TCP/IP Router Performance

Scott Bradner
Harvard University

Acknowledgments

Dan Lanciani produced all of the software and was instrumental in deciding what tests to run.

Kent England was critical in navigating the obscure waters of the Proteon "user" interface.

Jerry Lotto captured the routing packets and gathered the data about the packet size distribution on the Harvard network.

What are we talking about?

Bridge:
A device that connects two or more networks and forwards packets between them. Bridges operate at the ISO physical layer. Networks connected by bridges operate as if they were a single network.

Router:
A device that connects two or more networks and forwards packets between them. Routers normally operate at the ISO network layer. Networks that are connected by routers operate as separate networks.

Gateway:
A device that connects two or more networks and forwards packets between them. Gateways operate above the ISO network layer. Gateways may be used to translate functions of one protocol to equivalent functions in another protocol. Networks that are connected by gateways operate as separate networks.

Brouter:
A device that can operate both as a bridge and as a router. Usually a brouter can operate as a router for specific network protocols while simultaneously acting as a bridge for others.

SO's view of the world:

7	Application
6	Presentation
5	Session
4	Transport
3	Network
2	Data Link
1	Physical

Life on a real world network: Harvard

Harvard's existing network is the result of largely unplanned interconnection of building LANs. A plan has been drawn up for a coordinated system involving a back-bone network of 7 nodes connected, at first with ethernet, then as traffic dictates, FDDI. The network supports both TCP/IP and DECnet.

Harvard's network, now:

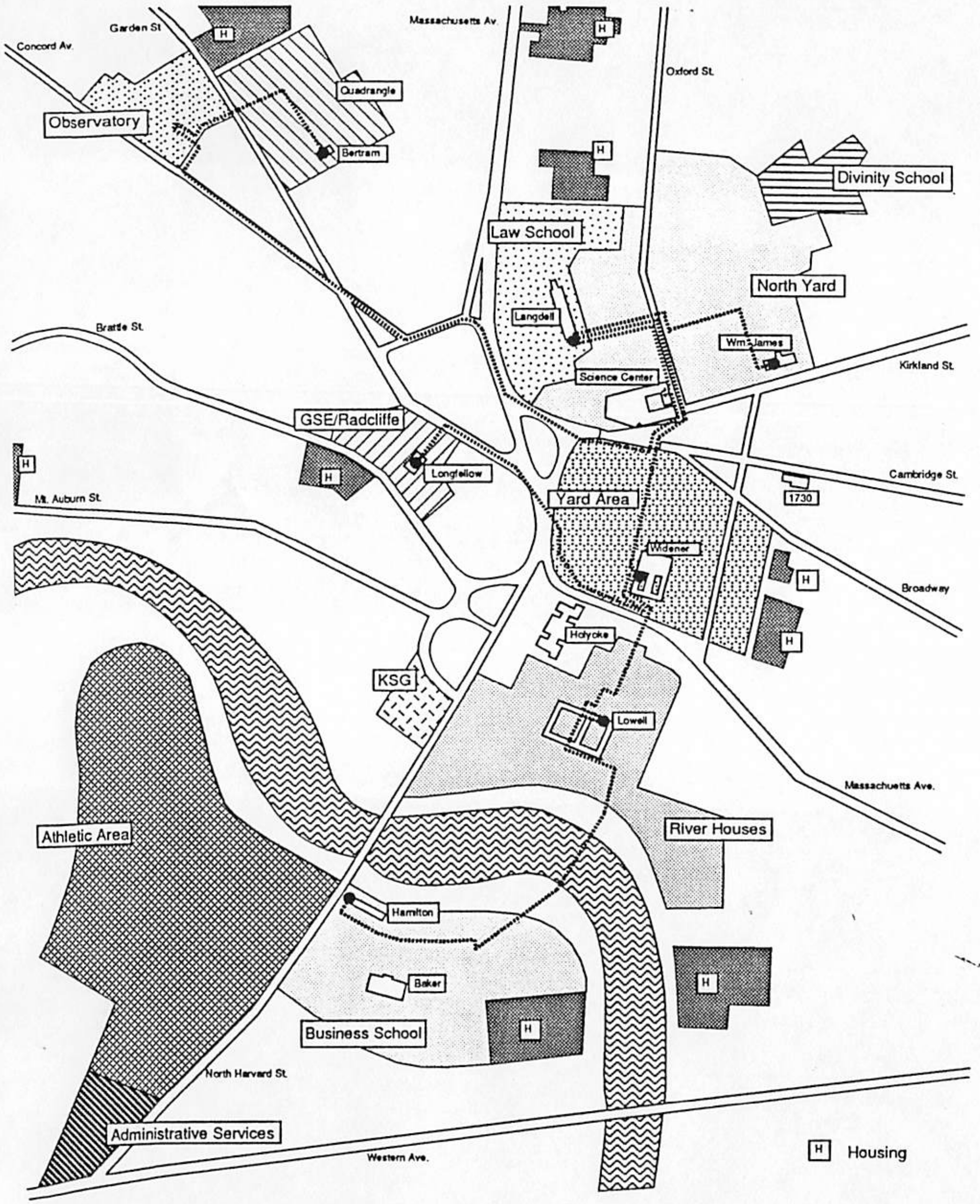
- About 600 ip nodes
- Many DECnet nodes
- About 30 subnets
- About 8 router connected nets
- Rest AppleTalk etc
- Lots of NFS
- Some PC clusters

