“A Maze of Twisty, Turney Passages” – Routing in the Internet Swamp (and other adventures)

David L. Mills
University of Delaware
http://www.eecis.udel.edu/~mills
mills@udel.edu

“When you are up to your ass in alligators, it is wise to remember you are there to drain the swamp.”
- R.M. Nixon

Background (not in the tutorial presentation)

- This was first presented as a tutorial at Harvard for SIGCOMM 99.
- There were four tutorials, including this one, presented over an 8-hour period. They were videotaped, but I don’t know where the tapes are.
- This is a personal retrospective, not a history archive, and covers topics important to me and which were my major research interests.
- From the perspective of the program managers, I was the “internet greasemonkey”.
- I chaired the Gateway Algorithms and Data Structures (GADS) and later the Internet Architecture (INARC) task forces and was a member of the Internet Control and Configuration Board (ICC) and later the Internet Activities Board (IAB).
- On my watch was gateway architecture, network and internetwork routing algorithms, subnetting and growing pains.
- The Internet History Project is at www.postel.org.

On the Internet cultural evolution

- "We have met the enemy and he is us." – Walt Kelly
- Maybe the most important lesson of the Internet was that the technology was developed and refined by its own users
  - There was a certain ham-radio mentality where users/developers had great fun making new protocols to work previously unheard applications
  - The developers were scattered all over the place, but they had a big, expensive sandbox with little parental supervision
  - There is no doubt that the enthusiasm driving the developers was due to the urgent need to communicate with each other without wasting trees or airplane fuel
- The primary motivation for the Internet model was the need for utmost reliability in the face of untried hardware, buggy programs and lunch
  - The most likely way to lose a packet is a program bug, rather than a transmission error
  - Something somewhere was/is/will always be broken at every moment
  - The most trusted state is in the endpoints, not the network

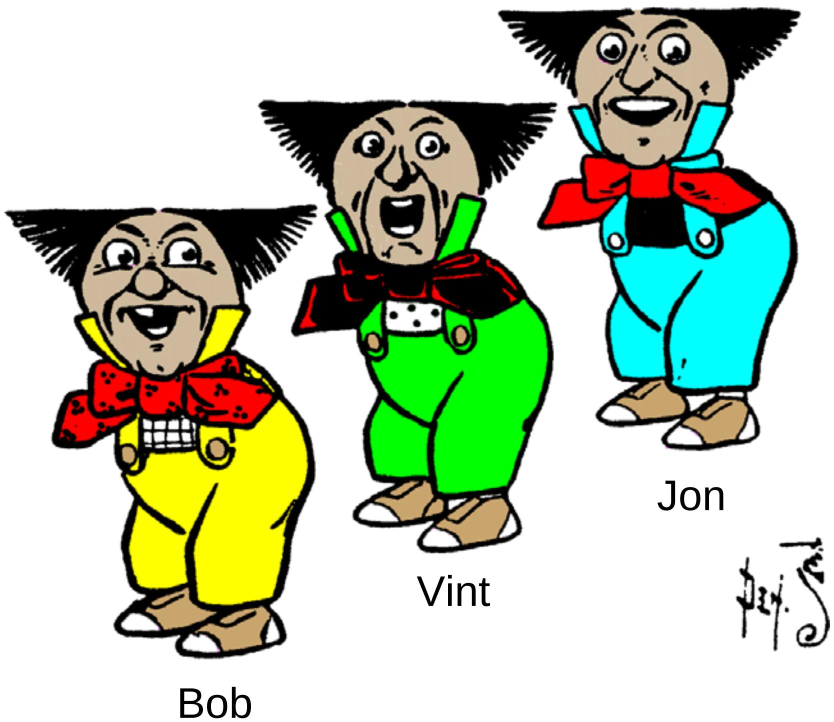
Milestones

- The IP/TCP coming-out party at NCP was a full day of presentations and demonstrations using ARPAnet and SATnet between Washington and London. This was in November 1979.
- The boys in the back room had been noodling the architecture and protocols and demonstrating interoperability at numerous bakeoffs since 1977.
- The original Internet address structure was a single 8-bit network number. My sandbox was net 29. We did this because we thought the evolved Internet would have only a few providers, like the telephone infrastructure.
- The Internet Flag day was 1 January 1982 when the Internet formally came into existence. We had been tunneling it over ARPAnet for five years. Some of the boys got “I survived the Internet” teashirts.
- At a meeting in London in 1981 the now familiar class A/B/C formats were approved. Subnetting and multicasting came later.

The day the Internet (almost) died

- There was hard feeling in the intenational (ITU) community, who believed networks should be evolved from ISO architectural concepts.
- We rascals were sneaking around in the bushes building IP/TCP and didn’t ask them for advice. The called us arrogant ARPAnaut pirates.
- The NAS convened a panel of experts to discuss what to do:
  - 1. Turn off the lights on IP/TCP and do it right now.
  - 2. Allow a couple of years to do (1), then put the ARPAnauts in jail.
  - 3. Turn off the lights on ISO.
- The decision was (2). Then, somebody asked where to buy ISO and the cupboard was bare. Meanwhile, Unix had IP/TCP and the AT&T license had elapsed.
- Funny thing is that many routers of that day to this could and can switch both IP and ISO at the same time.

Intermission 1977-1983

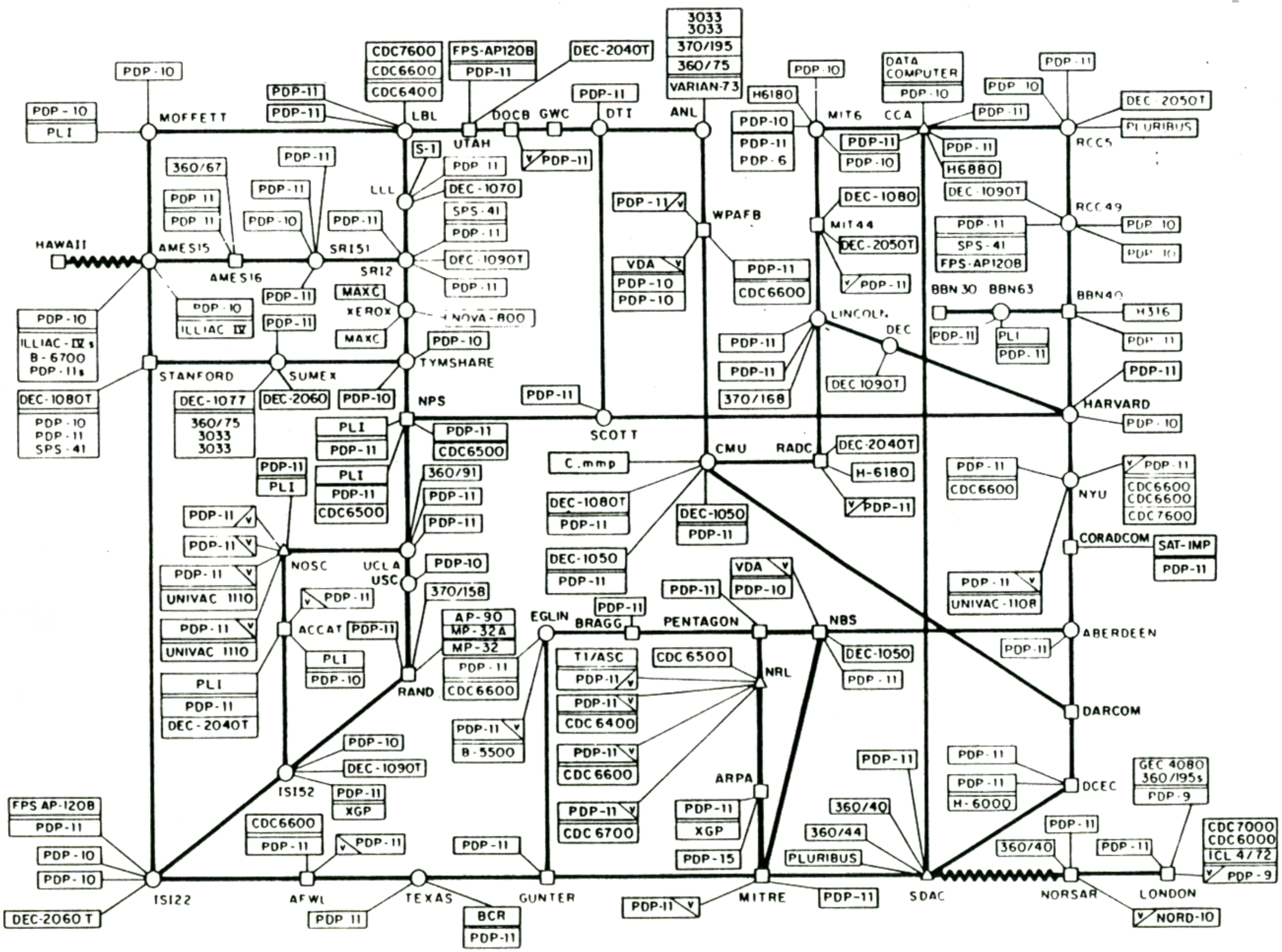


- Getting the word out
- The ARPAnet as the first Internet backbone network
- Internet measurements and performance evaluation
- The GGP routing era
- Evolution of the autonomous system model

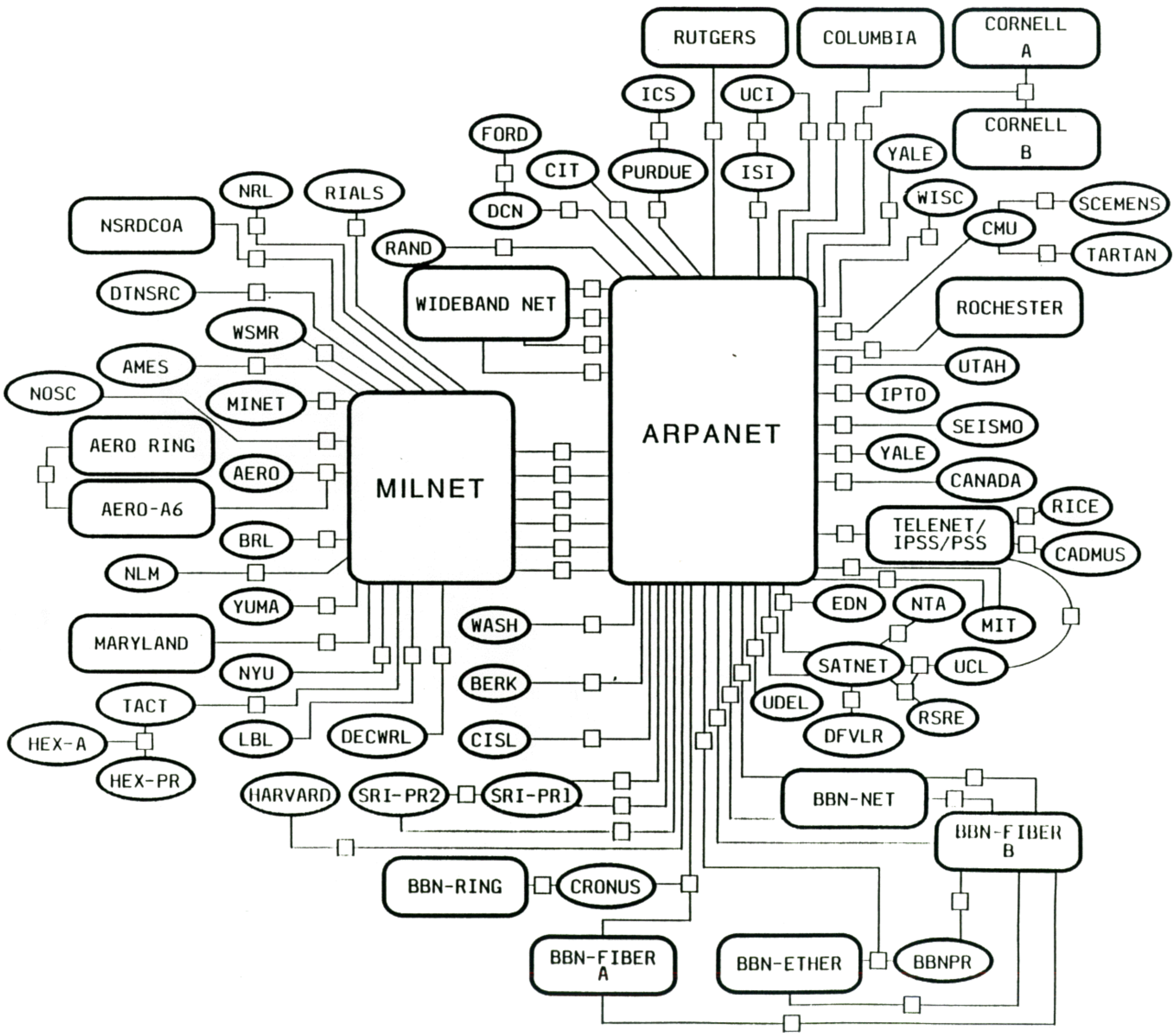
On the Internet and the ARPAnet life cycle

