# Proceedings of the 8-9 April 1986 DARPA Internet Engineering Task Force

Prepared by:
Phillip Gross

SECOND IETF

The MITRE Corporation
MITRE $C^3I$ Division
Washington $C^3I$ Operations
1820 Dolley Madison Boulevard
McLean, Virginia 22102

# TABLE OF CONTENTS

# APPENDIX A

## Hardcopy of Presentation Slides

| *Author* | *Title* |
|---|---|
| R. Hinden | Gateway Status Report |
| M. Brescia | GGP Software Bug |
| H. W. Braun | Overview Of NSFnet |
| P. Gross | Recent LSI/Mail Bridge Gateway Performance |
| M. Gardner | Mail Bridge Traffic and Performance |
| M. St. Johns | Proposed Amendment To RFC 904, EGP Specification |
| N. Chiappa | New ICMP Messages and Host Attachment |
| Z. Zhang | IP Congestion Control Proposal |
| E. Cain | Recent Internet Pathology |
| M. Tasman | On The IP TTL |

# APPENDIX B

## Additional Papers Distributed At The Meeting

| *Distributed By:* | *Paper* |
|---|---|
| R. Hinden | *The Internet Through The Ages* |
| D. Mills | *Requirements For NSF Gateways* |
| N. Chiappa | *Interconnection Of A Host And The Internet* |
| H. W. Braun | *NSFnet Briefing Slides* |

# Meeting Notes For Internet Engineering Task Force

# Internet Engineering Task Force

8-9 April 1986

Prepared by

Phill Gross
MITRE Corp.

# Table of Contents

## 1. Introduction

The first full meeting of the DARPA Internet Engineering Task Force was held Tuesday and Wednesday, 8-9 April 1986, at the Ballistics Research Laboratory in Aberdeen, Maryland. The meeting was hosted by Ron Natalie.

The notes for this meeting will be distributed initially by electronic mail and then in "Proceedings" format with the presentation slides.

A grateful acknowledgement goes to Pat Keryeski for editing the meeting notes and compiling the hardcopy Proceedings.

## 2. Attendees

| Name | Organization | Net Address |
|---|---|---|
| John Anderson | DCEC | janderso@ddn2 |
| Hans-Werner Braun | U of Mich | hwb@gw.umich.edu |
| Mike Brescia | BBNCC | brescia@bbnccv |
| Noel Chiappa | Proteon,MIT | jnc@proteon.com |
| Mike Corrigan | OSD | corrigan@sri-nic |
| Marianne Gardner | BBNCC | mgardner@bbncc5 |
| Phill Gross | MITRE | gross@mitre |
| Ken Harrenstien | SRI | klh@sri-nic |
| Robert Hinden | BBNCC | hinden@bbnccv |
| Steve Holmgren | CMC | sfh@edn-unix |
| Mike Karels | UCBerkeley | karels@berkeley.edu |
| Bob Knight | SRI | knight@sri-nic |
| David Mills | Linkabit | mills@isid.arpa |
| Ron Natalie | BRL | Ron@brl |
| Carl-Herbert Rokitanski | DFVLR | roki@isid |
| Mike St. Johns | DCA/B612 | stjohns@sri-nic |
| Zaw-Sing Su | SRI | zsu@sri-tsc |
| Mitchell Tasman | BBNCC | mtasman@bbncct |
| Shukri Wakid | NBS | wakid@nbs-vms |
| Stephen Wolff | BRL | steve@brl |
| Lixia Zhang | MIT-LCS | lixia@xx.mit.edu |

## 3. Agenda

(As distributed prior to the meeting)

**Tuesday, 8 April**

Morning

- o Introduction, Charter/Goals (Corrigan)
- o Gateway Status Report (Hinden)
- o Plans to Add New Host Communities (Braun)

Break

- o Recent Internet Performance
  - Case Study 1 (Gross)
  - Case Study 2 (Cain)
  - Open Discussion (led by Cain)
  - (Presentations of other findings, particularly by BBN and BRL, are encouraged)

Afternoon

- o Requirements for Internet Gateways (Mills)
- o Open Discussion of Gateway Requirements for Improved IP Performance (led by Mills)

Break

- o EGP Background (Mills)
- o Proposed EGP Modifications (St. Johns)
- o Open Discussion (led by St. Johns)

**Wednesday, 9 April**

Morning

- o Host IP Requirements (Chiappa)
  - Routing
  - Congestion Avoidance
- o Open Discussion (led by Chiappa)

Afternoon

- o Open Discussion: Revised Areas of Concern, next Agenda (Corrigan)

## 4. Relevant Documents

The following RFC's are available via anonymous FTP from the <RFC> directory at SRI-NIC.ARPA.

RFC896 - "Congestion Control In IP/TCP Internetworks", Nagle.

RFC950 - "Internet Standard Subnetting", Mogul.

RFC970 - "On Packet Switches With Infinite Storage", Nagle.

RFC975 - "Autonomous Confederations", Mills.

The following three papers have been provided to members separately:

"Internet Engineering Task Force Agenda and Meeting Notes", Gross.

"Requirements For Internet Gateways", Mills.

"Interconnection Of A Host With The Internet", Chiappa.

The following article provides background on several efforts that promise to rapidly increase the size of the Internet.

"Computer Networking for Scientists",
D. M. Jennings, L. H. Landweber, I. H. Fuchs, D. J. Farber, W. R. Adrion,
Science, 28 February 1986, pp. 943-950.

## 5. Meeting Notes

### 5.1 April 8, 1986

Hinden began the presentations with a gateway status report, including a Butterfly deployment schedule. He reported that there are currently ~130 active networks and ~85 gateways, passing 160M packets per week. (Note: Presumably this is an estimate for all gateways, since the Gateway Throughput Report for that week gave a total of 106M packets per week.) He gave the development plan for the LSI gateways as follows:

Current Release #1007, handles 120 networks.

Release #1008, 11/23's with memory management, will handle up to 150 networks, available by the end of April.

Planned final LSI Release #1008.1, will handle 300 networks.

He also reported on Butterfly status. Eight have been currently deployed, with twenty-one more to be installed on Satnet, Suran or PDN by Fall '86. Work has started on the "Mail Bridge" Butterfly, which will include EGP Access Control in addition to the normal mail bridge functionality. The thought of EGP "access control" gave several members heartburn and Corrigan suggested off-line clarification.

Brescia reported on a serious software bug in the LSI routing code that led to routing loops over the past few weeks. The feeling was that this led to the widely reported increase in traffic and decrease in performance during this period.

Mills voiced major concern. He felt that 1) the problem should have been diagnosed more readily and 2) once diagnosed, should have been corrected more quickly. All in all, he felt this was a strong indication of the rickety state of the Internet. On the second point, it was explained that there is inertia in reloading the gateways, since they are not all under the direct administration of BBN.

With this as an introduction, Braun presented an overview of NSFnet; an effort, he stressed, that will further tax an already overburdened Internet. In its initial phase, NSFnet will utilize existing networks (Arpanet, CSnet, BITnet, campus networks) to provide supercomputer access. This will include a significant increase in Arpanet sites and nodes. Future developments will include very high speed links and migration to ISO protocols. He also gave details of several ongoing pilot projects. (Note: Additional details can be found in the Science article cited above.)

Gross then presented a perspective on traffic and performance in the LSI gateway system. Using information processed from the weekly Gateway Throughput Reports, he produced graphs of traffic sent and traffic dropped in both the Mail Bridges and the entire LSI system. The graphs showed sharp and consistent increases, which began in December '85. In summary:

| | Dec 85 | Apr 86 |
|---|---|---|
| Traffic Sent by Mail Bridges | ~27M | ~35M (Packets/week) |
| Traffic Sent by LSI System | ~90M | ~105M (Packets/week) |
| Traffic Dropped by Mail Bridges | ~3% | ~6% |
| Traffic Dropped by LSI System | ~2% | ~4% |

He further estimated that the portion of the traffic that represented successful user data had dropped from 55% to around 45%.

Gardner followed this with an interesting presentation of more detailed Mail Bridge data and gave an additional cause for poor Mail Bridge performance. She pointed out that:

- Each host-to-host connection requires a Connection Block (CB) in the source and destination PSNs.

- Gateway pings occupy a CB.

- Mail Bridge PSNs are short of CBs.

- When no CB is available, the oldest in use is torn down. When the host is actually a gateway, this is a suboptimal policy, since the oldest connection is probably just getting ready to re-ping.

An upgrade, which was due to be complete by the end of April, from PSN 3/4 to PSN 5, will increase the CBs from 73 to 255. She felt that this will greatly increase Mail Bridge performance.

After lunch, Mills discussed his NSFnet Gateway Requirements document. One goal of this paper is to bring together all RFC references that specify gateway IP requirements. It will be distributed as an RFC and used for NSF gateway procurement.

In brief, NSF needs packet switch and gateway functionality. This may reside in separate boxes as long as the price is bundled. There is also a requirement for a monitoring center and eventual high throughput gateways (~2000 pakets/sec) over T1 links. He hoped to get comments on his document from Berkeley (Re: Ethernet interface) and DoD (Re: Arpanet interface). Due to a possible need for multi-vendor Autonomous Systems, Mills also wanted comments on routing protocols. He pointed out that GGP is the only interior routing protocol that is currently documented as an RFC.

Mills next opened a discussion on EGP. He presented the model and gave some suggestions for improvements. Although he felt that major changes to the protocol were necessary, he also felt that there were modifications that could be made that amounted only to a "re-interpretation" of the current specification. When asked for priorities, he said that the most important thing was to reduce the amount of information exchanged between hosts. Chiappa disagreed, saying that relaxing topological restrictions was more important.

### 5.2 April 9, 1986

St. Johns opened the morning session with a proposal for EGP version negotiation. Mills questioned this, saying that different versions may be acceptable within the same model but that he wants to see a "different trust model". Mills opened up a lively discussion by asking who supports EGP. It was widely agreed that, although many separate implementations exist (e.g., CMU, MIT, BBN, MITRE, Mills), Kirton's is the de-facto standard because it is distributed with Berkeley Unix. Hinden questioned the development cycle for Kirton's EGP and suggested that we needed a better method of releasing patches.

Mills made an action list for EGP modifications. They included:

- Partial updates
- Cache management of routing table
  - Separate TTLs for each net
  - Use LRU replacement
- Remove topology restrictions
- Event driven updates
- Password neighbor acquistion

- Version negotiation

Hinden felt we needed to "engineer" EGP. He pointed out that the update will exceed 576 bytes at 130 networks and will exceed the Arpanet message size of 300 nets. This kicked off a long discussion on restricting access to new users, during which it was noted that connectivity is not guaranteed to new subscribers. When the question of response time was brought up, Chiappa claimed that there are really only two types of "users": humans and mailers. While humans demand shorter response time, mailers could accept longer update cycles.

Cain reported on four pathological incidents and gave a convincing impromptu demonstration of a translating gateway.

Chiappa presented several proposed changes and clarifications to IP/ICMP. The goal was to specify host IP requirements for connection to the Internet to faciliate both routing and fault isolation. A document (included with the hardcopy "Proceedings") has been previously produced, which captured many of these points (e.g., eliminating all Redirect types except "per host/TOS").

Chiappa's proposal for several new ICMP messages generated a lengthy discussion. The proposed messages are:

> Initial Gateway Discovery - Used by a host on a local net upon startup to discover a local gateway. This captures an interesting capability of the ISO proposed ES-IS protocol.
>
> Find Gateway Next Hop - envisioned as a diagnostic tool to debug "black hole" routing problems.
>
> Best Host Address - used to discover the preferred address of a muti-homed host.

When Mills questioned the scope of the effort, Chiappa replied that we needed an improved model for host attachment. Gardner suggested that perhaps a separate protocol was called for.

Zhang presented some very interesting thoughts on IP congestion control. She views it as a feedback control problem, in which the network/gateway system must respond to the level of load offered by hosts. For any feedback scheme to work well, the system response time must be much less than typical changes in the load. She pointed out that a poorly designed or poorly tuned feedback system will be either overdamped or underdamped, that is, it will either not work well or it will oscillate, both of which are symptoms displayed by Source Quench.

Zhang questioned whether we understand either the network load characteristics or the network response time well enough to attack the problem. As a start, we should develop a better understanding of traffic patterns. If, for example, it turns out that data transfers are too bursty, we may need to give up on adaptive control and simply do a better job of sizing the networks.

She listed requirements for retrofitting a congestion control scheme into the current IP. For the host, control must be at the IP level and no overhead should be imposed in the absence of congestion. The gateway must send very specific control information to hosts and have a capability to selectively punish hosts that do not comply.

She sketched a possible scheme; but the real question was how to compute the information passed to the host (e.g., transmit rate and expiration time). In conclusion, she proposed that we collect some detailed traffic measurements and perform some control experiments. Mills suggested that this discussion be continued at the next Internet Architecture Task Force.

Tasman wrapped up the day with comments on the IP "Time To Live" parameter. He recounted

Nagle's observation that, properly utilized, the TTL effectively bounds the length of the output queue. However, as Hinden had mentioned earlier, the gateways check and decrement TTL only once. This amounts to using TTL as a "hop count", rather than as it was intended. This, in turn, allows the queue length to grow, which contributes to round-trip variance. He also pointed out that TTL should not be significantly longer than the retransmission timer, since this leads to multiple copies of the datagram flying around the Internet at once.

Nagle's observation that, properly utilized, the TTL effectively bounds the length of the output queue. However, as Hinden had mentioned earlier, the gateways check and decrement TTL only once. This amounts to using TTL as a "hop count", rather than as it was intended. This, in turn, allows the queue length to grow, which contributes to round-trip variance. He also pointed out that TTL should not be significantly longer than the retransmission timer, since this leads to multiple copies of the datagram flying around the Internet at once.

# APPENDIX A

## Hardcopy of Presentation Slides

# APPENDIX A

## Hardcopy of Presentation Slides

| *Author* | *Title* |
|---|---|
| R. Hinden | Gateway Status Report |
| M. Brescia | GGP Software Bug |
| H. W. Braun | Overview Of NSFnet |
| P. Gross | Recent LSI/Mail Bridge Gateway Performance |
| M. Gardner | Mail Bridge Traffic and Performance |
| M. St. Johns | Proposed Amendment To RFC 904, EGP Specification |
| N. Chiappa | New ICMP Messages and Host Attachment |
| Z. Zhang | IP Congestion Control Proposal |
| E. Cain | Recent Internet Pathology |
| M. Tasman | On The IP TTL |

# GATEWAY STATUS REPORT

## CURRENT INTERNET

131 Networks

85+ Gateways

160,000,000 Packets/Weeks

R. HINDEN
BBNCC
4/9/86

# LSI-11 GATEWAYS

Release 1007

120 Networks

Release 1008

150 Networks

256K bytes Memory

Improved EGP

Release 1008.1

300 Networks

# BUTTERFLY GATEWAY

8 Installed

BBN-WB, SRI-WB, CMU-WB, BBN-UCL,
MIT-WB, MIT, ISI, IPTO

Deployment Schedule

7 Satnet — April - June '

14 SURAN & X.25 — Summer - FALL

Release 3 — Completed

SPF Routing
EGP
Neighbor Op/Down
Interface Up/Down
SATNET HDH
HDLC
1000 Networks (250, 750)

MAILBRIDGE 1009

Load Sharing
Access Control
EGP Access Control

# GGP BUG

- ROUTES TO SOME NETS (ESP. UNREACHABLE ONES) FOUND VIA DOWN NEIGHBORS
  (e.g. YALE via FIBER A)

- ROUTES TO EXT. NETS FOUND AT INT. DISTANCES AND RAPIDLY CHANGING (e.g. CSS 24 UPDATES/MIN)

o SUSPECT PDP-11 SIGN EXTENSION (SIGNED VS. UNSIGNED CHAR) WHEN RUNNING MORE THAN 127 NETS

+ (FINE TOOTH COMB APPROACH) FOUND

  BICB #177400, P ;(NOP)

  REFERRING TO RT MATRIX - WRITE DISTANCE IN WRO
  NBR. COL.

- ROUTING CYCLES STILL FOUND -

• TRAP "redundant route in update"
  (finding RT MATRIX entry ALREADY. SMALL DISTANCE)

+ REUSING NET SLOTS - NET ROW OF RT MATRIX NOT CLEARED WHEN REUSED

+ NET TABLE NOT MARKED IN USE

  EFFECT → NEW NET USED OLD DISTANCE (EXT. DIST INT. DIST

# PATCH WED 4/2 4-7 PM
  (rt. updates 3/min)

M. BRESCIA 4/8/86
BBNCC

Table 1. NSF supercomputer centers.

| Center | Supercomputers | | |
|---|---|---|---|
| | 1984–85 | 1985–86 | 1987 |
| | | *Resource* | |
| Purdue | Cyber 205 | Cyber 205 | |
| Minnesota | Cray 1A | Cray 2/ Cyber 205 | |
| Boeing | Cray 1S | Cray X-MP/24 | |
| AT&T Bell Labs | | Cray X-MP/12 | |
| Colorado State | | Cyber 205 | |
| Digital Productions | | Cray X-MP/22 | |
| | | *National* | |
| JVNC–Princeton | | Cyber 205 | ETA-10 |
| SDSC–San Diego | | Cray X-MP/48 | Cray X-MP/48 |
| NCSA–Illinois | | Cray X-MP/24 | Cray X-MP/48 |
| Theory–Cornell | | IBM 3084/ FPS 264s | IBM 3084/ FPS |
| Pittsburgh | | Cray 1S | Cray 1S |

existing supercomputer centers at Purdue University, the University of Minnesota, Boeing Computer Services, AT&T Bell Laboratories, Colorado State University, and Digital Productions (Table 1). By the end of 1985, a total of 30,000 hours of supercomputer time had been allocated under this program to approximately 800 users, and more than 9000 hours had been consumed. In 1984 also, the OASC issued a project solicitation for national supercomputer centers. As a result, four new NSF centers were funded in 1985—the John von Neumann Center (JVNC) at Princeton University, the San Diego Supercomputer Center (SDSC) on the campus of the University of California at San Diego, the National Center for Supercomputer Applications (NCSA) at the University of Illinois, and the Theory Center, a production and experimental supercomputer center at Cornell University. More recently a fifth center has been established in Pittsburgh, to be run by Westinghouse, Carnegie-Mellon University, and University of Pittsburgh (Table 1).

The NSFnet networking activities were initiated in December 1984 when a panel of the OASC confirmed that networking was a fundamental component of the supercomputer initiative, and, moreover, that a network could be designed to meet the requirements of this initiative while providing the basis for a future, general purpose, national academic research network (*5*). The report proposed a two-phased approach for the development of the network: phase 1 to connect supercomputer users to the supercomputer centers and to each other; and phase 2 to provide a general high-speed network, with speeds of 1.544 megabits per second (Mbps), commonly called "T1 speed," or greater. In addition, a variety of experiments to understand better how to utilize and integrate a number of network topologies and usage modalities are to be initiated.

The general strategy recommended by the networking panel report was that the NSFnet should begin by taking advantage of the existing academic networks. NSFnet should be built as a "network of networks" rather than as a separate new computer network. This general approach is based on the experience gained by the Depart-

Table 2. NSFnet. List of planned member institutions. Key: ARPANET, an existing or planned ARPANET site; SDSC, a San Diego consortium network site; JVNC, a Princeton (JVNC) consortium network site; NCAR, a National Center for Atmospheric Research (NCAR) satellite network site; Illinois, a direct 56-kbps connection to the Illinois Supercomputer Center; backbone, a supercomputer center on the NSFnet backbone network.

| Institution | Network | Institution | Network |
|---|---|---|---|
| Agouron Institute | SDSC | University of North Carolina | ARPANET |
| University of Arizona | JVNC | North Carolina State University | ARPANET |
| AT&T Bell Labs, New Jersey | ARPANET | Northwestern University | ARPANET, Illinois |
| University of California, Berkeley | ARPANET, SDSC | Ohio State University | ARPANET |
| Boeing Computer Services | ARPANET | Oregon State University | NCAR |
| Brown University | JVNC | University of Pennsylvania | ARPANET, JVNC |
| California Institute of Technology | ARPANET, SDSC | Pennsylvania State University | JVNC |
| Carnegie-Mellon University | ARPANET | University of Pittsburgh | ARPANET |
| University of Chicago | Illinois | Princeton University | JVNC |
| Colorado State University | ARPANET, NCAR | Purdue University | ARPANET |
| University of Colorado | JVNC, NCAR | Rice University | ARPANET |
| Columbia University | ARPANET, JVNC | University of Rochester | ARPANET, JVNC |
| Cornell University | ARPANET, backbone | Rutgers University | ARPANET, JVNC |
| City University of New York | ARPANET | Salk Institute | SDSC |
| University of Delaware | ARPANET | San Diego Supercomputer Center | ARPANET, SDSC, backbone |
| Duke University | ARPANET | University of California, San Diego | SDSC |
| Harvard University | ARPANET, JVNC | San Diego State University | SDSC |
| University of Hawaii | SDSC | University of California, San Francisco | SDSC |
| Institute for Advanced Studies, at Princeton University | JVNC | University of California, Santa Barbara | ARPANET |
| University of Illinois, Urbana | ARPANET, NCAR, backbone, Illinois | Scripps Clinic and Research Foundation | SDSC |
| University of Illinois, Chicago | Illinois | Scripps Institute of Oceanography | SDSC |
| Indiana University | ARPANET, Illinois | Southwest Fisheries | SDSC |
| John von Neumann Center | ARPANET, JVNC, backbone | Stanford University | ARPANET, SDSC |
| Kitt Peak Observatory | SDSC | State University of New York, Stony Brook | ARPANET |
| Lawrence Berkeley Laboratory | ARPANET | University of California, Los Angeles | ARPANET, SDSC |
| University of Maryland | ARPANET, SDSC, NCAR | University of Texas, Austin | ARPANET |
| University of Miami | NCAR | University of Utah | ARPANET, SDSC |
| University of Michigan | ARPANET, SDSC, NCAR | University of Washington | ARPANET, SDSC, NCAR |
| University of Minnesota | ARPANET | Westinghouse (Pittsburgh) | ARPANET, backbone |
| Massachusetts Institute of Technology | ARPANET, JVNC | University of Wisconsin | ARPANET, SDSC, NCAR |
| National Center for Atmospheric Research | ARPANET, NCAR, backbone | Woods Hole Oceanographic Institution | NCAR |
| National Science Foundation | ARPANET | Yale University | ARPANET |
| New York University | JVNC | | |

Fig. 2. The 1985 Configuration of the Computer Science Research Network, CSNET; which has three major components: (■) ARPANET sites; (◆) X25NET sites connected to the public X.25 data networks Telenet and UNINET; (●) Phonenet sites with dial-up connections to a central mail relay service at the CSNET Coordination and Information Center (CIC) run by Bolt, Beranek, and Newman (BBN). CSNET provides remote terminal access, file transfer, and electronic mail services to ARPANET and X25NET sites. Electronic mail is the only service available to Phonenet sites. [Courtesy of the CSNET CIC]

in accessing pertinent technical information and in attracting faculty and students.

In October 1985, NSF and DARPA, with DOD support, signed a memorandum of agreement to expand the ARPANET to allow NSF supercomputer users to use ARPANET to access the NSF supercomputer centers and to communicate with each other. The immediate effect of this agreement was to allow all NSF supercomputer users on campuses with an existing ARPANET connection to use ARPANET. In addition, the NSF supercomputer resource centers at Purdue University and the University of Minnesota, and the national centers at the University of Illinois and Cornell University are connected to ARPANET. In general, the existing ARPANET connections are in departments of computer science or electrical engineering and are not readily accessible by other researchers. However, DARPA has requested that the campus ARPANET coordinators facilitate access by relevant NSF researchers (Table 2).

As part of the NSFnet initiative, a number of universities have requested connection to ARPANET. Each of these campuses has undertaken to establish a campus network gateway accessible to all campus researchers, thus ensuring that individual researchers will, in due course, be able to use the ARPANET to access the NSF supercomputer centers, from within their own local computing environment (Table 2). Additional requests for connection to the ARPANET are being considered by NSF.

*CSNET*. Establishment of a network for computer science research was first suggested in 1974, by the NSF advisory committee for computer science. The objective of the network would be to support collaboration among researchers, provide resource sharing, and, in particular, support isolated researchers in the smaller universities.

In the spring of 1980, CSNET, the computer science network, was defined and proposed to NSF as a logical network made up of several physical networks (*10*) of various power, performance, and cost. NSF responded with a 5-year contract for development of the network under the condition that CSNET was to be financially self-supporting by 1986. Initially CSNET was a network with five major components—ARPANET, Phonenet (a telephone-based message-relaying service) (*11*), X25Net (support for the TCP-IP protocol suite over X.25-based public data networks), a public host (a centralized mail service), and a name server (an on-line database of CSNET users to support transparent mail services). The common service provided across all these networks is electronic mail, which is integrated at a special service host, which acts as an electronic mail relay between the component networks. Thus CSNET users can send electronic mail to all ARPANET users and vice versa. CSNET, with DARPA support, installed ARPANET connections at the CSNET development sites at the universities of Delaware and Wisconsin and Purdue University.

In 1981, Bolt, Beranek, and Newman (BBN) contracted to provide technical and user services and to operate the CSNET Coordination and Information Center. In 1983, general management of CSNET was assumed by UCAR—the University Corporation for Atmospheric Research, with a subcontract to BBN. Since then, CSNET has grown rapidly and is currently an independent, financially stable, and professionally managed service to the computer research community (Fig. 2). In the beginning, the need for CSNET was not universally accepted within the computer science community. However, the momentum created by CSNET's initial success caused the broad community support it now enjoys. More than 165 university, industrial, and government computer research groups now belong to CSNET (*12*).

A number of lessons may be learned from the CSNET experience (*12*). (i) The network is now financially self-sufficient, showing that a research community is willing to pay for the benefits of a networking service. (Users pay usage charges plus membership fees ranging from $2,000 for small computer science departments to $30,000 for the larger industrial members.) (ii) While considerable benefits are available to researchers from simple electronic mail and mailing list services—the Phonenet service—most researchers want the much higher level of performance and service provided by the ARPANET. (iii) Providing a customer support and information service is crucial to the success of a network, even (or perhaps especially) when the users are themselves sophisticated computer science professionals. Lessons from the CSNET experience will provide valuable input to the design, implementation, provision of user services, and operation and management of NSFnet, and, in particular, to the development of the appropriate funding model for NSFnet.

CSNET, with support from the NSFnet program, is now developing the CYPRESS project which is examining ways in which the level of CSNET service may be improved, at low cost, to research departments. CYPRESS will use the DARPA protocol suite and provide ARPANET-like service on low-speed 9600-bit-per-second (bps) leased line telephone links. The network will use a nearest neighbor topology, modeled on BITNET, while providing a higher level of service to users and a higher level of interoperability with the ARPANET. The CYPRESS project is designed to replace or supplement CSNET use of X.25 public networks, which has proved excessively expensive. This approach may also be used to provide a low-cost connection to NSFnet for smaller campuses.

*BITNET*. In 1981, City University of New York (CUNY) surveyed universities on the East Coast of the United States and Canada, inquiring whether there was interest in creating an easy-to-use, economical network for interuniversity communications. The response was positive. Many shared the CUNY belief in the importance of computer-assisted communication between scholars. The first link of the new network, called BITNET, was established between CUNY and Yale University in May 1981.

The network technology chosen for BITNET was determined by the availability of the RSCS software on the IBM computers at the initial sites. [The name BITNET stands for Because It's Time NETwork (*13*).] The RSCS software is simple but effective, and

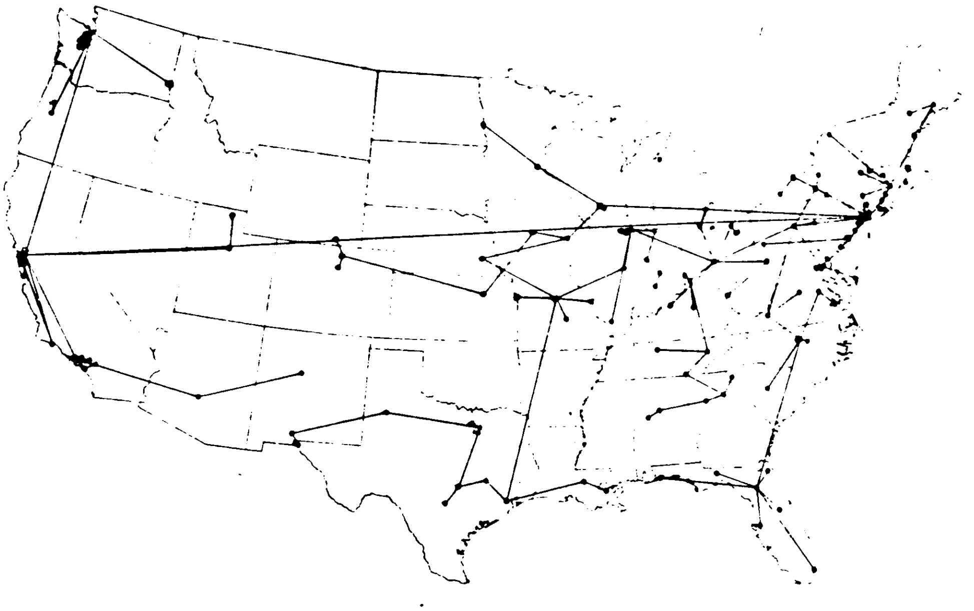


Fig. 3. The 1985 BITNET configuration. BITNET is a store-and-forward network with files and messages sent from host computer to host computer across the network. Services provided include electronic mail, file transfer, and remote job entry. The standard BITNET links are leased telephone lines running at 9600 bps. Electronic mail relays at the University of California at Berkeley and at the University of Wisconsin–Madison provide communication between BITNET, ARPANET, and CSNET users. [Courtesy of Texas A&M University]

most IBM VM-CMS computer systems have it installed for local communications, supporting file transfer and remote job entry services. The standard BITNET links are leased telephone lines running at 9600 bps. Although all the initial nodes were IBM machines in university computer centers, the network is in no way restricted to such systems. Any computer with an RSCS emulator can be connected to BITNET. Emulators are available for Digital Equipment Corporation (DEC) VAX-VMS computer systems, for VAX-UNIX systems, and for Control Data Corporation Cyber systems and others. Today, more than one-third of the computers on BITNET are non-IBM systems.

BITNET is a store-and-forward network with files and messages sent from computer to computer across the network. It provides electronic mail, remote job entry, and file transfer services, and supports an interactive message facility and a limited remote logon facility. Most BITNET sites use the same electronic mail procedures and standards as the ARPANET, and as a result of the installation of electronic mail gateway systems at the University of California at Berkeley and at the University of Wisconsin–Madison, most BITNET users can communicate electronically with users on CSNET and the ARPANET.

BITNET has expanded extremely rapidly—a clear indication that it is providing service that people need and want. The simplicity of connection to the network—acquiring a 9600-bps leased line to the nearest neighboring computer node and installing an additional line interface and a modem—provides the service at the right price. By the end of 1985 the number of computers connected was expected to exceed 600, at more than 175 institutions of higher education throughout the United States (Fig. 3). BITNET is open without restriction to any college or university. It is not limited to specific academic disciplines, and may be used for any academic or administrative purpose. However, use for commercial purposes is prohibited. In special cases, connection of commercial organizations may be sponsored by universities. A particular case is the connection of Boeing Computer Services to BITNET, as part of the NSFnet initiative, to provide remote job entry services to their Cray X-MP/24 to NSF supercomputer grantees who have access to BITNET.

Until recently BITNET had no central management structure, and was coordinated by an executive board consisting of members from the major institutions participating. This worked because most of the computers connected were managed and operated by professional service organizations in university computer centers. However, the growth in the network made it impossible to continue in this ad hoc fashion, and a central support organization was established with support from an IBM grant. The central support organization, called the BITNET network support center (BITNSC), has two parts: A user services organization, the network information center (BITNIC), which provides user support, a name server and a variety of databases, and the development and operations center (BITDOC) to develop and operate the network. A major question facing the members of BITNET is how the funding of this central organization will be continued when the IBM grant expires in 1987.

BITNET, with support from the NSFnet Program, is now examining ways to provide ARPANET-like services to existing BITNET sites. The project, which is similar to the CSNET CYPRESS project, will explore a strategy to provide an optional path to the use of the TCP-IP procedures on existing 9.6-kbps leased lines. The possibility of upgrading these lines to multiple alternate links, providing higher reliability and availability, or to higher speed 56-kbps links is also being studied. The project will offer a higher level of service to BITNET sites choosing this path and also enable a low-cost connection to NSFnet.

*MFENET*. The DOE's magnetic fusion energy research network (MFENET) was established in the mid-1970's to support access to the MFE Cray 1 supercomputer at the Lawrence Livermore National Laboratory. The network uses 56-kbps satellite links, and is designed to provide terminal access to the Cray time-sharing system (CTSS), also developed at the Livermore Laboratory. The network currently supports access to Cray 1, Cray X-MP/2, Cray 2, and Cyber 205 supercomputers. The network uses special-purpose networking software developed at Livermore, and, in addition to terminal access, provides file transfer, remote output queuing, and electronic mail, and includes some specialized application procedures supporting interactive graphics terminals and local personal computer (PC)-based editing. Access to the network is in general restricted to DOE-funded researchers. Recently the network has been expanded to include the DOE-funded supercomputer at Florida State University. MFENET (Fig. 4) is funded by DOE and managed by Livermore.

MFENET has been successful in supporting DOE supercomputer users. However, the specialized nature of the communications protocols is now creating difficulties for researchers who need advanced graphics workstations that use the UNIX BSD 4.2 operating system and the TCP-IP protocols on LAN's. For these

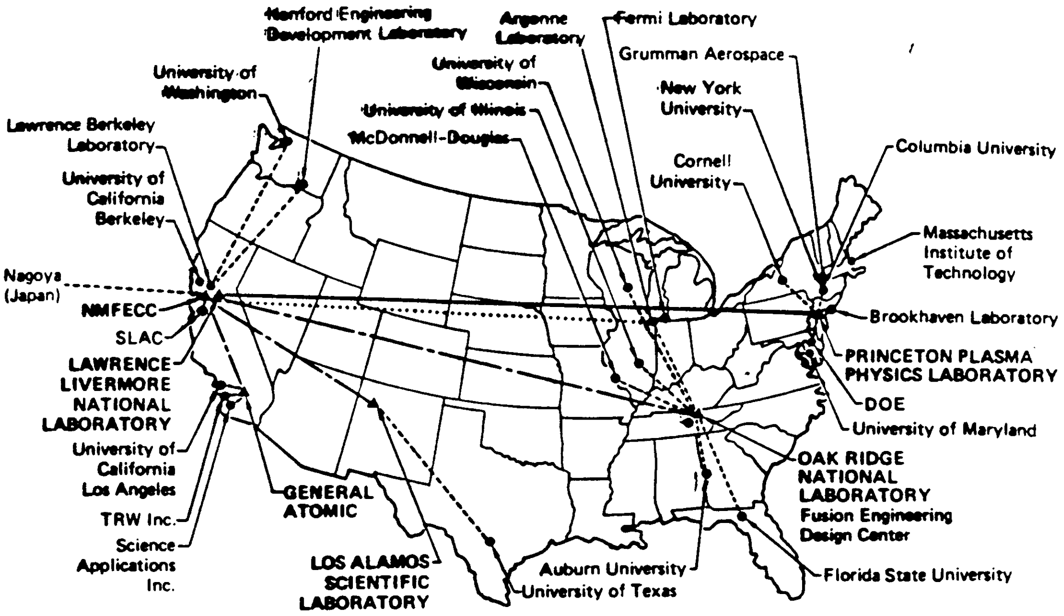


Fig. 4. The 1985 Configuration of DOE's Magnetic Fusion Energy researchers network (MFENET). The network uses dual satellite links at 112 kbps (solid line) and 56 kbps (dashed lines) and terrestrial links at 56 kbps (dotted lines) and 9600 bps (short dashes). The network was developed at the Lawrence Livermore National Laboratory to provide access to supercomputers running the CTSS, also developed at the Livermore Laboratory. The network uses special-purpose networking software developed at MFE. Services include terminal access, file transfer, remote output queuing, and electronic mail. Abbreviations: SLAC, Stanford Linear Accelerator site; NMFECC, National Magnetic Fusion Energy Computer Center. [Courtesy of NMFECC]

and other reasons, DOE is examining how best to migrate MFENET to the TCP-IP, and later to the OSI, protocols.

