Copy 1

UNITED STATES AIR FORCE

SUMMER RESEARCH PROGRAM -- 1992

SUMMER RESEARCH EXTENSION PROGRAM FINAL REPORTS

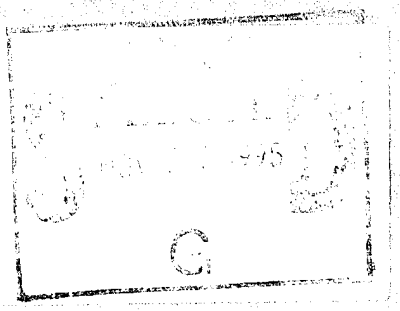
VOLUME 3

ROME LABORATORY

RESEARCH & DEVELOPMENT LABORATORIES

5800 UPLANDER WAY

CULVER CITY, CA 90230-6608



SUBMITTED TO:

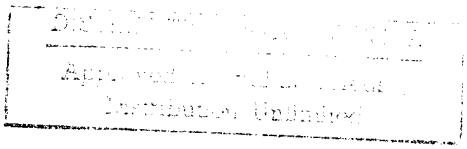
LT. COL. CLAUDE CAVENDER
PROGRAM MANAGER

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

BOLLING AIR FORCE BASE

WASHINGTON, D.C.

MAY 1993



REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	28 Dec 92	Annual 1 Sep 91 - 31 Aug 92

4. TITLE AND SUBTITLE

1992 Summer Faculty Research Program (SFRP)
Volume: 3C SREP

5. FUNDING NUMBERS

F49620-90-C-0076

6. AUTHOR(S)

Mr Gary Moore

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Research & Development Laboratoreis (RDL)
5800 Uplander Way
Culver City CA 90230-6600

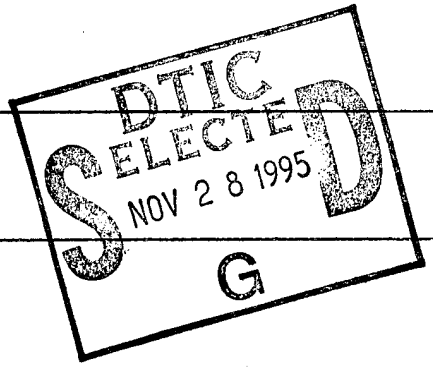
8. PERFORMING ORGANIZATION REPORT NUMBER

AFOSR-TR-95

0732

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR/NI
110 Duncan Ave., Suite B115
Bldg 410
Bolling AFB DC 20332-0001
Lt Col Claude Cavender



11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The purpose of this program is to develop the basis for cintinuing research of interest to the Air Force at the institution of the faculty member; to stiumlate continuing relations among faculty members and professional peers in the Air Force to enhance the research interests and capabilities of scientific and engineering educators; and to provide follow-on funding for research of particular promise that was started at an Air Force laboratory under the Summer Faculty Research Program.

During the summer of 1992 185 university faculty conducted research at Air Force laboratories for a period of 10 weeks. Each participant provided a report of their research, and these reports are consolidated into this annual report.

19951127 021

14. SUBJECT TERMS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

UNITED STATES AIR FORCE

SUMMER RESEARCH PROGRAM -- 1992

SUMMER RESEARCH EXTENSION PROGRAM FINAL REPORTS

VOLUME 3

ROME LABORATORY

RESEARCH & DEVELOPMENT LABORATORIES

5800 Uplander Way

Culver City, CA 90230-6608

Program Director, RDL
Gary Moore

Program Manager, AFOSR
Lt. Col. Claude Cavender

Program Manager, RDL
Scott Licoscos

Program Administrator, RDL
Gwendolyn Smith

Submitted to:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

Bolling Air Force Base

Washington, D.C.

May 1993

PREFACE

This volume is part of a five-volume set that summarizes the research of participants in the 1992 AFOSR Summer Research Extension Program (SREP). The current volume, Volume 3 of 5, presents the final reports of SREP participants at Rome Laboratory.

Reports presented in this volume are arranged alphabetically by author and are numbered consecutively -- e.g., 1-1, 1-2, 1-3; 2-1, 2-2, 2-3, with each series of reports preceded by a 22-page management summary. Reports in the five-volume set are organized as follows:

VOLUME	TITLE
1A	Armstrong Laboratory (part one)
1B	Armstrong Laboratory (part two)
2	Phillips Laboratory
3	Rome Laboratory
4A	Wright Laboratory (part one)
4B	Wright Laboratory (part two)
5	Air Force Civil Engineering Laboratory, Arnold Engineering Development Center, Frank J. Seiler Research Laboratory, Wilford Hall Medical Center

1992 SUMMER RESEARCH EXTENSION PROGRAM FINAL REPORTS

Rome Laboratory

1992 SUMMER RESEARCH EXTENSION PROGRAM (SREP) MANAGEMENT REPORT

1.0 BACKGROUND

Under the provisions of Air Force Office of Scientific Research (AFOSR) contract F49620-90-C-0076, September 1990, Research & Development Laboratories (RDL), an 8(a) contractor in Culver City, CA, manages AFOSR's Summer Research Program. This report is issued in partial fulfillment of that contract (CLIN 0003AC).

The name of this program was changed during this year's period of performance. For that reason, participants' cover sheets are captioned "Research Initiation Program" (RIP), while the covers of the comprehensive volumes are titled "Summer Research Extension Program" (SREP). The program's sponsor, the Air Force Office of Scientific Research (AFOSR), changed the name to differentiate this program from another which also bore its original name.

Apart from this name change, however, the program remained as it has been since its initiation as the Mini-Grant Program in 1983. The SREP is one of four programs AFOSR manages under the Summer Research Program. The Summer Faculty Research Program (SFRP) and the Graduate Student Research Program (GSRP) place college-level research associates in Air Force research laboratories around the United States for 8 to 12 weeks of research with Air Force scientists. The High School Apprenticeship Program (HSAP) is the fourth element of the Summer Research Program, allowing promising mathematics and science students to spend two months of their summer vacations at Air Force laboratories within commuting distance from their homes.

SFRP associates and exceptional GSRP associates are encouraged, at the end of their summer tours, to write proposals to extend their summer research during the following calendar year at their home institutions. AFOSR provides funds adequate to pay for 75 SREP subcontracts. In addition, AFOSR has traditionally provided further funding, when available, to pay for additional SREP proposals, including those submitted by associates from Historically Black Colleges and Universities (HBCUs) and Minority Institutions (MIs). Finally, laboratories may transfer internal funds to AFOSR to fund additional SREPs. Ultimately the laboratories inform RDL of their SREP choices, RDL gets AFOSR approval, and RDL forwards a subcontract to the institution where the SREP associate is employed. The subcontract (see Attachment 1 for a sample) cites the SREP associate as the principal investigator and requires submission of a report at the end of the subcontract period.

Institutions are encouraged to share costs of the SREP research, and many do so. The most common cost-sharing arrangement is reduction in the overhead, fringes, or administrative changes institutions would normally add on to the principal investigator's or research associate's labor. Some institutions also provide other support (e.g., computer run time, administrative assistance, facilities and equipment or research assistants) at reduced or no cost.

When RDL receives the signed subcontract, we fund the effort initially by providing 90% of the subcontract amount to the institution (normally $18,000 for a $20,000 SREP). When we receive the end-of-research report, we evaluate it administratively and send a copy to the laboratory for a technical evaluation. When the laboratory notifies us the SREP report is acceptable, we release the remaining funds to the institution.

2.0 THE 1992 SREP PROGRAM

SELECTION DATA: In the summer of 1991, 170 faculty members (SFRP associates) and 142 graduate students (GSRP associates) participated in the summer program. Of those, 147 SFRPs and 10 GSRPs submitted SREP proposals; 88 SFRP SREPs and 7 GSRP SREPs were selected for funding (total: 95).

	Summer 1991 Participants	Submitted SREP Proposals	SREPs Funded
SFRP	170	147	88
GSRP	142	10	7

The funding was provided as follows:

Contractual slots funded by AFOSR	75
Laboratory-funded	13
Additional funding from AFOSR	7
Total	95

Seven HBCU/MI associates from the 1991 summer program submitted SREP proposals; five were selected (one was lab-funded; four were funded by additional AFOSR funds).

By laboratory, the applications submitted and selected show in the following table:

	Applied	Selected
Air Force Civil Engineering Laboratory	6	4
Armstrong Laboratory	34	20
Arnold Engineering Development Center	12	2
Frank J. Seiler Research Laboratory	5	3
Phillips Laboratory	30	18
Rome Laboratory	16	11
Wilford Hall Medical Center	1	1
Wright Laboratory	53	36
TOTAL	157	95

Note: Phillips Laboratory funded 2 SREPs; Wright Laboratory funded 11; and AFOSR funded 7 beyond its contractual 75.

ADMINISTRATIVE EVALUATION: The administrative quality of the SREP associates' final reports was satisfactory. Most complied with the formatting and other instructions RDL provided to them. In the final days of December 1992 and in the first two months of 1993, several associates called and requested no-cost extensions of up to six months. After consultation with our AFOSR Contracting Officer's Representative, RDL approved the requests but asked that all such associates provide an interim report to be included in this volume. That caused an AFOSR-approved delay beyond the 1 April 1993 submission of this report. The subcontracts were funded by $1,893,616 of Air Force money. Institutions' cost sharing amounted to $948,686.

TECHNICAL EVALUATION: The form we used to gather data for technical evaluation and the technical evaluations of the SREP reports are provided as Attachment 2. This summary evaluation is shown by SREP number. The average rating range was from 3.1 to 5.0. The overall average for those evaluated was 4.6 out of 5.00. The three rating factors with the highest average scores were:

- The USAF should continue to pursue the research in this RIP report.
- The money spent on this RIP report was well worth it.
- I'll be eager to be a focal point for summer and RIP associates in the future.

Thus it is clear that the laboratories place a high value on AFOSR's Summer Research Program: SFRP, GSRP, and SREP.

3.0 SUBCONTRACTS SUMMARY

Table 1 lists contractually required information on each SREP subcontract. The individual reports are published in volumes as follows:

Laboratory	Volume
Air Force Civil Engineering Laboratory	5
Armstrong Laboratory	1
Arnold Engineering Development Center	5
Frank J. Seiler Research Laboratory	5
Phillips Laboratory	2
Rome Laboratory	3
Wilford Hall Medical Center	5
Wright Laboratory	4

TABLE 1: SUBCONTRACTS SUMMARY

Researcher's name / Institution / Location	Highest Degree / Department / Amount	Subcontract Number	Duration / Sharing
Abbott, Ben A Vanderbilt University Nashville, TN 37235	MS Dept of Electrical Engineering 19966.00	135	01/01/92-12/31/92 0.00
Acharya, Raj State University of New York, Buffalo Buffalo, NY 14260	PhD Dept of Electrical & Comp Engrg 20000.00	151	01/01/92-12/31/92 0.00
Adams, Christopher M Oklahoma State University Stillwater, OK 74078	PhD Dept of Chemistry 20000.00	68	01/01/92-12/31/92 0.00
Anderson, Richard A University of Missouri, Rolla Rolla, MO 65401	PhD Dept of Physics 20000.00	50	01/01/92-12/31/92 5000.00
Arora, Vijay K Wilkes University Wilkes-Barre, PA 18766	PhD Dept of Electrical & Comp Engrg 19996.00	3	10/01/91-09/30/92 36208.00
Ball, William P Duke University Durham, NC 27706	PhD Dept of Civil & Environmental Eng 20000.00	71	01/01/92-12/31/92 26747.00
Battles, Frank P Massachusetts Maritime Academy Buzzard's Bay, MA 025321803	PhD Dept of Basic Sciences 20000.00	152	01/01/92-12/31/92 22000.00
Bieniek, Ronald J University of Missouri, Rolla Rolla, MO 65401	PhD Dept of Physics 19945.00	147	01/01/92-12/31/92 4000.00
Blystone, Robert V Trinity University San Antonio, TX 78212	PhD Dept of Biology 20000.00	127	01/01/92-12/31/92 14783.00
Cha, Soyoung S University of Illinois, Chicago Chicago, IL 60680	PhD Dept of Mechanical Engineering 20000.00	011	01/01/92-12/31/92 3842.00
Chandra, D. V. Satish Kansas State University Manhattan, KS 66506	PhD Dept of Electrical Engineering 20000.00	89	01/18/92-10/17/92 11170.00
Chenette, Eugene R University of Florida Gainesville, FL 32611	PhD Dept of Electrical Engineering 20000.00	106	01/01/92-12/31/92 0.00
Christensen, Douglas A University of Utah Salt Lake City, UT 84112	PhD Dept of Electrical Engineering 19999.00	83	01/01/92-12/31/92 5000.00

Chubb, Gerald P Ohio State University Columbus, OH 43235	PhD Dept of Aviation 20000.00	26 7600.00	01/01/92-12/31/92
Courter, Robert W Louisiana State University Baton Rouge, LA 70803	PhD Dept of Mechanical Engineering 20000.00	8 445.00	10/01/91-09/30/92
Dey, Pradip P Hampton University Hampton, VA 23668	PhD Computer Science Department 19921.00	120 0.00	01/01/92-12/31/92
Draut, Arthur W Embry Riddle Aeronautical University Prescott, AZ 86301	PhD Computer Science Dept 19431.00	133 0.00	01/06/92-05/08/92
Dreisbach, Joseph University of Scranton Scranton, PA 185104626	PhD Dept of Chemistry 20000.00	108 4000.00	12/01/91-12/01/92
Dror, Itiel Harvard University Cambridge, MA 02138	BS Dept of Psychology 20000.00	76 0.00	01/01/92-12/31/92
Drost-Hansen, W. University of Miami Coral Gables, FL 33124	PhD Dept of Chemistry 20000.00	124 12000.00	12/01/91-12/01/92
Dunleavy, Lawrence P University of South Florida Tampa, FL 33620	PhD Dept of Electrical Engineering 20000.00	41 6463.00	01/01/92-12/31/92
Evans, Joseph B University of Kansas Lawrence, KS 66045	PhD Dept of Electrical & Comp Engrg 20000.00	96 0.00	01/01/92-12/31/92
Flowers, George T Auburn University Auburn, AL 368495341	PhD Dept of Mechanical Engineering 19986.00	73 12121.00	01/01/92-12/30/92
Gantenbein, Rex E University of Wyoming Laramie, WY 82071	PhD Dept of Computer Science 20000.00	22 26643.00	01/01/91-12/31/92
Garcia, Ephrarim Vanderbilt University Nashville, TN 37235	PhD Dept of Mechanical Engineering 20000.00	32 9659.00	12/01/91-11/30/92
German, Fred J Auburn University Auburn University, AL 36830	PhD Dept of Electrical Engineering 20000.00	49 0.00	01/01/92-12/31/92
Gould, Richard D North Carolina State University Raleigh, NC 276957910	PhD Dept of Mech and Aerospace Engrg 20000.00	87 14424.00	01/01/92-12/31/92
Gove, Randy L University of Alabama, Huntsville Huntsville, AL 35899	MS Dept of Physics 20000.00	122 3469.00	01/01/92-12/31/92
Grabowski, Marek University of Colorado, Colorado Springs Colorado Springs, CO 809337150	PhD Dept of Physics 19700.00	92 0.00	01/01/92-12/31/92

Gunaratne, Manjriker University of South Florida Tampa, FL 33620	PhD Dept of Civil Engrg & Mechanics 19994.00	90 10062.00	01/01/92-12/31/92
Hall, Ernest L University of Cincinnati Cincinnati, OH 452210072	PhD Dept of Robotics Research 19975.00	134 0.00	01/01/92-12/31/92
Hamilton, William L Salem State College Salem, MA 01970	PhD Dept of Geography 20000.00	47 32000.00	01/01/92-12/31/92
Hamilton, Kirk L Xavier University of Louisiana New Orleans, LA 70125	PhD Dept of Biology 20000.00	57 16100.00	01/01/92-12/31/92
Harris, Harold H University of Missouri, St.Louis St. Louis, MO 63121	PhD Dept of Chemistry 19300.00	94 8600.00	01/01/92-12/31/92
Hartung, George H University of Hawaii Honolulu, HI 96822	PhD Dept of Physiology 20000.00	46 7530.00	01/01/92-12/31/92
Hatfield, Steven L University of Kentucky Lexington, KY 40506	BS Dept of Materials Science & Engrg 20000.00	23 28625.00	01/01/92-12/31/92
Hedman, Paul O'Dell Brigham Young University Provo, UT 84602	PhD Dept of Chemical Engineering 19999.00	17 6928.00	01/01/92-12/31/92
Heister, Stephen D Purdue University West Lafayette, IN 47907	PhD School of Aero & Astronautics 20000.00	5 4419.00	01/01/92-12/31/92
Hess, David J University of Texas, Austin Austin, TX 78713	BA Dept of Psychology 19914.00	149 8784.00	01/01/92-12/31/92
Hoffman, R. W Case Western Reserve University Cleveland, OH 44106	PhD Dept of Physics 19770.00	99 0.00	01/01/92-12/31/92
Huerta, Manuel A University of Miami Coral Gables, FL 33124	PhD Dept of Physics 20000.00	62 1207.00	01/01/92-12/31/92
Hui, David University of New Orleans New Orleans, LA 70148	PhD Dept of Mechanical Engineering 20000.00	116 0.00	01/01/92-12/31/92
Iyer, Ashok University of Nevada, Las Vegas Las Vegas, NV 89154	PhD Dept of Electrical & Comp Engrg 20000.00	74 18549.00	01/01/92-12/31/92
Khonsari, Michael M University of Pittsburgh Pittsburgh, PA 15260	PhD Dept of Mechanical Engineering 20000.00	53 32958.00	01/01/92-12/31/92
Kibert, Charles J University of Florida Gainesville, FL 32611	PhD Dept of Fire Testing & Research 20000.00	2 6928.00	01/01/92-12/31/92

Klarup, Douglas G University of Montana Missoula, MT 59812	PhD Dept of Chemistry 20000.00	84 0.00	01/01/92-12/31/92
Koblasz, Arthur J Georgia Institute of Technology Atlanta, GA 30332	PhD Dept of Civil Engineering 19956.00	145 0.00	01/01/92-09/30/92
Kornreich, Philipp Syracuse University Syracuse, NY 13244	PhD Dept of Electrical & Comp Engrg 20000.00	35 0.00	10/01/91-09/30/92
Kuo, Spencer P Polytechnic University Farmingdale, NY 11735	PhD Dept of Electrical Engineering 20000.00	59 9916.00	01/01/92-12/31/92
Langhoff, Peter W Indiana University Bloomington, IN 47402	PhD Dept of Chemistry 20000.00	115 35407.00	01/01/92-12/31/92
Lee, Byung-Lip Pennsylvania State University University Park, PA 16802	PhD Dept of Engrg Science & Mechanics 20000.00	93 8173.00	01/01/92-12/31/92
Leigh, Wallace B Alfred University Alfred, NY 14802	PhD Dept of Electrical Engineering 19767.00	118 18770.00	01/01/92-12/31/92
Liddy, Elizabeth Syracuse University Syracuse, NY 132444100	PhD Dept of Information Studies 20000.00	104 0.00	01/01/92-12/31/92
Liu, Cheng University of North Carolina, Charlotte Charlotte, NC 28270	PhD Dept of Engineering Technology 20000.00	6 0.00	11/01/99-12/31/92
Main, Robert G California State University, Chico Chico, CA 959290504	PhD Dept of Communication Design 20000.00	28 7672.00	01/01/92-06/30/92
Mains, Gilbert J Oklahoma State University Stillwater, OK 74078	PhD Dept of Chemistry 19071.00	52 8746.00	01/01/92-12/31/92
Marathay, Arvind S University of Arizona Tucson, AZ 85721	PhD Dept of Optical Sciences 20000.00	51 0.00	01/01/92-12/31/92
Martin, Charlesworth R Norfolk State University Norfolk, VA 23504	PhD Dept of Physics & Engineering 20000.00	125 0.00	01/01/92-12/31/92
Mayes, Jessica L University of Kentucky Lexington, KY 405034203	BS Dept of Material Science & Engrng 20000.00	16 28625.00	01/01/92-12/31/92
Mulligan, Benjamin E University of Georgia Athens, GA 30602	PhD Dept of Psychology 19895.00	54 13677.00	01/01/92-12/31/92
Munday, Edgar G University of North Carolina, Charlotte Charlotte, NC 28223	PhD Dept of Mechanical Engineering 20000.00	38 11638.00	10/01/91-10/30/92

Nurre, Joseph H Ohio University Athens, OH 45701	PhD 56 Dept of Electrical & Comp Engrg 19842.00	01/01/92-12/31/92 15135.00
Orkwis, Paul D University of Cincinnati Cincinnati, OH 452210070	PhD 14 Dept of Engineering Mechanics 19966.00	10/01/91-10/30/92 23017.00
Patra, Amit L University of Puerto Rico Mayaquez, PR 00681	PhD 69 Dept of General Engineering 20000.00	01/01/92-12/31/92 2750.00
Peters II, Richard A Vanderbilt University Nashville, TN 37235	PhD 160 Dept of Electrical Engineering 20000.00	01/01/92-12/31/92 0.00
Pollack, Steven K University of Cincinnati Cincinnati, OH 452200012	PhD 31 Dept of Materials Sci & Engrg 20000.00	01/01/92-12/31/92 14877.00
Prescott, Glenn E University of Kansas Lawrence, KS 66045	PhD 72 Dept of Electrical Engineering 20000.00	01/01/92-12/31/92 8000.00
Price, James L University of Iowa Iowa City, IA 52242	PhD 48 Dept of Sociology 20000.00	01/01/92-12/30/92 8600.00
Qazi, Salahuddin SUNY, Utica Utica, NY 13504	PhD 129 Dept of Electrical Engineering 20000.00	01/01/92-12/31/92 25000.00
Rappaport, Carey M Northeastern University Boston, MA 02115	PhD 58 Dept of Electrical & Comp Engrng 19999.00	01/01/92-06/30/92 0.00
Rawson, Jenny L North Dakota State University Fargo, ND 58105	PhD 144 Dept of Electrical Engineering 19997.00	01/01/92-12/31/92 19826.00
Riccio, Gary E University of Illinois, Urbana Urbana, IL 61821	PhD 80 Dept of Human Perception 20000.00	01/01/92-12/31/92 0.00
Rotz, Christopher A Brigham Young University Provo, UT 84602	PhD 136 Dept of Manufacturing Engineering 20000.00	12/01/91-12/31/92 11814.00
Schwartz, Martin University of North Texas Denton, TX 762035068	PhD 55 Dept of Chemistry 20000.00	01/01/92-12/31/92 18918.00
Senseman, David M University of Texas, San Antonio San Antonio, TX 78285	PhD 77 Dept of Information 20000.00	12/01/91-11/30/92 19935.00
Sensiper, Martin University of Central Florida Orlando, FL 32816	BS 15 Dept of Electrical Engineering 20000.00	11/01/91-05/31/92 0.00
Shamma, Jeff S University of Texas, Austin Austin, TX 78713	PhD 70 Dept of Electrical Engineering 20000.00	01/01/92-12/31/92 0.00

Shively, Jon H California State University, Northridge Northridge, CA 91330	PhD 140 Dept of CIAM 20000.00	01/01/92-12/31/92 14553.00
Singh, Sahjendra N University of Nevada, Las Vegas Las Vegas, NV 89014	PhD 79 Dept of Electrical Engineering 20000.00	01/01/92-12/31/92 20595.00
Smith, Gerald A Pennsylvania State University University Park, PA 16802	PhD 63 Dept of Physics 20000.00	07/01/92-07/01/93 0.00
Stephens, Benjamin R Clemson University Clemson, SC 29634	PhD 114 Dept of Psycology 19988.00	01/01/92-12/31/92 4250.00
Sudkamp, Thomas Wright State University Dayton, OH 45435	PhD 97 Dept of Computer Science 20000.00	01/01/92-08/31/92 18739.00
Sydor, Michael University of Minnesota, Duluth Duluth, MN 55804	PhD 11 Dept of Physics 20000.00	01/01/92-12/31/92 0.00
Tankin, Richard S Northwestern University Evanston, IL 60208	PhD 44 Dept of Mechanical Engineering 20000.00	01/01/92-12/31/92 29103.00
Taylor, Michael D University of Central Florida Orlando, FL 32816	PhD 141 Dept of Mathematics 20000.00	05/01/92-07/31/92 1587.00
Teegarden, Kenneth J University of Rochester Rochester, NY 14627	PhD 98 Dept of Optics 20250.00	01/01/92-12/31/92 60600.00
Tew, Jeffrey D Virginia Polytech Instit and State Univ Blacksburg, VA 24061	PhD 137 Dept of Industrial Engineering 17008.00	03/01/92-09/30/92 4564.00
Tipping, Richard H University of Alabama Tuscaloosa, AL 35487	PhD 81 Dept of Physics & Astronomy 20000.00	01/01/92-05/31/92 15000.00
Tripathi, Ram C University of Texas, San Antonio San Antonio, TX 78249	PhD 105 Dept of Mathematics 20000.00	01/01/92-12/31/92 2274.00
Wells, Fred V Idaho State University Pocatello, ID 83209	PhD 155 Dept of Chemistry 20000.00	01/01/92-12/31/92 8000.00
Whitefield, Phillip D University of Missouri, Rolla Rolla, MO 65401	PhD 25 Dept of Chemistry 19991.00	01/01/92-12/31/92 25448.00
Wolfenstine, Jeffrey B University California, Irvine Irvine, CA 92717	PhD 18 Dept of Mechanical Engineering 20000.00	01/01/92-12/31/92 11485.00
Wolper, James S Idaho State University Pocatello, ID 83209	PhD 138 Dept of Mathematics 20000.00	01/15/92-09/30/92 4828.00

Zavodney, Lawrence D Ohio State University Columbus, OH 43210	PhD 148 Dept of Engineering Mechanics 20000.00	01/01/92-12/31/92 0.00
Zimmerman, Wayne J Texas Women University Denton, TX 76204	PhD 111 Dept of Mathematics 19990.00	01/01/92-12/31/92 8900.00

ATTACHMENT 1:

SAMPLE SREP SUBCONTRACT

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
1993 SUMMER RESEARCH EXTENSION PROGRAM SUBCONTRACT 93-36

BETWEEN

Research & Development Laboratories
5800 Uplander Way
Culver City, CA 90230-6608

AND

University of Delaware
Sponsored Programs Admin.
Newark, DE 19716

REFERENCE: Summer Research Extension Program Proposal 93-36
Start Date: 01/01/93 End Date: 12/31/93
Proposal amount: $20000.00

(1) PRINCIPAL INVESTIGATOR: Dr. Ian W. Hall
Materials Science
University of Delaware
Newark, DE 19716

(2) UNITED STATES AFOSR CONTRACT NUMBER: F49620-90-C-09076

(3) CATALOG OF FEDERAL DOMESTIC ASSISTANCE NUMBER (CFDA): 12.800
PROJECT TITLE: AIR FORCE DEFENSE RESEARCH SOURCES PROGRAM

(4) ATTACHMENTS 1 AND 2: SREP REPORT INSTRUCTIONS

*** SIGN SREP SUBCONTRACT AND RETURN TO RDL ***

1. BACKGROUND: Research & Development Laboratories (RDL) is under contract (F49620-90-C-0076) to the United States Air Force to administer the Summer Research Programs (SRP), sponsored by the Air Force Office of Scientific Research (AFOSR), Bolling Air Force Base, D.C. Under the SRP, a selected number of college faculty members and graduate students spend part of the summer conducting research in Air Force laboratories. After completion of the summer tour participants may submit, through their home institutions, proposals for follow-on research. The follow-on research is known as the Research Initiation Program (RIP). Approximately 75 RIP proposals annually will be selected by the Air Force for funding of up to $20,000; shared funding by the academic institution is encouraged. RIP efforts selected for funding are administered by RDL through subcontracts with the institutions. This subcontract represents such an agreement between RDL and the institution designated in Section 5 below.

2. RDL PAYMENTS: RDL will provide the following payments to RIP institutions:

- 90 percent of the negotiated RIP dollar amount at the start of the RIP Research period.
- the remainder of the funds within 30 days after receipt at RDL of the acceptable written final report for the RIP research.

3. INSTITUTION'S RESPONSIBILITIES: As a subcontractor to RDL, the institution designated on the title page will:

a. Assure that the research performed and the resources utilized adhere to those defined in the RIP proposal.
b. Provide the level and amounts of institutional support specified in the RIP proposal.
c. Notify RDL as soon as possible, but not later than 30 days, of any changes in 3a or 3b above, or any change to the assignment or amount of participation of the Principal Investigator designated on the title page.
d. Assure that the research is completed and the final report is delivered to RDL not later than twelve months from the effective date of this subcontract. The effective date of the subcontract is one week after the date that the institution's contracting representative signs this subcontract, but no later than January 15, 1992.
e. Assure that the final report is submitted in the format shown in Attachment 1.

f. Agree that any release of information relating to this subcontract (news releases, articles, manuscripts, brochures, advertisements, still and motion pictures, speeches, trade association meetings, symposia, etc.) will include a statement that the project or effort depicted was or is sponsored by: Air Force Office of Scientific Research, Bolling AFB, D.C.

g. Notify RDL of inventions or patents claimed as the result of this research in a format specified in Attachment 1.

h. RDL is required by the prime contract to flow down patent rights and technical data requirements in this subcontract. Attachment 2 to this subcontract contains a list of contract clauses incorporated by reference in the prime contract.

4. All notices to RDL shall be addressed to:

RDL Summer Research Program Office
5800 Uplander Way
Culver City, CA 90230-6608

5. By their signatures below, the parties agree to the provisions of this subcontract.

Abe S. Sopher RDL Contracts Manager	Signature of Institution Contracting Official
	Typed/Printed Name
Date	Title
	Institution
	Date/Phone

Attachment 1

Final Report Format

1. All RIP Principal Investigators will submit a final report of the research conducted.

2. One copy of the report is due to RDL no later than twelve months after the effective date of the RIP subcontract. At the same time, submit one copy to the Air Force laboratory focal point.

3. The title page should contain the title of the research, the Principal Investigator and or other co-investigators, the month and year of issue, the university with department and address, and acknowledgement of sponsorship by AFOSR (see clause 3f of this subcontract).

4. For text, use a font that is 12 characters per inch (elite) and as close to letter quality as possible. Start with the title in all caps one and one-half inches from the top of the first page; if the title requires two or more lines, single space it. Double space below the title, and then center and type the researcher's title and name. Then space twice and begin the double-spaced text.

   Use a one-and-one-half-inch left margin and a one-inch right margin for the body of the text. Center page numbers at the foot of each page, one inch from the bottom. Each page should have a one-inch margin at the top. The format should be that of a standard research paper: it should begin with a one-paragraph abstract (on its own page) summarizing your work and should be followed by an introduction, a discussion of the problem, a results section, and a conclusion. Since multiple copies of your report may be required, assure that all pages can be readily copied to a black-and-white 8 1/2" by 11" page. (No colors, such as blue or green, that don't photocopy well, and no foldouts, please.)

5. The report must be accompanied by a separate statement on whether or not any inventions or patents have resulted from this research. If yes, use a DD Form 882 (supplied by RDL on request) to indicate the patent filing date, serial number, title, and a copy of the patent application, and patent number and issue date for any subject invention in any country in which the subcontractor has applied for patents.

Attachment 2

Contract Clauses

This contract incorporates by reference the following clauses of the Federal Acquisition Regulations (FAR), with the same force and effect as if they were given in full text. Upon request, the Contracting Officer or RDL will make their full text available (FAR 52.252-2).

FAR CLAUSES	TITLE AND DATE
52.202-1	DEFINITIONS (APR 1984)
52.203-1	OFFICIALS NOT TO BENEFIT (APR 1984)
52.203-3	GRATUITIES (APR 1984)
52.203-5	COVENANT AGAINST CONTINGENT FEES (APR 1984)
52.304-6	RESTRICTIONS ON SUBCONTRACTOR SALES TO THE GOVERNMENT (JUL 1985)
52.203-7	ANTI-KICKBACK PROCEDURES (OCT 1988)
52.203-12	LIMITATION ON PAYMENTS TO INFLUENCE CERTAIN FEDERAL TRANSACTIONS (JAN 1990)
52.204-2	SECURITY REQUIREMENTS (APR 1984)
52.209-6	PROTECTING THE GOVERNMENT'S INTEREST WHEN SUBCONTRACTING WITH CONTRACTORS DEBARRED, SUSPENDED, OR PROPOSED FOR DEBARMENT (MAY 1989)
52.212-8	DEFENSE PRIORITY AND ALLOCATION REQUIREMENTS (MAY 1986)
52.215-1	EXAMINATION OF RECORDS BY COMPTROLLER GENERAL (APR 1984)
52.215-2	AUDIT - NEGOTIATION (DEC 1989)
52.222-26	EQUAL OPPORTUNITY (APR 1984)
52.222-28	EQUAL OPPORTUNITY PREAWARD CLEARANCE OF SUBCONTRACTS (APR 1984)
52.222-35	AFFIRMATIVE ACTION FOR SPECIAL DISABLED AND VIETNAM ERA VETERANS (APR 1984)
52.222-36	AFFIRMATIVE ACTION FOR HANDICAPPED WORKERS (APR 1984)

52.222-37	EMPLOYMENT REPORTS ON SPECIAL DISABLED VETERANS AND VETERANS OF THE VIETNAM ERA (JAN 1988)
52.223-2	CLEAN AIR AND WATER (APR 1984)
52.232-6	DRUG-FREE WORKPLACE (MAR 1989)
52.224-1	PRIVACY ACT NOTIFICATION (APR 1984)
52.224-2	PRIVACY ACT (APR 1984)
52.225-13	RESTRICTIONS ON CONTRACTING WITH SANCTIONED PERSONS (MAY 1989)
52.227-1	AUTHORIZATION AND CONSENT (APR 1984)
52.227-2	NOTICE AND ASSISTANCE REGARDING PATENT AND COPYRIGHT INFRINGEMENT (APR 1984)
52.227-10	FILING OF PATENT APPLICATIONS - CLASSIFIED SUBJECT MATTER (APR 1984)
52.227-11	PATENT RIGHTS - RETENTION BY THE CONTRACTOR (SHORT FORM) (JUN 1989)
52.228-6	INSURANCE - IMMUNITY FROM TORT LIABILITY (APR 1984)
52.228-7	INSURANCE - LIABILITY TO THIRD PERSONS (APR 1984)
52.230-5	DISCLOSURE AND CONSISTENCY OF COST ACCOUNTING PRACTICES (SEP 1987)
52.232-23	ASSIGNMENT OF CLAIMS (JAN 1986)
52.237-3	CONTINUITY OF SERVICES (APR 1984)
52.246-25	LIMITATION OF LIABILITY - SERVICES (APR 1984)
52.249-6	TERMINATION (COST-REIMBURSEMENT) (MAY 1986)
52.249-14	EXCUSABLE DELAYS (APR 1984)
52.251-1	GOVERNMENT SUPPLY SOURCES (APR 1984)

DoD FAR CLAUSES	TITLE AND DATE
252.203-7001	SPECIAL PROHIBITION ON EMPLOYMENT (MAR 1989)
252.203-7002	STATUTORY COMPENSATION PROHIBITIONS AND REPORTING REQUIREMENTS RELATING TO CERTAIN FORMER DEPARTMENT OF DEFENSE (DoD) EMPLOYEES (APR 1988)
252.223-7500	DRUG-FREE WORK FORCE (SEP 1988)
252.225-7001	BUY AMERICAN ACT AND BALANCE OF PAYMENTS PROGRAM (APR 1985)
252-225-7023	RESTRICTION ON ACQUISITION OF FOREIGN MACHINE TOOLS (JAN 1989)
252.227-7013	RIGHTS IN TECHNICAL DATA AND COMPUTER SOFTWARE (OCT 1988)
252.227-7018	RESTRICTIVE MARKINGS ON TECHNICAL DATA (OCT 1988)
252.227-7029	IDENTIFICATION OF TECHNICAL DATA (APR 1988)
252.227-7034	PATENTS - SUBCONTRACTS (APR 1984)
252.227-7037	VALIDATION OF RESTRICTIVE MARKINGS ON TECHNICAL DATA (APR 1988)
252.231-7000	SUPPLEMENTAL COST PRINCIPLES (APR 1984)
252.231-7001	PENALTIES FOR UNALLOWABLE COSTS (APR 1988)
252.231-7003	CERTIFICATION OF INDIRECT COSTS (APR 1986)
252.251-7000	ORDERING FROM GOVERNMENT SUPPLY SOURCES (APR 1984)
252.271-7001	RECOVERY OF NONRECURRING COSTS ON COMMERCIAL SALES OF DEFENSE PRODUCTS AND TECHNOLOGY AND OF ROYALTY FEES FOR USE OF DoD TECHNICAL DATA (FEB 1989)

7 November 1991

AFOSR/PKO
Bldg. 410, Room C-124
Bolling AFB, DC 20332-6448

Attn: Ms. Kathleen Wetherell

Dear Ms. Wetherell:

Enclosed for your approval is the model subcontract for the Research Initiation Program under the Summer Research Programs (Contract F9620-90-C-0076). The blanks will be filled by merging information from our dBase IV database.

Sincerely,

Abe S. Sopher
Contracts Manager

cc: AFOSR/NI (Lt. Col. Cavendar)

ATTACHMENT 2:

SAMPLE TECHNICAL EVALUATION FORM AND TECHNICAL EVALUATION SUMMARY

1992 RESEARCH INITIATION PROGRAM TECHNICAL EVALUATION

RIP NO: 92-2
RIP ASSOCIATE: Dr. Charles Kibert

Provided are several evaluation statements followed by ratings of (1) through (5). A rating of (1) is the lowest and (5) is the highest. Circle the rating level number you best feel rates the statement. Document additional comments on the back of this evaluation form.

Mail or fax the completed form to:

RDL
Attn: 1992 RIP TECH EVALS
5800 Uplander Way
Culver City, CA 90230-6608
(Fax: 310 216-5940)

1.	This RIP report has a high level of technical merit	1 2 3 4 5
2.	The RIP program is important to accomplishing the lab's mission	1 2 3 4 5
3.	This RIP report accomplished what the associate's proposal promised	1 2 3 4 5
4.	This RIP report addresses area(s) important to the USAF	1 2 3 4 5
5.	The USAF should continue to pursue the research in this RIP report	1 2 3 4 5
6.	The USAF should maintain research relationships with this RIP associate	1 2 3 4 5
7.	The money spent on this RIP effort was well worth it	1 2 3 4 5
8.	This RIP report is well organized and well written	1 2 3 4 5
9.	I'll be eager to be a focal point for summer and RIP associates in the future	1 2 3 4 5
10.	The one-year period for complete RIP research is about right	1 2 3 4 5

****USE THE BACK OF THIS FORM FOR ADDITIONAL COMMENTS****

LAB FOCAL POINT'S NAME (PRINT): _____

OFFICE SYMBOL: _____ PHONE: _____

TECHNICAL EVALUATION SUMMARY

Technical Evaluation Questionnaire Rating Factors

Subcontract no.	1	2	3	4	5	6	7	8	9	10	Average
135	5	4	5	4	4	4	4	4	5	5	4.4
50	4	4	5	4	4	4	4	3	5	5	4.2
3	4	3	3	3	3	3	3	3	3	4	3.2
71	4	4	4	4	3	5	5	4	5	5	4.3
152	3	4	3	4	4	3	4	3	4	5	3.7
147	5	5	5	5	5	5	5	5	5	4	4.9
011	4	4	5	4	5	5	5	4	5	4	4.5
106	5	5	4	5	5	5	5	5	5	5	4.9
83	5	4	5	5	5	5	5	5	5	4	4.8
26	5	4	4	5	5	5	5	5	4	4	4.6
8	5	3	4	4	5	5	5	3	5	5	4.4
120	1	5	2	4	5	3	2	1	4	4	3.1
133	3	2	4	5	5	4	3	4	3	5	3.8
108	5	4	4	5	5	5	5	5	5	5	4.8
76	5	5	5	5	5	5	5	5	5	3	4.8
122	5	5	5	5	5	4	5	5	5	5	4.9
92	4	5	5	5	5	5	5	5	5	5	4.9
47	5	5	5	5	5	4	4	5	5	5	4.8
57	4	4	4	5	5	4	4	4	4	2	4.0
17	5	5	5	5	5	5	5	5	5	5	5.0
5	5	3	4	4	4	5	5	5	4	3	4.2
62	5	4	5	4	4	5	5	5	5	5	4.7
74	4	3	4	4	4	4	5	4	4	5	4.1
53	4	3	4	4	3	4	3	5	3	4	3.7
84	5	4	4	5	5	5	5	5	5	4	4.7
145	4	4	5	4	5	5	5	5	5	4	4.6
35	5	5	5	5	5	5	5	5	5	5	5.0

	Technical Evaluation Questionnaire Rating Factors										
Subcontract no.	1	2	3	4	5	6	7	8	9	10	Average
59	5	4	5	5	5	5	5	5	5	5	4.9
115	5	5	5	5	5	5	5	5	5	5	5.0
118	4	5	5	5	5	5	5	4	5	4	4.7
104	5	3	4	3	5	4	5	5	4	5	4.3
6	3	5	5	5	3	5	5	4	5	3	4.3
28	5	4	5	5	5	4	5	4	4	4	4.5
51	5	5	4	5	5	5	5	5	5	4	4.8
16	5	5	5	5	5	4	5	5	5	5	4.9
54	5	4	5	4	5	4	5	5	5	5	4.7
56	3	3	5	4	5	3	4	5	5	5	4.2
69	4	5	4	5	5	4	5	5	5	5	4.7
72	5	5	5	5	5	5	5	5	5	5	5.0
129	5	5	5	5	5	5	5	5	5	5	5.0
58	3	4	5	4	3	4	5	4	4	4	4.0
144	5	5	5	5	5	5	5	5	5	5	5.0
80	5	5	5	5	5	5	5	5	4	4	4.8
136	5	4	5	5	5	5	5	5	5	4	4.8
55	5	5	5	5	5	5	5	5	5	4	4.9
77	5	4	3	4	3	4	4	4	5	4	4.0
15	5	4	5	5	5	5	5	4	5	5	4.8
70	5	4	4	5	5	5	5	5	5	4	4.7
140	5	5	5	5	5	5	5	5	5	5	5.0
79	4	3	5	4	5	4	5	5	4	5	4.4
63	5	5	5	5	5	5	5	5	5	5	5.0
97	5	4	4	5	5	5	5	5	5	5	4.8
11	5	4	4	4	4	5	4	4	5	3	4.2
44	5	5	5	5	5	5	5	5	5	5	5.0
141	5	4	5	4	4	5	5	5	5	4	4.6
98	5	5	5	5	5	5	5	5	5	5	5.0

PARALLELISM IN USER ASSISTED INFORMATION EXTRACTION

Pradip Peter Dey
Associate Professor
Computer Science Department

Hampton University
Hampton, VA 23668
Email: dey@cs.hamptonu.edu
Fax: (804)727-5390

Final Report for:
Research Initiation Program
Rome Laboratory

This research was sponsored by Air Force Office of Scientific Research, Bolling AFB, D.C.

December, 1992

PARALLELISM IN USER ASSISTED INFORMATION EXTRACTION

Pradip Peter Dey
Associate Professor
Computer Science Department
Hampton University

ABSTRACT

Since information extraction from formatted and unformatted messages cannot be fully automated with current technology, user assisted information extraction is considered a reasonable strategy for military message processing. User assisted information extraction systems are good candidates for parallel execution; because (1) these are highly complex systems that run slowly on sequential computers, (2) these systems must respond extremely fast, (3) these systems exist in a multicomputer (distributed) environment where proper utilization of resources of several computers is a reasonable strategy. This report presents results of a systematic study of parallel processing architectures, algorithms, data and control structures for a user assisted information extraction system called Generic Intelligence Processor which is being developed in evolutionary stages by Rome Laboratory.

PARALLELISM IN USER ASSISTED INFORMATION EXTRACTION

Pradip Peter Dey

1. INTRODUCTION

Certain military applications require rapid, precise and error-free information extraction (from formatted and unformatted messages) which can be delivered neither by fully automated systems nor by fully manual systems. User assisted information extraction is considered a reasonable strategy for these applications. User assisted information extraction systems are good candidates for parallel execution; because (1) these are highly complex systems that run slowly on sequential computers, (2) these systems must respond extremely fast, (3) these systems exist in a multicomputer (distributed) environment where proper utilization of resources of several computers is a reasonable strategy.

This report presents a systematic study of parallel processing architectures, algorithms, data and control structures for a user assisted information extraction system called Generic Intelligence Processor (GIP) that has been procured by the IRDS branch of Rome Laboratory.

One of the goals of IRDS/RL has been to develop and test GIP in the real world situation. For this purpose an enhanced version of GIP is under development now. In addition, further long-term enhancements of GIP are being planned. With GIP, IRDS has undertaken a challenging task of putting an emerging technology into work. GIP is proposed to be developed in evolutionary stages gradually adding new components and enhancing the system as the technology develops. GIP has enormous potential for technology transfer in some crucial areas as it provides an environment for transferring both short and long-term developments. This proposed research focuses on long-term enhancement of GIP rather than immediate enhancement that is already undertaken. The specific long-term enhancement that is to be addressed here is the parallelization problem of GIP.

2. GENERIC INTELLIGENCE PROCESSOR (GIP)

A detailed description of GIP is available in a technical proposal written by the Knowledge Systems Concepts, Inc. (KSC 1991). The main goal of GIP is to extract information from formatted and unformatted messages and transmit the results to downstream processes such as expert systems. GIP is a toolkit for developing, testing, and delivering message processing applications. When fully developed GIP will provide the following capabilities (a) user selectable input sources, (b) automatic message parsing, (c) automatic/user assisted information extraction, (d) automatic output formatting for development systems, (e) utilities for developing new algorithms. For processing messages with substantial natural language texts user assistance will be crucial for GIP, because, natural language

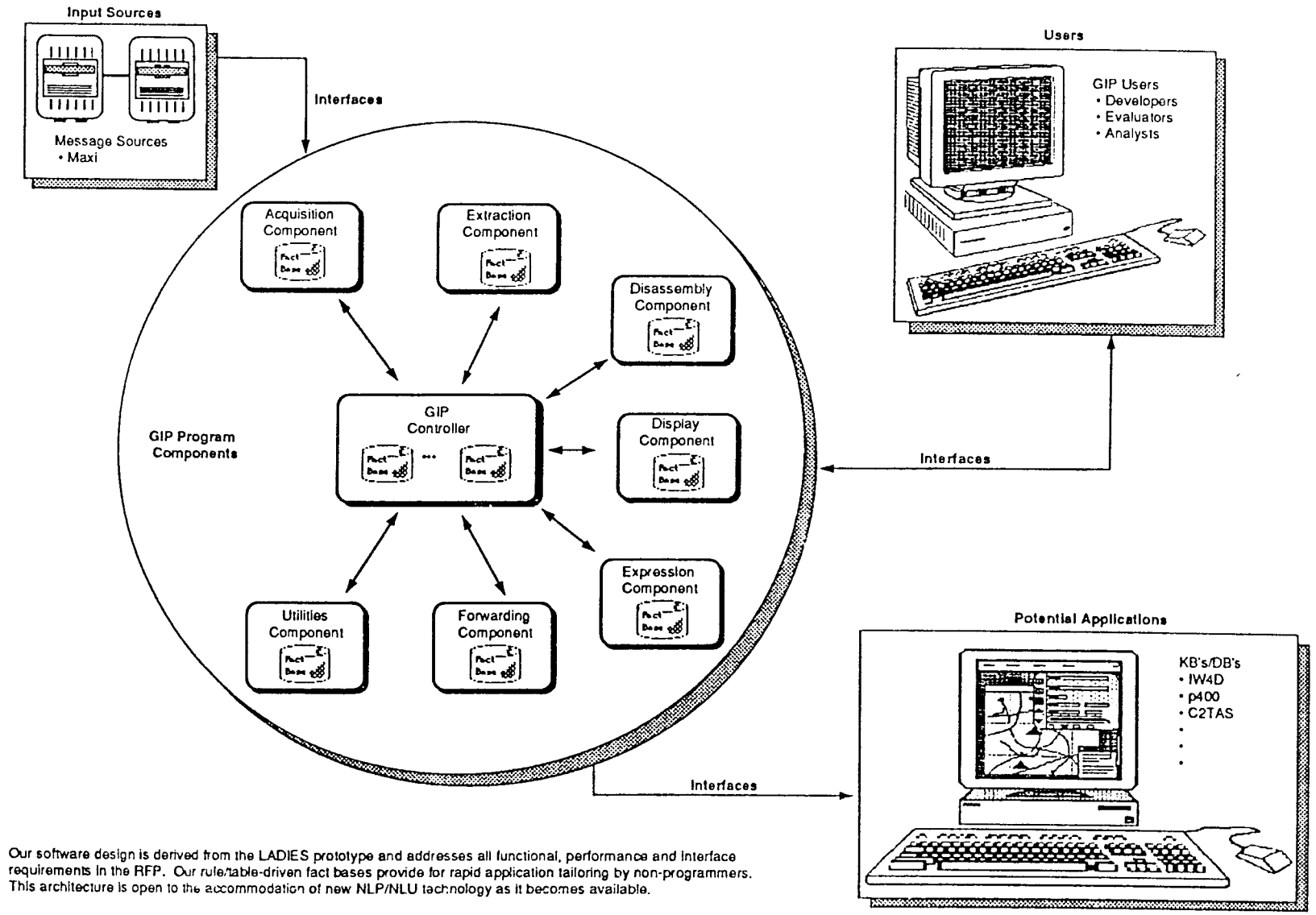


Figure-1: GIP Program Components (From KSC 1991)

processing technology is not mature enough for fully automated information extraction in an operational environment. Therefore, there is considerable emphasis on user interface in the GIP design. The GIP program components are reproduced in Figure-1.

As mentioned earlier, GIP is a complicated system requiring enormous computing power. Efficient execution of GIP is possible only by clever utilization of computing resources in the GIP environment.

3. PARALLEL ARCHITECTURES FOR GIP

There are three basic architecture types: (1) SISD = single instruction single data stream. Traditional sequential computers are of this type. A single instruction is executed at a time on a single piece of data. (2) SIMD = single instruction multiple data stream. A single instruction is executed in a number of processors on different pieces of data. Synchronization is automatic in this architecture. A processor either executes the same instruction with other processors or remains idle. The connection machine is an SIMD machine (Hillis 1985). (3) MIMD = multiple instruction multiple data stream. Different instructions can be executed in different processors on different pieces of data. There are two types of MIMD computers: (a) Multicomputers = loosely coupled systems. In this architecture each processor has its own local memory and communications between the processors are achieved through an interconnection network. They are also known as distributed systems. The n-cube computers are of this type (Athas and Seitz 1988).

(b) Multiprocessors = tightly coupled systems. In these systems, the memory is common to all processors. The processors share a number of global variables stored in the common memory. These variables enable the processors to communicate efficiently through the shared memory if the number of processors is not too large. Two examples of commercially available shared memory multiprocessors are the BBN Butterfly system (Rettberg and Thomas 1986) and the Sequent Balance system (Osterhaug 1985). Algorithms designed for multiprocessors are known as asynchronous parallel algorithms. An asynchronous algorithm is a collection of processes, some or all of which are executed simultaneously on a number of available processors. When the parallel algorithm begins execution on a processor, it creates a number of processes to be performed. If free processors are available, these processes are assigned to the processors to perform the necessary computations. Otherwise, a process is queued and waits for a processor to be free. When a processor completes execution of a process, it becomes free.

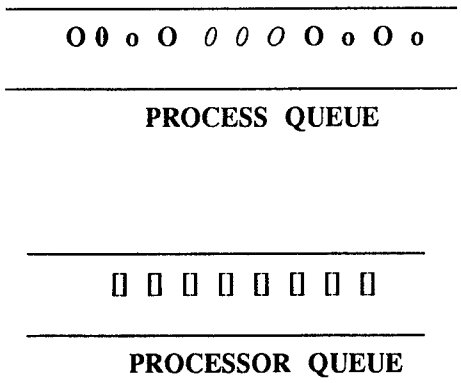


Figure 2. Two Main Queues of a Multiprocessor

Multiprocessors are very popular because they are easy to program. Multiprocessors work with two queues: a process queue and a processor queue. The main thing a programmer has to do is create an optimal number of processes out of a program. The processes are queued in the process queue. Free processors wait in the processor queue. If processes are waiting in the process queue then they are assigned to available processors from the processor queue. After a process is executed, the processor becomes free and is queued back to the processor queue.

Besides these basic architecture types described above there are reconfigurable architectures that can be conveniently configured as shared memory or distributed memory according to computational requirements of specific applications.

There is another interesting but specialized parallel architecture of computation known as the neural network in which a large number of processors (similar to neurons) are interconnected and processing is done by spreading activation (McClelland et al 1986). By passing values processors change the connective strengths among themselves and propagate activations through the network. This propagation constitutes processing by the system. The neural network is an attractive model of parallel computation for building trainable systems (Howells 1988).

When fully developed GIP would be a very large and complex system that would probably require a hybrid architecture for efficient execution of its varied components. For example, neural network architecture may be needed for some of the learning components of GIP; whereas a multicomputer architecture could be used for the natural language component. The shared memory architecture would not suitable for GIP for the communication bottleneck the architecture creates when the number of processors is increased. GIP requires a scalable architecture because the message processing load of GIP would vary from time to time. A cost effective parallel architecture for GIP is a network of work stations with client server computing capabilities. A crucial problem of this architecture is the difficulty in parallelization of software. An experimental parallel execution module has been developed in this architecture using a network of three sun workstations.

4. PARALLEL EXECUTION MODULE

A parallel execution module has been developed that runs on a network of three Sun workstations. It allows execution of any three precompiled programs on three Sun workstations in parallel. When a user invokes the parallel execution module, it creates three distinct processes and executes them on 3 different workstations and obtain the results back to the client machine. The user has the freedom to choose the programs to be executed in parallel. The remote execution module is implemented using rcp and rsh protocols of the Sun Microsystems. This module has been delivered and demonstrated to Rome Laboratory in August 1992. This module produces large grain parallelism at a high level. Distributed execution of large GIP components by this module is highly desirable for adequate speedup. It takes about 4 seconds for a process to migrate to a remote server site with its data. Therefore, there would not be substantial speedup by parallel execution of small GIP modules. The level-1 data flow diagram (DFD) of the parallel execution module is presented in Figure 3.

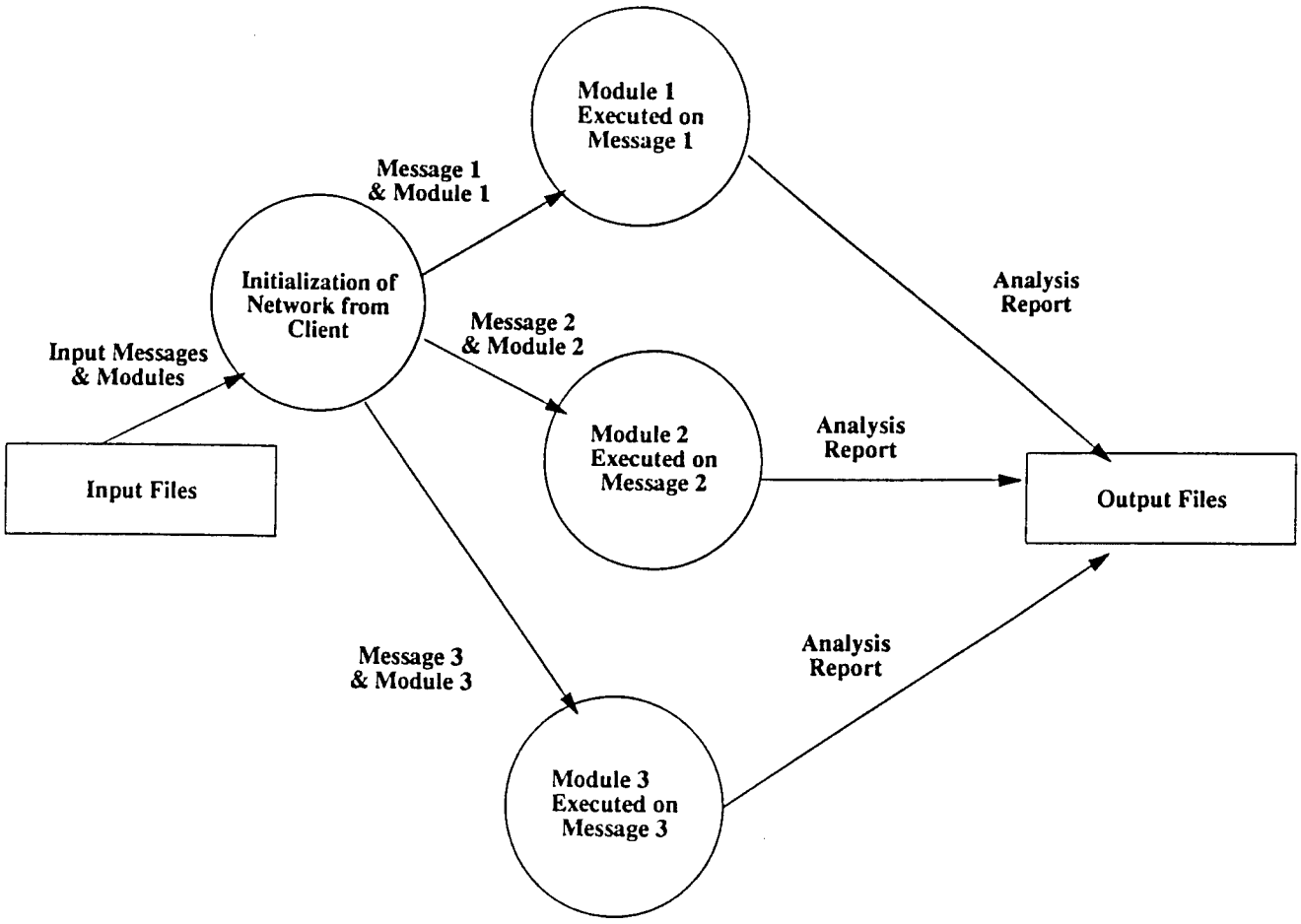


Figure 3. Data Flow Diagram of the Parallel Execution Module

5. DATA STRUCTURES

Input to GIP is a sequence of characters given from text files. After a preliminary error checking process the data is broken into tokens. Structural and semantic analyses are then performed for information extraction. The extracted information is then arranged in attributes and values. The output information is formatted according to the down stream processes. The parallel execution module can adequately deal with the data structures required by GIP.

6. NATURAL LANGUAGE PROCESSING IN THE GIP ENVIRONMENT

Although Natural Language Processing (NLP) technology is not mature enough for fully automated systems, it is mature enough to be put in certain real world use with appropriate user assistance. Information extraction from messages is one of the most promising fields for this technology. User assistance would be required in the extraction process until some future break-through makes totally automated natural language understanding possible. GIP requires robust, user-oriented, modifiable, parallelizable NLP modules. GIP is recommended to have at least two alternative NLP modules so that users would be able to compare the results of the two NLP modules. These two NLP modules should be run in parallel for obtaining efficient response time so that user assistance can be planned within reasonable time limit.

Robustness and Scalability should be among the major goals of the NLP modules. NLP systems require large knowledge bases for achieving robustness. Some learning methods should be used to partially automate the system development process. Learning methods should contribute to scalability by making it feasible to develop the knowledge bases efficiently. Learning methods should also contribute to enhancibility of the system. The requirements of the NLP modules of GIP can be briefly specified as follows:

(1) The GIP NLP modules should have partially automated knowledge acquisition methods. That means some learning algorithms should be used for acquiring domain knowledge. GIP would then be able to deal with the diversity of domains.

(2) GIP should use some statistical methods to guide processing. Statistical methods can be successfully used in dealing with ambiguity.

(3) GIP would require well-documented NLP modules. In order to integrate the NLP modules, documents like algorithm descriptions, software design documents and user's manual would be necessary. Software development models used for the modules should be briefly described.

(4) The NLP component should be properly modularized. The interfaces between modules should be well-defined.

(5) The NLP modules should allow parallelization for efficient execution.

(6) Scalability should be considered as an essential feature of the NLP modules. It is often required to incorporate this provision in the design level; otherwise, the modules have to be redesigned.

(7) The NLP modules should be modifiable, otherwise GIP will not be able to deal with the emerging technology.

9. BIBLIOGRAPHY

Aho, A.V. and J. D. Ullman. 1972. The theory of parsing, translation and Compiling: volume 1. Englewood Cliffs, N.J.: Prentice-Hall.

Akl, S. 1985. Parallel Sorting Algorithms, Academic Press.

Alshawi, Hiyan 1990. Resolving Quasi Logical Forms, Computational Linguistics, 16, 133-144.

Alshawi, Hiyan and J. van Eijck 1989. Logical Forms in the Core Language Engine, 27th Annual Meeting of the Association for Computational Linguistics.

Athas, W. C. and C. L. Seitz (1988) "Multicomputers: Message Passing Concurrent Computers" *IEEE Computer* 21: 9-24.

Barton, G. E., Berwick, R., and Ristad, E. 1987. *Computational Complexity and Natural Language*, MIT Press.

Beizer, B. 1990. *Software Testing Techniques*, Van Nostrand Reinhold.

Bobrow, D.G. and A. Collins 1975, (eds.) 1975. *Representation and Understanding*, New York: Academic Press.

Chomsky, N. 1957. *Syntactic structures*, Mouton, The Hague.

Chomsky, N. 1965. *Aspects of the theory of syntax*, Cambridge, Ma: MIT Press.

Chomsky, N. 1981. *Lectures on Government and Binding*, Dordrecht: Foris.

Conery, J. S., 1987, The AND/OR Process Model for Parallel Execution of Logic Programs. Kluwer Academic Publishers, Boston.

Cottrell, Garrison W. 1984, A Model of Lexical Access of Ambiguous Words, Proceedings of AAAI-84.

Dey, P., and Hayashi, Y. 1990, "A multiprocessing model of natural language processing," Theoretical Linguistics, 16, 11-23.

Dey, P., Bryant, B. and Takaoka, T. 1990, "Lexical ambiguity in tree tree adjoining grammars," Information Processing Letters, 34, 65-69.

Dey, P., Whitlock, D. and Lim, J. S. 1991a, "Parallel Best-first Search Strategies," Proceedings of the 4th International Conference on Parallel and Distributed Computing and Systems, 369-373.

Dey, P. 1991b, "User Assisted Information Extraction", Report on the 1991-AFOSR research in the IRDS branch of Rome Laboratory submitted to RDL.

Earley, J. 1970. "An efficient context-free parsing algorithm". Communications of the ACM 13, 94-102.

Gazdar, G. 1981. "Unbounded dependencies and coordinate structure". Linguistic Inquiry 12, 155-184.

Hass, A. 1990. Sentential Semantics for Propositional Attitudes", Computational Linguistics, 16, 1990, 213-233.

Hillis, D. and G. L. Steele Jr. 1986. "Data Parallel Algorithms," Communications of the ACM, 29, 1170-1183.

Hirst, G. J. 1984 Semantic Interpretation against ambiguity, Ph.D. diss. Brown University.

Howells, T. (1988) "Vital - A Connectionist Parser" *Proceedings of the 10th Annual Meeting of the Cognitive Science Society*, Canada.

Jackendoff, R. S. 1972. *Semantic Interpretation in generative grammar* , MIT Press.

Jamieson, L. H., D. B. Gannon and R. J. Douglass (eds). 1987. The Characteristics of Parallel Algorithms, MIT Press.

Joshi, A. K. 1985. "Tree Adjoining Grammars: How much context-sensitivity is required to provide reasonable structural descriptions?" Dowty, et al (eds.) Natural Language Parsing.

Kaplan, D. 1975. "Quantifying in", in Davidson et al eds. *The Logic of Grammar,*

Keenan, E. L. and L. M. Faltz 1985, *Boolean Semantics for Natural Language,* Dordrecht: D. Reidel.

Kowalski, R. A., 1974, "Predicate Logic as a Programming Language." In Information Processing 74, IFIP No. 5, 733-742.

KSC 1991. Generic Intelligence Processor (GIP), Technical Proposal, Volume III, Knowledge Systems Concepts Inc. , Rome, NY.

Kuck, D. J. 1977, "A Survey of Parallel Machine Organization and Programming," ACM Comput. Survey, 9, 29-59.

Kumar, V. and Kanal, L. N. 1984, "Parallel Branch and Bound Formulations for understanding and synthesising AND/OR TREE search. IEEE Trans. Pattern Analysis and Machine Intell. 6, 768-778.

Kung, H. T. 1980. The Structure of Parallel Algorithms. Adv. Comput. 19, New York: Academic Press, 65-112.

Lenat, D. et al 1990. "CYC: Toward Programs With Common Sense", Communication of the ACM, 33, 30-49.

Marcus, M. 1980. A theory of syntactic recognition for natural language. Cambridge, MA: MIT Press.

May, R. 1985. Logical Form. Cambridge, MA: MIT Press.

McCawley, J. D. 1968. "The role of semantics in a grammar". In E. Bach and R.T. Harms (eds), Universals in Linguistic Theory, New York: Holt, Rinehart and Winston.

Milne, R. 1986. Resolving Lexical Ambiguity in a Deterministic Parser", Computational Linguistics, 12, 1-12.

Morell, L. 1990. A Theory of Fault-based Testing, IEEE transactions on Software Engineering, 844-857.

Ottmann, T. A., Rosenberg, A. L. and J. L. Stockmeyer. 1982, "A Dictionary Machine (for VLSI)," *IEEE Transactions on Computers,* 31, 9, 892-897.

Raynal, M. 1985. *Algorithms for Parallel Processing,* MIT Press.

Schank, R. and L. Birnbaum 1980. "Memory, Meaning, and Syntax," Report 189, Dept. of Comp. Science, Yale Univ.

Schmidt, D. A., 1986, *Denotational Semantics*, Allyn and Bacon, Inc., Boston.

Schubert, L. and F. J. Pelletier 1982, "From English to Logic: Context-Free Computation of 'Conventional' Logical Translation," Computational Linguistics 8, 27-44.

Snider, C. 1979. Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse. Ph.D. Diss. MIT.

Somani, A. K. and V. K. Agarwal. 1984, "An efficient VLSI dictionary machine," *Proceedings of the 11th Annual ACM Int. Symp. Comput. Architecture,* 142-150,

Steinberg, D. and L. A. Jakobovits (eds) 1971, *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology.*

Stockwell, R. P., Schachter, P. and B. H. Partee 1973, *The Major Syntactic Structures of English.* New York: Holt, Rinehart and Winston.

Tennant, H. 1981. Natural language processing. New York: Petrocelli.

Voas, J., Morell, L. and K. Miller 1991. Prediction Where Faults can Hide from Testing, IEEE Software, 41-48.

Waltz, D. and J. Pollack. 1984. "Phenomenologically Plausible Parsing", AAAI-84. 335-339.

Warren, D. S. and J. Friedman 1982. "Using Semantics in Non-Context-Free Parsing of Montague Grammar," Computational Linguistics 8, 123-138.

Warren, H. D. and F. C. N. Pereira 1982, "An Efficient Easily Adaptable System for Interpreting Natural Language Queries," Computational Linguistics 8, 110-122.

Webber, B. 1979. *A Formal Approach to Discourse Anaphora,* New York: Garland.

Winograd, T. 1972. *Understanding Natural Language,* New York: Academic Press.

Winograd, T. 1983. *Language as a cognitive process: Syntax.* Reading, Mass.: Addison-Wesley.

Woods, W.A. 1970. "Transition Network Grammars for Natural Language Analysis". Communications of the ACM, 13, 591-606.

Woods, W.A. 1968. "Procedural Semantics for a question answering machine", Fall Joint Computer Conference, 457-471.

NOISE MODELING OF MILLIMETER-WAVE FIELD EFFECT TRANSISTORS

Steven Lardizabal, Graduate Student
and
Lawrence P. Dunleavy, Assistant Professor
Department of Electrical Engineering

University of South Florida
4202 East Fowler Avenue
Tampa, Florida 33624

Final Report for:
Research Initiation Program
Rome Laboratory, RL/ERAC

Sponsored by:

Air Force Office of Scientific Research
in cooperation with Rome Laboratory
Hanscom Air Force Base, Boston, MA

and

University of South Florida

December 1992

NOISE MODELING OF MILLIMETER-WAVE FIELD EFFECT TRANSISTORS

Steven Lardizabal, Graduate Student
and
Lawrence P. Dunleavy, Assistant Professor
Department of Electrical Engineering
Tampa, Florida 33624

Abstract

An efficient noise modeling method for millimeter-wave field effect transistors is described. The method differs from conventional techniques by avoiding complications associated with the use of a variable impedance/admittance tuner at the input of the FET. Noise figure is measured at one known source admittance state over several frequencies. A knowledge of the FET equivalent circuit model is assumed, and an extraction of the model parameter values is demonstrated on both a low noise MESFET and PHEMT. Noise correlation matrix techniques are applied to express the noise figure equation in a generalized format that allows various intrinsic FET noise source representations. Noise parameter data is measured at microwave frequencies using an established tuner based system for model comparisons. The sensitivity of the noise modeling method and measurements are addressed.

NOISE MODELING OF MILLIMETER-WAVE

FIELD EFFECT TRANSISTORS

TABLE OF CONTENTS

I. Executive Summary

New noise modeling research is presented to eliminate expensive hand tuning and improve the effectiveness of computer-aided-design of millimeter-wave (mm-wave) FET low noise amplifiers. This research is very relevant to the goals of the Air Force and of Rome Laboratory towards the advancement of high performance mm-wave command and control communications and intelligence (C^3I) systems. These systems should benefit substantially from the inclusion of low noise amplifiers (LNA's) that employ state-of-the-art mm-wave FETs (including MESFETs and HEMTs), and monolithic microwave/mm-wave integrated circuit (MMIC) technology. However, these potential benefits cannot be realized until accurate mm-wave FET noise characterization becomes possible. Accurate mm-wave FET noise characterization is not currently possible with conventional techniques used at microwave frequencies (i.e. below 30GHz).

The research completed here supports on-going in-house Air Force research in the Component Technology Branch (ERAC) at Rome Laboratory. The modeling technique complements, and may eventually replace, the difficult experimental noise characterization of mm-wave Indium Phosphide High Electron Mobility Transistors (HEMTs). These HEMTs are currently being used by RL/ERAC in 60GHz monolithic low noise amplifiers. Related RL/ERAC research will continue with 60GHz monolithic receive circuits and 120GHz LNAs. The noise modeling technique reported here is expected to contribute to both of these efforts.

The research advances a new method for noise modeling of mm-wave field effect transistors. The new noise modeling method was the subject of a 1991 USAF-AFOSR Summer Research Program. The method avoids the complications of a variable impedance tuner, and requires only the knowledge of a small signal equivalent circuit, and noise figures measured across a range of frequencies for a single known source impedance. Noise parameters derived from this method have demonstrated promising results in comparison with measured data taken at microwave frequencies. However, the method required further enhancement and verification for reliable application to mm-wave noise modeling development. The specific tasks of the 1992 research initiation program are shown in Figure 1 and

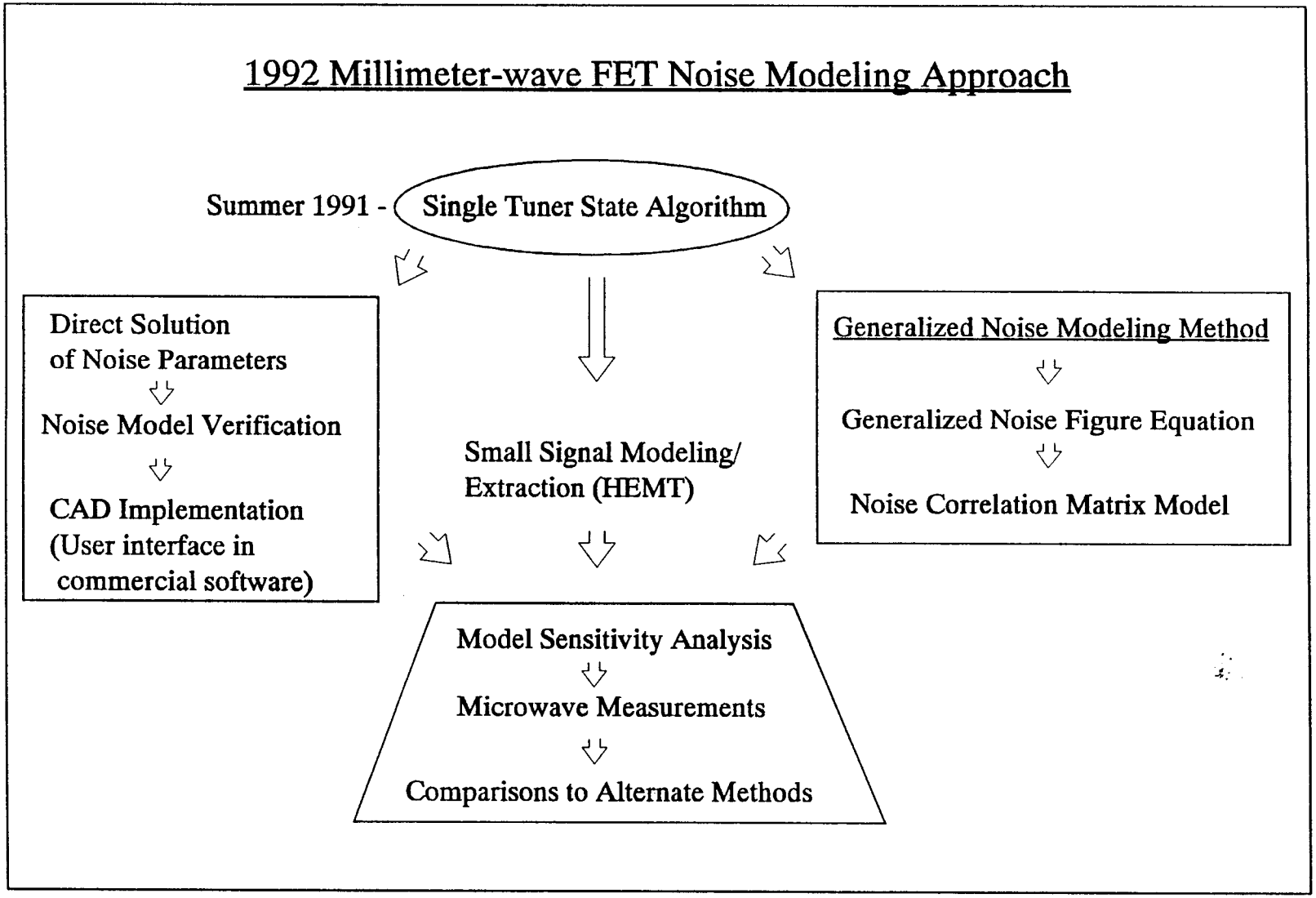


Figure 1 Millimeter-wave noise modeling development plan for 1992.

were carried out as five tasks as follows:

1) Implement necessary improvements to the method, including utilization of a more general formulation, and a modified computer algorithm.

2) Investigate the method's sensitivity to measurement errors, and (if warranted) explore the use of alternate computational techniques for improving the method's tolerance to measurement errors.

3) Implement an improved small signal modeling method to separate the intrinsic and extrinsic FET equivalent circuit parameters in a more physically meaningful and consistent way.

4) Perform microwave noise figure measurements for method verification, and to examine the validity of frequency extrapolation of the derived noise model.

5) Conducted a theoretical investigation of alternative mm-wave noise modeling techniques.

This research provides a much needed mm-wave FET noise characterization method. The use of FET models derived from this method will reduce design costs, and yield high performance MMIC LNAs that, in turn, will improve the performance and reduce the costs of the Air Force's Future C^3I systems.

II. OBJECTIVES AND ACCOMPLISHMENTS

Problem: Conventional microwave noise modeling methods utilize a variable impedance (or admittance) tuner, which presents a sequence of "known" source impedances Z_s (admittances Y_s) to a device's input terminals. The tuner must be able to present a range of impedances that represents a good coverage of the Smith Chart, including near the edges of the chart. Hence, the tuner must have very low loss. The ability to fabricate, and characterize, such a tuner becomes increasingly more difficult as frequency is increased. Research and commercial developments are being pursued in the industry and at the Air Force towards implementing mm-wave tuners for noise parameter measurements. Because of the inherent difficulties with this approach, however, it is essential that alternatives be explored to the reliance on a variable impedance tuner for mm-wave FET noise characterization.

Research Objectives: The objectives of the performed work were to: advance and verify a new method for FET noise model development formulated as part of a Summer research project[1]. Long range goals for the anticipated multi-year research initiated with this project are to investigate FET noise modeling methods applicable at mm-wave frequencies (i.e. frequency > 30GHz). Also to be investigated are the inclusion of temperature effects in microwave and mm-wave noise models.

Scope of Research: The performed work comprised the first year of a multi-year investigation of microwave and mm-wave device noise characterization. Advancement of the new FET noise modeling method discussed above was carried out at the University of South Florida. The work involved computer program development , FET measurements, CAD program use, a theoretical investigation, and numerical experiments.

Specific Objectives vs. Accomplishments: A summary of the tasks performed and the objectives that were accomplished are given as follows:

1) Objective - Advance and refine the new noise modeling method.

Accomplishment - A direct solution for the noise parameters was derived using

noise correlation matrix methods replacing the synthesized tuner measurement of the previous algorithm (See Section III. A.). In addition, a generalized noise modeling method was developed using the correlation matrix methods learned by implementing the direct noise parameter solution (See Section III. B.).

2) Objective - Investigate method sensitivity to errors.
Accomplishment - Variations of the noise coefficients P, R, and C were studied with respect to sets of statistically perturbed sets of input data (noise figure and reflection coefficient). See Section V.

3) Objective - Implement an improved small signal modeling method.
Accomplishment - After a review of FET small signal model extraction techniques it was concluded that the extraction of the device parasitic components was most important to improving existing extraction techniques. Therefore, a paper study was conducted with corresponding measurements for the extraction of MESFET and HEMT extrinsic parameter element values. Subsequently, a procedure was documented, where small signal measurements were made and model element values extracted for both a low-noise MESFET and a PHEMT (See Section IV). Comparisons were made to a conventionally derived "brute force optimized" model. Microwave noise figure and source reflection coefficient measurements were performed at U.S.F. and used to compare the resulting P, R, and C values using both small signal models.

