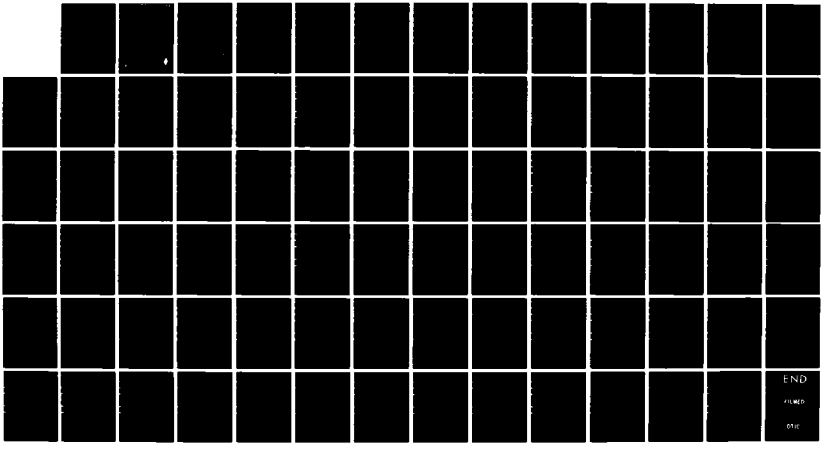
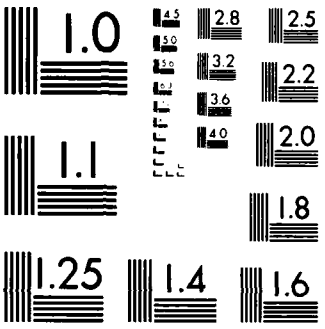
AD-A158 768 SPECIFIC RECOMMENDATIONS FOR IMPROVING NETWORK FEEDBACK 1/1
TO HOSTS(U) SPARTA INC MCLEAN VA 22 JUN 85
URW-5-DCA-039 DCA100-84-C-0085

UNCLASSIFIED F/G 9/2 NL





MICROCOPY RESOLUTION TEST CHART

Specific Recommendations

for

Improving Network Feedback to Hosts

REPORT: URW-5-DCA-039

Specific Recommendations
for
Improving Network Feedback to Hosts

22 June 1985

PREPARED FOR:

DEFENSE COMMUNICATIONS ENGINEERING CENTER
1860 Wiehle Avenue
Reston, VA 22090

Contract DCA100-84-C-0085

SPARTA, INC.
7926 Jones Branch Drive, Suite 1070
McLean, VA 22102
(703) 448-0210

SECURITY CLASSIFICATION OF THIS PAGE AD-A158 768

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188
Exp Date Jun 30 1986

1a REPORT SECURITY CLASSIFICATION: UNCLASSIFIED	1b RESTRICTIVE MARKINGS: NONE
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION/AVAILABILITY: Approved for public release; UNLIMITED Distribution Unlimited
2b DECLASSIFICATION/DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S): URW-5-DCA-039	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION: SPARTA, INC.	6b OFFICE SYMBOL (If applicable)
6c ADDRESS (City, State, and ZIP Code): 7926 Jones Branch Drive, Suite 1070 McLean, Va 22102	7a NAME OF MONITORING ORGANIZATION: Defense Communication Engineering Center
	7b ADDRESS (City, State, and ZIP Code): 1860 Wiehle Ave Reston, VA. 22090
8a NAME OF FUNDING/SPONSORING ORGANIZATION: DCEC	8b OFFICE SYMBOL (If applicable): R640
	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: DCA100-84-C-0085
8c ADDRESS (City, State, and ZIP Code): 1860 Wiehle Ave Reston, VA. 22090	10 SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO; PROJECT NO; TASK NO; WORK UNIT ACCESSION NO.

11 TITLE (Include Security Classification): Specific Recommendations for Improving Network Feedback to Hosts (U)

12 PERSONAL AUTHOR(S): SPARTA, INC.

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
FINAL	FROM _____ TO _____	85,06,22	78

16 SUPPLEMENTARY NOTATION

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number): Computer Networks, Defense Data Network, Gateways, Internetworking

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

This report provides recommendations for improving the feedback from the DDN network switching resources, IMPs, to their attached gateways/hosts so that the internetwork as a whole can function more efficiently and effectively. The report begins with a general description of the current internetworking environment. Problem areas in designing a efficient packet switching network or internetwork are identified along with potential improvements to current internetworking control mechanisms in the areas of flow control, congestion control, type of service routing, uses of multipath routing and the impact/use of the new IMP end-to-end protocol. Emphasis is placed on the uses of the IMP end-to-end protocol to provide statistics about the network and on the DDN X.25 access protocol as the medium for passing this information to the gateways and hosts. Several approaches and techniques were identified which could improve the efficiency and effectiveness of the DoD internetwork and the DDN specifically. The main information required by the hosts is readily available within the IMP, or at least will be available as the new IMP end to end protocol is introduced later this year. In order to develop new control algorithms, delay/capacity statistics need to be collected and

(Continue on attached page)

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED
22a NAME OF RESPONSIBLE INDIVIDUAL: Mr. Ed Cain	22b TELEPHONE (Include Area Code): (703) 437-2578 / 22c OFFICE SYMBOL: R640

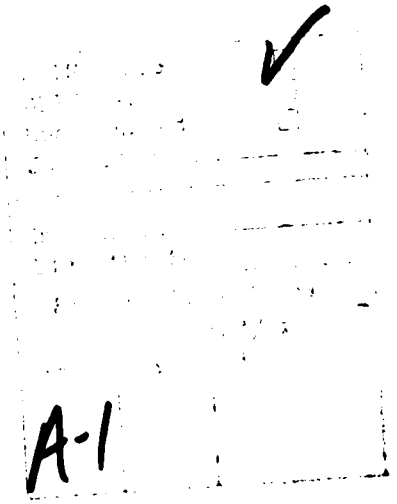
DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

Specific Recommendations for Improving Network Feedback to Host (U) Cont.

19 ABSTRACT Cont.:

analyzed. Network feedback mechanisms are proposed which would allow operational host/gateways to obtain the network delay/capacity parameters needed for optimal route selection, flow control, congestion control, etc.



A-1

Table of Contents

1.0 Introduction

This report is one of four tasks to be accomplished for DCA under the topic of Analysis and Resolution of Packet Switching Issues. The four tasks are:

1. Design an Area Routing Scheme for the DDN
2. Develop a Gateway Functional Requirements Document
3. Identify the Requirements for the Next Generation Packet Switch
4. Develop Specific Recommendations for Improving Network Feedback to Hosts

The purpose of this report, Task 4, is to provide DCA with recommendations for improving the feedback from the DDN network switching resources, IMPs, to their attached gateways/hosts so that the network as a whole can function more efficiently and effectively. The basic premise as stated in the RFP is: "As DDN hosts develop more sophisticated network software, the network should be able to provide them with more detailled information about conditions in the network. This would allow the hosts to make more intelligent choices about routing and traffic flows. Such information can be particularly useful to gateways." We make specific recommendations in this report which will allow the DDN to become more efficient. However equally important is that these recommendations were derived from a review of the internet

issues and should also allow the entire internet to become more efficient.

The DDN should be viewed as just one part of a wider internet, even though the DDN itself may consist of several component networks. In fact most users of the DDN will eventually not be directly connected to DDN switches but will indirectly use the DDN via gateways from LANs, packet radio nets, or other networks. It is important therefore, not just to make the DDN more efficient, but to make the entire internet more efficient and effective.

The methodology we used in pursuing this task starts with the global internet. We first looked at ways to improve the overall internet performance and how to make an internet efficient for the ultimate users. This analysis identified the information needed by gateways and hosts to help accomplish these improvements. Once we identified the problem areas and information needed for an efficient internet, we then looked at ways the supporting networks (e.g. DDN) could acquire and distribute the necessary information to the gateways and hosts. And finally we identified the specific DDN network mechanisms/protocols along with the changes or improvements, needed to get the data to the gateways/hosts. By approaching the task in this manner we were assuring that the ultimate goal of developing an efficient internet would not be subverted by

suboptimizing for certain improvements in the DDN specific network. Any such improvements in the DDN network would most likely have been illusory since inefficiencies in other portions of the internet would have eventually backed up into and been apparent within the DDN and caused degraded service to the all users. There are many examples of this occuring even today in the ARPANET/MILNET split and the problems apparent when addressing hosts on the other net.

An analogy we found useful throughout our study was in comparing the global internet problems and proposed improvements to the DDN/ARPANET problems and proposed improvements. The DDN consists of switches, trunks, and end terminations (hosts, TACs, gateways, etc.). Similiarly the internet consists of gateways (analogous to switches), networks (analogous to trunks), and end terminations (which are the hosts). The general problem areas are also analogous: i.e. congestion control, flow control, type of service support, routing, etc. Just as the DDN is a richly interconnected network of IMPs connected by trunks of certain capacity and delay, so the internet is expected to be a rich interconnection of gateways connected by networks of varying capacity and delays. We found many of the latest ideas for improving the network performance (i.e. multipath routing) are applicable to improving the internet performance.

Of course many of the problems faced by the DDN and/or the

internet are caused by legitimate but conflicting user requirements. That is, some users need minimum delay or high bandwidth or maximum reliabilty or survivablity or priority service. All users have cost constraints, though they may differ. Networks can be designed to meet subsets of these requirements. However when trying to meet multiple requirements such as these, tradeoffs and compromises are required. One of the best approaches in addressing conflicting requirements is to maintain maximum flexibility and adaptibility to be able to respond to dynamically changing conditions.

