Proceedings of the Tenth Internet Engineering Task Force

June 15-17, 1988 in Annapolis, MD

Edited by

Gladys Reichlen

Allison Mankin

October 1988

TENTH IETF

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

ACKNOWLEDGMENTS

Producing the Proceedings for the quarterly plenary sessions of the IETF requires a concerted effort on the parts of many. Fortunately, there were many contributors to the effort.

Allison Mankin, Gladys Reichlen, and Anne Whitaker, (all from MITRE) wrote the meeting notes and working group presentation summaries (except for those submitted by working group chairs). Gladys Reichlen finished compiling and assembling the final version. Allison Mankin and Phill Gross (MITRE) proofread and edited the final document for technical accuracy.

Additionally, I would like to thank all the Working Group chairpersons who responded with their timely reports. Reporting of Working Group activity is extremely important, especially for those whose schedules would not permit them to attend a specific meeting. For this IETF plenary, five of the seventeen Working Groups contributed to the Proceedings.

Finally, I'd like to thank Terry Slattery of the US Naval Academy for hosting the June 15-17, 1988 meeting. Hosting an IETF meeting has evolved to the point where it now requires a great deal of preparation and we are all grateful for Terry's efforts.

TABLE OF CONTENTS

TABLE OF CONTENTS (Concluded)

1 CHAIRMAN'S INTRODUCTION

The meeting in Annapolis was filled with energy and activity. There were approximately 120 attendees and thirteen of the (then) 17 Working Groups met and reported. Since that time, the number of Working Groups has both swelled and receded. Several new groups have been formed and five have retired after completing at least the current phase of their charter.

The fifteen current active groups and their status is listed in the table below. Not all of the WG reports were compiled as part of this preliminary version of the Proceedings. The final version which will be provided to the NIC will have all current WG reports.

Let me again thank all those who have contributed to making the IETF a successive group. There is an incredible amount of collective energy channeled through the IETF toward the resolution of Internet issues. I am constantly amazed at how successful you have all made this effort.

Active Working Groups	Charter? (Form 2)	RFC or IDEA?	Met at USNA?	Current Report?	Meeting at Ann Arbor?
Authentication	Yes	Yes	Yes	Yes	Yes
CMIP-over-TCP (CMOT)	Yes	Yes	-	-	Yes
Interconnectivity	Yes	-	NA	NA	Yes
InterNICs	-	-	-	Yes	Yes
Host Requirements	Yes	-	Yes	Yes	Yes
Internet MIB	Yes	Yes	Yes	-	Yes
Open SPF-based IGP	Yes	Yes	Yes	Yes	Yes
Open INOC	Yes	-	Yes	Yes	-
Open Systems Routing	Yes	Yes	Yes	-	-
PDN Routing Group	Yes	Yes	Yes	Yes	Yes
Performance and CC	-	-	Yes	Yes	Yes
Pt-Pt Protocol	Yes	Yes	Yes	-	Yes
RIP Advisory Group	Yes	NA	NA	NA	NA
ST and CO-IP	Yes	Yes	NA	NA	Yes
TELNET Linemode	Yes	Yes	Yes	Yes	Yes
Groups with completed missions					
Domain	-	Yes	Yes	-	NA
EGP3	Yes	Yes	-	-	NA
OSI Technical Issues	Yes	Yes	-	-	NA
Short Term Routing	Yes	Yes	Yes	Yes	NA
SNMP Extensions	-	Yes	Yes	-	NA

2 IETF ATTENDEES

The following is a list of people who attended all or part of the June 1988 IETF meeting. All organizational affiliations are listed as submitted, and for brevity have not been expanded (Example: DCA vice Defense Communications Agency).

Name	Organization	Email Address
Adkins, Sherrill	DCA	sra@edn-vax.arpa
Alterman, Peter	HHS/PHS	————————
Almes, Guy	Rice University	almes@rice.edu
Beasley, Larry	USNA	————————
Berggreen, Art	ACC	art@acc.arpa
Biviano, John	MITRE	gateway.mitre.org
Blake, Coleman	MITRE	cblake@gateway.mitre.org
Boivie, Rick	IBM	rboivie@ibm.com
Borman, David	Cray Research	dab%hall.cray.com@uc.msc.umn..edu
Bosak, Len	Cisco	Bosack@methom.cisco.com
Bostwick, Bill	DOE	bostwick@nmfecc.arpa
Braden, Bob	USC/ISI	braden@venera.isi.edu
Bradley, Terry	Wellfleet Comm	linus!wellflt!tbradley
Braun, Hans-Werner	U of Michigan	hwb@mcv.umich.edu
Brescia, Mike	BBNCC	brescia@park-street.bbn.com
Brim, Scott	Cornell Theory Ctr	swb@tcgould.tn.cornell.edu
Brooks, Charles E.	DAC	uunet!cos!stubby!ceb
Cain, Ed	DCA	cain@edn-unix.arpa
Callon, Ross	BBNCC	rcallon@bbn.com
Case, Jeff	Univ of Tenn	case%utkvxl.decnet@utkcs2.cs.utk.edu
Cavallini, John	HHS	————————
Chiappa, Noel	MIT	jnc@xx.lcs.mit.edu
Chinoy, Bilaz	Merit/NSFNET	bnc@merit.edu
Choy, Joe	NCAR/USAN	choy@windom.ucar.edu
Collins, Michael	LLNL	collins@nmfecc.arpa
Curley, John	NRC	curley@nrcmol.bitnet
Davin, James	Proteon	jrd@monk.proteon.com
Disque, Robert	USNA	disque@usna.mil
Fedor, Mark	NYSERNET	fedor@nic.nyser.net
Feinstein, Hal	MITRE	gateway.mitre.org
Finkelson, Dale	MIDnet	dmf@fergvax.unl.edu
Fischer, Allan	USNA	allan@usna.mil
Fonash, Pete	DCA	fonash@edn-vax.arpa
Foster, Robb	BBNCC	robb@park-street.bbn.com
Frerer, Troy	Proteon	twf@monk.proteon.com
Garcia-Luna, J. J.	SRI	garcia@sri.com
Gerich, Elise	Merit/NSFNET	epg@merit.edu

Greifner, Mike	DCEC	greifner@edn-vax.arpa
Gross, Martin	DCEC	martin@edn-unix.arpa
Gross, Phill	MITRE	gross@gateway.mitre.org
Hahler, Thomas L.	Intermetrics	hahler@inmet.inmet.com
Hahn, Jack	U of MD	hahn@umdc.umd.edu
Hain, Tony	LLNL	hain@nmfecc.arpa
Hastings, Gene	PSC	hastings@morgul.psc.edu
Hedrick, Charles	Rutgers University	hedrick@aramis.rutgers.edu
Hitchcock, Dan	DOE	hitchcock%b.mfenet@mfenet.arpa
Hobby, Russell	UC-Davis	rdhobby@ucdavis.edu
Hooper, Bill	MITRE	hooper@gateway.mitre.org
Jacobsen, Ole	ACE	ole@csli.stanford.edu
Jacobson, Van	LBL	van@lbl-csam.arpa
Karels, Mike	UC Berkeley	karels@ucbvax.berkeley.edu
Kramer, Michael	NYNEX	mike@nynexst.com
Kunis, Gary	Boeing	kunis@nwboel.boeing.com
LaBarre, Lee	MITRE	cel@mitrebedford.arpa
Lekashman, John	NASA/NAS	lekash@orville.nas.nasa.gov
Lepp (Gardner), Marianne	BBNCC	mgardner@park-street.bbn.com
Levy, Stuart	MN Supercomputer Ctr	slevy@uc.msc.umn.edu
Little, Mike	M/A-COM	little@macom.arpa
Lottor, Mark	SRI NIC	mkl@sri-nic.arpa
Lougheed, Kirk	Cisco Systems	lougheed@cisco.com
Mamakos, Louis	Univ of MD	louie@trantor.umd.edu
Mankin, Allison	MITRE	mankin@gateway.mitre.org
Mathis, Matt	PSC	mathis@faraday.ece.cmu.edu
McCloghrie, Keith	TWG	kzm@twg.arpa
Medin, Milo	NASA/NAS	medin@ames-titan.arpa
Melohn, Bill	Sun Microsystems	mehohn@sun.com
Mills, Dave	U of Del	mills@udel.edu
Mockapetris, Paul	USC/ISI	pvm@venera.isi.edu
Morris, Don	NCAR	morris@windom.ucar.edu
Moy, John	Proteon	jmoy@monk.proteon.com
Nakassis, Tassos	NBS	nakassis@icst-ecf.arpa
Natalie, Ron	Rutgers Univ	ron@rutgers.edu
Nitzan, Rebecca	LLNL	nitzan@nmfecc.arpa
Partridge, Craig	BBNCC	craig@nnsc.nsf.net
Perkins, Drew	CMU	ddp@andrew.cmu.edu
Petry, Mike	Univ of MD	petry@trantor.umd.edu
Poh, Susan	IBM/SID	poh@ibm.com
Prindeville, Philip	McGill Univ	philipp@cs.mcgill.ca
Pullen, Mark	DARPA	pullen@vax.darpa.mil
Rehkter, Jacob	IBM	yakov@ibm.com
Reichlen, Gladys	MITRE	reichlen@gateway.mitre.org
Reilly, Brendan	TFI	reilly@wharton.upenn.edu

Rochlis, Jon	MIT	jon@athena.mit.edu
Rock, Mary	MITRE	gateway.mitre.org
Rodriguez, Jose M.	UNISYS	jose@kauai.msl.unisys.com
Rokitansky, Carl-Herb.	DFVLR, West Germany	roki@isia.edu
Rowlett, Tom	DOE	————
Sanford, Dave	ARINC	————
Satz, Greg	Cisco	satz@mathom.cisco.com
Schiller, Jeff	MIT	jis@bitsy.mit.edu
Schofield, Bruce	DCEC	schofield@edn-unix.arpa
Showalter, Jim	DCEC	gamma@edn-unix.arpa
Singh, Aditya	Nynex S&T	singh@nynexst.com
Slattery, Terry	USNA	tcs@usna.mil
Staudt, Dave	NSF	dstaudt@note.nsf.gov
St. Johns, Michael	USAF	stjohns@sri-nic.arpa
Stone, Geoff	Network Sys. Corp.	stone@orville.nas.nasa.gov
Su, Zaw-Sing	SRI	zsu@tcsa.ista.sri.edu
Swanson, John	Unisys	swanson@mcl.unisys.com
Thompson, Kevin	MITRE	gateway.mitre.org
Tontonoz, James	DCA/DCEC	tontonoz@edn-unix.arpa
Tribble, Dave	MITRE	gateway.mitre.org
Tsuchiya, Paul	MITRE	tsuchiya@gateway.mitre.org
Van Bellegham, Dan	NSF	dvanbell@note.nsf.gov
Veach, Ross	Univ, of Illinois	rrv@uxc.cso.uiuc.edu
Waldbusser, Steve	CMU	waldbusser@andrew.cmv.edu
Waldfogel, Asher	Wellfleet Comm	linus!wellflt!awaldfog
Wasley, David	UCBerkeley/BARRNET	dlw@berkeley.edu
Whitaker, Anne	MITRE	whitaker@gateway.mitre.org
Wolff, Stephen	NSF	steve@note.nsf.gov
Woodburn, Robert	M/A-COM	woody@macom.arpa
Zahavi, Ron	MITRE	rzahavi@gateway.mitre.org

3 FINAL AGENDA

WEDNESDAY, June 15

9:00 Opening Plenary (Introductions and local arrangements)

9:30 Working Group Morning Session

- Host Requirements (Braden, ISI)
- SNMP (Rose, TWG)
- Open Routing (Callon, BBN and Hinden, BBN)
- Open SPF IGP (Petry, UMD and Moy, Proteon)
- TELNET Linemode (Dave Borman, Cray)

12:00 Lunch

1:30 Working Group Afternoon Session

- Host Requirements (Braden, ISI)
- Landmark Routing (Tsuchiya, MITRE)
- Short-Term Routing (Hedrick, Rutgers)
- Open INOC (Case, UTK)

5:00 Recess

THURSDAY, June 16

9:00 Opening Plenary

9:15 Working Group Session

- Management Information Base (Partridge, BBN)
- Authentication (Schiller, MIT)
- PDN Routing (Rokitanski, DFVLR)
- Performance and Congestion Control (Mankin, MITRE)
- Domains (Mamakos, UMD)

11:30 Lunch

1:00 Network Status Reports

- Arpanet/Internet Report (Brescia/Lepp (Gardner), BBN)
- Status of the New NSFnet (Braun, UMich/Rekhter, IBM)
- FRICC Initiatives (Bostwick, DOE/Pullen, DARPA/Wolff, NSF)
- Canadian Research Networking (Curley, NRC of Canada)
- Switched Multi-Megabit Data Service (SMDS) (Kramer & Singh, NYNEX)

5:00 Recess

FRIDAY, June 17

9:00 Working Group Reports and Discussion

12:00 Lunch

1:30 Technical Presentations

- TCP Performance and Other Unconfirmed Rumors (Van Jacobson, LBL)
- Bellringing, Clock Punching and Gongferming (Mills, UDel)
- Cray TCP Performance, An Update (Borman, Cray)
- Issues in Canadian Networking (Prindeville, McGill)

4:45 Concluding Plenary Remarks

5:00 Adjourn

4 NETWORK STATUS REPORTS

As has become tradition, the afternoon of the second conference day was reserved for status reports from the various networks.

4.1 Status of the NSFNET (Braun, UMich/Rekhter, IBM)

Hans-Werner Braun treated the plenary group to a slide-show of computer room views of the Ann Arbor Nodal Switching Subsystem (one node of the new NSFNET backbone). A surprising amount of equipment fits into those small cabinets.

He reported that the all the nodes were up and running, with the complete cutover still due to occur July 1. A bug discovered in IP TTL was the only glitch. Six regional sites were doing EGP simultaneously with the NSS and the old backbone, and the NSS EGP appears to be in good shape. Network monitoring data from the backbone will be shared with the regionals, to allow good cooperative management.

Jakob Rekhter reported some initial performance measurements of the backbone. Pings stopping once at all the nodes (using source routes?) had 170-385 millisecond maximums. Unmodified 4.3 FTP attained 24-47Kb/second transfer rates.

These figures were viewed by the IETF members as rather unsatisfactory, given that this is with minimal or no background traffic. Rekhter pointed out that these measurement cases had seven hops, whereas the routing worst case in the backbone normally is 3 hops. It is possible as well that some undetected routing bugs contributed to the high delays. It takes 40-50 milliseconds for a packet of the same size as the pings to go cross-country on the raw MCI links, not passing through any NSS. And it is known that the delay contributed by each IDNX component is 4.5 ms. independent of packet size. There is not saturation of the T-1 links in the ping and FTP experiments, so better network-level performance is expected with tuning.

4.2 FRICC Initiatives (Bostwick, DOE/ Pullen, DARPA/Wolff, NSF)

Bill Bostwick (DOE) reported on the purpose and composition of the Federal Research Internet Coordinating Committee (FRICC). The FRICC is composed of five government agencies that currently fund network research, network operations, or both. There may be other agencies joining the consortium in the future, but, at present, the members are the National Science Foundation (NSF), the Department of Energy (DOE) , the Defense Advanced Research Projects Agency (DARPA), the National Space and Aeronautics Administration (NASA), and Health and Human Services (HHS).

The FRICC is an outgrowth of the recommendations of the congressionally chartered Federal Coordinating Committee on Science, Engineering, and Technology (FCCSET). FCCSET was formed with the charter to make recommendations to Congress on funding science and technology. One of the recommendations was to establish a national computer network (or internet) for the use of scientific researchers. The five agencies of the FRICC were all part of the original study, and acting with the support of the FCCSET, formed the FRICC to begin acting immediately and cooperatively on these recommendations.

Bostwick discussed several of the FRICC initiatives, which included establishing the Research Internet Backbone (RIB) and pursuing efforts in Directory Services and Policy-based Routing.

Mark Pullen (DARPA) discussed the transition of the Arpanet into the Defense Research Internet (DRI), using a portion of the RIB bandwidth to achieve the first step of the transition.

The transition of the ARPANET to the DRI is a three-phased operation:

1) transfer of leased lines to T-1 coast-to-coast lines forming the RIB;
2) upgrade to T3 backbone capacity; and
3) start of research into the configuration and use of a network providing gigabit/second throughput.

Phase 1 has a further breakdown, relating to the effect of these changes on current ARPANET users: first DARPA will cut out the most expensive links in the ARPANET, beginning with the cross-country terrestrial links. Next the RIB part of the ARPANET will go in. ARPANET users will be encouraged to find alternatives for the support of their interconnection. LosNettos on the West Coast is a model for such alternatives.

The DRI will suport C^3 requirements and the DARPA sponsored gigabit research. Subscribers to the DRI must be approved by DARPA with emphasis on supporting federal agencies. The FRICC will provide a paper in the near future on the criteria for policy-based routing, which is necessary due to the inter-agency character of the DRI.

4.3 BBN Report (Lepp (Gardner)/Brescia, BBN)

Marianne Lepp talked about the reduction of ARPANET internal links due to the DRI steps. These reductions come at a time when the ARPANET is experiencing a sharp rise in transit traffic.

BBN consulted with DARPA on how to reduce DARPA's payments for the ARPANET operations, and came up with the idea of using the existing Wideband satellite network capacity in place of the terrestrial cross-country links, which are very expensive. Three Wideband channels are replacing the trunks as a temporary measure until the RIB is in place.

A Wideband to PSN interface was developed. Previously the connection has been through a gateway, while this new interface is an encapsulation. An issue was that the PSN parameters were tuned for fixed-speed links. The Wideband is variable speed and has other characteristics that may cause perceptible changes in performance after the change. Lepp stated that the best cross-country transit would be around 600 ms. Finally, she noted that, since Wideband has always been experimental, BBN may have some trouble keeping the lines up at first.

Lepp also reported on the status of the hardware for the Research Internet Backbone (RIB) to ARPANET connections that are scheduled. Nothing had been procured yet, but BBN had proposed a T-1 product called the T/500. This is manufactured by a company, NSS, bought by BBN a year ago. ARPANET users should not expect that T-1 service is coming their way. Parallel 56K channels are planned for the indefinite future.

Mike Brescia continued the BBN status report, but presented his piece on Friday morning. He announced that SATNET would be dismantled in July. Its shared channels are to be replaced by dedicated 64K satellite or fiber channels. UCL, one of the major SATNET sites, is to join the NSFNET. The replacement connections for another of the major sites, RSRE, are more complex, as it will become a defense network switching center.

The removal of the ARPANET cross-country links resulted in there being one less mailbridge. The Butterfly mailbridges would be installed in July, and tested in August. The cutover from the LSI-11s would be announced in September. They are to be removed in December. The Butterfly EGP service is scheduled to start by December. Brescia restated that these schedules are changeable and that the EGP transition would be advertised on EGP-PEOPLE.

Responding to a couple of questions, Brescia explained the new Autonomous System number issue again. The Butterflies will not be AS 1, and code that assumes this is the AS number of the core should be fixed. EGP mandates the peer with the lower AS number to run as active, so there is a rule to follow to handle the new core's AS number of 60. He shared the current plans as to filtering by the mailbridges: filtering is not to be turned on right away, but after a grace period, inbound TELNET from the ARPANET (that is ARPA users logging in to MILNET systems) will be filtered out.

4.4 Canadian Research Networking (Curley, NRC)

John Curley of the Canadian National Research Council spoke on the status of Canada's Internet. The Canadian Research Network resembles the NSFNET in topology and protocols, and plans also to transition to OSI. There exists a "coast-to-coast" Canadian fiber backbone and proposals from telecommunications companies are being sought.

4.5 Switched Multi-Megabit Data Service (SMDS) (Cramer/Singh, NYNEX)

SMDS is a joint effort by BELLCORE and the RBOC's to provide a uniform, data service in the early 1990's. It is intended to offer LAN-like performance over Metropolitan areas. SMDS is a service concept, not a new technology, for high speed, public, packet-switched data communications.

A feature of the SMDS is the Subscriber Network Interface (SNI). A goal of SNI is to contribute to end-to-end low delay which will be achieved by a new 3 layer access protocol (not equivalent to OSI layering). Layer 3 will provide a network service with variable length PDU's of $<$ 8K bytes. Layer 2 provides framing for PDU's with error detection not correction. Layer 1 provides the physical transmission interface. Initially this will be a DS3 interface, with a possible future switch to SONET. SONET is a BELLCORE proposed optical and electrical interface with a 50 megabit/second baseline. SONET is open-ended, but so far has been defined to a top speed of 1.2Gb. One SNI will use the ISDN numbering scheme and can have multiple addresses. Provisions for multicasting, closed communities, and costing by access class are currently being studied.

NYNEX is also working on a proposal for IEEE 802.6 for MAN access in a public network. The proposed standard is the Distributed Queue Dual Bus. It can support both isochronous (fixed bandwidth and delay, video) and non-isochronous (data) service simultaneously. Singh gave a stimulating description of this shared media access switching method.

5 WORKING GROUP REPORTS

The first day and a half of the IETF meeting was divided into three half day sessions, during which individual working groups gathered. Of the currently active IETF Working Groups, thirteen met in Annapolis and fourteen report on their activities. They are listed below with their spokesperson.

- Internet Management Information Base (MIB) (Craig Partridge, BBN)
- Authentication (Jeff Schiller, MIT)
- Domains (Louie Mamakos, UMD)
- CMIP-based Net Management (NETMAN) (Lee LaBarre, MITRE)
- Internet Host Requirements (Bob Braden, ISI)
- Landmark Routing (Paul Tsuchiya, MITRE)
- Open SPF-based IGP (Mike Petry, UMD)
- Open Systems Internet Operations Center (Jeff Case, UTK)
- Open Systems Routing (Ross Callon, BBN)
- PDN Routing (Carl-Herbert Rokitanski, DFVLR)
- Performance and Congestion Control (Allison Mankin, MITRE)
- Short-term Routing (Chuck Hedrick, Rutgers)
- SNMP Extensions (Marshall Rose, TWG)
- TELNET Linemode (David Borman, Cray)

5.1 Internet MIB

Craig Partridge reported on the success of his group in producing an initial Internet Management Information Base (MIB). He siad that there remains some unresolved areas about the MIB, such as how to divide it below IP, but that the group has decided to reserve judgement until some experience is collected with the draft MIB.

It is important to point out that the definition of a 'MIB' is meant to be independent of the Network Management protocol which would carry the information. In other words, the MIB defined by Craig's group will be used by both SNMP and CMOT. He stressed that work on the second generation MIB for TCP-IP would begin in the Fall.

5.2 Authentication

Jeff Schiller restated the goals of the group to be two-fold: 1) to specify the format that authentication information could be in network/internet protocols, to specify an appropriate crypto checksum, and not to specify procedures for verification; 2) to demonstrate a proof-of-concept which could include the use of SNMP, SPF IGP, and NTP plus authentication.

The group's objective is to produce an RFC which will identify the format, cost benefits of authentication, and guidelines for including authentication in protocol implementations. A second RFC will discuss key distribution using Kerberos as the example security service.

Jeff concluded by stating the group's focus is on end-to-end security not just network security. Dave Mills asked that authentication be considered in the network layer so as to verify source quench and redirects.

Phill Gross asked the group to consider only unclassified information exchange.

5.3 Domains

The work of this group is winding down. A document, "PHASE II OF THE MILNET DOMAN NAME IMPLEMENTATION" will be distributed shortly as a DDN Management Bulletin. It addresses the MILNET naming transition, and includes the specification of name resolution hosts for MILNET. All MILNET, ARPANET and Internet hosts must be registered in a domain other than ".ARPA".

It was recommended that the host name and address information be updated daily and that hosts use retry rates exceeding 5 minutes. It was allowed that the domain system still had problems with the user interface as well as basic functionality within the service itself. Notably, the new root name servers seem to be working well. Score one success here.

The group discussed using the domain name system to perform Network Name —> Network Number, and Network Number —> Network Name lookups. It would also be desirable to have the mechanism for doing this work with subnets. A note describing the issues in more detail, and soliciting input should appear on either the TCP-IP or NAMEDROPPERS mailing list.

The group recommended that the Host Requirements working group REQUIRE that host software implement the domain name system. It would be up to the user of the machine to choose to use it or not. The somewhat modified adage "like minds travel in the same packet" was verified, as they chose to adopt this view independently.

Something to think about: For a given domain name, should the server randomly order records of the same type (i.e. more than one NS record)?

Yet another, hopefully the last, draft of the Responsible Person resource record IDEA was circulated. This will be prepared as IDEA0008-01 available soon. Comments will be welcomed.

5.4 CMIP-based Net Management (NETMAN)

The major emphasis of the NETMAN group at this time is focused on the demonstration for the September TCP/IP Interoperability Conference. The demonstration will consist of monitoring a LAN with workstation traffic. In addition the group hopes to provide draft Implementation Agreements at the conference.

Further development is awaiting the achievement of DIS status for CMIS/CMIP.

Phill Gross commented that the CMIP balloting was complete and that a number of NO votes with comments were recorded. It was his opinion that without major changes, the comments could be addressed and that the NO votes would be changed to YES votes on the next ballot. [Note: DIS status was voted by ISO in August.]

Issues that remain are authentication, access control and event management.

5.5 Internet Host Requirements

The goal of this group is to produce an RFC by December 1988 and thereafter dissolve the group. However, a section on TELNET must still be written, and a contributor would be most welcome.

5.6 Landmark Routing (Tsuchiya, MITRE)

The first meeting of this working group covered the major features of LM and Assured Destination Binding in a seminar-like fashion.

5.7 Open SPF IGP

Reported by John Moy, Proteon.

The main purpose of this group's meeting was to review the first part of the OSPFIGP specification. That document had been distributed to all interested IETF members approximately two weeks before the meeting.

The following general comments on the specification were received:

- There needs to be support for networks having no broadcast capabilities. An X.25 network is a good example. We decided to treat these similarly to the way broadcast networks are treated in the spec: there will be a Designated Router for the network and it will generate the network's link state advertisement. There needs to be some additional configuration information in order to discover the Designated Router on these networks. For more details see below.
  - The protocol should run directly over IP, instead of over UDP. A checksum field was therefore added to the general OSPFIGP header.
  - There should be a capability to authenticate all packet exchanges. (Currently we are just authenticating the creation of adjacencies). For this reason the authentication field has been added to the general protocol header.
  - We were not sure that it was a good idea for the protocol to specify the use of IP multicast. For the moment we are going to specify local-wire broadcast instead. We will discuss our particular concerns in this area with Steve Deering.
- There should be an appendix to the specification concerning metric assignment strategies. The protocol specifies only a dimensionless metric. This could be configured by the AS administrators to mean weighted hop count, delay, bandwidth, etc. A discussion of metric assignments should include how the protocol's equal cost multipath would be affected.

A rough, incomplete draft of the rest of the specification was then handed out at the meeting. This draft included detailed packet formats. After some discussion the following changes were made to the detailed parts of the specification:

- We were worried about the size of AS external links advertisements. OSPFIGP relies on IP fragmentation to deal with large packets, and we want to avoid large packets as much as possible. Also, when a single AS external route changes, we would like to not have to reflood all routes. So we made each AS external route into its own link state advertisement. This is very similar to the EGP-3 strategy. Note that in each hop of the flooding procedure, multiple link state advertisements may be contained in a single Link State Update Packet.
- A change was made to the Designated Router selection on broadcast networks. We want to avoid changing Designated Router as much as possible, so when a router's interface first comes up, it will wait some period of time to see whether or not a Designated Router has already been selected for that network. If so, the new router will defer to that Designated Router, regardless of who has higher priority. This does mean that it will sometimes be hard to predict who will be the Designated Router on a network.
- On networks with no broadcast capability (like X.25) the Designated Router will be selected as follows. A small number of routers on the network will be configured as eligible to become Designated Router. Each one of these routers will have a configured list of all routers attached to the network. Each router in this list that is eligible for becoming Designated Router will also have a configured Router Priority.
- If a router (that is eligible to become Designated Router) loses all adjacencies to routers of higher priority, it will become Designated Router, establishing adjacencies with all routers of lower priority. These adjacencies will be broken if a higher priority router is again heard from.
- It would be helpful if the lower level protocols on these networks provide an indication that a neighboring router has become unreachable.
- All references to the Dijkstra algorithm will be moved to an appendix. The references to Dijkstra in the main body of the specification should refer instead to the building of a shortest path tree. Many different algorithms can be used to build such a tree.
- Subnet masks were added to the Hello packets. This will aid in the detection of inconsistent configurations.
- There was quite a bit of discussion concerning authentication. The authentication issues dealt with were:
  - An authentication type field was added to the protocol header so that multiple authentication schemes can be supported.
  - One of the authentication schemes should be a simple password. This will keep new routers from be indiscriminately turned on — they will have to discover the simple password first.
  - There should be an option for no authentication.
  - There was no need seen for replay protection, and so time synchronization was not seen as an issue.

- There is a strong desire to separate the authentication procedure as much as possible for the operation of the routing protocol. It was proposed that to implement a Kerberos-like scheme, a router would act only as a host until it has obtained the session key from the Kerberos server. This would mean that the distribution of session keys would fan out from the Kerberos server.
- There was alot of discussion on how to use a Kerberos-like scheme. A couple of packet types would need to be added to distribute session keys. There is also a desire to have a single key per network, and this does not seem to fit perfectly with the Kerberos model for a session.

A first draft of the complete OSPFIGP specification should be available by late July. At that time we would like to have a meeting to discuss prototyping the protocol.

5.8 Open Systems Internet Operation Center

Reported by Jeff Case, UTK.

The charge of the OINOC WG is to:

Define

* duties and activities of NOC personnel
  - questions they need to answer
  - problems they need to solve
  - reports they need to generate
* information they need to do the above
* data they need to produce the information above
* sources of the data above
* tools and applications needed to acquire and process these data
* architectures for the development of these tools and applications including the structural relationships between NOCs and NOC-NOC communications

The OINOC Working Group compliments other working groups in the general area of network management in that it focuses on goals and architectural issues while leaving to other groups more focused efforts such as the development of protocols.

Tasks:

1. Define a model for combining elements of network monitoring and control into a total system.
   (a) Define the roles of an Internet Network Operations Center (INOC)
      i) a point of controlled access to information including protecting monitored entities from excessive/redundant requests
      ii) provide proxy services for non-IP entities
      iii) provide appropriate levels of security for data integrity and authorization of access
   (b) provide mechanisms for exchange of information across administrative domains

2. Database
    (a) define needs
    (b) mechanisms for information storage and retrieval
3. Information required to do network management *
    (a) MIB
    (b) input from other WGs like congestion/control, host req
4. Define application needs
    (a) real time status monitoring
    (b) fine-fighting
    (c) report generation
    (d) standard application interface

* This task was reassigned to the MIB Working Group as a result of the IAB actions outlined in RFC 1052.

There have been several important events related to network management since the San Diego IETF meeting. They include:

* March 21 IAB Meeting

  SNMP until CMIP
  MIB WG Formed
  SNMP WG Formed

* MIB WG Products

  IDEA 0023-00 SMI
  IDEA 0024-00 MIB

* SNMP WG Products

  IDEA 0011-01 SNMP

* SNMP/MIB/SMI Implementation Activities

* CMIP Failure (so far) to reach DIS

* Network management issues related to new NSFNET backbone

* Many new OINOC WG meeting attendees

The pressing issues before the group include:

1. We need to form a consensus as to what is "Network Mangement"?
2. We need to agree how to accomplish network management/monitoring, especially fault management, in the context of multiple administrative domains and redundant/distributed NOCs. This is in light of the following:
   (a) network managers tend to be conservative in what they are willing to make available
   (b) need a balance between usability and security
   3. The relationship between policy based routing and network management aspects of NOC-NOC communications

5.9 Open Systems Routing

A requirements document for interautonomous systems routing service is finished. Functional specification of the protoco has begun. Probably the biggest concern is how to do "external routing constraints" (also known as policy routing). The problem can be divided into 1) the trust model, 2) access control, and 3) information hiding. Also impacting the functional specification is the issue of scale. We have no working experience for the worldwide internetwork that is envisioned; the EGP model is just about to fail at the size the DoD Internet has reached.

The group discussed a few existing specifications, such as the Dissimilar Gateway Protocol and Landmark Routing. There are significant overlapping and compatible ideas, but it is unclear yet "how to put it all together into an elephant that acually walks around and does things."

Overall, the ways to do interautonomous system routing will probably require fairly drastic measures. First, it needs a new addressing format that allows variable length and is more or less hierarchical, but does not have one top-level. Second, it needs link state routing that allows information hiding, in other words, a new approach to link state routing. Finally, it will call for entry point routing, where some entity in the first domain is responsible for pulling together the whole route. IP and ISO IP Source Routes will not hold enough information for this. Route setup will probably be the answer. All of these measures are overkill for many routing situations, so a simple forwarding paradigm will coexist for those.

5.10 PDN Routing

A significant feature of the PDN routing scheme is "cluster addressing" among clusters of public data networks in Europe. Another feature of the PDN Internet which this working group will be addressing is a transport bridge between TCP and TP0.

A paper on cluster addressing will be submitted to ICCC 88 and to the IETF as a new IDEA. The content will include X.121 address resolution protocol, reverse charging for internal calls, and routing metrics.

5.11 Performance and Congestion Control

Reported by Anne Whitaker (MITRE).

At our meeting on June 16, the performance working group took our first pass through a rough draft of our paper. Seven authors contributed sections. The paper is currently titled "Internet Performance Recommendations." It will describe work to-date in protocol enhancements and in improved protocol implementations that have resulted in internet system performance improvements. There is still a requirement for editorial review, original contributions, and improvement in focus of the document. Work pressures on a number of the group members dictate that it will not be completed until about January 1989.

Our early discussion involved questions about the relationship between the performance work and the development of the MIB. We did not all agree that measurement standards were within the concerns of our paper. However, the current draft has a section on metrics, and it is hoped that network management variables will be developed in coming months that allow performance monitoring.

Van Jacobson (LBL) gave the working group meeting a brief status report on his current Berkeley-based performance work. He has added a diagnostic path via a raw socket, generalizing the calls that access kernel data structures as well as allowing packet logging. He has completely revamped the mbuf system. The diagnostic socket, but probably not the new mbuf code, will be included in the next Berkeley UNIX release.

MITRE then spoke about their extension of Van's TCP instrumentation to a per connection basis and its incorporation into an instrumented host and gateway for congestion control experiments.

The group had a lengthy discussion about gateway time-to-live decrements, queuing strategies and packet dropping criteria. We got hints from Van about gateway interactions with his TCP interactions, such as that the random dropping he is leaning toward should not wait for a queue to form. Aside from Time-to-Live, where the paper can make a strong recommendation that it be a hop count, we need to do a great deal more work on our gateway performance recommendations.

Attendees were: Art Berggreen (ACC), Coleman Blake (MITRE), David Borman (Cray Research), Ross Callon (BBN), Michael Collins, Troy Frerer (Proteon), Bill Hooper (MITRE), Van Jacobson (LBL), Allison Mankin (MITRE), Rebecca Nitzan, Jose Rodriguez (UNISYS), Bruce J. Schofield (DCEC), Geof Stone (NASA), James Tontonoz (DCEC), Anne Whitaker (MITRE)

5.12 Short-term Routing

Reported by Chuck Hedrick (Rutgers).

This was a somewhat odd period for this group to meet. Our primary goal is to look at the overall operation of the Internet, specifically at the interconnections between its major pieces. At the moment this means primarily the links between DDN, the NSFNET backbone, and the regionals. Since the NSFNET was about to change over to a new technology, detailed examinations of the old NSFNET backbone and its connections with the regionals did not seem overly useful.

One person observed that routes within the ARPANET core seemed unstable. In particular, metrics seemed to be changing in ways that did not look appropriate. It did not seem likely that this was a new phenomenon. Problems with GGP are well-known. What was perhaps more interesting is that MIT has a proposed workaround. Rather than taking metric information from the core at face value, they attempt to pick gateways based on what is known about the way the core works. There are two main rules:

1) in order to stabilize routing, and also to avoid unnecessary transcontinental hops, the nearest of the 3 core gateways is given priority in routing. That is, they declare BBN as their primary EGP gateway. If they hear of routes from both the primary gateway and another, they prefer the route that they heard about from the primary.

2) in order to avoid the extra hop problem, they use a heuristic. When extra hop happens, it always follows a very specific form: one of the EGP core gateways claims that the route to a network is through another one of the core gateways, whereas another core gateway has the correct route. So if

   - two different EGP peers propose different routes to a given network,
   - one of those routes is via one of their EGP peers
   - the other route is via a gateway that is not one of their peers the route that is not via an EGP peer is preferred. (They peer with all 3 EGP core gateways.)

Note that these rules cause them to ignore the EGP metrics.

Another issue involving ARPANET routing was announcement of routes for NSFNET sites into the ARPANET core. Until recently there were only a few NSFNET/ARPANET gateways. In order to provide redundancy, it made sense for a gateway to announce all of the NSFNET networks. There are now enough that it makes sense to be selective. Rutgers is a typical example. We have a T1 connection to JvNC. JvNC has an IMP. Obviously we'd like to people to use JvNC to talk to us, and not PSC's already overloaded gateway. It's not even clear that we need a backup. If jvnca.csc.org is down, we can't get anywhere outside Rutgers anyway. I believe everyone at the meeting agrees that we need to reengineer the NSFNET/ARPANET connections, more or less as follows: Campus network managers should have control over who announces them to the ARPANET. In most cases, a single gateway will do so, or one gateway and a backup. Depending upon whether the network has its own connection to the ARPANET, the metric may be 0, 3, or a primary with 0 and a backup with 3. All gateway managers should make sure that they are announcing only networks that should be announced. I think in most cases this will be handled by negotiations among the regionals, since in general the regionals will know what their members want done. (If not, they should find out!) Obviously we don't want every gateway manager to have to talk directly to every campus served by NSFNET. At the meeting the feeling was that the default should now be that a given network is announced only by the nearest ARPANET gateway, unless the campus network manager has authorized a backup. It's not entirely clear what we do to implement this sort of thing, but most of the gateway managers were at the meeting, and I trust that this message will reach the rest.

We are still getting a lot of reports of connections closing, in situations where the site is still reachable. Most people believe that this is due to brief transient unreachable conditions. Unfortunately, there is no one thing that can be done to fix this. The most important is that TCP implementations must not close connections when they receive

ICMP unreachables. This is a common bug, unfortunately. System managers to whom robustness matters should check their implementation to see whether it has this problem. If so, get your vendor to fix it. However there are a number of other things that can be done to reduce this problem. Here are some examples of known problems:

- gated routing transitions between EGP and RIP routes can leave a brief period during which the route is unreachable
- Proteon routers with routing turned off (all but one line down?) apparently do not issue redirects. Proteon may not be alone in this. Boxes with only one operational interface tend to think they are not gateways. Since they are not, it might be inappropriate for them to issue ICMP's. There can be similar problems during booting. When a gateway comes up, before it has received routing information from all of its neighbors, there are a lot of places that it thinks are unreachable. It may tend to issue unreachable messages during this time. I heard a complaint about this from a Proteon user. I verified that cisco routers do the same. I believe the correct behavior is that for the first N minutes of uptime, a gateway should not issue unreachables. Frankly, with things the way they are now, I'd prefer it if systems stopped issuing unreachables entirely.
- when a route goes down, it may time out at different times different places, so a gateway that knows it is down may sent an ICMP unreachable back through a path that a nearer gateway thinks is still up. (Sounds like a routing implementation that doesn't do flash update.)
- hosts may not be able to change from a failed gateway to one that is still up. 4.2 had only the most limited ability to do this. 4.3 is better, but even in 4.3 it is not clear what to do with UDP. Apparently by the end of this year, Sun's NFS will do the right thing, so if your most critical UDP application is NFS (which is the case for us), you'll be in fairly good shape. A complete solution probably also requires the ICMP where's-my-gateway/here's-your-gateway messages, which are just now being put into an RFC or IDEA.

In general, IP implementations still do not deal with routing changes smoothly enough to prevent connections from breaking. If you expect to avoid breaking connections, you must make sure that your vendor is following all the developments in 4.3 technology, or doing equivalent work, and you should follow the progress of the ICMP gateway messages.

The rest of the meeting was a review of the implications of the changeover to the new IBM/Merit NSFNET backbone. There was no one from IBM or Merit present at the meeting. (This will not be allowed to happen again.) However a number of sites reviewed their configurations in detail, and came up with a list of issues to pursue with the IBM/Merit folks. They were collared at a later meeting, which became a de facto extension of the short-term routing group.

The new NSFNET backbone has as a goal doing policy-based routing. What this means at the moment is that any network manager can choose which gateways will handle routing for his networks. The implementors chose to combine this with hierarchical routing. They are using the autonomous system number to provide the second level in the hierarchy. This leads to a system that uses AS numbers in a manner that is not entirely consistent with their normal interpretation. The decision to do that seems to have been based on the fact that EGP was the only practical way to get routing information from the regionals, and the AS number was the only thing they could get out of it that could be coerced into providing second level information. At any rate, the primary routing within the NSFNET backbone is an SPF algorithm, where the objects being routed are AS numbers. There are static tables indicating which network numbers should be handled through which AS's. For example, Rutgers could declare that 128.6 should be handled through JvNC if possible, and next through PSC. Each gateway into the backbone has a set of AS's that it can get to. In addition to the normal routing packets that keep track of routing among the AS's, each gateway advertises which networks it can get to (through which AS, I believe). Routing works as follows: to get to a network, find the first AS number in its list that shows that network as reachable. Then use the best route to that AS number (i.e. using the SPF routing, take the best route to the nearest exit gateway in that AS). Round-robin alternation is done among equally good routes.

