Proceedings of the Ninth
Internet Engineering Task Force
March 1-3, 1988 in San Diego

Edited by

Phillip Gross

Allison Mankin

May 1988

NINTH IETF

The MITRE Corporation
Washington C^3I Operations
7525 Colshire Drive
McLean, Virginia 22102

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Concluded)

ACKNOWLEDGMENTS

As you can tell from the size of this document, producing the Proceedings for the quarterly plenary sessions of the IETF is no longer a trivial matter. Fortunately, there were many contributors to the effort.

Allison Mankin, Coleman Blake, Phill Gross and Anne Whitaker (all from MITRE) wrote selected sections of the meeting notes. Allison compiled the initial draft of the document, while John Biviano (MITRE) and Phill Gross finished compiling and assembling the final version. Several presenters (in particular, Van Jacobson (LBL), Dave Borman (Cray) and Rob Hagens (UWisc) either contributed to or proof read the description of their talks. Phill Gross and Richard Wilmer (MITRE) proofread and edited the final document.

Reporting of Working Group activity is becoming increasing important. For this IETF plenary, eight of the ten Working Groups contributed to the Proceedings. Charles Hedrick (Rutgers) and Allison Mankin deserve particular credit for producing timely reports after the plenary and distributing them to the IETF mailing list. We encourage all Working Groups to be responsive to the Internet community in this way in the future (see Chapter 1, the Chairman's Introduction).

Finally, I'd like to thank Paul Love of the San Diego Supercomputer Center (SDSC) for hosting the March 1-3, 1988 meeting. As the size and activity of the IETF has grown, hosting the plenary has also become a non-trivial undertaking. Paul and SDSC were model hosts, with ample meeting rooms, access to terminals with Internet connectivity, timely refreshments, and an interesting tour of the facilities.

1.0 CHAIRMAN'S INTRODUCTION

The IETF has been both blessed and cursed with success. Over the last year and a half, the group has greatly expanded in size and scope. The combined mailing lists (ietf-tf@isi.edu and ietf-interest@isi.edu) now contain over 250 names with over a dozen secondary mail exploders. The IETF has become a focus for a number of very important Internet efforts (e.g., EGP3, the Host Requirements document, and Network Management of TCP/IP-based Internets to name only three). Because of the importance and visibility of its work, the IETF has a responsibility to the whole Internet community.

There are now 17 IETF Working Groups (WGs). Some groups are now concluding their mission, while others are just getting started. The current groups are:

Working Group	Chair
Authentication	stjohns@sri-nic.arpa
CMIP-based Network Management (NETMAN)	cel@mitre-bedford.arpa
Domains	louie@trantor.umd.edu
EGP3	mgardner@alexander.bbn.com
InterNICs	feinler@sri-nic.arpa
Internet Host Requirements	braden@isi.edu
Internet Management Information Base	craig@bbn.com
Landmark Routing	tsuchiya@gateway.mitre.org
OSI Technical Issues	mrose@twg.com
Open SPF-based IGP	petry@trantor.umd.edu/jmoy@proteon.com
Open Systems Internet Operations Ctr	case@utkux1.utk.edu
Open Systems Routing	hinden@bbn.com
PDN Routing Group	roki@isi.edu
Performance and Congestion Control	mankin@gateway.mitre.org
Short-Term Routing	hedrick@aramis.rutgers.edu
SNMP Extensions	mrose@twg.com
TELNET Linemode	dab%oliver.cray.com@uc.msc.umn.edu

As originally conceived, WGs were meant to have a clearly defined objective and a possibly fixed (i.e., short) life span. The groups were meant to be somewhat autonomous, meeting independently of the quarterly IETF plenary meetings and setting up their own mailing lists. Several groups have done this. In the interest of progress, WG Chairs could stipulate that membership to the group was either open or closed. Most importantly, WGs would promptly report status and progress back to the the full IETF. For example, this might be done as a written report to the IETF mailing list after each occasion that the WG meets.

I encourage all groups to follow these guidelines and would particularly emphasize that each group should keep the full IETF informed of its progress. If a group meets at an IETF plenary, the group should submit a report to include in the Proceedings for that meeting (eight of ten groups from the last meeting have submitted reports for these Proceedings). If a WG meets between IETFs, it is important that a (possibly, brief) set of meeting notes be submitted to the full IETF list (ietf@isi.edu).

I also encourage WGs to meet between IETF meetings, if that is appropriate. Much of the work being done is important enough that it should have more activity than four meetings a year. Again, several groups have already done this and I think this is a good sign. This would also make the Plenary meetings less hectic and reduce the frustration when many of the interesting WGs overlap.

To further help with IETF administration, I sent out a request for information from each working group. This information included such boilerplate info as name and mailing list, but it also asked for more dynamic info like projected WG lifetime and status. I have received this information from most of the 17 WGs. This information will be collected and issued as an IDEA to make the information widely available. The information will be periodically updated to help in tracking progress.

I would be remiss in this message if I did not also take the opportunity to thank all those who have contributed so much to the many successful IETF activities over the last year. There are so many that I won't try to list them here for fear of leaving someone out. With their continuing help, I'm not worried about the "curse" of IETF growth.

2.0 IETF ATTENDEES

Name	Organization	Email Address
Almquist, Phillip	Stanford University	almquist@jessica.standford.edu
Almes, Guy	Rice University	almes@rice.edu
Baker, Peter	UNISYS	baker@jove.cam.unisys.com
Ben-Artzi, Amatzia	Sytek	amatzia@miasma.standford.edu
Berggreen, Art	ACC	art@acc.arpa
Blake, Coleman	MITRE	cblake@gateway.mitre.org
Borman, David	Cray Research	dab%hall.cray.com@uc.msc.umn..edu
Bosak, Len	Cisco	Bosack@methom.cisco.com
Braden, Bob	USC/ISI	braden@venera.isi.edu
Braun, Hans-Werner	U of Michigan	hwb@mcv.umich.edu
Brescia, Mike	BBNCC	brescia@park-street.bbn.com
Broersma, Ron	NOSC	ron@nosc.mil
Brim, Scott	Cornell Theory Ctr	swb@tcgould.tn.cornell.edu
Brinkley, Don	Unisys	don@mcl.unisys.com
Brown, Alison	Cornell Theory Ctr	alison@tcgould.tn.cornell.edu
Brunner, Thomas Eric	SRI International	brunner@span.istc.sri.com
Callon, Ross	BBNCC	rcallon@bbn.com
Case, Jeff	Univ of Tenn	case%utkvxl.decnet@utkcs2.cs.utk.edu
Cerf, Vint	Nat'l Research Initiatives	cerf@a.isi.edu
Chiappa, Noel	MIT	jnc@xx.lcs.mit.edu
Clark, Pat	Ford	paclark@ford-cos1.arpa
Crumb, Steve	NCSA	scrumb@newton.ncsa.uiuc.edu
Davin, Chuck	Proteon	jrd@monk.proteon.com
Deering, Steve	Stanford University	deering@pescardero.stanford.edu
Dunford, Steve	UNISYS	dunford@jove.cam.unisys.com
Enger, Robert	CONTEL SPACECOM	enger@bluto.scc.com
Fedor, Mark	NYSERNET	fedor@nic.nyser.net
Foster, Robb	BBNCC	robb@park-street.bbn.com
Gross, Phill	MITRE	gross@gateway.mitre.org
Hagens, Robert	U of Wisconsin	hagens@cs.wisc.edu
Hammett, Jeff	UNISYS	---------------
Hedrick, Charles	Rutgers University	hedrick@aramis.rutgers.edu
Heker, Sergio	JVNC	heker@junca.csc.org
Hobby, Russell	UC-Davis	rdhobby@ucdavis.edu
Jacobsen, Ole	ACE	ole@csli.stanford.edu
Jacobson, Van	LBL	van@lbl-csam.arpa
Joshi, Satish	ACC	satish@acc-sb-unix.arpa
Karels, Mike	UC Berkeley	karels@ucbvax.berkeley.edu
Karn, Phil	Bellcore	karn@thumper.bellcore.com
LaBarre, Lee	MITRE	cel@mitrebedford.arpa
Larson, John	Xerox PARC	JLarson.pa@xerox.com
Lekashman, John	NASA/NAS	lekash@orville.nas.nasa.gov
Lepp (Gardner), Marianne	BBNCC	mgardner@park-street.bbn.com

Lottor, Mark	SRI NIC	mkl@sri-nic.arpa
Lynch, Dan	ACE	lynch@a.isi.edu
Mamakos, Louis	Univ of MD	louie@trantor.umd.edu
Mankin, Allison	MITRE	mankin@gateway.mitre.org
Mathis, Jim	Apple	------------
McCloghrie, Keith	TWG	kzm@twg.arpa
Medin, Milo	NASA/NAS	medin@ames-titan.arpa
Melohn, Bill	Sun Microsystems	mehohn@sun.com
Messing, Judy	UNISYS.COM	judy@MCL.UNISYS.COM
Mockapetris, Paul	USC/ISI	pvm@venera.isi.edu
Morris, Don	NCAR	morris@windom.ucar.edu
Moy, John	Proteon	jmoy@monk.proteon.com
Mundy, Russ	DCA	mundy@ddn1.arpa
Nakassis, Tassos	NBS	nakassis@icst-ecf.arpa
Natalie, Ron	Rutgers Univ	ron@rutgers.edu
Partridge, Craig	BBNCC	craig@nnsc.nsf.net
Perkins, Drew	CMU	ddp@andrew.cmu.edu
Petry, Mike	Univ of MD	petry@trantor.umd.edu
Ramakrishnan, K.	DEC	rama%erlang.dec@decwrl.dec.com
Reynolds, Joyce	USC/ISI	jkrey@venera.isi.edu
Robertson, Jim	Bridge	------------
Rochlis, Jon	MIT	jon@athena.mit.edu
Rokitansky, Carl-Herb.	DFVLR, West Germany	roki@isia.edu
Rose, Marshall	TWG	mrose@twg.arpa
Satz, Greg	Cisco	satz@mathom.cisco.com
Schiller, Jeff	MIT	jis@bitsy.mit.edu
Schoffstall, Marty	Nysernet	schoff@nisc.nyser.net
Schofield, Bruce	DCEC	schofield@edn-vax.arpa
Singh, Aditya	Nynex S&T	singh@nynexst.comm
Stahl, Mary	SRI-NIC	stahl@sri-nic.arpa
St. Johns, Michael	USAF	stjohns@sri-nic.arpa
Stone, Geoff	Network Sys. Corp.	stone@orville.nas.nasa.gov
Su, Zaw-Sing	SRI	zsu@tcsa.ista.sri.edu
Trewitt, Glenn	Stanford University	trewitt@amadeus.stanford.edu
Tsuchiya, Paul	MITRE	tsuchiya@gateway.mitre.org
Veach, Ross	Univ, of Illinois	rrv@uxc.cso.uiuc.edu
Waldbusser, Steve	CMU	waldbusser@andrew.cmv.edu
Whitaker, Anne	MITRE	whitaker@gateway.mitre.org
Zhang, Lixia	MIT	lixia@xx.lcs.mit.edu

3.0 FINAL AGENDA

TUESDAY, March 1

8:30 am Opening Plenary (Introductions and local arrangements)
8:45 am Working Group meetings convene

- Open IGP (Petry, UMD/Moy, Proteon)
- Open Systems Routing (Callon, BBN)
- Open Systems Internet Operations Center (Case, RPI)
- Authentication (Schoffstall, RPI)
- Internet Host Requirements (Gross, Mitre/Braden, ISI)
- Short-Term Routing (Hedrick, Rutgers)

5:00 pm Recess

WEDNESDAY, March 2

8:30 am Opening Plenary
8:45 am Working Group meetings convene

- Domains (Mamakos, UMd)
- Performance and Congestion Control (Mankin/Blake, Mitre)
- EGP3 (Lepp, BBN)
- OSI Technical Issues (Rose, TWG)

1:00 pm Detailed Report on the New NSFnet (Braun, UMich/Rekhter, IBM)
3:15 pm Status of the Adopt-a-GW Program (Enger, Contel/Gross, Mitre)
3:45 pm BBN Report (Brescia/Lepp, BBN)

5:00 pm Recess

THURSDAY, March 3

8:30 am Opening Plenary

8:45 am Working Group Reports and Discussion

- Domain (Mamakos, UMd)
- EGP3 (Lepp, BBN)
- Open Systems Internet Operations Center (McCloghrie, TWG)
- Authentication (Schoffstall, RPI)
- Performance and Congestion Control (Blake, Mitre)
- OSI Technical Issues (Rose, TWG/Callon, BBN/Hagens, UWisc)
- NetMan (Rose, TWG)
- Short Term Routing (Hedrick, Rutgers)
- Open Routing (Callon, BBN)
- Open IGP (Petry, UMD/Moy, Proteon)
- Host Requirements (Braden, ISI)

1:00 pm Technical Presentations

- Routing IP Datagrams Through X.25 PDNs (Rokitansky, DFVLR)
- Internet Multicast (Deering, Stanford)
- TCP Performance Prototyping and Modelling (Jacobson, LBL)
- Cray TCP Performance (Borman, Cray Research)
- DCA Protocol Testing Laboratory (Messing, Unisys)

5:00 pm Adjourn

4.0 MEETING NOTES

4.1 Tuesday, March 1

4.1.1 Working Groups

The first one and a half days were devoted to meetings of the Working Groups. Reports from these meetings are reproduced in Section 5.

4.2 Wednesday, March 2

After a morning of Working Group meetings, Wednesday afternoon was devoted to presentations on Internet status. Two of these reports, on NSFnet and BBN activities, have become regular features of the IETF Plenary.

4.2.1 Report on the New NSFnet: Hans-Werner Braun (UMich), Jakob Rekhter (IBM)

The architecture and design of the new NSFNET backbone have been developed by MERIT, Inc., MCI, and IBM. Hans-Werner Braun gave an overview of the network and milestones. Jakob Rekhter's talk was on technical issues of the backbone nodes.

The structure of the NSFNET starts with a backbone of IP packet switches. Connected to this backbone are regional networks. The regionals then provide interconnection to campus-level networks. The new NSFNET backbone will provide a T1 speed service. Braun gave a functional overview of the backbone. Please see the MERIT proposal document, "Management and Operation of the NSFNET Backbone Network" and Braun's and Rekhter's presentation slides in Section 6.

The backbone was designed with upward growth in mind. There are "hooks" for T3, which Braun hopes will come in 1990, though it is not funded now. The backbone nodes have an open architecture, so that faster switches also can be brought on as they become feasible.

Network management is part of the backbone design. It is based on IBM Netview and PC/Netview as the management applications. Information from backbone nodes will be gathered for the applications by an agent using the interim Internet network management protocol, SNMP. Input is needed from the Internet community about what services the NSFNET Network Information Center should provide. It was asked who will be handling user end-to-end problems. Braun replied that he and Steve Wolff are interested in what the IETF InterNIC Working Group can come up for the problem of fault-isolation in a decentralized network. The NSFNET Network Service Center, located at BBN, which has acted as an ad hoc problem clearing-house, will not be going away.

The transition to the new backbone has the full cutover scheduled for July, 1988. A four-node research network with full T1 links was scheduled to begin service in April. In initial tests, dynamic bandwidth reconfiguration capabilities provided by MCI (including the ability to create multiple, unconnected subnets) are to be exercised.

It was asked if MERIT knew where to begin to tune the backbone, given so much flexibility. Braun answered that the reason for the research network was to develop tuning procedures.

Jakob Rekhter presented the architecture and some protocol engineering aspects of the backbone's packet switching nodes, the Nodal Switching Subsystems (NSS). Each NSS is made up of a number of processors connected by one or more IBM token ring LANs (two currently). IP packet switching and route processing are done by IBM PC RT's running a modified version of BSD UNIX 4.3. Each Packet Switching Processor (PSP) could have a T1 link from MCI's multiplexor. In response to audience questions, Rekhter said that the IBM proprietary interface card currently can only push data at a half T1 speed, but that IBM plans to improve this later. In answer to further questions, he stated that every token-ring interface in the NSS has its own IP address. However, passing through an NSS decrements the IP TTL on a datagram only once; the NSS is one hop.

The Intra-NSS communications are over TCP. A Routing Control Processor (RCP) communicates with the PSPs in master-to-slave mode, maintaining current routing tables in each PSP. If the RCP goes down, the PSPs revert to static routing information. Currently no redundancy is planned. A PSP in each node runs EGP.

An adaptation of the ANSI IS-IS protocol runs between nodes. Rekhter said it is close to IDEA0005. An discussion of NSFNET routing can be found in two other IETF working documents (issued after this meeting) IDEA0021, "EGP and Policy Based Routing in the New NSFNET Backbone" by Jakob Rekhter, and IDEA0022, "The NSFNET Routing Architecture" by Hans-Werner Braun. The Inter-NSS protocol is implemented over Level 2 on the trunks. It has some capability for load-splitting in that it can identify a set of equal-cost paths. Its metric is intended to reflect link speed and delay. The metric is static; that is, upon bandwidth reconfiguration using the MCI capabilities, an operator must manually change the metric. It was asked if it will be possible to monitor the overhead of the routing protocol. Rekhter said that it won't be, but that the worst case has been determined.

As far as the interaction between the nodes and the regionals, Rekhter said that "very simple" policy-based routing would be put in place, starting July 18. Its goals are to allow no bogus networks, and to protect campus networks from unwanted representations. The mechanism is the EGP metric. Each campus will select one or more regionals to represent them to the backbone. The regional which is selected as the campus's primary representative will advertise the campus with a metric of 0, the secondary representative will advertise a metric of 1, and so on. The choices will be done by the network administrators. The EGP implementation in the backbone will have a gated-like protection capability, checking that the campus is advertised with low metrics only by its chosen representatives.

It was asked if any one node was going to have two regionals coming in. Rekhter said this was possible and that a second EGP-speaking packet switch would be run in such a node.

As research-oriented issues, Rekhter discussed some congestion control plans for the NSSs. These plans are influenced by Dave Mills' experience with preemptive queue disciplines, and include giving routing protocol datagrams highest priority, issuing soft ICMP quenches, and dropping first the excess datagrams from hosts to whom the most quenches have been sent. Audience members urged Rekhter to reconsider using host preemption since some hosts may legitimately require more capacity than others, but Rekhter argued that the techniques will discriminate mainly against bad TCP implementations. Rekhter said further study would be done.

4.2.2 The Adopt-A-Gateway Program: Bob Enger (Contel), Phill Gross (Mitre)

Bob Enger (Contel) gave an overview of the history, motivation, and status of the "Adopt-A-Gateway" program. He presented convincing data showing both the poor performance prior to, and the improved performance after, the inception of the program. Phill Gross (MITRE) showed data from a different source that supported Enger's conclusions.

The "adoption" program began at the November IETF meeting in Boulder. During a presentation in which the continuing plight of the Internet was being discussed, Enger casually suggested that we might see an improvement if the Core gateways were upgraded from LSI-11/23 to LSI-11/73 processors. The audience sat in stunned silence over the naive implication in the suggestion. As we all knew, the length of a typical procurement cycle would stand in the way of this type of short-term solution. Undeterred by the facts, Enger suggested that many institutions must surely have surplus 11/73's sitting dusty in their spare parts bins. He pointed out that the LSI-11 architecture was no longer quite state-of-the-art. He suggested that we collect "loaner" boards from willing foster parents and then contact DCA about getting them installed.

Enger reported that between the November meeting and the March meeting, five of the six Core EGP servers and one of the Core Mailbridges had been upgraded in this way to 11/73's with a full complement of memory. The foster parents are:

- BBN
- Contel
- University of Illinios
- Thinking Macines, Inc.
- University of Maryland

Enger acknowledged Annette Bauman of DCA for her help in getting the equipment installed. (Note: following the March IETF, Phil Karn of Bellcore arranged for the loan of processors and memory to upgrade the remaining EGP server and remaining Mailbridges.)

Enger had made 'before' and 'after' Ping measurements. His data show that the EGP servers were simply overwhelmed by the well known extra-hop problem. He proved that the long delays were not in the subnet by making measurements to other hosts on the same PSN's as the EGP servers. While the EGP servers showed extraordinarily long delays, hosts on the same PSN often had much more resonable delays. After the upgrade to 11/73 (with more memory), these delays were reduced considerably. (See his presentation slides in Section 6 for his complete set of measurements.)

Gross also showed data that supported Enger conclusions. He had plotted various data from the weekly BBN Core Gateway Throughput Reports. (See the presentation slides in Section 6.) He showed that in the weeks prior to the Core gateway upgrades, the packet drop rate was rising at an alarming pace. This caused the overall traffic through the system to decline. In the weeks after the upgrade, the drop rate was significantly reduced and the overall traffic increased. He said his and Enger's data showed that the upgrade resulted in "more packets faster"—a double win.

4.2.3 BBN Status Report: Mike Brescia, Marianne (Gardner) Lepp (BBN)

The BBN report at this meeting featured a tour of the BBN gateway system, given by Mike Brescia, and then a status report on PSN 7, by Marianne Lepp.

Butterfly gateways are gradually replacing the LSI-11's. The LSI-11 core gateways, fortified by the processors and memory donated in the Adopt-A-GW Program, are reaching their upper limit of the table and update sizes. The last kludge in GGP, by Steve Atlas, will allow 500 networks to peer with the core. The number of networks peering with the core has been doubling annually, and there is nothing to indicate a slowing-down now.

The Butterfly Shortest Path First (SPF) routing protocol replaces GGP. The table limits of the core will be eased and the extra-hop problem will vanish; Marianne Lepp observed that the traffic on the EGP servers caused by the extra-hop is from 40-80%. With the new core gateway system, there is still a need for the EGP fixes that have been specified in EGP 3 (IDEA0009), but tasking for a Butterfly implementation and the transition to this new version is not in place.

Brescia presented a rough plan for the Butterfly core conversion, in which there would be parallel Butterfly and LSI-11 mailbridges and EGP servers until testing of the Butterfly EGP is complete. The start of this conversion has been delayed, and cannot be precisely scheduled for several reasons, the paperwork about PSN ports being the major one. Administrators of external gateways (those running EGP) should watch for an announcement of the new EGP servers and mailbridges in EGP-PEOPLE@BBN.COM. At that time, they should begin to peer with new servers, but continue to peer with the old ones as well. It was asked if the Autonomous System number of the new core would remain 0, as there are networking implementations that assume this. Those implementations should be fixed, because the AS number of the Butterfly core will be 60.

The new End-to-end protocol is the key item in PSN Release 7. Tailored to interact better with X.25 host interfaces, the new EE has more a efficient acknowledgment policy. Also important to its performance is the elimination of resource reservations. A higher level performance change is that it permits multiple PSN connections between host pairs.

In the new EE, messages that arrive when there are no resources for them are dropped by the destination, and the source retransmits. The blocking to await reservations that hosts and gateways saw in the old protocol is gone. Lepp presented new EE performance statistics, from a collection made from 12/5 to 2/14. A new collection method was used, making the statistics useful for evaluating the function of the new EE policies, but not for comparing the performance of the new and the old protocols.

BBN finds that 85% of traffic in the ARPANET is single-packet messages. In the old EE, almost all single packets obtained resources without delay, but 38% of multi-packet messages had to wait, blocking the host for all traffic until the resource was available. In the new EE, retransmissions (indicating any failure to obtain resources) are rare, fewer than 1 in 2500 messages. For those aware of the work on retransmit timers by Van Jacobson and others, Marianne noted that the new EE retransmit timers are not dynamic. They are configured during installation.

Other results from the statistics include an increase of about 20% in trunk utilization. This can be attributed to the new acknowledgment policies.

4.3 Thursday, March 3

Working Groups gave their status reports at Thursday morning's plenary session. The NetMan Working Group presented a status report based not on a meeting at this IETF, but on its activities in the weeks prior to the IETF. Presentation slides from these reports are contained in Section 6 of these Proceedings. Written reports from these meetings are in Section 5.

Thursday afternoon contained a very full lineup of technical presentations.

4.3.1 Routing IP Datagrams Through X.25 PDNs: Carl-Herb. Rokitansky (DFVLR)

Carl-Herbert Rokitansky, of the West German Aerospace Research Institute (DFVLR), discussed the routing problems of the European TCP-IP Internet. It was surprising to hear the extent to which TCP-IP is developing in Europe. Thirty-six vendors (including the Deutsche Bundespost!), demonstrated TCP-IP at the Munich Systems Multinet Show last October, and sixty were expected at the Hanover Computer Show in April. Someone in the audience speculated that the demand for networking capabilities has arisen from publicity for OSI, but since many OSI products are not yet available, the market has grown for TCP-IP products instead.

Rokitansky noted that there is no central administration of network numbers accompanying this growth. Internetting will come, though, so the routing of IP through the European national PDNs needs to be engineered now. In the U.S. Internet, the ARPANET/MILNET connects several hundreds of networks, but the situation is completely different in Europe: the only network which could be used as a backbone to allow interoperation between the many local area networks in Europe now subscribing to the DoD TCP/IP protocol suite would be the system of Public Data Networks (PDN). Yet no algorithms have been developed to dynamically route internet datagrams through X.25 public data networks.

The high cost of X.25 call setup means that hosts within Europe, connected by PDNs, need to see all the national PDNs together as one network. Hosts reaching the PDN-connected networks from outside Europe need to see multiple networks, in order to choose the right Value-Added Network (VAN) Gateway the first time. To let the national PDNs appear to hosts on them as one network, Rokitansky has defined the Cluster Mask. The national PDNs should all be assigned a Class B address with the same bits in the high order byte of the Internet address. Hosts within the cluster apply the mask 255.0.0.0 to this net address and send datagrams without using a gateway, while hosts do not apply the mask and compute routes to individual PDNs. It would be necessary to reserve a block of Class B addresses for the PDN cluster.

Other requirements would include:

- Cluster masking software for the intra-cluster hosts.
- An address resolution protocol for the intra-cluster hosts to use to map IP addresses to X.121 PDN addresses.
- Cluster software, modified IP source route, modified EGP for the VAN gateways.
- No modifications would be required in Internet hosts outside the cluster.

An IETF Working Group will be established to work on the Cluster Mask scheme and other aspects of Internetting with PDNs. Some of its broader interests include the ISO-migration of the cluster scheme, research into routing metrics, especially in tune with PDN costing issues, and support of other IETF routing work.

4.3.2 Internet Multicast: Steve Deering (Stanford)

Steve Deering from Stanford University gave a presentation on multicast addressing using IP. Interest in this capability stems from packet minimization needs and a more efficient use of bandwidth in a congested environment. The basic design of IP multicasting requires a new address class (D) for a destination host group whose members can reside throughout the Internet, and whose membership is unbounded and dynamic.

The upper layer protocol must specify the destination host group and a time-to-live value of at least 1 for internet routing. Upon receipt of this information, IP then engages local multicast distribution within the subnet to which the source host is directly attached or sends the packet to a multicast router at a well-known address for distribution to another network. Multicast routers relay the packet to the destination subnet where final distribution is made by the local multicaster router. Basic requirements for implementation for multicasting via IP are multicast ES-IS, multicast IGP, and multicast EGP.

Section 6 contains a complete set of slides for this presentation. RFC 1054, *Host Extensions for IP Multicasting*, is now available from the NIC, and an implementation is planned for preliminary release to researchers via 4.3 BSD.

4.3.3 TCP Performance Prototyping and Modelling: Van Jacobson, (LBL)

The first part of Van's talk described a "little hack" that he and Mike Karels developed that allows TCP to run at 8 Mbps. Since there were no slides for this part of the talk, we edited, and are including, an note from Van to the tcp-ip mailing list that describes the technique.

The paper by Butler Lampson mentioned in the note was published in Operating Systems Review, volume 17 number 5, October 1983.

The second part of the talk presented an analysis of the effects of random packet loss on the throughput and the equilibrium window size of slow-start TCP. A lossy net will reduce the throughput of slow-start TCP since the window is closed in response to dropped packets. Until the window opens to full size, the throughput of the connection will be reduced. It is also possible that packet loss could cause the equilibrium window size to be smaller than the maximum, again reducing throughput.

Van's analysis showed that packet loss had a minor effect on throughput and that the equilibrium window size was limited by buffer constraints and not packet loss rate.

Since there were slides for this part of the presentation, it does not suffer from our editing. Van's edited note follows.

Van Jacobson and Mike Karels at LBL have developed a TCP that gets 8Mbps between Sun 3/50s. The throughput ranged from 7Mbps to 9Mbps because the Ethernet exponential backoff makes throughput very sensitive to the competing traffic distribution when the connection is using 100% of the wire bandwidth. The throughput limit seemed to be the Lance chip on the Sun since the CPU was showing 10-15% idle time. This number is suspect and needs to be measured with a microprocessor analyzer but the interactive response on the machines was pretty good even while they were shoving 1MB/s at each other.

Most of the VMS Vaxen did crash while running throughput tests but this had nothing to do with Sun's violating protocols. The problem was that the DECNET designers failed to use common sense. A 1GB transfer (which finished in 18 minutes) caused the VMS 780 to reboot when it was about halfway finished. The crash dump showed that it had run out of non-paged pool because the DEUNA queue was full of packets. It seems that whoever did the protocols used a *linear* backoff on the retransmit timer. With 20 DECNET routers trying to babble the state of the universe every couple of minutes, and the Suns keeping the wire warm in the interim, any attempt to access the ether was going to put a host into serious exponential backoff. Under these circumstances, a linear transport timer just does not work. There were 25 retransmissions in the outbound queue for every active DECNET connection.

The other Sun workstations were not all that happy about waiting for the wire either. Every Sun screen in the building was filled with "server not responding" messages but none of them crashed. Later most of them were shut down to keep ND traffic off the wire while they searched the upper bound on xfer rate.

Two simultaneous 100MB transfers between 4 3/50s verified that they were gracious about sharing the wire. The total throughput was 7Mbps, split roughly 60/40. The tcpdump trace of the two conversations has some holes in it (tcpdump can not quite achieve a packet/millisecond, steady state) but the trace does not show anything weird happening.

Quite a bit of the speedup comes from an algorithm that they developed called "header prediction". The idea is that if you are in the middle of a bulk data transfer and have just seen a packet, you know what the next packet is going to look like: it will look just like the current packet with either the sequence number or acknowledgment number updated (depending on whether you are the sender or receiver). Combining this with the "Use hints" epigram from Butler Lampson's classic "Hints for Computer System Design" you start to think of the tcp state (rcv.nxt, snd.una, etc.) as hints about what the next packet should look like.

If you arrange those hints so they match the layout of a tcp packet header, it takes a single 14-byte compare to see if your prediction is correct (3 longword compares to pick up the send & acknowledgment sequence numbers, header length, flags and window, plus a short compare on the length). If the prediction is correct, there is a single test on the length to see if you are the sender or receiver, followed by the appropriate processing. For example, if the length is non-zero (you are the receiver), checksum and append the data to the socket buffer, then wake any process sleeping on the buffer. Update rcv.nxt by the length of this packet (this updates your "prediction" of the next packet). Check if you can handle another packet the same size as the current one. If not, set one of the unused flag bits in your header prediction to guarantee that the prediction will fail on the next packet and force you to go through full protocol processing. Otherwise, you are finished with this packet. So, the *total* tcp protocol processing, exclusive of checksumming, is about 6 compares and an add. The checksumming goes at whatever the memory bandwidth is so, as long as the effective memory bandwidth at least 4 times the ethernet bandwidth, checksumming is not a bottleneck. The 8Mbps transfer rates were attained with checksumming on.

This same idea can be applied to outgoing tcp packets and most everywhere else in the protocol stack. In other words, if you are going fast, this packet probably comes from the same place the last packet came from so 1-behind caches of pcb's and arp entries are a big win if you are right and a negligible loss if you are wrong.

As soon as the semester is over, they plan to clean up the code and pass it out to hardy souls for beta-testing.

The header prediction algorithm evolved during attempts to make a 2400-baud SLIP dial-up send 4 bytes per character rather than 44. After staring at packet streams for a while, it became obvious that the receiver could predict everything about the next packet on a TCP data stream except for the data bytes. Thus all the sender had to ship in the usual case was one bit that said "yes, your prediction is right" plus the data. There is a lesson here for high speed, next-generation networks. Research to make slow things go fast sometimes makes fast things go faster.

4.3.4 Cray TCP Performance: Dave Borman (Cray Research)

Dave Borman described a series of improvements to the TCP/IP implementation for UNICOS that increased the throughput over a HYPERCHANNEL link from the 1-2 Mbps range to over 100 Mbps. These improvements also reduced or eliminated panic, crashes, and hangs caused by the implementation. He also described the direction of future work that may raise the throughput to as much as 400 Mbps.

The original code (a port of a Wollongong port of 4.2 BSD) could only attain 1-2 Mbps between machines and 8 Mbps in software loop-back mode. The main problems were a character oriented checksum which was very slow on the word oriented Cray, a limited number of buffers (2) in the driver, data copies from/to mbuf chains, and no compaction of the TCP reassembly queues which caused rapid depletion of mbufs and lead to panics and crashes. In addition, the HY driver did not perform retries, requiring packets dropped by the HYPERCHANNEL to be retransmitted by TCP.

To correct these problems, several fixes were developed and installed. A word-oriented checksum routine with an optimized, assembly language inner loop was written. The driver code was rewritten to increase the number of buffers and add dynamic buffers and headers. The mbuf code was rewritten, the TCP reassembly code was fixed, and retries were added to the HY driver.

The effect of these changes was to increase the throughput between machines to over 60 Mbps with checksumming on and 85 Mbps with checksumming turned off. The software loop-back speed increased to 118 Mbps. The crashes and panics caused by running out of mbufs were also eliminated.

There is still substantial room for improvement. The rewritten checksum routine still takes almost 500 microseconds (which is a lot of time on a Cray) for a 32K packet. This will be reduced by vectorizing the checksum routine. There are also 296 microseconds (or 70,000 clock ticks on a Cray) unaccounted for in the transfer of a 24K block. Future versions of the code will attempt to identify this slack and remove it. Other enhancements such as TCP window scaling to allow large (Mbyte size) windows to be sent and Van Jacobson's header prediction algorithm should also increase performance, possibly raising throughput as high as 400 MBps.

4.3.5 The DCA Protocol Testing Laboratory: Judy Messing (Unisys)

Judy Messing from UNISYS gave a presentation on the DCA Protocol Certification Laboratory built by UNISYS. The laboratory was implemented under contract to DCA (DCEC in Reston, VA) to provide a facility for vendors and contractors to test their DoD Military Standard protocol implementations. The basic testing criteria for the lab are:

1) To test correctness of MLSTD services implemented.

2) To test correctness of optional services implemented.

3) To test correct handling of erroneous input.

Tests can be executed on a single function and can be executed in a repeatable manner. In addition, an audit trail of protocol exchanges is provided, and results of all tests are available.

The Test Facility consists of a reference host that is remotely accessible via DDN by the testing host. Both hosts must implement a control protocol by which the reference host initiates and conducts the protocol tests on the remote testing host. A log file of the test scenario and accompanying results (which are available to the tester) is maintained.

A complete set of slides for the presentation is included in this proceeding and inquiries about the lab are to be directed to Judy Messing (sdjudym@protolaba.arpa).

5.0 WORKING GROUP REPORTS

This section gives the reports of the March 1-3 Working Group meetings (some were previously distributed by electronic mail).

In three cases (MIB, NETMAN, and SNMP), the reports are from meetings that took place after the March 1-3 plenary.

Reports in this section from the March 1-3 plenary:

- Authentication (Reported by St. Johns, DCA)
- EGP3 (Reported by Petry, UMD)
- Internet Host Requirements (Reported by Braden, ISI)
- OSI Technical Issues (Reported by Rose, TWG/Callon, BBN/Hagens, UWisc)
- Open SPF-based IGP (Reported by Moy, Proteon)
- Open Systems Routing (Reported by Callon, BBN)
- Performance and Congestion Control (Reported by Mankin, MITRE)
- Short-term Routing (Reported by Hedrick, Rutgers)

Reports in this section from meetings after the March 1-3 plenary:

- Internet Management Information Base (MIB) (Reported by Partridge, BBN)
- CMIP-based Net Management (NETMAN) (Reported by LaBarre, MITRE)
- SNMP Extensions (Reported by Rose, TWG)

5.1 Authentication

(These notes of the Authentication group meetings at, and after, the March 1-3 IETF were submitted by Capt. Mike St. Johns, DCA.)

Immediately after the SDSC IETF meeting, the "THEM" subgroup of the Authentication working group met in Menlo Park at the NIC for an afternoon. Present were Jon Rochlis and Jeff Schiller of MIT, Steve Kent of BBN, and Mike St. Johns of DCA (DDN Program).

This was a follow-up meeting to the meeting held at BBN a few weeks previously, and was originally intended to gather all the people who had missed that meeting because of snow. What it ended up being was a re-evaluation of how to authenticate properly various network services.

After much discussion of various approaches, the group consensus gradually centered on divorcing authentication from access control and key management. The group felt the approach was reasonable because of work in progress on the ANSI side of the world.

The basic design for authentication would use the DES as the crypto method for wrapping data, either by checksumming it, or by encrypting the entire package of data. The two entities that want to be authenticated to each other would share a secret—in this case a DES key. The problem of how they each get a copy of the key would reside in a standard network protocol for access control and key distribution. For authentication, this would be a black box with well defined interfaces. The group believed we should concentrate on defining those interfaces, defining what portions of data need to be protected, and what is considered adequate protection for various classes of applications.

Most of the progress in the ANSI arena centers around certificate-based authentication and access control. This in turn depends on various public-key crypto methods.

5.2 EGP3

(Notes of the March 2 meeting at the San Diego IETF were prepared by Mike Petry, University of Maryland.)

The EGP3 group met on Wednesday March 2, 1988. The attendees were:

- Marianne (Gardner) Lepp (Chair)
- Mike Karels
- John Moy
- Mike Petry
- Jeff Schiller
- Michael St. Johns

The meeting consisted of a detailed review of the current Idea 9 draft. The bulk of the time was spent examining the state variables and pseudo code. Some parts of the document were reorganized and extended to provide addition clarification with respect to state variable usage and definition. The pseudo code was felt to be both correct and an important aid in understanding the new database structure of EGP3 vs. EGP2. The document will have the above changes made and be resubmitted as a revised IDEA.

5.3 Performance and Congestion Control

(These notes of the Performance and Congestion Control group the March 1-3 IETF were prepared by Allison Mankin, MITRE.)

The IETF Performance/Congestion working group met in San Diego for the morning of March 2. Those attending were: Art Berggreen (ACC), Coleman Blake (MITRE), David Borman (Cray Research), Robb Foster (BBN), Van Jacobson (LBL), Phil Karn (Bellcore), John Larson (Xerox PARC), John Lekashman (NASA/GE), Allison Mankin (MITRE), Keith McCloghrie (Wollongong), K.K. Ramakrishnan (DEC), Bruce Schofield (DCEC), Aditya Singh (Nynex S&T), Geof Stone (Network Systems Group), Zaw-Sing Su (SRI), Steve Waldbusser (CMU), Anne Whitaker (MITRE), and Lixia Zhang (MIT-LCS).

The working group's agenda is to produce a paper recommending quick fixes for Internet congestion problems. A quick fix is one which:

1) Improves performance.

2) Can be retrofitted into host or gateway protocol implementations.

3) Allows interoperation with "unfixed" implementations.

In the March 2 meeting, the outline of the paper was developed. Section volunteers were found or extorted. In addition, Van Jacobson led an extended discussion.

The outline for the paper follows, with indications of who is working on individual sections. As of June 10, we had a first draft of most of the sections. The group will meet in Annapolis with the roughly edited first draft of the paper in hand. After that, we plan work by E-mail and to have an offline meeting to produce the IDEA. The mailing list for work on the paper is:

ccpaper@gateway.mitre.org.

1. Introduction

   A. Improved performance in a computer network. (Ramakrishnan, Mankin)

   B. Background of this paper's recommendations. (Mankin)
   Trials and implementation experiences that have
   given confidence in the fixes to be recommended.

2. Recommended Short-term Fixes for TCP

   A. Getting the retransmit timer right. (Blake)

Timer implementation is extremely important and is easy to get wrong. The approach taken in the publicly available Berkeley TCP code will be documented: algorithms for obtaining an accurate mean and variance of round trip time, for calculating the round-trip timeout, and for backing off.

B. Small packet avoidance revisited. (Karn)
Implementing the Nagle algorithm so that it works even when the peer offers a huge window.

C. The XTCP/CUTE congestion control algorithms. (Schofield)
A specification of the algorithms due to Jain et al, Van Jacobson and Mike Karels, which have been implemented in the publicly available Berkeley TCP code. The goal is to facilitate independent implementations and procurement specifications of these fixes.

3. Recommended Short-term Fixes for Gateways

A. Random dropping. (Ramakrishnan)
When a gateway must drop packets, dropping the last in tends not to penalize the ill-behaved connections whose large windows are responsible for congestion. Random preemption is simple to implement, requires little overhead, allows a very timely control of congestion, and is probably as good at penalizing bad guys as fair preemption.

B. Managing gateway X.25 VCs. (Berggreen)
How to trade off between gateways' bursty use of large numbers of VCs and the possible destruction of data when reclaiming a VC.

4. Recommended Short-term Fixes for Higher Layers

A. SMTP message reduction. (Karn)
Useful and safe batching of protocol messages.

B. Line-at-a-time TELNET. (Borman)
Documentation of how to negotiate this within the current TELNET spec (how Borman's 4.3BSD TELNET does it), and with a proposed new TELNET option.

C. Domain improvements. (Larson)
Quick fixes that improve caching (e.g.), plus an assessment of the limits of what short-term fixes can do.

5. Further Study or Can't Recommend

A. Source quench
Both when to generate it and how to react to it remain controversial.

B. DEC congestion avoidance (This does not belong under Can't Recommend!)
DEC's feed-forward approach using a bit in the IP header is probably not be retrofittable to our current network.

C. Fair service
Gateway algorithms that try to enforce equal shares of bandwidth for all connections will hurt connections that legitimately need extra shares (e.g. those of mail-relay hosts). This area requires further study and policy consideration.

D. Selective retransmission
A proposal exists for implementing this with a TCP option, but further study is needed.

E. Rate-based congestion control
Methods of bandwidth discovery and control of rate-based protocols are at too early a stage to be recommended now.

Coordination of this paper with the document being written by the IETF Host Requirements Group has been undertaken by John Lekashman.

A few further notes on the outline: in general, we defined short term fixes as those which have high assurance of success. Gateway random dropping algorithms require more testing; the group decided to recommend them as an approach. We should probably also write about more stateful gateway algorithms.

5.4 Short-term Routing

(These notes of the Short-Term Routing group from the March 1-3 IETF were prepared by Charles Hedrick, Rutgers.)

Present were: Charles Hedrick, Guy Almes, Steve Deering, Noel Chiappa, Ross Veach, Joyce Reynolds, Jon Rochlis, Russ Hobby, Bob Braden, Don Morris, Sergio Heker, Scott Brim, and Hans-Werner Braun.

First, we reviewed the problems noted at the previous meeting, to see what has been accomplished:

- Problems with ACC DDN X.25 connections - Traffic from NSFnet to the Arpanet was going through a few gateways. Many of these gateways used VAXes with ACC's X.25 board. This board (or its device driver) has a limit to the number of X.25 virtual connections, and that limit was being exceeded. Apparently a fix is now known and in testing, but is not yet in the field. However the problem has largely been avoided by splitting the load among a larger number of Arpanet gateways, including Maryland and later Illinois and Rice. Some sites that could handle traffic are still waiting for IMP's to come up. JvNC has been waiting over a year.

- Wrong gateways advertising NSFnet networks into the Arpanet via EGP - A number of network managers want to be able to control which gateways advertise their networks. There was a suspicion that inappropriate gateways (i.e. those with slow-speed links) were advertising. Code has been put into the fuzzballs to allow control over this. Reports were mixed on what the results were. Apparently the code was tried and works, but there are indications that NSFnet performance as a whole suffers drastically when the controls are turned on. No details were available, and no one seemed to know the current state of this knob.

- RIP Routing Information Protocol) hop counts greater than 16 - This has largely been solved, by a combination of things. This includes metric reconstitution at AS boundaries and some interesting tricks. We have been moving slowly to an AS-style routing strategy. Backdoors are tending to be closed down, to prevent routing loops. I get the impression that routing changes are being done on an ad hoc basis by each regional, rather than in some overall planned way, but that progress is being made. One interesting discovery is that one can route a network with diameter 31 using RIP. The trick is to have a gateway in the middle of the network advertise itself as a default route. If a packet needs to get from one end of the network to the other, it starts out at a point where the destination is > 16 and so is not visible. The default route, however, is visible, and the packet starts going through the network in the direction of the default. By the time the packet gets half-way across the network, it comes to the gateways that can see the final destination, and begins to be routed correctly. In summary, reports suggest that some routing instabilities remain, but that this is no longer a serious problem (at least not in comparison with the new problems).

Now we come to the new problems. There are really only two new problems: serious performance problems with the existing NSFnet backbone and uncertainties in staging the transition to the new backbone.

- Performance problems with the existing backbone - Several regionals report that routes from the backbone are flapping in a major way. That is, whole groups of routes will vanish and come back. At some locations, NSFnet is said to be unusable. From detailed descriptions of the behavior, most of us concluded that the LSI-11's have simply run out of CPU. It is likely that we have reached the capacity of the 56Kb lines that form the backbone. But the Arpanet has been at capacity for years, and things just slow down. The current NSFnet status is reported as being more serious, in that routing breaks down. (Note that I am simply passing on reports from the regionals here. I have no way to gather data on this myself, and detailed, BBN-style reports have never been given for NSFnet.) The best guess is that this is simply a result of traffic increases. We heard of increases like a factor of 4 in some areas. This should not be a great shock. Within the last couple of months, many networks have come online, including BARRnet. When you double the number of networks, you probably increase the traffic by a factor of 4. Suppose we have two groups of networks, A and B. Previously only traffic from A to A could be handled. Now we can get traffic from A to A, A to B, B to A, and B to B. If we have reached the limits of the fuzzballs, the obvious solution is to use something more powerful. The problem is that we are about to replace the backbone completely, so it is not clear whether there is enough time left for this to make sense. However if there is, two different vendors are willing to lend us 68000-based gateways to use in place of the fuzzballs (either all of the fuzzballs or a subset of them that are carrying the heaviest load—the details are open for negotiation).

- Transition issues - The contract for the existing NSFnet backbone expires at the end of March 88. Apparently the contract for the new backbone does not require interim support of the existing configuration, or at least is not unambiguous in doing so. The official cutover date is July 1, but many people are inclined to think that full production is going to be a few months later than that. So in principle, we could be without a backbone for 4 to 8 months. Nobody really believes this is going to happen, but there are reportedly many vigorous negotiations occurring among various groups within NSF and its contractors. Even if a solution is reached, the uncertainties affect the network badly, because they prevent us from being able to choose an approach to the current performance problems. We don't know whether the network after April will use the existing 56K lines, new lines from MCI, or whether we will fall back on some kludge cobbled up out of back-door lines. So it is impossible to do any serious planning. We identified several feasible approaches for the interim:

  - Get somebody to pay to continue the existing configuration. At that point, we still have to deal with the current performance problems. If we know this is going to be the alternative, we should examine the vendor offers to loan us new gateways.

  - Use the existing gateways, but using the new lines. The MCI lines are multiplexed, so it would in principle be possible to arrange a 56K network equivalent to the existing one. This would still leave enough bandwidth to test the new equipment. The best estimate is that the equipment needed to do this would be in place by May 1, so it would still be necessary to continue funding the existing lines for at least an additional month. This still leaves the performance problems with the fuzzballs, though faster lines might reduce the demands on the gateways and buy us enough additional time to survive.

  - If all else fails, the regionals are going to have to find ways to rebuild the NSFnet connectivity using lines other than the backbone. We identified connections to all of the regionals, mostly back doors, USAN, etc. It is clear that if all else fails, attempts will be made to use these lines. However it is likely that the results will be somewhere between unpleasant and disastrous. These lines are already being used for traffic, so the existing backbone traffic would not fit on them. And current routing technology would not be able to handle them. The routing chaos from last time was solved largely by simplifying routes through use of the backbone. It is likely that people would resort to fixed routes, and might handle only high-priority customers. Of course priorities would likely vary from site to site, with the obvious result.

In my view, the most prudent approach is to do some experiments immediately. See if we can find some places where the MCI equipment is ready, and try running an inter-fuzzball connection over one such line. Try a slightly higher speed than 56K, and see if it helps the fuzzball's performance. Try replacing one fuzzball with a commercial router to see how much trouble we run into with incompatibility. the primary decisions, however, involve money and politics, and there is not much this group can do about that. I will make sure that the people involved in those decisions get a copy of this report and probably some additional, more focused, recommendations.

There was a brief discussion of the scenarios that regionals will see with the new backbone. The IBM routers will use EGP to the regionals. Most regionals will end up talking EGP to both the NSFnet backbone and the Arpanet. They will probably have to leak routes that they get from the NSFnet backbone into their internal IGP. Regional network managers should examine their network configurations to see how they would set this up. They should make sure that vendors are alerted to any new capabilities that may be needed. The IBM routers will ignore metric information they get from regionals. They will use EGP only for reachability. Each end network will register with the backbone, and will declare primary, secondary, and tertiary interfaces. (That is, Rutgers might tell the backbone that 128.6 will normally come to the backbone via JvNC, but if that is down, could come via NYsernet.) The backbone will replace the metric they hear from the regional with the metric from their database, and will ignore reachability from any regional that is not listed as one of the authorized interfaces for that network. The hope is that this will tend to make the system less vulnerable to routing loops and other unexpected behavior.

Another issue: RIP continues to hang around my neck like the fabled albatross. We convoked a brief meeting of the RIP subcommittee to answer a question posed by a NYSERnet member to Proteon. Present at the meeting were Hedrick, John Moy, and Mike Karels. The question was: Proteon routers support static routes. They pass these routes on to other gateways via RIP. they do not, however, send the static route out the interface to which the static route points, because of split horizon. A user complained that he wanted static routes to be advertised out all interfaces. The subcommittee concluded:

1) Static routes are really a form of lying. While there are often good reasons to lie in complex networks, the RIP specifications were not intended to specify the details of the features that vendors may choose to support for such purposes.