- The original ARPAnet was actually a terminal concentrator network so lots of dumb terminals could use a few big, expensive machines
- In the early Internet, the ARPAnet became an access network for little IP/TCP clients to use a few big, expensive IP/TCP servers
- In the adolescent Internet, the ARPAnet became a transit network for widely distributed IP/TCP local area networks
- In the mature Internet, the ARPAnet faded to the museums, but MILnet and clones remain for IP/TCP and ITU-T legacy stuff
- ARPAnet clones persist today as the interior workings of X.25 networks used for credit checks and ATM networks.

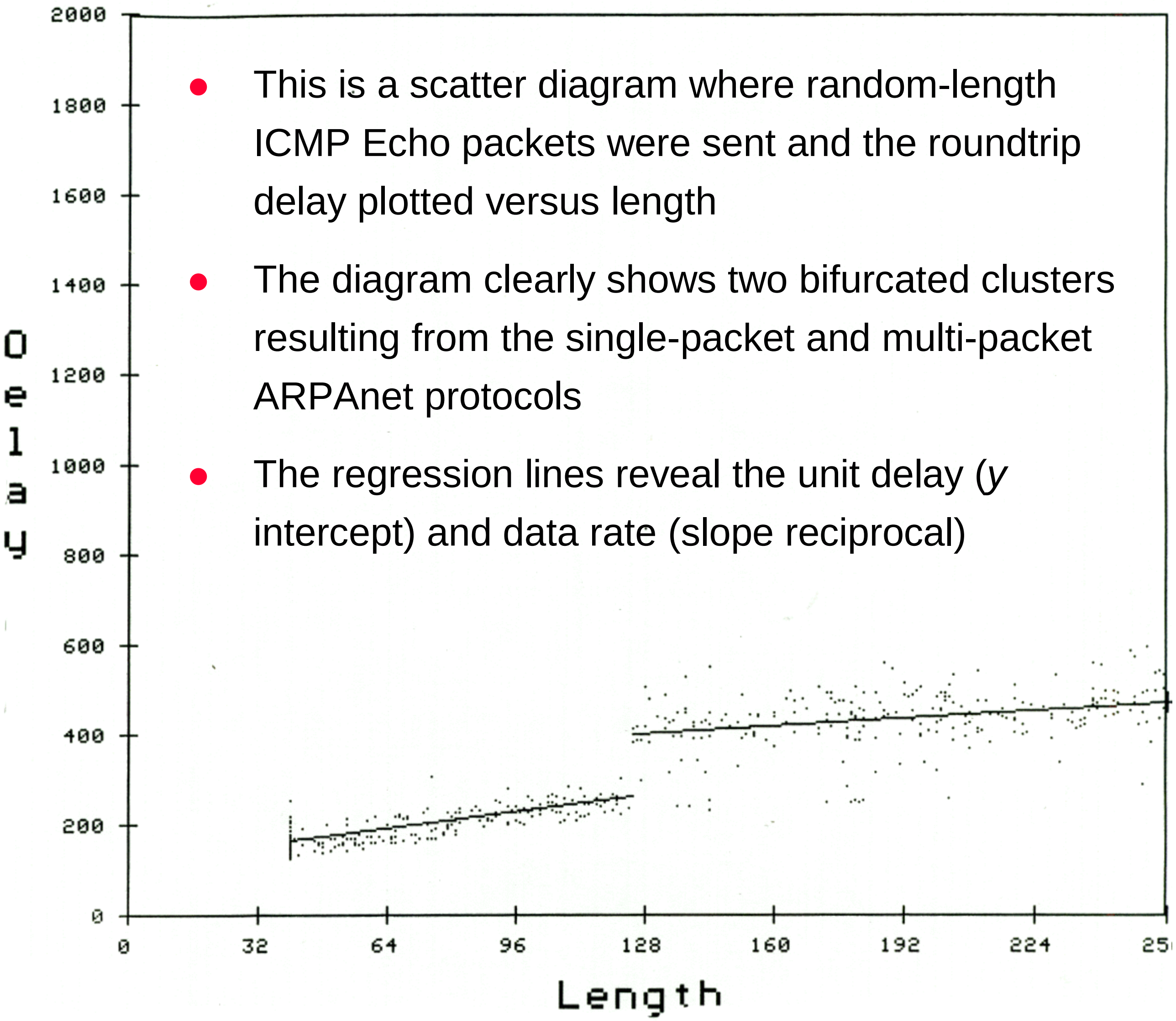
ARPAnet topology March 1979



ARPANet/MILnet topology circa 1983



Graphical means to estimate ARPAnet performance



- This is a scatter diagram where random-length ICMP Echo packets were sent and the roundtrip delay plotted versus length
- The diagram clearly shows two bifurcated clusters resulting from the single-packet and multi-packet ARPAnet protocols
- The regression lines reveal the unit delay (*y* intercept) and data rate (slope reciprocal)

DARPA packet radio network

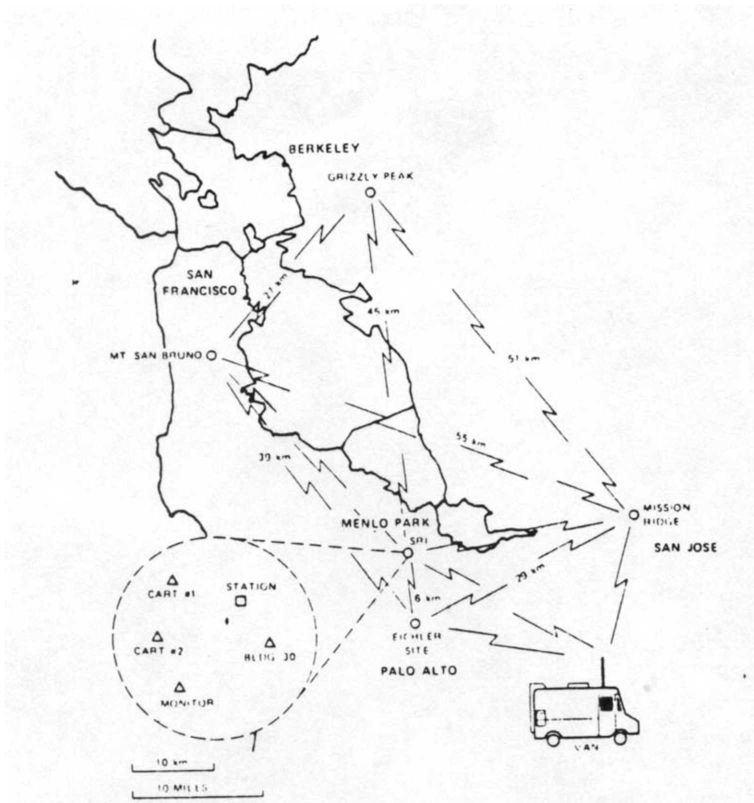
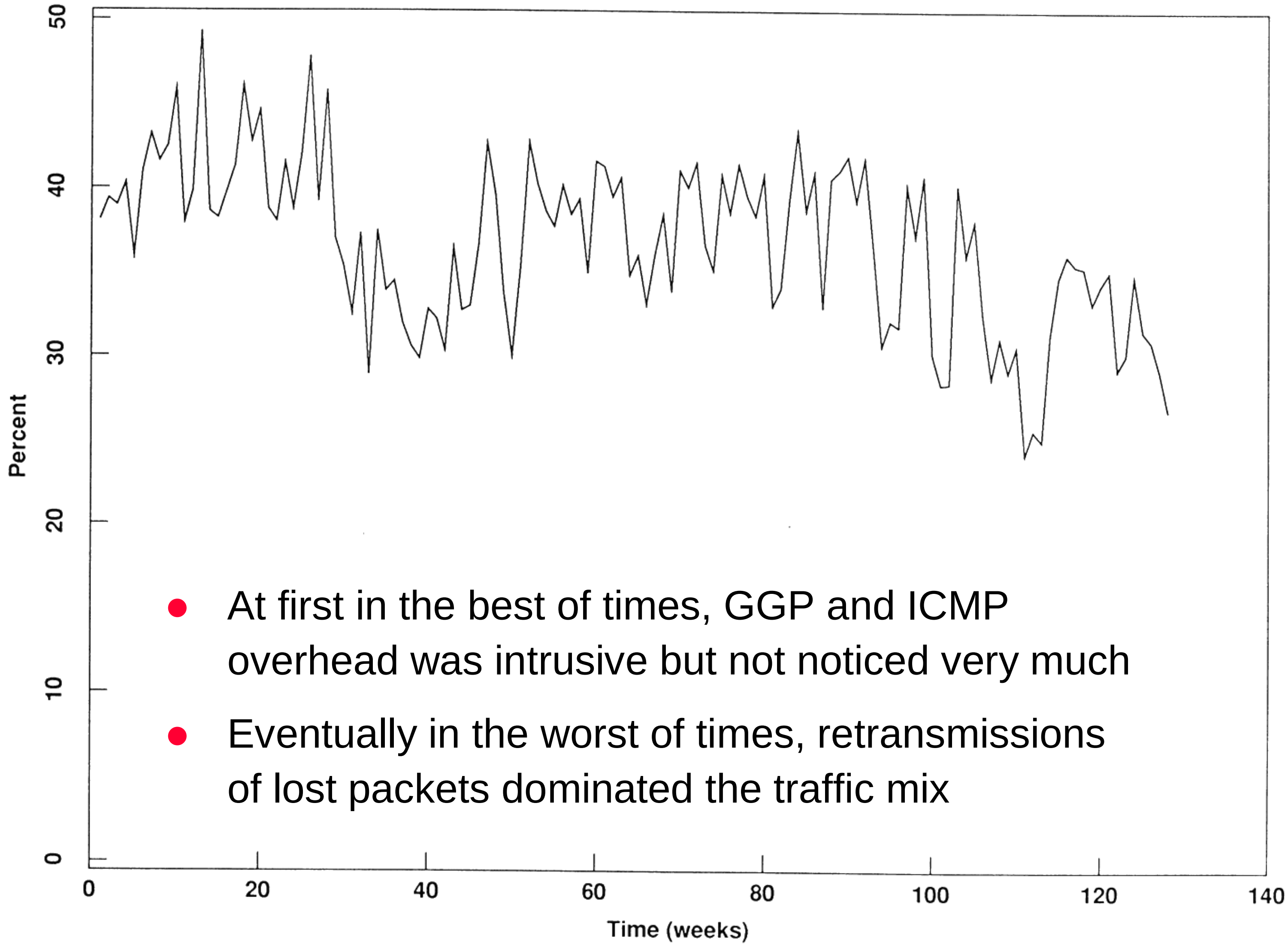


Fig. 12. Location of major elements of the packet radio testbed during 1977.

