Trans. on Database systems 13(1), ACM Press, pp. 23-52, 1988.

[21] R. Quinlann, C4.5: Programs for Machine Learning, Morgan Kaufmann, USA.

[22] Thuraisingham, B. (Ed.), Handbook of Data Management 1998, Auerbach Publications, Fl, USA, 1998.

[23] H.Turtle, W. Croft, Inference Networks for Document Retrieval, Computer and Information Science, Univ. of Massachusetts, USA.

[24] J. Ullman, Principles of Database and Knowledge-Base Systems, Vol 2.: The New Technologies, Computer Science Press, USA, 1989.

[25] A. de Vries, Content and Multi-media Database Management Systems, Ph.D. thesis, Univ. of Twente, Enschede, the Netherlands, 1999.

[26] Klomp, J., Zetten, H. van, Army Organic Air Defense: Effective and Affordable after 2000?, Militaire Spectator 167(3), 1998 (in Dutch).

This page has been deliberately left blank

_____

Page intentionnellement blanche

Information Processing as a Key Factor for
Modern Federations of Combat Information Systems

Dr. Stefan Krusche, Dr. Andreas Tolk
Industrieanlagenbetriebsgesellschaft mbH
Einsteinstr. 20
85521 Ottobrunn
Germany

Keywords: Combat Information Systems, Data Mediation Techniques, Federation, Standardized Data Exchange Objects

Summary

Building flexible collaborations of different heterogeneous military units just in time is one of the key factors to perform joint and combined operations successfully. It is one of the most important prerequisites to and challenges for information systems to support these collaborations with user-adapted information to the warfighter where it is needed.

For information systems, this is a requirement to adequately build federations of different and heterogeneous data and information sources on the basis of the existing data bases. This objective is only achieved by a new approach towards information sharing based on data mediation techniques which enable the configuration for different information systems towards a global information source to support military business processes across systems, nations and unit borders.

One of the key factors for data mediation techniques is, that it is not an isolated technical solution to gain interoperability between different information systems, but is integrated in an overall data management process, which produces standard business objects for data exchange and standard mediation rules to configure the technical solution. In NATO, this process already has started in form of respective NATO Data Administration Group (NDAG) activities, that aims at the development of standardized data elements.

Data Mediation means to establish federations of heterogeneous data sources on the basis of a common data exchange format while the data and systems itself are kept where and as they are. In other words, on the basis of the data administration and management processes the integration of legacy systems becomes possible without having to change the systems itself. This paper provides an overview of this new integration technique and its relation to already ongoing NATO activities.

Introduction

After catchwords like Enterprise Resource Planing (ERP), Supply Chain Management (SCM), and Customer Relationship Management (CRM), the free market of E-Commerce is actually looking at a new category of software strategy, the so called Collaborative Product Commerce (CPC) [Aberdeen, 1999]. Basic idea is, that instead of the old make-to-stock production model the market is demanding more and more new build-to-demand models. Instead of long planing and introducing procedures, the products are built by tying together the multitude of heterogeneous and geographically dispersed computing systems. This results in faster-to-the-market products that can be influenced during the production process by direct interaction with the customer. Thus, it is possible to innovate with products the customer wants and needs, and of which the respective sources can be delivered.

The technologies that are making CPC possible have been developed over just the last two years. A key attribute of CPC is a loosely coupled integration of data and application functionality, i.e., a common shared data model that does not rely on data commonality for individuals to collaborate. In other words, the information is shared using a common "language" that is not closely tied to one of the participating computer systems or applications.

The Aberdeen Group forecasts great benefits for partners using CPC. It is seen as a key methodology for gaining the best out of the opportunities of the Internet technology.

On the first look, this seems to be a strange introduction for a paper dealing with future Command, Control, Consultation, and Intelligence Systems (C3IS). What does Command and Control have in common with the E-Commerce methodologies and the CPC?

In the eyes of the authors, there are not only commonalties between C3IS and CPC. Obviously, there is a paradigm shift in systems development going on affecting also C3I systems. Componentware and middleware solutions enables the system developer for the first time to build systems coupling reusable components using standardized integration platforms by vertical component integration. Respective management processes insure the alignments of architecture and the semantic consistency of the interchange data. To summarize this, systems development becomes mainly integration of applications delivering the necessary functionality.

Already today, modern C3I systems are comprising increasingly commercial off the shelf products (COTS),

thus, it has to be taken into account what are the main integration strategies of the commercial sector.

In addition, when looking at the operational requirements in military joint and combined operations (including operations other then war – OOTW), the answer to the question of commonalties is as follows:

- **Military Requirement:** The requirements for C3IS are changing rapidly, nearly from operation to operation. The warfighter needs a system that meets his actual needs on time. In addition, the systems must be deliverable in short periods.

  CPC: Innovating with products customers want and sourcing can deliver.

- **Military Requirement:** The requirements have to be brought into the production circle as fast as possible. Many requirements become obvious not before the first actions within an operation. This is especially true for OOTW. Thus, a close connection between the warfighter and the supplier is necessary.

  CPC: Customers can directly interact with the system and the product developers having access to the different prototypes within an evolutionary software development loop.

- **Military Requirement:** All allies and partners from different countries with different systems have to work together interoperable in order to reach the common goal. It will not be likely that everyone is using the same system for "command and control" in a broad sense (e.g., in common operations with the Red Cross, etc.). In addition, Command and Control Systems as well as Consultation Systems may have to stay in the supporting home nation and are not available in the operation area.

  CPC: Tying together the multitude of heterogeneous and geographically dispersed computing systems used today to create, build, and service a product without requiring the enterprise to scrap those systems and start new.

- **Military Requirement:** The systems have to be delivered fast without great additional costs.

  CPC: Faster time to market solutions.

This list should make obvious that the solutions of CPC can be of tremendous benefit to meet the requirements of modern federations of Combat Information Systems. The key technologies that are making CPC possible are available to C3I system developers also:

- High-speed, reliable, secure communications,
- Browser-based, standardized User Interfaces,
- Java or other object-oriented technologies enabling the implementation of reusable components,
- URL-style location transparency,
- Secure portals,
- Server and storage high-end scalability,
- Business functions (applications) delivered as services on demand.

This papers describes how Collaborative Information Processing as a Key Factor for Modern Federations of Combat Information Systems can be made reality for the next generation of C3I systems and how legacy systems can be migrated into this new world.

Data Modeling and Shared Data Models

The CPC is mainly based on a common understanding of the data to be interchanged between the participating systems. In order to benefit from the ideas, a common shared data model for military applications for actual and future operations is needed. Fortunately, a lot of work has been done on this field already that is ready to be implemented.

As pointed out in [Krusche and Tolk, 1999], generally each organization in the domain of defense depends on access to information in order to perform its mission. It must create and maintain certain information that is essential to its assigned tasks. Some of this information is private, of no interest to any other organization.

Most organizations, however, produce information that must be shared with others, e.g., operation plans, location and activity of a given unit, information on the logistics, etc. This information must be made available, in a controlled manner, to any authorized user who needs access to it.

At present, almost every defense information infrastructure exists as a collection of heterogeneous, non-integrated systems. This is also true for C3I systems, and – when trying to bring them together in common joint combined operations – the problem of interconnections even increases. This is due to the fact that each organization builds systems to meet its own information requirements, with little concern for satisfying the requirements of others, or of considering in advance the need for information exchange.

If any information exchange takes place, however, as a rule this information exchange is based on *ad hoc* interfaces. The result is an extremely rigid information infrastructure that costs months and millions to be changed or extended, and, which cannot cope with the increasing demand for widely integrated data sharing between multiple mission-related applications and systems. Actual solutions cannot solve these problems, thus, new ways have to be found in the era of joint and combined operations.

The Shared Data Environment (SHADE) fully described in [DoD, 1996] is a strategy to promote command and control systems' interoperability through a global view on the data of the battlespace, which is made available, in a controlled manner, to any authorized user who needs access to it. The objective is to define a *global infosphere*, that supplies a fused, real-time, true representation of the battlespace, to allow for an integrated data sharing between multiple mission-related applications and systems. The SHADE's technical focus and priorities are driven by near term systems' integration, migration and interoperability requirements that are identified in the Defense Information Infrastructure (DII) Common Operational Environment (COE) context. The main conceptual features of the SHADE address data interoperability for federations of system components and systems in general, not restricted to C3I systems.

Within SHADE, standard data elements (SDE) are defined for information exchange. In order to be able to manage this SDEs, a common shared data model is needed comprising all SDEs and giving them a semantic context.

A data model being able to cope with the requirements for the common shared data model has to have the following qualities:

- It must capture the information requirements of a wide range of battlefield functional areas. A common shared data model is best characterized as a "to-be" model of the required battlefield information rather than a model that is constructed with direct reference to existing current needs for information exchange.

- For flexible integration of future information (exchange) requirements, the data model must be constructed in a way that future information elements simply extend the model while its existing structure remains unchanged.

At has been shown in several publications, e.g., [Krusche and Tolk, 1999; Tolk, 1999], the ATCCIS Generic Hub [NATO, 1996] meets both requirements quite well, as it has been designed to meet exactly these requirements by data modeling experts of almost all nations in NATO during the last 10 years.

As has been pointed out, the definition of standard data elements (SDE) required for information exchange, the coordination and control of their implementation and use within systems have to be the central objectives of an overall data management organization. They may not longer be under the responsibility of system managers who's legal and understandable objective is to optimize their system and, logically, neglecting often the requirements of the superimposed federation of systems.

In general, data management is planning, organizing and managing of data by defining and using rules, methods, tools and respective resources to identify, clarify, define and standardize the meaning of data as of their relations. This results in validated standard data elements and relations, which are going to be represented and distributed as a common shared data model.

The overall objective to be reached by introducing a data management is, to coordinate and to control the numerous system projects technically and organizationally, in order to improve the integrity, quality, security and availability of standard data elements. Due to this objective, the following central tasks of the data management organization are proposed:

- Definition of standard data elements across system boundaries,
- Evolutionary development of a common shared data model as a reference representation for standard data elements,
- Representation of standard data elements through a common shared data model,
- Definition of rules and methods for
  - access, modification and distribution of standard data elements,
  - introduction of new information exchange requirements,
  - Coordination and Control of system projects using the standard data elements in order to assure their consistent use and interpretation within different applications and systems.

To summarize, in order to reach the objective of a common shared data model comprising the standard data elements of the application domain, a common data management organization is essential.

A Framework for Collaborative Information Processing

After having agreed on a common shared data model and the mapping rules for harmonization defined and distributed by the system independent data management organization, data mediation in the sense of automatic translation of system's data into standardized date elements and vice versa becomes possible. Thus, in order to achieve a collaborative information processing based of a common understanding of the data to be interchanged, the key factor is a data management agency, which is responsible for harmonizing legacy data models with the respective common shared data model and the validation and accreditation of the harmonization results.

When using an appropriate toolkit,[1] these results can be used to directly configure a software layer interconnecting the data access layers of different systems with heterogeneous data interpretations. It should be pointed out that the data mediation layer is not an isolated technical solution to gain interoperability between different

---

[1] Respective Toolkits have been developed and applied by the authors in German harmonisation projects.

information systems, but is integrated in an overall data management process. The following figure illustrates the concept. The data management agency harmonizes the heterogeneous data models of the legacy systems. The results are used to configure the data mediation layer that enables the systems to interchange information based on the common shared data model.

To this end, data mediation can be interpreted to be a strategy to implement data interoperability through data transformation mechanisms. These promote common data interchange by mediating individual data representations into a semantic equivalent shared data model representation, i.e., SDEs, and vice versa. The objective is to enable separate systems and system components, which have an overlap of interest, to interchange or share data in a common data representation independent from any system implementation.
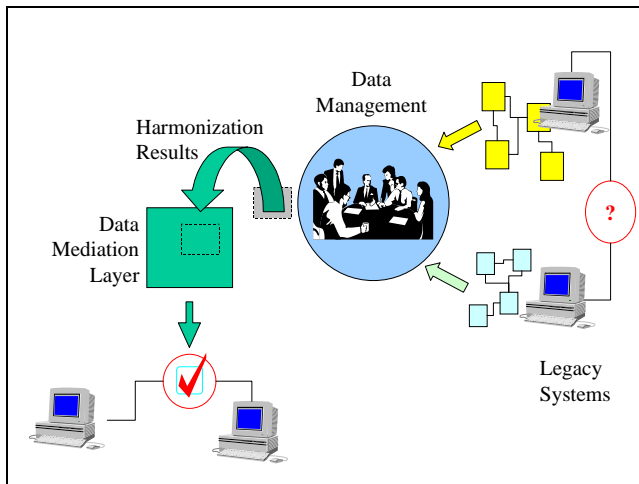
Fig. 1: Data Harmonization in the Integration Process

The data mediation implementation described in this document favors a software framework which may be linked as an additional software layer to existing systems and system components. This framework is a common platform to migrate existing systems and system components and integrate future ones into a shared data model based interconnection network.

It is characterized by the observation that integrating different data representations (schemas) as a prerequisite to build a database system federation, is also a prerequisite for data transformation. Data mediation then is almost equivalent to navigate through such an integrated schema.

The data mediation approach is derived from database federation techniques, thereby, extending these techniques. The database federation approach enables global applications to access different database systems transparently while interacting with a common (global) database schema.

The different underlying database schemas are integrated using a five level schema approach (component, local,

export, global and external schema) as shown in the next figure.[2]

The data mediation approach extends the database federation approach by harmonizing any data representation with an object-oriented shared data model (e.g., ATCCIS) representation using the same integration levels. Data sources are no longer restricted to database systems. Any software component which produces and consumes data is considered as a "data storage medium". With this approach the framework is a common shell for any system component summarizing these aspects in a common software platform architecture.
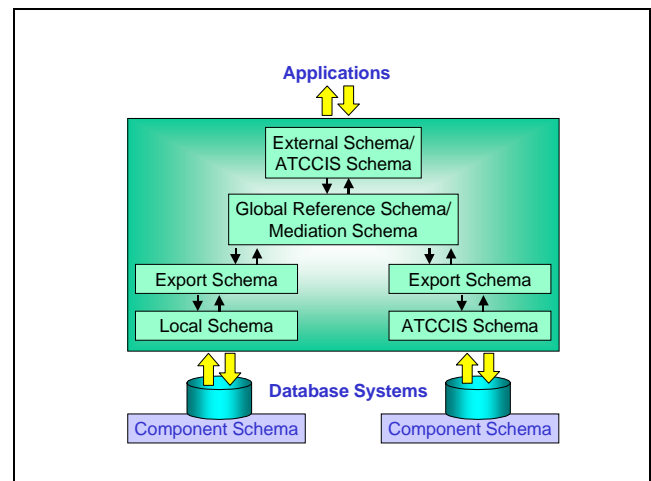
Fig. 2: Five-Level-Schema for Schema Integration

The actual approach implements the framework for collaborative processing as a virtual object-oriented database system, based on a mediation schema (built from an individual data representation and the ATCCIS representation), where the location of the data is completely transparent to the considered systems and system components. An individual system and system component may interact with the data mediation framework using a standard object-oriented database interface. The corresponding data may come from a local database system, from another system component or from a remote system which is interconnected to the current

---

[2] As has been pointed out, we recommend to use ATCCIS as the shared data model, thus, the ATCCIS schema is referenced in the respective figures.

system. Existing systems and system components, linked to the data mediation framework, become encapsulated and provide an common data representation.

Therefore the described framework approach not only permits to build system federations on top of heterogeneous system components and systems but also decouples components of large-scale systems to enable these components to evolve independently.

The approach given in this paper enables any system component and any system with an individual data representation to be represented by a shared data model representation. This, however, requires to first harmonize individual data representations with the agreed standard schema, which has to be done by the system independent data management agency described in the former section.

From an architectural point of view, the framework is divided into a common mediation kernel and multiple interconnection cartridges as shown in figure 3. These two architecture element types can be defined as follows:

- **Common Mediation Kernel.**
  The mediation kernel manages the object-oriented mediation schema and provides services to navigate from an individual system or system component schema to an standardized schema (e.g., ATCCIS) and vice versa.

- **Interconnection cartridges.**
  The interconnection cartridges are customized to support interconnection to applications and multiple database systems, and communication middleware products such as the Common Object Request Broker Architecture (CORBA) or the Runtime Infrastructure (RTI).[3] For ATCCIS-based database replication, the ATCCIS replication mechanism (ARM) is adopted.
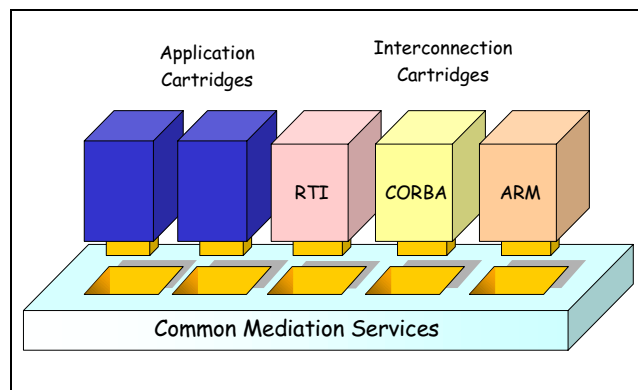
Fig 3: Interconnection Cartridges

---

[3] The RTI is defined within the High Level Architecture (HLA) as one element of the Common Technical Framework (CTF) being the standard for interoperable simulation systems [NATO, 1998].

This finally reflects the design goals of the data mediation framework as an open and flexible add-on to functionally enrich existing communication middleware architectures like CORBA or the RTI, enabling a general approach for systems interconnection.

Federated Solutions of Heterogeneous Systems

The information techniques and management procedures having been described so far enables a new way of C3I system development.

As a first step, federations of heterogeneous information systems can be build. In order to do so, the information to be interchanged while an ongoing operation has to be specified using the agreed common shared data model. The nest step is the harmonization of the exterior data view of the respective participating systems.

Using the appropriate toolkit, a software layer can be implemented, that enables the respective system to send and receive information using standardized data elements. Thus, the needed functionality can be reached by coupling different systems together in a loose way.

In the operational context, this means that the different partners within an alliance can use there own C3I systems for all functions that are supported and other systems for the rest. If, e.g., an additional communications server is needed, this one can be coupled with the rest of the federation using standardized data elements. As long as the information exchange requirements are within the scope of the external data view of a system, they can be fulfilled.

This is not only true for military information systems, but also for IT support of humanitarian relief organizations (e.g., Red Cross), governmental or non-governmental organizations, and every potential partner. Thus, the framework enables interoperability on the semantic information level between systems.

Next step should be the use of the ideas for intra-operability also, i.e., different applications supporting the warfighter by offering a special required functionality can be integrated into a new information system that can be adapted to the needs of a given operation. If an additional functionality is needed, the supporting application can be enriched with the data mediation software layer. Being then able to exchange information using SDEs, integration by respective cartridges is easy.

Finally, using these ideas consequently, the differences between inter- and intra-operability vanish. It doesn't matter any longer whether a needed function is implemented by the own system or an allied one, as the techniques bringing all applications together in the heterogeneous federation are equal for systems and applications.

The following figure exemplifies this. The framework for collaborative information processing can be used to couple the legacy systems A and B, hence coupling the respective comprised functionality, as well as to couple new applications to legacy systems or building new systems comprising only new applications.
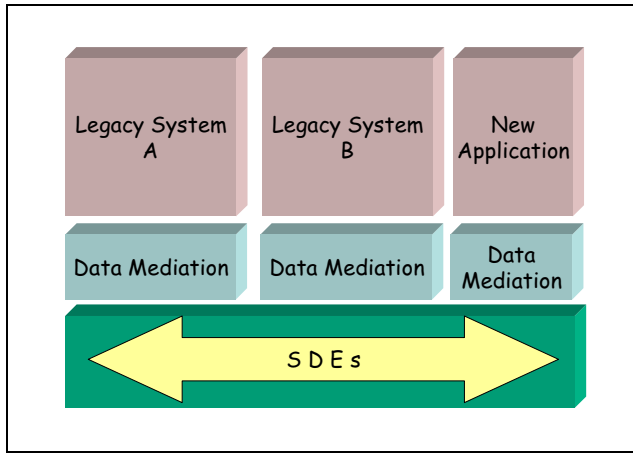
Fig 4: Intra- and Inter-Operability

To summarize these ideas: The intelligent use of standards is leading to synergy out of heterogeneity. Standards are not equalizers but reference concepts to be used to reach a common understanding of what's going on. Thus, participating allies within an operation do not have to change their systems, and it is not necessary to build a new system for every type of operation, but adaptable and configurable open and extendable solutions comprising reusable components become possible. On the long term, a library of functionality will become available to be used to couple together exactly the type of information system that is needed for a given operation. This can be even done "on the fly", i.e., from the home nations during an ongoing operation.

The Next Generation of IT Systems

There still may be the opinion that all of this becomes obsolete with the next generation of IT systems to support the warfighter. The authors want to point out that this will definitely not be the case. The need for data harmonization, data management, and data administration will not vanish, even if the next IT system generation comprises common data distribution and communication facilities for a set of common functions and applications. Looking, e.g., at the architecture worked out within the Defense Information Infrastructure (DII) Common Operating Environment (COE), all applications are capsulated and integrated into the new system. No application has to deal any longer with external data consistency, storage of data, transmitting data, etc. All this is done by global common functions of the DII COE.

However, there is still the need of data translation between the functional applications and the core data. Again, the

ideas of data mediation can be introduced to bridge the gap of semantic interpretation between the application and the core data. In addition, the requirements having to be fulfilled by the core data model are quite the same as having been mentioned before:

- All data needed by any application has to be stored.
- The integration of applications should leave the already stored data as it is.
- The introduction of new data and data types should be possible in order to fulfil all information exchange requirements.

Therefore, the idea is obvious to use the same information structures having been developed to support the data management for operational systems as the core data model also. The NATO C3 Data Model [NATO 1997] uses the principle of properties and propertied concepts being only loosely coupled within the data model also. Taking this idea further and fusing it with the advantages of the ATCCIS data model and the insights gained from the experiences with data management leads to a new concept of operational data models that are able to be configured within the operation without any shut-down or re-boot of the system, New applications can be introduced, new data interpretations can be down-loaded during the operational use. To do so, the same mechanisms having been introduced to configure the data mediation facilities are now used to configure the operational data base.
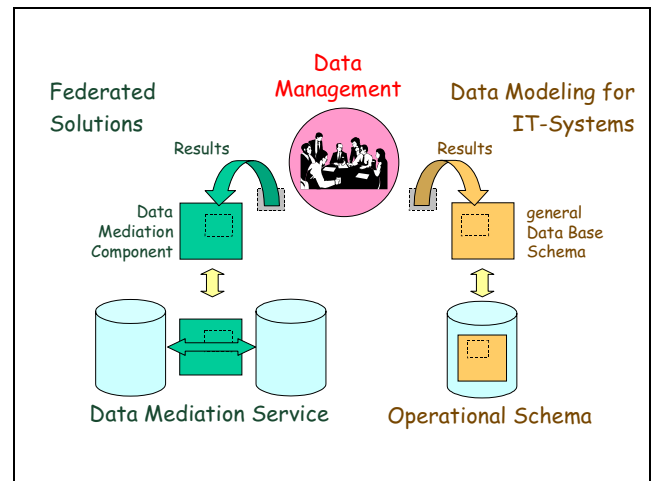
Fig 5: Configuring Operational Schemas

To summarize this: The authors have found a way to meet the information needs of the military user whenever they are defined, may it be in advance of an operation, in the preparation phase, or even when the system already is in use. Therefore, new applications – be they specially developed, brought in by partners or allies, or being commercial off the shelf products – can be integrated at any time within bringing the system back to the industry.

The UK has already made positive experiences with a very similar approach, and Germany is making first efforts to

introduce this new data modeling technology into their new operational systems.

Conclusions

The ideas of CPC are supposed to have a revolutionary effect on the Internet market enabling heterogeneous institutions and systems to develop systems with the consumer in the loop in a very efficient way. The underlying new technologies can be used also for collaborative information processing in military operations.

Kernel idea is the definition and use of a common shared data model for information exchange. This idea is not new to the military community. The SHADE [DoD, 1996] is a strategy to promote systems interoperability through a global view on the data of the battlespace. The SHADE's technical focus and priorities are driven by near term systems' integration, migration and interoperability requirements. Its main conceptual features address data interoperability for any federations of system components and systems in general, and thus, also provide an adequate approach towards the interconnection of C3I systems. Furthermore, in this paper, the necessity of an overall data management organization to effectively manage standard data as an operational asset has been stressed.

A framework for collaborative information processing as a common software platform, to meet the migration requirements of existing system components and systems, has been introduced. It implements standard data and mappings and allow users to access and interchange *as-is* data without knowing information about the common, standard data representation. On a mid term, the technique of data mediation will improve the migration of legacy systems in a cost efficient way.

It should be pointed out that only if all three columns of shared solutions are supported, this will finally lead to the desired success. These three prerequisites are:

1. A system independent data management agency has to be responsible for the standardized data elements and the common shared data model as well as the validation of the harmonization results of mapping legacy data to respective SDEs.
2. The toolkit family used by the data management agency and other people doing harmonization must be able to export the mapping results in a form readable to the framework for collaborative information processing.
3. All systems and applications must be integrated using the additional data mediation layer as well as the framework for collaborative information processing.

This leads to real federated solutions as a new C3I development paradigm meeting the new requirements emerging from joint and combined operations including operations other than war.

The ideas of data management are not to become obsolete when a single common system – e.g., the next generation of the Global Command and Control System (GCCS) or similar solutions resulting from the DII COE efforts – are introduced as one common systems to be used by all military services and/or all nations. It becomes even essential to ensure the meeting of user requirements for flexible IT support even in unpredictable operations by flexible and configurable core data capabilities. In this sense, the data management would offer the tools and procedures to enable the operational schema to support every function and application with the data it needs in the form it needs the information.

To conclude this paper, the last recent events concerning data administration and management activities in NATO should be given. Actually, ATCCIS is prepared to become the NATO Standard (STANAG) ADatP-32 [NATO, 2000]. In order to do so, in February 2000, the NATO Information Systems Sub-Committee (ISSC) tasked the NATO Data Administration Group (NDAG) to use ATCCIS as the reference model for the NATO Corporate Data Model (NATO CorpDM). Until the end of this year, these efforts are planned to lead to the first joint version of the NATO Corporate Data Model based on the ideas presented in this paper.

References

Following books and articles are referenced in the paper:

[Aberdeen, 1999] Aberdeen Group, Inc. *Collaborative Product Commerce: Delivering Product Innovations at Internet Speed.* Market Viewpoint Volume 12/Number 9, Boston, Massachusetts, October 7, 1999

[DoD, 1996] Defense Information Infrastructure (DII) Shared Data Environment (SHADE), CAPSTONE DOCUMENT, U.S. DoD, DISA, 11 July 1996

[ISO, 1990] ISO Standard IS10027:1990. *An Information Resource Dictionary System (IRDS) Framework*. 1990

[Krusche and Tolk, 1999] Stefan Krusche, Andreas Tolk. *A SHADE Approach for Coupling C4I Systems and Simulation Systems.* Paper 99F-SIW-004, Proceedings of the Simulation Interoperability Workshop Fall 1999 (SIW F99), Orlando, Florida, September 1999

[NATO, 1996] ATCCIS Working Paper 5-5 Draft 1.0. *ATCCIS Battlefield Generic Hub 3 Data Model Specification*. ATCCIS Permanent Working Group, SHAPE, Belgium, November 1996

[NATO, 1997] The NC3DM Data Model, Draft version 0.2, 28 February 1997, Task Group for Tasks 1.C.2 and 1.C.3, NATO Ad Hoc Working Group on Data Management

[NATO, 1998] NATO Document AC/323 (SGMS) D/2. *NATO Modelling and Simulation Master Plan (Version 1.0).* NATO HQ, Brussels, August 1998

[NATO, 2000] ADatP-32, Edition 2.0. *The Land C2 Information Exchange Data Model.* NATO HQ, Brussels, March 2000

[Tolk, 1999] Andreas Tolk. *Using ATCCIS as an Information Layer to couple CGF Federates and Closed Combat Simulations.* Paper 99F-SIW-008, Proceedings of the Simulation Interoperability Workshop Fall 1999 (SIW F99), Orlando, Florida, September 1999

Contact

E-mail:	sk@dm-forum.org	(Dr. Stefan Krusche)
	at@dm-forum.org	(Dr. Andreas Tolk)
URL:	www.dm-forum.org	
Mail:	IABG	
	Einsteinstr. 20	
	85521 Ottobrunn	
	Germany	

Architecture for Flexible Command and Control Information Systems (INFIS)

M. Wunder

Forschungsgesellschaft für Angewandte Naturwissenschaften (FGAN)
Neuenahrerstrasse 20
D-53343 Wachtberg-Werthhoven
GERMANY
Tel: +49 228 9435 511
Fax: +49 228 9435 685
E-mail: wunder@fgan.de

1   Processes of Change

Next generation CCIS will be leaner and more flexible. The previous trend to decentralise the computer power increased the complexity of local devices. Now it is more and more difficult to achieve the requirements of higher mobility and simpler access to information services.

The available bandwidths will raise. Communication via radio transmission will be extended. Robuster and more mobile devices use radio interfaces and will be established as standard. Due to the quick deployment of the internet a new network oriented client/server technology arises. The requirements for front-end devices change. Information services were transferred to decentralized server machines. All required services can be accessed from different local points by authorized users. Miscellaneous inhomogeneous technical characteristics of various information pools, that are time consuming and very annoying for users will no longer be relevant. The necessary expense for the administration of local devices will decrease and will be similar to the currant situation of mobile telephone devices – almost zero.

The change of the political conditions for military engagements causes the reduced importance of geographical distances. Decision cycles will decrease. In all sorts of military orders specialists are flexibly gathered for short time engagements. The importance of mobility will increase.
More and more often, specialists are gathered together at short notice for fast progressing operations. These persons should be supplied with easy to use but robust IT systems which require a minimum of expenditure for installation and configuration and support an operation of a partially mobile group distributed spatially.

The information processing will face the following requirements: Information have to be available in a short-term. Nearly any information must be accessible to the military leader. All different kinds of information, whether structured/unstructured or formatted/ unformatted or data-based that are necessary for the complete analysis of any particular situation, must be individually linkable.

Requirements for Command and Control Information Systems

- **Maintenance**
  Minimum or zero expense for system management, lean client

- **Interoperability**
  Among heterogeneous CCIS of other countries

- **Mobility**
  Any information accessible from any location via mobile devices.

- **Reliability**
  High safeguarding against failure, stabil Architecture

- **Portability**
  Plattform independent access to the CCIS

- **Cost**
  Minimum operation prerequisites, browser and access to internet / Intranet

- **Extendability**
  High life span, easy reaction to new functionality, easy replacement of COTS products

Picture 1

That requires an extension of common CCIS data. The data for the description of e.g. "Situation of Own Troops" must be completed by links to corresponding e-mails, faxes, videos, graphics etc. These links must be variable and managed automatically. From the user's point of view, a homogeneous application system provides a complete and consistent information about all relevant facts.

The information exchange between military units must be manageable and should run automatically for horizontal and vertical information flows. Depending on the purpose, the used front-end devices must be scalable from laptops to e.g. mobile telephones. The IT-administration ought to be unnecessary. Mobile devices must be robust, simple to use, easily exchangeable and rapidly employable.

2 Chances with Internet-Technology

The internet technology enables new scenarios for communication and cooperation of user teams that interact across long distances. It offers attractive means to access a CCIS. Nearly any kind of electronic storable information is accessible through an intranet or the internet. Variant types of communication are no longer limited to particular media. The internet handles voice communication, video telephoning, emailing, file exchange etc. The browser (COTS) is a universal tool within the web and easy to use. It enables the similarity of user interfaces. So the expense for teaching and operating of heterogeneous applications can be reduced.

Due to the working standardisation, the open interfaces, the open protocols and the open data formats within the arising systems and communication platforms are widely unique. Products of various suppliers are widely combinable and exchangeable. The communication between systems of different suppliers works quite well. This kind of a working and complete standardisation occurs not very often in the history of information processing.

Another relevant development, which is closely linked to the internet, is the standardisation of Java by the "Open Group", where all important suppliers and large user companies of information and communication technologies participate. The reason for Java's success is especially the platform independence of Java applications. Appropriate environments ("virtual machines") are available for almost every platform. All these virtual machines can operate the same Java-Byte-Code, that was compiled once from a single source code.

The internet technology offers considerable improvement potentials for all kind of business processes. But as in the case of all new technologies, we have to apply the means very carefully, because the enthusiasm of the technicians and customers which easily fall in love with the new things should not lead into the fatal situation, that the technique is in the foreground of considerations, but not the operational process that has to be supported by applying a technology.

3 Interoperability

The interoperability of heterogeneous applications can be eased by using internet technologies. For the lowest level of interoperability different applications can run in windows of a single browser. A further integration can be achieved when data of different data bases are collected and processed within a XML page. A sort of an intensive integration of heterogeneous applications is possible by using distributed agents. The properties of agents can be: intelligent, learning, interacting and if necessary mobile. In order of a user they can independently handle complex tasks.

But no one should expect miracles. Even if the integration of information from different heterogeneous sources is technically possible, one problem has to be solved always: the identical interpretation of information by the sender and by the receiver. In this context it seems to be more than doubtful, that it is possible some day to have a "machine" automatically interpret and automatically translate more than trivial facts.

In opposite to this consideration the consequently operated standardisation of logical semantics is a certain way to achieve interoperable systems. In that case ATCCIS[1] (Army Tactical Command and Control Information System, NATO-project) has established developing specifications for a unique replication protocol to share data automatically between different command control systems based on an agreed conceptual data model.

4 Future Front-End-Devices

Few years ago the NC (Network Computer) was supposed to reduce the necessary management efforts. These efforts came up due to the increasing decentralization and number of installed PCs. Frequent troubles among the PC environment caused by instable operation systems and poor network software required intensive personal efforts for system management tasks. An increasing complexity of applications needed higher performing devices though just a small amount of functions was actually used.

There are several reasons for the actual silence around the NC. On one hand the current operation systems are more stabile and better management tools support the remote maintenance of local devices.
On the other hand all requested facilities have already become reality due to the increased utilization of internet technologies. It is possible to choose between local, server based or web based

work. This is important for the data administration and the access to the application. There is an actual movement in the direction of web based work.

5 Transmission

Future networks must transmit nearly all sorts of information: voice, data, video etc. Today we can observe a competition between ATM and IP-based networks. Also a combination of IP via ATM might have its way. But probably the importance of pure telephone networks will decrease. Future investments will extend the capacities for combined voice and data transmission.

Fiber technologies can enable bandwidths of more than 100 Gbit/s in the future. In tests transmission capacities of several tera-bits/s could already be achieved for distances of several kilometres. Due to actual plans an enormous increase of capacities seems to be expectable.

Particularly the radio transmission is going to obtain an increasing importance. It provides advantages for handling and flexibility. The GSM (Global System for Mobile Communication) technology, which is used for mobile telephones offers a bandwidth of 9,6 kbit/s. That is sufficient for voice communication but not enough for further requirements. The package oriented GPRS technology (General Packet Radio System) supplies up to 115 kbit/s. An advantage for data transmission is expected because the transmitted volume can be charged instead of the time for a connection.

Today different technical equipments have to be used, depending on the available radio cells. In example DECT is used in small cells like an apartment. Bundle transmission is used on a production site, GSM over land and INMAR-Sat over sea.

Starting in 2002 UMTS (Universal Mobile Telecommunications System) is supposed to supply one technique for all kinds of radio cells. Bandwidths up to 2 Mbit/s are planned for devices. This commerce driven technology is an incentive for an increasing usage of mobile devices.

6 Architecture Proposal

The proposed architecture is an "Integration Platform for Flexible CCIS". It is called **INFIS**[3] (in German: **IN**tegrationsplattform für **F**lexible Führungs**I**nformations**S**ysteme).

The INFIS architecture is based on COTS products (hardware, database, operational system, network, middleware, browser, programming-language and referring tools) and standard concepts for the database (ATCCIS).

INFIS includes a CCIS as the essential component. It essentially contains the features: situation management, replication mechanism, flexible contract management for the replication mechanism, message handling, mobile agent controlling for the distribution of client software, database handling.

The internet technology enables the platform independent[2] access of any stationary or dislocated devices to any INFIS service like the database applications of the leading CCIS or any other application e.g. for message management, graphical situation display and in the future for integrated office applications, document management and so on.
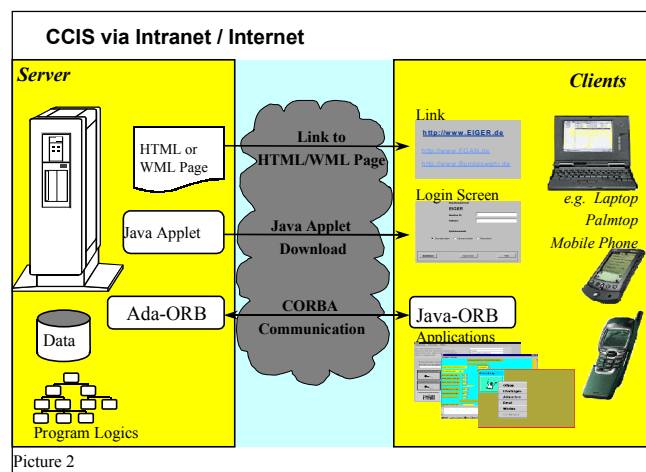
Picture 2

The first action to access INFIS on all sorts of devices is the download of an applet from the server (pict. 2) via internet / intranet. From that time on the communication runs via the middleware CORBA (Common Object Request Broker Architecture).

The use of CORBA within the structure of a flexible system provides a relative independence of server and client. The encapsulation on side of the server enables the usage of any hardware, any operation system, any database operating system and any programming language. Even legacy applications, that still meet the actual requirements, might be operated and used via modern graphical user interfaces on side of the client.

A browser enables the access to the application logic, which itself controls the access to the structured database. The applet, downloaded first, consists of the client functions like reasonable checks, format handling, data handling. It also handles the communication to the server. Now the format specifications and data can be sent from server to the client. If a local storage is practical, the format specifications, which have been downloaded in former sessions, are used (pict.3).

Picture 3
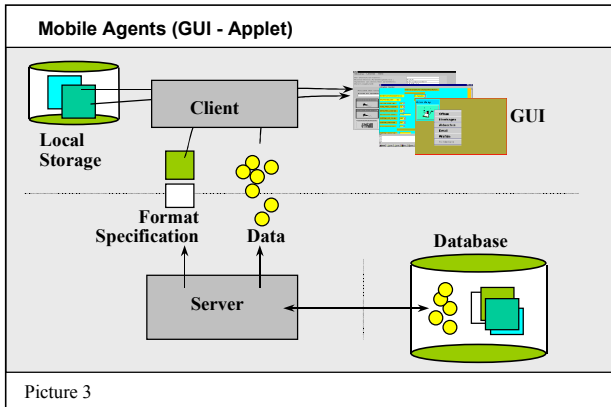
Due to that architecture, the perpetual transported data volume between server and client is small. However, this advantage faces a uniquely longer charging time during the first download of the applet onto the client.

If in case of any software maintenance the database or the application is extended or changed, there is nothing to do on side of the client. This advantage supports the mobility and independence of the users.

As mobile devices conventional laptops, palmtops or mobile telephones can be used. The choice of the device depends on the required functionality. In all cases the single prerequisite on the device is a common browser – any installation of further software on the device can be omitted .

Conventional laptops use conventional browsers like Netscape or Explorer (pict. 4). At present the WAP technology (Wireless Application Protocol) is applicable for small devices to transfer graphics or texts. WAP is comparable to the common internet technology. WAP uses WML pages (Wireless Markup Language). The structure and usability of WML is similar to HTML (Hyper Text Markup Language) pages.

Nowadays the used path (Applet or WML) depends on the selected device. The retrieval and the data input into the structured data base is possible in both cases.

Picture 4
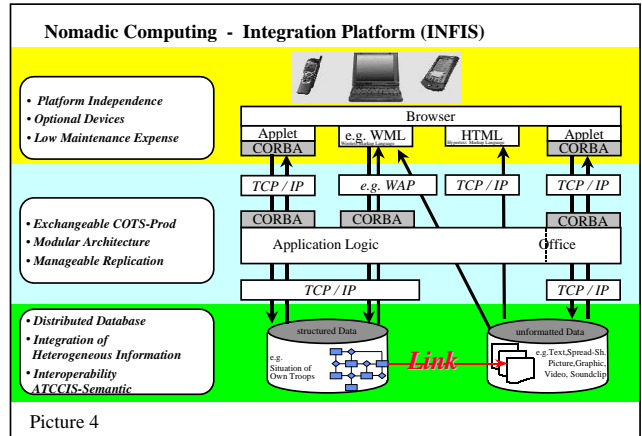
For the near future the mobile telephone producers have announced a Java Virtual Machine running on their devices. If that environment is available on mobile telephones, the WAP technology can be replaced by the well known internet technology. That means, that the same program techniques can be used for every device. Programmers must only consider the size of the display.

To create or manipulate unformatted data (text, pictures, graphics, videos, sound clips, ...), an appropriate application e.g. an office software is required. At present an extra installation of a proprietary office software (e.g. MS-WORD) on the local device is required. Several suppliers intend to ship office products that can be used via internet in the future. These products require only a browser on the local device. In that case it is possible to retrieve documents via a browser and also to manipulate documents via the browser without (or at least with minimum) requirements for local software installation.

The INFIS architecture enables the extensibility through other applications for particular purposes (pict. 5): office products, e-mailing, document management and so on.

The CORBA connection encapsulates the additional systems. The communication between INFIS and each of these COTS systems is implemented in pairs of CORBA interfaces[2].

The prerequisites for the extension are:

1. the additional COTS systems have a 3 tier architecture,
2. they are able to run in a browser and
3. they have an common programming interface or better a CORBA interface.

If one of these COTS products should be replaced in future by a more powerful one, the pair of

CORBA interfaces on at least one side must be adapted.

Faced with the fast evolution of internet technology and the circulation of Java applications, it seams to be likely, that in future many COTS products meet these requirements.

Extensible Architecture of INFIS

Picture 5
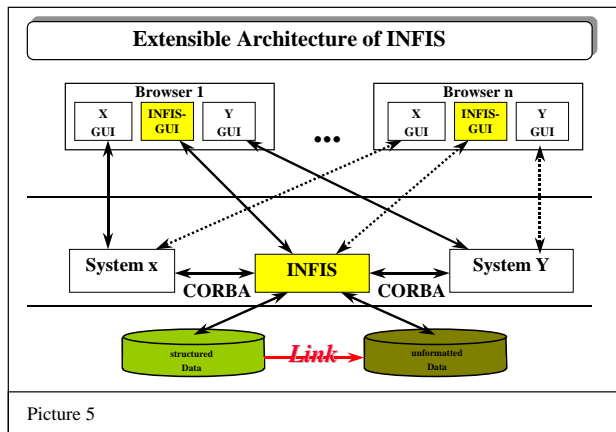
It is planned to advance the INFIS architecture for the following purpose: The user should be allowed to manage individual links to any information. These information must be flexibly linkable depending on the particular operational situation. In this case, the leading system within the INFIS-architecture, the CCIS must manage the links in connection to referring situations in the database. I.e. it might be helpful for the military leader to link the daily internet weather report to the corresponding operational situation or to link a video clip from an area of interest with the operational planning.

Features of the INFIS - Concept

- **Cost effective**: any (even mobile) Devices, Browser, access via Internet / Intranet
- **Minimal/none Expenditure for Administration** due to Lean Clients, (in the future: Mobile Phones too)
- Integration of **heterogeneous Information** (structured, unformatted)
- **International Interoperability** to heterogeneous Systems (ATCCIS-Standard: Replication Mechanism, Conceptual Core Data Model).
- Dynamic Architecture and easy Replacement of COTS-Products: **high Life Span**, flexible Maintenance, fast Reaction to new Operational Pequirements
- **Portability** due to Hardware- and Operation-System Independence

Picture 6

The regard of international standards – especially in the area of semantic standardisation of objects occurring on the battlefield and their relation with each other – promotes the interoperability to other systems. INFIS implements the concepts for the ATCCIS data base and the replication mechanism.

A consequent modular design enables a flexible architecture and the easy integration and the replacement of single (COTS-) components. That allows the easy maintenance and fast reaction to new requirements coming up in the future. That again is the preposition for a long term actuality and utility of a CCIS.

Due to the typical causalities of procurement and life cycles of military information systems this feature is considerably more important than in the case of civilian systems.

Via a military intranet, the access to sensible information can be controlled and the management of the information pool can be well coordinated.

If an e-mail system is part of the mentioned architecture, the information distribution (push-principle) to determined receivers and the active receipt (pull-principle) of requested information from a distinct source is possible. Command and control processes may become leaner, quicker and over all more efficient.

7   Future Developments

Computers get smaller and more powerful. Hard drive capacities increase while their size gets smaller. Nearly every equipment gets a better performance and more memory. In opposite to these enhancements the man machine interfaces of computers are still very poor. Computers are forcing users to accept its rules, which often are rather complex. The man machine interfaces of other very complex technical devices which surround us all day, are better adapted to the ergonomic needs of human beings. E.g. a lot of electronic gadgets control the driving unit in a modern car. The actual load of the motor, the fuel quality, the compound of the exhausted gas, the acceleration and so on. All these values must be computed, but they remain concealed for the driver. He "programs" his car just with the movement of his right foot. That is in fact an ergonomic interface.

In accordance with the mobility of computers additional requirements arise. It is still necessary, that palmtops or laptops have to be operated intentionally. Another activity must be interrupted to lead the full attention to the computers man machine interface.

Future computers can be used while other activities are carried out. An actual, primitive example is the

usage of mobile telephones while driving a car. The voice control is an improvement to the user interface. The driver uses the telephone services and keeps on driving just with little distraction.

This is just the beginning. Computers that are imperceptible and wearable like watches on the arm were controlled by the voice or collect input parameters automatically. The output might be handled via transparent visors or via a synthetic voice. These kinds of interfaces can support a new kind of interaction with computers.

Further considerations have to be made in order to develop architectures that support flexible military information systems, which allow an information management of dislocated users with mobile and easy manageable systems .

8   Literature

1   Army Tactical Command and Control Information System (permanent), SHAPE Policy & Requirements Division, Mons (Belgium)

2   Bühler, Fassbender (1999); Applying Ada, Java, and CORBA for Making a Command and Control Information System Platform Independent; SIGAda'99   10/99 Redondo Beach, CA, USA

3   Wunder (2000); Architektur für flexible Führungsinformationssysteme,   IT-Report   2000, Report Verlag, Bonn-Frankfurt

The Rabi Quantum Computer

Rudolph A. Krutar
Advanced Information Technology[†]
U.S. Naval Research Laboratory, code 5583
Washington DC 20375-5320
USA

Summary: Thomas Carlysle once wrote: "Our main business is not to see what lies dimly at a distance, but to do what lies clearly at hand." Totally ignoring that advice, I set about to examine how a quantum computer might eventually be exploited. (A quantum computer uses strange but real effects in Quantum Mechanics to explore many possibilities at the same time with the same hardware.) There are currently just four known quantum computer algorithms. The possibilities are staggering, but are not likely to be realized by attending only to what lies clearly at hand. Yes, we might be able to factor large numbers, search databases in fewer steps, evaluate simple global properties of arbitrary predicates, and simulate quantum systems. But we might also be able to factor complicated problems, search databases of large images and sounds, evaluate interesting properties of complicated functions, and simulate sound and weather systems. No one currently has any idea how to do anything that complicated on a quantum computer. The research community therefore needs to focus on what has to be done to liberate the imaginations of students everywhere, so they can rework the foundations of their fields to exploit the exponential potential of quantum computers.

Prognostication: Our crystal ball is cloudy. Looking through it smears the image. Nevertheless, I will make some guesses using humor as a rhetorical tool. I am more interested in conveying the awesome potential of quantum computing than in explaining all the details, many of which are well explained in the cited references. Someday these guesses may be improved through formal forecasting techniques such as the Delphi Method[1]. Eventually quantum computers may help make such guesses more accurate. Let the reader be aware!
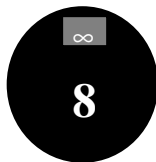
> By ignoring the longer time horizon we may hope that additional options or solutions to currently unsolved problems will materialize, that the need to make a decision will vanish, or that the responsibility for a decision will be in other hands. -- Harold A. Linstone[2]

Another book on forecasting methods[3] claims that people have an innate ability to foresee the future, embodied not so much in our ability to predict surprises, but in our ability to plan ahead. It also cites an old Arab proverb that "he who predicts the future lies even if he tells the truth." The lies I am about to tell you may be true.

Quantum Effects: The universe at quantum scales is weird, as shown by the well known double slit experiment. Shining light of a given color through two separated slits in a barrier onto a screen behind the barrier results in a pattern of alternating light and dark stripes on the screen (see Fig. 1). This pattern is known as a diffraction pattern and results from the constructive and destructive interference of the light waves from the two slits. This is as expected, however experiments[4] have shown astounding results. If we dim the light source so that only one photon is emitted at a time, the photon strikes the screen at some random point with probability related to the intensity of the diffraction pattern at that point. In other words, if we replace the screen with photographic film, and send a stream of photons one at a time, we record the diffraction pattern. As Klein explains on p.120 of his book[5], the diffraction pattern disappears if we block one of the slits each time a photon is sent or if we somehow determine which slit the photon goes through. For example, the standard textbook theory[6] is that any particle has an associated probability wave that limits its possible trajectories, and this probability wave explores all possible paths. This theory is not an explanation. Some aspect of each photon somehow goes through both slits at the same time, provided we cannot determine which slit it goes through. This astounds me!

How Can That Be: No one knows for sure how Nature can possibly behave this way, but it does[7]. There are several theories and many books that expound on those theories, including Klein's book[5] that examines the role of paradoxes in the development of physics. The foremost theory is the

---

[†] This paper was written while the author was undertaking long-term training at North Carolina State University.

Copenhagen school: "Don't ask, just do the math!" Many physicists disparage my favorite, the Multiple Universes (Multiverse) Theory, but some show its simplicity[4]. We sometimes cannot know what really did happen, simply because Nature forgot. Nevertheless, we all assume the past is unique in most of our thinking. We soon forget what we had for lunch a week ago, but we are confident that the answer is unambiguous. The exact number of carbon atoms we ingested may be ambiguous, and we have no way of knowing exactly what the course of events (or food) really was. We live in a world of approximations. The past may not be unique. What other unwarranted assumptions do we blithely make?

That mathematicians throughout the ages should have made various mistakes about matters of proof and certainty is only natural. The present discussion should lead us to expect that the current view will not last forever, either. But the confidence with which mathematicians have blundered into these mistakes and their inability to acknowledge even the possibility of error in these matters are, I think, connected with and ancient and widespread confusion between the *methods* of mathematics and its *subject-matter*. -- David Deutsch[4].

What's a Qubit?: We can replace the photons in the two-slit experiment with electrons, protons, atoms, even buckeye-balls, and still get a diffraction pattern. Similar ambiguities in Nature result from other phenomena, such as the polarization of light and the spins of subatomic particles. These spins can be controlled by resonant microwave radiation. Any such ambiguity can be exploited to construct a quantum bit or 'qubit', which is like a conventional bit in that it can store one of two states, except that a qubit can store a mixture of two states as long as we do not know which state it stores. When it is observed, a qubit assumes one of the two states.

Quantum Computers: A quantum computer is a device that exploits qubits (however constructed) to explore several possibilities at the same time with the same hardware. Williams and Clearwater[8] have explained the theory well. Whereas one qubit has a superposition of two states, two qubits have a superposition of four states, three qubits have a superposition of eight states, and so on, so that N qubits have a superposition of 2^N states. The qubits can maintain this superposition without interacting with each other as long as outside forces do not disrupt the coherence.

NMR Quantum Computers: Quantum computers can be built in several ways, and ultimately in ways not yet considered. Nuclear magnetic resonance (NMR) has been used to build some of the first functional quantum computers. The idea is that each molecule in a solution is a quantum computer with some of its atoms constituting the qubits. For example, alanine has three carbon atoms, which can be replaced by carbon-13 atoms with spin $1/2$. Measuring the spins of these atoms in a strong magnetic field will show them aligned either with or against the magnetic field. Each atom in the molecule has a resonance frequency, and applying resonant microwave radiation at that frequency can modify its spin. All the atomic spins normally precess in their local magnetic fields as affected by the spins of neighboring atoms. An NMR quantum computer program therefore consists of a sequence of microwave pulses at specified frequencies and in specified directions each followed by a delay of a specified duration to allow coupling between qubits.

Quantum Algorithms: Researchers have developed four algorithms for quantum computers. Williams and Clearwater[8] explain these algorithms in detail. Feynman[9] predicted that physics could be simulated on a quantum computer more readily than on a conventional computer. Recent developments in quantum harmonic oscillators[10] show how practical Feynman simulators might develop. The Shor Algorithm[11] shows in principle how to factor large numbers quickly (but no quantum computer has yet factored 15). Grover's Algorithm[12] shows how to search unstructured databases, with modest success in searching a four-bit database. The Deutsch-Jozsa Algorithm[13] shows how to measure a global property of a function (such as whether a predicate of a four-bit number is constant or true on half of all possible inputs) by executing it on all possible inputs simultaneously.

Entanglement: The notion of entangled qubits is currently a hot topic[14] and is likely to lead to improvements in communications technology. Some sophisticated experiments[15] have shown that two particles with correlated quantum states can maintain their correlation over great separation distances. Entanglement happens whenever a system can exist in a superposition of just some of the possible states. For example, three qubits are entangled if they could never be observed in the same state; that is, one must differ from the other two. If the state of a qubit is not determined until it is measured, how can one qubit know that the other two have the same state? Although entanglement is not required in all quantum algorithms, it may be very important in building large quantum computers. Eventually entanglement may be used to communicate quantum states between widely separated parts of a networked quantum computer.

Bit-Parallel Algorithms: Some algorithms for quantum computers, including the Deutsch-Josza Algorithm, work like bit-parallel algorithms. For example, a 5-bit input has 32 possible values, so we assign a bit position in a 32-bit word to each of those

values. Every possible predicate of that input corresponds to some 32-bit signature. To negate a predicate, complement its signature. To require all of several predicates, take the logical intersection (and) of their signatures. To require any of several predicates, use logical union (or), These operations can be carried out in parallel for all possible 5-bit inputs. To evaluate a predicate for any specific 5-bit input, just look at the corresponding bit of the signature of the predicate.

Complexity: The complexity of a problem is the scale of its difficulty measured as the growth rate of the logistical resources required to solve it as a function of the size parameters of the problem. We do not need to be overly specific here, and we especially do not need to define terms like NP-complete. A simple scale is enough (see Fig. 2):

- Easy -- a solution costs pennies, you do it yourself;
- Nontrivial -- a solution costs dollars, you buy it;
- Hard -- a solution requires research, someone learns something;
- Intractable -- costs double for a fixed increase in size, versus a tractable problem for which costs double for some percentage increase in size.
- Noncomputable -- there is proof that no general solution is possible.

Turing Tarpit: Theorems about limits on what can be done have a chilling effect on research. Teach a bright student about Turing noncomputability (proofs that some functions are inherently not computable on conventional computers) and that student will later recognize certain problems as being noncomputable and not even try, although partial solutions could be extremely valuable. He appears to be mired in the Turing Tarpit[16] and the deeper one's understanding, the harder it is to ignore limits. For example, we know that there cannot be a proof procedure that determines whether an arbitrary program ever terminates, but we can design proof procedures that work on a class of programs large enough to include all acceptable programs by definition. Reliable programs tend to be simple. For another example, a recent paper[17] claims that NMR quantum computers as currently constructed cannot demonstrate entanglement. The paper does not refute the assumption that each molecule in solution in an NMR computer attains all its allowed states simultaneously, but shows that the approximations used in small NMR quantum computers do not demonstrate entanglement. We cannot expect to escape the Turing Tarpit by shallow thinking. We need to understand all the assumptions that determine various limits.

Tractability: Quantum computers and conventional computers can theoretically simulate each other. Therefore what is not computable for one is not computable for the other. Quantum computers have an exponential advantage however; so what will always be intractable for a conventional computer may become tractable for a quantum computer. A tractable problem is theoretically practical. We turn hard problems into nontrivial problems through research, and nontrivial problems into easy problems through education.

Grand Challenges: The Office of Naval Research has posted four Grand Challenges[18], problem areas that the Navy currently sees as very significant:
- Battle Space Awareness
- Naval Materials by Design
- Electric Power Sources
- Intelligent Naval Sensors

Quantum computers and related technology may someday contribute substantially to these challenges. They are likely to contribute to meeting the first three challenges through improved simulators. Intelligent naval sensors will benefit most when quantum computers help artificial intelligence succeed. Artificial intelligence may be the ultimate beneficiary of quantum computing because many of its failures have resulted from the intractability of the problems it faced.

Moore's Law: Many charts show the dramatic exponential growth of computer technology throughout its history. Gordon Moore predicted that this growth in 1963 when he had only three data points. He has since said that his rule was not a law, but a self-fulfilling prophecy[19]. The silicon industry adopted Moore's Law as a guideline in establishing an industry roadmap[20]. Manufacturers who were behind the curve had to allocate more resources to stay competitive, but those who were ahead could relax a little. Many progress charts have been prepared and are available on the web. One of the best charts[21] shows the evolution of computer power over cost compared to evolution of human brainpower.

Limits: Moore's Law cannot continue to hold for conventional computers. The speed of light, the Heisenberg Uncertainty Principle, and the Rayleigh Resolution Criterion limit conventional computers. Quantum computers hold forth the possibility of side stepping these limits by performing computations in many parallel universes simultaneously. The various limits do not constrain the computation until a measurement is attempted. We do not know what other limits will be discovered on quantum computation.

Imagery: How could a quantum computer use its enormous state space? That depends on the kinds of data structures that are developed for quantum computers. For example, a $2^A \times 2^B$-pixel image could

be mapped through Fock-state preparation[22] onto the superposed states of A+B qubits. Specifically, one HDTV screen image (1024x1024 pixels) could be mapped onto 20 qubits, and a four-hour movie (2^{14} seconds) at 64 frames per second could be mapped onto 40 qubits. That is not to say that we could get the images back again because a quantum computer with N qubits will only be able to answer N yes/no questions. That problem is partially addressed by using a great number (say 10^{18}) of very small (molecular) quantum computers running the same program. Even so, the state space grows much faster with additional qubits than the possibility of deploying enough quantum computers to render the quantum state on a conventional computer. For example, just 700 qubits would be enough to map a Euclidean universe the size and duration of our own down to the Planck scale (10^{-35} cm). There would be a substantial input/output problem.

Processing: The problem is not how to extract the quantum state from the qubits, but how to process mapped images so much simpler questions can be answered. The assumption necessary for exploiting quantum computers is that each quantum computer assumes all of its allowed states in each run. The initial and final states may be small, but the computation may proceed through an extremely large intermediate state space that is not measured. For example, could a quantum computer locate an image of a face or a weapon in a collection of a thousand one-hour movies? The answer needs 10 bits to say which movie plus 18 bits to say which frame, not the 2^{20} bits needed to render that frame. For another example, a quantum computer should be able to simulate certain physical systems, such as the weather or the propagation of underwater sound. I say 'should' for sound (and radar) because the various wave equations are time symmetric up to the inclusion of attenuation, and I have run such simulations backwards[23]. The fundamental operations of a quantum computer are unitary (time symmetric) transformations, except for making observations, starting up, and shutting down. These exceptions prove that the full operation of a quantum computer need not be time symmetric. Time symmetry is just a means of improving performance.

Education: Quantum computers will help solve many interesting and worthwhile problems only when enough researchers have the tools and expertise to tackle them. Not only will these researchers have to master their problem domains, they will have to understand and rework the assumptions in those domains. Nahin did so for time travel[24] by his scholarly and comprehensive analysis of our assumptions about time. The purpose of many assumptions is to make certain solutions tractable. Any technology that changes what is tractable will require re-examination of the underlying assumptions of any field that might use that technology. We must not only train future researchers to use basic techniques, but to invent them. We must also assume that productive programming environments for general-purpose quantum computers can be developed. Having quantum computers that are accessible most of the time will greatly contribute to the education of many experts in programming them.

Direction: We need a reference point for future analysis, a design that is well ahead of the state of the art, so that we can make future estimates about the development of quantum computers. To freeze the reference point, we will use a very old design[25], one that will not change because it has not changed. The Rabi Quantum Computer is named after Isador Isaac Rabi, an Austrian-born American physicist who discovered that resonant microwave radiation could affect the spins of subatomic particles. It is specified to be 300 qubits long, 50 qubits wide, and 30 qubits high with a 1 qubit high grid on top for an interface to a windows system (see Fig. 3). This design could help us survive an information flood that makes our current one look like an April shower. It could surely take on the complete genome for two of every kind of animal in the world, because it is the RQC (pronounced "ark, you see" in English). However, the imminent use of this design does not depend on actually building it or on faith in its Designer, but on its uncontested age, so that an estimate of when it could be constructed will be commensurate with future estimates.

Schedule: When could an RQC be built? Current estimates are necessarily very vague. We need to progress from the current state of the art of short linear 8-qubit chains to large chains and grids. Some sixteen doublings are required to build the RQC as specified with $450000 \sim 8*2^{16}$ qubits. I expect the following stages will take place:

- 1-2 years: QC concepts proven;
- 2-5 years: Some QC is up all the time;
- 5-10 years: Remote QC access for study;
- 10-25 years: Practical QC grids available;
- 25-50 years: Cheap QC's in use everywhere;
- 50-100 years: An RQC can be built.

Shortcuts: Advances in technology sometimes take surprising leaps when supporting technology is available. Perhaps someone will figure out how to exploit the magnetic fields in old core memories to control qubit grids. Perhaps someone will couple CCD grids (as in camcorders) to qubit grids, so that a complicated quantum computer program can be prepared as a video clip. An RQC could conceivably be built in twenty years. At least by then we will have 2020 hindsight.

Connections: Although a linear chain of qubits is enough because the quantum states of two adjacent qubits can be exchanged, more complicated networks are probably desirable. There are indications even now that we will be able to create three-dimensional qubit grids using DNA structures[26]. How those qubits are interconnected is not specified. Connecting each qubit to its six nearest neighbors is surely overkill. Connecting 15000 chains of 50 qubits to 300 chains of 30 qubits in the I/O grid, each connected to one chain of 50 qubits may be awkward because moving quantum states around such a network may lose many of the benefits obtained from quantum computing.

Perfect Shuffle: A connection topology that has minimal direct connections for easy implementation and maximal indirect connections for rapid movement of quantum states will be desirable. One perfect shuffle network (see Fig. 4) connects each cell directly with just three other cells, but indirectly with some 2^k cells in k steps. For an example that does not quite meet the RQC specification, consider connecting M=131071 cells in loops of P=17 cells each with the remaining cell linked to itself. Number the cells in one such collection so that cell n is connected to cell 2n (modulo M). Number the cells in another such collection so that cell n is connected to cell 2n+1 (modulo M). Entangling qubits in correspondingly numbered cells from each collection creates a perfect shuffle network. All M cell pairs are connected in a single long chain by uniform sequences of steps (back-shuffle-forward-shuffle). Other uniform sequences of steps connect distant parts of that long chain. To get halfway around the chain (π=65576), use a different step sequence (forward-shuffle-back-shuffle). Every cell is connected to just three cells, but has a tree of all other cells both below it and above it. This would be great for artificial intelligence applications, which are frequently recursive.

Stimulation: An ulterior purpose of the design of the RQC is to stimulate imaginations. The more researchers dream about the possibilities in the exponential potential of quantum computers, the sooner quantum computers can be profitably applied to real problems. What might the future bring?

My Dreams: I hope that this paper will help liberate the imaginations of many students who will help other students learn to make quantum computers useful in many fields. I also want to persuade the National Oceanic and Atmospheric Administration (NOAA) to focus its efforts in Quantum Computing by adopting a long-range goal of building the RQC for use in weather forecasting. ☺

References:

[1] Harold A. Linstone, *Murray Turoff, The Delphi Method,* Addison-Wesley Publishing Co. (Reading MA) ©1975; ISBN 0-201-04294-0, ISBN 0-201-04293-2 pbk.

[2] Ibid., p.374.

[3] Peter Schwartz, *The Art of the Long View: Planning for the Future in an Uncertain World,* Doubleday (New York) ©1991, 1996.

[4] David Deutsch, The Fabric of Reality, The Penguin Press (New York) ©1997.

[5] Etienne Klein, *Conversations with the Sphinx: Paradoxes in Physics,* Souvenir Press (London), translated 1996; ISBN 0 285 63305 8 (hardback).

[6] D. Halliday, R. Resnik, K. Krane, *Physics, Volume Two, Extended Version,* Fourth Edition, John Wiley & Sons (New York) ©1992; p.1064.

[7] Gerald P. Milburn, *The Feynman Processor,* Perseus Books (Reading MA) ©1998; p.20 citing Feynman.

[8] Colin P. Williams, Scott H. Clearwater, *Explorations in Quantum Computing,* Springer TELOS ((Santa Clara CA) ©1998; ISBN 0-387-94768-X.

[9] Richard P. Feynman, "Simulating Physics with Computers", *International Journal of Theoretical Physics,* vol. 21, nos. 6/7, 1982, pp. 467-488.

[10] C.R. Nave, "Quantum Harmonic Oscillator", Georgi State University; http://230nsc1.phy-astr.gsu.edu/hbase/quantum/hosc.html

[11] P W. Shor, "Algorithms for quantum computation: Discrete logarithms and factoring," in *Proceedings of the 35th Annual IEEE Symposium on Foundations of Computer Science* (IEEE Computer Society Press, Los Alamitos CA, a994), pp. 124-134.

[12] L. K. Grover, "Quantum Mechanics Helps in Searching for a Needle in a Haystack", *Phys. Rev. Lett.* **79**, 325 (1997).

[13] D. Collins, K. W. Kim, W. C. Holton, H. Sierzputowska-Gracz, and E. O. Stejskal, "NMR Quantum Computation with Indirectly Coupled Gates", *e-print* quant-ph/9910006.

[14] Michael Brooks (ed.), *Quantum Computing and Communications,* Springer-Verlag (London, …) ©1999; ISBN 1-85233-091-0.

[15] Ibid., p.123.

[16] A.J. Perlis, *personal communications,* circa 1970; there is no consensus as to what Perlis meant by the Turing tarpit.

[17] R. Fitzgerald, "What really gives a quantum computer its power?" , *Physics Today,* 2000 Jan, pp.20-22.

_____

[18] Office of Naval Research, "S&T Grand Challenges", http://www.onr.navy.mil /sci_tech/chief/GrandChal.htm

[19] Gordon Moore, (speech), 50th Anniversary of the ACM, 1997.

[20] (a global community of researchers, manufacturers, and suppliers), *International Technology Roadmap for Semiconductors: 1999 Edition,* Semiconductor Industry Association (San Jose CA) 1999; http://notes.sematech.org/ntrs/PublNTRS.nsf/

[21] Jeff Sutherland, "Evolution of Computer Power/Cost", http://www.jeffsutherland.org/objwld98/futur e.html

[22] G. Harel and G. Kurizki, "Fock-State Preparation from Thermal Cavity Fields by Measurements on Resonant Atoms", *Phys. Rev. A.* 54, 5410 (1996).

[23] R. A. Krutar, S. K. Numrich, et al., "Computation of Acoustic Field Behavior Using a Lattice Gas Model," *Proc. Oceans 91 Conference* (Honolulu, 1991), Vol. 1, pp. 446-452, IEEE.

[24] Paul J. Nahin, *Time Machines: Time Travel in Physics, Metaphysics, and Science Fiction,* Springer-Verlag (New York) ©1999, 1993; ISBN 0-387-98571-9.

[25] *Genesis* 6:14-16.

[26] C. Wu, "DNA Strands Connect the Quantum Dots", *Science News,* vol. 156, no, 12, 1888 Sep 16.

Figure 1: The Two-Slit Experiment, which shows interference patterns, standing waves, and Fresnel diffraction (behind the screen).
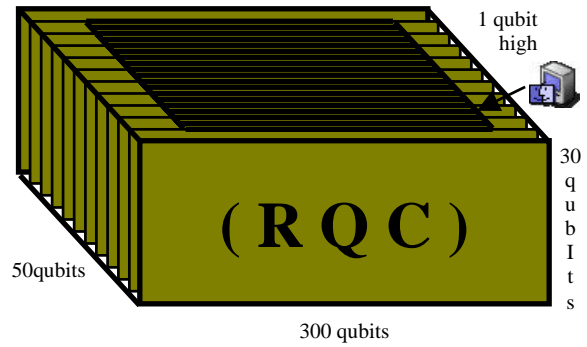
Figure 3: The Rabi Quantum Computer, which serves as a landmark in the dim future because it is well beyond the state of the are and its dimensions are frozen.
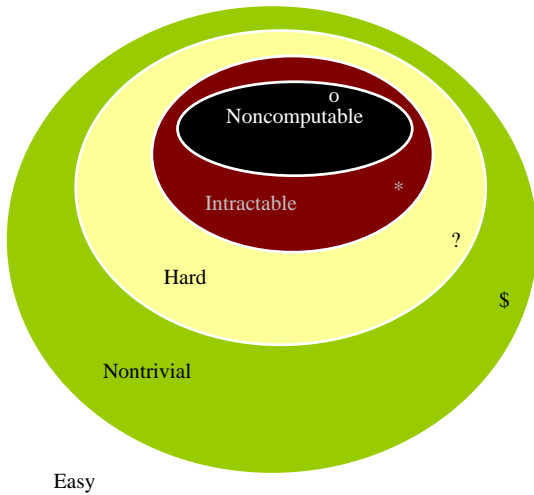
Figure 2: Levels of Difficulty, showing a simplified class structure for computational problems.
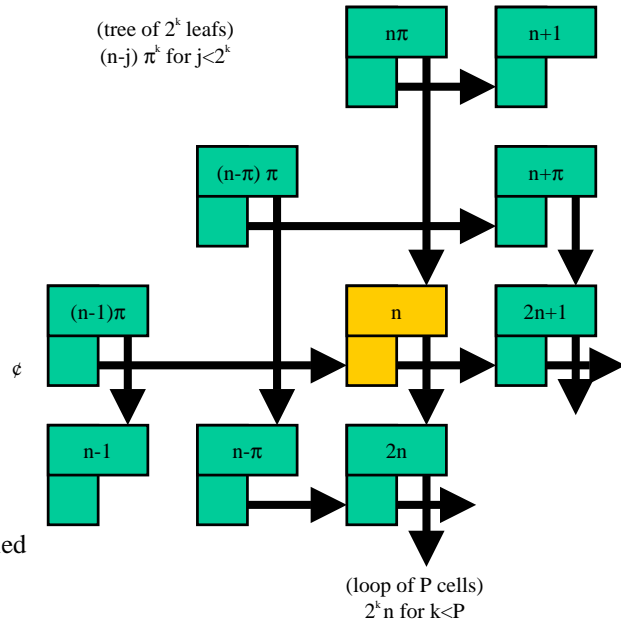
Figure 4: A Perfect Shuffle Grid, which serves as an example of a minimalist connection grid with maximal reach for any prime P, arithmetic modulo $M = 2^P - 1$, $\pi = 1/2 = (M+1)/2$, every two-tile cell participating in a vertical loop and a horizontal loop.

This page has been deliberately left blank

Page intentionnellement blanche

Challenges for Joint Battlespace Digitization (JBD)

S. Hamid, I. White and C. Gibson
Defence Evaluation and Research Agency (DERA)
Portsdown, Fareham, PO17 5EU, ENGLAND
shamid/iwhite/cgibson@dera.gov.uk

Abstract

This paper highlights several important areas in achieving extensive integration of military command and control systems. The discussion in the paper focuses on two areas where technology per se is deficient, and that must be considered carefully against the stronger tides of technology push in systems design and acquisition. These areas are the human factors aspects of system design, and the challenge of information management

1. INTRODUCTION

1.1 Background

Several of the NATO countries have proposed more integrated defence IT infrastructures to help achieve information superiority. It is, of course, a truism that information superiority will give a party an advantage in any conflict, and historic examples abound, perhaps the most famous from a British perspective being the use of the Enigma decodes during WW2 [1].

It is the nature of defence acquisition that the information systems inventory will comprise a range of systems of different capabilities and technology vintages. At the present time, the UK inventory includes old bespoke military systems coupled with modern IT systems having a high commercial off-the-shelf (COTS) element. Information superiority requires, amongst other things, that this mix of systems be integrated into an interoperable, 'seamless' whole.

This vision of seamless Defence IT has spawned several national initiatives that seek to help achieve it. In the UK this is embodied in the Joint Battlespace Digitization (JBD) initiative that seeks to provide a wide range of interoperability benefits that will take UK defence well towards the information superiority goal.

The goal of information superiority requires a whole range of technical and non-technical changes in all aspects of defence capability - doctrine, command processes, organisations, user-requirements specification, architecture definition, technology development and exploitation, procurement, training and operational use.

In this paper some of the challenges are examined and the prospects for progress reviewed. The emphasis is on the definition of systems and their implementation in modern technologies. This encompasses the definition of a systems concept, an integration architecture, and the ability to realise many of its elements using COTS services, systems and elements. By integration architecture is meant a collection of rules, recommendations, process definitions and standards that apply across all systems.

The paper highlights in particular the issues of human factors (HF) and information management (IM). These have been chosen here because there are still many unknown and therefore significant risks in these areas.

1.2 Major Technical Issues for JBD

At the technical level many major problems have been identified that need to be resolved through defence-sponsored research. Others are better left to the commercial world to resolve. The decision about which issue is in which category is itself a critical technical risk! A shortlist of the issues being addressed within the UK defence research programme are grouped under the following headings:

- Future command concepts and business processes
- Human Factors
- Information management
- Information technology and systems
- Systems Integration

In this short paper it is impossible to describe all the solutions being pursued within the UK to meet these issues. This paper therefore gives more attention to human factors and to information management than the other topics. There is a degree of euphoria surrounding new capabilities in information and communications technologies (ICT) and the related, the exploitation of COTS developments. It can be easily forgotten that command information Systems (CIS) are to meet human needs, and that the system design aspects that focus on this are still very poorly understood, and inadequately applied. First the key issues concerned with user needs are outlined.

2.   USER NEEDS

2.1   Comment

It is a half-truism that the in-theatre user lacks the vision to see the technological possibilities of the future, and the visionary lacks the experience of real users to provide what is really needed. This belies the real situation whereby with careful liaison between these groups a reasonable prediction of needs can be attempted. For success, users' aspirations and technological possibilities need to be aligned, but the chain of dependencies is not well understood. The issues are understanding:

- distributed operations
- the needs for 'synchronysity' (making the right actions occur at the right place and the right time over a very distributed network of decision makers and organisations)
- satisfying manourvist warfare principles.

Current design techniques are based on workshop methods (story boarding and ad hoc diagrammatic representations, that often retain a strong degree of ambiguity). What technologies exist to help us here? There are a range of soft-system methodologies prescribed for the better definition and elucidation of such issues [2], [3], [4]. DERA has invested substantially in high level information flow models to represent the information exchange behaviours of emergent new social structures of command that are consequences of the changes in both technology and in warfare

concepts. This work however would need another paper to describe it.

3.   HUMAN FACTORS

3.1   Cognitive Dominance

Central to the achievement of information superiority and military objectives is the very human concept of *Cognitive Dominance* [5] – going beyond information dominance to have the ability to exploit this superiority. This concept has the following six aspects:
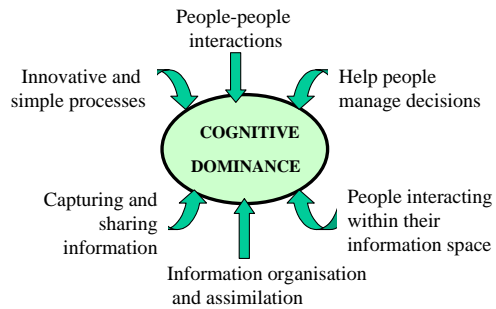
Figure 1 – The six aspects of the concept of Cognitive Dominance

ICT (i.e. digitization) is a key enabler of cognitive dominance, but in realising it a balanced organisational change process is needed. As Figure 2 shows, technology is but one of four forces of change that the Joint Battlespace Digitization (JBD) initiative aims to bring into dynamic stability [6].
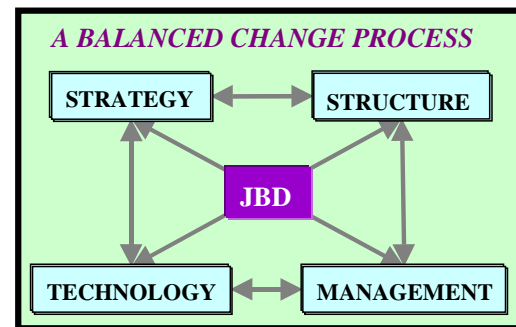
Figure 2 – The four forces the JBD change process needs to balance

Important human factors issues permeate all four forces. These are discussed below according to the schema in Figure 2; Figure 3 indicates the scope of each of theses forces, as discussed below and shows how they inter-relate.
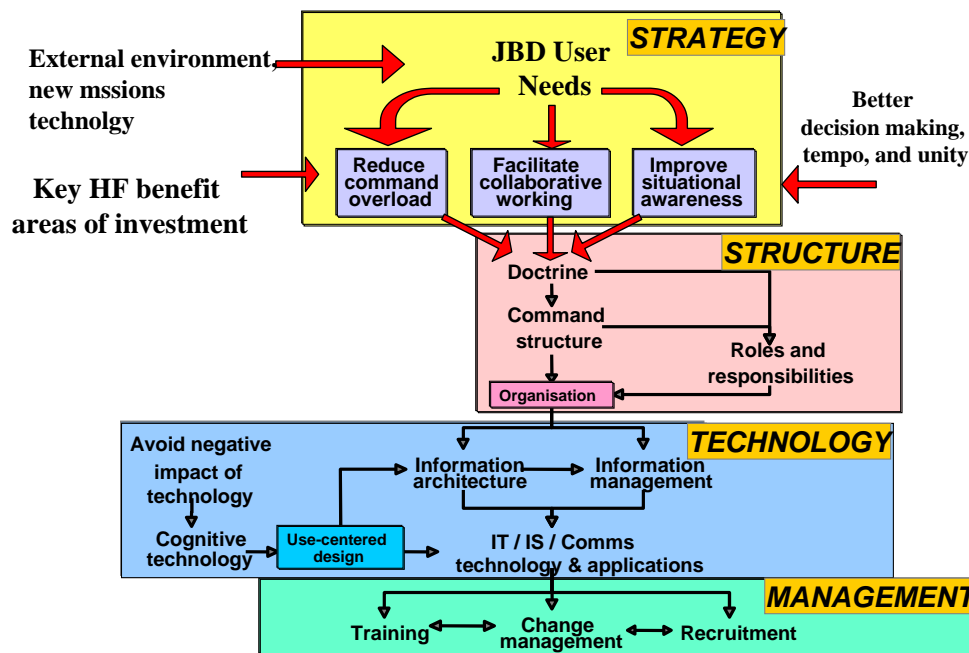
Figure 3 – JBD - A view of the inter-relationships between the four forces

3.2 Strategy

In essence the aim of JBD is to enhance perception and situational awareness (i.e. lift 'fog') and increase action options (reduce 'friction'), thus allowing command to cycle more quickly (i.e. better decisions and tempo). This leads to the desired force agility and flexibility. If however, as is highly likely, 'fog' and 'friction' are still key problems of future military command, we should expect the emphasis to shift to the qualities of the commander, in particular to his cognitive abilities. Such commanders must be able to synthesise information from a great variety and number of sources (e.g. political and military), particularly in complex diplomatic-military operations in a multi-national context, to achieve a true sense of what is going on.

Related to the above JBD aim from a human factors perspective, the three key benefit areas of investment in JBD are that it should [7]:

- reduce command overload,
- facilitate collaborative working, *and*
- improve situational awareness.

These are part of the JBD User Needs at the level of strategy (see Figure 3). In combination these three benefits contribute to the achievement of all six aspects of cognitive dominance, that should lead to better decision making and operational tempo, and unity of effort. Some of the command level problems induced by technology are outline later.

Reduce command overload: This is best achieved by reducing the amount of cognitive processing required and the number of activities an individual has to attend to. Measures include presenting information in an integrated fashion, simplifying processes (e.g. reducing the number of manual steps), and having standard operating procedures and doctrine for dealing with a situation or issue. Often we cannot tell in advance what information may prove to be highly pertinent, so that staff and commanders must operate in surveillance mode, scanning and sifting for useful items. To some extent this dilemma can be addressed through collaboration among commanders and staff or through organisational procedures and systems. Intelligent systems that are able to highlight changes and key factors will also help here.

Facilitate collaborative working: Joint and Multinational operations demand effective teamwork towards shared military objectives. The key issue here is providing support to those activities related to communication, co-ordination and the development of shared understanding within a team, including both distributed and ad hoc teams. Technological support for such teams is currently immature. Recent improvements of the understanding of team processes and teamwork challenges encountered by military personnel operating in distributed and ad hoc teams can be used to highlight technological design imperatives for computer-supported collaborative working.

Forms of data fusion support are required that are more in tune with the socio-cognitive processes by which humans collectively integrate information, and the fundamentals of military practice.

Improve situational awareness: This has three components of equal importance: the ability to understand how a situation has developed; the ability to comprehend the current situation; and the ability to predict how the situation might develop. What is important is having the right information and cues at the right time. From an information presentation perspective, use of tactical picture agents for monitoring and alerting as part of the strategy for the management of operator attention can help.

3.3 Structure

The new missions in the changing operational environment, together with digitization technology make ad hoc 'teams' more attractive as a way to get jobs done – the challenge is to facilitate effective teamwork in such ad hoc situations. Flatter command structures to maximise agility and force flexibility enabled by the information age and are seen as inevitable. Driven by these pressures, it is not surprising that the armed forces should consider more loosely-based federations of functions to perform a mission in a self-synchronous way; noting that organisational and team loyalty are still important to the bonding process and team resilience. Future CIS structures may therefore be loose aggregations of autonomous units rather than rigid hierarchies, since the CIS organisation requirement depends on the ratio between CIS speed and battle speed. As noted in [8] "the more fluid the circumstances, the lower the decision level should be set"

Digitization allows integration of organisational functions at all levels, both within and between organisations. In order to achieve this, management must effect changes in organisational strategy, structure, processes and culture (possibly only at the artefact and attitude level rather than at a deeper social level). Information and communications technologies (ICTs) have implications for the boundaries of the organisation, and the ability of managers to control the flow of information. If information is power, people are likely to be reluctant to give it away. Management must

therefore create the environment in which people will share information. The speed and extent to which organisations may move in this direction is constrained by their current structures, process, skills and expertise, and their investment in existing (legacy) information systems. However, the greatest inhibitor may be the legacy organisational culture and practices.

Organisational structures are affected by the changes in the way work is done as a result of ICT developments – for example, new specialists, fewer unskilled and semi-skilled employees. There is a need to define and delineate the new roles and responsibilities within a digitized organisation, taking into account recruitment and retention problems. Clearly, organisational re-structuring should be aligned with doctrine, and tools, techniques and procedures. A use-centred philosophy should be employed where the emphasis is on simplifying the processes, and using technology to support skilled staff to enhance their productivity [9].

3.4 Technology Implications from HF

JBD architecture from an HF perspective: The JBD information architecture encompasses a cognitive dimension as an overarching top layer. This is because at the cognitive level individual and group perceptions provide understanding and they influence decisions and people. Furthermore, meaning is often extracted by the humans from information outside of, and ahead of, the supporting information system. Information management is therefore not merely a question of computer systems and information and communication technologies. The following four inherent limitations and shortcomings of computer-based information systems further illustrate this point [10]:

- There is a large amount of key information and knowledge that is not captured by or represented in these computer-based information systems.
- Important but unpredictable or anomalous organisational processes are unlikely to be supported by a computer-based information system.

- Current ICT systems only present to the decision-maker that information that has been identified as being of value by the designers and procurers of the system. They are unlikely to have thought of every eventuality and circumstance.
- Information gathered by current ICT systems is historical - they have little predictive capability.

The focus of information management should be on management processes and organisational implications. The emphasis of technology considerations should therefore be on the use to which technology is put in organisations. A use-centred information management considers the ways that managers and staff *actually* use information to drive the development of computer-based information systems, rather than imposing some idealised or normative technical solution on the people in the organisation [10].

Communications systems design must remember that communications between people is naturally a human process, where:

- The message itself is not necessarily unproblematic – the sender may be struggling to express it precisely.
- The context is very important as other messages may be competing for attention.
- The expectations or perceptions of the recipient can generate particular interpretations.
- Likewise, the medium used to convey a message can affect perceptions.
- Non-verbal elements are very important in human communications.

Organisations cannot function without managing their information processes effectively. Within the organisation, knowledge is within the heads of the workforce, is embedded in tools and machines, as well as being captured and represented in organisational processes and heuristics.

The ability to learn is crucial, so that information processes have to be dynamic. Information processing requirements should be derived from CIS requirements on critical information needs, team interaction requirements, display and software design imperatives, physical configuration of command post, decision support tools etc.

Automation of task and data management has a strong impact on human behaviour and task requirements. Automation often gives rise to new tasks (such as management of information) and responsibilities (such as data ownership, access and control of data). Adaptation behaviours also give rise to systems being used differently from that anticipated.

There is still a huge amount to understand about the relative overall costs of handling information in different ways, and the ease or difficulty of ensuring up-to-date and accurate information under different arrangements. These questions are inextricably bound up with the crucial people issues of motivation, job design and staff development.

Command problems exacerbated by technology: There are three potential command problems that may be made worse by inadequate future CIS [11]:

Cognitive overload: Future digitized combat threatens to stretch commanders' cognitive resources much more than before, while still being just as stressful physically and emotionally. Cognition capacity can be enhanced through appropriate method of presentation of information and by the experience and training of the individual who is trying to encode the information. Future CIS interfaces should present information in an appropriately aggregated and individually configurable manner to commanders. A further risk is 'information pursuit' – the more readily assimilable information that is provided the more the commander may pursue additional information, often to the detriment of the operation.

Over-controlling command style: In collective training, due to increased availability of information through digitization, unit commanders have repeatedly been observed using an over-controlling command style. As a consequence junior commanders' initiative is stifled and the decision cycle is slowed due to unnecessary upward referral of low-level decisions. If from the senior commander's perspective 'perfect control' is possible, it may lead to micro-management that overall will have a negative effect on performance.

Big-picture blindness: Not only does this mean that a commander focused on detail will be more likely to miss the big picture, it may also mean that his method of coming to a decision is altered by digitization. Although under normal circumstances the brain manages

to integrate the parallel functions of its two hemispheres (on detailed and global processing) to produce coherent behaviour, systems could be designed and used to reduce the risk of big-picture blindness.

Technology pitfalls summary: Many attempts to support the command process with technology fail because the design of the latter is founded upon an inappropriate decision-making paradigm [12]. The design of decision support systems should be based upon an understanding of how <u>people</u> actually make decisions in the real uncertain world, and also upon what they need to do to make these decisions. There is a need to ensure that systems provide information in a manner that is cognitively compatible with the user's mental model and decision strategies. Common problems with inappropriate technological support for command have been found to be:

- User frustration at feeling tethered to workstations.
- Digital mapping used in parallel with the paper systems they are supposed to replace.
- Automated position plotting on digital maps reducing user engagement and thus situational awareness.
- Teams having difficulty in brainstorming around a digital support system.
- Teams having difficulty in using digital maps for situation and mission briefings.
- Lack of system transparency, i.e. lack of understanding of where digital information comes from, how it has been integrated and how it has changed over time. This can thus reduce understanding about how the situation has evolved, and lead to surprise.
- Individuals spend their time 'driving' systems, including information pursuit, that significantly reduces the time available for thinking and talking about problems they are dealing with.

Systems that are difficult to use will not be used widely, particularly under combat stress. This leads to longer terms problems of complacency, skills decay and job dis-satisfaction.

3.5 Exploitation of cognitive science and technology

Recent advances in understanding of the mechanisms of cognition offer opportunities for developing more effective military processes and systems. A satisfactory model of how humans separate and integrate information can be used in the design and development of all systems in which the operator applies cognitive effort. Through an understanding of how sequences of events are separated and integrated by the brain into a 'cognitive stream', guidelines for the presentation of multi-modal sources of information can be formulated.

Cognitive design for human-computer interface (HCI) is concerned with the presentation of information in a manner that converts potential cognitive tasks into perceptual tasks. This is a critical issue underling the design of HCI for use in situations of high workload. Visual and perceptual tasks can require little or no apparent cognitive effort. Thus for an HCI that is used frequently, in a time pressured situation, a simple saving in one element of a display may reduce significantly the user's overall cognitive burden. This leaves more effort available for dealing with (i.e. thinking about) uncertainty and risk in decision making.

Beyond information presentation is the possibility of using real-time adaptive automation and adaptive decision aiding [13]. These use *inter alia* cognitive control theory to consider the effects of time pressure and uncertainty on decision making. The aim is to enable the decision-maker to concentrate cognitive capabilities on the important aspects (i.e. dealing with uncertainty) whilst off-loading routine activities to automation (i.e. alleviating time pressure). This allows the decision-maker to remain in a feed-forward loop whilst feedback can be automated using decision aiding. The main features of such a tasking interface are:

- a shared mental model between man and machine;
- the ability to track goals, plans and tasks; and
- the ability to communicate intent.

Use-centred design as the way ahead:

Two fundamental observations can been made concerning the development of technological support systems to-date [8]:

- Much of the technology that has been, and is being developed, for supporting joint and strategic command comes from a technology 'push'. Consequently, the organisation often ends up being 'fitted' around the technology rather than the organisation establishing and defining its technology requirements.
- The design process for the production of technological support systems rarely takes proper account of the critical human factors identified in this paper.

The solution to these two problems is easily summarised, but more difficult to provide in prescriptive form:

- Develop the support technology from a holistic perspective. In other words, understand the nature, values and goals of the whole organisation; the key tasks that it undertakes; the relationship between technological and human function allocation. This can be used as a basis for driving technological development.
- Adopt design approaches that are 'use-centred'. If appropriate, use specialised methods to understand the nature of all the tasks that the humans undertake (including cognitive tasks) and provide this understanding as an input into the more traditional constructs managed by other systems analysis methods. For success, it is imperative to exploit the synergies arising from the convergence of information science, business science and cognitive psychology in the context of JBD.

Empirical studies are essential to validate conclusions on technology design and use, e.g. on organisational and physical design of HQs. Such designs should be based on meeting the requirements of the command function, e.g. physical layout, allocation of function including adaptive automation, distributed and ad hoc team working etc.

3.6 Management

In parallel with the developments in strategy, structure and technology, there is a need to develop cultural acceptance and user confidence. This is concerned with the way individuals, teams and organisations adapt to system changes. Some effects will be cognitive, such as skills fade with the need for amelioration (e.g. through training), whilst other effects will be emotional, such as change in the levels of trust and system dependency.

There will be a requirement for changes in skills and knowledge across the entire armed forces. This will be in part due to the use of new technology, and in part due to the need to conduct command decision making in new ways (including liaison with political, non-government, and other nations' groups). In addition to these higher level changes, indirect skill requirement may take the form of critical thinking and meta-cognition skills, reversionary mode skills, and alternative staff management skills. One of the most commonly cited reasons underlying failures of technology when introduced into military organisations is inadequate training, which results in delayed or reduced uptake and exploitation of technology.

There seems to be an assumption that we will need to recruit and retain more IT specialists as systems managers of the new digitized command, control, communications, and information systems. However, there are doubts that sufficient numbers of people will be available, and, if available, whether they will be easy to retain. Fortunately, there is an alternative approach: as a matter of design principle, make the systems so simple to use and maintain, or support them with intelligent aiding systems, that the need for true IT specialists is correspondingly small. This has two advantages: first, the requirement for specialists would be more likely to be tailored to the very small potential supply; second, the systems would be more usable by individuals and groups who are short of sleep, hungry, thirsty, frightened, and angry – in short, who are experiencing combat stress.

Finally, a major challenge for management will be to lead their organisations through the transformation necessary to prosper in the globally competitive environment. When the issue at hand is organisational transformation, enabled by technology, it appears particularly important to invest sufficient time and effort in getting the organisation to understand where it is going and why. The people issues are critical in the transformation process. One root cause for the reported lack of impact of ICT on the improved performance of commercial organisations is an organisation's unwillingness to invest heavily and early enough in human resources. The armed forces

must therefore learn this lesson and invest in new skills, in psychological ownership of the change process, and in a safety net under the employee so that there is no fear of taking prudent risks. These investments are required throughout the defence organisation as management itself is part of the required change.

4. INFORMATION MANAGEMENT

4.1 IM Definition

The need to meet the HF aspects in new systems design and operation has its counterpart in the definition of, and management of, information. With the creation of distributed operational staffs, powerful communications and powerful IT have highlighted our poor ability to manage the vast amounts of information to which we have access,. Much of this information is stored in a variety of formats, and with little consistency in terms of structure of storage, time stamping, formatting. presentation, precision and communications form.

The definition of IM is a source of argument, but the following covers the scope addressed by this paper:

Information Management is the control (creation, direction, filtering, presentation and deletion) of information between users in all defined domains at all communications, middleware and semantic levels (as defined below).

To be fully effective it is easily shown that management at each of these three levels is needed, and that there must be strong management interactions between these levels.

4.2 The IM Problem

The IM problem is simply that the wide variety of national CIS systems (land, sea, air, joint and coalition), in a JBD environment must inter-operate to varying degrees. Not only are these CIS constructed to represent data in ways matched to the specific 'sub-culture' of their users, the technical solutions are also widely different. Database schemas and their update and configuration control regimes and HCI vary. Furthermore the systems cover a range of procurement periods, ranging from full legacy to modern systems using a wider range of concepts and technologies to provide information to CIS users. IM is essential to manage and distribute such diversity of

information to those who need it in an integrated, timely and expedient way, taking full account of the HCI issues outlined in section 3.

4.3 An IM Model

An IM model for examining information management is:
- a society of agents, who have some degree of common culture (agents may be men or machines)
- a repository of information (static and dynamic, created, deleted, growing and decaying)
- a set of goals (also static and dynamic).

The achievement of these goals is facilitated by information agents having the right information at the right time in the right place, and in the right form. This is of course an ideal posing many difficult questions. IM needs to act at three levels:
1.   Semantic
       - Context correlations
       - Domains (whose members collectively invent sub-languages)
2.   Applications
3.   Communications/networking

This is shown in Figure 4. The diagram is explained from lower to the upper levels.
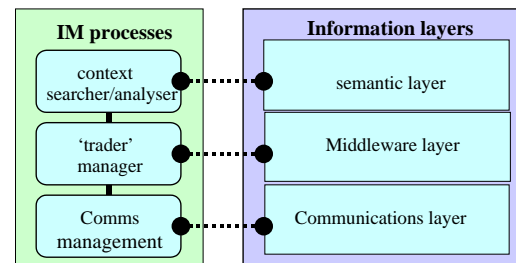
Figure 4   Layered model of IM

Communications layer management (CLM) can monitor who is communicating with whom, and at what rates, can note fault conditions, and can detect unusual traffic patterns, security violations etc., and control the parameters of the communications network. Communications management must interact with middleware management.

Middleware layer management (MLM) can monitor and control applications interactions, who is using what applications, with whom, what types of file are being interchanged. It can assign and administer

priorities, and control other middleware functions such as data caching. MLM can effectively implement load-balancing between information and communications levels, referred to here as 'trading'. MLM can also detect anomalous behaviours.

Semantic layer management (SLM) is the control and monitoring of all the information interactions (deriving from the meaning and intentionality) of the society of agents who seek to meet that society's goals. The SLM must interact with the MLM. This level of information management is concerned fundamentally with the range of human factors issues described in section 3.

Management at each of these different levels contributes to the management of information. Some are better understood than others. The integration of the layer interchanges, and other more sophisticated interactions have not been extensively studied.

4.4 The Problem with IM Models

The problem with this IM, and similar schemes, is that the development of IM into a fully inter-working schema is seen as a long-term task, requiring the solution of some hard problems. Some of the tools from the COTS world are of benefit in implementing IM, notably agent technologies, object schema such as CORBA and advanced mark-up languages such as XML. The work needed to develop and embed such schemes across a wide and diverse range of CIS is nonetheless a major undertaking.

In the following sections the three layers are described further, working form the lowest layer to the highest. This is in order of increasing difficulty

4.5 Communications Layer

Instance of Communications: It is fundamental that a society of agents can only be a society by the cohesive process of communication. The International Organisation for Standards' (ISO) open systems interconnection (OSI) 7-layer reference model, defines all communications around the concept of an *instance of communication*. Each *instance of communication* is described by a set of basic features [14]:

- Sender (identified by an address)
- Receiver (identified by an address)

- Communications set up
- Transfer of data
- Communications clear down
- Duplex/half duplex/simplex
- Duration
- Quality of Service (QoS).

The model also represents the communications process in a layered sense, with the higher layers being more abstract, the lower ones more physical. The ISO model of agent interaction (i.e. the instance of communication) is a fundamental element for a society of interacting agents.

Communications Management: In a given network the various data flows between users compete for the available communications resource[1]. When communications resources are not available, specific communications may be delayed or destroyed. Communications protocols have in-built protective mechanisms to control congestion, most of which involve the loss of information. The criteria for optimising the flows of information are based on various definitions of data exchanges, e.g.:

- Continuous bit rate [hard real time][2]
- Variable bit rate
- Available bit rate.

Protocols exist for establishing reserved paths through networks, but these are not usually an efficient means of using communications resource. The key routes to managing communications resources include:

- Circuit switching
- Packetisation
- Dynamic routing
- Store & forward
- Data shedding (throwing data away)
- Priority and Pre-emption (P&P) mechanisms.

Priority and Pre-emption: P&P requires the user to determine the priority of a message (i.e. instance of communication), and the network to then adjust the bandwidth and the queuing order according to rules based on these priorities. There are many incipient problems with such priority schemes including:

---

[1] Bandwidth, switching capacity, QoS provision, notably time critical delivery, and error characteristics.
[2] Definitions have been developed by ISO for Time Critical Applications (TCA's)

- What is the QoS of each prioritised communication?
- How large is each communication (how does this impact on its allowable priority)?

Commercial products and protocols exist for most of these processes, and all aspects are subject to continuing development.

The issues raised by implementing P&P in networks has important implications on the middleware layer. P&P is of particular interest for military networks because the concept is long standing, has been implemented in some military communications networks, and is still a requirement. This is discussed further in the Middleware Layer section, 4.6.

Network management status: Network Management (NM) concepts and technology are relatively well developed, and are generally effective, provided the overall mission needs are well understood and mapped into a sound network management strategy. Standards and related protocols are available from ISO, ITU, and IETF, the Telecommunications management forum (MTForum) and there are many useful products UK work is concentrating on developing a concept of operations for network management. The civil concept of NM is focused on service delivery, service growth, and revenue and market capture. Accordingly the development of public domain process models (e.g. within the TMF) is focused on these objectives. The military users must accept that these are not their objectives, and accordingly these need different process models. The implication is that the military must develop the required models themselves.

Interoperation with the MLM; Inter-operation requires that values of priority labels, and of QoS, are available from the middleware layer, and that status messages from the network are communicated to the middleware layer. More sophisticated interchanges may also be needed, as discussed further in below.

4.6 Middleware

Middleware Layer – scope: The term middleware is used here to indicate processes above the level of communications, and below the level of information comprehension and use (i.e. the semantic layer); see Figure 4. Many of the middleware operations and concepts assume information need is defined

somewhere, but does not itself need to 'know' what that information definition is. Thus:

- QoS must be defined
- Priority labels need to be defined in terms of some functional range (e.g. routine/urgent/flash)
- Data flows over the whole network can be minimised by using file caching strategies, related to information exchange needs. These strategies must be formulated.
- The models of information exchange are logically similar to the connection oriented communications definitions, that are based on the concept of *an instance of communication*.

All of these ideas are discussed here under the heading of 'middleware'. It is accepted that some readers may feel a particular middleware topic is really a communications, or semantic layer issue.

Quality of Service (QoS): The various parameters defining QoS must be related to the users' needs. These can be defined in semantic terms, then middleware terms and finally in communications QoS terms, in a way that is technically meaningful to communications systems designers and service providers. An issue that has received little attention is the extent to which users are prepared to negotiate QoS when the communications or middleware layers cannot provide the QoS requested.

Assigning Priority labels: Across a community of users each 'instance of communication' needs to have some *priority label* attached to it. The rules for attaching such labels will depend upon organisational factors such as:

- Seniority of the person using the priority label
- Significance of the role of the agent undertaking some task.

Note that these are cultural parameters of the society of agents (e.g. what context determines a *Flash* priority, and what service does the user expect *Flash* to imply?). Assigning priorities must itself be a process that is also assigned, i.e. individuals will be given the right to assign priorities. These rights may, from time to time, be changed, revoked, issued etc. and be further subject to restrictions based on any of subject, file-type, file-size, and classification.

The assignment of priorities does not necessarily result in a more equitable data flow

through a communications network. It is generally true that provided only a small proportion of the total data within a network is allowed high level P&P labels, an overall gain is made in communications efficiency.

It is fundamental that a good P&P scheme, in which priorities are assigned at the user middleware level, includes interaction with both the communications network, to establish a balance between what can be communicated, and the semantic level, to accommodate what the set of all users wish to communicate.

Data flow monitoring and caching: By observing data flow, not just at the level of bits or packets, but in terms of higher level entities such as files, or logical documents, substantial savings in data flow can be achieved by data caching. The classic example is web browsing, where local servers can hold frequently called pages and save significantly on the demands on the wide-area communications. With such schemes servers need only be sent updates of such pages.

Balancing Mechanisms at the Middleware/Communications Layer Boundary: These are needed when there is some mismatch between agents needs, normally achieved by some feedback mechanism. In the case of the middleware/communications layer boundary, data sources need information on network loading/congestion. For example P&P is in effect assigning a crude value-measure to messages, by setting a high value on the delivery of high priority messages. It is clear that for a JBD society, rich in information, a simple priority system is not sufficient, because it does not guarantee delivery QoS, nor do all message interchanges have similar QoS needs, e.g.:

- What defines satisfactory delivery? (milliseconds, seconds 10's minutes, 1 hour etc.)
- What is the influence of message size? (short ADAP-P3 message; digital image etc.).

These differing criteria must have effects on the overall P&P process and they need to be taken into account in any IM schema.

Middleware schemes for prioritisation can allow each independent server to provide arbitration between users' needs and communications capability, for all users

wishing to send information to other users. What these schemes do not usually have (in any depth) is interaction with the communications network (or its management system) nor with other servers on the network. Commercial bandwidth managers for IP (internet protocol) networks do include, in some cases, the ability for different bandwidth managers to communicate, and perform some degree of overall network load balancing.

Given that the various QoS schemes, prioritisation labelling schemes and the information exchange requirements (IERs) and network topology and capacity are defined, information management, at the levels of communications and middleware, can be established by analysis and optimisation modelling.

4.7 Semantic/Middleware Layer Boundary

Instance of information exchange: Just as communications can be described in terms of an instance of communications, so we can also extend the concept to an instance of information exchange. This requires a minimum of two intelligent agents, a common context for information exchange, a process for exchange and a communications channel. The idea is illustrated in Figure 5.
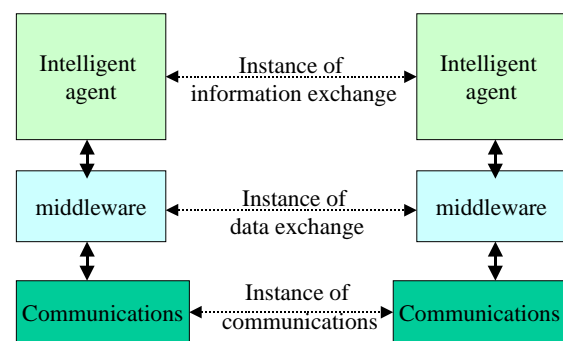
Figure 5 'Instance of Exchange' model

There are various social and motivational models of information exchange:
- Pull (I need to know about *X*)
- Push (I deem it appropriate for you to be informed of *Y*)
  - Agent to agent/agent to object
  - Agent to many (list or broadcast)
  - Acknowledged/non acknowledged.

The implications of information exchange are varied and may be merely informative; more generally information exchange is a mediating

process in inter-agent activity to achieve common understanding and fulfil common goals. These models of *push* and *pull* are further elaborated in [15].

5. SYSTEM INTEGRATION

5.1 Introduction

Integrating the design results from HF, IM, interoperability, management, security and vulnerability is an immature engineering process. This process must also recognise that we are now building what should be regarded as 'living' systems, with rapidly changing users and user needs, different sociological models of how they are used, as well as the more obvious dynamics in their topologies of their use, placement and communications. This new dynamics of systems use, and systems acquisition and through use support is posing many questions that in the demanding environment of the military have not yet been properly answered.

5.2 Keys to the future

Architecture: It has long been the belief of systems developers, designers and those developing long term requirements that any future system must be built according to well understood and widely agreed architectural principles. Unfortunately at the higher levels of abstraction, systems are what are technically called 'soft'. They are not amenable to strong, or hard theorising. For JBD we are nonetheless developing a technical architecture, that is intended to inform both acquisition and commercial developers of military systems. The need here is for a System of Systems (SoS) architecture, and for its communications needs a network of networks (NoN). Such an architecture must also provide the ability to inter-work with national legacy systems and coalition systems.

Interoperability: A useful approach towards understanding interoperability is to evaluate interoperability matrices that relate different interactions between users, and then to quantify these interactions, in terms of need, service description, and finally in terms of the protocol and physical interface needs. This can be undertaken in both a technical and a cost-effectiveness sense. Major problems are that HF and IM are not mature, and the need to inter-work with legacy systems. The policy being advocated for legacy systems is to interface to a civil standard in the case of communications. For software the approach is

to pursue a variety of interface options; CORBA object wrapping, the use of interface description language (IDL), and HTML interfaces (web pages) and examination of more powerful schemas based on semantic labelling concepts, using XML.

Systems and Network Management: It is now widely recognized that unless a SoS is in some sense manageable as a complete whole, then control of the system is likely to be lost in adverse or complex circumstances. This is being pursued by invoking a wide range of management capabilities in new systems, and by specific development of inter-system management interfaces. Although longer-term systems using agent based distributed management are attractive, they are unlikely to find application in military systems for some time, because of a lack of standards for these applications and security concerns. Whilst self-healing and self-adaptive systems are of course attractive, especially if a SoS becomes so large that its overall complexity is not within the conceptual grasp of a management function, the implications for denial of service and other security weaknesses demand much more research.

Security and Vulnerability: Security and vulnerability over a SoS poses many problems, notably the use of differing system security policies, different cryptographic and related boundary protection devices, different schemas for their management (placement, local and remote control, key management etc). This problem is difficult enough within a single nation, but when coalition interoperability is needed the difficulties are far greater. Difficulties arise due to the absence of common agreement on the architecture, processes, and the function and performance of boundary devices from different nations.

As a SoS, or a NoN, is expanded, increasing difficulties occur in management, in guaranteeing its security, and in both controlling and meeting the needs of the increasingly diverse community of users. The result is that vulnerabilities and shortfalls present in each system may become manifest over much wider domains and be easier for adversaries to exploit. The interfacing mechanisms themselves may also admit new vulnerabilities. These problems are an important part of the technology challenge for new C2 systems.

6. CONCLUSIONS

The provision of an integrated, readily protected CIS infrastructure that can provide a wide range of resilient services worldwide is far from straightforward.

The ready availability of a wide range of low-cost and powerful COTS products provides the apparent ability to build large powerful complex CIS. However our poor understanding of the human–machine inter-relationship, and the need to include human factors in a far more powerful way than is currently the case, remain substantial risks to the ambition of truly joint, truly seamless, and truly effective coalition command and control.