Note that these algorithms are going to tend to require you to use more AS numbers than you might otherwise need. For example, suppose a regional has two connections to the backbone. If they use the same AS number for each, problems can ensue. If a network is reachable via any of those gateways, it will be shown as reachable through that AS. Traffic for that network will then go to the nearest exit gateway for the AS. If the network is accessible only through some of those gateways, some traffic will go into a black hole. Thus separate AS numbers should be used for each gateway. There were also questions about how the IBM routers would deal with situations where they were talking to several routers at the same site. It is fairly common that the IBM router will be put on an Ethernet with several other routers. Quite often one of those routers will be closer to a given destination network than the others. You'd like the IBM router to pick the right one. You would not like to have to use a different AS number for each router at your site. As a result of this meeting, IBM agreed that they would pay attention to the metrics at a single location. These metrics will not be passed on to the rest of the backbone. But once their routing algorithm has sent a packet to a given exit gateway, it will then send the packet to a directly-connected router with the lowest metric for the destination network.

Present at the meeting were (subject to possible misreadings of their handwriting):

Gene Hastings, Pitt. Supercomputer Center, hastings@morgul.psc.edu
Geof Stone, Network Systems Corp, stone@orville.nas.nasa.gov
Don Morris, NCAR/UCAR, morris@windom.ucar.edu
Kirk Lougheed, cisco Systems, lougheed@cisco.com
Dale Kinkelson, Univ. of Nebraska and Midnet, dmf@fergvax.unl.edu
Ross Veach, Univ of Illinois, rrv@uxc.cso.uiuc.edu
Allan Fischer, US Naval Academy, allan@usna.mil
Stuart Levy, Minn. Supercomputer Center, slevy@uc.msc.umn.edu
Gary Kunis, NorthWestNet, kuns@nwboe1.boeing.com

Matt Mathis, Pitt. Supercomputer Center, mathis@faraday,ece.cmu.edu
Susan Poh, IBM/SID, Poh@ibm.com
David Wasley, Univ. of Calif, Berkeley, dlw@berkeley.edu
Jeff Schiller, MIT, jis@bitsy.mit.edu
Mark Fedor, Nysernet, fedor@nisc.nyser.net
Gary Almes, Rice and Sesquinet, almes@rice.edu

5.13 SNMP Extensions

IDEA011 will be updated so as to align with MIB criteria, to meet the short-term network management needs of the Internet. Currently, there are two server implementations of SNMP, one at University of Tennessee-Knoxville, and one at NYSER, Inc. The group plans to submit the IDEA011 as an RFC and disband when the latter state is achieved. [Ed. That has now happened.]

5.14 TELNET Linemode

David Borman restated the group's goal, which is not to deal with "local emulation of remote terminals", but rather to enhance the TELNET option set. The group discussed the relationship between IDEA00016 and RFC 1053, and reached the conclusion that the RFC must be labelled experimental and not pursued. The RFC author, Steve Levy, was in agreement.

6 TECHNICAL PRESENTATIONS

6.1 TCP Performance and Other Unconfirmed Rumors (Van Jacobson, LBL)

In order to develop the gateway side of his congestion control algorithms, Van stated, he is now in the process of developing some "wild theories" about why ping data during network congestion can show packet delays varying from 20 to 200 seconds. Where do packets stay for so long, and what circumstances bring about this kind of variability?

Van analyzed a data set from a DECNET routing problem he found at LBL some time ago. A phenomenon of self-organization shown by these data may be a start towards the necessary theory.

DECNET routers broadcast a Hello message every 15 seconds and a routing update every 120 seconds. Using a variant of his program tcptrace, Van recorded the times at which the routing broadcasts of a group of DECNET routers occurred. He started the data collection following a power failure. The assumption was that this crash should have randomized the updates, because each router would come up slowly and become able to function again at a different time. However, Van's graphs show that by three hours after the crash the routers were very close to synchronization, and by six hours after, they were astonishingly synchronized (see the vugraphs).

[Editor's Note: it is difficult to do justice to the clarity of Van's presentation, but here goes...] The explanation of the phenomenon begins with drifts of the individual router's interval timers. An individual routing process wakes up after an interval, processes incoming updates, broadcasts its own update, resets its interval timer, and goes back to sleep. It resets its interval timer from the time when it completes all its processing.

From the random time at which each router starts following the crash, a combination of events begins to clump the routing broadcasts together. At first, all that is needed is any slight drift caused by operating system (scheduling) or Ethernet access noise. This eventually causes two routers' processes to overlap in the following way: one process awakens while another process is doing its broadcast. Incoming traffic (i.e. the broadcast from the earlier-starting of the overlapping processes) has priority in the DECNET protocol, so the later-starting process (A) delays its broadcast by the amount of overlap. This delay is preserved in A's new interval timer calculation. Meanwhile, B is shifted too, because it stays awake to process the update from A. The resulting close synchronization of A and B will persist because of their interaction each time they awaken.

The synchronized routing processes awaken and broadcast at lower frequency than the unsynchronized processes. Any noise or accident that increases the timer interval of as yet unsynchronized processes tends to move them toward overlapping with those that have become synchronized. Someone in the audience described this as as making "a black hole which then goes off hunting". Van also called it an "aggregation exponential." Further discussion identified the fact that it takes a while for the aggregation of 40 millisecond process runs to occur, since they have 120 second intervals to take place in, but once aggregation starts, it happens faster and faster. This acceleration was labelled "a potential well."

Noel Chiappa asked if the DECNET nodes were homogeneous (all DEC routers). Two of them were Proteon gateways doing the DECNET protocol. Noel said this strengthened the data set, since Proteons are very different from DECs in their operating system characteristics, such as interrupt priorities and process scheduling.

Chuck Hedrick asked if the problem would be eliminated if the interval timers were calculated from the rising edge instead of the falling. It would slow down the synchronization, but not stop it. Changing the timer parameters also just prolongs the process. Next the discussion dealt with the idea that the routers could have varied interval parameters that are not multiples of each other. This would be hard to implement with the coarse clock resolution available from the typical systems.

The randomization features of RIP would help. It was pointed out that a similar study is infeasible for RIP, since there would not be one Ethernet on which all the routers' updates could be observed. However, Mike Karels said he did not see evidence of aggregation of the timers during his tests of the RIP randomization code.

The rest of Van's talk described theories relating the self-synchronization of the DECNET routers to IP in the Internet. He has identified several roads to synchronization of IP packets passing through gateways. One is that TCP connections produce IP packets at fairly regular intervals, reflecting the round trip time and the use of acknowledgements to clock out packets. Several TCP connections passing through a gateway interact in the frequency of their interpacket intervals: when any packet gets queued, it is shoved back in time, and nothing can restore the original interval of the packets.

An important extra impetus to "clumping" of Internet packets is the way a reliable subnet such as the ARPANET, by not reordering, keeps once-together packets from a connection together at later queues. It is this factor that possibly changes a linear, and not too persistent, effect into an exponential effect that is hard to break up. The tendency of the reliable subnet to keep together packets that have started out together also accounts for the observation that connections keeping large windows full get very few source quenches. They gain a "slot" because of the advantage the system gives to their clumped-together packets.

It appeared likely to Van from reasoning like the above, that the ARPANET behaved like a token ring. Gateway queue data Van collected met this expectation. It showed that packets clocked out on a TCP connection in response to a round of acknowledgements wait together in the gateway queue, then leave the gateway together. They move in this burst at bottleneck bandwidth. As a result of these unintended send bursts, the next acknowledgments also come in a burst. These bursty acknowledgments are a problem for Van's TCP send algorithms, as they lead to a too-high sending rate,

Overall, synchronization effects by gateways and the ARPANET cause non-uniform utilization of links and other network resources. Are there ways to regain some of the lost efficiency? Van said the he would approach this, with the help of a mathematician post-doc, by modelling the problem using diffusion equations, such as the Smoluchowski equation. Diffusion equations include constants corresponding to how far in time packets can shift randomly and how much they interact. With a combination of modelling and gateway measurement, Van hoped it would be possible to find rules for how fast Internet systems aggregate and gateway algorithms to combat the effects of aggregation.

6.2 Bellringing, Clock Punching, and Gongferming (Dave Mills, UDEL)

Dave Mills emphasized the importance of accurate time keeping across the Internet. He described his most recent work on the Network Time Protocol (NTP), which is currently accomplishing such synchronized timekeeping.

He presented some very nice graphs of the NTP accuracy over several different hosts. One type of graph of 'offset vs delay', which he termed the 'wedge diagram' (see slides), turned out to have a secondary function. It was able to show packets traversing different paths through the Internet.

He also suggested that there must be many sources of accurate time. There are 6 services now serving 20-40 clients having about 10 millisecond precision.

6.3 Cray TCP Performance (Borman, Cray Research)

David Borman updated the IETF on the results he presented in San Diego (the top rate then was 150Mb). His recent kernel modifications of TCP in Cray's BSD UNIX-based UNICOS operating system have resulted in phenomenal TCP throughput, 175 Megabytes per second! The network medium for these throughputs is the Cray-proprietary 800 Mb HSX channel, connecting two Cray. It can also be used to connect Crays with high-speed graphics output devices. In software loopback, Dave reported that the top rate now is 247Mb.

The improvements from San Diego were obtained by incorporating Van Jacobson's slow-start algorithms. Van's high speed improvements using header prediction are still to come.

6.4 Issues in Canadian Networking (Prindiville, McGill)

Philip Prindeville described Canadian interests in networking, which are planned to involve universities, high technology firms, R&D facilities and government. He discussed a proposal he has drafted for the Canadian National Research Council's network procurement and how it might fit with the US Internet.

7 PRESENTATION SLIDES

This section contains the slides for the following presentations made at the June 15-17, 1988 IETF meeting:

- Tenth Internet Engineering Task Force (Gross, MITRE)
- IETF NETMAN (LaBarre, MITRE)
- Arpanet/Internet Report (Hinden/Lepp (Gardner), BBN)
- Status of the New NSFnet (Braun, UMich/Rekhter, IBM)
- FRICC Initiatives (Wolff, NSF/Bostwick, DOE)
- Canadian Research Networking (Curley, NRC of Canada)
- Switched Multi-Megabit Data Service (SMDS) (Singh, NYNEX)
- TCP Performance and Other Unconfirmed Rumors (Van Jacobson, LBL)
- Cray TCP Performance, An Update (Borman, Cray)
- Issues in Canadian Networking (Prindeville, McGill)
- Bellringing, Clock Punching and Gongferming (Mills, UDel)
- Switched Multi-megabit Data Service (Kramer, NYNEX)
- Performance and Congestion (Mankin, MITRE)
- Domains (Mamakos, UMD)
- SNMP Extensions (Rose, TWG)

7.1 Tenth Internet Engineering Task Force—Gross, MITRE

The Tenth Internet Engineering Task Force (USNA, Annapolis)

Phill Gross

MITRE

Introduction

- Local Arrangements -- Terry Slattery
- Proceedings
- IDEAS
- Working Group Groundrules
- Internet Problem Description Forms

IDEAS

- Internet Design, Engineering and Analysis Series
- Meant as IETF document management tool
- 'Pre-RFC' and Stand-alone D, E, or A Note
- NOT meant to be used as standards!

Working Group Groundrules

- On the Up side: People are starting to take the IETF seriously
- On the Down side: People are starting to take the IETF seriously
- Rules are easy:
  - Fill out IETF Form 2
  - Submit Working Group Reports for the Proceedings

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

IETF Form 2

1) Statement of the charter and goal of the group

2) Expected duration of the group

3) Is membership to the WG open or closed?

4) List of members.

5) Mailing lists for the group? (open or closed?)

6) When was your last meeting?

7) Accomplishments To Date

Internet Problem Description Forms

- This is IETF Form 1
- These might even be used

IETF Internet Problem Description

1) **Name of Submission:**

2) **Category of Submission** (Network Engineering, Protocol Engineering, or Research):

3) **Problem Description:**

4) **Suggested Approach** (optional):

5) **Cost** (optional, but tied to Suggested Approach, if given):

6) **Time Frame** (Short-term, Mid-term, or Long-term):

7) **Responsible Group** (i.e., funding authority):

8) **Date of Submission:**

7.2 Arpanet/Internet Report—Hinden/Lepp (Gardner), BBN

ARPANET & INTERNET

INTERNET GROWTH

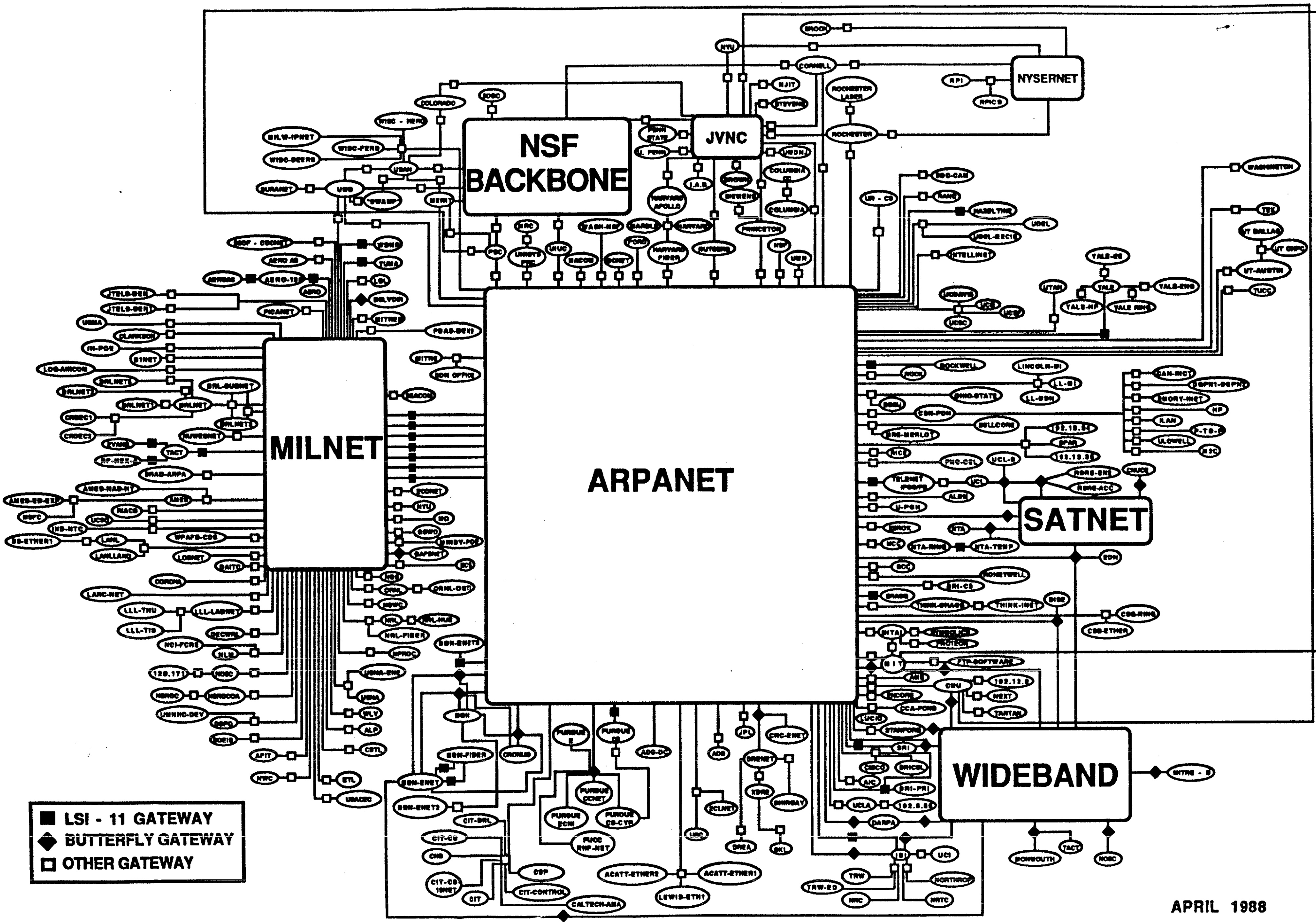
ARPANET TRUNKING

GATEWAY INSTALLATION PLANS (BBN BUTTERFLY)

- SATNET REPLACEMENT WITH MODEM LINES
- ARPANET-MILNET GATEWAY ("MAILBRIDGE") REPLACEMENT

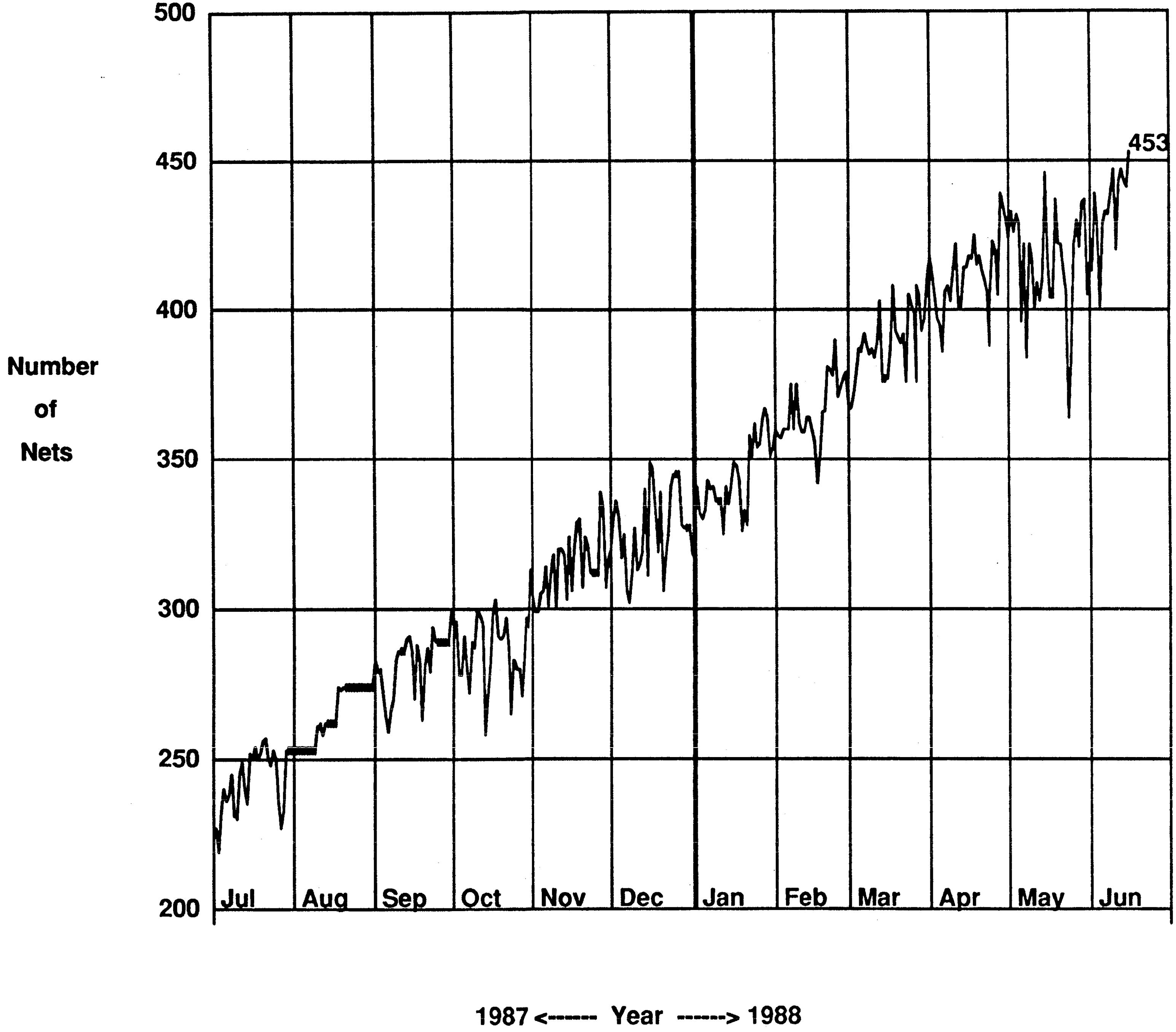
GATEWAY DEVELOPMENT

- PACKET RADIO IPR (1822)
- INTERNET MULTICAST
- X.25 CERTIFICATION BY TELENET
  - 2 "VAN" GATEWAYS TO BE INSTALLED
  - "PDN ROUTING"

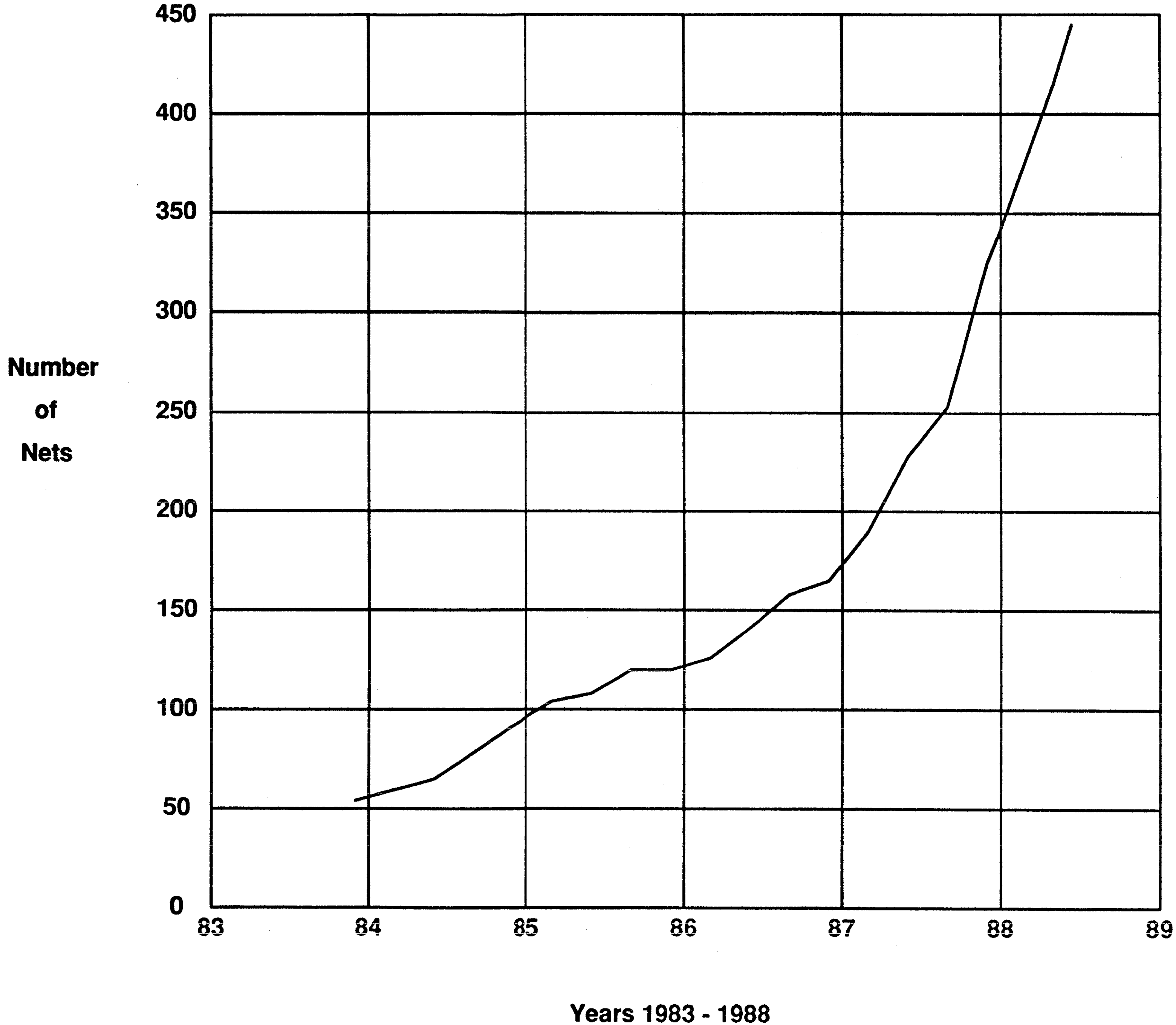


BBN Communications Corporation

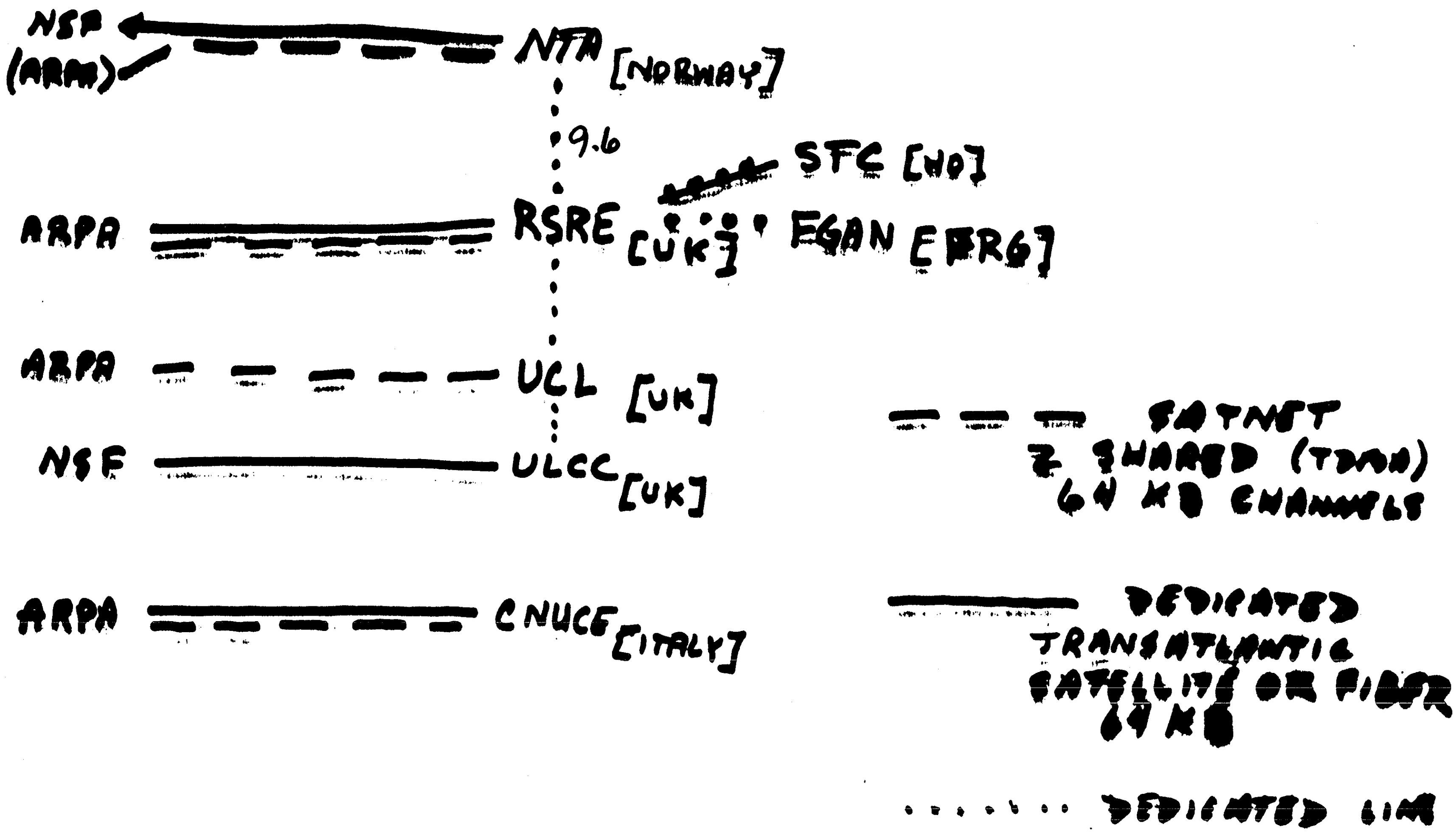
Internet Growth in Networks



Internet Growth in Networks

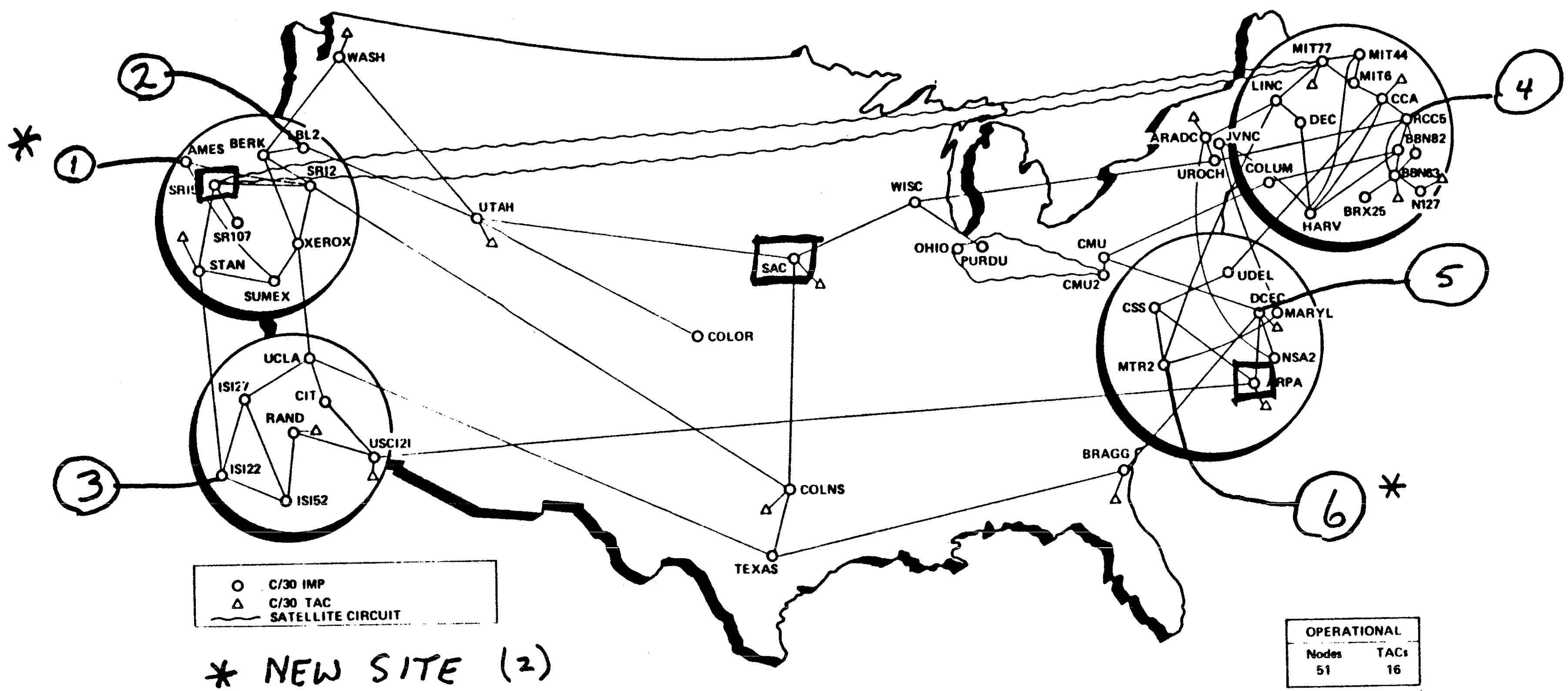


SATNET REPLACEMENT



ARPANET Geographic Map, 30 April 1988

6 BUTTERFLY MAILBRIDGE GATEWAYS



* NEW SITE (2)

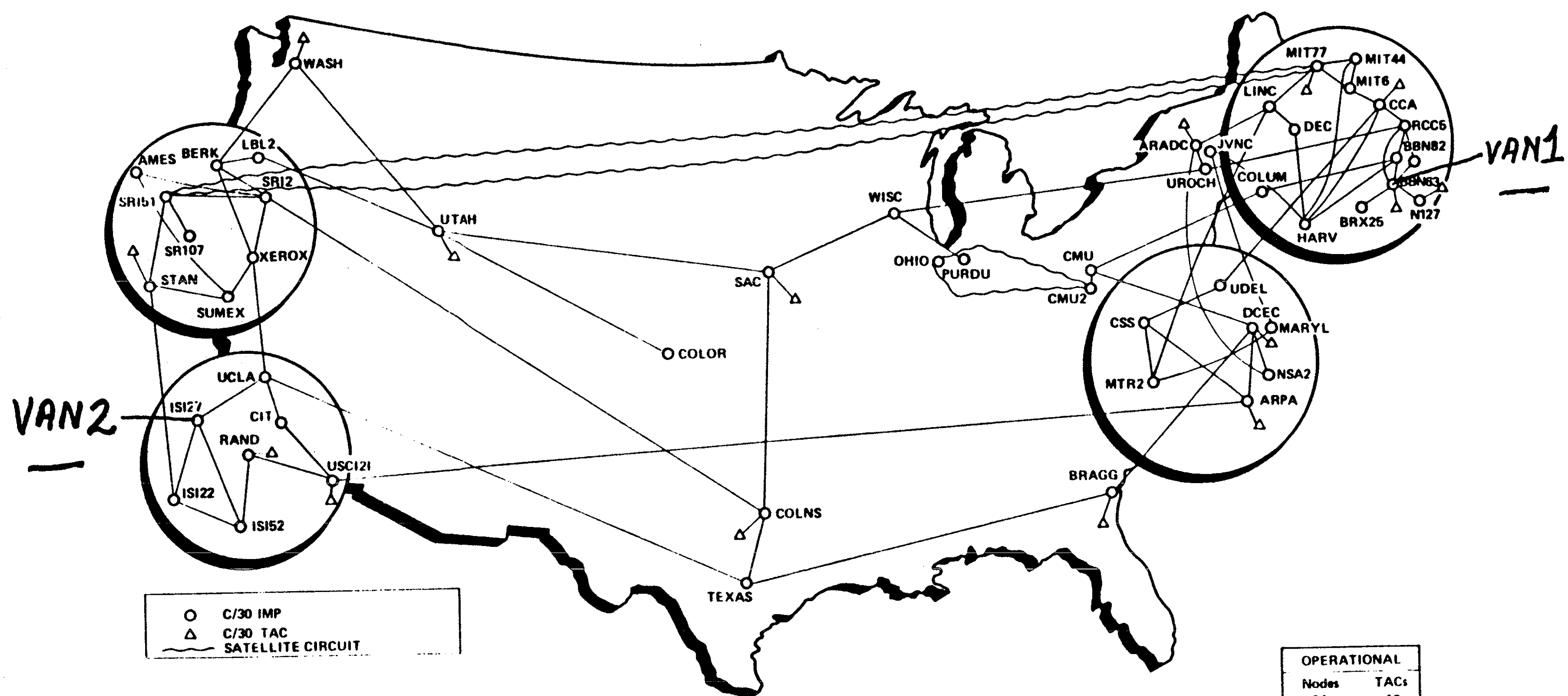
[o] OLD SITE (3)

MAILBRIDGE REPLACEMENT
LSI-11 OUT / BUTTERFLY IN

- DELIVER, INSTALL	JULY 88
- AVAILABLE FOR TESTING	AUG 88
- ANNOUNCE CUTOVER MAILBRIDGE (ARPA-MIL) EGP SERVER	SEPT 88
- PERIPHERAL LSI-11'S BECOME 'STUB'	OCT 88
- REMOVE LSI-11 MAILBRIDGES	DEC 88
- REMOVE LSI-11 EGP SERVERS	DEC 88

ARPANET Geographic Map, 30 April 1988

2 BUTTERFLY VAN GATEWAYS



7.3 Status of the New NSFnet—Braun, UMich/Rekhter, IBM

NSFNET Backbone Routing

Jacob Rekhter (yakov@IBM.COM)
T.J. Watson Research Center
IBM Corp

Routing Traffic

① SPF Hello Packets

Hello – I-H-U every 10 seconds
consumed bandwidth 112 bits/sec

② LSP Traffic

Router Link PDU }
ES PDU } ⟺ Sequence Number PDU

Average exchange rate – 2 per minute
(with 17 NSS's up)

Router link PDU < 100 bytes
Sequence Number PDU < 100 bytes
ES PDU – large (~ number of ES)

Formal Model

$BB = \{R_i\}$

$R_i = \{A\$_j\}$ $\quad \exists\, m,n \quad R_m \cap R_n \neq \emptyset$

$net_j = (A\$_1, A\$_2, \ldots, A\$_n)$ $\quad \forall k,i \quad A\$_i \,!= A\$_k$
$k \,!= i$
$1 \leq k,i \leq n$

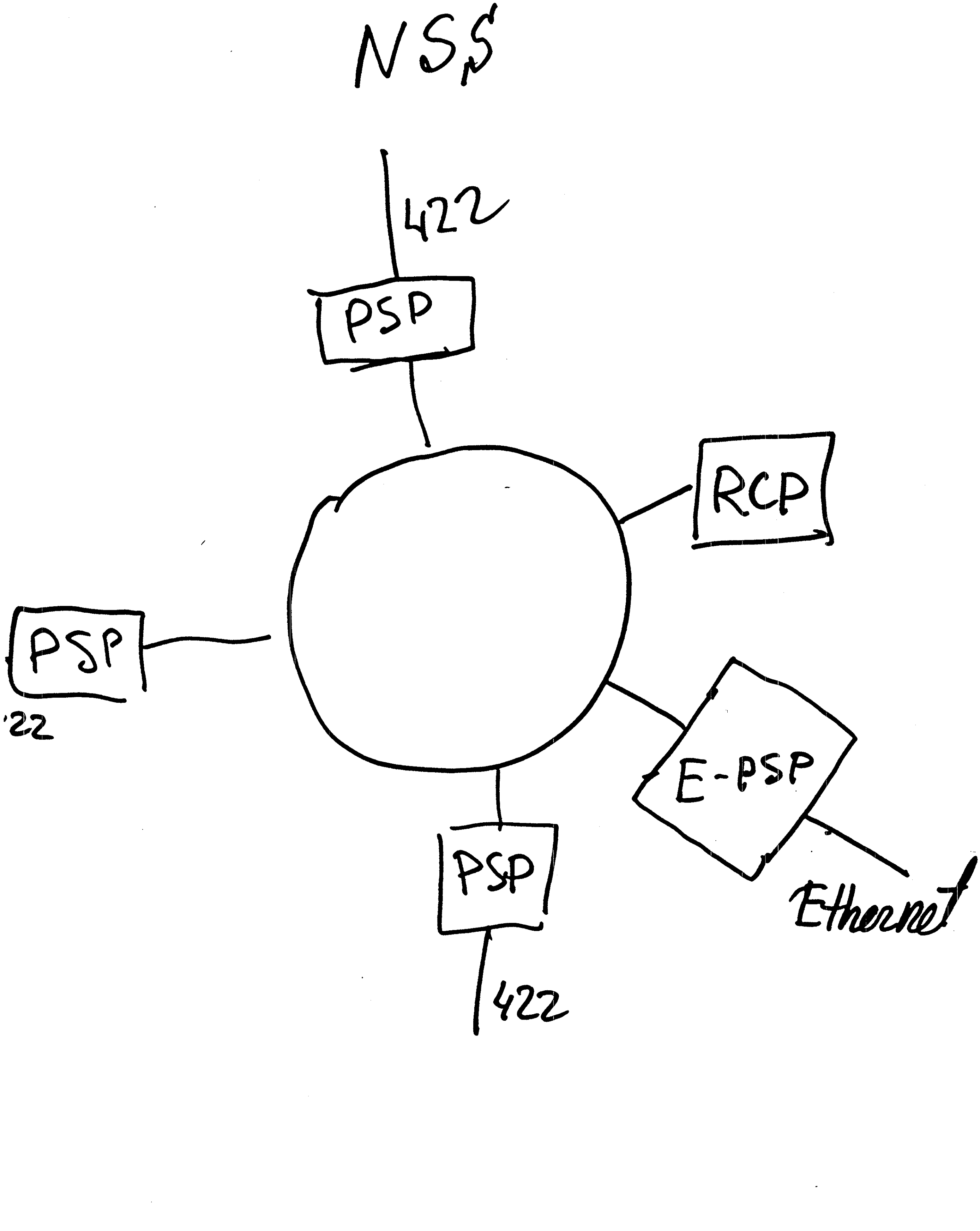
$BB \times BB \rightarrow BB\ metric$

<u>Algorithm:</u>

```
/* given net_A, R_s find Rexit */
cost = ∞
Rexit = ?
net_A = (A$_1, A$_2, ... A$_k)
BB = {R_1, R_2 ... R_n}
for i=1 to k {
    for j=1 to n {
        if A$_i ∉ R_j
            continue
        if cost < BB metric[s, j]
            continue
        cost = BB metric[s, j]
        Rexit = R_j
    }
    if cost < ∞
    break
}
```

Gateway Policy Routing Group

```
GatewayPolicyRoutingGroup {
  AS in   sequence of integers     -- valid in AS's
  valid_AS sequence of {
      net  IpAddress               -- particular network
      AS   integer                 -- member of ...
      metric  integer              -- primary, secondary...
  }
  AS out  sequence of integer      -- valid out AS's
}
```


