DRAFT

Proceedings of the

April 22-24, 1987

Internet Engineering Task Force

Phillip Gross

July 1987

SIXTH IETF

**This document was prepared for authorized distribution.
It has not been approved for public release.**

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

DRAFT

TABLE OF CONTENTS

1.0 Introduction

The Internet Engineering Task Force met at Bolt, Beranek and Newman (50 Moulton Street, Cambridge Massachusetts) for the three days of April 22 through April 24, 1987. The meeting was hosted by Bob Hinden.

The second day and the morning of the third day were devoted to a joint meeting with the ANSI X3S3.3 Network and Transport Layer standards group. Lyman Chapin (Data General, X3S3.3 Chair) was instrumental in coordinating the agenda for the joint portions of the meeting.

Bob Stine (MITRE) is gratefully acknowledged for his assistance in producing the meeting notes in Section 4. Various working group Chairs contributed to the reports in Section 5. Individual contributions are noted there.

2.0 Attendees

2.1 IETF Affiliates (47)

Name	Organization	Email Address
Bosack, Len	Cisco Systems	bosack@hplabs.hp.com
Boule, Rich	Proteon	rfb@proteon.com
Braden, Bob	ISI	braden@isi.edu
Braun, Hans-Werner	Univ of Michigan	hwb@mcr.umich.edu
Brescia, Mike	BBNCC	brescia@bbn.com
Brim, Scott	Cornell Univ	swb@devvax.tn.cornell.edu
Callon, Ross	BBN	rcallon@suran.bbn.com
Chao, John	BBNCC	jchao@bbn.com
Coltun, Rob	MITRE	rcoltun@gateway.mitre.org
Corrigan, Mike	DoD	corrigan@ddn3.arpa
Fedor, Mark	Cornel Univ	fedor@devvax.tn.cornell.edu
Feinler, E. (Jake)	SRI (NIC)	feinler@sri-nic.arpa
Garcia-Luna, Jose	SRI	garcia@istc.sri.com
Gross, Phill	MITRE	gross@gateway.mitre.org
Hastings, Gene	PSC	hastings@morgul.psc.edu
Hedrick, Charles	Rutgers Univ	hedrick@rutgers.edu
Heker, Sergio	JVNCC	heker@jvnc.csc.org
Hinden, Robert	BBN	hinden@bbn.com
Jacobsen, Ole	ACE	ole@sri-nic.arpa
Jacobson, Van	LBL	van@lbl-csam.arpa
Karels, Mike	UC-Berkeley	karels@berkeley.edu
Kingston, Doug	BRL	dpk@brl.arpa
LaBarre, Lee	MITRE	cel@mitre-bedford.arpa
Lazear, Walt	MITRE	lazear@gateway.mitre.org
Lottor, Mark	SRI (NIC)	mkl@sri-nic.arpa
Love, Paul	San Diego Supercmp	loveep@sdsc-sds.arpa
Mallory, Tracy	BBNCC	tmallory@bbn.com
Mamakos, Louis	Univ of MD	louie@trantor.umd.edu
Mantiply, Stan	Ungermann Bass	mantiply%engr.ub.com@relay.cs.net
McCloghrie, Keith	ACC	kzm@acc-sb-unix.arpa
Medin, Milo	NASA/Ames	Medin@orion.arpa
Mockapetris, Paul	ISI	pvm@isi.edu
Morris, Donald	NCAR	morris@scdsw1.ucar.edu
Moy, John	Proteon	jmoy@proteon.com
Nakassis, Tassos	NBS	nakassis@icst-ecf.arpa
Partridge, Craig	BBN Labs	craig@sh.cs.net
Perkins, Drew	CMU	ddp@andrew.cmu.edu
Petry, Michael	U of MD	petry@trantor.umd.edu
Reilly, Brendan	NSF	reilly@note.nsf.gov
Rodriguez, Jose	Unysis	jrodrig@edn-vax.arpa
Schoffstall, Martin	RPI	schoff@csv.rpi.edu
Stahl, Mary	SRI (NIC)	STAHL@SRI-NIC.ARPA

Stine, Robert	MITRE	stine@gateway.mitre.org
StJohns, Mike	DCA(B612)	stjohns@sri-nic.arpa
Su, Zaw-Sing	SRI	zsu@istc.sri.com
Tasman, Mitch	BBNCC	mtasman@bbncct.arpa
Tontonoz, James	DCA/DCEC	tontonoz@edn-unix.arpa
Topolcic, Claudio	BBN Labs	topolcic@bbn.com
Zhang, Lixia	MIT	lixia@xx.lcs.mit.edu

2.2 X3S3.3 Affiliates (31)

Name	Organization	Email Address
Carneal, Bruce	Mentat	!ihnp4!mentat!blc
Chapin, Lyman	Data General	chapin@a.isi.edu
Chiu, Da-Ming	DEC	chiu%erlang.dec@decwrl.dec.com
DiCecco, Steve	Codek	
Greenwood, Jerry	Data General	
Gruchersky, Steven	Unisys	
Guyer, Kay	Mentat	!ihnp4!mentat!kay
Hemrick, Chris	Bellcore	cfh@sabre.bellcore.com
Hagens, Rob	U of Wisconsin	hagens@rsch.wisc.edu
Hall, Nancy	U of Wisconsin	nhall@rsch.wisc.edu
Ilnicki, Ski	Hewlett-Packard	
Jain, Raj	DEC	jain@marlboro.dec.com
Katz, Dave	U of Michican	dave_katz@um.cc.umich.edu
Kelly, Dave	U.S. Navy	
LaBarre, Lee	MITRE	cel@mitre-bedford.arpa
Langdon, Steve	Amdahl	sjl@amdahl.amdahl.com
Lemon, John	Tandem Computers	
Merala, Mark	Tandem Computers	
Mills, Kevin	NBS	mills@icst-osi.arpa
Montgomery, Doug	NBS	dougm@icst-osi.arpa
Nakassis, Tassos	NBS	nakassis@icst-ecf.arpa
Obert, Carl	NCR	
Oran, David	DEC	oran%oran.dec@decwrl.dec.com
Piscitello, David	Unisys	piscitello@a.isi.edu
Ramakrishnan, K.	DEC	rama%erlang.dec@decwrl.dec.com
Reddy, Nelluri	CDC	
Stern, Ed	Foxboro	
Su, Zaw-Sing	SRI	zsu@sri-ucl.arpa
Taylor, Ed	IBM	
Trinchieri, Mario	Honeywell Bull	trinchieri.hdsa@kis_phoenix_multics.arpa
Tsuchiya, Paul	MITRE	tsuchiya@gateway.mitre.org

2.3 Other (10)

Name	Organization	Email Address
Abram, Len	BBNCC	labram@bbn.com
Atlas, Stephen	BBNCC	satlas@bbn.com
Caloccia, William	BBNCC	caloccia@bbn.com
Davin, Chuck	Proteon	jrd@proteon.com
Greifner, Michael	DCA	Greifner@edn-vax
Kaufman, David	Proteon	dek@proteon.com
Kullberg, Alan	BBNCC	akullberg@bbn.com
Little, Mike	M/A-COM	little@macom2.arpa
Park, Philippe	BBN Labs	ppark@bbn.com
Westcott, Jill	BBN	westcott@bbn.com

3.0 Final Agenda

Wednesday, April 22

Morning

- Welcome, Task Force Reorganization	Gross (MITRE)
- Enhanced AHIP	StJohns (DDN)
- BBN Report	Hinden/Gardner (BBN)
- Progress Report on	
- Congestion Control Simulation	Stine (MITRE)
- Arpanet Performance Measurement	Gross (MITRE)
- TCP Performance Enhancement	Jacobson (LBL)

Afternoon

- Gateway Monitoring	Partridge (BBN)
- Management Architecture	LaBarre (MITRE)
- Internet Problem Descriptions	Groups

Thursday, April 23

Joint X3S3.3/IETF Meeting on Gateways and Routing

Morning

- Welcome	Chapin/Gross
- IETF status/overview	Gross (MITRE)
- FCCSET report	Gross (MITRE)
- ANSI/ISO status/overview	Chapin (Data General)
- ANSI routing architecture	Tsuchiya (MITRE)
- NSF gateway requirements	Braden (ISI)
- Routing Directions at SRI	Su and Garcia (SRI)
- NBS Routing Proposal	K. Mills (NBS)

Afternoon

- Burroughs Integrated Adaptive Routing	Piscitello (Unisys)
- DECnet Phase V Routing	Oran (DEC)
- Discussion & questions	

Friday, April 24

Joint X3S3.3/IETF Meeting on Gateways and Routing (Con't)

Morning

- SPF Routing in the Butterfly Gateways	Mallory (BBN)
- Other Advanced Routing Work at BBN	Gardner (BBN)
- Congestion Avoidance	Jain, et. al. (DEC)
- Adjourn Joint Session	

Afternoon

- Parallel IETF Working Groups
    - EGP2 RFC (Petry)
    - Name Domain Planning (Kingston)
    - Performance and Congestion Control (Stine)
    - Gateway Monitoring (Partridge)
    - NSF Routing (Hedrick)
    - Misc. Issues (StJohns)

4.0 Meeting Notes

4.1 Wednesday, April 22

4.1.1 AHIP Enhancements: Mike StJohns (DCA-DDN)

Mike StJohns presented a report on the coming enhancements for AHIP. The new AHIP will allow growth in the subnet. In addition, it will provide logical addressing functionality, and subnet congestion feedback. It is also expected that type of service (TOS) routing will be provided; the interface specification for TOS routing is under development. The enhanced AHIP will replace 1822L.

4.1.2 BBN report: Bob Hinden, Marianne Gardner (BBN)

Bob Hinden and Marianne Gardner reported on current status in the Internet and the ARPANET. Hinden noted that the Internet has been growing rapidly: from 160 nets in January, the Internet grew to 211 nets in April. Accompanying the growth, there has been evidence of EGP fluctuations(????).

On BBN's current gateway work on the core gateways, Hinden reported that implementing IP reassembly should be completed by mid-June. Other progress in the core system is that Butterflies gateways have been installed as "mail bridges"—the gateways between the ARPANET and the MILNET.

Marianne Gardner reported on ARPANET performance, especially the high network delays seen in late 1986. Gardner reported that "the performance crisis has passed." A source of the problem was unstable routes in the ARPANET. The thrashing was due primarily to inadequate cross-country trunking, which led to congestion, resulting in high delays, which in turn caused route recomputation.

4.1.3 Congestion Control Simulation: Robert Stine (MITRE)

to be supplied

4.1.4 Arpanet Performance Measurements: Phill Gross (MITRE)

to be supplied

4.1.5 TCP Enhancements: Van Jacobson

Van Jacobson proposed two improvements to TCP implementations which would improve Internet performance. He also discussed the manner in which TCP traffic tends to organize itself in a way that is detrimental to performance.

4.1.5.1 Slow Start Algorithm. Jacobson noted a problem that can occur during bulk data transfers, when large windows are used. In this situation, Jacobson reported that he has observed TCP performing in a stable, highly inefficient fashion, as follows:

1. The sending TCP transmits a large window of data, and then quiesces while awaiting acknowledgements.

2. One or more packets from the interior of the window are dropped.

3. The segments corresponding to the dropped packets time out, and another blast is transmitted.

The above behavior is inefficient for two reasons: TCP is quiescent for long periods, and most segments are transmitted several times.

Jacobson reasoned that TCP would avoid the above blast, wait, and retransmit scenario, if it could only start out operating right. To achieve this, he proposed the use of a "slow start" algorithm for TCP.

The slow start algorithm works by having TCP implementations open their initial send windows gradually, as acknowledgements are received. At the beginning of a connection, a TCP would transmit a single segment Max Segment Size (MSS). Upon receipt of each ack, the send window will be opened by another MSS, up until it is fully opened. Jacobson maintained that utilized window size will actually increase logarithmically over time, since queuing delays in the Internet will tend to clump the acks.

Jacobson reported that when using send windows of 16kb, the slow start algorithm improved throughput by 30%, and reduced retransmissions by a factor of 8. With 4Kb windows, retransmissions were reduced by a factor of 3.

4.1.5.2 Estimating RTT. Another TCP enhancement proposed by Jacobson is the use of Box-Jenkings autoregressive techniques for predicting the roundtrip time (RTT) of TCP segments. The Box-Jenkins models use previous observations, and sometimes previous predictions, for estimation. For example, the autoregressive model of order 2 (AR(2)) uses two previous observations to predict the next observation:

$$x_t = \xi + \phi_1 x_{t-1} + \phi_2 x_{t-2}$$

The weights given to the previous observations and the constant term ξ can be estimated (e.g., by a least squares fit), based on the history of previous observations. These estimations can be performed recursively: a new estimate is a function of the most recent observation and the last estimate.

4.1.5.3 Traffic characteristics. After presenting his proposed TCP enhancements, Jacobson discussed some interesting Internet traffic characteristics that result from the bandwidth mismatch between LANs and the ARPANET. One of these is that observed RTT exhibits sharp increases and gradual declines. (note: another factor here is that few packets are dropped by the ARPANET. Apparently, however, there is an occasional suspension of some sort of service within the system. Acks pile up in queues during these blockages. When the blockage ends, the queues are emptied in a nearly deterministic fashion, so that acks arrive at their destination hot on each other's heels. But, several acks arriving simultaneously at a host will result in the perception that RTT declines at the interval between segment transmissions. For example, if segments are transmitted at times 0, 3, 6, and 9, and their acks arrive at time 20, 21, 22, and 23, then the observed RTT values will be 20, 18, 16, and 14).

Acks also tend to be bunched. An effect of this is that random traffic on the Internet tends to organize itself in a destructive way: TCP connections will begin to transmit in unison. This increases the probability of exceeding a gateway's resources. One way that gateway's may defend themselves from this behavior is to reintroduce randomness in traffic. Fair queuing is a means to accomplish this.

4.1.6 Gateway Monitoring: Craig Partridge (BBN)

Craig Partridge described the status of his work on a High-level Entity Monitoring Program (HEMP).

In the HEMP system, queries will be in AS1 format. A design goal is to keep query processing as simple as possible.

HEMP is at the Applications level. A remaining issue in the HEMP design is the selection of a transport protocol. Since HEMP may require exchanges of high volumes of data, its transport protocol must be reliable. Hence, UDP is unsuitable. TCP, however, imposes a high overhead. HMP and RDP were discussed as candidate transport protocols.

Another issue in the design of HEMP is the use of traps for monitoring network entities. Ill-conceived use of traps could degrade performance of network entities, and also generate copious data. There was also discussion on the need to predefine a set of traps for use by HEMP.

4.1.7 Management Architecture: Lee LaBarre (MITRE)

Lee LaBarre reported on the activities of the newly formed IAB System Working Group. As a basis for developing system management concepts, LaBarre offered a strawman management architecture, in which the standard protocol stack is overlayed with a corresponding management stack. This approach is similar to the ISO management framework.

The immediate tasks that LaBarre has defined for his working group are:

1. Define a system management framework.

2. Define the scope of system management.

3. Specify the management information that the system will collect.

4. Specify a management protocol.

LaBarre said that his group will attempt to form liaisons with other groups, such as Partridge's, working on network management problems. LaBarre's group plans to hold monthly meetings for the next 6 to 12 months.

4.2 Thursday, April 23

4.2.1 Welcome: Chapin/Gross

4.2.2 IETF Overview and FCCSET Report: Phill Gross (MITRE)

to be supplied

4.2.3 ANSI/ISO Overview: Lyman Chapin (Data General)

Chapin described the process of developing standards. He also briefed the relations of the various standards bodies. In addition, he presented a bibliography of pertinent network protocol standards.

4.2.4 Routing Architecture: Paul Tsuchiya (MITRE)

Paul Tsuchiya reported the status of an internet routing architecture under development by X3.S3.3. He described several categories for classifying groups of Intermediate Systems (IS's, a.k.a "gateways"). (??? and ES's???) "Domains" are groups of IS's that use a common routing algorithm. If domains use hierarchical routing, then they are divided into clusters. The hierarchy is of addressing authorities, and does not entail the use of separate, syntactically distinct components in an address that correspond to each authority level.

Another routing concept included in the routing architecture is "dominions," which are autonomous system of IS's. A "common dominion" is a set of dominions that have agreed-upon routing procedures.

4.2.5 Gateway Requirements: Bob Braden (ISI)

to be supplied

4.2.6 Routing Directions at SRI: Zaw-Sing Su, Jose Garcia-Luna (SRI)

Zaw-Sing Su and Jose Garcia-Luna described a routing algorithm that they are developing at SRI. Su began by presenting the motivation for their research.

The two major classes of routing algorithms for long-haul nets or internets are the Bellman-Ford and the Dijkstra algorithms. Bellman-Ford algorithms (a.k.a. distance vector algorithms) share too little information. They are notoriously susceptible to the "count to infinity" when routers go down. The Dijkstra algorithm (a.k.a. link state, SPF) requires tight coupling, since each router must maintain a database of the status of all links in the system. This could result in routers maintaining and exchanging a large amount of information that they never use, and could entail problems for very large, heterogeneous networks.

SRI's goal for its routing algorithm is to find a middle ground, which avoids the count to infinity but only requires lose coupling. Su characterized the algorithm as a "nonhierarchical area scheme."

The algorithm itself was described by Jose Garcia-Luna. He noted that it is similar to the split horizon concept.

4.2.7 NBS Routing Proposal: Kevin Mills (NBS)

to be supplied

4.2.8 Burrough's Integrated Adaptive Routing: David Piscitello (Unisys)

to be supplied

4.2.9 DECnet Phase V Routing: David Oran (DEC)

to be supplied

4.3 Friday, April 24

4.3.1 SPF Routing in the Butterfly Gateways: Tracy Mallory (BBN)

Tracy Mallory described BBN's implementation of Dijkstra's SPF routing algorithm as an Interior Gateway Protocol of the core Butterfly gateways. SPF is a link state protocol: each router maintains a database of all links in the system. As implemented in the Butterflies, each link is assigned a fixed cost.

Between gateways on an internet, a "link" consists of a path through a single network. The existence of a link is determined by neighbor up/down protocol similar to the Neighbor Reachability protocol of EGP. Gateways on the ends of a link engage in a master/slave exchange of "Hello" "I Hear You" messages. Sequence numbers are used to enhance reliability.

When core Butterflies gateways establish links, they exchange link state data bases. As members of the system, each core Butterfly "floods" link state updates upon detection in a topological change, or every 8 minutes.

4.3.2 Other Advanced Routing Work at BBN: Marianne Gardner (BBN)

to be supplied

4.3.3 Congestion Avoidance: Raj Jain

Raj Jain, K. Ramakrishnan, and Da-Ming Chiu described DEC's approach to congestion avoidance for use by hosts and routers in a connectionless network or internet in which the transport protocol uses windowing for flow control.

As an introduction, several issues concerning congestion were presented. Congestion was defined as a network state in which delay increases at a high rate as throughput drops to zero. It was explained that over-engineering is not necessarily a solution to the congestion problem. For example, if very fast switches are used throughout the network, aggregate traffic from several switches could overwhelm the resources of a single switch.

It was maintained that current congestion control algorithms focus on recovery from congestive collapse. The DEC scheme, however, attempts to avoid congestion. It could be used in conjunction with congestion recovery procedures.

As a measurement for congestion, the DEC scheme introduces the application of system power to communications systems; this is defined as average throughput divided by average delay. Congestion can be avoided and throughput maximized if the system is operated at maximum power. This maximum corresponds to the critical "knee" in the relation between delay and system load, in which additional load results in sharp increases in delay.

The high-level goals of DEC's congestion avoidance scheme are:

1. Efficient operation: a high ratio of throughput to delay should be achieved at low overhead.

2. Fairness: users on the same path should experience the same throughput.

3. Responsiveness: as system capacity changes, the offered load should respond.

4. Convergence: as a control system, the congestion avoidance scheme should result in stable, optimal loads.

5. Robustness.

6. Distributed operation.

7. Maximum information entropy: each congestion control message should contain as much information as possible.

8. Simplicity.

In a nutshell, DEC's congestion avoidance scheme operates as follows: each packet in the network has a single bit to indicate congestion status. Network routers set this "cc bit" if their average queue lengths exceed some threshold. Hosts will adjust their receive windows according to the number of packets they receive that have the cc bit set. If a critical density of packets have that bit set, then the receive window will be narrowed. On the other hand, if arrivals of packets with the cc bit set are sparse, then the receive windows will be opened.

The computation of average queue length by the routers is only performed when the system is busy. As for deciding whether a queue length indicates congestion, there are two design alternatives: either a simple threshold may be used, or hysteresis can be induced (i.e., a higher threshold would be used to initiate setting the cc bit than would be used to cease setting it). It was reported that simulations showed power to be maximized with no hysteresis, and an threshold of average queue length at 1.

In the DEC scheme, there were several design alternatives for the window adjustment algorithm. A problem with strictly additive increases and decreases in window size was reported: new users never get a share of the network's bandwidth. The only fair approach found for window management was to use additive increases, and multiplicative decreases for window adjustment. It was reported that simulations show that system oscillation is minimized if windows are increased by 1 segment and decreased by a factor of 0.875.

4.3.4 Working Groups

DEC's congestion control presentation concluded the morning's session and the joint meeting of the IETF and X3S3.3. The final afternoon was devoted to working group meetings. Reports from these meetings are given in the next section.

5.0 Working Group Reports

On the final afternoon of the IETF meeting, Friday April 24th, the following groups met:

Group	Convened by:
- Name Domain Planning	Doug Kingston (BRL)
- Miscellaneous MilSup Issues	Mike StJohns (DDN PMO)
- EGP Enhancements	Mike Petry (UMd)
- Management/Monitoring	Craig Partridge (BBN)
- Short-Term Routing	Charles Hedrick (Rutgers)
- Performance and Congestion Control	Bob Stine (MITRE)

This section reproduces the combined report from these working group meetings (previously distributed by electronic mail).

5.1 Name Domain Planning

Convened by Doug Kingston (BRL)
Reported by Doug Kingston (BRL) and Mary Stahl (NIC)

Participants:

Paul Mockapetris (ISI),
Mark Lottor (NIC),
Doug Kingston (BRL),
Louis Mamakos (UMD),
Steve Dyer,
Rob Austein (MIT),
Jake Feinler (NIC),
Mary Stahl (NIC)

1) The charter of this Working Group is to look into the problems and concerns of the military community about using the domain system, with the goal of producing a MILNET nameserver white paper.

There are two basic issues to be discussed. One, what changes need to be made in host software in order two work well in a nameserver based environment where information is not always available (nameserver timeouts). The second issue is integrity in the nameserver data. This is of importance to all of us, and we need to determine what rules need to be followed to prevent spoofing or nameserver pollution.

A white paper would serve as a transition strawman plan for MILNET hosts by presenting recommendations to the appropriate organizations (eg, OSD, PSSG, DCA) for approval/response.

2) Coordinated with this in the short-term, work is in progress on 3 proposed RFCs:

o Domain Admin. Guide (ie, how to set up master file) - Mary Stahl and Mark Lottor (NIC)
o System Admin. Guide (ie, how to sign up) - Mary Stahl and Mark Lottor
o Root document to tie above together, point to sample implementations and possibly give dates - Walt Lazear (MITRE)

and updates to the Domain RFC's with current and planned changes.

The three new documents will be coordinated with DDN and MITRE as part of a Milnet Domain Transition Plan. The RFC authors may meet in late May to discuss progress. When complete, these proposed RFCs will be put online in SRI-NIC:<IETF> to get feedback from IETF members.

3) Name string discussion: any string should be allowed in a name, but we should probably warn domain administrators that names used for receipt of mail may require all names to begin with an alpha character. This has been discussed on the bind/namedroppers mailing lists as well. There is agreement in our group that there is no reason for the nameserver specification to preclude using numbers or any other ascii character in a domain name (except perhaps for NULL(0), for a string terminator) since the nameserver was designed to be a general facility. As for bind, anything is useable except for NULL. The key to this issue is that the domain/nameserver specification is only one of many that govern domain/hostnames. The assumption we make is that your choice of domain name for a given entity is governed by the intersection of the limitations imposed by the RFCs you may operate under. This probably allows things such as hostname vax.3com.com but precludes hostname 3com.com. If they are willing to abide by this, fine. Practicality may dictate otherwise.

4) Other miscellaneous changes to domain server software, such as negative caching was discussed. Discussion will continue online.

5) Motivated by a paper by Louis Mamokos, a strawman proposal was developed for a "Responsible Person" record in name servers. We should have something firm by the next meeting for others to comment on.

6) We need to consider adding another root server on ARPANET on East coast. If we do so, it might make sense to do so at BBN where it could be dual homed near ARPANET and MILNET. Suggestions are welcome. The need is not critical if the network stays healthy.

7) There needs to be an education process for DCA. Folks like the Arpanet and Milnet Managers might profit attending some of the Domain meetings and participating in the related discussions. Alternatively, tutorial sessions could be conducted at the PMO.

5.2 Miscellaneous MilSup Issues

Convened and Reported by Mike StJohns (DDN PMO)

A working group met to consider the ideas documented in a draft RFC which considered several ideas to augment the functionality of the Internet. The draft for this document is in the IETF archive on SRI-NIC.

The first part of the RFC deals with augmenting RFC822 (mail formats) to handle precedence and security within the mail system. General consensus was that the precedence stuff was useful, but needs more work and that the security marking stuff was useful only in a military environment.

Second part of the RFC deals with assigning default types of service to specific protocols. This part will be broken out as a separate RFC and expanded to include more protocols. There was some disagreement about what "reliability" meant in the TOS context and I'll be making some changes in the text to reflect the comments.

The third part of the RFC deals with several TCP and IP Options. The general consensus was "Why bother?". I'm going to issue these for comment anyway as their own RFC and see what flak I get.

5.3 EGP Enhancements

Convened and Reported by Mike Petry (UMd)

Participants:

Scott Brim, Cornell Univ
Marianne Gardner, BBNCC
Mike Karels, UCB
Tracy Mallory, BBNCC
John Moy, Proteon
Mike Petry, Univ of MD
Jose M Rodriguez, Unisys
Mike St. Johns, DCA

The charter of this Working Group is to review and make necessary changes to a draft EGP2 RFC document produced by Mike StJohns and Jose Rodriguez. The text for this draft RFC is in the IETF archive on SRI-NIC. Most of the changes involved clarifications and suggested implementation algorithms prompted by an extensive set of comments by Marianne Gardner. These included:

- Transit Autonomous System must be defined. This is needed in order to require that all transit Autonomous Systems must implement EGP (RFC904) and EGP2.

- A required change must be made to RFC904. The change is to return an error on a EGP version mismatch. The new code, 6 shall be defined as EGP version level mismatch. This is required for future levels of EGP to perform version negotiation. This change should be done now.

- At neighbor aquisition, the starting sequence number will be sent back in the response. This avoids treating the zero sequence number as a special case.

- More detail on the concept of the sequenced routing database must be added. Clarification is needed on how to age the data and estimates for TTL values and how TTL should propogate to other neighbors.

- Unsolicited requests will only be generated for the following reasons; gateway up/down, gateway change. A change of metric is NOT grounds for an unsolicited update.

Issues of controversy that were unresolved are:

- How/When to generate unsolicited updates?
- Does HELLO still serve a useful purpose?
- Are the draft metrics complete/useful?

5.4 Management/Monitoring

Convened and Reported by Craig Partridge (BBN)

The charter of the Management/Monitoring Working Group is to develop a framework and protocols for managing and monitoring Internet components. The initial focus is the current draft documents for the High-Level Entity Monitoring System (HEMS) by Craig Partridge (BBN) and Glenn Trewitt (Stanford). These documents are based on technical discussions in both the IETF and NSF groups prior to the formal establishment of IETF working groups. During the April 24, 1987 meeting, coordination with the new Network Management effort being set up by Lee LeBarre (Mitre) was also discussed.

Regarding HEMS, the group discussed transport protocol issues. A list of issues and possible solutions was developed and was sent to the GWMON mailing-list shortly afterwards for comment. Several people expressed interest in using HEMS to monitor LAN bridges, which may be possible (HEMS only requires a reliable link between an application and an entity being monitored). Marty Schoffstall expressed concern about whether HEMS was on a timetable consistent with NSFNET's immediate needs.

Lee LeBarre talked a little bit about his view about approaches to network management, and there was a short discussion about how Lee's group might interact with the existing IETF working group. It was observed that Lee was talking about a working schedule requiring meetings every few weeks, which was more often than most Internet researchers are willing or able to meet on a regular basis. It was urged that Lee's group arrange quarterly meetings in conjunction with the IETF meetings.

5.5 Short-Term Routing

Convened and Reported by Charles Hedrick (Rutgers)

Participants:

Bob Braden, USC-ISI
Hans-Werner Braun, Univ of Mich
Mark Fedor, Cornell Univ
Jose Garcia-Luna, SRI
Gene Hastings, PSC
Sergio Heker, JVNC
Charles Hedrick, Rutgers Univ
David Kaufman, Proteon
Paul Love, San Diego Superc.
Stan Mantiply, Ungermann-Bass
Don Morris, NCAR
Jeffrey Schiller, MIT
Zaw-Sing Su, SRI
Lixia Zhang, MIT-LCS

The charter of this group is to address short-term routing issues, particularly problems that have shown up on the NSFnet backbone and the regionals, but not restricted to these. Note the term short-term. There will be a separate Working Group charged to deal with routing technology. This Working Group presumes that the routing technology Working Group will do its job, and routing technology will be developed that can deal with the full complexity of the Internet. However it also presumes that any major change in routing technology will take at least a year to implement. This Working Group was charged with looking into how we survive that year. The intent is that any suggestions this group makes should be implementable almost immediately, probably within weeks for gated, and within months in commercial implementations.

The existing routing structure involves several national backbones (Arpanet, Milnet, and the NSFnet backbone), regionals (e.g. JvNC and NYsernet), and campus or other institutional networks. Currently, the backbones use their own private routing technology (Arpanet and Milnet) or Hello (NSFnet). Regionals and campuses seem to be using mostly RIP, though there are other strategies as well. The interfaces between Arpanet/Milnet and other networks use EGP. The interfaces between Hello-speaking backbones and other networks seems to use Mark Fedor's gated program. Gated translates metrics, at least between Hello and RIP. Thus the 16-hop maximum in RIP applies to the entire set of connected gateways, not just the individual regional or campus network.

Here are the major problems presented by this structure:

- the RIP maximum of 16 is being exceeded. Networks are inaccessible because of this.

- there seem to be large unexplained changes in metrics of some networks. Metric changes are happening more often than one would expect.

- a single user making a mistake can cause bad routing information to propagate through substantial portions of the network. This is not confined to NSFnet and the regionals. Rutgers was recently unable to reach Milnet because one Milnet host started sending inappropriate RIP packets to the Rutgers Arpanet gateway.

Several different approaches were discussed in the meeting:

- increased compartmentalization of routing. The word "firebreak" was used a lot. This could take the form of an arms-length protocol such as EGP at boundaries between local networks and regionals or regionals and backbones. Administrative controls on which routing information can pass a boundary are also possible.

- information hiding. Another word that was used a lot was "autonomous system". Many speakers made a convincing case that the details of campus routing should not be visible outside the campus, and that metric information for distant networks might not be needed. Various ideas were tossed around, but the only concrete proposals for how to implement this involved playing games with metrics at AS boundaries. (See detailed proposals below.)

- metric changes in RIP. There was strong support for relaxing the upper bound of 16 in RIP. There was some support for changing the RIP metric in other ways.

- algorithmic improvements. Although there was little discussion about this, there seemed to be general agreement that all implementations of RIP (and Hello?) should agree on such features as split horizon and hold-downs, and should use the same constants for timeouts and hold-downs.

The use of EGP was discussed several times during the meeting. EGP is intended to provide isolation among autonomous systems. Thus it seemed reasonable to think of using EGP to provide the necessary isolation, and to avoid having single RIP systems large enough that they exceed the 16-hop maximum. The problem with this is that EGP is really just a communications protocol. If the Internet is broken up into pieces that communicate via EGP, routing technology will be needed to route among those pieces, and to determine the routes and metrics to be communicated to the pieces via EGP. EGP alone will not provide that technology. Creating this technology appears to be beyond the scale of change that can be considered by this group. Several references were made to the Autonomous Confederations RFC. The approach taken by that RFC is quite compatible with what was being discussed in the meeting. Unfortunately, the RFC describes only the format for communicating routes, not an actual routing technology. So again, it does not specify a solution usable in this timeframe.

The only real product of the meeting was a set of changes to be made to RIP and gated. There was considerable question expressed at the meeting as to how useful such changes would be. Gateway vendors may find the overhead of making any change at all large enough that it is nearly as hard to do these short-term fixes as to implement major new technology. If that is true, then short-term fixes may not be very attractive. Many of these changes can be introduced without causing any incompatibility with existing implementations of RIP. The most controversial suggestion is the one for changing the metric. Anyway, here are the suggestions made at the meeting:

Suggestions that would not introduce incompatibilities into RIP:

- administrative tools should be added to allow metrics to be hacked when they cross AS boundaries. E.g. a regional might want to treat all routes obtained from the NSFnet backbone as metric 5, so that by the time they propagate to the far end of the regional, they do not exceed the limit of 16. Ideally it should be possible to exert controls on metrics both coming into and going out of an AS, and it should be possible to control the metrics of individual routes. That is, it should be possible for MIT to say that the AI Lab is to be advertised to NSFnet with a metric of 1, Proteon with a metric of 3, and no other networks are to be advertised. This is not quite a fixed route. Should AI or Proteon be inaccessible, it would not be advertised. However if it is accessible, the metric would be replaced with 1 or 3 respectively.

- timeouts, holddowns, and other algorithms should be specified, and constants agreed upon on a network-wide basis. No details were given. I presume that the draft RIP RFC that will be circulated shortly is what is intended here.

- administrative controls such as those currently implemented in gated should be considered. Gated allows the administrator to specify per interface and per protocol (i.e. RIP, EGP, or Hello) a list of networks for which routing information will be either accepted or excluded. If a list of networks to be accepted is specified, then information on all other networks will be ignored. Gated also allows the administrator to specify a list of acceptable peers. Routing information from other gateways will be ignored.

Changes to the RIP metric. The following changes would effectively create a new routing protocol, since it would be dangerous to allow new implementations to talk to old ones. This means that implementations that follow these suggestions should use a different UDP port from the old RIP.

- infinity should be increased from 16 to something that allows the entire 32-bit field to be used. It is suggested that a metric be used that is compatible with Hello's, so that conversion into and out of the NSFnet backbone is not needed. This means that the metric should nominally represent milliseconds of delay. Probably we should follow gated's lead in suggesting a default increment of 100 for each link. As long as link costs are set by the system administrator, the semantics of the metric are really not determined by the protocol. That is, the only real change here is removing the maximum of 16. Whether the metric is changed to present a delay is strictly up to the network administrators.

- it should be possible to set the cost to be used for a given link. The intent is to allow a metric that distinguishes between faster and slower links. Note that we are suggesting only static settings. We do not believe RIP in its current form is capable of supporting real-time delay measurements, etc.

- the algorithms used should otherwise be the same as normal RIP. We assume that any implementation that follows these guidelines will be prepared to accept both the old and new RIP, on different ports. Common code should be used. The only difference would be that the value of infinity would be different for conversations on the two ports. Gateways at the boundary of an AS may find themselves speaking old RIP in one direction and new RIP in the other direction. They will need to use the ability to set metrics, described in the previous section. (One might also allow the Sstem administrator to define a conversion factor to be applied to metrics going between old RIP and new RIP. The recommended conversion factor is 100.)

Gated should implement the suggestions in the first group as soon as possible, and vendors should do so as soon as practical. In particular, the ability to turn all metrics into a specified constant may be needed to allow existing routing structure to survive the addition of Suranet, because the diameter of Suranet is expected to be large.

Extended followup discussion has ensued online and a proposed RFC on RIP has been distributed.

5.6 Performance and Congestion Control

Convened and Reported by Bob Stine (MITRE)

Participants:

Bosack, Len, Cisco Systems
Callon, Ross, BBN
Chiu, Da-Ming, DEC
Coltun, Rob, MITRE
Gross, Phill, MITRE
Jacobson, Van, LBL
Jain, Raj, DEC
Ramakrishnan, K., DEC
Stine, Robert, MITRE

The major goal of the Working Group is to produce a white paper on congestion control techniques, which will

- Enumerate and evaluate various congestion control schemes that have been proposed for the Internet, and

- Suggest techniques that could be employed by hosts and gateways to control or avoid congestion.

The major topics discussed during the meeting were:

1. DEC's congestion control scheme.

2. The employment of Van Jacobson's TCP "slow start" algorithm.

3. The use of Box-Jenkins time series analysis for improved RTT estimation.

4. Modifications to gateway software for reducing congestion.

5. Several longer-range or more esoteric techniques for controlling or avoiding congestion.

This summary presents the discussions on the above issues topically, rather than in the strict chronological order in which they occurred.

DEC's Congestion Avoidance Procedure

The DEC scheme, developed by Raj Jain, is characterized as a congestion avoidance scheme. The features of this scheme were presented in detail to a plenary session of the IETF and the ISO X3S3.3 group earlier that day. The technique is intended for use in communications systems which have datagram service at layer 3, and which use windowing for flow control at layer 4. It employs a feed-forward mechanism: if a packet traverses a highly loaded Intermediate System (IS), then a bit is set in packet. If the End System (ES) receives a high enough rate of incoming packets with the congestion bit set, then the receiving window is narrowed.

The DEC scheme has several appealing characteristics. Among these are that it requires very low overhead, since the window size info must be transmitted, whether or not congestion avoidance is implemented. The scheme was also reported to have performed well in simulations.

During the working group meeting, there was some discussion on whether and how an IS might feed back congestion information, so that information could get to the message transmitter sooner. There are, however, problems with having an IS piggy-back a congestion control message to send to a message source. First, an IS does not usually examine the source address of a packet. Hence, there would be a significant amount of additional processing for each packet that traversed the IS. Furthermore, packet flows between sources and destinations may be asymmetric, leading to fewer opportunities for the IS to inform the message source of traffic conditions. Finally, the return path to a message source may not be the same as its outbound path. Hence, the congestion level reported on a return path might not accurately reflect the congestion level of an outbound path. In light of these points, the consensus of the group seemed to be that employing a feed-back scheme would not be profitable.

Several observations were made concerning the fact that the DEC congestion avoidance scheme only uses one bit per packet. It was noted by Len Bosack that one bit is sufficient for indicating congestion, in particular since hosts actually process a series of bits. The use of more bits, so that a level of congestion could be indicated, would merely increase the rate at which the system could adapt to changes in traffic rates. It would not, however, alter the overall characteristics of the control system behavior. A member of the working group suggested that it would probably be better for a congestion control scheme to sample more often, rather than at greater fidelity. In addition, it was suggested that it require less processing to obtain a 1-bit congestion level, rather than a rate. The point was made, however, that this would not necessarily be true if if the congestion control bit is set by computing a congestion level and then comparing it against a threshold.

Finally, it was noted that there are unused bits in the IP header, which could allow the DEC scheme congestion avoidance scheme to be retrofitted into the existing DoD protocol suite.

Slow Start in TCP

During the Working Group meeting, Van Jacobson's slow start algorithm for TCP was also discussed. This idea has been presented on the Internet Engineering email list, as well as at the April 22 session of the IETF meeting. In a nutshell, the slow start algorithm has TCP open its transmission windows gradually, only as acks are received. In the event of retransmission, the windows are narrowed.

There was not much discussion on the slow start algorithm; the consensus seemed to be that it is a necessary bug fix. The use of slow start could perhaps complement a DEC-like congestion avoidance, though there would probably be some undesired control interaction. Nevertheless, in order to eliminate the worst effects of blasting gateways and spurious retransmissions, it seems that TCP ought to use a slow start algorithm, regardless of the other congestion control or avoidance schemes employed.

Box-Jenkins techniques for improved RTT

During the Working Group session, another TCP enhancement that Van Jacobson suggests was discussed: the employment of AR (Auto-Regressive) or ARMA (Auto-Regressive, Moving Average) models to predict RTT (round trip time; the time between a segment's transmission and the receipt of its acknowledgement). The algorithm which the TCP spec suggests for use in estimating RTT is exponential smoothing, which computes a mean. The problems with this technique for estimation are:

1. It assumes that successive observations are independent;

2. It assumes that there is no trend, and it reacts slowly to change;

3. The mean does not characterize the dynamics of the system.

Van Jacobson reported that he had obtained very large improvements in estimation with the simple AR 1 model, and even better performance with the ARMA 3 and ARMA 4 models. The ARMA models were reported to work well if the net is congested. So, the times that they do not work well, the Internet is not congested, and it is a much less serious a matter that TCP is using an inaccurate RTT estimate.

In a related discussion, the question of how to measure RTT if there have been retransmissions was aired. Measuring from the first packet sent will tend to overestimate RTT, while measuring from the last packet sent will tend to be overly optimistic. It was briefly discussed whether or not second-order sequence numbers might be used to indicate the times a packet had been retransmitted. It was suggested that this is probably not necessary, since "... packet exchange gives the same information." The question of biased estimator not seen as important; it is much more critical to have the estimators track the observed values more closely.

Also in this discussion it was noted that in current TCP implementations, if the initial RTT estimate is too low, the system will never correct itself, since the RTT estimate is not adjusted during a retransmission.

Gateway tactics for congestion control

It was noted that given their narrower administrative span and fewer absolute numbers, quick fixes might be easier to implement in gateways than in the thousands of hosts. The major congestion control techniques proposed for gateways were designed to randomize the order packets it receives, so that the synchronization of datagram blasts which Van Jacobson has reported could be avoided. Len Bosack reported that in experiments, random selection of packets for service has been seen to be the best for desynchronizing internet traffic, and that fair queuing works almost as well. (If prioritizing service is a requirement, then separate randomizing can occur at the differing priority levels.) The interaction of fair queuing with ARPANET connection setup characteristics (e.g., if too much time elapses between transmission to a destination PSN, then a connection block must be reallocated) was not seen as a problem.

In the discussion on gateway queuing discipline, a Working Group member observed that once a gateway using random selection "queues" a large number of packets, it becomes something close to a stack. If the randomizing is accomplished by inserting an arriving packet in a random location in the queue, then for those packets placed near the tail of the queue, it is highly likely that new packets will be inserted before them. To insure progress through the queue, it was suggested that there be a short section at the front of the queue into which insertions could not be made.

Another discussion topic was that gateways could also reduce congestion if they employed better strategies in selecting which packets to drop. In particular, they should drop the packets they have held the longest, rather than those that have most recently arrived. In addition, Van Jacobson suggested that by attempting to drop at most one packet per connection could reduce the load offered to a gateway several fold, while dropping more than one packet from a single connection increases traffic.

Longer-range congestion control techniques

As at his IETF briefing, Van Jacobson noted that the RTT varies in a saw-tooth like cycle: there tends to be a large step increase in RTT, followed by a gradual decay. Using ARMA techniques to estimate RTTs, hosts might also try to determine at what point they are in the RTT cycle, and schedule their transmissions accordingly.

Another point raised was that hosts could obtain a wealth of traffic information from gateways: estimated subnet reliability, subnet MTU, available bandwidth, and other data. The point was made that it would be simpler to request such information instead of estimating it.

APPENDIX A

Presentation Slides

This section contains the slides for the following presentations made at the April 22-24, 1987 IETF meeting:

- Enhanced AHIP StJohns (DDN)
- BBN Report Hinden/Gardner (BBN)
- Congestion Control Simulation Stine (MITRE)
- Arpanet Performance Measurement Gross (MITRE)
- TCP Performance Enhancement Jacobson (LBL)
- Gateway Monitoring Partridge (BBN)
- Management Architecture LaBarre (MITRE)
- IETF status/overview Gross (MITRE)
- FCCSET report Gross (MITRE)
- ANSI routing architecture Tsuchiya (MITRE)
- NSF gateway requirements Braden (ISI)
- Routing Directions at SRI Su and Garcia (SRI)
- NBS Routing Proposal K. Mills (NBS)
- Burroughs Integrated Adaptive Routing Piscitello (Unisys)
- DECnet Phase V Routing Oran (DEC)
- SPF Routing in the Butterfly Gateways Mallory (BBN)
- Congestion Avoidance Jain, et. al. (DEC)

Enhanced AHIP StJohns (DDN)

Arpanet / Milnet Interface Changes:

- Needed Due to size of MILNET
- ...Due to additional Functionality

Size

By Late 1988 at current installation rates, the MILNET will consist of more than 256 packet switches.

Current interface protocols (X.25 + AHIP ... 1822) have a limit of 256 packet switches in their addressing schemes.

AHIP Changes

1) Expand field to permit 1024 packet switches

2) Merge functionality of 1822L into the standard interface

3) Provide as an option

   a) Subnet congestion feedback

   b) logical addressing name server

   c) precedence

   d) Type-of-service

4) Provide version #s to distinguish between old/new.

Actual Changes (min)

AHIP-E:

Copy 3 IP bytes to AHIP header (Instead of zeroing center)

X.25: modify IP mapping

Grab 2 extra bits to calculate HAH

PS:<ietf> Ahip-e.txt

on the NIC

BBN Report · Hinden/Gardner (BBN)

ARPANET APRIL 1987

JANUARY CRISIS PASSED

WHAT WE DID

- o PSN ROUTING PARAMETERS ADJUSTED

  stabilize cross-country routes

- o LSI/11 PARAMETERS ADJUSTED

  requested ping rate 1 min
  requested poll rate 3 min
  delay added to nbr's poll rate

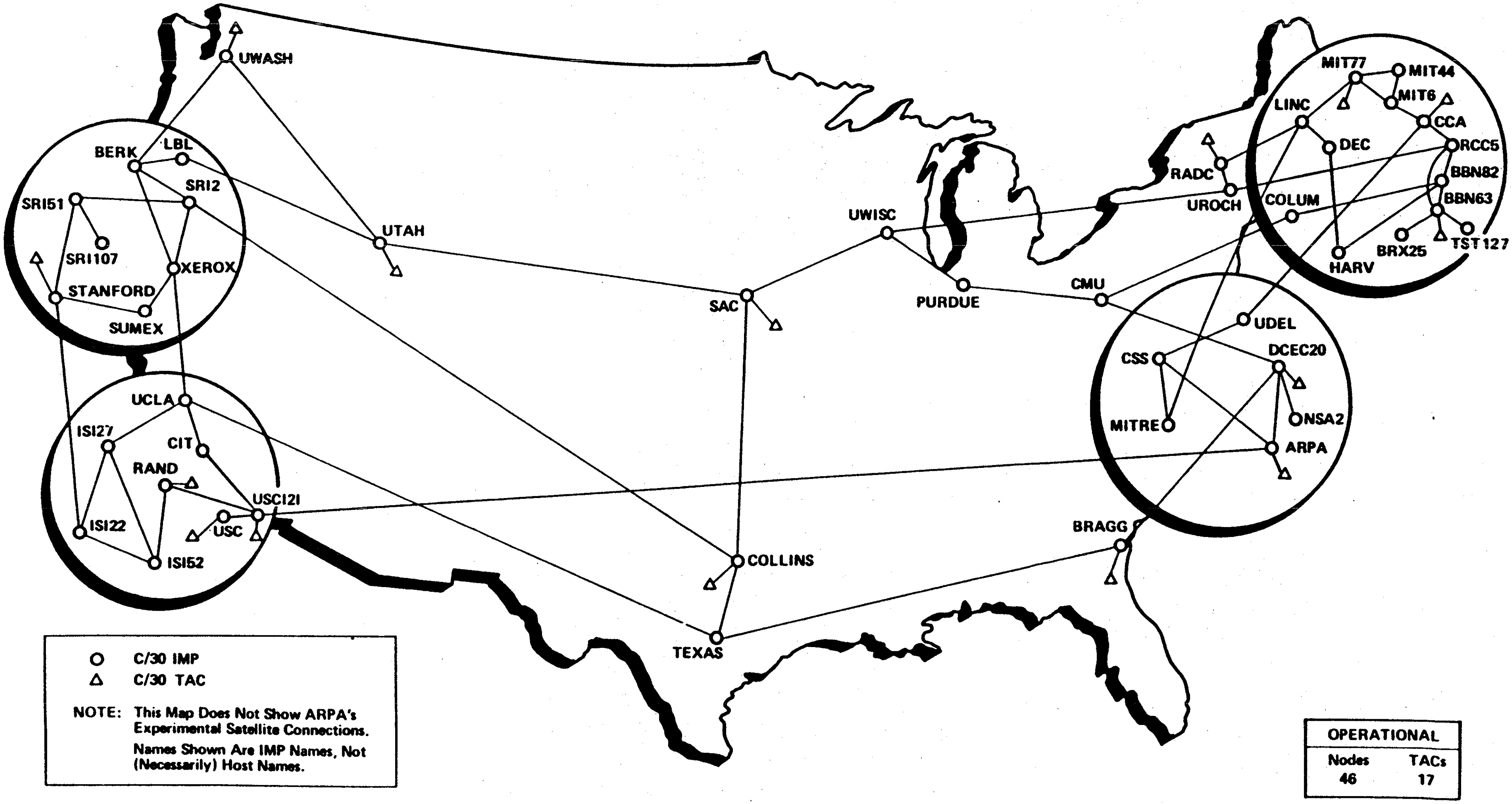
- o C300 UPGRADES COMPLETE

  UCLA-1
  INOC-5
  ISI-27
  SRI-51
  MIT-77
  UWISC-94

NODE UTILIZATIONS

PSN	Measured Peak 7-minutes Jun86	 Nov86	 Feb87	 Apr87
-----	-----	-----	-----	-----
ISI27	73	58	61	22
WISC	81	60	105	37
RCC5	74	65	96	40
UCLA	81	51	56	25
SRI51	46	49	61	24
MIT77	57	77	90	30

ARPANET Geographic Map, 31 January 1987



○ C/30 IMP

△ C/30 TAC

NOTE: This Map Does Not Show ARPA's Experimental Satellite Connections.

Names Shown Are IMP Names, Not (Necessarily) Host Names.