The provision of JBD, including coalition interoperation, requires not just technical solutions, but different management and procurement processes for its success. These are sociological issues of system-building, that are every bit as critical as the technical aspects.

We are now developing a more effective socio-technical systems-integration capability, that is bringing together the ideas of human information sharing and use, its control and communication by machines, and the man - machine relationship.

References

1. Montefiore, H., Enigma; *The Battle for the Code*, Weidenfeld and Nicolson, 2000.

2. Addison, D. and Fitzgerald G., *Information Systems Development: Methodologies, Techniques and Tools*, McGraw Hill, 1995.

3. Jarayaytna, N. *Understanding and Evaluating Methodologies*, McGraw Hill, 1994.

4. Stapelton, J., *Dynamic Systems Development Method*, Addison Wesley, 1997.

5. Sleigh, A. and Pratley, P. (Col); *'Achieving a Balanced Capability for Future Operations'*, June 1999, unpublished MoD paper.

6. Boynton, A. C. and Victor, B.*, 'Beyond Flexibility: The Dynamically Stable Organisation'*, California Management Review, 34/1, Fall 1991, pp 53-66.

7. McGahan, C., Fernall, R., and Campbell, J., *'Joint Battlespace Digitization Statement of User Needs Human Factors Benefits Analysis'*, October 1999, unpublished DERA paper.

8. Moffat, J and Prins, G., *'A revolution in military thinking? – issues from the 1999 DERA Senior Seminars'*, July 2000, Journal of Defence Science, Vol. 5, No. 3.

9. Hamid, S., *'Digitization – Human Sciences Research Strategy and Co-ordination Mechanism'*, March 2000, unpublished DERA paper

10. Davenport, T. H., *'Saving IT's soul: human-centred information management'*, March 1994, Harvard Business Review, pp 119-131.

11. Rooney, D.,*'Mission Command and Battlefield Digitization'*, March 1998, unpublished DERA paper.

12. Henderson, S.,*'Shared Situational Awareness in PJHQ Teams"* May 1999, unpublished DERA paper.

13. Bonner, M., Taylor, R., Fletcher, K. and Miller, C., *'Adaptive Automation and Decision Aiding in the Military Fast Jet Domain'*, unpublished DERA paper..

14. Black, U., *OSI: A Model for Computer Communications Standards*, Prentice Hall, 1990.

15. Davenport, T.H., *Information Ecology: Mastering the Information and Knowledge Environment,* Oxford University Press, 1997.

This page has been deliberately left blank

_____

Page intentionnellement blanche

Information System for Logistics – Modern Tool for Logisticians

Zdeněk Buřival, Grad. Eng.
AURA Ltd.
Úvoz 56
602 00 Brno
Czech Republic
E-mail: burival@aura.cz
Internet: www.aura.cz

Lt. Col. Jaroslav Řeha, Grad. Eng.
Director of ISL Project
General Staff of the Army of the Czech Republic
Prague
Czech Republic
Internet: www.army.cz

SUMMARY

The article in its first part characterises main features, capabilities and architecture of the Information System for Logistics (ISL). The second part describes the way and main principles used for building the ISL.

The authors used in the article their experience acquired during their work in the joint development team of the ISL built for the Ministry of Defence and the Army of the Czech Republic.

1. HISTORY AND PRESENT

The Army of the Czech Republic (ACR) has more than twenty-five years long experience in building and using information systems for logistics. Information systems were built to support various areas of logistics. The materiel management was individual for each from about 30 groups of materiel and as well information systems were built separately mostly using mainframe computers. This way of non-uniform materiel record keeping had many disadvantages and led to ineffective materiel economy.

The ACR made a courageous decision to start building a new integrated Information System for Logistics (ISL) in 1994. The aim was to build up a system that will assure compatibility with NATO in logistics benefiting from modern logistic approaches of NATO (e.g. the NATO Codification System) and will increase efficiency in logistic processes saving human and materiel resources. Decision makers of the Ministry of Defence (MoD) and the ACR defined the main objectives of the new ISL:

- To provide a tool for economical control in peace, threat period and in wartime
- To provide a uniform support for all services in the ACR with unified management of all materiel items
- To assure a clear visibility on units' performance
- To allow maximal control over expenditures: man-hours and funds
- To enhance flexibility and reliability
- To achieve interoperability with external systems (NATO, Parliament, other military information systems)

The new ISL should use up-to-date ways of developing information systems and modern means of information technologies.

Since that time the ACR overcame a long way and now a crucial part of the ISL is finished and after completion of the test run and security certification it will be put into full operation.

Thanks to the decision made a long time before joining NATO, the ACR has now the functional, unified, integrated Information System for Logistics – one of basic prerequisites for logistic interoperability within NATO.

2. ISL CAPABILITIES

The Information System for Logistics (ISL) provides support for military logistics in all important areas for both **consumer logistics** and **production logistics**. It ensures uniform support for all services of the armed forces in all over the territory of the country.

2.1. ISL Position and Connections

The diagram (Fig. 1) shows schematically the ISL position and interfaces to other organisations and information systems. There are shown existing connections and potential interfaces prepared for a time, when other information systems will be put into operation.

♦ **Financial Information System**

Connection to the Financial Information System (FIS) allows using the uniform materiel identification according to the NATO Codification System, which is determinant for all armed forces and is assured by the ISL module MC CATALOGUE. Further it is used for sending accounting data about completed materiel transactions from the ISL to the FIS and an interchange of information related to making and consuming a budget.

♦ **NATO Maintenance and Supply Agency (NAMSA)**

The connection to NAMSA is very important. It is used for electronic data interchange of materiel codification

data (transactions) between countries using the NATO Codification System.

♦ **Staff Information System**

It allows a flow of information important for operations planning and controlling.

♦ **Personal Information System**

Interface to the Personal Information System allows interconnection important for an area of preparation and employment of logistics operatives.

♦ **State Information System**

The State Information System defines standards for a uniform information interchange between governmental departments.

♦ **Parliament**

There is assured a possibility to provide Parliament (e.g. the Defence and Security Committee) with reports about a status of crucial weapon systems.

♦ **Logistic Functional Area Sub-System (LOGFASS)**

This interface will allow exchanging the logistics related data of formations assigned to NATO plans by nations.

♦ **Stockholding and Asset Requirements Exchange (SHARE)**

The electronic data interchange with SHARE will allow mutual informing about materiel asset availability and requirements for future joint procurement actions and mutual support.

2.2. ISL Functional Structure

The diagram (Fig. 2) shows the functional structure of the ISL. The diagram depicts the basic functions only, which support mainly materiel, supply and maintenance functions of logistics. The ISL is planned to be enlarged in the second phase to support also other logistic functions, e.g. service, movement and transportation, engineering, etc. Basic characteristics of ISL sub-systems will be described in the following text.

♦ **MC Catalogue**

The *Materiel Codification Catalogue* (*MC Catalogue*) is a basic module of the ISL. It is a tool for materiel codification according to the rules of the *NATO Codification System* (*NCS*) and for logistics categorisation. It allows electronic data interchange with other NATO and non-NATO countries via the NATO Mail Box System (NMBS). The *MC Catalogue* is the heart of the ISL and serves to all other modules and sub-systems as a source of data about materiel. The *MC Catalogue* could be operated as a standing alone system or as a module of the ISL.

The main objectives of the *MC Catalogue*:

− To create a uniform catalogue of materiel for all services of the armed forces

− To provide support for codification of materiel according to the rules of the NCS

♦ **Logistic Requirements**

The *Logistic Requirements* provide tools for work with organisational structures of the armed forces (commanding, logistic, financial) derived from the strategic doctrine and logistic methods of the armed forces. It contains the database of all organisational parts of the armed forces (formations and units) including their links and sub-ordinations.

By means of this module, materiel norms used in the armed forces are also maintained.

Materiel norms define entitlements and specific needs of the organisational parts and their specific needs for various activities. The *Logistic Requirements* also ensure connections between the general materiel norms and standards and specific parts and on that basis it makes it possible to calculate the entitlement of each part with respect to specific materiel items.

♦ **Materiel Record Keeping**

The *Materiel Record Keeping* (*MRK*) is the second basic module of the ISL. It is a tool for the management of central materiel record keeping and creates a basis for a uniform supply system. It follows movements and shipments and the stock level of materiel enrolled in the *MC Catalogue*. It enables transmission of data concerning materiel transactions to an accounting information system (FIS).

The main objectives of the *Materiel Record Keeping*:

− To create a basis for a uniform supply system

− Central materiel record keeping management

− Transmission of data concerning materiel movements and shipments to accounting

♦ **Supply Management Sub-system**

The *Supply Management Sub-system* (*SMS*) covers the **supply function of logistics** that is armed forces materiel assurance. The objective of this area is, within the determined financial and materiel limits, to create optimal materiel conditions for the fulfilment of the armed forces' tasks. The *SMS* is composed of two partial sub-systems: the *Inventory Management* and the *Distribution Management*.

• **Inventory Management Sub-system**

It supports in particular the planning and procurement of materiel. The *Inventory Management Sub-system* is composed of the following modules (according to the DoD terminology called Computer Software Configuration Items – CSCI):

◊ *Provisioning* – One of the most important activities of military logistics is to provide the armies with materiel and equipment. There is an immense amount of labour associated with assurance of the operability of the equipment with spare parts and in-time supply of ammunition for the individual organisational parts, while minimising the warehouse inventories which freeze considerable financial and human resources. Therefore, objective needs must be identified. This is done by means of analysing the present situation by:

   − Comparing the current inventory level (*Materiel Record Keeping*) with the norms for the individual organisational parts (*Logistic Requirements*)

   − Including expected consumption derived from experience of past consumption (the *Materiel Record Keeping*) with known past and future activity of the armed forces (the *Equipment Maintenance Sub-system*)

- Taking into account the expected time of delivery
- Taking into account the time of usability of materiel (the *Materiel Record Keeping* – expiration), percentage of reparability of the damaged materiel (the *MC Catalogue*), etc.

The result consists of an objective evaluation of the future needs of the armed forces. This result is subsequently modified according to the budget resources assigned (taken from the Planning, Programming and Budgeting System – PPBS). After such a modification it is submitted for implementation (the *Acquisition and Procurement Directions*).

Information concerning the surplus or useless materiel identifies inputs of the *Disposal*, where it is further processed.

The main objectives of the *Provisioning*:
- Forecasting the needs of the armed forces and comparing them with resources; offering objective data for decision-making by an item manager
- Providing data needed for the generation of an acquisition and procurement direction
- Signalling the existence of surpluses and suggesting a method of processing

◊ *Acquisition and Procurement Directions* – The purpose of this module is to support decision-making concerning which materiel will be procured for the armed forces and in what amounts (purchase, manufacture). The main input consists of data created by the *Provisioning*.

The directions created in the *Provisioning* with respect to the assurance of the materiel needs of the armed forces, are connected together to form requirements that are approved within the scope of the hierarchical structure of the armed forces and hand over to the acquisition centre. Appropriate contracts are signed with suppliers. These contracts are followed in the *Acquisition and Procurement Directions*.

The information concerning the individual expected supplies is distributed to the individual local servers (in warehouses equipped with the *Receiving*) that on this basis implement receipt of materiel from suppliers.

After physical receipt of the supply, data concerning the gradual fulfilment of the contracted supplies are sent from the *Receiving* to the *Acquisition and Procurement Directions* at the central server.

The main objectives of the *Acquisition and Procurement Directions*:
- Creating, approving and checking the acquisition and procurement directions
- Record keeping of the contracts entered into and monitoring of their execution
- Providing the data from the contracts necessary for the receipt of materiel

◊ *Distribution Directions* – Their purpose is to create directions for the complementing of the inventory level according to the amount standard applicable to the individual organisational parts of the armed forces.

Distribution directions are created based on requirements for materiel from the organisational parts of the armed forces (entered at local servers by means of the *Distribution Directions*), or based on a parametrisable automated calculation done by comparing the current inventory level (the *Materiel Record Keeping*) with standards for the individual organisational parts (the *Logistic Requirements*).

The distribution direction is sent to the local server of the issuing warehouse (central warehouse or armed forces formation) to be implemented.

The main objectives of the *Distribution Directions*:
- Enabling control of the distribution process in the armed forces
- Creating distribution directions automatically and manually
- Managing materiel requirements

- **Distribution Management Sub-system**

The *Distribution Management Sub-system* (*DMS*) supports in particular supply and materiel management (supplies to armies according to the defined standards, creation of the necessary inventory, supplies of consumer materiel, definition of conditions for materiel storage, materiel handling, sale and disposal of surplus and useless materiel). The *DMS* is composed of the following modules:

◊ *Movements and Shipments* – the basic tool for the implementation of physical movements and shipments within the armed forces.

◊ *Receiving* – the basic tool for the implementation the physical movements and shipments from civil organisations to the armed forces and from the armed forces to civil organisations.

◊ *Storage* – supports storage of materiel at the place of its location, stocktaking, maintenance of materiel in long-term storage, etc.

◊ *Issuing* – supports the activities connected with the collecting of items from locations in warehouses on the basis of directions from other modules.

◊ *Disposal* – supports the process of retrieval of items to be disposed of, the approval of the individual recommendations for materiel to be disposed of, and the monitoring of the process of the implementation of physical disposal or sale of the materiel.

♦ **Logistic Management Sub-system**

The *Logistic Management Sub-system* (*LMS*) serves in particular top-level logisticians. It is composed of the following modules:

◊ *Control (Performance Indicators)* – is designed for the top level of logistics. It serves to evaluate the logistic performance and effectiveness of logistic entities (formations, warehouses, bases) and processes (supply, maintenance, storage, etc.). It defines the objectives, intentions, measurable norms and performance indicators. Based on these, it compares the actual performance with the relevant norms and presents the result to the user.

The *Control (Performance Indicators)* contains tools that make it possible to combine the performance indicators into groups for evaluation of comprehensive situations. It also makes it possible to perform several types of result analyses in order to support decision-making.

◊ *Operation Logistic Support Planning* – provides support for the logistic assurance of operations of the armed forces. It provides a comprehensible overview of the existing situation during the entire operation and supports continuous planning of the logistic support. Information concerning the achieved status of logistic assurance and the existing requirements are transmitted between the individual commanding levels of the armed forces by means of logistic reports.

Logistic support planning takes place with a close direct link to the command and control of the operation.

♦ **Equipment Maintenance Sub-system**

The *Equipment Maintenance Sub-system* (*EMS*) is designed to support the activities related to the planning and executing of the equipment maintenance. A substantial part of the *EMS* is formed by the monitoring of the operation of equipment. The *EMS* is composed of the following modules:

◊ *Standards, Norms and Procedures* – create a database containing a source of information necessary for the majority of processes of the two remaining EMS modules. The database contains in particular information concerning specifications of maintenance and constant tables of various codes with corresponding textual descriptions. The module also makes it possible to monitor the operation, maintenance and composition of specific items of equipment. (Specific items of equipment are explicitly defined in the *Standards, Norms and Procedures* by means of a catalogue number and a record keeping number.)

◊ *Maintenance Planning* – supports the user in creating long-term and detailed plans of maintenance, in evaluating the use of maintenance resources with respect to their capacities, and in developing working plans for repair facilities.

The module also makes it possible to monitor the actually implemented operation of the specific items of equipment and compare it with the annual plan of operation.

◊ *Maintenance Execution and Control* – supports monitoring of maintenance (both scheduled and unscheduled) executed both in the armed forces' repair facilities and in repair facilities outside the armed forces. By means of a link to maintenance, it is possible to keep records concerning the resources consumed (standard hours, spare parts, services purchased). Information so obtained is used to evaluate the norms of resource consumption for individual types of maintenance and for financial definition of costs of the maintenance executed.

The module also makes it possible to record information concerning defects. Such information is then used in the calculation of performance indicators by means of which it is possible to evaluate the reliability of the equipment.

2.3. Life Cycle of Item of Supply

The capabilities of the ISL support an item of supply during the whole life cycle of materiel. The basic scheme of the item life cycle supported by the ISL functions is depicted on the diagram (Fig. 3).

− The *Provisioning* compares the current inventory level (from the *Materiel Record Keeping*) with the norms for units, formations, etc. (from the *Logistic Requirements*) and as the result calculates the future needs of the armed forces. This result is corrected according to a budget taken for example from the PPBS (Planning, Programming and Budgeting System). The corrected result is submitted for implementation to the *Acquisition and Procurement Directions* and information concerning the surplus or useless materiel enters the *Disposal*.

− The *Acquisition and Procurement Directions* sum the directions created in the *Provisioning* according to possible suppliers and create requirements for the acquisition centre. The data about a concluded contract enter the *Acquisition and Procurement Directions* where they are followed. The information about expected supplies is then distributed to the *Receiving* on the individual servers.

− The *Receiving* supports the process of receiving of materiel from civil organisations to the armed forces. It compares incoming materiel with contracts and creates reports about fulfilment of the contracts for the *Acquisition and Procurement Directions*. The *Receiving* also generates data, so called transaction documents, which are sent using modules of the *Materiel Record Keeping* to the Financial Information System for accounting purposes.

− The *Storage* supports activities performed in warehouses like storing items in a proper location, stocktaking, etc.

− The *Disposal* processes materiel to be disposed of.

− The *Distribution Directions* create directions for the complementing of the inventory level according to the amount standard defined for each organisational unit of the armed forces. These directions are the input for the *Issuing*.

− The *Issuing* creates plans for issuing according to the directions obtained from the *Distribution Directions*. It supports a process of collecting materiel from a warehouse and prepares data about the collected materiel for the *Movements and Shipments*.

− The *Movements and Shipments* support the process of packing, completion and shipping shipments. It monitors physical movements of materiel amongst the organisational units of the armed forces and

provides the *Materiel Record Keeping* with data – transaction documents, for accounting in the Financial Information System.

– Issued and distributed materiel is used, maintained and consumed in the armed forces and information about the actual inventory level is used by the *Provisioning* for new calculation of armed forces' needs.

3.   ISL ARCHITECTURE

The ISL manages logistic activities, items, stocks and transactions. The stocks can be found in warehouses, in units, in movement between warehouses and units, and in the process of receiving (stock comes into the armed forces) or process of consumption and disposal (stock gets out of the armed forces). Logistic activities (like maintenance, provisioning, etc.) are also done at different levels of the armed forces.

The main concept is to allow independent work in different sites, but managing logistics in a central armed forces system. This concept lets each site (warehouse, unit or troop recording department) manage stocks and record transactions without regarding to the availability of communications to the central site. On the other hand, the central files will be updated (on-line or by batch) for every transaction, thus allowing to maintain a centralised logistics management, and capability for the provisioning and acquisition of stocks.

To meet the above mentioned concept the ISL is designed as a centralised distributed system. A network of servers is deployed over all the country including extraterritorial locations. The servers are interconnected by WAN, telephone lines or, when electronic connection is not available, data are distributed by a suitable memory medium like CD ROM. Databases in all servers are updated by the sophisticated utility "Data Distribution" so that information in all locations is always consistent and up-to-date. The scheme of the *Supply Management Sub-system* architecture is depicted on Fig. 4.

4.   TECHNICAL BACKGROUND

The state-of-the-art information technology is used for ISL development. The application is built using the multi-tier client / server architecture with a graphic user interface. The whole process of software development is supported by so called "technological line" composed of a set of software tools like CASE (Computer Aided Software Engineering), source code generators, tools for automated software testing and of methodologies and working procedures. Developers can focus their effort to creative work and routine activities are carried out by supporting tools.

4.1.   Technological Line

Considering other industrial branches automation in software development is still little used. It is obviously given by it that the major part of software development activities has creative nature and it is difficult to automate creative work. Nevertheless, even in software development there is number of tasks that can be supported by various software tools. Thus the development team can focus its effort mainly to creative activities and to reach high effectiveness.

For large projects other reason for using suitable tools for software development arises. A large team numbering several tens of members cannot be composed of only top analysts and programmers and badly there will be number of only average members. Therefore it is preferable for a project manager to establish a narrower team of top creative employees whose task is to develop and establish from selected components such a support for the whole team that even average analysts and programmers could work effectively and generate high-quality software.

It is appropriate to use commercial-off-the-shelf (COTS) software for establishing the development environment. However COTS software usually does not meet fully all requirements of project methodology and it is necessary to modify it or to develop additions. Therefore one of the most important features for software selection is their openness.

The selected COTS software, its additions and modification must be then completed by working procedures corresponding to chosen methodology covering all stages of a project life cycle. By creating the working procedures defining what must be done in what time, what are inputs and outputs, what tools will be used and what professions will participate a software company gets closer to production in other industrial branches. A software company acquires thus a technological line for production of software and creates prerequisites successful solution of even large software projects.

To create a high-quality technological line is not a simple task. It is a gradual process based on collecting experiences and ongoing "tuning" of all components of a technological line to optimal configuration.

A high-quality technological line should comprise of tools supporting the whole project life cycle. The itemization of key tools used in particular stages of the project life cycle follows:

♦ **Analysis**
–   Tools supporting record-keeping of software requirements (advantage if it is a part of a CASE tool)
–   CASE tool
♦ **Design**
–   CASE tool
♦ **Programming**
–   CASE tool
–   Program languages, the main program language should have a bridge from a CASE tool supporting automated source code generation
♦ **Testing**
–   Tools for automated testing software supporting also manual testing
–   Tools supporting the record-keeping of software requirements (allowing monitoring verification of specified requirements)
–   Tool supporting software configuration management

♦ **Maintenance**
– For software modification same tools as for development are used
– Tool for the record-keeping of problem reports (advantage if it is a part of a tool for automated testing)
– Tool supporting software configuration management

5.   CONCLUSION

Success of large software projects is given on one hand by a clear vision of a customer as a future user of an information system and on the other hand by a capability of a development team to realise customer's vision and to develop high-quality software meeting users' needs.

It could be stated that in the case of the Information System for Logistic both prerequisites where fulfilled and the success has been reached. It is necessary to admit that the way was not always easy and both the customer and the developer had to learn and look for an optimal solution of difficult situations.

On the Development of Command & Control Modules for Combat Simulation Models on Battalion down to Single Item Level

Prof. Dr. Hans W. Hofmann
Dr. Marko Hofmann
University of the Federal Armed Forces Munich
Department of Computer Science
Institute for Applied Systems Science and Operations Research (IASFOR)
Werner-Heisenberg-Weg 39
D-85577 Neubiberg, Germany
e-mail: hofmann@informatik.unibw-muenchen.de

Summary: *The paper contains an **overview on the design principles and main characteristics of a family of new, strictly object-oriented combat simulation models called COSIMAC** (CO_mbat_ SI_mulation_ M_odel with_ A_uto-mated_ C_ontrol), developed at our Institute since 1995. They are designed as closed models which means, that a detailed modeling of the highly complicated C³I processes is indispensible. Additionally, the option of an interactive man/machine interface is implemented, which offers the possibility of manual control on different command & control levels for playing against computer generated (and controlled) forces, for experimenting with "unconventional" decisions, and for developing and improving the rule system in a trial-and-error fashion. Furthermore, the paper describes **a general architecture for the design of command & control modules**, which offers the possibility of describing tactical/operational intentions and concepts of operation in a kind of battle management language (**multilayer tactical language concept**), the **terrain representation, attrition and movement modeling**, the **development of terrain and situation assessment modules**, which are - together with a set of so-called **planning functions and spatial and procedural templates** - a prerequisite for the rule systems that generate adequate tactical missions and orders for the assigned units or simulated objects, and, finally, the **main results, conclusions** and **future developments** of the project.*

1 On the Development of the New Combat Simulation Model Family COSIMAC

1.1 Background and Main Characteristics of the New Models

As a contribution to the NATO RSG.18 study on Stable Defense (see [Hofmann et al. 95]) and in context with two doctoral thesises at our Department (see [Tolk 95] and [Schnurer 96]) more than 30,000 simulation experiments on division/brigade level have been performed with the closed, rule driven battle simulation model KOSMOS over a period of four years. They cover more than 340 different scenarios featuring, attacks by two different types of generic divisions against three different types of defending brigades under different situational conditions involving three types of terrain, two visibility conditions, up to three degrees of defense preparation and different combat modes (e.g., with or without a preceeding delaying battle). Thus, in addition to addressing the primary questions raised, e.g., by the RSG.18, the experiments offered a unique opportunity for testing a rather complex model in the light of results, leading to a continuous improvement primarily of the rule sets controlling the tactical and operational decisions.

One of the experiences we gained with the KOSMOS simulation experiments was, that - for a further substantial improvement of the rule sets that control the assigned combat and combat support units - the rule system must know the tactical/operational intentions and the concept of operation of the superior command and control level in order to react adequately. This particularly applies to cases of surprising events. Otherwise the lower units would react unadequately or - in the view of the superior command - in a rather self sufficient way.

For example, the battalion commander must know the concept of operation of the superior brigade to react adequately in the view of the brigade. In reality, the battalion commander knows that from the operational order or the context of the situation (or common held exercises etc.).

Therefore, one of the objectives of the following COSIMAC project was to describe tactical/operational intentions and options, and the corresponding concepts of operation for a set of relevant scenarios, combat modes and command levels in a computer readable way (e.g., with a kind of tactical/operational battle management language which generates spatial and procedural templates) so that the rule system could operate dynamically in accordance with the (long term) intentions of the higher command even in case of rapidly changing situational conditions and/or failure of the communication system. With this approach it should also be possible to model the

(so-called) German "Auftragstaktik", a special type of mission-type-tactics which offers, among others, the lower command levels a comparatively high degree of independence and flexibility in exploiting favourable opportunities.

Furthermore, the new combat simulation system should provide a higher resolution than KOSMOS, to enable modeling on the basis of physically measurable input data and thus being no longer dependent on the insertion of aggregated data (i.e., Lanchester-coefficients), that had been derived beforehand by running a high resolution simulation model. (Regarding the problems incurred in deriving Lanchester-coefficients on the basis of results obtained by running a high resolution battle simulation model see [Schaub 91].

Fig. 1.1 summarizes the main characteristics of the new combat simulation models COSIMAC in comparison with the KOSMOS model.

- Higher resolution of the battlefield (down to single weapon systems, no Lanchester-approach for attrition modeling for the major weapon systems)

- Interactive and/or closed model version at the user's disposal

- Realistic representation of military C^3I-processes with C^2-modules down to lower command & control levels (single item, platoon, company, battalion)

- Development of sophisticated terrain analysis modules

- Data bank oriented

- Running on PC's and workstations (hardware independent)

- Strictly object-oriented with SMALLTALK or C++

Figure 1.1: Main Characteristics of the New Combat Simulation Models COSIMAC in Comparison with the KOSMOS Model

1.2 Design of the Central Simulator

Experiences obtained in developing combat simulation systems in the past at our institute have shown that a flexible architecture requires a **strict and rigorous separation of the pure combat processes** of the basic combat (or combat support) elements (objects at the lowest level of resolution; in our models, e.g., platoons in COSIMAC-P or single weapon systems in COSIMAC-WS) **from the command and control processes which control these objects**.

A major problem was constituted in "hidden" assumptions regarding the tactical behavior of these combat objects. Actually such assumptions are often firmly connected with these objects.

In order to create a rigorous separation between different processes the architecture of the new models follows a concept of layers (see Fig. 1.2). The main layers are established by the *central simulator* and a set of *command & control modules*. Besides there exist explicit *interface layers* that serve as an additional abstraction and explicit *user interface layers*.

In the **Central Simulator** (see Fig. 1.3) the terrain, environmental data (e.g., weather data) and the basic combat (or combat support) elements as well as their associated models are being administrated. Furthermore the central simulator controls the simulation. The combat elements are described by a set of mainly physical and/or technical (input) data. The elementary processes of these objects are implemented in the associated models. These encompass reconnaissance, attrition, movement, communication, manipulation of the environment, and transport of combat elements. Furthermore, the associated models imply basic knowledge for command & control processes, e.g., target acquisition and selection, route planning etc.

The **layer of the command & control modules** consist of the C^3I-modules for the different C^2-authorities and/or an interactive user interface. Furthermore, the C^2-modules administrate the individual perceived situation of a C^2-authority.
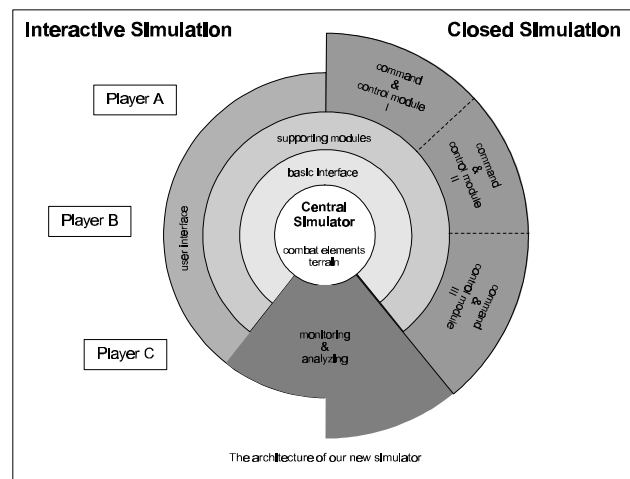
Figure 1.2: General Architecture of COSIMAC

The Central Simulator allows the access to the basic combat (or combat support) elements only by a set of elementary orders by which these elements are being controlled. This set of elementary orders is not identical with the battle management language as described in *Chapter 3*, and which will be used only between and within the command & control modules and moreover will also be much more extensive. For this very reason the command & control modules do not communicate directly with their associated combat elements but rather via a further interface object. Conversely, the communication of the combat elements addressed to their command & control modules also works via an interface. Thereby a strong logical separation between the combat elements and command & control modules can be achieved thus providing an easier exchangeability, extension and optimizing of the command & control functions. (In older models the firm

connection of command & control knowledge with the basic combat elements had revealed itself as an impasse.) Moreover, this design principle also offers the possibility of a "physical" separation between the command & control modules and the combat elements, i.e., the possibility of running the simulation on several computers is considerably facilitated.
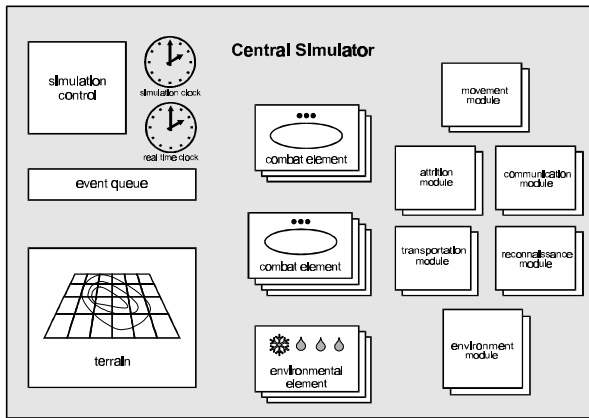
Figure 1.3: Architecture of the Central Simulator at Platoon Level (COSIMAC-P)

Nevertheless, elementary command & control knowledge has been implemented in the models associated with the basic combat elements.

The reason for this is that a strict separation of all elementary command & control processes from the basic combat elements would cause an extremely high flow of information via the interface between the Central Simulator and the command & control modules. Since, however, tactically adequate behavior of the basic combat elements is highly dependent on the specific situation and location it must be exchangeable during the running period. Therefore, a single combat element is not only allotted to one associated model for an aspect such as, e.g., target aquisition or movement, but rather to several models, one of which is the currently used one.

Since the tactical behavior of the basic combat elements used in the Central Simulator is defined by terrain to a high degree, the representation of terrain is explained in the following chapter.

1.3 Terrain Representation

For model development, implementation and testing we presently employ the digital maps used already in the BASIS simulation experiments[1]. They comprise three different pieces of real terrain in Germany and resemble the following terrain types:

- mountainous/wooded (Furth im Wald, 8*14 km^2),

- rolling hills/partly covered (Bubach, 6*10 km^2),

- flat/open (Grettstadt, 6*10 km^2).

They contain elevation, terrain vegetation and trafficability data for square grid sizes of 25, 50 or 100 m. The altitude resolution is 10 cm, resp. 1 m.

The altitude of the different combat vehicles as well as the altitude of the different terrain cells are considered when checking the line of sight. Additional visibility barriers are taken into account. These obstacles may be static (e.g., buildings, vegetation) or change dynamically during the simulation (e.g., to represent smoke screens dispensed by artillery or combat vehicles to protect themselves).

Moreover, vector data referring to roads and rivers were used in order to determine a set of *cell transfer velocities* that are subject to the four orthogonal directions and that rise values for each terrain cell calculated in connection with vegetation and altitude data. This enables us to generate and store these values in advance, according to mounted or dismounted basic combat elements and, if mounted, to the different types of vehicles (i.e., wheel, chain, air borne, etc.).

In addition to these static values taken mainly from the natural conditions of the regarded terrain, the _concept of cell transfer velocities_ can also be extended by dynamic aspects such as obstruction and fire. This allows us to build simple but extremely flexible movement models for our basic combat elements in combat situations that are capable of considering natural terrain obstacles as well as obstructions and other artificial obstacles and moreover even obstacles caused by enemy fire.

Disregarding the cellular terrain representation on which most of the internal simulation processes are based with respect to the basic combat elements in an off-road combat mode, our display concept also permits the superimposal of pure bitmap- or vectordata-displays.

In addition to the old BASIS terrain data we have digitized the terrain of the CMTC HOHENFELS in a similar way to get the possibility of comparing some scenarios (initial situation and combat dynamics) of real held exercises in the CMTC with replayed scenarios performed with the COSIMAC models.

1.4 Attrition and Movement of the Combat Elements

Movement Modeling

---

[1] BASIS is a high resolution, stochastic Monte Carlo -type combat simulation model developed 1982 - 84 in PL1 for a mainframe computer at our University. It permits the detailed simulation of battalion-size ground forces defending against a sequence of regimental-size attacker forces. Resolution goes down to every combat vehicle or weapon system. It is a closed, script driven model (without C^2-models) and was extensively used for a study on „Non-offensive Defense Options" in 1984 - 85 and the derivation of Lanchester-coefficients for simulation experiments with the KOSMOS model. In 1992 it was re-implemented in PASCAL on an Apollo workstation and in 1996 in C++ on a PC. Presently it is used for test purposes, e.g., for comparing the results of the pure combat processes of the new model COSIMAC with those obtained with the BASIS model. For further information on the BASIS model see [Hofmann et al. 84].

The modeling of movement of the low level combat elements is a decisive process in every high resolution combat simulation system, especially if the system is able to conduct **automated route planning**.

Since COSIMAC is based on a square grid terrain model we first implemented the algorithms which are most commonly used to optimize routes in grid models, the <u>Dynamic Optimization algorithm</u> or a specific version of <u>Dijkstra's algorithm for route planning</u> (see, e.g., [Fould 1992]). However, in order to improve the performance of the Central Simulator we currently use a <u>special algorithm</u>, which turns out as a mixture of Dynamic Optimization, Dijkstra's algorithm and Branching & Bounding. The basic idea of this approach is to reduce the number of nodes (or terrain cells) permanently labelled by the spirit of Branching & Bounding by taking into account not only the time (or cost of movement) from the starting to the regarded terrain cell but also some estimate of the further distance from the regarded node to the target terrain cell. Furtheron, the possible routes of a regarded combat element (or unit) are confined beforehand by assigning mobility corridors. In other words: The combat elements can only move within predefined corridors which correspond to predefined combat sectors. Additionally, we made some research on simplified heuristic versions of this algorithm which will not find precisely the real optimum in any case, but are much faster.

In many combat simulation systems usually only one criteria is chosen for the route planning:

- <u>time</u>, e.g., the goal of optimization is to minimize the time a combat element needs to reach a (given) target cell.

We consider three additional criterias:

- <u>path length</u>, e.g., minimizing the length of the path between the current position and the target cell, taking regard of given constraints,

- <u>concealment</u>, e.g., maximizing the protection against sight (visual detection) given by vegetation or buildings, and

- <u>altitude</u>, e.g., trying to find a path that minimizes the altitude of all cells you tread on the way to the target cell. (This criteria can be understood as an attempt to maximize cover.)

These extensions and the combination of them not only allow us to automate route planning, they do also provide us with the ability to model appropriate tactical behavior, for instance to move a combat element or unit "like water flows" or to choose routes which avoid weapon engagement zones, detected or supposed mine fields, etc. Furthermore, this approach makes it possible to design a route planning algorithm for helicopters, e.g., using a modified version of the "altitude-oriented" algorithm.

Attrition Modeling

In accordance with the different classes of weapon systems the combat simulation system COSIMAC contains different kinds of attrition modeling approaches. The most important are roughly described in the following chapters:

- **Attrition model for the direct fire weapon systems**

  Regarding the direct fire weapons the attrition model in the single item version (COSIMAC-WS) operates on the (individual) single shot approach. For the platoon level (COSIMAC-P) neither a pure single-shot model on the single weapon system level with individual fire control nor an aggregated Lanchester-equation model is implemented. Since in the regarded version platoons are the basic combat element, it is assumed that all major weapon systems of a platoon fire at the same moment under the control of the platoon leader (or leading weapon system). This simplification can be justified by the fact that the fire unit of the combat troops is normally the platoon that contains similarly equipped vehicles. Keeping in mind this abstraction, the attrition model for the direct fire weapon systems can be roughly described as a multi-shot model on the combat element level. Moreover, COSIMAC is able to model unguided rockets, antitank guided missiles and fire-and-forget missiles of every range, taking regard of possible countermeasures during flight time.

- **Attrition model for the high-angle-fire**

  The model for the high-angle-fire (mortars, artillery systems and rocket launchers) operates on a single shot approach. For all kinds of targets (combat or combat support units) a special spatial distribution (spatial template) for the individual weapon systems is assumed, depending on their activity (see *Chapter 2.2*). With this approach we can base the calculation of losses caused by high-angle-fire on the commonly approved concept of lethal areas around the impact points of the fired shells.

2 C³I-Modeling

Since 1997 we have been working with emphasis on the development of first, simple command & control modules and rule sets for the single weapon system, platoon, company and battalion level. A prototype version for the combat modes *attack* and *defense* is ready for use. The next chapters describe the main design principles and first solutions.

2.1 On the Development of Terrain and Situation Assessment and Planning Modules

2.1.1 Terrain Assessment and Force Deployment Modules

Terrain assessment and subsequent force deployment are

major and indispensable tasks of every military commander. Therefore any C²-module must be able to perform at least some of their subtasks. This is especially true and a demanding task for a detailed terrain model.

We started with the development of a simple, interactive terrain assessment and force deployment tool for defense operations on single item, platoon, company and battalion level for pieces of terrain as described before.

Input data were:

- forward edge of the battlefield (FEBA),
- boundaries of the defense area (defense sector),
- number and type of own troops/friendly forces,
- point of main defense effort,
- areas of field fortifications and target areas for supporting prearranged artillery fire.

These data are given by the op-order for defense.

In a first step, the algorithm figures out all <u>points or areas of interest for a defender</u> such as, e.g.:

- large, interrelated areas of open terrain, of forests and of urban (built up) areas (towns or villages),
- the rims of such areas with respect to the main defense direction, which offer defilade, protection against artillery fire and free firing zones for the low angle fire of the own combat elements (especially important for dismounted infantry),
- hills and other points or areas with good visibility conditions such as observation points or places for weapons with long range direct fire, e.g., long range anti tank positions.

In the next step the program deploys the own basic combat elements (platoons or weapon systems), which are available for the defender in the regarded defense area, into their <u>initial defense positions</u> on or near the FEBA (or into the security line or into positions in the depth) by considering (besides the points/areas of interest of the regarded terrain) the following items:

- type and maximum, minimum, and effective firing range of the combat elements and their effective (terrain dependent) firing zones in the proposed positions,
- the degree of overlapping between the different firing zones in order to minimize the number and size of dead firing zones,
- the point of main defense effort (where the degree of overlapping fire should be extremely high), and
- the (initial) spatial template for defense operations, which divides the defense area in an area on or near the security line, an area on or near the FEBA, an area of positions in the depth, and a rear defense zone.

Meanwhile, this module was extended and integrated into the simulator. With this module it is presently possible to assess terrain on battalion level for defense purposes and to deploy given forces to their initial defense positions. Furthermore, we extended the module to comprise also terrain assessment and force deployment for attack operations. Input data for this module are:

- boundaries of the attack area,
- line of departure,
- objective of the attack, number and type of own troops/friendly forces and reconnoited enemy troops,
- areas of friendly and (supposed or reconnoited) hostile field fortifications, and - as an option -
- one or more intermediate objective(s).

2.1.2 Situation Assessment and Planning Modules

These modules comprise a large number of mathematical functions that may be used in situation (and/or threat) assessment and operations planning.

Examples for <u>situation assessment functions</u> are, e.g.,

- force ratios,
- force concentrations,
- deep penetrations into the defense sector,
- open flanks, etc.

(Many of them are derived from the combat geometry.)

Examples for <u>ops-planning functions</u> are, e.g., estimations with respect to

- speed, space, and time requirements,
- availability of forces,
- losses,
- loss-exchange ratios for planned operations, etc.

A lot of these functions can be taken from the KOSMOS model. But there is still a considerable amount of them which must be newly developed, especially for the lower command levels. An example is elaborated in *Chapter 2.4*.

2.2 On the Development of Spatial and Procedural Templates for Generic Weapon Systems, Units and Force Stuctures, and a General Approach for the Assessment of Tactical Options

2.2.1 Definition of Generic Weapon Systems, Units, and Force Stuctures

With regard to the scope of this project to reflect on general solutions and in accordance with the object-oriented design principle of the simulation system COSIMAC we

are primarily interested in a <u>general scenario design principle</u> which covers a large variety of different kinds of weapon systems, unit types, and force structures. Therefore, we have applied for this project the <u>modular force design principle</u> of the KOSMOS simulation experiments and extended it to the lower command levels. (For more information see [Hofmann et al. 95] or [Hofmann, Hofmann 98].)

2.2.2 Spatial Templates

Spatial templates are defined as models of how objects (e.g., weapon systems, platoons, companies etc.) are positioned and oriented relatively to other objects. On all command levels, they describe (approximately) the spatial arrangement (grouping, formation) of units and sub-units depending on the state (e.g., type of combat) and situational conditions.

Examples:

- column or double column formation,

- line formation (permits excellent fire to the front),

- wedge, inverted wedge or Vee-formation ("Keil", "Breitkeil", formations used when the enemy situation is vague and the leader requires firepower to the front and the flanks),

- two-up (a formation with two elements disposed abreast, the remaining elements in rear).

Fig. 2.2 shows - as an example - the wedge formation ("Keil"). It was designed as a "pulling" template which means, that the leading 1. platoon - which advances to an objective area on the best route as described in *Chapter 1.4* - pulls the following two or three other platoons. They follow dynamically in predefined areas looking for best positions (with respect to cover or field of fire) for their own, whereas the overall heading, depth and width of the template is ordered by the company (or leading platoon).
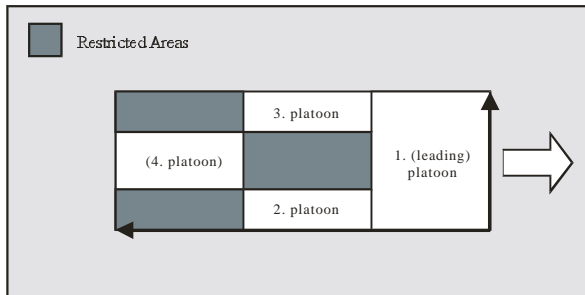
Figure 2.2: Wedge Formation ("Keil")

Fig. 2.3 depicts the inverted wedge or Vee-formation ("Breitkeil") which turns out to be a "pushing" template. In this case the 1. platoon "pushes" the other platoons before him leading and controlling them as described before on the route to the objective.
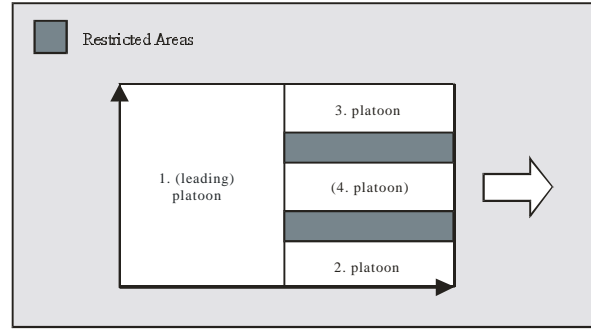
Figure 2.3: Inverted Wedge Formation ("Breitkeil")

Fig. 2.4 finally shows an example for a <u>change of formation</u> from the wedge over the double column to the inverted wedge formation including a change in direction of the whole formation. Intention is to avoid that the platoons interfere with others by overcrossing and/or outpacing the movements of other platoons.

2.3 Procedural Templates

Procedural Templates are defined as models on tactics, techniques and/or procedures which describe how objects or units on all command levels typically operate and work together in different combat modes (Order of Battle, employment of forces, activities, time schedules, combat dynamics, etc.).
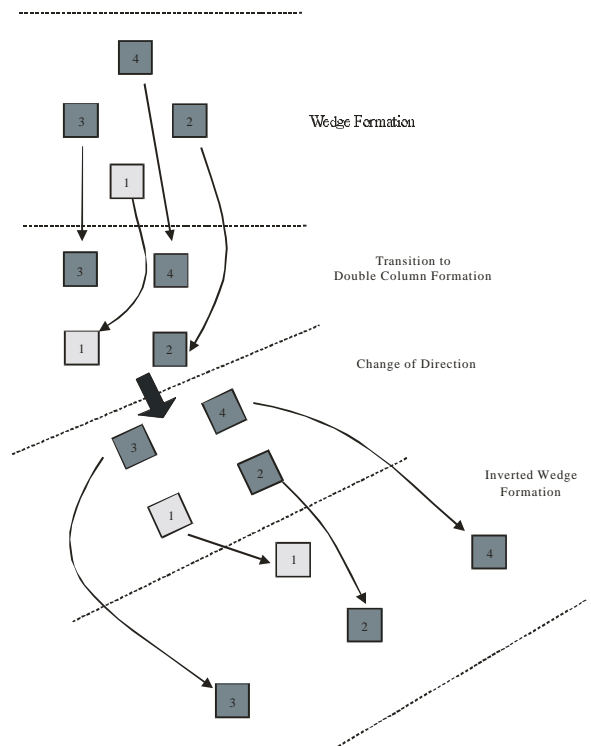
Figure 2.4: Change of Formation and/or Direction

Examples:

- leapfrogging (e.g., one combat module moves, the

other one fires (überschlagendes Vorgehen)) or

- advance in an accordion-like manner (raupenförmiges Vorgehen), but also

- schematic representations of the Order of Battle, employment of forces etc. for the different kinds and phases of an operation as described in the Field Manuals (see, e.g., FM 100-5 or HDv 231/100).

Leapfrogging and advance in an accordion-like manner are implemented on company level.

As an example for defense operations Fig. 2.5 depicts the schematic organization of a prepared defense by a *Mixed Mechanized Infantry Battalion* with two Mechanized Infantry Companies side by side in front-line positions and the Tank Company as a reserve. The scenario was elaborated by the Tactical Center of the German Army (Taktikzentrum des Herres) and published recently in [TRUPPENPRAXIS 9/97].

It shows, as an example, the "implementation" (realisation) of a given schematic organization of a prepared defence according to the respective German Field Manual (HDv 231/100) into an assumed "real" situation taking into account the perceived enemy situation, situation of own troops, terrain and a variety of further situational conditions.
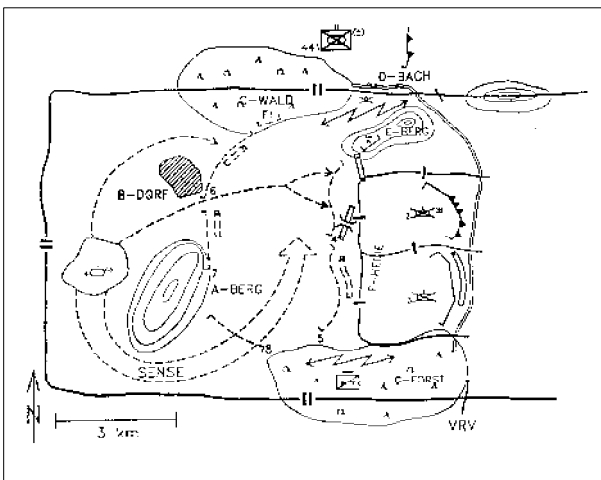
Figure 2.5: Excerpt from the Operation Plan of a Prepared Defence for a Mixed Mechanized Infantry Battalion [TRUPPENPRAXIS 9/97].

2.4 A General Approach for the Allocation of Fire and Forces and the Assessment of Tactical Options

Even though each C²-authority and the different branches or functional staff areas have their own C²-modules (see *Chapter 3),* a general method for the design of C²-modules should be mentioned at this moment: **the appliance of the multi-dimensional utility theory** for solving a large variety of decision problems from the target allocation problem up to the assessment of the effectiveness of tactical options, orders and missions, a concept that has already proven its usefulness in the KOSMOS model.

Background for this approach was the experience we made when asking military experts for rule sets for problems like, e.g., target allocation to artillery batteries, employment of reserves etc. It was very hard for them to formulate general decision rules in a precise "if - then - else" manner. Most often the answer was "that depends on the situation", specifying a set of more than 10 or 20 different influence factors.

The multi-dimensional utility theory approach reduces the general problem of fire or force allocation on all command levels to a m*n allocation problem based on a m*n-utility-matrix with each value representing the expected utility, calculated by using specific, multi-dimensional utility functions. Subsequently, the allocation of fire or forces could, for example, be made in the sequence of the utility values, taking also into account, e.g., the marginal utility.

Multi-dimensional utility criteria for the target allocation problem of artillery batteries regarding one allocation period may be, e.g.,

- expected, weighed damage (expected number of destroyed weapon systems, weighed with their values in the specific situation),

- expected effects of target suppression (often important with respect to infantry),

- cost (negative utility) of the target engagement (e.g., cost of ammunition, "cost" of being detected, etc.),

- tactical/operational aspects or urgency for allocating the target[2].

For more information see [Schnurer 96].

The corresponding criteria for the assessment of tactical options, allocation of forces, employment of reserves may be, e.g.,

- degree (or probability) of performing the given orders, missions (or intents of the higher command & control authority),

- expected weighed losses of adversary forces,

- expected weighed losses of own forces.

With this approach a large variety of different situational conditions can be considered.

However, the main problem of this approach is to get proper estimates of these (situation depending) expected values. This holds true especially for the evaluation of (given) tactical/operational options on the higher

---

[2] In reality, we are confronted with a very complex n-stage, 2-person-zero-sum (game theoretic) allocation problem which is far too complex for an algorithmic solution. In order to consider at least some aspects of the dynamic (or n-stage) dimension of the problem this aspect was introduced as an additional criterion. Furthermore, it offers the possibility of considering the tactical/operational intentions of the higher command in the allocation process.

command levels, at which a kind of "simulation in advance" to get these values would be indispensable.

One possibility for solving the problem would be the use of the same detailed stochastic model for the evaluation process one takes for the simulated ground truth. But that would be very (running) time consuming, especially if one considers that a large number of replications would be necessary to get expected (or mean) values. A second possibility is the development of own aggregated, expected value models for the different evaluation processes. (But this would cost a lot of time for development.)

We voted for the second option and designed and implemented

- a <u>deterministic (expected value) model for the target allocation problem</u> for direct and high angle fire weapons. (It operates mainly on the same input data that are also used for the simulated ground truth by the detailed Monte Carlo simulation model.)

- a comparatively simple <u>aggregated deterministic Lanchester model</u>, which offers the possibility for a dynamic "simulation in advance" for a limited time frame to get estimated results for the (given) tactical options to be assessed. The model operates on the perceived situation of enemy forces. (For calibration purposes it also offers the possibility of working with the real ground truth.)

Up to now, one important common principle for the design of terrain evaluation, situation assessment and planning modules within the C^2-modeling approach has been elaborated: **the appliance of classical algorithmic approaches, optimization techniques and geometric analysis as far as possible in order to exploit the advantages of modern computers for numerical solutions, to attain robust and highly efficient solutions, and to reduce the number of "if - than - else" decision rules.** But this is not sufficient. In the next chapter some further important aspects and design principles for modern command decision modeling are described.

3 On the Development of a Tactical/Operational Battle Management Language and a General Architecture for the Design of Command & Control Modules

3.1 Overview

The Tactical/Operational Battle Management Language should offer the possibility of describing (in a formalized, computer readable manner) tactical/operational intentions and concepts of operation (as described in an Operation Plan) and deliver the prerequisites for breaking down a concept of operation of a higher command level to individual missions and battle orders for the subordinate combat and combat support units, and further down to the simulated objects at the end of the command and control chain. And this should be done - as much as possible - automatically.

Taking into consideration the different levels of aggregation between a battalion operation plan and a platoon order (which are the elementary orders in our Central Simulator at platoon level), we are convinced that the main effort to realize an operation plan is connected with the task of breaking down this plan into individual missions and orders for the simulated objects.

3.2 On a General Architecture for the Design of C^2-Modules

3.2.1 Introductory Considerations

In most of the combat simulation systems realized hitherto, the modeling of the command & control process is confined to the battalion and higher command levels. This is due to the fact that on these levels analytical and geometrical approaches are (regarded as) sufficient to model such processes. But for future systems it is impossible to circumvent the modeling of at least some aspects of command & control at the company, platoon and weapon system level.

In principle, it would be possible to design a Combat Simulation System (CSS), that only copes with command & control processes at the lower echelons, but we consider that to be inappropriate, since some of the most compelling questions concerning the automation of decision making at the lower command levels are closely linked with the corresponding events at the higher levels, for instance:

- breaking down the operation plan of a battalion into concrete orders at the platoon or weapon system level,

- taking into account the higher commander's intent in unexpected situations (with disturbed communication) and

- taking advantage of favorable developments of the situation without neglecting the overall plan of the superior command level.

Therefore, we advocate for a comprehensive modeling of both lower and higher command levels (at least battalion) C^2-processes within one CSS. The nexus of all these demands is flexibility: it should be possible to replace human decision makers (man in the loop) with C^2-modules wherever you like in the simulation. This attribute is essential especially when CSS are applied within combat training exercises.

Facing these challenges we have developed a new architecture for combat simulation systems for the last four

years, that enables us to design command & control modules for each command level. This architecture is founded on the following concepts:

- separation (as far as possible) of the elementary combat processes (movement, reconnaissance, attrition, etc.) from the C^2-processes (**central simulator vs. C^2-processes concept**),

- design of specific tactical languages for every command level and for different branches and staff functions to describe the set of options provided by the system (**multilayer tactical language concept**),

- development of command & control modules for every command level and branch based on the corresponding tactical languages (**concept of the tailor-made C^2-modules**),

- division of the types of combat and the general battlefield tasks into specific phases to reduce complexity (**phase concept**),

- assignment of (a limited number of) options to each phase, which are, in a first approach, given to the system and later on generated automatically (**option concept**),

- evaluation of these options and missions with a **generalized utility theory approach**,

- implementation of a special function to recall superior command automatons with restricted information in order to model "actions in accordance with the higher commander's intent" after failure of the communication system (**recall-function concept**, for more details see [Hofmann, Hofmann 98]),

- implementation of an **exception-interruption concept** to model a second aspect of mission type tactics: the exploitation of favorable situations, and

- strongly **object oriented programming.**.

3.2.2 The Multilayer Tactical Language Concept in Detail

In *closed combat simulation systems* every tactical language (developed for a certain echelon and a certain branch) can be defined as the set of instructions, that is used to conduct the course of the battle at the corresponding level. These instructions could be simple orders, missions or even (graphical) operation plans. Thus in closed combat simulation systems the tactical language defines exactly the interface between a given echelon and its superior C^2-automaton.

In *interactive combat simulation* systems the tactical language consists of the menu of instructions, which is at the human decision maker's disposal.

Ultimate objective of the programming of a general simulation system is the exact correspondence of the "human" and the "computer" tactical language, since this is the paramount precondition for the employment of a CSS for command post exercises on different echelons without additional (service) personal. Otherwise the exchange of humans and C^2-automaton within the system in justifiable

time would be impossible. Figure 3.1 shows the corresponding interface between two command levels according to the multilayer tactical language concept. (X, Y and Z designate a declining order of echelons, for instance: brigade, battalion and company.)
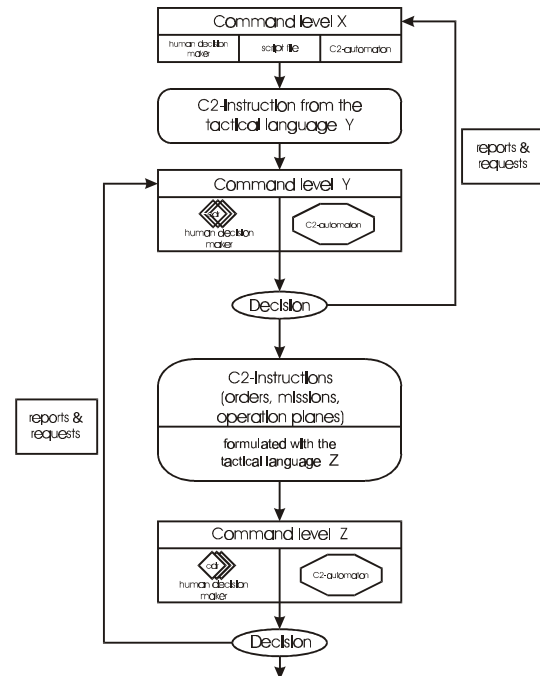
Figure 3.1: Transformation of Tactical Decisions into Concrete Instructions

This concept forces both human decision makers and C^2-automatons to formulate their *instructions* (sets of orders and missions) to subordinate units with expressions being part of the corresponding tactical language.

In general, human commanders will first make a decision and afterwards translate it into a sequence of instructions. Within an automaton the processing of the data can also lead directly to concrete instructions skipping an explicit decision.

At the lowest echelon modeled in the CSS, the decision will be transformed into a set of elementary orders, thereby controlling the elementary combat objects (platoons, weapon systems) of the central simulator.

Following this procedure, the operation plan of a unit, for instance a brigade, will be transformed into more detailed orders step by step, taking account of the capabilities and competencies of the respective echelons and finally deriving instructions to command the elementary combat objects.

The performance of such a multilayer tactical language system depends mainly on the scope of the different languages. Hence to improve the system their extension is a supreme task. Since human decision makers (military leaders) participate in the interactive version of the CSS it is also advisable to carry out this extension in a way

approaching the usual military custom.

3.2.3 The Concept of Tailor-made C²-automatons

In general, the C²-automatons for the closed version of the CSS cannot be designed before the corresponding tactical languages are developed. This is due to the fact that the automatons are tailored for the languages: Most of the modules of an automaton are created to perform the transition of instructions from the higher level tactical language into (more detailed) expressions of the lower level tactical language. Usually, to realize this transition a certain amount of supporting modules must be implemented. These modules include part of the tactical knowledge of a human decision maker. With no doubt, this is the most crucial step in the whole development of the simulation system, which can only be done by means of gradually improved prototypes.

To the extent the tactical languages differ from each other, the automatons will differ too. In fact, a priori there are no constraints at all for the architecture of any automaton (like, for example, a general scheme of the military decision making process); the design and implementation of the automatons depend only on the requirements of the languages.

This is why we call this approach the concept of tailor-made C²-automatons. It provides us with the ability to design and implement a variety of different C²-modules *within one CSS*. Therefore, such a system can be considered as a **general test environment for the development of C²-automatons.**

3.2.4 Modeling Command & Control on the Battalion Level

In the following a view is given on the basic modules of a decision making automaton at the battalion level. Although it can only be a first sketch, we think that this list gives a helpful guideline. What we need is:

- a mission analysis module (considerably simplified by the tactical language concept, since it only searches for key words to trigger the assessment modules),

- an intelligence estimation module which builds up the perceived situation and must finally encompass a comprehensive enemy situation and threat assessment,

- a module to analyze and project the own tactical situation,

- a module to analyze and project the own logistical situation,

- a terrain assessment module,

- a module that administers pre-developed options (courses of action) or generates these options by itself according to the different phases of combat,

and

- an explicit decision making module, which connects all the information provided by the evaluation modules with the different options, weighs these options up and eventually chooses one of them for a decision.

Figure 3.2 shows how these modules can constitute a prototype of a battalion C²-automaton.
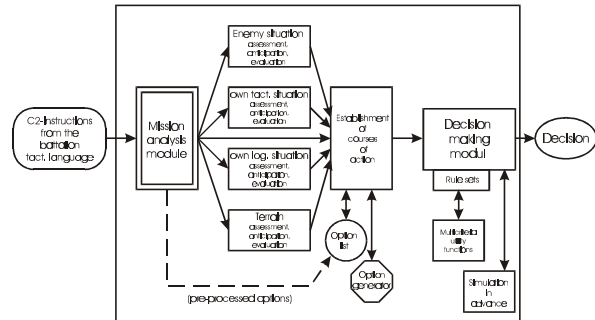
Figure 3.2: Possible Architecture for a Battalion C²-automaton

Of course this is only one exemplary solution, but it certainly comprises the main elements necessary to model military decision making at the battalion echelon in general.

3.2.5 The Transformation of the Battalions Commander's Intent on the Company Level

The Company Tactical Language

The company tactical language (CTL) mainly serves to translate the battalion commander's decision into company level instructions. Thus the question is: Which aspects of combat are usually performed at the company level, respectively: What are the corresponding orders to transform the commander's decision?

Above all the company is responsible for the *immediate coordination* of fire and movement - even when facing enemy fire and difficult terrain - and all the arrangements to perform it. Consequently, the company tactical language must allow the battalion commander to set the stage for this coordination which means that the CTL will mostly comprise the following instructions:

- a concrete *movement order*, if needed specified with
  - an objective,
  - if need to be: intermediate objectives,
  - a line of movement,
  - a velocity,
  - an exact time to start the movement
  - an order and maybe
  - a formation,

- a *fire control order* (addressing chiefly the clearance of fires) and

- a *communication order* (inevitable for the reward passage of lines).

A next step to extend the scope of a general CTL surely includes:

- an *emplacement order*, subdivided in
  - a detailed fighting position order and
  - an order to occupy a battle area,

- a set of *logistical orders* to organize supply and maintenance,

- an *order to attach/detach platoons*,

- an *order to allot sectors of observation and fire* and

- a set of *orders to command the combat of dismounted <u>and</u> mounted fighting forces.*

With this set of orders it should be possible to "translate" most of the battalion commander's decision into concrete instructions for the company.

Realization of the Company Tactical Language Instructions within the Company C²-automaton

As mentioned before decision making at the company level differs markedly from its higher level counterparts. Instead of being driven by a general information processing leading to a choice among different options, low level command resembles frequently a permanent adaptation to the current situation, using proven actions and arrangements.

Therefore, a straightforward processing of the battalion orders is seldom feasible. A very instructive example for this difficulty is the tactical movement appropriate to terrain and situation. Without a notion of the combat formations (wedge, column, Vee-formation, etc.) and the possibilities to advance (by bounds, by echelon, leapfrogging or accordion-like) any C²-automaton will fail to produce reasonable commands for the platoon level.

In order to solve this problem we endowed the company automaton with a set of supporting modules, reflecting exactly this kind of skill and knowledge. Since calling these modules is quite similar to calling the whole automaton with tactical language instructions, we have named the totality of these modules the *company tactical realization language* (CTRL). It is essential to notice that the orders triggering these support modules are not part of the company tactical language (CTL), which implies that they cannot be used at the battalion level to specify the battalion commander's decision. In fact, they are not accessible for him at all (see Figure 3.3). Notwithstanding this constraint the object-oriented design of the simulation system permits to reuse most of the elements of the tactical realization language in different automatons.
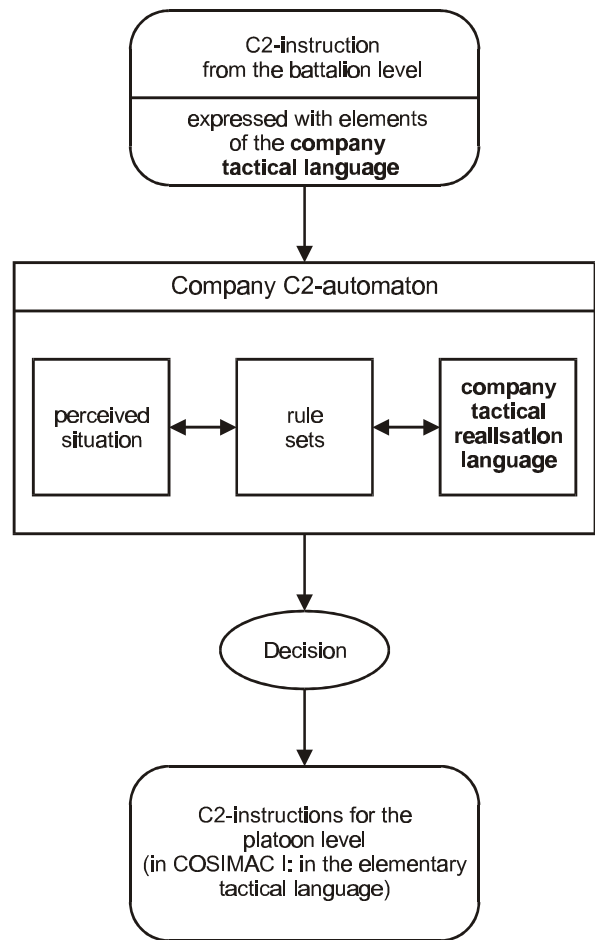
Figure 3.3: Company Tactical Language and Company Tactical Realization Language

4 Main Results, Conclusions and Future Developments

One of the main efforts is the development of robust and highly efficient rule sets for combat simulation systems (CSS) with automated control. For higher command levels we realized this task with the CSS KOSMOS. During the last five years we focused our attention to lower echelons and designed and implemented the COSIMAC models with a completely new object oriented architecture in order to evaluate different command & control automatons within one simulation system.

Among others, we have

- developed a concept for the description of tactical/operational intentions and concepts of operation with a battle management language depending on the different command & control levels and branches or functional staff areas,

- developed and implemented an architecture for the design of C²-modules to break down these concepts into individual mission and battle orders for the subordinate combat and combat support units,

- developed and implemented some detailed modules for route planning, terrain assessment for defense and attack operations, and some spartial and procedural templates on battalion, company and platoon level,

- implemented a general algorithm based on a multi-dimensional utility theory approach for solving the general allocation problem of fire and forces in order to come on robust and highly efficient solutions, and to reduce the complexity and the number of "if - then - else" decision rules.

Altogether, we come to the conclusion that the development of C^2-modules at different command levels within one combat simulation system seems to be a feasible task; even sophisticated aspects of mission-type-tactics like

- acting in accordance with the higher commander's intent or

- the exploitation of favorable unanticipated situations

are not out of reach of modern combat simulation systems with automated control.

One of the main problems of this approach is not only the complexity of the problems to be solved, which reveals the limits of what seems possible today. In our view, another major problem simply consists in the enormous amount of work, which leads to the real limits.

But the project is going on. Our MoD was pleased with the concept and decided to support the further development of COSIMAC as a research and study project, that concentrates on command decision modeling for different combat modes on the battalion, company, platoon and single weapon system level.

In the long run we think about a <u>PC-based training or a risk evaluation and decision support tool</u>, by means of which officers may test a range of tactical options in assumed scenarios in training or real combat situations providing them a better understanding of the regarded tactical/operational situation and assisting their decision making. More details on the project are documented in [Hofmann, Hofmann 98] and [Hofmann 00].

References

[FM 100-5] Department of the Army: Operations. Washington, DC, June 1993

[Foulds 92] Foulds, L. R.: Graph Theory Applications. Springer-Verlag, New York, 1992

[HDv 231/100] BMVg, FüH I 6: Das Panzergrenadierbataillon. Bonn, März 1988, VS-NfD

[Hofmann et al. 86] Hofmann, H.W., Huber, R.K., Steiger, K.: On Reactive Defense Options - A Comparative Systems Analysis of Alternatives for the Initial Defense Against the First Strategic Echelon of the Warsaw Pact in Central Europe. In Huber (Ed.): Modeling and Analysis of Conventional Defense in Europe (pp. 97 - 140). New York, 1986

[Hofmann et al. 92] Hofmann, H.W., Rochel,T., Schnurer, R., Tolk, A.: KOSMOS - Ein Gefechtssimulationsmodell auf Korps-/Armee-Ebene (Version 3.0). Band 1: Beschreibung des Gefechtsmodells. IASFOR-Report Nr. S-9208, Fakultät für Informatik, Universität der Bundeswehr München, Neubiberg, 1992

[Hofmann et al. 95] Hofmann, H.W., Schnurer R., Tolk, A.: Kosmos Simulation Experiments on Stable Defense. In: Christensen T. (Ed.): Stable Defense - Final Report. Appendix 3 to Annex IV to Technical Report AC/243 (Panel 7) TR/5. NATO RSG 18, Brüssel, 1995

[Hofmann, Hofmann 98] Hofmann, H.W., Hofmann, M.: Formal Description, Modeling and Simulation of Tactical/Operational Intentions and Concepts of Operation - Final Report. IASFOR-Report Nr. S-9803, Fakultät für Informatik, Universität der Bundeswehr München, Neubiberg, September 1998

[Hofmann 00] Hofmann, M.: Zur Abbildung von Führungsprozessen in hochauflösenden Gefechtssimulationssystemen. Dissertation, Fakultät für Informatik, Universität der Bundeswehr München, Neubiberg, 2000

[Rochel 90] Rochel,T.: Zur Architektur geschlossener Gefechtssimulationsmodelle höherer Abbildungsebene unter besonderer Berücksichtigung der Modellierung und Implementierung von Führungsprozessen. Dissertation, Fakultät für Informatik, Universität der Bundeswehr München, Neubiberg, 1990

[Schaub 91] Schaub, T.: Zur Aggregation heterogener Abnutzungsprozesse in Gefechtssimulationsmodellen. Dissertation, Fakultät für Informatik, Universität der Bundeswehr München, Neubiberg, 1991

[Schnurer 96] Schnurer, R.: Zur Abbildung von Führungsprozessen in geschlossenen Gefechtssimulationsmodellen. Dissertation, Fakultät für Informatik, Universität der Bundeswehr München, Neubiberg, 1996

[Tolk 95] Tolk, A.: Zur Reduktion struktureller Varianzen - Einsatz von KI in geschlossenen Gefechtssimulationsmodellen. Dissertation, Fakultät für Informatik, Universität der Bundeswehr München, Neubiberg, 1995

[TRUPPENPRAXIS 9/97] Taktikzentrum des Heeres: Damit die Zeit nicht davonläuft (Teil1) - Beurteilung der Lage und Entschluß anhand von Leitfragen. In: TRUPPENPRAXIS 9/97

The Czech Approach in the Development of a NATO Interoperable Ground Forces Tactical Command and Control System

Milan Šnajder, Jaroslav Horák
Military Technical Institute of Electronics
Pod Vodovodem 2, Prague,
Czech Republic
msnajder@vtue.cz, jhorak@vtue.cz

Václav Jindra, Ladislav Nesrsta
DelINFO, s.r.o.
Chodská 15, Brno
Czech Republic
vjindra@delinfo.cz, lnesrsta@delinfo.cz

Abstract - This paper describes systems engineering and system architecture design and development of the Ground Forces Tactical Command and Control System (GF-TCCS) of the Army of the Czech Republic. The design objective of the GF-TCCS is to provide automation support to commanders and their staff, based on the mission and phase of operations. The objective system will use a high proportion of commercial-off-the-shelf networking software, GIS products and government-off-the-shelf equipment, including military mobile radios and switches, tactical platforms (e.g. trucks, containers, armored personal carriers).

Introduction

At present the Army of the Czech Republic (ACR) doesn´t have an integrated, automated ground forces tactical command and control system. Commanders and staffs generally perform their mission using a manual system, augmented by some commercially available hardware and software systems. Some automation and communications systems operated in mutually isolated manner do not provide the mobility, functional flexibility, security, survivability, and interoperability required for ACR's GF-TCCS.

At the and of 1997, the Military Technical Institute of Electronic (MTIE), Prague was selected to provide a system integrator´s mission for the GF-TCCS of the ACR programme. For technical advice, the ACR called on the U.S. Army´s systems engineers for digitization, the MITRE Corporation.

Objective

The MTIE objective, in partnering with MITRE, was to jump-start common effort at developing a tactical command and control (C^2) system for Czech ground forces - by bringing to bear on the ACR´s requirements MITRE´s expertise in C^2 and MITRE´s years of experiences in engineering the U.S. Army´s tactical C^2 systems.

Operational requirements

GF-TCCS has to provide seamless connectivity from the lower tactical (squad/mobile plaform, platoon) level to the Operational Commands (Ground Forces Command and Territorial Forces Command). GF-TCCS will be used regularly within garrison, during deployment, and in the field to maintain the soldier's proficiency at the level required to respond to the broad range of potential missions.

GF-TCCS vertically and horizontally integrates information from the squad/mobile plaform to operational commands level. This requires GF-TCCS to fully comply with the seamless data architecture described in the Staff Information Sytem (SIS) ACR concept.

The GF-TCCS operational capabilities will allow the commander and staff to:

- Collect, process and organize large amount of battle information.
- Combine information from multiple sources to create more complete and useful information.
- Process information to analyze trends, detect unusual activities, or predict a future situation.
- Develop courses of action based on situational factors.

- Exchange information efficiently among and within command posts on the battlefield.
- Present information as graphic displays and textual summaries.

Fundamental to the GF-TCCS operational concept and relevant *common tactical picture* is a single entry, near-real-time information, and automated interoperability between each battlefield information system. The GF-TCCS must provide sufficient interoperability that data entered at any node in the architecture is distributed to all other nodes requiring that data without the need to copy or re-enter data. The elapsed time from initial data entry to receipt at other nodes in the architecture shall be consistent with the needs of the operational mission being supported by the data.

System architecture

The GF-TCCS is the integration of five plus one plus three subsystems: five plus one battlefield functional area command and control systems plus three supporting tactical communications and management systems commonly exploited by all six C^2 systems.

Five plus one battlefield functional area C^2 systems provides situational information and decision support to commanders and staffs in the execution of the operational/tactical battle at operational echelon (operational group) and below. Within this integration of systems, the force level database first takes form at the battalion (or battalion task force) to meet the tactical commanders' requirements for common battlefield picture and situational awareness. The GF-TCCS command and control subsystems are heavily oriented toward combat operations.

The five plus one GF-TCCS command and control subsystems are linked by Tactical Area Communications System and by the Combat Net Radio System. Combat forces, weapon systems and battlefield automated systems will be supported by the Integrated Management and Control System that will provide managing of the tactical communications.

GF-TCCS will be linked directly to the SIS ACR, providing the framework for seamless connectivity from the battalion to the General Staff echelons. Objectively, its five tactical C2 subsystems will merge into a single, coherent, interoperable system binding the combined arms battlefield operating systems together within a unifying open system environment (OSE).

The five plus one plus three GF-TCCS's subsystems as shown in Figure 1.
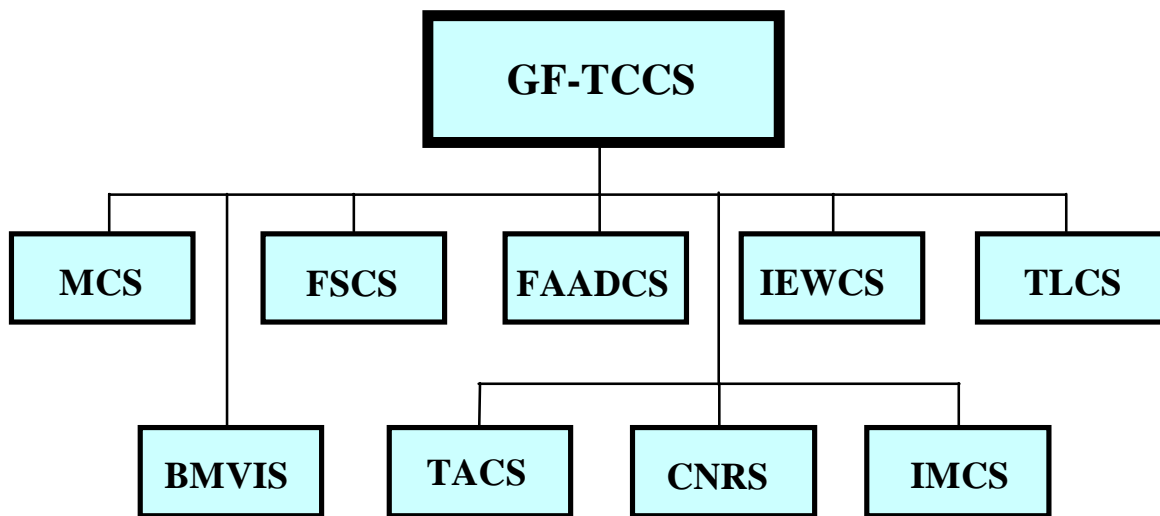
Figure 1: Subsystems of the GF-TCCS

MCS   -Maneuver Control System
FSCS   -Fire Support Control System
FAADCS -Forward Area Air Defence Control System
IEWCS   -Intelligence and Electronic Warfare Control System
TLCS   -Tactical Logistics Control System

BMVIS   - Battle Management Vehicular Information System

TACS   -Tactical Area Communications System
CNRS   - Combat Net Radio System
IMCS   - Integrated Management and Control System

4		***Common/Unique Applications***	Maneuver control Engineer support planning Fire support planning Chemical support planning Forward area air defence control Medical support planning Intelligence and EW control Other special tasks Logistics support planning
3	O S E	***Common Support Software Modules***	Office automation Simulations Convoy planning (MS Office) Friendly situation and control Database management/ Enemy Situation Other common tasks administration Operational plan, (prognosing of losses, File management operational order resources completing, Message handling Terrain and weather combat readiness Multimedia support effects evaluation recovery etc.) CP's LAN administration Human (soldier) Security management resources evaluation Alert and warning Supply/equipment services status
2		***Open Architecture Software***	Operating system Graphical User Interface Database Management Other COTS system software products
1		***Common Hardware Suite***	Computers (servers, desktop and notebook PCs, handheld machines) Additional functional hardware modules Peripheral equipment

Figure 2: Layered GF-TCCS Architecture

The common and unique applications (not part of OSE) will be products embodying specific functions such as movement control, terrain evaluation, operation plan/operation order, etc. The requirements are derived from the various subsystem operational programs. These products will be developed through a rapid prototyping developmental strategy which envisions an incremental, iterative build process, involving close coordination among the user and combat and system developers.

GF-TCCS will use, via communication protocols in the OSE, prepared new ACR's Tactical Area Communications System and Combat Net Radio System. It shall also be able to use strategic(static) MoD/ACR's Common User Communications Network/Army Data Network (CUCN/ADN).

Example of the architectural approach of the MCS

Maneuver Control System (MCS), as a core tactical forces information system, provides commanders and staffs with the capability to collect, coordinate, and act on near-real-time battlefield information. This allows the commander easy access to information, access to display current situation reports, that assess enemy strength, weakesses, movement, and the status of friendly forces. The MCS also aids the

battle staff in rapidly disseminating the commander's orders.

Through the MCS, the commander transmits critical battlefield information, courses of action, schemes of maneuver, warning orders, operation orders, priorities, intelligence requests, and air operations requests. The MCS helps the commander maximize combat power at the appropriate time and place, respond to threats, and anticipate a developing situation.

From battalion through operational commands (or their operational groups), the rapid exchange of information through the MCS gives all command posts the same picture of battlefield. This, along with the capability to query both local and remote databases, helps commanders to synchronize the battle. Commanders at these echelons can make decisions supporting mesh with the decisions and capabilities of other commanders.

Commanders of the tactical forces will use the combat support elements as force multipliers to enhance the combat power of his maneuver units. Digitally equipped combat support elements will use enhanced decision aids and increased situational awareness provided by digital MCS means. Table overview of the MCS major applications:

Application	Functions
Unit Task Organization (UTO)	UTO forms for: UTOs, UTO management, new UTO and copy UTO
Reports	Generate different reports of unit readiness
Electronic Maps and Overlays	Displays maps and associated overlays
Formatted Messages	Creates, edits, and transmits standardized formatted messages
Operation Orders	Creates, edits, transmits and authenticates operation orders (OPORDs) and operation plans (OPLANs)
Synchronization Matrix	Graphic display of unit missions as they relate in time. Can be used to develop Courses of Action (COA).
Command and Control (C2) Products	The file manager functions provide interfaces for the C2 product windows, for example: InBox messages, unread reports, archives files, etc.

Battle Management Vehicular Information System (BMVIS) is a core battalion/brigade and below information system and a key element in the effort to digitize the battlefield.
As organizational/technical arrangement of GF-TCCS computer/communications (CNRS) components at lowest levels inside battle mobile platforms is intended to provide automated capabilities providing time, form and place utility to critical combat information.

The BMVIS will provide the user to access all information collected by sensors in and around his combat vehicle by integrating the information into a single source. Data such as vehicle position, targeting data, chemical contaminants and range to targets will be integrated into the BMVIS situation map and reports which are provided to the commander/user through the interactive display and are transmitted via radio (CNRS) data transmission to all or selected BMVIS in the battalion task force. At the later stages sensors can report on-board logistics (such as fuel, ammunition).

The BMVIS will be a typically MCS-type C2 componets arrangement and will have automated/ programmable interfaces with the co-operating systems of the battlefield functional areas. The BMVIS will also interface with the some NATO

armies automated vehicular C2 systems such as US Army FBCB2/IVIS and FRG IFIS.

Evolutionary systems development and rapid prototyping

The GF TCCS was specified as system of systems (subsystems). Is almost impossible from different view of point to develop all systems together in one time. We took experiences of world´s advanced development organizations and we accomodate in GF-TCCS project an evolutionary systems development approach.

This development methodology is based on two main fundamentals:
1. **Incremental development.** Both whole complex and each subsystem is (and will be) developed in a sequence of increments. From the first increment, each of them is fully operable and fully integrated with all preceding increments
2. **Rapid prototyping**. Every planned element and prepared systems is prepared like a prototype and consecutively tested in development labs, in special testbed and in the field as well.

For all teams, that will participate in development process was defined one mandatory methodological standard. The unique methodological basis for GF TCCS is RUP – Rational Unified Process (by Rational Rose). Rational Rose modeling tools are close linked with this methodology and are fully acceptable, because of object orientated system architecture.

Rapid prototyping together with experimentation provides an effective tools for resolving issues, experimental data collection and reducing risk early, and the determining the adequacy of requirements, design, and new GF-TCCS's system capabilities before committing major resources.

The attributes of the proposed approach to evolutionary GF-TCCS development include:

- use of software development environments/tools for rapid prototyping of functionality
- object-oriented design that allow rapid integration of COTS software
- NATO OSE software standards and profiles practices

- documentation appropriate for expansion into formal specifications and
- continuous interaction and feedback from the military end-users

GF- TCCS actual structure

In the year 1999, there were developed starting prototypes of two first subsystems - MCS and BMVIS. MCS is prepared as a base environment for all other subsystem, is oriented to higher tactical and operational-tactical echelons (battalion, brigade, …). BMVIS is more linked with lower tactical echelons (company, platoon, …), but some common elements will be used in mobile platforms of higher echelons as well. In the detail analysis phase of MCS and BMVIS, **four main application elements** were specified – TAGIS, ELMET, OTS and FBD.

TAGIS – Tactical Geographical Information System

TAGIS is a proprietary solution of the integrated, geographic and information system designed for the operational and tactical use. It enables the all-round use of the drawing and editing possibilities over the electronic map, which is based on standard graphic data formats (JPEG, VMAP, DTED, CADRG, common used ESRI formats). It is developed especially for the work in the military environment and for the military users (Fig. 3). TAGIS was designed like a multi-purpose system, which can be used both separately and in the computer network operation. It is working over the common database of TCCS objects. It is one of the main supporting means of most of the software applications included into TCCS. It can be also used for supporting the other software applications that respect the defined data interfaces.

In addition to the standard properties of the commercial GIS (graphic information system) TAGIS enables above all:

- The full-value entry of the operational and tactical situations into map with the use of all objects included into these situations in accordance with the standards of the Czech Army and NATO
- The work with the operational and tactical marks and objects according to NATO/APP-6 standards. Creating and editing the new

operational and tactical marks including their text description.

• Interconnecting the plotting of the operational and tactical situation with the database of the forces and means of the friendly troops and the hostile troops. This feature enables to fully support most of the operational and tactical solutions (hereinafter referred to as OTS).

• interconnecting the plotting of the operational and tactical situation and the database of the forces and means with the combat documents.

• The work with the raster and vector data at respecting most of the normally used data interfaces - the raster map basis, the digital model of the terrain and the relief of the terrain, the electronic products of the terrain analysis, the digitized aerial pictures etc.
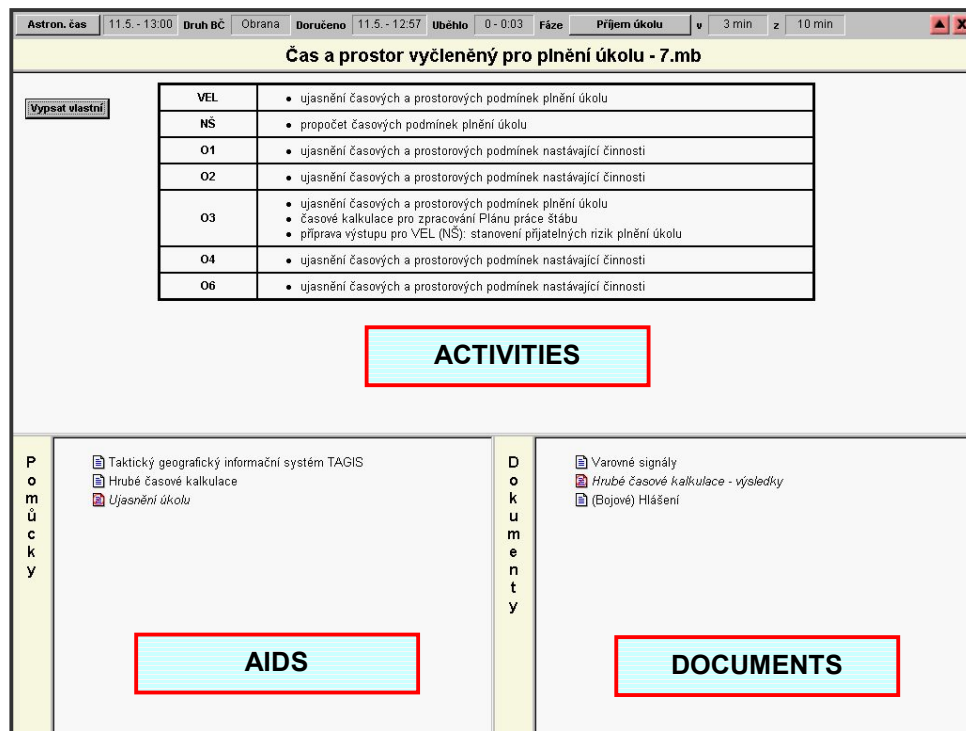
Figure 3

ELMET – Electronic Methodology of Decision Making Process

ELMET is an author's solution of the applied software system which is designed for the control and co-ordination of the all-round support of the decision making process of the commander and the staff from the side of the operational and tactical system (hereinafter referred to as OTS) in the process of planning and conducting the combat activities and as well during performing "out of combat" activities. Actual solution stage covers all main sorts of DMP - Deliberate Decision-Making, Combat Decision-Making, Quick Decision-Making.

ELMET works in the environment of the local computer network of the staff. It is designed as a multi-purpose system and by having been loaded with the appropriate data it can be used not only by the ground forces of the Czech Army but also by the staffs of the Czech Air Force, the Territory Defense Force, the Logistics Headquarters and by the management authorities of the civil institutions dealing with the crisis management.
It is fully compatible with the other software applications of OTS above all with TAGIS application and it is closely co-operating with them. When respecting the defined data interfaces it can co-operate with the other software applications in addition to OTS.

Main properties of ELMET:

- ELMET is a GroupWare solution for the entire local network of the staff. The synchronization of the work at the working places under ELMET is carried out at the time when the chief of staff is formulating the task for the staff.
- Each phase of ELMET has the navigation screens with the standard structure and specific contents defined for the specific staff officer. The navigation screen contains three basic fields (Fig. 4):
- Field of activities – it defines the standard activities for the individual officers of the staff

- Field of aids – it contains the software applications to support the standard activities listed in the field of activities. The user can directly activate these applications by using the buttons
- Field of documents – it contains the list of the formalized documents available with the user during the standard operations defined in the field of activities. The documents can be activated in the same manner as the aids.
- The authorized officer may distribute, in case of need, the various kinds of signals in the staff network by means of ELMET. These signals are always displayed above the active software applications and the precise evidence of their sending out and confirmation is maintained.
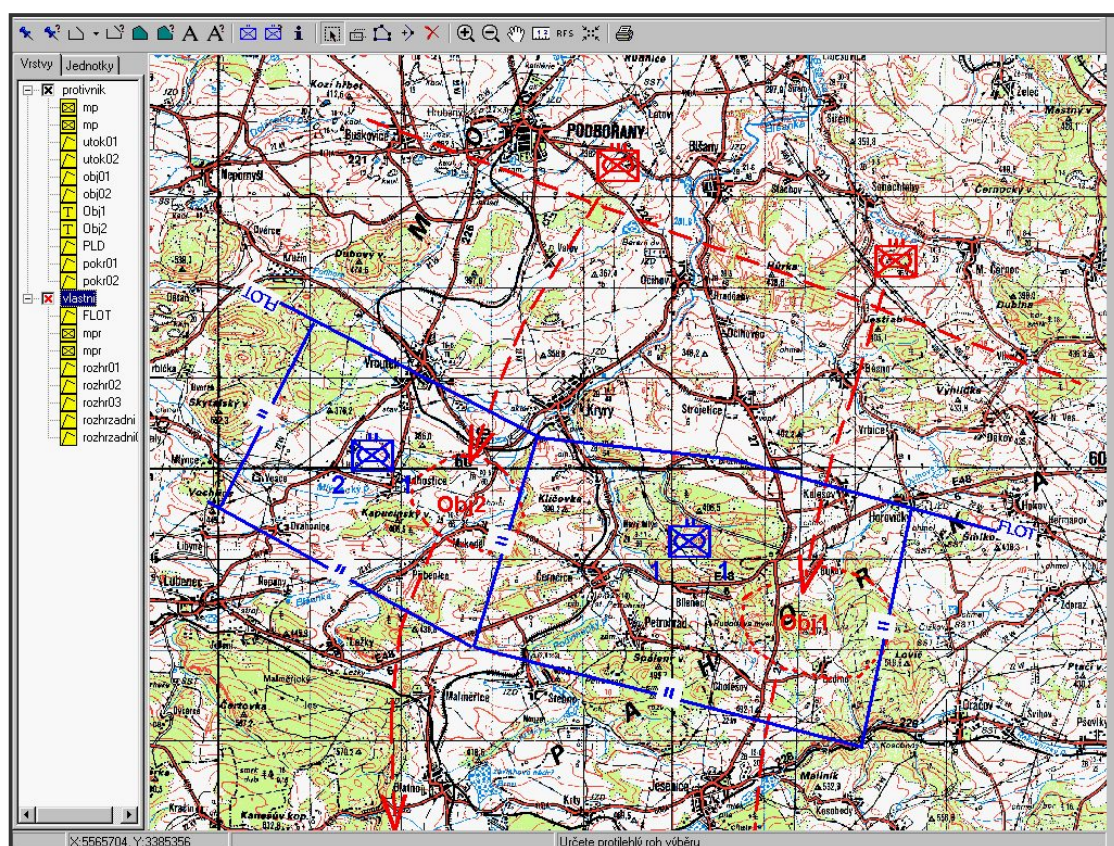
Figure 4

OTS – Operational-Tactical Solutions

This group of applications supports all commanders and staff officer tasks. Total amount of defined OTS has dozens of items. In actual solution stage OTS has about 10 specific applications for supporting typical commander and staff activities:

- Rough Time calculations
- Ratio of power
- Transport calculation
- Electronic methodology of the Topographic data (electronic maps) demand
- Optimum variant selection
- Chemical situation assessment (Fig. 5)

- Protection buildings calculation
- Prediction of a radio-relay communication
- Warning Order (WARNO) preparation
- Situation report (OWNSITREP)
- Distribution of Warning signals
- Management of the tactical database
- Management of applications and users

All of these applications are working over common tactical database and most of them are using TAGIS like a ground for data entry and output. All applications can be lunched from ELMET environment, some of them (non grouping applications) could be run like stand-alone applications.
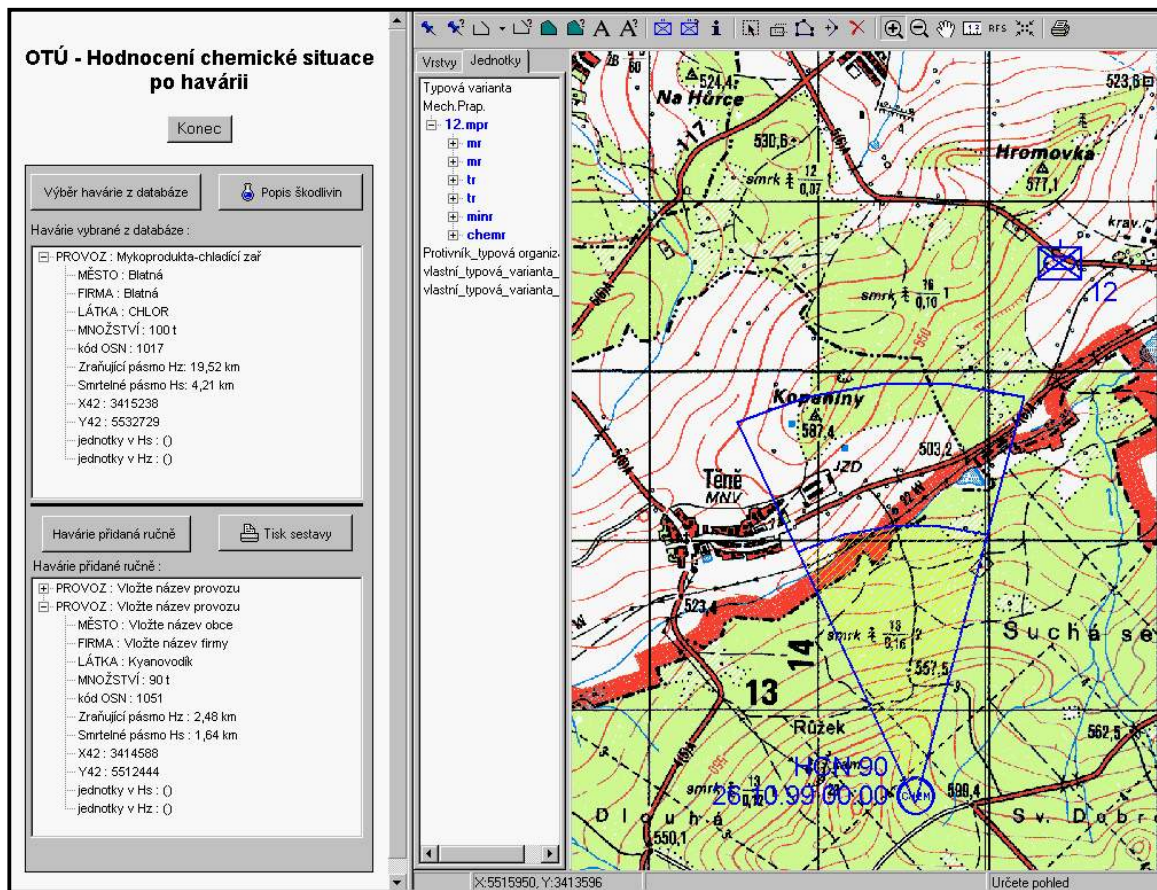
Figure 5

FBD – Formalized Battle Documentation

The application is designed for the group processing of the formalized combat documents. It is a base of the formalized messaging according to STANAG 2014, STANAG 2434 (APP-9), STANAG 5500 (Formets-AdatP3). The structure of the formalized documents being created ensures the operational interoperability of the combat documents according to STANAG 2014. The formalized document can be prepared in the Czech and English language. It ensures the operational interoperability with NATO with respect to the fact that it has the formalized

structure (Fig. 6). The full procedure interoperability with NATO in accordance with the AdatP3 standard will be ensured after completing the coding tables of message items.

The application has the following basic features:
- It enables the distributed processing of the documents by the authorized officers of the staff
- It has the direct data relationship to ELMET.
- It receives the data from the centralized database of TCCS.
- It is using the HTML technology, but it complies with the requirements of the standard XML.

Figure 6

Conclusion

A NATO - interoperable, tactical C2 systems is achievable, in the near term, using off-the-shelf software and hardware. The architecture of the systems will be net-based and state-of-the-art.

GF-TCCS will be the keystone of a future digitized battlefield, providing the commander an integrated digital information network that supports warfighting system and ensures command and control decision cycle strength.

This page has been deliberately left blank

_____

Page intentionnellement blanche

Principles and Application of Geographic Information Systems and Internet/Intranet Technology

Prof. Dr.-Ing. Wolfgang Reinhardt
Wolfgang.Reinhardt@unibw-muenchen.de
http://agis.bauv.unibw-muenchen.de/staff/reinhardt/home.htm

University of the Federal Armed Forces Munich
Institute for Geo Information and Land Development
Werner-Heisenberg-Weg 39
D-85577Neubiberg
Germany

key words: GIS, Internet/Intranet

Abstract

The paper presented consists of three main parts. In the first part we roughly outline the state of the art of Geographic Information Systems (GIS) mainly with respect to technology and data. Within this part we also give some examples where GIS is applied in military applications.

The second part gives an overview of architectures and some technical aspects of GIS-Internet/Intranet solutions, compares the different approaches and discusses the potential of this technology in general. Furthermore some examples demonstrate the practical use of GIS and Internet/Intranet. As the World Wide Web (WWW) gains more and more significance and there is a large demand of GIS applications in the Internet/Intranet we introduce the main principles of this technology. Especially we explain how Geographic Information Systems can be connected to the world wide web and which extensions are necessary to transfer and to view Geographic data. In this part we also show how the Virtual Reality Modelling Language (VRML) can be used in this field. Part of this is based on results of some projects conducted for the AMilGeo (Amt für Militärisches Geowesen) of the German Federal Armed Forces as well as on other civilian projects.

In the third part of the paper we demonstrate the potential of the GIS and Internet/Intranet technology for civilian and military applications. Besides we discuss the main advantages of GIS and Internet/Intranet, such as Ease of Use or the possibility to access up-to-date information in various databases. Furthermore we show how the connection of Internet and telecommunication can be used in GIS.

1 State of the art of Geographic Information Systems

Geographic Information Systems are Information Systems which are extended to handle geographic data which often also is called spatial data. That means a GIS in addition to the common IS functionality offers specific data types, data access methods and spatial data analysis methods.

Due to the fact that more than 80% of all data are somehow related to a geographical position GIS is used in very many fields today, civilian and military too.

Geographic Information Systems consist of hardware, software and data. As hardware platform today standard PCs are used. GIS software packages are offered from more than 500 companies world-wide, but about ten of the most important vendors share more than 60% of the market. Data is not only the most important part of a GIS but also the most costly factor. This is due to the generally high costs of data

acquisition. Figure 1 depicts the approximately cost relation of hardware, software and data which has been verified in various projects.

These days GIS data is available in many countries, often countrywide and in various forms.
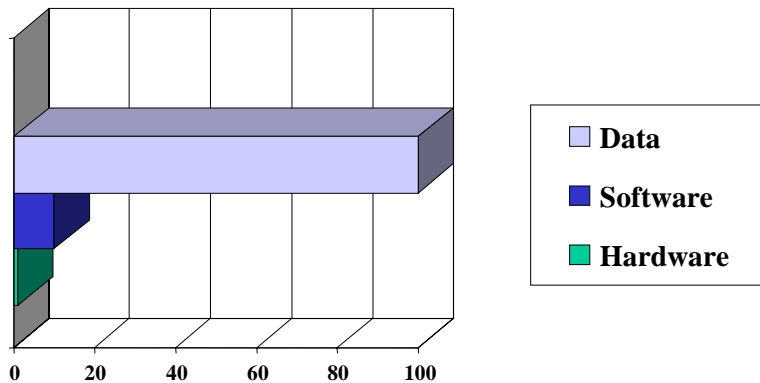
Figure 1: Cost relation of hardware, software and data in a GIS

1.1   Technology and trends

From a technical point of view we can state that state of the art GIS hardware and software are standard IT products. Some of the important points:

- GIS support the client / server architecture. Often in a GIS a standard data base is included. These data bases are mostly relational data bases. Object oriented data bases are used very seldom
- The programming languages used are mainly C++ or Java, for customising purposes also Visual Basic
- Internet/Intranet technology has become a key factor in GIS during the last years (for details see chapter 2). This technology enables the introduction of GIS based services which can be used in Intranets and in the Internet. For these services recently also eCommerce techniques and products have been utilised.
- Based on the connection of Internet and telecommunication new services have been introduced which can be used in principal from mobile phones and other mobile equipment (´mCommerce applications´). Mainly so called location based services are discussed widely nowadays. These services e.g. allow for queries to show the way to a point of interest (see figure 2)

More information to these items are given in the contributions in [Reinhardt, 2000]

Figure 2: Example of a location based service

1.2   Interoperability

As already mentioned there is a large number of data sets available in most countries of the world. These data sets are available mainly on state, county or community level as well as on enterprise level.

These GIS in general have been designed in a proprietary manner based on proprietary systems. Due to this fact data exchange or integration without loosing information is not easy. For this reason several activities have been started during the last decade which are aiming on an improvement of this situation. The most remarkable activities – from the point of view of the author – are the activities of the

- International Standardisation Organisation (ISO), TC 211 [http://www.statkart.no/isotc211/] and the
- Open GIS Consortium (OGC) [http://www.opengis.org]

The activities of the OGC are consequently following the idea of interoperability of geographic information. OGC and ISO are working together very closely (´first class liaison´). E.g. OGC uses ISO specifications (international standards, draft international standards) within their development process to produce implementation specifications.

Because of the strong impact of OGC work on GIS in general it shall be outlined roughly:

The Open GIS Consortium has been founded as a non profit organisation in 1994 in the US. The consortium now has more than 200 members. All relevant computer hardware, data base and GIS vendors as well as other IT companies, consulting companies, GIS data providers and users from all over the world are working in OGC. The aim is to allow for a interoperability of GIS data without any loss of information. For that goal abstract and implementation specifications are produced in working groups which are in general implemented by the vendors within their GIS software. OGC´s work of the last two years is strongly focused on the use of internet technology.

Roughly speaking the idea of interoperability can be reached in two ways:

- By direct access to distributed data bases (on-line) over common networks. For this purpose standard access methods are defined which are published on OGC´s homepage.
- By means of file transfers (off-line). For this purpose standard data description languages are defined and used respectively. With these languages data schemes and the data itself can be described. The file can be transferred via networks like the internet, of course.

The approach of the second way is outlined now shortly:

In this approach GIS features are described in a specific language, the OpenGIS Geography Markup Language (GML). This GML is based on the well known XML language of the w3 consortium [http://www.w3.org/]. More information on GML can be found at [https://feature.opengis.org/rfc11/GMLRFCV1_0.html]

With this approach the geographic data is transferred in ASCI files. The representation of a point is like the following:

```
<?xml version="1.0" standalone="yes" ?>
<!DOCTYPE Point (View Source for full doctype...)>
- <Point name="location" srsName="epsg:3567">
<CList>445.12,345.71</CList>
</Point>
```

For a graphical display in a standard browser one of the common graphics formats can be used, these are:

SVG (Scalable Vector Graphics, http://www.w3.org/Graphics/SVG/Overview.htm8)
VML (Vector Graphics Markup Language, http://www.w3.org/TR/NOTE-VML)
VRML (Virtual Reality Markup Language http://www.vrml.org/)
PNG (portable network graphic http://www.libpng.org/pub/png/)

If a GIS data server is able to deliver data in this GML format and a converter to one of the graphics formats mentioned is available the data of the server can be accessed from a client using a standard browser.

2 Geographic Information Systems and Internet Technology

2.1 Some remarks on web technology

Webbrowsers, like Netscape Communicator or Microsoft Internet Explorer, can nowadays be referred as standard equipment of a PC. With an appropriate network authorisation it is possible to connect to the WWW with a webbrowser, either by clicking on a link in an HTML[*] (Hypertext Markup Language) page or by typing a Uniform Resource Locator[**] (URL). The request is transferred to the addressed webserver through the Hypertext Transfer Protocol (HTTP Protocol). The according HTML-page is invoked on the server and transferred to the webbrowser where it is displayed. A detailed description of the WWW functionality is presented in (Assfalg et al., 1998).

The following sections present and discuss briefly different possibilities to link GIS to the WWW. These solutions run partly on the server and partly on the client side. The webbrowser is considered as a client, which sends requests to the webserver. First we present solutions particularly based on the server side and then on the client side.

2.2 GIS linkage to the WWW

Web applications in general follow a so called 3tier architecture. In our case we have a Geo data server a web or internet server and a web or internet client (browser) as depicted in figure 3.

Internet client   Internet server   Geo data server
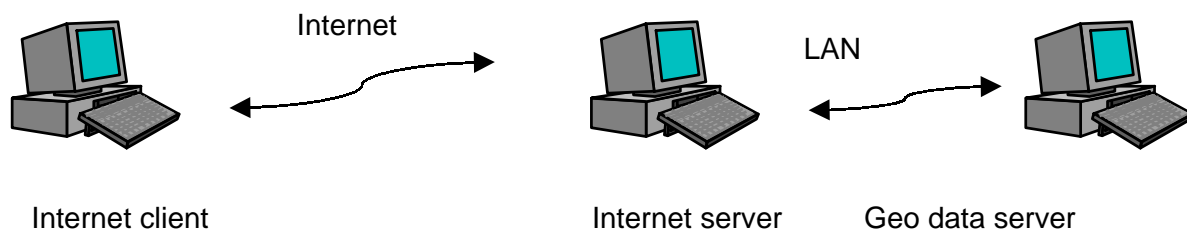
Figure 3: 3 tier architecture

---

[*]   HTML: Description resp. Edition language for WWW-pages. It consists of commands, which formats a document for the presentation with a Browser.

[**] URL: In WWW-Terminology the adress of a document is named URL. Normally that is the domain adress of the computer, which provides the requested document.

Figure 4 now shows schematically the communication between webbrowser, webserver and GIS server. On the webserver side there are basically five possibilities to realise the GIS-connection to the World Wide Web: Common Gateway Interface (CGI), Webserver Application Programming Interface (API), Active Server Pages (ASP), Java Server Pages (JSP) and Java-Servlets. The user on the client side does not need knowledge about the linkage of the IMS at the server side, but the system administrator respectively application developer should be familiar with these techniques.
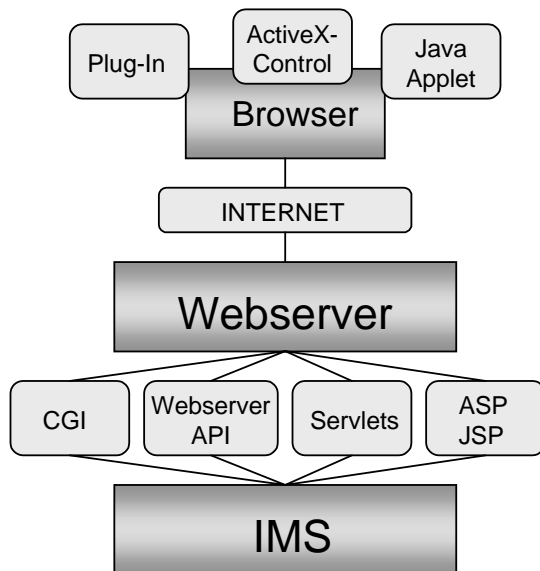
For a detailed discussion of these techniques please refer to [Reinhardt and Leukert, 2000]

On the client side: In general a webbrowser can handle HTML-documents and embedded raster images in the standard formats GIF, JPEG or PNG. To deal with other data formats like vector data, video clips or music files the browser's functionality has to be extended. This can be done by using e.g. plug-ins, activeX controls or Java applets. For a description of these techniques also refer to [Reinhardt and Leukert, 2000]

Figure 4. GIS linkage to the WWW

2.3   Type of transferred geo data

A decisive question for using GIS in the Internet is the form of data (vector or raster) which is used to transfer the data to the client. In principle it is possible to use raster as well as vector data. It should be mentioned that the transfer format is independent from the format the data is stored on the server. In most applications geo data is stored as structured objects in form of vector data in a proprietary format. For the data transmission to the client the map is converted in raster or a suitable vector format.