The combination of the CTSS operating system and the MFENET protocols creates an effective interactive computing environment for researchers using Cray supercomputers. For this reason, two of the new NSF national supercomputer centers—San Diego (SDSC) and Illinois—have chosen the CTSS operating system. In SDSC's case, the MFENET protocols have also been chosen to support the SDSC Consortium network. In Illinois's case, a project to implement the TCP-IP protocols for the CTSS operating system has been funded by the NSFnet program, and these developments will be shared with SDSC (and with DOE) to provide a migration path for the SDSC Consortium network.

*UUCP and USENET*. The UUCP network was started in the 1970's to provide electronic mail and file transfer between UNIX systems. The network is a host-based store and forward network using dial-up telephone circuits and operates by having each member site dial-up the next UUCP host computer and send and receive files and electronic mail messages. The network uses addresses based on the physical path established by this sequence of dial-up connections. UUCP is open to any UNIX system which chooses to participate. There are "informal" electronic mail gateways between UUCP and ARPANET, BITNET, or CSNET, so that users of any of these networks can exchange electronic mail.

USENET is a UNIX news facility based on the UUCP network that provides a news bulletin board service. Neither UUCP nor USENET has a central management; volunteers maintain and distribute the routing tables for the network. Each member site pays its own costs and agrees to carry traffic. Despite this reliance on mutual cooperation and anarchic management style, the network operates and provides a useful, if somewhat unreliable, and low-cost service to its members. Over the years the network has grown into a worldwide network with thousands of computers participating.

## Other Wide-Area Networks

Of necessity this discussion of wide-area networks has been incomplete: Other networks of interest include the Space Plasma Analysis Network (SPAN)—a network of DEC VAX computers using 9.6-kbps links and the DECNET protocols for National Aeronautics and Space Administration's researchers; the planned Numerical and Atmospheric Sciences (NAS) network centered at Ames Research Center—a network that is expected to use existing and planned NASA communications links and the TCP-IP protocols; and the planned high-energy physics network—a network based largely on VAX computers and using the standard X.25 network level protocols plus the so-called "coloured books" protocols developed in the United Kingdom. Also, many high-energy physicists, at the Stanford Linear Accelerator, at the Lawrence Berkeley Laboratory, and at Fermi Laboratory, among others, have used DECNET to connect their DEC VAX computers together.

## State Networks

A number of states have over the years developed state-wide networks to provide access to shared computing facilities and to support exchange of information among researchers. The best known of these is the Merit Computer Network in Michigan, which links the campuses of the University of Michigan and of Oakland, Michigan State, Wayne State, and West Michigan universities. This is an extensive network, providing terminal access to a wide variety of resources, and is based on the use of the X.25 network level protocols.

Other states are beginning to examine the development of a statewide research network. An example is the proposal for a New York State education and research network (NYSERNet). This network is envisaged by the proposers to provide a computer communications infrastructure for both the academic research institutions, and for high technology industrial research laboratories in the state. The network is designed not only to support the development of research activities between the academic researchers and existing industry, but also to provide the basis for the attraction of new high-technology industry to the state.

NYSERNet is to be based on multiple redundant T1 (1.544 Mbps) links, and high-performance switches, with gateways to every campus. The network will support the DARPA protocol suite, and the host and campus gateways will run the TCP-IP protocols. The plan envisions that each campus will install a campus-wide network—a model that is entirely consistent with the NSFnet model—and that each individual researcher will be equipped with a powerful graphics workstation. All computing and information resources on the network, including the new NSF national supercomputer center at Cornell, will be accessible from those workstations. NYSERNet, will also be gatewayed to the NSFnet, and will become an integral part of the evolving national research network.

*Supercomputer consortia and "Backbone" networks, NSFnet pilot projects*. Two of the NSF national supercomputer centers are consortia endeavors. The JVNC center was proposed by the Princeton consortium, and the SDSC center by the San Diego consortium. Each proposed a network to link the members of their consortium to their supercomputer center.

*The Princeton consortium network*. The Princeton consortium comprises 13 schools, mostly along the East Coast of the United States (Table 2). The planned consortium network is a star network linking the member campuses to the JVNC. The network uses T1 circuits (1.544 Mbps) in most cases, and each link will be terminated at a campus gateway system, providing connection to a campus-wide network—a model consistent with the NSFnet model. The campus gateway systems and the front-end computers at the JVNC will run the DARPA protocol suite, so the Princeton consortium network is, in fact, an integral part of the NSFnet. Researchers on the consortium campuses will be able to access the JVNC Cyber 205 (and by mid-1987 the ETA-10 system), and, via the consortium network, the other national supercomputer centers and the other campuses on NSFnet, from within their own local computing environments. The Princeton consortium network should be operational by June 1986.

*The San Diego consortium network*. The San Diego consortium comprises 19 schools, mostly along the West Coast of the United States (Table 2). The consortium network is also a star network linking the consortium member campuses to the San Diego center. The network uses 56-kbps circuits, of various types, with each link terminated at a campus remote user access system (RUAC), providing access to the supercomputer for campus researchers—a model somewhat similar to the NSFnet model. Because the SDSC will operate a CRAY X-MP/48 system running the CTSS operating system, the consortium network will initially use the MFENET protocols providing terminal access, file transfer, remote output queuing, interactive graphics, and electronic mail. Although the network will not be an integral part of the NSFnet, a migration to the DARPA protocol suite is planned and is expected to take place during 1987. As an interim measure a gateway/relay system will be installed at the SDSC, which will be accessible to the consortium users, and which will be connected to the NSFnet. Thus consortium users will be able to access the other national supercomputer centers, and other users on the NSFnet will be able to access the SDSC. The San Diego consortium network should be completed by August 1986.

*The supercomputer "backbone" network*. To connect the supercomputer consortia networks to all the NSF national supercomputer centers, including the long-established National Center for Atmospheric Research (NCAR) and to facilitate cooperation between the centers (such as for file transfer, data sharing, or load balancing), NSF is installing a supercomputer "backbone" network, as part of the development of NSFnet (Fig. 5). Initially, this network will be based on multiple 56-kbps circuits, with low-speed switches and gateways, but it is envisioned that the network will be upgraded to T1 circuits as the volume of user to supercomputer traffic and file-transfer traffic between supercomputer centers grows. This backbone will be integral to the NSFnet internet. The network may be expanded to include connections to other supercomputer centers and to the larger campuses.

*NSFnet pilot projects*. In addition to the CSNET CYPRESS project, the BITNET migration project, and the Illinois project to develop the TCP-IP procedures for the CTSS operating system, the NSFnet program will include a number of pilot networking projects. The objective will be to explore the use of new networking technologies and to gain experience to assist with the design of the phase 2 NSFnet.

Although it is expected that several substantial projects will be funded over the next few years, to date only one pilot project has been funded—the NCAR satellite experiment. This project will utilize Ku-band (12 to 14 GHz) satellite equipment developed by the Vitalink Corporation to link together Ethernets in several locations in the United States. The central or "hub" site will be located at NCAR in Boulder, Colorado, and will broadcast at 224 kbps to several remote sites [the universities of Illinois, Maryland, Miami, Michigan, and Wisconsin, Oregon State University, and the Woods Hole Oceanographic Institution (Table 2)]. Each remote site will be able to receive data addressed to it by the hub site at up to 224 kbps, and each will have a dedicated 56-kbps return satellite path to NCAR. In addition, 56-kbps terrestrial links will be installed to Colorado University and Colorado State University. The Ku-band Earth stations used are relatively inexpensive.

The objective of the NCAR pilot project is to explore the use of the shared broadcast channel to provide high-speed communications to remote supercomputer users, to investigate the optimization required to efficiently use the satellite network with the TCP-IP protocols, and to develop the experience necessary to evaluate the more extensive use of Ku-band satellite channels and the Vitalink technology in the phase 2 NSFnet.

## Campus Networks

The same factors that have motivated the development of wide-area networks—access to a variety of computing facilities and communication amongst researchers—have also motivated the development of campus networks. Until recently, these developments

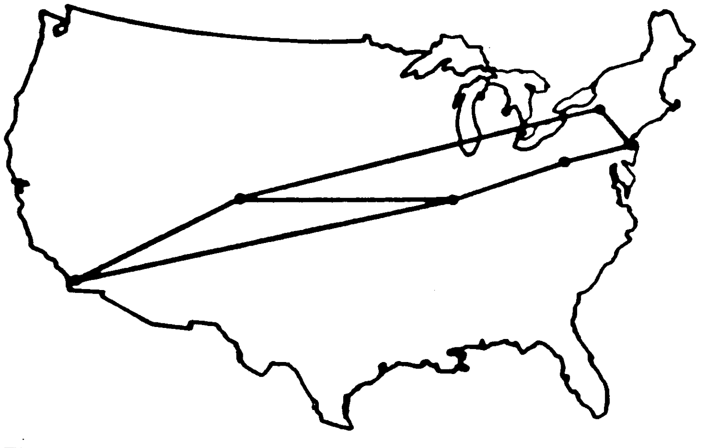


Fig. 5. Planned backbone network connecting NSF-sponsored supercomputers at Cornell University, the John von Neuman Center, at the University of Pittsburgh, the University of Illinois, the National Center for Atmospheric Research, and the San Diego Supercomputer Center. The links will be 56-kbps terrestrial digital circuits connecting network gateways at each site. The supercomputer front-end computers will run the NSFnet standard protocols (TCP-IP and associated application protocols). The NSFnet backbone network will be connected to the ARPANET, to various regional and state networks, and to the planned NSF supercomputer center networks to provide NSF-sponsored supercomputer users with access to all the NSF supercomputer centers. [Courtesy of the NSF's Office of Advanced Scientific Computing]

# CYPRESS *

# A New CSNET

# Technology

192.12.63

* CYPRESS is a joint effort of

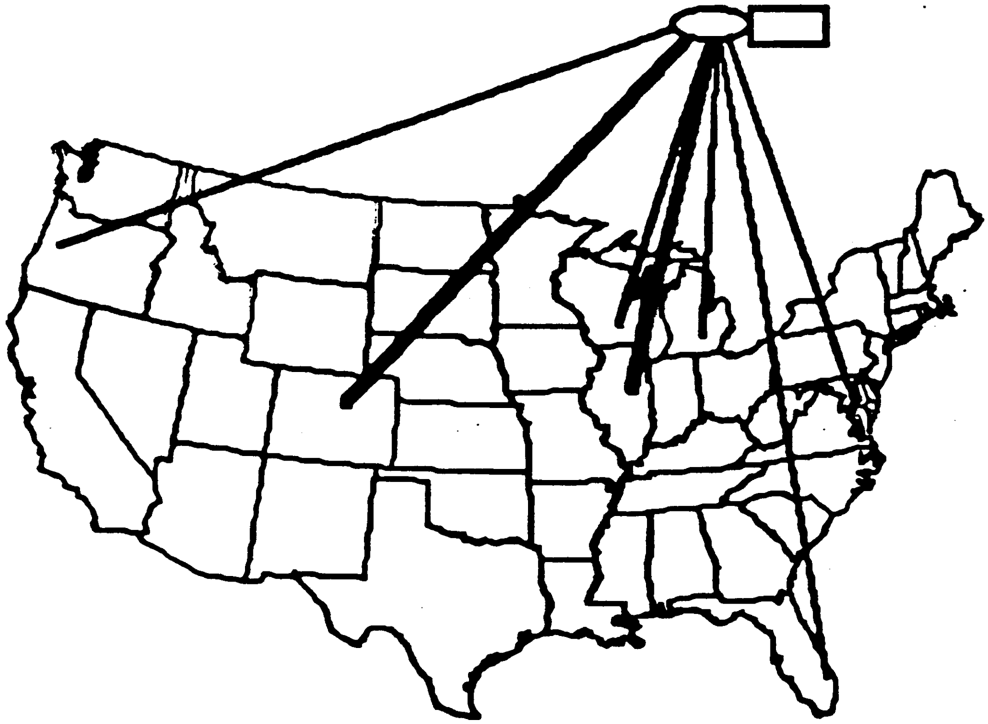
Digital Equipment Corp., CSNET, and Purdue University

# POSSIBILITIES

- Increase backbone capacity
- Dynamic routing updates
- TAC type access at nodes
- Phonenet relays at some sites
- Biconnected topology
- Satellite broadcast to leaf nodes

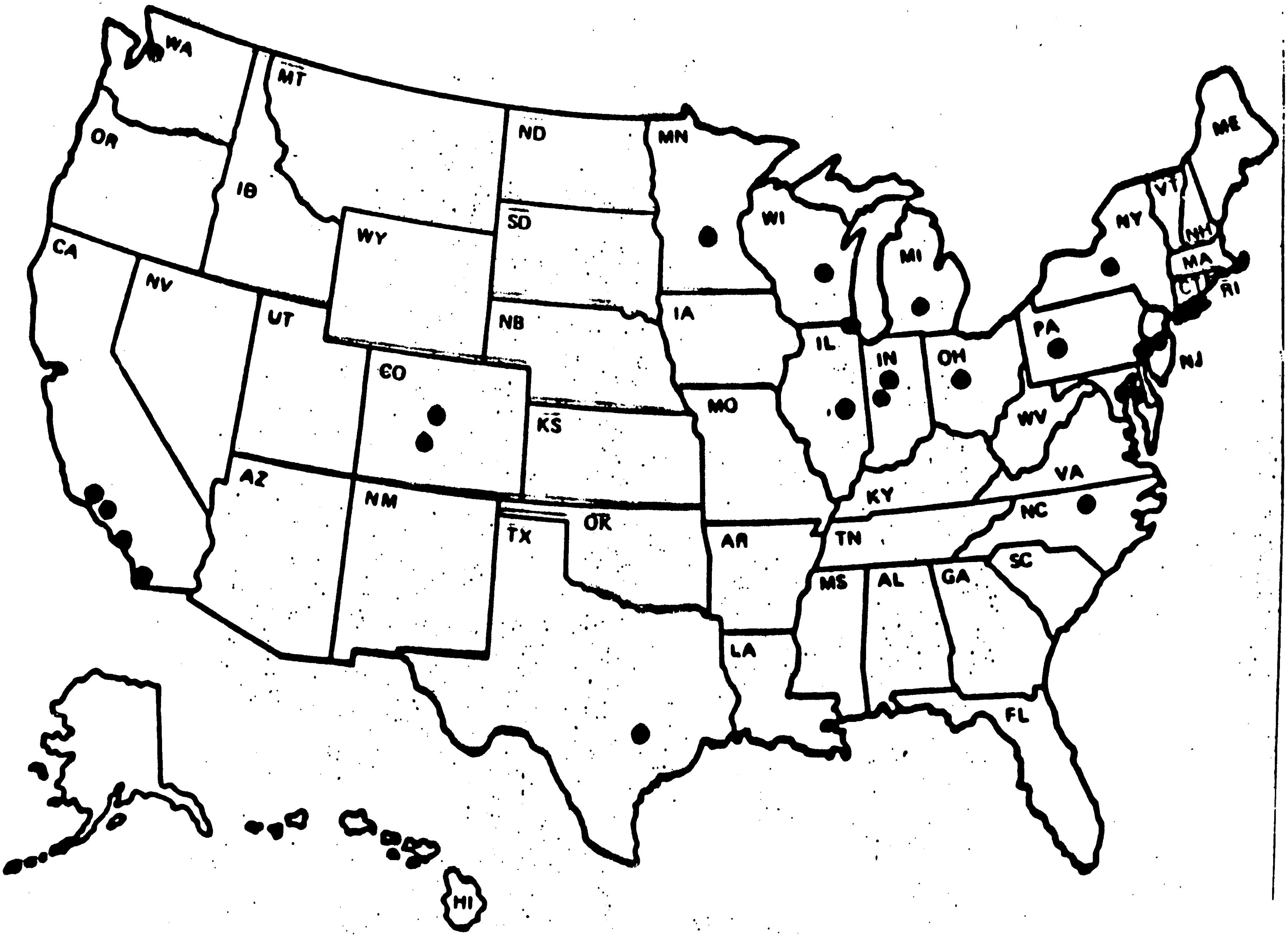
# USAN

## University Satellite Network Project

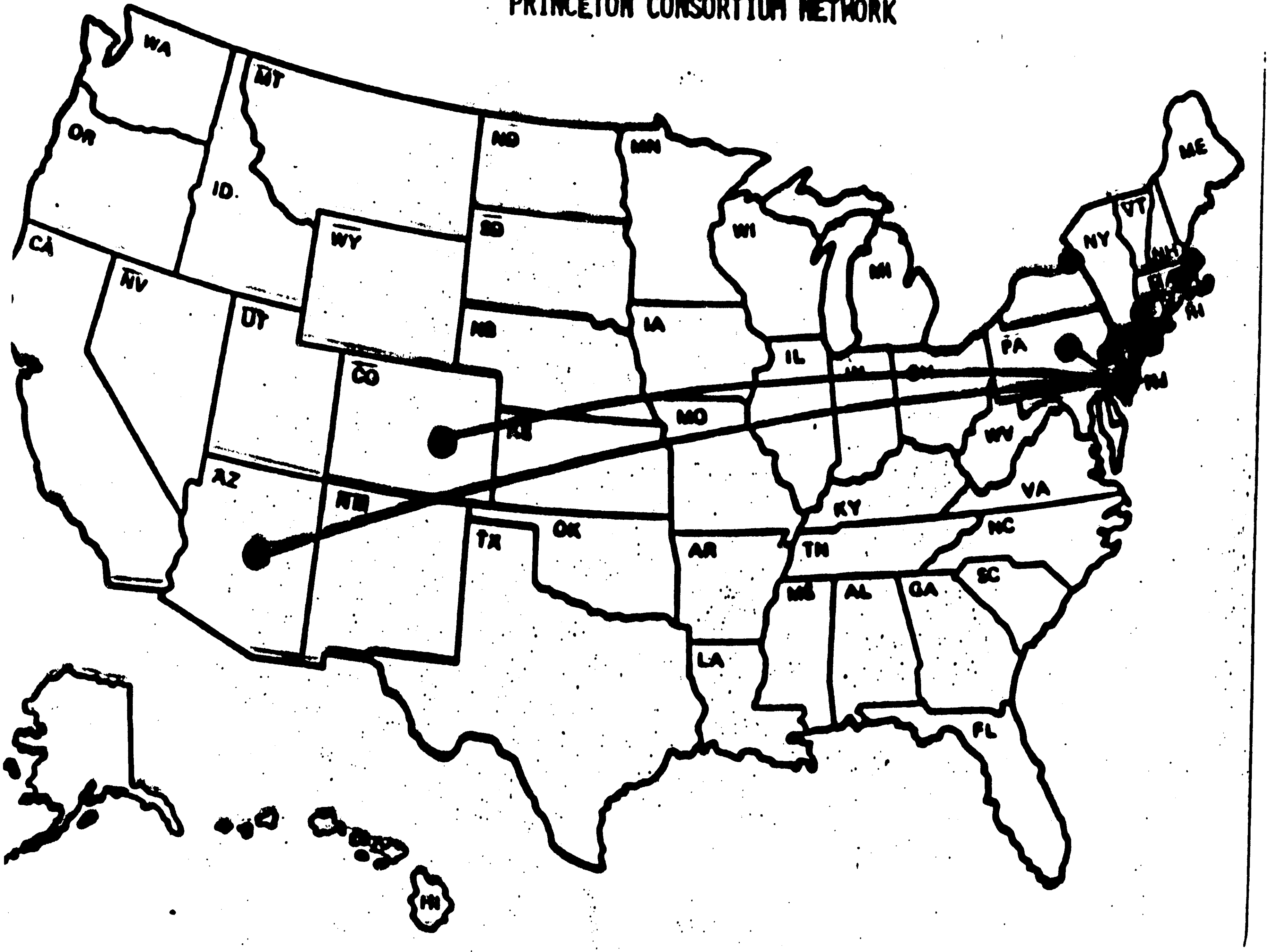


NCAR, Boulder, Colorado
Oregon State University, Corvallis, Oregon
University of Illinois, Urbana, Illinois
University of Maryland, College Park, Maryland
University of Miami, Miami, Florida
University of Michigan, Ann Arbor, Michigan
University of Wisconsin, Madison, Wisconsin

# ARPANET EXPANSI. : 1ST PHASE

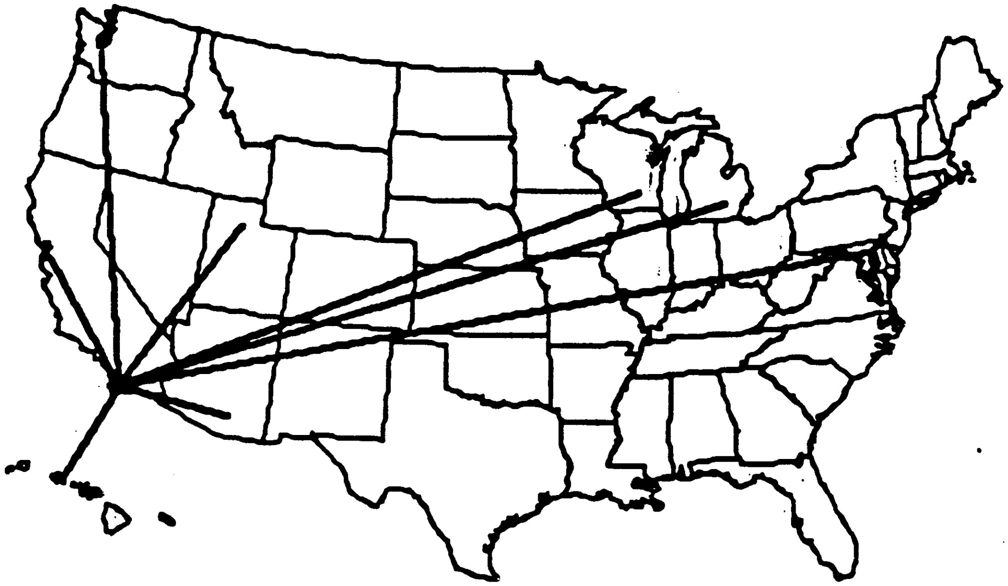


# PRINCETON CONSORTIUM NETWORK



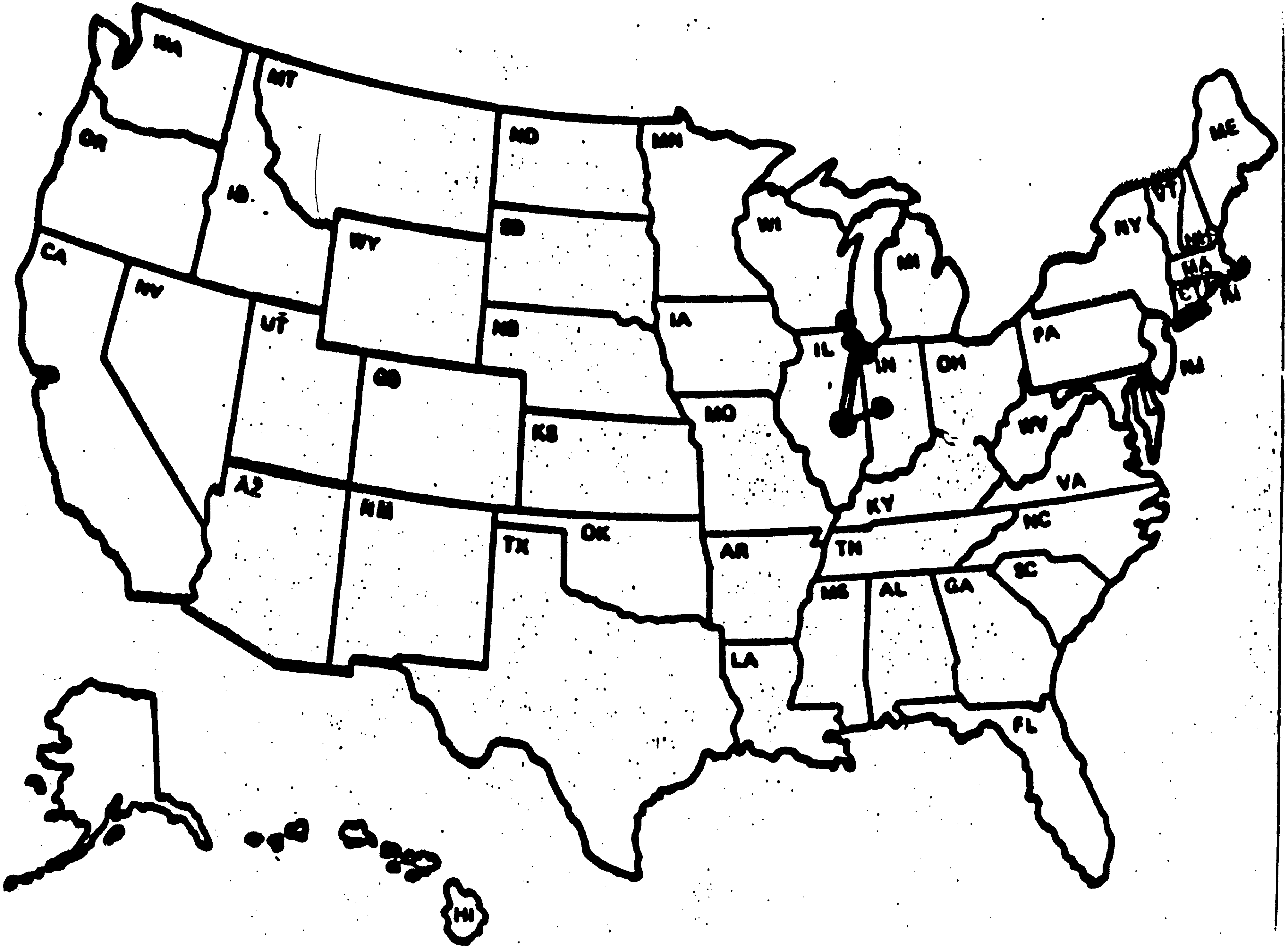
# SDSC

## San Diego Supercomputer Consortium

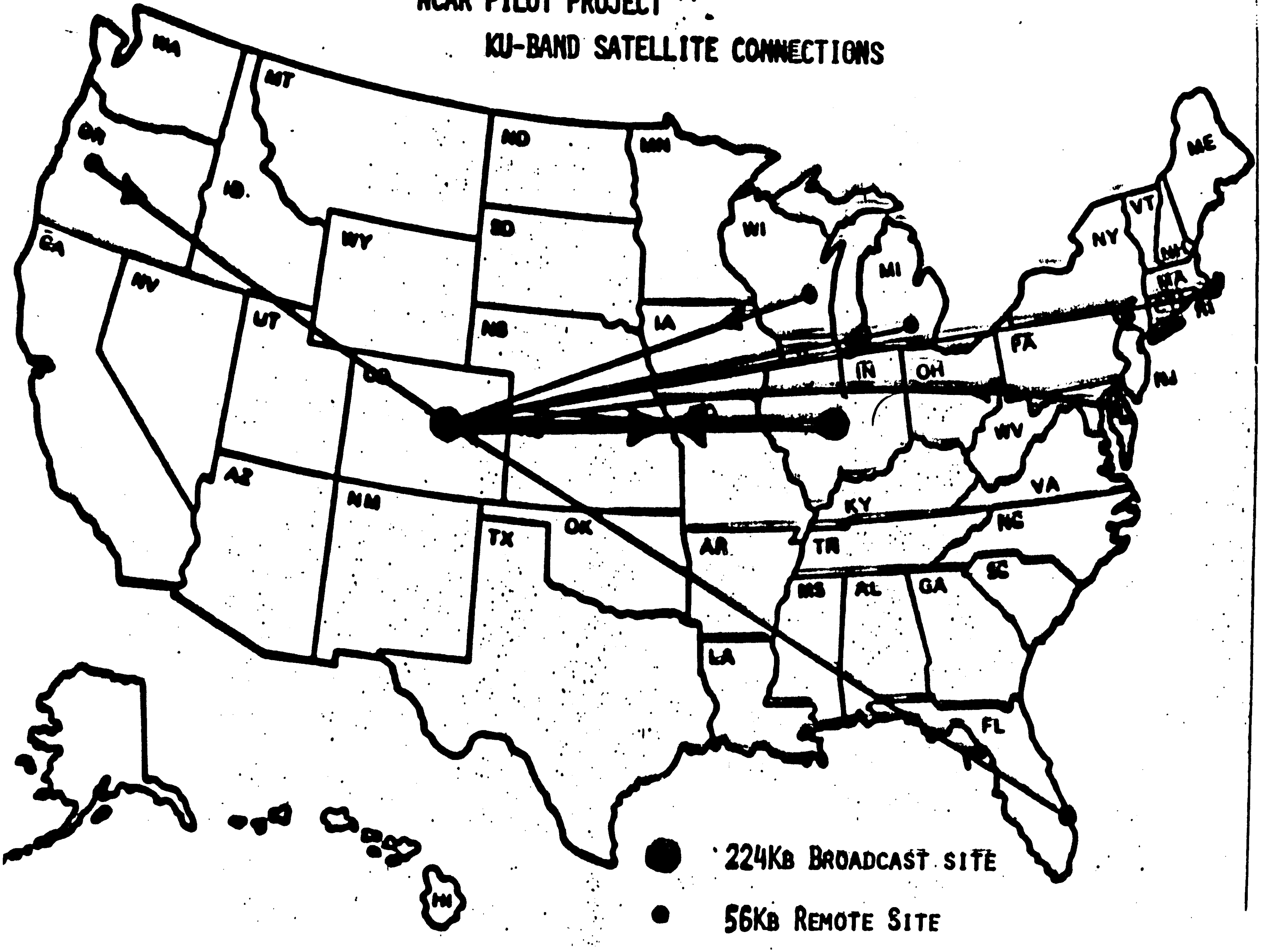


Agouron Institute, La Jolla, California
California Institute of Technology, Pasadena, California
National Optical Astronomy Observatories, Tucson, Arizona
Research Institute of Scripps Clinic, La Jolla, California
Salk Institute for Biological Studies, San Diego, California
San Diego State University, San Diego, California
Scripps Institute of Oceanography, La Jolla, California
Southwest Fisheries Center, La Jolla, California
Stanford University, Stanford, California
University of California -- Berkeley, Berkeley, California
University of California -- Los Angeles, Los Angeles, California
University of California -- San Diego, La Jolla, California
University of California -- San Francisco, San Francisco, California
University of Hawaii, Honolulu, Hawaii
University of Maryland, College Park, Maryland
University of Michigan, Ann Arbor, Michigan
University of Utah, Salt Lake City, Utah
University of Washington, Seattle, Washington
University of Wisconsin, Madison, Wisconsin

# ILLINOIS CENTER NETWORK



# NCAR PILOT PROJECT
## KU-BAND SATELLITE CONNECTIONS



# INITIAL EXPERIMENT

5 sites

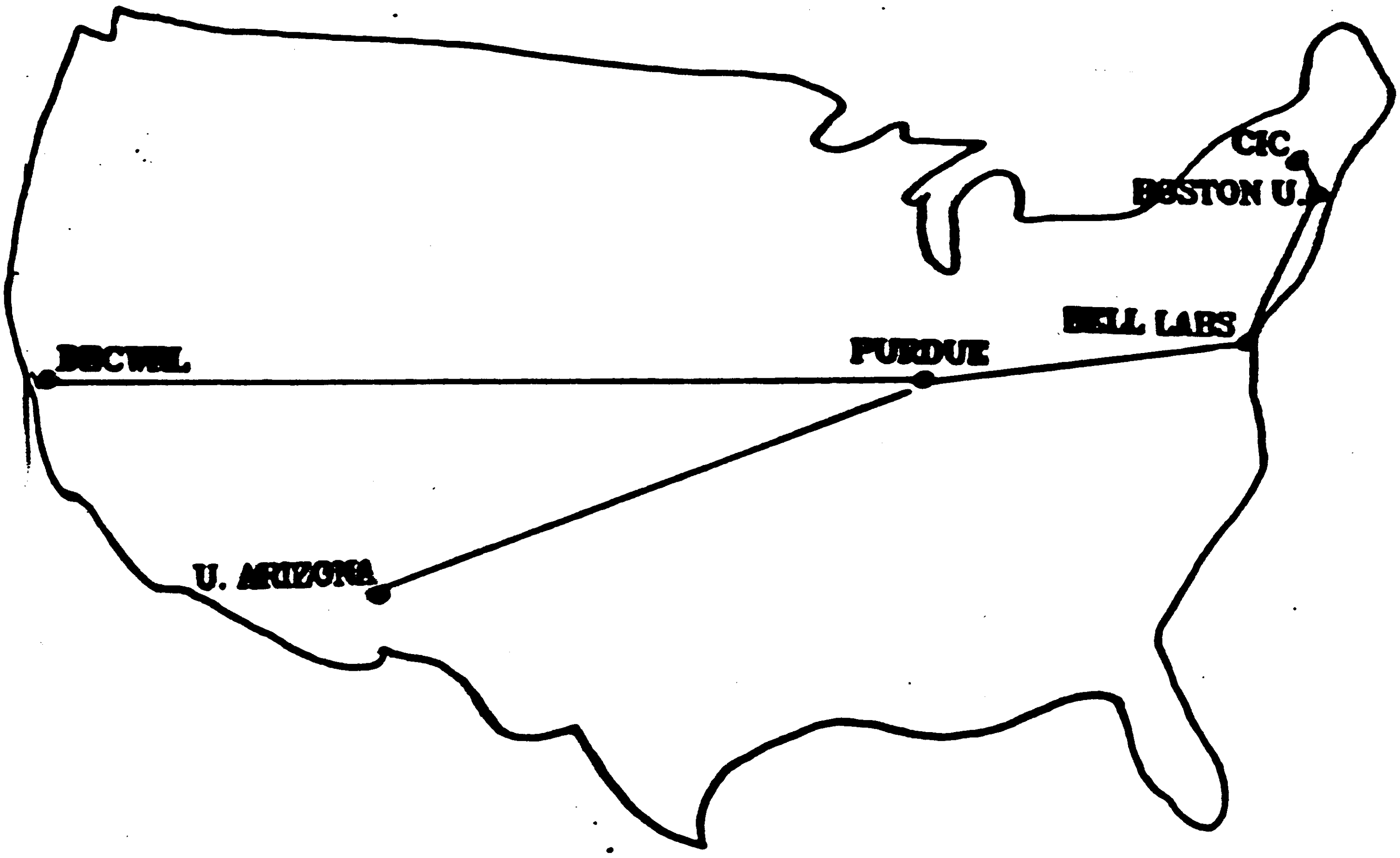
9.6 Kbaud connection

DEC VAX 11/725

4.2bsd (or ULTRIX)

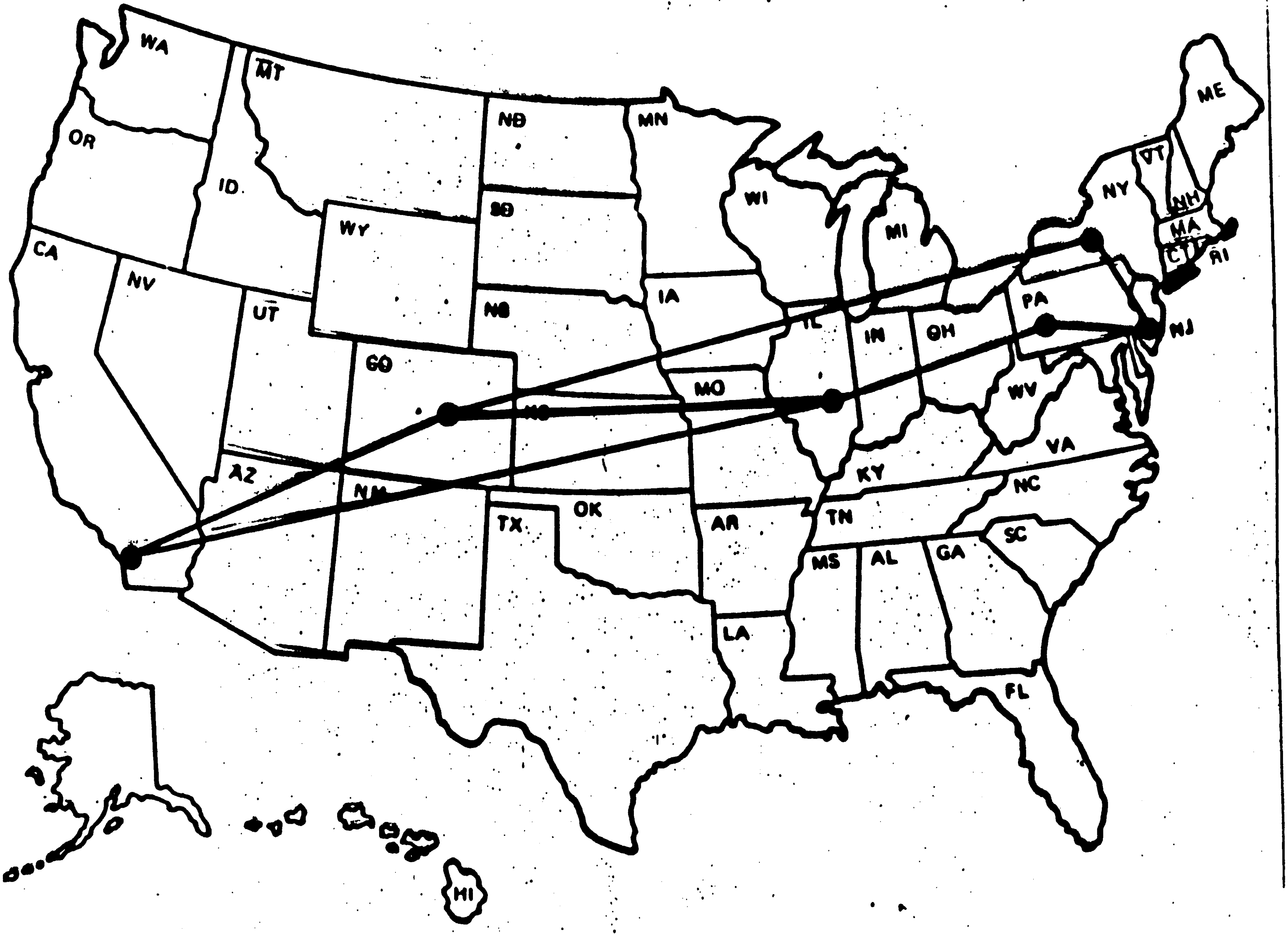
Ethernet load area network connection

Serial IP long-haul connections



Initial CYPRESS Topology

# NSFNET BACKBONE NETWORK



# NSFNET

## PHASE 1: PILOT PROJECTS

OBJECTIVE: USE EXISTING TECHNOLOGIES TO EXPLORE THE ENHANCEMENT OF USER TO SUPERCOMPUTER COMMUNCATIONS.