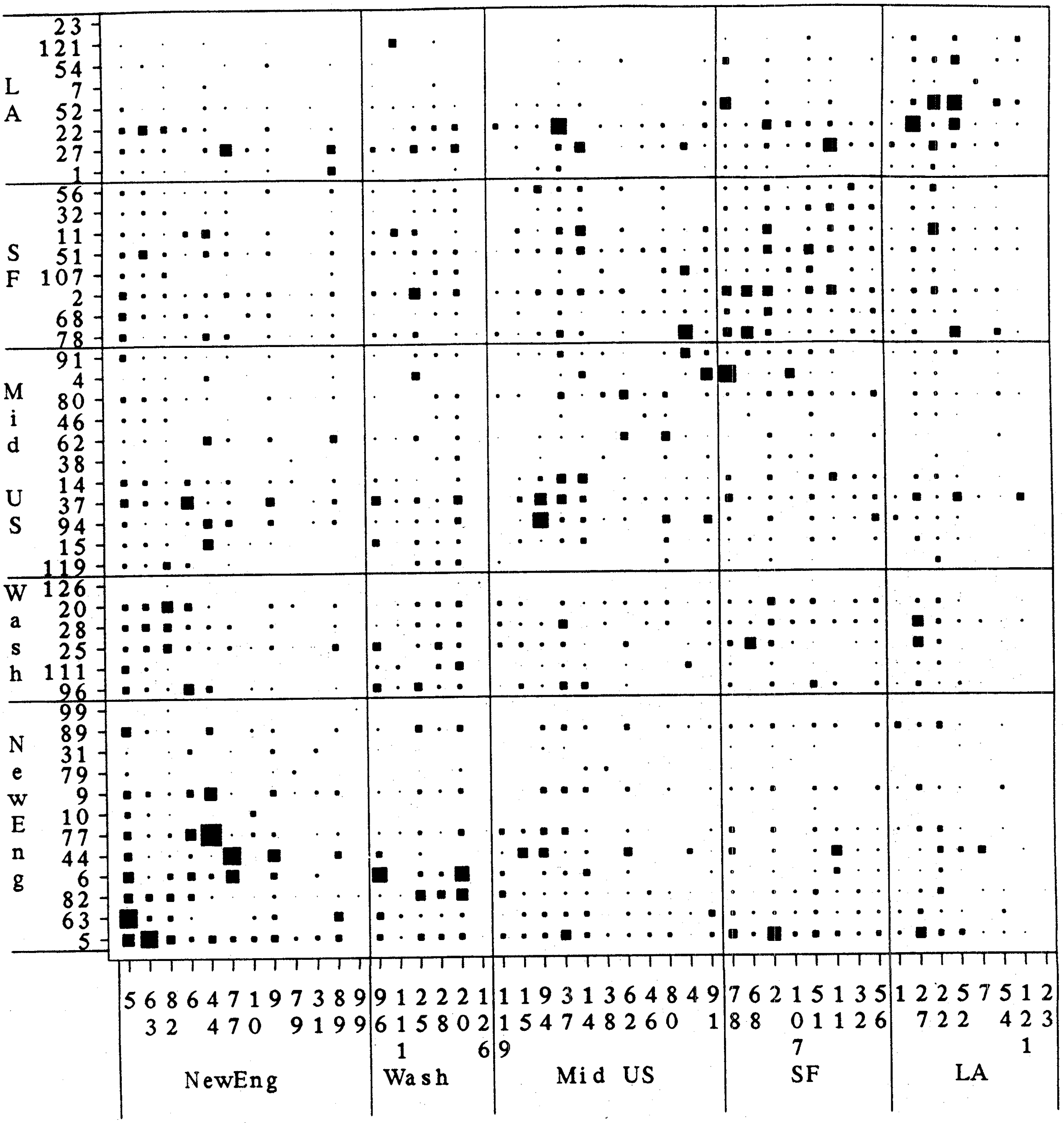
OPERATIONAL	
Nodes	TACs
46	17

TRUNKING REMAINS AS MAJOR PROBLEM

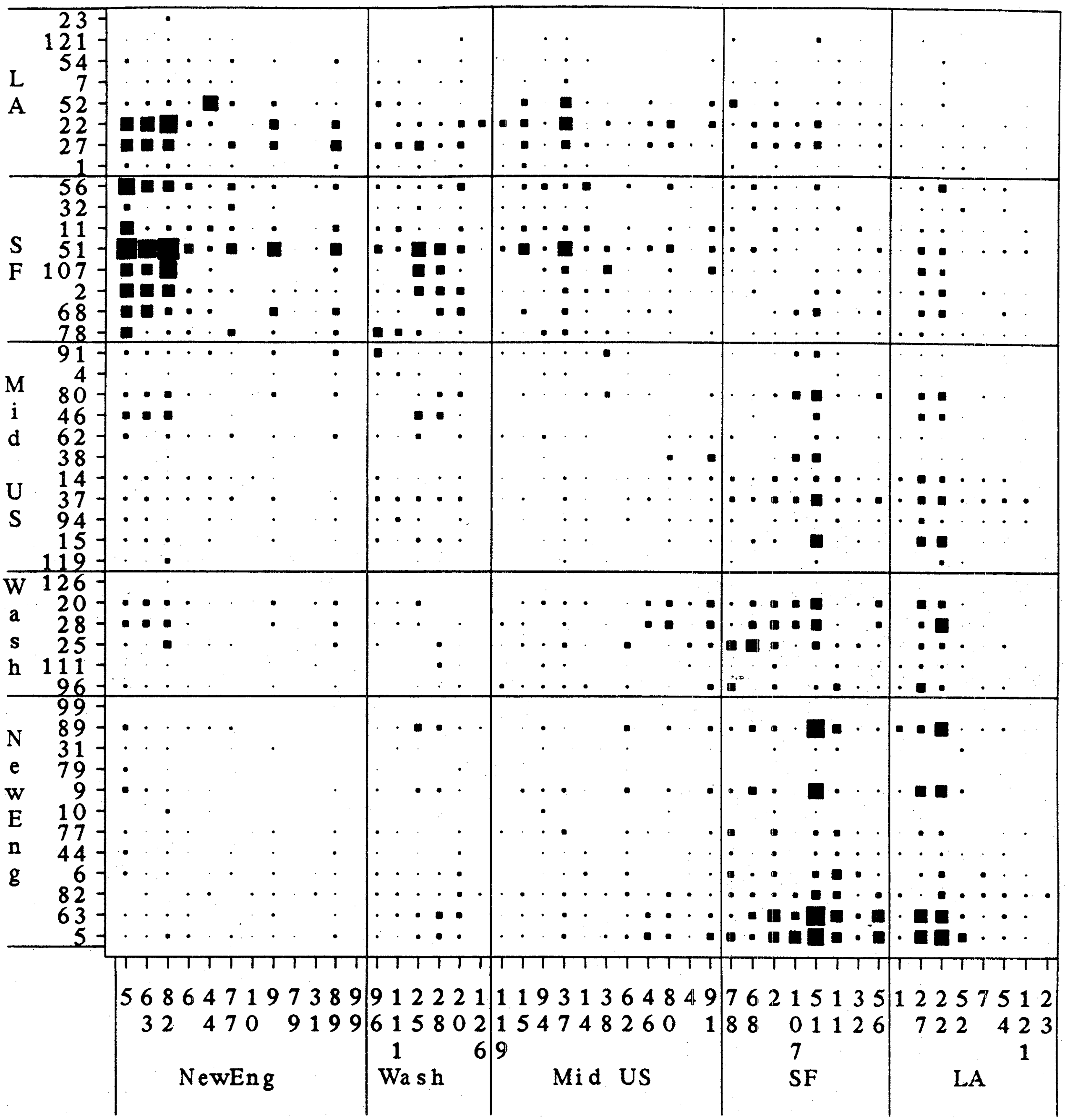
- VSAT LINE MIT44-SRI51 SOON
- CHARACTER OF TRAFFIC HAS CHANGED
  - TRAFFIC IS UP SLIGHTLY
  - MEAN PATH LENGTH HIGHER

```
                              TOTAL TRAFFIC
                       PACKETS PER WEEK (MILLIONS)
1986*
JAN02    87.6 **************************
JAN09   117.1 ***********************************
JAN16   122.8 *************************************
JAN23   121.5 ************************************
JAN30   132.9 ****************************************
FEB06   124.9 *************************************
FEB13   135.3 ****************************************
FEB20   145.1 *******************************************
FEB27   139.8 ******************************************
MAR06   148.9 ********************************************
MAR13   146.3 ********************************************
MAR20   158.2 ***********************************************
MAR27   111.1 *********************************
APR02   112.5 **********************************
APR10   143.8 *******************************************
APR17   148.4 ********************************************
APR24   139.2 ******************************************
MAY01   126.1 **************************************
MAY08   142.9 *******************************************
MAY15   142.8 *******************************************
MAY22   138.7 *****************************************
MAY23   128.6 **************************************
JUN05   142.9 *******************************************
JUN12   107.2 ********************************
JUN19   152.1 *********************************************
JUN26   193.0 **********************************************************
JUL03   128.9 ***************************************
JUL10   109.9 *********************************
JUL17   133.1 ****************************************
JUL24   134.6 ****************************************
JUL31   174.8 ****************************************************
AUG07   140.6 ******************************************
AUG14   140.2 ******************************************
AUG21   109.4 *********************************
AUG28   129.1 ***************************************
SEP04   113.9 **********************************
SEP11   123.8 *************************************
SEP18   100.1 ******************************
SEP25   147.5 ********************************************
SEP25   147.5 ********************************************
OCT02   159.7 ************************************************
OCT09   114.3 **********************************
OCT16   134.4 ****************************************
OCT23   137.8 *****************************************
OCT30   138.3 *****************************************
NOV06   148.8 ********************************************
NOV13   152.1 *********************************************
NOV20   156.0 ***********************************************
NOV27   154.7 **********************************************
DEC04   134.0 ****************************************
DEC11   149.7 *********************************************
DEC18   163.0 *************************************************
DEC24   112.6 **********************************
1987*
JAN01    91.9 ****************************
JAN08   126.6 **************************************
JAN15   150.3 *********************************************
JAN22   150.8 *********************************************
JAN29   151.7 *********************************************
FEB05   141.3 ******************************************
FEB12   167.6 **************************************************
FEB19   171.8 ***************************************************
FEB26   181.0 ******************************************************
MAR05   168.9 ***************************************************
MAR12   115.8 ***********************************
MAR19   166.2 **************************************************
MAR26   181.4 ******************************************************
APR02   178.9 ******************************************************
```

ARPANET Peak Hour Traffic Matrix, 2115-2215, 25 Feb 87



ARPANET 63-Traps for 13 Jan 87



Congestion Control Simulation

MITRE Corporation, W-31
Robert Stine

Briefing Overview: Congestion Control Simulation

- Description of Model
  - Underlying assumptions.
  - Approach to modeling:
    - Transport connections.
    - Windows.
    - Network access.
- Preliminary results
- Conclusions.

Gateway Congestion Control Experiments

- Part of MITRE's Internet Engineering Program
  - — Provide policy recommendations, coordinated approach to internet enhancements.
  - — Congestion control study funded by DARPA.
- Goals of congestion control study:
  - — Examine suspected causes of congestion on Internet.
  - — Evaluate approaches proposed for improving performance under high traffic conditions.
- Method: Discrete event simulation.

Gateway Congestion Control Simulation

- Previous generation:
  - Stub gateway topology.
  - Simple traffic model: segments generated by Poisson process.
- Current generation: Detailed model of Transport connections.
- Next generation: Internet topology.
  - Gateway-Gateway protocols may be modeled.
- Current status: beginning experiments with current simulation

Underlying Assumptions of Model

- Dominant congestion-inducing effects in stub topology:
  - — Dropped packets at gateway.
  - — Gateway queuing delays at slow interface.

- Negligible effects in stub topology:
  - — LAN contention.
  - — Traffic from long-haul net to LAN.
  - — Non-transport traffic.

- Model limitations
  - — Congestion in long-haul net not modeled.
  - — Traffic from long-haul net to LAN hosts not modeled.
  - — Non-transport traffic not modeled.
  - — Single gateway between LAN and long-haul net.

Model of Transport Connections

- Each host may have several individual Transport connections.
- Transport connection characteristics
  - — Type: bulk traffic or interactive.
  - — Volume of traffic.
  - — Ratio of long to short segments.
  - — Maximum window size.

Modeling stochastic traffic

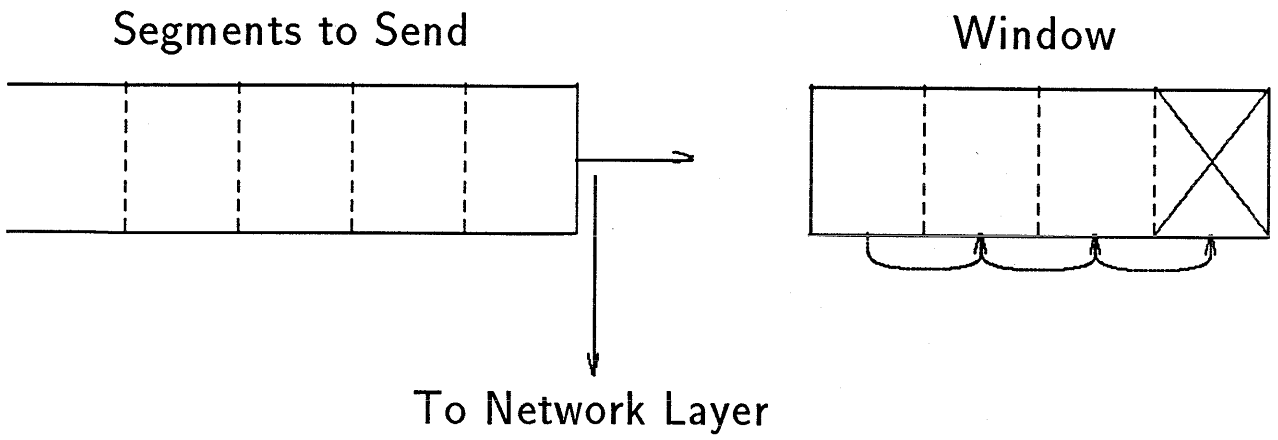
- Connection generation:
  - Poisson; λ may be different for each host.
- Deciding between bulk and interactives, & long and short packets:
  - Bernoulli trials.
- For interactive connections:
  - Segment generation: Poisson.
  - Length (in time) of connection: normal.
  - Average transmission rate: normal.

Modeling stochastic traffic (continued)

- Bulk connections
  - Number of segments: normal, rounded to nearest integer.
- Net-to-gateway round trip
  - "Truncated" normal - has lower bound.

Model of Transport Windows

- For each connection:
  - — Separate SRTT maintained.
  - — Windows divided into segment-sized slots.
  - — RTO computed for each slot.
  - — Window slots freed as segments at edge are acked.



Model of network access

- Simulation tick: time to send average packet across LAN.
- From hosts to gateway, aggregate total no more that 1 packet/tick.
- Hosts queue packets at net interface.
- Hosts polled, round-robin, for packets to send.
- Hosts may regulate interpacket transmission intervals
  - If unthrottled, may send up to 1 packet/tick.
  - If throttled, hosts space packet transmissions.

Simulation parameters

- Control options -
  GW queue division (e.g. Fair Queuing vs. monolithic queuing);
  Host retransmission backoff and RTO limits.
  Host response to SQ; SQ throttle and hold-down values.

- Traffic density -
  Rate of connection generation; Size of bulk transfers;
  Interactive transmission rate; Packet length mixes.

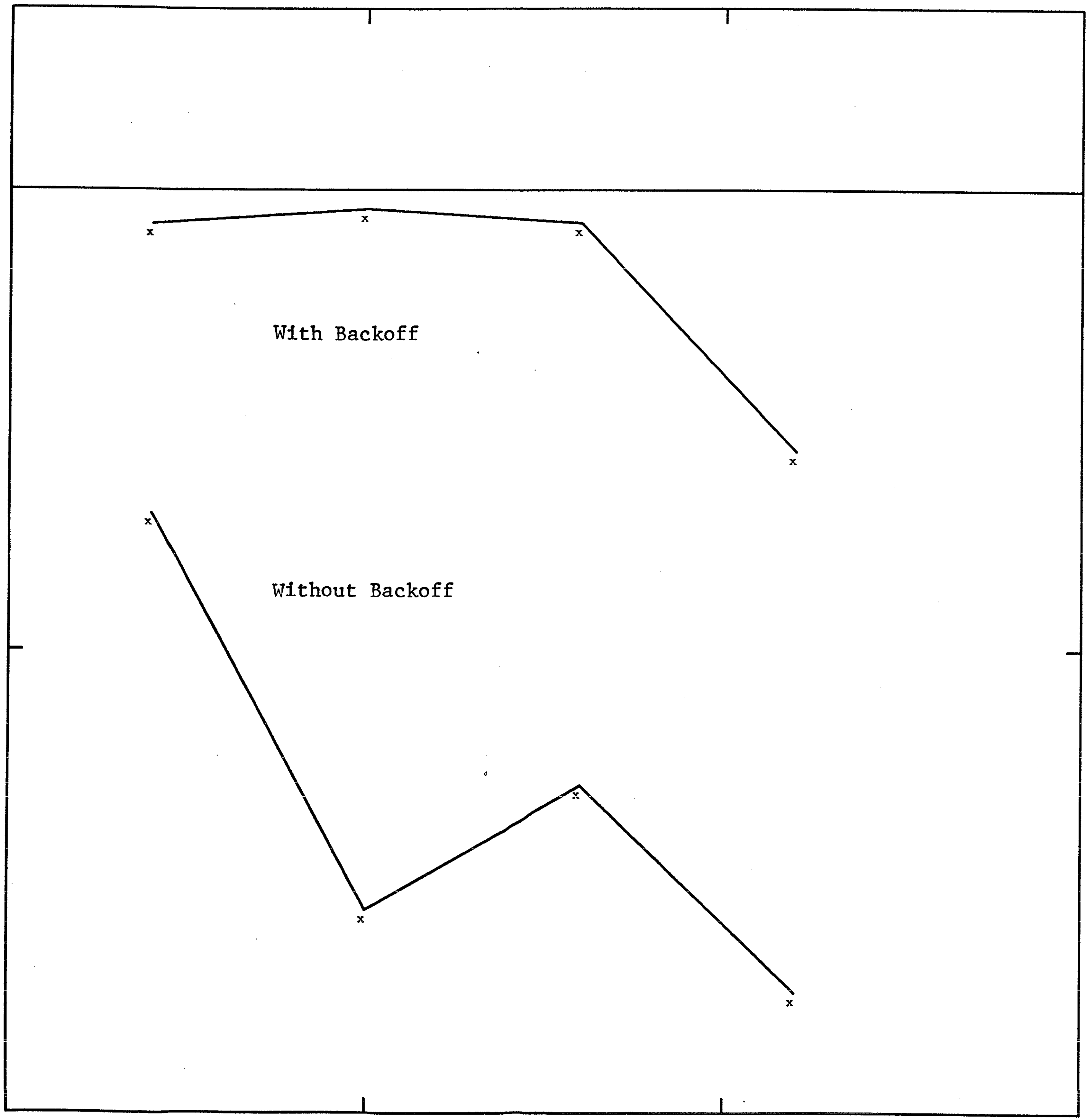
- Delays -
  LAN & net bandwiths; Average net delay, std dev;
  SQ "timeout" in GW; SQ delay to reach host;

- Other system characteristics: GW buffer pool; Packet length, etc.

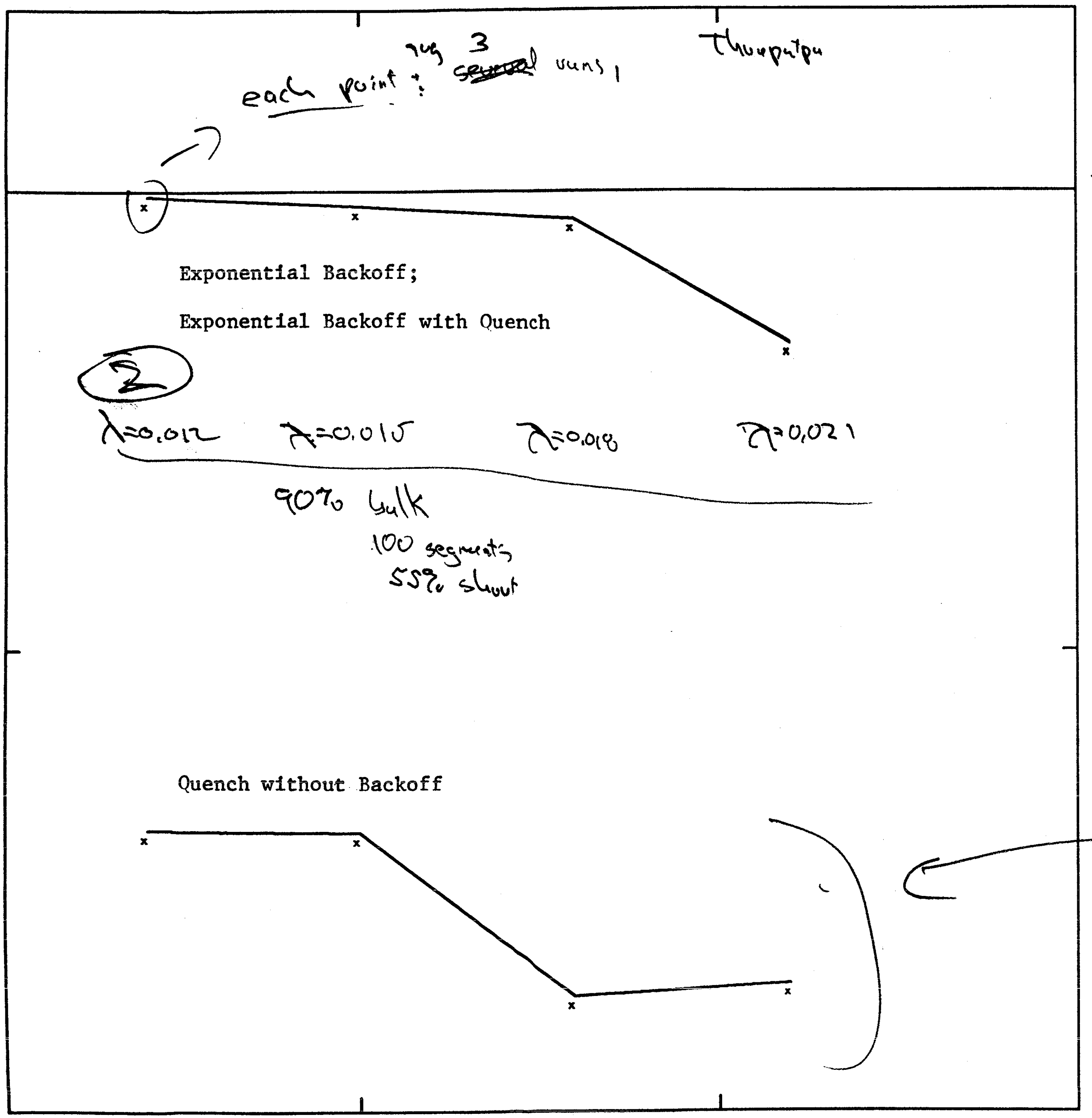
Experiments with the simulation

- Definitions:
  - End-to-End load: volume of segments handed to transport layer from above.
  - Thruput: Segments arriving at destination, after filtering for duplicates.
  - One measure of efficiency: ratio of thruput to end-to-end load.

- Sample Cases: use of source quench with and without retransmission backoffs.



Source Quench with Fair Queuing and Incremental Backoff



Source Quench with Monolithic Queuing and Exponential Backoff

1

1.0

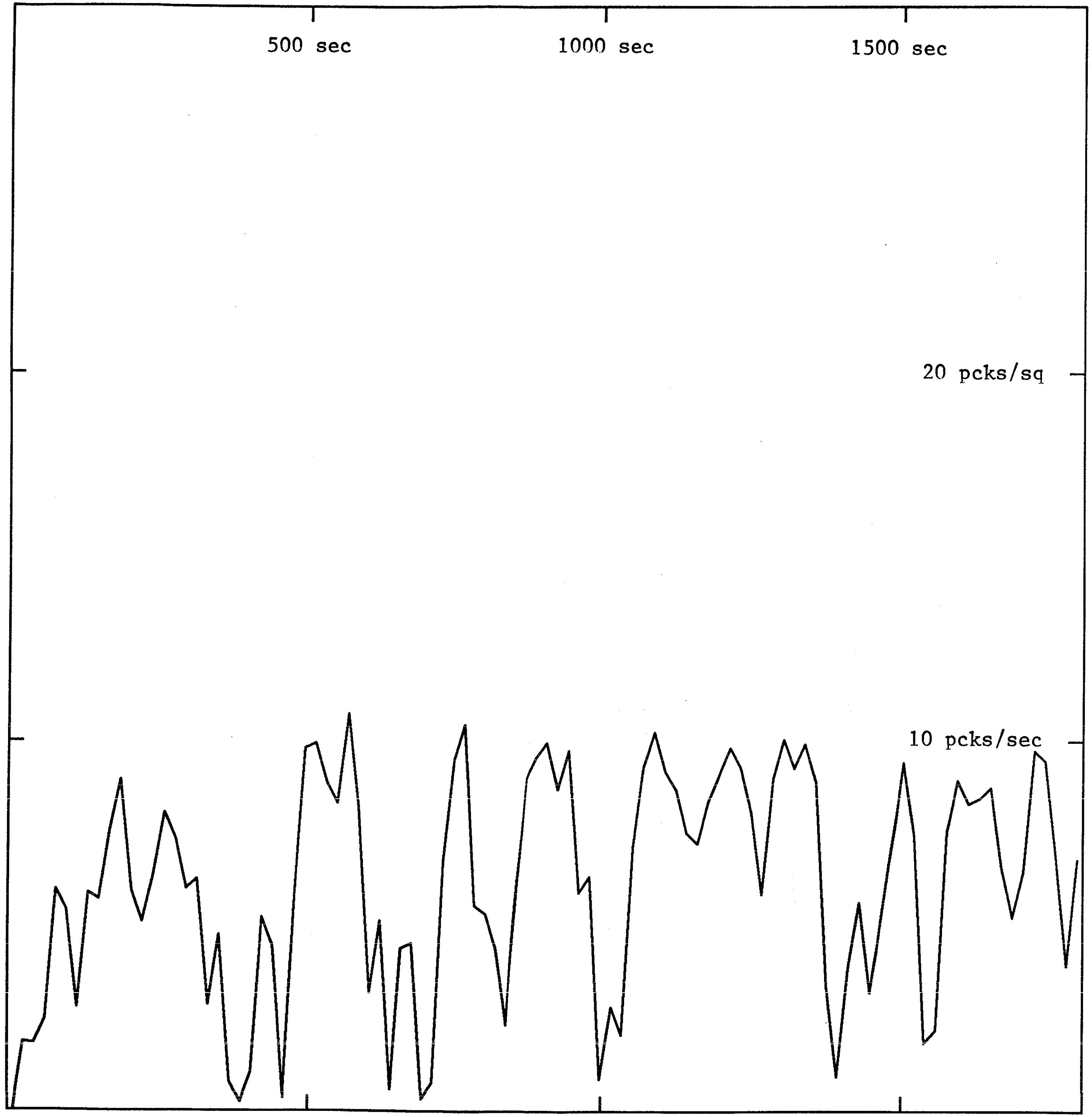
Runs: 30 simulated minutes
- 5 Mb lan
- 56 Kb net interface

5 hosts, identical chars.

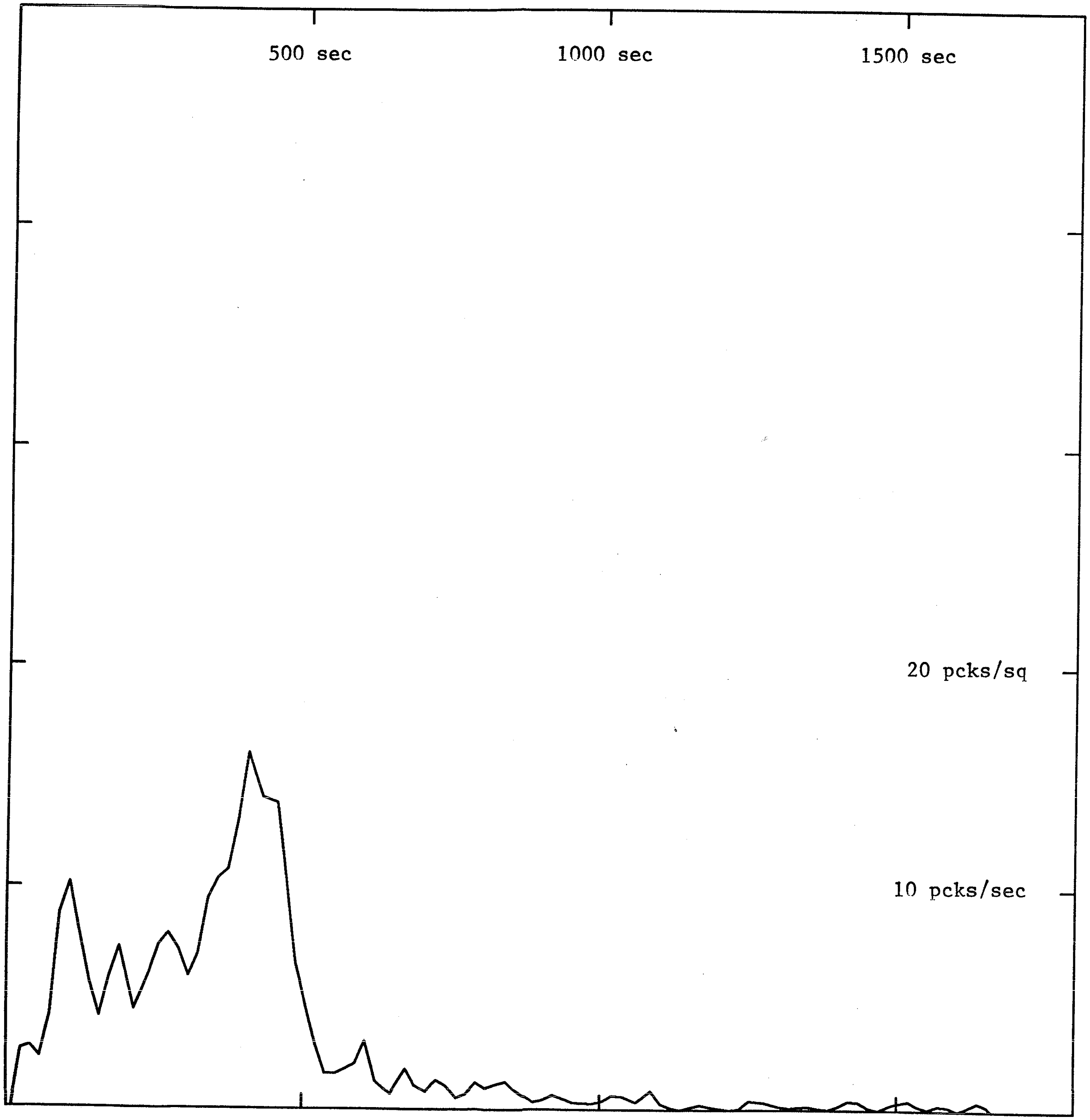
3

High variance

Single run.



FQ, SQ, INCR, l=0.015, default seed: thru_baud.pts



FQ, SQ, no backoff, def seed, l=0.015: thru_pcks.pts

Concluding Observations

- As configured, systems failed hard:
  - — For borderline systems, wide swings in performance. → Given same avg traffic, Stochastic variation can make/break performance.
  - — Phenomenon could be an artifact of the simulation.
- Preliminary results support views that:
  - — Study of retransmission algorithms may have high payoff.
  - — If SQ is implemented, it should be integrated with Transport layer.

Preliminary Data

- Distribution of RTT's

- Parameters vs. hops
    - Median
    - Variation
    - Skew

- Short term variation over single measurement
    - Median
    - single trace

→ Difference in Data presented Today

- Medians
- Distributions
- Include HDH, more hops

→ Still to do

- DCEC C30 local net
- Talk to BBN about how PSN's work

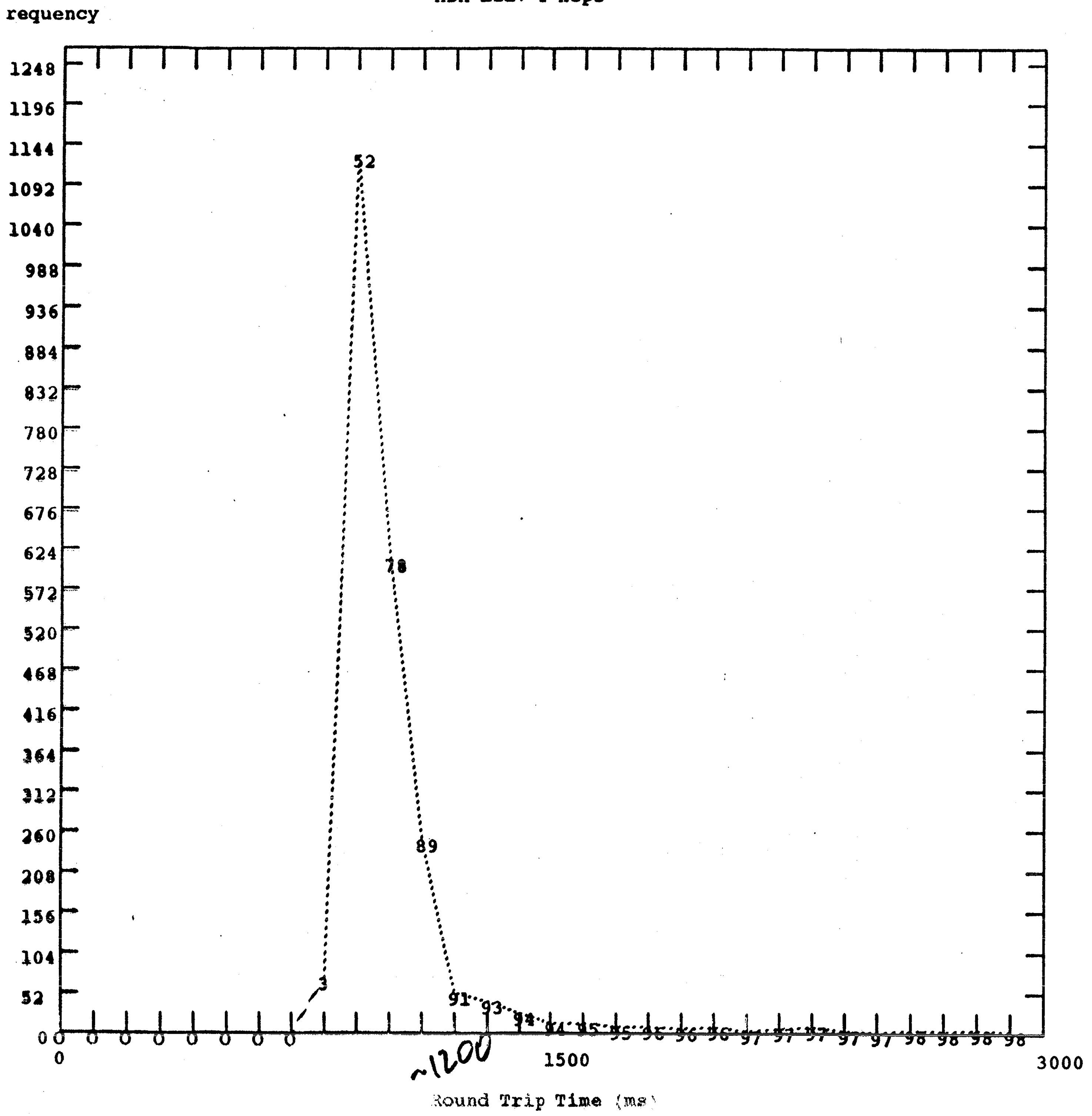
Arpanet Delay Measurement Progress

- ICMP Echo
- 3 interfaces (1822, X.25, HDH)
- PSN Distance (0-9 hops, 8 hosts in all)
- 6 cases
  - short/long Packets
  - High/Low Traffic Period
  - Short/long gap between packets
  - denoted as

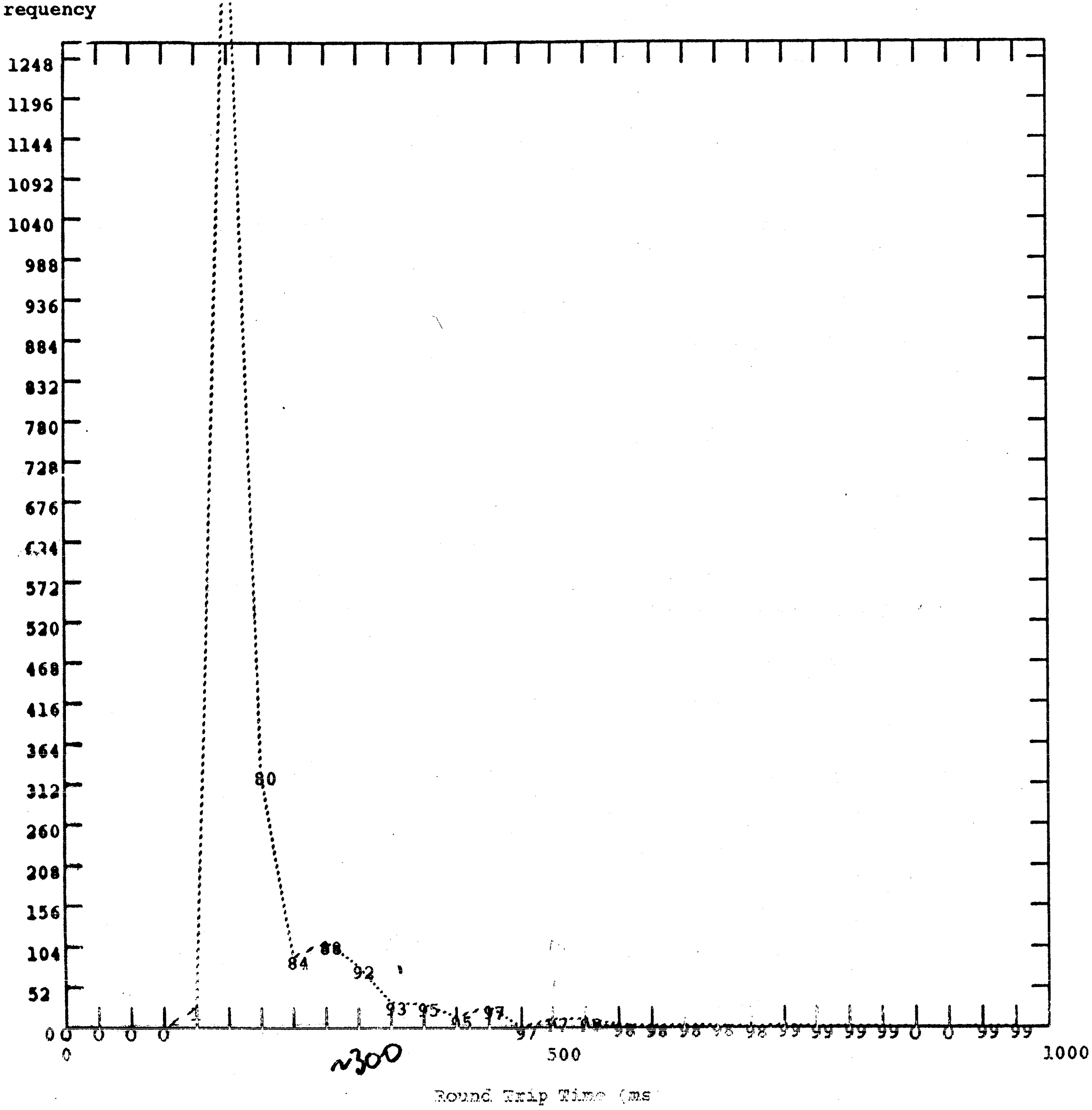
    S O L ← Low traffic

    short packet (S), short gap (O)
  - LOL + LOH don't work

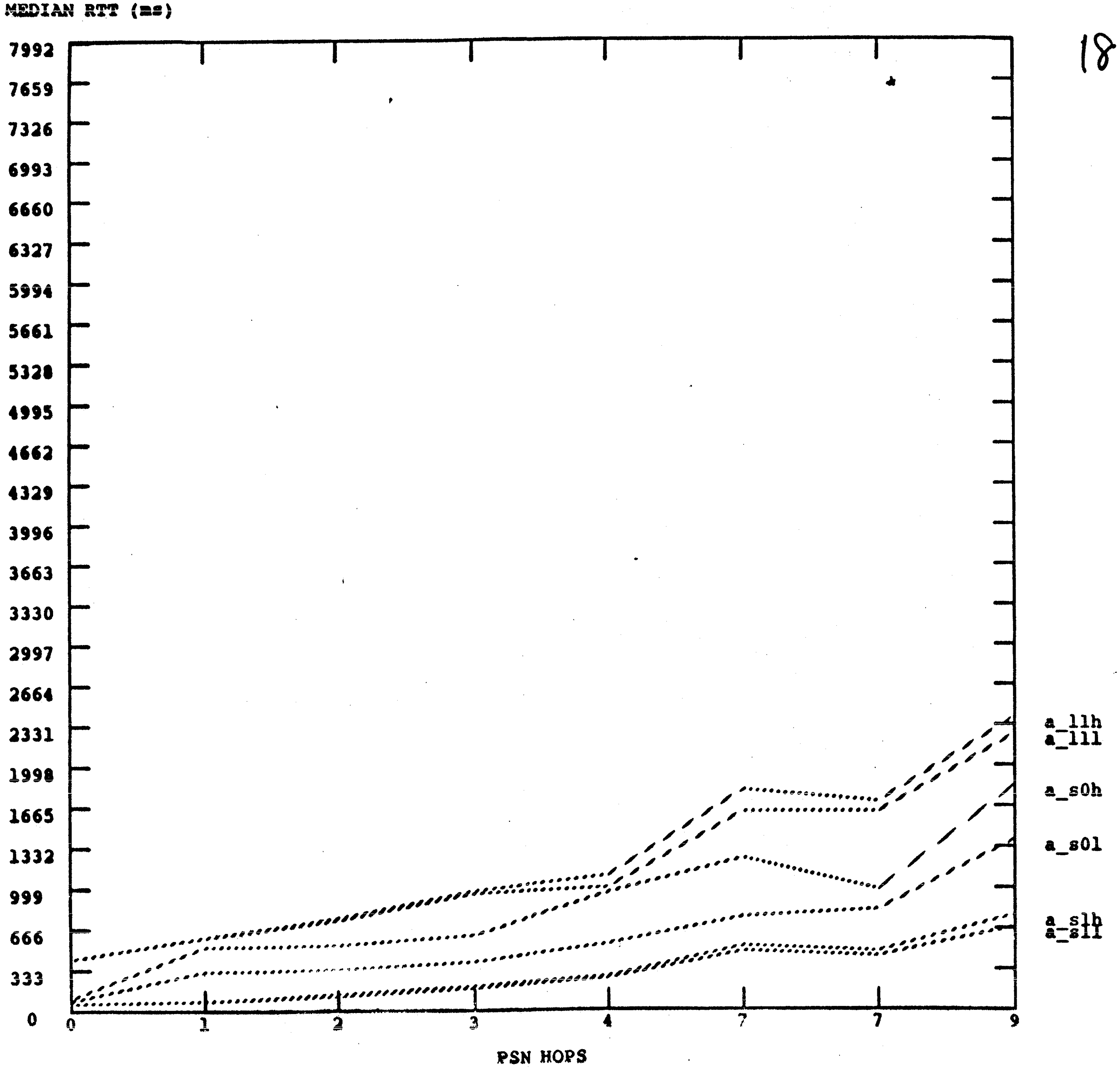
HDH L1L: 2 hops



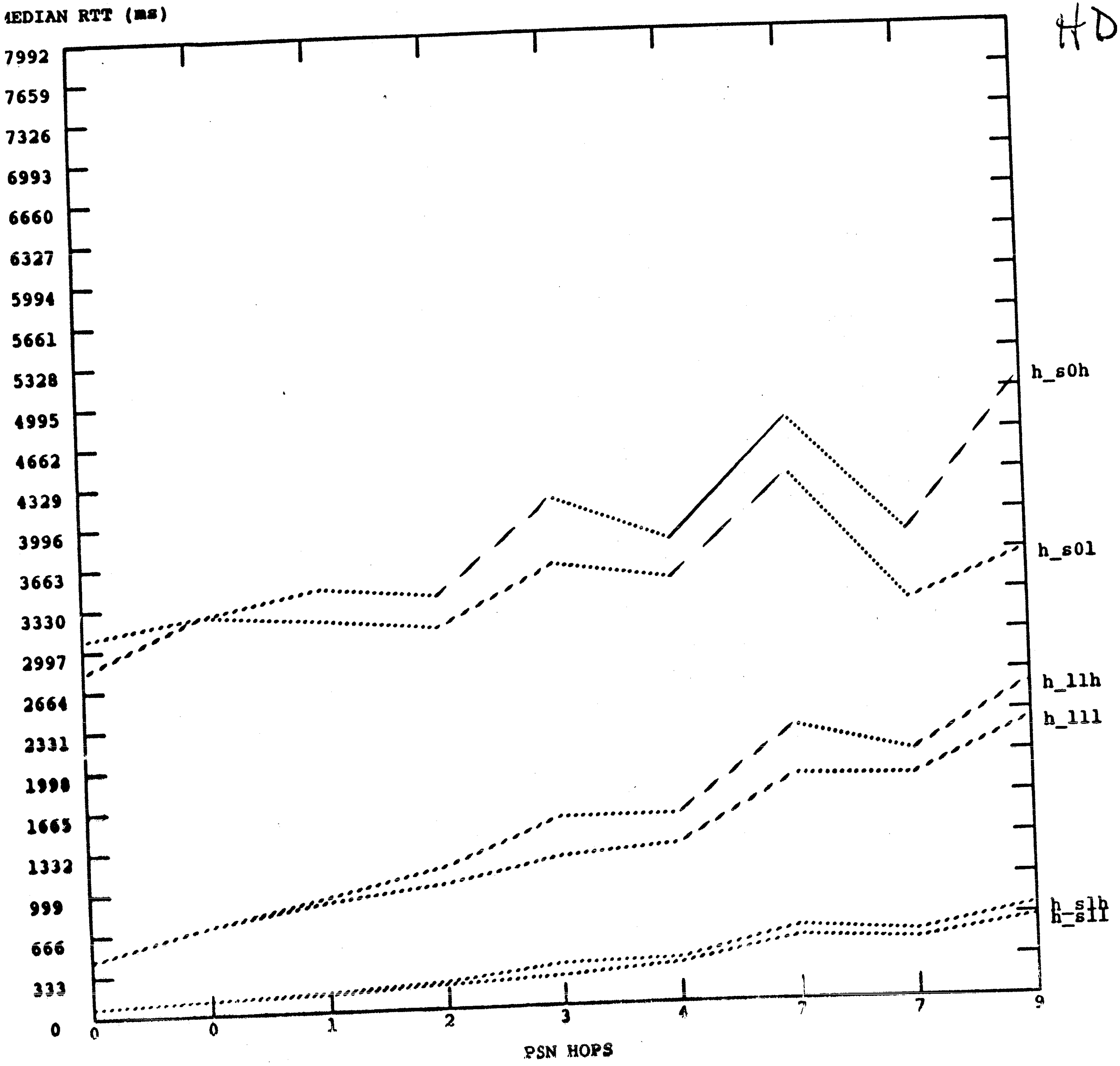
HDH S1L: 2 hops



1822

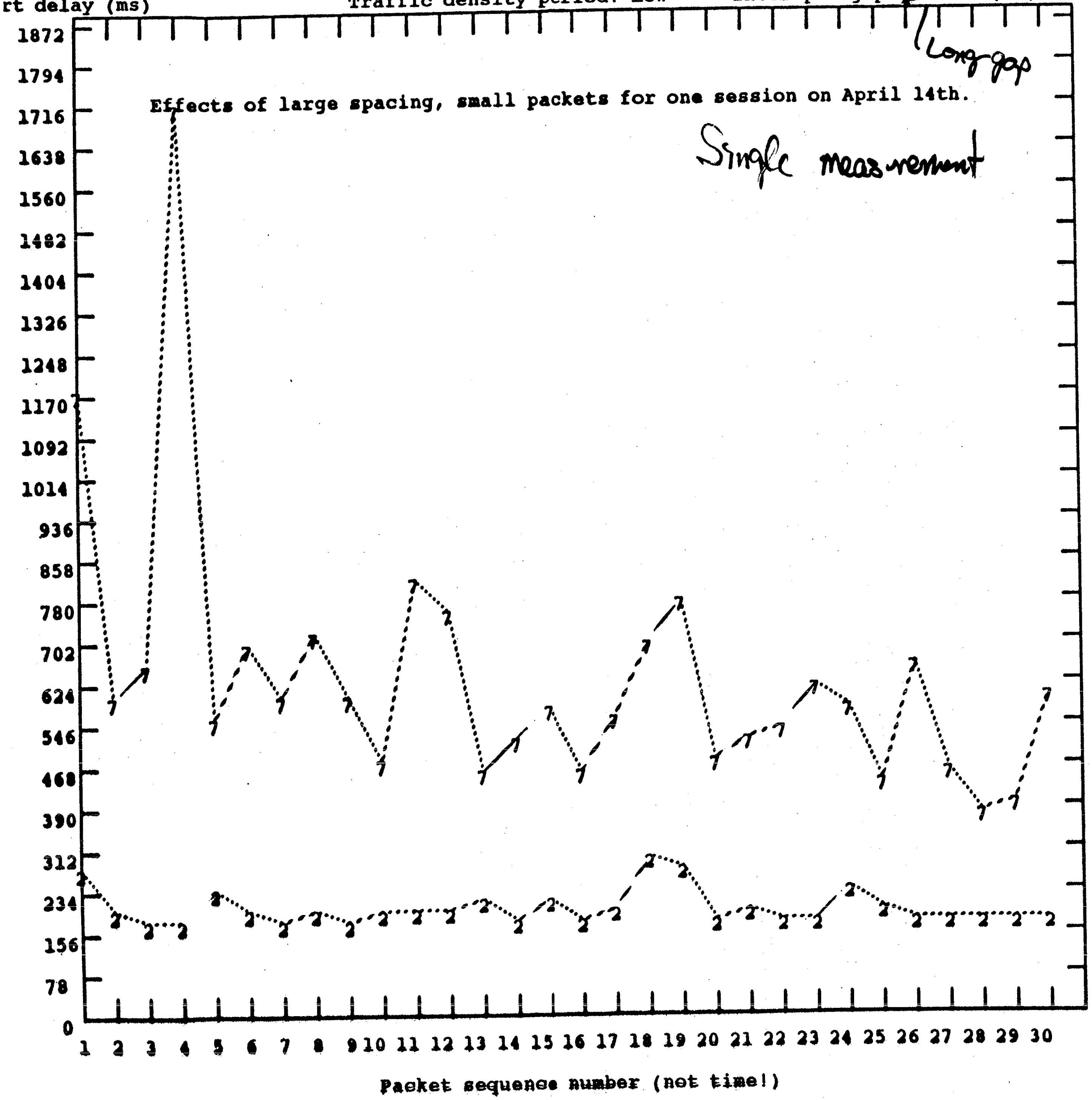

MEDIAN RTT (ms)
7992
7659
7326
6993
6660
6327
5994
5661
5328
4995
4662
4329
3996
3663
3330
2997
2664
2331
1998
1665
1332
999
666
333
0
0
1
2
3
4
7
7
9
PSN HOPS
a_llh
a_lll
a_s0h
a_s0l
a_slh
a_sll

HDH


MEDIAN RTT (ms)
7992
7659
7326
6993
6660
6327
5994
5661
5328
4995
4662
4329
3996
3663
3330
2997
2664
2331
1998
1665
1332
999
666
333
0
0
0
1
2
3
4
7
7
9
PSN HOPS
h_s0h
h_s01
h_11h
h_111
h_s1h
h_s11

Source: TERP.ARPA [X.25]
8 hosts tested (0 - 7 PSN hops),
Traffic density period: Low

Packet size: 56 bytes
30 packets sent to each
Inter-pkt gap: ~~140563~~ (ms)



Effects of large spacing, small packets for one session on April 14th.