When raster data is transferred a standard webbrowser without extension can be used since webbrowsers can display GIF and JPEG. Only a kind of screenshot in form of a raster image is send to the client. That means the data on the server has to be converted to a raster format. The data volume is due to the known image size of X * Y pixels estimable and the original data on the server is safe as only an image is sent to the client. A disadvantage of using raster data is the lack of comfort of handling. Single objects cannot be highlighted by moving over them with the mouse. In addition a server contact is necessary per each request from the client but with a high performance infrastructure, e.g. Intranet, that does not cause problems.

Vector data can handled only in a standard webbrowser with extended functionality (e.g. plug-in). The user gets a more comfortable handling with vector data. For example single objects can be selected directly or highlighted. One more advantage using vector data is the possibility of local processing, it is not necessary to contact the server per executed browser action. Disadvantages of vector data are manufacturer dependence as well as changing data volume because the amount of data can vary depending of the selected area. Transferring vector data may endanger the copyright of the owner of the original data, since with tricks a user could store the transferred data locally.

Basically the presentation at the client can be realised with vector as well as raster data. The choice of the transferring data form should consider the application and the existing infrastructure. Software products which offer optional transferring of vector or raster data may provide advantages. They may allow a preselection with raster data and afterwards loading of the actual vector data with the possibility of subsequently processing locally.

Different consortia develop future standard formats for transferring data over the Internet as shown in chapter 1. These formats probably will be used in near future.

2.4   An example for 3D visualisation of the terrain using VRML and Java

2.4.1   Basic idea

The Virtual reality modelling language VRML has been developed by the internet community and is a well known possibility to generate 3d worlds which can be used for walk throughs in common browsers extend by plug-ins which are available for free. More information on VRML and its usage are given in [Koppers, 1998]. In several projects we applied VRML. One prototype application was the 3D visualisation of the terrain.

The basic idea of the application is pretty easy and will be explained in this chapter.

Starting from different kinds of terrain related data (which could be included optionally) like:

➢ Digital Terrain Models (grid or TIN)

➢ Digital Map or GIS data (vector form)

➢ Aerial photographs, Remote sensing data (raster form)

we developed some procedures to integrate the different data and transform it more or less automatically to VRML format.

We have chosen VRML because of the following advantages:

➢ It is in principle platform independent

➢ It uses standard COTS software and it can be used over the Internet in principle

➢ The standard Internet Browsers (Netscape, Microsoft) can be extended – e.g. by cosmo player – to interpret VRML data. Using this extended Internet browsers one can e.g. view the terrain from various points of view, walk through the terrain, and use a couple of other navigation modes.

➢ This is a low cost solution because it runs on standard PC´s with cheap/free of charge software

That means only the data integration and transformation step has to be done to be able to generate a 3D model of the terrain, to view it and to walk through it.

But for serious applications there is one considerable disadvantages: this solution doesn't provide an orientation where the ´walker´ is in the terrain and in what direction he is looking. How we could get writ of this lack will be described next.

2.4.2 The application

To provide a georeferenced orientation for the ´walker´ we developed an application using Java language with extends the VRML browser mainly by an orientation window which displays additionally a map in which the position of the walker and his viewing direction is displayed.

The application includes the following components:

➢ The VRML window

➢ The orientation window which in general includes a topographical map. This map shows the position and the viewing direction of the ´walker´

➢ The navigation board to control the ´walker´

This application always provides a synchronisation of walk in the 3d world and the display of the actual position and viewing direction. This is obtained by a communication of the internet browser and the Java applet.

The application supports different navigation modes:

➢ Walk mode, allows for a walk through the scene (left, right, forward, backward with constant height)

➢ Slide mode, allows for an incline of the viewing direction

➢ Examine mode, allows for movement on a sphere surface

➢ Point mode. Allows the movement to a specific position

A more detailed discussion of the application can be found in [Koppers, 1998]

3 Potential of Internet/Intranet technology for GIS

As already mentioned Internet/Intranet technology offers quite a number of possibilities to open up new fields of applications for GIS or to improve existing applications considerably.

Internet technology was first used to establish so called map servers, where internet users could download maps (jpeg or GIF images) for example for tourism purposes. In a second step these maps have been connected to a information, for example the location of hotels or other points of interest, which could be visualised in the map. Furthermore results of complex analysis could be visualised using this technology. Please refer to [http://www.esri.com] for examples.

During the last years in several countries applications have been developed which allow for a distribution of geographic data via the internet. Figure 5 shows an example view from the USGS homepage. There interest people can download relevant geographic data [http://www.usgs.gov].
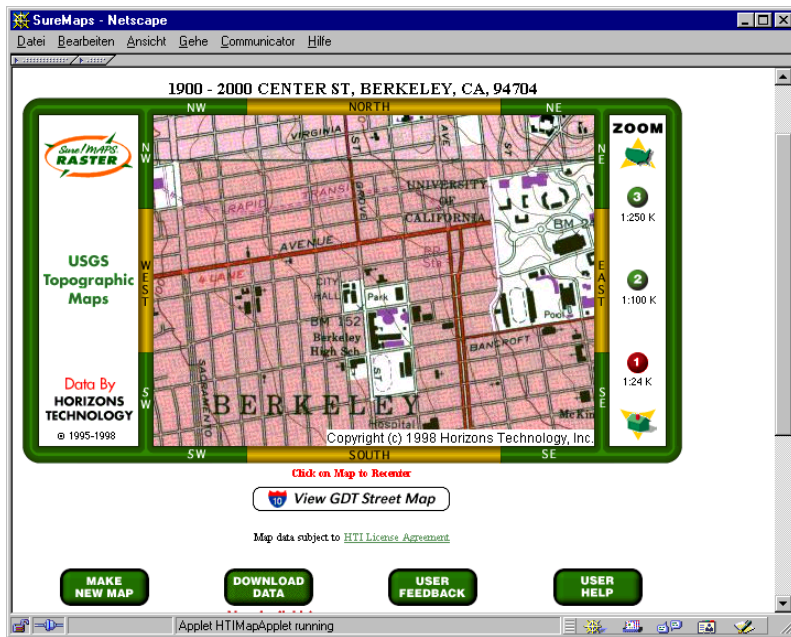
Figure 5: GIS and Internet example of USGS (screenshot)

Some of these application are using eCommerce techniques and products to allow for a more comfortable handling.

Implementations of Geographic Information Systems based on Internet/Intranet technology can be found in various fields and on very many places. In the following an example is presented which demonstrates how this techniques can be used in a meaningful manner.

This example comes from BASF enterprise (Fig. 6) and includes their main company site in Ludwigshafen / Germany which is larger than 7 km^2 and includes more than 2000 buildings, 115 km of roads and 211 km of railway tracks. The GIS application shall be used finally by the 20 000 Intranet users in Ludwigshafen and more than 10 000 Internet users world wide. All these people are just using a standard browser such as Internet Explorer or netscape communicator. Therefor the graphical data is transmitted in raster form (jpeg). It allows for queries for specific buildings, which can be searched by address and the result is presented alpha numerically and graphically. Furthermore queries concerning the location of roads / railroads under construction and closed roads are supported.

At some places such applications are used at the entrance of an enterprise to be able to produce specific maps for visitors with the route they have to follow to reach their destination on it.
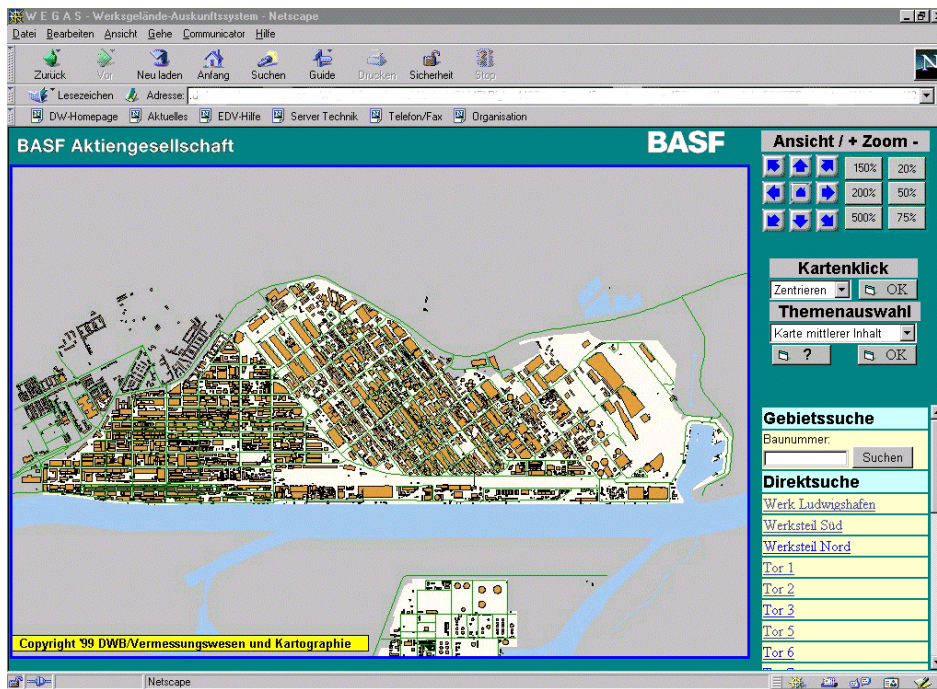
Figure 6: Screenshot from an example of an internet based GIS application

Recently there are also GIS based and Internet services available which can be used also from mobile phones and other mobile equipment (in combination with positioning equipment such as GPS) via the Wireless Application Protocol (WAP). Figure 7 shows the principle architecture of such applications. For details refer to the contributions in [Reinhardt, 2000] or to [http://www.mogid.com].

4   Conclusion

In this paper the principles of GIS and the Internet/Intranet technologies have been outlined. By some examples also the potential of this technology has been demonstrated. The main advantages of applications based on GIS and Internet/Intranet are:

- Internet / Intranet nowadays is a standard technology which is widely available and commonly used.
- Based on this technology applications can be provided which are easy to use also for non GIS experts.
- Via Internet or Intranet many users can access actual data stored in data servers.
- Information stored on different servers can be linked together easily using the hyperlink technique.
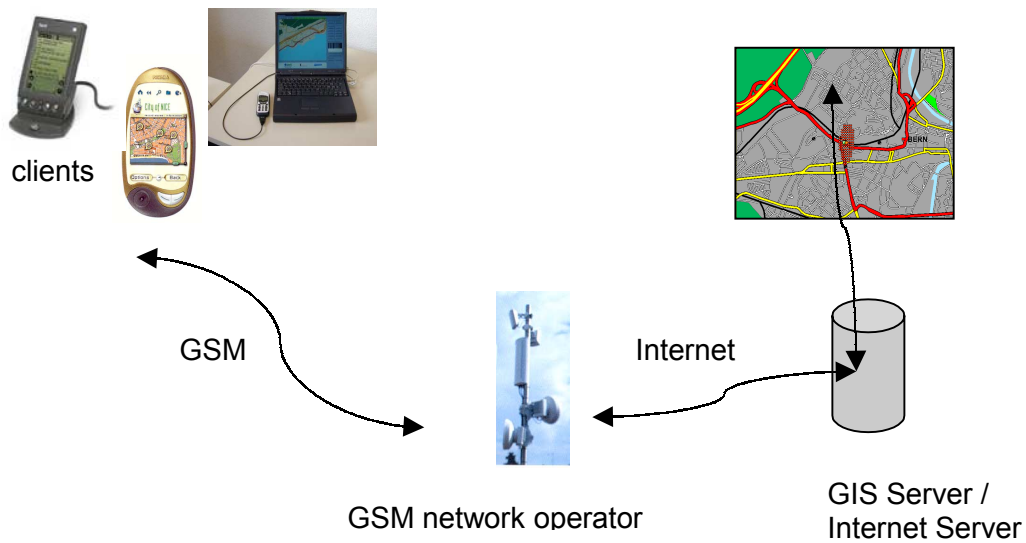
Figure 7: Principle architecture of GIS based mobile services

Literature

Assfalg, Goebels, Welter 1998: Internet Datenbanken - Konzepte, Methoden, Werkzeuge, Addison-Wesley

Koppers, 1998: 3D-Map - Virtual reality and Geodata, EOGEO ´98, Salzburg

Leukert, K. and Reinhardt, W,: GIS Internet Architecture, Paper presented at the XV ISPRS Congress, Amsterdam 2000

Reinhardt, W. (editor), 2000: GIS and Internet/Intranet technology – proceedings of a seminar held at university FAF (in German)

Assessing Survivability Using Software Fault Injection

Jeffrey Voas

Reliable Software Technologies

21351 Ridgetop Circle, #400

Dulles, VA 20166

jmvoas@rstcorp.com

Abstract

In this paper, we present an approach and experimental results from using software fault injection to assess information survivability. We define information survivability to mean the ability of an information system to continue to operate in the presence of faults, anomalous system behavior, or malicious attack. In the past, finding and removing software flaws has traditionally been the realm of software testing. Software testing has largely concerned itself with ensuring that software behaves correctly — an intractable problem for any non-trivial piece of software. In this paper, we present "off-nominal" testing techniques, which are not concerned with the correctness of the software, but with the survivability of the software in the face of anomalous events and malicious attack. Where software testing is focused on ensuring that the software computes the specified function correctly, we are concerned that the software continues to operate in the presence of faults, unusual system events or malicious attacks.

1 Introduction

Our motivation for researching advanced software assessment techniques fits in line with the following comments made by the committee that wrote the 1998 *Trust in Cyberspace* report:

1. "The absence of standard metrics and a recognized organization to conduct assessments of trustworthiness is an important contributing factor to the problem of imperfect information. In some industries, such as pharmaceuticals, regulatory mandate has resolved this problem by requiring the development and disclosure of information."

2. "A consumer may not be able to assess accurately whether a particular drug is safe but can be reasonably confident that drugs obtained from approved sources have the endorsement of the US Food and Drug Administration (FDA) which confers important safety information. Computer system trustworthiness has nothing comparable to the FDA. The problem is both the absence of standard metrics and a generally accepted organization that could conduct such assessments. There is no Consumer Reports for [software and information] Trustworthiness."

These statements highlight two key problems facing software users and consumers alike: (1) a lack of sound metrics for quantifying that information systems are trustworthy, and (2) the absence of an organization (such as an Underwriter's Laboratory) to apply the metrics in order to assess trustworthiness. In fact, if these problems were solved, software vendors who sought to provide reliable products would also benefit.

Note, however, that these two problems are not of equal size. Problem (1) is the more difficult and problem (2) can be achieved more easily, but only after problem (1) is solved.

The lack of sound, fair, and quantitative metrics for software safety, reliability, security, and fault-tolerance have contributed to the distrust of Cyberspace mentioned in the report. There is a deeper problem here however, and that is that software quality is more difficult to assess than it is to achieve. This problem is unique to software; physical systems do not experience it. For example, it is far easier to determine if a ball bearing has been perfectly manufactured via an electron microscope than it is to produce perfect ball bearings. Such a situation is not true for software.

Our software research projects over the last 4 years have focused on creating automated technologies and metrics to assess software trustworthiness. Our belief is that enough emphasis has been applied to process improvement methods to improve software quality (even though those processes are often ignored). If

we can better assess the quality of software systems, then hopefully the distrust can be reduced and as a side-benefit, we will be able to assess the return-on-investment from software process improvement.

We acknowledge, along with the report, that the US Government has not ignored the software assessment problem. They have invested heavily in software testing research for the past 20 years. Software testing is still the most common approach for determining whether software will behave as desired. Unfortunately, however, the outcome of that research is not applicable to the large-scale survivability problems endemic to the Internet.

As noted in the *Trust in Cyberspace* report, this research has focused more on testing "in the small" than testing "in the large." While this enables better subsystems, it does not address the interaction problems that weaken survivability:

> "Much of the research in testing has been directed at dealing with problems of scale. The goal has been to maximize the knowledge gained about a component or subsystem while minimizing the number of test cases required. Approaches based on statistical sampling of the input space have been shown to be infeasible if the goal is to demonstrate ultra-high levels of dependability [5], and approaches based on coverage measures do not provide quantification of useful metrics such as mean time to failure. The result is that, in industry, testing is all too often defined to be complete when budget limits are reached, arbitrary milestones are passed, or defect detection rates drop below some threshold. There is clearly room for research - especially to deal with the new complications that MIS brings to the problem: uncontrollable and unobservable subsystems."

Therefore research is needed to increase the observability of "ilities" such as safety, security, reliability, and survivability. In this paper we describe two areas of research that use off-nominal testing for survivability.

2 Off-Nominal Testing for Survivability

In this paper, we present an approach and experimental results from using software fault injection to assess information survivability. We define information survivability to mean the ability of an information system to continue to operate in the presence of faults,

anomalous system behavior, or malicious attack. In the past, finding and removing software flaws has traditionally been the realm of software testing. Software testing has largely concerned itself with ensuring that software behaves correctly — an intractable problem for any non-trivial piece of software. In this paper, we present "off-nominal" testing techniques, which are not concerned with the correctness of the software, but with the survivability of the software in the face of anomalous events and malicious attack. Where software testing is focused on ensuring that the software computes the specified function correctly, we are concerned that the software continues to operate in the presence of faults, unusual system events or malicious attacks.

The off-nominal testing approach uses fault injection analysis to determine how survivable a program is to unusual events that can occur during field operation. Fault injection is the process of perturbing program behavior by corrupting a program state during program execution. Corrupting program states can affect program control flow as well as corrupt program data. We use fault injection analysis to assess information survivability under three different scenarios:

- software flaws in program source code,

- malicious attacks against programs,

- anomalous behavior from third party software.

To assess the survivability of a program, we must know how robust it is under flawed software conditions. Since most programs today contain on average one defect for every 6000 lines of source code, we know that today's systems are deployed with a great number of undiscovered software flaws that may be triggered in the field at anytime [8]. If we knew *a priori* where these flaws exist, we would be able to locate and fix them. However, since we do not know where these flaws are, we simulate their effects by automatically corrupting program state at as many program locations as possible and assessing the effect on survivability of a program state corruption at a particular location. The effect on security and safety of software flaws has been documented in great detail in BugTraq[1] and in [7].

The technique to simulate software flaws uses program state corruption. Since, the range of possible effects on program state is too great to use specific program corruptions, we use random program corruptions

---

[1] See www.securityfocus.com for BugTraq archives.

for specific program state types. For instance, we can corrupt program memory by using random number selection based on the program data type. Program control flow can be corrupted by corrupting Boolean conditions in control flow constructs.

In the second scenario, we are interested in assessing the impact of malicious attacks against programs. In this scenario we can use directed fault injection techniques that subject a software program to the types of well-known attacks it may experience in the field. The most common attack by far is the buffer overrun attack. We have developed specific fault injection functions to test the vulnerability of program buffers to "stack-smashing" buffer overrun attacks. On occasion, testing using random program state corruption to simulate software flaws will sometimes result in unveiling a security flaw. Examples of using these techniques against commonly used network servers are presented later in this paper.

Finally, we are interested in assessing the impact of failing third party software on information survivability. This topic is important to gauge survivability of an information system because today's software is almost always built using third party software such as libraries and commercial off-the-shelf (COTS) components. In the preceding two analyses, we use the source code of the program to perform the fault injection analysis. In assessing the impact of third party software failures on system survivability, we cannot assume access to source code for the third party software (such as proprietary operating system code or COTS software components). As a result, we have developed a technique we call Interface Propagation Analysis (IPA) that gives us the ability to assess the impact of failing third party software in the system under consideration. It is briefly described in Section 4.

3 Source-Code–Based Fault Injection

Fault injection can be applied to software source code by inserting instrumentation "hooks" into the original program source. The idea is to be able to observe program state and corrupt either control flow or data flow at particular locations within the source code. By corrupting program state, we can assess the impact on system survivability to inadvertent flaws or deliberate attacks against the program.

In the fault-error-failure model of software, a fault is introduced by a programmer, known as a "bug" in common parlance. The fault may be an error in the design of an algorithm or a simple coding error, such as an unconstrained buffer array. The fault is innocuous until it is activated (or triggered) by some input. At this point, the error is manifest. An error is

only manifest when the resulting program state is incorrect (according to some correct specification) based the preceding program state and the current input. In other words, if the program state is correct, then the error is not manifest and the fault is inconsequential for the moment. Once the error is manifest, the program, or more generally, the system may continue to perform correctly or it may fail. If the system continues to perform correctly (or at least acceptably), then the error is either latent or it has been masked. If the system fails due to the error, then the error has been manifested as a failure.

We use fault injection to manifest errors. Thus, we are not introducing true faults in the fault-error-failure model sense; rather, we are injecting errors. A closer match to fault injection in the sense of the fault-error-failure model is mutation testing, where program code is selectively "mutated" or altered in order to determine if test cases can distinguish between good and flawed code [3]. Since we cannot know *a priori* where all program faults are, we manifest program errors by corrupting program states. If the errors we introduce during fault injection analysis cause system failure, then we have a measure for how survivable the

3.1 Implementation approaches for fault injection

The hypothesized errors that software fault injection uses are created by either: (1) adding code to the code under analysis, (2) changing the code that is there, or (3) deleting code from the code under analysis. One key requirement from these processes, however, is that the code that is either added, modified, or deleted must change either the software's output or an internal program state for at least one software test case. (Different applications of software fault injection will guide the decisions as to which of these two alternatives applies.) Without this requirement, the hypothesized errors will have had no semantic impact to the original code base and thus were meaningless (they were not anomalies at all). In mutation testing (a type of fault injection that we will discuss later), this is the dreaded "equivalent mutant" problem. The difficulty stems from the fact that equivalent mutants are often undetectable, forcing the costs to perform mutation testing to be much greater than they should be [9].

Figure 1 shows the software fault injection process. Code that is added to the program for the purpose of either simulating errors or detecting the effects of those errors is called *instrumentation code*. To perform fault injection, some amount of instrumentation is always necessary, and althrough this can be added

manually, it is usually performed by a tool. Instrumentation code can be placed on top of input or output interfaces to the software or directly into the logic of the software.

Instrumentation can be added into a variety of code formats: source code, assembly code, binary object code, etc. In short, any code format that can be compiled, interpreted, or that is ready for execution can be instrumented.

There are two key approaches for simulating errors: (1) directly changing the code that exists (this is referred to as code *mutation*), or (2) modifying the internal state of the program as it executes. We will now walk through an example of each approach beginning with code mutation.

Suppose a program has the following code statement:

```
a = a + 1;
```

This statement could be mutated as follows:

```
a = a + a + 1;
```

(provided that a does not have the value of zero). We could also modify the statement to:

```
a = a + 10;
```

And we could delete the statement as well. Note that all of these mutations change the resulting value of a from what it would have had not we not mutated the code (and for every test case that allows this statement to be executed).

The concept of forcefully changing the internal state of an executing program is a slight variation on the code mutation examples just shown. Clearly, each of the mutations above will change the state of the program after they are executed. But note that that is not necessarily true for all mutants. There are code mutants that although they are exercised will not modify the software's internal state. That would be the case if the value of a before the mutant a = a + a + 1 was executed is zero. (This would be an example of a transient fault using the definitions provided by Carriera *et al.*)

To forcefully modify a program's internal state to a value different than the one it currently has, we will add a function call to the code that overwrites the current internal value of a portion of the program's state. Typically, we overwrite programmer defined variables or the data that is being passed to or from function calls. By modifying this data, we are simulating the internal effects of faulty logic or any other anomalous

event that could possibly affect the software's internal state.

The function calls we add to overwrite internal program values are termed *perturbation functions*. Perturbation functions are code instrumentation. When perturbation functions are applied to programmer defined variables, they typically either: (1) change the value of the variable to a value based on the current value, or (2) they pick a new value at random (independent of the original value). Also, they can simply return a constant replacement value if it is suspected that any fault placed at that point in the code would likely result in one particular value regardless of what the current value was. When non-constant replacement values are used, the perturbation function will produce random values based on the current value and a *perturbing distribution*. Non-constant perturbing distributions include all of the continuous and discrete random distributions. The perturbation function

```
newvalue(x)= equilikely(
floor(oldvalue(x)*0.6),
floor(oldvalue(x)*1.40))
```
is an example of a discrete distribution that perturbs a value by substituting an equilikely random value on the interval of 40% more and 40% less than the expected value. This function however leaves the possibility of returning `newvalue(x) = oldvalue(x)`. Conditions are placed in the code that executes this function to avoid this.

For example, if we wanted to change a's value to something close to what it has after this computation,

```
a = a + 1;
```

we would replace the original statement with the following code chunk:

```
a = a + 1;
a = newvalue(a).
```

The code for `newvalue()` would also be added somewhere into the program and would look like the following pseudo-code:

```
int newvalue(int a)
{
  counter = 1;
  oldvalue = a;

  do
  {
    a = equilikely( floor(oldvalue * 0.6),
                    floor(oldvalue * 1.4) );
    counter++;
  }
```

```
while ( (a == oldvalue) && (counter < 100) );

if ( (counter == 100) && (a == oldvalue) )
{
    a = oldvalue - 1;
}

return (a);
}
```

(Note that 0.6 and 1.4 can be modified to however tight or loose of an interval as is desired. For example, 0.0001 and 10000 could be used to widen the interval of choices.)

Because this function could result in an infinite loop while trying to find a different value, a counter is added to ensure that after 100 attempts, the loop terminates and simply decreases the value of a by one. (We could have just as easily decided to program it to increase the value by one or even flip a coin as to which it does.)

Note that we can also use fault injection to modify the time at which code is executed by adding function calls that slow down the software. For example, in Ada, we can add a delay(5) statement to stop a process from executing for 5 milliseconds. And we can even simulate events such as the software's state, stored in memory, having its bits toggled due to radiation or other electromagnetic corruption. The flipBit function which will now be described provides this capability.

flipBit

The perturbation function flipBit toggles specific bits. The first argument to flipBit is the original integer value and the second argument is the bit to be toggled (we assume little-endian notation). The function flipBit is then written in C as follows and linked with the executable. Note that the ^ represents the XOR operation in C and the << operator represents a SHIFT-LEFT of y positions.

```
void flipBit(int *var, int y)
{
    *var = *var ^ (1 << y));
}
```

flipBit can serve as the underlying engine from which other perturbation function can be created. For example, to toggle two or more randomly selected bits in the integer, we can employ flipNbits:

```
void flipNbits(int *var, int n)
{
    int bits = 0;
```

```
    int bitPos = 1;
    int i,j,k;
    int xbit;
    for (i = 0; i < n; i++)
    {
        bits |= bitPos;
        bitPos <<= 1;
    }
    for (j = 0; j < sizeof(int) * 8; j++)
    {
        xbit = lrand48()
        if ((!!(bits & (1 << xbit))) !=
            (!!(bits & (1 << j))))
        {
            flipBit(&bits, xbit);
            flipBit(&bits, j);
        }
    }
    for (k = 0; k < sizeof(int) * 8; k++)
        if (bits & (1 << k))
            flipBit(var, k);
}
```

3.2   Fault Injection Security Tool

We will now discuss our Fault Injection Security Tool (FIST). The tool automates the analysis of security-critical software and requires program inputs, fault injection directives (meaning information about how to corrupt program states), and assertions written in C and C++ (that define when security of the software has been compromised). A schematic diagram of FIST is shown in Figure 1.

The fault injection engine provides a developer or analyst the ability to perturb program states randomly, append or truncate strings, attempt to overflow a buffer, and perform a number of other numerical fault injection functions. The security policy assertion component provides a developer or analyst the ability to code the security policy of the program under analysis as well as system security constraints.

Using FIST is a four step process: instrument, compile, execute, and analyze. The source code is instrumented with assertions and perturbation functions using a source code browser component. The browser tells the user all the legal points in the source where instrumentation can be attached. The user places instrumentation according to the desired analysis, then the instrumented code is compiled. Next, the instrumented program is executed repeatedly, once for each perturbation function that was encountered during an unperturbed run of the program. In each execution, only one location is perturbed. Any assertions that fire during the runs are noted. Relative security met-
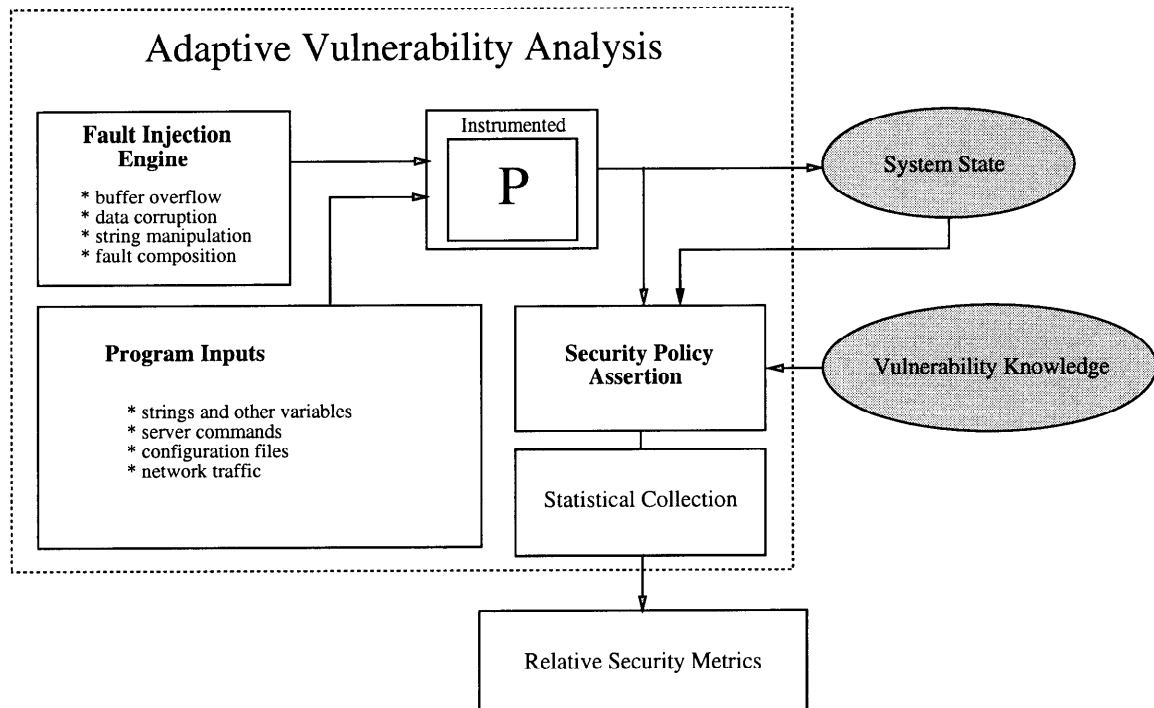
Figure 1: Overview of the Fault Injection Security Tool. A program, P, is instrumented with fault injection functions and assertions about its security policy (based on the vulnerability knowledge of the program). The program is exercised using program inputs. The security policy is dynamically evaluated using program and system states. If a security policy assertion is violated during the dynamic analysis, the specific input and fault injection function that triggered the violation is identified. Algorithm 1 is used to collect statistics about the vulnerability of the program to the perturbed states. One output from the analysis is the relative security metric $\hat{\psi}_{alPQ}$.

rics are accumulated for each program location that indicate the percentage of runs where a fault injection function at that location resulted in a security violation. The user can browse the result of the experiments using a results browser that links results to the original source code.

A fault injection engine has been implemented to support injection of anomalous states as well as specific exploits to test for vulnerability to known malicious threats during the execution of the program. Fault injection functions are instrumented by default in every viable program location to permit analysis of software flaws anywhere in the program source code. The reasoning is that without prior knowledge of where actual flaws exist, simulating their effects everywhere during automated analysis can identify which locations are most likely to impact security. Recall from the algorithm that program states are perturbed singly in each test run in order to assess the effect of a single flaw in a given location.

Fault injection is useful for simulating a variety of anomalous program behavior that would otherwise be very difficult, if not impossible, to simulate using standard testing. The main use of fault injection functions for vulnerability analysis is to determine where potential weaknesses exist in a software program that can be leveraged into security violations. Fault injection also reveals the relative importance of variables, statements, or whole functions on the output (and security) of a program. For example, perturbing the result of a display function may have little or no effect on the output of a program. On the other hand, perturbing the result of a function that parses user input, may well affect the output and perhaps even the security of the application. Finally, fault injection can be used to simulate malicious threats against a software application such as buffer overrun threats. We describe these uses of fault injection in the Section 3.3.

FIST includes numerous fault injection functions for all primitive data types ranging from simple

Boolean state flips, to string mangling, to "stack smashing" buffer overflow functions. These functions include the ability to corrupt Booleans, characters, strings, integers, and doubles. The Boolean perturbation function applies a logical negation operation to an unperturbed value. The character perturbation function returns a character randomly selected from the ASCII table. String perturbation functions provide the ability to truncate strings, concatenate a random string, concatenate a fixed string, generate a new string of random characters, and replace strings with a string randomly selected from a file. In addition to simple fault injection functions, FIST supports composition of fault injection functions from a combination of selected basic fault injection functions. For example, a user can append a fixed string with a random character fault perturbation, thus building a new fault injection function.

The buffer overflow function overwrites the return address of the stack frame in which the buffer is allocated with the address of the buffer itself. By tracing the frame pointer back through the stack, the fault injection function is able to determine where to overwrite the return address. The opcodes for machine instructions are written into the buffer being perturbed. Eventually, the activation record containing the modified return address will be popped off the program stack and the program will jump to the machine instructions embedded by the fault injection function. These instructions will be executed as if they were a part of the normal operation of the program. Because different platforms implement different forms of program stacks, the buffer overflow fault injection functions are platform-dependent. Linux x86 and Sparc are the two platforms currently supported.

Unsafe languages such as C make buffer overflow attacks possible because of input functions such as `gets`, `strcat`, and `strcpy` that do not check the length of the buffer into which input is being copied. If the length of the input is greater than the length of the buffer into which it is being copied, then a buffer overflow can result. Safe programming practices that read in constrained input can prevent a vast majority of buffer overflow attacks. However, many security-critical programs in the field today do not employ these safe programming practices. In addition, many of these programs are still coded in commercial software development labs in unsafe languages today.

FIST detects the potential for buffer overflow attacks to be successful regardless of how the input is read. Searching for unsafe functions such as `strcat` and `strcpy` is one technique for detecting potential

problems; however, it is insufficient by itself. Programmers often write their own dangerous input functions that read in unconstrained input. FIST attempts to overflow buffers regardless of whether the buffer is used in a known dangerous function or is used in a custom-written input function. Furthermore, FIST can overflow buffers for variables that are not pushed on the stack. While this type of perturbation may not result in the execution of arbitrary program code, it may have side effects that compromise program security by corrupting other variables used for access/privilege decisions. If the fault injection function results in a security policy breach, the programmer must either ensure that the vulnerable buffers cannot be overflowed from user input or use safe programming practices to ensure that the buffer overflow cannot occur. Once patched, FIST can be re-run to determine if the patch is resilient to attack.

As an alternative to the source-code–based analysis approach, StackGuard, a `gcc` compiler variant for Linux developed by the Oregon Graduate Institute, attempts to protect buffers from stack smashing attacks by aborting the program if the return address pushed on the stack is overwritten [2]. Stack Guard will not protect programs against all buffer overflow attacks, but can prevent stack smashing attacks from running arbitrary code embedded in user input. For example, buffer overflow attacks that overwrite local variables that were never intended to be user changeable can result in security violations not prevented by StackGuard [1].

The Fuzz tool [4] can be used to overflow buffers, too, but with inconclusive results. Because the input is randomly generated, the vulnerability of the program to executing user-defined code cannot be assessed. FIST implements specific fault injection functions that determine the program's vulnerability to specially-crafted buffer overflow attacks.

FIST integrates with the normal build process of the application under analysis. Any source file that is compiled using the FIST pre-processor at build time is instrumented. Libraries can be instrumented using FIST and then linked to applications, but only if the source code for the library is available. Uninstrumented libraries can also be linked to instrumented applications.

The security-policy–monitoring component of FIST allows users to specify what constitutes a security violation for the software application under analysis. Using assertions to encode this policy, the policy is monitored during the dynamic analysis to determine if it has been violated. The nature of violations will vary

from application to application, and the types of violations the user will seek to detect will generally be dependent on both the input to the program and fault injection functions. As a result, the analyst must determine the security policy for the program being analyzed. A number of pre-defined assertion functions have been developed from which a user can specify the security violations for internal program variables, environment variables, and external system states.

Perhaps the broadest assertion function FIST provides allows the user to develop any expression in C to represent a violation assertion. This expression is evaluated during execution to determine if a violation has occurred. If the result of the expression is non-zero, then the violation is assumed to have occurred. This function has been developed for a sophisticated user who does not want to be constrained by the pre-packaged functions provided in the tool. Assertion functions are placed at locations in the source code during the instrumentation step. FIST also provides a mechanism for external assertion monitoring.

The external assertion monitor runs in parallel with the instrumented program and uses a subset of the built-in assertion functions. It is able to monitor files on the system, checking for modifications and/or accesses. For the buffer overflow functions, FIST checks for side effects of the mycmd program. The assertion is coded such that a file called touch.out should not be modified during the execution of the instrumented program. This assertion will be violated if the buffer overflow succeeds and the mycmd program is executed, which in turn will open touch.out and modify it. So when checking for buffer overflows, the security policy is simple: touch.out should never be modified.

3.3 Case studies of security-critical software

FIST analysis was performed on five different network services. Network service daemons are interesting case studies from a security standpoint because they provide services to untrusted users. Most network daemons typically allow connections from anywhere on the Internet, leaving them vulnerable to attack from malicious users anywhere. Network daemons sometimes run with super-user, or root, privilege levels in order to bind to sockets on reserved ports, or to navigate the entire file system without being denied access. Successfully exploiting a weakness in a daemon running with high privileges could allow the attacker complete access to the server. Therefore, it is imperative that network daemons be free from security-related flaws that could permit untrusted users access to high privilege accounts on the

server.

The programs examined were NCSA httpd version 1.5.2.a, the Washington University wu-ftpd version 2.4, kfingerd version 0.07, the Samba daemon version 1.9.17p3, and pop3d version 1.005h. The source code for these programs is publicly available on the Internet. Samba, httpd, and wu-ftpd are popular programs and can be found running on many sites on the Internet. The analysis of those programs was performed on a Sparc machine running SunOS 4.1.3_U. The other programs, pop3d and kfingerd, are Linux programs found in public repositories for Linux source code on the Internet. The analysis of those programs was performed on a Linux 2.0.0 kernel. The programs were instrumented with both simple fault injection functions as well as the buffer overflow functions where applicable.

A summary of results from the analysis is shown in Table 1. The table shows the total number of instrumented locations together with the number of simple perturbations and buffer overflow perturbations that resulted in security violations. The last column shows the percentage of the functions in the source code that were executed as a result of the test cases employed. Higher coverage results may result in more potential security hazards flushed out through the analysis. The results should not be interpreted to mean that the locations identified in the analysis are necessarily exploitable, only that they require closer examination from the software's developers to determine if they can be exploited from input and whether fault-tolerant mechanisms should be employed. It is worth mentioning, however, that one of the potential buffer overflow vulnerabilities found in wu-ftpd v2.4 and published in [6] was later reported in CERT Coordination Center, Pittsburgh, PA, CERT Advisory CA-99-03, "FTP Buffer Overflows" (see www.cert.org).

4 Interface Propagation Analysis

Much of our research during the past 4 years has been geared toward increasing the observability of large-scale information systems. The main "ilities" that our work has addressed are security and safety.

The premise of our approach is as follows: since it is rarely possible to guarantee "correct" behavior at the system or component level, we should instead focus on guaranteeing levels of "acceptable" behavior. In essence, we should work to thwart system level failures that are the most undesirable and ignore the rest.

Our approach is simple. Start from an assumption about the worst behaviors from a component and observe how that will affect the full system. If the effect is negligible, ignore the component. If the impact is

Program	Instrumented Locations	Successful Simple Perturbations	Successful Buffer Overflows	Function Coverage
Samba v1.9.17p3	1264	12	15	45.5%
NCSA httpd v1.5.2a	463	27	3	40.14%
wu-ftpd v2.4	476	11	3	58.62%
pop3d v1.005h	73	2	1	63.64%
kfingerd v0.07	146	12	5	38.1%

Table 1: **Results from FIST analysis of network daemons.**

large, it is clear that the component is one that needs scrutiny. The bottom line is that we do not care how poorly subsystems behave as long as their behaviors do not jeopardize the integrity of the full system.

Given that resources are always too few, this perspective provides an intelligent way to allocate component testing resources, i.e., to components that have demonstrated a capacity to cause undesirable, system-wide problems.

The approach we have developed is termed Interface Propagation Analysis (IPA). IPA is a fault injection-based technique that simulates component and subsystem failures.

IPA is normally applied once the system is completed. IPA can also be applied before a component is built, provided there exists a specification for what the component is expected to do. (Components that do not yet exist are termed "phantom components"). And finally, IPA can also be used to test the robustness of individual components.

IPA is made of two software fault injection algorithms: "Propagation From" (PF) and "Propagation Across" (PA). PF corrupts the data exiting a real component (or phantom component) and observes what it does to the remainder of the system (i.e., what type of system failures ensue, if any). PF can also observe whether other subsystems fail and how. Thus, PF is an advanced testing technique that provides the raw information needed to measure the semantic interactions between components in order to measure their tolerance to one another.

PA corrupts the data entering a component. This process simulates the failure of system components that feed information into the component in order to see how it reacts. These simulated failures mimic human operator errors, failures from hardware devices, or failures from other software subsystems. After the component under analysis is forced to receive corrupt input, PA observes whether the component chokes on the bad data and fails. Note that PA is very similar to PF. The only difference is scale: PA is focused on standalone components and PF is focused on compo-

nent interactions.

5 Conclusions

In this paper, we described the use of an off-nominal testing approach — fault injection analysis — to test the survivability of an information system to three different types of events:

- software flaws in program source code,

- malicious attacks against programs,

- anomalous behavior from third party software.

Source-code–based fault injection analysis can be applied either to open source software after software is released or to software during development by software vendors. The earlier in the software lifecycle off-nominal testing techniques are used, the cheaper the cost to find and correct bugs. The Fault Injection Security Tool supports testing of the first two scenarios above: simulation of software flaws and malicious attacks against programs. The tool was applied to several commonly deployed open source systems. Even with the low levels of code coverage, several potential security-related hazards were demonstrated, one of which was later independently found and reported to the CERT CC.

The third scenario is becoming increasingly important. Software developed and released today is heavily dependent on third party or COTS software. Anomalous behavior from third party software can result in system-wide failure. Interface Propagation Analysis addresses the survivability of a system composed of custom and third-party components by using fault injection analysis at component interfaces. The fault injection analysis can determine the effect of failing or anomalous behavior of third party software on system survivability. This technology is a key step from moving from "testing in the small" to "testing in the large".

5.1 Acknowledgements

Earlier versions of this paper can be found at:

1. A. Ghosh and J. Voas. "Innoculating Software for Survivability," *Communications of the ACM*, 42(7):38-44, July, 1999.

2. J. Voas and A. Ghosh. "Software Fault Injection for Survivability", In *Proc. of the DARPA Information Survivability Conference and Exposition*, January, 2000.

References

[1] S. Bellovin. Re: Stackguard: Automatic protection from stack-smashing attack. On-line. Bugtraq archives. See http://www.geek-girl.com/bugtraq/1997_4/0514.html, December 19 1997.

[2] C. Cowan, C. Pu, D. Maier, H. Hinton, J. Walpole, P. Bakke, S. Beattie, A. Grier, P. Wagle, and Q. Zhang. Stackguard: Automatic adaptive detection and prevention of buffer-overflow attacks. In *Proceedings of the 7th USENIX Security Symposium*, pages 63–78, San Antonio, TX, January 1998.

[3] M. Daran and P. Thevenod-Posse. Software error analysis: A real case study involving real faults and mutations. In *Proceedings of the 1996 Int'l Symp. on Software Testing and Analysis*, pages 158–171. ACM Press, January 1996.

[4] B.P. MILLER ET AL. Fuzz revisted: A re-examination of the reliability of UNIX utilities and services. Technical report, University of Wisconsin, Computer Sciences Dept, November 1995.

[5] R. BUTLER AND G. FINELLI. The infeasibility of experimental quantification of life-critical software reliability. In *Proceedings of SIGSOFT '91: Software for Critical Systems*, pages 66–76, New Orleans, LA, December 1991.

[6] A.K. Ghosh, T. O'Connor, and G. McGraw. An automated approach for identifying potential vulnerabilities in software. In *Proceedings of the 1998 IEEE Symposium on Security and Privacy*, pages 104–114, Oakland, CA, May 3-6 1998.

[7] J. VOAS AND G. McGRAW. *Software Fault Injection: Inoculating Programs Against Errors*. John Wiley and Sons, New York, 1998.

[8] J. D. MUSA, A. IANNINO, AND K. OKUMOTO. *Software Reliability Measurement Prediction Application*. McGraw-Hill, 1987. ISBN 0-07-044093-X.

[9] R. A. DEMILLO, R. J. LIPTON, AND F. G. SAYWARD. Hints on test data selection: Help for the practicing programmer. *IEEE Computer*, 11(4):34–41, April 1978.

Model-Based Design of
Information-Rich Command Organizations

Daniel Serfaty
Aptima, Inc., 600 West Cummings Park, Suite 3050, Woburn, MA 01801, USA
Tel: (781) 935-3966, Fax (781) 935-3966 <www.aptima.com>

INTRODUCTION

Command organizations and teams[1] are not usually "designed" in a formal sense. Instead, organizational structures and individual roles for team members evolve over time, based on previous structures and roles, through an ad hoc process of trial, error, and adjustment. For military teams in recent years, however, a combination of rapidly evolving technology and frequently changing missions have created the need for more rapid and efficient ways to create team structures that take maximum advantage of the capabilities of technology for accomplishing mission goals.

The influx of technology on the battlefield has altered the nature of military missions. Today's military missions are complex processes executed by networked individuals, supported by highly sophisticated hardware, all functioning in dynamic and uncertain environments. They require extensive communications, coordination, synchronization, and information management. This rapid development of advanced information technology and the resulting concepts of "information-centric warfare" demand changes to communication and collaboration at both the individual and organizational levels within the military. This changing environment has created the need for innovative methods for designing effective military teams.

This paper describes a breakthrough organization/team design method—a systematic, formal, quantitative approach to designing a team that best fits the mission to be accomplished. The Team Integrated Design Environment (TIDE) is a tool set designed to support this method, enabling the quantitative definition of requirements for command teams operating in complex mission environments. The TIDE methods and tools represent a powerful methodology to create novel organizational structures, based on operational mission variables, using quantitative methods. We know of no other methods that provide a similar formal framework for this type of

design. This paper explains what it means to design a team or an organization and describes the TIDE method for team design. Then it presents some initial empirical results that indicate that optimally designed teams can outperform teams that use more traditional organizational structures, and discusses how the team design process must be altered to focus on different concerns, depending on the nature of the team being designed and the environment in which that team must function.

WHAT DOES IT MEAN
TO "DESIGN" A TEAM?

The military's need for effective teams and other organizational structures has led to considerable progress in the last decade on methods for improving the performance of teams (see Serfaty, Entin, Deckert, and Volpe, 1993; Brannick, Salas, and Prince, 1997; Salas, Bowers, and Cannon-Bowers, 1995; Salas, Dickinson, Converse, and Tannenbaum, 1992; Swezey and Salas 1992). A useful product of this research has been the development of a shared definition of what constitutes a team. Salas, Dickinson, Converse, and Tannenbaum (1992) define a team as having the following characteristics:

- There is dynamic, interdependent, and adaptive interaction.
- There is a common goal, mission, or objective.
- There is some organizational structure of the team members.
- Each individual team member has specific tasks or functions.
- Task completion requires the dynamic interchange of information, the coordination of task activities, and constant adjustment to task demands.

The TIDE design approach produces a "team" in the sense that it is defined above. Based on the mission objectives for the team, we specify the specialized roles and functions of each team member, the information exchange and coordination interactions that must take place among the team members based on those roles and functions, and the organizational structure for the team.

The focus of much prior team research has been on improving team performance through training to im

---

[1] In this paper, we use the terms "organization" and "team" in an interchangeable manner. In the literature, they usually have different definitions. The organizational design methods described in this paper have been applied to the design of both small teams and larger command organizations.

prove team competencies and through collaborative tool technology. However, we suggest that there is a third major facet that can be manipulated to improve team performance—the *team structure*. Figure 1 illustrates the three facets underlying team performance and the tools and processes available to support them. Team competencies are addressed through selection, assessment and training to ensure that individuals with the right knowledge, skills and abilities are selected for the team, and that they receive adequate training in both taskwork and teamwork skills. A substantial body of research has addressed these issues for military environments (see Salas, Bowers, and Cannon-Bowers (1995) for a review). We suggest that the definition and measurement of team competencies is interdependent with the team structure, defined as the tasks performed by each team member, the information needed for those tasks, the nature of the interdependencies among tasks, the coordination and communication required between team members because of the interdependency of their tasks, and the hierarchical command structure of the team.
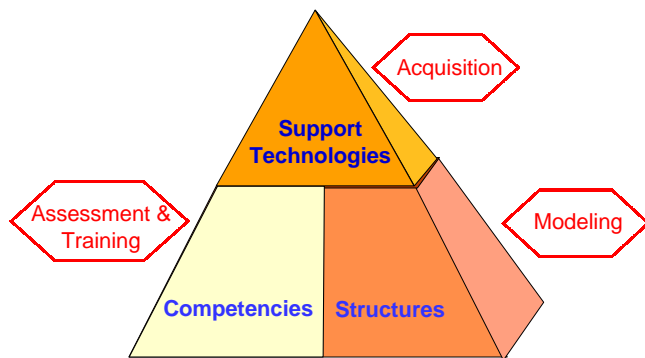
Figure 1. Three Facets of Team Performance

We believe that the most effective support technologies for the team—communication links, shared displays, and other technologies to support collaborative work—will depend on both team competencies and team structure. Requirements definition during the acquisition of support technologies should, therefore, be based on an understanding of both team competencies and team structure.

Research focused on collaborative support technology or on team training and assessment usually takes the team structure, the third element in Figure 1, as a given. Our work, in contrast, focuses on designing the best team structure for given set of goals and tasks (the team's mission), based on the application of optimization algorithms to a model that relates team structure to team performance. The method for team design described in this chapter is model-based, using a mathematical representation of the mission tasks to suggest an optimal team structure for performing those tasks.

The process of designing a team structure is far more complex than simply specifying an organization chart or "wiring diagram." The team structure specifies both the structure and the strategy of team, including who owns which resources, who takes which actions, who uses what information, who coordinates with whom and the tasks about which they coordinate, and who communicates with whom. It includes role definitions for each of the team members as well as a specification of a command structure for the team.

WHY DESIGN A TEAM?

The TIDE approach to team organizational design is *model-based* in the sense that it represents the mission, tasks, and functions to be accomplished by the team, the demands of those tasks and the resources required to accomplish them, the constraints on the team structure, and the performance goals for the team in a mathematical structure. This mathematical structure can then be manipulated to create a team design that is optimized for specified criteria. As illustrated in Figure 2, a mathematical representation of a complex problem such as the accomplishment of a military mission, is, by necessity, a simplification of a complicated, messy, and uncertain world. As the saying goes, "the map is not the territory"—it is only a representation of that territory. The ultimate criteria for the usefulness of such a simplification is whether it produces answers to questions that are useful when they are fed back into the real world and put to use. A map is useful if it helps you get to your destination. A mathematical model of a mission is useful if it can be manipulated to produce a team structure that functions effectively for the mission for which it was designed.

What is the advantage of using a mathematical representation of the mission to formally design a team structure? The team design problem seems to fall into that area of complex, interdependent, dynamic, and ambiguous problems characterized as "wicked" design problems (Rittel and Webber, 1973; Vicente, Burns, and Lawlak, 1997) for which seasoned practitioners suggest that perhaps the best design solution may be a matter of "muddling through (Lindlblom, 1953; Vicente, Burns, and Lawlak, 1997). In fact, the usual approach to creating a team structure is exactly this—to make small evolutionary changes to an existing structure, not to start "from scratch" with a blank sheet of paper in order to design a new team.

We argue that model-based team design has value even though it involves, by necessity, the simplification of a complex problem. First, it provides a way to approach the design of a team for a radically different mission or for radically different organizational constraints. If the mission, environment, or design constraints differ enough from those for which there are known, existing solutions, the strategy of taking an existing solution and modifying it becomes less useful. For example, the U.S. Navy is currently developing designs for the next generation of surface ships. A major goal for these ships is to drastically reduce the number of crew members required to operate the ship, moving from a crew of three to four hundred down to a crew size of fewer than 100 people (Bush, Bost, Hamburger, and Malone, 1998). This goal is to be achieved through the introduction of automated capabilities throughout the ship, along with a planned redesign of the traditional roles and responsibilities of crew members, taking into account the restructuring of tasks that will be brought about by advanced automation. Small changes based on historical structures will not address the team design goal for these ships. Model-based team design offers a way to approach this complex problem and to generate innovative possible solutions that are not overly rooted in previous ways of doing business.

The map is not the territory….
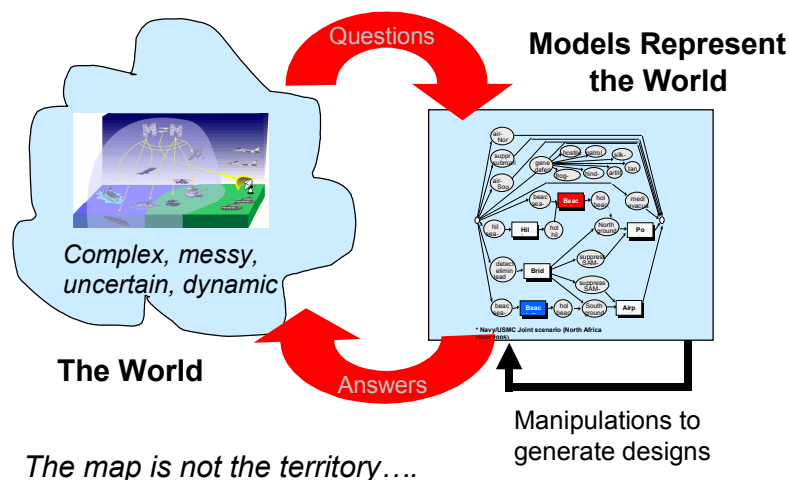
Figure 2. The Model-Based Design Problem

In another example, a TIDE-based redesign of a Joint Task Force team structure suggested it would be more efficient (requiring less coordination and communication) to locate the control of "joint" assets at much lower levels of a command hierarchy than is the current practice, e.g., a lower-level commander at the scene controls both Air Force and Navy assets (Levchuk, Pattipati, and Kleinman, 1998; Entin, Serfaty, and Kerrigan, 1998; Entin, 1999). Experienced

commanders reviewing the design commented that they could see the value of the new organizational design, but that individuals did not currently exist with the expertise to effectively exercise control of both types of assets.

This example highlights a second advantage of the TIDE approach—it provides a way for new mission requirements to drive the design of new selection and training requirements or new collaborative technology requirements. Of course, team structures that vary radically from the traditional introduce a new set of issues regarding team competencies (hence the interrelationship shown in Figure 1). If tasks are grouped into roles in a new way, the knowledge and skills needed for those new roles, and the associated training requirements for individuals on the team, must be radically redefined. Although the costs of changes in selection and training requirements must be factored into overall design feasibility and cost, the TIDE design method provides a way to generate innovative ideas for further exploration and testing.

Similarly, the technologies required to support new team structures must take into account the new roles for team members. For example, command centers are often designed to provide the capability for multiple large screen displays to provide a "common picture" for command team members. What information should be on these displays, and who needs to see that information? These questions should drive the design of the displays, but they are often asked after the display technology has been acquired, not before. With model-based design, the shared information requirements among team members are defined and used as part of the design process, providing a basis for display requirements specification during acquisition.

New technologies enter into the team design process in two major ways. First, the introduction of new technologies will alter the nature of the tasks that must be performed by humans. For example, it is rapidly becoming a truism of human factors that introducing automation does not simply offload tasks from the human, it changes the nature of the tasks to be performed and can radically alter the way humans define their goals and think about what they can achieve in a given situation (Woods, 1997). The capabilities of new technologies such as new sensor systems, new weapons systems, and new information fusion algorithms must be taken into account in defining the tasks that are the basic building blocks of the TIDE approach.

Second, the nature of the collaborative technologies available to the team serves both as a constraint to the design and as a requirement that is a product of the design. By collaborative technologies we mean the team's ways of obtaining and sharing information: physical proximity, electronic communication links within and outside the team, and individual and shared information displays. These technologies can serve as constraints on the design of the team, e.g., if there will be no real-time communication link available between two team members, then those individuals should not be assigned tasks that require rapid coordination. The team design can also drive the *need* for collaborative technology, e.g., if two team members are performing tasks that utilize the same set of information and require frequent communication, then it may be advantageous to have the two individuals co-located, or to provide a very reliable high-bandwidth communication link between them, and for both individuals to share the same display or to have a linked display that is visible in two locations in order to communicate clearly in reference to the same information. For example, one of the challenges for AWACS Weapons Directors (WDs) providing intercept information to fighter pilots is that the WDs and the pilots view very different radar displays, but need to communicate very rapidly and accurately about the same objects (enemy aircraft). This necessitates a specialized and precise verbal communication protocol to ensure that the WDs and the pilots are, in fact, talking about the same object. In a TIDE-based team design, the nature of the coordination between tasks drives the allocation of those tasks to individuals, and the need to share information in order to perform the coordinated tasks, in turn, drives the communication links required.

THE TIDE APPROACH TO TEAM DESIGN

Team design requires, in essence, the specification of "who does what when." The central thesis of our team-design method is that a set of interdependent, interrelated tasks that must be completed under time constraints has an underlying quantitative structure that can be exploited to design the "best" team for accomplishing those tasks.

At the core of our method is a systems-engineering approach that describes organizational performance criteria as a multi-variable objective function to be optimized. This approach is based on a three part allocation model, presented in Figure 3, that considers: 1) the tasks that must be accomplished and their interrelationships (the "mission"); 2) the external resources needed to accomplish those tasks (e.g., information, raw materials, or equipment); and 3) the

human decision makers who will constitute the team. The team design process is, in simplest terms, an algorithm-based allocation between these three parts.

First, a quantitative model describing the mission and the existing organizational constraints is built. Then, one or more objective functions for the design are specified. Finally, an organization is designed to optimize the objective function(s). When the objective function includes several non-commensurate criteria, the organizational design problem is treated as a multi-objective optimization problem. The power of quantitative modeling lies in describing a great variety of phenomena underlying the structure of a mission and of an organization by a relatively limited set of fundamental elements, parameters, variables, laws, and principles. These laws and principles specify the functional interdependencies among the structural elements and the dynamics of system parameters and variables. The algorithms that are fundamental to this team design method (Levchuk, Pattipati, and Kleinman, 1998) were originally developed under the sponsorship of the Office of Naval Research for the Adaptive Architectures for Command and Control (A2C2) program (Serfaty, 1996).

Figure 3. Three Part Allocation Model for Team Design

Inputs From Subject Matter Experts

Our team design method is algorithm-based, but it relies on heuristics and on the judgment of subject matter experts to frame the design problem in a meaningful way, including decomposing an overall mission (or goal) into specific tasks, specifying the relationships between tasks, specifying the resources needed to complete the tasks, and specifying the criteria to be optimized for the team. Subject matter experts in the area of application are also needed to review and revise the organization and structures suggested by the model. The design method is iterative. Typically, review of the team designs suggested by

the algorithms reveals adjustments and corrections to be made in the task structure, the organizational constraints, or the optimization criteria.

The team-design methodology is goal- or mission-driven. That is, the model uses a detailed scenario that specifies the tasks required to accomplish a goal and the resources available to accomplish those tasks, and uses algorithms to optimally allocate these tasks and resources to team members to create an organizational structure for best accomplishing the goal. To capture the operational elements in a scenario, we rely on expert insight from subject-matter experts who develop scenarios. The interaction between operational experts and modeling specialists at this stage is essential for the design process.

Of course, subject matter experts do not always agree in their characterization of the mission, their descriptions of relevant scenarios, or their opinions about the effectiveness of resources for different tasks. An area that has been problematical for the TIDE approach, and for which we hope to develop more consistent and reliable methods in the future, is the resolution of differing SME opinions. We also need better methods for allowing SMEs to see the consequences of their inputs for the team design, thus allowing them to judge whether the strength of their opinion is sufficient to warrant its effects on the design. For example, in a recent design of a Navy team, we used SME input to constrain the availability of external communication links to team members, i.e., only one team member was able to control each external communication link. This constraint turned out to be a major driver of the team design that was produced by the algorithms, but the importance of the constraint was not obvious to the SMEs who were reviewing the design. After extensive review and discussion, it was decided that this constraint should be relaxed so that alternative designs could be generated and explored.

In addition to the selection or development of a scenario (or multiple scenarios), it is necessary to create a detailed model of the mission that serves as the input for the method. An essential question that underlies all organizational design processes is "Who does what?" This requires that a mission be described in terms of its tasks (the "what" independent of the "who"). There are multiple ways to decompose a mission, and this process relies on interaction between the designer and domain experts. Mission analysis, functional decomposition, and subsequent function allocation must be driven by design goals.

After multi-dimensional task decomposition is used to identify mission elements, specific modeling techniques are applied to capture the internal structure of the mission. The mission decompositions are used to define parallelism, sequence, and structure for the mission tasks. These task interdependencies are used to create a hierarchical structure among mission tasks which is represented by a mission task dependency graph.

There are two major inputs for the team design method, the quantitative mission structure just described (e.g., parallelism or required sequencing of tasks, time needed to complete the task, required completion times driven by external constraints, resources available to the team, and the estimated effectiveness of resources for completing the task), and a set of organizational constraints. Organizational constraints include the specific resources and technologies available for accomplishing the tasks as well as any restrictions on how tasks are assigned to team members, based on specifications by subject matter experts who understand the domain of application. Team size may be set as an organizational constraint, or allowed to vary as part of the optimization. Other organizational constraints may specify, for example, the need to group certain tasks together because they require a specified level of authority (e.g., weapon release) or the need *not* to group certain tasks together because the knowledge and skills required to perform them are so disparate that attempting to have one individual perform them would create insurmountable selection or training problems for the organization.

Steps In The Design Process

Figure 4 shows the steps followed in a typical team design process. The first stage is mission representation,[2] which depends heavily on inputs from subject matter experts. At this stage, we define the tasks that must be completed in order to accomplish the mission and specify their interdependencies. Tasks may be triggered by events (e.g., the appearance of a new air track triggers the task of identifying that track) or they may be triggered by other tasks (e.g., once a track is identified, its intent must be evaluated). Still

---

[2] In this section, the terms "representation" and "model" are used interchangeably to indicate the exact specification of what must be accomplished during the mission. For example, certain tasks may need to be performed before other tasks can be initiated (e.g., identify an aircraft as hostile before targeting it). This sequential interdependence of tasks may be represented in a matrix form, where values in each cell of the matrix indicate that the task in the column may not be started until the task in the row is completed.

other tasks are on-going, independent of events or other tasks (e.g., the need to continually monitor for new tracks).
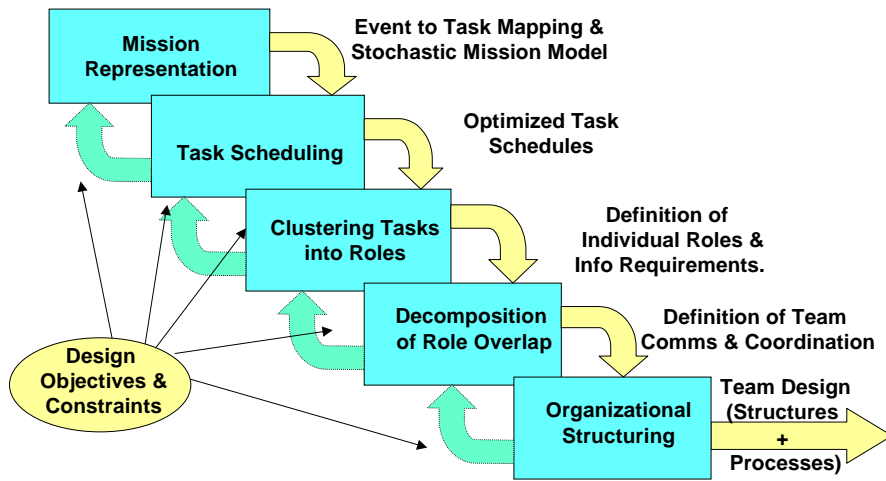
Figure 4. Steps in Designing a Team

Typically we work with one or many mission scenarios in designing the team. If possible, we develop a stochastic mission model, which specifies the scenario in terms of the probabilities of various events occurring, rather working from a single deterministic scenario.

At the mission representation stage we define "attributes" for the tasks to be accomplished. The task attributes of greatest interest will vary depending on the nature of the team design problem (as discussed in more detail below), but typical attributes that are considered include the workload associated with the task, the time needed to complete the task, the information needed to accomplish the task, and the communication/coordination links that exist among tasks due to the nature of the work being performed (e.g., the planning of air-to-ground strikes requires coordination with the planning of ground troop movements).

At the mission representation stage, we also specify the resources that *could* be used to accomplish the task, if the problem is resource constrained. Resources include, for some types of teams, assets such as sensor or weapons systems. Some types of assets can only be used at one place and at one time (e.g., an artillery unit) while other types of assets can be used simultaneously by many people in many locations (e.g., information). Depending on the domain of application, there may be multiple ways to accomplish the same task with different combinations of assets (e.g., ships, amphibious units, and aircraft may all be involved in a mission task such as "take the beach"). If there are multiple ways to accomplish a task, we

specify (based on subject matter expert input) the relative effectiveness of each of the possible combinations of assets for accomplishing the task.

The next step in team design is task scheduling. This step is accomplished by optimization algorithms that determine the optimal way to use the available assets in order to accomplish the tasks given an overall objective, e.g., to minimize the time needed to accomplish the mission or to maximize mission effectiveness. The importance of the task-scheduling step in team design depends on the nature of the mission domain. If there are a number of assets that can only be used in one place at one time, and a number of different ways that assets can be combined to accomplish tasks, this step may be extremely important in team design. In contrast, if there is relatively little competition for assets, or only one way to accomplish a task with those assets, then task scheduling is not a major factor in the design of the team.

The output of this step in the design process is an optimized task schedule for using the available assets to accomplish the mission. At this stage, human roles have not yet been considered. The schedule produced at this stage is "optimal" only under the assumption that all of the team members can do any of the tasks and that there are no constraints on the amount of work any individual can do. Obviously these assumptions are unrealistic, and there is an iterative process in which the results of the next stage, in which tasks are assigned to individuals, are used to adjust the optimal schedule to take human capabilities into account.

The next step, and the central one for team design, is to create roles for individuals by clustering tasks (and the resources needed to accomplish them) in such a way as to optimize an objective function. Task clustering is often done on the basis of two (potentially competing) criteria: the goal of equalizing workload across the team members, and the goal of minimizing the amount of communication/coordination required between team members. The tension between these two criteria can been seen from a simplified example: the best way to minimize the need for coordination is to assign all of the tasks to one individual, but this obviously directly contradicts the goal of equalizing the workload across the team.

Because workload is often central for team design, the definition of the workload associated with tasks is an important issue for the design. This is an area in which the TIDE approach could benefit from improved data collection methods for working with subject matter experts to elicit workload information. At the simplest level, workload is defined by the number of tasks being performed by an individual. Obviously this is an unsatisfactorily crude definition, since tasks can differ widely in their demands on the human. A more sophisticated approach, used in the design of Navy command center teams, is to ask SMEs to rate tasks, based on a description of the task, on a workload scale for four dimensions: visual, auditory, cognitive, and psychomotor workload. The correct method for combining these ratings into a single workload rating for the task is far from obvious, however. An additional issue is the workload "overhead" associated with an individual performing multiple tasks simultaneously (Adams, Tenney, and Pew, 1994). As tasks accumulate, workload does not simply accumulate linearly, there is an added workload for "juggling" multiple tasks that is a function of the number of tasks being juggled. The best approaches for defining workload, developing workload estimates, and establishing thresholds for workload tolerance are critical areas in which additional research could enhance the TIDE method.

While the goal of equalizing workload (or keeping workload below a tolerable threshold) is a relatively intuitive one, the goal of minimizing the need for coordination requires further explanation. It is not that coordination is, in itself, "bad." However, if communication is required in order to achieve that coordination, then that communication takes up the time and attention of team members. Therefore, the need to coordinate through communication can have a negative effect on performance in conditions where there is a high task load (i.e., workload imposed from outside the team). While it is always good to have information about what other members of the team are doing, there may be a cost to acquiring that information. Communication can be good or bad for team performance, depending on when it occurs and what else is going on at that time.

Team theory suggests that if individuals on a team have a good "mental model" of what each of the other team members is doing and a good shared mental model of the situation, then this mental model allows them to anticipate the needs of the other team members (MacIntyre, Morgan, Salas, and Glickman, 1988; Cannon-Bowers, Salas and Converse, 1990; Kleinman and Serfaty, 1989; Orasanu, 1990). This mental model can be acquired through communication and

planning during periods of low workload ("here's how we are going to handle it when…") (Orasanu, 1990) or through cross training (each team member receives training in the other's job) (Travillian, Volpe, Cannon-Bowers, and Salas, 1993; Baker, Salas, Cannon-Bowers, and Spector, 1992) or simply through experience.

In periods of high workload, these mental models allow members of the team to anticipate the needs of other team members so that they can coordinate "implicitly" (with less need for communication) rather than coordinating explicitly (requiring communication of the form "send me this" or "do this now"). Implicit coordination reduces the need for communication under high task load, freeing team members up to do other things, and causing the team to perform better (Serfaty, Entin, and Volpe, 1993; Serfaty, Entin, and Johnston, 1998). So, it is not that either coordination or communication is poor, it is just that, especially under high task load, teams often perform better if they can coordinate without the need for frequent communication.

For team design, assigning tasks to minimize the *need* for coordination (to the extent possible, without overloading any of the team members) reduces the amount of knowledge the team members need to have about each other's roles, and the amount they need to communicate. This is most critical, and probably will have the most effect on performance, when the team is in high stress, high task load conditions.

The product of the clustering step in team design is to define roles for individuals in terms of the tasks for which they will be responsible. Associated with those roles, based on the attributes of the tasks, is a specification of the information that will be used by each team member, the resources that each individual will control in order to accomplish the tasks, and the need for coordination among team members (based on the interdependencies of tasks). Another product of the clustering is a prediction of each individual's workload over time, based on the tasks assigned to that individual and the timing of the tasks in the mission scenario. Note that if workload is a major concern for the team design, we also include an estimate of the "overhead" workload associated with managing multiple tasks simultaneously.

The results of the clustering step must be fed back into the optimized task schedule to determine if that schedule is feasible given the assignment of tasks to individuals. We might discover, for example, that the "optimal" schedule requires an individual to accomplish too many tasks simultaneously, and will there-

fore need to delay tasks or to change the task assignments as a result. We may also specify as a constraint on the model, that certain tasks should not be grouped together to be done by one individual because they require such disparate knowledge or skills that it would be too difficult or costly to select or train a single individual with the needed skill set.

For some team designs, it will be possible to assign tasks to individual team members in such a way that no one team member is overloaded. For other teams, such an assignment may not be possible, and it may be necessary to assign the same task to multiple individuals, creating an overlap in task responsibilities. If so, this creates a need for communication and coordination among the individuals with overlapping responsibilities, which must then be factored back into calculations of the workload for each of the affected team members.

The final step in the design process, once individual roles have been defined, is the specification of an organizational structure (e.g., a command hierarchy) for the team. For military teams, this is usually straightforward, driven primarily by the need to designate a team commander. The workload associated with being the team commander must also be fed back into the workload calculations, however, to ensure that command responsibility has not been placed on an individual who is already at a maximum workload ceiling.

The final output of the team design process is a specification of both a team structure and a team process associated with that structure. The team design specifies which team member (or members) accomplishes each task, what resources are controlled by each team member, what information is used by each team member, and who needs to coordinate with whom (and about what). Depending on the criteria used to optimize the team and the attributes defined for the tasks, the final design can also produce predictions about the team's performance and the workload that will be experienced by each of the individuals on the team.

EXPERIMENTAL EVALUATION OF TEAM DESIGNS

The ultimate test of the model-based optimal team design method is the performance of the teams that have been designed using this method. Initial empirical evidence is available from the Adaptive Architectures for Command and Control (A2C2) program (Serfaty, 1996) on the effectiveness of model-based team design. In the A2C2 program, innovative mission-based Joint Task Force (JTF) team structures

were designed using the model-based optimization method. As a comparison, a group of subject matter experts also generated team structures for the same JTF mission.

The two team structures were "played out' in a simulation-based experiment, with 10 six-person teams of military officers from the Naval Postgraduate School in Monterey (Entin, 1999). Each team participated under both architectures, with the order counterbalanced to control for learning effects. Figure 5 shows the results of the experiment. Two types of summary performance measures are shown: simulation-based measures, which come directly from the simulation testbed, and observer-based measures, which were prepared by subject matter expert observers rating team behavior during the experiment sessions. For both types of performance measures, the performance of the six-person team designed using the model-based method was superior to the performance of the six-person team using a more traditional team structure developed by subject matter experts. The model-based method was also used to design a reduced-staff four-person team, shown as "model reduced" in Figure 5. The performance of this four-person model-based team was at the same level as (not significantly different from) the performance of the six-person team designed by the experts.
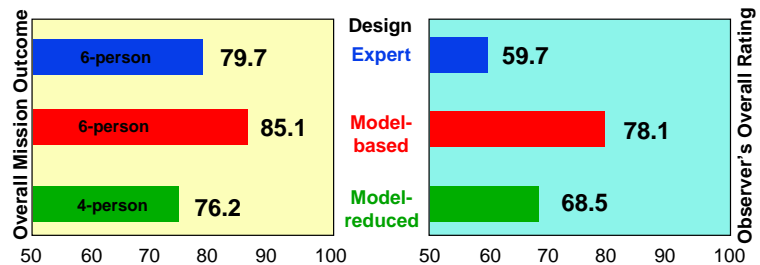
Figure 5. Performance in model-based (optimized) versus traditional (designed by subject matter experts) team organizational structures.

The optimized team was designed to reduce the need for communication and coordination among team members, and the results in Figure 6 show that it was successful in this objective.

The six-person optimized team achieved higher performance levels with fewer coordination actions and a lower communication rate. The six-person optimized team also had a higher "anticipation ratio." This anticipation ratio measures the ratio of information transfers over requests for information. Higher values of the anticipation ratio indicate that team members were "pushing" information without having to be

asked, also indicating more effective coordination (i.e., coordination with less communication).

The innovative team structures developed using the optimal design method resulted in superior performance only if the teams were thoroughly trained in the new team structure prior to using that structure in the experiment, however. Earlier experiments (Entin, Serfaty, and Kerrigan, 1998) in which subjects received less training on the innovative team structures failed to find significant differences between model-based and traditional team designs.

Figure 6. Communication and coordination measures for model-based and expert-designed team structures

An interesting feature of the JTF team designs produced by the model-based method was that the algorithms tended to push the "jointness" of the control of resources down to much lower levels in the command structure than is current military practice (e.g., a lower-level commander might control both Navy and Air Force resources). Although the military domain experts working on the project commented that the expertise to handle this combination of resources does not currently exist at lower levels of command, they admitted that such an organization would probably be more efficient than current practice.

Overall, the results of the A2C2 experiments indicate that the optimized, model-based team design method can produce innovative team structures in which teams can perform at a higher level than they do under more traditional structures. This improved performance is observed only if teams receive sufficient training in how to function in the new structures, however.

DESIGN FOCUS BY DOMAIN OF APPLICATION AND TYPE OF ORGANIZATION

We are currently engaged in applying the TIDE team design approach described in this chapter in a number of different military domains. Each domain presents different challenges for team design, and requires adaptation of the method and emphasis on different aspects of the design process.

Joint Task Force (JTF) Command Team

A primary issue for the design of JTF command teams (see results above) is the control of resources. In the mission being analyzed, the JTF teams orchestrated the use of Navy, Air Force, Marine, and Army resources (ships, planes, infantry units, satellite sensors, etc.) to recapture a port that was being occupied by the enemy. Many of the tasks depended on the success of the previous task (e.g., "advance to the airport" could not be initiated until "take the beach" was accomplished). There were often a number of ways in which a particular task could be accomplished with the available resources, but a resource being used in one geographical area could not be used immediately in another. For this application, the optimal (requiring least coordination) control of resources was a driving factor for the design, leading to the creation of team structures in which each team member directly controlled many if not all of the resources needed to accomplish his or her tasks. Note that the model-based approach produced team designs that are quite different from traditional JTF designs, with joint control of Navy/Army/Air Force assets at much lower levels of the command hierarchy than is currently the case.

Next Generation Navy Surface Ships Command Team

For this application, the goal is to design the next generation of Navy ships to take advantage of automation and to operate with a much smaller crew than is currently required. The goal is to reduce the number of individuals needed in the shipboard command center by half, from 20 or more to approximately 10. In this application, the control of scarce and geographically dispersed resources is not the driving issue for team design, as it was for JTF team design.

The major resource needed by the shipboard command team is information, which can be made available to everyone simultaneously with the planned technology. The primary concern for this team is balancing workload in order to keep workload below a manageable threshold for all team members. The fundamental question is: Can 10 people, aided by technology, handle a mission that previously required 20? For this design effort, we are working with more detailed workload data and developing new methods for modeling workload, including methods for calculating the workload effects of multi-tasking.

AWACS Command And Control Team

Teams on board Air Force AWACS planes direct air traffic and monitor for hostile aircraft from an airborne command center. Because the team is airborne, and must fit into limited space, the number of crew positions needed is a critical concern. With the introduction of new sensor technology, some of the tasks previously performed by the crew will be automated. The primary issue for this team redesign problem is how the responsibilities of the team members should be reallocated now that some tasks have been eliminated, and whether it may be possible to reduce the number of positions needed on board the aircraft.

Uninhabited Air Vehicle Control Operations Center

Current uninhabited air vehicles (UAVs) require a team of multiple operators on the ground to control one UAV in the air. Future concepts call for a reversal of this ratio, with a small team of operators on the ground controlling many UAVs in the air. Our focus is the design of roles for the ground controller team. Preliminary analysis shows that the major problem for designing this team involves the sequencing of waves of aircraft and the patterns in which the aircraft will be flown. The workload associated with the control of the UAVs varies enormously at various stages in the UAV's flight. The challenge will be to develop deployment patterns for the UAVs that do not result in the creation of infeasible workload peaks for the team in the control center.

Air Operations For Time Critical Targets (JFACC) Team

A theater-level air campaign such as the one just conducted in the Balkans requires the generation and execution of Air Tasking Orders (ATOs), typically on a daily basis. These ATOs specify targets as well as the aircraft and weapons to be used to strike these targets. Difficulties arise when new target information is received, however, or when some aspect of the plan proves unworkable (e.g., a tanker that was scheduled to perform airborne refueling has mechanical problems and must return to base). In these situations, the speed with which the Joint Force Air Component Commander (JFACC) air operations organization can respond to new information, modify plans, and execute those new plans, becomes critical. In previous operations, the time needed to strike a "time critical target" (e.g., a SCUD launcher not likely to remain in position for very long) was too long for effective action. A critical concern in developing new architectures for the JFACC is therefore the speed of response of the organization. Workload is not a primary concern. Instead, the focus is on optimizing the organization for quick reaction to changing information.

CONCLUSIONS

The TIDE model-based method for optimal team organizational design has shown promise for generating innovative team structures that can provide insight into how military organizations can best take advantage of changes in technology. With the enormous increases in network capability, many tasks in an organization can be done in almost any location, although some are still geographically constrained. The TIDE approach provides tools for working with subject matter experts in a domain to specify the tasks that must be accomplished, then producing optimized organizational structures for accomplishing those tasks. A major advantage of the approach is that it is not necessarily constrained by how things are done now, and can generate new ideas and new approaches. While these ideas may not be workable for a variety of practical reasons (e.g., the training costs for a new position may be too great), they provide a innovative starting point for rethinking military team structures. Initial empirical evidence indicates that the model-based approach has value for the Joint Task Force domain. Considerably more research and empirical testing is needed in other domains. Also, the applicability of the approach to the redesign of organizational structures in nonmilitary environments should be explored. Commercial organizations face many of the same problems as the military in adapting their organizational structures to take advantage of new technologies. The less-structured nature of many commercial missions and tasks is presenting new challenges for the TIDE method.

REFERENCES

Baker, C.V., Salas, E., Cannon-Bowers, J.A., & Spector, P. (1992). The effects of interpositional uncertainty and workload on team coordination skills and task performance. Presented at the annual meeting of the Society for Industrial and Organizational Psychology, Montreal, Canada.

Brannick, M.T., Salas, E., & Prince, C. (Eds.) (1997). *Team Measurement and Performance: Theory, Methods, and Applications*, Mahway, NJ: Lawrence Erlbaum Associates.

Bush, T., Bost, J.R., Hamburger, T., & Malone, T.B. (1998). *Optimized Manning on DD-21*. Presented at the International Symposium Warship '98: Surface Warships—the Next Generation. London, UK.

Cannon-Bowers, J.A., Salas, E., & Converse, S. (1990). Cognitive psychology and team training: Training shared mental models of complex systems. *Human Factors Bulletin, 33 (12)*, 1-4.

Entin, E.E. (1999). Optimized Command and Control Architectures for Improved Process and Performance. *Proceedings of the 1999 Command and Control Research and Technology Symposium,* Newport, RI.

Entin, E.E., Serfaty, D. & Kerrigan, C.K. (1998). Choice and performance under three command and control architectures. *Proceedings of the 1998 Command and Control Research and Technology Symposium,* Monterey, CA.

Kleinman, D. L. and Serfaty, Daniel (1989). Team Performance assessment in distributed decision-making. Proceedings of the Symposium on Interactive Networked Simulation for Training, pp. 22-27, Orlando, FL.

Levchuk, Y., Pattipati, C., and Kleinman, D. (1998). Designing Adaptive Organizations to Process a Complex Mission: Algorithms and Applications. *Proceedings of the 1998 Command and Control Research and Technology Symposium* (11-32) Naval Postgraduate School, Monterey, CA.

Lindblom, C.E. (1959). The science of "muddling through." *Public Administration Review, 19,* 78-88.

McIntyre, R. M., & Salas, E. (1995). Team performance in complex environments: What we have learned so far. In R. Guzzo & E. Salas (Eds.), *Team effectiveness and decision making in organizations* (9-45). San Francisco: Jossey-Bass.

Orasanu, J. M. (1990). Shared Mental Models and Crew Decision Making, CSL Report 46. Princeton, NJ: Cognitive Science Laboratory, Princeton University.

Rittel, H.W.J. & Webber, M.M. (1973). Dilemmas in a general theory of planning. *Policy Science, 4,* 155-169.

Salas, E., Bowers, C.A., & Cannon-Bowers, J.A. (1995). Military team research: 10 years of progress. *Military Psychology*, 7, 5575.

Salas, E., Dickinson, T. L., Converse, S. A. and Tannenbaum, S. I. (1992) Toward and Understanding of Team Performance and Training in *Teams: Their Training and Performance*, Eds. Robert W. Swezey and Eduardo Salas, Ablex Publishing Company, Norwood, NJ.

Serfaty, D. (1996). Adaptive Architectures for Command and Control (A2C2): An Overview. *Proceedings of the 1996 International Command and Control Research and Technology Symposium* (272-276) NDU, Washington, DC.

Serfaty, D., Entin, E. E., & Johnston, J. H. (1998). Team coordination training. In J. A. Cannon-Bowers & E. Salas (Eds.), *Making decisions under stress: Implications for individual and team training* (pp. 221-245). Washington, D. C.: American Psychological Association.

Serfaty, D., Entin, E.E., & Deckert, J.C., & Volpe, C (1993). Implicit coordination in command teams. In *Proceedings of the 1993 Symposium on Command and Control Research*, NDU, Washington, D.C, 53-57.

Swezey, R.W. & Salas, E. (Eds.) (1992) *Teams: Their Training and Performance*, Ablex Publishing Company, Norwood, NJ.

Travillian, K.K., Volpe, C.E., Cannon-Bowers, J.A., & Salas, E. (1993). Cross training highly interdependent teams: Effects on team processes and team performance. *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting,* 1243-1247.

Vicente, K.J., Burns, C.M., & Pawlak, W.S. (1997). Muddling through wicked design problems. *Ergonomics in Design, 5, 1,* 25-30.

Woods, D. (1997). Beyond Function Allocation: Transformation, Distributed Systems, Post Conditions, Proof by Contradiction. Presented at the 41st Annual Meeting of the Human Factors and Ergonomics Society.

This page has been deliberately left blank

Page intentionnellement blanche

Using of Fault Tolerant Distributed Clusters in the Field of Command and Control Systems

Prof. SERB AUREL, Ph. D.
Bd. George Cosbuc, Nr.81-83
Military Technical Academy
Bucharest-Romania

aserb@mta.ro

Prof. PATRICIU VICTOR VALERIU, Ph. D.
Bd. George Cosbuc, Nr.81-83
Military Technical Academy
Bucharest-Romania

vip@mta.ro

ABSTRACT: The open and distributed systems, that are the most important systems used in the field of command and control systems, must never fail. But only ideal system would be perfectly reliable and never fail. Fault tolerance is the best guarantee that high-confidence systems will not succumb to physical, design, or human-machine interaction faults.

A fault tolerant system is one that can continue to operate reliably by producing acceptable outputs in spite of occasional occurrences of component failures.

A fault tolerant cluster is a cluster with a set of independent nodes, connected over a network, and always with external storage devices connected to the nodes on a common input/output bus. The cluster software is a layer that runs on top of local operating systems running on each computer. Clients are connected over the networks to a server application that is executing on the nodes. The nodes of a cluster are connected in a loosely coupled manner, each maintaining its own separate processors, memory, and operating system. Special communications protocols and system processes bind these nodes together and allow them to cooperate to provide outstanding levels of availability and flexibility for supporting mission critical applications.

One of the most important problems in implementing fault tolerant system is the identification of single points of failure and elimination of these single points of failure by using replaceable units. When a component becomes unavailable, fault tolerant cluster software detects the loss and shifts that component's workload to another component in the cluster. The failure recovery is done automatically, without any human intervention.

The need for interoperability between the M&S world and the Command and Control world has been formulated in several publications. The challenge even increases when NATO and PfP Nations demands to train using their own simulation systems as well as their own command and control systems. The key issue for the Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance (C4ISR) community is the interoperability between live or real C4ISR systems and modeling and simulation systems.

Today, some of the architectural key words that are in common in the modern command and control systems and fault tolerant systems are the following:

-Open and distributed systems;
-Networks;
-High level operating systems;
-Hierarchical architecture;
-Interoperability and reusability;
-High availability systems.

1 INTRODUCTION

In the past, command and control systems were designed and developed to implement a constrained, prescribed system specification. Each organisation built systems to meet its own information requirements, with little concern for satisfying the requirements of others, or for considering in advance the need for information exchange. Incorporating the potential for growth and change were not essential criteria. Also, design of command and control systems was driven by available technology. Limitations in areas such as bandwidth, peripherals, and reliability of the components did not allow the realization of distributed systems that to be used for command and control.

But, command and control systems must continue to evolve and adapt to new requirements over an extended period of time. The architecture of these systems should be flexible enough to handle any component or subcomponent changes that are required to make the system, as a whole, meet all requirements, and to offer the flexibility to support multiple configurations of the architecture needed for specific objectives, and the ability to sustain changes in design over the program life cycle.

Now, advances in information technology permit that these systems to be designed around function and flexibility, not hardware. What is important this time in command and control systems design is not the technology, but the information exchange requirements, the mission organizational rules and functions.

2 FAULT TOLERANCE IN HIGH AVAILABILITY CLUSTERS

2.1 FAULT TOLERANCE

The open and distributed systems, which are the most important systems used for command and control, must never fail. But only ideal system would be perfectly reliable and never fail. This, of course, is impossible to be achieved in practice, because the systems fail for many reasons. Fault tolerance is the best guarantee that high-confidence systems will not succumb to physical, design, or human-machine interaction faults, or will allow viruses and malicious acts to disrupt essential services.

A fault tolerant system is one that can continue to operate reliably by producing acceptable outputs in spite of occasional occurrences of component failures.

The basic principle of fault-tolerant design is the use of redundancy, and there are three basic techniques to achieve fault tolerance: spatial (redundant hardware), informational (redundant data structures), and temporal (redundant computation).

Redundancy costs both money and time. The designers of fault-tolerant systems, therefore, must optimize their designs by trading off the amount of redundancy used and the desired level of fault tolerance. Computational redundancy usually requires recomputation and the typical result is a slower recovery from failure, compared to hardware redundancy. On the other hand, hardware redundancy increases hardware costs, weight, and power requirements.

The classical hardware and software fault tolerant techniques are modular redundancy, N-version programming, error-control coding, checkpoints, rollbacks, and recovery blocks.

2.2 REPLACEABLE UNITS

Modern systems are partitioned at several levels based on functions provided by specific subsystems. A fault-tolerant system displays similar functional partitioning, but in addition it contains redundant components and recovery mechanisms which may be employed in different ways at different levels. It is reasonable to view a fault-tolerant system as a nested set of subsystems each of which may display varying levels of fault tolerance. Recovery from a fault within a redundant partition may be effected within the domain itself, or may require action by higher levels within the system.

Fault tolerant architectures package these redundant partitions into replaceable units. A replaceable unit is a unit of failure, replacement and growth - that is, a unit that fails independently of other units, which can be removed without affecting other units, and can be added to a system to augment its performance, capacity, or availability.

The desired result of system partitioning and subsystem design is an integrated set of local, intermediate, and global fault tolerance functions that serve as a protective infrastructure to ensure the timely and correct delivery of system services.

2.3 CLUSTERS

A distributed system is a collection of computers (called nodes) that communicate with each other through a communication medium. Under the control of systems software, the nodes can co-operatively carry out a task. An open system allows system integration, so the customers can choose various hardware and software components from different vendors and integrate them to create a custom configuration suiting their needs and cost requirements.

A cluster is a set of loosely coupled, independent computer systems, connected over a network that behave as a single system. The cluster software is a layer that runs on top of local operating systems running on each computer. Client applications interact with a cluster as if

it is a single high-performance, highly reliable server. System managers view a cluster much as they see a single server. Most applications will run on a cluster without any modification at all. And only standard-based hardware components such as SCSI disks and Ethernet LANs are used to create a cluster.

Clustering can take many forms. A cluster may be nothing more than a set of standard personal computers interconnected by Ethernet. At the other end of the spectrum, the hardware structure may consist of high-performance symmetric multiprocessor systems connected via a high-performance communications and I/O bus. In both cases, processing power can be increased in small incremental steps by adding another commodity system.

If one system in a cluster fails, its workload can be automatically dispersed among the remaining systems. This transfer is frequently transparent to the client.

2.4 FAULT TOLERANT CLUSTERS

A fault tolerant cluster is a cluster with a set of independent nodes, connected over a network, and always with external storage devices connected to the nodes on a common input/output bus. Clients are connected over the networks to a server application that is executing on the nodes. The nodes of a cluster are connected in a loosely coupled manner, each maintaining its own separate processors, memory, and operating system. Special communications protocols and system processes bind these nodes together and allow them to cooperate to provide outstanding levels of availability and flexibility for supporting mission-critical applications. Fault tolerant clusters maintain strict compliance to the principles of open systems. There are no proprietary application programming interfaces that force vendor lock-in and require substantial development investment. Most applications will run on a fault tolerant cluster without any modification at all.

The top-level software of a fault tolerant cluster can be designed to maximize the flexibility of configurations within a local cluster. Clusters may be formed with a different number of nodes. This flexibility in system selection and cluster configuration protects customer investments in installed systems and allows the processing power of each node to be matched with the specific requirements of each application service.

Fig. 1 shows a sample configuration for a fault tolerant cluster.
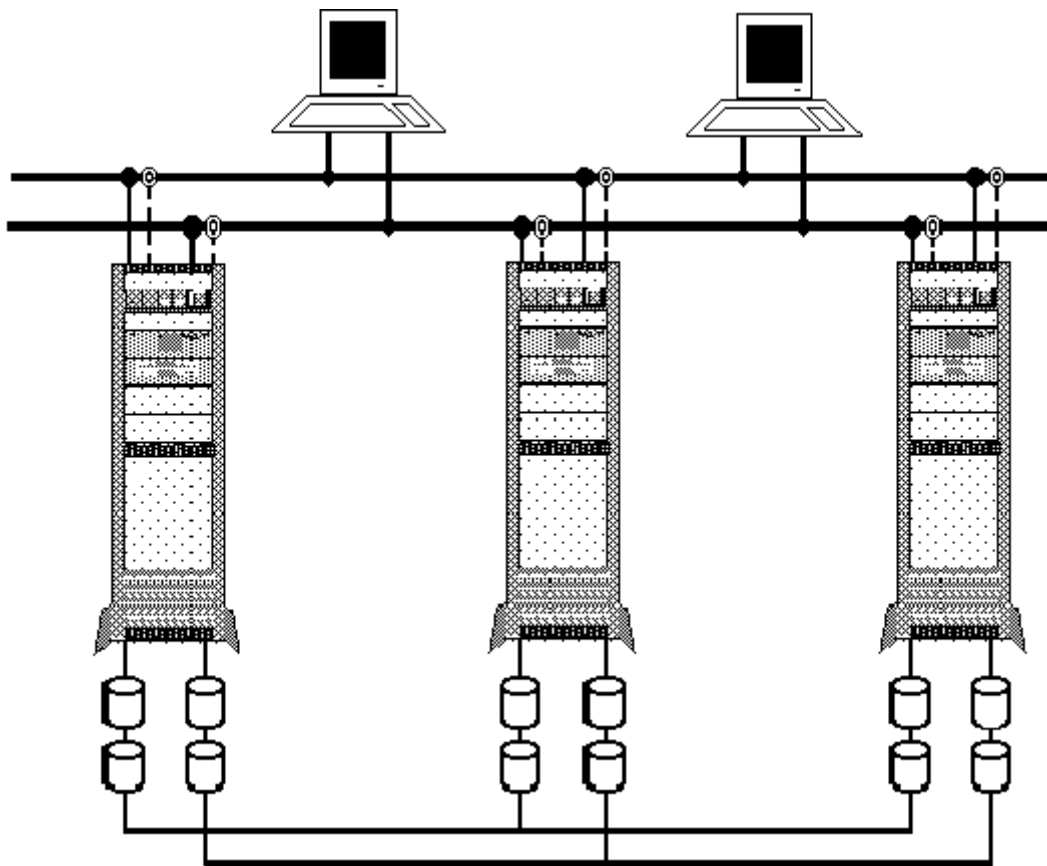
Fig. 1 A sample configuration for a fault tolerant cluster

If the failure of any component in a cluster results in the unavailability of service to the end user, this component is called a single point of failure for the cluster. One of the most important problems in implementing fault tolerant system is the identification of single points of failure and elimination of these single points of failure by using replaceable units.

The elimination of a single point of failure, by using replaceable units, always has a cost associated with it. Usually, what can be done is only to attempt to make a service highly available if the cost of losing the service is greater than the cost of protecting it.

The possible single points of failure that a cluster could have are:
-Nodes in the cluster,
-Disks used to store application or data, adapters, controllers and cables used to connect the nodes to the disks,
-The network backbones over which the user are accessing the cluster nodes and network adapters attached to each node,
-Power sources,
-Applications.

A high availability cluster is a grouping of servers having sufficient redundancy of software and hardware components that a failure will not disrupt the availability of computer services. To develop a complete high availability solution, is necessary to be maintained high availability within a hierarchy of system levels, some of which go beyond the cluster level. Failure at all levels must be detected quickly and a fast response provided. When a component becomes unavailable, fault tolerant cluster software detects the loss and shifts that component's workload to another component in the cluster. The failure recovery is done automatically, without any human intervention. At the same time, planned maintenance events at all levels must be possible with minimum disruption of service.

2.5 ELIMINATING NODES AS SINGLE POINTS OF FAILURE

The node in a fault tolerant system consists of a group of components, any of which can fail. The most important components are:
-One or more central processing units,
-Memory boards,
-Input/output controllers.

The use of cluster architecture lets the system eliminate a node as a single point of failure without losing service.

The nodes are connected to each other by a local area network, which allows them to accept client connections and to transmit messages that confirm each other's health. If one node fail, the failured node is removed

from the cluster and, after only a brief delay, its resources are taken over by the node configured to do so, so called the takeover node. This process is known as failover. The process of failover is handled by special high availability software running on all nodes in the cluster. Different types of clusters use different cluster management and failover techniques. There are specific differences in cluster types and their high availability software.

In fault tolerant clusters, disks containing data are physically connected to multiple nodes on a common input/output bus. When a node that owns a disk fail, a surviving node assumes control of the disk, so that the critical data remains available.

Clients and other devices are connected over the networks to the nodes. After the failover, all the clients and network devices connected to the failed node can access the second node as easily as the first. When a node failed during the running of a critical application, a takeover node can restart that application so that the service is not lost.

2.6 ELIMINATING DISKS AS SINGLE POINTS OF FAILURE

A fault tolerant solution improves data availability by allowing that a number of nodes to share the same hard disks within a cluster. When a node in the cluster fails, the fault tolerant cluster software will recover and disperse the work from the failed node to another node within the cluster. As a result, the failure of a system in the cluster will not affect the other systems, and in most cases, the client applications will be completely unaware of the failure.

Each node in a cluster has its own root disks, but each node may also be physically connected to several other disks in such a way that multiple nodes can access the same data. On such systems, this cluster-oriented access is provided by a software cluster component called Logical Volume Manager. Access may be exclusive or shared, depending on the kind of cluster created. Redundancy is necessary to prevent the failure of disk media or a disk controller.

Different fault tolerant configurations provide a range of solutions that address varying levels of fault protection requirements, including multi-site resiliency solutions. There are solutions that combine local fault tolerant cluster configurations with Fibre Channel mass-storage devices in order to provide a disaster-tolerant solution for a clustering environment up to 40 kilometers apart.

The most important two methods available for providing disk redundancy are: using disk arrays in a RAID configuration and using software mirroring. Each approach has its own advantages.

RAID (Redundant Array of Inexpensive Disk) is a disk technology that is designed to provide improved availability, security and performance over conventional disk systems. While appearing logically to the operating system as a single disk drive, a RAID array is actually made up of several disks, which have their data spread across the drives in any of several different methods. The group of disks that function together in a well-defined arrangement is known as RAID level. RAID Level 1 allows hardware mirroring, while others provide protection through the use of parity data. The RAID levels allow the array to reconstruct lost data if a disk fails.

In addition, arrays can be configured in independent mode, which means that each member of the array is seen as an independent disk.

Some of the advantages of using disk arrays for protected data storage are as follows:
-Data redundancy.
-Redundant power supplies and cooling fans.
-Online maintenance.
-Highest storage connectivity.
-Flexibility in configuration (different modes available).
-On some devices, dual controllers can eliminate additional single points of failure.

An alternative technique for providing protected data storage is the use of software mirroring, which allows that a single logical filesystem to be implemented on multiple physical copies in a way that is transparent to users and applications. It means that if a disk or sectors of a disk, containing a copy of the data, should fail the data will still be accessible from another copy on another disk. Note that the mirror copy is on a separate input/output bus. This arrangement eliminates the disk, the input/output card and the bus as single points of failure.

2.7 ELIMINATING NETWORKS AS SINGLE POINTS OF FAILURE

Networks are configured and used in clustered systems for access to an application by clients or other systems, and for communication between cluster nodes. In a fault tolerant cluster, the software establishes a communication link knows as a heartbeat. It is recommended that the fault tolerant cluster to be designed with more than one network, so that high level cluster software has at least one network at all times that it can use to monitor the status of cluster nodes.

This special use of networking must itself be protected against failures. Points of failure in the network include the LAN interfaces and cables connected to each node. In the cluster the entire communication link from the client system to the application server is subject to failures of various kinds. Depending on the type of LAN

hardware, failures may occur in cables, interface cards, network routers, hubs, or concentrators.

All these single points of network failure can be eliminated by providing fully redundant LAN connections, and by configuring local switching of LAN interfaces. To protect against network adapter failure, a second network adapter would be configured to the same network backbone. If the fault tolerant cluster is designed with more than one network, two network adapters will be used for all the network backbones. For eliminating the loss of client connectivity, can also be configured redundant routers or redundant hubs through which clients can access the services of the cluster. Another way to eliminate points of failure is to configure local switching, which means shifting from a configured LAN interface card to a standby.

2.8 ELIMINATING POWER SOURCES AS SINGLE POINTS OF FAILURE

Different methods can be used for eliminating power sources as single points of failure. The use of multiple power circuits with different circuit breakers reduces the likelihood of a complete power outage. An uninterruptible power supply provides standby in the event of an interruption to the power source. Small local uninterruptible power supply can be used to protect individual system processor units and data disks. Large power passthrough units can protect the power supply to an entire computer system.

2.9 ELIMINATING APPLICATIONS AS SINGLE POINTS OF FAILURE

The software of a fault tolerant cluster is a layer that runs on top of local operating systems running on each computer. The cluster management software provides services like as failure detection, recovery, load balancing, and the ability to manage the servers as a single system. This high level software monitors local hardware and software subsystems, tracks the states of the nodes, and quickly responds to failures in a way that eliminates or minimizes applications downtime, and provides a number of important other benefits, including improved availability, easier manageability, and cost-effective scalability.

The critical applications and data are housed on disk devices that are physically cabled to cluster nodes. This shared physical connection allows the ownership of shared logical volumes and their contents to be quickly switched from one node to another. Load balancing technique allows the performance of a server-based program to be scaled by distributing its client requests across multiple servers within the fault tolerant cluster. The load balancing management software can specify the load percentage that it will handle, or the load can be equally distributed across all of the hosts. If a host fails, the load balancing mechanism dynamically redistributes

the load among the remaining hosts. Load balancing technique is used to enhance scalability, which boost throughput while keeping response times low.

The high level software of a fault tolerant cluster operates in a fully transparent manner to both server applications and to TCP/IP clients. It lets users employ off-the-shelf software components, such as existing WWW, FTP, or proxy servers and other popular Internet applications, and enhances fault-tolerance and scales performance transparently to the TCP/IP protocol, to server applications, and to clients.

When a host fails or goes offline, the high level software of a fault tolerant cluster automatically reconfigures the cluster to direct client requests to the remaining computers. In addition, for load-balanced ports, the load is automatically redistributed among the computers still operating, and ports with a single server have their traffic redirected to a specific host. While connections to the failed or offline server are lost, once the necessary maintenance is completed, the offline computer can transparently rejoin the cluster and regain its share of the workload. This robust fault tolerance is enabled by a unique distributed architecture, which avoids the single points of failure or performance bottlenecks of other load balancing solutions.

If there is a node failure, it shuts down, and the cluster reconfigures itself; services that were on the failed node are made available on another system. There are different methods used for providing services after the shutting down of a node.

One way is to have another node that take over the applications that were running on the failed system. By using the high-level cluster software, application services and all the resources needed to support the application can be putted together into special entities called application packages. This application packages are the basic units that are managed and moved within the fault tolerant cluster. Packages simplify the creation and management of highly available services and provide outstanding levels of flexibility for workload balancing. When a package is failed over between nodes, all of the contents of the package are moved from the failed node to a new node. The ability to easily move application packages within a fault tolerant cluster provide outstanding availability during system maintenance activities such as hardware or software upgrades. Packages can be moved from node to node with simple operator commands, allowing scheduled maintenance to be performed on one node of a cluster while other nodes continue to provide support for critical applications. When the maintenance is complete, the node rejoins the cluster and assumes its normal workload of application packages. The same method can also be used to perform rolling operating system upgrades across a cluster.

Another approach for providing services after the shutting down of a node is to provide different instances of the same application running on multiple nodes so that when one node goes down, users need only reconnect to an alternate node.

In both cases, the use of clusters makes recovery from failure possible in a reasonably short time.

3 MODELING AND SIMULATION SYSTEMS AND COMMAND, CONTROL, COMMUNICATIONS, COMPUTERS, INTELLIGENCE, SURVEILANCE, AND RECONNAISSANCE SYSTEMS

3.1 INTEROPERABILITY

Warfighter battlespace is complex and dynamic, and information related to the battlespace must flow quickly among all tactical, strategic, and supporting elements. There is an unprecedented increase in the amount of data and information necessary to conduct operational planning and combat decision-making. The ability of the command, control, communications, computers, intelligence, surveillance, and reconnaissance systems supporting these operations to interoperate—work together and exchange information—is critical to their success.

Several elements are common to almost all systems used for modeling and simulation. According with the ideas of the NATO "Modelling and Simulation Master Plan" [8], simulation systems are used in the following application domains:
-Analyses;
-Simulation based Acquisition;
-Training and Exercises;
-Decision Support.

The need for interoperability between the Modeling and Simulation (M&S) world and the Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance (C4ISR) world has been formulated in several publications. The challenge even increases when NATO and PfP Nations demands to train using their own simulation systems as well as their own command and control systems. The key issue for the C4ISR community is the interoperability between live or real C4ISR systems and M&S systems.

According to Joint Publication 1-02, DoD Dictionary of Military and Related Terms, interoperability is defined in two context as follows:
(1) "The ability of systems, units, or forces to provide services to and accept services from other systems, units, or forces, and to use the services so exchanged to enable them to operate effectively together";
(2) "The condition achieved among communications-electronic systems or items of communications-

electronic equipment when information and services can be exchanged directly and satisfactorily between them and/or their users. The degree of interoperability should be defined when referring to specific cases"

Within the simulation community, the new and promising approach of using "High Level Architecture" and "Synthetic Environment Data Representation and Interchange Specification" is promoted to gain interoperability and reuse of the systems. For realising the interoperability, the C4ISR community is moving to standardize the Joint Technical Architecture (JTA) and the Defense Information Infrastructure Common Operating Environment (DII COE). Unfortunately, over the last decade, uncoordinated standards for M&S-to-C4ISR interoperability have been and are currently being developed by both communities.

3.2 HIGH LEVEL ARCHITECTURE AND SYNTHETIC ENVIRONMENT DATA REPRESENTATION AND INTERCHANGE SPECIFICATION

The intention of the High Level Architecture (HLA) and Synthetic Environment Data Representation and Interchange Specification (SEDRIS) is to provide a common architecture for modeling and simulation, and to offer a structure that will support the reuse and interoperability of simulations.

The High Level Architecture comprises three components: the rules, the interface specification, and the object model template.
-High Level Architecture Rules are a set of rules that must be followed to achieve proper interaction of federates during a federation execution. These describe the responsibilities of federates and of the Runtime Infrastructure in High Level Architecture federations.
-The Interface Specification defines the standard services and interfaces to be used by federates in order to support efficient information exchange when participating in a distributed federation execution and the reuse of the individual federates.
-The Object Model Templates prescribes the format and syntax for recording the information in High Level Architecture object models, for each federation and federate.

The basic components of the High Level Architecture are the simulations themselves, or more generally, the federates. The High Level Architecture requires that all federates incorporate specified capabilities to allow the objects in the simulation to interact with objects in other simulations through the exchange of data supported by services implemented in the Runtime Infrastructure. The Runtime Infrastructure is a distributed operating system for the federation which provide a set of general purpose services that support federate-to-federate interactions and federation management and support functions.