4) Method verification with microwave noise measurements.
Accomplishments - An automatic noise parameter test set (ATN's NP5a) was assembled and configured as part of this project for operation with on-wafer microwave S-parameter measurement capabilities at U.S.F. Two devices were characterized using the ATN system over a 2.0 to 18.0GHz range. The noise parameters of a MESFET and a PHEMT were used for noise modeling purposes.

III. Advancement/Refinement of the Noise Modeling Method

The development of a noise modeling algorithm that only requires knowledge of the FET equivalent circuit parameters and the measurement of noise figure over several frequencies at one known source impedance has been described[1]. This technique can significantly simplify the determination of device noise parameters at millimeter-wave frequencies, because the technique is not reliant upon a variable source impedance tuner. The previously developed algorithm [1,2] is shown in Figure 2.

The previous approach proceeds with a determination of the noise coefficients P, R, and C, through a least squares fit of data to a linearized equation for the noise figure. Then the device noise parameters are solved for by numerically synthesizing a tuner based measurement data set and calculating the noise parameters using a standard technique[3]. This artificial measurement was implemented, because a direct solution for the noise parameters was not obvious. This is due to the additional algebraic complexity that arises when the circuit elements C_{dc} and C_{dg} are incorporated in the FET model.

In this report a direct solution is obtained for the noise parameters. The solution involves a matrix based representation of the equivalent circuit of Figure 3, which applies to various topologies and noise source representations. The noise correlation matrix method[4] is used to develop a direct solution for the noise parameters, that simplifies the algorithm of Figure 2.

A Direct Solution for the Noise Parameters

The parameters P, R, and C in the FET noise model of Figure 3. are necessary to describe the intrinsic gate and drain noise sources i_g and i_d. The noise correlation matrix, previously described[4], allows for Z, Y, and ABCD matrix

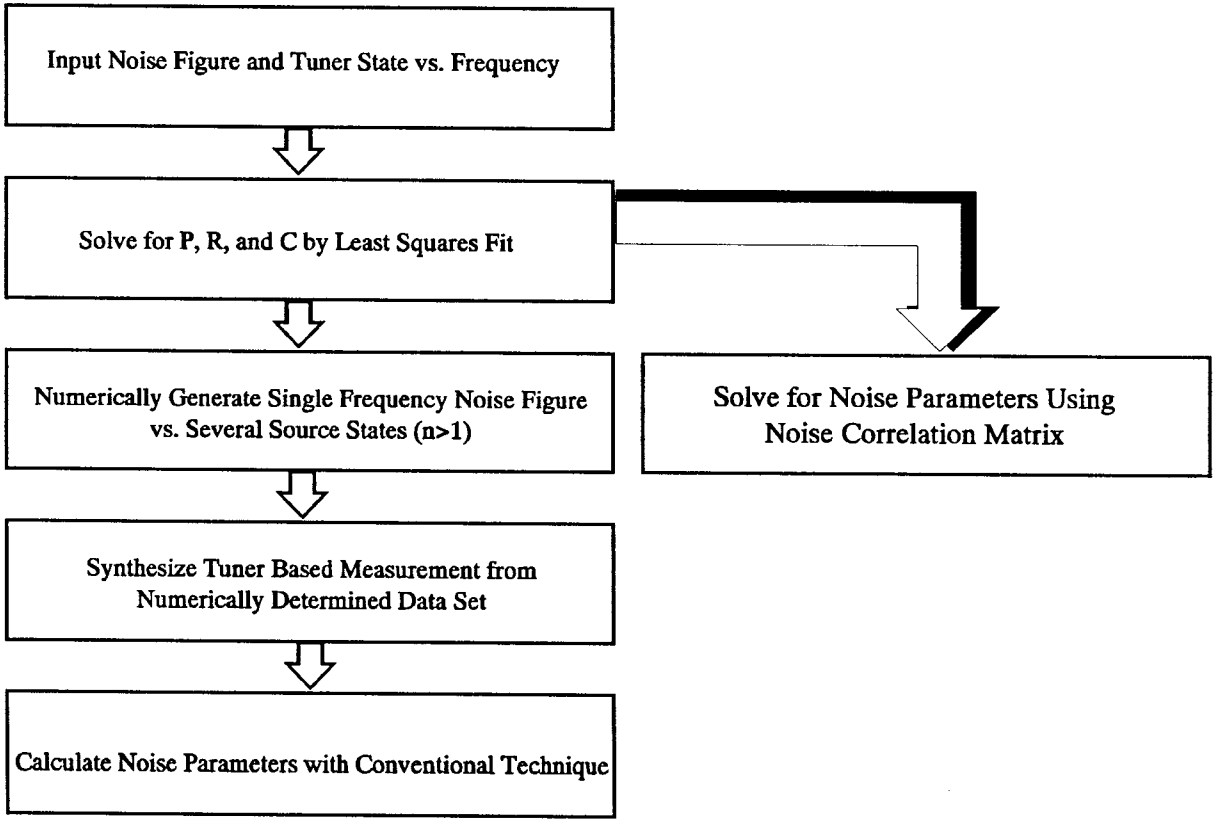


Figure 2 Comparison of previous noise modeling algorithm (left) and the algorithm resulting from this work (right).

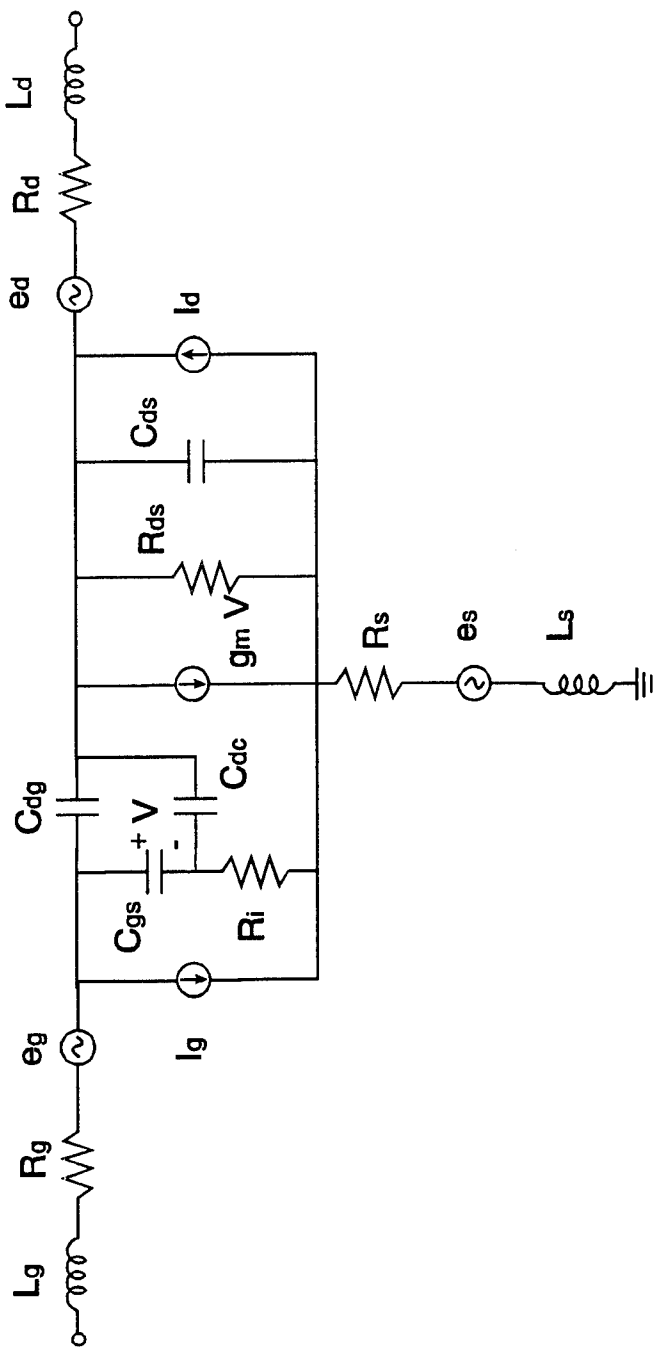


Figure 3 Field effect transistor small signal and noise model used for this work.

representations of the noise power spectral densities related to the noise sources and their correlation. Consider a set of noise matrices associated with the interconnection of two-ports shown in Figure 4. These noise matrices can be transformed between the various representations with the appropriate transform matrix, and a knowledge of the circuit elements through which the transform is applied. Therefore, with P, R, and C determined by the single tuner state algorithm, a synthesis of the various two-ports that comprise the equivalent circuit model defines the noise correlation matrix of the total extrinsic FET model.

The total extrinsic noise correlation matrix is developed by first describing the intrinsic FET, and then interconnecting the parasitic networks according to Figure 4. This is done by determining the coefficients P, R, and C associated with the intrinsic network. The intrinsic FET correlation matrix is then represented in a Y parameter form as:

$$C_{yi} = \begin{bmatrix} \dfrac{\omega C_{gs} R}{g_m} & j\omega C_{gs} C\sqrt{RP} \\ -j\omega C_{gs} C\sqrt{RP} & g_m P \end{bmatrix} \qquad (1)$$

The analysis for the interconnection of noisy two-ports parallels that of linear network theory. The details of converting and connecting noise matrices is addressed by Hillebrand and Russer[4]. This interconnection of noisy two-ports proceeds as follows:

1. Form the series connection of the intrinsic FET network and the source parasitic network to develop the noise matrix $C_{Z,is}$ and the related network matrix Z_{is} .
2. Transform the Z-parameter matrices of step 1. into ABCD-parameter forms.
3. Form the ABCD noise and network matrices for the gate and drain parasitic networks.
4. Cascade the interconnected linear two-port ABCD matrices to form the complete model network matrix and noise correlation matrix C_A.

Certain network parameter representations may not exist in general for all

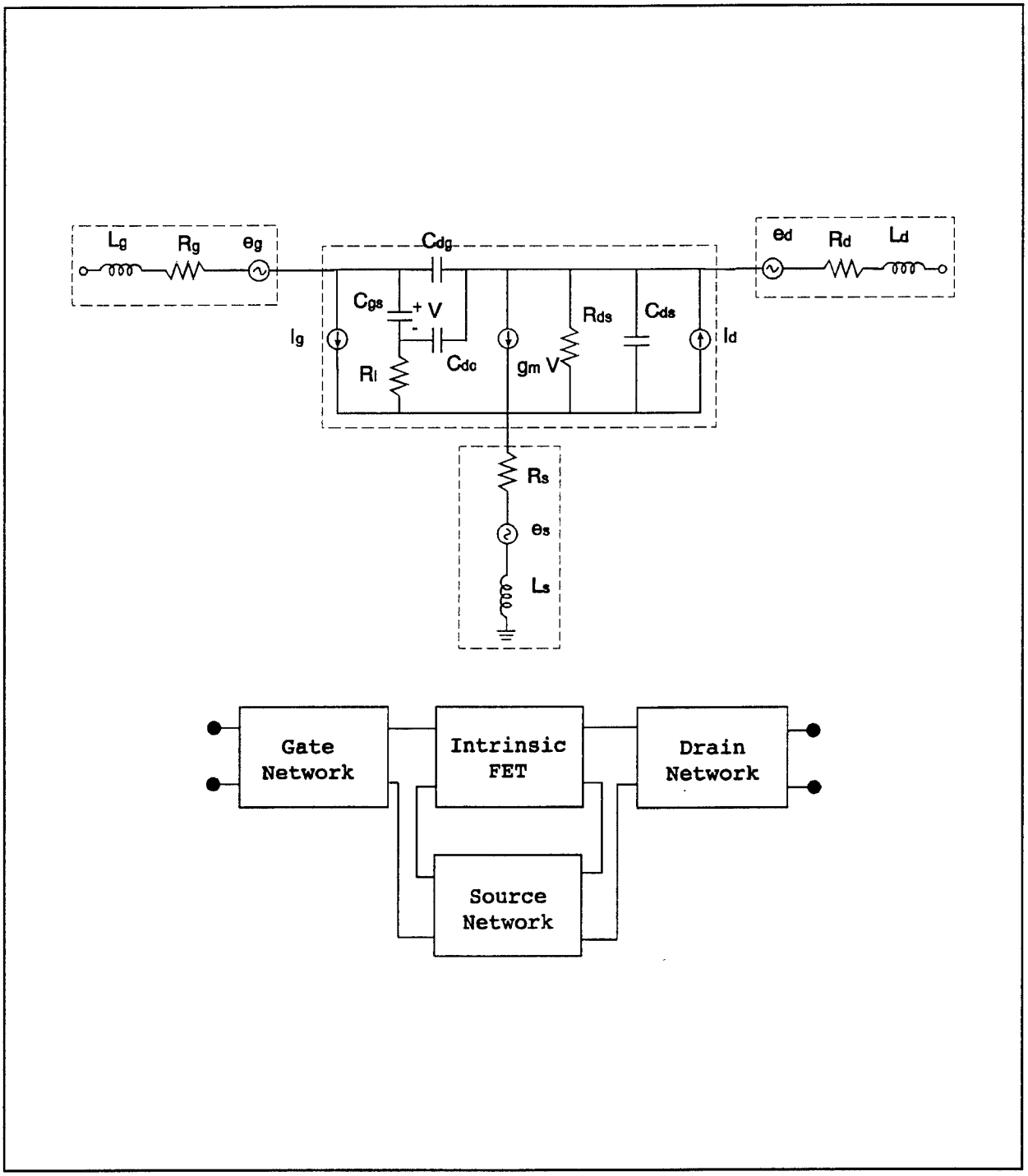


Figure 4 The small signal equivalent circuit element model (top) is redrawn as an interconnection of two-ports (bottom). Each two-port has a related noise correlation matrix and a network parameter matrix.

elements. For example, the series Z-parameter representation of the gate and drain parasitics does not exist. Therefore, the computer algorithm reflects the basic transforms required to achieve the final ABCD or chain matrix representation. The noise correlation matrices given here have been normalized by a factor of 2kT to simplify the expressions, this term ultimately factors out of the final noise parameter equations. The resulting chain matrix, C_A, is given in terms of the FET noise parameters by[4]:

$$C_A = \begin{bmatrix} R_n & \frac{F_{\min} - 1}{2} - R_n Y_{s,opt} \\ \frac{F_{\min} - 1}{2} - R_n Y_{s,opt}^* & R_n |Y_{s,opt}| \end{bmatrix} \tag{2}$$

The noise parameters may now be solved for directly with the following results[4,5]:

$$\begin{aligned} F_{\min} &= 1 + 2\left(C_{A,12} + C_{A,11} Y_{s,opt}^*\right) \\ R_n &= C_{A,11} \\ Y_{s,opt} &= \sqrt{\frac{C_{A,22}}{C_{A,11}} - \left[I_m \frac{C_{A,12}}{C_{A,11}}\right]^2} + j\left[I_m \frac{C_{A,12}}{C_{A,11}}\right] \end{aligned} \tag{3}$$

It is notable that, for the field effect transistor model commonly used at microwave and millimeter-wave frequencies, the matrix C_A forms a standard noise model, through which various noise source representations can be compared. That is, a particular intrinsic noise source representation can then be implemented within this model, provided that these sources are known. The following section extends the use of the correlation matrix technique and the formulation of the single source state noise modeling method by generalizing the noise figure equation from which the intrinsic sources are described.

A Generalized Noise Figure Equation

Essentially, the noise modeling method presented above, has two input data sets: 1)the small signal model parameters and 2)the noise coefficients P, R, and C. These coefficients are calculated from a linearized form of the noise figure equation, which was derived from a nodal circuit analysis[1,2]. In this work an equation for device noise figure is derived using noise correlation matrices.

The noise figure of an active two-port is described in terms of its noise parameters by

$$F = F_{\min} + 4\frac{R_N}{Z_o}\frac{\left|\Gamma_s - \Gamma_{opt}\right|^2}{\left|1 + \Gamma_{opt}\right|^2\left(1 - \left|\Gamma_s\right|^2\right)} . \qquad (4)$$

The noise figure associated with a particular source tuner state, Γ_s, is the most directly measurable parameter in (4). Automated systems make use of redundant sets of noise figure data for the analysis of device noise parameters[3,6], avoiding the alternative search for F_{min} and Γ_{opt}[7]. Again, the method described here requires a small signal equivalent circuit model and a measurement of noise figure over several frequencies at only one source admittance. The intrinsic FET noise model is characterized by the equivalent gate and drain noise current sources i_g and i_d respectively[8]. These intrinsic noise sources are in general defined by the user. These are the intrinsic sources that were described by P, R, and C in the noise correlation matrix of the direct solution. The noise figure is written is terms of the total FET noise correlation matrix as,

$$F = 1 + \frac{Z^{+}\left[C_A\right]Z}{\mathbf{R}\{Z_{gen}\}} \qquad (5)$$

where the intrinsic FET correlation matrix is embedded within the extrinsic equivalent circuit model matrix Z. Therefore, the FET modeling may in general be implemented with a two parameter noise model using noise temperatures T_g and T_d[9] or a four parameter model with sources e_g and i_d having complex correlation[5]. For the present sources, i_g and i_d, a fit is made to the measured noise figure by the parameters P, R, and C[1,2]. The parameter C is complex, but assumed only

imaginary for this analysis[8].

The noise figure equation of a pair of correlated noise sources with spectral intensities $|i_g^2|$, $|i_d^2|$, $|i_d i_g^*|$ and $|i_g i_d^*|$ is in general a quadratic form that for the case of i_g and i_d may be represented in terms of the noise coefficients P, R, and C as

$$F = M_1 + M_2 R + M_3 P + M_4 C\sqrt{RP}. \tag{6}$$

By identifying a noise correlation matrix for the intrinsic FET model and the embedding parasitics, a generalized algorithm is readily developed for FET noise modeling. Correlation matrices provide a direct means of determining the "M" coefficients of equation (6) for various small signal equivalent circuit model topologies. The following analysis will reformulate the coefficients M_1, M_2, M_3, and M_4 in the noise figure equation through application of the correlation matrix C_A, which represents the complete extrinsic FET noise correlation matrix. The noise figure equation in terms of C_A is given by (5) where

$$Z = \begin{bmatrix} 1 \\ Z_{gen} \end{bmatrix}.$$

Now, in equation (7) the noise correlation matrix of the intrinsic FET, Czi,

$$F = 1 + \frac{Z^+ \left(T + \tilde{B} C_{zi} \tilde{B}^+\right) Z}{\mathbb{R}\{Z_{gen}\}} \tag{7}$$

results from the correlated gate and drain noise current sources of the small signal FET model. This intrinsic noise correlation matrix was previously written in equation (1), in admittance form, for a three parameter model with noise coefficients P, R, and C.

The $\tilde{B}$ elements of (7) are noise transformation matrices for a Z-matrix to chain-matrix transform. For the FET model all the components of (2) are 2x2 matrices and, the contribution of the extrinsic parasitic networks are elements of the matrix T. Combining (5) and (7), the coefficient M_1 is written in terms of the matrix T :

$$M_1 = 1 + \frac{Z^+ T Z}{\mathbb{R}\{Z_{gen}\}} \tag{8}$$

Before solving for the coefficients M_2, M_3, M_4, by expanding the intrinsic term $\tilde{B}\, C_{zi}\, \tilde{B}^+$, let $\bar{C}$ represent this intrinsic term. From equation (7) comes the quadratic form:

$$\bar{C} = \bar{C}_{11} + \bar{C}_{21} Z^*_{gen} + \bar{C}_{12} Z_{gen} + \bar{C}_{22} Z_{gen} Z^*_{gen} \tag{9}$$

which, is used to relate the expansion of $\tilde{B}\, C_{zi}\, \tilde{B}^+$ to the $M_{2,3,4}$ coefficients. Using the noise correlation matrix representation the intrinsic term is:

$$\bar{C} = A_g\, \tilde{A}_{is}\, Z_i\, C_{yi}\, Z_i^+\, \tilde{A}_{is}^+\, A_g^+ \tag{10}$$

where, A_g= gate parasitic network chain-matrix, Z_i = intrinsic FET impedance matrix, and A_{is} = the Z to chain-matrix noise transform of the intrinsic FET and the series connected source network. For simplicity (10) is rewritten as:

$$\bar{C} = X\, C_{yi}\, Y \quad , where\ X = A_g\, A_{is}\, Z_i\ and\ Y = X^+ . \tag{11}$$

Equating (11) with (9) provides a solution for the $M_{2,3,4}$ coefficients:

$$M_2 = \left[X_{11} Y_{11} + X_{11} Y_{12} Z_{gen} + X_{21} Y_{11} Z^*_{gen} + X_{21} Y_{12} Z_{gen} Z^*_{gen}\right] \frac{(\omega C_{gs})^2}{g_m} \frac{1}{\mathbb{R}\{Z_{gen}\}} \tag{12}$$

$$M_3 = \left[X_{12}Y_{21} + X_{12}Y_{22}Z_{gen} + X_{22}Y_{21}Z_{gen}^* + X_{22}Y_{22}Z_{gen}Z_{gen}^*\right] \frac{g_m}{\mathbb{R}\{Z_{gen}\}} \quad (13)$$

$$M_4 = \left[\begin{array}{c} (X_{11}Y_{21} - X_{12}Y_{11}) + (X_{11}Y_{22} - X_{12}Y_{12})Z_{gen} + \\ (X_{21}Y_{21} - X_{22}Y_{11})Z_{gen}^* + (X_{21}Y_{22} - X_{22}Y_{12})Z_{gen}Z_{gen}^* \end{array}\right] \frac{j\omega C_{gs}}{\mathbb{R}\{Z_{gen}\}} \quad (14)$$

With the M_1 and $M_{2,3,4}$ determined, the coefficients P, R, and C are determined from the noise figure data, which is modeled by a least squares fitting of a linearized form of equation (2). Once the internal noise sources are de-embedded and characterized, the extrinsic FET correlation matrix is determined over the simulation frequency range by embedding the characterized sources within the noise model as presented in the previous section.

This development can be quickly extended to a four parameter model with noise coefficient $C = C' + jC''$. The resulting model would include a value M_5 corresponding to the real part, (C') of the coefficient C. Additionally, the two parameter model of Pospieszalski[9] could be implemented with parameters T_g and T_d representing equivalent gate and drain noise temperatures. Coupled with the direct noise parameter solution, a general noise modeling algorithm has been developed that can be used to implement various noise source representation within the single source state modeling method.

IV. Small Signal Modeling/Extraction/Extrapolation

Measurements and extraction procedures required to develop accurate millimeter-wave FET small signal models are discussed in this section. An accurate extraction of small signal equivalent circuit parameter values (ECP) relates directly to the integrity of the resulting noise model. Here a brief discussion of several small signal model investigations is first given. Then a following section presents the results of a measurement based modeling method.

How high in frequency can the extracted ECPs be used for extrapolation of S-parameters beyond the range of measured data? At issue is the model and measurement accuracy and how sensitive the resulting representation is. Early work by Yeom et. al[10] and Curtice and Camisa[11] describes the resulting quality of an extrapolation versus frequency and suggest an accurate predictability up to 50 GHz for an optimized model. Other work by Vaitus[12] investigated S-parameter sensitivity to variation of the model ECP values, but the modeling lacked a physical significance with which to establish a means of calculating an expected error versus an extrapolation of S-parameters over frequency. While the more recent extraction methods for determining model ECP values have improved the confidence of the small signal modeling procedure in predicting S-parameters, the uncertainties in these techniques are critical to the extrapolation of device performance. One of these uncertainties addressed here is the "Cold FET" extraction of extrinsic resistance values.

Continuing with a review of FET small signal model extraction, it is notable that uncertain or inadequate models mis-calculate certain figures of merit. Several groups including work by Trew and Steer[13] and H. Vickes[14] have investigated not only the modeling of S-parameters, but have included k-factor, unity current gain frequency f_t, and maximum frequency of oscillation F_{max}. These parameters are used as figures of merit. Successful prediction of these quantities is not guaranteed by an S-parameter fit, particularly when measured data is taken below the expected value of a figure such as F_{max}. More detailed studies in this respect include work by Vickes[6], Hughes and Tasker[15], P. Roblin[16], and M. Feng[17] which, provide measurement and modeling techniques that highlight the effects of the parasitic ECP values on model performance and frequency

extrapolations. In particular a general partitioning of the gain between the intrinsic and extrinsic circuit shows the ability to predict several possible "regions" for device performance at extrapolated frequencies due to the interrelation of the intrinsic and extrinsic elements[6]. This modeling approach is supported both by actual device measurements[5] and by physics based numerical simulation[9].

In Fulfilling our aim to further the development of the noise modeling procedure[1], an ECP extraction procedure is used along with on-wafer S-parameter measurements. As with all the present extraction methods, an improvement over "brute force optimization" is required to determine accurate parasitic element values. A review of published extraction techniques reveals that no single technique appears suited for all devices. This is particularly true for the case of physical differences, such as those between HEMTs and MESFETs, are considered. What is sought is a method which allows an extraction of parameter values that remain valid under frequency extrapolation of device small signal and noise performance. Also important, is an understanding of the uncertainty of the extracted values and also an understanding of the sensitivity of model predictions to the element variations.

FET Equivalent Circuit Parameter Extraction

Improvement of the noise modeling method is expected through implementation of direct extraction of equivalent circuit parameter values (ECPs). Extraction of FET equivalent circuit parameter values from S-parameter data, using forward gate bias along with normal operating bias conditions on the FET, are referred to here as "Cold FET" extractions. As demonstrated by Dambrine et. al.[18] and others. Improvements[19] to Cold FET techniques have shown frequency independent parameter extraction to 26.5 GHz. Limitations for application of these methods to short gate length devices and HEMTs are attributed to the sensitivity (or lack thereof) in the S-parameter measurements[20].

The USF[21] extraction technique is first demonstrated here on a 0.5 x 600μm GaAs MESFET provided by M/A-COM. While this device is designed with a power

process, it was provided with process parameter data that allows an investigation of the extraction technique. The technique employed is similar to that described by Dambrine and proceeds as follows:

1. Perform Cold FET measurements. The Cold FET bias levels used are between 11.0 mA and 18.0 mA of gate current. Assume that the channel resistance, R_{ch}, is negligible (this will be discussed in detail later). Then, extrapolate a value of R_s+R_g from a plot of Re[Z11] versus $1/I_g$.
2. Extract parasitic resistances from the highest forward biased data set.
3. Then for this example, use the R_s+R_g value that was extrapolated in step 1. to extract an approximate value of the channel resistance, R_{ch}. Where the values of R_s and R_g are supplied in process data sheets.

While process variation of the internal MESFET may be expected, the extrinsic inductance and resistance values are assumed accurate. A second small signal model is also extracted, for comparisons, from measured S-parameter data, but the supplied parasitic resistance values are used in the analysis. Proceeding with step 1, from the extrapolation to the origin of the Re[Z11] versus $1/I_g$ the value $R_g+R_s+R_{ch}/3$ is obtained. A value of R_g+R_s= 3.2 Ohms was determined, for R_{ch} = 0.0 as an initial approximation. The extraction of the parasitic values R_s, R_g, R_d, L_s, L_d and L_g proceeded with the values given in Table 1. Using the M/A-COM supplied extrinsic resistances a value of the channel resistance is calculated from $R_s + R_g + R_{ch}/3 = 3.2$, such that R_{ch} = 3.81 Ohms. This value is large enough to affect a mis-calculation of the drain parasitic resistance, R_d, which is dependent upon the full value of R_{ch} not $R_{ch}/3$, see Dambrine[18]. The resulting model parameter values are compared in Table 1 versus the typical process parameter values supplied by M/A-COM and the set of parameters derived from the M/A-COM extrinsic resistance values and the USF measured S-parameter data. Comparing the two extracted parameter data sets the values of g_m, R_{ds}. and τ are most affected by the modeling of the extrinsic resistances. The extracted inductance values agree well with the supplier process data. It is apparent that the un-modeled channel resistance in the USF extrinsic modeling

technique is significant and this resulted in a larger S-parameter modeling error.

Table 1 Small Signal Model ECP Values: Compares supplier average process values versus USF extraction using supplier extrinsic resistance values and also baseline USF extraction of both parasitic and intrinsic element values.

Parameter	Normal Bias Conditions	M/A-COM Process Data	USF/ M/A-COM	USF Extracted	Unit
C_{gs}	4V, 50% I_{dss}	0.606	0.649	0.716	pF
C_{gd}		0.048	0.052	0.050	pF
C_{ds}		0.150	0.162	0.182	pF
R_i		0.732	1.532	1.203	Ohm
g_m		78.0	66.878	73.5	mS
R_{ds}		135	175	148	Ohms
τ_d		4.0	3.36	2.88	pSec
R_s		0.75	0.75	1.936	Ohms
R_d		1.104	1.104	2.88	Ohms
R_g		1.18	1.18	1.264	Ohms
L_s		0.003	0.0065	0.006	nH
L_d		0.05	0.060	0.060	nH
L_g		0.04	0.046	0.046	nH
Avg. S-parameter Modeling Error		0.11142	0.00410	0.01066	

The resulting FET model is compared in Figure 5a. versus the measured data. The difference in modeled S-parameters is increasing with frequency. K-factor, $|h_{21}|$, and unilateral gain, U, compare well as shown in Figures 5b. and 5c.

Results for a low noise 0.5μm MESFET are given in Table 2. Here the channel resistance value was initially set to zero, but due to a lack of parameter process data was determined through a numerical search. A value of R_{ch} = 1.3 Ohms was used for the extraction and the average S-parameter modeling error reflects the improvement noted in the M/A-COM device model. Further optimization given in Table 2. does show a significant (2 Ohm) change in the extracted value of R_d, but it is suspected that a frequency dependence that is observed in several of the intrinsic parameters (particularly R_{ds}) may affect the

optimization. This dispersion is not currently accounted for in the extraction

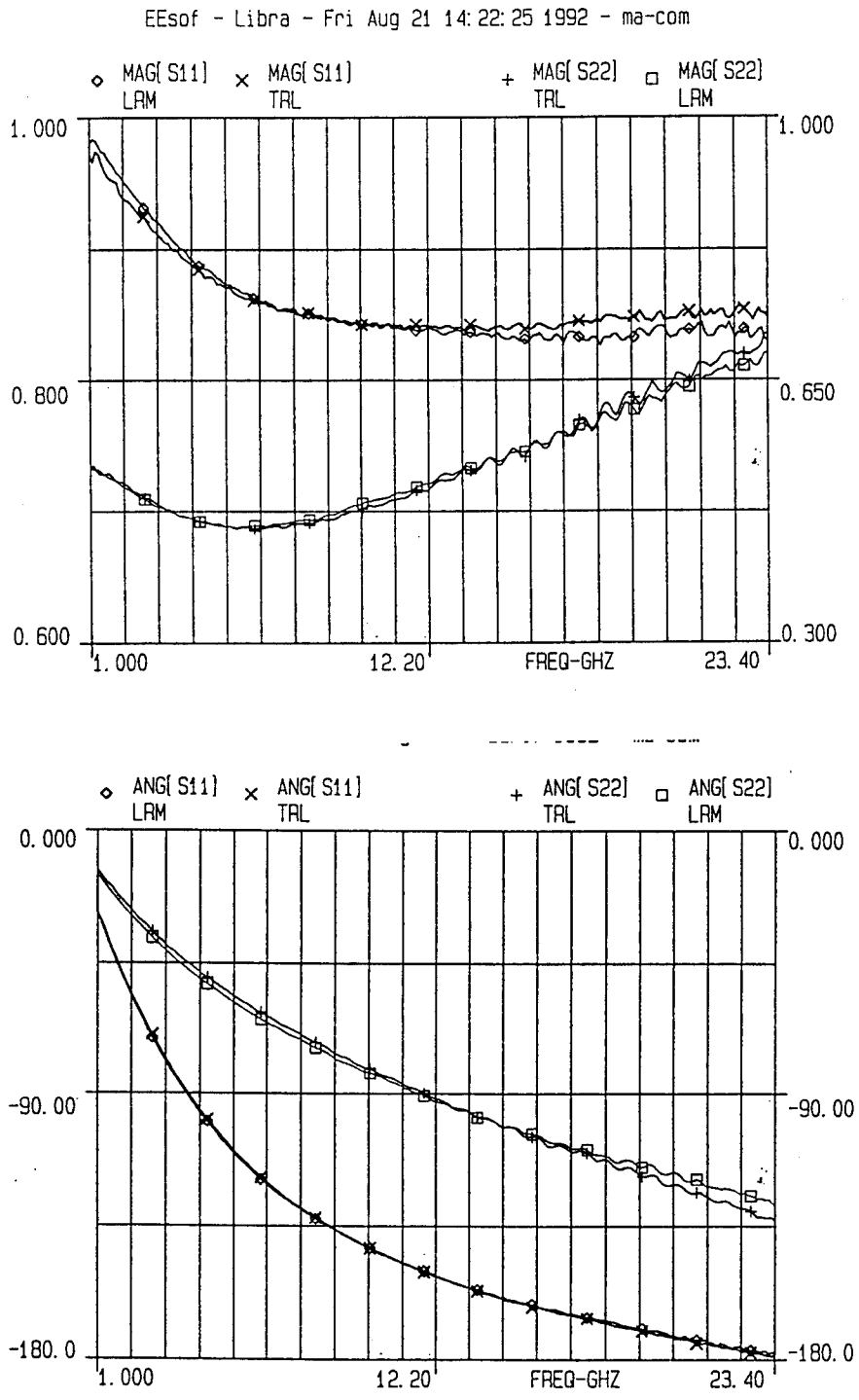


Figure 5a Extracted Small Signal Model Performance (MESFET) versus measured data.

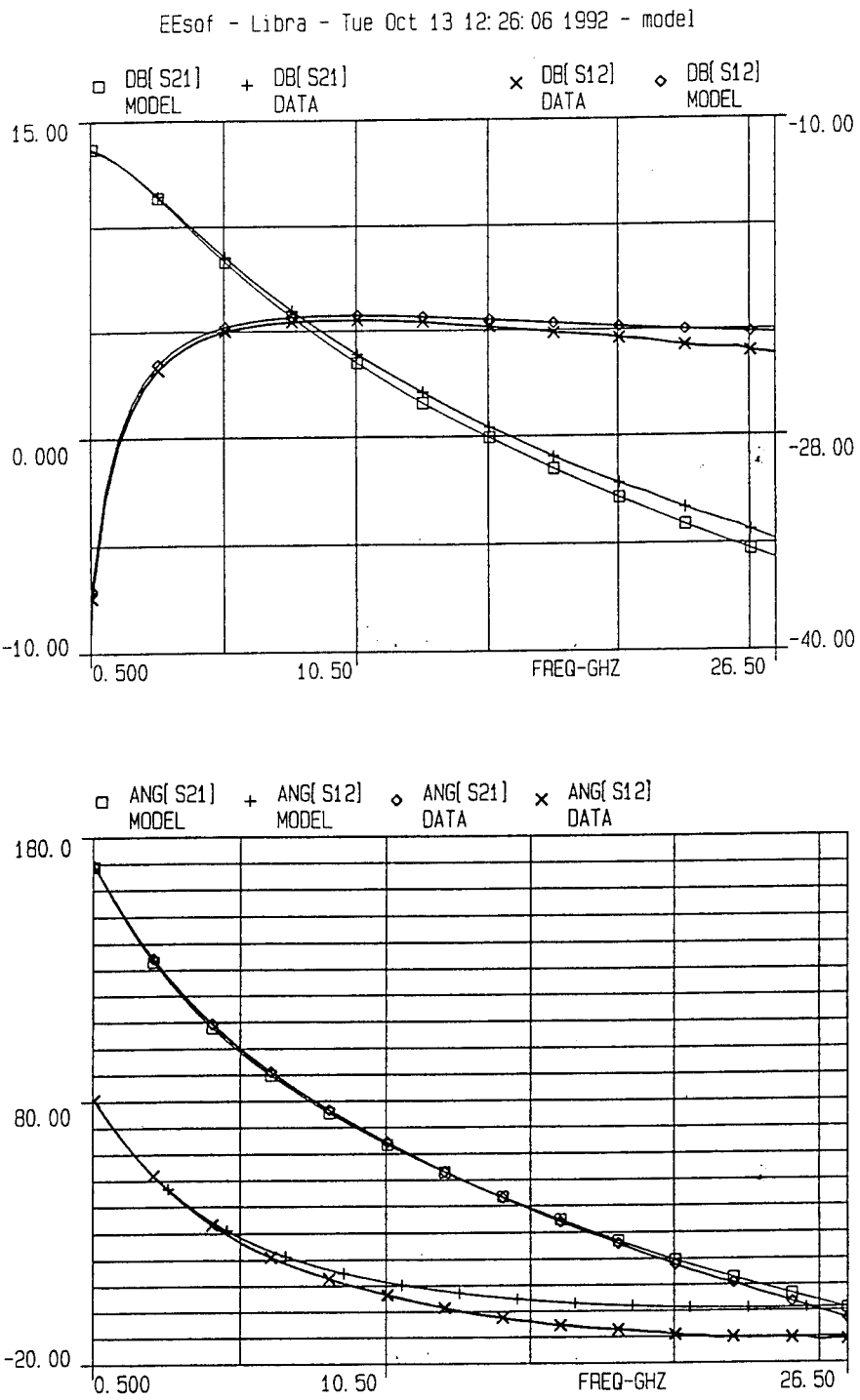


Figure 5b Extracted Small Signal Model Performance (MESFET) versus measured data.

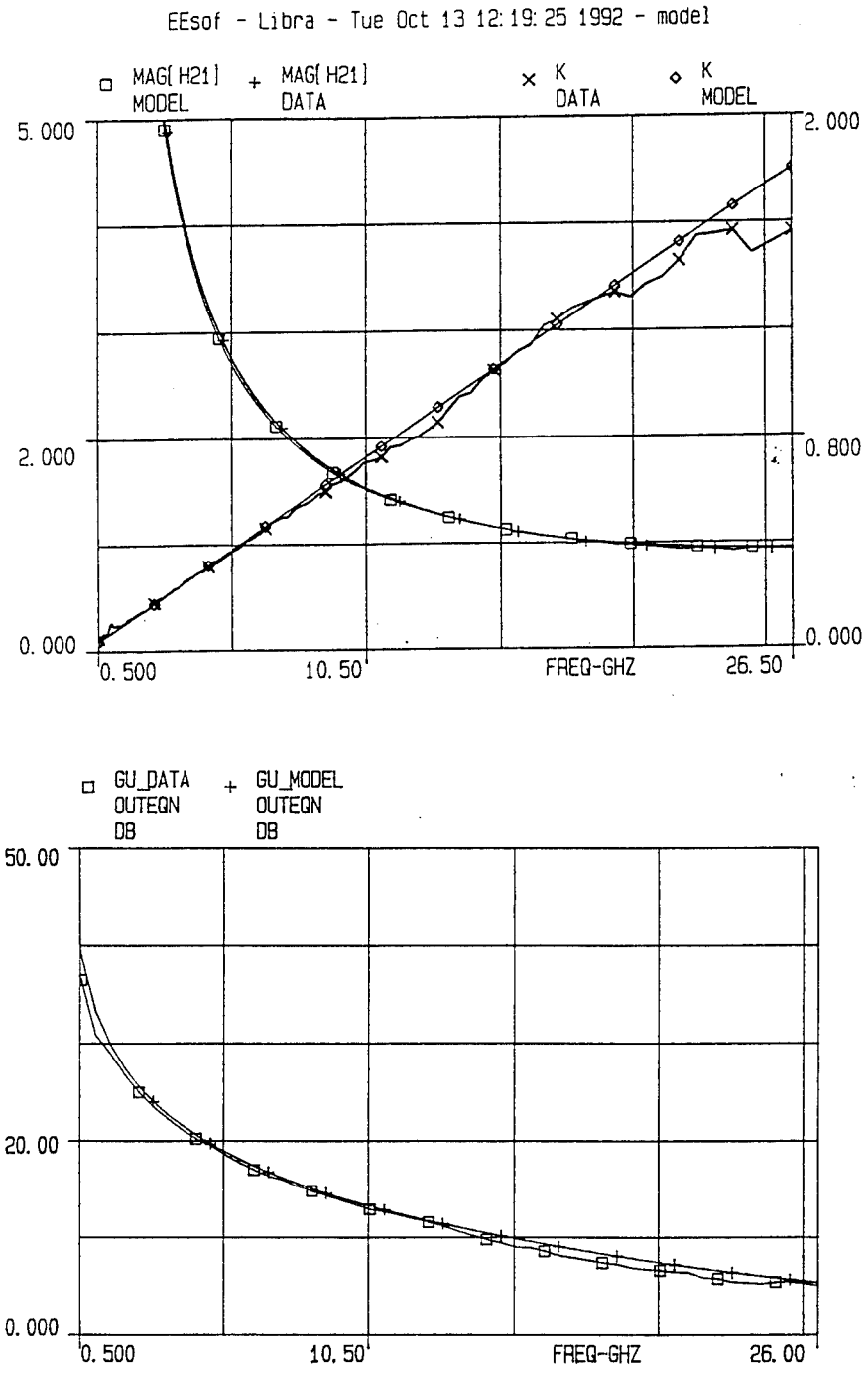


Figure 5c Extracted Small Signal Model Performance (MESFET), k-factor, Mag. h21, and unilateral gain versus measured data.

procedure. An improved fit of S_{22} is achieved as a result of optimization, using the extracted model as a starting point.

Table 2 Small Signal Model ECP Values (MESFET LN300): Compares optimized ECP values versus USF extraction.

Parameter	Normal Bias Conditions	USF Extracted	USF Optimized Extracted	Unit
C_{gs}	4V, 20.0mA	265.0	251.8	pF
C_{gd}		37.0	39.74	pF
C_{ds}		114.0	105.88	pF
R_i		2.9	3.335	Ohm
g_m		35.0	32.63	mS
R_{ds}		289.8	328.33	Ohms
τ_d		2.4	3.0	pSec
R_s		2.75	2.065	Ohms
R_d		2.80	0.813	Ohms
R_g		1.18	1.6	Ohms
L_s		2.89	2.065	nH
L_d		68.8	60.04	nH
L_g		63.8	61.77	nH
Avg. S-parameter Modeling Error		0.0067	0.00054	

The resulting FET model compares well with the measured data.

Results for a 0.25μm PHEMT are given in Table 3. Cold FET extrinsic values were extracted at 15.0mA of gate current or 3.0 X $10^8 A/m^2$. For a zero valued channel resistance the extracted values of the extrinsic resistances are given in Table 3. under the USF Extracted column heading. The phase of S_{22} and the magnitude of S_{11} differed from the measured data by 4%. Additionally, the unilateral gain and K-factor deviated from the measured data.

In this case the initial error function was comparable to that achieved for the M/A-COM device when the process determined extrinsic values were used for modeling. Therefore, the un-modeled channel resistance and/or un-modeled effects at the PHEMT source and drain contacts were considered to be less dominant. That is the values of the extrinsic resistances are large compared to the value of the channel resistance and its fractions. The S-parameters are predicted accurately (Figure 6a and 6b), but the unilateral gain and K-factor differ from the measured

data. A brief gradient optimization of the model (less than 15 iterations) quickly reduced these differences in the S-parameters k-factor and unilateral gain.

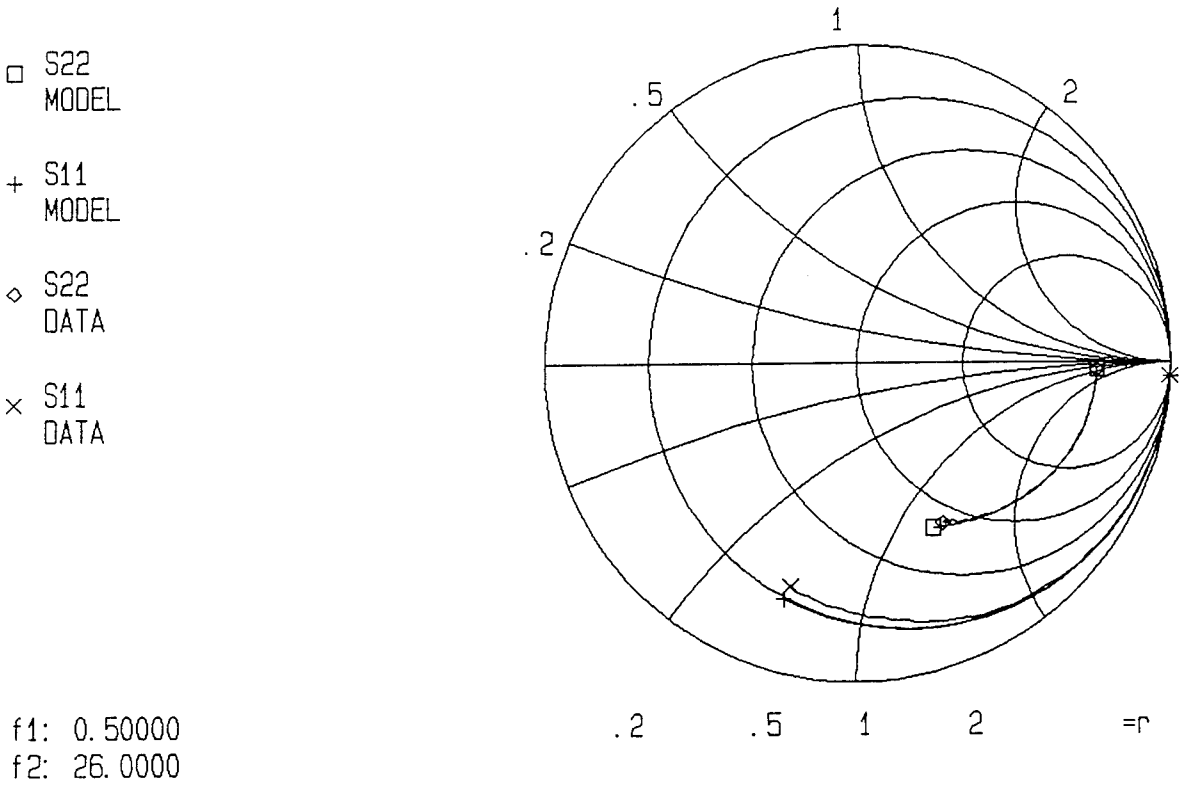
Table 3 Small Signal Model ECP Values (PHEMT): Compares optimized ECP values versus USF extraction.

Parameter	Normal Bias Conditions	Brute Force Optimized	USF Extracted/ Optimized	Unit
C_{gs}	3V, 13.0mA	89.7	105.51	pF
C_{gd}		21.12	16.26	pF
C_{ds}		29.82	48.51	pF
R_i		1.26	1.44	Ohm
g_m		40.23	45.84	mS
R_{ds}		373.26	305.24	Ohms
τ_d		0.82	0.348	pSec
R_s		1.32	4.47	Ohms
R_d		2.64	6.39	Ohms
R_g		4.29	4.22	Ohms
L_s		6.10	3.32	nH
L_d		66.37	50.34	nH
L_g		49.19	51.57	nH
Avg. S-parameter Modeling Error		0.0026	0.00029	

Discussion

A small signal modeling procedure was investigated to update the "brute force" optimization procedure followed in the previous noise modeling work[1]. The determination of equivalent circuit parameter values by a method similar to Dambrine et. al.[18,21] was experimentally tested. The final resulting models are considered to be more accurate for noise modeling purposes, because the extrinsic element values are determined directly from measured data. These cold FET extracted models have a lower average S-parameter modeling error than the brute force optimized models. Consistency in the ability of the Cold FET extracted models to predict and extrapolate small signal gain parameters, such as the unilateral gain is shown.

The uncertainty that exists in the extraction of parasitic resistance values was investigated, for Cold FET measurements. As Dambrine et. al.[12] note,



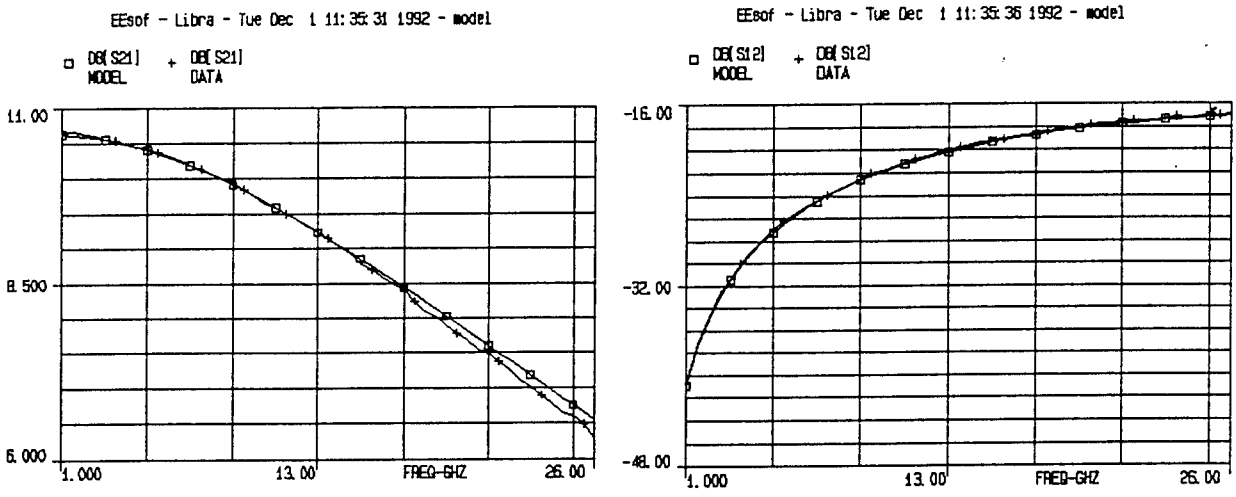


Figure 6a Extracted Small Signal Model Performance (PHEMT).

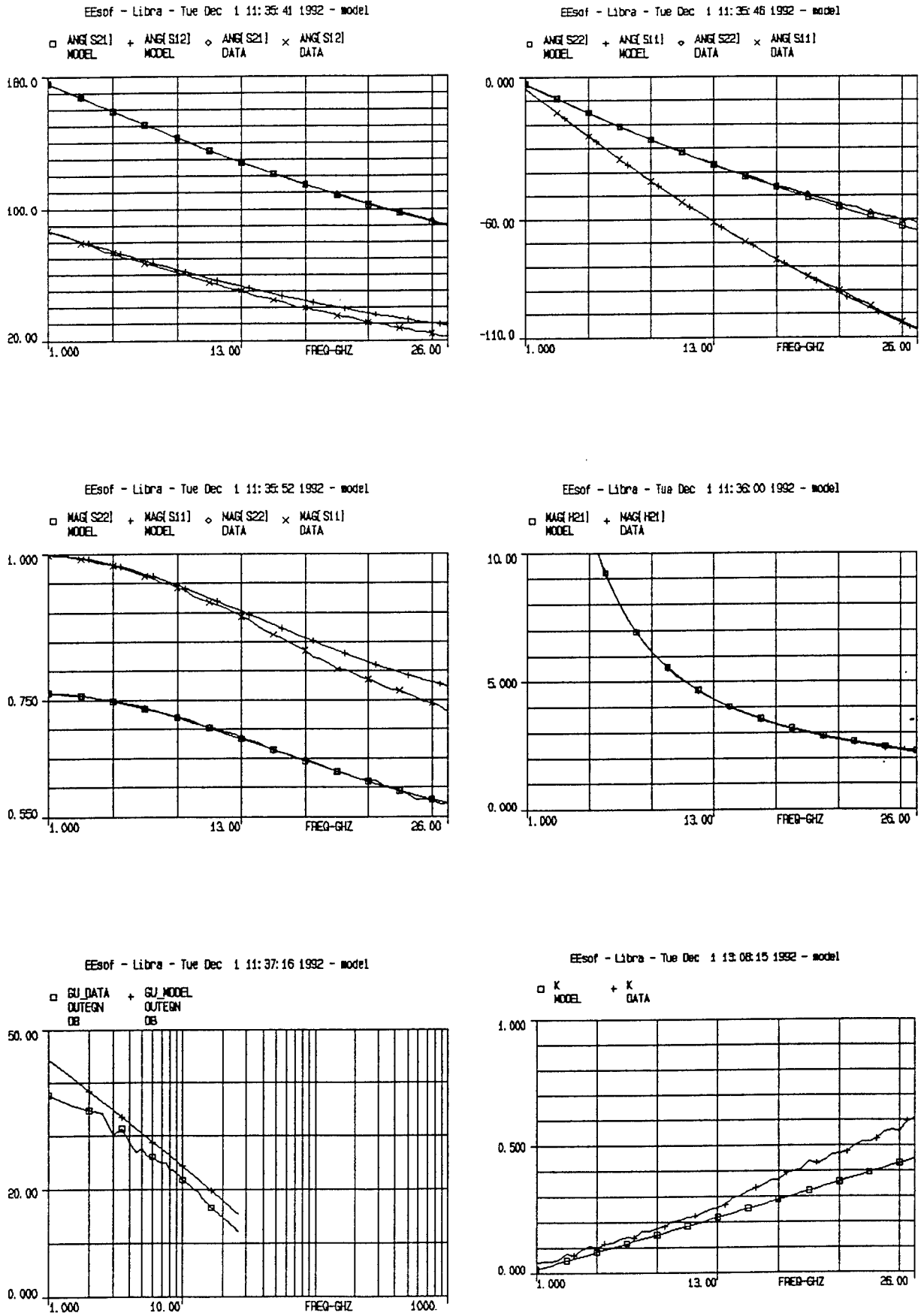


Figure 6b Extracted Small Signal Model Performance (PHEMT).

an additional relation (technological process data) or measurement (DC data) is needed to account for the channel resistance, R_{ch}. The determination of the parasitic resistance is uncertain, because the Cold FET technique establishes a system of three independent equations[18], but along with the parasitic resistors the channel resistance is a fourth unknown. Without additional information about R_{ch} a numerical search for appropriate values was investigated.

Additional measurements in this regard are suggested, although for the case of the PHEMT the standard DC measurements may be inadequate. The concern here is to avoid parallel conduction in the doped layer (AlGaAs for GaAs HEMTs)[H. Morkoc]. Of the published literature the work of Fu et. al. directly addresses the measurement and modeling of short (length < 0.1μm) gate length HEMTs.

Additional uncertainty also exists in choosing the extracted values from a set of data that exhibits a dispersion of the parameter values versus frequency. While the extrinsic resistance data (Re[Z11]) may vary by less than 0.5 Ohms from 1.0 to 26.0 Ghz, the values of R_{ds} may vary by 30%. Also, a better "choice" of the extracted value versus frequency is needed, as shown with the PHEMT, where slight optimization changes were needed to improve the overall model performance.

Noise Parameter Sensitivity Analysis

In the previous work[1] the noise modeling algorithm was highly dependent upon the input data set. This data set consists of the source reflection coefficient and the associated FET noise figure measured at several frequencies. This work initiates an investigation of the sensitivity of the data processing required to determine the coefficients P, R, and C. Initially, a perturbation of the input data set parameters was investigated.

The results of a perturbed set of nominally 50 Ohm source impedance noise figure measurements is given in Figure 8. In this case the magnitude and angle of the reflection coefficient and the noise figure were separately varied by 10%. The variation in the noise coefficients was then plotted versus variations in the input parameters.

Our investigation of noise parameter measurement sensitivity continues with the least-squares fitting procedure described by Lane[3]. It is known that the matrix of equations, used to solve for the coefficients of Lane's linearized form of the noise figure equation, could be ill-conditioned[22]. An IRE Subcommittee had also published standards on measuring noise in linear two-ports which defined a procedure for tuner based measurements that helped avoid such ill-conditioning[7].

Before proceeding, it is often noted in mathematical treatments of such problems that certain data sets may result in matrix singularity. Ill-conditioning is a more qualitative reference to the case where, although the matrix is not singular, the data may be related to the singular matrix such that it is ill-conditioned to complete the analysis without errors. Therefore, the degree of ill-conditioning may result in an inaccurate measurement of the noise parameters and in the worst case no solution at all (a singular coefficient matrix).

In determining the noise parameters, Lane suggested that improvement in the data processing technique was needed to lower error sensitivity. Sannino then reported on an investigation of loci of Y_s on the Smith Chart which produced singularities and experimentally exhibited an updated procedure for selecting source tuner states[22]. It was concluded that a measurement of data from tuner

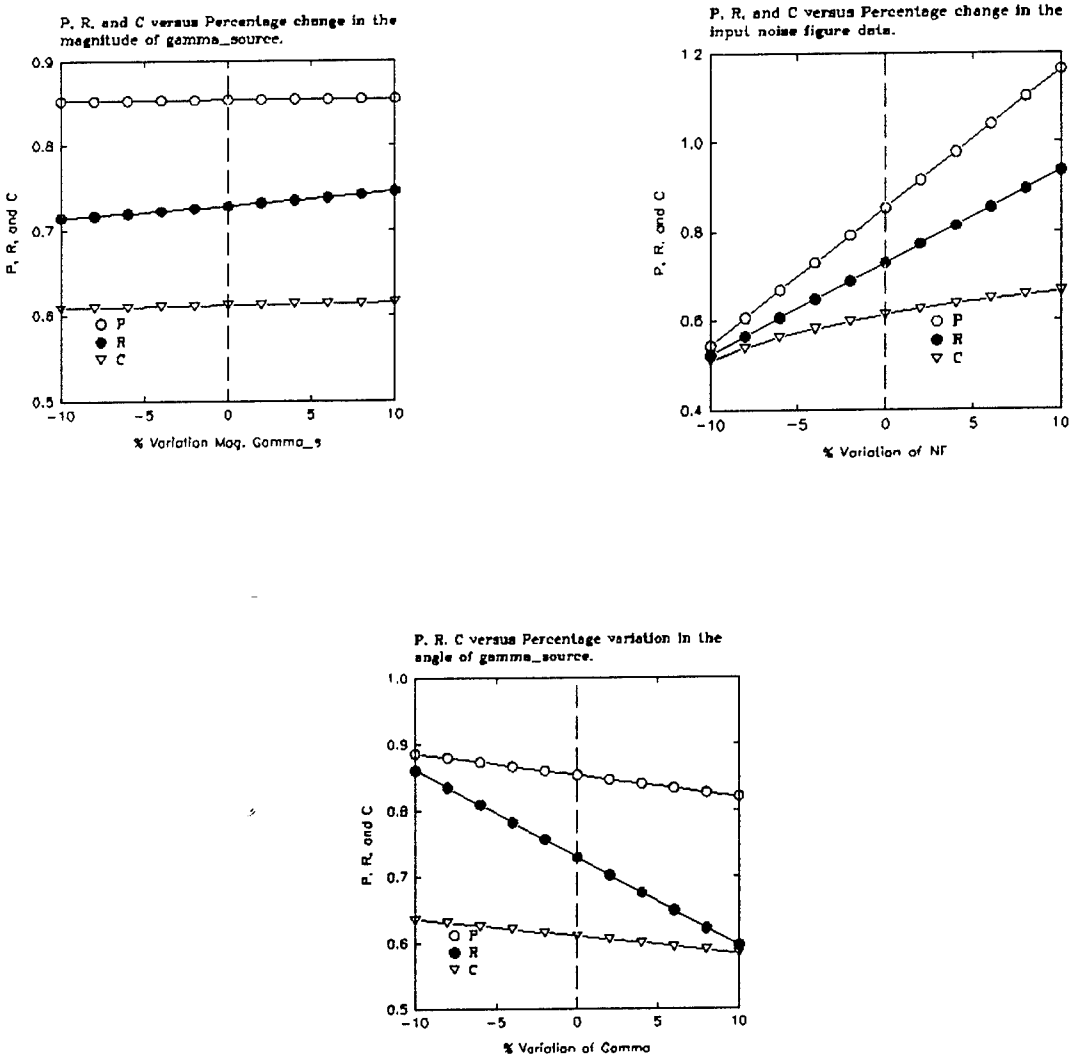


Figure 7 Variation of coefficients P, R, and C versus 10% variation in the the 50 Ohm input tuner data set (noise figure, source state magnitude and angle).

states that belonged to at least two families of the singular loci was adequate for a solution that characterized the surface described by the noise figure equation. The singular loci are circles on the G_s, B_s plane of the Smith Chart. One of the singular locus conditions is when G_s, B_s and $|Y_s|^2$ equal constant values. Also at this time, Mitama and Katoh[23] improved the least-squares fitting algorithm of Lane by accounting for the least-squared error due to the source state components G_s and B_s.

More recently, O'Callaghan and Mondal used a vectored error correction approach was applied to the measurement of noise parameters[24]. The resulting method uses a linearized noise figure equation similar to that of Mitama and Katoh, that is the error sensitivity due to noise figure as well as G_s and B_s is included. Because the vectored method is built upon a projection theorem argument, a quantitative as well as qualitative analysis of the error sensitivity is shown. A coefficient matrix for the linearized set of equations is developed. However, O'Callaghan and Mondal proceed to solve for the coefficients of the linearized equations in closed form. The least squares fit of these coefficients is avoided by a decomposition of the matrix (Cholesky's Method). The singularity of the matrix can then be addressed by standard vector methods. Additionally, the type of successful vectored error correction that is common to S-parameter network analyzer calibration is applied to the projection of the vectors which comprise the linearized noise figure equation. The result is that a quantitative analysis of the measurement error sensitivity and the degree of ill-conditioning of the equation matrix is addressed.

Upon reviewing the details of these analyses several questions concerning the single tuner state algorithm are given as follows:

1. Will a least-squares error fit that also includes Y_s errors prevent model coefficients P, R, and C from being non-physical (negative values or $C > 1$)?
2. What are the linearized equation matrix conditions on 50 ohm source state data and on a data set from a constant arbitrary state, Γ_s?
3. What are the effects on the conditioning of the matrix when using a fewer number of frequencies (data points) in the present modeling procedure?

A complete sensitivity analysis of the noise modeling algorithm presented in Section III was not available at this time. Such an experimental analysis requires a device that is characterized by the small-signal modeling technique

presented in Section 4. However, noise parameters for a MESFET and a PHEMT were measured and a preliminary analysis gives the following results:

1. A study on the ill-conditioning of the coefficient matrix associated with the linearized noise figure equation was done. For the 50 Ohm source state the matrix is not singular, and values for the coefficients P, R, and C are developed. The resulting values are dependent upon the size of the data set.
2. In the case that the data set is chosen with the magnitude of the input reflection coefficient, $|\Gamma_s|$, constant the matrix is singular. For the case that the source state varies about the Smith Chart a solution is calculated, but the resulting noise parameters vary with different data sets.

VI. Microwave Measurements and Model Comparisons

Microwave noise parameter measurements performed from 2.0 to 18.0 GHz on three devices each of a 0.5 μm GaAs MESFET and a 0.25 μm PHEMT. The ATN NP5A noise parameter test-set and measurement software was configured at U.S.F. as part of this work. Results for the MESFET are given below.

VD=4.000 V ID= 20.000mA
Measurement Date: 5 Dec 1992 Time: 05:25:27
MESFET DEVICE

FREQUENCY (Ghz)	FMIN (dB)	GAMMA OPT Mag	Ang	Rn (NORMILIZED)		Ga (Associated Gain) (dB)
2.000	.72	.92	16.3	1.02	!	16.69
2.960	.94	.88	23.9	1.09	!	15.18
3.920	1.15	.85	31.3	1.15	!	13.86
5.040	1.39	.81	39.6	1.19	!	12.54
6.000	1.60	.78	46.6	1.21	!	11.57
6.960	1.81	.76	53.4	1.22	!	10.74
7.920	2.02	.74	60.0	1.21	!	10.03
9.040	2.27	.71	67.6	1.19	!	9.34
10.000	2.48	.70	73.9	1.16	!	8.84
10.960	2.69	.68	80.0	1.12	!	8.41
11.920	2.90	.67	86.0	1.08	!	8.04
13.040	3.14	.67	92.8	1.03	!	7.66
14.000	3.35	.66	98.5	.98	!	7.35
14.960	3.56	.66	104.0	.92	!	7.05
16.080	3.81	.67	110.3	.86	!	6.69
17.040	4.02	.67	115.6	.79	!	6.36
18.000	4.23	.68	120.8	.73	!	5.98

The plots of Figure 8 represent the measured and modeled noise parameters of a MESFET[1] versus frequency. Each model is derived from 50 Ohm tuner state data. P, R, and C are determined using the same algorithm for each model. A

direct solution of the noise parameters is compared to the "synthesized measurement" of the previous algorithm. A subroutine uses correlation matrices to provide the solution. Inputs to this subroutine are frequency, P, R, C and the FET equivalent circuit parameter values.

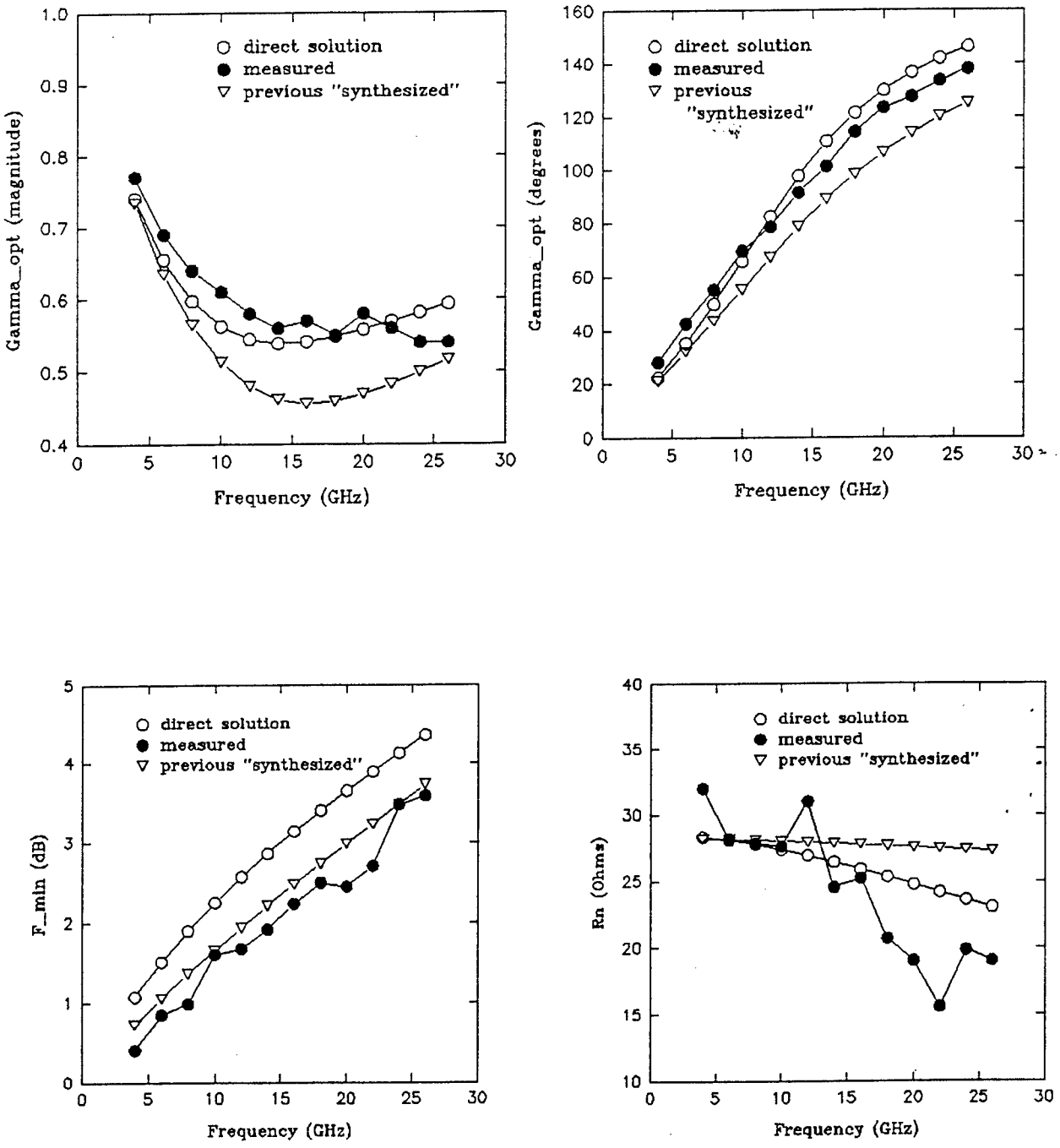


Figure 8 These plots represent the measured and modeled noise parameters of a MESFET versus frequency. A direct solution of the noise parameters is compared to the "synthesized measurement".

VIII. Millimeter-wave Noise Modeling Conclusions and Extensions

In furthering the development of the noise modeling method discussed in this report, topics of continuing interest include:

1. Verification of the method at millimeter-wave frequencies.
2. A continued analysis of the least squares determination of coefficients P, R, and C and their sensitivities.
3. Small signal modeling improvements in the determination of the parasitic resistances.
4. An experimental investigation of the temperature dependencies of the intrinsic FET model noise sources.

Item (2) may be approached in two ways. One investigation would study the conditioning of the coefficient matrix associated with the least squares solution of P, R, and C. As a result, a consistently accurate procedure for noise parameter modeling with the single source state method would be defined. The other area of approach concerns the approximation of the noise figure equation in terms of P, R, and C. Here it has been linearized for the least squares fit. This is of interest for measurement sensitivities, because the actual non-linear system of equations will have different singularities from the linearized equation.

As part of this investigation several DC and lower frequency measurement techniques were studied for the determination of the FET parasitic resistances. The experimental implementation of such a measurement is a natural extension of this work, that will experimentally provide the additional information needed to complete a small signal model extraction.

The temperature dependence of the intrinsic noise sources is of interest to U. S. F. and the Air Force. This work helps to address the goal of determining the noise source temperature dependencies. In particular, the general model derived here allows the implementation and study of both new and existing intrinsic model noise source representations.

VIII. Summary and Conclusions

This noise modeling investigation addresses several items necessary to the

development of a single source state noise modeling method. Emphasis was placed on a general formulation of the model. This resulted in the derivation of an efficient noise modeling algorithm from noise correlation matrix methods. The algorithm is useful for CAD implementation, because the model topology and intrinsic noise sources are described by an interconnection of cascaded matrices. The small signal model extraction procedure adopted here updates the "brute force" optimization previously used to determine model ECP values. Small signal models were extracted for both a MESFET and a PHEMT with good agreement to measured data. In addition, a conventional on-wafer noise parameter measurement system and related software was configured at U. S. F. Noise parameter measurements were performed with this conventional system over the 2 to 18 GHz range. For the present single source state model, as with conventional measurement techniques, a sensitivity to measurement errors and coefficient matrix conditioning was studied. Initiation and continuation of this sensitivity analysis is needed to improve the least squares fitting procedure. As a result, certain data sets were identified with respect to the ill-conditioning of the solution, and further steps towards realizing a consistently accurate modeling procedure were presented.

REFERENCES

1. L. Dunleavy and S. Lardizabal, "Millimeter-wave Noise Modeling Investigation", 1991 USAF-AFOSR Summer Research Program Final Report, July 19, 1991.

2. W. Patience, "A Simplified Method of Determining Noise Parameters of High Frequency MESFET's", 1989 USAF-AFOSR Summer Faculty Research Program Final Report, Sept. 6, 1989.

3. R. Lane, "The Determination of Device Noise Parameters", Proc. IEEE, vol. 57, pp.1461-1462, Aug. 1969.

4. H. Hillebrand and P. Russer, "An Efficient Method for Computer Aided Noise Analysis of Linear Amplifier Networks", IEEE Trans. on Circuits and Systems, CAS-23, pp.235-238, 1976.

5. R. Pucel, W. Strubble, R. Hallgren, and U. Rhode, "A General Noise De-embedding Procedure for Packaged Two-Port Linear Active Devices", IEEE Trans. on Microwave Theory and Tech., Nov. 1992.

6. V. Adamian and A. Uhlir, "A Novel Procedure for Receiver Noise Characterization", IEEE Trans. on Instrumentation and Meas., pp.181-182, June 1973.

7. H. Haus ed., "IRE Standards on Methods of Measuring Noise in Linear Two-Ports", Proc. IRE, vol. 48, pp.60-68, Jan. 1960.

8. A. van der Ziel, "Thermal Noise in Field Effect Transistors", Proc. IRE, pp.1808-1812, Aug. 1962.

9. M. Pospieszalski, "Modeling of Noise Parameters of MESFET's and MODFET's and Their Frequency and Temperature Dependence", IEEE Trans. Microwave Theory Tech., vol.MTT-37, pp.1340-1350, Sept. 1989.

10. K. Yeom, T. Ha, and J. Ra, "Frequency Dependence of GaAs FET Equivalent Circuit Elements Extracted from the Measured Two-Port S-Parameters", *Proc. of the IEEE*, vol. 76, pp. 843-845, July 1988.

11. W. Curtice and R. Camisa, "Self-Consistent GaAs FET Models for Amplifier Design and Device Diagnostics", *IEEE Trans. on Microwave Theory and Tech.*, vol. MTT-32, pp.1573-1578, Dec. 1984.

12. R. Vaitus, "Uncertainty in the values of GaAs MESFET equivalent circuit elements extracted from measured two-port scattering parameters", IEEE Microwave Theory Tech., vol. MTT-32, pp.1573-1578, Dec. 1984.

13. M. B. Steer and R. J. Trew, "High Frequency Limits of Millimeter-Wave Transistors", *IEEE Electron Device Lett.*, vol. EDL-7, pp. 640-642, Nov. 1986.

14. H. O. Vickes, "Gain Partitioning: A New Approach for Analyzing the High-Frequency Performance of Compound Semiconductor FET's", *IEEE Microwave Theory and Tech.*, vol. 39, pp. 1383-1390, August 1991.

15. P. Tasker and Brian Hughes, "Importance of Source and Drain Resistance to the Maximum f_t of Millimeter-Wave MODFET's", *IEEE Electron Device Lett.*, vol. 10, pp. 291-293, July 1989.

16. P. Roblin and S. C. Kang, "Unilateral Power Gain Resonances and Roll-Off with Frequency for the Velocity-Saturated MOSFET/MODFET Wave Equation", *IEEE Trans. on Electron Dev.*, vol. 39, pp. 1490-1494, June 1992.

17. M. Feng, C. Lau, and C. Ito, "A Technique for Correction of Parasitic Capacitance on Microwave f_t Measurement of MESFET and HEMT Devices", *IEEE Trans. Microwave Theory Tech.*, vol. 39, pp. 1880-1882, November 1991.

18. G. Dambrine, A. Cappy et. al., "A New Method for Determining the FET Small Signal Equivalent Circuit Parameters", *IEEE Trans. on Microwave Theory Tech.*, pp.1151-11159, July 1988.

19. M. Berroth and R. Bosch, "Broad-band determination of the FET small-signal equivalent circuit," *IEEE Trans. Microwave Theory Tech.*, vol. 38, pp. 891-895, July 1990.

20. S. Fu and M. Das, "Determination of Equivalent Circuit Parameters for Short-Gate-Length HEMTs," *IEEE Trans. Electron Dev.*, pp. 888- , April 1990.

21. A. Fernandez, L. Dunleavy, P. Abernathy, and J. Rollman, "An Improved FET Model Extraction Technique with Applications to Temperature-Dependent FET Modeling", 4th Annual Florida Microelectronics Conference Digest, Tampa, Florida, May 1992.

22. M. Sannino, "On the Determination of Device Noise and Gain Parameters", Proc. IEEE, vol. 67, pp.1364-1366, Sept. 1979.

23. M. Mitama and H. Katoh, "An Improved Computational Method for Noise Parameter Measurement", IEEE Trans. Microwave Theory Tech., vol. 27, pp. 612-615, June 1979.

24. J. O'Callaghan and J. Mondal, "A Vector Approach for Noise Parameter Fitting and Selection of Source Admittances", IEEE Trans. Microwave Theory Tech., vol. 39, pp.1376-1381, Aug. 1991.

NOVEL DIGITAL SPEECH COMPRESSION TECHNIQUES

Joseph B. Evans, Assistant Professor
Bernard Yu, Graduate Research Assistant
Anthony Zaide, Graduate Research Assistant
Telecommunications and Information Sciences Laboratory
Department of Electrical & Computer Engineering

The University of Kansas
1013 Learned Hall
Lawrence, KS 66045-2228

Final Report for:
Research Initiation Program
Rome Laboratories

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

December 1992

NOVEL DIGITAL SPEECH COMPRESSION TECHNIQUES

Joseph B. Evans, Assistant Professor
Bernard Yu, Graduate Research Assistant
Anthony Zaide, Graduate Research Assistant
Telecommunications & Information Sciences Laboratory
Department of Electrical & Computer Engineering
University of Kansas

Abstract

This research involved increasing the capabilities of the Sinusoidal Transform Coder (STC), a low bit rate digital speech coder. During this project, the STC software was revised to allow easier modification during this project and in future research. New FFT windowing techniques were studied in order to reduce the processing delay of the STC. Multirate parameter transformation techniques were studied, and coder configuration parameters for various rates were selected. Finally, a voice mail application which demonstrates the utility of the transformation techniques for speech data compression and storage was developed. This application, which can be used by any host running TCP/IP and directly connected to the Internet, allows voice messages to be transmitted from a sender to a recipient host using a novel negotiation protocol.

NOVEL DIGITAL SPEECH COMPRESSION TECHNIQUES

Joseph B. Evans
Bernard Yu
Anthony Zaide

Introduction

This report describes research performed during the AFOSR Research Initiation Program related to digital speech coding using sinusoidal representations. This work involved modifications and extensions to the Sinusoidal Transform Coder (STC) developed by R. J. McAulay at MIT Lincoln Laboratory under the support of the Rome Laboratories Digital Speech Laboratory [6, 8, 9, 10]. The STC modifications and extensions provide significant increases in flexibility and utility, as well as usability of the code for research purposes.

This work concentrated on two related areas. The first of these areas is the multi-stage transformation of encoded speech from one rate to a lower rate. This transformation technique provides flexibility not available in other low and medium bit rate coders. During this project, the existing STC software, including the transformation code, was restructured to allow future modifications to be made in a straightforward manner. In addition, vector quantization of the spectrum was added in order to track ongoing work at MIT Lincoln Laboratory. FFT windowing techniques for a low delay coder were explored, as was selection of parameters for a medium rate STC coder. The second major area of effort for this project was the development of a voice mail application for use on the Internet. This application demonstrates the utility of the coding and transformation techniques for data storage management. This application runs under the UNIX ® operating system on most common workstations. It allows voice-based messages to be exchanged over data networks and stored efficiently on the recipient computers.

®UNIX is a registered trademark of AT&T.