Long gap

Single measurement

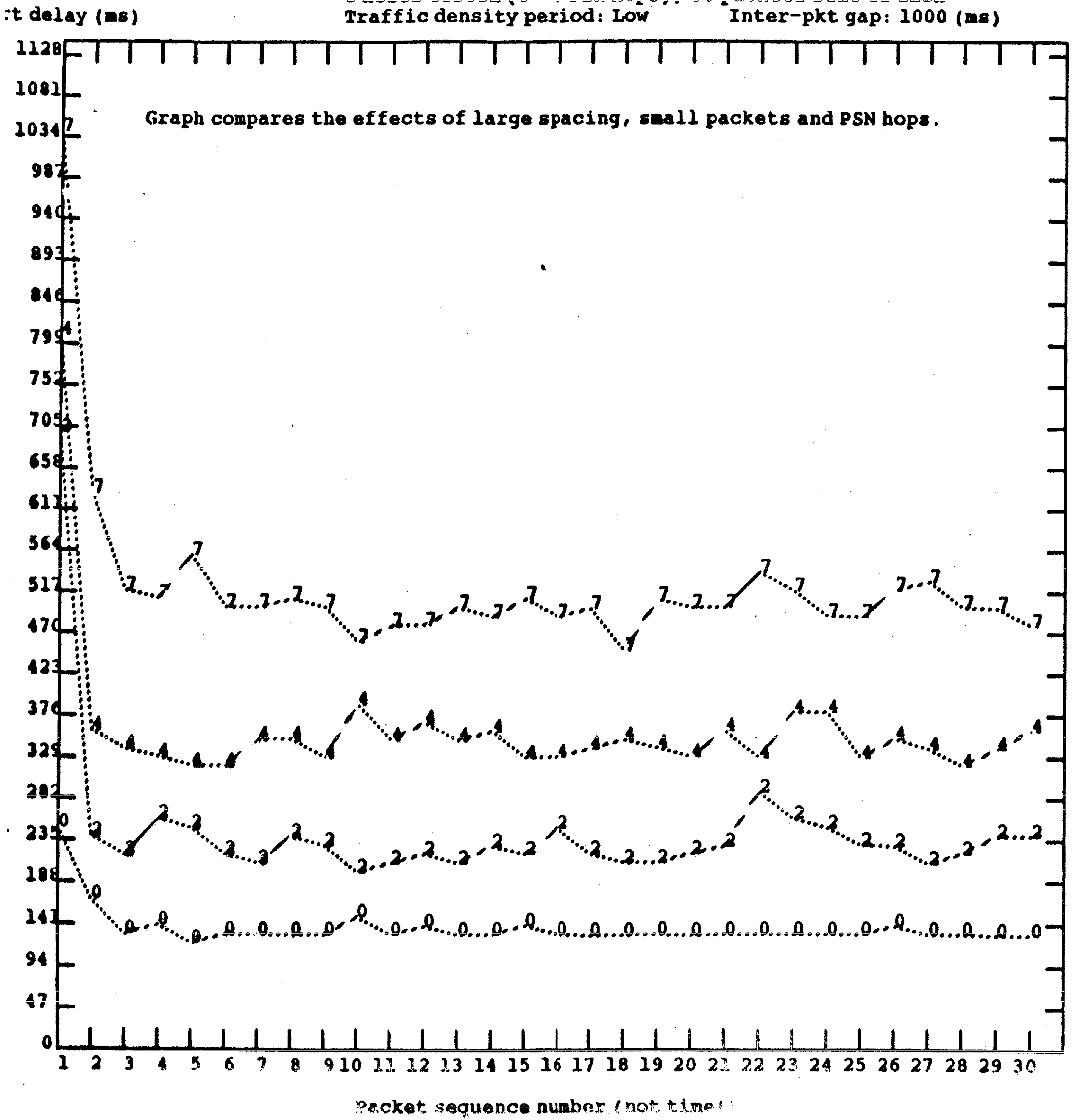
2 PLOT:
Destination: UDEL-HUEY.ARPA

Total sent: 30
Median rt delay: 199
Median rt bits/second: 3778

7 PLOT:
Destination: ISI-VAXA.ARPA

Total sent: 30
Median rt delay: 600
Median rt bits/second: 1253

8 hosts tested (0 - 7 PSN hops), 30 packets sent to each
Traffic density period: Low Inter-pkt gap: 1000 (ms)



0 PLOT:
Destination: DCN1.ARPA

Total sent: 1982
Median rt delay: 130
Median rt bits/second: 6030

2 PLOT:
Destination: UDEL-HUEY.ARPA

Total sent: 1877
Median rt delay: 230
Median rt bits/second: 3408

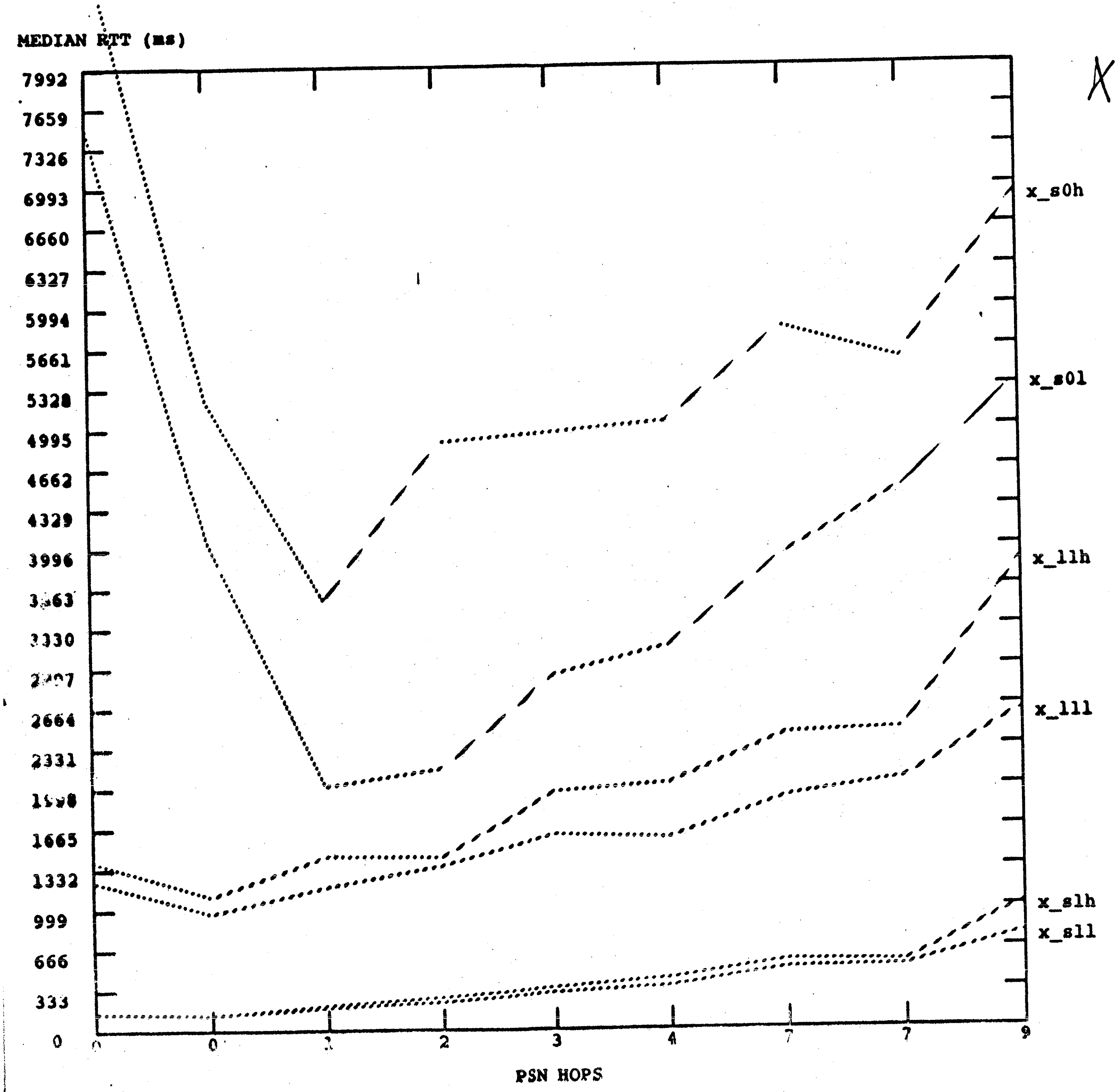
4 PLOT:
Destination: CSNET-SH.ARPA

Total sent: 2118
Median rt delay: 350
Median rt bits/second: 2240

7 PLOT:
Destination: ISI-VAXA.ARPA

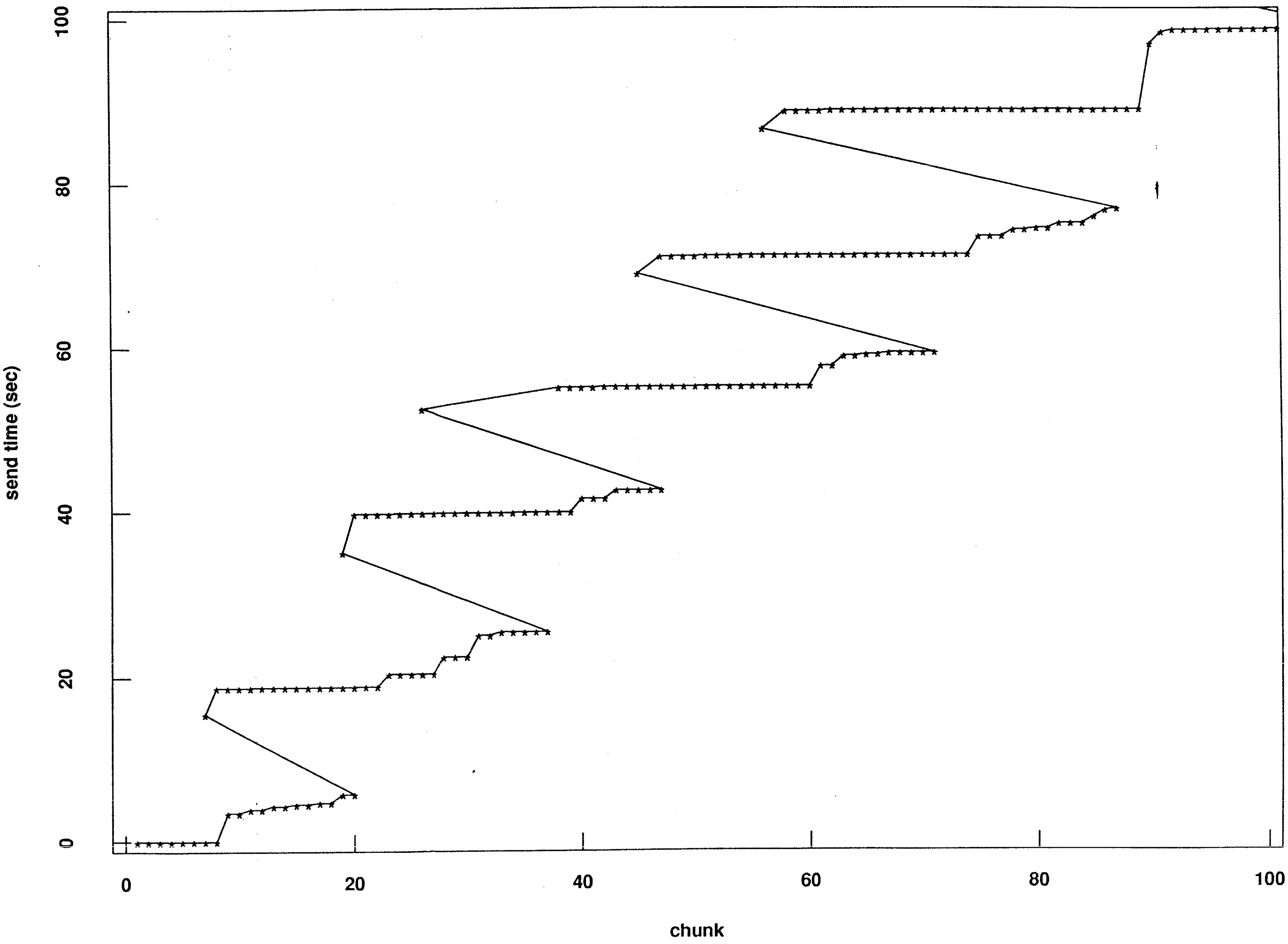
Total sent: 1967
Median rt delay: 510
Median rt bits/second: 1537

X.25


MEDIAN RTT (ms)
7992
7659
7326
6993
6660
6327
5994
5661
5328
4995
4662
4329
3996
3663
3330
2997
2664
2331
1998
1665
1332
999
666
333
0
0
0
1
2
3
4
7
7
9
PSN HOPS
x_s0h
x_s01
x_l1h
x_l1l
x_s1h
x_s1l

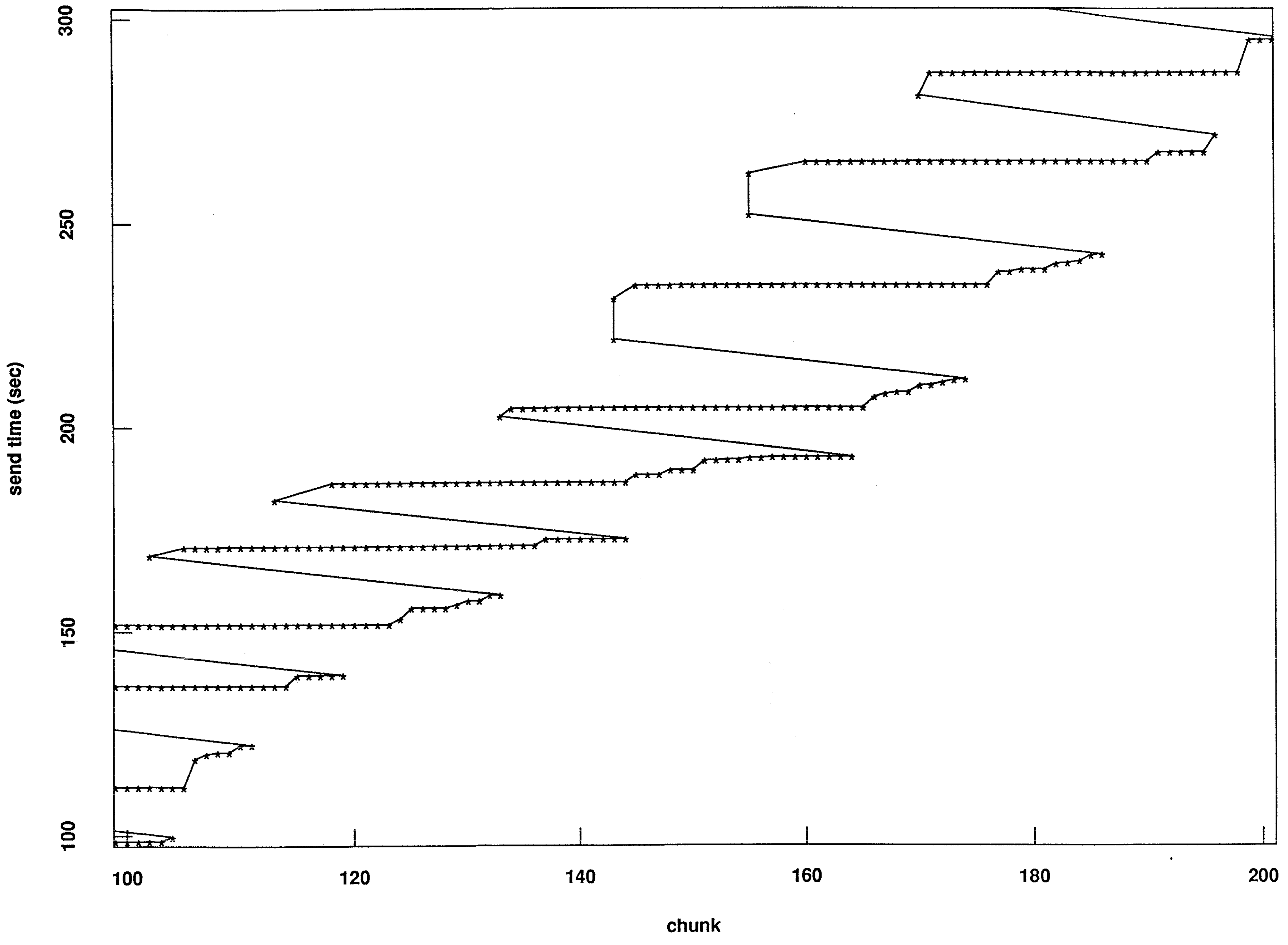
TCP Performance Enhancement Jacobson (LBL)

Original 4.3 Behavior, 16KB Window

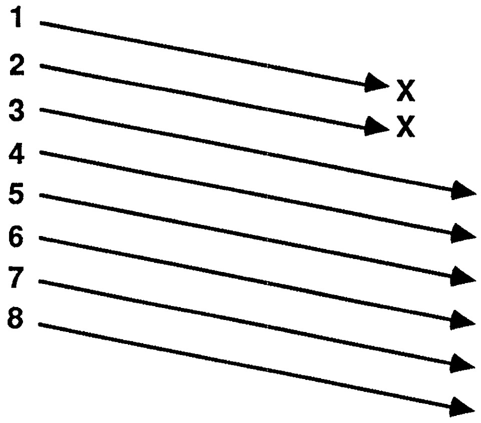


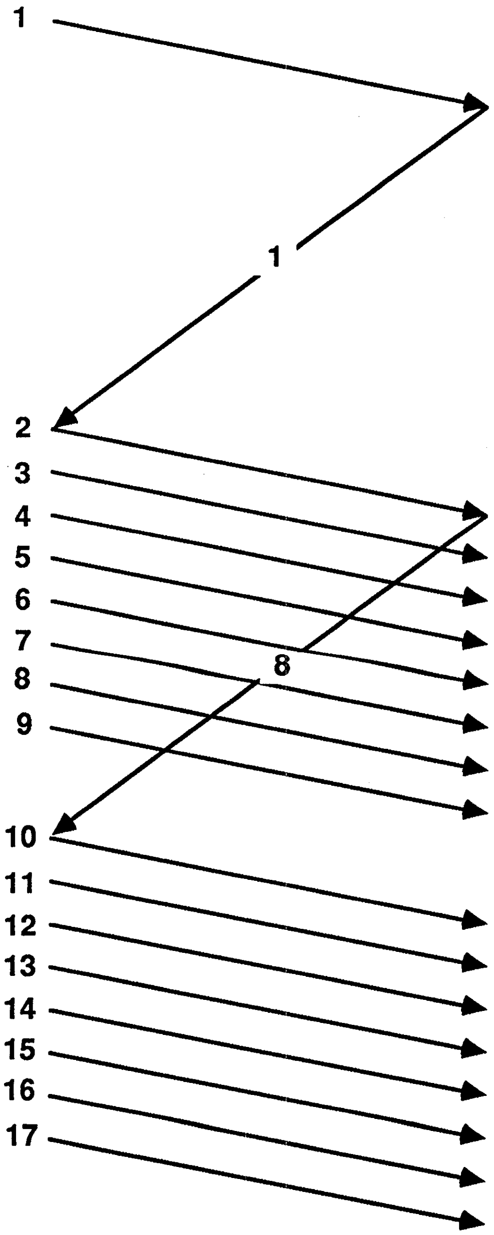
VJ (1)

Original 4.3 Behavior, 16KB Window



VJ (2)

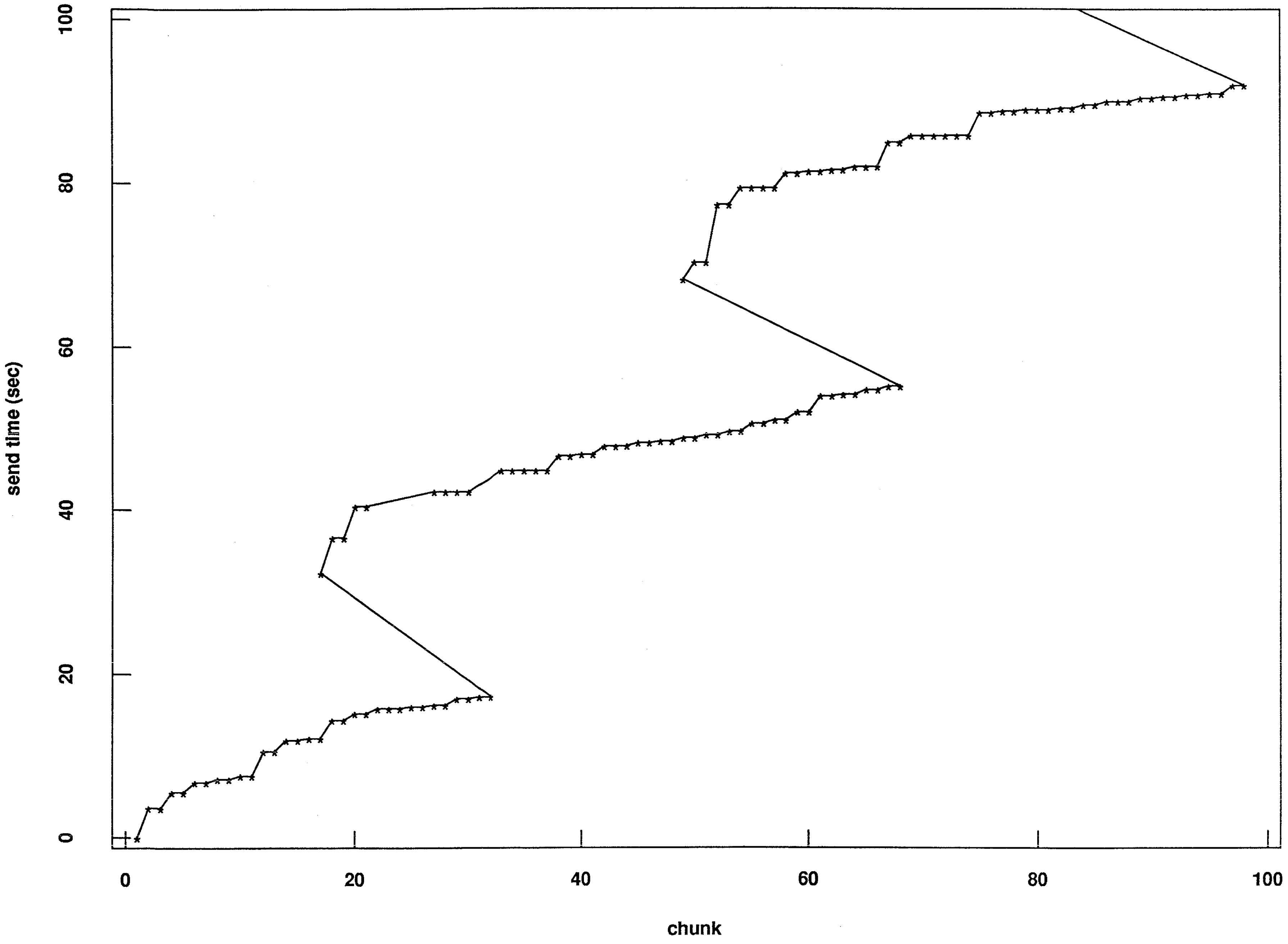




VJ

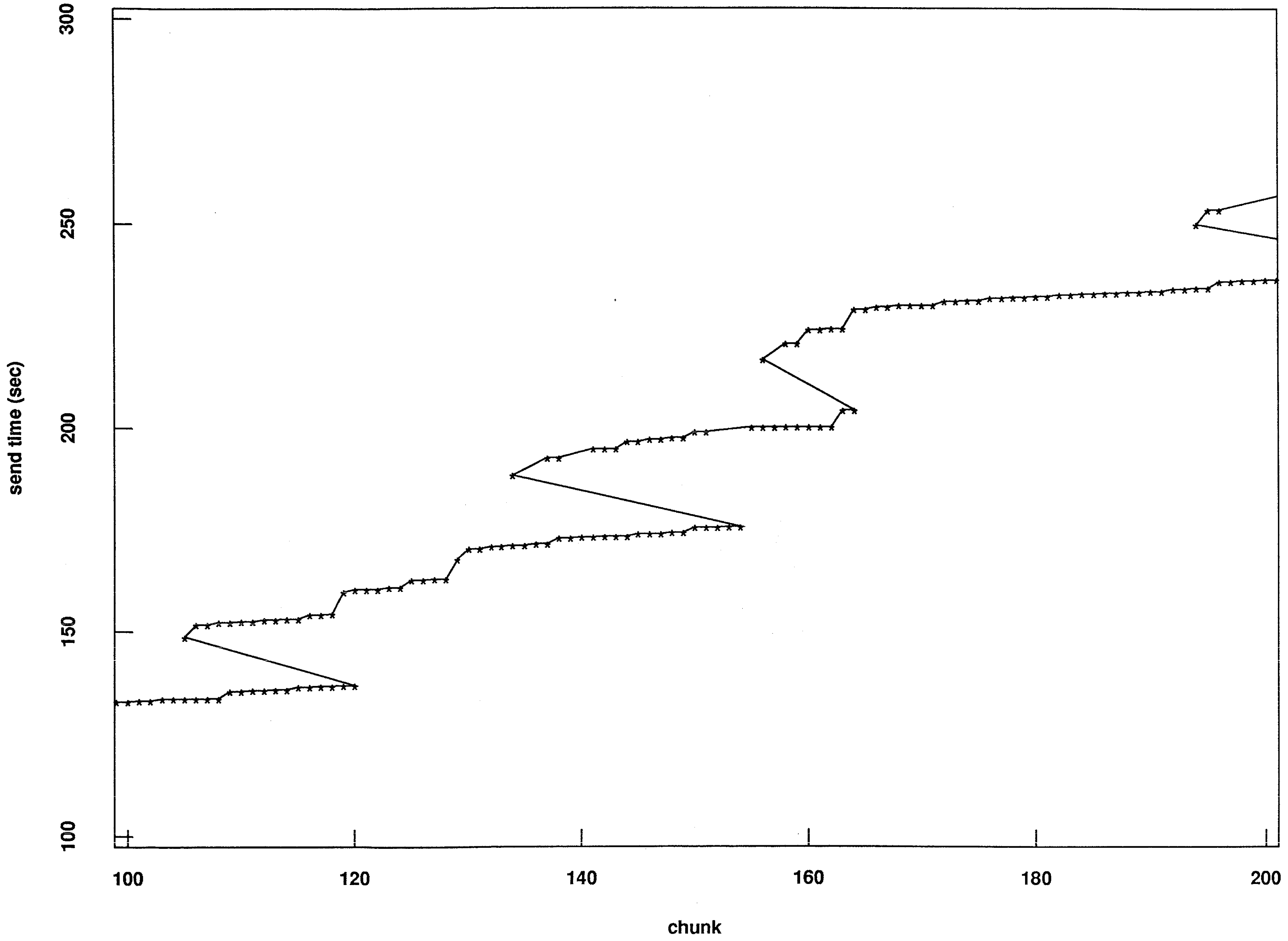
3

Slow Start Behavior, 16KB Window



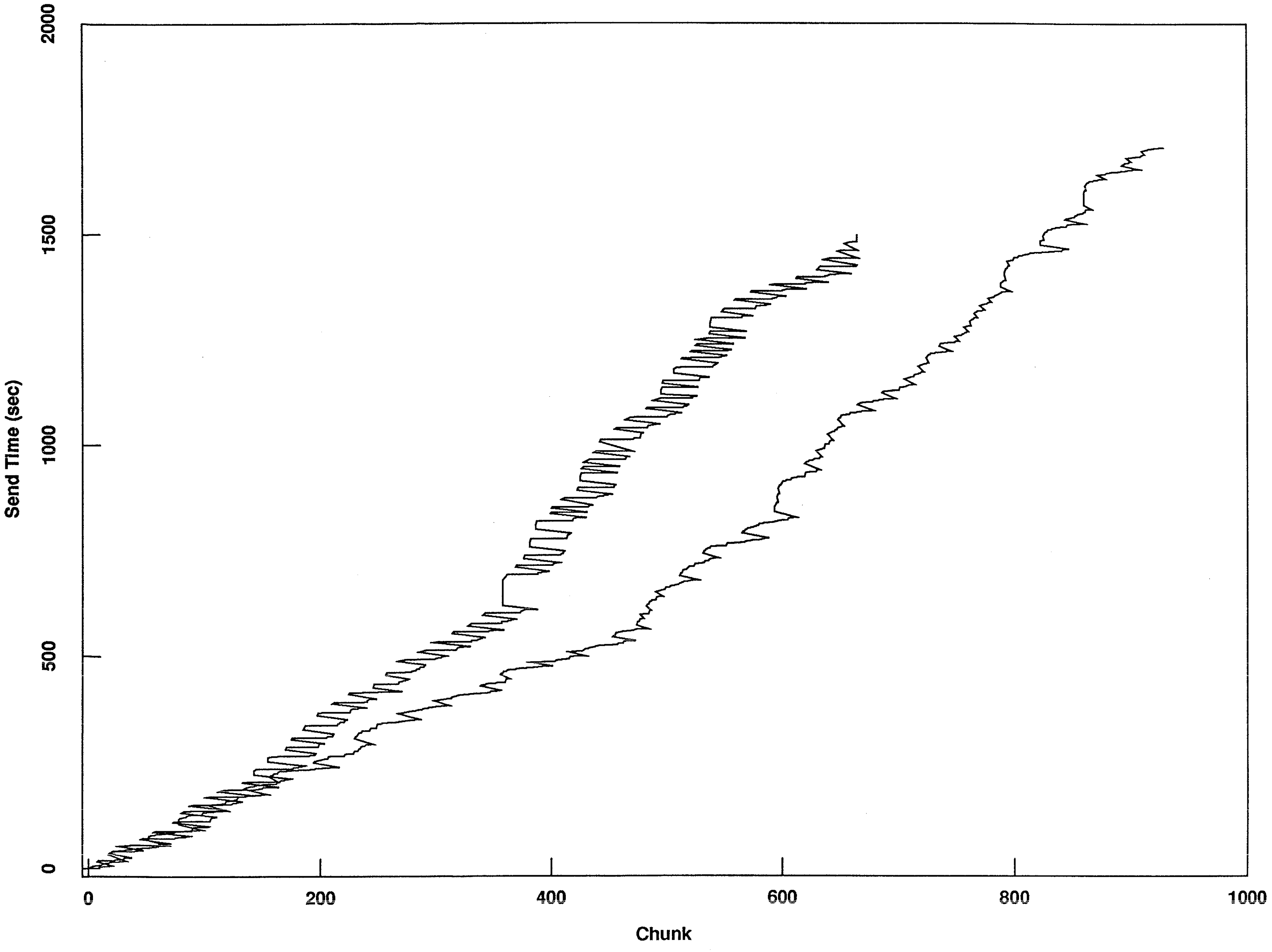
VJ (4)

Slow Start Behavior, 16KB Window



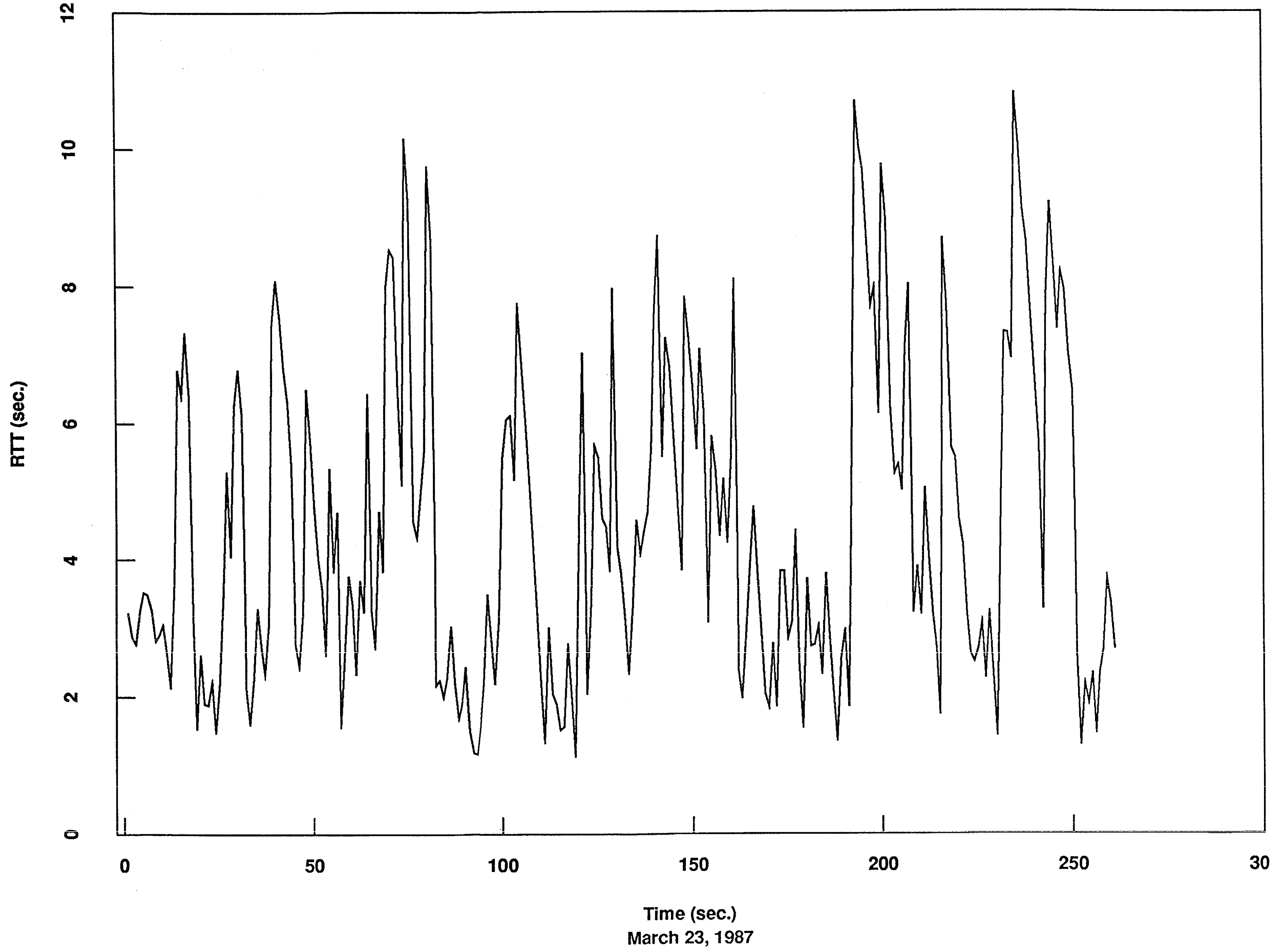
VJ 5

Slow Start and Original 4.3



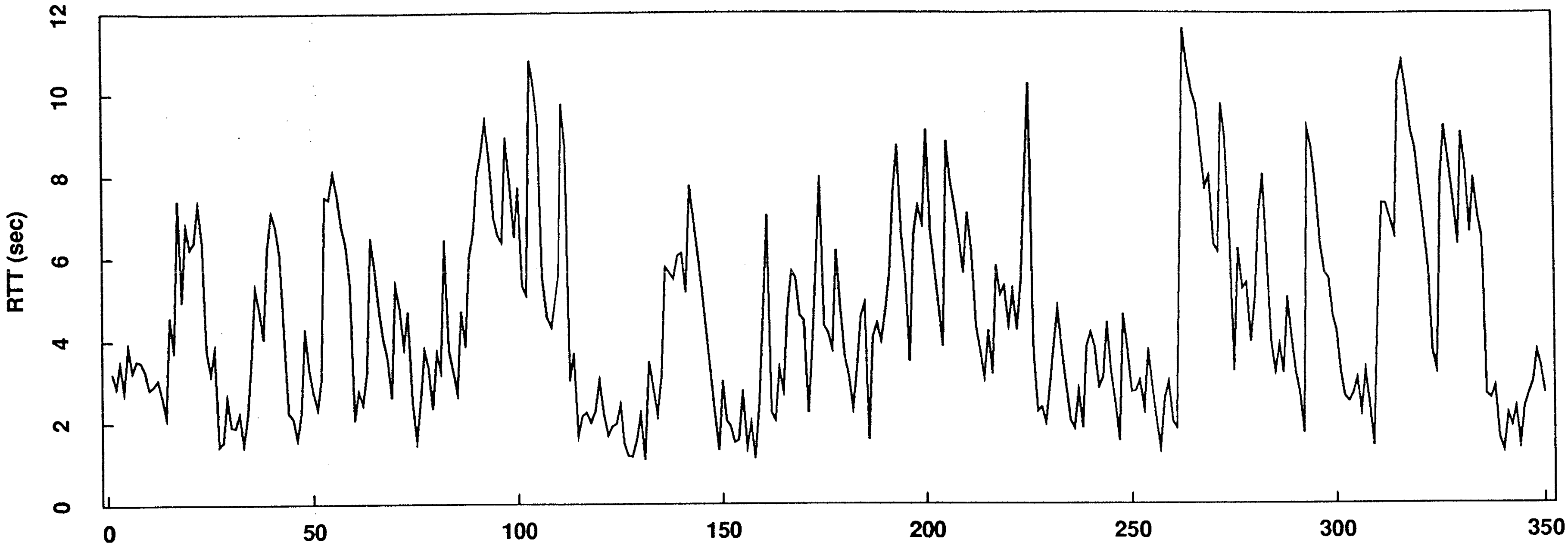
VJ 6

lbl-rtsg - mit-ai ping data

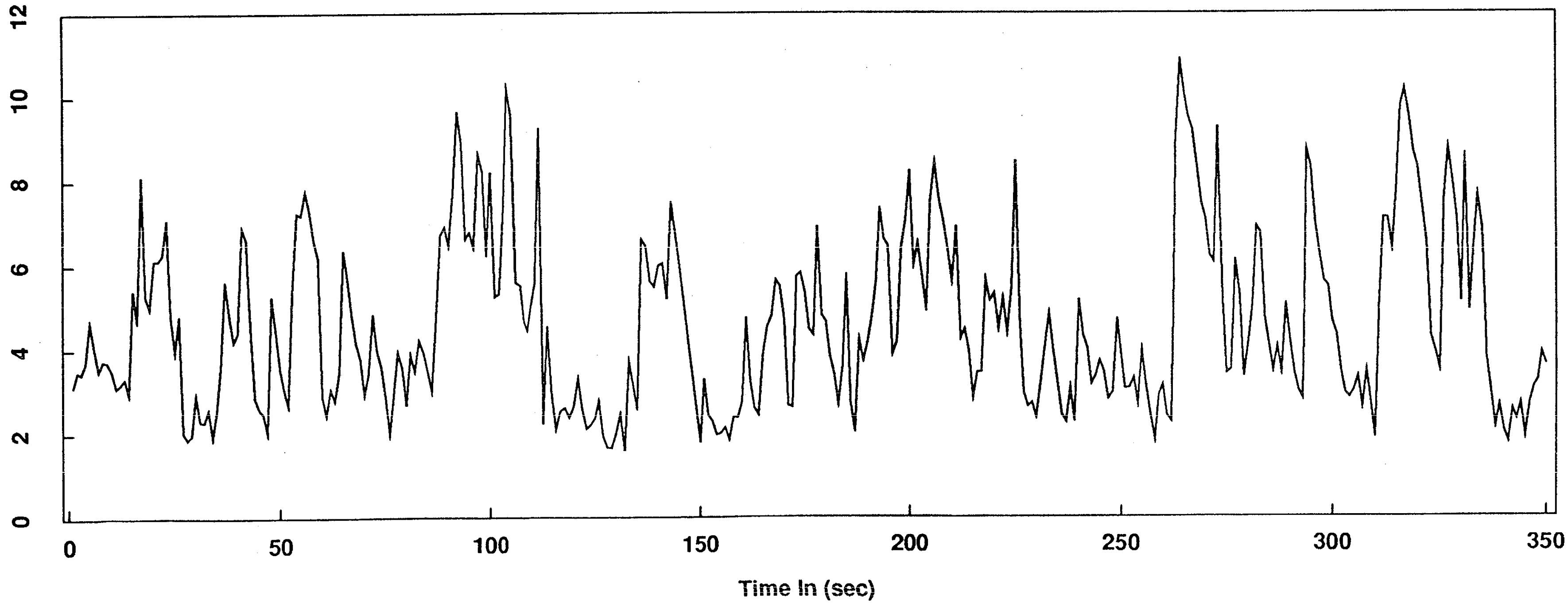


VJ (7)

LBL to MIT-AI RTT Raw Data

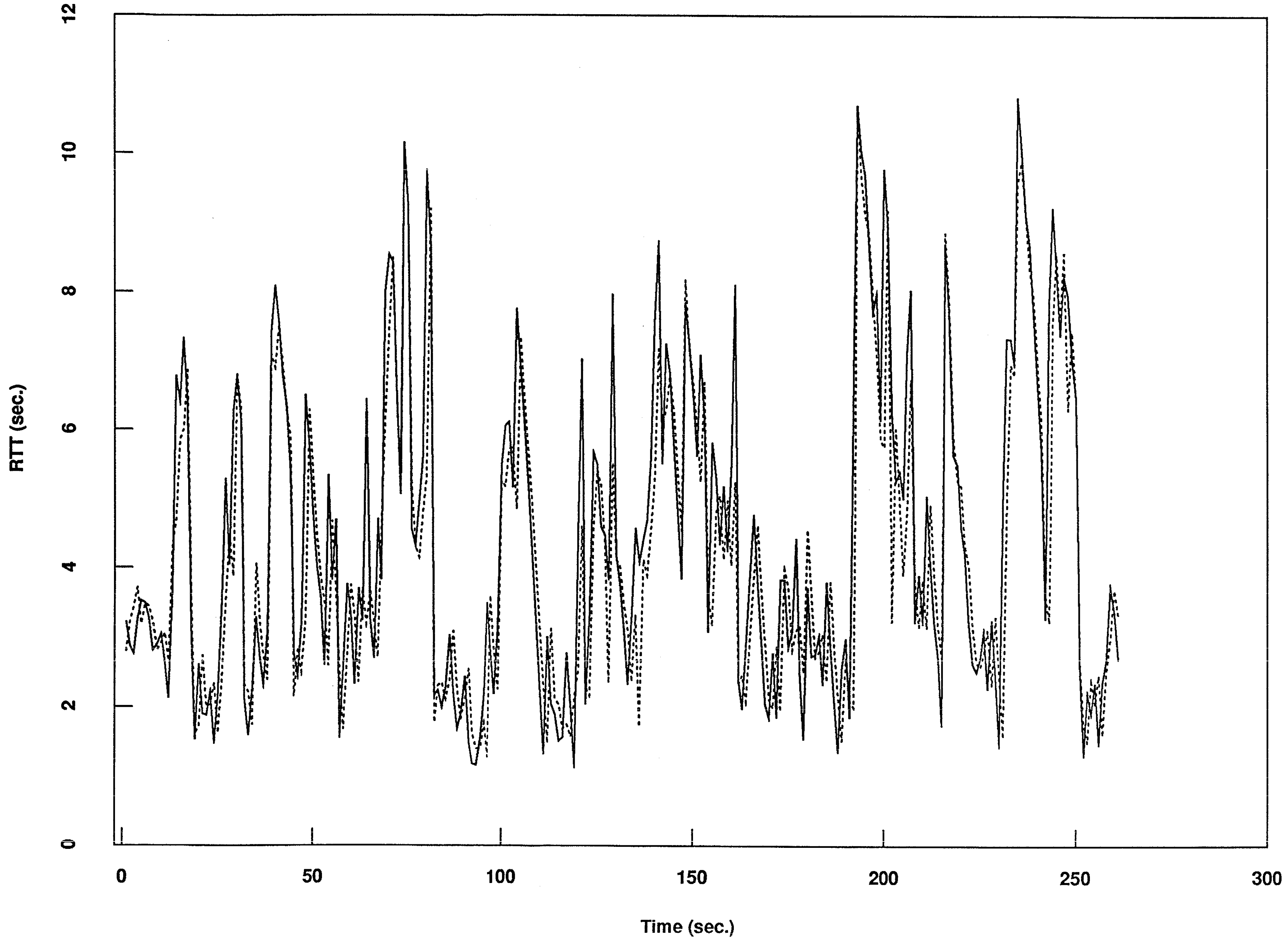


AR(1) Fit to LBL - MIT-AI Data



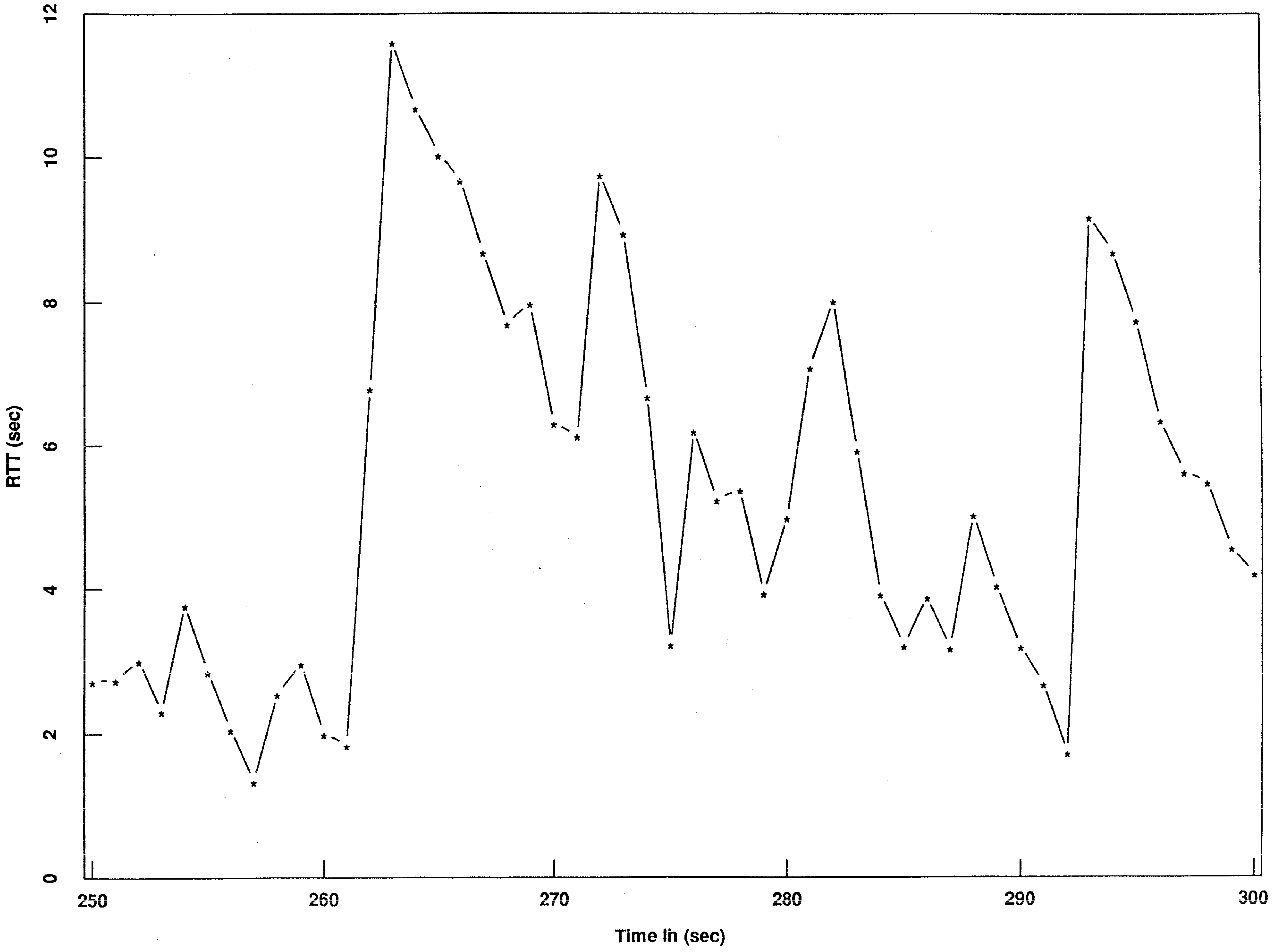
VJ (8)

ping data: raw and fit
(solid line is raw data)



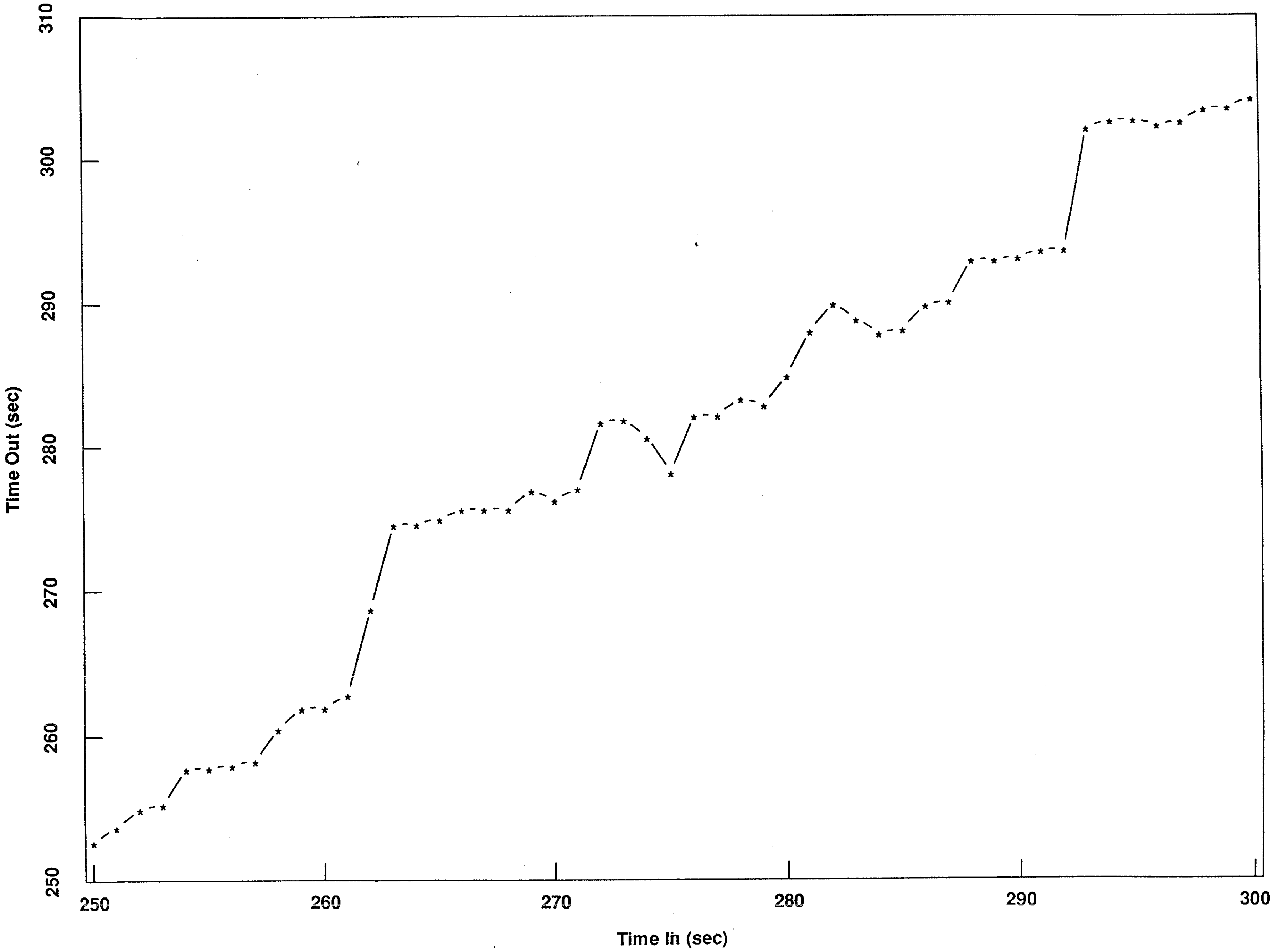
VJ 9

LBL to MIT-AI RTT Data (expanded)



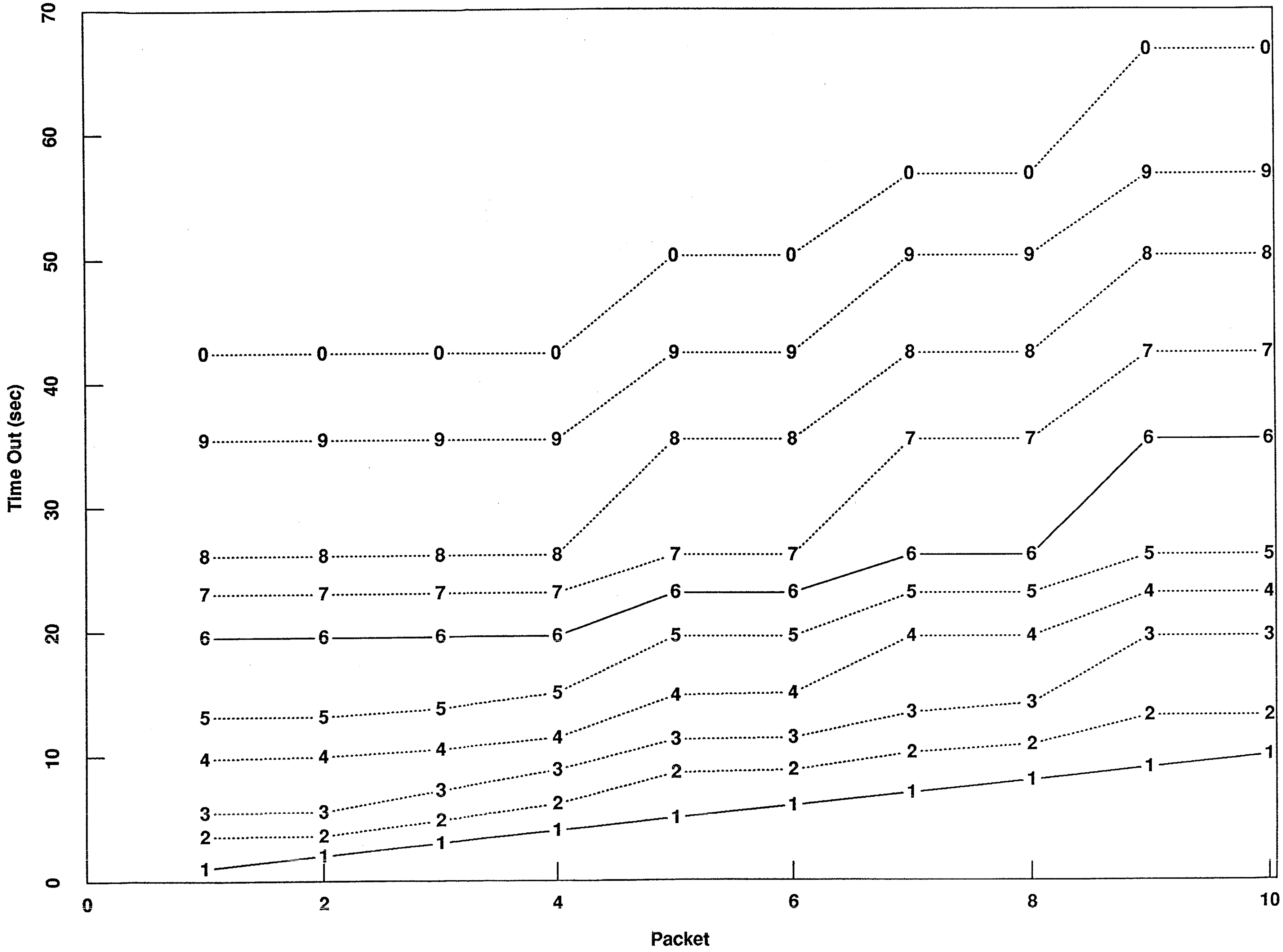
VJ (10)

LBL to MIT-AI Time In vs. Time Out



VJ 11

LBL - MIT-AI Packet Clumping



VJ (12)

Gateway Monitoring Partridge (BBN)

The High-Level Entity Monitoring System

- Query processor at any (every?) entity with an IP address.