We found that the current implementations of network protocol software are not very adaptive to changing network conditions. Rather they must be fine tuned for special applications. Part of the reason for this is that protocol parameters are often influenced by the worst case considerations. Consequently we considered implementations in which the values for key parameters (e.g. timeouts, segment sizes, etc.) can be dynamically changed in response to observed network conditions such as instantaneous delay, degree of network utilization, etc. The information requirements of adaptive protocol implementations were examined, as was the availability of this information from the DDN IMPs. Where information was available within, or could be collected by the IMPs, recommendations are made for the reporting of this information to the hosts such that more intelligent decisions could be made by the hosts or gateways.

In the end we found several approaches for improving the feedback of information to gateways/hosts which would help the network respond more efficiently and effectively to the user community. Some of these techniques require more research before they could be implemented in an operational network, but most of them could be implemented with relative ease in conjunction with on-going upgrades to the DDN.

The remainder of this report is organized as follows. After this introduction we give a general description of the current internetworking environment. Our purpose is to describe the basic resources, mechanisms, and protocols, how they are used, *and essentially a theory of operation of the internet with* particular attention to the status of the DDN portion. New developments with their potential impacts are also discussed. Following a description of the internet environment, we then provide some background on fundamentals of efficient network design and utilization. Here we discuss some of the basic problem areas in designing a packet switching network or internetwork, the inherent conflicts in user requirements, and some general guidelines on ways to resolve these problem areas. In section 4 we then present the results our studies and identify some potential improvements to internetworking control mechanisms. In particular we cover the areas of flow control, congestion control, type of service routing, uses of multipath

routing and the impact/use of the new IMP end to end protocol. Section 5 then specifies how to implement these improvements in the DDN, by describing changes to the current DDN access mechanisms and protocols. Here we have placed particular emphasis on the uses of the IMP end to end protocol to provide statistics about the network and on the DDN X.25 access protocol as the medium for passing this information to the gateways and hosts. Section 6 then summarizes our findings and makes final recommendations on what and how modifications should be made to the DDN.

2. Environment

2.1 Introduction

This section describes the overall environment (architecture) within which the DDN protocol services are implemented, the goal of this section is to provide enough information for the reader to understand the assumptions we have made about the DDN internet in developing our recommendations.

To productively examine Defense Data Network (DDN) protocol services and their effect on efficient resource usage, we must examine those "things" which constitute the network (or internet) resource pool. Todays DDN is not a single homogeneous network, the DDN is the aggregation of many different networks, an internet. The resources of this internet are its constituent networks, the gateways connecting them, and the hosts transmitting and receiving information across the internet. Each network has its own set of resources, communications channels, switches, and hosts and gateways, which it uses to provide the services necessary to support the internet. The access points of the network, hosts and gateways, themselves have resources they can call upon to perform their specific services; e.g., providing computation or data base management to a user, or routing a message from one network to another. In this section we will examine what it is that makes up the DDN and talk about

the service goals which the network must optimize around.

We will also examine the DoD internet, networks, and hosts/gateways from the perspective of the DDN backbone, with a view towards the services necessary to provide efficient and effective service to DDN backbone users. The purpose of this report is to recommend improvements in the information provided by the DDN to its attached hosts and gateways so they may make better decisions about routing and traffic flows.

Some of the networks which participate in the DDN, such as Local Area Networks (LANs) or packet radio networks, are not part of the 'backbone'DDN network and instead simply provide access channels which hosts may use to get to the DDN. The access networks are very important to the overall usefulness of the internet, but since they are not part of the general transport system, the backbone, we have not addressed them as part of our study. However, we do expect that the recommendations we make for the DDN will apply to these other networks as well.

2.2 Internet

2.2.1 Description

The DoD internet consists of many heterogeneous networks connected to each other via gateways. In order to facilitate

internet communications the DoD has standardized on the Internet Protocol (IP) as the method of communicating from one network, i.e., an Ethernet, to another incompatible network, i.e., the MILNET. The translation from the Ethernet to the MILNET takes place in a gateway, known as an IP gateway because it connects two systems at the IP level. Today the IP gateway is the primary method of allowing two incompatible networks to communicate. Though LANs, packet radio nets, satellite packet nets, etc., are full members of the internet our thrust in this report is to examine only what information may be available from the backbone DDN which will allow the hosts and gateways to make improvements in their routing and flow allocation decisions.

2.2.2 Resources

The internet, as described above, is a communications service provider to its subscribers, the hosts. The service goals of the internet are: to transfer information from one subscriber another with as little delay as possible; to provide a minimum service to every subscriber; and, to maintain the network so as to support the first two goals. There are a number of resources which the internet can call upon to perform it's services, and it is the efficient and effective management of these resources which provides acceptable service to the internet subscribers.

Analogous to a single network, the resources of the internet

consist of communications links, the member sub-networks, switches, the gateways, and hosts. The subnets perform the basic trunking for the internet and are characterized according to their throughput and delay characteristics, just as a communications link in a single network. Gateways perform routing functions, just like a switch. Hosts are the sources and sinks of information within the internet.

Though an internet's constituent networks at a high level approximate communications channels, they are obviously much more complex. In a network a communications channel is characterized according to its delay, how long it takes to send a message to the other end, throughput, how many data units the channel can accept in a period of time, and error rate, how many data units will be correctly received. In a lightly loaded internet, throughput and delay are the only items of interest to the gateways, and these parameters are relatively static, just as they are in a network. As the internet becomes more highly utilized the constituent networks may tend to become congested, that is their delay and throughput characteristics change with time. The ability to accurately measure the change in throughput and delay is absolutely necessary for efficient internet resource usage. The mechanisms available to the subnet to relay this information to the gateways is discussed in sections 2.2, and 5.0 below.

Gateways are quite similar to the switches in a network. The primary service goal of the internet gateway is to route internet datagrams to their destination networks. In the case of a lightly loaded internet the fulfillment of this goal is quite similar to the functioning of an IMP. Unlike an IMP the gateway connects to "trunks" which have differing characteristics. These differing characteristics cause severe flow and congestion control problems as the internet becomes more highly utilized.

Gateways take in datagrams from their attached networks and attempt to route the datagrams to the proper network. If a gateway becomes overutilized, congested, it will begin simply throwing away datagrams if it does not have a buffer to store it in. The gateway relies on the higher level host to host protocols to recover from the loss of these datagrams. This strategy would work well in a lightly loaded internet where the congestion is local to the gateway, but in the heavily loaded internet, once any gateway begins throwing away packets all hosts going through that gateway begin retransmitting their datagrams, congesting the local network and, because there is no internet congestion control, the internet. The gateway must become an active party in the management of the information flow between hosts to solve the congestion and flow control problems mentioned above. Further discussion of the role of the gateway in internet management and control is in Sections 3 and 4.

2.2.3 Protocols within the Internet

In order to mee it's performance goal the internet has established a suite of protocols. The following are those currently in use and most germane to this task.

2.2.3.1 IP

The Internet Protocol is used between hosts and gateways to deliver independent packets over an internet. It provides a datagram service and therefore, it contains no flow control, sequencing, or error control for data. If desired, these services are assumed to be at the next higher, end-to-end, protocol level (the transport layer).

2.2.3.2 Internet Control Message Protocol

The Internet Control Message Protocol (ICMP) is an extension of IP used to relay control information between gateways and hosts, or hosts and hosts. Typically, ICMP is used to report errors in the processing of datagrams. Such errors could include the parameters specified in the IP header are incorrect for the destination network, the destination is unreachable, that is the gateway cannot find a path from itself to the destination network, reroute traffic to another gateway, or source quench, i.e., stop sending traffic to this gateway.

Information which gateways may have which is of interest to hosts could be relayed using ICMP. Information of this type is discussed in Section 4.

2.2.3.3 Gateway to Gateway Protocol

The internet gateways use the Gateway-to-Gateway Protocol (GGP) to determine connectivity to networks and neighbor gateways. It is also used in the implementation of a dynamic, shortest-path first routing algorithm.

2.2.3.4 Exterior Gateway Protocol

Within the internet the Exterior Gateway Protocol (EGP) is used to determine connectivity to networks and neighbor gateways. EGP is intended for use in gateways which connect "stub" networks to the internet. EGP is used to determine, between a pair of exterior neighbors, which networks may be reached from the other. EGP enables each network to have an independent routing algorithm whose operation cannot be disrupted by failures of other networks.

2.2.4 Internet Evolution

The internet is still evolving, today most gateways within

the internet are interior gateways, that is, they all are part of the same giant internet. This situation is getting more and more difficult to handle as radically different gateways try to participate in a common routing algorithm. Maintenance and fault isolation is becoming nearly impossible. This problem is being addressed in part by the exterior gateway protocol, because new gateways can perform their routing as they wish and not participate directly in GGP routing. However, this is only a temporary fix. There is work going on to define how "autonomous systems" may be interconnected, essentially giving us an internet of internets. The evolutionary nature of the internet sould be recognized. We believe that the recommendations given in Sections 3. and 4. will prove applicable to the internet as it evolves.

2.3 Packet Switching Network

2.3.1 Description

For the purposes of this report a DDN packet switching network is simply any ARPAnet IMP based network. These networks provide basic transport of information from one network access point, a gateway or host, to another. These networks provide no service above and beyond that necessary to move information from one point to another, i.e., servers are not considered part of the network.

2.3.2 Resources

The most basic resource of the network is its communications channels, the communications channels provide the upper limit on the information carrying capacity of the network. Ideally, the communications channels are perfect, they can carry any amount of traffic, with no delay, and it will be received error free. In practice, though, any number of types of communications channels may be used, from low delay moderate throughput telephone lines, to high delay, high throughput satellite links. Management of this resource is rather straight forward; the maintenance of the channels, predicting future bandwith needs and procuring the service from the common carriers.

The IMPs are responsible for inserting and removing information from the communications channels and spreading the load over the available bandwith according to the policies of the network embodied in the IMPs routing algorithm. In the current DDN the loading policy is to send user data over the lowest delay path available.