NSS
422
PSP
RCP
PSP
'22
E-PSP
Ethernet
PSP
422

IS-IS management

RouterCounterGroup {

RouterLinksPDUin	counter
RouterLinksPDUout	counter
ESLinksPDUin	counter
ESLinksPDUout	counter
SequenceNumberPDUin	counter
SequenceNumberPDUout	counter
CorruptedPDU	counter
IS-ES Hello in	counter
IS-ES Hello out	counter
IS-IS Hello in	counter
IS-IS Hello out	counter

}

IS-IS management

```
Router Group {
   MaximumRouterLSPGenerationInterval integer
  MaximumEndSystemLSPGenerationInterval integer
 MinimumLSPTransmissionInterval integer
 MinimumLSPGenerationInterval integer
}

Neighbor Group {
   HelloTimer    integer
   Cost          integer
   State         {On(1), Off(0)}
  NeighborStatusChange   event
  Neighbor address   IpAddress
  Hold-time          Integer
}
```

Current Parameters

- Hello Interval (SPF Hello)
  10 seconds
- Hold Time (SPF)
  40 seconds
- Minimum LSP Generation Interval
  30 seconds
- Minimum LSP Transmission Interval
  5 seconds
- Maximum LSP Generation Time
  15 minutes

Software

- Total 17,000 lines of source code

  ~7,000 lines EGP

  ~7,000 lines SPF

  ~3,000 lines "route distribution to PSN

NSFNET Backbone Routing

- Introduced new layer of hierarchical routing – route to AS

- First implementation (partial) of ANSI IS-IS protocol

- Rudimentary Policy Based Routing based on EGP

```
Script started on Wed Jun 15 19:35:07 1988
rcp-3-1:/usr/nss: netstat -nr
Routing tables
Destination          Gateway              Flags    Refcnt Use        Interface
129.140.1            129.140.3.13         UG       0      0          lan0
129.140.2            129.140.3.13         UG       0      3224       lan0
129.140.3            129.140.3.1          U        6      364716     lan0
129.140.5            129.140.3.13         UG       0      0          lan0
129.140.6            129.140.3.13         UG       0      0          lan0
129.140.7            129.140.3.13         UG       0      0          lan0
129.140.8            129.140.3.13         UG       0      0          lan0
129.140.9            129.140.3.13         UG       0      0          lan0
129.140.10           129.140.3.13         UG       0      0          lan0
129.140.11           129.140.3.13         UG       0      0          lan0
129.140.13           129.140.3.13         UG       0      0          lan0
129.140.14           129.140.3.13         UG       0      0          lan0
129.140.15           129.140.3.13         UG       0      0          lan0
129.140.16           129.140.3.13         UG       0      0          lan0
129.140.17           129.140.3.13         UG       0      0          lan0
129.140.45           129.140.3.12         UG       0      0          lan0
129.140.46           129.140.3.11         UG       0      496        lan0
rcp-3-1:/usr/nss:
rcp-3-1:/usr/nss:
rcp-3-1:/usr/nss:
rcp-3-1:/usr/nss:
rcp-3-1:/usr/nss:
```

```
Script started on Wed Jun 15 20:29:45 198
rcp-3-1:/usr/nss: ftp rcp-1-1
Connected to rcp-1-1.
220 rcp-1-1 FTP server (Version 4.108 Wed Jan 20 23:40:05 PST 1988) ready.
Name (rcp-1-1:nss): ibmykt
331 Password required for ibmykt.
Password:
230 User ibmykt logged in.
ftp> bin
200 Type set to I.
ftp> cd /
250 CWD command successful.
ftp> get vmunix
200 PORT command successful.
150 Opening data connection for vmunix (129.140.3.1,1707) (954368 bytes).
226 Transfer complete.
local: vmunix remote: vmunix
954368 bytes received in 40 seconds (24 Kbytes/s)
ftp> 221 Goodbye.
rcp-3-1:/usr/nss: ftp rcp-7-1
Connected to rcp-7-1.
220 rcp-7-1 FTP server (Version 4.108 Wed Jan 20 23:40:05 PST 1988) ready.
Name (rcp-7-1:nss): ibmykt
331 Password required for ibmykt.
Password:
230 User ibmykt logged in.
ftp> bin
200 Type set to I.
ftp> cd /
250 CWD command successful.
ftp> get vmunix
200 PORT command successful.
150 Opening data connection for vmunix (129.140.3.1,1711) (954368 bytes).
226 Transfer complete.
local: vmunix remote: vmunix
954368 bytes received in 70 seconds (13 Kbytes/s)
ftp> 221 Goodbye.
rcp-3-1:/usr/nss:
script done on Wed Jun 15 20:34:08 198
```

```
rcp-3-1:/usr/nss: ping rcp-10-1
PING rcp-10-1: 56 data bytes
64 bytes from 129.140.10.1: icmp_seq=0. time=237. ms
64 bytes from 129.140.10.1: icmp_seq=1. time=237. ms
64 bytes from 129.140.10.1: icmp_seq=2. time=236. ms
?
----rcp-10-1 PING Statistics----
3 packets transmitted, 3 packets received, 0% packet loss
-ound-trip (ms)  min/avg/max = 236/236/237
rcp-3-1:/usr/nss: ping rcp-11-1
PING rcp-11-1: 56 data bytes
64 bytes from 129.140.11.1: icmp_seq=0. time=174. ms
64 bytes from 129.140.11.1: icmp_seq=1. time=174. ms
64 bytes from 129.140.11.1: icmp_seq=2. time=180. ms
?
----rcp-11-1 PING Statistics----
3 packets transmitted, 3 packets received, 0% packet loss
-ound-trip (ms)  min/avg/max = 174/176/180
-cp-3-1:/usr/nss: ping rcp-12-1
PING rcp-12-1: 56 data bytes
?
----rcp-12-1 PING Statistics----
3 packets transmitted, 0 packets received, 100% packet loss
-cp-3-1:/usr/nss: ping rcp-13-1
PING rcp-13-1: 56 data bytes
64 bytes from 129.140.13.1: icmp_seq=0. time=346. ms
64 bytes from 129.140.13.1: icmp_seq=1. time=345. ms
64 bytes from 129.140.13.1: icmp_seq=2. time=345. ms
?
----rcp-13-1 PING Statistics----
3 packets transmitted, 3 packets received, 0% packet loss
ound-trip (ms)  min/avg/max = 345/345/346
cp-3-1:/usr/nss: ping rcp-14-1
ING rcp-14-1: 56 data bytes
4 bytes from 129.140.14.1: icmp_seq=0. time=284. ms
4 bytes from 129.140.14.1: icmp_seq=1. time=285. ms
4 bytes from 129.140.14.1: icmp_seq=2. time=284. ms

---rcp-14-1 PING Statistics----
 packets transmitted, 3 packets received, 0% packet loss
ound-trip (ms)  min/avg/max = 284/284/285
cp-3-1:/usr/nss: ping rcp-15-1
ING rcp-15-1: 56 data bytes
4 bytes from 129.140.15.1: icmp_seq=0. time=186. ms
4 bytes from 129.140.15.1: icmp_seq=1. time=179. ms
4 bytes from 129.140.15.1: icmp_seq=2. time=179. ms

---rcp-15-1 PING Statistics----
 packets transmitted, 3 packets received, 25% packet loss
ound-trip (ms)  min/avg/max = 179/181/186
cp-3-1:/usr/nss: ping rcp-16-1
ING rcp-16-1: 56 data bytes
 bytes from 129.140.16.1: icmp_seq=0. time=190. ms
 bytes from 129.140.16.1: icmp_seq=1. time=190. ms
 bytes from 129.140.16.1: icmp_seq=2. time=196. ms

--rcp-16-1 PING Statistics----
packets transmitted, 3 packets received, 0% packet loss
und-trip (ms)  min/avg/max = 190/192/196
p-3-1:/usr/nss: ping rcp-17-1
NG rcp-17-1: 56 data bytes
 bytes from 129.140.17.1: icmp_seq=0. time=91. ms
```

```
Script started on Wed Jun 15 20:27:16 1988
PING rcp-1-1: 56 data bytes
64 bytes from 129.140.1.1: icmp_seq=0. time=77. ms
64 bytes from 129.140.1.1: icmp_seq=1. time=77. ms
64 bytes from 129.140.1.1: icmp_seq=2. time=77. ms
^C
----rcp-1-1 PING Statistics----
3 packets transmitted, 3 packets received, 0% packet loss
round-trip (ms)  min/avg/max = 77/77/77
rcp-3-1:/usr/nss: ping rcp-2-1
PING rcp-2-1: 56 data bytes
64 bytes from 129.140.2.1: icmp_seq=0. time=34. ms
64 bytes from 129.140.2.1: icmp_seq=1. time=33. ms
64 bytes from 129.140.2.1: icmp_seq=2. time=33. ms
^C
----rcp-2-1 PING Statistics----
3 packets transmitted, 3 packets received, 0% packet loss
round-trip (ms)  min/avg/max = 33/33/34
rcp-3-1:/usr/nss: ping rcp-5-1
PING rcp-5-1: 56 data bytes
64 bytes from 129.140.5.1: icmp_seq=0. time=234. ms
64 bytes from 129.140.5.1: icmp_seq=1. time=234. ms
64 bytes from 129.140.5.1: icmp_seq=2. time=235. ms
64 bytes from 129.140.5.1: icmp_seq=3. time=234. ms
^C
----rcp-5-1 PING Statistics----
4 packets transmitted, 4 packets received, 0% packet loss
round-trip (ms)  min/avg/max = 234/234/235
rcp-3-1:/usr/nss: ping rcp-6-1
PING rcp-6-1: 56 data bytes
64 bytes from 129.140.6.1: icmp_seq=0. time=382. ms
64 bytes from 129.140.6.1: icmp_seq=1. time=382. ms
64 bytes from 129.140.6.1: icmp_seq=2. time=382. ms
^C
----rcp-6-1 PING Statistics----
3 packets transmitted, 3 packets received, 0% packet loss
round-trip (ms)  min/avg/max = 382/382/382
rcp-3-1:/usr/nss: ping rcp-7-1
PING rcp-7-1: 56 data bytes
64 bytes from 129.140.7.1: icmp_seq=0. time=146. ms
64 bytes from 129.140.7.1: icmp_seq=1. time=146. ms
64 bytes from 129.140.7.1: icmp_seq=2. time=146. ms

----rcp-7-1 PING Statistics----
3 packets transmitted, 3 packets received, 0% packet loss
round-trip (ms)  min/avg/max = 146/146/146
rcp-3-1:/usr/nss: ping rcp-8-1
PING rcp-8-1: 56 data bytes
64 bytes from 129.140.8.1: icmp_seq=0. time=132. ms
64 bytes from 129.140.8.1: icmp_seq=1. time=132. ms
64 bytes from 129.140.8.1: icmp_seq=2. time=132. ms

----rcp-8-1 PING Statistics----
3 packets transmitted, 3 packets received, 0% packet loss
round-trip (ms)  min/avg/max = 132/132/132
rcp-3-1:/usr/nss: ping rcp-9-1
PING rcp-9-1: 56 data bytes
64 bytes from 129.140.9.1: icmp_seq=0. time=176. ms
64 bytes from 129.140.9.1: icmp_seq=1. time=177. ms
64 bytes from 129.140.9.1: icmp_seq=2. time=176. ms
```

7.4 FRICC Initiatives—Wolff, NSF/Bostwick, DOE

(Slides unavailable for Preliminary Draft)

7.5 Canadian Research Networking—Curley, NRC of Canada

National Research Council

Canada's national science and technology institution

- has 3000 employees, $400M/yr budget
- performs fundamental and applied research
- develops codes and standards
- maintains national facilities: wind tunnels, wave basins, etc.
- has a technology transfer program
    - Canada Institute for Scientific and Technical Information
    - Industrial Research Assistance Program
- has major links to int'l research community

Relationship to other networks

- NetNorth(BITnet): e-mail and file transfer
  - to universities, some gov't and private sector
  - using low speed lines and restrictive IBM protocols
- CDNnet: provides electronic mail to
  - university/private sector/government
  - using UBC developed X.400 EAN software
- by contrast, NRCnet would
  - allow new functions such as remote computer access
  - serve a large multi-sector community
  - use high speed lines and widely available protocols
  - provide a migration path for NetNorth and CDNnet
  - serve as test bed for new protocol development.

Evidence of demand

- strong positive reaction to NRCnet proposal
- success of NetNorth/CDNnet despite low line speeds and restrictive protocols
- rapid development of regionals – e.g., BCnet, CRIM
- success of US networks NSFnet, NYSERNet
- increasing tendency to link south

Issues: protocols; self sufficiency

- NRCnet is committed to international standards
  - ISO IP will supercede IP over time
  - Both protocols will be supported
  - RSCS, DECnet through encapsulation
- Backbone self-sufficiency
  - Strategic technology needs startup funds
  - User-pay would be phased in over 5 year period
  - Regional networks would be independantly funded and managed

The need for partners

- requirements exist
    - for technical/management resources
    - at campus/regional/nat'l/intnat'l levels
- one five-year scenario shows $23M cost:
    - $8M backbone (5 years)
    - $15 regional/campus (5 years)
    - breakdown: 35% people, 65% commx lines
- want partners to help implement backbone
    - high visibility, low cost, low risk
    - NRC initially prime contractor
    - operated by consortium when self-sustaining
- Productive discussions with
    - Universities: for network support services
    - Industry (Northern T'com, IBM, T'com Canada, etc.)
    - OGD's
    - NetNorth and CDNnet
    - consultant will assess potential industry involvement

Relationship to other federal programs

- NRC's research and technology transfer programs
- Research programs of OGD's – EMR, DOC, DFO, Environment
- Granting councils: NSERC, MRC, SSHRC
- DIST
- Space Agency
- Centres of Excellence

7.6 Switched Multi-Megabit Data Service—(SMDS) Singh, NYNEX

EARLY AVAILABILITY
BROADBAND SWITCHING TECHNOLOGIES
for the SUPPORT of SMDS

Eddie Singh
Broadband Communications and Services Laboratory
NYNEX Advanced Technology Development

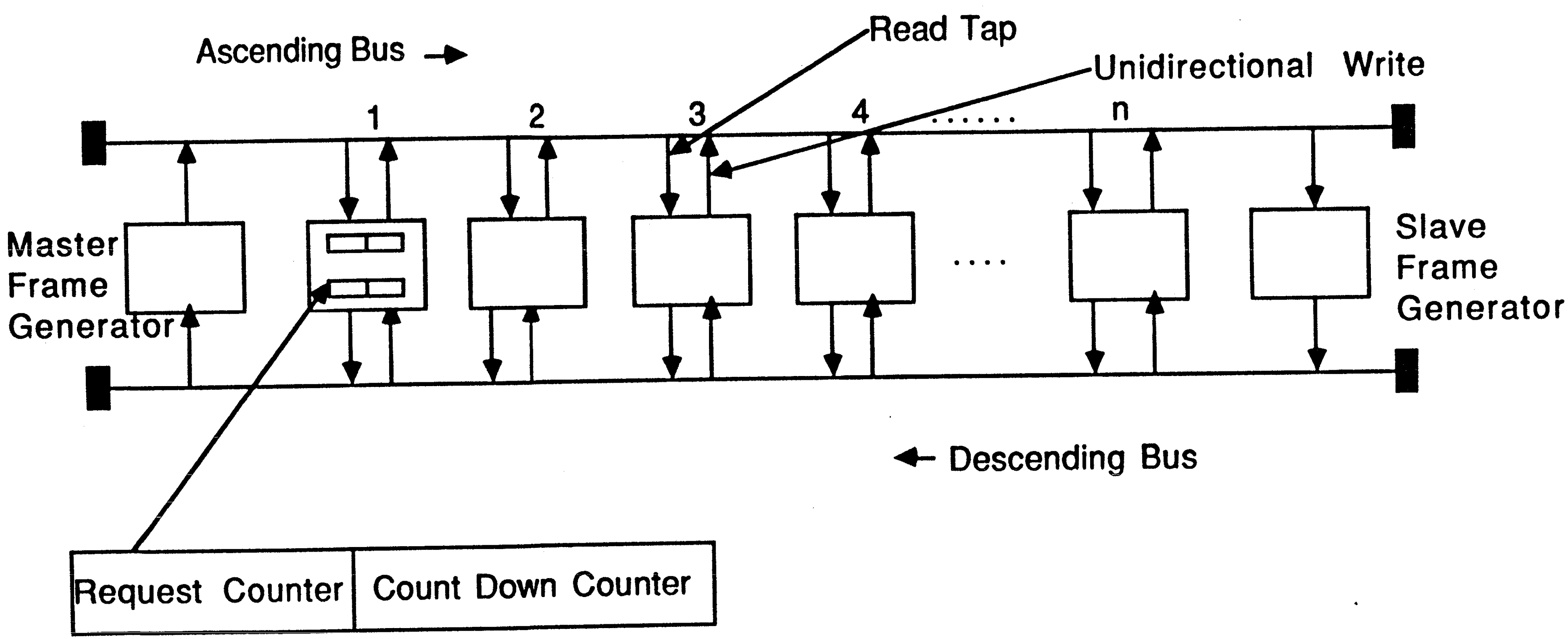
June 16, 1988

[redacted]

DISTRIBUTED QUEUE DUAL BUS (DQDB)

E. Singh, NYNEX-ATD, 3/21/88

DISTRIBUTED QUEUE DUAL BUS

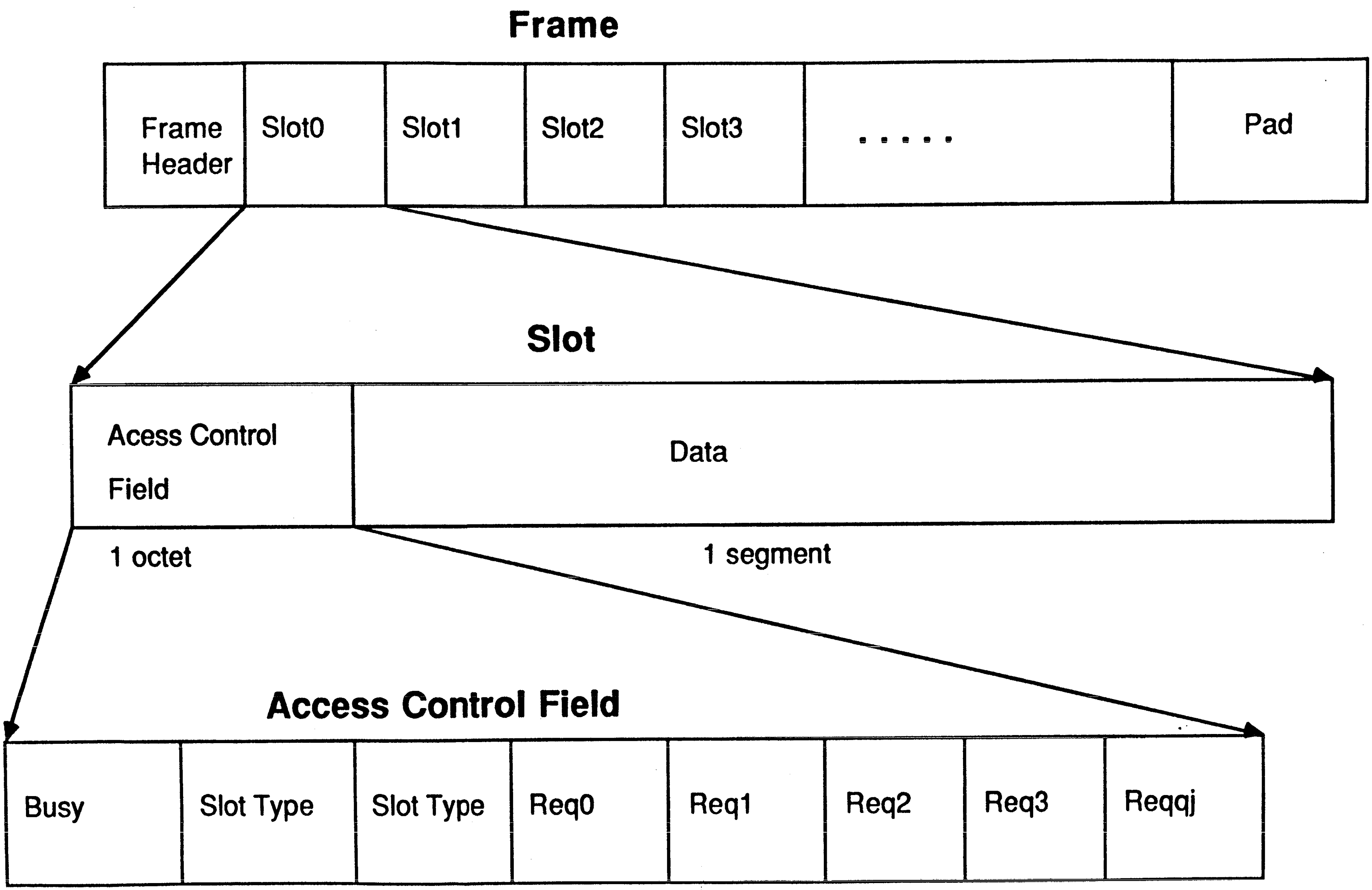


E. Singh, NYNEX-ATD, 3/21/88

OPERATION

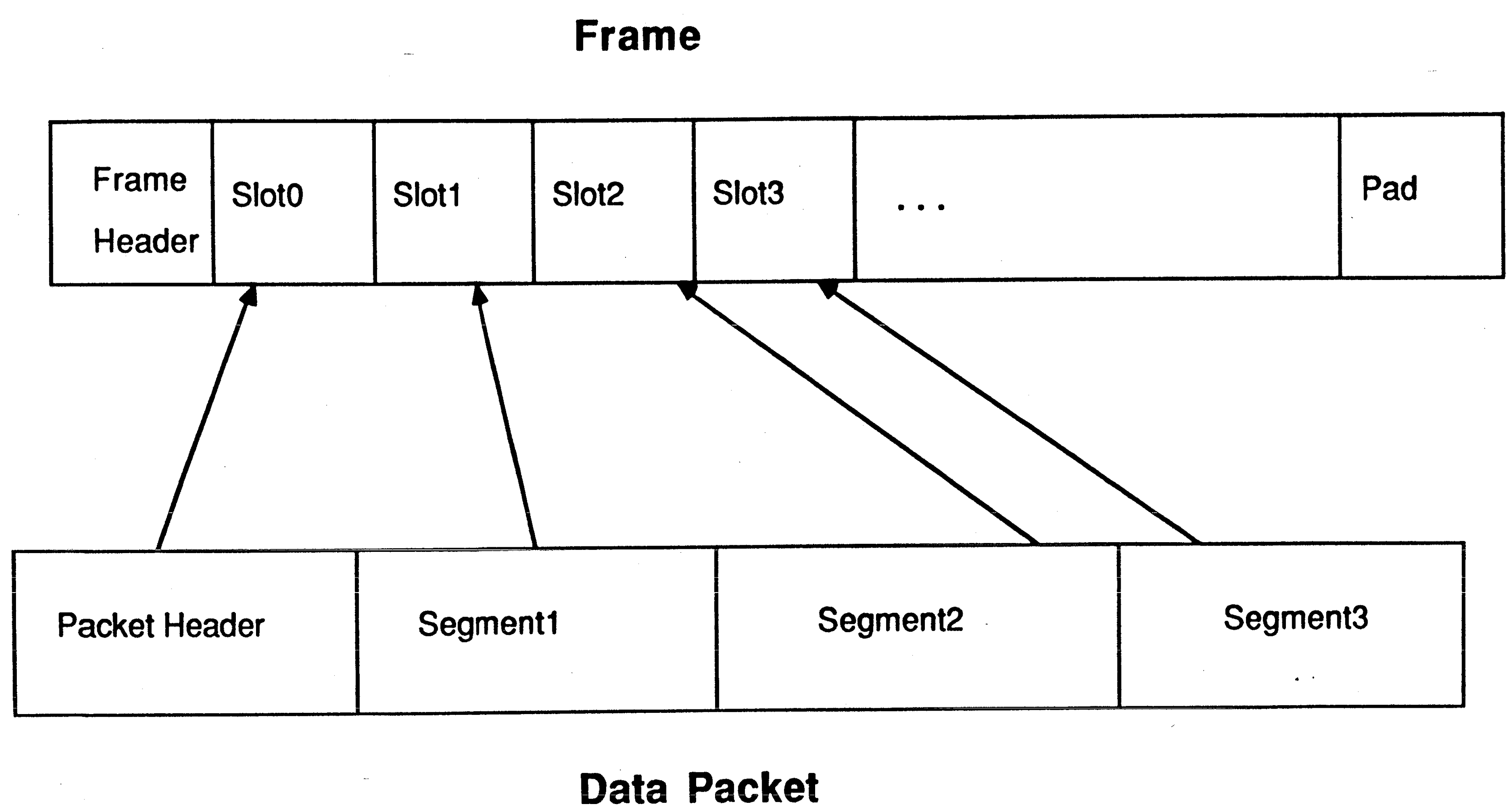
- Two unidirectional buses
- Read Tap, Unidirectional Write connections
- Slotted frames every 125 microseconds
- Nodes reserve slots
- Bandwidth access by Distributed Queueing Protocol
    - Counters maintained at each node

Dual Bus Slot Format (Non-Isochronous)



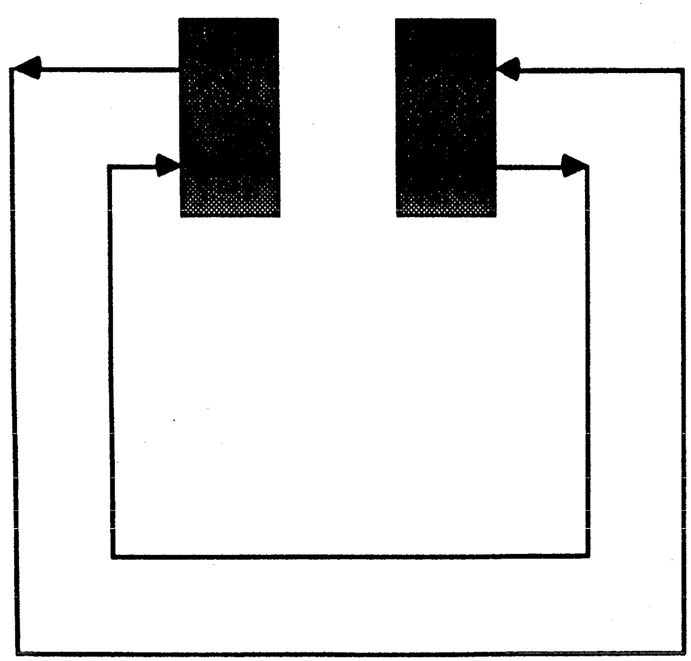
E. Singh, NYNEX-ATD, 3/21/88

Dual Bus Data Packet Transfer

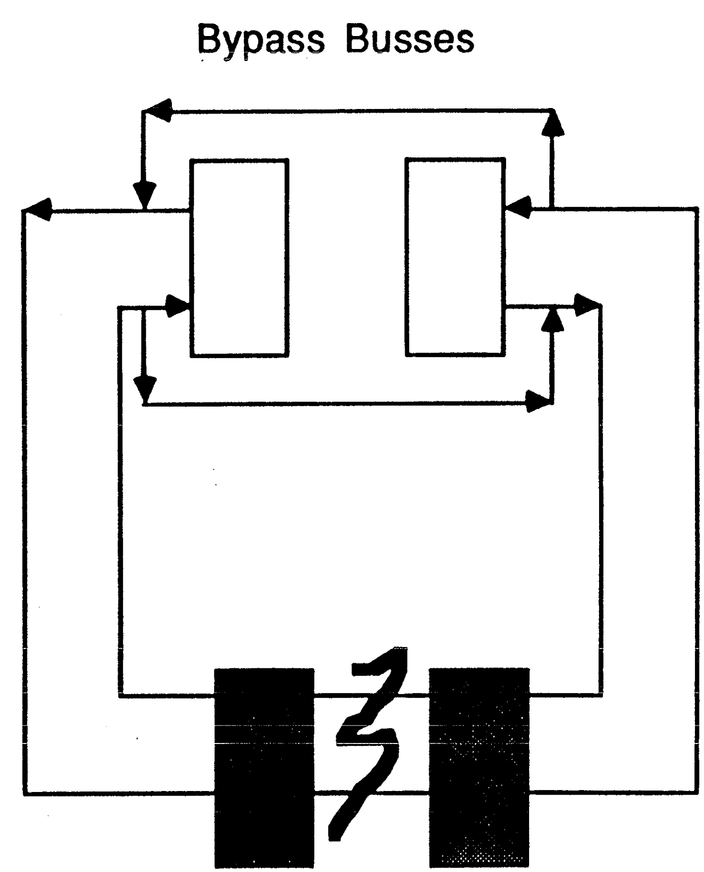


E. Singh, NYNEX-ATD, 3/21/88

Network Reconfiguration



Normal Operation



Healed Operation

E. Singh, NYNEX-ATD, 3/21/88

DQDB FEATURES

- Efficient utilization of bandwidth
- Fair access of bandwidth
- No inherent distance limitation
- Reliable - Self Healing

FIBER DISTRIBUTED DATA INTERFACE (FDDI)

E. Singh, NYNEX-ATD, 3/21/88

FDDI

- Proposed American National Standard

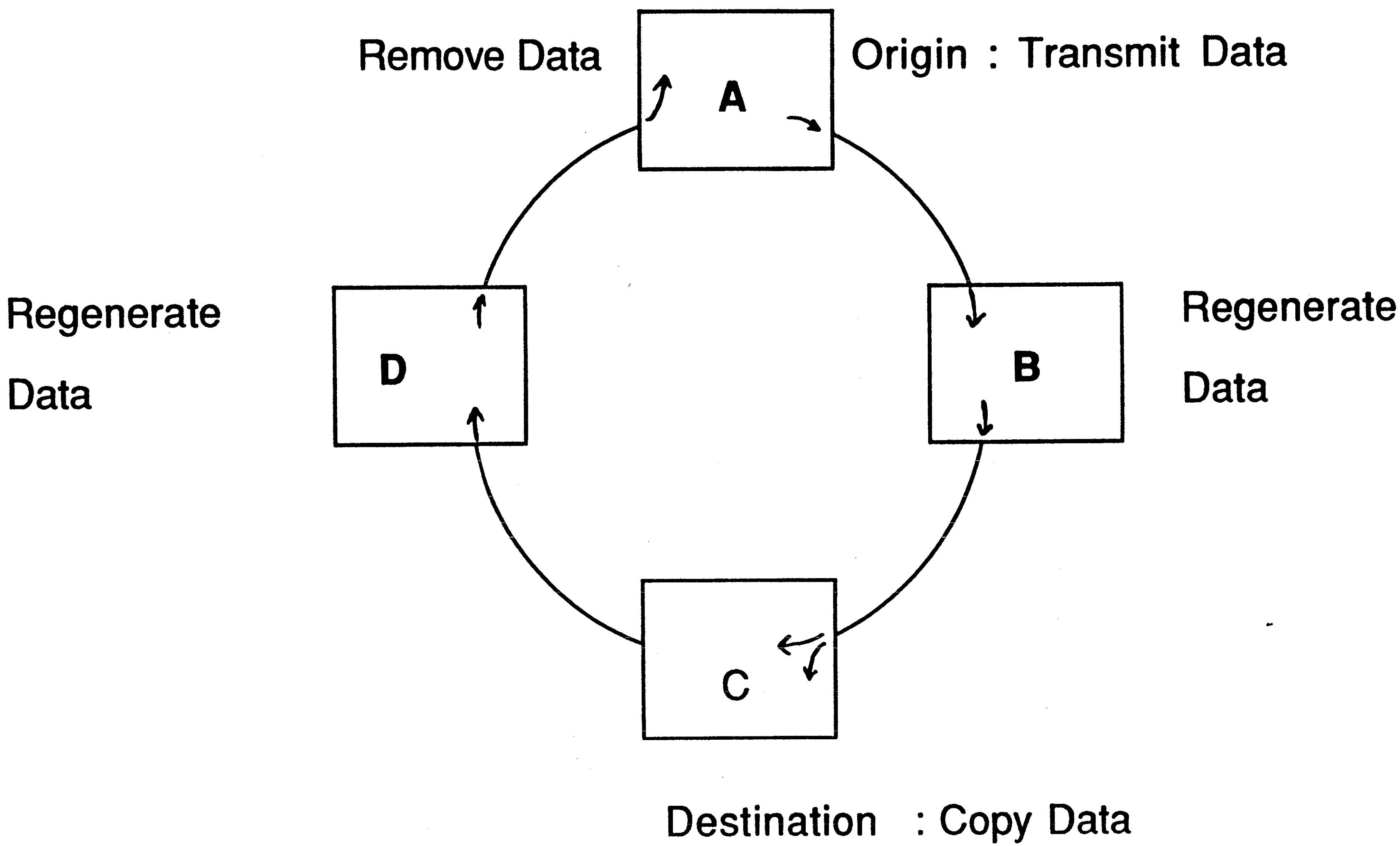
- Designed primarily for LAN environments

- Two classes of service

  - Synchronous traffic
  - Asynchronous traffic (restricted , non restricted)

- 100 Mbps token ring, fiber optics medium

FDDI Operation



E. Singh, NYNEX-ATD, 3/21/88

OVERVIEW OF OPERATION

- Information transmitted sequentially as a stream of symbols(4 bits of data)
- Each station regenerates and repeats each symbol
- The addressed destination station(s) copies the data as it passes on the ring
- Originating station removes the data from the ring

MEDIA ACCESS

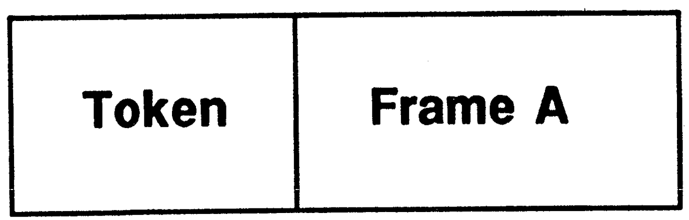
- How does a station gain the right to transmit information ?

  - Detect a Token (unique symbol sequence)
  - Remove Token from ring
  - Transmit information
  - Issue a new Token

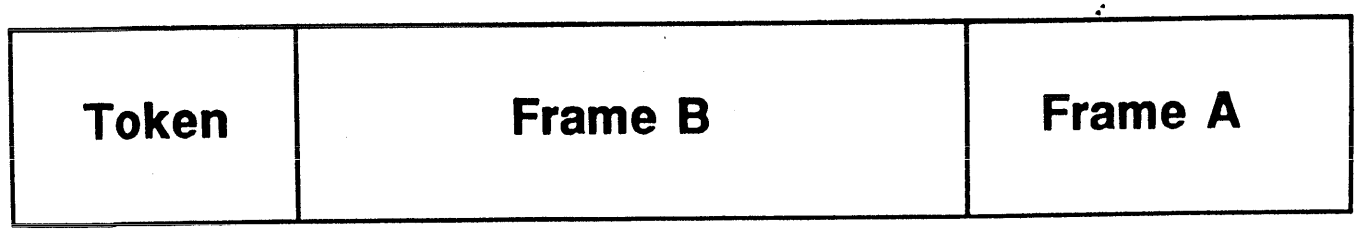
FDDI TOKEN OPERATION

Station A :
- Detects Token

Station A :
- Strips Token
- Transmits Frame A
- Issues Token



Station B :
- Strips Token
- Transmits Frame B
- Issues Token

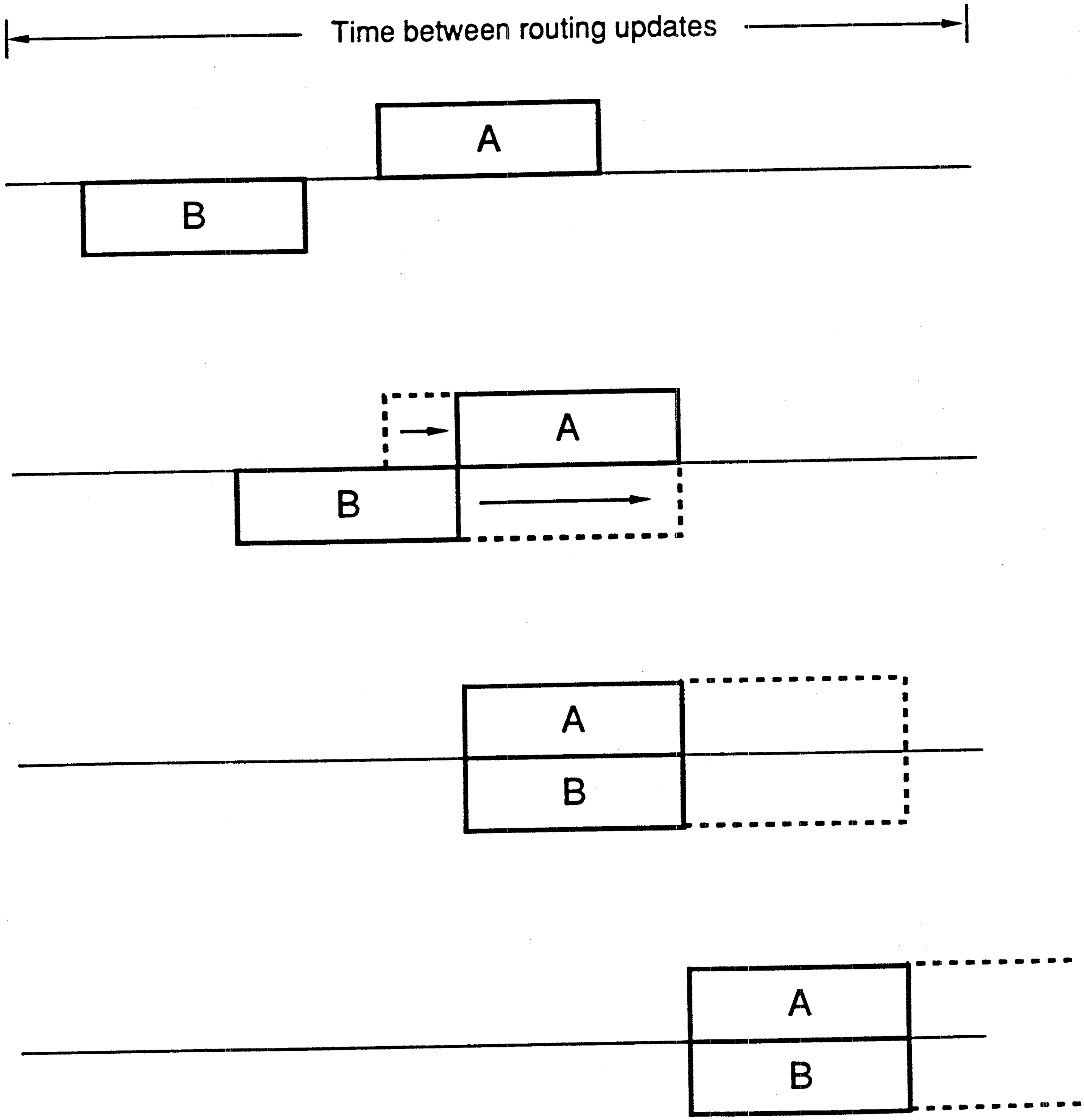


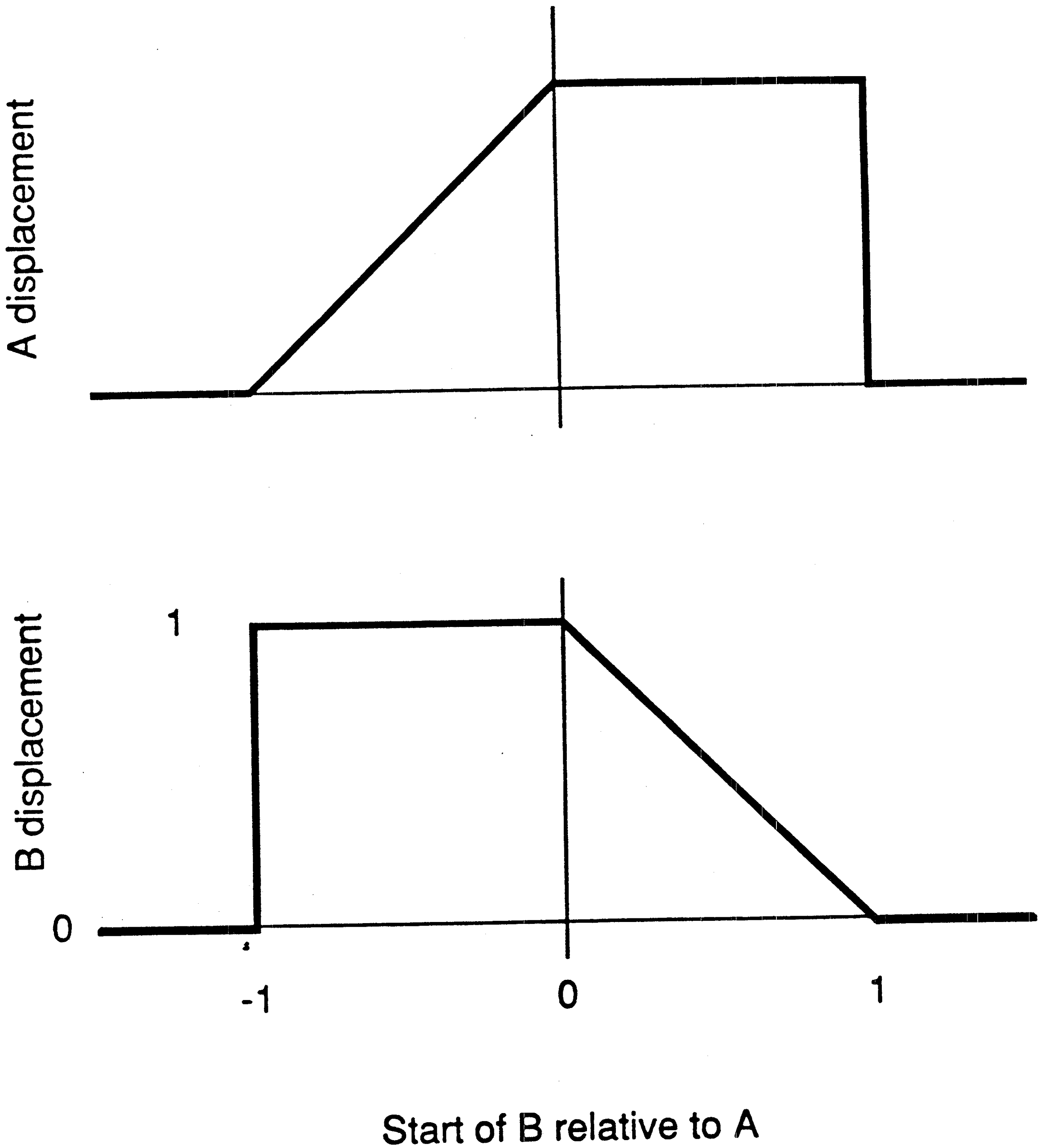
E. Singh, NYNEX-ATD, 3/21/88

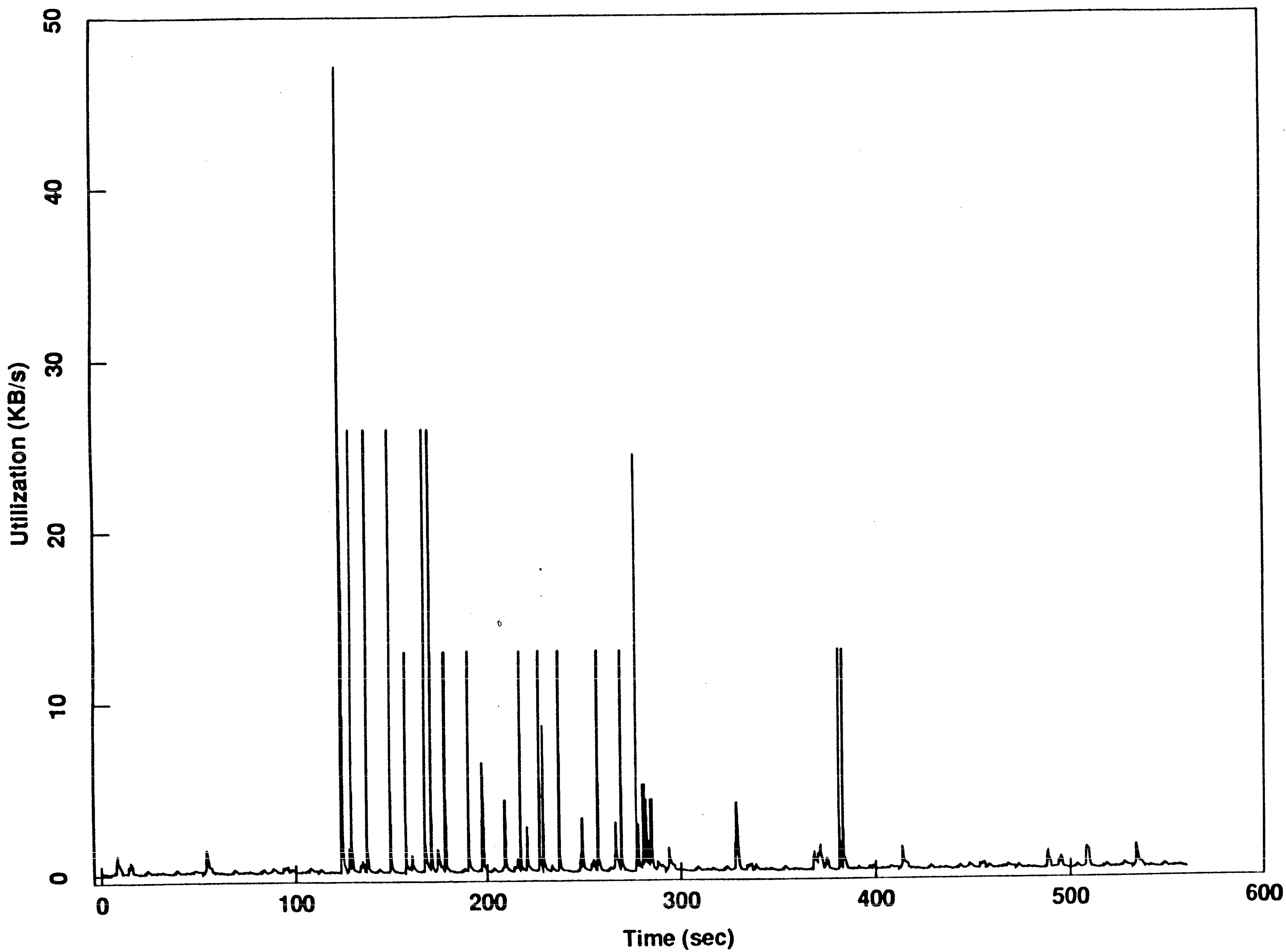
FDDI FEATURES

- Guaranteed bandwidth and average response time
- Maximum configuration of 500 stations, 100 km
- Reliable
  - Counter Rotating Ring
  - Station Bypass Switch