- Smart applications that know how to query entities.

- Trap mechanism for unsolicited notification.

Query Processor

- Designed to be as simple as possible consistent with supporting effective monitoring. Don't want to make it too expensive to put in gateways.

- Takes queries which request values to be returned. Answers are immediate - no deferred or repeated replies.

- List of supported values is standardized on a per-entity basis (gateways, hosts, access machines) with a mechanism for adding entity-specific values.

- Queries and replies use ASN.1 (X.409) format. Query processor only has to support limited set of ASN.1 types, so processor should not have to support an ASN.1 compiler.

Query Language

- Queries contain a simple data extraction language. Making query processor intelligent appears to be a win; yields both simpler implementation and greater flexibility.

- Form of language still under debate. Either we will use a stack machine architecture, or a object-oriented approach. Either implementation is straightforward.

Applications

- Expected to be more powerful than query processor. (Need to be able to discover, retrieve and display entity-specific values).

- Probably need to include an ASN.1 interpreter.

Communication Protocols

- An application-specific protocol (HEMP) runs over an existing transport protocol.

- Choices of transport protocols is large. Best choice appears to be one of HMP (the current monitoring protocol), VMTP (a transaction protocol), and RDP or TCP (robust reliable transport protocol). The current plan is to use VMTP as the initial protocol for the initial experimentation and then re-evaluate choice based on initial experience.

Traps

- Critical nodes (like gateways) often want to be able to relay useful status information to monitoring centers.

- Reliability useful but not critical.

Current Schedule

- Overview RFC in late draft stage. Available first week of May?

- HEMP RFC draft available but needs reworking. Available late May?

- Query processor specificiation and list of values for gateways in early draft stage. Available in June?

- Test implementations planned for this summer.

- RFC to be revised this fall based on experience.

Management Architecture LaBarre (MITRE)

IAB SYSTEMS MANAGEMENT WG

PURPOSE:

Develop set of RFCS defining tools for managing systems of multiple vendor TCP/IP products.

FOCUS:

Define management framework

Define parameters to control/monitor the network

Define management protocol

Background

Stan Ames

successful initially vendor coop to develop Methods RF [illegible]

try again to do same thing with rel ergonomics

all said yes, etc conditions

then one vendor said ok I would want it

· HA - Vendors then said since

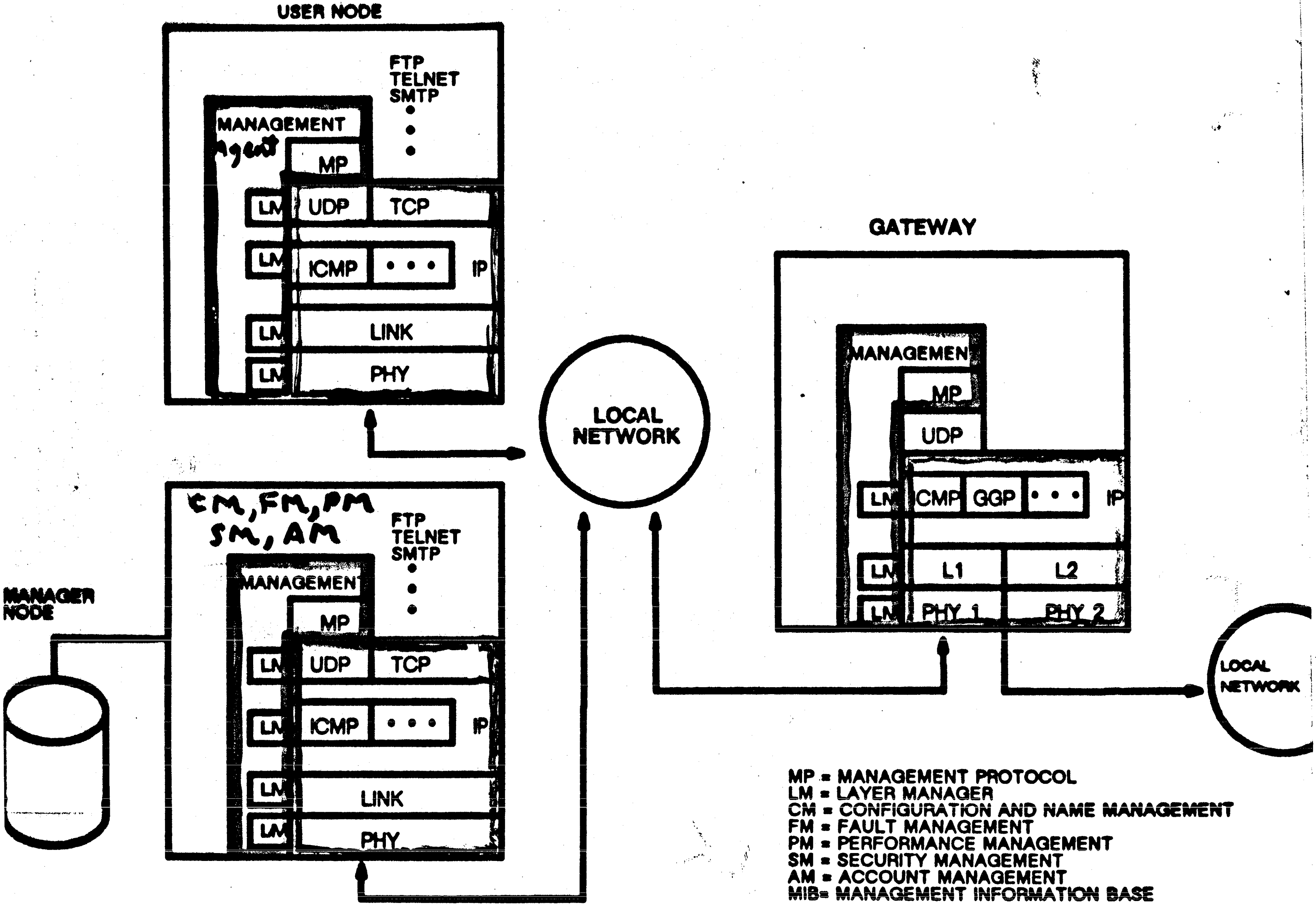
have vendor independent group

Stan Ames proposed that Don Lynch and I act as Lightning Rods

TASKS

- AGREE ON SYSTEMS MANAGEMENT FRAMEWORK
- AGREE ON SCOPE OF MANAGEMENT
- DEFINE MANAGEMENT INFORMATION
- DEFINE MANAGEMENT PROTOCOL
- DEVELOP RFC(S)

STRAWMAN MANAGEMENT FRAMEWORK



MP = MANAGEMENT PROTOCOL
LM = LAYER MANAGER
CM = CONFIGURATION AND NAME MANAGEMENT
FM = FAULT MANAGEMENT
PM = PERFORMANCE MANAGEMENT
SM = SECURITY MANAGEMENT
AM = ACCOUNT MANAGEMENT
MIB= MANAGEMENT INFORMATION BASE

(1) Assume IP gateway connecting 2 networks

(2) We show one host node &

one manager node

(3) Nodes contain the Internet protocols as outlined in green

(4) We impose a management framework ~~as~~ follows (in Red)

(A) a <u>management agent</u> in each node

(B) layer management (LM) entity for ~~each~~ layer

(C) each layer may have ~~internal~~ management protocols (eg ICMP, GGP for network layer)

~~5~~B (D) a manager station that contains management applications (CM, FM, PM, SM, AM)

(E) Distributed management ~~information~~ information Base (MIB)

(5) The manager interacts with agents via a management protocol (MP) to <u>Get</u>, <u>Set</u> parameters, ~~and~~ receive <u>events</u> generated due to errors and initiate actions
(Shown on UDP)

APPROACH

- Don't Reinvent the Wheel
- Identify, form liaisons, other mgmt activities
- Agree on framework
- Solicit input on mgmt info for TCP/IP ...
- Examine, select, augment a management protocol

  CMIP/ROS, HMP, IEEE 802.1
  ASN.1

- Generate RFCs

- SCHEDULE

  monthly meetings for 6-12 months

Kickoff Meeting

Date: May 5

Time: 9-4

Place: TECHMART
SUITE 125
5201 Great America Parkway
Santa Clara, CA

Contact: Dan Lynch (408) 996 2042 (lynch@a.isi.edu)
Lee LaBarre (617) 271 8507 (cel@mitre-bedford

MEETING AGENDA

- Systems Management Framework Proposal
- Strawman scope, goals, schedule for WG
- Proposal for TCP/IP Systems Management
- TCP/IP management parameters
- Protocols for management exchanges
- Report on IETF activities
- Organizational Details

IETF status/overview Gross (MITRE)

DARPA Task Forces

- Outgrowth of DARPA Internet Research Program
- Govemed by Internet Activities Board (IAB)
- Composition of IAB

  Dave Clark (MIT) - Chair

  Jon Postel (ISI) - Co-Chair

  Task Force Chairs

  Agencies - Darpa, NSF

Task Force List

* Internet Architecture (INARC)

* Internet Engineering (INENG)

Autonomous Networks

Privacy

Supercomputers

End-End

* Five-Year

Testing

Tactical Internet

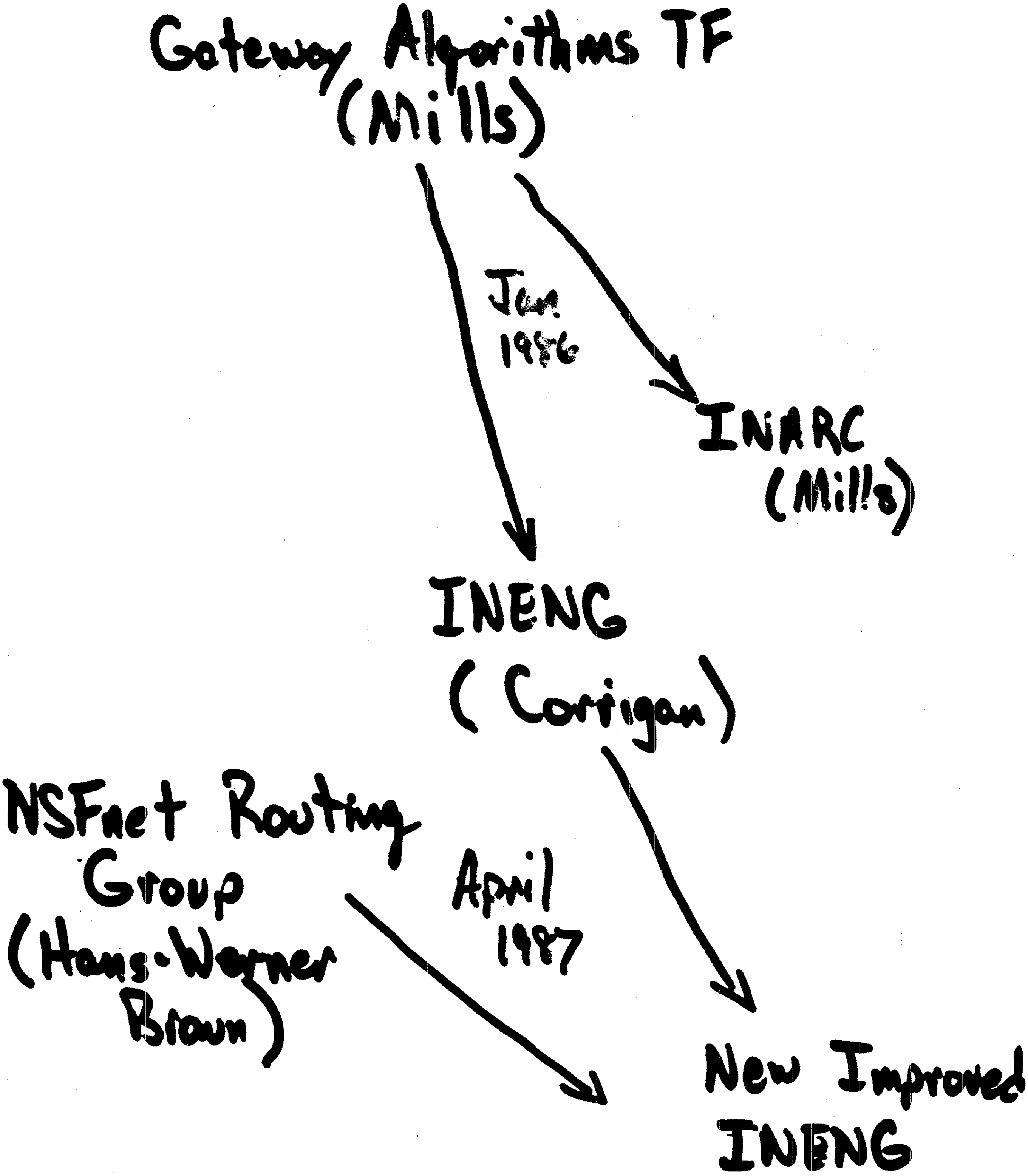
Generic TF Charter

- Identify areas for development
- Propose funding
- Guidance to ongoing work
- Position papers on approach

INENG Charter

- Short to intermediate problems in current Architecture
- Spans O+M to short range applied research

INENG History



Gateway Algorithms TF
(Mills)

Jan.
1986

INARC
(Mills)

INENG
(Corrigan)

NSFnet Routing
Group
(Hans-Werner
Braun)

April
1987

New Improved
INENG

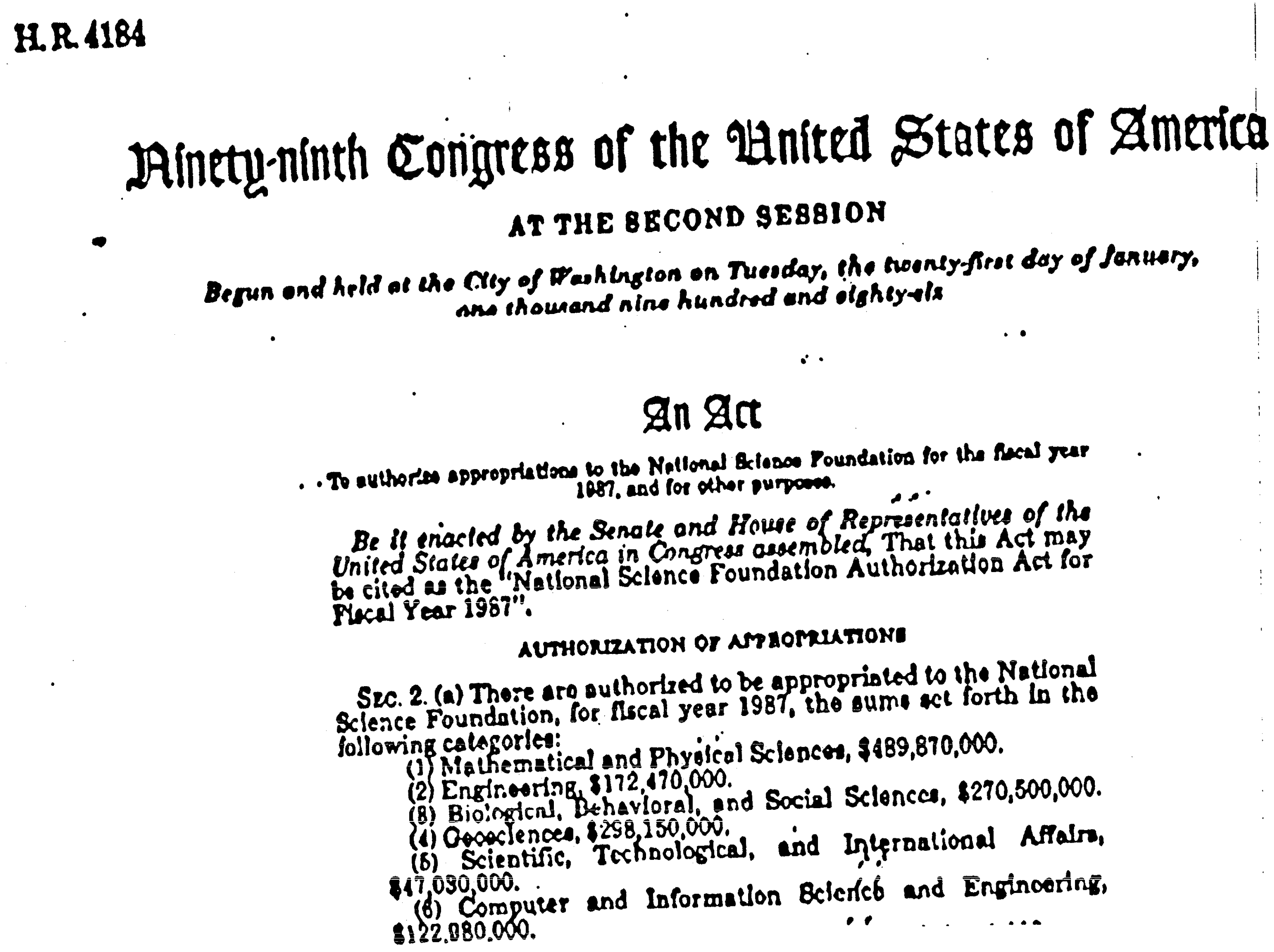
INENG Direction

- Under Review
- Working Group Format
- Current Groups
  - ISO Transition
  - Name Domain Planning
  - Routing
  - Monitoring and Management
- X3S3.3

Overview of FCCSET Computer Network Study

MITRE Corporation,
P.Gross
23 Apr 87

Origin of FCCSET Computer Network Study



H.R. 4184

Ninety-ninth Congress of the United States of America

AT THE SECOND SESSION

Begun and held at the City of Washington on Tuesday, the twenty-first day of January, one thousand nine hundred and eighty-six

An Act

To authorize appropriations to the National Science Foundation for the fiscal year 1987, and for other purposes.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled, That this Act may be cited as the "National Science Foundation Authorization Act for Fiscal Year 1987".

AUTHORIZATION OF APPROPRIATIONS

SEC. 2. (a) There are authorized to be appropriated to the National Science Foundation, for fiscal year 1987, the sums set forth in the following categories:

(1) Mathematical and Physical Sciences, $489,870,000.

(2) Engineering, $172,470,000.

(3) Biological, Behavioral, and Social Sciences, $270,500,000.

(4) Geosciences, $298,150,000.

(5) Scientific, Technological, and International Affairs, $47,030,000.

(6) Computer and Information Science and Engineering, $122,980,000.

Networking Requirements and Future Alternatives Panel

- Vendors
  - Contel, MCI, AT&T, BBN
- Users
  - DoD, DOE, NASA, Gov't Labs
- Other Experts

Report and Recommendations

- Advocate establishment of interagency networking facility with 15 year mission to
  - — Ensure U.S. scientists have most advanced networking facilities available
  - — Ensure U.S. networking technology maintains position of world leadership

- For second year of study, establish interagency coordinating committee
  - — identify and resolve short-time implementation issues

Attachment

- Invitation Letter
- Congressional Bill
- San Diego Agenda
- Panel Descriptions
- Draft of Future Requirements Panel Report

ANSI routing architecture Tsuchiya (MITRE)

ANSI X3S3.3 Draft Routing Architecture

MITRE Corporation,
Paul F. Tsuchiya

Outline

- Where this architecture fits in
- Routing Hierarchy Structure
- Addressing

Where Architecture Fits In

- Currently only ANSI document
  - — Gives our current thinking on the global routing problem
  - — Provides common terminology and ideas for discussion
  - — Provides GUIDANCE in specifying routing protocols (Not chiseled in stone)

- Sister document, Routing Framework, is what is seen in ISO
  - — For tutoring ISO people
  - — For mollifying opposition

- Not clear what status of Architecture will be in ISO
  - — We hope it takes back seat, if any seat at all

Main Principles Behind Architecture

- Assumes Internet Architecture (ISs and ESs and subnetworks)

- Assumes more ISs than ESs, and ISs are more complex than ESs
  - — Put protocol complexity in ISs, let ESs be dumb and simple

- Different sets of ISs and ESs need autonomy (Autonomous Systems)
  - — Multiple (some of them non-standard) routing "islands" glued together by common routing protocol (sound familier?)

- Administrative entities very important
  - — Routing Firewalls between administrations

Elements of Routing Hierarchy

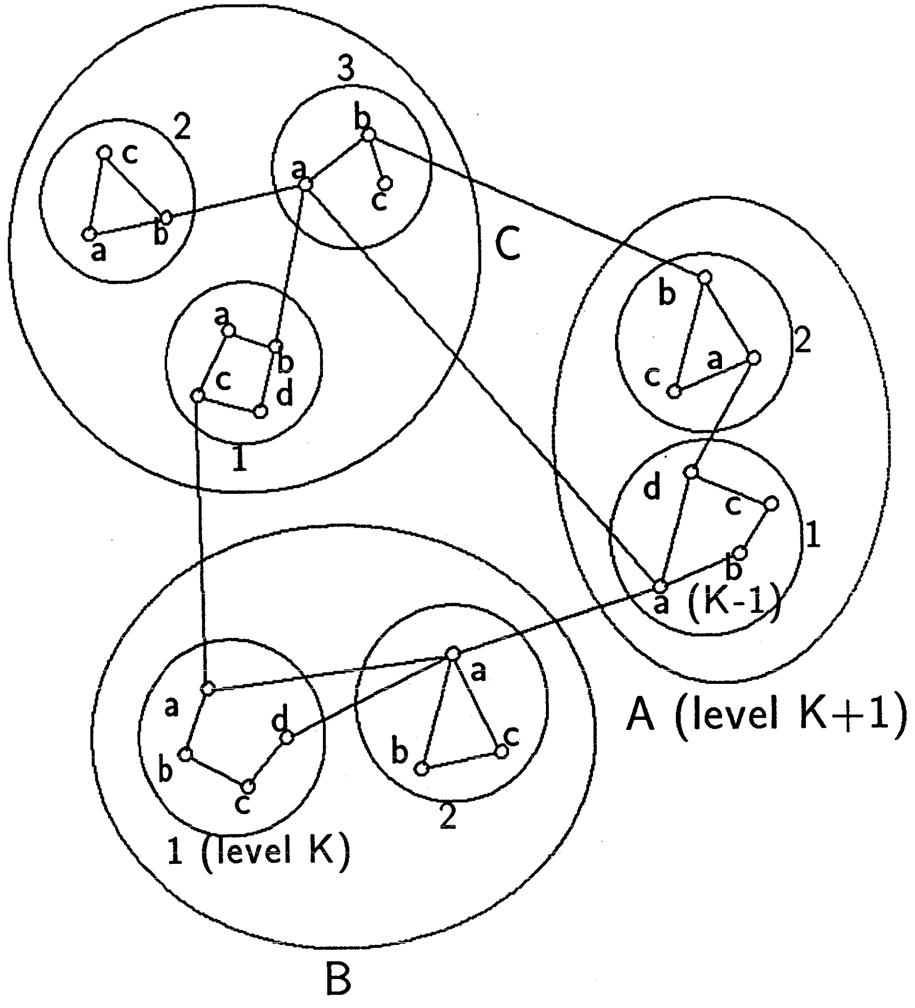
- Three-tiered functional routing structure (n-tiered hierarchical routing structure)
  - — Tier 1 - ES-IS routing
    - - Full trust. Lets ESs and ISs find each other. Simple.
  - — Tier 2 - Homogeneous IS-IS
    - - Full Trust. ISs run same Routing Procedures. High quality routing (based on network distance, not administrative distance)
  - — Tier 3 - Heterogeneous IS-IS
    - - Autonomy required. May be little or no trust. Firewalls important. Routing based more on administrative distance than network distance.

Hierarchical Structure

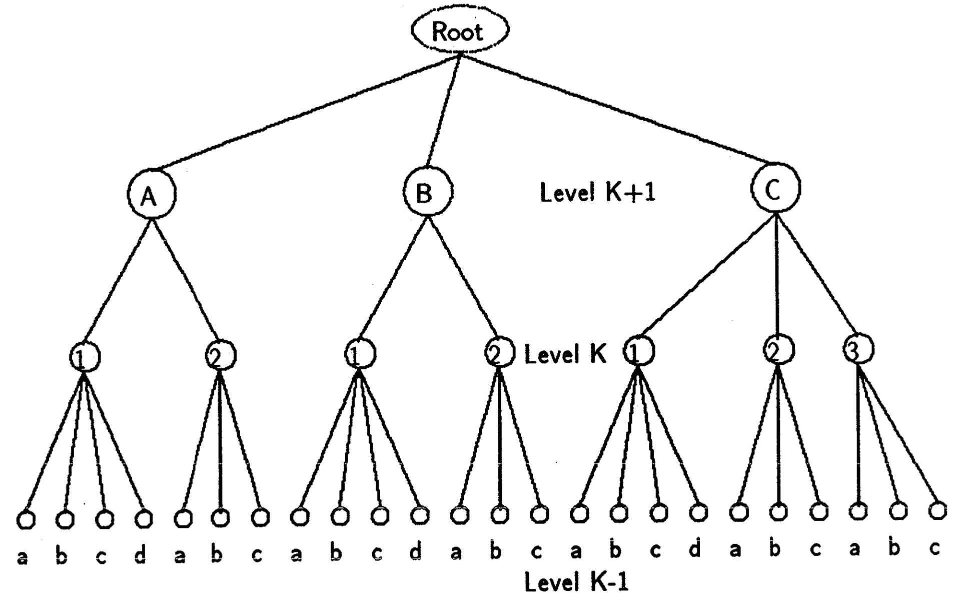
- Graph-theoretically speaking, all levels of hierarchy are the same
  - — Same rules for structuring hierarchy
  - — Same rules for addressing elements of hierarchy

- Functionally speaking, levels of hierarchy have different functions
  - — More-or-less in line with 3 functional tiers, but more detailed

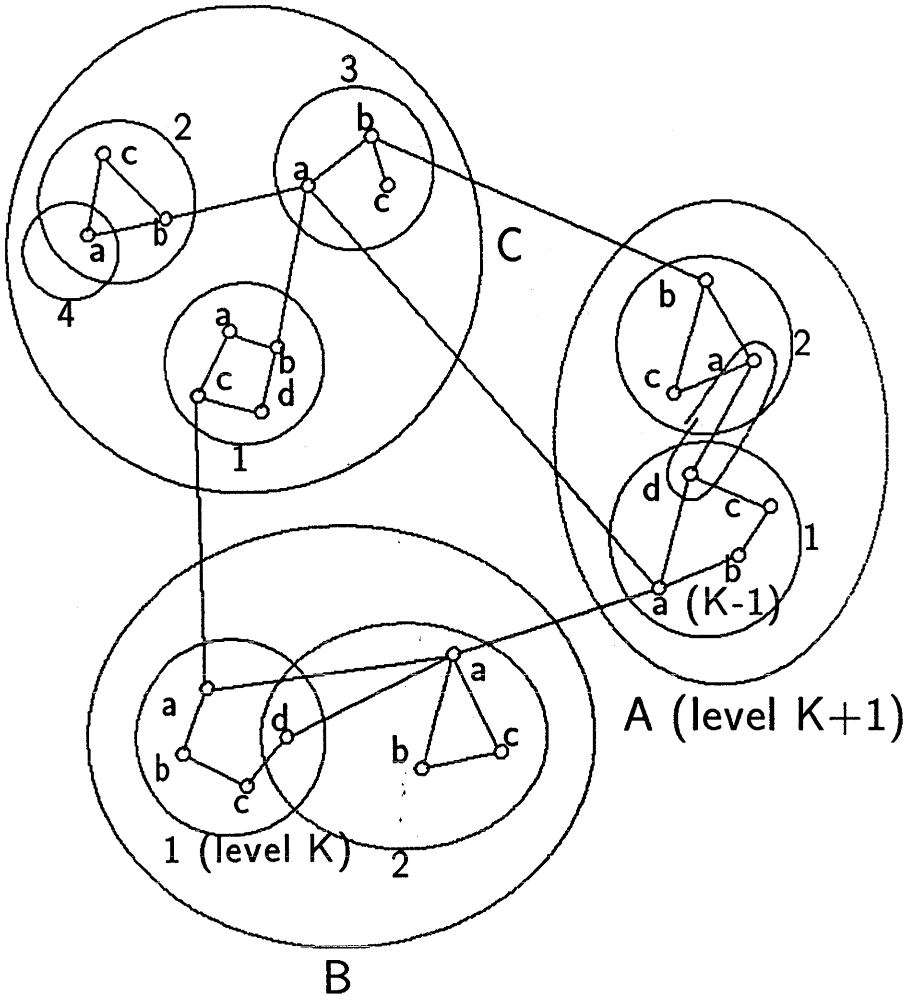
Hierarchy - Topological Representation
No Variations



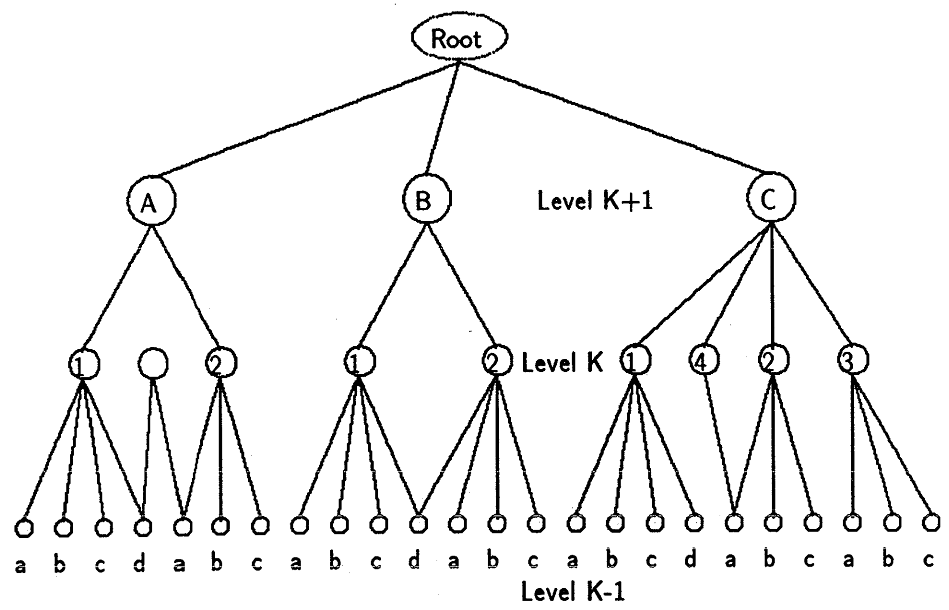
Hierarchy - Logical Representation
No Variations



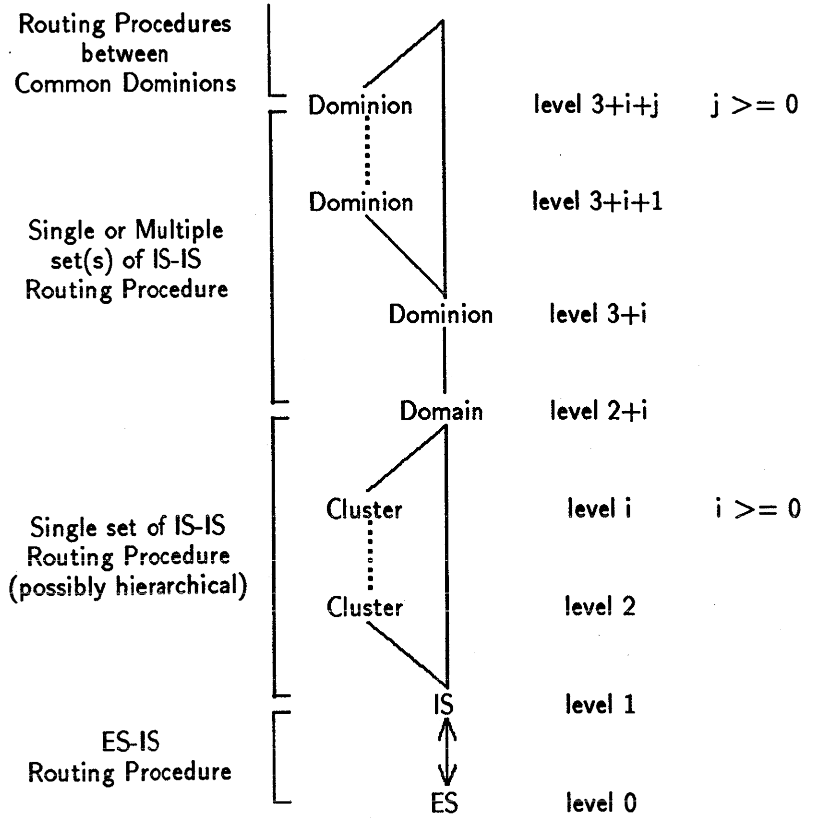
Hierarchy - Topological Representation With Variations



Hierarchy - Logical Representation With Variations



Levels of Hierarchy - Picture



Routing Procedures between Common Dominions

Single or Multiple set(s) of IS-IS Routing Procedure

Single set of IS-IS Routing Procedure (possibly hierarchical)

ES-IS Routing Procedure

Dominion level 3+i+j j >= 0

Dominion level 3+i+1

Dominion level 3+i

Domain level 2+i

Cluster level i i >= 0

Cluster level 2

IS level 1

ES level 0

Levels of Hierarchy

- Lowest Level - End System (mandatory)
- Next Level - Intermediate System (mandatory)
- Cluster (optional)
  - — Group of ISs, formed within a Domain for the purpose of accommodating large numbers of ISs.
- Domain (mandatory)
  - — Group of ISs (or Clusters) formed primarily because they share a common set of Routing Procedures

Levels of Hierarchy (cont)

- Dominion (at least one, may be Common Dominion)
  - — Group of one or more Domains. Formed because all Domains in Dominion fall under control of single Routing Authority
- Additional Dominions (optional)
  - — Formed of lower level Dominions to reflect administrative hierarchy
- Common Dominion (mandatory)
  - — Highest level hierarchy element. Formed of lower level Dominions or Domain(s). No higher Routing Authority exists for this group of ISs and ESs (i.e., there is no single "root" to the global routing hierarchy).

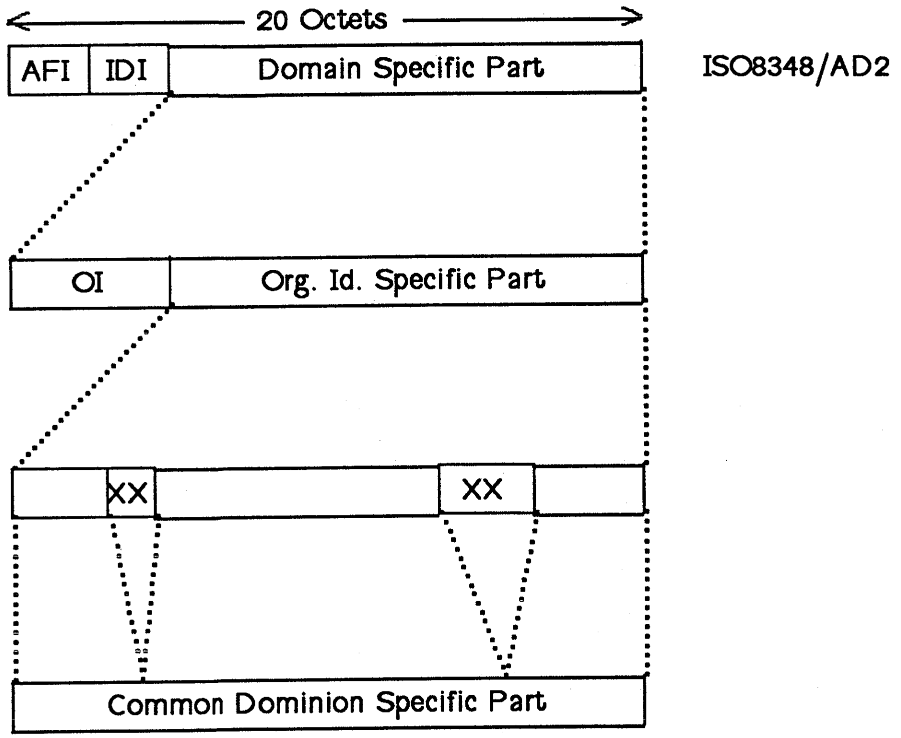
Routing Hierarchy Example

COMMON DOMINION	International Widgets Inc.
DOMINION	National Offices
DOMINION	Sales, Manufacturing, etc.
DOMAIN	Office Network, Mobile Sales Network
CLUSTER	Building A, Building B, etc.
ESs and ISs	ESs and ISs in building A

Addressing - Common Dominion Level

- Currently Network Service Access Point Addresses (NSAPA) are defined only to be unique
  - Administrative address assigning hierarchy used to make assignment of unique addresses manageable
- Due to address assignment process, address spaces which define Common Dominions will not come from the same bits of the NSAPA
  - As such, an NSAPA and a mask are needed to fully define address space in which ESs and ISs in Common Dominion fall

Common Dominion Level Address Example



Address Assignment Within Common Dominion

- Address further partitioned by Routing Authorities
  - — Now address carries Routing Hierarchy semantics
- Routing Authority has complete autonomy in assigning its own address space
  - — Routing decisions in one Domain/Dominion not concerned with internal structure of another Domain/Dominion
- *Therefore, unified, global address structure not needed*
  - — (Subnetting concept generalized over the whole address space)

NSF gateway requirements Braden (ISI)

GATEWAY REQUIREMENTS RFC

[RFC-985 Update]

RFC-????

Bob Braden
Jon Postel

In Preparation

GATEWAY REQUIREMENTS RFC

- TARGET: Gateway Vendors

- GOALS:

  - TELL VENDORS WHAT WE NEED IN GATEWAYS

  - DESCRIBE CURRENT INTERNET ARCHITECTURE

    - Clarify the intent of the architects
    - Fill in some gaps
    - Scrape off a few barnacles...

  - CONSERVATIVE --

    - Don't invent new architecture !

- COMPREHENSIVE
  - Gather together everything about gateways
- SELECTIVE
  - Host Requirements is ANOTHER RFC!
- INCORPORATE CURRENT EXPERIENCE AND CONCERNS

  Examples:

  - O&M Facilities
  - Martian Filtering
  - Routing Protocols

OUTLINE

1. Introduction

2. Protocols Required for Gateway

3. Constituent Network Interfaces

4. Gateway Algorithms

5. Operation and Maintenance

Appedix A -- Technical Details

Appendix B -- NSFnet Specific Requirements

GENERAL PRINCIPLES

- REQUIRE IMPLEMENTATION OF ALL FEATURES

  –> E.G. Timestamps, Address Mask,
  Info Request/Reply, Record Route

- KEEP STRICT PROTOCOL LAYERING

  –> between local network and IP

- SHARPEN GATEWAY / HOST DISTINCTION

HOST – Independently managed and operated

– Depends on gateways for routing

– Responsible for higher–level protocols

vs.

GATEWAY – Managed and operated as part of a SYSTEM (AS)

– Handles IP datagram routing for hosts

INTERNET GATEWAYS

- A Gateway interfaces to its connected networks as a host.

- A Gateway may be built using any of:

  - Special–purpose hardware.

  - General–purpose CPU dedicated to gateway function.

  - Gateway software embedded in host operating system.
    [e.g., BSD Unix]

BUT . . . embedded gateways may have a conflict between the host role and the gateway role.

Being a gateway in the Internet
is SERIOUS business,
NOT for amateurs.

TERMINOLOGY

- *Gateway* == *IP router*

- *MAC router* (**instead of** *bridge*
  or *level–2 router*)

- *Datagram* (**IP protocol data unit**)
  vs.
  Packet (**Physical network data unit**)

- *Proxy ARP* (**instead of** *ARP hack*
  or *promiscuous ARP*)

SETTLED (?) ISSUES

- ICMP REDIRECTS

  -> Send Host Redirect, not Network Redirect

- SOURCE QUENCH

  -> Must implement something [placeholder]

  -> Configuration parameters to control:

  When to send?

  Maximum frequency to send?

  *** NEED DEFINITE RECOMMENDATION ***

- TTL

  -> Gateway must decrement TTL by

  max(SecondsDelay, 1)

- REDIRECTS TO A GATEWAY

  -> Allowed -- on a technicality (part of IGP)

SETTLED (?) ISSUES (cont'd)

- BROADCAST RULES FOR GATEWAYS

  -> Filter on IP address, not local net address (strict layering)

  -> Don't forward to network 0 or -1

  -> Recommend configurable filters for Martians and other badness

- DIRECTED BROADCASTS

  -> Allowed but limited (indirectly)

- SUBNETS

  -> Allow different subnet masks within same subnetted network

  -> Allow (but recommend against) non-contiguous subnet bits in mask

UNSETTLED ISSUES

- EGP (!!)

  –> Specs are in terrible shape

  –> Can/should we document core's use of the EGP metric?

  –> Does every gateway need EGP?

- Gateways REQUIRED to implement reassembly ?

- Multiple networks/subnets per wire ?

- Implications of general subnetting ?

- Default routes –– good / bad ?

- Hold–downs ?

MAJOR HOLES

- EGP

  -> Need EGP Revision !

- Gateway monitoring [and control] protocol standard

- Serial line protocol standard

- Open IGP --

  -> More generally, recommendations on routing protocols

- DGP

ISSUES FOR A LATER REVISION
[RFC-???? +]

- Fair Queueing

- Type of Service Routing

- How to do ISO-IP and IP together
  (at constituent network level)

- Provision for load-sharing lines
  (at constituent network level)

TRANSPARENT GATEWAYS

(Address–Sharing Gateways)

- Current Example: ACC Product
- Generic Example: SRI Port Expander
- Also related to Jon Postel's Magic Box
- Box between PSN port and Ethernet
  - Hosts are on Ethernet
  - Use Proxy ARP
  - Multiplex on "logical host" field of PSN address

Routing Directions at SRI   Su and Garcia (SRI)

SOME WORK ON SHORTEST-PATH ROUTING AT SRI

ZAW-SING SU
JOSE GARCIA-LUNA

BELLMAN-FORD:

COUNTING-TO-INFINITY

DIJKSTRA:

TIGHT COUPLING

A LOOSELY-COUPLED
MIN-HOP ALGORITHM
W/O COUNTING-TO-INFINITY

* DISTANCE METRIC / TOS

* EGP/AS

* GROWTH

A NEW MINIMUM-HOP ROUTING ALGORITHM

J.J. Garcia-Luna-Aceves

Information Sciences and Technology Center

OUTLINE

- Existing distributed, adaptive algorithms for minimum hop computation and their problems
- The new routing algorithm
- Performance
- Conclusions

BELLMAN-FORD ALGORITHM

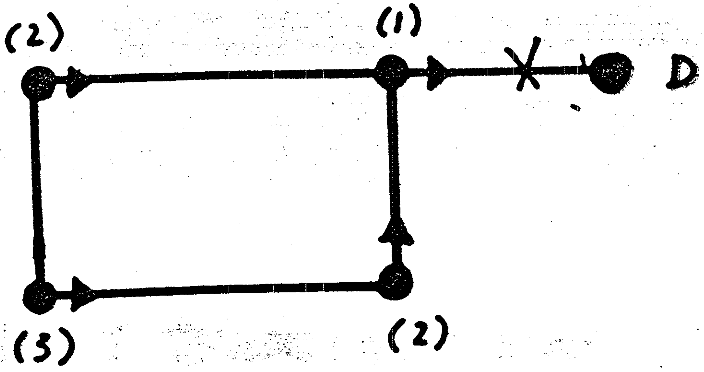
- Routing table entry = node ID, next hop, and shortest distance
- Update messages sent only to neighbors
- Entry in update message = destination ID and shortest distance to it

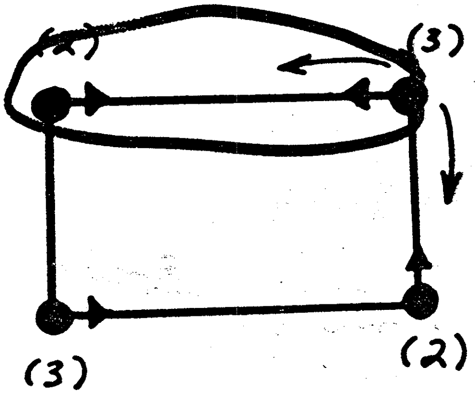
- *Updated shortest distance = shortest distance reported by* ANY NEIGHBOR.
- *Updated next hop* = ANY NEIGHBOR *reporting the shortest distance*

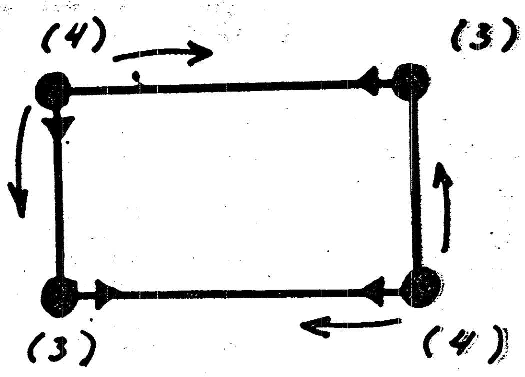
- Example: OLD ARPANET ROUTING ALGORITHM (shortest path)

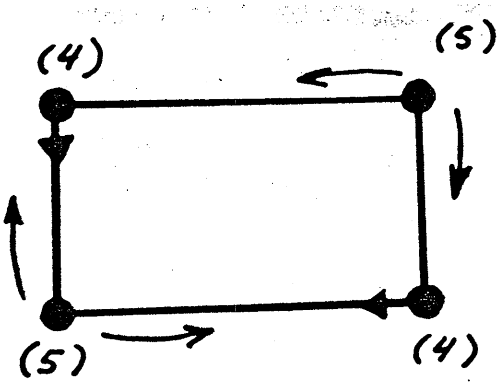
BELLMAN-FORD ALGORITHM

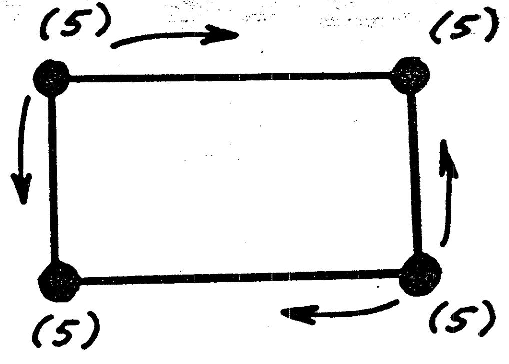
- Update activity....





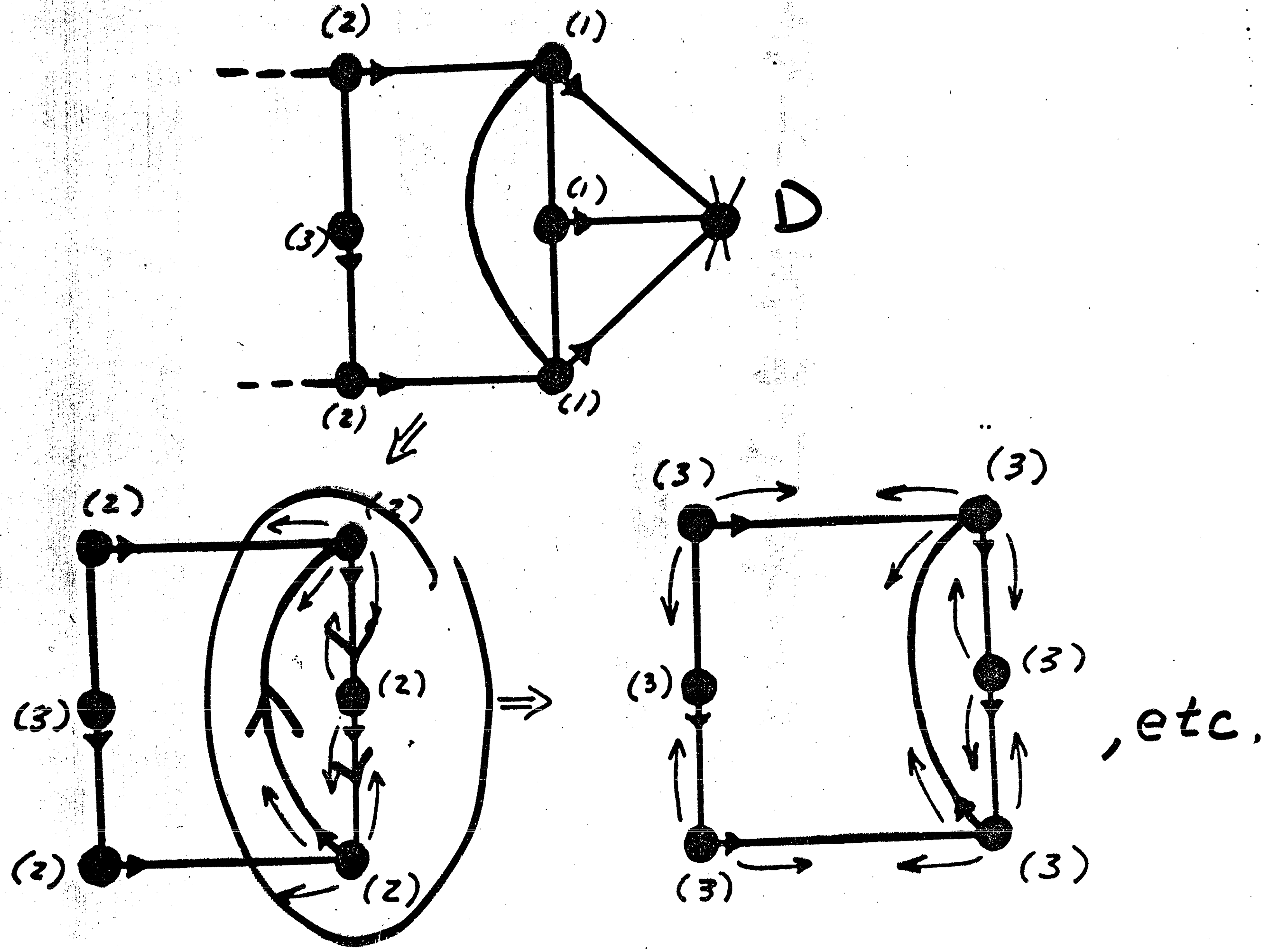






- PROBLEMS: PING-PONG LOOPING AND COUNTING TO INFINITY

COUNTING TO INFINITY
More than Ping-Pong Looping



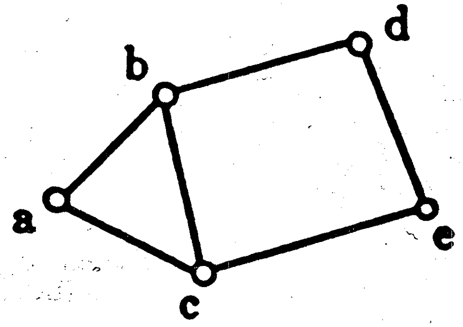
EXISTING ALTERNATIVES

- **New ARPANET routing algorithm:**
  - Each node must know entire topology
  - Each link failure/addition must be broadcast to all nodes
- **Algorithms by Jaffe and Moss,* and Merlin and Segall:**
  - Tight internodal update coordination
  - In *, questionable loop freedom
- **Shin and Chen:**
  - Require complete path from source to destination in updates and routing tables.