Gateway-Gateway Protocol (GGP)

- Used in early Internet of wide area (ARPAnet), packet radio (PRnet) and international satellite (SATnet) networks
- Implemented by BBN and COMSAT in tiny PDP11 computers
- Used node-state Bellman-Ford routing algorithm similar to early ARPAnet routing algorithm
- Shared all deficiencies known with node-state algorithms
  - Becomes unstable in large networks with intermittent connectivity
  - Vulnerable to routing loops (counts to infinity)
  - Does not scale to large Internet (single packet updates)
  - Burdened with network information functions, later divorced to ICMP
  - Problems with interoperable implementations
  - First instance of hello implosion – hosts should not ping gateways
- Lesson learned: the Internet was too vulnerable to scaling and interoperability issues in the routing infrastructure

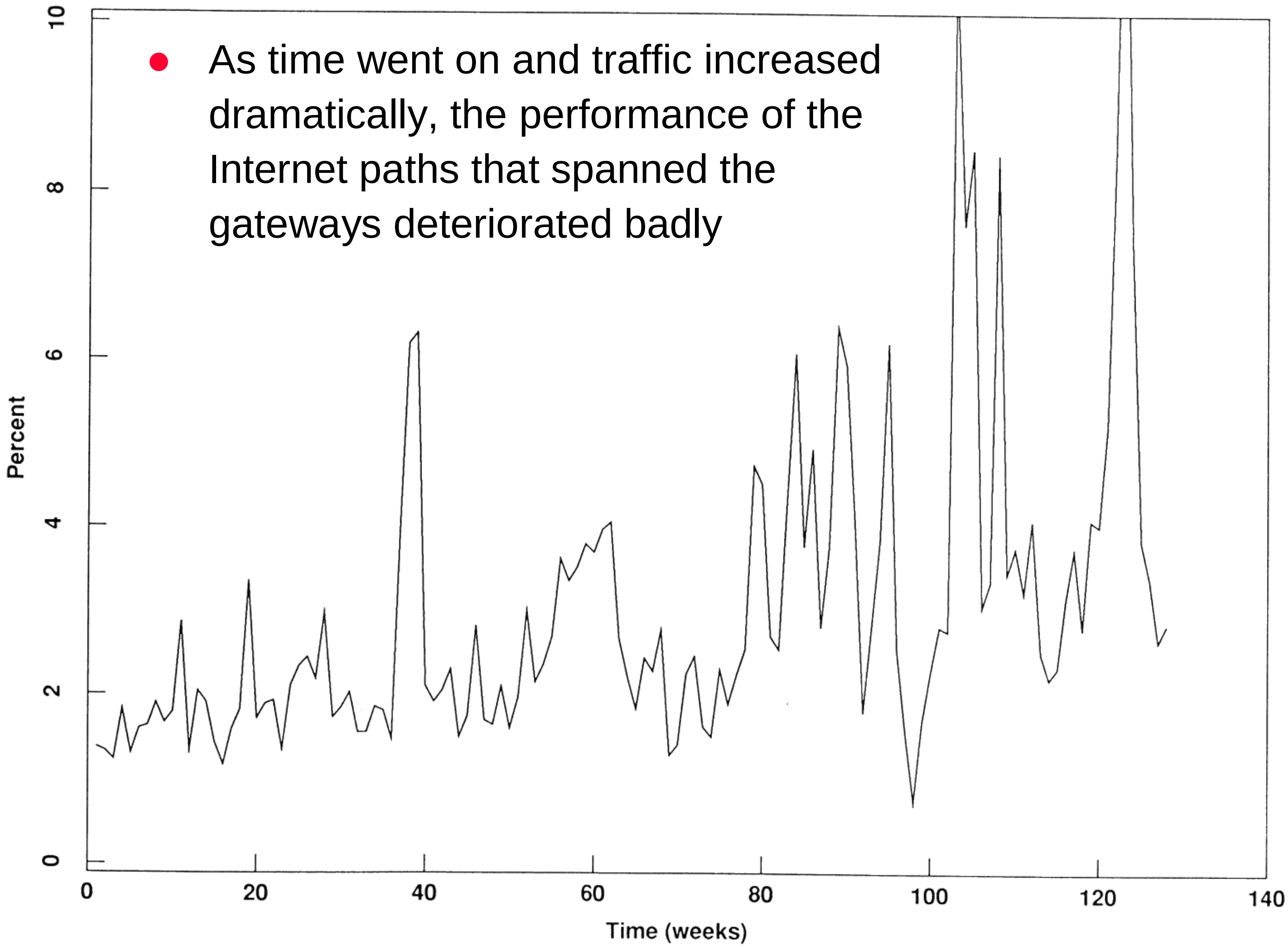
Overhead at GGP ARPAnet/MILnet gateways



- At first in the best of times, GGP and ICMP overhead was intrusive but not noticed very much
- Eventually in the worst of times, retransmissions of lost packets dominated the traffic mix

Packet loss at GGP ARPAnet/MILnet gateways

- As time went on and traffic increased dramatically, the performance of the Internet paths that spanned the gateways deteriorated badly

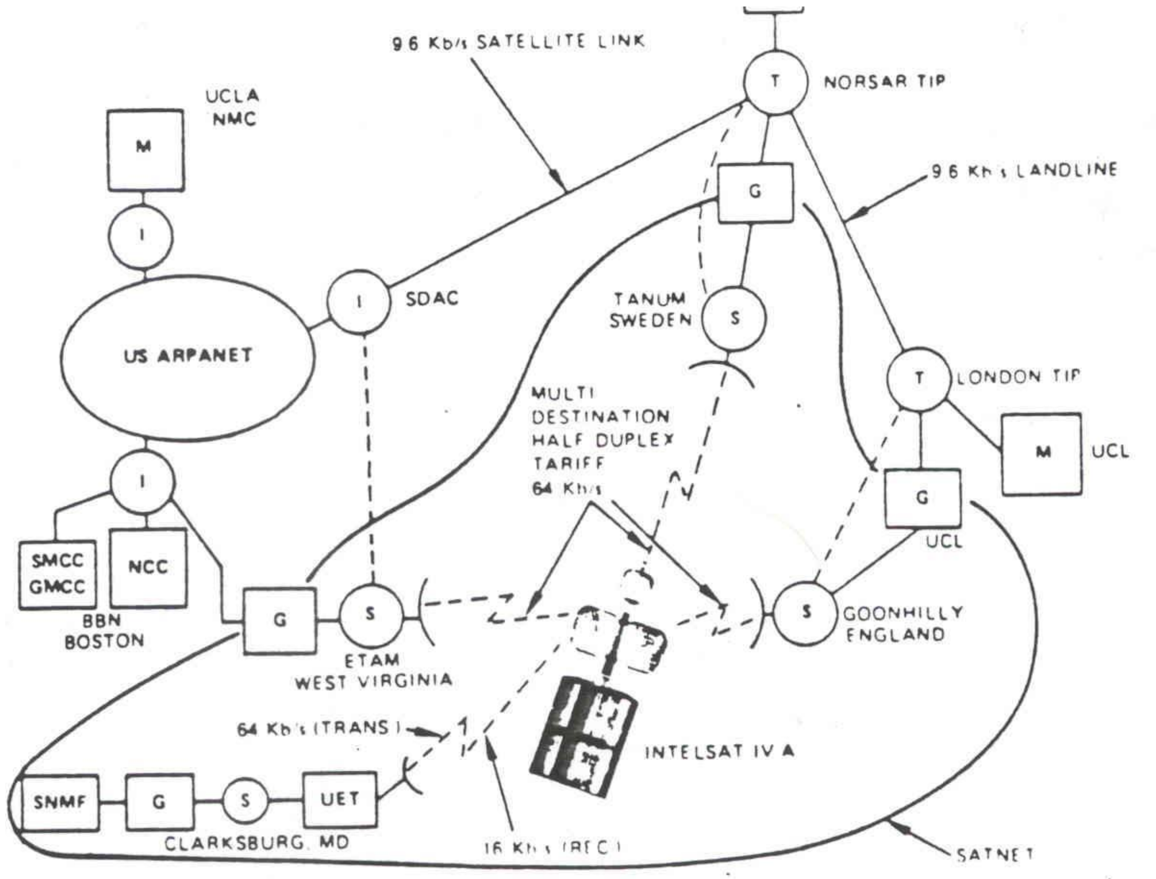


Internet measurements and performance evaluation

TCP a fine mouthwash available in Britain

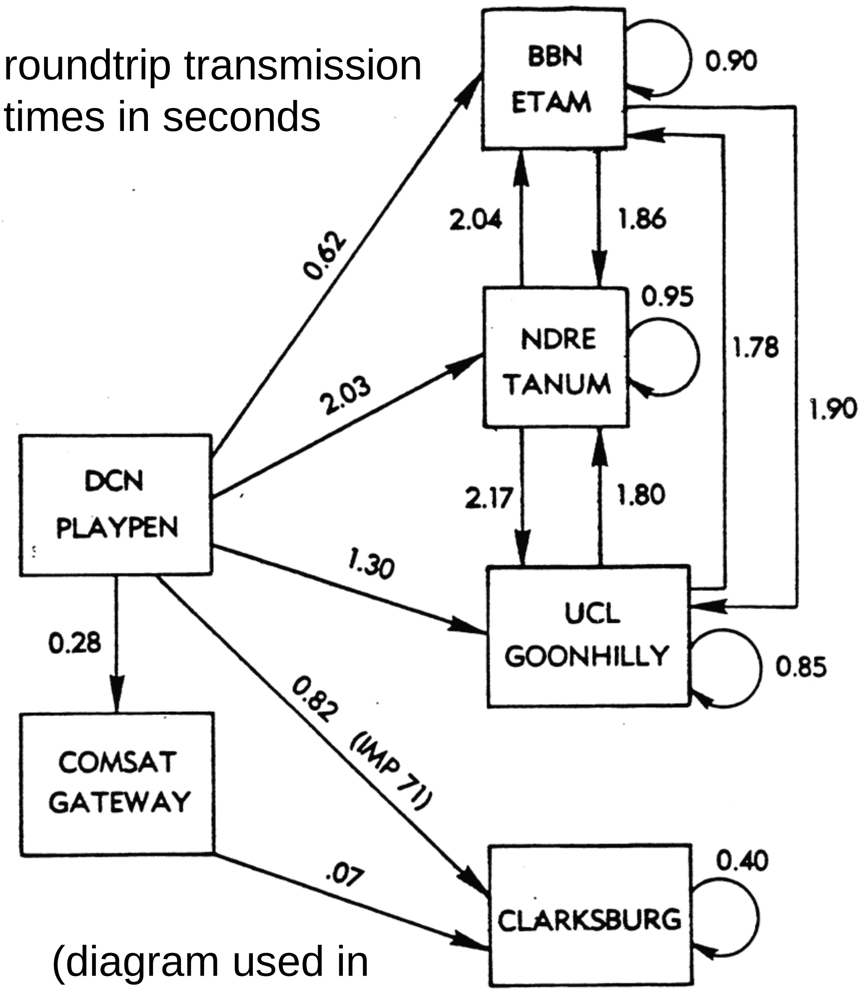
- While ARPAnet measurement tools had been highly developed, the Internet model forced many changes
- The objects to be measured and the measurement tools could be in far away places like foreign countries
- Four example programs are discussed
  - Atlantic Satellite Network (SATNET) measurement program
  - IP/TCP reassembly scheme
  - TCP retransmission timeout estimator
  - NTP scatter diagrams
- These weren't the last word at all, just steps along the way