IMPs are abundant both to service a large community and to assure survivability/availability of the network. The IMPs are designed to keep a dynamic map of the network so they can continue to perform their services in the event of the failure of

one of their brethren. An IMP which has become congested and no longer is on the lowest delay path between two subscribers will be routed around until the congestion clears up. The management of the IMP system entails assuring that all IMPs are using the current network policies, and that they are being updated correctly.

Hosts are the subscribers to the local network. There are several types of hosts but to the network they all look the same, none are more equal than others. There are basically three types of host: normal, or user, hosts; gateways; and, terminal access controllers (TACs).

Normal hosts, or user hosts, are general purpose computers, they are used to perform general processing services such as word processing and data base management. User hosts may use the network (and internet) to provide to users the appearance of a host which is capable of much more computation than the basic machine. A weather system is capable of presenting to a user the appearance that it is everywhere at once through the use of the network. A user host may also be a specialized processor which is used by other user hosts to perform certain tasks such as numerical calculation, data base searching, etc.

Gateways are specialized processors which present the services of a host to the network and, as discussed above in

Section 2.1, the services of a switch to the internet. A gateway is different form a normal host because it may receive information from any of a large number of hosts and so may become an internet bottleneck if many hosts from one net wish to talk to a single host on another. Although from a network view hosts and gateways are the same, a user on a host may not communicate directly with a gateway.

TACs are specialized processors also, but their purpose is simply to provide a way to multiplex many terminal users through the same access line on an IMP. Today TAC is an economic necessity if cost effective dial-up service is to be provided. A TAC, like the gateway, may receive or transmit information to any of a large number of hosts.

2.3.3 Sub-network Protocols

Subnetwork protocols are the part of the network layer of the DoD (ISO) Protocol Reference Model which provide to the internet sub-layer data transfer, status and congestion control services within a homogeneous communications subnetwork. Within the subnetwork these protocols provide routing, local (node-to-node) flow control, local congestion control, status reporting, and subnet control (recovery from IMP failure, routing table updates, etc.). Access protocols also fall within the sub-network layer in the DoD Reference Model, but due to their importance in the

relaying of network status and control information from the network to hosts and gateways, section 5 has been devoted to the discussion of these protocols.

We will examine two subnet protocols, the existing IMP-to-IMP protocol and the new proposed IMP end-to-end protocol. We also look at the Multipath technology and describe its potential for increased subnet efficiency.

2.3.3.1 IMP to IMP Protocol

The current IMP-to-IMP protocol provides an unreliable transport service to the higher level protocols which are then responsible for checking the ordering and integrity of the data units.

The current IMP-IMP protocol has a limit of 8 outstanding messages between any host pair, regardless of the connections across which the messages are being sent. This is a particularly severe problem for TACs and gateways because of the potential for many processes needing the resources between two hosts but on different connections. In a case where two users are sending information between a host pair where one user is interactive and the other is performing file transfer the file transfer user may cause the interactive user to experience a delay because the file transfer is using the entire outstanding message allocation.

The current IMP-IMP protocol relies on buffering taking place at the destination IMP so packets lost due to intermediate node failure will cause the sending host to retransmit the entire message. This is wasteful of the network resources to retransmit an entire message, up to 8056 bits, when only the lost subnet packet, up to 1008 bits, needs to be sent to complete the transfer.

Type of service (TOS) is supported only by a single bit, the precedence bit, which is quite hard to map to the 4 precedence levels of DDN X.25 or the 256 "handling types" of the 1822 host access protocol.

2.3.3.2 IMP End-to-end Protocol (EE)

DCA has recognized the problems with the existing IMP-IMP protocol and has begun work on a new IMP End-to-end Protocol (EE). The goal of EE is to increase the efficiency of the subnet for all users and, in partucular, for X.25 and connection oriented 1822 hosts. EE will also provide the subnet with more of the information which can be used to improve host and gateway knowledge of the local network.

The major feature of the new IMP End-to-end protocol (EE) is that it is a reliable subnetwork protocol. Unlike the current

IMP-to-IMP protocol, EE provides to higher level protocols a reliable connection oriented subnetwork transport service. By providing a reliable subnetwork service the network can make more efficient use of its resources by not requiring entire messages (X.25 packets) to be retransmitted because of the loss of a single subnet packet. EE will retransmit missing packets at the IMP level, not requiring retransmission from the host.

EE will allow for more outstanding messages per host pair than the existing IMP-to-IMP protocol in two ways; first, by measuring outstanding messages on a per connection basis, rather than per host pair; and secondly, by allowing from one to 128 outstanding messages on each connection, rather than eight. This feature will make the EE compatible with X.25, which allows flow control windows up to 127 X.25 packets in length. This feature will be particularly useful for gateways and TACs.

The level three protocols, X.25 and 1822, in the IMP are the mechanisms used for flow control on each of their connections. Congestion metrics are also maintained EE for each connection, as well as each IMP, these metrics, local and global, respectively, are updated in each acknowledgement sent from destination to source. Based on these metrics EE can tell the level three protocol to slow down or stop sending traffic due to congestion in the source or destination node. If the network in general is congested EE detects this by slow response from the store and

forward section of the IMP and will tell level three to slow down or stop.

The ability to route based on type of service (TOS) considerations has long been a goal of network designers, particularly in the DoD, the new EE allows for 4 precedence levels and other type of service indications which may be used for routing decisions, as well as connection block preemption. X.25 also contains 4 precedence levels so routing and preemption decisions at the X.25 level map directly to the subnet.

2.3.3.3 Multipath

Multipath is not a subnet protocol but a research topic which shows much promise for increasing the efficiency of the network in handling a wide range of loads. We mention it here because most of the work to date has been done on Multipath routing in ARPAnet IMPs within the subnet. As well as providing better service, higher throughput for those users who need it, and low delay for low volume users, the Multipath technology will provide the IMPs with more information about the network which may be of use to the hosts/gateways.

Within the network the multipath routing algorithm would optimize throughput by measuring the capacity of each path from source to destination and then using the paths in order of least

delay. Once the current path reaches its allocated capacity additional information is routed across the next shortest delay path, this procedure continues until either all paths are utilized to capacity or the information flow stabilizes. Through this technique high volume users will be able to transmit large amounts of information with delays proportional to the volume. A low volume user, a user who requires less than the capacity of the shortest path, will experience no greater delay than currently experienced.

To arrive at the sets of paths from any point to any other a number of items, currently not available, need to be made available to the IMPs. The capacity of any path will be limited to the available capacity of the slowest link along the path, to decide what that capacity is the IMP must know the capacities of all links in the network. Since available capacity varies with time, links go up and down, nodes become congested, etc., how, when, and how often to perform measurement and reporting are crucial issues. The compilation of capacity information is still a topic for research, but this information is potentially of significant value to hosts and gateways.

Within the subnet, capacity information will allow hosts and gateways to make more intelligent routing decisions. In an internet environment a host chooses the first gateway of an internet route based on the shortest internet path first, that is

which gateway will allow access to the destination in the shortest number of gateway hops.

3. Fundamentals of Efficient Network Utilization

3.1 Introduction

In this section we will provide some background information on the basic problems and conflicting requirements faced by network designers in developing the current packet switched internet. We will focus on three primary areas: congestion control, flow control, and routing (including type of service routing). We will also identify the current mechanisms and protocols designers can use in the near term to make modifications and improvements to the internetwork and DDN network performance.

Flow control attempts to match the rate at which flows are introduced into the network with the rate at which they are being delivered to the destination hosts. Flow control is basically a host to host problem across the network. Properly implemented flow control will limit the host to host communications to the slower of the two hosts. The philosophy of the packet switching internet is to use flow control to force the source host to buffer information, rather than have the network buffer information intended for the slower destination host. This allows more efficient usage of the available internet resources. Our purpose in this paper is to provide the source and destination hosts or gateways with enough information about the status of the network so that they can implement an appropriate

flow control methodology.

Congestion control attempts to keep the subnetwork or internetwork itself from being overrun by levels of traffic that it is not equipped to handle. Congestion control is closely related to flow control, however congestion control is considered to be the network's problem as opposed to a host problem. Congestion is over utilization of various network resources, usually buffers, which may cause severe degradation in response time, throughput, and availability, and may possibly cause lock up. Internet congestion can arise within one of the component networks or at gateways. Network congestion can arise in any one of three places: 1) the source IMP; 2) the network; or 3) the destination IMP. Our goal in this report is to identify new feedback mechanisms or protocols whereby the subnetwork can provide the gateways and hosts information about the status of the network so that they can make their flow and routing decisions accordingly, thus reducing congestion in the network. Our goal was not to identify new protocols within the subnetwork, except to the extent of identifing information which could/should be gathered from the network in order for the gateways and hosts to make informed decisions.

Route selection, in particular gateway route selection is the key to making efficient use of the internetwork resources. However here we run into a basic problem of conflicting user

requirements. What criteria should be used to select the "best path"? Is it minimum delay, maximum throughput, global or network load sharing, fairness in terms of access to network resources, priority or other type of service requirements? The answer is yes! Different users have different requirements. Our emphasis in this task was how do we provide the gateways and hosts with a measure of what the network capabilities are at any instant, so that they can adjust their offered/potential load accordingly, and how do the hosts provide the networks information about their instant/future load characteristics.

3.2 Internet Route Selection

In this section we will discuss the issues of distributed control, route optimization, the information currently available from the net to the gateways and alternate sources of information which are/may be available.

Some of the fundamental problems with distributed control of the internet or network are the timeliness and accuracy of status information and normalization of the units of measure, e.g., distance metrics. Gateways in particular see a constantly changing view of both their attached networks and the offered load. There is significant feedback today to the gateways. The method of gaining this feedback is primarily by introducing more overhead onto the networks in the form of "pinging" or status

messages between gateways.