NEW ALGORITHM
Information at Each Node

Distances are measured in HOPS

Distance table = routing table info reported by neighbors



(a)

Distance Table						
j	N^a_{jb}	D^a_{jb}	F^a_{jb}	N^a_{jc}	D^a_{jc}	F^a_{jc}
a	a	1	0	a	1	0
b	b	0	0	b	1	0
c	c	1	0	c	0	0
d	d	1	0	b	2	0
e	c	2	0	e	1	0

Routing Table		
j	N^a_j	D^a_j
a	a	0
b	b	1
c	c	1
d	b	2
e	c	2

(b)

NEW ALGORITHM
Assumption

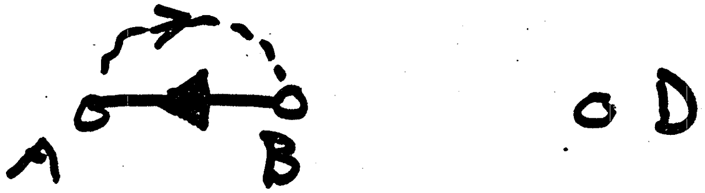
- A link-level service that assures that ALL messages traverse a link reliably and in the proper order
- For unreliable links, feasibility flag must be included in update messages.

NEW ALGORITHM
Table Update Rules

- *UPDATED DISTANCE* at node **A** to **D** = 1 + minimum distance reported by a *FEASIBLE NEIGHBOR* **B** to **D**

- *UPDATED NEXT HOP* at **A** to **D** =
    - Same IF NO CHANGE in distance to **D**
    - Any other neighbor reporting *FEASIBLE NEIGHBOR* to **D** IF CHANGE in distance

- *FEASIBLE NEIGHBOR to* **D** = **B** =>
    - Distance from **B** to **D** is decreasing or constant with respect to previous value reported by **B** (if constant, next hop is also constant).
    - Next hop from **B** to **D** ≠ **A**



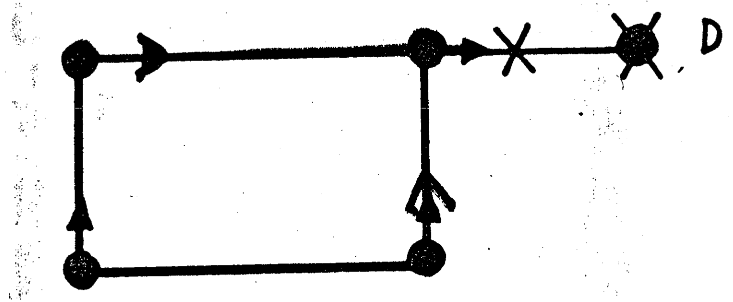
NEW ALGORITHM
Message Exchange Rules

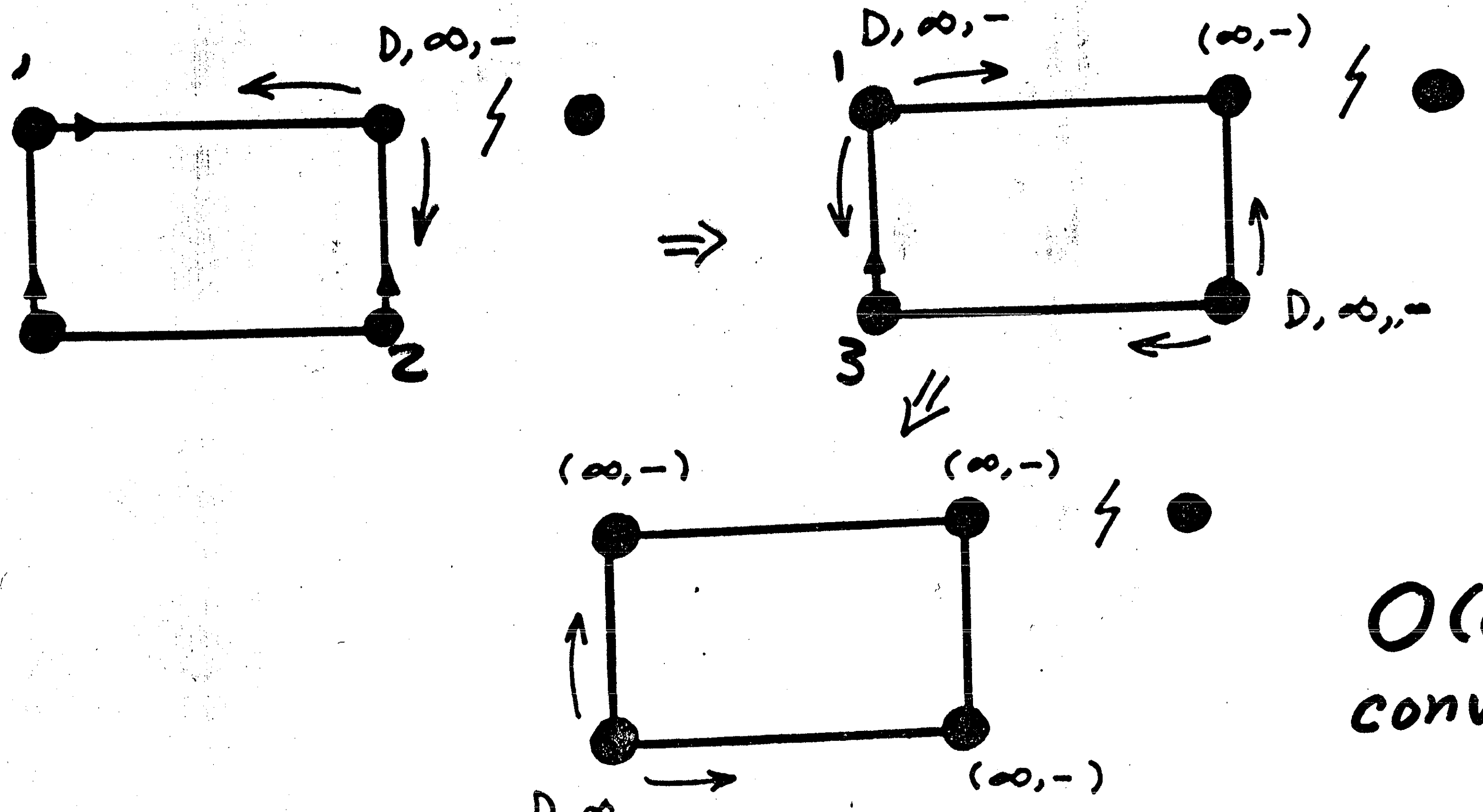
- If routing table (RT) changes, broadcast updates to all neighbors
- If only distance table (DT) changes after processing message from neighbor, send REPLY to that neighbor
- If DT and RT do not change, neighbor reports infinite distance to D, and distance to D in RT is finite, send REPLY to neighbor
- REPLY contains info about same entries in neighbor's original update

NEW ALGORITHM
Update Messages

- Can be event driven or periodic
- Each update is sent only to neighbors
- Each update contains next node and shortest distance to one or more destinations.

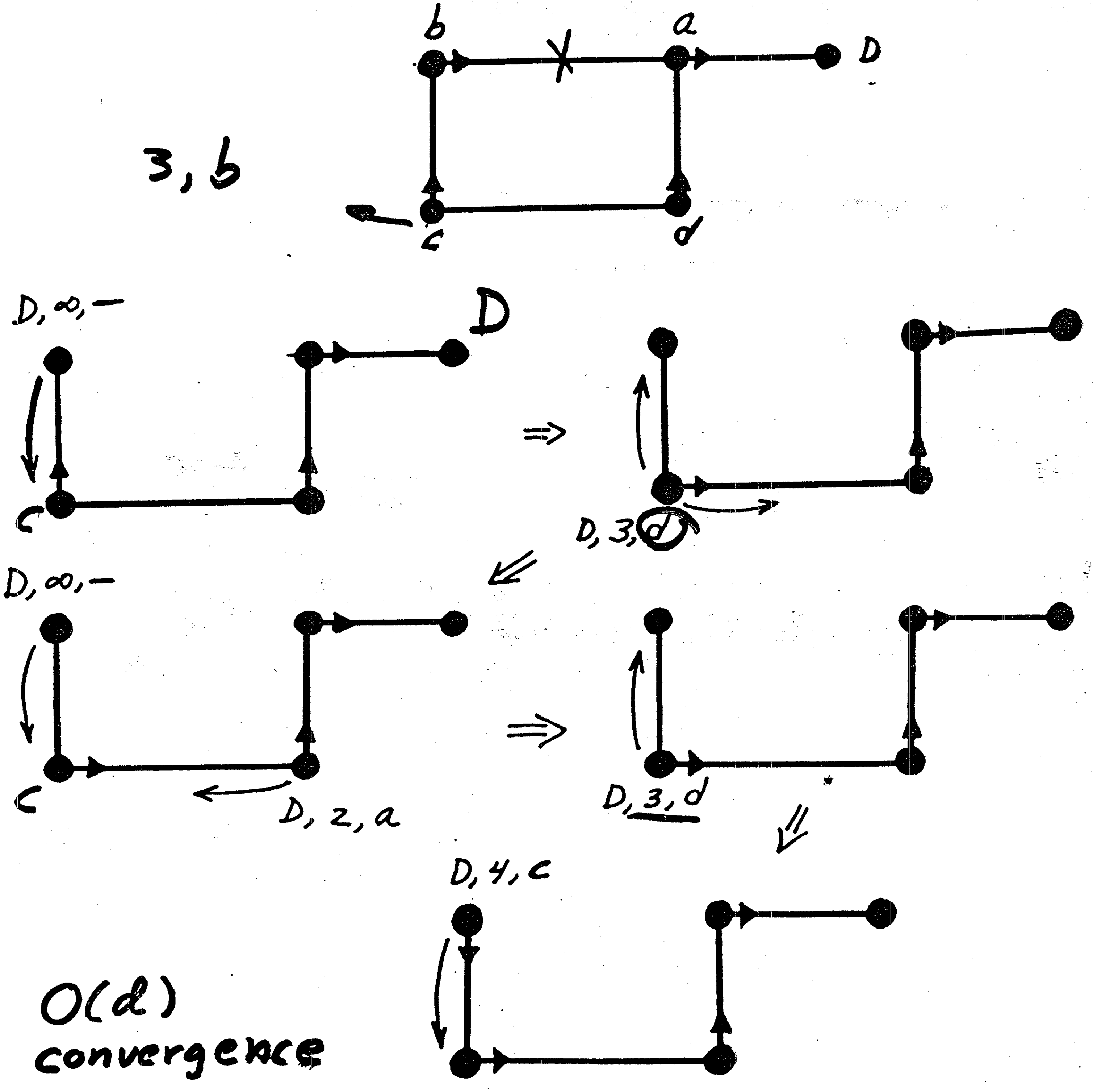
HOW ALGORITHM WORKS
(Node Fails)





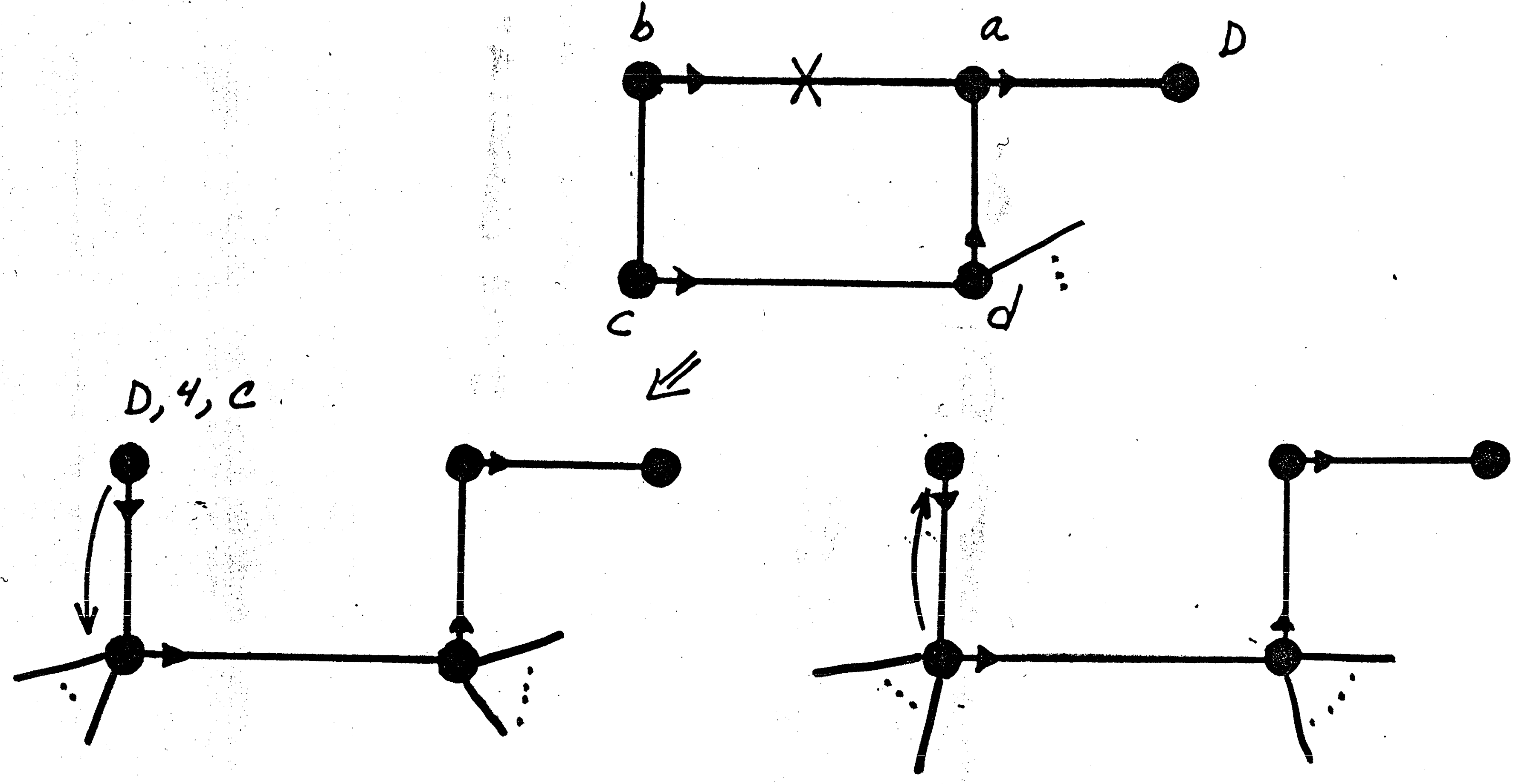
O(d)
convergence

HOW ALGORITHM WORKS
(Link Fails, Bad Case)



O(d) convergence

HOW ALGORITHM WORKS
(Link Fails, Good Case)



Convergence time << O(d)

CONVERGENCE

- Proved to obtain correct routes for all destinations after the network suffers arbitrary topology changes and then remains stable for a sufficiently long time interval

→ - Only assumptions: each node knows its neighbors, link-level service

- Proof is similar to Tajbnapis's and Hagovel's proofs.

PERFORMANCE

- Synchrony assumption--all links have same delay, neighbors of a resource detect its failure at the same time
- Maximum number of update cycles after failure is $\textcircled{d} + 3$
- Maximum number of update messages after failure is smaller than $\textcircled{6|E|}$
- Maximum number of update cycles and messages after recovery are $\underline{d} + 1$ and $\underline{2|E|}$

PERFORMANCE

Algorithm	Worst-Case Number of Steps for Recovery	Worst-Case Number of Messages for Recovery
BF algorithms (e.g., Tajibnapis)	$O(\|N\|)$	$O(\|N\|^2)$
SPF	$O(d)$⊙	$O(2\|E\|)$*
Merlin/Segall	$O(d^2)$ [JAFF-82]	$O(\|N\|^2)$ [SCHW-86]
Jaffe/Moss	$O(x)$ [redacted] [JAFF-82]	
New Algorithm	$O(d)$	$O(6\|E\|)$ → but see examples!

where: |N| = number of network nodes

d = diameter of network

x = number of nodes 'uptree' of failure on shortest-path tree = |N| if node fails!

|E| = number of links

⊙Applies to each topological change, not only worst case

FEATURES OF NEW ALGORITHM

- Inherently distributed
- Simple to implement
- Eliminates ping-pong looping and counting-to-infinity problem
- Proved to be correct
- Outperforms other existing adaptive, min-hop routing algorithms (combined communication, storage, and processing overhead).

NBS Routing Proposal K. Mills (NBS)

NBS Committment to OSI

- Standards Work Since 1979
- Implementors Workshops Since 1983
- GOSIP 1987
- DoD Transition
- Priority on Dynamic Routing Protocols

ROUTING DOMAIN INTERCONNECTION MODEL, SERVICE, & PROTOCOL

OBJECTIVE: PROVIDE ROUTING BETWEEN DOMAINS

CRITERIA:

1) REFLECT ADMINISTRATIVE BOUNDARIES AND ALSO THE REAL COMMUNICATION NEEDS BETWEEN DOMAINS

2) KEEP ROUTING DATABASES SMALL

3) DECOMPOSE THE HARD ROUTING PROBLEMS (e.g., loop avoidance and detection) INTO A TRACTABLE SET OF EASILY CONTAINABLE, LOCAL PROBLEMS

PROPOSED SOLUTION

- A SET OF ONE OR MORE ISs SHALL BE DEPLOYED FOR THE PURPOSE OF ROUTING BETWEEN DOMAINS (A COMMUNITY)

- COMMUNITIES ARE DISJOINT SETS OF THE ABOVE ISs

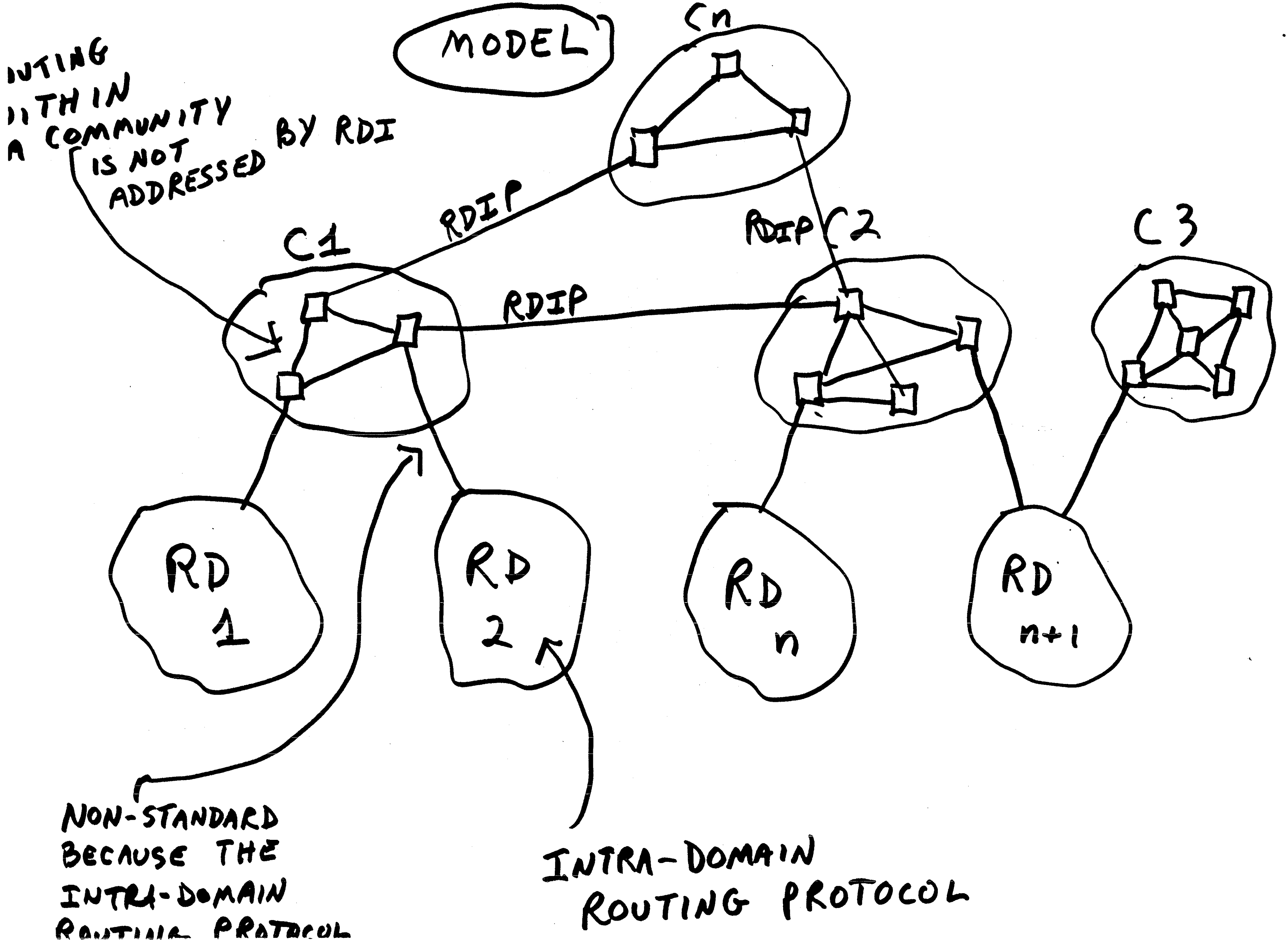
- PRIVATE ROUTING AGREEMENTS MAY BE MADE BETWEEN COMMUNITIES

- BOTH ROUTING INFORMATION FLOWS AND THE FLOW OF IPDUs SHALL BE RESTRICTED BY AN ESTABLISHED SET OF RULES (UP, HORIZONTAL, DOWN)

MODEL CONCEPTS

- COMMUNITY*
- ROUTING DOMAINS*
- ROUTING DOMAIN INTERCONNECTION* PROTOCOL
- INTRA-DOMAIN ROUTING PROTOCOL
- INTRA-COMMUNITY ROUTING PROTOCOL

*ADDRESSED BY RDI PROPOSAL



UTING
ITHIN
A COMMUNITY
IS NOT
ADDRESSED BY RDI

NON-STANDARD
BECAUSE THE
INTRA-DOMAIN
ROUTING PROTOCOL

INTRA-DOMAIN
ROUTING PROTOCOL

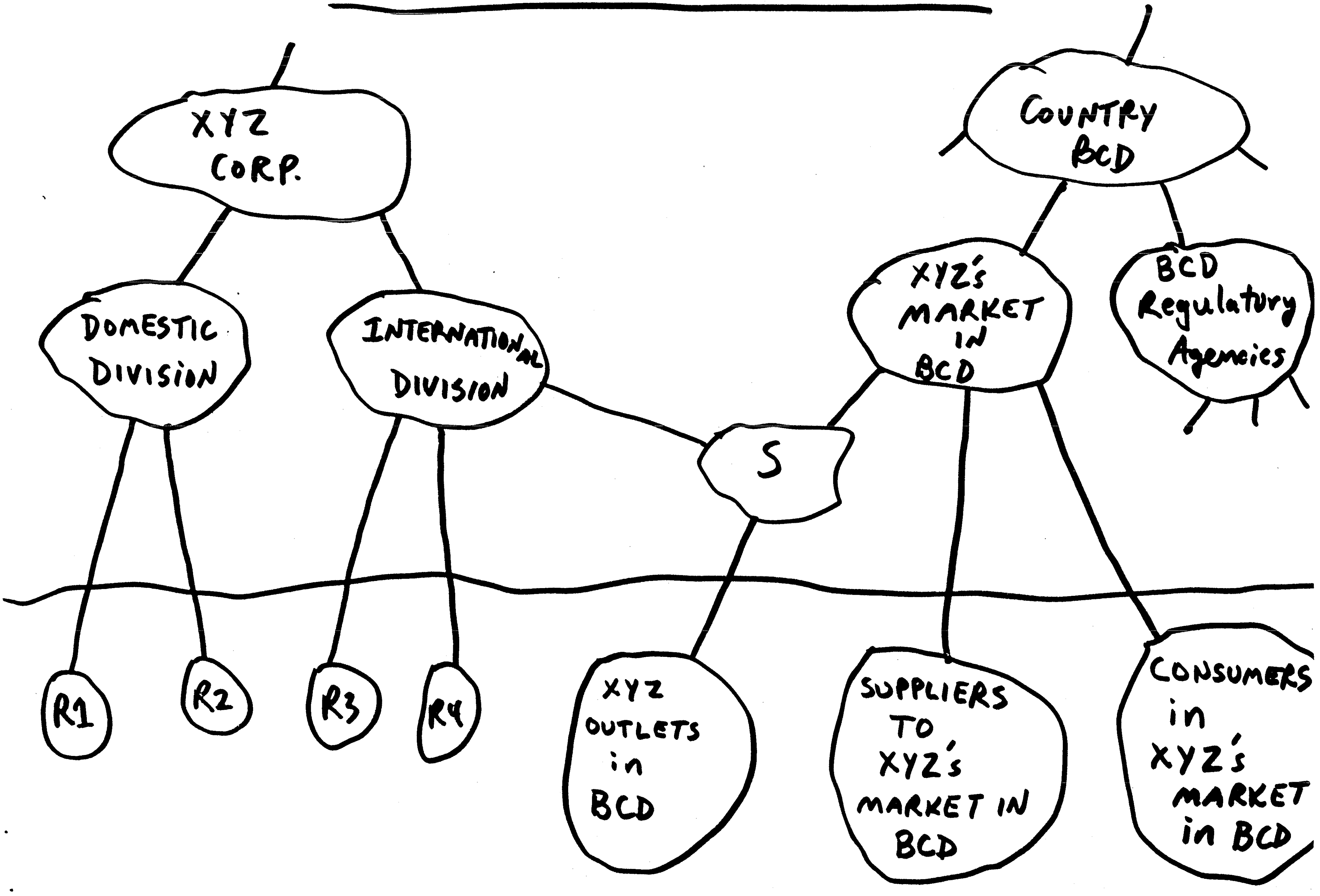
DESIRABLE CHARACTERISTICS OF THE MODEL

- IS COMMUNITIES CAN BE CREATED AND ENTERED INTO APPROPRIATE AGREEMENTS - ALLOWING ROUTING DOMAINS TO PARTICIPATE IN MULTIPLE, DISTINCT COMMUNITIES OF INTEREST

- ENABLED INFORMATION FLOWS ARE SUCH THAT INTER-COMMUNITY LOOPS ARE NOT POSSIBLE

- WHEN RIBs ARE STABLE AND CONSISTENT, IPDUs ARE ROUTED ALONG THE INVERSE OF THE INFORMATION FLOWS

EXAMPLE - MULTI-NATIONAL CORP.



ROUTING DOMAINS

INFORMATION FLOW RULES

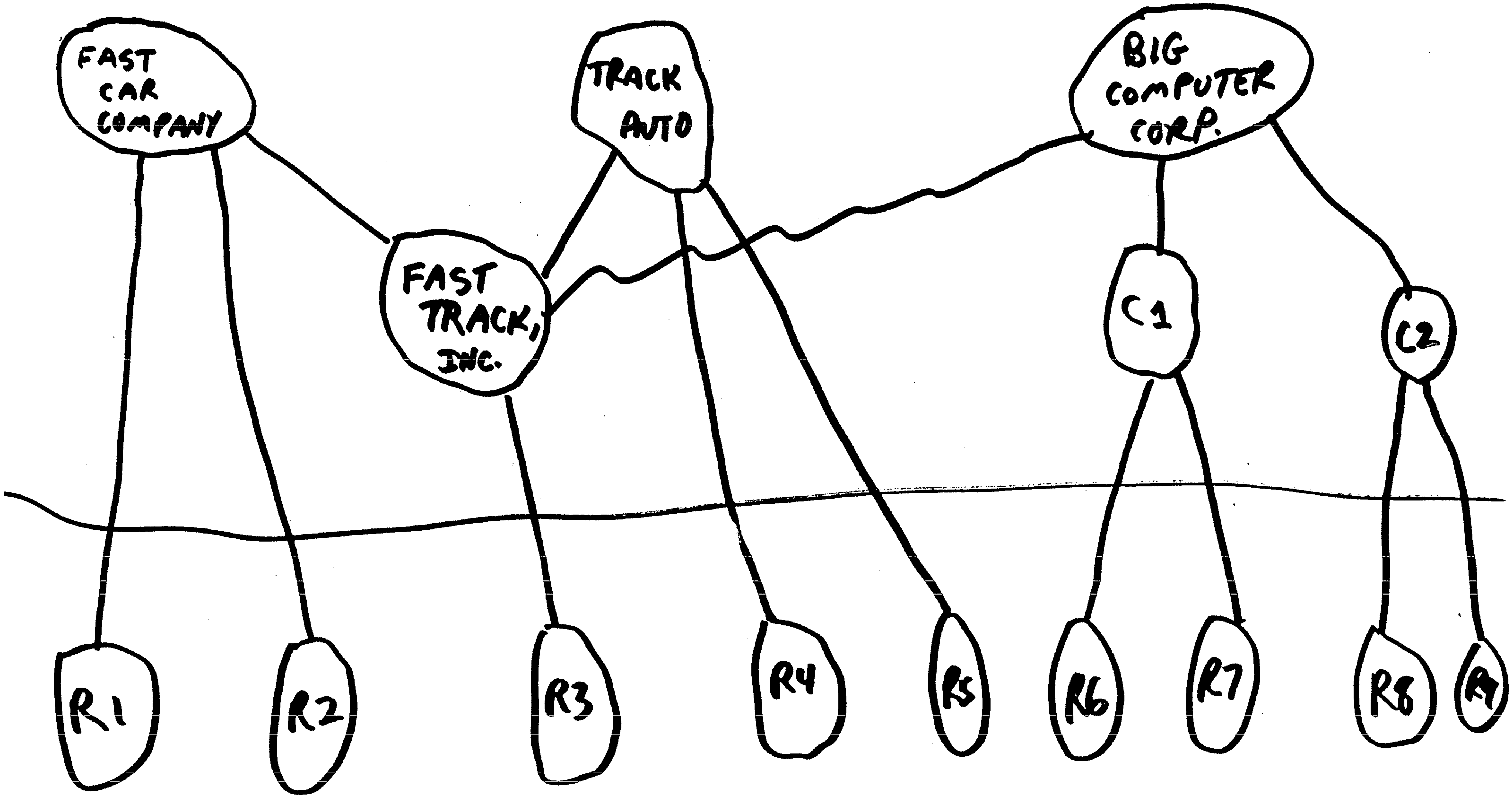
- IF C_A is beneath C_B in the partial order, then C_A is a subcommunity of C_B.
- $C_A \rightarrow C_B$ INFORMATION FLOW IS UP. $C_B \rightarrow C_A$ INFORMATION FLOW IS DOWN.
- BILATERAL AGREEMENTS MAY BE MADE BETWEEN COMMUNITIES. INFORMATION FLOWS BETWEEN THESE COMMUNITIES IS HORIZONTAL.
- COMMUNTIES NEED NOT REVEAL INFORMATIO
- FLOW WITHIN A COMMUNITY IS INTERNAL.

INFORMATION FLOW (con't)

- IF INFORMATION REACHES CA VIA UP OR INTERNAL FLOW, THEN CA MAY DISTRIBUTE THE INFORMATION IN ANY MANNER.

- IF INFORMATION IS RECEIVED VIA DOWN OR HORIZONTAL FLOW, IT MAY ONLY BE SENT OUT ON DOWN OR INTERNAL FLOWS

EXAMPLE - JOINT VENTURE



MERITS OF SCHEME

- COMMUNTIES OF INTEREST CAN BE EASILY AND EFFICIENTLY CONSTRUCTED SUCH THAT NO INDIVIDUAL COMMUNITY WILL BE COMPROMISED

- COMMUNITIES OF INTEREST CAN EASILY EVOLVE WITHOUT A DIRECT EFFECT ON UNINVOLVED ROUTING DOMAINS

- THE ESTABLISHED COMMUNITIES OF INTEREST ARE UNKNOWN TO UNINVOLVED ROUTING DOMAINS

- MORE FLEXIBLE THAN A STRICT HIERARCHY - MORE EFFICIENT THAN A FLAT ARCHITECTURE

MUNDANE PROTOCOL ISSUES ARE ALSO ADDRESSED

- NEIGHBOR ACQUISITION AND MAINTENANCE
  - SOLICIT & ACQUIRE NEIGHBOR
  - NEGOTIATE AGREEMENTS BETWEEN NEIGHBORS
- ROUTING DATA EXCHANGE PHASE
  - ROUTING DATA UPDATES
  - HELLO NEIGHBOR
  - ROUTING DATA QUERY
- NEIGHBOR RELEASE
  - AGREEMENT VIOLATION
  - PERFORMANCE OPTIMIZATION
  - NO WORD
  - FAILURE TO RESPOND DURING NEGOTIATION
  - CANCELLATION OF AGREEMENTS

ISSUES FOR FURTHER STUDY

- MULTIPLE LOGICAL ENTITY INSTANCES OVER THE SAME SNPA

- DETECTION OF INTRA-COMMUNITY LOOPS

- FORM OF THE ROUTING INFORMATI DATA BASE

- ILLEGAL ROUTING DETECTION (due to transient changes)

SUMMARY

- MODEL OF ROUTING DOMAIN INTERCONNECTION USING COMMUNITIES OF ISs SUCH THAT THE COMMUNTIES ARE RELATED VIA A PARTIAL ORDERING
- RESTRICTION ON INFORMATION FLOW BETWEEN COMMUNITIES
- DOES NOT ADDRESS:
  - INTRA-DOMAIN ROUTING
  - ROUTING BETWEEN DOMAINS AND COMMUNITIES
  - INTRA-COMMUNITY ROUTING
- RDI PROPOSAL COULD BE EXTENDED TO ADDRESS INTRA-COMMUNITY ROUTING

Burroughs Integrated Adaptive Routing Piscitello (Unisys)

BNA BIAS ROUTER PRESENTATION

INTRODUCTION

BNA

NETWORK LAYER

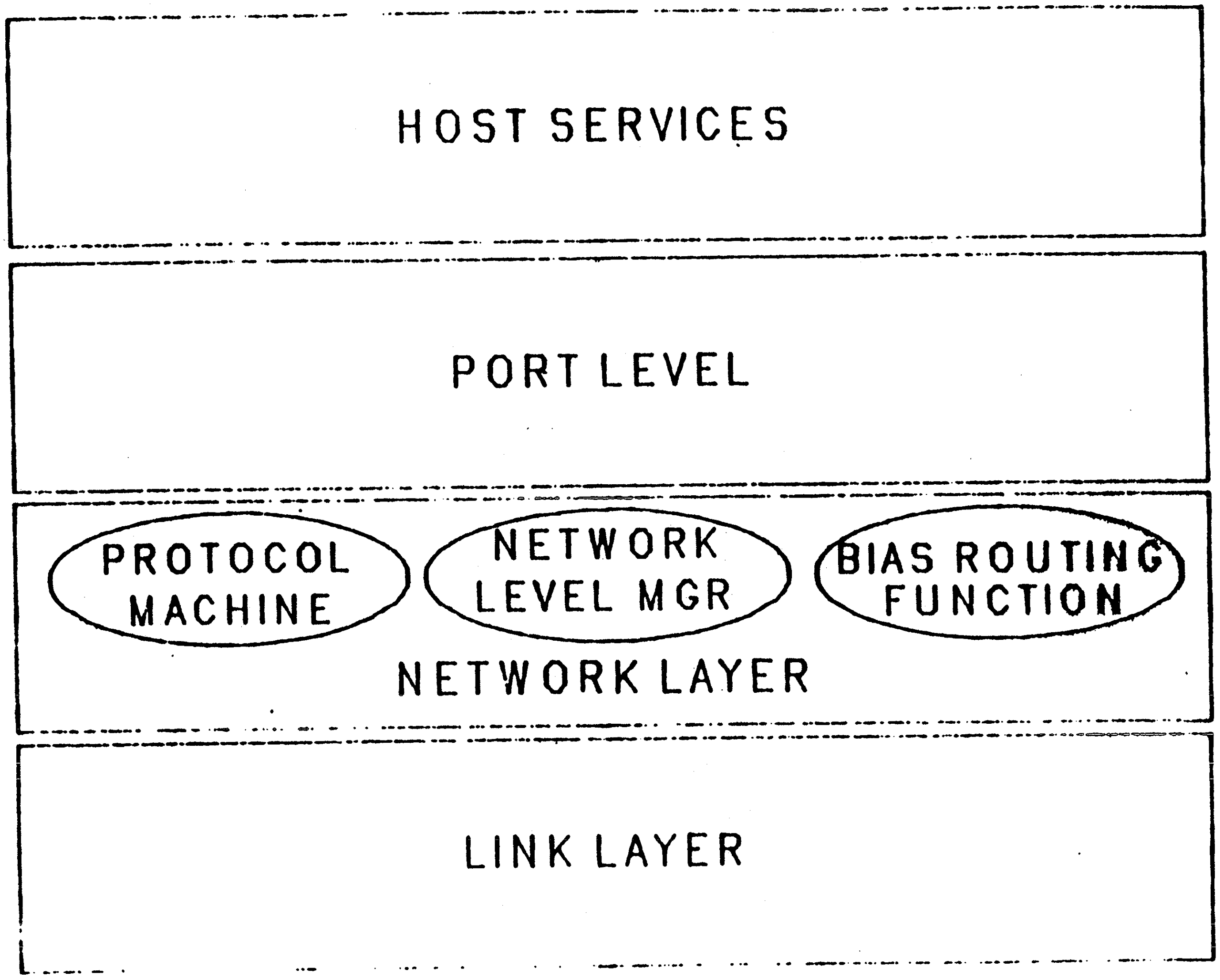
BIAS ROUTER

FUNCTION

CONSTRAINTS AND LIMITATIONS

OVERVIEW

EXAMPLES



BNA OVERVIEW

BNA NETWORK LAYER

CONNECTIONLESS INTERNET SUBLAYER IN ISO TERMS USES ISO 8473 AS DATA PDU

USES BIAS ROUTING PROTOCOL

OPERATES OVER MANY LINK/SUBNET TYPES

BURROUGHS INTEGRATED ADAPTIVE ROUTING SYSTEM (BIAS)

DECENTRALIZED, DETERMINISTIC SYSTEM

BASED ON THE ALGORITHM USED FOR THE MERIT COMPUTER NETWORK

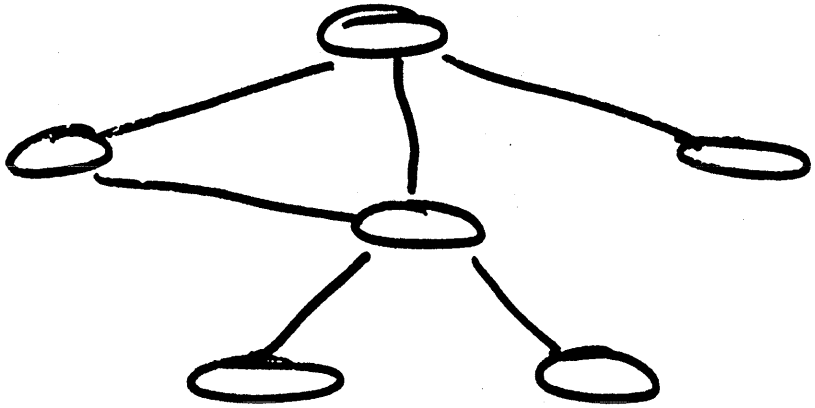
REFINEMENTS TO ORIGINAL ALGORITHM

- O KLEINROCK/KAMOUN
- O JAFFE/MOSS

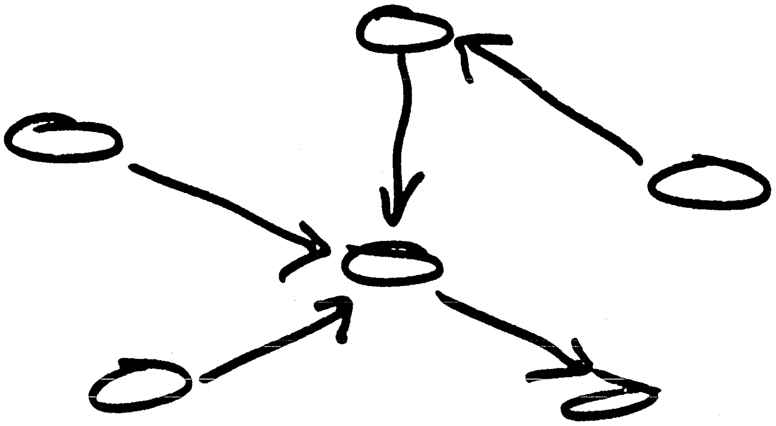
IMPROVEMENTS IN RESPONSIVENESS
USING JAFFE/MOSS

ALGORITHM	WORST CASE # STEPS
1ST GENERATION IUP	O (N)
CUP	O(X)
CUP + FAILURE RECOVERY	O(h)

WHERE N = # NODES IN NETWORK
X = # NODES UPTREE OF FAILURE
h = height of shortest path tree



A: Sample Network Configuration



B: SPANNING TREE OF A

FUNCTION

RESPOND/ADAPT TO CHANGES IN NETWORK TOPOLOGY

- o ADDITION/DELETION OF A NODE
- o ADDITION/DELETION OF A LINK

VERSION PROCESSING

VERIFICATION OF NEIGHBOR IDENTITIES

HIERARCHICAL ADDRESSING

GUARANTEED DELIVERY OF RCFs

LOAD SPLITTING

CONGESTION AVOIDANCE MECHANISM

PARTITION HANDLING

PRIORITY TRAFFIC

CONSTRAINTS AND LIMITATIONS

CONNECTIONLESS SERVICE ONLY

DOES NOT CHANGE ROUTES IN REACTION TO TRAFFIC FLOW

DOES NOT ALLOW UPPER LAYER TO DICTATE [redacted] PATH (NO SOURCE ROUTING)

DOES NOT PROTECT THE NETWORK FROM THE OPERATOR

ASSUMES NODE ADDRESSES ARE UNIQUE

ADDRESSING

- HIERARCHICAL

- FOUR LEVELS OF ADDRESSING "COMPONENTS":

  NODE (CLUSTER LEVEL Ø)
  SIMPLE CLUSTER ↓
  SUPER CLUSTER
  SUBNETWORK (CLUSTER LEVEL 3)

- CURRENTLY USE LOCAL AFI FROM ISO 8348/Z

GREETINGS AND LINKCHANGES

CR:ETINGS:

- o USED TO VERIFY COMPATIBILITY AND ESTABLISH IDENTITIES
- o NEIGHBOR NODE VALIDATION
  - - USER MAY PREDEFINE ACCEPTABLE NEIGHBORS
- o NEIGHBOR NODE AUTHENTICATION
  - - NEIGHBOR MAY PROVIDE PASSWORD TABLE TO PROTECT AGAINST INTRUDERS

LINKCHANGES:

- o INDICATES THE STATUS OF THE LINK BETWEEN TWO NODES.