DARPA Atlantic satellite network (SATnet)



SATnet measurement program

roundtrip transmission times in seconds



(diagram used in 1982 report)

- Earth stations in several countries were connected by a packet-switched INTELSAT satellite channel
- Stations supported scripted message generators and measurement tools
- Scripts were prepared transmitted via IP/TCP to experiment control program EXPAK, which ran in a designated ARPAnet host
- Once initiated, EXPAK launched the scripts and collected the results

TOPS-20 IP/TCP reassembly scheme

Seq	ID	Start	Length	Window	Offset
26250	36497	0	536	376	0
38321	36497	-536	536	912	0
39630	36634	536	376	912	0
40195	36498	0	536	0	0
41539	36648	0	536	376	0
54795	36648	-536	536	912	0
56096	36649	0	536	376	0
3695	36705	0	536	0	0
8939	36880	96	536	912	0
10263	36881	632	536	912	0
16224	36705	-440	536	0	0
17664	36961	256	536	912	0
27057	36881	-280	536	120	0
43698	36881	-1072	536	344	0
44825	37049	0	344	568	0
45623	37055	0	536	376	0
47021	37062	0	536	376	0
65365	37062	-536	536	912	0
1046	37063	0	536	376	0
2308	37148	0	536	0	0

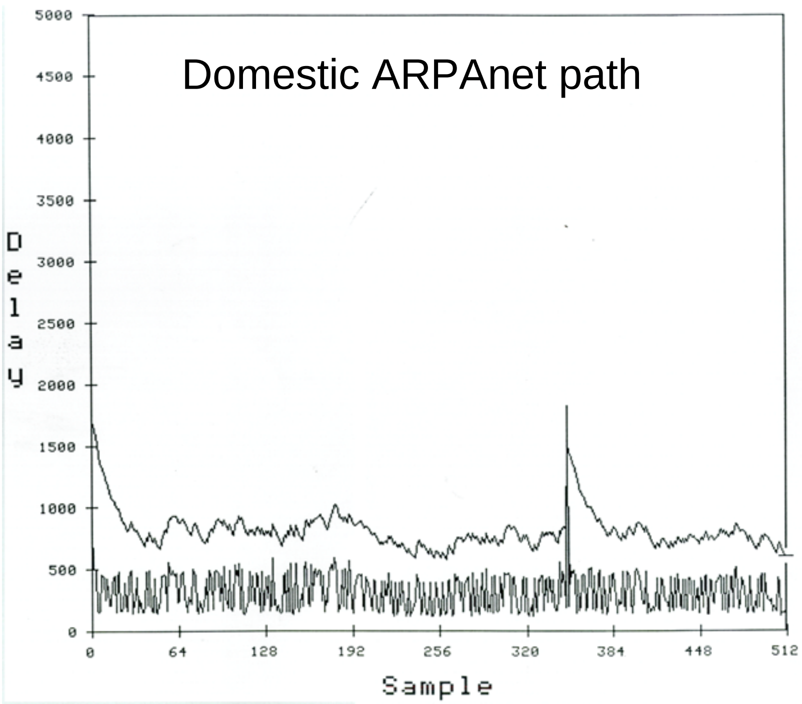
FTP: TOPS-20 - fuzzball 1200-bps device

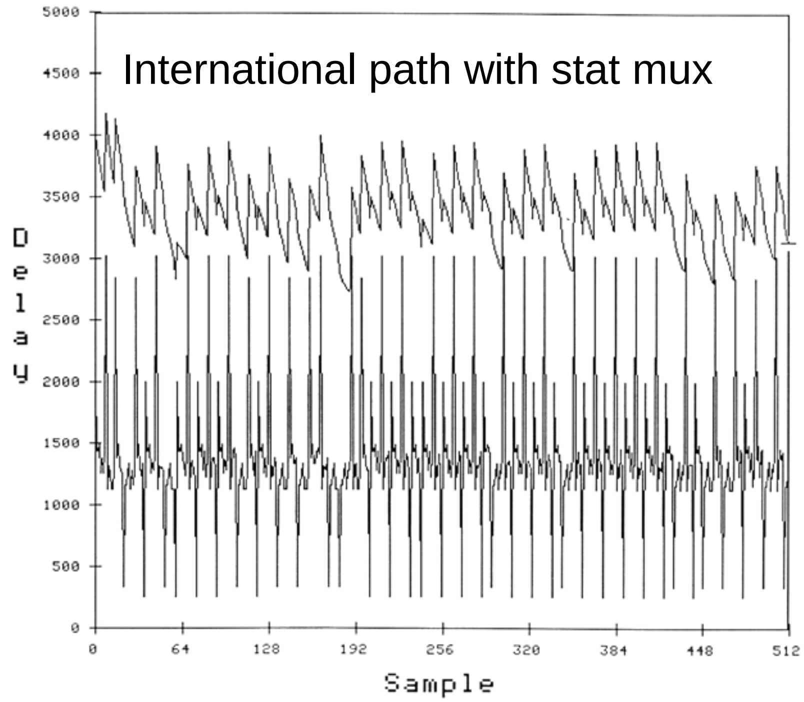
Seq = time of arrival (ms)
ID = IP sequence number
Start = packet start SN
Length = packet length
Window = size after store
Offset = ignore

(data shown circa 1980)

- Data shows TCP segments arriving via a seriously congested SATnet, which used 256-octet tinygrams
- A negative value in the `Start` field means an old duplicate
- A positive nonzero value means a lost packet and resulting hole
- TOPS-20 always retransmits the original packet and sequence number, which helped IP reassembly plug holes due to lost packets
- So far as known, this is lost art

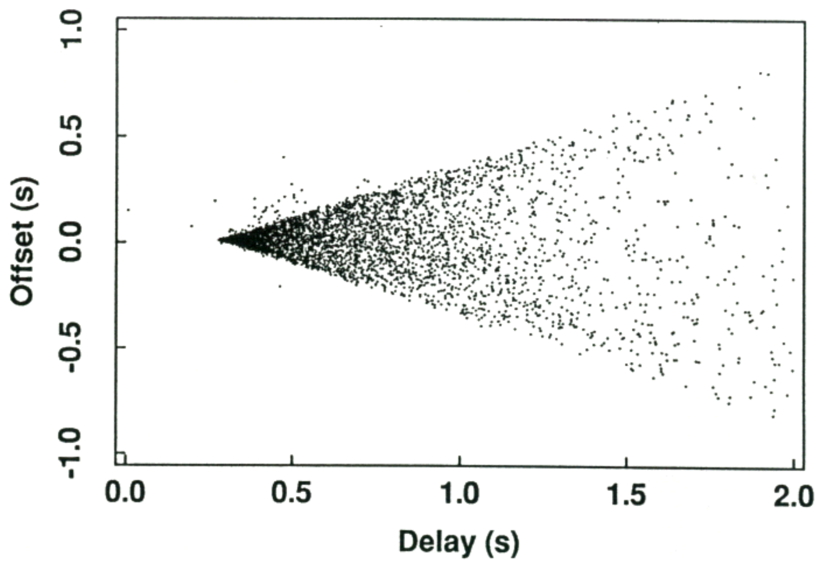
TCP retransmission timeout estimator

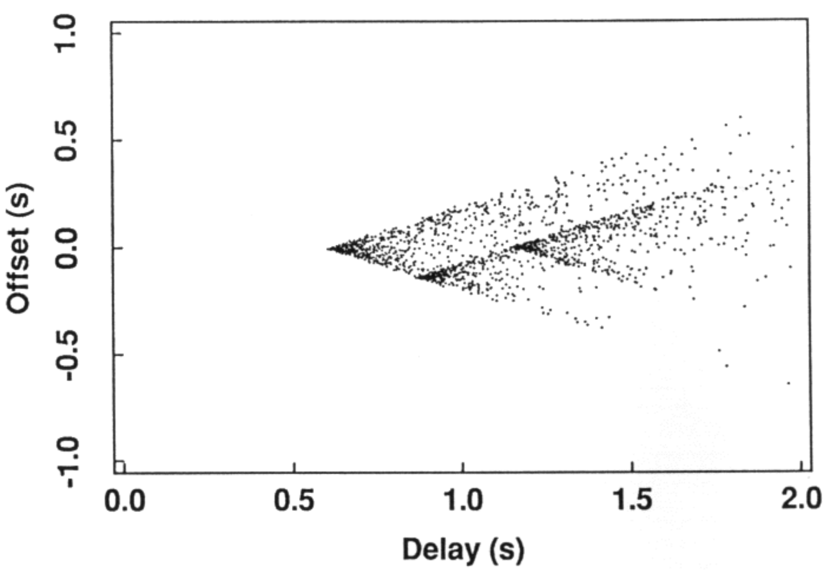




- These graphs show TCP roundtrip delay (bottom characteristic) and transmission timeout (top characteristic) for two different Internet paths
- The left diagram shows generally good prediction performance
- The right diagram shows generally miserable prediction performance
- The solution was to use different time constants for increase/decrease

NTP scatter diagrams





- These wedge diagrams show the time offset plotted against delay for individual NTP measurements
- For a properly operating measurement host, all points must be within the wedge (see proof elsewhere)
- The top diagram shows a typical characteristic with no route flapping
- The bottom diagram shows route flapping, in this case due to a previously unsuspected oscillation between landline and satellite links