SEVERAL PROPOSALS BEING EVALUATED

E.G. o VITALINK TRANSLAN
  o DARPA WIDEBAND
  o WORKSTATION PROJECTS
    - SCIENTISTS WORKBENCH

ALSO, DISCUSSIONS WITH SEVERAL COMMUNICATIONS CARRIERS.

# NSFNET

## PHASE 1: PILOT PROJECTS

OBJECTIVE: USE EXISTING TECHNOLOGIES TO EXPLORE THE ENHANCEMENT OF USER TO SUPERCOMPUTER COMMUNCATIONS.

SEVERAL PROPOSALS BEING EVALUATED

E.G.
- o VITALINK TRANSLAN
- o DARPA WIDEBAND
- o WORKSTATION PROJECTS
  - - SCIENTISTS WORKBENCH

ALSO, DISCUSSIONS WITH SEVERAL COMMUNICATIONS CARRIERS.

NSFNET

## NETWORK DEVELOPMENT STRATEGY

PHASE 1:

GOAL: Provide Access to Supercomputers

- o Community Networks
- o Consortia Networks
- o Pilot Projects
- o Studies

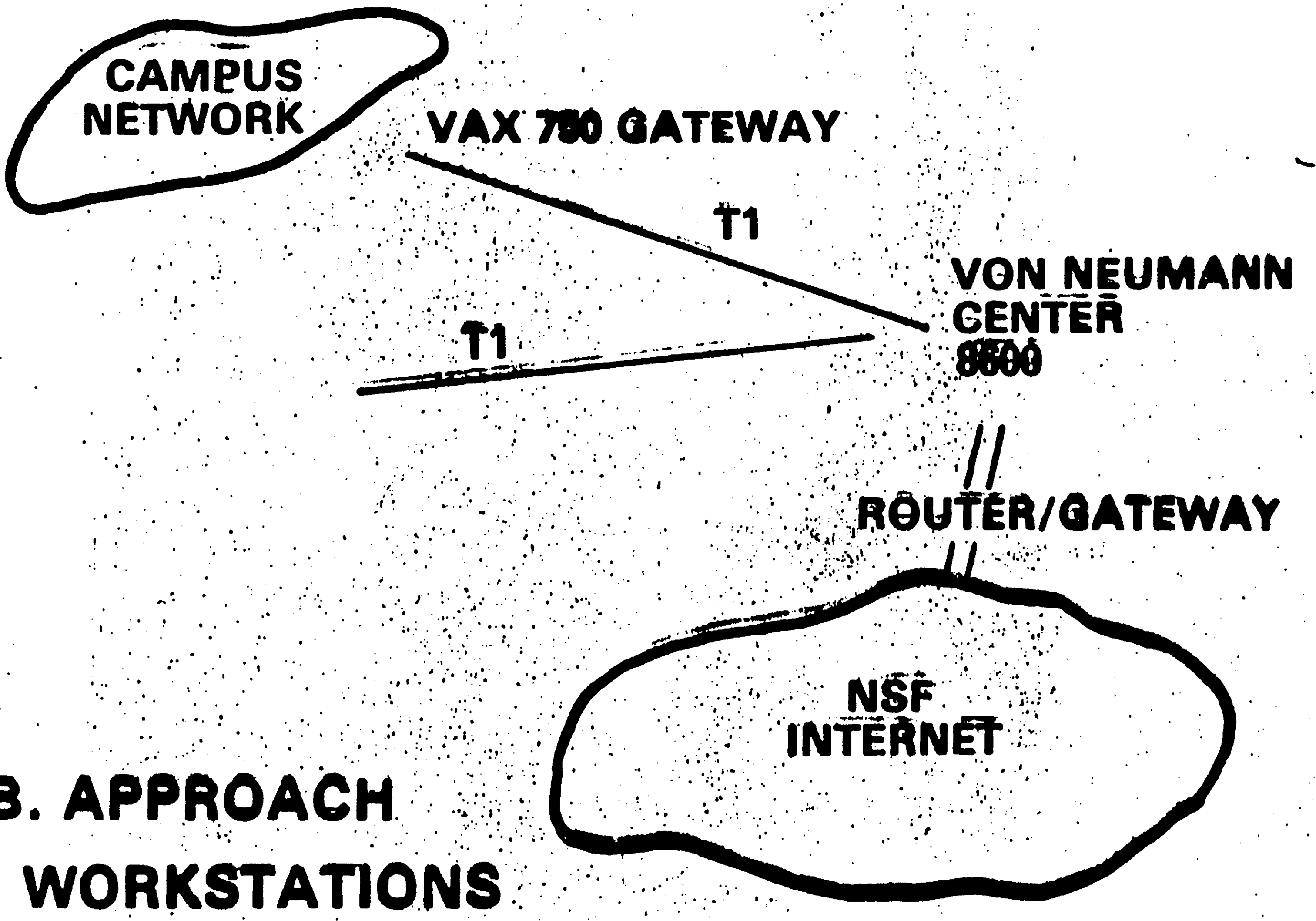
PHASE 2:

GOAL: High Speed Access to Supercomputers

- o Based on Phase 1 Experience

## NSFNET

## PRINCETON CONSORTIUM NETWORK



CAMPUS NETWORK

VAX 750 GATEWAY

T1

T1

VON NEUMANN CENTER 8600

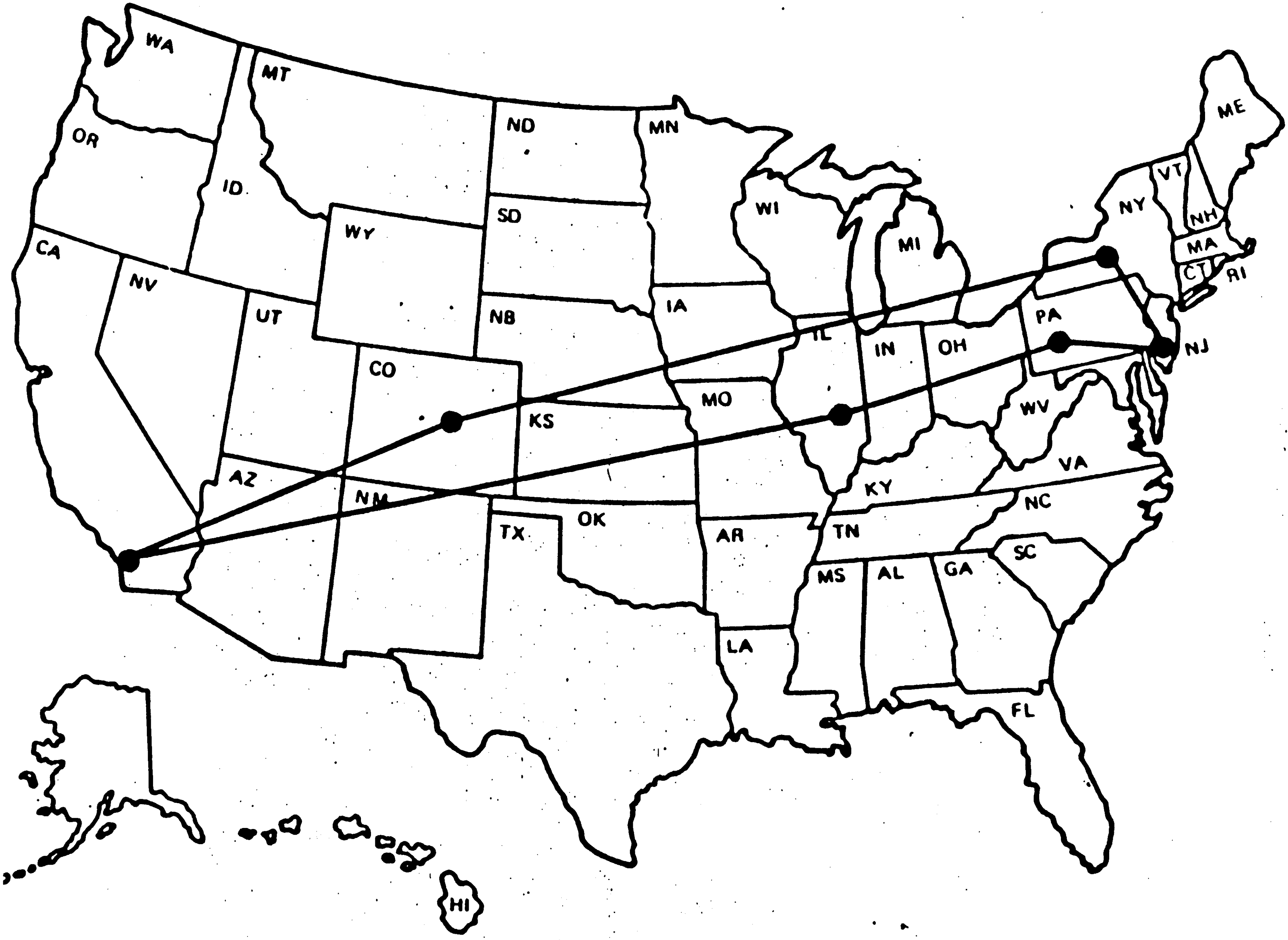
ROUTER/GATEWAY

NSF INTERNET

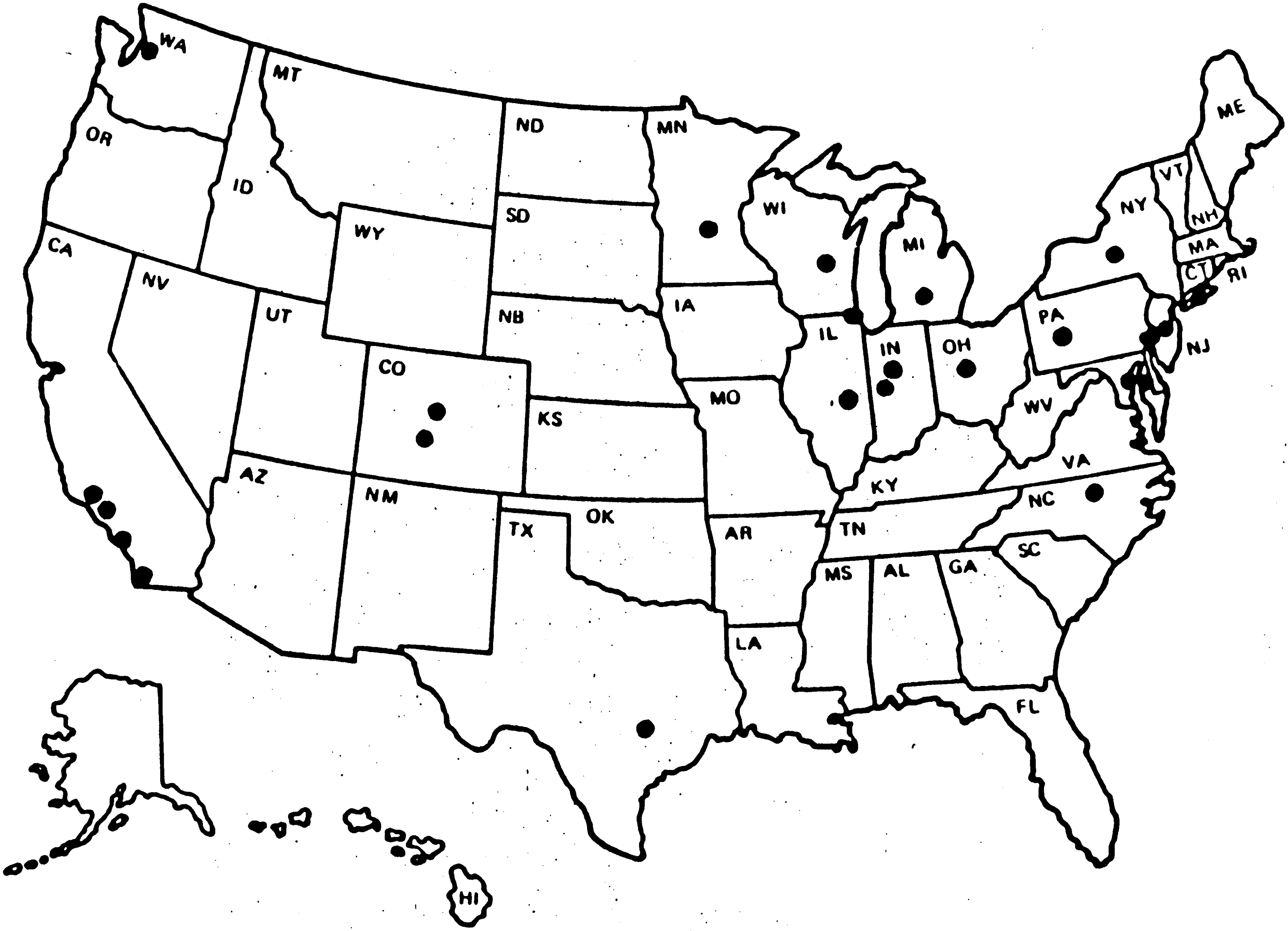
## N.B. APPROACH

- WORKSTATIONS
- POWERFUL F/E FILE & COMMS. SERVERS
- SUPERCOMPUTERS AS CYCLE SERVER
  (BATCH MACHINES RATHER THAN INTERACTIVE)

# NSFNET BACKBONE NETWORK



ARPANET EXPANSION: 1st PHASE



## NSF - INTERNET

## NETWORKING

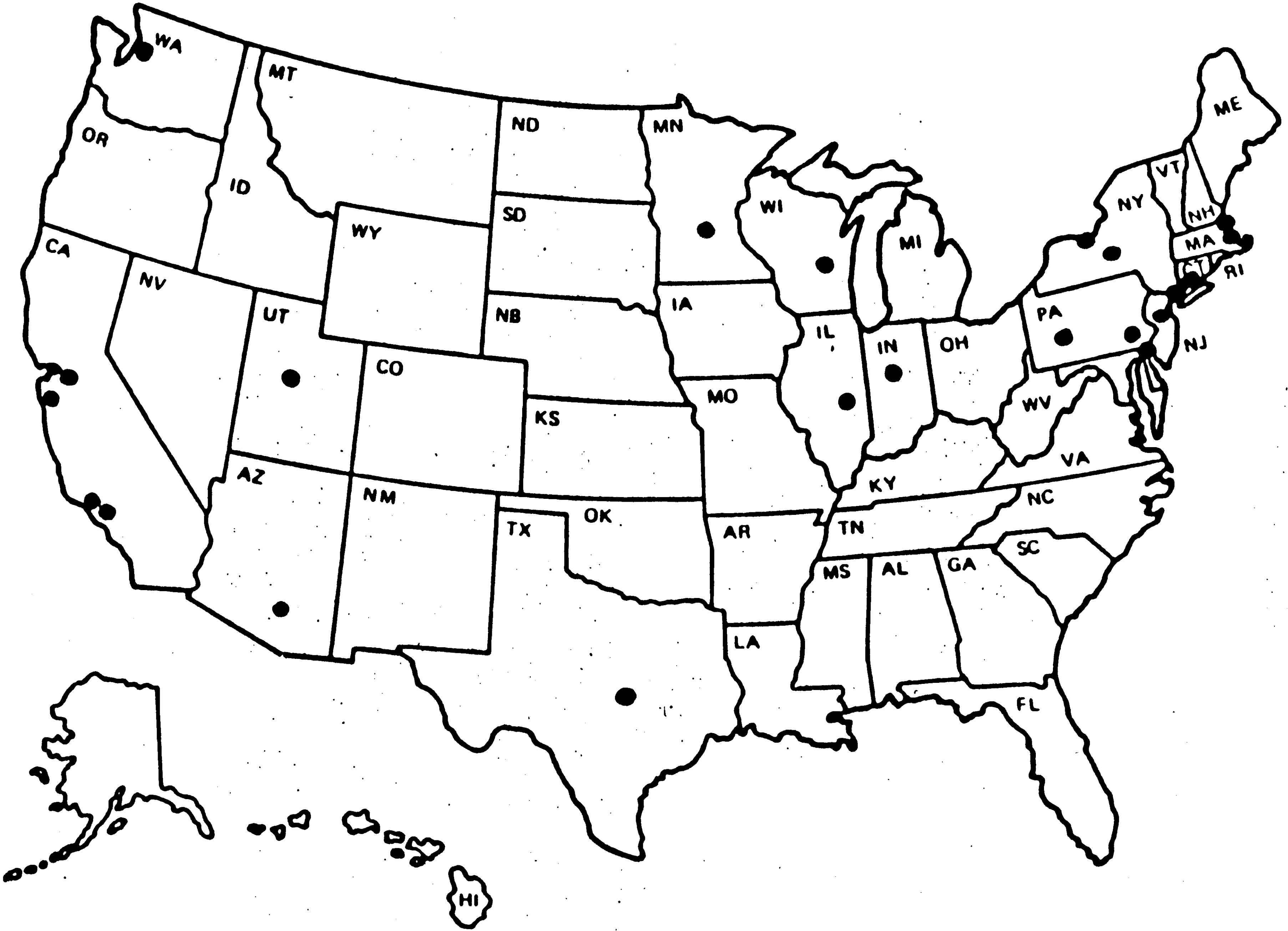
**BACKGROUND:**

- Seen as a Fundamental Component of the Supercomputer Initiative

**GOALS:**

- Access to Supercomputers
- Information Exchange Infrastructure for Supercomputer Users
- Basis of National Research Network

# ARPANET EXISTING SITES



SAN DIEGO

o AGOURON INSTITUTE

o CALTECH

o KITT PEAK

o SCRIPPS CLINIC

o SALK INSTITUTE

o SAN DIEGO STATE U.

o SW FISHERIES CENTER

o STANFORD

o UCLA

o UC SAN DIEGO

o UC SAN DIEGO - SCRIPPS

o UC SFO

o U. HAWAII

o MARYLAND

o MICHIGAN

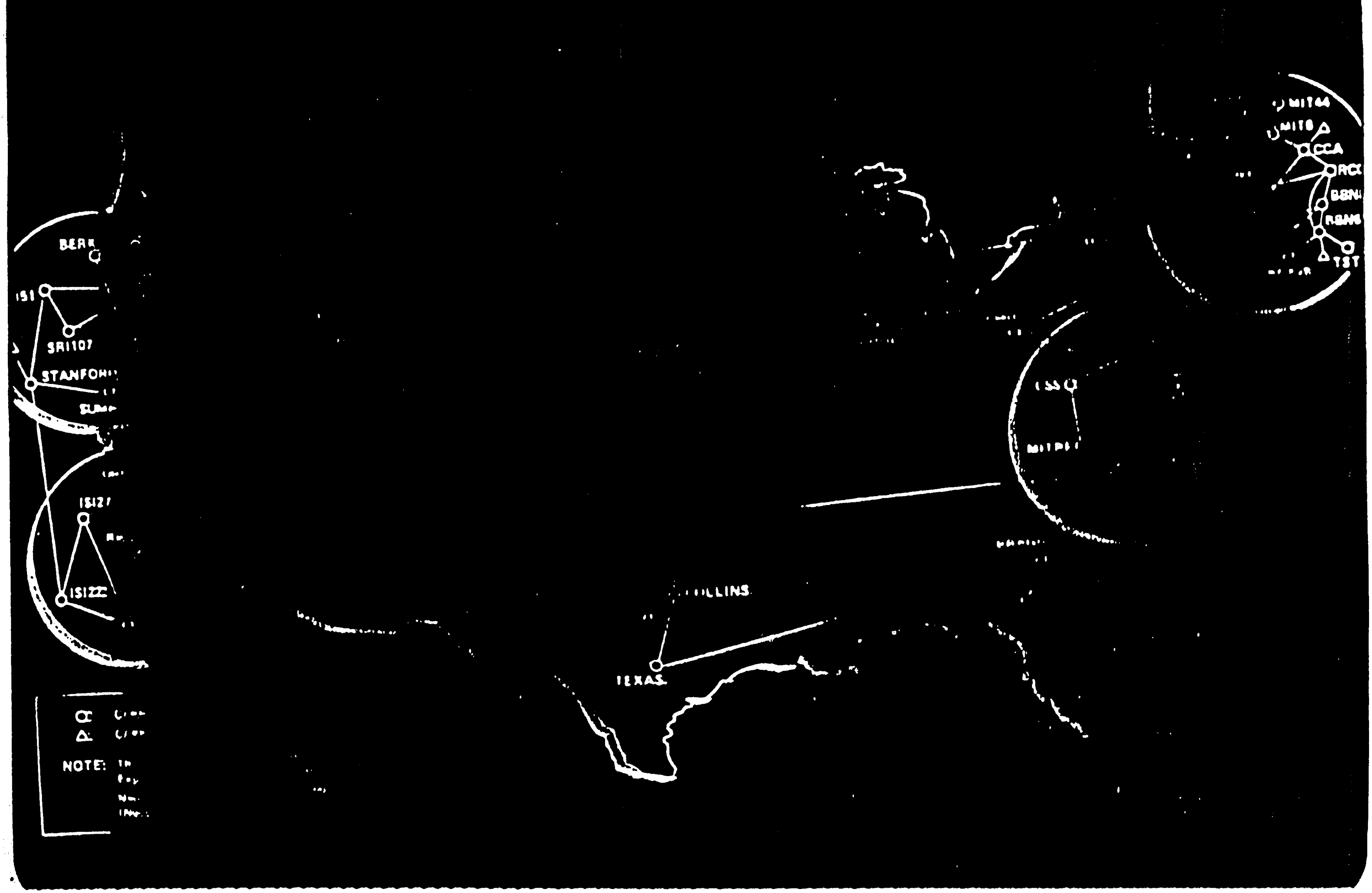
o UTAH

o WASHINGTON

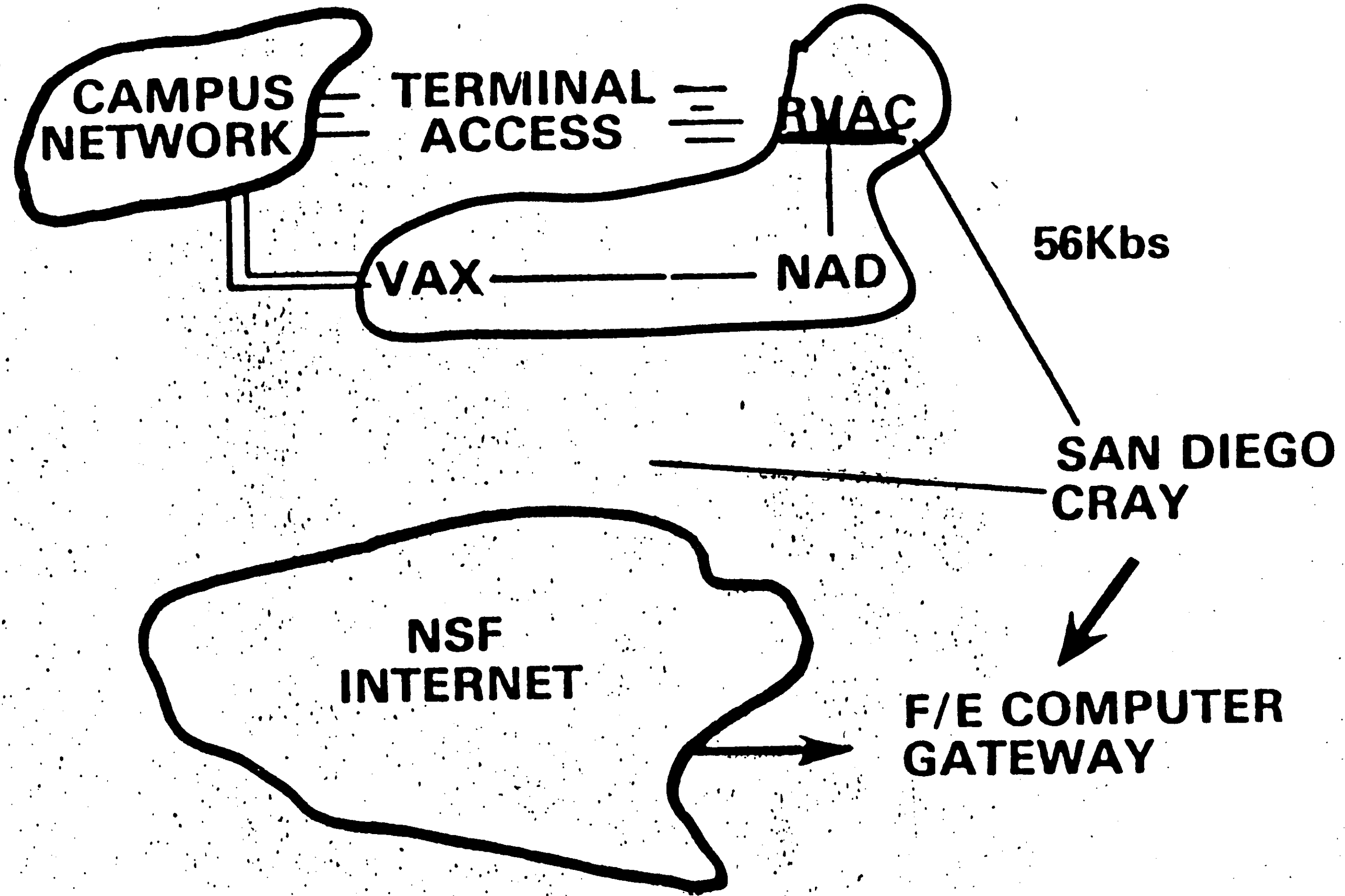
o WISCONSIN

o BERKELEY

# ic Map, 31 May 1985



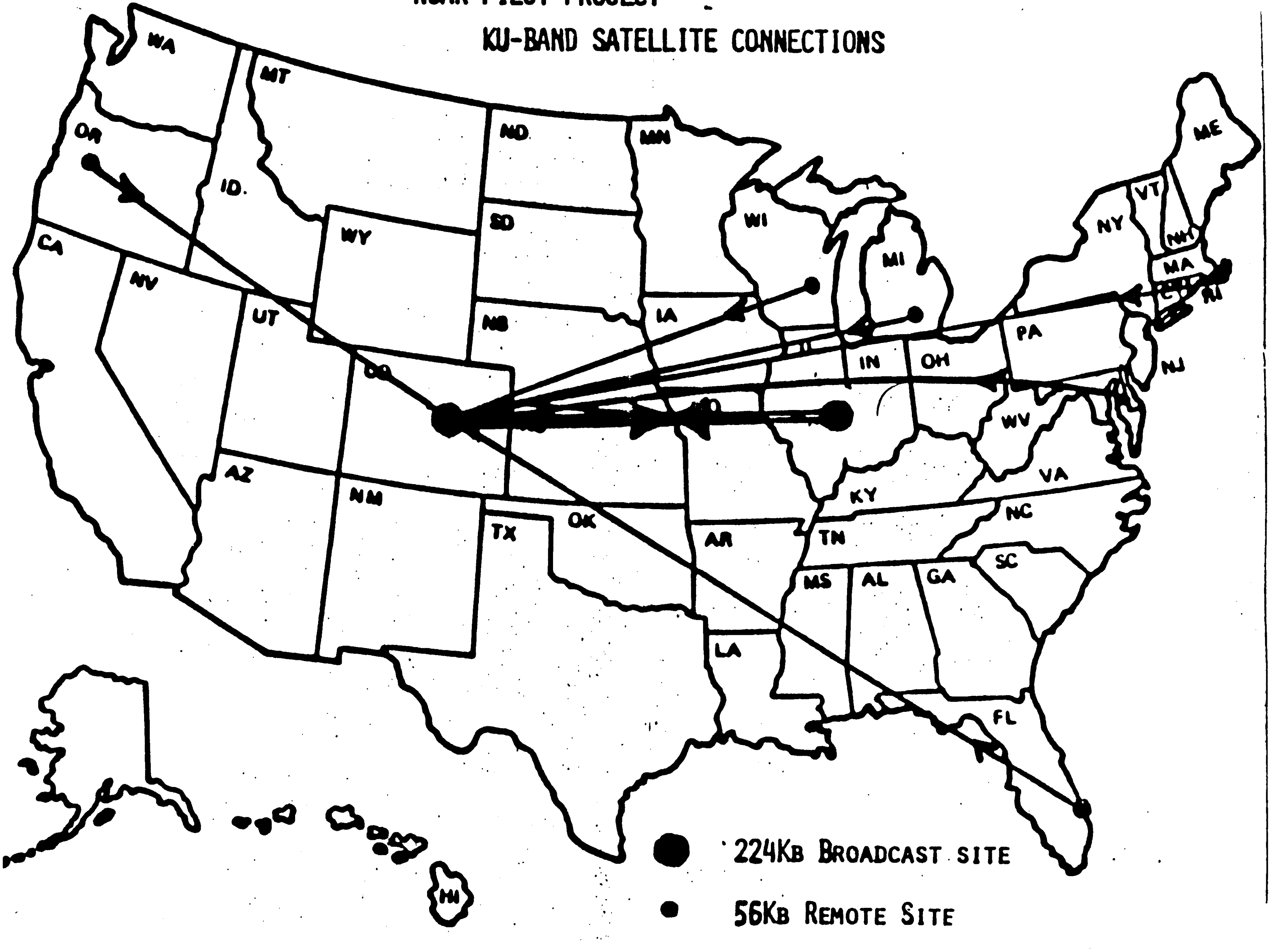
# NSFNET
# SAN DIEGO CONSORTIUM NETWORK



N.B. APPROACH

- TERMINALS TO INTERACTIVE MAINFRAME TIME SHARING SYSTEM

## NCAR PILOT PROJECT
## KU-BAND SATELLITE CONNECTIONS



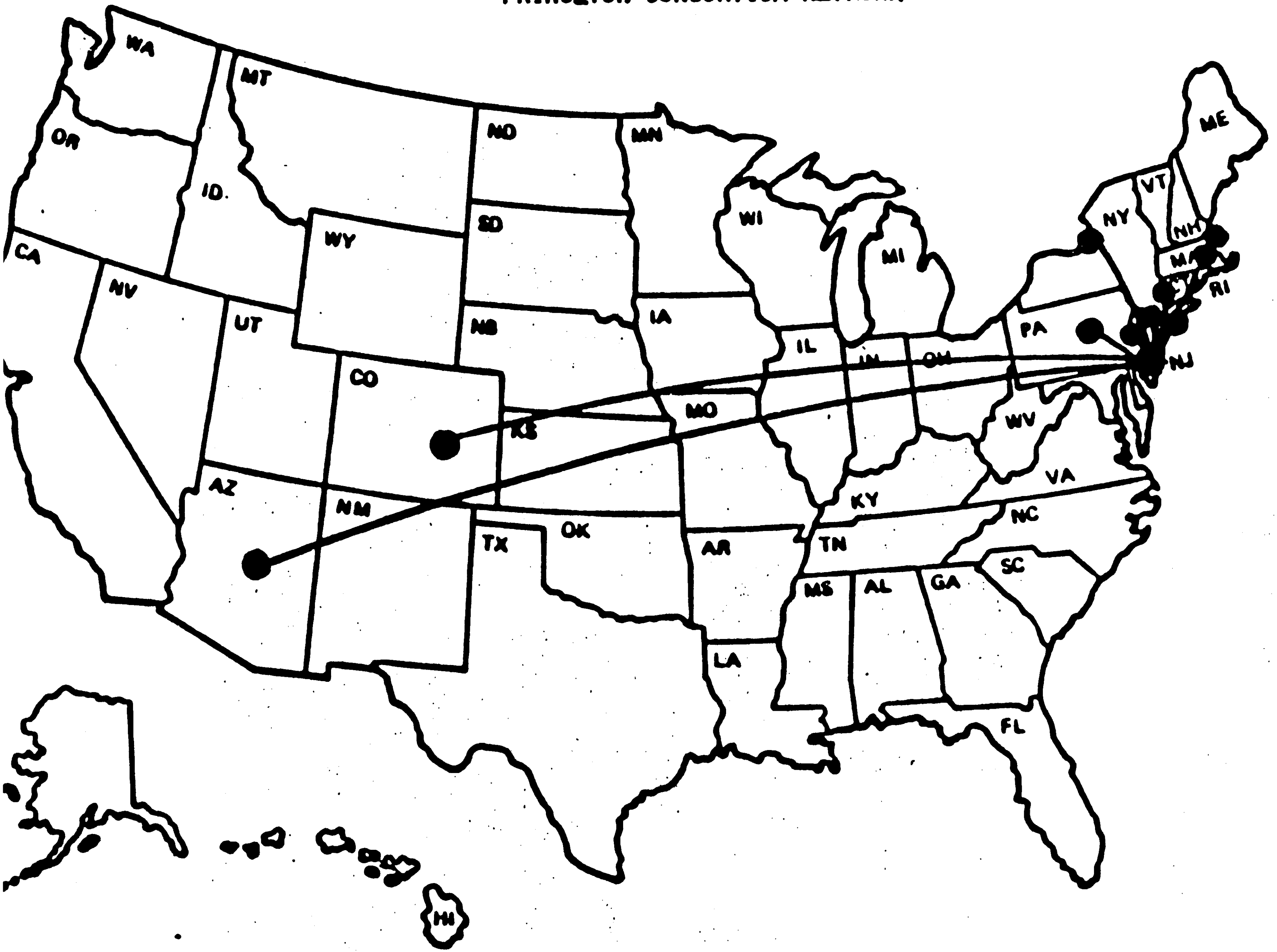
NSFNET

## CONSORTIA NETWORKING

PRINCETON

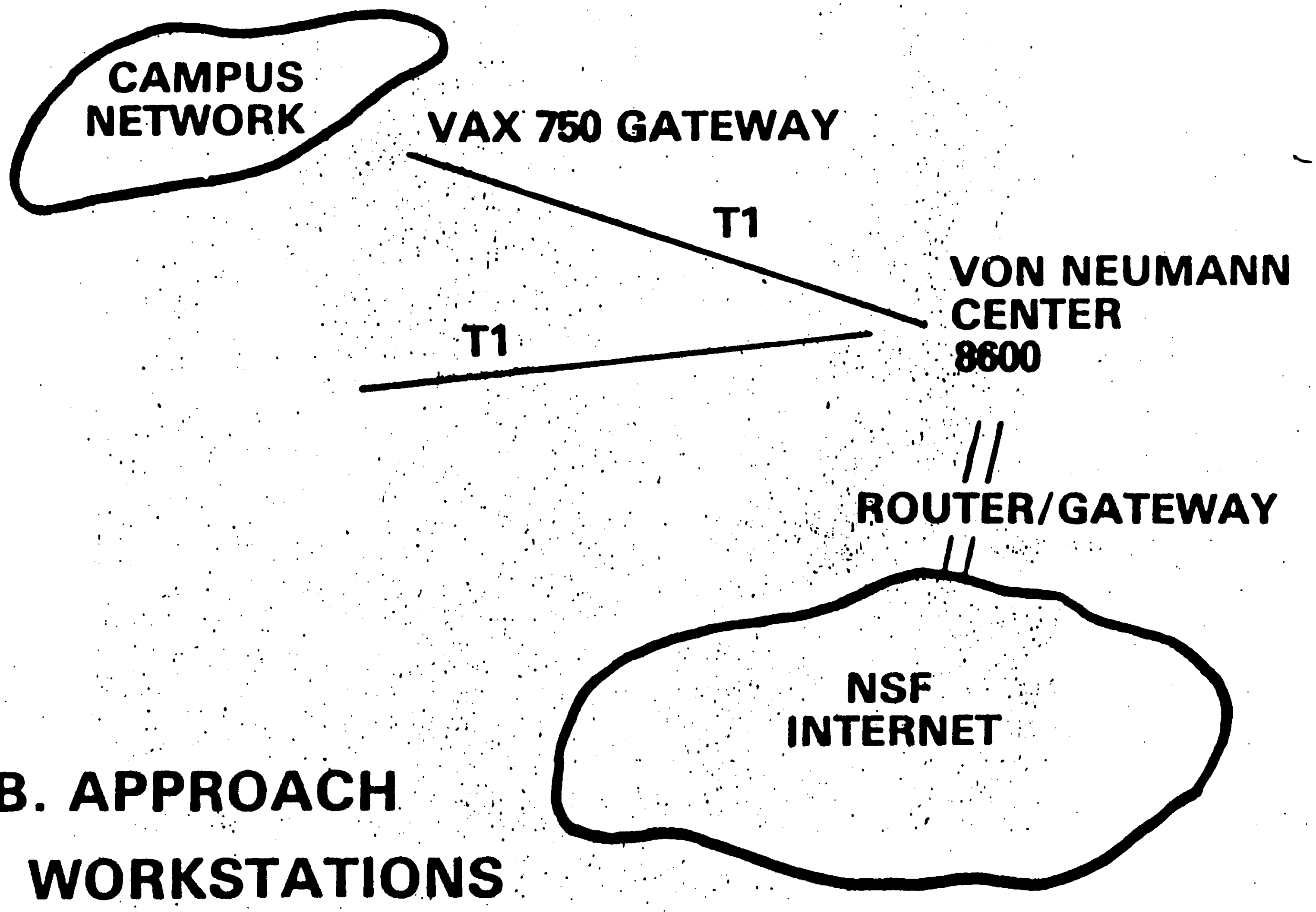
- ARIZONA
- BROWN
- COLORADO U.
- HARVARD
- INSTITUTE FOR ADVANCED STUDY
- MIT
- NYU
- PENN STATE
- PRINCETON
- ROCHESTER
- RUTGERS
- COLUMBIA
- PENNSYLVANIA.