UPDATE ALGORITHM

MAINTAINS THE BEST ROUTINGS TO EACH DESTINATION

MULTIPLE IN-USE ROUTINGS

    O PRIMARY ROUTING

    O SECONDARY ROUTING(S)

PERFORMED INDEPENDENTLY FOR EACH DESTINATION CLUSTER LEVEL

INDEPENDENT UPDATE PROCEDURE

    O DISTRIBUTION OF GOOD NEWS

COORDINATED UPDATE PROCEDURE

    O DISTRIBUTION OF BAD NEWS

INDEPENDENT UPDATE PROCEDURE (IUP)

LOCAL NODE SENDS NETCHANGE TO NEIGHBOR

O RESISTANCE FACTOR TO ITSELF IS ZERO

O MINIMUM RESISTANCE FACTORS TO ALL THE OTHER DESTINATIONS

NEIGHBOR RETURNS NETCHANGE

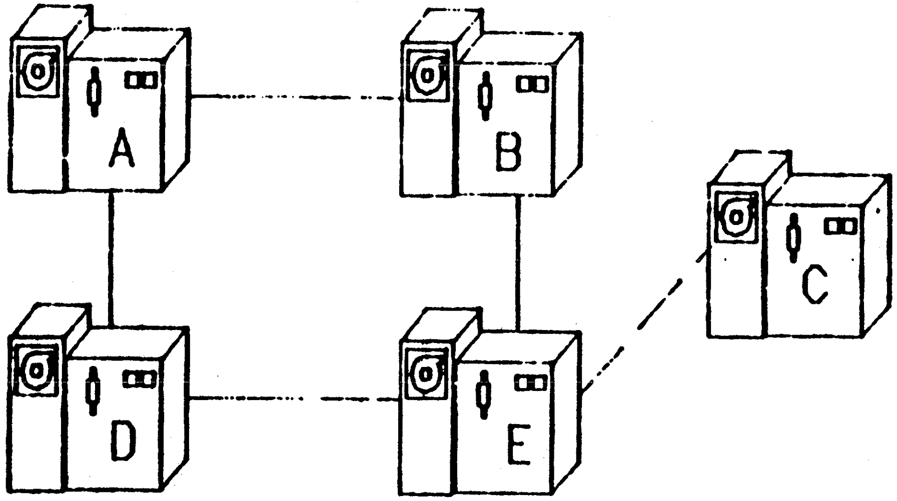
LOCAL NODE ADDS LINKRF OF LINK THAT JOINS IT TO NEIGHBOR TO THOSE DESTINATIONS INDICATED IN THE NETCHANGE MESSAGE

DETERMINES BEST ROUTING TO EACH DESTINATION

O UPDATES ITS IN-USE ROUTING(s)

IF GOOD NEWS RECEIVED RESULTS IN NEW IN-USE ROUTING(s). [redacted]
NEIGHBOR REPEATS PROCESS

An IUP Example



1) Link EC comes up
   (GREETINGs and LINKCHANGE sequences exchanged)
2) E sends to C about E,B,A,D
3) C sends to E about C (AND A,B,D,E)
4) E enters C in routing tables, and
   sends to B and D about C

COORDINATED UPDATE PROCEDURE (CUP)

LOCAL NODE ENTERS THE FREEZE STATE

- O MAY UPDATE ITS ROUTING TABLE WITH RESPECT TO THE DESTINATION
- O MAY NOT CHANGE ITS NEIGHBOR ON ITS IN-USE ROUTING TO THE DESTINATION

LOCAL NODE NOTIFIES ALL NEIGHBORS OF THE BAD NEWS

EACH NEIGHBOR:

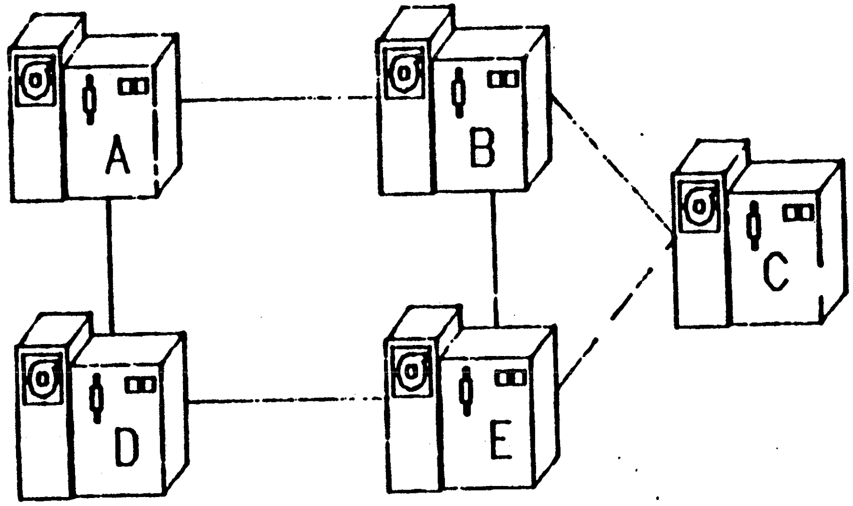
- O CHECKS ITS IN-USE ROUTING TO THE DESTINATION
- O IF AFFECTED BY THE BAD NEWS, ENTERS THE FREEZE STATE AND NOTIFIES ITS NEIGHBORS
- O IF IN-USE ROUTING TO DESTINATION IS NOT AFFECTED, THE LOCAL NODE UPDATES ITS ROUTING TABLE AND RETURNS AN ACK TO NEIGHBOR THAT SENT BAD NEWS

AFTER EACH DOWN TREE NEIGHBOR RECEIVES ACKs FROM ALL ITS UPTREE NEIGHBORS TO WHOM IT SENT BAD NEWS:

- O EXIT THE FREEZE STATE
- O UPDATES ITS IN-USE ROUTINGs BY SELECTING A NEW BEST PATH TO THE DESTINATION

EACH NODE THEN PERFORMS THE IUP (IF APPROPRIATE) AS IT RECOGNIZES THE NEW BEST ROUTING TO THE DESTINATION

A CUP Example



1) Link EC fails
2) E discovers IN – USE ROUTING to C fails and FREEZEs
3) E sends, to B and D, BAD NEWS about C
4) B ACKs to E; its IN – USE ROUTING to C does not change
5) D FREEZEs and sends BAD NEWS to A
6) A ACKs D; its IN – USE ROUTING to C does not change
7) D UNFREEZEs and ACKs E; E UNFREEZEs
8) E picks new IN -- USE ROUTING (thru B)
   to C and sends GOOD NEWS

GUARANTEED DELIVERY

- PROTECT AGAINST LOSS OF NETCHANGE OVER UNRELIABLE LINK
- PRESERVE SEQUENCE
- USES CONNECTIONLESS WITH ACKNOWLEDGEMENT PROCEDURES A LA IEEE 802.2 TYPE 3 LLC

LOAD SPLITTING

LOAD SPLITTING OCCURS WHEN:

O THE FIRST HOP UN EACH PATH IS TO A SEPARATE NEIGHBOR

O THE RESISTANCE FACTORS OF EACH ROUTING ARE EQUAL

O THE MAXIMUM SEGMENT SIZE ON EACH PATH IS EQUAL

SURGE CONTROL

RESPONDS TO AMOUNT OF CONGESTION ON THE NEXT HOP

FRAMES OF LOWER PRIORITY ARE TANKED UNTIL FURTHER NOTICE

PARTITION HANDLING

A WAY OF RESOLVING "BROKEN" [redacted] CLUSTERS

NODES WITH NEIGHBORS IN OTHER CLUSTERS REGISTER WITH DIRECTORY SERVICES AND ACT AS VIRTUAL NEIGHBORS OF A GIVEN CLUSTER WITH OTHER NODES THAT ARE ALSO REGISTERED

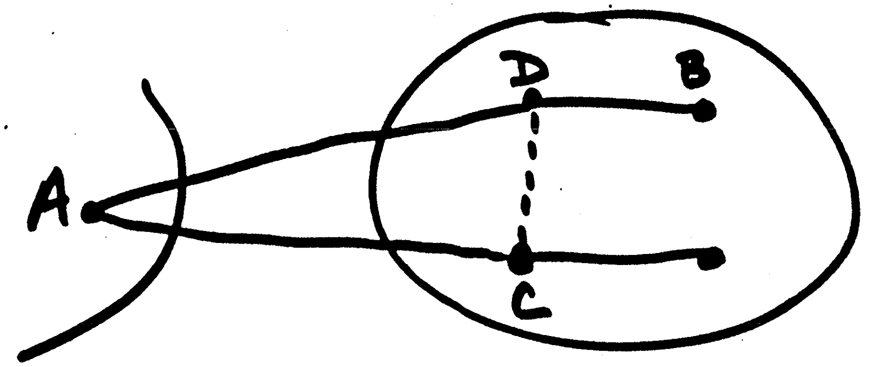
TYPES OF PARTITIONS HANDLED:

o BRIDGES

o COCOONS

PARTITION HANDLING

BRIDGING:



D, C KNOWS:

1) A IS NEIGHBOR

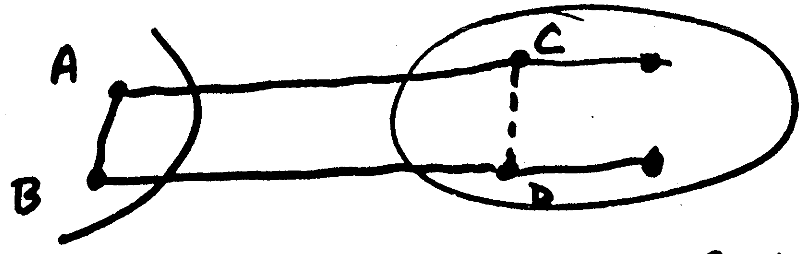
2) A IN DIFFERENT CLUSTER; SO

D, C SENDS CLUSTER DETAILS TO A

- WHEN LINK (D,C) FAILS, C SENDS BAD NEWS (RESISTANCE = ∞) TO A
- A FOLLOWS CUP PROCEDURE (ACKS)
- A SENDS GOOD NEWS

PARTITION HANDLING

"COCOONING"



- C KNOWS A IS A NEIGHBOR AND IN DIFFERENT CLUSTE
- D KNOWS SAME ABOUT B
- A, B ENROLL IN DIRECTORY AS VIRTUAL NEIGHBORS OF C, D'S CLUSTER
- A, B KNOW EACH OTHER AS VNs OF C, D'S CLUSTE

WHEN LINK (C, D) FAILS:

C SENDS BAD NEWS TO A (D DOES SAME TO B)

A'S BEST PATH TO D AFFECTED ; A FREEZES

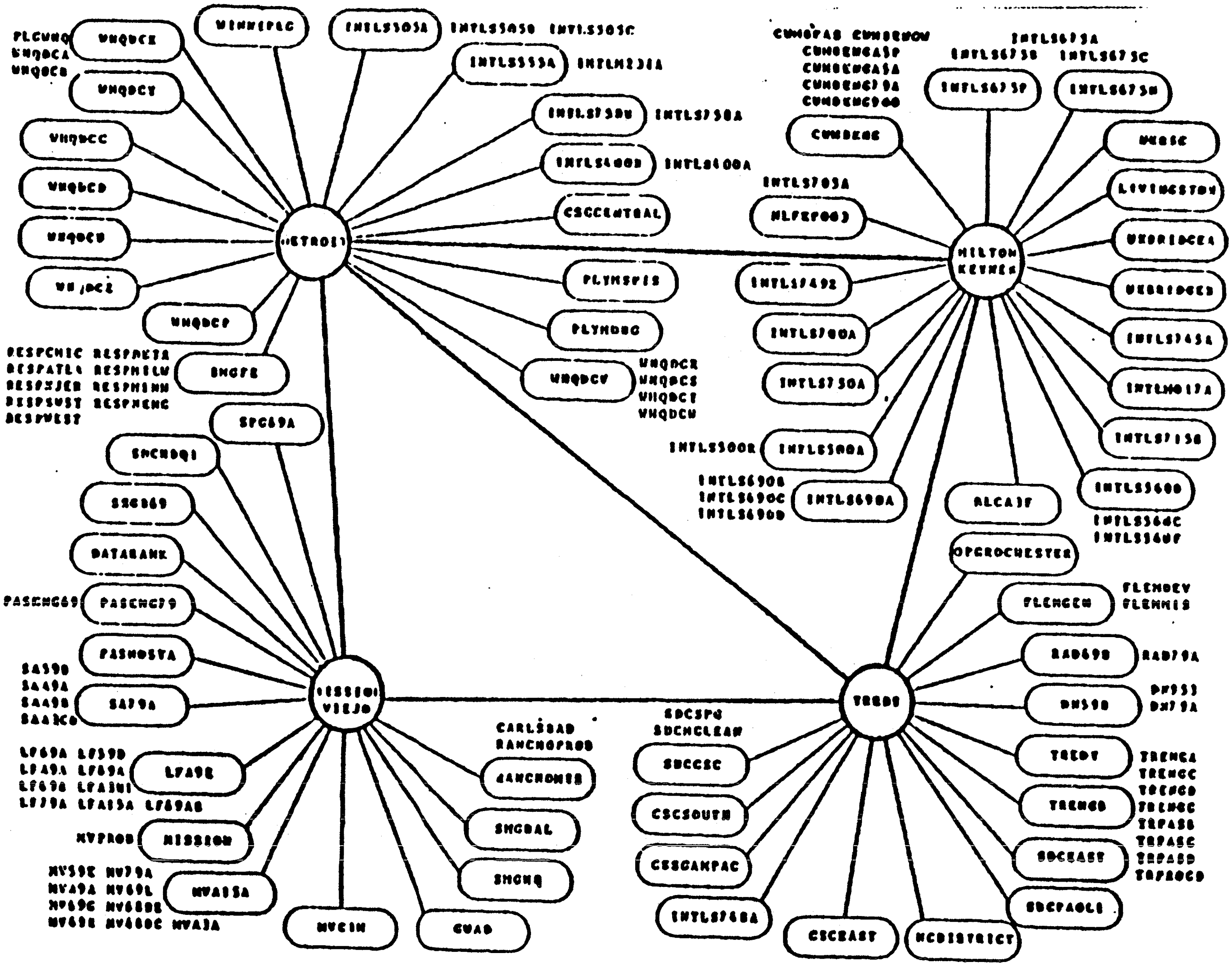
A SENDS BAD NEWS TO B ; B ACKS

A UNFREEZES AND ACKS C

B SENDS GOOD NEWS TO A ABOUT D'S PART OF CLUSTER

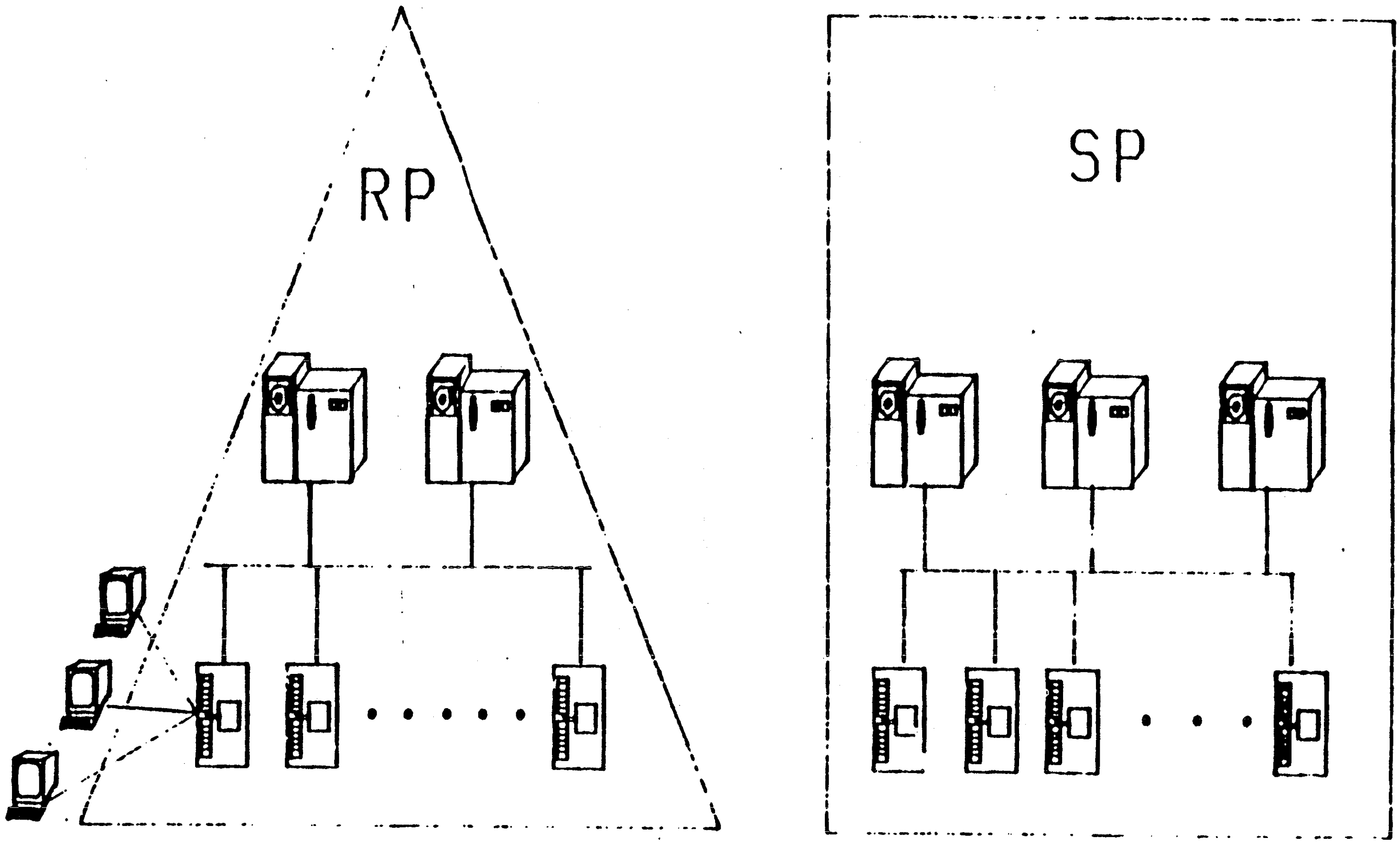
A SENDS GOOD NEWS TO B ABOUT C'S PART OF CLUSTER

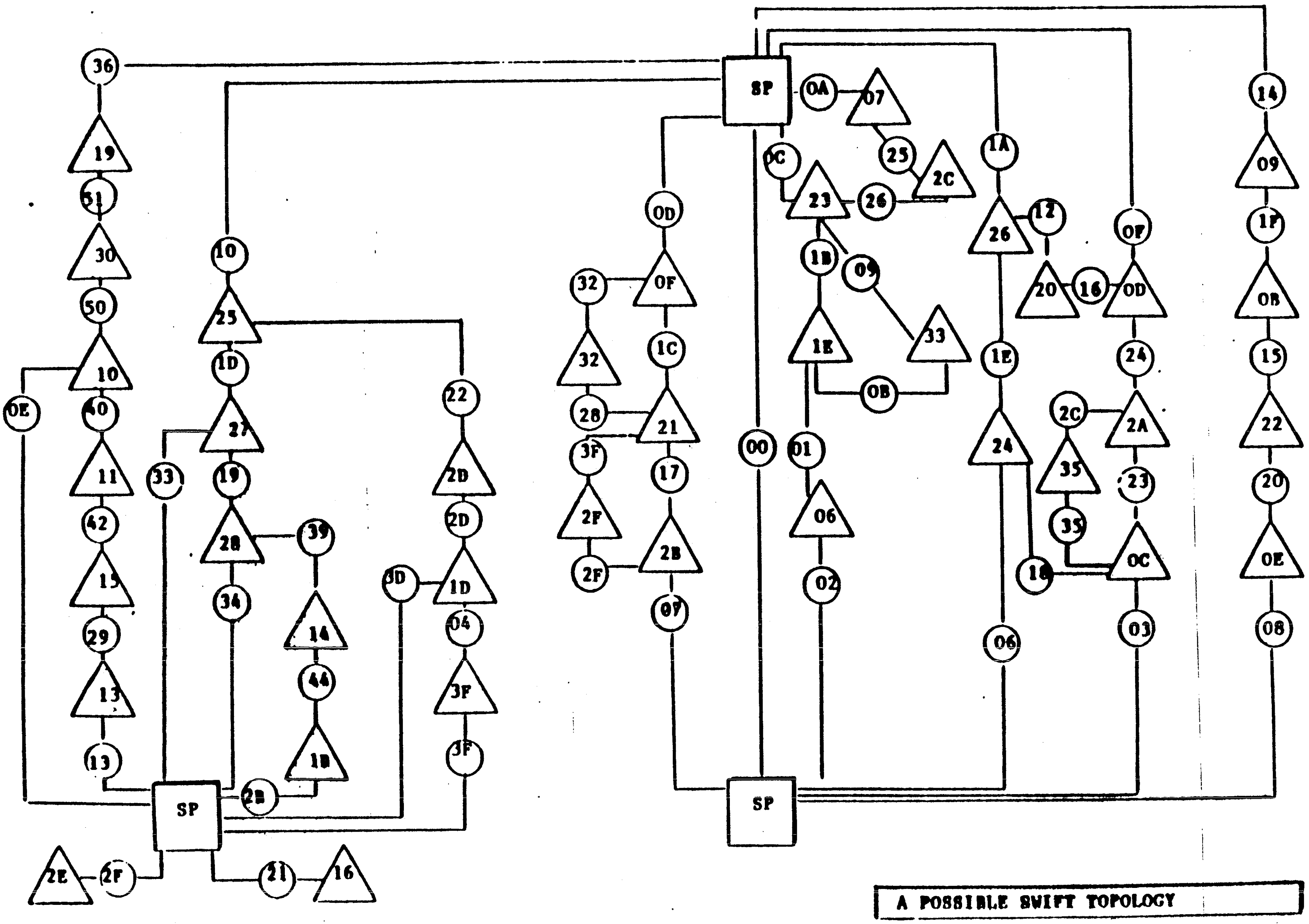
A SENDS GOOD NEWS TO C ABOUT D'S PART OF CLUSTER



RNA HOSTS FEBRUARY 1987

CORPORATE TECHNICAL SYSTEMS & SERVICES


RP
SP



A POSSIBLE SWIFT TOPOLOGY

DNA Routing Overview

David Oran
Digital Equipment Corporation
April 23, 1987

digital

Basic Characteristics

- A scheme for Intra-Dominion Routing
- Has some facilities for Inter-Dominion Routing
- Can handle Dominions of up to 100 million Network Entities
- Uses a Link State Routing Algorithm which adapts to topology changes or management-initiated metric changes. It is not a fully dynamic algorithm which reacts in real-time to changes in the traffic matrix.
- Uses ISO standard addressing as specified in ISO 8348/AD2. All address formats are supported.
- Designed to work together with ISO8473 (Connectionless Network Protocol) and DP9542 (End System to Intermediate System Routing Protocol)
- Extensive network management built in, based on CMIS/CMIP

Topology Features

- Handles arbitrary topologies of LAN's, PDN's, private circuit or packet-switched networks, POTS, and leased circuits
- Two-level Clustering Hierarchy
  - An *Area* can consist of up to 100,000 End systems and Intermediate systems, although $\approx$10000 is a recommended maximum for robustness reasons
  - A Domain can consist of up to 10000 areas, although $\approx$1000 is a recommended maximum
- Allows multiple Domains to be interconnected
  - Inter-Domain Routing handled via static tables at boundary IS's
  - Other domains handled by a special addressing data structure called an *Address Prefix*
  - Routing Algorithm propagates information about exit points to other domains through normal routing method

Basic Routing Features

The Basic routing scheme uses a ~~GGP~~ SPF-like link state routing algorithm. Routing within an area is called *Level 1* routing. Routing among areas is called *Level 2* Routing. At Level 1, each IS in an area has a total map of the area. At Level 2, each IS has a total map of the "level 2 net".

- Uses controlled flooding to build the network map, like ~~GGP~~ SPF but with slightly different duplicate suppression techniques
- Uses a variant of Dijkstra SPF to compute shortest paths
- All known bugs in ~~GGP~~ SPF fixed; especially computational complexity and sequence space problems
- Uses a single, arbitrary Routing metric for each link, assigned by the network manager, or measured locally. The scheme could be extended without too much difficulty to support multiple routing metrics.
- High-connectivity links (802.3 LANs) do not result in N^2 paths for the routing algorithm.

Basic Routing Features (cont.)

- Does not need an initialization hold-down timer
- Computational complexity: $O(E)$ where E is the number of links in the area/domain. Note that basic Dijkstra SPF runs in $O(N^2)$, and previous optimizations have lowered this to $O(E \log N)$.
- Highly robust against hardware and software failures, including memory corruptions. Can tolerate nearly any non-Byzantine failure.

Fancy Routing Features

- Can repair partitions of an area dynamically. Repair of level 2 partitions by "tunneling" through a level 1 area is possible but not implemented (it's *very* hairy)
- Can split traffic over any number of equal-cost paths. Round Robin queueing tends to preserve packet ordering thus reducing CPU overhead in Transport. Option exists to suppress path splitting for traffic in which sequence preservation is more important than throughput.
- Detects and reports congestion via the "Congestion Experienced bit" in CLNP. Congestion collapse prevented via a square-root limiter hueristic
- Optimizes routing for End Systems with multiple SNPAs on the same subnet (we call these *multi-link end systems*).

Fancy Routing Features (cont.)

- Complete autoconfiguration using ES-IS and IS-IS initialization exchanges once ISs are assigned to areas.
- Reduces Routing overhead on LANs by using pseudo-node technique, and the election of a *Designated IS* for the LAN/area.

Addressing

The basic DNA Routing address format is as follows:

IDP	Pad	Area	ID	Nsel

where:

IDP is one of the allowable *Initial Domain Parts* from ISO8348/AD2.

Pad is 0–6 octets, used to pad the address to its maximum length. This makes hashing addresses for forwarding efficiency much easier.

Area is a two-octet integer, assigned by the dominion manager to the area in which this IS logically resides

ID is a 6 octet system identifier. DEC uses Ethernet absolute host ids in this field to ease address administration. The only requirement, however, is that the ID be unique within an area for level 1 ISs and ESs, and unique within the domain for level 2 ISs.

NSel is a 1-octet NSAP Selector, used for discriminating CLNP users within a network entity. The network entity itself is identified by using the reserved Nsel value of zero.

Features for Handling Connection-oriented Networks

All of the functions called out in ISO8473/AD1 are available. In addition, the use of ISO 8208 (X.25) networks is coupled to the routing algorithms to improve performance:

- VCs may be brought up when traffic arrives and torn down on a timer without cranking the routing algorithm
- Three forms of routing over connection-oriented facilities:

  Static Routing uses manually-entered addresses. The path to the destination is always declared "up" by the routing algorithm and is handled just like a point-to-point datalink

  Dynamic Connection Management VCs are set up to pass routing and data traffic. IS-IS routing PDUs are sent over the circuit at low frequency to ensure against bad routing or black holes.

Dynamic Assignment The circuit is brought up upon receipt of traffic and the DTE address to call is determined dynamically. There are different dialed/undialed costs for the circuit thus avoiding multiple calls from different ISs to the same destination. The SNPA to call is determined based either on the destination address, (if the SNPA is derivable from the NSAP address), or via static tables (similar to the inter-dominion routing tables) configured in the IS.

PDUs and their uses

IS-IS LAN Initialization Hello Used to initialize all ISs on a LAN, detect transitivity of link, elect the designated IS for a LAN, determine which ISs are level 1 and which are level 2, and to label a LAN pseudo-node

IS-IS PT-PT Initialization Hello Used to initialize the two ISs on a pt-pt link (leased link, X.25 VC, etc.) Similar in function to LAN Initialization but much simpler. Could be carried in proposed IS-IS Initialization field in the ES-IS Protocol.

Link State Reports the status of a link. There are four flavors of these:

- **Level 1 IS** Reports all Level 1 IS neighbors on a link
- **Level 1 ES** Reports all neighbor ESs on a link
- **Level 2 IS** Reports neighbor level 2 ISs and contains information concerning area partition repair
- **Level 2 ES** Propagates static information on other domains/systems

Sequence Numbers Used to resynchronize link state databases periodically to recover from memory corruptions or faulty ISs

Current Status

- Breadboards running in house
- Extensive simulations and performance analysis of central algorithms has been done.
- No products have been shipped yet with this routing scheme
- DEC is willing to make the algorithms, data structures, and protocols public for standardization without fees or formal licensing.
- We can have a base document prepared for discussion at the next ANSI X3S3.3 meeting in July

SPF Routing in the Butterfly Gateways Mallory (BBN)

SPF ROUTING IN THE BUTTERFLY GATEWAY

Tracy Mallory

BBN Communications Corporation

OVERVIEW

- Background
  - General features of routing algorithms
  - Problems with LSI-11 GGP
  - SPF in the Arpanet
- SPF in the Butterfly Gateway
  - Metric - neighbor connections
  - Updating procedure - flooding
  - Path calculation - SPF
  - Forwarding - entry/exit model
- Next

GENERAL FEATURES OF ROUTING ALGORITHMS

- Metric
- Updating procedure
- Path calculation
- Forwarding

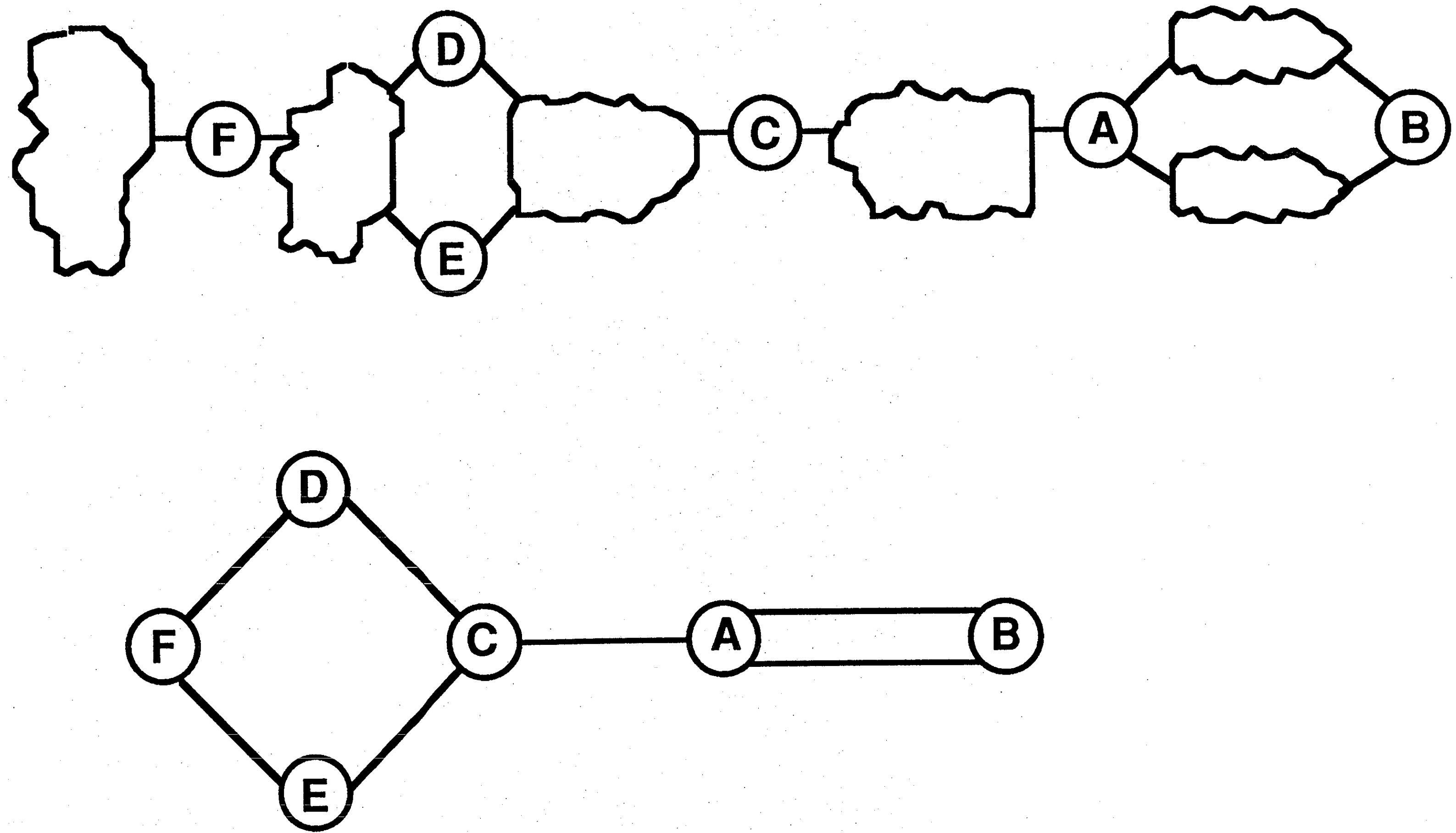
PROBLEMS WITH CURRENT GGP ROUTING

- "Counting to infinity"
- Not enough information to allow special features
- Updates potentially very large

SPF IN THE ARPANET

- Single path
- Delay based
- Global knowledge of topology and delays
- Hold-down to ensure complete database

ADAPTING SPF TO THE INTERNET



METRIC - NEIGHBOR CONNECTIONS

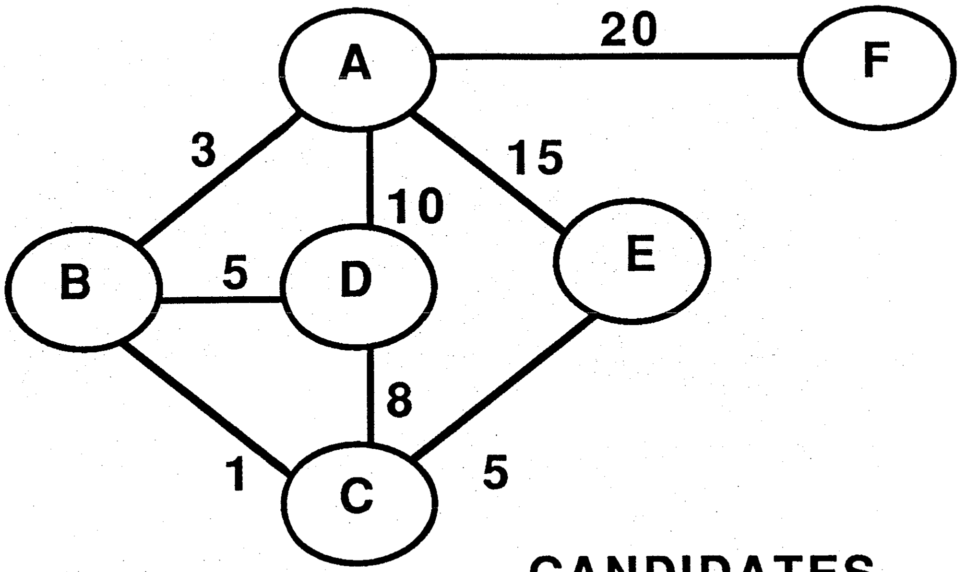
- Initially "fixed cost" per link
  - Per network
  - Per neighbor
- Neighbor up/down
  - Hello/IHU
  - Master/slave
  - Sequence numbers
  - Database exchange

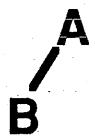
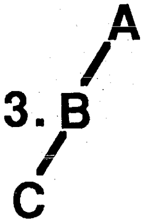
UPDATING PROCEDURE

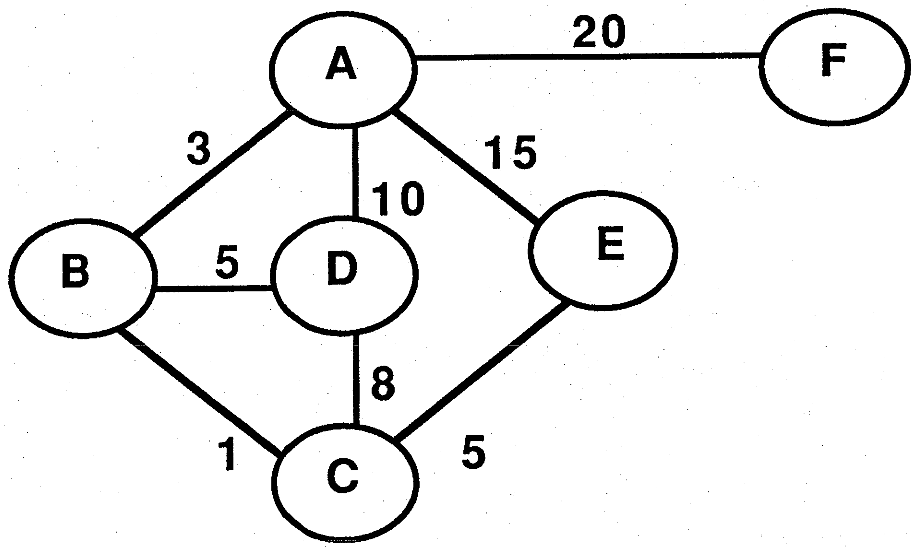
- Updates contain network interfaces, neighbors, and cost to each neighbor
- Ability to package multiple updates in one message
- Update triggered by topology change or once/8 minutes
- Sequence numbers
- Flooding
- Receipt of update is acknowledgement
- Explicit acks
- Retransmissions
- Aging of information

PATH CALCULATION

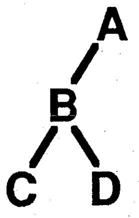
- Dijkstra's algorithm
  (shortest path first)
- Shortest path tree
- Incremental updating of tree
- Minimizes a single cost



TREE	CANDIDATES
1. A	B(A,3) D(A,10) E(A,15) F(A, 20)
2. 	C(B, 4) D(B, 8) E(A, 15) F(A, 20)
3. 	D(B, 8) E(C, 9) F(A, 20)

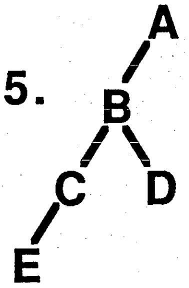


4.



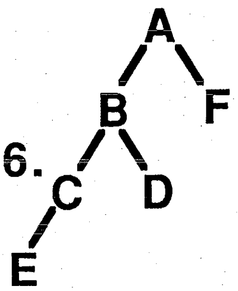
E(C, 9)   F(A, 20)

5.



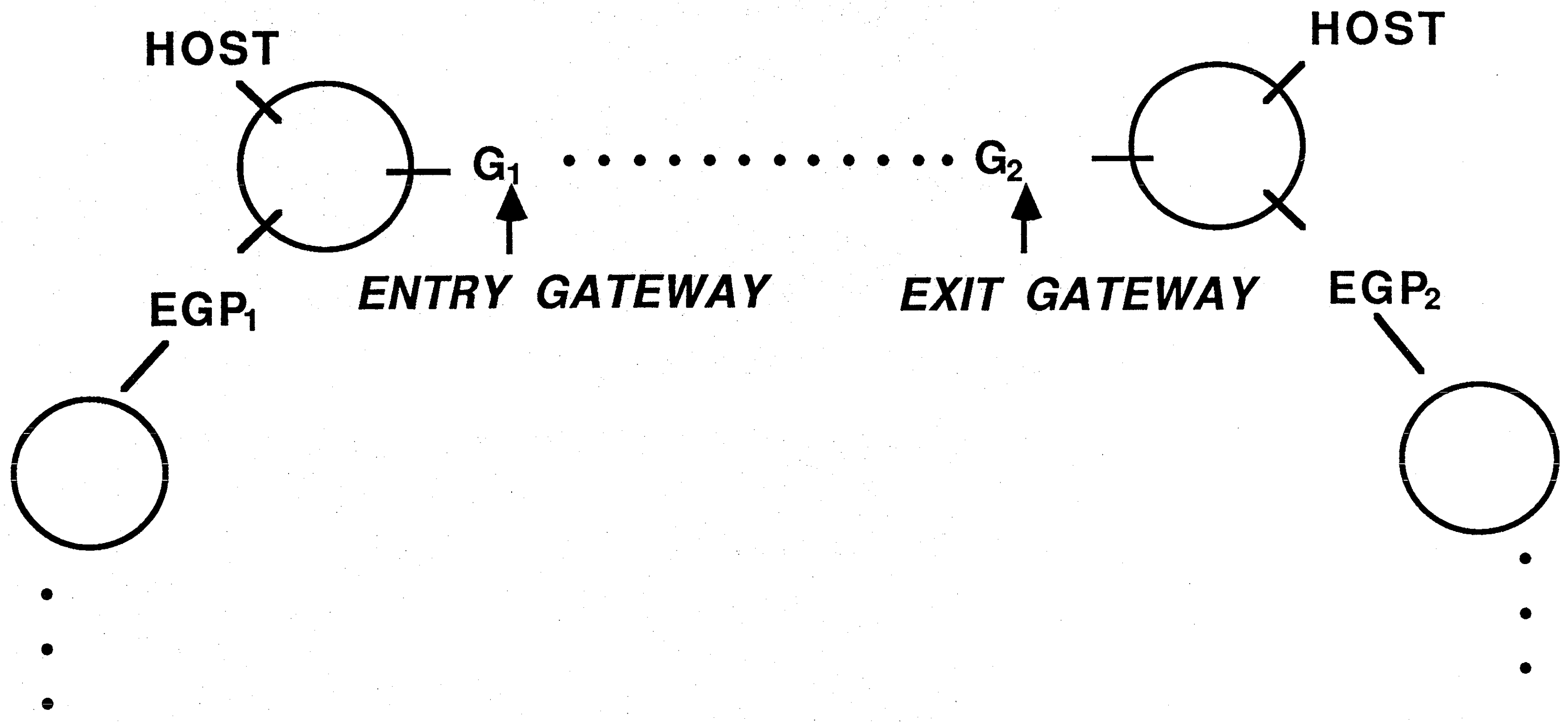
F(A, 20)

6.



FORWARDING - ENTRY/EXIT

- If entry = exit, choose interface to destination network
- Otherwise, choose interface to next hop gateway, specify next hop address, and add GG-header with exit gateway number



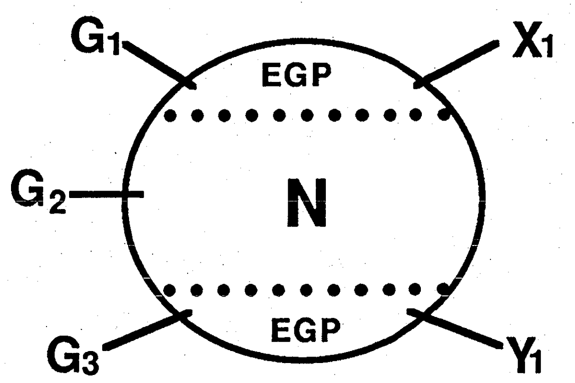
EXTERIOR ROUTING

THE MODEL

- Definition: we say that a gateway borders an exterior Net E with respect to Net N if it has an exterior neighbor, direct or indirect, on Net N who reports a path to E.

- All gateways on a net border the same set of exterior networks

- All gateways not on a given Net can think of the exterior networks that are bordered by gateways on that Net as being directly connected to those gateway



Reports
Nets A & B

Reports
Nets C & D

G_1 , G_1 , and G_3 all Border A, B, C & D

UPDATING PROCEDURE - ENAP

- Exterior Network Advertising Protocol
- Indirect Neighbor Update (INU)
- Exterior Network Update (ENU)

INUs

- Guarantee that all gateways on a net know about all exterior neighbors on a net and all nets reachable through them (all gateways on a net border the same exterior nets with respect to that net)
- Sent only by gateways with direct exterior neighbors
- Sent only to other gateways on the Net the INU is about
- Lists exterior neighbor(s) and their nets
- Sequenced

ENUs

- Distribute exterior information to all gateways
- One gateway on each Net that has any exterior gateways sends ENUs (arbitrarily, lowest numbered)
- ENU lists all exterior networks bordered by the sender
- Sending rate limited
- Flooded

PATH CALCULATION

- Choose lowest cost route for each bordered net
- If we don't border a net, then choose lowest cost exit gateway which borders the net
- Could choose lowest exterior cost first
- Could combine interior and exterior metrics

Congestion Avoidance in Computer Networks with a Connectionless Network Layer

Raj Jain, K. K. Ramakrishnan, Dah-Ming Chiu
Digital Equipment Corp.
550 King St. (LKG1-2/A19)
Littleton, MA 01460-1289

ARPAnet: Jain%Erlang.dec@DECWRL.DEC.COM
Rama%Erlang.dec@DECWRL.DEC.COM
Chiu%Erlang.dec@DECWRL.DEC.COM

April 24, 1987

Myths

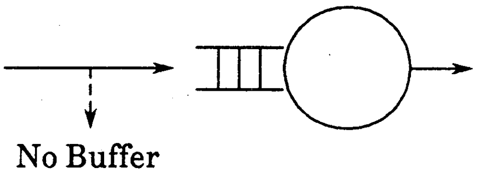
Congestion control will be solved when:

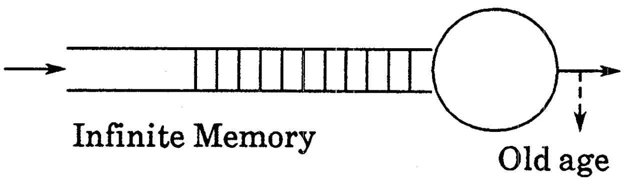
1. Memory becomes cheap (Infinite memory)

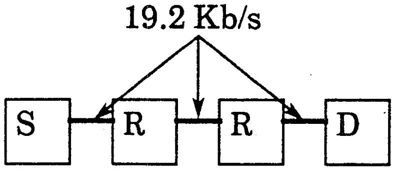
2. Links become cheap (Very high speed links)

3. Processors become cheap (High speed processors)

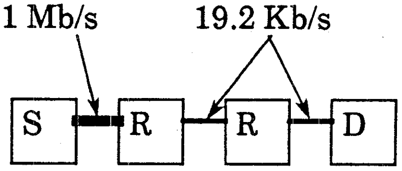
4. All of the above



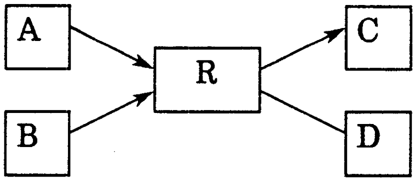




Time to transfer a file = 5 minutes



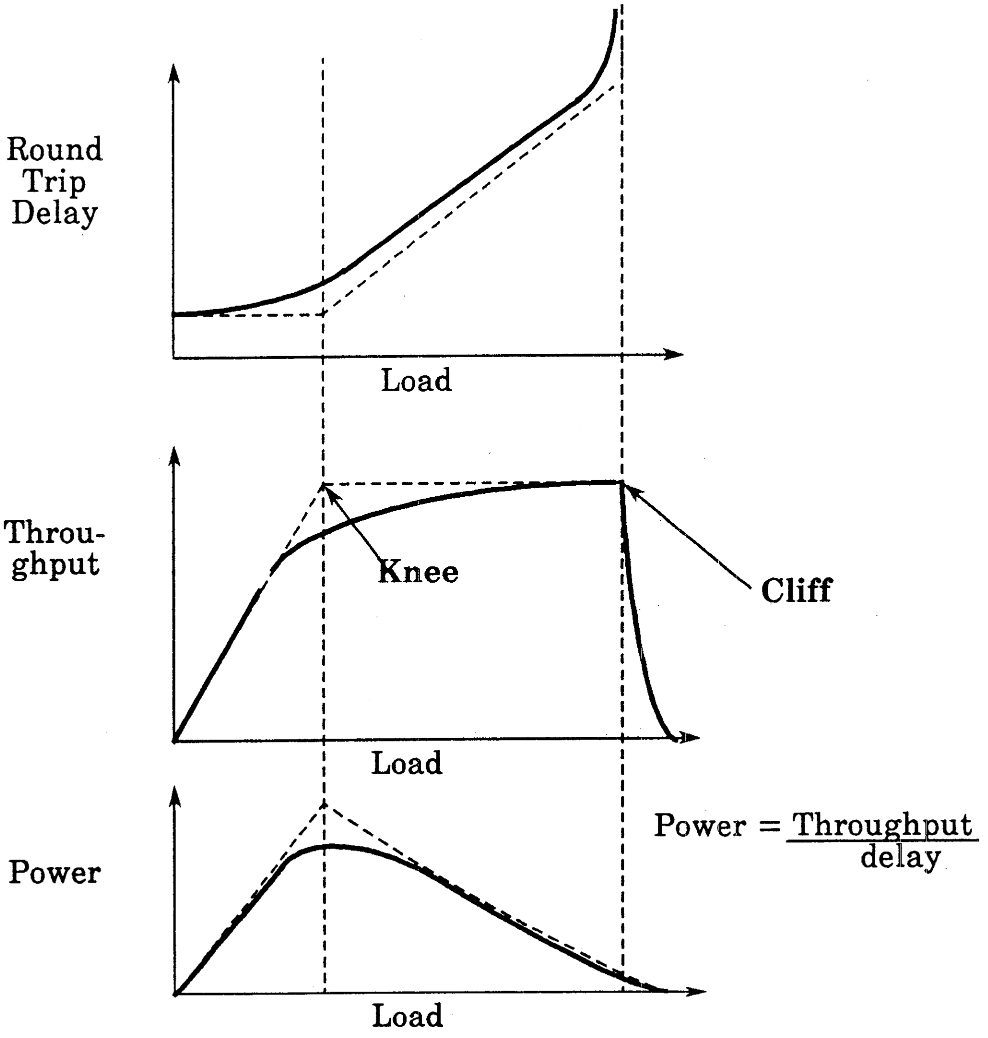
Time to transfer the file = 7 Hours



Balanced Configuration: A links 1 Mb/sec

Conclusions: 1. Congestion is a dynamic problem. Static solutions are not sufficient.
2. Bandwidth explosion
⇒ More unbalanced Networks.

Congestion Avoidance



Congestion Control Mechanisms:

Recover from *zero* throughput and *infinite* delay zone

Left of Cliff (Depends on # of buffers)

Congestion Avoidance Mechanisms:

Keep in *high* throughput and *low* delay zone

At knee (Independent of bufs)

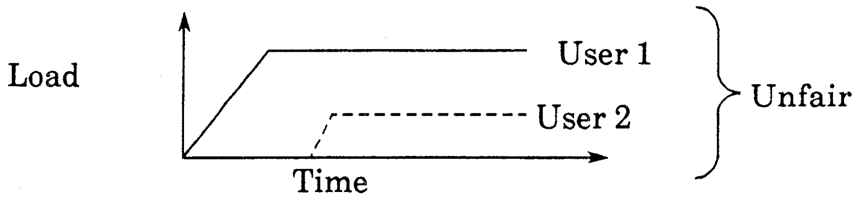
Ref: Jain, IEEE JSAC, Oct 1986
Ramakrishnan, Networking Symp. Nov 1986

4/24/87

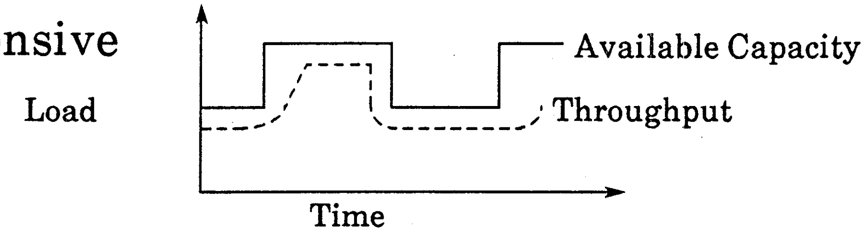
Goals

1. Efficient: Network Power/Network Power at knee

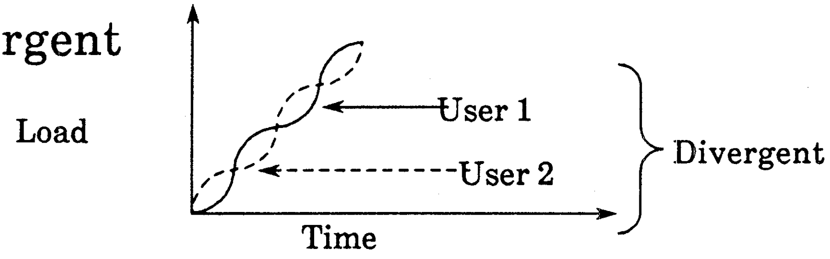
2. Fair: Users sharing the same path get the same throughput.



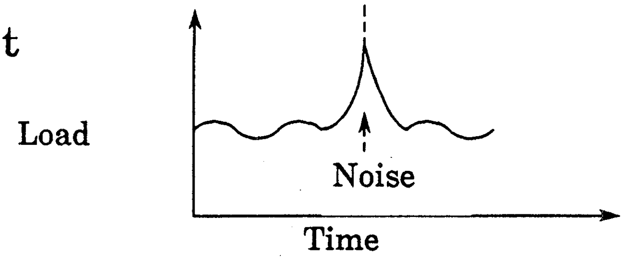
3. Responsive



4. Convergent



5. Robust



6. Distributed

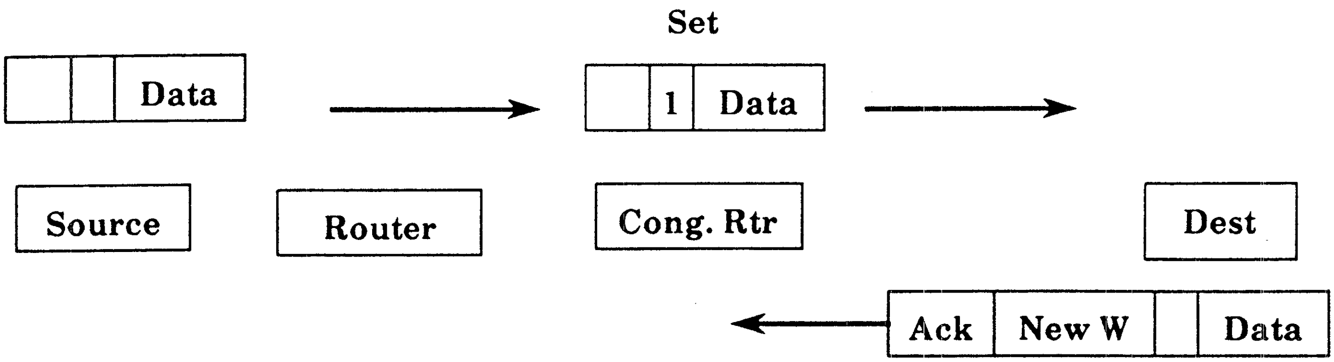
7. Maximum Information Entropy

8. Simple

4/24/87

The Binary Feedback Scheme

Routers provide explicit feedback information when congested.



Set

Data

1 Data

Source

Router

Cong. Rtr

Dest

Ack New W Data

Transport layer characteristic: Window flow control.

Transport entity adjusts window in response to congestion

Network Layer: Connectionless;
No additonal traffic when congested;
challenge - only 1 bit available to indicate congestion.

Issues:

1) When to start setting/stop setting the bits?

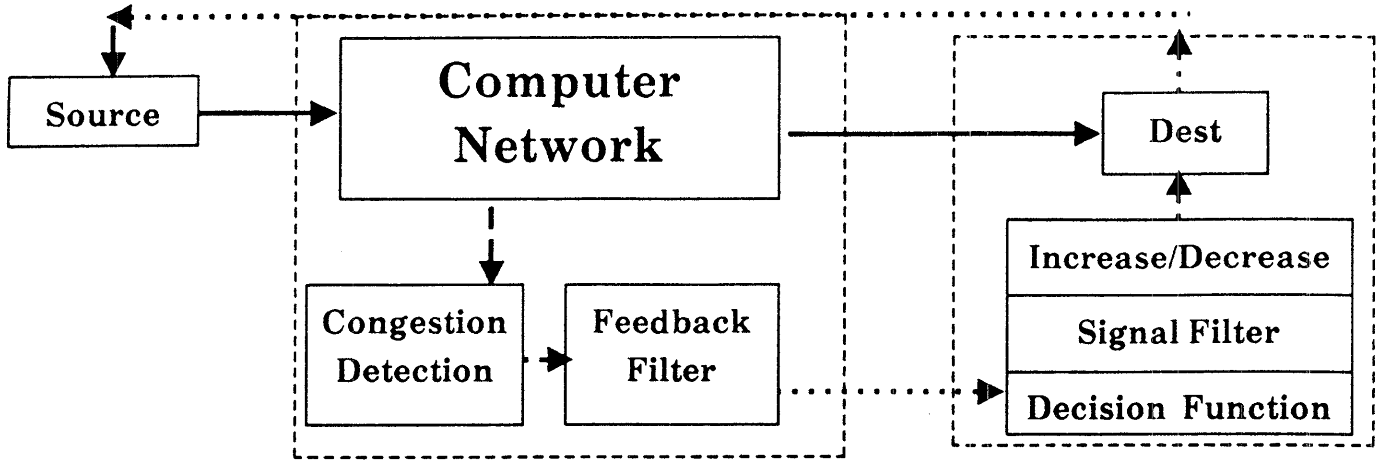
2) What do the users do with the bits?

how many of these bits should we look at?

How to dynamically adjust the window size?

Modeling Approach

Network and users: modeled as a feedback control system.



Source

Computer Network

Dest

Congestion Detection

Feedback Filter

Increase/Decrease

Signal Filter

Decision Function

- **Network Policy:**

  (1) Each router averages number of packets queued.

  (2) Feedback signal - Set the congestion avoidance bit when average queue length ≥ threshold C.

$$C = 1$$

- **User Policy:**

  (1) Frequency of Update :
  $\mathbf{W}_{\mathbf{prev}} + \mathbf{W}_{\mathbf{cur}}$ packets have been acked.

  (2) Filtering by user: Examine $\mathbf{W}_{\mathbf{cur}}$ bits.

  cut-off policy: If > = 50% of congestion avoidance bits are set cause window to decrease. Otherwise increase.

  (3) Increase/Decrease: Fairness considerations - increase : +1, decrease : aW, $(0<a<1)$.