The Synthetic Environment Data Representation and Interchange Specification objectives are to:
-Articulate and capture the complete set of data elements and associated relationships needed to fully represent the physical environment.
-Support the full range of simulation applications (e.g., computer-generated forces, manned, visual, and sensor systems) across all environmental domains (terrain, ocean, atmosphere, and space).
-Provide a standard interchange mechanism to pre-distribute environmental data (from primary source data providers and existing resource repositories) and promote data base reuse and interoperability among heterogeneous simulations.

3.3 JOINT TECHNICAL ARCHITECTURE

The role of Joint Technical Architecture (JTA) is that of providing the foundation for interoperability among all tactical, strategic, and combat support systems. The JTA provides the minimum set of standards that, when implemented, facilitates the flow of information in support of the warfighter.

According to [10] the Joint Technical Architecture must be:
-A distributed information-processing environment in which applications are integrated.
-Applications and data independent of hardware to achieve true integration.
-Information-transfer capabilities to ensure seamless communications within and across diverse media.
-Information in a common format with a common meaning.
-Common human-computer interfaces for users, and effective means to protect the information.

Joint Technical Architecture wants to reach a consensus between a working set of standards and to establish a single, unifying technical architecture that will become binding on all future C4I acquisitions so that new systems can be born joint and interoperable, and existing systems will have a baseline to move towards interoperability.

3.4 INTEROPERABILITY BETWEEN M&S SYSTEMS AND C4ISR SYSTEMS

A key task for the M&S community is to link with live or real C4ISR systems. Within the C4ISR community there is a similar pressing need to link C4ISR equipment with simulations. Recent work promises the cooperation of these communities in developing a unified approach to linking simulations and C4ISR systems.

According to [11] in the near term simulation control is basically one-way with the simulations initializing the real C4I system databases. In the mid term, it can be expected to see the HLA linking constructive and virtual simulations on the simulation side and, via common

components found in C4ISR systems, the HLA also allowing simulations and C4ISR systems to exchange both data and messages. Simulation initialization will be two-way with real system databases providing information to the simulation side. Ultimately, it can be expected to have full two-way via common databases, thus achieving a measure of seamless interoperability. Finally, an interoperable M&S and C4ISR architecture is based on a common conceptual reference model accommodating common mediation techniques and shared data and object models, all linked via a common information management process providing common solutions for the C4ISR and simulation community.

4 MODELING AND SIMULATION SYSTEMS, COMMAND AND CONTROL SYSTEMS, AND FAULT TOLERANT CLUSTERS

Today, can be identified some key words that are in common in modern modeling and simulation systems, command and control systems and in fault tolerant clusters. Some of them can be:
-Open and distributed systems;
-Networks;
-High level operating systems;
-Hierarchical architecture;
-Interoperability and reusability;
-High availability systems.

4.1 OPEN AND DISTRIBUTED SYSTEMS

All modern systems used for command and control must be open and distributed systems. They must be flexible and extensible, able to be kept up-to-date with state-of-the-art technology, and to offer the best capabilities for reuse and interoperability. The architecture of all modern fault tolerant systems is that of a cluster, which is one of the best open and distributed system.

4.2 NETWORS

A fault tolerant cluster is a set of independent computers (nodes) connected over a network, and always with external storage devices connected to the nodes on a common input/output bus. Clients are connected over the networks to a server application that is executing on the nodes. The nodes of a cluster are connected in a loosely coupled manner, each maintaining its own separate processors, memory, and operating system. Special communications protocols and system processes bind these nodes together and allow them to cooperate to provide outstanding levels of availability and flexibility for supporting mission-critical applications.

The basic High Level Architecture protocol establishes that the communications path between any federates is over the network. There are no opportunities for back-channel data paths to corrupt the purity of the architecture. This rigor requires substantial effort to design the models in the federate and the common functions of High Level Architecture for each federate the interface data structure and the message transactions or services required for this highly object-oriented architecture. The resulting architecture, however, offers the flexibility to support multiple configurations of the architecture needed for specific modeling and simulation, and command and control objectives, and the ability to sustain changes in design over the program life cycle.

4.3 HIGH LEVEL OPERATING SYSTEMS

In a fault tolerant system the nodes of the cluster are connected in a loosely coupled manner, each maintaining its own operating system. The cluster software is a layer that runs on top of local operating systems running on each computer. The high availability applications in the fault tolerant cluster run at the top level cluster software.

In the High Level Architecture the Runtime Infrastructure is defined as a distributed operating system for federates and federations.

The top-level cluster software can be a good support for Runtime Infrastructure, and for the command and control systems.

4.4 HIERARCHICAL ARCHITECTURE

A complex simulation or command and control centers must be considered as a hierarchy of components of increasing levels of aggregation. At the lowest level is the model of a system component. This may be a mathematical model, a discrete-event queuing model, a rule-based model, etc.

In HLA the model is implemented in software to produce a simulation. When this simulation is implemented as part of an HLA-compliant simulation, it is referred to as a federate. HLA simulations are made up of a number of HLA federates and are called federations. Simulations that use the HLA are modular in nature allowing federates to join and resign from the federation as the simulation executes.

Based on functions provided by specific subsystems, all fault-tolerant clusters are partitioned at several levels, but in addition it contains redundant components and recovery mechanisms which may be employed in different ways at different levels. It is reasonable to view fault-tolerant clusters as a nested set of subsystems, each of which may display varying levels of fault tolerance. Recovery from a fault within a redundant partition may be effected within the domain itself, or may require action by higher levels within the system.

At top of any fault tolerant cluster, command and control, and High Level Architecture compliant system there is a distributed operating system that runs on top of local operating systems running on each computer or on top of federates and federations.

4.5 INTEROPERABILITY AND REUSABILITY

More work is needed today for being in the happy case of buying hardware and software components from various vendors, integrating them in applications to form complete systems. There is a need for software components to be able to communicate with each other using "standard" mechanisms and "open" interfaces for an effective integration to occur. As software and hardware systems get more complex, the need for interoperability among different components becomes critical.

The High Level Architecture can be conceptual "software bus" that allow applications to communicate with one another, regardless of who designed them, the platform they are running on, the language they are written in, and where they are running. High Level Architecture also enables the building of a plug-and-play component software environment.

The fault tolerant cluster can offer a good architecture for command and control systems and High Level Architecture compliant systems to work with these applications. The fault tolerant cluster is an open and distributed system, flexible and extensible, and its architecture offer compliance to the principle of reusability and interoperability.

4.6 HIGH AVAILABILITY SYSTEMS

The military systems must not succumb to different faults and must continue to operate reliably in spite of occasional occurrences of component failures. A fault tolerant solution improves data and applications availability by allowing that a number of nodes to share the same hard disks within a cluster. When a node in the cluster fails, the fault tolerant cluster software will recover and disperse the work from the failed node to another node within the cluster. As a result, the failure of a system in the cluster will not affect the other systems, and in most cases, the client applications and data will be completely unaware of the failure.

In command, control, communications, computers, intelligence, surveillance, and reconnaissance systems information related to the battlespace is complex and dynamic and must flow quickly among all tactical, strategic, and supporting elements. There is an unprecedented increase in the amount of information necessary to conduct operational planning and combat decision-making. For the command and common systems, and for the modeling and simulation systems high availability of data and applications is very important.

Fault tolerance is the best guarantee that the systems will be available, and the essential services will be offered in real-time to the users. The modern fault tolerant clusters are able to eliminate all single points of failure in the nodes of the cluster, the disk used to store applications or data, the networks, the power sources, the data, and the applications and to offer the best high availability architecture for command and control systems, and modeling and simulation systems.

REFERENCES

[1]. AUREL SERB
"Sisteme de calcul tolerante la defectari"
Academia Tehnica Militara.
Bucuresti, 1996

[2]. DAVID A. PATTERSON and JOHN L. HENNESSY
"Computer Organization & Design. The Hardware/Software Interface"
Morgan Kaufmann Publishers, Inc.
San Francisco, California, U.S.A., 1998

[3]. ANDREW S. TANENBAUM
"Structured Computer Organization", 4^h Ed.
Prentice-Hall, Inc.
New Jersey, 1999

[4]. PETER WEYGANT
"Clusters for High Availability. A Primer of HP-UX Solutions".
Prentice Hall Pt.,
Upper Saddle River, New Jersey, U.S.A., 1996

[5]. High Level Architecture Interface Specification, v1.3.
Defense Modeling and Simulation Office.
5 February 1998
http://hla.dmso.mil/hla/tech/ifspec/if1-3d9b.doc

[6]. High Level Architecture Object Model Template, v1.3.
Defense Modeling and Simulation Office.
5 February 1998
http://hla.dmso.mil/hla/tech/omtspec/omt1-3d4.doc

[7]. High Level Architecture Rules, v1.3.
Defense Modeling and Simulation Office.
5 February 1998
http://hla.dmso.mil/hla/tech/rules/rules1-3d2b.doc

[8]. Modeling and Simulation (M&S) Master Plan.
Department of Defense.
October 1995
http://www.dmso.mil/dmso

[9]. Levels of Information Systems Interoperability
(LISI).
Architectures Working Group.
30 March 1998

[10]. Joint Technical Architecture. Version 4.0 Draft 1.
Department of Defense
14 April 2000

[11]. JOSEPH LACETERA and DON TIMIAN
Interim Report Out of the C4I Study Group
Simulation Interoperability |Workshop
26-30 March 2000

[12]. AUREL SERB
Fault tolerance in systems used for computer assisted
exercises.
NATO's Research & Technology Organization PfP
Symposium on Computer Assisted Exercises for Peace
Support Operations,
The Hague, the Netherlands, 28-30 September 1999.

Security Architectures for COTS based Distributed Systems

Pierre Bieber, Pierre Siron
ONERA-CERT
BP 4025, 2 avenue E. Belin
31055 Toulouse Cedex 4
France
Pierre.Bieber@cert.fr, Pierre.Siron@cert.fr

Abstract

The paper describes two experiments in the design of security architectures for distributed systems that are implemented with Commercial Off The Shelf components.We added security components to protect information exchanged in a Distributed Interactive Simulation environment. We added a role-based access control component to a Workflow tool implemented with CORBA technologies. The two experiments followed the same approach that includes four steps (threat analysis, security policy definition, selection of security components and architecture efficiency evaluation).

1   Introduction

Economical incentives are forcing the use of COTS (Commercial Off The Shelf) components in the design of complex distributed systems in the military sector. In this context, we are interested in securing COTS based systems.

The use of COTS components introduces several difficulties. First, we cannot use existing COTS components to enforce security because COTS components generally offer limited security services. Another difficulty is that we have to guarantee security even if we lack a detailed knowledge of how the components are implemented. COTS components developers often provide poor technical documentation. COTS component source code is generally not released with the notable exception of "open source software". Although this new breed of components is very promising we did not consider it in this paper because the systems we studied did not rely on free components. Finally, a common feature of COTS based system is that security is addressed when the development of the system is almost finished. Hence, the components used by the system can no longer be modified to guarantee security

Our job is to analyse the "software architecture" of a distributed system in order to extend it with components that guarantee security. By software architecture we mean a description of the components the system is made of and a description of how these components interact. Generally a description of how components are physically distributed over a network is available. But we need a more detailed description such as an object model that would list the classes of objects the system use and a set of scenarios that would explain how the objects interact.

In the following of this paper we describe our approach to design security architectures for COTS based distributed systems. We illustrate our approach with the security architectures we designed for two distributed systems. The first one is a Distributed Interactive Simulation environment. This system let interoperate simulations developed by various companies or military organisations. This kind of system is used to simulate new weapons or to conduct virtual manoeuvres. The second system is a collaborative workflow tools. This system let users locate relevant resources to perform a task. This system can be used to mechanise administrative tasks.

In the first part of the paper we explain the common steps in the design of a security architecture. Then we illustrate them on the Distributed Interactive Simulation and the Collaborative Workflow tool. Finally we discuss the similarities of these two security architectures and we describe the impact of the new compoents on original security architectures.

2   Design of Security Architectures

The first step is the analysis of the software architecture in order to find out what threats could be applied to the system. We consider two classes of threats:

- Attacks directed at the interaction between components. of a system. During their transit on the communication links between components (such as the Internet or a LAN) messages are subject to possible eavesdropping, or even worse blocking, replaying or forging.
- Attacks directed at the components. As components of the system were not developed with security requirements in mind, they could be used in order to disclose confidential data or in order to modify illicitly critical information.

We have to consider what attacks should really be taken into account. This involves assuming that some components are trusted not to perform attacks. For instance, some interaction links will be considered as secure because no component will try to listen to it. These assumptions generally remain unverified. One possibility would be to certify the components with respect to security evaluation criteria such as the ITSEC [15] or the new Common Criteria [12]. But the certification process is very expansive. So this option does not seem compatible with the cost-reduction imperatives that led us to rely on COTS components.

The second step is the definition of the security policy that lists the security objectives that should be satisfied. An obvious security objective is that all the attacks should be appropriately countered. So we have to define what information disclosure or modification are authorised with respect to the application under consideration. Although in previous papers, we have considered security policies in the context of multi-company co-operation where the main security objective is the protection of "Intellectual Property" and hence a confidentiality concern. In this paper we stress on applications where data has to be shared by civilian and military organisations. This involves both a confidentiality requirement: military private data should not be disclosed to civilian components and an integrity requirement: civilian components should not introduce erroneous information in military components.

The third step is the selection of components that enforce the security objectives that were selected during the previous step. For economical reasons we will try to use SCOTS (Security Components Off-The Shelf) when possible and try to limit the development of ad-hoc security components. SCOTS that protect the communication links implement security protocols such as Secure Socket Layer (SSL) [?] or Generic Security Service (GSSAPI) [7]. To implement access controls we can use SCOTS that perform IP packet filtering included in Firewall [17] toolkits such as Gauntlet from TIS.

The fourth and final step in the design of a security architecture is the evaluation of its efficiency. As stated previously, we think that it is unlikely that individual components of a COTS based distributed system will be certified. But, ITSEC evaluation principles could be applied in order to assess the assurance-efficiency of a security architecture. A completeness analysis could be conducted to see whether the threats are correctly by the selected security components. A consistency analysis could test whether the security components interact properly. Finally, an impact analysis could measure what is the impact of the new components on the original architecture.

3   Security Architecture for HLA/RTI

3.1   CERT HLA/RTI prototype architecture

ONERA/CERT has developed a prototype of distributed interactive simulation environment conforming to the HLA RTI standard (see [1] and [5]). In the following, we will use federate to denote an individual simulation and federation to denote a group of federates. Our prototype (see [2] and [3]) is made of several components: each federate interacts with a RTI Ambassador component (RTIA) and RTIA components interact with the RTI Gateway (RTIG) component. The RTI architecture is depicted by Figure 1.
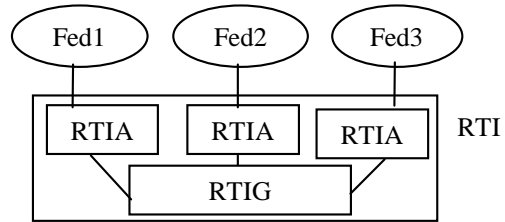
Figure 1.   RTI architecture

The RTIA components are processes that exchange messages over the network, in particular with the RTIG process, via TCP (and UDP) sockets, in order to run the various distributed algorithms associated with the RTI services. RTIG is a centralisation point in the architecture. It uses the Federation Object Model (FOM) that describes the classes of data that federates can exchange during an execution. The RTIG records the identity of federates willing to publish data belonging to a class of the FOM or subscribe to a class of the FOM. The RTIG uses this information to forward messages in a multicast approach.
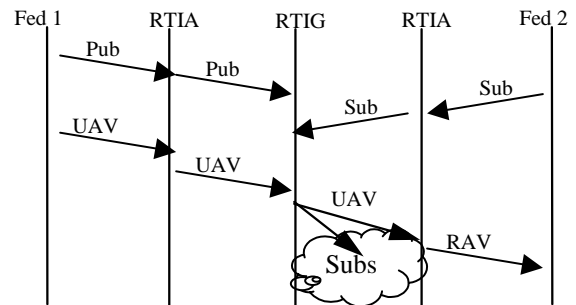
Figure 2.   RTI data-transfer scenario

Figure 2 illustrates the message exchanges involved when federate 1 wants to inform other federates of the new value of an object. We suppose that in a previous step, federate 1 informed the federation that it was willing to publish values for that a class of objects in the FOM. An UAV (Update Value) message is sent to the RTIA and forwarded to the RTIG. The RTIG forwards this message to the RTIA of federates that subscribed to this class. If federate 2 has subscribed to this class, its ambassador will send it a RAV (Reflect Value) message as soon as the delivery condition holds.

3.2   Threat Analysis

We are considering a federation where simulations of both civilian and military organisation interoperate. We call FedM federates belonging to the military organisation and FedC the civilian federate.
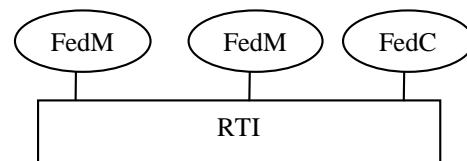
Figure 3.   A dual federation

A detailed security analysis was performed, it is described in [3]. We distinguish two channels that Civilian federates can use to obtain Military information.

The first disclosure channel is related to attacks directed at the communication links between components of the RTI. It is likely that organisations will prefer that their federates run on hosts that belong to their local area network. These federates have to use a Wide Area Network such as the Internet to exchange messages with other components of the RTI. We assume that a federate and its RTIA process are executed on the same host so that we will not consider that the interaction link between them might be attacked. So we should only protect the communication link between a RTIA and the RTIG.

The second channel is the leak of information via the RTI services. A malicious federate could try to use its services in order to gain knowledge about a private object. One insecure scenario occurs when federate FedC subscribes to a class that happens to be regarded as private by the Military. The normal behaviour of RTI will be to forward values of objects in the private class to federate FedC. So HLA/RTI data transfer services could be used to disclose confidential data. We suppose that the RTIG process is under the control of a third party that is trusted by all the organisations. For instance, this third party could be a public organisation running a simulation facility. So the RTIG will not disclose private date intentionally. An organisation may trust the federate component it has written, it might also trust components of the RTI such as its RTIA or RTIG. But it would certainly not trust federate components developed by other organisations, so we should forbid a federate from one organisation to use the RTI services in order to learn private information belonging to another organisation.

3.3 Security Policy

In order to limit data exchanged using HLA/RTI data transfer services such as Update Value / Reflect Value, we propose to associate a security label with objects and federates of the federation. The RTI will control the messages according to the security labels of the object and of the federate. Security labels we have considered are PrivateMil and Public. PrivateMil dominates security label Public.

The description of the federation must be completed with Federation Execution Data that include security label information. Figure 4 gives an example of security label association. In this federation, the class Aircraft is public whereas its sub-class Fighter is regarded as Private by the Mil organisation. We consider two federates: one modelling an air traffic controller that has security label Public and another one that models a military controller that has security label PrivateMil.

```
(fed
(objects
(class "Aircraft"
      (sec_level "Public")
      (attribute "Position")
    (class "Fighter"
      (sec_level "PrivateMil")
      (attribute "WeaponLoad")
…)))
(federate              "AirTrafficControl"
"Public")
(federate "MilControl" "PrivateMil")
…)
```

Figure 4: Federation security labels

The RTI should allow the Air Traffic Controller federate to observe the current position of a fighter but the RTI should not authorize this federate to know the current status of the weapon loaded on the fighter.

3.4 Selection of Security Components

With regard to the first threat, the security function needed should build a secure association between the RTIA and the RTIG. This association should guarantee authentication and confidentiality of the communication. The second threat can be resolved by adding access controls within the RTI services that should restrict the message exchanged between federates of two companies.
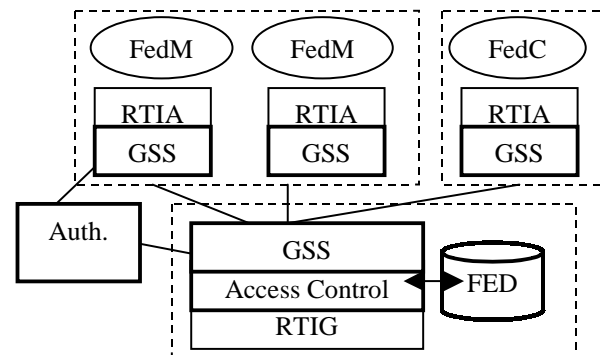
Figure 4. HLA/RTI Security Architecture

3.4.1 GSS-API security services

In a WAN context, protecting communication links can be done by using cryptographic techniques. Rather than developing a new cryptographic protocol, we selected the Generic Security Services Application Programming Interface (GSS-API) [7]. It is an Internet Engineering Task Force standard that defines cryptographic services that are useful to secure a client-server application. The services include the management of encryption keys, the distribution of shared key or using distributed keys to enforce the confidentiality, authentication or integrity of exchanged messages.

The GSS-API interface hides the details of the underlying security mechanism leading to better application portability. In particular, this would allow changing the underlying security mechanism if a security flaw is encountered in it. Several popular security protocols offer a GSS-API interface such as Kerberos [8] or SESAME [9].

We used the GSS-API implementation developed at DTSC in Australia [10]. GSS-API was integrated to the RTIA and RTIG processes to secure their communication. We extended the Socket class that is used to exchange any messages within RTI. Ambassadors of remote federates will use sub-class SecureSocket that includes the appropriate calls to the GSS-API services whereas Ambassadors of local federates will use class Socket.

The SecureSocket class hides several steps that should be performed to protect a communication link. Prior to the execution of a federation, every federate ambassador has to contact the Authentication server that will provide the ambassador with credentials (an encryption key tied together with the identity of the federate and a date). When a federate wants to join a federation its RTIA has to contact the RTIG and authenticate itself using the credentials. This involves several exchanges of messages between RTIA and RTIG. If this step succeeds, a security association is created between RTIA and RTIG (a session key is distributed to both processes). Then, RTIA and RTIG can protect the HLA/RTI messages they exchange by using cryptographic functions. For instance, if integrity has to be enforced an elaborate message digest (the MD5 algorithm is used) will be appended to each message, if confidentiality has to be enforced all exchanged messages will be encrypted (the DES algorithm is used) with the distributed session key.

3.4.2   Publish/subscribe access controls

Publication and subscription messages are controlled rather than data-transfer messages. A federate will be authorised to subscribe to a class if its security label dominates or is equal to the security label of the requested class.

This control is performed by the RTIG because, in our architecture, this component is already in charge of recording the publication and subscription. So the RTIG will now check the security labels of the federate and of the class whenever this federate has sent a subscription message for this class. The RTIG will record for each published class a list of authorised subscribers. As the RTIG transmits Update Value messages only to authorised subscriber RTIA, a federate from one company will never receive Reflect Value messages for a private object of another company because its subscription request are blocked by the security label control in the RTIG.

The RTIG is extended with new services that manage the security labels. These services use the Federation Execution Data to associate security labels with federates as they join a federation, and with the classes in the FOM. Furthermore, a function comparing two security labels has

been implemented (this function is quite independnet from the security labels used in the security policy), it is used to check whether a subscription message should be discarded. In this case, we generate an exception.

4   Security Architecture for CIDRIA
4.1   CIDRIA Architecture

The CIDRIA workflow tool was developed by France Telecom/CNET. It is described in [11]. CIDRIA is implemented as a CORBA [16] server that handles various objects : *Agent* objects that represent users or resources and a *Cooperator* object that supervises Agent objects.

One benefit of using CORBA to implement CIDRIA is that each class is associated with an IDL interface that lists the methods it offers. In the following, we illustrate how these methods can be used. For that purpose, we suppose that CIDRIA is used to implement a flight planning system. Pilots would use this system to prepare their flight mission. A map provider resource called IGN is connected to CIDRIA as well as a commander that will give the mission goal to the pilots. We suppose that the map provider is a civilian organisation whereas pilots and commanders belong to a military organisation
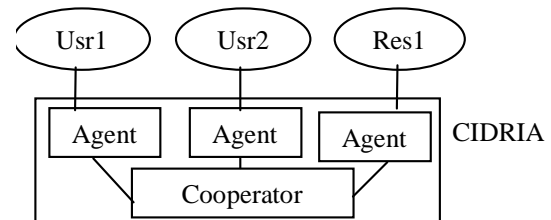
Figure 5.   CIDRIA architecture

First of all, when a resource or a user client logs on CIDRIA, it invokes method `connect` of object Cooperator. This creates an object of class Agent, then the client interacts exclusively with this object. When a pilot wants to prepare a mission, it will invoke method `add_request` of its agent with parameter `map` to request a map. The agent submits this request to the cooperator that tries to locate an agent that could provide a map.
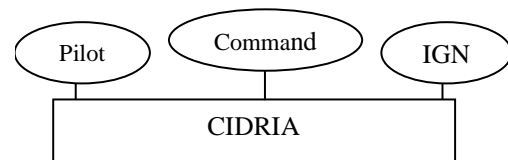
Figure 6.   Flight planning application

The Cooperator is able to locate resource IGN if it previously registered its competence on the topic map using method `add_competence` of its agent. Then the request is sent to IGN that can decide to answer it (and provide the requested map) or not using methods

reply_request or reject_request of its agent. If the request is rejected the Cooperator will try to locate another resource providing maps.
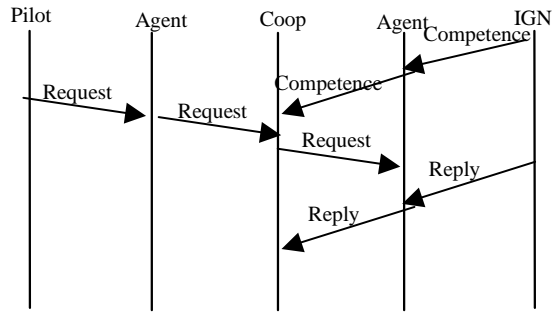
Figure 7.   Request resolution scenario

4.2   Threat Analysis

As in the case of HLA/RTI we consider two disclosure channels. The first disclosure channel is related to attacks directed at the communication link between CIDRIA and its clients. In the case of CIDRIA, messages exchanged between CIDRIA and the clients contain requests, responses or object localisation information. Hence, these attacks could allow an intruder to invoke any method of CIDRIA. We should protect the communication link between a client and the CIDRIA.

The second channel is the leak of information via CIDRIA services. We suppose that the CIDRIA server is operated by an organisation that does not trust its clients to behave correctly. A malicious client could try to use its services in order to gain knowledge about a private data or to provide erroneous information. For instance, a malicious client could claim to be a commander using method add_competence then the normal behaviour of CIDRIA is to forward the pilot requests for mission goals. So a malicious client could reply to these requests and provide erroneous mission goals to the pilots. A malicious client could also send a request for mission goals using method add_request with parameter goal. CIDRIA would forward the request to the commander. If the commander replies to the request, then a malicious client could learn what is the mission goal, which is likely to be a confidential information.
We should control that a client is authorised to claim a competence on a topic. For instance, a civilian client is not authorised to claim its competence on mission goals. We should also control that a client is authorised to request data on a topic. For instance, a civilian client is not authorised to ask for mission goals.

4.3   CIDRIA Security Policy

We selected "Role-based Access control" RBAC policy (see [13] and [14]) to describe CIDRIA security objectives. In a RBAC policy, a set of roles is associated with each user according to the tasks this user should perform within the organisation. A role is set of operations that a user playing this role is authorised to perform.

In order to define CIDRIA roles we consider that CIDRIA operations are methods that appear in the interface of classes Cooperator or Agent with their parameters. Example of methods is add_competence or add_request, and examples of parameters are: maps or goals. We define the following roles: Pilot, Commander and MapProvider. A Pilot is authorised to invoke method add_request with parameter maps or goals. A commander is authorised to invoke method add_competence with parameter goals. And a Map Provider is authorised to invoke method add_competence with parameter maps.

A role access control component checks that a client is authorised to play the role it selects when it connects to CIDRIA. And the access control component should control that the method invoked by a client is authorised according to the role it is playing.

4.4   Selection of security components

The threat analysis showed that components that protect communications between a client and CIDRIA should be added. We could select the GSS-API component just as in HLA/RTI. So we do not detail the protection of communication links for CIDRIA and we focus on access control components.
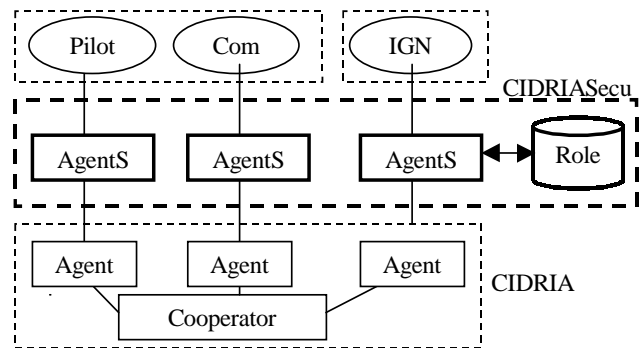
Figure 8.   Secure Components added to CIDRIA

4.4.1   Role-based Access Control Server

The CORBA standardisation body has defined a set of security services that include access-control. We decided not to use it because these services are not currently available in the most common CORBA implementations. Instead we decided to define a new server called CIDRIASecu that manages two new classes CooperatorS and AgentS. The interfaces of these classes are similar to class Cooperator and class Agent interfaces. Our goal is that clients invoke methods of the new CIDRIASecu component instead of CIDRIA server methods. When a client invokes a method of CIDRIASecu, this method first checks that the method with its parameters is an

authorised operation according to the role that the client is playing. If the operation is authorised then the synonymous method of server CIDRIA is called otherwise an exception is generated.

The access controls performed by the methods of class AgentSecu are based on the description of the roles. A Prolog database contains predicates that describe authorised operations for a role, and authorised roles for a client. The prolog database is encapsulated within a C++ object that can be used by all the AgentS components.

4.4.2 Border Access Control: Firewall and Wonderwall

In the last section, we made the assumption that clients would invoke the methods of server CIDRIASecu rather than the methods of CIDRIA. Of course a malicious client would certainly try to bypass the controls performed by CIDRIASecu and try to invoke directly the methods of CIDRIA. To avoid this situation, we propose that CIDRIA and CIDRIASecu servers run in a protected area called an enclave. Inside the enclave all the components are trusted and outside the enclave the components such as the clients are supposed to be malicious.

So we should add functions that protect the border of the enclave (i.e. the interactions between clients and the components inside the enclave should controlled). We want to forbid clients to invoke methods of server CIDRIA. For that purpose, we used a SCOTS called Wonderwall [18] (produced by IONA). Wonderwall controls what objects are accessed on a server and what methods are invoked on these objects. In our security architecture, Wonderwall will allow accesses to objects in classes AgentS and CooperatorS and it will forbid accesses to objects in classes Agent and Cooperator. We added a Firewall that filters any communication whose destination is not the Wonderwall.
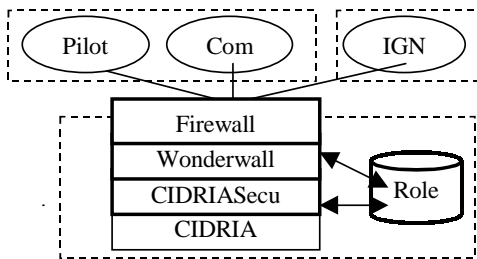
Figure 9.   CIDRIA Security Architecture

The previous figure shows the interaction between a client, CIDRIASecu and CIDRIA. A request from the client proceeding from Internet goes through the firewall then Wonderwall filters it. If the request is authorised, Wonderwall forwards it to CIDRIASecu. If the request passes the role controls, CIDRIASecu forwards it to CIDRIA. The response to this request takes the same path in the reverse way.

5   Conclusion: Security Architecture Evaluation

In this concluding section, we try to evaluate efficiency of the two architectures we have designed. We first explain the similarities between both security architectures then we evaluate their impact on the original architectures.

5.1   Similarities of the two security architectures

The two systems we have considered could be regarded as rather simple instances of COTS based system. Both systems were designed in an object-oriented way, hence we could use an appropriate description of their software architecture. Furthermore, the systems were developed by partners (ourselves for HLA/RTI and France Telecom/CNET for CIDRIA) so we could get all the details we needed to know on the behaviour of the components. Nevertheless, we applied our approach to secure systems according to the rules stated in the introduction: we did not modify any existing components nor did we use our knowledge of the components implementation. We mainly added new components. We used several SCOTS: a firewall, GSS-API security protocol and the Wonderwall CORBA messages filtering tool. So we believe, that our approach could be valid in less forgiving cases. For instance, our approach would still be valid if we totally ignore the source code of components.

For both systems we had to develop ad-hoc security components that control whether the services offered by the distributed system are correctly used. It was not possible to use SCOTS because these security components depend on the application security policy we want to consider and on the services of the distributed system we want to protect.

5.2   Impact

Once the security architecture was designed and implemented we were able to study the impact of the new components we added on the overall behaviour of the system. As the two applications have quite different purposes we did not study the security impact in the same way. HLA/RTI is rather focused on real-rime performances whereas CIDRIA is an information system tool where ease of management matters.

To assess the impact of security components on the real-time performances of HLA/RTI we used a toy federation modelling three aircraft trajectories. We compared the number of simulation steps performed by the simulation with and without security components.

As access control is only performed during the subscription stage of a federation, this component has no influence on the main stage of a federation that generally involves a lot of data-transfer message exchanges. So we have not observed any significant decrease of real-time performances when we added access controls.

The first experiments we made showed a 25% decrease of the performances when GSS-API is used to protect message integrity and much more when confidentiality is protected. The federates use very simple trajectory computation and communicate a lot, so the performance of SecureSocket functions have a huge impact on the overall performance of the federation. So we expect that performance decrease would not be as great in a realistic simulation with federates performing more complex computations.

The CIDRIA tool has a complex administration procedure that should be performed before clients connect to the server. An administrator client should invoke CIDRIA methods in order to create or destroy agents, to add or remove request topics. The CIDRIASecu server we added should also be administered in this way, furthermore the server needs a description of the roles. So we had to introduce a security officer that invokes CIDRIASecu methods that add, modify or remove a role, an operation or an identity. This security officer is also in charge of creating instances of AgentS objects and creating request topics that are managed by CIDRIASecu. The resulting administrative procedure is cumbersome so we decided to simplify. When a role is defined by the security officer, an instance of Agent and Agent S are automatically created. Similarly when operation is defined with a new request topic it is automatically added in CIDRIA and CIDRIASecu.

6   Acknowledgements

7   References

[1] Department of Defense, "High Level Architecture Interface Specification, Version 1.3 Draft 7", January 1998.

[2] P. Siron, "Design and Implementation of a HLA RTI Prototype at ONERA", the Fall Simulation Interoperability Workshop, 1998.

[3] P. Bieber, J. Cazin, P. Siron, G. Zanon "Security extensions to ONERA HLA RTI Prototype", the Fall Simulation Interoperability Workshop, 1998.

[5] DMSO, "HLA/RTI web page", http://hla.dmso.mil

[7] J. Linn, "Generic Security Service Application Programming Interface", Internet RFC 2078, January 1997.

[8] J.Steiner, B. Neuman, J. Schiller, "Kerberos: An Authentication Service for Open Network Systems", proceedings of The USENIX Winter Conference, February 1988.

[9] P. Kaijser, J. Parker, D. Pinkas, "SESAME: The solution to security for Open Distributed Systems", Computer Communications, July 1994.

[10] D.P. Barton, L.J. O'Connor, "Implementing Generic Security Services in a Distributed Environment", Technical Report, CRC for Distributed Technology, Brisbane, Australia, April 1995.

[11] Bruno Dillenseger, François Bourdon, « Modélisation de la coopération et de la synchronisation dans les systèmes d'information – Une expérience de Workflow basée sur les nouvelles technologies », Calculateurs parallèles, volume 9, n 2, 1997.

[12] Common Criteria, http://csrc.nist.gov/cc

[13] Ravi Sandhu, Edward Coyne, Hal Feinstein, Charles Youman, « Role-based Access Control Models », Computer, February 1996, IEEE Computer Society Press,.

[14] D. Ferraiolo, J. Cugini, R. Kuhn, « Role Based Access Control: Features and Motivations », Proceedings 10th Annual Computer Security Applications Conference, IEEE Computer Society Press, 1994. http://waltz.ncsl.nist.gov/rbac/rbac/newpaper/rbac.html

[15] ITSEC, Information Technology Security Evaluation Criteria, Union Européenne, 1991.

[16] OMG, "CORBA Services". http://www.omg.org

[17]D. Brent Chapman, Elizabeth D. Zwicky, « La sécurité sur Internet – Firewall », Editions O'Reilly International Thomson, 1996.

[18] IONA, "Wonderwall", http://www.iona.com/products/orbix/wonderwall.html

This page has been deliberately left blank

Page intentionnellement blanche

Design Aspects in a Public Key Infrastructure for Network Applications Security

(August 2000)

Prof. Dr. VICTOR-VALERIU PATRICIU
Cdor.Prof. Dr. AUREL SERB
Computer Engineering Department,
Military Technical Academy,
Bd. G.Cosbuc nr.81-83, sect.4, Bucharest,
Romania

Abstract: Computer security is a vitally important consideration in modern systems. Typically, the military and banking areas have had detailed security systems. This paper will concentrate on an interesting area of software security based on public key cryptographic technology. The Public Key system makes it possible for two parties to communicate securely without either having to know or trust the other party. This is possible because a third party that both the other parties trust identifies them, and certifies that their keys are genuine. This third party is called the Certification Authority, or CA. CA guarantees that they are who they claim to be. The CA does this by registering each user's identification information, and issuing them with a set of Private keys and a set of Public Key Certificates. A worldwide Public Key Infrastructure (PKI) that supports international, government, and state policies/regulations will not be available in the near future. In the meantime, organizations and corporations can utilize this security technology to satisfy current business needs. Many organizations are choosing to manage their own Certificate Authority (CA) instead of outsourcing this function to a third party (i.e., Verisign, Thawte, GTE CyberTrust, GlobalSign). Our paper try to analyse the main design issues for a Public Key Infrastructure (PKI), needed to secure the most important network applications: Web access authentication and server-client communication confidentiality, VPN over Internet implementation, secure (signed) document and e-mail interchange.

1. INFORMATION SECURITY

Information security is now the major issue facing today's electronic society. For instance, e-mail now carries not only memos and notes, but also orders, contracts and sensitive information. The Web is being used not only for publishing corporate brochures but also for placing some organization's sensitive information, needed in the decision process. Virtual private networks (VPNs) are extending organization networks onto the Internet for remote network access. Extranets turn the Internet into dedicated, secure links between organization (military) partners for information exchange. Moreover, e-commerce is a competitive imperative for business worldwide; it is the fastest growingchannel for marketing, selling, documenting and distributing products and services, most of them needed for military purposes.

Why do we need information security? Although all nations have their own classified estimations of the threats their computer systems face, the following quotes from unclassified sources provide a strong indication of the magnitude of that threat:

- "More than 120 countries already have or are developing computer attack capabilities." Defence Science Board;
- "It is estimated that the DoD is attacked about 250,000 times a year …" Defence Information Systems Agency (DISA);
- "Computer attacks have also become easier to carry out due to the proliferation of readily-available hacker information, tools, and techniques on the Internet." General Accounting Office (GAO);
- "Any marginally computer literate individual can use the Internet itself to quickly obtain basic information on the tools and techniques needed to become a computer hacker" GAO.

Given the severity of the threat, it is clear that unprotected communication and information systems are at risk. If they are not protected, organizations in **Romanian Armed Forces (RAF)** will experience (1) exposure of classified information to unauthorized persons, (2) destruction of critical data or, just as problematic, loss of confidence in the correctness of the data and (3) a potential loss of control over its forces. Finally, the performance of inadequately protected CIS could be degraded or reduced to zero at critical points in time by adversaries.

Ensuring adequate information security for military CIS systems require the development and evolution of an effective Information Security (INFOSEC) architecture that is based on a thorough understanding of the threat, system vulnerabilities and availability of counter-measures for protecting his own CIS. Another principal guidelines used in creating integrated security architecture are the need to ensure adequate protection of NATO classified information that is shared with Romania. As an alliance of independent sovereign states, NATO depends on the cooperation of its members to ensure adequate levels of security for shared information.

2. SECURITY ARCHITECTURE

When developing architecture, it is important to understand the underlying security principles, the basic security services and associated mechanisms and specific building blocks that can be used as the basis for creating the architecture. *Absolute total security of all military CIS resources simply cannot be afforded.* A corollary of this principle is that *total multi-level security (MLS) solutions are not realistic*–no one can afford to put a trusted workstation on every desktop. This fact is driven by the excessive cost and time required to produce products which have the necessary level of trust to be judged multi-level secure. All security system designers, managers and decision-makers need to be prepared to deal with residual risk. In other words, the use of operational procedures, audit reduction and monitoring, and other techniques will be needed to handle those risks that cannot be totally overcome by the technical design of CIS security features.

Since each and every desktop, server and network cannot be protected in an absolute sense, a better alternative is to provide strong protection at the enclave level. In its simplest terms this means providing the strongest security at the boundaries of an enclave where it is connected to less trusted or untrusted networks. Internal to an enclave less stringent technical security measures would be used, backed up by operational procedures and other techniques.

Highly interconnected networks and systems are a fact of life. The World Wide Web and related technologies are driving a trend toward new ways of accessing information, new information services and distributed processing that the military cannot afford to ignore technically or operationally. However, with all of these new capabilities come increased risks via network interconnections. In this environment, if a hostile agent can gain access to and subvert one workstation or server on a network that is "trusted," it can use that access to penetrate the remainder of the systems. Therefore, it is important that the "weakest link" be protected.

The level of protection afforded a system should be based on the value of the information that it contains, or the function that the system performs. In most military systems the value is a direct function of the classification of the data on that system and its military mission category.

There are a number of security risks. To reduce these risks, some **security services** have evolved over time. Table 1 lists these services along with the **security mechanisms** that can be used to provide the service.

Table 1. Security Services and Mechanisms

Service	Mechanism
Confidentiality	Encryption (link, bulk, E3) Access control (MAC/DAC)
Integrity	Digital signature Access control (MAC/DAC)
Availability (Denial of Service)	Encryption (link, bulk, E3) Digital signature Access control (MAC/DAC)
Authentication	Encryption (digital signature)
Non-repudiation	Encryption (digital signature)

- **Confidentiality** services ensure that the contents of the message have not been disclosed to third parties and data is not accessed, seen or otherwise available to unauthorized users whether it stored on a workstation or server or in transit over a network. Confidentiality requirements are enforced by using access control mechanisms on computers and by encrypting data while it is in transit over a network and sometimes while it is stored on disk. There are many types of encryption including link, bulk and end-to-end encryption (E3).

- **Integrity** services proof that the message contents have not been altered or destroyed, deliberately or accidentally, by an unauthorized action. Mechanisms used to protect the integrity of data include message hashing, encryption and access controls. Message hashing is a technique that creates a "checksum" based on a "one-way" function and attaches it to the data. Digital signatures are a special encryption technique; the encryption process does not encrypt the "text," but instead encrypts the message hash and other data designed to prevent replay and other types of attacks. Access controls limit access to data to authorized personnel that prevents system users who aren't authorized access to that data from altering the data.

- **Availability** is focused on ensuring that a particular resource is accessible and useable upon demand by authorized personnel, i.e., that they are not denied access and use by an adversary. Again, encryption is used to prevent sophisticated attacks against networks and computer systems over communication links while access controls are used to prevent unauthorized personnel from shutting them down.

- **Authentication** is a mechanism by which a user proves he is who he says he is. In computer systems and networks, some mechanism is needed to ensure that the identification supplied is in fact the real identity of the individual. Historically, this has been done with simple passwords but that has proven to be ineffective against today's hackers. There are many techniques being used in modern identification and authentication systems but all of the strong ones depend on encryption and many depend on digital signatures.

- **Non-repudiation** is a service that prevents entities involved in a communication exchange from denying that they participated in that exchange. For example, non-repudiation can be used to prove that a certain user originated a message and that another user received that message. Again, digital signatures provide a strong technical solution for this requirement.

Development of a **top-level security architecture** is easier if one uses generic components to create it. These components fall into *two main generic categories*:

- *Communication security (COMSEC)*, based on the use of encryption (both symmetric and asymmetric) and associated security protocols and key management, protects data in transit;
- *Computer security (COMPUSEC)*, refers security techniques embedded in computer systems that enable those systems to be "trusted".

3. NEED FOR PKI

Having established the need for security, looked at how NATO views security and reviewed some basics of security architectures, it is time to consider the *technologies that are available to put together a meaningful architecture and system design*. Computer security technology involves understanding what it means to "trust" a system and how one can achieve the requisite level of trust. The "*Common Criteria*" provides a framework for achieving trust. *Public key cryptography* is a relatively new technology that provides a number of important security capabilities that can be used by security architects. *Protect, detect and react tools* are being driven by the marketplace and the ever expanding use of the global Internet for commerce. These tools are extremely useful in countering many of the security threats that are a natural outgrowth of the "openness" of Internet technologies and the impossibility of writing totally correct software. *Firewalls and Guards* are two examples of technology solutions designed to provide a *perimeter defence.*

The new approach in modern cryptography, based on public keys, enables the provision of a *digital signature* capability, non-repudiation, strong identification and authentication, secure key exchange and ad hoc secure communications. In order to provide these services, a **Public Key Infrastructure (PKI)** is required. This infrastructure depends on certificate authorities to create and sign certificates for all of the users of the public key system. These certificates, which are signed by the certificate authority, provide the public keys of users and generally are entered into a public directory so that anyone can access them. In general, *public key cryptography is a computationally intensive technology and is not suitable for the encryption of files, long messages, etc.* It is, however, *suitable for digital signature and key exchange*. Symmetric keys are suitable for encrypting data but not for digital signature or ad hoc exchange of keys. Consequently, in the vast majority of applications a combination of public key and symmetric

key cryptography is used to best advantage. More western countries, U.S. and NATO are in the process of selecting and fielding PKI systems at this time.

PKI as defined herein refers to the framework and *services that provide the following*:
- generation, production, distribution, control, revocation, archive and tracking of public key certificates,
- management of keys,
- support to applications providing confidentiality and authentication of network transactions,
- data integrity, and
- non repudiation.

The organizational (military) PKI shall provide an integrated public key infrastructure that supports a broad range of commercially based security-enabled applications and provides secure interoperability with the military and commercial partners while minimizing overhead and impact to operations. It is the objective of the PKI to provide *certification services that have the following characteristics*:
- support multiple applications and products;
- provide secure interoperability throughout the military organizations and with its partners such as government agencies, industry and academia;
- standards based;
- uses commercial PKI products to minimize the investment and the manpower to manage the PKI;
- support digital signature and key exchange applications;
- support key recovery;
- employs centralized certificate management and decentralized registration;
- support Federal Information Processing Standards- FIPS compliance requirements.

4. PKI COMPONENTS

In a PKI, there are several different entities or components. These components may be implemented separately, but are commonly integrated and delivered through what are called **Certificate Servers**.

1. **Certificate Authority (CA)** is the most fundamental component that will authorize and create digital certificates. A certificate authority (CA) server issues, manages, and revokes certificates. The CA's certificate (i.e., public key) is well known and trusted by all the participating end entities. The CA can delegate its authority to a subordinate authority by issuing a CA certificate, creating a certificate hierarchy. This is done for administration (e.g., different issuance policies) and performance reasons (e.g., single point of failure and network congestion). The ordered sequence of certificates from the last branch to the root is called a certificate chain. Each certificate contains the name of that certification's issuer (i.e., this is the subject name of the next certificate in the chain). A self-signed certificate means that the signer's public key corresponds to it's

private key (i.e., the X.509v3 issuer and subject lines are identical).

2. The second core component of a PKI is the **Registration Authority (RA)**, which provides the mechanism and interface for submitting users' public keys and identifying information in a uniform manner, in preparation for signing by the CA.

3. The third component is a **Repository (Directory Server)** in which certificates and certificate revocation lists are stored in a secure manner for later retrieval by systems and users. *Lightweight Directory Access Protocol (LDAP)* was originally designed to make it possible for applications running on a wide array of platforms to access X.500 directories. LDAP is defined by RFCs 1777 and 1778 as an on-the-wire bit protocol (similar to HTTP) that runs over TCP/IP. It creates a standard way for applications to request and manage directory information (i.e., no proprietary ownership, or control of the directory protocol). The directory entries are arranged in a hierarchical treelike structure that reflects political, geographic, and/or corporation boundaries.

4. **PKI Applications** are those use public-key technology. In most cases, the application would provide underlying cryptographic functions (e.g., public/private key generation, digital signature, and encryption) and certificate management. Certificate management functions include creating certificate requests, revocations, and the secure storage of a private key(s). Examples of PKI applications include Netscape's SSL 3.0 browser/server, Deming's Secure Messenger, and GlobeSet's Secure Electronic Transaction (SET) Wallet, Microsoft Outlook mail system.

5. PKI COMPONENT SECURITY REQUIREMENTS

PKI components each share a set of security requirements (i.e., baseline) with each other. The baseline corporate PKI security requirements are as follows:

- Reliable software (i.e., a comfortable level of assurance that security software is implementing the cryptographic controls properly).
- Secure/trusted communications between components (e.g., IPSec, SSL 3.0).
- PKI specific security policies that are derived from the existing set of corporate security policies.

Most PKI software/hardware is built upon cryptographic toolkits (e.g., RSA's B-Safe). The application that calls the lower level functions in the toolkit is still prone to human errors. Every other month Microsoft and Netscape release bug fixes for their Internet product sets. If the browser wars continue, there will be shorter quality assurance cycles to meet the current time to market constraints, hence produce a lower quality of software. PKI components require authenticated and private communication among each other. This prevents active or passive threats (e.g., eavesdropping, spoofing) from

occurring. Most current implementation of PKI components supports SSL 3.0. Each component has a security criterion it must meet to be part of a PKI. This criterion is based on the level of protection necessary to perform the business objectives within the acceptable level of risk. The security mechanisms used to meet this criterion usually falls into physical, platform, network, and application categories. These categories are not all included in the PKI applications and have to be supplemented. Examples of these are network firewalls, disabling NFS exports, authenticated naming services, and tight administrator controls (e.g., root user).

Certificate Authority

The certificate authority security requirements are:

- Certificate generation, issuance, inquiries, revocation, renewal, and storage policies.
- Certification Practice Statement (CPS).
- Certificate attributes or extension policies.
- Certificate administration, audit journal, and data recovery/life-cycle support.
- Secure storage of private keys.
- Cross certification agreements.

The applicability and/or usage of the certificate the CA manages are defined in the **Certificate Policy (CP).** A security policy must exist for each CA function (e.g., generation, issuance, revocation list latency, etc.). These policies are the foundation upon which all the CA security related activities are based on **Certification Practice Statement (CPS)** is a detailed statement by the CA as to its certificate management practices. The certificate end entities and subscribers need to be well aware of these practices before trusting the CA. The CPS also allows the CA to indemnify itself to protect its relationships.

One of the major improvements to version 3 of X.509 is the ability to allow flexible extensions to the certificate structure. These extensions include additional key and policy information, user and CA attributes, and certification path constraints. The CA must document, by way of a policy, the certificate attributes and extensions it supports. In addition, to allow interoperability outside the corporation, one must register the extension object identifiers (OID) with the American National Standards Institute (ANSI).

The CA must maintain an audit journal of all key management operations it performs. All certificate management functions must be audited (e.g., issuance, revocation, etc.) in case of a dispute. In conjunction with this auditing function, a data recovery and certificate life cycle plan must also exist. The CA administrator interface must enforce the least privilege principal for all administrator actions.

The certificate authority must provide for the adequate protection of the private key that it uses to sign certificates. The machine that the CA runs on must be protected from network and physical intrusions. Optionally, the CA's private key used to sign certificates can be stored in a tamperproof hardware module (e.g., meets FIPS PUB 140-1 level 3).

Cross-certification certificates are issued by CAs to form a non-hierarchical trust path. Two certificates are

necessary for a mutual trust relationship (i.e., forward, and reverse directions). These certificates have to be supported by an agreement between the CAs. A cross-certification agreement details the obligation of liability between partners if a certificate turns out to be false or misleading.

Directory Server

The directory server security requirements are as follows:

- Supports network authentication through IP address/DNS name, and user authentication through LDAP user name and password, or a X.509 version 3 public-key certificate.
- Controls the users' ability to perform read, write, search, or compare operations down to the attribute level.
- Provides message privacy (SSL) and message integrity for all communications.

The directory server contains corporate and user personal attribute information. Access to this information must be controlled at the most granular level possible. Directory administrators must be able to restrict particular users from performing specific directory operations (e.g., read, write, search, and compare). Authentication must support conventional username/passwords and/or certificates. Additional filtering should be provided using IP address/DNS name. Network access to the directory server must be able to be protected between all PKI components.

PKI Clients

All PKI clients, at a minimum, must be able to generate digital signatures and manage certificates. PKI client requirements are as follows:

- Generate a public/private key pair.
- Create a certificate request (PKCS#10).
- Display certificate.
- Verify certificate.
- Delete certificate.
- Enable or disable multiple certificates.
- Request a certificate revocation.
- Secure storage of certificates (e.g., password, and hardware).
- Secure exporting certificates (e.g., PKCS #12).
- Select algorithm, key strength, and password controls.
- Configure security options (e.g., sign/encrypt whenever possible).

The process begins with a PKI client generating a public/private key pair locally. The software used to generate the public/private key pair must use a non-deterministic algorithm. Once the key pair is generated, the public portion needs to be bound inside a certificate structure. The PKI client must then generate a certificate request adhering to the PKCS#10 syntax and submit that information to a CA. Once the CA fulfils the request, the message response sent back to the client is in PKCS#7 syntax (i.e., signed envelope). All network traffic is kept private between the client and the CA.

The PKI client must have the ability to manage multiple certificates. This includes viewing the certificate structure (e.g., subject, issuer, serial number, fingerprint, and validity dates); deleting it, if necessary; choosing (i.e., enabling) what certificate to use or query the user; or requesting the CA to revoke it.

A large portion of public cryptography is based on the protection of the private key. The PKI client must protect their private key commensurate with the risk associated with the loss of all the transactions it processes. This will require encrypted storage of the key using an application authentication challenge (e.g., organization compliant password), or hardware token or smart card, and the user physically protecting their desktop (e.g., password protected screen saver).

Due to the infancy of this technology, certificates are bound to the PKI client application software and hence the host that the software resides on. An emerging public key cryptographic standard (PKCS) called personal information exchange syntax standard (i.e., PKCS #12) details the transfer syntax for personal identity information. This includes private keys, certificates, miscellaneous secrets, and extensions. This will allow PKI clients to import and export personal identity information across multiple platforms and applications. The most secure method includes a privacy and integrity mode that requires the source and destination platforms to have trusted public/private key pairs available for digital signatures and encryption. The least secure method protects personal identity information with encryption based on a password.

6. CERTIFICATES AND REVOCATION LISTS

The certificate tipical forma is ITU's **X.509v3**. PKI clients support their own **certificate types** (e.g., mail, browser). Before a certificate can be requested from a CA it is necessary to have access to the CA's certificate in the PKI client. Typical certificates are the following:

- *Certificate Signer/Provider*-These external institutions provide an outsourced CA function. They are usually preloaded into the PKI client application (e.g., VeriSign Class 2 Primary CA, GTE CyberTrust).
- *Sites/Hosts*-This is a list of sites/hosts that the PKI client has stored locally. The importing process for this varies depending on the application vendor. Netscape's browsers allow importing site/host certificates from non-certificate signers/providers over SSL. Microsoft requires a more stringent trust model that most CA vendors use insecurely.
- *Code Signer/Provider*-The code signer/provider certificate allows Java applets or ActiveX scripts to verify message authentication (i.e., data integrity) before they are executed.
- *Cross –Certification*- A cross-certification certificate defines a one-way trust path between CAs.
- *Personal*-Personal certificates are used to identify one's self to other PKI clients that require authentication and/or privacy.

- *File*- A file certificate is used to encrypt or sign local files. This certificate is only shared with a key recovery server.
- *Key Recovery*- A certificate used between the key recover servers.
- *PKCS #12*- PKCS#12 optionally requires a pair of certificates; one for encryption and the other for signing each host that requires the secure transfer of private key information.

7. ARCHITECTURE REQUIREMENTS FOR CA

The target PKI employs centralized certificate management and decentralized registration shown in Figure 1 and uses common processes and components to minimize the investment as well as the manpower required to manage and operate the PKI.
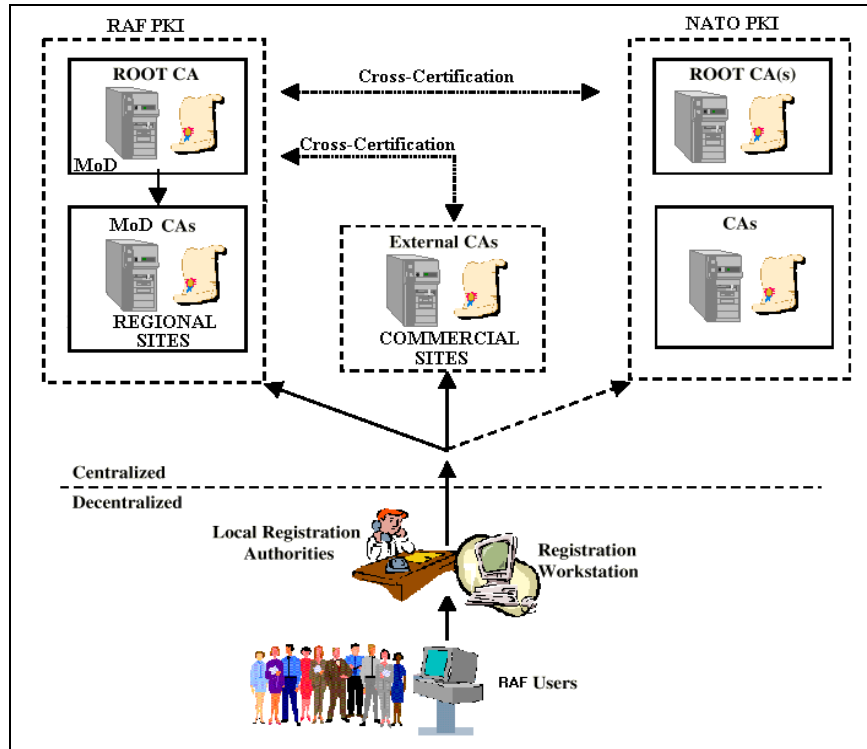
Figure 1 PKI CA Architecture

The centralized portion of the certificate management process shown in Figure 1 is comprised of a combination of military owned and operated components. The Defense Minister will manage and operate its Root CA s. The Root CAs is responsible for managing subordinate CAs and External CAs (ECAs) and cross certifying with other domains for interoperability. The Root CA s will be operated as offline devices with maximum physical personnel and procedural security protections. A standards based certificate request format (e g PKCS #10 or RFC 2511) will be used to interface with the Root CA s and register subordinate CAs into the system in a trusted out of band process.

Based on current technology limitations it is envisioned that it will require separate CAs on each of its networks, similar to the current implementations today where identical PKIs are replicated on each network. The CAs that support classified mission critical command and control applications will be under the direct control of the Root CA s and will be owned and operated by the Defense Ministry. They are networked devices supporting a standards based secure interface for the Local RAs for user registration. They will be operated

with the technical physical personnel and procedural security protections as defined in the military regulations. It is expected there will only be a small number of CAs located at several regional sites. The target PKI plans to eventually achieve secure interoperability with non-military entities through a process called "direct cross certification" which establishes a policy and process for recognizing third party CAs.

The registration function is decentralized in the target PKI with **Registration Authorities (RA**s) responsible for user identification. The military Services and Agencies will manage registration. It depicts a common workstation and web based application with hardware token. A common registration workstation with unified ordering and delivery software will be based on commercial standards and technologies. The target envisions a common set of processes and tools so that the only differences between assurance levels from the RAs' and users' perspective are the user identification procedures and tokens protecting the keys. This will allow users to register with the appropriate CA server

through a single RA. This single registration workstation should be able to transparently register users into CAs commercial certificate service providers or other external CAs as needed.

End users commonly referred to, as end entities can be a person a machine such as a router or a process running on a computer such as a firewall. The target PKI will need to provide support for all end entities including non-human Registration of end entities will use a common registration application to securely register with the infrastructure. During registration the user's token will generate a digital signature key pair public and private key and send the public key to the CA. Once the CA returns the certificate, the user can then load the certificate onto the token (smartcard, Universal Serial bus USb device or personal computer PC card).

User registration process employs pre-registration and direct delivery of the certificate and key information to the end user or equipment. As an example one potential implementation is the following:

1. The RA making use of this common web based registration application securely authenticates e g SSL to the appropriate CA servers via a common KMI management front end and registers the user s.
2. Next the RA identifies the user as required by the policy and provides the necessary information for the user to authenticate to the CA server.
3. The CP specifies the authentication requirements process for the various assurance levels. After receiving the authentication information the user can use a common registration application to securely connect to the appropriate CA server and request a certificate.
4. During registration the user's token or software application will generate a digital signature key pair public and private key and send the public key to the CA server.
5. The CA server processes the request verifies possession of the private key generates the certificate posts it to the directory system and returns the certificate to the user.
6. The user can then load the certificate onto the token (e. g. smartcard, Universal Serial Bus USB device,

or personal computer PC card). This token certificate can be used in a variety of applications allowing a single registration to support multiple applications. Once the user has a digital signature certificate he she can use that certificate to request additional certificates such as encryption or attribute certificates at the same assurance level.

8. CASE STUDY: RSA KEON CERTIFICATE SERVER

This case study is an implementation of a PKI using RSA Security suite of products, particularly *RSA Keon Certification Server (KCS)*. The PKI clients were Netscape's Navigator 4.1 and Microsoft Internet Explorer 5.0 and Outlook Express. The outcome of the effort fields an organization level PKI, including client authentication and secure e-mail (S/MIME) using a self-signed root. This solution addresses the security problems of an organization and prepares the steps necessary for creating a *PKI pilot centre*.

RSA Keon is a family of products of RSA Data Security S.A., which can be applied to deliver organization level security through the application of public/private key-based cryptography. The RSA Keon product family expands beyond this definition to include an RSA Keon Security Server, RSA Keon Desktop, RSA Keon Agents. The components that provide for the creation and management of public/private keys and the associated digital certificates make up a PKI. Across vendors, there may be some variation in the components that make up a PKI, but the most common set of components is:

- A *Certificate Authority* (CA), an entity which issues certificates;
- A *repository* for public key certificates and *Certificate Revocation Lists* (CRLs) usually based on a Lightweight Directory Access Protocol (LDAP)-enabled directory service;
- A *management function*, typically implemented via a management console (RA).

RSA Keon Certificate Server integrates the functions of a CA, RA and Repository into a single system, as shown in Figure 2.
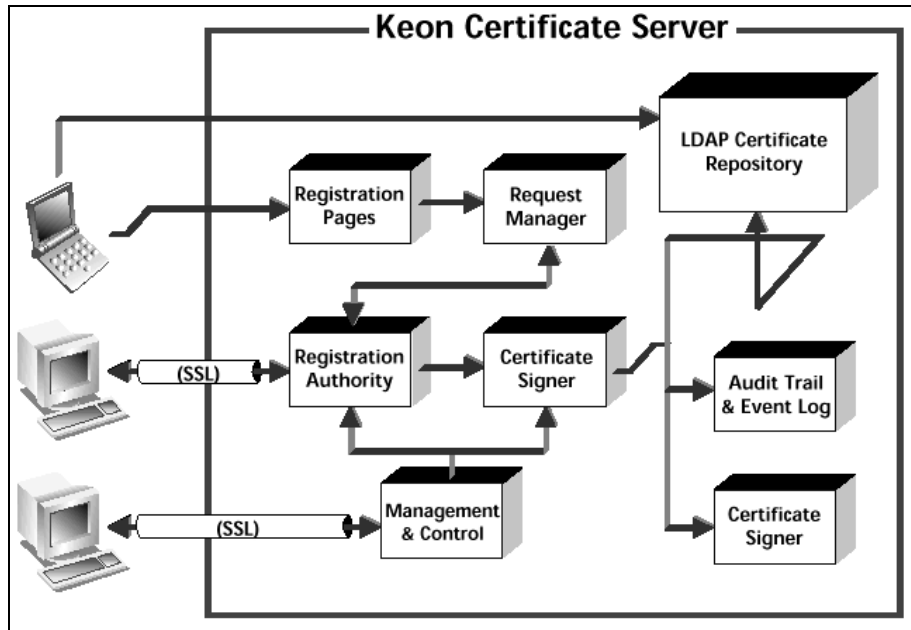
Figure 2 RSA Keon Certificate Server Architecture

The Keon Certificate Server is based strictly on open-standards. Organizations can rest assured that they are

Instead, companies using the Keon Certificate Server can deliver certificates that will interoperate with PKI solutions from any vendor that follows the popular PKI standards in existence, such as LDAP, PKCS #7, PKCS #10, X.509v3, and PKIXv1. RSA Keon Certificate Server was designed to inter-operate with standard, off-the-shelf, PKI ready applications from **Netscape** and **Microsoft**. At the RSA Keon Certificate Server, certificates and CRLs can be published to the bundled Netscape LDAP directory or to other LDAP compliant directories. As a result the RSA Keon Certificate Server can be replaced with other Certificate Servers that implement standard X.509 v 3 certificates, such as **VeriSign OnSite**.

Secure administration of KCS is handled through simple Web interfaces protected by digital certificates and SSL. The certificate delivery and granting process can be automated so that no administrator intervention is required. In addition, a full suite of complementary tools is available for integrating existing and new applications into the PKI. The BSAFE line of software development kits (SDKs) from RSA Security can be used to PKI-enable applications.

RSA Keon Certificate Server, acting as a trusted entity known as a **Certificate Authority (CA),** include two entities:

   -**System Administrator**- which designates Certificate Administrators and configure components of system- *Signers*, *Signers Hierarchies*, *Jurisdictions*, *Directory* and

   -**Certificate Administrators**-, which manages certificates for subscribers.

Figure 3 explains the process of certificate administration in a Keon Certificate Server context. The Keon CA:

not tied to a vendor's proprietary solution, which will only work with other products from that same vendor.

- control over **who has access to your organization's information** on Intranet /Extranet,
- control over **access to sensitive data,**
- establish & publish the **Statement of Practice** document,
- control over **who signs certificates** by specifying signers and jurisdictions,
- control over **certificate granting,**
- **issues** and **manage** certificates,
- **digitally signs** subscriber's certificate,
- **made up** of entities within your organization or within an external source, a **trusted third party** (eg.**Verisign**).

In this context, **Certificate Server** & **System Administrator** act as CA and **Certificate Administrator** acts as agent of CA. RSA Keon Certificate Server (KCS) enables **policies** and **practices** of:

   -certification,
   -certificate revocation,
   -certificate management,
   -additional certificate activities.

KCS is accessed through a 2 **Web browser**:

   -**CA Center** for System Administrator,
   -**Control Center** for Certificate Administrator.

KCS generates and distributes **3 distinct types of certificates**:

   -**Personal Certificates**
      >Netscape & Microsoft Personal
      >CSR (Certificate Signing Request) Personal,
for applications that produce certificate requests.
   -**Secure Server Certificates** (Web server)
   -**Internet Protocol Security (IPSec) Certificates** (between routers- IP layer).

Using the concept of **jurisdiction**, KCS divides Subscriber population into groups. Each Jurisdiction is managed by a different Certificate Administrator (e.g. Engineering Jurisdiction, Finance Jurisdiction, IT Jurisdiction) and simplifies the certificate management.

Each Jurisdiction may be configured to use only specified certificate types. Several Signers may be allocated to a Jurisdiction but only one Signer can be allocated per certificate type.
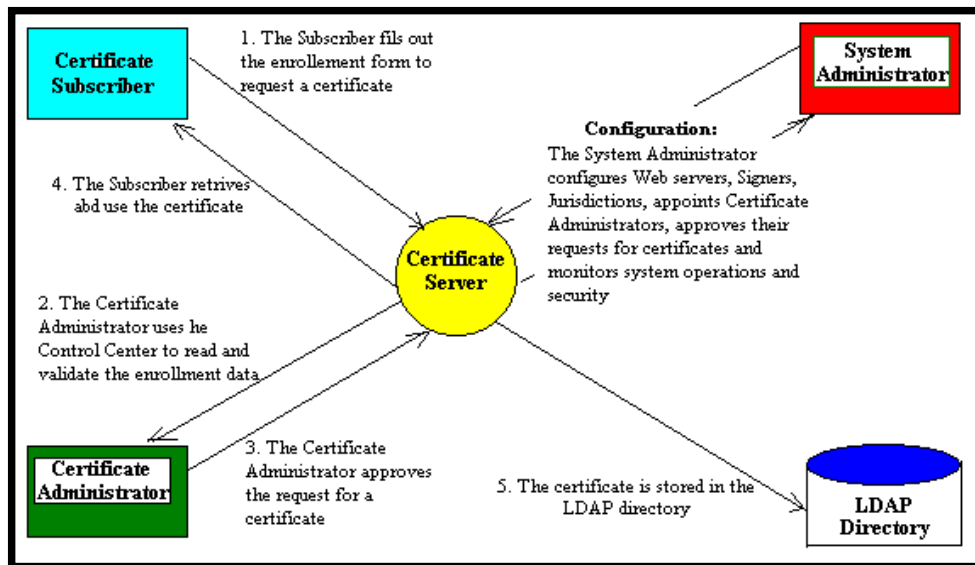
Figure 3 Certificate Management Process

Another notion used by KCS is **signer**: a cryptographic key that the signing software in KCS uses to sign a certificate. Using the Signer, KCS digitally signs the information that a subscriber entered in the *Certificate*

Enrollment Form. The result is a certificate. Each Signer has a unique name that appears in each certificate. Figures 4 illustrate the signer's hierarchy.
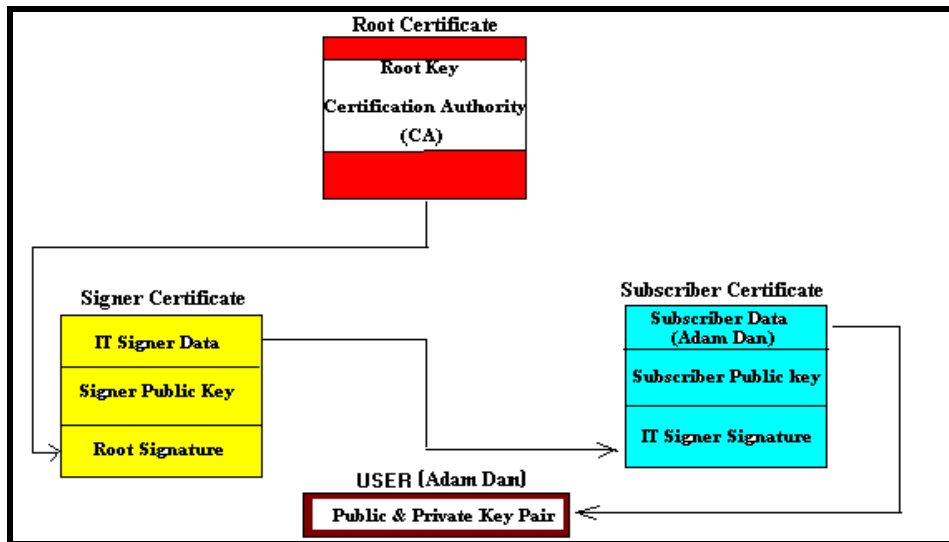
Figure 4 Hierarchies of signers

System Administrator has the following *responsibilities*: installing & configuring, managing signers, managing jurisdictions, configuring, Directory Server interface, and scheduling the creation of CRL. His *day to day tasks* are: appointing qualified Certificate Administrators, auditing & monitoring system security, managing System Administrator certificates, providing support for Certificate Administrators.

Certificate Administrator has the following *responsibilities*: configuring the enrollment pages, customizing the Subscriber agreement. His *day to day tasks* are: validating the identities of applicants, approving or rejecting Certificate requests, assigning requests to other Administrators, revoking certificates, providing support for Certificate Subscribers.

There are many **applications that require a Certificate Server in order to function**. We will examine three of them: Web access authentication and server-client communication confidentiality, secure (signed) document and e-mail, and virtual private networks (VPN).

1. Secure E-mail and Messaging

Many organizations rely on e-mail to distribute sensitive information. Unsecured e-mail can be intercepted, read and modified while in transit from sender to receiver, without the knowledge of either party. Worse still, today's e-mail systems make it very easy for attackers to create phone messages, create false orders, and disrupt business operations. An organization PKI based on digital certificates from a Certificate Server is the solution to these problems. In secure e-mail systems, a Certificate Server supplies digital certificates that plug in to existing mail clients including Microsoft Outlook, Netscape Messenger, and Eudora. These clients are all able to leverage the power of digital certificates and the S/MIME secure messaging standard out-of-the- box or through plug-ins. In addition to e-mail, message-based applications such as EDI and electronic billing are useful because they provide asynchronous delivery of documents. Security for these other applications can be accomplished using S/MIME technology in combination with a Certificate Server.

2. Web Applications

The Web plays a huge role in organization today. However, security continues to be a problem. In the case of Intranets, physical security measures and firewalls are sufficient to keep external parties from accessing sensitive internal Web pages. However, a means needs to exist to keep sensitive information from other employees within the organization firewall. The technology that solves these problems and enables secure Web is the Secure Sockets Layer (SSL) protocol. The protocol allows Web servers and Web clients to securely share information across networks, both inside and outside of the firewall. Any Web application or server that contains an SSL engine for secures connections only needs a digital certificate from a trusted Certificate Server to enables Web security. In fact, the ability to perform secure Web access through SSL already exists in Web clients such as Netscape Navigator and Microsoft Internet Explorer. Web servers such as Netscape Enterprise Server and Microsoft IIS are already SSL-capable. In virtually all organizations today, the only component missing from activating secure Web access is digital certificates from a Certificate Server. Using a Certificate Server, organizations can issue digital certificates to employees, Web servers, customers, and partners that enable access to Web resources from inside and outside the firewall.

3. Virtual Private Networks (VPNs)

Military organizations have employees and facilities distributed across long distances. Employees should be able to access shared data stores as securely and easily as if they were sitting at a desk in the home office. However, leased data lines are very expensive and of doubtful security for confidential information. Expenses are huge when connecting distributed offices across the world via this method. Ideally, the organization could use the Internet to securely connect their distributed networks and individual users at the cost of only the connection fee to a local Internet Service Provider (ISP), thus only paying for access to the Internet at a fraction of the cost of leased data lines. Virtual Private Network (VPN) technology enables organizations to leverage the Internet, allowing users and networks distributed across the world to securely access resources. VPNs are enabled through the use of specialized hardware, such as routers, and also network drivers that allow VPN access. The protocol for VPN is based on the IPSec standard. IPSec-compliant software and hardware need digital certificates from a trusted Certificate Server in order to securely extend organization networks to distributed offices and remote users.

9. Activities for implementing an organizational PKI

An *organizational public key infrastructure* is defined as a PKI that is used by an organization to support its own processes, which may be of a command, manage or business nature. We try to identify some important steps in the process of implementing such PKI.

Analysis of Operational Requirements PKI should be implemented to meet clearly defined operational objectives. The primary objective of this activity is to determine the processes that can be supported by the PKI, and the nature of those PKI services. It must also decide to implement a PKI in conjunction with the implementation of an application that will use PKI-based security services. If you do so, you should still consider the overall requirements of your organization to ensure that the PKI you are about to implement would meet more than the requirements of a single application.

Analysis of Certificate Policy Requirements It must define the quality of the PKI security services needed in order to support your operational processes. Quality is normally established by specifying certificate policies:

- Methods to identify and authenticate applicants who receive keys and certificates
- Obligations and liabilities to be expected of the subscribers, relying parties, and the certificate authority (CA)
- Procedures for a variety of topics, including certificate application and revocation, the collection of audit records, and records archival
- Technical aspects of key generation, delivery, and usage, cryptographic modules, activation data, and computer and network security
- Physical and procedural controls, which cover such topics as physical safeguards for the CA facility, specification of CA roles, and key changeover and recovery

- CA personnel security controls, including specifications for security qualifications, experience, and training
- Profile of certificates and certificate revocation lists.

Analysis of PKI Solutions There is a number of PKI solutions available on the market. You must review these products and identify those that can meet your requirements. Examples of things to consider are: maturity of the product, number and types of installations, users' level of satisfaction, conformance to standards (PKI, cryptography, communications, directory), product functionality, availability of integration toolkits and their ease of use, availability and quality of product training, vendor's product evolution philosophy and release strategy, etc. At the end of this activity, you should have an idea of the potential PKI solutions.

Analysis of Network Infrastructure The next step is to review your network infrastructure and identify the work required to integrate the potential solutions. You must analyse communications protocols, directory protocol, Internet connection for accessing PKI services from the Internet, entry points protected by a firewall, etc. Answers to such questions will provide you with information that will serve as input to the cost-benefit analysis.

Analysis of Cost-benefit: The analysis should clearly demonstrate the costs for your organization to implement, operate, and maintain its own PKI, versus the cost of purchasing PKI services from an established CA. You must select the PKI implementation option that best meet your organization's needs.

Integration of PKI Applications A PKI gives you the ability to use public key-based security services to support you operational functions. Before you can benefit from these services, however, you need to integrate the security service requests into your applications, or you must replace them by, or upgrade them to, applications with such requests already integrated. It is important to devise early a strategy for integrating PKI-based security services into your environment. Integrated applications should be introduced gradually into your environment, possibly starting with e-mail and other office automation applications.

Analysis of Policies and Standards This activity consists in reviewing policies and procedures that apply to your organization's activities, and all related security policies, standards, and procedures, for the purpose of understanding the framework under which the certificate policies are to be developed.

Development of Certificate Policies Results of the operational requirements analysis, the certificate policy requirements analysis, and the policies and standards review should provide you with the information you need to either adopt existing certificate policies, or to develop your own. Certificate policies contain specifications for security controls, CA practices, certificates, and keys, which must be implemented within the infrastructure. If you need to accept certificates issued by another CA, it might be wise to adopt policies that conform to, or to develop policies consistent with, the IETF's framework.

PKI Architecture The PKI architecture is typical system architecture. It should specify such things as hardware, software, and communications components, communications protocols, directory software and structure, cryptographic standards, and CA facility security.

Threat and Risk Assessment Organisations with established IT security risk management policies and standards may want to assess the suitability of their PKI architecture and the technical and administrative safeguards they are about to implement so as to determine the level of risk to their CA operations. Depending on the results of your threat and risk assessment, you may have to go back and modify the certificate policies and your PKI architecture so as to reduce the risk to an acceptable level.

PKI Design Your PKI's design depends on the certificate policies your CA will be supporting and the PKI solution you have selected. During this activity, you will describe in detail your PKI implementation, including the specification and configuration of network segments, and hardware and software components. You should prepare an itemized inventory of servers, workstations, routers, hubs, cables, network interface cards, firewalls, un-interruptible power supplies, and any other hardware and software components that you need to purchase. The PKI design document should also contain an inventory of changes to existing components.

CA Facility Design Based on your certificate policies, you must select and design a facility to house your CA's main components. Your design should cover construction requirements and the specification of physical and environmental safeguards. Deliverables from this activity should include an inventory of items to purchase and an inventory of changes to existing elements.

Personnel Selection The trustworthiness of your CA's operations will depend largely on the personnel you assign to the various roles. You need to select your PKI personnel with care according to the relevant stipulations of your certificate policies. Selecting your team at this stage will ensure their participation in the implementation activities, which provides an opportunity for knowledge transfer and learning.

CA Operations Manual The manuals typically supplied by vendors with their products rarely offers sufficient documentation, as they do not contain the operational procedures specific to your implementation. It is therefore recommended that you develop a manual containing detailed procedures covering all of the day-to-day operations. The manual should also cover maintenance and support.

Certification Practice Statement To support legal requirements, you probably have to prepare and publish a certification practice statement (CPS), which is a statement of the practices that your CA employs in issuing its certificates. The CPS describes the equipment,

the policies, and the procedures you have implemented to satisfy the specifications of your certificate policies. Like the certificate policies, your CPS should be consistent with the IETF PKIX Part 4. It will contain high-level statements from the PKI and CA design documents and the CA operations manual, as well as the general provisions expressed in the certificate policies.

PKI Implementation Plan As for any IT system, you should prepare a detailed implementation plan that covers activities for acquisition, installation, configuration, testing, certification, accreditation, and training. The plan should also contain a detailed schedule, complete with tasks, resources, and start and end dates.

Hardware and Software Acquisition Time to order your hardware and software components. You may wish to start this process as early as possible to avoid unnecessary delays.

Installation, Configuration, and Testing During this activity, you will construct the CA facility, and install and configure the hardware and software components as per your PKI design documents and your CPS.

PKI Training You should develop a training plan for your PKI personnel. Training should cover operations, maintenance, and support. Your PKI solution provider will probably have an adequate training program that can be tailored to include all of the procedures specific to your implementation. PKI personnel should also participate in the installation, configuration, and testing activities.

PKI Certification and Accreditation Certification of your PKI is equally important, a process by which you measure your PKI's actual implementation against its design. This type of certification may be conducted internally; however, having an independent and qualified firm conduct it for you will add credibility to the process and help establish the trustworthiness of your CA's practices.

Operations It may now begin to receive subscriber applications and issue certificates.

10. CONCLUSIONS

Internet is changing the way military activities are conducted. PKI is the enabling technology that simplifies the management and security of this process. With the right PKI implementation, military organizations can spend less time worrying about security, and more energy on their main activities. For example, confidential documents no longer need to wait for days to be physically shipped. Instead, they can be securely sent through e-mail. Web servers can allow secure access for only designated users, eliminating the need for human intervention. Military organization networks can securely extend over the Internet, eliminating expensive leased data lines. PKI's possibilities are limitless. For *Romanian Armed Forces*, the Public Key Infrastructure (PKI) capability may adopt the following components:

- Certificate Authorities,
- Local Registration Authorities,
- Certificate Directory,

and principles:

- use commercial products,
- use smartcards for protection of cryptography, digital signature, access control, keys and certificates.

REFERENCES

Burr W. E.. "*Public Key Infrastructure Technical Specification*", NIST, 1997.

DoD PKI Program Management Office, "*X.509 Certificate Policy for US DoD*", version 5.0, 1999.

DoD PKI Program Management Office, "*PKI Roadmap for DoD*", version 3.0, 1999.

Ford Warwick, Baum Michael , "*Secure Electronic Commerce – Building Infrastruture for Digital Signatues and Encryption*", Prentice Hall, 1997.

Gerk E., "*Overview of Certification Systems – X.509, CA, PGP and SKIP*", Meta-Certificate Group, 1998.

King, C., "*Building a Corporate PKI*", INFOSEC Engineering, 1999.

Marinier, F., "*25 Steps to the Implementation of a Corpoate PKI*", Labcal Technologies, 1999.

Patriciu,V.V., **Pietrosanu M., Bica I., Cristea C**., "*Securitatea informatică în Unix şi Internet*", Ed Tehnica, Bucureşti,1998;

Patriciu,V.V., **Pietrosanu M., Bica I., Voicu N**., **Vaduva C,** "*Securitatea comertului electronic*", Ed All, Bucureşti, 2000;

Patriciu,V.V.,"*Semnarea electronică a documentelor*", PC-Report, dec., 1998;

RSA Security S.A. , "RSA Keon Certificate Server Product Overview", 1999.

Schneier B., "*Applied Crypography*", John Wiley & Sons, 1996.

Developing Correct Safety Critical, Hybrid, Embedded Systems*

Alexander Pretschner, Oscar Slotosch, Thomas Stauner

Institut für Informatik, Technische Universität München

Arcisstraße 21, 80290 München, Germany

{pretschn,slotosch,stauner}@in.tum.de

Abstract

Several aspects of the development process of correct safety critical discrete and hybrid embedded systems are discussed. The general process and its support by the CASE tool AUTOFOCUS is outlined. This is illustrated along the lines of a simplified version of NASA's Mars Polar Lander. It is argued that specific aspects of hybrid systems do require the modification of classical theories on software development, and these modifications are discussed. The paper concludes by focusing on one part of the development process, namely testing. A novel approach to the automated generation of test cases for discrete as well as hybrid systems is presented. The Mars lander's crash serves as an example for the derivation of meaningful test cases.

Keywords. Reactive Systems, Validation, Development Process, Automatic Test Case Generation, CASE

1 Introduction

Safety critical systems. Developing correct safety critical software for hybrid, embedded systems is a difficult and error prone task. The functional reliability of the resulting systems is at least as important as security aspects. High quality of the resulting systems can only be achieved using a well structured development process. We present a development process that integrates many methods for quality assurance for discrete systems. In this paper, *discrete* systems refer to discrete event systems. However, those discrete event specification techniques we consider also have a discrete-time execution model. For mixed discrete-continuous systems (or "hybrid systems") we discuss elements necessary to obtain a similar integrated process, taking into account discrete as well as continuous aspects. The process for discrete systems is an extension of the V-model [22]. It is based on system models that are validated by

---

means of formal techniques.

AutoFocus. Discrete systems are modeled using graphical description techniques for structure, behavior, and interaction. In this paper we shortly present AUTOFOCUS, a tool prototype for modeling discrete embedded systems that we will use to model a hybrid system, namely a simplified version of NASA's crashed Mars Polar Lander. The models are based on a common formal semantics and can therefore be used to support the development process from the requirements engineering phase throughout the test phase. The existing features of AUTOFOCUS suffice to model discretizations of hybrid systems. These discretizations, however, usually alter the model which reduces the set of properties that can be derived from a system model.

Hybrid systems. The development of hybrid systems is an interdisciplinary task. Usually engineers from different disciplines are involved and must discuss their designs. Graphical description techniques are one element very useful to support this communication. Just as for safety critical discrete systems, it is furthermore desirable to apply a high degree of mathematical rigor in the development of safety critical hybrid systems. Today formal methods for hybrid systems are still an active area of research, and there are hardly any tools available which could yet be used in industrial practice. In this paper we therefore outline how formal tools for discrete systems (such as AUTOFOCUS) can be extended with aspects for continuous systems in a development process for hybrid systems that is feasible today. We discuss shortcomings of this method and outline how our work on hybrid modeling and validation leads to an *integrated* development process for hybrid systems which is close to the current AUTOFOCUS approach for discrete systems.

Testing. Much work has been devoted to checking validity and consistency of a specification. By now, testing is the only practicable, scalable means of validating the conformance of an implementation w.r.t. its specification, even though Dijkstra's popular remark that testing can only reveal the presence but never the absence of errors undoubtedly holds true. We discuss the role of testing as a complement to formal methods and present a novel approach

based on Constraint Logic Programming to automatically generating test cases for discrete as well as hybrid systems. Experimental results from a case study of a safety critical system, the Mars Polar Lander, are discussed. We also take into account the integration of testing processes into the development process of safety critical systems and our method's benefits and shortcomings.

Overview. The remainder of this paper is organized as follows. In section 2, we briefly discuss principles of a software development process for reliable discrete systems. Section 3 describes shortcomings of this classical approach w.r.t. hybrid systems and suggests modifications that remedy these shortcomings. Specifically, we argue for integrated description techniques right from the beginning. Section 4 then describes the CASE tool AUTOFOCUS and its description elements. Since so far there is no tool support for the integrated development process of Sec. 3, section 5 exemplifies the use of a discrete modeling tool for a hybrid system, the Mars Lander. Section 6 discusses the generation of test cases for discrete as well as hybrid systems along the lines of our example. Section 7 concludes this paper. Related work is cited in the respective sections.

2 Notes on the Development Process

For safety critical systems, we advocate a development process that heavily relies on *models*. In this process, models are developed in phases known from conventional software development processes. These models are then validated, and the last step is to generate code from them in order to get an executable implementation. A last validation step consists of testing the developed systems in interaction with their environment, for example together with other components or hardware. Model validation is the key factor for producing highly reliable programs for safety critical systems. Useful models should describe the developed system from different views by means of various hierarchical diagrams. The diagrams can be used to capture requirements, architecture, design and implementation decisions, and to represent test sequences.

Model validation may be seen as checking consistency. It can be applied at several levels: syntactic consistency (checking names), completeness consistency (checking references and types), semantic consistency (checking refinement relations), and adequacy [28, 4], i.e., conformance of a model with possibly informal requirements. Many available tools perform a syntax check with fixed built-in routines. For this purpose, modern tools use the object constraint language OCL [38] of UML which, in principle, could also be used to check completeness of the models. At present, CASE tools with useful semantic checks are not available. Recently, model checking tools have been connected to tools based on statecharts or SDL, but due to the complex semantics of these languages without practical relevance. A systematic generation of test sequences is also not available. Checking adequacy is reduced to simulation facilities.

AUTOFOCUS, on the other hand, also allows for semantic validation. These validation techniques are based on the simple and intuitive semantics of AUTOFOCUS [21], and can be used to support model-based development steps according to the V-model within all phases; in particular, testing is supported at the design, implementation, integration and system requirements levels.

Modeling hybrid systems with AUTOFOCUS is done by a simple discretization of the continuous behavior, and this approach allows to integrate discrete and continuous parts in a single model (as will be shown in the example in Section 5).

3 Hybrid Systems

In this section we outline how a conventional formal tool for discrete systems, such as AUTOFOCUS, can be used within the development of hybrid systems. We discuss advantages and drawbacks of such an approach, and present a more visionary approach not yet supported by tools. The new approach is supposed to prevent these drawbacks. It results from carrying over ideas like graphical specification with different systems views and model based validation based on formal methods to hybrid systems. A central characteristic of the proposed approach is that it is based on notations that have a clearly defined semantics.

A conventional development process. A conventional development process for hybrid systems builds upon isolated description techniques for purely discrete and purely continuous components. Popular in industry are tool couplings such as using Statemate together with $Matrix_X$ or the MATLAB/Simulink/StateFlow environment [11, 9]. For the development of safety critical systems we advocate the use of formal methods and notations wherever possible. This hinders the use of current commercial tools like Statemate, ObjectGeode, Rational RoseRT or Stateflow. Their notations only have a formal syntax, but *the semantics remains imprecise and ambiguous, or very complex.* A semantics for the coupling with continuous tools in not defined anyway. However, tools like AUTOFOCUS are available which support the design of discrete systems based on formal notations such as architecture descriptions, extended automata and MSC dialects [23]. For continuous systems there also are analysis and simulation tools based on block diagram notations, e.g. MATLAB [36]. Note that we regard block di-
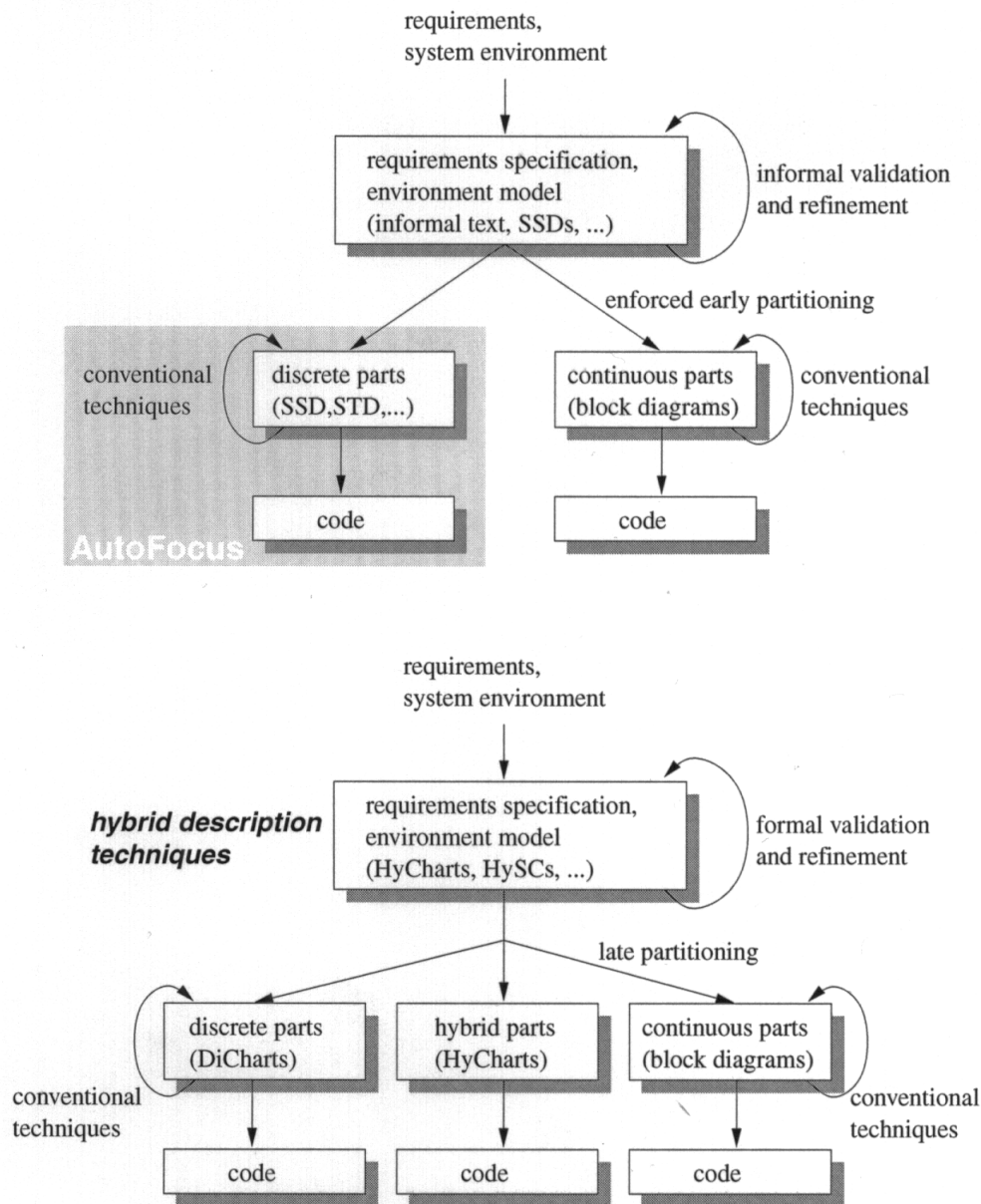
Figure 1: A conventional development process (top) and an integrated development process with hybrid description techniques (bottom).

agram descriptions of continuous systems as formal here, because a mathematical model can be associated with individual blocks and their interconnection in a straightforward manner. (Nevertheless, the user has to keep in mind that the selection of integration algorithms for simulation can have a great impact on simulation results and can cause them to differ strongly from the mathematical model.) As soon as the system under development is partitioned into discrete and continuous parts, a formal specification can be written down using these existing tools, see Figure 1, top. Well-known techniques from the discrete and continuous world can then be applied to the respective parts of the model. For instance, model checking and automatic test case generation may be used for the discrete part and analysis of eigenvalues for the continuous part.

Drawbacks. So far, the only currently available technique for examining properties of the mixed system is simulation. There are hardly any analytical methods regarding the mixed model and there are no techniques which support design modifications that affect both parts of the model. In fact such modifications could necessitate a redesign of the whole model.

Furthermore, in such a development process a designer has to perform a number of development steps informally, i.e., without documenting them with clearly defined notations, before a clearly documented process can start, i.e., a process relying on formal description techniques. In particular, these steps include partitioning the design into discrete and continuous parts which may involve an (implicit) discretization of some parts. This is unsatisfactory since

the partitioning decisions may be difficult to alter later on. Apart from that, the resulting coupled discrete continuous model often is not natural for some components of a hybrid system. For example, analog to digital (AD) and digital to analog (DA) converters, and in some systems the environment, are inherently hybrid components.

Recommendation. As no formal hybrid notations with tool support that is suitable in practice are available today, there currently is no real alternative to the outlined conventional development process. We therefore propose to use informal text coupled with formal descriptions where possible in this process: Writing down mathematical formulae expressing hybrid behavior directly is hardly reasonable for bigger systems. In the context of AUTOFOCUS we recommend using architecture descriptions (SSDs, see Sec. 4) already during the requirements capture phase and to describe the behavior of system components informally with text or, where practicable, with mathematical formulae, until the partitioning into discrete and continuous components has been performed. Figure 1, top, outlines the conventional process and its support by AUTOFOCUS. For the discrete part the model checking and testing techniques already implemented in the tool can be employed. They enable a more rigorous analysis of the discrete part than what is possible with other tools for discrete systems that do not have a formal semantics.

Outlook: An integrated development process. In a development process with hybrid description techniques, such as the one depicted in Figure 1, bottom, the designer is able to formally specify mixed discrete continuous models at early stages of the development process. In the context of formal methods, we refer to "refinement" as altering (or augmenting) a system's functionality without violating properties that have already been established. If validation and transformation techniques, such as simulation and refinement, are available for these description techniques, the model can be *systematically* designed to meet those system requirements which affect its discrete as well as its continuous aspects. Rudimentary versions of such techniques already exist and are an area of current research (e.g. [3, 10]). In later steps the model can be refined into discrete, continuous, and possibly some remaining hybrid submodels. For the discrete and continuous submodels conventional techniques can then be used to realize those properties which only affect the respective part. Thus, the availability of formal hybrid description techniques and supporting methods for them pushes the point at which systematic development, i.e. development with formal description techniques, can begin towards the beginning of the analysis phase. A partitioning into discrete and continuous submodels can be postponed towards subsequent development phases. Such a development process with hybrid description techniques

allows to obtain greater confidence in the model before a partitioning. Namely, testing and model checking techniques can be used to analyze requirements and refinement techniques can be used to guarantee some requirements by construction. By postponing implementation related questions changing requirements can more easily be taken into account. Thus, errors made in the initial development phases can be found earlier and are therefore cheaper to correct.

The development process we propose in Figure 1, bottom, is based on description techniques developed within our group in the last years. For requirements specification and environment modeling it uses the MSC-like notation *HySC* [15], and the combination of architecture diagrams and a hybrid automata variant which is subsumed in *HyCharts* [16]. A methodological transition from HySCs to HyCharts is ongoing work (for similar work on discrete systems see [24]). Succeeding steps in the figure refer to HyCharts rather than to HySCs. As notations for the discrete and the continuous part we propose DiCharts [17], a discrete-time variant of HyCharts, and (continuous time) block diagrams, repectively, taht can be integrated easily into the HyChart notation.

Note that the aspect of postponing the partitioning of a system into discrete and continuous parts is related to the area of hardware/software codesign [6]. There, the decision on which parts of a system are implemented in hardware and software is postponed to later phases. However, unlike hardware/software codesign the partitioning into discrete and continuous components proposed here does not yet imply how the components are implemented. The discrete part could be implemented in software or on digital hardware, the continuous part can be turned into a discrete-time model and implemented in software (or digital hardware), or it could be implemented in analog hardware.

While there is hardly any tool support for the integrated process today, a close coupling of discrete and continuous notations in the HyChart style is implemented in the *MaSiEd* tool [1], which also allows simulation. The *HyTech* tool [18] (or other tools, e.g., Uppaal or Kronos) which offers model checking of hybrid models is another element needed as support for an integrated development process. Presently, however, its application is limited due to scalability problems and deficits of the underlying hybrid automata model [29]. Promising tool approaches for the future should couple analysis algorithms like those implemented in *HyTech* with modular graphical description techniques, e.g. HyCharts, in comprehensive tool frameworks, such as accomplished with AUTOFOCUS for discrete systems.

Note that the development process for hybrid system proposed in [9] can be regarded as an intermediary between the two processes outlined here. There, the authors propose to complement block diagrams

and automata-based notations with formal specifications using Z [34].

4 AUTOFOCUS

AUTOFOCUS [19, 20] is a tool for graphically specifying embedded systems. It supports different views on the system model: structure, behavior, interaction, and data type view. Each view concentrates on a fixed part of the specified model.

Structural view: SSDs. In AUTOFOCUS, a distributed system is a network of components, possibly connected one to another, and communicating via so-called channels. The partners of all interactions are components which are specified in *System Structure Diagrams* (SSDs). Figure 4 shows a typical SSD. In this static view of the system and its environment, rectangles represent components, and directed lines visualize channels between them. Both are labeled with a name. Channels are typed and directed, and they are connected to components at special entry and exit points, so called *ports*. Ports are visualized by filled and empty circles drawn on the outline (the *interface*) of a component. As SSDs can be hierarchically refined, ports may be connected to the inside of a component. Accordingly, ports which are not related to a component are meant to be part of unspecified components which define the *outside world* and thus the component's interface to its environment. Components can have local variables to store values; these variables can be used to describe the behavior and the interaction of components.

Behavioral view: STDs. The *behavior* of an AUTOFOCUS component is described by a *State Transition Diagram* (STD). Figures 5 and 6 show typical STDs. Initial states are marked with a black dot. An STD consists of a set of *control states*, *transitions* and *local variables*. The set of local variables builds the automaton's *data state*. Hence, the internal state of a component consists of the automaton's control as well as its data state. A transition can be complemented with several annotations: a label, a precondition, input statements, output statements, and a postcondition, separated by colons. The precondition is a boolean expression that can refer to local variables and transition variables. Transition variables are bound by input statements, and their life-cycle is restricted to one execution of the transition. Input statements consist of a channel name followed by a question mark and a pattern. An output statement is a channel name and an expression separated by an exclamation mark. The expression on the output statement can refer to both local and transition variables. A transition can *fire* if the precondition holds and the patterns on the input statements match the values read from the input. After execution of the transition the values in the output

statements are copied to the appropriate ports and the local variables are set according to the postcondition. Actually the postcondition consists of a set of actions that assign new values to local variables, i.e., the assignments set the automaton's new data state.

Communication semantics. AUTOFOCUS components have a common global clock, i.e., they all perform their computations simultaneously. The cycle of a composed system consists of two steps: First each component reads the values on its input ports and computes new values for local variables and output ports. After the clock tick, the new values are copied to the output ports where they can be accessed immediately via the input ports of connected components and the cycle is repeated. This results in a *time-synchronous* communication scheme with buffer size 1. Values on the output ports are copied over the channels to the appropriate input ports and the cycle is repeated. This results in a non blocking synchronous communication.

Interaction view: MSCs. Message Sequence Charts (MSCs) are used to describe the interaction of components. In contrast to Message Sequence Charts as defined in [23], AUTOFOCUS MSCs refer to time-synchronous systems. In the following, the term MSC always denotes these time-synchronous sequence charts. Progress of time is explicitly modeled by ticks which are represented by dashed lines. All actions between two successive ticks are considered to occur simultaneously, i.e., the order of these actions is meaningless. An action in an MSC describes a message that is sent via a channel from one component to another.

MSCs can be used to describe requirements, simulation traces, counter examples (from model checking), and test sequences. Figures 7 and 8 show a typical test case specification as well as a satisfying test case that satisfies it.

Datatype view: DTDs. For the specification of user defined data types and functions AUTOFOCUS provides Data Type Definitions (DTDs). Definitions in DTDs are written in a functional style. For hybrid systems functions with continuous ranges can be defined, for instance:

```
const GMars = 3.73;
fun speed(last:Float,dt:Float)
        = last+GMars*dt.
```

Features of AUTOFOCUS. In addition to its modeling capabilities, AUTOFOCUS allows for checking consistency between views as well as simulating models (using OCL). For the German BSI (Federal Agency for Security in Information Technology) several validation techniques have been integrated into AUTOFOCUS [33]. The result is a model validation framework that supports

- model checking and bounded model checking to

check temporal properties (invariants) [31],

- abstraction techniques to *safely* reduce complex models to simpler ones,

- interactive theorem proving techniques to verify arbitrary security requirements, and

- systematic generation of test sequences and test cases [39, 26, 27].

All these validation techniques are based on the simple and intuitive semantics of AUTOFOCUS [21], and can be used to support model-based development steps. This framework also allows to generate Java and C code from the validated models in order to get executable implementations.

5   The Mars Lander

This section describes our example, a spaceship similar to the doomed Mars Polar Lander that allegedly did not complete its mission due to a faulty design [8]. The discretization technique we apply is common in the design of discrete-time control systems [30]. The next section then shows how automated test generation could have helped in avoiding this problem.

Behavior.   The hybrid system and its environment may be described by four main distinct states: The lander may be orbiting (orbiting) or falling freely after leaving its orbit (Rockets Off). Furthermore, the lander may have ignited its retro rockets (Rockets On) in order to slow down its vertical movement, and ant it may have landed (Landed). While orbiting, the lander's vertical speed, $v_\ell(t)$, is

$$v_\ell(t) \approx 0.0 \quad (1)$$

$$v_\ell(t) - v_\ell(t_0) \approx \int_{t_0}^{t} g_{mars} \, dt = g_{mars} \cdot (t - t_0) \quad (2)$$

$$v_\ell(t) \approx \int_{t_0}^{t} \left(g_{mars} - \frac{\dot{m}_\ell}{m_\ell(t)} \cdot v_f\right) dt + v_\ell(t_0) \quad (3)$$

$$\dot{m}_\ell \approx \int_{t_0}^{t} (c_1 \cdot (v_\ell(t) - v_{req}) + c_2 \cdot \dot{v}_\ell) dt \quad (4)$$

Figure 2: Mars Lander orbiting/landed (1), falling without (2) and with (3,4) rockets.

zero (Fig. 2-1). This behavior is also exhibited when the lander has landed. Once it has left its orbit, we assume it is freely falling without friction. Its behavior can thus be described by Fig. 2-2 where the planet's gravity, g_{mars}, is assumed to be constant. Strongly simplifying, the system's dynamic behavior in state RocketsOn may be modeled as follows. Let $F_w(t) \approx g_{mars} \cdot m_\ell(t)$ be the lander's weight force with $m_\ell(t)$ being the lander's mass. Furthermore, let $F_{thr}(t) = \dot{m}_\ell \cdot v_f$ be the rocket's thrust

force where $\dot{m}_\ell$ is the lander's (negative) change of mass and v_f the (negative, approximately constant) rocket's exhaust speed. By ignoring friction effects and letting $h(t)$ denote the lander's height one derives $\ddot{h} \approx g_{mars} - \frac{\dot{m}_\ell}{m_\ell(t)} \cdot v_f$ and thus the lander's vertical speed (Fig. 2-3) from $F(t) \approx m_\ell(t) \cdot \ddot{h} = F_w(t) - F_{thr}(t)$. For simplicity's sake, horizontal forces have been ignored. Oversimplifying again, we furthermore assume a simple PD controller for the lander, modeled by $\dot{m}_\ell \approx c_1 \cdot (v_\ell(t) - v_{req}) + c_2 \cdot \dot{v}_\ell$ for adequate gains c_1, c_2 and the lander's required speed, v_{req}. This yields the system's second descriptive equation (Fig. 2-4) for this state. The control variable is thus $\dot{m}_\ell$ (or $\ddot{m}_\ell$, respectively) which reflects an increase or decrease in fuel to be burnt.

In order to model the system with AUTOFOCUS, the above equations have to be discretized (i.e., linearized). Being the results of two AUTOFOCUS simulations, the curves in Fig. 3 have been obtained after discretization with step size $\Delta t = .01$. It shows height and velocity for two behaviors of the spaceship. After a certain time, it is caused to start its landing procedure by leaving the orbit (event "enter"; events are symbolized by long vertical arrows). When a maximum speed (45 m/s) is reached, the rockets are ignited ("rocketsOn"). Some time later, the legs are caused to open. From now on, the behaviors differ. The intended behavior is that when the spaceship actually lands, it should turn off its rockets (trajectories with annotation "rockets on"). *Ground contact is inferred from a shock in the legs.*

If, on the other hand, opening the legs causes the rockets to be switched off, velocity immediately increases which results in a crash (trajectories annotated with "rockets off"). This is what allegedly happened to the real spaceship: Opening and adjusting the legs caused some sensors in the lander to believe the spaceship had landed *for the legs sensed a shock.* (According to [8], engineers were well aware of this problem. When testing the system, they encountered a wiring problem, fixed it, and did not re-run their tests. Nonetheless, we will use this example as a motivation for a semi-automated generation of test cases in Sec. 6.)