7.7 TCP Performance and Other Unconfirmed Rumors—Van Jacobson, LBL


Time between routing updates
A
B
A
B
A
B
A
B

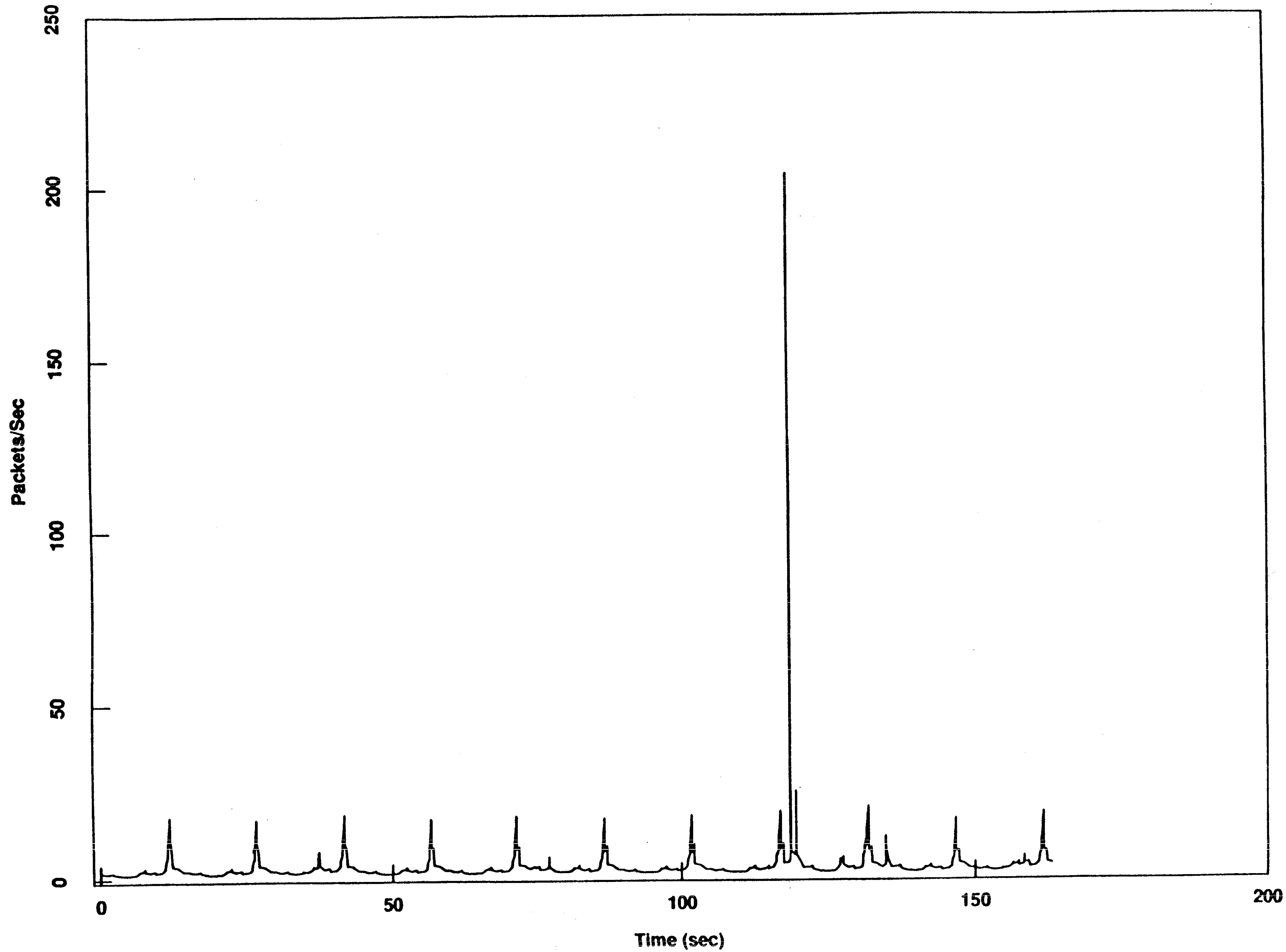

A displacement
B displacement
1
0
-1
0
1
Start of B relative to A



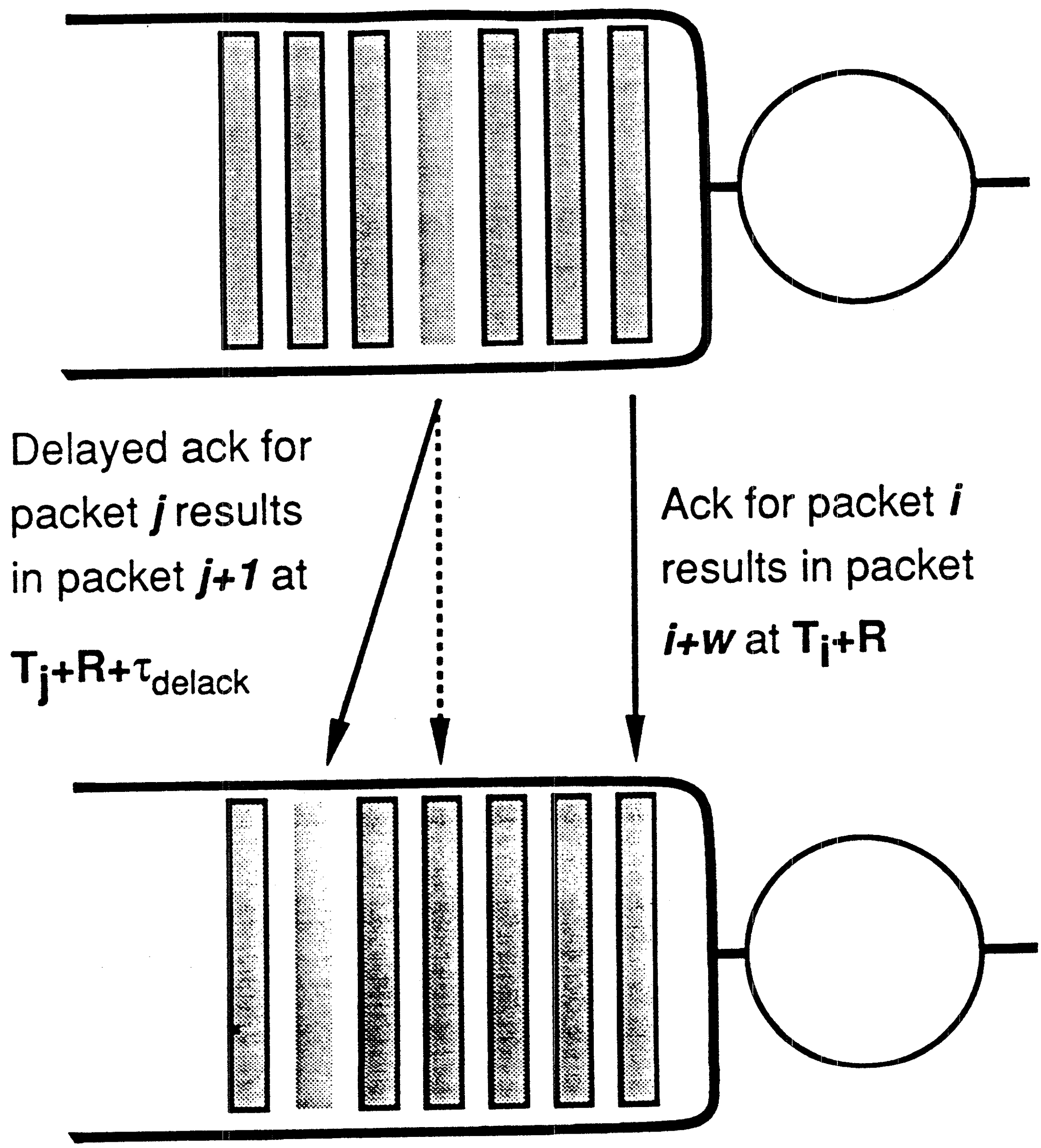
Utilization (KB/s)
0
10
20
30
40
50
0
100
200
300
400
500
600
Time (sec)

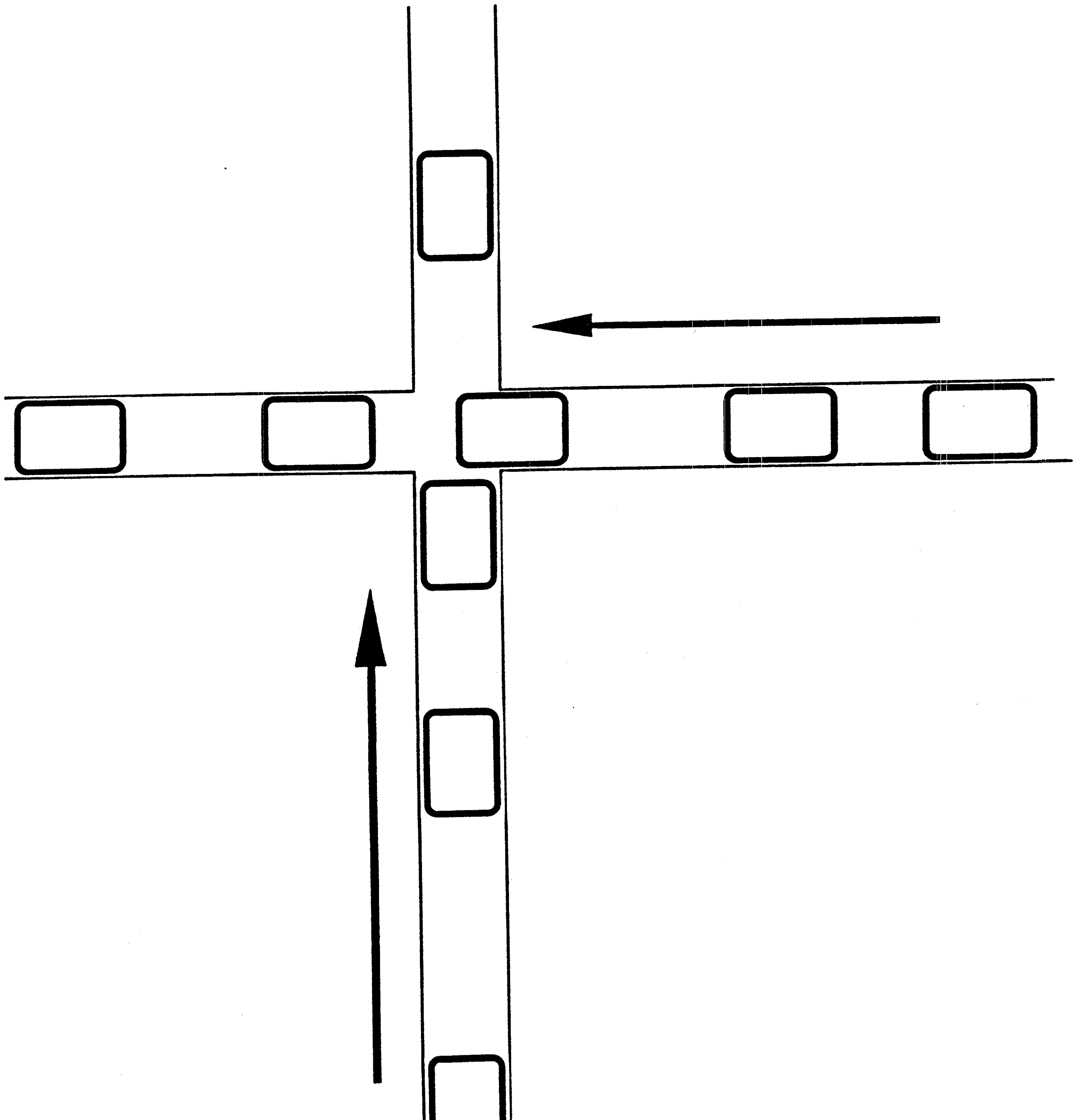
3 Minutes of LBL Ethernet Traffic
6am, Monday, May 18th, 1987

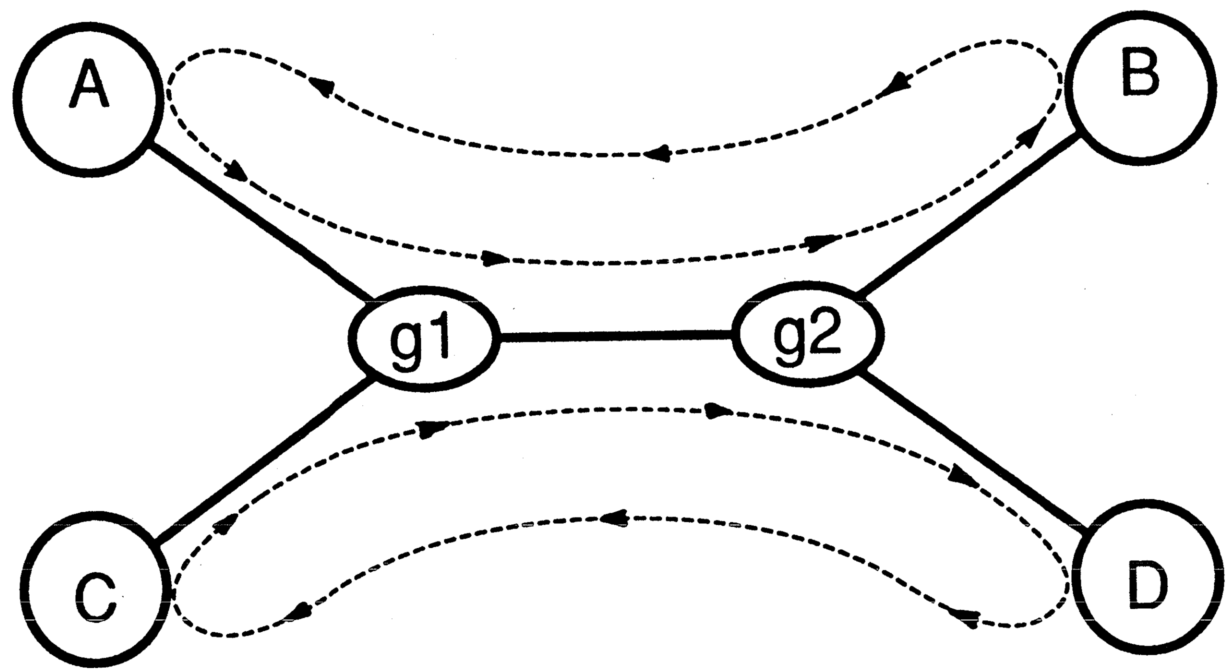
(1[illegible] clock-driven 15s & 120s updates)

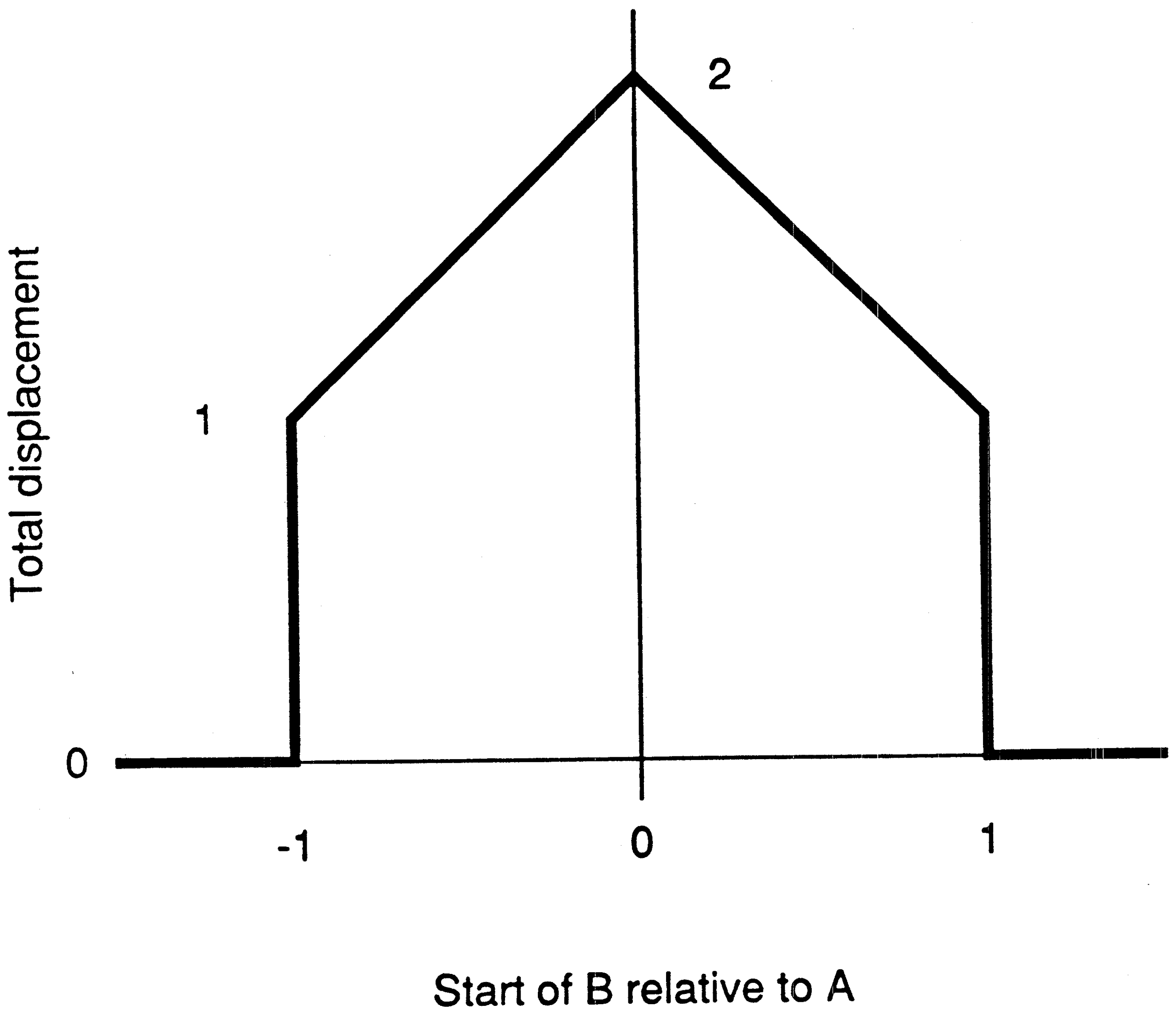


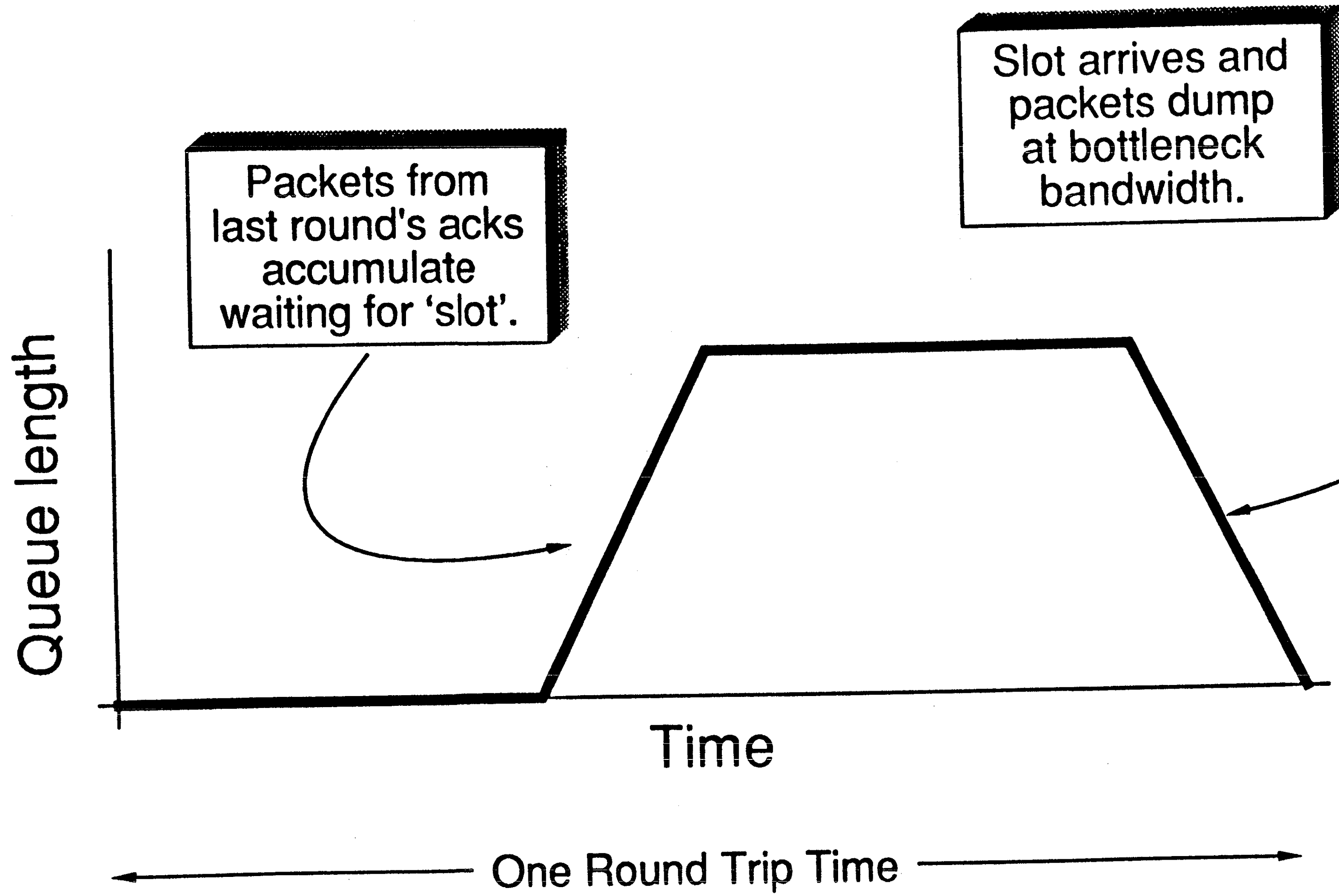
VJ


Delayed ack for packet *j* results in packet *j+1* at
$T_j+R+\tau_{delack}$
Ack for packet *i* results in packet *i+w* at T_i+R




A
B
g1
g2
C
D


2
1
0
Total displacement
-1
0
1
Start of B relative to A


Packets from last round's acks accumulate waiting for 'slot'.
Slot arrives and packets dump at bottleneck bandwidth.
Queue length
Time
One Round Trip Time

The Fokker-Planck equation for packet (probability) density ρ at position x and time t is:

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial x} m\rho + \frac{1}{2} \frac{\partial^2}{\partial x^2} \sigma^2 \rho$$

If the system is "viscous" ($d^2x/dx^2 \approx 0$), this simplifies to the Smoluchowski equation:

$$\frac{\partial \rho}{\partial t} = \frac{1}{2} \frac{\partial^2}{\partial x^2} \sigma^2 \rho$$

Some variant of the Smoluchowski equation shows up in many physical "agregation" processes. E.g., the coagulation of a colloidal suspension.

Given an initial particle concentration of C_0, diffusion coefficient D and reaction distance R, the equation can be solved to give the rate of growth of "clumps" of size k, relative to the initial concentration:

$$C_k = C_0 \frac{(C_0 \tau)^{k-1}}{(1 + C_0 \tau)^{k+1}}$$

where the time-scale $\tau = 4\pi DRt$.

7.8 Cray TCP Performance, An Update—Borman, Cray

HSX Transfer rate:

- 75 nanosecond / word (8 bytes)
- User to User RTT 860 μsec.
- 604.8 μsec + 430 μsec for 63 Kbyte Transfer, or 498.7 Mbits/sec
- at 32 kbytes, 355.6 Mbits/sec

- HSX transfer rate
  ⇒ 75 nanosec/word
  ⇒ 230 usec/24K block

- HSX User to User RTT: 860 usec
  ⇒ Assume 430 usec one way
  ⇒ 430 + 230 usec = 660 usec for transfer
  ⇒ 2166 - (1210 + 660) = 296 usec (~70000 clocks) not yet accounted for.

HSX transfer rate
307 μsec/32 K block

⇒ 430 + 307 μsec = 737 μsec for transfer
⇒ 1916 μsec - (861 + 737) = 318 μsec (~75000 clocks)

```
./mcli -tcp -f -kb 512k localhost 100 512k
Transfer: 100*524288 bytes from          to localhost
             Real   System              User            Kbyte    Mbit(K^2) mbit(1+E6
  write   1.6750   0.3324  (19.8%)   0.0015  ( 0.1%) 30567.16 238.806    250.406
   read   1.7140   0.9913  (57.8%)   0.0048  ( 0.3%) 29871.65 233.372    244.709
    r/w   3.3890   1.3237  (39.1%)   0.0063  ( 0.2%) 30215.40 236.058    247.525
  5120:       1   15363:       10   23555:       1   27651:       12
 32771:      26   33792:        2   43008:      10   51200:        1
 68608:      10  205824:        1  210944:      10  218115:        1
219136:      12  220160:        1  224256:      11  226305:       12
227327:      25  227328:       12  228352:       2  229373:       50
243712:      26  246784:       12  249856:      12  254976:       13
254977:      13
# _
```

Usertoken: 32K
MTU : 32K

Software loopback

```
# ./mcli -tcp -f -kb 256k localhost 200 256k
Transfer: 200*262144 bytes from          to localhost
             Real   System              User            Kbyte    Mbit(K^2) mbit(1+E
  write   1.7750   0.4014  (22.6%)   0.0030  ( 0.2%) 28845.07 225.352    236.299
   read   1.7630   0.9201  (52.2%)   0.0056  ( 0.3%) 29041.41 226.886    237.907
    r/w   3.5380   1.3215  (37.4%)   0.0086  ( 0.2%) 28942.91 226.116    237.100
  5120:      17    9216:       17   15363:      17   23555:       17
 27651:      17   32771:        1   33792:      19   84992:       17
194560:      17  201728:        8  219136:      16  220160:        1
222208:       7  223231:        7  223232:      24  224257:        1
227327:       7  228352:       26  229373:      53  229377:        1
230400:       8  237568:        8  244736:      16  245760:        1
246785:       7  253953:        1  254977:       7
# _
```

Usertoken: 32K
MTU : 32K

Software loopback

```
s3# ./mcli -tcp -f -kb 256k snql-hsx 200 256k
Transfer: 200*262144 bytes from             to  snql-hsx
           Real   System              User              Kbyte    Mbit(K^2)  mbit(1+E
  write   2.3550  0.2934  (12.5%)   0.0038  ( 0.2%)  21740.98  169.851    178.102
   read   3.8370  0.4000  (10.4%)   0.0258  ( 0.7%)  13343.76  104.248    109.312
    r/w   6.1920  0.6934  (11.2%)   0.0296  ( 0.5%)  16537.47  129.199    135.475
 16160:       1  32840:   1596
s3# _
```

Usertokern: 32K   Cray2 ↔ Cray2

MTU: sn2012 → snql 32K
snql → sn2012 16K

```
s3# ./mcli -tcp -f kb 512k snql-hsx 200 256k
Usage: mcli [-d] [-c] [-f] [-kb ###]
            [-tcp [host]] [-udp [host]] [-unix] [-pipes]
            [count] [size] [port]
s3# ./mcli -tcp -f -kb 512k snql-hsx 200 256k
Transfer: 200*262144 bytes from             to  snql-hsx
           Real   System              User              Kbyte    Mbit(K^2)  mbit(1+E
  write   3.4500  0.2888  ( 8.4%)   0.0101  ( 0.3%)  14840.58  115.942    121.574
   read   3.8390  0.4005  (10.4%)   0.0258  ( 0.7%)  13336.81  104.194    109.255
    r/w   7.2890  0.6894  ( 9.5%)   0.0359  ( 0.5%)  14048.57  109.754    115.086
 16160:       1  32840:   1596
s3# _
```

Usertokern: 32K   Cray2 ↔ Cray2

MTU: 16K

```
s3# ./mcli -tcp -f -kb 256k snql-hsx 200 256k
Transfer: 200*262144 bytes from             to  snql-hsx
           Real   System              User              Kbyte    Mbit(K^2)  mbit(1+E
  write   2.3550  0.2933  (12.5%)   0.0038  ( 0.2%)  21740.98  169.851    178.102
   read   2.3790  0.4002  (16.8%)   0.0258  ( 1.1%)  21521.65  168.138    176.305
    r/w   4.7340  0.6935  (14.6%)   0.0296  ( 0.6%)  21630.76  168.990    177.199
 16160:       1  32840:   1596
s3# _
```

Usertokern: 32K   Cray2 ↔ Cray

MTU: 32K

```
# ./mcli -tcp -f -kb 256k snql-hsx 100 128k
Transfer: 100*131072 bytes from     snql to  snql-hsx
          Real   System             User             Kbyte    Mbit(K^2)  mbit(1+E6)
  write  1.0240  0.1453 (14.2%)  0.0036 ( 0.4%) 12500.00   97.656   102.400
   read  1.0430  0.6171 (59.2%)  0.0159 ( 1.5%) 12272.29   95.877   100.535
    r/w  2.0670  0.7624 (36.9%)  0.0196 ( 0.9%) 12385.10   96.759   101.459
 19112:     1  24648:    332  49296:      96  73944:      1
 98592:     1
# _
```

User tokern: 32K
HSX MTU: 24K

Hardware loopback

```
# ./mcli -tcp -f -kb 256k snql-hsx 100 128k
Transfer: 100*131072 bytes from              to  snql-hsx
          Real   System             User             Kbyte    Mbit(K^2)  mbit(1+E6)
  write  0.9910  0.2122 (21.4%)  0.0037 ( 0.4%) 12916.25  100.908   105.810
   read  1.0140  0.5863 (57.8%)  0.0119 ( 1.2%) 12623.27   98.619   103.410
    r/w  2.0050  0.7984 (39.8%)  0.0156 ( 0.8%) 12768.08   99.751   104.596
 32840:   265  36880:      1  65680:      65  98520:      1
# _
```

Usertokern: 4K
HSX MTU: 32K

Hardware loopback

```
# ./mcli -tcp -f -kb 256k snql-hsx 200 256k
Transfer: 200*262144 bytes from              to  snql-hsx
          Real   System             User             Kbyte    Mbit(K^2)  mbit(1+E6)
  write  3.3700  0.3978 (11.8%)  0.0072 ( 0.2%) 15192.88  118.694   124.460
   read  3.3890  2.1698 (64.0%)  0.0527 ( 1.6%) 15107.70  118.029   123.762
    r/w  6.7590  2.5676 (38.0%)  0.0600 ( 0.9%) 15150.17  118.361   124.110
 16160:     1  32840:   1297  65680:     148  98520:      1
# _
```

Usertokern: 32K
HSX MTU: 32K

Hardware loopback

Things to do:

- Add TCP window shift option
- Add Van Jacobson header prediction code
- Analyze flowtrace of Kernel to identify possible areas of improvement.

Measurements:

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

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

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

124 Mbits	32k	473	198	1916	hsx lo
177 Mbits	32k	675	100	1381	hsy
247 Mbits	32k	944	198	861	softlo[illegible]

Changes since Feburary:

- Checksum vectorized.
  - Scalar to vector crossover point
    - Save vectors: 800 bytes
    - Don't save vectors: 180 bytes
- Larger copies from user to kernel, into large mbufs (32k for these numbers)
- Bug fix in wayout() code in CRAY-2 Kernel.

7.9 Issues in Canadian Networking—Prindeville, McGill

Users in Canada

- Universities
- High-Tech Firms
    - Computer
    - Telecom
    - Aerospace
    - . . .
- Research Facilities:
    - Libraries & Databases
    - Medical
    - Space
    - Physical Sciences
    - National Resources:
        - Fisheries
        - Mines
        - Logging
        - . . .
- Government (other)

Groups

NetNorth	- BITNET North
CDNnet	- Commercial X.400 mail service
Interneters	- McGill, Toronto, UBC...

Needs

TCP/IP	
RSCS/SNA	- NetNorth
DECnet	- SPAN/DAN, HEPNET
ISO?	

Network Requirement

- Rapid deployment
- Existing standards & technology
- High bandwidth
- Production oriented
- Three tier organization:
  national, regional, local
- Transition to ISO later
- Privatization in 5 years

The Players

Vancouver	- BCnet
Calgary	- (Supercomputer facility)
*Saskatoon	
Toronto	- ONet, SC facility
Ottawa	- Feds, telcos
Montreal	- CRIM, SC facility
Fredrickton	
*St. John's	

Toronto/IBM

- TCP/IP suite
- NSS-like technology
- 56k; 1.5mbps later
- off-the-shelf technology
- get it running today
- free (IBM grant)
- unifying force for various camps:
    common denominator technology
        (minimal functionality)
    wide range of implementations
- solid networking experience
- good research resources

UBC

- X.25 service (undisclosed switch)
- 56k - 1.5mbps
- no network (DoD or ISO IP) or
  transport (TP0) support
- minimal NOC(s)
- good commercial track-record

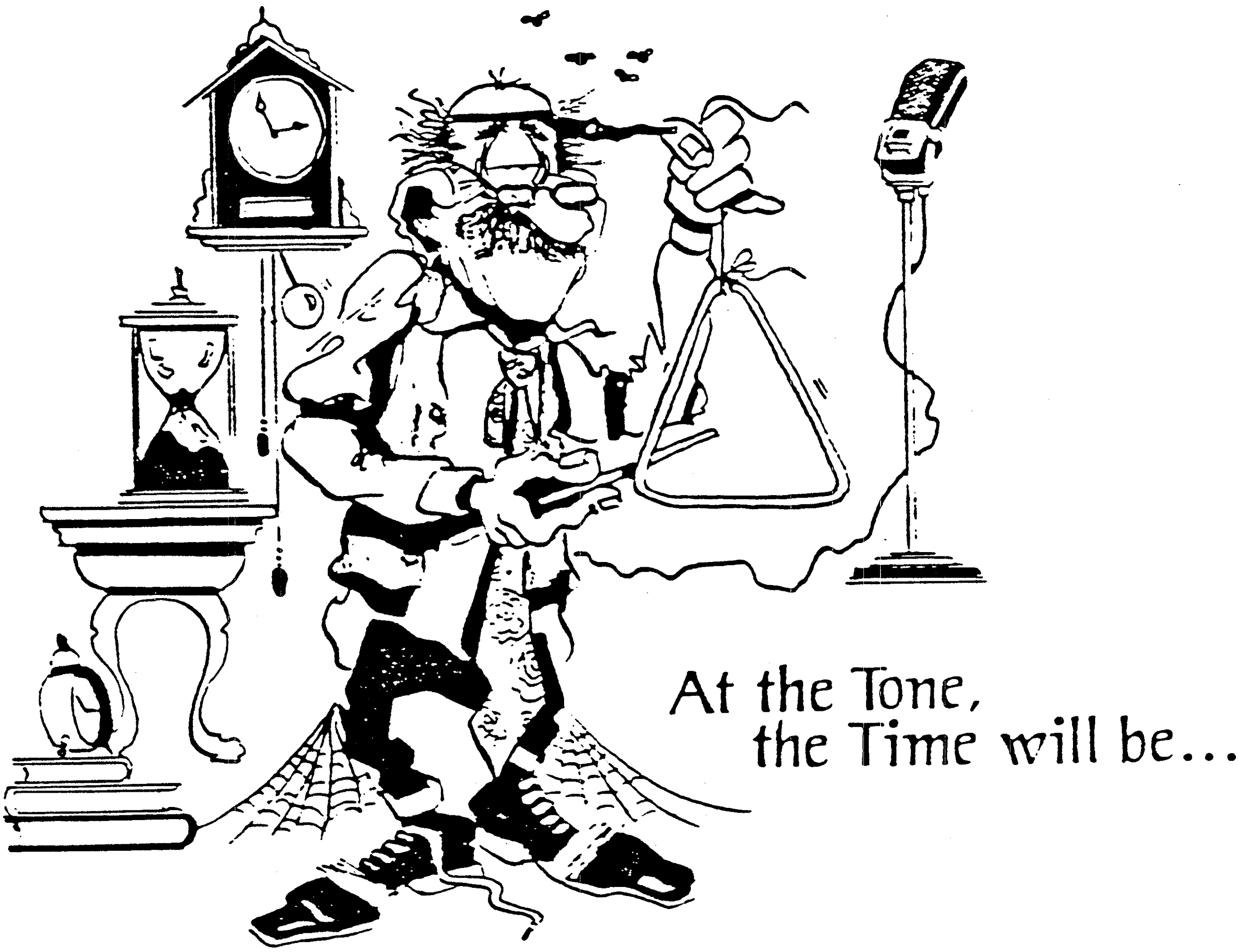
AlterNet

- get it running today
    (before lunch?)
- "disposable" technology (off-the-shelf routers)
- start with 1.5mbps
- strong support for:
    - regional development
    - NOC(s)
    - further research...
- develop switching technology
    - T3 and up
    - multiple protocol support (TCP/IP, ISO, DECnet, RSCS V2)
    - off-the-shelf technology (VMEbus?)
    - involvement of telecom manufacturers
    - participation in standards process
- good connectivity with NSFnet, DRI, IRI, EARN, RARE, JUNET...