# PRINCETON CONSORTIUM NETWORK



# NSFNET

# PRINCETON CONSORTIUM NETWORK



CAMPUS NETWORK

VAX 750 GATEWAY

T1

T1

VON NEUMANN CENTER
8600

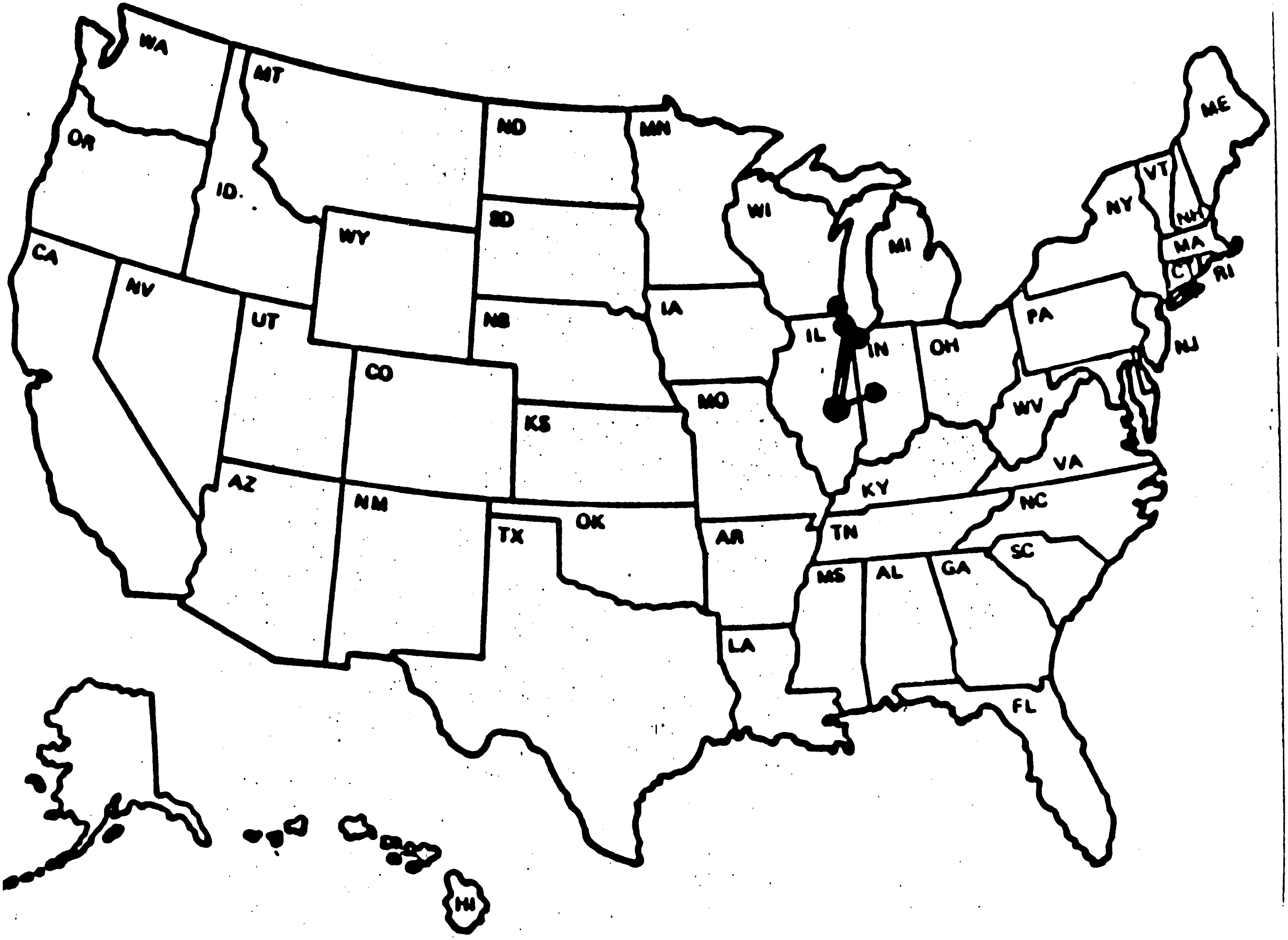
ROUTER/GATEWAY

NSF INTERNET

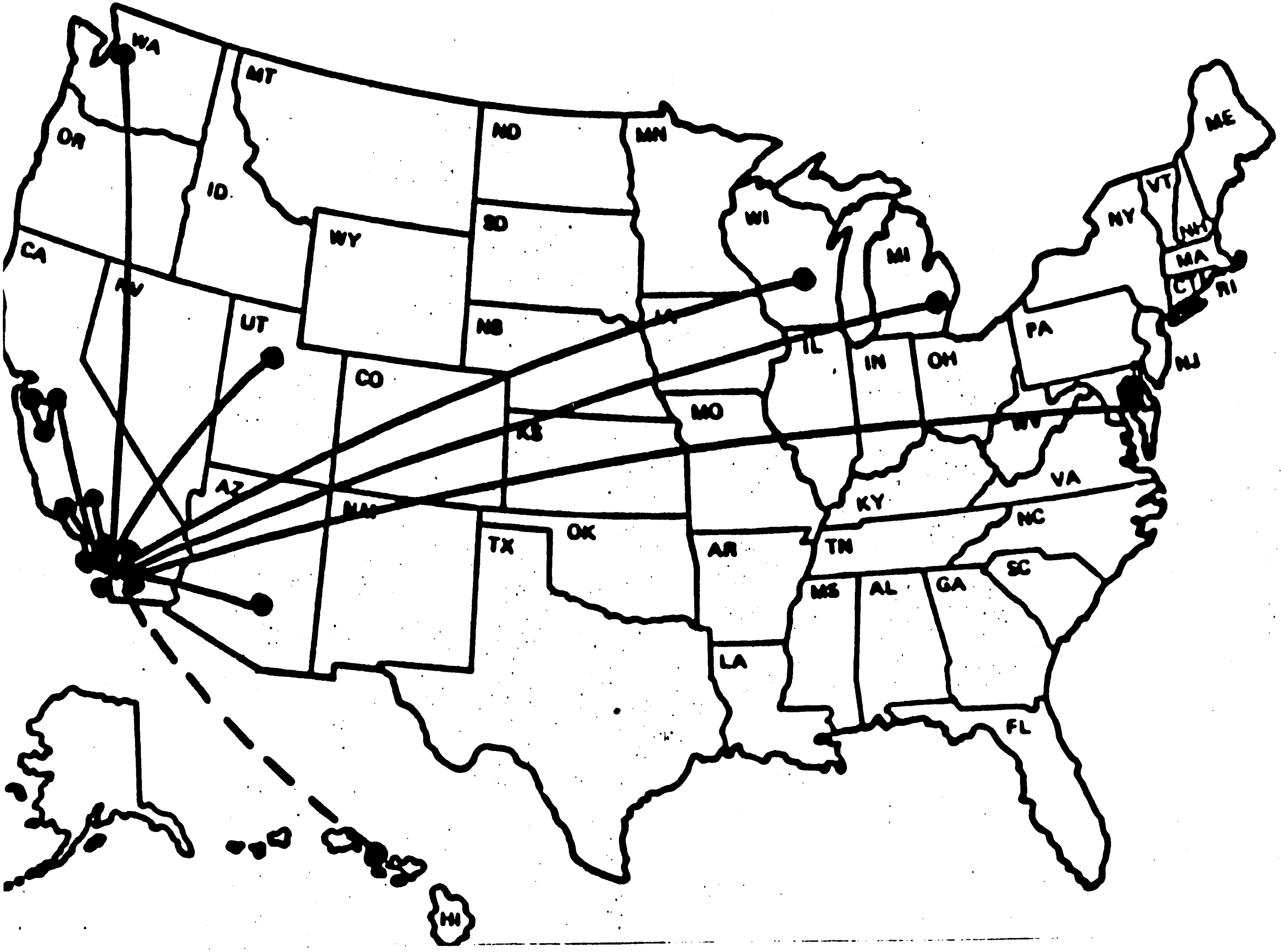
## N.B. APPROACH

- WORKSTATIONS
- POWERFUL F/E FILE & COMMS. SERVERS
- SUPERCOMPUTERS AS CYCLE SERVER
  (BATCH MACHINES RATHER THAN INTERACTIVE)

# ILLINOIS CENTER NETWORK



# SAN DIEGO CONSORTIUM NETWORK



## NSFNET

## COMMUNITY NETWORKS

- ARPANET
- BITNET
- CSNET

| SUPERCOMPUTER CENTER | X.25 ~~TELENET~~ | DIAL-UP | BITNET | ARPANET |
|---|---|---|---|---|
| BOEING | ✓ | ✓ | ✓ | |
| MINNESOTA | ✓ | ✓ | (✓) | ✓ |
| PURDUE | ✓ | ✓ | (✓) | ✓ |
| AT&T | [illegible] | ✓ | | |
| CSU | ✓ | ✓ | ✓ | |
| DIGITAL | | | | |
| CORNELL | [illegible] ✓ | ✓ | ✓ | ✓ |
| PRINCETON | PLANNED | ✓ | ✓ | |
| ILLINOIS | ✓ | ✓ | ✓ | (✓) |
| SAN DIEGO | TEST (TYMNET) | ✓ | | |
| (PITTSBURGH) | .. | | ✓ | |
| NCAR | UNINET | ✓ | | |

# EXISTING ARPANET SITES

BERKELEY
CALTECH
CMU
COLUMBIA
CORNELL *
DELAWARE **
HARVARD
ILLINOIS *
LBL
U.C. LOS ANGELES
MINNESOTA *

MIT
PENNSYLVANIA (WHARTON)
PURDUE **
ROCHESTER
RUTGERS *
STANFORD
TEXAS, AUSTIN
UTAH
WASHINGTON **
WISCONSIN **
YALE *

SFNET

VERALL NETWORK STRATEGY:

"INTERNET"

- COLLECTIONS OF NETWORKS WITH THE SAME ADDRESSING STRUCTURE, AND THE SAME PROTOCOLS

GOAL: ALL RESOURCES ADDRESSABLE, IN THE UNIFORM FASHION, ACCROSS THE COLLECTION OF NETWORKS, FROM WITHIN THE USERS OWN COMPUTING ENVIRONMENT.

# ARPANET EXPANSION: 1ST PHASE

AT&T
BOEING
CALTECH
CMU/PITT*
CSU
CORNELL *
CUNY
ILLINOIS *
INDIANA
U. C. LOS ANGELES
MARYLAND
MICHIGAN
MINNESOTA
NSF
NCAR
NORTH WESTERN
OHIO STATE
PRINCETON
PURDUE *
SANTA BARBARA
SAN DIEGO
WASHINGTON
WISONSIN *

# STANDARDS

AN INTERNETWORKING STANDARD REQUIRED

DECISIONS:

INTERIM
- DOD INTERNET
- TCP/IP
- APPLICATIONS

GOAL
- ISO/OSI

A MIGRATION STRATEGY WILL BE REQUIRED.

# NSFNET

## NEXT ACTIONS

- RFI/RFQ/SOLICITATION
  - NETWORK
  - SWITCHES/GATEWAYS

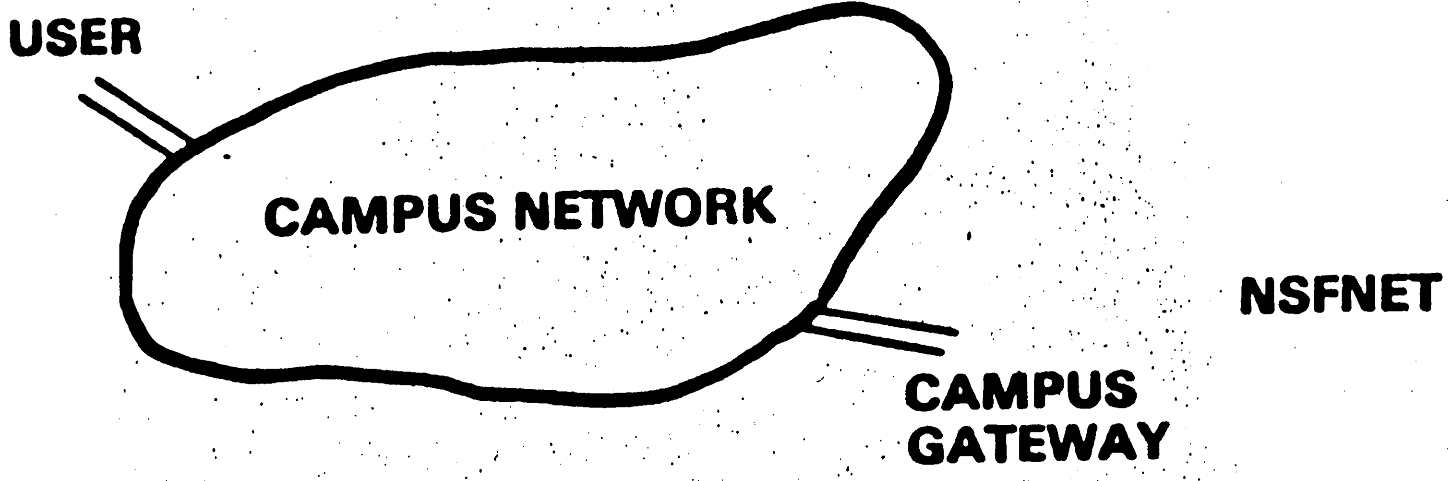
> STUDIES

- RFQ/SOLICITATION
  - NET MANAGEMENT
  - NET OPERATIONS
  - NET USER SERVICES

PRODUCED USING BY LEHMANN

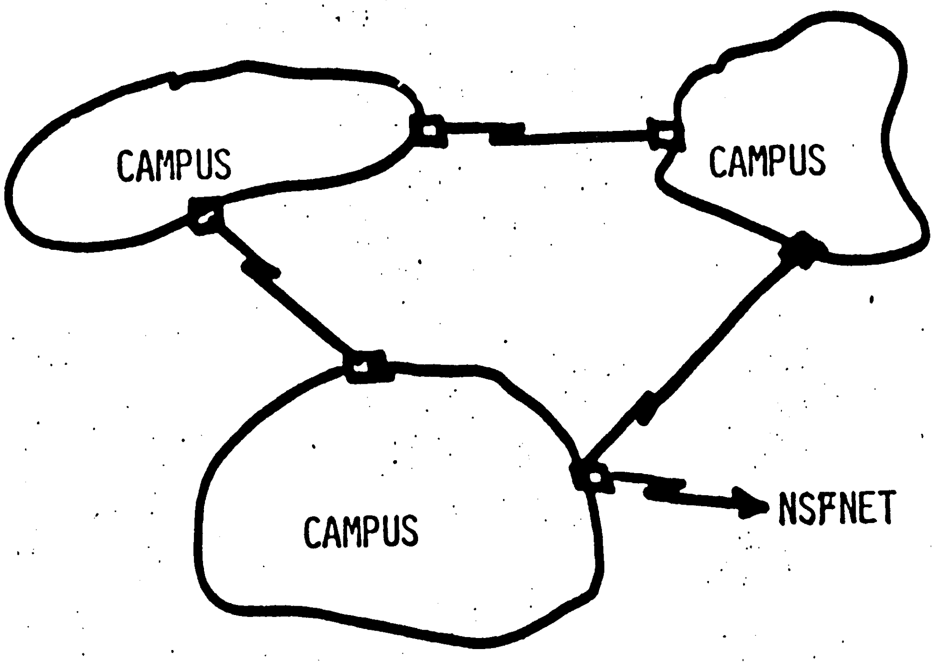
AGOURON INSTITUTE
ARIZONA
AT&T
BERKELEY
BOEING
BROWN
CALTECH
CMU
COLORADO U.
COLORADO STATE
COLUMBIA
CORNELL
CUNY
DELAWARE
HARVARD
HAWAII U.
IAS
ILLINOIS
INDIANA
KITT PEAK
LBL
MIAMI
MARYLAND
MICHIGAN
MINNESOTA
MIT
NCAR
NSF

# NSFNET
# CAMPUS NETWORKS

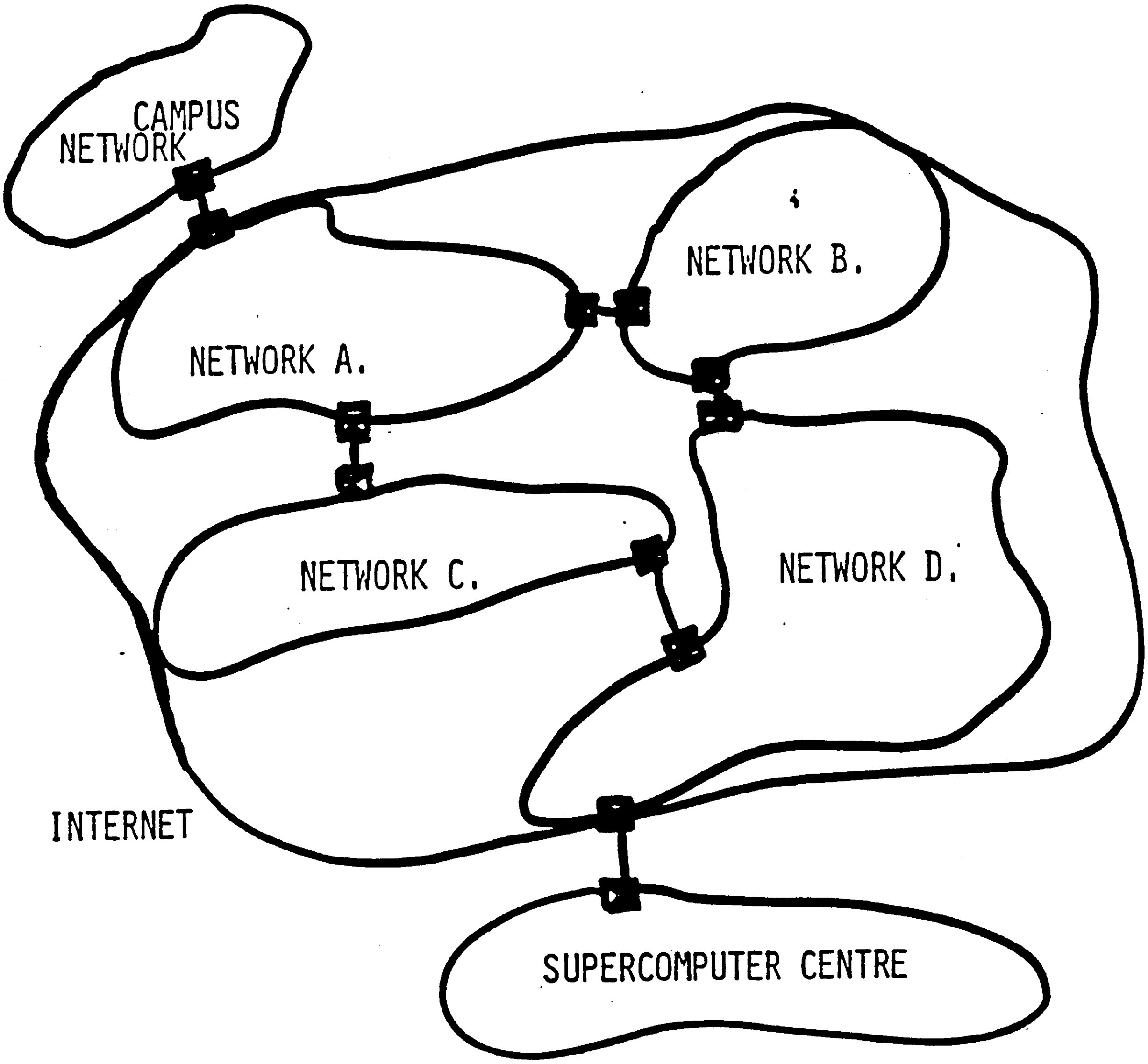


- ENCOURAGING CAMPUS NETWORKS
  - SERVICE ORGANIZATION FOR NETWORKING
  - SUPPORTS USERS
  - PROVIDES GATEWAY

# STATE NETWORKS



## NSFNET MODEL



ORTH WESTERN
YU
HIO STATE
REGON STATE
ENN U.
ENN STATE
ITTSBURGH
RINCETON
URDUE
OCHESTER
UTGERS
ANTA BARBARA
ALK INSTITUTE
AN DIEGO U.

SAN DIEGO STA
SCRIPPS CLINI
SAN DIEGO SCR
SW FISHERIES
SAN FRANCISCO
STANFORD
TEXAS, AUSTIN
UCLA
UTAH
WASHINGTON
WESTINGHOUSE
WISCONSIN
YALE

SDSC STEERING COMMITTEE
JANUARY 13, 1986
BOTTOM

# NSF NATIONAL SUPERCOMPUTER CENTERS

- John Von Neumann Center
  (Princeton)
- National Center for Supercomputing Applications (Ill.)
- Production Supercomputer Facility (Cornell)
- San Diego Supercomputer Center
  (U. of Calif. @ San Diego)

# NSF ACCESS TO SUPERCOMPUTERS

- AT&T Bell Laboratories
  Cray X-MP/24
- Boeing Computer Services
  Cray X-MP/24
- Colorado State University
  Cyber 205
- Digital Productions
  Cray X-MP/22
- University of Minnesota
  Cray 2, Cyber 205
- Purdue University - Cyber 205

|  | SDS | ILL. | JVNC | CORNELL |
|---|---|---|---|---|
| FINANC | APP. | APP. | UNDER REVIEW | N/A |
| CONSTR RENOV | NEW BLDG PRTL CMPLT | RENOV PRTL CMPLT | NEW BLDG UNDER CONSTR | RENOV CMPLT |
| HRDWRE INSTLD | 11/85 | 8/85 | 3/86 | 11/85 |
| LMTD SRVC | 12/85-<br>1/86 | 9/85-<br>12/85 | 12/85-<br>4/86 | 5/85-<br>10/85 |
| FULL SRVC | 1/86<br>~~2~~/86 | 1/86 | 5/86 | 11/85 |

# CAPABILITIES OF THE NATIONAL SUPERCOMPUTER CENTERS

| | SAN DIEGO | ILLINOIS | VON NEUMANN | CORNELL |
|---|---|---|---|---|
| SUPERCOMPUTER | CRAY X-MP/48 | CRAY X-MP/24 | CYBER 205 | IBM 3084 QX<br>W/5 FPS 264<br>& 1 164 W/MAX |
| OPERATING SYSTEM | CTSS | CTSS | VSOS | VM/CMS |
| DATA STORAGE | 8 MW MEMORY<br>10 GBYTES DISK | 4 MW MEMORY<br>32 MW SSD<br>8 GBYTES DISK<br>55 GBYTE MASSTOR | 4 MW MEMORY<br>6 GBYTES DISK<br>40 GBYTES ON 8600'S | 16 MW MEMORY (3084)<br>2 MW EA. 264<br>5.1 GBYTES TOTAL DISK |
| COMMUNICATIONS AND FRONT-END PROCESSORS | PDP 11 (1) | VAX 11/785 (1) | VAX 8600 (4) | GOULD 9000 (1) |
| AVAILABLE MACHINE HOURS PER YEAR | 30,000 | 15,000 | 7,500 | 7,000 CRAY X-MP/1 EQUIVALENT HOURS |
| TYPES OF NETWORK CONNECTIONS | ARPANET<br>BITNET<br>TELENET<br>MFENET<br>56 KB TO CENTERS & CONSORTIA | ARPANET<br>BITNET<br>TELENET<br>56 KB TO CENTERS, NCAR, & REGIONAL UNIVERSITIES | APRANET<br>BITNET<br>TELENET<br>56 KB TO CENTERS<br>T1 LINES TO CONSORTIA | APRANET<br>BITNET<br>TELENET<br>56 KB TO CENTERS<br>NYSERNET |
| FULL SERVICE DATE | 6/1/86<br>1/1/86 | 1/1/86 | 6/1/86 | 10/31/85 |
| PLANNED UPGRADES | UPGRADE UNDER REVIEW | X-MP/48 & 132 MW SSD (10/86) | ETA 10 (3/87) | IBM 3090/400 W/ 4 VECTOR FACILITIES (10/86) |
| RESEARCH PROGRAM/ CENTER | | INTERDISCIPLINARY RESEARCH CENTER | | THEORY CENTER RESEARCH INSTITUTE |

# SUPERCOMPUTER ALLOCATIONS FOR PHASE II

- Open to US Researchers, if
  - o Computations are Deemed of Scientific Merit
  - o Research Accomplishments are Accessible by Community
- Exception: Limited Amount for Proprietary (non-secure) Research (up to 10% of total)

# SUPERCOMPUTER SUMMER INSTITUTES

- 3 Institutes (2-4 weeks long)
  - Boeing Computer Services
  - University of Minnesota
  - NCAR
- 90+ Attendees
- $500,000 Support
  - $220K OASC Centers Program
  - $250K DOD
  - $ 30K NSF Phys. Ocean. Prog.

# PROCESS ENCOURAGES

- Researchers to link Requests with Proposals Submitted to NSF
- Requests for Start-Up Research
- Requests for Education & Training
- Requests from Researchers NOT Funded by NSF
- Coordination within Institutions

# OVERSIGHT OF ALLOCATION PROCESS

- OASC Advisory Committee will annually review the entire allocation process,

- Large allocations (>250 SU's) within both the NSF & Center's share will be discussed at each OASC Adv. Cmt. meeting,

- Balance between disciplines will also be reviewed at each OASC Adv. Cmt. Meeting.

- Boeing Computer Services (4 Wks)
  - o Partic. Optim. Working Code
  - o Focus on Comput. Physics, Fluid Dynmics, Control Sys., Modeling in Life Sciences
- University of Minnesota (4 Wks)
  - o Access to Workstations, Cray-1, Cray 2, Cyber 205
  - o Lectures in Polymer Modeling Molecular Dynm., Astrophys.

# DISTRIBUTION OF
# EACH CENTER'S RESOURCES

- 60% to NSF for Allocation by Program Directors
- 40% to Center for Allocation by Multidisciplinary Panel
  - Up to 25% of Center's Share Available for Proprietary Computational Research
  - Block Allocations to Institutions Permitted for Start-Up & Training

# PLANS FOR FY 1986

- 4 to 6 Summer Institutes
- Open Competition
- Seek Other Agency Support
- 2 Levels
    - o Novices on Supercomputers
    - o Advanced & Specialized Applications

- National Center for Atmospheric Research (2 Wks)
  - o Emphasis on Atm. Sci, Oceanography, Solar Physics
  - o Access to Cray-1's & Boeing Cray **X**-MP

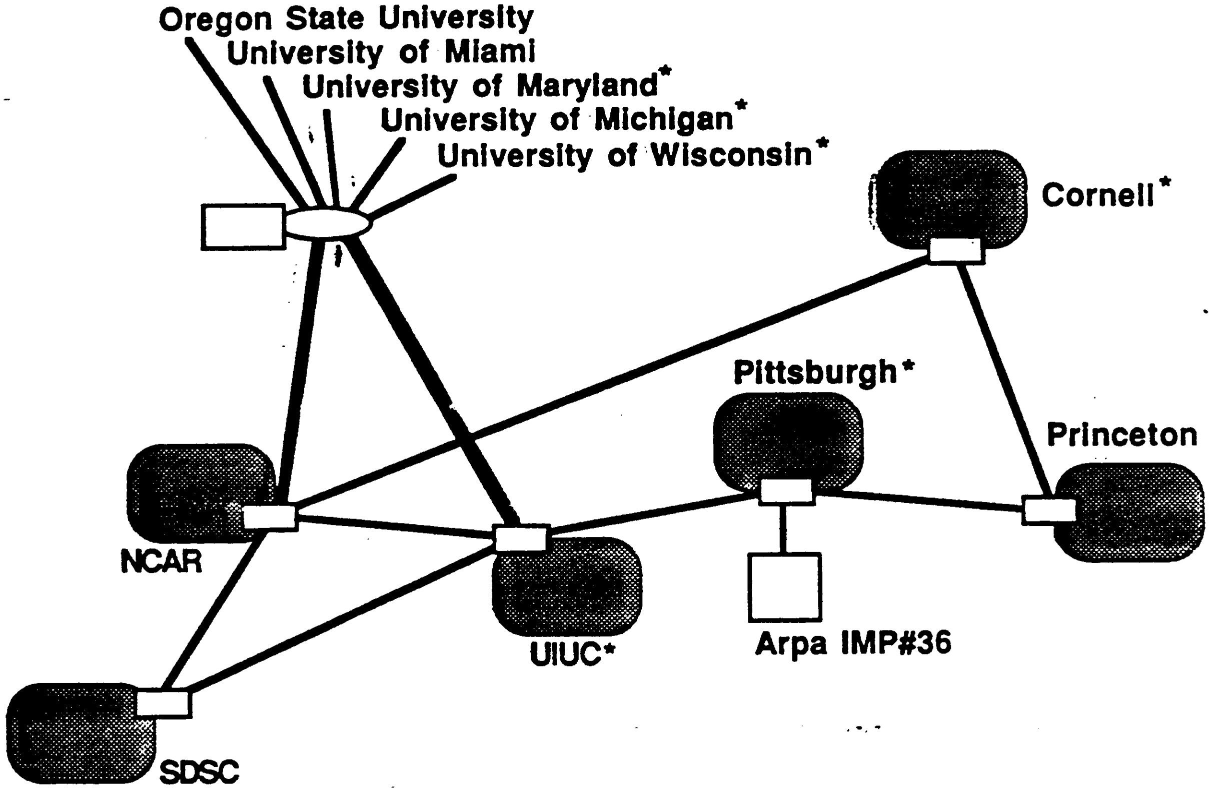
# CENTER ALLOCATION PLANS

| | SDSC | Illinois | JVNC | Cornell |
|---|---|---|---|---|
| Processors | 4 | 2 | 1 | 4 |
| Avail. service units | 30,000 | 15,000 | 7,500 | 7,000 (CRAY X-MP single proc. equiv) |
| NSF/Ctr | 60/40 | 60/40 | 60/40 | 60/40 |
| Center specific features | Priorities<br>o 3,800 su's to consortia<br>o bundled block req from consort & non-consort encouraged<br>Other<br>o requests to SDSC > 200 su's refer to NSF | Priorities<br>o breakthrough<br>o IRC<br>o comput that push facility<br>o education<br>o requests that fall through cracks | Features<br>o 2,600 su's to consortia<br>o thru 4/86 no 60/40<br>o 100 hour + requsts to NSF<br>o 100% over-allocation<br>o thru 8/86 mnthly alloc<br>o quarterly alloc after 9/86 | Features<br>o all requests to NSF for peer rev except:<br>- second stage req ie, time beyond NSF allocation<br>- exceptional cases eg, interdis. or other agency<br>Reviewed by Alloc. Subcommittee along w/ allocations by NSF > 500 su's |

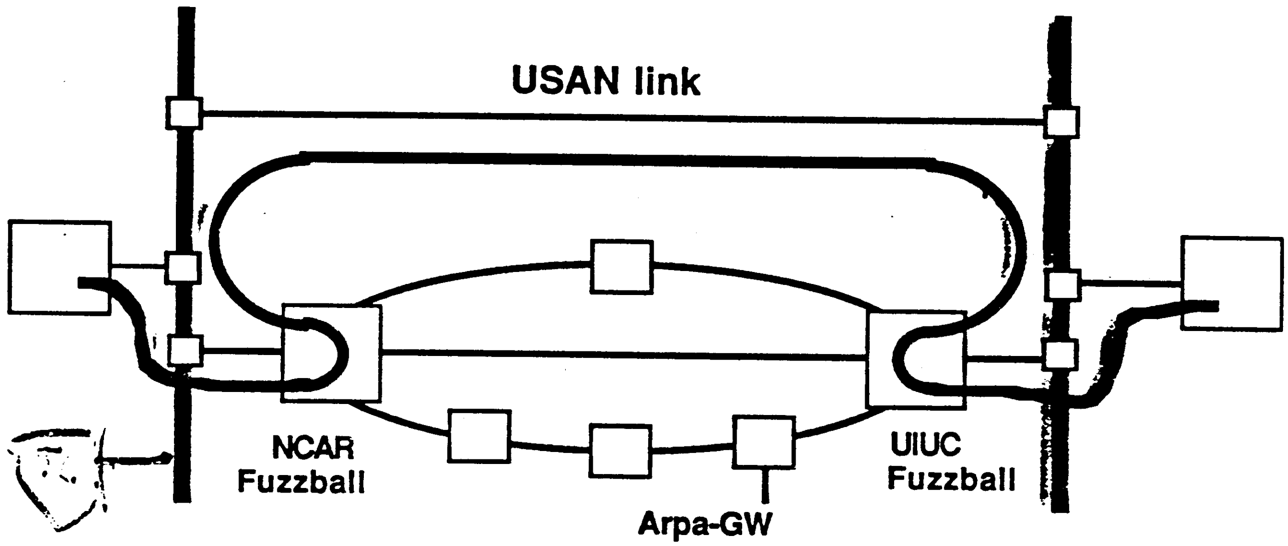
Projection Screens
As Speaker Faces Audience