Structure.   The above equations show that two main variables are involved, namely the change of mass, $\dot{m}_\ell$, and the lander's vertical speed, $v_\ell(t)$. This motivates the systems top level structure as described by the SSD in Fig. 4 that consists of three components: a Lander, a Physics, and a Controller component. All components receive the current time via channel T from the environment. The value of T is assumed to be present throughout every time slice, and to be increased by a constant value. The controller sends control commands to the other components, in order to switch the lander's rockets on and off, to enter the landing phase, and to open the legs for landing. It has a local variable CState to
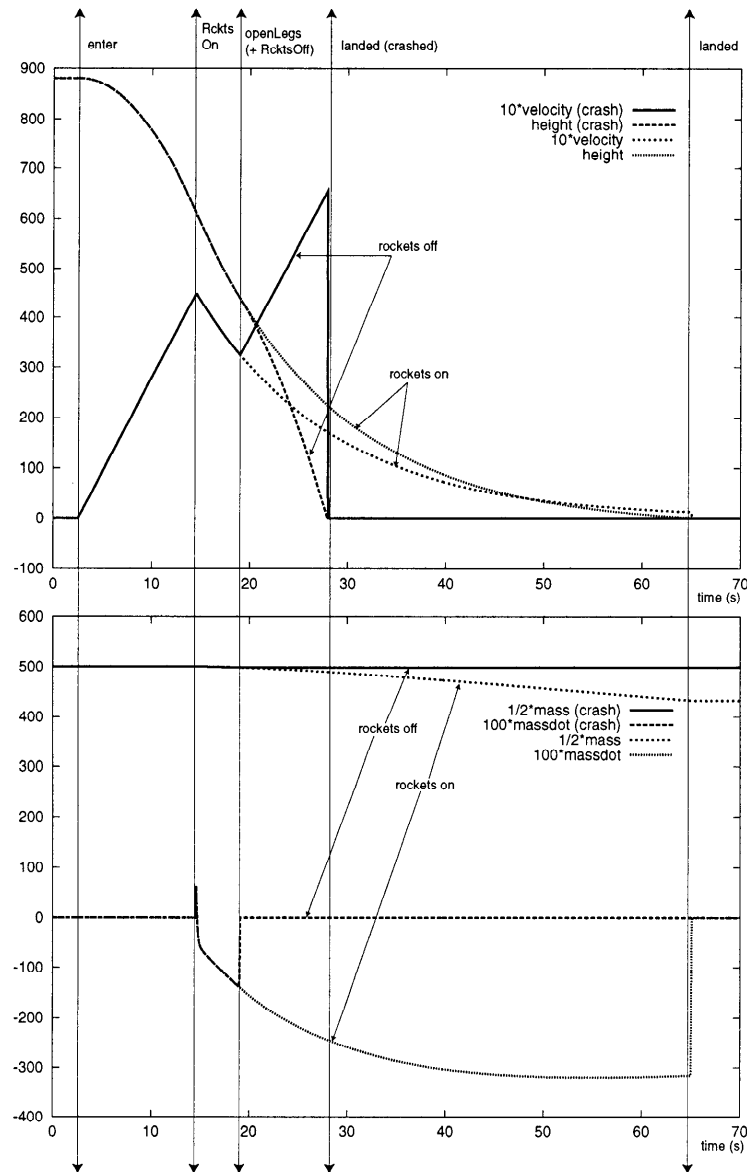
Figure 3: Spaceship crashes and lands.

record the spaceship's state since it needs to know when to issue which command. The initial value of this variable is Waiting; the type of this variable is a DTD data CState = Waiting | Ron | Roff. When port Sensor receives True, this should (!) indicate that the lander has sensed ground contact, and in turn its rockets will be switched off.

The differential equations of Fig. 2-3 and 2-4 are mutually dependent. In order to compute the values independently, the computation has to be separated into two subcomponents, namely Lander and Physics. The main interaction is between Lander, and Physics: the environment sends the current velocity to the lander (channel V), and the lander, in turn, sends its change in mass to the environment (channel Mdot). Channel Speed is only necessary for the initial value of the control process. Lander and Physics could have been grouped together into one hierarchic component. This is advisable if systems

become more complex. The behavior of component Physics is separated into two states (see Fig. 5). State Control Off just outputs the current speed (and does not react to changes of mass), whereas in state Control On, the new speed is computed from the last speed and the actual change in mass (Eq. 3). Component Physics has several local variables to store past values, used for integration, and differentiation, LastT:Float, for instance, to compute time differences. The denotation of the transition labels in Fig. 5 consists of functional terms computing new values according to the discretized equations. The behavior of component Lander is separated into four states (see Fig. 6), each representing a differential equation or the respective computation of new values (mainly for the local variables: LastT, LastV, LastM, LastH, and LegsOut) and for the output Mdot. In the initial state Orbiting the lander waits for the command enter from the controller (received at port
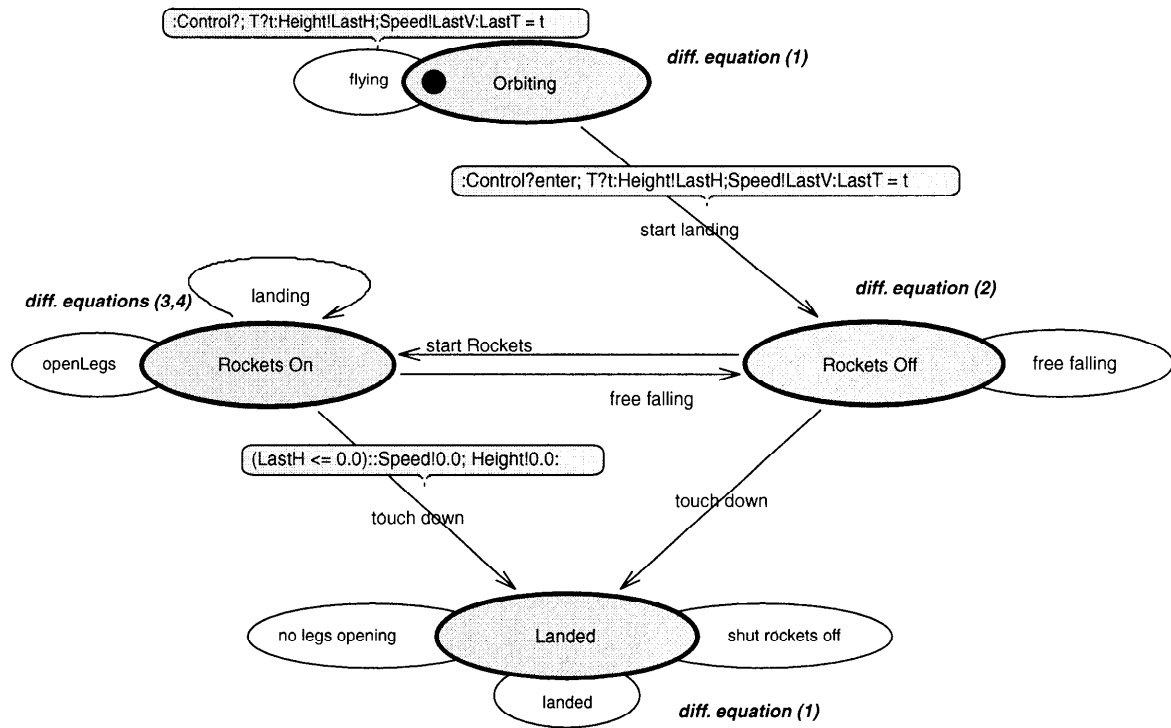
Figure 6: Behavior of Component `Lander`
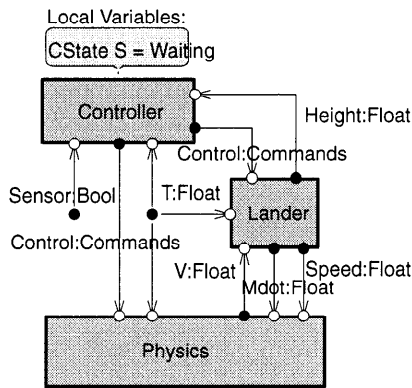
Figure 4: System Structure Diagram
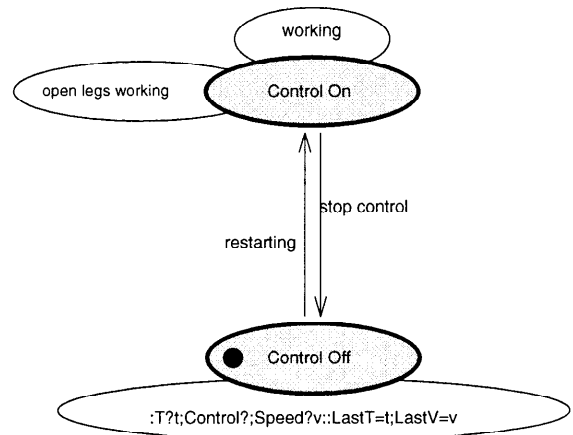
Figure 5: Behavior of Component `Physics`

`Control`). Unless no `enter` command is present, initial values are sent to the environment. This is done within the transition labeled `flying` and the following semantics:

- input pattern `Control?` denotes no input on port `Control`,

- input pattern `T?t` denotes that the input on port `T` is bound to transition variable `t`,

- output patterns `Height!LastH`, and `Speed!LastV` denote the sending of the current values of the variables to the output ports,

- action `LastT = t` stores the value `t` into the local variable `LastT`.

This last transition has no precondition (since `t>LastT` is a general assumption on time).

The states `Rockets On` and `Rockets Off` control the lander during the landing phase (with and without boosters); control commands `RocketsOn` and `RocketsOff` from port `Control` can be used to switch between the two respective equations. If the height is less or equal to zero in one of the states, the lander reaches the final state landed.

Shortcomings. Suitably discretizing a continuous model is a difficult problem. We chose a simple piecewise linearization with trapezoidal approximation. Problems with this approach include a conservative determination of Δt as well as meaningful

error estimations. Thus far, we use the same Δt for all components (in accordance with user-defined step sizes for each component). This approach may result in efficiency problems, but it solves the problem mentioned below for communicating components integrating over a same variable in the case this variable is t. The automatic generation of discretized systems from continuous equations is subject of ongoing work. Especially methods are developed to break systems of differential equations into single components, to determine appropriate discretization methods, and to to find good timing rates for the components.

6 Testing

Approaches to ensure a system's reliability include validating a model w.r.t. its specification as well as checking an implementation's conformance w.r.t. the specification. Formal methods, such as model checking, allow for determining a system's correctness in terms of user defined properties usually formulated in an (unintuitive) logic, e.g., the Linear time Temporal Logic LTL. Without suitable (and usually hard to determine) abstractions, model checking is restricted to finite state spaces which, for instance, typically grow exponentially with the number of variables involved. Not surprisingly, industrial applicability has not yet been achieved. In the following, we describe how a classical approach to quality assurance, namely testing, is supported by AUTOFOCUS. We advocate an integration of mathematically complete techniques (model checking) with testing. In addition to specifying test cases during the design phase, testing should also be done interactively, for certain errors can only be revealed by "playing around" with the model. This kind of testing may thus be seen as a debugging aid. This is, in fact, the case for most of the spectacular software faults the model checking/theorem proving communities use as a motivation for their work. The discussion of test management strategies and particular techniques such as mutation analysis and fault injection is beyond the scope of this article and thus omitted.

Applicability and Terminology. We distinguish between possibly informal requirements, a specification which is called a *model* if it is written down formally (e.g., in AUTOFOCUS), and an implementation. Testing an implementation is usually done w.r.t. its specification, e.g., [32, 26, 27]; the specification is thus considered to be correct. Obviously, this is a strong and usually unrealistic assumption. However, we think it is *one* necessary step. The techniques sketched below and explained in more detail in [26, 27, 39] allow for the determination of test sequences on the grounds of a test case specification. A test case specification is the formalization of some test purpose, i.e., reach a particular state or cause the system to throw a particular exception. Test

case specifications can, for instance, be written down as mathematical formulas [12], formal specifications [37, 5, 32], as MSCs [13, 26, 27, 39], as partial I/O traces or constraints over them [26, 27, 7]. A test case is an artifact that satisfies a given test case specification and may be formulated in the same forms as test case specifications. A test sequence, finally, is an executable test case, e.g., an I/O trace. [27] discusses this terminological framework.

Our work aims at (semi-) automatically deriving test cases from test case specifications that may be used for both, interactively white box testing a specification and (semi-)automatically black box testing an implementation. As indicated above, the interactive part plays an important role in the development process. Even though it is undoubtedly true that a system should be thoroughly thought over before it is implemented or modeled, we believe that simulation and interactive testing help in understanding a model. This is related to the rapid prototyping approach in software engineering; we even see simulation as a specialization of the testing process [26, 27].

However, it is worth emphasizing that most likely none of the commonly used approaches to quality assurance will do it alone. In contrast to formal methods testing is an inherently incomplete process. As formal methods yet do not scale to real size applications, this deficit has to be accepted but borne in mind. Dijkstra's popular remark that testing can only reveal the presence but never the absence of errors also applies to formal methods: One can only check properties that have been formulated by a human. This process, however, obviously is also necessarily incomplete.

Test case specification. The specification of test cases or properties to be checked requires intuitive and, if possible, graphical description techniques. One problem with formal techniques surely lies in the fact that without an intense formal education properties are hard to express in formalisms such as LTL or the Temporal Logic of Actions TLA [25]. We hence advocate the use of a variant of Message Sequence Charts [23] for the specification of test cases [13, 39, 26, 27]. MSCs (HySCs) are augmented with elements for talking about states in condition boxes [15] as well as constructs for expressing iteration and the necessity of certain transitions to fire. The identification of typical test purposes, e.g., causing the system to output certain values, reaching states, executing transition sequences [39], led to the incorporation of these language constructs.

An important concept is that of negation (negating transitions, the reachability of states, or forbidding certain inputs or outputs). However, a suitable semantics for MSCs in the context of test cases seems to be incomplete in the sense that between two elements in an MSC, arbitrarily many others may be present. Apparently the formal definition of a se-

mantics for negation in this context is not obvious [24] and subject of ongoing work.

In AUTOFOCUS, test cases may be specified by both LTL formulas and MSCs. In the following, we focus on the derivation of test cases from system and test case specifications. In the remainder of this section, the system specification should be thought of as an AUTOFOCUS model, and the test case specification is formulated using MSCs. Computed test cases (I/O sequences) are displayed in the form of MSCs themselves for inspection by a human (or comparison with expected test results, i.e., correspondence of the model's output with the output as described in the test case specification). Note that in this paper we concentrate on testing a specification and do not take into account testing implementations even though computed test sequences can be fed into an implementation for conformance testing with the specification.

Testing discrete systems. This paragraph briefly describes the generation of test cases from test case specifications by means of Constraint Logic Programming (CLP) as well as of propositional logic. These methods *automatically* derive test sequences from system and test case specifications.

CLP is the result of integrating two declarative programming paradigms, namely logic and constraint programming. Distinctive features include invertability of functions, the use of free (logical) variables that may be bound during program execution, built-in search mechanisms – backtracking –, and a semantics based not only on terms but rather on arbitrary domains. It turned out that AUTOFOCUS models can very naturally be translated into CLP languages. The idea is to feed the executable model with partial I/O traces and make the test case generation system create actual test cases (possibly partial I/O sequences subject to certain constraints, e.g. ranges for variables) by relying on the above mentioned built-in search mechanism and by using logical variables. By imposing constraints (e.g., in the forms of MSCs) on the set of all possible system execution, the search space can significantly be reduced. Further analyses such as automated interval analyses or (manually derived) classification trees [14] for variables then allow for the determination of meaningful test sequences (taking into account, for instance, range boundaries that yield equivalence classes to be tested). [26, 27] contain a more detailed description of this approach.

Testing based on propositional logic is suitable only for small finite systems (in particular, for systems with small, finite variable ranges). AUTOFOCUS models as well as test case specifications are translated into propositional logic and combined into a single formula which is fed into a propositional solver. The results (binding of free variables in traces) are translated back into MSCs. A detailed description of this approach which is related to bounded model
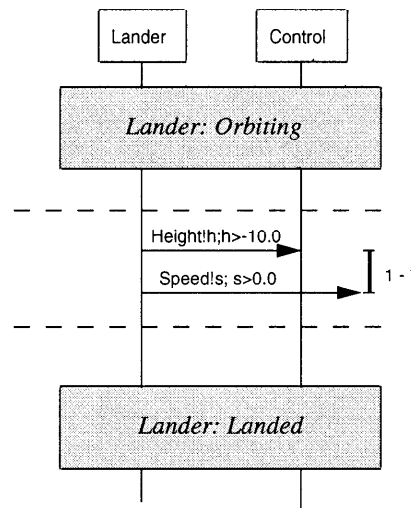
Figure 7: Test case spec.: Reach state *landed.*

checking [2] can be found in [39].

Testing hybrid systems. In principle, the above automatic CLP based generation of test sequences is also applicable to mixed discrete-continuous systems, for numerical or algebraic solvers can easily be connected to the CLP system. Yet, assuming that continuous activities take place within particular states of the system and that there is a *continuous* data flow between components, a number of problems arise. First of all, it is not clear how a continuous data flow can be simulated on ordinary computers (in control systems, however, there indeed is a continuous flow of data). Secondly, numerical solvers also discretize differential equations and solve these equations with different, possibly even dynamic, integration step sizes. It is not clear how to handle the situation where one component triggers a transition dependent on, e.g., the global time. Assume that two components run at different speeds, i.e., with different integration step sizes. If one component integrates over a common variable and meanwhile receives a value for exactly this variable that has been determined according to an earlier time, it has to stop its integration process and to step back. This results in severe methodical as well as efficiency problems, both of which are subject of ongoing work, based on (1) the semantics for hybrid systems as defined in [35] and (2) a modification of the AUTOFOCUS semantics where continuous activities do not take place on transitions but rather within states. This applies only to hybrid testing since real time simulation forbids re-calculating certain variable boundaries.

Example: Testing the lander. In accordance with these considerations, so far the methods for deriving test cases described above have only been implemented for discrete systems. As AUTOFOCUS is based on an inherently time-discrete semantics, this paragraph illustrates the derivation of test sequences for the discretized model of the Mars lander. Due

Figure 8: Test case: Lander crashes.

scenarios, one of which consisted of test case specifications with verdicts and has been created independently of the modeling process. The second scenario is closer to the area of rapid prototyping, where testing is seen as a debugging aid. In the case of Fig. 8, both scenarios may apply. However, there obviously is need for an engineer who derives from the test sequence that using a shock in the legs as ground detection mechanism is a bad idea!

7 Conclusion

A central aim of our work is the support of a systematic design of correct safety-critical hybrid embedded systems. For discrete systems we think that a number of effective validation and verification techniques has been integrated within the AUTOFOCUS framework. In this paper we presented an example of an ad-hoc discretization of a hybrid system, using discrete formal models. The model allowed an improved validation; in particular, important test scenarios have been derived. Secondly, precise requirements for dealing with hybrid systems in the context of discrete CASE tools have been obtained: (1) systematic discretization support, and (2) extending formal modeling and validation methods with continuous features to hybrid methods. Thirdly, a new development process for hybrid systems has been proposed and discussed.

In the future we will further evaluate how AUTO-FOCUS can be applied in the development of safety critical avionic systems and what is necessary to make it compatible with the certification process required for such systems. Furthermore, a testing medhodology (which test cases to choose, how many, etc.) is the subject if future work.

Acknowledgment. We would like to thank Michael van der Beeck for helpful comments on this paper.

References

[1] J. Albert and J. Tomaszunas. Komponentenbasierte Modellbildung und Echtzeitsimulation kontinuierlich-diskreter Prozesse. In *Proc. of VDI/VDE GMA Kongreß Meß- und Automatisierungstechnik*, 1998.

[2] A. Biere, A. Cimatti, E. Clarke, and Y. Zhu. Symbolic Model Checking without BDDs. In W. Cleaveland, editor, *Proc. TACAS/ETAPS'99*, LNAI 1249, pages 193–207, 1999.

[3] M. S. Branicky. Stability of switched and hybrid systems. In *Proc. 33rd IEEE Conf. Decision and Control*, 1994.

[4] P. Braun, H. Lötzbeyer, B. Schätz, and O. Slotosch. Consistent integration of formal methods. In *Proc. 6th Intl. Conf on Tools and Algorithms for the Construction and Analysis of Systems (TACAS'00)*, 2000.

to space limitations, we concentrate on just one test case specification which is, however, sufficient to convey the principal idea. Figure 7 shows the graphical specification for the test case "find a system run that makes component *lander* reach state *landed*". A derived corresponding test case specifying this specification is depicted in Fig. 8. Note the close relationship with the State Transition Diagram of Fig. 5. This system run makes the lander crash for its final velocity when touching the ground is much too high (approximately 67 m/s). Obviously, this is just *one* test case for the given specification. Another successful run leaves the rockets ignited until the spaceship actually has had ground contact. Both possible runs are depicted in Fig. 3 in forms of the respective variables' trajectories. Note that in case of the crash there is no automated means for assessing the outcome of a test case, nor some help in order to detect the fault. Above, we described two possible test

[5] E. Brinksma. A theory for the derivation of tests. In S. Aggarwal and K. Sabnani, editors, *Proc. 8th Intl. Conf. on Protocol Specification, Testing, and Verification*, pages 63–74, 1988.

[6] K. Buchenrieder and J. Rozenblit. Codesign: An overview. In *Codesign – Computer-aided HW/SW Engineering*. IEEE Press, 1995.

[7] A. Ciarlini and T. Frühwirth. Using Constraint Logic Programming for Software Validation. In *5th workshop on the German-Brazilian Bilateral Programme for Scientific and Technological Cooperation*, Königswinter, Germany, March 1999.

[8] CNN News. NASA: Premature engine shutdown likely doomed Mars lander. 28.3.00, www.cnn.com/2000/TECH/space/03/28/lander.report.02/.

[9] M. Conrad, M. Weber, and O. Müller. Towards a methodology for the design of hybrid systems in automotive electronics. In *Proc. of ISATA'98*, 1998.

[10] DFG. Priority program KONDISK (analysis und synthesis of continuous-discrete systems). www.ifra.ing.tu-bs.de/kondisk/, 2000.

[11] M. Fuchs, M. Eckrich, O. Müller, J. Philipps, and P. Scholz. Advanced design and validation techniques for electronic control units. In *Proc. of the International Congress of the Society of Automotive Engineers*. SAE International, 1998.

[12] M. Gaudel. Testing can be formal, too. In P. Mosses, M. Nielsen, and M. Schwartzbach, editors, *Proc. Intl. Conf. on Theory and Practice of Software Development (TAPSOFT'95)*, LNCS 915, pages 82–96, Aarhus, Denmark, May 1995.

[13] J. Grabowski. *Test Case Generation and Test Case Specification with Message Sequence Charts*. PhD thesis, Universität Bern, 1994.

[14] M. Grochtmann and K.Grimm. Classification trees for partition testing. *Software Testing, Verification, and Reliability*, 3:63–82, 1993.

[15] R. Grosu, I. Krüger, and T. Stauner. Hybrid Sequence Charts. In *Proc. of ISORC 2000*. IEEE, 2000.

[16] R. Grosu, T. Stauner, and M. Broy. A modular visual model for hybrid systems. In *Proc. of FTRTFT'98*, LNCS 1486. Springer-Verlag, 1998.

[17] R. Grosu, G. Ştefănescu, and M. Broy. Visual formalisms revisited. In *Proc. International Conference on Application of Concurrency to System Design (CSD'98)*, 1998.

[18] T. Henzinger, P.-H. Ho, and H. Wong-Toi. A user guide to HyTech. In *TACAS 95: Tools and Algorithms for the Construction and Analysis of Systems*, LNCS 1019. Springer-Verlag, 1995.

[19] F. Huber, S. Molterer, A. Rausch, B. Schätz, M. Sihling, and O. Slotosch. Tool supported specification and simulation of distributed systems. In B. Krämer, N. Uchihira, P. Croll, and S. Russo, editors, *Proc. Intl. Symp. on Software Engineering for Parallel and Distributed Systems*, pages 155–164. IEEE, 1998.

[20] F. Huber, S. Molterer, B. Schätz, O. Slotosch, and A. Vilbig. Traffic Lights - An AutoFocus Case Study. In *1998 International Conference on Application of Concurrency to System Design*, pages 282–294. IEEE Computer Society, 1998.

[21] F. Huber, B. Schätz, and G. Einert. Consistent Graphical Specification of Distributed Systems. In J. Fitzgerald, C. Jones, and P. Lucas, editors, *Industrial Applications and Strengthened Foundations of Formal Methods (FME'97)*, LNCS 1313, pages 122–141. Springer Verlag, 1997.

[22] IABG. Das V-Modell. www.v-modell.iabg.de, (documents also available in English), 2000.

[23] ITU. ITU-T Recommendation Z.120: Message Sequence Charts (MSC), November 1999.

[24] I. Krüger. *Using MSCs for design and validation of distributed software components*. PhD thesis, Technische Universität München, 2000.

[25] L. Lamport. The temporal logic of actions. *ACM Transactions on Programming Languages and Systems*, 16(3):872–923, 1994.

[26] H. Lötzbeyer and A. Pretschner. AutoFocus on Constraint Logic Programming. In *Proc. (Constraint) Logic Programming and Software Engineering*, London, July 2000.

[27] H. Lötzbeyer and A. Pretschner. Testing Reactive Systems with Constraint Logic Programming. In *Proc. 2nd workshop on Rule-Based Constraint Reasoning and Programming*, Singapore, September 2000. To appear.

[28] B. Müller. Unterstütung von Entwicklungsschritten auf Objekten mit unterschiedlichen OCL-Konsistenzanforderungen. Master's thesis, Institut für Informatik, TU München, 2000.

[29] O. Müller and T. Stauner. Modelling and verification using linear hybrid automata - a case study. *Mathematical and Computer Modelling of Dynamical Systems*, 6(1):71–89, 2000.

[30] K. Ogata. *Discrete-Time Control Systems*. Prentice Hall, 1987.

[31] J. Philipps and O. Slotosch. The quest for correct systems: Model checking of diagrams and datatypes. In *Proc. IEEE Asian Pacific Software Engineering Conference (APSEC'99)*, pages 449 458, 1999.

[32] S. Sadeghipour. *Testing Cyclic Software Components of Reactive Systems on the Basis of Formal Specifications*. PhD thesis, TU Berlin, 1998.

[33] O. Slotosch. Overview over the project Quest. In *Proc. of FM Trends 98*, LNCS 1641. Springer-Verlag, 1998.

[34] J. Spivey. *The Z Notation: A Reference Manual*. Prentice Hall, 2nd edition, 1992.

[35] T. Stauner and G. Grimm. Prototyping of hybrid systems - from HyCharts to Hybrid Data-Flow Graphs. In *Proc. of WDS'99 (satellite workshop to the 12th International Symposium on Fundamentals of Computation Theory, FCT'99)*, Electronic Notes in Theoretical Computer Science 28. Elsevier Science, 1999.

[36] The MathWorks Inc. MATLAB. www.mathworks.com/products/matlab/, 2000.

[37] J. Tretmans. Test generation with inputs, outputs and repetitive quiescence. *Software–Concepts and Tools*, 17(3):103–120, 1996.

[38] J. Warmer and A. Kleppe. *The Object Constraint Language*. Addison-Wesley, 1998.

[39] G. Wimmel, H. Lötzbeyer, A. Pretschner, and O. Slotosch. Specification Based Test Sequence Generation with Propositional Logic, December 2000. J. Software Testing, Verification & Reliability (STVR): Special Issue on Specification Based Testing. To appear.

Wireless Tactical Networks in Support of Undersea Research

Alessandro Berni and Lorenzo Mozzone
NATO SACLANT Undersea Research Centre
Viale S. Bartolomeo, 400
19138 La Spezia
Italy

Introduction

Emerging concepts for Anti-Submarine Warfare (ASW) and Rapid Environmental Assessment (REA) increasingly rely on communication technology, in order to implement distributed information networks and to exchange information between naval units and military commands ashore. The necessary communication links could be accomplished using a variety of solutions: our main focus is on radio frequency (RF) links, which offer easy deployment and flexible operations.

Requirements (such as transmission data rate) change from one specific application to another. There are however a number of prerequisites that are shared by all applications and users: they include, but are not limited to, reliability, availability and security.

The biggest challenge derives from the fact that those requirements are countered by either natural factors, such as thermal noise and multipath interference, or by hostile activity aimed at disrupting the integrity of the lines of communication.

This document illustrates how spread-spectrum techniques can be adopted to substitute and enhance existing communications systems, to permit the deployment of distributed, scalable networks of ships and sensors, characterized by reliable performance (resistance to hostile jamming and environmental interference) and low probability of interception. An overview of real applications in ASW and REA is presented.

Wireless communications at sea

Sophisticated sensors and information technology require the transfer of large flows of data. Current communication systems are limited in their throughput capability not only by technical limits, but above all by regulations about frequency assignment and by interference inherent to the frequency bands in use. For example, frequencies ranging from 2 MHz to 2 GHz, where the majority of naval communication systems are located, are crowded with signals generated by commercial users and radar emissions. In addition to that, the bandwidth that is made available to a single user is limited by practical reasons and by international regulations.

Frequencies above 2 GHz are less densely populated, and wider RF bandwidths are available. Operating at those high frequencies solves the problem of bandwidth availability and limits the adverse effects of thermal noise: it is a well known fact that the sky noise temperature is minimum at frequencies between 1 and 10 GHz, in the so-called *microwave window* [Skla-88], as shown in figure 1. Moreover, stronger atmospheric absorption reduces interference from other users of the same channels. On the other hand, a large number of issues are still to be considered: for example, resistance to multipath interference and hostile jamming. Efficient use of radio communication resources is also important, to allow scalable and efficient operations with dynamic deployment topologies for ships and sensor platforms.
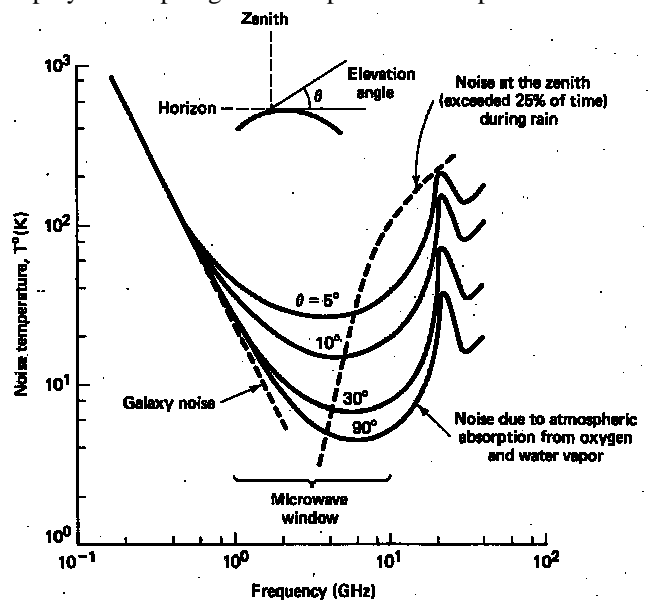
Figure 1 – Sky Noise temperature versus frequency (from [Skla-88])

Line of sight communications take place with low grazing angles; they correspond to the flattest curve of the above plot ($\theta = 5$ degrees). The most interesting frequency band appears to be in the 2 GHz – 10 GHz range. Further system parameters (e.g. antenna size, data rate, range to be achieved) need to be considered in the final choice of the operating frequencies.

Benefits of Spread-Spectrum communications

In the study of digital communications systems, spectral and power efficiency represent important qualifying factors: in general terms, it is necessary to exploit effectively available radio frequency bandwidth and transmission power. In addition to that, technical shortcomings could be present, such as operation from a platform where available electrical power is limited (e.g. a battery-powered buoy).

Spectral efficiency is practically defined as:

$$n_s = \frac{R}{B_W} = \frac{transmission\_rate}{required\_bandwidth}$$

where

n_s = Spectral efficiency or bandwidth efficiency defined in b/s/Hz

R = Transmission rate in b/s

B_W = Bandwidth

Traditional modulation schemes aim at *maximizing* spectral efficiency. However, interesting advantages can be obtained using a very large radio frequency bandwidth. Such a transmission technique is called *spread-spectrum* (SS) and has been developed since the mid-1950's in support of military applications, including secure tactical communications. In SS systems, the bandwidth spread is achieved using a code that is independent of the data: subsequent despreading and data recovery at the receiver is performed with the same code, in association with synchronization techniques. Spreading the spectrum in the appropriate way leads to multiple simultaneous benefits. Some of these are:

- Improved interference rejection
- Antijam capability
- Code division multiplexing for multiple access applications
- Low-density power spectra for covert transmission
- High-resolution ranging and timing
- Secure communications

The techniques by which the signal spreading can be accomplished are various: the most common and interesting techniques are termed *direct-sequence* and *frequency hopping*. In *direct-sequence spread-spectrum* (DSSS) a fast pseudo-random sequence, termed *pseudo-noise* (PN) *sequence*, is applied to the signal that needs to be transmitted, causing phase transitions in the data carrier. In *frequency-hopping spread-spectrum* (FHSS) the PN sequence forces the carrier to rapidly drift its frequency in a pseudo-random way. The use of a larger portion of RF bandwidth is compensated for by the interference advantages anticipated above. Signal energy becomes so diluted that the amount of power density present in any point within the spread signal is very limited. The dilution may result in the signal falling below the noise floor, and thus becoming invisible to receivers that are not sharing an appropriate despreading code.

DoD and Commercial-off-the-shelf implementations

The U.S. Department of Defense is in the process of specifying and developing a new wireless networking radio capable of transmitting voice, video, and data between its mobile vehicles, aircrafts, and ships. Among the most promising systems that are being evaluated are the Department of Defense Near Term Digital Radio (NTDR), GEC Marconi Hazeltine VRC-99, as well as

Commercial-Off-The-Shelf CSMA/CA systems, such as those defined by standard IEEE 802.11. Specific research projects are being conducted at the Naval Research Laboratory and at SPAWAR Systems Center, San Diego [8].

The NTDR is a tactical radio developed by ITT for mobile networked Internet protocol data applications. The NTDR handles data packet information at a burst rate of 375 kbps. d. The resulting coded signal is modulated onto a Direct Sequence Spread Spectrum (DSSS) waveform at 500 kbps and transmitted at up to 20 Watts (+43 dBm) in the 225-450 MHz band. For a simple point-to-point connection the maximum throughput is about 250 kbps.

The VRC-99 is a direct sequence spread spectrum radio manufactured by GEC-Marconi Hazeltine guaranteeing reliable, simultaneous, multichannel voice, data, imagery, and video transmission. Data rate is 625 kbps with options of adaptive data rate operation from 157 kbps to 10 Mbps. Low probability of intercept and jamming resistance is achieved through specialized direct-sequence and frequency-hopping spread-spectrum techniques. Operation is conducted in the 1.2-2 GHz frequency band. The VRC-99 supports IP data from a standard Ethernet local area network (LAN) and up to four simultaneous voice telephone links.

In addition to the above systems, which are specifically conceived for military use, Commercial-Off-The-Shelf systems are being studied to assess their effectiveness. COTS systems are usually based on open standards, which implies availability of products from multiple suppliers on a shorter time scale, at a lower cost. It is possible either to purchase a complete turnkey system or to build up a custom made one to match specific requirement, using available chipsets and RF components.

COTS equipment is typically configured to operate in the ISM (Industrial, Scientific and Medical) frequency band (2.4 GHz and 5.7 GHz). License-free operation is granted for transmission in the ISM band using spread-spectrum equipment, provided limits on transmit power are respected (typically 0.1 W). Operation at higher transmit power is usually subject to approval by the national authorities.

"802.11" is the first true industry standard for Wireless Local Area Networks, or "WLANs". Developed by the Institute of Electrical and Electronics Engineers (IEEE), 802.11 can be compared to the 802.3 standard for Ethernet wired LANs. The goal of 802.11 is to provide a standard set of operational rules so that WLAN products from different manufacturers interoperate in the same way that Ethernet equipment does today.

The physical layer of IEEE 802.11 includes three alternatives: DFIR (Diffuse Infra-Red), DSSS, and FHSS. Both the DS and FH spread spectrum specifications utilize the 2.4 GHz radio frequency band. The 2.4 GHz band was chosen because it is available for unlicensed operation worldwide and because it is possible to build low cost, low power radios in this frequency range that operate at LAN speeds.

The shared access of several users to a single frequency band is implemented by ad-hoc protocols.

IEEE 802.11 and other relevant WLAN products, such as P-Com Datametro II, use the CSMA/CA (Carrier Sense Multiple Access/Collision *Avoidance*) access method, which is, in turn, derived from the Ethernet standard, CSMA/CD (Carrier Sense Multiple Access/Collision *Detection*). In a nutshell, CSMA/CA applies a "listen before talk" approach that can minimize collisions by using request to send (RTS), clear-to-send (CTS), data and acknowledge (ACK) transmission frames, in a sequential fashion. This results in higher system throughput and better frequency utilization.

Applications

Naval operations and field experiments both require underlying communication architectures capable of delivering the required services. The following table lists some key services that are of interest to naval users, together with the associated data rates.

Service	Data rate (kbps)
Automated real-time decision aides	>64
Concurrent, distributed data bases	>64
Data fusion	>64
Data transfer (vertical arrays, sonobuoys, DUSS)	Up to 10 Mbps
Distributed computing	Up to 10 Mbps
Distributed white board	>64
High-resolution imagery	>64
High-resolution maps	>64
Voice	>2.4
Web browsing	>64

Important requirements relative to availability, reliability and security are summarised in the following list:

1. Guarantee of data integrity
2. Protection from denial of service
3. Low battery power requirements
4. Low probability of intercept
5. Source authentication

Spread-spectrum with no doubt fulfils requirements 2, 3 and 4.

Low battery power requirements and low probability of intercept are strongly correlated: spread-spectrum transmission is stealth by its own nature. In addition to that, effective data transmission is accomplished with limited transmission power. Hostile interference is not fought by brute force (increasing the transmission power), but by more advanced techniques. The result is that spread spectrum equipment can be easily operated from a battery: Commercial-Off-The-Shelf spread-spectrum radios are capable of delivering data rates in the order of several Megabits per second up to the distance of 20 n.mi., with transmission power lower than 1 W.

Requirements 5 and 6 cannot be satisfied totally by simply adopting spread-spectrum techniques: better results would probably be obtained operating at the application level, using appropriate authentication schemes which should be supplemented by strong encryption. Nevertheless, since a spread-spectrum wireless network node can communicate with the other nodes only by using the appropriate PN code, it can be said that SS plays a non marginal role in satisfying source authentication requirements.

Rapid Environmental Assessment

To accomplish their mission, naval forces rely on all kinds of environmental data. Weather and climatology information, tide atlases, satellite images of cloud cover, and oceanographic databases are regularly generated and distributed: this activity of preparation, collection and delivery of standard environmental products represents the standard level of service offered to navy customers. A situation may however arise, when oceanographic support centres are not able to supply the environmental information that is needed by naval commanders and planners. In such a case, a request is issued to the NATO military oceanography (MILOC) group, to conduct specific surveys in the area of interest to acquire the necessary data.

In the past, the whole process of data collection, processing and distribution of a final environmental report always took several years. Such a long delivery time was not deemed satisfactory by the end users, which formulated a requirement to obtain operationally relevant products within a tactically relevant time frame after the identification of a particular area. The definition of "tactically relevant time frame" ranges from several months, during the operational planning phase, to few days, during navy operations.

Acceleration of traditional methodologies (e.g. re-assignment of priorities in data collection processing) can reduce the delivery times from several years to few months, which is not even close to meeting the most stringent requirements.

Rapid Environmental Assessment (REA) is a set of state of the art methodologies and techniques that enable the collection, processing and distribution of environmental data and products within a compressed time frame, in the order of days, if not hours.

The REA concept requires the collection of in-situ data by one or more survey vessels, as illustrated in figure 2. Survey vessels perform the necessary environmental measurements and relay the information to a data fusion centre afloat (e.g. a command ship), which can be positioned several miles away. Fused data are then transferred to a fusion centre ashore, where they are made available to the REA community (data providers, product developers, and customers) using wide-area computer networks. [Sell-99], [TFB-99].
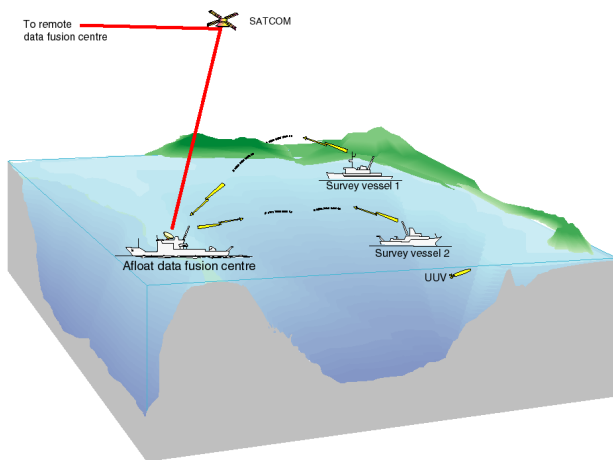
Figure 2 - Transmission of REA information from survey vessels to a data fusion centre

REA experiments conducted so far constitute proof of the effectiveness of the REA concept: however, additional efforts are still necessary to define a communication architecture suitable for use in operational conditions. There is no doubt that the availability of reliable and scalable ship-to-ship data links is of paramount importance to the effectiveness of REA surveys.

During a crisis the access to commercial land-based communication infrastructures (such as cellular phone networks) will be easily denied. It is very important that the communication architecture that is defined is independent from local infrastructures. On the basis of the above consideration, the deployment of a wireless LAN connecting the survey vessels is a practical solution.

The WLAN, implemented using medium or high data rate RF links with a high resistance to multipath interference and hostile jamming, is the first level of the REA tactical network. The second level is the long-range communication link, typically implemented using SATCOM, which connects the survey group to the commands ashore.

Spread-spectrum makes an excellent candidate to implement the REA tactical network. When operating near the shore, a major challenge comes from the multipath interference that is caused by reflections of RF waves by the sea surface and by the land. The characteristics of robustness to multipath interference of spread-spectrum enable the delivery of reliable wireless data communication at the required high data rate. The LPI and LPD properties of spread-spectrum, together with resistance to hostile jamming, may prove extremely useful for adoption in an operational situation.

Computer simulations using the AREPS model [Patt-98] show that ship-to-ship communications can achieve a maximum range of 18 n.mi. (21 n.mi. when directive antennae are used) with only 1 W of transmit power, operating in the 2 GHz frequency range with antenna masts 33 m high, as summarized in the following table.

Transmit power	Antenna type	Range (n.mi.)
1 W	Omnidirectional	18
1 W	Directive (*)	21

(*) requires tracking system to ensure antenna alignment

Tests have been conducted during SACLANTCEN experiment **Advent 99** in May 1999, to demonstrate the effectiveness of the concept. Italian Navy ship Ciclope towed a CTD chain and transferred data in quasi-real time to NRV Alliance, using a DSSS link with a data rate of 128 kbps and a transmit power of 0.6 W. NRV Alliance was positioned at a maximum distance of 10 n.mi.
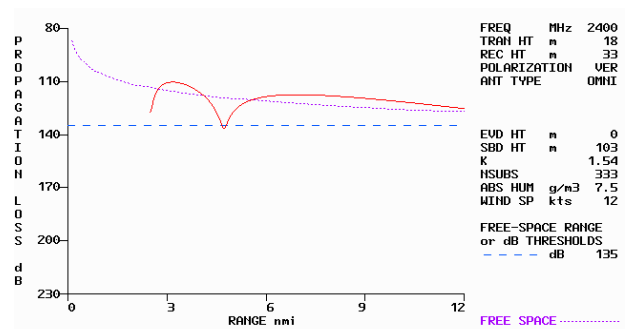
Figure 3 - Propagation loss (red) for the Advent 99 DSSS radio link. The blue dashed line represents the communication threshold

The spread-spectrum communication system was judged extremely reliable and contributed to the success of the experiment.

A brief interruption of the link has been observed, as a consequence of the multipath effects due to reflections from sea surface. This is in accordance with RF

propagation loss predictions computed using the EREPS model and the methodologies described in [MBG-99]. The result is summarized in figure 3: multipath induced propagation loss brings the system below communication threshold at the range of 4.7 n.mi. Loss of signal is observed for an additional 0.1 n.mi, where communications is resumed. From the range of 4.8 n.mi onwards, communication is stable and reliable, up to a range exceeding 12 n.mi.

The first operational application of the concept took place during ACLANT exercise Linked Seas 2000, in May 2000. A REA precursor phase has been conducted, in support of amphibious operations: HMS Roebuck, SPS Don Carlos and FS La Perouse were linked by a WLAN implemented using P-Com Datametro units, as illustrated in fig. 2.

All data collected in the course of the exercise (in the order of 30 MB) were distributed among the participating vessels and transferred to the SACLANTCEN data fusion centre using SATCOM.

Deployable Underwater Surveillance Systems

ASW nowadays places very demanding requirements on sonar capabilities, in an era of shrinking budgets. Protection of worldwide commercial flows on sea routes from the attacks or mining operations of hostile submarines have become the highest priority. The areas of interest have now moved to shallow littoral waters and choke points, characterized by heavy (and noisy) shipping traffic, poor and complex sound propagation, strong reverberation. The submarines that need to be countered are now small, diesel electric vessels, already too silent for easy passive detection, and now undergoing major technological improvements like sound absorbing cladding, air independent propulsion, modern passive sonars (towed and flank arrays) and navigation systems. Finally, operational requirements have become very strict in terms of probability of detection, completeness and accuracy of coverage, risk for naval units and personnel, discretion, interoperability.

SACLANTCEN is conducting theoretical and experimental investigations on Deployed Undersea Surveillance systems to meet these requirements. The new concept consists of multiple acoustic sources and receivers: small, inexpensive, expendable elements easily deployed from any air or sea platform (Figure 4).

Sonar units can be drifting, moored to the bottom at chosen depths or laid directly on the seabed. Battery powered, they are autonomous and transmit data in compressed form via radio, satellite link, optic fiber or other. The moderate gain receivers, which may be used also in passive mode, have the advantage of no own ship noise and can be laid to form a large network extended over wide areas. The covertness of receivers suits operations in critical areas and improves chances of

detection of unaware submarines (that optimize their course only with respect to the transmitter position) with either a favorable aspect angle or Doppler shift.

Figure 5 illustrates the advantages of aspect diversity.
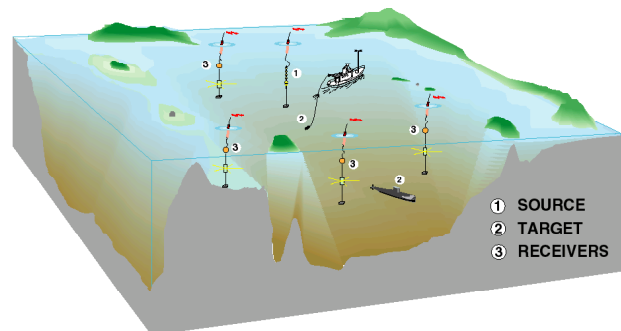
Figure 4 - Pictorial view of the DUSS concept.
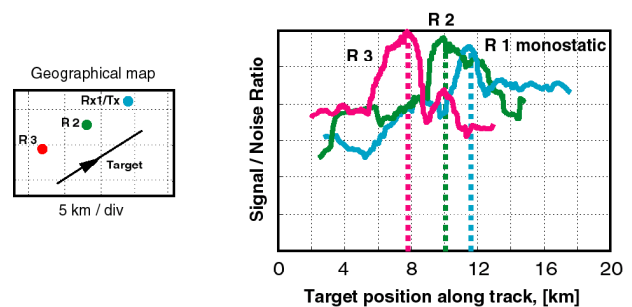
Figure 5 - Target detection opportunities are multiplied by the deployment of additional multistatic receivers.

Elementary detection volumes can be overlapped: each node has an independent opportunity to detect the target, while inter-sensor data fusion reduces false alarms and exploits deployment geometry to enhance localization and tracking.

A field of moored elements facilitates detection of any moving object, as the background remains relatively stationary. Towed sources, on the other hand, are not subject to the same critical constraints in power and endurance, while the towing ship does not need to be the master operation center.

Figure 6 shows raw sonar displays for three separate receivers.

The feasibility of the whole concept largely relies on the recent advances of communication, geographical localization and digital processing techniques, which permit distributed sonar processing on small autonomous nodes and transmission of the data collected in compressed form. Different degrees of data reduction before transmission are possible, according to the tasks to be executed. The following sections address the two

cases of scientific experiments and of demonstrators approaching operational conditions.
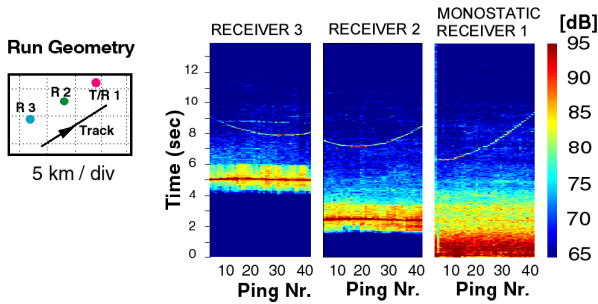
Figure 6 - Raw data on sonar displays for three DUSS receivers.

DUSS radio link specifications
Accurate collection of scientific data requires hydrophone data to be stored with a large dynamic range. The whole bunch of information collected by the sensors is transmitted to the laboratory on NRV Alliance and maintained for successive off-line processing and analyses that can not be anticipated during the experiments. Data reduction is limited, in this case, to base-band filtering and shifting. The follow on estimates requirements for DUSS.

Raw data: recordings of raw data on board Alliance are mandatory for scientific work. Therefore they will always be required as a first – priority output for all experiments. In-buoy recordings are also being considered, and represent a safe backup resource. Expected rates are 24 bits * 64 Hydr. * 1365 Hz = **2 Mbps**

Compressed data: an operational DUSS may resorts to in-buoy processing and implement a distributed - knowledge, distributed-processing network. Pre-processed contact information is passed through the net in a packet switching fashion. Data rates are reduced. This section tries to produce an estimate of expected reduction ratios by considering very simple approaches.

- Geographical map of cells: range resolution reduced to 50 m via pre-processing. Max range: 30 km. 600 cells. 8 bits per cell (e.g. level coded in dB of SNR after normalization and thresholding between 0 and 32 dB in 0.5 dB steps). 64 beams. Total: **5 kbp**s.
- Markov Random Field processing: this method transforms the raw sonar display into a list of OBJECTS. Each of them is described by time, bearing, size, level (i.e. 16 bytes). Analyses of experimental data (DUSS and SWAC data) estimated an average reduction between 1% and 0.1% of the number of raw sonar cells to objects. Assuming 2 bytes/cell and 16 bytes/object, 1% data

rate reductions represent a reasonable conservative guess for reverberating environments after full development of this method. Total: **20 kbp**s.
- Thresholding: The experimental probability of false alarm is plotted versus level in dB with respect to normalized background. Contacts above 6 dB represent 1 % of the total. Their range needs to be stored in sparse files. Their level can be recorded with just 1 byte. The result is a 1 % compression ratio. Total **20 kbp**s.

Estimated rates sum up together when a Local Area Network (LAN) structure is used, with nodes that forward data through the network together further to broadcasting their own contacts.

- Bit error rate (BER) requirements for raw data acquisition range from 10^{-3} to 10^{-5}. In fact, as shown by field experiments, the isolated error bursts that occur on radio data links can easily be detected, and do not affect the measurements.
- On the contrary, compressed data require lower BER, around 10^{-8}, thus partially losing the advantage of a lower data rate.
- Typical working ranges, necessary for significant multistatic sonar tests, are of 10 n.mi.

Accurate collection of scientific data requires hydrophone data to be stored with a large dynamic range. The whole bunch of information collected by the sensors is transmitted to the laboratory on NRV Alliance and maintained for successive off-line processing and analyses that can not be anticipated during the experiments. Data reduction is limited, in this case, to base-band filtering and shifting. The radio link needs to face a massive flow of data, from 2 to 6 Mbps, with bit error rate requirements ranging from 10^{-3} to 10^{-5}. At the same time, long ranges are necessary for significant multistatic sonar tests (10 n.mi).

DUSS radio link experiments
A conventional radio link has been implemented and successfully tested at sea. Antenna mast height above the water, transmit power, receive antenna gain represent the most critical issues. Two different working frequencies have been tested, 0.4 and 2.3 GHz, with equivalent overall performance, but different features. Figs. 9 and 10 show the reconstruction of propagation conditions after model validation with experimental data. The objective of 10 n.mi. ranges at 2 Mbps was accomplished by both frequencies with powers of 6 W and antenna gains of 22 dB (2 GHz) or 20 W with antenna gains of 7 dB (400 MHz).

As discussed above, spread-spectrum techniques and other signal modulation and coding schemes deserve serious attention. Their interaction with the peculiar environment of the present application (air-sea

boundary) may result in remarkable improvements of system performance and reliability. The most relevant limiting factor derives from multipath and destructive interference from the sea surface reflected signal. Spread spectrum is expected to counter such effects, providing at the same time a better rejection of man made interference.

Computer simulations have been conducted with AREPS to determine the practical transmission range for DUSS with spread spectrum techniques, in three representative cases. Case A accounts for an elevated communications relay station (33 m above sea level), that could either be the mast of a support ship (e.g. NRV *Alliance*), a moored buoy with a communications payload connected to a tethered balloon, or an antenna installed ashore, in case of littoral surveillance (e.g. a harbor). Cases B and C are representative of LOS buoy-to-buoy communications, with antennae 6 m and 9 m above sea level, respectively.

	Transmit power	Range (n.mi)
Case A (33 m)	0.1 W	6.1
Case A (33 m)	1 W	10
Case B (6 m)	0.1 W	3
Case B (6 m)	1 W	4.7
Case B (6 m)	6 W	6.8
Case C (9 m)	0.1 W	4
Case C (9 m)	1 W	6.6
Case C (9 m)	6 W	9

Validation of this estimation has been produced during engineering tests conducted with the following link:

Lucent WaveLAN IEEE Turbo at 0.5 Mbps. 2.4 GHz, 100 MHz passband, 1 Watt RF power.

- 9 dBi omni antenna at 33 m on Alliance, 11 dBi omni co-linear antenna at 30 m on Formica Grande island. Link up to 12 km range, with periodical signal loss.
- 9 dBi omni antenna at 33 m on Alliance, 15 / 24 dBi directive Yagi / parabolic antenna at 20 m on Formica Grande island. Link up to 25 km range, with periodical signal loss.
- 9 dBi omni antenna at 33 m on Alliance, 7 dBi omni co-linear antenna at 7 m on buoy. Link up to 6 km range, with periodical signal loss.

The result is that, adopting spread-spectrum transmission, it is possible to communicate at a range compatible with DUSS operations, using a limited transmit power (1 W). This result is extremely important when operating from a battery-powered platform, where a trade-off exists between transmit power and battery duration. The execution of practical tests at sea is warmly recommended, due to the variability and uncertainty introduced by the environment, and is made easier by the reduced costs and efforts involved.
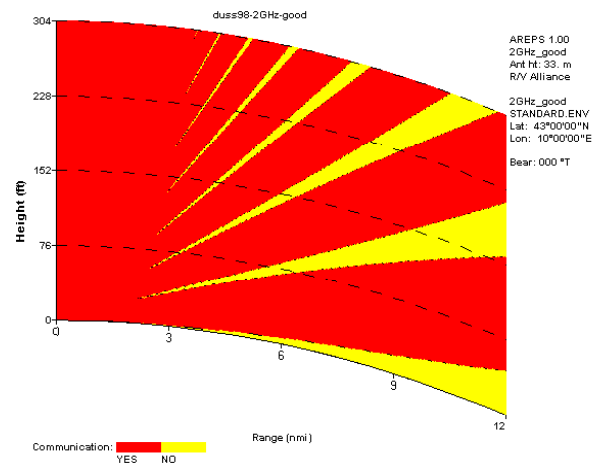
Figure 7 - AREPS output for 2.28 GHz link.
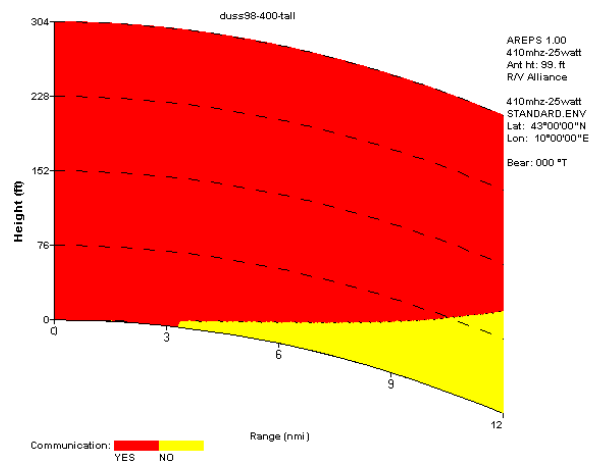
Figure 8 - AREPS output for 400 MHz link.

The concept of a "repeater" buoy is also promising, in the attempt to align radio ranges to acoustic performance. The following figure 9 illustrates a tactical network of DUSS nodes configured according to Case A, described above. A repeater buoy is used to reach buoys positioned beyond line of sight (LOS).

The availability of cost-effective off-the-shelf systems also represents a very attractive issue for scientific applications. Such systems can be very easily interfaced to a computer-controlled buoy, thus providing additional advantages in terms of remote operation of the system. Finally, no special licenses are required for the access to existing channels, thus making easier the organisation of tests at sea.
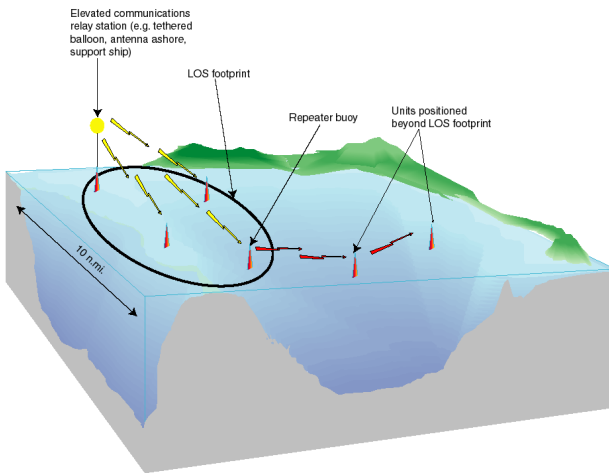
Figure 9 - Example of DUSS WLAN deployment pattern, using an elevated communications relay to extend LOS range and a repeater node to reach nodes beyond LOS (drawing not to scale)

DUSS and radio links for operational concept demonstration

This case presumes the capability to reduce the flow of data transmitted across the transmission channels by an in-buoy processor. At the same time, the physical radio link is required to set up a multi-point wireless LAN. Each buoy operates in a partial-knowledge/partial-connection configuration, exchanging the minimum necessary information with neighbouring buoys in order to identify and track contacts in their progress through the sonobuoy field.

The bi-directional exchange of contact information between neighboring buoys becomes therefore vital for the existence of DUSS as a system. This traffic overlaps to the flow of information packets towards the surveillance command site. DUSS becomes a true wireless network of independent, interacting sonar surveillance units.

Work on smart data reduction is going on, towards the definition of a minimum information size and structure for the identification and integration of contacts without performance loss. The performance and methodologies of data exchange therefore become a key factor for the determination of sonar surveillance performance.

All the typical features of packet switching networks can be inherited by DUSS, as redundancy, fault tolerance, and diversity of communication paths through the network towards the final destination(s) of surveillance information. Techniques, protocols and commercial systems developed for the civilian communication market can be profitably applied to the present application. The limited traffic generated by each unit makes it possible to consider both satellite and classical radio links. Spread-spectrum radios provide the necessary sharing of bandwidth, as well as covert, robust transmission. Techniques and experience derived from the Internet and cellular phone applications are expected to provide valuable contributions both in terms of concept development and of implementation of a demonstrator.

Another concept that is being studied is the adoption of wireless ship-to-ship local area networks in support of Low Frequency Active Sonar (LFAS) in multistatic configurations. The WLAN will enable the near real-time exchange of contact information between the participating platforms, to achieve better accuracy through sonar display correlation. At sea experiments are scheduled for November 2000 as a preparation of the Cerberus experiment, to take place in the second half of 2001.

Is SATCOM a suitable alternative to a spread-spectrum tactical network?

One last word is spent to discuss the role of SATCOM in the fulfilment of the communication requirements discussed so far. In particular, it is interesting to assess whether Low Earth Orbit (LEO) SATCOM systems can efficiently substitute ad-hoc wireless networks deployed on the field. The constraints associated to SATCOM are the narrow bandwidth that is presently made available for data communications and the high associated cost. As an example, providing 24 hours a day connectivity to 10 platforms (e.g. buoys, ships) using the Globalstar system would cost between $12000 and $18000. Those costs could be reduced activating the link on demand, instead of keeping it open 24 hours a day.

Supported data rates are also very limited, in the order of 2.4 kbps. This is acceptable for less demanding applications, such as transmission of positioning information, but is hardly adequate for REA or DUSS operations, unless radical data compression/data reduction schemes are adopted.

Until more performing and cost-effective alternatives are presented, the most practical solution is to deploy tactical wireless networks granting full coverage of the geographical area of interest, using SATCOM gateways to ensure interconnection with land-based infrastructures (e.g. wide area networks).

Conclusions

- Using spread-spectrum, transmit bandwidth can be traded for transmit power (good for battery operation)
- Resistance to multipath, derives from the narrow auto-correlation of the spreading function of Direct Sequence techniques or from the short time slice duration of Frequency Hopping techniques.
- Other features that are not offered by classical systems are low probability of detection and low probability of intercept.

- Spread-spectrum systems play a vital role in all modern telecommunication systems (both military and commercial). They also play a central role in current US DoD research projects on wireless networking in support of the *network-centric warfare* concept, in which operational advantage is achieved from the efficient networking of a geographically dispersed force. This means a focus shift from single autonomous platforms to an integrated network approach.

- The introduction of spread-spectrum communications in Rapid Environmental Assessment and Deployable Underwater Surveillance Systems permits the deployment of distributed and scalable wireless tactical networks of ships and sensors, characterized by reliable performance (survivability, resistance to hostile jamming and environmental interference) and low probability of interception. The high data rates that can be delivered by spread-spectrum systems (up to 10 Mbps) are sufficient to accommodate the most demanding applications that are presently supported.

- The specific requirements of wireless ad-hoc networks include issues such as frequency re-use and multiple access to the same channel, dynamic reconfiguration of network topology, complex variable network structure (including support for moving platforms, such as ships), scalability and quality of service, communications and information security. Further studies will concentrate on this vast and complex subject. Particular emphasis will be put on the issues or resource reservation, quality of service, authentication, and encryption.

- An alternative to spread-spectrum techniques is represented by traditional narrow-band modulation schemes, enhanced by multi-carrier modulation and adaptive equalization techniques. However, implementation of such systems is still at the prototyping phase: additional studies and tests are required before systems become available to end-users.

References

[MBG-99] Mozzone L., Berni A., Guerrini P., Long range, large throughput radio data link for DUSS (Deployable Underwater Surveillance Systems). SACLANTCEN SM-360. La Spezia, Italy, NATO SACLANT Undersea Research Centre, 1999.

[Patt-98] Patterson, W.L., Advanced Refractive Effects Prediction system (AREPS). Version 1.0 User's Manual. TD-3028, Space and Naval Warfare systems Center, San Diego, CA, 1998.

[Sell-99] Sellschopp, J. "Rapid Environmental Assessment", Naval Forces 3/1999.

[Skla-88] Sklar, B., Digital Communications, Fundamentals and Applications, Prentice Hall, 1988 [ISBN 0-13-211939-0]

[TFB-99] Trangeled A., Franchi P., Berni A., Data Communication and Data Fusion Architectures for Rapid Response 96-98, SACLANTCEN SR-296. La Spezia, Italy, NATO SACLANT Undersea Research Centre, 1999.

This page has been deliberately left blank

————————————

Page intentionnellement blanche

Telesonar Signaling and Seaweb Underwater Wireless Networks

J. A. Rice

Space and Naval Warfare Systems Center, San Diego
Acoustics Branch, Code D857
San Diego, CA 92152
United States
rice@spawar.navy.mil

Abstract— **Seawebs '98, '99, and 2000 are experiments incrementally advancing telesonar underwater acoustic signaling and ranging technology for undersea wireless networks. The constraints imposed by acoustic transmission through shallow-water channels have yielded channel-tolerant signaling methods, hybrid multi-user access strategies, novel network topologies, half-duplex handshake protocols, and iterative power-control techniques. Seawebs '98 and '99 respectively included 10 and 15 battery-powered, anchored telesonar nodes organized as non-centralized bi-directional networks. These tests demonstrated the feasibility of battery-powered, wide-area undersea networks linked via radio gateway buoy to the terrestrial internet. Testing involved delivery of remotely sensed data from the sea and remote control from manned command centers ashore and afloat. Seaweb 2000 introduces new telesonar modem hardware and a compact protocol for advanced network development. Sublinks '98, '99, and 2000 are parallel experiments that extend Seaweb networking to include a submerged submarine as a mobile gateway node.**

I. INTRODUCTION

Digital signal processor (DSP) electronics and the application of digital communications theory have substantially advanced the underwater acoustic telemetry state of the art [1]. A milestone was the introduction of a DSP-based modem [2] sold as the Datasonics ATM850 [3,4] and later identified as the first-generation *telesonar* modem. To promote further development, the U.S. invested small business innovative research (SBIR) funding and Navy laboratory support with expectations that energy-efficient, inexpensive telesonar modems would spawn undersea wireless networking methodologies embodied by the *Seaweb* concept [5]. Steady progress resulted in the second-generation telesonar modem [6], marketed as the Datasonics ATM875. Encouraged by the potential demonstrated with the ATM875, the Navy funded the advanced development of a third-generation telesonar modem [7] designated the Benthos ATM885. Seaweb functionality implemented on commercial off-the-shelf (COTS) telesonar hardware shows enormous promise for numerous ocean applications.

Off-board Seaweb nodes of various types may be readily deployed from high-value platforms including submarine, ship and aircraft, or from unmanned undersea vehicles (UUVs) and unmanned aerial vehicles (UAVs).
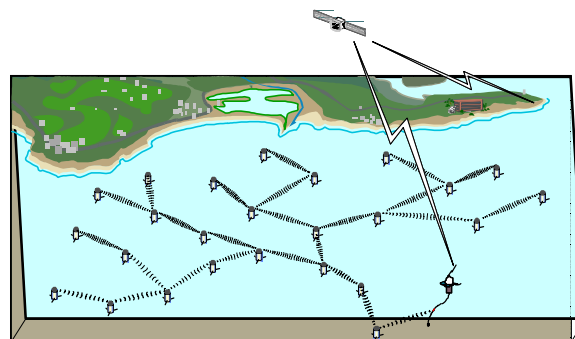
Fig. 1. Seaweb underwater acoustic networking provides digital wireless links enabling deployable autonomous distributed systems (DADS). Gateways to manned control centers include radio links to space or shore and telesonar links to ships.

The architectural flexibility afforded by Seaweb wireless connections permits the mission planner to allocate an arbitrary mix of node types with a node density and area coverage appropriate for the given telesonar propagation conditions and for the mission at hand.

The initial motivation for Seaweb is a requirement for wide-area undersea surveillance in littoral waters by means of a deployable autonomous distributed system (DADS) such as that depicted in Fig. 1. Future sensor nodes in a DADS network generate concise anti-submarine warfare (ASW) contact reports that Seaweb will route to a master node for field-level data fusion [8]. The master node communicates with manned command centers via gateway nodes such as a sea-surface buoy radio-linked with space satellite networks, or a ship's sonar interfaced to an on-board Seaweb server.

The DADS application generally involves operation in 50- to 300-m waters and node spacings of 1 to 5 km. Primary network packets are contact reports with about 1000 information bits [9]. DADS sensor nodes asynchronously produce these packets at a variable rate dependent on the receiver operating characteristic (ROC) for a particular sensor suite and mission. Following ad hoc deployments, DADS relies on the Seaweb network for self-organization including node identification, clock synchronization on the order of 0.1 to 1.0 s, node geo-localization on the order of 100 m, assimilation of new nodes, and self-healing following node failures. Desired network endurance is up to 90 days.

DADS is the natural initial use of Seaweb technology because it forms a fixed cellular network grid of inexpensive interoperable nodes. This architecture is well suited for supporting autonomous oceanographic

sampling network (AOSN) concepts [10] and various autonomous operations, including navigation, control, and telemetry of UUV mobile nodes.

II. CONCEPT OF OPERATIONS

Telesonar wireless acoustic links interconnect distributed undersea assets, potentially integrating them as a unified resource and extending "net-centric" operations into the undersea environment.

Seaweb is the realization of such an undersea wireless network [11] of fixed and mobile nodes, including intelligent master nodes and various interfaces to manned command centers. It provides the command, control, and communications infrastructure for coordinating appropriate assets to accomplish a given mission in an arbitrary ocean environment.

The Seaweb *backbone* is a set of autonomous, stationary nodes (*e.g.*, deployable surveillance sensors, sea mines, relay stations).

Seaweb *peripherals* include mobile nodes (*e.g.*, UUVs, including swimmers and crawlers) and specialized nodes (*e.g.*, bi-static sonar projectors).

Seaweb *gateways* provide connections to command centers submerged, afloat, aloft, and ashore. Telesonar-equipped gateway nodes interface Seaweb to terrestrial, airborne, and space-based networks. For example, a telesonobuoy serves as a radio/acoustic interface permitting satellites and maritime patrol aircraft to access submerged, autonomous systems. Similarly, submarines can access Seaweb with telesonar signaling in the WQC-2 underwater telephone band or by using other organic sonars. Seaweb provides the submarine commander several options for secure, digital connectivity at speed and depth, including bi-directional access to all Seaweb-linked resources and distant gateways.

A Seaweb *server* resides at manned command centers and is an interface to the undersea network as shown in Fig. 2. The server archives all incoming data packets and provides read-only access to client stations via internet. A single designated "super" server controls and reconfigures the network.

Seaweb quality of service is limited by low-bandwidth, half-duplex, high-latency telesonar links. Occasional outages from poor propagation or elevated noise levels can disrupt telesonar links [12]. Ultimately, the available energy supply dictates service life and battery-limited nodes must be energy conserving [13]. Moreover, Seaweb must ensure transmission security by operating with low bit-energy per noise-spectral-density (E_b/N_0) and by otherwise limiting interception by unauthorized receivers. Seaweb must therefore be a revolutionary information system bound by these constraints.

The Seaweb architecture of interest includes the physical layer, the media-access-control (MAC) layer, and the network layer. These most fundamental layers of communications functionality support higher layers that will tend to be application specific.

Simplicity, efficiency, reliability, and security are the governing design principles. Half-duplex handshaking [14] asynchronously establishes adaptive telesonar links [15] as described in Fig. 3. The initiating node transmits a request-to-send (RTS) waveform with a frequency-hopped, spread-spectrum (FHSS) [16] series or direct-sequence spread-spectrum (DSSS) [17] pseudo-random carrier uniquely addressing the intended receiver. (Alternatively, the initiating node may transmit a universal pattern for broadcasting or when establishing links with unknown nodes.) The addressed node detects the request and awakens from a low-power sleep state to demodulate. Further processing of the request signal provides an estimate of the channel scattering function and signal excess. The addressed node then

Fig. 2. Seaweb extends modern "net-centric" interconnectivity to the undersea realm. Wireless underwater networks include gateway nodes with radio, acoustic, wire, or fiber links to manned command centers where a Seaweb server provides the required user interface. Command centers may be aboard ship, submarine, aircraft, or ashore. They may be geographically distant and connected to the gateway node via space satellite or terrestrial internet. At the designated command center a "super" Seaweb server manages and controls the undersea network. All Seaweb servers archive Seaweb packets and provide data access to client stations.

Fig. 3. Seaweb handshake protocol for data transfer involves Node A initiating a request-to-send (RTS) modulated with a channel-tolerant, spread-spectrum pattern uniquely associated with intended receiver Node B. So addressed, Node B awakens and demodulates the fixed-length RTS packet. Node B estimates the channel parameters using the RTS as a probe signal. Node B responds to A with a fixed-length clear-to-send (CTS) that fully specifies the modulation parameters for the data transfer. Node A then sends the data packet(s) with optimal source level, bit-rate, modulation, and coding. If Node B receives corrupted data, it initiates a selective automatic repeat request (ARQ) exchange.

acknowledges receipt with a FHSS or DSSS acoustic reply. This clear-to-send (CTS) reply specifies appropriate modulation parameters for the ensuing message packets based upon the measured channel conditions. Following this RTS/CTS handshake, the initiating node transmits the data packet(s) with nearly optimal bit-rate, modulation, coding, and source level.

At the physical layer, an understanding of the transmission channel is obtained through at-sea measurements and numerical propagation models. Knowledge of the fundamental constraints on telesonar signaling translates into increasingly sophisticated modems. DSP-based modulators and demodulators permit the application of modern digital communications techniques to exploit the unique aspects of the underwater channel. Directional transducers further enhance the performance of these devices [18].

The MAC layer supports secure, low-power, point-to-point connectivity, and the telesonar handshake protocol is uniquely suited to wireless half-duplex networking with slowly propagating channels. Handshaking permits addressing, ranging, channel estimation, adaptive modulation, and power control. The Seaweb philosophy mandates that telesonar links be environmentally adaptive [19], with provision for bi-directional asymmetry.

Spread-spectrum modulation is consistent with the desire for asynchronous multiple-access to the physical channel using code-division multiple-access (CDMA) networking [20]. Nevertheless, the Seaweb concept does not exclude time-division multiple-access (TDMA) or frequency-division multiple-access (FDMA) methods and is in fact pursuing hybrid schemes suited to the physical-layer constraints. In a data transfer, for example, the RTS/CTS exchange might occur as an asynchronous CDMA dialog in which the data packets are queued for transmission during a time slot or within a frequency band such that collisions are avoided altogether.

Optimized network topologies are configured and maintained under the supervision of master nodes [21] as outlined in Fig. 4. Seaweb provides for graceful failure of network nodes, addition of new nodes, and assimilation of mobile nodes. Essential by-products of the telesonar link are range measurement, range-rate measurement, and clock-synchronization. Collectively, these features support initialization, navigation, and network optimization.

III. DEVELOPMENTAL APPROACH

Given the DADS performance requirements, Seaweb research is advancing telesonar modem technology for reliable underwater signaling by addressing the issues of (a) adverse transmission channel; (b) asynchronous networking; (c) battery-energy efficiency; (d) transmission security; and (e) cost.

Despite an architectural philosophy emphasizing simplicity, Seaweb is a complex system and its development is a grand challenge. Given the high cost of sea testing and the need for many prototype nodes, the natural course is to perform extensive engineering system analysis following the ideas of the previous section.

Simulations using an optimized network engineering tool (OPNET) with simplified ocean acoustic propagation assumptions permit laboratory exploration of candidate Seaweb architectures and refinement of networking protocols [22]. Meanwhile, controlled experimentation in actual ocean conditions incrementally advances telesonar signaling technology [23].

Seaweb development applies the results from these research activities with a concentration of resources in prolonged ocean experiments. These annual Seaweb experiments are designed to validate system analysis and purposefully evolve critical technology areas such that the Seaweb state-of-the-art makes an advance toward greater reliability and functionality. The objective of the Seaweb experiments is to implement and test telesonar modems in networked configurations where various modulation and networking algorithms can be exercised, compared, and conclusions drawn. In the long-term, the goal is to provide for a self-configuring network of distributed assets, with network links automatically adapting to the prevailing environment through selection of the optimum transmit parameters.

A full year of hardware improvements and in-air network testing helps to ensure that the incremental developments tested at sea will provide tractable progress and mitigate overall developmental risk. In particular, DADS relies on the annual Seaweb engineering experiments to push telesonar technology for undersea wireless networking. After the annual Seaweb experiment yields a stable level of functionality, the firmware product can be further exercised and

Preparation 10 days	Analyze mission requirements
	Measure or predict environment
	Model transmission channel
	Predict connectivity range limits
	Specify spacing, aperture, and mix
	Pre-program master node only
Installation 1 day	Obey spacing constraints
	Test master node link to gateway
Activation 1 hr	Awaken network nodes
	Handshake with neighbors
Initiation 2 hrs	Obtain reciprocal channel response
	Perform 2-way ranging
	Synchronize clocks
	Measure own depth
	Set power and shape spectrum
Registration 1 hr	Report link & node parameters
	Assimilate data at master node
Optimization 1 hr	Choose primary/alternate routes
	Assign protocols
Operation 90 days	Monitor energy, links, and gateways
	Optimize life, security, and latency
	Assimilate new and mobile nodes

Fig. 4. An automatic process of self-organization follows an ad hoc deployment of telesonar nodes. These methods are tested as OPNET simulations and incrementally implemented in Seaweb experiments.

refinements instituted during DADS system testing and by spin-off applications throughout the year. For example, in year 2000, Seaweb technology was implemented in the May Sublink 2000 and in the April ForeFRONT-2 and June FRONT-2 experiments [24]. These applications afford valuable long-term performance data that are not obtainable during Seaweb's aggressive engineering activities when algorithms are in flux and deployed modems are receiving frequent firmware upgrades.

The Seaweb '98, '99, and 2000 operating area is the readily accessible waters of Buzzards Bay, MA, framed in Fig. 5. An expanse of 5- to 15-meter shallow water is available for large-area network coverage with convenient line-of-sight radio contact to Datasonics and Benthos facilities in western Cape Cod. A shipping channel extending from the Bourne Canal provides periodic episodes of high shipping noise useful for stressing the link signal-to-noise ratio (SNR) margins. The seafloor is patchy with regions of sand, gravel, boulders, and exposed granite.

Seaweb '98, '99, and 2000 modem rigging is illustrated in Fig. 6. Testing is performed during August and September when weather is conducive to regular servicing of deployed network nodes.

A representative sound-speed profile inferred from a conductivity-temperature depth (CTD) probe during Seaweb '98 is plotted in Fig. 7. For observed August and September sound-speed profiles, ray tracing suggests maximum direct-path propagation to ranges less than 1000 m as seen in Fig. 8. Beyond this distance, received acoustic energy is via boundary forward
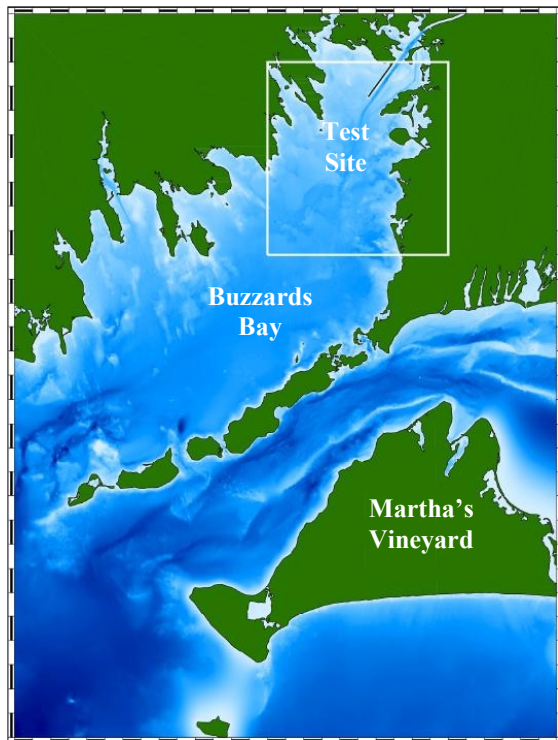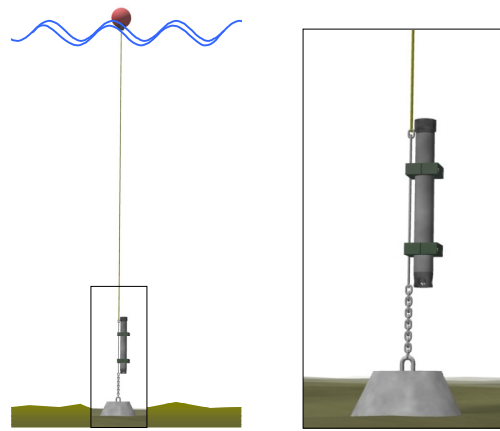
Fig. 6. Seaweb '98, '99, and 2000 modems are deployed in Buzzards Bay with concrete weight, riser line, and surface float. The shallow water and simple rigging permit a small craft to rapidly service the network, including battery replacement and firmware downloads.

scattering. Ray tracing further indicates that received signal energy at significant ranges is attributable to a very small near-horizontal continuum of projector elevation launch angles. Fig. 9 presents predicted impulse responses for 10 ranges each revealing multipath spreads of about 10 ms [25]. All ranges are considered "long" with respect to water depth. Summer afternoon winds and boat traffic regularly roughen the sea surface, increasing scattering loss and elevating noise levels.

IV. TBED '96

A Seaweb predecessor called "telemetry buoy environmental data" (TBED) involved a brute-force networking approach using unmodified COTS Datasonics ATM850 modems. The ATM850 is one of the earliest DSP-based modems, and hence is identified as the "first-generation" telesonar modem. TBED networking used the ATM850's M-ary frequency-shift-keying (MFSK) modulation in a TDMA format wherein all member nodes would sequentially report a compete matrix of data for up to three environmental sensors per node. TBED was significant in that it was the first undersea acoustic digital network with a non-centralized architecture. That is, the network did not involve a central master node with direct links to all slave nodes. Non-centralized architectures are a Seaweb hallmark because of network expandability and area coverage not constrained by point-to-point links. To achieve data forwarding, every TBED transmission was a broadcast including all data from all sensor nodes, thereby permitting receiving nodes to update any stale matrix elements for retransmission in their respective TDMA slots. Thus, the data would reliably but inefficiently be disseminated through the network, ultimately reaching a gateway node. The Navy successfully tested this concept with four nodes in the Gulf of Mexico in 1996. The tested TDMA format could accommodate up to 10 nodes.

Fig. 5. The test site for Seawebs '98, '99 and 2000 is northern Buzzards Bay, MA. Water depth is 5 to 15 m.
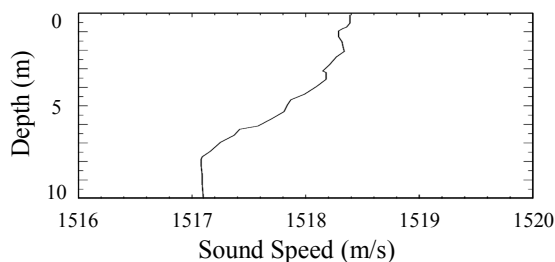
Fig. 7. Sound-speed profiles calculated from conductivity and temperature probes are generally downward refracting during August-September at the Seaweb '98, '99, and 2000 site. This sound-speed profile, 1 of 14 obtained during Seaweb '98, is typical.
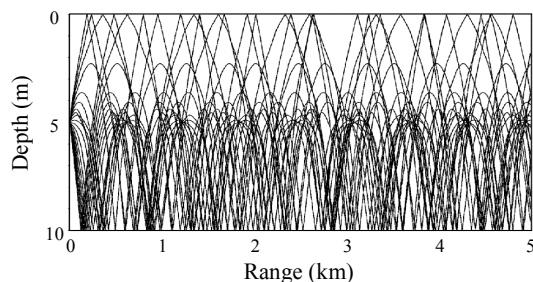
Fig. 8. Seaweb '98 propagation refracts downward in response to vertical sound-speed gradients caused by sea-surface warming. The sound channel is modeled above as rays traced from a ±2.5° fan of elevation angles launched from a transmitter at 5-m depth. A parametric modeling study assessing the dependence of modem depth for this environment confirmed the general rule that long-range signaling in downward-refracting, non-ducted waters is favored by modems placed nearer the seafloor. Hence, Seaweb '98, '99, and 2000 modem transducers are generally about 2 m above the bottom.

V. SEAWEB '98

Seaweb '98 led off a continuing series of annual ocean experiments intended to progressively advance the state of the art for asynchronous, non-centralized networking. Seaweb '98 used the Datasonics ATM875 second-generation telesonar modem [26] recently available as the product of a Navy SBIR Phase-2 contract.

The ATM875 normally uses 5 kHz of acoustic bandwidth with 120 discrete MFSK bins configured to carry 6 Hadamard codewords of 20 tones each. The codewords are interleaved to provide maximum resistance to frequency-selective fading and the Hadamard coding yields a frequency diversity factor of 5 for adverse channels having low or modest spectral coherence. This standard ATM875 modulation naturally supports 3 interleaved FDMA sets of 40 MFSK tonals and 2 codewords each. To further reduce multi-access interference (MAI) between sets, half the available bandwidth capacity provided additional guardbands during Seaweb '98. Thus, only 20 MFSK tonals composing 1 Hadamard codeword were associated with each FDMA set. The Seaweb '98 installation was three geographic clusters of nodes with FDMA sets "A"
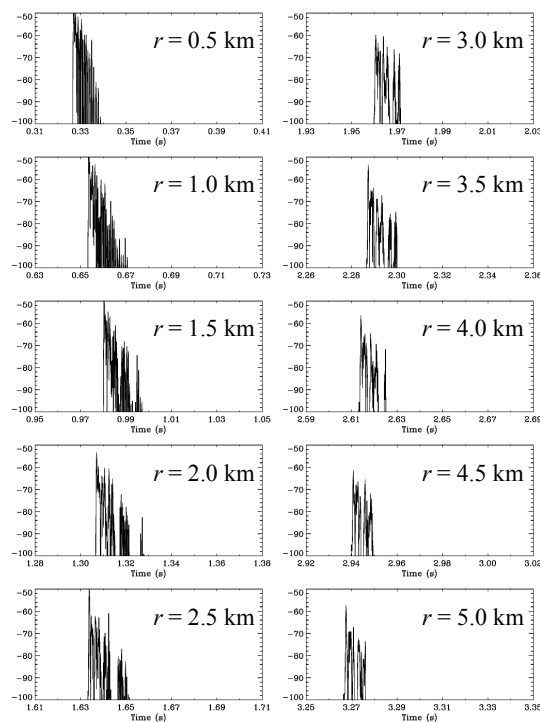
Fig. 9. For a 10-m deep Seaweb '99 channel, a 2-D Gaussian beam model predicts impulse responses for receivers located at 10 ranges, r. Response levels are in decibels referenced to a 0-dB source. Multipath spread is about 10 ms. Note the Seaweb '98, '99, and 2000 working ranges are hundreds of times greater than the water depths and boundary interactions are complex. For rough sea floor and sea surface, the 2-D model approximation must give way to 3-D forward scattering and the predicted response structures will instead be smeared by out-of-plane propagation. Seaweb 2000 testing includes channel probes designed to directly measure channel scattering functions with receptions recorded at various ranges by telesonar testbeds. These channel measurements are used to calibrate an experimental 3-D Gaussian beam model under development for telesonar shallow-water performance prediction and to support analysis of experimental signaling.

through "C" mapped by cluster. For example, all nodes in cluster A were assigned the same FDMA carrier set for reception. Each cluster contained a commercial oceanographic sensor at a leaf node asynchronously introducing data packets into the network. This FDMA architecture was an effective multi-access strategy permitting simultaneous network activity in all three clusters without MAI [27]. A drawback of FDMA signaling is the inefficient use of available bandwidth. Seaweb '98 testing was based on a very conservative 300 bit/s modulation to yield a net FDMA bit-rate of just 50 bit/s. This was an acceptable rate since the Seaweb '98 objectives were to explore networking concepts without excessive attention to signaling issues. Within a cluster, TDMA was the general rule broken only by deliberate intrusion from the command center.

The gateway node is an experimental Navy Racom (radio acoustic comms) buoy pictured in Fig. 10. The "master" node was installed approximately 1500 m from the gateway node. These nodes formed cluster C, meaning both received and demodulated only the FDMA

carriers of set C. The link between gateway and master nodes was extensively exercised during various multi-hour and multi-day periods to gather link statistics and to specifically improve the wake-up and synchronization schemes in the modem acquisition subsystem. Link reliability was monitored at the command center, with performance statistics tallied manually. This point-to-point testing identified specific suspected problems in the fledgling ATM875 implementation, and firmware modifications improved acquisition success from 80% to 97% of packets acquired.

Next, a 3-node subset of cluster A was installed as a relay branch around Scraggy Neck, a peninsula protruding into Buzzards Bay. An Ocean Sensors CTD produced data packets relayed via each of the intervening A nodes to the master node, and then on to the gateway node. Each relay link was about 1500 m in range. Direct addressing of cluster-A nodes from the gateway node confirmed the existence of reliable links to all but the outermost node. Remarkably, a reliable link existed between two nodes separated by 3.6 km in spite of shoaling to 1 to 2 meters in intervening waters! Various network interference situations were intentionally and unintentionally staged and tested until this simple but unprecedented relay geometry was well understood.

An unexpected benefit of the gateway node was discovered during these early tests. The gateway node was accessible from the radio-equipped workboat. Thus, functionality of a newly installed node could be immediately verified. Field personnel would use a deck unit and the gateway node for an end-to-end network circuit test including the new modem as an intermediate node, or they would bidirectionally ring the new modem via just the gateway route. Effectively, the work boat was a mobile node in the network equipped with both telesonar and gateway connections.

At this point, associate engineers from National Oceanic and Atmospheric Agency (NOAA) and Naval Surface Warfare Center (NSWC) visited Seaweb '98.

Fig. 10. In Seawebs '98, '99, and 2000, a Racom buoy provides a very reliable line-of-sight packet-radio link to Seaweb servers at the ashore command center and on the work boat. The radio link is a 900-MHz spread-spectrum technology commercially known as Freewave. In Seawebs '99 and 2000, additional gateway nodes using cellular modems linked via Bell Atlantic and the Internet provide even greater flexibility and provide access by Seaweb servers at various locales across the country.

They were brought by boat far into Buzzards Bay and a hydrophone was deployed over the side with a deck unit programmed to act as such a mobile network node. The visitors were permitted to type messages which were transported through the network and answered by personnel at the ashore command center.

Next, a branch was added to cluster A with a Falmouth Scientific 3-D current meter and CTD. Network contention was studied by having the two cluster-A sensor nodes generate packets at different periods such that network collisions would occur at regular intervals with intervening periods of non-colliding network activity.

Finally, cluster B was introduced to the network with inter-node separations of 2 km. A third device generated data packets. With all available network nodes installed and functioning, the remaining few days involved a combination of gradually arranging network nodes with greater spacing as charted in Fig. 11, and of doing specialized signal testing with the telesonar testbed [28]. In addition, the telesonar testbed was deployed in the center of the network for 5 data-acquisition missions and recorded 26 hours of network activity. The testbed also included a modem, permitting it to act as the tenth network node and giving ashore operators the ability to remotely control and monitor testbed operations. The testbed node provides raw acoustic data for correlation with automatic modem diagnostics, providing opportunity to study failure modes using recorded time series.

Seaweb '98 demonstrated the feasibility of low-cost distributed networks for wide-area coverage. During the three weeks of September testing, the network performed reliably through a variety of weather and noise events. Individual network links spanned horizontal ranges hundreds of water depths in length. The Seaweb '98 network connected widely spaced autonomous modems in a binary-tree topology with a master node at the base and various oceanographic instruments at outlying leaf nodes. Also connected to the master node was an acoustic link to a gateway buoy, providing a line-of-sight digital radio link to the command center ashore. Data packets acquired by the oceanographic instruments were relayed through the network to the master node, on to the gateway node, and thence to the command center. The oceanographic instruments and modems generally operated according to pre-programmed schedules designed to periodically produce network collisions, and personnel at the command center or aboard ship also remotely controlled network nodes in an asynchronous manner.

The most significant result of Seaweb '98 is the consistent high quality of received data obtained from remote autonomous sensors. Data packets were delivered to the command center via up to four acoustic relays and one RF relay. About 2% of the packets contained major bit-errors attributable to intentional collisions at the master node. The quality of data was very high even after the network was geographically expanded. From the gateway node, reliable direct
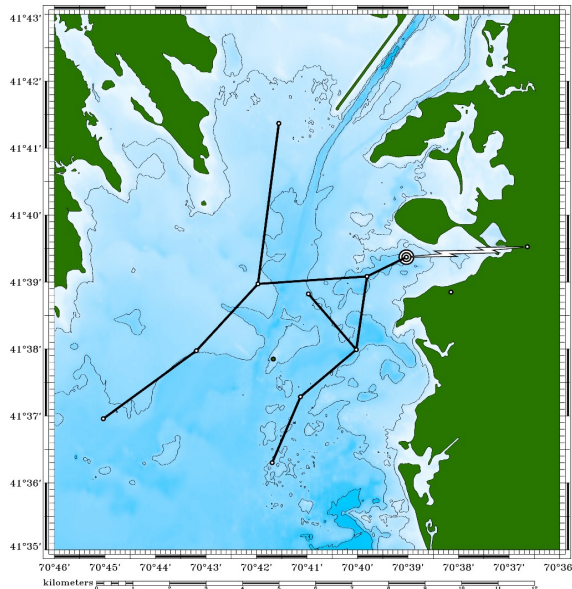
Fig. 11. Seaweb '98 demonstrated store and forward of data packets from remote commercial sensors including a CTD, a current vector meter, and a tilt/heading sensor (at the most northerly, westerly and southerly leaf nodes) via multiple network links to the Racom gateway buoy (large circle). Data packets are then transmitted to the ashore command center via line-of-sight packet radio. An FDMA network with three frequency sets reduced the probability of packet collisions. Following extensive firmware developments supported by this field testing, the depicted topology was exercised during the final days of the experiment. Isobaths are contoured at 5-m intervals.
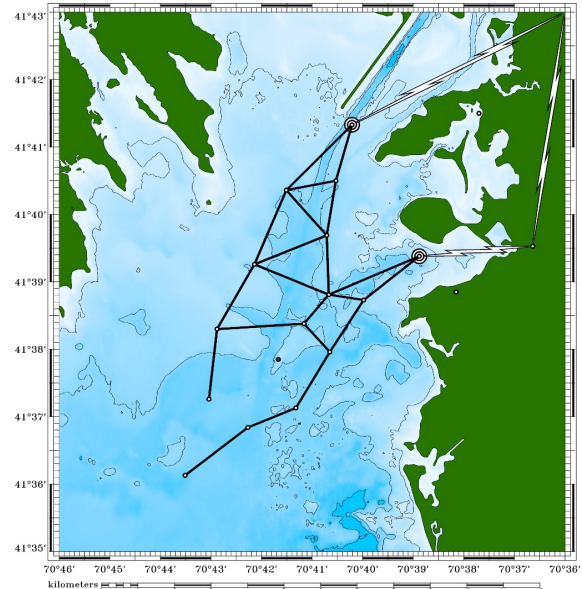
Fig. 12. Seaweb '99 explored the use of handshaking and power control. An ADCP sensor node, a tilt/heading sensor node, and a CTD sensor node generated data packets and the network routed them through various paths. The Racom gateway (easterly large circle) again provided a solid link to shore. A second gateway (northerly large circle) installed on a Coast Guard caisson near the Bourne canal provided a Bell Atlantic cellular modem link to the internet and thence to the command center. The Seaweb server running on a laptop PC managed both gateway connections and archived all network activity.

communications to a node nearly 7 km was achieved, suggesting the network could be expanded considerably more, in spite of the non-ducted 10-m deep channel. Seaweb '98 experience suggests that this environment could have supported 4-km links using the same ATM875 modems and omni-directional transducers. Attesting to the channel-tolerant nature of the MFSK modulation, a 3-km link was maintained during an early phase of testing between two nodes separated by a 1- to 2-m deep rocky shoal. Consistent network degradation occurred during most afternoons and is attributable to summer winds roughening the sea-surface boundary and thus scattering incident acoustic energy. Automated network operations continued during heavy rains and during large ship transits through the field.

Seaweb '98 demonstrated the following network concepts: (a) store and forward of data packets; (b) transmit retries and automatic repeat request; (c) packet routing; and (d) cell-like FDMA node grouping to minimize MAI between cells. In addition, the following DADS concepts were demonstrated: (a) networked sensors; (b) wide-area coverage; (c) acoustic/radio interface (Racom gateway); (d) robustness to shallow-water environment; (e) robustness to shipping noise; (f) low-power node operation with sleep modes; (g) affordability; and (h) remote control. Finally, Seaweb '98 resulted in dramatic improvements to the ATM875 modem and improved its commercial viability for non-networked applications.

Seaweb '98 observations underscore the differences between acoustic networks and conventional networks. Limited power, low bandwidth, and long propagation times dictate that Seaweb networks be simple and efficient. Data compression, forward error correction, and data filtering must be employed at the higher network levels to minimize packet sizes and retransmissions. At the network layer, careful selection of routing is required to minimize transmit energy, latency, and net energy consumption, and to maximize reliability and security. At the physical and MAC layers, adaptive modulation and power control are the keys to maximizing both channel capacity (bits/s) and channel efficiency (bits-km/joule).

VI. SEAWEB '99

Seaweb '99 continued the annual series of telesonar experiments incrementally advancing the state of the art for undersea wireless networks. During a 6-week period, up to 15 telesonar nodes operated in various network configurations in the 5-15 meter waters of Buzzards Bay. Network topologies involving compound multi-link routes were deployed and exercised. All links used a rudimentary form of the telesonar handshake protocol featuring an adaptive power-control technique for achieving sufficient but not excessive SNR at the receiver. Handshaking provided the means for resolving

packet collisions automatically using retries from the transmitter or repeat requests from the receiver.

The multi-access strategy was a new variation of FDMA wherein the six available 20-tone Hadamard words provided 6 separate FDMA sets, A through F. Rather than clustering the FDMA sets as in Seaweb '98, the notion here was to permit the server to optimally assign FDMA receiver frequencies to the various nodes in an attempt to minimize collisions through spatial separation and the corresponding transmission loss. This approach represents an important step toward network self-configuration and prefigures the future incorporation of secure CDMA spread-spectrum codes to be uniquely assigned to member nodes during the initialization process.

Node-to-node ranging was performed using a new implementation of a round-trip-travel time measurement algorithm with 0.1-ms resolution linked to the DSP clock rate. Range estimation simply assuming a constant 1500 m/s sound speed was consistently within 5% of GPS-based measurements for all distances and node pairs.

A very significant development was the introduction of the Seaweb server. It interprets, formats, and routes downlink traffic destined for undersea nodes. On the uplink, it archives information produced by the network, retrieves the information for an operator, and provides database access for client users. The server manages Seaweb gateways and member nodes. It monitors, displays, and logs the network status. The server manages the network routing tables and neighbor tables and ensures network interoperability. Seaweb '99 modem firmware permitted the server to remotely reconfigure routing topologies, a foreshadowing of future self-configuration and dynamic network control. The Seaweb server executes as a graphical set of LabView virtual instruments implemented under Windows NT on a laptop PC. An important function of the server was illustrated when operators bypassed server oversight and inadvertently produced a circular routing where a trio of nodes continuously passed a packet between themselves until battery depletion finally silenced the infinite loop.

In Seaweb '99, the server simultaneously linked with a Bell Atlantic cellular digital packet data (CDPD) gateway node via the internet and with the packet-radio Racom gateway link via a serial port. A milestone was the establishment of a gateway-to-gateway route through the Seaweb server that was exercised automatically over a weekend.

Another test examined networking of automatic uplink sensor packets while simultaneously issuing server-generated downlink commands to deliberately poll sensors. In preparation for the "Front-Resolving Observation Network with Telemetry" (FRONT) application, large acoustic Doppler current profiler (ADCP) packets were synthesized and passed through the network with TDMA scheduling.

For every packet received by a Seaweb '99 node, the modem appended link metrics such as bit-error rate (BER), automatic gain control (AGC), and SNR. These diagnostics aided post mortem system analysis.

Performance correlated strongly with environmental factors such as refraction, bathymetry, wind, and shipping although no attempt was made to quantify these relationships in Seaweb '99.

The ATM875 second-generation telesonar modem again served as the workhorse modem for all network nodes. During the last phase of the experiment, progress was thwarted by memory limitations of the Texas Instruments TMS320C50 DSP. A firmware bug could not be adequately resolved because of lack of available code space for temporary in-line diagnostics. As a result, the final days of the test reverted to a prior stable version of the Seaweb '99 code and the 15-node network charted in Fig. 12 covered a less ambitious area than intended. These limitations plus the desire to begin implementing FHSS and DSSS signaling motivated the initiation of ATM885 third-generation telesonar modem development for Seaweb 2000.

VII. SEAWEB 2000

Seaweb 2000 includes major hardware and firmware advances.

Use of the ATM875 modem during Seawebs '98 and '99 continually thwarted progress in firmware development because of limited memory and processing speed. The ATM885 modem depicted in Fig. 13 overcomes these shortcomings with the incorporation of a more powerful DSP and additional memory. Now, telesonar firmware formerly encoded by necessity as efficient machine language is reprogrammed on the ATM885 as a more structured set of algorithms. The ForeFRONT-1 (Nov. 1999), FRONT-1 (Dec. 1999), ForeFRONT-2 (April 2000), Sublink 2000 (May 2000), and FRONT-2 (June 2000) experiments hastened the successful transition of Seaweb '99 firmware from the ATM875 to the ATM885. These intervening Seaweb applications were stepping stones toward achieving basic ATM885 hardware readiness prior to instituting Seaweb 2000 upgrades.

Seaweb 2000 implements in firmware the core features of a compact, structured protocol. The protocol efficiently maps network-layer and MAC-layer functionality onto a physical layer based on channel-tolerant, 64-bit utility packets and channel-adaptive, arbitrary-length data packets. Seven utility packet types are implemented for Seaweb 2000. These packet types permit data transfers and node-to-node ranging. A richer set of available utility packets is being investigated with OPNET simulations, but the seven core utility packets provide substantial networking capability.

The initial handshake consists of the transmitter sending an RTS packet and the receiver replying with a CTS packet. This round trip establishes the communications link and probes the channel to gauge optimal transmit power. Future enhancements to the protocol will support a choice of data modulation methods, with selection based on channel estimates derived from the RTS role as a probe signal. A "busy" packet is issued in response to an RTS when the receiver

node decides to defer data reception in favor of other traffic. Following a successful RTS/CTS handshake, the data packet(s) are sent. The Seaweb 2000 core protocol provides for acknowledgments, either positive or negative, of a data message. The choice of acknowledgment type will depend on the traffic patterns associated with a particular network mission. Seaweb 2000 begins exploring the factors that will guide this application-specific choice.

A "ping" utility packet initiates node-to-node and node-to-multinode identification and ranging. An "echo" packet is the usual response to a received ping.

In Seaweb 2000, FDMA architectures are superseded by hybrid CDMA/TDMA methods for avoiding mutual interference. FDMA methods sacrifice precious bandwidth and prolong the duration of a transmission, often aggravating MAI rather than resisting it. Furthermore, the use of a small number of frequency sets is viewed as an overly restrictive networking solution. It should be noted that all of these drawbacks are well known and FDMA was employed in Seawebs '98 and '99 primarily for ease of implementation as a simple extension to the rigid ATM875 telesonar machine code. The ATM885 permits a break from those restrictions.

Seaweb 2000 execution fully incorporates the experimental approach tried in Seaweb '99 of establishing two parallel networks—one in air at the command center and one in the waters of Buzzards Bay.
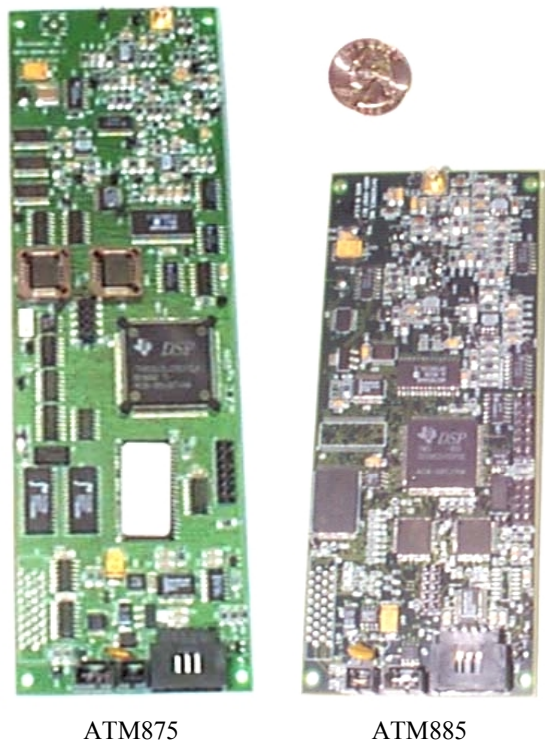
ATM875   ATM885

Fig. 13. The TMS320C5410-based ATM885 telesonar modem overcomes the hardware limitations of the TMS320C50-based ATM875s. The ATM875s served Seaweb '98 and '99. Seaweb 2000 will use ATM885s and networking development will benefit from faster processing, lower power draw, and increased memory. The ATM885 supports 100 MIPs and 320K words of memory compared with 25 MIPs and 74K available from the ATM875.

This approach minimizes time-consuming field upgrades by providing a convenient network for troubleshooting deployed firmware and testing code changes prior to at-sea downloads.

As a further analysis aid, all modems now include a data-logging feature. All output generated by the ATM885 and normally available via direct serial connection is logged to an internal buffer. Thus, the behavior of autonomous nodes can be studied in great detail after recovery from sea. To take maximum advantage of this capability, Seaweb 2000 code includes additional diagnostics related to channel estimation (*e.g.*, SNR, multipath spread, Doppler spread, range rate, etc.), demodulation statistics (*e.g.*, bit-error rate, automatic gain control, intermediate decoding results, power level, etc.), and networking (*e.g.*, data packet source, data packet sink, routing path, etc.). For Seaweb applications, the data-logging feature can also support the archiving of data until such time that an adjacent node is able to download the data. For example, a designated *sink* node operating without access to a gateway node can collect all packets forwarded from the network and telemeter them to a command center when interrogated by a gateway (such as a ship arriving on station for just such a data download).

Increasing the value of diagnostic data, the C5410 real-time clock time is maintained even during sleep state. Although this clock may not have the stability required for certain future network applications, its availability permits initial development of in-water clock-synchronization techniques.

The new ATM885 modem also includes provision for a *watchdog* function hosted aboard a microchip independent of the C5410 DSP. The watchdog resets the C5410 DSP upon detection of supply voltage drops or upon cessation of DSP activity pulses. The watchdog provides a high level of fault tolerance and permits experimental modems to continue functioning in spite of system errors. A watchdog reset triggers the logging of additional diagnostics for thorough troubleshooting after modem recovery.

An aggressive development schedule following Seaweb '99 and preceding Seaweb 2000 matured the Seaweb server as a graphical user interface with improved reliability and functionality consistent with Seaweb 2000 upgrades.

Recent telesonar engineering tests have played host to an applied research effort known as SignalEx [29]. This research uses the telesonar testbeds to record high-fidelity acoustic receptions and measure relative performance for numerous signaling methods. Seaweb 2000 will host SignalEx during the second week of testing. The advantage of coupling SignalEx research with Seaweb engineering is that both activities benefit— SignalEx gains resources and Seaweb gains added empirical test control. By the fifth week, the major Seaweb 2000 engineering developments will reach a level of stability and several experimental network tests will commence. These tests will explore the use of acoustic navigation methods for node localization, cost

functions for optimized network routing, and statistics gathering for network traffic analysis.

Seaweb 2000 doubles as an engineering test for the FRONT-3 experiment to occur in 25-m to 50-m continental shelf waters. In keeping with the developmental approach, FRONT-3 will exercise stabilized Seaweb 2000 technology for an important oceanographic application.

In summary, the specific implementation objectives of Seaweb 2000 are: (a) packet forwarding through network, under control of remotely configurable routing table; (b) 64-bit header; (c) improved software interface between network layer and modem processing; (d) improved wake-up processing, *i.e.*, detection of 2-of-3 or 3-of-4 tones, rather than 3-of-3; (e) improved acquisition signal, *i.e.*, one long chirp, rather than three short chirps; (f) improved channel estimation diagnostics; (g) logging of channel estimates; (h) RTS/CTS handshaking; (i) configurable enabling of RTS/CTS handshake; (j) configuration of power control algorithm; (k) watchdog; (l) automatic-repeat-request (ARQ) feedback; (m) packet time-stamping; and (n) a simple form of adaptive modulation restricted solely to parameter selection for Hadamard MFSK modulation.