Problems/Issues

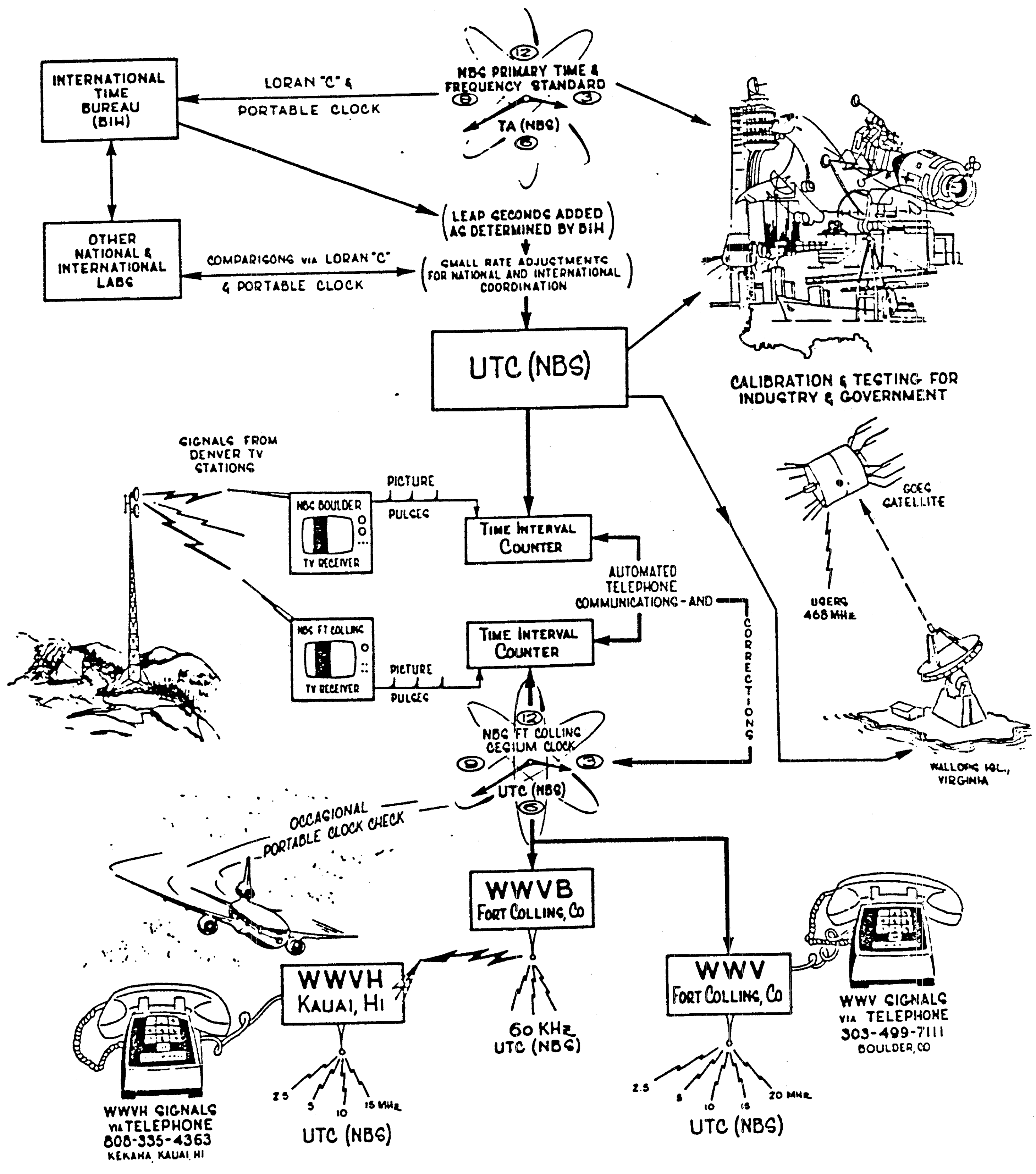
- Communications regulation (CRTC)
  - Canada is larger area with smaller population
  - Largely monopoly; slow to offer new services
  - Heavy cross-subsidization of residential and loop service
  - Cheaper to drop lines south and go cross-continent in U.S.
- Lot of "dark fibre" (unused bandwidth)
- Multi-protocol support
  - coercion or extortion?
  - management headache
- Multiple carriers and type-of-service routing
  - FTP/mail via satellite
  - TELNET via terrestrial
- Policy-based routing
  - stay in Canada if possible, otherwise use U.S. path
- ISO development, possibly using TCP/IP transport (ISODE)

7.10 Bellringing, Clock Punching and Gongferming—Mills, UDel

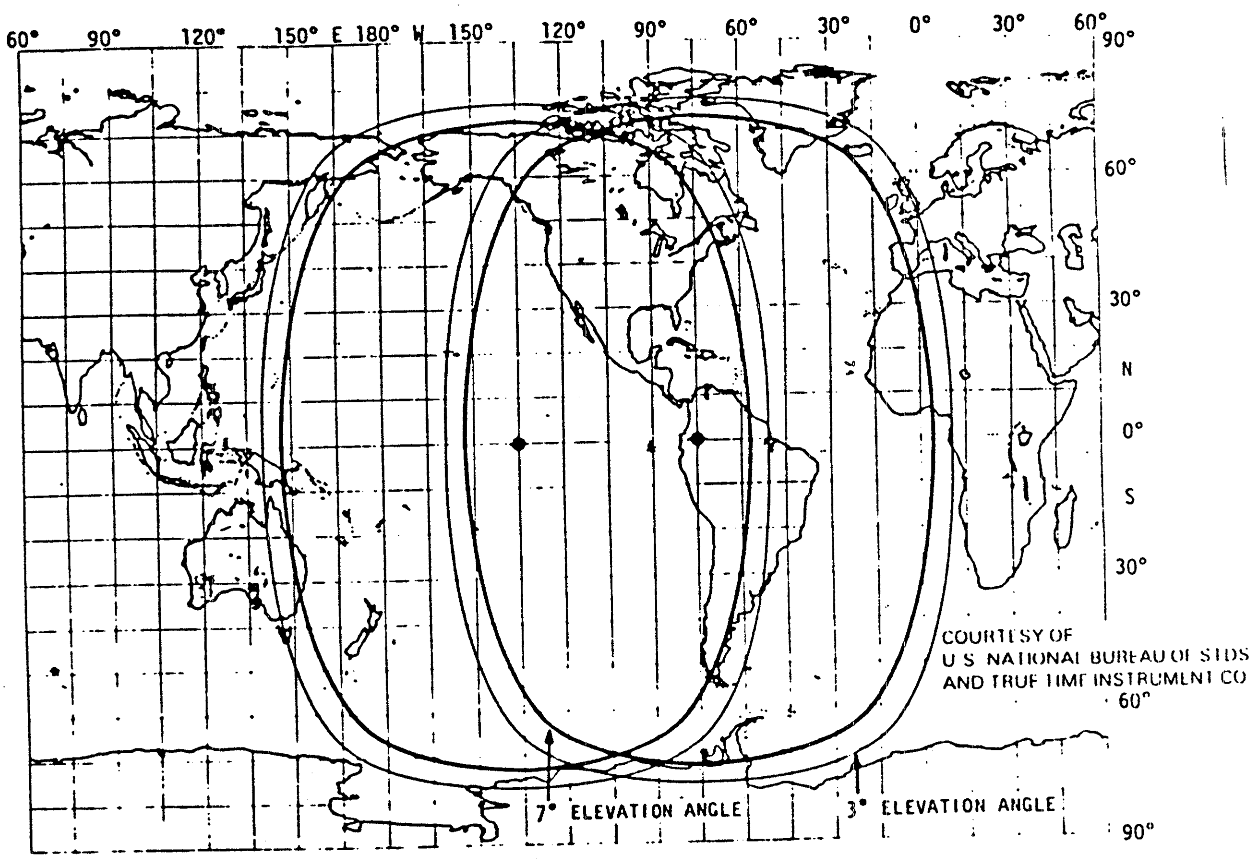


Network Time Protocol

NATIONAL BUREAU OF STANDARDS
FREQUENCY AND TIME FACILITIES



GOES SATELLITE LOCATIONS AS OF JUNE 1st, 1980



There are three GOES satellites in orbit, two in operation and the third serving as an in-orbit spare
The two operational units are located as shown above and covering the areas indicated

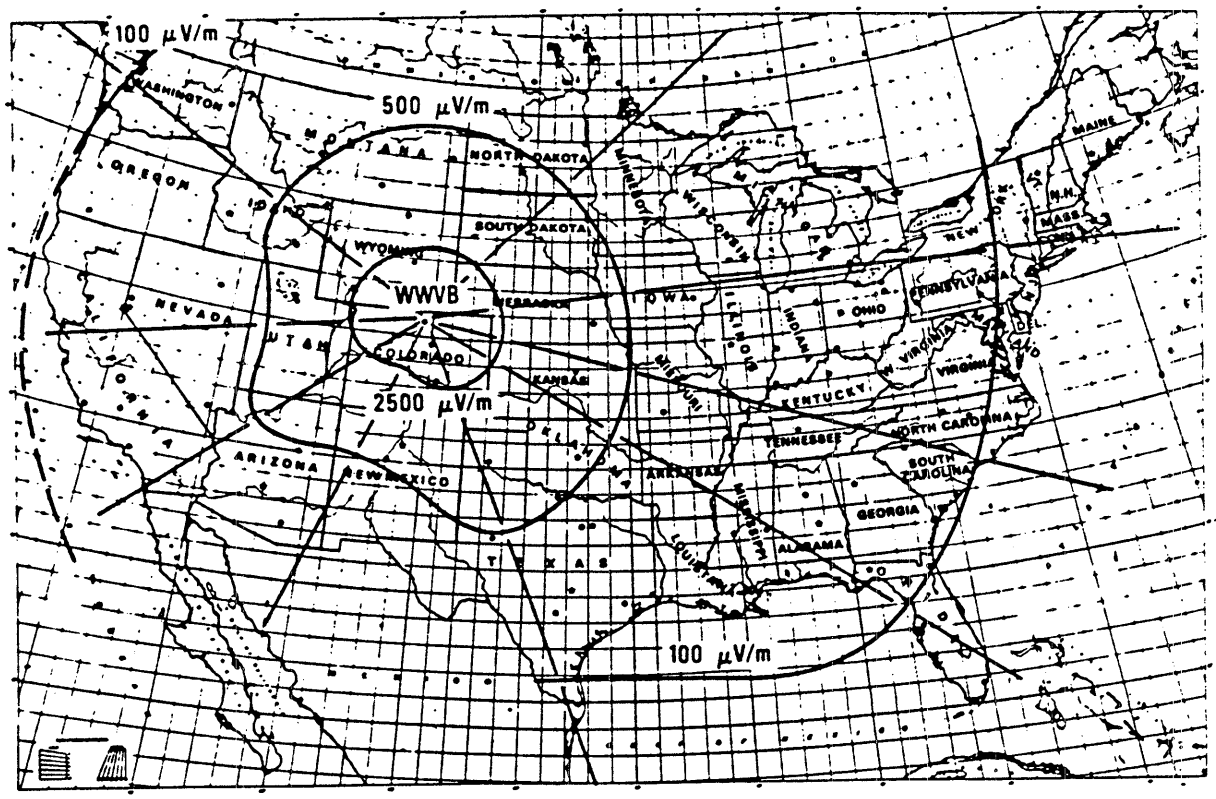
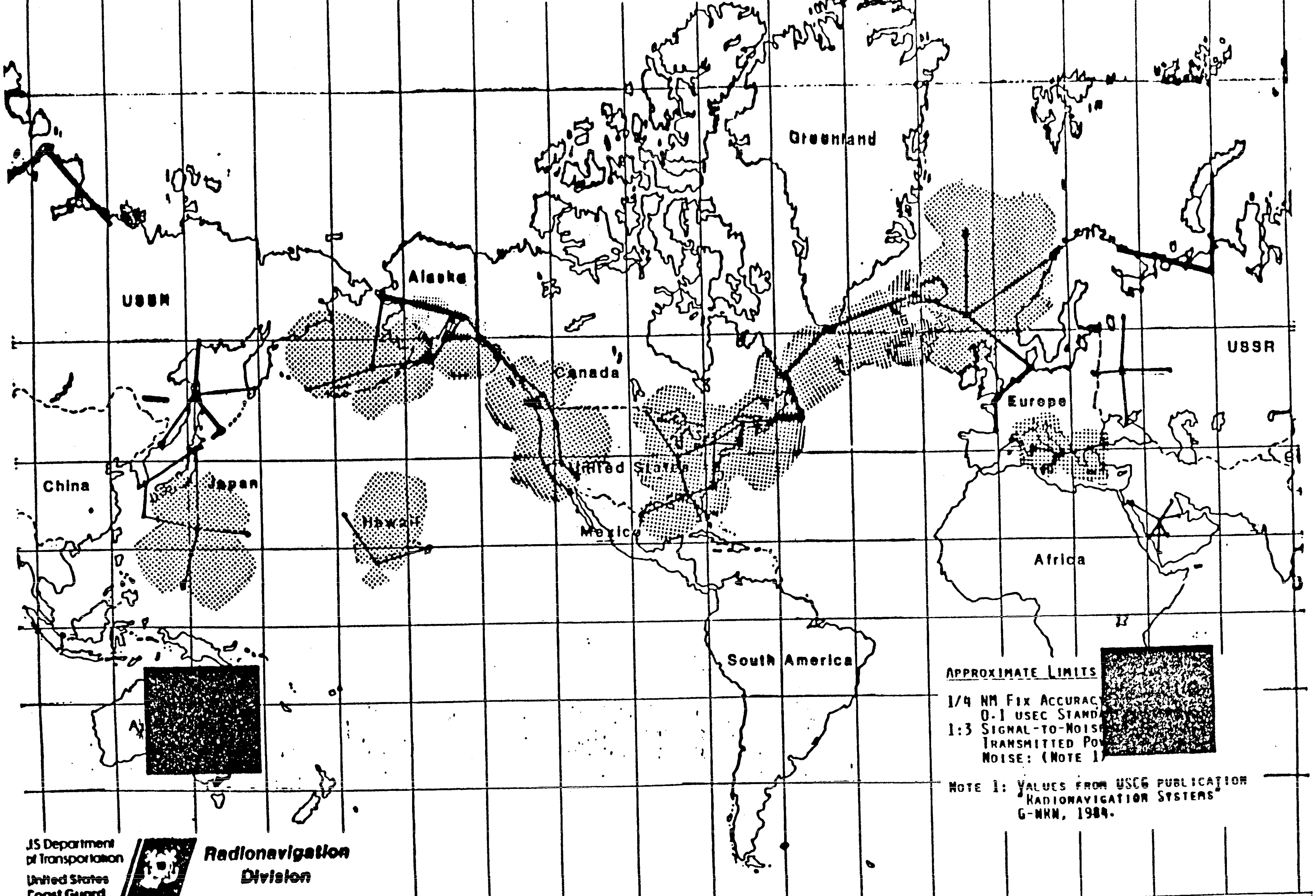
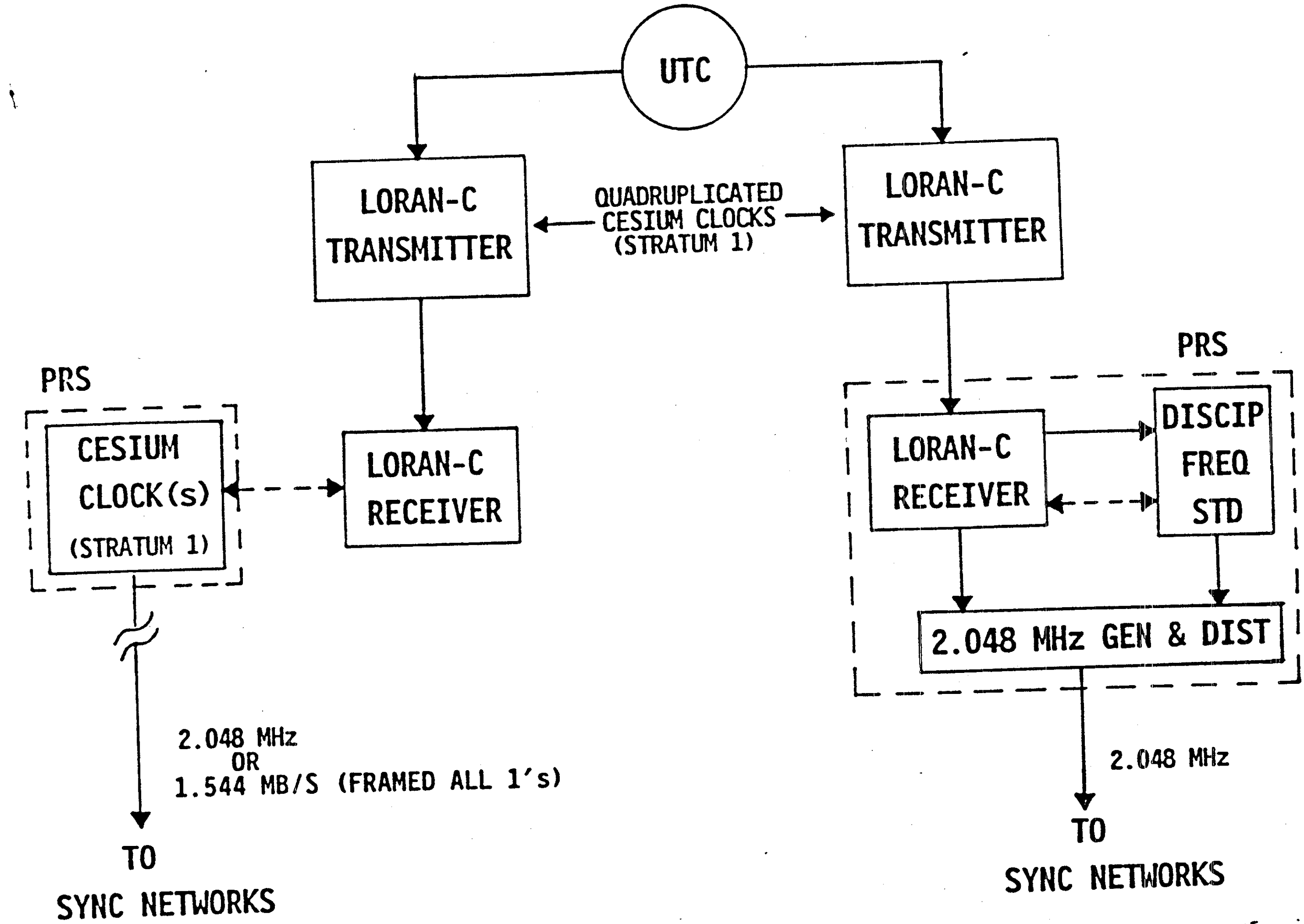
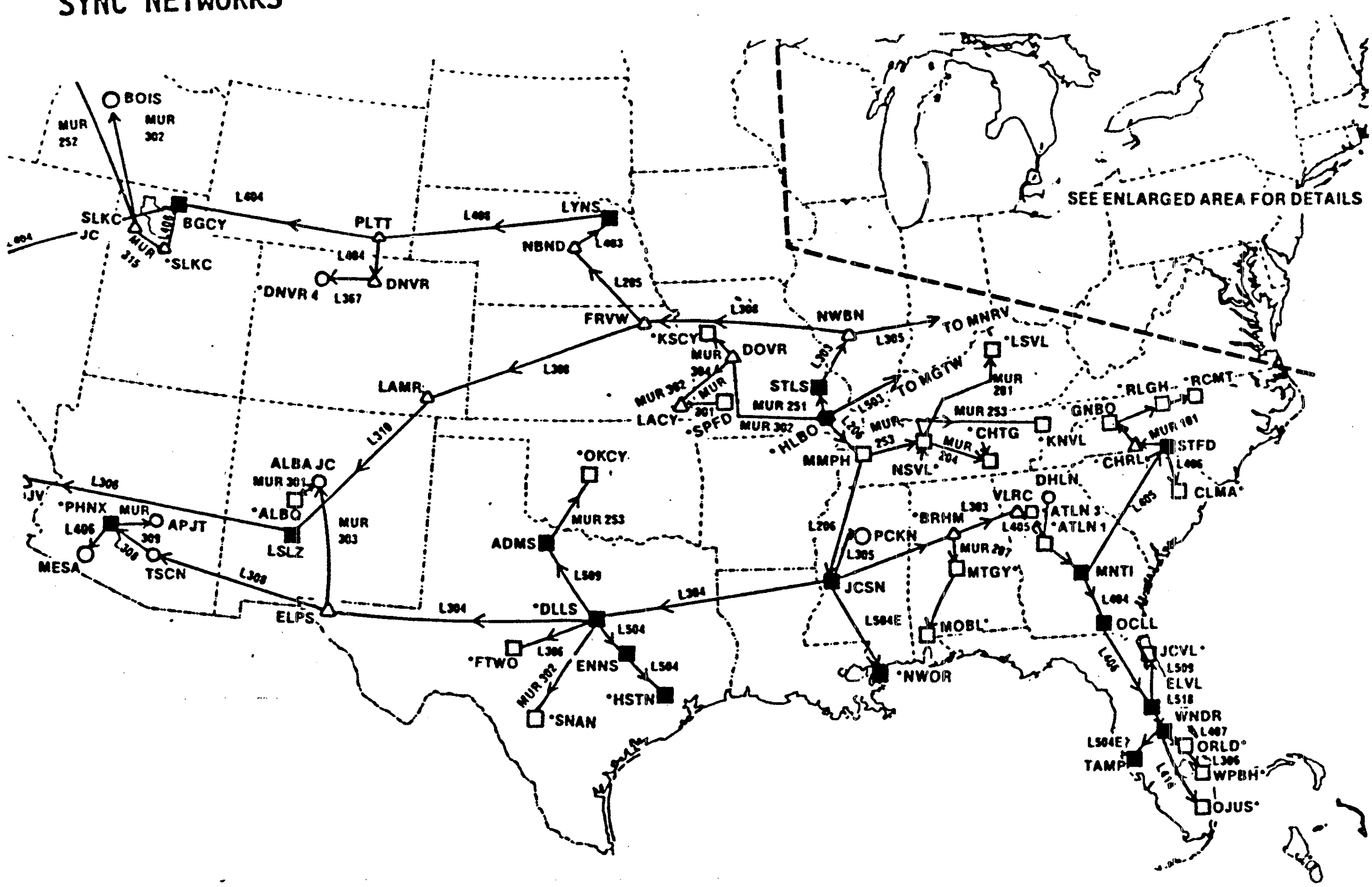


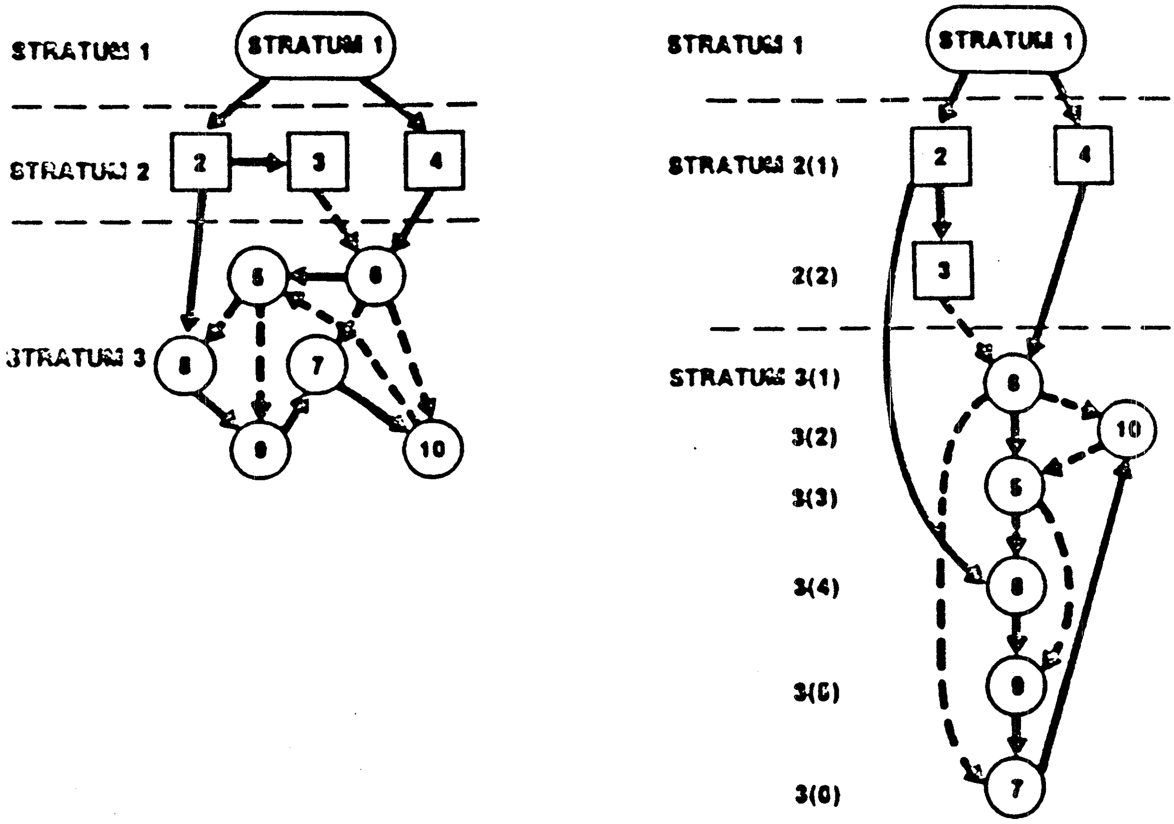
FIG. 1-3 MEASURED FIELD INTENSITY CONTOURS: WWVB @ 13 KW ERP

HYPERBOLIC LORAN-C COVERAGE



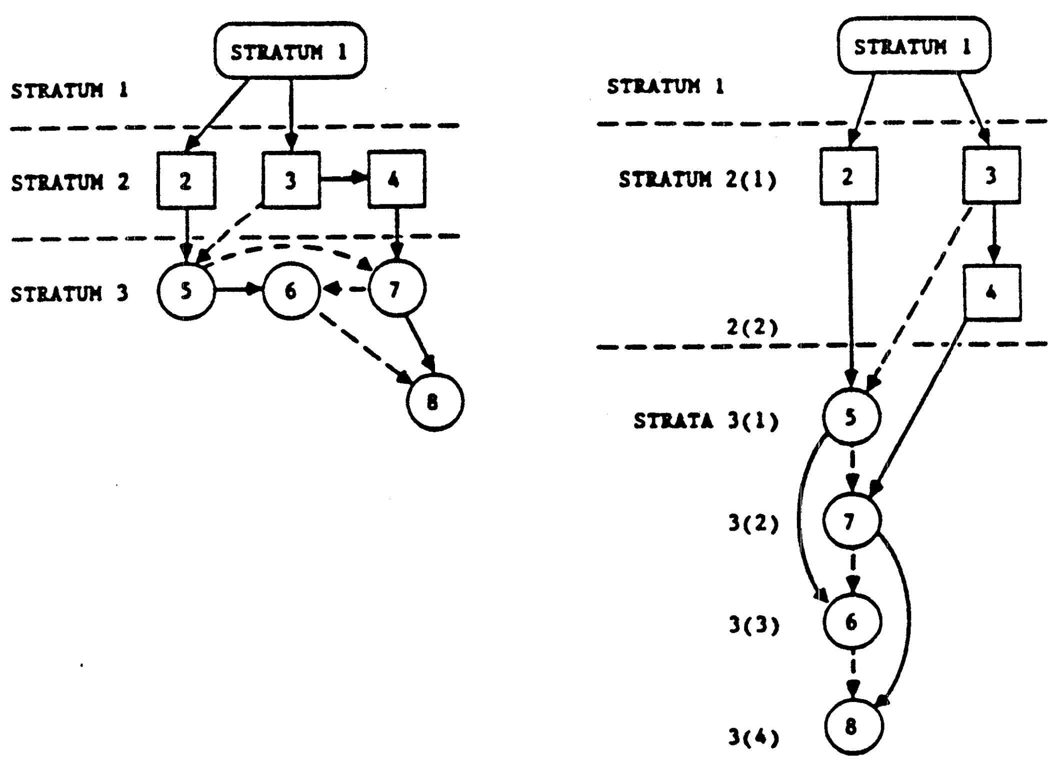

UTC
LORAN-C TRANSMITTER
QUADRUPLICATED CESIUM CLOCKS (STRATUM 1)
LORAN-C TRANSMITTER
PRS
CESIUM CLOCK(s) (STRATUM 1)
LORAN-C RECEIVER
PRS
LORAN-C RECEIVER
DISCIP FREQ STD
2.048 MHz GEN & DIST
2.048 MHz OR 1.544 MB/S (FRAMED ALL 1's)
TO SYNC NETWORKS
2.048 MHz
TO SYNC NETWORKS


SEE ENLARGED AREA FOR DETAILS
BOIS
SLKC
BGCY
PLTT
LYNS
NBND
DNVR
FRVW
KSCY
DOVR
NWBN
TO MNRV
LSVL
STLS
TO MGTW
LAMR
LACY
SPFD
HLBO
MMPH
NSVL
CHTG
KNVL
GNBO
RLGH
RCMT
STFD
CHRL
CLMA
ALBA JC
ALBQ
LSLZ
PHNX
APJT
MESA
TSCN
ELPS
OKCY
ADMS
DLLS
FTWO
ENNS
HSTN
SNAN
PCKN
JCSN
BRHM
MTGY
MOBL
NWOR
VLRC
DHLN
ATLN 3
ATLN 1
MNTI
OCLL
JCVL
ELVL
WNDR
ORLD
TAMP
WPBH
OJUS



(A) CONVENTIONAL GEOGRAPHIC LAYOUT (B) LAYOUT WITH SUBSTRATA

PRIMARY REFERENCE SECONDARY REFERENCE



9(A) CONVENTIONAL GEOGRAPHIC LAYOUT 9(B) LAYOUT WITH SUBSTRATA

PRIMARY REFERENCE SECONDARY REFERENCE

- Previous version described in RFC-958

- Evolved over five-year period

- Based on Hellospeak LAN routing protocol

- Related technology
  - Unix timed - uses election protocol to establish master, then master polls slaves, redistributes timestamps
  - Xerox - broadcasts timestamps, uses convergence algorithm to adjust each clock independently
  - IBM - slot-synchronizes entire network, assigns unique time to each slot
  - Others - based on interactive convergence and consistency algorithms; status not known

- Survey conducted in early January 1988 of 5498 hosts and 224 gateways listed in Network Information Center tables:

46	Network Time Protocol
1158	TIME Protocol
1963	ICMP Timestamp Message

  Plus many more listed only in domain-name system or not at all

- o Primary Service Network (Fuzzball)
  - U Delaware (Newark, DE), WWVB
  - U Maryland (College Park, MD), WWVB
  - NCAR (Boulder, CO), WWVB
  - Ford Research (Dearborn, MI), GOES
  - ISI (Marina del Rey, CA), WWVB

- o Primary Backup Servers (Fuzzball)
  - U Michigan (Ann Arbor, MI), WWV
  - Backroom (Newark, DE), WWV

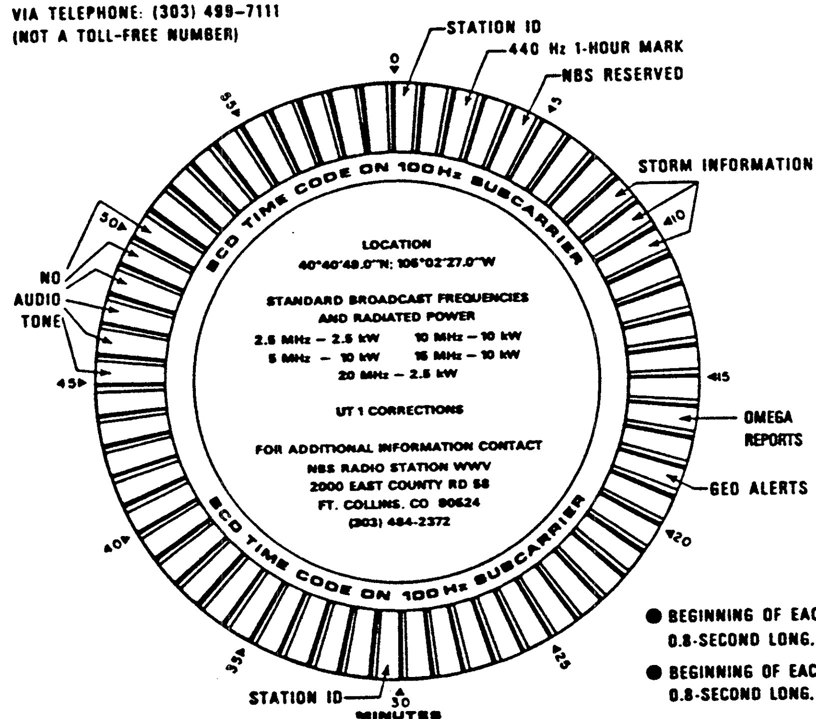
- o Secondary Service Network (Fuzzball)
  - Rice University (Houston, TX)
  - M/A-COM Government Systems (Vienna, VA)
  - Ford Research (Dearborn MI)
  - DEC Western Reseach Labs (Palo Alto, CA)
  - NASA/AMES (Sunnyvale, CA)
  - University of Hawaii (Honolulu, HA)
  - USECOM Patch Barracks (Stuttgart, FRG)
  - DFVLR (Oberpfaffenhofen, FRG)
  - CNUCE (Pisa, Italy)
  - NTA - RE (Oslo, Norway)
  - UK MoD - RSRE (Malvern, UK)
  - SHAPE Technical Centre (den Hague, Holland)

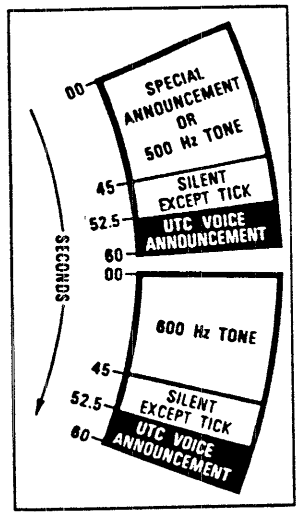
- o Secondary Service Network and retail distribution (Unix 4.3bsd NTP daemons)
  - About two dozen peers using present servers
  - Present implementation manages local time and date

Present Deployment Status

WWV BROADCAST FORMAT

VIA TELEPHONE: (303) 499-7111
(NOT A TOLL-FREE NUMBER)

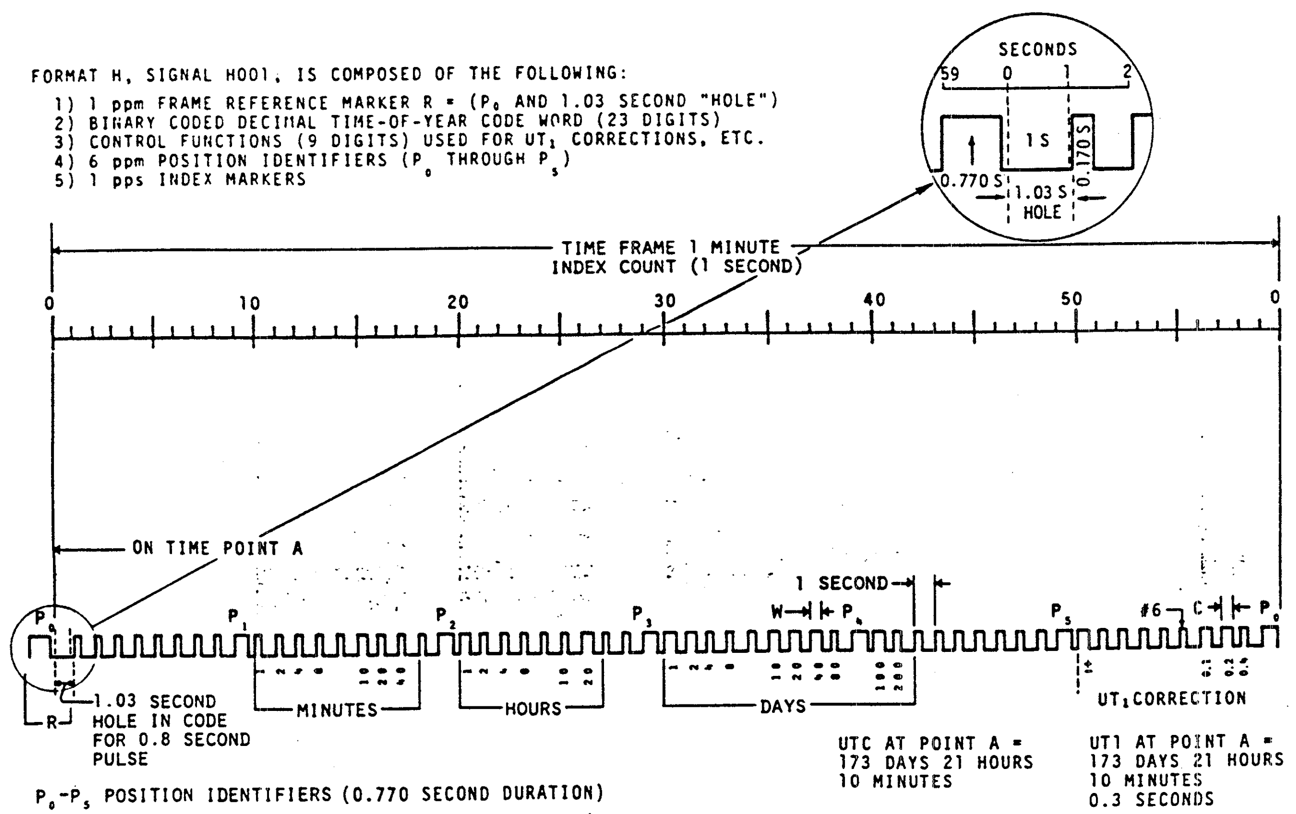




- BEGINNING OF EACH HOUR IS IDENTIFIED BY 0.8-SECOND LONG, 1500-Hz TONE.
- BEGINNING OF EACH MINUTE IS IDENTIFIED BY 0.8-SECOND LONG, 1000-Hz TONE.
- THE 29th & 59th SECOND PULSE OF EACH MINUTE IS OMITTED.