2) There were probably better ways to solve this user's problems than what he requested.

3) In any case advertising static routes out the interface they pointed to was likely to result in routing loops, and so Proteon was wise in enforcing split horizon.

4) We saw no objection to Proteon providing an option to disable split horizon in such cases, should they wish to do so. However we strongly suggest that any such option should default to off, and that appropriate warnings should be placed in the documentation.

5.5 Open Routing

(These notes of the Open Routing group from the March 1-3 IETF were prepared by Ross Callon, BBN.)

The Open Routing Working Group met on Monday February 29th, the day before the full IETF meeting started. We also met for a half day on Tuesday March 1. Ross Callon acted as chair in the absence of Bob Hinden, who was unable to attend.

The first day was a general discussion about how we might do inter-autonomous system routing. Marianne (Gardner) Lepp started with a strawman architecture protocol approach. This was discussed and modified in real time. Two possible approaches emerged, which are not necessarily mutually exclusive. We also had a discussion of addressing issues.

Pat Clark handed out a brief description of DGP on Monday. We then had a "for information only" discussion of DGP on Tuesday morning. This was very useful in giving a better understanding of what DGP does and how it operates. We did not attempt to evaluate the applicability or feasibility of the protocol at this time. Separating the task of group discussion towards improved understanding of the protocol from evalation of the protocol was felt useful in maximizing the effectiveness of the meeting.

Tuesday afternoon we had an open meeting to allow IETF as a whole to comment on IDEA007, "Requirements for Inter-Autonomous Systems Routing" There were no major changes required, but a number of minor improvements and clarifications were discussed. These comments will be combined with others received (particularly from ANSI X3S3.3) to guide future revision of IDEA007.

5.6 Open SPF IGP

(These notes of the Open SPF IGP (OIGP) group from the March 1-3 IETF were prepared by John Moy, Proteon)

The IETF OIGP working group met in San Diego on March 2. The morning session was an open meeting to solicit comments on IDEA 005. The room was crowded, with about 40 people. The afternoon session was a working meeting to discuss details in the design of the OIGP. The afternoon session was attended by: Milo Medin, Mike Karels, Paul Tsuchiya, Phil Almquist, Louis Mamakos, K. K. Ramankrishnan, Mike Petry and John Moy.

1. The morning session

The first comment was that the organization of IDEA 005 is poor. General design guidelines are mixed in with the requirements. It was also noted that the requirements seemed to be written with the specific solution already in mind. This is a valid comment. To rectify this, IDEA 005 will be split into 2 documents: a requirements document and the protocol design document (specification).

A related comment was that there are other routing technologies (other than SPF) that can also solve the problems that the OIGP is trying to solve. The technologies mentioned specifically were Ford-based algorithms and Landmark routing. The chair (Mike Petry) pointed out that the OIGP group was formed with the idea of developing an SPF based protocol, and that there is room in the Internet architecture for several IGPs. It is assumed that there will not be a single standard IGP for the Internet. The suggestion was made to change the name of the group to OSPFIGP (for Open SPF-based IGP).

A number of people then asked "why not just implement DEC's IS-IS proposal?" The response of the chair was that we saw a number of problems with the DEC proposal that we attempted to enumerate in IDEA 005, and that also we thought that the differences between the IP and ISO architecture would force the two protocols to be distinct. For example, IP subnetting will be fully integrated into the OIGP. It is however assumed that there will be a large common base of ideas between the DEC IS-IS and the OIGP. John Moy promised to write a separate document detailing the problems we see in the DEC IS-IS.

There was some confusion on how the OIGP would operate in the presence of external routing information. This part of IDEA 005 needs to be rewritten including the following requirements:

- Link state information will be advertised separately from externally derived routing information. This externally derived information may be advertised by any border gateway. One should think of this external information as being configured in the border gateways. The metrics describing the external routes are not comparable to the link state metric.

- When a router then calculates its routing table, it does the SPF calculation on its link state database. This will calculate the shortest (internal) distance to each of the networks, subnets, and gateways present in the AS. Then, for those networks still not reachable, the external routing information is examined. For these networks, the gateway is found that advertises the shortest external route, and the route to that gateway is installed as the path to the network. When multiple gateways advertise the same shortest route, the gateway is chosen that is closest via link state information.

- The reason for this method is that we do not want to be forced into comparing external and internal metrics. It is also assumed that it will usually be desirable to route within the AS as much as possible.

- At this meeting we added a new external metric type, that would work like the internal metric. External routes using this new metric type will be considered first after the link state information is processed. In this case the border gateway will be chosen whose combined internal and external distance is shortest.

Many people were unhappy with the dimensionless link state metric. This is an area that needs more thought. The possibility was mentioned that we could get some help from the Open Routing group in this area.

Finally, some people were concerned that the OIGP is not trying to support the complicated topologies that we are seeing in NSF land. The OIGP is staying with the model where all gateways in an AS speak the same IGP. Some of the hard problems are being left to EGPs replacement (the protocol connecting the AS's) to solve.

Other comments included:

- The proposed link state graph takes only metrics on the outbound of interfaces into account. Maybe the input side should also have a metric associated to it (Scott Brim).

- Low-speed serial lines (down to 9600 baud) are not going away in the near future and should be supported (Chuck Hedrick).

- Nagel wrote a paper on a better way to distribute routing information than flooding. We should look at it (Ron Natalie).

2. Afternoon session

The afternoon began with the creation of a mission statement. We ended with the following:

- Our goal is the design and development of a multi-vendor SPF IGP. We plan to take ideas from the existing SPF technology, such as the BBN work and the DEC IS-IS proposal.

- A short list of requirements for the IGP includes: stability of the protocol in a large, heterogeneous system, TOS support, authentication of participants, and a precise specification of how the protocol will react with parts of the IP architecture such as subnetted networks and the presence of externally derived routing information. We realize that the requirements can probably be met by routing technology other than SPF.

- We now have an IDEA that discusses requirements and general design issues. We hope to have a preliminary protocol specification by the next meeting, with trial implementations in the summer.

We then discussed alternatives to the designated router of the DEC IS-IS scheme. The designated router performs two functions: it allows dead gateways to be detected quickly, and it ensures that the gateways connected in the link state graph can actually talk to each other. The obvious alternative is for a gateway to advertise its list of neighbors in the link state packets along with its interface state. This was rejected because of the increased size of link state packets and SPF database, along with the increased SPF processing time, that this would involve.

We could not think of any alternative to the designated router. We did list some good reasons not to have one:

- It would be nice not to have to perform the election algorithm needed to select the designated router for each LAN.
- Proper operation of the designated router is required for any gateway on that LAN to use the LAN for thru traffic, regardless of whether or not the designated router itself was the next hop.

The following things were also discussed briefly:

- Requirements for authentication. More work needs to be done here.
- Physical multicast should be used on networks that support it, instead of broadcast.
- When supporting unnumbered serial lines, the possibility exists for a gateway having no IP addresses assigned to its interfaces. Such a gateway will need to be assigned an OIGP identifier in order to participate in the protocol.
- Host routes should be fully supported by the OIGP. They should not be condensed into network-level routes at subnet boundaries.

3. Goals for next meeting

The goal is to produce three documents by the next meeting: a revision of IDEA 005 that contains only requirements, a document detailing the questions we have concerning the DEC ANSI proposal, and the OIGP protocol specification.

5.7 Host Requirements

(These notes, and update, of the Host Requirements group from the March 1-3 IETF were prepared by Bob Braden, ISI)

This working group is tasked with writing an RFC documenting the requirements for an Internet host, paralleling RFC-1009 on gateway requirements.

1. The writing assignments handed out at the San Diego IETF meeting have mostly been carried out, and the results have been assembled into an RFC draft by the editor. Major text contributions came from Noel Chiappa, Craig Partridge, Paul Mockapetris, John Lekashman, and James Van Bokkelen. A number of other committee members have contributed substantial editorial input, especially Steve Deering, Phil Karn, Keith McCloghrie, and Mark Lottor.

2. As editor, Bob Braden has been devoting a significant amount of time to smashing the contributed text together into a consistent format and organization, and tightening up the wording when necessary.

3. The group held a one day meeting to discuss the draft, using the ISI/BBN packet-video teleconference setup. We are immensely grateful to Steve Casner at ISI and his peers at BBN for the work they put into this. A total of 13 people participated at the two ends. John Lekashman served as meeting secretary.

4. The group intends to meet at the Annapolis IETF meeting. After that meeting, we hope that the results will be in good enough shape to receive public exposure as an IDEA.

The draft document has grown to 80+ pages in length. It is generally organized in accordance with the layers of the Internet protocol stack. Specifically, the current outline is as follows:

1. Introduction

2. Link Layer (this is small, mostly points to RFC-1009)

3. IP Layer (IP and ICMP)

4. Transport Layer (TCP and UDP)

5. Application Layer (SMTP, FTP, TFTP, and Telnet)

6. Support Programs (Network Management, Booting)

7. Appendix: Checklists

5.8 ISO Technical Issues

The ISO Working Group met for the first time at the March 1-3 IETF. The Chair is Marshall Rose (TWG). These notes were compiled by Phill Gross (MITRE) from submissions by Rob Hagens (UWisc), Ross Callon (BBN), and Marshall Rose.

A focus of discussion for this meeting was the DoD/OSI addressing structure proposed by Ross Callon in IDEA 003. This is important for at least two reasons: the DoD OSI planning will very likely use the addressing format specified by this group, and the University of Wisconsin, which is planning to do some collaborative experiments in sending OSI CLNP datagrams through the DoD/NSF Internet, would also use this addressing format.

During the Working Group reports on the final day of the IETF, there were two presentations that covered most of what was discussed in the ISO group. These presentations were:

- Addressing for the ISO IP in the DoD Internet (Ross Callon, BBN).
- The Use of the DARPA/NSF Internet as a Subnetwork for Experimentation with the OSI Network Layer (Hagens, UWisc.).

In addition, Marshall Rose presented a summary of current efforts within the IETF CMIP-based Network Management (NETMAN) group. He also gave an overview of his proposal in IDEA 017 for "ISO Presentation Services on Top of TCP/IP-based Internets".

The following notes are based on Ross Callon's summary of the discussions at the recent ANSI meeting, as well as the IETF meeting.

There has been enough varied discussion of addressing that the basic ideas on which each of the previous proposals was designed will be summarized below. The specific proposal that Ross is advocating is near the end of these notes.

The basis for RFC 986 was:

- Use the ICD value assigned to DoD Internet.
- Encode user protocol field.
- Encode current DoD Addresses to make use of current routing and address assignment.
- Allow for a version field, since we know the RFC's addressing-scheme is not sufficient for the long term.
- This results in a three part field:
  - AFI/ICD/version (4 octets, fixed)
  - DoD IP address (4 octets)
  - User Protocol (1 octet)
- All parts of address are in fixed location.

This approach suffers from two serious problems: (1) It is incompatible with the desire of the EON to experiment with the ANSI routing proposal now; (2) It is very much temporary, and will clearly become inadequate sometime in approximately the next 5 years or less. When it is time to change it, there will be a large installed base which will make it very expensive to fix.

The basis for IDEA 003 was:

- Choose an address scheme which can work for a longer time.
- Use the ICD value assigned to DoD Internet.
- Encode user protocol field.
- Encode current DoD Addresses to make use of current routing and address assignment.
- Routing by network number will become infeasible as Internet grows.
- AS number is convenient "higher level" address which has already been assigned.
- The number of ASs is growing rapidly, so we will probably also need a "higher-level" area.
- These requirements result in a five part field:
  - AFI/ICD/version (4 octets, fixed)
  - global area (2 octets)
  - AS # (2 octets)
  - DoD IP address (4 octets)
  - Use Protocol (1 octet)
- All parts of address are in fixed location.

The basis for Ross' presentation was:

- Address scheme needs to work long term, etc...
- Selector field does not have to be identical to DoD IP user protocol field, but is functionally similar.
- Some autonomous systems may want to use different address format internally. For example EON wants to use DEC/ANSI scheme, and other IGPs may use current DoD IP addresses.
- Therefore use AS specific address for local routing.
- These requirements result in a five part field:
  - AFI/ICD/version (4 octets, fixed)

- global area (2 octets)
- AS # (2 octets)
- IGP specific (variable)
- selector (1 octet)

- All "Inter-AS" parts of address are in fixed location.
- "Intra-AS" parts of address are NOT fixed, depend on AS (only gateways familiar with a particular AS know how its part of address is parsed).

Issues Raised at IETF:

- It would be useful if DoD part of address is always in the same place (This seems at first to conflict with proposal to have an "IGP specific" part of the address).
- It would be useful if some of the lower-level fields (AS # or DoD Address) are globally unique.
- Why should the next higher level thing from "network number" in address be exactly equal to current AS numbers? We are likely to want to have a single "routing domain" which consists of what is currently several AS's.
- It would be computationally more efficient if we always padded addresses to 20 octets. This would not increase address lengths by much in any case.

NOTE: The first 4 octets (AFI, ICD, and version) may be used to determine that the rest of the address is according to our format. The fact that we will in the future need to interact with Systems using other formats (such as addresses assigned via ANSI or ECMA) implies that this test will eventually be needed in any case. The next 4 octets (or the entire first 8 octets, if the first four octets contain a valid value) could be treated as a flat field identifying the routing domain or autonomous system. Thus the only thing that that cannot already be treated as a flat field in any case is the DoD address. We will consider schemes which will allow people to find the DoD address and treat it as flat.

Two other possible address schemes:

(These other possible schemes will use the term "routing domain" instead of "AS number" in the address. This implies that we will not require that the domains into which the Internet is divided will be precisely the same as the AS's currently assigned).

1) Change "AS #" to "Routing Domain" pad to 20 octets, otherwise leave the same.

   - This padding now makes it a six-part field with a total of 20 octets (variable parts must add to 11 octets):

- AFI/ICD/version (4 octets, fixed)
- global area (2 octets)
- routing domain (2 octets)
- padding (variable)
- IGP specific (variable)
- selector (1 octet)

With this scheme, gateways which route ISO IP packets are required to look at the Routing Domain number (possibly by treating the first 8 octets as a flat number), and only route according to the IGP part of the address if they are familiar with the routing domain (i.e., the routing domain is either those gateways or another set of gateways which they are familiar with by some a priori agreement).

2) Temporarily limit the allowed IGP specific address parts, all of which must include the DoD address just before the selector. Pad so that DoD part of address is always in the same place. This is the same as the previous option, except for a temporary guarantee of where the DoD address can be found. When this guarantee is phased out, then it will probably be necessary to change the version number.

This would embed the DoD IP 4-octet address in the 6-octet identifier in the addresses from the ANSI routing scheme. The guarantee that the DoD IP Address is embedded in this manner would be temporary only, and would be phased out when a new inter-AS routing scheme is in place.

This results in the same addresses as above, except that the IGP-specific part can be further subdivided into zero or more octets which are truly IGP specific, plus 4 octets of DoD IP address.

Ross proposes that we should adopt this approach. The version number should probably be set initially to 2, on the basis that some implementations may exist that implement RFC 986 (with version = 1), but no implementations should exist yet that implement any other scheme (for example, IDEA003 should not be implemented already).

Other Possible Ideas:

It has been suggested that we encode the length of the part of the address which is needed to determine the domain in the version number. This would allow current implementations which only understand early versions of the address to still be able to route to the destination domain, if they know that the fifth through eighth octets may be treated as domain number. There are several ways which this can be accomplished:

(1) We could specify that address versions up through some number (say, version 15) will always use the fifth through eighth octets to specify the domain.

(2) We could use some number of bits (4 to 6) for the version, and some number (2 to 4) for the length of domain field.

In any case, gateways in a domain can only route to addresses which they have been informed of in some way. Thus, when a gateway sends a message to the effect of "I have a route to addresses beginning with this prefix" the prefix probably includes the version number, and the length of the prefix is just the length of field needed to specify the domain or other entity which the route can reach. An approach similar to this will be necessary in any case when the Internet is connected to other internets (such as private, or European internets) which use different address structures (not assigned from the DoD Internet address space). This implies that a priori knowledge that a particular address version has a known location in which the domain can be found is of only limited usefulness in the long term.

Following Ross's presentation at the IETF, Rob Hagens presented an overview of the Experimental OSI-based Network (EON), which proposes to use the DARPA/NSF Internet as a subnetwork for experimentation with the OSI network layer. What follows is a brief overview of an RFC proposed by Robert Hagens and Nancy Hall (from the Computer Sciences Department at the University of Wisconsin - Madison) and Marshall Rose (from The Wollongong Group).

Since the International Organization for Standardization (ISO) Open Systems Interconnection (OSI) network layer protocols are in their infancy, both interest in their development and concern for their potential impact on internetworking are widespread. This interest has grown substantially with the introduction of the US Government OSI Profile (GOSIP), which describes the configuration of any OSI product procured by the US Government in the future. The OSI network layer protocols have not yet received significant experimentation and testing. The status of the protocols in the OSI network layer varies from ISO International Standard to "contribution" (not yet a Draft Proposal). It is critical that thorough testing of the protocols and implementations of the protocols should take place concurrently with the progression of the protocols to ISO standards. For this reason, the creation of an environment for experimentation with these protocols is timely.

Thorough testing of network and transport-layer protocols for internetworking requires a large, varied, and complex environment. While an implementor of the OSI protocols may, of course, test an implementation locally, few implementors have the resources to create a large enough dynamic topology in which to test the protocols and implementations well.

One way to create such an environment is to implement the OSI network-layer protocols in the existing routers in an existing internetwork. This solution is likely to be disruptive due to the immature state of the OSI network-layer protocols and implementations, coupled with the fact that a large set of routers would have to implement the OSI network layer in order to do realistic testing.

The proposed RFC suggests a scenario that will make it easy for implementors to test with other implementors, exploiting the existing connectivity of the DARPA/NSF Internet without disturbing existing gateways.

The method suggested is to treat the DARPA/NSF Internet as a subnetwork, hereinafter called the "IP subnet." This is done by encapsulating OSI connectionless network-layer protocol (ISO 8473) packets in IP packets, where IP refers to the DARPA/NSF Internet network-layer protocol, RFC 791. This encapsulation occurs only with packets travelling over the IP subnet to sites not reachable over a local area network. The intent is for implementations to use OSI network-layer protocols directly over links locally, and to use the IP subnet as a link only when necessary to reach a site that is separated from the source by an IP gateway. While it is true that almost any system at a participating site may be reachable with IP, it is expected that experimenters will configure their systems so that a subset of their systems will consider themselves to be directly connected to the IP subnet for the purpose of testing the OSI network layer protocols or their implementations. The proposed scheme permits systems to change their topological relationship to the IP subnet at any time, also to change their behavior as an end system (ES), intermediate system (IS), or both at any time. This flexibility is necessary to test the dynamic adaptive properties of the routing exchange protocols.

A variant of this scheme is proposed for implementors who do not have direct access to the IP layer in their systems. This variation uses the User Datagram Protocol over IP (UDP/IP) as the subnetwork.

The experiment based on the IP subnet is called EON, an acronym for "Experimental OSI-based Network" The experiment based on the UDP/IP subnet is called EON-UDP.

5.9 Internet Management Information Base (MIB)

(These notes of the meeting of 5/9-5/10/88 at Advanced Computing Environments were prepared by Craig Partridge, BBN)

Attendees:

- Greg Satz - Cisco Systems
- Karl Auerbach - Epilogue Technology
- Jim Robertson - 3COM/Bridge
- Phill Gross - MITRE
- Marshall T. Rose - The Wollongong Group

- Lawrence Besaw - Hewlett-Packard
- Mark Fedor - Nysernet
- Jeff Case - Univ. Tennessee
- James Davin - Proteon
- Unni Warrier - Unisys
- Robb Foster - BBN Communications Corporation
- Lou Steinberg - IBM
- Keith McCloghrie - The Wollongong Group
- Lee LaBarre - MITRE
- Bent Torp Jensen - Convergent Technologies
- Craig Partridge - BBN (Chairman)

As with the last set of minutes, instead of discussing all the issues in detail, I have chosen to mention the major issues that came up and their resolution. I have also listed action items.

The entire meeting was devoted to review of the proposed SMI and MIB documents developed by Marshall Rose and Keith McCloghrie of the Wollongong Group. The SMI document was in its second reading, having been completely reviewed at the first meeting in Boston. The MIB document was going through its first complete reading although some portions had been discussed in Boston.

The first morning was spent reviewing the first half of the MIB document. Our first action was to revise the list of criteria for inclusion in the MIB developed at the Boston meeting. The criteria we finally settled on was:

(1) Any object in the MIB should be useful for either fault or configuration management.

(2) Only weak control variables were permitted, because we felt that the current generation of management protocols did not have strong enough authentication mechanisms.

(3) We require evidence that these variables had been used in some networking system already (i.e. evidence of utility was required).

(4) The initial MIB could not contain more than approximately 100 objects. This goal was established to make sure that implementation of the instrumentation required by initial MIB was not onerous on vendors.

(5) Variables whose value could be derived from others would not be included.

(6) Implementation specific (e.g. BSD UNIX) values would not be included.

A seventh criteria was developed later in the review process:

(7) Keep counting to a minimum in main-line code. In other words, we did not want to be responsible for notably slowing down implementations by requiring massive instrumentation in heavily used code.

The review of the MIB document, although slow, went quite well. In general, the group was able to reach consensus on most objects to include or exclude from the MIB. In only a few cases was the chairman forced to take a vote. One important contribution to making the process go faster was Jeff Case's insistence that we draw flow diagrams of the various layers on a whiteboard and label where the flows were counted. These diagrams, promptly dubbed "Case diagrams" proved invaluable for determining where the important flows were and how best to count them. Entire pages of definitions were resolved with a few minutes of sketching on the board. One important change in the MIB document that had effects on the SMI was that we decided not to keep track of the time of day, but to keep timestamps only in terms of 100ths of a second since the system was last rebooted.

The afternoon of the first day was taken up reviewing the SMI document from the last meeting. This was expected to be a short run-through but proved to take the entire afternoon. Chuck Davin presented a scheme to simplify object naming in the SMI, and after substantial debate, it was adopted. Some changes were made in the SMI to reflect the MIB use of timestamps. Lee LaBarre withdrew his proposal from the last meeting to include thresholds in the initial MIB and so they were left out of the SMI. Furthermore, members of the group were concerned that we needed to define how the MIB and SMI were to expand and grow in a backward compatible way -- so the SMI was changed to include a section defining how the ways they should (and should not) be changed.

For the morning of the second day we returned to the MIB document and actually finished the review. Again, Case diagrams proved key to finishing it up. Keith McCloghrie plans to revise the draft and circulate it to the group late next week for review. Unless there prove to be major disagreements we propose to report this document to the IETF late this month.

In the afternoon, we sat down with the SMI document we had revised the previous day (thanks to fast work by Chuck Davin and Marshall Rose) and approved it for release to the IETF as an IDEA.

We also developed a schedule for making the documents into RFCs:

- The working documents will be released in the next couple of weeks as IETF IDEAs. Members of the IETF will be given until the last day of the IETF meeting in June to report comments to Craig Partridge (craig@nnsc.nsf.net).

- After the IETF, the Working Group will review the comments received and make appropriate changes (if any). The revised IDEAs will then be sent to IAB and Jon Postel as the official reports of the IETF MIB WG by the end of June, with the request that they be made into RFCs as soon as possible. (Phill Gross reports that the IAB is in the midst of a debate about how to make documents into Internet standards. If this looks like it will hinder release of our documents, we will ask they be released simply labelled as RFCs, otherwise as standards).

Finally, the chairman was given the task of writing up short report listing the recommendations of the MIB Working Group to the IAB. Beyond recommending that the SMI and MIB documents be made RFCs, this report will recommend that the IAB:

- Create a long-term organization to:
  - review proposed management documents
  - control the issuance of MIB version numbers
  - direct future research
  - advise on management protocol transition issues (e.g. SNMP -> CMIP)

- Require that no protocol be approved as an Internet standard without accompanying recommendations about how the protocol be instrumented for network management.

No further meetings of the MIB WG are planned unless there is controversy over the revised MIB document or a need to review IETF comments on the MIB and SMI documents.

5.10 IETF CMIP-Based Net Management (NETMAN)

(These notes of the meeting of 5/11/88 at Advanced Computing Environments were prepared by Lee LaBarre, MITRE)

The IETF NETMAN Working Group met the afternoon of May 11 at Advanced Computing Environments in Mountain View, CA. This meeting was held subsequent to a two day meeting of the IETF MIB Working Group om May 9-10, and a meeting of the NETMAN Demo subgroup meeting on the morning of May 11.

Since the Demo subgroup participants were the same set of people that attended the NETMAN WG meeting, the discussions often switched context between the long term NETMAN requirements and the detailed requirements for the Fall demonstration. Described below are the salient aspects of both meetings that relate to NETMAN as a whole.

The MIB-WG meeting results were discussed and the intent to use the structure and identification of the management information (SMI), and the near-term management information base (MIB) defined by that group was reaffirmed. Lee LaBarre was tasked to send a liaison statement to the MIB-WG informing them of this intent.

Structures not in the SMI and parameters not in the near-term MIB will be defined by NETMAN. For example, thresholds and event structures and additional TCP and data link (802.3) parameters. After some experience is gained in their use and their value ascertained, they will be proposed as extensions to the SMI and near-term MIB.

The structure of the CMIP MgmtInfoId field and its relation to the CMIP ObjectClass and ObjectInstance fields was discussed at length. A complex structure of the MgmtInfoId field was proposed to satisfy the requirement that it be possible to operate on attributes in different objects within a single CMIP PDU. The two options discussed were a doublet and triplet form as described in the ANSI X3T5.4 contribution attached to these minutes. It was decided that the triplet form was preferred because of assumed savings in encoding. The decision of which form to use for the fall demo was left to Unisys.

Lee LaBarre of MITRE and Amatzia Ben-Artzi of 3-Com/Bridge were tasked to take the NETMAN requirements and proposed structure of the MgmtInfoId to the ANSI X3T5.4 meeting of the following week May 16-20. It turns out that the triplet encoding is also preferred because of ISO compatibility considerations. This will be discussed in a separate report on the ANSI X3T5.4 meeting.

The need was identified to have a separate SMI document to replace IDEA013 which incorporates the MIB-WG SMI results, NETMAN extensions, and CMIP protocol specific aspects. This document would be referenced in implementors agreements. Lee LaBarre agreed to begin the effort.

The next NETMAN meeting is scheduled to coincide with the September IETF meeting. At that time it is expected that sufficient experience will have been gained through the demo effort, and sufficient stability will be in the CMIP protocol to make stable implementors agreements on the ISO based Internet management effort (Is ISOIME, or IMEISO, a good acronym for the effort?).

The NETMAN Demo subgroup will meet throughout the summer.

As a follow up on the assigned work items:

1. A distribution list has been established for participants of the fall demo, called nmdemo88@gateway.mitre.org.

2. The MIB-WG liaison statement has been sent out.

3. The NETMAN requirement for operations on attributes in different objects, and the MgmtInfoId proposal were taken to ANSI X3T5.4. The results will be distributed soon in a separate message.

4. The NETMAN SMI document is in progress.

5.11 SNMP Extensions

(These notes of the meeting of 5/12/88 at Advanced Computing Environments were prepared by Marshall Rose, TWG)

The SNMP Extensions Working Group was formed as a response to RFC1052. The Chair is Marshall Rose (TWG). The first meeting of the WG was held May 12, 1988 at ACE in Mountain View, CA. Based on the progress of the group, the second day of the meeting was cancelled.

A new baseline document was introduced along with the draft Internet-standard SMI and parts of the MIB. The document was then reviewed in detail by the committee over the entire course of the day. Consensus was reached on a number of issues. The action items resulting from this meeting are:

- A small subset of the working group will incorporate the group's comments on the document into the baseline;

- This baseline will be sent to the snmp-wg and eventually to be installed as an IDEA [Note: this has been done as IDEA0011-01, i.e., the first revision of the previously released SNMP document.]

- Members of the working group with SNMP technology currently running will attempt implementation of the resulting document (only a subset of the MIB will be supported); and,

- At the next IETF, the group will meet again. The comment period on the document will close. Assuming no implementational difficulties remain, the document will be submitted as an RFC.

6.0 PRESENTATION SLIDES

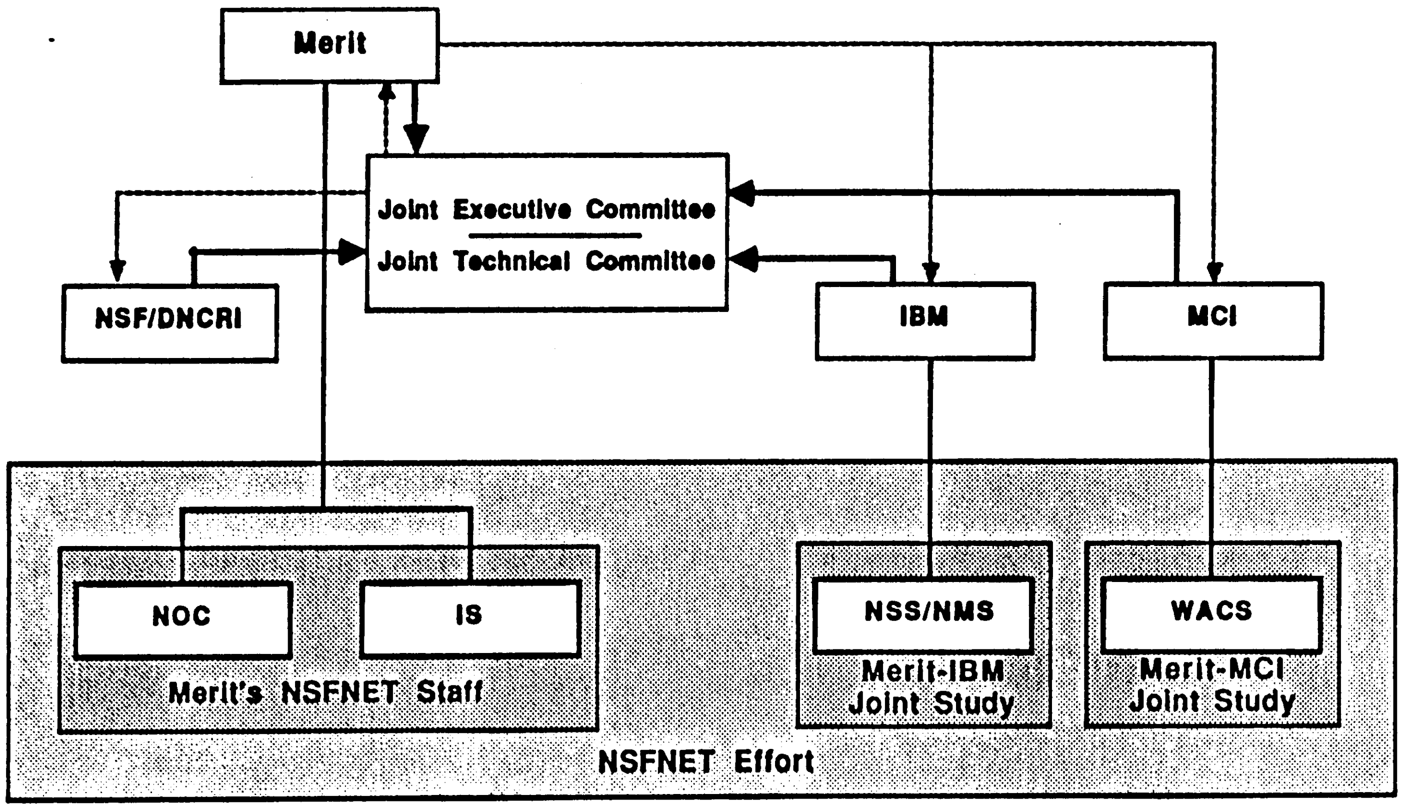
This section contains the slides for the following presentations made at the March 1-3, 1988 IETF meeting:

- Report on the New NSFnet (Braun, UMich/Rekhter, IBM)
- Status of the Adopt-A-GW Program (Enger, Contel/Gross, MITRE)
- BBN Report (Brescia/Lepp, BBN)
- Domain Working Group (Lottor, SRI-NIC)
- EGP3 Working Group (Lepp, BBN)
- Open Systems Internet Operations Center WG (Case, UTK)
- Authentication WG (Schoffstall, RPI)
- Congestion Control WG (Blake/Mankin, MITRE)
- OSI Technical Issues WG (Callon, BBN/Hagens, UWisc/Rose, TWG)
- Open Routing WG (Hinden/Callon, BBN)
- Host Requirements WG (Braden, ISI)
- Routing IP Datagrams through Public X.25 Nets (Rokitansky, DFVLR)
- Internet Multicast (Deering, Stanford)
- TCP Performance Prototyping and Modelling (Jacobson, LBL)
- Cray TCP Performance (Borman, Cray Research)
- DCA Protocol Testing Laboratory (Messing, Unisys)

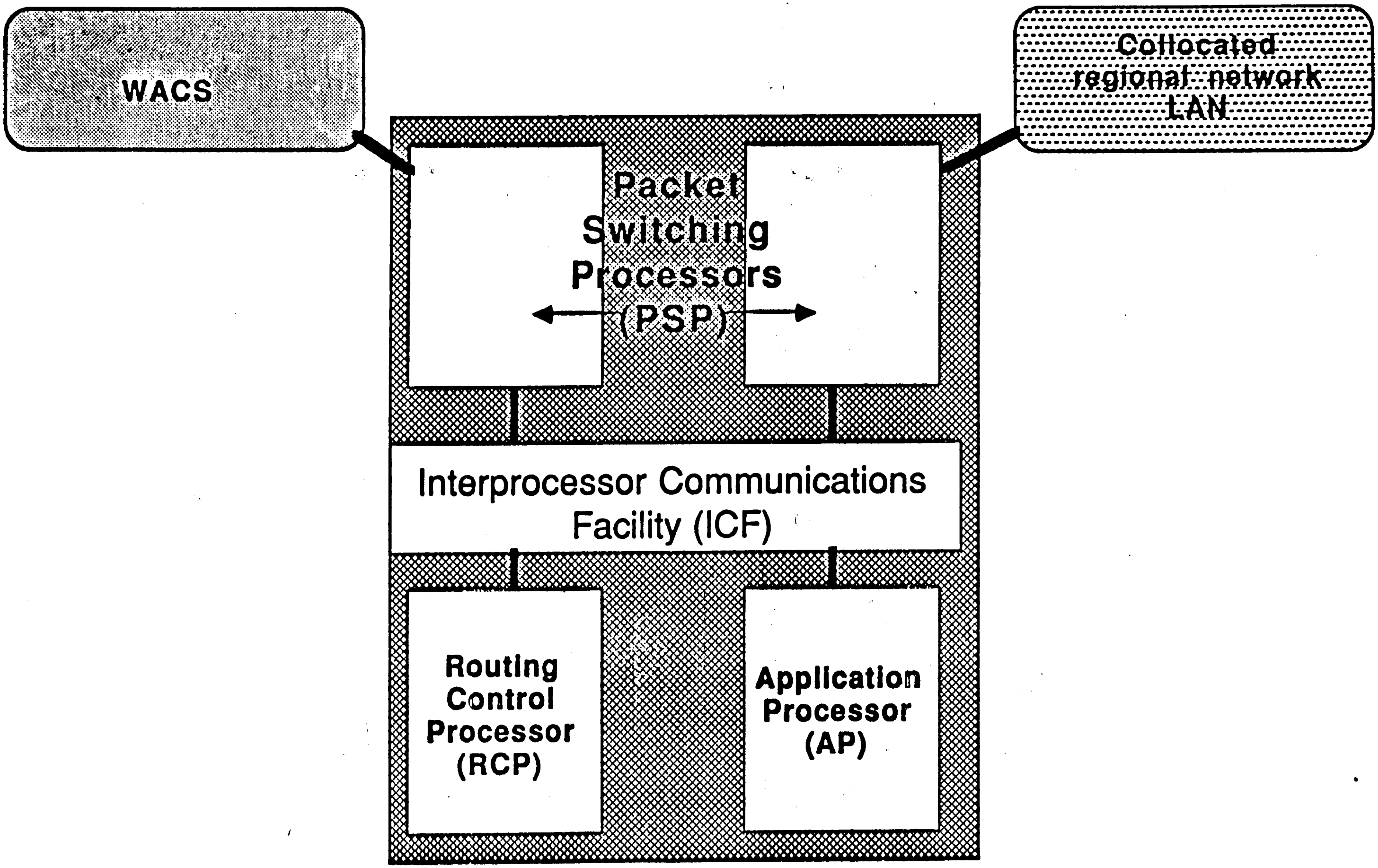
6.1 Report on the New NSFnet—Hans-Werner Braun, UMich

Major components of NSFNET project

- Network management
- Information services
- Advisory role to NSF
- Research

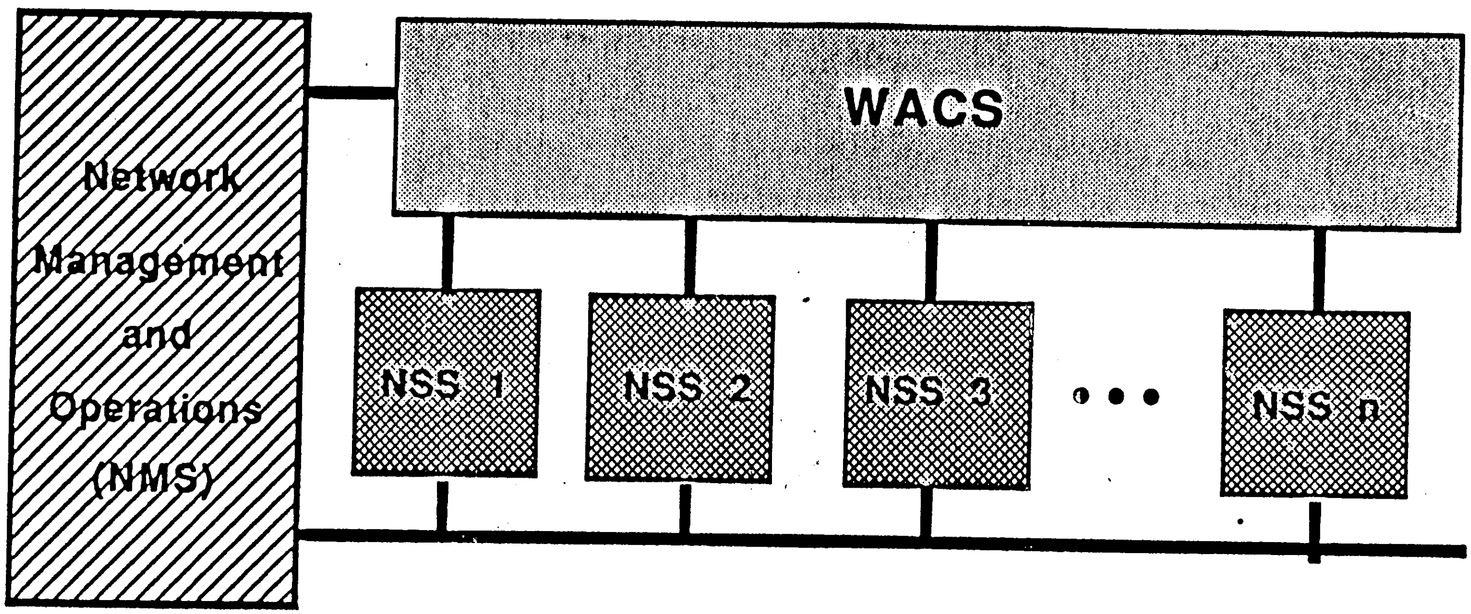


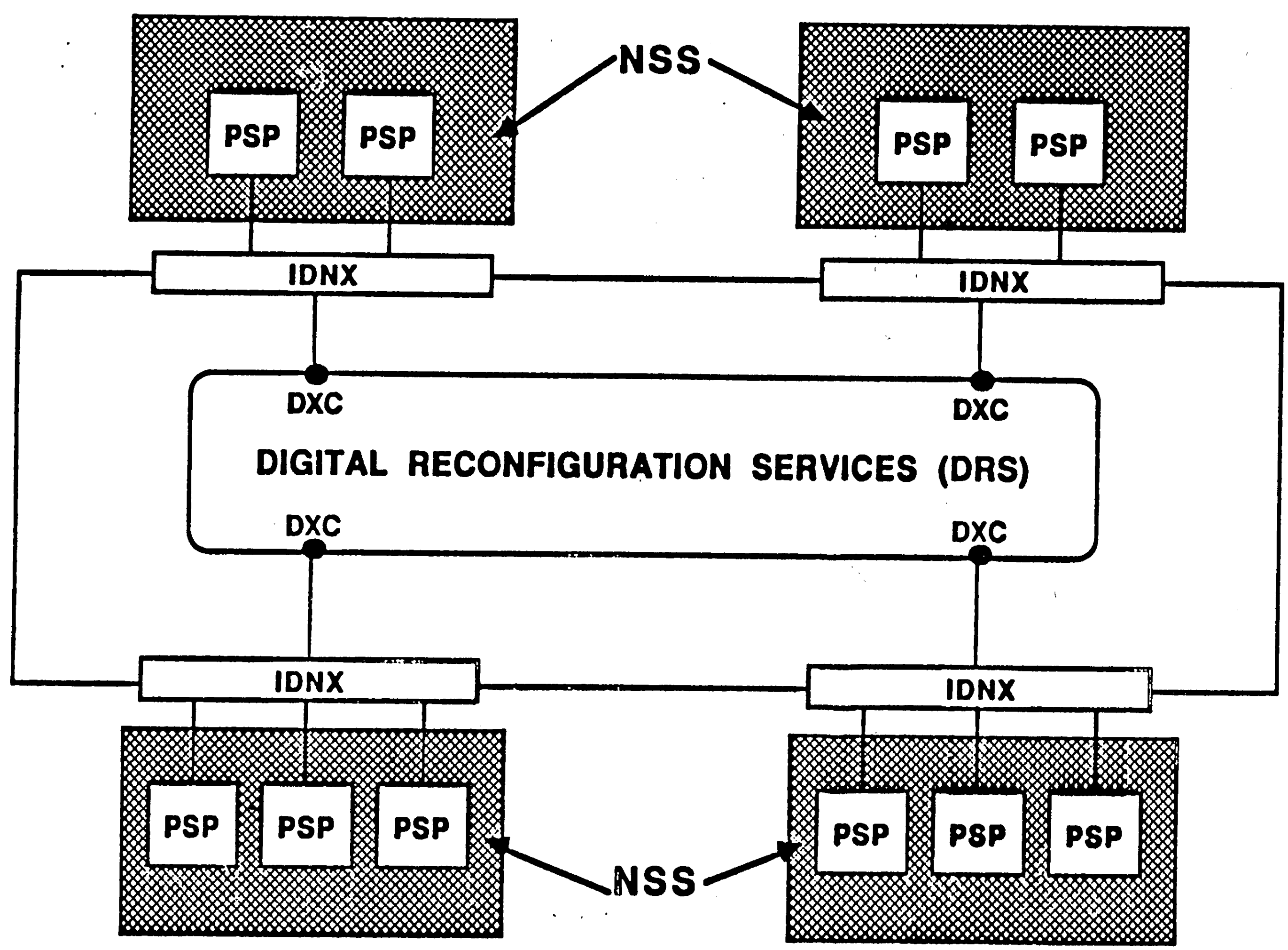
Merit NSFNET Proposed Organization



Simplified nodal switching subsystem (NSS) architecture

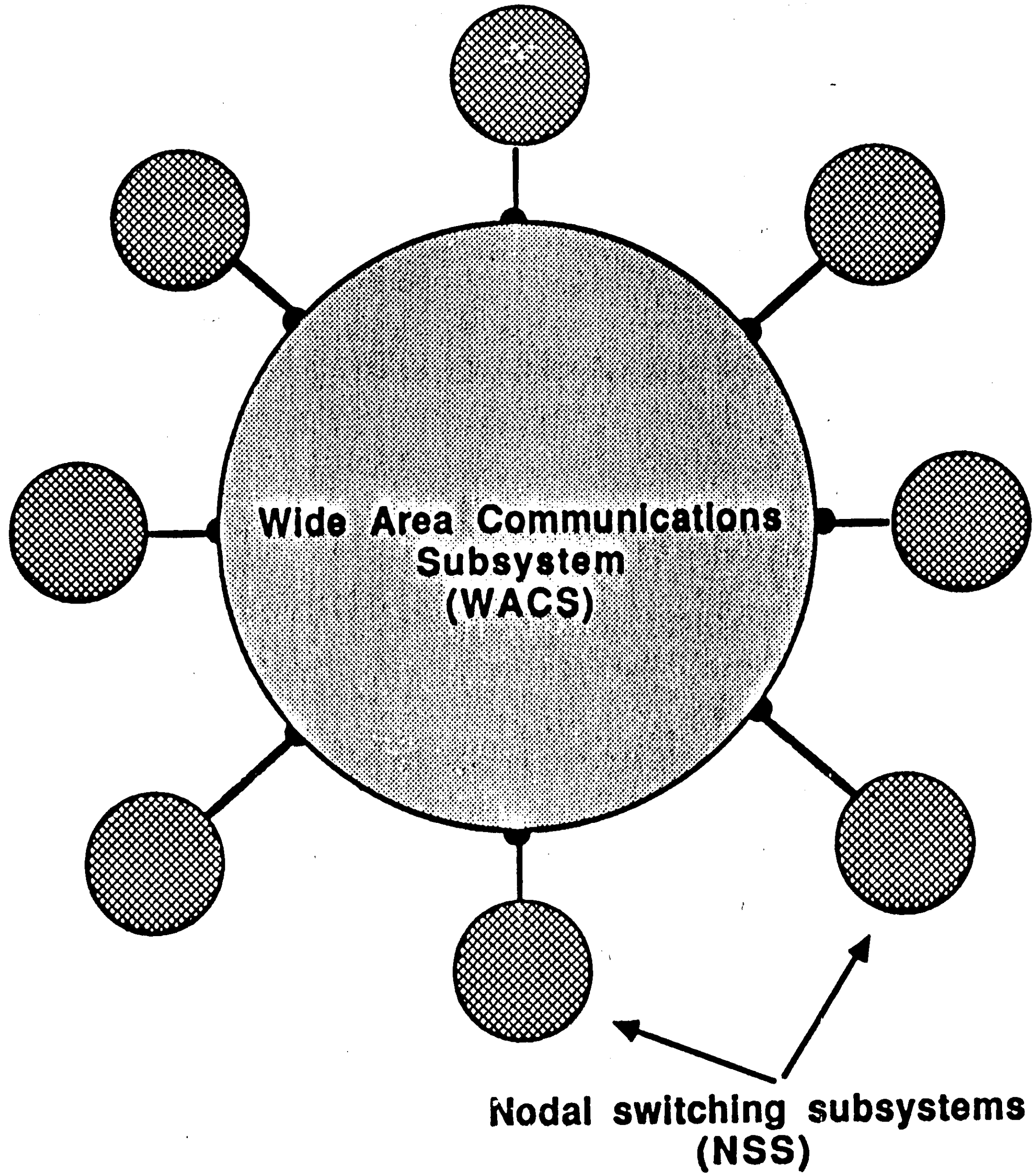
NSFNET subsystems

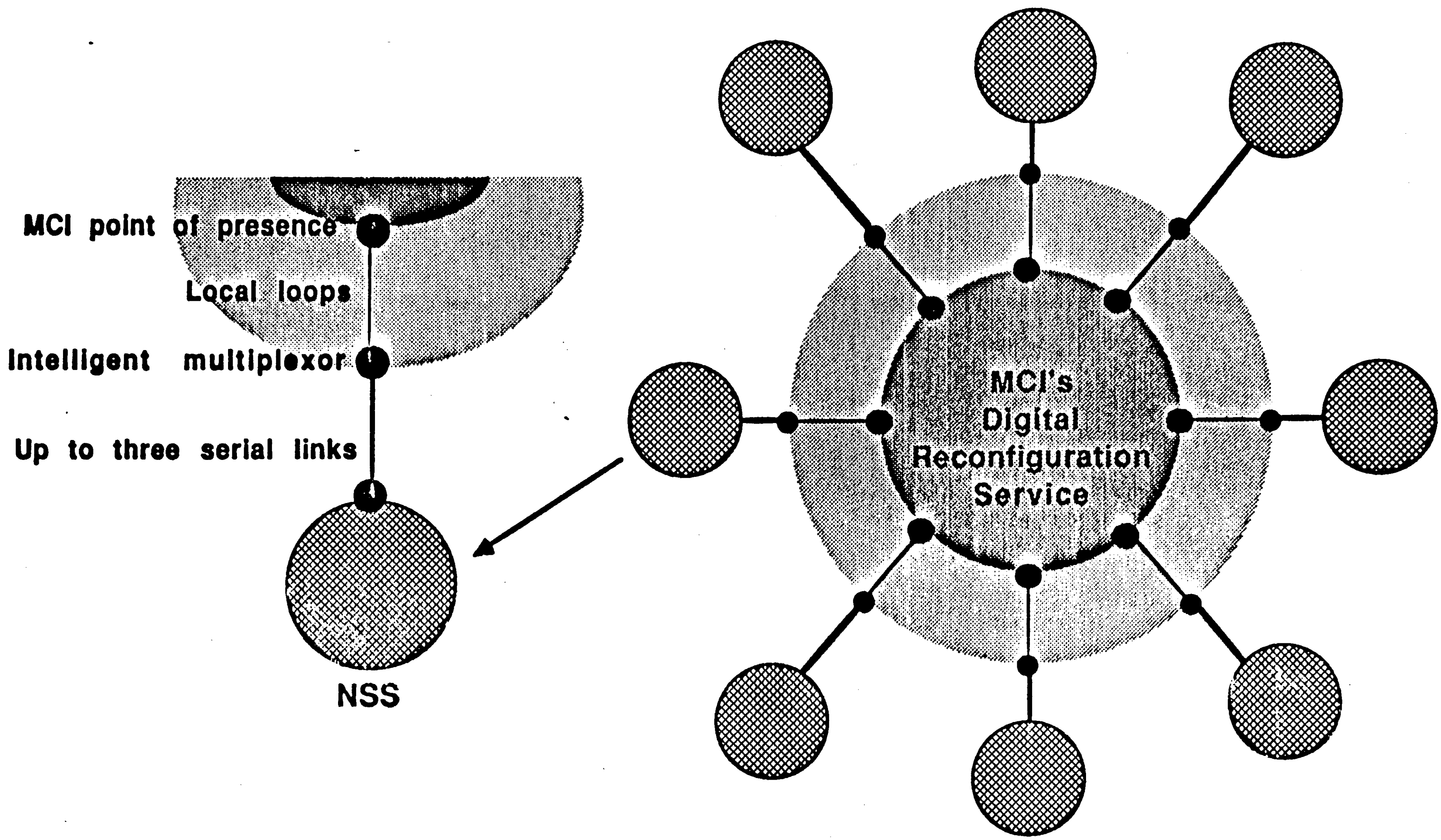




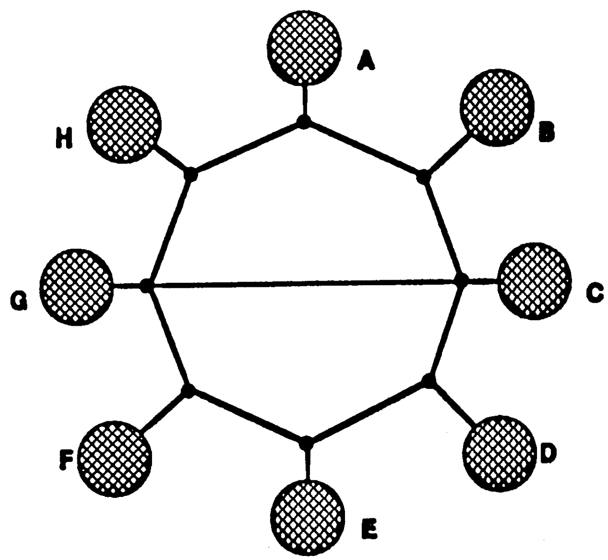
Wide Area Communications Subsystem Architecture