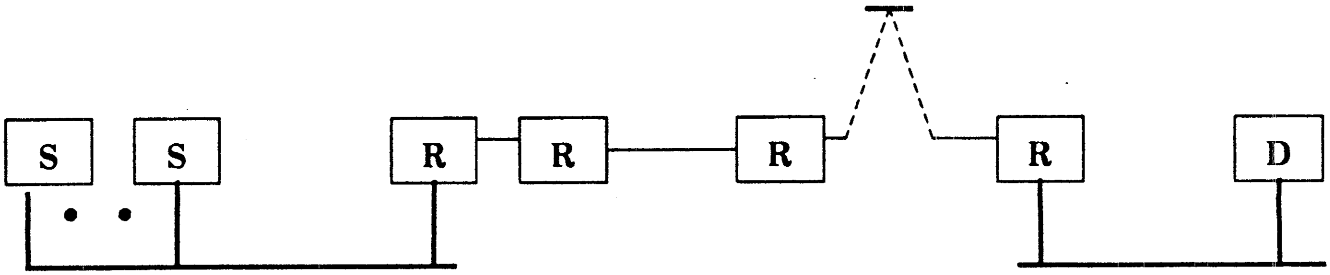
Methodology

Analytical methods to study of aspects of policies in isolation

Detailed simulation to study policies as a whole in network:

Multiple users of the network.sharing the same resources.
users have abundant packets to transmit.
Transport characteristics of window flow control, time-outs and retransmissions are modeled.

Routers - single server queues.
Service times may be deterministic or random.
Individual router service times may be different.
Model satellite links



Limitations:

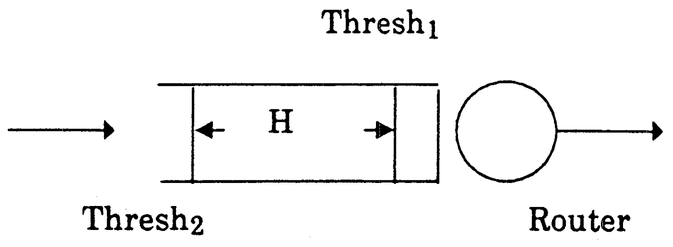
Don't simulate overhead for window updates

No path splitting.

no traffic in the reverse direction.

Congestion Detection

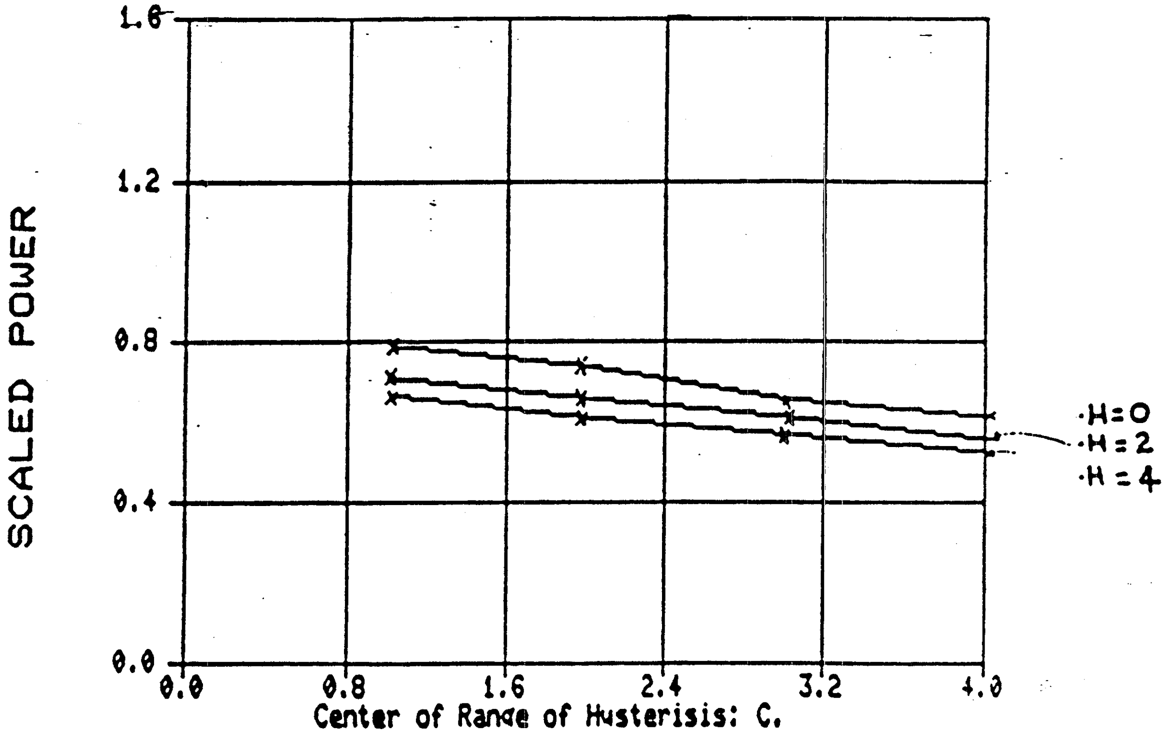
Feedback signal generated by setting "congestion avoidance" bit in routing layer header.



Policy Alternatives:

1) Simple Threshold: Queue size ≥ Thresh$_1$, set bit.

2) Hysterisis Policy: Queue size ↑ and ≥Thresh$_2$, set bit. Continue to set bit till queue size ↓ and ≤Thresh$_1$.



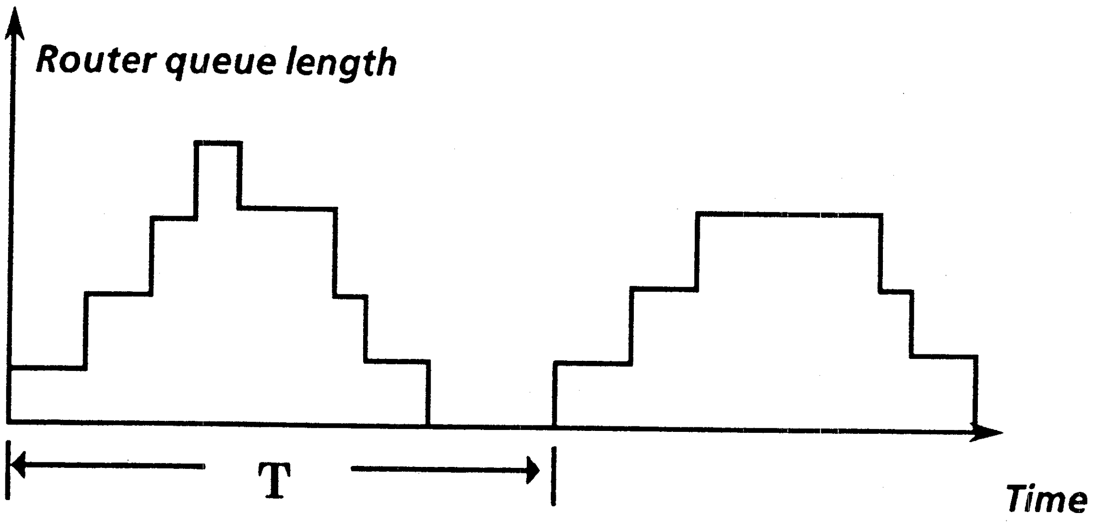
Observation: optimal power at Thresh$_1$ = 1, no hysterisis.

Feedback Filter

Congestion detection based on instanteous queue sizes results in feedback due to transient changes at router.

Filter to provide consistent signal to users from network.

Adaptive Averaging of Queue Length.



Determine cycle time T, at router. A cycle is (busy + idle) interval.
Compute average queue length over the cycle.
Use average to set bit for packets arriving in subsequent cycle.

Refinements to account for certain cases, e.g., long busy periods.
Average over (previous cycle + part of current cycle.)
Averaging performed as each packet arrives at router.

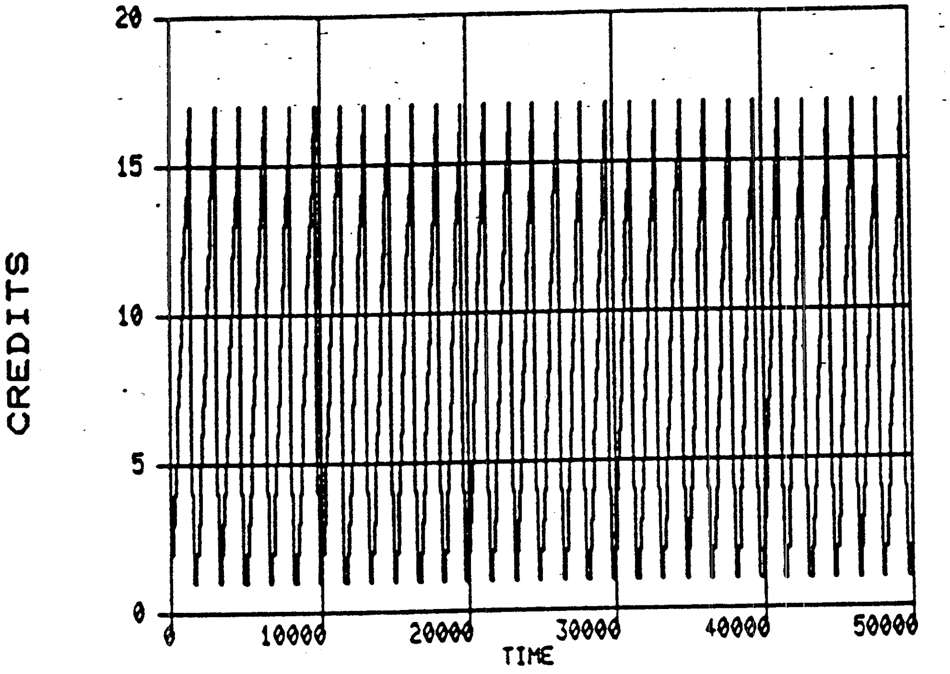
KKR 4/24/87

digital

Decision Function

How Frequently should the decisions be made?

1) Every Acknowledgement



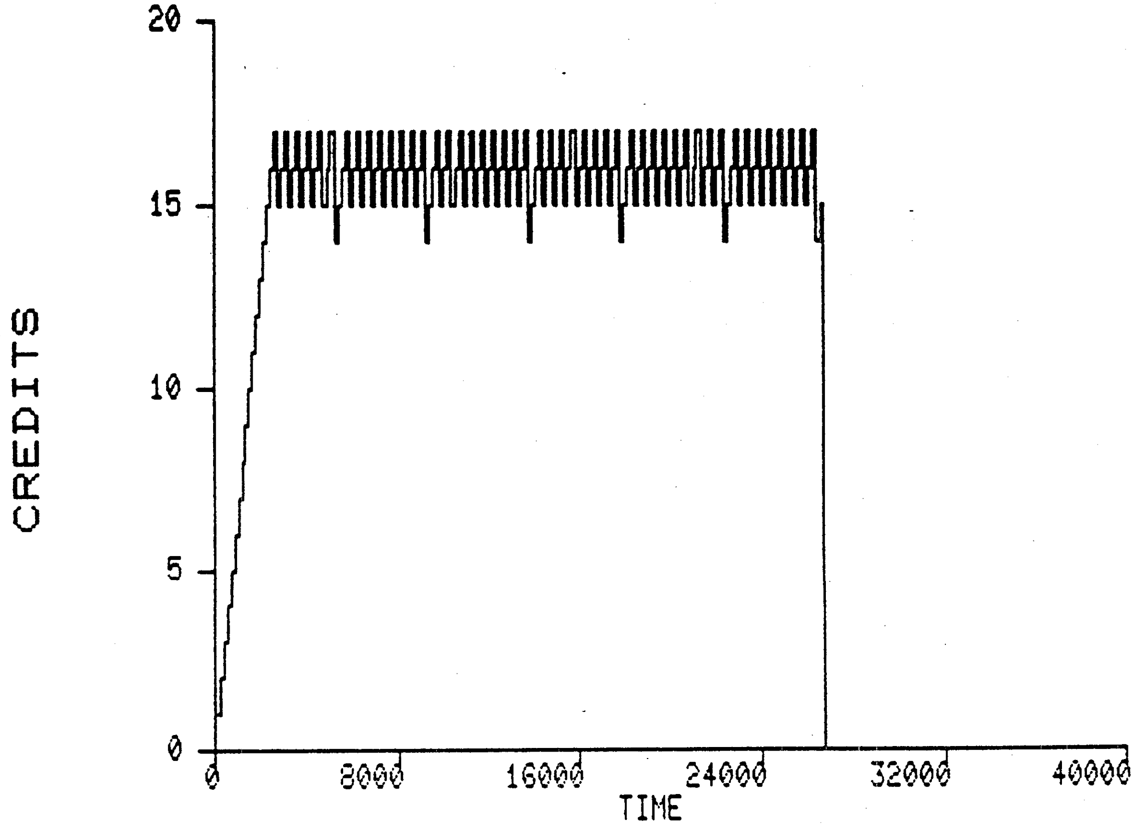
Observation: Considerable oscillation: over-correction.

Maintaining history of 'bits' from previous window
also causes over-correction.
erase old information, after a window update.

Frequency of Decision Making

Based on Current Window. Update every nW acks., n=1, 2,..

Decision frequency: allow control to take effect. Then monitor effect of change.
1st W (=1 round trip delay) for new window to take effect.
congestion avoidance bits received relate to previous window .
Next W bits received based on new window.



Conclusion: **Overall performance improves. Window size less oscillatory. Update Freq. = $W_{prev} + W_{curr}$.**

KKR 4/24/87

d i g i t a l

Signal Filtering

Proportion of bits received set dependent on distribution of packet size and router threshold.

deterministic service times:

above the knee : 100% of the bits set at the router.

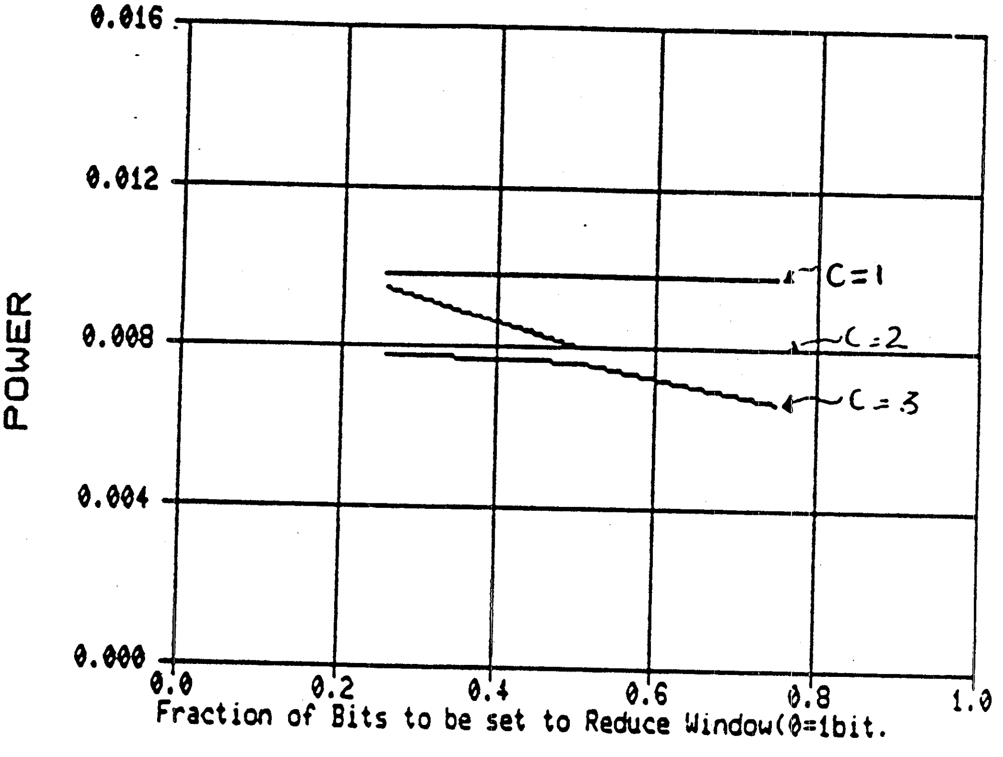
exponential service times, and utilization of router = 0.5.

C = 1: (1-P(0)) = ρ = 0.5. Thus, 50% of bits set at router.

C = 2: (1-P(0)-P(1)) = ρ^2 = 0.25. Thus, 25% of bits set at router.

Router threshold and signal filtering by user related.

Policy: A single cut-off determined by % bits received by the user being set/not set causes change in window.

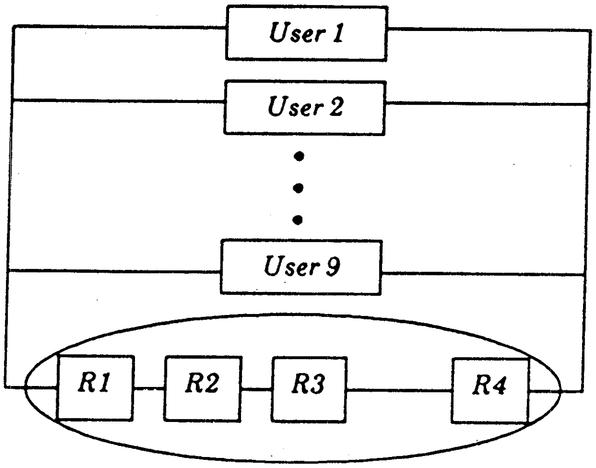


Observation:Cut-off at 50%bits of congestion avoidance bits set to trigger decrease of window

KKR 4/24/87

Decentralized Window Adjustment

(Increase and Decrease Algorithm)



Common Binary Feedback

Common Objective

Increase

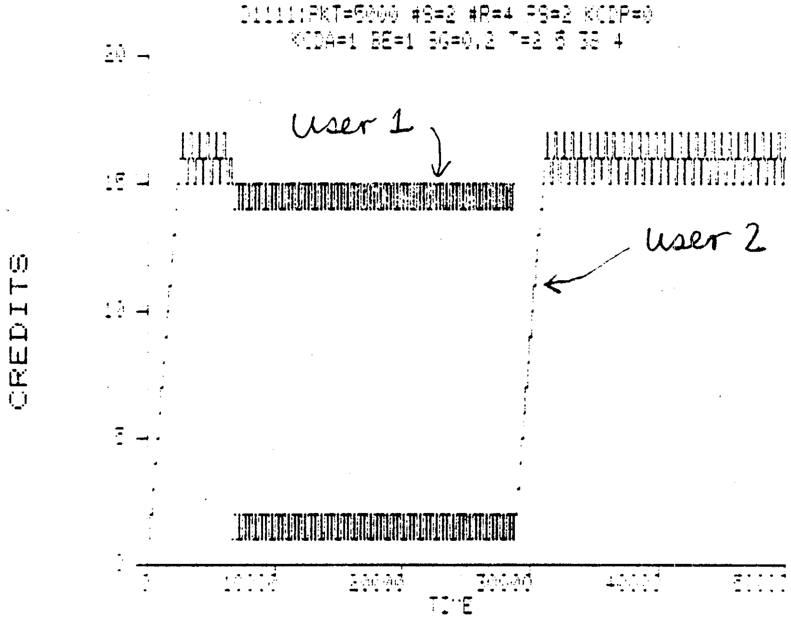
Additive: $W = W + a$

Multiplicative: $W = cW, c > 1$

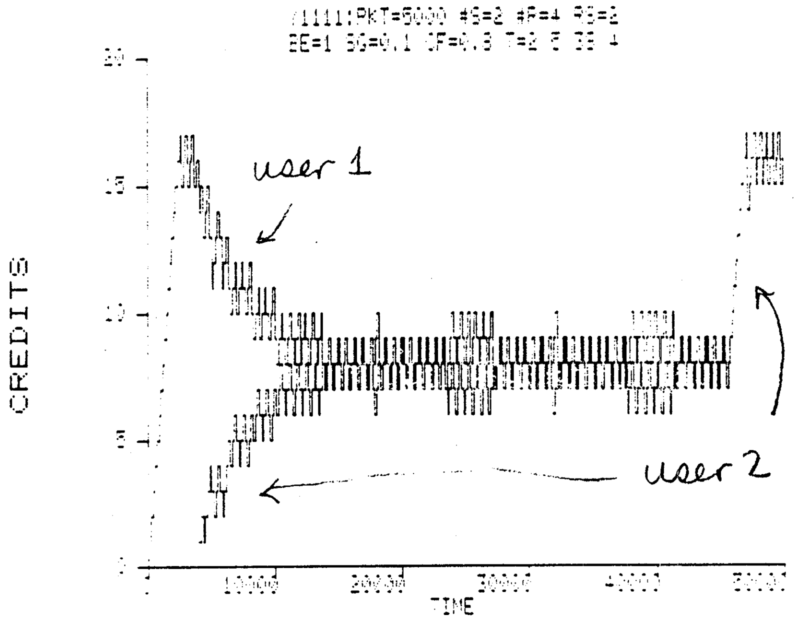
Decrease

Additive: $W = W - b$

Multiplicative: $W = dW, d < 1$

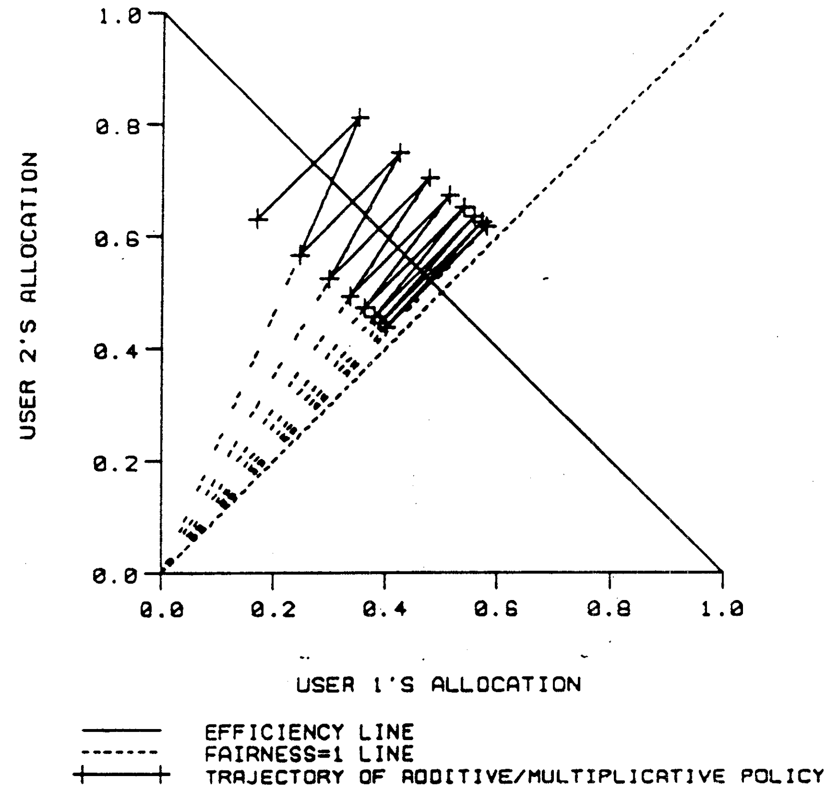


Additive Increase and Decrease



Additive Increase Multiplicative Decrease

Vector Representation of the algorithm



satisfies:

1. Sufficient conditions for convergence

2. Fewest parameter and most insensitive to configurations

Bounds and Rounding

Increase

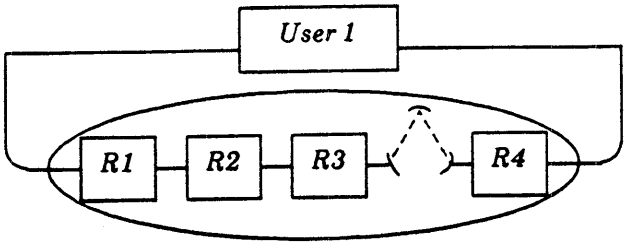
$$W_{new} = min(round(W_{old} + 1), W_{max})$$

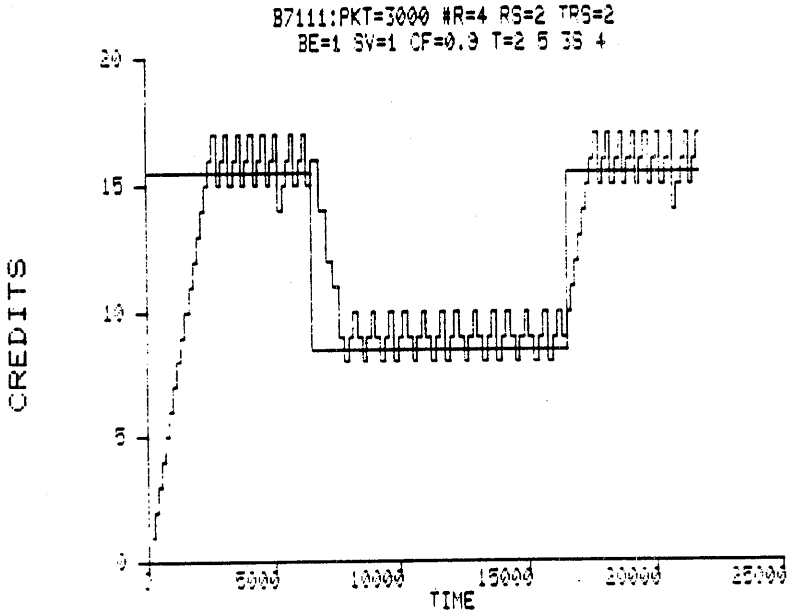
Decrease

$$W_{new} = max(round(min(0.875 * W_{old}, W_{old} - 1)), 1)$$

Responsiveness

Configuration: A user passing through four routers and a satellite link. R2's service time changes temporarily from 5 to 10.



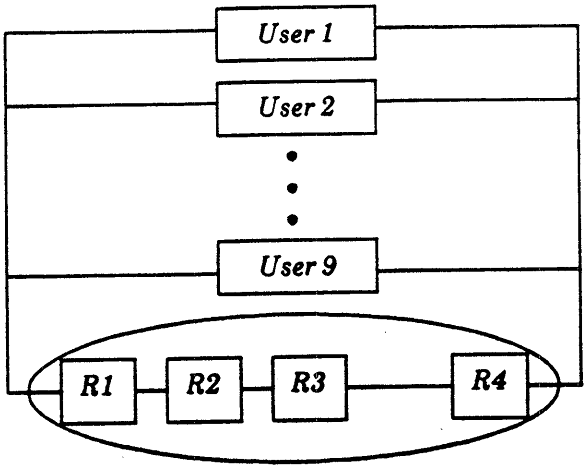


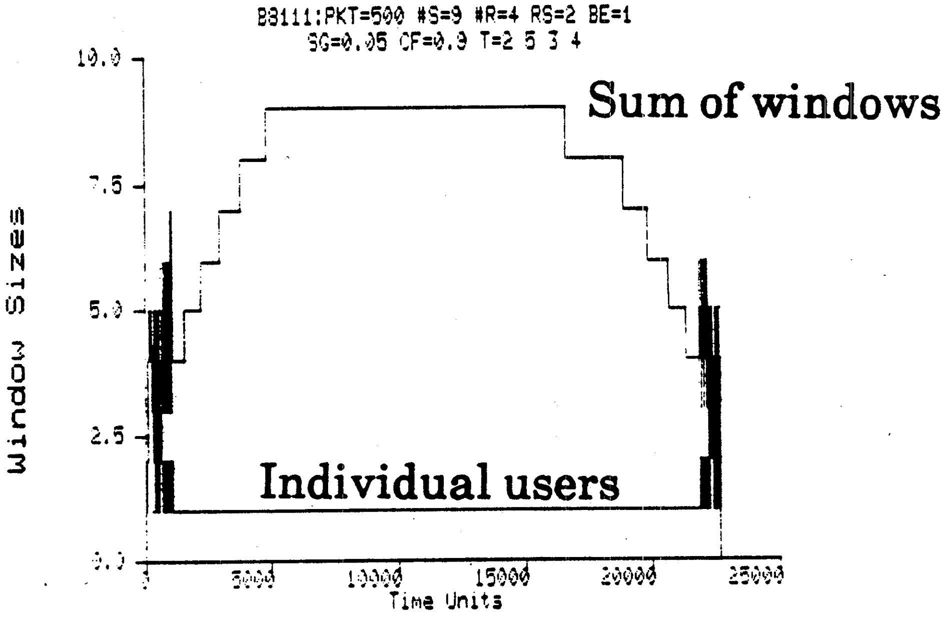
Conclusion: Yes! The binary feedback scheme is responsive.

4/24/87

Convergence

Configuration: Nine users sharing four routers.
Optimal window at knee = 3 ⇒ 1/3 per user

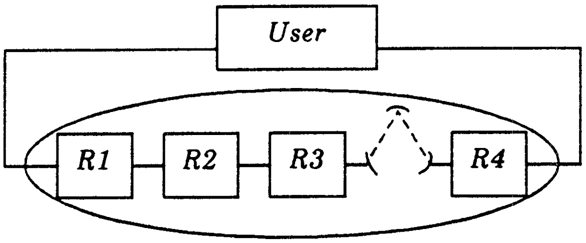




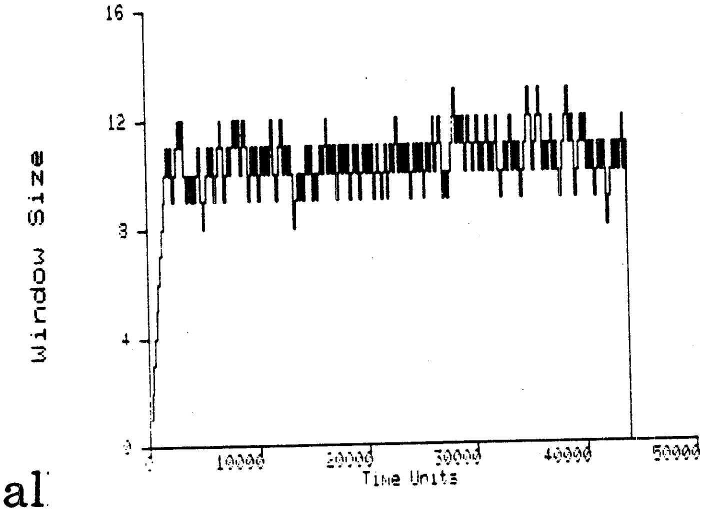
Conclusion: Yes! The binary feedback scheme converges.

Random Service Times

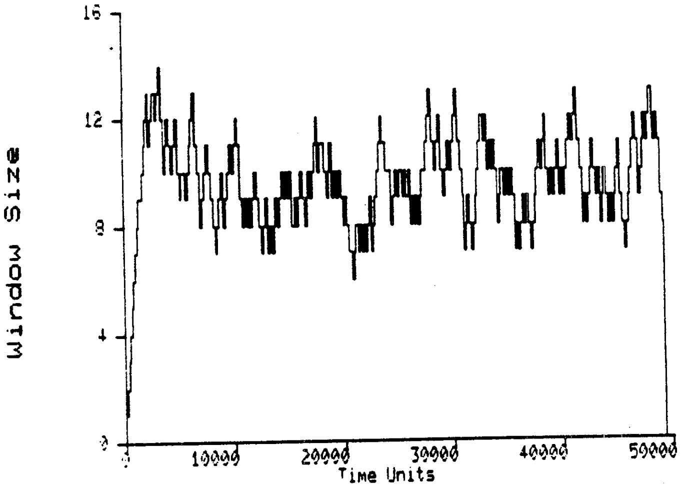
Configuration: A single user passing through four routers and a satellite link. Router service times are random.



Uniform



Exponential



Conclusion: Yes! The binary feedback scheme is robust.

4/24/87

Features of the Scheme

1. No new packets

   During overload or underload.

2. Distributed control

3. Low parameter sensitivity

   $Q_{threshold} = 1$

   Cutoff Percentage (% of bits set) = 50%

4. Minimum Oscillation Size

   Increase Amount = 1

   Decrease Factor = 0.875

5. Maximum information entropy.

   P(bit = 1) = P(bit = 0) = 0.5

   $Q_{threshold} = 2$, Cutoff Percentage = 25% ⇒ Less entropy

6. All parameters are dimensionless.

   No time values ⇒ Good for all link speeds and network sizes.

7. No prior reservation of resources

   Resources not used by one user are allowed to be used by others.

Summary

1. Congestion is not a static problem.

2. Congestion Avoidance:

   Operation with low delay and high throughput

   Independent of number of buffers.

3. Congestion can be avoided in connectionless networks.

4. Binary Feedback Scheme:

   User Policies: Decision fn (Collect $W_{old} + W$ bits,
   Examine the last W bits)
   Signal filter (up if < 50% bits set)
   Increase/Decrease ($W + 1, 0.875W$)

   Router Policies: Congestion Detection ($Q_{avg} = 1$)
   Feedback filter (Avg since last cycle)

5. The proposed scheme is efficient, fair, responsive, convergent, and robust.

APPENDIX B

Distributed Documents

The following documents/papers were distributed at the meeting.

- Excerpt of FCCSET Document
- GOSIP FIPS Statement
- Standards Listing

Excerpt of FCCSET Document

- 600
#593

NATIONAL SCIENCE FOUNDATION
WASHINGTON, D.C. 20550

nsf

OFFICE OF THE
ASSISTANT DIRECTOR
FOR COMPUTER AND INFORMATION
SCIENCE AND ENGINEERING

December 19, 1986

TO: Distribution

Dear Colleague

The National Science Foundation authorization act for the fiscal year 1987 (PL 99-383) requested a study of critical problems and current and future options regarding communications networks for research computers, including supercomputers, at universities and Federal research facilities in the United States. These computer network activities are funded and managed in several agencies of the government and your participation through the FCCSET committee networking activities provides an excellent mechanism for interagency cooperation. In this regard then, I am inviting you to participate on a panel to perform a study of communications networks for research computers as requested by PL 99-383 (see attached).

The NSF supported San Diego Supercomputer Center has agreed to host a Workshop on Computer Networks in San Diego on February 17-19, 1987. It is my hope and plan that this workshop will be a timely and beneficial forum for gathering and exchanging information across government, industry, and academic organizations.

This study represents an opportunity to survey network research needs, to surface computer network issues, to seek consensus on future goals, and to present these important areas to the Congress. I want to personally thank you for lending your valuable time and support to this effort.

Sincerely,

Gordon Bell

Gordon Bell
Assistant Director

Federal Coordinating Council on
Science, Engineering and Technology

Computer Network Study

WORKSHOP

Holiday Inn Embarcadero
San Diego, California
February 17-19, 1987

Workshop Agenda

Monday, February 16

3:00 pm - 7:00 pm Registration, Lobby Foyer

Tuesday, February 17

8:00 am - 9:00 am Registration, Convention Foyer

9:00 am - 10:00 am Coffee & Danish — Introduction to Workshop - Pacific BC off Convention Foyer James Burrows and Gordon Bell

10:00 am - 12:00N Planning Group and Working Group Meetings in the following rooms:

Room	Working Group
Captain 1	Group A - Internet Concepts Chair: Lawrence Landwebber
Pacific D	Group B - Networking Requirements and Future Alternatives Chair: Sandy Merola
Captain 2	Group C - Future Standards and Services Requirements Chair: Richard des Jardins
Captain 3	Group D - Security Issues Chair: Dennis Branstad
Captain 4	Group E - Government Role in Networking Chair: Jesse Poore
Captain 5	Group F - Special Requirements for Supercomputer Networks Chair: Robert Borchers
Circulate to Working Groups	Group G - Planning Group Chair: James Burrows

12:00 N - 1:00 pm	Deli buffet; Pacific BC
1:00 pm - 5:00 pm	Continuation of Group Meetings
5:00 pm - 7:00 pm	Cocktail Party; Pacific BC

Wednesday, February 18

8:30 am - 12:00 N	Group Meetings; coffee & danish
12:00 N - 1:00 pm	Seafood buffet Pacific A
1:00 pm - 5:00 pm	Continuation of Group Meetings Development of outline, summaries, and recommendations by each group
5:30 pm - 7:00 pm	Tour of San Diego Supercomputer Center and wine and cheese reception; bus transportation will be available from front of the Holiday Inn

Thursday, February 19

8:00 am - 12:30 pm	Working Group summary presentations to Planning Group: coffee & danish
12:30 pm - 1:30 pm	Sit down Luncheon, Pacific BC
1:30 pm - 5:00 pm	Discussion of Working Group reports and development o final report by Planning Group; Pacific D

Support arrangements: A terminal for electronic mail, a small copier and a thermofax machine will be available in the hospitality suite (Room 218). A macintosh Plus will be available for each Working Group during workshop and for a few hours in the early evening.

The San Diego Supercomputer Center is hosting the luncheons, cocktail party, and wine and cheese reception.

GROUP A

INTERNET CONCEPTS

This Working Group will cover:

1. Review of current networking activities at agencies sponsoring advanced scientific computing facilities.

2. Development of an interagency internet 1987-1992.

   - Technical issues
   - Management issues
   - Funding model
   - User services
   - Obstacles to interoperability

3. Vision/Goals for the future -- 1992-2000

   - Computing environment paradigm
   - Identification/Integration of new technologies and user services

CHAIR: Lawrence Landweber, University of Wisconsin

Members:

Vinton G. Cerf, Corporation for National Research Incentives
Henry Dardy, NRL
David Farber, University of Delaware
James Green, NASA Goddard Space Flight Center
Paul Green, IBM Hawthorne Research Laboratory
Anthony Lauck, DEC
James Leighton, LLNL
Barry Leiner, Research Institute for Advanced Computer Science
Richard Mandelbaum, University of Rochester
Ravi Mazumdar, Columbia University
John Morrison, Los Alamos National Laboratory
Jonathan Postel, University of Southern California, Information Sciences Institute

GROUP B

NETWORKING REQUIREMENTS AND FUTURE ALTERNATIVES

The Working Group will collect and analyze network information for research computer networks, and examine the methodology and feasibility of interconnecting existing network resources (including the possible use of fiber optic systems). Emphasis is on the five year timeframe. The planned process is as follows:

1. The networking needs of U.S. academic and Federal research programs will be collected and analyzed.

   Information to be submitted is expected to include both current network usage data as well as future networking planning information. Three independent groups from the DOE, NASA, and NSF communities will analyze the data collected. The individual groups performing the analyses will be represented at the San Diego workshop.

2. Each of the three groups performing an analysis of the data will distribute their reports by January 25, 1987.

   During the period of time preceding the mid-February workshop, workshop members are expected to examine and analyze all reports in the context of the Working Group Charter specified above.

3. The workshop scheduled for February 17-19, 1987, will constitute the major, perhaps only, meeting of this working group. The agenda will facilitate:

   - survey presentations by those agencies performing a network analysis;
   - industry trends and cost/capacity reports by corporate committee members;
   - optional initial "point of view" talk by any committee member;
   - open discussion, with emphasis on alternatives for addressing future networking needs;
   - preparation of a consensus analysis;

4. Generation of a final Work Group Report.

GROUP B: NETWORKING REQUIREMENTS AND FUTURE ALTERNATIVES

CHAIR: Sandy Merola, Lawrence Berkley Laboratory

Members:

Allison Brown, Cornell University
Paul Deitz, BRL Aberdeen Proving Grounds
Fred Fath, Boeing Computer Services
John Fitzgerald, Lawrence Livermore National Laboratory
Dennis Hall, Lawrence Berkeley Laboratory
Jack Haverty, BBN Communications Laboratory
Charles Kennedy, BRL Aberdeen Proving Grounds
Thomas Lasinski, NASA Ames Research Center
Fred McClain, San Diego Supercomputer Center
Pat McGregor, Contel Business System
Hugh Montgomery, Fermi National Laboratory
Sushil G. Munshi, United Telecom
Glenn Ricart, University of Maryland
Richard T. Roca, AT&T
Stan Ruttenberg, UCAR
Dave Stevens, Lawrence Berkeley Laboratory
Bob Wilhelmson, National Center for Supercomputing
Applications

GROUP C

FUTURE STANDARDS AND SERVICES REQUIREMENTS

The Future Standards and Services Requirements Working Group will develop a statement of trends and recommendations for standards and services requirements for future research networking in the 1990s. The assumption for the work is that widespread availability and low cost of specific network services depends on standards, but the standards in turn affect implementability and TCP/IP software in Berkeley UNIX, and the rapid growth of electronic mail using SMTP. The question to be asked is whether and how this interaction between standards and services should affect future research networks.

Issues to be addressed include the following:

1. What role should standards play in future network services developments?

2. What are the principal standardization trends that should be taken into account in planning future research networks?

3. How should standardization and networking research interact in the future?

4. How does internationalism affect this question?

Each working group member is requested to bring the workshop a brief white paper (2-5 pages) addressing one or more of these issues or identifying other issues important to this theme. Coordination and dissemination of white papers by telephone and electronic mail prior to the workshop is encourgaged. Each member will have 10 minutes (plus discussion) to give a 1-3 viewgraph presentation summarizing the key ideas in his/her white paper, as a springboard to opening the identification and discussion of the issues. We will then proceed by discussion and consensus (including minority views) in the remainder of the workshop to agree on what the issues are and how to address them in the working group report. Writing assignments will be given following the workshop to complete the working group report by mid March. Depending on the final number and distribution of people on the working group, we may break up into two groups, one on Hosts, Workstations and Network Services, and the other on Trunks, Access Links and Wiring the Campus.

GROUP C: FUTURE STANDARDS AND SERVICES REQUIREMENTS

Chair: Richard desJardins, CTA

Members:

Michelle Arden, Sun Microsystems
John Day, CODEX
Debbie Deutsch, BBN, Inc.
John Katz, The Analytic Sciences Corporation
Rich Pietravalle, DEC
Marvin Sirbu, Carnegie Mellon University
George Sullivan, DCA
Ash Trividi, Bell, Northern Research

GROUP D

SECURITY ISSUES

This Working Group will address issues such as:

- Isolation of researchers within an installation to data that they are authorized to access.
- Access of foreign researchers involved in cooperative projects while presenting unauthorized access or disclosure of sensitive information.
- Protection of communications media and planning for emergency mode communications.
- Security services needed for commercial, academic, and government environments.
- Security architectures.
- Laws, rules and policies governing computer security.
- Cost effectiveness of controls that respond to threats.

Government and private sector experts have been invited to contribute to Work Group discussions.

Chair: Dennis Branstad, National Bureau of Standards

Members:
*Roger Callahan, National Security Agency
Michael Corrigan, Department of Defense
Dorothy Denning, SRI, Inc.
Peter Dunningham, CRAY Computer
*Dave Golber, SDC/UNISYS
Dave Gomberg, MITRE Corporation
Gary Johnson, Department of Treasury
*Steve Kent, BBN, Inc.
Noel Matchett, Information Security Inc.
*Dan Nessett, Lawrence Livermore Labs
Gerry Popek, UCLA, Department of Computer Science
*Miles Smid, NBS
Douglas Price, Sparta, Inc.
*Joseph Tardo, Digital Equipment Corporation
Steve Walker, Trusted Information Systems

*Will attend San Diego Workshop; others will attend session at NBS on March 4.

GROUP E

GOVERNMENT ROLE IN NETWORKING

This working group will consider the issues involved in the role of local state and Federal governments in computer networking, access to the networks and in training of personnel to operate and use networks. The essential issues concern the governmental role in coordination, procurement, management and operation of networks and the association with universities. The role of Federal agencies in developing standards, in providing research resources, and in providing operating expenses will be addressed.

Academic, governmental, and private sector participants will be involved. Perspectives from several existing and planned large scale computing centers will be sought. This working group must interact strongly with the other working groups because the issues are dependent on the technical and other aspects of their recommendations.

Chair: Jesse Poore, University of Tennessee

Members:

Jane Alexander, U.S. Senate
Saul Buchsbaum, AT&T
Paul Huray, OSTP
Jim Infante, University of Minnesota
Bob Johnson, Florida State University
Robert Kahn, CNRI
John Killeen, Livermore MFECC
Ken Kliewer, Purdue University
Ken Wilson, Cornell University

GROUP F

SPECIAL REQUIREMENTS FOR SUPERCOMPUTER NETWORKS

The Working Group will include experts from government, academia, and industry to address two important supercomputer access issues:

1. The special networking requirements that must be addressed to provide meaningful access to supercomputers.

2. The status of supercomputer access for U.S. researchers and also with regard to network availability.

Chair: Dr. Robert Borchers, Lawrence Livermore Laboratory

Members: Charles Crum, National Cancer Institute, FCRF
Dennis Duke, Florida State University
Dieter Fuss, LLNL
Sid Karin, GA Technology
Larry Lee, Cornell University
Michael Levine, Pittsburgh Supercomputer Center
Norm Morse, LANL
Ari Ollikainen, NASA - Ames
Harry Reed, BRL Aberdeen

REPORT OF THE SUBPANEL ON NETWORKING REQUIREMENTS AND FUTURE ALTERNATIVES

March 1987

Ms. Alison Brown Assoc. Director Advanced Computing & Networks Theory Center, Cornell University
alison@tcgould.tn.cornell.edu

Dr. A. Fredrick Fath Communication Systems & Services Group, Boeing Computer Services

Mr. John Fitzgerald Assistant Director Planning and Finance National MFE Computer Center Lawrence Livermore National Laboratory
fitzgerald#j@mfe.mfenet

Mr. Philip Gross Technical Staff, Mitre Corporation
gross@mitre.arpa

Mr. Dennis Hall Head, Advanced Development Projects, Lawrence Berkeley Laboratory
hall@lbl-csam.arpa

Dr. Jack Haverty Director of Systems Engineering, BBN Communications Corp.
haverty@bbn.arpa

Mr. Charles Kennedy Director of Ballistics, Ballistic Research Laboratory
Kermit@brl.arpa

Dr. Thomas Lasinski NAS Workstation Systems Manager, NASA Ames Research Center
lasinski@ames-nas.arpa

Mr. Fred McClain Manager of Programming and Software Services, San Diego Supercomputer Center

Dr. Patrick McGregor Vice President of Engineering, Contel Business Networks

Mr. Sandy Merola Deputy Division Head Information and Computing Sciences Div. Lawrence Berkeley Laboratory
merola@lbl-csa3.arpa

Dr. Hugh Montgomery Head, Computing Department Fermi National Accelerator Laboratory
mont@fnal.bitnet

Dr. Sushil G. Munshi Vice President, Technology Planning, United Telecom

Professor Glenn Ricart Director, Computer Science Center
glenn@umd5.umd.edu

Mr. Richard T. Roca Director, Data Architecture Center AT&T Bell Laboratory
ihnp4!arch3!rtr@seismo.css.gov

Dr. Stan Ruttenberg Executive Director, CSNET, UCAR
stan@sh.cs.net

Mr. David Stevens Office of Computing Resources Lawrence Berkeley Laboratory
stevens@lbl-csa3.arpa

Dr. Bob Wilhelmson Associate Director, National Center for Super Computing Applications
wilhelm@ncsavmsa.bitnet

ABSTRACT

The subpanel recommends creation of an international, interagency networking facility for science, whose fifteen year mission is to: (a) Ensure that U. S. scientists have available the most advanced wide area networking facilities in the world. (b) Ensure that U. S. wide area network technology maintains a position of world leadership. A minimum of 1.5 Mbps access to major government and academic research centers should be provided. Such a network would greatly benefit the competitive position of the United States in scientific research. It would also place the United States in a leadership position in utilization of high bandwidth, wide area networks. United States industries supporting wide area networks technologies would gain a significant competitive advantage. An ongoing program of research and development into both wide area network technology and network management is necessary for this endeavor to be successful. As part of the second year study, the subpanel recommends an interagency coordinating committee be established to identify short term implementation issues that can be investigated and resolved in parallel with long term issues. This would provide immediate benefit to the nation's scientific community.

BACKGROUND

Many scientific research facilities in the U. S. consist of a single, large costly installation such as a synchrotron light source, a supercomputer, a wind tunnel or a particle accelerator. These facilities provide the experimental apparatus for groups of scientific collaborators located throughout the country. The facilities cannot be duplicated in all states because of cost. Wide area networks are the primary mechanism for making such facilities available nationwide. Examples include government supported wide area networks such as ARPAnet, HEPnet, MFEnet, MILnet, NASnet, NSFnet, SPAN, and so on, as well as commercial facilities such as Tymnet, BITnet and AT&T leased lines. The cost of such networks is generally much less than the cost of the facility.

Congress recently enacted legislation calling for an investigation of the fifteen year future networking needs for the Nation's academic and federal research computer programs. The Federal Coordinating Council on Science Engineering and Technology (FCCSET) formed a Network Study Group to coordinate investigation of the benefits, opportunities for improvements, and available options with particular attention to supercomputing. Within the Network Study Group, the Subpanel on Network Requirements and Future Alternatives was formed to identify network demand during the next five years and to recommend a strategy for meeting that demand. This document is the subpanel's report.

APPROACH

The following approach was taken:

+ The networking plans of the U. S. research community were analyzed, creating a five year network demand summary;
+ Corporations that provide telecommunications services were surveyed, with particular attention to the possible use of fiber optics and related cost/capacity gains;
+ Issues related to interagency sharing of network facilities were identified;
+ Alternative methodologies for meeting total network demand were considered;
+ A five year networking strategy was developed and presented to the FCCSET Network Study Group.

NETWORK DEMAND SUMMARY

Four methods of estimating network demand were used:

+ Analysis of existing network utilization:

Wide area networks are used by scientists to access unique remote facilities (supercomputers, accelerators, analysis software and databases) and as a critical mechanism for communication and coordination among the large geographically distributed U. S. and international scientific collaborations ([11], [12]). High speed local area networks are being connected to lower speed wide area networks throughout the research community. 1.5 Mbps (Megabits per second) technology, digital data service (DDS) and packet networks have been introduced to wide area networks, and their use has become widespread. Nevertheless, wide area networking capacity has not kept up with the levels found in local area networks. Some wide area networks handle both high data volume and highly interactive traffic over the same communications links. This results in suboptimal performance. At the functional level, wide area network user interfaces have not kept up with their counterparts in local area networks.

The subpanel heard presentations of current and planned networking in DOD, DOE, NASA, and NSF. Many scientific research centers funded by these agencies are physically connected to more than one network. The backbones for the major networks are similar in topology, and existing network links throughout the community are generally fully utilized. Some of these networks suffer severe overloading, resulting in significant performance degradation. Additionally, more ubiquitous access is needed by the University research community, especially at smaller institutions. For example, there is a clear unmet need for nationwide, high speed access to large scientific databases. The subpanel noted that in many cases demand for capacity seriously exceeded current supply ([4], [5], [6]).

+ Estimation based on typical site:

A direct estimation of network demand was made using a major NSF university site as a basis. Network usage included wide area network facilities for supercomputer access as well as an extensive local area network. An absolute level of network demand for the next five years was estimated using three different models: task, user, and external flow. The task model focused on the network load generated by typical network tasks. The user model identified demand as a function of typical university network users. The external flow model centered on the university as an entity, and estimated networking demand between it and other external locations. The three values of predicted network traffic were in agreement within an order of magnitude. They indicated a thousandfold increase in needed capacity over current network resources [10].

- Extrapolation from experience with local area networks:
  This method also projected need for a thousandfold increase in wide area network capacity over the next five years. A remote supercomputer access scenario was presented to demonstrate how network transparency can increase the speed and accuracy with which engineering decisions can be made. It was argued that one order of magnitude is needed to create a nationwide distributed file system on an existing 56 Kbps network, another order of magnitude is needed to provide interactive monochrome graphics ([2], [3]) and a third order of magnitude is needed to accommodate expected increases in basic computer speeds. As more users are added, further increases in demand are anticipated.
- Estimation based on expanded user community:
  The above analyses estimate load increases for existing network topologies. There is an important additional need to extend network service to the smaller universities throughout the nation. This would add another factor of two to three to the above estimates. Since by definition, these research sites are not currently connected to an existing wide area network, this represents a demand for more communications lines rather than an increase in line speeds [4].
  There is a further need to extend network service to international sites. Access to overseas scientific collaborations would significantly enhance the quality of U. S. science by providing researchers with access to remote experimental apparatus, data, and personnel. It would also enhance U. S. prestige in the scientific research community by providing overseas collaborators with access to U. S. facilities, data, and personnel. The effect on network traffic would be negligible, but network size would be increased dramatically.