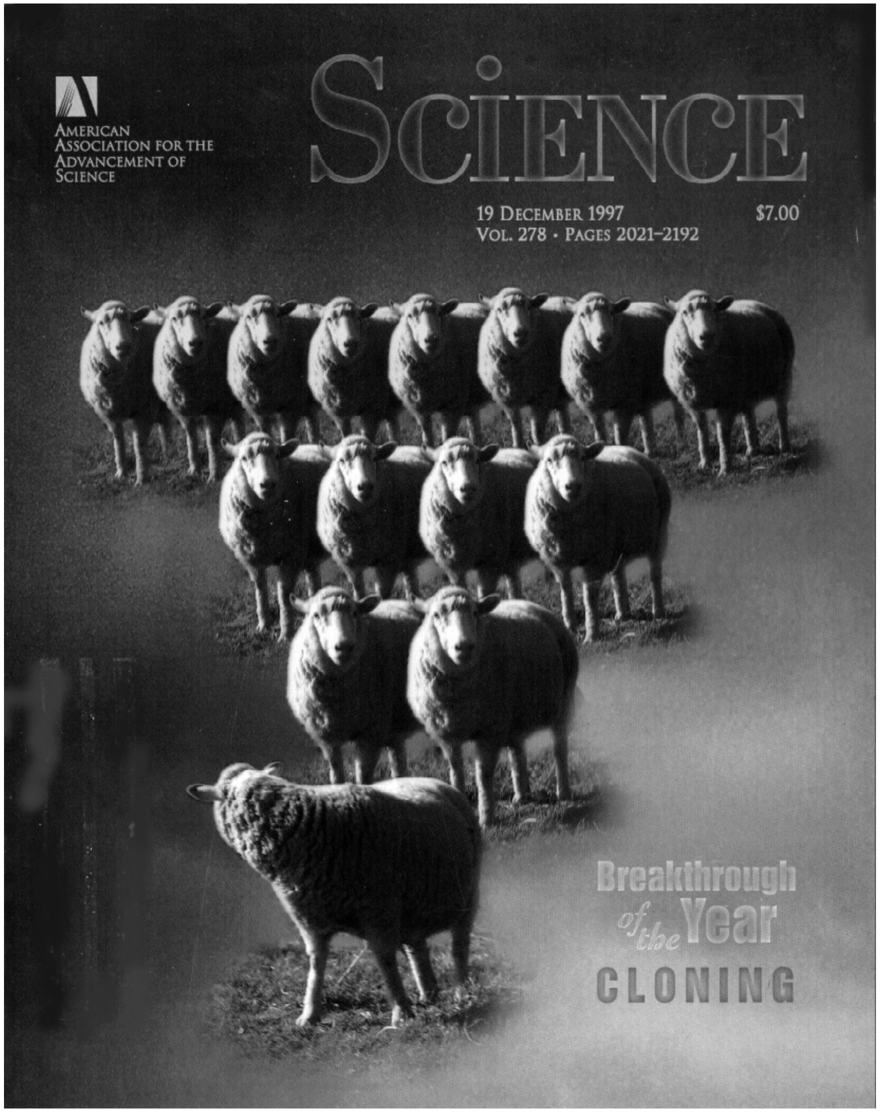
Autonomous system model

- There was every expectation that many incompatible routing protocols would be developed with different goals and reliability expectation
- There was great fear that gateway interoperability failures could lead to wide scale network meltdown
- The solution was thought to be a common interface protocol that could be used between gateway cliques, called autonomous systems
  - An autonomous system is a network of gateways operated by a responsible management entity and (at first) assumed to use a single routing protocol
  - The links between the gateways must be managed by the same entity
- Thus the Exterior Gateway Protocol (EGP), documented in rfc904
  - Direct and indirect (buddy) routing data exchange
  - Compressed routing updates scalable to 1000 networks or more
  - Hello neighbor reachability scheme modeled on new ARPAnet scheme
  - Network reachability field, later misused as routing metric

Unicore routing

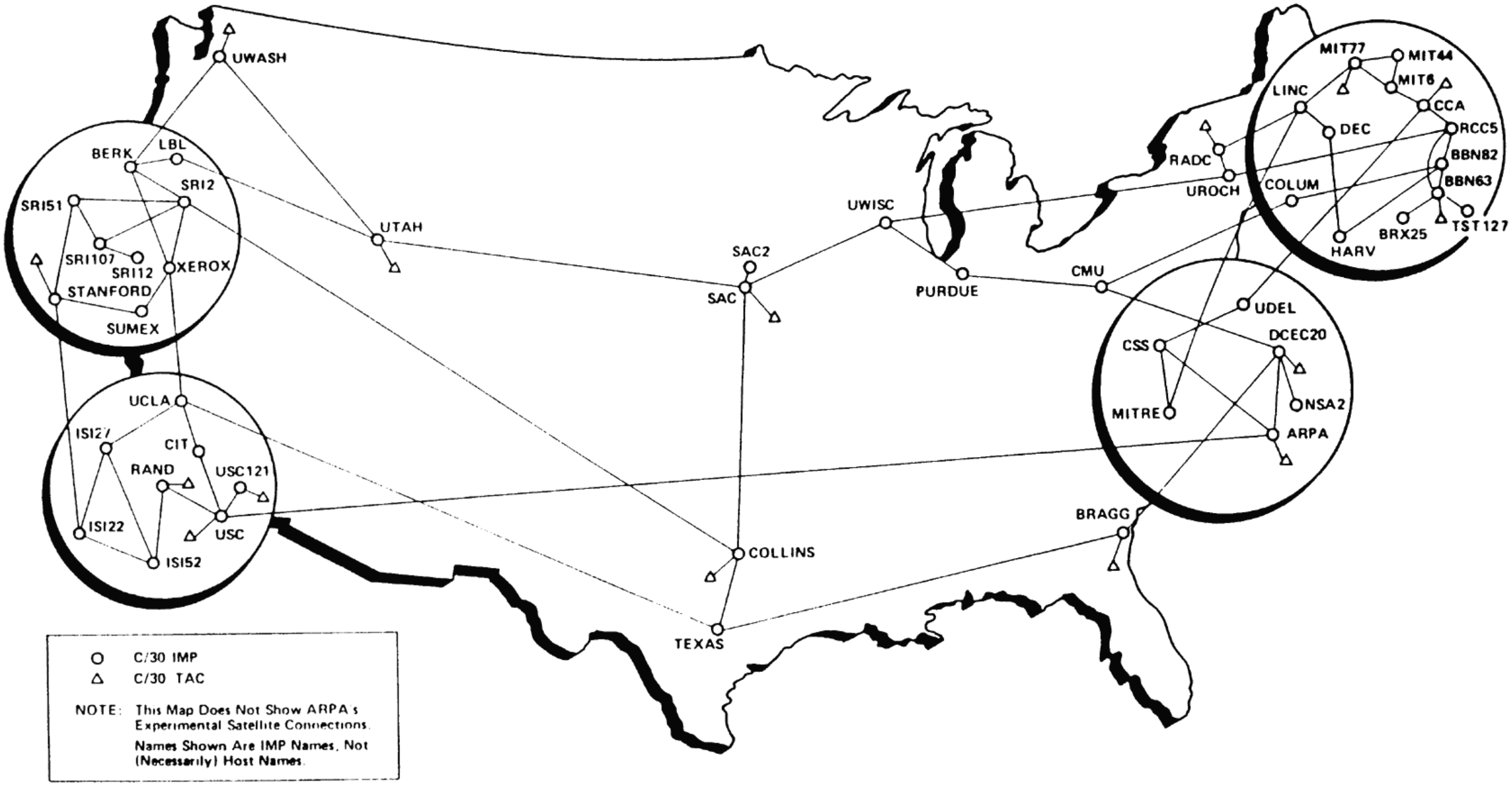
- The ICCB didn't trust any autonomous system, except a designated core system, to reveal networks not directly reachable in that system
  - The primary fear was the possibility of destructive, intersystem loops
  - A secondary fear was the possibility that not all network operating centers could detect and correct routing faults with equal enthusiasm
- This principle required that non-core gateways could not reveal networks reachable only via gateways of other systems
- While the unicore model insured stability, there were many problems
  - All traffic to systems not sharing a common network must transit the core system
  - All systems must have a gateway on a core network
  - Ad-hoc direct links between non-core systems could not be utilized by other systems
- While the unicore model was extended to multiple, hierarchical core systems (rfc975), this was never implemented

Intermission 1983-1990



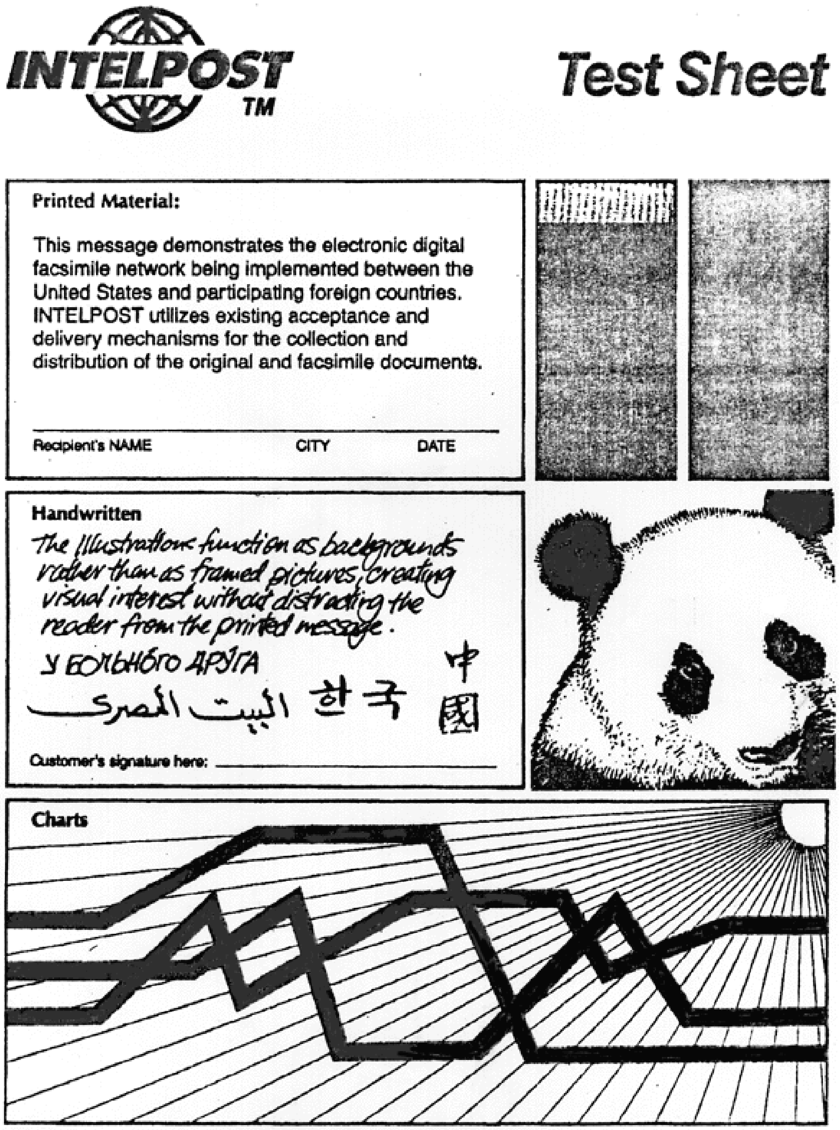
- Cloning the technology
- Decline of the ARPAnet
- INTELPOST as the first commercial IP/TCP network
- Evolution to multicore routing
- The NSFnet 1986 backbone network at 56 kb
- The NSFnet 1998 backbone network at 1.5 Mb
- The Fuzzball
- Internet time synchronization

ARPAnet topology August 1986



- ARPAnet was being phased out, but continued for awhile as NSFnet was established and expanded

INTELSAT network



- The first known commercial IP/TCP network was the INTELPOST fax network operated by the US, Canada and UK
- It was gatewayed to the Internet, but the only traffic carried past the gateway was measurement data
- The panda in the test sheet was originally scanned in London and transmitted via SATnet to the US during a demonstration held at a computer conference in 1979
- The panda image was widely used as a test page for much of the 1980s

Evolution to multicore routing

- NSF cut a deal with DARPA to use ARPAnet connectivity between research institutions until a national network could be put in place
- Meanwhile, NSF funded a backbone network connecting six supercomputer sites at 56 kb, later upgraded to 1.5 Mb
- The Internet routing centroid shifted from a single, tightly managed system to a loose confederation of interlocking systems
- There were in fact two core systems, the ICCB core and NSF core
  - The ICCB core consisted of the original EGP gateways connecting ARPAnet and MILnet
  - The NSF core consisted of Fuzzball routers at the six supercomputing sites and a few at other sites
- Other systems played with one or both cores and casually enforced the rules or not at all