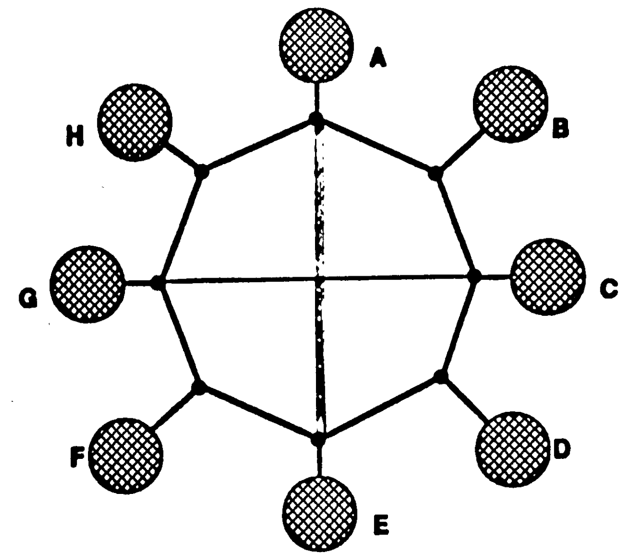
Example eight node network

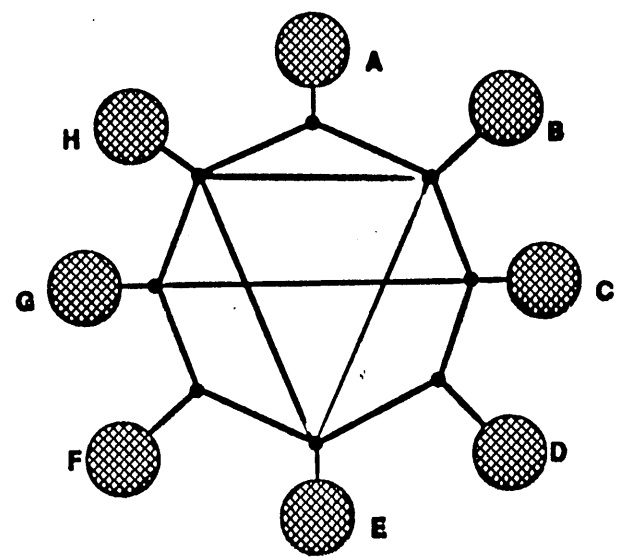


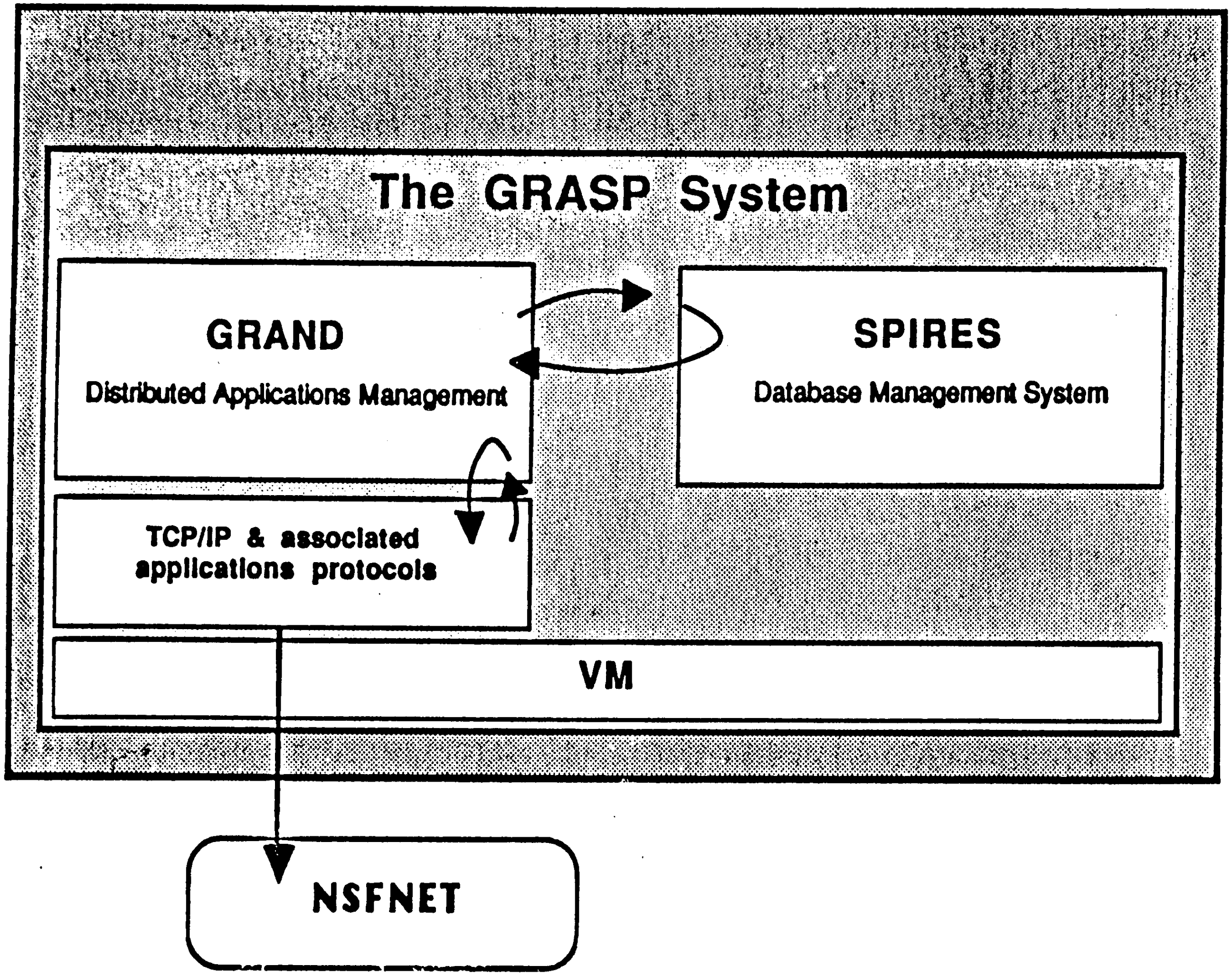

MCI point of presence
Local loops
Intelligent multiplexor
Up to three serial links
NSS
MCI's
Digital
Reconfiguration
Service

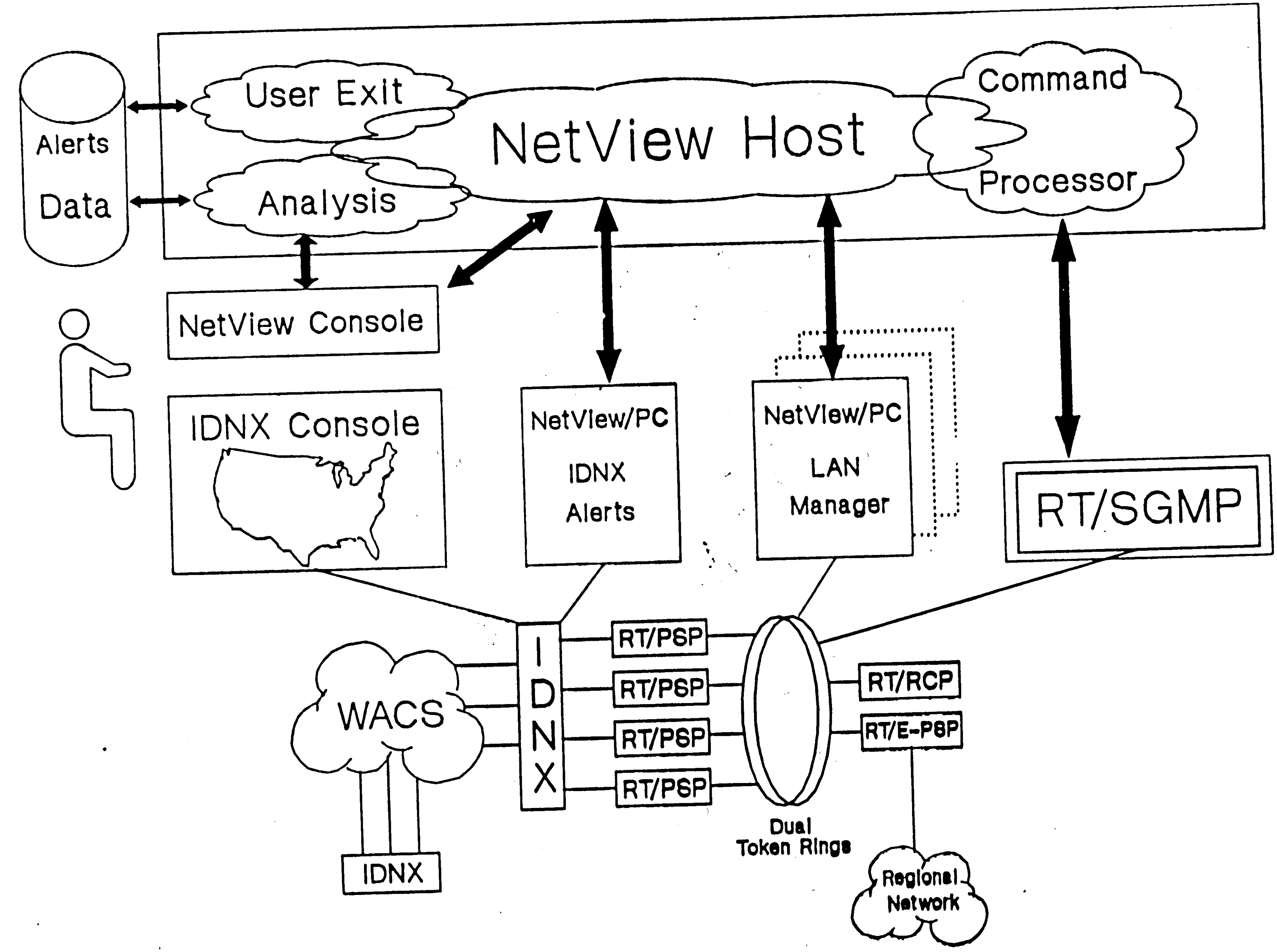
Examples of bandwidth control



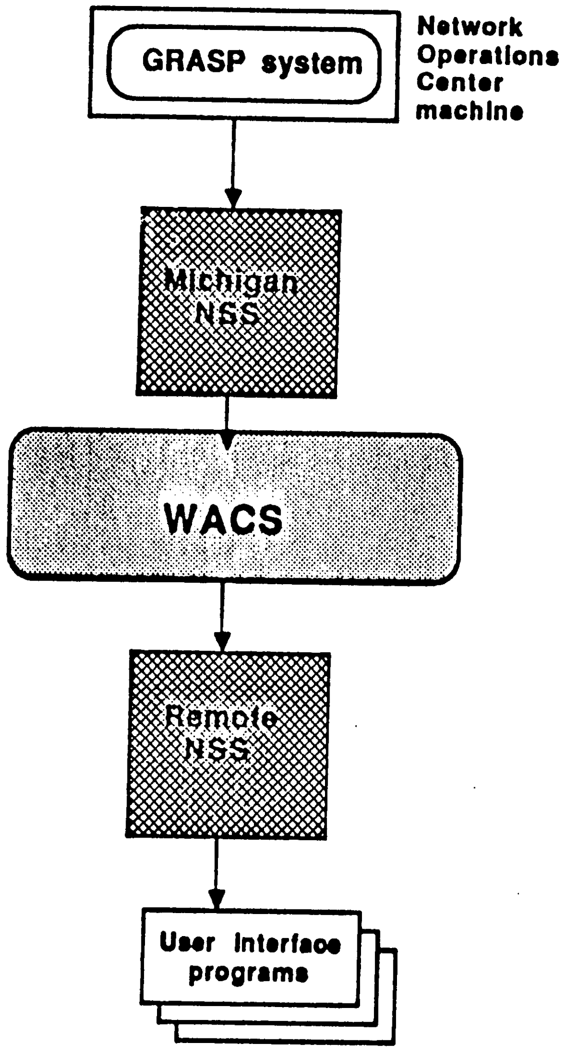





The GRASP System
GRAND
Distributed Applications Management
SPIRES
Database Management System
TCP/IP & associated applications protocols
VM
NSFNET

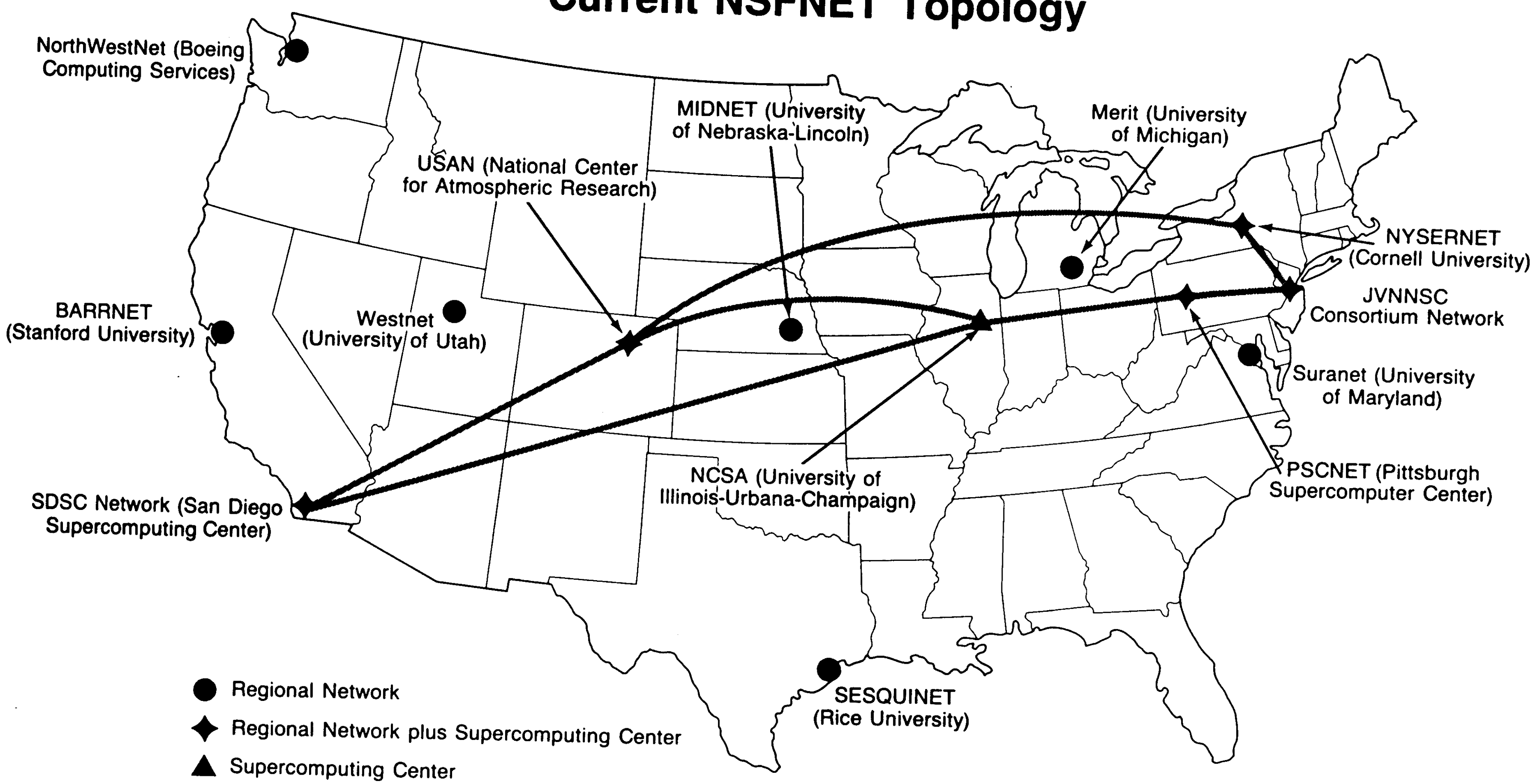


Alerts
Data
User Exit
Analysis
NetView Host
Command
Processor
NetView Console
IDNX Console
NetView/PC
IDNX
Alerts
NetView/PC
LAN
Manager
RT/SGMP
WACS
IDNX
IDNX
RT/PSP
RT/PSP
RT/PSP
RT/PSP
Dual
Token Rings
RT/RCP
RT/E-PSP
Regional
Network



GRASP links to users

Current NSFNET Topology



Center for Cartographic Research and Spatial Analysis, Michigan State University, 2/88

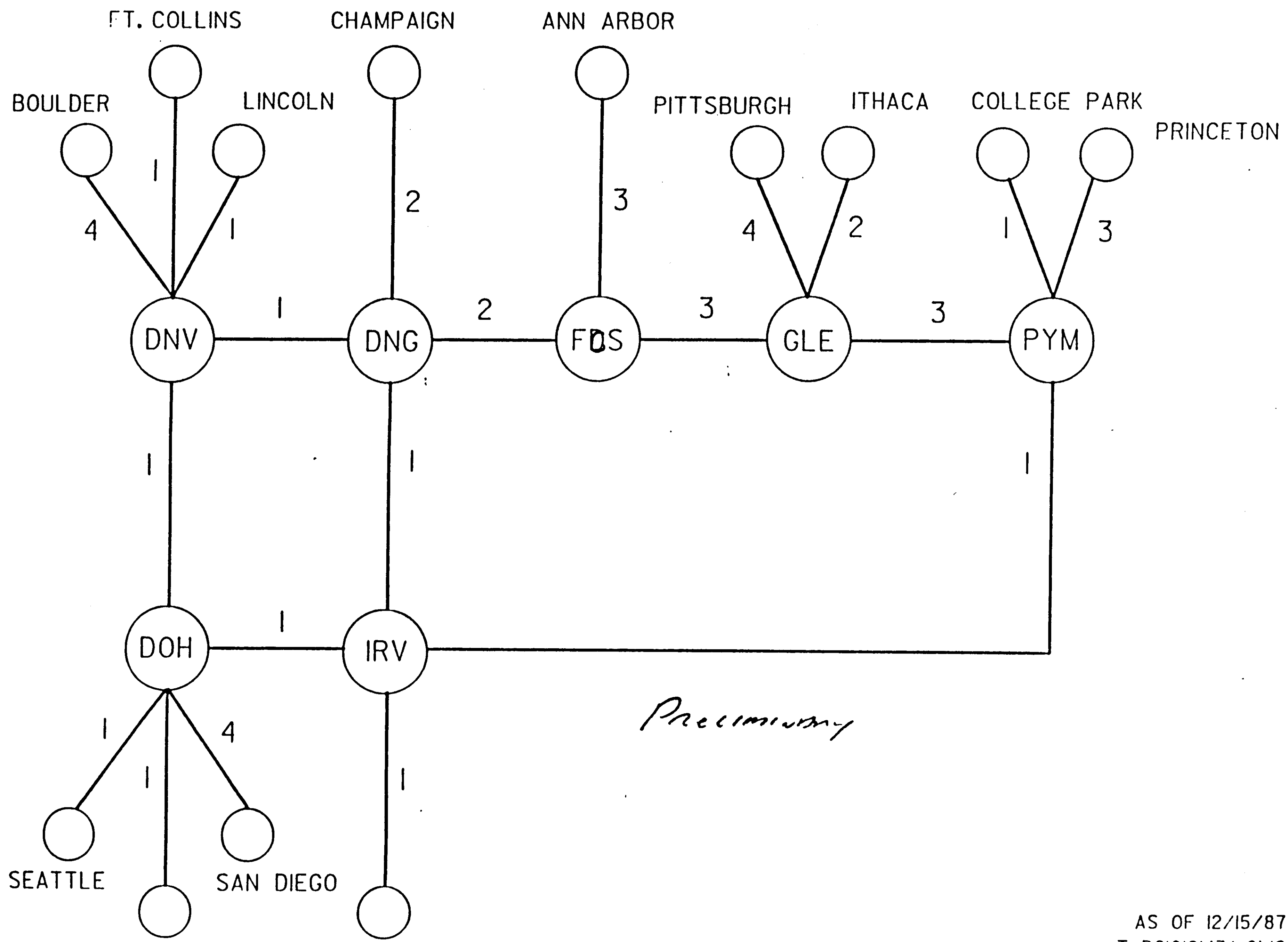
NSF NETWORK
POINT-TO-POINT REQUIREMENTS

NSF CKT#	FROM	TO
1	ANN ARBOR, MI	PRINCETON, NJ
2	PRINCETON, NJ	ITHACA, NY
3	ITHACA, NY	PITTSBURGH, PA
4	PITTSBURGH, PA	ANN ARBOR, MI
5	ANN ARBOR, MI	BOULDER, CO
6	BOULDER, CO	SAN DIEGO, CA
7	SAN DIEGO, CA	CHAMPAIGN, IL
8	CHAMPAIGN, IL	PITTSBURGH, PA
9	SEATTLE, WA	SAN DIEGO, CA
10	PALO ALTO, CA	SAN DIEGO, CA
11	FT. COLLINS, CO	BOULDER, CO
12	LINCOLN, NE	BOULDER, CO
13	COLLEGE PARK, MD	PRINCETON, NJ
14	HOUSTON, TX	PITTSBURGH, PA

TEST NETWORK REQUIREMENTS

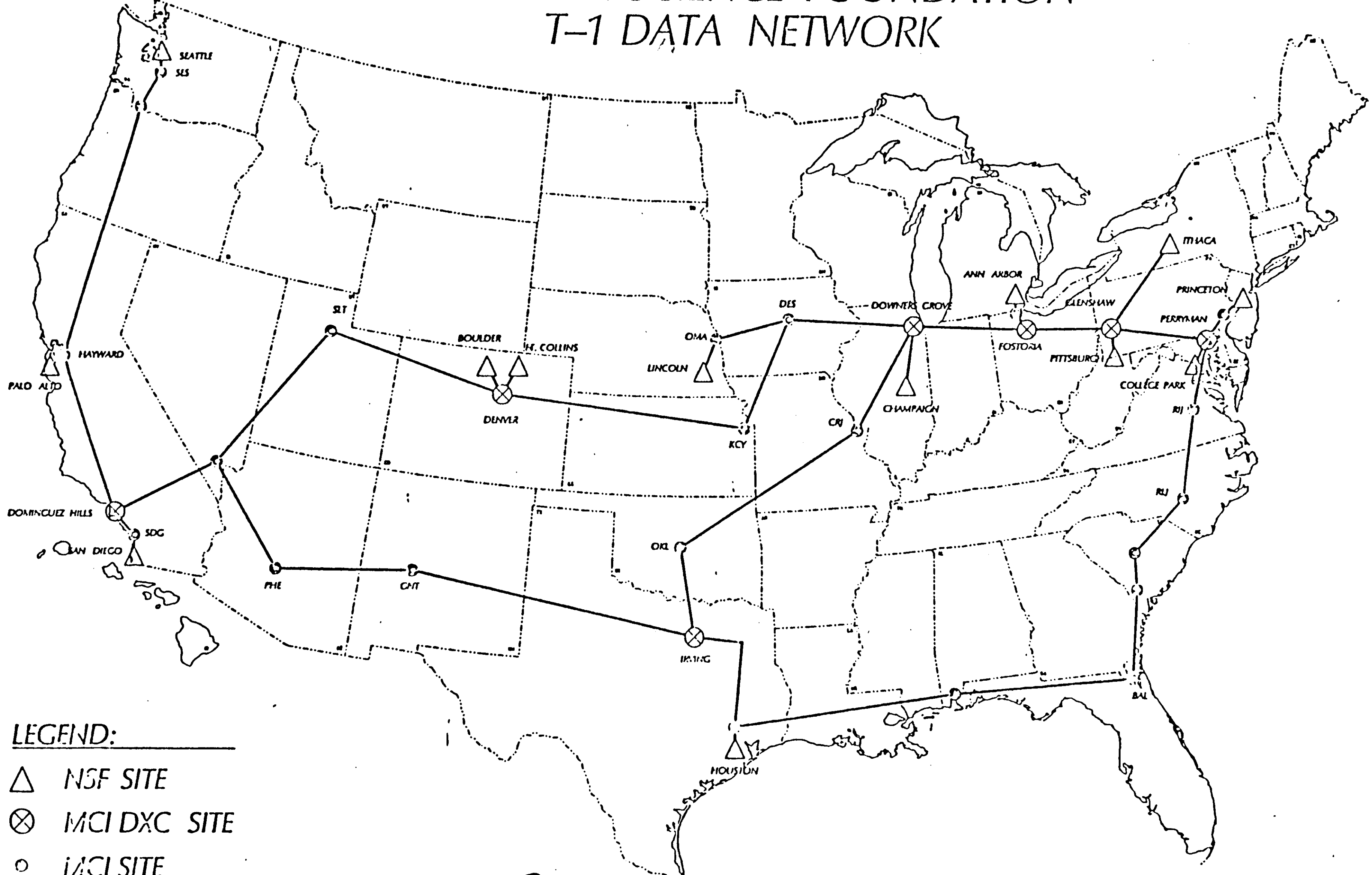
	FROM	TO	
1	YORKTOWN, NY	RESTON, VA	INSTALL TO DEMARC
2	YORKTOWN, NY	MILFORD, CT	INSTALL TO DEMARC
3	ANN ARBOR, MI	MILFORD, CT	
4	ANN ARBOR, MI	RESTON, VA	

NSF NETWORK



AS OF 12/15/87
T DGI0I0I43A.0LIS

NATIONAL SCIENCE FOUNDATION
T-1 DATA NETWORK



LEGEND:

△ NSF SITE

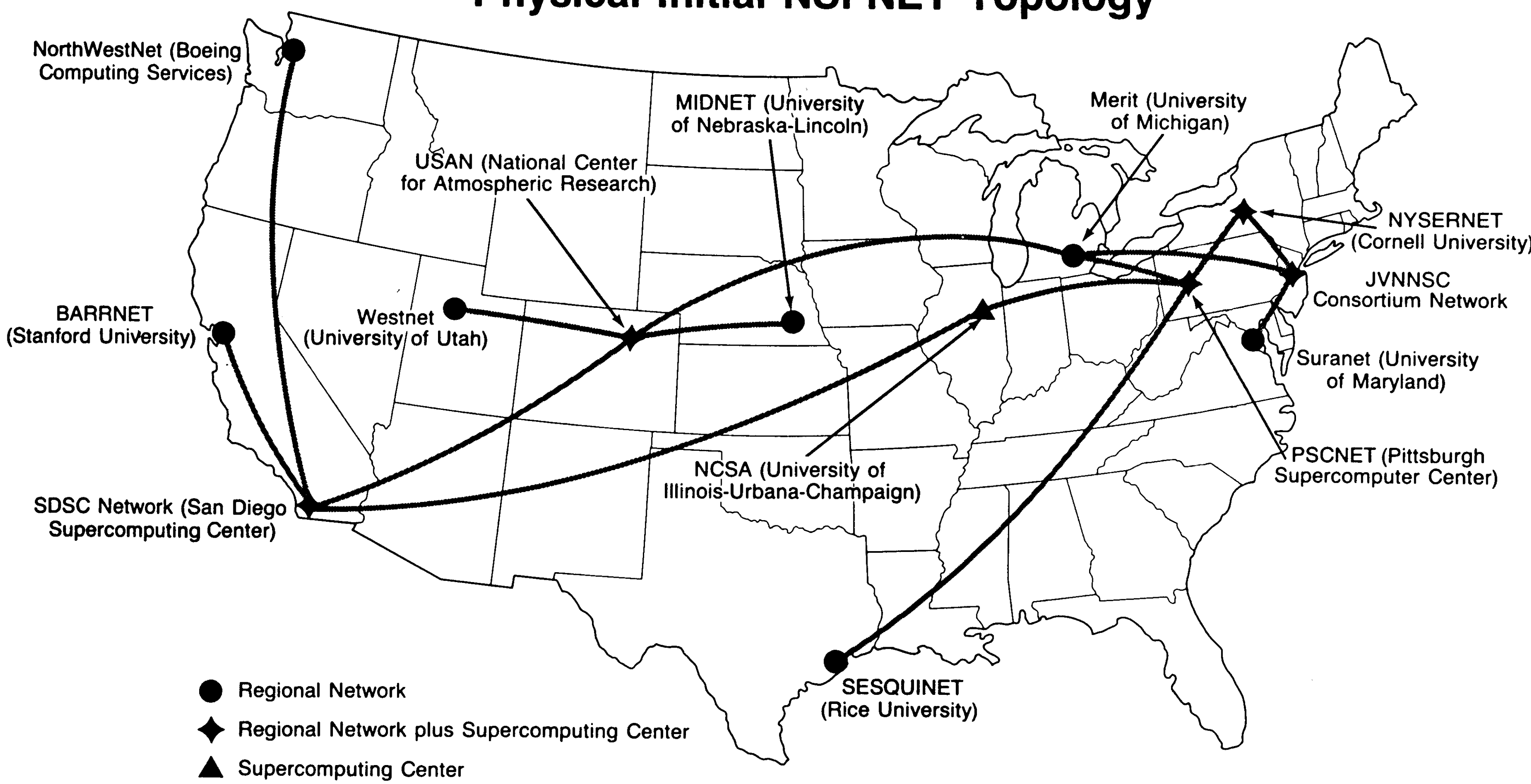
⊗ MCI DXC SITE

○ MCI SITE

Preliminary

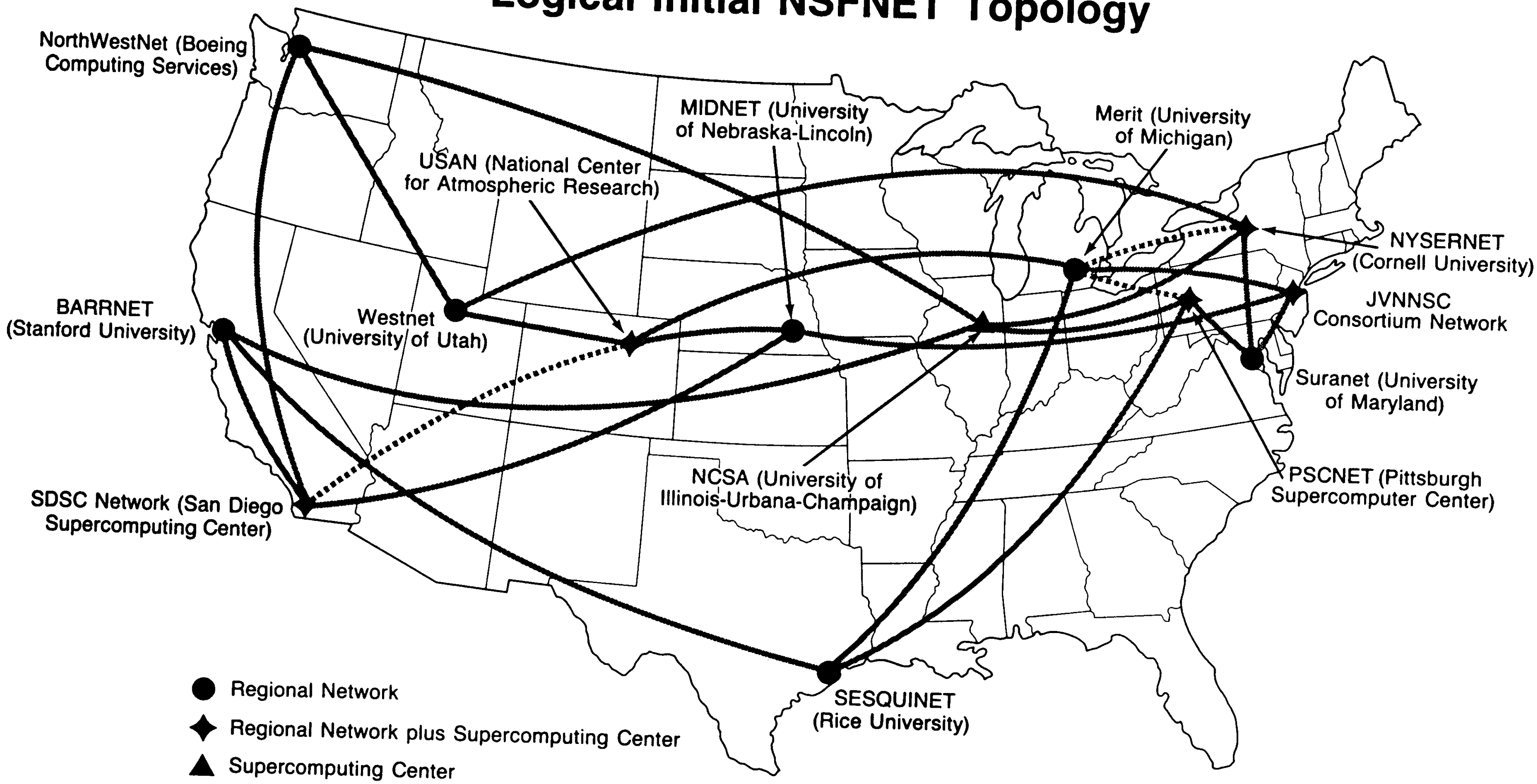
T MP#36004D.OLIS

Physical Initial NSFNET Topology

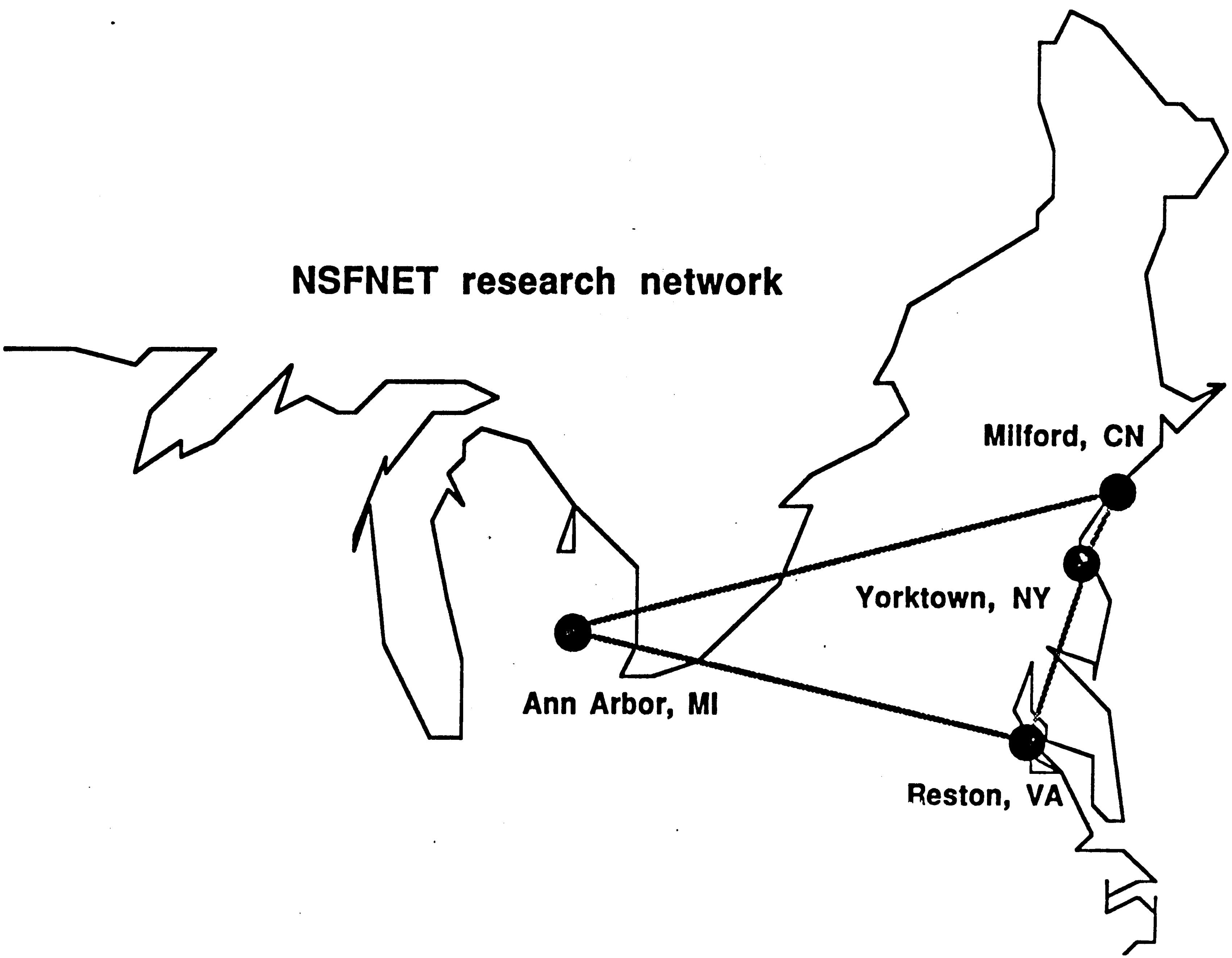


Center for Cartographic Research and Spatial Analysis, Michigan State University, 2/88

Logical Initial NSFNET Topology



Center for Cartographic Research and Spatial Analysis, Michigan State University, 2/88


NSFNET research network
Milford, CN
Yorktown, NY
Ann Arbor, MI
Reston, VA

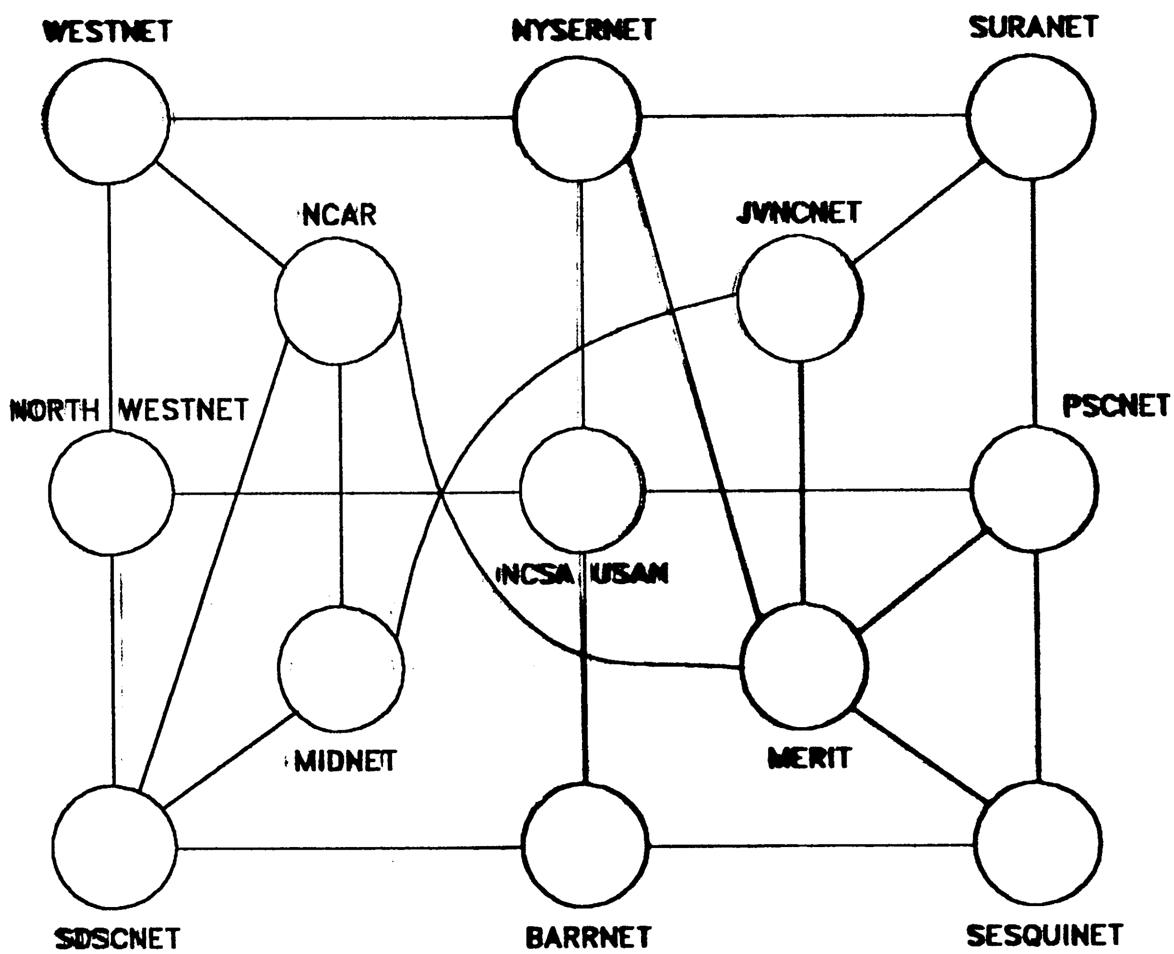
6.2 Report on the New NSFnet (Cont.)—Jacob Rekhter, IBM

Topics

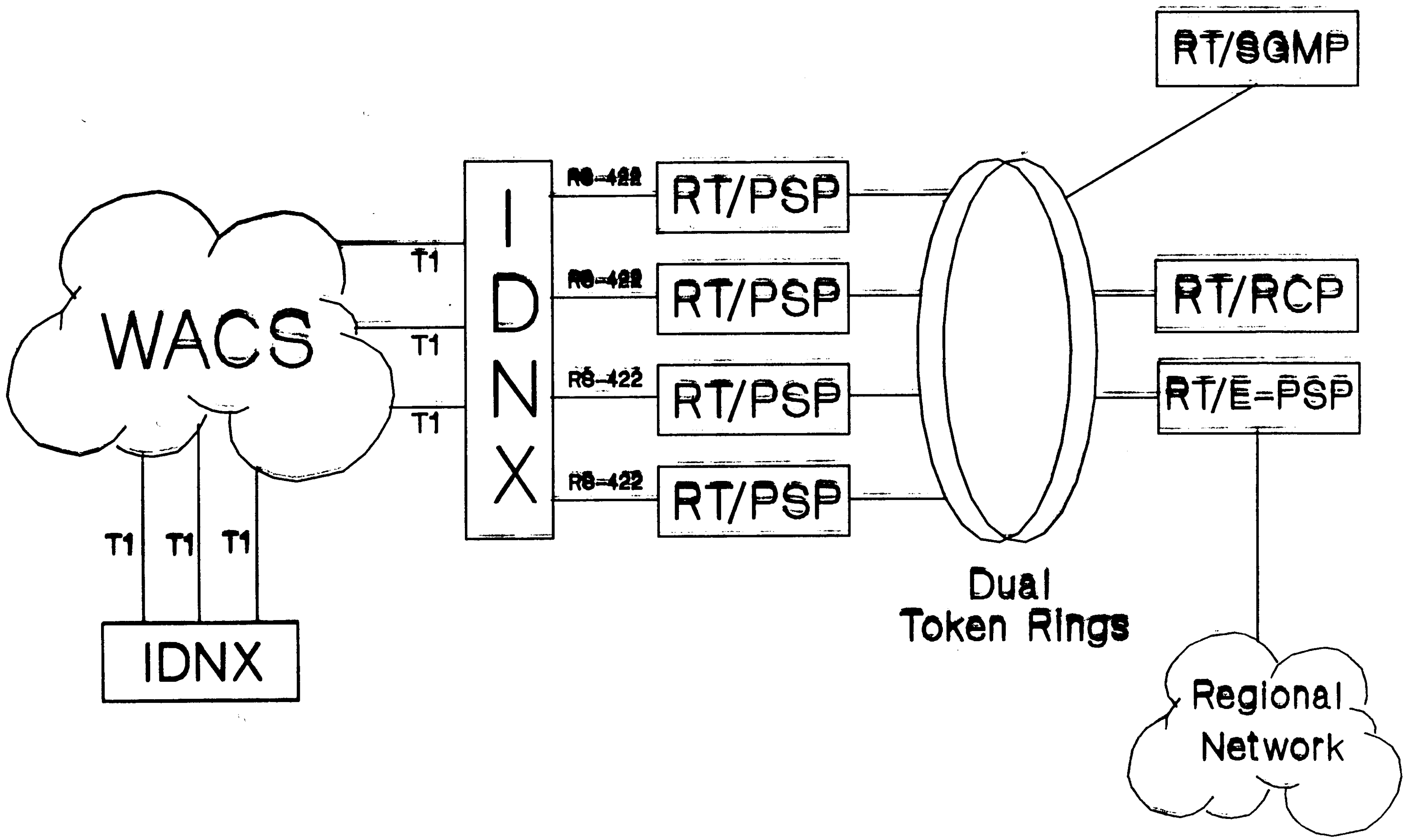
- NSS architecture
- Intra-NSS software
- Inter-NSS software
- NSS-Regional interaction
- Others ....