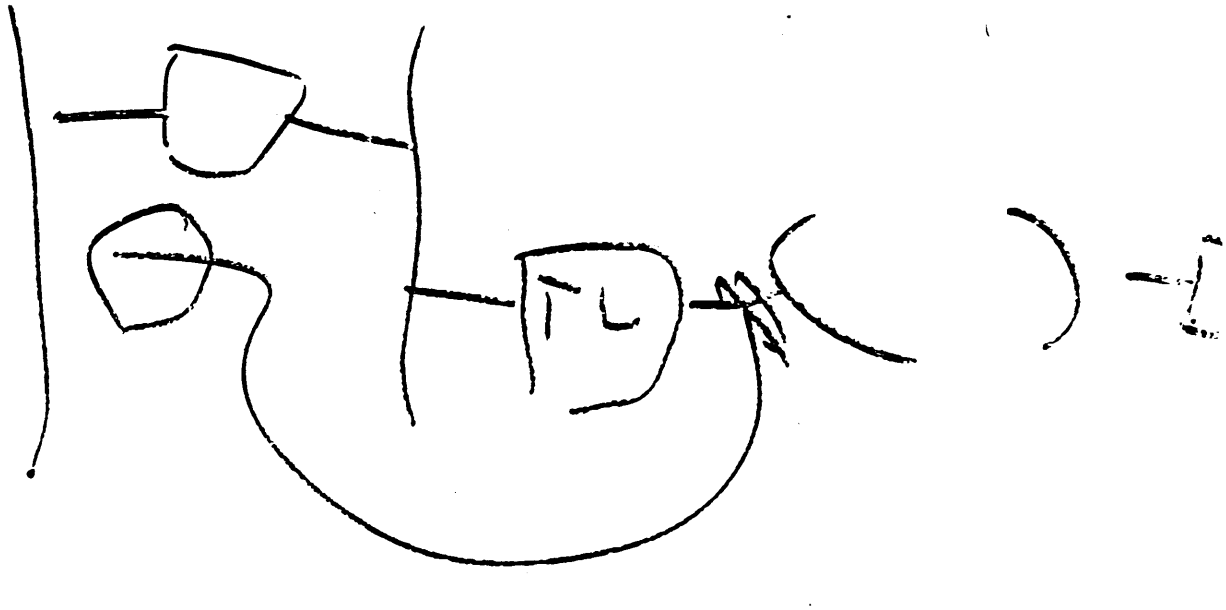
| | SDSC | ILL. | JVNC | CORNELL |
|---|---|---|---|---|
| STAFF | 50/65 | 46/50 | 13/47 | 43/60 |
| ALLOC NSF/CTR | 325/0 | 890/0 | 75/0 | 275/ 1800^ |
| ALLOC CMMTE | QTR 1/86* | QTR 2/86* | QTR 12/85* | QTR 1/86* |
| MAJOR DONAT (NON- MNFRME | SCS40<br>8MW<br>$1.0M<br>6/86 | WORK STATNS<br>$1.5M | NEGOT. W/ VENDORS | FPS164<br>(2)<br>$1.0M |

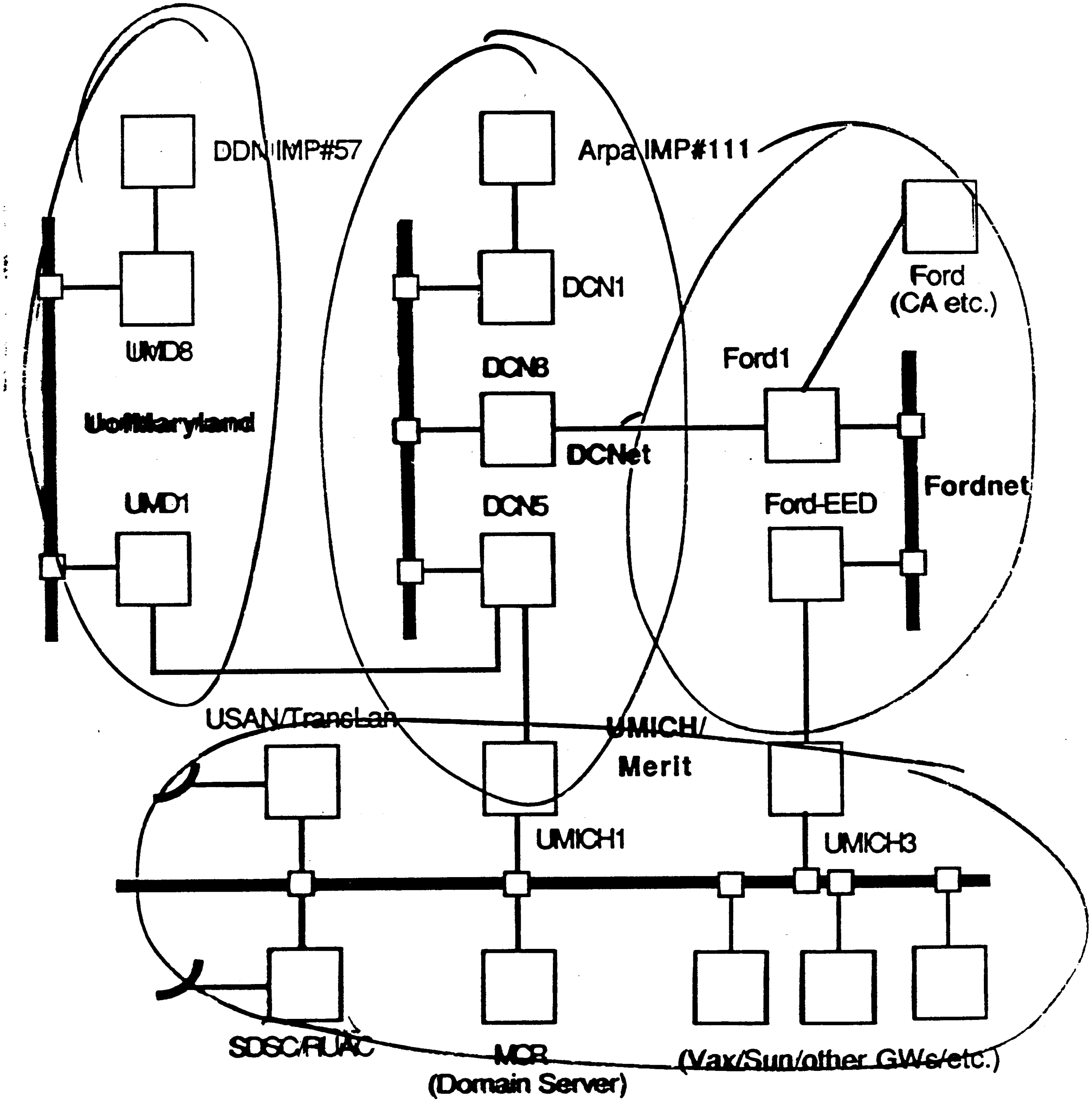
*FIRST MEETING
^TRANSFER FROM THE INTERIM FACILITY


Oregon State University
University of Miami
University of Maryland*
University of Michigan*
University of Wisconsin*
Cornell*
Pittsburgh*
Princeton
NCAR
UIUC*
Arpa IMP#36
SDSC

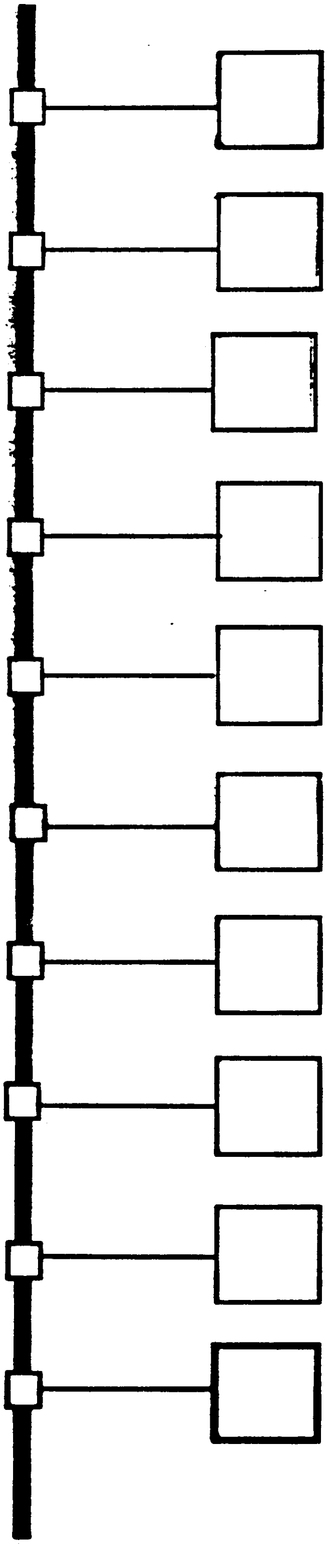
# Fuzzballs as routing agents with ICMP-Redirect avoidance







(Other anticipated UMICH additions include a direct Arpa-IMP connection as well as a link to CICnet and to CMU (via Case Western).)

<br>
GW.UMICH.EDU (link to DCNet)

UMICH2 (IBMPC (Ethernet Monitor))

UMICH3 (link to Fordnet and packet radio subnet))

UMICH4 (Merit SCP and 23 bit wide subnet))

UMICH5 (Sun3)

SATURN.UMICH.EDU(gateway to two concatenated rings of about 140 Apollo workstations)

MCR.UMICH.EDU (Sun3, Domain Name Server)

UMICH9 (for packet radio and NSFnet experiments, linked via Ethernet, HDLC link to UMICH1 and packet radio channel to UMICH3)

MADVAX.UMICH.EDU (Vax)

Anticipated:

. Direct Arpanet (IMP) connection
. USAN TransLan
. SDSC RUAC
. Appletalk/Ethernet gateway (tested)
. Gateway to the EECS department subnet
. other PCs

# IP TTL

- GATEWAY OUTPUT QUEUE LENGTH IS BOUNDED, BUT ONLY IF IP DATAGRAMS HAVE A FINITE TIME TO LIVE (TTL).
- THE UNIT MEASURE FOR IP TTL IS SECONDS; FOR TCP, TTL IS CURRENTLY SPECIFIED AS 60 SECONDS.
- GATEWAYS CURRENTLY USE TTL ONLY AS A HOP COUNT. THUS, QUEUE LENGTHS ARE UNBOUNDED.
- GATEWAYS SHOULD COUNT DOWN TTL. (THIS WILL NOT TAKE INTO ACCOUNT SUBNET TRANSIT TIME; THUS, TTL WILL BE A LOWER BOUND ON DATAGRAM LIFETIME.)

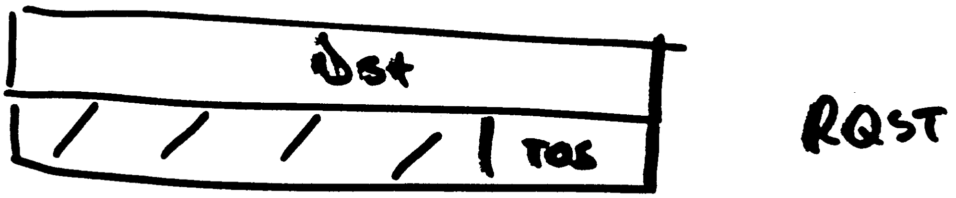
- ALLOWING GATEWAY QUEUE LENGTHS TO GROW UNNECESSARILY CONTRIBUTES TO A WIDE VARIANCE IN ROUND-TRIP TIME
- FOR A TCP SEGMENT, THE USEFUL DATAGRAM LIFETIME IS OFTEN MUCH LESS THAN 60 SECONDS. ASSUMING A 5-SECOND RETRANSMISSION TIMER, FOR EXAMPLE, 12 RETRANSMISSIONS COULD OCCUR DURING A 60-SECOND PERIOD
- AS A RESULT, TCP MAY NEED TO COMMUNICATE "USEFUL DATAGRAM LIFETIME" TO IP, ON A PER-SEGMENT BASIS, RATHER THAN USING A CONSTANT 60-SECOND VALUE.

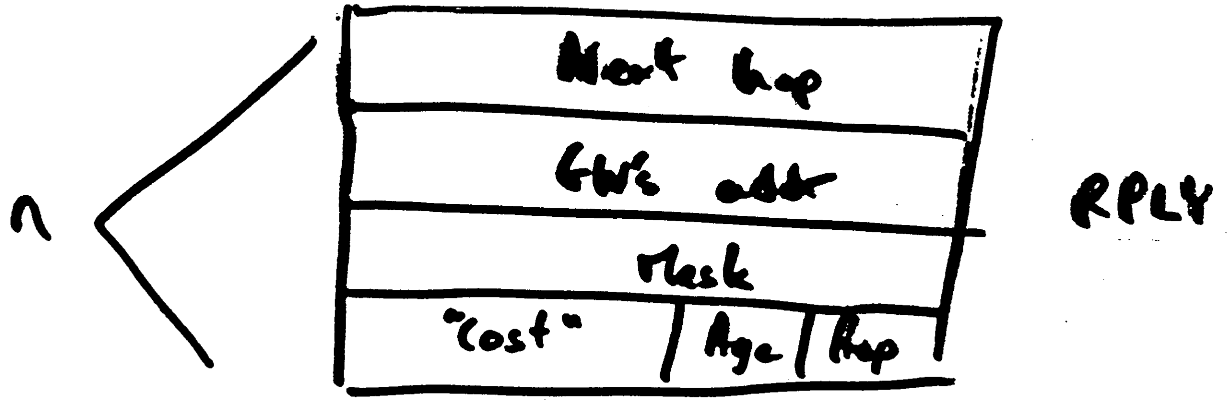
- WE SHOULD LEARN FROM DECNET'S EXPERIENCE: A VARIANCE OF AT LEAST 4 TO 5 SHOULD BE EXPECTED WHEN TRANSMITTING ACROSS ONE OR MORE WIDE-AREA NETWORKS.
- THIS CONFLICTS WITH CURRENT TCP IMPLEMENTATIONS, WHICH TYPICALLY ASSUME A VARIANCE OF 1.5 - 2.
- NOTE THAT VARIOUS SCHEMES FOR DELAYING TCP ACK'S ALSO CONTRIBUTE TO VARIANCE IN ROUND-TRIP TIME.

Use appropriate address.

- Broadcast
- Multicast for GW's

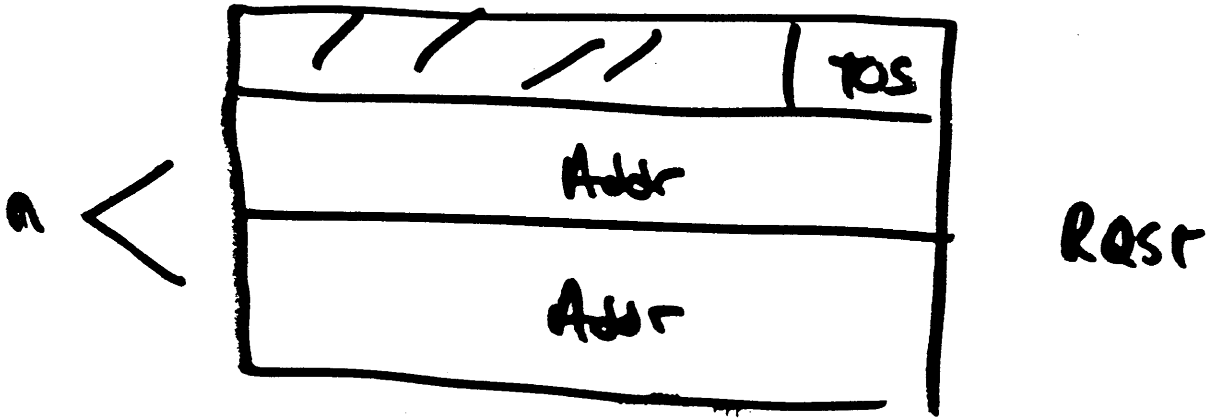
a) Find GW

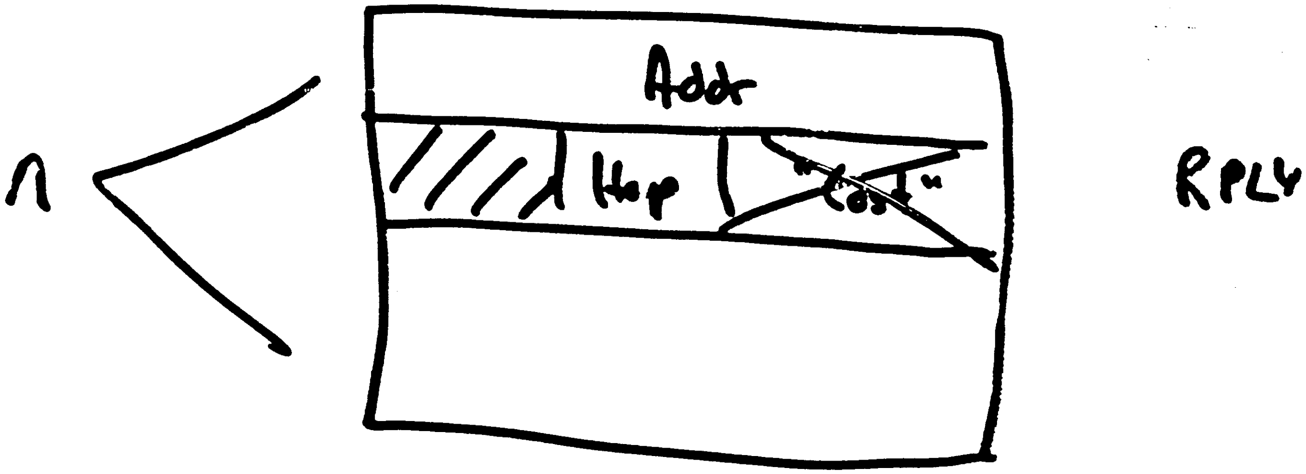




b) Find GW Next hop

- Last prefix or first
- [illegible] OK
- ID (port)
- Other route used





- If dest on lcl, ask host
- Pm to get redir to best lcl gw?

c) Best Host Address

1) New ICMP messages

   a) Find GW
   b) Next hop
   c) Best host address

2) Dead GW detection

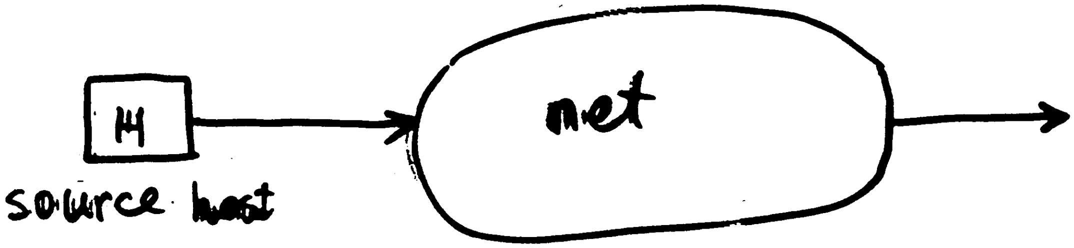
   - No Pings !! (unless 2 counters)

3) Per Host / TOS Redirect only

4) Dumb hosts

5) General comments on proto - RFC?

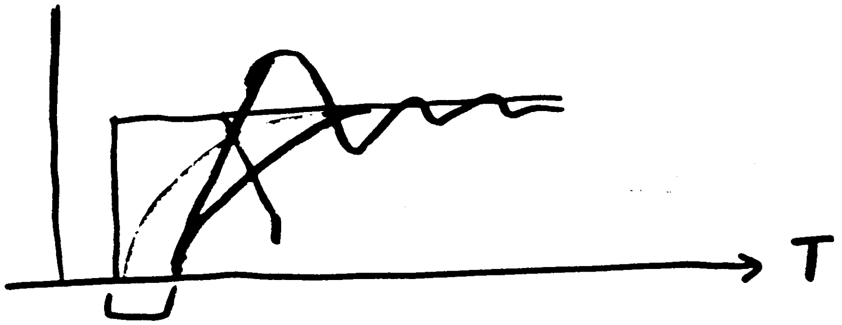
# Limitations of IP congestion fix



a feedback control system

Limitation of feedback control system:

- system response time
  (control delay)



- require load change cycle >> response time

Network load dynamics:

a function of {
- transfer duration
- fluctuation
- speed

Network control response time:

depend on
- averaging period length
- net transmission delay

Do we know both ?

Traffic measurement:

How long do most host-host transfers last?

How is a congestion created?

Gateway load histogram:

- How heavily is it loaded?
- How fast does the load change

Network delay

(stop here?)

If data transfers are too short to control,

- give up control
  identify bottlenecks and add more
  cpu power or bandwidths

- make major changes to IP

# Requirements to IP congestion fix

- It must be able to survive till the next generation of internet protocol.
- It must be simple while effective, must work well when piecemeal changes going through the net
- It must not reduce the robustness of current internet.
- It must be fair.

# Host side requirement

- The control must be at IP level, in order to control all traffic.
- Window scheme cannot be used
- Add a patch to the current IP functionality.
- No overhead when there is no congestion.

# Gateway side requirements

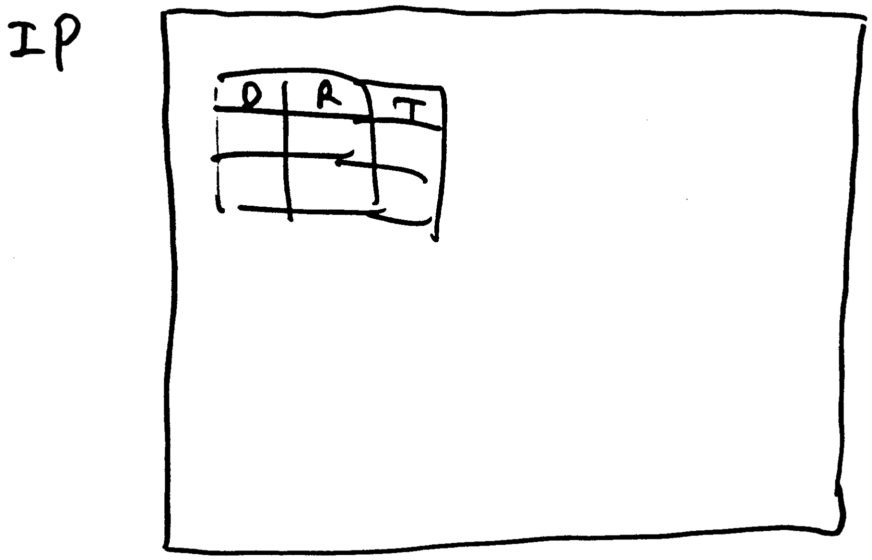
- The control messages must bring specific control information to the host on what to do.
- Hosts should not be allowed to self-backup.
- The control hasto be in rate
- The gateway should selectively punish offending hosts.

# Host changes

IP source quench

| transfer rate |
|---|
| expiry time |

~ pkts/sec

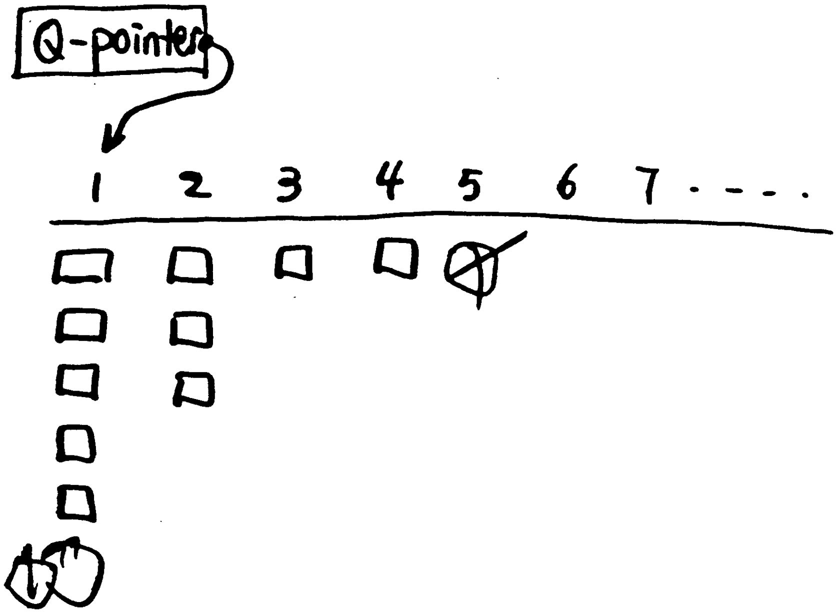


Record IP source quench messages received

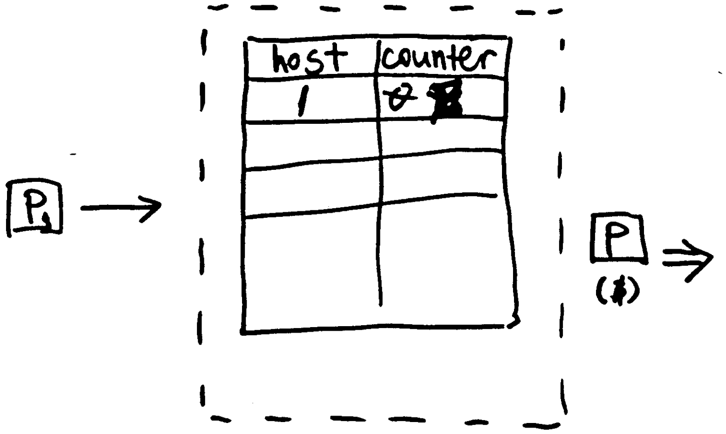
Limit transfer rate to the quenched destination(s)

Erase the quench when it expires

# Control at gateway



each source host can only put one packet at each queue



control box

- The function of this queue structure
- How each pkt is put into queue
- How to compute the control rate & expiry tim
- renew control messages when needed
-

# Work to do

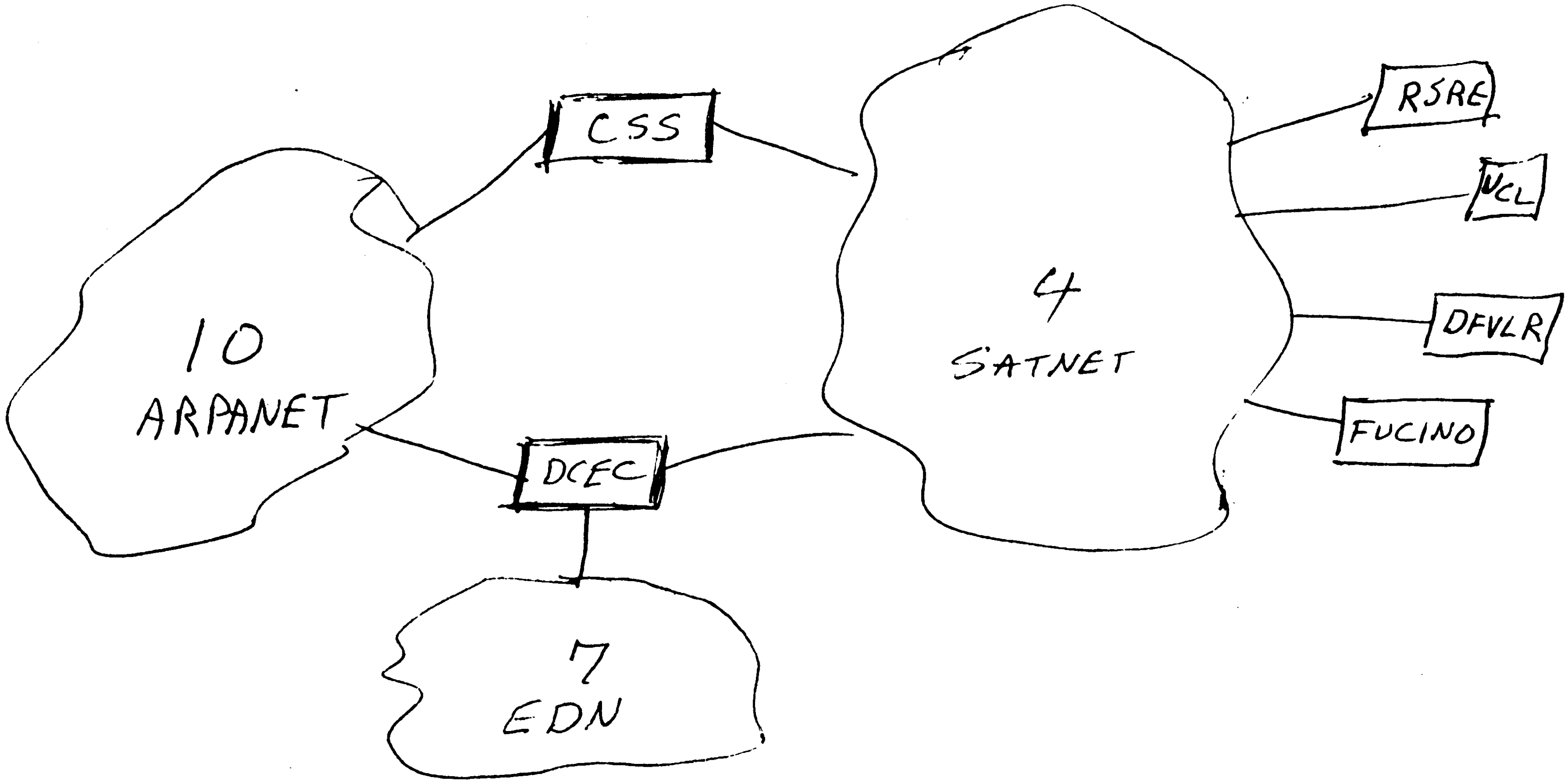
- traffic measurements
- control experiments

# PSN 6 CRASH

IMP:  CRASH AT 133175
RA = 11
CODE = 20

## HDLC I-FRAME WITH NULL I-FIELD

```
T    2    01    P    SABM
R    2    01    P    UA
R    7    03         IFRAME N(S)=0 N(R)=0 10 00 FB 07 82
T    2    01         IFRAME N(S)=0 N(R)=0
T    2    01    P    DISC
T    2    01    P    SABM
```



SATNET ROUTING BUG

Internet Engineering Task Force
Request for Comments:
Amends: RFC904

M. StJohns
April 1986

# Exterior Gateway Protocol
# Incremental Changes

Status of this memo:

This RFC is the first in a series of incremental changes to EGP. It describes the negotiation of versions between two EGP entities. This RFC specifies a revised standard for the DARPA and DDN communities. Gateways which implement an EGP on the ARPA-Internet must take steps to conform to this standard.

Introduction:

It has become obvious in recent months that there are some deficiencies in the current Exterior Gateway Protocol (EGP) as defined by RFC904. This RFC is the first in a series of 'band-aids' or 'hacks' to improve and extend the usefullness of the current EGP until its successor can be designed and implemented.

Each extension in this series will be designed so that it provides additional functionality for those who implement it without abridging the usefullness of those implementations that conform to RFC904.