Pinging is a particularly annoying technique whereby gateways constantly send (at certain intervals) messages to other gateways to which it thinks it is attached to see if they are still up and who they are connected to. Hosts also ping gateways as a technique to determine their connectivity points to the outside internet. Pinging works okay on underutilized nets such as the early ARPANET with an excess of capacity. However as networks are maturing and being used more to their capacity for operationally critical traffic, alternatives to techniques such as pinging with its excessive overhead must be found. Pinging even became a problem for the underutilized ARPANET. However the solution (increase the time interval between pings and limit the number of gates/hosts pinged) was less than desireable. This solution reduced the timeliness and accuracy of the network status available to the gateway and was only a temporary patch since as the number of gateways and hosts continue to increase we will soon reach the point where even this restricted pinging is too much.

Therefore there is still a problem of developing accurate models at any particular instant in time, while keeping the network overhead to a minimum. Particularly critical is how or if it is necessary to have synchronized consistent models of the internet or network. Earlier experiments on the ARPANET

demonstrated that asynchronous models worked much better than expected with significantly less overhead burden placed on the IMP subnetwork. The granularity and significance of the measurement intervals is also a major factor. For example do you just report changes to previous conditions or do you periodically submit complete status updates between nodes. This must be balanced against the projected time period between reporting intervals to determine the accuarcy and timeliness of the information. Better yet alternative sources of information should be found which could reduce the overhead and provide more efficient control of network resources.

Optimization is the name of the game in any network design and in particular in selecting the best route for traffic. It is also critical to determining the most optimal internet route to assure not only the best service to the customers, but to also protect the internet and networks from unnecessary looping or congestion which suboptimal network routing could cause. For example a current popular topic in the internet community is the problems being experienced in passing traffic between the ARPANET and MILNET (particular large files) through the gateways. In certain instances hosts are overrunning the gateways by providing traffic to the gateway faster than the gateway can get rid of it. In these cases the gateway simply trashes the incoming traffic without telling anyone (after all it is not supposed to provide reliable service at the IP level) assuming recovery will occur at

the higher protocol levels. This causes excessive delays, and ultimately timeouts in the higher level protocols resulting in connection resetting and retransmission of the entire file, etc. Some of the suggested changes like increasing window sizes, or increasing gateway power/memory are self defeating. The main is that suboptimal subnetwork routing decisions are being made based only on subnetwork data instead of taking a more global view of the internet and determining the most optimal internet route.

Of course how we define optimality depends on which users you are talking to. The ARPANET was and still is optimized primarily for minimum delay. This meets the needs of the interactive community nicely but does not support the large file transfer type users very well. And recognize that it is the large file users who supply much of the justification and funding for the ultimate capacity of todays networks. These and other bulk users want maximum end to end throughput. Military command and control and intelligence users in addition want to assure survivability (via redundancy and network adaptibilty) and priority/preemption in times of crises to assure that they can get their critical traffic through when required.

Load leveling is an approach proposed by some network designers as a way to optimize the use of network resources. However how to "level" the load, how much capacity is held in reserve, and how to allocate that capacity, are problem areas

which must be resolved before such a strategy could be implemented for the internet.

Multipath routing is a relatively new approach to network design and its concepts are equally valid to the internetworking problem. Multipath routing attempts to level the load on a network by allocating the offered load from any source over several routes depending on the link capacities of the subnet. The multipath routing technique is far from perfected at this time and many critical issues are still undefined, however, early experimentation by BBN has shown some promising results. The original intent of the multipath concept was to maximize throughput (as opposed to the current SPF shortest path first algorithm which minizes delay). The proposed technique chooses the first (of multiple) paths between any two points based on the traditional SPF approach. However it also takes note of the capacity of this path to determine a match vis a vis the offered load from a host. The algorithm then proceeds to determine the next best path, via SPF, after having eliminated the link(s) which could be congested or a bottleneck for a high throughput user. This process repeated as necessary to acheive a certain throughput. The best part of this algorithm is that it introduces a measure of capacity into the routing algorithm which is not there today, yet at the same time continues to make its best path determination based on delay oriented SPF. In considering capacity in its routing algorithm new questions (as

yet unanswered) are raised like how does one measure capacity (instantaneous or averaged over some time period), how does one allocate the potential capacity among several gateways or hosts, how does the network assure fairness, i.e. fair access to all users of at least some portion of the network resources, and is the current SPF algorithm and granularity good enough to give an accurate measure of link or network delay (especially when used on an internetwork basis). Despite these unknown areas the proposed multipath routing approaches hold a lot of promise for improvements in efficiency of future network protocol implentations.

Type of service (TOS) routing has long been a recognized goal of network and internetwork designers. Precious little progress has been made in terms of ways to use TOS desighnations to affect routing decisions, however. The DoD standard IP has identified certain types of service designators for precedence, delay, throughput, and reliability. Thus the protocol available to the hosts and gateways is structured to support this concept. No real progress has been made beyond this point. One of the main problems is that the underlying subnetworks are not capable of supporting different types of service nor are they adaptive enough to be tuned for different usages even if they could understand TOS designations. At least most new protocols recognize the need for TOS use in the future and are reserving space in their protocol structures to accomodate it.

What is the information needed for route optimization? First a consistent measure of delay which is valid across the internet. In particular a way to measure gateway to gateway or host to gateway delay is required. Next a way to measure and predict the maximum end to end troughput is required. Such a measure should take into account the internet topology, the currently available subnet capacities, and the current or projected traffic load offered to the net or internet. A view of the intranet loading and internet flows is also important in order to assure load leveling and better efficiency within the internet. TOS matching between the projected host/gateway requests and the current network status is needed. And of course the current network and gateway status is needed to determine the ability of the network to adapt to changing conditions and offered loads.

There are several information sources potentially available to gateways from which they could glean more information. From the internet, gateways have access to IP (including ICMP), GGP, EGP, as well as communications to/from network (someday internetwork) monitoring and control stations. From their attached subnetworks gateways and hosts can aquire information via the host/network access protocols such as 1822 HDH, DDN X.25, etc. Routing information is also currently maintained by the gateways including; neighbor gateway connectivity and addresses, a minimum distance measure to the network and neighbor gateways, and a

routing table for directly connected networks. Ways to use the above protocols to improve the dissemination of information will be discussed in the next section. Some of the key features of each protocol applicable to this task are summarized below:

IP - IP is the primary source of information from a host to the gateway. Besides the obvious source/destination addresses, IP can provide TOS request to the gateway. IP can give the source host a role in routing via source routing parameters and could also aid in congestion control by implementing a better scheme for using the time to live field for a datagram.

ICMP - This is an optional protocol used at the internet level on top of IP. Although this protocol is not reliable it is a source of information from a gateway back to a host. ICMP messages could be used to have a host redirect messages along a better internet path, to "quench" or throttle the source host, and to provide other feedback information yet to be determined.

GGP - GGP is a protocol for neighboring gateways to pass information about their directly connected networks, updated network or internetwork routing tables, etc. GGP is also the protocol used in the notorious "pinging" approach to determine the status of neighboring gateways.

EGP - The exterior gateway protocol was designed to reduce the load on GGP and reduce the loading on the internet. Its primary purpose is to determine connectivity within the internet and facilitate fault isolation and recovery.

HAP/NAP - A variety of host and network access protocols are addressed in the section 5. These are the protocols for providing information to/from the directly connected network, i.e. IMP, and the host or gateway. These protocols and recommended modifications thereto are the ultimate purpose of this task, since these are the protocols through which any improved feedback will occur. Since DDN X.25 is the perferred HAP for the future DDN we have concentrated on this protocol.

3.3 Congestion and Flow Control

As mentioned earlier congestion and flow control are closely related. As used here congestion occurs within networks or the internet when its resources, usually buffers at either the IMPs or gateways, are overrun by receiving more data in than they can put out. Congestion control refers to the network's attempt to resolve such problems by adaptive routing or flow controlling the hosts. Thus flow control applies to hosts and their participative attempt to match their offered load and needed service to the current or projected network capabilities. By way of background, this section will introduce the problems of congestion, ways one can control congestion and how the network can assist.

Congestion ocntrol, like routing, is basically an optimization problem. the global view is to make the most efficient and effective usage of teh current internetwork capabilities by matching them with the offered load and requirements from the hosts. Compounding this problem is that what we really have is a distributed optimization problem with each gateway making decisions based on its own limited view of the world. Gateways have a limited ability to sense the global environment, and an even more limited ability to predict congestion and take steps to avoid it. Basically all we have today is very crude mechanisms which react to congestion after it

has occured and even at that these are rather drastic (such as source quench or just throwing datagrams away and assuming higher level protocols will be able to recover).

An optimal control algorithm depends on the type of congestion a gateway or network is experiencing. For example the gateway must determine if it is congested (i.e. throwing away traffic), if it is causing congestion at some other point in the internet by providing more traffic than that point can handle, or if some other gateway/host is congesting the resources that this gateway needs to deliver its traffic. In all of these cases the customer's traffic is not being delivered and even worse valuable network resources may be being used to continually (re)attempt to deliver traffic which is being discarded at a remote congested point in the internet, thus causing further congestion in other nets and gateways.

Therefore the first issue in congestion control is how to detect and predict the onset of congestion within the net or internet. The first step a gateway should take is to watch its own buffers, i.e. is it taking in more than it is putting out. Next it can monitor other gateways connected to the same nets, and theoretically it should be able to monitor its local IMP to see if it is causing the congestion. To predict congestion within the internet is much more difficult since it requires more monitoring of internet resource usage and a measure of capacity

along different internet paths. The gateway must then be able to anticipate overutilization either of itself or of specific paths.