FORMAT H, SIGNAL H001, IS COMPOSED OF THE FOLLOWING:

1) 1 ppm FRAME REFERENCE MARKER R = (P_0 AND 1.03 SECOND "HOLE")
2) BINARY CODED DECIMAL TIME-OF-YEAR CODE WORD (23 DIGITS)
3) CONTROL FUNCTIONS (9 DIGITS) USED FOR UT_1 CORRECTIONS, ETC.
4) 6 ppm POSITION IDENTIFIERS (P_0 THROUGH P_5)
5) 1 pps INDEX MARKERS



P_0-P_5 POSITION IDENTIFIERS (0.770 SECOND DURATION)

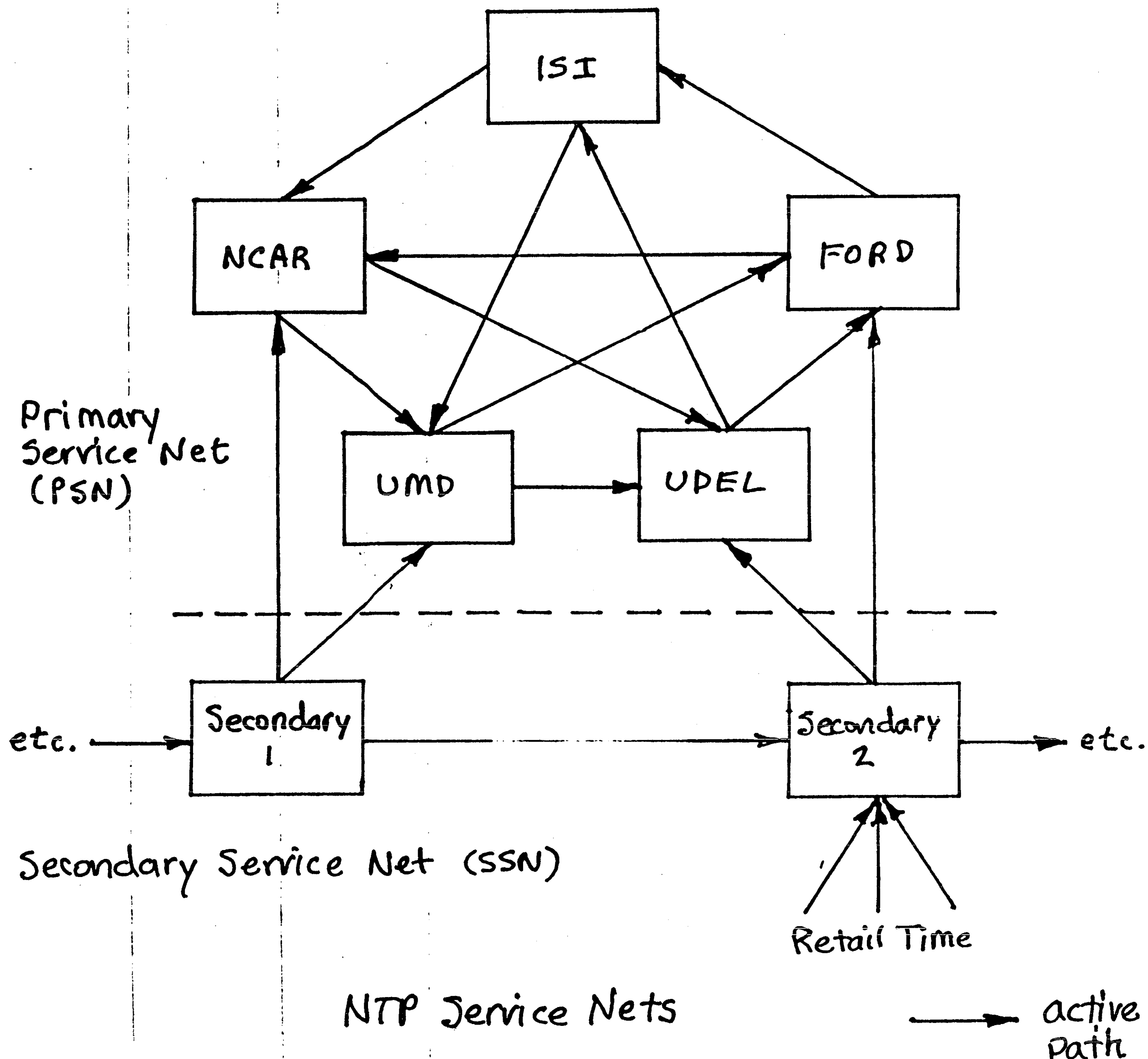
W WEIGHTED CODE DIGIT (0.470 SECOND DURATION)

C WEIGHTED CONTROL ELEMENT (0.470 SECOND DURATION) CONTROL FUNCTION #6 {BINARY ONE DURING 'DAYLIGHT' TIME / BINARY ZERO DURING 'STANDARD' TIME

DURATION OF INDEX MARKERS, UNWEIGHTED CODE, AND UNWEIGHTED CONTROL ELEMENTS = 0.170 SECONDS

NOTE: <u>BEGINNING</u> OF PULSE IS REPRESENTED BY POSITIVE-GOING EDGE.

9/75

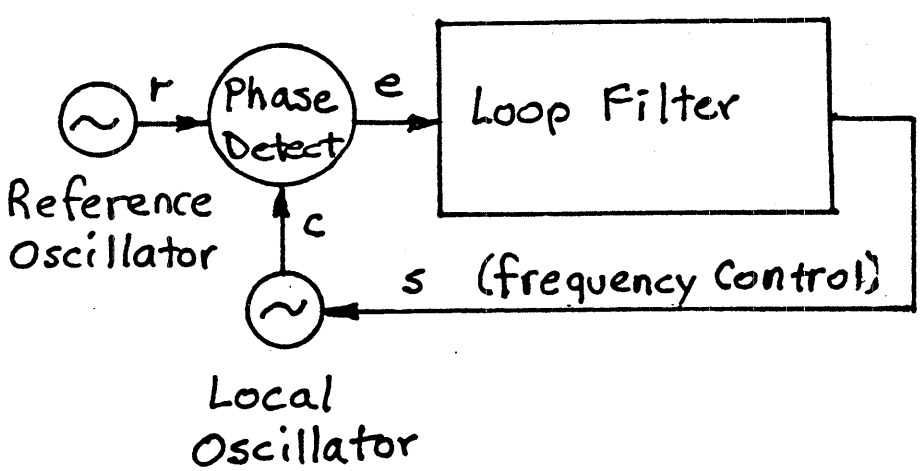


NTP Service Nets

Peer 1		Peer 2
t_1	------------>	t_2
t_4	<------------	t_3
	. . .	
t_{i-3}	------------->	t_{i-2}
t_i	<-------------	t_{i-1}

$$\text{delay} = (t_i - t_{i-3}) - (t_{i-1} - t_{i-2})$$

$$\text{offset} = [(t_{i-2} - t_{i-3}) + (t_{i-1} - t_i)]/2$$



- Primary server is LSI-11 CPU with disk (for support and monitoring) running Fuzzball operating system designed for highest accuracy (typically 1 ms relative to primary reference)

- Primary clock derived via NBS LF radio (WWVB) or UHF satellite (GOES); backup clock derived via NBS HF radio (WWV/WWVH)

- Normal synchronization is via primary or backup clock or, in case of failure, is via other primary servers or secondary/backup servers

- Completely connected tolopogy for robustness
  - PSN can survive loss of up to four radio clocks while delivering reliable time to all customers
  - Surviving PSN continues service as long as a single synchronization path is available to a radio clock
  - PSN delivers reliable time when a clock or server turns falseticker, even when another pimary server is lost

Primary Service Network (PSN)

- Secondary servers include both Fuzzball and Unix 4.3bsd with ntpd NTP daemon
- Normal synchronization is via either of two PSN servers or, in case of failure, via another SSN server with different primary servers
- Non-completely connected topology for load sharing
  - Surviving SSN continues service as long as a single synchronization path is available to a radio clock
  - SSN server delivers reliable time for all failure modes except when both primary servers turn falseticker

Secondary Service Network (SSN)

- o Distributed, multiple-process, multiple-host organization

- o Self-organizing subnetwork
  - Minimum spanning tree rooted on primary servers
  - Distributed Bellman-Ford routing algorithm
  - Metric based on stratum and delay
  - Synchronizes only to equal or greater stratum

- o Symmetric datagram protocol
  - Based on periodic, variable-rate polling (64-1024 s, depending on sample quality)
  - Does not require reliable delivery, sequencing or duplicate detection
  - Uses simple association management for state variables (timestamps, polling variables)

- o Time scale
  - Synchronized to Atomic Time (TA) on 1 January 1972
  - Corrected to UTC by NBS radio WWVB, GOES
  - NTP timestamp format 32-bit integer part plus 32-bit fraction part, zero corresponds to 0000 hours UTC January 1900, precision 0.2 ns, maximum 136 years

- o Time distribution
  - Returnable time (reversible)
  - Automatic distribution of leap-second corrections
  - Hierarchical master-slave by stratum:
    - 0 unknown (LAN synchronized)
    - 1 primary (independently synchronized)
    - 2..n secondary (NTP synchronized)

NTP Characteristics

- NTP produces a continuous sequence of samples $< d_i , c_i >$, where d_i is the measured delay and c_i the measured clock offset

- The clock filter algorithms operate on a window of k samples $[< d_i , c_i > , < d_{i-1} , c_{i-1} > , \ldots , < d_{i-k+1} , c_{i-k+1} >]$ saved in a shift register ok k stages

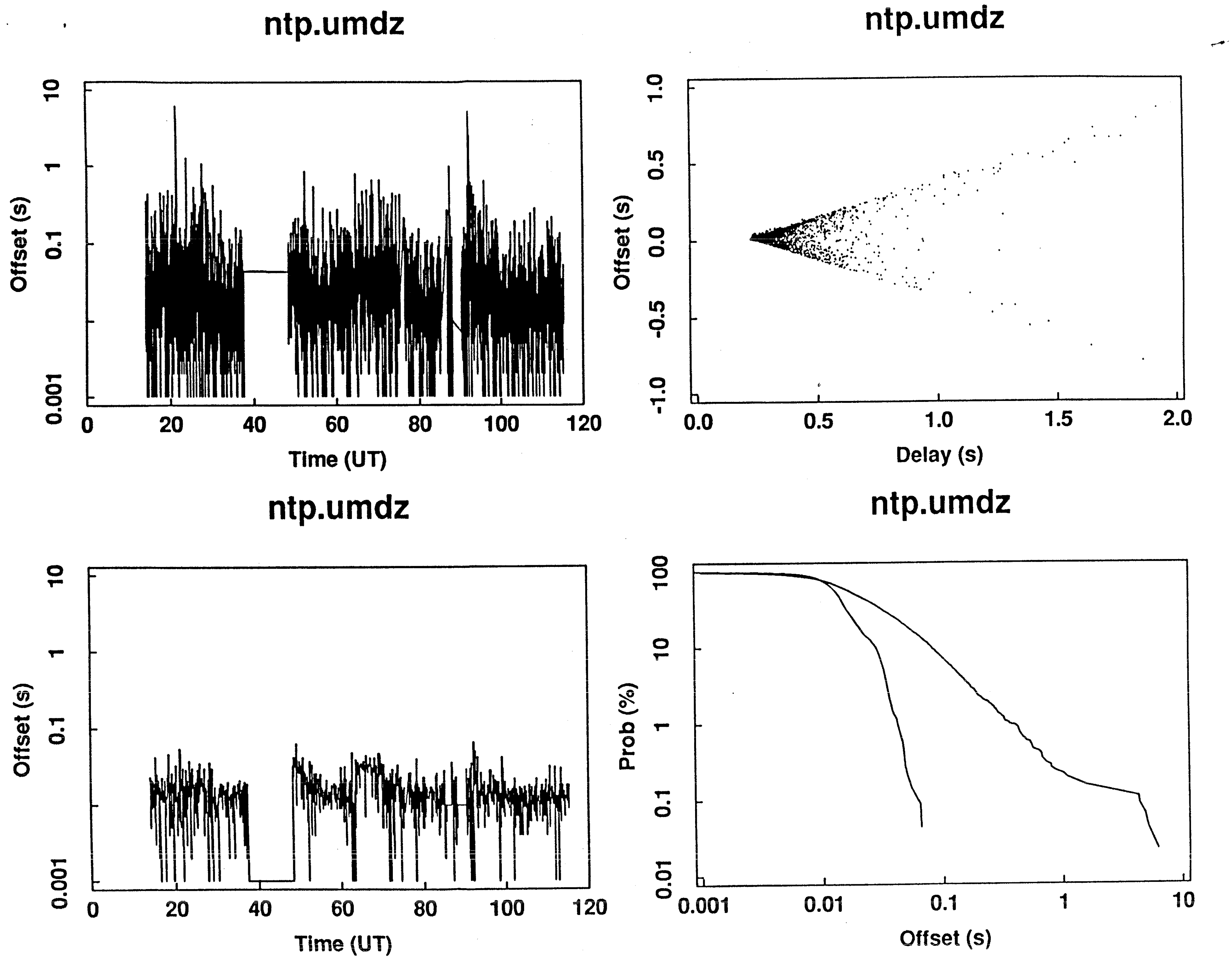
- Mean filter
  - Output mean of offset samples as offset estimate
  - Does not use delay samples
  - Is vulnerable to occasional large excursions in offset

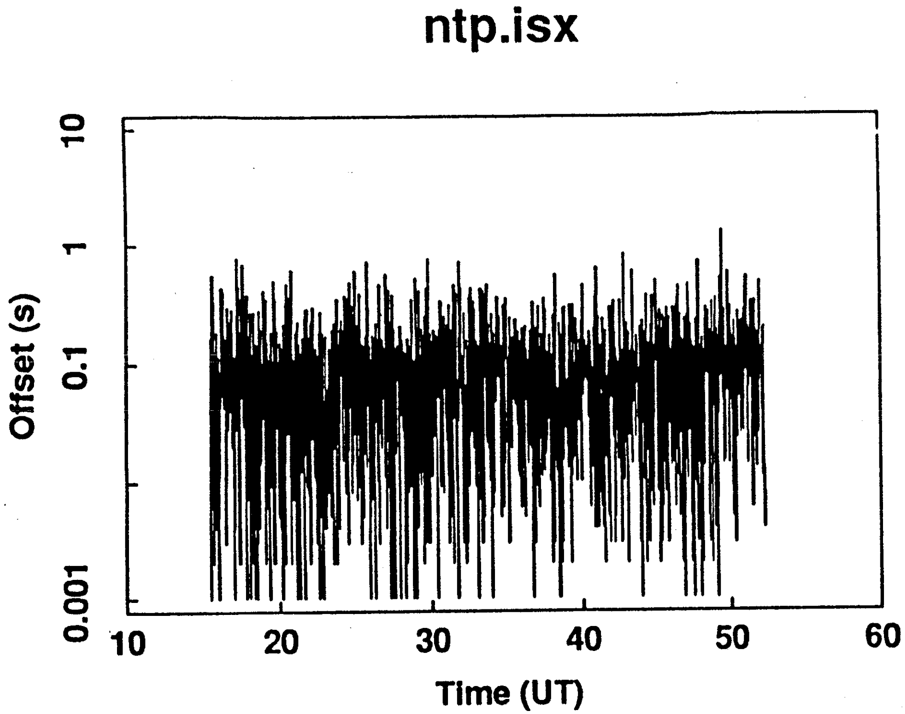
- Median filter
  - Output median of offset samples as offset estimate
  - Does not use delay samples
  - Experiments show this results in disappointing accuracy

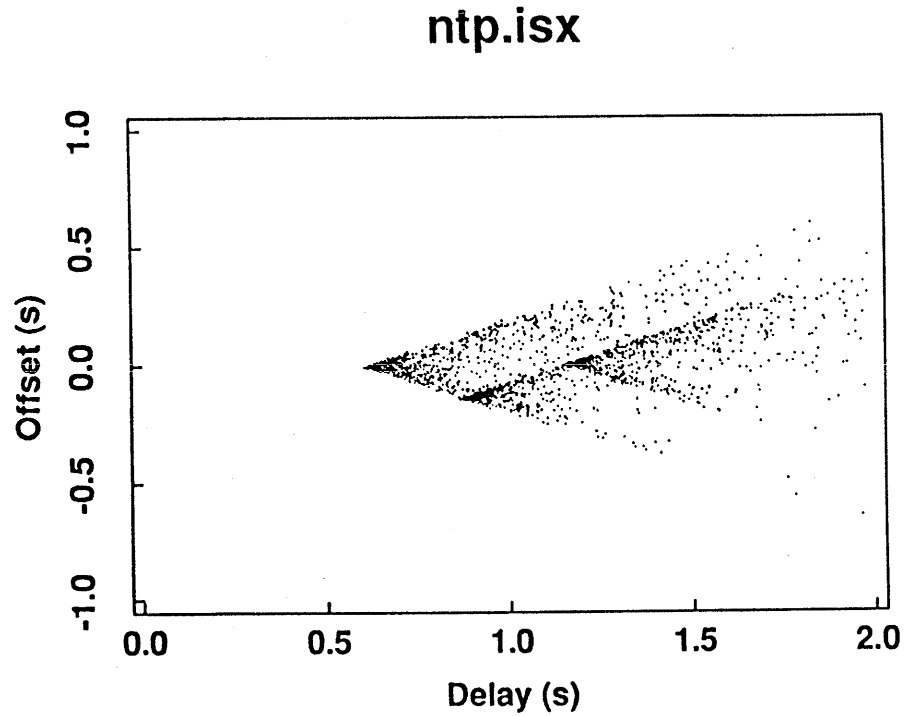
- Modified median filter (old Fuzzball algorithm)
  - Compute median of remaining samples in the shift register, discard extreme outlyer and repeat until only one left
  - Output remaining sample as offset estimate
  - Experiments show accuracy can be improved

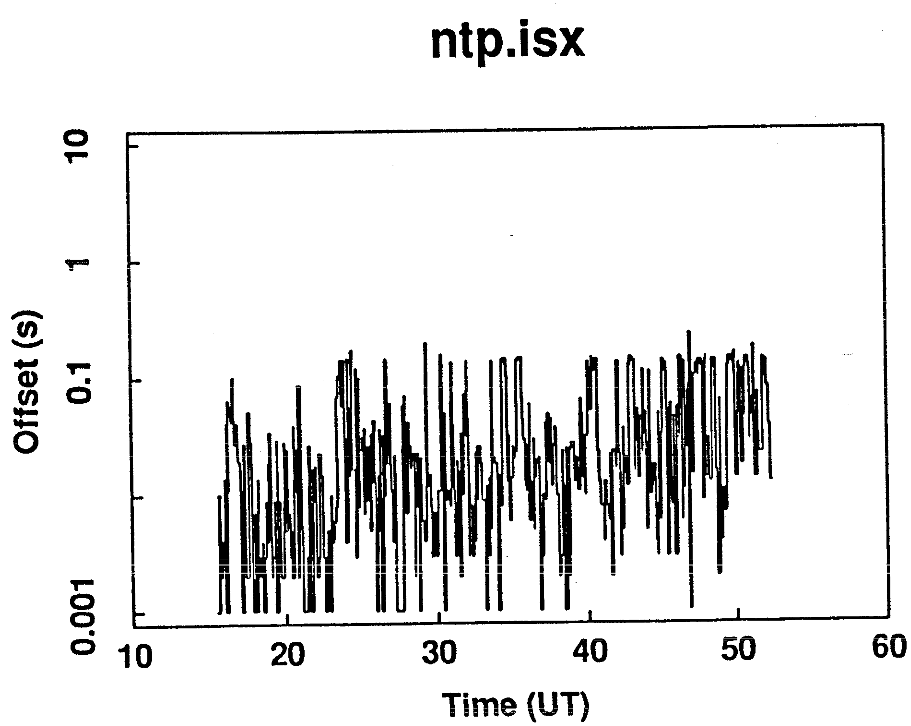
- Minimum filter (new Fuzzball algorithm)
  - Sort $< d_i , c_i >$ pairs in order of increasing d_i
  - Output c_0 of first pair as offset estimate C
  - Output $\sum_{i=0 \ldots k-1} (| d_0 - d_i | w^i)$ as dispersion estimate S
  - Output suppressed unless $D < T$ threshold
  - Present system uses $w = 2$, $T = 500$

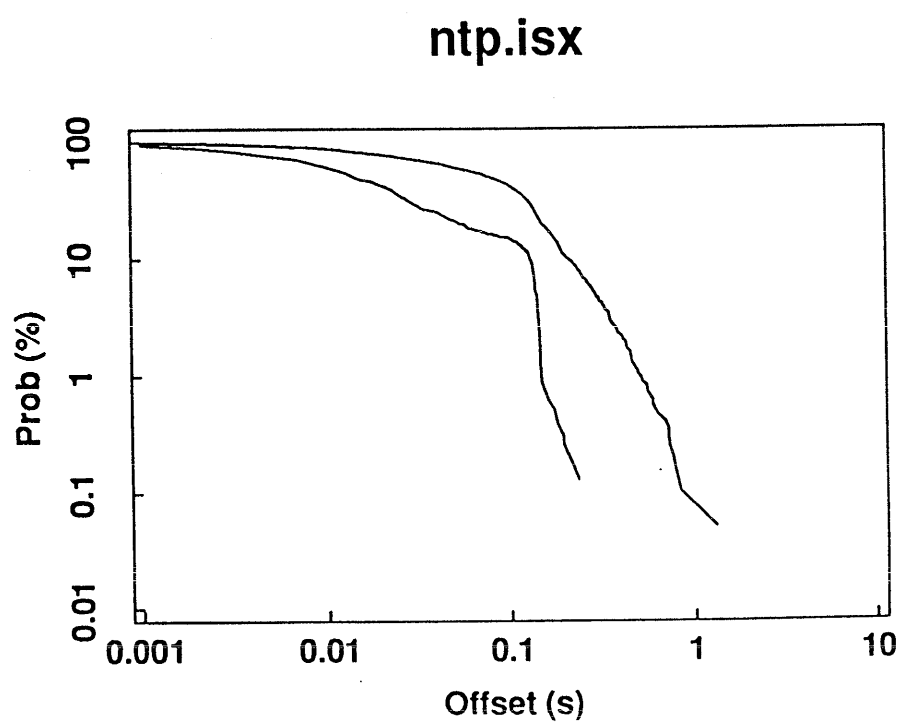
Clock Filter Algorithms


ntp.umdz
Offset (s)
10
1
0.1
0.001
0
20
40
60
80
100
120
Time (UT)
ntp.umdz
Offset (s)
1.0
0.5
0.0
-0.5
-1.0
0.0
0.5
1.0
1.5
2.0
Delay (s)
ntp.umdz
Offset (s)
10
1
0.1
0.001
0
20
40
60
80
100
120
Time (UT)
ntp.umdz
Prob (%)
100
10
1
0.1
0.01
0.001
0.01
0.1
1
10
Offset (s)

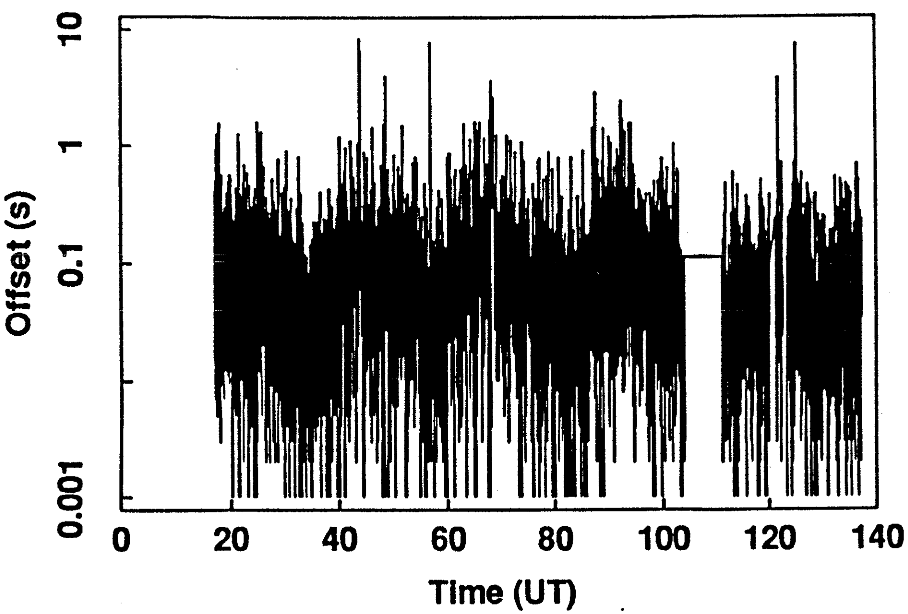

ntp.isx
Offset (s)
10
1
0.1
0.001
10
20
30
40
50
60
Time (UT)


ntp.isx
Offset (s)
1.0
0.5
0.0
-0.5
-1.0
0.0
0.5
1.0
1.5
2.0
Delay (s)

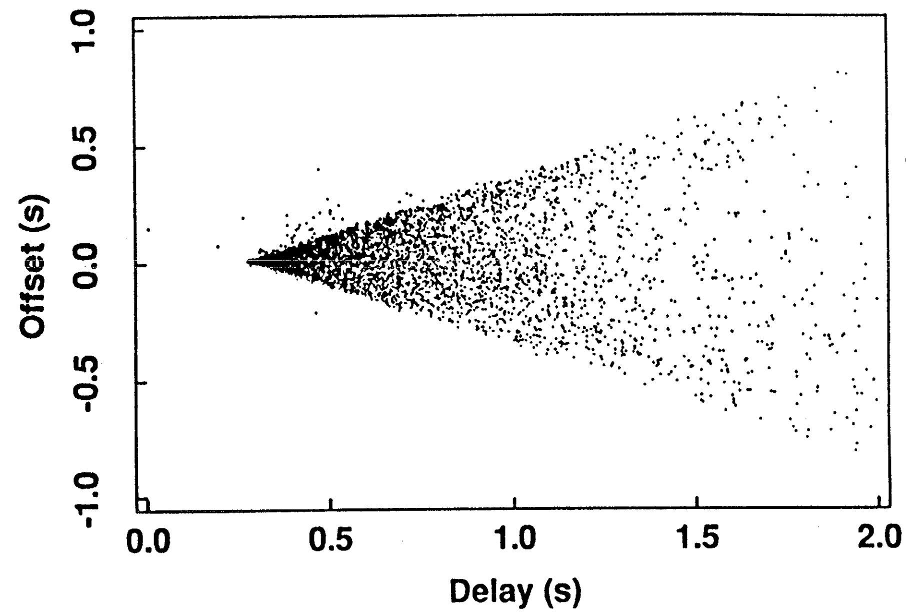

ntp.isx
Offset (s)
10
1
0.1
0.001
10
20
30
40
50
60
Time (UT)


ntp.isx
Prob (%)
100
10
1
0.1
0.01
0.001
0.01
0.1
1
10
Offset (s)

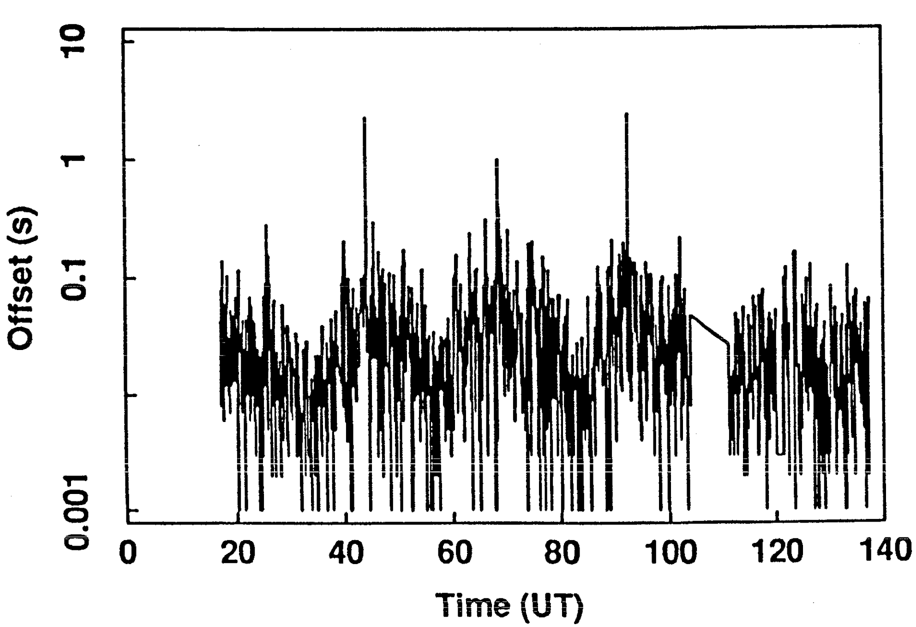
ntp.pscy



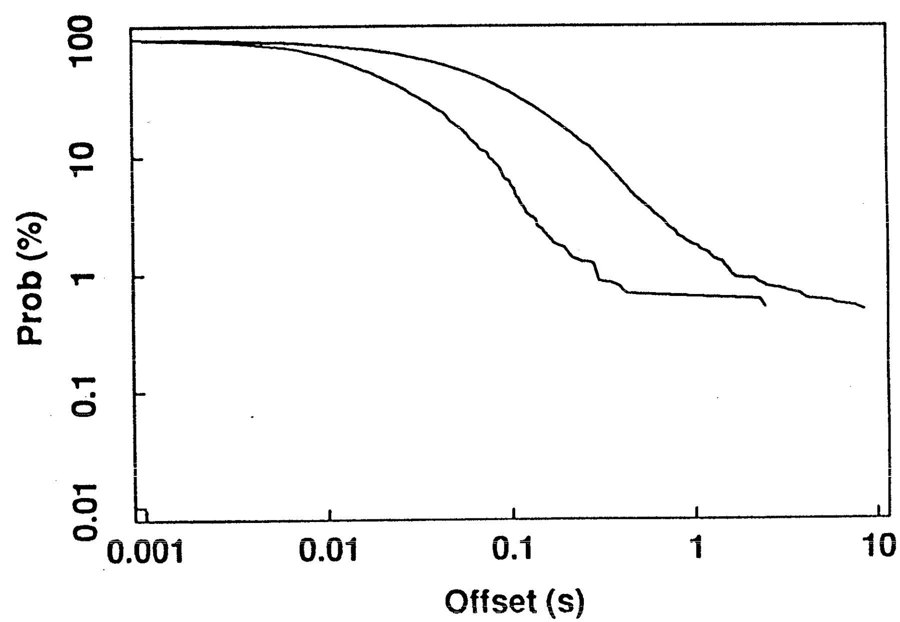
ntp.pscy



ntp.pscy



ntp.pscy



- Clock filter algorithm produces offset estimates Cj for each of p clocks

- Clock selection algorithm selects candidate clocks on the basis of reasonable criteria

- Each clock asigned a sixteen-bit sort key K_j
  - High-order three bits are current stratum
  - Low-order thirteen bits are current total delay (delay computed to clock plus its delay to primary server)

- Pairs $< C_j , K_j >$ are saved in a list L and sorted in order of increasing K_j

- For each pair j remaining in the list of size q calculate

  $$\underset{i = 0 \ldots q-1}{\text{sum}} (| C_j - C_i | w^i) \text{ as dispersion of } j$$

  - Discard clock with highest dispersion and repeat until only a single clock left
  - Output offset of surviving clock as best estimate
  - Present system uses w = 0.75, which is chosen so that an ambiguity between two clocks at a stratum can be resolved by a clock at the next lower stratum

Clock Selection Algorithm

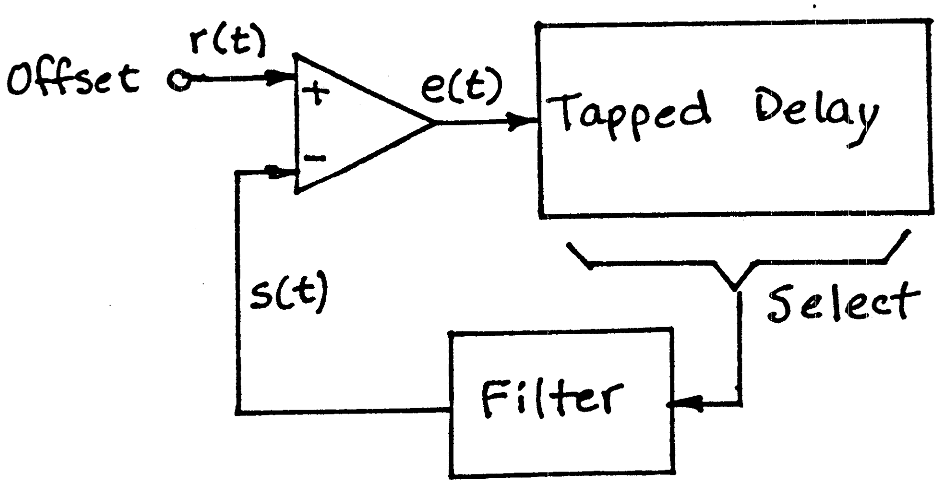
- o UTC time-of-day in 1-ms increments, wraps at 2400 hours; UTC day relative to 1 January 1972

- o Disciplined oscillator uses first-order phase-lock loop
  - Optimized for crystal-stabilized and mains-derived clocks
  - Implemented with several types of clock interfaces in Fuzzball and also in Unix 4.3bsd ntpd daemon

- o Typical error LAN paths 1 ms, Internet paths 20 ms

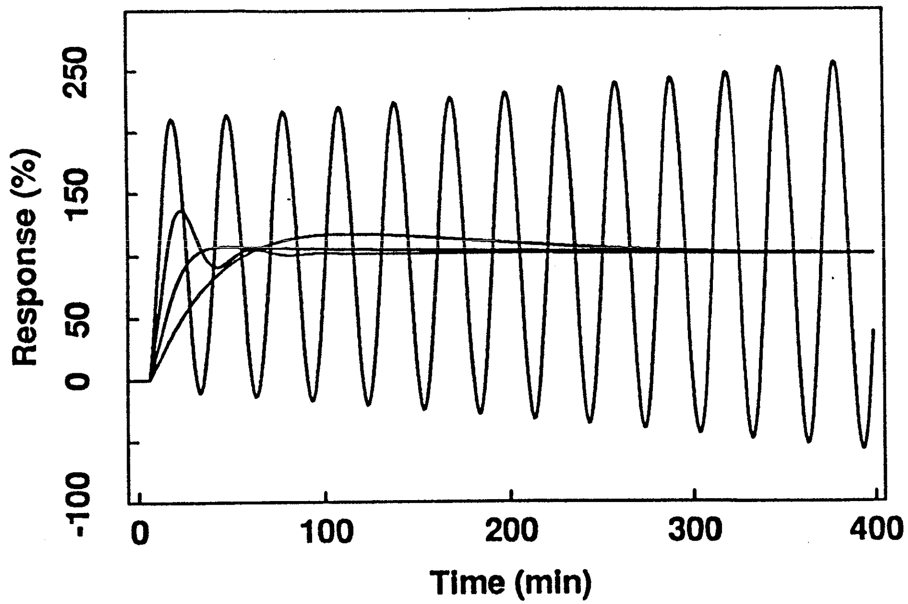
- o Max drift 1 ppm (86 ms/day), typical drift <0.1 ppm