The new ATM885 hardware and the Seaweb 2000 protocols are major strides toward the ultimate goal of a self-configuring, wireless network of autonomous undersea devices.

VIII. SUBLINKS '98 AND '99

An associated series of annual tests is exploring submarine participation as a mobile node in Seaweb networks. Sublinks '98 and '99 involved acoustic signaling between the research submarine *USS Dolphin* (AGSS 555), telesonar testbeds, gateway buoys, stationary autonomous bottom nodes, and the *R/V Acoustic Explorer*. The experiments are demonstrating digital, acoustic, underwater signaling to and from a submerged, moving submarine using developmental telesonar technology.

Sublink testing measures communication figures of merit as a function of controlled and measurable environmental parameters for candidate signaling modes. *Dolphin* executes free and controlled tracks around and away from the telesonar testbeds, varying her depth, speed, and telesonar transducer selection.

Aboard *Dolphin*, a COTS underwater telephone (EDO 5400) includes a fully integrated telesonar modem. This electronic system is interfaced to WQC-2 underwater telephone transducers on the sail (EDO SP23LT), keel (EDO SB31CT), and foredeck (EDO SB31CT). The Seaweb server control and monitoring station is located in a forward lab space adjacent to the sonar room with a serial interface to the underwater telephone. The lab space also accommodates a signal analysis station and a multi-channel digital audio tape recording system. This installation permits experimental wireless networked communications with autonomous, off-board nodes as illustrated in Fig. 14. Sublinks '98 and '99 used the

ATM875 modem and Sublink 2000 used the new ATM885. All Sublink acoustic transmissions fall within the 8 to 10.5 kHz band for compatibility with the WQC-2 underwater telephone sonar system response. This "half-band" implementation was readily achieved by compressing the standard 5-kHz telesonar band and prolonging the MFSK signal chips by a factor of 2 to correspondingly tighten the spectral response.

Acoustic Explorer supports telesonar testbed operations and is the afloat command center. *Acoustic Explorer* remains moored south of the testbeds during *Dolphin* dives and monitors test transmissions with an over-the-side transducer and deck modem.

These experiments occurred on the Loma Shelf in the vicinity of 32°36′N, 117°21′W, 10 km west-southwest of Pt. Loma, San Diego, in waters 150- to 250-m deep. The site is conveniently accessed from the port of San Diego and is environmentally well characterized through historical surveys (*e.g.*, Seabeam bathymetry, geoacoustic inversions), prior ocean acoustics testing (*e.g.*, SwellEx), oceanographic measurements (*e.g.*, CTD), geometric measurements (*e.g.*, GPS), and acoustic channel probes. This operating area has a relatively flat bottom, sloping approximately 1° downward to the west, and is consistent with range-dependent assumptions for numerical propagation modeling. Most prescribed test tracks overlie a region with bottom slopes less than 0.5° extending roughly 10 km north and south, with a width of approximately 3 km. When range-dependent geometries are desired, the adjacent Loma Canyon and Coronado Bank offer complex bathymetry.

The submarine sonar is a Seaweb network gateway with the on-board Seaweb server providing an interface for the submarine command center. The server interprets
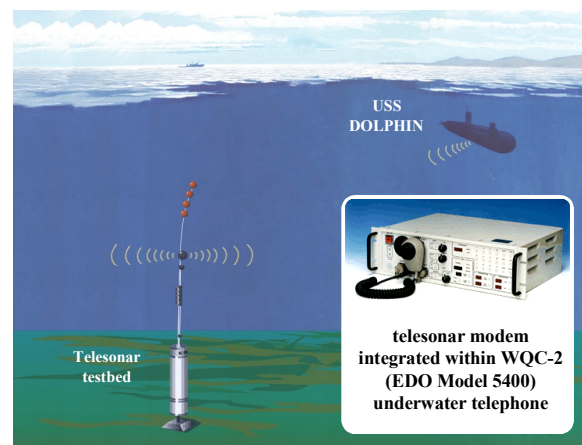
Fig. 14. Sublinks '98, '99, and 2000 demonstrated feasibility of telesonar links between submerged submarines and autonomous offboard devices. Autonomous devices included seafloor nodes such as the telesonar testbeds and surface buoys such as the Racom. Aboard the *USS Dolphin* research submarine, a telesonar modem was integrated with the pictured COTS underwater telephone electronics. These electronics provide connections to standard underwater telephone sonars operating in the 8- to 11-kHz band. The digital modem uses the analog sonar much as a computer modem uses the plain old telephone system. A Seaweb server running on a laptop computer aboard the submarine provides the user interface to the telesonar link.

messages and commands destined for the telesonar network, converts this information into bit strings compatible with telesonar modems, appends necessary headers and routing instructions, and directs the transmissions to a gateway node. The server interprets traffic from gateway nodes, time stamps the messages, logs the traffic, provides a graphical user interface, and maintains a database accessible to client stations.

IX. SUBLINK 2000

Sublink 2000 involved acoustic signaling between *Dolphin*, seafloor-deployed telesonar testbeds, a moored Racom-3 gateway buoy, telesonar listener nodes, and nodes suspended over the side of the moored *Acoustic Explorer* as shown in Fig. 15. Network gateways at the Racom and at *Acoustic Explorer* provided packet radio links to the internet via an ashore radio repeater. Links between all combinations of network nodes were tested while varying several different signaling and channel geometries. Highlights of Sublink 2000 include:

• An ATM885 telesonar modem was successfully integrated with an EDO 5400 underwater telephone and the *Dolphin* WQC-2 sonar. Three single-element WQC-2 transducers were individually tested. The ATM885 "half-band" mode was exercised using the 8-10.5 kHz band. This band is compatible with the WQC-2 sonars and with the telesonar/acomms interoperability standards established jointly by Navy research labs.

• Two autonomous telesonar testbeds were deployed to the seafloor and recovered 12 times using a new acoustically activated release method. High-fidelity transmission, reception, and data acquisition were verified.

• Six *Dolphin* dives were executed, each for approximately 6-8 hours. Telesonar communications from *Dolphin* to testbeds were achieved at maximum test ranges of 10 km. Channel-tolerant MFSK receptions were experienced at testbed nodes in spite of strong downward refraction and absence of ducts causing received signal energy to interact repeatedly with channel boundaries. Testbed-to-*Dolphin* links performed less reliably at the maximum ranges, largely because of lower source levels and higher receiver noise levels both contributing to a relatively lower SNR.

• Feasibility of *Dolphin* as the source of synthetic ASW contact reports was demonstrated, thus giving the green light for the planned Seaweb 2001 experiment in support of littoral ASW future naval capabilities.

• A complicated transmission cycle involving three transmitter platforms and multiple experiments was executed flawlessly in an interleaved TDMA schedule. Preprogrammed testbed missions were performed autonomously for the full duration of each experimental event.

• SignalEx transmissions were performed every three minutes using composite probe and eight different waveform suites, including contributed signals from Naval Undersea Warfare Center (NUWC), Polytechnic
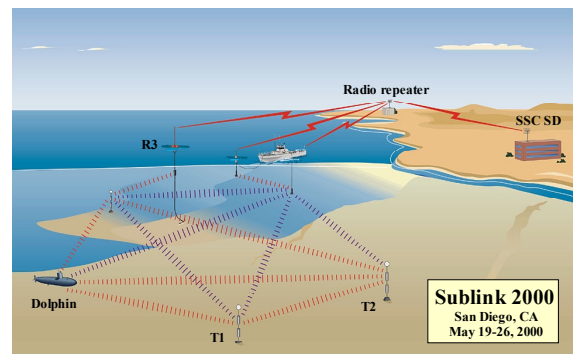
Fig. 15. Sublink 2000 was a TDMA network in 200-m waters with submarine connectivity to two telesonar testbeds (nodes T1 and T2) at ranges up to 10 km and with a Racom gateway buoy (node R3) at ranges up to 4 km. As a demonstration, standard email messages were generated by *USS Dolphin* at speed and depth and delivered to ashore users via internet.

University, Northeastern University, Woods Hole Oceanographic Institution, Science Applications International Corporation (SAIC), Benthos, and Delphi Communication Systems (DCS) SignalEx transmissions used the 8-11 and 8-16 kHz bands.

• A new ATM885 "autobaud" implementation was exercised by means of a new ATT9 command causing the modem to sequentially transmit 14 64-bit Seaweb utility packets, each with a different modulation, coding, bit-rate or source level format. The receiving modem automatically and repeatedly processes all 14 modes.

• SignalEx and ATT9 data sets were collected between two testbeds on seafloor with separations of 3, 5 and 7 km. SignalEx and ATT9 data sets were collected between testbeds and submarine at 200-400 ft depths and 3-5 kt speeds over range-independent (200-m depth SWellEx benchmark channel) and range-dependent bottoms (over Loma Canyon and Coronado Bank).

• A Racom buoy with telesonar modem and Freewave radio provided gateway links to shore and ship. Shipboard personnel monitored all telesonar transmissions using an over-the-side telesonar modem and the Racom buoy 5 km distant.

• Seaweb servers aboard *Dolphin* and *Acoustic Explorer* controlled remote modems and archived incoming packets.

• Emails from transiting, submerged *Dolphin* were delivered via telesonar, Racom gateway buoy, and Seaweb server to the Office of Naval Research (ONR), Submarine Development Squadron Five, and to the family of a young sailor.

• GPS navigation data were recorded on *Acoustic Explorer* and inertial navigation data were recorded on *Dolphin*. All testbed deployment stations were well documented and well within specified placement tolerances.

• Excellent environmental (CTD) measurements revealed a prevalent thin layer of warm surface water overlaying a downward refracting sound-speed gradient.

• At-sea channel modeling and propagation prediction were performed using numerical physics-based telesonar

propagation models. Analysis of SignalEx channel probes confirmed predicted channel impulse responses and validated numerical channel models.

• The Racom-3 buoy, presumed lost after failure of a mooring line, was successfully recovered following a night-time search-and-rescue effort relying solely on experimental telesonar ranging technology.

• Several new Seaweb 2000 network functions were implemented on the ATM885 modem and exercised as incremental developments prior to the June FRONT-2 and August Seaweb 2000 experiments.

• ATM885 diagnostics were automatically logged, including SNR, AGC, and the number of corrected and uncorrected errors. Modems were intentionally driven to failure by systematic reduction in source level. The large cache of telesonar performance data with appropriate ground-truth measurements will support detailed comparative studies and parametric analyses.

X. FUTURE WORK

In 2001, stabilized Seaweb 2000 functionality will support the FRONT application (ForeFRONT-3 and FRONT-3 experiments) and the DADS application (Fleet Battle Experiment "India").

In late summer, Seaweb 2001 is scheduled to occur in a very large expanse of 30- to 300-m waters adjacent to San Diego, CA, and will incorporate several new fixed and mobile undersea systems as network nodes.

The annual Seaweb and Sublink experiments will continue to extend area coverage, resource optimization, network capacity, functionality, and quality of service. Active research feeding new technologies into Seaweb includes spread-spectrum signaling, directional transducers [30], in situ channel estimation, adaptive modulation, ad hoc network initialization, and node ranging and localization.

XI. CONCLUSION

Undersea, off-board, autonomous systems will enhance the war-fighting effectiveness of submarines, maritime patrol aircraft, amphibious forces, battle groups, and space satellites. Wide-area sensor grids, leave-behind multi-static sonar sources, mine-hunting robots, swimmer-delivery systems, and autonomous vehicles are just a few of the battery-powered, deployable devices that will augment high-value space and naval platforms. Distributed system architectures offer maximum flexibility for addressing a wide array of ocean environments and military missions.

Telesonar is an emerging technology for wireless digital communications in the undersea environment. Telesonar transmission channels include shallow-water environments with node-to-node separations hundreds of times greater than the water depth. Robust, environmentally adaptive acoustic links interconnect undersea assets, integrating them as a unified resource.

Seaweb offers a blueprint for telesonar network infrastructure. Warfare considerations stipulate the network architecture support rapid installation, wide-area coverage, long standoff range, invulnerability, and cross-mission interoperability. Seaweb is an information system compatible with low bandwidth, high latency, and variable quality of service. Seaweb connectivity emphasizes reliability, flexibility, affordability, energy efficiency, and transmission security. Network interfaces to manned command centers via gateways such as those demonstrated by Sublink and Racom are an essential aspect of the Seaweb concept. Command, control, and communications via Seaweb supports common situational awareness and collective adaptation to evolving rules of engagement. Seaweb revolutionizes naval warfare by ultimately extending network-centric operations into the undersea battlespace.

DCS and Datasonics involvement in Seaweb '98 was sponsored by an SBIR topic N93-170 Phase-2 contract administered by Phil DePauk and Linda Whittington of SPAWAR. DCS and Datasonics involvement in Seawebs '99 and 2000 was sponsored by an SBIR topic N97-106 Phase-2 contract administered by Al Benson, Tom Curtin, Don Davison, Dave Johnson, and Doug Harry of ONR. In addition, Datasonics invested independent research and development (IR&D) resources in support of Seawebs '98 and '99.

The officers and crew of *USS Dolphin* and *R/V Acoustic Explorer* expertly served the needs of the Sublink experiments, ensuring safe and successful testing under demanding conditions.

Dan Codiga, Philip Bogden, Dennis Arbige, and Adam Houk of University of Connecticut Department of Marine Sciences contributed to Seawebs '99 and 2000 and implemented the cellular-modem gateway and the ADCP sensor node. John Newton (Polar Associates, Inc.) performed environmental analysis in support of Seaweb '98 and Sublink '98. Daryl Huggins (Edo Corp.) participated in Sublinks '98, '99, and 2000. Kent Raysin (SAIC) participated in Seaweb '99. Jerry Schwell (NUWC) participated in Sublink '98. Kim McCoy (Ocean Sensors, Inc.) participated in Sublink '99. Bob Peloquin (NUWC), Mike Porter (SAIC), Sherman Havens (Predicate Logic, Inc.), and Mark Hogue (SAIC) participated in Sublink 2000.

Mike Cox and Tim McCombs of NSWC Coastal Systems Station performed the TBED '96 experiment.

REFERENCES

[1] D. B. Kilfoyle A. B. Baggeroer, "The State of the Art in Underwater Acoustic Telemetry," *IEEE J. Oceanic Eng.*, Vol. 25, No. 1, pp. 4-27, Jan. 2000

[2] J. A. Catipovic, M. Deffenbaugh, L. Freitag, and D. Frye, "An Acoustic Telemetry System for Deep Ocean Mooring Data Acquisition and Control," *Proc. IEEE Oceans '89 Conf.*, Sept. 1989

[3] S. Merriam and D. Porta, "DSP-Based Acoustic Telemetry Modems," *Sea Technology*, May 1993

[4] D. Porta, "DSP-Based Acoustic Data Telemetry," *Sea Technology*, Feb. 1996

[5] J. A. Rice and K. E. Rogers, "Directions in Littoral Undersea Wireless Telemetry," *Proc. TTCP Symposium on Shallow-Water Undersea Warfare*, Halifax, Nova Scotia, Canada, Vol. 1, pp. 161-172, Oct. 1996

[6] K. F. Scussel, J. A. Rice, and S. Merriam, "New MFSK Acoustic Modem for Operation in Adverse Underwater Acoustic Channels," *Proc. IEEE Oceans '97 Conf.*, Halifax, Nova Scotia, Canada, pp. 247-254, Oct. 1997

[7] M. D. Green, "New Innovations in Underwater Acoustic Communications," *Proc. Oceanology International*, Brighton, U.K., March 2000

[8] E. Jahn, M. Hatch, and J. Kaina, "Fusion of Multi-Sensor Information from an Autonomous Undersea Distributed Field of Sensors," *Proc. Fusion '99 Conf.*, Sunnyvale, CA, July 1999

[9] S. McGirr, K. Raysin, C. Ivancic, and C. Alspaugh, "Simulation of underwater sensor networks," *Proc. IEEE Oceans '99 Conf.*, Seattle WA, Sept. 1999

[10] T. B. Curtin, J. G. Bellingham, J. Catipovic, and D. Webb, "Autonomous Oceanographic Sampling Networks," *Oceanography*, Vol. 6, pp. 86-94, 1993

[11] E. M. Sozer, M. Stojanovic, and J. G. Proakis, "Underwater Acoustic Networks," *IEEE J. Oceanic Eng.*, vol. 25, no. 1, pp. 72-83, Jan. 2000

[12] J. A. Rice, "Acoustic Signal Dispersion and Distortion by Shallow Undersea Transmission Channels," *Proc. NATO SACLANT Undersea Research Centre Conf. on High-Freq. Acoustics in Shallow Water*, Lerici, Italy, pp. 435-442, July 1997

[13] J. A. Rice and R. C. Shockley, "Battery-Energy Estimates for Telesonar Modems in a Notional Undersea Network," *Proc. MTS Ocean Community Conf.*, Vol. 2, pp. 1007-1015, Baltimore MD, Nov. 1998

[14] P. Karn, "MACA—A New Channel Access Method for Packet Radio," *Proc. ARRL/CRRL Amateur Radio 9th Computer Network Conf.*, Sept. 1990

[15] J. A. Rice and M. D. Green, "Adaptive Modulation for Undersea Acoustic Modems," *Proc. MTS Ocean Community Conf.*, Vol. 2, pp. 850-855, Baltimore MD, Nov. 1998

[16] M. D. Green and J. A. Rice, "Channel-Tolerant FH-MFSK Acoustic Signaling for Undersea Communications and Networks," *IEEE J. Oceanic Eng.*, Vol. 25, No. 1, pp. 28-39, Jan. 2000

[17] E. M. Sozer, J. G. Proakis, M. Stojanovic, J. A. Rice, R. A. Benson, and M. Hatch, "Direct-Sequence Spread-Spectrum–Based Modem for Underwater Acoustic Communication and Channel Measurements," *Proc. IEEE Oceans '99 Conf.*, Seattle WA, Sept. 1999

[18] N. Fruehauf and J. A. Rice, "System Design Aspects of a Steerable Directional Acoustic Communications Transducer for Autonomous Undersea Systems, *Proc. Oceans 2000 Conf.*, Providence RI, Sept. 2000

[19] J. A. Rice, V. K. McDonald, M. D. Green, and D. Porta, "Adaptive Modulation for Undersea Acoustic Telemetry," *Sea Technology*, Vol. 40, No. 5, pp. 29-36, May 1999

[20] M. Stojanovic, J. G. Proakis, J. A. Rice, and M. D. Green, "Spread-Spectrum Methods for Underwater Acoustic Communications," *Proc. IEEE Oceans '98 Conf.*, Vol. 2, pp. 650-654, Nice France, Sept. 1998

[21] J. G. Proakis, M. Stojanovic, and J. A. Rice, "Design of a Communication Network for Shallow-Water Acoustic Modems," *Proc. MTS Ocean Community Conf.*, Vol. 2, pp. 1150-1159, Baltimore MD, Nov. 1998

[22] K. Raysin, J.A. Rice, E. Dorman, and S. Matheny, "Telesonar Network Modeling and Simulation," *Proc. IEEE Oceans '99 Conf.*, Sept. 1999

[23] V. K. McDonald, J. A. Rice, M. B. Porter, and P. A. Baxley, "Performance Measurements of a Diverse Collection of Undersea Acoustic Communication Signals," *Proc. IEEE Oceans '99 Conf.*, Seattle WA, Sept. 1999

[24] D. L. Codiga, J. A. Rice, and P. S. Bogden, "Real-Time Delivery of Subsurface Coastal Circulation Measurements from Distributed Instruments using Networked Acoustic Modems," *Proc. IEEE Oceans 2000 Conf.*, Providence RI, Sept. 2000

[25] P. A. Baxley, H. P. Bucker, and J. A. Rice, ""Shallow-Water Acoustic Communications Channel Modeling Using Three-Dimensional Gaussian Beams," *Proc. MTS Ocean Community Conf.*, Vol. 2, pp. 1022-1026, Baltimore MD, Nov. 1998

[26] M. D. Green, J. A. Rice, and S. Merriam, "Underwater Acoustic Modem Configured for Use in a Local Area Network," *Proc. IEEE Oceans '98 Conf.*, Vol. 2, pp. 634-638, Nice, France, Sept. 1998

[27] M. D. Green, J. A. Rice, and S. Merriam, "Implementing an Undersea Wireless Network Using COTS Acoustic Modems," *Proc. MTS Ocean Community Conf.*, Vol. 2, pp. 1027-1031, Baltimore MD, Nov. 1998

[28] V. K. McDonald and J. A. Rice, ""Telesonar Testbed Advances in Undersea Wireless Communications," *Sea Technology*, Vol. 40, No. 2, pp. 17-23, Feb. 1999

[29] M. B. Porter, V. K. McDonald, J. A. Rice, and P. A. Baxley, "Relating the Channel to Acoustic Modem Performance," *Proc. European Conf. Underwater Acoustics*, Lyons, France, July 2000

[30] A. L. Butler, J. L. Butler, W. L. Dalton, and J. A. Rice, "Multimode Directional Telesonar Transducer," *Proc. IEEE Oceans 2000 Conf.*, Providence RI, Sept. 2000

This page has been deliberately left blank

_____

Page intentionnellement blanche

THE TURKISH NARROW BAND VOICE CODING AND NOISE PRE-PROCESSING NATO CANDIDATE

Ahmet Kondoz   Hasan Palaz*

TÜBİTAK-UEKAE National Research Institute of Electronics & Cryptology
P.O. Box 21, 41470, Gebze, KOCAELI, TURKEY.
*E_mail : palaz@mam.gov.tr

ABSTRACT

Robust and low power communication systems are essential for battle field environment in military communication which require bit rates below 4.8kb/s. In order to benefit from the new advances in speech coding technologies and hence upgrade its communication systems, the NATO has been planning to select a speech coding algorithm with its noise pre-processor. In this paper we describe a speech coder which is capable of operating at both 2.4 and 1.2kb/s, and produce good quality synthesised speech. This coder will form the basis of the Turkish candidate which is one of the three competing. The rate of the coder can be switched from 2.4kb/s to 1.2kb/s by increasing the frame length for parameter quantisation from 20ms to 60ms. Both rates use the same analysis and synthesis building blocks over 20ms. Reliable pitch estimation and very elaborate voiced/unvoiced mixture determination algorithms render the coder robust to background noise. However in order to communicate in very severe noisy conditions a noise pre-processor has been integrated within the speech encoder.

1. INTRODUCTION

Speech coding at low bit rates has been a subject for intense research over the last 2 decades and as a result many speech coding algorithms have been standardised with bit rates ranging from 16kb/s down to 2.4kb/s. The standards covering the bit rates down to around 5kb/s are based mainly on CELP derivatives and the standards below 5kb/s are based mainly on frequency domain vocoding (harmonic coding) models such as sinusoidal coding [1]. Although in principle a harmonic coder should produce toll quality speech at around 4kb/s and good communications quality at around 2.4kb/s and below, various versions may have significantly different output speech quality. This quality difference comes from the way the parameters such as pitch and voicing are estimated/extracted at the analysis and the way parameters are interpolated for smooth evolution of the output speech during the synthesis process. A further difference is the parameter update rates and quantisation methods used. In this paper we focus on the split-band LPC (SB-LPC) approach to achieve a mode switchable 2.4-1.2kb/s coding rates with high intelligibility and good quality output speech, even during high background and channel noise conditions. Both versions of the algorithm work on 20ms analysis blocks and use the same analysis/synthesis procedures where a novel pitch detection algorithm and an elaborate voicing mixture determination are

used which are essential for good speech quality. Although this algorithm performs well in background noise conditions, if the noise is too high (SNR<10dB) the use of a noise pre-processor (NPP) helps to improve the speech intelligibility as well as enabling perceptually more comfortable speech quality. We have therefore incorporated a NPP in the encoder.

In the following we present the description of the speech analysis/encoding, parameter quantisation followed by decoding/speech synthesis building blocks. This is then followed by the description of the NPP, and finally test results and the conclusions of the paper are presented.

2. SPEECH ANALYSIS

The Split-Band LPC Vocoder has been presented in detail in [2]. In this new version we have used a novel pitch estimation and a multiple input time/frequency domain voicing mixture classification algorithms. Residual spectral magnitudes are extracted by selecting the harmonic peaks for the voiced part of the spectrum and computing the average noise energy in each fundamental frequency band for the unvoiced part. During the extraction of the residual spectral magnitudes we are only interested in the relative variations of magnitudes and not their absolute values. A separate energy control factor is computed from the input speech for proper scaling of the signal at the output of the synthesiser. Speech analysis and synthesis are based on 20ms frames but parameters are quantised every 20ms for 2.4kb/s and every 60ms for 1.2kb/s versions respectively.

2.1 PITCH ESTIMATION ALGORITHM

The pitch estimation algorithm consists of three parts. First a frequency domain analysis is performed. The most promising candidates from this first search are then checked by computing a time domain metric for each. Finally one of the remaining candidates is selected based on the frequency and time domain metrics, as well as the tracking parameters.

Frequency domain pitch analysis is performed using a modified version of the algorithm described by McAulay [4] which determines the pitch period to half sample accuracy. The speech is windowed using a 241 point Kaiser window ($\beta = 6.0$), then a 512 point FFT is performed to obtain the speech spectrum. The fundamental frequency is the one that produces the best periodic fit to the smoothed spectrum. In order to reduce complexity, only the lower 1.5 kHz of this spectrum is used for the pitch

algorithm. To further reduce complexity, only integer pitch values are used above the pitch value of 45 samples.

However, this initial pitch estimate is not always correct. In particular doubling and halving of the pitch frequency can occur. In order to avoid these problems, a certain number of candidate pitch values are selected for further processing. In addition, the range of possible values for ω_0 is divided into 5, corresponding in pitch lags of: [15-27],[27.5-49.5],[50-94.5],[95-124.5] and [124.5-150]. In each of these intervals, the best candidate is also selected, if it is not already selected in the first stage. These intervals are selected so that no pitch candidate can double in a given interval.

All candidate pitch periods determined above are re-examined using a metric which measure the RMS energy variations with respect to the energy computation block length which takes the values given by the candidate pitch periods. The RMS energy fluctuation is minimum when RMS computation block length equals the correct pitch period or its integer multiples.

After the elimination of some candidates based on the time domain metric, if more than one pitch candidates are left, the final decision process operates as follows: For each candidate a final metric is computed, which takes into account both the time- and frequency- domain measures: The candidate with the best combined final metric is then selected as a pitch estimate. In order to avoid pitch doubling, a sub-multiple search is performed. If there is a remaining candidate close enough to being a sub-multiple of the current pitch estimate, and whose final metric is above a certain threshold (typically 0.8 times the final metric of the current pitch estimate), then it is selected as the new current pitch estimate. The sub-multiple search is then repeated using this new value.

The pitch algorithm described above is usually reliable in clean speech conditions. However, it occasionally suffers from pitch doubling and halving when the pitch is not clearly defined, or in heavy background noise conditions. To overcome this problem we have used a mild pitch tracking. In order to be able to update the tracked pitch parameters during speech only frames a simple voice activity detector which is explained in section 5 is used. After the computation of the time and frequency domain metrics, before the start of the elimination process, each candidate which is close to the tracked pitch has its metrics biased to increase its chances of being selected as the final pitch.

The VAD also determines the signal to background noise ratio of the input samples which controls the amount of tracking used. The bias applied by tracked pitch on the metrics is more for noisy speech than in clean speech conditions.

In clean speech conditions this pitch estimation algorithm exhibits very few errors. They only occur when the pitch is not clearly defined and only extra look-ahead could improve this. It is also very resilient to background noise, and still operates satisfactorily down to SNR of 5 dB. At higher noise levels errors start to occur occasionally but the algorithm still manages to give the correct pitch value most of the time.

2.2 LP EXCITATION VOICING MIXTURE

Many low bit rate vocoders now use the assumption that the voicing content of the speech can be represented by only one cut-off frequency below which the speech is considered harmonic and above which it is considered stochastic. This has the advantage of requiring only a very small number of bits to quantise the voicing information, as opposed to transmitting one bit per harmonic band. If performed accurately, the distortion induced by this assumption will be very limited and acceptable for low bit rate speech coders. It is however very important to correctly determine the cut-off frequency as errors will induce large distortions in the output speech quality.

In SB-LPC, for accurate voicing extraction the speech is first windowed using a variable length Kaiser window. Four different windows are used, from 121 to 201 samples in length, depending on the current pitch period, so as to have the smallest possible window covering at least 2 pitch cycles. In the next step the limits of each harmonic band across the spectrum is determined. This is done by refining the original pitch estimate down to a more accurate fractional pitch. The original pitch accuracy is at half a sample accuracy up to the pitch value of 45 samples and integer for bigger values. Moreover the pitch has been determined using only the lower 1.5 kHz of the spectrum. The spacing of the harmonics might be slightly different in the higher part of the spectrum. Hence it is necessary to refine the pitch using the whole of the 4 kHz spectrum.

A threshold value is then computed for each band across the spectrum, based on various time- and frequency domain factors. The general idea being that if the voicing value is above the threshold value for a given band, then it is probably voiced. Finally for each possible quantised cut-off frequency, a matching measure is computed using the threshold and voicing measures for each band, and the final quantised cut-off frequency is selected as the one which maximises this matching.

If a harmonic band is voiced, then its content will have a shape similar to the spectral shape of the window used to window the original speech prior to the Fourier transform, whereas unvoiced bands will be random in nature. Hence voicing can be determined by measuring the level of normalised correlation between the content of the harmonic band and the spectral shape of the window. The normalized correlation lies between 0.0 and 1.0, where 0.0 and 1.0 indicates unvoiced and voiced extremes respectively.

For the decision making this normalized correlation is compared against a fixed threshold for each band across the spectrum. Since the likelihood of voiced and unvoiced is not fixed across the frequency spectrum, and may also vary from one frame to the next, the decision threshold value needs to be adaptive for accurate voicing determination. When determining a voicing threshold value for each frequency band (harmonic) we have used additional factors some of which are listed in [3]. A threshold value is computed for each band based on the following variables:

- the peakiness (ratio of the L1 to L2 norms),

- the cross-correlation value at the pitch delay,

- the ratio of the energy of the high frequencies to energy of the low frequencies in the spectrum

- the ratio between the energies of the speech and of the LP residual

- the ratio between the energy of the frame and the tracked maximum energy of the speech, E_s/E_{max}.

- the voicing of the previous frame

- a bias is added to tilt the threshold toward more voiced in the low frequencies.

Having computed a voicing measure and a threshold for each harmonic band we now need to find the best quantised cut-off frequency for this set of parameters. For each possible quantiser value a matching measure is computed taking into account the difference between the correlation value and the corresponding threshold, as well as the energy in a given harmonic band. A bias which favors voiced decisions over unvoiced decisions is also used. A typical quantiser for the voicing is a 3 bits quantiser, representing 8 cut-off frequencies spaced between 0 and 4 kHz.

3. PARAMETER QUANTISATION

Table 1. shows the bit allocation for the 2.4 and 1.2kb/s versions.

Bit Rate	2.4 kb/s	1.2 kb/s		
Update rate (in ms)	20	60		
LPC	21	44		
Pitch	7	3	6	3
Voicing	3	3		
RMS energy	6+1	6+6		
Spectral Magnitudes	9	0	0	0
Sync. bit	1	1		
Total	48	72		

Table 1: Bit allocation for the different rates of the Split-Band LPC Vocoder

In the case of 2.4kb/s 47 bits are used to quantised the parameters every 20ms. The LP parameters are quantised in the form of line spectral frequencies (LSF) with a multi-stage vector quantisation (MSVQ) which has three stages of 7,7,7 bits. However, before the MSVQ, a first order moving average (MA) prediction with 0.5 predictor is applied to remove some of the correlation in the adjacent LP parameter sets. The RMS frame energy is quantised with a 6-bit scalar quantiser after a similar MA prediction with 0.7 predictor plus one bit protection. Only the 64 levels out of the 128 (6+1 bits) are used for encoding by ensuring that in case of channel errors, the codewords that could potentially result in large gain changes are not used. This process ensures that the errors introduced will have minimum damaging effect. The pitch is quantised non-uniformly with 7-bits, covering the range from 16 to 150 samples. Since the residual spectral

magnitudes under the formant regions are more important, during magnitude quantisation the most important 7 magnitudes followed by the average value of the rest is vector quantised using a 9-bit codebook.

In the case of 1.2kb/s, a frame of 60ms is used where it is split into three 20ms sub-frames. The LP parameters are multi stage vector quantised using 44bits after a similar MA prediction process. For the pitch, voicing and energy computations, 20ms sub-frame length is used and repeated 3 times per frame. Pitch of the first and third sub-frames are quantised with respect to the pitch of the middle sub-frame using 3-bits each. The middle sub-frame's pitch is quantised using 6-bits. The voicing mixtures of all three sub-frames are jointly quantised using 3-bits. Similarly the RMS energies are jointly quantised with a gain shape vector quantiser using 6 bits for the gain and 6 bits for the three element shape vector.

4. DECODING AND SPEECH SYNTHESIS

4.1 Parameter Decoding

In the 2.4kb/s mode, each 20ms frame has its own LP parameters, pitch, voicing mixture and the RMS frame energy which are sufficient for good quality speech synthesis. During the decoding process of LSFs the usual stability checks are applied. When decoding the RMS energy, channel error effects are minimised by using only 64 possible combinations of the 7 bits representation with proper robust index assignment [5]. For the pitch and voicing no channel error checks are applied.

In the case of 1.2kb/s no error checks are applied to any of the parameters, except the usual LSF stability check and robust index assignment [5].

4.2 Speech Synthesis

In order to improve the speech quality, at the decoder we introduce half a frame delay for both 2.4 and 1.2kb/s versions. In the case of 2.4kb/s first half of 20ms frame is synthesised by interpolating the current parameters with the preceding set and the second half uses the parameters interpolated between the current and the next sets. Similar interpolation is applied for the 1.2kb/s version where each 20ms sub-frame is assumed to be a 20ms frame. The actual interpolation is applied pitch synchronously and the contribution of the left and right hand side parameters is based on the centre position of each pitch cycle within the synthesis frame. The actual synthesis of both voiced and unvoiced sounds is performed using an IDFT with pitch period size. The voiced part of the spectrum has only the magnitudes with zero phases and the unvoiced part of the spectrum is filled with both unvoiced magnitudes and random phases. If desired a perceptual enhancement process is applied where the valley regions of the excitation spectrum are suppressed [2]. The resultant excitation is then passed through the LP synthesis filter which has its parameters interpolated pitch synchronously. Finally the output signal which may have arbitrary energy is normalised per pitch cycle to match the interpolated frame energy.

5. NOISE PRE-PROCESSOR

The SB-LPC speech coder with the above detailed parameter analysis and quantisation techniques operate well within background noise environments. However, both speech quality and intelligibility in heavy noise conditions can be improved if a suitable noise suppression/pre-processing technique (NPP) is used before speech analysis is applied. We have used a noise pre-processing technique to suppress the background noise before encoding [8][9]. A significant reduction of the background noise level improved the parameter estimation process which improved the overall synthesized speech quality in the presence of noise. Furthermore reduction of the overall noise enables a more comfortable listening level which is very significant in terms of the tiredness it may cause to the user. The performance of the NPP is dependent on the speed of adaptation of its parameters and correct voice activity detection (VAD). The VAD used in [8] compares the ratio of the current frame's power and the accumulated noise power against a pre-set threshold which works well in reasonably high SNR conditions (typically 10dB or greater). When the SNR worsens this VAD makes occasional mistakes in declaring noise as speech mixed with noise, and speech mixed with noise as noise only. The former reduces the speed of adaptation of the background noise which is not very serious. The latter on the other hand updates background noise while speech present which causes significant distortion in the output speech quality.

We have used an energy-dependent time-domain VAD technique, which helps in better tracking speech and noise levels during harsh background noise conditions. This VAD algorithm estimates the levels of various energy parameters - instantaneous energy E_0, minimum energy E_{min}, maximum energy E_{max} - that are, in turn, used to indicate the SNR estimate of the current frame. The role of E_{max} is to track the maximum value of the input signal, which is done by a slow descending and sharp ascending adaptation characteristic. E_{min} tracks the minimum energy of the input signal and is therefore characterised by a sharp descending and slow ascending gradient. The SNR_{est} represents the ratio between the maximum and the minimum energy for any given frame.

The importance of the SNR_{est} is that its level controls the energy thresholds for the VAD. Namely, the VAD operates according to the ratio:

$$VAD = \begin{cases} 0, & (E_0 / E_{min}) < E_{th} \\ 1, & (E_0 / E_{min}) \geq E_{th} \end{cases}$$

where the value of E_{th} depends on the SNR estimate and is adaptively constrained to be within a limited range of 1.25-2.0.

Another important feature of the SNR_{est} is that it defines the speed of adaptation for the NPP parameters.

In order to reduce the overall NPP+speech encoding/decoding delay, the NPP frame size (up-date rate) must be same as or integer sub-multiple of the speech frame. The NPPs usually have 256 sample window and FFT building blocks which are shifted by 128 samples (up-date rate). A Hanning window is usually preferred since the synthesis process becomes a simple overlap and add. However the up-date rate of 128 samples is unsuitable for the 20ms speech frames. We have therefore used 80 samples up-date rate (176 samples overlap) and applied two NPP processes per speech frame. Since the overlap of the two adjacent NPP processing stages is more than 50%, during the NPP cleaned speech synthesis the two adjacent blocks are first de-windowed (to remove the analysis windowing effect) and then a trapezoidal window is used before overlap/add is executed.

6. SIMULATIONS

In order to assess the performance of the designed coder we have used subjective listening tests. In the tests 2 male and 2 female speakers with two sentences from each were used. The input sentences were also added with noise at 10 and 5dB. Three types of noise were used, helicopter, vehicle and bable. The input level of the signal was set to nominal -26dB during all testing. In the tests A and B comparisons were made. Each sentence was played twice one produced by our coder and one produced by the reference coder. We have used two reference coders, the DoD CELP at 4.8kb/s [6] and MELP at 2.4kb/s [7]. During the comparisons 22 trained subjects were asked to grade their preferences using 2, 1, 0, -1 and -2 to indicate better, slightly better, the same, slightly worse and worse respectively. They were also asked to describe the reasons for their choice.

The coders were numbered as C1, C2 and C3 for SB-LPC at 2.4kb/s, 1.2kb/s and 2.4kb/s+NPP respectively. The reference coders were numbered as R1 and R2 for CELP and MELP respectively.

Comparison	Clean Speech	Noisy Speech
C1 vs. R1	11	-2
C1 vs. R2	9	13
C1 vs. C2	2	1
C1 vs. C3	0	-10

Table 2: Subjective comparison results

As can seen from the results in Table 2, in clean speech there is a clear preference for SB-LPC as compared with DoD CELP. The main reason for not preferring CELP was its rather noisier output quality. The quality of the SB-LPC has been preferred due its cleanness and less muffling. In noisy speech however the preference of CELP was found to increase. There were two main reasons for this. Firstly the reproduction of the background noise by CELP had a more pleasant nature and it was easier to recognize the noise type. The second reason is that since the voicing classification of the SB-LPC was tuned to favor voiced, during the noise only parts some voiced declarations caused periodic components which were found to be unpleasant.

When compared against MELP under clean background conditions SB-LPC was preferred again. The main reason for this was that MELP had occasional artifacts which was found to be annoying and had more metallic nature. Under background noisy conditions the difference was more noticeable. The reason for this difference was that MELP voicing decision mistakes caused roughness in its output speech quality. Some on-sets and off-sets

where the relative noise level was high, were declared as unvoiced.

After the comparison of the 2.4kb/s SB-LPC against the two DoD standards it was then compared against its 1.2kb/s version. In clean speech input case, there was a slight preference for the 2.4kb/s. In the noisy conditions, as expected, the two rates were found to be very similar. The comparison of the 2.4kb/s with and without NPP clearly showed the NPP's effectiveness in noisy conditions. Finally the 2.4kb/s version was informally tested under 1% random bit errors and 3% frame erasures. Although, the random bit errors caused slight degradations, owing to accurate frame substitution methods, frame erasures did not caused noticeable distortions.

7. CONCLUSIONS

In this paper we have presented a split-band LPC based speech coder which is capable of operating at two modes of 2.4 and 1.2kb/s. Both of the modes use the same core analysis and synthesis blocks. The rate halving is obtained by increasing the encoding delay to have efficient quantisation of the parameters with fewer bits. A noise pre-processor has also been integrated with the speech encoder to improve the performance during noisy background conditions.

The coder was tested in two stages. In the first stage the 2.4kb/s version was compared against DoD CELP and MELP algorithms operating at 4.8 and 2.4kb/s respectively. In the second stage two modes of the coder were compared to quantify the degradation incurred in halving the bit rate. In clean input condition the 2.4kb/s version was preferred against both references but in noisy speech condition CELP was found to be slightly better. In the case of 1.2kb/s very similar speech quality to the 2.4kb/s version was produced for both clean and noisy inputs. The use of a NPP at the encoder increased the performance of the coder for noisy input samples. Both speech intelligibility and quality was improved significantly. The 2.4kb/s version was also tested against channel errors at 1% random bit error rates and 3% frame

erasure rates. The random bit errors were found to cause slight quality reductions. However by protecting the RMS energy with a single bit possible blasts were eliminated. The 3% frame erasures did not cause noticeable degradation.

8. REFERENCES

[1] R.J. McAulay, T.F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation", *IEEE Trans. on ASSP*, 34 pp 744-754, 1986.

[2] I.Atkinson, S.Yeldener, A.M.Kondoz, "High Quality Split-Band LPC Vocoder Operating at Low Bit Rates" *ICCASP-97*, Volume 2, pp 1559-1562.

[3] J.P Campbell, T.E Tremain "Voiced/unvoiced classification of speech with applications to the U.S. Government LPC-10E Algorithm", *ICASSP -1986, pp 9.11.1-9.11.4.*

[4] R.J. McAulay, T. F. Quateri, "Pitch Estimation and Voicing Decision Based Upon A Sinusoidal Speech Model", *ICASSP-90*, Vol. 1, pp 249-252.

[5] K.Zeger, A.Gersho, "Pseudo-Gray Coding", *IEEE Trans. On Communications*, 38, no 12, pp 2147-2156, 1990.

[6] J. P. Campbell, T. Tremain, V. C. Welsh, "The DoD 4.8kbps Standard (Proposed Federal Standard 1016)", *Speech Technology*, Vol. 1(2), pp 58-60, April 1990.

[7] A. McCree et.al. "A 2.4kb/s MELP Coder Candidate for the New U.S. Federal Standard", *ICASSP-96*, pp 200-203.

[8] Y. Ephraim, D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", IEEE Trans. On Acoustics, Speech and Signal Processing, Vol. ASSP-32, No.6 pp. 1109-1121, December 1984.

[9] R.J. McAulay, M.L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter", IEEE Trans. On Acoustics, Speech, and Signal Processing, Vol. ASSP-28, NO.2, April 1980, pp 137-145.

This page has been deliberately left blank

————————————

Page intentionnellement blanche

Data Communication and Data Fusion in Rapid Environmental Assessment: state of the art

(May 2000)

Alex Trangeled, Frederik H. Vink[1], and Alessandro Berni
NATO SACLANT Undersea Research Centre
Viale S. Bartolomeo, 400
19138 La Spezia, Italy

1. Introduction

In the evolving world of international security, NATO faces new challenges. In early 1990's the Alliance studied the new security situation and revised its strategically concept. Future conflicts will occur in a wider area of interest with lower intensity and on a regional scale; the area of operation is in many aspects unknown. An active effort is necessary to be able to combat future threats.

The mission of the SACLANT Undersea Research Centre (SACLANTCEN), based in La Spezia, Italy, is to conduct research in support of NATO's maritime operational requirements. Considerable efforts are being made to identify and counter the threats related to underwater warfare. SACLANTCEN performs operations research and analysis, research and development in the field of Anti Submarine Warfare (ASW), Mine Counter Measures (MCM) and Military Oceanography (MILOC).

Rapid Environmental Assessment (REA) is one of the five thrust areas of SACLANTCEN's Scientific Programme of Work (SPOW). The goal of the Centre's REA program is to research methods for providing warfighters and planners with tactical relevant information in a tactical relevant timeframe.

This document concentrates on the technological aspects of data processing, fusion and transmission, illustrating the evolution of the techniques adopted and their innovative impact on MILOC activities.

2. Rapid Environmental Assessment: supporting NATO's Crisis Response Doctrine

The concept of Rapid Environmental Assessment (REA) has emerged in recent years as one of the most interesting research topics in MILOC.

REA is defined as:
"*The acquisition, compilation and release of tactically relevant environmental information in a tactically relevant time frame*".

The definition of "*tactically relevant time frame*" can range from several months, during the operational planning phase, to a few hours, during naval operations.

2.1 Planning

Conventional planning of naval operations commenced months prior to the execution. When the time of execution comes near, the mission is planned in detail and at that point a near real time information flow becomes more important.

In the past, MILOC efforts have been concentrated at gathering information from a strategically important static area. Data was collected over a long period, and several months would pass before actual reporting of the results occurred. When the need emerged, it would eventually find its way into a Tactical Decision Aid (TDA) such as NATO's Allied Environmental Support System (AESS), to produce Environmental Briefing Dockets (EBD). Unfortunately, this process would typically take two to three months, so that the end product would be neither timely or current [1].

---

[1] Seconded from the Royal Netherlands Naval College (RNLNC)

In consideration of the new NATO security scenario that has to cope with multiple risks, including crisis management and humanitarian operations in littoral areas, there is a high probability that the designated operation theatre is an area for which the availability of *a priori* knowledge is minimal. Environmental information should however be available within a crisis response time scale of a few weeks, in order to gain some tactical advantage.

"The tactical advantage will probably depend not on who has the most expensive, sophisticated platforms but rather on who can most fully exploit the natural advantages gained by thorough understanding of the physical environment."

Rear Adm. W. G. "Jerry" Ellis, U.S. Navy, Oceanographer of the Navy, 1999

What is considered to be essential in the planning phase is a prediction capability to provide the planner with a stop light decision aid: green, yellow and red to assign the possibility of executing a mission. Predictive capabilities are useful to foresee a change in the environment that may influence the tactical deployment of people, material or systems.

Prior to and during the executional phase of an operation, knowledge of the environment should be valid and up to date. This knowledge should express tactical useful information; the newly obtained environmental information should be made available for assimilation.

The principal aim of SACLANTCEN's REA investigates efforts is to provide a framework for the prediction of sonar parameters and to supply ASW, MCM and Amphibious Warfare (AW) commanders with environmental information within a time scale compatible with tactical operations. Methodologies and techniques that enable the collection, processing and distribution of environmental data and products within a compressed time frame are implemented, integrating traditional methods of information gathering in MILOC with modern communications and data processing techniques.

3. Data fusion

Environmental information is made available to the warfighter using dependable information technology assets for information gathering, transmission and presentation: state of the art Commercial-Off-The-Shelf (COTS) solutions constitute the basis for REA research and development efforts.

The data collected during an REA survey are transferred to a central location, termed the Data Fusion Center (DFC), from where it can be easily retrieved by the customers. Standard Internet protocols are used for data exchange, for maximum interoperability and platform-independence. Data uploads to the DFC are done by File Transfer Protocol (FTP), data presentation and retrieval is handled by a standard Hyper Text Transfer Protocol (HTTP) server, allowing customers to browse the archive using a World Wide Web (WWW) -browser.

Type-, time- and space-domain search engines have been implemented, presenting the customer with a list of data sets pertaining to his specific area of interest.

Up to now, the main emphasis has been on delivering unclassified data over the Internet, using IP address and password authentication for access control. The unclassified contents have been subsequently transferred to the NATO Initial Data Transfer System (NIDTS), where classified products have been added to the server contents.

4. Communications in support of at-sea experiments

SACLANTCEN started exploiting Internet technologies in support of field experiments since 1994, with real-life concept tests during surveys Yellow Shark 95 and Winter Sun 95 [2]. The following *Rapid Response* (RR) operations involved the evaluation of a wide range of COTS communication methodologies to satisfy the REA requirements. The resulting infrastructure was used to transfer data (both raw and processed) from at-sea platforms to ashore centres and vice versa.

Present research is aimed at the definition of turnkey solutions that can be deployed on site with small advance notice. A *REA in A Box* (RIAB) prototype has been prepared and demonstrated during NATO exercise *Linked Seas 2000*. Furthermore, the goal of future warfighter consultation is to define and prioritize the products that really are considered helpful during the planning and execution of a naval operation, in order to define the necessary REA data flow.

5. Rapid Response REA operations

The *Rapid Response* series of operations, conducted between 1996 and 1998, demonstrated how COTS Internet technologies could be successfully integrated to build *ad-hoc* networks in support of REA surveys. *Rapid Response* and the other experiments conducted so far constitute proof of the effectiveness of the REA concept.

5.1 REA Data and products

During *Rapid Response*, a variety of data types were distributed to survey participants and customers, ranging from simple American Standard Code for Information Interchange (ASCII) files containing Expendable Bathythermograph (XBT) data to large image files containing high resolution satellite remote sensing data.

In consideration of the high volume of data that was generated by survey data contributors, standardization in file formats was found to be essential, to ensure the data consistency and prompt data fusion. File format standardization was not limited to naming issues, but was extended to the data set structure and to the attachment to the file of geographic/time information, in order to ease subsequent retrieval. The header files were subject to minor changes between *Rapid Response 96* and *Rapid Response 98*.
The following sections, extracted from [3], provide a comprehensive listing of supported data types.

5.1.1 Atmosphere

- Meteorological ship observations
- Upper air observations by weather balloons
- Drifting meteorology buoys, deployed from aircraft

5.1.2 Beach and hinterland

- Landsat satellite images
- Systeme Pour l'Observation de la Terre (France) (SPOT) satellite images
- Aerial photographs
- Photogrammetry
- Beach photographs
- Trafficability measures by hand-held bottom penetrometer
- Trafficability assessed by conventional methods
- Maps
- Reports
- Beach profiles
- Surf measurements
- Numerical surf predictions

5.1.3 Ocean surface

- Tidal water level
- Sea surface height from satellite altimeter
- Wave height from satellite altimeter
- Wave height and spectrum from wave rider buoys
- Ocean features by radar images from satellites
- Surface roughness and microwave emission indicating surface wind
- Radiation temperature images of the sea surface
- Ocean color
- Lagrangian current measurements by surface drifters

5.1.4 Water column

- Deep currents by drifters dragged to several hundred meters depth
- Eulerian current measurements by moored current meters
- Current profiles by acoustic current profilers (ADCP) on the ocean bottom
- Current profiles underway by ship borne ADCPs
- Temperature profiles by ship deployed expendable probes (XBT)
- Temperature profiles by air dropped probes (AXBT)
- Temperature, salinity and derived parameters by Conductivity Temperature Depth (CTD) probes
- High resolution parameter fields by towed CTD chains
- Water samples for laboratory analysis

- Transparency by Secchi discs
- Transparency from multi-color satellite imagery
- Chlorophyll from multi-color satellite imagery
- Shipping density (for noise assessment) by naval patrol aircraft
- Spectral ambient noise by sonobuoys
- Directional noise by towed hydrophone arrays
- Reverberation levels by towed hydrophone arrays
- Transmission loss

5.1.5 Ocean bottom

- SPOT satellite images for depth and bottom type in shallow waters
- Airborne laser system for depth in shallow waters
- Single beam echo sounding
- Multibeam area mapping
- Side scan sonar imaging
- Assessment by video cameras
- Bottom grabs
- Cores
- Mechanical bottom parameters by Expendable Bottom Penetrometers (XBP)
- Seismic reflection profiles
- Sound velocity in the bottom layers
- High frequency bottom reverberation
- Inverse modeling of bottom parameters

5.2 Rapid Response 96

Operation *Rapid Response 96* (RR96) took place in a portion of sea between Sicily, Tunisia and Sardinia. It was the first time the experimental concept of REA was validated in an operational context. It was held in support of the Commander in Chief South Atlantic (CINCSOUTHLANT) annual maritime LIVEX, Dynamic Mix 96 in addition to the associated MCM exercise Damsel Fair.
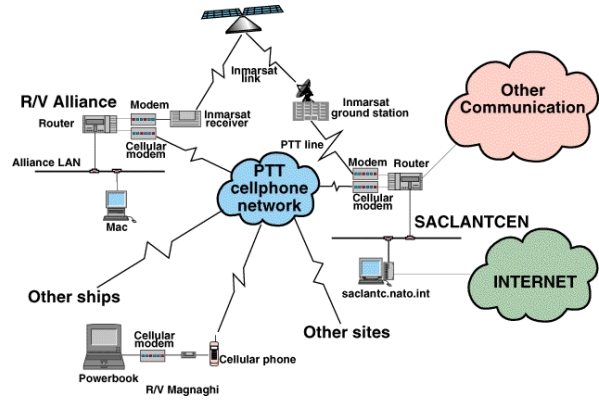
Figure 1: Network Architecture for Rapid Response 96-98 [4]

A total of 6 vessels took part in the REA survey. SACLANTCEN participated with NATO Research Vessel (NRV) Alliance and NRV Manning; other participants were USNS Pathfinder, HMS Herald, FS La Gazelle and the Italian research vessel Magnaghi. The survey vessels transmitted the collected data by standard FTP to the SACLANTCEN Data Fusion Center using Italian ETACS or GSM cellular telephone networks.

The cellular phones were encapsulated in a watertight box and mounted directly on the mast, in order to minimize cable loss. For the ETACS phones, ranges of up to 100 Km from shore were obtained. GSM range was limited to approx. 32 Km.
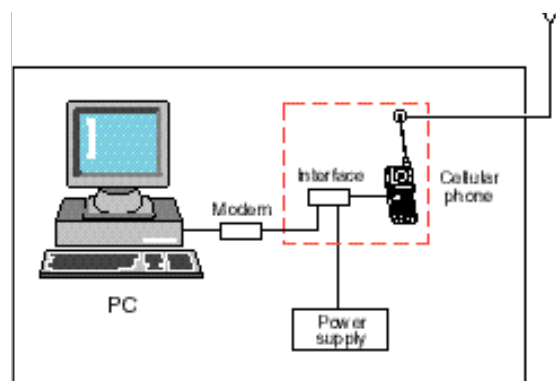
Figure 2: Cellular phone installation RR 96-98 [4]

Maritime reconnaissance for ambient noise evaluation and AXBT measurements were conducted from aircraft based at NAS Sigonella in Sicily. From here, the data were transmitted to the DFC at SACLANTCEN via dial-in through the NATO telephone network. A WWW server at SACLANTCEN was used to present data in an organized structure to external customers.

A copy of all data was transferred using Satellite Communication (SATCOM) to USS La Salle for use during Dynamic Mix 96.

5.3 Rapid Response 97

Rapid Response 97 (RR97), the second exercise of the series, was conducted to support Dynamic Mix 97 and Damsel Fair, taking place in a large area, including the Strait of Messina, the Ionian, Adriatic and Aegean Seas.

NRV Alliance was the leading ship of a fleet of 8 survey vessels. The other vessels were FS D'Entrecasteaux, WFS Planet, HNLMS Tydeman, HMS Roebuck, USNS Pathfinder, HENA Pytheas and ITS Crotone.

Where the data fusion took place exclusively at SACLANTCEN in 96, during RR97 the task was undertaken at sea by NRV Alliance, where most data processing was performed. A two-way mirroring process of both raw and processed data was implemented to/from the SACLANTCEN Data Fusion Center via Inmarsat B, with updates at least every 12 hours.

A detailed database was maintained at both ends to ensure that all parties had a fully up to date data set. The other survey ships had access to the Data Fusion Center via ETACS dial-up connections and Inmarsat, enabling them to exchange data with the server mirror at SACLANTCEN.
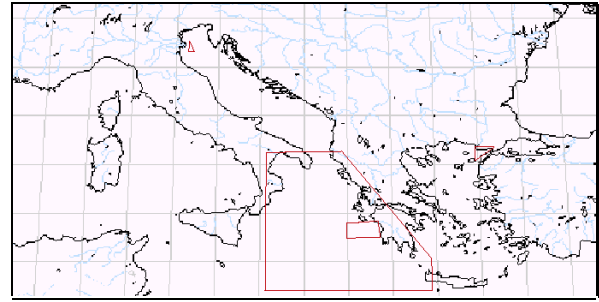
Figure 3: Operational areas for Rapid Response 97 [4]

5.4 Rapid Response 98

Rapid Response 98 (RR98) took place in the Atlantic area south of the Iberian Peninsula. This third and final exercise of the series focused on coordinated environmental reconnaissance in support of Strong Resolve 98. The REA activities integrated air, sea and satellite remote sensing operations with archive data searches to acquire essential oceanographic and atmospheric data for Mine Warfare (MW), AW and ASW.

Participants to the exercise included NRV Alliance, HMS Roebuck, USNS Pathfinder, WFS Planet, SPS Tofino and FS D'Entrecasteaux.
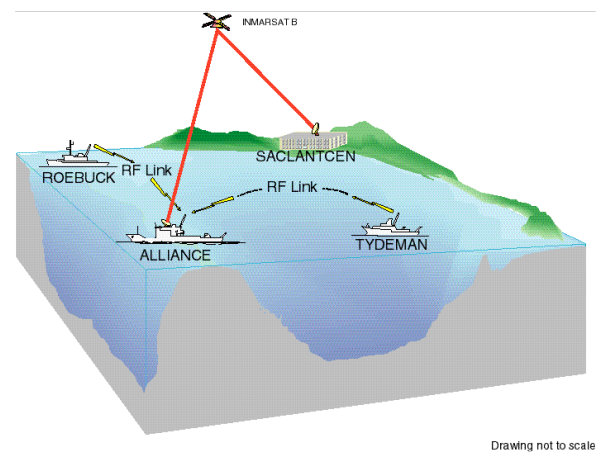
Figure 4: Summary of Network Architecture for RR98 [4]

As in the previous year, NRV Alliance hosted the Data Fusion Centre for the survey. The other survey vessels could interact with NRV Alliance by radio links using standard Internet protocols, or access the mirror at SACLANTCEN via GSM dial-in. At the end of REA phase, the data fusion operations control was transferred to

SACLANTCEN, until the completion of Strong Resolve. A mirror of the REA unclassified web was made available to NATO users also via NIDTS.

6. Current efforts

The *Rapid Response* series of demonstrated how COTS network technologies could be successfully integrated to build *ad-hoc* networks in support of REA surveys. *Rapid Response* and the other experiments conducted so far constitute proof of the effectiveness of the REA concept.

This initial data communications infrastructure relied heavily on cellular phones. However, cell phones are not an option in a denied area. In addition to that, the RF link used for ship-to-ship communications in RR98 did not provide sufficient bandwidth for effective transmission of high-resolution data.

The complexity in configuration and operation of both data processing and communication systems was originally mitigated focusing on the human factor, that is, relying at all times on experienced technical and scientific personnel. Clearly this can not be the case in an operational scenario. The availability of reliable and scalable ship-to-ship links and data fusion architectures is of paramount importance to the effectiveness of REA surveys: present efforts are therefore concentrated on defining a general architecture suitable for use in operational conditions.

6.1 RIAB: REA in A Box

In the present context, where interoperable SATCOM is not readily available on all vessels, environmental measurements are relayed via spread-spectrum line-of-sight wireless data links to a data fusion centre afloat (e.g. a command ship), which can be positioned several miles away. Data are then transferred to a fusion centre ashore using a SATCOM gateway, where they are made available to the REA community (data providers, product developers, and customers) using wide-area computer networks (WAN).

The formalization of the above concept is leading to the definition of a *REA-in-a-box* (RIAB) system: RIAB is a fully-featured solution that can be deployed on site with little advance notice, to provide a low-cost, easy to use system for data exchange between survey participants.

The RIAB system is made of two entities, client and servers, that interoperate to provide an end-to-end solution to the REA requirements.

6.1.1 RIAB Client

The RIAB client system is a package integrating a Personal Computer (PC) and a Spread Spectrum (SS) wireless router for ship-to-ship communications, to be installed on every vessel participating to the REA survey.
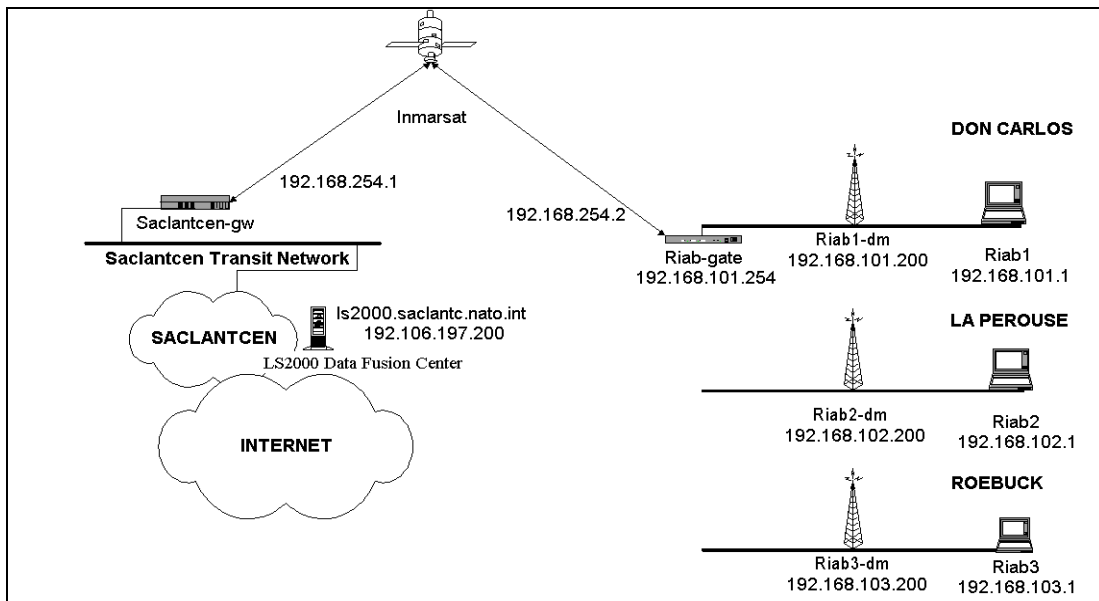
Figure 5: RIAB System for Linked Seas 2000 - Wireless Network

6.1.2 RIAB Server

The RIAB server system is packaged as the client, but is supplemented by an Inmarsat-B High-Speed Data (HSD) satellite link. The transition to military SATCOM systems is a fairly straightforward process.

6.1.3 RIAB Data flow

The RIAB software implementation is focused on robustness and user-friendliness. An elaborate mirroring scheme handles the actual relaying of data, so that all survey vessels maintain copies of all data by recursive data pull. This means that any RIAB system is capable of delivering not only its own data, but also data from any other client, with which a successful data exchange (synchronization) has been effected.
All data received by the RIAB server by the different clients will be subsequently packaged, compressed and transmitted to the remote Data Fusion Center.

6.1.4 Linked Seas 2000

The first field test of a RIAB prototype was effected in April-May 2000, during the REA precursor phase to the NATO *Linked Seas 2000* exercise. The present implementation has been implemented on COTS portable computers using the Linux operating system.

Three survey vessel were equipped with RIAB equipment: NRP Don Carlos (PT), FS Laperouse (FR), and HMS Roebuck (UK), with Don Carlos acting as the RIAB server.
Practical communication ranges between ships were in the order of 12 nautical miles. Overall excellent results were obtained, since all data from all survey vessels (in the order of 30 MB) was delivered with a maximum delay of 24 hours time. Customer feedback was quite positive; the RIAB end-users were able to operate the equipment successfully after a 20-minute brief. Data exchange could be monitored continuously through a normal web-browser.

The few problems that were encountered during LS2000 were diagnosed and corrected remotely during a routine data transfer, without specialist personnel onboard the vessels involved.

7. RIAB - the concept evolving

Since REA support systems such as RIAB are evolving towards implementation suitable for operational use, a continuous interaction with the warfighters, the people who actually have to utilize this information, is necessary.

Raw data are fairly immediately available and can be presented nicely on screen; the products are presented to the warfighter after post-processing and assimilation of the available information. Efforts from multiple and complementary expertise such as oceanographers, meteorologists, operations research, communications and information technology experts and last but not least the customer at sea, the warfighter, need to be fused to clearly define the REA tactical product list. To facilitate this, a two-way information flow is required between the DFC and the naval units.

Future developments provide for this two-way capability. They involve the coupling of RIAB type systems to classified military networks and provide access from the participating units to the full contents of the Data Fusion Center. The fusion center could be located anywhere, including the Command Unit responsible for the survey area, as long as it has the necessary satellite uplink capability, be it permanent or on-demand (from both peers). Such an infrastructure provides for near real-time scientific support and remote counseling from experts in various fields. Naval units are able to receive the most up-to-date environmental products and Tactical Decision Aids through this two-way network.

8. Conclusions

The scientific community and the warfighter should address points of importance in the naval REA product, to fully exploit its value of tactical oceanography to provide battle space awareness for the warfighter at sea.

A general culture aspect should also be addressed. Warfighters and operators of sonar systems have done their job for many years to the best of their ability: now the availability of REA methodologies can radically change the way in which they work. To fully exploit the power of REA, navies need to become acquainted with the new concept. Future development efforts mandate

an involvement of a warfighter perspective to validate progress and prompt operational needs.

RIAB, in its first implementation, has vastly improved REA data management and distribution. Further efforts are required however to arrive at a turnkey MILSPEC system.

Currently endeavors are undertaken to verify the military impact of REA in depth. A dialog between the scientific community and the warfighter is established and will remain important for further REA product development.

9. References

[1] Hammond, N. Rapid Response 96, A demonstration of NATO's Rapid Environmental Assessment capability. *In*: Poulinquen, E., Kirwan, A.D. *and* Pearson, R.T., *editors.* Rapid Environmental Assessment, SACLANTCEN Conference Proceedings Series CP-44, NATO UNCLASSIFIED. La Spezia, Italy, NATO SACLANT Undersea Research Centre, 1997: pp 21-23 [ISBN 88-900194-0-9]

[2] Bovio, E., Max, M.D., Spina, F. *and* Berni, A., Communication Technology in support of SACLANTCEN Programme of Work. *In*: Poulinquen, E., Kirwan, A.D. *and* Pearson, R.T., *editors.* Rapid Environmental Assessment, SACLANTCEN Conference Proceedings Series CP-44, NATO UNCLASSIFIED. La Spezia, Italy, NATO SACLANT Undersea Research Centre, 1997: pp 21-23 [ISBN 88-900194-0-9]

[3] Sellschopp, J. Experience gained and lessons learnt in three years of Rapid Response, *In:* Undersea Defence Technology Europe 98, London, UK, Nexus Media Ltd.: pp 69-73 [ISBN 1899919 28 7]

[4] Trangeled, A., Franchi, P., Berni, A., Data Communication and Data Fusion Architectures for Rapid Response 96-98, SACLANTCEN SR-296. La Spezia, Italy, NATO SACLANT Undersea Research Centre, 1999.

An Overview of Information Fusion©

G D Whitaker
Defence Evaluation and Research Agency (DERA)
St. Andrew's Road
Malvern, Worcestershire, WR14 3PS
United Kingdom

Telephone: +44 (0)1684 895822
Facsimile: +44 (0)1684 894384
E-mail: D.Whitaker@signal.dera.gov.uk
WWW: http://www.dera.gov.uk

1. Abstract

This paper provides an introduction to, and overview of, the field of information fusion within a wider data and information fusion and processing context. It starts by considering the aims and objectives of research and development programmes in this area. In particular, asking what are we trying to achieve by such fusion from the end user (military commander?) point of view. The main emphasis of the paper is on military systems and reference is made to work at the United Kingdom's Defence Evaluation and Research Agency (DERA) for examples but the paper has more general relevance.

Some of the common operational and logistical difficulties associated with current information fusion systems are highlighted. In other words, "Why is making sense of data difficult?" The differences and similarities between data and information and between their fusion and processing are discussed. The rôle of information fusion systems is to address some or all of these difficulties and so provide more effective systems for a range of different applications and users. The means by which this is accomplished is then described in terms of fusing information at various levels of abstraction. Reference is made to models, architectures and frameworks that have been developed independently within the USA and the UK and that help structure and clarify the whole process.

Current research aims to further improve our capabilities in this important, force-multiplying technology. Some people and nations aspire to information dominance in modern conflicts and the same can be said for modern businesses. To achieve or even approach this goal requires a vigorous and healthy research programme. Some of the current key research activities in this area are summarised. Most of this research is targeted on specific, near-term applications. The paper concludes with a personal perspective on the main future, longer-term research challenges.

2. Introduction

The United Kingdom's (UK) Defence Evaluation and Research Agency (DERA) is an agency of the UK's Ministry of Defence (MoD). It has an annual turnover of over £1Bn and around 12,000 staff – around 75% of whom are scientists and engineers. The organisation has 15 major sites in the UK and many smaller sites both in the UK and in other countries. As such, it is one of the largest research organisations in Europe.

DERA was formed from many previous non-nuclear MoD defence research establishments and supports the MoD mainly by providing impartial advice and by undertaking research. DERA is increasingly encouraged to undertake work for other organisations on commercial terms. Such work must conform with the framework of its corporate and business plans as agreed by the Secretary of State for Defence but, even so, this extends the range of the research activities.

A survey in the late 1990s showed that there were over 100 MoD funded projects that were involved in data or information fusion. Many of these only had a small interest in this subject but there were still many that had a major data fusion activity. One might assume that with this number of projects each researching the same subject, there must be considerable overlap and therefore scope for savings measures. On investigation, however, there was actually very little duplication. Each of these projects differed from the others in some crucial way. It also illustrated the depth and breadth of DERA's research in this field.

Some of these projects were addressing land domain problems, others concentrated on maritime (surface or sub-surface), others were concerned with the air domain. This could apply to both the sensor platform and to the targets of interest. Different combinations of sensors were also being used or investigated. Some applications needed real time solutions, others could process off-line.

Some projects were challenged to improve detection capability, others to improve location or identification or tracking. Others were compiling whole pictures of the battlespace of interest whilst others were looking at more abstract concepts such as situation or threat assessment, decision support or even decision making. Some were concerned more with the implementation of such decisions by investigating logistics and planning.

3. Aims and Objectives

There are many military platforms and command and control (C&C) centres that need to fuse information. Initially, the aims and objectives of the different platforms within the various domains appear quite different. Submarines have the difficulties of acoustic propagation and lots of background noise. Surface ships have to deal with fast, sea skimming missiles that are difficult to detect amidst the waves. Land conflicts involve relatively large numbers of platforms each of which has considerable autonomy and is difficult to detect or predict – they are called soldiers. In the air, platforms move rapidly and can perform tight manoeuvres. In space, the main difficulties are the distances involved and observing through the atmosphere.

In most battles, there will be at least one command and control centre overseeing information across the whole battlespace but at higher levels of abstraction than most platforms. Much of the processing at this level relies on very capable neural networks colloquially referred to as human brains.

Is there anything that these domains and platforms have in common? In order to answer this question, we need to look at the underlying situation and ask why these entities wish to make sense of data. Each of these platforms, like each of us, exists within an "outside-world". We use sensors and non-sensor derived information to construct a picture of our perception of that outside world. For the moment, let us regard the data and information as entering a "black box" that contains all the processing and fusion. Coming out of the black box are generally commands and instructions that impact on the outside world. So the simple answer to the question, "Why make sense of data?" is that we wish to find out about that outside world in which we exist. This is rarely the end of the answer because, usually, we wish to influence that outside world to benefit ourselves in some way or another.

The military doctrines refer to an "OODA" loop. That is:

- they **observe** the outside world using sensors and collateral information;

- they then construct a picture representing their perception of that outside world in order to **orient** themselves within it;

- they then **decide** if and what and when and how they wish to influence that outside world; and

- finally they **act** by implementing their decision.

This act may be observed directly and / or the impact on the outside world may be observed and the loop cycles around again.

This OODA loop represents a challenge for military commanders and for the providers of technology to support or to undertake the various functions. In real life, and particularly in modern conflicts, the situation can be much more complex. There is generally an enemy or competitor as well as several neutral parties involved, each of whom is following their own OODA loop. The challenge is to make sure that your OODA loop is better and / or faster than the OODA loop of your opponent(s) or competitor(s) in order to give you the best chance of being successful in your objectives.

4. Current difficulties

There are many difficulties facing specific platforms or specific functions. There are, however, some difficulties common to many of them. One is data or information deluge. Firstly, there is a huge and increasing amount of data and information available to us. Sensors are seeing further and in more detail and in more conditions than ever before. There are also more sensors available. Non-sensor derived information such as books, doctrines, best practices, news feeds, and the internet are also growing at high rates. Secondly, researchers in the communications field have been so successful in recent years that the information is reaching us ever quicker.

In addition to the deluge challenge, information is being generated and provided by a range of diverse sources. There are significant differences in information originating from simple range and bearing sensors, compared with imaging sensors, compared with map data, compared with textual intelligence messages, compared with expert analyses, compared with information gained from past experience. Each of these provides a different challenge in its processing and then the information from the various sources needs combining or fusing.

A particularly important but often overlooked difficulty is that of specifying what is required from a system. Similarly, most providers of such systems would have difficulty in accurately assessing how effective their systems are. It can be even harder to predict the performance of sub-systems and then to aggregate these to predict the overall effectiveness for systems that are yet to be built.

An illustration of this was when the author researched and developed air defence systems. An important person from overseas was visiting the establishment. An

equally important host was showing the visitor around. The visitor was shown the models and simulations of air defence systems and seemed suitably impressed. He then asked the deceptively simple question, "So how good is the UK air defence system?" The host thought for a while and then replied, "About 5." The visitor looked quizzical for a moment and then asked the supplementary question, "On what scale is that?" To which the reply was, "On any scale you choose."

Challenges that face many systems include the detection of objects or events, their location in space and / or time, the identification and recognition of those objects or events, the assessment of the situation, the making of a decision, and the planning of the implementation of that decision.

More specific challenges affecting smaller numbers of projects or domains include: low-observable targets, high noise, manoeuvres, data incest, multi-platform architectures (get data to right place at right time - push versus pull), data fusion / processing architecture map to C&C or vice-versa, conflict resolution (C&C versus own platform priorities), robust against information warfare, fusion of numeric, categorical and "soft" data, interoperability, bandwidth, confidence factors, variable latency, variable reliability, variable quality (accuracy, uncertainty, duplicates, omissions, contradictions, ambiguities).

At this point it is worth making a few observations about digital data links and about digitisation in general. Both of these are vital both in their operational significance and in the work being undertaken to develop or further improve their capabilities. They are not, however, the total solution. In particular, conformity with the data link protocols does not guarantee that platforms can meaningfully interoperate. Similarly, just because data are stored digitally, does not mean that they have been fused or that they are accurate. Whenever data are being transferred, there will always be some latency or time delay. Bandwidth is not, and never will be, infinite and so it will not be possible to transfer all the data that are available. Even if the latency were zero and the bandwidth were infinite, there would still be the issue of the data deluge and the need to extract the **relevant** data from that deluge. Hence the observation that, even with the improvements that data links and digitisation provide, data and information fusion, processing and management algorithms will be essential.

To summarise the requirement: military commanders (like everyone else) need **appropriate** information on which to base their command decisions. This information must be necessary, in that unnecessary information just adds to the confusion. It must be sufficient or the commander may be missing the crucial fact that will undermine his or her decision. It must be timely as information arriving a day late, an hour late or even a second after the decision has been made is of limited, if any, use and may even be detrimental to the future battle. Finally, it must be in a suitable format. If the information is being received by a computer and it is not in a suitable format then the receiving computer will, probably, ignore it – or crash. If people are receiving the information, then they may be tired, hungry, frightened, and uncomfortable (especially if they are in a battle where they are being fired upon). Military commanders under such stress may politely (!) suggest that such information is reformat so that they can more readily absorb it.

Data and information are provided by sensors (such as radar, sonar, optical and infra-red devices, aerial and satellite photographs) and other sources (such as intelligence, open-source, background knowledge). Modern systems outstrip the abilities of human operators, even well trained expert military staff, to absorb, process, and make all the correct inferences based on that data. This is especially the case when the information is "difficult" in some way. For instance when it is incomplete, uncertain, ambiguous or contentious. Hence there is a growing need for automated support.

5. Rôle of data and information fusion

Many people spend inordinate amounts of time discussing possible definitions of data and information fusion. For the purposes of this paper, the following, quite open, definition will be used:

> *Data fusion is the combining of data.*

Strictly speaking, this is all that data fusion is. This alone, however, does not justify their embarking on lengthy programmes attempting to fuse their data. Such programmes might be fun and may provide considerable interest, enjoyment and satisfaction for those undertaking the work. Given that funds and time are generally limited, however, one would hope that there was a purpose or objective to justify the development of this fusion capability.

This objective is usually improved performance in some way. For instance, increased accuracy, increased robustness, decreased time, increased resilience to counter-measures, less false alarms, or more effective decisions.

Consequently, a more realistic definition of data fusion is:

> *Data fusion is the combining of data ....*
> *..... to achieve an objective.*

From this open definition it can be seen that data can be combined or fused in many different ways. It can be fused:

- **Within a single sensor at each moment in time**. For example, within an imaging sensor,

adjacent pixels can be combined to form elements of the total scene. Another example is a radar that generates a range and a bearing that can be combined to deduce a geographical location.

- **Over time.** A series of observations can be used to predict a target's movements or that an engine is about to fail.

- **Among similar sensors.** A large part of the battlespace can be covered with a series of adjoining or overlapping sensors. For instance, siting radars along a coastline to detect incoming aircraft or arrays of sonobuoys attempting to detect submarines.

- **Among disparate sensors.** Using a sensor with a wide field of view to cue a more focused or accurate sensor. Exploiting complementary features of, say, infra-red and visible light to improve the identification of targets.

- **Sensors with non-sensors.** Data from battlefield sensors combined with a knowledge of doctrine may help assess enemy threats. Intelligence reports could cue local sensors to search for targets in likely regions or at specific times. News feeds and the internet can also provide a wealth of information.

- **At various levels of abstraction.** This will be expanded upon later but, for now, raw sensor data can be fused as can probabilities or decisions.

Finally,

> ***Information*** *is data that is relevant to this application.*

Thus, one application's or one person's **information** is merely irrelevant **data** to another application or person. This definition also implies that there is much overlap between the underlying techniques for **data fusion** and those for **information fusion**. Consequently, the definition and expansion of **data fusion** can be repeated for **information fusion**.

To determine the rôle of data and information fusion, we need to return to the "black box" introduced in section 3. Information about the outside world was taken as input, and commands and instructions were issued as output. Now we consider the content of that black box. First, however, it should be observed that this black box may contain people and / or it may contain computers. Ideally it will contain the most effective mixture of people and computers for the particular application being undertaken. In many effective systems there is good synergy between humans and machines.

Our interest is not in whether there are people or computers but in the key components of the black box.

What are the sub-functions and how do they relate to each other? It should be noted that, theoretically, such a black box need not have pre-determined sub-components. Research into self-organising systems suggest that appropriate structures can be discovered or emerge automatically without external supervision (Reference [1]).

Several groups have considered this question. Perhaps most notable is the United States of America (USA) Joint Directors of Laboratories (JDL) with their data fusion model. This has evolved over the years and an interpretation of a recent version can be seen in Figure 1.
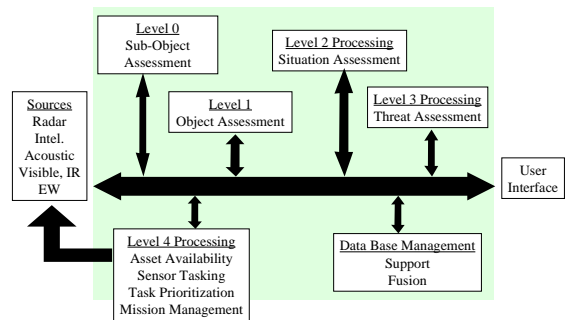
Figure 1 - USA JDL Data Fusion Model

Developed at around the same time as the JDL model but updated for this paper was one from the UK. This can be seen in Figure 2.
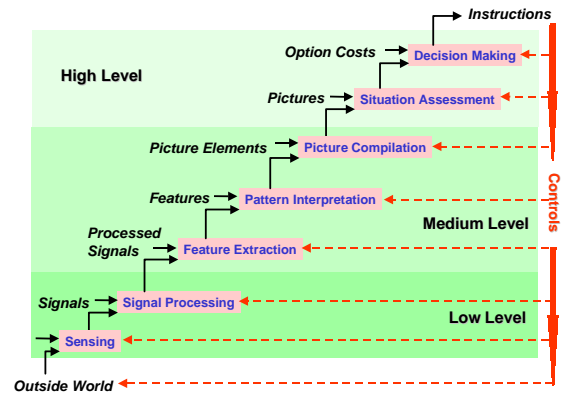
Figure 2 – A UK Data Fusion Model

Both models serve the excellent purpose of clarifying understanding and facilitating communications in this complex technical field. The USA JDL model better covers the context within which data fusion sits. For example, the important aspects of data base management and user interfaces appear explicitly. On the other hand, even in comparison with this later USA JDL version, the UK model is more explicit as to the contents of the "black box".

It may be helpful to consider the meaning of these key components:

- **Signal processing.** The initial analysis of the raw sensor data. Usually provided as a part of the sensor system and relying heavily on a good understanding of the sensor and the physics of the environment in which it operates.

- **Data feature extraction.** The extraction or selection of significant attributes from the processed data. It is important to determine the most useful features that support subsequent processing and fusion whilst ignoring irrelevant or less relevant aspects of the data.

- **Pattern processing.** Analysis of the features to estimate the elements of the total picture. These elements could be individual targets.

- **Picture compilation.** The combination of all of the picture elements into a total picture. That is, a local perception of what is happening in the outside world.

- **Situation assessment.** The determination of what the picture means from the viewpoint of interest; the determination of possible futures based on options available to each of the participants; and the evaluation of those possible futures from the viewpoint of interest.

- **Decision making.** Selecting the action such that the resulting sub-set of possible futures is desirable. The entity making the decision has a number of options that have been identified during *situation assessment*. One of those options needs to be selected (doing nothing is an option) and this should be chosen such that the future is more desirable than that resulting from any of the other options. This could be the future in which the entity is in the most favoured position. It could, however, be a good option that can be identified quickly - even if it may not be the best. This might keep up the battle tempo and get inside the enemy's OODA loop. It could be a good option that is robust to uncertainty in the picture or uncertainty in the assessment or to countermeasures. The choice will depend on the strategy, terms of engagement etc. being followed by the entity making the decision.

It should also be noted that higher levels can often beneficially influence lower levels. A simple example is the deployment of a sensor to provide further valuable information.

These components represent increasing levels of abstraction. Levels of abstraction also exist within command and control (C&C) systems. That is not to say that each and every level identified here can be found in C&C systems or that the levels can be separated out in this way. By considering such hierarchies, however,

sensible questions can be posed as to where fusion should take place in order to maximise the decision making capability of the total system whilst taking account of practical constraints. Fusing at a low level provides the best prospect for optimum use of the data but requires higher bandwidths so may be limited to fusion onboard a single platform. Fusing at higher levels reduces the bandwidth requirements, but relies critically on the accuracy of information provided by the lower levels. Intermediate levels can offer compromises on these extremes.

C&C systems often have hierarchies within them. For example, a platform will, generally, have a number of sensors and data links providing input that is to be combined with on-board databases and the like. Typically there are several platforms engaged in any operation. These platforms are controlled by, for instance, an airborne C&C (such as AWACS – Airborne Warning and Control System) or a ground C&C (such as UKADGE – UK Air Defence Ground Environment). These in turn provide information and receive commands from a higher level, such as a Combined Air Operations Centre, CAOC. Thus, this example C&C system has multiple levels. The sensors will be operating on, for instance, range and bearing or pixels. This level of detail will rarely be of interest to the commander based in the AWACS who will be more concerned with the locations and trajectories of hostile targets. Even this may need to be abstracted to, for instance, the numbers of aircraft on combat air patrol in certain regions of the battlespace. The commander at the CAOC will, typically, find this too detailed and be more interested in approximate sizes of enemy forces, their likely targets, and the status of defence capabilities that could be brought to bear on this threat.

Another perspective on these hierarchies can be found in multi-platform data fusion. A package of aircraft may consist of a main attack squadron with a flight of helicopters dealing with ground forces and back-up fighters in the rear. Each of these groups will be exchanging information internally and the groups within the package also need to share summary information with the other groups. The whole package may be in communications with air and / or ground command and control stations who, in turn, will be relaying information with headquarters.

Even within one layer in this example, there are hierarchies. A single air platform will have a variety of sensor systems (visual, radar, electro-optic, identification-friend-or-foe – IFF etc.), often more than one communications system (Link 16, SATCOM, on-board communications etc), several weapons systems, tactical and geographic databases, intelligence sources etc. Consequently, each of these systems should be regarded as a system of systems in its own right.

Now consider the fusion of information at some of the levels within such a hierarchy. Firstly, fusion at the sub-system level – for example a ground based radar. The radar will **observe** by making detections that can be confirmed by temporal fusion of data and by combining data from different modes of operation. The output from the radar may well be tracks. The radar data processing system may then **orient** such data into, for instance, latitude and longitude. The system will then **decide** if mode changes are required to improve this local picture. If a change is desirable, then the system will **act** on that decision and implement the change.

Secondly, fusion at the platform level – for instance a ship or fixed-wing aircraft. Such platforms will **observe** by associating the inputs that they receive from their various sensors and communications links to form tracks. The platform system will then **orient** these tracks by fusing them and perhaps combining them with maps to form a local recognised air or sea picture. A **decision** will then be made on any action to take such as moving the platform to gain further information or to avoid a threat or to switch-on or redirect a sensor. The platform with then **act** by implementing these decisions.

Finally, fusion at the command and control centre (C&C). C&C will **observe** the inputs from its own controlled sensors and platforms. It will then **orient** by fusing this input and forming a wider picture. It will **decide** on what actions to take, and then plans will be formulated and taskings prepared. The **act** is to issue these taskings as commands to the sensors and controlled forces.

In each case and at each level of these hierarchies, there is an OODA (**o**bserve, **o**rient, **d**ecide, **a**ct) loop being enacted. Looking for levels of abstraction and OODA loops can help with the understanding of the processes. This in turn can lead to more effective command and control systems.

6. Future challenges

What of the future - where will the challenges lie? In the near term future, the challenges lie in addressing and solving the current difficulties identified earlier in section 4. That answer is a little too glib and does not aid thinking of the longer term direction. Obviously, no-one can know with certainty what new difficulties and challenges we will face. The author's personal, current opinion is as follows.

An all pervading goal is to get a better grip on specifying requirements, assessing systems and measuring effectiveness. The author makes no apologies for repeating this item from the earlier list of current difficulties. Much further progress is necessary in this vital area to set everything else in context.

More specific challenges that were not listed earlier include:

- **Improved models.** Many models make gross assumptions in order to make progress or to achieve real time performance. Such assumptions include linear motions and Gaussian noise or distributions. Where these work for specific applications, this is acceptable. Generally, however, the real world consists of non-linear systems and non-Gaussian distributions. Ask a fighter pilot if, when being targeted by a heat seeking missile, they would be happy to fly straight, level and at a constant velocity! As computers operate ever faster, the performance costs of more accurately modelling such real world characteristics become less critical.

- **Consistent uncertainty handling.** Some systems still assume there is no uncertainty in the underlying data or their resulting decisions. Denying subsequent systems this vital additional information can make the difference between a good decision and a bad decision; between surviving and being killed; being victorious and suffering defeat. Providing accurate, or at least improved, estimates of the uncertainty can, in many cases, be more important and more useful than gaining the last iota of accuracy in the decision itself. A "hard" decision, one with the uncertainty or confidence removed, can act as a veto over other systems that may be being more honest over their uncertainty and so negate many of the desired benefits from information fusion. Further advantages can be gained if each system handles uncertainty in a consistent manner. There are a number of well known methodologies for dealing with uncertainty in a sound and rigorous manner. For example, Bayes, Dempster Shafer and Fuzzy Logic.

- **Mixed type fusion.** Increasingly, modern systems are introducing fusion of information but usually at the same or a similar level of abstraction. For instance, combining arrays of pixels from imaging sensors, or range and bearings from radars, or text from intelligence reports. More of a challenge is to combine information from different levels. Techniques to do this, such as Bayesian Belief Networks or Graphical Information Models, are emerging from research laboratories but they are still limited. Particular difficulties are experienced when trying to fuse dynamic and especially fast changing data.

- **Higher levels of abstraction.** Fusion is becoming ubiquitous at lower levels of abstraction. Higher, more abstract levels are still, predominantly, left to humans. The first

challenge at these higher levels is to extract the more symbolic information from the lower level data. Once this has been achieved, there still remain further difficulties. Many of the usual assumptions that we make about data no longer apply. For instance, size and proximity do not, necessarily, make sense when dealing with categorical data. Is an *armoured personnel carrier* nearer to a *truck* or nearer to a *main battle tank*? In the past, humans were being overloaded by low level data and could not see the overall picture without automated assistance. Now, people are getting a better view of the entities in the battlespace but cannot always extract and assess the situation or threat or their defensive options. There is an appropriate saying that "we cannot see the wood for the trees".

- **Resource management.** People, sensors, weapons and equipment are examples of resources that may be scarce, valuable and / or expensive. It is critical to modern, high tempo operations that such resources are well utilised to maximise the effectiveness of command and control systems. For example, sensors may be providing more data that simply re-enforce current perceptions of the outside world. It may be more effective to relocate or redirect such sensors to gain information about an area of ignorance. Similarly, improved planning of the deployment of a suite of sensors that takes into account the current state of knowledge and ignorance could reduce the time taken to detect and track a target whilst using less sensors. Other examples include the scheduling of requests for use of a resource within tight time constraints or optimising movements across route networks.

- **Frameworks and architectures.** In an ideal world, from the perspective of automating the decision making process, all data would be made available at a central point, processed and fused, and a decision made. Unfortunately, real life does not permit this. Each sensor system and each platform will undertake fusion and processing. Any information passed on to other participants will suffer a time delay. There will also be information lost due to the limited bandwidth and to fixed message types. This situation will be aggravated with each onward transmission. In a stable environment, an **architecture** can be devised that takes into account the ideal goals of the decision making process and tempers them with reality to compromise on a pragmatic system that performs effectively within the constraints. The old "cold-war" was, arguably, such an environment. Modern conflicts and operations, however, are more difficult to predict. More

attention should, perhaps, therefore be given to identifying a kit of parts, or **framework**. Such a framework needs to identify useful components and suitable mechanisms for joining those parts together. When a new situation, for conflict or peace keeping or something else, occurs, the appropriate parts can be selected and connected. This can take place rapidly, safe in the knowledge that the parts will work together to form an effective (note that this is *effective* rather than *perfect*) command and control system. A consequence of this approach is that more attention will need to be paid to the interfaces between the components than hitherto. An additional benefit of this type of approach is that the individual components can be procured separately – even several competing versions of the same part could be developed if this were thought desirable – which may be attractive both financially and in reducing risk.

7. Summary

This paper has provided an overview of information fusion. It has started with the basic need for data and information fusion. There are many application specific requirements to be satisfied but the underlying need is to find out and to influence. Generic aims and objectives have been identified and discussed and some application specific requirements listed to give a flavour for the challenges facing this technology.

The rôle of data and information fusion has been described, including two models that aim to clarify thinking and aid communications for workers in this domain. The relationship between this technical hierarchy of abstractions and military command and control systems has been explored with illustrative examples at a variety of these levels. Finally the paper has indicated the author's view on the challenges that we all face in the future. Whether this is an accurate prediction or not, remains to be seen. What is clear, is that information fusion will become increasingly important in future military operations to the extent that such operations will become infeasible without increasingly automated support. Information fusion is, without doubt, a crucial force multiplier in our modern world.

8. Acknowledgements

Many of these projects have involved staff from DERA's pattern and information processing group and it is they especially who have helped form the author's views over many years of patient explanations and shared enthusiasm for the work.

In addition, special thanks are due to Martin Ferry who contributed the original versions of some of the air warfare viewfoils on which parts of this paper are based. Similarly, many other colleagues across DERA have made contributions either directly or indirectly.

9. References

[1]   Luttrell S P and Webber C J S, 1999, DERA/S&P/SPI/TR990564/1.0 (local ref. DERA/S&P/SPI/651/FUN/STIT/5_17/1.0), Using self-organising neural networks to discover structure in data.

Processing and Fusion of Electro-Optic Information

I. Davies
DERA,
D30 Portsdown West,
Portsdown Hill Road,
Fareham, Hampshire.
PO17 6AD.
England.

Introduction

The UK Defence Evaluation and Research Agency (DERA) has been researching over many years the use of knowledge-based techniques for the automation of information fusion within combat management systems functions. All-source automated data fusion techniques have successfully been demonstrated at the platform level and are currently embodied in a testbed called CMISE (Combat Management Integrated Support Environment). This makes use of own platform sensor data and tracks from other platforms via datalink for the automatic construction of the platform's tactical picture.

The Data Fusion Module (DFM) within CMISE correlates at two levels, track and multi-track. Track correlation joins tracks from similar sources to form multi-tracks and multi-track correlation joins multi-tracks (from dissimilar sources) to form vehicles. Tracks and multi-tracks are correlated by a rule-based system using multi-hypothesis techniques supported by probability based algorithms.

The data sources currently correlated by CMISE are radar, Electronic Support Measures (ESM), datalink, sonar, Identification Friend or Foe (IFF), plans and geographic information. This paper describes the modelling of an EO sensor and the effects of including data from such sensors in a fused tactical picture.

DERA has been evolving the capabilities of CMISE in support of the applied research programme for over ten years. The requirement for a substantial increase in the level of automated support system comes from:

- a rapid increase in the amount of data available to Command. More sensors are available, producing more data;

- in the drive to improve the extent and quality of tactical information, automated methods are potentially faster, more reliable and more consistent than manual methods;

- increases in hostile target mobility and weapon lethality particularly in the littoral battlespace, stressing the importance of accurate and timely identification of targets;

- pressures to reduce platform through-life costs, particularly through reduction of manning.

As well as addressing the above issues, automating the tactical picture compilation process allows operators to focus attention on situation assessment and resource allocation (actually fighting the ship) instead of being consumed by the mundane and repetitive track fusion and identification tasks for which automation is more suited.

Data Fusion concept

Data fusion in this context is the process of combining multiple elements of data from disparate sources in order to produce information of tactical value to the Command, hence reducing the information load on operators and improving the tactical picture quality. This data, both real-time and non real-time, includes ESM, radar, IFF, infrared, sonar, intelligence information, Operating Procedures and Own Force Plans. Sources may be similar, such as radars, or dissimilar such as electronic emissions and infrared.

Data fusion usually occurs either at the plot (measurement) level or at the track level. At the plot level data is fused using the raw sensor output. At the track level data is fused after a track extraction and state estimation process. Different sensor types produce different types of data, for example position and velocity for radar, bearing and emitter parameters for ESM and bearing and acoustic signature for passive sonar. Fusion between plots or tracks is only possible if two sets of data contain measurement of at least one similar attribute, e.g. fusion of radar tracks with ESM tracks by bearing analysis.

In DFM new tracks always form a new multi-track and (tentative) vehicle (V'). The track is then compared against other vehicles' tracks and multi-tracks in an

attempt to correlate it. If the correlation is possible a tentative link is established between the new track and the associated multi-track. New track updates either confirm correlation or the correlation fails. A track can have tentative links with more than one multi-track. A correlation link will confirm when only one tentative link remains. If there are no tentative correlations the track is classified as an established vehicle (V). Figure 1 shows this process.
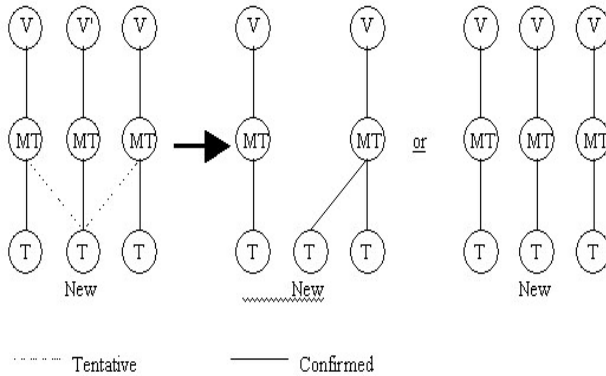
Figure 1. Track correlation.

Other, more complex, processes have also been developed to implement multi-track correlation, repair, track confirmation and the inclusion of collateral data, such as plans, geographic, etc.

Following track and multitrack correlation, a function then combines identity evidence associated with each contact to establish its platform identity and hostility. Many categories of stored information are used to identify contacts, such as structural models to define the relationship between measured contact attributes (acoustic/radar signatures, radio frequency emissions, etc) and contact classes, and behavioural models to relate the temporal behaviour (velocity, altitude, etc) and spatial behaviour (contact formations, weapons ranges) to contacts and events.

The production of the fused tactical picture makes available the following types of information:

- position and velocity,

- identification of contacts and associated uncertainty or ambiguity between multiple possible contacts,

- situations of military importance resulting from individual contact locations, behaviour or aggregate behaviour of multiple contacts,

CMISE modes of operation

The CMISE test bed can produce the real-time tactical picture using data from either live real world sensors (and collateral sources) or recorded real world sensor data or simulated sensor data. CMISE receives tracks from all sensors when in live mode. A simulator/stimulator system called the Object Oriented Programming scenario data generator (OOPSDG) is used when simulated data is required. This system stores and maintains the position and identification of contacts in a scenario together with relevant environmental data. The system uses sensor models (plus contact and environment information) to produce tracks that are output to CMISE. OOPSDG also contains clutter models for each sensor type to produce random clutter tracks (clutter is already present if recorded sensor data is used).

OOPSDG currently contains sensor models for radar, sonar, ESM and IFF. This paper will now describe recent work towards developing an additional electro-optic (EO) sensor model within OOPSDG. It describes performance estimates found prior to producing the completed model.

EO sensor model requirements

The drivers for production of an EO sensor model for inclusion in OOPSDG in order to stimulate CMISE are:

- to investigate the benefits of inclusion of EO sensor data in the fused tactical picture,

- to determine the tactical picture requirements of an EO sensor specification.

Previous work under the current project, investigated the potential use of EO sensors in a naval context. This study concluded that a great deal of tactically significant information is available from EO sensor systems, especially during low intensity operations. An EO sensor model for stimulation of CMISE was proposed as a means to investigate the possible benefits through improved tactical picture quality. All types of EO sensor were investigated for potential to improve their contributions. However, the initial EO sensor model was based on that most likely to be fitted to near future Royal Navy (RN) warships ensuring that the modelled sensor capability matched that of future RN platforms.

The model will be used to determine the effects on automated tactical picture compilation of such attributes as update rate, field of regard, false alarm rate, bearing and elevation accuracy, and detection, recognition and classification range requirement limits, (i.e. the minimum values necessary for improved tactical picture quality).

EO model description

The initial EO model will be based on an infra-red search and track (IRST) sensor, which is primarily

designed for detection of sea-skimming missiles. For this reason, the IRST field of view is concentrated on a region a few degrees either side of the horizon and image processing to detect point source targets (point source targets are objects that are at sufficient range to fill only one pixel on the IRST detector array).

An initial equation was provided by the EO sensors group within DERA to calculate the infrared signal strength from a generic sea-skimming missile as a function of atmospheric path attenuation, target IR signature, range and processing threshold, Equation 1.

$$Signal = \frac{\left(T_h + T_s(\sin\theta)^P\right)\exp(-\sigma R)}{S\,R^2} - Th$$

Equation 1. IR signal strength.

Where T_h and T_s are the target IR signals at zero range for head-on and side-on views respectively, P is a signature modification factor, R is the range between target and sensor, σ is the atmospheric absorption, θ is the viewing aspect of the target measured from head-on, Th is the processing threshold of the sensor and S is the sensor noise equivalent irradiance.

The equation used to determine target signal strength is accurate to a first approximation. The equation does not account for secondary factors affecting IR signal strength such as scintillation or solar heating of target surfaces. These factors produce changes in IR signal strength smaller than the errors in Equation 1 and were therefore neglected. Using Equation 1 to calculate object detection range gave a maximum error (in range) of 15%. A low fidelity model was developed as it best typified EO sensor detection behaviour at a level good enough for tactical picture fusion.

Detection of a real world object is not deterministic and has a random component. Objects at a given distance from a sensor have a certain probability of <u>not</u> being detected even if the signal strength given by Equation 1 is greater than the detection threshold (and also a non-zero probability of being detected when the signal strength is below the detection threshold). The statistical nature of object detection is modelled by the cumulative normal distribution, Equation 2. Standard statistical tables of the normal distribution give critical values of signal strength for given detection probabilities.

$$Pd = \int_{-\infty}^{signal} \frac{1}{\sqrt{2\pi}}\, e^{-\left(\frac{x^2}{2}\right)}dx$$

Equation 2. Probability of object detection.

We desire the interval of range values over which the probability varies significantly. Equation 1 cannot be solved explicitly to find the range for a given signal and threshold. An iterative formula was therefore used to solve the equation and determine the minimum and maximum range for likely first detection of an object (i.e. the 99% and 1% detection probability ranges respectively), Equation 3.

$$R_{i+1} = R_i - \frac{R_i^2 A - e^{-\sigma R_i}}{2AR_i + \sigma e^{-\sigma R_i}}$$

Equation 3. Detection range iteration formula

Where $A = S(Signal+Th)/(T_h+T_s(\sin\theta)^P)$ and R_i tends towards the correct range as i increases.

Iterative calculations are computationally time consuming and Equation 3 was therefore not implemented as the sensor equation in the OOPSDG EO model. Target IR signature data was obtained and used with Equation 3 to produce tables of detection range (for both 99% and 1% detection probabilities) as a function of target viewing aspect, atmospheric absorption, target IR signal and processing threshold. An estimated equation for a curve of best fit for Equation 3 (not using iteration) was obtained. A least squares approximation was performed for each target type at threshold intervals of one (from three to ten). The least squares approximation produced the coefficients for the detection range equation of best fit, Equation 4.

$$DetectionRange = C_1\alpha^4 - C_2\alpha^3 + C_3\alpha^2 - C_4\alpha + C_5 - C_6\theta^2 + C_7\theta$$

Equation 4. Generic detection range equation.

Where C_i is constant i, α is the atmospheric absorption and θ is the target viewing aspect measured from head-on. Equation 4 will be implemented in OOPSDG to model generic detection of targets. A linear relationship between detection probability and range will be assumed between the calculated values for 99% and 1% detection probability ranges. It was found that Equation 4 was a best fit for modelling detection probabilities of fixed wing aircraft and missile target types. Equation 5 was found to best model detection probabilities for ship and helicopter target types.

$$DetectionRange = C_1\alpha^{-C_2} - C_3\alpha^2 + C_4\theta - C_5$$

Equation 5. Generic detection range equation for ships and helicopters.

Limitations of EO model

The initial EO model described has the following limiting factors:

- The model applies to point source targets only; a point source target is detected when the target signal

strength at the detector is greater than the detector threshold, i.e. point source target detection is based only on IR signal strength. Objects that fill more than one pixel in the detector array (extended objects) may be detected using techniques other than IR signal strength measurement, for example object detection based on target shape. Object detection models using methods other than IR signal strength will require different equations.

- Equation 1 applies only for target viewing aspect from head-on to side-on (0-90°); the initial range detection equation was derived for a generic sea-skimming missile. For such target types it is a reasonable assumption that the target will most likely be viewed head-on or near to head-on (i.e. target travelling towards the sensor on Ownship). Equation 1 has been modified to account for rear-on IR signals for all target types.

- Secondary factors affecting target IR signal strength have been neglected; it is planned to enhance the fidelity of the EO sensor model by accounting for secondary factors as the next iteration of the model.

- IR clutter has not been modelled; IR clutter, from cloud edges, sea glint, birds etc, are a major limiting factor for automating target detection using a real EO sensor. An IR clutter model is necessary for the completeness of the sensor model. Such a model is to be included in the next phase of the project.

EO sensor model proof of concept

The equations described previously have been used to produce a PC version of the EO sensor model. The PC version was coded in order to verify the sensor model concept. This version of the EO sensor model used a look-up table containing constants for four different target types as inputs to Equation 4. Equation 4 was used to determine the 99% and 1% detection probability ranges for a target. One percent of detectable targets were randomly undetected and one percent of undetectable targets were randomly detected to reflect the statistical nature of the target detection process. A linear relationship between detection probability and range was assumed for ranges between the calculated 99% and 1% detection probability ranges.

Validation of the PC EO model (and therefore the supporting model equations) was achieved by comparison of model calculated detection ranges and real target trials recorded detection ranges. The atmospheric absorption coefficient was not available for trials recorded data. It was approximated by a value of 0.1 representing good IR transmission through the atmosphere (or 'good' weather conditions, i.e. a low amount of water vapour content in the atmosphere) through to a value of 0.9 for poor IR atmospheric transmission. Differences between calculated and measured detection ranges were within the errors of the calculated and measured values for target data available.

Data fusion model investigation of EO contribution to tactical picture

A faster than real time data fusion model separate from CMISE, has been developed in order to rapidly assess data fusion performance prior to use of CMISE. The data fusion model uses simplified sensor models for radar, ESM, sonar and IR sensors to produce a tactical picture [1]. Targets are given statistically random positions and motions. The simplified sensor model equations are then used to determine target detections. Tracks are fused in a similar manner as the data fusion process of CMISE described previously. The data fusion model outputs measures of tactical picture quality, such as picture completeness, picture correctness, correct correlations, etc.

The data fusion model has been used to perform a preliminary investigation of the effects of including EO sensor data in a fused tactical picture. Two sets of data were obtained. One set of data corresponded to all sensor data including EO. The second set omitted the EO sensor. Averages for each tactical picture quality metric were calculated for ten, twenty minute 'scenarios', both with and without EO sensor input.

Results show that tactical picture quality was improved with the inclusion of EO sensor data.

Inclusion of EO sensor data improved tactical picture correctness[1] owing to the accurate angular measurement of EO sensors (compared to that of other sensors): the accurate bearing (and elevation) data from EO sensors restricts the volume (and thus number of possible incorrect associations) considered in the fusion process resulting in a more correct tactical picture. This fact was verified by observing a directly proportional relationship between EO sensor bearing accuracy and the number of correct associations (and hence, tactical picture correctness).

The inclusion of EO sensor data reduced tactical picture completeness[2]; the inclusion of additional sensor data in the tactical picture results in increased numbers of associations that have to be made for a complete tactical picture. The reduced tactical picture completeness shows

---

[1] Tactical picture correctness is given as the number of correct pairwise associations made by the fusion system divided by the total number of pairwise associations made.

[2] Tactical picture completeness is the mean number of correct objects, where an object is considered correct if: all the objects tracks come from the same real world object; all the real world object associated tracks are associated with the picture object; and at least one track supporting the picture object has a current sensor report.

that a smaller proportion of the increased number of associations were achieved when EO sensor data was included in the tactical picture compilation process.

Inclusion of EO sensor data increased the number of incorrect and missed correlations. Inclusion of EO sensor data increases the number of possible correlations between tracks. The resulting percentage of incorrect and missed correlations was lower (improved) when EO sensor data was included in the tactical picture fusion process. The relationship between inclusion of EO sensor data and the percentage of incorrect and missed correlations was verified by increasing EO sensor bearing accuracy and observing increased incorrect and missed correlation percentages (as well as increased total number of possible correlations).

Future Work

EO sensor model enhancement

The EO sensor model described in this paper is to be further developed in the following aspects:

- Further investigation of EO sensor data fusion contribution; it is proposed that the fusion of EO sensor data with other sensor sources be investigated in pairs of sensors, e.g. radar and EO or ESM and EO,

- Full investigation of EO sensor data fusion; it is proposed that the contribution of EO sensor data to the tactical picture be investigated using CMISE and the sensor models in OOPSDG,

- Extension of EO sensor model; imaging sensors, such as EO, offer improved situation awareness to Command as a result of Command being able to actually see targets (captured in an image). The potential benefits to Command includes accurate target classification and identification, target behaviour and intentions assessment and battle damage assessment. Measurements of this type are most accurate for extended targets in an image. It is proposed to enhance the current EO sensor model to include sensor functionality for extended targets.