Assuming you can detect or predict congestion, how then does a gateway respond? Currently the only options available to gateways are to reroute traffic from a host, throttle the source of the traffic, and/or throw the traffic away and rely on higher level recovery mechanisms within the hosts. Of course in todays environment it is not at all clear what another host or gateway should/will do with source quench messages nor how a gateway should decide which datagrams to discard. We will discuss and suggest alternative mechanisms to alleviate these problems in the next section, but note that any solution is going to require a measure of capacity and a method to allocate that capacity.

Current congestion control mechanisms are quite primitive, partly because of the lack of good descriptive information about the global state of the avaialability/utilization of network resources. Better information would permit hosts and gateways to alter their mix of traffic and routes to best match the available resources. There are some new approaches being researched and implemented which are a start to collecting and providing better network status information on which to make control and routing decisions. Already mentioned was the multipath routing algorithm which is a throughput oriented algorithm and collects data on the capacities within a net in addition to the delay measurements and

makes its choice of multiple routes optimally based on this data. This approach could be extended to the internet problem by having the gateways query their local attached net, gather this information, and relay the information to other attached gateways as part of their table updates. The basic philosophy of the multipath routing approach seems to apply to either the network or internetwork routing.

Another current development which we understand is going to be implemented throughout the DDN by 1986 is the new end to end (source IMP to destination IMP) protocol. This protocol in effect sets up a reliable connection between source and destination IMP and in a sense reserves a certain capacity allocation through the local net. This will be discussed later with recommendations about how it could be used to provide additional feedback, to help the internet congestion control and flow control problems.

Given that the internet congestion control problem occurs at either gateways or subnets, the following are some of our basic conclusions about what is needed to resolve these problems. First to help the gateways, it appears a new distributed gateway capacity control algorithm is needed. The current GGP and EGP are strictly delay oriented and assume there is an excess of capacity. Even with some of the new intranetwork protocol developments just mentioned, the gateways have no way to make use

of the additional information they could have access to. The current controls such as source quench are much too coarse. Additional research is needed into new congestion control algorithms for the internet, in particular to address throttle mechanisms, the exchange of information among gateways (and hosts), and a definition of the information collection requirements. This is beyond the scope of this report; however we do recommend the type of feedback the subnetworks can and should provide to the gateways to use as an input to such a control algorithm. Another interesting architectural concern is the question of fairness to/for gateways versus other hosts. Currently the subnetworks and IMPs treat all attached entities the same. Yet we suggest that gateways really are "more equal" than other hosts because of their role as a communication switch within the internet, and as such should get preferential treatment by the subnetworks in terms of capacity allocation, priority, etc.

Another premise we tried to follow was that gateways should attempt to poll the attached networks for status and not attempt to duplicate services by end to end polling. In many cases the subnetwork must already collect similiar data in order to accomplish its own routing and congestion control functions. The current pinging technique should be avoided except in special cases where practically nothing is known about the attached net or internet connectivity. The goal is to get better feedback

from the networks, reduce internet traffic to collect similiar statistics, and report exceptions or significant events.

In general the current approaches to (inter)network routing and control are reactive rather than predictive. The current control mechanisms are too drastic such as discarding packets or shutting off the source hosts/gateways. Some of the new techniques being investigated as part of multipath routing and implemented in the new IMP end to end protocol may help. The next section will recommend improvements which can be made to the DDN network protocols and ways to use some of the newer techniques to collect and distribute additional information to the hosts/gateways for better control.

4. Potential Improvements to Internet Control Mechanisms

This task was initiated under the assumption that the current internet control mechanisms can be improved upon if better network status information is; a) collected, and b) made use of by hosts and gateways. This section identifies potential ways in which the internet control mechanisms can be improved. In each area of improvement, the information requirements of the corresponding internet control mechanisms are defined. Next, the role that the DDN backbone network could play in providing the required information is examined and compared to alternate ways in which the same or similiar information could be obtained by the internet. For that information which could best be provided by the DDN backbone network, a recommendation is made for the feedback of this information to the host/gateway. After deciding what information should be fed back, section 5 then examines the DDN host access protocols to determine how best to get this information to the host and gateways.

The major theme of this report has been that the current internet control mechanisms are inadequate, and that better mechanisms can be defined. The current internet control mechanisms were developed to demonstrate that a high degree of interoperability among a wide variety of networks could be acheived. This interoperability goal has been successfully demonstrated, and the results of those research efforts are now

being applied to operational networks. However with the the evolution from research testbed to operational network, changes are inevitable. In an operational network the ability to support a much larger user community and remain cost effective is a paramount concern. Mechanisms are needed for the operational DDN backbone network to make better utilization of the existing (and future) transmission capacity. This contrasts with the mechanisms that were developed during the research phase in an environment of excess capacity.

A similiar situation is evolving for the internet. As increased usage is made of the internet for operational communications, the need arises to provide this communication capability in an efficient and more cost effective manner, i.e. a manner in which excess capacity is not a design assumption. The design rule should be to make better use of the capacity you have rather than to simply buy more capacity to compensate for inefficient network design.

There are several efforts underway to improve the control mechanisms within the DDN backbone, and make better use of the network resources (switches and communications lines). These improvements generally are aimed at optimizing the way network resources are used to handle the offered host to host load.

From the standpoint of the network, the offered load

presented by hosts represents a given, and the network control mechanisms are designed to handle this given load requirement in an optimal manner. The internet control mechanisms also get a fixed offered load howver it has the ability to distribute that load among it's different switches (i.e.gateways). What is desired is a way to perform global optimization of the internet and individual network mechanisms by providing an optimal offered load combined with network mechanisms that efficiently handle the offered load. Other studies are looking at how the subnet can best handle a given offered load. This report examines the larger question of how a host can determine the optimal load to give to the network.

The goal of the internet control mechanisms is to match the offered internet load (on a true end host to end host basis) with the internet resources in the most efficient and effective manner. The internet has three basic ways of varying its usage of the internet resources. It can distribute the load among component networks (varying the magnitude of the offered load into each network through gateways). It can distribute the load among hosts and gateways off the same network (varying the offered load in terms of network end to end flows). Thirdly, it can throttle the sources of data on a true end to end basis.

This view of the internet gives rise to the "internet as a network" model mentioned earlier. In this model, (see Figure

4.0) one can view the internet gateways as switches which are interconnected via virtual trunks (intranetwork end to end paths). The internet optimization problem then reduces to a more familiar network optimization problem, albeit a more complicated one in which neither the switches nor the trunks are homogeneous, and in which the data handling capacities of these switches and trunks can vary dynamically with considerable variance. However, these problems are becoming more common in the network domain (e.g. a variety of switches are used, a variety of media are used to interconnect switches, etc.) such that network solutions are being developed and can be readily adapted to the internet.

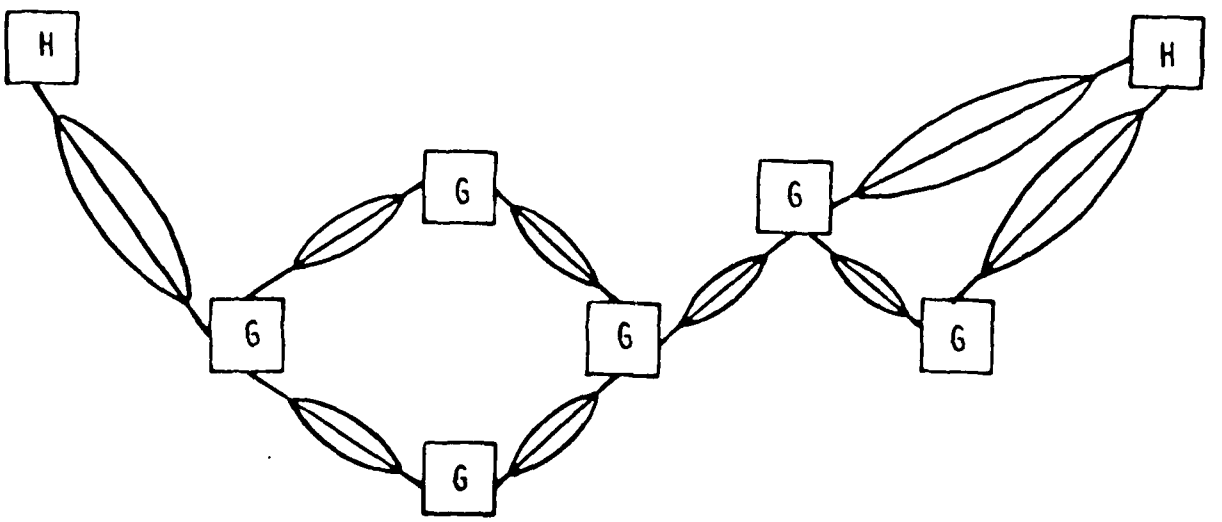


Figure 4.0 Internet Modeled as a Network

This section discusses four areas in which the existing internet control mechanisms can be improved upon; 1) type of service routing, 2) multipath routing, 3) internet flow control, and 4) internet congestion control. Note that in all cases analogies can be made to similiar mechanisms which are being developed for intranetwork optimization of resource utilization. In our view the network and internetwork control mechanisms are distinct but complementary: the internet mechanisms provide an optimal mix of traffic to the network(s), while the network mechanisms handle the host to host intranetwork traffic in an optimal fashion.

The following sections address how improved internet control algorithms should operate, and in particular what information these algorithms would require. The actual specification of these internet control algorithms is being addressed in other developments. We next looked at ways this information could be provided, either directly (by measurement) or by approximation (by measuring something that approximates that which we need, but are unable to directly measure). The quality of the control mechanisms is often a function of the fidelity, accuracy, and timeliness of the correlation between the unit actually measured to the unit it approximates.