Local Clock Algorithm

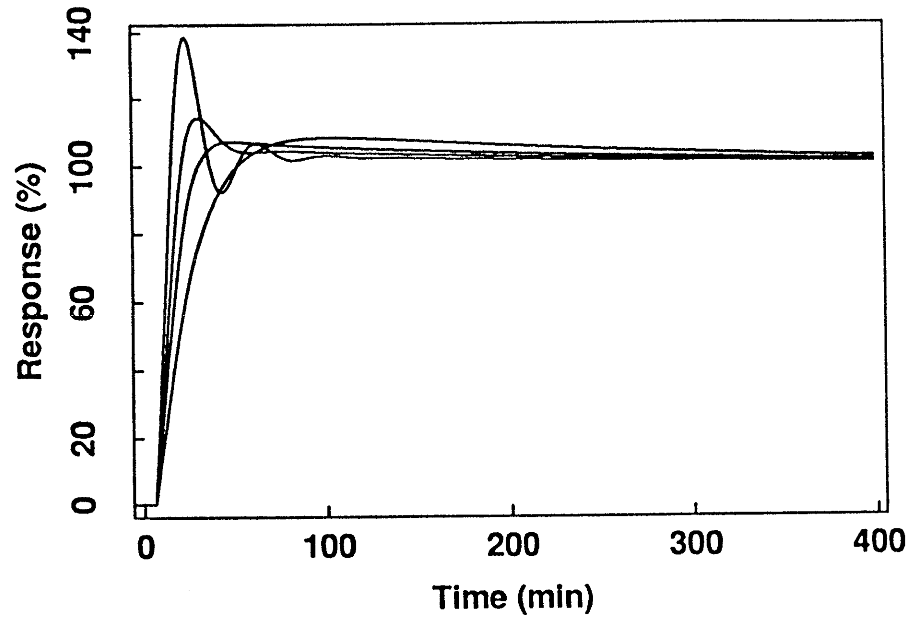


$$s(t) = a\,e(t-\tau) + b\int_{\tau}^{t} e(y-\tau)\,dy$$

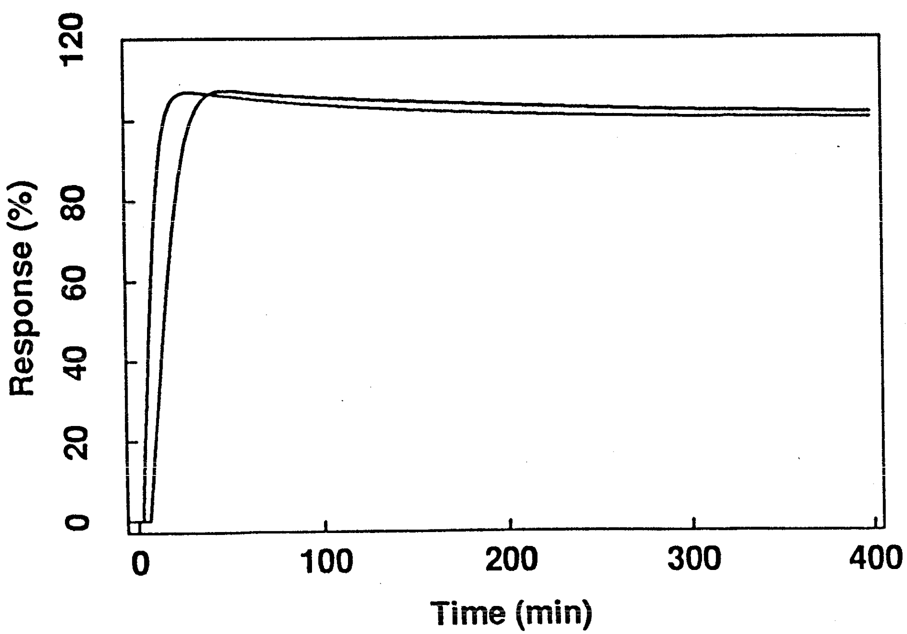
filt.1



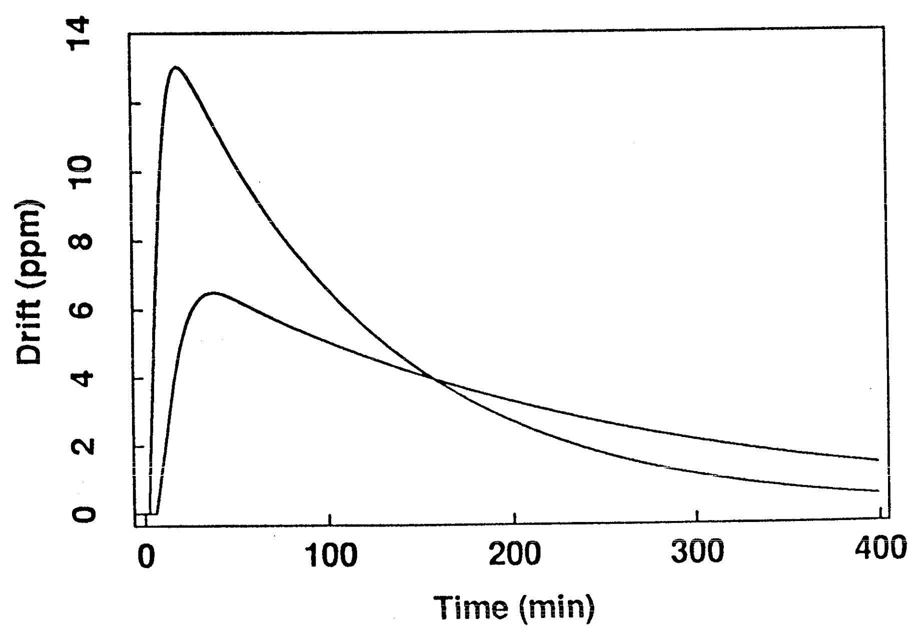
filt.3



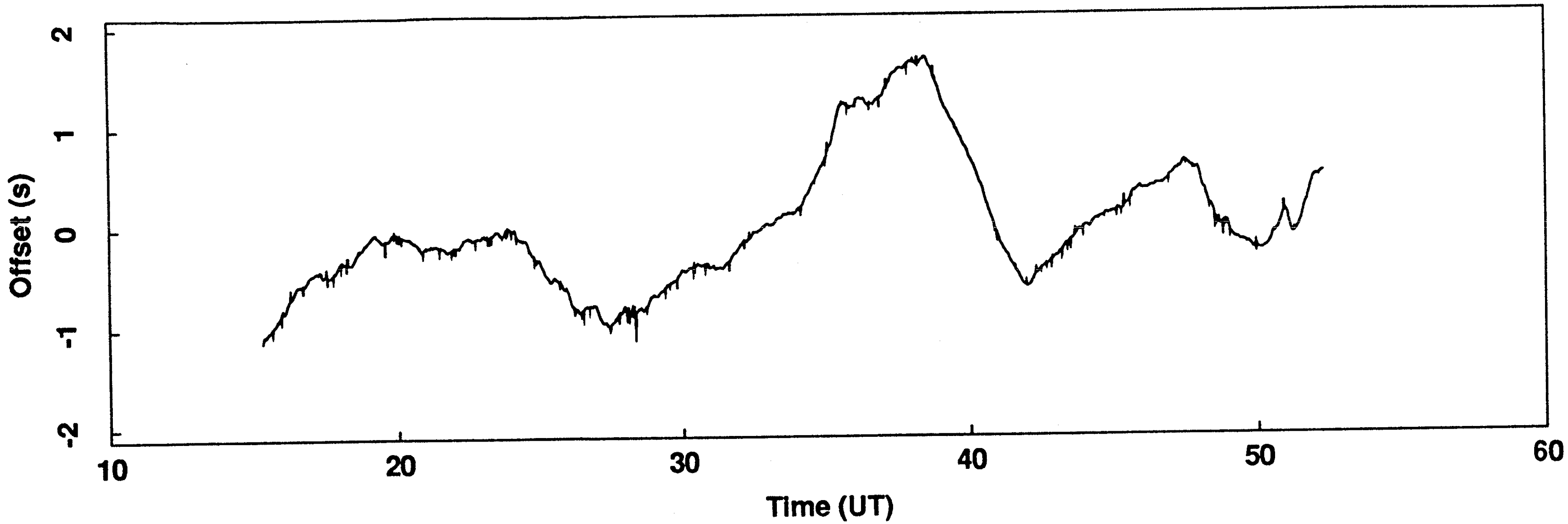
filt.4



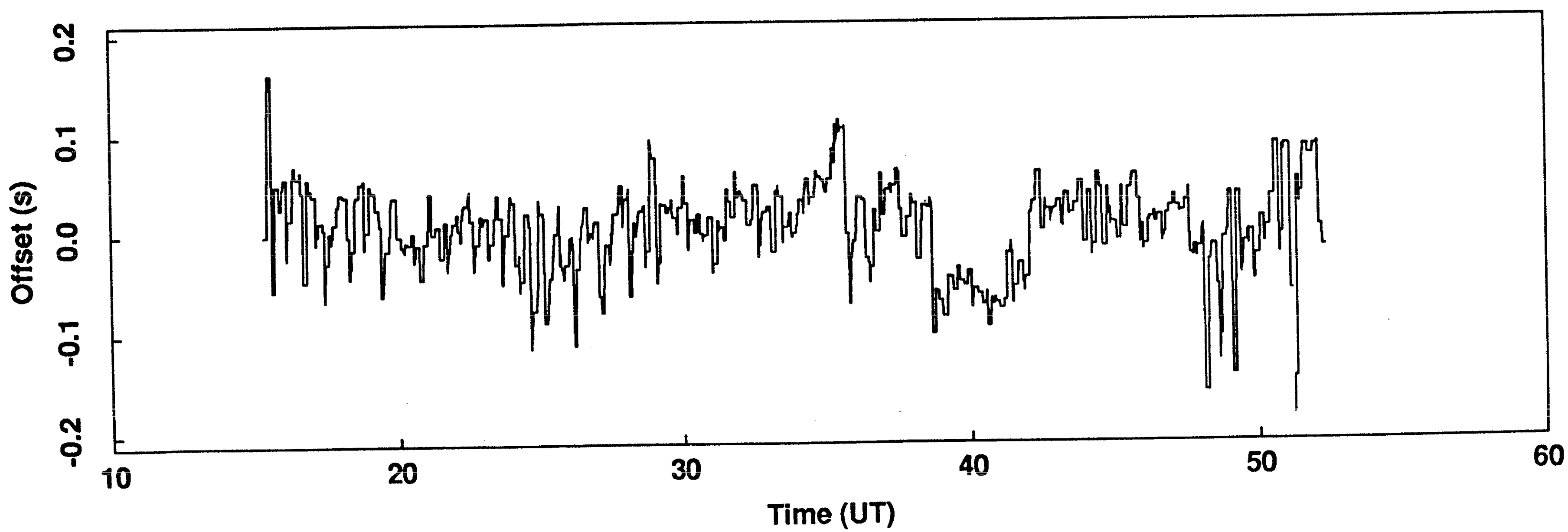
filt.4



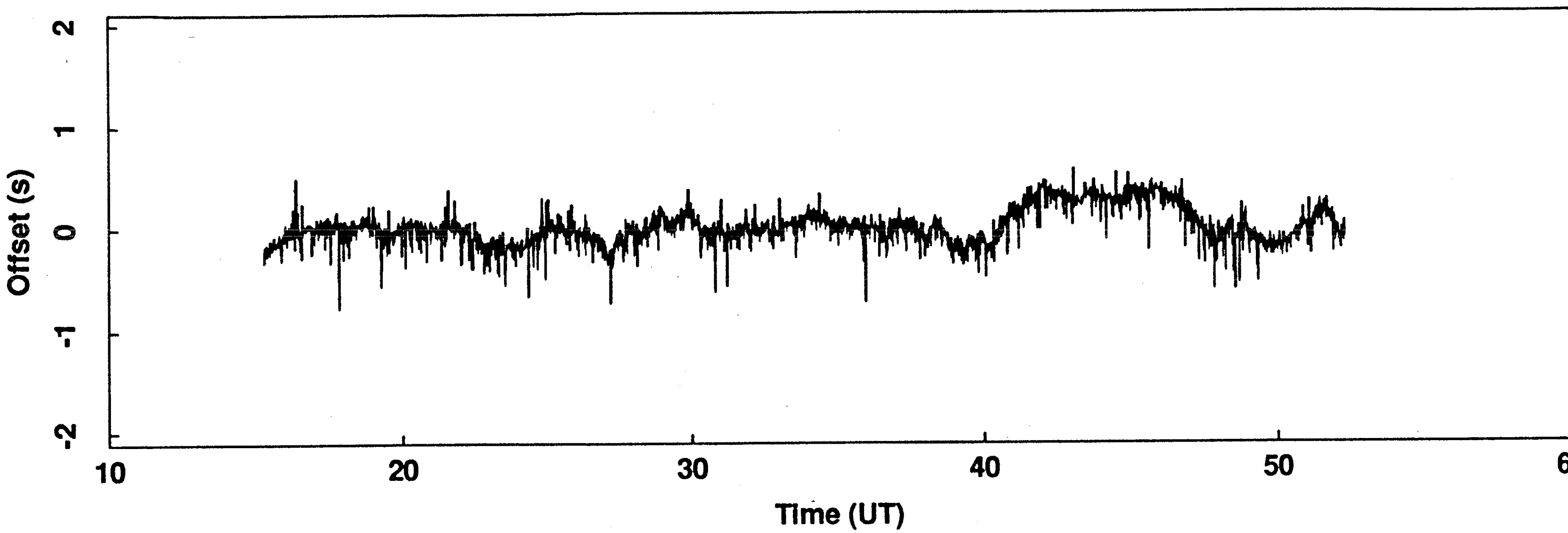
min.eas



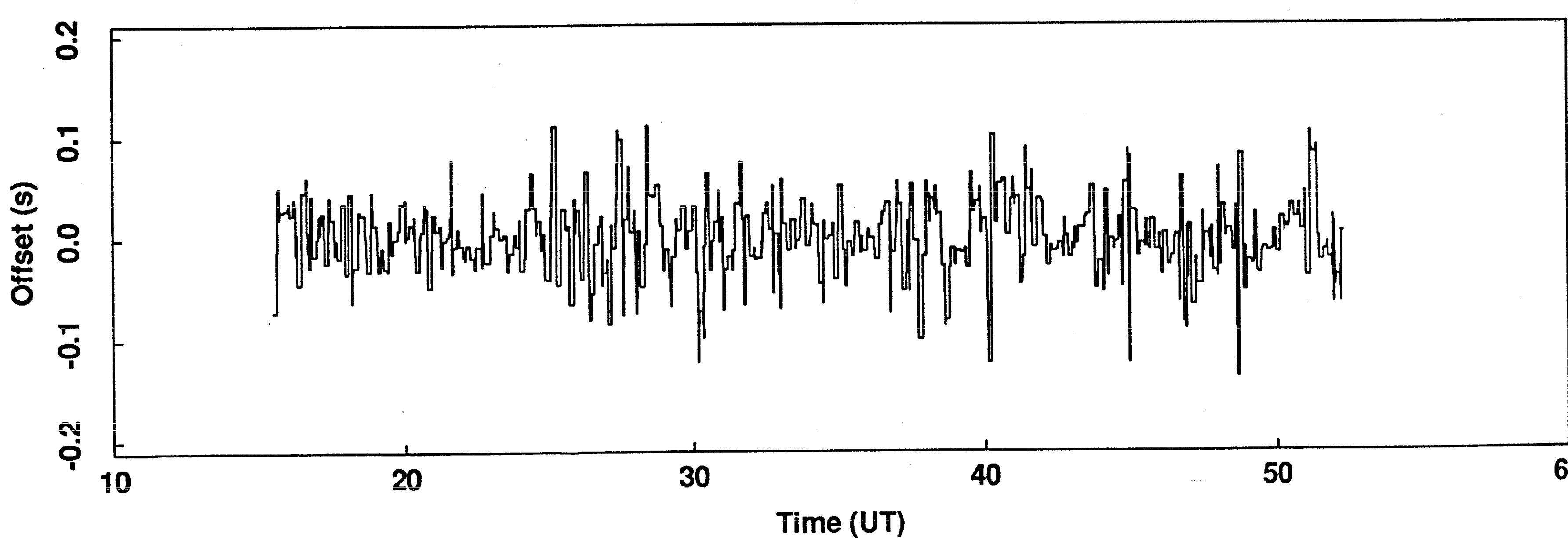
min.eas



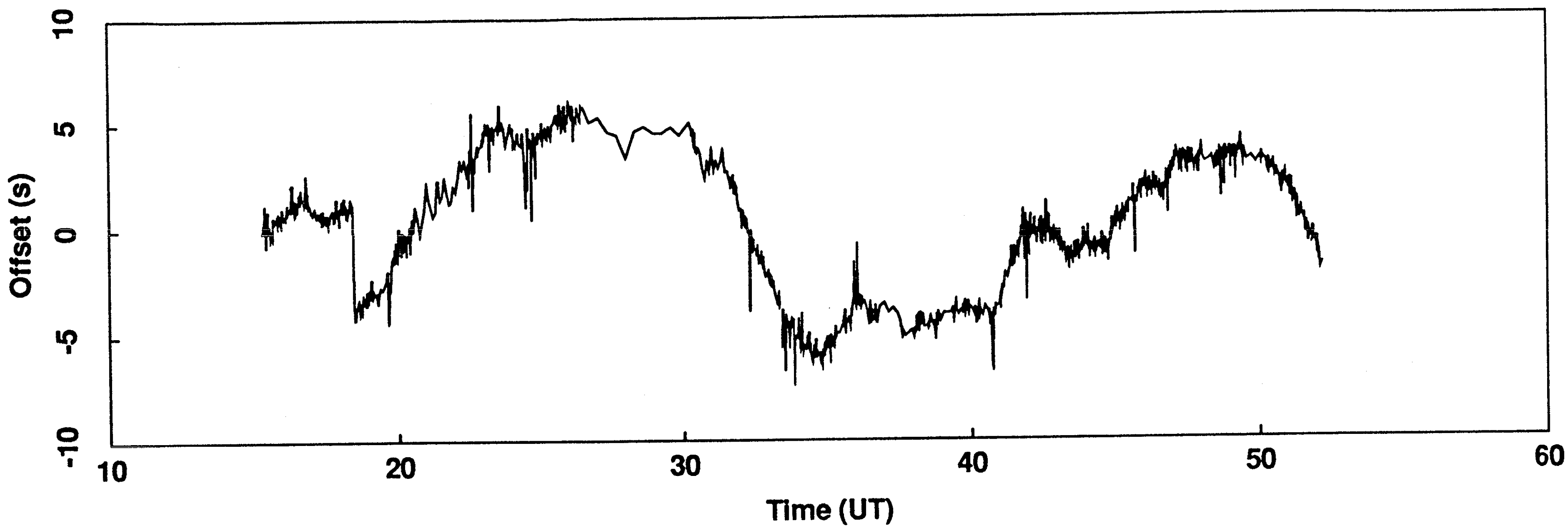
min.cal



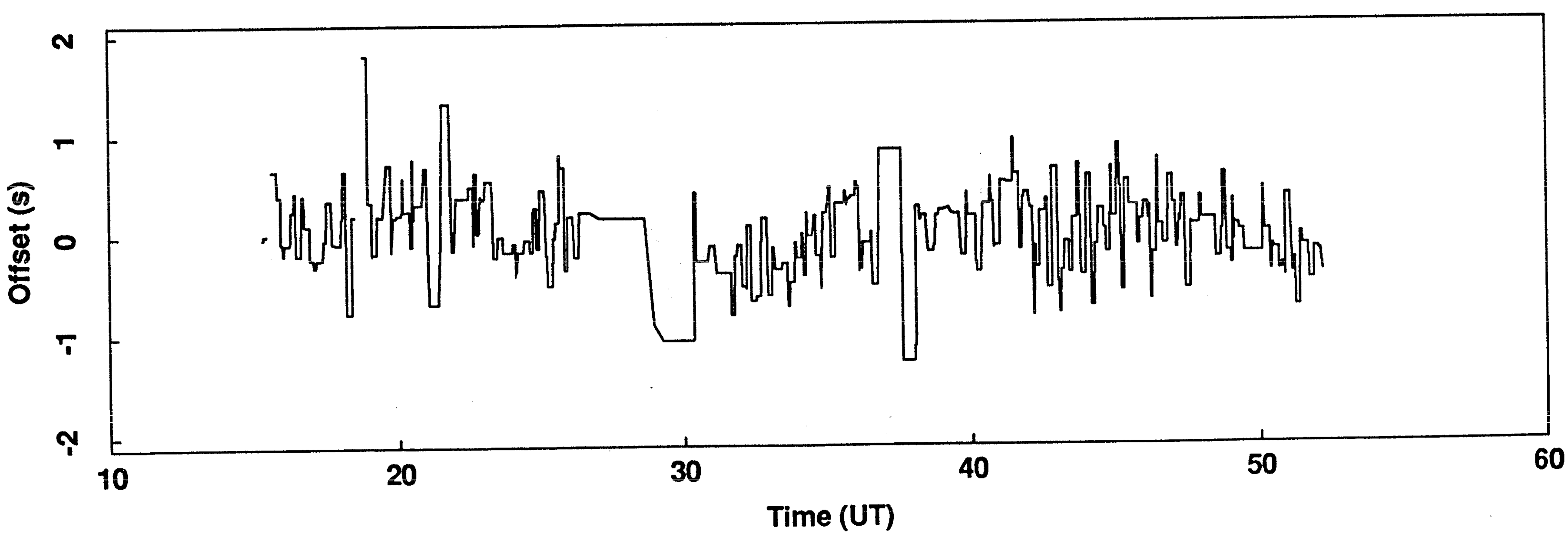
min.cal

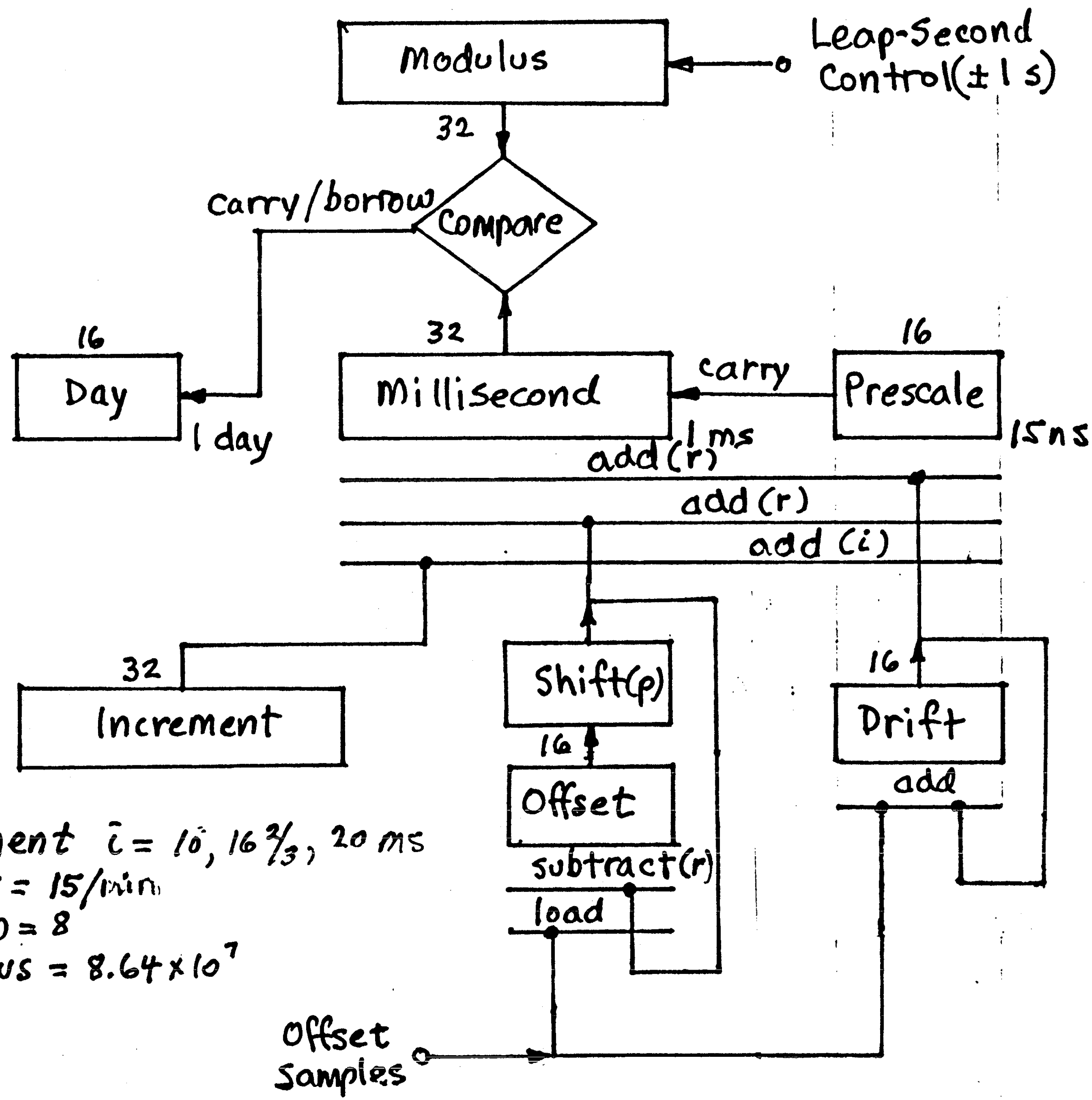


min.eur



min.eur





increment $i = 10, 16\frac{2}{3}, 20$ ms
rate $r = 15$/min
shift $p = 8$
modulus $= 8.64 \times 10^7$

Local Clock

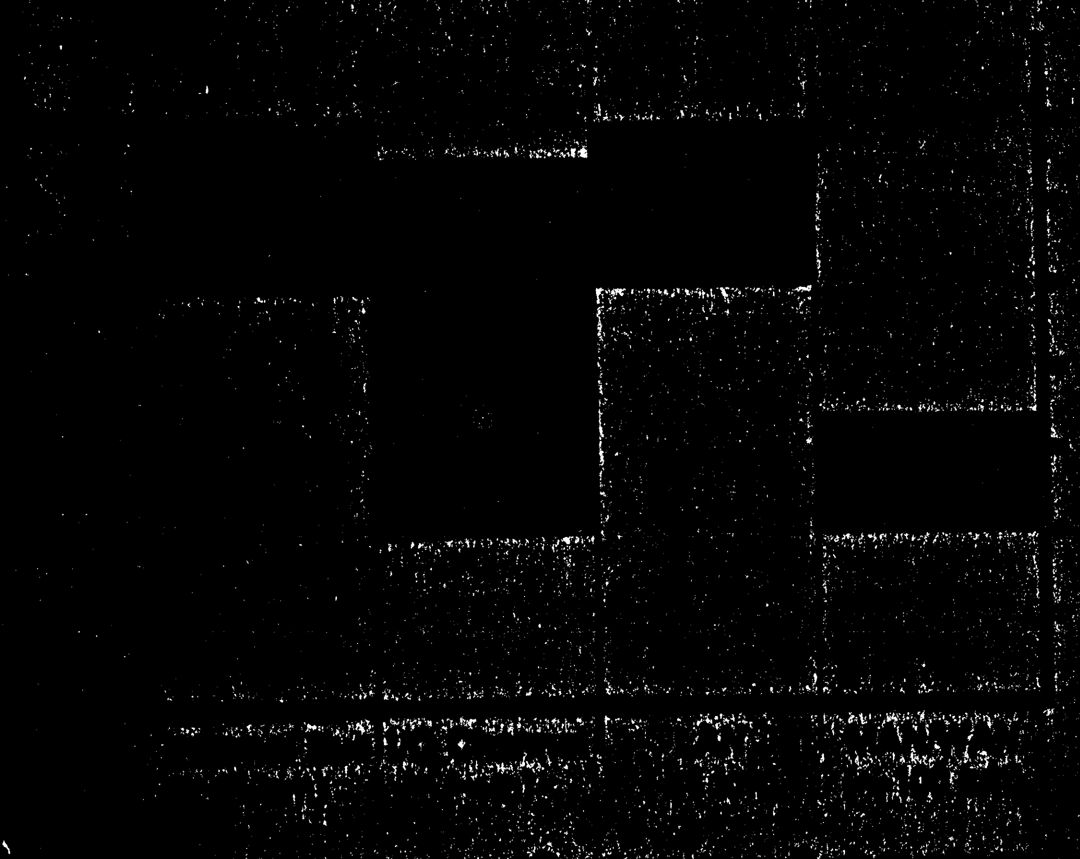
7.11 Switched Multi-megabit Data Service—Kramer, NYNEX

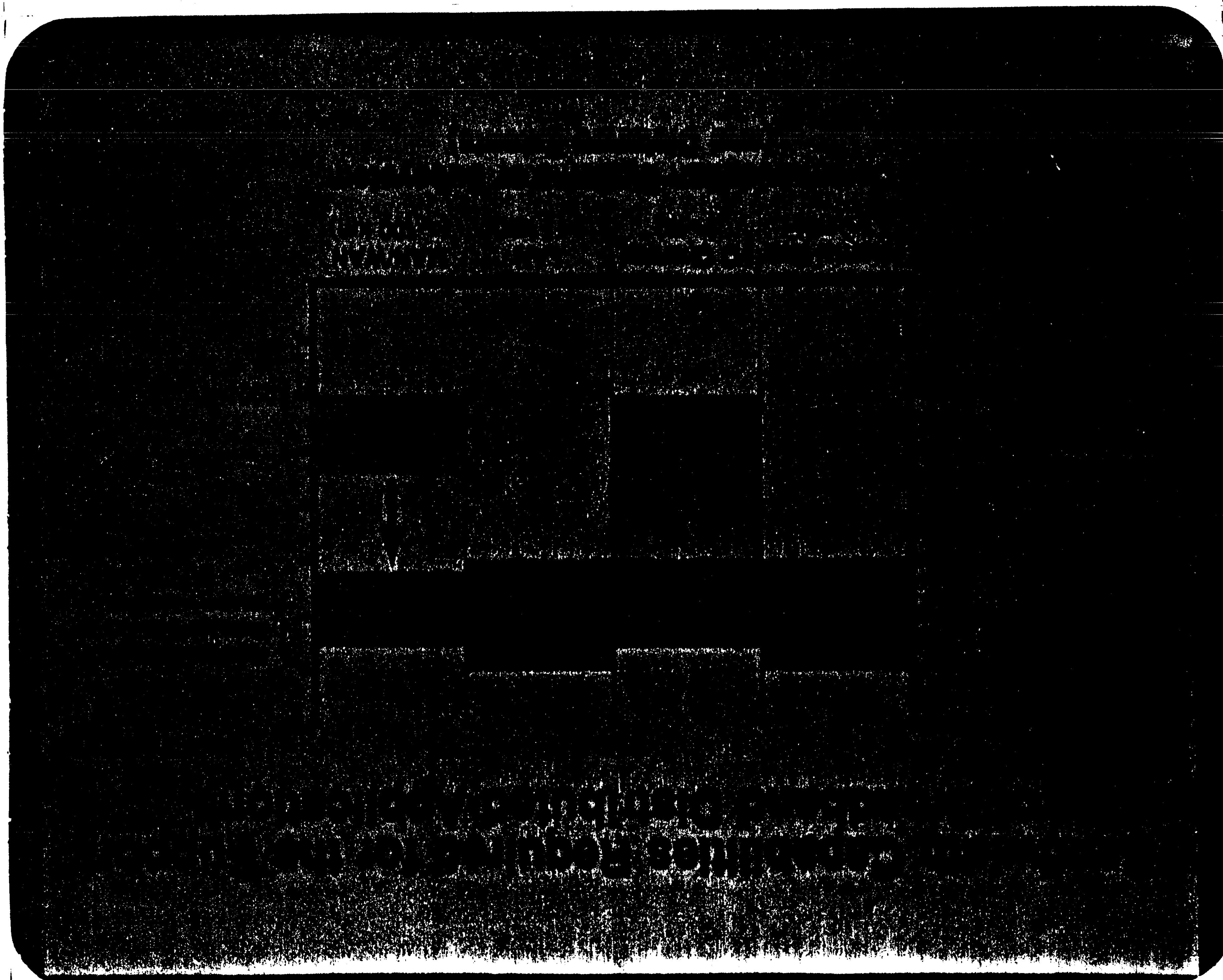
SWITCHED MULTI-MEGABIT DATA SERVICE
(S M D S)

Michael Kramer
Broadband Communications and Services Laboratory
NYNEX Advanced Technology Development

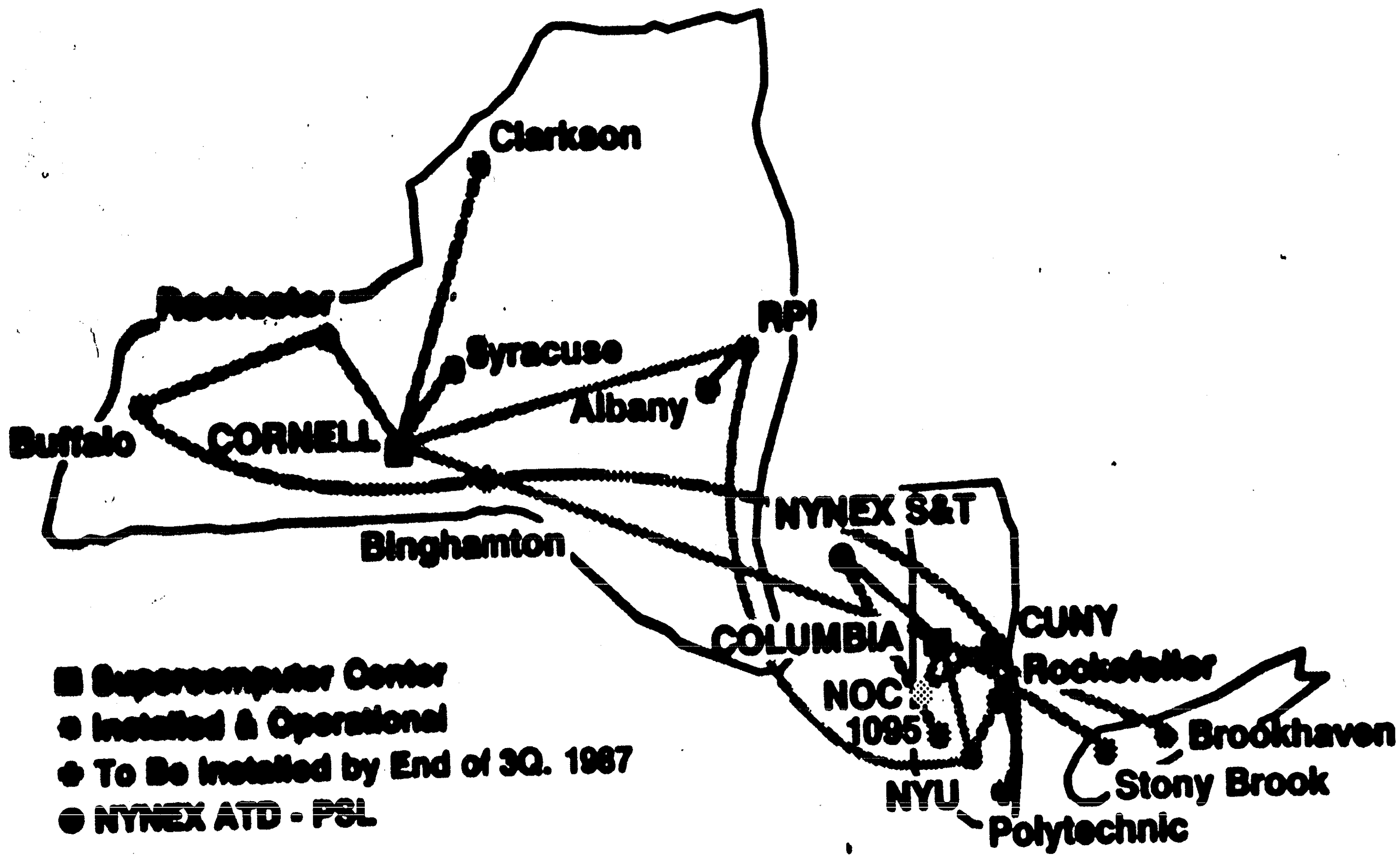
May 27, 1988

Bandwidth Capabilities of
... Communications Sub-Systems





NYSERNet
New York State Education and Research Network



PUBLIC BROADBAND DATA SERVICE REQUIREMENTS

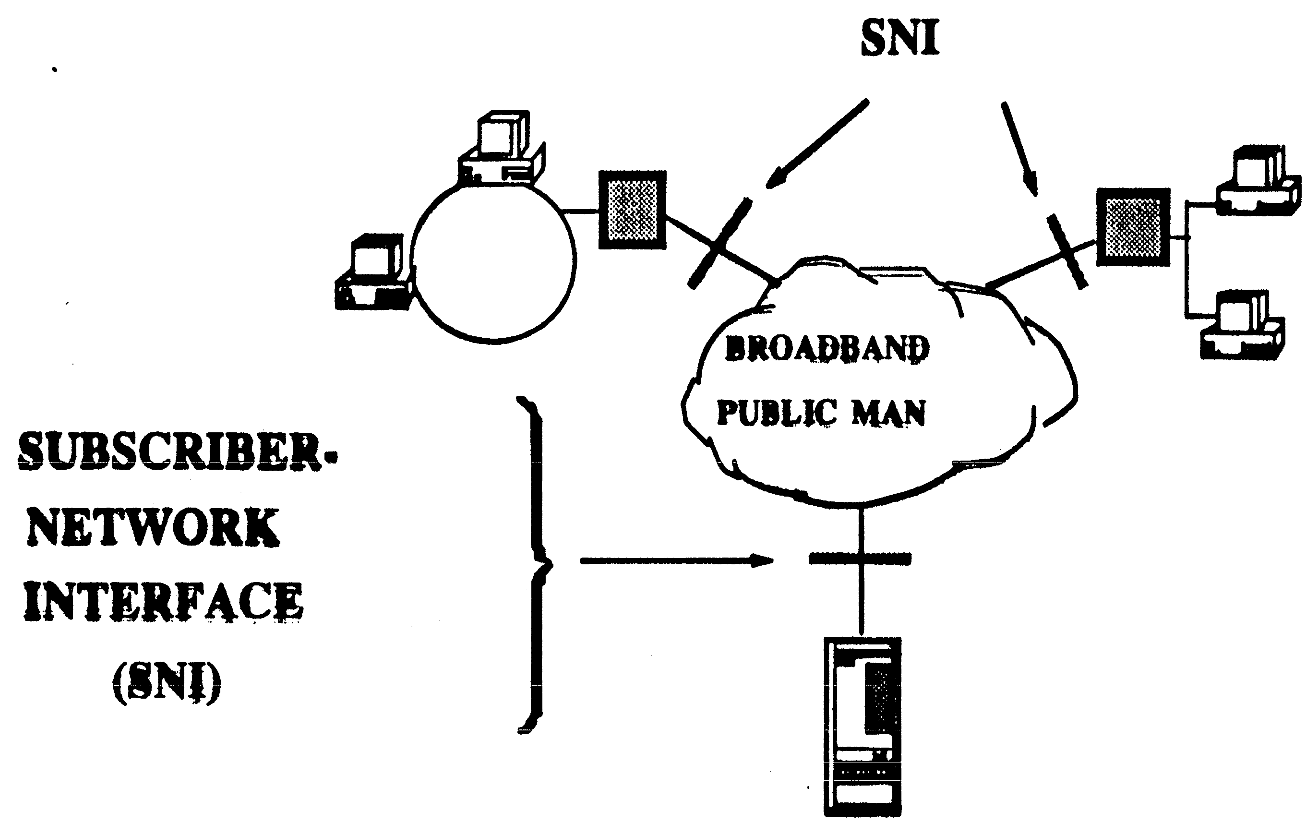
- Public Network Architecture.
- "LAN-like performance over a Metropolitan Area".
- Simple interface to end systems and the end-user.
- Allow for easy integration into customer's existing applications.
- Stimulate the development of new applications
- Suitable for early service introduction.
- Evolvable to a WAN Service

Broadband Network

WHAT IS SMDS ?

- A service concept (not a technology) for public, packet switched high speed data.
- Supportable by several technologies and architectures including "early availability technologies".
- Near term service capability which is evolvable to Broadband ISDN.
- Connectionless Packet Data Service
- Provide support for typical applications
  - LAN interconnection
  - workstation to host communication
  - host to host communication

BROADBAND MAN SERVICES ARCHITECTURE



Broadband ISN

SMDS INTERFACE PROTOCOL
(S I P)

SIP LEVEL 3	Network Services Level Variable Length PDU (Maximum Size 8227 bytes)
SIP LEVEL 2	High performance DLC Provides Framing for Level 3 PDU Error detection (NOT Correction)
SIP LEVEL 1	Physical (Transmission) Interface Based on standard hierarchy DS3 interface

SIP LEVEL 3 PDU

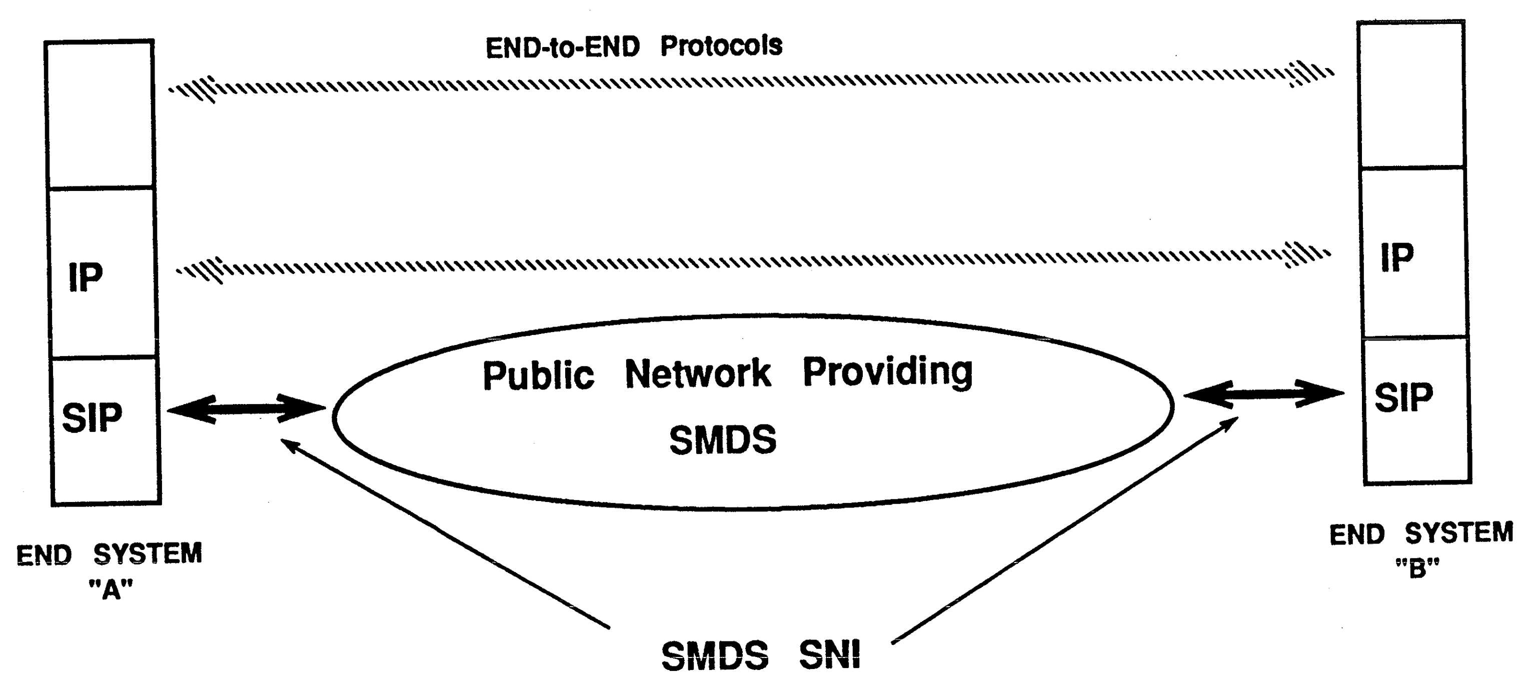
Control Indicator	Destination Address	Source Address	Reserved	Carrier Indicator	Carrier Select	User Data
(1 octet)	(8 octets)	(8 octets)	(2 octets)	(1 octet)	(2 octets)	(≤ 8191 octets)

Control Indicator → Provision for Group Addressing

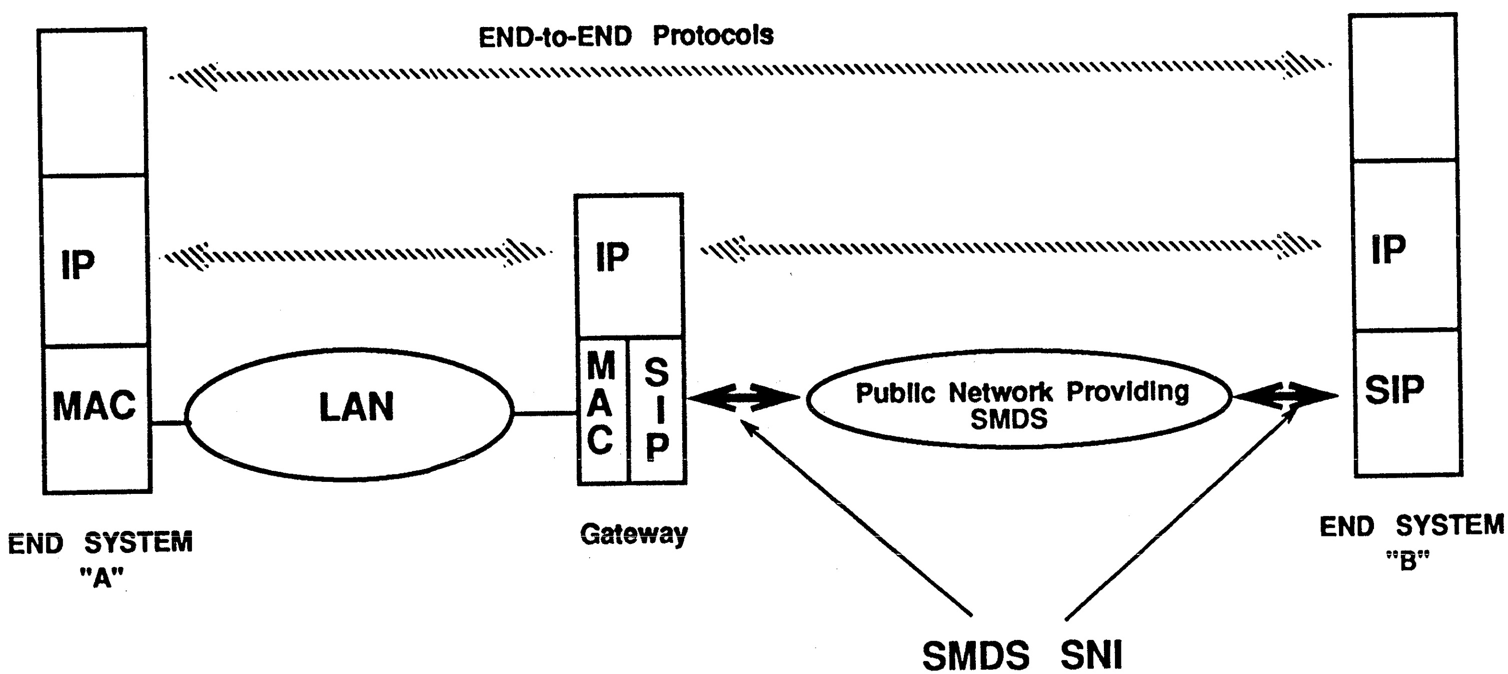
Destination Address and Source Address: **CCITT E.164 (ISDN) Addresses** (Up to 15 BCD encoded digits)

Carrier Indicator and Carrier Select: For Inter-LATA routing utilizing Inter-Exchange Carriers (IEC's)

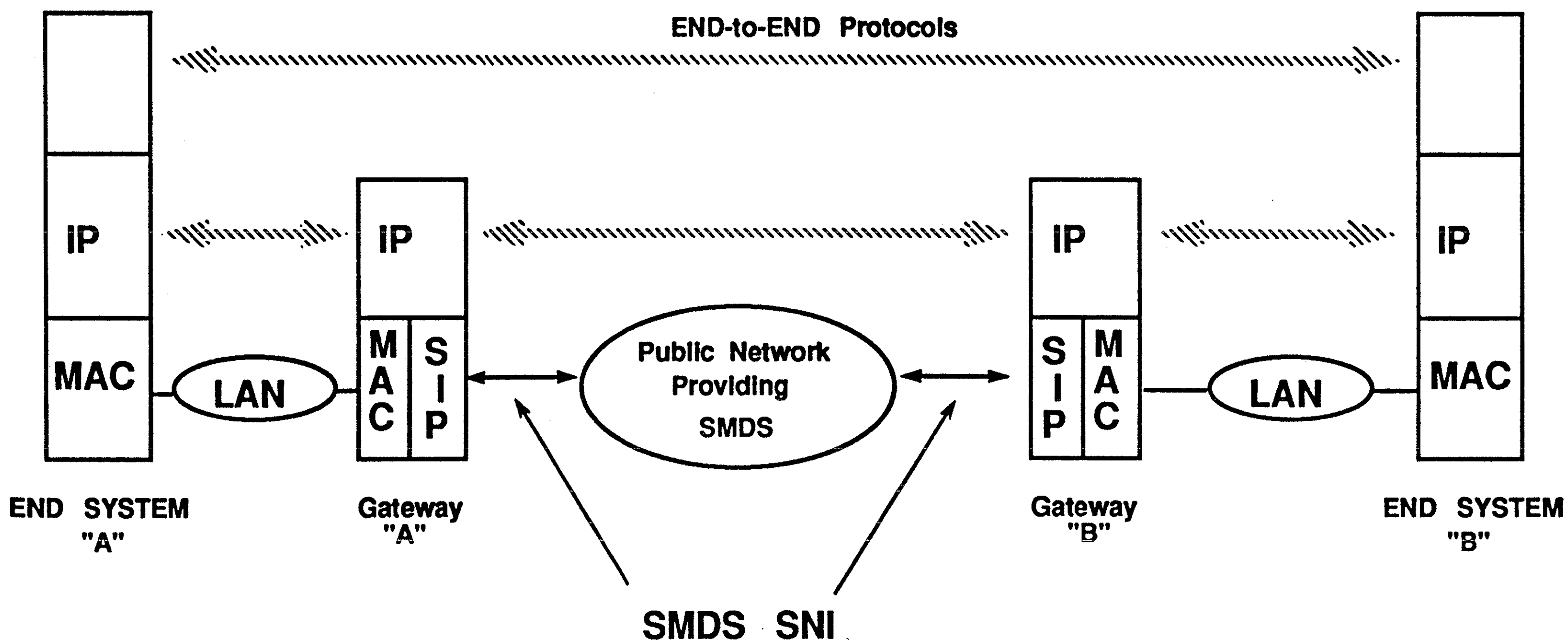
A View of SMDS from the End System Perspective



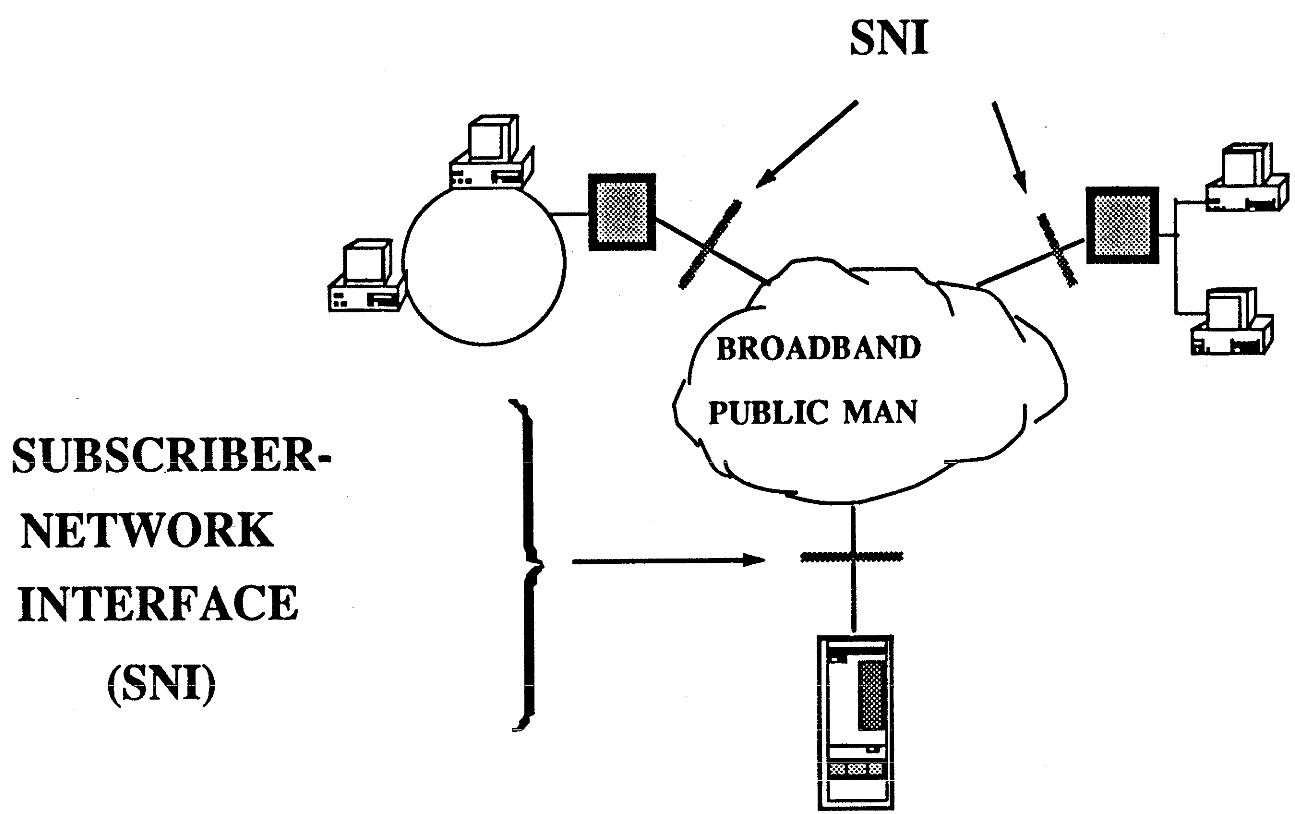
A View of SMDS from the End System Perspective
SMDS Subnet Role (II)



A View of SMDS from the End System Perspective
SMDS Network Subnet Role



BROADBAND MAN SERVICES ARCHITECTURE



ACCESS CLASS ENFORCEMENT

At each SNI:
A Credit Manager controls the rate of information flow into and out of the network.