Background[1]

Speech communications systems can be evaluated on the basis of many interrelated criteria, including intelligibility, perceived quality, and robustness. Several of these measures are clearly more significant than others, but clearly speech coding algorithms do not exist in isolation from the entire voice communications system. It is therefore important that flexibility of application and robustness to severe data transmission environments be considered when evaluating speech coding methods.

Speech Coding via Sinusoidal Representations

A variety of voice communications systems require data rate compression of the speech signal. Approaches to this problem can be separated into two categories, algorithms which make few assumptions about the nature of the input data, such as adaptive Huffman or Lempel-Ziv-Welsh (LZW) coding, and those which are optimized for the expected data, in this case speech, such as ADPCM and LPC-based coders. The first type of compression technique reduces the memory storage requirements with no information loss, while the second type involves some reduction in information as compared to the original data. Both types of coders reduce the resource usage by a fixed percentage for specified data. The coder used in this project allows for resource usage to be balanced against speech quality in a dynamic manner.

The basic coding technology used in this project is the Sinusoidal Transform Coder (STC) algorithm [6, 7, 8, 9, 10]. The STC model has been shown to be effective at various discrete rates from 8 kb/s to 2.4 kb/s [9]. Further, this coding technology is extremely robust to manipulation in the parameter space, as shown by the previous work in STC parameter space transformations [2]. The parameter space transformation method allows the most optimum coding for this coder type to be used for a given compression ratio. The

[1]Portions of this section are drawn from [4].

technique which forms the basis of this work allows compression in multiple stages of any selected size, with no interaction with the source.

Multirate Parameter Transformations

Because of the robust nature of the representation used by the STC algorithm, not only can the coder operate at various rates [9], but transformations in the parameter space can be used to change the bit rate of the coded data.

The parameter transformation technique allows the coding for all spectrum, excitation, and frame length parameters to be selected at each rate to give the optimum coding for that rate. Among the parameters which may be varied at each rate are the number of cepstral coefficients, the frequency range for coding, the method of representation (frame-fill or the actual value) of the mid-frame excitation parameters, and the number of bits used for each parameter. The method for changing the spectral representation (that is, the number of cepstral coefficients) is shown in Figure 1. The cepstral coefficients derived from the coded parameters are used to derive a spectral envelope, which is then used to generate a new set of cepstral coefficients, which can then be coded. The parameter temporal interpolation technique used for frame length changes at each stage is depicted in Figure 2. The various parameters at each new frame boundary (that is, at T'_{k-1}, T'_k, T'_{k+1}), are linearly interpolated according to the distance of the new frame boundaries from the closest old edges ($T_{k-1}, T_k, T_{k+1}, T_{k+2}$). Provision is also made for interpolating mid-frame and frame-fill parameters in a similar manner.

Intelligibility tests have verified that the transformation algorithms were producing output speech of good quality. The 2.4 kb/s test data was generated by performing the initial STC analysis at a 16 kb/s rate with a 10 ms frame size. The rate was then reduced via the transformation to 8 kb/s with a 20 ms frame length, then to 4.8 kb/s with a 20 ms frame, and finally to 2.4 kb/s with a 30 ms frame length. The results are summarized in Table 1. Most listeners find the the perceived speech quality to be good, only slightly less

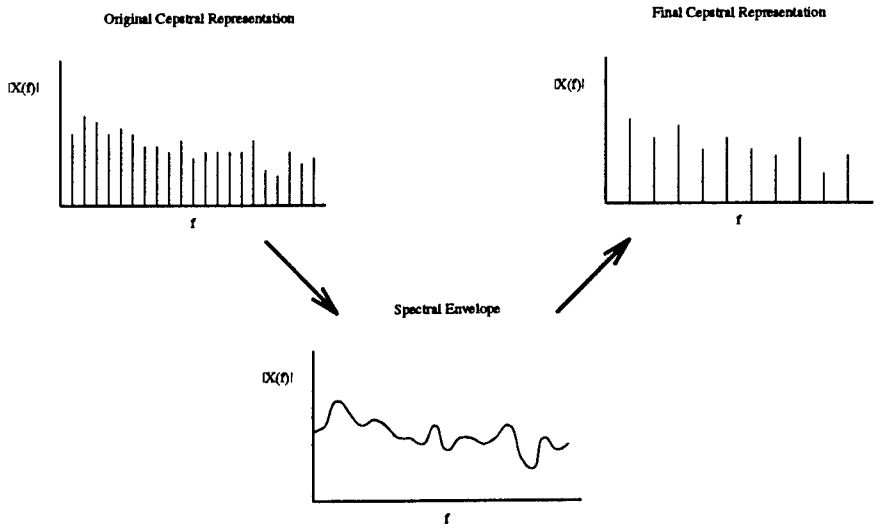


Figure 1: Technique for Changing the Spectral Representation

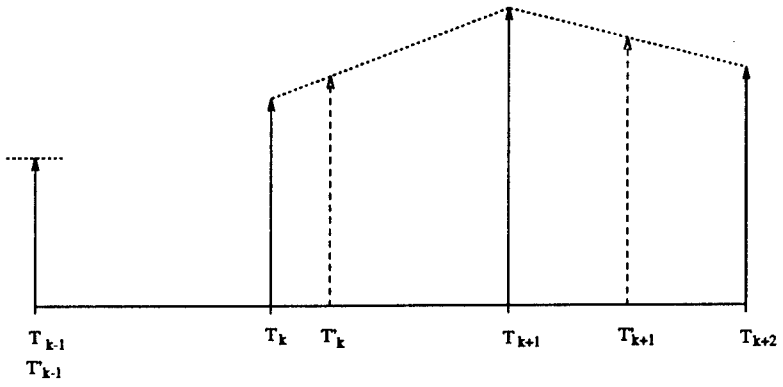


Figure 2: Interpolation of Parameters for New Frame Lengths

Table 1: DRT Test Results

Speaker(s)	Results
JE	91.93
CH	95.44
RH	91.67
Total for 3 Males	93.01

pleasing than the untransformed 2.4 kb/s STC.

Applications

The method presented above is ideal for many applications. A typical application for this multi-level compression technique, a voice mail system, was created to test the overall concept [4]. This project extends that rough prototype.

Because of the rapid development of audio hardware for personal computer systems and workstations, voice mail transmission through the national network known as the Internet is becoming common. Almost every modern workstation and personal computer system supports some type of audio file format. For such systems, corresponding audio hardware is provided to record and play audio data.

The operation of a typical voice mail system is illustrated in Figure 3. The mail is generated by a sender and transmitted to the recipient via a computer network. The storage system processes the arriving packets to insure that message consistency is maintained. The speech data is then passed to the appropriate decoder and played on the recipient host's audio interface.

The original application documented in [4] used the Sun Microsystems Remote Procedure Call (RPC) protocol, which in turn used the standard TCP/IP protocol stack for data transmission.

There are two types of file formats: "raw" formats, where the device parameters and encoding are fixed, and self-describing formats (e.g., .wav format used by MicroSoft

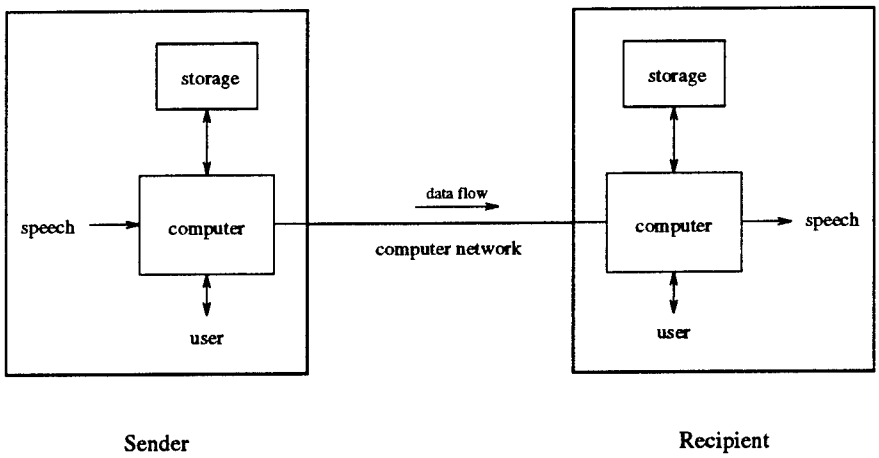


Figure 3: Voice Mail Application

Windows), where the device parameters and encoding are made explicit in some form of header. The following table presents an overview of popular file formats.

The main characteristics of audio hardware are described by the sample rate (Hz) and the resolution. For current audio hardware, the sample rates vary from 6 - 50 kHz and the resolution is 8 or 16 bit linear or μ-law. The audio hardware of a modern computer usually provides multiple sampling rate and quantization options.

Table 2: Some audio self-describing file formats

Extension or Name	Origin	Parameters
.au, .snd, .vox	NeXT, Sun	rate, #channels, encoding, info string
.aif(f), .AIFF	MAC	rate, #channels, sample width, lots of info
.iff, .IFF/8SVX	Amiga	rate, #channels, instrument info (8 bits)
.voc	Soundblaster	rate, (8 bits/1 ch; uses silence deletion)
.sf	IRCAM	rate, #channels, encoding, info
HCOM	MAC	rate (8 bits/1 ch; uses Huffman compression)

One audio format may be supported by different audio hardware and be implemented in more than one computer system with a different range of parameters. In general, it is possible to define a conversion procedure between almost any pair of file formats - often with loss of information, however.

In addition, a variety of data compression methods (LZW, ADPCM, LPC and STC [4]) are used to reduce the requirement of the network bandwidth and hard disk storage. It is generally impossible to convert compressed audio files to different audio formats or a file compressed by different compression methods without losing significant amounts of information. The variety of audio hardware, audio file formats and compression methods makes the exchange of audio files between computers on the Internet very complicated.

In comparison with e-mail, voice mail transmission requires much higher throughput and storage (500 times higher than the e-mail for low quality voice mail and 10000 times higher for CD quality voice mail). Therefore, throughput and storage may be a problem in future voice mail systems. To prevent unnecessary high quality audio files flooding a system, the voice protocol must be able to get information about out-going audio files and determine whether to accept them or not automatically.

Existing voice mail protocols such as NVP (Network Voice Protocol, RFC741, [3]) and VFIP (Voice File Interchange Protocol, RFC978, [12]) were not designed originally for the variety of hardware and audio formats of current personal computer systems and workstation systems. MIME (Multipurpose Internet Mail Extensions, [1]) was designed for future multimedia mail communication. It is not clear whether MIME can provide a common base for modern audio voice mail and support the wide a variety of audio formats, at least in the near term. Furthermore, available commercial voice mail packages used on the Internet are usually designed to convert audio files from one commercial format to another one and not designed to check compression methods in order to transfer compressed audio files.

It is therefore important that complexity and diversity of audio hardware and audio formats used in the Internet environment be considered when a new voice protocol is designed.

Methodology

The two main thrusts of this project, coder development and voice mail application development, were able to proceed relatively independently. These efforts interacted only during the instantiation of the coder into the voice mail application. Different personnel were involved in the segments.

Speech Coding Methods

The STC speech coding research began by addressing the difficulty of working with the existing STC software. The STC source coder has been used as a research tool at MIT Lincoln Laboratory and Rome Laboratories, and was not designed to be production code. As a result, the code contains the remnants of many years of research which are not relevant to the current work. The code was thus revised to reflect current requirements and the need for well-structured code appropriate for use by novice researchers. The research in this area then proceeded to use this code to find STC parameter settings appropriate for medium (16 kb/s) bit rates. Further, this code was used to investigate the performance of low delay FFT windows. Informal listening tests were performed to evaluate the performance of the various parameter settings as well as the windowing functions.

Voice Mail Application

The voice mail application was developed in the UNIX environment commonly available on Sun and DEC workstations. The code development was segmented according to the functional blocks necessary for the application. In particular, code was developed and tested for the server and client daemons for the mail protocol which need to run on each host supporting this system, as well as for the graphical and non-graphical user interface programs. Complete system testing has also been performed, and more extensive off-site testing is being investigated.

Results

Speech Coding Research

Coder Revisions

Several modifications were made to the source code of the STC speech coder in an attempt to improve its usability by researchers. This allowed this project to more easily extend the original version, and should provide benefits to future research in this area.

The source code was restructured for improved efficiency. For example, all parameters related to the pitch were declared as members of a structure called pitch. The voicing parameters became members of a structure called probability, etc. The structures are declared in a single file that was included with every function. Many of the global variables were eliminated. This method clarified parameter passing to various routines.

Most simulations were based on zero channel spacing, so the channel warping was eliminated.

A single synthesizer was developed, eliminating the need for a separate one in the case of rate transformation. It has been tested extensively and is producing correct audio speech files.

Some setup files have been eliminated and the rest configured so that there is one for the analyzer, one for synthesizer, and one for the transformation. For this, a variable called "prog_type" has been defined in the main programs where prog_type is 1 for the analyzer, 2 for the synthesizer, and 3 for the transformation. This allowed the identification of the parameters passed to the pitch and voicing coding/decoding.

Instead of a single median() function, 2 separate ones were created: median_pitch() and median_voicing. This was a consequence of the requirement for proper parameter passing. The functions are part of "f0subs.c".

All STC programs now use a single channel coding program "code_gs.c", thus eliminating all the ones related to the transform programs. This necessitated the definition

of a pre-processor variable named "XFORM". This variable is set to 1 if the transform programs are used and 0 if they are not.

All STC programs now use uniform data files. When a transform program is used, the data file for the initial rate is entered at the command line followed by the data file for the final rate. This eliminated the need for the two extra data files used previously. A single line is included in the data file indicating whether the number of bits will be specified for the final rate in transform programs, so this in combination with the global "XFORM" dictates the flow of control in this data file. This also conforms with the new code_gs.c routine structure.

All references to SPOX have been deleted form the updated files, and so goes the graphics code.

The vector quantization code developed by MIT Lincoln Laboratory for use in their current version of the STC was added to the new version developed by the University of Kansas, so that this major breakthrough could be available in the restructured software.

Medium Coder Parameters

Various coder parameters were informally tested for a medium rate coder. While various combinations produced good speech quality, the increase in quality over an example 8 kb/s coder was not great. This observation implies that alternate coding schemes such as transmission of phase information should be used to exploit the additional bandwidth. Attempts to implement such schemes lead to the previously described coder restructuring effort. Phase coding methods continue to be pursued.

Low Delay Windows

Because of the robust nature of the representation used by the STC algorithm, not only can the coder operate at various rates [9], but transformations in the parameter space can be used to change the bit rate of the coded data. The usefulness of the STC at higher rates, however, is somewhat limited by the large delay imposed by processing requirements,

including that incurred by FFT windowing. In order to address this concern, an asymmetric window was incorporated into the code. It is basically a two-pole window [5] that has the form

$$w(n) = b_1^n - ab_2^n \tag{1}$$

where $a = 0.85$, $b_1 = 0.97$, and $b_2 = 0.95$ have been shown to provide good performance. Its z-transform is

$$W(z) = \frac{(1-a) + (ab_1 - b_2)z^{-1}}{1 - (b_1 + b_2)z^{-1} + b_1 b_2 z^{-2}}. \tag{2}$$

A Matlab program was written in order to analyze the frequency and phase response of the window. The frequency response was very smooth and decreased monotonically with increasing frequency. No ripple was observed. The phase response was extremely close to linear. At lower frequencies, it showed a slight non-linearity, but this is irrelevant since a low-pass filter will be used. Other combinations of windows (i.e. rectangular and Hamming) were tried but the results obtained from the 2-pole window produced the best linear phase approximation. The exponential window's first sidelobe was 27 dB less than the peak, and the second sidelobe down 7 dB. The centroid for this window (480 points) was calculated to be at (46.421, 0.082) for a delay of 5.80 ms when using a sampling frequency of 8 kHz. The code for this window was included in the file "stft_ana.c" under the anlwin() function. The control parameter specified in the data file will allow its use ("exp" instead of "cos" for the Hamming/cosine window).

The window can be shown to produce an increase in the signal-to-noise ratio for most sentences, along with a crispier speech than the speech observed previously with a Hamming window for other coders [5]. The speech audio files resulting from the 2-pole window produce a slight chirping at the end of each utterance. This is believed to be as a result of the voicing probability produced from the voicing rule function. An attempt to

vary the hard-coded parameters within each voicing rule, such as the SNR, noise level, and the upper and lower limits on the allowed SNR is currently underway.

In order to evaluate the delay of the coder with the new window, a program was written to produce an impulse function, where a single value was non-zero. This impulse was then used as an input to the STC coder in order to evaluate the delay introduced in the coding process. The delay observed from running simulations on a speech input was 40 ms. However, the delay observed from the impulse did not make sense because the result indicated a 110 ms delay. The SoundTool program available on SPARCstations was used to determine the delay; the sample at which a speech waveform feature appeared was identified and the time to that sample was measured. This technique is not very accurate, but it was the only one available at the moment.

Internet Voice Mail Application Development

A new voice mail protocol which guarantees correct transmission of audio files between computer systems with different audio hardware and audio formats through the Internet was developed. A voice mail application based on this protocol was implemented on several UNIX workstation systems. This voice mail application also provides a graphic user interface based on the X-Windows Athena widget set, and a non-graphic user interface based on the UNIX "curses". By using these user interfaces, users can play and record audio files, compress audio files to reduce the requirement for the network bandwidth and hard disk storage, and send and receive audio files to and from a remote computer system with different audio hardware.

The Protocol

The Internet Voice Mail Protocol (IVMP) can be used on low-end computers with one type of audio hardware (called a computer with single audio-mode or single mode computer) as well as on high-end workstations with several types of audio hardware (a multi-mode computer system). All audio formats are divided into two categories: standard

audio formats and non-standard formats.

The server for a local network (LAN) may provide voice mail service for several types of computer systems such as IBM PCs or MacIntoshes, which are connected to this local network. This server can be considered as a computer with multi-audio modes.

The conversion between any pair of standard audio formats is well defined, and the conversion between any pair which includes one or two non-standard audio formats is experimental, mainly provided by programmers who use such new audio hardware, new audio file format or new compression method.

The implementation of this voice mail protocol includes a database of standard audio formats, compression methods, and a library of conversion functions between standard file formats. The higher the version of the implementation, the larger the database and the library. The local programmers are required to configure the conversion table of their computer system. The entries in this conversion table are determined by the configuration of the computer system and the types of audio files the "server" can accept. Local programmers can add non-standard conversion relationships into this conversion table as well.

The lowest version of the implementation of this protocol is defined as 1. Local programmers can define the version of the "server" as zero when they want to force all clients to convert audio files before they are sent.

This voice protocol provides two services: 1) the client sends information about the queued audio files to the remote site and gets information regarding audio formats the remote computer system can accept before the transmission of audio files start; 2) audio files are transmitted through Internet network if these audio files are acceptable by the remote site.

The negotiation between a client and a server is completed by the exchange of control messages. A simple algorithm provides basic negotiation capabilities between the client and the server. This algorithm can be considered as the core of a more complicated negotiation

process which may add more criteria in order to set different priorities to different types of audio files or different clients.

Eleven control messages are specified in the protocol:

#1 Connection Required Message

#2 Connection Response Message

#3 Standard Information Message

#4 Standard Information Response Message

#5 Nonstandard Information Message

#6 Nonstandard Information Response Message

#7 Disconnect Message

#8 Resend Information Message

#9 Short Connect Require Message

#10 Renegotiation Message

#11 Multi-mode Information Message

In this simple negotiation algorithm, the negotiation processes are divided into 5 modes based on the parameters of the audio messages and the acceptable parameters of the server. For mode I, the server sends its parameter to the client and later converts audio files from one standard format to another standard format if necessary, or determines whether a compressed audio file should be sent. Mode II is similar to mode I but the client and the server exchange roles. Mode III is also similar to mode I but both sites support non-standard formats. In modes IV and V, only one site supports a given standard audio format. The client supports non-standard formats for mode IV, and the server supports non-standard modes for the mode V. In modes IV and V, the site which supports the non-standard formats does the conversion of the audio files. Mode I is designed for most audio file transfers and the rest of the negotiation modes are designed for special cases.

IVMP Implementation

An Internet voice mail application for the UNIX operating system is shown in Figure 4. This system is designed to complete the following operations: play and record audio files; compress audio files; and send and receive audio files through the Internet to/from a remote site with different audio hardware. This system consists of the user interface (graphical and non-graphical) module, the data compression module, the audio hardware module, the voice mail protocol module, and the database module.

A versatile tool for converting between various audio formats is Sound Tools [11] (not to be confused with the NeXT library of the same name). This is a library for reading and writing various types of audio files and applying special effects (e.g., echo, or channel averaging) to buffered audio data. In this implementation, the conversions of audio files between current audio formats are completed by calling the Sound Tools Library.

All information needed in the user interface module, data compression module, and the voice mail protocol module are stored in the database module. Database files in the database module can be modified through the user interface or through any text editor. The voice mail protocol provides the "client" and "server" which both are daemon processes and always run in the background. The parameters of an outgoing audio file are sent from either the graphic user interface "xivmp-ui" or the non-graphical user interface "ivmp-ui" to the daemon process "client" through the BSD UNIX socket interface.

The voice mail protocol is designed to exchange information regarding audio file formats and audio hardware between the "client" and the "server" before the transmission starts. Information regarding particular audio files is stored in a special database file when an audio file is created.

The graphical user interface provides a friendly way to control this voice mail application, but it requires many resources. The non-graphical user interface provides an alternative on UNIX systems. It is designed to match the graphical user interface in

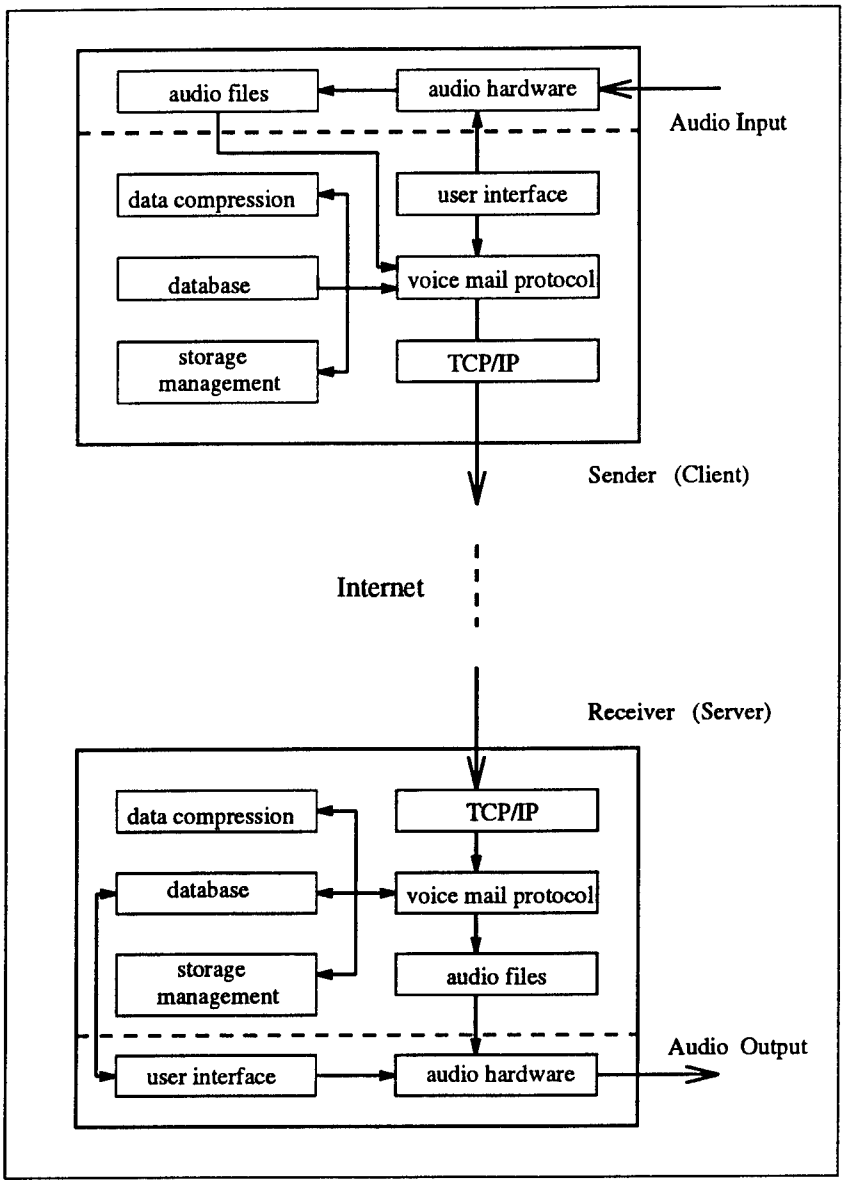


Figure 4: Voice Mail Application

functionality.

In the voice mail protocol, "client" and "server" programs are implemented. Each consists of three parts: a connection process based on BSD sockets; a negotiation process based on control messages and the algorithms mentioned earlier; and a file sending or file receiving process based on the TCP/IP suite. When the "client" receives a set of parameters for an audio file from a user, it forks off a child process to handle the negotiation process and the transmission process between "client" and remote "server". After successful negotiation, the "client" will signal the owner of this audio file using e-mail. Similarly, the "client" will signal the owner of errors when the audio file cannot be sent. Except for the case of a failed connection to the "server", the "client" will delete audio file, modify the database, and terminate the child process after signaling success or error. When the "server" accepts a connection, it forks off a child process to handle the connection. The parent process continues to wait for more connections. After completing the transmission, the "server" modifies corresponding database, announces the acceptance of this audio file, and terminate the child process.

The compression module uses the STC analyzer to code the data in the STC format, if it is not already using that representation (this encoding provides a 4-to-1 compression over standard 64 kb/s μ-law coding). After the initial compression, the data is stored to disk. The storage management logic then checks the storage space remaining available, and performs rate transformations as required. The memory management control logic is shown in Figure 5. This simple scheme reduces the bit rate of all of the stored messages when a specified threshold is reached. In this manner, the total memory usage is reduced, with corresponding quality reductions in all of the messages. More advanced control mechanisms, such as transformation of previously checked or old messages before new messages, could be applied if desired.

This implementation of the voice mail application has been extensively tested on Sun

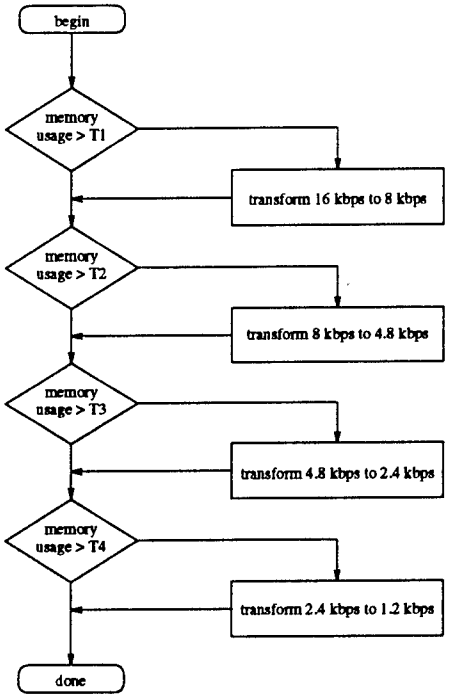


Figure 5: Message Compression Logic

SPARCstation-1+ and SPARC IPX hosts, and Digital Equipment Corporation DECStation-5000 workstations.

Summary of Results

This research has extended speech coder algorithms and produced utilities for networked workstations which provide increases in the utility of those coders. The modified STC coder and documentation, and the voice mail application and documentation, will be supplied to Rome Laboratories under a separate cover.

The results of this and related research have been described in the following papers:

J. B. Evans and T. G. Champion, "Robust Speech Coding and Reconstruction Techniques", *1992 Int. Conf. Signal Processing Appl. & Tech.*, Boston, Massachusetts, Nov. 1992.

B. Yu and J. B. Evans, "An Internet Voice Mail Protocol", *1992 Midwest Elec. Tech. Conf*, Ames, Iowa, March 1992.

Further publications based on this work are expected.

Conclusion

The work performed under this program has explored algorithms for improved capability for speech data storage and transmission, and resulted in computer network applications demonstrating those algorithms.

A multi-stage speech compression algorithm using parameter space transformations, which was developed as part of the Air Force Office of Scientific Research Summer Faculty Program, was investigated. This new method allows speech quality to be balanced against storage requirements, while maintaining high quality for each particular rate. This algorithm provides for more efficient use of resources; some quality can be sacrificed as conditions require, while normal quality can be maintained when resource usage is light. In this project, parameter selections for a medium rate (16 kb/s) STC were studied. The results of these studies suggest that additional work in phase coding would be valuable for obtaining higher quality. This project also investigated new techniques for reducing the delay of the STC algorithms, so that they can be used in environments where that characteristic is important. A low delay FFT window was evaluated for use with the STC, and was found to provide acceptable speech quality with some modification to various STC parameters.

The utility of the parameter transformation technique was demonstrated by its use in an example voice messaging application. This application runs under the UNIX operating system, and makes use of the TCP/IP protocol suite for reliable transport of the speech data. Two user interfaces were developed, one based on an X-Windows graphical interface and the other based on standard dumb terminals. A novel negotiation and transmission protocol was developed for this application. This application uses the STC transformation technique to allow dynamic management of disk resources on recipient hosts. Additional research into the application of these techniques will certainly demonstrate their value for easing integrated network management and control as well as host resource management.

References

[1] N. Borenstein and N. Freed. MIME (Multipurpose Internet Mail Extensions) Mechanisms for Specifying and Describing the Format of Internet Message Bodies. Internet Working Group Request for Comments 1341, USC/Information Sciences Institute, Marina del Rey, California, June 1992.

[2] T. G. Champion. Theory of parameter space transformation techniques. Tech. rep., Rome Laboratories, to be published.

[3] D. Cohen. Specifications for the Network Voice Protocol (NVP). Internet Working Group Request for Comments 741, USC/Information Sciences Institute, Marina del Rey, California, Nov 1977.

[4] Joseph B. Evans. Sinusoidal transform coder parameter manipulation techniques and their use in network and data storage applications. TISL Tech. Rep. 1660-6, University of Kansas, Sept 1991.

[5] V. Iyengar and P. Kabal. A low delay 16 kb/s speech coder. *IEEE Trans. Acoust., Speech, Signal Processing*, 39(5):1049–1057, May 1991.

[6] R. J. McAulay and T. G. Champion. Improved interoperable 2.4 kb/s LPC using sinusoidal transform coder techniques. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 641–643, 1990.

[7] R. J. McAulay and T. F. Quatieri. Magnitude-only reconstruction using a sinusoidal speech model. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 27.6.1–27.6.4, 1984.

[8] R. J. McAulay and T. F. Quatieri. Mid-rate speech coding based on a sinusoidal representation of speech. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*,

pages 945–948, 1985.

[9] R. J. McAulay and T. F. Quatieri. Multirate sinusoidal transform coding at rates from 2.4 kb/s to 8 kb/s. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 1645–1648, 1986.

[10] R. J. McAulay and T. F. Quatieri. Speech analysis-synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34(4):744–754, Aug 1986.

[11] L. Norskog. SoundKit documentation. Usenet distribution, July 1991.

[12] J. Reynoids, R. Gillmann, W. Rackernridge, A. Witkowski, and J. Postel. Voice File Interchange Protocol (VFIP). Internet Working Group Request for Comments 978, USC/Information Sciences Institute, Marina del Rey, California, Feb 1986.

A FRAMEWORK FOR ADAPTIVE FAULT MANAGEMENT IN SURVIVABLE C^3I SYSTEMS

Rex E. Gantenbein
Associate Professor
Department of Computer Science

University of Wyoming
P.O. Box 3682
Laramie WY 82071

Final Report for:
Research Initiation Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington DC

and

University of Wyoming

December 1992

A FRAMEWORK FOR ADAPTIVE FAULT MANAGEMENT IN SURVIVABLE C^3I SYSTEMS

Rex E. Gantenbein
Associate Professor
Department of Computer Science
University of Wyoming

Abstract

Most strategies for fault management are effective only for a narrow range of fault classes. In survivable distributed systems, a wide range of operating environments may be encountered that require different strategies to be used at different times. This report discusses how adaptive fault management can be used to select the most appropriate methods for assuring survivability under conditions that can suddenly and drastically change.

There are three basic issues in this research. First, a taxonomy for fault management mechanisms must be established that allows a system to differentiate among possible choices. Second, a characterization of operating environments must be developed that can be used by a system to dynamically match a fault management mechanism to the current environment. Finally, a distributed testbed must be constructed that will support a C^3I application and the collection of metrics that fit into the models constructed for the abovementioned mechanisms and environments. This paper describes the results of our investigation into these three issues and our plans for future work.

A FRAMEWORK FOR ADAPTIVE FAULT MANAGEMENT IN SURVIVABLE C^3I SYSTEMS

Rex E. Gantenbein

INTRODUCTION

Survivable computing systems are those designed to continue to perform adequately in the face of various kinds of adversity [Neumann92]. The increased dependency of the military on computer-based command, control, communication, and intelligence (C^3I) systems has made the survivability of such systems an important research problem. A complex system that can justifiably be trusted to perform its duties under less-than-ideal conditions must include some form of *fault management* to allow the system to detect and respond to failures due to these conditions.

Unfortunately, most strategies for fault management consider only a very few specific classes of faults to which they can respond. The assumptions about a system's operating environment inherent in its fault management strategy cannot be violated when the system is deployed. If this happens, the system may be unable to cope with the conditions encountered and, consequently, be in danger of catastrophic failure. This presents a problem for many military systems, as operating environments can rapidly or drastically change under combat conditions.

To prevent system failure due to a rapidly changing environment, a system can employ *adaptive fault management*, a design and operation paradigm in which the fault management mechanisms can be changed or modified to respond to variations in the system's behavior, environment, or requirements. Unlike static fault management, which typically uses a single mechanism to respond to faults in the system, adaptive fault

management allows the system to bring into play any of a number of alternative mechanisms. By adapting to changing conditions, the system can continue to perform adequately in a variety of environments, including significant loss of processing, data, or communication resources.

Over the past year, researchers from the University of Wyoming Department of Computer Science have attempted to establish a foundation for the study of adaptive fault management as a method for achieving survivability in distributed C^3I systems. Our goal has been to define the major issues in this approach and to propose some solutions. In this report, we present the results from our research so far.

RESEARCH ISSUES IN ADAPTIVE FAULT MANAGEMENT

We present our research in the context of distributed computing systems, since many C^3I applications are designed to be distributed over several processing nodes, geographically dispersed and loosely coupled by communication links. Each node may contain resources for processing, storing, and transmitting information, although not all resources need be present at every node. Operations to access instances of each of these resource classes are provided by server processes on behalf of (potentially remote) clients, which exchange messages over the communication system to request and carry out resource accesses.

We also assume that the system is dedicated to a single, multi-user application composed of a number of cooperating client processes, which may communicate among themselves as well as with servers. This application may be partially layered, so that one component of the application provides a higher-level service interface to another. In this model, an *error* occurs when the application (or any component) is unable to meet its

requirements with respect to a service, and *failure* is the consequence of an error manifesting itself in a server's inability to deliver its required service to one or more of its users. Survival, in this case, thus means the avoidance of failure in critical services at the highest (end-user) layer of the application.

The success of an adaptive approach to fault management is dependent on the system being able to dynamically determine how it will attempt to avoid failure. Our approach uses an adaptation data base containing information about the mechanisms that may be employed for this purpose [Gantenbein92]. This information consists of:

- a rule base that supports decisions about the appropriateness of each available strategy under existing internal and external conditions, and
- a data base that contains characterizations of the mechanisms that can be used to implement these strategies.

During operation, the system continuously collects information about the current external and internal conditions to interpret its environment. An *adaptive behavior manager* in the system uses the rule base to relate this information to the data base of mechanisms and to select the fault management mechanism that can best maintain "adequate" behavior under adverse conditions.

There are a number of research tracks inherent in an adaptive approach to avoiding failure. Our work has concentrated on the following three tracks:

- characterizing mechanisms for fault management to differentiate them for adaptive decision making;
- defining system requirements so that the system can efficiently determine whether

it is performing "adequately" under a variety of adverse conditions; and

- selecting a testbed that can support both development of reliable distributed applications and collection of run-time data that can be used in making decisions about the need for change in the fault management mechanism.

These three research areas are fundamental to the adaptive fault management approach to survivable systems. Our work has concentrated on developing models that can be used to evaluate proposed solutions for each track. We present the results of our work in the next section.

RESEARCH RESULTS

The University of Wyoming's involvement in adaptive fault management began with the author's participation in the 1991 AFOSR Summer Faculty Research Program at Rome Laboratory. Here, collaboration with members of the Computer Systems branch of the Command, Control, and Communication Division led to a successful proposal to the AFOSR Research Initiation Program for 1992. Additional support for this work came from the University of Wyoming Planetary and Space Science Center and from a second year of participation in the Summer Faculty Research Program during 1992.

Characterizing fault management mechanisms

The four basic phases of fault management are error detection, damage assessment, damage confinement, and response. Most previous classification schemes for fault management mechanisms have grouped the mechanisms according to the type or amount of hardware or software redundancy required and the phase at which the redundancy is applied.

Since we are more interested in the operating conditions in which a given mechanism is most effective, we instead define three general fault management strategies of which a mechanism may be an instance [Gantenbein91]:

- *optimistic* strategies are those used when resources (including time) are sufficiently plentiful to allow explicit detection and recovery from errors that occur without significant effects on the other requirements of the system;
- *pessimistic* strategies are used when the number, type, or frequency of errors requires the additional allocation of system resources to fault tolerance in order to prevent errors from occurring and causing loss of critical services; and
- *ultrapessimistic* strategies are used only when conditions have degraded to the point where too few resources are available to maintain all critical services as well as the fault tolerance mechanisms.

These categories serve as the basis of a taxonomy for fault management techniques. By classifying a technique into one of these three categories, we can effectively determine whether it matches a particular set of operating conditions. We can broadly apply these categories to the three basic resource classes in a distributed system [Kohler81]: processing (generating or transforming information), data (storing information), and communication (transmitting information). We classify individual techniques as optimistic, pessimistic, or ultrapessimistic depending on the resources needed to detect, confine, assess, and respond to errors in the system.

Optimistic processing techniques depend on explicit error detection, which requires a significant amount of *a priori* information, both about the application and the errors that

can occur in its operation. Response to a detected error typically involves restoration of an error-free state and an attempt to repeat the operation or invoke an alternate component. System services are suspended during the response, but little information about the application or error is needed to carry it out.

Pessimistic strategies typically require replication of a server or resource, both to detect errors and to provide redundant information that allows an error to be masked. This involves significant resource allocation to fault management. Compared to the optimistic techniques, either new allocations will need to be made or some degradation in other services will occur. Since errors are not explicitly detected and an error-free state is created dynamically by the replicated resources, very little *a priori* information is required for either error detection or response.

The resource overhead of both optimistic and pessimistic strategies allow a system to continue to perform adequately under adverse conditions, as long as sufficient resources exist. However, in some situations, the resources may not be available for either of these approaches to be effective. If the system is to survive, detection and response must be carried out using information about the application and the error to set the system state to a default or approximated error-free state so that the system can continue to provide at least some portion of its expected service. Unfortunately, an approximation or default is likely to cause the system to violate its specification, at least temporarily. Therefore, the expectations on the service must be relaxed to avoid the degraded service delivery from being detected as an "error" and causing further degradation. This requires significant *a priori* knowledge about the application.

Most well-known fault management techniques fall into one of these three strategy classes. Optimistic and pessimistic techniques have been widely used to detect and recover from and mask errors in both hardware and software. Ultrapessimistic techniques are less widely used, probably because of their application-specific nature. In Figure 1 below, we present a number of fault management techniques, taken from [Armstrong91], as examples of these general strategies for the processing, data, and communication resource classes.

	Classification Characteristics	Processing	Data Storage	Communication
1	**Optimistic:** devotes resources primarily to computation; needs *a priori* information for error detection; recovers state by recomputation	Standby versions: Cold hardware sparing, recovery blocks	transactions, checkpointing	retransmission
2	**Pessimistic:** devotes resources largely to error detection and response; needs little *a priori* information for either task; recovers state by masking	Multiple versions: N-modular redundancy, N-version programming, hybrids	replicated data, error-correcting codes	error-correcting codes, multicasting, redundant media
3	**Ultrapessimistic:** devotes resources to critical services; relaxes system requirements; requires *a priori* information for both error detection and response; recovers state by approximation	cold sparing without checkpoints	compensating transactions	network partitioning

Figure 1. Examples of fault management strategies.

Monitoring the operating environment

The identification of these three strategies for fault management allows us to differentiate among a number of mechanims and determine which are most effective in a given operating environment. Howver, the need for dynamic selection of a fault management mechanism is created by a change in operating conditions, which must

somehow be signaled to the adaptive behavior manager. Thus, the system must be able to monitor the operating environment and determine whether a change has taken place that indicates the need for change in the fault management strategy.

Monitoring operating conditions at run time requires an efficient, quantifiable metric that can measure changes in the environment and, more importantly, their effect on a system's ability to deliver its services. This metric, the *objective function*, is a partial specification of a service that uses derived or observed attributes representing the environment to evaluate the delivery of the service in that environment [Gantenbein91].

This function should express all of the objectives that a system service must meet to be survivable. Objectives may include such things as:

- functionality, the extent of the delivery or accessibility of the expected service;
- performance, a measure of the timeliness of the delivery of the service;
- consistency, the agreement among a set of servers (or among successive invocations by a single server) on the contents or attributes of a service;
- precision, the latitude that will be accepted in the delivery of a service;
- security, a measure of the service's resilience to deliberately introduced faults; and
- safety, the risks associated with delivery of the service.

Objectives such as these can be defined as an expected region for the value of the objective function, which may be a single function or a composition. The value of the function then measures how closely the delivery of the service matches its expectations, and can thus be used by the adaptive behavior manager to decide when a new strategy is needed and which strategy to apply next.

To illustrate, we can define an objective function to measure the number of active replications of a process in a distributed system of m nodes with $n \leq m$ tasks that must be executed on some subset of those nodes. Let $[X_{ij}]$ be the n by m matrix of zeros and ones, where $X_{ij} = 1$ if task i is scheduled on node j and 0 otherwise. The number of processors to which task i is to be assigned is the replication factor r_i of the task. We can evaluate whether the scheduler has met the requirement that each task be assigned to $0 \leq r_i \leq m$ nodes using the function

$$\sum_{j=1}^{m} X_{ij}$$

in which $[X_{ij}]$ is updated by the scheduler and i and m are system design factors. With this function, the adaptive behavior manager can determine the replication of any task in the system and thus evaluate the effectiveness of the current scheduling algorithm.

To carry out this kind of analysis for all the dependability attributes of a distributed C^3I system, we must define the kinds of environmental changes to be considered in each of the resource classes identified in the previous section. From this, we can develop metrics that will allow us to monitor the changes and determine when a change in the fault management strategy is necessary.

According to our model, change in the fault management strategy is necessary when the strategy's assumptions about the operating conditions do not match the current conditions. Methods for detecting such a situation include monitoring the conditions for such things as resource availability, error rates, communication delays, etc., and match the conditions against the rule base to determine which mechanism would be most appropriate for the conditions. In this way, the system can predict problems in the system before they

occur.

However, the objective function also allows us to explicitly detect errors in the system. If the computed value of an objective function does not fall into the expected region, then we say that an *anomaly* has occurred in that objective. An anomaly is, obviously, the internal manifestation of an error in the system, so we can model errors in service delivery in terms of anomalies in the objective for the service. The advantage of this model is that, since we have a region for objective values for any service, we allow the system some tolerance in meeting it. Furthermore, tradeoffs between objectives can be specified in this model by policies such as maximizing the number of objective functions in their expected regions or minimizing the likelihood that an objective will not be met.

In order for this kind of adaptivity to be effective, it is necessary that, first, the system designer be able to specify the boundaries of adequate behavior for a given environment. The definition of these boundaries comes from the system requirements, which specify the expected behavior of a server and can thus be used to evaluate the server's effectiveness in the current environment. These boundary values [Kim90] are used by the objective function to indicate that a change in the environment has violated the assumptions of the current fault management strategy.

Matching the regions of adequate behavior to an appropriate fault management mechanism is a complex problem for a nontrivial system. One approach to doing this is to define the regions as failure classes that are likely to occur in the system (independent of the cause of the failure), and then link each failure class to a given mechanism that can prevent it. Any detected anomaly in the system would be mapped into a failure class, which

in turn would suggest the mechanism to recover from it. This level of granularity of error recovery presumes an efficient dynamic selection algorithm, which could be based upon an exception handling approach similar to that proposed in [Liu92].

Identification of the failure classes and the mapping to a given mechanism could be done using fault tree analysis. A fault tree illustrates the logical interrelationships of basic events that lead to an undesired event at the root of the tree [Hansen89]. This approach can be used to define the environment conditions that would lead to a failure. The design-phase characterization of the environment in terms of its failure classes, coupled with the run-time collection of information about the environment, gives an implementation of the objective function that can be used to handle both anticipated (explicitly detected) and unanticipated (predicted by mismatches between the environment and the mechanisms) errors.

Valuable insight into this problem may come from a proposed technique for automatic error recognition [Iyer90]. The methodology for this technique consists of:

(1) recognition of errors among which relationships can exist;

(2) confirmation of the relationships among the errors; and

(3) quantifying the strength of those relationships.

The rate at which system errors occur identifies the error states among which relationships occur by the identification of similar error patterns. Frequent instances of errors within a short time period describes the same error condition or, at the least, highly related errors. These error groups can be used to identify the symptoms of persistent problems and the times of their occurrence. This approach can be used for on-line error detection.

In the adaptive fault management model, such an approach can be used to differentiate persistent from transient errors and to separate detected errors from each other. However, there are some inherent drawbacks in this method. Repeated occurrences of errors can appear quite different, since the same fault could affect different components and thus cause the system to classify the error in different groups. Also, the method depends heavily on statistical techniques, which could be a disadvantage in defining a broad framework. Finally, the method here is directed largely at hardware, rather than software errors, which may not be so easily analyzed. More consideration is needed.

The task of creating such a set of metrics for the services in a complex system is difficult. The dependency between the properties of a metric and its usage are very important, and so the properties of the metric must be examined and tested as to what they measure [Kearney86]. Validation is also an issue; nonparametric statistical methods (such as the use of contingency tables) appear to have an important role in evaluating metrics against fixed validity criteria. We are exploring methods for validating objective functions as a measure of a fault management mechanism's effectiveness, as well as for error detection. Among the criteria that being considered are association, consistency, discriminative power, tracking, predictability, and repeatability [Schneidewind92].

Choosing a testbed for adaptive system evaluation

The issues to be considered in using an adaptation paradigm for constructing survivable systems include the system structure, the operational profile, the operating environment, and the status of the system once a failure (actual or potential) has been discovered. Abstract analysis of these factors is an extremely difficult problem, particularly

in a real-time environment, where the survivability requirements may conflict with the timing constraints [Shin91]. We therefore intend to develop a prototype survivable C^3I system in order to address those issues most important to the construction of these systems in a distributed real-time environment. The value of such a *testbed* approach, where validation of proposed techniques is based on their experimental application and the measurement of the results, has been shown in several research projects involving reliable real-time systems (e.g., [Kim89] and [Wedde89]).

The first step in constructing this prototype is the selection of a development platform that can support dependable distributed applications that exhibit the features of a C^3I system. A number of different platforms are available, and the features offered vary. Based on our intent to use this prototype as a feasibility study as well an experimental evaluation testbed, we have defined a set of six criteria for the testbed platform:

- The software system should allow applications to execute in native mode on the platform. No interpreter or secondary system should be necessary. Any secondary system would influence the experimental data we intend to collect from the prototype.
- The system should support data abstraction and information hiding in a well-defined form such as objects or modules. These abstractions should exhibit location transparency, migration, and replication.
- The system should support transactions or some other form of atomic action.
- The system should either be fault tolerant itself or provide tools for building fault-tolerant applications. Automatic recovery is desirable, although the prototype should

be allowed to make its own decisions about what components to make fault tolerant and which method is used.

- Graceful degradation should be possible, as should noninterrupting repair or replacement of objects in the system.
- The system should support real-time processing to the extent that hard and soft deadlines can be set and alarms raised if the deadlines are missed. High performance is not a direct issue, but the ability to specify a level of performance in terms of deadlines is essential.

Selection Matrix

OS	Native	Objects	Atomic	Fault Tolerant	Graceful Degradation	Real-Time
ISIS	N	y/n	Y	Y	Y	Y
Mach	Y	N	N	N	N	N
Clouds	Y	Y	Y	Y	Y	N
Amoeba	Y	Y	?	Y	Y	Y
Chaos	Y	Y	N	?	?	Y
Alpha	Y	Y	Y	Y	Y	Y
Cronus	Y	Y	Y	Y	Y	N
Maruti	N	Y	Y	Y	Y	Y
Rose	N	y/n	Y	N	N	N
Eden	N	Y	Y	y/n	N	N
NextStep	Y	Y	N	N	N	N
Galaxy	Y	Y	Y	Y	N	N
Dunix	N	Y	N	N	Y	N
ZGL	Y	N	N	N	N	N
V	Y	Y	Y	y/n	N	Y
Thoth	Y	N	N	N	N	N
Locus	Y	N	N	y/n	y/n	N

Figure 2. Results of platform evaluation.

We have used these criteria to evaluate a total of 17 different platforms in this project. The results are tabulated in Figure 2. Based on this evaluation, we chose three systems for further study.

ISIS [Birman91] is a distributed systems development toolkit that extends any of a number of UNIX dialects. While not directly supporting objects, a distributed object known as a *group* can be defined that can support replication, migration, and fault-tolerant communication. ISIS also includes a package called META that allows creation of soft real-time applications [Marzullo91]. META can also be used as a monitoring tool for performance and failure in an application. Upon failure, a process can be restarted using information kept by a log manager.

Chaos [Schwan90] is an operating system designed for development of large, complex real-time applications. Chaos supports, and is composed of, objects, which can contain multiple processes distributed over several nodes. The VCA feature of Chaos adapts an executing real-time application by dynamically moving it along a set of programmer-specified versions of the application, with each version having different functionality and performance. The goal of the Chaos system is to achieve the programmer-specified level of performance at all costs. Its fault-tolerant support is limited.

Clouds [Dasgupta91] is another distributed operating system that is object-based. The compilers in the system create and distribute objects to data servers, which are responsible for the storage of objects. The same object can be concurrently invoked by threads at different compute servers, supporting replication and consistency. No mention of a migration facility was made in the source material. Atomicity is achieved, however, by specialty threads defined by a locking and recovery subsystem. Support for graceful degradation is provided for multiple objects and threads, as is transparent replacement. There does not appear to be strong support for real-time processing.

As shown in Figure 2 above, each of these three systems has some advantages and disadvantages in terms of the criteria for an evaluation platform. Among these three candidates, however, ISIS appeared to have most of the features we required, particularly the support of fault tolerance. Since ISIS is a toolkit rather than a native-mode operating system, ISIS applications run directly atop UNIX, using the ISIS facilities mostly for communication and coordination. We were also impressed by the stated capabilities of META for run-time monitoring of both the environment (as required by the objective function) and system attributes (as part of the evaluation of different techniques for fault management.)

In addition, we wanted the system to be compatible with systems available at Rome Laboratory, so that our experiments can be ported and demonstrated there. The Air Force has shown interest in ISIS as one of the platforms for its survivable system experiments, and the commercial version of the system is available to academic institutions at a significant discount. We therefore selected ISIS as the platform for development of the testbed.

PROJECT STATUS

Additional details of the work described in this report are documented in three papers that are currently under review:

- a description of the taxonomy for fault management strategies [Gantenbein93a];
- a summary of the use of the objective function as a mechanism for relating system requirements to system behavior [Gantenbein93b]; and
- an overview of the adaptive fault management approach to survival in distributed systems as an "innovative idea" report [Gantenbein93c].

An fourth paper is planned for submission to the next Symposium on Reliable Distributed Systems (SRDS) early in 1993.

The SRDS and the annual International Symposium on Fault-Tolerant Computing (FTCS) are very highly regarded conferences for researchers working in dependable computing. The principal investigator attended both of these conferences in 1992 with the support of the AFOSR Research Initiation Program grant. The contacts with others working in the field was very valuable; in particular, discussions with two individuals involved with the adaptive fault tolerance project -- Dr. K.H. (Kane) Kim of the University of California at Irvine and Dr. Lawrence Alexander of GE Aerospace -- provided motivation and insight for our work and opened avenues for further coordination and collaboration.

At this time, our work continues along the research tracks already identified. With internal support from the University of Wyoming through May 1993, we intend to complete the population of the taxonomy and refine the categories for better separation among the mechanisms. We also plan to formalize the definition of the objective function and extract the most important features of our current model for evaluation in a prototype distributed application. We have acquired version 2.2 of ISIS and will soon install it on 20 nodes through the Wyoming campus network. This configuration will be used as the platform for a prototype distributed application, which will simulate hostile and friendly targets entering and leaving the range of a set of radar stations. The system will be able to simulate both mission changes and resource loss to allow us to evaluate different strategies for survival and demonstrate the feasibility of the adaptation mechanism we have proposed.

CONCLUSIONS AND FUTURE RESEARCH

We have in this project established a solid framework for the further study of adaptive fault management as an approach to building survivable C^3I systems. Our next goal is to develop the characterizations of fault management mechanisms and operating conditions to the extent that the mappings between them can be quantified and implemented using META functions in our prototype application. Once these metrics are included in the prototype, we will be able to simulate various scenarios in the testbed and measure the prototype's ability to adapt to and survive rapid and drastic changes in the operating environment.

Our research, in combination with other work sponsored by Rome Laboratory, forms a cohesive effort in the area of adaptive, survivable systems. The relationships among these projects can be described by the research schema shown in Figure 3. In the short term, our work is helping build a general framework for the application of adaptive fault management to survivable C^3I systems. Development of the models for the environments and fault management mechanisms will lead to evaluation metrics for these systems, while the testbed study will identify ISIS features for adaptivity and survival that will be used by Rome Laboratory and its subcontractors, GE Aerospace and SRI Associates, in developing demonstration systems.

In the long term, we see this research as leading towards a generalized methodology for the dynamic redefinition of system behavior with respect to an operating environment that may change over time. Not only fault management, but other system components (such as task scheduling or resource management) become less effective when the assumptions

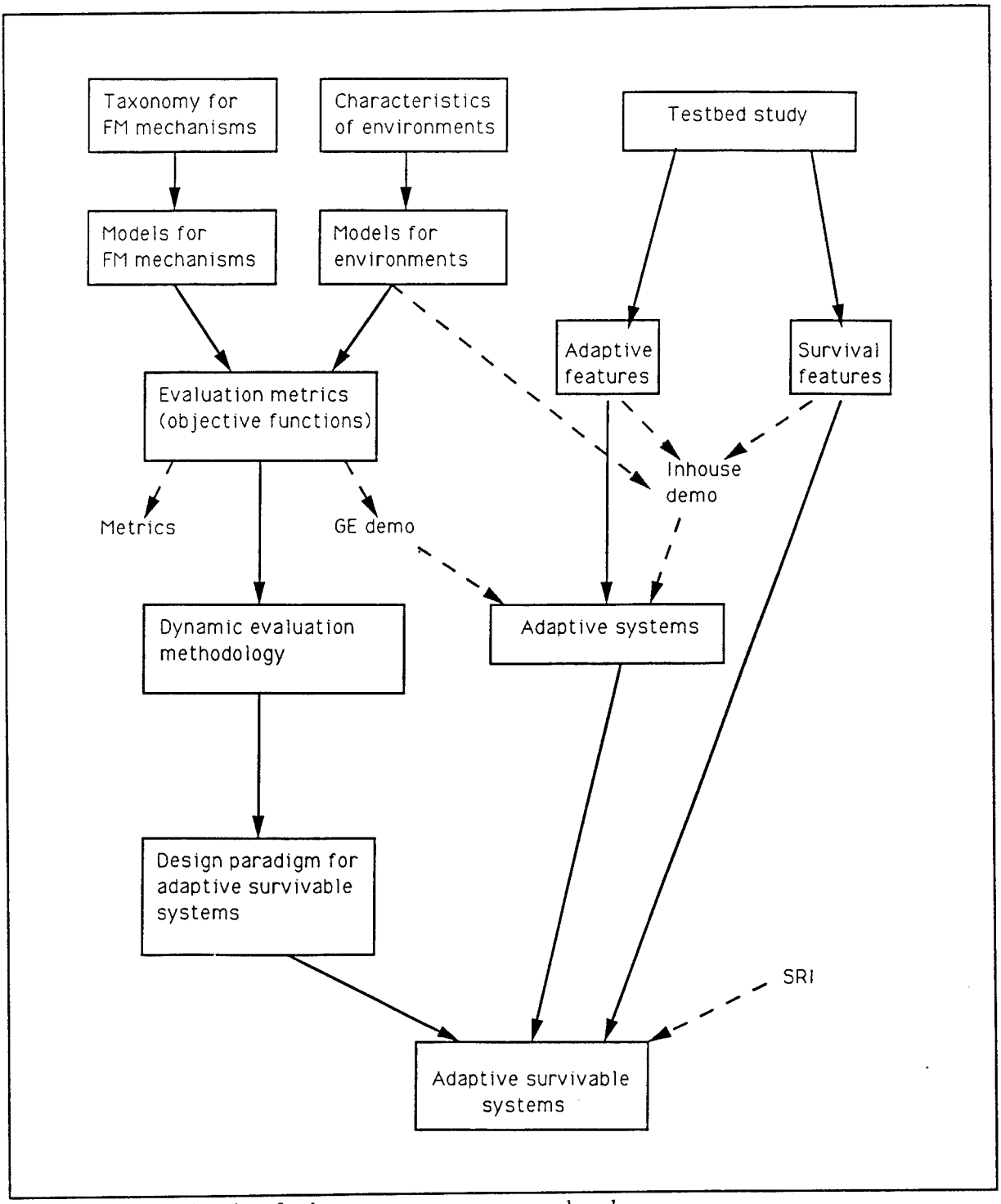


Figure 3. An adaptive fault management research schema.

inherent to them are violated. Eventually, we foresee a formal design paradigm for adaptive systems, supported by this methodology. With the increased reliance to be placed on

computer-driven systems in the future, we believe that this approach can make these systems increasingly dependable.

ACKNOWLEDGEMENTS

The author wishes to thank Tom Lawrence of Rome Laboratory for his vision and his help in this project. Thanks are also due to the other people at Rome Laboratory whose discussions and advice helped to make my two summers there worthwhile: Dick Metzger, Jerry Dussault, Mary Denz, Gary Craig, Mark Foresti, Tony Newton, and Scott Huse. Special appreciation also goes to Waleed Smari of Syracuse University, with whom I shared both office space and these ideas during my two summers in Rome.

Much of the work described in this report was the product of graduate students in the Department of Computer Science at the University of Wyoming. The author recognizes the valuable contributions of Subirendu Mukherjee, Joseph Gradecki, V. Gopalakrishnan, and Dinesh Fernandopulle. Dr. Sung Yun Shin, a Wyoming graduate and faculty member at South Dakota State University, also collaborated on this project.

BIBLIOGRAPHY

[Armstrong91] L.T. Armstrong and T.F. Lawrence, Adaptive Fault Tolerance, *Proc. 1991 Systems Design Synthesis Technology Workshop* (September 1991), Naval Surface Warfare Center.

[Birman91] K. Birman and R. Cooper, The ISIS Project: Real Experience with a Fault-Tolerant Programming System, *Operating Systems Review* 25,2 (April 1991), 103-107.

[Dasgupta91] P. Dasgupta et al., The Clouds Distributed Operating System, *Computer* 24,11 (November 1992), 34-43.

[Gantenbein91] R.E. Gantenbein, *A Taxonomy for Adaptive Fault Management in Survivable C^3 Systems*, Final Report, 1991 AFOSR Summer Faculty Research Program (August 1991).

[Gantenbein92] R.E. Gantenbein, *Issues in Adaptive Fault Management for Survivable C^3I Systems*, Final Report, 1992 AFOSR Summer Faculty Research Program (August 1992).

[Gantenbein93a] R.E. Gantenbein, T.F. Lawrence, and S.Y. Shin, Adaptive Fault Management in Survivable Distributed Systems, submitted to *Computer* (September 1992).

[Gantenbein93b] R.E. Gantenbein and S.Y. Shin, Adaptive Fault Management for Dependability in Distributed Systems, submitted to International Conference on Computer Applications in Design, Simulation, and Analysis (September 1992).

[Gantenbein93c] R.E. Gantenbein. T.F. Lawrence, and S.Y. Shin, Adaptive Fault Tolerance for Survivability in Distributed Systems, submitted to International Symposium on Fault-Tolerant Computing (December 1992).

[Hansen89] M.D. Hansen, Survey of Available Software Safety Analysis Techniques, *Proc. Annual Reliability and Maintainability Symposium* (1989), 46-49.

[Iyer90] R.K. Iyer, L.T. Young, and P.V. Iyer, Automatic Recognition of Intermittent Failures: An Experimental Study of Field Data, *IEEE Trans. on Computers* 39,4 (April 1990), 525-537.

[Kearney86] K.J. Kearney et al., Software Complexity Measurement, *Comm. of the ACM* 29,11 (November 1986), 1044-1050.

[Kim89] K.H. Kim, An Approach to Experimental Evaluation of Real-Time Fault-Tolerant Distributed Computing Schemes, *IEEE Trans. on Software Eng.* 15,6 (June 1989), 715-725.

[Kim90] K.H. Kim and T.F. Lawrence, Adaptive Fault Tolerance: Issues and Approaches, *Proc. 2nd IEEE Workshop on Future Trends of Distributed Computing Sys.* (September-October 1990), 38-46.

[Kohler81] W.H. Kohler, A Survey of Techniques for Synchronization and Recovery in Decentralized Computer Systems, *Computing Surveys* 13,2 (June 1981), 149-183.

[Liu92] C. Liu, A General Framework for Software Fault Tolerance, *Computer* 25,7 (July 1992), 38-46.

[Marzullo91] K. Marzullo et al., Tools for Distributed Application Management, *Computer* 24,8 (August 1991), 42-51.

[Neumann92] P.G. Neumann, Inside Risks: Survivable Systems, *Comm. of the ACM* 35,5 (May 1992), 130.

[Schneidewind92] N.F. Schneidewind, Methodology for Validating Software Metrics, *IEEE Trans. on Software Eng.* 18,5 (May 1992), 410-421.

[Schwan90] K. Schwan and A. Gheith, CHAOS^{arc}: A Kernel for Predictable Programs in Dynamic Real-Time Systems, *Real-Time Systems Newsletter* 6,2 (Spring 1990), 11-19.

[Shin91] K.G. Shin, G. Koob, and F. Jahanian, Fault Tolerance in Real-Time Systems, *Real-Time Sys. Newsletter* 7,3 (Summer 1991), 28-34.

[Wedde89] H.F. Wedde et al., Operating System Support for Adaptive Distributed Real-Time Systems in DRAGON SLAYER, *Operating Sys. Review* 23,3 (July 1989) 126-140.

1992 FINAL RL REPORT

WAVEGUIDES IN GLASS SUBSTRATES, WAVEGUIDE LASERS, AND TUNNELING ELECTRO-OPTIC **AND** GATES

by
Philipp Kornreich

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
Syracuse University
Syracuse N. Y. 13244
(315) 443 4447

December 1992

1 INTRODUCTION

We tested optical waveguides in glass substrates. These waveguides will be used in ultra high speed network pulse multiplexing optical circuits and as **C**oupled **W**aveguide **L**ight **A**mplifiers (CWLA) and waveguide lasers. They can also be used as **E**xternal **O**ptical **M**odulators (EOMs). The coupled laser can be controlled by applying bias voltages across the lasing waveguides. Both free space and fiber optic communication channels have wider bandwidth of the order of 10^{14} HZ, than any other communication medium. At present, the bandwidth of optical communication channels is limit by the electronic devices employed at the input and output of the channels. Current electronic devices are limited to a bandwidth of about 50 GHz (5×10^{10} Hz). This is a small fraction of the available optical bandwidth. These devices, in conjunction with passive optical waveguides, can be used as all optical multiplexers and demultiplexers that would convert multi channel parallel "slow" electronic input signals to high speed serial optical signals and and high speed optical signals to parallel multi channel "slow" electronic signals. As stated above, the coupled laser can be controlled by applying bias voltages across the lasing waveguides. The application of a voltage across one lasing waveguide will shift the frequency of the light from the particular waveguide with respect to the light from the other waveguide. The difference frequency of the light beams from the two waveguides would be in the mm-wave frequency or micro-wave frequency range. The electrical bias allows continuous control of the difference frequency. Such light sources have applications in **O**ptical **P**hased **A**rray **A**ntenna **S**ignal **P**rocessors (OPAASPs).

We also tested Tunneling Devices for Optical AND Gate applications. These type of gates are used in Ultra High Speed Fiber Optical Networks.

1 OPTICAL WAVEGUIDES IN GLASS SUBSTRATES

a) Planar and Ridge Waveguides

We fabricated two types of waveguides. We fabricated both planar and ridge waveguides. Planar waveguides are constructed by doping channels in a glass substrate just below the surface as shown in Fig. 1. Ridge waveguides have channels of doped glass that protrude above the substrate surface. Planar waveguides can only guide light around large radii, typically about 10 cm in radius. Ridge waveguides can guide light around substantially smaller radii of only a few millimeters. The light input and output to these waveguides is via single mode fibers. Indeed all tests to date on the waveguides were performed using light coupled into the waveguides through a single mode fiber.

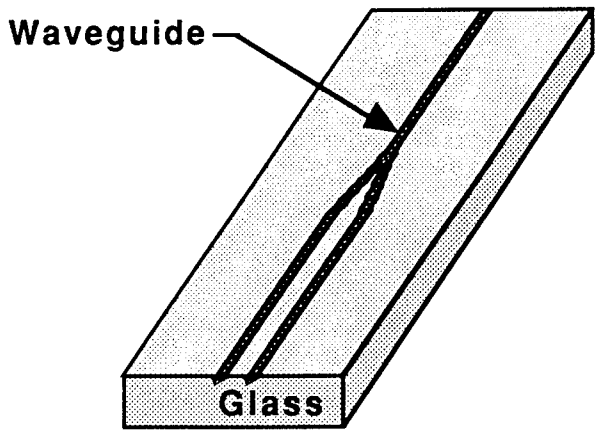


Fig. 1. Planar waveguide.

b) Waveguide Application

These waveguides have many applications. One of which is in **pulse multiplexing and demultiplexing networks**. A multiplexing network consists of a single input waveguide that branches into a number of waveguides. Each waveguide paths has a different optical lengths before

the path are recombined. The differences in the optical path length are integer multiples of the shortest path length difference. Consider a pulsed optical signal with a repetition rate longer than the longest difference in optical delay. Each path delays the pulse by a different time interval. When the delayed pulses are recombined a signal with a much higher pulse repetition rate results. The realization using a single glass substrate results in the construction of precise delays. Also since all the delay elements are located on the same substrate they expand and contract with temperature in unison.

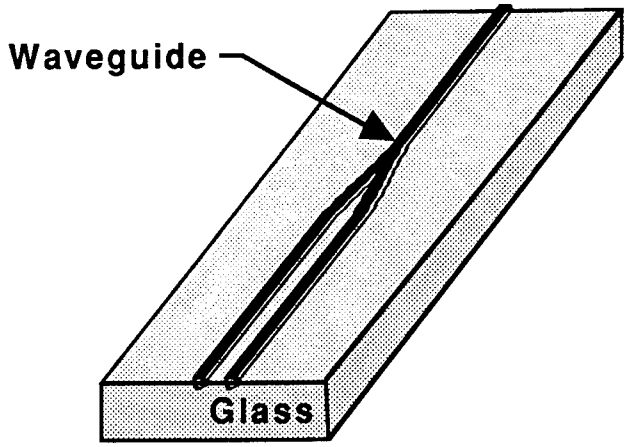


Fig. 2. Ridge waveguide.

However, to make such a device useful the pulses have to be not only delayed but also modulated. Therefore each path must also contain a modulator capable of switching the pulse through the particular waveguide guide on and off. These modulators can be fabricated by either depositing ZnO on the waveguide and patterning the ZnO into a Mach-Zhender interferometer modulator configuration or attaching a $LiNbO_3$ Mach-Zhender modulator onto the waveguide. Since both $LiNbO_3$ and ZnO have substantially higher indices of refraction than glass the light will switch from the glass into the ZnO or $LiNbO_3$.

We built and tested branching planar and ridge waveguides on glass substrates.

c) Waveguide Fabrication

The waveguide is fabricated by doping channels in the surface of optical quality glass with silver. Silver increases the index of refraction of the glass. Thus, light will be confined to the doped channels in the glass. Since the increase of index of refraction of the glass is small, only about 0.1 to 0.05% the light will be weakly guided. This restricts the curvature of the guides to large radii. On the other hand, since there are large regions on the sides of the waveguides where the light intensity decreases exponentially, it is easy to construct evanescently coupled waveguides.

The doping was accomplished as follows: First, a chromium film was deposited on the optical quality glass substrate. A layer of photoresist was deposited on the chromium film. The photoresist was exposed through a mask and developed. This leaves photoresist only where where the waveguide will not be. The chromium film is next etched into the waveguide pattern. Thus, this chromium film served as a mask. Next, a silver metal film was vacuum deposited on the chromium mask and the bare glass waveguide pattern. The silver film was covered with a second vacuum deposited chromium film. Another chromium film was deposited on the bottom of the glass substrate. Electrodes were attached to the top and bottom chromium films. A voltage of about 10 Volts was applied between the chromium film electrodes while the substrate was heated to 350 ^{o}C for 2 hours. During this process the silver diffuses into the glass forming a channel with a higher index of refraction than the surrounding glass. Silver diffuses much more readily than chromium in glass. This process is known as electric field assisted diffusion. After the diffusion process is completed the chromium films were etched off. This process results in a planar waveguide.

Ridge waveguides are fabricated by first fabricating a planar waveguide. Photoresist is, again, deposited on the glass substrate. The photoresist is exposed through the same mask and developed. The mask is aligned with the waveguide using alignment marks. This leaves photoresist on the waveguide. The glass substrate is etched using HCl in H_2O in a 1 to 4 ratio for one hour. This is followed by etching in a 1 to 15 HCl solution in H_2O

for 5 minutes to smooth the waveguide edges. Lastly the remaining photoresist is removed. This process leaves an elevated ridge waveguide on the glass substrate. Ridge waveguides can guide light around much smaller radii of curvature than planar guide can.

d) Waveguide measurements

We have measured the performance of the waveguides. Waveguides in the shape of a "Y" were fabricated. Both ends of the substrates were polished to provide flat smooth input and output surfaces. A single mode fiber was attached to the single leg of the "Y". The fiber was used as the light input from a HeNe laser. The other end of the substrate was observed with a CCD camera connected to a computer. The computer generated three-dimensional plots of the light intensity as a function of position. The pattern produced by a ridge waveguide is shown in Fig. 3 and the output from a planar waveguide is shown in Fig. 4. We note that in both cases distinct, approximately, Gaussian beam profiles are observed.

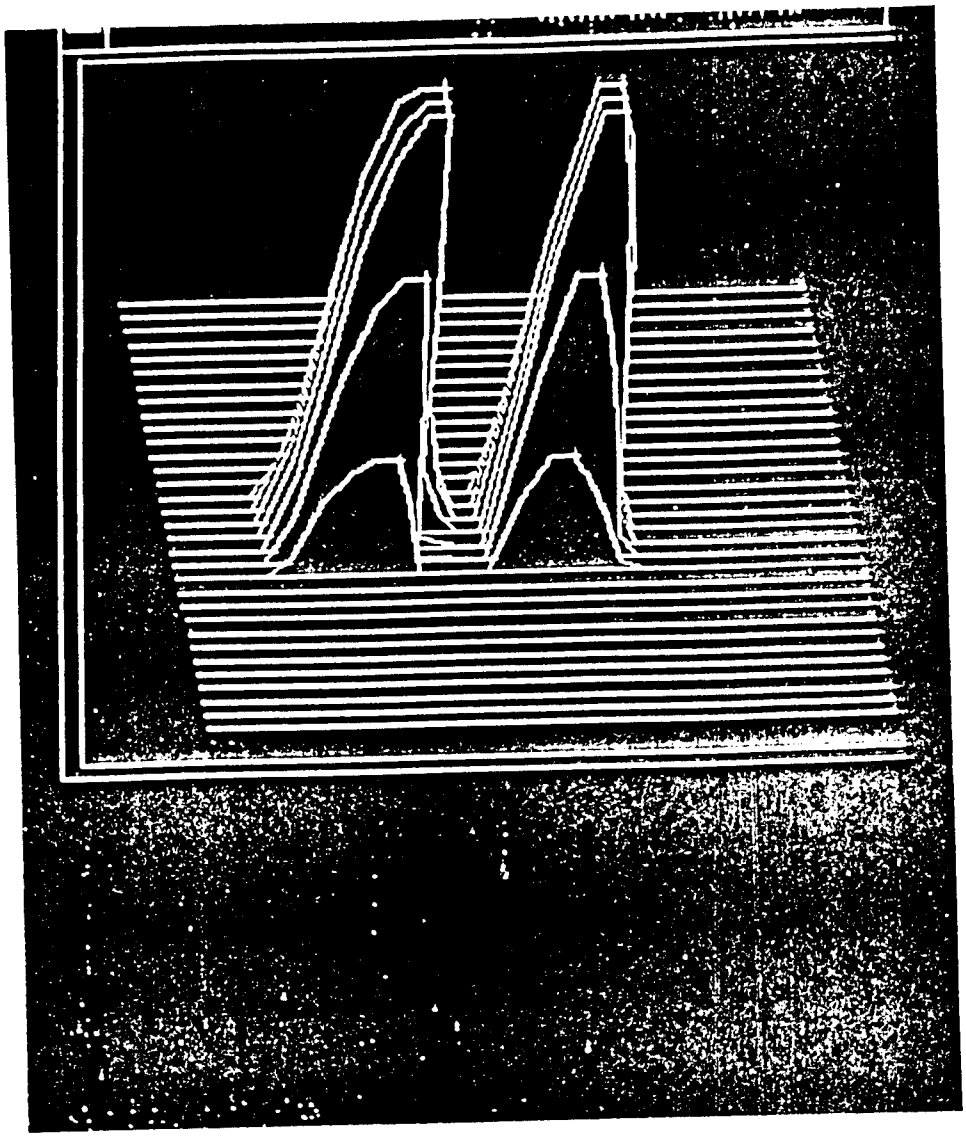


Fig. 3. Output light intensity profile of "Y" shaped planar waveguide.

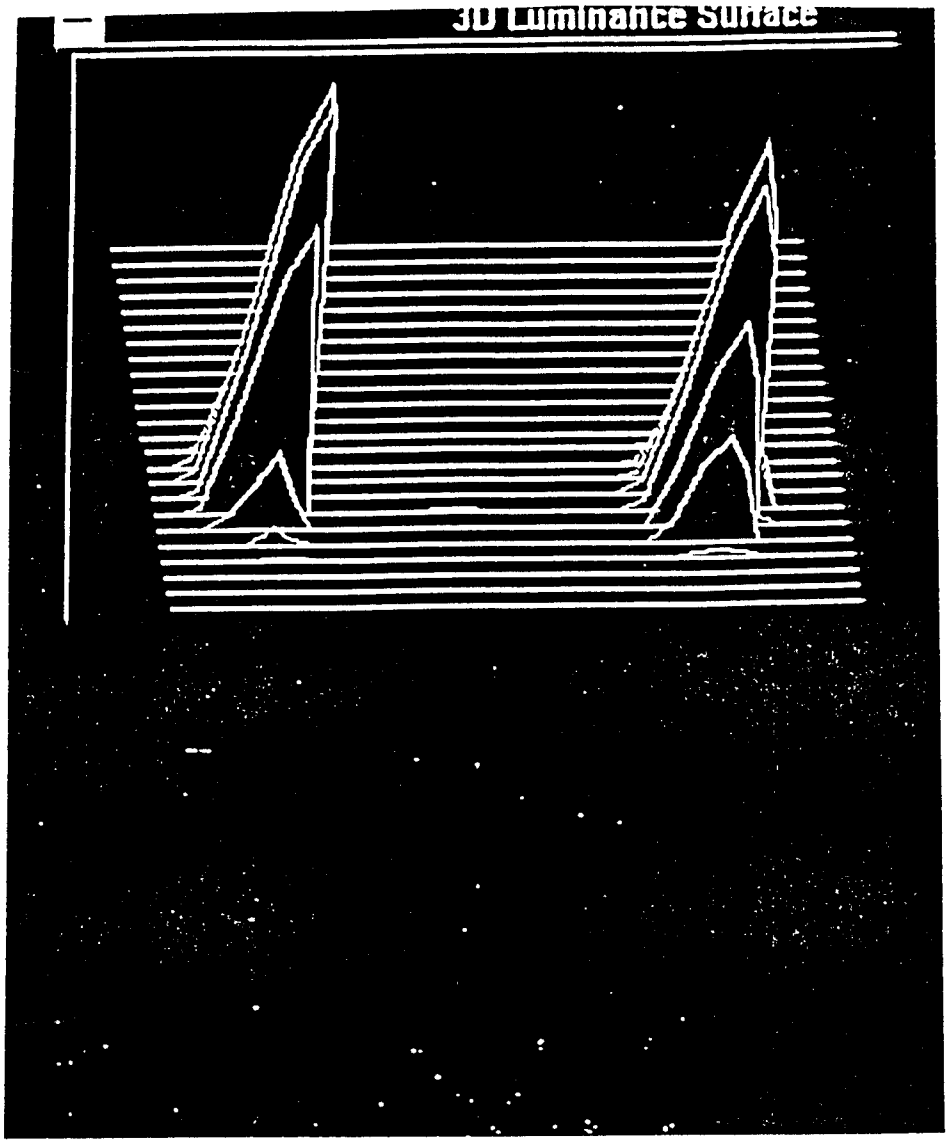


Fig. 4. Output light intensity profile of "Y" shaped ridge waveguide.