With regard to the latter point, it is important to realize that a number of network and internetwork control mechanisms in

use today require information that cannot be readily or accurately measured. An example is the approximation of trans network delay by hop count. Today the delay between switches is approximated to be 1 hop. This works adequately as long as all trunks within a network are the same speed and capacity. However in the internet, delay and capacity between switches are highly variable and thus hops are not a good approximation.

Because the internet is far less homogeneous than most networks (e.g. inter-gateway delay is far more variable than network inter-switch delay), the accuracy of correlation is a bigger issue in the design of internet control algorithms.

4.1 *Type of Service Routing*

The current Internet Protocol contains provisions for requesting different types of network services based on application requirements. For example, interactive traffic desires low delay but generally does not involve much volume and therefore does not need much throughput. Bulk data transfer, on the other hand, can tolerate high delays on individual packets as long as an overall high effective data rate is maintained.

The IMP switching network currently has no provision for providing different types of service, although there is some thought being given to ways of doing this. Concurrently,

consideration is being given to the use of different media, e.g., satellite links, high speed terrestrial links, low speed telephone lines, etc., for inter-IMP trunks, such that there would be alternate end to end paths between hosts. As the DDN evolves and dissimiliar trunks are supported, the IMP could allocate traffic based on the type of transmission media available; thus interactive traffic could use low delay land lines, while bulk data traffic could be switched when necessary to high capacity satellite links.

The current task, network feedback to hosts, is oriented toward improving the internet control mechanisms. Currently, gateways have the option of routing data through the IMP network or through an alternate network, but they do not have the ability to specify type of service to the network. When type of service routing is supported by the network, the internet gateways will be modified to match IP type of service requests to network type of service offerings. This section examines how type of service routing could be performed by the internet, and what information would be required from the network.

4.1.1 Description of Gateway TOS Routing Mechanisms

Consider the following problem. A gateway recieves a datagram destined for some host with a request for low delay routing. How does it choose the next internet hop to route the

datagram along?

The gateway would need to maintain tables to describe the topology of the internet as well as the delay characteristics of each path between internet nodes, i.e.. true endpoint hosts and intermediate gateways. For the example problem there are two cases: the destination host is a member of the same net as the gateway; or the destination host can only be reached through other gateways. In the former case, the gateway need only specify the appropriate type of service and give the packet to the network for delivery.

In the latter case the gateway has to make a decision as to which next gateway to route the datagram to. To do this, the gateway would presumably calculate a delay metric for the paths through other gateways to the destination and to then select the next gateway along the shortest path. A possible way to calculate the path lengths is to have each gateway determine the path length from itself to each destination network by adding the path from itself to neighbor gateways and the path from the neighbor to the destination net. This process requires the determination of path lengths to neighbor gateways.

The path length (delay) to neighbor gateways is the delay along the network end to end path to the neighbor. This delay varies dynamically. There are two ways for a gateway to acquire

this information: direct measurement, by pinging neighbor gateways, or by peeking at the IMP's internal tables.

The IMP maintains tables that it uses for making its own routing decisions. The IMP's tables contain delay estimates to each destination network address, including neighbor gateways. This information is currently maintained in the form of IMP to IMP hops, which is a first order linear approximation to delay. When heterogeneous trunks are supported by the IMPs, the IMP tables will continue to contain some approximate representation of end to end delay, probably using weigthed hops. Thus the IMP currently has an approximate measure of end to end delay which could be made available to an IMP's local host upon request (*where "could be" means that the IMPs could be modified to* provide this information).

The gateway would then be able to obtain path lengths to neighbor gateways (after converting IMP hops to some standard delay metric such as milliseconds). With this added information from the network, the gateway could compute internet end to end delays and choose the appropriate end to end path.

A similar argument can be made for throughput oriented routing decisions. The gateways could compute end to end capacities to destination nets and choose the path that offers the greatest capacity. To do this computation, the gateways

would need to be able to determine the capacity of each network end to end path to neighbor gateways.

At the current time this can only be done rather crudely. The capacity to all neighbor gateways has an identical maximum upper bound of the constant maximum throughput through the net. This constant can be compared to similar constants for other networks to distinguish among paths that go through different networks; no information is available to distinguish among alternate network end to end paths to neighbor gateways connected to the same net. Direct measurement would be required to estimate alternate path capacities.

When the new end to end protocol is implemented, however, there will be additional information maintained by the IMP which can be used to estimate end to end path capacities. Because of the way the new end to end protocol implements flow control for network connections, the average window size can be combined with the end to end delay estimates to compute an estimate of end to end capacity. An algorithm could be developed to combine the gateway to neighbor gateway path capacities with neighbor gateway to destination net capacities to yield an estimate of gateway to destination net capacity for each path through neighbor gateways. Thus type of service routing for high throughput services can be implemented.

Given the basic parameters to work from, other forms of type of service, e.g., priority, security level, etc. could be supported. For instance, the fastest, least delay, path at a given priority level could be computed.

4.1.2 Information Required from the Network

In order to distinguish among paths through a network to alternate neighbor gateways, a gateway needs to have some information that describes the characteristics of the paths to the neighbors.

The delay characteristics are currently available within IMPs and could be made available to hosts. The host (or gateway) may need to convert network units (i.e., hops) to a standard internet unit (e.g., milliseconds) to avoid comparing apples and oranges.

The capacity characteristics are not currently available, but could be derived when the new end to end protocol is implemented. Network end to end path capacity equals the window size divided by the window advancement delay (e.g., round trip delay). Again, some normalization may be needed.

Both of these parameters could be provided by the network to a locally attached host by extension of the network access protocol. Section 5 will discuss such extensions.

4.2 Multipath Routing

Multipath routing techniques can be used to provide increased throughput by concurrently using parallel paths from source to destination. Multipath seeks to optimally use available capacity by spreading the traffic load across all available resources. This is in contrast to type of service routing which selectively routes traffic over matched routes rather than sharing the load among multiple routes in multipath.

In a network context, multipath attempts to use parallel paths through different IMPs. For instance, if there are two parallel paths, through IMP2 and IMP3, respectively, between IMP1 and IMP4, IMP1 may choose to route half of the IMP1 to IMP4 traffic through IMP2 and the other half through IMP3. If the two parallel paths can be used concurrently, the queueing delays at IMP2 and IMP3 occur in parallel, for an effective one half reduction in the queueing delay. Further, if the available capacity from IMP1 to IMP4 is C1 through IMP2, and C2 over the path through IMP3, a throughput of C1+C2 can be attained from IMP1 to IMP4.

The multipath concept can be carried over to the internet case fairly easily. Consider the internet gateways as being analogous to network switches which are connected by virtual

trunks (end to end network paths). Each gateway determines the capacity of each of its trunks to neighbor gateways, and then uses some algorithm to apportion data flows over paths through its neighbors.

4.2.1 Multipath Routing Algorithms

A variety of algorithms are being investigated for apportioning traffic loads across a set of parallel communications paths. Generally, an iterative approach is used whereby the unloaded capacity of each path is determined, with the capacity decremented as traffic is sent. The updated topological representation is reexamined, a path is selected, and the path capacity is again decremented; this process is repeated periodically. Variations on this theme involve how paths are selected (e.g., shortest path, highest capacity path, etc.) and how frequently the topological representation is updated.

4.2.2 Information Required from the Network

All of the capacity allocation algorithms start with a topological representation of the network which contains capacity information about each link between nodes. As mentioned earlier it is important to keep this information current. In the internet case, the link capacity between internet nodes is the network end to end capacity between these nodes (hosts).

The end to end network capacity between hosts can be detemined from window sizes and delays. In the existing DDN, this would be the eight packet window size and the neasured delay between RFNMs. In the new end to end protocol, the window size is much broader ranged and more dynamic; the end to end window parameters are internal to the IMP. Capacity can be estimated from the average window size and the end to end delay; both statistics should be available within the IMP and could be made available to a locally connected host.

4.3 Internet Flow Control

The purpose of flow control in the (inter)network is to control the rate of information flow across the network. One of the goals of flow control is to match the rate of the sender to that of the receiver. This prevents the internet from becoming a virtual buffer between the source and destination, which would reduce the resources available to other hosts in the network.

Another goal in the use of flow control mechanisms is to allocate network capacity fairly among network subscribers. Such allocation is useful when excess capacity cannot be assumed and where sources must be throttled.

Therefore we view the internetwork as a network of switches

(gateways) which are interconnected by variable capacity trunks (intranetwork end to end paths). Because excess capacity cannot be assured, mechanisms are desired for allocating end to end subscriber traffic among the internet resources, and for throttling end point hosts fairly when demand exceeds capacity.

Of course the concept of fairness means different things to different users. In the current DDN, hosts and gateways compete equally for network services, while the penalty for reduced service is not equal. That is, if a gateway gets blocked, the traffic then gets backed up into the previous network(s) causing multiple retransmissions in the network(s). This will affect multiple hosts and networks whereas, if a host gets blocked only that host is affected. Generally a gateway is supporting multiple hosts (who in turn may have multiple connections). Some supported hosts will be from the local network and some will be using this gateway as a transit media to other nets. If a host/gateway sends packets into a net and they get blocked, then they will attempt to retransmit. If they still cannot get through or the process takes too long, then the source host will time out and retransmit the entire datagram or TCP message. This will occur for all hosts blocked directly or indirectly by that one gateway. Thus blocking a gateway can have much more severe consequences than blocking an individual host.

Of course this is not the only factor to consider, since an

implication of this philosophy is to give priority to the through traffic (gateway to gateway). Although in most cases this is peferred, in extreme situations this could result in local traffic being delayed in deferrence to transit traffic from other nets, and at some point the local populace will legitimately complain about unfair access to their local net. Therefore the philosophy of giving priority to gateway to gateway traffic may have to be moderated in practice by limiting gateway traffic to a certain percentage of the of the available network resources.