For each direction of flow:
A Credit Balance, C, is regulated by a set of Flow Enforcement Parameters

Each customer subscribes to an "Access Class" as defined by this set of Flow Enforcement Parameters.

Flow Parameters characterize the ***average information flow rate*** and the maximum allowed ***burstiness of flow*** across the SNI

ACCESS CLASS CREDIT MANAGER

IU = An information unit

IU_size = number of octets / IU

C = Credit Balance, measured in IUs

Δt = Credit increment interval, measured in seconds

C_{max} = Maximum value that credit balance is allowed to attain

} **ACCESS CLASS**

C_{max} controls burstiness

Δt controls the maximum sustained information rate

ACCESS CLASS CREDIT MANAGER - An example

IU_size = 8192 octets / IU

Δt = 2 ms

C_{max} = 128 IU's

MaxBurst ≅ 1 Megabyte
Sustained Flow Rate (SFR) ≅ [illegible] 33 Mbps

Example 2:

IU_size = 8192 octets / IU

Δt = 250 ms

C_{max} = 128 IU's

MaxBurst ≅ 1 Megabyte

SFR ≅ $\frac{1}{4}$ Mb/s

ADDRESSING

- CCITT E.164 (ISDN) Adressing
- Address Validation
- Group Addressing
- Address Screening:

  Security / Virtual Private Network Applications

  •source address screening

  •destination address screening

SMDS: Some Additional Aspects

- minimum performance objectives (delay, lost packets, etc.)

- requirements for a network supporting SMDS
  - -MSS definition
  - -internal interfaces- IMSSI, SI
  - -operations and maintenance
  - -internal protocol architecture

- SIP level 2
- Transmission specifications

Broadband Network

7.12 Performance and Congestion—Mankin, MITRE

Performance and Congestion Control W.G.

Paper - Interim Draft in Group yesterday - work continue by E-mail

Goals - Document performance pitfalls

Describe practices

Current Outline -

Intro: Improved Performance in a Computer Network

- Framework (No numbers)
- Measurement methods

Congestion Handling in Internet Gateways

" " " " Hosts

Performance Issues in DNS

" " " Telnet

Mailing list - Work on paper

ccpaper@gateway.mitre.org

7.13 Domains—Mamakos, UMD

6/14/88

DOMAINS AND HOSTS
REGISTERED WITH DDN NIC

Top-level domains	=	33
2nd-level domains	=	513
Hosts in.CA	=	2
Hosts in.COM	=	421
Hosts in .EDU	=	2436
Hosts in .GOV	=	325
Hosts in .IL	=	1
Hosts in .IT	=	3
Hosts in .MIL	=	199
Hosts in .NET	=	20
Hosts in .NL	=	2
Hosts in .NO	=	3
Hosts in .ORG	=	21
Hosts in .UK	=	11
Hosts in .US	=	1
Hosts still in .ARPA	=	2642

143 (net 10)

1729 (net 26)

770 (other nets)

6/9/88

DDN Growth

Network Naming and Addressing Statistics

	May 1987	May 1988	Increase
Internet Hosts (includes ARPA/MIL)	4,178	5,639	35%
ARPANET/MILNET Hosts	820	1717	110%
ARPANET/MILNET TACs	148	189	28%
ARPANET/MILNET GWs	134	180	34%
Internet Gateways (includes ARPA/MIL)	182	240	32%
ARPANET/MILNET Nodes	217	259	19%
Connected Networks	637	915	44%
Domains (top-level, 2nd-level)	328	546	67%
Hostmaster online mail	1231	1526	24%

(Size of current host table = 607,577 bytes)

Name Server stats
for TERP.UMD.EDU

Thu Jun 16 20:52:58 1988

36231	time since boot (secs)
36231	time since reset (secs)
30350	input packets
28409	output packets
28041	queries
3	iqueries
97	duplicate queries
2574	responses
2373	duplicate responses
14062	OK answers
13830	FAIL answers
1	FORMERR answers
1	system queries
8	prime cache calls
1	check_ns calls
0	bad responses dropped
2	martian responses
0	Unknown query types
11196	A querys
4405	NS querys
23	invalid(MF) querys
652	CNAME querys
157	SOA querys
3	WKS querys
6532	PTR querys
6	HINFO querys
1393	MX querys
110	AXFR querys
3563	ANY querys

Thu Jun 16 21:33:04 1988

		rate per second over the last 40 ~~[illegible]~~ minutes	644 minutes
38637	time since boot (secs)		
38637	time since reset (secs)		
32161	input packets	4.46	0.832
30073	output packets	4.09	0.778
29697	queries	4.07	0.768
3	iqueries		
104	duplicate queries	0.017	0.0026
2747	responses	0.426	0.071
2536	duplicate responses	0.4	0.0656
15126	OK answers	2.62	0.391
14412	FAIL answers	1.43	0.373
1	FORMERR answers		
1	system queries		
8	prime cache calls		
1	check_ns calls		
0	bad responses dropped		
2	martian responses		
0	Unknown query types		
11987	A querys	1.95	0.310
4645	NS querys	0.59	0.120
25	invalid(MF) querys		
	CNAME querys	0.07	0.0176
167	SOA querys	0.02	
3	WKS querys		
6802	PTR querys	0.665	0.176
6	HINFO querys		
1476	MX querys	0.204	0.038
118	AXFR querys		
3786	ANY querys	0.549	0.0979

7.14 SNMP Extensions—Rose, TWG

IETF SNMP EXTENSIONS WORKING GROUP:

Status Report

Marshall T. Rose
The Wollongong Group, Inc.

June 15–17, 1988

A NEW WORKING GROUP

- CHARTERED BY RFC1052
- UPDATE IDEA11:
  - TO ALIGN WITH THE OUTPUT OF THE MIB WORKING GROUP
  - TO MEET SHORT-TERM NETWORK MANAGEMENT NEEDS OF THE INTERNET
- FIRST (AND ONLY) MEETING HELD IN MAY.
- OUTPUT WAS MODIFIED IDEA (IDEA011-01) MEETING ABOVE GOALS
- WORKING GROUP WILL DISSOLVE ONCE NEW IDEA BECOMES RFC

7.15 NETMAN—LaBarre, MITRE

IETF NETMAN Working Group Report

Lee LaBarre

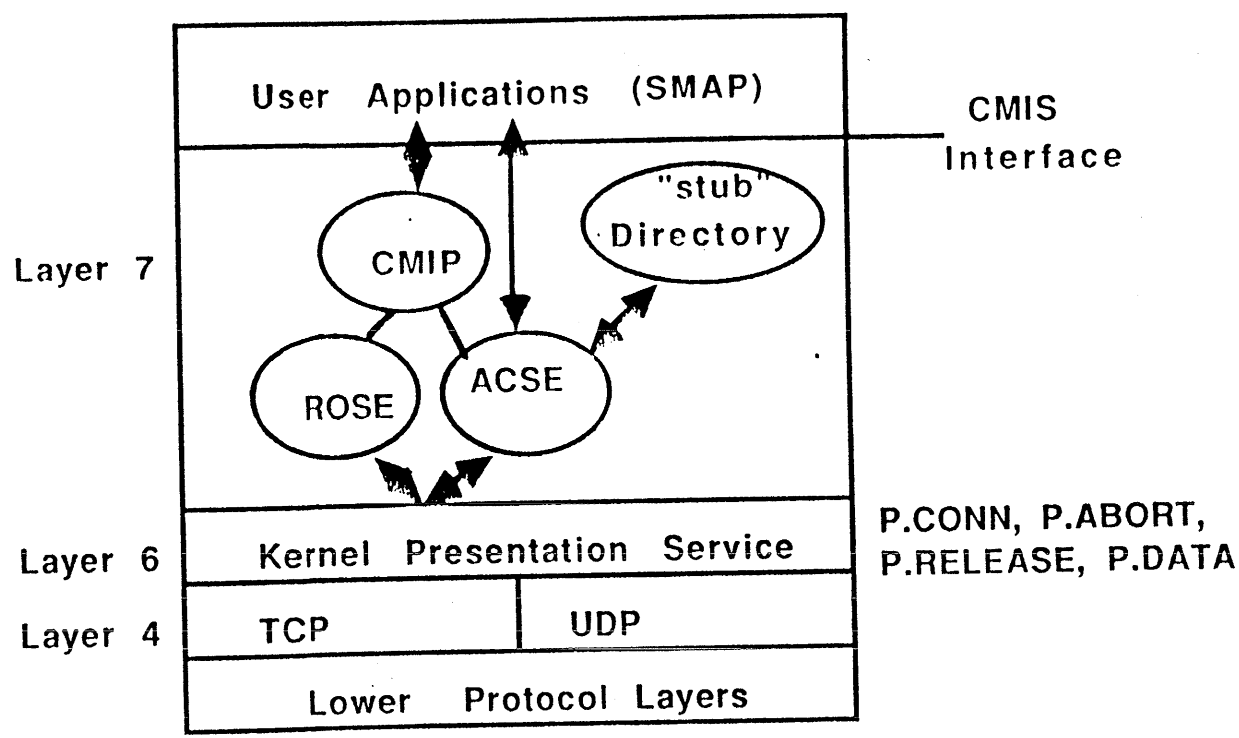
JUNE 17, 1988

MITRE

IAB Decision on TCP/IP Network Management
RFC1028

- SNMP is short term solution
- ISO based approach (NETMAN) is long term solution
- MIB-WG formed (Craig Partridge)
  - IDEA023 (SMI)
  - IDEA024 (MIB)
- SNMP WG formed (Marshall Rose)

Protocol Architecture for ISO Network Management on TCP/IP



MITRE

Status

- Awaiting CMIS/CMIP DIS
  - Editors meeting (August)
  - ISO WG4 meeting (Nov-Dec)

- Concentrating on Sept. TCP/IP Interoperability Conference Demo
  - Implementation experience (CMIP/RO/ACSE/Thin presentation)
  - SMI additions (thresholds, events)
  - MIB additions (IP, TCP, System, Ethernet)

Documents

- Goals and Scope 5/28/87
- IDEA012 "Network Management for TCP/IP Network: An Overview"
- IDEA013 "Structure and Identification of Management Information for the Internet"
  ----> IDEA023
- IDEA017 "ISO Presentation Services on Top of TCP/IP-based internets"
- Several Papers on MIB for Various Layers
  ---> IDEA024
- IDEA018 "System Load"
- IDEA0xx "TCP/IP Network Management Implementors Agreements for the Third TCP/IP Interoperability Conference Demonstration"

Major Issue

- When to stabilize implementors agreements on CMIS/CMIP and SMI
  - Track ISO Standard Schedule
    (no feedback to ISO as called for in 1028)
  - Take snapshot in Sept (after demo and editors meeting)
  - Take snapshot in Dec (after ISO WG4 meeting)

11.0 Acknowledgements

This memo was heavily influenced by the work of many people including the NETMAN committee, and the IETF MIB Working Group.

It is the result of the suggestions, the discussions, and the compromises reached by the members of the NETMAN Demo Sub-committee:

Brain Adley, Apollo Computer
Roald Adolfsen, Ungerman-Bass
Bill Anderson, The MITRE Corporation
Karl Auerbach, Epilogue Technology
K. Ramesh Babu, Excelan
Amatzia Ben-Artzi, 3Com
Lawrence Besaw, Hewlett-Packard
Anthony Chung, Sytek
George Cohn, Ungerman-Bass
James Davidson, The Wollongong Group
Phill Gross, The MITRE Corporation
Steve Holmgreen, CMC
Bert Jensen, Convergent
Lee Labarre, The MITRE Corporation
Anne Lam, Unisys
Dan Lynch, Advanced Computing Environments
Keith McCloghrie, The Wollongong Group
Dave Mackie, 3Com/Bridge
George Marshall, 3Com/Bridge (Chair)
Lynn Monsanto, Sun Microsystems
John Mullen, CMC
Gary Nitzberg, Network General
Anita Patton, Unisys
Jim Robertson, 3Com/Bridge
Milt Roselinsky, CMC
Harry Saal, Network General
Lou Steinberg, IBM
Unni Warrier, Unisys
Ken Chapman, Digital Equipment Corp.

7.16 Internet Host Requirements—Braden, ISI

Host: venera.isi.edu

Path:

pub/ietf-hosts.rfc.txt

Mailing list:

ietf-hosts-request@nnsc.nsf.net

ICMP GATEWAY DISCOVERY

- Hosts learn default gateways
- Hosts learn about dead gateways

Gateway Discovery Report:

- Gateway(s) multicast/broadcast
- Contains:
  Default gateway (list) and
  Address Mask

Gateway Discovery Query:

- Host may multicast/broadcast
  when it initializes

MODELS FOR USE:

- MODEL A:

  Each gateway broadcasts its own address every (5 ± random#) mins.

- MODEL B:

  One designated gateway (e.g. highest IP address) broadcasts list of all gateways that are up, every 5 mins.

- MODEL C:

  One designated gateway broadcasts list of all up gateways every 15 seconds.

Host Requirements RFC

THE BIG WORDS . . .

---> ---> ---> M U S T <--- <--- <---

S H O U L D

(or: RECOMMEND)

M A Y

(or: OPTIONAL)

Host Requirements RFC

OUTLINE

[13] 1. INTRODUCTION

- 1.1 General Caveats
- 1.2 Internet Architecture
- 1.3 Reading this Document
- 1.4 References

[1+] 2. LINK LAYER

(=> RFC-1009)

[36] 3. IP LAYER -- IP and ICMP

[19] 4. TRANSPORT LAYER -- TCP and UDP

5. APPLICATION LAYER

[10] 5.1 SMTP and RFC-822

[5+] 5.2 FTP

[3+] 5.3 TFTP

[0+] 5.4 Telnet

6. SUPPORT SERVICES

[6+] 6.1 Domain System

[0+] 6.2 Booting

[1+] 6.3 Management

7. APPENDICES

[0+] A. Checklist

EXAMPLE

3. IP LAYER

. . .

3.3 SPECIFIC ISSUES

3.3.1 Routing Outbound Datagrams

3.3.2 Reassembly

3.3.3 Fragmentation

3.3.4 Multihomed Hosts

3.3.5 Mis-addressed Datagrams

3.3.6 Error Reporting

3.3.7 IP Multicasting

TYPICAL ORGANIZATION

x.1 INTRODUCTION

x.2 PROTOCOL WALK-THROUGH

Contains exceptions, errors, requirements, suggestions, and pitfalls, keyed to section/page of protocol specification document(s).

x.3 SPECIFIC ISSUES

Discusses important general topics for the protocol(s).

x.4 INTERFACES

Discusses service interface.

x.4 REFERENCES

The documents every implementor MUST read . . .

Host Requirements RFC

OUTLINE

1. INTRODUCTION
   - 1.1 General Caveats
   - 1.2 Internet Architecture
   - 1.3 Reading this Document
   - 1.4 References

2. LINK LAYER
   (=> RFC-1009)

3. IP LAYER -- IP and ICMP

4. TRANSPORT LAYER -- TCP and UDP

5. APPLICATION LAYER
   - 5.1 SMTP and RFC-822
   - 5.2 FTP
   - 5.3 TFTP
   - 5.4 Telnet

6. SUPPORT SERVICES
   - 6.1 Domain System
   - 6.2 Booting
   - 6.3 Management

7. APPENDICES
   - A. Checklist
   - B. Checksums

8 PAPERS DISTRIBUTED AT IETF

- Monitoring Data Exchanges Between the NSF Backbone Network and its attached Regional Clients

Monitoring Data Exchanges between the NSFNET Backbone Network and its attached Regional Clients

Merit Computer Network
University of Michigan
June 1988

This report is the result of a meeting held 20 May 1988 to resolve questions about the availability of monitoring data and to discuss formats for data representation. The document is intended to form a base for further discussions and to provide an initial framework for policies covering the availability and exchange of monitoring data.

The May meeting was held following initial discussions between Merit, NSF, and the regional clients via electronic mail discussing initial monitoring data availability for the IP components of the backbone to regional network operations centers. Discussions of these issues between Merit and IBM also occurred prior to the meeting to explore the technical feasibility of various monitoring options.

Attending the meeting from Merit were Eric Aupperle, Hans-Werner Braun, Bilal Chinoy, Elise Gerich, Steve Gold, Dave Katz, Dave Martin, Rick Schmalgemeier, and Jessica Yu. Also attending were Jack Drescher, the NSFNET project manager within IBM, Craig Partridge (BBN/NNSC), and Guy Almes (Sesquinet/FARNET). Guy Almes, Craig Partridge, and Jacob Rekhter (IBM) reviewed an earlier draft of this document. Jacob Rekhter also made several suggestions for augmentation of the MIB, which were forwarded to Craig Partridge for consideration for the Internet MIB.

It should be noted that in the preceding months, the first priority has been development of NSS capabilities essential for implementing a full production network operation within the scheduled time frame. Additional features not required by the project solicitation, such as monitoring data interfaces to regional networks, were assigned a lower priority. While NSS development efforts are continuing, more resources are now being focused on implementing monitoring facilities within the network, both for the Merit/NSFNET Network Operation Center (NOC) and for regional network operation centers.

1. INITIAL IMPLEMENTATION PLAN FOR SGMP IN THE NSFNET BACKBONE

Three categories of individual needs for monitoring data were identified. These are:

Those that need immediate, real-time monitoring capabilities

Those that need composite information updated on a periodic basis

Those that need long-term data for research or long-term planning

Initially, SGMP will provide the monitoring facilities within the network. The proposed implementation will provide monitoring in which the entire Nodal Switching Subsystem (NSS) will appear as a single host to SGMP. Although each NSS is composed of nine IBM

RT/PCs, for the user the NSS appears as a single multi-processor system. This image needs to be retained to allow for a more logical view of backbone structure and to assure that later changes in NSS technology will not conflict with external views of the system.

Given that SGMP queries are relatively expensive, the ideal architecture would locate processor-intensive components (like ASN.1) outside of packet-forwarding processes (i.e., the Packet Switching Processors or PSPs within the NSS) while still allowing direct access to all critical data. One logical place to locate the SGMP query processor would be on the Routing Control Processors (RCPs), as RCPs are not involved in time-sensitive, packet-forwarding processes. The ASN.1 work can then be done internally by the RCP in a way not unlike the EGP peers, where EGP packets sent to the E-PSP are internally forwarded to the RCP. Alternatively the SGMP session can be set up with the RCP Internet address providing the same result. Use of the RCP would also facilitate future integration of the routing daemon with network management. The RCP will then be able to request monitoring information from the other local processors. As proposed, the query processor will be able to request data of system components of the NSS in real time.

With this system in place, a regional client may send SGMP queries to the local NSS via the regional network interface and will get responses from the same address. As long as regional clients only exchange SGMP traffic with the local NSS, the impact of excessive SGMP queries will be felt first by the regional network, rather then contributing to congestion in the overall network.

This model will work well for monitoring the backbone as seen by the local NSS. There may be instances where regional network operators would also like to query a remote NSS. This can be implemented by addressing an inquiry to the external IP address of an E-PSP in a remote NSS, i.e., the IP address of either the Ethernet interface or RCP. This service should be possible provided the additional traffic does not have a negative performance impact on the operation of the backbone.

Some upper limit of the query frequencies can be achieved by the use of session names within the SGMP servers. One or more session names can be assigned per regional network and to people with a need for access to real-time-monitoring data. The session names would be known to all the backbone nodes. Session names will provide security to the backbone by limiting SGMP queries and therefore, session names should be changed regularly. An accounting mechanism would be implemented to keep usage tables ordered by session names. Counts will include uses per session.

Initially there will be no broader public access to real-time monitoring. Depending on how the operation of the backbone is or is not impacted by the real-time-monitoring-data access, access privileges could be reviewed and changed if the need for such a re-evaluation arises.

2. WHAT IS NEEDED TO SATISFY THE MONITORING NEEDS OF THE REGIONAL NOCs?

Prior to the meeting, Guy Almes sent a summary of a MIB to Merit, including a prioritization of the entries. It was generally felt that this would be a minimum of data that would be useful to the regional networks. Guy Almes' list was modified slightly during the meeting. The adjusted list is included in the appendix of this document, with the entries of the MIB prioritized as high, medium, or low priority for the early phases of operation. Furthermore a MIB extension suggested separately by Jacob Rekhther of IBM to satisfy the policy-based routing as well as the IS-IS monitoring needs is also attached to the appendix.

In summary, those entries receiving a high priority are:

System Group

Interfaces Group--just the virtual interfaces in and out of the NSS are included

IP Group

IP Gateway Group

EGP Group - entries concerning EGP neighbors are essential, others are only medium priority

Those entries receiving a medium priority are:

Much of the Interfaces Group

Address Translation Group

UDP Group (need due to SGMP)

EGP Group - In/Out msgs and In/Out errors

Those entries receiving a low priority are:

ICMP Group

Those entries that need not be available at all:

TCP Group

In addition, it was agreed that since SGMP will give real-time data to regional NOCs, there is no need for them to have login accounts on the NSS. A well-working transaction protocol appears to be preferable.

3. CONCLUSIONS

Real-time monitoring facilities will be provided by SGMP servers close to the regional networks. It should be possible for designated SGMP clients at regional NOCs to query remote backbone nodes as need be.

Summarized monitoring data for non time-critical needs should be available on line from the Merit Information Services (IS) machine. This may also include data which is not available via

SGMP (like IDNX monitoring).

Monitoring data should be kept by the Merit NOC and should be available from the IS machine for researchers.

There may be improved database support for monitoring data available on the IS machine at a later stage of the project.

There was recognition of the importance to implementing time synchronization between networking components, so that monitoring data and other events from different network entities can be correlated with each other.

4. Appendices

Appendix 1

Suggestions sent by Craig Partridge prior to the Ann Arbor meeting:

To: hwb@mcr.umich.edu
To: almes@rice.edu
Cc: nnsc@NNSC.NSF.NET
Subject: Monitoring Information
Date: Wed, 18 May 88 11:18:45 -0400
From: Craig Partridge <craig@NNSC.NSF.NET>

Hans-Werner and Guy,

I've spent a little time this morning trying to pull together my thoughts on making network management information available to people outside MERIT. Here are my general views -- which are subject to change at the meeting.

First, my inclination is to divide the community of interest into two groups: researchers, who want to examine the network information as a test of ideas, and operational folk, who want to examine network information to help diagnose network performance problems (or failures). I think the two groups have very different needs.

I've talked with the NOC here about what long term information they make available to researchers. It turns out to be very little. There's a lot of detailed information that stays around on the INOC host for short periods (under a week) and a certain amount of summary information that is kept for up to three years. But detailed data isn't available for further back. Apparently the summary information is good enough for most people's purposes.

But personally, I'd like to encourage you to keep better records than that. I'd love it if it were possible to order a tape of detailed network management information (possibly as much as hourly dumps of the complete MIB on each machine) for any time in the history of the backbone. (For example, I'd like to be able to call up and say, "can I have the tapes for March of each year of operation?"). Given that tape archiving and tape copying is cheap, and 6250bpi holds a fair amount of information, I think this isn't an outrageous idea.

In the short term, of course, accredited researchers can long into INOC and get the information they want. That's fine, except how much do you want researchers pinging on your network?

As for operational folk -- they usually want up to date current information. Again the problem is how much do you want them pinging on your network, and how much do they need to ping on your network.

I can make a strong case that operational people never should need to monitor the backbone itself, and that you should only let them do so if you believe it will help you run the backbone better. (Note that it probably will help you run the backbone better because they'll catch some problems faster than you will -- but there's a tradeoff here).

The argument that operational folk never need to monitor the backbone is. The classic problem is figuring out what's wrong with connectivity from point X on one regional to point Y on another. (Note that since, to the outside world, the backbone only takes IP traffic, no node on the backbone will be X or Y.) So the real question is do operational folks need to monitor the backbone to track down the connectivity problems between X and Y. I don't think so.

Consider that both regional networks can monitor their gateways connecting them to the backbone (this from Lou Steinberg) so they can confirm that their connection to the backbone is sound. A simple ICMP ping will confirm that they can get through the backbone. After they've confirmed they can get through the backbone, then the connectivity problem is a matter of using SNMP within the regionals to track the problem, not a matter of looking at the backbone.

But, one fly in the ointment. Assume that an ICMP across the backbone shows that they cannot get across the backbone, or that backbone round-trip times are highly variable. Would you prefer that they track the problem further and then call MERIT, or that MERIT be notified and track the problem itself? If they do the research, you save a lot of staff time -- but will have to spend time educating people into how the backbone works.

If you prefer their help, you need an open backbone (anyone can monitor it if they have the right SNMP password). (Note that having an INOC they can log into is a partial help, but you cannot assume that they can reach INOC -- the failure may be between them and your INOC). Otherwise, you can tell them just call MERIT at signs of backbone trouble.

Politically this may be touchy so you'd have to release a detailed technical explanation of why you are doing this.

Finally, on MIB information -- my view is that you should make everything in the MIB visible to people. The idea is that the MIB contains information useful to external people. So hiding it is a bad idea. Also, you should conform to the core MIB being developed by the IETF (yes I'm biased here).

Does this help start things???

Craig

Appendix 2

Suggested prioritized MIB for the initial monitoring:

```
System Group
h       sysID           Octet String
h       sysObjectId     Object Identifier
h       sysClock        NetworkTime
h       sysLastInit     Integer(seconds)

Interfaces Group
h       ifNumber        Integer
h       ifTable         sequence of IfEntry, where
        IfEntry is sequence {
        m       ifPhysAddress   Octet String
        h       ifIpAddress     IpAddress
        h       ifMtu           Integer
        h       ifNetMask       IpAddress
        h       ifInPkts        Counter
        h       ifOutPkts       Counter
        m       ifInDropped     Counter
        m       ifOutDropped    Counter
        m       ifInBcastPkts   Counter
        m       ifOutBcastPkts  Counter
        m       ifInErrors      Counter
        m       ifOutErrors     Counter
        h       ifOutQLen       Gauge
        l       ifName          Octet String
        h       ifStatus        Integer{reserved, testing, down, up}
        h       ifType          Integer{reserved, 1822hdh, 1822, fddi, ddn-x25,
                                        rfc877-x25, starLan, proteon-10MBit,
                                        proteon-80MBit, ethernet,
                                        88023-ethernet, 88024-tokenBus,
                                        88025-tokenRing, pointToPointSerial}
        h       ifSpeed         Gauge(b/s)
        m       ifMediaErrors   Counter
        h       ifUpTime        NetworkTime
        }

Address Translation Group
m       atTable         sequence of AtEntry, where
        AtEntry is sequence {
        m       atPhysAddress   Octet String
        m       atIpAddress     IpAddress
        }

IP Group
h       ipInDatagrams   Counter
m       ipInErrors      Counter
h       ifInDropped     Counter
h       ipOutDatagrams  Counter
m       ipOutErrors     Counter
h       ifOutDropped    Counter
m       ipFragRcvd      Counter
m       ipFragDropped   Counter
m       ipFragTimedOut  Counter
h       ipFragmented    Counter
h       ipRoutingTable  sequence of IpRoutingEntry, where
```

```
        IpRoutingEntry is sequence {
        h       ipRouteMetric1  Gauge
        h       ipRouteMetric2  Gauge
        h       ipRouteNextHop  IpAddress
        h       ipRouteType     Integer{nowhere, direct, remoteHost,
                                        remoteNetwork, subNetwork}
        h       ipRouteAuthor   IpAddress
        h       ipRouteProto    Integer{other, local, icmp, egp, ggp, hello,
                                        rip, proprietaryIGP, netmgmt}
        }

IP Gateway Group
h       gwCoreRouter    Integer{leaf, internal}
h       gwAutoSys       Integer
h       gwForwDatagrams Counter

ICMP Group
l       icmpInStats     IcmpStats
l       icmpOutStats    IcmpStats, where
        IcmpStats is sequence {
        l       icmpMsgs        Counter
        l       icmpErrors      Counter
        l       icmpDestUnreach Counter
        l       icmpTimeExcd    Counter
        l       icmpParmProb    Counter
        l       icmpSrcQuench   Counter
        l       icmpRedirect    Counter
        l       icmpEcho        Counter
        l       icmpEchoRep     Counter
        l       icmpTimestamp   Counter
        l       icmpTimestampRep Counter
        l       icmpInfo        Counter
        l       icmpInfoRep     Counter
        l       icmpAddrMask    Counter
        l       icmpAddrMaskRep Counter
        }

TCP Group
n/a     tcpRtoAlgorithm Integer{other, constant, rsre, vanj}
n/a     tcpRtoMin       Integer
n/a     tcpRtoMax       Integer
n/a     tcpMaxConn      Gauge
n/a     tcpConnAttempts Counter
n/a     tcpConnOpened   Counter
n/a     tcpConnAccepted Counter
n/a     tcpConnClosed   Counter
n/a     tcpConnAborted  Counter
n/a     tcpInOctets     Counter
n/a     tcpOutOctets    Counter
n/a     tcpInSegs       Counter
n/a     tcpDupSegs      Counter
n/a     tcpOutSegs      Counter
n/a     tcpRetransSegs  Counter
n/a     tcpListens      sequence size (256) of Integer{idle, listening}

UDP Group
m       udpInDatagrams  Counter
m       udpInErrors     Counter
m       udpOutDatagrams Counter
```

```
EGP Group
m       egpInMsgs        Counter
m       egpInErrors      Counter
m       egpOutMsgs       Counter
m       egpOutErrors     Counter
h       egpNeighborTable sequence of EgpNeighborEntry, where
        EgpNeighborEntry is sequence {
        h       egpNeighborState Integer{idle, acquisition, down, up, cease}
        h       egpNeighborAddr IpAddress
        }
```

Appendix 3

Initial draft of policy based routing and IS-IS MIB extensions as suggested by Jacob Rekhter; neither considered complete or final:

```
Gateway Policy Routing Group {
    ASin sequence of Integer
    validAS sequence of {
         net  IpAddress
         AS   Integer
         metric    Integer
    }
    Egpmetricout sequence of {
         EgpNeighborAddr     IpAddress
         metric         Integer
    }
    Egpmetricin sequence of {
         EgpNeighborAddr     IpAddress
         metric         Integer
    }
}
```

```
IS-IS Group {
    RouterLinksPDUin    Counter
    RouterLinksPDUout   Counter
    ESLinksPDUin        Counter
    ESLinksPDUout       Counter
    SequenceNumberPDUin Counter
    SequenceNumberPDUout     Counter
    CorruptedPDUin      Counter
    IS-ESHelloin        Counter
    IS-ESHelloout       Counter
    IS-ISHelloin        Counter
    IS-ISHelloout       Counter
    IS-ISneighborTable sequence of IS-ISneighbor, where
    IS-ISneighbor is sequence {
         IS-ISneighborAddr   IpAddress
         cost           Integer
         hold-time      Integer
    }
}
```

Appendix 4

Example gated EGP peer

```
#
# Gated conf for exchanging routing information with NSFnet backbone
#

traceflags internal external egp route

RIP yes
HELLO no
EGP yes

# No RIP on exterior net
noripoutinterface 192.35.82.34
noripfrominterface 192.35.82.34

# Allow NSFnet learned routes to be protogated to the campus
sendAS 145      ASlist 26

# Ignore Merit from campus in favor of EGP learned route from NSS
donotlisten 35           intf 128.84.248.34      proto rip

# Cornell's autonomous system number
autonomoussystem 26

# Peer with NSS
egpneighbor 192.35.82.100       ASin 145        nogendefault validate

# Nets that we will listen to from NSS
validAS 35               AS 145  metric 24564
validAS 129.140          AS 145  metric 24564
validAS 192.35.161       AS 145  metric 24564
validAS 192.35.162       AS 145  metric 24564
validAS 192.35.163       AS 145  metric 24564
validAS 192.35.164       AS 145  metric 24564
validAS 192.35.165       AS 145  metric 24564
validAS 192.35.166       AS 145  metric 24564
validAS 192.35.167       AS 145  metric 24564
validAS 192.35.168       AS 145  metric 24564
validAS 192.35.169       AS 145  metric 24564
validAS 192.35.170       AS 145  metric 24564

# Nets that we will advertize to the NSS
announce 192.35.82       intf all                proto rip egp   egpmetric 1
announce 128.84          intf all                proto rip egp   egpmetric 1
announce 128.253         intf all                proto rip egp   egpmetric 1

# Nets that we will advertize to the campus
announce 129.140         intf 128.84.248.34      proto rip
announce 192.35.161      intf 128.84.248.34      proto rip
announce 192.35.163      intf 128.84.248.34      proto rip
```

Appendix 5

Example NSS routing configuration file corresponding to the gated.conf file in Appendix 4

```
RIP     no
HELLO   no
EGP     yes
#
#traceflags internal external route egp update is-is es-is
traceflags internal external route update is-is
#
autonomoussystem 145
egpneighbor      192.35.82.238 nogendefault egpmetricout 128 ASin 26 validate
egpneighbor      192.35.82.34  nogendefault egpmetricout 128 ASin 26 validate
#
egpmaxacquire  2
#
validAS 128.84          AS 26 metric 1          # Cornell
validAS 128.253         AS 26 metric 1          #
validAS 192.35.82       AS 26 metric 1          #
#
sendAS 26 ASlist 145
#
backbone 129.140.74.9   metric  10
backbone 129.140.74.12  metric  10
backbone 129.140.74.15  metric  10
#
regional 192.35.82.100
#
```

Appendix 6

Example routing configuration file for another regional network

```
RIP     no
HELLO   no
EGP     yes
#
#traceflags internal external route egp update is-is es-is
traceflags internal external route update is-is
#
autonomoussystem 145
egpneighbor      128.121.54.71 nogendefault egpmetricout 128 ASin 97 validate
egpneighbor      128.121.54.72 nogendefault egpmetricout 128 ASin 97 validate
#
egpmaxacquire  2
#
validAS 128.121        AS 97   metric 1        # JvNC
validAS 128.112        AS 97   metric 1        # Princeton
validAS 192.16.204     AS 97   metric 1        # IAS
validAS 128.6          AS 97   metric 1        # Rutgers
validAS 18             AS 97   metric 1        # MIT
validAS 128.103        AS 97   metric 1        # Harvard
validAS 128.148        AS 97   metric 1        # Brown
validAS 192.12.216     AS 97   metric 1        # Stevens
validAS 192.26.148     AS 97   metric 1        # UMdNJ
validAS 128.235        AS 97   metric 1        # NJIT
validAS 128.119        AS 97   metric 1        # UMass Amherst
validAS 129.170        AS 97   metric 1        # Dartmouth
validAS 129.10         AS 97   metric 1        # Northeastern
validAS 128.197        AS 97   metric 1        # Boston  U.
validAS 129.133        AS 97   metric 1        # Wesleyan
validAS 192.26.88      AS 97   metric 1        # Yale
validAS 128.36         AS 97   metric 1        # Yale
validAS 128.118        AS 97   metric 1        # Penn State
validAS 128.91         AS 97   metric 1        # UPenn
validAS 128.122        AS 97   metric 1        # NYU
validAS 128.151        AS 97   metric 1        # Rochester
validAS 128.59         AS 97   metric 1        # Columbia
validAS 128.196        AS 97   metric 1        # Arizona
validAS 128.138        AS 97   metric 1        # Colorado
validAS 192.31.28      AS 97   metric 1        # Steward Obs
validAS 128.128        AS 97   metric 1        # Woods Hole
validAS 128.180        AS 97   metric 1        # Lehigh
validAS 129.25         AS 97   metric 1        # Drexel
validAS 129.32         AS 97   metric 1        # Temple
#
backbone 129.140.72.9   metric  10
backbone 129.140.72.16  metric  10
backbone 129.140.72.17  metric  10
#
regional 128.121.54.1
#
sendAS 97 ASlist 145
#
```