DRAFT

## A.6 Negotiate Command/Acknowledgement

```
     0                   1                   2                   3
     0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    | EGP Version # |     Type      |     Code      |    Status     |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |    Checksum                   |       Autonomous System #     |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
    |           Sequence #          | Max Version   |  Min Version  |
    +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

| | | |
|---|---|---|
| Type | 10 | |
| Code | 0 | Negotiate |
| | 1 | Negotiate Ack |
| Status | 0 | unspecified |
| | 1 | version accepted |
| | 2 | unimplemented version |

Other proposed changes:

1. Partial updates
2. Variant formate updates
3. Local polling
4. Distance Metric

DRAFT

Changes to RFC904:

?

## 4.6 Version Negotiation

Version negotiation may take place at any time between two EGP entities, but if at all possible take place before any other traffic. Version negotiation is independent of any other traffic. To maintain compatability between versions, version negotiation messages are always in the baseline version format. In the case of EGP, this is version 2. An EGP entity must always be able to communicate using the baseline version. The version number will always be octet 1 of an EGP message.

The sequence of events for negotiating a common version follows:

1. Gateway A sends a Negotiate command to Gateway B. The Negotiate command is ALWAYS version 2. It indicates the minimum and maximum versions it is prepared to deal with.

   a. If gateway B simultaneously sends a Negotiate command to gateway A, the autonomous system numbers are compared. The gateway with the lower number becomes gateway A and must remain silent for a P3 interval before retransmitting the Negotiate command. The gateway with the higher number becomes gateway B and responds according to step 2 below.

   b. If gateway B does not support version negotiation, it returns an error indication and gateway A may not try to renegotiate versions for a P4 interval.

   c. If no response is heard from gateway B, gateway A must wait a P3 interval before retransmitting the Negotiate command. If there is no response after a P5 interval, gateway A must not try to negotiate versions for a P4 interval.

2. Gateway B sends a Negotiate Ack response to Gateway A.

   a. If the offered versions in the Negotiate command do not overlap the set of versions gateway B is prepared to deal with, gateway B returns an Ack with a status code indicating an unimplemented version. In this case, the default version (version 2) is used between these two entities.

   b. If there is an overlap between the two version sets, gateway B sends an ack with the maximum version field set to the highest common version between the two entities, and with the status code set to version accepted.

DRAFT

3. Gateway A sends a Negotiate Ack with a version accepted status.
    a. Gateway B may start using the negotiated version as soon as it receives this message. Gateway A must be prepared to handle the old version messages until it receives a new version message. As soon as it receives the new version message, it may ignore any old version message it gets.
    b. Gateway B must send some sort of message back to gateway A within a P3 interval.
4. Gateway B sends ANY message to Gateway A using the new version.
    a. Gateway A MUST wait until it gets a message in the new version from gateway B before it starts using new version messages itself.
    b. If gateway A receives a message in the old version, it retransmits the Negotiate Ack.
    c. If gateway A does not receive ANY message in a P3 interval, it retransmits the Negotiate Ack.
    d. If a P5 interval has gone by from the initial transmission of the Negotiate Ack, gateway A goes silent for two P5 intervals awaiting any message. If no message is received, gateway A may restart the negotiation from step 1.

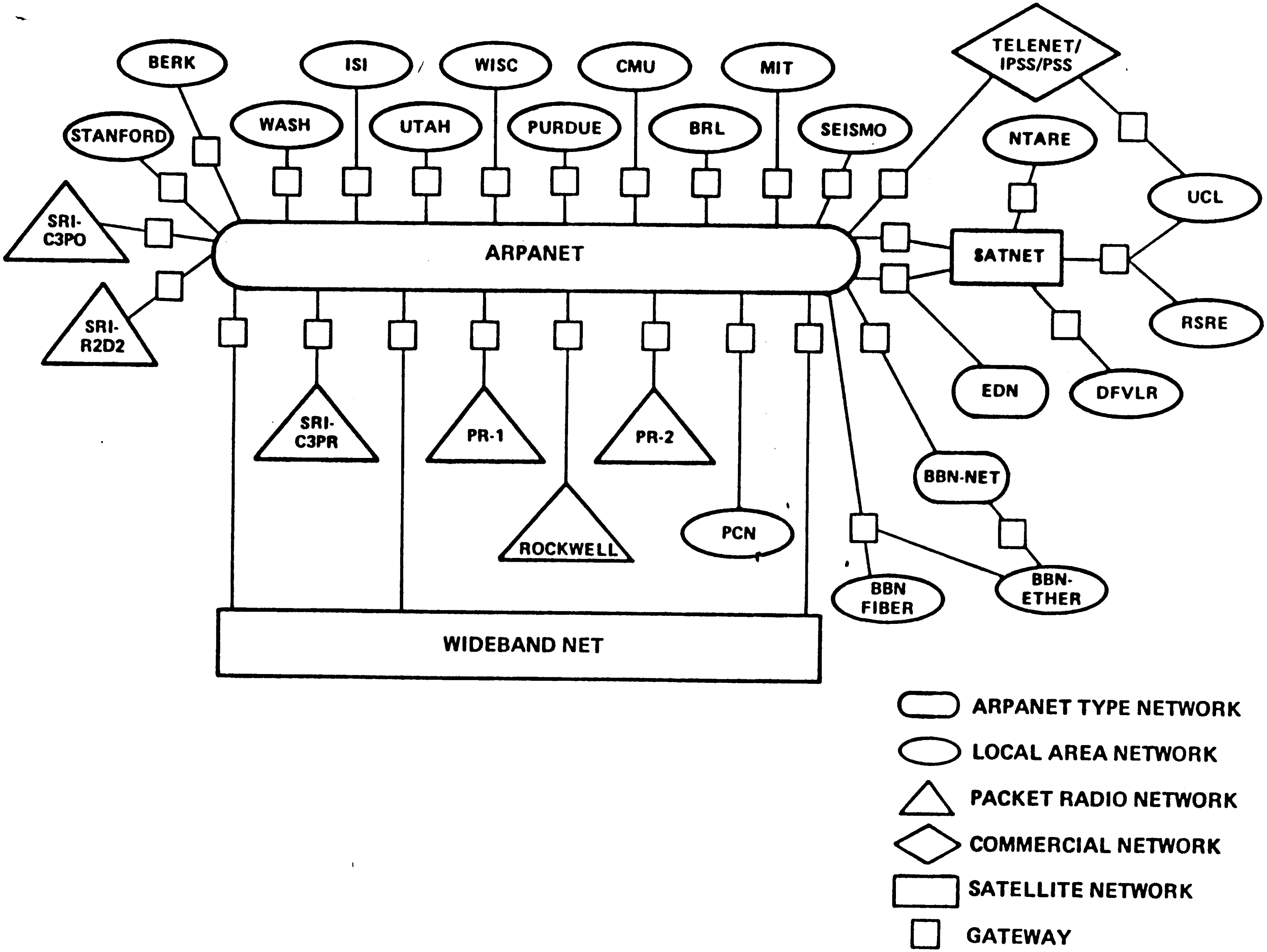
# APPENDIX B

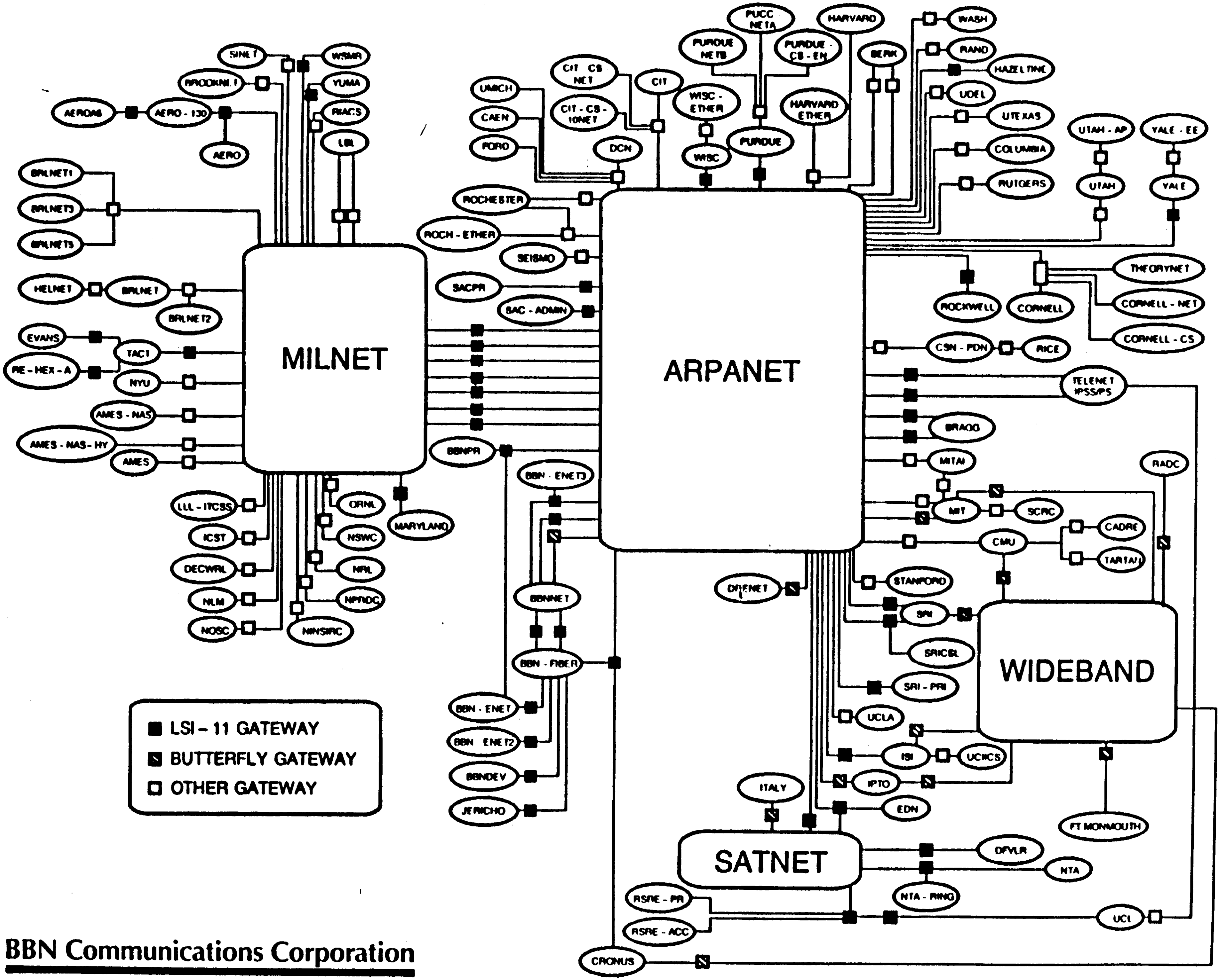
## Additional Papers Distributed At The Meeting

# APPENDIX B

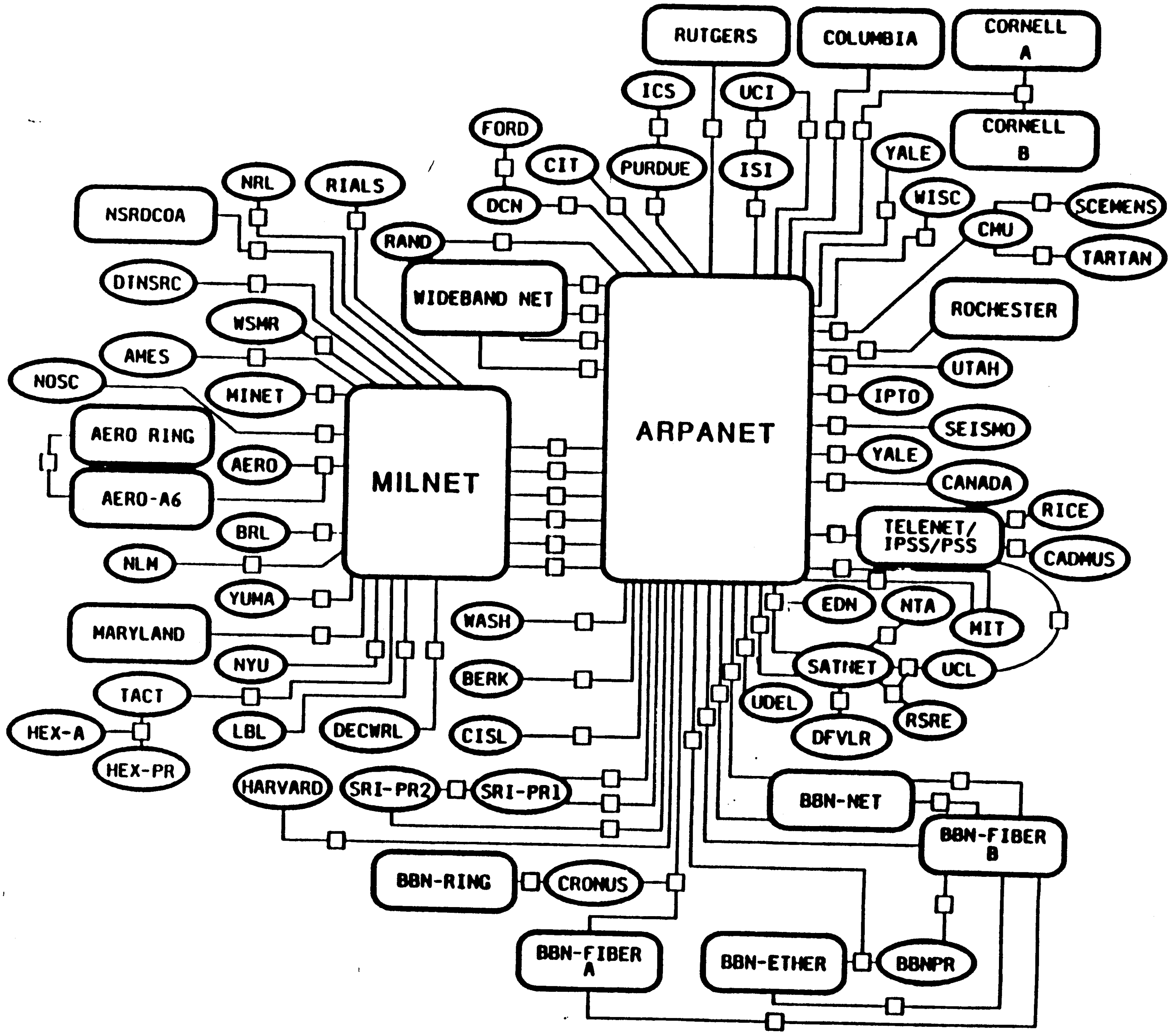
## Additional Papers Distributed At The Meeting

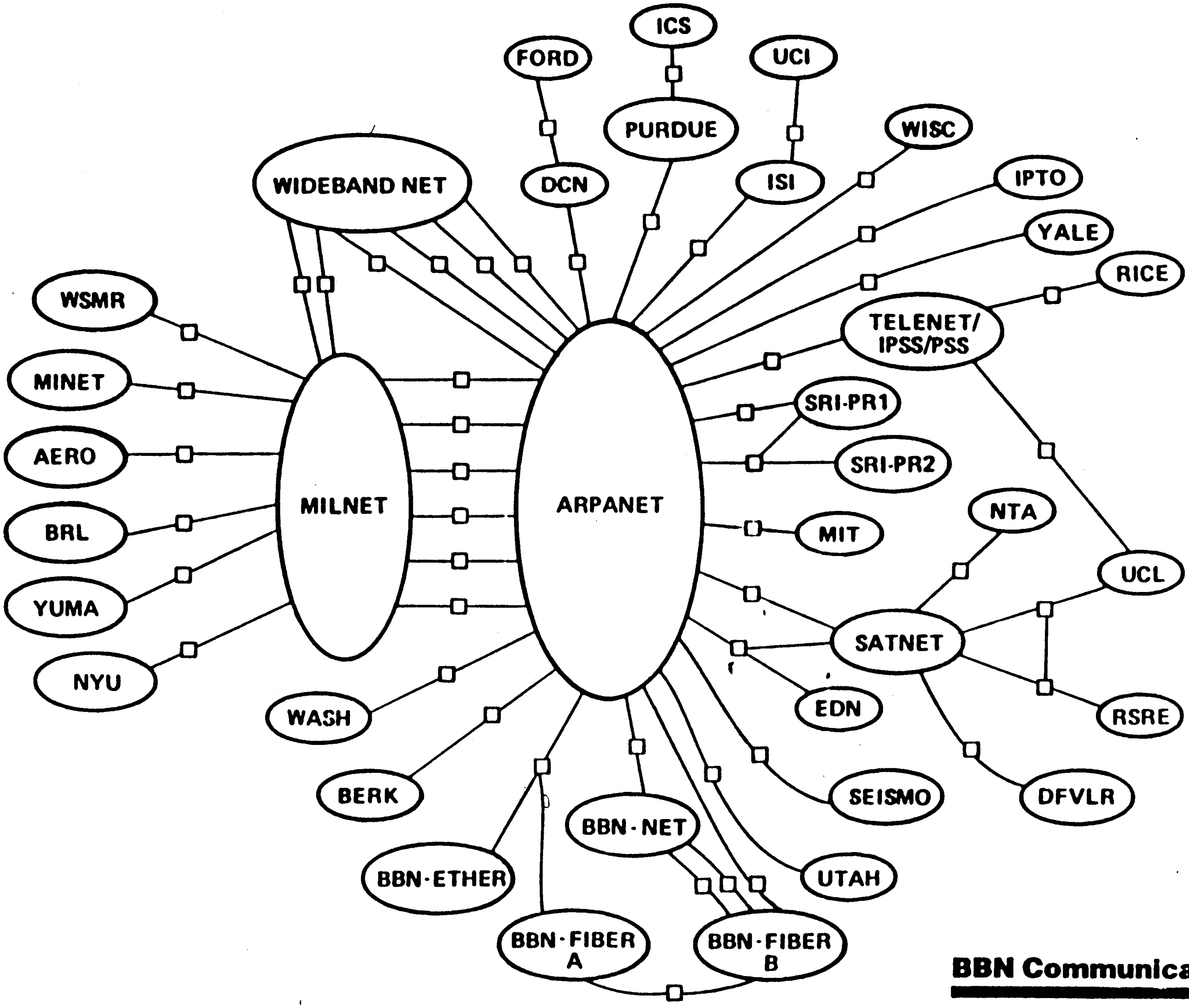
| *Distributed By:* | *Paper* |
|---|---|
| R. Hinden | *The Internet Through The Ages* |
| D. Mills | *Requirements For NSF Gateways* |
| N. Chiappa | *Interconnection Of A Host And The Internet* |
| H. W. Braun | *NSFnet Briefing Slides* |


BERK
ISI
WISC
CMU
MIT
TELENET/
IPSS/PSS
STANFORD
WASH
UTAH
PURDUE
BRL
SEISMO
NTARE
UCL
SRI-
C3PO
ARPANET
SATNET
RSRE
SRI-
R2D2
EDN
DFVLR
SRI-
C3PR
PR-1
PR-2
BBN-NET
ROCKWELL
PCN
BBN
FIBER
BBN-
ETHER
WIDEBAND NET
ARPANET TYPE NETWORK
LOCAL AREA NETWORK
PACKET RADIO NETWORK
COMMERCIAL NETWORK
SATELLITE NETWORK
GATEWAY



**BBN Communications Corporation**


RUTGERS
COLUMBIA
CORNELL A
CORNELL B
ICS
UCI
FORD
CIT
PURDUE
ISI
YALE
NRL
RIALS
WISC
NSRDCOA
DCN
CMU
SCEMENS
RAND
TARTAN
DINSRC
WIDEBAND NET
ROCHESTER
WSMR
AMES
UTAH
NOSC
MINET
IPTO
ARPANET
SEISMO
AERO RING
YALE
AERO
CANADA
MILNET
AERO-A6
RICE
TELENET/ IPSS/PSS
BRL
CADMUS
NLM
EDN
NTA
YUMA
WASH
MIT
MARYLAND
SATNET
UCL
NYU
BERK
TACT
UDEL
RSRE
HEX-A
LBL
DECWRL
CISL
DFVLR
HEX-PR
HARVARD
SRI-PR2
SRI-PR1
BBN-NET
BBN-FIBER B
BBN-RING
CRONUS
BBN-FIBER A
BBN-ETHER
BBNPR



**BBN Communications**

Network Working Group
Request for Comments: QQSV

National Science Foundation
April 1986

# Requirements for NSF Gateways

## Status of this Memo

This RFC summarizes the requirements for gateways to be used on networks supporting the DARPA Internet protocols and National Science Foundation research programs and was prepared by the Gateway Requirements Subcommittee of the NSF Network Technical Advisory Group. It requests discussion and suggestions for improvements. Distribution of this memo is unlimited.

The purpose of this document is to present guidance for vendors offering products that might be used or adapted for use in an NSF network. It ennumerates the protocols required and gives references to RFCs and other documents describing the current specifications. In a number of cases the specifications are evolving and may contain ambiguous or incomplete information. In these cases further discussion giving specific guidance is included in this document.

## 1. Introduction

The following sections are intended as an introduction and background for those unfamiliar with the DARPA Internet architecture and the Internet gateway model. General background and discussion on the Internet architecture and supporting protocol suite can be found in the documents "Internet Protocol Transition Workbook" and "Protocol Implementor's Guide," available from SRI International. Readers familiar with these concepts can proceed directly to Section 2.

### 1.1. The DARPA Internet Architecture

The DARPA Internet system consists of a number of gateways and networks that collectively provide packet transport for hosts subscribing to the DARPA Internet architecture. These protocols include the Internet Protocol (IP), Internet Control and Monitoring Protocol (ICMP), Transmission Control Protocol (TCP) and application protocols depending upon them. All protocols use IP as the basic packet-transport mechanism. IP is represented by a datagram, or connectionless, service and includes provision for service specification, fragmentation/reassembly and security information. ICMP is considered an integral part of IP, although it is architecturally layered upon it. ICMP provides error reporting, flow control and first-hop gateway redirection. Reliable data delivery is provided in the protocol suite by TCP, which provides end-end retransmission, resequencing and connection control.

The aggregate Internet community presently includes several thousand hosts connected to over 373 networks using over 127 gateways. There are now well over 2300 hosts registered in the ARPA domain alone and an unknown number registered in other domains, with the total increasing at about ten percent each month. Many of the hosts, gateways and networks in the Internet community are administered by civil organizations, including universities, research laboratories and equipment manufacturers. Most of the remainder are administered by the US DoD and considered part of the DDN Internet, which presently consists of three sets of networks: the experimental segment, or ARPANET, the unclassified segment, or MILNET, and the classified segment, which does not yet have a collective name.

The Internet model includes constituent networks, called local networks to distinguish them from the Internet system as a whole, which are required only to provide datagram (connectionless) transport. This requires only best-effort delivery of individual packets, or datagrams. Each datagram carries 32-bit source and destination addresses, which are encoded in three formats providing a two-part address, one of which is the local-network number and the other the host number on that local net. According to the Internet service specification, datagrams can be delivered out of order, be lost or duplicated and/or contain errors. In those networks providing connection-oriented service the extra reliability provided by virtual circuits enhances the end-end robustness of the system, but is not strictly necessary.

Local networks are connected together in the Internet model by means of Internet gateways. These gateways provide datagram transport only and normally seek to minimize the state information necessary to sustain this service in the interest of routing flexibility and robustness. In the conventional model the gateway has a physical interface and address on each of the local nets between which it provides forwarding services. The gateway also participates in one or more distributed routing or reachability algorithm such as the Gateway-Gateway Protocol (GGP) or Exterior Gateway Protocol (EGP) in order to maintain its routing tables.

## 1.2. The Internet Gateway Model

An Internet gateway is a self-contained, stand-alone packet switch that performs the following functions:

1. Interfaces to two or more packet-switching networks, including encapsulation, address transformation and flow control.

2. Conforms to specific DARPA Internet protocols specified in the document, including the Internet Protocol (IP), Internet Control Message Protocol (ICMP), Exterior Gateway Protocol (EGP) and others

as necessary.

3. Supports an interior gateway protocol (IGP) reachability or routing algorithm specific to a class of gateways operating as a system. Supports the EGP reachability algorithm to exchange routes between systems, in particular the DARPA "core" system operated by BBN.

4. Receives and forwards Internet datagrams consistent with good engineering practice in the management of resources, congestion control and fairness. Recognizes various error conditions and generates ICMP error messages as required.

5. Provides system support facilities, including loading, debugging, status reporting, exception reporting and control.

In some configurations gateways may be connected to packet-switching local nets that provide generic local-net routing, error-control and resource-management functions. In others gateways may be directly connected via serial lines, so that these functions must be provided by the gateways themselves.

There are three typical scenarios that should be addressed by gateway vendors:

1. National or regional network. Gateways of this class should be capable of switching multiple continuous flows in the 1.5-Mbps range at rates to several thousand packets per second. They will be high-performance, redundant, multiple-processor devices, probably procured as a system and operated remotely from a regional or national monitoring center. The design of these gateways should emphasize high aggregate throughput, throughput-sensitive resource management and very high reliabilty. The typical application would be an NSF backbone net or one of the consortium or regional nets.

2. Campus network. Gateways of this class should be capable of switching some burst flows at 10-Mbps (Ethernets, etc.), together with some flows in the 64-Kbps range or lower, at rates to perhaps a thousand packets per second. They will be medium-performance devices, probably competitively procured from different vendors for each campus and operated from a campus computing center. The design of these gateways should emphasize low average delay and good burst performance, together with delay and type-of-service sensitive resource management. Their chief function might be to interconnect various LANs and campus computing resources, including a high-speed interconnect to a national or regional net. An important factor will be a very flexible routing mechanism, since these gateways may have to select among several backbone nets based on cost/performance considerations.

3. Terminal network. Gateways of this class should be capable of

switching a small number of burst flows at 10-Mbps (Ethernets, etc.), together with a small number of flows in the range 64-Kbps or lower, at rates of a few hundred packets per second. They will be medium-performance devices procured from a variety of vendors and used for protocol-matching, LAN repeaters and as general utility packet switches. They will probably be locally maintained by the various users and not be used as transit switches.

It is important to realize that Internet gateways normally operate in an unattended mode, but that equipment and aoftware faults can affect the entire Internet. While some of the above scenarios involve positive control of some gateways from a monitoring center, usually via a path involving other networks and Internet gateways, others may involve much less formal control procedures. Thus the gateways must be highly robust and be expected to operate, possibly in a degraded state, under conditions of extreme congestion or failure of network resources.

## 2. Protocols Required

The Internet architecture uses datagram gateways to interconnect networks and subnetworks. These gateways function as intermediate systems (IS) with respect to the ISO connectionless network model and incorporate defined packet formats, routing algorithms and related procedures. In the following it is assumed the protocol implementation supports the full protocol, including all required options, with exceptions only as noted.

### 2.1. Internet Protocol (IP)

This is the basic datagram protocol used in the Internet system. It is described in RFC-791 and also MIL-STD-1777, both of which are intended to describe the same standard, but in quite different words.

With respect to current gateway requirements the following can be ignored: Type of Service field, Security option, Stream ID option and Timestamp option. However, if recognized, the interpretation of these quantities must conform to the standard specification.

Note that the Internet gateway model does not require that the gateway reassemble IP datagrams with destination address other than the gateway itself. However, in the case of those protocols in which the gateway directly participates as a peer, including routing and monitor/control protocols, the gateway may have to reassemble datagrams addressed to it. This consideration is most pertinent to EGP.

### 2.2. Internet Control Message Protocol (ICMP)

This is an auxilliary protocol used to convey advice and error messages and is described in RFC-792.

The distinction between subnets of a subnetted network, which depends on an arbitrary mask as described in RFC-950, is in general not visible outside that network. This distinction is important in the case of certain ICMP messages, including the ICMP Destination Unreachable and ICMP Redirect messages. The ICMP Destination Unreachable message is sent by a gateway in response to a datagram which cannot be forwarded because the destination is unreachable or down. A choice of several types of these messages is available, including one designating the destination network and another the destination host. However, the span of addresses implied by the former is ill-defined unless the subnet mask is known to the sender, which is in general not the case.

The ICMP Redirect message is normally sent by a gateway to a host in order to change its first-hop gateway address for a designated net; however, this message can in principle be sent in other cases as well. A choice of four types of messages is available, depending on whether it applies to a particular host, network or service. As in the previous case, these distinctions may depend upon the subnet mask.In both of the above cases it is recommended that the use of ICMP messages implying a span of addresses (net unreachable, net redirect) be avoided in favor of those implying specific addresses.

The ICMP Source Quench message has been the subject of much controversy. It is not considered realistic at this time to specify in detail the conditions under which this message is to be generated or interpreted by a host or gateway.

New implementations are expected to support the ICMP Address Mask messages described in RFC-950. It is highly desirable, although not required, to provide correct data for ICMP Timestamp messages, which have been found useful in network debugging and maintenance.

## 2.3. Exterior Gateway Protocol (EGP)

This is the basic protocol used to exchange information with other gateways of the Internet system and is described in RFC-904. However, EGP as presently specified is an asymmetric protocol with only the "non-core" procedures defined in RFC-904. There are at present no "core" procedures specified, which would be necessary for a stand-alone Internet. RFC-975 suggests certain modifications leading to a symmetric model; however, this is not an official specification.

In principle, a stand-alone Internet can be built with non-core EGP gateways using the EGP distance field to convey some metric such as hop count. However, the use of EGP in this way as a routing algorithm is discouraged, since typical implementations adapt very slowly to changing topology and have no loop-protection features.

If a routing algorithm is operated in one or more gateways, its data base must be coupled to the EGP implemntation in such a way that, when a net is declared down by the routing algorithm, the net is also declared down via EGP to other autonomous systems. This requirement is designed to minimize demand and insure fairness on the core-system resources.

There are no peer-discovery or authentication procedures defined in the present EGP specification and no defined interpretation of the distance fields in the update messages, although such procedures may be defined in future (see RFC-975). There is currently no guidance on the selection of polling parameters and no specific recovery procedures in case of certain error messages (e.g. "administratively prohibited"). It is recommended that EGP implementations include provisions to initialize these parameters as part of the monitoring and control procedures and that changing these procedures not require recompilation or rebooting the gateway.

## 2.4. Address Resolution Protocol (ARP)

This is an auxilliary protocol used to manage the address-translation function between Ethernet addresses and Internet addresses and described in RFC-826. However, there are a number of unresolved issues having to do with subnets and response to addresses not in the same subnet or net. These issues, which are intertwined with ICMP and various gateway models, are discussed in Appendix A.

## 3. Subnets

The concept of subnets was introduced in order to allow arbitrary complexity of interconnected LAN structures within an organization, while insulating the Internet system against explosive growth in network numbers and routing complexity. The subnet architecture, described in RFC-950, is intended to specify a standard approach that does not require reconfiguration for host implementations connected to an Ethernet, regardless of subnetting scheme. The document also specifies a new ICMP Address Mask message, which a gateway can use to specify certain details of the subnetting scheme to Ethernet hosts and is required in new host implementations.

The current subnet specification RFC-950 does not describe the specific procedures to be used by the gateway, except by implication. It is recommended that a (sub)net address and address mask be provided for each network interface and that these values be established as part of the gateway configuration procedure. It is not usually necessary to change these values during operation of any particular gateway; However, it should be possible to add new gateways and/or (sub)nets and make other configuration changes to a gateway without taking the entire network down.

## 4. Local Network Interface

The packet format used for transmission of datagrams on the various subnetworks is described in a number of documents summarized below.

### 4.1. Public data networks via X.25

The formats specified for public data subnetworks via X.25 access are described in RFC-877. Datagrams are transmitted over standard level-3 virtual circuits as complete packet sequences. Virtual circuits are usually established dynamically as required and time out after a period of no traffic. Retransmission, resequencing and flow control are performed by the network for each virtual circuit and by the LAPB link-level protocol, however, multiple parallel virtual circuits are often used in order to improve the utlitization of the subscriber access line, which can result in random resequencing. The correspondence between Internet and X.121 addresses is usually established by table-lookup. It is expected that this will be replaced by some sort of directory procedure in future.

### 4.2. ARPANET via 1822 Local Host, Distant Host or HDLC Distant Host

The formats specified for ARPANET subnetworks via 1822 access are described in BBN Report 1822, which includes the procedures for several subscriber access methods. The Local Host (LH) and Very Distant Host (VDH) methods are not recommended for new implementions. The Distant Host (DH) method is used when the host and IMP are separated by not more than about 2000 of cable, while the HDLC Distant Host is used for greater distances where a modem is required. Retransmission, resequencing and flow control are performed by the network and by the HDLC link-level protocol, when used. While the ARPANET 1822 protocols are widely used at present, they are expected to be eventually overtaken by the DDN Standard X.25 protocol and the new PSN End-to-End Protocol described in RFC-979.

Gateways connected to ARPANET/MILNET IMPs must incorporate features to avoid host-port blocking (RFNM counting) and to detect and report (as ICMP Unreachable messages) the failure of destination hosts or gateways.

### 4.3. ARPANET via DDN Standard X.25

The formats specified for ARPANET subnetworks via X.25 are described in the "Defense Data Network X.25 Host Interface Specification". This document describes two sets of procedures, the DDN Basic X.25 and the DDN Standard X.25, but only the latter is suitable for use in the Internet system. The DDN Standard X.25 procedures are similar to the public data subnetwork X.25 procedures, except in the address mappings. Retransmission, resequencing and flow control are performed by the network and by the LAPB link-level protocol.

### 4.4. Ethernets

The formats specified for Ethernet subnetworks are described in RFC-894. Datagrams are encapsulated as Ethernet packets with 48-bit source and destination address fields and a 16-bit type field. Address translations between Ethernet addresses and Internet addresses is managed by the Address Resolution Protocol, which is required in all implementations. There is no explicit retransmission, resequencing or flow control. although most hardware interfaces will retransmit automatically in case of collisions on the cable.

It is expected that amendments will be made to this specification as the result of IEEE 802 evolution. See RFC-948 for further discussion and recommendations in this area. Note also that the IP broadcast address, which has primary application to Ethernets and similar technologies that support an inherent broadcast function, has an all-ones value in the host field of the IP address. Some early implementations chose the all-zeros value for this purpose, which is presently not in conformance with the definitive specification RFC-922.

See Appendix A for further considerations.

### 4.5. Serial-Line Protocols

Gateways may be used as packet switches in order to build networks. In some configurations gateways may be interconnected with each other and some hosts by means of serial asynchronous or synchronous lines, with or without modems. When justified by the expected error rate and other factors, a link-level protocol may be required on the serial line. While there is no requirement that a particular standard protocol be used for this, it is recommended that standard hardware and protocols be used, unless a convincing reason to the contrary exists. In order to support the greatest variety of configurations, it is recommended that full X.25 be used where resources permit; however, X.25 LAPB would also be acceptable where requirements permit. In the case of asynchronous lines no clear choice is apparent.

## 5. Interoperability

In order to assure interoperability between gateways procured from different vendors, it is necessary to specify points of protocol demarcation. With respect to interoperability of the routing function, this is specified as EGP. All gateway systems must include one or more gateways which support EGP with a core gateway, as described in RFC-904. It is desirable that these gateways be able to operate in a mode that does not require a core gateway or system. Additional discussion on these issues can be found in RFC-975.

With respect to the interoperability at the network layer and below, two points of protocol demarcation are specified, one for Ethernets and the other for serial lines. In the case of Ethernets the protocols are as specified in Section 4 of this document. For serial lines between gateways of different vendors, the protocols are specified as full X.25. Exceptions to these requirements may be appropriate in some cases.

## 6. Subnetwork Architecture

It is recognized that gateways may also function as general packet switches to build networks of modest size. This requires additional functionality in order to manage network routing, control and configuration. While it is beyond the scope of this document to specify the details of the mechanisms used in any particular, perhaps proprietary, architecture, there are a number of basic requirments which must be provided by any acceptable architecture.

### 6.1. Reachability Procedures

The architecture must provide a robust mechansim to establish the operational status of each link and node in the network, including the gateways, the lines that connect them and, where appropriate, the hosts connected to the network. Ordinarily, this requires at least a link-reachability protocol involving a periodic exchange of hello messages, which might be intrinsic to the link-level protocols used (e.g. DDCMP). It is in general ill-advised to assume a host or gateway is operating correctly if the link-reachability protocol connecting to it is operating correctly. Additional confirmation is required in the form of an operating routing algorithm or peer-level reachability protocol, such as used in EGP.

Failure and restoral of a link and/or gateway are considered network events and must be reported to the control center. It is desireable, although not required, that reporting paths not require correct functioning of the routing algorithm itself.

### 6.2. Routing Algorithm

It has been the repeated experience of the Internet community participants that the routing mechanism, whether static or dynamic, is the single most important engineering issue in network design. In all but trivial network topologies it is necessary that some degree of routing dynamics is vital to successful operation, whether it be affected by manual or automatic means or some combination of both. In particular, if routing changes are made manually, the changes must be possible without taking down the gateway for reconfiguration and, preferably, be possible from a remote site such as a control center.

It is not likely that all nets can be maintained from a full-service control center, so that automatic-fallback or rerouting features may be required. This must be considered the normal case, so that systems of gateways operating as the only packet switches in a network would normally be expected to have a routing algorithm with the capability of reacting to link and other gateway failures and changing the routing automatically. Following is a list of features considered necessary:

1. The algorithm must sense the failure or restoral of a link or other gateway and switch to appropriate paths within an interval bounded from above by a constant times the network diameter.

2. The algorithm must never form routing loops between neighbor gateways and must contain provisions to avoid and suppress routing loops that may form between non-neighbor gateways. In no case should a loop persist for longer than an interval bounded from above by a constant times the network diameter.

3. The control traffic necessary to operate the routing algorithm must not significantly degrade or disrupt normal network operation. Changes in state which might momentarily disrupt normal operation in a local area must not cause disruption in remote areas of the network.

4. As the size of the network increases, the demand on resources must be controlled in an efficient way. Table lookups should be hashed, for example, and data-base updates handled piecemeal, with only the changes broadcast over a wide area. Reachability and delay metrics, if used, must not depend on direct connectivity to all other gateways or the use of network-specific broadcast mechanisms. Polling procedures (e.g. for consistency checking) should be used only sparingly and in no case introduce an overhead exceeding a constant times the network diameter.

5. The use of a default gateway as a means to reduce the size of the routing data base is strongly discouraged in view of the many problems with multiple paths, loops and mis-configuration vulnerabilities. If used at all, it should be limited to a discovery function, with operational routes cached from external or internal data bases via either the routing algorithm or EGP.

6. This document places no restriction on the type of routing algorithm, such as min-hop, shortest-path-first or any other algorithm, or metric, such as delay or hop-count. However, the size of the routing data base must not be allowed to exceed a constant times the network diameter. In general, this means that the entire routing data base cannot be kept in any particular gateway, so that discovery and caching techniques are necessary.

## 7. Operation and Maintenance

Gateways and packets switches are often operated as a system by some organization who agrees to operate and maintain the gateways, as well as to resolve link problems with the respective common carriers.

In general, the following requirements apply:

1. Each gateway must operate as a stand-alone device for the purposes of local hardware maintenance. Means must be available to run diagnostic programs at the gateway site using only on-site tools, which might be only a diskette or tape and local terminal. It is desireable, although not required, to run diagnostics via the network and to automatically reboot and dump the gateway via the net in case of fault. In general, this requires special hardware.

2. It must be possible to reboot and dump the gateway manually from the control site. Every gateway must include a watchdog timer that either initiates a reboot or signals a remote control site if not reset periodically by the software. It is desireable that the data involved reside at the control site and be transmitted via the net; however, the use of local devices at the gateway site is acceptable. Nevertheless, the opeation of initiating reboot or dump must be possible via the net, assuming a path is available and the connecting links are operating.

3. A mechanism must be provided to accumulate traffic statistics including, but not limited to, packet tallies, error-message tallies and so forth. The preferred method of retrieving these data is by explicit request from the control site using a standard protocol such as TCP.

4. Exception reports ("traps") occuring as the result of hardware or software malfunctions should be transmitted immediately (batched to reduce packet overheads when possible) to the control site using a standard protocol such as UDP.

5. A mechanism must be provided to display link and node status on a continuous basis at the control site. While it is desireable that a complete map of all links and nodes be available, it is acceptable that only those components in use by the routing algorithm be displayed. This information is usually available local at the control site, assuming that site is a participant in the routing algorithm.

The above functions require in general the participation of a control site or agent. The preferred was to provide this is as a user program suitable for operation in a standard software environment such as Unix. The program would use standard IP protocols sucvh such as TCP and UDP to control and monitor the gateways. The use of specialized host hardware and software requiring significant additional investment is strongly discouraged; nevertheless, some vendors may elect to provide the control agent as an integrated part of the network in which the gateways are a part. If this is the case, it is required that a means be available to operate the control agent from a remote site using Internet protocols and paths and with equivalent functionality with respect to a local agent terminal.

Remote control of a gateway via Internet paths can involve either a direct approach, in which the gateway suports TCP and/or UDP directly, or an indirect approach, in which the control agent supports these protocols and controls the gateway itself using proprietary protocols. The former approach is preferred, although either approach is acceptable.

## Appendix A. Ethernet Management

Following is a summary of procedures specified for use on an Ethernet.

### A.1. Hardware

A packet is accepted from the cable only if its destination Ethernet address matches either the assigned interface address or a broadcast/multicast address. Presumably, this filtering is done by the interface hardware; however, the software driver is expected to do this if the hardware does not. Fuzzballs incorporate an optional feature that associates an assigned multicast address with a specific subnet in order to restrict access for testing, etc. When this feature is activated, the assigned multicast address replaces the broadcast address.

### A.2. IP datagram

In case of broadcast/multicast (as determined from the destination Ethernet address) an IP datagram is rejected if the source IP address is not in the same subnet, as determined by the assigned host IP address and subnet mask. It is desirable that this test be defeatible by a configuration parameter, in order to support the infrequent cases where more than one subnet may coexist on the same cable.

### A.3. ARP datagram

An ARP reply is rejected if the destination IP address does not match the local host address. An ARP request is rejected if the source IP address is not in the same subnet. It is desirable that this test be defeatible by a configuration parameter, in order to support the infrequent cases where more than one subnet may coexist on the same cable. An ARP reply is generated only if the destination protocol IP address is reachable from the local host (as determined by the routing algorithm) and the next hop is not via the same interface. If the local host functions as a gateway, this may result in ARP replies for destinations not in the same subnet.

### A.4. ICMP redirect

An ICMP redirect is rejected if the destination IP address does not match the local host address or the new target address is not on the same subnet. An accepted redirect updates the routing data base for the old target address. If there is no route associated with the old target address, the redirect is ignored. Note that it is not possible to send a gratuitous redirect unless the sender is possessed of considerable imagination.

When subnets are in use there is some ambiguity as to the scope of a redirect, unless all hosts and gateways involved have prior knowledge of the subnet masks. It is recommended that the use of ICMP network-redirect messages be avoided in favor of ICMP host-redirect messages instead. This requires the original sender (i.e. redirect recipient) to support a general IP address-translation cache, rather than the usual network table. However, this is normally done anyway in the case of ARP.

An ICMP redirect is generated only if the destination IP address is reachable from the local host (as determined by the routing algorithm), the next hop is via the same interface and the target address is defined in the routing data base.

ICMP redirects are never forwarded, regardless of destination address. The source IP address of the ICMP redirect itself is not checked, since the sending gateway may use one of its addresses not on the common net. The source IP address of the encapsulated IP datagram is not checked on the assumption the host or gateway sending the original IP datagram knows what it is doing.

## Appendix B

The following sections discuss certain issues of special concern to the NSF scientific networking community. These issues have primary relevance in the policy area, but also have ramifications in the technical area.

### B.1. Interconnection Technology

Currently the most important common interconnection technology between Internet systems of different vendors is Ethernet. Among the reasons for this are the following:

1. Ethernet specifications are well-understood and mature.

2. Ethernet technology is in almost all aspects vendor independent.

3. Ethernet-compatible systems are common and becoming more so.

These advantages combined favor the use of Ethernet technology as the common point of demarcation between NSF network systems supplied by different vendors, regardless of technology. It is a requirement of NSF gateways that, regardless of the possibly proprietary switching technology used to implement a given vendor-supplied network, its gateways must support an Ethernet attachment to gateways of other vendors.

It is expected that future NSF gateway requirements will specify other interconnection technologies. The most likely candidates are those based on X.25 or IEEE 802, but other technologies including broadband cable, fiber-optic or other protocols such as DDCMP may also be considered.

### B.2. Proprietary and Extensible Issues

Internet technology is a growing, adaptable technology. Although hosts, gateways and networks supporting this technology have been in continuous operation for several years, vendors users and operators should understand that not all networking issues are fully understood. As a result, when new needs or better solutions are developed for use in the NSF networking community, it may be necessary to field new protocols. Normally, these new protocols will be designed to interoperate in all practical respects with existing protocols; however, occasionally it may happen that existing systems must be upgraded to support these protocols.

NSF systems vendors should understand that they also undertake a committment to remain aware of current Internet technology and be prepared to upgrade their products from time to time as appropriate. As a result, These vendors are strongly urged to consider extensibility and periodic upgrades as fundamental characteristics of their products. One of the most productive and rewarding ways to do this on a long-term basis is to participate in ongoing Internet research and development programs in partnership with the academic community.

## B.3. Multi-Protocol Gateways

Although the present requirements for an NSF gateway specify only the Internet protocol suite, it is highly desirable that gateway designs allow future extensions to support additional suites and allow simultaneous operation with more than a single one. Clearly, the ISO protocol suite is a prime candidate for one of these suites. Other candidates include XNS and DECnet.

Future requirements for NSF gateways may include provisions for other protocol suites in addition to Internet, as well as models and specifications to interwork between them, should that be appropriate. For instance, it is expected that the ISO suite will eventually become the dominant one; however, it is also expected that requirements to support other suites will continue, perhaps indefinately.

Present NSF gateway requirements do not include protocols above the network layer, such as TCP, unless necessary for network monitoring or control. Vendors should recognize that future requirements to interwork between Internet and ISO applications, for example, may result in an opportunity to market gateways supporting multiple protocols at all levels through the application level. It is expected that the network-level NSF gateway requirements summarized in this document will be incorporated in the requirements document for these application-level gateways.

## B.4. Access Control and Accounting

There are no requirements for NSF gateways at this time to incorporate specific access-control and accounting mechanisms in the design; however, these important issues are currently under study and will be incorporated into a redraft of this document at an early date. Vendors are encouraged to plan for the early introduction of these mechanisms in their products. While at this time no definitive common model for access control and accounting has emerged, it is possible to outline some general features such a model is likely to have, among them the following:

1. The primary access control and accounting executive mechanisms will be in the service hosts themselves, not the gateways, packet switches or workstations.

2. Agents acting on behalf of access control and accounting executive mechanisms may be necessary in the gateways, packet switches or workstations. These may be used to collect data, enforce password

protection or mitigate resource priority and fairness. However, the architecture and protocols used by these agents may be a local matter and not possible to specifiy in advance.

3. NSF gateways may be required to incorporate access control and accounting mechanisms based on packet source/destination address, as well as other fields in the IP header, internal priority and fairness. However, it is extremely unlikely that these mechanisms would involve a user-level login to the gateway itself.

Network Working Group
Request for Comments: XXX

J. Noel Chiappa
Lab for Computer Science, M.I.T.
February 1985

# INTERCONNECTION OF A HOST AND THE INTERNET

## 1 STATUS OF THIS MEMO

This is a draft of an RFC which is under consideration as a standard. This RFC will specify a standard for the DARPA Internet community. Hosts on the ARPA-Internet will be expected to adopt and implement this standard. Distribution of this memo is unlimited.

This RFC will be expanded over time as additional insights in differing areas are gained. At present the topics covered are: Routing. It is being released in this partial form now because a standard in the area of routing is needed, and it is not possible to complete the entire document in a timely fashion.

## 2 OVERVIEW

A lot of disparity exists in the functionality present in the IP layer in various host implementations. This is a severe problem in an evolving system like the Internet, since it is not clear what changes to the basic architecture will affect hosts, and how major the effects will be. Host IP layers often contain more functionality than they need; it is desirable for them to be as minimal as possible, to minimize the chances of being caught up in changes.

This memo sets general standards for the functionality which must be present in a host IP layer. It outlines and strongly recommends an implementation guideline so that all host IP layers will be similar. It will also tend to insulate the hosts from changes in the architecture.

Much useful information along these lines is contained in the set of RFC's by Dave Clark, [1], [2], and [3], which deal with the IP layer. These are reprinted in the Implementation Guide available from the Network Information Center [5]. However, do be aware that these documents have not been revised and may contain inconsistencies with this document. In these cases, this document should be taken as superseding the ones listed.

## 3 ACKNOWLEDGMENTS

Address masks were talked about during a conversation with Dave Moon, and the scheme outlined here to use them to insulate the hosts was

delineated by Dave Clark.

## 4 INTRODUCTION

### 4.1 Routing

When routing is considered, hosts need to decide whether a destination can be reached directly by the locally attached transmission medium (called the 'local net' from here on, even though it may not be a network of the type known by that cognomen), or whether it has to be reached by forwarding at the IP level through some intermediate entity. If a forwarding step is needed, then the intermediate entity needs to be chosen.

However, hosts often provide a more sophisticated IP layer than necessary, and they become involved in routing decisions that should properly be left to the gateways. Such involvement is undesirable, since then changes to the basic architecture that involve routing have repercussions in the hosts. Since the system is fluid in this area, we wish to remove this dependency. In addition, the new approach has the characteristic that it is general enough so that future changes to the Internet architecture to attack other problems will, in many instances, not require any further changes to the hosts.

## 5 DETAILS

### 5.1 Routing

#### 5.1.1 Basics

Clearly, the first step in handling an outbound packets is to decide if the destination is on the local net or not. At the moment, this is done by parsing the destination address to extract a network number, and then seeing if it matches the network number of the attached network.

We would like to make this step more general so that parsing the address is no longer necessary. That way, if the form of an address changes, it will not be necessary to change the address handling code. All hosts should consider IP addresses as featureless 32 bit numbers. A simple algorithm is needed to effect the decision above; one that is simple, but provides considerable flexibility, is the use of a bit mask.

If the part of the destination address under the mask matches the part of the host's address under the mask, then the destination is on the same local net and the packet should be transmitted directly. If not, then the destination is elsewhere and it must be routed through a gateway. It should be possible to set the bit mask as part of the host configuration information, like the host address. (Presumably, the part of the host address under the mask will be computed once and stored, for efficiency reasons.)

The second step is that if the destination is not local, a gateway must be picked and the packet routed through that gateway. Typically, many hosts now maintain a network routing table; this is a database which relates destination network numbers to next hop gateways. This is filled in in a variety of ways, including configuration tables as well as dynamically, from the net, via ICMP Redirects. Such a table is undesirable, since once again this is including routing functionality in the hosts.

The appropriate method, rather, is to keep a cache of gateways for individual distant host addresses (once again considered as featureless 32 bit numbers). When a packet must be sent to a host which is not on the local net, and which does not have an entry in the routing cache, a gateway (one of the set of gateways already known) is chosen and the packet sent there. If the gateway is not a good next hop for the destination in question, it will send an ICMP Host Redirect message back to the originator. The originator should use this information to update the entry for that particular host in the routing cache.

### 5.1.2 Route cache maintenance

One possible implementation choice is to have a single default gateway, and only make entries in the routing cache for hosts which do not go through that gateway. Another is to not have a distinguished default gateway; when a route to a host not in the cache is needed, a new entry is created and filled with a gateway randomly picked from the set of gateways already known.

An important point is how entries are discarded. If a gateway goes down, cache entries that point to it will cause packets to be discarded, perhaps undetectably. Somehow, the fact that the entry points to a dead gateway must be rectified. If the local net has a low level method for indicating dead hosts, that can be used to invalidate entries, but this method cannot be the only one since some nets do not provide that information. This is most important: failure to have some mechanism to address this question will lead to hosts becoming unreachable even though a viable path exists. (This topic is covered in some detail in [3].) The host may also wish to recycle entries which have not been used recently, but this is optional.

One possible strategy not mentioned there is to age cache entries; when one is used it is marked as recently used. The host should periodically go through and send an ICMP ping to all cache entries marked as active. Gateways which do not respond should have all their cache entries deleted. This presents a lower load to the net than simply polling all the gateways all the time, but is better than nothing, and would be applicable in cases where the host higher level client software is not structured to use an entry point where the client can advise the lower level that a particular route may be dead.

One final thing to note is that although it is permissible for host IP layers to look up a route on every packet, the implementor striving for an efficient implementation may wish to keep a 'hint' in any protocol with connections, which does away with this unnecessary overhead. (This is a point that is really orthogonal to whether a conventional routing table or a per host cache is used; it is discussed in [4].)

### 5.1.3 Dynamic configuration on broadcast nets

Another important point is use of broadcast nets to reduce the amount of configuration information the host must be provided with manually. There are three pieces of configuration information which will be needed: the host address, the address mask for the local net, and one (or more) gateways to 'bootstrap' the routing cache. On a broadcast net, broadcast ICMP packets can be used to gather all but the first piece of information. This technique is only available as an option, since the code must be able to work on nets that do not support broadcast.

One thing to note with this approach is that if it is used it may require extra care and resources to function reliably. Consider the case in which it is desired to operate a cluster of hosts on a single isolated broadcast net without any gateways; if the code insists on finding at least one gateway to start the IP gateway cache with before it will function, the hosts will be unable to operate. Conversely, the hosts may proceed and assume they are on an isolated net, when they are in fact on a net with a single gateway that is temporarily down. In this case, they may be unable to contact sites off net even if the gateway starts functioning. Clearly, there is some trade-off between the certainty of getting the correct information and the resources used to get it.

To combat this, hosts should assume an all zero mask initially, which will act as if all possible destinations are on the local net, until a ICMP Mask Reply is received. The host should also be prepared to accept and act on an unsolicited Mask Reply, to cover the case where the gateway starts up at some later time. Likewise, the routing cache (and default gateway, if any) should be initially empty; whenever a packet needs to be sent off the local net and the cache is empty an ICMP Gateway Request should be broadcast. (Also, if a default gateway is being used, and the entry is empty, an unsolicited Gateway Reply should fill it). If none is received, then it should be assumed that the local net is isolated.

### 5.1.4 Multi-homed hosts

In general, hosts should not attempt to be gateways if they are multi-homed. The reason for this is that the functions of a gateway, and the protocols by which they communicate, are not stable, and will have to continue to change as the system grows larger. Attempting to have a general gateway function as part of the host IP layer will thus force the maintainer to track these changes. Also, a larger pool of gateway implementations will make coordinating the changes more difficult. For these reasons, providing host IP layers with the capability to be gateways is in general not advised.

There are some tricky questions when dealing with multi-homed hosts. For instance, when you do an address lookup on such a host (perhaps using the Name Server Protocol, [6]), you are returned several IP addresses. Obviously, any one should function correctly, but some hosts may wish to pick the 'best' address, the one the use of which will produce the best performance path. To tell (from that list) which one is the best to use, there is a new ICMP query/response pair by which a host can get a gateway on the local net to pick the 'best' address from the list.

### 5.1.5 Future Directions

The approach shown here will allow us to attack a variety of problems in the future, without any change to the hosts. For example, consider partitioned nets and mobile hosts. Per host route caches in the hosts allow us to attack these problems without affecting the hosts. Additional mechanism will be required in the gateways, but the changes should be invisible in the hosts.

### 5.1.6 Notes

This section contains notes about what will have to change elsewhere in the IP specs if this spec is adopted. This section will be removed before release.

- Clearly, gateways which now send ICMP Network Redirects will have to be changed to send Host Redirects. ICMP will eventually be changed to remove Network Redirects. (Should gateways find the need for something like Network Redirect, it should be made part of some inter-gateway protocol, since ICMP should contain only things needed for the interaction of hosts and gateways.) To ease the conversion, gateways should first be changed to send both Redirect types; this will allow the conversion of hosts at leisure. When the hosts are done, Network Redirects can be removed from the gateways.
- Also, several new ICMP packet types will need to be set up, for use in finding configuration information on broadcast nets. These will be Mask Request, Mask Reply, Gateway Request and Gateway Reply. These will look much like the Information Request/Reply messages. Also, for handling multi-homed hosts, an Optimal Address Request and Reply will have to be defined. (The hosts do this rather than the name servers since the name

servers may not be on the same local net, and thus will not have knowledge of the gateways local to the requester, which are the ones which will know which address is 'best'.) Note that the Optimal Address request should include a TOS field, as well a list of addresses, to allow the gateways to consider TOS as well when choosing the 'best' address.

- One thing is left unspecified about the internals of multi-homed hosts. The problem is that there is a slight routing problem on outbound connections; how does the host pick which of its addresses to use (and thus, which interface to send packets out over) when initiating a new connection? I cannot think of any easy solution other than having the host ask gateways on each attached net what they think the 'cost' of getting somewhere is and using the net with the lowest cost. (Clearly, if the destination is on a net the host is also attached to that net is the one to use.) One difficulty here is that if a host is attached to two different autonomous systems, they may not use identical routing metrics, and thus comparing the routes may be impossible.

- Also, there is a question on whether hosts can reply to Gateway or Mask Request. This might alleviate some problems with the 'isolated net case above', but my feeling is that it is not needed. The mask or gateway info is only needed for communication off the local net, and if there isn't a gateway to send the Mask Reply you don't need it. I don't see any reason to burden the hosts with doing this, since the gateways will have to do it anyway. I suppose that hosts could be allowed to do so optionally, but what's the point? Why allow it if it doesn't buy you anything? It might get us in trouble later on. Thus, gateways will simply send the replies on startup. This has the additional advantage of saving the hosts the overhead of having them poll continuously for a possible state change. Note that the gateways should send these several times (especially the Mask Reply) so that all the hosts are fairly certain to get the information.

## REFERENCES

1. Clark, David D. Names, Addresses, Ports and Routes. Network Working Group Request for Comments RFC 814, DARPA Network Working Group, July, 1982.

2. Clark, David D. IP Datagram Reassembly Algorithms. Network Working Group Request for Comments RFC 815, DARPA Network Working Group, July, 1982.

3. Clark, David D. Fault Isolation and Recovery. Network Working Group Request for Comments RFC 816, DARPA Network Working Group, July, 1982.

4. Clark, David D. Modularity and Efficiency in Protocol Implementation. Network Working Group Request for Comments RFC 817, DARPA Network Working Group, July, 1982.

5. Internet Protocol Implementation Guide. August 1982 edition, Network Information Center, SRI International, Menlo Park, CA, 1982. Available from the NIC by sending network mail to NIC@NIC.

6. Postel, J. Internet Name Server. Network Working Group Internet Experiment Note IEN 116, DARPA Network Working Group, August, 1979.

## Analysis of Gateway Throughtput Report for Mar 24 to Mar 30 1986
(40 total gateways, time covered 6.71 days)