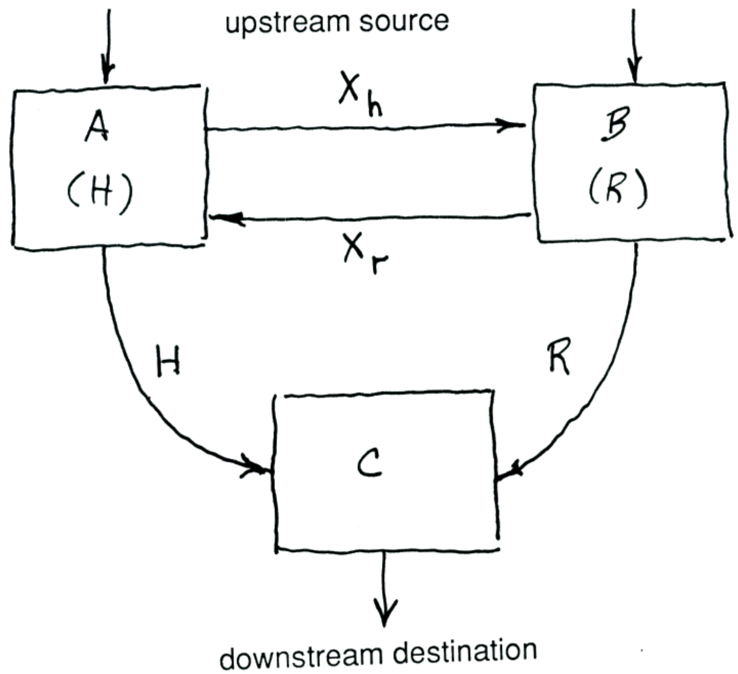
NSF 1986 backbone network

- The NSFnet phase-I backbone network (1986-1988) was the first large scale deployment of interdomain routing
- NSF supercomputing sites connected to the ARPAnet exchanged ICCB core routes using EGP
- Other NSF sites exchanged routes with backbone routers using Fuzzball Hello protocol and EGP
- All NSF sites used mix-and-match interior gateway protocols
- See: Mills, D.L., and H.-W. Braun. The NSFNET backbone network. *Proc. ACM SIGCOMM 87*, pp. 191-196

Septic routing – a dose of reality

- The NSF Internet was actually richly interconnected, but the global routing infrastructure was unaware of it
- In fact, the backbone was grossly overloaded, so routing operated something like a septic system
  - Sites not connected in any other way flushed packets to the NSF backbone septic tank
  - The tank drained through the nearest site connected to the ARPAnet
  - Sometimes the tank or drainage field backed up and emitted a stench
  - Sites connected to the ARPAnet casually leaked backdoor networks via EGP, breaking the third-party core rule
  - Traffic coming up-septic found the nearest EGP faucet and splashed back via the septic tank to the flusher's bowl
- Lesson learned: the multiple core model had no way to detect global routing loops and could easily turn into a gigantic packet oscillator

Metric transformation constraints



if $x =< F_r(F_h(x))$ and $y =< F_h(F_r(y))$

then

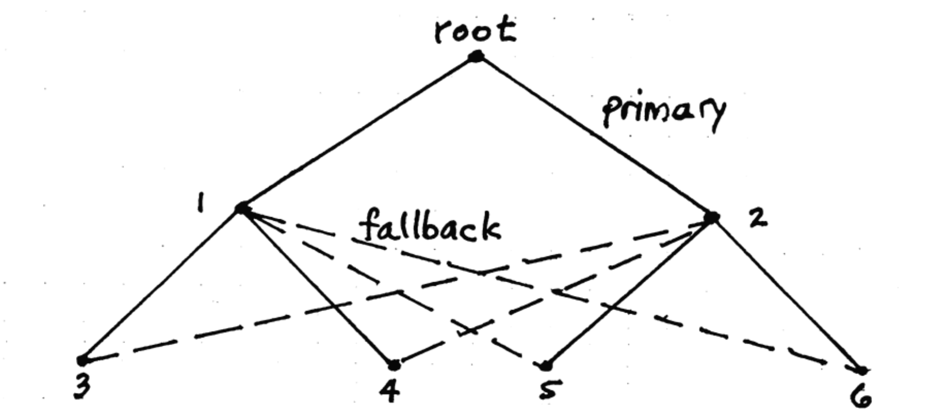
$X_h + F_h(R) < H$ implies $R < F_r(H) + X_r$

transformation constraints

(diagram used in
1986 presentation)

- The problem was preventing loops between delay-based Hello backbone routing algorithm and hop-based RIP local routing algorithm
- The solution diagrammed a left was a set of provable metric transformation constraints
- This didn't always work, since some nets were multiply connected and didn't present the same metric for the same network
- One should never have to do this, but it does represent an example of panic engineering

Fallback routing principle

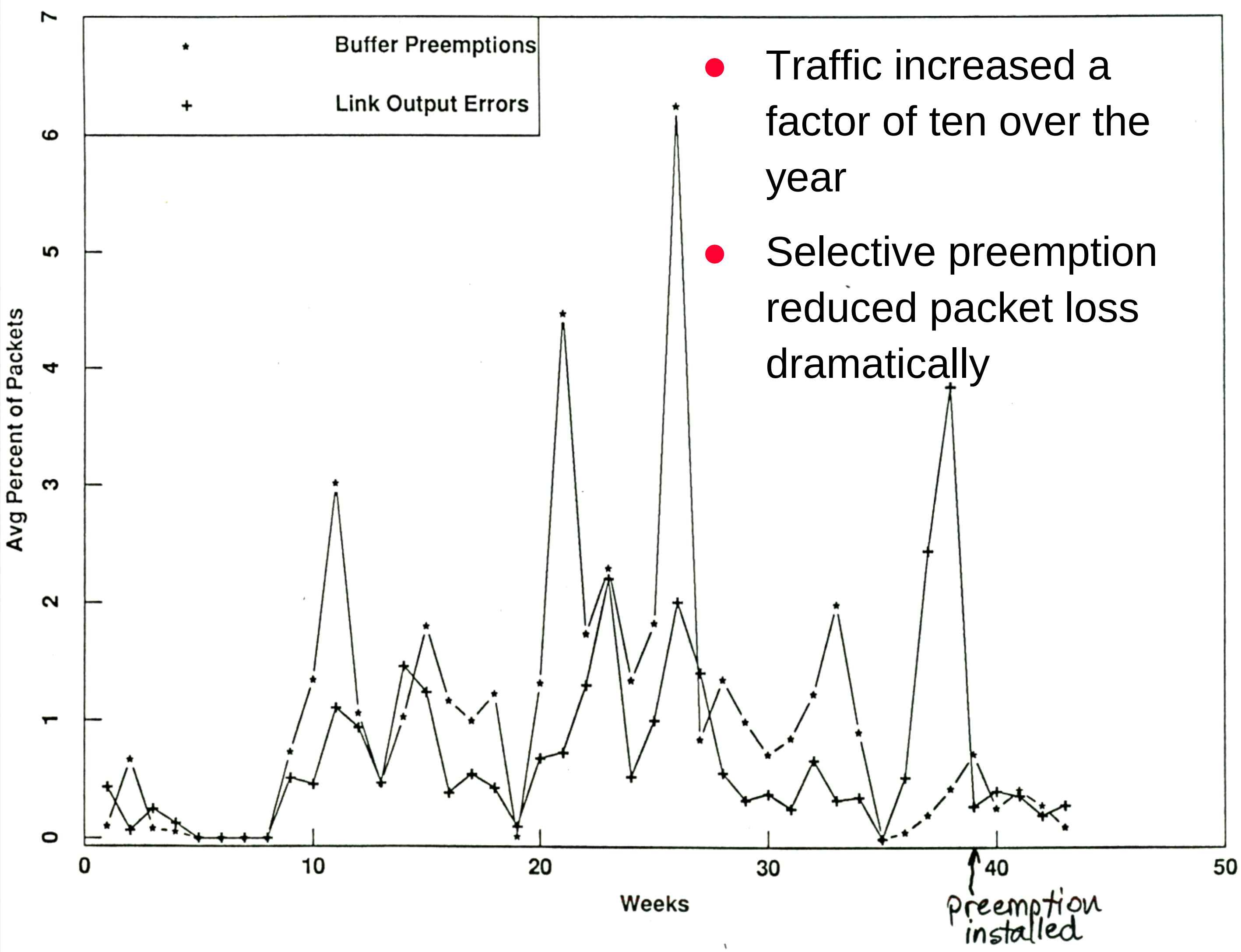


- Reachability (EGP model): spanning tree is static and pre-engineered; link up/down states are determined dynamically
- Fallback (EGP practice): primary and fallback routes are pre-engineered to form a spanning tree (avoid loops) under all failure scenarios; link up/fallback/down states are determined dynamically
- Comprehensive: spanning trees are computed from measured link data according to specified metric

(diagram used in 1986 presentation)

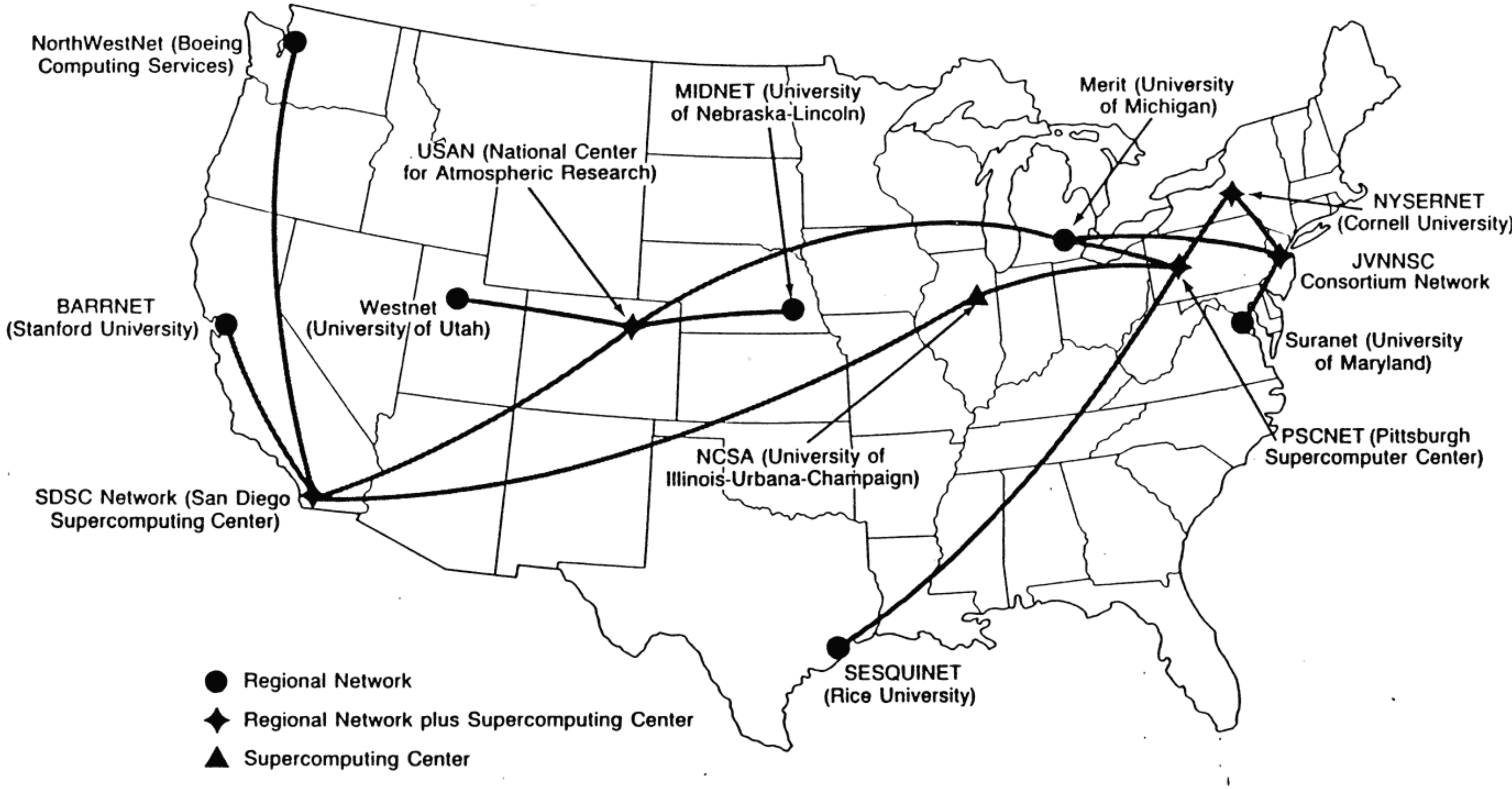
- The problem was how to handle routing with the ICCB core and the NSFnet core, so each could be a fallback for the other
- The solution was to use the EGP reachability field as a routing metric, but to bias the metric in such a way that loops could be prevented under all credible failure conditions
- Success depended on a careful topological analysis of both cores
- But, we couldn't keep up with the burgeoning number of private intersystem connections