NSFNET Work Products

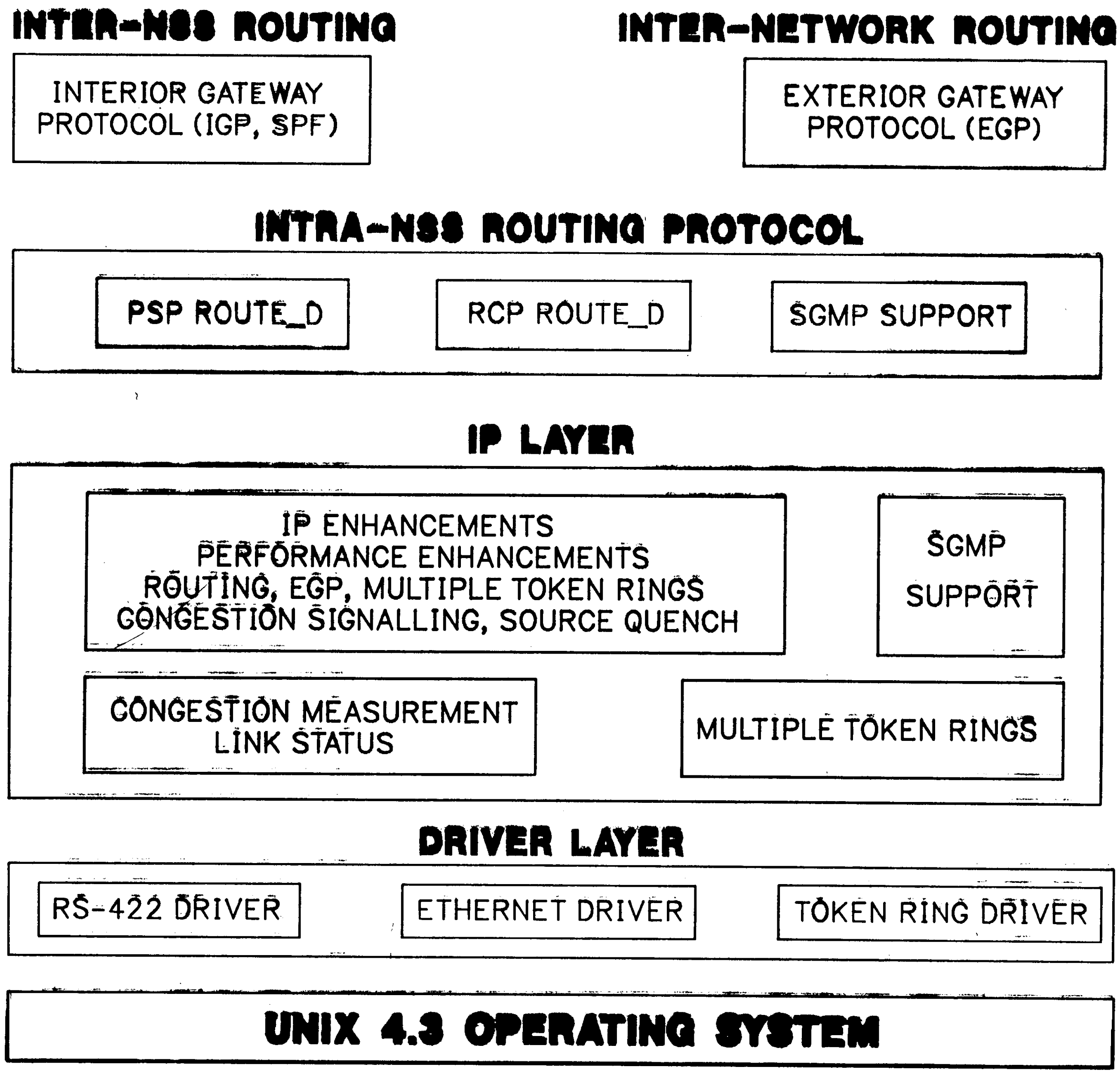
Wide Area Communications Subsystem (WACS)
Logical Topology



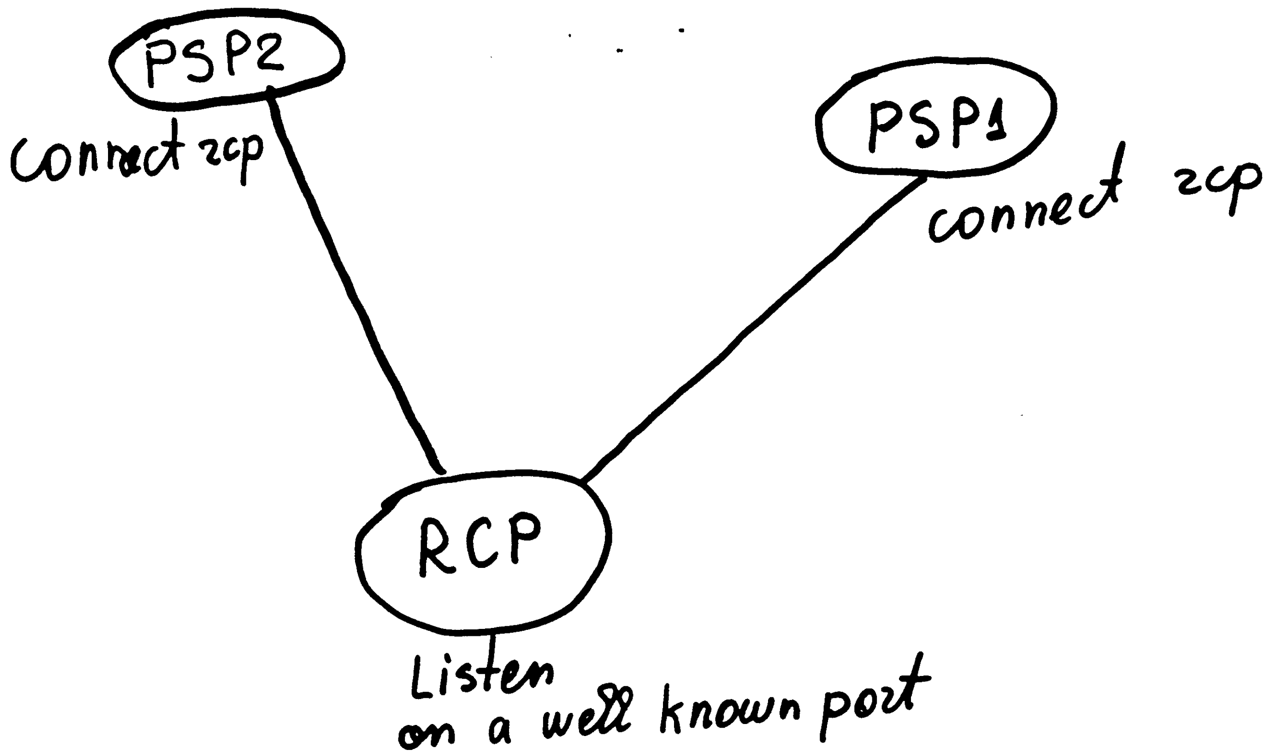
Packet Switching Logical Topology


WACS
T1
T1
T1
T1
T1
T1
IDNX
IDNX
RS-422
RS-422
RS-422
RS-422
RT/PSP
RT/PSP
RT/PSP
RT/PSP
Dual
Token Rings
RT/SGMP
RT/RCP
RT/E-PSP
Regional
Network

NSS SOFTWARE



Intra-NSS software



TCP based connections

Intra-NSS software

- Master (RCP) - Slave (PSP) model
- Request/Reply
- Unsolicited Reply
- Passthrough in PSP

Intra-NSS software

- Interfaces
  - address
  - status
- Route add/delete/change
- Interface status changed
  - unsolicited
- Interface queue length changed
  - unsolicited
  - threshold
  -

Inter-NSS software

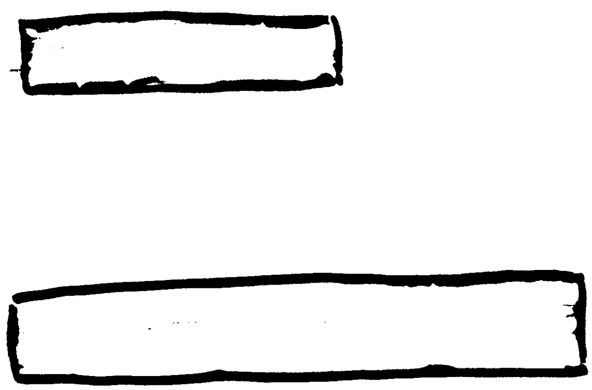
- Based on ANSI IS-IS routing
- Close to IDEA005.TXT

Inter-NSS software

- IS-IS routing
  - 119 pages document
  - Intra-Domain Routing
  - Hierarchical - 2 levels
  - Algorithm is SPF

Inter-NSS software

- IS-IS Routing Protocol
  - Subnetwork Independent Functions
    - + Level 1
    - + Level 2
    - + Repairing partitioned areas
  - Subnetwork Dependent Functions
    - + Point-to-point
    - + ISO 8208
    - + Broadcast subnetworks



Inter-NSS Software

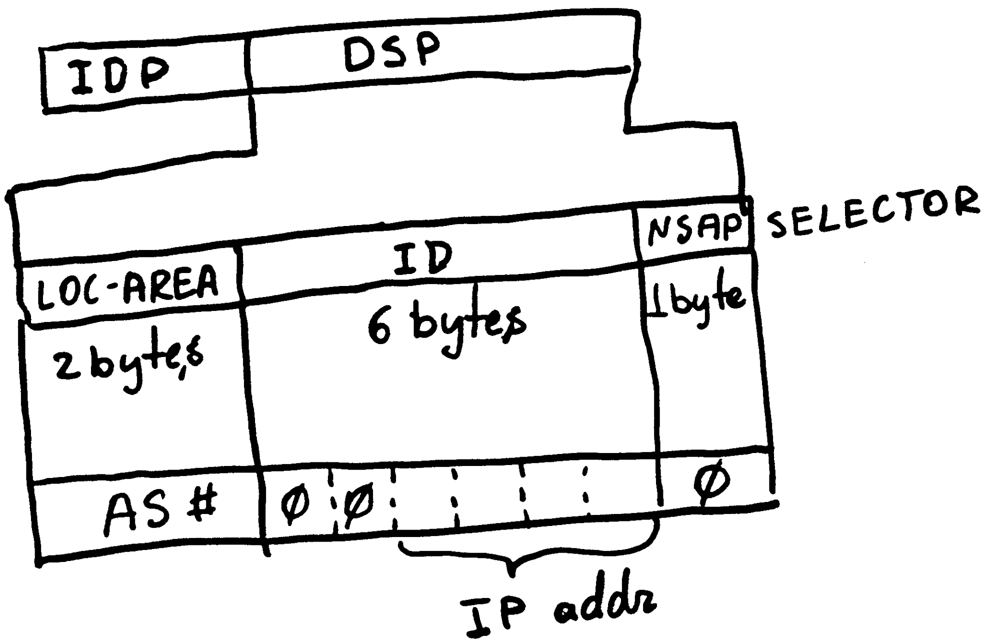
- Algorithm – SPF modified
  - Complexity O(e)
  - permits load splitting by identifying a set of equal cost paths to each destination

Inter-NSS Software

- Protocol
  - PDU - Protocol Data Unit
  - Router Links PDU
  - End System Links PDU
  - "Flooding"
  - Runs on top of Link Layer (in NSFNET on top of IP)

Inter-NSS Software

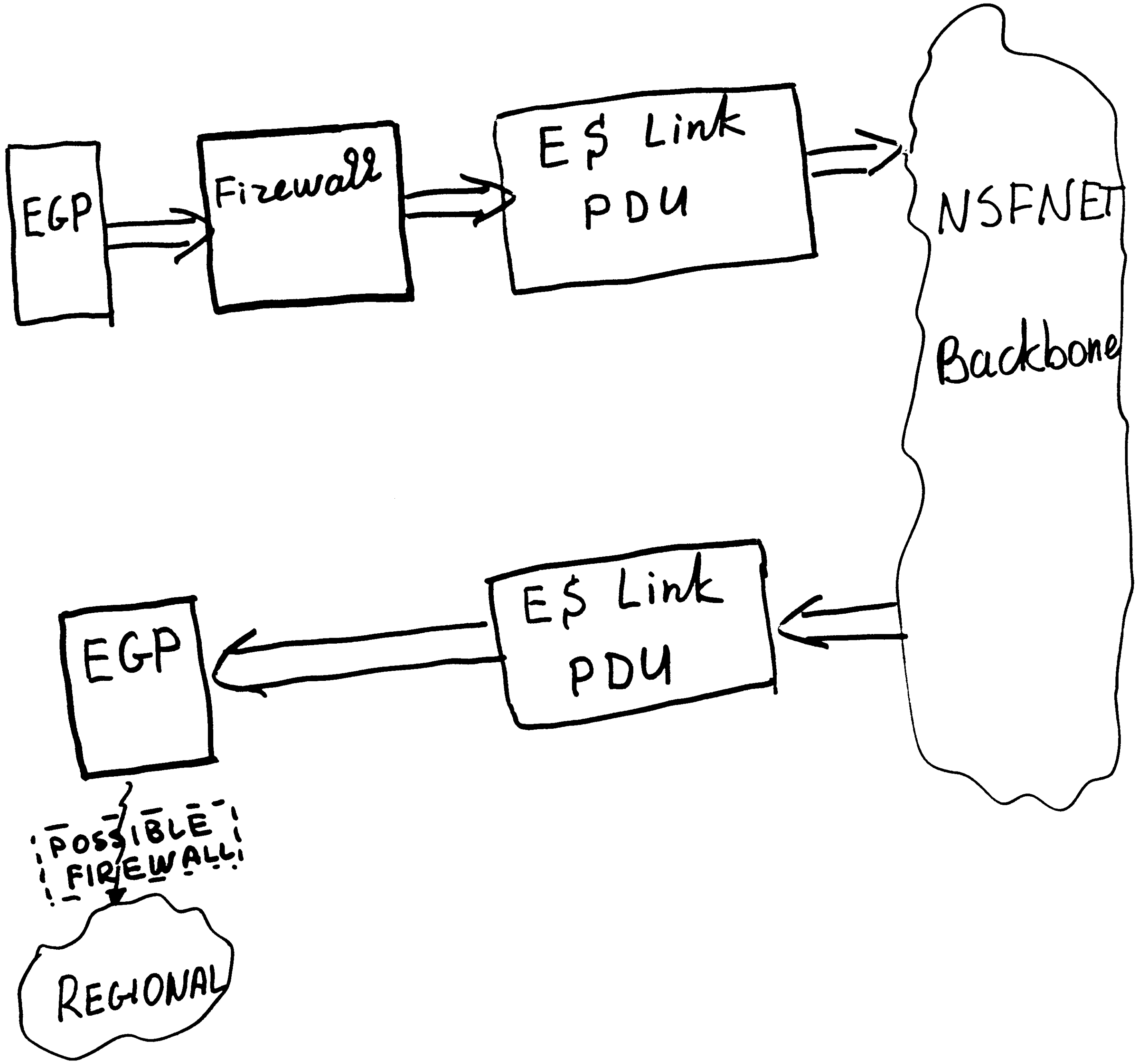
- IP mapping
  - ISO address



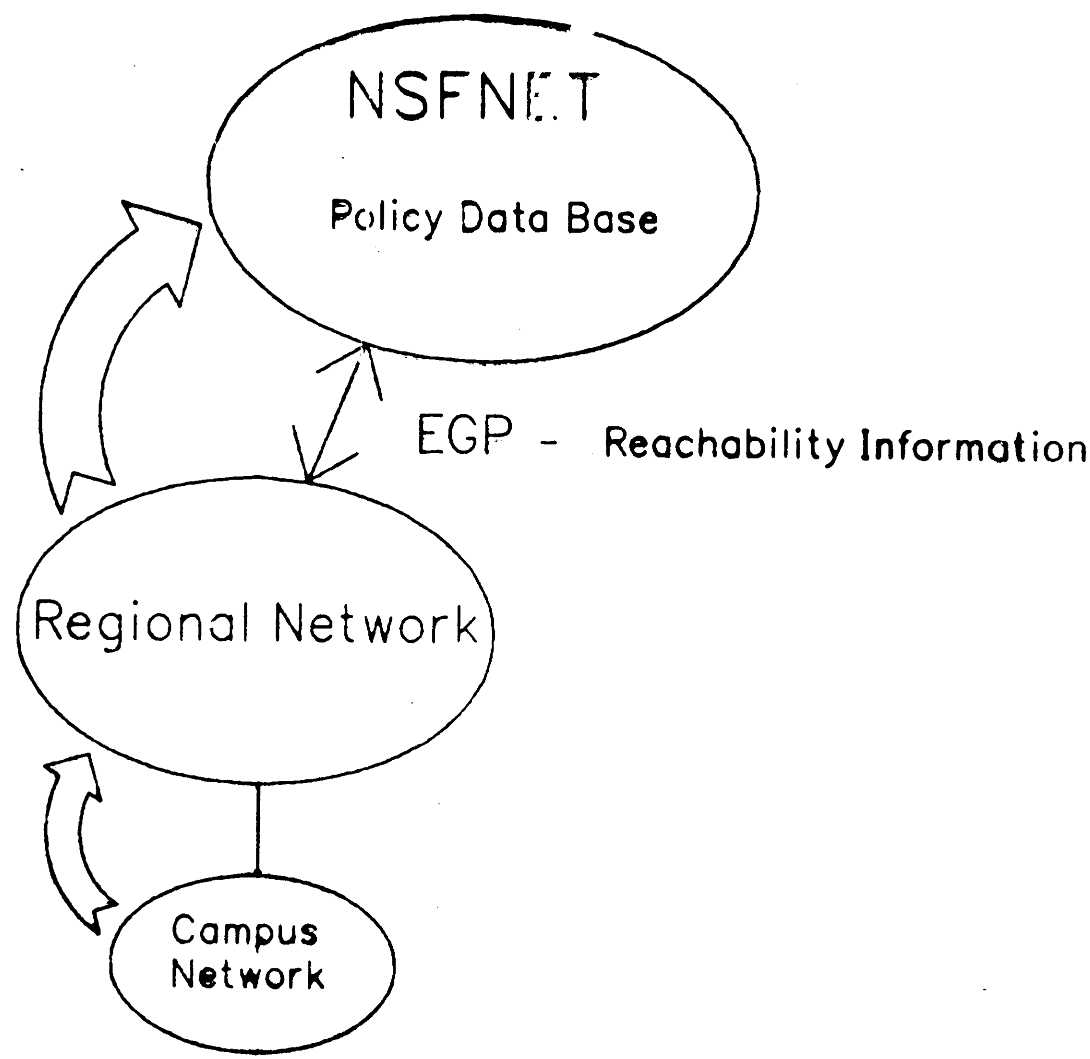
NSS-Regional Interaction

- EGP as REACHABILITY ONLY
- EXPLICIT FIREWALL
  (similar to GATED)
- UTILIZES ES Link PDU
  (Level 2)

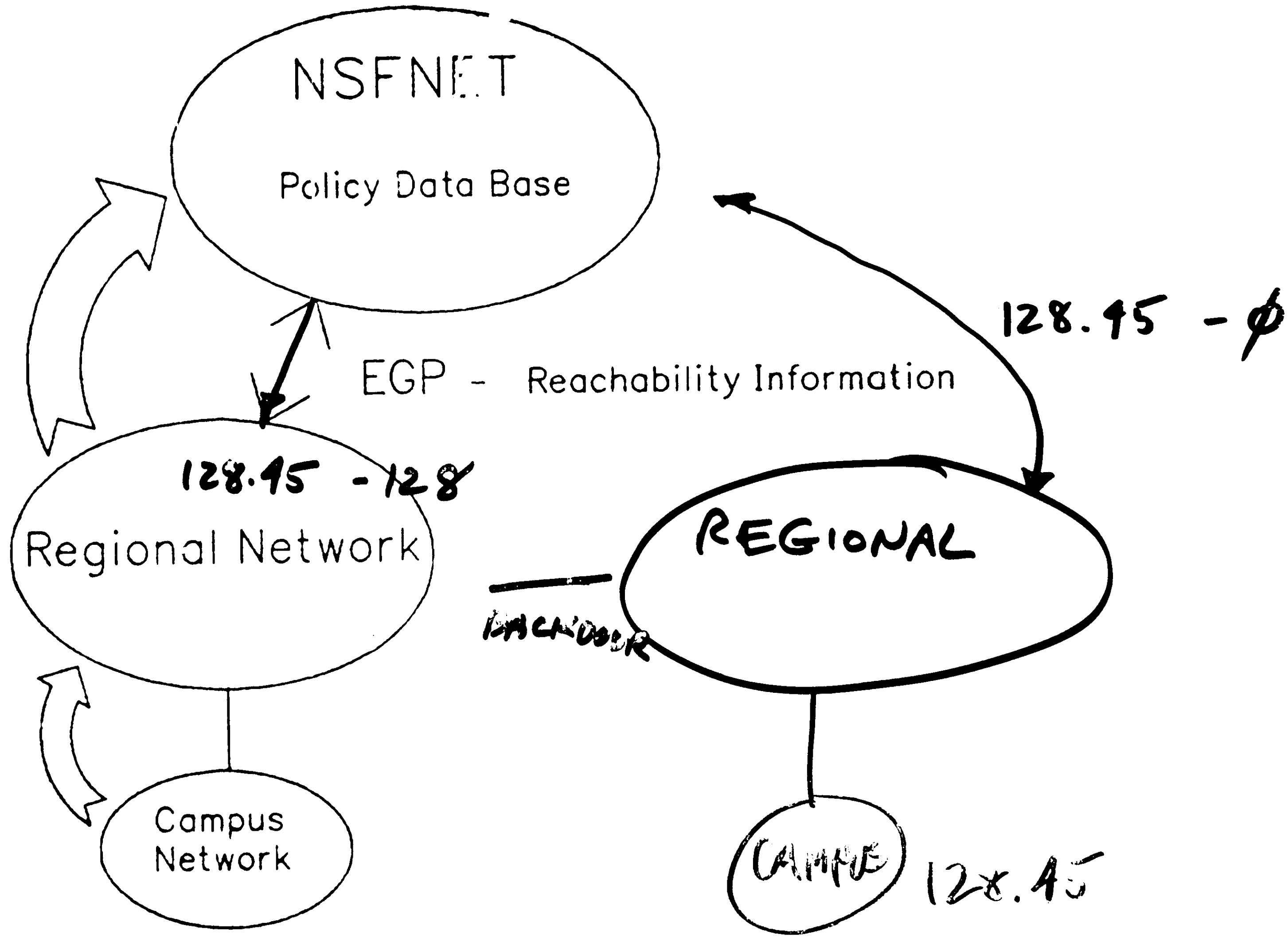
NSS - Regional Interaction



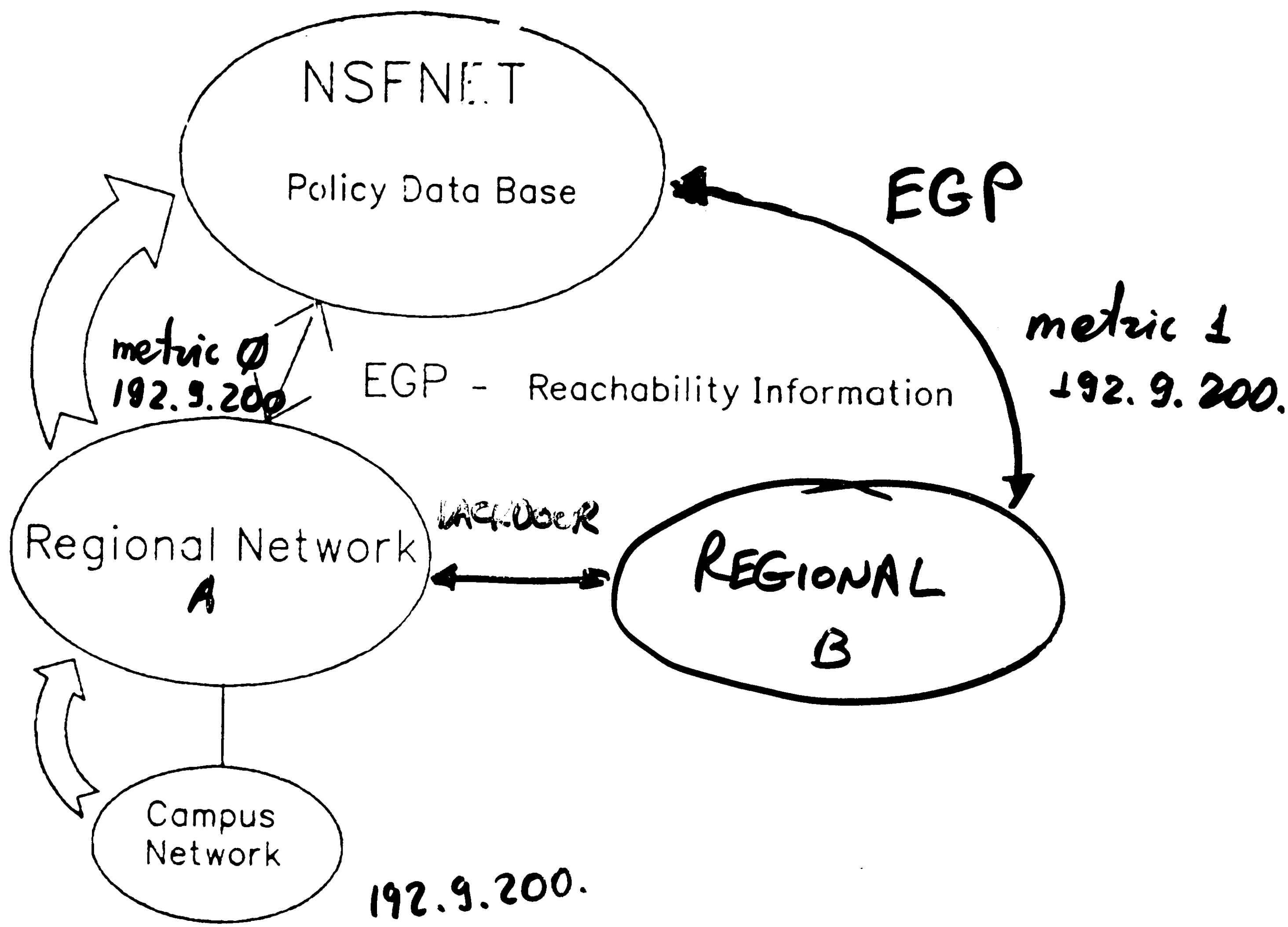
Policy Based Routing



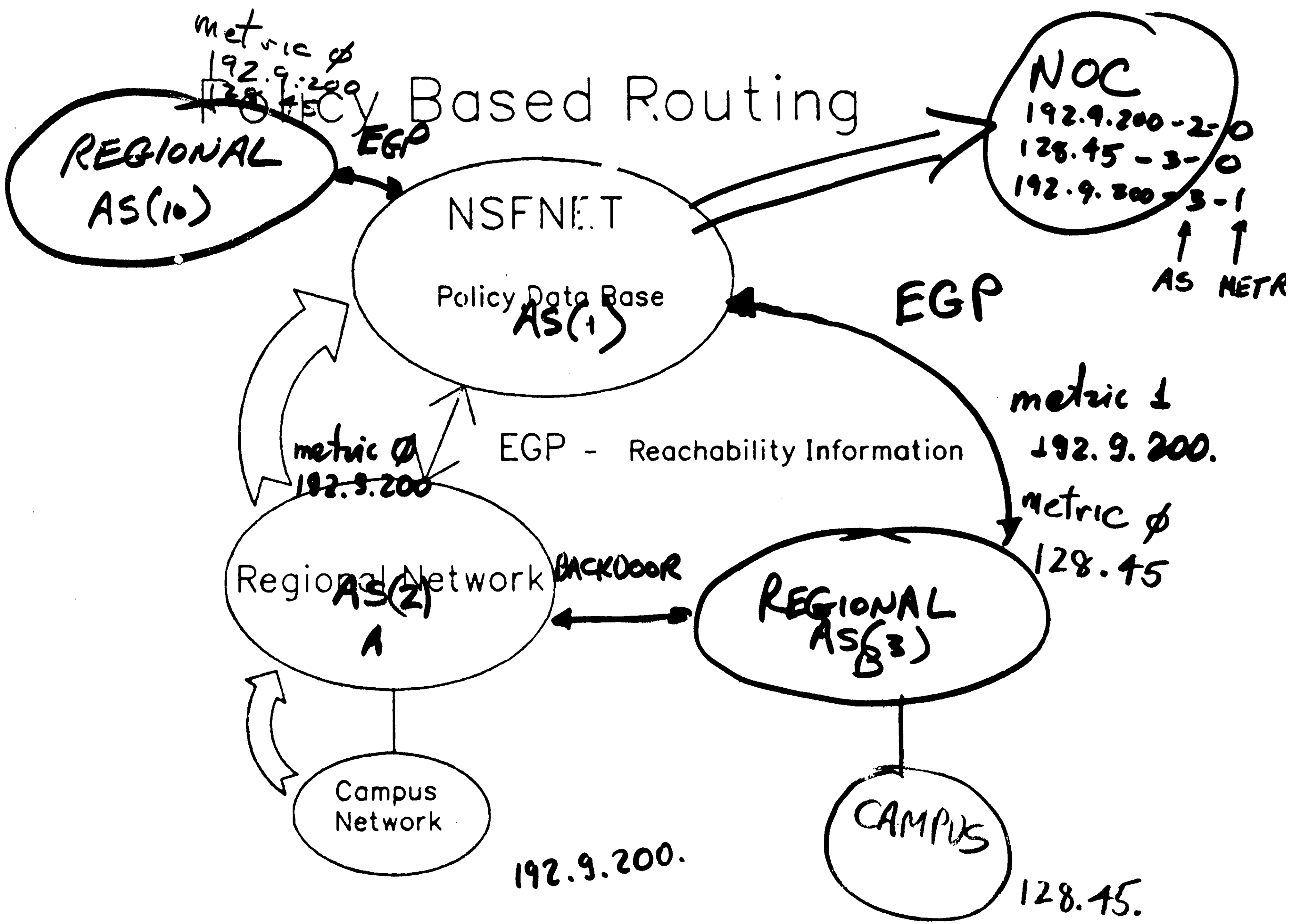
Policy Based Routing



Policy Based Routing



Policy Based Routing



Other issue:

- High priority for Routing Packets
- ICMP Source Quench
- Preemption
  - encourage "social behavior"
  - fairness ?
  - TOS support

6.3 Status of the Adopt-A-GW Program—Bob Enger, Contel

Motivation:	Internet performance seemed poor
	Felt there was trouble right here in River City
Problem:	Core gateways underpowered
	EGP processing slows down EGP core servers
	EGP extra hop

Why wasn't anyone dealing with the situation?

ADOPT -A-GATEWAY

An Interim Solution

Purpose: To improve our Internet standard of living

Methods: Upgrade the hardware platforms of critical core gateways.

Bypass all formalities.

Incur zero cost.

Procedure: Advertise for the free loan of hardware.

Install the boards in the machines.

Pray they work.

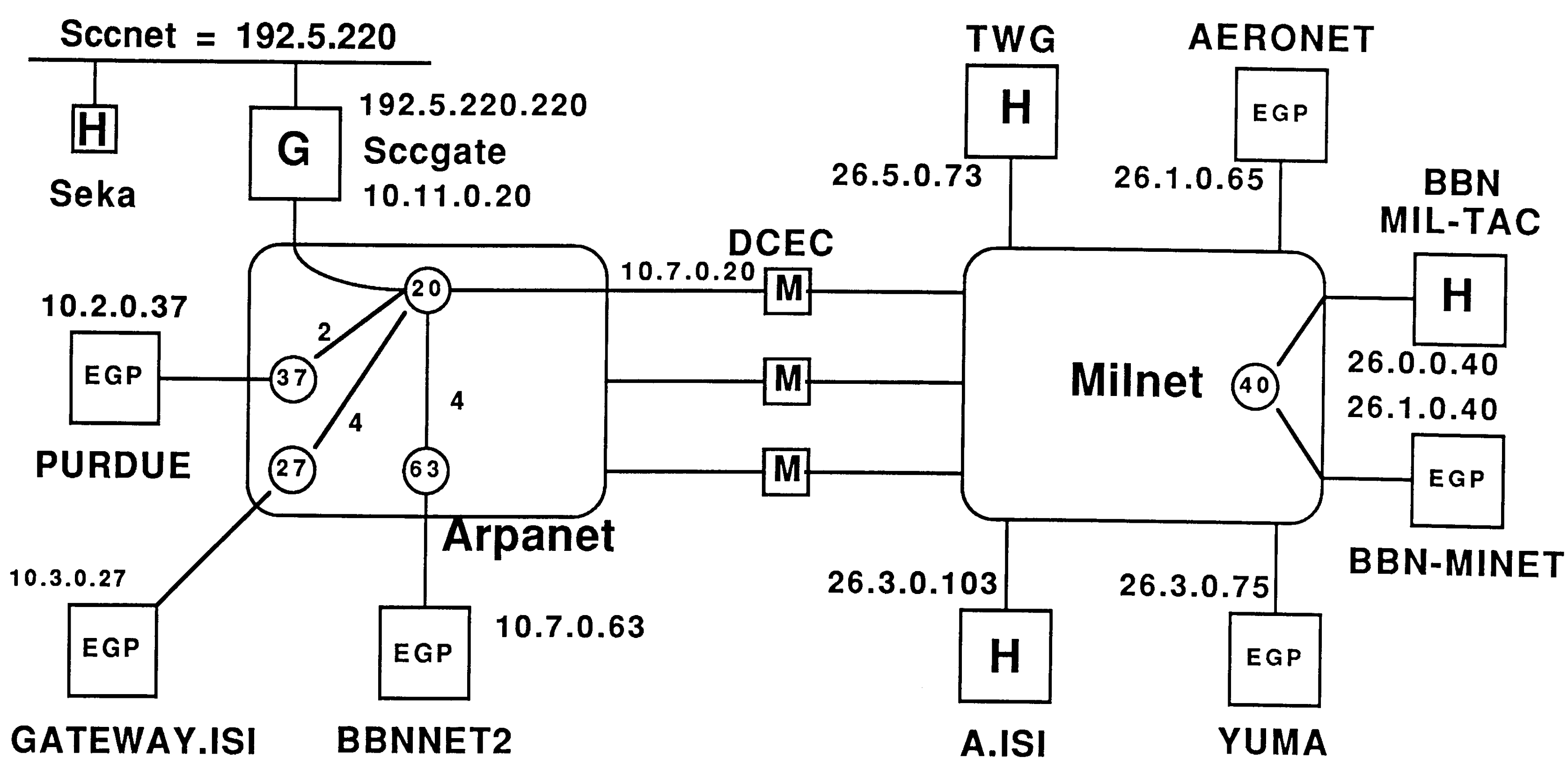
ADOPT-A-GATEWAY

Co-conspirators:	Robert Enger, Contel Phill Gross, Mitre Corp. Annette Bauman, D.C.A.
Supporting Cast:	Steve Atlas, BBN Mike Karels, U.C. Berkeley Jerry Scott, Wollongong

FOSTER PARENTS

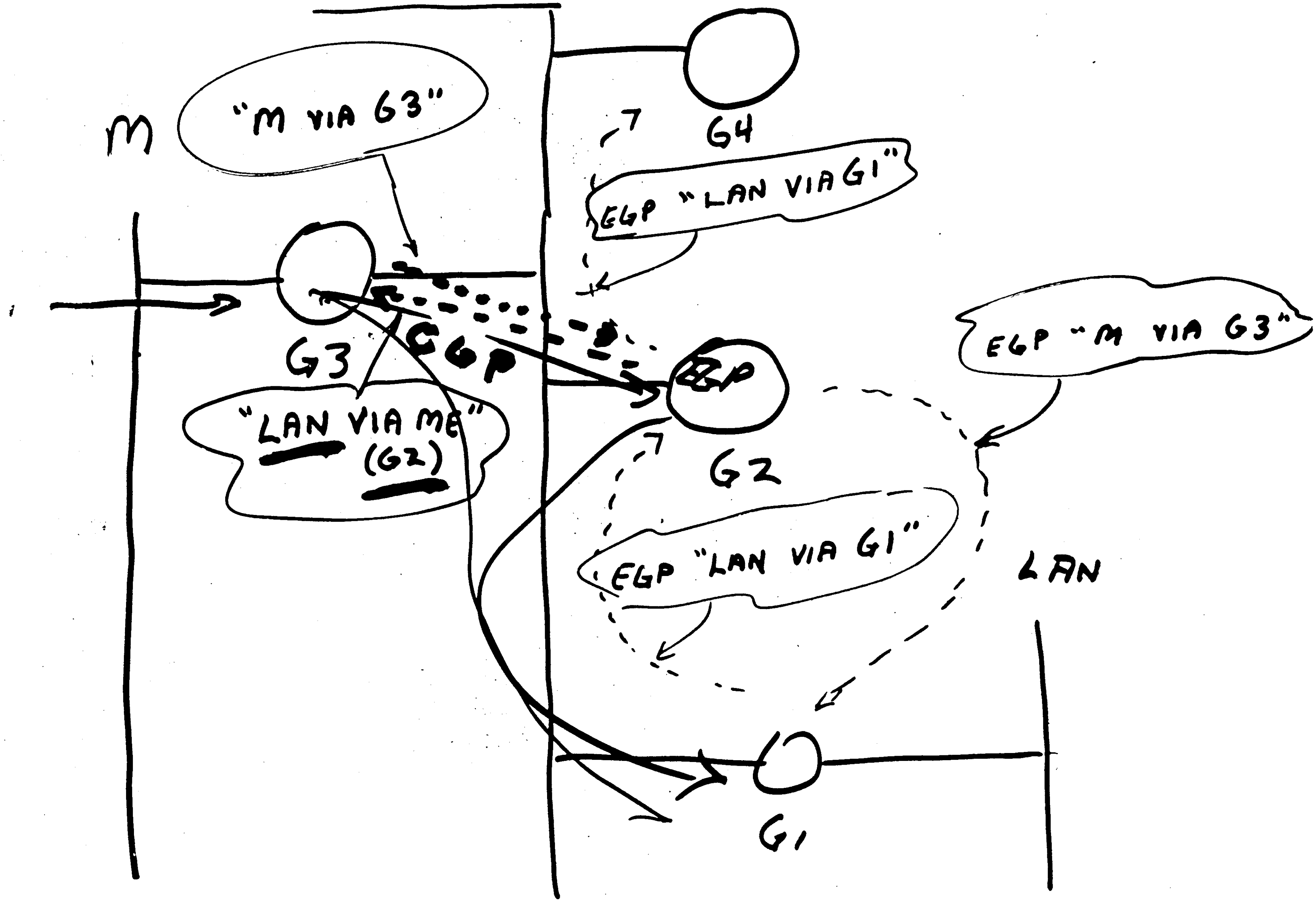
Foster Parent	Equipment / Location	Date Installed
Steve Atlas, BBN	CPU to BBNNET2-ARPANET-GW	9 Nov 87
Mike Petry, U. of Md.	CPU to PURDUE-CS-GW	18 Nov 87
Louie Mamokus, U. of Md.	CPU to GATEWAY.ISI.EDU	23 Nov 87
Robert Enger, Contel	CPU to DCEC-MILNET-GW	23 Nov 87
	Memory to DCEC-MILNET-GW	23 Nov 87
	Memory to BBN-MILNET-GW	Date - Unknown
	Memory to YUMA-GW.ARPA	
Bill Nesheim, Thinking Machines	CPU to AERONET-GW.ARPA	24 Dec 87
	Memory to AERONET-GW	13 Jan 88
Paul Pomes, U. of Ill.	CPU to YUMA-GW.ARPA	12 Jan 88

TESTING TOPOLOGY



IETF 26 FEB 88
CONTEL: Enger

EXTRA HOP



MIKE BRESCIA, BBN

PRE-ADOPT-A-GATEWAY

Ping Destination	Ping Originator	Last Hop GW to Me	Average Delay (ms)	% Loss
TWG.ARPA	SCCGATE	SRI-MILNET	3173	5
	SEKA	BBNNET2	8042	26
DCEC-MILNET	SCCGATE	DCEC-MILNET	1872	0
	SEKA	BBNNET2	14954	53
A.ISI.EDU	SCCGATE	MILNET-GW.ISI	4343	17
	SEKA	BBNNET2	19509	89

Observations:
First hop from LAN GW (sccgate) is to DCEC-MILNET.
Shows extra hop through EGP server BBNNET2.
Shows queuing in PSN feeding into EGP server (per S. Atlas)

PING TESTS TO ARPANET-EGP SERVERS

SCCGATE 10.11.0.20 Sun-3/260	BBNNET2 10.7.0.63 11/73	PURDUE-CS 10.2.0.37 11/73	GATEWAY.ISI 10.3.0.27 11/23	D.ISI.EDU 10.0.0.27 DEC2060
11/19/87	300/ 887 /4509	200/ 754 /4540	380/ 34401 /156300	
2:00PM	1	0	35	
11/19/87	320/ 2008 /14300	180/ 1938 /12100	39400/ 94287 /149180	
2:15PM	14	9	76	
11/19/87	320/ 1444 /6880	200/ 628 /2240	14680/ 30677 /56106	340/ 450 /1020
2:25PM	5	0	44	0

Observations: Low delay times from D.ISI show subnet not at fault.

IETF 26 FEB 88
CONTEL: Enger

MORE PING TESTS TO ARPANET-EGP SERVERS

SCCGATE 10.11.0.20 Sun-3/260	BBNNET2 10.7.0.63 11/73	PURDUE-CS 10.2.0.37 11/73	GATEWAY.ISI 10.3.0.27 11/73
2/26/88	360/ 721 /1700	220/ 630 /1860	360/ 723 /2220
12:30PM	0	0	0
2/26/88	340/ 712 /2620	240/ 603 /2960	360/ 757 /900
1:30PM	0	0	0
2/26/88	360/ 744 /2140	239/ 738 /3580	380/ 893 /2500
2:20PM	0	0	0

Observations: 11/73 CPU present in GATEWAY.ISI.
New End-to-End software running in subnet.
DMA limit raised to 8 per interface (per Steve Atlas).

PING TESTS TO MILNET EGP SERVERS

From TWG. 26.5.0.73	YUMA-GW 26.3.0.75 CPU=11/73	AERONET 26.1.0.65 CPU=11/73	BBN-MINET 26.1.0.40 CPU=11/23
1/13 /88 6:45PM	470/ 828 /4140 0	440/ 799 /9520 0	750/ 3194 /8650 0
1/13/88 10:40PM	460/ 666 /1920 1	440/ 539 /1290 0	750/ 2732 /8100 0
2/24/88 6:40PM	470/ 1026 /3860 0	440/ 689 /1830 0	740/ 6121 /19470 3
2/24/88 7:00PM	470/ 926 /2900 0	440/ 792 /3110 0	710/ 2784 /7820 0

MORE PING TESTS TO MILNET EGP SERVERS

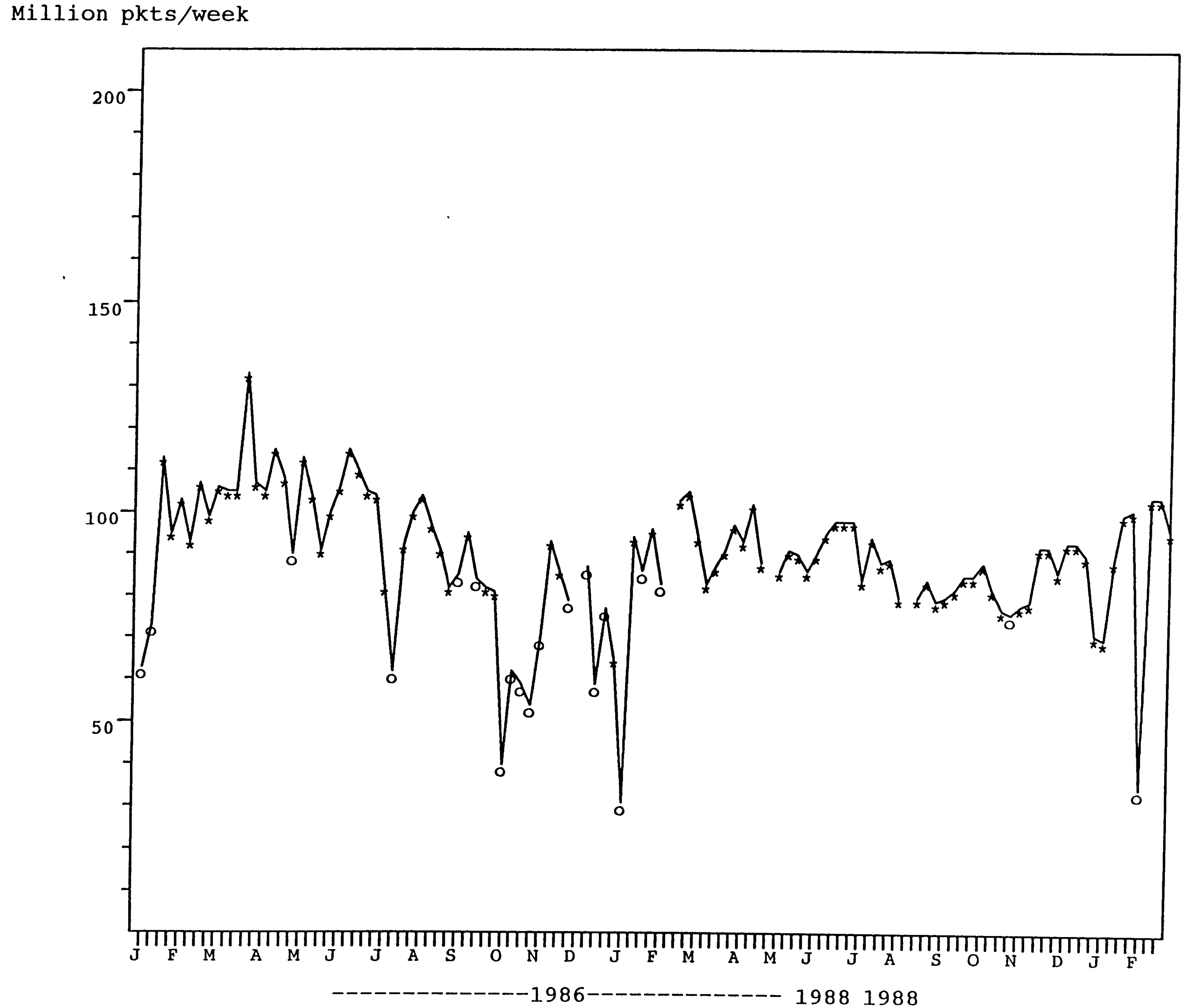
TWG. 26.5.0.73	YUMA-GW 26.3.0.75 CPU=11/73	AERONET 26.1.0.65 CPU=11/73	BBN-MINET 26.1.0.40 CPU=11/23	BBN-MIL-TAC 26.0.0.40
2/26/88	480/ 1262 /5160	440/ 658 /2920	1250/ 8767 /18430	
12:50PM	0	0	2	
2/26/88	480/ 1319 /4080	440/ 1309 /4460	1110/ 4993 /12140	
2:20PM	1	0	2	
2/26/88	480/ 1605 /7860	440/ 846 /2420	820/ 14710 /34310	770/ 935 /1330
3:40PM	3	0	2	0
2/26/88	480/ 1145 /5760	440/ 1153 /4470	770/ 3520 /12960	760/ 904 /1580
4:30PM	0	2	0	0

Trivia Quiz

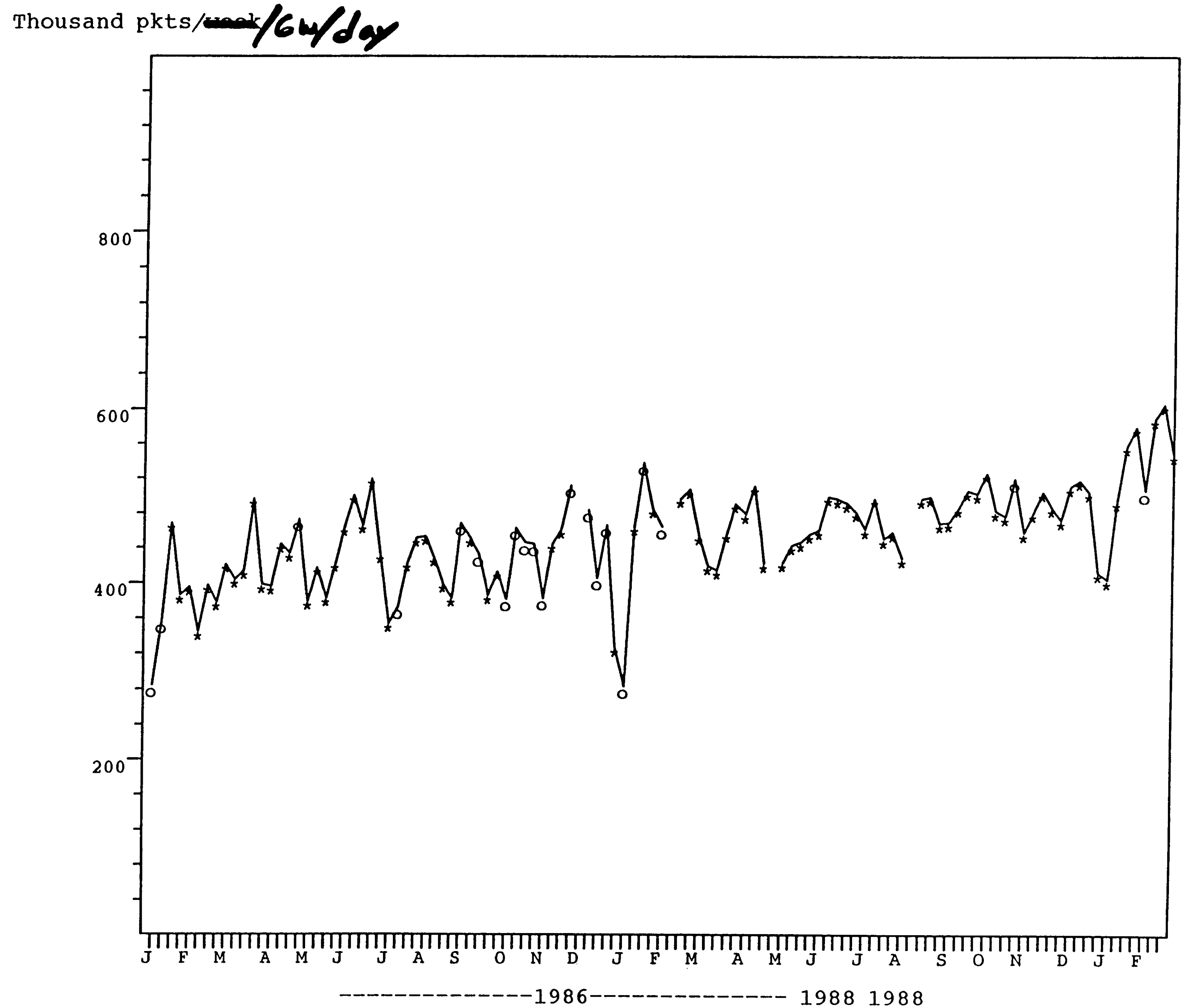
Where did our butterfly slide come from?