2 TUNNELING AND GATE

We have fabricated and tested several photon assisted tunneling diodes. The devices consisted of a 0.5 mm wide aluminum strip vacuum deposited on a glass substrate. The aluminum strip was deposited through a mask. The aluminum film was oxidized in ambient air. SiO insulating edge protection layers were deposited through a mask onto the edges of the aluminum strip. Thin, 80 Å thick, strips of aluminum were vacuum deposited across the oxidized aluminum strip with edge protectors. These top aluminum contacts are thin enough to be transparent to light. Copper contacts were deposited on the thin aluminum strips on either side of the oxidized aluminum strip, see Fig. 5. The edge protection prevents shorting

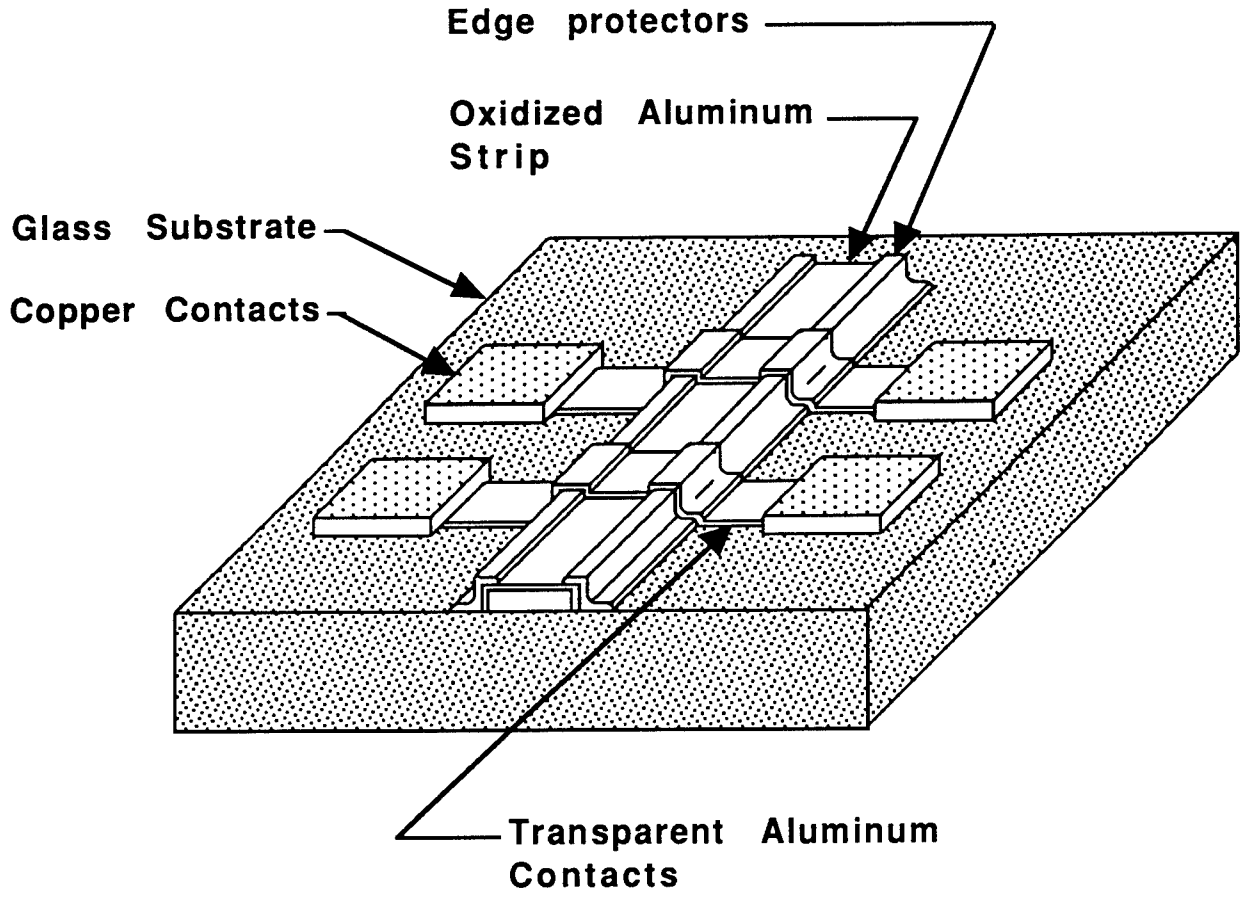


Fig. 5. Tunneling diodes.

at the edges of the bottom aluminum strip.

We measured the tunneling current. All tunneling currents measured were quite low. The oxide layers on all devices fabricated were much to thick. However, we were able to observe an increase in the tunneling current with illumination from a HeNe laser. The signal obtained were rather small. The reason for this is that 98% of the light is reflected and only 2% is absorbed by the device and that the oxide layer was very thick. The light absorption can greatly be increased by using an anti-reflection coating. Since the device is made out of a metal a standard anti-reflection coating could not be used. However, an etalon consisting of a thin transparent film polystirene layer coated with a 50 Å thick aluminum layer was tried. Light absorption is increased by a factor of about ten by the use of the etalon. However, the thick oxide layer still limited the tunnelling current.

Since the late 1950s various attempts have been made to fabricate **Tunneling Transistors**. This device consists of a thick metal emitter, a 30 to 50 Å thick insulating oxide layer, a thin (20 to 100 Å thick) metal base, another 30 to 50 Å thick insulating oxide layer, and a metal collector electrode. This is a **M**etal, **O**xide, **M**etal, **O**xide, **M**etal (MOMOM) device. Such a device could be very fast since the quantum mechanical tunneling process is very fast. However, to date no successful MOMOM transistor has been build. The reason these device did not work is that if the metal film that form the base is thick enough to conduct horizontally it will be to thick for electrons to penetrate the film from insulator to insulator. The electrons are scattered in the metal. It is possible to make very thin metal base layers. However, such layers don't conduct horizontally. We propose to build MOMOM devices with exceedingly thin metal base layers only 10 to 20 Å thick. We plan to access these layers optically through a thin (100 Å thick) collector electrode rather than trying to make electrical contact to them. We are proposing to try to build a **Photon Assisted Tunneling Transistor** (PATT). Two of these devices can be connected in series to form a **Photon Assisted Tunneling AND Gate**.

DEFECT PROPERTIES OF PHOTOREFRACTIVE AND NON-PHOTOREFRACTIVE BSO

Wallace B. Leigh, PhD
Assistant Professor
Division of Electrical Engineering

Alfred University
26 N Main Street
Alfred, NY 14802

Final report for:
Research Initiation Program
Rome Laboratories
Hanscom AFB

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, DC

December, 1992

DEFECT PROPERTIES OF PHOTOREFRACTIVE AND NON-PHOTOREFRACTIVE BSO

Wallace B. Leigh, PhD
Assistant Professor
Division of Electrical Engineering
Alfred University

Abstract

Defect properties of various types of single-crystal $Bi_{12}SiO_{20}$ (BSO) were investigated to determine photorefracting and defect properties of these crystals. Crystals investigated included photorefracting and non-photorefracting BSO grown by Czochralski (Cz) technique, BSO grown by Hydrothermal (HTL) technique, and Cz BSO which had been irradiated by gamma radiation. Above room-temperature optical absorption measurements indicate an anomalous high shift of the absorption edge with temperature for the HTL material, a shift attributed to changes in the localized environment of the oxygen-residing lone pair electrons. Fit of absorption data to the Urbach rule indicates the presence of different impurity bands in the photorefracting and non-photorefracting Cz material; impurity bands which are not present in the HTL material. Furthermore it is demonstrated that fit of the absorption band tails to the Urbach rule may be used as a method of characterizing impurity content in BSO. The irradiated Cz material had been investigated in an attempt to improve the photorefracting efficiencies and response of BSO using red and infra-red pump and probe beams rather than blue and red. Clear samples of Cz and HTL BSO were observed to be photorefracting using 632 nm pump beams, with efficiencies of only $\sim 10^{-5}$ and response times ranging from seconds for HTL material to several minutes for Cz samples. Gratings formed in this manner were observed to fix in the Cz samples, causing the long erase times. Fixing was not observed in the HTL material. These photorefracting properties were not observed to change in irradiated BSO and it is shown that the small percentage change in absorption due to irradiation cannot be expected to improve photorefracting properties in the long-wavelength regions.

DEFECT PROPERTIES OF PHOTOREFRACTIVE AND NON-PHOTOREFRACTIVE BSO

Wallace B. Leigh

Introduction

Photorefraction is a method of writing volume holograms which are dynamic in nature; that is, the volume gratings eventually decay following cessation of exposure. Photorefraction is observed in cubic insulators such as $Bi_{12}SiO_{20}$ (BSO) and $Bi_{12}GeO_{20}$ (BGO) ; in ferroelectrics such as $BaTiO_3$; and in cubic semi-insulators such as InP:Fe, and GaAs:Cr. The applications of photorefraction in these materials include many promising new advances in image processing and optical information processing. Some of these applications are already bearing fruit with new products.

Many of the photorefractive crystals mentioned are being grown at Rome Laboratories, Hanscom AFB and their photorefractive properties are being investigated there as well. Introductory studies of defects in BSO and their relation to photorefraction were begun by this author at Rome under an Air Force Summer Faculty Research program. This report gives continuing progress on research initiated in that study, on properties of different types of photorefracting BSO. The purpose of this study was twofold: The first is a continuation of the original study set out to investigate the properties of BSO grown by the Cz and HTL methods at Rome. The second, in an attempt to lower the pump and probe beam energies to red and infrared regions, the photorefractive properties of Cz, HTL, and irradiated Cz BSO were investigated at these longer wavelengths.

Cz and HTL BSO

Several samples of Cz and HTL BSO were investigated to help identify why HTL BSO is transparent in the visible and, as undoped, has different photorefracting properties compared to Cz BSO. Used in this study were two samples of photorefracting Cz BSO grown from 4-9s purity sources. These samples were observed to be photorefracting using 455nm radiation as a pumping source and 632 nm radiation as a probe, with efficiencies of 10^{-4} and response times of several milliseconds. In this report, we will define "short wavelength photorefraction" as photorefraction experiments conducted using blue pump and red probe beams. "Long wavelength photorefraction" will be defined as that measured under red pump beams and infrared probe beams, and results of these investigations will be addressed later in the report. Undoped HTL material is not photorefracting under short wavelength conditions. Also under study were two samples of low purity Cz BSO, grown with only 3-9s purity sources. This material has not been observed as photorefracting under any conditions, however neither of the 3-9s purity samples were sufficiently clear for proper photorefractive measurements. Finally, two samples of HTL BSO were used , one of which is photorefracting at long wavelengths, the other, not clear for photorefracting measurements.

An earlier thermally stimulated current (TSC) study[1] of these samples was conducted and results are shown in Fig. 1. This data shows the concentration of traps of activation energy < 0.7 eV in HTL material is approximately a factor of 10^3 smaller than that of Cz BSO. At least five different defects were identified in the TSC measurements.

While a large concentration of shallow traps were always found in the Cz material, the identity of the dominant shallow trap is not always the same. In that same study, high temperature electrical conductivity data indicated a thermally activated carrier excitation of energy ~1.3 eV. The Arrhenius plot of this data is shown in Fig (2). This thermally activated process was observed for both Cz and HTL material, although trap identity and concentration measurements were not available.

Temperature Dependent Absorption

In an attempt to identify absorption/conduction mechanisms in the different types of BSO, temperature dependent optical absorption studies were undertaken. While the absorption edge for the transparent HTL material is believed, but not, as yet confirmed, to be due to band-to-band transitions, that of the Cz material is unknown. Absorption studies as a function of temperature help identify different absorption and conduction processes for the different material types. Another motivation for temperature dependent absorption studies is the odd behavior of HTL material at high temperatures. It has been noted that when transparent HTL material is heated to moderate temperatures, it tends to color, and when it cools it becomes transparent again. Cz material does not produce this phenomenon, and instead remains the same yellow color through heating.

To analyze this phenomenon as well as the general conduction mechanisms of the three types of samples, transmission/absorption data was measured as a function of temperature. A standard transmission apparatus is used for this purpose, with the sample mounted on a copper block containing a heating element. The block is milled to fit the sample in a fashion to help minimize convection currents. This apparatus can

be used to measure transmission data for temperatures ranging from RT to 500^{o}C. A quartz-halogen light source is used, along with a 1/4 meter monochrometer with a bandpass of 0.4 nm. A silicon photodetector and lock-in detections system is used, and the system is computer controlled.

For the absorption studies, three different types of samples of BSO were used: Three samples were of high purity BSO, two samples of low purity BSO and one HTL BSO sample. Absorption data has been determined from the relation

$$T = \frac{(1-R)^2 e^{-\alpha t}}{1-R^2 e^{-2\alpha t}} \qquad (1)$$

for reflection from two surfaces, and plotted on a semilog plot for the three sample types. In Eq. (1), t is the sample thickness and R is the reflectivity. The reflectivity is determined by measuring the transmission in the mid-ir and assuming no absorption at this wavelength.

Figure 3 shows the spectral absorption curves for different temperatures for three samples. BSO 9 (fig(a)) is 3-9s pure Cz grown, BSO110H (fig(b)) is 4-9s pure Cz grown and BS110A(fig(c)) is HTL grown material. All values of α are in m^{-1}. The change of color in HTL BSO at high temperatures is shown to be caused by the shift in the absorption edge with temperature. Heating up to ~400^{o}C, the sample turns from yellow to green as the absorption edge moves into the visible region of the spectrum. In Figures (a) and (b) it is observed that the shift with temperature is not as great for Cz BSO. The shift for HTL BSO is observed to be linear with temperature , and is measured to be ~2X10^{-3} eV/^{o}C. The absorption edge was determined from the intercept of transmission curves with the x axis. For Cz material the shift is much

smaller, $\sim 10^{-4}$ for the Cz BSO samples. This can be compared to $\sim 10^{-4}$ eV/^{0}C for most materials[p95l]. Furthermore, for Cz BSO the shift was not linear although no concise relationship could be determined from these samples.

While it is difficult to determine the nature of the absorption shift for HTL BSO, it is likely that the lone pair electrons of oxygen play a role in determining the band structure. It has been found that in amorphous semiconductors containing group VI elements, anomalous pressure and temperature variations of the absorption edge, refractive index and other optical properties have been explained by a change in the local environment of the group VI element which contains lone-pair electrons[3,4]. In Cz material, the absorption is mechanism is different, characterized by the impurity band of that material.

It is noted that the "tails", which is the area defined as the band-edge (the high-energy regions of the curves in Fig. 3) can be modeled as an exponential function. This is the so-called "Urbach rule" and the data of this region are fit to the equation

$$\alpha(E,T) = \alpha_0 \, e^{-\sigma(E_0-E)/kT} \qquad (2)$$

Where $\alpha(E,T)$ is the absorption coefficient and E is the photon energy. The main explanation for the dependence given by Eq(2) is that the energy of the excited electron states varies due to lattice deformation by lattice vibrations. Any increase in the deformation has an effect on the measured value of σ. In this way, any perturbation, such as disorder or the presence of impurity tails, can be determined in a somewhat qualitative fashion from the dependence of σ and E_0 with temperature.

Figure 4 shows values of σ with temperature for the three sample

types. Again, BSO 9 is 3-9s pure Cz grown, BSO110H is 4-9s pure Cz grown and BS110A is HTL grown material. There is a linear dependence of σ with temperature for all three sample types, and the slope of this dependence falls with increasing impurity type.

The effect of impurity concentration on absorption tails has been considered in Ref[5-7]. For crystals containing impurity bands, and for a certain concentration range of impurities, there is an effect on the variation of σ with temperature. Depending on the shift of the absorption edge with impurity concentration, the effects of impurities in a band is to either decrease or increase the linear dependence of σ with temperature. Furthermore, it was shown in reference [5] that this shift is linear with impurity concentration. These effects are due to phonon-assisted absorption processes. As the impurity band represents an added deformation potential, the change in sigma can be noted.

Obviously, from Fig. 4 this condition holds for BSO. The HTL BSO has been shown by the TSC measurements to be of less impurity content compared to Cz BSO. The relation of σ with temperature is linear in the range of BSO being grown at Hanscom. In the less pure Cz material, the slope of the line is less, correlating with an added impurity concentration. A more detailed study may prove a correlation between impurity content and the σ vs. T variation. This technique may prove a rapid means for characterization of BSO.

It appears from Fig. 3 that there are two absorption processes, one at longer wavelengths and one at shorter ones, separated by a "kink" in the absorption curve. The reason for the "kink" in the absorption curve is as yet unknown. It is interesting to observe that in the high-purity BSO at ~400^{o}C, one absorption process disappears in favor of the other

and the "kink" disappears.

Photorefraction at longer wavelengths

Most experiments used to characterize the photorefractive properties of BSO are performed with the use of two lasers: a "pump" laser source and a "probe" laser source. Photorefraction is most efficient when a pump beam of ~500nm (from an Argon ion source) is used with a 633 nm probe beam of a Helium-Neon laser. Nevertheless, photorefraction does occur using pump and probe beams of longer wavelengths[8,9] despite the small (~0.4cm^{-1}) absorption coefficient at these frequencies. In the literature it was postulated that some of the more shallow levels were responsible for the photorefractive properties at these wavelengths.

Two samples of Cz BSO were irradiated with 50Mrad of gamma-ray radiation in order to analyze radiation-created defects and to move the absorption edge to longer wavelengths. Fig. 5 shows the effect of irradiation on the absorption edge for Cz BSO. By moving the absorption edge closer to 633nm, a He-Ne pump beam used to write photorefractive gratings would have higher efficiencies than non-irradiated BSO.

The photorefractive setup was a typical non-degenerate 4 beam shown in Fig. 6. A 10mW 633nm polarized HeNe laser was used as a pump, and a 5mW 806nm laser diode was used as a probe. A silicon detector and lock-in detection system was used to detect the diffracted probe beam. The set-up allows an incident pump angle of 9.7^{o} which corresponds to the formation of holographic gratings of ~2 μm spacing. The incident pump irradiance was 500 mW/cm^2. Chopping frequency

was 250 Hz.

Initially observed phenomena with the irradiated material included the tendency for the damaged area to cure. Indeed, after ~3 weeks time the irradiated material had returned to its original unirradiated condition. Furthermore, curing seemed to be accelerated in the presence of the 633nm light. No additional defects were observed in TSC measurements of the irradiated material, and the only major difference observed between irradiated and unirradiated BSO is in the transmission curves of Fig. 5.

All BSO material tested was observed to be photorefractive using the set-up in Fig. 6, with the exception of the low-purity Cz material. This includes the irradiated and non-irradiated Cz material as well as HTL material. The major limitation in long-wavelength photorefraction is in obtaining a material which is totally free of occlusions, cracks or other physical defects from which the infrared light may scatter. The efficiencies of photorefraction at these wavelengths tend to be quite low, on the order of 10^{-5}, which is on the order of the limits of the detection for the system of Fig. 6.

As the laser diode has a higher irradiance than the pump beam, there is a tendency for the probe to erase the grating once it is formed. The order of the experiment then, is as follows: First the pump beams are allowed to form the grating with the probe beam off. The probe beam is next turned on and efficiencies are then recorded. After equilibration is maintained, one of the pump beams is turned off and the grating is allowed to decay. This is then recorded by measuring the decay in the diffracted probe beam. Care was taken to eliminate reflections and scattering of the IR light. Several measurements were

taken at different spots on the sample at different times.

Figure 7 shows the recorded diffracted beam from a Cz sample where red beams are used as the pumping source. Initial rise is quite swift, followed by a slow decline as the probe erases the grating. After several minutes, one of the red beams is blocked and the signal from the diffracted probe beam is allowed to decay. In Cz material, this decay can vary anywhere from several minutes to several tens of minutes, varying from sample to sample and from different probed areas of the sample. This long delay time can be explained as a "fixing" of the holograms by compensated charges. These charges make up the so-called "complementary" grating which exists long after the primary grating has dispersed. Fixing can occur from charge compensation by free carriers[9,10] or by drift of ions[11] under the influence of the space charge electric field set-up by the primary grating. In short-wavelength photorefraction efficiencies of the complimentary gratings can be much lower than that of the primary, but since primary efficiencies are low under long wavelength conditions, it is more apparent in our system.

An interesting development of photorefraction is that a transparent HTL sample, which is not photorefracting at short wavelengths, is potorefracting at longer wavelengths. Furthermore, the fixing of the primary by a complimentary grating is not observed in HTL material. Thus erasure of the the grating in the HTL sample occurs at a much faster response time than that of the Cz material. Figure 8 shows the initial rise of the diffracted beam followed by the erasure of the beam from HTL BSO when one red pump beam is blocked. The relative efficiency of photorefraction using this system is about the same for Cz

and HTL material.

Although it appears in the figure that not all the grating has erased, several other measurements on this sample could not confirm that a complimentary grating ever existed, and usually the signal returns to zero once the primary grating is erased. This represents a change of several orders of magnitude in the time to erase the grating as compared to Cz material.

Since the defect population as determined by TSC is very low for levels of thermal activation energy less than 0.7 eV, it is apparent that a trap level deeper than this must be involved in the HTL photorefraction illustrated in Fig. 8. Figure 2 shows high temperature electrical conductivity data which indicates a thermal activated carrier excitation of energy ~1.3 eV. This thermally activated process is observed for both Cz and HTL material, and could very well be the responsible absorption process for photorefraction at long wavelengths.

As for the absence of thermal fixing in HTL material, either the mechanism of photorefraction is different in HTL samples or the ions responsible for the fixing of gratings in the Cz material are of not of the same concentration or diffusivity in HTL material.

Conclusions

Spectral dependence of absorption at different temperatures has proven that hydrothermal-grown BSO has a large shift of the absorption edge with temperature. This shift is quite possibly characterizing the change in bandgap with temperature. The magnitude of the shift is caused by changes in the localized environment of the lone pair electrons residing on oxygen. Fit of the absorption data to Urbach's rule has shown a way to characterize impurity content in Cz and HTL BSO. Photorefraction at long wavelengths in both Cz and HTL BSO has been demonstrated. Long-wavelength photorefraction may be characterized by a fixing of the holographic gratings in Cz material, which is not observed in HTL material.

Bibleography

1. W. B. Leigh, *Final Report, Air Force Summer Faculty Research Program, 1991.*

2. T. S. Moss, Optical Properties of Semi-conductors. Academic, NY (1959) p 43.

3. S. Onari, T. Inokuma, H. Kataura, T. Arai, *Phys. Rev.* B, 35 4373 (1987)

4. M. Kastner, *Phys Rev* B 7, 5237 (1973).

5. A. F. Lubchenko, *Phys. Stat. Sol.* 27, K73 (1968).

6. D. E. McCumber, *Phys Rev.* 135, 1676 (1964).

7. H. Mahr, *Phys Rev* 132, 1880 (1963).

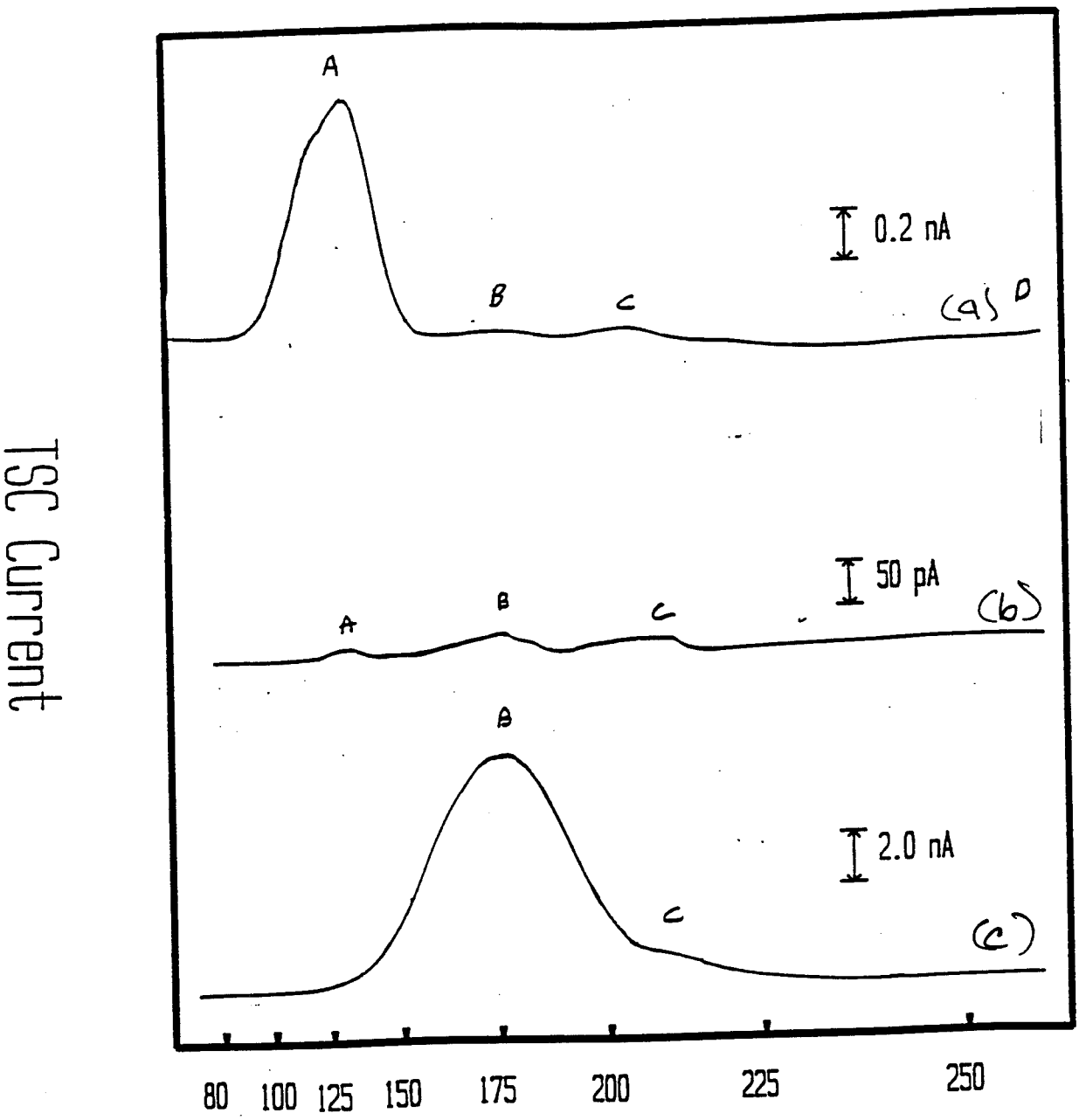
8. M. A. Powell, C. R. Petts, *Optics Letters* 11, 36 (1986).

9. M. C. Bashaw, T.-P. Ma, R. C. Barker, S. Mroczkowski, and R. R. Dube, *Phys Rev B* 42, 5641 (1990).

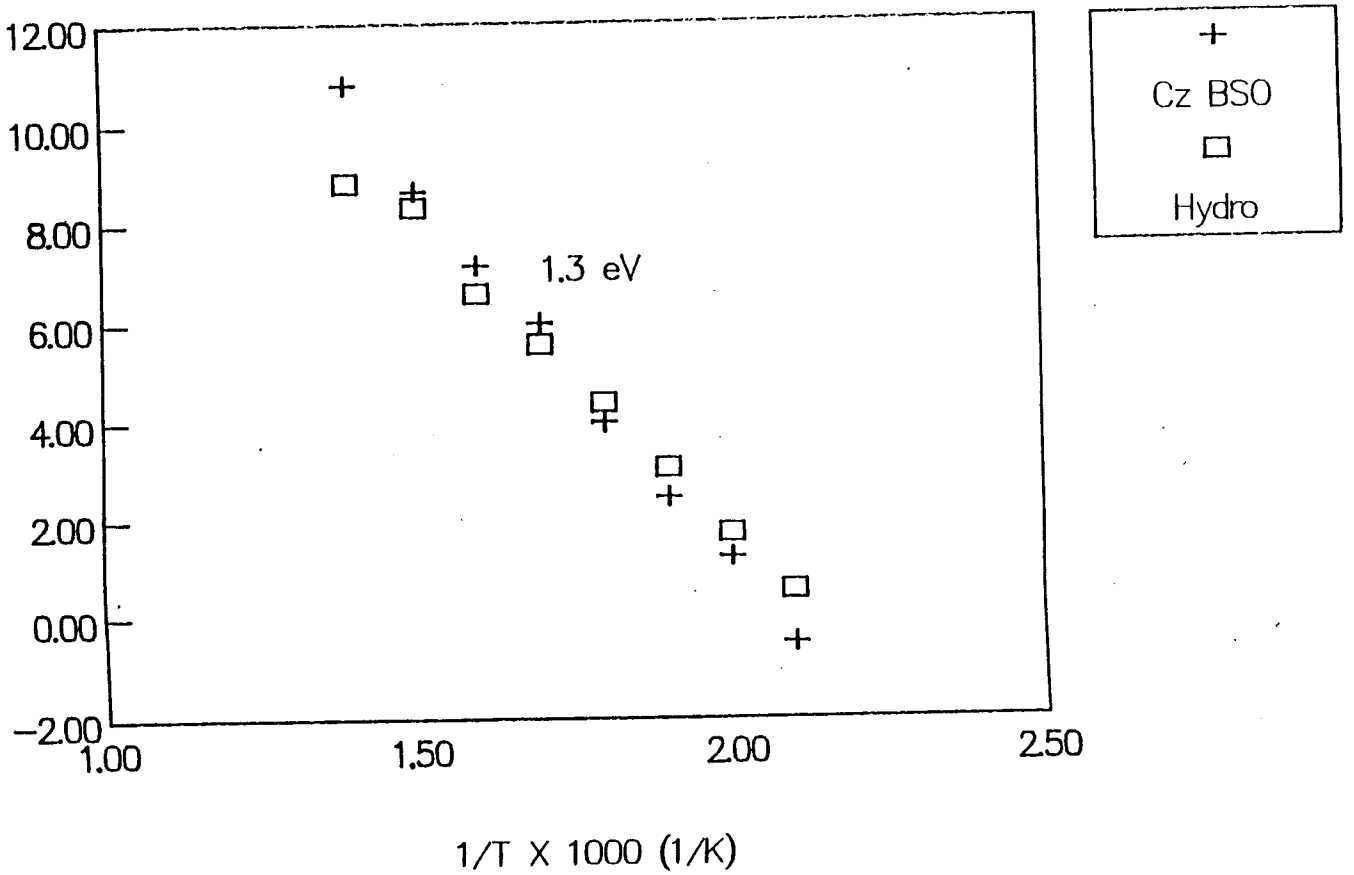
10. J. P. Herriau and J. P. Huignard, *Appl. Phys. Lett* 49, 1140 (1986).

11. L. Arizmendi, *J. Appl. Phys.* 65, 423 (1989).

List of Figures

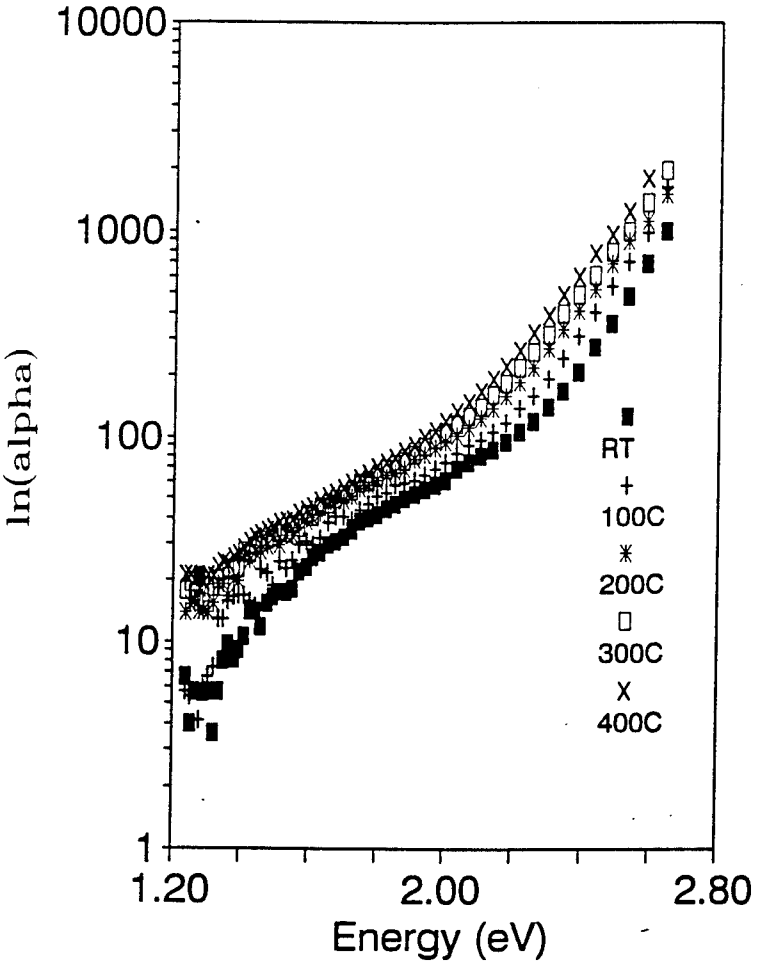

A
B
C
0.2 nA
(a)
A
B
C
50 pA
(b)
B
C
2.0 nA
(c)
80
100
125
150
175
200
225
250
TSC Current
Temperature (K)

Leigh,
Fig 1

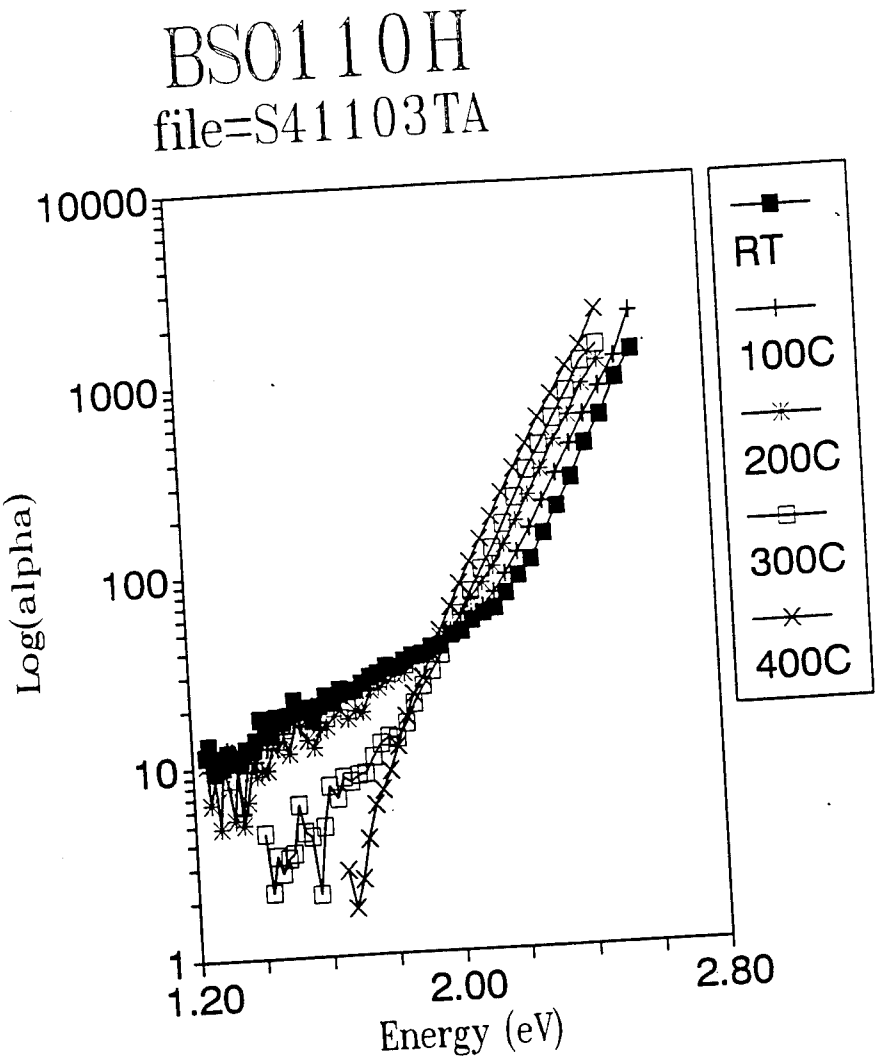

12.00
10.00
8.00
6.00
4.00
2.00
0.00
−2.00
1.00
1.50
2.00
2.50
ln (nA)
1.3 eV
Cz BSO
Hydro
1/T X 1000 (1/K)

Leigh,
Fig 2

BS09

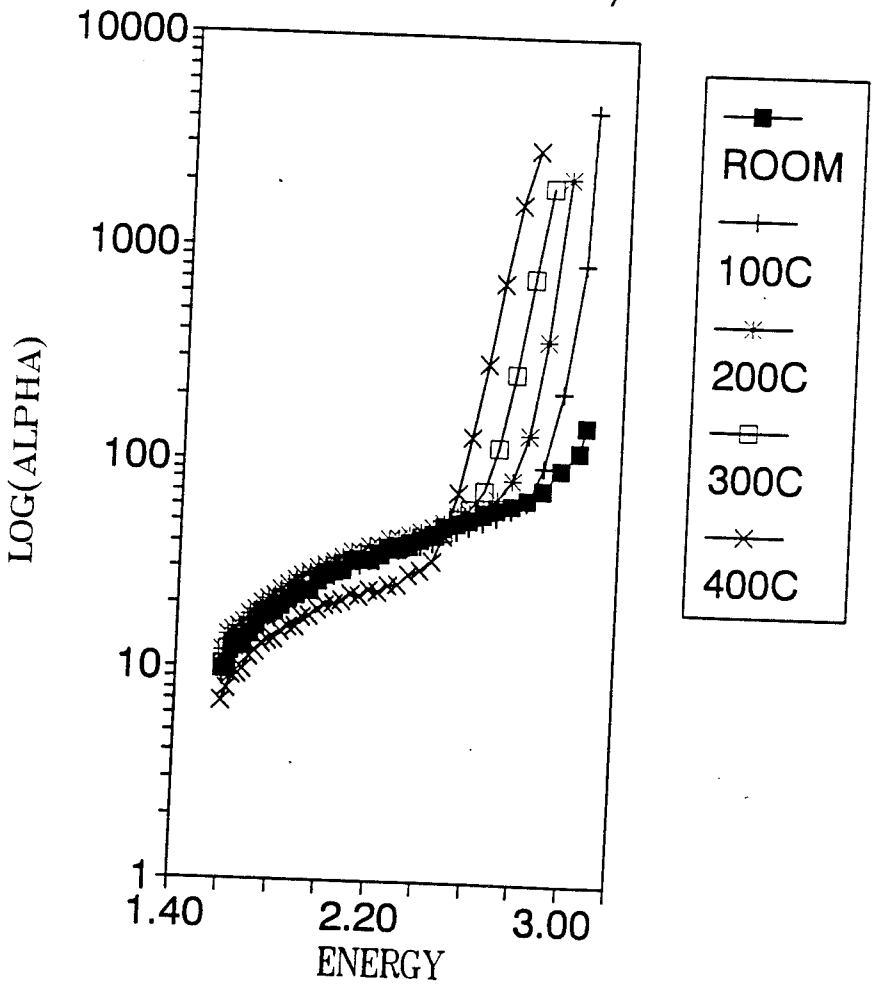


Leigh,
Fig 3(a)

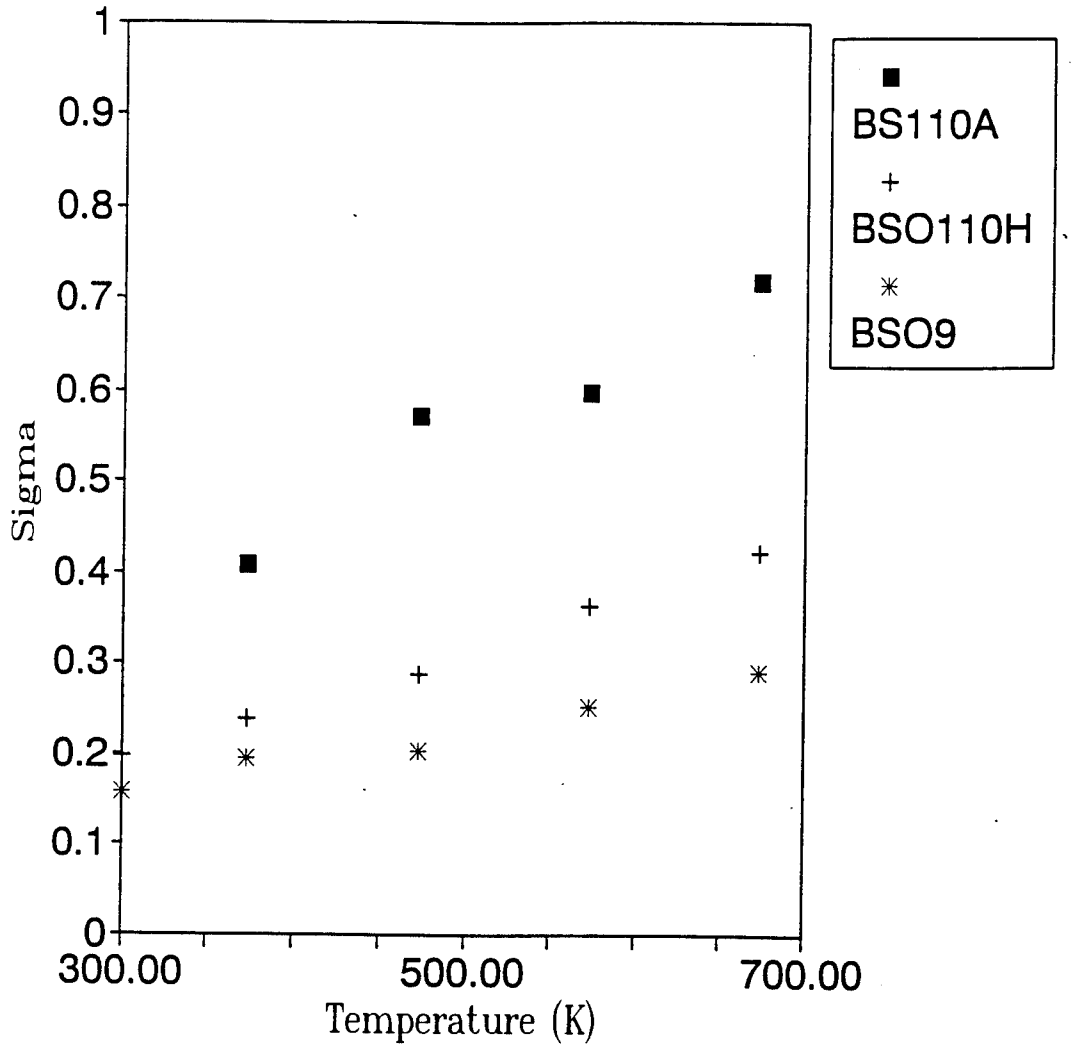

BS0110H
file=S41103TA
10000
1000
100
10
1
Log(alpha)
1.20
2.00
2.80
Energy (eV)
RT
100C
200C
300C
400C

Leigh,
Fig 3(b)

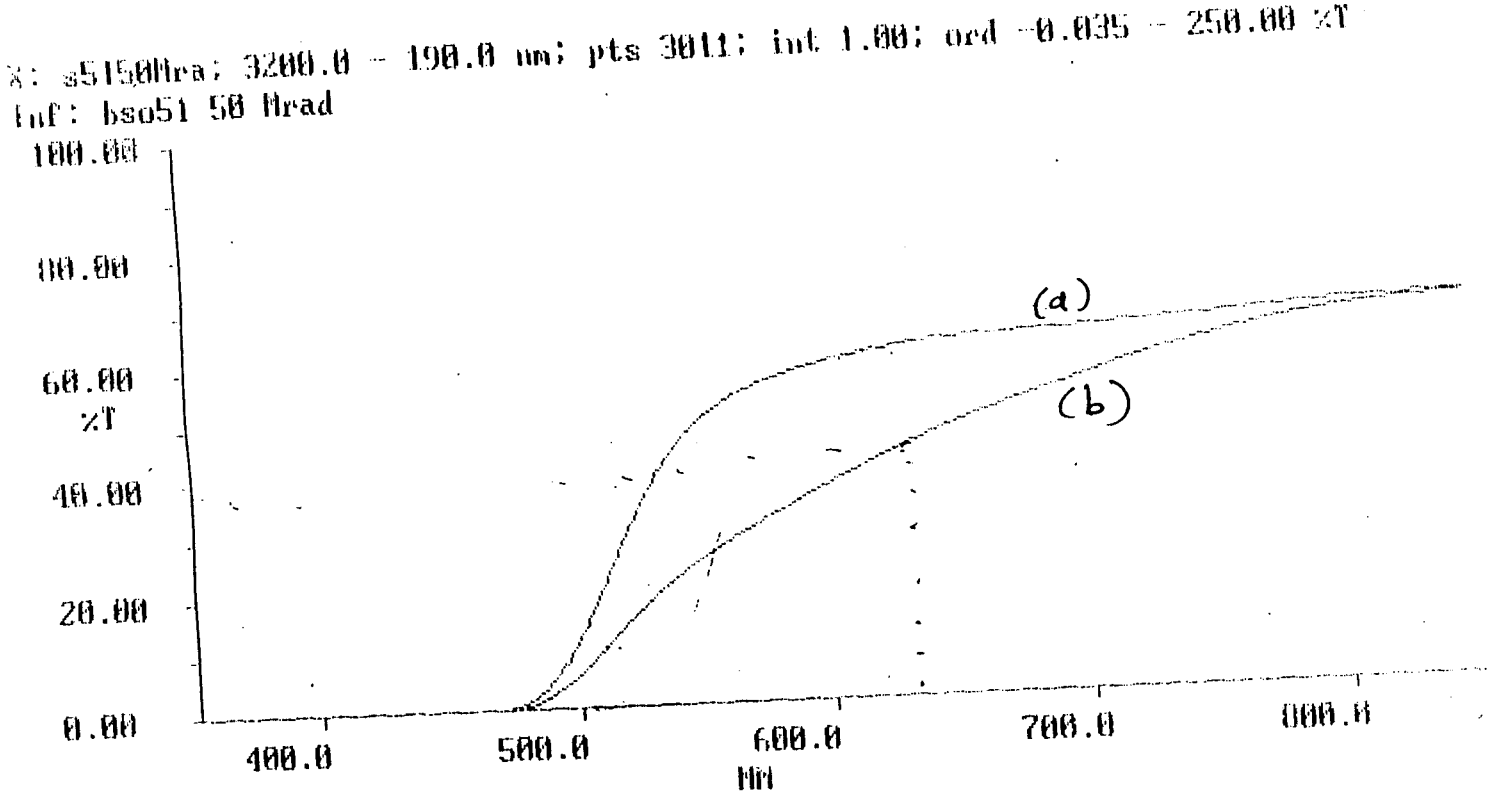
BS110A 10/20



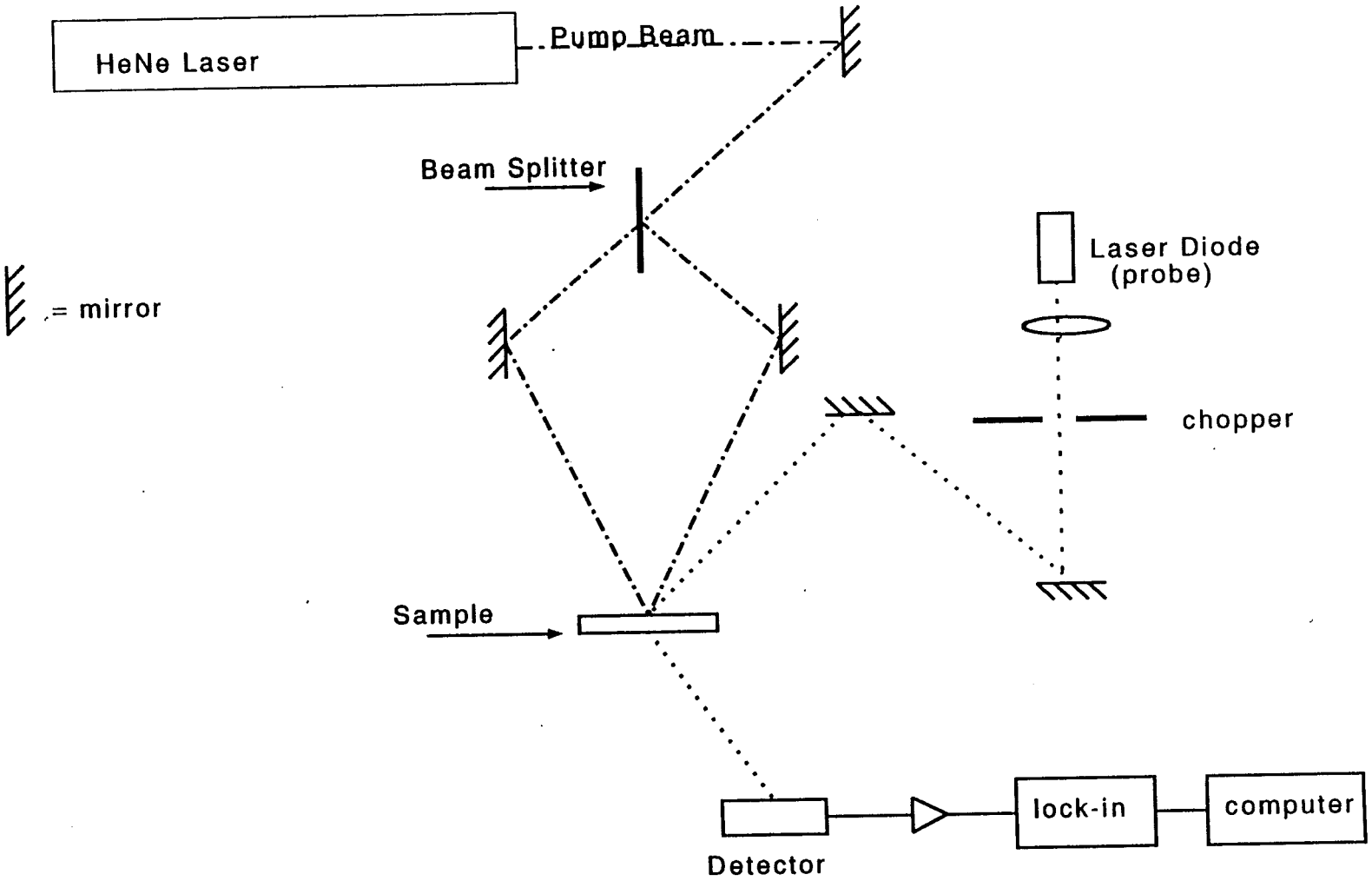
Leigh,
Fig 3(c)


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Sigma
300.00
500.00
700.00
Temperature (K)
BS110A
BSO110H
BSO9

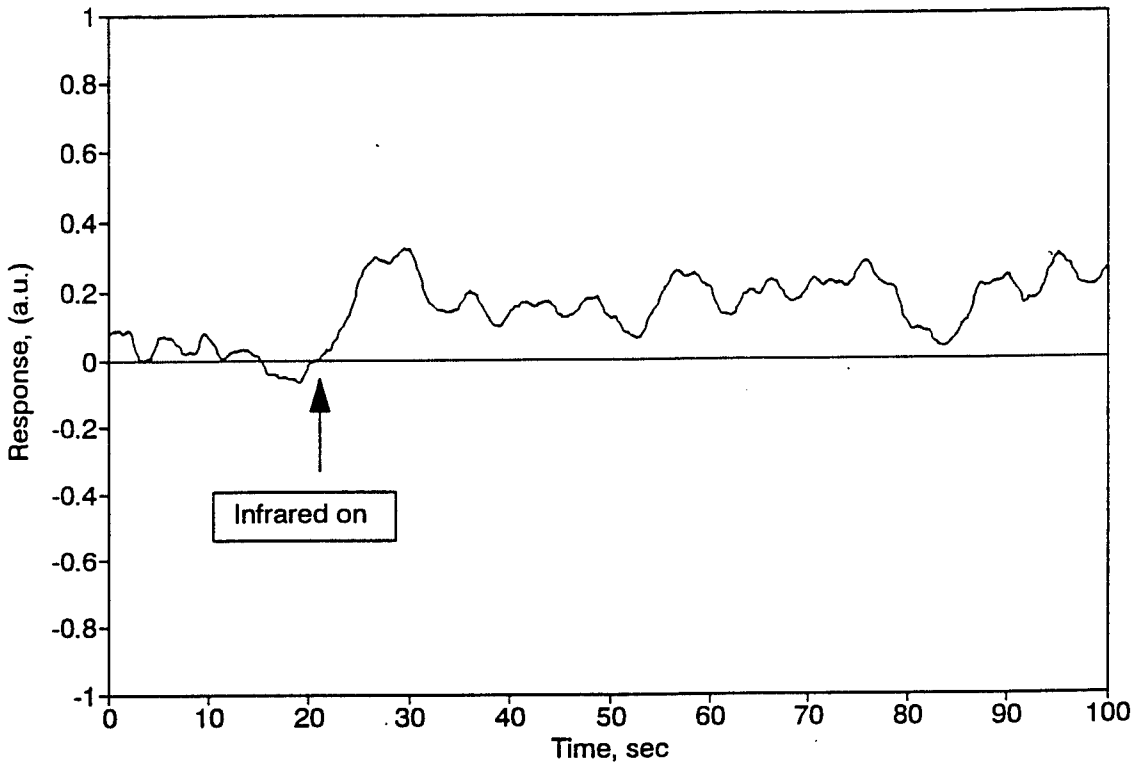
Leigh,
Fig 4


X: s5150Nra; 3200.0 - 190.0 nm; pts 3011; int 1.00; ord -0.035 - 250.00 %T
Inf: bso51 50 Nrad
100.00
80.00
60.00
%T
40.00
20.00
0.00
400.0
500.0
600.0
700.0
800.0
NM
(a)
(b)

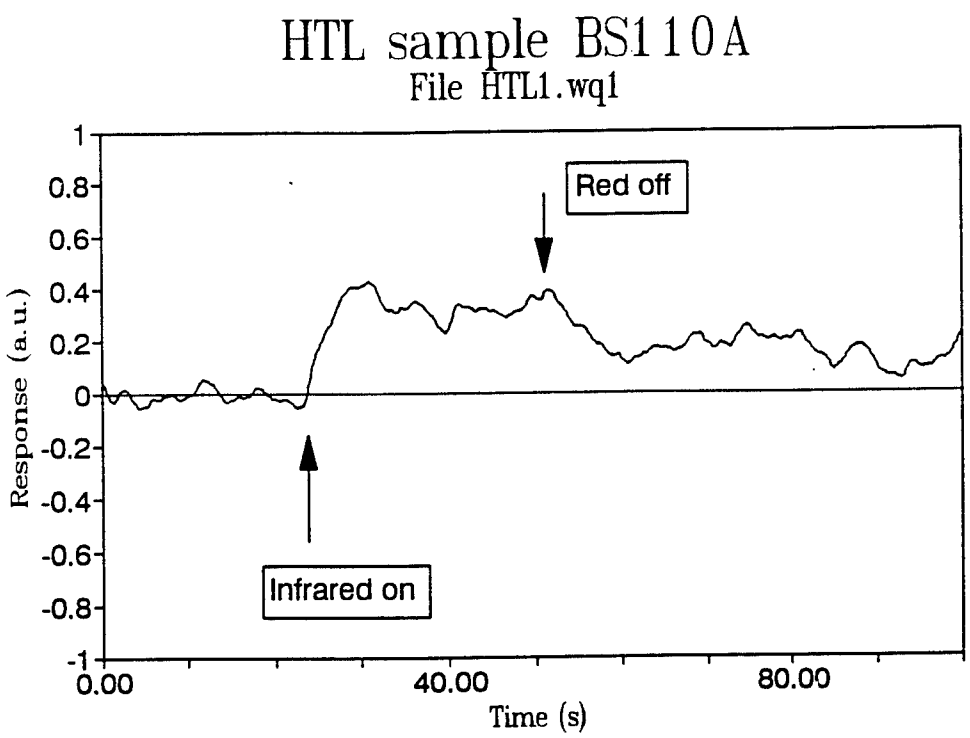
Leigh,
Fig 5


HeNe Laser
Pump Beam
Beam Splitter
= mirror
Laser Diode
(probe)
chopper
Sample
Detector
lock-in
computer

Leigh,
Fig 6


Response, (a.u.)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
10
20
30
40
50
60
70
80
90
100
Time, sec
Infrared on

Leigh,
Fig 7


HTL sample BS110A
File HTL1.wq1
Response (a.u.)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0.00
40.00
80.00
Time (s)
Red off
Infrared on

Leigh 1
Fig 8

AUTOMATIC CONSTRUCTION of a SEMANTIC LEXICON for use in NATURAL LANGUAGE PROCESSING SYSTEMS[1]

Elizabeth D. Liddy, Asst. Prof.
Diana M. Lauterbach, MLS

School of Information Studies
Syracuse University
4-206 Center for Science & Technology
Syracuse, New York 13244-4100

Final Report for Research Initiation Program
Rome Lab

January, 1993

[1]Research reported herein was sponsored by the AFOSR and Rome Lab.

AUTOMATIC CONSTRUCTION of a SEMANTIC LEXICON for use in NATURAL LANGUAGE PROCESSING SYSTEMS

Elizabeth D. Liddy, Asst. Prof.
Diana M. Lauterbach, MLS

School of Information Studies
Syracuse University
4-206 Center for Science & Technology
Syracuse, New York 13244-4100

Abstract

The underlying goal of the proposed research is to contribute substantively to the solution of a major problem facing the development of realistic Natural Language Processing (NLP) systems, namely the construction of reasonably-sized lexicons. The knowledge for use in NLP systems must be acquired and represented in such a manner that its implicit semantics are then available for the necessary inferencing of the host system. The methodology exploited in the research herein reported, provides one plausible alternative. It is called *knowledge extraction* (Liddy, Kwasnik & Myaeng, 1989) and is a means whereby pre-existing textual sources are processed in as automatic a manner as possible with the goal of extracting the knowledge encoded in natural language repositories. This approach is based on the existence of detectable linguistic patterns which predictably reveal the presence of semantic relations and derives from sublanguage theory (Sager et al, 1987) which has shown that within a particular text-type, predictable syntactic and semantic regularities develop over time, and these regularities are used consistently. The meaning to be extracted consists of *concepts* and the *relations* among these concepts. Machine Readable Dictionaries (MRDs) form an excellent and readily available textual source for performing semantic knowledge extraction.

Our efforts have been focused on developing patterns that can be used in algorithms for the automatic recognition of semantic relations in definitions for inclusion in a semantic representation for use by NLP systems. The specific focus of this research effort was to develop means for automatically determining in a verb definition which of a verb's senses was intended by the lexicographer who created the entry. For example, the most frequently used verbs in a dictionary are used in thousands of definitions. However, each verb has several senses and it is important when creating a semantic representation of a definition to know which sense of that verb is being used in defining the verb being represented. For example, the word 'make' is used in approximately eight hundred definitions in our machine-readable dictionary (LDOCE). However, 'make' has several senses (PRODUCE, PERFORM, or CAUSE) and it is important when creating a semantic representation of a definition to know which of the senses of 'make' is being used in defining the verb being represented. We refer to the patterns we developed to recognize specific senses as Relation Revealing Formulae (RRFs). RRFs couple a semantic relation to the particular linguistic features of definitions which reliably indicate the presence of that semantic relation. We analyzed the seventeen most frequently used verbs in LDOCE and created recognition templates which contain RRF which we developed. These RRF will be applied to unanalyzed definitions to determine in each definition, which sense of the defining verb is intended. The RRF are based on 1,912 verb definitions which we analyzed.

AUTOMATIC CONSTRUCTION of a SEMANTIC LEXICON for use in NATURAL LANGUAGE PROCESSING SYSTEMS

Elizabeth D. Liddy
Diana M. Lauterbach

Problem

The underlying goal of the proposed research is to contribute substantively to the solution of a major problem facing the development of Natural Language Processing (NLP) systems, namely the construction of reasonably-sized lexicons. As stated by Velardi et al (1991, p. 157), "automatic or semi-automatic acquisition of the lexicon is a critical factor in determining how widespread the use of natural language processors will be in the next few years". The lack of reasonably sized, semantically rich lexicons is posing one of the major bottlenecks in NLP development (Lenat et al, 1986) and is due to the absence of a method for the efficient acquisition of semantic knowledge of the world.

The knowledge for use in NLP systems must be acquired and represented in such a manner that its implicit semantics are then available for the necessary inferencing of the host system. Knowledge acquisition, carried out by knowledge engineers, is the most commonly used approach to acquiring this knowledge, but it has proven time-consuming and difficult to contain. The methodology exploited in the research herein reported, provides one plausible alternative to this inefficient process. It is called *knowledge extraction* (Kwasnik, Liddy & Myaeng,1989), and is a means whereby pre-existing textual sources are processed in as automatic a manner as possible with the goal of extracting the knowledge encoded in natural language repositories.

This approach is based on the existence of detectable linguistic patterns which predictably reveal the presence of semantic relations and derives from sublanguage theory (Sager et al, 1987) which has shown that within a particular text-type, predictable syntactic and semantic regularities develop over time, and these regularities are used consistently. Sublanguages develop amongst a group of language users who share a common purpose and who must repeatedly produce instances of a particular type of text (Kittredge & Lehrberger, 1982; Grishman & Kittredge, 1986). Since lexicographers are a specialized group working on a common task, the linguistic features of dictionary entries can be analyzed using sublanguage methodologies which have shown themselves successful in other text-types consisting of brief,

fairly standardized utterances (Liddy et al, 1989; Liddy et al, 1991).

The meaning to be extracted consists of *concepts* and the *relations* among these concepts. In the English language, concepts are most frequently expressed in the form of noun phrases while relations among concepts are commonly expressed by either verb phrases or prepositions. The *concepts* present in dictionary definitions are easier to recognize automatically than the semantic *relations* that exist among concepts, because the language used in dictionary definitions only implicitly reveals these relations. For example, a definition may not expressly state "the purpose of X is Y", but our results show that there is a delineable set of phrases such as "X is an instrument for Y-ing" in which the phrase "is an instrument for" reveals the PURPOSE relation between X and Y. Relations define the nature of the interaction, dependency, influence or simply co-occurrence which holds between the entities.

Machine Readable Dictionaries (MRDs) form an excellent and readily available textual source for performing semantic knowledge extraction because: 1) MRDs are culturally validated sources of much of our commonplace knowledge of the world as accepted by native language speakers over many years; 2) MRDs provide raw data which is unbiased by a particular individual's beliefs, views, knowledge or experience, and; 3) MRDs span a wide range of subject areas, thereby providing a shallower, but more complete coverage of the language than would be available in other corpora.

MRDs provide a rich potential for knowledge extraction, although it has proven to be a very difficult task. Even more pessimistically stated by Alshawi (1989), "...providing complete analyses of arbitrary dictionary definitions is beyond the current state of the art in computational linguistics." We have not found the task impossible, but have found it very slow-going. And although the results are quite intellectually pleasing, the automation of the task is not to the point we had originally expected. The size of the task can perhaps be best understood when one realizes that full, well-received research papers have been devoted to simply understanding dictionary phrases of the form "to VERB with NP" (Ravin, 1993).

For our research, we used the 1987 version of Longman's Dictionary of Contemporary English (LDOCE). The 1977 version of this MRD has been the object of a substantive body of work (Alshawi, 1987; Boguraev & Briscoe, 1987; Wilks et al, 1987) which has revealed the

potential richness of the data encoded on it. LDOCE is particularly appropriate for the present work because it was developed to be used as a learner's dictionary by non-native English speakers and, to accommodate this use, is based on a defining vocabulary of approximately 2000 words. Most of the definitions are written using only these terms and any word used in a definition that is not on this list is indicated by a type-font code. This controlled defining vocabulary contributed substantively to the tractability of our research.

Figure 1 provides some examples of how semantic relations are implicitly expressed in LDOCE definitions. The terms following the parentheses exist in relation to the term being defined, in the relation stated in capital letters within the parentheses. The name of the relation within the parentheses has been added following intellectual analysis of the definition. The phrases in each definition presented in bold print are the lexical clues used in the definition which suggest the nature of that particular relationship. For example, in the definition of *diet*, the GOAL relation exists between *diet* and 'become thinner' and is indicated by the phrase **in order to**.

---

depress - **to make** [CAUSE] less active or strong
smooth - **having** [CHARACTERISTIC] an even surface
pitcher - **a** player **who** [AGENT] throws the ball toward the batter
stab - **to** strike forcefully **with** [INSTRUMENT] the point of something sharp
diet - to eat according to a special diet, esp. **in order to** [GOAL] become thinner
shoot - an area of land **where** [LOCATION] animals are shot for sport
nail file - a small **instrument** with a rough surface **for** [PURPOSE] shaping fingernails

Fig. 1: Sample Definitions from LDOCE

Our efforts have been focused on developing patterns that can be used in algorithms for the automatic recognition of semantic relations in definitions for inclusion in a semantic representation for use by NLP systems. The specific focus of this research effort was to develop means for automatically determining in a verb definition which of a verb's senses was intended by the lexicographer who created the entry. For example, the most frequently used verbs in a dictionary are used in thousands of definitions. However, each verb has several senses and it is important when creating a semantic representation of a definition to know which sense of that verb is being used in defining the verb being represented.

We refer to the patterns we developed as Relation Revealing Formulae (RRFs). RRFs couple a semantic relation to the particular linguistic features of definitions which reliably indicate the presence of that semantic relation. The three levels of RRF which we worked with are presented in Figure 2 and each is followed by a sample RRF template of that level.

LEXICAL RRF: Require straightforward lexical string matches wherein the exact word or phrase which has been established as a lexical clue to the relation is matched in the definition. More complex LEXICAL RRF require the recognition of non-adjacent phrase patterns.

RRF: LOCATION = 'where' + < >

SYNTACTIC RRF: Consist of either sequences of words of specified parts-of-speech or combinations of lexical string matches and specified parts-of-speech and takes advantage of the grammar codes available on the LDOCE tape.

RRF: RESULTANT = 'to make' + <adjective>

SEMANTIC RRF: Depend on information about semantic features of words (e.g. animacy) in the formulae being available in the dictionary entry box codes. It is typical for a single Semantic RRF to include lexical and syntactic matching as well as semantic matching.

RRF: EXPERIENCER = 'a'/'an' + <animate noun> + 'who' + <stative verb>

Fig. 2: Three Levels of Relation-Revealing Formulae

Methodology

The determination of the predictable patterns which reveal semantic relations in definitions, entails a labor-intensive, rigorous, intellectual analysis of large sets of definitions for the inductive discernment of semantic relations which are predictably indicated by the three levels of linguistic regularities described above.

For this research effort, we focused our attention on verbs. Verbs were chosen for several reasons. Firstly, our research at Rome Lab last summer began the current effort and we had chosen verbs for that effort, partially because Michael McHale, a Rome Lab researcher was developing a principle-based parser (McHale, 1993) which requires knowledge of the semantic

role of nouns for various verbs. A more basic reason is the fact that verbs are central to many NLP analysis schemes, such as case-frame analysis. Also, much of the early work on better representation of dictionary definitions had focused on nouns and we wished to extend this work into a new grammatical category. A frequency count of the individual verbs which occur in the verb definitions in LDOCE determined the set of verbs on which we would focus our attention. We analyzed the seventeen verbs with the highest frequency of occurrence in LDOCE definitions. They are: **be**, **become**, **bring**, **cause**, **do**, **force**, **form**, **give**, **go**, **have**, **make**, **move**, **place**, **put**, **show**, **take**, and **use**. Rather than analyzing all the occurrences of each of these verbs, we looked at just half of the definitions, reserving the other half as a database against which we will next test the RRF developed during this project.

We developed RRF which could distinguish the various senses of each verb. The algorithms require the RRF applicator to consult the online LDOCE file to determine the parts of speech of the words in the definitions. In addition, semantic information that is available in the Box Codes on the machine-readable LDOCE is included in the RRF for a fuller conceptual representation of the definitions. Fuller detail of the process and results of developing RRF for the verb 'make' are available in my final report to RDL in 1991, entitled "Development of a Methodology for Extracting Semantic Relations from Definitions". Figure 3 provides a few definitions which use the verb 'make' and indicate the type of semantic information which can be added to a definition by applying the RRF.

conscript	to make [**CAUSE**] someone [**PATIENT**] serve in one of the armed forces by law
damp	to make [**CAUSE**] a stringed musical instrument [**AFF.OBJ**] sound less loudly, e.g. by using a damper
defuse	to make [**CAUSE TO BE**] less dangerous [**RES.ATT**] or harmful [**RES.ATT**]
convey	to make [**CAUSE TO BE**] feelings, ideas, thoughts, etc. [**AFF.OBJ**] known [**RES.ATT**]
copy	to make [**PRODUCE**] a copy [**RES.CONC**] of
crackle	to cause to make [**PRODUCE**] small sharp repeated sounds [**RES.ABS**]

AFF.OBJ = Affected Object RES.ATT = Resulting Attribute
RES.CONC = Result, Concrete RES.ABS = Result, Abstract

Fig 3: Definitions with relations detectable by RRF for 'to make'

Procedures

The RRF were developed for one verb at a time, using the following procedures. The verb definitions were extracted from LDOCE and Key Word in Context (KWIC) programs were run for each verb's set of definitions, to enable us to look for similarities in syntax. Any definitions that were for the noun sense of a verb (e.g. 'cause' is both a noun and a verb) or used the verb later in the definition and not as the defining verb were eliminated. The intellectual analysis began by first grouping definitions based on shared sense of the defining verb. The like definitions were then grouped together based on similar syntax. This work was guided by checking the senses found in LDOCE and other dictionaries to see the senses for the verb being analyzed. This means that if the verb under analysis was 'cause' and the hundreds of definitions being analyzed were those in which 'cause' occurred, then the possible senses of 'cause' were checked in the dictionaries to serve as a possible guide to senses that might possibly occur in the definitions which used 'cause' as their defining verb. Generally, we found that only a subset of the possible senses of a verb found in the dictionaries were actually used for defining other verbs.

After determining the senses of a verb which occurred in the definitions, an initial Relation Revealing Formula (RRF) for each sense was developed. If RRF's were similar within a verb, an attempt was made to combine RRFs into one. This initial breakdown of senses was done for each of the seventeen verbs. If the sense developed was more complex than the verb it was trying to define, then further analysis was done to see if another, simpler sense could be substituted.

Twenty definitions verbs were left as 'questionnables' at the end because the definitions did not fit into any one sense. This is a relatively very small proportion of the total definitions that were analyzed. Other verbs had 'idiom' senses. For example the verb 'place' has an 'associate' sense because the definition uses 'to place a particular meaning on'. Idioms are those senses which have idiosyncratic meanings which are not based on a compositional analysis of the words used in the definition.

At the same time that a sense was being selected for each definition, the remainder of the definition was being analyzed to determine the semantic roles played by other words and phrases in the definitions. In several instances, the sense of the verb was based on the semantics and

syntactics of the words following the verb. For example, in definitions containing 'to be' + ADV/ADJ, the adverb or adjective conveys the CHARACTERISTIC relation and the sense of the verb BE is EXIST.

After all the verbs were broken down by senses and initial RRF's were written, a matrix was created to present an overview model of how verbs share senses and to show how the RRF for one sense of a verb differs from the RRF for other senses of the same verb. The idiom senses were separated out into a separate matrix. Some verbs took "higher level" verbs (i.e. the sense is the same as a lexical verb) as a sense. Also, the verb 'place' has a PUT sense and vice versa. Although this may be confusing initially, we could not find any way around it, as the lexical verbs which were used as senses convey the meaning of that sense better than any other term we could find. This problem is similar to the circularity problem encountered by lexicographers for all dictionaries. Appendix 1 is the final verb matrix and presents in summary form the individual verbs that were analyzed and the senses for each verb that were used in the dictionary definitions. Appendix 2 is the matrix for the idiom senses of the verbs.

The lexical expressions (e.g. EXIST, CHANGE PLACE, CAUSE TO BE) used to label or characterize the senses were chosen from several sources. The first, and most often used source, was the verb definitions from the print version of LDOCE. Other sources referred to included New Webster's Dictionary and the Merriam Webster's Thesaurus.

Once a verb sense was identified and an initial Relation Revealing Formula (RRF) created, a decision was made whether the sense was an idiom. From a total of 34 different senses, sixteen are designated as idiomatic. Eight of the seventeen verbs analyzed have an idiomatic sense. Four verbs have one idiomatic sense, one verb has two idiomatic senses, two verbs have three idiomatic senses and one verb has four idiomatic senses.

In overviewing the senses found in this analysis (the column labels in the Appendices), the following summary interpretation is offered. The sense EXIST has four subcategories - EXIST as a CHARACTERISTIC, CONDITION/POSITION, ROLE/FUNCTION, or STATE. The difference of each subcategory is designated syntactically in the RRF. For example, EXIST as a CHARACTERISTIC is followed in a definition by either an ADJ or ADV. For the subcategory CONDITION/POSITION, the part of speech found after the verb is the preposition 'in/as/at/on' plus a Noun Phrase.

The senses CAUSE and CAUSE TO BE were found to be the senses occurring most frequently across the seventeen verbs, with seven verbs having CAUSE senses and five verbs having CAUSE TO BE senses. CAUSE TO BE is a sense for five verbs when the definition contains a specific action and outcome. For example, the CAUSE TO BE formula for the verb **BRING** contains 'to bring under/ into' plus a NP. The verb **PUT** also uses a similar RRF for the CAUSE TO BE sense. Four verbs use both CAUSE and CAUSE TO BE senses. Two verbs have two separate senses of CAUSE which are distinguished by their RRF. One verb has two separate senses of CAUSE TO BE also distinguished by its RRF. This distinction is necessary since each of these senses comprise the largest number of verb definitions.

There are four manifestations of a CHANGE sense. First, two verbs use CHANGE which entails changing something or someone into something else. Two other uses of the underlying semantics of the notion of change are CHANGE PLACE and CHANGE POSITION. CHANGE PLACE indicates a change of location while CHANGE POSITION indicates an adjustment in physical position but the location remains the same.

Another manifestation of the notion of change is CAUSE TO CHANGE PLACE which refers to a change of location as caused by someone or something. The verb 'move' uses the three senses CHANGE PLACE, CHANGE POSITION, and CAUSE TO CHANGE PLACE.

The verb **MOVE** also uses a TRANSFER sense, as do two other verbs. The TRANSFER sense involves the relocation of someone or something from one place to another. The verb definition usually includes the place something was moving from or to. The relations respectively are LOCATION-SOURCE and LOCATION-GOAL.

However, the majority of senses were used by only one verb. For example, the senses EFFECT, PROVIDE, PLACE, PUT, CAUSE TO POSSESS, EXPERIENCE, PERFORM, PRESENT, INDICATE, MUST, PUSH, EMPLOY, and REMOVE were only used once.

Most idiom RRF's included wording that was very consistently repeated throughout the definitions containing the idiomatic sense. For example, for the idiomatic sense ASSOCIATE for the verb **PLACE** every definition contain the words 'a particular meaning on'. This wording then

became part of the RRF and the sense ASSOCIATE is idiomatic for the verb **PLACE** since the wording is not found in any other definitions.

RESULTS

The results that are reported herein are very detailed and likely of interest only to those who have been involved in or plan to undertake a similar analysis of how different senses of defining verbs can be detected in definition. The cursory reader is directed to the Verb-Sense Matrix in Appendix 1 and the Verb-Idiom Matrix in Appendix 2. In the next few pages, our results will be presented for each verb analyzed. First, the actual RRF for the verb will be presented in brief formulae, to be followed by a more narrative explanation of the findings that supported the development of these RRF.

First, a few lines of explanation of how the various fonts, capitalization, bracketing, and abbreviations are used in the following detailed report of the results. The verb being reported on occurs in bold, capital letters and is followed by a number indicating the frequency of occurrence of that verb as the defining verb in one-half of the definitions in LDOCE. The remaining half of the definitions are reserved to provide the database against which these rules, when automated, will be tested.

The senses are in all caps to the left of the equal sign; parts of speech are also in all caps but they appear only to the right of the equal sign and are more familiar to readers; semantic relations are in all caps within square brackets; the words as they actually appear in definitions in LDOCE are in single quotes; optional items are in curly brackets; alternatives are separated by diagonals. As an example, the first set of RRF are for the verb **BE** has six RRF.These are based on the 200 definitions in one-half of LDOCE in which BE served as the first and defining verb. The first RRF is for the sense REPRESENT. This sense can be recognized in LDOCE definitions by the occurrence of the words 'to be' followed by one of the following expressions: 'a sign' or 'a picture' or 'a symbol'. Each of these three expressions may be followed by either the word 'that' or the word 'of'. In all instances, the next item in the definition will be a NP (Noun Phrase) which exists in the semantic relation to the word being defined as an [ABSTRACT OBJECT]. Optionally, the NP will be followed by a PP (Prepositional Phrase).

BE (defining verb in 200 definitions)

REPRESENT = 'to be' + 'a sign/picture/symbol {that/of}' + NP [ABSTRACT OBJECT] + {PP}

EXIST = 'to be' + ADJ/ADV [CHARACTERISTIC] + {'to' + VB [ACTION]} + {NP [OBJECT/ PATIENT]} + {PP}

EXIST = 'to be' + ADJ/ADV [CHARACTERISTIC] + {PP}

EXIST = 'to be' + PAST PARTICIPLE [STATE] + {PP/ADV}

EXIST = 'to be' + 'in/as/at/on' + NP [CONDITION/POSITION]

EXIST = 'to be' + NP [ROLE/FUNCTION] + [P.P.]

The verb **BE** occurred in 200 definitions and conveys two basic senses in all its occurrences, one of which is idiomatic. Of the total number of occurrences, only 13 are idiomatic. When 'to be' plus the words 'sign/picture/symbol are used the verb being defined conveys the idiomatic sense of REPRESENT. For example, the word 'betray' is defined as 'to be a sign of [REPRESENT] 'something one would like to hide'.

The remaining definitions convey the sense of the verb **BE** as EXIST which is itself sub-categorized into four senses and detectable by five RRF. The RRF are separable by the different parts of speech following 'to be.' If it is an ADJ or ADV, then the sense is EXIST-CHARACTERISTIC]. For example, the word 'linger' is defined as 'to be [EXIST] slow [CHARACTERISTIC] to disappear.' Another example is the word 'suffice' which is defined as 'to be [EXIST] enough [CHARACTERISTIC]'. There are two RRF's for the EXIST-CHARACTERISTIC sense, since some definitions also contain a VB [ACTION] and NP [OBJECT/PATIENT].

If the definition contains a PAST PARTICIPLE, then the sense is 'exist' [STATE]. For example, the word 'cook' is defined as 'to be [EXIST] prepared [STATE] in this way [MANNER].' As another example, the word 'sell' is defined as 'to be [EXIST] bought [STATE].'