Fuzzball selective preemption strategy



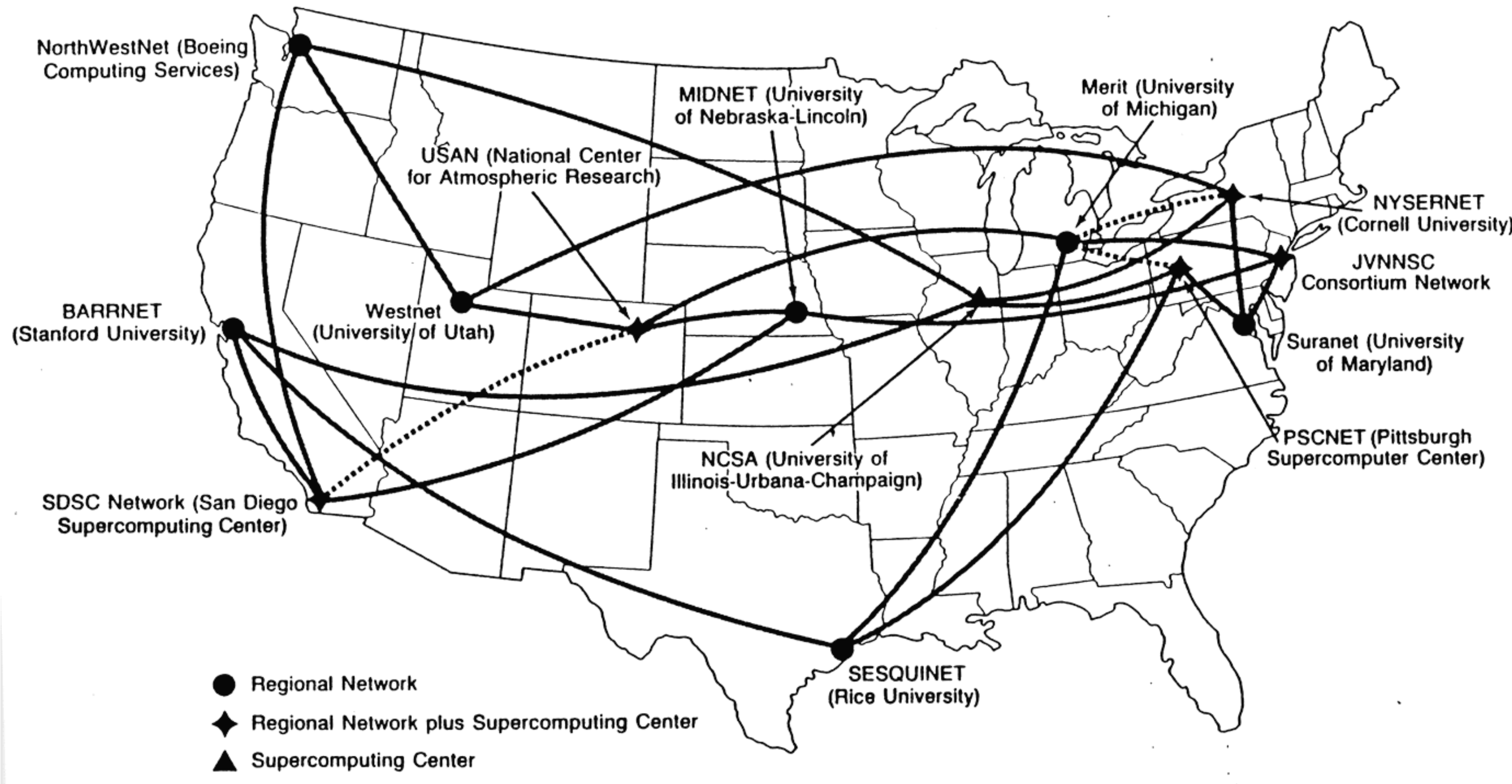
- Traffic increased a factor of ten over the year
- Selective preemption reduced packet loss dramatically

NSFnet 1988 backbone physical topology



- This physical topology was created using T1 links as shown
- All sites used multiple IBM RT routers and multiplexors to create reconfigurable virtual channels and split the load

NSFnet 1988 backbone logical topology



- This logical topology was created from the T1 virtual channels and backhaul, which resulted in surprising outages when a good ol' boy shotgunned the fiber passing over a Louisiana swamp
- Backhaul also reduced the capacity of some links below T1 speed

Things learned from the early NSFnet experience

- We learned that finding the elephants and shooting them until the forest is safe for mice was the single most effective form of congestion control
- We learned that managing the global Internet could not be done by any single authority, but of necessity must be done by consensus between mutual partners
- We learned that network congestion and link level-retransmissions can lead to global gridlock
- We learned that routing instability within a system must never be allowed to destabilize neighbor systems
- We learned that routing paradigms used in different systems can and will have incommensurate political and economic goals and constraints that have nothing to do with good engineering principles
- Finally, we learned that the Internet cannot be engineered – it must grow and mutate while feeding on whatever technology is available

The Fuzzball

Dry cleaner advertisement
found in a local paper

- The Fuzzball was one of the first network workstations designed specifically for network protocol development, testing and evaluation
- It was based on PDP11 architecture and a virtual operating system salvaged from earlier projects
- They were cloned in dozens of personal workstations, gateways and time servers in the US and Europe

Mommy, what's a Fuzzball?

- On the left is a LSI-11 Fuzzball, together with control box and 1200-bps modem. Telnet, FTP, mail and other protocols were first tested on this machine and its friends at ISI, SRI, MIT and UCL (London).
- On the right is the last known Fuzzball, now in my basement.
- More at www.eecis.udel.edu/~mills and the citations there.

Rise and fall of the Fuzzball

- From 1978, PDP11 and LSI-11 Fuzzballs served in Internet research programs
  - as testbeds for all major IP and TCP protocols and applications
  - in numerous demonstrations and coming-out parties
  - as measurement hosts deployed at SATnet terminals in the US, UK, Norway, Germany and at military sites in several countries
- During the period 1986-1988 they served as routers in the NSFnet phase-I backbone network
- The IP/TCP and routing code was deployed in the INTELPOST network operated by the US, Canada and UK postal services and COMSAT
- Fuzzballs were increasingly replaced by modern RISC machines starting in 1988. The last known one spun down in the early 90s
- See: Mills, D.L. The Fuzzball. *Proc. ACM SIGCOMM 88*, pp. 115-122

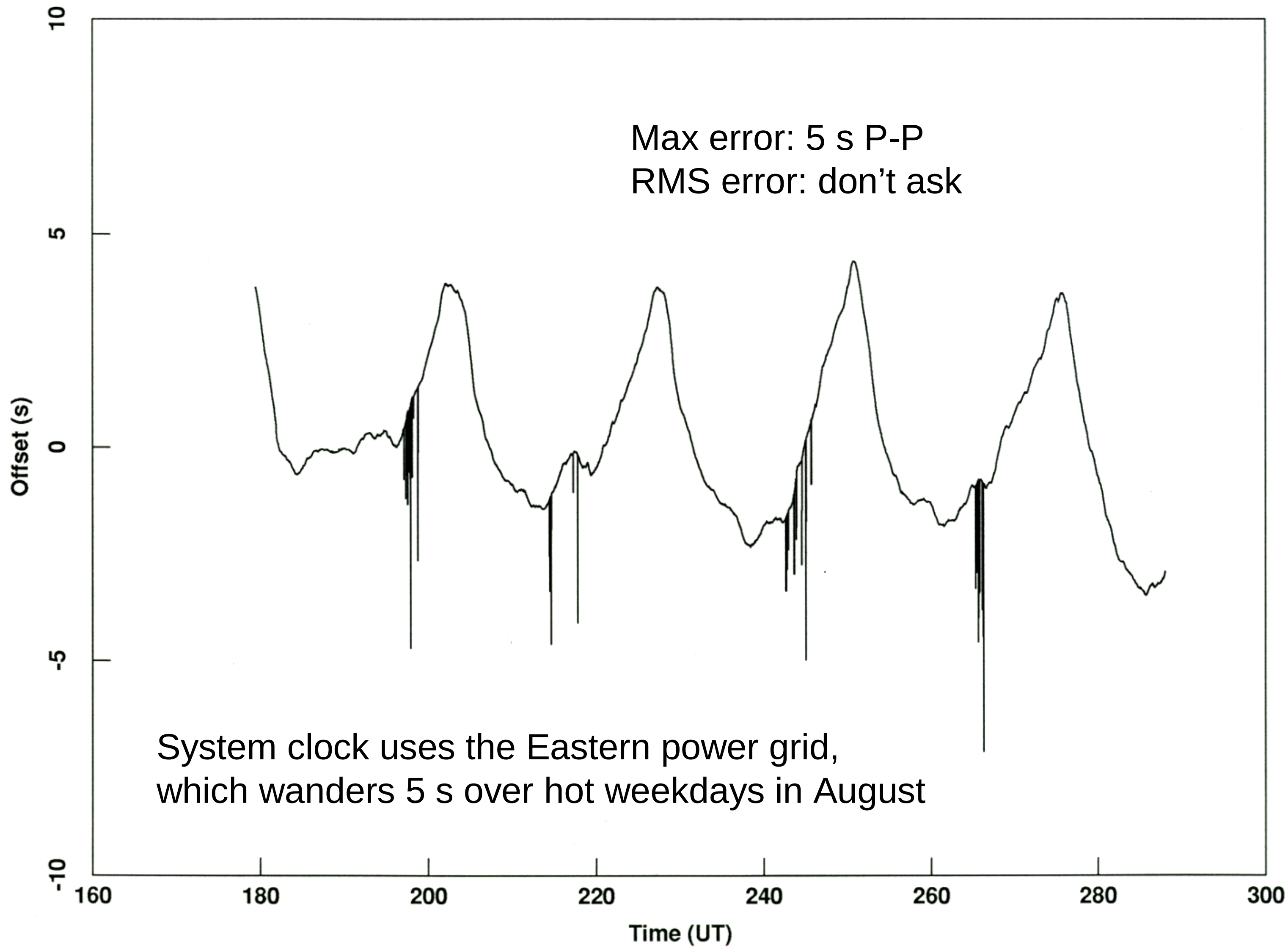
Internet time synchronization

- The Network Time Protocol (NTP) synchronizes many thousands of hosts and routers in the public Internet and behind firewalls
- At the end of the century there are 90 public primary time servers and 118 public secondary time servers, plus numerous private servers
- NTP software has been ported to two-dozen architectures and systems