6.4 Status of the Adopt-A-GW Program (Cont.)—Phill Gross, MITRE

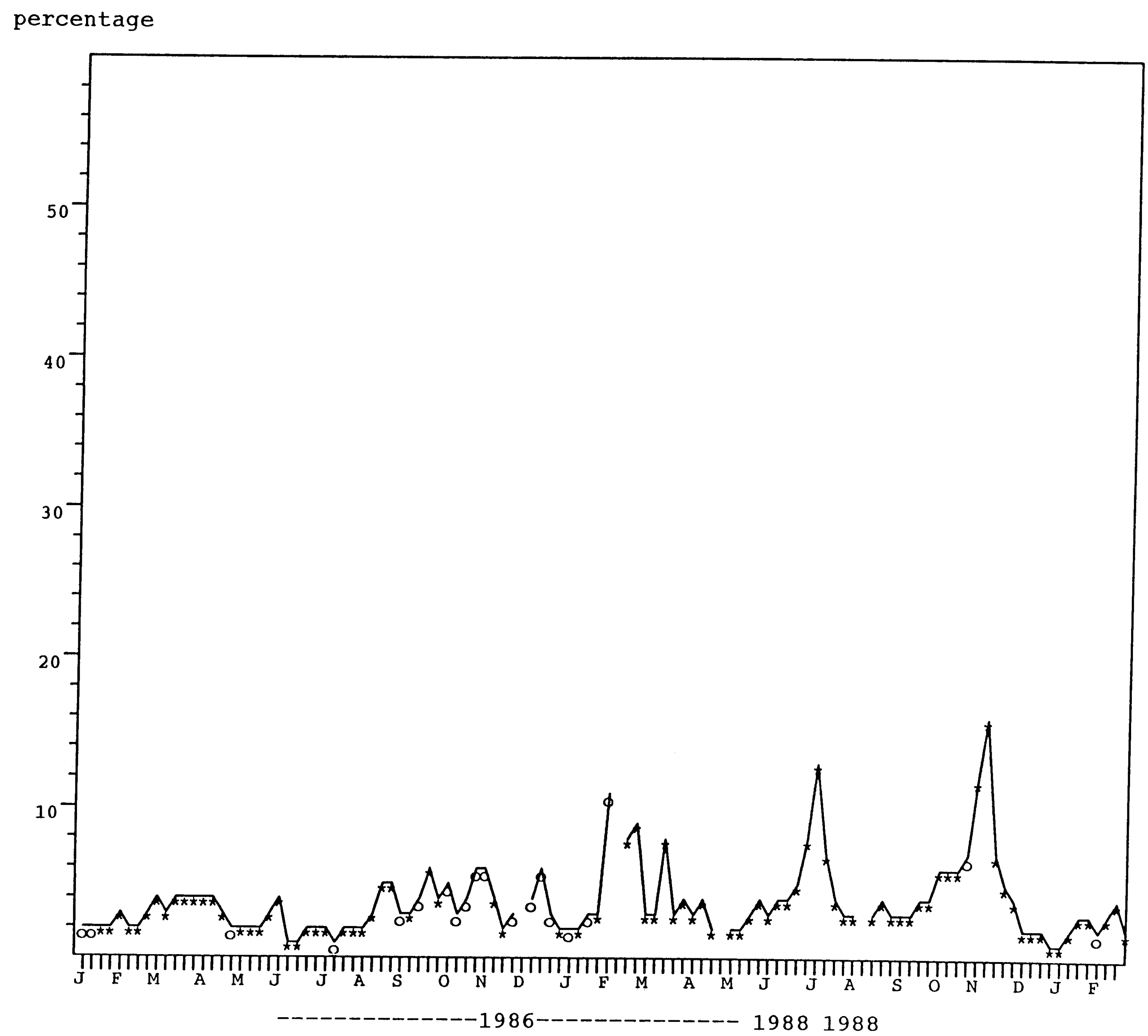
Traffic Sent by Core Gateways



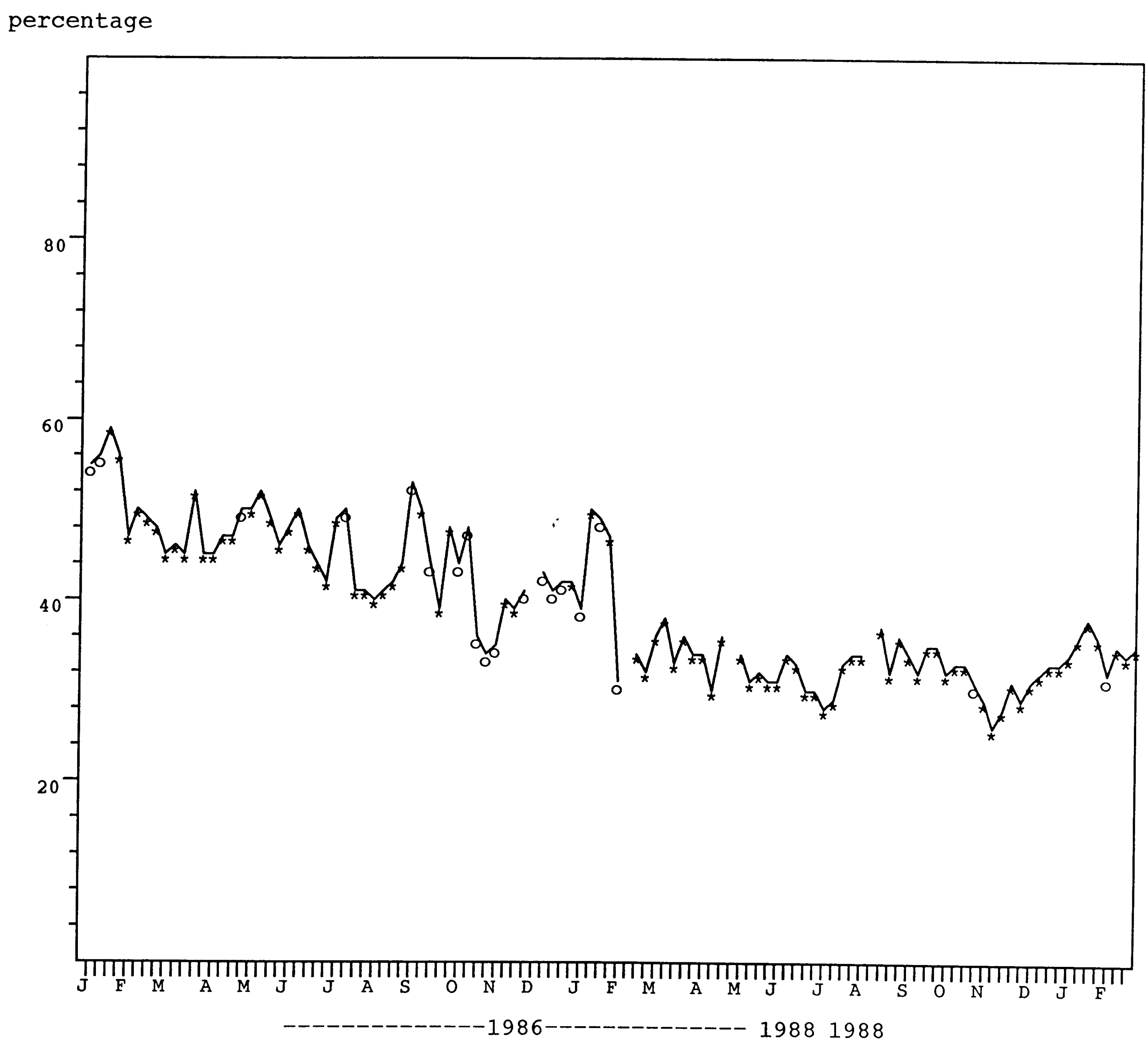
Traffic Sent by Core Gateways



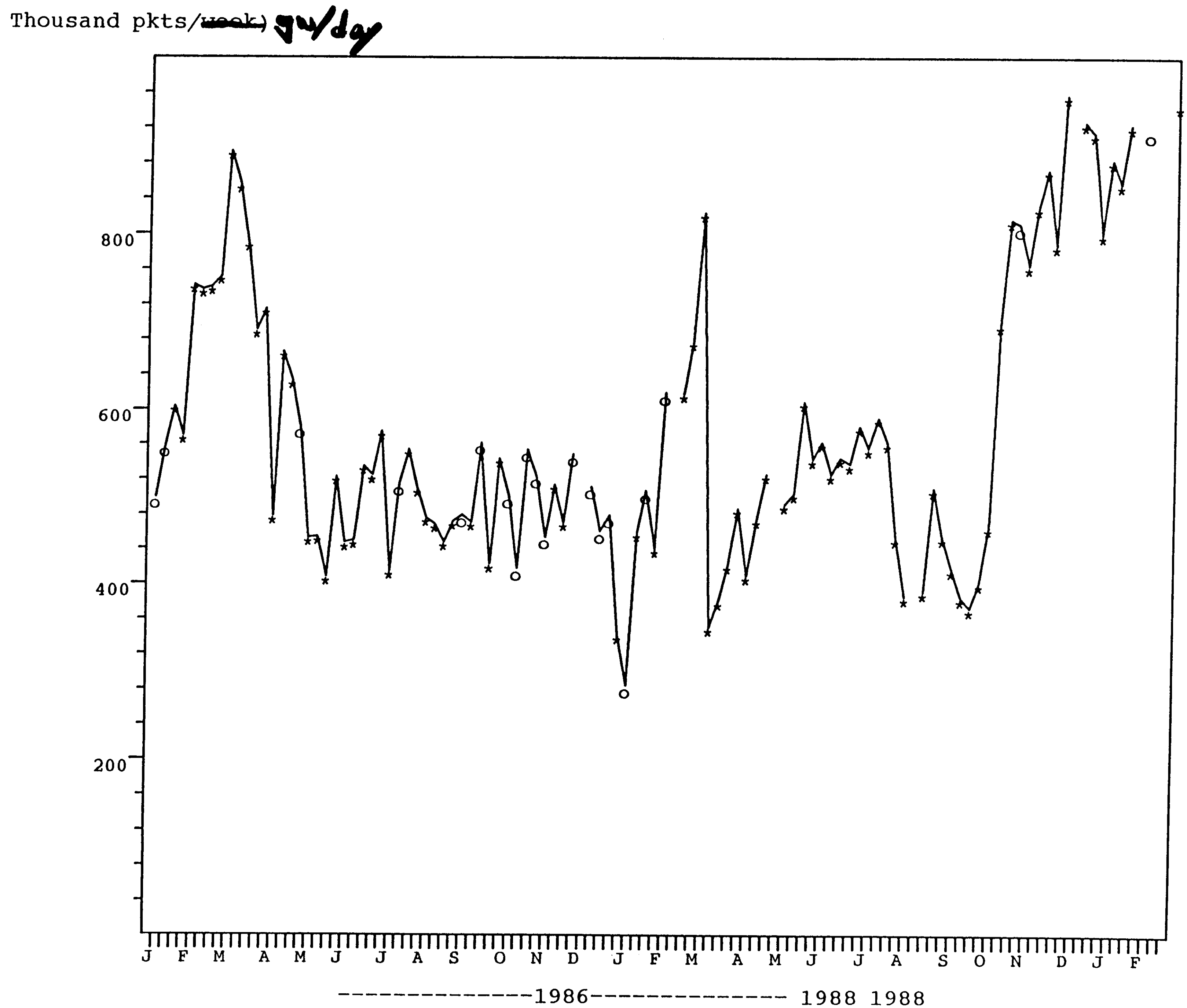
Sent Traffic Dropped by Core Gateways



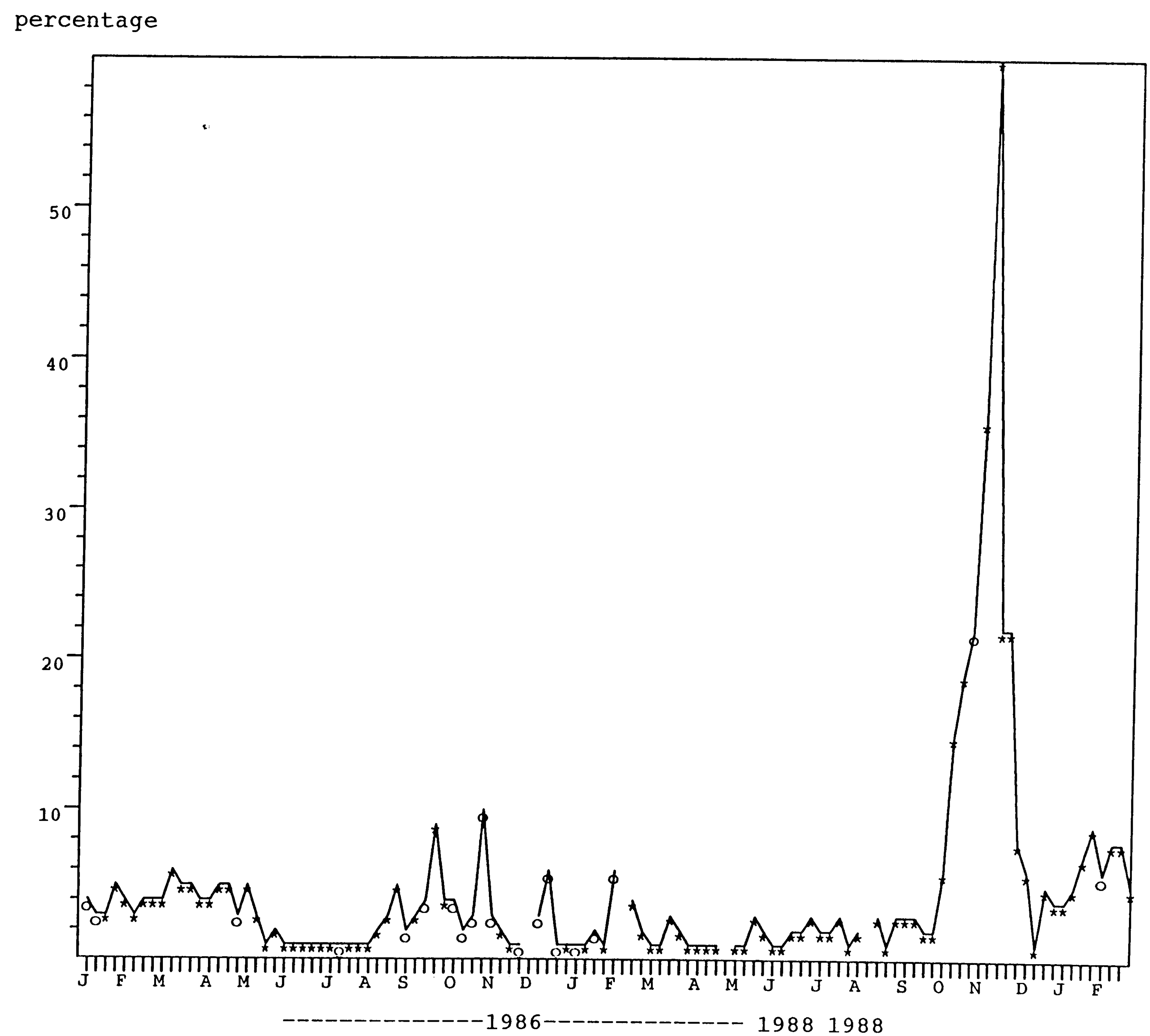
User Data in Gateway Received Traffic



Traffic Sent by Mail Bridges



Sent Traffic Dropped by Mail Bridges



6.5 BBN Report—Mike Brescia, BBN

PROGRESSION OF GATEWAYS

- 1978 Digital PDP 11/40 128K Memory

  SATNET, ARPANET, and Packet-Radio
  6 Packets Per Second

- LSI-11 56K Memory

  Faster OS
  50 Packets Per Second
  20 Buffers

- LSI-11 128K - 256K Memory

  200 Buffers

ADD'L INTERFACES
ETHERNET
RING
X.25
HDH
FIBERNET
WIDEBAND SAT.

- Butterfly First Fielded October 1985

  Expanded Buffering, Functionality,
  Shortest Path (SPF) Routing

INSTALLED GATEWAYS

- ~~20~~ ~~27~~ 26 Butterfly Gateways

  Interface to ARPA, ETHER, SATNET, "Wideband~~s~~"

- ~~39~~ 29 LSI-11 Gateways

  Interface to ARPA, ETHER, Ring, Packet-Radio, X.25

- LSI-11 Traffic ~~180~~ 100 Million Packets Per Week
- "Core" Central Routing for ~~200~~ 400 Nets
- Future Butterfly Installations

  7 Replace LSI-11 between ARPANET and MILNET

  20 Other DARPA Sites

BUTTERFLY GATEWAY TECHNICAL DESCRIPTION

- Hardware - Butterfly

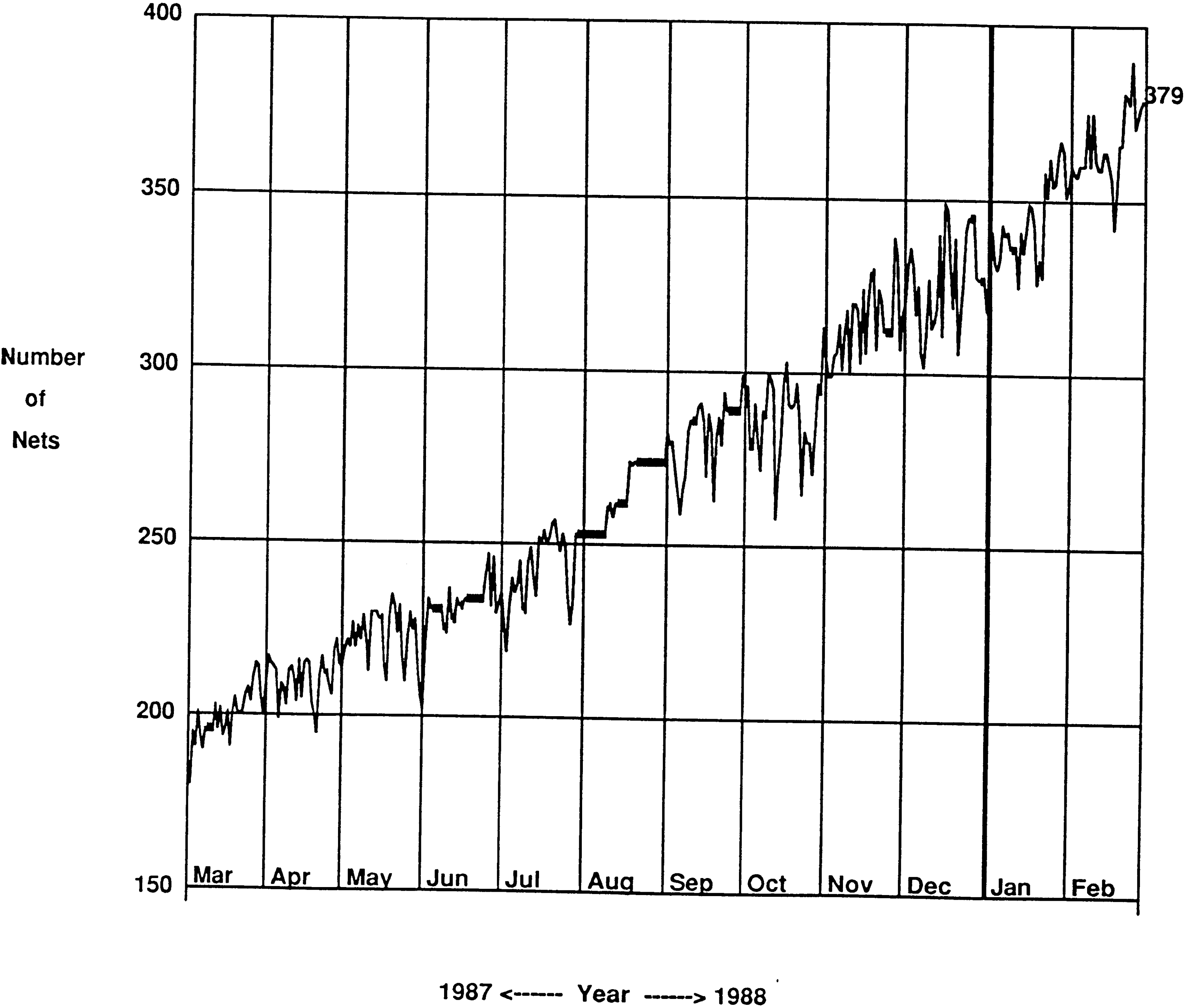
  Multiprocesor, 1 Mbyte Per Processor (68000)
  Fully Interconnected
  Gateway has 2 to 5 Processors

  IO — MULTIBUS - ETHER, ARPA 1822, 100 KB HDLC (others); SPECIAL - 2 MB HDLC

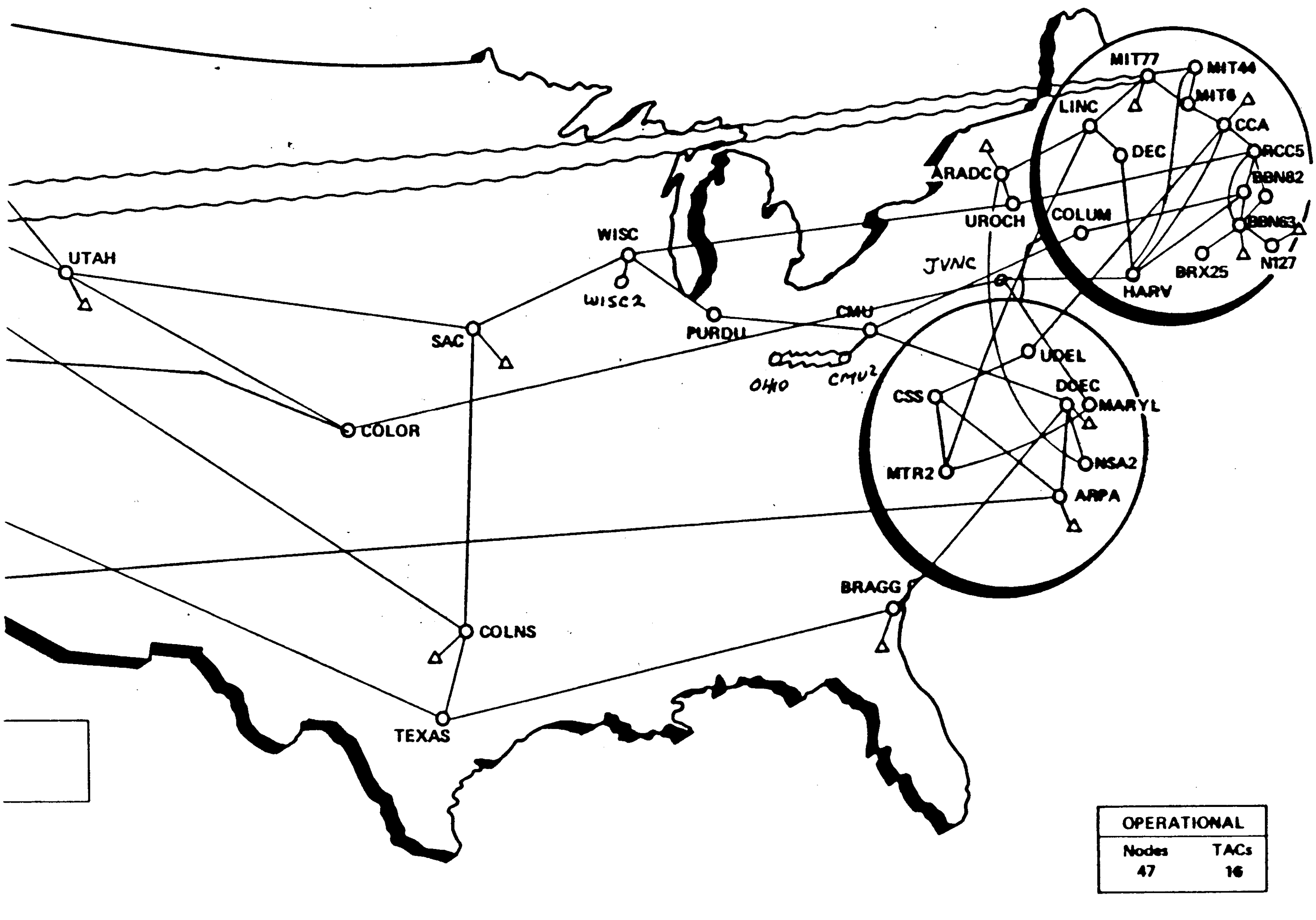
- Software

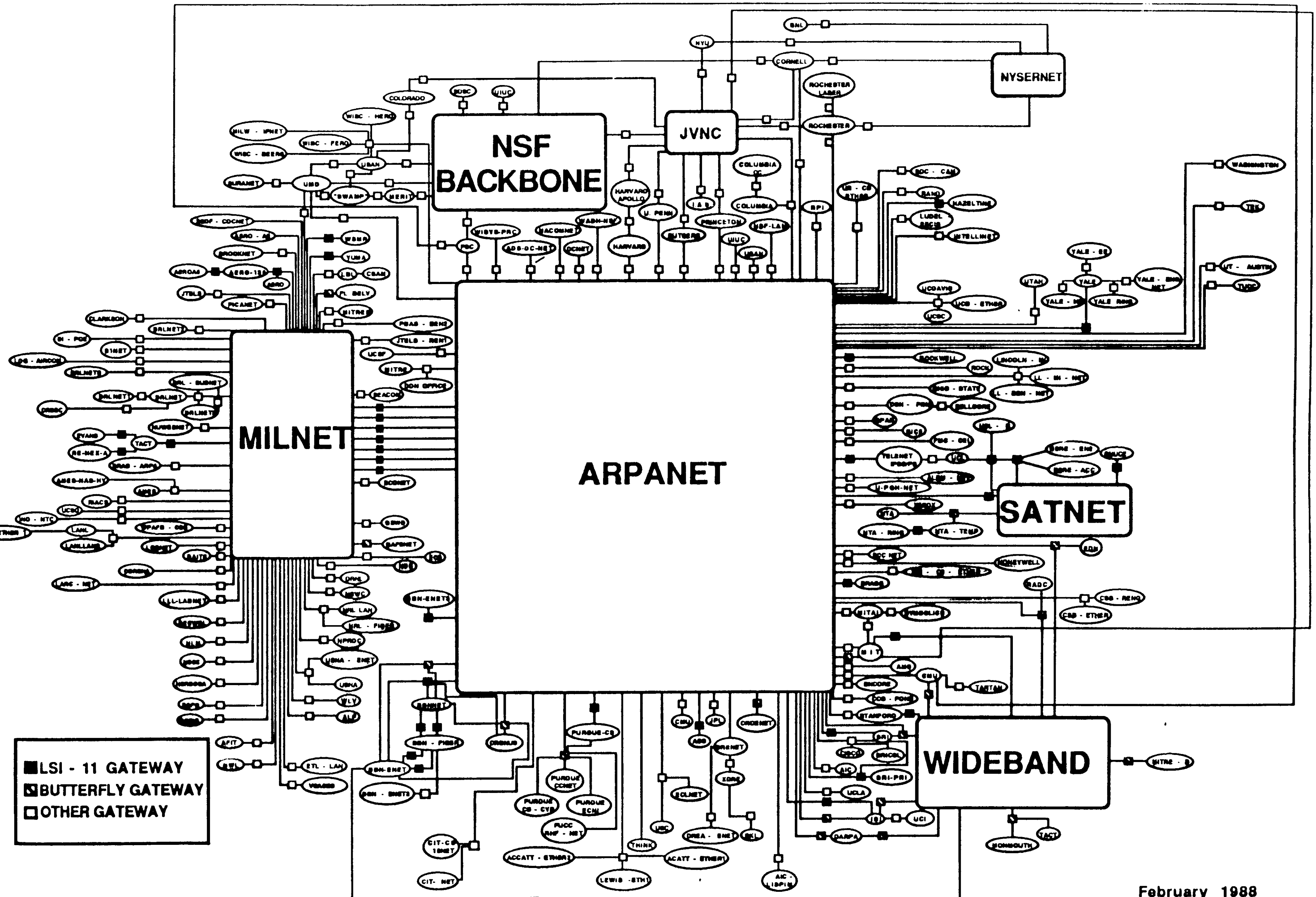
  Message Passing Between Net Layer Interfaces
  and Central Processes (Router, Database Distributor)

Internet Growth in Networks

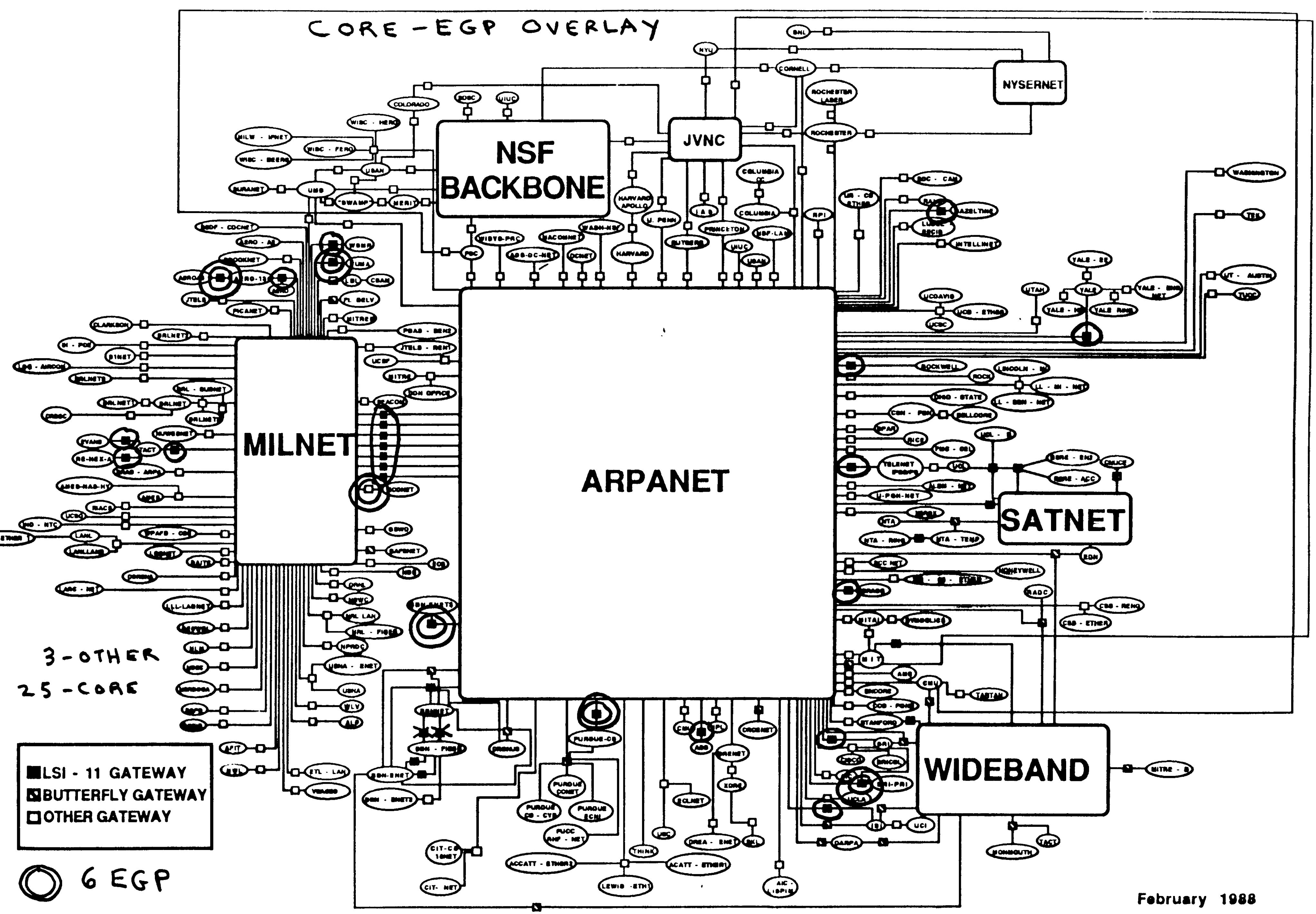


PANET Geographic Map, 31 January 1988 / 29 Feb.

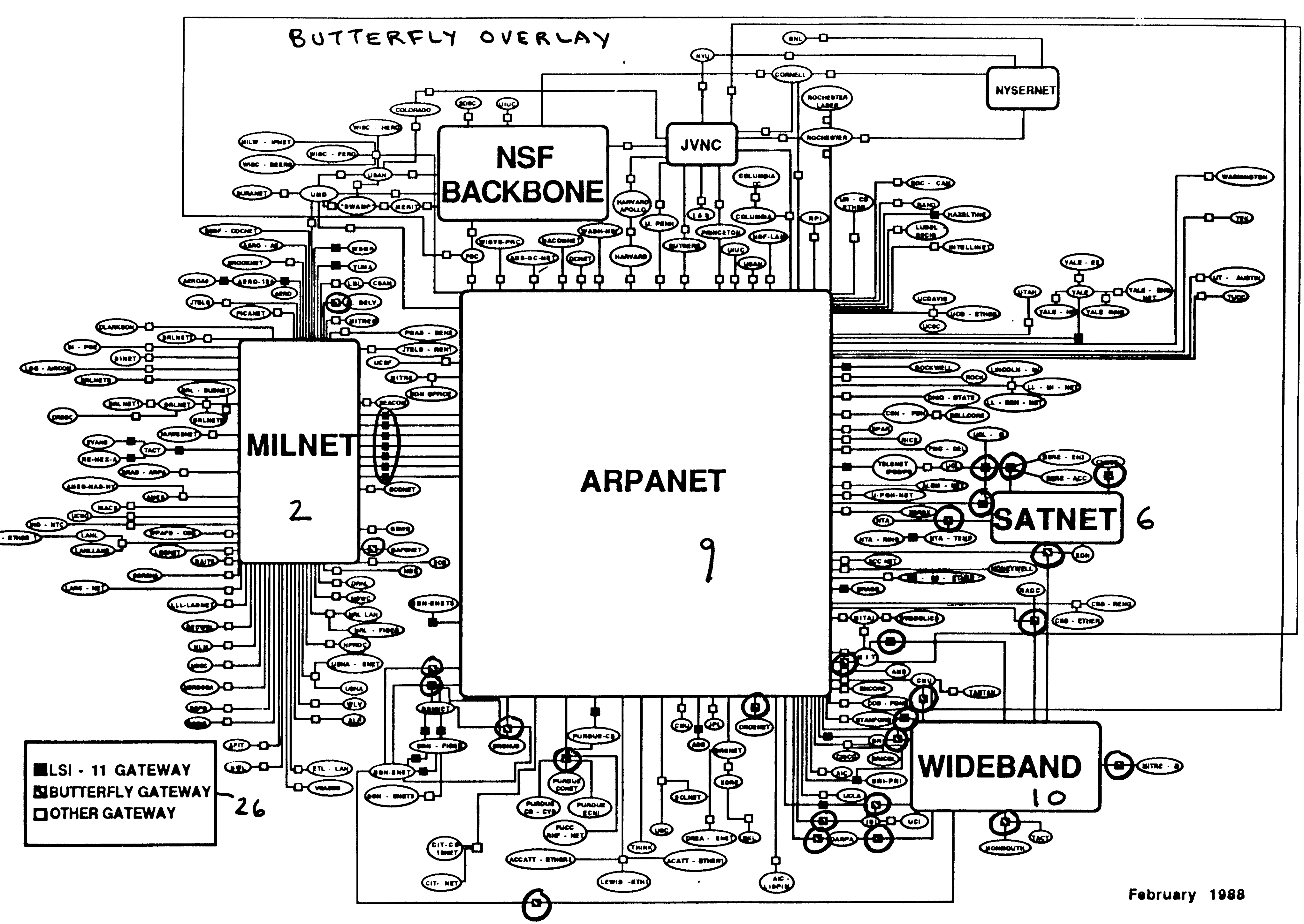




February 1988



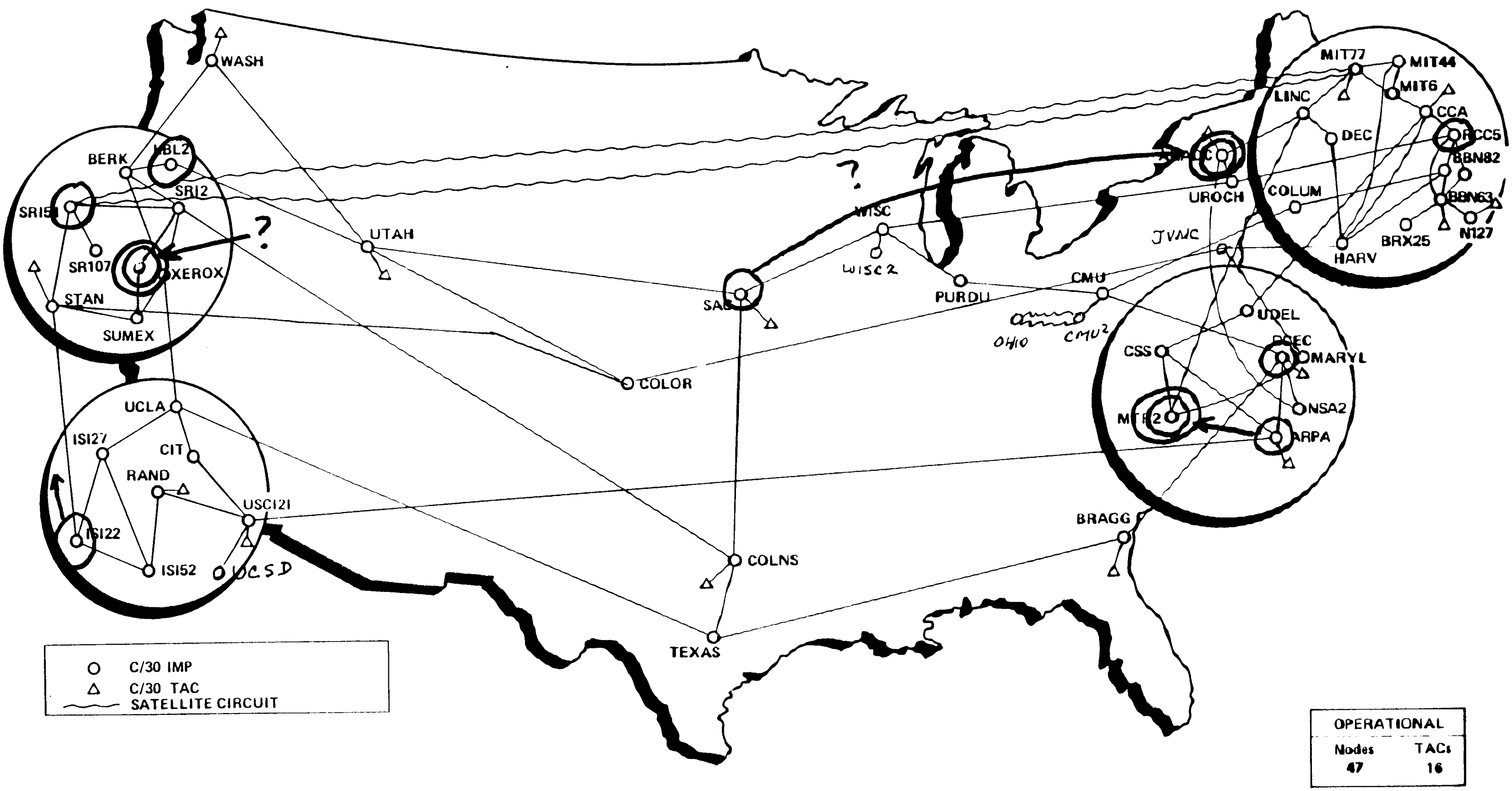
CORE-EGP OVERLAY
NSF BACKBONE
JVNC
NYSERNET
ARPANET
MILNET
SATNET
WIDEBAND
LSI - 11 GATEWAY
BUTTERFLY GATEWAY
OTHER GATEWAY
6 EGP
3-OTHER
25-CORE
February 1988


BUTTERFLY OVERLAY
NSF BACKBONE
JVNC
NYSERNET
ARPANET
9
MILNET
2
SATNET
6
WIDEBAND
10
LSI - 11 GATEWAY
BUTTERFLY GATEWAY
OTHER GATEWAY
26
February 1988

MAIL-BRIDGE OVERLAY

ARPANET Geographic Map, 31 January 1988

/ 29 Feb.



"MAILBRIDGES"

7

GATEWAY ROUTING

- Gateway Based Routing
- Access/Departure Model
- Shortest Path First Routing Algorithm
- Exterior Gateway Protocol

SHORTEST PATH FIRST ROUTING ALGORITHM

- Based on DDN SPF
  - Operational for over 8 Years
- Replaces Gateway to Gateway Protocol (GGP)
- Improvements
  - Smaller Routing Updates
  - Complete Routing Database
  - Extensible to Multiple Metrics
  - Cost Based Routing Metric
  - [redacted]

CORE PROBLEMS

- DISTANCE VECTOR ROUTING (GGP) 1000 BYTE UPDATE FOR 400 NETS
- EGP INFORMATION LOST IN GGP (EXTRA HOP)
- LSI 11 MEMORY RESTRICTION
  - " SINGLE PROCESSOR
  - " DEVELOPMENT TOOLS
  - " ASSEMBLY LANGUAGE

BUTTERFLY REPLACEMENT

- LINK STATE ROUTING - SMALL UPDATE
- EGP INFO CARRIED IN FULL NO EXTRA HOP
- 68000 (...20 ...) UNIFORM MEMORY ADDRESSING
  - " MULTIPROCESSOR - EXPANSION
  - " DEVELOPMENT TOOLS
  - " C LANGUAGE

NOW: ARPANET

ARPANET		
	PURDUE	10.2.0.37
	ISI	10.3.0.27
	BBN	10.7.0.63
MILNET		
	AEROSPACE	26.1.0.65
	BBN	26.1.0.40
	YUMA	

"SOON": THE ARPANET-MILNET GATEWAYS (MAILBRIDGES) WHEN BUTTERFLY REPLACES LSI-11

WATCH FOR ANNOUNCEMENT IN "EGP-PEOPLE@BBN.COM"

BUTTERFLY CORE CONVERSION

1- INSTALL BUTTERFLY GW (SOME) IN PARALLEL WITH LSI-11 MAILBRIDGES
   - HOSTS BEGIN TESTS
   - GATEWAYS ~~TEST~~ WITH EGP

2 - LSI-11 MAILBRIDGES REDIRECT TRAFFIC TO BUTTERFLIES

3 - ANNOUNCE CUTOVER TO NEW MAILBRIDGES "LOAD SHARING" FOR HOSTS, AS IN DDN-MANAGEMENT-BULLETIN-33

4 - ~~PLUG~~ REPLACE REST OF MAIL BRIDGES

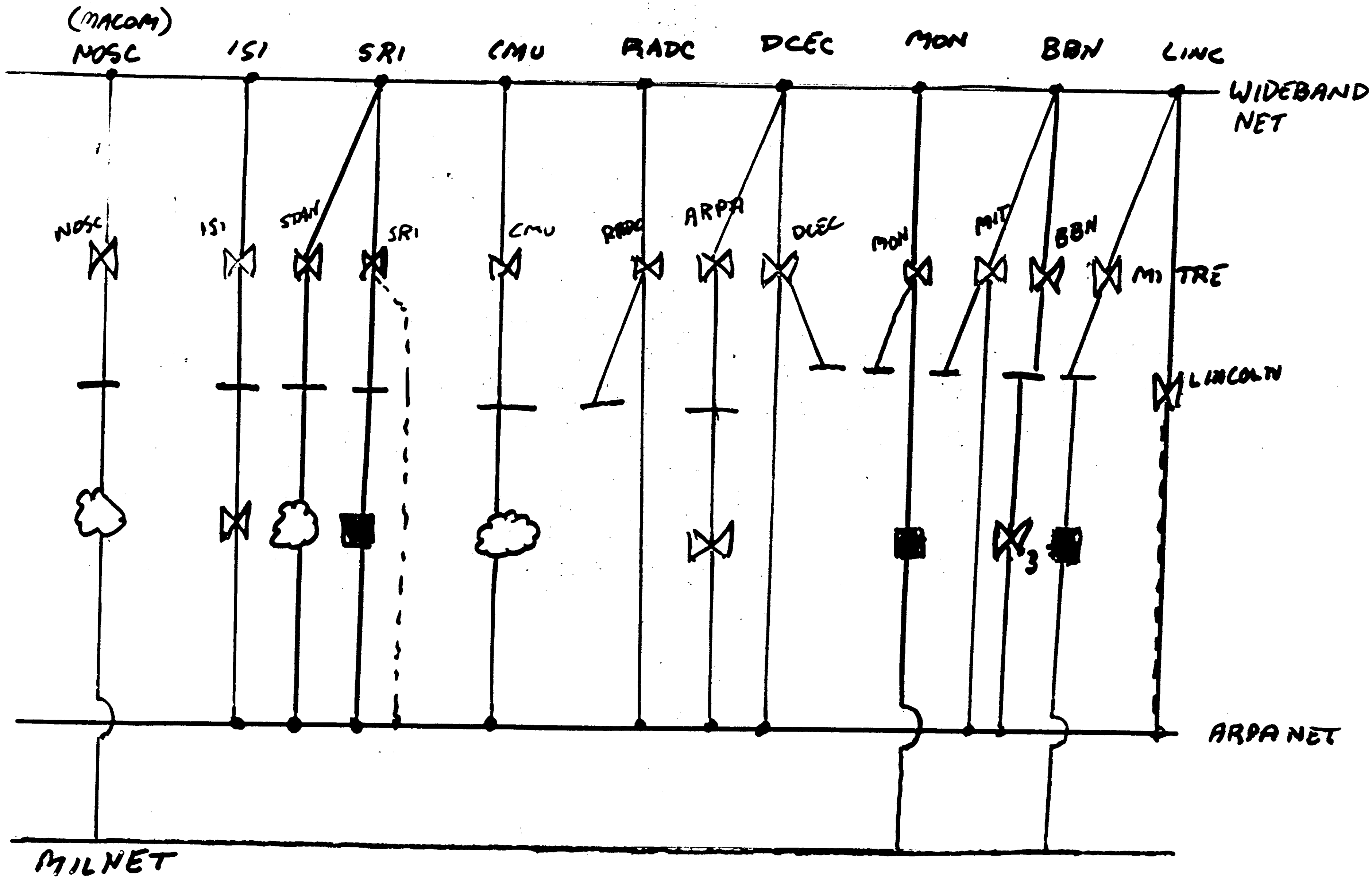
5 - ANNOUNCE CUTOVER TO NEW EGP SERVERS VIA "EGP-PEOPLE" MAILING LIST

6. REMOVE PARALLEL LSI-11 MAILBRIDGES

   REMOVE LSI-11 EGP SERVERS

? - OTHER LSI-11 GATEWAYS RUN EGP TO NEW CORE

WIDEBAND GATEWAYS AND CONNECTED NETS



6.6 BBN Report (Cont.)—Marianne (Gardner) Lepp, BBN

PSN7 Status Report

NEW END-END PROTOCOL

- virtual circuit protocol between source & destination PSNs
- connection setup/tear down
- reassembly
- sequencing
- retransmission

Key features of new E-E

- tailored to X.25
- more efficient ack. policy
- no resource reservations
- multiple connections per host/destination pair

Acknowledgement Policy

- IACK
- EACK
- NETAA

Statistics

	Rate (per second)		
Ack type	peak hr	typical hr	typ. da
IAcks piggybacked	49.57	57.06	30.67
EAcks piggybacked	188.46	173.11	128.6
non-NETAA packets	613.96	653.01	447.5
IAcks on NETAAs	12.58	12.15	10.12
EAcks on NETAAs	101.60	93.49	86.6
NETAA msgs	81.73	75.02	71.6

source buffering:

Oldee
- 85% ARPANET traffic is single packet messages
- almost all single packet messages obtained resources w/out delay
- 15% traffic is multi-packet
- 38% of this required a REQ8/ALL
- host blocked during REQ8/ALL8 exchange

Newee
- <1 in 2500 messages retransmitted over busiest hou

Performance

Protocol	msg/sec from hosts	pkts/sec	NETRAN/ REMOTE	Trunk usage Kb/s
– Peak hours –				
newee	432.3	602.3	81.7	1860.8
oldee	357.2	567.7	357.2	1867.3
– Typical hours –				
newee	412.9	615.2	75.0	2098
oldee	314.0	498.3	314.0	1977
– Typical days –				
newee	309.2	443.9	71.6	1456.0
oldee	276.7	432.9	276.7	1581.4

6.7 Domain Working Group—Mark Lottor, SRI-NIC

DDN Growth

Network Naming and Addressing Statistics

	Feb 1987	Feb 1988
Internet Hosts (includes ARPANET/MILNET)	3,807	5,392
ARPANET/MILNET Hosts	668	1514
ARPANET/MILNET TACs	139	170
ARPANET/MILNET GWs	130	168
Internet Gateways	170	224
ARPANET/MILNET Nodes	209	245
Connected Networks	568	824
Domains (top-level, 2nd-level)	269	485
Hostmaster online mail	1064	1394

(Size of current host table = 579,780 bytes)

Domains and Hosts
Registered with DDN NIC
27 Feb 88

Top-level domains	=	32
2nd-level domains	=	452
Hosts in.COM	=	411
Hosts in .EDU	=	2461
Hosts in .GOV	=	186
Hosts in .IL	=	1
Hosts in .MIL	=	141
Hosts in .NET	=	17
Hosts in .ORG	=	21
Hosts in .UK	=	9
Hosts still in .ARPA	=	2538

146 (net 10)

1500 (net 26)

892 (other nets)

6.8 EGP3 Working Group—Marianne (Gardner) Lepp, BBN

EGP3 is here

features:

- version negotiation
- incremental routing update
- no explicit Hello/IHUs
  replaced by Polls/update
- Polls contain routing data
- EGP2 – new error msg
  Reason 6
- metric – { hop count, reach, explicit down }

non features:

- Does not turn EGP into a routing protocol

IDEA 9

6.9 Open Systems Internet Operations Center WG—Jeff Case, UTK

INOC Working Group

Chair : Jeff Case, UTK.

2nd Meeting

AIMS

1. Define model for combining elements of network monitoring and control into a total system:

   a. define the roles of a INOC:

      i. point of controlled access to information, including protecting monitored entities from excessive/redundant requests.

      ii. provides proxy services for non-IP entities.

      iii. provides appropriate level of security for data integrity & authorization of access.