If the definition contains one of the following PREPOSITIONS 'in/as/at/on', plus a NP, then the sense is 'exist' [CONDITION]. For example, the word 'conflict' is defined as 'to be [EXIST] in opposition [CONDITION].'

If the definition contains a NP, then the sense is 'exist' [ROLE/FUNCTION]. For example, the

word 'preside' is defined as 'to be [EXIST] in charge of [FUNCTION].' Whereas, the word 'reign' is defined as 'to be [EXIST] the king or queen [ROLE].' The NP representing a role can also identify a person.

BECOME (defining verb in 46 definitions)

CHANGE = 'to become'

CHANGE = 'to become' + NOUN [I.D.] + {PP}

CAUSE TO BE = 'to become' + {ADV} + ADJ [CHARACTERISTIC] + {ADV/PP}

CAUSE TO BE = 'to become' + {ADV} + PAST PARTICIPLE [STATE] + {NP} + {PP}

The verb **BECOME** has two senses, with each sense having two RRFs. The first sense is CHANGE. CHANGE's first RRF consists simply of the phrase 'to become.' For example, the word 'get' is defined as 'to become [CHANGE].' The second RRF for the sense CHANGE contains a NP labeled as an I.D. For example, the word 'father' is defined as '(of a man) to become [CHANGE] the father [I.D.] of.'

The base sense for the verb BECOME is 'cause to be.' Its two RRF's are similar in structure, except one contains an ADJ which indicates the presence in the definition of a CHARACTERISTIC and the PAST PARTICIPLE reveals a STATE relation. For example, the first RRF occurs in the definition for 'languish', which is defined as 'to become [CAUSE TO BE] weak [CHARACTERISTIC], or unhappy [CHARACTERISTIC] through desire [MEANS].' In the second RRF, the past participle is labeled a STATE. For example, the word 'crumble' is defined as 'become [CAUSE TO BE] ruined [STATE].'

BRING (defining verb in 55 definitions)

CHANGE PLACE = 'to bring' + {NP [OBJ]} + ADV [DIRECTION/LOCATION] + ...

TRANSFER = 'to bring' + {NP [CONC.OBJ]} + 'from' + {NP [LOC-SOURCE]} + {'to'} + {NP [LOC-GOAL]}

CAUSE = 'to bring' + {NP [OBJ]} + 'to' + NP [STATE] + {PP/ADV}

CAUSE = 'to bring' + NP [STATE] + 'to/on/in' + {NP (PATIENT)}

CAUSE TO BE = 'to bring' + 'under/into' + NP [STATE] + {PP}

RETURN = 'to bring' + {NP [PATIENT]} + 'back' + {(NP) [OBJ]} + {'to/into'} + {NP [STATE/LOC-GOAL]}

The verb **BRING** has five senses, one of which is idiomatic. Three of the RRF's are quite similar, in that the senses they indicate are themselves quite similar and basic variants of the base sense CHANGE PLACE. The RRF for the base sense differs in that it denotes a direction or location. For example, the word 'pull' is defined as 'to bring [CHANGE PLACE] something [OBJECT] along behind one [DIRECTION] while moving [MANNER].'

A similar RRF is for the sense TRANSFER, but instead of DIRECTION, the RRF contains a LOCATION-SOURCE and a LOCATION-GOAL relation. For example, the word 'get' is defined as 'to bring [TRANSFER] from one place [LOC-SOURCE] to another [LOC-GOAL].'

Another sense for the verb **BRING** is CAUSE and is represented by two RRFs. The first RRF signifies bringing an OBJECT to a STATE. And the second RRF signifies bringing a STATE 'to/on/in' a PATIENT. For example, the word 'promote' is defined as 'to bring [CAUSE] goods [OBJECT] to public notice [STATE] in order to encourage people to buy [REASON].' An example for the second RRF is the word 'cure' which is defined as 'to bring [CAUSE] health [STATE] to a person [PATIENT] in place of disease or illness [CONCESSION].'

A similar sense for the verb **BRING** is CAUSE TO BE. The NP in the RRF is also labeled a STATE. For example, the word 'disrupt' is defined as 'to bring [CAUSE TO BE] into disorder [STATE].'

The idiomatic sense for the verb **BRING** is RETURN and is represented in the RRF by including the word 'back' in each definition. For example, the word "restore" is defined as 'to bring back [RETURN] into use or existence [STATE].' This sense is idiomatic since the locational sense of back is not at all what is intended in this definition.

CAUSE (defining verb in 377 definitions)

EFFECT STATE = 'to cause to' + {NP [OBJECT/PATIENT]} + {ADV} + VERB [STATE] + {NP [ABSTRACT OBJECT]} + {ADV/PP}

EFFECT ACT = 'to cause to' + {NP [OBJECT/PATIENT]} + VERB [ACT] + {NP [CONCRETE OBJECT]} + {ADJ/ADV/PP}

PRODUCE = 'to cause' + NP [OBJECT] + 'on/to'

The verb **CAUSE** can covey two basic senses, but one of these is sub-categorized further, so there are three RRFs as a result. The base sense is to EFFECT but is differentiated by the definitions either effecting a STATE or an ACT. For example, the word 'render' is defined as 'to cause to [EFFECT] be [STATE].' Or the word 'persuade' is defined as 'to cause to [EFFECT] believe [STATE] or feel certain [STATE].'

The RRF for EFFECT-ACT is similar syntactically but the definitions containing this RRF are distinctly different in meaning conveyed. For example, the word 'vaporize' is defined as 'to cause to [EFFECT] change [ACTION] into vapour [CONCRETE OBJECT].' Or the word 'rain' is defined as 'to cause to [EFFECT] fall [ACTION] like rain [CONCRETE OBJECT].'

The second sense for the verb **CAUSE** is PRODUCE and occurs in only two definitions. An example is the word 'bruise' which is defined as ' to cause [PRODUCE] a bruise [ABSTRACT OBJECT] on.'

DO (defining verb in 34 definitions)

PERFORM = 'to do' + {ADV} + {NP [ACTIVITY]} + {ADV/PP}

The verb **DO** occurs in only 34 definitions and its uses in definitions always convey the sense PERFORM. Since 'to do' denotes an action, an action word was chosen to for the sense as well. As an example, the word 'experiment' is defined as 'to do [PERFORM] an experiment [ACTIVITY].'

FORCE (defining verb in 22 definitions)

CAUSE = 'to force' + {NP [PATIENT]} + {ADV} + {"to" + VERB [ACTION]} + {NP [RESULT]} + {PP/ADV}

CAUSE = 'to force' + {NP [(OBJ)} + {ADV} + {'into'} + {NP [RESULT]} + {PP/ADV}

PUSH = 'to force' + {(NP) [PATIENT]} + {ADV [MANNER]} + {NP [OBJ]} + {PP}

The verb **FORCE** can convey two senses CAUSE and PUSH with the majority of the definitions conveying the CAUSE sense. Also, there are two RRF's for the CAUSE sense. One RRF contains an optional VERB and includes a PATIENT performing an ACTION or an ACTION causing a RESULT. Even though the word **FORCE** seems have a connotation of action, the basic sense of the word is CAUSE For example, the word 'frogmarch' is defined 'to force [CAUSE] a person [PATIENT] to

move [ACTION] forward [DIRECTION] with the arms held together firmly from behind [MANNER].' And the word 'muzzle' is defined as 'to force [CAUSE] to keep [ACTION] silent [RESULT].'

The second RRF for the CAUSE sense of **FORCE** entails an OBJECT forced into a RESULT. For example, the word 'compress' is defined as 'to force [CAUSE] a substance [OBJECT] into less space [RESULT]' or 'squash' is defined as 'to force [CAUSE] into a flat shape [RESULT].'

The other sense of **FORCE** is PUSH, which conveys the manner in which the **FORCE** occurs. For example, 'dislodge' is defined as 'to force [PUSH] out [MANNER] of a position [LOCATION-SOURCE]' or 'overtax' is defined as 'to force [PUSH] beyond [MANNER] a limit [RESULT].'

FORM (defining verb in 26 definitions)

PRODUCE = 'to form' + {'into'} + {NP [RESULT]} + {ADV/PP} + {'by' [MEANS]}

CAUSE = 'to form' + NP [PATIENT] + VERB [ACTION] + NP [RESULT]

The verb **FORM** has two senses, PRODUCE and CAUSE, but the basic sense is PRODUCE. The sense CAUSE occurs in but one definition and is the only definition containing the verb **FORM** with a person in the definition. The word 'dress' is defined as 'to form [CAUSE] soldiers [PATIENT] to form [ACTION] a straight line [RESULT].' The RRF for the sense PRODUCE signifies there is a RESULT from the action **FORM**. For example, the word 'stack' is defined as 'to form [PRODUCE] into a neat pile [RESULT].' Or the definition may not contain the specific RESULT but tells how the implied result is accomplished via the MEANS relation. For example, the word 'forge' is defined as 'to form [PRODUCE] by heating and hammering [MEANS].'

GIVE (defining verb in 155 definitions)

CAUSE [AN EFFECT] = 'to give' + NP [EFFECT] + {ADV/PP}

PROVIDE = 'to give' + {NP [RECIPIENT]} + {VERB} + {NP [OBJECT]} + {PP/ADV}

PROVIDE = 'to give' + NP [ABSTRACT OBJ] + {'to'} + NP [RECIPIENT] + {PP}

TRANSFER = 'to give' + {NP [CONCRETE OBJECT]} + {NP [RECIPIENT]} + {PP}

EMIT = 'to give' + 'out' + {NP [ABSTRACT OBJECT]} + {PP}

FORGO = 'to give' + 'up' + {ADV} + {(NP)}

BIRTH = 'to give' + 'birth" + {ADV} + NP [PATIENT]

The verb **GIVE** has six different senses, three of which are idiomatic. The three base senses are CAUSE, PROVIDE, and TRANSFER. The three idiomatic senses are EMIT, FORGO, and BIRTH. The base sense of the verb **GIVE** is PROVIDE which is revealed by two RRFs. The first RRF signifies someone [RECIPIENT] is being provided with something [OBJECT]. For example, the word 'promote' is defined as 'to give [PROVIDE] someone [RECIPIENT] a higher position or rank [OBJECT].' The other RRF signifies something [OBJECT] is being given to someone [RECIPIENT]. In many definitions the recipient is implied. For example, the word 'inform' is defined as 'to give [PROVIDE] information or knowledge [OBJECT] to.' The word 'shelter' is defined as 'to give [PROVIDE] shelter [OBJECT] to.' The OBJECT can be either abstract or concrete.

Another sense of **GIVE** is TRANSFER and its meaning as well as its RRF differs from PROVIDE because the object being given to the RECIPIENT is CONCRETE and sometimes includes a LOCATION-SOURCE and LOCATION-GOAL. For example, the word 'hand' is defined as 'to give [TRANSFER] from one's own hand [LOC-SOURCE] into someone else's [LOC-GOAL].' The word 'present' is defined as 'to give [TRANSFER] something [OBJECT] away.' The TRANSFER sense contains an implied movement of an object.

The last sense of **GIVE** is CAUSE. The definitions show that 'to give' an ABSTRACT OBJECT can be seen as to CAUSE AN EFFECT. For example, the word 'tint' is defined as 'to give [CAUSE] a slight or delicate color [EFFECT] to.' The word 'feel' is defined as 'to give [CAUSE] or produce the stated sensation [EFFECT].'

As for the idiomatic senses of **GIVE**, the RRF for EMIT contains the idiomatic phrase 'to give out.' For example, the word 'glow' is defined as 'to give out [EMIT] heat [OBJECT] and/or soft light [OBJECT] without flames or smoke [MANNER].' The next sense FORGO is the largest of the idiomatic groups and the RRF contains the phrase 'to give up.' For example, the word 'yield' is defined as 'to give up [FORGO] control [OBJECT] of.' The last idiomatic sense BIRTH has only three definitions. The RRF contains the phrase 'to give birth.' For example, the word 'foal' is defined as 'to give birth [BIRTH] to a foal [PATIENT].'

GO (defining verb in 73 definitions)

CHANGE PLACE = 'to go' + {NP [OBJ]} + {ADV [DIRECTION]} + {PP} + {NP [OBJ/PATIENT/ LOC-GOAL]}

DEPART = 'to go away' + {ADV [MANNER]} + {PP}

DEFY = 'to go against' + NP [OBJ]

EXCEED = 'to go' + ADV [DEGREE] + {NP [ABST-OBJ]}

The verb **GO** has four distinct senses, three of which are idiomatic. The base sense for the verb GO is CHANGE PLACE and has a complex RRF, with many variations. For example, the word 'sneak' is defined as 'to go [CHANGE PLACE] quietly and secretly, [MANNER] so as not to be seen [PURPOSE].' Or the word 'decline' is defined as 'to go [CHANGE PLACE] from a better [LOC-SOURCE] to a worse position [LOC-GOAL].'

The first idiomatic sense of **GO** is DEPART and is represented in the RRF by the words 'to go away.' For example, the word 'begone' is defined as 'to go away [DEPART] at once [MANNER].' The second idiomatic sense is DEFY and is represented in the RRF by the words 'to go against.' For example, the word 'disoblige' is defined as 'to go against [DEFY] the wishes [OBJECT] of.' The third idiomatic sense is EXCEED and contains an ADV labeled as a DEGREE in the RRF. For example, the word 'overflow' is defined as 'to go [EXCEED] beyond [DEGREE] the limits [ABSTRACT OBJECT] of.'

HAVE (defining verb in 63 definitions)

POSSESS = 'to have' + {NP [OBJECT]} + {PP}

CAUSE = 'to have' + NP [EFFECT] + {'on'} + {NP [PATIENT/OBJECT]} + {PP}

EXPERIENCE = 'to have' + NP [ABST OBJ] + {NP [PATIENT]} + {PP}

MUST = 'to have to' + {VERB [ACTION]} + {P.P.}

CAUSE TO BE = 'to have' + PAST PARTICIPLE [ACTION] + {PP} + {NP}

The verb **HAVE** has five senses, one of which is idiomatic. The base sense is POSSESS and is signified in the RRF by the pattern 'to have' + OBJECT. For example, the word 'merit' is defined as 'have [POSSESS] a right to [OBJECT].' Or, the word 'suit' is defined as 'to have [POSSESS] the right qualities [OBJECT].'

Another sense for the verb **HAVE** is CAUSE. When this sense is intended, 'have' is followed in the definition by a NP representing an EFFECT. For example, the word 'impact' is defined as 'to have [CAUSE] an impact [EFFECT] on. A similar sense is CAUSE TO BE but is different because the definition contains a PAST PARTICIPLE which is labeled as an ACTION. For example, the word 'deserve' is defined as 'to have [CAUSE TO BE] earned [ACTION] by one's actions [MANNER].'

The second largest group of definitions containing **HAVE** convey the sense EXPERIENCE. The definitions are different from the previous senses because the word being defined conveys more of an action than a state. For example, the word 'ache' is defined as 'to have [EXPERIENCE] a continuous, but not violent pain [OBJECT].' Or, the word 'violate' is defined as 'to have [EXPERIENCE] sex [OBJECT] with a woman [PATIENT] by force [MEANS].'

The idiomatic sense of HAVE is MUST and is distinguished in the RRF by having 'to have to' in the definition. For example, the word 'owe' is defined as 'to have to [MUST] pay [ACTION] for something already done [PURPOSE].'

MAKE (defining verb in 426 definitions)

PRODUCE = 'to make' + NP [ABSTRACT OBJECT/CONCRETE OBJECT]

CAUSE = 'to make' + NP + VERB

CAUSE TO BE = 'to make' + NP + ADJ

The verb **MAKE** conveys three distinct senses and three distinct RRFs. The base sense of the verb is PRODUCE and can be the production of either a CONCRETE OBJECT or an ABSTRACT OBJECT. For example, the word 'draw' is defined as 'to make [PRODUCE] pictures [CONCRETE OBJECT] with a pencil or pen [INSTRUMENT.]' Or the word 'grin' is defined as 'to make [PRODUCE] a wide smile [ABSTRACT OBJECT].'

The second sense for the verb **MAKE** is CAUSE and includes a second VERB in each of the definitions. For example, the word 'enforce' is defined as 'to make [CAUSE] something [OBJECT] happen [ACTION] esp. by threats or force [MEANS].' Or the word 'shape' is defined as 'to make [CAUSE] a piece of clothing [CONCRETE OBJECT] fit the body [ACTION] closely [MANNER].'

The third sense for the verb **MAKE** is CAUSE TO BE and contains an ADJ or PAST PARTICIPLE to

distinguish the RRF. For example, the word 'enable' is defined as 'to make [CAUSE TO BE] possible [RESULTING ATTRIBUTE].' Or the word, 'declare' is defined as 'to make [CAUSE TO BE] known [RESULTING ATTRIBUTE] publicly or officially [MANNER].'

MOVE (defining verb in 117 definitions)

CHANGE POSITION = 'to move' + {NP [OBJECT/(PATIENT)]} + {ADV/PP [MANNER]}

CAUSE TO CHANGE PLACE = 'to move' + {NP [OBJ/PATIENT]} + {ADV} + PP

TRANSFER = 'to move' + {NP [OBJECT]} + {ADV} + PP [LOC-SOURCE/LOC-GOAL]

CHANGE PLACE = 'to move' + ADV [DIRECTION] + {NP [LOC-GOAL]} + ..

The verb **MOVE** is represented by four quite similar senses: CHANGE POSITION, CAUSE TO CHANGE PLACE, TRANSFER, and CHANGE PLACE. The largest group of definitions is represented by the sense CHANGE POSITION. For example, the word 'wiggle' is defined as 'to move [CHANGE POSITION] in small side-to-side, up-and-down, or turning movements [MANNER].' Or the word 'rocket' is defined as 'to move [CHANGE POSITION] at very great speed [MANNER].'

The next largest group of definitions containing **MOVE** is represented by the sense CAUSE TO CHANGE PLACE. For example, the word 'occupy' is defined as 'to move [CAUSE TO CHANGE PLACE] into and hold possession of (a place), [OBJ] e.g. by military force [MEANS].'

The other two senses, TRANSFER and CHANGE PLACE represent approximately the same number of definitions. For the sense TRANSFER, the word 'jump' is defined as 'to move [TRANSFER] suddenly [MANNER] from one point [LOC-SOURCE] to another, [LOC-GOAL] often missing out what comes in between [RESULT]. And, for the sense CHANGE PLACE, the word 'recede' is defined as '(of a thing) to move [CHANGE PLACE] back or away [DIRECTION].'

PLACE (defining verb in 13 definitions)

PUT = 'place' + {NOUN [PATIENT/OBJ] + {ADV/PP}

GIVE = 'to place' + Noun Phrase [CONCRETE OBJECT]

ASSOCIATE = 'to place' + 'a particular meaning on'

The verb **PLACE** has three senses, one of which is idiomatic. All three senses have a small group

of definitions, with the base sense being PUT. For example, the word 'settle' is defined as 'to place [PUT] (someone or oneself) [PATIENT] so as to be comfortable [PURPOSE].' Most definitions for the sense PUT have a PATIENT following the verb.

Another sense of **PLACE** is GIVE and its RRF contains the VERB plus a CONCRETE OBJECT. For example, the word 'commission' is defined as 'to place [GIVE] a special order [CONC OBJ] for, or appoint someone to do (something).'

The idiomatic sense for the verb **PLACE** is ASSOCIATE and its RRF contains the words 'a particular meaning on.' For example, the word 'interpret' is defined as 'place [ASSOCIATE] a particular meaning on.'

PUT (defining verb in 143 definitions)

PLACE = 'to put' + {NP [OBJ]} + 'under/in/into/through/onto' + NP [OBJECT/LOC-GOAL]

CHANGE = 'to put' + (NP) [ABSTRACT OBJECT] + {ADV} + 'into' + NP [RESULT]

CAUSE = 'to put' + {ADV [DEGREE]} + NP [CONDITION] + 'to/on' + ..

CAUSE TO BE = 'to put' + {(NP) [PATIENT/OBJ]} + 'into/under/ out/in' + NP [CONDITION]

COVER = 'to put' + 'on' + NP [OBJ}

COVER = 'to put' + NP [OBJ] + 'on/over/onto' + {NP [LOC-GOAL]} + {...}

COMBINE = 'to put' + 'together' + {...}

The verb **PUT** has six senses, two of which are idiomatic. Just as the base sense for **PLACE** is PUT, the base sense for **PUT** is PLACE. The definitions for the PLACE sense usually contain the preposition of 'under/in/into/through/onto' as indicated in the RRF. For example, the word 'immerse' is defined as 'to put [PLACE] deep into a body of liquid [LOC-GOAL].' Or, the word 'blockade' is defined as 'to put [PLACE] under a blockade [LOC-GOAL].'

The second sense of **PUT** is CHANGE. The definitions which convey the CHANGE sense all contain an ABSTRACT OBJECT which is represented in the RRF. For example, the word 'encode' is defined as 'to put [CHANGE] a message [ABSTRACT OBJECT] into a code [RESULT].'

Two other senses of **PUT**, CAUSE and CAUSE TO BE are somewhat similar but have two distinct

RRFs. Both senses may contain a CONDITION, but definitions for CAUSE TO BE may have two NPS. For example, the word 'disgrace' is defined as 'to put [CAUSE TO BE] a public person [PATIENT] out of favour [CONDITION].' Or, the word 'restore' is defined as 'to put [CAUSE TO BE] back into a former position [CONDITION].' The CAUSE sense may have an ADV signifying a DEGREE. For example, the word 'overtax' is defined as 'to put [CAUSE] too great [DEGREE] a tax [CONDITION] on goods [OBJECT].'

The first idiomatic sense of **PUT** is COVER and is detected in definitions by two semantically very similar RRFs. The first RRF signifies putting on something; and the second RRF signifies putting something 'on/over/onto' something. For example, the word 'dress' is defined as 'to put [COVER] on formal clothes [OBJECT] for the evening [REASON].' And the word 'plaster' is defined as 'to put [COVER] wet plaster [OBJECT] on.'

The second idiomatic sense of **PUT** is COMBINE and is indicated by the word 'together' in definitions. For example, the word 'string' is defined as 'to put [COMBINE] together onto a thread [OBJECT] so as to form a string [PURPOSE].'

SHOW (defining verb in 40 definitions)

INDICATE = 'to show' + {ADV} + {NP [OBJ/PATIENT]} + {PP}

PRESENT = 'to show' + (NP) + ADV

The verb **SHOW** has two senses, with one sense occurring only one definition. The base sense is INDICATE. For example, the word 'chart' is defined as 'show [INDICATE] on a chart [LOC-GOAL].' Or, the word 'relate' is defined as 'to show [INDICATE] a connection between [RELATIONSHIP.' The other sense is PRESENT and the example word defined for this sense is 'rerun', 'to show [PRESENT] a film or recorded broadcast [OBJECT] again [FREQUENCY].'

TAKE (defining verb in 90 definitions)

POSSESS= 'to take' + {NP [ABSTRACT OBJECT/PATIENT]} + {PP}

CAUSE TO POSSESS = 'to take' + {NP [OBJECT/PATIENT]} + {PP/ADV}

REMOVE = 'to take' + {NP [OBJ]} "away/off/out/from/back" + {NP [OBJECT/PATIENT]} + {PP/ADV}

PARTICIPATE = 'take part in' + Noun Phrase [ABSTRACT OBJECT] + {P.P.}

INGEST = 'to take' + {NP [OBJ]} + 'in/into/through' + {NP [LOC-GOAL]} + {PP}

CHOOSE = 'to take' + NP [OBJ]

GRASP = 'to take hold" + {ABST OBJ} + {ADV/PP}

The verb **TAKE** can convey seven senses, four of which are idiomatic. The POSSESS sense is the simplest and most frequent. In the definitions conveying the POSSESS sense, 'to take' is most usually, but not always followed by a NP which is the ABSTRACT OBJECT or PATIENT of the POSSESS sense. For example, the verb 'conquer' conveys the POSSESS sense via the definition of 'to take' (land)[OBJ] by force [MEANS].

The CAUSE TO POSSESS sense of **TAKE** covers those definition senses in which the POSSESS act is a bit more removed as in the definition of 'colour' where 'to take on [CAUSE TO POSSESS] takes as its OBJ, 'colour'.

The sense of the verb **TAKE** which is REMOVE occurs in definitions which can contain one of several different prepositions. For example, the word 'deduct' is defined as 'to take away [REMOVE] an amount [OBJECT] from a total [LOC-SOURCE].' Or the word 'undress' is defined as 'to take [REMOVE] the clothes [OBJECT] off someone [PATIENT].'

Three of the idiomatic senses of **TAKE** contain specific wording in their RRF. They include: PARTICIPATE, INGEST, CHOOSE and GRASP.

The first idiomatic sense PARTICIPATE includes the phrase 'take part in' in each definition. For example, the word 'join' is defined as 'to take part in [PARTICIPATE] an activity [ABSTRACT OBJECT] as a member of a group [MANNER].'

The second idiomatic sense of **TAKE**, INGEST includes one of the prepositions 'in/into/through' in the definitions. For example, the word 'inhale' is defined as 'to take [INGEST] cigarette smoke [OBJECT] into the lungs [LOC-GOAL].' Or the word 'sniff' is defined as 'to take [INGEST] a harmful drug [OBJECT] through the nose [LOC-PATH].'

The third idiomatic sense CHOOSE does not contain any specific wording in the RRF, but is

definitely idiomatic. For example, the word 'fork' is defined as 'of a person [AGENT] to take [CHOOSE] one fork of a road [OBJECT].'

The fourth idiomatic sense GRASP contains the words 'to take hold' in some format in the definitions. For example, the word 'grip' is defined as 'to take hold [GRASP] of the attention [ABSTRACT OBJECT].'

USE (defining verb in 32 definitions)

EMPLOY = 'to' + {ADV} + 'use' + {ADV} + {NP [OBJ]} + {PP}

CONSUME = 'to use' + {'up'} + {NP [OBJ]} + {PP/ADV}

The verb **USE** has two senses, one of which is idiomatic. The base sense is EMPLOY and the idiomatic sense is CONSUME. For example, the word 'defend' is defined as 'to use [EMPLOY] arguments [OBJECT] to support, protect, or show the rightness of [REASON].' Or, the word 'waste' is defined as 'to use [EMPLOY] wrongly [MANNER].'

The definitions for the idiomatic sense CONSUME usually contain the word 'up.' For example, the word 'exhaust' is defined as 'to use up [CONSUME] completely [DEGREE].' But CONSUME is also used for the word 'drink' which is defined as 'to use [CONSUME] alcohol [OBJECT].'

Conclusion

The development of a methodology for producing both syntactic and semantic level RRF which can produce semantically tagged representations of definitions, as well as the RRFs themselves, will aid in bringing the field of computational linguistics much closer to its goal of developing automatic procedures for extracting semantic relations from text. As all NLP systems are reliant on large, rich lexicons of the common vocabulary of everyday communication, this effort provides a very valuable resource.

The work will be continued, hopefully, if time and resources allow. What needs to be done immediately is an empirical testing of the RRF on the other half of LDOCE and then on other dictionaries to determine the generalizeability of the RRF herein developed.

References

Alshawi, H. (1987). Processing dictionary definitions with phrasal pattern hierarchies. Computational Linguistics, 13 (3-4), 195-202.

Alshawi, H. (1989). Analysing the dictionary definitions. In Boguraev, B. & Briscoe, T. (Eds.). Computational lexicography for natural language processing. London: Longman.

Boguraev, B. & Briscoe, T. (1987). Large lexicons for natural language processing: Utilising the grammar code system of LDOCE. Computational Linguistics, 13 (3-4), 203-218.

Grishman, R. & Kittredge, R. (1986). Analyzing language in restricted domains. Hillsdale, NJ: Lawrence Earlbaum Associates.

Kittredge, R. & Lehrberger, J. (1982). Sublanguage: Studies of language in restricted semantic domains. Berlin: de Gruyter.

Kwasnik,B. H., Liddy, E. D. & Myaeng, S. H. (1989). Automatic knowledge extraction from dictionary text: Project development. Syracuse, NY: CASE Center Technical Report #8911.

Lenat, D., Prakash, M. & Shepherd, M. (1986). CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. AI Magazine, 65-85.

Liddy, E. D., Jorgensen, C. L., Sibert, E. & Yu, E. S. (1989). Processing natural language for an expert system using a sublanguage approach. Proceedings of the 52nd Annual Meeting of the American Society for Information Science.

Liddy, E.D., Jorgensen, C.L., Sibert, E. & Yu, E.S. (1991). Sublanguage grammar in natural language processing. Proceedings of RIAO '91 Conference. Barcelona.

Liddy, E. D. & Paik, W. (1991a). Automatic recognition of semantic relations in text. Informatics 11. London: ASLIB.

Liddy, E. D. & Paik, W. (1991b). Automatic semantic relation assigner: Preliminary work. In Natural Language Learning Workshop: IJCAI-91. Sydney, Australia.

McHale, M. (1993). Principle based parsing using a machine readable dictionary: Dissertation proposal. (unpublished manuscript).

Ravin, Y. (1993). Disambiguating and interpreting verb definitions. In Jensen, K., Heidorn, G. & Richardson, S. (Eds.). Natural language processing: The PLNLP Approach. Boston: Kluwer.

Sager, N., Friedman, C., & Lyman, M. (1987). Medical language processing. Reading, MA: Addison Wesley.

Velardi, P., Pazienza, M. T., & Fasolo, M. (1991). How to encode semantic knowledge: A method for meaning representation and computer-aided acquisition. Computational Linguistics. 17(2), pp. 153- 170.

Wilks, Y., Fass, D., Guo, C., McDonald, J., Plate, T. & Slator, B. (1987). A tractable machine dictionary as a resource for computational semantics. In: Proceedings of the Workshop on Natural Language Technology Planning. Blue Mountain Lake, NY: 1-27.

	Exist (Characteristic)	Exist (Condition/ Position)	Exist (Role/Function)	Exist (State)	Cause to Be	Cause to Be	Cause	Cause
BE	Adj/Adv + {Verb} + {NP} + {PP}	'in/as/at/on' + NP	NP + {PP}	Past Part. + {PP/Adv}				
BECOME					{Adv} + Adj + {ADV /PP}	{Adv} + Past Part. + {NP} + {PP}		
BRING					'under/into' + NP + {PP}		{NP} + 'to' + {NP} + {PP/Adv}	{NP} + 'to/on/ In' + {NP}
CAUSE								
DO								
FORCE							{NP} + {Adv} + {Verb} + {NP} + {ADV/PP}	{NP} + {Adv} + {'into'} + {NP} + {ADV/PP}
FORM							NP + NP	
GIVE							NP + {ADV/PP}	
GO								
HAVE					Past Part. + {PP} + {NP}		NP + {'on'} + {NP} + {PP}	
MAKE					NP + Adj		NP+Verb	
MOVE								
PLACE								
PUT					{NP} + {ADV} + 'into /under/out/In' + NP		{Adv} + NP + PP	
SHOW								
TAKE								
USE								

APPENDIX 1 - Verb Matrix

	Change	Change Place	Change Position	Cause to Change Place	Transfer	Effect (State)	Effect (Act)	Provide
BE								
BECOME	Noun [I.D.] + {PP}							
BRING		{NP} + Adv + {PP}			{NP} + 'from' + {NP} + {'to'} + {NP}			
CAUSE						{NP} + {Adv} + Verb + {NP} + {Adv/PP}	{NP} + Verb + {NP} + {Adj/Adv/PP}	
DO								
FORCE					{NP} + 'out' + {PP}			
FORM								
GIVE					{NP} + {NP} + {PP}			{NP} + {VERB} + {NP} + {PP/Adv}
GO		{NP} + {Adv} + {PP} + {NP}						
HAVE								
MAKE								
MOVE		Adv + {PP}	{NP} + {Adv} + {PP}	{NP} + {Adv} + PP	{NP} + {Adv } + PP			
PLACE								
PUT	(NP) + {Adv} + 'into' + NP + {PP}							
SHOW								
TAKE								
USE								

	Provide	Give	Put	Place	Possess	Cause to Possess	Experience	Perform	Present
BE									
BECOME									
BRING									
CAUSE									
DO								{Adv} + {NP} + {Adv/PP}	
FORCE									
FORM									
GIVE	{NP} + {'to'} + {NP} + {PP}								
GO									
HAVE					{NP} + {PP}		NP + {NP} + {PP}		
MAKE									
MOVE									
PLACE		NP	{NP} + {Adv/PP}						
PUT				{NP} + 'under/In/ through/Into/onto' + NP					
SHOW									(NP) + Adv
TAKE					{NP} + {PP}	{NP} + {PP/Adv}			
USE									

	Indicate	Push	Employ	Remove	Produce
BE					
BECOME					
BRING					
CAUSE					NP + 'on/to'
DO					
FORCE		{(NP)} + {Adv} + {NP} + {PP}			
FORM					{'Into'} + {NP} + {Adv/PP} + {'by'}
GIVE					
GO					
HAVE					
MAKE					NP
MOVE					
PLACE					
PUT					
SHOW	{Adv} + {NP} + {PP}				
TAKE				{NP} + 'away/off /out/from/back' + {NP} + {PP/Adv}	
USE			{Adv} + 'use' + {Adv} + {NP} + {PP}		

	Represent	Emit	Forgo	Birth	Cover	Combine	Associate	Participate	Grasp	Depart	Return
BE	sign,picture, symbol {of} + {NP} + {PP}										
BRING											{NP} + 'back' + {(NP)} + {'in/into'} + {NP}
GIVE		'out' + {NP} + {PP}	'up' + {Adv} + {(NP)}	'birth' + {Adv} + NP							
GO										'away' + {Adv} + {PP}	
HAVE											
PLACE							'a particular meaning on'				
PUT					{'on'} + NP + 'on/over/ onto' + {NP} + {PP}	'together' + {PP}					
TAKE								'part in' + NP + {PP}	'hold' + {NP} + {ADV/PP}		
USE											

APPENDIX 2 - Idiom Matrix

	Defy	Consume	Ingest	Choose	Exceed	Must
BE						
BRING						
GIVE						
GO	'against' + NP				Adv + {NP}	
HAVE						'to' + {Verb} + {PP}
PLACE						
PUT						
TAKE			{NP} + 'in/ into/through' + {NP} + {PP}	NP		
USE		{'up'} + {NP} + {PP/Adv}				

OPTICAL COMMUNICATION SYSTEM BASED ON TUNABLE DOPED FIBER OSCILLATOR

Salahuddin Qazi
Associate Professor
Department of Electrical Engineering Technology

State University of New York Institute of Technology
P.O. Box 3050, Marcy Campus
Utica, New York 13504-3050

and

Kenneth J. Teegarden
Professor
Institute of Optics

University of Rochester
Rochester, New York 14627

Final Report for:
Research Initiation Program
Photonics Laboratory of Griffiss Air Force Base
Rome, New York

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

State University of New York Institute of Technology
Utica, New York

December 1992

OPTICAL COMMUNICATION SYSTEM BASED ON TUNABLE DOPED FIBER OSCILLATOR

Salahuddin Qazi
Associate Professor
Department of Electrical Engineering Technology
State University of New York Institute of Technology

Kenneth J. Teegarden
Professor
Institute of Optics, University of Rochester

Abstract

A tunable ring laser and an optical amplifier based on erbium doped silica fiber as a gain medium were developed to be used for the wavelength division multiplexing system in fiber optic communications. The fiber ring laser uses a GRIN lens and a Fabry-Perot etalon to tune over a range of 23 nm wavelength and is centered around 1.5316 nm. The laser was pumped by a strained quantum well InGaAs laser, which delivered 20 mw of power in the fiber with a consequential output power of 10 micro-watt. The ring laser was based on a 4.75 meter erbium doped fiber amplifier developed first to establish an experimental model. Peak gain of 20 db was obtained for the amplifier.

OPTICAL COMMUNICATION SYSTEM BASED ON TUNABLE DOPED FIBER OSCILLATOR

Salahuddin Qazi

and

Kenneth J. Teegarden

INTRODUCTION

Optical amplifiers, based on erbium as dopant in a single mode silica fiber, are increasingly used to amplify weak signals due to attenuation created by long fiber transmission distances. The amplifier is also finding applications in lightwave networks operating in a broadcast mode where the signal from the central station is split into many outputs to go to multiple stations. A fiber amplifier can compensate for losses as the signals are split and also for propagation losses in the branches of network over extended distances of 10 to 25 kilometers, permitting the range of the network to be increased (1). Gain in the fiber amplifier is based on a certain erbium transition at around 1.55 micron which has attracted great interest as a result of its enormous potential for amplifiers in the third telecommunications window.

Erbium doped fibers have also been used as optical oscillators or lasers, employing the same gain medium as amplifiers, but their applications in optical communications is not as widely investigated (2). For increasing the amount of information to be transmitted in a wideband optical communication system, wavelength division multiplexing

is used. In wavelength division multiplexing a number of laser diodes of of different wavelength are independently modulated as a signal source, and all the outputs are launched into a single fiber using optical couplers (3). A tunable optical laser, based on the erbium doped silica fiber, can be effectively used as a signal laser for different wavelengths in a wavelength division multiplexed optical fiber communication systems. Erbium doped fiber lasers based on ring configuration offer a large variety of features such as large tuning range, high output power, single longitudinal mode operation and mode locking at repetition rates byond 10 GHz (4). Such lasers are proposed for use in wideband fiber optic communications where they would replace solid state diode lasers as the primary signal source.

METHODOLOGY

In order to implement an all-fiber communication system at 1.5 micron, it was essential to develop and construct a fiber amplifier and an oscillator based on erbium doped fiber as a gain medium. This was achieved in the following manner:

1. A fiber amplifier, based on erbium doped fiber, as shown in figure 1 was implemented. It consists of a strained quantum well InGaAs laser diode, operating at 0.98 micron, used as a pump laser and a wavelength multiplexer which combines an input signal at 1.5 microns with the pump signal. A wavelength demultiplexer at the output of the amplifier is used to separate the amplified output signal from any remaining pump signal. Light from the pump laser is suitably collimated

and focused with microscope objective onto the fiber end at the input of the multiplexer. Erbium doped fiber has a core diameter of 8 micron and an outer diameter of 100 micron, and was supplied by GTE Laboratories of Massachusetts. Various lengths of fiber were used to obtain suitable gain from the fiber amplifier. The overall gain of the amplifier as a function of pump power and signal power for the pump wavelength was measured. Two methods were used to make these measurements. They are shown in the experimental set-up of figure 2.

In the first method, the peak output power was measured by using an optical spectrum anlayser for a given peak input power. In the second method, a commercially available lightwave transmitter at 1.5 micron was direct intensity (amplitude) modulated with a 33 MHz sinusoidal RF signal. The resulting modulated signal was used as an input to the fiber amplifier, and the output signal was taken at the output of the lightwave receiver. The input and output signals were measured using an oscilloscope. In both the cases the input signal was measured at the output end of the demultiplexer, hence losses in this device were excluded from the determination of overall gain. Measurement of pump power launched into the amplifier were made with a Newport power meter at the output end of the demultiplexer. The emission spectrum of the amplifier as a function of pump power was measured using an optical spectrum analyser.

2. A fiber ring laser as shown in figure 3 was constructed using erbium doped fiber as the gain medium at wavelengths around 1.5 micron. The ring or the feedback loop is completed by connecting 90% of the output of the fiber amplifier, via the output coupler (splitter), to the input of the wavelength division multiplexer with a signle mode fiber. A GRIN lens and a Fabry-Perot etalon consisting of about 1 mm microscopic slide, is included in the feedback loop as a tuning and wavelength selecting element for the ring laser. Further tuning was achieved by a second Fabry-Periot etalon consisting of 0.3 mm glass plate.

Emission spectra of the laser output, with and without feedback, was obtained by using an optical spectrum analyser. Output of the laser with GRIN lens cavity and Fabry-Perot etalon was obtained by using an optical spectrum analyser. This was repeated without the Fabry-Perot etalon.

RESULTS AND DISCUSSION

Amplifier - Emission spectrum of the erbium doped fiber amplifier pumped with 20 mw is shown in figure 4. The spectrum showed a periodic ripple superimposed on the slowly varying envelope of the flourescence, whose shape and amplitude depends on the amount of pump power and the way pump power was launched. It was also noticed that at a certain connector position at higher power, lasing (oscillations) were induced as shown in figure 5. These oscillations, which causes instabilities in the gain spectrum of the amplifier, were minimized by careful alignment

of the connectors and also by using optical isolators in the amplifier design. The overall gain of an amplifier with an erbium doped fiber of length 4.75 meter is shown in figures 6a and 6b. This indicates that the peak of the gain curve for the amplifier lies at about 1.5303 micron, while the signal wavelength did not coincide with the wavelength for maximum gain. The gain obtained from the ratio of the output to the input signal is about 9.0 or 9.5 db. The overall gain of an amplifier using 4.75 meter, measured by this method as a function of launched power at 0.98 micron pump power for an input signal power of 30 nw, is shown in figure 7. These results indicate that higher gain would be obtained by using fiber of length greater than 4.75 meters. Therefore, length of fiber was increased to 7.75 meters, and measurements of overall gain were made by applying a 1.5 micron optical signal modulated with a 33 MHz RF signal. The results of these measurements, as illustrated in figure 8, show higher gain, strong dependence on launched power, and considerable saturation at maximum signal power. Finally, the gain of the erbium doped fiber amplifier was estimated by correcting the maximum overall gain obtained for losses incurred in the bulkhead connectors and the fusion splices between the silica fiber patch cords and the erbium doped fiber. The peak gain cross section of the fiber was about twice that of the cross section at the signal wavelength. The peak gain of the amplifier with longer fiber was found to be approximately 20 db.

Ring Laser - The emission spectrum of the fiber laser without the feedback loop at 20 mw pump power in figure 9 shows maximum output centered at 1.533 micron. Figure 10 shows emission and lasing spectrum on the same laser output with the feedback loop, indicating the start of lasing around 1.555 micron. The output of the fiber laser with a single mode silica fiber in the feedback loop shows certain instabilities (figure 11) and mode mixing, which became relatively stable when the GRIN lens and Fabry-Perot etalog using a microscopic slide of 1 mm was included. The output in figure 12 showed a lesser number of modes centered aroung 1.5316 micron. Further tuning of the fiber laser output was achieved by adjusting GRIN lens and microscopic slide of 0.3 mm thickness as shown in figure 13. The results in this case show the presence of only two modes of lower power (about 160 nw) separated by about 23 nm wavelength, which is the tuning range. Fiber laser output as a function of pump power with and without etalon, as shown in figure 14, indicates a linear rise in the output power after the threshold current. The output of the fiber laser was also observed to contain an unmodulated signal with a relatively narrow spectrum of several modes and a periodic signal consisting of a series of short pulses associated with the relatively large spectral linewidth. This behavior is due to the spontaneous mode locking occurring in the output of this erbium doped fiber ring oscillator.

CONCLUSIONS

The project shows that the gain medium used in fiber amplifier can also be used in fiber lasers, which are proposed to replace solid state lasers as a signal source in wavelength division multiplexed fiber optic communications systems. Although higher gain in the fiber amplifier is desirable, such a gain produced instabilities at the amplifier output. It was also noted that some oscillations were produced at the output of the fiber laser due to bulkhead connectors used in the fiber ring oscillator. The output, however, became more stable when the connectors were replaced with fusion splices. The manually adjustable Fabry-Perot etalon, and tuning elements used to select wavelength and tune the laser, can be improved by using commercially available precision-controlled components. The relatively large spectral line width associated with the spontaneous mode locking can have a limiting effect on the tunability of the fiber oscillator used in wavelength division multiplexed communication systems. It is important to understand the mechanism of such mode locking and study its limitations on the tuning of fiber lasers.

REFERENCES

1. Emmanuel Desurvire, "Lightwave Communication: The Fifth Generation." Scientific American, p. 92-103, Jan. 1992.

2. K.J. Teegarden, S. Qazi, "Optical Fiber Amplifiers and Oscillators." AFOSR Summer Faculty Research Program, 1991.

3. B.E.A. Saleh and M.C. Teich, "Fundamental of Photonics." John Wiley & Sons, Inc., 1991.

4. Th. Pfeiffer, H. Schmuck, "Widely Tunable Actively Modelocked Erbium Fiber Ring Laser." Optical Amplifiers and Their Applications Topical Meeting, CO, 116/Th B4-1, July 24-26. 1992.

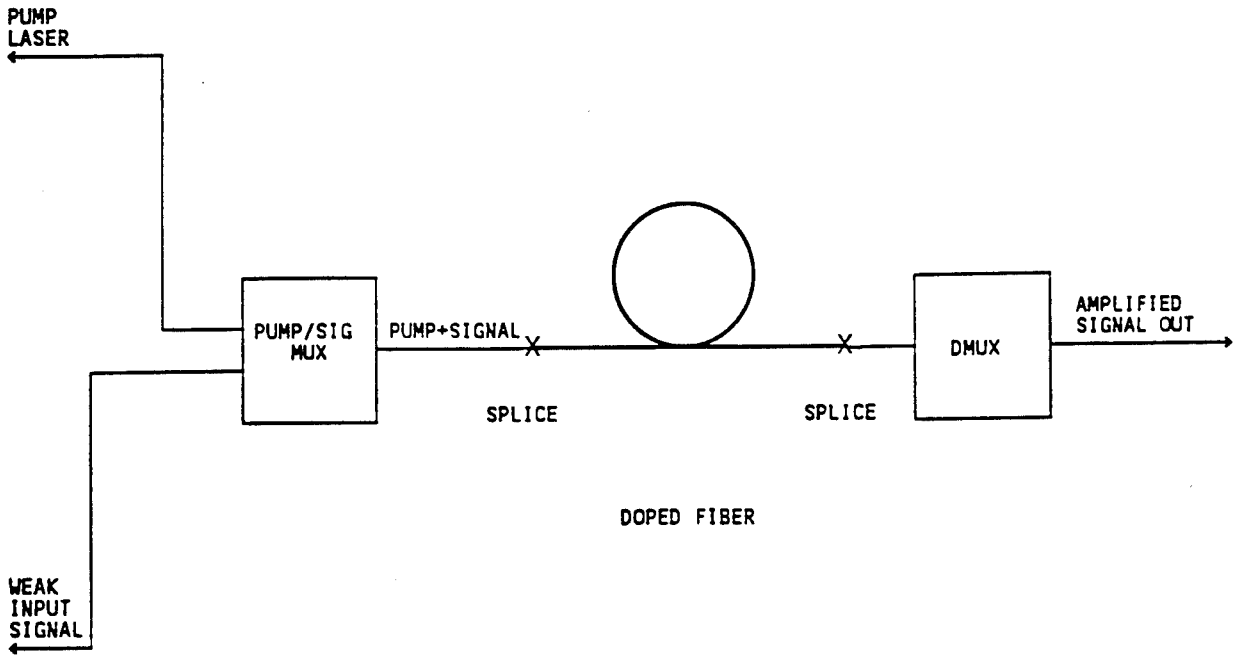


Figure 1 Erbium doped fiber amplifier configuration

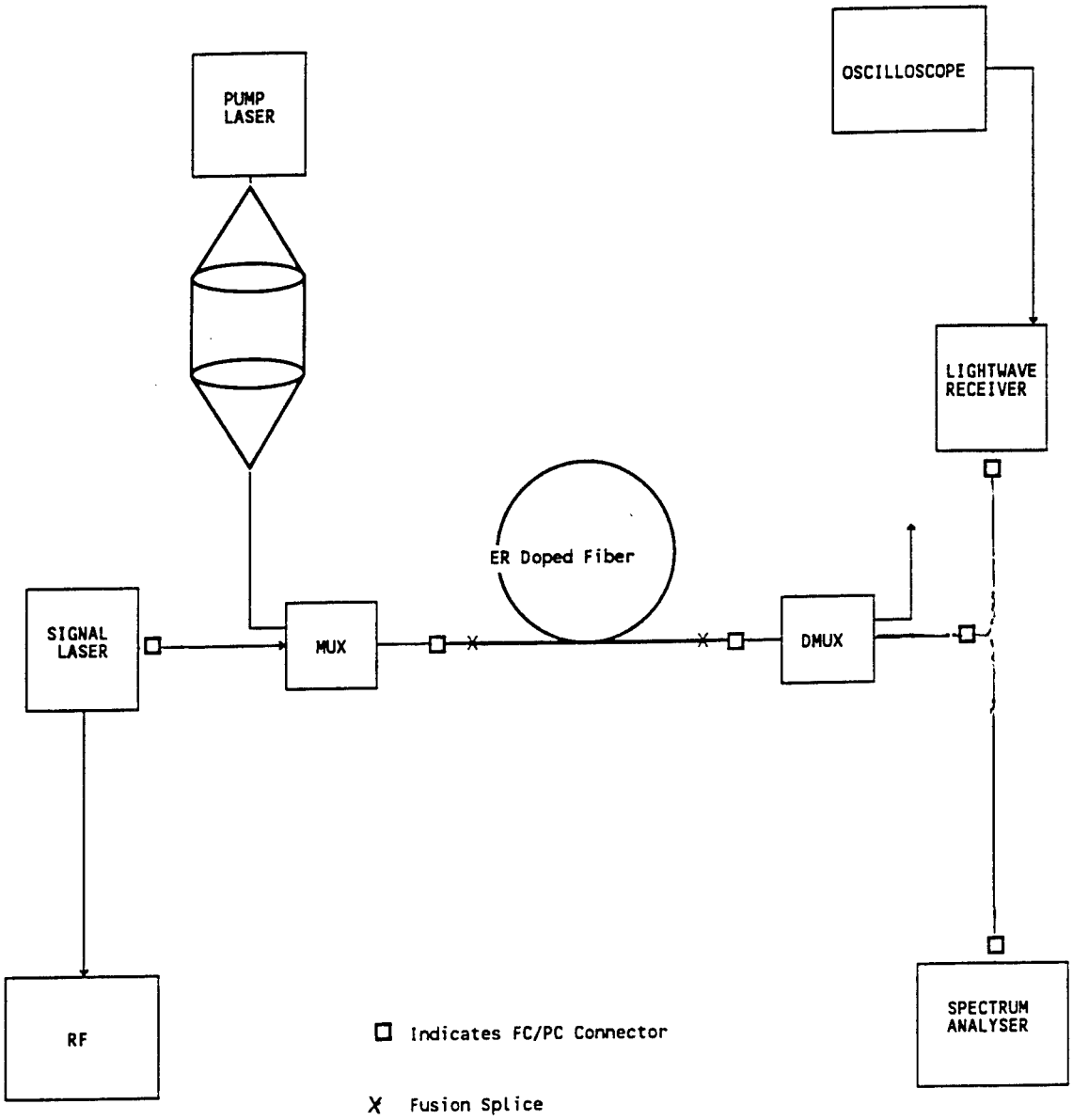


Figure 2 Experimental setup of erbium doped fiber amplifier system

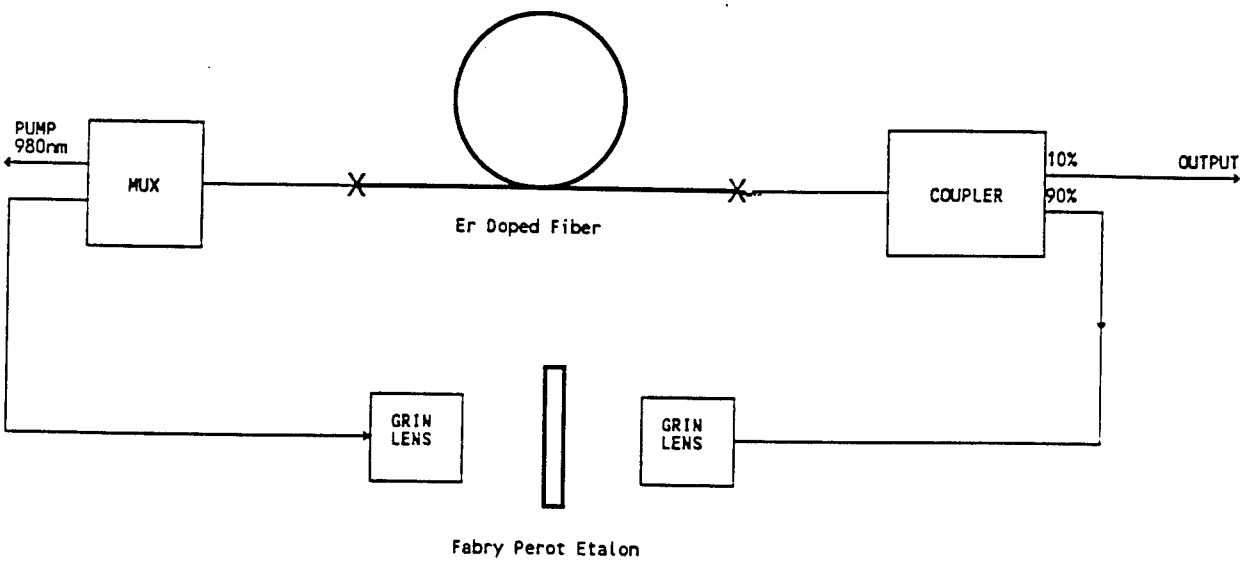


Figure 3 Tunable erbium doped fiber ring laser

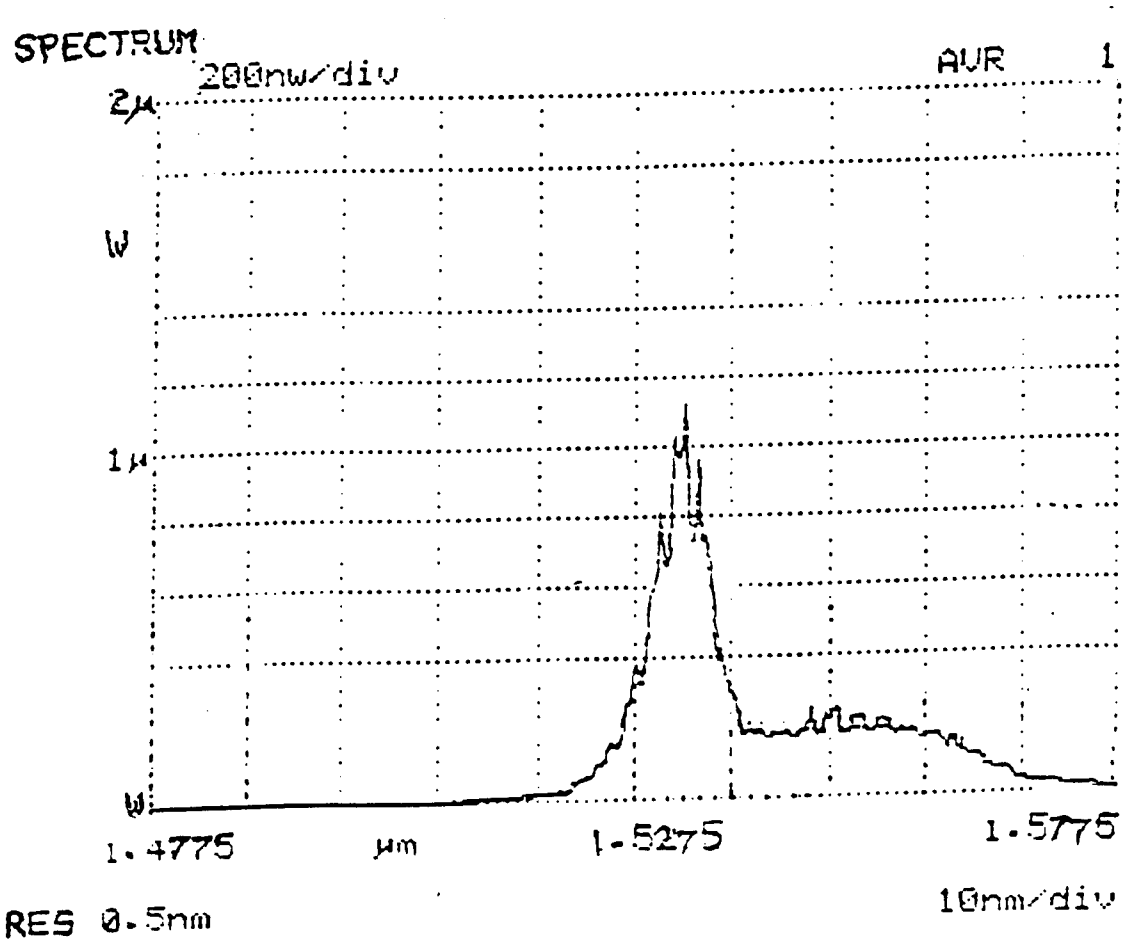


Figure 4 Emission spectrum of erbium doped fiber amplifier (wavelength versus output power)

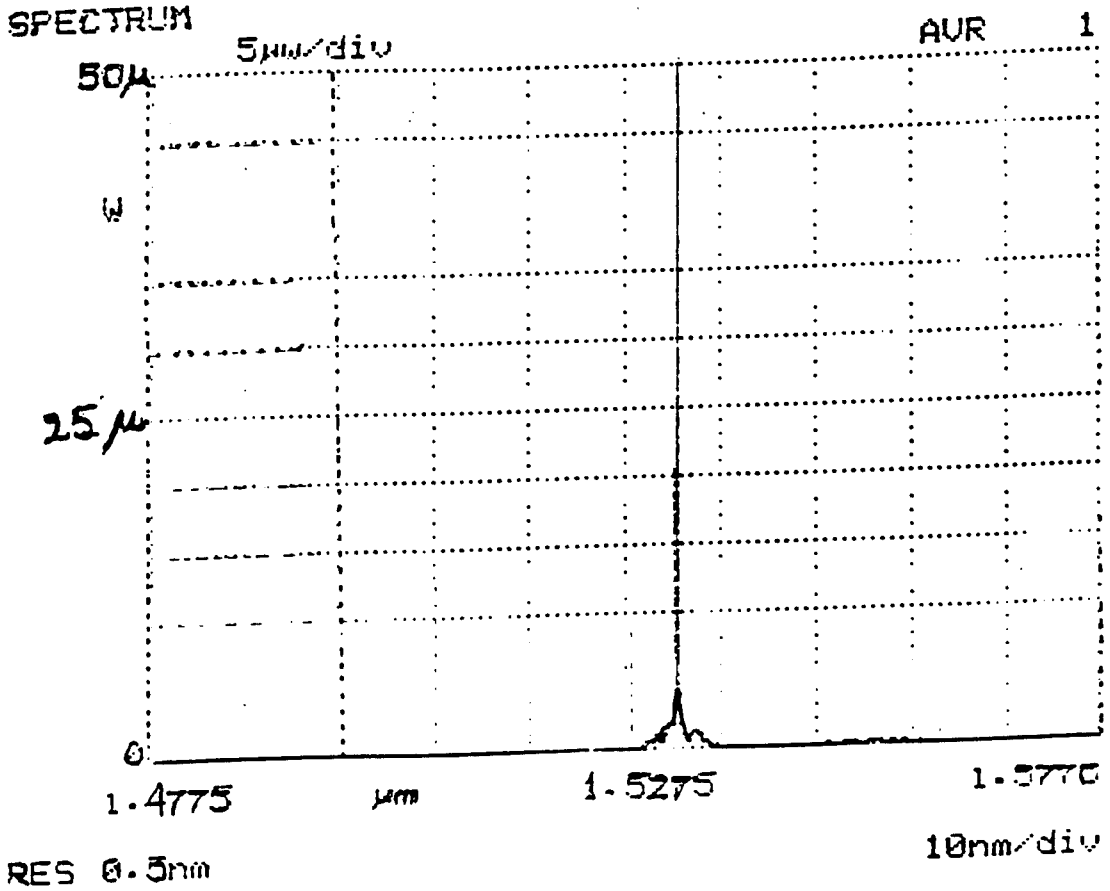


Figure 5 Optical spectrum of the fiber amplifier showing lasing at high pump power

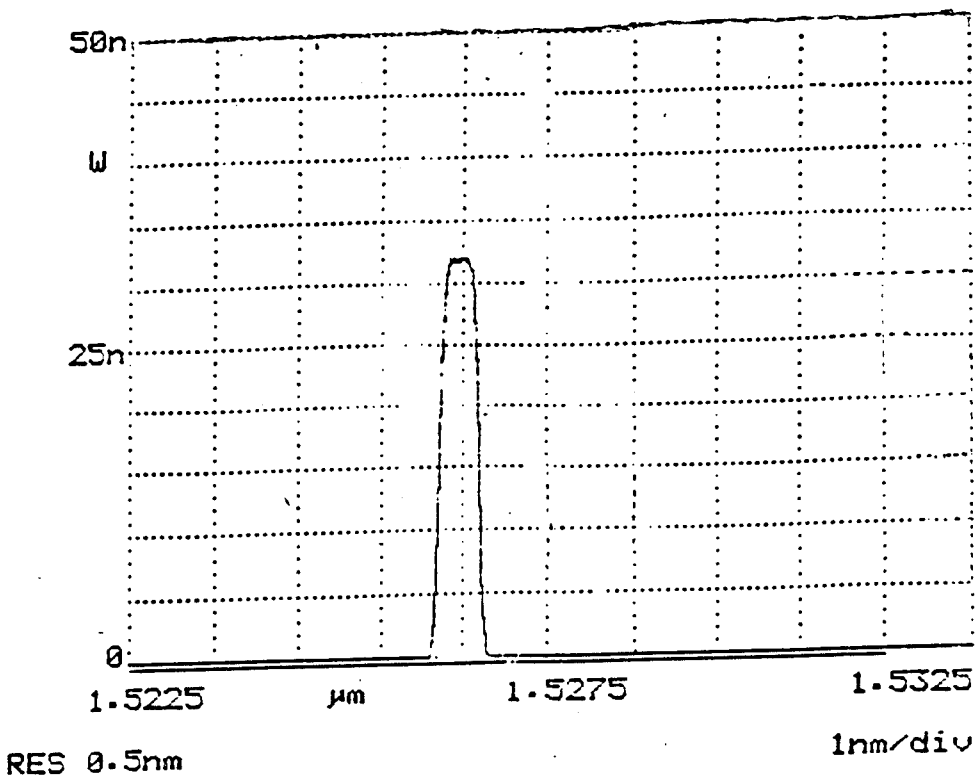


Figure 6a Optical spectrum of the input signal to the fiber amplifier

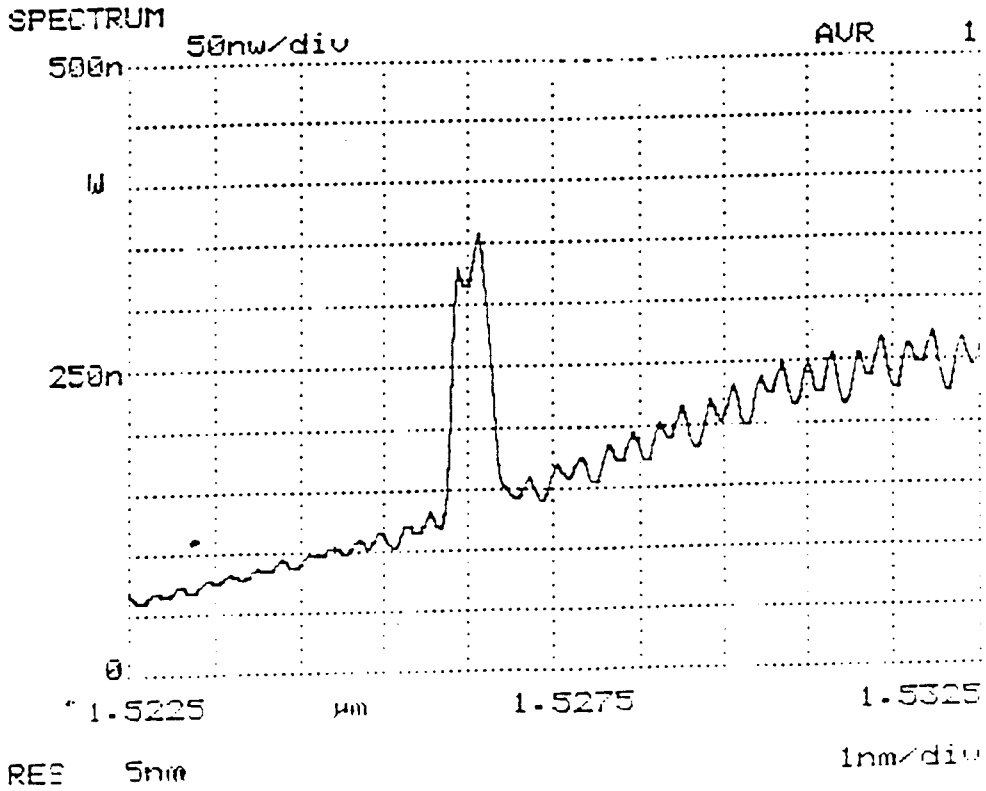


Figure 6b Optical spectrum of the output signal to the fiber amplifier

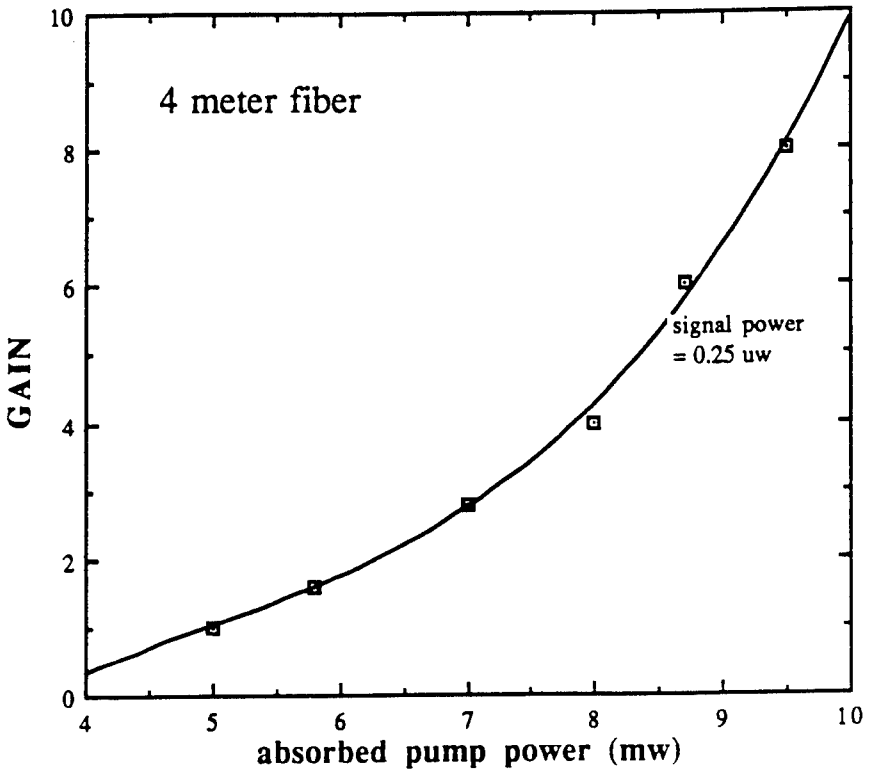


Figure 7 Amplifier gain versus pump power for 4 meter fiber

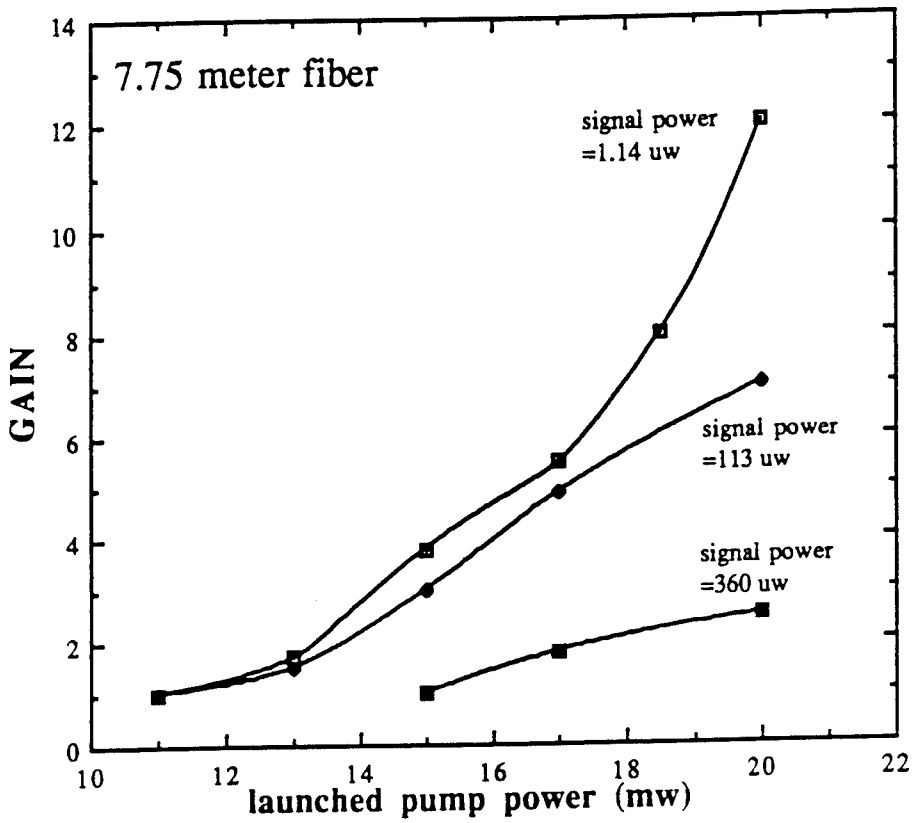


Figure 8 Amplifier gain versus pump power for 7.75 meter fiber

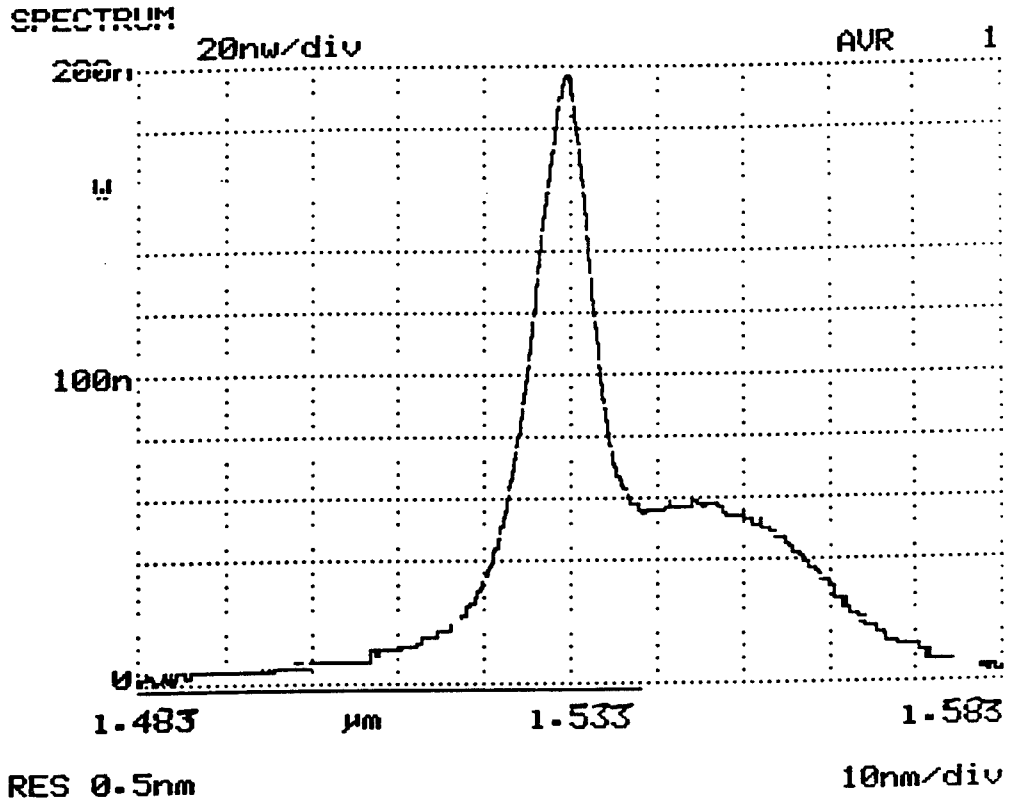


Figure 9 Emission spectrum of the fiber ring laser without the loop

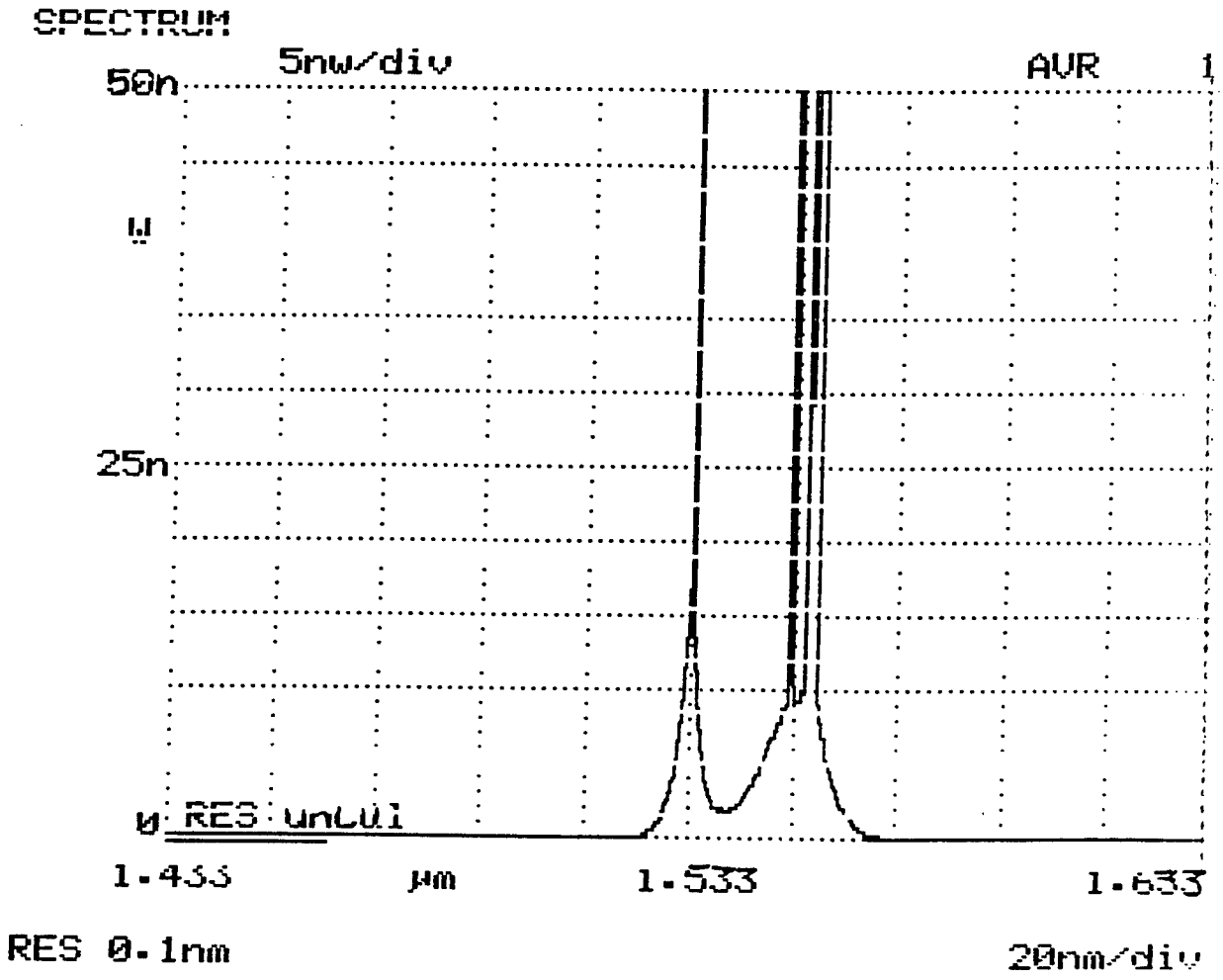


Figure 10 **Optical spectrum of the fiber ring laser showing emission and lasing**

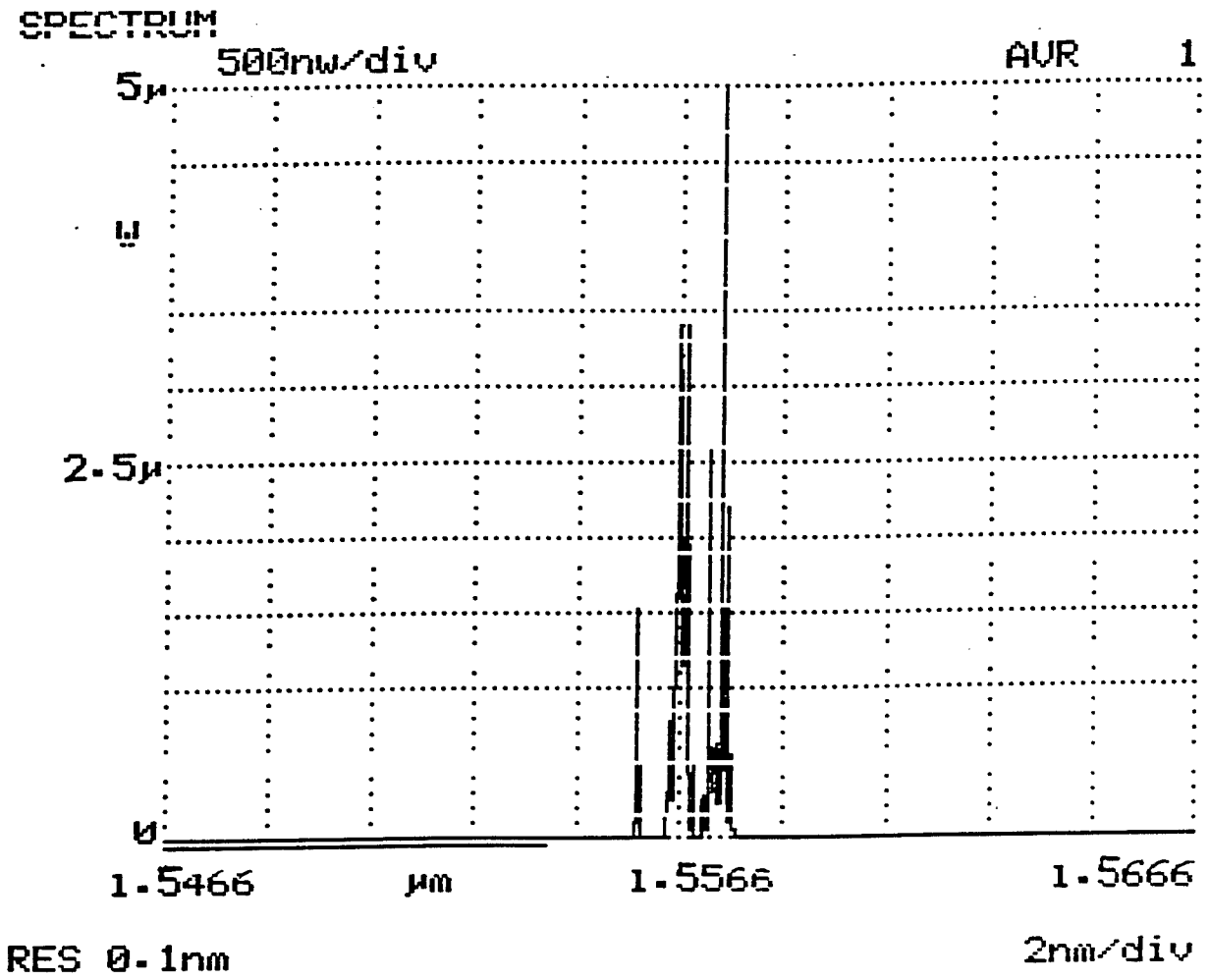


Figure 11 Optical spectrum of the fiber ring laser showing mode mixing and instabilities