SUPPLY

Several major U.S. telecommunications corporations were represented on the panel. They jointly provided a summary of expected industry-wide technological trends over the next five years ([1], [7], [8], [9]). Cost/capacity forecasts and opportunities for use of fiber optic technology in the U.S. scientific research community were also presented.

The leading trends in U.S. telecommunications technology are the decreasing cost of component materials and the widespread, though not ubiquitous, availability of fiber optics [14]. The transport capabilities of the U.S. telecommunications industry will greatly increase during the next five years, as witnessed by the following observations: Packet switching rates are expected to rise to 10,000 packets per second (25 Mbps). Digital circuits are widely available at 56 Kilobits per second (Kbps) today. Within the next five years ISDN switched and non switched circuits ranging from 64 Kbps to 1.5 Mbps will be available in the larger metropolitan areas of the U.S. The digital interexchange transmission rates available to users are at 1.5 Mbps in general, and will rise to 45 Mbps between larger metropolitan areas. 150 Mbps service could be made available by special arrangement. ISDN 64 Kbps service will be present in about 20% of the U.S. market by the end of the five year period. The ability of the user to customize service (such as time of day conversion and simultaneous coordinated voice and data), as well as the availability and general use of applications services (such as X.400 mail and electronic document interchange) will dramatically increase.

Fiber optic technology is driving media costs downward. The cost of basic private line telecommunications services could fall by a factor of 20% to 50% during the upcoming five years. Any expectation that fiber would more dramatically reduce costs to the typical telecommunications user must be balanced by the recognition that the fiber itself is only one component of total transmission service cost.

It was recognized that the combination of fiber optic technology and the large amount of aggregate interagency demand may offer the scientific research community unique opportunities to acquire increasingly cost effective bandwidth. This is only possible in the case of a long term lease of very high bandwidth circuits. This ensures industry recovery of capital investment costs. If such a national network infrastructure were established as a long term interagency goal, migration to such a topology is possible using existing standard telecommunications technologies, including satellite, microwave, copper, and fiber optic transmission media.

ALTERNATIVES

+ Supplying capacity:

The need to increase wide area network capacity by a thousandfold is justified both by increased opportunities for scientific breakthroughs and by maintaining the nation's position of world leadership in wide area network technology. While industry projections indicate the necessary bandwidth will certainly be available as a national backbone, the required bandwidth will not be available all the way to the end user's site. The subpanel felt the most cost effective way to proceed would be to provide the needed bandwidth in stages.

The subpanel recognized that a factor of about thirty could be achieved simply and cost-effectively by: (a) tuning existing protocol implementations and managing access, (b) installing smarter congestion control algorithms. (c) upgrading existing 56 Kbps trunks to 1.5 Mbps and 45 Mbps lines in a judicious manner, and (d) providing type-of-service routing for efficient performance on high data volumes as well as highly interactive traffic [6].

Beyond that, another factor of thirty is needed to meet the projected demand. The subpanel identified two promising approaches: (a) develop more optimal distribution of network services between user systems and server systems to make more efficient use of the available bandwidth, and (b) develop powerful gateway computers that compress data entering wide area networks and decompress it at its destination. Such machines could also provide encryption without significant additional overhead. The two approaches are entirely complementary. Thus, each might contribute a factor of 5 or 6, for a combined factor of 30X. However, optimal distribution software is not available today, and data compression computers are only available for video compression. Therefore, applied research in these and other promising approaches is required.

+ Improved usability:

The subpanel agreed that an interagency, international network would significantly enhance the U. S. scientific research environment. To ensure ease of use, some peripheral issues must be addressed:

Global management and planning: The ARPAnet provides valuable experience in operating connected networks without global management. For example, ARPAnet management reported that traffic generated by external networks created internal performance problems that are unmanageable. Similarly, inefficient protocol implementations cannot be prevented, since no central authority exits. The effect is to reduce network performance for all users. ARPAnet management concluded that global management is essential to provide guaranteed performance. The subpanel agreed with this conclusion.

User services: Consulting help and documentation are necessary for any facility accessed directly by end users. However, most scientists are not interested in networks per se, but only in the resources they make

available. If a network could be made transparent or nearly so, the need for consulting help and documentation would be significantly reduced.
Reliable: Wide area networks in scientific research must be more reliable than many existing networks because of their critical role in supporting operation of remote experiments.
Extensible: The network will grow significantly in the next fifteen years. It must be possible to expand it incrementally and to join it with other networks, both national and international.
Evolutionary: To prevent obsolescence, the network must be tolerant of change. It must be designed in such a way that new protocols and services can be added without significantly disrupting existing services. This ensures the nation's scientists will keep a competitive edge in advanced networking technology. The rich environment for development of new products, ensures that the technology itself maintains a competitive edge.

RECOMMENDATIONS

(1) An interagency scientific network facility should be created whose 15 year mission is to: (a) Ensure that U. S. scientists have available the most advanced networking facilities in the world. (b) Ensure that U. S. wide area network technology achieves and maintains a position of world leadership.

(2) A phased implementation plan should be developed to provide these advanced network facilities to the nation's scientists. Rough guidelines should be to increase the effective capacity of existing networks tenfold in three years, a hundredfold in five years and a thousandfold in ten years:
(a) Existing wide area scientific networks should be overhauled to provide 56 Kbps service to end users at about 30% of maximum load. 1.5 Mbps or 45 Mbps trunk lines would be necessary in some areas to provide the needed bandwidth to end users. Existing protocol implementations should be checked and tuned to eliminate unnecessary congestion from inefficient implementations. Networks from all U. S. government agencies funding academic and federal scientific research would be upgraded.
(b) Modern networking facilities such as wide area network file systems, distributed scientific databases, distributed window systems, and distributed operating

systems should be developed and installed, along with facilities for users to find and use network resources from remote sites. Existing communications facilities should be upgraded tenfold to 1.5 Mbps speeds to end users as necessary to handle anticipated increases in load. Very high bandwidth trunk lines may be necessary in some areas to provide the needed 1.5 Mbps service to end users.
(c) More advanced facilities such as wide area color graphics capabilities, and remote control of experiments should be developed and introduced. Existing communications capacity should be upgraded tenfold to handle the load increase by using hardware and software technology developed as a result of applied research.
(d) To handle an anticipated increase in hardware speeds, existing communications links should be upgraded another tenfold as newer faster computers become available in the mid 1990s.
(e) New local area network facilities should be tracked so that the more promising new products can be made available in wide area networks.
(f) Coverage should be expanded so that most colleges and universities in the U. S. will have access to the network in ten years, and the remainder in fifteen years.

(3) An applied research and development program in advanced communications and network techniques should be implemented to:
(a) Provide the technology needed to increase the effective bandwidth of communications links. (i) More optimal distribution of functions between local hosts and remote hosts to minimize the need for raw network bandwidth. (ii) High performance systems that compress data entering a wide area network and decompress it at its destination. (iii) Development of gateway technology in general. (iv) Utilization of formal language theory and other innovative techniques to design components that fail in a diagnosable manner.
(b) Provide better ways to access remote resources thereby increasing opportunities for scientific breakthroughs. Local area networks are the only cost effective testbed for such facilities today. As capacity of wide area networks increases, a new source for network innovations can be expected to emerge.
(c) Provide better tools and techniques for management of networks.

(4) An ongoing basic research program into future network architectures to ensure continued leadership in use of scientific networks, as well as national leadership in wide area network technology.

(5) The panel recommends that issues of network design, cost analysis, management authority, and implementation plans be addressed by the second year study. Within this framework, an interagency coordinating committee should be established to identify issues that can be investigated and resolved in the short term. An important short term issue is implementation of the first factor of thirty improvement to existing networks. This can provide immediate benefit to the nation's scientific community.

BENEFITS

Implementation of the above recommendations would provide the U. S. scientific research community with a significant competitive advantage. Modernization of the nation's wide area networks by increasing speed, functionality, and size increases opportunities for research advances significantly ([2], [3]). Greater network speed can reduce the time required to perform a given experiment, and increase both the volume of data and the amount of detail that can be seen. Scientists accessing supercomputers would benefit particularly, because access speed is often critical in this work. Improved functionality frees scientists to concentrate directly on their experimental results rather than on operational details of the network. Increased network size extends these opportunities to tens of thousands of individuals located at smaller academic institutions throughout the nation. These modernization measures would significantly enhance the nation's competitive edge in scientific research.

The components of a shared network infrastructure obviously benefit from global management. The positive effects of such an approach are widespread. Centralized administration of research in wide area networks would minimize duplication of effort and rapid resolution of identified high priority problems. A global management structure would also allow a matrix approach to this distributed network expertise.

The U. S. communications industries would also gain a significant competitive advantage. Development of modern, low cost distributed computing facilities for wide area networks would help maintain the United States position of world leadership in networking technology. Use of these products in support of science will accelerate the development of newer products by U.

S. industry to meet challenges from both Europe and Japan. The United States would thus gain a position of world leadership in utilization of wide area, high bandwidth networks. This would increase the nation's competitive edge in communications technology as well as scientific research. As a spinoff, it would help maintain the U. S. leadership position in computer architectures, microprocessors, data management, software engineering, and innovative new networking facilities.

GOSIP FIPS Statement

FIPS PUB

FEDERAL INFORMATION
PROCESSING STANDARDS PUBLICATION

U.S. DEPARTMENT OF COMMERCE/National Bureau of Standards

GOVERNMENT OPEN SYSTEMS INTERCONNECTION PROFILE (GOSIP)

CATEGORY: SOFTWARE AND HARDWARE STANDARD
SUBCATEGORY: COMPUTER NETWORK PROTOCOLS

FIPS PUB

U.S. DEPARTMENT OF COMMERCE, Malcolm Baldrige, Secretary
NATIONAL BUREAU OF STANDARDS, Ernest Ambler, Director

Foreword

The Federal Information Processing Standards Publication Series of the National Bureau of Standards is the official publication relating to standards, guidelines, and documents adopted and promulgated under the provisions of Public Law 89-306 (Brooks Act) and under Part 6 of Title 15, Code of Federal Regulations. These legislative and executive mandates have given the Secretary of Commerce important responsibilities for improving the utilization and management of computers and automatic data processing in the Federal Government. To carry out the Secretary's responsibilities, the NBS, through its Institute for Computer Sciences and Technology, provides leadership, technical guidance, and coordination of Government efforts in the development of standards, guidelines and documents in these areas.

Comments concerning Federal Information Processing Standards Publications are welcomed and should be addressed to the Director, Institute for Computer Sciences and Technology, National Bureau of Standards, Gaithersburg, MD 20899.

James H. Burrows, Director
Institute for Computer Sciences and Technology

Abstract

This Federal Information Processing Standard (FIPS) specifies the use of the Government Open Systems Interconnection Profile (GOSIP) for the acquisition of networks and services. GOSIP defines a common set of data communications protocols which enable systems developed by different vendors to interoperate and enable the users of different applications on these systems to exchange information.

KEY WORDS: computer communications; Federal Information Processing Standards Publication; GOSIP; information exchange; International Standards Organization; interoperability; networking; open systems; protocols; protocol standards.

Natl.Bur.Stand. (U.S.) Fed.Info.Process.Stand.Publ.(FIPS PUB)
______, ______ pages.
(1987)
CODEN:FIPPAT

For sale by the National Technical Information Service, U.S. Department of Commerce, Springfield, VA 22161.

Federal Information
Processing Standards Publication ______

(date)

Announcing the Standard for

GOVERNMENT OPEN SYSTEMS INTERCONNECTION PROFILE (GOSIP)

Federal Information Processing Standards Publications are issued by the National Bureau of Standards pursuant to the Federal Property and Administrative Services Act of 1949, as amended, Public Law 89-306 (79 Stat. 1127), and as implemented by Executive Order 11717 (38 FR 12315, dated May 11, 1973), and Part 6 of Title 15 Code of Federal Regulations (CFR).

Name of Document. Government Open Systems Interconnection Profile (GOSIP).

Category of Document. Hardware and Software Standards, Computer Network Protocols.

Explanation. This Federal Information Processing Standard adopts the Government Open Systems Interconnection Profile (GOSIP). GOSIP defines a common set of data communication protocols which enable systems developed by different vendors to interoperate and enable the users of different applications on these systems to exchange information. These Open Systems Interconnection (OSI) protocols were developed by international standards organizations, primarily the International Organization for Standardization (ISO) and the Consultative Committee on International Telephone and Telegraph (CCITT). GOSIP is based on agreements reached by vendors and users of computer networks participating in the National Bureau of Standard (NBS) Workshop for Implementors of Open Systems Interconnection.

Approving Authority. Secretary of Commerce.

Maintenance Agency. U.S. Department of Commerce, National Bureau of Standards (Institute for Computer Sciences and Technology).

Cross Index.

a. NBSIR 87-3353, Final Implementation Agreements for Open Systems Interconnection Protocols, NBS Workshop for Implementors of Open Systems Interconnection, March 1987.

b. NBSIR 87-3354, FTAM (File Transfer, Access, and Management) Phase 2 Implementation Agreements, NBS Workshop for Implementors of Open Systems Interconnection, March 1987.

Related Documents. Related documents are listed in the Reference Section of the GOSIP document.

Objectives. The primary objectives of this standard are to:

- to achieve interconnection and interoperability of computers and systems that are acquired from different manufacturers in an open systems environment
- to reduce the costs of computer network systems by increasing alternative sources of supply

- to facilitate the use of advanced technology by the Federal government

- to stimulate the development of commercial products compatible with Open Systems Interconnection (OSI) standards

__Specificatiens.__ GOSIP (affixed).

__Applicability.__ GOSIP is to be used by Federal government agencies when acquiring computer network products and services and communications systems or services that provide equivalent functionality to the protocols defined in the GOSIP documents. Currently, GOSIP supports the Message Handling Systems and File Transfer, Access and Management applications. GOSIP also supports interconnection of the following network technologies: CCITT Recommendation X.25; Carrier Sense Multiple Access with Collision Detection (IEEE 802.3); and Token Bus (IEEE 802.4). Additional applications and network technologies will be added to later versions of the GOSIP document.

__Implementation.__ This standard is effective ____________ (six months following publication). For a period of eighteen months after the effective date, agencies are permitted to acquire alternative protocols which provide equivalent functionality to the GOSIP protocols. Agencies are encouraged to use this standard for solicitation proposals for new network products and services to be acquired after the effective date. This standard is mandatory for use in all solicitation proposals for new network products and services to be acquired after ____________ (eighteen months after the effective date). OSI protocols providing additional functionality will be added to GOSIP as implementation specifications for these protocols are developed by the NBS Workshop for Implementors of OSI. For a period of eighteen months after these new protocols are included in GOSIP, agencies are permitted to acquire alternative protocols which provide equivalent functionality. After the eighteen month period, the new protocols should be cited in solicitation proposals when systems to be acquired provide equivalent functionality to the protocols defined in the GOSIP document.

For the indefinite future, agencies will be permitted to buy network products in addition to those specified in GOSIP and its successor documents. Such products may include other non-proprietary protocols, proprietary protocols, and features and options of OSI protocols which are not included in GOSIP.

Waivers. Heads of agencies may waive the requirements of this standard in instances where it can be clearly demonstrated that there are significant performance or cost advantages to be gained and when the overall interests of the Federal government are best served by granting the waiver. Waivers may be requested for special purpose networks which are not intended to interoperate with other networks. Waivers may also be requested for products supporting network research.

A request for waiver generated within an agency shall include:

a. a description of the existing or planned ADP system for which the waiver is being requested,

b. a description of the system configuration, identifying those items for which the waiver is being requested, and including a description of planned expansion of the system configuration at any time during its life cycle, and

c. a justification for the waiver, including a description and discussion of the significant performance or cost disadvantages that would result through conformance to this standard as compared to the alternative for which the waiver is requested.

Agency heads may act only upon written waiver requests. Agency heads may approve requests for waivers only by a written decision which explains the basis upon which the agency head made the required finding(s). Within thirty (30) days of approving a waiver, a copy of each such decision, with procurement sensitive or classified portions clearly identified, shall be sent to the Director, Institute for Computer Sciences and Technology, National Bureau of Standards, Gaithersburg, MD 20899. Also, a notice of the waiver determination shall be published in the Commerce Business Daily.

A copy of the waiver request, any supporting documents, the document approving the waiver request and any supporting and accompanying document(s), with such deletions as the agency is authorized and decides to make under 5 U.S.C. Sec. 552(b), shall be part of the procurement documentation and retained by the agency.

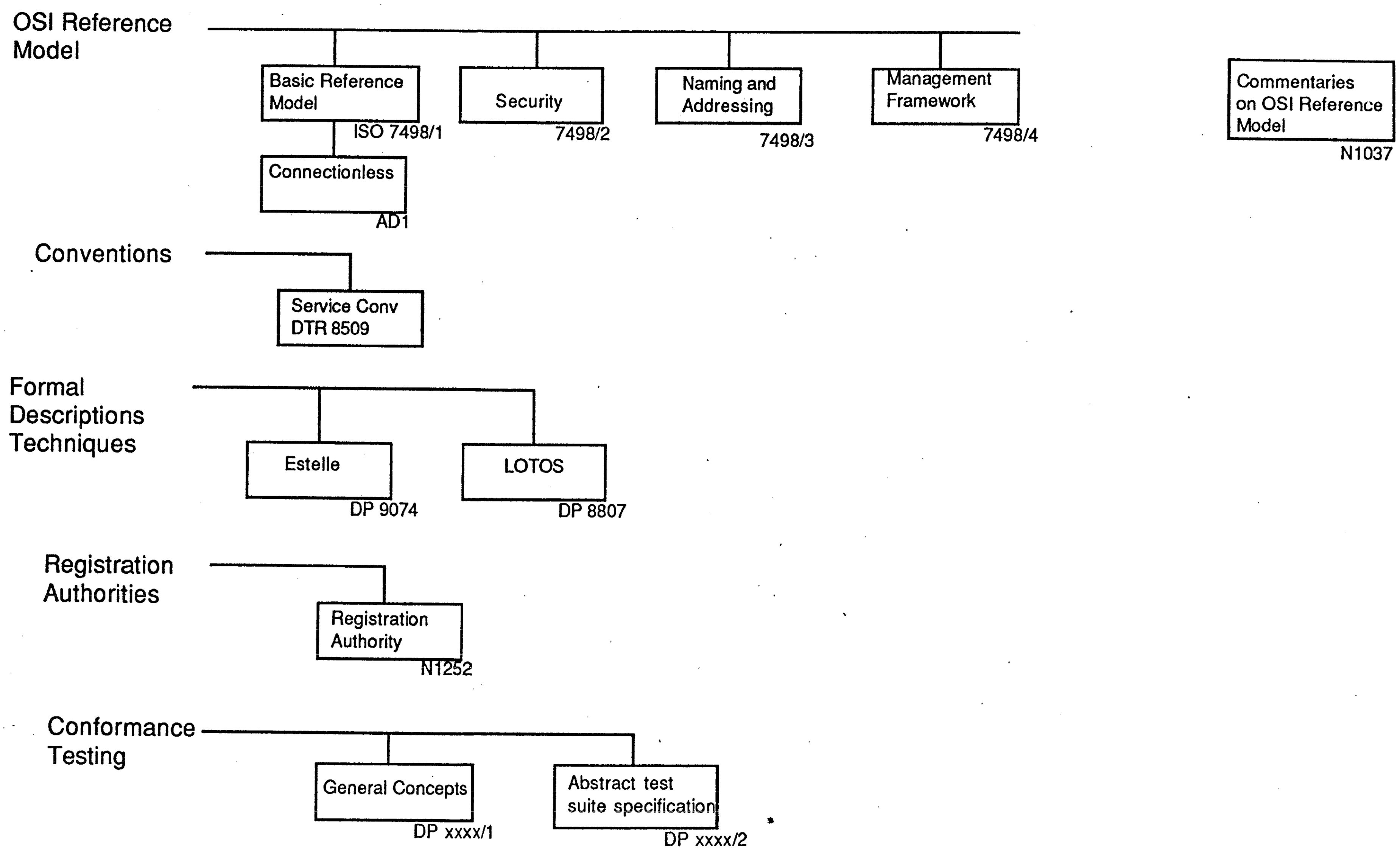
Special Information. The appendices to the GOSIP specification describe advanced requirements for which adequate profiles have not yet been developed. Federal government priorities for meeting these requirements and the expected dates that work on these priorities will be completed are also provided. As these work items are addressed and completed by the NBS Workshop for Implementors of OSI, addenda will be inserted into the GOSIP document.

Where to Obtain Copies. Copies of this publication are for sale by the National Technical Information Service (NTIS), U.S. Department of Commerce, Springfield, VA 22161. When ordering, refer to Federal Information Processing Standards Publication _____ (FIPSPUB_____), and title. Specify microfiche if desired. Payment may be made by check, money order, or NTIS deposit account.

Standards Listing

Supplement ~~from~~ to Chapter 5 two

Layer Independent Standards



OSI Reference Model

Basic Reference Model ISO 7498/1

Connectionless AD1

Security 7498/2

Naming and Addressing 7498/3

Management Framework 7498/4

Commentaries on OSI Reference Model N1037

Conventions

Service Conv DTR 8509

Formal Descriptions Techniques

Estelle DP 9074

LOTOS DP 8807

Registration Authorities

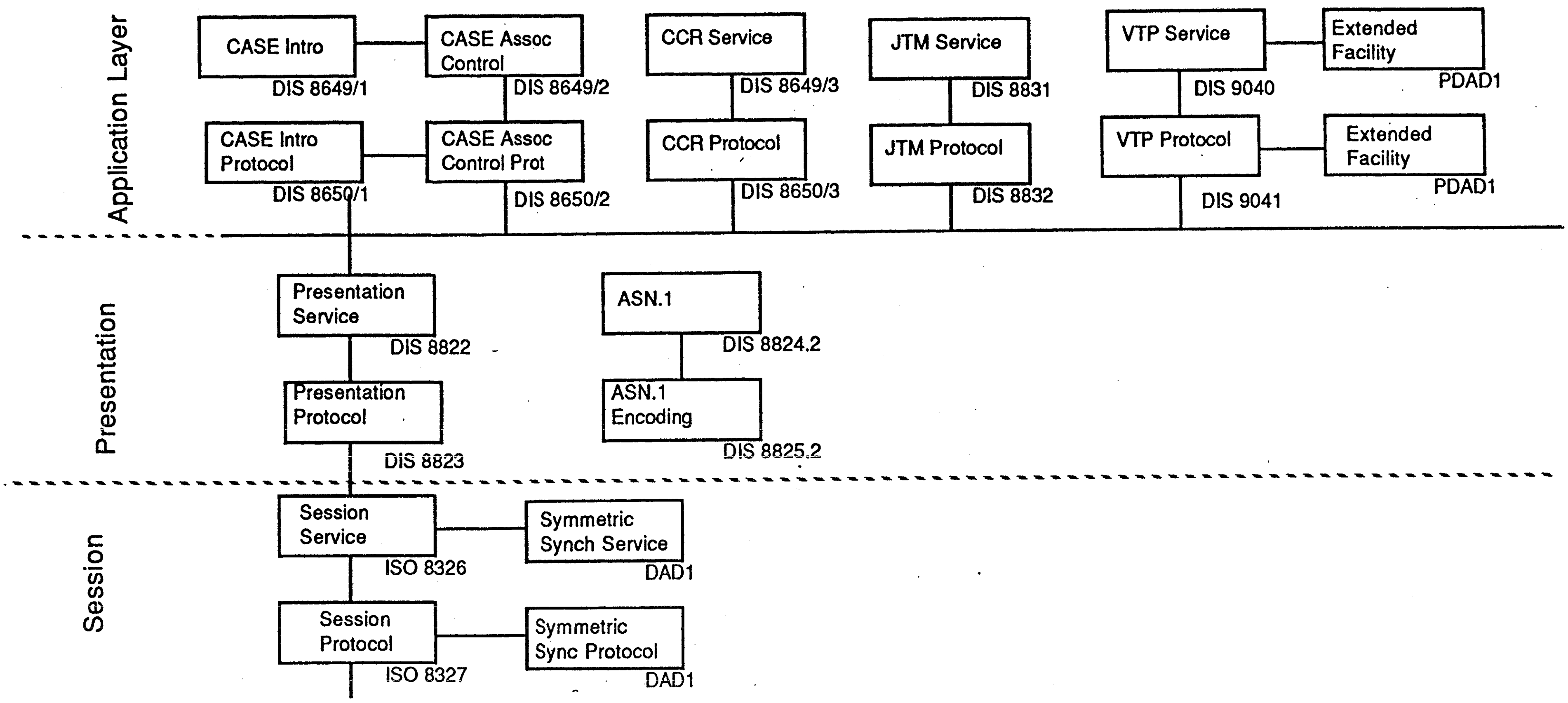
Registration Authority N1252

Conformance Testing

General Concepts DP xxxx/1

Abstract test suite specification DP xxxx/2

Upper Layer Standards



Application Layer

CASE Intro
DIS 8649/1

CASE Assoc Control
DIS 8649/2

CCR Service
DIS 8649/3

JTM Service
DIS 8831

VTP Service
DIS 9040

Extended Facility
PDAD1

CASE Intro Protocol
DIS 8650/1

CASE Assoc Control Prot
DIS 8650/2

CCR Protocol
DIS 8650/3

JTM Protocol
DIS 8832

VTP Protocol
DIS 9041

Extended Facility
PDAD1

Presentation

Presentation Service
DIS 8822

ASN.1
DIS 8824.2

Presentation Protocol
DIS 8823

ASN.1 Encoding
DIS 8825.2

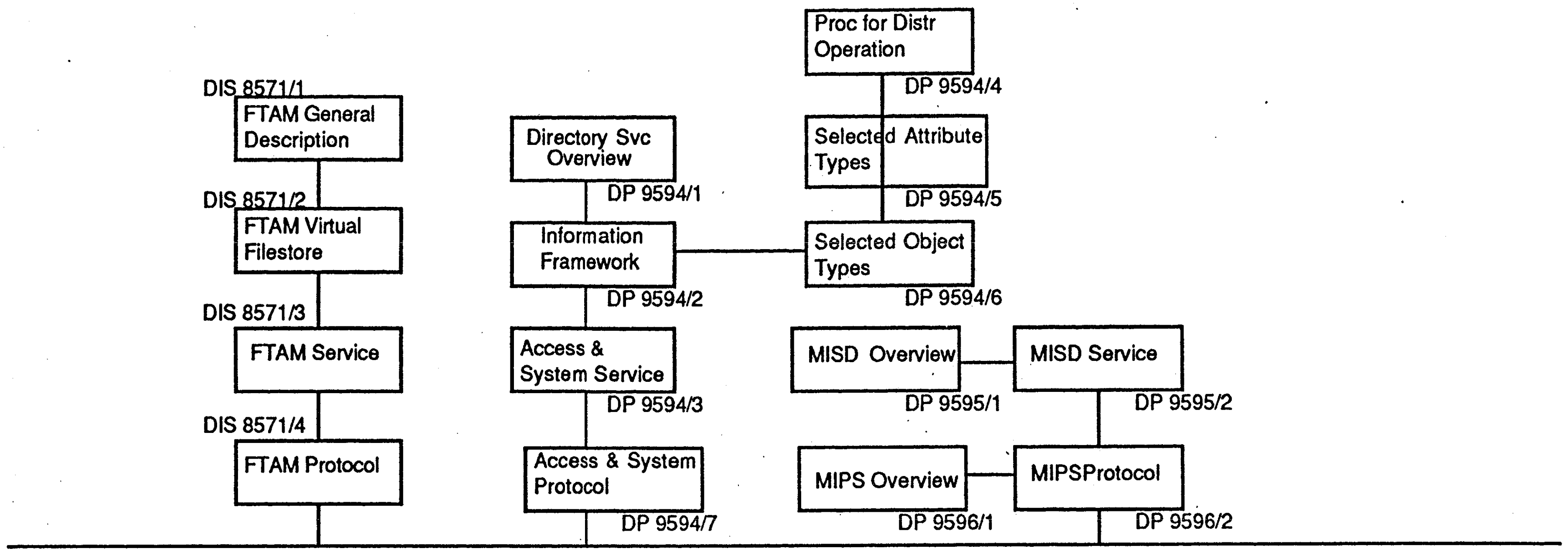
Session

Session Service
ISO 8326

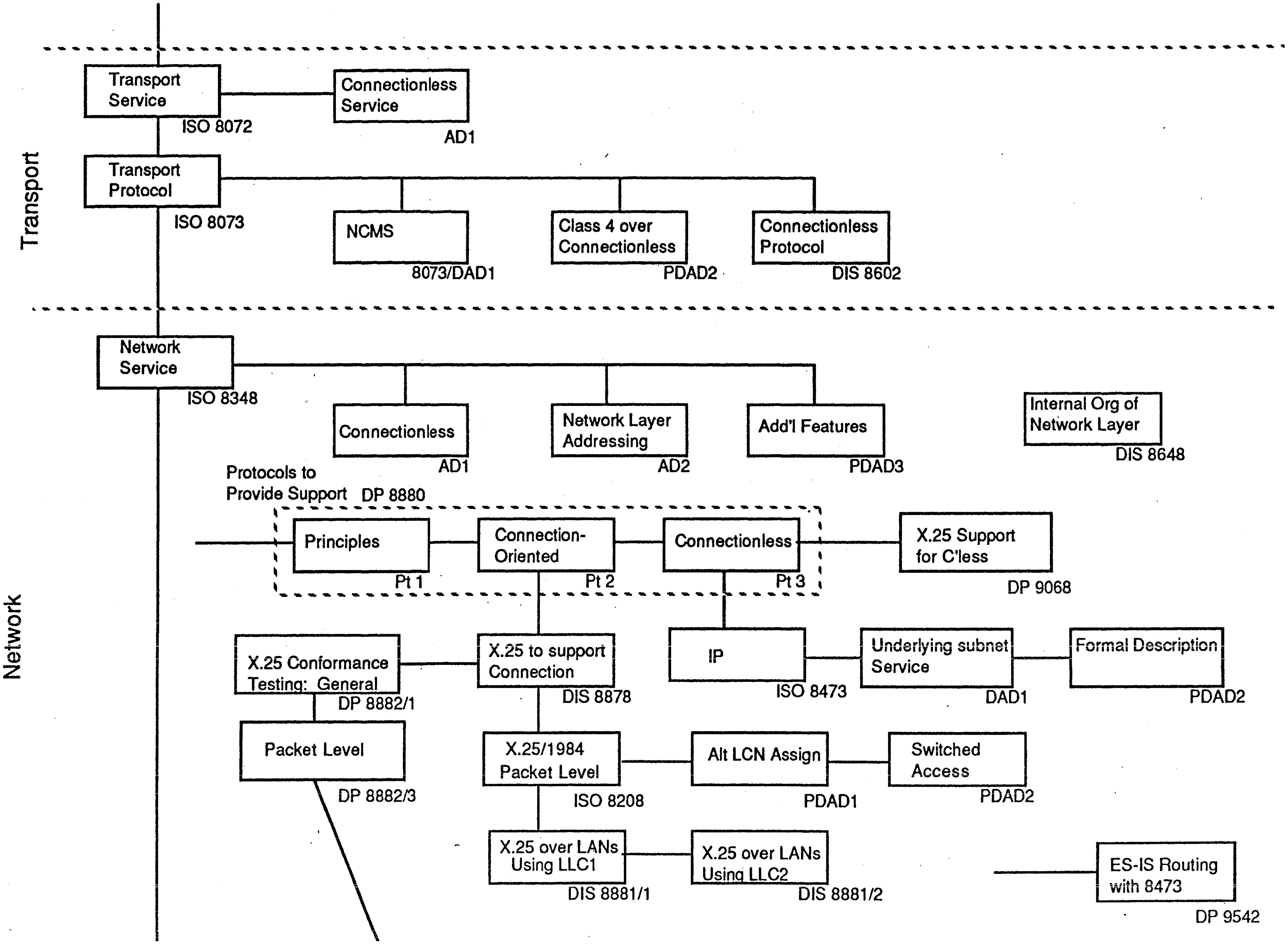
Symmetric Synch Service
DAD1

Session Protocol
ISO 8327

Symmetric Sync Protocol
DAD1


DIS 8571/1
FTAM General Description
DIS 8571/2
FTAM Virtual Filestore
DIS 8571/3
FTAM Service
DIS 8571/4
FTAM Protocol
Directory Svc Overview
DP 9594/1
Information Framework
DP 9594/2
Access & System Service
DP 9594/3
Access & System Protocol
DP 9594/7
Proc for Distr Operation
DP 9594/4
Selected Attribute Types
DP 9594/5
Selected Object Types
DP 9594/6
MISD Overview
DP 9595/1
MISD Service
DP 9595/2
MIPS Overview
DP 9596/1
MIPSProtocol
DP 9596/2

Lower Layer Standards



Transport

Transport Service ISO 8072

Connectionless Service AD1

Transport Protocol ISO 8073

NCMS 8073/DAD1

Class 4 over Connectionless PDAD2

Connectionless Protocol DIS 8602

Network

Network Service ISO 8348

Connectionless AD1

Network Layer Addressing AD2

Add'l Features PDAD3

Internal Org of Network Layer DIS 8648

Protocols to Provide Support DP 8880

Principles Pt 1

Connection-Oriented Pt 2

Connectionless Pt 3

X.25 Support for C'less DP 9068

X.25 Conformance Testing: General DP 8882/1

Packet Level DP 8882/3

X.25 to support Connection DIS 8878

IP ISO 8473

Underlying subnet Service DAD1

Formal Description PDAD2

X.25/1984 Packet Level ISO 8208

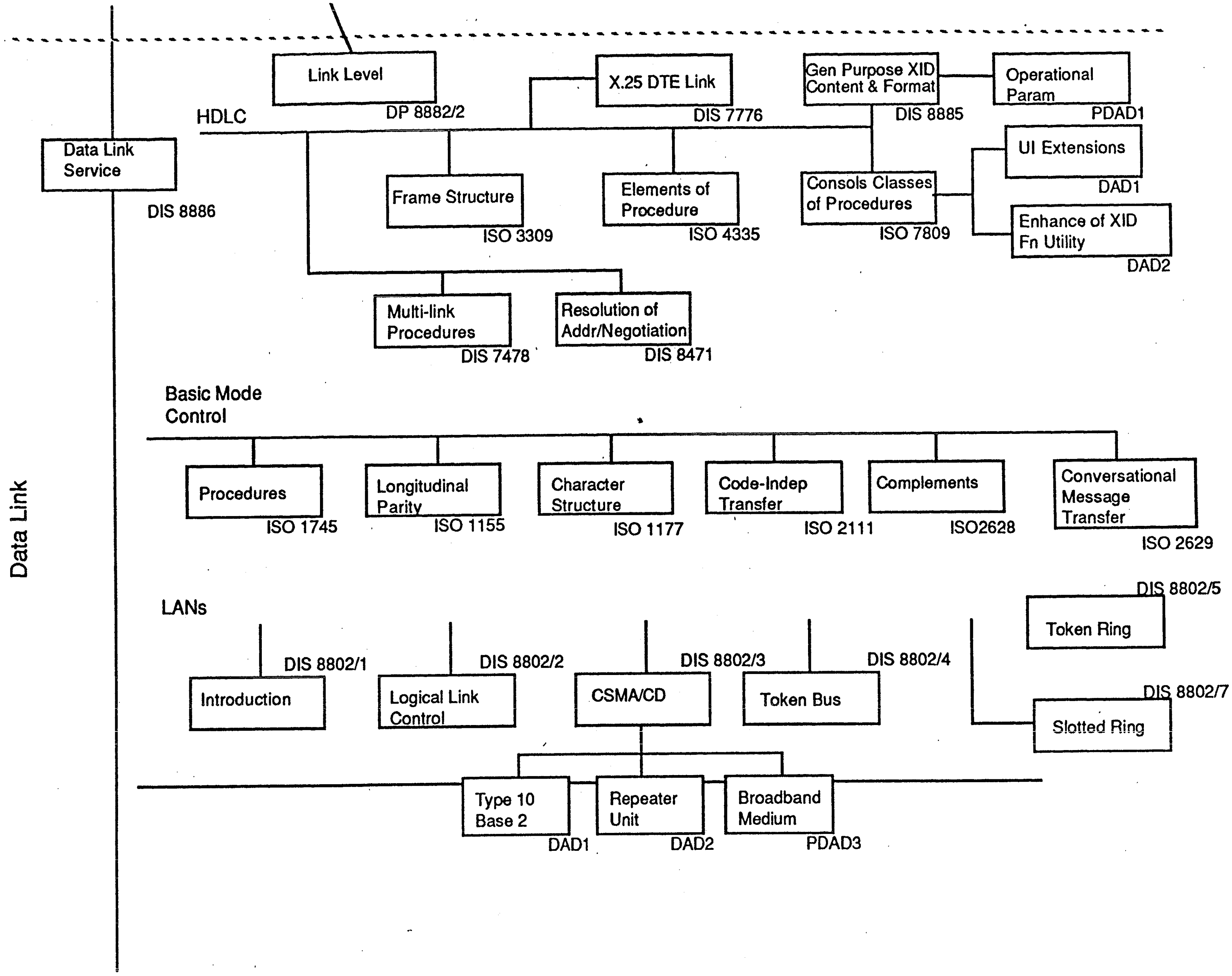
Alt LCN Assign PDAD1

Switched Access PDAD2

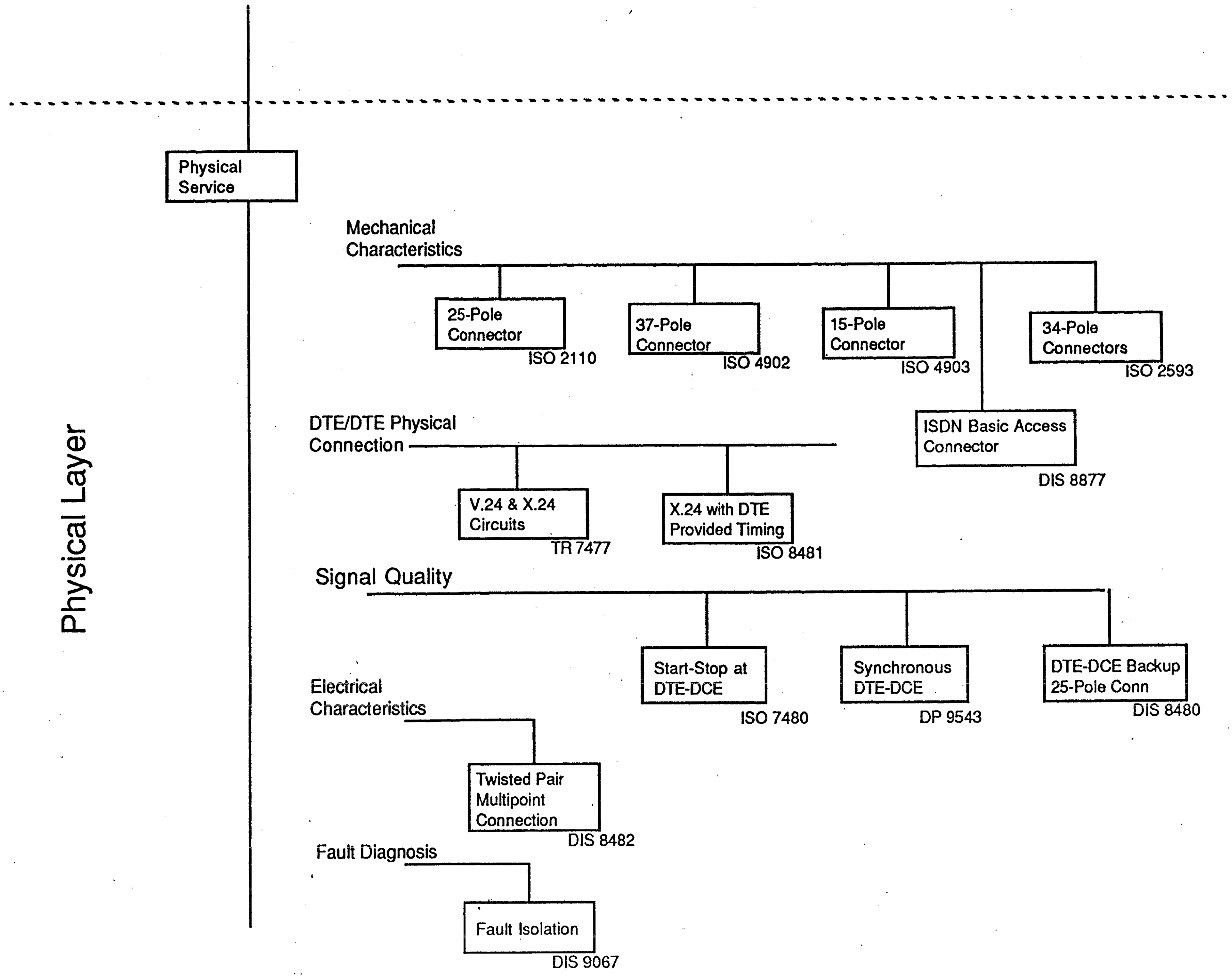
X.25 over LANs Using LLC1 DIS 8881/1

X.25 over LANs Using LLC2 DIS 8881/2

ES-IS Routing with 8473 DP 9542


Data Link
Data Link Service
DIS 8886
HDLC
Link Level
DP 8882/2
X.25 DTE Link
DIS 7776
Gen Purpose XID Content & Format
DIS 8885
Operational Param
PDAD1
Frame Structure
ISO 3309
Elements of Procedure
ISO 4335
Consols Classes of Procedures
ISO 7809
UI Extensions
DAD1
Enhance of XID Fn Utility
DAD2
Multi-link Procedures
DIS 7478
Resolution of Addr/Negotiation
DIS 8471
Basic Mode Control
Procedures
ISO 1745
Longitudinal Parity
ISO 1155
Character Structure
ISO 1177
Code-Indep Transfer
ISO 2111
Complements
ISO2628
Conversational Message Transfer
ISO 2629
LANs
Introduction
DIS 8802/1
Logical Link Control
DIS 8802/2
CSMA/CD
DIS 8802/3
Token Bus
DIS 8802/4
Token Ring
DIS 8802/5
Slotted Ring
DIS 8802/7
Type 10 Base 2
DAD1
Repeater Unit
DAD2
Broadband Medium
PDAD3

Physical Layer



IETF Internet Problem Descriptions

At the first IETF meeting in January 1986, a list of Internet problems was developed covering short, intermediate and long range issues. At the most recent IETF meeting in February 1987, an attempt was made to develop such a list in a more rigorous fashion. The IETF membership was divided into groups with the goal of compiling problem descriptions in particular areas. The resulting Internet Problem Descriptions are contained in this appendix and are a mixture of intermediate range protocol issues and very short range O&M issues.

Problems were listed in the following format:

Problem Description:

Severity: (low, medium, high)

Time Frame: (time until problem becomes critical)

Owner: (Responsible Agency or group)

Plan/Options:

The original forms have been edited to combine or eliminate redundant descriptions. The problem list is not exhaustive and further work will 1) develop a more complete list,

2) divide into categories by timeframe and

3) prioritize within category.

Problem Description:

Internet doesn't work under heavy load (ie, Congestion)
For example, existing DDN protocols can't efficiently handle gateways between networks of grossly different band-width (e.g., Ethernet- Arpanet)

Severity: High

Time Frame: Immediate

Owner: DDN/DARPA

Plan/Options:

Short term: Add capacity to existing infrastructure

Intermediate term:
1) Develop congestion control for DoD IP
2) Investigate existing solutions outside the DDN community.

Long Term Research: Look at new Internet schemes; eg, Internet Connection Oriented Protocol

Problem Description:

1) Lack of ISO Connection-Less Internet Protocol in current Internet Gateways.

2) Lack of ES-IS

Severity: Low now, grows to severe in 2 years

Time Frame: 2 years

Plan/Options:

1) Set/define "standards" for how ISO IP should be used

2a) Start funding contractors to implement ISO/IP in gateways

2b) Purchase gateways with ISO/IP

3) Deploy in Internet Infrastructure starting in 6 - 18 months

4) Run some applications (FTAM, etc) to gain experience. Modify standards goto 2)

Also: Work with Standard's Organization to apply DoD IP experience into ISO/IP

Problem Description:

MILNET domain adoption plan

Severity: Low - now; Medium - 6 mo; High - 1-2 years

Time Frame: (see Severity)

Owner: DDN/OSD

Plan/Options: Plans needed for vendor documentation and advice, administrator documentation, migration plan and RFC updates.

Problem Description:

Short-term Internet Routing Problems; eg,
Extra-Hop, table space (routing) performance, buffering limitations in LSI-11, mail bridges (gateways)

Severity: High

Time Frame: Immediate

Owner: DDN/BBN

Plan/Options:

1) Deploy Butterfly Mailbridge Gateways in Parallel with LSI-11 GW's in about 6 months

2) Transition Core to Butterflies MB's

3) Remove LSI-11's

Requires SW/HW to be deployed before configuration mgt and testing is completed.

Problem Description:

Internet Information Management; eg, Much duplication; needless distribution info; congestion problems

Severity:

Time Frame:

Owner: DDN

Plan/Options:

- reconvene the group "! %@"
- include regional NIC reps
- look at what info is needed
- look at what is duplicated
- create info "way stations"
- share tools; techniques
- keep general centers informed of who to hand off users to
- distribute data collection
- SRI-NIC acts as reference and replicates data strategically

Basic model - interlibrary loan system for traditional libraries; everybody contributes; everybody wins; nobody pays too much or foots the whole bill; some systems are shared others are translated; the general NICs hand off to the specialized ones

Coordinate host liaison, host administrators, etc., by holding meetings; getting input for net administrators

Problem Description:

Name Servers

(1) Get root servers off heavily loaded hosts

(2) See that name servers are well distributed

(3) Migration of name service to login hosts
(service then part of backbone service) and
(equipment maintained by backbone)

Severity: Medium

Time Frame: 6 months

Owner: DDN

Plan/Options:

(1) Can negotiate immediately to get servers off heavily loaded hosts; evenly distributed throughout net

(2) NIC can coordinate Berkeley to get good BIND (UNIX/VAX version) of domain service

(3) Bring NIC into BARNET so we are on NSF net

(4) Need more capacity in login hosts; needs $ but easy to solve

Problem Description:

No organization exists to attend to problems which transcend network boundaries, Internet O&M is not defined

Severity: High

Time Frame: Immediate

Owner: NSF/DARPA/NASA/DDN

Plan/Options:

(1) Define network and Internet O&M at next IETF

(2) Determine organization suitable to do O&M

(3) Draft RFC defining Internet O&M

Problem Description:

IOP Facility in PSN 6 can drop messages. The current IOP module in PSN Release 6 behaves very much like a gateway; if an 1822 host sends messages faster than a standard X.25 host can receive them, some percentage of the messages will be dropped. The impact of this feature on future Internet performance should be considered.

Severity: Low

Time Frame: (Next PSN Release)

Owner: DDN/BBN

Plan/Options:

a) Determine whether this feature will exist in future PSN releases.

b) If so, evaluate potential impact on Internet performance as standard X.25 gateways are more widely deployed.

Problem Description:

Lack of protocol testing is a severe problem in gateways and hosts. Incompatible implementations abound.

Severity: High

Time Frame: Immediate

Owner: OSD/DCA

Plan/Options:

1) Accept the situation - ISO is coming anyway

2) Establish testing center(s) funded by

   a) gov't
   b) vendors
   c) private enterprise

3) If none of 2) can get funded then spend money on advertising who the apparent "winners" are anyway; i.e., let the marketplace decide

Problem Description:

Networking research must continue into the forseeable future. Should its operational base be TCP/IP or ISO? TCP/IP is more accessible for manipulation, but ISO will be more prevalent and thus more realistic in terms of providing the problems to be researched. But will ISO implementations be "modifiable" for/by researchers? and how will vendors track the research?

Severity: High

Time Frame: 5 years

Owner: IAB

Plan/Options: Establish a study group IETF to outline the problem and report to all interested parties: gov't, researchers, vendors, users. While this looks like it overlaps with FCCSET, if they don't succeed in addressing it, the problem won't disappear.

Problem Description:

Although several agencies have cross-country trunks, some of these are seriously congested while others are unused. Sharing of under-utilized trunking may help solve network congestion.

Severity: ARPA 10
NASA 0
NSF 2

Time Frame: Immediate

Owner: IAB

Plan/Options: Interagency agreements? IRI?

Problem Description:

Procedures for making changes in DDN and the internet are too cumbersome; eg,

o Line-in/line-out coordination;
o line-at-a-time acquisition leasing wastes available leverage;
o new nodes, new hosts, additional circuits.

Severity: High

Time Frame: Immediate

Owner: DECCO

Plan/Options: Review of current administration procedures by sponsoring agencies. Develop new management organization. Study NASA trunking concept.

Problem Description:

Insufficient processing and memory capacity at some some Arpanet PSNs. Several sites are either memory- or CPU-bound because of the growth of users and gateways

Severity: High

Time Frame: Immediate

Owner: DDN

Plan/Options: Upgrade approporiate nodes from C3OE's to C300's. The sites are

o SRI 51,
o ISI 27,
o RCC 5,
o WIS 94

Problem Description:

Insufficient cross-country bandwidth on ARPANET. Highly utilized lines induce retransmissions at the store and forward level resulting in long delays for traffic between the two coasts. This in turn increases the congestion and resource use seen at the source and destination of the traffic.

Severity: High

Time Frame: Immediate

Owner: DDN

Plan/Options: Install 2 new cross-country links:
o MIT44 - SRI151;
o ISI22 - Columbia (APL)

Problem Description:

Internet audit trail/billing sharing

Severity: Low

Time Frame:

Owner: IAB

Plan/Options: This is probably part of larger Network Operations. Toward this end

- We can share audit trail/billing system
- Cooperate in building a useful interagency billing system
- Make capacity planning reports available for ARPANET, MILNET, etc.

Problem Description:

EGP

Severity: High

Time Frame: 6-12 Months

Owner: DDN/DARPA

Plan/Options: Draft of EGP2 by Jose Rodrigues (SDC) and Mike StJohns (DDN) for next IETF

Problem Description:

EGP Topology Restrictions

A) Common metric

B) Core gateway computation load

C) Information hiding by cores leads to lost information and suboptimal routes

D) Political restrictions - autonomy

Severity: High (very important to NSF)

Time Frame: Immediate

Owner: NSF/DARPA

Plan/Options:

1) Remove 3rd party routing restrictions

2) Increase base of trusted gateways/autonomous systems

Likeliest is new (unspecified) protocol

Problem Description:

Gateway authentications

A) What's a real gateway?

B) What routes can a gateway advertise?

Severity: Low

Time Frame: 2 years

Owner: OSD, IETF

Plan/Options: Non-authenticated gateways present denial-of-service threats, as well as wiretapping traffic

Problem Description:

Interior Gateway Protocol Problems; eg,

A) GGP traffic volume

B) GGP/EGP interactions

C) Common metrics, algorithmically converted to EGP common metric

D) Current IGP's not published (RIP, SPF, CISCO)

Severity: Medium

Time Frame: 12-24 Months

Owner: IETF, DDN, BBN

Plan/Options: 1. Document existing IGP's

2. Define standard (suggested/example) IGP

Problem Description:

Mail Bridges

1) Administrative restrictions/routing interactions

2) Name servers use Mail Bridges

Severity: Medium

Time Frame: 12 months

Owner: DDN

Plan/Options: When Mail Bridges are shut down to non-mail transit traffic, there will be a furor aimed at DCA.

Problem Description:

Gateway Redirection

Intermediate gateway decides that an alternate route is better, has no way to inform previous gateway.

Severity: Medium

Time Frame: 12-24 Months

Owner: DDN/DARPA

Plan/Options: Develop improved Internet routing/ICMP model.