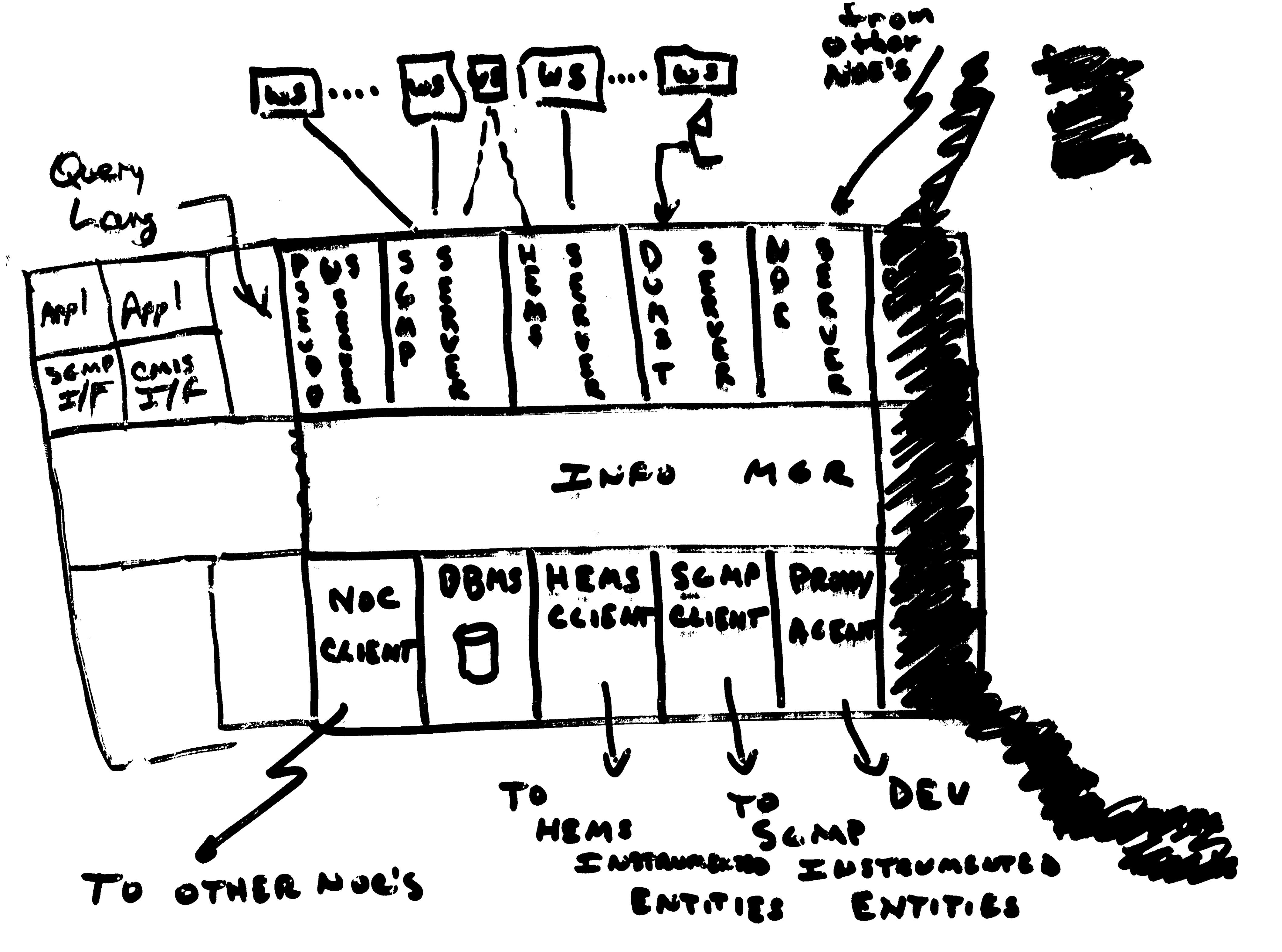
   b. provide mechanism for exchange of information across administrative domains.

AIMS (contd)

2. Database
   a. define needs
   b. mechanisms for information storage & retrieval

3. Information required to do network management
   a. MIB
   b. input from performance/congestion-control needs.

4. Define application needs
   - real-time status monitoring
   - fire-fighting
   - report generation
   - standard application interface



WS
WS
WS
WS
WS
from other NOC's
Query Lang
Appl
Appl
SGMP I/F
CMIS I/F
PSEUDO WS SERVER
SGMP SERVER
HEMS SERVER
DUMB T SERVER
NOC SERVER
INFO MGR
NOC CLIENT
DBMS
HEMS CLIENT
SGMP CLIENT
PROXY AGENT
TO OTHER NOC's
TO HEMS INSTRUMENTED ENTITIES
TO SGMP INSTRUMENTED ENTITIES
DEV

6.10 Authentication WG—Marty Schoffstall, RPI

Authentication

Requirements (Again):

1) short term or interim (time to field)
2) Scope of application (network management + EGP3?)
3) support private + public KDC's
4) no "strategic" applications
5) exportable
6) support gourantee of information
   optionally no read of info

Two teams assigned:

1) MIT / Kerberos
   St John's / "anything else"
2) Presentation in one month
   Albany NY
3) study to throw darts
4) Kerberos DOC issued as IDEA
5) decision to prototype in one month

6.11 Performance/Congestion Control—Coleman Blake, MITRE

OUTLINE CC SHORT-TERM DOCUMENT

I. INTRO

KKR — A. WHAT **IS** IMPROVED PERFORMANCE?
Ramakrishnan, DEC

B. CURRENT FIGURES

C. TARGET FIGURES

D. BACKGROUND OF RECS / REFERENCES

II. RECOMMENDED SHORT-TERM FIXES

A. END SYSTEMS

CB
Blake, MITRE 1. RTT ESTIMATIONS, RTO CALC.

PK
Karn, Bellcore 2. SMALL PKT AVOIDANCE, NAGLE ALGORITHM, NEW BIG WINDOW PROBLEM

Mankin, MITRE
CB, AM, BS
Schofield, DCEC 3. XTCP / CUTE AS A WHOLE

B. GATEWAYS

KKR 1. RANDOMIZATION

AB
Berggreen, ACC 2. GUIDELINES FOR X.25 VIRTUAL CIRCUIT ALGORITHMS

C. "APPLICATIONS"

PK 1. MESSAGE REDUCTION
- SMTP

DB 2. TELNET DATA ACCUMULATION AND LINE AT A TIME MODE
Borman, Cray

JL 3. DOMAIN CACHING, NEGATIVE CACHING
Larson, Xerox PARC

JLe COORDINATION W/ HOST REQUIREMENTS DOCUMENT
Lekashman, PSC

III. FURTHER STUDY

A. SQ
B. DEC CONGESTION AVOIDANCE
C. FAIRNESS IN GATEWAYS
D. RATE-BASED CONTROL OF CONNECTIONLESS TRAFFIC

6.12 OSI Technical Issues WG—Ross Callon, BBN

ADDRESSING FOR THE ISO IP IN THE DoD INTERNET

R. CALLON

3/88

OBSERVATIONS + REQUIREMENTS

- DoD IS MOVING TO ISO/OSI
  - WHAT ADDRESSES SHOULD WE USE?

- OSI ROUTING PROTOCOLS ARE COMING
  - INTRA-DOMAIN (IGP)
  - INTER-DOMAIN (EGP)
  - NOT DONE YET
  - IGP'S MAY REQUIRE SPECIFIC ADDRESS FORMAT

- ADDRESSES ARE HARD TO CHANGE
  - SCHEME SHOULD LAST
  - FLEXIBILITY

GROWTH

- INTERNET IS GROWING RAPIDLY
  - CURRENTLY ≈ 330 NETS (700+ ASSIGNED #s)
  - DOUBLING ≈ EVERY YEAR
  - PROBABLY >10,000 Nets in 5 to 10 YEA
- ROUTING BY NET # WILL BECOME INFEASIBLE
  - AS # &/or AREA NEEDED
  - ALSO USEFUL FOR POLICY ROUTIN
- # OF AS's ALSO GROWING RAPIDLY

IDEA 003 ADDRESSES

A. F. I.	1
I.D.I. = I.C.D.	2
VERSION	1
GLOBAL AREA	2
AUTONOMOUS SYS.	2
DOD IP ADDR.	4
USER PROTOCOL	1
(total	13 octets)

POSSIBLE MIGRATION PLAN

- SHORT TERM: USE "DOD IP ADDR."
- MEDIUM TERM
  - IGP LOOKS AT "DOD IP ADDR"
  - NEW EGP LOOKS AT A.S.#
- LATER
  - NEW EGP LOOKS AT A.S.# & Global Are
  - LOW ORDER PART DEPENDS ON IGP
  - "ONE AS AT A TIME" TRANSITION
- IMPLICATIONS
  - VARIABLE LENGTH ADDRESSES
  - PER-HOST (NSAP) REDIRECTS
  - "ROUTING DOMAIN" MODEL

MIGRATION TO ANSI ROUTING
IN ONE A.S.

A.F.I. I.D.I. = I.C.D. VERSION GLOBAL AREA A.S. #	SAME
LOCAL AREA 6 BYTE ID SEL	ACCORDING TO ANSI

(total 17 octets)

OTHER SUGGESTIONS

- NSAP SELECTOR ≢ USER PROTOCOL

- START WITH 17 OCTET FIELDS
  - 4 BYTE DOD ADDRESS ENCODED IN 6 OCTET ID
  - REST OF ID & LOC-AREA TO ZERO

- START WITH LONGER TERM SOL'N
  - EGP & IGP PARTS SEPARATE
  - IGP PART VARIABLE LENGTH

RECOMMENDED APPROACH

A.F.I.

IDI = ICD

VERSION

GLOBAL AREA

A.S. #

IGP PART (VARIABLE)

SELECTOR (1 OCTET)

SUMMARY

- WE NEED ADDR. FORMAT COMPATIBLE WITH ISO/OSI STANDARD
- USE ICD VALUE ASSIGNED TO DOD
- ALLOW FOR
  - GROWTH
  - COMPATIBLE WITH "FIRST CUT" ROUTIN
  - MIGRATION TO FUTURE ROUTING
- SOLUTION
  - SEEMS LIKE OVERKILL FOR SHORT TE
  - FULFILLS REQUIREMENTS FOR SHORT, MEDIUM, & LONG TERM.

6.13 OSI Technical Issues WG (Cont.)—Rob Hagens, UWisc

Motivation

Emphasize Experimental
Not transition

- Goal: Experiment with ISO lower layers as they progress through the standardization process

- Examples:
  - TP4/CLNP
  - ES-IS
  - IS-IS

- Experimentation includes:
  - Performance tuning
  - Interoperability testing

Requirements

- "Typical" datagram service:
  - possible packet loss
  - duplication
  - corruption
  - re-ordering
  - congestion
  - variable delay
  - etc.

- A complex topology
  - heterogeneous subnetworks
  - multiple paths
  - varying link and media characteristics
  - etc.

- In short, a national CLNP-based Internet

An Observation

- Where have we seen this before?

  ..... the DARPA/NSF Internet

- The Internet meets all of the requirements except one:

  It is IP-based rather than CLNP-based

Possible Solutions

- Implement the OSI CLNL in Internet routers
  - ° disruptive
  - ° requires many routers to change

- Emulate CLNL packets on top of IP packets

  UDP – [illegible] access
  IP – many people have access to IP

  - ° non-disruptive
  - ° utilizes entire topology

EON

- **E**xperimental **O**SI-based **N**etwork

  ° treats the DARPA/NSF Internet as a CLNL subnetwork.

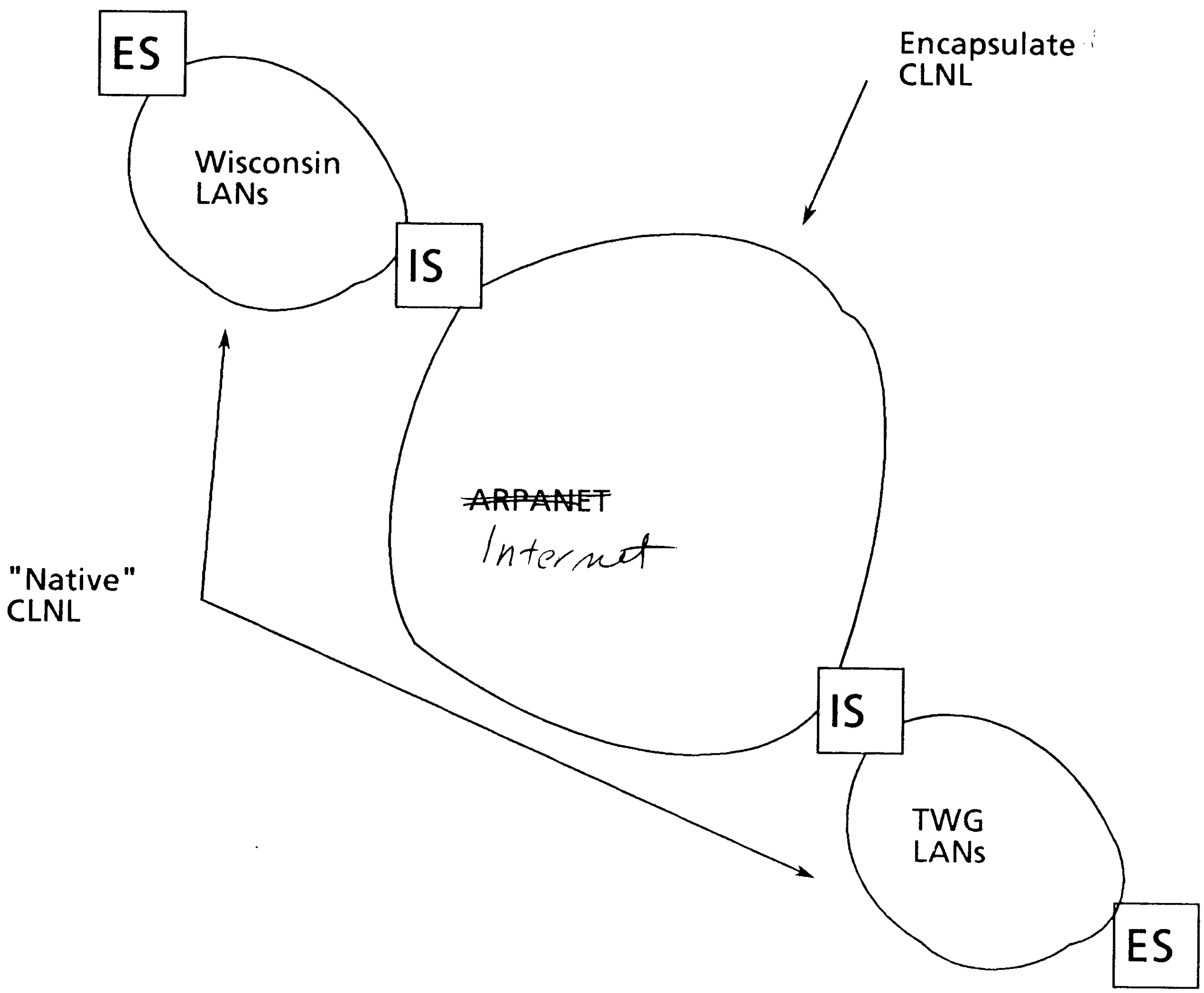
  ° eg: the IP-subnet

- Participating IP-nodes form a logical ISO subnet.

- Several logical ISO subnets may exist in the DARPA/NSF Internet

Example

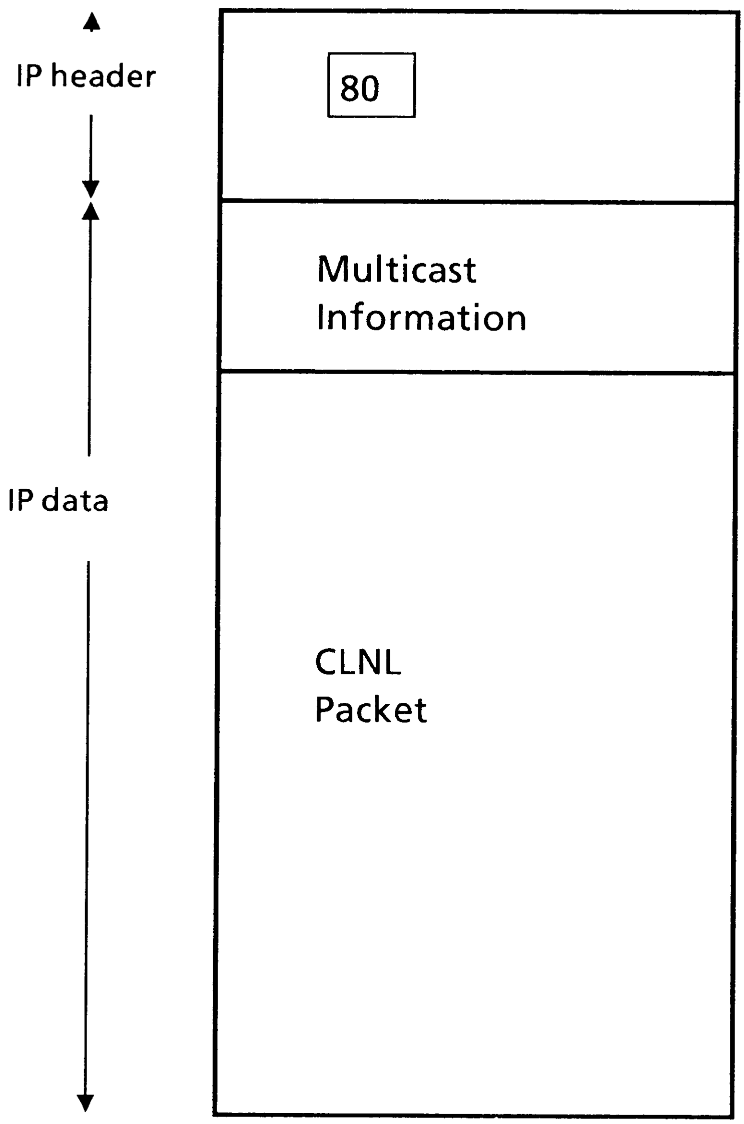
Explain these
ES-IS
IS-IS



EON Defines

- Procedures for encapsulation NPDUs
- NSAP address format
- NSAP address -> SNPA address mapping
- Procedures for wide-area multicasting
- Mechanism for dissemination of topological information

Encapsulation



IP Packet

- Fragmentation
- UDP

Multicasting

- Required by ISO ES-IS, IS-IS
  - "all end systems"
  - "all intermediate systems"
- Realized by sublayer: SNAcP
  - holds table of "core" systems
  - unicast: sends to specified destination
  - multicast: sends to every "core" system
- SNAcP header (4 bytes):
  - version
  - semantics (unicast, multicast, broadcast)
  - checksum

Status

- New. Submitted as RFC, not yet published

- TWG & Wisconsin expect to begin participating as soon as NSAP address format issues are resolved.

6.14 OSI Technical Issues WG (Cont.)—Marshall Rose, TWG

PART V

NETWORK MANAGEMENT WORKING GROUP STATUS REPORT

- NETWORK MANAGEMENT IN AN OSI FRAMEWORK
- SEPTEMBER DEMONSTRATION

NETWORK MANAGEMENT
IN AN OSI FRAMEWORK

- ADOPTED THE ISO

  COMMON MANAGEMENT INFORMATION SERVICE

  AS THE MODEL

- WORK IN TWO AREAS:

  DEFINING THE MANAGEMENT INFORMATION BASE FOR TCP/IP NETWORKS

  MAKING CMIP RUN ON TOP OF TCP/IP

COMMON MANAGEMENT INFORMATION SERVICE

- CONNECTION-ORIENTED USE OF REMOTE OPERATIONS
- CURRENTLY A 2nd DP IN ISO

RFCs IN PREPARATION

- NETWORK MANAGEMENT FOR TCP/IP NETWORKS: AN OVERVIEW
- STRUCTURE AND IDENTIFICATION OF MANAGEMENT INFORMATION FOR THE INTERNET
- LAYER MANAGEMENT INFORMATION FOR: TCP, UDP, IP, MAC802.3
- SYSTEM MANAGEMENT ENTITY: MANAGEMENT INFORMATION
- ISO PRESENTATION SERVICES ON TOP OF TCP/IP-BASED INTERNETS
- TCP/IP NETWORK MANAGEMENT IMPLEMENTORS' AGREEMENTS

  TWO VERSIONS: SHORT-TERM AND LONG-TERM

LIGHTWEIGHT PRESENTATION PROTOCOL

- USED TO PROVIDE THE GLUE BETWEEN THE ISO APPLICATION LAYER AND TCP/IP
- APPEARS TO BE THE ISO PRESENTATION SERVICE, BUT IS REALLY IMPLEMENTED ENTIRELY DIFFERENTLY
- SUPPORTS USE OF EITHER TCP AND UDP DEPENDING ON REQUESTED QUALITY OF SERVICE
- INDEPENDENT OF NETWORK MANAGEMENT

USEFUL FOR ANY SMALL OSI APPLICATION

SEPTEMBER DEMONSTRATION

- MULTI-VENDOR NETWORK MANAGEMENT TO BE DEMONSTRATED AT

  THE 3rd TCP/IP INTEROPERABILITY CONFERENCE

- MANAGEMENT OF SOME OF THE FLOOR TCP/IP NETWORK

- USING THE ISO NETWORK MANAGEMENT FRAMEWORK

PART VI

ISO PRESENTATION SERVICES ON TOP OF TCP/IP-BASED INTERNETS

INTRODUCTION

- [RFC1006]* SPECIFIES A MECHANISM FOR PROVIDING THE ISO TRANSPORT SERVICE ON TOP OF THE TCP

- PERMITS ANY CONNECTION-ORIENTED OSI APPLICATION TO RUN IN A TCP/IP-BASED INTERNET

- SIMPLY IMPLEMENT

  ISO SESSION,

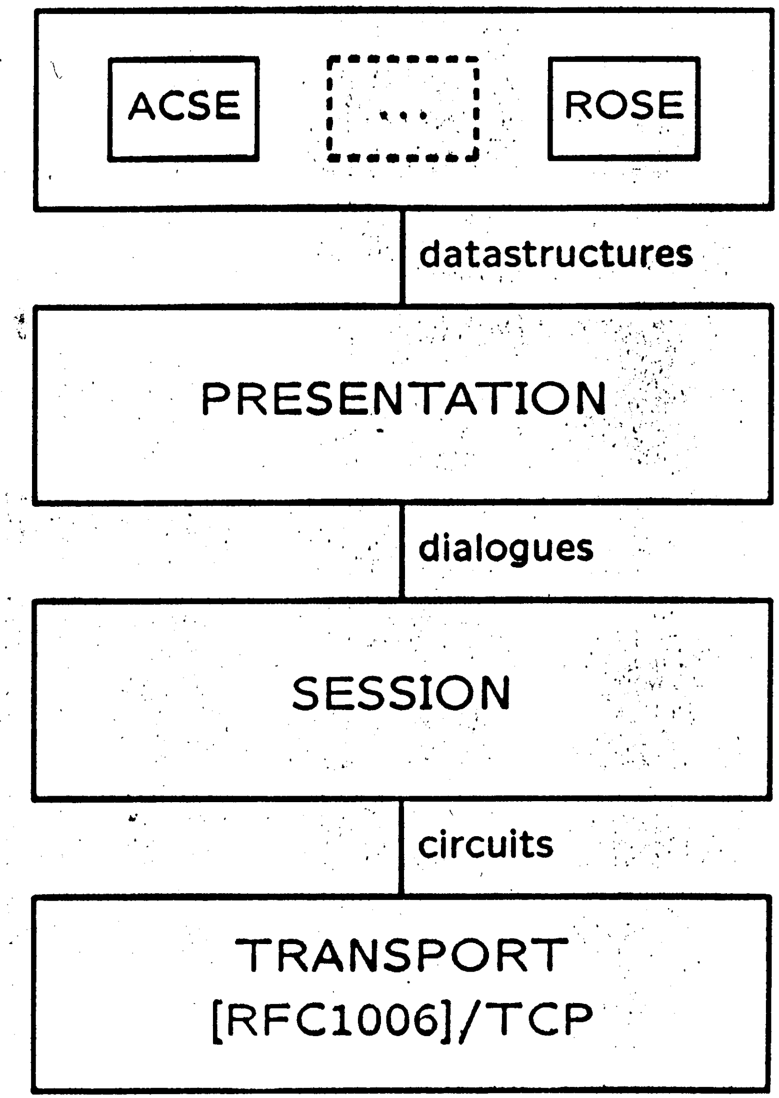
  ISO PRESENTATION, AND

  ISO APPLICATION

  SERVICES ON TOP OF [RFC1006]

*ISO Transport Services on top of the TCP, May 1987

THE UPPER-LAYER ARCHITECTURE



THE PROBLEM

- FOR SOME ENVIRONMENTS, THIS APPROACH IS IMPRACTICAL

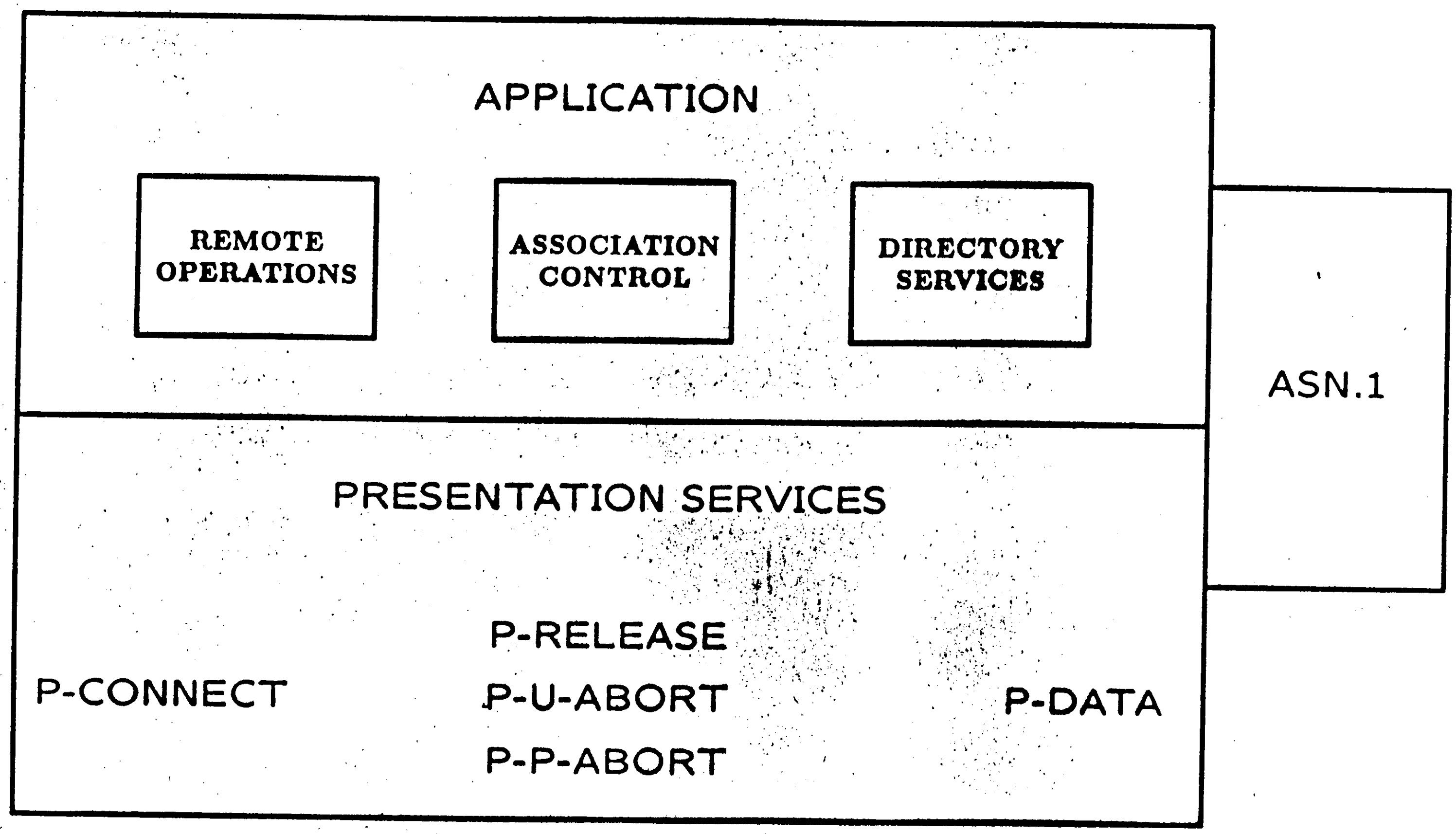
    - TOO MUCH SOFTWARE INFRASTRUCTURE

- WHAT CAN BE DONE TO SUPPORT A LIMITED CLASS OF OSI APPLICATIONS, THOSE WITH "MINIMAL" APPLICATION CONTEXTS

    - ASSOCIATION CONTROL SERVICE ELEMENT (ACSE)

    - REMOTE OPERATIONS SERVICE ELEMENT (ROSE)

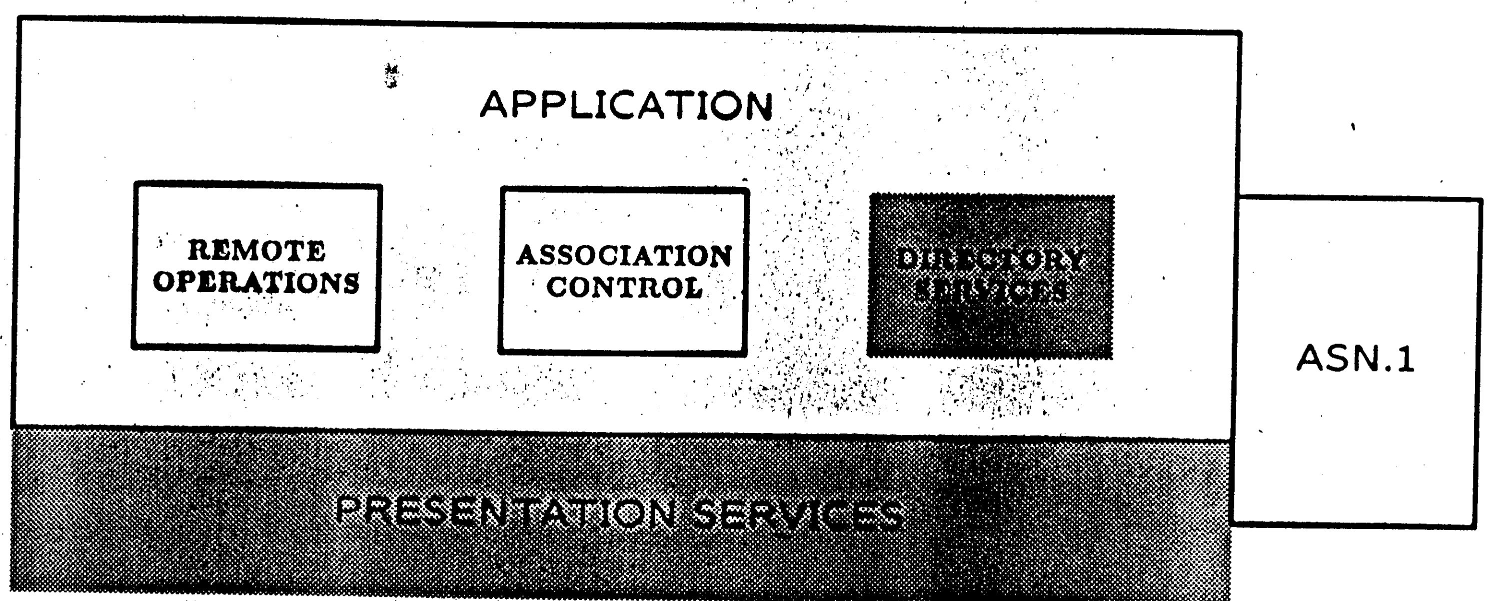
APPLICATION USE OF UPPER-LAYER SERVICES



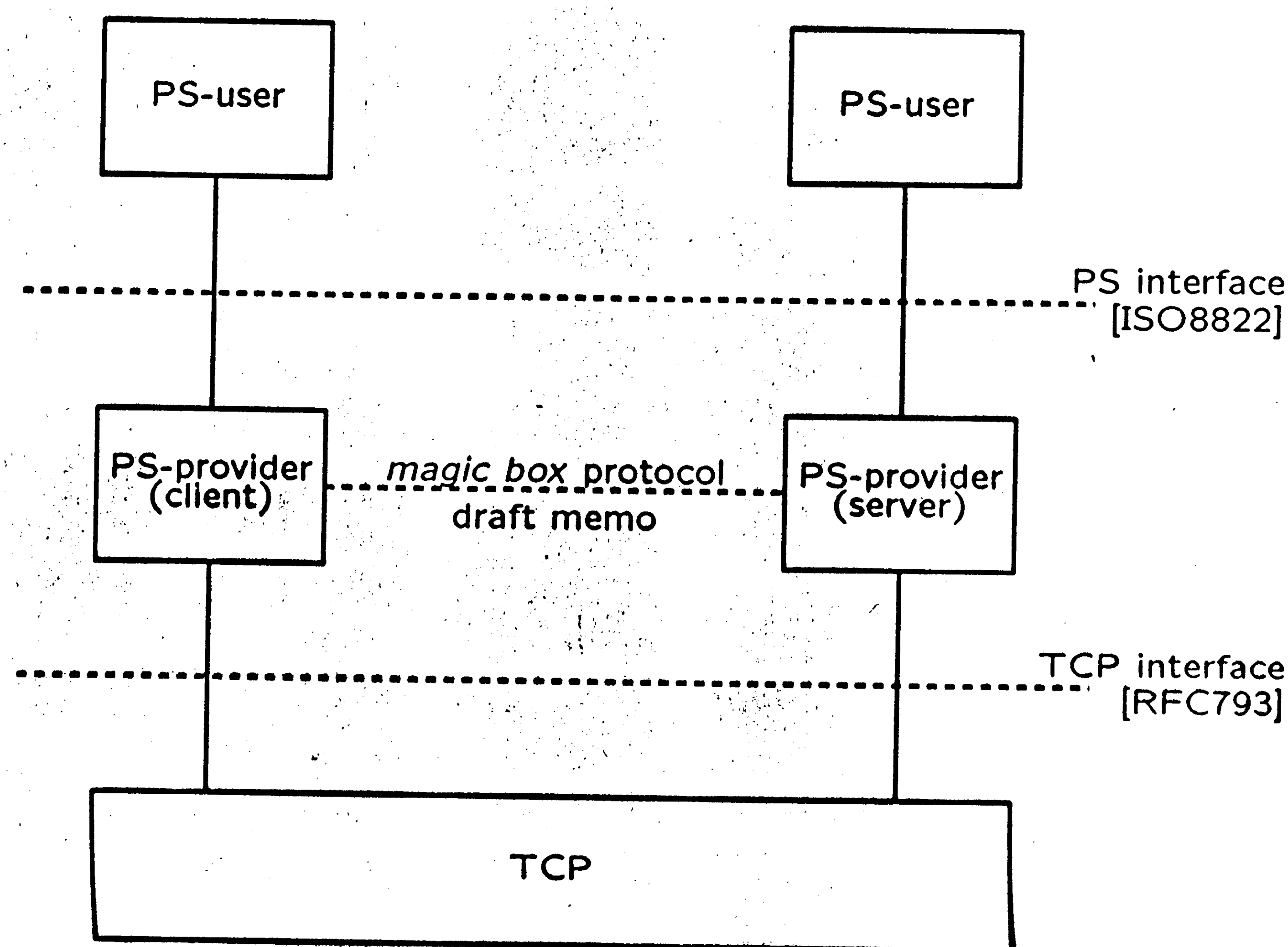
APPROACH

- ISO PRESENTATION SERVICE DEALS WITH ASN.1 OBJECTS
- THESE OBJECTS WHEN SERIALIZED ARE SELF-DELIMITING
- THE TCP IS A STREAM-ORIENTED TRANSPORT PROTOCOL
- THE NEGOTIATION COMPLEXITIES OF THE ISO PRESENTATION PROTOCOL CAN BE AVOIDED

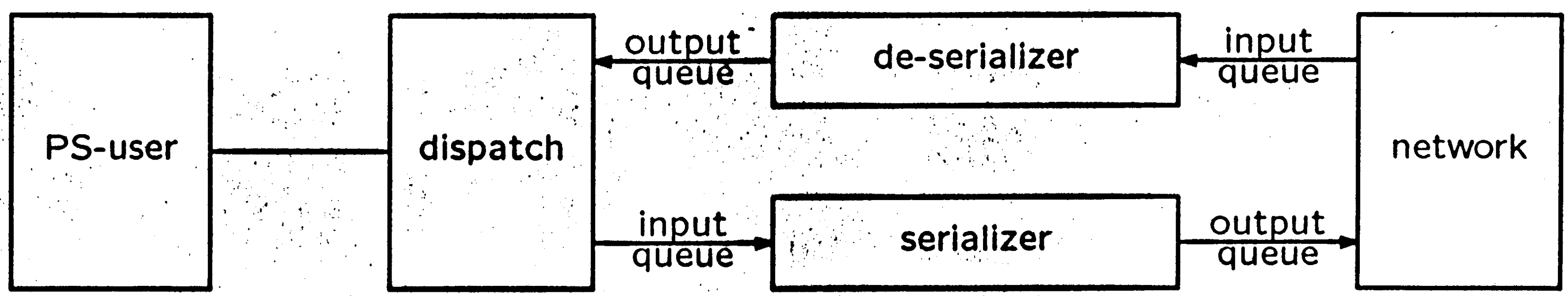
APPLICATION USE OF UPPER-LAYER SERVICES WITH PSEUDO-SERVICE PROVIDERS



OVERALL ORGANIZATION



PSEUDO-PRESENTATION PROVIDER: SOFTWARE ARCHITECTURE



FUNDAMENTAL PARAMETERS

- PRESENTATION ADDRESS
  - 1 OR MORE NETWORK ADDRESSES
  - T-, S-, AND P-SELECTORS
- OUR INTERPRETATION
  - NETWORK ADDRESS
    - 32–BIT IP ADDRESS
    - SET OF AVAILABLE TRANSPORT SERVICES (e.g., TCP)
    - 16–BIT PORT NUMBER
  - T-SELECTOR, S-SELECTOR
    - NULL
  - P-SELECTOR (OPTIONAL)
    - OCTET STRING

FUNDAMENTAL PARAMETERS (cont.)

- PRESENTATION CONTEXT LIST
  - PRESENTATION CONTEXT IDENTIFIER (PCI)
  - ABSTRACT SYNTAX NAME
  - ABSTRACT TRANSFER NAME

- OUR INTERPRETATION

PCI	ASN	ATN
1	SASE PCI	ASN.1
3	ACSE PCI	ASN.1

- PCI FOR SPECIFIC APPLICATION SERVICE ELEMENT (SASE) IS USED BY ROSE
  - THE ROSE APDUs CARRY THE SASE APDUs

CHOICE OF TRANSPORT SERVICE

- USERS MAY WISH TO FORM ASSOCIATIONS WHICH ARE LOW-COST IN THEIR CONSUMPTION OF CONNECTION-RELATED RESOURCES

- THESE ASSOCIATIONS MAY BE USED INFREQUENTLY AND HAVE MINIMAL RELIABILITY CHARACTERISTICS

- FOR EXAMPLE

  A GATEWAY MAY HOURLY REPORT STATISTICS

  FOR 1000 GATEWAYS, IT'S OKAY TO LOSE SOME OF THESE REPORTS

  ALSO IT IS "EXPENSIVE" TO USE HIGH-QUALITY CONNECTIONS

CHOICE OF TRANSPORT SERVICE (cont.)

- THE QUALITY OF SERVICE PARAMETER IS A COLLECTION OF ELEMENTS

- BASED ON THE VALUE OF THE "TRANSPORT MAPPING" ELEMENT, WE CHOOSE A DIFFERENT TRANSPORT MECHANISM

- CHOICES ARE:

  TCP-BASED (HIGH-QUALITY), AND

  UDP-BASED (LOW-QUALITY)

- UDP-BASED SERVICE TRIES NOT TO RE-INVENT THE TCP

ELEMENTS OF PROCEDURE

- STATE MACHINES FOR TCP- AND UDP-BASED SERVICE VERY SIMILAR

- WITH EXCEPTION OF HANDLING OF COLLISION ON P-RELEASE SERVICE, PRESENTATION SERVICE IS IDENTICAL

ELEMENTS OF PROCEDURE:
LOW-QUALITY SERVICE

- CONNECTIONS DISTINGUISHED BY ADDRESSES, PORTS, AND SESSION CONNECTION IDENTIFIER
- SESSION CONNECTION IDENTIFIER CONTAINS "COOKIE" TO DISTINGUISH AMONG PRESENTATION CONNECTIONS
- OPERATIONS WITH A HANDSHAKE

  (CONNECTION ESTABLISHMENT AND RELEASE)

  USE A SIMPLE RETRANSMISSION STRATEGY
- OTHER OPERATIONS TAKE THEIR CHANCES!

PSEUDO-DIRECTORY SERVICES ELEMENT

- PROVIDE TWO MAPPINGS:

  SERVICE NAME TO AN APPLICATION ENTITY TITLE

  APPLICATION ENTITY TITLE TO PRESENTATION ADDRESS

- OUR INTERPRETATION

  SERVICE NAME: "**`<designator>-<qualifier>`**"

  `<designator>` DENOTES A DOMAIN NAME OR IP ADDRESS

  `<qualifier>` DENOTES THE TYPE OF APPLICATION ENTITY

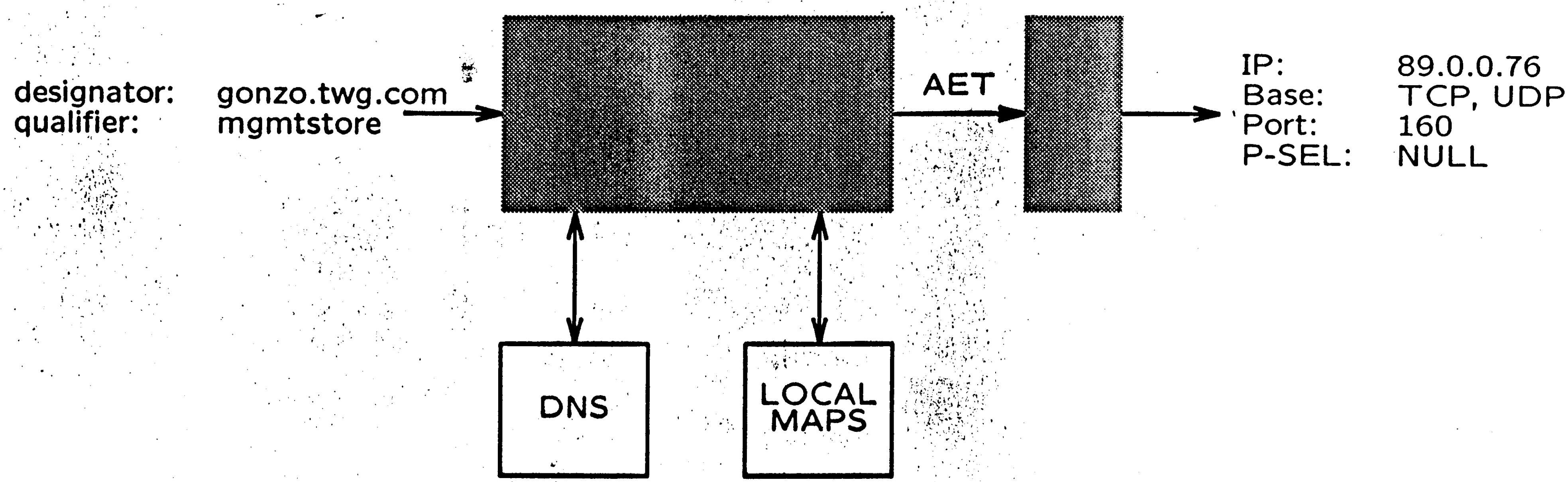
  APPLICATION ENTITY TITLE: OPAQUE

  PRESENTATION ADDRESS: IP ADDRESS, PORT NUMBER, P-SELECTOR

- P-SELECTOR USED TO CONVEY ADDITIONAL ADDRESSING INFORMATION

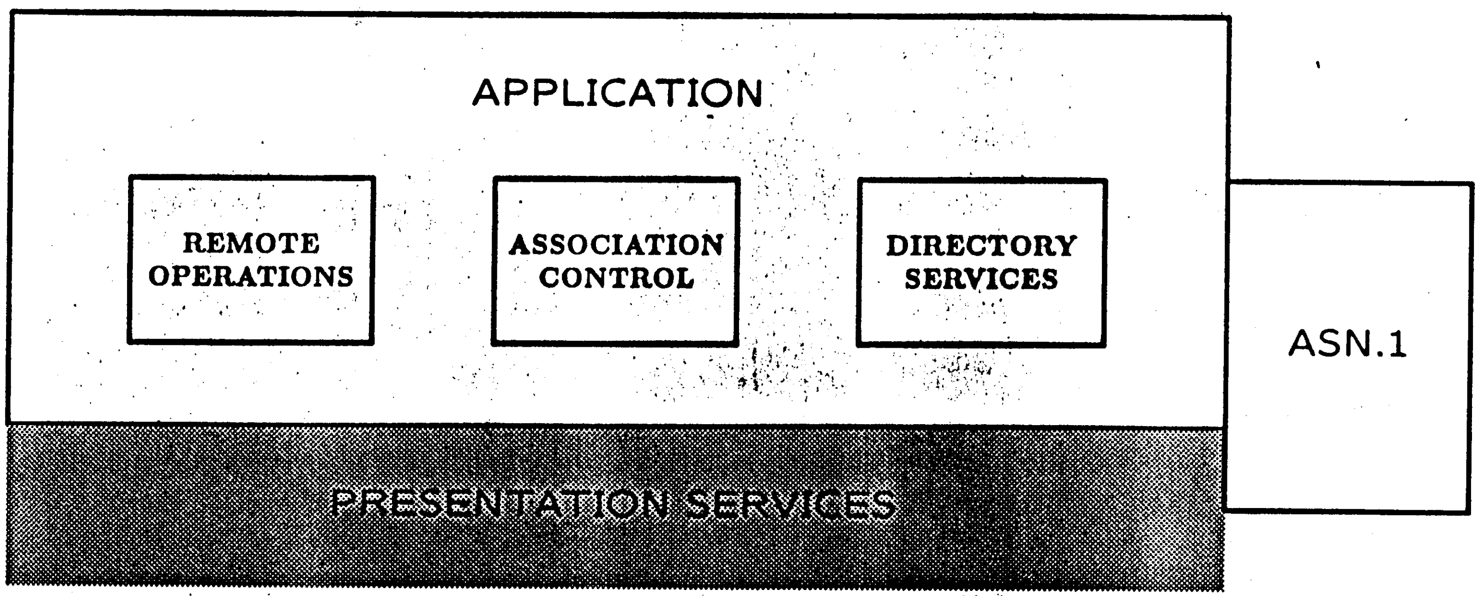
  e.g., FOR PROXY NETWORK MANAGEMENT

DSE MAPPINGS



PUTTING THE PIECES TOGETHER

- AN OPENLY AVAILABLE IMPLEMENTATION OF THE APPLICATION SERVICE ELEMENTS EXISTS
- NEEDED IS AN IMPLEMENTATION OF THE PSEUDO-PRESENTATION PROVIDER



ISODE AVAILABILITY:
USPS

- VERSION 3 AVAILABLE OCTOBER 14, 1987

  CONTAINS ACSE, ROSE, DSE, and ASN.1

- SEND CHECK OR INVOICE FOR $200 US DOLLARS TO:

  ISODE DISTRIBUTION
  DEPARTMENT OF ELECTRICAL ENGINEERING
  UNIVERSITY OF DELAWARE
  NEWARK, DE 19716

  TELCO: 302-451-1163

- DISTRIBUTION CONTAINS:

  1600bpi TAR TAPE

  3 VOLUME DOCUMENTATION SET

ISODE AVAILABILITY:
DARPA/NSF INTERNET

- VERSION 3.2 (BETA) AVAILABLE JANUARY 4, 1988

  CONTAINS ACSE, ROSE, DSE, and ASN.1

  ALONG WITH A STUB-GENERATOR

- USE ANONYMOUS FTP

HOST	louie.udel.edu
FILE	portal/isode-beta.tar.Z
MODE	binary

- FILE IS A COMPRESSED TAR IMAGE

- NEED LaTeX AND A LASER PRINTER

DISCUSSION GROUPS

- THE GROUP:

  ISODE@NRTC.NORTHROP.COM

- LIST ADDITIONS:

  ISODE-REQUEST@NRTC.NORTHROP.COM

REMARKS

- IF YOU CAN AFFORD IT, [RFC1006] IS THE PREFERRED APPROACH
- IF YOU CAN'T, MINIMAL SIZE OF SOFTWARE INFRASTRUCTURE IS KEY
- OBVIOUSLY, THE TWO APPROACHES ARE NOT COMPATIBLE
- A HIGH-DEGREE OF THE WORK DONE WILL BE RE-USABLE IN "REAL" OSI SYSTEMS

6.15 Open Routing WG—Ross Callon, BBN

STATUS REPORT

OPEN ROUTING W.G.

R. CALLON

3/88

GOALS

- ROUTING BETWEEN A.S.'s (REPLACE EGP)
- REQUIREMENTS
- ARCHITECTURE
- PROTOCOLS
- INTERACTION WITH ANSI DESIREABLE

NEXT STEPS

- DEVELOP APPROACHES OFFLINE
- DISCUSS AT NEXT MEETING
- UPDATE REQ'S DOCUMENT

APPROACHES, CONT'D

- TYPES OF SOURCE ROUTING
- ⇒ TWO MAIN APPROACHES (NOT EXCLUSIVE)
- ADDRESSING
- DGP (INFO. ONLY)

DISCUSSION OF POSSIBLE APPROACHES

- ARCH. OUTLINE
- POLICY
- INFO. SUMMARIZING &
  INFO. HIDING
- APPROACHES TO T.O.S.

COMMENTS

- SERVICE
  - "KNOB" FOR COST vs LATENCY
  - T.O.S. METRICS, PRECEDENCE, & QOS
  - MEMORY/CPU/BANDWIDTH

- POLICY
  - MUST BE DEALT WITH

COMMENTS

- OVERALL
  - RELATION WITH ANSI/ISO/ECMA
  - EVALUATE & GUIDE PROPOSAL
  - PROPOSALS ARE LIKELY TO BE IMPERFECT (TRADEOFFS)

- ARCHITECTURE
  - GWYS IN 2+ A.S.'s
  - REQ's ON IGP's

CURRENT ACTIVITIES

- COMMENTS ON IDEA 007 ("REQUIREMENTS FOR INTER-AUTONOMOUS SYSTEMS ROUTING")
- DISCUSSION OF POSSIBLE APPROACHES

6.16 Host Requirements WG—Bob Braden, ISI

INTERNET HOST REQUIREMENTS

Objective: Host analog to Gateway Requirements RFC-1009.