Thus if the internet is viewed as a two tiered switching network, a gateway performs a switching function which should be recognized as such by the network. From the standpoint of internetwork efficiency, some improvement in internetwork utilization could be achieved by having the network offer better service to internet switching components (i.e. gateways) than to hosts.

What is really needed is a global picture of internet resource requirements. A logical first step is for the network switch to allocate capacity among it's hosts in proportion to the recent history of each host's offered load. However it must also consider the impact of delaying/discarding traffic (i.e. how much additional traffic will result from retransmissions). The problem is how to measure this at the internet level such that parameters can be defined for an internet optimization algorithm.

With some assumptions, a first order approximation can be made. First, assuming that a host's retransmission queue is proportionate in length to the rate at which it introduces IP datagrams into the internet, and that this proportionality carries over to the host's (true host or gateway) introduction of packets into the network, the magnitude of packet retransmission that would result from throttling a source can be viewed as proportional to the offered load presented by that host to the network. Thus if a host's recent offered load has been lower than it's potential capacity, one would expect a small number of retransmissions if that host was throttled. Whereas if another host has recently been attempting to send at it's maximum allocated capacity, one would expect a relatively large number of retransmissions (and thus significant network impact) if that host was throttled.

However there is a second order effect which is caused by the spreading of congestion. If throttling of a gateway causes not only the delay of some datagrams through the gateway (with subsequent retransmission by the source), but also in the congestion of other internet resources (i.e. other gateways), then other traffic through the internet can be delayed. For example if a network attempts flow control on a gateway to gateway (end to end) basis, throttling the connection between gateway A and gateway B could congest gateway A (the source in this case), with the result of impacting the traffic between

host/gateway C and gateway A.

The conclusion is that the network needs to distinguish between hosts and gateways, and to favor gateways in allocating resources. Gateways should also monitor the capacities of it's local network connections and factor this information into an internet capacity allocation algorithm. In a similiar vein gateways need to distinguish between hosts and gateways. Gateway resources should be allocated to its offered load in relation to the expected number of retransmissions. Gateway to gateway (internet trunking) traffic should be favored over host to gateway (internet access line) traffic in order to minimize congestion.

All of this suggests an internet flow control strategy that allocates resources according to the expected retransmission impact (i.e. throttle those ends which when throttled will cause the least retransmissions). This in turn suggests that internet gateways keep track of two things: capacities of internet trunks and access lines, and the recent history of the offered load for each trunk and access line. The latter can be measured directly by the gateways. The former corresponds to the network end to end capacities which should correspond closely to the window size associated with each network end to end connection. Thus this information could be obtained from the network.

4.3.1 Anticipated Flow Control Mechanisms

Currently there are there are very limited and mostly ineffectual flow control mechanisms at the internet layer. Thus congestion of internet resources is a frequent problem. The new network end to end protocol will provide flow control on a host/gateway to host/gateway basis. From the internet view, flow control will be provided independently for each inter-gateway "trunk" and each host to gateway "access line".

The maximum data flow rate between a pair of end to end hosts is limited to the slowest internet trunk/access line in the path between the hosts, or in multipath routing the limit is a function of the aggregate capacity of the parallel paths. If the source host were to provide data addressed to the destination host at a rate greater than this maximum, the internet would not be able to keep up and the result would be congestion of network resources. It is therefore desireable to restrict the source-destination data flow to the available internet capacity.

There are a several ways of providing flow control. One way is to define a windowing scheme on a true end to end basis, with dynamic computation of the available capacity of the end to end path. Another way is to make the source host responsible for

flow control, tell the host what the available end to end capacity is, and let the host allocate the capacity as it chooses among outgoing traffic. This further suggests independent retransmit timeout values for each destination, where the timer does not start running until the datagram is actually sent, as opposed to queued for sending, to the network. The host algorithm would have to keep track of datagram inter-transmittal times to maintain the end to end flow within limits.

In any internet flow control scheme, what is needed is a good picture of the capacity of each of the internet trunks and access lines. These capacities are the network end to end path capacities between hosts and gateways (internet access lines) and among gateways (internet trunks).

The new end to end network protocol provides flow control on a host to host basis through use of a window mechanism. The IMP computes a window size for each connection based on buffer availability and negotiated load requirements. Given the window size and path delay, the path capacity available to the host can be calculated. The window size and path delay information is available within the IMP. It is proposed that this information be made available from the IMP to the network hosts.

The network hosts (true hosts and gateways) would then use this information to compute end to end internet path capacities,

with mechanisms provided to limit the rate at which source hosts originate traffic destined for each destination host.

4.3.2 Information Required from the Net

The new end to end network protocol uses a windowing scheme to limit the number of outstanding packets between hosts. Given the window size W, and the end to end delay, D, before the W+1 packet can be sent, the effective capacity is W/D.

Provisions should be made for providing a network host with the W and D parameters for selected trans-network paths (i.e., for those paths that lead to gateways). These parameters provide the basic information needed to compute path capacities, which could form the basis for the computation of internet path capacities. Flow control mechanisms can then be defined which restrict a host's offered load in accordance with the available capacities.

4.4 Internet Congestion Control

The previous section addressed the topic of congestion avoidance through use of flow control mechanisms to limit the traffic load introduced into the internet by a host. These flow control mechanisms should go a long way towards reducing the internet congestion that is experienced today.

It should be noted, however, that congestion will still occur, and must be taken into account in the design of internet control mechanisms. This section examines current internet congestion control mechanisms and postulates potential improvements. It then identifies the information needed to implement the improved mechanisms.

4.4.1 Internet Congestion Control Mechanisms

Congestion is a common occurance in the internet today much of it due to ineffectual end to end flow control. To date, most work in internet congestion control has concentrated on how to use the ICMP Source Quench messages, and the *proper response* by a host when it receives a Source Quench.

There are major problems in controlling congestion today because the existing mechanisms come too late (i.e. they react rather than predict), result in too extensive a reduction in traffic, and are too indiscriminate with respect to the optimal party to select for throttling. Part of the problem stems from the fact that ICMP treats congestion as a binary state: a gateway is either fully available or fully congested. This in turn leads to gateways being hesitant to announce congestion until it sets in, by which time the effects are spreading rapidily throughout the network and internetwork.

Part of the difficulty in addressing congestion comes from the fact that the effects of a congested resource spill over to adjacent resources. Because it is difficult for individual gateways to determine the cause of congestion, it is difficult to know what steps to take to alleviate congestion. Thus the common approach is to throw traffic away indiscriminately and to tell everyone to stop (or slow to a crawl) until things get cleaned out.

Because the effects of Source Quench are so extreme, an alternate graduated response mechanism is desired to gracefully slow down the source hosts. That is the basis for the internet flow control mechanisms discussed above.

Given that flow control will not always prevent congestion, the question arises as to whether there's a better way to respond to congestion than discarding randomly selected packets. And, if there is a better strategy, could the network provide any information which could improve the gateway's response?

Some objectives of a congestion control strategy could be to localize the congestion (keep it from spreading), to select the packets to be discarded in such a way so as to minimize the number of subscribers affected (e.g., discarding eight packets from one TCP connection is better than discarding one each from

eight connections), and to select packets in a way that minimizes the number of retransmitted packets (e.g., discarding a packet that ACKs eight other packets will cause all eight to be retransmitted). The latter two goals can be accomplished if desired with information in the gateway; no additional information is required from the net.

With the new network end to end protocol, the most probable cause of congestion will be mismatches in available channel capacities into/out of the two (or more) networks due to different end to end window sizes in the two nets, i.e., the gateway thinks it has a 50 Kbps channel out but is only able to get 10 Kbps out. As gateway buffers fill up, it will have to either discard traffic of refuse to accept new incoming packets, which will in turn congest the previous node.

If the gateway can predict the beginning of congestion, it could take steps early on to avoid it by using Redirect, Source Quench, or a new (to be determined) mechanism that attempts a graduated (as opposed to all or nothing) source quench. The gateway currently can monitor buffer queue lengths which does reflect outgoing channel capacity, albeit with some time lag. The question is whether a better predictor can be obtained from information that could be provided from the net.

In order to predict the onset of congestion, it is necessary

to consider the rate of data flow into and out of the gateway, and keep up with the changes in the rate at which packets come in and can be sent out. The current strategy of sending a source quench when packet buffers exceed a threshold might be improved upon by instead using the first derivative of queue size with respect to time, dQ/dt.

The rate of growth of a queue is the difference of the input and output rates (e.g. 30 packets/second in, 20 packets/second out, means the queue grows at 10 packets/seconds). The queue length is the integral of the queue growth. Thresholding the queue length provides an indication that a problem exists, but does not reflect the magnitude of the problem, and therefore cannot be used to provide a graduated response.

An alternative approach is to monitor the rate of change in queue length (dQ/dt) and perhaps even the acceleration of this rate (d^2Q/dt^2). This would allow us to determine the magnitude of the problem (e.g. queues growing at 1 packet/second can be addressed differently than queues growing at 10 packets/second), and would also allow the prediction of future problems (e.g. if the queue keeps growing at this rate, we need to use a mechanism that reacts within n seconds or we will be congested). Monitoring the rate of queue growth would allow various gradations in the response mechanisms to match the severity of the predicted problem.

Thus for the purpose of this task, let us assume that someone wants to develop an adaptive predictive congestion control algorthim that will attempt to predict channel capacities, and use these predictions to trigger some response mechanism. Now let us examine the information needs of such an algorithm.

4.4.2 Information Required From the Net

In the previous section it was noted that the first and second derivatives of queue lengths over time could be used to predict congestion. These derivative parameters reflect the rate, and change in rate, of observed (recent history) packet arrival and transmission. It would be interesting to also consider a predictive scheme, i.e., to be able to forecast future rates, such that appropriate action can be taken in time. A predictive scheme would require that future channel capacities be estimated, such that future packet transmission rates can be estimated. The question is, what is the best parameter to use in estimating future channel capacity.