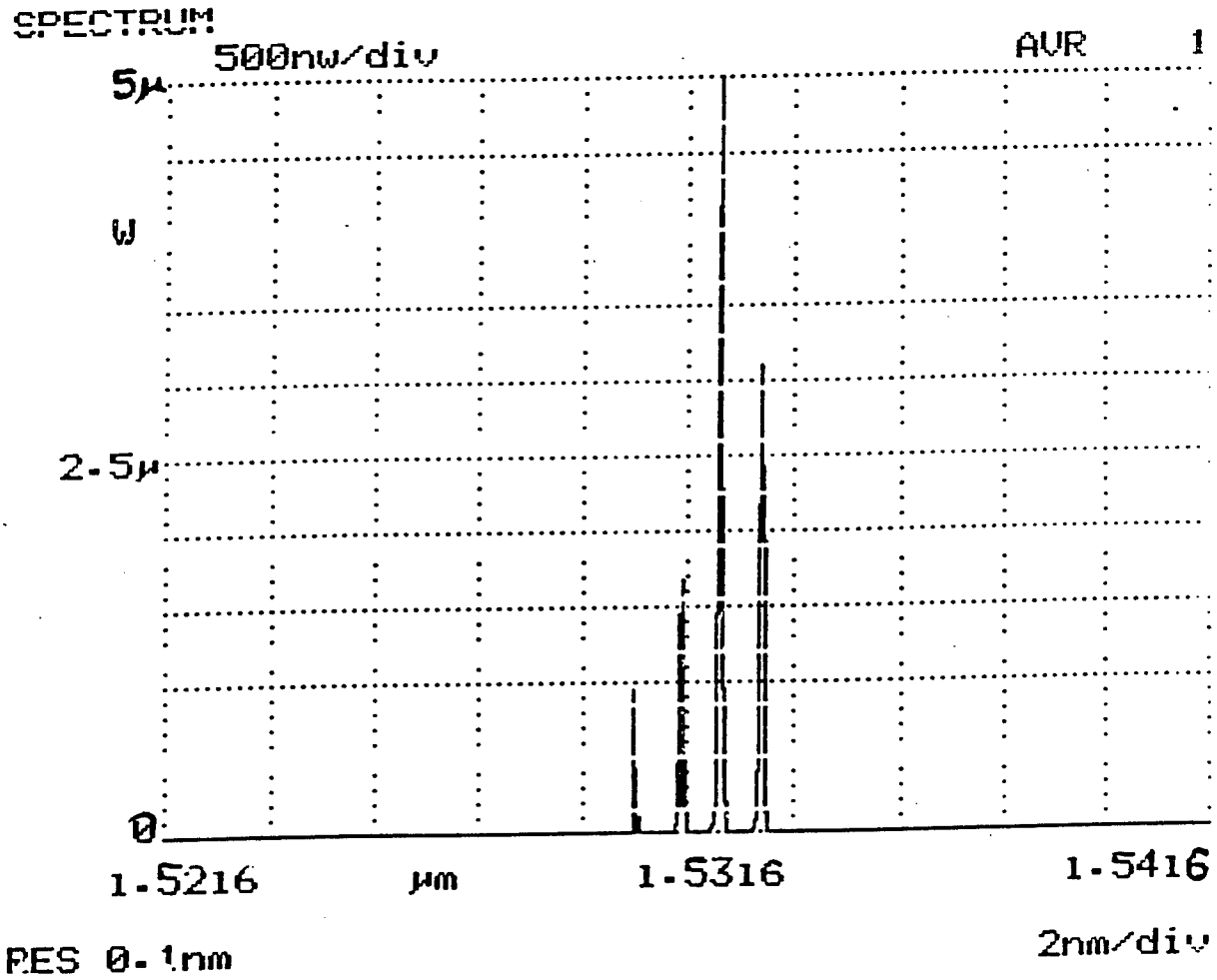


Figure 12 Optical spectrum of the fiber ring laser with the tuning elements

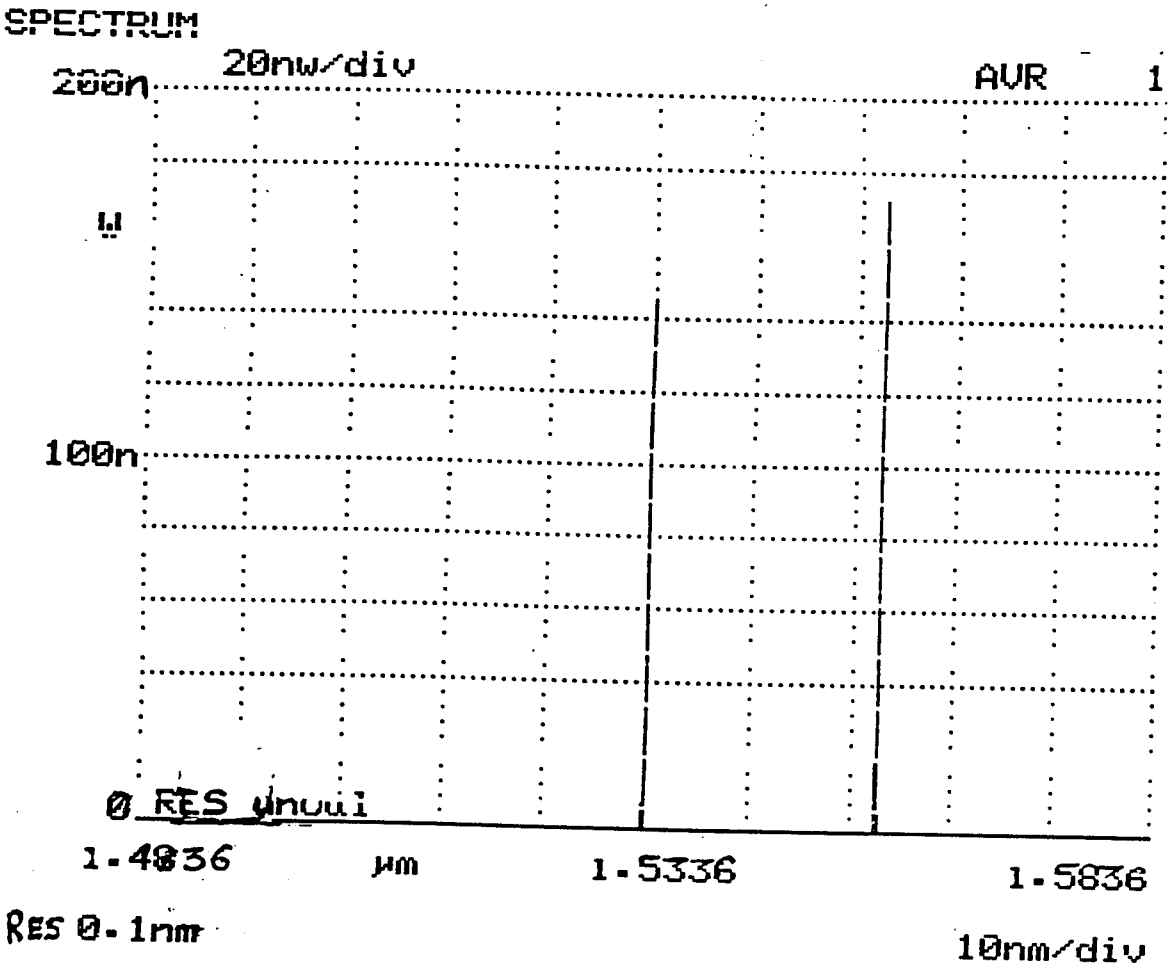


Figure 13 Optical spectrum of the fiber ring laser showing two tuned modes seperated by 23 nanometer

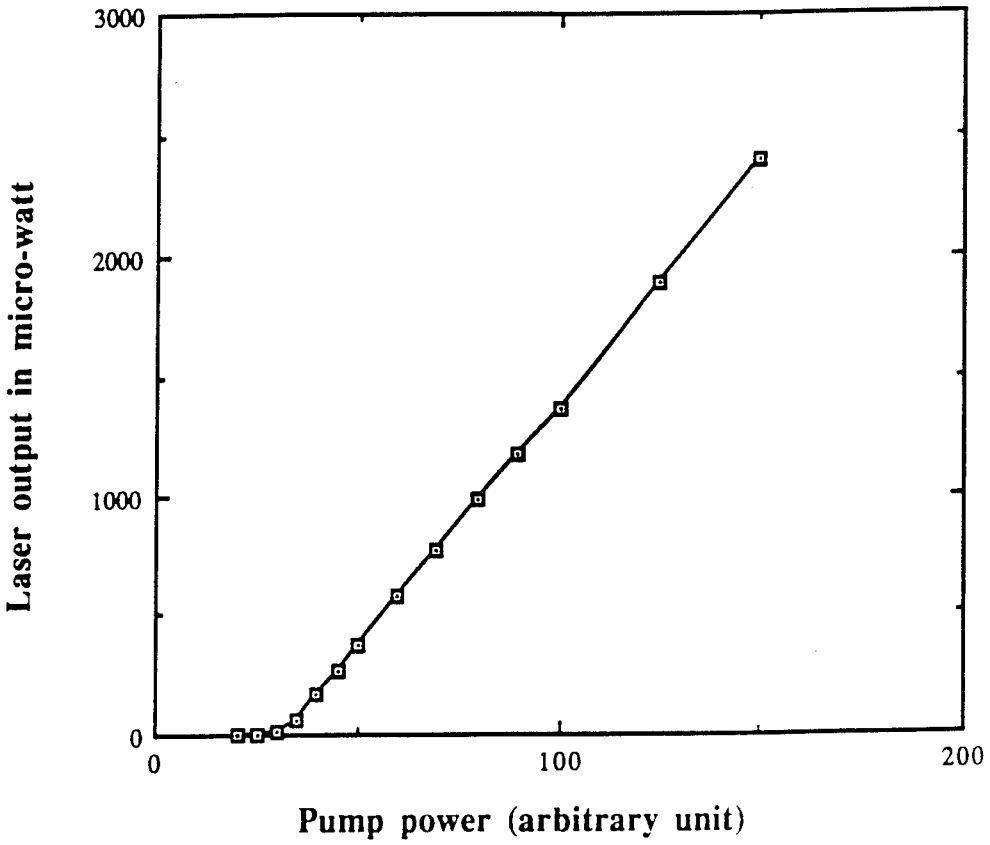


Figure 14 Laser output as a function of pump power

FDTD ANALYSIS OF THE RADIATION PROPERTIES OF A PARABOLIC CYLINDER ILLUMINATED BY A VERY SHORT PULSE

Carey M. Rappaport
Assistant Professor
Department of Electrical and Computer Engineering
Northeastern University
Boston, MA 02115

Final Report for:
Research Initiation Program
Rome Laboratories
Hanscom Air Force Base

Sponsored by:
Air Force Office of Scientific Research
Boiling Air Force Base, Washington, DC

January 1993

FDTD ANALYSIS OF THE RADIATION PROPERTIES OF A PARABOLIC CYLINDER ILLUMINATED BY A VERY SHORT PULSE

Carey M. Rappaport
Assistant Professor
Department of Electrical and Computer Engineering
Northeastern University
Boston, MA 02115

ABSTRACT

This final report completes the analysis of the problem, begun at Rome Laboratories at Hanscom Air Force Base during the Summer of 1991, concerning extremely short-time pulse excitation of a parabolic reflector antenna. We consider a pulse that is short compared to the time required for it to traverse the depth of the reflector, so that only a fraction of the reflector is illuminated at any given time. The best means of analysis of this transient problem is the Finite Difference Time Domain (FDTD) method.

The computer code written during the summer assignment has been modified and improved, and can now simulate a realistic, short, high-frequency modulated Gaussian amplitude profile cylindrical pulse. Field distributions were previously simulated for a deep parabola with an unmodulated Gaussian source in both transmit and receive modes. The current results show that for a source with a high-frequency modulating carrier, the two-dimensional parabola behaves just as expected, with no strange time artifacts causing complex behavior. Thus, the often perceived assumptions of differential transient reflection are inaccurate. A short-time pulse radiates from a parabola aperture in the same fashion as if it were radiating from comparably sized hole in a conducting screen.

One additional accomplishment of this project is in graphics visualization of highly oscillating waves. Because of the richness of available field data, it had previously been impossible to depict the modulated waves travelling in space and time. Using a smoothing scheme, the wave envelope is now available. This low-pass spatial filtering allows more easy interpretation of large scale wave propagation—an important aspect of this radiation analysis.

FDTD ANALYSIS OF THE RADIATION PROPERTIES OF A PARABOLIC CYLINDER ILLUMINATED BY A VERY SHORT PULSE

Carey M. Rappaport

INTRODUCTION

As signal generation technology improves, the need to understand the behavior of shorter and shorter pulses becomes increasingly important. When the duration of the pulse becomes comparable to—or shorter than—the typical length scale of an antenna system, new ways of studying electromagnetic field propagation become necessary.

One important problem that has recently been studied [1,2] is the excitation of a parabolic reflector antenna by a pulse that is shorter than its focal length. The traditional analysis method of tracing rays is brought into question since the rays do not all illuminate the reflector simultaneously.

In fact, Hansen first states [1] that the leading and trailing edges of a short pulse have different antenna patterns, leading to a differential transient radiation characteristic. The leading portion of the transmitted wave is first reflected by the reflector surface in a circle nearest the vertex while the reflector rim is unilluminated, while subsequently, the trailing pulse edge has left the vertex it still illuminates an annular region near the rim. See Figure 1. The conclusion is that the initial radiation pattern is that of a smaller diameter aperture, while the final pattern is that of an annulus with constant outer diameter.

This view was later reversed [2] with an argument that rays diffracted from the edges of the reflector leave the antenna first, before reflected rays from the vertex. The transient pattern should instead be that of an annular aperture first, followed by the gradually shrinking circle.

The object of the current study is to show that neither conclusion is accurate, but instead that the parabolic reflector has no characteristic transient behavior other than a simple aperture.

The Finite Difference Time Domain (FDTD) method is often used to simulate the time behavior of short pulses. It is not based on extending frequency domain analysis across a broad frequency spectrum, but instead formulates the wave propagation completely in the time domain. The shorter the time pulse, the easier it is to model the field interaction.

Although it must be cautioned that using numerical methods to prove theoretical assertions is dangerous, it is the transient behavior that is of concern here, and only a procedure which emphasizes the different temporal aspects of the field can be effectively employed.

A previous report [3] analyzed both transmit and receive pulses that had unmodulated Gaussian spatial amplitude profiles, and a modulated Gaussian plane wave receive pulse, all assumed to be transverse magnetic (TM) polarization. The results for the 2-dimensional parabolic reflector were consistent with long established intuition, that the parabola converts a circularly cylindrical wave to a plane wave regardless of the time duration of the incident pulse.

This report describes the final, most realistic, and most important case: a high-frequency modulated short-time Gaussian pulse, transmitted from a source at the parabola focal point. The problem is again cast as 2-dimensional, with TM excitation. We expect a relatively uniform plane wave reflected from the parabola.

THE FDTD METHOD OF WAVE ANALYSIS

To study the transient wave behavior, a time domain numerical method: Finite Difference Time Domain (FDTD) is used. This method was first proposed on a cubical lattice by Yee [4], and then, more recently, popularized by others [5-7]. Other Finite Difference techniques, which make use of triangular, conformal meshes in both the time and frequency domains have been presented [8-10].

The idea behind FDTD is that Maxwell's partial differential equations are replaced by multi-dimensional centered difference equations in space and time. First, space is discretized into two sets of interlocking cubical meshes, with the cube corners of one mesh coinciding with the cube centers of the other. The magnetic flux of Faraday's Law is solved at the center of every cube face in terms of the electric field on the edges that bound that face; and the electric flux of Ampere's Law is solved at the center of every cube face of the complimentary mesh in terms of the magnetic fields on the edges that bound it. With isotropic media the electric flux and the electric field are related by a scalar permittivity constant, ϵ, so that knowing the first from Ampere's Law allows the computation of the magnetic flux from Faraday's Law. The field values on one mesh are computed at one-half a time step ahead of those of the other, so that the difference equation in time at one centered space point can be equated to the difference equation in space at the one centered time point of the complimentary mesh.

One important detail, which must be carefully addressed for exterior, unbounded geometries, is simulating the wave's radiation to infinity. Although the computational domain terminates at the end of the array lattice, the wave must act as if it were continuing outward. In other words, the wave must not reflect off the artificial boundary. The boundary must absorb all incident waves; hence it is referred to as an Absorbing Boundary Condition (ABC). In principle, no numerical condition can absorb waves from every incident angle, but there have been several ABC's proposed which absorb those near normal incidence [11-15] which makes use of pseudo-differential annihilation operators. A new ABC method which is based on simulated anechoic chamber absorber has recently been presented [15]. This method appears to be preferable for very large scatterers and higher frequencies.

NUMERICAL SIMULATION ANALYSIS

For the parabolic reflection problem, a two-dimensional FDTD method was implemented on a vector-processor enhanced Digital Equipment Corporation VAX 9000 computer. The ABC used is the Engquist-Majda [12] planar condition on forward, backward, right and left edges of the computational domain. The parabolic segment is chosen to be four times as wide as deep ($F/D = 0.25$). This is typical of deep reflector antennas, with focus in the plane of the reflector rim.

The Courant Condition, that the time step multiplied by phase velocity of the wave must be less than the spatial step, is satisfied by selecting: $r = c\Delta t/\Delta x = 0.5$. This means that a wavefront propagates $n/2$ space steps in n time steps. The computational domain used is 600 by 600 space steps, so for a pulse to travel from the first row to the last row and then back again would require 2400 time steps, amounting to a total of 864,000,000 computations each of E_z, H_x, and H_y. It is clear that supercomputers are particularly useful for this type of problem. Massively parallel platforms are well suited to FDTD calculation, since the three field components are calculated for the entire grid at each time step.

The circular cylindrical source pulse is a Gaussian modulated with a cosine, by given at $t = 0$ by:

$$\overline{E} = \hat{z}\frac{e^{-(\frac{\rho-\rho_0}{W})^2}}{\sqrt{\frac{\rho}{\rho_0}}}\cos k(\rho - \rho_0) \tag{1}$$

and for magnetic field:

$$\overline{H} = \hat{\phi}\frac{e^{-(\frac{\rho-\rho_0}{W})^2}}{\eta\sqrt{\frac{\rho}{\rho_0}}}\cos k(\rho - \rho_0) \tag{2}$$

The parameters chosen for the simulation are: $\Delta x = \Delta y = \Delta$, $W = 20\Delta$, $\rho_0 = 100\Delta, \eta = 377$ Ohms, and the modulating wave number, $k = \pi/8$ (corresponding to 16 grid points per wavelengths).

These initial wave conditions are approximations to an actual cylindrical wave at a given time sample. They are the asymptotic (farfield) cylindrical pulse wave solutions to the wave equation, normalized to unity at $\rho = \rho_0$. The wave representations do not have to be exact for

a cylindrical wave to propagate. Once the algorithm starts, the laws of wave propagation are strictly followed. Any imperfections in the initial conditions will lead to an inward-travelling wave, or a change in the outward cylindrical wave shape. In fact, a small inward-travelling wave with about 5% of the initial amplitude is generated. This counter-propagation artifact can be eliminated using an absorbing circular disk of a matched lossy medium with a smaller radius than $\sim \rho_0 - 3W$ [16] coaxially located with the pulse.

Figure 2 shows a three-dimensional surface plot of the electric field, E_z, given in Equation (1). Only the field data sampled every 12 points in x and y are shown. Clearly, this highly modulated cylindrical pulse is not well depicted. In particular, the wave appears discontinuous, and the maximum value in the plot of Figure 2 is only 85 % of its true value. The next section describes an improved graphics alternative which alleviates this deficiency.

The parabolic reflector is given by the equation:

$$y = \frac{x^2}{4f} - f \tag{3}$$

where $f = 200\Delta$ is the focal length. It extends across 2/3 of the width of the computational domain with an aperture of 400 spatial points. The focal point is positioned at the center of the grid at point $(i, j) = (300, 300)$.

GRAPHICAL PULSE SMOOTHING

In the first attempts to analyze the cylindrical, modulated Gaussian pulse, the sheer volume of data made it quite difficult to visualize the wave behavior. The standard method of graphical display, three-dimensional surface plots, while effective in showing unmodulated pulses, fails when a high-frequency carrier is present. The problem comes from the trade-off between displaying all the data, resulting in an impossibly dense grid with 600 by 600 crossing lines; and an insufficient undersampling of the data, which misses many of the essential details. It was decided that a 50 by 50 3-dimensional surface grid is the maximum density for easy visualization, implying that only every twelfth point is displayed.

An effective means of identifying the large-scale behavior of a modulated wave is to detect its envelope by low-pass filtering. This is the same well-established technique sometimes used in AM radio signal demodulation. In the current application, the wave signal field data is available for spatial points rather than as a function of time, so the smoothing involves low-pass spatial filtering.

Rather than using a formal Fourier transform or convolution to filter the spatial data, the envelope can be simply detected by finding the absolute values of the modulation peaks, and then connecting these values with straight line segments. As long as the modulation is fast enough so that the changes between peaks are small relative to the peak values, using straight segments to join the peaks does not introduce much error. There are 5 peaks within the full Gaussian pulse width, $2W$, so the error is minimal. The line segments are found in both x- and y-directions, yielding first-order finite element patches covering the absolute value of the modulated 2-dimensional spatial waveform.

Once the surface patches are found, the intermediate sample points are easy to find using bi-linear interpolation. The result of smoothing and interpolating the data of Figure 2 is shown in Figure 3. Note that the wave is now circular, continuous, and almost attains its true maximum value: 97.9 %. This plot does not depict the modulation at all, but shows the overall envelope shape. Used in conjunction with Figure 2, the smoothed data gives considerable information about this highly modulated extremely numerically dense wave.

NUMERICAL RESULTS

Figures 4 through 15 show the progression of the modulated cylindrical pulse as it expands, interacts with the parabolic reflector, and then forms a backward propagating plane wave. Each page shows two plots of the same time "snap-shot" of the wave. The even numbered figures are the undersampled surface plots, while the odd numbered figures show the smoothed envelope of the wave. The time samples occur every 150 time steps. Note that the rate of expansion of the wave between Figures 2 and 4 (for example) is about an eighth of the total computational grid, or about 75 space steps, as expected. It is also seen that even though the envelope plot is much more smooth than the modulated plot, the Gaussian is still quite narrow compared to the entire domain, and as such appears to have a ragged peak.

At time step $300\Delta t$ the pulse has encountered the central portion of the reflector. Figure 6 clearly shows that an interaction is occurring, but one must inspect the envelope in Figure 7 to see that a plane wave is being reflected. The maximum values at the edges of the reflected plane wave are due to the standing waves formed when an inverted reflected peak constructively interferes with the negative portion of the incident wave. In Figure 8, the reflected plane wave is clearly formed, with uniformity across a single modulation peak. Also visible are the interactions of the incident wave with the domain boundaries. Most of the wave incident on the boundaries is absorbed with the Engquist-Majda ABC, seeming to just vanish. Figures 10 and 11 however, show that there are slight reflection from these edges. At $750\Delta t$, shown in Figures 12 and 13, the incident cylindrical wave is no longer apparent. All that remains is the parabolically reflected uniform plane wave and the ABC artifacts.

This sequence of plots clearly indicates that the parabola generates an almost uniform, transmitted plane wave with envelope width comparable to that of the incident source radial pulse width, much as would be expected.

CONCLUSIONS

An analysis of the reflection characteristics of parabolic reflectors to short pulses has been completed. Contrary to published theory, the parabola does not produce any unexpected transient effects beyond the standard dispersive effects of free-space propagation of a short pulse.

The FDTD method was used to analyze the transient behavior of the two-dimensional parabolic reflector illuminated with a short TM pulse. In the numerical case considered in this report, as well as in the three previous cases, the wave behaved as expected, focusing to high intensity at the parabola focal point, or forming a planar transmitted phase front. The reflector seems to have no differential effects on the various time samples of the pulse—the reflection of the initial, leading edge of the pulse joins correctly with the subsequently reflected trailing portions of the pulse.

There will be the same transient behavior of a wave pulse for parabolic reflection as for propagation in free space, but the parabola itself does not cause any transient non-uniformities in phase.

The enhancement of envelope detection by low-pass spatial filtering improves the clarity of graphic visualization. Using this option, the bulk behavior of the wave is more clearly visible. The high resolution, detailed images are still required, however, to give the more precise information required for careful analysis of individual wave peak interactions with the reflector.

There may be some concerns about the choice of a two-dimensional reflector illuminated with transverse magnetic field. It is much more complicated to develop the three-dimensional FDTD algorithms needed to analyze a circular paraboloid of revolution. Furthermore, the transverse electric (TE) polarization is hard to model accurately, since the electric field becomes infinite at sharp edges (like the edge of the reflector). These are important limitations of the current study. However, the claim that there is a differential transient effect due to the reflector [1,2] also does not address wave polarization, but relies solely on a path length

analysis. For this transient claim to be valid, it must apply to a 2-dimensional parabolic reflector and to all polarizations. Our FDTD analysis provides a clear counter-example to the claim.

REFERENCES

1. Hansen, R., "Short-Pulse Excitation of Reflector Antennas," *IEE Proceedings*, vol. 134, Pt. H, no. 6, December 1987, pp. 557-559.

2. Hansen, R., "Short-Pulses in Reflectors Revisited," *IEE Proceedings*, Pt. H.

3. Rappaport, C., "FDTD Analysis of the Radiation Properties of a Parabolic Cylinder Illuminated by a Very Short Pulse," *1991 AFOSR Summer Research Program, Final Report*, Rome Labs, Hanscom AFB, September, 1991.

4. Yee, K.S., "Numerical Solution of Initial Boundary Value Problems Involving Maxwell's Equations in Isotropic Media", *IEEE Transactions on Antennas and Propagation*, vol. AP-14, 1966, pp. 302-307.

5. Taflove, A., and Umashankar, K., "The Finite-Difference Time-Domain (FDTD) Method for Electromagnetic Scattering and Interaction Problems," *Journal of Electromagnetic Waves and Applications*, vol. 1, 1987, pp. 243-267.

6. Borup, D., Sullivan, D., and Gandhi, O., "Comparison of the FFT Conjugate Gradient Method and the Finite-Difference Time-Domain Method for the 2-D Absorption Problem," *IEEE Transactions on Microwave Theory and Techniques*, vol. MTT-35, 1987, pp. 383-395.

7. Luebbers, R., Hunsberger, F., and Kunz, K., "FDTD Formulation for Frequency Dependent Permittivity," *AP-S Sym. Digest* June 1989, pp. 50-53.

8. Ling, R., "Application of Computational Fluid Dynamics Methods to a Numerical Study of Electromagnetic Wave Scattering Phenomena," *Journal of Applied Physics*, vol. 64, 1988, pp. 3785-3791.

9. Rappaport, C. and McCartin, B., "FDFD Analysis of Electromagnetic Scattering in Anisotropic Media Using Unconstrained Triangular Meshes," *IEEE Transactions on Antennas and Propagation*, vol. AP-39, no. 3, March 1991, pp 345-349.

10. Rappaport, C. and Smith, E., "Anisotropic FDFD Computed on Conformal Meshes." *Fourth Biennial IEEE Conference on Electromagnetic Field Computation*, October 1990, p. BB-10.

11. Bayliss, A., Gunzburger, M., and Turkell, E., "Boundary Conditions for the Numerical Solution of Elliptic Equations in Exterior Regions," *SIAM Journal of Applied Mathematics*, vol. 42, 1982, pp. 430-451.

12. Engquist, B., and Majda, A., "Absorbing Boundary Conditions for the Numerical Simulation of Waves," *Mathematical Computation*, vol. 31, 1977, pp. 629-651.

13. Kriegsmann, G.A. and Morawetz, C.S., "Solving the Helmholtz Equation for Exterior Problems with Variable Index of Reflection", *SIAM Journal of Sciences, Statistical Computation*, vol. 1, 1980, pp. 371-385.

14. Lee, C., Shin, R., Kong, J., and McCartin, B.J., "Absorbing Boundary Conditions on Circular and Elliptic Boundaries," *Prog. in Electromag. Research Sym. Proceedings*, July 1989, pp. 317-318.

15. Rappaport, C., and Bahrmasel, L., "An Absorbing Boundary Condition Based on Anechoic Absorber for EM Scattering Computation," *Journal of Electromagnetic Waves and Applications*, vol. 6, no. 12, December 1992, pp. 1621-1634.

16. Rappaport, C., "Preliminary FDTD Results form the Anechoic Absorber Absorbing Boundary Condition," *1992 IEEE AP-S Symposium Digest*, July 1992, pp. 544-547.

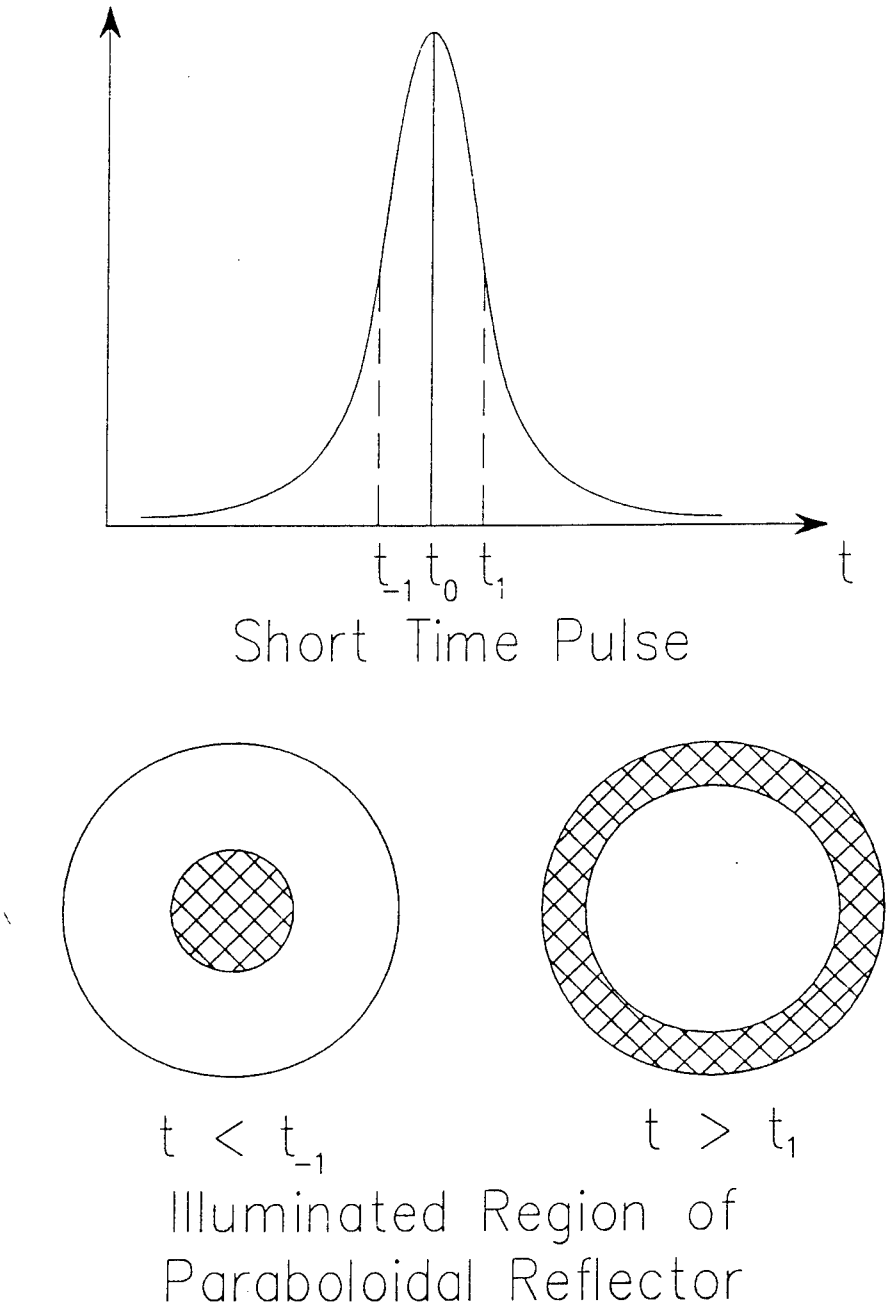


Figure 1: A short pulse in time, and the different illuminated regions of a paraboloidal reflector surface.

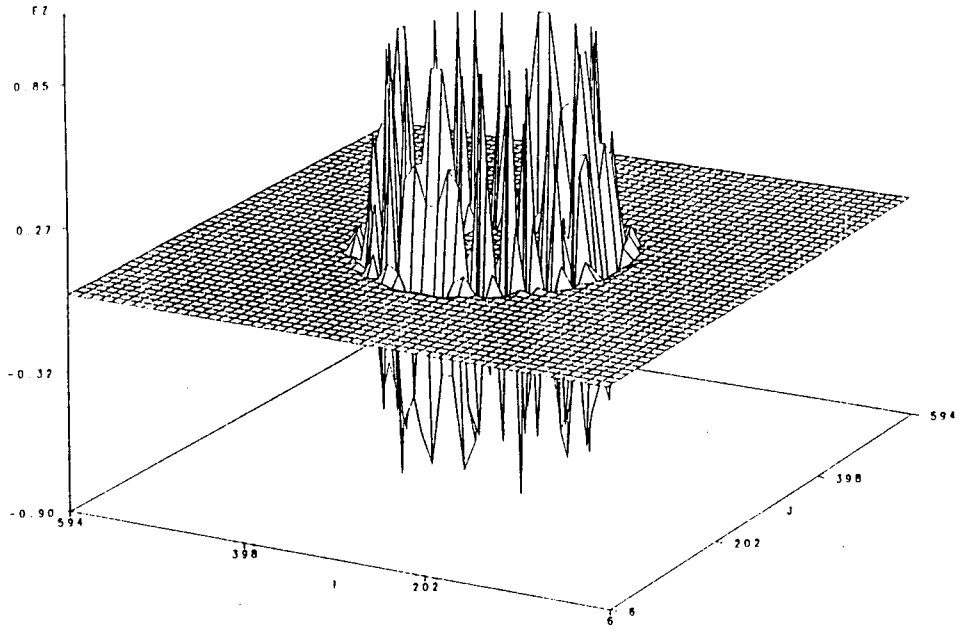


Figure 2: The initial ($t = 0$) modulated Gaussian cylindrical pulse, centered at the parabola focal point, sampled at every twelfth data point.

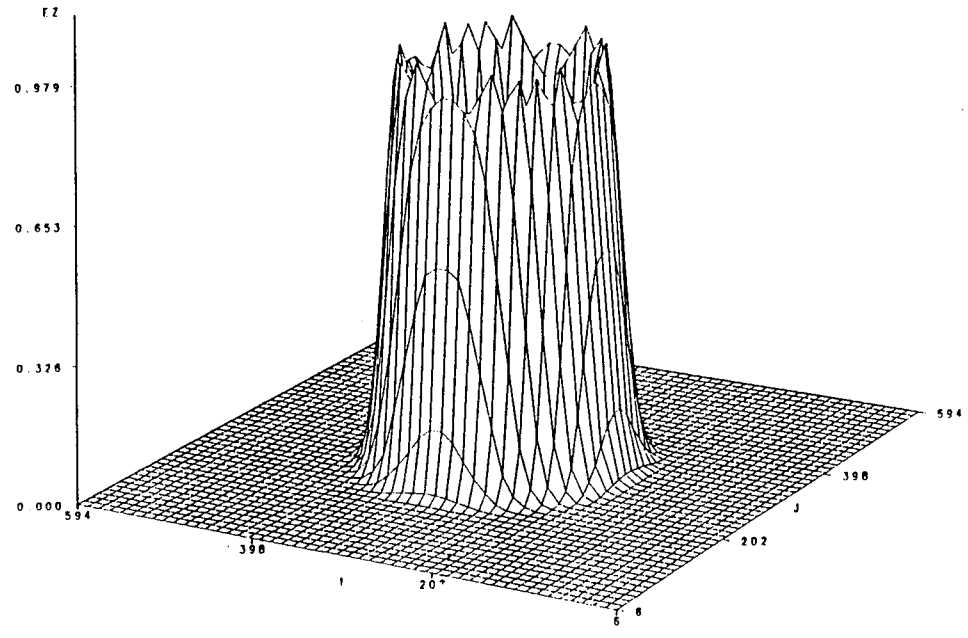


Figure 3: Low-pass spatially filtered and interpolated data of Figure 2, indicating the Gaussian envelope of the cylindrical wave.

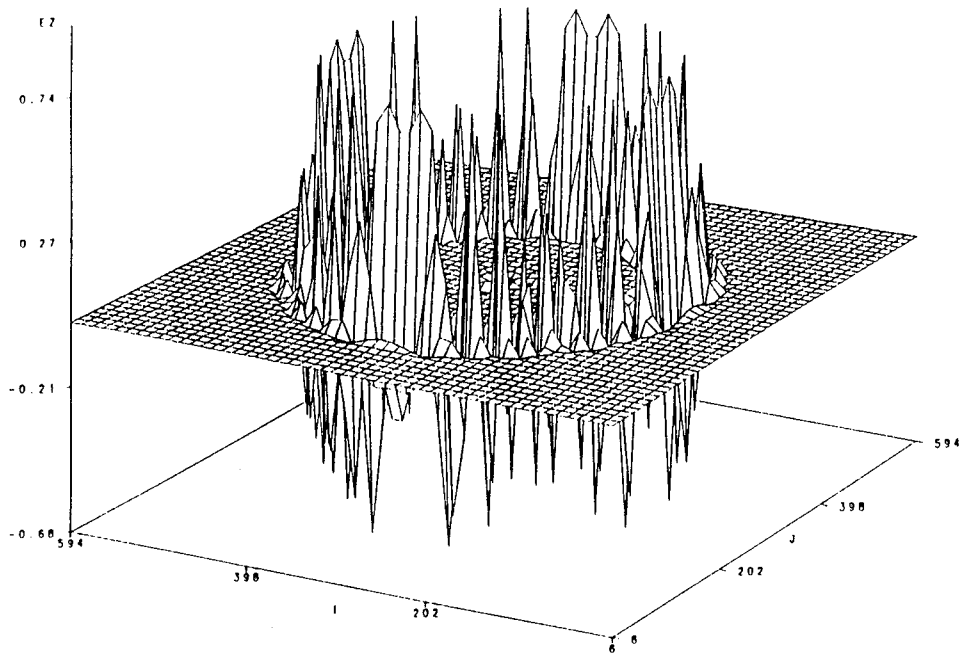


Figure 4: Sampled modulated wave at $t = 150\Delta t$.

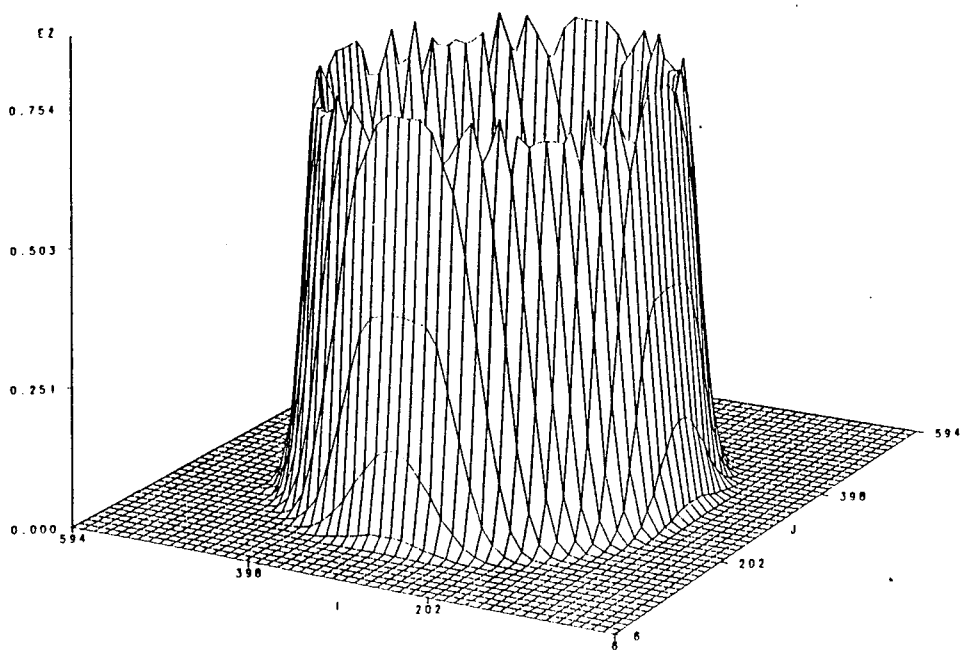


Figure 5: Smoothed version of the wave at $t = 150\Delta t$

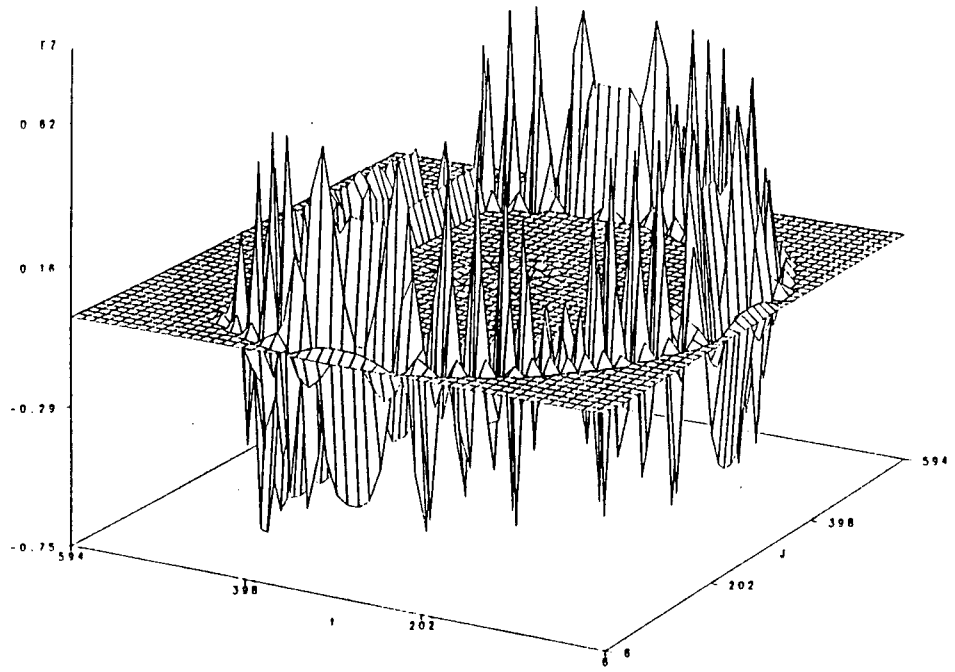


Figure 6: Sampled modulated wave at $t = 300\Delta t$.

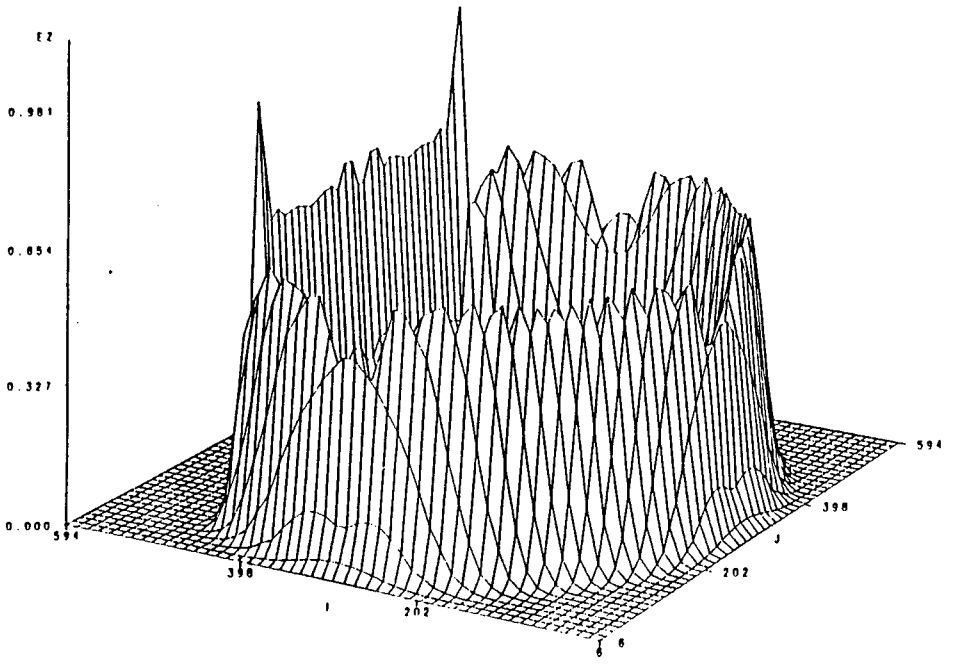


Figure 7: Smoothed version of the wave at $t = 300\Delta t$

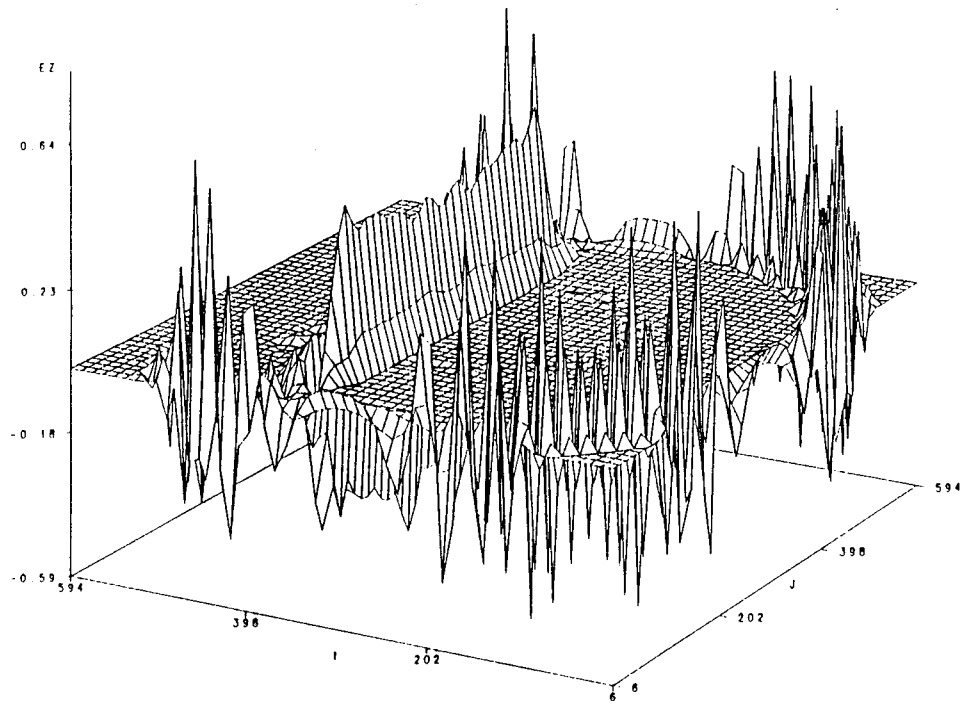


Figure 8: Sampled modulated wave at $t = 450\Delta t$.

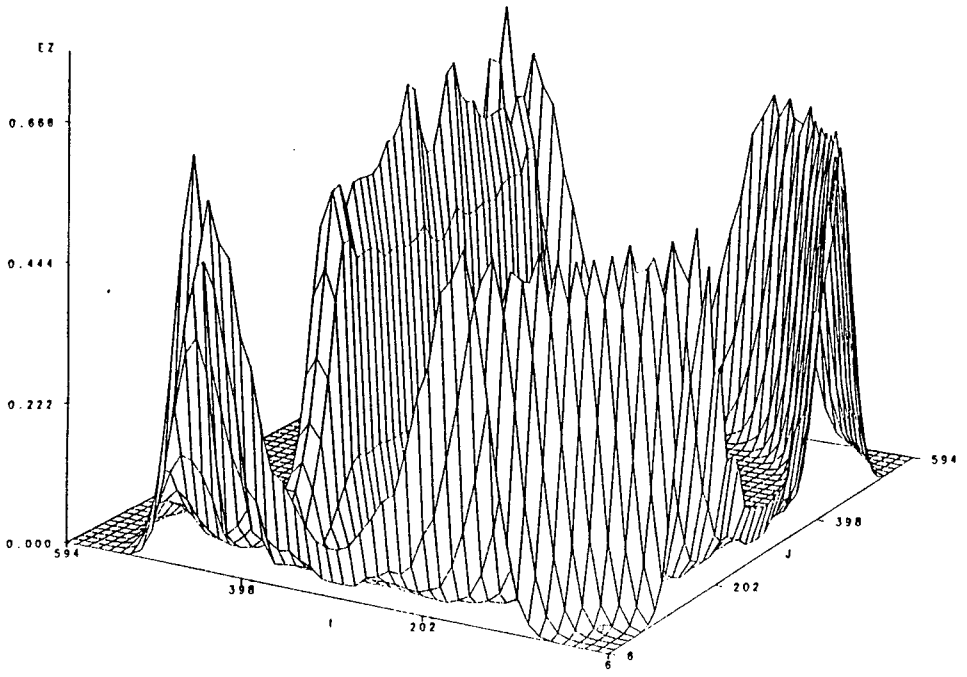


Figure 9: Smoothed version of the wave at $t = 450\Delta t$

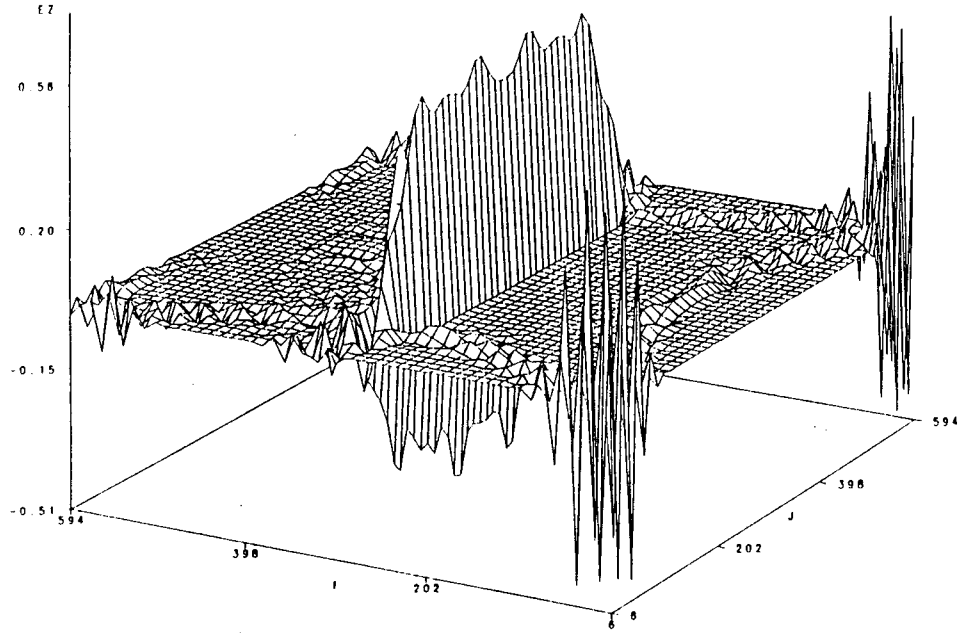


Figure 10: Sampled modulated wave at $t = 600\Delta t$.

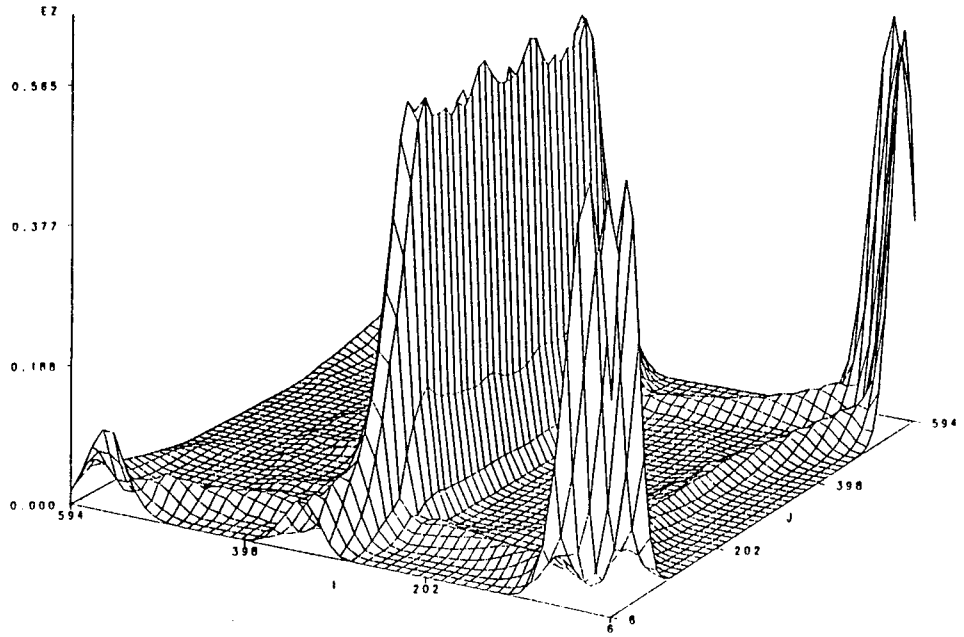


Figure 11: Smoothed version of the wave at $t = 600\Delta t$

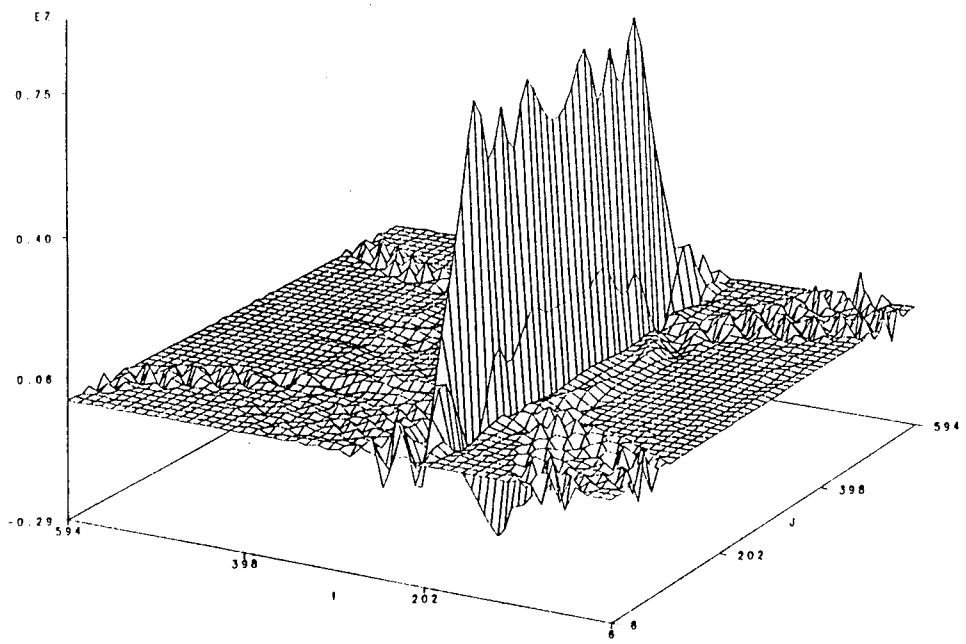


Figure 12: Sampled modulated wave at $t = 750\Delta t$.

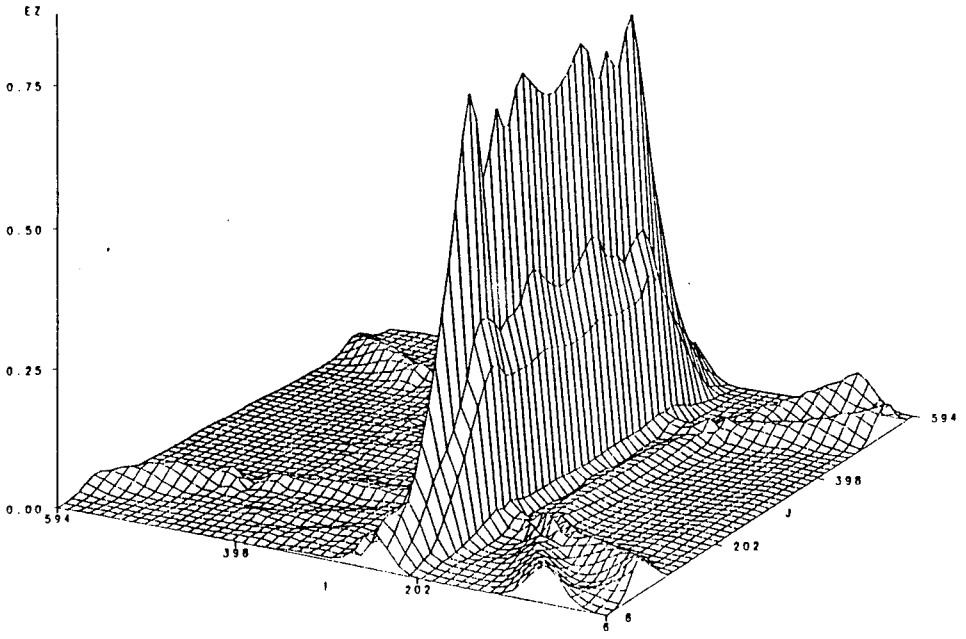


Figure 13: Smoothed version of the wave at $t = 750\Delta t$

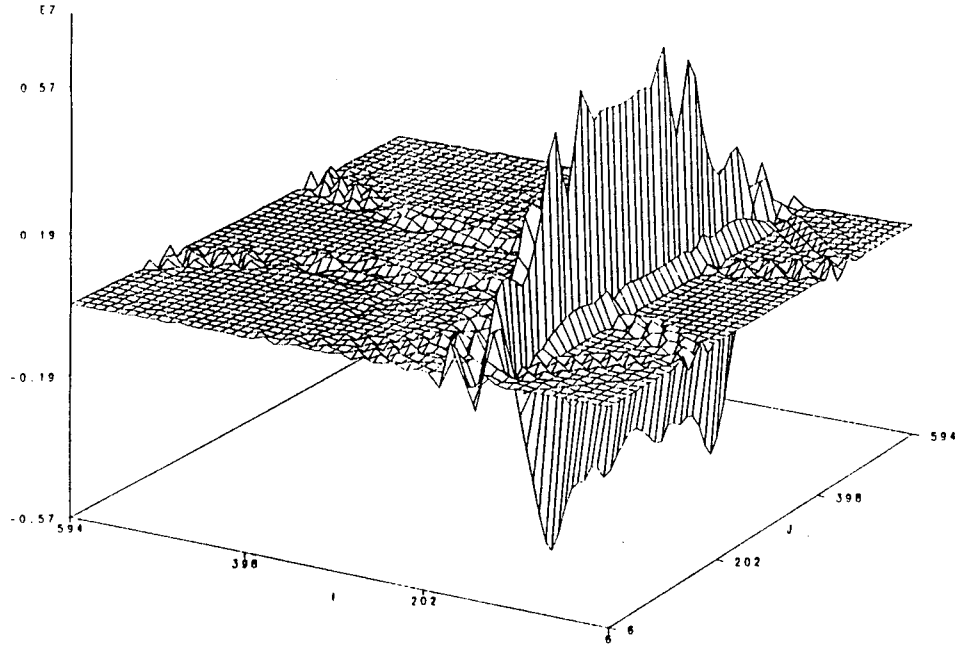


Figure 14: Sampled modulated wave at $t = 900\Delta t$.

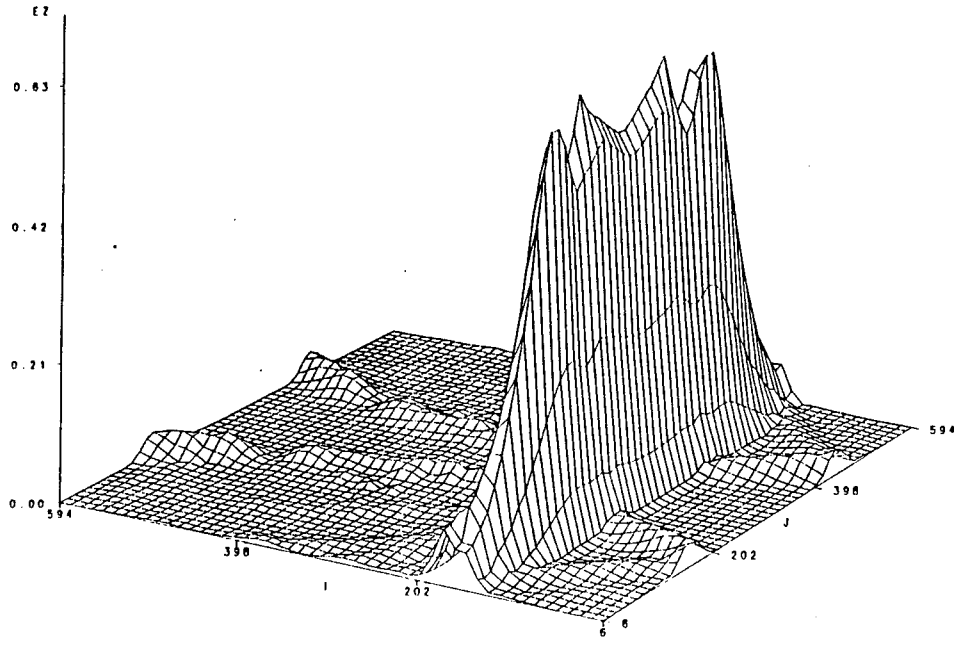


Figure 15: Smoothed version of the wave at $t = 900\Delta t$

C^1 APPROXIMATION TECHNIQUES FOR INVESTIGATING AND TRAINING NEURAL NETWORKS

Michael D. Taylor
Professor
Department of Mathematics

Unversity of Central Florida
P. O. Box 25000
Orlando, FL 32816

Final Report for:
Research Initiation Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Griffiss Air Force Base, Rome, New York

and

University of Central Florida

July 1992

C^1 APPROXIMATION TECHNIQUES FOR INVESTIGATING AND TRAINING NEURAL NETWORKS

Michael D. Taylor
Professor
Department of Mathematics
University of Central Florida

Abstract

The most popular method for training neural networks is the error back-propagation algorithm. In principle this method of training ought not to be applicable to many widely used networks because the functions which describe them are not continuously differentiable (C^1). In practice this difficulty can be overcome by constructing a second neural network which approximates the behavior of the first network but having the property that the functions which describe the behavior of the second network are C^1 functions. One can train the second network and transfer the results of the training to the first network. Systematic methods of constructing such approximating networks were studied. In the case where the original network is describable in terms of continuous functions, particular attention was paid to a method of constructing approximations by the use of convolutions. In the case where the original network is describable by discontinuous functions, it was found that there are fundamental mathematical obstacles to constructing sastisfactory approximations, and some progress was made in overcoming these obstacles.

C^1 APPROXIMATION TECHNIQUES FOR INVESTIGATING AND TRAINING NEURAL NETWORKS

Michael D. Taylor

INTRODUCTION

Neural nets have emerged as a possible alternative or complement to standard computer technology. They are believed to have a potential for performing such tasks as image processing, pattern recognition, pattern reconstruction, and optimal path determination in a swifter and more satisfactory fashion than standard computers. Two well-known introductions to the area are [6] and [12], and a good survey of recent advances can be found in [9] and [10].

One very important characteristic of neural nets and one reason for hoping they can perform some otherwise difficult tasks is that neural nets have the potential to be trained rather than having to be programmed. The most widely used method of training neural nets is the error back-propagation algorithm. The back-propagation algorithm has been described many times in the literature on neural nets. (See, for example, [13] or [15].)

The description of the back-propagation algorithm has usually been given with certain restrictions in mind, namely that computations of cell states will involve multiplying connection weights by other cell states, adding a number of results of this kind together, then adding (or subtracting) a "threshold" value, and finally applying a so-called "squashing function" or "transfer function" to the result. That is, the value of the ith cell in the network is given by a function of the form

$$s_i = f\left(\sum_j w_{ij}s_j + \theta_i\right)$$

where the sum is taken over all j for which the jth cell has an input to the ith cell. This means that a very particular architecture is being assumed. However it is straightforward to extend back-propagation to layered, feedforward networks of a more general sort than those which are usually considered.

There is also a certain technical restriction on the method of back-propagation: The algorithm requires that the ways in which units of a neural net depend on one another all be describable in terms of what are called continuously differentiable functions (also known as C^1 functions). Not all nets meet this requirement, so that any attempt to apply back-propagation to them lacks theoretical justification.

It should be noted in connection with these last two points that there is recognition in the neural network community of the desirability of studying neural networks with nonstandard architectures. This point is specifically brought out in the DARPA study on neural networks, [3]. Another example occurs in [11] where nonstandard architectures of interest in image processing are mentioned as worthy of further investigation; this last example makes use of operations typical of fuzzy analysis.

A particular motivation for this work is interest in neural nets having an architecture suggested by concepts occuring in fuzzy analysis. Fuzzy analysis is a method of dealing with uncertainty, particularly uncertainty arising from vagueness rather than randomness. (Example: When is a pear ripe? There is more than one possible answer.) There has recently been a great deal of interest in using fuzzy analysis in control and classification problems; for example, as described in [7], [1], and [5], an automatic train operating system has been successfully tested in Japan, fuzzy control logic has been applied to a glass melting furnace, and fuzzy set theory has been used for classifying pyrometallurgical industry products. Fuzzy analysis has also been turned to the problem of pattern recognition (see [2]) with possible applications to military intelligence and medical data.

Very little seems known about neural net architectures which derive from fuzzy analysis concepts. (But an example of such an architecture can be seen in [8].) One difficulty in investigating such architectures is that the functions one typically uses in fuzzy analysis (maximum and minimum) are nondifferentiable. This means the application in this area of the back-propagation algorithm, our most widely used tool in training a neural net, lacks firm theoretical justification.

THE PROBLEM

Working in the 1991 AFOSR Summer Research Program, [14], this author began

to develop a way – at least in theory – to extend the back-propagation algorithm to many networks for which its application would not normally be justified. The idea is essentially this: Suppose we are given a neural net whose workings are described by continuous functions. We construct a second neural net whose behavior approximates that of the first net very closely – as closely as we may wish – but whose workings are described by C^1 (that is, continuously differentiable) functions. Let us call this second network a C^1 network. (The fact that the approximating C^1 network can be constructed is a straightforward application of some well-known facts from mathematical analysis. A proof of the construction and a particular way to carry it out are given in [14].) The back-propagation algorithm can now be applied to train the C^1 network. The results of this training – connection weights and threshold values of cells – can then be transferred to the original network. Since the two networks behave in very similar ways, the original network then behaves as though trained.

Two remarks are now appropriate.

First, in the work just described these C^1 approximation techniques have been studied only in the case when the neural network is describable by <u>continuous</u> functions. The case of discontinuous functions was not examined. Yet the idea of approximating a neural network by a C^1 network and training the C^1 network by back-propagation is implicit in the literature from early on, and it is precisely this case in which the functions are discontinuous which is dealt with. (See for example [13] in which the nodes of the networks take on only the values $+1$ and -1.) Moreover this discontinous function case is one in which, as we shall show later, serious difficulties might emerge.

Second, we may now formulate the main goals of this work. On the one hand we wish to find a way, given a neural network, to construct an approximating C^1 network and to be able to state with confidence what sort of approximation is meant and how good it is. On the other hand we wish to develop a means of investigating nonstandard neural net architectures, particularly those suggested by concepts from fuzzy analysis. Achievement of the first goal can be used as a means of attaining the second goal. Much progress has been made toward the first goal in the continuous case. Only partial progress has been made in implementing this work in computer simulations, in investigating the discontinuous case, and in the investigation of nonstandard architectures.

THE CONTINUOUS CASE

Suppose we are dealing with a feed-forward, multilayer, synchronous neural network. Suppose the input layer has N_0 cells, the next layer has N_1, the next N_2, and so on to the output layer which has N_r cells. We assume the state of each cell is given by a real number (or by an element of some subset of the real numbers such as $\{-1,1\}$) so that the state of the ith layer at time t may be considered to be a point in $\mathbb{R}^{N_i}$. There must be a function $F_i: \mathbb{R}^{N_i-1} \to \mathbb{R}^{N_i}$ with the property that if x is the state of the $(i-1)$th layer at time t, then $F_i(x)$ is the state of the ith layer at time $t+1$. It follows that the neural net architecture is, in effect, described by a sequence of functions thus:

$$\mathbb{R}^{N_0} \overset{F_1}{\to} \mathbb{R}^{N_1} \overset{F_2}{\to} \dots \overset{F_r}{\to} \mathbb{R}^{N_r}.$$

Notice that what one might call the total response or ultimate response of the network, namely the output arising from a given input, is given by the composite function $F_r \circ F_{r-1} \circ \cdots \circ F_1$.

What we call the continuous case is the one in which all the F_i's are continuous functions. This takes in functions such as max and min which occur in fuzzy analysis (hence by implication some interesting nonstandard architectures) but not such functions as

$$\operatorname{sgn}(x) = \begin{cases} -1 & \text{if } x \le 0 \\ 1 & \text{if } x > 0 \end{cases}$$

which occurs quite frequently in the literature.

The back-propagation algorithm can be applied provided each F_i is a C^1 function. In practice this requirement is often not satisfied. The usual response in the literature to such a situation seems to be the replacement of each F_i by a C^1 function G_i which behaves, in some sense, approximately the same as F_i. For example, $\operatorname{sgn}(x)$ might be approximated by

$$g_\alpha(x) = \frac{e^{\alpha x}-1}{e^{\alpha x}+1} \text{ where } \alpha > 0.$$

The replacement is apparently done on an ad hoc basis; experimentation shows which replacements work and and which do not. In order for this procedure to be a reasonable

one, it is essential that $F_r \circ F_{r-1} \circ \cdots \circ F_1$ be closely approximated by $G_r \circ G_{r-1} \circ \cdots \circ G_1$. As we shall show in the next section, situations can easily arise in which this does not happen.

In [14] a particular class of approximations is explored in a systematic way for what is essentially the continuous case. We describe this procedure in a somewhat simplified form: For each F_i construct a C^1 function $\phi_i: \mathbb{R}^{N_{i-1}} \to \mathbb{R}$ satisfying

$$\text{supp } \phi_i \subseteq [-\delta_{i-1}, \delta_{i-1}]^{N-1},$$
$$\phi_i \geq 0,$$

and

$$\int \phi_i = 1,$$

where supp ϕ_i is the support of ϕ_i and δ_{i-1} is a given positive number. Then set $G_i = F_i * \phi_i$ where $*$ is the operation of convolution. (Since F_i is vector-valued and ϕ_i is real-valued, $F_i * \phi_i$ is the vector-valued function obtained by convoluting every component of F_i with ϕ_i .) In [14] it was shown how, by properly choosing the δ_i's, one may construct G_i's so that $G_r \circ G_{r-1} \circ \cdots \circ G_1$ differs in norm from $F_r \circ F_{r-1} \circ \cdots \circ F_1$ by less than any previously assigned ϵ. (The norm used in [14] was the sup norm. That is, if $x = (x_1, x_2, \cdots, x_n)$, a point in $\mathbb{R}^n$, then the sup norm of x is $\|x\| = \sup\{|x_i| : 1 \leq i \leq n\}$. In effect this measures the difference in performance between two networks by comparing the values on corresponding cells, a pair at a time, and taking the norm to be the largest magnitude of the difference of values of corresponding cells. This might also be thought of as an $\mathcal{L}^\infty$ norm.)

In a somewhat expanded version of this work which has been submitted to the Air Force as a technical report, [4], it is brought out that there is a difficulty in running computer simulations of these C^1 approximation techniques. Convolutions tend to be procedures which require a great deal of computation. Since training a neural network by back-propagation can require a huge number of training runs, to combine this procedure with convolutions is to create a process which takes far too long to run on a computer. However there are several things which can be done about this.

First of all one can restrict one's attention to the special but still interesting case in which the state of the ith cell is given by

$$s_i = f\left(\sum_j w_{ij} s_j + \theta_i\right)$$

where the sum is taken over all j for which the jth cell has an input to the ith cell.

Assuming that f is a continuous, nondifferentiable function, one can then form a convolution not with $f\left(\sum w_{ij}s_j + \theta_i\right)$, a function of many variables, but with f, a function of a single variable. Not only does this greatly speed up the process of running a computer simulation, it turns out that the theorem in [14] which estimates the "goodness" of approximation of the approximating C^1 network still holds with the same epsilons and deltas. This strategy has been successfully implemented in [4] for the XOR problem using for f the function

$$f_T(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } 1 \leq x. \end{cases}$$

Second it should be noted that this difficulty we are considering does not say anything is wrong with the theory of using convolutions to construct C^1 approximating networks, only that it is hard to simulate the application of such a theory on a serial computer. One might try to find particulary fast ways to compute convolutions. Alternatively, when dealing with a specific neural net architecture and set of functions defining cell states, one might hope to construct unusually efficient convolution techniques for that particular case. Finally one might consider constructing special hardware to implement the convolution/approximation concept; there exist in engineering literature well-developed digital and analog techniques for computing convolutions with great rapidity.

Third, the use of convolutions is only one possible technique for constructing approximating C^1 networks. Its great virtue is its generality and the ease of theoretical analysis. There are other means of constructing C^1 approximations to given functions. Here are some examples of approximations of functions of interest in the study of neural networks:

Recall that

$$f_T(x) = \begin{cases} 0 \text{ if } x \leq 0, \\ x \text{ if } 0 \leq x \leq 1, \\ 1 \text{ if } x \geq 1. \end{cases}$$

$$H(x) = \begin{cases} 0 \text{ if } x \leq 0, \\ 1 \text{ if } x > 0. \end{cases}$$

$$\operatorname{sgn}(x) \;=\; \begin{cases} -1 \text{ if } x \le 0, \\ 1 \text{ if } x > 0. \end{cases}$$

f_T is the threshold function, H is the Heaviside function, and sgn is the signum function.

It is widely recognized that

$$H(x) \;\approx\; \frac{1}{1+e^{-\alpha x}} \quad \text{where } \alpha > 0.$$

In this case the approximation is not to be understood as sup norm or uniform, since H has a discontinuity at 0, but in some other sense, perhaps, for example, as an $\mathcal{L}^1$ approximation or in terms of convergence in measure or pointwise. This particular approximation is widely used in the study of neural nets.

Since $\operatorname{sgn}(x) = 2H(x) - 1$, we have

$$\operatorname{sgn}(x) \;\approx\; \frac{2}{1+e^{-\alpha x}} - 1 \;=\; \frac{e^{\alpha x}-1}{e^{\alpha x}+1} \quad \text{where } \alpha > 0.$$

Again the approximation should be probably be understood in an $\mathcal{L}^1$ sense or in terms of convergence in measure or as pointwise convergence.

Both min and max (or $\wedge$ and $\vee$) are of interest in fuzzy analysis and in some areas of application such as image processing. Note that

$$|\,x\,| \;=\; x\,\operatorname{sgn}(x) \;\approx\; x\left(\frac{e^{\alpha x}-1}{e^{\alpha x}+1}\right).$$

These should be understood as pointwise or $\mathcal{L}^\infty$ approximations (with respect to some given, bounded interval). Since

$$x \wedge y \;=\; (1/2)\,(x+y-|\,x-y\,|)$$
$$x \vee y \;=\; (1/2)\,(x+y+|\,x-y\,|)\,,$$

this leads to the following approximations for $\wedge$ and $\vee$:

$$x \wedge y \;\approx\; \frac{1}{2}\left\{ x+y-(x-y)\left(\frac{e^{\alpha(x-y)}-1}{e^{\alpha(x-y)}+1}\right)\right\}$$

$$x \vee y \approx \frac{1}{2}\left\{ x+y+(x-y)\left(\frac{e^{\alpha(x-y)}-1}{e^{\alpha(x-y)}+1}\right)\right\}.$$

These approximations are to be understood in a pointwise sense or $\mathcal{L}^\infty$ sense with respect to some bounded interval.

Finally we note that

$$f_T(x) \approx \frac{1}{\alpha}\ln\left(\frac{1+e^{\alpha x}}{1+e^{\alpha(x-1)}}\right) \text{ where } \alpha > 0.$$

Again this is a pointwise or $\mathcal{L}^\infty$ approximation. As a matter of fact, if we set

$$g_\alpha(x) = \frac{1}{\alpha}\ln\left(\frac{1+e^{\alpha x}}{1+e^{\alpha(x-1)}}\right),$$

then it can be shown that

$$\max_{x\in\mathbb{R}} | f_T(x) - g_\alpha(x) | = \frac{1}{\alpha}\ln\left(\frac{2}{1+e^{-\alpha}}\right) < \frac{1}{\alpha}\ln 2 .$$

THE DISCONTINUOUS CASE

If we are given a neural network describable by a sequence of functions thus,

$$\mathbb{R}^{N_0} \overset{F_1}{\to} \mathbb{R}^{N_1} \overset{F_2}{\to} \cdots \overset{F_r}{\to} \mathbb{R}^{N_r},$$

where the F_i's may now be discontinuous, then we wish to construct an "approximating" network

$$\mathbb{R}^{N_0} \overset{G_1}{\to} \mathbb{R}^{N_1} \overset{G_2}{\to} \cdots \overset{G_r}{\to} \mathbb{R}^{N_r}$$

where the G_i's are C^1 functions. Presumably each F_i will be approximated by G_i in some appropriate sense. Also, it is absolutely essential that $F_r \circ F_{r-1} \circ \cdots \circ F_1$ be approximated by $G_r \circ G_{r-1} \circ \cdots \circ G_1$. What is a reasonable sort of approximation to use? Since we are dealing with discontinuous functions, we cannot talk about uniform or $\mathcal{L}^\infty$ approximations. We present some examples which shed a little light on the situation:

Example 1. Let

$$H(x) = \begin{cases} 0 & \text{if } x \le 0, \\ 1 & \text{if } x > 0, \end{cases}$$

and

$$g_\alpha(x) = \frac{1}{1+e^{-\alpha x}} \text{ where } \alpha > 0.$$

It is readily seen that as $\alpha \to \infty$, we must have $g_\alpha \to H$ in $\mathcal{L}^1(\lambda)$ and in measure (where the measure in question, λ, is Lebesgue measure on $\mathbb{R}$).