Schedule:

- People signed up to write sections
- Meeting early April
- Draft by early May, as IDEA

OUTLINE

1. Introduction

2. Media Support
   [~ same as RFC 1009]
3. Internet Layer
   IP, ICMP, Addressing, & Routing
4. Transport Layer
   UDP, TCP
5. Support Services
   BOOTP, DNS, Net mgt., RARP
6. Applications
   Telnet, FTP, SMTP

Appendix:
   Checklist

6.17 Routing IP Datagrams Through X.25 Nets—C-H Rokitansky, DFVLR

INTERNET ENGINEERING TASK FORCE
SUPER COMPUTER CENTER, SAN DIEGO
MARCH 1-3 '88

ROUTING OF INTERNET DATAGRAMS THROUGH X.25 PUBLIC DATA NETWORKS

CARL-HERBERT ROKITANSKY

GERMAN AEROSPACE RESEARCH ESTABLISHMENT
(DFVLR)

MARCH 1988

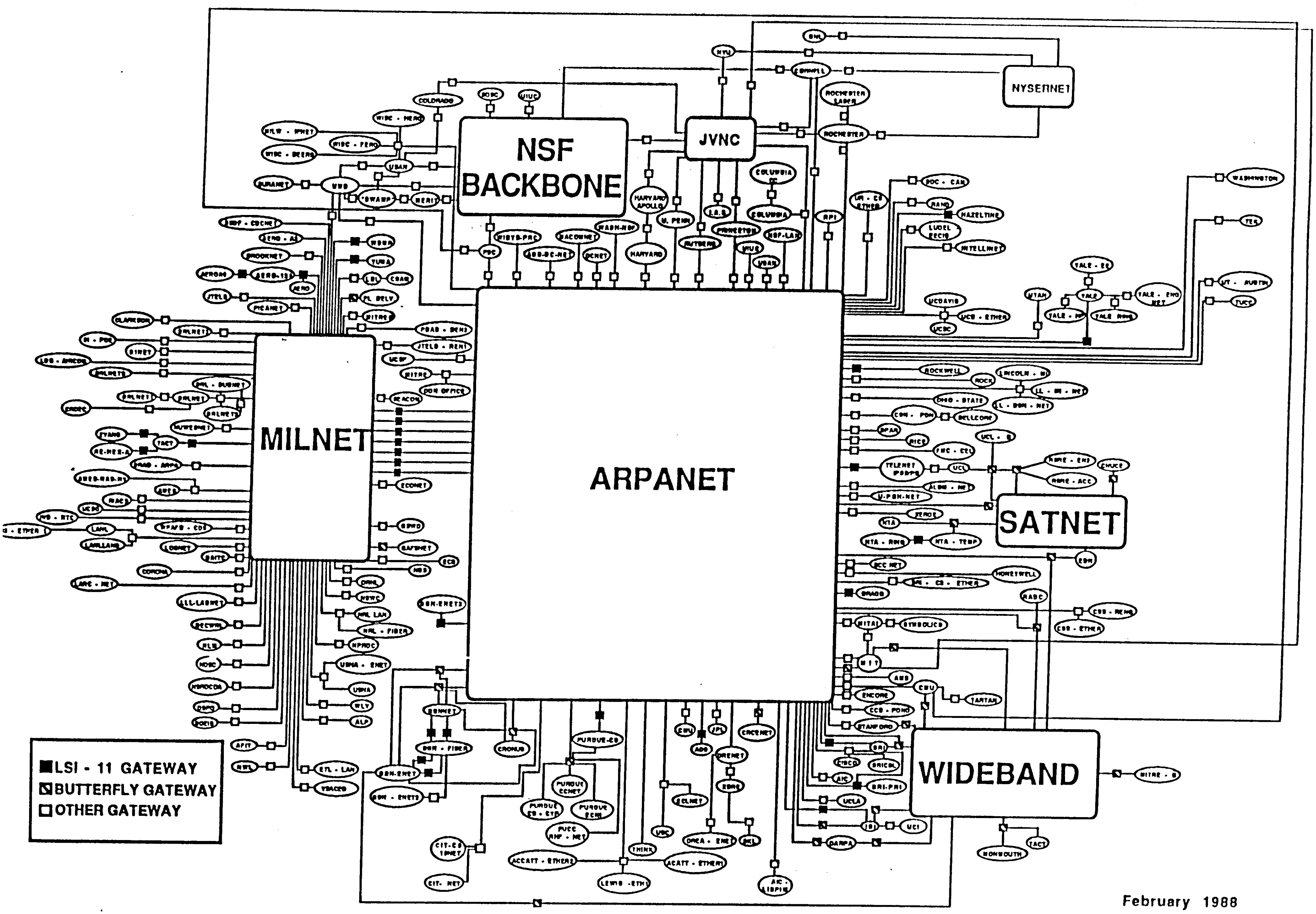
ROUTING OF INTERNET DATAGRAMS THROUGH X.25 PUBLIC DATA NETWORKS

- EUROPEAN TCP/IP STATUS
- CLUSTER-ADDRESSING SCHEME (OVERVIEW)
- INTERNET/X.25 RESEARCH ISSUES

SYSTEMS MULTINET (TCP/IP), MUNICH OCTOBER '87:

36 MANUFACTURES (VENDORS):

AEG KABEL
ALLEN - BRADLEY
APOLLO DOMAIN
APPLE COMPUTER
CADTRONIC
COMCONSULT
CONVEX
DANET
DEUTSCHE BUNDESPOST
DIGITAL EQUIPMENT
ELTEC
FIBRONICS
GEI RECHNERSYSTEME
GOULD
HEWLETT-PACKARD
HIRSCHMANN
IBM
KÖPKE
KRONE
NCR
NIXDORF
PAN DACOM
PCS COMPUTERSYSTEME
POSITRONIKA
SCHNEIDER & KOCH
SIEMENS
STEMMER ELEKTRONIK
STOLLMANN
SUN MICROSYSTEMS
SYMBOLICS
SYNELEC DATENSYSTEME
TEKTRONIX
TELEMATION
UNISYS
WESTERN DIGITAL
WETRONIC

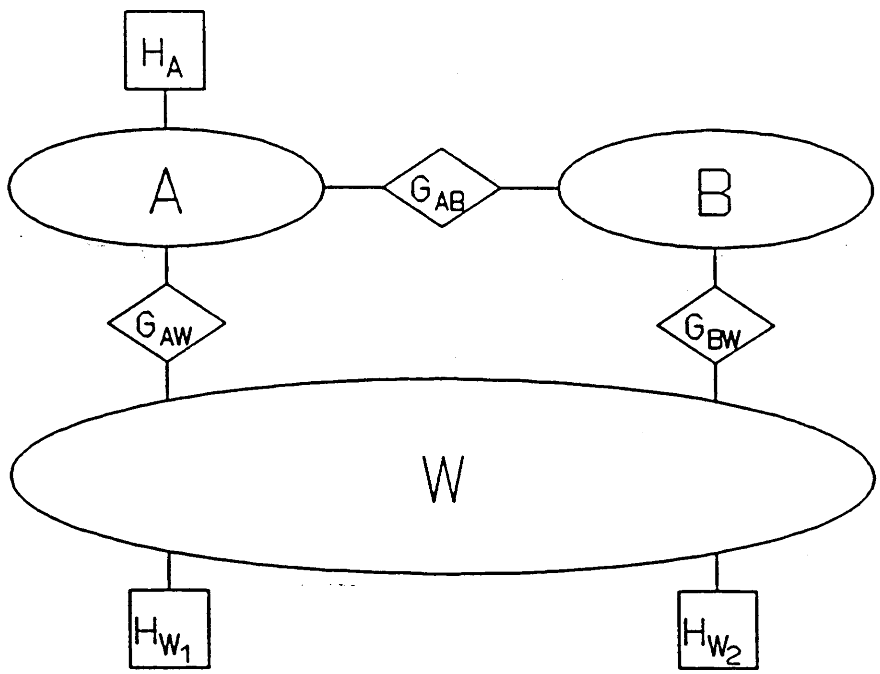


February 1988

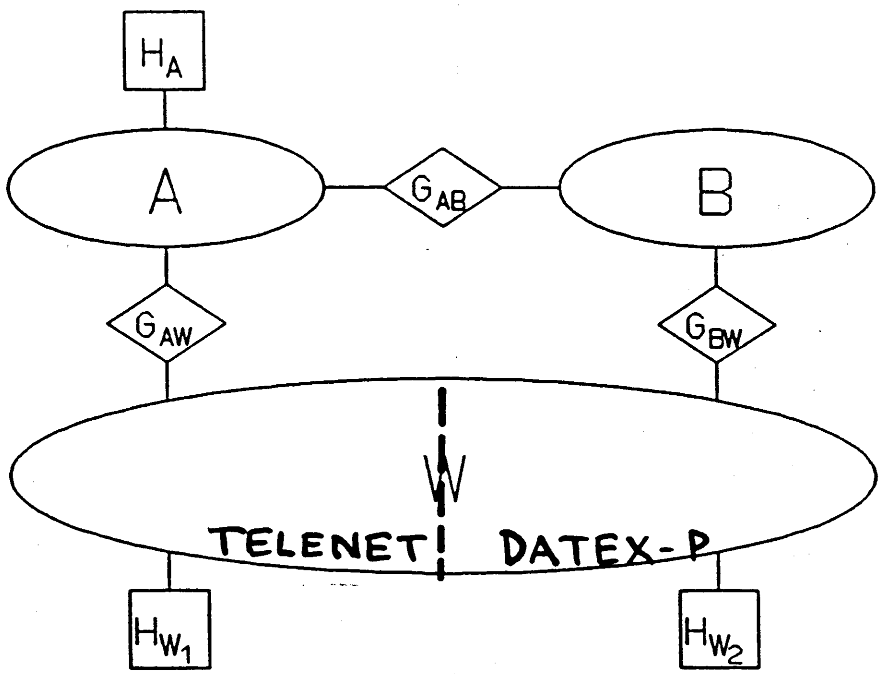
X.25 FIXED COSTS (VIRTUAL CIRCUIT)

EUROPE (GERMANY) : ~ 120 US $ / MONTH

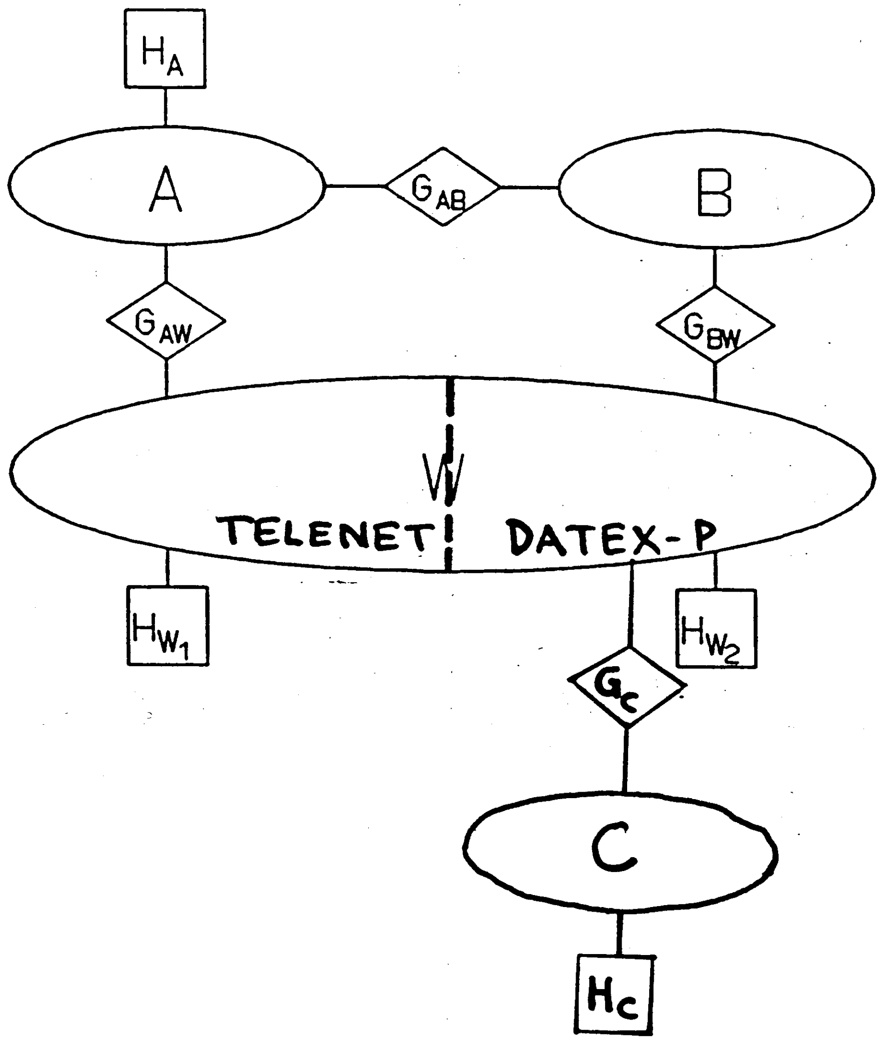
U.S. : ~ 800 TO 1200 US $ / MONTH



H_A
A
G_AB
B
G_AW
G_BW
W
H_W1
H_W2



H_A
A
G_{AB}
B
G_{AW}
G_{BW}
W
TELENET
DATEX-P
H_{W_1}
H_{W_2}



H_A
A
G_{AB}
B
G_{AW}
G_{BW}
W
TELENET
DATEX-P
H_{W_1}
H_{W_2}
G_C
C
H_C

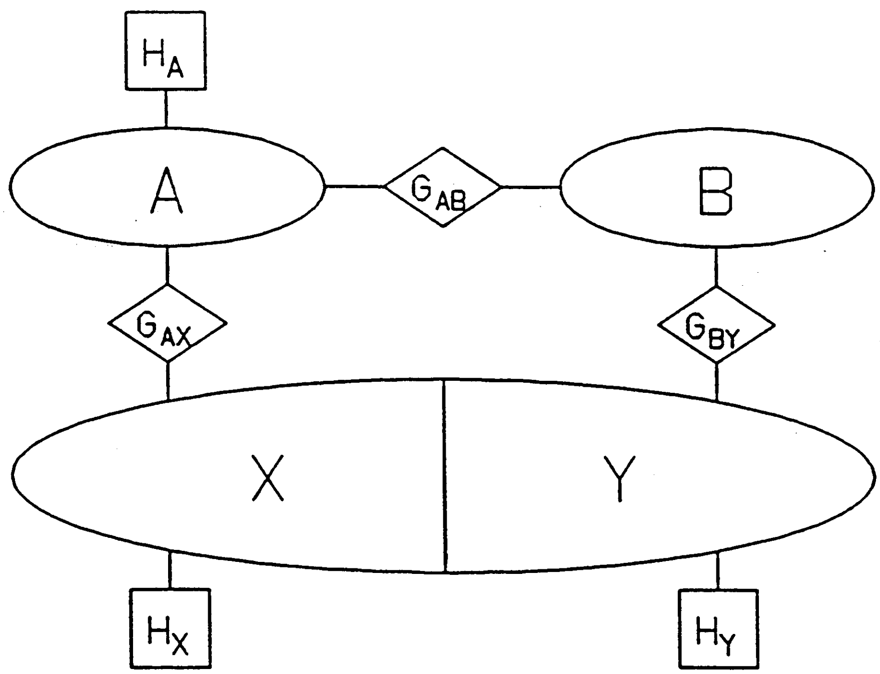


Figure 3-2

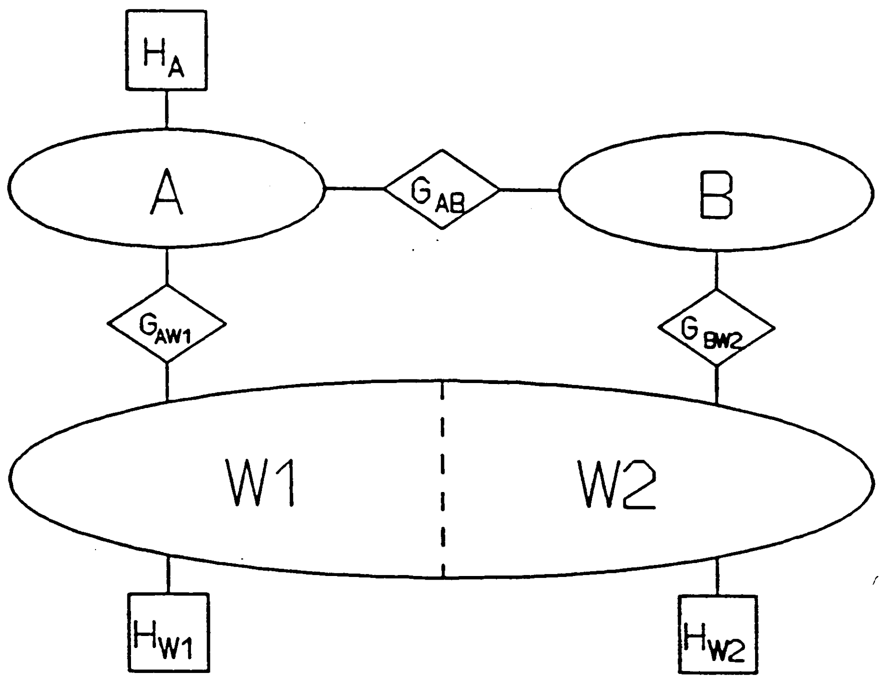


Figure 3.1-1

CLUSTER - ADDRESSING SCHEME - CONCEPT

- SPECIFIC INTERNET NETWORK NUMBERS ARE ASSIGNED TO A SET OF NETS BETWEEN WHICH DIRECT CONNECTIONS CAN BE ESTABLISHED WITHOUT TRANSITING A GATEWAY

- THESE NETWORKS ARE ASSOCIATED TO AN "INTERNET CLUSTER"

- AN ADDRESS-MASK, CALLED „CLUSTER-MASK" IS USED FOR ROUTING DECISIONS WITHIN THE CLUSTER

Measurements:

- Client/Server pair
  - ⇒ Memory to Memory transfer rates
  - ⇒ Bi-directional
  - ⇒ Many options for setting various buffer sizes
- Latest numbers:128k send/receive space, 64K window

Driver	MTU	Checksum	Usertokern	Xfer Rate
hsx	24K	on	4K	62.3 Mbits
hsx	24K	on	24K	67.8 Mbits
hsx	24K	off	24K	85.1 Mbits
lo	32K	on	4K	118.3 Mbits

Xfer Rate	Xfer Size	Pkts per sec	Check-sum (usec)	Time packet(usec)
118Mbits	32K	451	990	1210
67Mbits	24K	340	734	2166
85Mbits	24K	430	0	2300

Cluster - Scheme

<INTERNET Address> ::= <NETWORK-Nr><RESTFIELD>

<NETWORK-Number> ::= <CLUSTER-Number><CLUSTER-NET>

Cluster - Mask

<INTERNET Address> ::=

<CLUSTER-Number><CLUSTER-NET><RESTFIELD>

1 . . . . 1 0 . . . . 0 0 . . . . 0 CLUSTE MASK

f.e.

255. 0. 0. 0. CLUSTER MASK

Public Data Networks (PDN) - Characteristics:

- Wide Area Network
- Complex of national public data networks
- International virtual circuits
- Different costs for international and national virtual circuits
- Costs depend on data volume and length of time of connection
- no broadcasting

Proposed Solution:

- INTERNET class B network numbers (with identical bits in the first (high-order) 8-bit field of the INTERNET address) are assigned to national public data networks.

- The national public data networks are assembled to form a cluster of networks ("PDN-Cluster")

- Use of a „Cluster-mark", thus all hosts within the „PDN-Cluster" appear to be reachable „locally"

- If necessary, WAN gateways are exchanging (modified) EGP messages on an "event driven" basis (i.e. No periodic updates (!))

- Mapping between the INTERNET address and X.121 address of PDN hosts is done by an „X.121 Address Server/Resolution Protocol"

DNIC Mapping (8 bits to specify the <cluster-net>)

- use cluster-mask <255.0.0.0>
- reserve network numbers 191.001 to 191.254 for the "PDN-cluster" (191.000 and 191.255 reserved)
- assign INTERNET network numbers to DNICs in order of request

Example:

DNIC	Public Data Network	INTERNET network #
3110	TELENET (USA)	191. 1
2342	IPSS (U.K.)	191. 2
2405	TELEPAK (Sweden)	191. 3
2041	DATANET (Netherlands)	191. 4
2624	DATEX-P (West Germany)	191. 5

DNIC	Public Data Network	INTERNET Network Number
3110	TELENET (USA)	191.1
2041	DATANET (Netherlands)	191.2
2342	IPSS (U.K.)	191.3
2405	TELEPAK (Sweden)	191.4
2624	DATEX-P (West Germany)	191.5
etc.		

ADVANTAGES OF THE CLUSTER-ADDRESSING SCHEME

- INTERNAL STRUCTURE OF THE PDN-SYSTEM BECOMES VISIBLE TO THE OUTSIDE WORLD (IMPORTANT FOR ROUTING DECISIONS)

- FACT THAT AN INTERNET CLUSTER HAS BEEN FORMED IS INVISIBLE OUTSIDE THE CLUSTER (→ NO CHANGES TO THE EXISTING INTERNET GATEWAY SYSTEM)

- ALL HOSTS (GATEWAYS) WITHIN THE SAME CLUSTER APPEAR TO BE REACHABLE DIRECTLY („LOCALLY")

- ICMP REDIRECT MESSAGES CAN BE USED WITHIN THE SAME CLUSTER

- NO (or only MINOR) CHANGES TO HOSTS SUPPORTING SUBNETS

- ICMP ADDRESS MASK REQUEST/REPLY MESSAGES CAN BE USED WITHOUT CHANGES

DISADVANTAGE: SPECIFIC INTERNET NETWORK NUMBERS MUST BE RESERVED FOR EACH CLUSTER

Neccessary Implementations to support the CLUSTER - Concept

INTERNET Hosts:

no changes

INTERNET Gateways:

no changes

PDN Hosts:

no changes (if software supports SUBNET /(CLUSTER) - Scheme)

VAN Gateways:

- Cluster Scheme
- IP Source Route Option (modified use)
- modified EGP (event driven)

X.25 (PDN) RELATED RESEARCH ISSUES

- MODIFIED EGP FOR VAN-GATEWAYS (EVENT DRIVEN BASIS)
- X.121 ADDRESS RESOLUTION PROTOCOL
- ROUTING METRICS ((COSTS, etc.)
- PERFORMANCE TESTS
- ISO MIGRATION

→ PDN INTERNET ROUTING GROUP

6.18 Internet Multicast—Steve Deering, Stanford

Internetwork Multicasting

Steve Deering

Computer Systems Laboratory

STANFORD UNIVERSITY

The Host Group Model

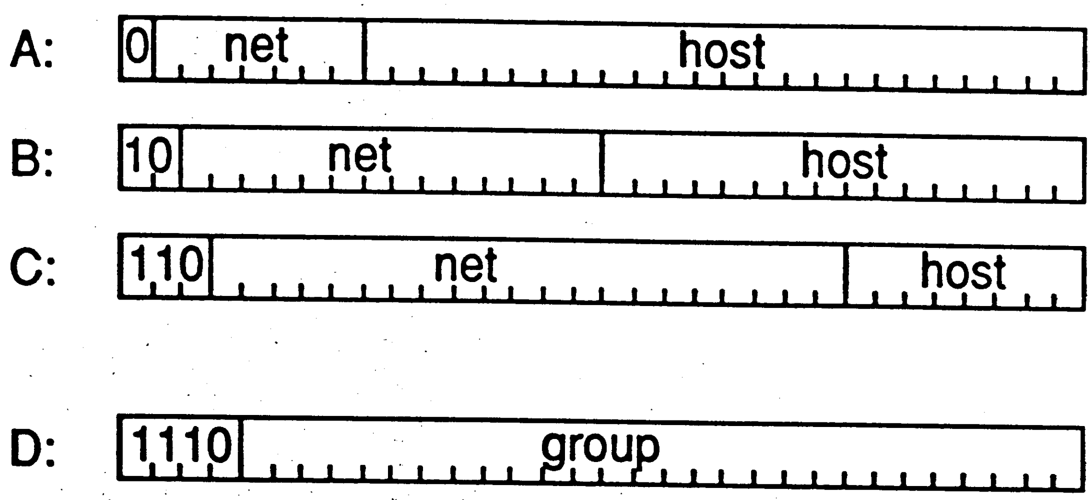
A *host group* is a set of zero or more hosts

- identified by one internetwork address (permanent or temporary)
- members anywhere in the internetwork
- dynamic and unbounded membership

A datagram sent to a host group address is delivered to all current members, *subject to:*

- unreliability of datagram delivery
- non-atomicity of membership changes
- delivery scope constraint (e.g., TTL)

Host Group Addresses



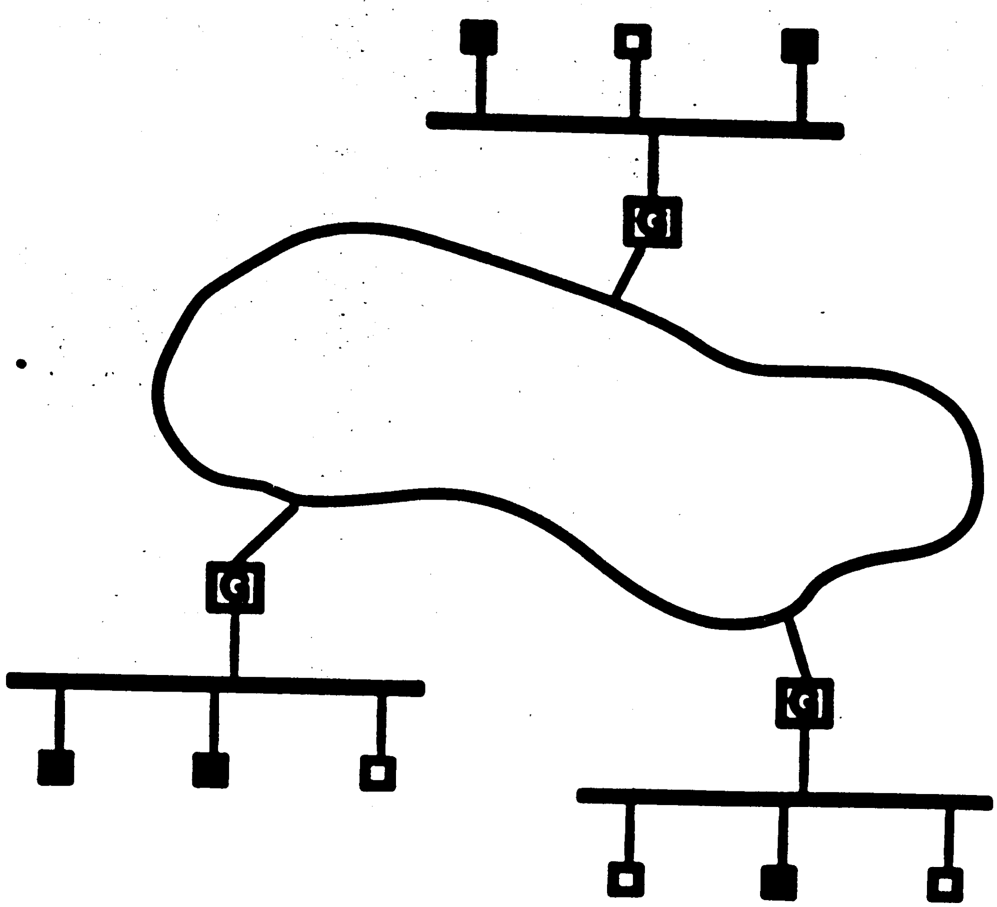
group addresses are independent of networks

some reserved for **permanent groups**
(e.g. name server group, gateway group)

rest available for **transient groups**
(e.g. conferences, distributed computations)

Delivery Strategy

- sender transmits as a local multicast
- first gateway forwards to gateways on other member networks (the "network group")
- remote gateways multicast to their own local members



THREE FAMILIAR SUBPROBLEMS:

① HOST REQUIREMENTS

INCLUDING "MULTICAST ES-IS PROTOCOL"

② INTER-GATEWAY MULTICAST ROUTING

"MULTICAST IGP"

③ INTER-DOMAIN MULTICAST ROUTING

(MULTIPLE ADMINISTATIONS OR DIFFERENT IGPs)

"MULTICAST EGP"

HOST REQUIREMENTS

- REVISED RFC-988 "HOST EXTENSIONS FOR IP MULTICASTING"
  - WILL MAKE AVAILABLE AS AN "IDEA"
  - POTENTIAL INTERNET STANDARD
  - "HOST" INCLUDES NON-MULTICASTING GATEWAYS.
- RELATIVELY MINOR MODIFICATIONS FOR:
  - SENDING MULTICAST IP DATAGRAMS
    (ROUTING DECISION, TTL CONTROL, MULTIHOMING)
  - RECEIVING MULTICAST IP DATAGRAMS
    (LOCAL MEMBERSHIP LIST, JOIN GROUP/LEAVE GROUP
  - LINK LEVEL ADDRESS MAPPING + FILTERING
    (SPECIFIED FOR ETHERNET, GUIDELINES FOR OTHER)

MODIFICATIONS ARE SUFFICIENT TO SUPPORT <u>SINGLE-NETWORK</u> IP MULTICASTING —

CAN MIGRATE CURRENT IP <u>BROADCAST</u> APPLICATIONS TO USE MULTICAST. E.G.:

- RWHO
- RIP
- BOOTP
- ICMP MASK REQUEST
- (ARP, RARP)

  ⋮

AND SUPPORT PROPOSED BROADCAST APPLICATIONS:

- ICMP GATEWAY REQUEST
- SPF LINK STATE FLOODING

  ⋮

"MULTICAST ES-IS"

INTERNET GROUP MANAGEMENT PROTOCOL (IGMP)

MANDATORY FOR ALL MULTICASTING IP IMPLEMENTATIONS

- MULTICAST ROUTERS PERIODICALLY SEND MEMBERSHIP QUERY TO "ALL NEIGHBOR HOSTS GROUP" (SPECIAL CASE)
- REPLIES DELAYED BY RANDOM AMOUNT TO AVOID "IMPLOSION".
- REPLIES ARE MULTICAST SO OTHER MEMBERS CAN HEAR THEM AND SUPPRESS THEIR OWN REPLIES

∴ NORMALLY ONLY ONE REPLY PER GROUP PRESENT.

- QUERY RATE CAN BE VERY LOW
- FAST QUERIES WHEN MULTICAST ROUTER START UP
- UNSOLICITED REPLY WHEN HOST JOINS NEW GROUP

NOTE: IF NO MULTICAST ROUTER PRESENT, ONLY IGMP TRAFFIC IS FROM NEW JOINS.

STATUS

ALL OF RFC-988+ IS IMPLEMENTED FOR BSD 4.3+

TO BE AVAILABLE FROM BERKELEY OR STANFORD

- ONLY CURRENT APPLICATIONS:

- PING
- RWHOD
- RESOLVER ROUTINES / NAMED
  (USES "EXPANDING RING SEARCH")

- COMMENTS ON DRAFT RFC,
  GUINEA PIGS FOR IMPLEMENTATION,
  AND NEW APPLICATIONS
  WOULD ALL BE APPRECIATED!

"MULTICAST IGP"

MANY POSSIBILITIES — BEST INTEGRATED WITH UNICAST IGP:

INDEPENDENT

- STATIC ROUTES (CURRENT IMPLEMENTATION)
- SINGLE SPANNING TREE (E.G. DEC LANBRIDGE PROTOCOL)

BASED ON DISTANCE-VECTOR ROUTING

- REVERSE PATH FORWARDING (NO PROTOCOL CHANGE)
- "AUTUMN" R.P.F. (NEW BIT IN ROUTING TUPLES)
  (NEXT TO IMPLEMENT ?)

BASED ON LINK-STATE ROUTING

- ANY OF THE ABOVE
- PER-SOURCE MULTICAST TREE
  - FLOOD MEMBERSHIP AS PART OF LINK STATE
  - COMPUTE TREES FROM GRAPH
  - ON-DEMAND COMPUTATION + CACHING OF:
    (SOURCE, GROUP) → OUTGOING LINKS

MULTICAST EGP

REQUIRED PROPERTIES OF A MULTICAST ROUTING DOMAIN:

- INJECTED MULTICASTS REACH ALL INTERIOR MEMBERS + ALL BOUNDARY ROUTERS (SUBJECT TO TTL)

- INTERNALLY-ORIGINATED MULTICASTS REACH ALL BOUNDARY ROUTERS (SUBJECT TO TTL)

- CAN DISCOVER INTERIOR MEMBERSHIPS BY ASKING ANY BOUNDARY ROUTER

∴ LOOKS JUST LIKE A BIG ETHERNET ⇒ CAN USE IGP PROTOCOLS HIERARCHICALLY.

CURRENT PROPOSAL:

USE WIDEBAND SATELLITE NETWORK AS A WIDE-AREA MULTICAST BACKBONE, LINKING MULTICAST DOMAINS.

- EACH DOMAIN ELECTS AN EXTERIOR MULTICAST ROUTER.
- EXTERIOR ROUTER "TUNNELS" TO NEAREST WIDEBAND GATEWAY.
- CAN MIGRATE TO HIGH-SPEED, LOW-DELAY TERRESTRIAL BACKBONE WHEN AVAILABLE

6.19 TCP Performance Prototyping—Van Jacobson, LBL

What happens to throughput of a vanilla TCP (no slowstart or congestion avoidance) as the loss rate goes up?

Say the loss rate is p. Say the round trip time is R and the window size is W so the no-loss throughput $X_0 = W/R$. Assume W is less the delay-bandwidth product. Assume $p \ll 1$ so we can ignore retransmits of retransmits.

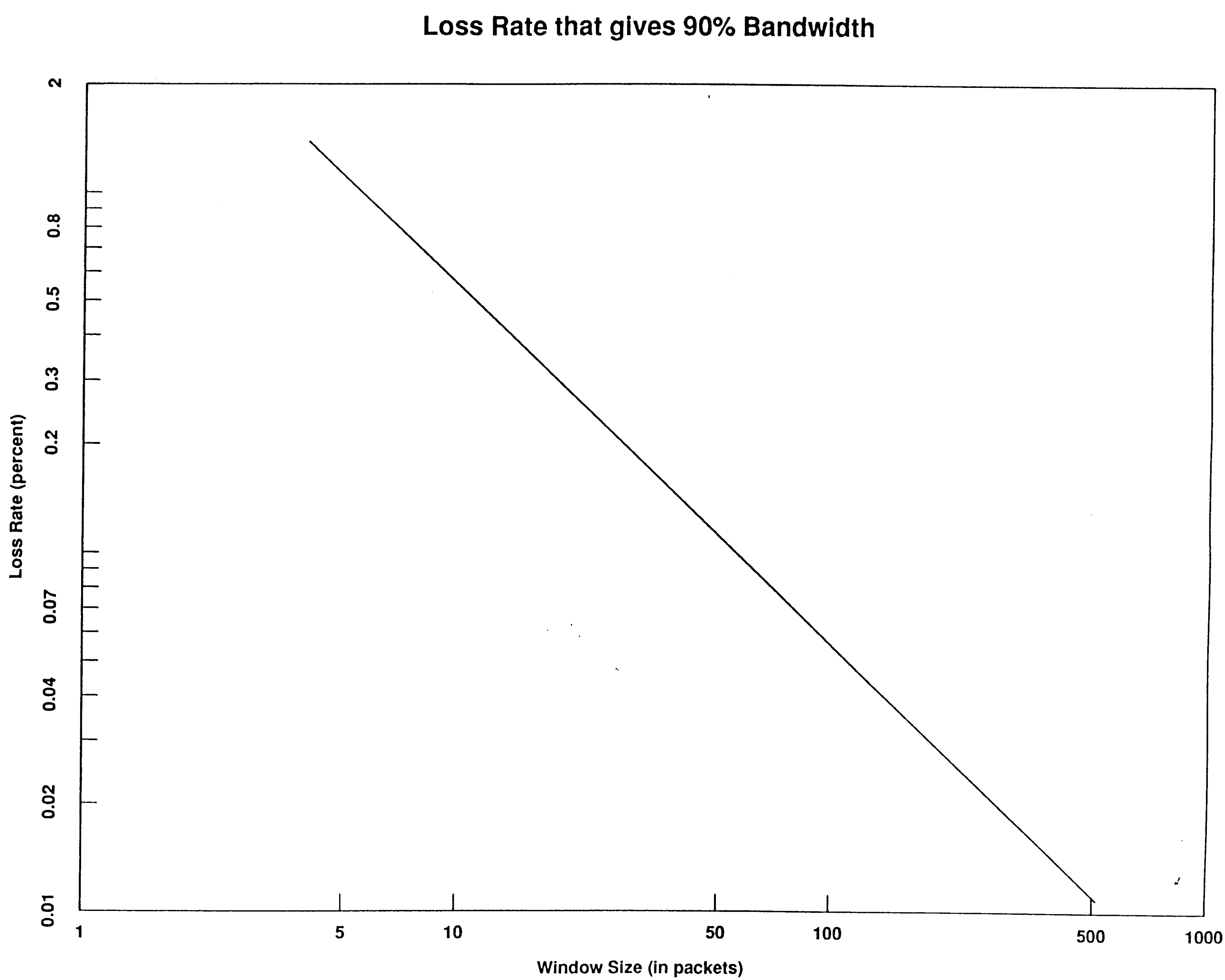
A loss rate of p means $1/p$ packets between losses. Since the bandwidth is W/R, it takes time $(1/p)/(W/R)$ to send those packets plus $2R$ to detect and retransmit the lost packet. Since effective throughput is packets over time, we get:

$$X(p) = \frac{1/p}{(1/p)/(W/R) + 2R}$$

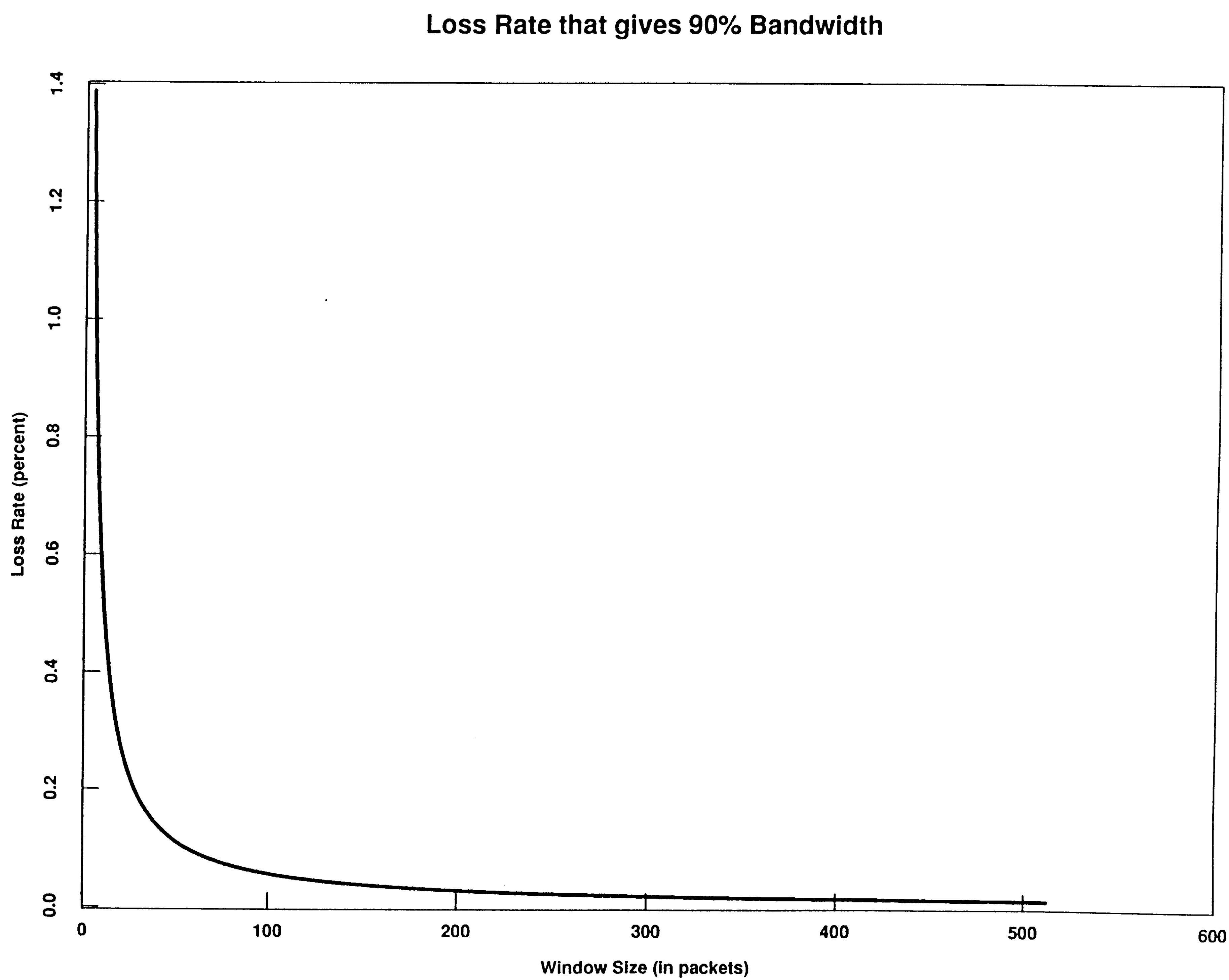
$$= \frac{1}{1/(W/R) + 2pR}$$

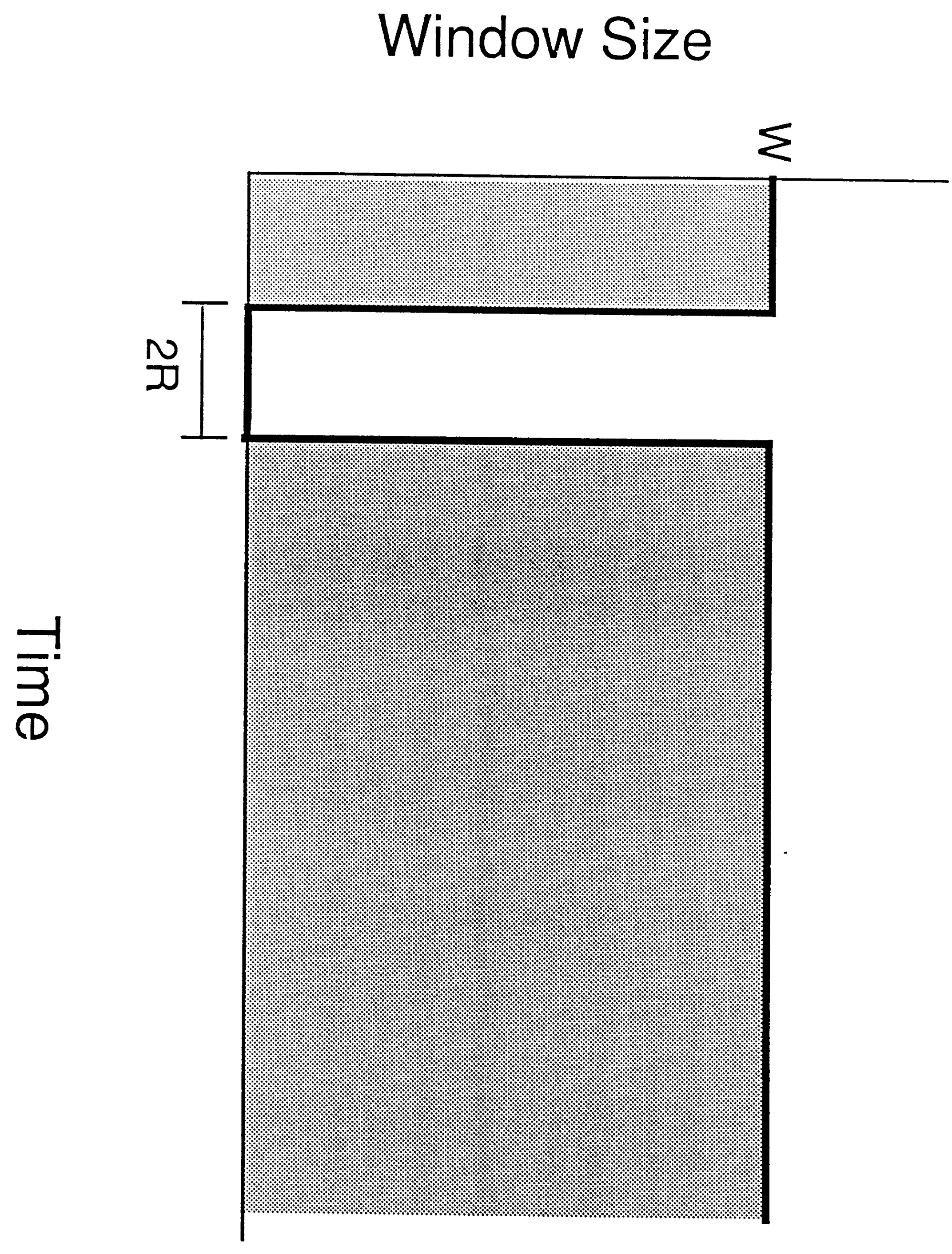
$$= \frac{W}{R} \frac{1}{1 + 2pW}$$

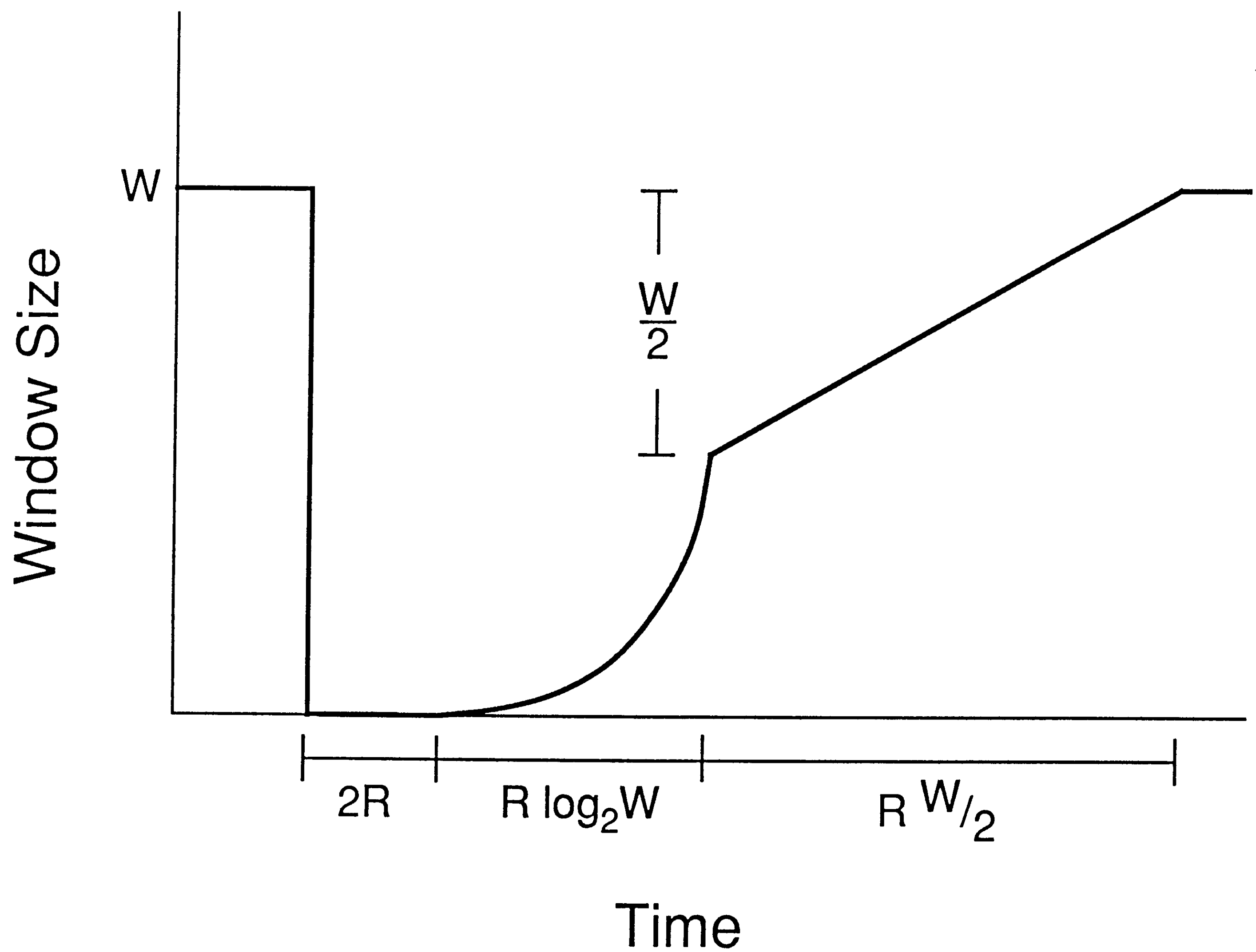
$$= X_0 \frac{1}{1 + 2pW}$$



Loss Rate that gives 90% Bandwidth
Loss Rate (percent)
2
0.8
0.5
0.3
0.2
0.07
0.04
0.02
0.01
1
5
10
50
100
500
1000
Window Size (in packets)


Loss Rate that gives 90% Bandwidth
Loss Rate (percent)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
100
200
300
400
500
600
Window Size (in packets)


Window Size
W
2R
Time


Window Size
W
W/2
2R
R log2W
R W/2
Time

Window Size this packet

Window Size next packet	2	3	4	5	6	7	8
2	**p**	**p**	**p**	0	0	0	0
3	**q**	0	0	**p**	**p**	0	0
4	0	**q**	0	0	0	**p**	**p**
5	0	0	**q**	0	0	0	0
6	0	0	0	**q**	0	0	0
7	0	0	0	0	**q**	0	0
8	0	0	0	0	0	**q**	**q**

probability of changing to next size

p = probability of packet loss

q = 1 – p = probability of no loss

Compute one-step window size distribution for starting size of six packets and limit of eight packets:

$$\begin{bmatrix} p & p & p & 0 & 0 & 0 & 0 \\ q & 0 & 0 & p & p & 0 & 0 \\ 0 & q & 0 & 0 & 0 & p & p \\ 0 & 0 & q & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & q & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & q & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & q & q \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ p \\ 0 \\ 0 \\ 0 \\ q \\ 0 \end{bmatrix}$$

In general, if A is the transition matrix and s is a window size distribution vector, the size distribution n steps in the future is $A^n s$.