There are three candidate parameters: rate of queue length change; rate of change in window size between host and IMP, i.e. X.25; and, rate of change of window size between source IMP and Destination IMP, i.e., network end to end protocol. The first two are currently available to the host (gateway); the third is

available within the IMP and could be made available to the host. It is suggested that this would be a useful topic for future study: determining if a predictive algorithm could be developed that would provide better results than simple thresholding.

At this point we recommend that the IMP be changed to provide the window statistics to the host on request such that studies could be performed to validate the concept of predictive congestion control. If it turns out that a predictive algorithm results in significantly improved performance, the IMP-provided statistics could then be used operationally.

5. Host Access Protocols

Host Access Protocols (HAPs) are the part of the network layer of the DoD Protocol Reference Model which provide data transfer across the host to IMP interface. These protocols provide the host to network and network to host status and control information. If any information is to be gathered by the host from the network this is the layer at which it will be relayed. We will look at the DDN standard HAPs and suggest how delay and window size information may be passed from the network to the host.

Within the DDN the method of access is via either of two HAPs, the ARPANET 1822 protocol, or DDN X.25. Currently DDN supports both protocols as coequals, but will phase out 1822 in the future. DDN X.25 provides two types of service, basic and standard. Basic service is equivalent to CCITT recommendation X.25 and FIPS-100/FS-1041, basic service is oriented towards hosts that have existing higher level protocol implementations, e.g., SNA, that require reliable packet delivery. Standard service is oriented towards hosts using DoD standard TCP/IP higher level protocols. The DDN will, in the near future, provide interoperability between 1822 and X.25 for users of DDN Standard X.25 service. Currently only DDN Basic X.25 service and 1822 are supported in the DDN.

5.1 1822

The 1822 protocol is a local host to IMP protocol, that is, 1822 messages have significance only over the local host to IMP interface. Flow control over this interface is performed using host interface blocking and RFNMs. The flow control window size currently is eight, if a host attempts to send a ninth message, with eight currently outstanding, the IMP will block the host until a RFNM is received. Although RFNMs carry the message ID of the message being acknowledged, the IMP only counts RFNMs and will continue to accept message as long as the number of messages outstanding is less than eight. If a message is lost an "Incomplete Transmission" message is returned in place of a RFNM.

With the adoption of the new IMP end to end protocol the "message in flight" limit of eight messages per host pair will be changed to a one to 128 messages in flight per connection per host pair limit which is negotiated between IMPs during connection establishment. Since an 1822 connection between a source host and a destination host is defined by the destination IMP, host and the handling type of a regular message, the most messages in flight possible from 1822 hosts will be 128.

As mentioned in Section 4, we believe that the network should provide to the hosts information on network window sizes, and network delay. To relay this information to an 1822 host both

items must be sent in an 1822 message from IMP to host. Appropriate changes to the 1822 protocol messages could be made consistent with our recommendations for DDN Standard X.25 service if window and delay information relay is desired for 1822 hosts. However, based on the DDN goal of converting to X.25 as the host access protocol for the DDN, we have not pursued specific recommendations for the implementation of these new message types in 1822.

5.2 X.25

DDN X.25 is the interface between hosts (or gateways) and IMPs for hosts operating in the packet mode on the DDN. The X.25 interface consists of three parts: the physical level interface; the link level interface; and, the packet level interface. In this section we describe the packet level interfaces of DDN Basic and DDN Standard X.25 service.

5.2.1 DDN Basic X.25 Service

DDN Basic X.25 service provides communication only between an X.25 host and other X.25 hosts implementing compatible higher level protocols. Basic service subscribers may not communicate with hosts which are not on the same subnet within the DDN internet because internet communication is performed via the Internet Protocol (IP). If a host wishes to use IP (and TCP)

then it should connect to the local DDN subnet via standard X.25.

Basic service provides end to end call management with significance as described in CCITT Recommendation X.25 and FIPS-100/FS-1041. Currently Basic service users may negotiate flow control windows of up to seven outstanding messages (X.25 packets), per connection. With the current IMP subnet protocol this negotiation is of little real value because the IMP does not take part in the negotiation and will block the host once it has eight packets unacknowledged on all connections to the distant host since the IMPs will allow only eight packets outstanding between host pair.

5.2.2 DDN Standard X.25 Service

DDN Standard X.25 service provides interoperable communication between an X.25 host and other DDN hosts, either 1822 or X.25 hosts, when used in conjunction with DoD standard protocols. Standard service subscribers which implement IP are full members of the internet unlike basic service users who may communicate only with in their local DDN subnet.

In order to intercommunicate with 1822 hosts, standard service, currently, has only local significance. Data from the local host is transmitted to the IMP using X.25 standard service where the IMP terminates the X.25 message. The information is

then carried through the network to its destination IMP via internal subnet protocols (see Section 2.3.3), where it is delivered to the destination host using the local access protocol. There is no X.25 level reliability of acknowledgment or delivery with standard service; reliability is provided by the DoD higher level protocols.

With the introduction of the new IMP end to end protocol, end to end significance will be supported for all connections, local significance may be added later. The new IMP end to end protocol will be able to distinguish between connections on a host and will account for outstanding messages on a per connection basis. The IMPs will become full partners in the end to end flow control window size negotiations and will be able to provide window sizes up to the X.25 maximum of 127.

In Section 4 we determined that the network should be able to make available to its attached hosts information on the network flow control window size and the delay associated with each host to host path through the network. As shown above, the new IMP end to end protocol will allow X.25 hosts to determine flow control window sizes simply by electing to negotiate flow control window sizes, this assumes that X.25 windows and EE windows are the same.

Delay information resides in the IMP in the form of distances

to destination IMPs. To get this information to a host we suggest that new DTE to DCE and DCE to DTE packet types be added to DDN Standard X.25 service. These packet types, STATUS and STATUS INDICATION respectively, which will have local significance only, are to be used to pass network status information to the DTE on request. Facilities should be provided to relay the window size and delay information as described below, other facilities may be added latter to provide information on type of service, congestion, or other network characteristics of interest to hosts or gateways.

The STATUS packet type will be patterned after the CALL REQUEST. There should be facilities provided to pass the two types of status information which we have identified in this report, window size and delay, for either a host to host path, or an existing virtual circuit or permanent virtual circuit. When host to host information is desired, the logical channel number shall be set to zero, and the appropriate facility specified, either window size or delay. Precedence may be specified, using the CALL REQUEST Precedence facility, in the STATUS packet type for host to host information. When information is desired about an existing virtual circuit or permanent virtual circuit, the logical channel number shall be specified and the calling and called DTE address shall be set to zero, and the appropriate facility specified.

The STATUS INDICATION packet type will be patterned after the CALL CONNECTED. When host to host window size information is requested the DCE shall calculate available window size in the same manner as it would if call setup were taking place and return the value in the Window Information facility of the STATUS INDICATION packet. When host to host delay information is requested the DCE shall examine its current routing table and return the value in the Delay Information facility. When precedence is specified in the STATUS packet type for host to host information, the window size calculation shall be performed assuming that lower precedence connections may be cleared, and delay information shall reflect any expeditious routing which may take place based on precedence. Information relayed in response to a request for virtual circuit or permanent virtual circuit information shall be gathered from the current state of the connection, no new calculation shall take place.

6. Conclusions/Recommendations

Throughout this report we have identified several approaches and techniques which could improve the efficiency and effectiveness of the DoD internetwork and the DDN specifically. Our methodlogy was to first look at the internet, identify concepts for improving its performance, and look at the information requirements those concepts would place on the component networks and resources. The end result was to specify the feedback requirements from the DDN network (IMPs), to the gateways and hosts, which would give them the raw data necessary for use in the generic algorithms in improve the network and internetwork performance.

The main information required by the hosts is readily available within the IMP, or at least will be available as the new IMP end to end protocol is introduced later this year. They are; 1) to give the gateways a measure of delay across the network to any other gateway or host; and 2) to provide an estimate of the source to destination data handling capacity.

The delay today is measured in hops, but in the future this should become more granular and accurate as different speed trunks are added to the net. The only measure currently available at the subnet level to allow us to approximate the capacity through the network is the flow control windows in the

new end to end protocol. The window sizes are changed dynamically depending on loading conditions and along with delay measurements should allow one to calculate an approximate capacity across the DDN. The host access protocol should be modified to allow the host to request and gather updated status information as necessary from the local IMP.

With good estimates of end to end delay and capacity, new mechanisms can be defined for potentially improving the efficiency and capacity of the internetwork. For example we could reduce the need for pinging among gateways, new algorithms are possible for predicting congestion, and better flow control techniques become viable.

While the complete definition of a new internet control mechanism was beyond the scope of this study, it is clear that better control mechanisms could be developed if the proper delay/capacity parameters can be obtained. In order to develop new control algorithms, delay/capacity statistics need to be collected and analyzed. Analytical studies could then be used to evaluate alternate control algorithms and these studies could later be validated by actual experimentation. Network feedback mechanisms would be useful to both collect data for analysis as well as to support experiments. After the conclusion of these experiments, the network feedback mechanisms may also be used by operational hosts/gateways to obtain the network delay/capacity

parameters needed for optimal route selection, flow control, congestion control, etc.

With better and more timely feedback, new and improved control techniques become viable. It does not appear that the impact of providing this feedback will be a burden to the IMPs and could be phased in with the on-going upgrades to the subnet. The hosts are just now beginning to implement the new standard X.25. Therefore now is an appropriate time to include the mechanisms for gathering this information from the network, even though the final implementation of improved control and routing algorithms may be further away.

END

FILMED

10-85

DTIC