- Modelling of different EO sensors; the current model simulates the output of an IRST sensor. Other EO sensors have a wider field of regard, detect target signals in different wavelengths and measure target range offering Command greater situational awareness [1]. The EO sensor model is to be extended to incorporate features of other EO sensors to investigate the potential contributions to the tactical picture and situational awareness.

- Extension of target database; the current target database is limited by the availability of measured target signal data. The current target database contains IR signature data for thirty target types. The target signal database will be extended in the availability of measured target EO signal data.

Summary

The automated production of the tactical picture in the CMISE test bed using sensors and collateral data has been outlined. The current lack of EO sensor data in the system has been identified and is being addressed by the work described.

The development of an EO sensor model as a track input to CMISE is reported. The current EO sensor model is based on the possible IRST system fit on future RN platforms. Limitations of the model have been discussed and future work to enhance the model has been described.

References

1. Miles, J.A.H. & Metcalfe, G. 'Picture Quality Control and Measurement' presented at 'Defence System and Equipments International' DSEi 99, Chertsey UK, September 1999.

This page has been deliberately left blank

————————————

Page intentionnellement blanche

Convoy Planning in a Digitized Battlespace

S. A. Harrison
Pattern and Information Processing
Defence Evaluation and Research Agency (DERA)
St. Andrews Road
Great Malvern
Worcestershire WR14 3PS
UK

tel.: +44 (0)1684 895686
fax.: +44 (0)1684 894384
email: SAHARRISON@DERA.GOV.UK

Summary: In this paper we present a formal specification of a convoy planning problem in terms of a time-space network. We apply advanced heuristic techniques to this model and evaluate the approach on a number of realistic scenarios based on the UK MoD's Scenario Advisory Group (SAG) settings. The results demonstrate that the method described is an effective approach for solving practical instances of convoy planning. We also describe an automated planning tool that has been developed, based on the techniques described in this paper and which has been used to plan simulated movements of realistic size. The tool runs on a laptop, is fast and reduces planning time from man-hours to a few seconds.

The value of the techniques described in this paper is not limited to this one application. Hence, we review a representative set of military applications where we expect these techniques to be equally beneficial.

1   Introduction

Moving men and materials in large numbers and quantities is a long-standing military problem faced by all arms. For land forces in particular, present-day military engagements emphasize the need for mobility more than ever. Thus, routing convoys so that they reach their correct destinations in the shortest time is important. But the planning task itself can be considerable, and must be carried out quickly if the tempo of operations is to be maintained. With this requirement in mind, we have examined the development of a planning tool to assist in the strategic routing of objects between specific origin-destination pairs, taking into account the sorts of restrictions that are likely to be met in practice. The planning tool is suitable for running on a laptop computer and is based on an optimisation formulation.

There are many real-world applications for which a similar approach to the one described in this paper could be adopted directly - for example, moving forces into a remote operational theatre, strategic-level routing of hazardous materials through a given route network, or the routing and scheduling of trains over a rail network.

Not surprisingly, there is an optimisation problem at the heart of many military decision processes. A detailed discussion of the construction of optimisation models from military applications, and methods used to solve them, are given in [15].

1.1   Document overview

In this paper, we develop a model for the convoy routing, based on an optimisation formulation that exploits the concept of a time-space network. We apply a Lagrangian relaxation to this model and show that the resulting Lagrangian dual function may be evaluated efficiently using an enhanced version of Dijkstra's shortest path algorithm that is applicable to very large, implicitly-defined graphs - see [10, 11, 12].

The remainder of the paper is structured as follows. In the next section, we provide some background and in section 3 we outline the benefits of an optimisation based approach to military planning. Then in section 4 we present a formal specification of the convoy routing problem. In section 5, we describe the time-

space network model for the convoy routing problem and in section 6, we describe a Lagrangian relaxation of the time-space model and discuss how the algorithm is implemented. In section 7 we describe a planning tool based on the techniques described in this paper and its application to some realistic scenarios. Section 8 discusses the effect of uncertainties, such as on route delays and third party disruptions, on the implementation of the plans obtained and how we can account for these uncertainties in the planning process. A number of other applications where we expect the methods described in this report to be beneficial are discussed in section 9. We finish with conclusions.

For a detailed technical discussion of the methods described in this paper the interested reader is referred to [1].

2   Background

In this paper we consider a strategic network routing problem which is applicable to situations where certain objects, or commodities, are to be shipped, or transported, between specified origin-destination pairs with restrictions on how objects may encounter each other en route. The problem we consider is motivated by, and is presented in terms of, an application in which the objects are military convoys. For this reason, we have christened our problem the **convoy movement problem**, or **CMP** for short. There are many real-world applications for which a similar approach could be adopted. For example, the deployment of a force into a remote operational theatre; strategic level routing of hazardous materials through a given route network [7]; or the routing and scheduling of trains over a rail network [2, 3, 8] could easily give rise to instances of the CMP.

Modern military doctrine places great emphasis on the generation of a fast level of operational tempo. The movement of one's own forces into their correct locations is regarded as a key function, the planning of which needs to be completed quickly if the tempo of operations is to be maintained. In the recent action in the Balkans, for example, an enormous amount of time will have been spent on detailed planning for the movement of convoys in Kosovo. More generally, such planning could be further complicated by the fact that the enemy will be attempting to disrupt one's own actions through the disruption or destruction of the road network. The aim of the work reported in this paper is to reduce the amount of time required for the planning process. It is intended to incorporate the algorithms resulting from this

research into the British Army Digitisation Programme. This programme aims to enhance the operational effectiveness of UK forces in joint and combined operations by using modern information technology to couple weapons, sensors, communications and information systems (CIS) across the battlespace and thus create an effective, robust, efficient and affordable federation of systems.

3   Military Benefits of Optimisation

At the heart of many aspects of the military decision process are constrained optimisation problems. Finding the optimal set of resources required to achieve a set of targets subject to restrictions (or constraints) on the possible alternatives is typical of the problems that military commanders are required to solve and that can be formulated as constrained optimisation problems.

Traditional optimisation approaches, such as branch-and-bound, rely upon linear methods that are unable to handle the full complexity of real-world problems effectively. Moreover, traditional methods do not allow for rapid re-planning or the investigation of "what if" scenarios in a timely manner.

In contrast, heuristic based optimisation techniques provide the capability for planning in real world applications with their associated complexities and uncertainties. Plans obtained from approaches based on heuristic based optimisation techniques are generally of high quality and rapidly obtained.

An approach based on formulating aspects of the decision making process as an optimisation problem and employing heuristics to obtain near optimal solutions rapidly leads to the development of automated decision making tools. The benefits of such automation include

- an increase in the speed of processing data,

- reduction in the workload and stress of staffs freeing them for other roles and functions;

- stable performance, of military staffs, with time as opposed to the unavoidable degradation of human performance with fatigue and stress; and

- the ability to cope with the highly complex scenarios associated with the modern battlespace with a greater level of accuracy and effectiveness.

All of which leads to an increases in the speed and tempo of operations, which in turn enables the commander to get inside the opposing forces decision making cycle.

4   Convoy Planning

In this section, we present a formal specification of the CMP. This is a slightly modified version of the specification given in Lee, McKeown and Rayward-Smith [9].

In a CMP, we are given a collection of military units (**convoys**). Each convoy consists of a collection of vehicles that must travel nose to tail in a pre-specified order with pairs of vehicles maintaining a spacing of between fifty and hundred metres. Associated with each convoy are an **origin** and a **destination** such that the convoy must move from the origin to the destination location across a limited route network. The objective is to find a set of paths (or a **movement**) such that the total movement cost for all of the convoys is minimised, where the movement cost is defined to be the summed completion times of all the convoys.

Different types of convoys are composed of different numbers and types of vehicles. Hence they may move at different speeds along the same parts of the network. Indeed, some types of convoy may not be able to use some parts of the network at all. Furthermore, a convoy of a given type may travel at a different speed when moving in one direction along part of the network from when moving in the opposite direction. For example, consider a convoy going up an incline and the same convoy going down the same incline. Further, a given type of convoy may be able to travel in one direction along some parts of the network, but not in the opposite direction. In this paper, we assume convoys are not allowed to stop en-route.

We define the route network in terms of a set of nodes, one per junction in the underlying road network, and a set of links. A link is defined between a pair of nodes if the corresponding junctions in the underlying road network are joined by a road that does not go through any other junctions en-route. Associated with each link and each convoy pairing is a cost. The cost denotes the number of time units required by the convoy to travel directly between the underlying junctions. The cost is either a positive, non-zero integer or ∞. If the cost is ∞ the convoy cannot travel along the underlying road. The cost, for some convoy, in one direction along a link is not necessarily equal to the cost, for the same convoy, in the opposite direction. An example route network is shown in Figure 1.

A **movement** consists of a set of paths across the route network, one per convoy. Paths followed by different convoys may use common parts of the route network but two convoys cannot occupy the same part of the route network at the same time. When two convoys attempt to use the same part of the route network simultaneously this is referred to as a **conflict**.

Figure 1: Example of a route network.

Associated with each convoy is a **time-window**. The time window is the time it takes for a convoy to pass through any point in the route network, although it can be interpreted as the time during which the convoy blocks a node in the route network. When a convoy is blocking a node no other convoy may enter the node. The time window represents the convoy's length.

In practice the time taken by a convoy to pass through a particular point in the route network will depend on the point and on the time when the point is reached. However, detailed knowledge of this nature for all points in the route network is unlikely to be available a priori, and in any case may be of a non-deterministic nature. Thus, for an approach to be of practical interest, it is essential to introduce a *simple* device that enables the user to adapt the convoy planning to varying circumstances. This is the purpose of the safety time-window, the value of which could be increased or decreased depending on the planner's confidence in the available data (intelligence). In particular, the size of the time-window will often be an over-estimate.

We also associate with each convoy an **earliest ready time**. This is the earliest time at which a convoy can start its movement and represents constraints imposed by earlier phases of an operation. The convoy does not have to start its movement at its earliest ready time. It can delay its movement as this will often allow the convoy to follow a quicker route whilst avoiding later conflicts. The **initial delay** on a particular convoy's movement is a variable whose value is to be determined during the planning process. A zero delay corresponds to a convoy starting its

motion at its earliest ready time. For simplicity we assume that a convoy's initial delay must be an integer multiple of some prescribed **waiting interval**. The waiting intervals may be of different durations for different origins and different convoys.

Hence, a movement consists of a set of paths and a set of initial delays, one path and one delay for each convoy. We refer to the pairing of a path and an initial delay as the **route**. The **completion time** of a particular convoy's route is then the convoy's earliest start time plus its initial delay plus the time it takes for the convoy to traverse its path plus the convoy's time-window. Including the convoy's time-window accounts for the time required to allow the entire body of the convoy to arrive at and enter the destination. We refer to the collection of all the routes as the **movement**. The **overall completion time** is then just the sum of the completion time for all the routes in the movement.

We also associate with each convoy a **finish time** (or deadline). A movement is said to be valid (or **feasible**) if there are no conflicts and each convoy's completion time is less than its finish time, that is, every convoy has met its deadline.

Therefore, the aim is to find a valid movement such that the overall completion time of the movement is **minimal** with respect to all the valid movements; that is, there is no valid movement with a shorter overall completion time.

5   Time-Space Networks

In this section, we present a model of the CMP in which we relax the constraints associated with conflict prevention. The constraints associated with conflict prevention are **complicating constraints**; that is, if these constraints are removed the resultant optimisation problem is relatively straightforward to solve.

We introduce the concept of a **time space network** in terms of a single convoy. The route, associated with the convoy, is represented on the vertical axis by distance along the route and time is indicated on the horizontal. The convoy's occupancy of the route, in time and space, is indicated by a skewed rectangle - as illustrated in Figure 2. We refer to the skewed rectangle as the convoy's **time space occupancy**.

As the source and destination points of the route cannot be changed, nor in this simple example can the route, then the only freedom available to schedule the convoy is the initial delay.

Changing the initial delay corresponds to sliding the convoy's time space occupancy horizontally between the vertical lines that indicate the earliest ready time and the finish time. Clearly for the route to be valid with respect to the earliest ready time and the finish time the convoy's time space occupancy must lie entirely between these two limits.
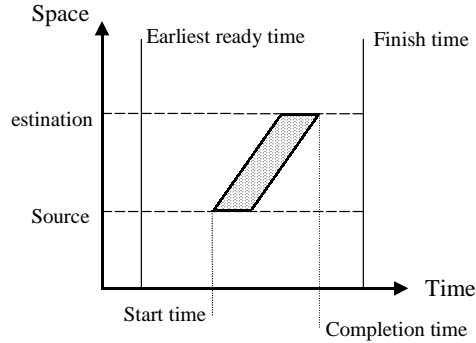
Figure 2: The time-space network.

The **time-space network** formulation is extended to multiple convoys and a route network as follows. First we define a **time frame**, for example, from the minimum earliest ready time to the maximum finish time over all the convoys, and we discretise the interval in some manner.

For each node in the route network and each time step in the discretised time frame we define a time-expanded copy of the node. For each link in the route network and each time-expanded copy of the head node we define a time-expanded copy of the link. We also add in additional links out of each origin node to model initial delays. We refer to resulting set of time-expanded nodes and links as the time-space network (associated with an instance of CMP).
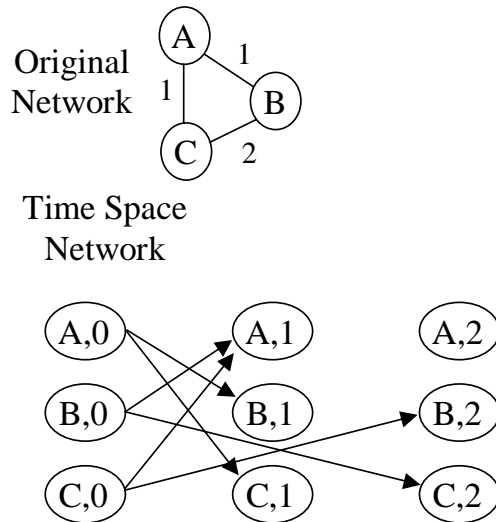
Figure 3: A time-space network

An illustration of a simple route network and an associated time space network is shown in Figure 3. The time space network is only expanded over the first few time steps to avoid an overly complex figure. The same structure of inter-connections will be repeated throughout the entire time space network unless links in the underlying network have time dependent costs.

It is easily seen that a path across the time space network from an origin node at the start of the time frame corresponds to a route. While it would be convenient if a set of disjoint paths on the time-space network defined a valid movement this is not quite the case. However, within the implementation we are able to straightforwardly generate sets of paths across the time-space network that correspond to valid movements.

The size of the resultant time-space network is a function of the number of convoys; the size of the route network; the size of the time frame and the granularity at which the planning is done. For realistic problems with hundreds (or thousands) of convoys; thousands of nodes; time frames stretching over tens of hours and movements planned down to the granularity of minutes the time-space network can easily become very large. Hence it is often impractical to store the network explicitly. In order to handles these very large time space networks in practice, the time-space network must be stored implicitly requiring an efficient implementation with sophisticated memory management techniques.

6   Relaxation Based Optimisation

As we have already stated if we were able to ignore the complicating constraints, that is, those associated with conflict prevention, the problem could be solved straightforwardly, by repeated application of Dijkstra's shortest path algorithm. However, we cannot ignore these constraints. Instead we can relax the constraints through a Lagrangian relaxation formulation of the problem.

Since the seminal work of Held and Karp [6], **Lagrangian relaxation** has enjoyed considerable success for solving combinatorial optimisation problems with large numbers of constraints. This heuristic techniques provides a framework for handling constraints whose presence complicates a mathematical programme the solution of which would otherwise be fairly straightforward. The essential idea of the approach is to price out the complicating constraints, in a systematic manner, using **Lagrange multipliers**. In this way, a Lagrangian **dual problem** is defined

corresponding to a given optimisation problem (referred to as the **primal problem**). Assuming the primal problem is a minimisation problem, to solve the Lagrangian dual problem we must maximise the corresponding Lagrangian dual function. Typically, this is done using a **subgradient optimisation** procedure [14].

The maximised dual function provides a **lower bound** on the minimal solution to the primal problem, that is, a value that is guaranteed to be no greater than the minimal solution's value. In general, the lower bound is close to the minimal value in which case it is said to be **tight**. Given the dual solution, corresponding to the maximised dual function, we can use the dual solution to heuristically construct a solution to the primal problem. The constructed solution is usually of high quality, where high quality corresponds to a small objective function value. If we compare the quality of the primal solution with the lower bound we have a bound on the quality of the primal solution with respect to the minimal solution. Often the quality of the resultant primal solution is found to be within a few percent of the minimal.

In the context of the CMP the complicating constraints are those associated with the conflict prevention. If we were to relax these constraints and allow conflicts the resulting problem could be solved by the repeated application of Dijkstra's shortest path algorithm to generate the shortest path for each convoy between origin and destination. We could then route each convoy along its shortest path starting at its earliest ready time. Clearly the overall completion time for this movement is a lower bound as no faster movement can exist. In general, this movement is not valid and the lower bound is not very tight. We refer to this as the *shortest paths* movement.

We define our Lagrangian dual function by pricing out the constraints associated with conflict prevention. To illustrate how we are able to price out the constraints associated with conflict prevention consider Figure 4.
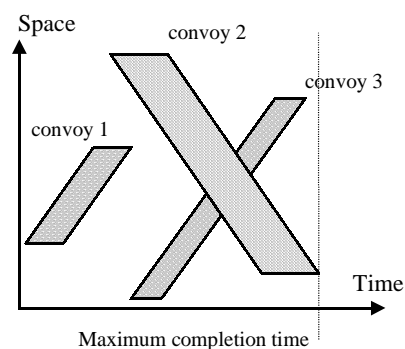
Figure 4: De-conflicting multiple convoys.

For the purpose of illustration we shall consider three convoys being moved along a single route. The overlapping of the time space occupancies associated with convoys 2 and 3 clearly indicates that these two convoys conflict en route. In this simple case, with a single route and a small number of convoys, the conflict can easily resolved by sliding the time space occupancies of the convoys until there is no overlap. However, for realistic scenarios where there are large numbers of convoys and multiple routes such an approach is no longer practical. In fact, representing such situations by a simple graphic, as in Figure 4, is generally no longer an option.

The graphical approach illustrated in Figure 4 can, however, be generalised and this leads to the approach described in this paper.

We refer to the area of the overlap of the time space occupancies of conflicting convoys as the **size of the conflict**. De-conflicting two convoys corresponds to reducing the size of the conflict to zero. Hence instead of attempting to obtain a valid movement we relax the conflict prevention constraints by replacing them with an objective of minimising the size of any conflicts. Clearly optimal solutions with respect to this new objective will be valid movements. However, we wish to simultaneously minimise the overall completion time and the size of any conflicts. Hence we must optimise with respect to some function of the two objectives.

Whilst it remains difficult to find solutions that are optimal with respect to this relaxed problem, the relaxed problem has a particular structure that means it is relatively straightforward to find near optimal solutions to the relaxed problem. The relaxed problem is formulated in such a manner that the overall completion time of the relaxed problem is a **lower bound** of the minimal overall completion time of the original problem. In other words, the overall completion time of the relaxed problem is guaranteed to be no greater than the minimal overall completion time of the original problem.

The near optimal solutions to the relaxed correspond to movements with small numbers of conflicts that can easily be resolved by simple heuristics. The resulting valid movements will tend to have near minimal overall completion times.

In particular, we define our Lagrangian dual function, with respect to some Lagrange multipliers as the overall completion time of the *shortest paths* movement over a time-space network with a modified cost function minus penalty terms for the any conflicts in the *shortest paths* movement.

Each node in the time-space network has a Lagrange multiplier associated with it and the modifications to the cost function of the time-space network are in proportion to the magnitude of associated Lagrange multipliers. To be specific for each edge in the time-space network we add to its original cost the average value of the Lagrange multipliers associated with the incident nodes. The *shortest paths* movement is then generated over the cost-modified time-space network. A **penalty term** is defined, for each node in the space-time network, as the product of the node's Lagrangian multiplier and the total magnitude of conflicts at the node in the *shortest paths* movement obtained for the given Lagrange multipliers. Clearly if there are no conflicts the penalty term is zero.

The resultant Lagrangian dual function can be shown to always be a lower bound. Hence, the Lagrangian dual problem is then to find a set of Lagrange multipliers that maximise the value of the Lagrangian dual function, that is, the lower bound. The greater the lower bound is the tighter it is. The multipliers are updated by a subgradient optimisation procedure that is guaranteed, at each iteration, to move closer to the Lagrange multipliers corresponding to the maximal value of the dual function.

By maximising the Lagrangian dual function we are minimising the penalty terms. Hence the magnitude of the conflicts in the corresponding *shortest paths* movement are minimised.

Employing a modified version of Dijkstra's shortest path algorithm accelerates the calculation of the Lagrangian dual function.

6.1 Benefits of relaxation methods

Relaxation methods, as described, coupled with other heuristic provide not only a high quality and de-conflicted movement but they also provide a measure of its quality with respect to the optimal movement, via the lower bound. This provides added value to the movement produced in that there is a measure of how much scope there is for further improvement in the planned movement. Hence, the commander is provided with a degree of confidence in the planned movement.

The combination of rapid planning and a measure of confidence is a feature that is not generally shared by the majority of automated decision aids. Moreover, this is combined approach is equally applicable to a wide range of military applications.

7 The Planning Tool

A prototype planning tool has been developed based on the techniques described above and implemented on a standard PC platform. The planning tool provides the user with the capability to develop optimised movements for various scenarios and to visualise the solutions.

The planning tool allows the user to perform three main functions:

- to iteratively generate a sequence of valid movements of increasing quality and to monitor the progress of this improvement measured in terms of their convergence towards the optimal;

- to stop the technique on any iteration and view the movements of all or any subset of the convoys obtained, and

- to configure the software to tune the algorithms so that convergence is as rapid as possible.

7.1 Planning tool capability

The planning tool offers the capability of

- rapid and automated planning;

- the investigation of "what if" scenarios; and

- rapid re-planning.

7.2 Planning tool functionality

The planning tool is PC based; is able to run on a standard laptop; and provides

- a graphical interface to the planning tool; and

- a graphical tool for visualising and interrogating the obtained movements.

In order to plan a movement it is assumed that the user has available a vectorised description of the route network as well as a data file describing each of the convoys; their objectives and constraints in terms of their origin, earliest ready time, destination and any deadline. These data files form the input to the planning tool. It is assumed that these data files have been obtained from other tools. Given the necessary data files the user is able to load the scenario and begin the planning algorithm.

Generally the first iteration of the planning algorithm will result in a valid movement which is within 10% (or better) of the optimal.

Once a valid movement has been obtained the planning tool provides the user with the ability to *playback* the generated movements on a graphical display. In *playback* convoys are represented as "worms" which progress over a representation of the route network.

The planning tool provides the user with the ability to playback

- movements in step mode or play mode;

- movements forwards and backwards via a slide control; and

- subsets of movements.

All functionality is available via menus, dialog boxes and easy to use controls. A detailed description of the functionality of the planning tool can be found in [13].

7.3 Results obtained

The planning tool was evaluated on a number of realistic scenarios based on the UK MoD's Scenario Advisory Group (SAG) settings.

Data sets	No. of nodes	No. of links	No. of convoys
P1	160	212	17
P2	530	724	25
Q1	932	2,482	166
Q2	1,145	3,058	333
Q3	7,232	15,496	1,817

Table 1: The test data sets.

The data sets used are summarised in Table 1. For all data sets, bar Q3, the planning tool is able to obtain movements that are within a few percent of optimal, or better, in the order of a few seconds to a few tens of seconds. In the case of data set Q3 tens of minutes are required to obtain a valid movement that is within a few percent of optimal. However, bearing in mind the size of data set Q3 even this performance is remarkable and of significant operational benefit.

7.4 Benefits of the planning tool

Such a planning tool, based on a combination of relaxation and heuristic based optimisation methods, reduces the effort required to plan movements from the order of man weeks to minutes of computing time. The speed of the planning tool also allows time for rapid replanning and the investigation of a number of "what if" scenarios.

Such a tool will enable military staffs to complete their movement plans far quicker than they can at present. This would lead to an

increase in the overall speed and tempo of operations. Such a tool would also enable staffs to quickly amend their plans if required. Once again this would serve to maintain a high operational tempo.

Our aim is to incorporate this planning tool into the British Army Battlespace Digitization Programme. The benefits of such a planning tool include

- automating the planning of large movements of military convoys, to be specific the automation of movement de-confliction; and

- the ability to visualise and interrogate the movements generated.

The automation of the planning process will result in benefits to the commander and his staffs of

- an increase in the speed of processing data;

- a reduction in the workload and stress of staffs freeing them for other roles and functions;

- the ability to cope with the highly complex scenarios associated with the modern battlespace with a greater level of accuracy and effectiveness; and

- the capability to rapidly re-plan when the original movement has been disrupted by third party action.

All of which leads to an increases in the speed and tempo of operations, which in turn enables the commander to get inside the opposing forces decision making cycle.

8   Planning Under Uncertainty

Optimal movements by definition attempt to minimise the planned delays on route. Such an approach will result in movements where once a convoy has cleared a junction the plan will specify that the next convoy will go through immediately. Of course such a plan leaves no margin for error. Simple experiments demonstrate that by introducing only small delays into an optimised movement rapidly leads to gridlock.

Therefore what we require are movements that are both optimised and robust. A movement is **robust** if it minimises the cost of disruption and re-planning if things do not go to plan. Simple mechanisms for producing robust movements are to introduce minimum inter-convoy spacings

or to artificially extend the convoy's time window in front and behind the convoy and route these extended virtual convoys. In the latter case this means that the true convoy can lie at any point within the virtual convoy during the plan's execution without disrupting the movements of other convoys. Inter-convoy spacings are already specified by many NATO nations when planning the movements of their convoys.

However, current research is investigating whether more sophisticated delay models and methods for planning in the presence of uncertainty will result in better plans in terms of their quality and robustness.

9   Other Applications

It has already been stated that constrained optimisation problems lie at the heart of many aspects of military decision making. Moreover the combination of techniques described in this paper are sufficiently generic that they can be applied to many of the constrained optimisation problems encountered in military decision making.

Therefore in this section we briefly review four further applications, of military relevance, where these techniques are either applicable or have already been applied by the Pattern and Information Processing group at DERA Malvern. For a further discussion of the construction of optimisation models from military applications, and methods used to solve them, the interested reader is referred to [15].

9.1   Depot outloading

One of the fundamental logistics roles of an armed force is the movement of men and equipments between storage locations and (sea and air) ports of embarkation coupled with the movement of men and equipments between ports of disembarkation and the operational theatre. Such movements are generally very large and require large amounts of planning in order to ensure all men and equipments are in the correct location at the correct time and that the appropriate logistical infrastructure is also available in the correct location at the correct time. We refer to this as **depot outloading**.

Not surprisingly depot outloading is a constrained optimisation problem. There are two objectives that can be considered in the depot outloading problem: either

- Outload a given number of men and equipments as quickly as possible; or

- For a given a deadline, outload as many men and equipments as possible by the deadline?

9.2 Earth observing satellites

Observation satellites have a limited window of opportunity for imaging or taking measurements of a given target area, dictated by orbit considerations. In low Earth orbit (typically 400 to 1200 km altitude) a ground object will be in view for a few minutes at most. Depending on the mode of operation of the satellite, the data taking window may range from a few minutes down to only a fraction of a second.

The scheduling problem then is one of achieving the maximum efficiency of use of a satellite. Traditionally this has been achieved by teams of ground planners who write, check and recheck procedures.

An initial feasibility study [4,5] has already demonstrated, for a simplified model of the satellite and its environment, that optimised taskings can be obtained in the lead times encountered in practice. A demonstration graphical planning tool has been developed.

9.3 Frequency assignment

Good communications are at the heart of successful military operations. The scope of modern radio communications technology is so attractive that ever more data is being exchanged on the battlefield, further down the command structure. But there is an overriding difficulty – the electromagnetic spectrum is physically limited, so congestion results. Bandwidth is becoming every bit as precious as fuel or ammunition.

Not surprisingly, the assignment of radio frequencies to transmitters under the constraints of bandwidth and interference described in the previous paragraph is a constrained optimisation problem with applications for cellular networks for mobile phones as well as the assignment of frequencies to networks of mobile transmitters in military operations. The application of heuristic based optimisation techniques to the assignment of radio frequencies is described in [16].

9.4 Collection management

Collection management, in the land domain, hinges on the ability of the military planner to satisfy a wide variety of requests for information (RFIs) each with a priority value associated to it. These RFIs must be satisfied by making an efficient utilisation of one or more of a varied array of assets. These assets include surveillance sources, such as ISTAR, satellites and various agencies, along with a potentially large array of reconnaissance sources, such as manned and unmanned aircraft, plus 'spot' modes on some surveillance sources. Adding to the complexity of the planning problem is the fact that not all the collection assets are suitable for dealing with all RFIs and some RFIs may require more than one asset to satisfy. Furthermore the RFIs come with time windows during which they are to be answered.

This all leads to a complex and highly constrained scheduling problem, which is discussed in greater detail in [17]. Whilst the constrained optimisation problem at the heart of collection management is extremely challenging, the techniques described in this paper offer the potential to automate many aspects of collection management and to further reduce the workload and stresses imposed on the military commander and their staffs.

10 Conclusions

The principal message of this paper is that the combination of relaxation and heuristic based optimisation techniques provide a powerful set of tools for the automation of many aspects of the military decision process. Thus reducing the workload and stress of military staffs releasing them to concentrate or more strategic level planning.

A planning tool based upon or incorporating optimisation methods has been shown to reduce the effort required for planning, in the case of convoy planning the effort was reduced from the order of man weeks to minutes of computing time on a laptop.

An operational planning tool based on the concepts described in this paper would provide the commander and his staffs with the ability to provide high level outlines of a planned movement and leave the detailed planning to the tool. The benefits of automating the detailed planning in this manner include

- a reduced workload;

- reduced stress;

- supporting an increase in the speed and tempo of operations;

- the rapid generation of alternative plans; and

- the generation of new plans when the original plan has been disrupted by enemy action.

The speed of the planning tool also provides the commander and his staffs with the opportunity to explore a number of "what-if" scenarios. This is clearly not an option when the planning of a movement requires a significant effort on the part of the commander and his staffs.

11   References

1. P. Chardaire, G. P. McKeown, S. A. Harrison and S. B. Richardson. Solving a Time-Space Formulation for the Convoy Movement Problem. *Operations Research* (submitted 1999).

2. K. Chih, M. P. Bodden, M. A. Hornung and A. L. Kornhauser. Routing and Inventory Logistics System: A Heuristic Model for Optimally Managing Intermodal Double-Stack Trains. *Journal of Transportation Research Forum* **31**:56 - 62, 1990.

3. M. Florian, G. Bushell and J. Ferland. The Engine Scheduling Problem in a Rail Network. *INFOR* **14**:121 - 138, 1976.

4. S. A. Harrison. *Task Scheduling for Satellite Based Imagery - An Initial Feasibility Study*. DERA Report DERA/S&P/SPI/TR990673, February 2000.

5. S. A. Harrison, M. E. Price and M. S. Philpott. Task Scheduling for Satellite Based Imagery. *Proceedings of the 18th Workshop of the UK Planning and Scheduling Special Interest Group* (PLANSIG-99), pages 64 - 78, University of Salford, UK, December 1999.

6. M. Held and R. M. Karp. The Travelling Salesman Problem and Minimum Spanning Trees. *Operations Research* **18**:1138 - 1162, 1970.

7. E. Iakovou, C. Douligeris, H. Li and L. Yudhbir. A Maritime Global Route Planning Model for Hazardous Materials. *Transportation Science* **33**(1):34 - 48, 1999.

8. O. K. Kwon, C. D. Martland and J. M. Sussman. Routing and Scheduling Temporal and Heterogeneous Freightcar Traffic on Rail Networks. *Transportation Research, Part E: Logistics and Transportation Review* **34**(2):101 - 115, 1998.

9. Y. N. Lee, G. P. McKeown and V. J. Rayward-Smith. The Convoy Movement Problem With Initial Delays, *in* V. J. Rayward-Smith, C. Reeves and G. D. Smith (editors) *Modern Heuristic Search Methods*, John Wiley, pp. 215 - 236, 1996.

10. A. Martelli. An Application of Heuristic Search Methods to Edge and Contour Detection. *CACM* **19**:73 - 83, 1976.

11. A. Martelli. On the Complexity of Admissible Search Algorithms. *Artificial Intelligence* **8**:1 - 13, 1977.

12. N. Nilsson. Problem *Solving Methods in Artificial Intelligence*, McGraw Hill, 1971.

13. S. B. Richardson. *User guide for the BSS convoy movement optimisation tool*. DERA Report DERA/LS(LSB2)/BSS(SCEN)/CMP/1. March 1999.

14. N. Z. Shor. Generalisations of Gradient Descent Methods for Nonsmooth Functions and their Applications to Mathematical Programming. *Econ. Math. Methods* **12**:337 - 356, 1976.

15. C. L. West. *Combinatorial Algorithms for Military Applications (CALMA)*. DERA Report DRA/CIS(SE1)/608/08/07/Final_1, DERA, 1996.

16. M. L. Williams. Making best use of the airways - an important requirement for military communications. *Electronics and Communications Engineering Journal* (to appear) 2000.

17. M. L. Williams. *Scheduling for Collection Management*. DERA Report DERA/S&P/SPI/TR000278/1.0

18. The Studies Assumptions Group: Scenarios and Assumptions for Studies, Interim SDR version issue 3, D/DFD/10/2/3.

An Information Filtering and Control System
To Improve the Decision Making Process
Within Future Command Information Centres

(January 2001)

Hans L.M.M. Maas, Sikke Jan Wynia
TNO - Physics and Electronics Laboratory
Maritime Command and Control
P.O. Box 96864, 2509 JG The Hague, The Netherlands

Dr. Morten Heine Sørensen
TERMA Elektronik AS
Info Systems Division
Bregnerodvej 144, DK-3460 Birkerod, Denmark

Dr. Maurice A.W. Houtsma
Hollandse Signaalapparaten B.V.
ASR
P.O. Box 42, 7550 GD Hengelo (Ov.), The Netherlands

Summary

This paper describes the achieved research results within several national and international C2 and information-management projects to develop concepts for balancing the information push with an operator's information need in order to meet the requirement to avoid / suppress information overload situations. The paper starts with an analysis and syntheses of the information overload problem. A model is used to describe the causes and the consequences of information overload on the operator's behaviour and performance in a command information centre of naval vessels. Research has shown that an increasing amount of time is needed for gathering and discriminating relevant information from the actual information push while less time is left for analysing the relevant information in more details and taking correct and original decisions. Information overload is seen as a serious threat for the quality and performance of mission execution. The blueprint for an adaptive information management support concept is based on merging several information management support approaches:

1. Approaches to estimate and/or measure and control the operator's information overload.

2. Information exchange concepts.

3. Information handling within several kind of tasks: Skill based, rule-based and knowledge-based tasks.

Based on the complexity of the problem, an information management concept is discussed to control and filter the information flows adaptively for skill and rule dominated tasks.

1. Introduction

Future Naval Command and Control organisations are characterised by its small staff while at same time naval vessels will operate in an increasingly complex and information rich environment with a high time pressure. Information overload is seen as a serious threat for the performance of future operations.

Current naval command and control concepts are based on an evolutionary continuation of the C2 concepts that have been developed several decades ago. In the last years we have seen that life-cycle costs have become an important design constraint for new command and control concepts. The need to decrease the life-cycle costs will increase in the coming decades. Those costs are to a large degree determined by personnel and exploitation costs. It is not an exception that the personnel costs amount to 40% of the total life-cycle costs. Application of the rapidly developing information technology has until now not yet resulted in a substantial reduction of the number of staffing in the Command information Centre (CIC), although CIC staffing and automation vary significantly in different navies.

A reduction of the number of CIC staffing implies at the same time a reduction of the human decision making capability which has to be counterbalanced by an increased capability of the supporting information management system. This factor especially calls for intelligent (support) systems that incorporate skills, knowledge and experience with naval warfare to substitute for the reduced number of operators. Some navies are addressing the issue of CIC staffing by research programmes on Reduced Manning Concepts for future command and control organisations. Examples of such research programmes are 'Smart Ship', 'SC-21' and the Dutch research programme on Future Naval Command and Control Concepts ([Maas and Keus, 1999], [Scott, 2000]). The increased mission complexity and the demand for personnel reduction require innovative solutions and choices in both the automation of processes as well as the management of the information flow in future C2 organisations.

This paper describes the analysis and the design of an advanced information management system to prevent / suppress information overload situations in future C2

organisations. The paper starts in chapter 2 with a description of a model that shows the causes and the consequences of information overload in the operator's behaviour and performance from an abstract point of view. The paper continues in chapter 3.2 with a projection of the information overload factors and characteristics on the human organisation in the command information centre of naval vessels. Interviews with operational Naval Officers have been used to analyse the model during military operations. Based on this analysis an overview of user requirements is described in chapter 3.3 for an advanced information management support system. A set of high-level system functions is analysed separately in chapter 4 and is translated to architecture of an information management support system. The paper concludes with a summary with respect to filtering and controlling the information flow in future command and control organisations.

2. Information load model

2.1 Introduction

This chapter describes the military implications of a model [Schneider, 1987] of the causes and the consequences of an information overload on the operator's behaviour and performance. This model has been used for criticising the information management concepts on their capabilities to avoid or to suppress information overload situations.

2.2 Description of the information overload process

This section discusses a simplified model of the causes and consequences of information overload in a command and control organisation (Figure 1). For easy understanding of the model we present the following definitions:

- Organisation: The combination of the computer systems, their users and the way they work together.

- User: Any person within the organisation (from operator up to commander), unless specifically indicated otherwise.

- Information overload: The condition at which the information processing requirements exceed the information processing mechanisms available, so that the organisation is unable to process the information adequately [Tushman and Nadler, 1978]. In short: the information-processing requirement is larger than the information processing capacity.

Figure 1 shows that two categories of aspects influence the sensitivity for information overload:

1. Organisational condition aspects.

2. Information overload factors and characteristics.
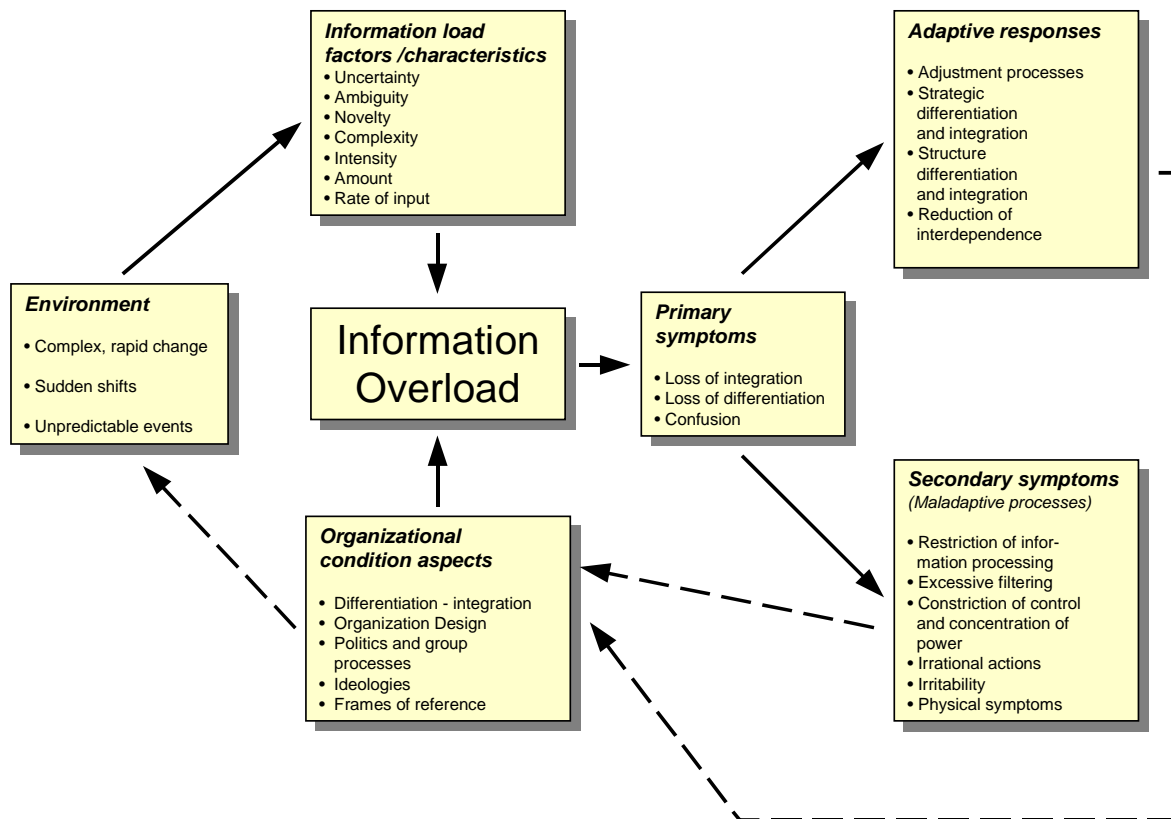
Figure 1. Information overload model [Schneider, 1987]

This first aspect is related to the Command and control organisation. Several organisational conditions influence the sensitivity of the organisation for information overload situations:

Organisation design: the organisation design affects basically how information is processed. For instance, a functionally designed organisation may restrict or distort information flow as it reaches higher levels which may even result in the failure of important information reaching the top of the organisation. A functional organisation requires information to be integrated at the higher levels. An insufficient integration will require extra information processing capabilities at the higher hierarchical levels of the organisation. Within a matrix organisation, the integration of information is mainly carried out at the lower levels. However, more information is generated as the amount of interaction between users increase. In addition, ambiguity may increase, due to different perspectives of the users.

Levels of differentiation and integration: Differentiation ensures that the required specific information becomes available. Defining specialised roles and functions within the organisation that acts as divisional and functional boundaries take care of this. These boundaries act as barriers for the information flows. The lack of boundaries can cause defective filtering, which will result in a failure to screen out irrelevant information.

Integration ensures that the correlation among specific information is recognised and brought to a higher level. The structure of the organisation must take care that the information flows, that carry the information that to be correlated, come together at the right level. Often, gathering information is not the problem, but the integration of information is.

Politics and group processes: Depending on own perceived power, the status of the sender or receiving user, and the extend to which it is seen as furthering or impeding goal attainment the information can be distorted, modified or re-routed. Further, the users may frame certain issues in light of personal or group interests, at the expense of organisational interests.

Frames of reference: Information may be processed incorrectly, at the wrong level, or not al all, because of the characteristic styles of collecting, analysing, and verifying information that is used in an organisation. Ultimately, this may lead to too much information, or less (but too ambiguous) information.

The second aspect is related to the information overload factors / characteristics. Organisations may operate in complex and turbulent environments, characterised by sudden shifts, unpredictable events and complex interdependencies. Such environments may be perceived as uncertain or ambiguous. The properties are not inherent to the environment, but are attributed to the environment by the organisation. The information need of the organisation determines which of the properties of the environment are relevant for the organisation. This implies that the information that must be processed by the organisation can suddenly change. If these changes can be predicted (even if only on a higher level), the organisation and/or the underlying information management system can better adapt to these changes. The information needs are a function of the

characteristics of the functional mission and strategy of the organisation, the individual users, group processes and the before mentioned organisational structure and schemes. The sensitivity for information overload situations is determined by a set of factors, which characterises the information flow:

Uncertainty: uncertainty refers to the quantity of information required versus the available information. Insufficient information often results in further information requirements, which often will still not provide the required information, but even increase the information overload even further. Note that, for instance, even a single message may incorporate uncertainty. This can still be regarded as a case of insufficient information, because the message does not contain enough information to make it certain.

Ambiguity: Possible different interpretations of the same information and/or too much information are in conflict with each other. This is often due to interpretation of the same information, in varying contexts.

Novelty: Information is considered to be novel when it does not conform to the current awareness of the situation. Information that is sufficiently novel will attract attention. Information, which is excessively novel, may be ignored because it appears irrelevant or unrelated to the present context.

Complexity: Complexity of information refers to the specific aspects of the environment that can have an impact on the organisation and reflects the inter-relatedness of these aspects.

Intensity: Intensity refers to the increase in rate of information, and/or the importance of certain information. Increased arrival rates could reduce the required time to process information and make correct decisions, thus inducing (potential) information overload. This situation often invokes stress that can stimulate a sense of urgency or overwhelm an individual.

Amount of information: The amount of information is related to the number of meaningful items.

An increase in one or more of the information load factors / characteristics will increase the information processing requirement, and thus may cause an information overload situation. The organisational condition aspects exert a major influence in the information processing capacity.

When an information overload occurs, the organisation will (try to) cope with the overload. This leads to changes (usually restrictions) in the information processing capacity of the organisation. These changes can for example be caused by narrowing of attention, simplification of information codes, or reduction of information channels (in number or in capacity). These changes in information processing are described by the so-called Primary Symptoms. These symptoms can usually not, or with great difficulty, be measured.

The organisational response to information overload may lead to successful adaptation or it may create temporary or perhaps enduring dysfunction. While the Primary Symptoms cannot be measured directly, they express themselves through the so-called Secondary Symptoms and through the Adaptive Processes. The Secondary symptoms are an expression of the maladaptive attempts

to cope with information overload. The symptoms themselves have a negative influence on the information processing capacity of the organisation (this is the negative feedback loop). The organisational responses to information overload situation that lead to a successful adaptation by the organisation are identified as the Adaptive Processes. Like the Secondary symptoms, they are triggered by the Primary Symptoms. The adaptive processes are the cause of a positive feedback to the organisational conditions.

2.3   Observations

The following observations can be made based on the analysis of the information overload model:

- Not all information overload situations can be solved by information technology solutions. The organisation of the human organisation will always be a vital link in controlling the information load of the operator.

- The fact that Primary symptoms are not measurable makes it very difficult to use a robust information flow control concept in the feedback loop to control information-overload situations. In most cases, the user performs several tasks simultaneously. This means that it is very hard to identify the critical user task that is responsible for an information-overload situation even if we could measure the most relevant information factors and characteristics.

- Measuring the Secondary Symptoms is an option for controlling information overload situations. However, we have to be aware that it is an emergency brake for preventing an escalation of the information overload problem.

- The information technology can show benefits in providing support in gathering and selecting the relevant information from the total amount of provided information. This should guarantee that only relevant information would be presented to the user at the right time and in the right way. However, this doesn't mean that information overload situation will not occur in case of presenting only relevant information to the user. It is still possible that relevant information or tasks will overload the user.

3.   Problem analysis and synthesis

3.1   Introduction

The previous chapter showed that the organisation and the characteristics of the information flow influence the sensitivity for information overload situations. The paper continues with the military domain analysis of the information load factors and characteristics and discussing technical solutions in next chapter. This chapter discusses the information overload problem in the Command Information Centres in naval vessels during tactical operations. This analysis was carried out by means of interviews and the monitoring of a command team during training sessions. The interviews were carried out on two levels of details. A storyboard was used to discuss the occurrence of information overload situations in general terms while more detailed

information was gathered from discussions of the bottlenecks in the execution of specific operational tasks. This paper discusses not the details of the information management problems for specific operational tasks, but they are used to underline the findings of the top-down analysis (generic analysis). This means that the results are not used to start a development process for a decision support tool / system for a specific situation, but that they are used to identify the military requirements that steer the research in generic information management support concepts.

Section 3.2 analyses the occurrence of information management bottlenecks within current and future naval Command Information centres during the execution of operational tasks. A set of user requirements is discussed in section 3.3 that should prevent the command team in getting into information specific overload situations.

3.2   Military analysis of information overload situations

The military domain analysis shows that the information availability and demand in future operations will greatly exceed those of current operations. The number of available information sources and their bandwidths will provide much more information than the current C2 organisations can handle. Information exchange among the C4I systems and their operators puts pressure on both the capacities of the communication links and the capability of the operators to survey and digest the incoming information.

The more information operators receive, the more time they need to assess the value and the relevance of the information in relation to their mission. All this information is converging inside the command information centre, resulting most probably in an information overload for the staff. Information overload situations will pose a serious threat for the quality and performance of mission execution.

Information overload occurs mainly in the execution of rule- and knowledge-based tasks within the orientation and the decision-making phases of the OODA loop. Research shows that (a lack of) time is seen as the most important bottleneck in the execution of these tasks. Too much time is needed to gather the required information and removing the non-relevant information from the total amount of information that is presented to the operator. The remaining time is not enough for operators to use the available information in a proper way to take correct and original decisions for situation assessment and decision making tasks.

Analysis shows that, uncertainty, amount of information, ambiguity, novelty, level of abstraction, time constraints, presentation media and task overload are seen as the most relevant factors for information (over)load. The last four factors are not identified as information overload factors in the literature. Especially the last two are factors that are not inherent to information, but are more related to the operator's status.

- Uncertainty in information severely restricts the possibility of the operator to locate, identify and recognise threats, and to determine and/or predict the intention(s) of the threats. In case of uncertainty, an operator will try to gather more information, or to

correlate existing information, to reduce the uncertainty. Depending on the amount of information that must be gathered and /or correlated, this may cause information overload situations. Uncertain information will only cause stress (and consequently possible information overload), when the information is considered as important.

- The amount of information is related to the total number and/or the rate of information items, in the sense that both have the effect that they can overwhelm the operator with information to be processed, which can result in an information overload situation.

- Ambiguous information only causes stress when the context is not clear and/or that the information conflicts with other information to determine the potential means of a threat to the ship or to the mission. An operator will try to gather more information, or correlate existing information, to reduce the ambiguity. Note that this has to be done at the level of interpretation of intentions, unlike the work that needs to be done to reduce uncertainty, which is at the level of gathering more (sensor) information and/or intelligence reports.

- Novel information causes stress when the information does not conform to the current awareness (and expectation) of the situation, and when the information is digested very late (or even too late). The last situation occurs in situations where the information is enclosed in already known messages. At first, the operator sees no reason to assess the offered information on its novelty.

- The level of abstraction influences the amount of time that is required to process the information. In most of these cases the provided information should be processed to bring the information to the right level of abstraction (which requires extra time) and in some cases the operator isn't able to process the information because of insufficient knowledge to interpret the information.

- Time constraints are not a real factor, in the sense that they are parameters that are attached to incoming information. However, they do influence stress due to the fact that most tasks have hard time constraints. In most cases the operator spends a lot of time in gathering the required information and separating the relevant information from the non-relevant information, and time is also needed to control the decision support tools to assess the information.

- Workload plays an important role within information overload situations. The time and attention that is spent on task execution that has impact on the time and the attention that is required to assess the offered information. For instance, the information gathering process is seen as a labor-intensive process.

- The operator has to deal with different information flows at the same time. For instance, the operator is confronted with three categories of information flow: information that is displayed on the tactical screen, voice information that is provided by his headphone and the presentation of textual messages on paper and in his mail box.

3.3 Military requirements to avoid military information overload situations

The analysis shows that the causes of the information overload can be divided into two main categories:
1. Information gathering: Too much time is required to gather the required information and as consequence there is not enough time left for a correct and extensive assessment of the information.
2. Information assessment and processing: An operator has to execute different tasks at the same time, each of which have their own priorities and information needs. The ensuing problem is that at a given time the operator is fully focussed at the execution of a task and is unable to detect other relevant or threatening events in his environment.

The Information gathering requirements are aimed to deliver support for the collection and aggregation of information from the various information sources. This support should be capable of composing information need descriptions and search support in order to increase the pace of the information gathering process. An example is the requirement for the system to be able to extract relevant information from textual messages and combine this with information contained in various databases. The information should then be available to the operator when his task preparation or execution requires it, thus alleviating the operator from searching for the information himself.

The task execution requirements focus on the reduction of the sensory input to the operator; both visual and auditive sources are covered. An example is the requirement for the system to adapt the level of detail of displayed information with respect to the operator's tasks. This could be executed by intelligent information filtering and displaying techniques that are able to aggregate information to a higher or a lower level of abstraction and placing relevant information on the front and moving the non relevant information to the background of operator's attention. For instance, the contacts that are close together are aggregated into one tactical symbol on the tactical display, and non-threatening contacts are dimmed. The system should be able to generate alerts to the operator, when changes in the tactical situation require him to switch his attention to the new event.

The resulting information management system should therefore address the following topics. The support system should provide a task planning in order to enable the system to optimally present the operator with the correct amount of information at the correct time. A filtering module must be able to derive from the tactical situation and the tasks at hand and the present workload of the operator, how much and in what form the information must be presented to the operator. This leads to an estimation or prediction of the effect of the information reduction on the actual overload.

To conclude, the system should provide support in controlling the actual information flow to avoid information overload situations. Passing only relevant information to the operator is not a guarantee that the operator is not confronted with information overload

situations. The Information management support system should be able to filter the amount of information with respect to the operator's preferences related towards information need and display requirements and his present level of work and information load.

4. System concepts and architectures

4.1 Introduction

This chapter discusses several conceptual visions on information gathering, filtering and information control. The chapter concludes with a describtion of an adaptive information management architecture that should be able to control the information flow between the computer system and its user by means of balancing the user's information need and the available information in the C4I system.

4.2 Conceptual views on information gathering, filtering and control

The information management architecture is based on the analysis of several information management support concepts / approaches:

1. Concepts that estimate and/or measure and control the operator's information load (Section 4.2.1).
2. Information exchange concepts (Section 4.2.2)
3. The kind of tasks that should be supported by the information management support system (Section 4.2.3)

4.2.1 Controlling the information load

There are basically three approaches to cope with and controlling information overload situations:

1. Feedback control
2. Prediction / estimation of the required information and information load
3. A combination of estimating and measuring the information load.

 The first concept (Fig 2a) sounds more or less as an obvious solution. The concept assumes that is possible to measure the actual operator's information load. Further, the information management support system should know exactly what kind of information is relevant for the operator and what is not. However, section 2.2 shows that it is very hard to detect the occurrence of an information overload situation by the fact that it is very difficult to measure the primary symptoms of an information overload and to determine which part of the provided information is mainly responsible for the information overload. Further, the information management system should be aware of what kind of information could be removed from and/or added to the information flow to decrease the required information processing capacity of the operator without causing confusion at the side of the operator. The concept shows benefits in situations where the information parts of the information flow could be processed separately and where the available information could be prioritised by the system in advance.
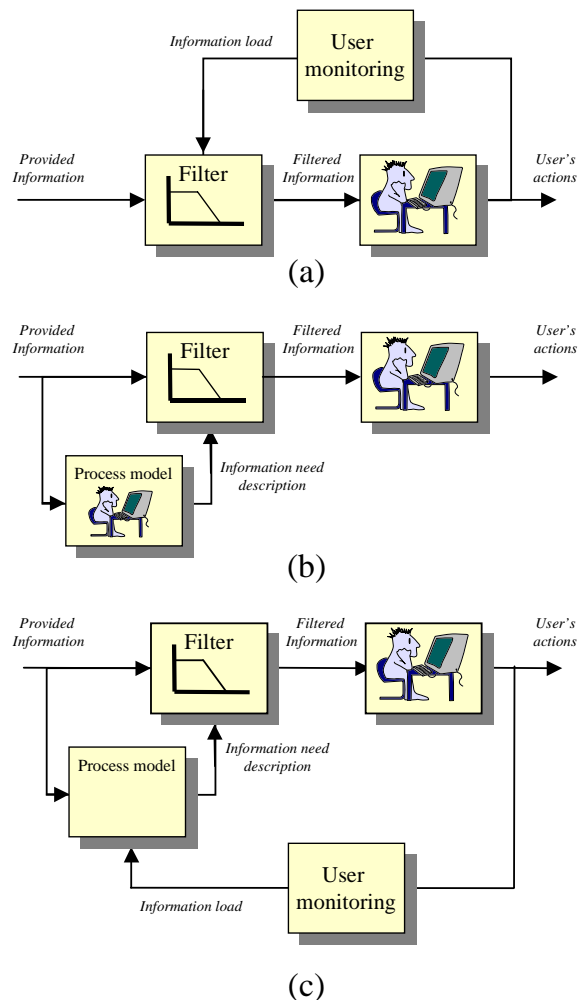
Figure 2. Information control concepts: (a) Feed-back, (b) Open loop and information load prediction control, (c) Combination of feed-back and prediction.

The second concept (Fig. 2b) estimates the required information processing capacity to process the provided information. The information management system should have a model of the processes presenting the tasks that have to be carried out by the operator and which information is needed to perform these tasks and how the information has to be presented to the operator. Basically, the model should contain information about the start conditions (stimuli) to execute a particular task. These conditions are mostly based on the contents of the provided information. For instance the detection of an approaching missile will result in the launch of a sequence of tasks that have to be carried out by the user. Each task will have its own priority and information need that should be made available to the operator while irrelevant information for these tasks should be moved to the background. The process model contains information about the sequence, the priorities and information need of each task, and should make an appropriate planning of which tasks could be carried out without overloading the user. The weakness of this concept is that it is very difficult to define all tasks and put the required details in the model. For instance, the estimation of required information processing capacity for a set of tasks that has be carried out simultaneously is not just a matter of adding the required information processing capacity of each task.

The third concept is a combination of the two above discussed concepts and addresses solutions for the weaknesses of these two concepts. The information filter will be controlled initially by the module that contains process models of the required information and task priorities of each task. A refinement in the control of the information flow is realised by the feedback loop. The concept ensures that the incompleteness and/or impreciseness of the process model can be adaptively refined or temporally modified by means of the feedback loop.

4.2.2 Information exchange concepts

The information exchange between the operator and the different information providers could be established by means of four different concepts (See Figure 3):
1. Report
2. Request
3. Server
4. Broker
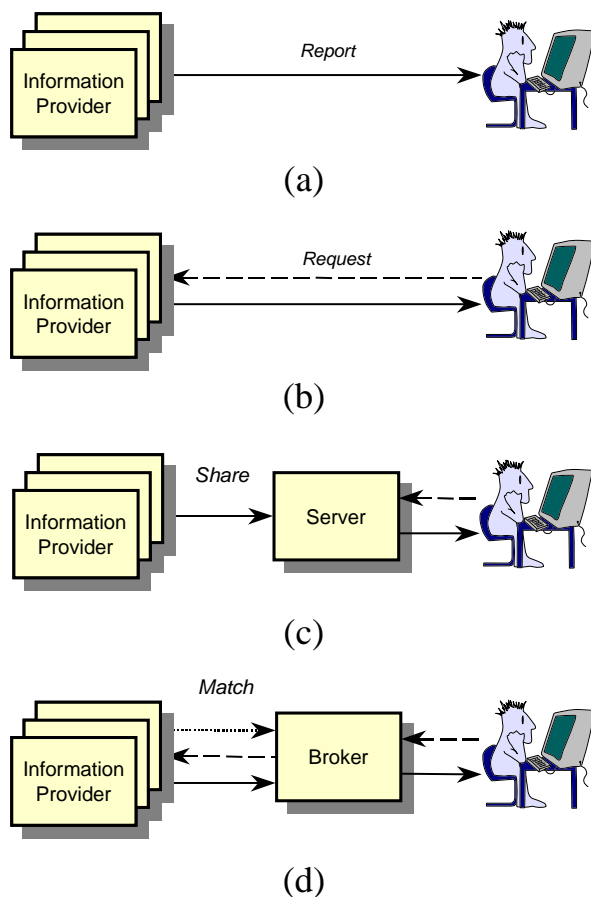
Figure 3. Information exchange concepts: (a) Report; (b) Request; (c) Server; (d) Broker

The first concept ('report'-concept) is the most plain information exchange concept of the four mentioned concepts. Within this concept, the information provider sends reports to the user at a fixed rate without knowing whether the distributed reports have enough relevance for the operator or not at the moment of distribution. This means that this concept is only applicable for

situations where the reported information has enough relevance to be pushed to the operator at all times.

The second concept ('request'-concept) differs from the first concept by the fact that the reports are distributed only to the operator in the event of an information request from the operator. The operator should be familiar with the information providers in order to put his information request or he should make his information request to the correct information provider, or he should make his information request public to all accessible information providers. The fact that the operator should know or should investigate what kind of information is available at each information provider makes it a time consuming process. Further, it is questionable whether the operator could reach the information provider to make his information request public all the time, and whether the information provider has the ability to provide the required information instantaneously. The concept shows benefits in cases of loose time constraints and where the operator doesn't have to spend much effort in locating the appropriate information provider.

The third concept ('server'-concept) shows similarities with the previous two concepts. The operator interacts with a local server like he should do with the information providers in the 'request'-concept. The difference is that the operator interacts with only one system (server). The server collects all information from the available information providers that might be used by the operator within a certain period of time. The fact that the operator is released from an external information search and gathering task is seen as a big advantage of the concept. However the disadvantage is that the communication links are not optimally used and could cause data overload situations on the communication links.

The fourth concept ('broker'-concept) meets the disadvantages of the 'server'-concept. The broker communicates with the different information providers whether they could deliver the required information with the required specifications and the information providers have the ability to pass a META-file of the available information to the broker. The information is passed to the broker in case of an agreement between the broker and the information provider. This concept is only applicable in cases where the broker is able to negotiate with the information providers on the information requirement specifications.

4.2.3 Determination of the required information need and support

Operators have to deal with different kind of tasks during the completion of the OODA-cycle (See Figure 4). Rasmussen [Rasmussen, 1983] distinguished three kind of behaviours in controlling the decision making process (See Figure 5): skill-/routine-, rule- and knowledge based behaviour. Using this distinction it is possible to gain better understanding of human errors in running complex processes in an information rich environment and reduce the likelihood of such errors with suitable information management support concepts.
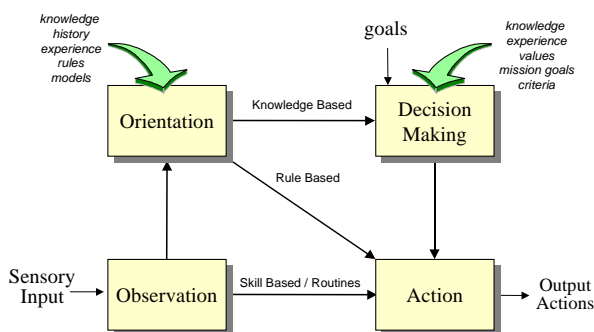
Figure 4. OODA cycle

Skill based behaviour includes expert sensorimotor performance which runs smoothly and efficiently without conscious attention. A dynamic mental model that depicts operator's movements and environment in real time controls this behaviour, and enables the operator to adjust rapidly to feedback from his actions. The tasks that require skill behaviour could be automated easily in most of the cases. However, several tasks require involvement of the operator for confirmation. The information management support system could support the operator by gathering and presenting the required information for getting confirmations from the operator. The information management support system should make it able to perform the skill based tasks more efficiently and rapidly than before.
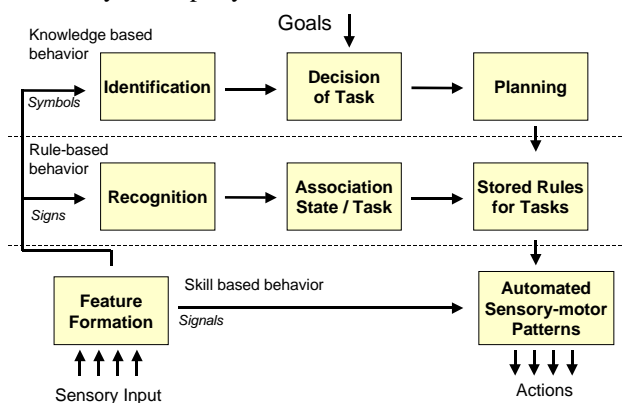
Figure 5: Three levels of human behaviour [Rasmussen, 1983]

Tasks and know-how that can be stated explicitly by the operator will be controlled by a stored rule or procedure. These rules or procedures may have been derived empirically during previous occasions, communicated from other persons know-how as instruction or a cookbook recipe, or it may be prepared on occasion by conscious problem solving and planning. The difference between skill and rule based behaviour depends on the extent to which the task is executed automatically or attentively. Information at the level of rule-based behaviour is processed as a kind of recognition, thereby invoking a rule that dictates the enactment of a certain behaviour (cue-task association) based in experience or formal training. The expertise and the familiarity of the operator define mainly whether the operator is operating at skill-based or rule-based level.

For instance, a defensive action on a single missile attack in open seas should be executed at skill-based level, while a defensive action on a multi-missile attack in coastal waters with many air and surface contacts could be executed at rule-based level (depending on the training and experience level of the operators). The information management support system should support the operator in guiding the operator through the decision making process by presenting all relevant information, and the different options and proposed advises (made by the computer system).

Knowledge-based behaviour is required for complex / novel situations where deeper understanding of the nature of the situation and explicit consideration of objectives and options are required. Information at the level of knowledge-based behaviour is processed as symbols, which are used to construct mental models representing causal and functional relationships in the environment. These models are constructed at different levels of abstraction and decomposition. The fact that the situation is novel and requires deeper understanding means that the information isn't able to determine the required information need in advance. The operator should interact with the information management support system to make the information need knowable to the system. The system could support the operator in the information gathering process and identification of the information need that has to be interpreted to reach a particular goal.

4.3 Design of an adaptive information management and support system

4.3.1 Positioning of the information management system in the C2-organisation

The problem analysis shows that the information management activities should support the operators in performing their situation assessment and decision-making functions. This support ranges from providing support defining the information requirements, in the information gathering process and balancing the presented information with the operator's information requirements. This support should guarantee that only relevant information is presented the right way and time to the right person. This demands a close co-operation with the other systems of the C2 organisation (See Figuur 6).

The communication between the information management support system is bi-directional. The information management support systems present of course the required information for the operators, but should also provide an interface to make the operator's information requirement constraints knowable to the information management system. These constraints consist of META - descriptions of the information need, the presentation formats for the information, and the conditions of passing the information to the operators.

Not all information is available in the database. In such situations, the information management should make an appeal to the C4I systems of the C2 organisation:

Decision Support: The raw information that is available in the database should be further analysed by the dedicated automated situation assessment and decision making analysis tools. The results of these analyses form the actual information need for the operator.
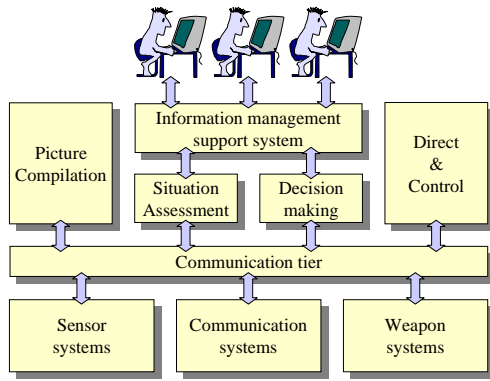
Figure 6. Positioning of the information management support system in relation to the main functionality's and systems of the C4I organisation.

Picture compilation: the available information isn't available in the tactical database and should be gathered from an adequate deployment of own sensor systems and / or should be gathered from other platforms. This requires an information search operation where the required information could be obtained, and /or the best use of all available data / info sources.

The above mentioned actions require a good co-ordination and communication with the involved C4I systems. The C4I systems should be able to react on triggers to start a particular analysis or information search operation, and the C4I systems should be able to notify the information management system that the requested information is available in the database.

4.3.2 The architecture

This section discusses the different processes of the information management support concept in more details. This model is based on results that are obtained within the IFICS project (See chapter 6). An overview of the different modules and their interrelationship is shown in Figure 7. There are five processes that could be identified as the core modules of and the thread for the information management concept:
1. Assessment of environment and provided information
2. Task monitoring and management module
3. User's information need module
4. Filter settings
5. Information Filter
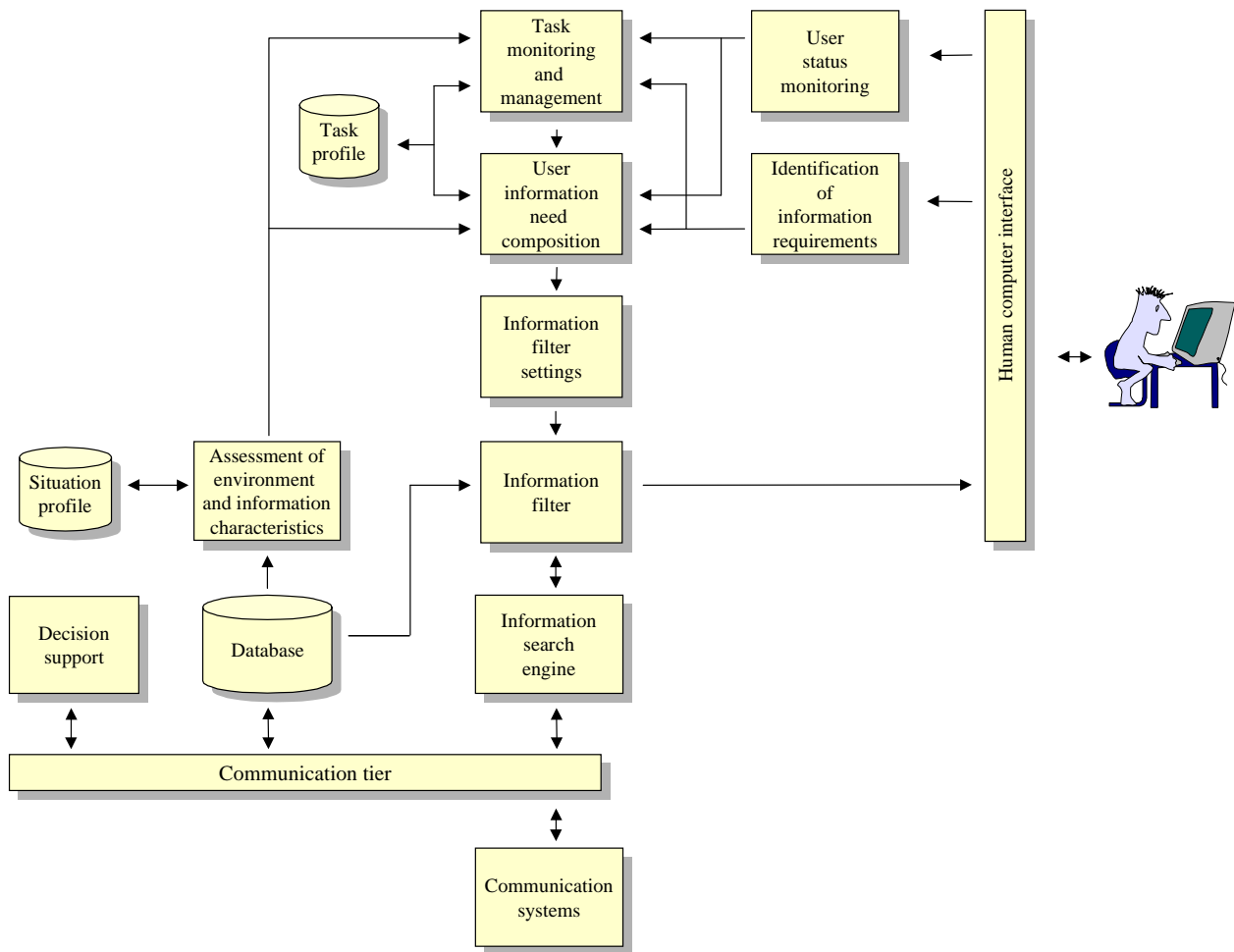
Figure 7. High level architecture of the information management support system

Three databases and four other modules support these five modules:

1. Databases
   a. Situation Profile database
   b. Task Profile database
   c. Tactical database
2. Supporting modules
   a. User status monitoring
   b. Identification of information requirements
   c. Information search engine
   d. Human computer interface

The information assessment module & Situation profile database

The module related to the assessment of the environment and information factors / characteristics assess the situation the environment of events that are important for the C2 organisation. The relevancy of the events is determined by the influence of that event on the mission of the platform. A set of event characteristics and their relevancy is stored in a situation profile database. The load of assessing the environment is expressing by measuring a limited set of the information load factors and characteristics as shown in Figure 1 in section 2.2.

Any system that employs filtering must decide which information to filter. The information that is relevant to the user's current tasks must be emphasised, whereas other information can be dimmed or left out entirely.

The relevance of the information in relation to the user's current tasks is expressed by the values of various attributes of the information. For instance, tracks normally have attributes identity, type (air, surface,…), range, speed, etc., and the user's current tasks will dictate that he is particularly interested in, say, hostile air tracks within a given range (this would be the case for a user performing an Ant Air Warfare task).

However, in some cases, the attributes that determine the relevance are not explicitly available in the information. For instance, the user may be particularly interested in threatening tracks, but the tracks database may not contain threat level information. In this case, it is necessary to interpret the behaviour of the track in order to estimate its threat level.

There may be other reasons to interpret the available information. For instance, the information presented to the user is driven by the events that occur in the environment. Some of these events may be easy to recognise, e.g. the presence of an unidentified air contact, whereas others may be more difficult, e.g. the presence of a formation aiming to protect a high-value unit, which requires an interpretation of the behaviour of several tracks.

The interpretation of information can be divided into interpretation of the environment, by which we have in mind mainly the tactical situation, and the provided information, by which we understand the text and voice messages obtained somehow by the user.

Interpretation of the environment covers recognition of events, e.g. deviation of a track from an established airway, the speed (or any other numeric attribute) exceeding a certain upper or lower threshold or changing a specified amount. Technically speaking, any query to the tracks database could correspond to an event. Release

and identification criteria are natural candidates for events. Other natural events include torpedo attack, submarine detection etc. Recognition of more complex patterns e.g. the presence of a formation protecting a high-value unit or various attack patterns, are also relevant. Several recent research projects, notably the EUCLID RTP 6.1 project, have studied a number of tactical analyses making use of advanced techniques from artificial intelligence, e.g. tactical threat evaluation, engagement co-ordination planning, and terrain analysis.

Concerning interpretation of provided information, it is interesting to employ message understanding technology to extract subject or priority information or structured or formatted contents from free-text messages or voice messages. For instance, position and other information might be extracted. In the case of voice messages, voice recognition must also be employed. Entire conferences are devoted to the message understanding field, providing valuable techniques.

Task monitoring and management module & Task profile database

The events detected as relevant will be a trigger for the operator to start and / or complete operator's tasks. The task monitoring and management module monitors the incoming events and determines the information need for every task that has to be carried out and distributes the tasks among the operators whom are qualified to carry out those tasks. The exact task sequence is based on the work procedures, the work-/information overload and the priority of every task. Information about the information-need, the load of each task and the procedures of performing each task is stored in a dedicated task profile database. A META language is used for expressing the information need and its presentation format in the task profile database.
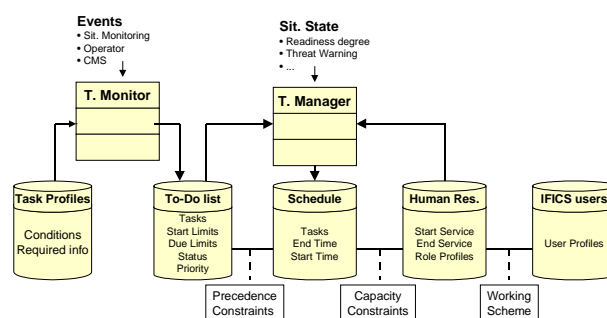
Figure 8: task monitoring and management module

Figure 8 shows the concept behind the task monitoring and management module. The task monitor runs continuously. It handles the incoming events and monitors the deadlines of the active tasks. New tasks are put on the 'To-Do'-list and accomplished tasks are removed from the 'To-Do'-list. The task monitor triggers the task manager if the content of the 'To-Do'-list has been changed. The task Manager calculates a new task schedule by taking the precedence and the capacity constraints of the human resources into account.

The contents of the task profile could be filled easily and in advance for skill and Rule-based tasks. Extra effort is required to define the information need for knowledge

based tasks. Such tasks could be divided into fragments of task sequences or the information need should be specified on-line and stored in task profile database. The operator could activate these task profiles at a later moment.

User's information need module

The user's information need module determines the total information need that has to be presented to each user, instead of determining the information need per task. The total information need is determined by the complete set of tasks that is (to be) executed by the user. The information need module has basically two ways of tuning the information load that is perceived by the user. The first option is to tune the actual amount of information, for instance, only confront the user with information that is required for top priority tasks. The second option is to tune the way in which information is presented: its presentation form. From experiments in the field of HCI, one has learned that some presentation forms are much more demanding than others; for instance, blinking tracks on a TDA attract much more attention than dimmed tracks, thus increasing the perceived information load.

Two aspects influence the information need (including presentation forms) that is determined for a user. The first aspect is the perceived tactical situation. If dramatic things happen in the perceived tactical situation, such as a missile attack on the user's ship, the presentation forms for all tasks may be influenced and it may be decided to give emphasis to the information need of top priority tasks, dealing with self-defence.

The second aspect is information load. As already depicted in Fig. 2, there are several ways to control information. The mechanism developed for IFICS is in accordance with option c) comprising both prediction and feedback. Prediction plays a role by having an information load attached to the information need per task. By summing up the information loads for all tasks, a prediction of the total information load is established. If this predicted information load reaches a particular threshold, it may be decided to remove the information needs for low priority tasks, until the predicted information load is beneath the threshold. Feedback plays a role by actually monitoring the user's behaviour, and delivering measures concerning the perceived information load to the information need module. Based on this feedback, it may now incorporate changes by either manipulating the total information need, or manipulating the presentation forms for each particular information need.

Filter settings module

The user's information need expressions are expressed in generic terms. This means that these expressions could not be used directly to access the used databases before any further processing of the information need expressions. The expressions will be parsed and translated into database queries to access the databases that are used within the C2 organisation.

Information filter module

The information need and presentation expressions are finally interpreted and used for accessing the databases of the C2 organisation, and passing the information to the operator's. But before doing that, the information search engine should provide clearness whether the required information is available within the database or that the information needs to be gathered by own sensors and/or from other platforms.

User status monitoring module

The task monitoring and management module gets feedback about the task progress from this module. The actions of the operator's are analysed to identify relevant events which refers to the start and/or end of tasks and to get a better insight in the actual information /workload of the operator in performing.

Identification of the information requirements module

Not all operators will have the same information requirement for the same tasks. This means that the operator should have the ability to adapt and refine the information need and presentation expressions as stored in the task profile database to meet the operator's wishes. In our concept, the operator is able to define new tasks and fill out the corresponding information requirements (information need, presentation, and start conditions) in a task template. This template will be activated on operator's demand or started automatically when the environmental conditions meet the start conditions as presented in the template of that particular task.

Information search engine

This module looks whether the required information can be found in the available tactical database or that the information should be provided by its own sensor systems and/or decision support systems, or that the other platforms should be asked to deliver the required information.

Human computer interface

The human computer is the physical interface between the operator and the information management system. The task of this module is directed towards the composition of the operator's display with:

1. A graphical tactical display.

2. An overview of all relevant / actual textual messages.

3. An overview of all relevant / actual interaction menus.

5.   The IFICS demonstrator

A demonstrator implementing the concepts that have been described above is in the process of being developed. Figure 9 shows a snapshot of the current system.

The demonstrator is composed of a number of agents [Knapik & Johnson] that are tied together in a proprietary agent framework based on Java RMI. Each agent implements one of the modules that were described in Section 4.3.2.

A simulator (whose interface appears in the lower right corner of the figure) is used to demonstrate the functionality of the system. In the snapshot a hostile missile appears on the tactical display (the red track). This has been recognised as an event by the tactical situation agent. Based on this, the task monitoring and management (TMM) agent has fired a missile self-defence task (consisting of a number of other subtasks). The information need definition agent has determined the information required by the user for these tasks. For instance, a pop-up display in the dialog areas prompts the user to select the weapon type to engage the missile, and time-to-first-fire and time-to-last-fire for the various weapon types of the platform are presented.

The system is highly configurable. For instance, it is a simple matter for the operator to change the information needs for each task.

Figure 9: IFICS Demonstrator

6.   Summary

The following statements summarise the obtained results:

The information overload problem will influence the performance of future command and control organisations. The structure of the organisation, the environment and the way the information is presented to the operator determine the sensitivity for information overload situations. Solutions for the information overload problem should be found in both in the structure of the organisation and applying advanced information technology concepts in supporting and optimising the information exchange between the computer system and the operator.

Operator support in gathering the required information and discerning the relevant information from the total available information would suppress but not completely avoid information overload situations.

A concept is postulated that could be used to filter and control the information flow among the computer system and the operator. The concept uses predefined templates containing the information need and the presentation formats for each task to support the filtering capabilities of the concept to ensure that only relevant information will be passed to the operator. Operator's load prediction and task progress measurement techniques are used to control the information load of the operator.

The concept is developed for supporting skill and rule based tasks, but extra research is required for supporting the operator in doing knowledge-based tasks. Work done so far shows that the information need and its presentation formats for skill and rule-based tasks could be expressed in advance.

7.   Overview of the involved projects

The contents of this paper is based on the research results that were obtained in two research projects:

1. Future Command and Control Concepts for naval vessels.

2. Information Filtering and Control System.

The first research project is directed towards the development of reduced manning concepts for future Command Information Centres of the Royal Netherlands Navy. TNO Physics and Electronics Laboratory and the TNO Human Factors institute carry out this research and are both part of the TNO Defence Research organisation.

IFICS (Information Filtering and Control System) is the name of a European research programme that is carried out as part of the EUCLID (European Co-operation for the Long-Term in Defence) RTP 6.11.1 programme. The IFICS consortium consists of five different companies from four countries: TNO-Physics and Electronics Laboratory, Hollandse Signaalapparaten B.V. (The Netherlands), TERMA (Denmark), DATAMAT (Italy) and INTRACOM (Greece). The research programme is carried out under the management of TERMA. The authors would like to thank the following people for delivering their contribution in the design and implementation of the IFICS demonstrator: Louwrens Prins (TNO-FEL), Lars Stavnem (TERMA), Domenico Pannucci, Luca Onofri (DATAMAT), Georgios Detsis, and Lefteris Dritsas (INTRACOM). The IFICS project started in 1998 to develop a system to avoid information overload situations in the human organisation by means of balancing the operator's information pull with the information push.

8.   References

[Maas and Keus, 1999] H.L.M.M. Maas, H.E. Keus, 'A methodological Approach to the Design of Advanced Maritime Command & Control Concepts'; Proceedings of the 1999 Command and Control Research and Technology Symposium'; Vol. 1, p. 174–191, June 29 – July 1 1999, Naval War College, Newport, Rhode Island.

[Scott, 2000] R. Scott, 'Danzig studies Dutch 'Lean manning'', Jane's Navy International, March 2000, Vol. 105, nr. 2.

[Schneider, 1987] S.C. Schneider, 'Information overload: Causes and consequences', Human Systems Management, 1987, vol 7, page 143-153.

[Tushman and Nadler, 1978] N.L. Tushman and D.A. Nadler, 'Information processing as an integrated concept on organizational design', Academy of Management Review, 3, page 613 – 624.

[Rasmussen, 1983] J. Rasmussen, 'Skills, Rules, Knowledge; Signals, Signs, and Symbols and Other Distinction in Human Performance Modelling', IEEE Transactions on Systems, Man and Cybernatics, SMC-13 (3), p. 257-267.

[Knapik & Johnson]: M. Knapik & J. Johnson, Developing Intelligent Agents for Distributed Systems, McGraw-Hill, 1998.

This page has been deliberately left blank

_____

Page intentionnellement blanche

Comprehensive Approach to Improve Identification Capabilities

Dr. Christoph Stroscher, Frank Schneider

Industrieanlagen-Betriebsgesellschaft mbH (IABG)

+49-89-6088 4065 / 3865
stroscher@iabg.de / fschneid@iabg.de

Keywords: Identification, Data Exchange, Data Fusion, Decision Support

Summary

The process and the prototype presented here, are dedicated to improve the overall identification capability. This is aimed to be achieved by making available all identification related information, - i.e. local and remote data -, fusing and interpreting them, and supporting the decision process by offering a recommendation together with all explanation that might be desired.

The paper presents a solution that uses the Identification Data Combining Process (IDCP) according to draft STANAG 4162 as baseline. The prototype, assisting in identifying airborne objects, is the result of an experimental system using simulated as well as live data in a German Control and Reporting Centre (CRC).

1. Introduction

Motivation The military identification function asks for the identity or at least for a classification of an object under consideration. The performance has essential impact on the success of military missions and a reliable and rapid identification can be seen as force multiplier. Identification is a tri service function which is performed by various host systems under various conditions. I.e., in different alert states from peace to war, under various scenario conditions - e.g. a mixture of civil and military activities -, and different not in advance known compositions of own or enemy forces in e.g. joint, combined operations. To maintain always a complete picture of activities, continuous identification covering all objects of interest and being available in time is required.

There are various sensors and sources, including procedures, delivering information from which the identity or the category of an object can be inferred. As long as coverage, reliability and timeliness are not guaranteed by one source alone, all information obtainable and related to an object needs to be combined in order to make the best assessment of all information available. Such a combination would improve the coverage and substantially reduce uncertainty by exploiting the synergy of various sensors and sources. Implemented as an automated, real time process it would rapidly deliver results and improve the situational awareness.

Challenge A concept to improve the identification capability as indicated above would require interoperability that allows to exchange and unambiguously interpret outputs from sensors and sources. Such a fusion system should be available for all host systems with (potential) identification tasks. It should cover the common kernel of identification functions, i.e. the fusion and interpretation of received information, and should be adaptable to any specific host system's needs like the lay-out of the identification function or the interaction with the operator. The fusion system needs to be flexible versus the various environmental conditions (scenarios, groups of objects to be distinguished) in order to make the best assessment of the obtainable information. In summary, such an approach would comprehensively improve identification capabilities by providing and assessing all obtainable information.

Approach The fusion concept presented here, complies with the requirements by employing Bayesian principles together with a flexible choice of what distinction is relevant for identification and how risks have to be assessed. It provides standard formats for the outputs of all sensors and sources, a fusion component, the interpretation of discriminated target attributes with respect to identification, and the derivation of a recommendation within the range of decision alternatives of the hosting system. It can be used within any scenario in being provided with data describing the perception of composition of forces.

2. The Concept

Subsequently the principles of the approach are mentioned.

The concrete employment of sources needs a careful analysis and sometimes specific adaptations in order to keep the process simple while making the best use of the principles.

Each source output needs to be associated to the object of consideration in the sense that the output refers to the object.

Conversion In order to express source outputs in a standardised format for further treatment, each source type is characterised by the set of target attributes it is able to discriminate. This set of attributes is used to convert each possible source output according to the Bayesian approach into an likelihood vector, a set of conditional probabilities referring to the attributes. In this way each source output is equivalently expressed by an likelihood vector describing what is discriminated and how good it is discriminated. Data or algorithms determining the likelihood vector values are contributing as a priori data to the process expressing the sources performance. These source dependend likelihood vectors reflect now the sources contribution to the process and are not yet interpreted with regard to allegiances or other specific operational, environmental aspects that are important for identification. Data exchange takes place at this level of not yet interpreted information in order to keep control on the data sources and not to loose pieces of information. As each source output corresponds to a predefined likelihood vector, the amount of data to be exchanged can be significantly reduced.
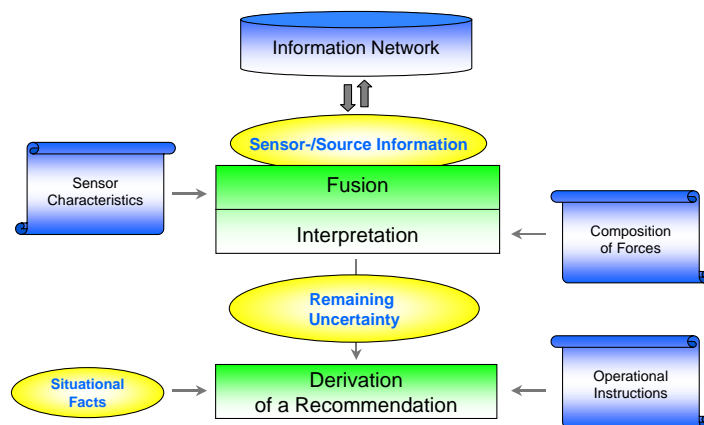
Data Flow and Processing

Fusion, Interpretation Fusion takes place at the level of source declarations – expressed in likelihood vectors - which might be locally or remotely obtained. The conditional independence of measurements and observations, i.e. source declarations, allows to combine them by multiplication of the corresponding likelihood vectors. To interpret the fusion result with regard to the attributes needed for identification (like allegiance) a new set of a priori data, used only once in the process, is needed. In principle this is an a priori distribution of the combination of all distinguished target attributes together with those needed for identification. Under some often justifiable assumptions (conditional independence of distinguished target attributes) the

granularity of such a distribution can be significantly reduced.

The result so far is a posterior distribution over target attributes required for identification like own, enemy, and non aligned forces or these in combination with civil and military allegiance. As the process is mathematically consistent, the posteriors exactly express the achievable reliability in terms of allegiance.

Risk Assessment The posteriors are an important basis for the identification decision, however, in addition the decision depends on some situational facts (like alert state, the targets position - in case of airborne objects - in the airspace, or collateral data), operational procedures and operators judgement. To support the operator, a risk assessment is performed that delivers a recommendation out of the set of possible decision alternatives. Situational facts, like alert state determine this set of alternatives. For each combination of situational facts a loss table needs to be prepared that contains for each combination of allegiance and decision alternative a loss value reflecting the loss (military risk) taking place if the allegiance would be true and the alternative selected. In this way operational assessments and procedures are reflected in the process. Subject to the situational facts, for each alternative the respective loss values are 'weighted' by the posterior likelihood vector to determine the risk for each alternative. After assessing the complexity of the decision situation, the alternative with the minimum risk value assigned will be recommended.

Peculiarities

The approach defines a specific data fusion system, which is specifically designed for identification purposes. It covers the common kernel of the identification function and can be easily laid out for the different purposes of identification. It is open for any source type including future ones. As fusion takes place at a level where only source information is combined, an optimal exploitation of information from local and remote sources is reached.

As it is mainly controlled by clear defined data (a prioris, loss values), it can easily be adjusted to any operational condition.

3. Adaptation

As indicated above, the adjustment to operational procedures and specific identification tasks can be easily achieved by the selection of appropriate structural elements and data.

Adaptation to Host Systems The system can be designed as a collection of modules with clearly defined interfaces. An implementation into a host system would mean to adapt it to the host system specific needs by adding some software parts. This would comprise adaptations to the HMI and the possibilities to interact with the system, as well as interfaces to various databases and to the various sources of

Example of Contributing Information
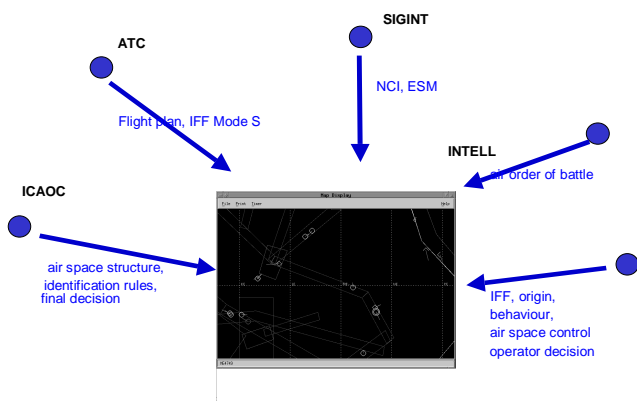
information. The latter could e.g. comprise the connection to a track source, and to locally as well as remotely available sources of identification information.

Adaptation to Architectures Any operational architecture can be supported by the system. It is possible to use the system as single node solution, treating only locally available data, it is however prepared to work in a netted environment in order make use of possible synergy. Different nodes in such a network could obtain the whole system or only components of it, depending on their operational function that might be assigned to them.

4. Experiences

The authors have had the opportunity to gain a lot of experience with the approach by supporting the development of the STANAG, by a feasibility study, and by developing the experimental system together with the prototype (see below).

The feasibility question focused on the availability of a priori data and their sensitivity, i.e. the impact of imprecise a prioris on the result. The answer was satisfactory: With known sensors and sources with at least some contribution to identification no problem occurs. Combining source outputs with inconsistent indications to allegiance may raise problems. These are, however, just the complex decision situations which are reliably detected by

internal control mechanisms of the process and, if desired, brought to the attention of the operator.

Summarising, the process may be characterised as follows:

- It provides the necessary interoperability to exploit the synergy within various sensors and sources. So it can be seen as system of systems.

- It is a consistent and reliable process making the best assessment of information available.

  - It is prepared to combine and assess any kind of information contributing to identification including future sensors and sources.

  - The process has a high flexibility to be adjusted to any host systems needs and any operational condition (mobile, joint and combined mission), supporting any operational architecture.

  - It is a substantial support to operators' tasks and reduction of load.

5. Experimental System

This paragraph is dedicated to a prototype system which allows assessments in terms of technical and operational performance.

It has been developed and adapted to a Control and Reporting Centre (CRC) in order to

- adapt the generic IDCP process to a typical operational environment,

- use common identification procedures in peace and war,

- use available simulated and live data,

- show the operational benefit,

- allow the reuse of the system in other environments or studies and

- allow a flexible adaptation and configuration to other requirements.

Results in terms of processing performance, flexibility, adaptivity, robustness, reliability, and operational benefit were won in the project.

Additionally the experimental system was installed at the NC3A.

The adaptation of the data fusion process to the operational environment consisted of three major steps:

- Architecture
  The definition of a system architecture considers interfaces to existing data sources, processes prior to data fusion like correlation

and association of the source data and processes after the data fusion e.g. result presentation, data fusion explanation and operator interaction.

- Configuration
  This includes the definition of a priori data and environment specific data bases to configure the system to a specific operational environment and scenario. This means, that operational procedures have to be analysed, whether they should impact the process of data fusion and how these should be reflected in the behaviour of the system. Furthermore the system has to present the received data and the results of the fusion in an way that the operating personal trusts and understands the system.
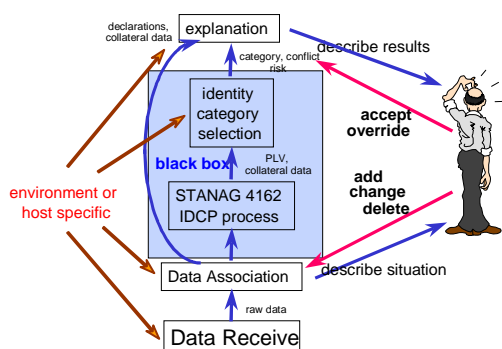
Embedded Data Fusion

- Testing
  Experience is gathered by testing the system. Live data behaves different than specifications may give the impression. The parameters controlling the processes depend on the quality of the received data and can be optimised only in intensive test periods. Software reliability can only be ensured with a careful analysis of all parts of the system using a high amount of source data.

Based on our experiences we would make the following conclusions concerning the implementation of an data fusion system in an operational environment:

- The generic data fusion process as defined in the STANAG 4162 can be easily adapted to the operational environment.

- Additional processes are necessary to embed this data fusion process in a system. These processes are not standardised and ensure that the adaptation to environment specific requirements is possible. The implementation might be complex and can contribute to the system performance in the same way as to the quality of the data fusion process itself.

- The design of a high performance data fusion system requires a deep understanding of operational requirements and procedures. A specific aspect is that the system provides only reasonable results if operational experience was successfully transferred in the behaviour of the system.

6. Way Ahead

Future steps in order to exploit the concept for a maximal operational benefit could be as follows:

- Gradual implementation, provision for new systems

- Expansion to new regimes like air - ground and ground - ground

- Definition of networks and forwarding information exchange requirements for tactical data links

- Use of future sensors and sources

References

Caromicoli, A., Kurien, T. Multitarget Identification in Airborne Surveillance. SPIE Vol. 1098 Aerospace Pattern Recognition 1989

DuPree, J., Antonil J. Information Measures Unify Decision Making. USAF Rome Laboratory/IRAP, April 1997

Fagin, R., Halpern, J. A new approach to updating beliefs. Uncertainty in Artifical Intelligence 6, Elsevier Science Publishers B.V: 1991

Hall, David L. Mathematical Techniques in Multisensor Data Fusion. ARTECH HOUSE, INC. 1992

Leung, H., Wu,J. (2000) Bayesian and Dempster-Shafer Target Identification for Radar Surveillance. IEEE Transactions on Aerospace and Electronic Systems Vol 36 No. 2 April 2000

NATO. (2000) Standardisation Agreement 4162 Technical Characteristics of the NATO Identification System (NIS), March 2000 version of Annexes C and D

Rehfeldt, M. Mathematical Description of an IDCP extension. IABG Working Paper, 1991

Schneider, F., Stroscher, C. Integration of an IDCP Experimental System into a CRC. IABG reports 1998, 2000

Stroscher, C. Series of IABG Working Papers in support of STANAG 4162 development. 1991 to 2000

Waltz, E., Llinas, J. Multisensor data fusion ARTECH HOUSE, INC. 1990

Automatic Detection of Military Targets utilising Neural Networks and Scale Space Analysis

A. Khashman

Chairman, Department of Computer Engineering
Near East University
Lefkosa, KKTC, Mersin 10
Turkey
E-mail: amk@neu.edu.tr or amk@ebim.net

Summary: This paper reports on a new approach to detecting military targets. The novel idea is based on combining neural network arbitration and scale space analysis to automatically select one optimum scale for the entire image at which object edge detection can be applied. Thus, introducing new measures to solve many of the problems existing in the discipline of image processing, such as: 1) poor edge detection in medium-contrast images 2) speed of recognition and 3) high computational cost. This new approach to edge detection is formalized in the Automatic Edge Detection Scheme (AEDS).

I.   Introduction

Recent operations in conflict areas around the world have made the need for accurate image processing and fast target detection for military systems more obvious. In Kosovo, for example, a civilian tractor convoy was mistakenly targeted as enemy military target. Therefore, there is a need for more advanced intelligent target recognition systems.

A novel approach to target detection is presented within this paper. It is based on combining the three fields of scale space analysis, edge detection and neural networks. The result is an automatic edge detection scheme (AEDS) that delivers very quick edge detection of objects within medium-contrast images, through the automatic selection of a *single optimum scale* for applying the scale space edge detection to an *entire image*. The computational cost is kept to a minimum through using a fast edge detection operator combined with the power of a successfully trained neural network that recognizes only one correct scale (referred to as the ideal sigma σ_{Ideal} in this paper) for the entire image, out of the many available scales possible in scale space. Noise sensitivity and scale dependency can be problematic in image recognition. In this novel approach, both phenomena have been utilised to create a criterion

upon which the ideal edge scale for optimum edge clarity will be chosen.

The proposed AEDS is expected to overcome the common problems that are experienced when implementing image recognition. These are: speed, computational expense, noise sensitivity and scale dependency, poor edge detection within medium-contrast images, large amount of training data required for the employment of neural networks and providing rapid automatic edge detection.

The AEDS is implemented to rapidly detect various military targets in low to medium contrast images. The camouflage targets comprise a military aircraft (JET), an armored tank (TANK), a military off-road vehicle (JEEP), a rocket launcher (SCUD) and a navy boat (BOAT). All five objects have had their images obtained in three different possible situations.

II.   Scale Space Analysis

There are two phases of implementing the AEDS. The first (the Preparation Phase) uses a fast format of the Laplacian of the Gaussian (FLoG) edge detection operator [1][2], as shown below in (1). The FLoG operator is convoluted with a number of images that represent the training set of images for the neural network, at seven scales in scale space. The standard deviation (σ) of the Gaussian function in the FLoG operator is variable and it dictates the amount of smoothing to be imposed on the image prior to edge detection [3]. The high computations involved in multiscale processing, can be marginally reduced if a suitable scale is found, and then used [4][5]. A criteria for selecting the 'ideal edge detection at one ideal scale (σ_{Ideal}) is set up, based on the convolution results using 3-dimensional objects [6]. The results of this phase represent the training data for the neural network arbitration in the second phase.

$$\nabla^2 G(x, y) = h(x)g(y) + g(x)h(y) \qquad (1)$$

whereby

$$h(\xi) = \sqrt{A}\left(1 - \frac{\xi^2}{\sigma^2}\right)e^{-\frac{\xi^2}{2\sigma^2}}$$

$$g(\xi) = \sqrt{A}e^{-\frac{\xi^2}{2\sigma^2}}$$

III.   Neural Network Arbitration

The second phase (the Training Phase) in the implementation of the AEDS is training a neural network to recognise σ_{Ideal} for an input image. This is based on the hypothesis that σ_{Ideal} is a function of noise [7], as in (2).

$$\sigma \propto \frac{1}{N} \qquad (2)$$

where, σ is the ideal scale (σ_{Ideal}) and N is the

amount of noise present within the image. The trend of this non-linear relationship is what the neural network will be trained to recognise. The basis of the methodology is that the alteration in the amount of noise present within the image causes a change in the choice of the ideal scale σ_{Ideal} and thus the ideal edge detection of the image [8]. The presentation of various images and their corresponding ideal scales σ_{Ideal} will teach the neural network this relationship that is impossible to solve using conventional techniques. Having learnt successfully, the neural network will be capable of selecting only one ideal scale at which scale space edge detection can be carried out.

IV.   The Training Data

Various three dimensional objects have been chosen and they represent potential real-life military camouflage targets. Their two-dimensional projections have been captured and used for the implementation of the neural system (AEDS).
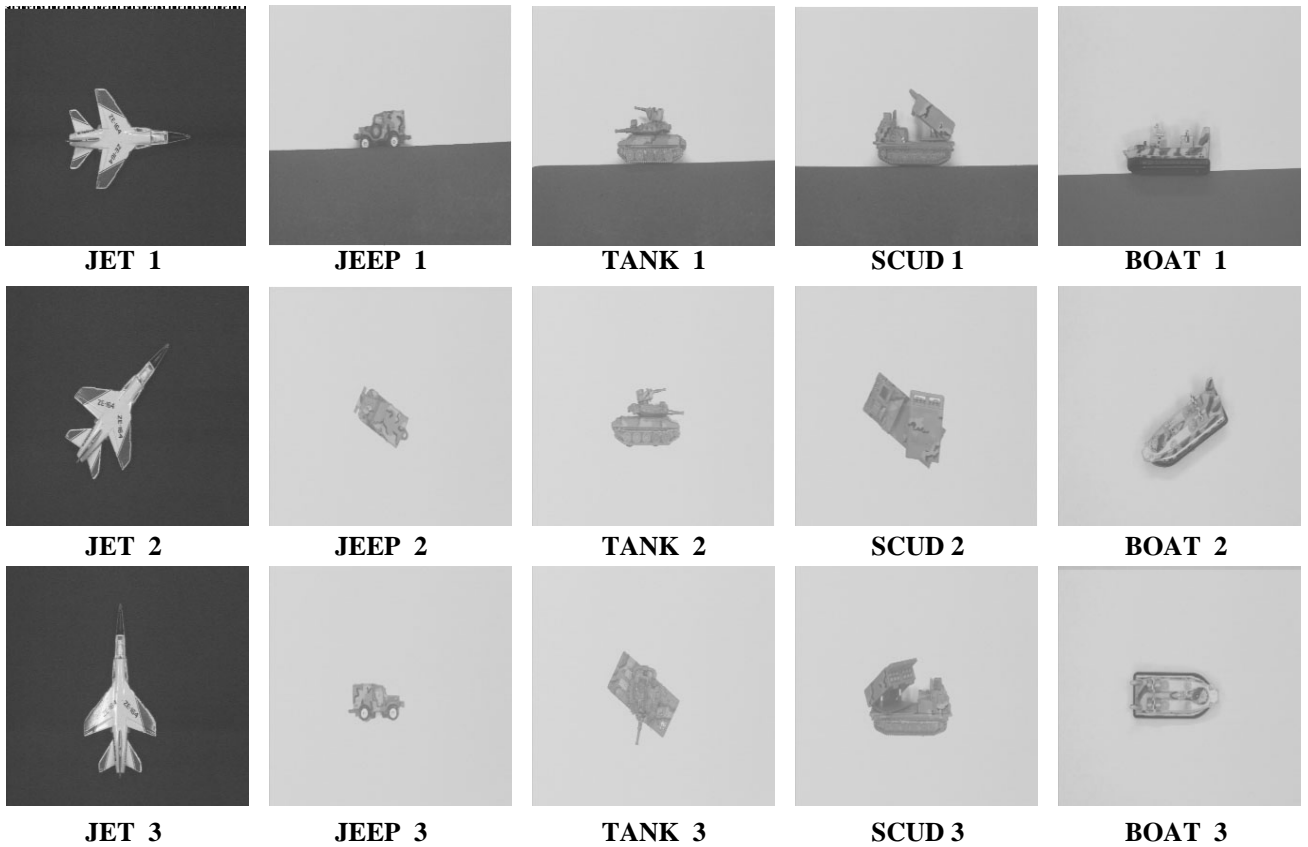
Figure 1. Images of military targets

The various objects comprise a military aircraft (JET), an armoured tank (TANK), a military off-road vehicle (JEEP), a rocket launcher (SCUD) and a navy boat (BOAT). All five objects have had their images obtained in three different possible situations, thus resulting in fifteen images available for the implementation of the automatic target detection system. Figure 1 shows the various original military target images.

A. The Preparation Phase
The first phase of the AEDS is implementing scale space analysis. Military images will be analysed using scale space, thus providing the ideal detection, represented through the ideal scales, σ_{Ideal}, for the various military targets.

The evolution of the military objects' edges in scale space can exhibit very many scales for certain images, and the scale space events that occur could lead to the disappearance of the object. This depends mainly on the background and how well camouflaged the target is. Figures 2 and 3 show two examples of the series of scale space events that occur on two of the targets. These are JET 2 and TANK 2 respectively.

The results of applying the first phase of the AEDS, that is the scale space analysis, have led to identifying the ideal scales at which the edge detection operator the Fast Laplacian of the Gaussian (FLoG) is to be applied. All the necessary data for the implementation of the second phase of the automatic edge detection scheme, which is the neural network recognition of the ideal scale (σ_{Ideal}),

have been prepared. Table 1 describes the fifteen military images together with their ideal scales (σ_{Ideal}).

B. The Training Phase
The second phase of the AEDS is implementing neural network arbitration. This phase involves training a neural network to recognise and select the ideal scale (σ) for the edge detection for any military image. The neural network will be trained to relate the noise and the intensity within the images to their ideal scales, σ_{Ideal}.

There are seven different scales used in our edge detection. These seven scales are sufficient to demonstrate the occurrence of scale space events on all the objects' edges available for the system implementation.

For the training purpose, ten out of the fifteen available images are used for training the neural network and for recalling. These ten images comprise the first two of each military object. That is the images with suffixes '1' and '2'. The remaining five images with suffixes '3' are used for testing the generalisation properties of the neural network to recognise the ideal scale for the military targets. For example; JET 1 and JET 2 are used for training, whereas JET 3 is used for generalising.