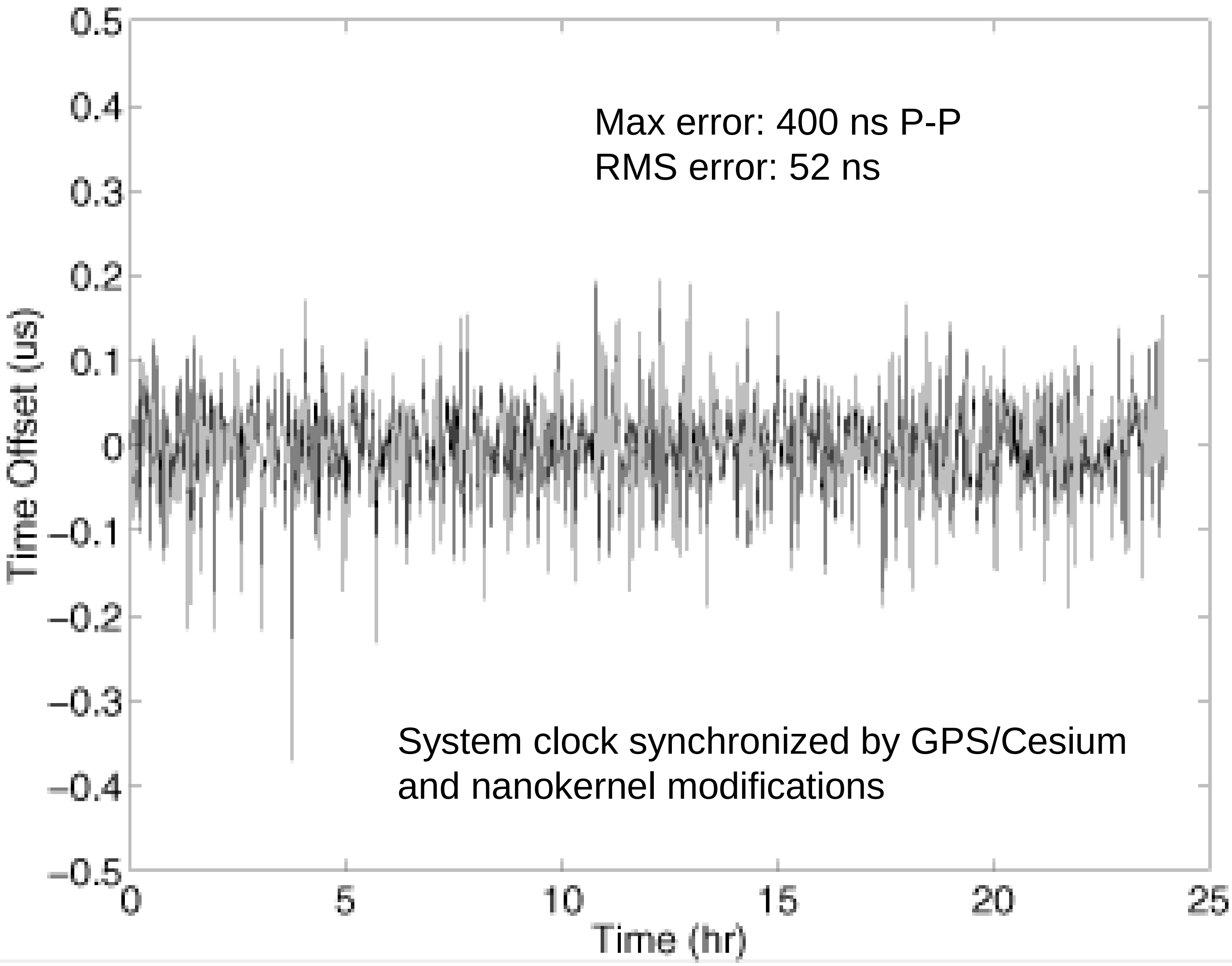
A brief history of network time

- Time began in the Fuzzball *circa* 1979
  - Fuzzball hosts and gateways were synchronized using timestamps embedded in the Hello routing protocol
  - Since 1984, Internet hosts and gateways have been synchronized using the Network Time Protocol (NTP)
  - In 1981, four Spectracom WWVB receivers were deployed as primary reference sources for the Internet. Two of these are still in regular operation, a third is a spare, the fourth is in the Boston Computer Museum
  - The NTP subnet of Fuzzball primary time servers provided synchronization throughout the Internet of the eighties to within a few tens of milliseconds
- Timekeeping technology has evolved continuously over 20 years
  - Current NTP Version 4 improves performance, security and reliability
  - Engineered Unix kernel modifications improve accuracy to the order of a few tens of nanoseconds with precision sources
  - NTP subnet now deployed worldwide in many thousands of hosts and routers of government, scientific, commercial and educational institutions

Timetelling in 1979



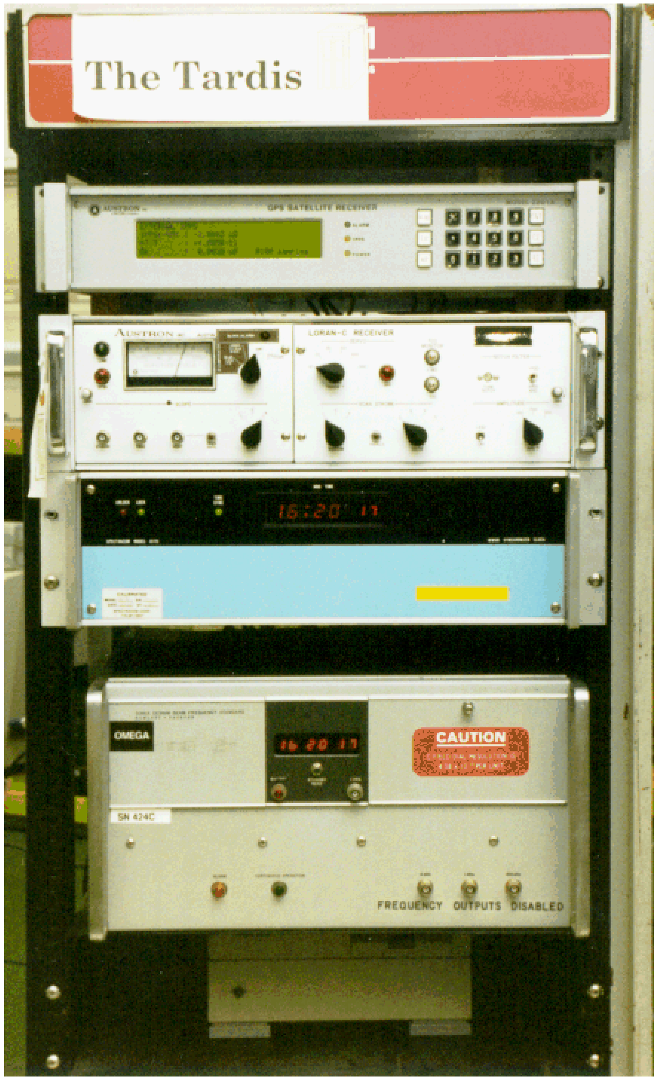
Timetelling in 1999



Lessons learned from NTP development program

- Synchronizing global clocks with submillisecond accuracy enables
  - the exact incidence of global events to be accurately determined
  - real time synchronization of applications such as multimedia conferencing
- Time synchronization must be extremely reliable, even if it isn't exquisitely accurate. This requires
  - certificate based cryptographic source authentication
  - autonomous configuration of servers and clients in the global Internet
- Observations of time and frequency can reveal intricate behavior
  - Usually, the first indication that some hardware or operating system component is misbehaving are synchronization wobbles
  - NTP makes a good fire detector and air conditioning monitor by closely watching temperature-dependent system clock frequency wander
  - Statistics collected in regular operation can reveal subtle network behavior and routing Byzantia
  - NTP makes a good remote reachability monitor, since updates occur continuously at non-intrusive rates

NTP Master Clock



Austron 2100A GPS Receiver
1988, $17K

Austron 2000 LORAN-C Receiver
1988, $40K

Spectracom 8170 WWVB Receiver
1981, $3K

HP 5061A Cesium Frequency Standard
1972, $75K

Selected bibliography

- NSFnet and the Fuzzball
  - Mills, D.L. The Fuzzball. *Proc. ACM SIGCOMM 88 Symposium* (Palo Alto CA, August 1988), 115-122.
  - Mills, D.L., and H.-W. Braun. The NSFNET Backbone Network. *Proc. ACM SIGCOMM 87 Symposium* (Stoweflake VT, August 1987),191-196.
- Lessons of the Internet, *inter alia*
  - Leiner, B., J. Postel, R. Cole and D. Mills. The DARPA Internet protocol suite. In: P.E. Green (Ed.), *Network Interconnection and Protocol Conversion*, IEEE Press, New York, NY, 1988. Also in: *IEEE Communications 23, 3* (March 1985), 29-34.
  - Mills, D.L. Internet time synchronization: the Network Time Protocol. *IEEE Trans. Communications COM-39, 10* (October 1991), 1482-1493. Also in: Yang, Z., and T.A. Marsland (Eds.). *Global States and Time in Distributed Systems. IEEE Computer Society Press*, Los Alamitos, CA, 1994, 91-102.
  - Additional references: www.eecis.udel.edu/~mills/ntp.htm

Selected bibliography (continued)

- Atlantic Satellite Network (SATnet) measurement program
  - Palmer, L., J. Kaiser, S. Rothschild and D. Mills. SATNET packet data transmission. *COMSAT Technical Review 12, 1* (Spring 1982), 181-212.
  - Cudhea, P.W., D.A. McNeill and D.L. Mills. SATNET operations. *Proc. AIAA Ninth Communications Satellite Systems Conference* (March 1982).
  - Mills, D.L. Internetworking and the Atlantic SATNET. *Proc. National Electronics Conference* (Chicago, Illinois, October 1981), 378-383.
  - Chu, W.W., D.L. Mills, et al. Experimental results on the packet satellite network. *Proc. National Telecommunications Conference* (Washington, D.C., November 1979), 45.4.1-45.4.12.
  - Kirstein, P., D.L. Mills, et al. SATNET application activities. *Proc. National Telecommunications Conference* (Washington, D.C., November 1979), 45.5.1-45.5.7.

Lessons of history and tall tales

- Epic Allegories and Didactic Legends
  - Cohen, D. The Oceanview Tales. Information Sciences Institute, Marina del Rey, CA, 1979, 17 pp.
  - Finnegan, J. The world according to Finnegan (as told to Danny Cohen and Jon Postel), Information Sciences Institute, Marina del Rey, CA, 1982, 34 pp.
- Milestone and planning documents
  - Federal Research Internet Coordinating Committee. Program plan for the National Research and Education Network. US Department of Energy, Office of Scientific Computing ER-7, Washington, DC, 1989, 24 pp.
  - Roessner, D., B. Bozeman, I. Feller, C. Hill, N. Newman. The role of NSF's support of engineering in enabling technological innovation. Executive Office of the President, Office of Science and Technology Policy, 1987, 29 pp.
    - Contains an extended technical history of the Internet and NSF involvement