Now let us consider what happens when we compose these maps. Note that $H \circ H = H$. But notice that if $x \ll 0$, then $g_\alpha(x) \approx 0$ and hence $(g_\alpha \circ g_\alpha)(x) \approx 1/2$. Thus $g_\alpha \circ g_\alpha$ does not converge to $H \circ H$ in any of the senses considered above.

This example reveals something very important. Namely, even if one chooses very reasonable -looking approximations $F_1 \approx G_1$ and $F_2 \approx G_2$, one cannot be sure that $F_2 \circ F_1 \approx G_2 \circ G_1$. This seems to be the central problem in constructing C^1 neural net approximations in the discontinuous case.

Example 2. Now let

$$\text{sgn}(x) = \begin{cases} -1 & \text{if } x \leq 0, \\ 1 & \text{if } x > 0, \end{cases}$$

and

$$f_\alpha(x) = \frac{e^{\alpha x}-1}{e^{\alpha x}+1} \text{ where } \alpha > 0.$$

As in the last example, as $\alpha \to \infty$, we see that $f_\alpha \to$ sgn in $\mathcal{L}^1(\lambda)$, pointwise, and in measure.

What happens with the compositions? We have $\text{sgn} \circ \text{sgn} = \text{sgn}$. As regards $f_\alpha \circ f_\alpha$, note first that it must be an increasing function with range between 0 and 1, so there is some hope it may converge to sgn. However $-1 < f_\alpha(-1) < (f_\alpha \circ f_\alpha)(x) < f_\alpha(1) < 1$, which implies the area lying between the graphs of $f_\alpha \circ f_\alpha$ and sgn must be infinite and thus there can be no $\mathcal{L}^1$ convergence as $\alpha \to \infty$. Now suppose x is fixed and positive. Notice that

$$(f_\alpha \circ f_\alpha)(x) = \frac{e^{\alpha f_\alpha(x)}-1}{e^{\alpha f_\alpha(x)}+1}.$$

If $\alpha \to \infty$, then we must have

$$f_\alpha(x) \to 1,$$
$$\alpha f_\alpha(x) \to \infty,$$

and hence

$$(f_\alpha \circ f_\alpha)(x) \to 1.$$

Similarly for x fixed and negative, we must have $(f_\alpha \circ f_\alpha)(x) \to -1$ as $\alpha \to \infty$. It follows from this and the geometry of the graphs that $f_\alpha \circ f_\alpha \to \text{sgn} \circ \text{sgn}$ in measure.

This example seems to say that $\mathcal{L}^1$ convergence is not an appropriate choice. And if not $\mathcal{L}^1$, then presumably neither is $\mathcal{L}^p$ for some other choice of p. Both pointwise convergence and convergence in measure, both which do work here, have nice, intuitive meanings. In the case of convergence in measure, it means that the approximating function has almost the same value as another one most of the time.

Example 3. We now return to the H of Example 1 and produce a modification of g_α such that its compositions will converge pointwise and in measure to compositions of H. As a matter of fact we shall show convergence holds for k-fold compositions. (By the k-fold composition of F with itself, F^k, we mean $F \circ F \circ \cdots \circ F$ where F occurs k times.) Convergence in $\mathcal{L}^1$ fails for the same reason as given in Example 2.

We construct the function

$$e_\alpha(x) = \frac{1}{1 + e^{-\alpha^2(x - (1/\alpha))}} \quad \text{where } \alpha > 0.$$

Note that ${e_\alpha}^k$ must be an increasing function whose values lie strictly between 0 and 1. Since $H^k = H$, we need only show that ${e_\alpha}^k \to H$ in measure. Because of the geometry of the graphs of the functions, it is sufficient to show that ${e_\alpha}^k(0) \to 0$ and ${e_\alpha}^k(x) \to 1$ for x fixed and positive as $\alpha \to \infty$.

We show by induction that for every natural number k we have $\alpha^2 {e_\alpha}^k(0) \to 0$ as $\alpha \to \infty$. (This gives us ${e_\alpha}^k(0) \to 0$.) Since

$$\alpha^2 e_\alpha(0) = \frac{\alpha^2}{1 + e^\alpha} \to 0 \quad \text{as } \alpha \to \infty,$$

we have the proposition for $k = 1$. Suppose we know $\alpha^2 e_\alpha{}^k(0) \to 0$ as $\alpha \to \infty$. From the definition of e_α we obtain

$$\alpha^2 e_\alpha{}^{k+1}(0) = \frac{\alpha^2}{1 + e^{\alpha - \alpha^2 e_\alpha{}^k(0)}},$$

which by the induction hypothesis must converge to 0 as $\alpha \to \infty$.

Now let x be fixed and positive. We induct on k to show $e_\alpha{}^k(x) \to 1$. This follows easily for $k = 1$, so now suppose we know it for some given k. Note that $\alpha - \alpha^2 e_\alpha{}^k(x) \to -\infty$ as $\alpha \to \infty$. Since

$$e_\alpha{}^{k+1}(x) = \frac{1}{1 + e^{\alpha - \alpha^2 e_\alpha{}^k(x)}},$$

it follows that $e_\alpha{}^{k+1}(x) \to 1$ as $\alpha \to \infty$. This establishes convergence pointwise and in measure.

What is revealed in this example is that even though "reasonable" choices of approximating functions may turn out to give bad results, there may be other choices which give good results. An obvious question then is, how can one know beforehand that a choice is a good one? Better yet, how can one construct good choices?

We present here, in partial answer to this question, a result concerning compositions and convergence of approximations to discontinuous functions.

A reasonably general class of functions to consider is the set of $f:\mathbb{R}\to\mathbb{R}$ which are of bounded variation. Every such function can be decomposed into a sum $f = g + h$ where g is continuous and h is a step function with at most a countable number of discontinuities. We already have some idea of how to deal with continuous functions from [14]. Recalling the defintion of the Heaviside function H, note that h can be written in the form

$$h(x) = \sum_i a_i H(x - x_i) + b$$

where b and the a_i's and x_i's are constants. We will look at functions of this last form and possessing only a finite number of steps.

Recall that

$$e_\alpha(x) \;=\; \frac{1}{1+e^{-\alpha^2(x-(1/\alpha))}} \quad \text{where } \alpha > 0,$$

or equivalently

$$e_\alpha(x) \;=\; \frac{1}{1+e^{\alpha-\alpha^2 x}}\,.$$

Let

$$F_j(x) \;=\; \Big(\sum_{i=1}^{m_j} a_{ij} H(x-x_{ij})\Big) + b_j \quad \text{for } j = 1, 2, \cdots, n$$

and

$$F_{j,\alpha}(x) \;=\; \Big(\sum_{i=1}^{m_j} a_{ij}\, e_\alpha(x-x_{ij})\Big) + b_j \quad \text{for } j = 1, 2, \cdots, n$$

where $\alpha > 0$, each a_{ij} , x_{ij} , and b_j is a constant, and $n \geq 2$. Let

$$G_k \;=\; F_k \circ F_{k-1} \circ \cdots \circ F_1$$

and

$$G_{k,\alpha} \;=\; F_{k,\alpha} \circ F_{k-1,\alpha} \circ \cdots \circ F_{1,\alpha}\,.$$

Then the following result shows that the $F_{j,\alpha}$'s are the "right" kind of C^1 functions with which to approximate the F_j's:

THEOREM. $\lim\limits_{\alpha\to\infty} G_{n,\alpha}(x) \;=\; G_n(x)$ for all x.

NOTE: It can be seen from the proof that the convergence is faster than that of α^2/e^α as $\alpha \to \infty$.

Proof. First of all we will prove by induction on k that

$$\lim_{\alpha\to\infty} \alpha^2\big(G_{k,\alpha}(x) - G_k(x)\big) \;=\; 0$$

and

$$\lim_{\alpha\to\infty} \alpha^2\left(e_\alpha(G_{k,\alpha}(x)+c) \;-\; H(G_k(x)+c)\right) \;=\; 0$$

where c is an arbitrary constant.

If $x \leq 0$, then

$$\alpha^2\left(e_\alpha(x) - H(x)\right) \;=\; \alpha^2\, e_\alpha(x)$$

$$= \frac{\alpha^2}{1+e^{\alpha-\alpha^2 x}} \to 0 \text{ as } \alpha \to \infty.$$

If $x > 0$, then

$$\alpha^2\left(e_\alpha(x) - H(x)\right) = \alpha^2\left(e_\alpha(x) - 1\right)$$

$$= \alpha^2\left(\frac{-1}{e^{\alpha^2 x - \alpha}+1}\right) \to 0 \text{ as } \alpha \to \infty.$$

Hence

$$\lim_{\alpha\to\infty} \alpha^2\left(e_\alpha(x) - H(x)\right) = 0.$$

From this we see that

$$\lim_{\alpha\to\infty} \alpha^2\left(G_{1,\alpha}(x) - G_1(x)\right) = \lim_{\alpha\to\infty} \sum_{i=1}^{m_1} a_{i1}\, \alpha^2\left(e_\alpha(x - x_{i1}) - H(x - x_{i1})\right) = 0.$$

Next notice that we can write

$$\alpha - \alpha^2\left(G_{1,\alpha}(x) + c\right) = \alpha - \alpha^2\left(G_1(x) + c\right) + \alpha^2\left(G_1(x) - G_{1,\alpha}(x)\right),$$

so that if we let $\alpha \to \infty$, we obtain

$$\alpha - \alpha^2\left(G_{1,\alpha}(x) + c\right) \to \infty \text{ if } G_1(x) + c \le 0$$

and

$$\alpha - \alpha^2\left(G_{1,\alpha}(x) + c\right) \to -\infty \text{ if } G_1(x) + c > 0\,.$$

Furthermore in each of these cases the divergence is at least as rapid as that of α. Now suppose $G_1(x) + c \le 0$ and consider what happens as $\alpha \to \infty$. We must have

$$\alpha^2\left(e_\alpha(G_{1,\alpha}(x) + c) - H(G_1(x) + c)\right) = \frac{\alpha^2}{1 + e^{\alpha - \alpha^2(G_{1,\alpha}(x) + c)}} \to 0$$

since the denominator diverges exponentially. If, on the other hand, $G_1(x) + c > 0$ and we let $\alpha \to \infty$, then

$$\alpha^2\left(e_\alpha(G_{1,\alpha}(x) + c) - H(G_1(x) + c)\right) = \frac{-\alpha^2}{e^{\alpha - \alpha^2(G_{1,\alpha}(x) + c)} + 1} \to 0$$

again because the denominator diverges exponentially. Therefore we have shown that

$$\lim_{\alpha\to\infty} \alpha^2 \left(e_\alpha(G_{1,\alpha}(x)+c) - H(G_1(x)+c)\right) = 0$$

which completes the first step in the induction.

Now suppose we know that

$$\lim_{\alpha\to\infty} \alpha^2\left(G_{k,\alpha}(x) - G_k(x)\right) = 0$$

and

$$\lim_{\alpha\to\infty} \alpha^2 \left(e_\alpha(G_{k,\alpha}(x)+c) - H(G_k(x)+c)\right) = 0.$$

It follows that

$$\lim_{\alpha\to\infty} \alpha^2 \left(G_{k+1,\alpha}(x) - G_{k+1}(x)\right) =$$

$$\lim_{\alpha\to\infty} \alpha^2 \left(F_{k+1,\alpha}(G_{k,\alpha}(x)) - F_{k+1}(G_k(x))\right) =$$

$$\lim_{\alpha\to\infty} \sum_{i=1}^{m_{k+1}} a_{i,k+1}\, \alpha^2 \left(e_\alpha(G_{k,\alpha}(x) - x_{i,k+1}) - H(G_k(x) - x_{i,k+1})\right) = 0.$$

It follows easily from this, in the manner exhibited in the $k = 1$ case, that

$$\lim_{\alpha\to\infty} \alpha^2 \left(e_\alpha(G_{k+1,\alpha}(x)+c) - H(G_{k+1}(x)+c)\right) = 0.$$

This completes the induction.

It follows trivially from

$$\lim_{\alpha\to\infty} \alpha^2 \left(G_{n,\alpha}(x) - G_n(x)\right) = 0$$

that

$$\lim_{\alpha\to\infty} G_{n,\alpha}(x) = G_n(x)\,. \qquad \square$$

SUMMARY AND FURTHER LINES OF INVESTIGATION AND ACTIVITY

In summary then, the problem of constructing approximating C^1 networks in a systematic fashion and the properties of such approximations were studied. In the case where the original network is describable in terms of continuous functions, a

theoretically satistfactory approximation can be constructed using convolutions. However it is difficult to perform computer simulations of this particular approximation method. Alternate methods of constructing approximations in the continuous case were considered. In the case where the network is describable in terms of discontinuous functions, fundamental difficulties of a mathematical nature arise. A partial solution is given to an important case of this last problem.

On the basis of what has been done so far, some interesting lines of inquiry suggest themselves.

1. Software. There is need to develop software for carrying out computer simulations of the concepts put forward here. A beginning in this direction has been made by Captain Jeff Fordon of the U. S. Air Force in [4]. A graduate student here at the University of Central Florida, Terence Durning, is presently at work to extend Captain Fordon's work.

2. "Fuzzy" networks. Depending on the success of 1, neural networks which are constructed using functions from fuzzy analysis should be trained on simple problems such as the XOR problem, addition base 2, encoding problems, etc. This would serve as a test of the C^1 network approximation techniques and simultaneously yield some knowledge of the capabilities of nonstandard neural networks.

3. The problem of spurious minima. One of the difficulties in applying back-propagation is the occurrence of local minima which produce false solutions of the training problem. The construction of C^1 neural nets to approximate given neural nets is a smoothing process. It is possible that such constructions might be used as a technique to get rid of local minima. (Following a remark in [13], base 2 addition is a candidate for experimentation here.)

Let us elaborate on this idea. Suppose we wish to investigate a neural net $\mathcal{N}$ and we construct a family of neural nets $\mathcal{N}_t$ having the property that $0 \leq t < \infty$, that each $\mathcal{N}_t$ is a C^1 neural net for $t > 0$, and $\mathcal{N}_t \to \mathcal{N}_0 = \mathcal{N}$ as $t \to 0$. (This would be straightforward using the convolution approximation described in [14].) We may arrange things so that the functions describing each $\mathcal{N}_t$ are "smoother" the larger t is. Then it is natural to ask, will it be easier to find a global minimum of the error surface or performance index for $\mathcal{N}_t$ which is "very smooth" than for an $\mathcal{N}_t$ which is only "moderately or a little smooth"? How will minima move across the error surfaces as $t \to 0$? Will global minima

transform to global minima as t increases?

Terence Durning of the University of Central Florida is currently carrying out work to test the feasibility of this idea in a computer simulation.

4. A fundamental mathematical problem is raised by this work. If we are given functions which approximate one another, say $F_i \approx G_i$ for $i = 1, 2, \cdots, r$, then when can we be sure that we will have $F_r \circ F_{r-1} \circ \cdots \circ F_1 \approx G_r \circ G_{r-1} \circ \cdots \circ G_1$? Is there a way to construct F_i's and G_i's to assure they will always have this property?

ACKNOWLEDGEMENTS

The author wishes to acknowledge the kind help of Capt. Jeff Fordon of Rome Laboratory, Griffiss Air Force Base, and Terence Durning, University of Central Florida, in carrying out programming which contributed to this work. He is also indebted to Pitor Mikusiński and Roy Choudhury, both of the Mathematics Department of the University of Central Florida, for many useful conversations and for their help in this research.

REFERENCES

1. S. Aoki and S. Kawachi, "Application of fuzzy control logic for dead-time processes in a glass melting furnace," Fuzzy Sets and Systems 38 (1990), 251-265.

2. J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms. New York and London: Plenum Press, 1981.

3. DARPA Neural Network Study. Fairfax, VA: AFCEA International Press, 1988.

4. Capt. J. Fordon and M. D. Taylor, "Approximating neural nets with C^1 neural nets," Rome Lab Technical Report RL-TR-92-84. To be published.

5. A. O. Ivanov and V. D. Ryvkin, "The application of fuzzy set theory for the classification of the product quality in rotary kilns," Fuzzy Sets and Systems 41 (1991), 133-143.

6. R. P. Lippmann, "An introduction to computing with neural nets," IEEE Transactions Acoust., Speech, Signal Proc. ASSP-4 (1987), 4-22.

7. S. Miyamoto, S. Yasunobu, and H. Ihara, "Predictive fuzzy control and its application to automatic train operation systems," in J. C. Bezdek, ed., Analysis of Fuzzy Information, Vol. II, Artificial Intelligence and Decision Systems. Boca Raton,

FL: CRC Press, Inc., 1987.

8. W. Pedrycz, "Neurocomputations in relational systems," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 13 (1991), 289-297.

9. Proceedings of the IEEE, Special Issue on Neural Networks, I, Sept. 1990.

10. Proceedings of the IEEE, Special Issue on Neural Networks, II, Oct. 1990.

11. G. X. Ritter, D. Li, and J. N. Wilson, "Image algebra and its relationship to neural networks," SPIE, Vol. 1098, Aerospace Pattern Recognition, (1989), 90-101.

12. D. Rumelhart and F. McClelland, eds., Parallel Distributed Processing, Vol 1. Cambridge, MA: M.I.T. Press, 1986.

13. D. Rumelhart, D. Hinton, and G. Williams, "Learning internal representations by error propagation," in D. Rumelhart and F. McClelland, eds., Parallel Distributed Processing, Vol 1. Cambridge, MA: M.I.T. Press, 1986.

14. M. D. Taylor, "Approximating neural nets with C^1 neural nets," 1991 AFOSR Summer Research Program.

15. P. J. Werbos, "Backpropagation through time: What it does and how to do it," Proceedings of the IEEE, Vol 78 (1990), 1550-1560.

A FIBER OSCILLATOR FOR OPTICAL COMMUNICATIONS

Kenneth J. Teegarden
Professor of Optics
Institute of Optics

University of Rochester
Rochester, NY 14627

Salahuddin Qazi
Institute of Technology
State University of New york
Utica, NY 13504

Final Report for:
Research Initiation Program
Rome Laboratory

Sponsored by:
Air Force Office of Scientific Research
Bolling Air Force Base, Washington, D.C.

and

The University of Rochester
State University of New York

December 1992

A FIBER OSCILLATOR FOR OPTICAL COMMUNICATIONS

Kenneth J. Teegarden
Institute of Optics, University of Rochester, Rochester New York, 14627

Salahuddin Qazi
State University of New York, Institute of Technology, Utica New York, 13504

Abstract

A tunable erbium doped fiber ring laser, pumped with a 980 nm InGaAs diode laser was constructed. Output power as a function of pump power and output wavelength for a given fiber length was measured. Spectral and temporal analysis of the signal showed mode-locked pulses of short duration and broad frequency content, as well as a CW component confined to a relatively narrow optical frequency range. The mechanisms for this type of mode-locking are discussed as well as the limitations placed on the lasing line width and spectral tunabiltiy.

A FIBER OSCILLATOR FOR OPTICAL COMMUNICATIONS

Kenneth J. Teegarden
Salahuddin Qazi

INTRODUCTION

The recent rapid development and optimization of optical amplifiers based on flourescent fibers has proceeded to the point that compact, well engineered, products are available from a number of companies. Attention has now turned to the development of oscillators which also employ erbium doped fibers as the gain medium. Flourescent fiber oscillators show great promise as sources of short pulses or solitons in the 1.55 micron telecommunications window. They can be passively mode-locked, and pulses of femtosecond duration have been reported with high peak powers. Most of the work so far has required high power pump lasers and specially designed amplifiers. The objective of the work reported in this paper was to determine if commercially available amplifiers, employing relatively low power diode lasers as pumps, could be used to develop oscillators with some of the short pulse characteristics mentioned above. In addition, the tunability of such oscillators and their performance in wave length multiplexed communication systems was explored.

EXPERIMENTAL

The experimental arrangement used in this work is shown in figure 1. The fiber oscillator consisted of a erbium doped fiber amplifier developed by Corning Inc., and a feedback loop assembled from single mode fiber and commercially available components. The model P3-35 amplifier contained approximately 20 meters of erbium doped fiber, as well as a thermoelectrically cooled semiconductor pump laser which developed a maximum of 100 mW and the WDM needed to combine both the pump and signal in the the doped fiber. The amplifier module had an output port through which the pump power launched into the gain medium could be monitored as a function of pump current. At the maximum rated pump current of 200 mA, the amplifier had a gain of 28 dB at 1550 nm. Its optical bandwidth was listed as 40 nm. The output of the amplifier was fed back into its input port through a splitter, which determined the

feedback and output coupling ratio, and an optical isolator. The characteristics of the resulting oscillator output were determined with either a optical spectrum analyzer, a fast detector and elecrical signal analyzer, or an optical power meter.

RESULTS

At a given value of amplifier pump power, the output of the oscillator consisted of two distinct components. These were an unmodulated (although noisy) signal with a relatively narrow spectrum containing several temporal modes, and a periodic signal consisting of a series of pulses with half widths less than one nanosecond associated with an optical spectrum of about 7.0 nm half width. At pump powers well above the threshold for oscillation, the output consisted of both types of signal. Below a certain critical value of the pump power the mode of oscillation characterized by the periodic short pulses existed alone. These results are illustrated in figures 2 through 8.

The average output power of the oscillator was measured as a function of pump power for two different values of the output coupler, using a power meter with a time constant too slow to follow rapid temporal fluctuations in the output. The result of these measurements is shown in figure 2. As noted above, the optical spectrum of the output changed with the pump power. At low pump power, near threshold, the output spectrum was characterized by the relatively broad spectrum shown in figure 4, while at maximum pump power, oscillation occurred mainly over a much narrower spectral range as shown in figure 6.

The temporal behavior of the output of the oscillator was examined as a function of pump power using a detector- amplifier combination with a band width of 500Mhz and an electical signal analyzer with a band width of 1.0 GHz. The results are shown in figures 3. and 5. At low pump powers, corresponding to the optical spectrum shown in figure 4., the output was found to contain the periodic train of short pulses shown in figure 3. As the pump power was increased, the width of the pulses increased and a second series of pulses developed between the components of the first series. The situation for the maximum pump current of

150 mA. is shown in figure 5. The corresponding optical spectrum is given in figure 6. A careful examination of this optical spectrum revealed that it consisted of the narrow lines shown in figure 6 superimposed on the broad structure shown in figure 3. This result can be stated in the following way. As the amplifier pump power is increased, oscillation first occurs in the form of a periodic series of short pulses with a broad spectral output having a band width of about 7.0 nm. The broad spectral output and pulsed behavior persists at higher pump powers but in addition a narrow component to the optical spectrum, consisting of several lines with widths less than 1.0 nm, appears.

The period of the pulses in the output of the oscillator was found to be equal to the round trip time of the amplifier and its external feed back loop. Figure 7 shows the result of inserting first one, and then two, 5 meter lengths of single mode fiber into the feedback loop. The first point on the curve shown in figure 7 was obtained by calculating the length of fiber in the amplifier and external feed back loop from a measured value of the period of the pulses, using a value of 1.50 for the effective index of refraction of the fiber. The last two points were obtained by using measured values of the period and fiber lengths obtained by adding 5.0 and 10.0 meters to the original length. The fact that these three points lie on a straight line indicates that indeed the period of the pulses is equal to the time required for a short pulse of light to make one round trip through the amplifier and its feed back loop.

The width of the pulses obtained at low pump powers was found to be limited by the band pass of the detection system used. With a detector band pass of 500 Mhz the width was measured to be 2.0 ns. When the detector band pass was increased to a nominal 1.0 GHz the result shown in figure 8 was obtained. Here a pulse width of about 600 ps is indicated. At maximum pump power the pulse width was measured to be 10.0 ns.

It should be noted that the results reported above for the case of a commercially available amplifier could be duplicated using several different amplifiers constructed in our laboratory. It thus appears that they may be characteristic of

any oscillator based on erbium doped fibers.

CONCLUSIONS

The evidence presented above indicates that self-starting harmonic mode locking occured in the output of this erbium doped fiber ring oscillator. Although others[1,2] have described a variety of similar effects obtained with the use of non-linear amplifying loop mirrors, no unusual effort had to be made to induce the mode locking in our case. In fact, the generation of short periodic pulses occurred spontaneously over a wide range of pump power and output power. Because of instrumental limitations no conclusions can as yet be drawn as to the exact width of of the mode-locked pulses obtained at low pump powers, although they appear to be sub-nanosecond in duration. The 7.0 nm optical band width of the pulses in the low power regimen indicates a minimum pulse width of about 2 ps. Since the power of the pulses seems to be clamped at some peak value, their width must increase at higher pump powers, as observed. A similar effect has been observed by Richardson, et. al.[3]

The relatively large spectral line width associated with the spontaneous mode locking has implications relative to tunabilty and the use of such oscillators in wavelength multiplexed communications systems. In the absence of mode-locking, the data given above indicate a possible spectral line width of less than one nanometer. Mode-locking widens the output spectrum at its base to 7.0 nm. Obviously this increased line width will lead to cross talk and a sizable reduction in the number of channels available in a multiplexed system.

Finally, the mechanism which spontaneously couples modes at low pump powers is not clearly understood. It should be remembered, however that erbium represents a three level laser system in which both gain and absorption loss can be present simultaneously. In these longitudinally pumped oscillators, part of the fiber may be operating as a saturable absorber while the rest acts as the gain medium. It would thus be interesting to determine if similar results are obtained in a four level gain medium, such as neodymium doped fiber.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the assistance of the GTE Research Laboratories in Waltham, Massachusetts, who supplied some of the erbium doped fiber used in this work. The support of the Rome Laboratories and the Office of Scientific Research of the United States Air Force is also greatfully acknowledged.

6. REFERENCES

1. I. N. Duling III, "All-fiber ring soliton laser mode-locked with a nonlinear mirror," Opt. Lett. 16, 539 (1991).
2. M. Nakazawk, E. Yoshida, and Y. Kimura, "Low threshold 290 fs erbium-doped fiber laser with a nonlinear amplifying loop mirror pumped by InGaAsP laser diodes," Appl. Phys. Lett. 59, 2073 (1991).
3. D. J. Richardson, R. I. Laming, D. N. Payne, M. W. Phillips, and V. J. Matsas, "Self starting , passively modelocked erbium fiber ring laser based on the amplifying Sagnac switch," Electron. Lett. 27, 542 (1991).

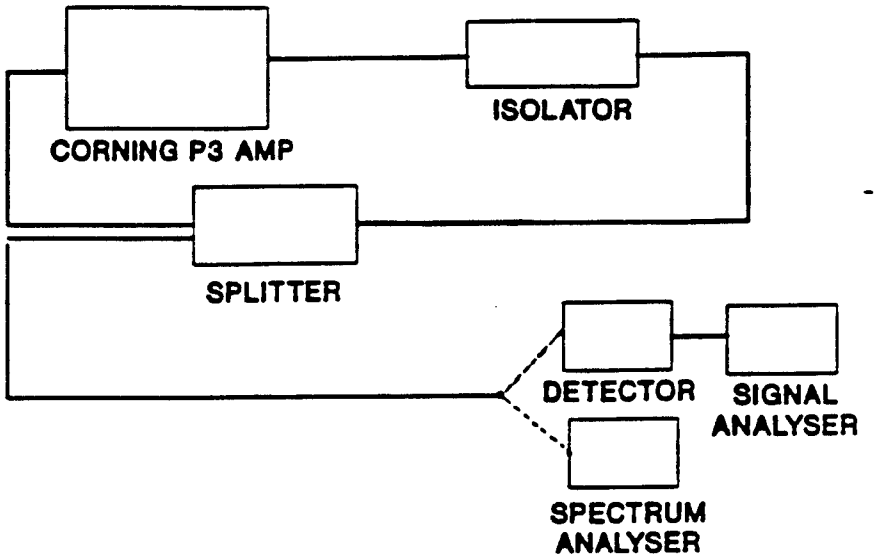


Figure 1. Experimental setup of fiber oscillator.

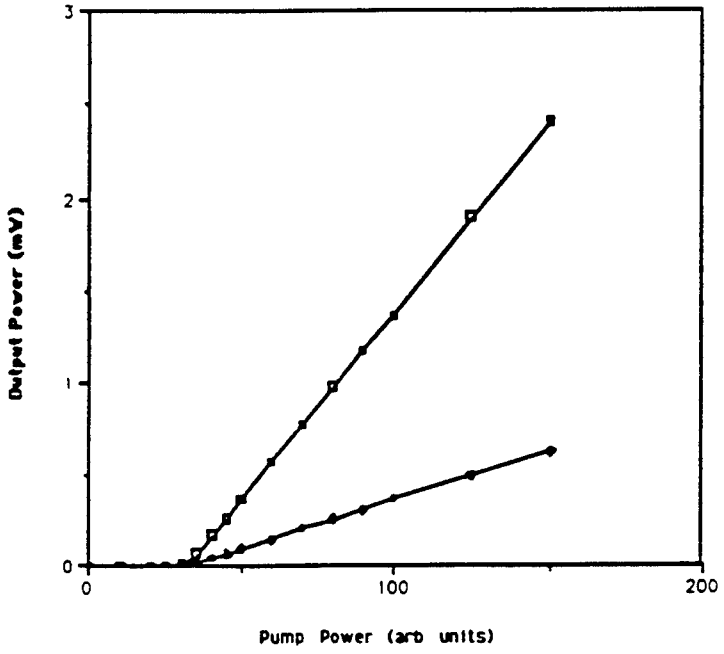


Figure 2. Oscillator output power as a function of pump power.

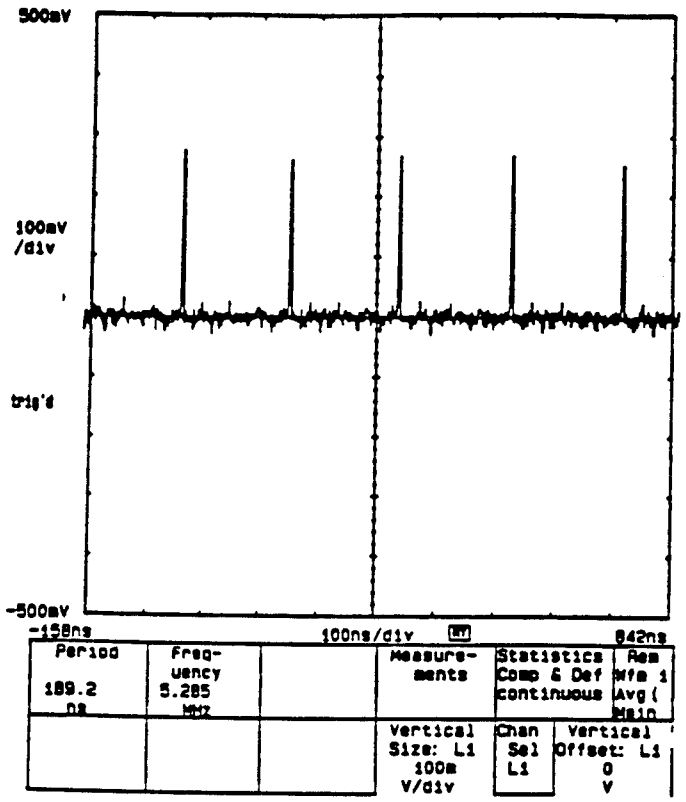


Figure 3. Output waveform of oscillator at low pump power.

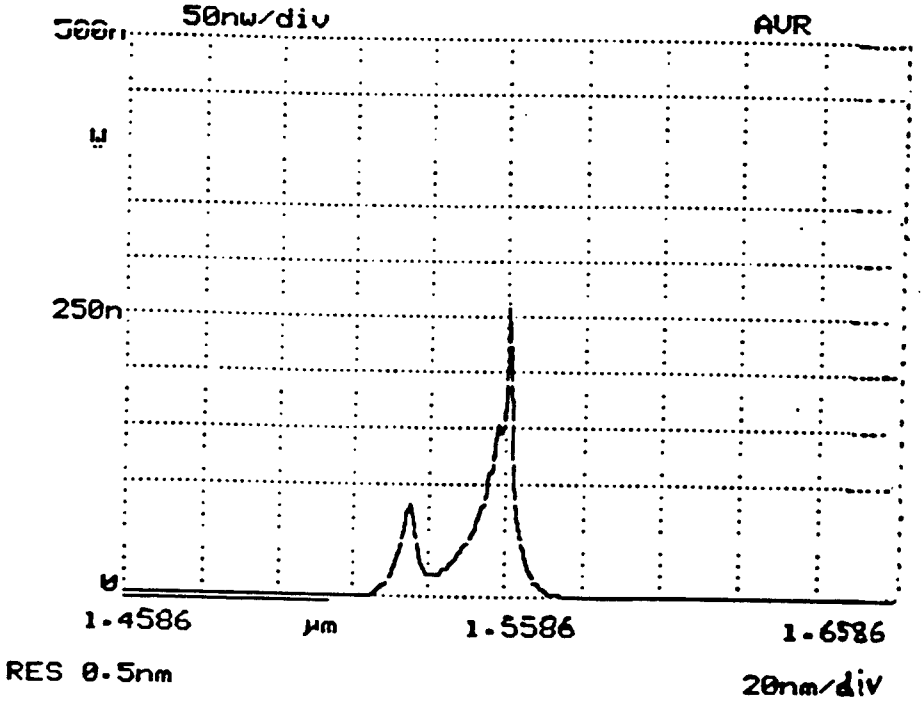


Figure 4. Optical spectrum of oscillator at low pump power.

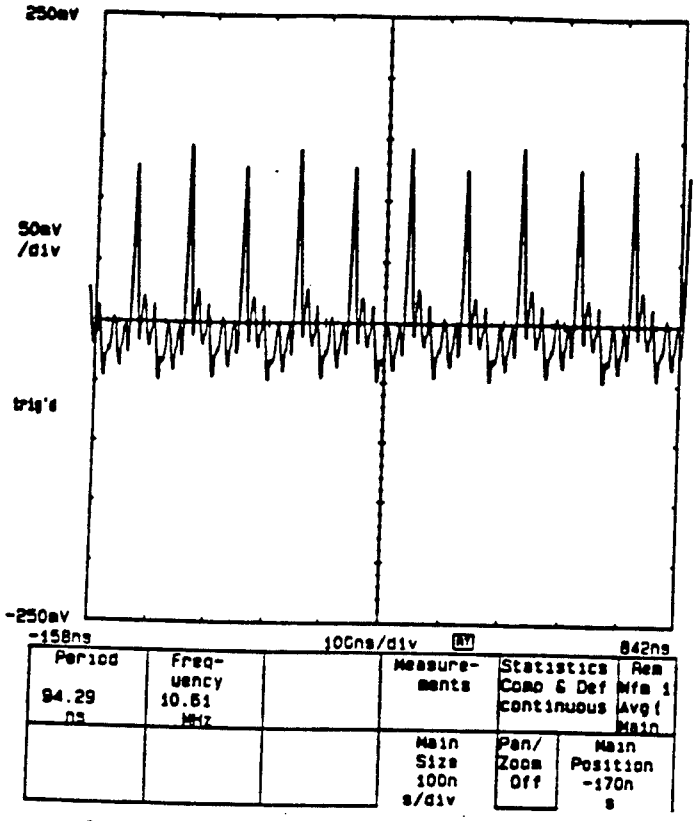


Figure 5. Output waveform of oscillator at high pump power.

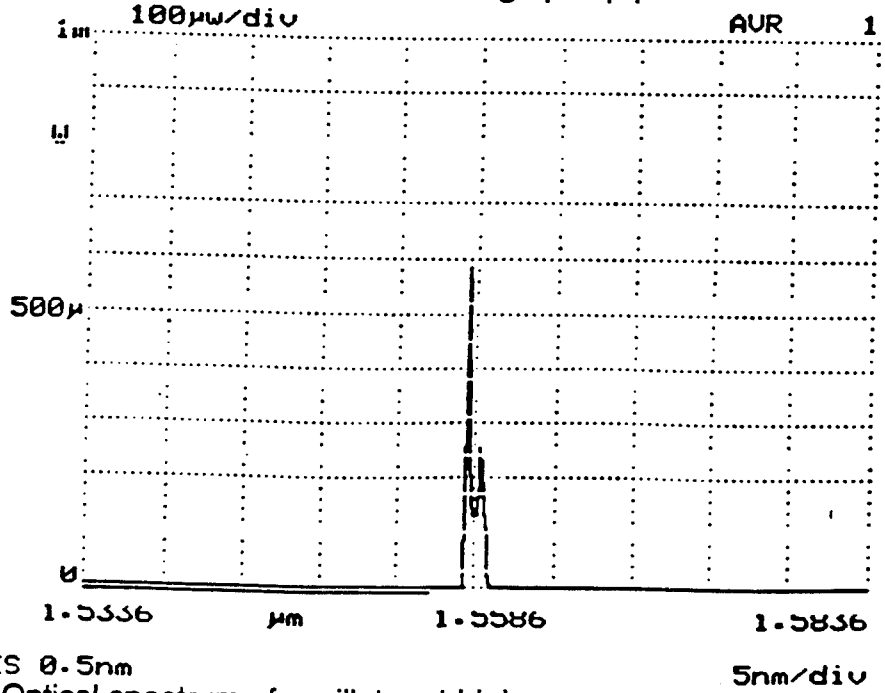


Figure 6. Optical spectrum of oscillator at high pump power.

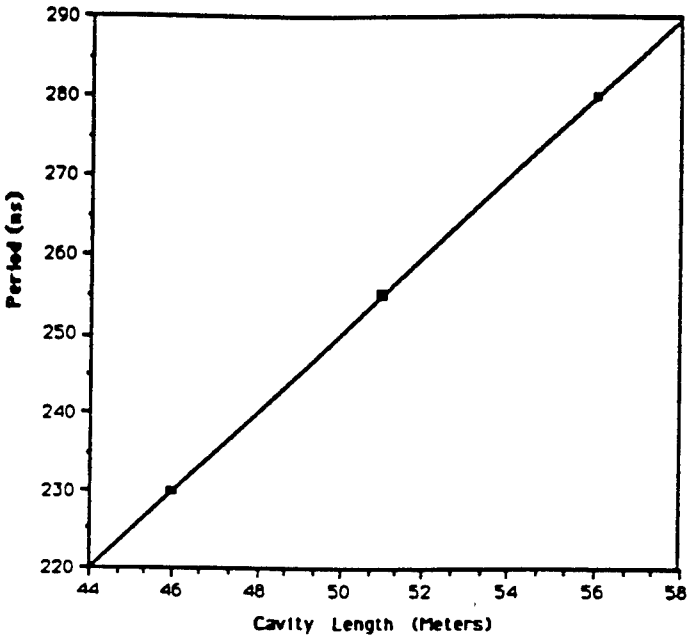


Figure 7. Period of mode-locked pulses as a function of cavity length.

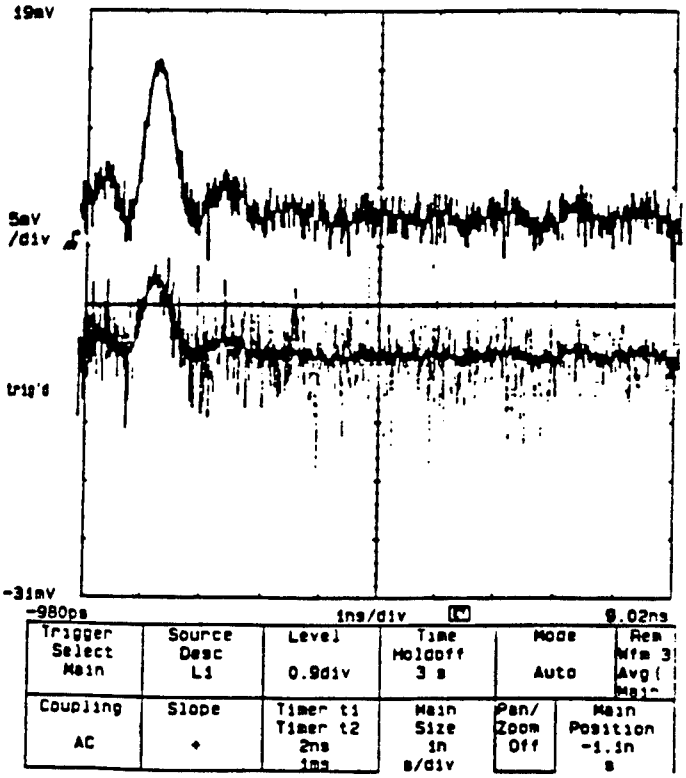


Figure 8. Width of output pulses using a 1.0 GHz detection system.

SIMULATION ANALYSIS OF A LOCAL AREA NETWORK FOR XP AT ROME LABORATORIES

by

Jeffrey D. Tew, Ph.D., Assistant Professor
Industrial and Systems Engineering Department
Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061

Final Report for: Research Initiation Program at Rome Laboratories

Sponsored by: Research Development Laboratories and Virginia Polytechnic Institute and State University

March 1993

SIMULATION ANALYSIS OF A LOCAL AREA NETWORK FOR XP AT ROME LABORATORIES

Jeffrey D. Tew, Ph.D., Assistant Professor

Industrial and Systems Engineering Department

Virginia Polytechnic Institute and State University

Abstract

This document reports on the results of a computer simulation analysis of a projected micro-computer network in the XP directorate at Rome Laboratories. The purposes of the simulation study were: (a) to effectively model the proposed Apple Talk network in XP via a computer simulation model, (b) to perform "what if" tests with the model in order to make recommendations for future additions to the network, and (c) to conduct preliminary tests on the effectiveness of various advanced simulation variance reduction techniques on the network model in order to form a baseline for future work in integrating numerous simulation models already in use at Rome Laboratories. The results of this study are: (a) a validated and accurate computer simulation model of the proposed Apple Talk network in XP at Rome Laboratories that includes a fully documented prescription of statistical analysis procedures for estimating the performance measures of interest from this model, (b) a preliminary study of the important "what if" scenarios to determine their relative value, and (c) a preliminary manual for implementing important variance reduction techniques with this model. The first two results provide XP with a good understanding of the operating characteristics of the Apple Talk network under a number of scenarios of interest and will provide them with the knowledge to determine how to develop the network in the future. The last result will lay a solid foundation for future work in integrating the simulation models at Rome Laboratories.

Simulation Analysis of a Local Area Network for XP
at Rome Laboratories

Jeffrey D. Tew

1. Introduction

This document is a report of a computer simulation analysis of a projected micro-computer network in the XP directorate at Rome Laboratories. There were three main purposes for conducting this study: (a) to effectively model the proposed installed Apple Talk network in XP via a computer simulation model, (b) to perform "what if" tests with the model in order to make recommendations for future additions to the network, and (c) to conduct preliminary tests on the effectiveness of various advanced simulation variance reduction techniques on the network model in order to form a baseline for future work in integrating numerous simulation models already in use at Rome Laboratories.

The construction of an effective computer simulation model of the Apple Talk network involved: (a) conducting an input modelling study of the random components in the model, (b) verification that the computer code which comprises the model is correct, (c) validating the performance of the model against the performance of the real network, and (d) constructing accurate estimators of the desired performance measure characteristics.

The process of conducting preliminary tests on the effectiveness of various variance reduction techniques on this model involved conducting numerous Monte Carlo comparision studies with the model in order to measure the percent variance reduction achieved with this model for each of the performance measures of interest and for each of the variance reduction techniques considered. This study focused on the variance reduction techniques of: (a) common random numbers, (b) antithetic variates, and (c) control variates.

Together, these three phases of the study will yielded: (a) a fully validated and accurate computer simulation model of the existing Apple Talk network in XP at Rome Laboratories that includes a fully documented prescription of statistical analysis procedures for estimating the performance measures of interest from this model, (b) an study of the important "what if" scenarios to determine their relative value, and (c) a preliminary manual for implementing important variance reduction techniques with this model. The last result will lay a solid foundation for future work in integrating the simulation models at Rome Laboratories.

In the next section, we give a brief description of the existing Apple Talk network in XP. In Section 3 we outline the development of the computer simulation model used in this study. Section 4 gives a brief discussion of how the "what if" tests were be performed as well as a discussion and analysis of the numerical results. Section 5 gives a detailed development of the variance reduction techniques that were developed as part of this project.

2. XP Computer Network

The XP directorate at Rome Laboratories is commissioned with the task of providing Rome Laboratories with cross-directorate strategic planning and technical program support. A critical component which enables the XP personnel to effectively carry-out that task is the ability to efficiently communicate on a technical software basis with one another and with those located in other directorates on base as well as other facilities throughout the world.

Currently, it is felt that the level this type of communication made available to XP personnel is inadequate. That is to say, that, increasingly, the XP personnel are finding it more difficult to communicate on a technical software basis with their colleagues. Consequently, it has been felt that improved computer communications capabilities among

the XP personnel would solve problem and bring about an immediate improvement in the level of technical software interfaceing among the XP personnel. To that end, the director of XP has initiated the installation of a 30 workstation Machintosh network operating under the Apple Talk protocal environment. This initial installation process began in August 1991 and was expected to be completed by 1 December 1991 (the network is unoperational at this writing). The installation process has involved 5 major steps indicated below:

1. Physical installation and connection
2. Setup of Macintosh equipment
3. Upgrade and setup of printing resources
4. Setup of DOS equipment
5. Connection to external networks

Figure 1 gives a diagramatic representation of the network. The network configuration is basically that of a backbone network with nodes connecting directly to the backbone. Traffic handling on the backbone is to be done by a central server called a Star Controller. Nodes on the network are of two types: (a) Macintosh machines and (b) DOS machines. Initially, most nodes will be Macintosh type machines (typically, Macintosh IIci with 4 Mb RAM). Although, provision will be made for interfacing with DOS-based machines, it is anticipated that most nodes in the future will continue to be Macintosh type machines. Traffic on the network will be of three main types: (a) data, (b) video, and (c) acknowledgement.

The Star Controller will also serve as a node on the Rome Laboratories interdirectorate LONEX computer network which in itself serves as a node on the Internet network.

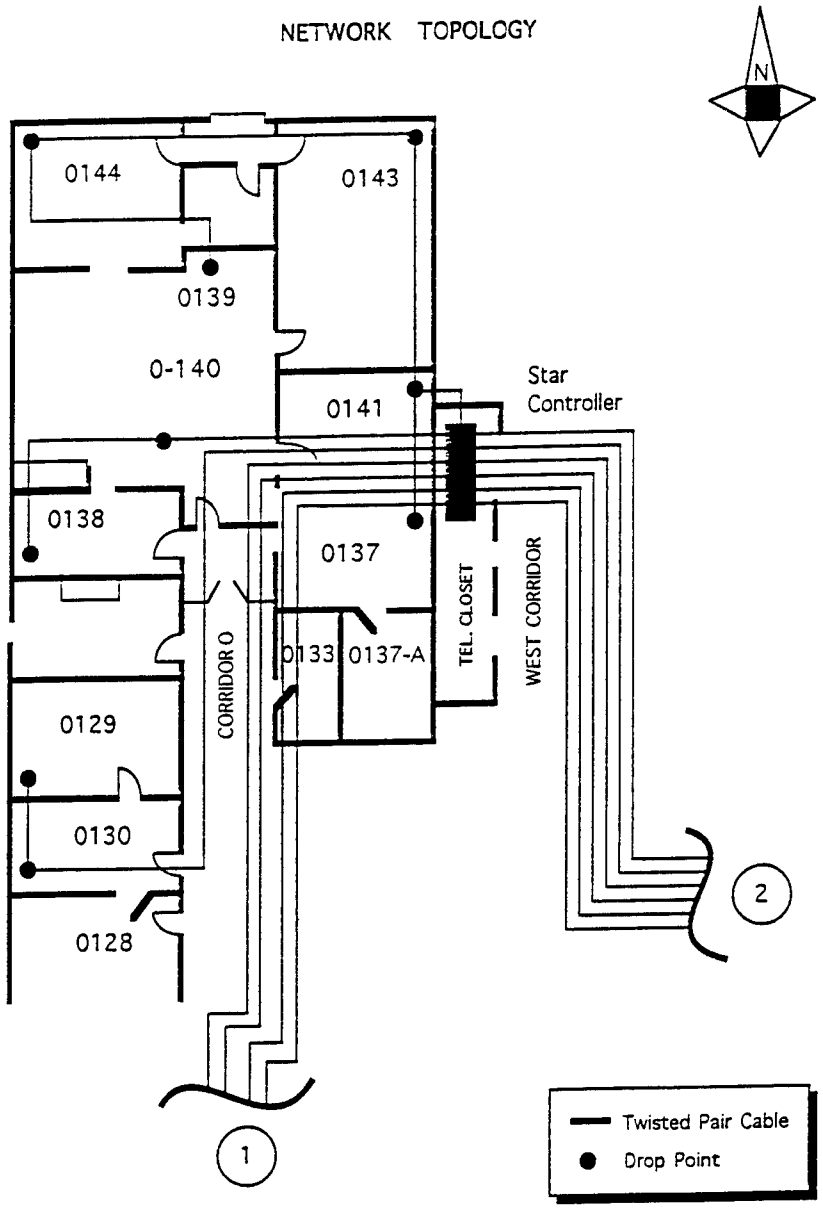


Figure 1(a): XP Macintosh Network

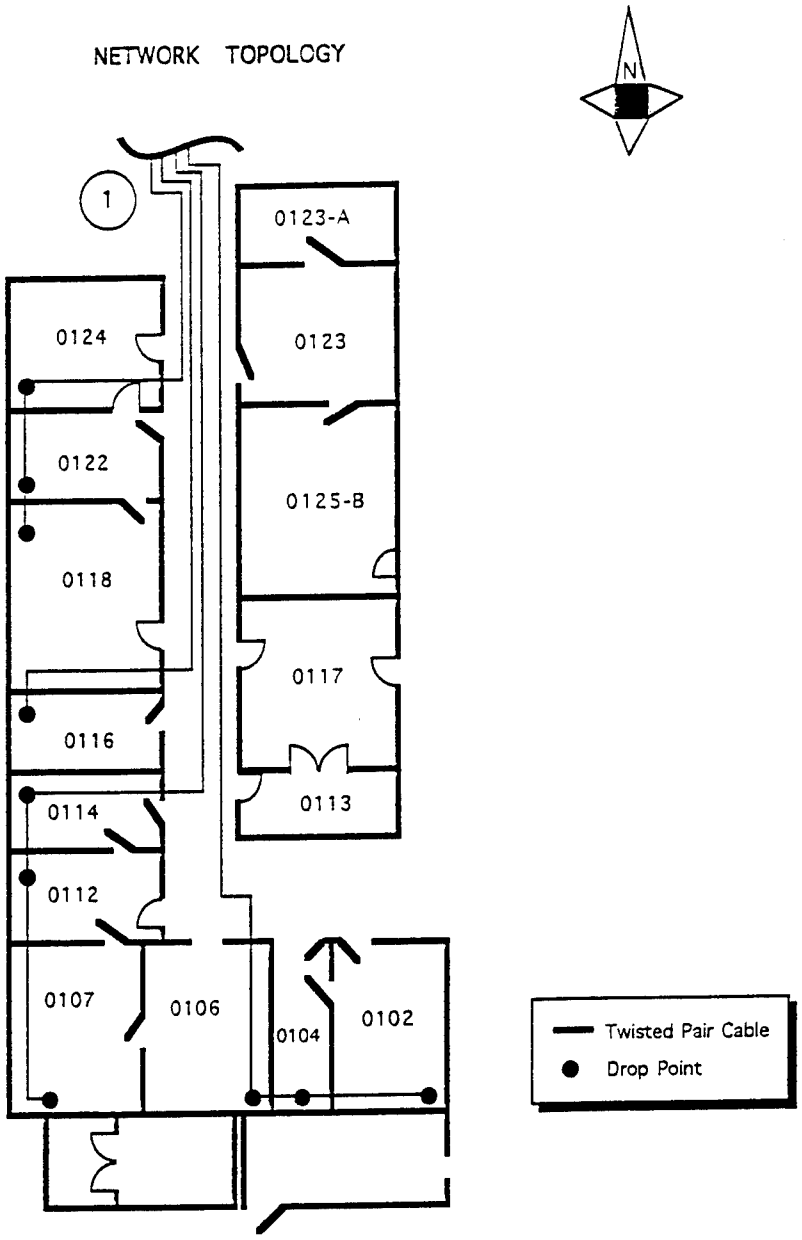


Figure 1(b): XP Macintosh Network

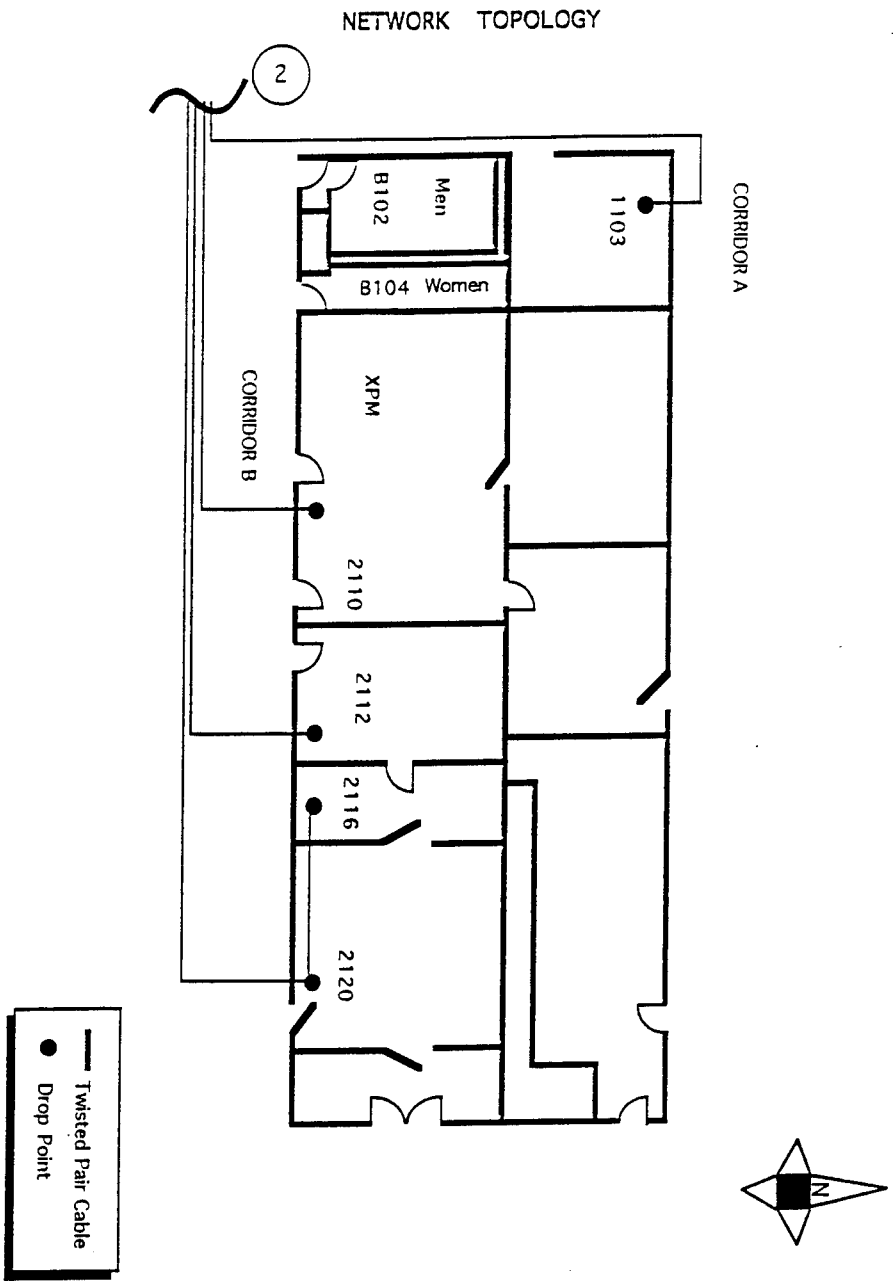


Figure 1(c): XP Macintosh Network

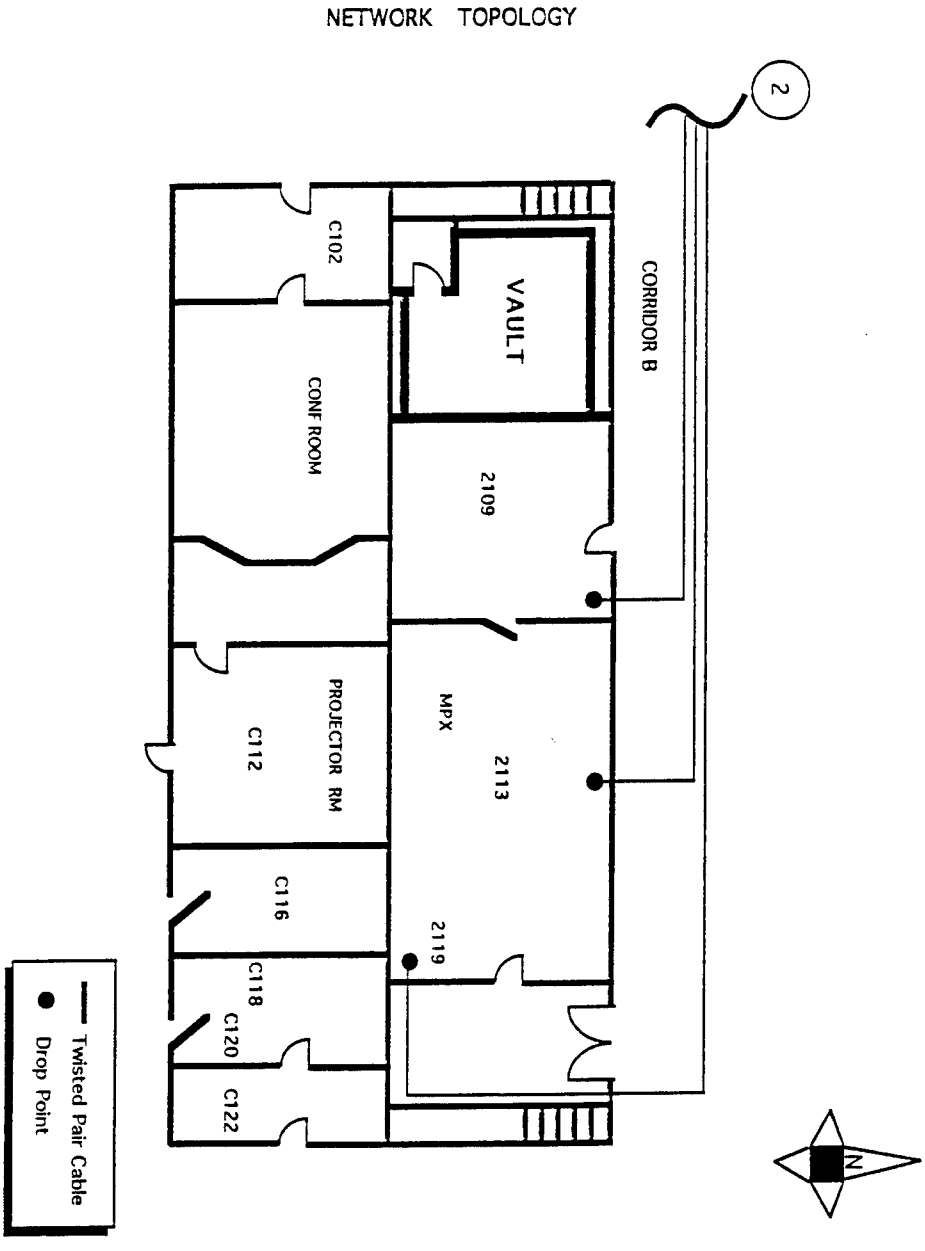


Figure 1(d): XP Macintosh Network

3. Simulation Model

The actual construction of the simulation model was carried out by performing the following eight steps:

1. Problem identification
2. Specification of objectives
3. Definition of the system
4. Validation of objectives and system definition
5. Model formulation
6. Model verification
7. Model validation
8. Model implementation

These eight steps are often recognized as standard means for building effective simulation models in the literature. Chapter 5 of Law and Kelton (1990), Chapter 3 of Pritsker (1986), and Chapters 1 and 4 of Bratley, Fox, and Schrage (1987) all offer state-of-the-art discussions and the topic of effective simulation model building. Basically, none of them differ by much from the presribed approach given above. Next we give a brief discussion of the implementation of each of these eight steps in the development of the simulation model.

1. **Problem Identification**: This step of the process involved extensive discussions with XP personnel in order to determine the actual structure of the network as well as the determination of which performance measure and input factors were

of most interest. Through these discussions, we were able to determine that: (a) the mean time in the network for all packets was the performance measure of most interest (y), (b) the mean think time between user commands on the network (x_1) and the number of CPU units at the Star Controller (x_2) were the two input factors of most interest, and (c) one hour's operation during a peak user's period was the time frame of interest.

2. **Specification of Objectives**: In this phase we determined that the mean packet time in the network was of most interest.

3. **Definition of System**: The logical flow of the packets was determined to be the following: packets are generated (created) at each user on the network. There are two kinds: (a) those that are to be routed to another user on the network and (b) those that are to be sent to the Star Controller. For purposes of this initial model it was determined that packet traffic leaving the network and coming into the network from other networks would not be included in the model.

4. **Model Formulation**: As indicated above, the simulation model was coded in the SLAM II simulation software. A description of this code is given later in this section. Since the system was not operational at the time of the simulation model development, we decided to use traffic composition distributions and interarrival time distributions provided in the literature.

5. **Model Verification**: Upon completion of the model code, we tested the model under different scenarios to determine if the code was performing correctly.

6. **Model Validation**: Again, since the network was not operational at the time of the development of the simulation model, we used the *face validity* of the model outputs as determined by some of the personnel at XP.

7. **Model Implementation**: This involved two primary parts: (a) selection and testing of appropriate statistical estimation methodologies for *each* performance measure of interest and (b) establishing tactical simulation run characteristics (that is, how long is each simulation run to be and how do we start each model run). The former is discussed in detail in Section 4. Consideration of the latter part resulted in determining that a simulation run was to be for one hour of simulation time with the model starting empty and idle and with a thirty minute burn-in period to control for initialization bias.

Next, we discuss the SLAM II code of the simulation model.

A copy of the SLAM II code for the simulation model of the XP micro-computer network is given in Appendix I. This model was developed using the Network modeling component in SLAM II (see Chapters 5, 6, and 7 of Pritsker 1986). We used 27 separate CREATE nodes to model the 27 nodes on the network (see Figure 1). Each used an Exponential distribution to model the time between arrivals (creations) of messages to the network. Upon creation, a number from a the Poisson distribution with mean 11 is generated and represents the number of *frames* in that particular packet. Each frame consists of 1526 bytes of information. Each message is sent immediately to the Star Controller for processing.

At the Star Controller messages are either: (a) processed and a return set of messages is sent back to the point of origination or (b) the newly created message is routed to its destination. Thus, we have two basic types of messages that are created at each of the 27 users (nodes) on the network: (a) those that are to be sent to other users on the network and (b) those that are to be sent to the Star Controller for processing. 25% of the created messages are to go to the Star Controller and 75% are to be routed to one of the other 26 nodes. In our model, we used a discrete uniform distribution to determine the destination of a message. Since the XP network was not operational during the time

of this study, all distributions used in this study were taken from the existing literature on successful simulation studies of similar Local Area Netorks (LAN) (see Mills 1988; Devai, Kerrigan, and Molloy 1990; and Kavi, Frost, and Shanmugan 1991). Waiting for processing at the Star Controller is done in a single queue until the message is processed and sent back onto the network to its point of termination (note, all transmission times to and from the Star Controller were taken to be .00003 seconds). Message processing time at the Star Controller is .00003 times the total number of bytes in the message and waiting in queue is done on a FIFO queue priority basis.

Upon reaching its point of termination, statistics on the time spent in the network and the total number of bytes contained in the message are collected at each termination nodel as well as accumulated across all termination nodes. In addition, the standard set of statistical outputs are generated by SLAM II Summary Report for this model. An example of such an output report is given in Appendix II. Among other things, this report gives statistics on the average queue length at the Star Controller queue, the average utilization of the Star Controller processor, and the total number of bytes processed , among others. In the next section, we discuss the experimentation that was done with this simulation model.

4. Simulation Experiments

Once the simulation model was developed, a number of experiments were performed. Two of these are summarized below. The personnel at XP were interested in: (a) the mean time spent in the network by a message (y), (b) the mean interarrival, in seconds, (creation) time of messages at each of the 27 user nodes on the network (x_1), and (c) the number of Star Controller CPU's (x_2). The first of these is the response variable of interest and the latter two are the input factors of interest.

First, we considered the relationship between y and x_1. For x_1, 4 different levels were

chosen that, together, represent the design space of interest to XP. The levels chosen were $x_1 = 5, 8, 11, 14$ seconds. For each level of x_1 a set of 5 simulation runs was made. The statistical software package SAS was used to perform a regression analysis on the data in order to ascertain whether a second-order model adequately described the relationship between $E[y]$ and x_1. Specifically, the SAS procedure PROC RSREG (see the SAS manual) was used to carry out the calculations. The results of this process are given in Figure 2.

From these results we see that a second-order model does describe the desired relationship well (the lack-of-fit test for the second-order model has a p-value of, approximately, 0.0000). Thus, we concluded that the relationship between $E[y]$ and x_1 is given as:

$$E[y] = 4354.54 - 476.14x_1 + 14.07x_1^2$$

From this fitted quadratic model, a canonical analysis was also performed to determine that the minimum response (y) was attained at the point $x_1 = 16.92$ and that it resulted in a minimum response value of $y_{\text{min}} = 326.95$ seconds.

Second, we considered the relationship between y and x_2. For x_2, again, 4 different levels were chosen that, together, represent the design space of interest to XP. The levels chosen were $x_2 = 1, 2, 3, 4$ seconds. For each level of x_2 a set of 5 simulation runs was made. The statistical software package SAS was used to perform a regression analysis on the data in order to ascertain whether a second-order model adequately described the relationship between $E[y]$ and x_2. Specifically, the SAS procedure PROC RSREG (see the SAS manual) was used to carry out the calculations. The results of this process are given in Figure 3.

As before, from these results we see that a second-order model does describe the desired relationship well (the lack-of-fit test for the second-order model has a p-value of, approximately, 0.0000). Thus, we concluded that the relationship between $E[y]$ and x_1

is given as:

$$E[y] = 4441.83 - 2444.37x_2 + 331.93x_2^2$$

From this fitted quadratic model, a canonical analysis was also performed to determine that the minimum response (y) was attained at the point $x_1 = 3.68$ and that it resulted in a minimum response value of $y_{\min} = -58.28$ seconds; which, of course, is not feasible. However, clearly a quadratic model of this relationship is appropriate and that an addition of one extra Star Controller will significantly reduce the mean time that a message spends in the system.

These two regression models, and others like them constructed from the existing SLAM II simulation code, can be used to guide XP in designing future additions to the network. Based on the results of presented here and under the current message traffic patterns on the network, we would recommend that at least two Star Controller CPU units be installed.

For each regression model described above, the variance reduction technique entitled Combined Method III was used to effecively reduce the variance of the estimated regression coefficients. The observed variance reduction was on the order of at least 76% in all cases. A full discussion of the variance reduction techniques developed as a part of this project is given in the next section.

5. Variance Reduction Techniques

In this section, we describe the variance reduction methodologies that were developed in order to improve the estimation of the regression model coefficients presented in the previous section.

Response Surface for Variable Y: MEAN TIME IN SYSTEM

Response Mean	1259.550000
Root MSE	65.910032
R-Square	0.9928
Coef. of Variation	5.2328

Regression	Degrees of Freedom	Type I Sum of Squares	R-Square	F-Ratio	Prob > F
Linear	1	9806292	0.9613	2257.4	0.0000
Quadratic	1	320804	0.0314	73.848	0.0000
Crossproduct	0	0	0.0000	.	.
Total Regress	2	10127097	0.9928	1165.6	0.0000

Residual	Degrees of Freedom	Sum of Squares	Mean Square	F-Ratio	Prob > F
Lack of Fit	1	68906	68906	223.0	0.0000
Pure Error	16	4944.000000	309.000000		
Total Error	17	73850	4344.132353		

Parameter	Degrees of Freedom	Parameter Estimate	Standard Error	T for H0: Parameter=0
INTERCEPT	1	4354.538889	136.730838	31.848
X1	1	-476.138889	31.422151	-15.153
X1*X1	1	14.072222	1.637548	8.593

Parameter	Prob > |T|	Parameter Estimate from Coded Data
INTERCEPT	0.0000	1101.237500
X1	0.0000	-939.450000
X1*X1	0.0000	284.962500

Factor	Degrees of Freedom	Sum of Squares	Mean Square	F-Ratio	Prob > F
X1	2	10127097	5063548	1165.6	0.0000

Figure 2(a): SAS Results

MEAN TIME IN SYSTEM VERSUS MEAN INTERARRIVAL TIME

Canonical Analysis of Response Surface
(based on coded data)

Factor	Critical Value Coded	Uncoded	
X1	1.648375	16.917687	MEAN INTERARRIVAL TIME

Predicted value at stationary point 326.954654

Eigenvalues	Eigenvectors X1
284.962500	1.000000

Stationary point is a minimum.

Figure 2(b): SAS Results

Response Surface for Variable Y: MEAN TIME IN SYSTEM

Response Mean	820.403500
Root MSE	69.938516
R-Square	0.9953
Coef. of Variation	8.5249

Regression	Degrees of Freedom	Type I Sum of Squares	R-Square	F-Ratio	Prob > F
Linear	1	15394001	0.8707	3147.2	0.0000
Quadratic	1	2203597	0.1246	450.5	0.0000
Crossproduct	0	0	0.0000	.	.
Total Regress	2	17597598	0.9953	1798.8	0.0000

Residual	Degrees of Freedom	Sum of Squares	Mean Square	F-Ratio	Prob > F
Lack of Fit	1	82902	82902	5268.3	0.0000
Pure Error	16	251.776120	15.736008		
Total Error	17	83154	4891.396085		

Parameter	Degrees of Freedom	Parameter Estimate	Standard Error	T for H0: Parameter=0
INTERCEPT	1	4441.830500	87.072751	51.013
X2	1	-2444.371300	79.434885	-30.772
X2*X2	1	331.933500	15.638728	21.225

Parameter	Prob > |T|	Parameter Estimate from Coded Data
INTERCEPT	0.0000	405.486625
X2	0.0000	-1177.055700
X2*X2	0.0000	746.850375

Factor	Degrees of Freedom	Sum of Squares	Mean Square	F-Ratio	Prob > F
X2	2	17597598	8798799	1798.8	0.0000

Figure 3(a): SAS Results

MEAN TIME IN SYSTEM VERSUS NUMBER OF CPU UNITS 3

Canonical Analysis of Response Surface
(based on coded data)

Factor	Critical Value Coded	Uncoded	
X2	0.788013	3.682020	NUMBER OF CPU UNITS

Predicted value at stationary point -58.281007

Eigenvalues	Eigenvectors X2
746.850375	1.000000

Stationary point is a minimum.

Figure 3(b): SAS Results

5.1 Notation

In a designed simulation experiment, the analyst is usually concerned with estimating the mean of a univariate response of interest based on the outputs of a carefully structured set of runs. Frequently, large-scale simulation studies require extensive experimentation to obtain acceptable precision in the estimator of interest. If we can reduce the variance of this estimator at little additional cost, then we can also reduce the requisite amount of experimentation. In this section we propose three ways of combining the two standard variance reduction techniques of antithetic variates and control variates to yield more precise estimators of the mean response.

For a single-model simulation experiment (i.e., one system configuration or design point), the methods of antithetic variates and control variates are probably among the most commonly applied variance reduction techniques (Law and Kelton 1991, Chapter 11). The method of antithetic variates assigns complementary streams of random numbers to pairs of simulation runs taken at a single design point to induce a negative correlation between the corresponding responses. Let y_1 and y_2 denote two responses obtained by antithetic replicates of a single design point. Suppose that we estimate $\mu_y = \mathrm{E}(y_i)$ $(i = 1, 2)$ by the sample mean response $\frac{1}{2}(y_1 + y_2)$. Then we observe that in general,

$$\mathrm{var}[\frac{1}{2}(y_1 + y_2)] = \frac{1}{4}\mathrm{var}(y_1) + \frac{1}{4}\mathrm{var}(y_2) + \frac{1}{2}\mathrm{cov}(y_1, y_2) = \frac{1}{2}\mathrm{var}(y_1) + \frac{1}{2}\mathrm{cov}(y_1, y_2).$$

In this equation, if the covariance between y_1 and y_2 obtained by antithetic replicates is negative, then the variance of the sample mean is less than that obtained by two independent replicates (for a more detailed discussion of antithetic variates see Section III.6 of Kleijnen 1974, Section 2.2 of Bratley, Fox, and Schrage 1987, and Section 11.3 of Law and Kelton 1991).

In contrast to the approach of antithetic variates, the method of control variates attempts to exploit *intrinsic* correlations between the target response and selected auxil-

iary outputs (control variates) when all of these quantities are generated within a single run. Let y_i and $\mathbf{c}_i$ respectively denote the response of interest and the $s \times 1$ vector of control variates obtained from the ith simulation run. We assume that ($i = 1, 2, \ldots, 2h$):

$$\begin{bmatrix} y_i \\ \mathbf{c}_i \end{bmatrix} \sim \mathrm{N}_{s+1}\left(\begin{bmatrix} \mu_y \\ \mathbf{0}_{(s\times 1)} \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \boldsymbol{\sigma}'_{y\mathbf{c}} \\ \boldsymbol{\sigma}_{y\mathbf{c}} & \boldsymbol{\Sigma}_{\mathbf{c}} \end{bmatrix} \right), \tag{1}$$

where σ_y^2 is the unconditional variance of y_i, $\boldsymbol{\sigma}_{y\mathbf{c}}$ is the $s \times 1$ covariance vector between y_i and $\mathbf{c}_i$, and $\boldsymbol{\Sigma}_{\mathbf{c}}$ is the $s \times s$ covariance matrix of $\mathbf{c}_i$. Also, without loss of generality, we assume $\mathrm{E}(\mathbf{c}_i) = \mathbf{0}_{(s\times 1)}$. In the context of performing $2h$ independent replications of the simulation, joint normality of the response and the controls ensures that the response can be represented by the following linear model:

$$\mathbf{y} = \mu_y \mathbf{1}_{(2h\times 1)} + \mathbf{C}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_{2h})'$, $\mathbf{1}_{(2h\times 1)}$ is a $2h \times 1$ vector of 1's, $\mathbf{C}$ is a $2h \times s$ control variate matrix whose ith row consists of $\mathbf{c}_i'$, μ_y is the mean response, $\boldsymbol{\alpha} = \boldsymbol{\Sigma}_{\mathbf{c}}^{-1}\boldsymbol{\sigma}_{y\mathbf{c}}$ is the $s \times 1$ vector of control coefficients, and $\boldsymbol{\epsilon}$ is the $2h \times 1$ vector of error terms (see Lavenberg, Moeller, and Welch 1982). The least squares estimators of $\boldsymbol{\alpha}$ and μ_y in the linear model (2) are given by, respectively:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{C}'\mathbf{P}\mathbf{C})^{-1}\mathbf{C}'\mathbf{P}\mathbf{y} \text{ and } \hat{\mu}_y = \bar{y} - \bar{\mathbf{c}}'\hat{\boldsymbol{\alpha}}, \tag{3}$$

where $\bar{y} = (2h)^{-1}\mathbf{1}'_{(1\times 2h)}\mathbf{y} = (2h)^{-1}\sum_{i=1}^{2h} y_i$ and $\bar{\mathbf{c}}' = (2h)^{-1}\mathbf{1}'_{(1\times 2h)}\mathbf{C} = (2h)^{-1}\sum_{i=1}^{2h}\mathbf{c}_i'$ respectively denote the sample mean of the responses and the control vectors computed across $2h$ replications, and $\mathbf{P} = \mathbf{I}_{(2h\times 2h)} - \frac{1}{2h}\mathbf{1}_{(2h\times 1)}\mathbf{1}'_{(1\times 2h)}$ (see Searle 1971, p. 341). Under the normality assumption in (1), the components of the error vector $\boldsymbol{\epsilon}$ are IID $N(0, \sigma^2_{y|\mathbf{c}})$ where

$$\sigma^2_{y|\mathbf{c}} = \mathrm{var}(y_i|\mathbf{c}_i) = (\sigma_y^2 - \boldsymbol{\sigma}'_{y\mathbf{c}}\boldsymbol{\Sigma}_{\mathbf{c}}^{-1}\boldsymbol{\sigma}_{y\mathbf{c}}) \tag{4}$$

(see Theorem 2.5.1 of Anderson 1984); and the least squares estimator $\hat{\mu}_y$ is an unbiased estimator for μ_y. Lavenberg, Moeller, and Welch (1982) showed that the unconditional variance of $\hat{\mu}_y$ is given by

$$\text{var}(\hat{\mu}_y) = \left(\frac{2h-2}{2h-s-2}\right)(1-R^2_{y\mathbf{c}})\frac{\sigma^2_y}{2h}, \tag{5}$$

where $R^2_{y\mathbf{c}} = \sigma_y^{-2}\boldsymbol{\sigma}'_{y\mathbf{c}}\boldsymbol{\Sigma}_{\mathbf{c}}^{-1}\boldsymbol{\sigma}_{y\mathbf{c}}$ is the square of the multiple correlation coefficient between y_i and $\mathbf{c}_i$. Lavenberg, Moeller, and Welch also defined the quantity $\frac{2h-2}{2h-s-2}$ as the loss factor due to the estimation of the unknown control coefficient vector $\boldsymbol{\alpha}$ in (2); and they identified $(1-R^2_{y\mathbf{c}})$ as the minimum variance ratio which represents the potential for reducing the variance of the estimator of μ_y by the control variates. Thus, the efficiency of control variates is measured by the product of the loss factor and the minimum variance ratio.