The multilayer perceptron neural network, which has been developed for the AEDS, is based on the back propagation learning algorithm, with the total number of three layers, comprising, input layer, hidden layer and output layer. The Acquisition of
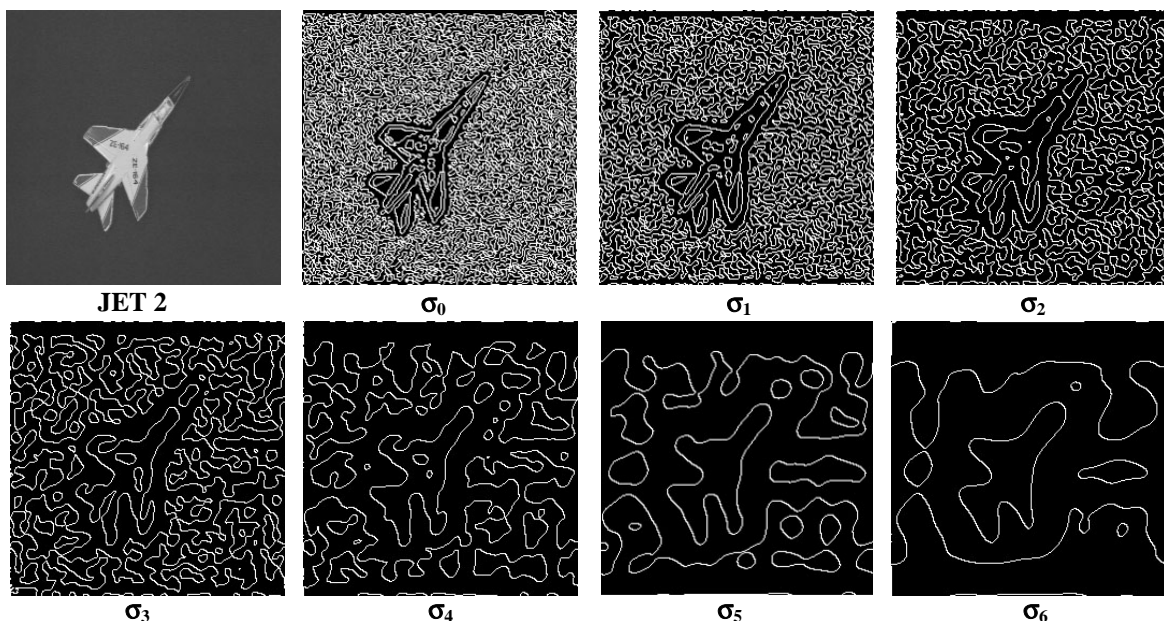
Figure 2. Scale space events occurring on JET 2 at various scales.
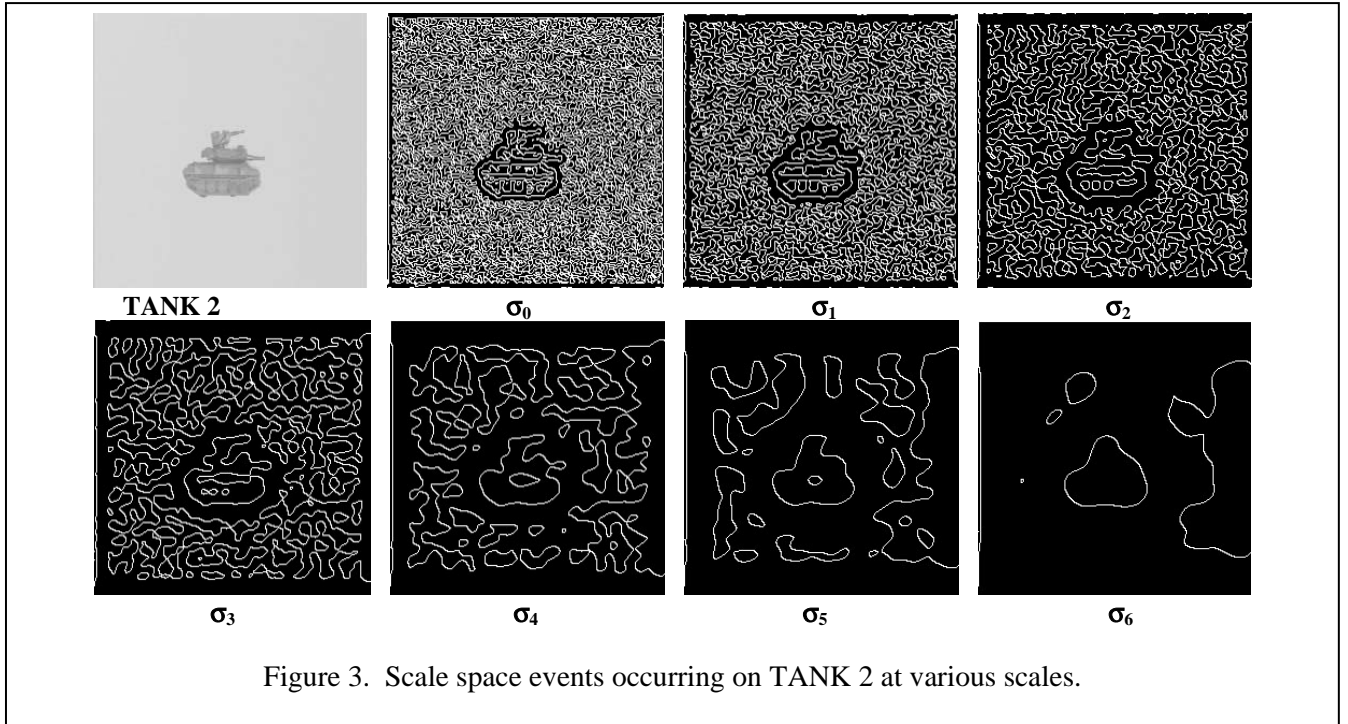
Figure 3. Scale space events occurring on TANK 2 at various scales.

Table 1. Ideal detection scales for the military images

JET 1	σ_5	JEEP 1	σ_3	TANK 1	σ_2	SCUD 1	σ_3	BOAT 1	σ_3
JET 2	σ_5	JEEP 2	σ_1	TANK 2	σ_3	SCUD 2	σ_3	BOAT 2	σ_3
JET 3	σ_5	JEEP 3	σ_1	TANK 3	σ_3	SCUD 3	σ_3	BOAT 3	σ_3

Table 2. Final neural network parameters for military targets.

Hidden Nodes (H)	Learning Rate (η)	Momentum Rate (α)	Initial Weights (W)	Error Level (ε)	Iterations (I)	Training Time (Tt)	Run Time (Rt)
70	0.004	0.20	[-0.3 - +0.3]	0.0074	3000	4 hrs	0.63 s

Table 3. AEDS total running time at the various scales.

Ideal Scale (σ_{Ideal})	σ_0	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6
Scale Recognition Time (seconds)	0.63	0.63	0.63	0.63	0.63	0.63	0.63
Edge Detection Time (seconds)	1	2.66	6.03	12.91	26.29	53.92	107.58
Total Time (seconds)	1.63	3.29	6.66	13.54	26.92	54.55	108.21

the training data and presenting it to the neural network is very important, and care should be taken when selecting the training data. Manipulating the large amount of data available, when dealing with images, can be very computationally expensive and hence can take a long time. However, the use of a Sun-Sparc 10 running the UNIX operating system, together with C-language source code, provided a quick and powerful tool to optimise training time.

Thus, complying with the objectives of the neural system.

C. Results
The neural network converged and learnt in 4 hours, whereas the running time for the generalised neural network was 0.63 seconds. Table 2 lists the final parameters of the successfully trained neural network. In order to keep the learning time down, a
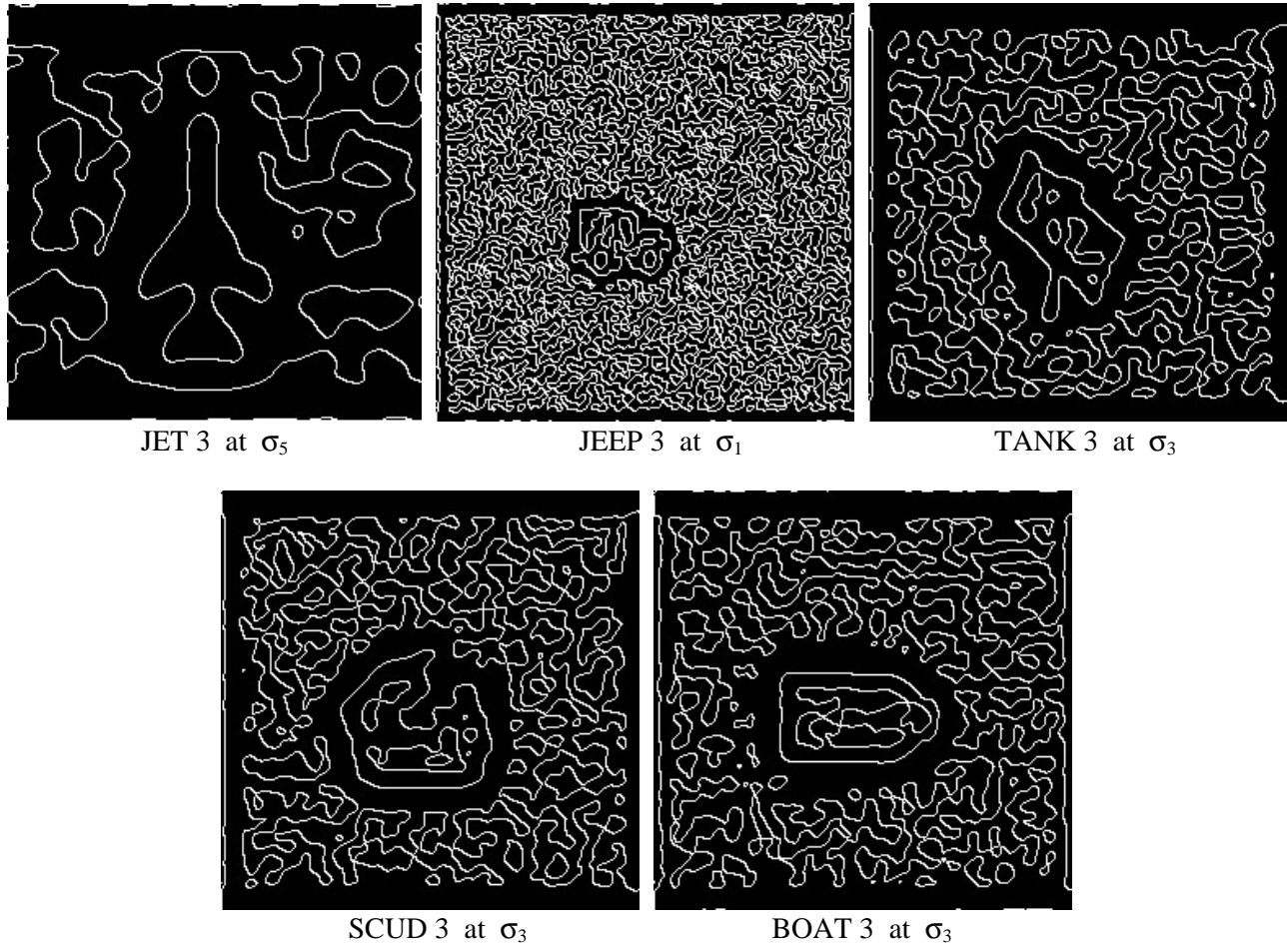
Figure 4. AEDS recognition of targets at their ideally-selected scales

minimum error of 0.0074 was regarded as adequate, as good recall and generalisation were obtained for the network. Table 3 shows the total running time for the AEDS at the various scales upon presenting an image to the system.

The robustness, flexibility and speed of the AEDS has been demonstrated through this application. The recalling of the ten training images was 100% successful, where all training images were allocated their correct ideal scales. The generalisation of the neural network, has also earned a similar success rate of 100%. The neural network recognised the correct ideal scales for the remaining 5 images, which it had not been trained on before.

The generalisation images, JET 3, JEEP 3, TANK 3, SCUD 3 and BOAT 3 can be seen in Figure 1. The ideal detection for the five targets as recognised by the neural system can be seen in Figure 4.

V.   Conclusions and Further Work

The automatic edge detection scheme AEDS has been successfully applied to the recognition of various camouflaged military targets. An implementation speed of 0.63 seconds is obtained when the neural network is used to generalise, as part of the complete automatic edge detection scheme. The total edge detection time including the automatic scale recognition time is in the range of (1.63 - 108.21) seconds; depending on the automatically selected ideal scale for edge detection. The necessary image data and information about enemy military targets can be, recurrently, used to train the AEDS in order to keep up-to-date information regarding any upgrades within the enemy military arsenal.

All the objectives which are outlined in the introduction section have been met, where:

- A remarkable scale recognition time of 0.63 seconds was achieved.
- High computational costs were reduced through using a fast edge detection operator that operates at one scale, rather than on a multiscale basis for an entire image.
- The AEDS is adaptable to the presence of noise and the variety of scale within the images. In fact, the novel technique utilises noise and scale information in order to produce an 'ideal' edge detection, thus making the AEDS ideal for implementation in other fields.

Further work can be carried out into implementing the AEDS in the rapid recognition of friendly military vehicles in action. This is necessary to avoid targeting own military arsenal, which could arise due to human misjudgment in real time.

VI.   References

[1] J.S. Chen, A. Huertas and G. Medioni, "Fast Convolution with Laplacian-of-Gaussian Masks", *IEEE Pat. Analys. and Mach. Intell.,* Vol. 9, pp. 584-590, 1987.

[2] Marr and E.C. Hildreth, "Theory of Edge Detection", *Royal Soc. London,* pp. 187-217, 1980.

[3] G.E. Sotak and K.L. Boyer, "The Laplacian-of-Gaussian Kernel: A Formal Analysis and Design Procedure for Fast, Accurate Convolution and Full-Frame Output", *Comp. Vision, Graphics and Image Processing,* Vol. 48, pp. 147-189, 1989.

[4] T. Lindeberg, "Detecting Salient Blob-Like Image Stuctures and Their Scales with a Scale-Space Primal Sketch: A Method for Focus-of-Attention", *Int. J. Computer Vision,* Vol. 11, pp. 283-318, 1993.

[5] T. Lindeberg, "Edge Detection and Ridge Detection within Automatic Scale Selection", *IEEE Conf. Computer Vision and Pattern Recognition,* pp. 465-470, 1996.

[6] A. Khashman and K.M. Curtis "Scale Space Analysis Applied to 3-Dimensional Object Recognition", *E-LETTER on Digital Signal Processing*, Georgia Institute of Technology, Atlanta, USA, Issue No. 24, June 1995.

[7] A. Khashman and K.M. Curtis, "A Novel Image Recognition Technique For 3-Dimensional Objects", *IEEE Int. Conf. (DSP'97)*, Santorini, Greece, 1997.

[8] A. Khashman, "AEDS: An Edge Detection Scheme Using Scale Space Analysis And Neural Network Arbitration", *Ph.D. Thesis, The University of Nottingham*, UK, 1997.

[9] A. Khashman, "Automatic Edge Detection of DNA Bands in Autoradiograph Images", *Proceedings of the IEEE International Symposium on Industrial Electronics (ISIE'99)*, Bled, Slovenia, July 1999.

Information Visualisation in Battle Management

Margaret Varga, Steve McQueen and Adrian Rossi
Defence Evaluation and Research Agency
St. Andrew's Road, Malvern, Worcs. WR14 3PS. UK
E-mail: varga@signal.dera.gov.uk, mcqueen@dera.gov.uk

Abstract

Visualisation is often thought of as simply the use of computer systems to display processed data graphically, often in a rather colourful and complex manner. More generally, however, visualisation is the human's capacity to utilise effectively and efficiently the output from a computer in order to understand data.

Military operations today depend heavily on the C^4ISR (Command Control, Communications, Computing, Intelligence, Surveillance and Reconnaissance) framework. Unfortunately many military systems make it difficult for users to develop a useful understanding of the information relevant to immediate requirements which is contained within the massive amount of data that flows from the various intelligence sources. The users may not be able to use the systems to extract the information from the data[1], or they may not be able to create displays that allow them to see what they need. Potential information sources may be ignored, or not well used, because techniques for extracting information are deficient. As a consequence, users of many current systems discard much data unassessed. Strategic and tactical actions, simulation and training are all seen to be significantly less efficient than they might be because commanders are not able to access, assimilate and exploit all the available information.

New technologies and data sources now envisaged will require radically improved ways for allowing users to interact with data. Interaction is critical, but at present information is usually presented to commanders, analysts and executives as a passive situation display. Effective visualisation requires the users to interact closely with the visual, auditory and perhaps haptic displays.

This paper describes the UK Master Battle Planner (MBP) [6]. The MBP is an Air Tasking Order planning tool, which aims to provide an adaptive, decision-centred, visualisation environment for UK Joint force commanders. The MBP's developing mission assessment component is also described.

---

[1] Information and data differ in that information is that which is useful or relevant to the task at hand, wheras data may or may not be relevant to the need.

Introduction

Military operations today depend heavily on the C^4ISR (Command Control, Communications, Computing, Intelligence, Surveillance and Reconnaissance) framework. This is concerned with the collection, dissemination, processing and interpretation of large volumes of data and information [11]. As battlefield operations become increasingly complex there is an increasing burden on commanders and operations room personnel to act as information assimilators and overseers. Recent conflicts have demonstrated the need for a revolution in the methods for handling the necessary information [10]. This has been found to be especially important for Joint and/or Combined operations where the larger tactical picture is of fundamental importance to the operation planner and controller. Such Joint/Combined operations are, of course, becoming increasingly likely.

An accepted model of conduct is the 'Observe, Orient, Decide, and Act' (OODA) cycle (also known as the Boyd Cycle). The OODA loop is a logical cycle of analysis, planning, implementation and assessment performed by battle commanders, see figure 1:

- *observe* the current state of the battlespace (friendly, hostile, neutral forces, weather, terrain etc.);
- *orient* own forces to adjust to the changes in the battlespace;
- *decide* what to do next (mission planning); and
- *act* on these decisions (e.g. fly the missions).

To date, research has tended to concentrate on supporting the '*decide*' and '*act*' stages of this cycle, e.g. the UK Master Battle Planer (MBP). There has, until now, been no research addressing the need for integrated and automated support for the first two stages of the OODA cycle, i.e. '*observe*' and '*orient*'. Developments in these latter two areas will allow the assessment of the actions (e.g. battle damage to targets, mission reports, enemy actions) to be fed back into the early stages of the next cycle in a much shorter time.

In this paper the UK Master Battle Planer (MBP) and, in particular, its developing mission assessment component are described. The mission assessment component provides the commander and his operations team with real-time feedback for next-step campaign mission
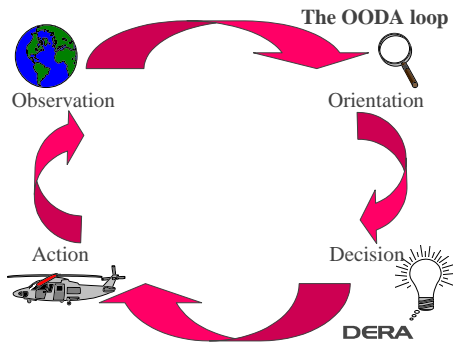
Figure 1: OODA Cycle

planning based on the performance of the accomplished campaign, i.e. the observe/orient/decide component of the OODA cycle.

Dynamic Visualisation of the battlespace will assist the battle commander and his team in projecting ahead from the orientation stage to decision making, supporting the development of shared mental models and situational awareness. Consequently, it should enable accurate perception of the environment and comprehension of the situation; it should also facilitate projection of future status. For example, if a mission was launched to destroy a bridge the commander will need to know:

- whether the bridge was hit;
- if so, which parts of the bridge were disabled;
- does the mission need to be repeated;
- factors for and against a repeat mission.

In campaigns such as Desert Storm and Kosovo, UK forces currently undergo the OODA loop every 72 hours. In other campaigns, e.g. against guerrilla combatants, or in famine relief, the OODA cycle will need to be much shorter. There is, therefore, an urgent need to be able to assess not only the success or failure of campaign, but also to monitor continuously the detail of campaigns and logistics in a ready and efficient manner, i.e. complete the OODA cycle in a short time (within the enemy's command cycle).

Military Needs

Of paramount importance to the military is the need for flexibility. The prime purpose of military systems is to deliver military force to achieve an objective, with the most important scenario being war, where failure would be catastrophic. However, there are Operations Other Than War (OOTW) and Low Intensity Conflicts (LICs)

where the systems designed for the war scenario lack the required flexibility. For example, an air planning system may enable missions to be planned to drop bombs on a target from 20,000 feet but cannot be adapted to famine relief operations where there are multiple 'targets' and the altitude for the 'package release' is a mere 6 feet.

There are many military Command and Control systems in use today that claim to assist the command team in the performance of their tasks. Unfortunately, the majority of these systems support the process that was prevalent at the time of their design and the systems cannot be changed (easily) to support an alternative process because the process is embedded within the systems. For example, the Improved UK Air Defence Ground Environment (IUKADGE) Command and Control System (ICCS) was designed in the late 1970s and implemented in the 1980s. It was accepted for military use in 1991 and is still in use in today, unchanged. This is despite the changing threat to the UK Air Defence Region, the requirements for more Out Of Area (OOA) operations, and the changing role of the operators it supports.

The above example highlights the problem with old, legacy systems. However, designers are still using the same concepts in new designs. For example, all US forces are mandated to use the Contingency Theater Automated Planning System (CTAPS) for air battle planning. This system is migrating to the Theater Battle Management Core System (TBMCS) which was scheduled to be delivered late 1999. These systems involve a great deal of operator-to-information interaction, but use the traditional method of allowing an operator to access a database, i.e. tables, which allows the operator no flexibility over how the information is presented. This potentially increases operator workload, forcing the user to divert cognitive resources towards operating the system and away from the primary the task, which could result in degraded task performance. Furthermore, the CTAPS and TBMCS systems are designed for US operations and therefore incorporate US doctrine. That is, it forces the operators to follow US rules and offers no flexibility. Furthermore, it is also not very scaleable. It is good for large US-style operations but not very good for smaller UK-led operations, as the overhead in machines and maintenance is tool high. The UK has procured TBMCS for the Pilot Joint Forces Air Component Headquarters (P-JFACHQ) system because in a US-led coalition CTAPS/TBMS will be the mandated system.

It has been accepted that the traditional interface offered to operators does not support them sufficiently, nor is it flexible. It is believed that a layered component-based structure architecture with the user interface as the highest level should be used instead. The user interface must be flexible and configurable (by the operator) to the task being undertaken.

A key element of visualisation is the interface by which the human interacts with the data and includes both the "how" as well as the "what, when, where, and why" of information presentation and control. Visualisation technologies include search engines, algorithmic processes, display and control devices, but the overall visualisation criteria is how these technologies enhance and allow the human to do his job [2,3,5]. A Nato IST-21/TG-007 representation of the overall process [7] is shown figure 2.
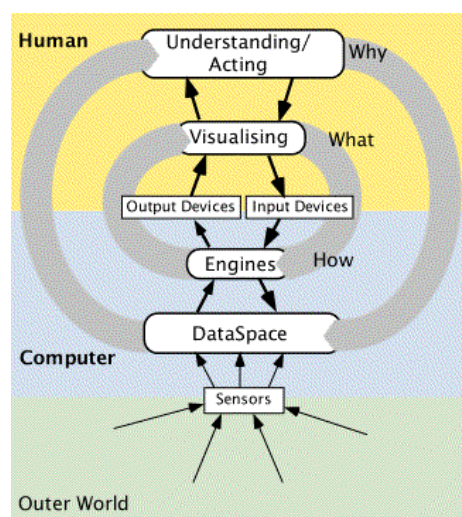
Figure 2: Nato IST-21/TG007 Visualisation Model

Master Battle Planner

The Master Battle Planner (MBP) is a prototype developed by DERA as a result of a study into the operational process of the UK CAOC (Combine Air Operation Centre). A technology gap was identified within the process and the MBP was developed to replace a single, manual procedure in developing the Master Air Attack Plan.

Existing air battle planning systems and CTAPS/TBMS operate on Unix platforms, and make use of large relational databases. At present the displays presented to the operator are still intended to mimic the layout of the database tables, i.e. rows of textual information.

The development of the MBP prototype investigated methods of improving the user interface. It was implemented as a map based system. As far as possible the system was designed to have the look and feel of a standard PC application.

By reducing the fidelity of information, e.g. the characteristics of aircraft and airbases, the need for a large database was removed. This, plus the intuitive design of the user interface, means that the lead-time in populating a scenario for a given operation can be drastically reduced.

A PC implementation also drastically reduces the hardware costs of the system. Whereas CTAPS/TBMCS require a minimum of 9 Unix servers supporting any number of Unix workstations, plus software licences for databases and graphics applications, the MBP can run on a single standard PC, or laptop, with the Windows operating system. This is an important consideration when deploying systems in theatre. A PC can be replaced at significantly less cost and overhead than a Unix platform.

MBP Functionality

The MBP is used to develop an Air Operations Plan. The system also provides the functionality to assist in the development of a defensive plan with the placement of CAPs (Combat Air Patrols) and AEW (Airborne Early Warning) situations.

It provides three stages to the planning:
- Visualise the scenario (figure 3)
- Produce the first cut plan(s) including packages and missions (figure 4) schedules
- Analyse and refine the plans (figure 5, and figure 6)

Visualisation is effective for achieving situation understanding. The scenario can be readily depicted, showing important information such as geographic locations, timing of flight paths, threats, etc. Figure 3 shows an example of this.

Representation of plans is important. Figure 4 shows the first cut plan, it provides key information such as the allocation of available resources and the management of the tasks, etc. It is possible, at a glance, to see if enough resources are available, any overlap or over tasking, etc.

Finally, a preview of the plan is available to analyse the planned mission, figure 7. This is achieved by using a play-mode so that the entire mission or particular package can be rehearsed (visualised) to ensure the success of the planned mission. This preview visualisation shows the mission in motion, it shows the interactions and brings out any mistakes or oversights.

The system can be used in two environments. The first is a large air campaign scenario where a CAOC is in operation for planning operations. In this scenario, the number of aircraft involved requires that high-level planning take place to define COMAO (COMposite Air Operation) packages etc. It is intended that the output from this process will be an ATO (Air Tasking Order) shell. The shell ATO can assist in the generation of the more detailed ATO outputs using available planning tools such as CTAPS or the Nato ICC (Integrated Command and Control).

Figure 3: Scenario

Figure 4: Plan of all missions

Figure 5: A Mission Plan
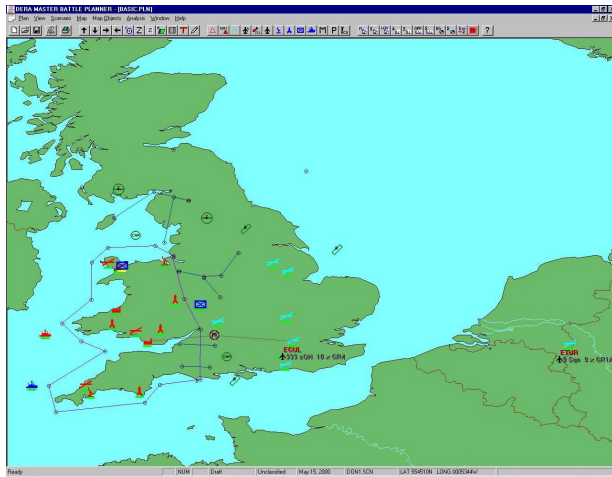
In the second operational environment, the system will be used in a small scenario with a small number of Air Units. This negates the need for a complex planning suite such as CTAPS or the ICC and the MBP tool will provide the required functionality to plan Air Operations.
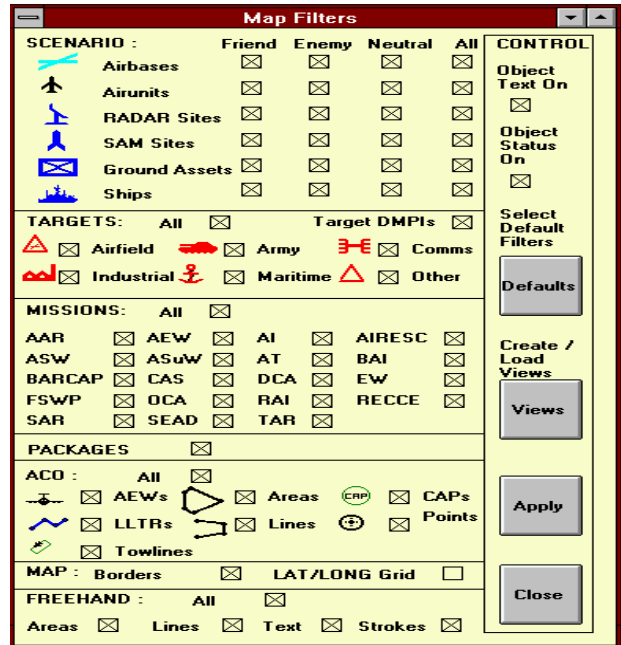
Figure 6: Mission Information
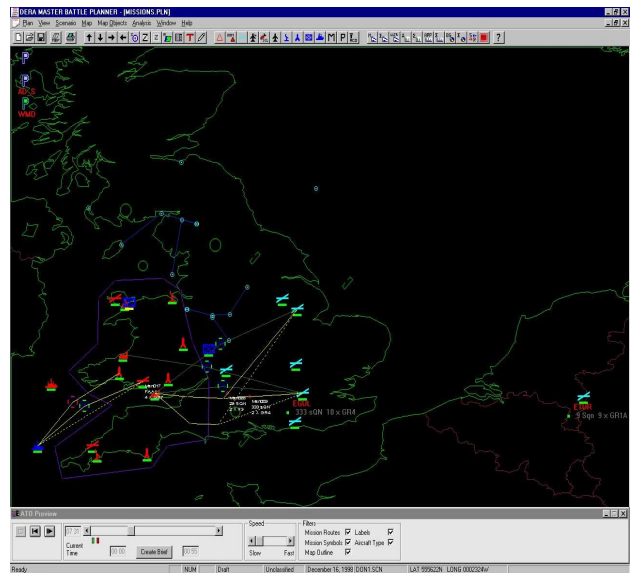
Figure 7: Preview of Mission Plan

Mission Plan

The output from the MBP system will contain sufficient information for it to be disseminated directly to the Wings or lower levels of command. The plans are produced in various formats:

- USMTF
- File importable by MS Office

An example ATO is shown below, it shows the exercise identification (DAIMON) followed by detail of the tasking for each unit. This can be up to 200 pages. During the Kosovo operations, ATOs were several hundred pages long, while ATOs produced during the Gulf campaign were so large that box loads had to be transported to the commanders.

EXER/\\DAIMON\users\hallam\Scenario
Backup\tfm.ATO//
MSGID/ATOCONF/-//
PERID/290000Z/TO:300000Z//
AIRTASK/UNIT TASKING//

TASKUNIT/15SQ/ICAO:LEUC//
MSNDAT/M004/1/OBERON/2GR1/SEAD/-/-/-/32222//
REFUEL/TARTAN67/M001/ESSO/ALT:190/291140Z/0
/<NOFREQ>/<NOFREQ>//
1MSNRTE/NAME/ENTRY TIME/ENTRY PT/EXIT
TIME/EXIT PT/TAS/ALT/INGRESS/291159Z/-
/291209Z/-/ALT:070/-//
ROUTE/291222Z/551400N0015700W//
ROUTE/291224Z/550200N0022000W//
ROUTE/291228Z/550800N0030000W//
ROUTE/291231Z/552000N0032800W//
ROUTE/291235Z/545200N0040300W//
ROUTE/291241Z/551300N0045300W//
ROUTE/291245Z/551300N0054000W//
ROUTE/291247Z/552200N0060000W//
ROUTE/291250Z/554700N0060000W//
ROUTE/291252Z/560700N0063000W//
TGTLOC/291254Z/291254Z/IONA/UNK/561900N0062
200W/-/IONA//
ROUTE/291256Z/563200N0055700W//
ROUTE/291258Z/562800N0053600W//
1MSNRTE/NAME/ENTRY TIME/ENTRY PT/EXIT
TIME/EXIT PT/TAS/ALT/EGRESS/291318Z/-
/291326Z/-/ALT:070/-//

The MBP system enables an operator to build a battle scenario containing airbases, targets, air units, aircraft types, ships, targets, radars, SAM sites, ground units, airspace measures and weapons configuration, using simple dialogs and point and click techniques for object placement on a map background (figure 6). The operator can then plan individual air missions or more complex COMAO packages using a drag-and-drop of objects on maps and data entry in dialog boxes. The system provides the operator with analysis tools to enable the planned operations to be assessed for the best utilisation of resources.

Combat Campaign Assessment

It has been recognised that in order to reduce the OODA cycle time it will be beneficial for the MBP to have direct

mission assessment support, so that the planning can be based on up-to-date information on the battlefield in relation to the executed missions.

The aim of the current Combat Campaign Assessment Component research is to investigate and develop technology to create an adaptive, decision-centred, visualisation environment for UK joint force commanders [9]. The commanders will have at their disposal a vast array of sensors, data sources and geographically distributed expertise. They will also be presented with dynamically updated models of the battlefield situation along with a suite of automated planning and decision-making tools. Military success will depend upon the commanders' ability to assimilate this information to understand and control the battlespace.

Vertical visualisation is defined to follow the chain of command. It will allow everyone in the same domain, e.g. in the air domain, to be aware of targets, threats and intentions that will have a direct effect on the deployment of the air forces. This can be achieved by presenting a filtered picture, i.e. a visualisation of the theatre airspace. A similar filtering mechanism can be used to provide a relevant picture to the maritime and land domains [4].

Horizontal visualisation will allow the component commanders to collaborate in Joint strategic planning. Currently there is no tool support to allow the Component Commanders to visualise the progress of a Joint campaign. Provision of accurate, real-time friendly location and combat status information will allow collaborative monitoring and will assist the disparate services to plan and execute a Joint operation towards a common aim.

It is necessary to have secure and responsive information that is available to the right user when needed, i.e. the right information must be delivered at the right time at the right place and in the right format [1,7,8].

Experimental Results

The development stage of the programme has been using an ICCS database. The initial aim has been to visualise the various component of an ATO especially what was planned and what was achieved. This enables the comparison/assessment of the accomplished mission's achievement.

The screenshot of the database, figure 8, shows the task components that were to be visualised and analysed for the next phase of the mission planning. They include:

- ATO_ID
- Mission Number
- Airborne
- Cancelled
- Lost
- Succ

- Unsucc
- Rcancel
- Rlost

Figure 8: Screen shot of the experimental database
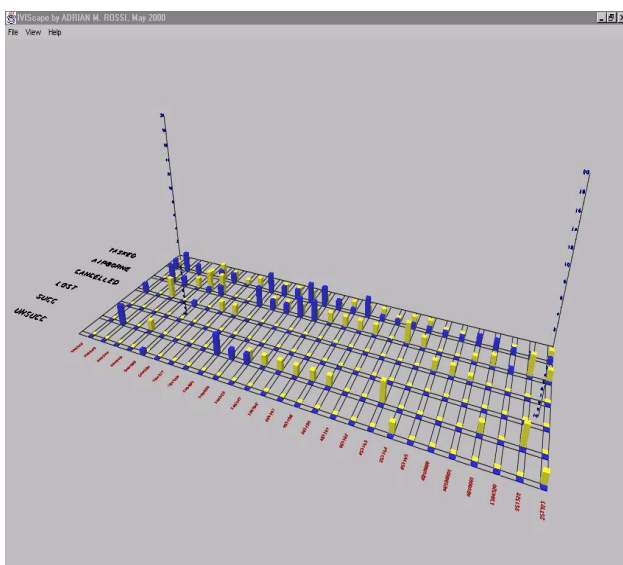
Figure 9: Visualisation of accomplished ATO (view 1)
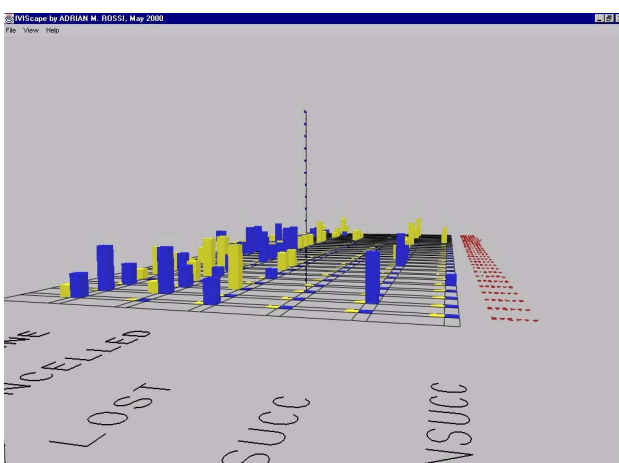
Figure 10: Visualisation of accomplished ATO (view 2)

The visualisation in figure 9 and 10 show the planned mission in blue and what is accomplished in yellow. At a glance one can see that what has been achieved differs from what was planned.

Conclusion

Initial results show that the developing Combat Campaign Assessment visualisation tool has produced encouraging results in providing information on the status of the completed missions within each Air Tasking Order. More work is required to integrate it into the MBP so that a real time mission assessment capability can be made available within the MBP. Thus closing the OODA loop and shorten the command cycle time.

References

[1] Andrews, K. Visualizing Cyberspace: Information visualisation in the Harmony Internet Browser, Proceedings of the IEEE Symposium on Information Visualisation, IEEE CS Press 97-105, 1995.

[2] Benbasat, I. & Todd, P.A., An Experimental Investigation of Interface Design Alternatives: Icon vs Text and Direct Manipulation vs Menus. International Journal of Man-Machine Studies 38(3): 369-402. 1993.

[3] Bennett, K.B & Flach, J.M., Graphical displays: Implication for divided attention, focused attention, and problem solving. Human Factors, 34(5),513-533, 1993.

[4] Herman, I, Delest, M. and Melancon, G., Tree Visualisation and navigation clues for information visualisation, Computer Graphics Forum, 17(20, 153-165, 1998.

[5] Lim, K.H, Benbasat, I., and Todd, P.A., An Experimental Investigation of the Interactive Effects of Interface Style, Instructions, and Task Familiarity of User Performance. TOCHI, 3(1),: 1-37, 1996.

[6] User Manual for the Master Battle Planner, DERA/LS/(SEC-M)/MBP/SUM/2.10, Ellis, M.R.K, 1997.

[7] Nato IST-21/RTG-007 Visualisation of Massive Military Datasets - Human Factors, Applications and Technologies - Part 1. 1998. http://www.visn-x.net/.

[8] Rossi, A. & Varga, M.J., Visualisation of Massive Retrieved Newsfeed in interactive 3-D, Proceedings of Information Visualisation Conference, 1999.

[9] Varga, M.J., V Outline Management Plan for Campaign Combat Information Management for Future Command, CRP/2000/S&P/SPI/17, March 2000.

[10] Weidenbacher, H. J. & Barnes, M. J., Target search in tactical displays with standard, single cue, and redundant coding. Displays, 18, 1-10.1997.

[11] Wentz, L., C4ISR systems and services in L. Wentz Lessons from Bosnia: The IFOR Experience. National Defence University, Washington, D.C., pp 355-368,1998.

This page has been deliberately left blank

_____

Page intentionnellement blanche

BARS: Battlefield Augmented Reality System

Simon Julier, Yohan Baillot, Marco Lanzagorta, Dennis Brown, Lawrence Rosenblum[1]
Advanced Information Technology (Code 5580)
Naval Research Laboratory
4555 Overlook Avenue SE
Washington, DC 20375, USA

Abstract

Many future military operations are expected to occur in urban environments. These complex, 3D battlefields are extremely demanding and introduce many challenges to the dismounted warfighter. These include limited visibility, lack of familiarity with the environment, sniper threats, concealment of enemy forces, ineffective communications, and a general problem of locating and identifying enemy and friendly forces. Better situational awareness is required for effective operation in the urban environment.

We believe that situational awareness needs cannot be met using traditional approaches such as radios, maps and handheld displays and more powerful display paradigms are needed. We are researching mobile augmented reality (AR) through the development of the Battlefield Augmented Reality System (BARS) in collaboration with Columbia University. The system consists of a wearable computer, a wireless network system and a tracked see-through Head Mounted Display (HMD). The user's perception of the environment is enhanced by superimposing graphics onto the user's field of view. The graphics are registered (aligned) with the actual environment. For example, an augmented view of a building could include a wireframe plan of its interior, icons to represent reported locations of snipers and the names of adjacent streets.

This paper describes the major challenges and the current implementation of BARS. In particular, we stress the need for high value graphical displays which provide the relevant, critical information for a user's current context. These displays should be precisely registered with the environment. There are three major research areas. First, an information distribution system is being developed which distributes to a
mobile user only a relevant subset of the common tactical picture. Second, to prevent information overload, we have developed an intelligent filter which selects and prioritizes the type of augmented information which is needed by a user's mission profile. Finally, high performance tracking and calibration systems are required to achieve accurate registration. We describe a general calibration framework that allows precision registration to be carried out in the field.

Introduction

Many future military operations are expected to occur in urban environments [CFMOUT-97]. These complex, 3D battlefields are very demanding and introduce many challenges to the dismounted warfighter. First, the environment is extremely complicated and inherently three-dimensional. Above street level, buildings serve many purposes (such as hospitals or communication stations) and can harbor many risks (such as snipers or mines) which can be located on many floors. Below street level, there can be a complex network of sewers and tunnels. Second, the cluttered environment makes it difficult to plan and coordinate group activities. In narrow, crowded streets it is virtually impossible for all members of a team to be in direct line of sight of one another. Third, the urban environment is highly dynamic and constantly changing. Dangers, such as the positions of snipers can change continuously. Furthermore, the structure of the environment itself can evolve. For example, damaged buildings can fill a street with rubble, making a once safe route impassable. These difficulties are compounded by the need to minimize the number of civilian casualties and the amount of damage to civilian targets.

In principle, many of these difficulties can be greatly reduced by providing greater situational awareness to the individual combatants. For example, if a user were shown the location of other members of his team, planning and coordination could be greatly simplified. A number of research programs exist which are testing digital maps or "rolling compass" displays [Gumm-98]. The United States Marine Corps, for example, tested a system called SUITE in the 1998 Urban Warrior Advanced Warfighting Experiment. SUITE was composed of a small laptop computer, a GPS receiver and a radio modem. The system provided users with a continuously updated map display. Although these types of displays have many advantages, including detailed information about the environment (through maps) which provide automatic and continuous updates of the location of entities in the environments, there are a number of important limitations. First, a map is an inherently two-dimensional display whereas the urban environment is inherently three-dimensional. Second, a user must switch attention between the environment and the handheld display. To overcome these difficulties, we propose the use of Augmented Reality (AR).

Figure 1: Image captured from a see-through augmented reality system. Various computer-generated annotations are overlaid directly on objects in the user's environment.

AR is, in effect, a "heads up display". The position and orientation of the user's head is tracked. The user wears a see-through head mounted display. Computer graphics are drawn into the display and these graphics align with objects in the user's environment. An example of the output from the prototype BARS is shown in Figure 1. Computer graphics are overlaid on a real building to annotate various structural features (such as windows) from a previously established database.

The first successful mobile augmented reality system was the Touring Machine [Feiner-97]. The system provided a user with labels and information about the Columbia University Campus. BARS, which builds directly upon this work and is being developed in collaboration with Columbia University [Höllerer-99], seeks to greatly extend this functionality to provide the user with high value graphical displays which provide the relevant, critical information for a user's current context. In [Julier-99], we argued that three major research thrusts had to be addressed: tracking (estimating where the user is located), user interface design (what the user sees) and user interaction (how does the user make requests, reports and queries from the system). In this paper, we describe the current implementation of BARS which partially addresses the first and second research challenges. The next section describes the information management system. We discuss the environment model and describe both the environment and the data distribution mechanism which propagates reports between multiple users. To prevent information overload, an information filter has been developed. This mechanism selects and prioritizes the type of augmented information which is needed by a user's mission profile, ensuring that only the most relevant information is presented to the user. The tracking and calibration framework is described next. Finally, we outline the current BARS prototype.

Information Management System

The information management system is responsible for describing the environment and disseminating this information to the remote user. It is built designed with the following assumptions:
1.   Any object of any type can, at any time become sufficiently "important" that it must be highlighted by the system.
2.   Certain types of objects (such as the location of enemy forces) are extremely important and should be known by all users all the time.
3.   Some objects (such as way points or objectives) are only critical to the mission profile of a particular individual.
4.   If an object has no "special properties", it should exhibit the following default behavior. The environment surrounding the user is known in the highest detail possible. As distance increases, the user might want to know progressively less and less information. At a significant distance, might only have the critical landmarks as well as the locations of known friendly and enemy forces.

Database Structure

All objects in the environment are considered to be "first class" entities which have a separate, identifiable

name, location and size. Examples of entities include physical objects (such as building, tree, tank, road, warfighter), spatial objects (such as areas or regions) and logical objects (such as waypoints and routes). The objects are organized hierarchically using the concept of containment. The top-level is a *City* entity which contains all other objects in the environment. The city entity contains buildings, streets, sewer systems and force units. In turn, each of these entities contains other sub-entities. A building, for example, can contain walls, floors, windows and doors.

Data Distribution

The design of the data distribution mechanism is guided by two important properties: the bandwidth is finite and the scheme should be robustness to network failures[2].

The basic approach is illustrated in Figure 2 and is an extension of the distribution concept initially described in [Brown-98]. The network is viewed as a collection of software objects which are non-fully replicated. When a particular object (such as a mobile user or a report of a sniper) is to be created, a single "master" copy is created on one computer which forms part of the information network. Although any computer can serve as the "master", we expect to only create these at a remote computing site. When another system wants to know about that object, it creates a "ghost" (or non-fully replicated) version of the master entity. The ghost version can be a highly simplified version of the master object. For example, the master might contain detailed track history information whereas the ghost could simply be a report of current location and an uncertainty ellipse. The ghost copies are used to perform tasks such as updating the graphics display. When the master entity changes, it automatically sends updates to all of its ghost copies. When a remote system wants to change the master, it sends a request to the master entity which processes the state change request and broadcasts the result.

This system meets the two requirements described above. First, the required bandwidth is reduced because a system only receives data about its ghost entities. Second, because each remote system maintains its own "ghost" entities, only the changes to these entities need to be distributed, greatly simplifying the type and kind of information which will be transmitted. Furthermore, these updates are

automatically pushed and there is no need to use a data polling mechanism. Finally, if a network connection fails, the remote system does not receive further updates of its ghost entities. However, these entities still exist in the remote system and can be used, for example, for display purposes.

An important feature of this system is that it must decide *what* objects will be distributed and *how* a system, when it enters the network, discovers what objects are available. This is achieved through the use of the Connection and Database Manager (ConMan). The ConMan knows on which machine each master entity and its ghosts are maintained, and the spatial position of each entity in the environment. Based on aura interactions [Greenhalgh-95] and arbitrary rules we can add to the ConMan, it directs objects to initiate or cease communications between themselves. Thus, communications are only set up between entities that need to know about each other, eliminating unnecessary network traffic.
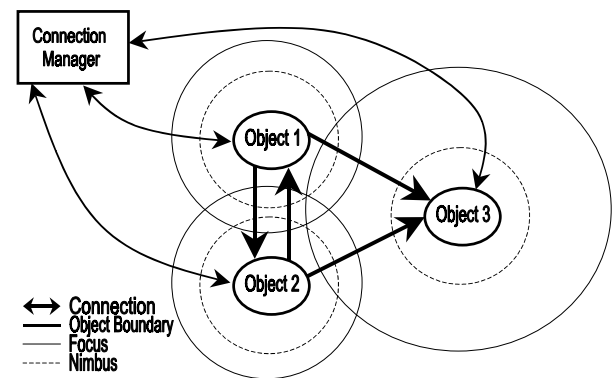
Figure 2: The Data Distribution Mechanism. The environment consists of three entities (Objects 1, 2, and 3). The focus of each entity is shown by a solid circle, and its nimbus by the dashed circle. Connections between each entity are brokered by the Connection Manager (as indicated by the lightweight connection arrows). Because the focus of Object 1 overlaps the nimbus of Object 2 and vice-versa, each entity contains a ghost copy of the other. The arrows show connections that are used to update the copies. Object 3 is a "stealth" viewer – its focus intersects the nimbuses of Objects 1 and 2 and so it receives copies of these objects. However, neither Object 1 nor Object 2 creates or receives updates of Object 3.

Entities may also join the simulation simply by existing in a relational database which is linked to the ConMan. The ConMan analyzes the entities stored in the database and calculates their nimbuses. When the ConMan determines that an entity in the simulation needs to interact with an entity in the database, the ConMan instantiates the object based on the information in its database record. For example, in a

---

[2] This research work does not focus on the design of robust data transport layers because a number of such layers (for example that used in SRI's InCON system[Seaton-98]) have already been developed. Our work focuses on what data will be transmitted over these layers.

simulation of a city, information about all buildings is stored in the database. The ConMan determines that the new mobile user needs to be aware of some buildings nearby, instantiates new objects for those buildings, and connects the user object with the building objects. As the user moves through the environment, he may encounter more buildings that will be instantiated by the ConMan. The instantiations of objects can be considered a collective write-through cache for the database. When an entity (that has a record in the database) is changed, the change is written to the database.

Figure 3: The effect of clutter. The top picture shows the view of a building. Several other buildings are visible behind, making the display extremely cluttered and confusing. The bottom picture shows the same dataset when the information filter is enabled.

Information Filter

Once the information has been distributed to a mobile user, the system must still plan and coordinate what information must be shown. The reason is illustrated in Figure 3. If the system simply shows the user all information which is known about the environment, the result can be a highly cluttered display which is difficult to interpret. To overcome this problem, BARS utilizes an intelligent information management filter to decide which entities have the highest priority and must be shown to the user [Sestito-00]. The filtering is a logical extension and refinement of the aura used with the object distribution.

To help the decision process and to take into account the variety of situations that can be encountered, the user can select a filtering mode according to his current mission. Current mission modes are: **stealth**, **reconnaissance**, **route**, and **attack**. So, for example, the route mode will show the user the position of enemy and friendly forces, his destination point, hazardous materials found in the way, zones for potential enemy ambushes. In addition to these basic modes an individual user is able to decide to increase or decrease the importance of certain objects.

Once the filtering mode has been selected, the filtering process is done in three stages:

1. In the first stage, all the objects that are of critical importance to the user are shown at all times. Objects in this category include enemy forces and hazardous zones such as mine fields. All the objects that are selected by this filtering stage are sent straight to the graphics system to be displayed to the user.
2. In the second stage, objects that are important to the mission are selected. For example in a route or reconnaissance mission the names of important streets and buildings are selected. The importance of each object is described in the database. A heuristic system based on the current Army field manuals for operations in urban terrain is used to designate importance values to each object. Due to the complexity of the missions being carried out, the complexity associated is a multi-dimensional vector. So, one component has the tactical importance for an offensive operation of the object, another component has the importance a civil target, and so on. While this filtering reduces the number of objects that are of potential interest to the user, a large number of objects are still selected. Thus, a third filtering stage is necessary.
3. In the third stage, the objects selected by the second stage are filtered again, but now according to their "region of influence" (RI). The RI specifies the volume in space where an object is of relevance for the mission, or the region where the object has any kind of influence in the development of the mission at hand. The region of influence is calculated by a heuristic formula in the BARS filter. It is a dynamical entity which

value depends on the mission mode, the importance of the object and other factors. If the object has been selected by the stage 2 of the filtering process, and its region of interest intersects the user's one, the object is displayed, otherwise, it is being held until the user steps into the object's RI.

The objects that have been selected at the end of the process are send to the graphic management system that displays the information to the user. The graphical management system is being designed to also determine in what part of the user's field of view the information will be displayed. This way, information will not be clustered in the center of attention of the user and it will be less distracting.

Calibration system

Once the set of objects have been determined, they must be drawn in such a way that they correctly align with the real world. The user viewing direction in the virtual world is determined at each frame using the position and orientation measured by the corresponding position and orientation sensors. Because of the characteristics of the display (such as field of view), the properties of the trackers (such as biases) and due to the fact that each user wears the display slightly differently, a precise calibration system which can be applied, while the user is in the field, must be used. Traditional approaches have used specially designed indoor environments. These are not appropriate, and a new framework has been developed. The framework solves two calibration parameters simultaneously. The first is the transformation from the world referential to the base referential and is equivalent to calculating the mapping the report from a sensing device to the true position and orientation of the sensor. To a first approximation, these parameters tend to be constant for a device and can be usually extrapolated from the tracker manual. For example, our inertial sensor measures orientation with respect to a referential having an axis aligned to the earth magnetic field and another axis pointing up. In this case the referential depends upon one's location on the earth. The second transformation maps the sensing unit referential to the viewpoint referential attached to one of the user' eyes. It depends on the way the sensing unit is attached to the HMD and the way the HMD is worn. To solve this unknown, each user must calibrate the system prior to using. This calibration is achieved by displaying in the user's head mounted display a wireframe representation of certain features in the

environment. The user turns their head until the virtual representation aligns with the real-world. At this point, the transformation can be determined and accurate registration is achieved.

Experimental System

A fully functioning prototype of the outdoors part BARS system has been implemented. The hardware system, is portable, wearable, and can operate both indoors and outdoors. The software system lets the user see information about the environment (e.g. building names and locations) superimposed upon the real world.
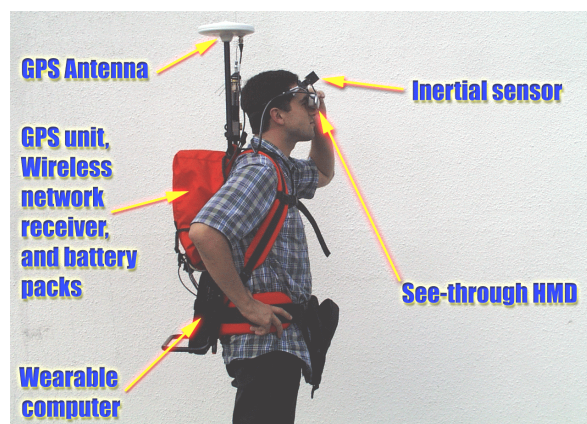
Figure 4: The hardware for the prototype BARS system.

Hardware Description

The system is composed of several interconnected off-the-shelf components which are shown in Figure 4. These components are:

- **A Dell Inspiron 7000 laptop.** This is a 366MHz Pentium II-based laptop computer which carries out the major tasks (data distribution, filtering, scene generation). Our initial trials suggest that, with a dedicated graphics card, a system with considerable less computational resources is sufficient.
- **Ashtech GG24-Surveyor GPS receiver**. This is a dual constellation (uses both the US NavStar and Russion GLONASS satellites) kinematic differential GPS receiver. With a base station, it is capable of providing (in clear areas), position measurements with centimeter level accuracy.
- **An InterSense IS300Pro inertial tracker**. This device uses gyroscopes and compasses to precisely determine the orientation of a body with

respect to a reference referential oriented along the earth magnetic field. The compasses are used to correct for the drift occurring with gyroscopes over time

- **A FreeWave Radio Modem**. This is capable of transmitting 115kbits/s over long ranges (greater than 20 miles). In an urban environment (NRL), we have demonstrated its capability to transmit from the interior of one building through two other buildings to a remote site over 50 m away. The data are fed to and read from the units using a serial connection at speeds up to 115 Kbps. A TCP/IP connection is opened between each radio pairs to allow for multiple channels of data through the same link using multiple socket connections. Currently the only information driven by the link is the correction messages emitted by the fixed GPS receiver. Later, the base station will transmit local versions of the database to the mobile units as needed. The mobile units will emit location information and will, in turn, receive the database part required. Additionally, changes to the global database will be possible allowing update messages from a remote station to the base station.

- **A Sony Glasstron Head-Mounted Display (HMD).** This provides a lightweight high-resolution solution to display graphics superimposed upon the real world for the BARS prototype. The display has its own battery and can display screens of resolution SVGA. The display is connected to computer via the SVGA port replicating the screen.

Portable batteries power all components. This enables the whole system to be wireless and to be transported by an individual. The processing unit (a laptop computer) is communicating with the GPS receiver, the radio unit, and the inertial tracker using a serial link. The BARS is composed of a stationary unit called base station and remote units transported by users in the field called mobile user.

The architecture of the system is shown in Figure 5. A typical output from the prototype BARS is shown in Figure 6.
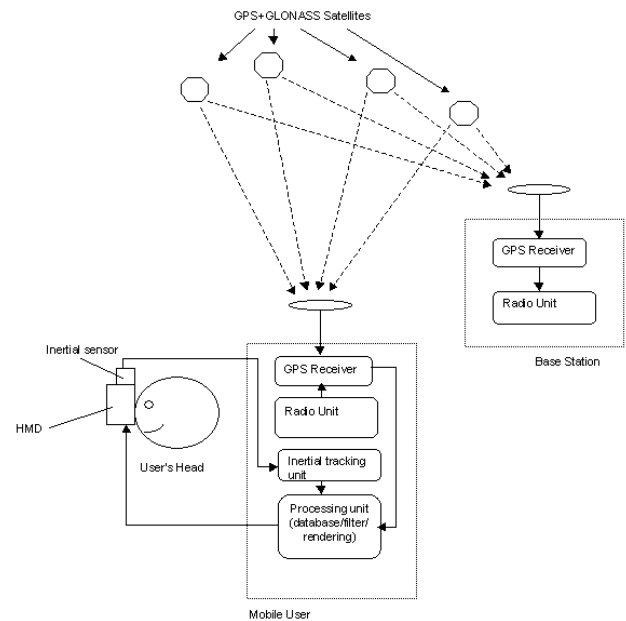
Figure 5: Architecture of the prototype BARS.

Summary

Augmented Reality has the capability to significantly change the way in which information can be delivered to the individual warfighter. Through registering high value graphical displays which provide the relevant, critical information for a user's current context, we believe that situation awareness can be greatly improved, leading to faster and more informed decisions. This paper has described some of the research issues and the current progress of the BARS.

Acknowledgements

References

[Brown-98] D. G. Brown, "An Architecture for Collaborative Virtual Environments With Enhanced Awareness," M.S. Thesis, University of North Carolina at Chapel Hill Department of Computer Science, 1998.

[CFMOUT-97] Concepts Division, Marine Corps Combat Development Command, "A Concept for Future Military Operations on Urbanized Terrain," approved July 1997

[Greenhalgh-95] Greenhalgh, Chris, and Steven Benford. 1995. MASSIVE: A Collaborative Virtual Environment for Teleconferencing. *ACM Transactions on Computer-Human Interaction* (September), 1995.

[Feiner-97] S. Feiner, B. MacIntyre, T. Höllerer and T. Webster, "A Touring Machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment", *Proceedings of the International Symposium on Wearable Computers, Cambridge MA,* October, 1997.

[Gumm-98] M. M. Gumm, W. P. Marshak, T. A. Branscome, M. Mc. Wesler, D. J. Patton and L. L. Mullins, "A Comparison of Soldier Performance Using Current Land Navigation Equipment With Information Integrated on a Helmet-Mounted Display", *ARL Report ARL-TR-1604, DTIC Report 19980527 081*, April, 1998.

[Höllerer-99] T. Höllerer, S. Feiner, T. Terauchi, G. Rashid and D. Hallaway, "Exploring MARS: developing indoor and outdoor user interfaces to a mobile augmented reality system*", Computers and Graphics*, Vol. 23, 779-785, 1999.

[Julier-99] S. Julier, S. Feiner and L. Rosenblum, "Augmented Reality as an Example of a Demanding Human-Centered System", First EC/NSF Advanced Research Workshop, 1-4 June 1999.

[Seaton-98] S. Seaton, "InCON: A New Concept in Real-Time Resource Management", talk presented at JavaONE Conference, March, 1998.

[Sestito-00] S. Sestito, S. Julier, M. Lanzagorta and L. Rosenblum, "Intelligent Filtering for Augmented Reality", *Proceedings of SimTecT 2000*, Sydney, Australia, Feb 28-Mar 2, 2000.

Figure 6: Output from the prototype BARS. The top picture shows an output from the system, the bottom shows an overview map display.

This page has been deliberately left blank

Page intentionnellement blanche

A Framework for Multidimensional Information Presentation using Virtual Environments

Sarah Monique Matzke, Ph.D.
and LCDR Dylan D. Schmorrow, Ph.D.

Office of Naval Research – Code 342
800 N. Quincy St.
Arlington, VA 22217
USA
schmord@onr.navy.mil

A variety of Virtual Environment (VE) systems exist today and there are as many terms to them. Synthetic environments, virtual reality (VR), and VEs, to name a few, all refer to a simulation of some operational environment. The gamut of VEs runs anywhere from desktop displays to fully immersive and interactive scenes. What these systems share is the need to capture, process, and present data. Presented effectively, the information can enable a user to make timely and informed decisions.

This paper will provide some examples of current VE technology and present a framework for VE systems. The paper will also address some of the future directions and challenges facing the evolution of VE technology.

1. Current VE Systems

The Office of Naval Research (ONR) sponsors basic research and advanced technology demonstrations in the area of VEs for training. The ONR VE program was founded partially upon the need for affordable, compact, deployable, and reconfigurable training systems. Additionally, the VE program was created to address reductions in manning and increasing training requirements. The training devices demonstrated to date include a pilot trainer for underwater remotely operated vehicles (ROV) and a trainer for submarine and shiphandling.

The ROV trainer is a desktop VE and illustrates at least three important characteristics of VEs. First, performance tests showed that trainees displayed a *positive training transfer* from the desktop training to actual operation. As will be discussed later, the utility of VE systems must be demonstrated in actual operative results. Second, the system, originally built upon an SGI platform, has been transferred to a PC version - an important step in developing portable systems for ease in *deployability*. Last, the desktop system has the potential for use in other types of remote operations and therefore meets the criterion of *reconfigurability*.

Another VE demonstration, the Conning Officer Virtual Environment (COVE) project at the Naval Air Warfare Center Training Systems Division, was designed to train shiphandling skills. This immersive VE incorporates a wide variety of tools including a head-mounted display (HMD), head tracking devices, voice recognition, cognitive modeling of spatial knowledge, and various task analyses.

Future systems will benefit from COVE for a number of reasons. For instance, performance measures, including metrics to compare performance of trainees using HMD versus using regular desktop displays, will be developed. Those types of analyses will help clarify the added value of immersive VE versus traditional training methods. Additionally, to the extent that the technology is available, the COVE project is integrating commercial off-the-shelf products, providing a more *affordable* and *reconfigurable* system than one designed with custom-built components. Finally, the COVE project will provide a platform for other vehicle operations including small craft and amphibious landing craft.

2. System description

As evident in the VE systems above, the general approach to the developing VE systems is multidisciplinary. The designs include aspects of psychology, computer modeling and simulation, user-interface design, and robotics. They may also include sophisticated hardware such as sensors, tracking devices, HMDs, and large-scale projection screens. The desired complexity of the VE will determine the need for the various components. The following description provides an overview of some of the key components of a VE as seen in Figure 1.

Environmental data capture
Whether the device is used for real-time operations or for mission rehearsal, the data driving these systems are initially captured from the external (or operational) environment. As situations or tasks require, environmental data can be stored and used to generate future scenarios for training, mission rehearsal, or debriefing purposes. Regardless of the intended use, the systems should accurately represent the operational environment (or be able to quickly generate simulations of the external environment). The level of fidelity that the system can present will depend upon the capacity of the sensors, the extraction of relevant information, and the method of data presentation.

System requirements
- Sensors - Systems rely on sensors to detect and capture data from the external world (e.g., radar, sonar, etc.). A misrepresentation of the external environment can lead to the development of a faulty mental model for the user. Subsequently, conclusions and decisions based on faulty information may lead to actions with catastrophic consequences. Therefore, the sensors must accurately capture and represent the details of the operational environment.
- Transcription - The system must have the capability to extract relevant information and discard irrelevant information ("noise") captured from the external environment. There should be built in capabilities to modify the level of granularity at which the data are presented. In some situations providing fine

detail may be distracting, whereas an overall picture may be more useful. Those issues need to be addressed through task analyses and comparison studies.
- Presentation – After the data are captured and transcribed appropriately, they must be presented to the user so that the information is useful. Human factor analyses can help determine how to exploit human attention and present information effectively.

The various modes of presentation in VEs can include any combination of the following:
- Visual – The arrangement of visual displays, the content of the display, and the ease in which the user can access specific information can greatly affect how quickly and effectively the user comes to a decision. Weeding through irrelevant information can slow the decision making process and lead to more stressful situations, which in turn can increase the room for error.

- Audio – Studies will need to be conducted to determine in what cases sound enables a decision instead of or in addition to vision. Such issues are important to address because spatial audio, virtual sound, and the integration of virtual with real sound may be among the future components of VEs. A few of the areas researchers will need to examine include whether sound in VEs enhances situational awareness, increases the accuracy of localizing objects, or aids in judging location relative to sound sources (i.e., to aid in navigation).

- Haptics – Haptics refer to the sense of touch derived from contact forces. Haptic interfaces must convey a sense of touch to a user exploring the environment to create a feeling of immersion. Haptics can aid in tasks such as locating and operating manual controls as well as providing pilots with haptic input for signaling their direction or location.

- The integration of visually, aurally, and haptically presented data will create more

realistic simulations for fully immersive VEs for training and other tasks. The various applications for multimodal, immersive VEs will be described below (see Section 3).

The human operator

These systems must be designed with the user in mind because certain limitations of human sensory processes will place restrictions on the capacity of the user to obtain useful information from the VE. The system design should also take into account the most effective user response. For example, will a verbal command be a more effective response than a keystroke or push of a button?

Feedback

Whatever the result of the user's action, the system should have the capability to provide feedback about the result. Feedback will enable the user to evaluate and learn from the exercise, whether it is real or training.

Networked systems

The above description applies to a single (VE) unit, however, multiple units will ultimately become networked over systems like local area networks or the internet. Networking VE systems will have an great advantage over traditional training systems in that they can be remotely operated, reach a greater number of trainees, and enable trainees to interact across multiple sites.

3. Applications

Military Training and Operational Tasks - VE systems will continue to be investigated for tasks such as expeditionary warfare in urban settings, mission critical duties, various types of reconnaissance (ground and air), artillery and surface fire support, and close air support (CAS). As mentioned previously, VE trainers will be applied to the piloting of ships, small craft, and amphibious landing craft. Individual VEs in the form of wearable, portable, and wireless devices have potential to facilitate maintenance tasks and tactical decision support.

Information centers – VEs and information displays (e.g., command information centers) will be combined to create multimodal, multifunctional workstations. As evidence of

designs for highly effective displays becomes available, the design can be incorporated into VEs and vice versa. To further such a joint venture, a decision support program sponsored by ONR will study through task analyses whether, for example, 3D versus 2D displays are more effective. The program will also use task analyses to determine what cognitive processes, what spatial arrangements, and what aspects of data are critical to decision support.

Commercial - The entertainment industry already uses VEs for a variety of functions, such as immersive video games. VEs are also implemented in telecommunications, information visualization, product design and manufacturing. Similarly, large-scale manufacturers can utilize the technology to prototype airplanes, ships, and other vehicles and test them prior to costly investment in the development of the actual product.

Medicine – VEs have a number of applications for medicine from planing neurosurgery[1] to using virtual patients to teach medical students how to deal with traumatic injuries[2] to teaching anatomy. Additionally, simulators have been created to deepen doctors' understanding of cancer-related fatigue[3]. Finally, modeling pharmaceutical compounds can aid in understanding their mechanism of action and facilitate the development of effective treatments.

Telecommunications - Items such as personal navigation devices or multimodal, wireless phones (3rd generation) will benefit from advancements in VE technology. Similarly, VE technology will benefit from advancement in wireless communications for networking VEs over multiple sites.

[1] Kockro, R. A., et al. (2000). Planning and Simulation of Neurosurgery in a Virtual Reality Environment. *Neurosurgery*, 46 (1), 118-137.

[2] Lurie, S. (2000). Innovation and service traditional at University of Michigan Medical School. *JAMA, 283 (7),* 865-866.

[3] Vastag, B. & Beidler, N. (1998). Tired out: patients find few easy answers for cancer-related fatigue. *Journal of the National Cancer Institute*, 90, 1591-1594.

4. Challenges

The promoters of VE technology have made several promises about the utility and importance of VEs. Before the developers can deliver on those promises, several technical and scientific issues must be resolved – some of which are outlined and described below.

- Visual - current visual displays are limited in terms of their physical and geometric fields-of-view, depth of focus, visual contrast, resolution, frame rate, weight, and display latency
- Representation of self (avatar) – methods to generate egocentric (a representation as one would normally view their body) or exocentric views need to be developed
- Direct interaction with objects (collision, control, manipulation) – haptic interfaces need to be improved
- Sound – methods to generate spatial audio and virtual sound without enormous expense need further development
- Multidisciplinary approach requires compatibility among system components (computer graphics, behavioral modeling, multimodal interfaces, etc.)
- Faster processors for real-time display of information and multimodal synchronization are needed
- More reliable and higher speed connections and multi-user interfaces are necessary for networking

In the context of training, VE systems must demonstrate efficacy in improving task performance. Further research is needed to determine the extent that 3D, immersive environments create greater situational awareness versus traditional methods. Effective metrics must be in place to evaluate the utility and efficacy of VEs as a teaching tool. Although the development of standard measures will require a concerted effort on the part of program managers and researchers, it is nonetheless achievable.

5. Summary

VEs are dynamic, complex, and powerful tools for information presentation. The capacity of VEs to supply multidimensional representations of data can dramatically change the manner in which those data are examined. The broad spectrum of application domains for VE technology has naturally generated interest from a number of disciplines. As such, one of the greatest challenges to advancing VE technology as whole is creating a collaborative environment where disparate fields can mutually benefit from one another and advance VE technology to its fullest potential.

6. General References

Durlach, N. I. & Mavor, A. S. (Eds.). (1995) *Virtual Reality: Scientific and Technological Challenges*. (Committee on Virtual Reality Research and Development, National Research Council). Washington, DC: National Academy of Sciences Press.

Macedonia, M. R., Zyda, M. J. & Pratt, D. R. (1995). Exploiting reality with multicast groups. *IEEE Computer Graphics & Applications*, *15,* 38-45.

Virtual reality comes of age. (1999). In *Funding a revolution: Government support for computing research.* Washington DC: National Academy Press.

Zyda, M & Sheehan, J. (Eds.). (1997). *Modeling and Simulation: Linking Entertainment & Defense.* Washington, DC: National Academy Press.
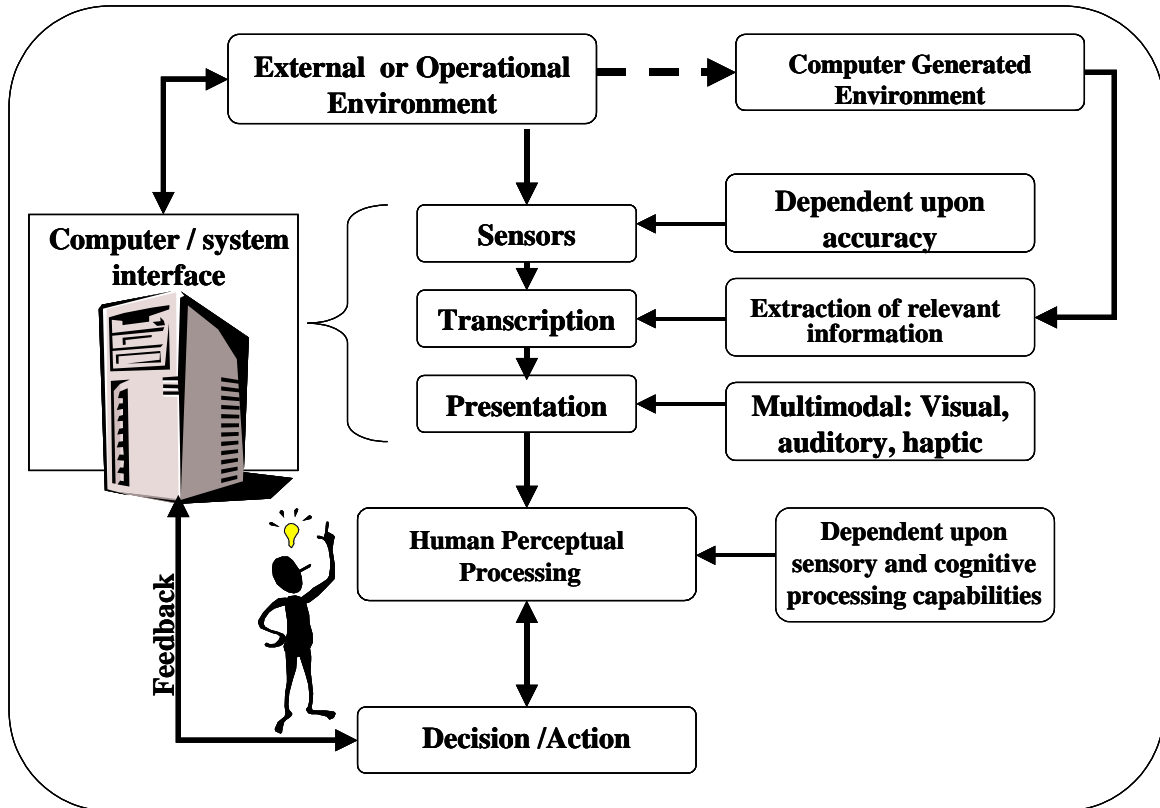
Figure 1. A framework for VE systems. See text for details.

This page has been deliberately left blank

Page intentionnellement blanche

Distributed Collaborative Virtual Reality Framework for System Prototyping and Training

Suleyman Guleyupoglu[‡]
Henry Ng

U.S. Naval Research Laboratory
4555 Overlook Avenue SW
Washington, DC 20375
United States of America

{suleyman, ng}@ait.nrl.navy.mil

Introduction

The significant impact Information Technology has made in the lives of most people in the last decade or two is undeniable. Coupled with other enabling technologies, many tasks get accomplished more efficiently and reliably nowadays. However, there are instances where the adoption of these technologies takes longer for variety of reasons. System prototyping and training of war fighters has been an area where enabling technologies can be better utilized.

One of the keys to the success of military operations with the least amount of casualties is a well-trained group of people fighting the war. It is also essential that they are well accustomed to the environment in which they operate and comfortable with it. To achieve this, live training would be ideal in almost all cases except where it would pose a significant safety risk. This, of course, is not possible in reality as the resources available for training is limited in the real world.

In this paper, we address the use of enabling technologies in the military and describe a framework that takes advantage of several state-of-the-art technologies to perform combat system design, evaluation and training. The framework uses virtual reality and distributed computing technologies to provide users (trainees, trainers, or combat system designers) an immersive environment to interact with the combat system and the other users. The framework allows participants to collaborate on the same mission or activity in the virtual environment without the need to be at the same geographical location.

There are numerous benefits to systems based on this framework. It is possible to bring together participants into the same virtual environment as opposed to doing this in the physical world that may be costly and sometimes difficult logistically. The cost of training is significantly cheaper since the initial cost and maintenance of computer equipment is usually less than the cost of real combat systems. The framework allows new weapon systems, for example a combat information center, to be built virtually so that users can conduct a walk-through and make suggestions on the design of the system. In other words, the end users can actively participate in the early design phase of a new combat system.

Virtual Engineering

We use the term *Virtual Engineering* to describe the engineering process primarily performed in a Virtual Reality (VR) environment. Virtual Engineering can be utilized at various phases of the development life cycle including requirements, design, development, testing and training. Virtual Engineering allows C4I concepts and combat systems to be evaluated before they are physically built, resulting in significant savings for the US Department of Defense.

There are a number of enabling technologies that we use to make Virtual Engineering effective (See Figure 1). At the top of the list is Virtual Reality. Virtual Reality technologies allow users to immerse themselves into the environment they are working in. What should be noted is that 3D graphics is not always Virtual Reality. What makes the experience almost real or virtually real is usually the human computer interface accompanied with realistic computer graphics.

[‡] Also affiliated with ITT Industries, Alexandria, Virginia, USA.

Modeling and simulation has found wide acceptance in many engineering fields. Using modeling and simulation, engineers can verify the validity of their design and optimize them. As such, modeling and simulation is an indispensable tool for engineers.

Distributed Computing or Simulations is another technology that is important for the Virtual Engineering of combat systems. Different simulators that the engineers need may have been designed to run on different hardware and they may be running at different physical locations. Therefore, the ability to link all these simulators seamlessly is significant for the Virtual Engineering infrastructure.

Parallel Processing is also an important aspect of modern engineering process. For high precision modeling or for relatively complicated problems, existing computing platforms can be used in parallel to solve problems faster than the alternative of using any one of the computers. Some problems are better suited for parallel processing than others are. However, significant run-time reductions can be obtained by parallel processing in general.

Finally, collaborative environments enable engineers to work on projects synchronously or asynchronously from geographically distant locations. Traditionally, the travel costs discourage teams of engineers who are at different cities or countries to get together and work together as frequently as they would, had they been at the same location. Collaborative environments offer a solution to the problem by allowing engineers to perform video teleconferencing and share documents or applications at the same time.

Virtual Engineering offers a number of advantages over traditional engineering practices. With Virtual Engineering, combat systems can be well understood before resources are committed for building the system. The users or the warfighters can be involved in all phases of the development process early on. Their feedback can save time and money by reducing development life cycle and total development time. Virtual Engineering serves as an integration tool for design, development, integration and testing of systems. In addition, it can be used to train warfighters on the system before the systems are actually built.

Virtual Engineering is well suited for Simulation Based Design and Acquisition (SBD/SBA) applications. Typical SBA life cycle can be divided into four phases: requirement gathering, design, development and testing (See Figure 2).

In the requirement phase, important issue decisions are made in the theater context. For example, what kind of force mix is considered, number of weapon systems to be used in the theater, their search and detection coverage, reaction time, fire power, capacity and weapon load.

In the design phase of the SBA, decisions are made in the platform context. For example, system configuration and connectivity decisions are made based on system requirements. These decisions are then used to determine the types of algorithms, logic and display components to use at element context. This requires simulations that start with simple logic for the requirement formulation phase in the early acquisition process. As the system engineering process progresses into design and development phases, the fidelity and number of entities in the simulation will be greatly increased to meet the needs of system design, development and testing. However, much of the model's fidelity often has to be sacrificed in order to reduce the computational turnaround time and meet the project schedule. In other situations, the fidelity of the model needs to be increased to gain more data and insights for the problem under study, increasing the model's running time. When running on a single processor the simulation faces the "Von Neumann Bottleneck". By taking advantage of recent advances in low cost, high speed computer chips, scalable parallel processors and new software languages, it is possible to have fast turnaround times while still allowing an increase in model fidelity and simulation scenario complexity. A composable simulation model which employs an optimistic parallel processing architecture in which the number of processing nodes can be scaled up or down as needed was developed to meet this need. This enables the analysts to run more experiments or parametric analyses to find the optimal system configuration design within a run time constraint.

The next phase of the SBA process is the testing of the system. An additional benefit of Virtual Engineering is that the testing system can also be used as a training tool for the end-users of the system and let the real users of the system provide feedback early in the life cycle.

It is important to evaluate different components with different characteristics in the development process. For example, to achieve a certain detection coverage characteristics a ship may carry different types and different number of radar systems. The simulation of these radar systems can help identify which system is

a better solution to satisfy the requirements. Being able run these simulations in parallel may accelerate system development time significantly.

To practice Virtual Engineering, an object-oriented set of classes has been developed at the Naval Research Laboratory (NRL).

Virtual Reality Framework

An object-oriented set of classes has been developed at the NRL to form a framework for Virtual Engineering and virtual reality applications. The objective of the framework is to allow rapid development of these applications with significant software reuse. We used this framework to demonstrate several concepts that we present in the paper.

Virtual reality is a natural interface between humans and computers, where communication can take the form of moving 3D imagery, sound, and even physical forces (from motion to touch). Virtual reality applications can be so responsive and render imagery in such high speed that the users can feel immersed in the synthetic environment. Many virtual reality applications share common tasks such as rendering geometric models, interfacing with the users, communicating with other users on the network, etc.

The Virtual Reality Framework (VRF) provides several classes that help carry out those tasks, independent of the type of application developed. For example, it implements classes that provide rendering functionality using SGI Iris Performer [1]. Similarly, it implements classes that let the application programmer to use High Level Architecture (HLA) Run-Time Infrastructure (RTI) for communicating with other networked applications or simulations with minimal understanding of the HLA RTI [2–4].

We present two applications that were built using the VRF.

Virtual Prototyping

We used VRF to develop a virtual prototyping application. This tool allows system designers to design and configure the equipment, such as combat consoles, large screen displays, to formulate the layout of room for the ship command center in a collaborative environment. System designers can run the software on a set of networked computers to work on the same system. This process is illustrated in Figure 3. Each person sees the workspace as shown in the lower left of the figure. They may have different types of user interfaces to interact with the software that provides different levels of immersion into the virtual environment. These interfaces range from a simple set of keyboard and mouse to a more advanced interface, such as a workbench or stereoscopic head mounted displays with hand and body movement tracking devices.

This can serve as a digital mockup of the ship command center. In addition, the components that are placed in the virtual ship command center is linked to simulations, so it is possible to interact with the consoles in the command center while running the simulations that the command center system depends on. In other words, it is possible to conduct virtual testing of the configuration using this virtual environment. The same system can also be used to train crewmembers. The costly physical mockup used for ship combat system design and development can be eliminated while reducing the development and testing time. More importantly, the system allows design experts from different disciplines as well as the end users (warfighters) to work as a team and to interact simultaneously during the design. Conflict and design errors can be detected, and resolved at the early stages of the design process. This is especially important because, as studies at the US Defense System Management College have shown, the majority of a system's life cycle cost is determined very early in the design phase of the program.

Having designed the ship command center in the virtual prototyping program, we are able to load the configuration file into another application for testing and training. We describe that application in detail next in a training tool context.

Virtual Combat Information Center Training System

Virtual reality technology has become a cost-effective training alternative with the advent of more powerful and cheaper graphics computers. This technology allows the user to be immersed into a simulated graphical environment where he or she sees the virtual environment through a computer monitor or a head mounted display and provides input through various devices such as a keyboard, mouse or more sophisticated alternatives. We used VRF to develop a distributed virtual CIC for the surface ship training. CIC crewmembers can then perform testing, team learning and training in a virtual ship environment. The virtual environment

offers a true interactive 3D view of the interior of the CIC as shown in Figure 4.

The visual simulation portion of the Virtual CIC is intended to provide the "look and feel" of the actual CIC through the use of extensive 3D models and sound clips. Each student appears in the environment as a 3D human representation, an avatar. A picture of each student's face is scanned and texture mapped onto his avatar so that each student can recognize the others inside the Virtual CIC. Networking capability of the Virtual CIC allows students at different geographical locations to train together in a unified virtual environment over a wide area network. This networking capability not only allows the crews in the same ship CIC to be distributed at different locations for training, it also provides the capability to simulate the whole virtual battle group operations as shown in Figure 5 so that different command and control structures can be developed and evaluated.

Various information visualization aids are incorporated into the virtual environment to help crewmembers in understanding and learning different tactical deployment of the combat systems. For example, a *holocube* allows students to visualize the entire battlespace in 3D as shown in Figure 6. This allows crewmembers to correlate sensor data with a visual 3D display of a god's-eye view, facilitating understanding of sensor capabilities and operations. This holographic-like display could also allow for visualization of sensor coverage, emission restrictions, operational boundaries and other doctrinal concepts and entities. As the crewmember changes his watchstation console mode, the holocube view could change to reflect the performance of that mode. This visual augmentation can assist the understanding and teaching tactical deployment of offensive and defensive capabilities. Virtual consoles are easily reconfigured or rearranged as needed for different ship configurations for maximum efficiency. By combining the virtual environment with a ship simulation model, operational procedures under various tactical scenarios can be explored and refined easily. System performance can be stressed in a realistic scenario without the risking accidents or tying up the real operational hardware. Incorporated information visualization aids allow ship crews to learn and understand the operations much faster than in a regular classroom environment. Virtual reality based training also allows students to train with advanced systems before the hardware has been built, or try out new physical layouts that do not currently exist, as well as experiment with and develop new operational procedures. The system described can be applied to land-based or mobile command centers as well as the ship CIC.

Summary

In this paper, we described the approach and benefits of virtual engineering in a distributed environment. We also illustrated two applications that have been developed based on a virtual reality framework to perform virtual engineering, system evaluation and training. The system provides an integrated environment for ship design, prototyping, and testing. This will enhance ship design and operations by exploring different configurations using virtual prototyping techniques coupled with the physics based models. It can also provide means to assess human computer interfaces (HCI) while designing various ship system components. More importantly, this same virtual environment can be used as (1) an assessment tool to evaluate a candidate design's performance; (2) as a means to letting the ship's crew have an early visualization and experience of the new design and provide feedback to the engineers; (3) as a training tool for the ship crew before and after the actual physical system is constructed. This virtual environment can support all phases of the ship life cycle. This approach will permit a much more realistic assessment of a candidate design and system configuration early in the design process. Changes can be made earlier, and therefore easier and cheaper, than in engineering production.

References

[1] IRIS Performer Programmer's Guide, Document Number 007-1680-040, SGI, Mountain View, CA, 1997.

[2] http://hla.dmso.mil.

[3] High Level Architecture Run-Time Infrastructure Programmer's Guide, RTI 1.3 Version 6, DMSO, March 12, 1999.

[4] High Level Architecture Interface Specification, Version 1.3, DMSO, April 2, 1998.

[5] http://www.microsoft.com/windows/netmeeting.

[6] http://manimac.itd.nrl.navy.mil/Ivox.
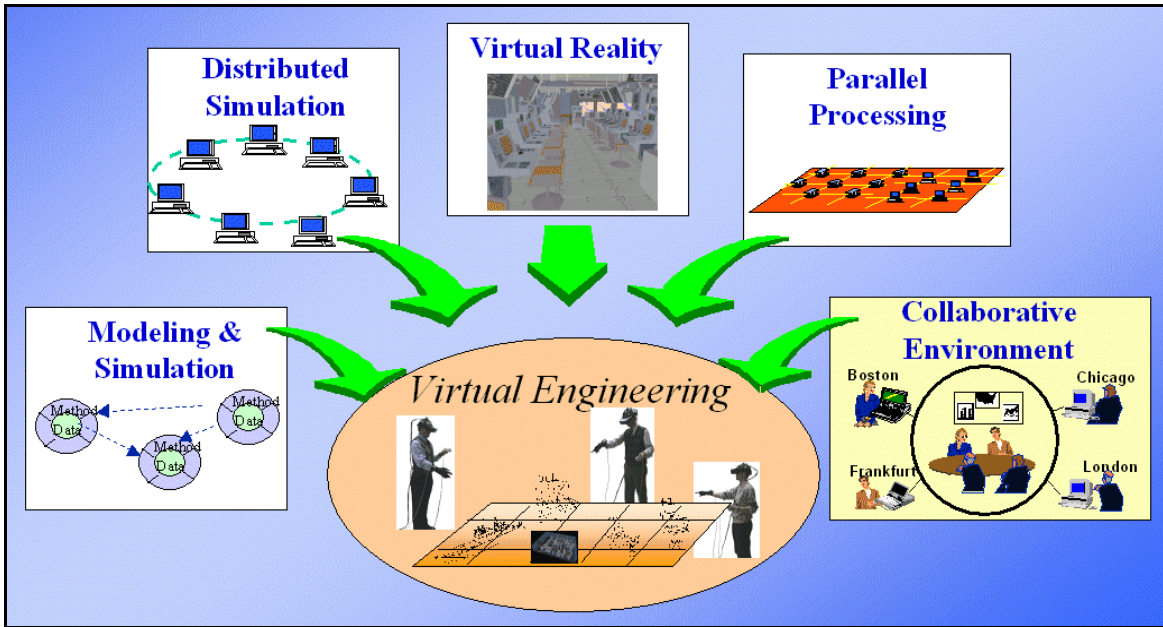
Figure 1: Virtual Engineering Technologies
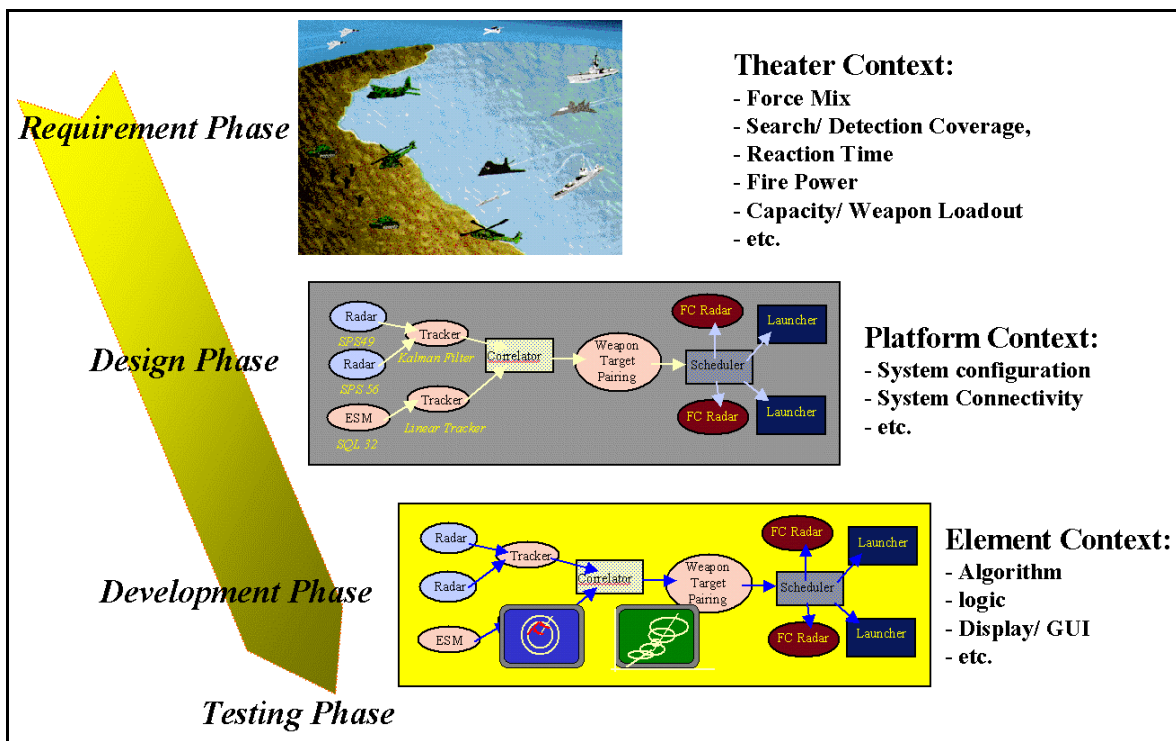
Figure 2: Virtual Ship Combat System Component Development for Acquisition Process
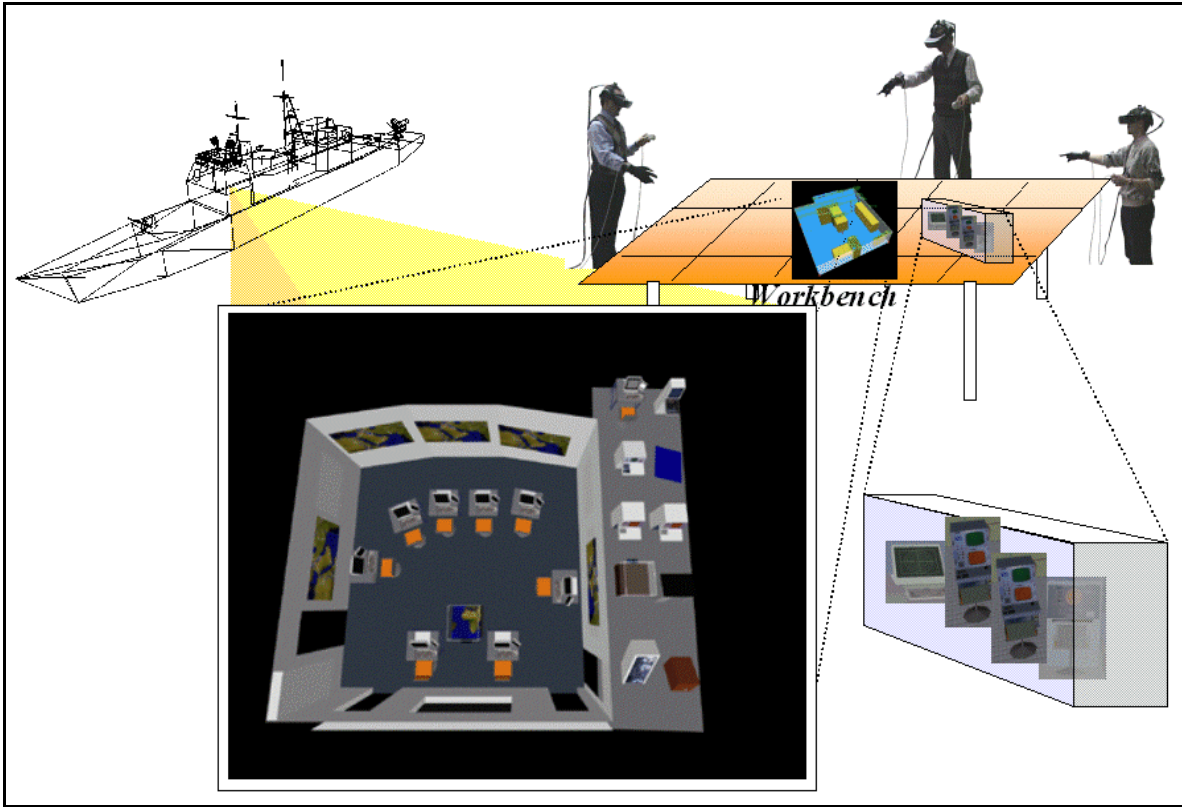
Figure 3: Virtual Prototyping in a Team Setting
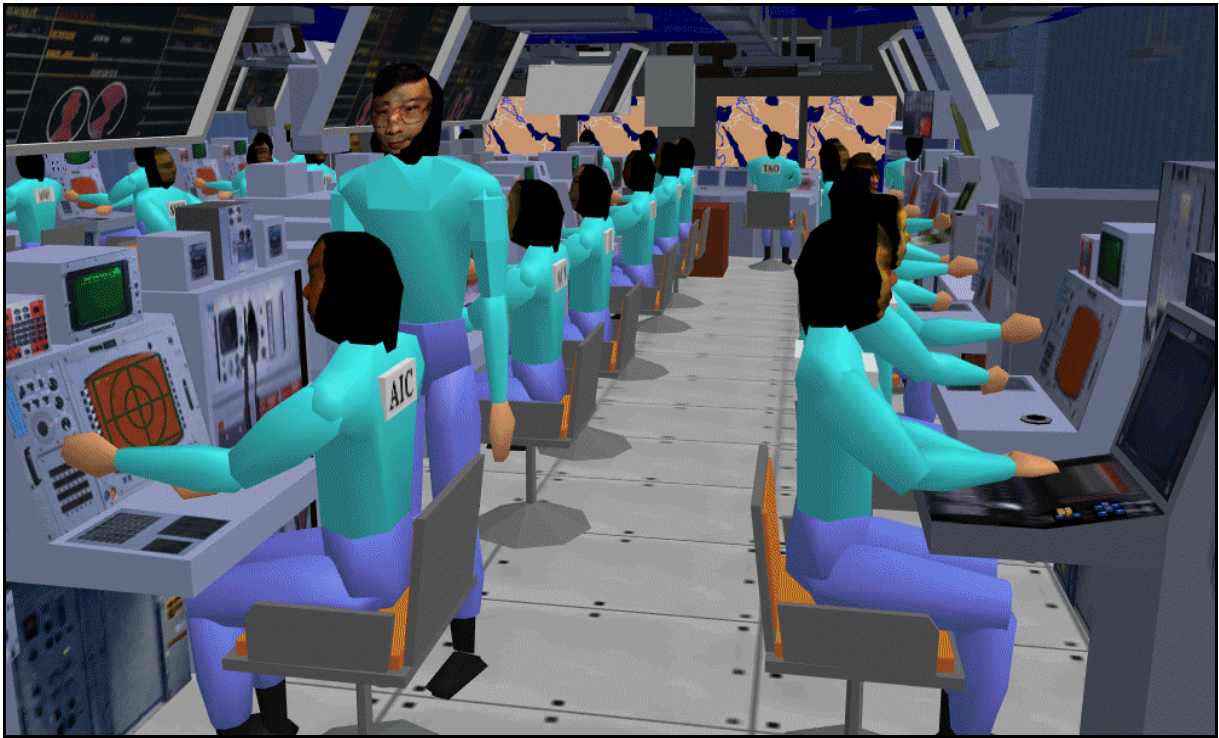
Figure 4: Virtual Ship Combat Information Center
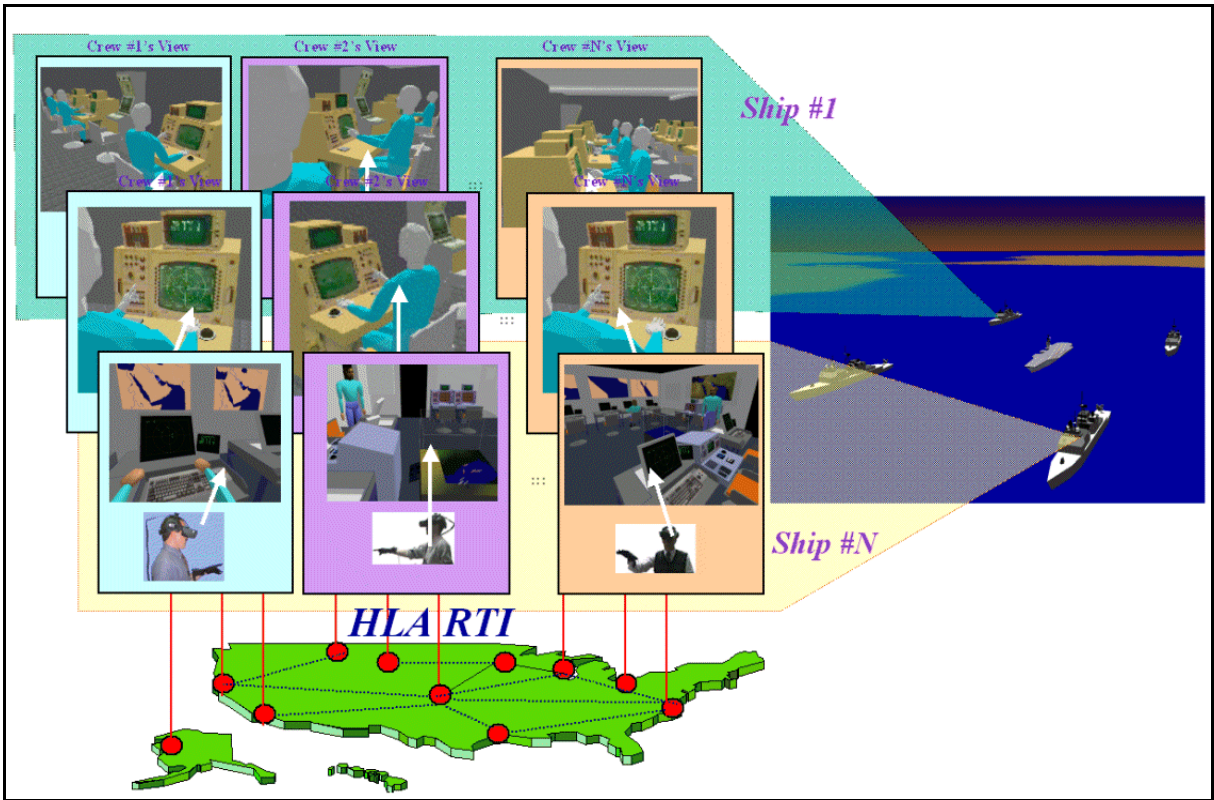
Figure 5: Virtual Battle Group
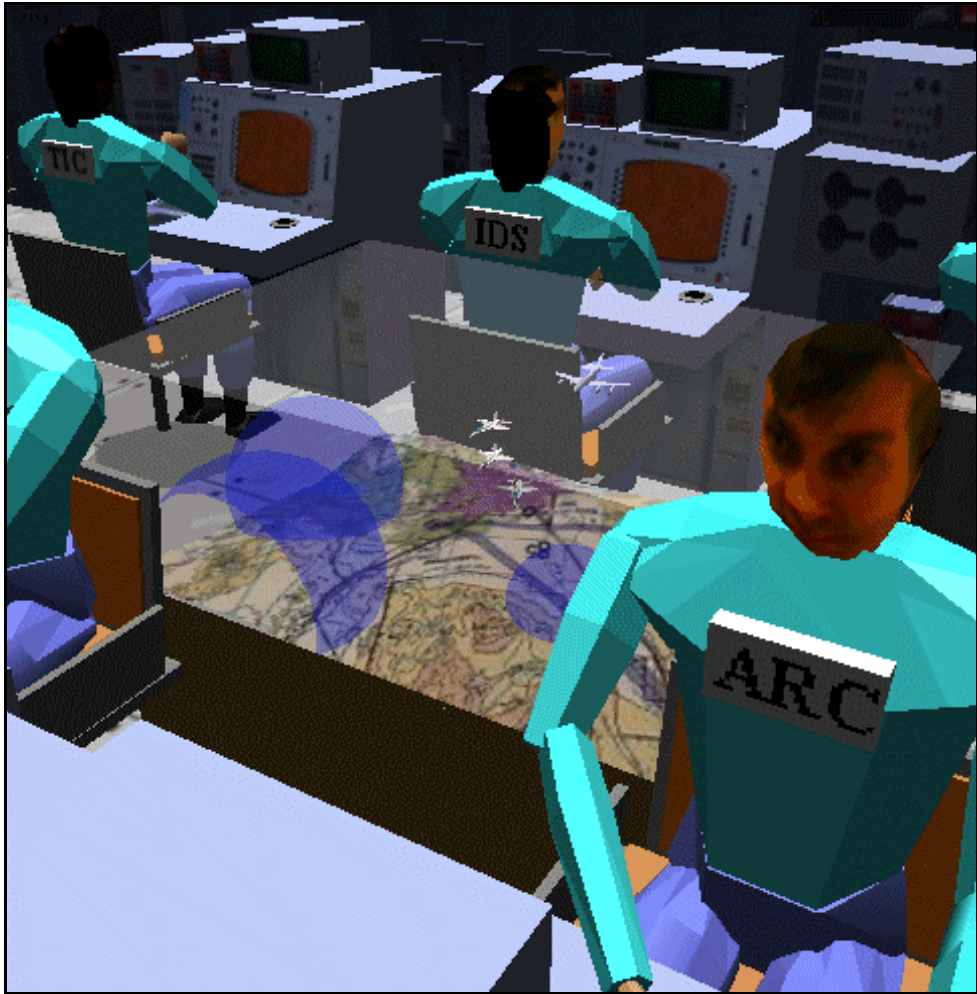
Figure 6: Holocube Visualization Aid

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's References	3. Further Reference	4. Security Classification of Document
	RTO-MP-049 AC/323(IST-017)TP/8	ISBN 92-837-1061-4	UNCLASSIFIED/ UNLIMITED

5. Originator	Research and Technology Organization North Atlantic Treaty Organization BP 25, 7 rue Ancelle, F-92201 Neuilly-sur-Seine Cedex, France

6. Title	New Information Processing Techniques for Military Systems

7. Presented at/sponsored by	the Symposium of the RTO Information Systems Technology Panel (IST) held in Istanbul, Turkey, 9-11 October 2000.

8. Author(s)/Editor(s)	9. Date
Multiple	April 2001

10. Author's/Editor's Address	11. Pages
Multiple	300 (text) 42 (slides)

12. Distribution Statement	There are no restrictions on the distribution of this document. Information about the availability of this and other RTO unclassified publications is given on the back cover.

13. Keywords/Descriptors

Information systems	Communications networks
Data processing	Data fusion
Military operations	Distributed systems
Battlefields	Neural nets
Command and control	Fuzzy sets
Operational effectiveness	Decision making
Real time operations	Man computer interface
Interoperability	Situation analysis
Logistics	Genetic algorithms
Secure communication	Virtual reality

14. Abstract

This volume contains the Technical Evaluation Report, 2 Keynote Addresses and 29 unclassified papers, presented at the Information Systems Technology Panel Symposium held in Istanbul, Turkey, 9-11 October 2000.

The papers were presented under the following headings:

- Information Systems and Techniques I

- Information Systems and Techniques II

- Security and Reliability

- Communications

- Detection, Fusion, Decision Support

- Virtual Reality and Human-Computer Interface

This page has been deliberately left blank

_____

Page intentionnellement blanche

NORTH ATLANTIC TREATY ORGANIZATION

R&T

RESEARCH AND TECHNOLOGY ORGANIZATION

BP 25 • 7 RUE ANCELLE

F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE

Telefax 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int

DISTRIBUTION OF UNCLASSIFIED
RTO PUBLICATIONS

NATO's Research and Technology Organization (RTO) holds limited quantities of some of its recent publications and those of the former AGARD (Advisory Group for Aerospace Research & Development of NATO), and these may be available for purchase in hard copy form. For more information, write or send a telefax to the address given above. **Please do not telephone**.

Further copies are sometimes available from the National Distribution Centres listed below. If you wish to receive all RTO publications, or just those relating to one or more specific RTO Panels, they may be willing to include you (or your organisation) in their distribution.

RTO and AGARD publications may be purchased from the Sales Agencies listed below, in photocopy or microfiche form. Original copies of some publications may be available from CASI.

NATIONAL DISTRIBUTION CENTRES

BELGIUM
Coordinateur RTO - VSL/RTO
Etat-Major de la Force Aérienne
Quartier Reine Elisabeth
Rue d'Evère, B-1140 Bruxelles

CANADA
Director Research & Development
Communications & Information
Management - DRDCIM 3
Dept of National Defence
Ottawa, Ontario K1A 0K2

CZECH REPUBLIC
Distribuční a informační středisko R&T
VTÚL a PVO Praha
Mladoboleslavská ul.
197 06 Praha 9-Kbely AFB

DENMARK
Danish Defence Research
Establishment
Ryvangs Allé 1, P.O. Box 2715
DK-2100 Copenhagen Ø

FRANCE
O.N.E.R.A. (ISP)
29 Avenue de la Division Leclerc
BP 72, 92322 Châtillon Cedex

GERMANY
Streitkräfteamt / Abteilung III
Fachinformationszentrum der
Bundeswehr, (FIZBw)
Friedrich-Ebert-Allee 34
D-53113 Bonn

GREECE (Point of Contact)
Hellenic Ministry of National
Defence
Defence Industry Research &
Technology General Directorate
Technological R&D Directorate
D.Soutsou 40, GR-11521, Athens

HUNGARY
Department for Scientific
Analysis
Institute of Military Technology
Ministry of Defence
H-1525 Budapest P O Box 26

ICELAND
Director of Aviation
c/o Flugrad
Reykjavik

ITALY
Centro di Documentazione
Tecnico-Scientifica della Difesa
Via XX Settembre 123a
00187 Roma

LUXEMBOURG
See Belgium

NETHERLANDS
NDRCC
DGM/DWOO
P.O. Box 20701
2500 ES Den Haag

NORWAY
Norwegian Defence Research
Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

POLAND
Chief of International Cooperation
Division
Research & Development
Department
218 Niepodleglosci Av.
00-911 Warsaw

PORTUGAL
Estado Maior da Força Aérea
SDFA - Centro de Documentação
Alfragide
P-2720 Amadora

SPAIN
INTA (RTO/AGARD Publications)
Carretera de Torrejón a Ajalvir, Pk.4
28850 Torrejón de Ardoz - Madrid

TURKEY
Millî Savunma Başkanliği (MSB)
ARGE Dairesi Başkanliği (MSB)
06650 Bakanliklar - Ankara

UNITED KINGDOM
Defence Research Information
Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

UNITED STATES
NASA Center for AeroSpace
Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320

SALES AGENCIES

NASA Center for AeroSpace
Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320
United States

The British Library Document
Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
United Kingdom

Canada Institute for Scientific and
Technical Information (CISTI)
National Research Council
Document Delivery
Montreal Road, Building M-55
Ottawa K1A 0S2, Canada

Requests for RTO or AGARD documents should include the word 'RTO' or 'AGARD', as appropriate, followed by the serial number (for example AGARD-AG-315). Collateral information such as title and publication date is desirable. Full bibliographical references and abstracts of RTO and AGARD publications are given in the following journals:

Scientific and Technical Aerospace Reports (STAR)
STAR is available on-line at the following uniform
resource locator:
   http://www.sti.nasa.gov/Pubs/star/Star.html
STAR is published by CASI for the NASA Scientific
and Technical Information (STI) Program
STI Program Office, MS 157A
NASA Langley Research Center
Hampton, Virginia 23681-0001
United States

Government Reports Announcements & Index (GRA&I)
published by the National Technical Information Service
Springfield
Virginia 22161
United States
(also available online in the NTIS Bibliographic
Database or on CD-ROM)

ST JOSEPH OTTAWA/HULL

Printed by St. Joseph Ottawa/Hull
(A St. Joseph Corporation Company)
45 Sacré-Cœur Blvd., Hull (Québec), Canada J8X 1C6