```
                               datagrams        bytes
LSI Gateway Rcvd Totals :       107.4 M      8916.1 M   (avg pkt len= 83.1 bytes)
Mail Bridge Rcvd Totals :        34.2 M      3305.6 M   (avg pkt len= 96.8 bytes)
MB percent of Rcvd Total:        31.8          37.1

Avg Traffic Rvcd/gateway:         2.7 M       222.9 M
Avg Traffic Rcvd/MB     :         4.9 M       472.2 M
MB percent of Average   :       181.9         211.9

LSI Gateway Sent Totals :       106.9 M      8668.9 M   (avg pkt len= 81.1 bytes)
Mail Bridge Sent Totals :        33.4 M      3351.0 M   (avg pkt len= 100.3 bytes
MB percent of Sent Total:        31.2          38.7

Avg Traffic Sent/gateway:         2.7 M       216.7 M
Avg Traffic Sent/MB     :         4.8 M       478.7 M
MB percent of Average   :       178.5         220.9

Mail Bridge Dropped   :           2.1 M    ( 6.2% of MB total sent)
LSI Gateway Dropped   :           3.9 M    ( 3.7% of LSI total sent)
MB percent of Dropped :          52.6

percent pkts addressed to gateways  = 41.66
percent pkts originating at gateways= 44.94
percent pkts forwarded to gateways  = 51.46

Total Packets to Gateways=   44.7M
Total Packets from GWs =     48.0M
Packets forwarded to GWs =   55.1M
Packets forwarded to Hosts=  51.9M
Packets received from Hosts= 52.3M


Conclusions:
  1) Hosts send 52.3M datagrams, of which
       - 3.94M are dropped (7.5%),
       - 0.00M are assumed to be redirects (0.0% of undrpd),
       - an undetermined amount are gw pings.

  2) Therefore, of 107.4M datagrams received by gateways:
       - no more than 48.4M are successful user data (45.1%),
```

# Gateway Traffic for Mar 24 to Mar 30, 1986

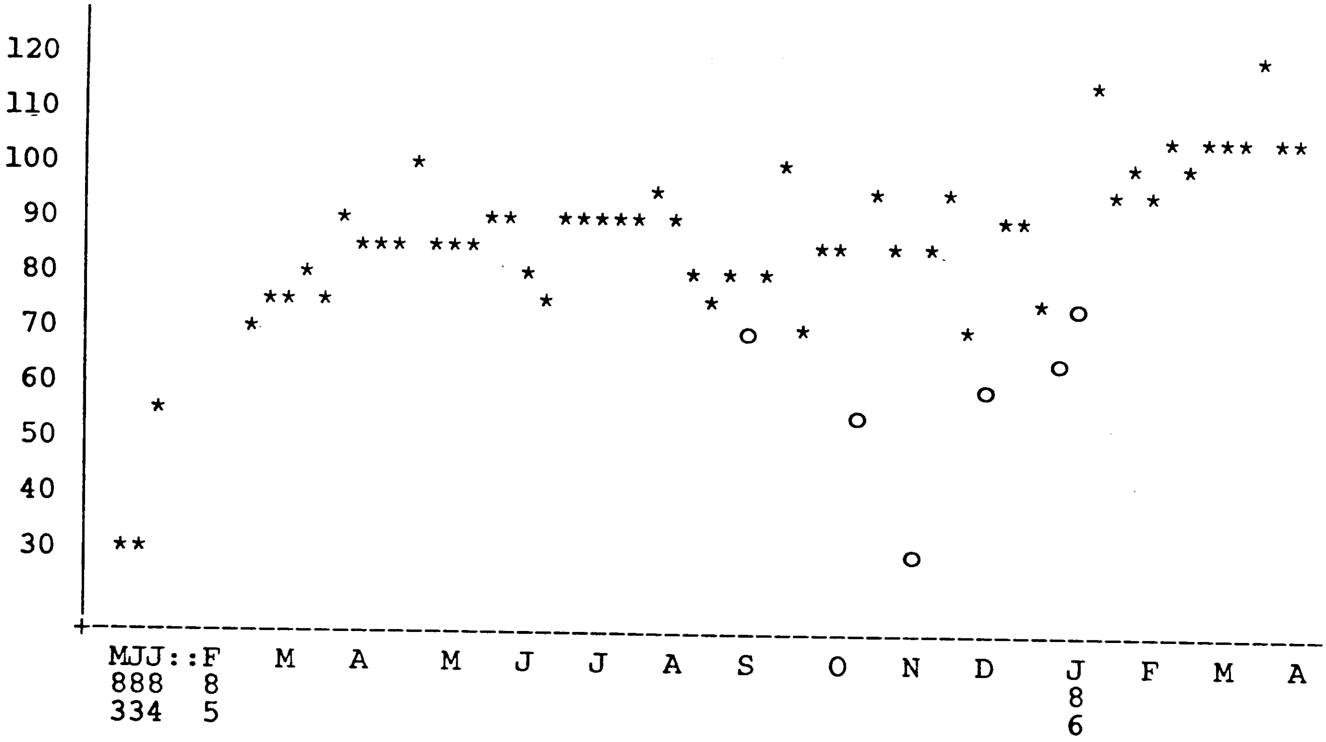
```
    ( 55.1M )    <--         <--         <--         <--    ( 55.1M )
         v                                                       ^
         v                                                       ^
         v                                                       ^
         v                                                       ^
         v                 LSI Gateway System                    ^
         v    +-------------------------------------------+      ^
         v    |  42% To GW= 45M      45% From GW= 48M     |      ^  51.5%
         v    |   ->->| GP              |->->->->->->->   |      ^ forwarded
              |  /    | H.Echo          |GP            \  |
  ( 107.4M ) --->                       |                ---> ( 106.9M )
         ^    |  \ Dropped Internally   |ICMP          /  |      v
         ^    |   ->->|   = 4.7M        |->->->->->->->   |      v
         ^    |    \  | (IP err, Unr,   | SQ,         /   |      v
         ^    |      \|  congestion)    | ER,DU=     /    |      v
         ^    |       \                 | Echo=    /      |      v
         ^    |        \                | RD=     /       |      v
         ^    |         \                        /        |      v
         ^    |          >->     ====>      ->->          |      v
         ^    |                 Thru GW                   |      v
         ^    +-------------------------------------------+      v
         ^                                                       v
         ^                                                       v
         ^                                                       v
         ^                                                       v
    =====^=======================================================v====
    | ( 52.3M )             Host  Population              ( 51.9M ) |
    ==================================================================
```

```
    Analysis of Gateway Throughtput Report for Mar 31 to Apr 6 1986
         (40 total gateways, time covered 6.63 days)

                              datagrams        bytes
LSI Gateway Rcvd Totals :      106.3 M       8834.9 M    (avg pkt len= 83.1 bytes)
Mail Bridge Rcvd Totals :       32.4 M       3050.5 M    (avg pkt len= 94.2 bytes)
MB percent of Rcvd Total:       30.5           34.5

Avg Traffic Rvcd/gateway:        2.7 M        220.9 M
Avg Traffic Rcvd/MB     :        4.6 M        435.8 M
MB percent of Average   :      174.1          197.3

LSI Gateway Sent Totals :      105.0 M       8460.2 M    (avg pkt len= 80.6 bytes)
Mail Bridge Sent Totals :       31.5 M       3122.1 M    (avg pkt len= 99.2 bytes)
MB percent of Sent Total:       30.0           36.9

Avg Traffic Sent/gateway:        2.6 M        211.5 M
Avg Traffic Sent/MB     :        4.5 M        446.0 M
MB percent of Average   :      171.2          210.9

Mail Bridge Dropped    :        2.0 M     ( 6.5% of MB total sent)
LSI Gateway Dropped    :        4.2 M     ( 4.0% of LSI total sent)
MB percent of Dropped :        49.2

percent pkts addressed to gateways  = 41.33
percent pkts originating at gateways= 45.23
percent pkts forwarded to gateways  = 51.42

Total Packets to Gateways=   43.9M
Total Packets from GWs =     47.5M
Packets forwarded to GWs =   54.0M
Packets forwarded to Hosts=  51.0M
Packets received from Hosts= 52.3M


Conclusions:
  1) Hosts send 52.3M datagrams, of which
       - 4.16M are dropped (8.0%),
       - 0.00M are assumed to be redirects (0.0% of undrpd),
       - an undetermined amount are gw pings.

  2) Therefore, of 106.3M datagrams received by gateways:
       - no more than 48.1M are successful user data (45.3%),
```

# Gateway Traffic for Mar 31 to Apr 6, 1986

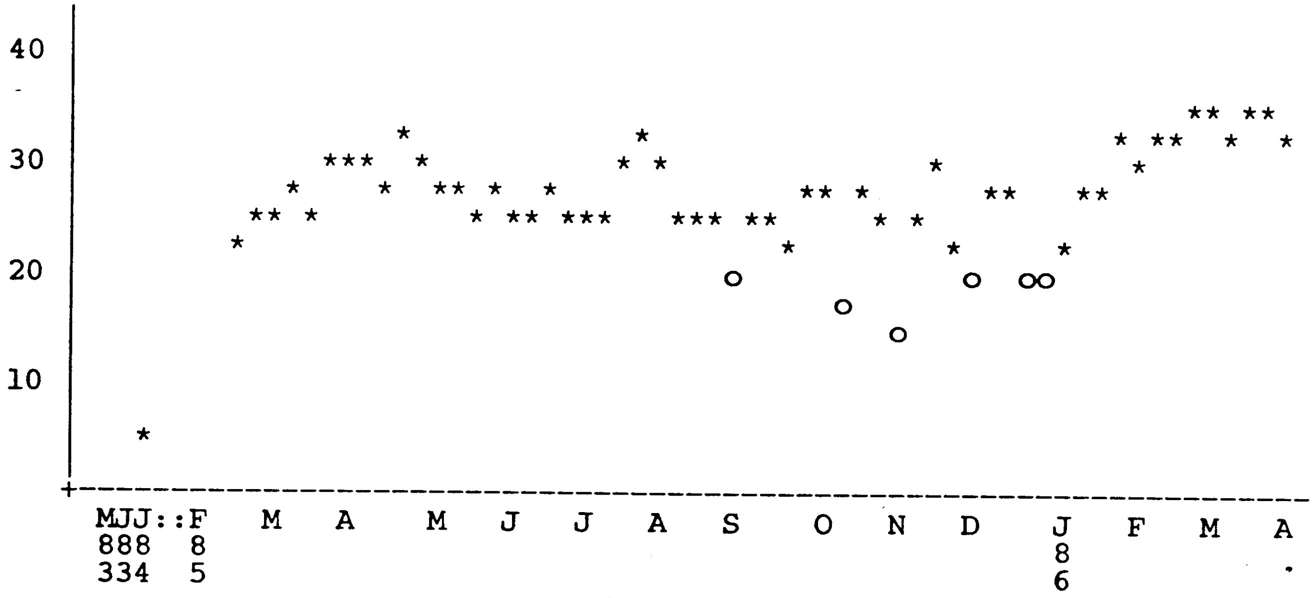
```
  ( 54.M )     <--         <--         <--         <--    ( 54.M )
      v                                                       ^
      v                                                       ^
      v                                                       ^
      v                                                       ^
      v                  LSI Gateway System                   ^
      v     +-------------------------------------------+     ^
      v     |  41% To GW= 44M      45%|From GW= 47.5M   |     ^    51%
      v     |   ->->| GP              |->->->->->->->-> |     ^  forwarded
            |  /    | H.Echo          |GP            \  |
( 106M  )  --->                       |                ---> ( 105M )
      ^     |  \ Dropped Internally   |ICMP          /  |
      ^     |   ->->|   = 5.6M        |->->->->->->->->/ |     v
      ^     |    \  | (IP err, Unr,   | SQ,        /    |     v
      ^     |     \ |  congestion)    | ER,DU=    /     |     v
      ^     |      \                  | Echo=    /      |     v
      ^     |       \                 | RD=     /       |     v
      ^     |        \                         /        |     v
      ^     |         >->      ====>      ->->          |     v
      ^     |                 Thru GW                   |     v
      ^     +-------------------------------------------+     v
      ^                                                       v
      ^                                                       v
      ^                                                       v
      ^                                                       v
                                                              v
  ====^=======================================================v====
  | ( 52.3M )            Host  Population               ( 51.0M ) |
  =================================================================
```

Traffic Sent by LSI Gateways (2/18/85 - 3/31/86)
(in Million pkts)



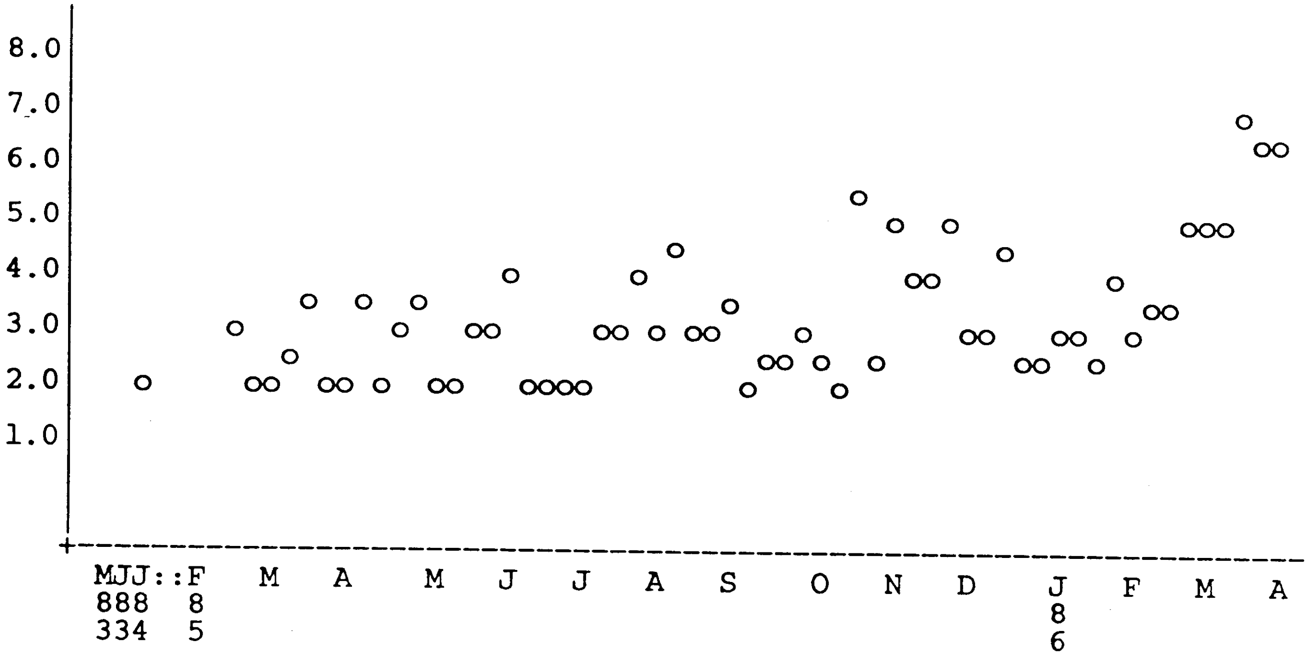
(o denotes incomplete data)

Traffic Sent by Mail Bridges (2/18/85 - 3/31/86)
(in Million pkts)

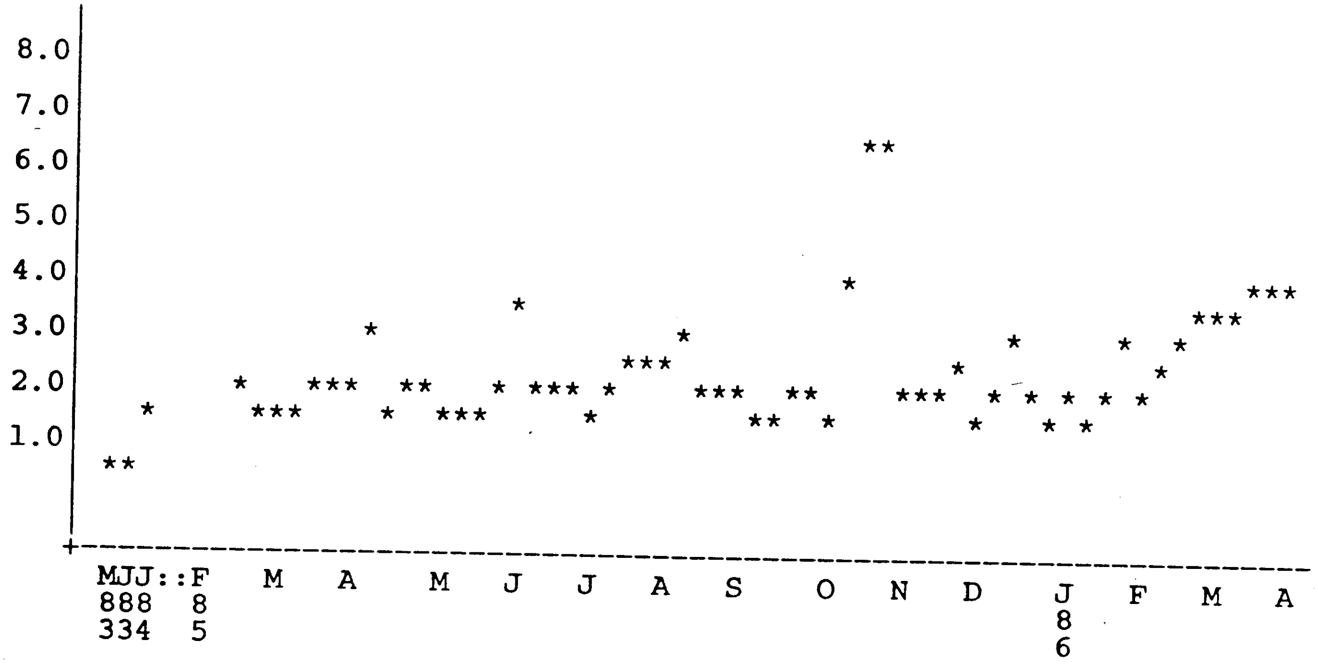


(o denotes incomplete data)

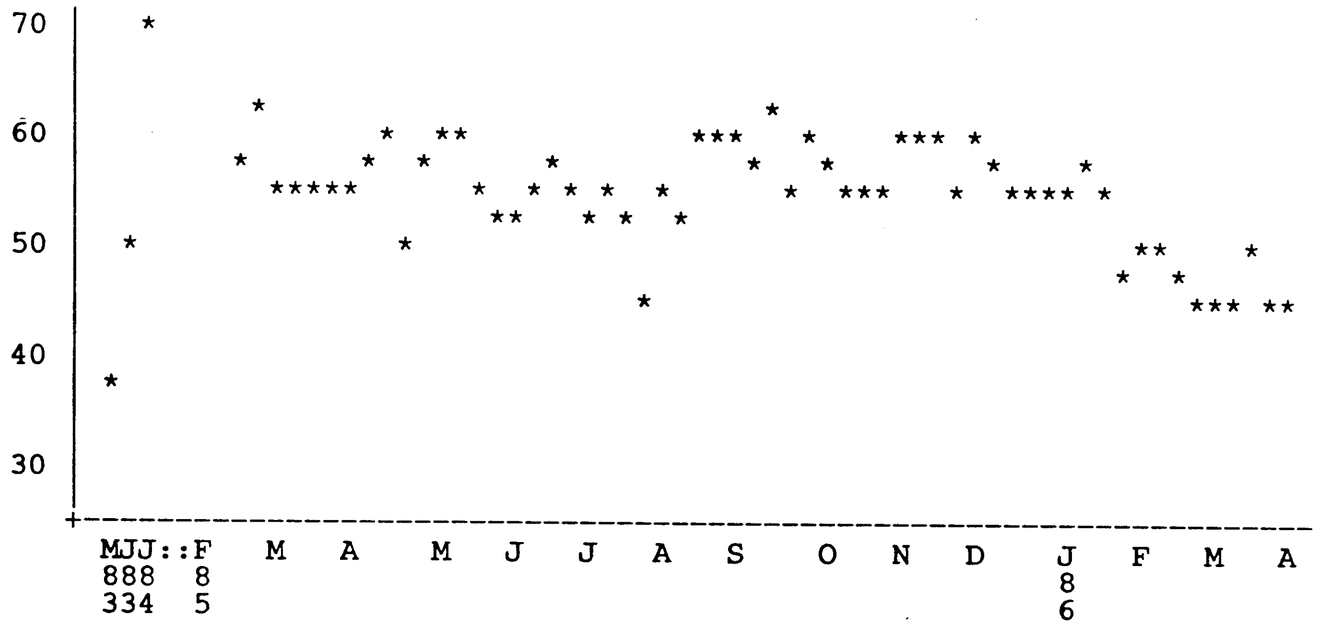
## Percent of Sent Traffic Dropped by Mail Bridges
## (2/18/85 - 3/31/86)



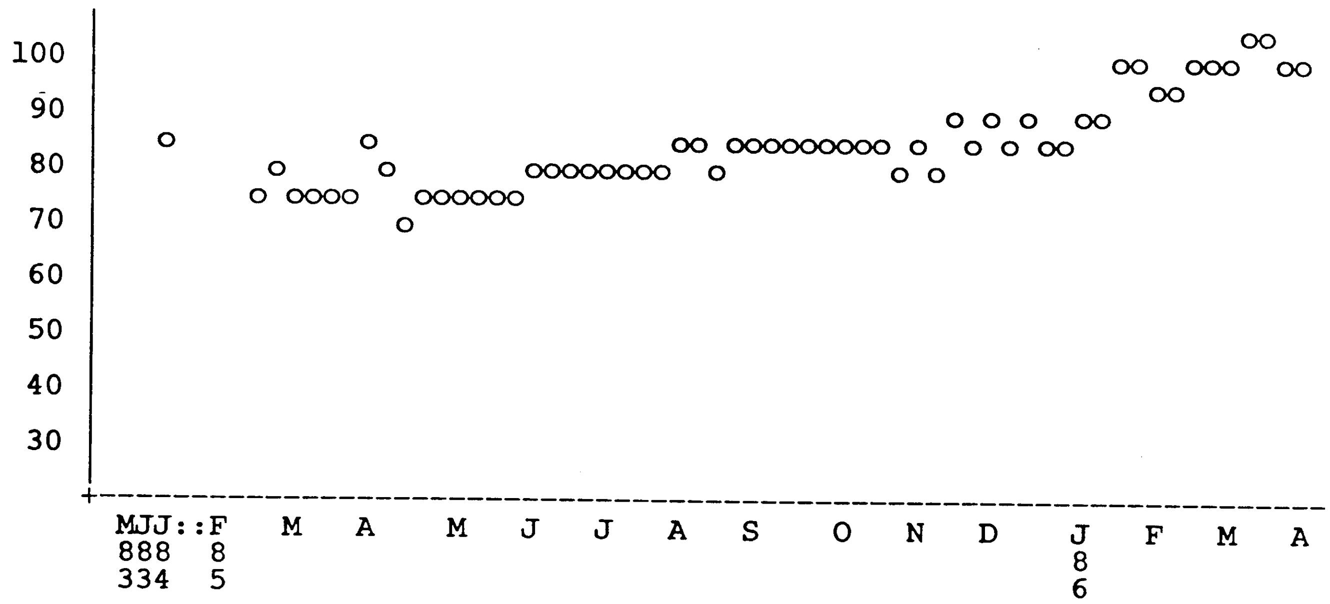
## Percent of Sent Traffic Dropped by LSI Gateways
## (2/18/85 - 3/31/86)



Percent User Data in Gateway Received Traffic
(2/18/85 - 3/31/86)



## Average Packet Length for Mail Bridges
## (2/18/85 - 3/31/86)



(Mailbridge Throughput, April 4, 1986)

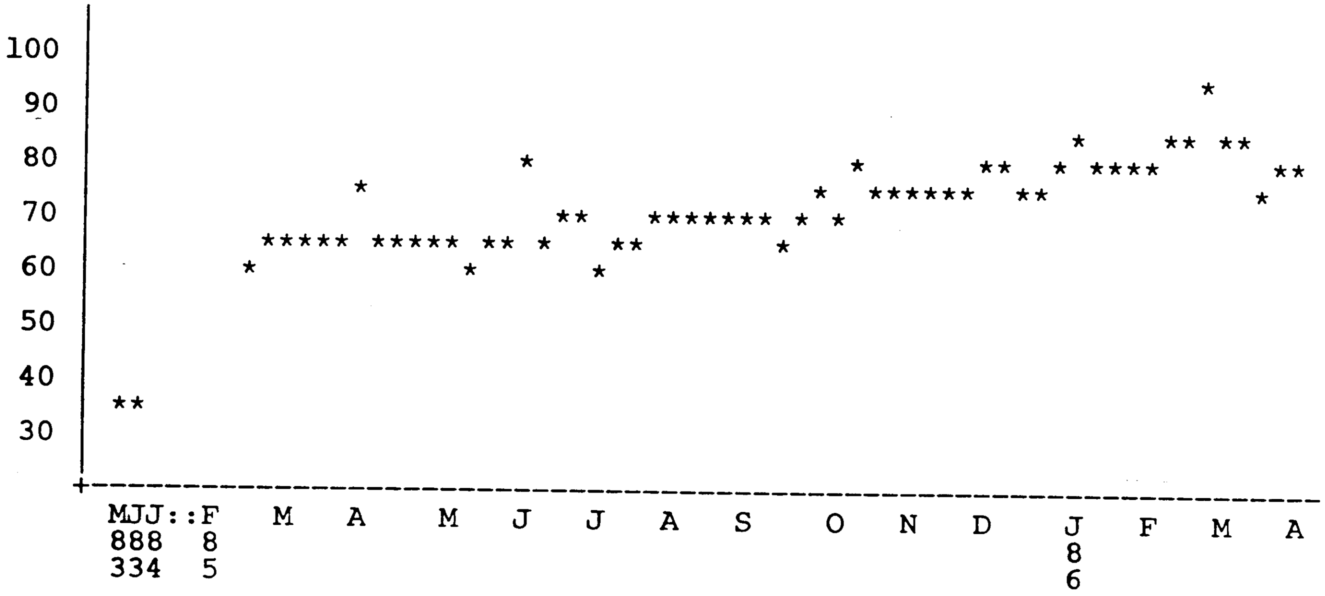
RATE (per second) and SIZE (bytes per datagram) TABLES

| GWY NAME | RCVD DGRAMS | RCVD BYTES | IP ERRORS | AVG BYTES PER DGRAM |
|---|---|---|---|---|
| MILARP | 11.59 | 1987.53 | 0.00 | 171.43 |
| MILBBN | 14.94 | 1448.52 | 0.00 | 96.96 |
| MILDCE | 11.41 | 1397.77 | 0.00 | 122.52 |
| MILISI | 11.09 | 1520.33 | 0.00 | 137.10 |
| MILLBL | 5.87 | 413.66 | 0.00 | 70.41 |
| MILSAC | 6.48 | 586.51 | 0.00 | 90.58 |
| MILSRI | 4.80 | 416.81 | 0.00 | 86.82 |

| GWY NAME | SENT DGRAMS | SENT BYTES | DROPPED DGRAMS | AVG BYTES PER DGRAM |
|---|---|---|---|---|
| MILARP | 10.85 | 1640.90 | 0.96 | 151.20 |
| MILBBN | 14.25 | 1396.16 | 1.06 | 97.96 |
| MILDCE | 11.28 | 1399.16 | 0.29 | 123.99 |
| MILISI | 10.83 | 1470.42 | 0.44 | 135.73 |
| MILLBL | 5.91 | 450.94 | 0.10 | 76.27 |
| MILSAC | 6.55 | 652.96 | 0.14 | 99.75 |
| MILSRI | 4.78 | 432.96 | 0.11 | 90.53 |

33.62% of all received packets are addressed to a gateway
37.47% of all sent packets originate at a gateway
53.36% of all sent packets are forwarded to another gateway

Average Packet Length for LSI Gateways
(2/18/85 - 3/31/86)



## INTERFACE SUMMARY
April 4, 1986