If A is regular, i.e., if the rows of A^n become identical for some n, then there is an equilibrium distribution of sizes and *any* size distribution will eventually turn into the equilibrium distribution.

Try a numerical experiment with the transition matrix for a 10% loss rate:

$$A = \begin{bmatrix} .1 & .1 & .1 & . & . & . & . \\ .9 & . & . & .1 & .1 & . & . \\ . & .9 & . & . & . & .1 & .1 \\ . & . & .9 & . & . & . & . \\ . & . & . & .9 & . & . & . \\ . & . & . & . & .9 & . & . \\ . & . & . & . & . & .9 & .9 \end{bmatrix}$$

Guess that somewhere between 10 and 20 iterations will be needed to forget the initial conditions. Fire up Macsyma and compute:

$$A^{15} = \begin{bmatrix} .01 & .01 & .01 & .01 & .01 & .01 & .01 \\ .03 & .03 & .03 & .03 & .03 & .03 & .03 \\ .09 & .09 & .09 & .09 & .09 & .09 & .09 \\ .09 & .09 & .09 & .09 & .09 & .09 & .09 \\ .08 & .08 & .08 & .08 & .08 & .08 & .08 \\ .07 & .07 & .07 & .07 & .07 & .07 & .07 \\ .63 & .63 & .63 & .63 & .63 & .63 & .63 \end{bmatrix}$$

Looks like $s = [\ 1 \quad 3 \quad 9 \quad 9 \quad 8 \quad 7 \quad 63\ ]$ is the equilibrium distribution. This says that window size will be 2 packets 1% of the time, 3 packets 3%, 4 and 5 packets 9%, 6 packets 8%, 7 packets 7%, 8 packets 63%.

The average window size must be size times the percent of time spent at that size, summed over all the sizes:

$$\sum_{i=2}^{8} i \times s_{i-1} = 6.9 \text{ packets}$$

So we take a $1 - 6.9/8 = 13\%$ throughput hit because of the congestion avoidance algorithm. (Compare this to the 38% hit any TCP takes because of the 10% loss rate.)

Life is simpler if we note that if s is the equilibrium window size distribution, it must be the case that

$$As = s$$

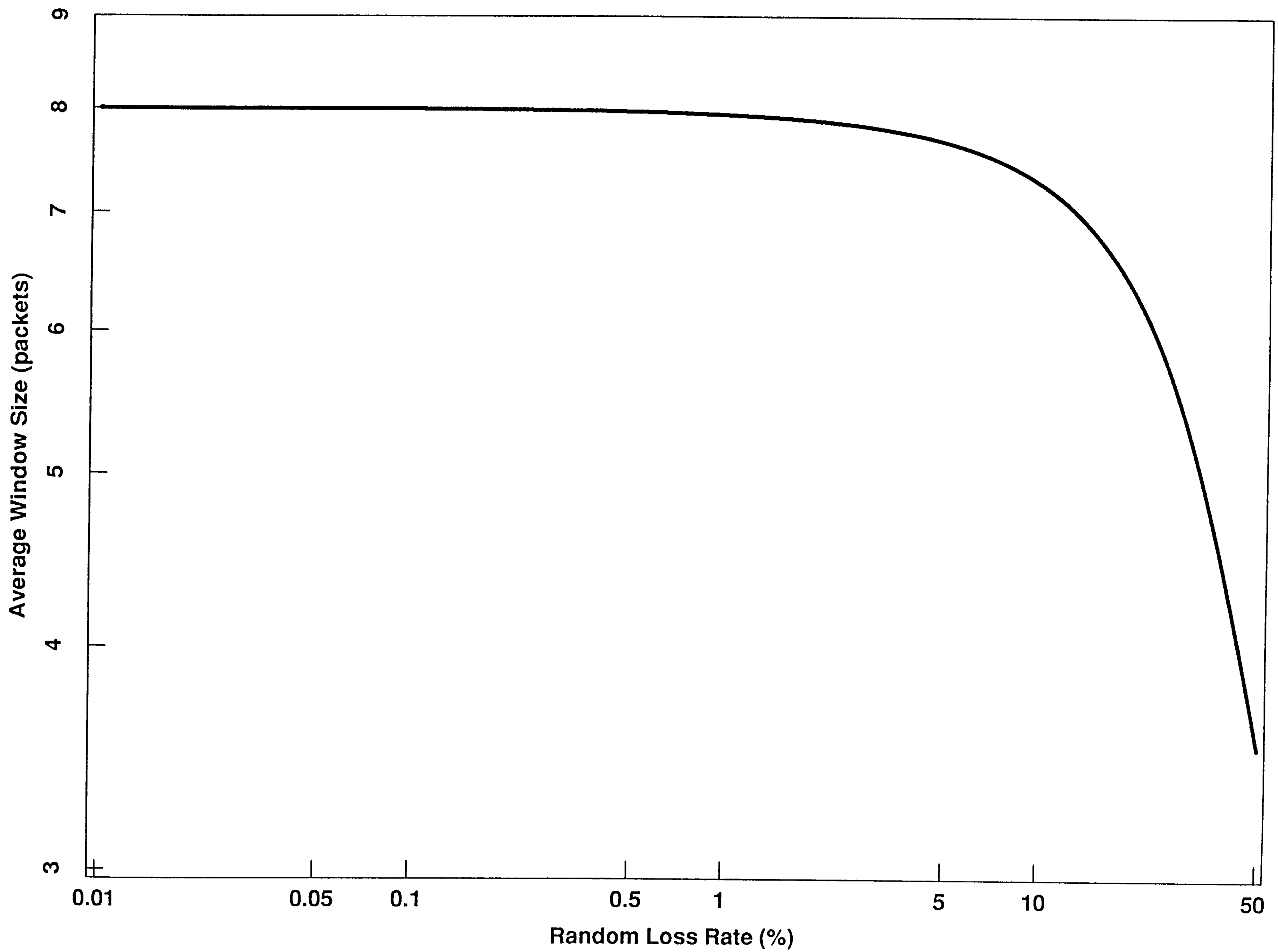
That is, s must be an eigenvector of A with eigenvalue 1.

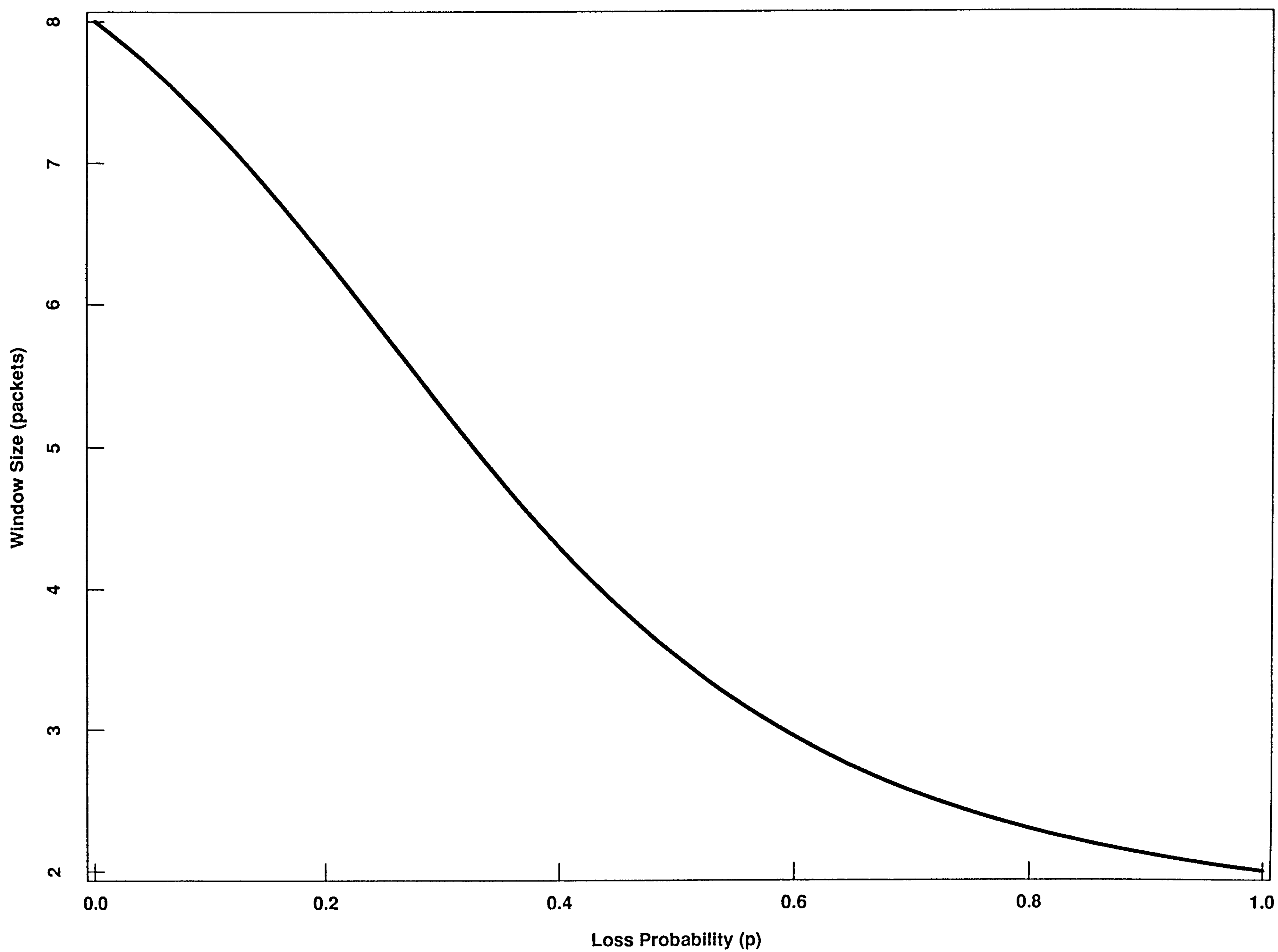
Remembering that Macsyma has an eigenvector package, we can solve for the equilibrium distribution as a function of p for any loss probability p. The average window size then just the inner product of the equilibrium distribution and a vector of sizes.

Computing this for an 8 packet window, we get the average window size, $\overline{W}$, as a function of the loss probability:

$$\overline{W}(p) = \frac{8 - 30p + 61p^2 - 70p^3 + 50p^4 - 21p^5 + 4p^6}{1 - 3p + 7p^2 - 8p^3 + 7p^4 - 4p^5 + p^6}$$

Effect of Random Packet Loss on Congestion Avoidance Algorithm




Window Size (packets)
8
7
6
5
4
3
2
0.0
0.2
0.4
0.6
0.8
1.0
Loss Probability (p)

What if there were no upper limit on the window size?

Let's use a more general adjustment rule:

On loss: $W_i = dW_{i-1} \qquad (d < 1)$

On no loss: $W_i = W_{i-1} + u$

We could solve the above as a stochastic difference equation but let's try to finesse it. Assume there's an equilibrium, $\overline{W}$. At equilibrium, the ups must cancel the downs. If the loss probability is p, there are p downs for every $1 - p$ ups. I.e., $(1 - p)/p$ ups for each down. For ups and downs to cancel, we must have

$$\overline{W} - d\overline{W} = \frac{1-p}{p}\,u$$

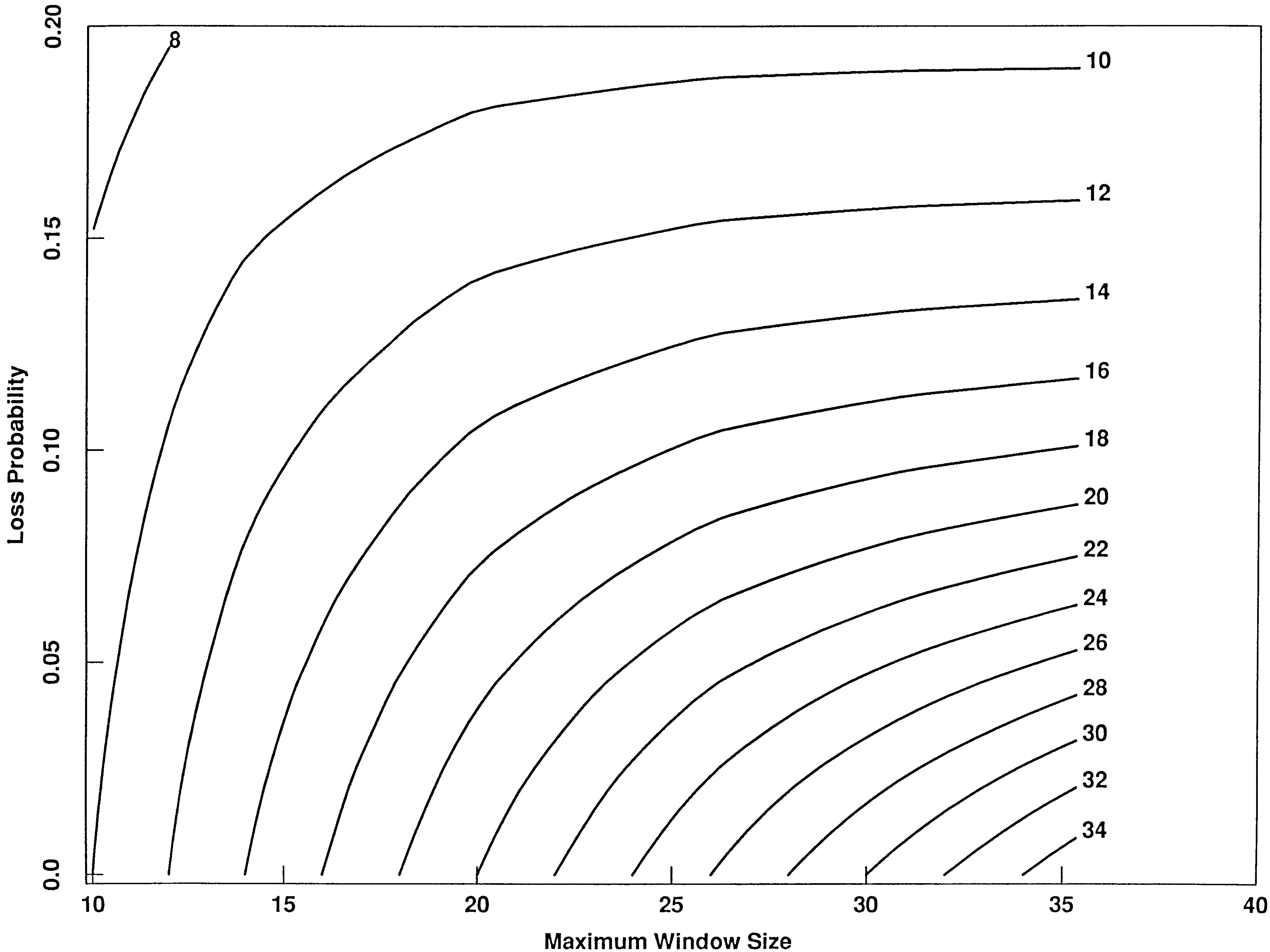
or

$$\overline{W} = \frac{1-p}{p}\,\frac{u}{1-d}$$

For us, $d = .5$ and $u = 1$ so

$$\overline{W} = 2\,\frac{1 - p}{p}$$

E.g., if we let XTCP chose its own window size and run it over a network with a 1% loss rate, no congestive loss and enormous buffer capacity, the window size will average 198 packets.



Loss Probability
0.20
0.15
0.10
0.05
0.0
10
15
20
25
30
35
40
Maximum Window Size
8
10
12
14
16
18
20
22
24
26
28
30
32
34

6.20 Cray TCP Performance—Dave Borman, Cray Research

TCP/IP Performance in the UNICOS Operating System

David A. Borman

Cray Research, Inc.
Networking and Communications
1440 Northland Drive
Mendota Heights, MN 55120

Original code:

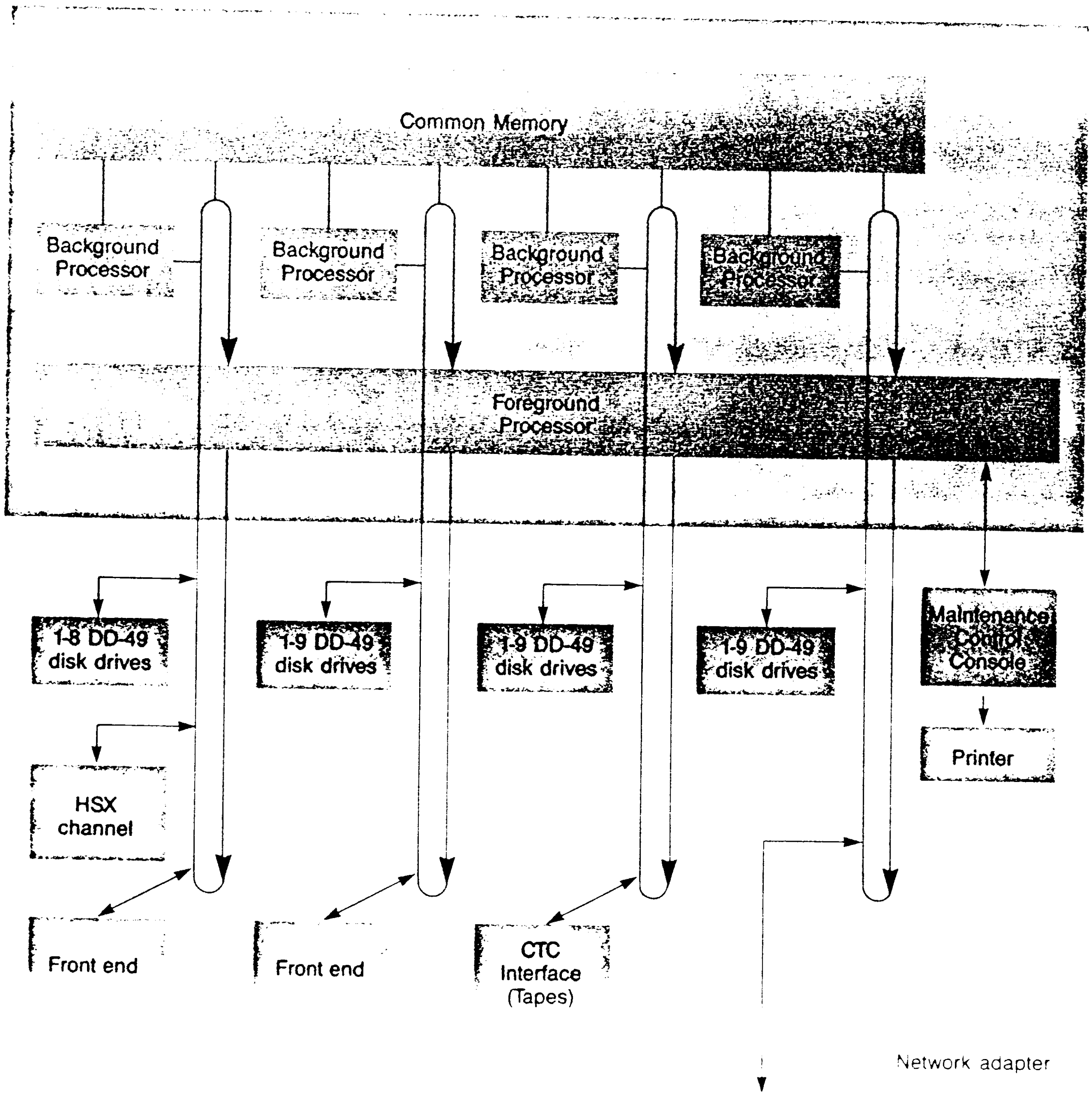
- Wollongong port of 4.2BSD to System V, ported to Cray 2 UNICOS OS.
- Checksum routine written in C, character oriented.
- Driver had 2 5K buffers

  ⇒ 1 for outgoing messages.

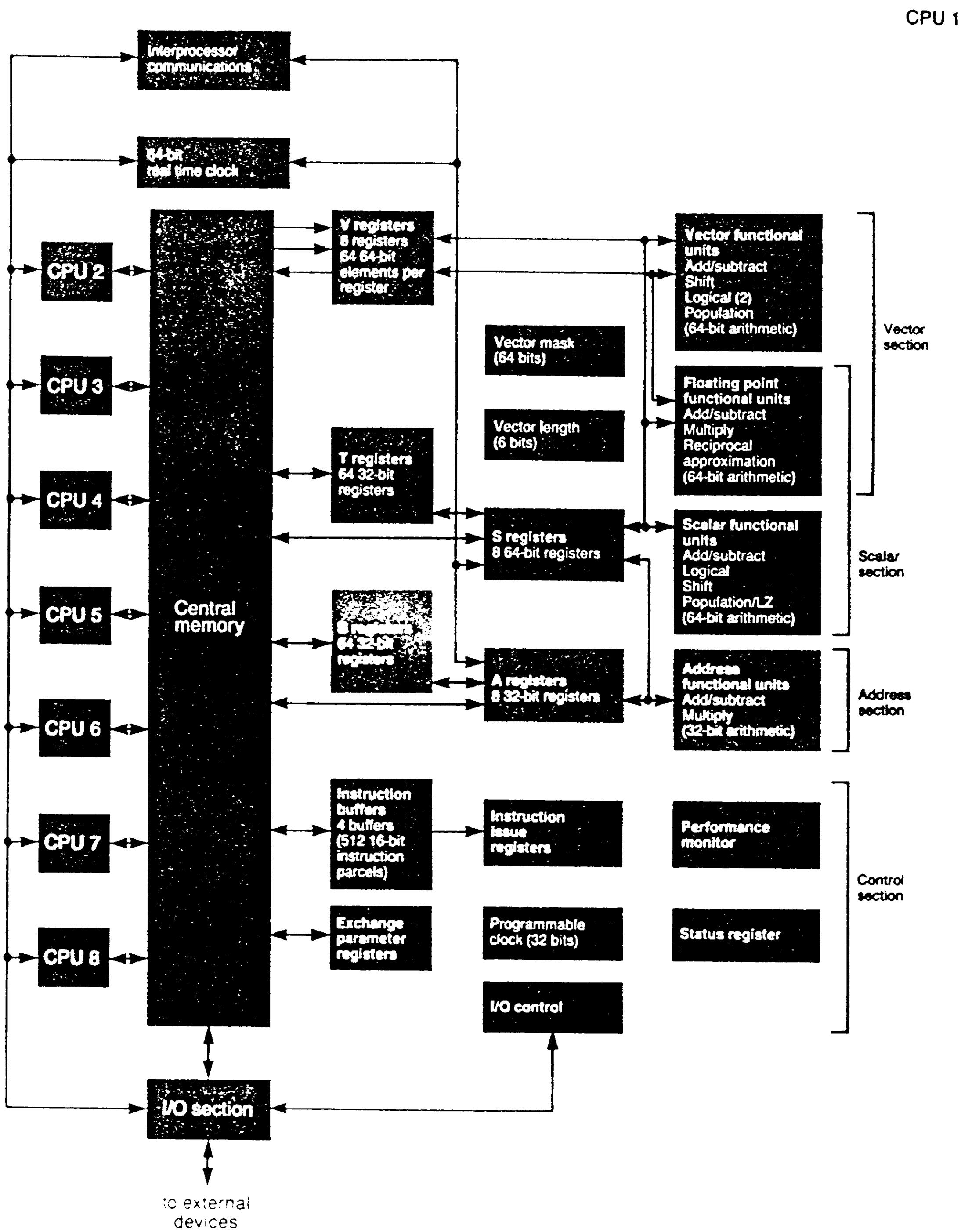
  ⇒ 1 for incoming messages.

  ⇒ Data was copied to/from mbuf chains.
- Mbufs were 1K long, with 4k external data areas.
- NSC HYPERchannel was the only medium available.
- HY driver on Cray 2 had no retries.

Sample CRAY-2 four-processor system configuration



CRAY Y-MP system organization



CRAY

Problems:

- Cray computers are word oriented, any character pointers are done in software, and thus quit slow.
- The system did not deal with running out of mbufs. (Usually caused a panic or crash)
- One busy remote adaptor could cause packets to be dropped, and tie up the local adaptor.
- NSC adaptor had problems with > 4K transfers.

Initial Fixes:

- Many known fixes to 4.2BSD were applied.
- Checksum routine was re-written to be word oriented, and then the inner loop was hand coded in CAL.
- Driver was expaned to have 3 incoming and outgoing buffers.
- Retry code for HY driver was added on Cray 2
- Fix code so that running out of mbufs no longer causes crashes.
- Fix TCP reassembly queue to do compaction, to keep from running out of mbufs.

Later work:

- Mbuf code was rewritten.
  - ⇒ Array of headers and 1K data areas.
  - ⇒ 1-1 mapping between headers and data
  - ⇒ Several mbufs can be linked together to form larger contiguous memory segments.
  - ⇒ Allocation/deallocation similar to V7 memory scheme.
- Static buffers in driver were removed, mbufs are now allocated on the fly.
  - ⇒ Eliminates copy on input
  - ⇒ Usually eliminates copy on output.
- Buffer headers were still static, hence only 3 input and 3 output packets allowed at any given time.
- Added dynamic buffer headers, allows up to 20 packets per interface to be queued up for output.

Current work:

- Using 4.3BSD + Van Jacobson code as base + local mods.
- Mbuf code keeps queues of mbufs of various sizes for fast allocation/deallocation.
  - ⇒ V7 scheme works ok for small mbufs (4K and less), but not for large mbufs (16K-64K)
- Sockets created by accept() inherit send/recv buffer sizes from socket that accept is being done on.
  - ⇒ Only have to reset buffer sizes once.
  - ⇒ MAXSEG is limited to 50% of receive buffer.

Need to do:

- Garbage collection of mbufs.
  ⇒ Go through all current active mbufs and truncate them, freeing up unused portions.
- Possibly eliminate dtom() and rewrite of mbuf code again.
- Have socket layer know about MTU of connection.
- Make TCP code biased to send data on mbuf boundries.
- Vectorize checksum routine
- Make code work with large buffers and large read/writes.
- Add TCP window scaling option
- Use .5Mbyte window, 64K MTU

- HSX transfer rate
  - ⇒ 75 nanosec/word
  - ⇒ 230 usec/24K block
- HSX User to User RTT: 860 usec
  - ⇒ Assume 430 usec one way
  - ⇒ 430 + 230 usec = 660 usec for transfer
  - ⇒ 2166 - (1210 + 660) = 296 usec (~70000 clocks) not yet accounted for.

```
Print screen:
CRAY-2 S/N 1 mendota heights                                   SCC 4.0.0-8222

UNICOS    505064 SECDED errors                  09:48:57 Sat Feb 27, 1988
<ASCII terminal keys> Esc PrtSc ^Home Alt-1

P             3c
S  23 37 00 0 00

1  System console.              Transfer file: file
  03435600       in$cksum$ox$prog
  03436000       in@cksum
  03440550       in$cksum$ox$strn
  05703240       in$cksum$ox$data
  05705360       ipcksum
  05706560       tcpcksum
  05710320       udpcksum
 013322360       in$cksum$obss
# ./mcli -tcp -f -kb 128k snql-hsxl 200 24k
Transfer: 200*24576 bytes from           to snql-hsxl
          Real   System           User           Kbyte    Mbit(K^2)  mbit(1+E6)
  write  0.5730  0.2942 (51.3%)  0.0049 ( 0.8%) 8376.96   65.445     68.624
   read  0.5870  0.1694 (28.9%)  0.0160 ( 2.7%) 8177.17   63.884     66.987
    r/w  1.1600  0.4635 (40.0%)  0.0208 ( 1.8%) 8275.86   64.655     67.796
    72:    196  14568:      1  22488:      1  22560:      1
 24576:    197
# _
```

user to kernel 24K checksum: on MTU 24K

user to kernel 24K checksum: off MTU 24K


```
Print screen:
CRAY-2 S/N 1 mendota heights                                   SCC 4.0.0-8222

UNICOS    505064 SECDED errors                  09:49:44 Sat Feb 27, 1988
<ASCII terminal keys> Esc PrtSc ^Home Alt-1

P             3c
S  23 37 00 0 00

1  System console.              Transfer file: file
 013322360       in$cksum$obss
# ./mcli -tcp -f -kb 128k snql-hsxl 200
Transfer: 200*24576 bytes from
          Real   System           User
  write  0.5730  0.2942 (51.3%)  0.0049
   read  0.5870  0.1694 (28.9%)  0.0160
    r/w  1.1600  0.4635 (40.0%)  0.0208
    72:    196  14568:      1  22488:
 24576:    197
```


```
# ./mcli -tcp -f -kb 128k snql-hsxl 200 24k
Transfer: 200*24576 bytes from           to snql-hsxl
          Real   System           User           Kbyte     Mbit(K^2)  mbit(1+E6)
  write  0.4520  0.1756 (38.9%)  0.0071 ( 1.6%) 10619.47   82.965     86.995
   read  0.4720  0.1509 (32.0%)  0.0162 ( 3.4%) 10169.49   79.449     83.308
    r/w  0.9240  0.3265 (35.3%)  0.0232 ( 2.5%) 10389.61   81.169     85.112
    72:    198  10320:      1  24576:    199
# _
```

```
rint screen:
RAY-2 S/N 1 mendota heights
                                                           SCC 4.0.0-8222
NICOS    505064 SECDED errors
ASCII terminal keys> Esc PrtSc ^Home Alt-1           18:05:45 Fri Feb 26, 1988

            3c
  23 37 00 0 00

  System console.              Transfer file: file
   r/w  0.7360  0.1372 (18.6%)  0.0030 ( 0.4%) 13043.48 101.902   106.852
49152:    100
 ./mcli -tcp -f -kb 128k localhost 100 64k
ransfer: 100*65536 bytes from            to localhost
          Real   System            User            Kbyte   Mbit(K^2) mbit(1+E6)
 write  0.4610  0.1142 (24.8%)  0.0014 ( 0.3%) 13882.86 108.460   113.728
  read  0.4720  0.0467 ( 9.9%)  0.0017 ( 0.4%) 13559.32 105.932   111.078
   r/w  0.9330  0.1609 (17.2%)  0.0031 ( 0.3%) 13719.19 107.181   112.388
65536:    100
 ./mcli -tcp -f -kb 256k localhost 200 128k
ransfer: 200*131072 bytes from            to localhost
          Real   System            User            Kbyte   Mbit(K^2) mbit(1+E6)
 write  1.7730  0.3785 (21.3%)  0.0029 ( 0.2%) 14438.80 112.803   118.283
  read  1.7730  0.1533 ( 8.6%)  0.0034 ( 0.2%) 14438.80 112.803   118.283
   r/w  3.5460  0.5319 (15.0%)  0.0063 ( 0.2%) 14438.80 112.803   118.283
31072:    200
 _
```

checksum: on   MTU 32K

user to kernel: 4K   checksum: on   MTU 24K

```
Print screen:
RAY-2 S/N 1 mendota heights
                                                           SCC 4.0.0-8222
NICOS    505064 SECDED errors                    09:43:52 Sat Feb 27, 1988
ASCII terminal keys> Esc PrtSc ^Home Alt-1

            4a
  23 37 00 0 00

  System console.              Transfer file: file
 ./netstat -i
ame  Mtu   Network    Address      Ipkts   Ierrs Opkts   Oerrs Collis
y0*  4144  none       none         0       0     0       0     0
yl*  4144  none       none         0       0     0       0     0
me2* 16432 none       none         0       0     0       0     0
me3* 16432 none       none         0       0     0       0     0
sx4  24688 101        snql-hsx     204     0     202     0     0
sx5  24688 101        snql-hsx2    202     0     204     0     0
o0   32808 loopback   localhost    7987    0     7987    0     0
 ./mcli -tcp -f -kb 128k snql-hsx1 200 24k
ransfer: 200*24576 bytes from            to snql-hsx1
          Real   System            User            Kbyte   Mbit(K^2) mbit(1+E6)
 write  0.6320  0.3679 (58.2%)  0.0034 ( 0.5%) 7594.94  59.335    62.218
  read  0.6510  0.1910 (29.3%)  0.0162 ( 2.5%) 7373.27  57.604    60.402
   r/w  1.2830  0.5589 (43.6%)  0.0196 ( 1.5%) 7482.46  58.457    61.296
   72:    197  15648:      1  19320:      1  24576:    198
 _
```

6.21 DCA Protocol Testing Laboratory—Judy Messing, Unisys

DCA PROTOCOL LABORATORY

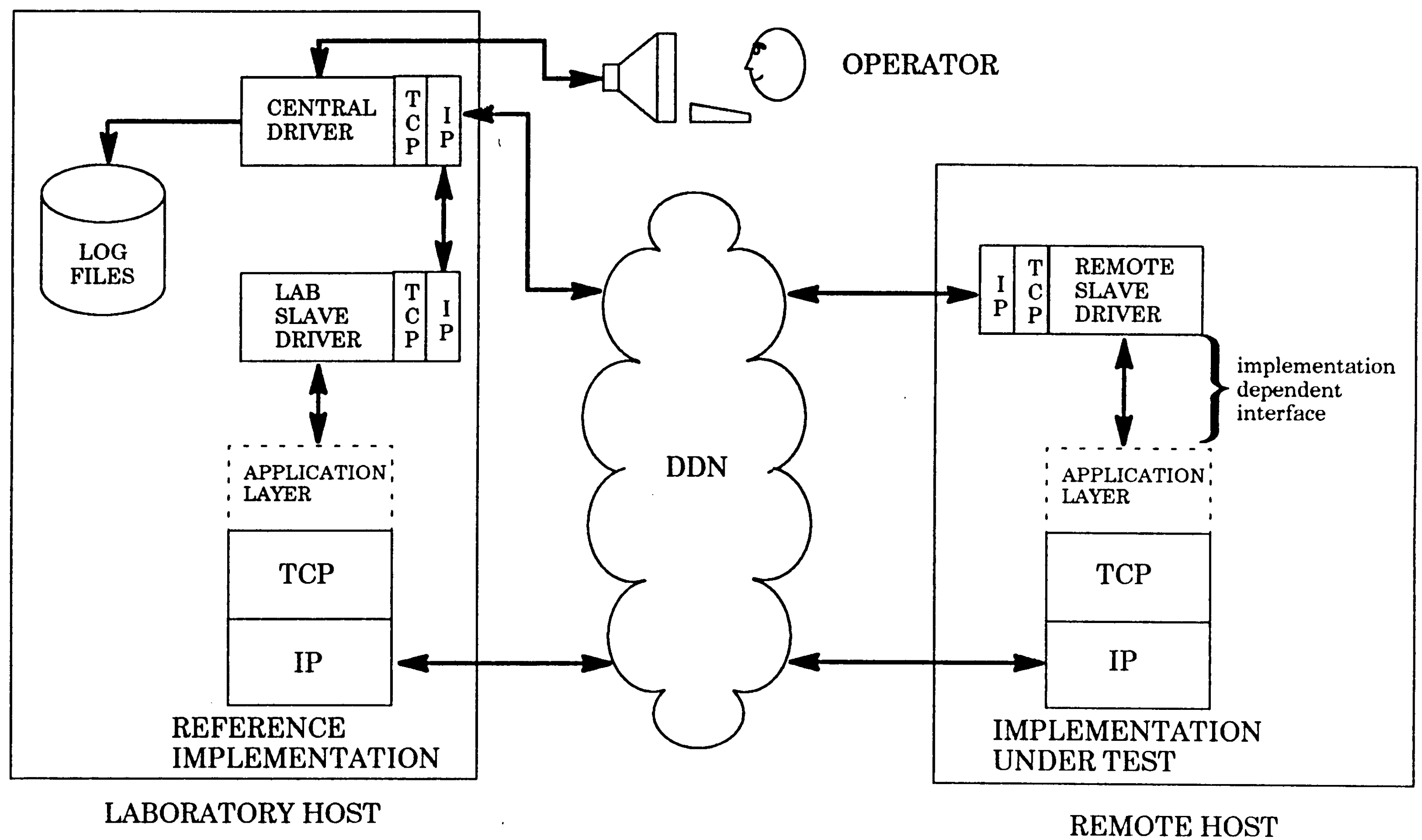
MIL-STD FUNCTIONAL TESTING

- *IMPLEMENTS REQUIRED SERVICES CORRECTLY*
- *IMPLEMENTS CORRECTLY ANY OPTIONAL SERVICES IT PROVIDES*
- *HANDLES ERRONEOUS INPUT WITHOUT ILL EFFECT*

TESTING PHILOSOPHY

- ***TEST ONE FUNCTION AT A TIME***
- ***PROVIDE REPEATABLE TESTS***
- ***PROVIDE AUDIT TRAIL OF PROTOCOL EXCHANGE***
- ***REPORT RESULTS PRECISELY IN SEVERAL CATEGORIES***
  - *OK*
  - *PROBLEM*
  - *OBSERVATION*
  - *INCONCLUSIVE*

Generic Testing Architecture



REFERENCE IMPLEMENTATIONS

- ***INSTRUMENTED TO SUPPORT PROTOCOL TESTING***
  - *TCP/IP MODIFIED FROM ULTRIX 1-1 RESIDE IN KERNEL*
  - *TELNET, FTP AND SMTP MODIFIED APPLICATION LEVEL PROCESSES*

CENTRAL DRIVER

- ***FUNCTIONS***
  - *TEST SYNCHRONIZATION*
  - *DRIVER COMMUNICATION*
  - *ANALYSIS OF TEST RESULTS*
  - *LOGGING*
- ***COMPILED SCENARIO***

DRIVERS

- ***LAB SLAVE DRIVER***
  - *INTERPRETS AND ACTS ON COMMANDS FROM CENTRAL DRIVER*
  - *RETURNS IUT'S PROTOCOL RESPONSES TO CENTRAL DRIVER*
  - *CAN SET UP TEST CONDITIONS ON CENTRAL DRIVER REQUEST*

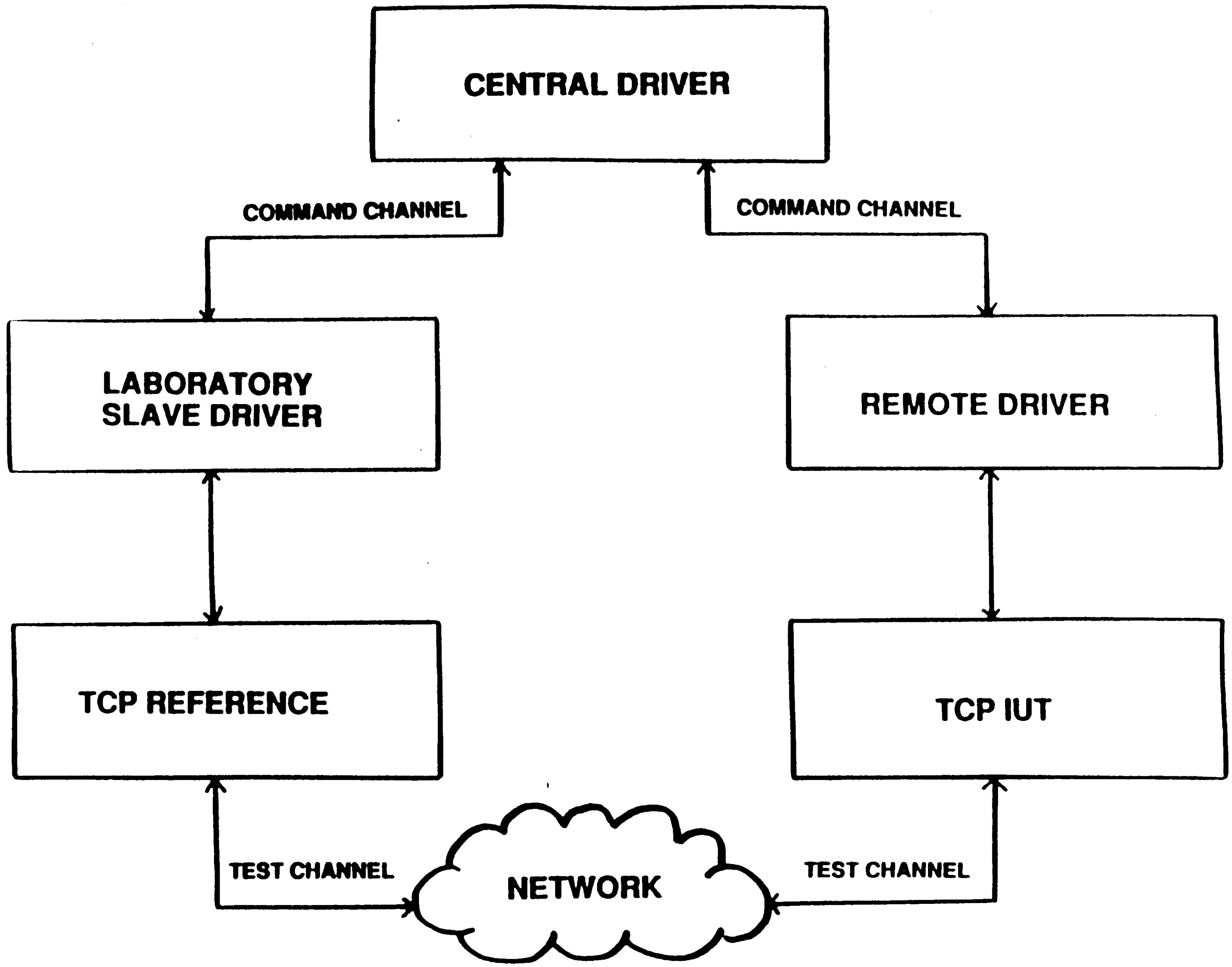
- ***REMOTE DRIVER***
  - *INTERPRETS AND ACTS ON COMMANDS FROM CENTRAL DRIVER*
  - *RETURNS IUT'S PROTOCOL RESPONSES TO CENTRAL DRIVER*
  - *PROVIDED BY IUT VENDOR*

TCP TEST SYSTEM OBJECTIVES

- ***TEST CONFORMANCE TO MIL-STD 1778***
- ***TEST IMPLEMENTATION POLICIES AND STRATEGIES***
  - *OBSERVATIONS FOR INTEROPERABILITY*
  - *OBSERVATIONS FOR QUALITY*

McLean Research Center

TCP TEST SYSTEM COMPONENTS



UNISYS

TCP TEST REFERENCE IMPLEMENTATION

IMPLEMENTS ALL MIL-STD 1778 FUNCTIONS VISABLE TO PEER

PROVIDES ALL MIL-STD 1778 SERVICE RESPONSES

ENHANCEMENTS ADDED TO ULTRIX 1-1

- ***ADDED***
  - *USER ABORT*
  - *ACTIVE OPEN WITH DATA*
  - *USER PUSH*
  - *USER SET CONNECTION PRECEDENCE AND SECURITY*
  - *PRECEDENCE AND SECURITY PROCESSING*
  - *TCP SERVICE RESPONSES*
- ***URGENT SERVICE***
  - *MORE THAN ONE BYTE*
  - *MULTIPLE CONNECTIONS*
- ***REMOVED: KEEP ALIVE TIMER***

CREATES REPEATABLE PROTOCOL EVENTS

- ***BOTTOM UP TESTING***
  - *INVALID SEGMENTS*
- ***TEST IUT USING***
  - *DATA PACKAGING*
  - *ACK STRATEGY*
  - *OFFERED WINDOW*
- ***PRESPECIFY WHETHER AND WHEN TO CORRECT***

CREATES REPEATABLE SIMULATED NETWORK PROBLEMS

- *DELAY*
- *LOSS*
- *DAMAGE*
- *DUPLICATION*
- *MISDIRECTION*
- *SMALL PACKET SUBNET*

COLLECTS PROTOCOL DATA

- ***MAINTAINS RECORD OF EVERY SEGMENT***
  - *INCOMING AND OUTGOING CHRONOLOGICAL ORDER*
  - *INCOMING (IUT) ERRORS ARE FLAGGED*
- ***RECORD ALSO CONTAINS***
  - *TCP HEADER*
  - *DATA LENGTH*
  - *IP PRECEDENCE AND SECURITY*
  - *CONTEXT INFORMATION*

A TCP SEGMENT RECORD

ref.3 <== dir 2 oflow 0 len 0

ref.3 <== TCP HDR: src port 3100 dst port 3000

ref.3 <== TCP HDR: seq 24721665 ack 11369090 off 6 flags 12

ref.3 <== TCP HDR: window 4096 checksum 2788 urg_ptr 0

ref.3 <== mtu 1024 prec 0 Sec:(type 0 len 0 lev 0 auth 000)

ref.3 <== win2 8192 rxmtf 0 withackf 0 fillwin 0

ref.3 <== zerof 0 ptr 0 timestamp 0

TCP LAB SLAVE DRIVER

- ***INTERPRETS COMMANDS FROM CENTRAL DRIVER***
  - *TO INITIATE TCP SERVICE REQUESTS*
  - *TO SET TCP TEST CONTROL PARAMETERS*
- ***MAINTAINS TCP SERVICE STATE MACHINE***
  - *PASSES ALL TCP SERVICE RESPONSES TO CENTRAL DRIVER*
- ***SUPPORTS UP TO 144 TEST CONNECTIONS***
- ***SELECTS AND FORMATS PROTOCOL DATA FOR CENTRAL DRIVER***
  - *SIMPLE DATABASE COMMANDS*

TEST COVERAGE AT BASIC LEVEL

Unspecified Passive Open Request

Active Open Request

Send Request and Deliver Response

Local Connection Name

Closing Handshake

Graceful Closing Function

Fully Specified Passive Open Request

Active Open With Data Request

Active Open With Data by Peer

Port Number Range

Abort Request

Remote Abort Service Response

Abort Function

COVERAGE OF TESTS AT BASIC LEVEL (CONT.)

Status Request

Allocate Request

Data Integrity Mechanisms

--Correction of Out of Order, Lost, Overlapping, Duplicate Data

Checksum Mechanism

Sequence Numbering Mechanism

Multiplexing Mechanisms

Precedence Level

Security Mechanism

COVERAGE OF TESTS AT FULL AND QUALITATIVE LEVEL (CONT.)

ULP Timeout Service

Push Service

Urgent Service

Reset Mechanisms

Maximum Segment Size Option

Retransmission Mechanism and Policy

Flow Control Window Mechanism and Policy

Maintenance of Large Number of Connections

EXAMPLE - CHECKSUM TEST DIALOG

.
. (Connection opened)
.

D_CMD to connection ref.2 with driver primitive SET_TEST
ref.2 ==> (cmd 19) 103 11 0 3 0 0
ref <== ACK

REF sends data with one segment having a bad checksum
the checksum is corrected on retransmission
D_CMD to connection ref.2 with driver primitive GEN_SND_TEXT
ref.2 ==> (cmd 1) x 1024
ref <== ACK
P_CMD to connection ref.2 with proto primitive SEND
ref.2 ==> (cmd 5) 103 0 0 0 0
ref <== ACK

Waiting for data...waiting...

.
. (Data delivered and connection closed)
.

Analyze REF report data to determine IUT behavior
D_CMD to connection ref.2 with driver primitive SHOW

```
ref.2 ==> (cmd   9) 103 IO F_CHKSUM SEQ LEN ACKNR
ref <== ACK
ref.2 <== STARTSHOW 103
ref.2 <==  :I SEQ=56430593 LEN=0 ACKNR=0
ref.2 <==  :O SEQ=56430657 LEN=0 ACKNR=56430594
ref.2 <==  :I SEQ=56430594 LEN=0 ACKNR=56430658
ref.2 <==  :O SEQ=56430658 LEN=512 ACKNR=56430594
ref.2 <==  :I SEQ=56430594 LEN=0 ACKNR=56431170
ref.2 <==  :O F_CHKSUM SEQ=56431170 LEN=512 ACKNR=56430594
ref.2 <==  :O F_CHKSUM SEQ=56431170 LEN=512 ACKNR=56430594
ref.2 <==  :O F_CHKSUM SEQ=56431170 LEN=512 ACKNR=56430594
ref.2 <==  :O F_CHKSUM SEQ=56431170 LEN=512 ACKNR=56430594
ref.2 <==  :O SEQ=56431170 LEN=512 ACKNR=56430594
ref.2 <==  :I SEQ=56430594 LEN=0 ACKNR=56431682
ref.2 <==  :I SEQ=56430594 LEN=0 ACKNR=56431682
ref.2 <==  :O SEQ=56431682 LEN=0 ACKNR=56430595
ref.2 <==  :O SEQ=56431682 LEN=0 ACKNR=56430595
ref.2 <==  :I SEQ=56430595 LEN=0 ACKNR=56431683
ref.2 <==  :ENDSHOW
```

OK: IUT ONLY ACKNOWLEDGED SEGMENTS WITH GOOD CKSUMS

AN IP DATAGRAM REPORT RECORD

ref.1 <== dir 1 oflow 0 type 1

ref.1 <== IP HDR: version 4 IHL 5 TOS 0 len 36 Id 5537

ref.1 <== IP HDR: frag_off 8192 TTL 15 prot 6 cksum 61853

ref.1 <== IP HDR: src 12c1fc0 dest 22c1fc0 optlen 0;

ref.1 <== TCP HDR: src port 3851 dst port 3951

ref.1 <== TCP HDR: seq 9803073 ack 0 off 5 flags 2

ref.1 <== TCP HDR: window 0 cksum 9040 urg_ptr 0;

ref.1 <== mtu 0 rxmtf 0 ptr 8000190b

ref.1 <== iptime 68780180 tcptime 0;

TEST CONDUCTED WITH UNISYS WEST COAST RESEARCH CENTER

- ***HOST ON AN ETHERNET GATEWAYED TO THE ARPANET***
  - *INTERNET ADDRESS SET IN TOOLS AT RUNTIME*
- ***LONG, VARIABLE NETWORK DELAYS***
- ***ALL COMPONENTS ROBUST***
- ***TEST SCENARIOS RAN WELL***
  - *IDENTIFIED IUT PROBLEMS CORRECTLY*
  - *RESULTS SAME REGARDLESS OF NETWORK CONDITIONS*