In this research, our main interest is to obtain a more precise estimator of the mean response μ_y by combining the variance reduction techniques of antithetic variates and control variates. These two methods utilize correlations between simulation outputs that are generated either from a single run or from different runs in a designed simulation experiment. Suppose that through correlated replications of a given simulation model, we get a reduced variance for the estimator of the mean response and yet we maintain the same correlation between the response y_i and the control vector $\mathbf{c}_i$ as we obtained with independent replications; then it is conjectured that we may apply the methods of antithetic variates and control variates jointly, thereby obtaining greater efficiency increases than could be obtained by applying either antithetic variates or control variates separately. (Note that Avramidis and Wilson (1992) establish general conditions for effective combination of the methods of control variates, correlation induction, and conditional expectation for finite-horizon simulation experiments.)

Based on this conjecture, this research focuses on developing methods for combining antithetic variates and control variates for the estimation of the mean univariate simula-

tion response. For this purpose, we consider three ways of utilizing correlations between: (a) multiple replications of the response of interest, (b) the response and a set of control variates observed on a single replication, and (c) multiple replications of the control variates. Some of these correlations are induced by an appropriate assignment of random number streams to different replications; and by using such correlation-induction strategies, we try to improve upon the simulation efficiency of the classical control variates method in which all replications are independent. The three combined methods differ only in which elements of the model that the method of antithetic variates is applied to. There are three obvious choices: (a) use control variates in conjuction with antithetic variates, where the method of antithetic variates is only applied to the response y, (b) use control variates in conjuction with antithetic variates, where the method of antithetic variates is applied to just the control variates, and (c) use control variates in conjuction with antithetic variates, where the method of antithetic variates is applied to both the response y *and* the control variates. These constitute the three combined methods discussed in Section 2.

In the next subsection we develop the three combined methods for jointly applying control variates and antithetic variates.

5.2 Simulation Efficiency of Combined Methods

In computer simulation, random number streams that drive a simulation model are under the control of the experimenter and *completely* determine the simulation output. Let the random number stream $\mathbf{r}_{ij} = (r_{ij1}, r_{ij2}, \ldots)'$ denote the potentially infinite sequence of random numbers used to drive the jth random component in the simulation model at the ith replicate, where $\{r_{ijl} : l = 1, 2, \ldots\}$ are IID $U(0,1)$ variates. The random components in a queueing simulation model may include, among other things, the sequence of service times sampled at a particular service center, the sequence of interarrival times sampled

from a given arrival process to a network, etc. Here, we are saying that, to each random component in the simulation model, we assign a separate random number stream to generate realizations of the corresponding stochastic simulation input process. We assume that g random number streams are required to drive the entire simulation model, and we let $\mathbf{R}_i$ denote the complete set of streams used for the ith replication of the model:

$$\mathbf{R}_i = (\mathbf{r}_{i1}, \mathbf{r}_{i2}, \ldots, \mathbf{r}_{ig}) \text{ for } i = 1, 2, \ldots, 2h.$$

We now consider the random number assignment strategy of jointly utilizing antithetic variates and control variates to estimate the mean response μ_y. To this end, we separate $\mathbf{R}_i$ into two mutually exclusive and exhaustive subsets of random number streams, such that $\mathbf{R}_i \equiv (\mathbf{R}_{i1}, \mathbf{R}_{i2})$ $(i = 1, 2, \ldots, 2h)$. The first subset, $\mathbf{R}_{i1}$, consisting of $(g - s^\star)$ random number streams, is used to drive the non-control-variate random components in the model so that $\mathbf{c}_i$ is independent of $\mathbf{R}_{i1}$. The second subset, $\mathbf{R}_{i2}$, consisting of $s^\star$ random number streams, is used to drive the control-variate random components in the model so that $\mathbf{c}_i$ is a function of $\mathbf{R}_{i2}$. These properties are summarized as follows:

$$\mathbf{c}_i, \mathbf{R}_{i1} \text{ are independent and } \mathbf{c}_i = \mathbf{c}_i(\mathbf{R}_{i2}) \quad \text{for } i = 1, 2, \ldots, 2h.$$

Applying the method of antithetic variates to the appropriate random components in the simulation model may induce correlations between: (a) responses, (b) control variates, and (c) responses *and* control variates, across replicates.

Specifically, we consider the following methods: (a) use antithetic variates for all random components except the control variates, (b) use antithetic variates on only the control variates, and (c) use antithetic variates for all random components. Through statistical analysis and simulation experimentation, we will explore how these methods may improve the simulation efficiency in reducing the variance of the estimator, and what conditions are necessary for each method to ensure an improvement in variance reduction.

5.2.1 Combined Method I

This section presents a method for combining antithetic variates and control variates based on correlated replicates utilizing only the non-control-variate random components in the model. Recall from the discussion given in the Introduction, the basic idea of this method (as well as the methods presented in the next two sections) is to group the replicates into h antithetic pairs. Within the jth pair of replicates, Combined Method I uses $(\mathbf{R}_{2j-1,1}, \mathbf{R}_{2j-1,2})$ as the input to run $2j-1$ and $(\bar{\mathbf{R}}_{2j-1,1}, \mathbf{R}_{2j,2})$ as the input to run $2j$, where $\mathbf{R}_{2j-1,1}$, $\mathbf{R}_{2j-1,2}$, and $\mathbf{R}_{2j,2}$ are mutually independent sets of random number streams, and $\bar{\mathbf{R}}_{2j-1,1}$ is the set of random number streams that are antithetic (complimentary) to those which comprise $\mathbf{R}_{2j-1,1}$ $(j = 1, 2, \ldots, h)$: $\bar{\mathbf{R}}_{2j-1,1} \equiv (\bar{\mathbf{r}}_{2j-1,k} : k = 1, 2, \ldots, g - s^\star)$, where $\bar{\mathbf{r}}_{2j-1,k}$ denotes the random number stream complimentary to $\mathbf{r}_{2j-1,k}$ so that

$$\begin{aligned} \bar{\mathbf{r}}_{2j-1,k} &\equiv \mathbf{1} - \mathbf{r}_{2j-1,k} \\ &= \begin{pmatrix} 1 - r_{2j-1,k,1} \\ 1 - r_{2j-1,k,2} \\ \vdots \end{pmatrix} \quad \text{for } k = 1, 2, \ldots, g - s^\star. \end{aligned}$$

Across pairs of replicates, this method uses independent random number streams. Thus, the jth pair of responses, y_{2j-1} and y_{2j} $(j = 1, 2, \ldots, h)$, are negatively correlated by the use of antithetic streams on the non-control-variate random components. However, across *all* $2h$ replicates, the control variates $\mathbf{c}_i$ $(i = 1, 2, \ldots, 2h)$ are independently generated by the assignment of randomly selected random number streams, $\mathbf{R}_{i,2}$, used to drive the control-variate random components for each replicate. Because we randomly select the random number streams to drive the control variates, the response y_i $(i = 1, 2, \ldots, 2h)$ is independent of the control variate vector $\mathbf{c}_k$ $(k = 1, 2, \ldots, 2h$ and $k \neq i)$. Based on the above discussion, we establish the following properties $(i = 1, 2, \ldots, 2h)$:

Property 1: homogeneity of response variances across replicates,

$$\text{var}(y_i) = \sigma_y^2, \quad i = 1, 2, \ldots, 2h. \tag{6}$$

Property 2: homogeneity of negative correlations across replicate pairs and independence of responses between replicate pairs,

$$\left.\begin{aligned} &\text{cov}(y_k, y_l) = -\rho_1 \sigma_y^2 \text{ for } k = l+1 \ \ (l = 1, 3, \ldots, 2h-1); \\ &\text{cov}(y_k, y_l) = 0 \text{ otherwise.} \end{aligned}\right\} \tag{7}$$

Property 3: homogeneity of response-control variates covariances across replicates and independence of response-control variates covariances between replicates,

$$\left.\begin{aligned} &\text{Cov}(y_i, \mathbf{c}_i) = \boldsymbol{\sigma}_{y\mathbf{c}}, \quad i = 1, 2, \ldots, 2h; \\ &\text{Cov}(y_i, \mathbf{c}_k) = \mathbf{0}_{(s\times 1)}, \quad \text{for } i \neq k. \end{aligned}\right\} \tag{8}$$

Property 4: homogeneity of control variates covariances across replicates and independence of control variates between replicates,

$$\left.\begin{aligned} &\text{Cov}(\mathbf{c}_i) = \boldsymbol{\Sigma}_{\mathbf{c}}, i = 1, 2, \ldots, 2h; \\ &\text{Cov}(\mathbf{c}_i, \mathbf{c}_k) = \mathbf{0}_{(s\times s)} \text{ for } i \neq k. \end{aligned}\right\} \tag{9}$$

We make the following assumption

$$\textbf{Assumption}: \quad 0 < \rho_1 < 1.$$

Under the four properties listed above, the variance of the mean response, $\bar{y}_j = \frac{1}{2}(y_{2j-1} + y_{2j})$, and mean control variate vector, $\bar{\mathbf{c}}_j = \frac{1}{2}(\mathbf{c}_{2j-1} + \mathbf{c}_{2j})$, within the jth replicate pair $(j = 1, 2, \ldots, h)$ are respectively given by:

$$\text{var}(\bar{y}_j) = \frac{1}{4}(\text{var}(y_{2j-1}) + \text{var}(y_{2j}) + 2\text{cov}(y_{2j-1}, y_{2j})) = \frac{1}{2}(1 - \rho_1)\sigma_y^2, \tag{10}$$

and

$$\text{Cov}(\bar{\mathbf{c}}_j) = \frac{1}{4}(\text{Cov}(\mathbf{c}_{2j-1}) + \text{Cov}(\mathbf{c}_{2j}) + 2\text{Cov}(\mathbf{c}_{2j-1}, \mathbf{c}_{2j})) = \frac{1}{2}\boldsymbol{\Sigma}_{\mathbf{c}}. \tag{11}$$

Also, the covariance matrix between $\bar{y}_j$ and $\bar{\mathbf{c}}_j$ is given by:

$$\mathrm{Cov}(\bar{y}_j, \bar{\mathbf{c}}_j) = \frac{1}{4}\mathrm{Cov}(y_{2j-1} + y_{2j}, \mathbf{c}_{2j-1} + \mathbf{c}_{2j}) = \frac{1}{2}\boldsymbol{\sigma}_{y\mathbf{c}}. \tag{12}$$

The joint normality assumption of the response and control variates gives the joint distribution of $\bar{y}_j$ and $\bar{\mathbf{c}}_j$ as follows:

$$\begin{bmatrix} \bar{y}_j \\ \bar{\mathbf{c}}_j \end{bmatrix} \sim \mathrm{N}_{s+1}\left(\begin{bmatrix} \mu_y \\ \mathbf{0}_{(s\times 1)} \end{bmatrix}, \frac{1}{2}\begin{bmatrix} (1-\rho_1)\sigma_y^2 & \boldsymbol{\sigma}_{y\mathbf{c}}' \\ \boldsymbol{\sigma}_{y\mathbf{c}} & \boldsymbol{\Sigma}_{\mathbf{c}} \end{bmatrix} \right). \tag{13}$$

Consequently, $\bar{y}_j$, given $\bar{\mathbf{c}}_j$, is normally distributed with expectation $\mathrm{E}(\bar{y}_j|\bar{\mathbf{c}}_j) = \mu_y + \bar{\mathbf{c}}_j'\boldsymbol{\alpha}$ and variance

$$\mathrm{var}(\bar{y}_j|\bar{\mathbf{c}}_j) = \frac{1}{2}((1-\rho_1)\sigma_y^2 - \boldsymbol{\sigma}_{y\mathbf{c}}'\boldsymbol{\Sigma}_{\mathbf{c}}^{-1}\boldsymbol{\sigma}_{y\mathbf{c}}) = \frac{1}{2}\tau_1^2, \tag{14}$$

where τ_1^2 is defined to be $((1-\rho_1)\sigma_y^2 - \boldsymbol{\sigma}_{y\mathbf{c}}'\boldsymbol{\Sigma}_{\mathbf{c}}^{-1}\boldsymbol{\sigma}_{y\mathbf{c}})$ (see Theorem 2.5.1 of Anderson 1984). As with the case of the linear relationship in (2), the $h \times h$ vector of mean paired responses, $\bar{\mathbf{y}}$, can be represented as:

$$\bar{\mathbf{y}} \equiv \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_h \end{bmatrix} = \mu_y \mathbf{1}_{(h\times 1)} + \bar{\mathbf{C}}\boldsymbol{\alpha} + \boldsymbol{\epsilon}^*, \tag{15}$$

where $\bar{\mathbf{C}}$ is a $h \times s$ matrix of control variates whose jth row is $\bar{\mathbf{c}}_j'$. Regression analysis on this linear model yields the following controlled estimator of the mean response:

$$\hat{\mu}_y = \frac{1}{h}\mathbf{1}'_{(1\times h)}(\bar{\mathbf{y}} - \bar{\mathbf{C}}(\bar{\mathbf{C}}'\mathbf{Q}\bar{\mathbf{C}})^{-1}\bar{\mathbf{C}}'\mathbf{Q}\bar{\mathbf{y}}) = \frac{1}{h}\mathbf{1}'_{(1\times h)}(\mathbf{I}_{(h\times h)} - \bar{\mathbf{C}}(\bar{\mathbf{C}}'\mathbf{Q}\bar{\mathbf{C}})^{-1}\bar{\mathbf{C}}'\mathbf{Q})\bar{\mathbf{y}}, \tag{16}$$

where $\mathbf{Q} = \mathbf{I}_{(h\times h)} - \frac{1}{h}\mathbf{1}_{(h\times 1)}\mathbf{1}'_{(1\times h)}$. From this expression for $\hat{\mu}_y$, we can easily show (see Section 5.2.4) that its unconditional variance is

$$\mathrm{Var}(\hat{\mu}_y) = \frac{\tau_1^2}{2h^2}\left[h + \frac{hs}{(h-s-2)}\right] = \frac{\sigma_y^2}{2h}(1 - \rho_1 - (R_{y\mathbf{c}}^{(1)})^2)\left(\frac{h-2}{h-s-2}\right), \tag{17}$$

where $R_{y\mathbf{c}}^{(1)}$ is the multiple correlation coefficient between y_i and $\mathbf{c}_i$ $(i = 1, 2, \ldots, 2h)$. Thus, provided the assumption that $0 < \rho_1 < 1$ holds, Combined Method I will result in a reduction of the variance of $\mathrm{Var}(\hat{\mu}_y)$, if the effects due to antithetic variates (ρ_1) and the control variates $(R_{y\mathbf{c}}^{(1)})^2$, together, compensate for the loss factor $\frac{h-2}{h-s-2}$.

5.2.2 Combined Method II

This section considers the second method for combining control variates and antithetic variates, which is based on correlated replications utilizing only just the control variates. Contrary to the random number assignment in the previous section, this method uses antithetic variates for the control-variate random components and independent random number streams for all other random components in the model. With this replication strategy, we induce negative correlations between the responses, between the control variates, and between the response and the control variates within h pairs of the responses and the control variates, respectively, obtained from $(\mathbf{R}_{2j-1,1}, \mathbf{R}_{2j-1,2})$ and $(\mathbf{R}_{2j,1}, \bar{\mathbf{R}}_{2j-1,2})$ $(j = 1, 2, \ldots, h)$. However, across pairs of replications, we get independent outputs. Based on the above discussion and the development given for Combined Method I, we note the following properties $(i = 1, 2, \ldots, 2h)$:

Property 1: homogeneity of response variances across replicates,

$$\mathrm{var}(y_i) = \sigma_y^2 \quad i = 1, 2, \ldots, 2h. \tag{18}$$

Property 2: homogeneity of negative correlations across replicate pairs and independence of responses between replicate pairs,

$$\left.\begin{array}{l} \mathrm{cov}(y_k, y_l) = -\rho_2 \sigma_y^2 \text{ for } k = l + 1 \ \ (l = 1, 3, \ldots, 2h - 1); \\ \mathrm{cov}(y_k, y_l) = 0, \ \text{ otherwise.} \end{array}\right\} \tag{19}$$

Property 3: homogeneity of response-control variates covariances across replicates, homogeneity of response-control variates covariances across replicate pairs, and independence of response-control variates covariances between replicate pairs,

$$\left.\begin{array}{l} \mathrm{Cov}(y_i, \mathbf{c}_i) = \boldsymbol{\sigma}_{y\mathbf{c}} \;\; i = 1, 2, \ldots, 2h; \\ \mathrm{Cov}(y_k, \mathbf{c}_l) = \boldsymbol{\sigma}_{y\mathbf{c}}^{(2)} \;\text{ for }\; k = l+1 \;\; (l = 1, 3, \ldots, 2h-1); \\ \mathrm{Cov}(y_i, \mathbf{c}_j) = \mathbf{0}_{(s\times 1)}, \;\; \text{otherwise.} \end{array}\right\} \tag{20}$$

Property 4: homogeneity of control variates covariances across replicates, homogeneity of control variate covariances across replicate pairs, and independence of control variates between replicate pairs,

$$\left.\begin{array}{l} \mathrm{Cov}(\mathbf{c}_i) = \boldsymbol{\Sigma}_{\mathbf{c}} \;\; i = 1, 2, \ldots, 2h; \\ \mathrm{Cov}(\mathbf{c}_k, \mathbf{c}_l) = \boldsymbol{\Sigma}_{\mathbf{c}}^{(2)} \;\text{ for }\; k = l+1 \;\; (l = 1, 3, \ldots, 2h-1); \\ \mathrm{Cov}(\mathbf{c}_k, \mathbf{c}_j) = \mathbf{0}_{(s\times s)}, \;\; \text{(otherwise).} \end{array}\right\} \tag{21}$$

We make the following assumption

$$\textbf{Assumption}: \;\; 0 < \rho_2 < 1.$$

Under the four properties given above, the variance of $\bar{y}_j$, the covariance of $\bar{\mathbf{c}}_j$ and the covariances between $\bar{y}_j$ and $\bar{\mathbf{c}}_j$ are, respectively $(j = 1, 2, \ldots, h)$:

$$\mathrm{var}(\bar{y}_j) = \frac{1}{2}(1 - \rho_2)\sigma_y^2, \tag{22}$$

$$\mathrm{Cov}(\bar{\mathbf{c}}_j) = \frac{1}{4}(\mathrm{Cov}(\mathbf{c}_{2j-1}) + \mathrm{Cov}(\mathbf{c}_{2j}) + 2\mathrm{Cov}(\mathbf{c}_{2j-1}, \mathbf{c}_{2j})) = \frac{1}{2}(\boldsymbol{\Sigma}_{\mathbf{c}} + \boldsymbol{\Sigma}_{\mathbf{c}}^{(2)}), \tag{23}$$

and

$$\begin{aligned} \mathrm{Cov}(\bar{y}_j, \bar{\mathbf{c}}_j) &= \frac{1}{4}\mathrm{Cov}(y_{2j-1} + y_{2j}, \mathbf{c}_{2j-1} + \mathbf{c}_{2j}) \\ &= \frac{1}{4}(\mathrm{Cov}(y_{2j-1}, \mathbf{c}_{2j-1}) + \mathrm{Cov}(y_{2j-1}, \mathbf{c}_{2j}) + \mathrm{Cov}(y_{2j}, \mathbf{c}_{2j-1}) + \mathrm{Cov}(y_{2j}, \mathbf{c}_{2j})) \\ &= \frac{1}{2}(\boldsymbol{\sigma}_{y\mathbf{c}} + \boldsymbol{\sigma}_{y\mathbf{c}}^{(2)}). \end{aligned} \tag{24}$$

Based on the joint normality assumption of the response and control variates and the properties given in (18), (19), (20) and (21), we have the joint distribution of $\bar{y}_j$ and $\bar{\mathbf{c}}_j$ is given as:

$$\begin{bmatrix} \bar{y}_j \\ \bar{\mathbf{c}}_j \end{bmatrix} \sim \mathrm{N}_{s+1}\left(\begin{bmatrix} \mu_y \\ \mathbf{0}_{(s\times 1)} \end{bmatrix}, \frac{1}{2} \begin{bmatrix} (1-\rho_2)\sigma_y^2 & \boldsymbol{\sigma}_{y\mathbf{c}}{}' + \boldsymbol{\sigma}'^{(2)}_{y\mathbf{c}} \\ \boldsymbol{\sigma}_{y\mathbf{c}} + \boldsymbol{\sigma}^{(2)}_{y\mathbf{c}} & \boldsymbol{\Sigma}_{\mathbf{c}} + \boldsymbol{\Sigma}^{(2)}_{\mathbf{c}} \end{bmatrix} \right). \tag{25}$$

Since $\bar{\mathbf{c}}_j$ $(j = 1, 2, \ldots, h)$ are independent, following the development of (17) we get by an analogous approach:

$$\mathrm{Var}(\hat{\mu}_y) = \frac{\tau_2^2}{2h}\left(\frac{h-2}{(h-s-2)}\right) = \frac{(1-\rho_2)\sigma_y^2}{2h}(1-(R^{(2)}_{y\mathbf{c}})^2)\left(\frac{h-2}{h-s-2}\right), \tag{26}$$

where τ_2^2 is defined to be

$$\tau_2^2 \equiv (1-\rho_2)\sigma_y^2 - (\boldsymbol{\sigma}_{y\mathbf{c}} + \boldsymbol{\sigma}^{(2)}_{y\mathbf{c}})'(\boldsymbol{\Sigma}_{\mathbf{c}} + \boldsymbol{\Sigma}^{(2)}_{\mathbf{c}})^{-1}(\boldsymbol{\sigma}_{y\mathbf{c}} + \boldsymbol{\sigma}^{(2)}_{y\mathbf{c}}) \tag{27}$$

and

$$(R^{(2)}_{y\mathbf{c}})^2 = ((1-\rho_2)\sigma_y^2)^{-1}(\boldsymbol{\sigma}_{y\mathbf{c}} + \boldsymbol{\sigma}^{(2)}_{y\mathbf{c}})'(\boldsymbol{\Sigma}_{\mathbf{c}} + \boldsymbol{\Sigma}^{(2)}_{\mathbf{c}})^{-1}(\boldsymbol{\sigma}_{y\mathbf{c}} + \boldsymbol{\sigma}^{(2)}_{y\mathbf{c}}), \tag{28}$$

which is the square of the multiple correlation coefficient between $\bar{y}_j$ and $\bar{\mathbf{c}}_j$. Thus, provided the assumption $0 < \rho_2 < 1$ holds, Combined Method II will result in a reduction of the variance of $\mathrm{Var}(\hat{\mu}_y)$, if the effects due to antithetic variates (ρ_2) and the control variates $(R^{(2)}_{y\mathbf{c}})^2$, together, compensate for the loss factor $\frac{h-2}{h-s-2}$.

5.2.3 Combined Method III

This section presents Combined Method III which jointly implements control variates and antithetic variates based on correlated replicates induced by utilizing *all* random components in the simulation model. Unlike the random number assignment strategies discussed in the previous two sections, we apply antithetic variates to *all* random components in the model. This assignment strategy induces correlations across h pairs of

both the response and the control variates. That is, negative correlations are induced between the reponses, between the control variates, *and* between the response and the control variates within h pairs of the replicates. However, the induced correlation between the response and the control variates is different from that of Combined Methods I and II. Across the h pairs of replicates, the mean reponse and the mean of the control variates (within a pair of replicates) are independently observed by the assignment of different sets of randomly chosen random number streams. As before, we identify a set of properties for this strategy with regard to the covariance structure of the responses and control variates in the experiment. Based on the above discussion and the developments for the other two combined methods, for Combined Method III we establish that $(i = 1, 2, \ldots, 2h)$:

Property 1: homogeneity of response variances across replicates,

$$\text{var}(y_i) = \sigma_y^2 \quad i = 1, 2, \ldots, 2h. \tag{29}$$

Property 2: homogeneity of negative correlations across replicate pairs and independence of responses between replicate pairs,

$$\left.\begin{array}{l} \text{cov}(y_k, y_l) = -\rho_3 \sigma_y^2 \text{ for } k = l + 1 \ \ (l = 1, 3, \ldots, 2h - 1); \\ \text{cov}(y_j, y_k) = 0, \ \ \text{(otherwise)}. \end{array}\right\} \tag{30}$$

Property 3: homogeneity of response-control variates covariances across replicates, homogeneity of response-control variates covariances across replicate pairs, and independence of response-control variates covariances between replicate pairs,

$$\left.\begin{array}{l} \text{Cov}(y_i, \mathbf{c}_i) = \boldsymbol{\sigma}_{y\mathbf{c}} \quad i = 1, 2, \ldots, 2h; \\ \text{Cov}(y_k, \mathbf{c}_l) = \boldsymbol{\sigma}_{y\mathbf{c}}^{(3)} \text{ for } k = l + 1 \ \ (l = 1, 3, \ldots, 2h - 1); \\ \text{Cov}(y_i, \mathbf{c}_j) = \mathbf{0}_{(s \times 1)}, \ \ \text{otherwise}. \end{array}\right\} \tag{31}$$

Property 4: homogeneity of control variates covariances across replicates, homogeneity of control variate covariances across replicate pairs, and independence of control variates between replicate pairs,

$$\left.\begin{array}{l} \mathrm{Cov}(\mathbf{c}_i) = \boldsymbol{\Sigma}_{\mathbf{c}} \;\; i = 1, 2, \ldots, 2; \\ \mathrm{Cov}(\mathbf{c}_k, \mathbf{c}_l) = \boldsymbol{\Sigma}_{\mathbf{c}}^{(3)} \;\; \text{for} \;\; k = l+1 \;\; (l = 1, 3, \ldots, 2h-1); \\ \mathrm{Cov}(\mathbf{c}_i, \mathbf{c}_k) = \mathbf{0}_{(s \times s)}, \;\; \text{otherwise.} \end{array}\right\} \tag{32}$$

We make the following assumption

$$\textbf{Assumption}: \;\; 0 < \rho_3 < 1.$$

Under the four properties given above we obtain analogous results to those given in Section 5.2.2 ($j = 1, 2, \ldots, h$):

$$\mathrm{var}(\bar{y}_j) = \frac{1}{2}(1 - \rho_3)\sigma_y^2 \tag{33}$$

$$\mathrm{Cov}(\bar{\mathbf{c}}_j) = \frac{1}{4}(\mathrm{Cov}(\mathbf{c}_{2j-1}) + \mathrm{Cov}(\mathbf{c}_{2j}) + 2\mathrm{Cov}(\mathbf{c}_{2j-1}, \mathbf{c}_{2j})) = \frac{1}{2}(\boldsymbol{\Sigma}_{\mathbf{c}} + \boldsymbol{\Sigma}_{\mathbf{c}}^{(3)}), \tag{34}$$

and

$$\mathrm{Cov}(\bar{y}_j, \bar{\mathbf{c}}_j) = \frac{1}{4}\mathrm{Cov}(y_{2j-1} + y_{2j}, \mathbf{c}_{2j-1} + \mathbf{c}_{2j}) = \frac{1}{2}(\boldsymbol{\sigma}_{y\mathbf{c}} + \boldsymbol{\sigma}_{y\mathbf{c}}^{(3)}). \tag{35}$$

Under the joint normality assumption of the response and control variates and the properties given in (29), (30), (31) and (32), the joint distribution of $\bar{y}_j$ and $\bar{\mathbf{c}}_j$ is given as:

$$\begin{bmatrix} \bar{y}_j \\ \bar{\mathbf{c}}_j \end{bmatrix} \sim \mathrm{N}_{s+1}\left(\begin{bmatrix} \mu_y \\ \mathbf{0}_{(s \times 1)} \end{bmatrix}, \frac{1}{2}\begin{bmatrix} (1-\rho_3)\sigma_y^2 & \boldsymbol{\sigma}_{y\mathbf{c}}' + \boldsymbol{\sigma}_{y\mathbf{c}}'^{(3)} \\ \boldsymbol{\sigma}_{y\mathbf{c}} + \boldsymbol{\sigma}_{y\mathbf{c}}^{(3)} & \boldsymbol{\Sigma}_{\mathbf{c}} + \boldsymbol{\Sigma}_{\mathbf{c}}^{(3)} \end{bmatrix}\right). \tag{36}$$

Since $\bar{\mathbf{c}}_j$ ($j = 1, 2, \ldots, h$) are independent, following the development of (26) we get by an analogous approach:

$$\mathrm{Var}(\hat{\mu}_y) = \frac{\tau_3^2}{2h}\left(\frac{h-2}{h-s-2}\right) = \frac{(1-\rho_3)\sigma_y^2}{2h}(1 - (R_{y\mathbf{c}}^{(3)})^2)\left(\frac{h-2}{h-s-2}\right), \tag{37}$$

where τ_3^2 is defined to be

$$\tau_3^2 \equiv (1-\rho_3)\sigma_y^2 - (\boldsymbol{\sigma}_{y\mathbf{c}} + \boldsymbol{\sigma}_{y\mathbf{c}}^{(3)})'(\boldsymbol{\Sigma}_{\mathbf{c}} + \boldsymbol{\Sigma}_{\mathbf{c}}^{(3)})^{-1}(\boldsymbol{\sigma}_{y\mathbf{c}} + \boldsymbol{\sigma}_{y\mathbf{c}}^{(3)}) \tag{38}$$

and

$$(R_{y\mathbf{c}}^{(3)})^2 = ((1-\rho_3)\sigma_y^2)^{-1}(\boldsymbol{\sigma}_{y\mathbf{c}} + \boldsymbol{\sigma}_{y\mathbf{c}}^{(3)})'(\boldsymbol{\Sigma}_{\mathbf{c}} + \boldsymbol{\Sigma}_{\mathbf{c}}^{(3)})^{-1}(\boldsymbol{\sigma}_{y\mathbf{c}} + \boldsymbol{\sigma}_{y\mathbf{c}}^{(3)}), \tag{39}$$

which is the square of the multiple correlation coefficient between $\bar{y}_j$ and $\bar{\mathbf{c}}_j$. Provided the assumption $0 < \rho_3 < 1$ holds, Combined Method III will result in a reduction of the variance of $\mathrm{Var}(\hat{\mu}_y)$, if the effects due to antithetic variates (ρ_3) and the control variates $(R_{y\mathbf{c}}^{(3)})^2$, together, compensate for the loss factor $\frac{h-2}{h-s-2}$.

5.2.4 Comparison of the Combined Methods and Control Variates Method

In this section we give a brief and formal comparison of the three combined methods presented in the three previous sections and the method of control variates. This comparison assumes the validity of the assumptions given for each of these four methods and is done with respect to the unconditional variances of the estimators for the mean response given in equations (5), (17), (26), and (37), respectively. (We can say something in general about the assumptions. If the simulation response y and the simulation model are structured so that y is *monotonic* in each random-number input (either nonincreasing or nondecreasing), then it is quaranteed that ρ_1, ρ_2, and ρ_3 are all nonnegative (see Brately, Fox, and Schrage 1987).

First, we consider the three combined methods presented earlier. Comparing Combined Methods I and II via equations (17) and (26) yields that Combined Method I is preferred to Combined Method II if

$$(1-\rho_1-(R_{y\mathbf{c}}^{(1)})^2) < (1-\rho_2)(1-(R_{y\mathbf{c}}^{(2)})^2). \tag{40}$$

Similarly, working with (17) and (37) yields that Combined Method III is better than Combined Method I, provided that

$$(1-\rho_1-(R_{y\mathbf{c}}^{(1)})^2) > (1-\rho_3)(1-(R_{y\mathbf{c}}^{(3)})^2). \tag{41}$$

Also, working with (26) and (37) yields that Combined Method III is better than Combined Method II if

$$(1-\rho_3)(1-(R_{y\mathbf{c}}^{(3)})^2) < (1-\rho_2)(1-(R_{y\mathbf{c}}^{(2)})^2). \tag{42}$$

As we discussed earlier, the loss factors for the three combined methods are the same; hence, they cancel when constructing the comparisons. Thus, the preference of the three methods is determined according to their minimum variance ratios given in (17), (26), and (37), respectively. Of course, other ordering schemes for the terms in equations (17), (26), and (37) could be conjectured. Clearly, it is not easy to identify an ordered relationship among $(R_{y\mathbf{c}}^{(1)})^2$, $(R_{y\mathbf{c}}^{(2)})^2$, and $(R_{y\mathbf{c}}^{(3)})^2$ since these terms involve the unknown elements $\boldsymbol{\Sigma}_{\mathbf{c}}$, $\boldsymbol{\Sigma}_{\mathbf{c}}^{(2)}$, $\boldsymbol{\Sigma}_{\mathbf{c}}^{(3)}$, $\boldsymbol{\sigma}_{y\mathbf{c}}$, $\boldsymbol{\sigma}_{y\mathbf{c}}^{(2)}$, $\boldsymbol{\sigma}_{y\mathbf{c}}^{(3)}$, ρ_2, and ρ_3.

Next, we compare these three combined methods to the method of control variates. A comparison of equations (5) and (17) yields that Combined Method I is better than the control variates method if

$$(1-\rho_1-(R_{y\mathbf{c}}^{(2)})^2)\left(\frac{h-2}{h-s-2}\right) < (1-R_{y\mathbf{c}}^2)\left(\frac{2h-2}{2h-s-2}\right). \tag{43}$$

Also, comparing equations (5), (26), and (37) shows that Combined Method II is better than the method of control variates if

$$(1-\rho_2)(1-(R_{y\mathbf{c}}^{(2)})^2)\left(\frac{h-2}{h-s-2}\right) < (1-R_{y\mathbf{c}}^2)\left(\frac{2h-2}{2h-s-2}\right); \tag{44}$$

and Combined Method III yields a better result than the method of control variates, provided

$$(1-\rho_3)(1-(R_{y\mathbf{c}}^{(3)})^2)\left(\frac{h-2}{h-s-2}\right) < (1-R_{y\mathbf{c}}^2)\left(\frac{2h-2}{2h-s-2}\right). \tag{45}$$

Note that the loss factor of each combined method is greater than that of the method of control variates. Hence, for preference of each combined method to the method of control variates, the associated minimum variance ratio of the combined method should, at least, compensate for an increase in the associated loss factor. The effects of antithetic variates and control variates on the minimum variance ratio for Combined Method I are represented by an additive form in reducing the variance of the estimator of the mean response.

5.2.5 Proof of Equation (17)

Given $\bar{\mathbf{C}}$, the conditional variance of $\hat{\mu}_y$ in (16) is:

$$\text{var}(\hat{\mu}_y|\bar{\mathbf{C}}) = \frac{1}{h^2}\mathbf{1}'_{(1\times h)}(\mathbf{I}_{(h\times h)}-\bar{\mathbf{C}}(\bar{\mathbf{C}}'\mathbf{Q}\bar{\mathbf{C}})^{-1}\bar{\mathbf{C}}'\mathbf{Q})\text{Cov}(\bar{\mathbf{y}}|\bar{\mathbf{C}})(\mathbf{I}_{(h\times h)}-\mathbf{Q}\bar{\mathbf{C}}(\bar{\mathbf{C}}'\mathbf{Q}\bar{\mathbf{C}})^{-1}\bar{\mathbf{C}}')\mathbf{1}_{(h\times 1)}. \tag{46}$$

Since $(\bar{y}_j, \bar{\mathbf{c}}_j)$ is obtained from the jth pair of replicates and the *pairs* of replicates are independent, we have from equation (14) that:

$$\text{Var}(\bar{\mathbf{y}}|\bar{\mathbf{C}}) = \frac{1}{2}[(1-\rho_1)\sigma_y^2 - \boldsymbol{\sigma}_{y\mathbf{c}}'\boldsymbol{\Sigma}_{\mathbf{c}}^{-1}\boldsymbol{\sigma}_{y\mathbf{c}}]\mathbf{I}_{(h\times h)} = \frac{1}{2}\tau_1^2\mathbf{I}_{(h\times h)}. \tag{47}$$

Subsituting for $\text{Var}(\bar{\mathbf{y}}|\bar{\mathbf{C}})$ into (47) gives:

$$\text{var}(\hat{\mu}_y|\bar{\mathbf{C}}) = \frac{\tau_1^2}{2h^2}(h + \mathbf{1}'_{(1\times h)}\bar{\mathbf{C}}(\bar{\mathbf{C}}'\mathbf{Q}\bar{\mathbf{C}})^{-1}\bar{\mathbf{C}}'\mathbf{1}_{(h\times 1)}), \tag{48}$$

since $\mathbf{Q}\mathbf{1}_{(h\times 1)} = \mathbf{1}'_{(1\times h)}\mathbf{Q} = \mathbf{0}_{(h\times 1)}$ and $\mathbf{Q}^2 = \mathbf{Q}$. From the result given in (13) and the last assumption in (9), the $(h \times s)$ random matrix $\bar{\mathbf{C}}$ has the matrix normal distribution:

$$\bar{\mathbf{C}} \sim \text{N}_{h,s}(\mathbf{0}_{(h\times s)}, \mathbf{I}_{(h\times h)}, \frac{1}{2}\boldsymbol{\Sigma}_{\mathbf{c}}),$$

where $\mathbf{0}_{(h\times s)}$ is a $(h \times s)$ matrix of zeroes (see Section 17.2 in Arnold 1981). From this result we have (see p. 454 and Theorem 17.2(d) of Arnold 1981):

$$\mathbf{Z} \equiv (\frac{1}{2}\boldsymbol{\Sigma}_{\mathbf{c}})^{-\frac{1}{2}}\bar{\mathbf{C}}' \sim \text{N}_{s,h}(\mathbf{0}_{(s\times h)}, \mathbf{I}_{(s\times s)}, \mathbf{I}_{(h\times h)}). \tag{49}$$

Now, it follows immediately that

$$\mathbf{Z}'\mathbf{Z} \equiv \bar{\mathbf{C}}(\frac{1}{2}\boldsymbol{\Sigma}_{\mathbf{c}})^{-1}\bar{\mathbf{C}}' \sim \mathrm{W}_h(s, \mathbf{I}_{(h\times h)}), \tag{50}$$

using the definition of the Wishart distribution (see equation (17.15) of Arnold 1981) and, by Theorem 17.7a in Arnold (1981), the $s \times s$ random matrix $(\bar{\mathbf{C}}'\mathbf{Q}\bar{\mathbf{C}})$ has the following Wishart distribution:

$$(\bar{\mathbf{C}}'\mathbf{Q}\bar{\mathbf{C}}) \sim \mathrm{W}_s(h-1, \frac{1}{2}\boldsymbol{\Sigma}_{\mathbf{c}}), \tag{51}$$

since $\mathbf{Q}$ is an idempotent matrix with rank $(h-1)$. We note that $(\mathbf{1}'_{(1\times h)}\bar{\mathbf{C}})$ and $(\bar{\mathbf{C}}'\mathbf{Q}\bar{\mathbf{C}})$ are independent (see Theorem 17.7b(2) of Arnold 1981).

Now, by equations (15) and (16), we have that

$$\begin{aligned} \mathrm{E}(\hat{\mu}_y|\bar{\mathbf{C}}) &= \frac{1}{h}\mathbf{1}'_{(1\times h)}[\mathbf{I}_{(h\times h)} - \bar{\mathbf{C}}(\bar{\mathbf{C}}'\mathbf{Q}\bar{\mathbf{C}})^{-1}\bar{\mathbf{C}}'\mathbf{Q}]\mathrm{E}[\bar{y}|\bar{\mathbf{C}}] \\ &= \frac{1}{h}\mathbf{1}'_{(1\times h)}[\mathbf{I}_{(h\times h)} - \bar{\mathbf{C}}(\bar{\mathbf{C}}'\mathbf{Q}\bar{\mathbf{C}})^{-1}\bar{\mathbf{C}}'\mathbf{Q}](\mu_y\mathbf{1}_{(h\times 1)} + \bar{\mathbf{C}}\boldsymbol{\alpha}), \end{aligned} \tag{52}$$

which, in turn, by multiplying through all terms and using the result $\mathbf{Q}\mathbf{1}_{(h\times 1)} = \mathbf{0}_{(h\times 1)}$ given earlier, yields

$$\mathrm{E}(\hat{\mu}_y|\bar{\mathbf{C}}) = \mu_y. \tag{53}$$

From this, we get

$$\mathrm{var}[\mathrm{E}(\hat{\mu}_y|\bar{\mathbf{C}})] = \mathrm{var}(\mu_y) = 0. \tag{54}$$

Thus, the unconditional variance of $\hat{\mu}_y$ can be written as:

$$\mathrm{var}(\hat{\mu}_y) = \mathrm{var}[\mathrm{E}(\hat{\mu}_y|\bar{\mathbf{C}})] + \mathrm{E}[\mathrm{var}(\hat{\mu}_y|\bar{\mathbf{C}})] = 0 + \frac{\tau_1^2}{2h^2}\mathrm{E}[h + \mathbf{1}'_{(1\times h)}\bar{\mathbf{C}}\mathrm{E}[(\bar{\mathbf{C}}'\mathbf{Q}\bar{\mathbf{C}})^{-1}]\bar{\mathbf{C}}'\mathbf{1}_{(h\times 1)}], \tag{55}$$

since $\mathbf{1}'\bar{\mathbf{C}}$ and $\bar{\mathbf{C}}'\mathbf{Q}\bar{\mathbf{C}}$ are independent. Theorems 17.6a and 17.15d in Arnold (1981) give, respectively,

$$\mathrm{E}[\bar{\mathbf{C}}(\frac{1}{2}\boldsymbol{\Sigma}_{\mathbf{c}})^{-1}\bar{\mathbf{C}}'] = s\mathbf{I}_{(h\times h)}, \text{ and } \mathrm{E}[(\bar{\mathbf{C}}'\mathbf{Q}\bar{\mathbf{C}})^{-1}] = \left(\frac{2}{h-s-2}\right)\boldsymbol{\Sigma}_{\mathbf{c}}^{-1} \text{ if } h > s+2. \tag{56}$$

Therefore, plugging the second equation in (57) into (56) finally yields

$$\mathrm{Var}(\hat{\mu}_y) = \frac{\tau_1^2}{2h^2}\left[h + \frac{2\mathrm{E}[\mathbf{1}'_{(1\times h)}\bar{\mathbf{C}}\boldsymbol{\Sigma}_{\mathbf{c}}^{-1}\bar{\mathbf{C}}'\mathbf{1}_{(h\times 1)}]}{(h-s-2)}\right] \tag{57}$$

which further reduces to equation (17), by equation (51) and Theorem 17.6(a) of Arnold (1981). The result in (17) indicates that the minimum variance ratio of this method is $(1-\rho_1-(R_{y\mathbf{c}}^{(1)})^2)$ and the loss factor is $\frac{h-2}{h-s-2}$.

6. Conclusions

In this study we have described a valid computer simulation model of the XP microcomputer network that is to become operational. We have also indicated how this simulation model can be used as an experimental device to create data that can be used to fit regression models between responses and input factors of interest and how these regression models can be used to help design the network. We have also developed a set of variance reduction methodologies that greatly improves the estimation of the regression model coefficient estimators.

In the past, both control variates and antithetic variates have been shown to often be effective variance reduction techniques for estimating the mean of a response of interest for simulation experiments. In Section 5 we have presented three methods that can easily be implemented in a simulation experiment without significant additional programming effort; each combines both control variates and antithetic variates. Under general assumptions, these combined methods were shown to yield superior performance, based on estimator variance, than control variates alone.

APPENDIX I

SLAM II CODE FOR SIMULATION MODEL

```
      PROGRAM MAIN
      COMMON/SCOM1/ATRIB(100),DD(100),DDL(100),DTNOW,II,MFA,
     *MSTOP,NCLNR,NCRDR,NPRNT,NNRUN,NNSET,NTAPE,SS(100),
     *SSL(100),TNEXT,TNOW,XX(100)
      COMMON/UCOM1/FX(27),XVAL(27)
      COMMON QSET(200000)
      DIMENSION NSET(200000)
      EQUIVALENCE (NSET(1),QSET(1))
      NNSET=200000
      NCRDR=5
      NPRNT=6
      NTAPE=7
      CALL SLAM
      STOP
      END

      SUBROUTINE INTLC
      COMMON/SCOM1/ATRIB(100),DD(100),DDL(100),DTNOW,II,MFA,
     *MSTOP,NCLNR,NCRDR,NPRNT,NNRUN,NNSET,NTAPE,SS(100),
     *SSL(100),TNEXT,TNOW,XX(100)
      COMMON/UCOM1/FX(27),XVAL(27)
C*************************************************************
C*    INITITALIZE THE CUMULATIVE DISTRIBUTION FUNCTION      *
C*    FX AND ITS ASSOCIATED VALUES XVAL THAT WILL DETERMINE*
C*    THE ROUTING SCHEME OF INCOMING MESSAGES               *
C*************************************************************
      FX(1)= .0370
      FX(2)= .0741
      FX(3)= .1111
      FX(4)= .1481
      FX(5)= .1852
      FX(6)= .2222
      FX(7)= .2593
      FX(8)= .2963
      FX(9)= .3333
      FX(10)= .3704
      FX(11)= .4074
      FX(12)= .4444
      FX(13)= .4815
      FX(14)= .5185
      FX(15)= .5556
      FX(16)= .5926
      FX(17)= .6296
      FX(18)= .6667
      FX(19)= .7037
      FX(20)= .7407
      FX(21)= .7778
      FX(22)= .8148
      FX(23)= .8519
      FX(24)= .8889
      FX(25)= .9259
      FX(26)= .9630
      FX(27)= 1.000
      XVAL(1)=1
      XVAL(2)=2
      XVAL(3)=3
      XVAL(4)=4
      XVAL(5)=5
      XVAL(6)=6
      XVAL(7)=7
```

```
      XVAL(8)=8
      XVAL(9)=9
      XVAL(10)=10
      XVAL(11)=11
      XVAL(12)=12
      XVAL(13)=13
      XVAL(14)=14
      XVAL(15)=15
      XVAL(16)=16
      XVAL(17)=17
      XVAL(18)=18
      XVAL(19)=19
      XVAL(20)=20
      XVAL(21)=21
      XVAL(22)=22
      XVAL(23)=23
      XVAL(24)=24
      XVAL(25)=25
      XVAL(26)=26
      XVAL(27)=27
      RETURN
      END

      FUNCTION USERF(I)
      COMMON/SCOM1/ATRIB(100),DD(100),DDL(100),DTNOW,II,MFA,
     *MSTOP,NCLNR,NCRDR,NPRNT,NNRUN,NNSET,NTAPE,SS(100),
     *SSL(100),TNEXT,TNOW,XX(100)
      COMMON/UCOM1/FX(27),XVAL(27)
      USERF=DPROB(FX,XVAL,27,1)
      RETURN
      END
//GO.SYSIN DD *
GEN,TEW,ROME NETWORK,10/6/92,1;
LIMITS,29,7,15000;
NETWORK;
;*********************************************************
;     THIS IS THE SLAM II NETWORK CODE FOR THE ROME     *
;     LABORATORY MACINTOSH ETHERNET NETWORK             *
;                                                       *
;     CREATE THE MESSAGES THAT WILL FLOW THROUGH THE    *
;     NETWORK AND INDICATE IN ATTRIBUTE (2) WHERE THAT  *
;     MESSAGE CAME FROM                                 *
;                                                       *
;     ALSO, LET ATRIB(4)=1 IF MESSAGE IS ROUTED TO      *
;     ANOTHER NODE AND LET ATRIB(4)=2 IF MESSAGE IS     *
;     IS TO GO TO THE STAR CONTROLLER                   *
;                                                       *
;     ATRIB(1)  IS THE ARRIVAL TIME OF MESSAGE          *
;     ATRIB(2)  IS THE ORIGINATION NODE OF MESSAGE      *
;     ATRIB(3)  IS THE DESTINATION NODE OF MESSAGE      *
;     ATRIB(4)  INDICATES IF MESSAGE RETURNS OR NOT     *
;     ATRIB(5)  IS THE LINE RESOURCE NUMBER             *
;     ATRIB(6)  IS THE NUMBER OF FRAMES IN A MESSAGE    *
;     ATRIB(7)  IS THE NUMBER OF BYTES IN THE MESSAGE   *
;*********************************************************

;*********************************************************
;     DEFINE RESOURCE ELEMENTS TO EACH LINE IN THE      *
;     NETWORK AS WELL AS THE STAR CONTROLLER            *
;*********************************************************
```

```
      RESOURCE/LINE1(1),29,15,17,18,19;
      RESOURCE/LINE2(1),29,13;
      RESOURCE/LINE3(1),29,14,16;
      RESOURCE/LINE4(1),29,11,12;
      RESOURCE/LINE5(1),29,8,9,10;
      RESOURCE/LINE6(1),29,7;
      RESOURCE/LINE7(1),29,4,5,6;
      RESOURCE/LINE8(1),29,1,2,3;
      RESOURCE/LINE9(1),29,20;
      RESOURCE/LINE10(1),29,22;
      RESOURCE/LINE11(1),29,23;
      RESOURCE/LINE12(1),29,25,26;
      RESOURCE/LINE13(1),29,21;
      RESOURCE/LINE14(1),29,24;
      RESOURCE/LINE15(1),29,27;
      RESOURCE/STAR(4),28;

;*******************************************************
; CORRIDOR O (CREATE MESSAGES AT EACH OF THE NODES     *
; IN CORRIDOR O)                                       *
;*******************************************************

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=1,ATRIB(5)=8,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(1),LINE8/1;
      ACT,.00003,,SC;

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=2,ATRIB(5)=8,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(2),LINE8/1;
      ACT,.00003,,SC;

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=3,ATRIB(5)=8,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(3),LINE8/1;
      ACT,.00003,,SC;

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=4,ATRIB(5)=7,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(4),LINE7/1;
      ACT,.00003,,SC;

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=5,ATRIB(5)=7,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(5),LINE7/1;
      ACT,.00003,,SC;

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=6,ATRIB(5)=7,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(6),LINE7/1;
      ACT,.00003,,SC;

      CREATE,EXP(5),0.0,1,,1;
```

```
ASSIGN,ATRIB(2)=7,ATRIB(5)=6,
       ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
AWAIT(7),LINE6/1;
ACT,.00003,,SC;

CREATE,EXP(5),0.0,1,,1;
ASSIGN,ATRIB(2)=8,ATRIB(5)=5,
       ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
AWAIT(8),LINE5/1;
ACT,.00003,,SC;

CREATE,EXP(5),0.0,1,,1;
ASSIGN,ATRIB(2)=9,ATRIB(5)=5,
       ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
AWAIT(9),LINE5/1;
ACT,.00003,,SC;

CREATE,EXP(5),0.0,1,,1;
ASSIGN,ATRIB(2)=10,ATRIB(5)=5,
       ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
AWAIT(10),LINE5/1;
ACT,.00003,,SC;

CREATE,EXP(5),0.0,1,,1;
ASSIGN,ATRIB(2)=11,ATRIB(5)=4,
       ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
AWAIT(11),LINE4/1;
ACT,.00003,,SC;

CREATE,EXP(5),0.0,1,,1;
ASSIGN,ATRIB(2)=12,ATRIB(5)=4,
       ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
AWAIT(12),LINE4/1;
ACT,.00003,,SC;

CREATE,EXP(5),0.0,1,,1;
ASSIGN,ATRIB(2)=13,ATRIB(5)=2,
       ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
AWAIT(13),LINE2/1;
ACT,.00003,,SC;

CREATE,EXP(5),0.0,1,,1;
ASSIGN,ATRIB(2)=14,ATRIB(5)=3,
       ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
AWAIT(14),LINE3/1;
ACT,.00003,,SC;

CREATE,EXP(5),0.0,1,,1;
ASSIGN,ATRIB(2)=15,ATRIB(5)=1,
       ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
AWAIT(15),LINE1/1;
ACT,.00003,,SC;

CREATE,EXP(5),0.0,1,,1;
ASSIGN,ATRIB(2)=16,ATRIB(5)=3,
       ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
AWAIT(16),LINE3/1;
ACT,.00003,,SC;

CREATE,EXP(5),0.0,1,,1;
```

```
      ASSIGN,ATRIB(2)=17,ATRIB(5)=1,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(17),LINE1/1;
      ACT,.00003,,SC;

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=18,ATRIB(5)=1,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(18),LINE1/1;
      ACT,.00003,,SC;

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=19,ATRIB(5)=1,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(19),LINE1/1;
      ACT,.00003,,SC;

;*********************************************************
; CORRIDOR A (CREATE MESSAGE AT EACH OF THE NODES       *
; IN CORRIDOR A)                                        *
;*********************************************************

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=20,ATRIB(5)=9,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(20),LINE9/1;
      ACT,.00003,,SC;

;*********************************************************
; CORRIDOR B (CREATE MESSAGES AT EACH OF THE NODES      *
; IN CORRIDOR B)                                        *
;*********************************************************

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=21,ATRIB(5)=13,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(21),LINE13/1;
      ACT,.00003,,SC;

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=22,ATRIB(5)=10,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(22),LINE10/1;
      ACT,.00003,,SC;

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=23,ATRIB(5)=11,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(23),LINE11/1;
      ACT,.00003,,SC;

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=24,ATRIB(5)=14,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(24)/LINE14/1;
      ACT,.00003,,SC;

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=25,ATRIB(5)=12,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
```

```
      AWAIT(25),LINE12/1;
      ACT,.00003,,SC;

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=26,ATRIB(5)=12,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(26),LINE12/1;
      ACT,.00003,,SC;

      CREATE,EXP(5),0.0,1,,1;
      ASSIGN,ATRIB(2)=27,ATRIB(5)=15,
             ATRIB(6)=NPSSN(11,2),ATRIB(7)=1526.,1;
      AWAIT(27),LINE15/1;
      ACT,.00003,,SC;

;*********************************************************
; STAR CONTROLLER                                        *
;                                                        *
; DETERMINE WHERE THE MESSAGE IS GOING TO THEN PUT IT    *
; IN THE STAR CONTROLLER QUEUE                           *
; ATRIB(3) GIVES THE DESTINATION OF THE MESSAGE          *
;                                                        *
; 25 PERCENT OF MESSAGES GO BACK TO POINT OF             *
; ORIGINATION AND 75 PERCENT GO TO ANOTHER NODE          *
;                                                        *
;*********************************************************

   SC ASSIGN,ATRIB(3)=USERF(1),
             ATRIB(4)=1,1;
      ACT,0.0,ATRIB(3)-ATRIB(2).LE.0.5.AND.
              ATRIB(2)-ATRIB(3).LE.0.5,SC;
      ACT,0.0;

      FREE,ATRIB(5)/1,1;
      ACT,0.0,.25,ASS;
      ACT,0.0,.75,GO1;
  ASS ASSIGN,ATRIB(3)=ATRIB(2),
             ATRIB(4)=2,1;
      ACT,0.0,,GO;

  GO1 GOON,1;
      ACT,0.0,ATRIB(3).LE.1.5,L8;
      ACT,0.0,ATRIB(3).LE.2.5,L8;
      ACT,0.0,ATRIB(3).LE.3.5,L8;
      ACT,0.0,ATRIB(3).LE.4.5,L7;
      ACT,0.0,ATRIB(3).LE.5.5,L7;
      ACT,0.0,ATRIB(3).LE.6.5,L7;
      ACT,0.0,ATRIB(3).LE.7.5,L6;
      ACT,0.0,ATRIB(3).LE.8.5,L5;
      ACT,0.0,ATRIB(3).LE.9.5,L5;
      ACT,0.0,ATRIB(3).LE.10.5,L5;
      ACT,0.0,ATRIB(3).LE.11.5,L4;
      ACT,0.0,ATRIB(3).LE.12.5,L4;
      ACT,0.0,ATRIB(3).LE.13.5,L2;
      ACT,0.0,ATRIB(3).LE.14.5,L3;
      ACT,0.0,ATRIB(3).LE.15.5,L1;
      ACT,0.0,ATRIB(3).LE.16.5,L3;
      ACT,0.0,ATRIB(3).LE.17.5,L1;
      ACT,0.0,ATRIB(3).LE.18.5,L1;
      ACT,0.0,ATRIB(3).LE.19.5,L1;
```

```
      ACT,0.0,ATRIB(3).LE.20.5,L9;
      ACT,0.0,ATRIB(3).LE.21.5,L13;
      ACT,0.0,ATRIB(3).LE.22.5,L10;
      ACT,0.0,ATRIB(3).LE.23.5,L11;
      ACT,0.0,ATRIB(3).LE.24.5,L14;
      ACT,0.0,ATRIB(3).LE.25.5,L12;
      ACT,0.0,ATRIB(3).LE.26.5,L12;
      ACT,0.0,ATRIB(3).LE.27.5,L15;
   L1 ASSIGN,ATRIB(5)=1,1;
      ACT,0.0,,GO;
   L2 ASSIGN,ATRIB(5)=2,1;
      ACT,0.0,,GO;
   L3 ASSIGN,ATRIB(5)=3,1;
      ACT,0.0,,GO;
   L4 ASSIGN,ATRIB(5)=4,1;
      ACT,0.0,,GO;
   L5 ASSIGN,ATRIB(5)=5,1;
      ACT,0.0,,GO;
   L6 ASSIGN,ATRIB(5)=6,1;
      ACT,0.0,,GO;
   L7 ASSIGN,ATRIB(5)=7,1;
      ACT,0.0,,GO;
   L8 ASSIGN,ATRIB(5)=8,1;
      ACT,0.0,,GO;
   L9 ASSIGN,ATRIB(5)=9,1;
      ACT,0.0,,GO;
  L10 ASSIGN,ATRIB(5)=10,1;
      ACT,0.0,,GO;
  L11 ASSIGN,ATRIB(5)=11,1;
      ACT,0.0,,GO;
  L12 ASSIGN,ATRIB(5)=12,1;
      ACT,0.0,,GO;
  L13 ASSIGN,ATRIB(5)=13,1;
      ACT,0.0,,GO;
  L14 ASSIGN,ATRIB(5)=14,1;
      ACT,0.0,,GO;
  L15 ASSIGN,ATRIB(5)=15,1;

   GO GOON,1;
      ACT,0.0;

;************************************
; THE STAR CONTROLLER               *
;************************************

      AWAIT(28),STAR/1,1;
      ACT/1,.00003*ATRIB(6)*ATRIB(7);
      FREE,STAR/1,1;
      AWAIT(29),ATRIB(5)/1,1;
      ACT,0.00003;
      GOON,1;
      ACT,.00003,ATRIB(3).LE.1.5,O102;
      ACT,.00003,ATRIB(3).LE.2.5,O104;
      ACT,.00003,ATRIB(3).LE.3.5,O106;
      ACT,.00003,ATRIB(3).LE.4.5,O107;
      ACT,.00003,ATRIB(3).LE.5.5,O112;
      ACT,.00003,ATRIB(3).LE.6.5,O114;
      ACT,.00003,ATRIB(3).LE.7.5,O116;
      ACT,.00003,ATRIB(3).LE.8.5,O118;
      ACT,.00003,ATRIB(3).LE.9.5,O122;
```

```
      ACT,.00003,ATRIB(3).LE.10.5,0124;
      ACT,.00003,ATRIB(3).LE.11.5,0129;
      ACT,.00003,ATRIB(3).LE.12.5,0130;
      ACT,.00003,ATRIB(3).LE.13.5,0137;
      ACT,.00003,ATRIB(3).LE.14.5,0138;
      ACT,.00003,ATRIB(3).LE.15.5,0139;
      ACT,.00003,ATRIB(3).LE.16.5,0140;
      ACT,.00003,ATRIB(3).LE.17.5,0141;
      ACT,.00003,ATRIB(3).LE.18.5,0143;
      ACT,.00003,ATRIB(3).LE.19.5,0144;
      ACT,.00003,ATRIB(3).LE.20.5,A103;
      ACT,.00003,ATRIB(3).LE.21.5,B109;
      ACT,.00003,ATRIB(3).LE.22.5,B110;
      ACT,.00003,ATRIB(3).LE.23.5,B112;
      ACT,.00003,ATRIB(3).LE.24.5,B113;
      ACT,.00003,ATRIB(3).LE.25.5,B116;
      ACT,.00003,ATRIB(3).LE.26.5,B120;
      ACT,.00003,ATRIB(3).LE.27.5,B119;

;*********************************************************
; COLLECT STATISTICS ON EACH MESSAGE BEFORE IT LEAVES    *
; THE NETWORK                                            *
;*********************************************************

 0102 COLCT,INT(1),0102;
      ACT,,,TERM;
 0104 COLCT,INT(1),0104;
      ACT,,,TERM;
 0106 COLCT,INT(1),0106;
      ACT,,,TERM;
 0107 COLCT,INT(1),0107;
      ACT,,,TERM;
 0112 COLCT,INT(1),0112;
      ACT,,,TERM;
 0114 COLCT,INT(1),0114;
      ACT,,,TERM;
 0116 COLCT,INT(1),0116;
      ACT,,,TERM;
 0118 COLCT,INT(1),0118;
      ACT,,,TERM;
 0122 COLCT,INT(1),0122;
      ACT,,,TERM;
 0124 COLCT,INT(1),0124;
      ACT,,,TERM;
 0129 COLCT,INT(1),0129;
      ACT,,,TERM;
 0130 COLCT,INT(1),0130;
      ACT,,,TERM;
 0137 COLCT,INT(1),0137;
      ACT,,,TERM;
 0138 COLCT,INT(1),0138;
      ACT,,,TERM;
 0139 COLCT,INT(1),0139;
      ACT,,,TERM;
 0140 COLCT,INT(1),0140;
      ACT,,,TERM;
 0141 COLCT,INT(1),0141;
      ACT,,,TERM;
 0143 COLCT,INT(1),0143;
      ACT,,,TERM;
```

```
 0144 COLCT,INT(1),0144;
      ACT,,,TERM;
 A103 COLCT,INT(1),A103;
      ACT,,,TERM;
 B109 COLCT,INT(1),B109;
      ACT,,,TERM;
 B110 COLCT,INT(1),B110;
      ACT,,,TERM;
 B112 COLCT,INT(1),B112;
      ACT,,,TERM;
 B113 COLCT,INT(1),B113;
      ACT,,,TERM;
 B116 COLCT,INT(1),B116;
      ACT,,,TERM;
 B119 COLCT,INT(1),B119;
      ACT,,,TERM;
 B120 COLCT,INT(1),B120;
      ACT,,,TERM;

 TERM FREE,ATRIB(5)/1,1;
      COLCT,INT(1),TOTAL TIME;
      ASSIGN,XX(1)=ATRIB(6)*ATRIB(7);
      COLCT,XX(1),TOTAL BYTES;
      TERMINATE;
      ENDNETWORK;
SEEDS,79725(1),93872(2);
INIT,0,3600;
SIMULATE;
FIN;
```

APPENDIX II

EXAMPLE OF SLAM II SUMMARY REPORT

S L A M I I E C H O R E P O R T

SIMULATION PROJECT ROME NETWORK BY TEW

DATE 10/ 6/1992 RUN NUMBER 1 OF 1

SLAM II VERSION AUG 87

GENERAL OPTIONS

PRINT INPUT STATEMENTS (ILIST):	YES
PRINT ECHO REPORT (IECHO):	YES
EXECUTE SIMULATIONS (IXQT):	YES
WARN OF DESTROYED ENTITIES:	YES
PRINT INTERMEDIATE RESULTS HEADING (IPIRH):	YES
PRINT SUMMARY REPORT (ISMRY):	YES

LIMITS ON FILES

MAXIMUM NUMBER OF USER FILES (MFILS):	29
MAXIMUM NUMBER OF USER ATTRIBUTES (MATR):	7
MAXIMUM NUMBER OF CONCURRENT ENTRIES (MNTRY):	15000

FILE SUMMARY

FILE NUMBER	INITIAL ENTRIES	RANKING CRITERION
1	0	FIFO
2	0	FIFO
3	0	FIFO
4	0	FIFO
5	0	FIFO
6	0	FIFO
7	0	FIFO
8	0	FIFO
9	0	FIFO
10	0	FIFO
11	0	FIFO
12	0	FIFO
13	0	FIFO
14	0	FIFO
15	0	FIFO
16	0	FIFO
17	0	FIFO

18	0	FIFO
19	0	FIFO
20	0	FIFO
21	0	FIFO
22	0	FIFO
23	0	FIFO
24	0	FIFO
25	0	FIFO
26	0	FIFO
27	0	FIFO
28	0	FIFO
29	0	FIFO

STATISTICS BASED ON OBSERVATIONS

COLCT NUMBER	COLLECTION MODE	IDENTIFIER	HISTOGRAM SPECIFICATIONS NCEL	HLOW	HWID
1	NETWORK	0102			
2	NETWORK	0104			
3	NETWORK	0106			
4	NETWORK	0107			
5	NETWORK	0112			
6	NETWORK	0114			
7	NETWORK	0116			
8	NETWORK	0118			
9	NETWORK	0122			
10	NETWORK	0124			
11	NETWORK	0129			
12	NETWORK	0130			
13	NETWORK	0137			
14	NETWORK	0138			
15	NETWORK	0139			
16	NETWORK	0140			
17	NETWORK	0141			
18	NETWORK	0143			
19	NETWORK	0144			
20	NETWORK	A103			
21	NETWORK	B109			
22	NETWORK	B110			
23	NETWORK	B112			
24	NETWORK	B113			
25	NETWORK	B116			
26	NETWORK	B119			
27	NETWORK	B120			
28	NETWORK	TOTAL TIME			
29	NETWORK	TOTAL BYTES			

RANDOM NUMBER STREAMS

STREAM NUMBER	SEED VALUE	REINITIALIZATION OF STREAM

1	21825	NO
2	19585	NO
3	1145661099	NO
4	1835732737	NO
5	794161987	NO
6	1329531353	NO
7	200496737	NO
8	633816299	NO
9	1410143363	NO
10	1282538739	NO

INITIALIZATION OPTIONS

BEGINNING TIME OF SIMULATION (TTBEG): 0.0000E+00
ENDING TIME OF SIMULATION (TTFIN): 0.3600E+04
STATISTICAL ARRAYS CLEARED (JJCLR): YES
VARIABLES INITIALIZED (JJVAR): YES
FILES INITIALIZED (JJFIL): YES

NSET/QSET STORAGE ALLOCATION

DIMENSION OF NSET/QSET (NNSET): 200000
WORDS ALLOCATED TO FILING SYSTEM: 165000
WORDS ALLOCATED TO VARIABLES: 15950
WORDS AVAILABLE FOR PLOTS/TABLES: 19050

INPUT ERRORS DETECTED: 0

EXECUTION WILL BE ATTEMPTED

INTERMEDIATE RESULTS

S L A M I I S U M M A R Y R E P O R T

SIMULATION PROJECT ROME NETWORK BY TEW

DATE 10/ 6/1992 RUN NUMBER 1 OF 1

CURRENT TIME 0.3600E+04
STATISTICAL ARRAYS CLEARED AT TIME 0.0000E+00

STATISTICS FOR VARIABLES BASED ON OBSERVATION

MEAN VALUE	STANDARD DEVIATION	COEFF. OF VARIATION	MINIMUM VALUE	MAXIMUM VALUE	NUMBER OF OBSERVATIONS

0102	0.4050E+03	0.2050E+03	0.5062E+00	0.1338E+02	0.7508E+03	283
0104	0.3898E+03	0.2228E+03	0.5716E+00	0.2106E+01	0.7484E+03	249
0106	0.4061E+03	0.2097E+03	0.5163E+00	0.1903E+02	0.7519E+03	275
0107	0.3941E+03	0.2059E+03	0.5226E+00	0.2009E+02	0.7510E+03	274
0112	0.3918E+03	0.2110E+03	0.5385E+00	0.4579E+00	0.7527E+03	272
0114	0.4288E+03	0.2156E+03	0.5027E+00	0.1255E+02	0.7468E+03	240
0116	0.4030E+03	0.2166E+03	0.5375E+00	0.1465E+01	0.7513E+03	269
0118	0.3987E+03	0.2054E+03	0.5152E+00	0.6638E+01	0.7522E+03	253
0122	0.3938E+03	0.2115E+03	0.5372E+00	0.9339E+01	0.7491E+03	265
0124	0.4057E+03	0.2079E+03	0.5124E+00	0.1413E+02	0.7472E+03	282
0129	0.3831E+03	0.2180E+03	0.5690E+00	0.1342E+02	0.7525E+03	252
0130	0.3933E+03	0.2245E+03	0.5709E+00	0.9705E+01	0.7466E+03	260
0137	0.4001E+03	0.2047E+03	0.5117E+00	0.1456E+02	0.7505E+03	255
0138	0.3525E+03	0.2093E+03	0.5937E+00	0.1271E+02	0.7477E+03	244
0139	0.3877E+03	0.2155E+03	0.5558E+00	0.1335E+02	0.7512E+03	263
0140	0.3896E+03	0.2128E+03	0.5461E+00	0.1021E+02	0.7528E+03	237
0141	0.3886E+03	0.2150E+03	0.5533E+00	0.1048E+02	0.7480E+03	254
0143	0.3911E+03	0.2152E+03	0.5503E+00	0.2610E+01	0.7490E+03	246
0144	0.3773E+03	0.2081E+03	0.5516E+00	0.1300E+02	0.7476E+03	281
A103	0.3932E+03	0.2235E+03	0.5683E+00	0.1312E+02	0.7503E+03	289
B109	0.4033E+03	0.2061E+03	0.5109E+00	0.3800E+01	0.7486E+03	243
B110	0.3758E+03	0.2149E+03	0.5720E+00	0.1282E+01	0.7515E+03	247
B112	0.3837E+03	0.2208E+03	0.5755E+00	0.8790E+01	0.7481E+03	288
B113	0.4015E+03	0.2098E+03	0.5225E+00	0.7371E+01	0.7531E+03	293
B116	0.3874E+03	0.2175E+03	0.5613E+00	0.7691E+01	0.7529E+03	292
B119	0.4023E+03	0.2241E+03	0.5572E+00	0.1310E+02	0.7528E+03	252
B120	0.3830E+03	0.2199E+03	0.5742E+00	0.5951E+01	0.7488E+03	268
TOTAL TIME	0.3928E+03	0.2139E+03	0.5446E+00	0.4579E+00	0.7531E+03	7126
TOTAL BYTES	0.1684E+05	0.4988E+04	0.2963E+00	0.0000E+00	0.3510E+05	7126

FILE STATISTICS

FILE NUMBER	LABEL/TYPE	AVERAGE LENGTH	STANDARD DEVIATION	MAXIMUM LENGTH	CURRENT LENGTH	AVERAGE WAITING TIME
1	AWAIT	0.0000	0.0000	1	0	0.0000
2	AWAIT	0.0000	0.0001	1	0	0.0000
3	AWAIT	0.0000	0.0001	1	0	0.0000
4	AWAIT	0.0000	0.0000	1	0	0.0000
5	AWAIT	0.0000	0.0001	1	0	0.0000
6	AWAIT	0.0000	0.0001	1	0	0.0000
7	AWAIT	0.0000	0.0000	1	0	0.0000
8	AWAIT	0.0000	0.0000	1	0	0.0000
9	AWAIT	0.0000	0.0001	1	0	0.0000
10	AWAIT	0.0000	0.0001	1	0	0.0000
11	AWAIT	0.0000	0.0000	1	0	0.0000
12	AWAIT	0.0000	0.0001	1	0	0.0000
13	AWAIT	0.0000	0.0000	1	0	0.0000
14	AWAIT	0.0000	0.0000	1	0	0.0000
15	AWAIT	0.0000	0.0000	1	0	0.0000
16	AWAIT	0.0000	0.0001	1	0	0.0000
17	AWAIT	0.0000	0.0001	1	0	0.0000

18	AWAIT	0.0000	0.0001	1	0	0.0000
19	AWAIT	0.0000	0.0002	1	0	0.0000
20	AWAIT	0.0000	0.0000	1	0	0.0000
21	AWAIT	0.0000	0.0000	1	0	0.0000
22	AWAIT	0.0000	0.0000	1	0	0.0000
23	AWAIT	0.0000	0.0000	1	0	0.0000
24	AWAIT	0.0000	0.0000	1	0	0.0000
25	AWAIT	0.0000	0.0000	1	0	0.0000
26	AWAIT	0.0000	0.0001	1	0	0.0000
27	AWAIT	0.0000	0.0000	1	0	0.0000
28	AWAIT	967.0222	521.9036	1877	1876	386.6797
29	AWAIT	0.0000	0.0000	1	0	0.0000
30	CALENDAR	28.0000	0.0031	44	28	0.7417

REGULAR ACTIVITY STATISTICS

ACTIVITY INDEX/LABEL	AVERAGE UTILIZATION	STANDARD DEVIATION	MAXIMUM UTIL	CURRENT UTIL	ENTITY COUNT
1	1.0000	0.0001	1	1	7126

RESOURCE STATISTICS

RESOURCE NUMBER	RESOURCE LABEL	CURRENT CAPACITY	AVERAGE UTILIZATION	STANDARD DEVIATION	MAXIMUM UTILIZATION	CURRENT UTILIZATION
1	LINE1	1	0.0000	0.0011	1	0
2	LINE2	1	0.0000	0.0005	1	0
3	LINE3	1	0.0000	0.0007	1	0
4	LINE4	1	0.0000	0.0008	1	0
5	LINE5	1	0.0000	0.0008	1	0
6	LINE6	1	0.0000	0.0005	1	0
7	LINE7	1	0.0000	0.0009	1	0
8	LINE8	1	0.0000	0.0009	1	0
9	LINE9	1	0.0000	0.0006	1	0
10	LINE10	1	0.0000	0.0005	1	0
11	LINE11	1	0.0000	0.0006	1	0
12	LINE12	1	0.0000	0.0008	1	0
13	LINE13	1	0.0000	0.0005	1	0
14	LINE14	1	0.0000	0.0006	1	0
15	LINE15	1	0.0000	0.0005	1	0
16	STAR	1	1.0000	0.0001	1	1

RESOURCE NUMBER	RESOURCE LABEL	CURRENT AVAILABLE	AVERAGE AVAILABLE	MINIMUM AVAILABLE	MAXIMUM AVAILABLE
1	LINE1	1	1.0000	0	1
2	LINE2	1	1.0000	0	1
3	LINE3	1	1.0000	0	1

4	LINE4	1	1.0000	0	1
5	LINE5	1	1.0000	0	1
6	LINE6	1	1.0000	0	1
7	LINE7	1	1.0000	0	1
8	LINE8	1	1.0000	0	1
9	LINE9	1	1.0000	0	1
10	LINE10	1	1.0000	0	1
11	LINE11	1	1.0000	0	1
12	LINE12	1	1.0000	0	1
13	LINE13	1	1.0000	0	1
14	LINE14	1	1.0000	0	1
15	LINE15	1	1.0000	0	1
16	STAR	0	0.0000	0	1

References

ANDERSON, T. W. 1984. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, New York.

ARNOLD, S. F. 1981. *The Theory of Linear Models and Multivariate Analysis.* John Wiley & Sons, New York.

AVRAMIDIS, A. N. AND J. R. WILSON. 1992. Integrated Variance Reduction Strategies for Simulation. submitted to *Operations Research.*

BRATELY, P., B. L. FOX AND L. E. SCHRAGE. 1987. *A Guide to Simulation*, Second Edition. Springer-Verlag, New York.

CHENG, R. C. H. 1982. The Use of Antithetic Variates in Computer Simulations. *Journal of the Operational Research Society*, **33**, 229–237.

DEVAI, J. J., T. C. KERRIGAN, AND M. K. MOLLOY. 1990. Simulation of the LAN Behavior in the Distributed HP-UX Environment. *Proceedings of the 1990 Winter Simulation Conference*, 805–812.

HUSSEY, J. R., R. H. MYERS, AND E. C. HOUCK. 1987a. Correlated Simulation Experi- ments in First-Order Response Surface Design. *Operations Research* **35**, 744–758.

HUSSEY, J. R., R. H. MYERS, AND E. C. HOUCK. 1987b. Pseudorandom Number Assignment in Quadratic Response Surface Designs. *IIE Transactions* **19**, 395–403.

KAVI, P., V. S. FROST, AND K. S. SHANMUGAN. 1991. Generic Approach to LAN Modeling. *Proceedings of the 1991 Winter Simulation Conference*, 716–724.

KLEIJNEN, J. P. C. 1974. *Statistical Techniques in Simulation, Part I.* Marcel Dekker, New York.

LAVENBERG, S. S., T. L. MOELLER AND P. D. WELCH. 1982. Statistical Results on Control Variates with Application to Queueing Network Simulation. *Operations Research*, **30**, 182–202.

LAW, A. M., AND W. D. KELTON. 1991. *Simulation Modeling & Analysis*, Second Edition. McGraw Hill, New York.

LEHMANN, E. L. 1966. Some Concepts of Dependence. *Annals of Mathematical Statistics*, **37**, 1137–1153.

MILLS, R. 1988. Telecommunication Network Analysis with COMNET II.5. *Proceedings of the 1988 Winter Simulation Conference*, 162–164.

MORRISON, D. F. 1976. *Multivariate Statistical Methods*, Second Edition. McGraw Hill, New York.

NOZARI, A., S. F. ARNOLD, AND C. D. PEGDEN. 1987. Statistical Analysis for Use with the Schruben and Margolin Correlation Induction Strategy. *Operations Research* **35**, 127–139.

PRITSKER, A. A. B. 1974. *The GASP IV Simulation Language*, John Wiley & Sons, New York.

PRITSKER, A. A. B. 1986. *Introduction to Simulation and SLAM II*, Third Edition. John Wiley & Sons, New York.

SEARLE, S. R. 1971. *Linear Models.* John Wiley & Sons, New York.

TEW, J. D. 1986. Metamodel Estimation Under Correlation Methods for Simulation Experiments. Ph.D. dissertation, School of Industrial Engineering, Purdue University, West Lafayette, Ind.

TEW, J. D. AND J. R. WILSON. 1992. Validation of Simulation Analysis Methods for the Schruben-Margolin Correlation-Induction Strategy. *Operations Research*, **40**, 87–103.

TEW, J. D. AND J. R. WILSON. 1993. Estimating Simulation Metamodels Using Combined Correlation-Based Variance Reduction Techniques. *IIE Transactions*, to appear.