```
Throughput summary for MILBBN Gateway

        Total time covered by data: 22 hours, 15 minutes in 89 messages

        First message received at Fri Apr  4 00:00:20 1986 (EST)
        Last message received at Fri Apr  4 23:50:16 1986 (EST)
        Total elapsed time = 23 hours, 49 minutes, 56 seconds

Datagrams dropped due to unreachable dest net:      18,334 (   0.23/sec   1.5
Datagrams dropped due to unreachable dest host:      3,706 (   0.05/sec   0.3

INTERFACE          RCVD         RCVD       IP    % IP    DGRAMS    % DGMS
                   DGRAMS       BYTES      ERR   ERR     LOOPED    LOOPED

10.5.0.5          602,242   52,095,722     200   0.03    48,527      8.06
26.2.0.49         594,446   63,930,334       0   0.00     3,686      0.62
TOTAL           1,196,688  116,026,056     200   0.02    52,213      4.36

INTERFACE          RCVD       RCVD     % DGRAMS      AVG. BYTES  % RCVD
                   DGM/SEC    BT/SEC   FOR SELF      PER DGRAM   HERE
10.5.0.5           7.52       650.38     26.59         86.50      50.33
26.2.0.49          7.42       798.13     24.09        107.55      49.67
TOTAL             14.94      1448.52     25.35         96.96     100.00

INTERFACE          BUFFER     % BUF
                   DROPPED    DROPPED
10.5.0.5           1,971       0.34
26.2.0.49          1,526       0.26
TOTAL                  0       0.00


INTERFACE          SENT          SENT      RFNM    QUEUE    % DGRAMS    % SENT
                   DGRAMS        BYTES     DROP     DROP    DROPPED     TO NBRS

10.5.0.5          629,500    68,664,005   38,683  20,466       8.59     69.22
26.2.0.49         512,111    43,168,488   19,231   3,179       4.19     32.92
TOTAL           1,141,611   111,832,493   57,914  23,645       6.67     52.94

INTERFACE          SENT       SENT     % DGRAMS      AVG. BYTES  % SENT
                   DGM/SEC    BT/SEC   FROM SELF     PER DGRAM   HERE
10.5.0.5           7.86       857.23     29.75        109.08      55.14
26.2.0.49          6.39       538.93     30.08         84.30      44.86
TOTAL             14.25      1396.16     29.90         97.96     100.00
```

# MAILBRIDGE THROUGHPUT REPORT
April 4, 1986

| GWY NAME | RCVD DGRAMS | RCVD BYTES | IP ERRORS | % IP ERRORS | DEST UNRCH | % DST UNRCH |
|---|---|---|---|---|---|---|
| MILARP | 208,691 | 35,775,596 | 12 | 0.00% | 1,713 | 0.82% |
| MILBBN | 1,196,688 | 116,026,056 | 200 | 0.02% | 22,040 | 1.84% |
| MILDCE | 903,565 | 110,703,268 | 69 | 0.00% | 10,976 | 1.21% |
| MILISI | 878,270 | 120,410,426 | 17 | 0.00% | 12,765 | 1.45% |
| MILLBL | 465,276 | 32,762,004 | 8 | 0.00% | 5,733 | 1.23% |
| MILSAC | 512,844 | 46,451,480 | 70 | 0.01% | 9,196 | 1.79% |
| MILSRI | 380,241 | 33,011,364 | 12 | 0.00% | 9,355 | 2.46% |
| TOTALS | 4,545,575 | 495,140,194 | 388 | 0.00% | 71,778 | 1.58% |

| GWY NAME | SENT DGRAMS | SENT BYTES | DROPPED DGRAMS | % DROPPED DGRAMS |
|---|---|---|---|---|
| MILARP | 195,343 | 29,536,210 | 17,223 | 8.10% |
| MILBBN | 1,141,611 | 111,832,493 | 85,056 | 6.93% |
| MILDCE | 893,710 | 110,813,222 | 22,865 | 2.49% |
| MILISI | 858,021 | 116,457,441 | 35,163 | 3.94% |
| MILLBL | 468,263 | 35,714,671 | 7,975 | 1.67% |
| MILSAC | 518,459 | 51,714,656 | 11,351 | 2.14% |
| MILSRI | 378,754 | 34,290,426 | 8,525 | 2.20% |
| TOTALS | 4,454,161 | 490,359,119 | 188,158 | 4.05% |

## SECTION OF DAILY TRAP REPORT

April 4, 1986

| | |
|---|---|
| GWY MILBBN T1006 Time expired | 657 |
| GWY MILBBN T1012 Received ICMP | 479 |
| GWY MILBBN T1014 Unusual 1822 reply | 925 |
| GWY MILBBN T1015 Unmatched 1822 reply | 328 |
| GWY MILBBN T1022 Sending Source Quench | 1398 |
| GWY MILBBN T1024 Overdue RFNM | 321 |
| GWY MILBBN T1048 ICMP -> ICMP | 6226 |
| GWY MILBBN T1509 Thrpt Meas. On | 2 |
| GWY MILBBN T1517 Thrpt Meas. Off | 2 |
| GWY MILBBN T1520 Lost traps | 389 |
| GWY MILBBN T2001 Neighbor down | 62 |
| GWY MILBBN T2004 Neighbor Up | 132 |
| GWY MILBBN T2008 Interface Up | 8 |
| GWY MILBBN T2011 New net | 33 |
| GWY MILBBN T2016 Redundant route | 5 |
| GWY MILBBN T2024 Nets full | 33 |
| TOTAL | 11000 |

## MAILBRIDGE THROUGHPUT, SHORT FORM

Data sorted by mailbridge

| DATE | HOURS COVERED | RCVD DGRAMS | DST UNRCH | DROPPED DGRAMS | RCVD PPS DGRAMS | SENT PPS DGRAMS |
|---|---|---|---|---|---|---|
| Gateway: MILARP | | | | | | |
| 3/31 (Mon) | 22.00 | 763,738 | 0.78% | 4.31% | 9.64 | 9.74 |
| 4/1 (Tue) | 21.50 | 870,606 | 1.17% | 4.79% | 11.25 | 11.23 |
| 4/2 (Wed) | 24.00 | 822,043 | 1.21% | 3.72% | 9.51 | 9.52 |
| 4/3 (Thu) | 15.50 | 505,128 | 0.71% | 2.05% | 9.05 | 9.12 |
| 4/4 (Fri) | 5.00 | 208,691 | 0.82% | 8.10% | 11.59 | 10.85 |
| Gateway: MILBBN | | | | | | |
| 3/31 (Mon) | 21.75 | 1,120,717 | 2.48% | 21.03% | 14.31 | 12.03 |
| 4/1 (Tue) | 22.50 | 1,464,874 | 1.96% | 21.89% | 18.08 | 15.21 |
| 4/2 (Wed) | 24.25 | 1,599,394 | 3.01% | 15.49% | 18.32 | 16.20 |
| 4/3 (Thu) | 23.00 | 1,186,268 | 1.51% | 7.15% | 14.33 | 14.02 |
| 4/4 (Fri) | 22.25 | 1,196,688 | 1.84% | 6.93% | 14.94 | 14.25 |
| 4/5 (Sat) | 24.00 | 861,729 | 2.11% | 1.61% | 9.97 | 10.04 |
| 4/6 (Sun) | 21.50 | 657,499 | 2.61% | 0.76% | 8.49 | 9.04 |

.....

## SUMMARIES

| DATE | ADDR TO | ORIG FROM | FORWARDED | TOTAL RCVD |
|---|---|---|---|---|
| 3/31 (Mon) | 48.63% | 55.28% | 60.51% | 5,358,788 |
| 4/1 (Tue) | 45.68% | 52.85% | 59.66% | 6,205,279 |
| 4/2 (Wed) | 40.34% | 46.17% | 56.03% | 6,160,927 |
| 4/3 (Thu) | 32.07% | 36.70% | 47.67% | 4,769,165 |
| 4/4 (Fri) | 33.62% | 37.47% | 53.36% | 4,545,575 |
| 4/5 (Sat) | 33.73% | 36.79% | 52.52% | 3,296,433 |
| 4/6 (Sun) | 36.41% | 41.49% | 48.92% | 2,052,551 |

......

| | | | | | | |
|---|---|---|---|---|---|---|
| date: 4/4 (Fri) | | | | | | |
| MILARP | 5.00 | 208,691 | 0.82% | 8.10% | 11.59 | 10.85 |
| MILBBN | 22.25 | 1,196,688 | 1.84% | 6.93% | 14.94 | 14.25 |
| MILDCE | 22.00 | 903,565 | 1.21% | 2.49% | 11.41 | 11.28 |
| MILISI | 22.00 | 878,270 | 1.45% | 3.94% | 11.09 | 10.83 |
| MILLBL | 22.00 | 465,276 | 1.23% | 1.67% | 5.87 | 5.91 |
| MILSAC | 22.00 | 512,844 | 1.79% | 2.14% | 6.48 | 6.55 |
| MILSRI | 22.00 | 380,241 | 2.46% | 2.20% | 4.80 | 4.78 |

```
                        25 BUSIEST HOSTS IN ARPANET

Host Throughput       From Fri Mar 21 00:00:45 1986
                      To   Fri Mar 28 00:00:45 1986

Host Name          {node/         Packets Received              Avg. Daily  Days
                    host} Inter-Node   Intra-Node    Total  Inter-Node

BBN-MILNET-GW     {  5/5}    5378046     436421    5814467
WISC-GATEWAY      { 94/0}    4273220      92506    4365726
DCEC-MILNET-GW    { 20/7}    3943193     257630    4200823
ISI-GATEWAY       { 27/3}    3637079      37904    3674983
ISI-MILNET-GW     { 22/2}    3136629      18914    3155543
PURDUE-CS-GW      { 37/2}    2751963     265814    3017777
ARPA-MILNET-GW    { 28/2}    2734553      14718    2749271
CSS-GATEWAY       { 25/2}    2713281      26328    2739609
MIT-MC            { 44/3}    2620395      25232    2645627
MIT-GW            { 77/0}    2547447      22237    2569684
MIT-AI-GW         {  6/3}    2254270      64268    2318538
CMU-CS-A          { 14/1}     269950    1925422    2195372
SRI-MILNET-GW     { 51/4}    1184252    1009017    2193269
YALE              {  9/2}    2062942     113304    2176246
GW.RUTGERS.EDU    { 89/1}    1916361      60114    1976475
COLUMBIA          { 89/3}     805962    1095844    1901806
USC-ISID          { 27/0}    1709510      10000    1719510
STANFORD-GW       { 11/1}    1702967      12776    1715743
UCB-VAX           { 78/2}    1608650      62115    1670765
SAC-MILNET-GW     { 80/2}    1611828      50133    1661961
LBL-MILNET-GW     { 68/0}    1622285      20609    1642894
CMU-GATEWAY       { 14/2}    1086649     516398    1603047
SEISMO            { 25/0}    1512925      34390    1547315
BBN-INOC          { 82/2}    1341485     163180    1504665
CSNET-RELAY       {  5/4}    1464005      31413    1495418
```

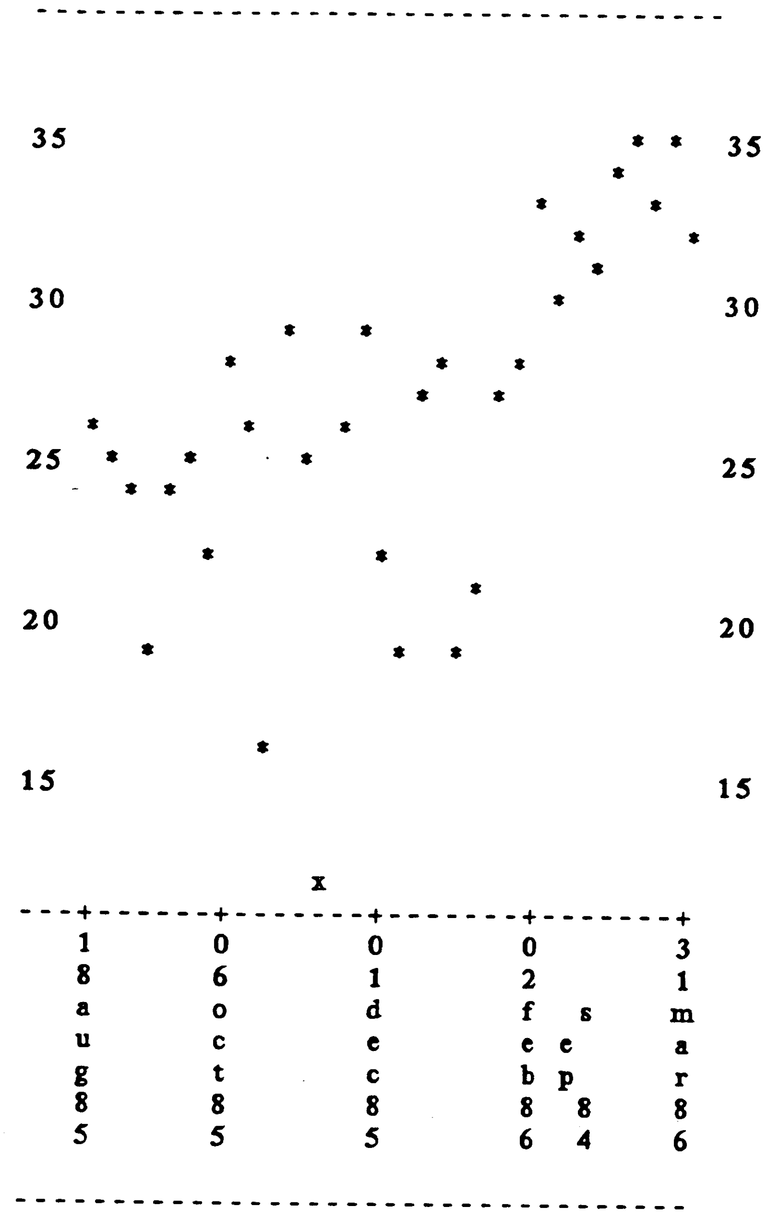
```
Host Throughput    From Fri Mar 21 00:00:45 1986
                   To   Fri Mar 28 00:00:45 1986

Host Name          {node/          Packets Received              Avg. Daily  Days
                    host} Inter-Node  Intra-Node     Total   Inter-Node


UCLA-TEST          {  1/0}     14917           0      14917
UCLA-CCN           {  1/1}     39991        9611      49602
UCLA-LOCUS         {  1/2}    374378        4371     378749
UCLA-ATS           {  1/3}     34709           2      34711
                           ---------   ---------  ---------
                              463995       13984     477979      66285      7

SRI-SPRM           {  2/0}     23221           0      23221
SRI-KL             {  2/1}    413611      141838     555449
SRI-CSL-GW         {  2/2}    815320       38002     853322
SRI-TSC            {  2/3}     98039       16826     114865
SRI-AI             {  2/4}    389464      309585     699049
SRI-IU             {  2/5}    299206      325317     624523
SRI-MCON-GW        {  2/6}         0           0          0
                           ---------   ---------  ---------
                             2038861      831568    2870429     291265      7

SAC-RPVAX          {  3/0}         0           0          0
SAC-RPGW-1         {  3/1}     43387      349142     392529
SAC-RPGW-2         {  3/2}     90277      349755     440032
                           ---------   ---------  ---------
                              133664      698897     832561      19094      7

UTAH-CS            {  4/0}    427928        4518     432446
FSNAP-GW           {  4/1}       134           0        134
UTAH-TAC           {  4/2}    118365         291     118656
UTAH-20            {  4/3}     61582       13291      74873
                           ---------   ---------  ---------
                              608009       18100     626109      86858      7

BBN-CLXX           {  5/0}    127425       11466     138891
BBNG               {  5/1}   1275714      132032    1407746
FIBER-ARPA-GW      {  5/2}         0           0          0
BBNA               {  5/3}    161498      159605     321103
CSNET-RELAY        {  5/4}   1464005       31413    1495418
BBN-MILNET-GW      {  5/5}   5378046      436421    5814467
BBN-PR-GW          {  5/6}   1356551      130946    1487497
BBN-PR-STATION     {  5/7}         1           0          1
                           ---------   ---------  ---------
                             9763240      901883   10665123    1394748      7
```

GATEWAY: MILBBN

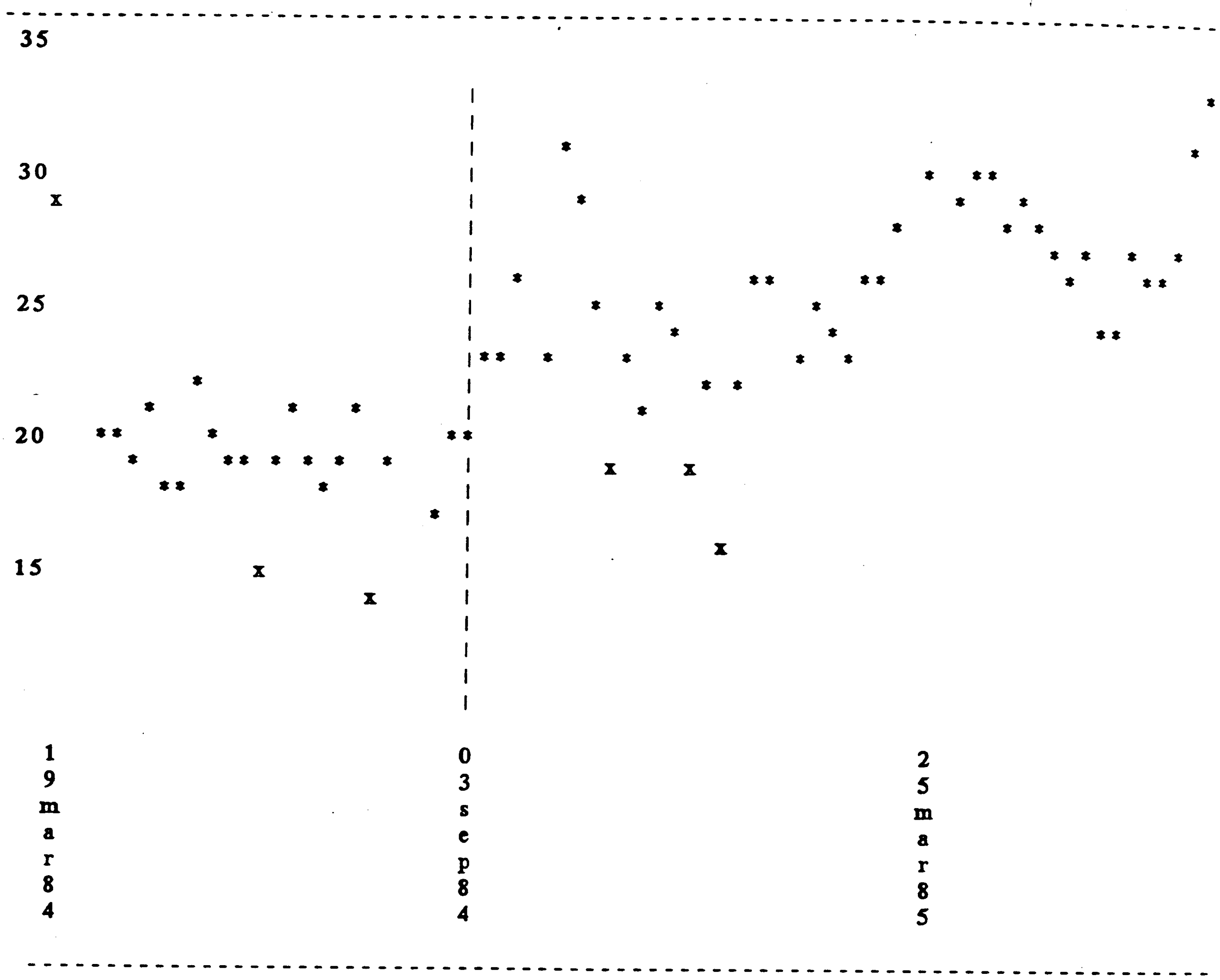
| DATE | NET UNR | %NET UNR | HOST UNR | %HOST UNR |
|---|---|---|---|---|
| 3/31 (Mon) | 17,690 | 1.58 | 10,085 | 0.90 |
| 4/1 (Tue) | 21,015 | 1.43 | 7,739 | 0.53 |
| 4/2 (Wed) | 41,369 | 2.59 | 6,788 | 0.42 |
| 4/3 (Thu) | 14,388 | 1.21 | 3,548 | 0.30 |
| 4/4 (Fri) | 18,334 | 1.53 | 3,706 | 0.31 |
| 4/5 (Sat) | 13,925 | 1.62 | 4,222 | 0.49 |
| 4/6 (Sun) | 12,194 | 1.85 | 4,969 | 0.76 |

| DATE | RCVD DGM | LOOPED | %LOOPED | BUF DROP | %BUF DROP |
|---|---|---|---|---|---|
| INTERFACE: | 10.5.0.5 | | | | |
| 3/31 (Mon) | 751,442 | 67,358 | 8.96% | 38,449 | 5.24% |
| 4/1 (Tue) | 945,753 | 106,324 | 11.24% | 61,658 | 6.53% |
| 4/2 (Wed) | 1,002,336 | 97,571 | 9.73% | 30,693 | 3.12% |
| 4/3 (Thu) | 694,557 | 65,609 | 9.45% | 3,539 | 0.53% |
| 4/4 (Fri) | 602,242 | 48,527 | 8.06% | 1,971 | 0.34% |
| 4/5 (Sat) | 487,227 | 38,782 | 7.96% | 7 | 0.00% |
| 4/6 (Sun) | 394,446 | 103,679 | 26.28% | 0 | 0.00% |
| INTERFACE: | 26.2.0.49 | | | | |
| 3/31 (Mon) | 369,275 | 23,812 | 6.45% | 32,582 | 8.69% |
| 4/1 (Tue) | 519,121 | 25,617 | 4.93% | 46,378 | 8.65% |
| 4/2 (Wed) | 597,058 | 16,643 | 2.79% | 22,174 | 3.73% |
| 4/3 (Thu) | 491,711 | 6,267 | 1.27% | 2,569 | 0.53% |
| 4/4 (Fri) | 594,446 | 3,686 | 0.62% | 1,526 | 0.26% |
| 4/5 (Sat) | 374,502 | 778 | 0.21% | 13 | 0.00% |
| 4/6 (Sun) | 263,053 | 2,859 | 1.09% | 0 | 0.00% |

| DATE | SENT DGM | RFNM DROP | %RFNM | QUE DROP | %QUE DROP | %DROP |
|---|---|---|---|---|---|---|
| INTERFACE: | 10.5.0.5 | | | | | |
| 3/31 (Mon) | 355,336 | bad data | | | | |
| 4/1 (Tue) | 556,439 | | | | | |
| 4/2 (Wed) | 641,140 | | | | | |
| 4/3 (Thu) | 584,082 | 27,592 | 4.32% | 26,459 | 4.15% | 8.47% |
| 4/4 (Fri) | 629,500 | 38,683 | 5.62% | 20,466 | 2.97% | 8.59% |
| 4/5 (Sat) | 422,252 | 6,832 | 1.59% | 1,680 | 0.39% | 1.98% |
| 4/6 (Sun) | 397,859 | 2,902 | 0.72% | 64 | 0.02% | 0.74% |
| INTERFACE: | 26.2.0.49 | | | | | |
| 3/31 (Mon) | 586,296 | 22,656 | 3.66% | 9,331 | 1.51% | 5.17% |
| 4/1 (Tue) | 675,310 | 45,897 | 6.16% | 24,267 | 3.26% | 9.41% |
| 4/2 (Wed) | 773,070 | 34,297 | 4.17% | 14,506 | 1.76% | 5.94% |
| 4/3 (Thu) | 576,387 | 25,806 | 4.26% | 3,394 | 0.56% | 4.82% |
| 4/4 (Fri) | 512,111 | 19,231 | 3.60% | 3,179 | 0.59% | 4.19% |
| 4/5 (Sat) | 445,178 | 5,638 | 1.25% | 0 | 0.00% | 1.25% |
| 4/6 (Sun) | 302,223 | 2,404 | 0.79% | 0 | 0.00% | 0.79% |



MILLIONS OF DATAGRAMS RECEIVED PER WEEK



MILLIONS OF DATAGRAMS RECEIVED PER WEEK

PROBLEM: POOR MAILBRIDGE PERFOMANCE

FACT1: EACH HOST-HOST CONNECTION NEEDS:

CONNECTION BLOCK AT THE SOURCE
RECEIVE BLOCK AT THE DESTINATION

FACT 2: GATEWAY PINGS OCCUPY A CONNECTION BLOCK

FACT 3: MAILBRDIGE IMPS ARE SHORT OF CONNECTION BLOCKS

FACT 4: WHEN A MESSAGE ARRIVES AND NO CONNECTION BLOCK IS AVAILABLE, ONE MUST BE TORN DOWN

FOUR MESSAGES EXCHANGED

LEAST RECENTLY USED

| Mailbridge | AUG-NOV-85 | JAN-86 |
|---|---|---|
| MILARPA | 6.4 | 6.4 |
| MILBBN | 10.4 | 11.3 |
| MILDCEC | 6.6 | 6.3 |
| MILISI | 9.7 | 9.4 |
| MILLBL | 5.7 | 4.8 |
| MILSAC | 5.0 | 5.1 |
| MILSRI | 3.6 | 3.6 |

| Mailbridge | FEB-86 | MAR-86 |
|---|---|---|
| MILARPA | 9.0 | 10.05 |
| MILBBN | 13.8 | 15.1 |
| MILDCEC | 7.4 | 8.5 |
| MILISI | 8.6 | 9.3 |
| MILLBL | 6.4 | 6.9 |
| MILSAC | 5.8 | 6.6 |
| MILSRI | 4.3 | 4.9 |

| Mailbridge | PRE-SPLIT | POST-SPLIT |
|---|---|---|
| MILARPA | 6-7 pps | 6-7.5 pps |
| MILBBN | 9-10 | 10-11.5 |
| MILDCEC | 3 | 6-7 |
| MILISI | 5-6 | 5-7 |
| MILLBL | ** | 4-5 |
| MILSAC | 3-4 | 4-5 |
| MILSRI | 5 | 3 |

| Mailbridge | MARCH-85 | SUMMER |
|---|---|---|
| MILARPA | 8-9 pps | 6-7 pps |
| MILBBN | 8-9 | 9-11 |
| MILDCEC | 8-9 | 7-8 |
| MILISI | 9-10 | 9.5-11.5 |
| MILLBL | 5-6 | 5-6.5 |
| MILSAC | 5-6 | 5-6 |
| MILSRI | 3-4 | 3.5-4 |

## FROM THE DECEMBER QUARTERLY STATISTICS:

95% OF ALL TRAPS RECEIVED ARE DUE TO LONG WAITS ( >3 SEC) FOR END-END RESOURCES

MOST OF THESE ORIGINATE FROM

SRI-2
RCC-5 (MILBBN)
STAN-11
DCEC-20 (MILDCEC)
ISI-22 (MILISI)
ISI-27
ARPA-27 (MILARP)
SRI-51 (MILSRI)
(BERK-78)
SAC-80 (MILSAC)
SR-107

MOST ARE DESTINED FOR

RCC-5
ISI-22
ISI-27
PURDU-37
SRI-51
BBN-82 (INOC)
(BBN-89)
WISC-94

NOTE: LBL-68 HAS ONLY ONE HOST ON IT:
MILLBL

| | |
|---|---|
| PSN 3/4 | 73 CONNECTION BLOCKS |
| PSN 5 | 255 CONNECTION BLOCKS |

SOLUTION: UPGRADE TO PSN 5

UPGRADE IN PROGRESS