Harvard's network, planned:

- About 2,000 ip nodes
- Few DECnet only nodes
- About 100 subnets
- About 50 router connected nets growing to 100s
- More NFS
- 100's of PC clusters

Copies of the RFI for this network can be FTPd from husc6.harvard.edu (128.103.1.56).

Harvard HSDN



Life on a real world network: NEARnet

NEARnet is a NSF regional network serving the northeast. Its backbone consists of a series of 10 MB microwave ethernet connections in the area around Boston Massachusetts. Branch nodes are connected to this backbone using leased lines at rates from 9.6KB to T1. The network supports TCP/IP only. Routers are used at the backbone nodes and at each member site.

NEARnet, now:

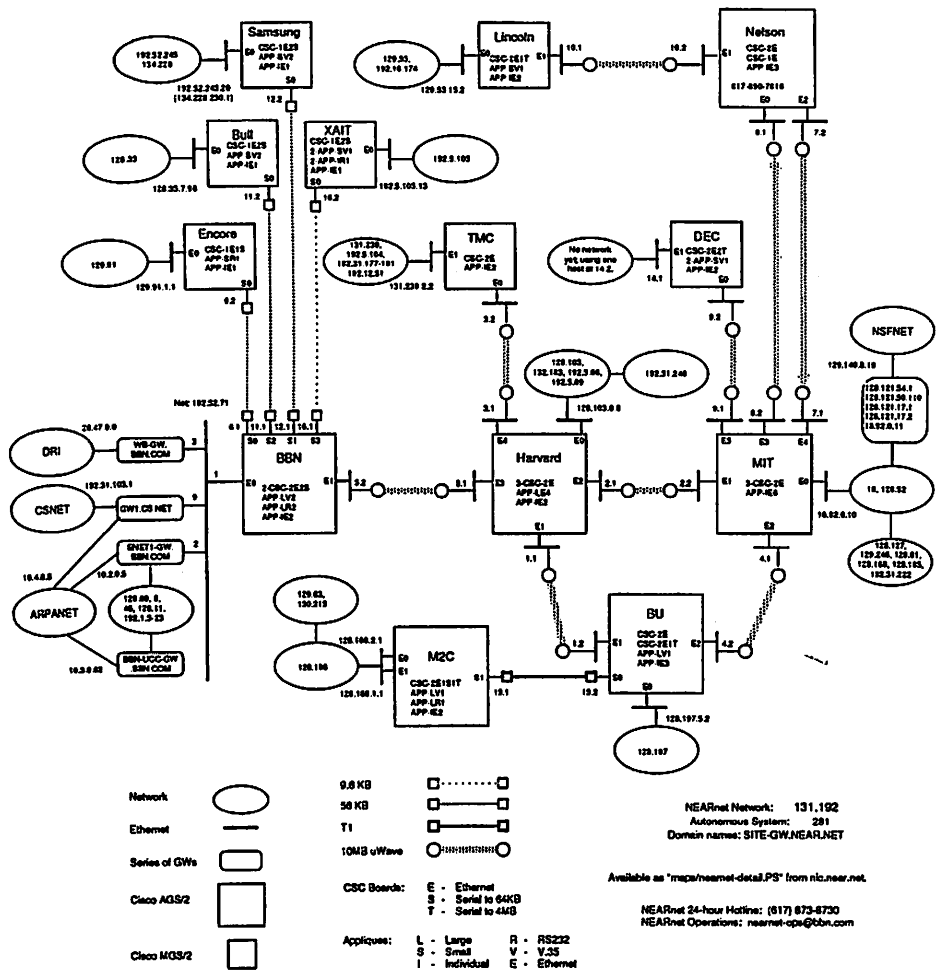
- 12 members
- 6 connected with 10MB microwave
- Rest from 9.6KB to T1

NEARnet, within a year:

- 50 members
- 10 at 10MB

A diagram of the current design of NEARnet can be FTPd from canapes.bbn.com (128.89.0.214).

NEARnet



Life on a real world network:

Pathological conditions:

Peak load

- arp storms
- broadcast storms
- rwho on diskless nodes
- bootp
- tftp booting

Back-to-back packets

- NFS traffic
- routing updates

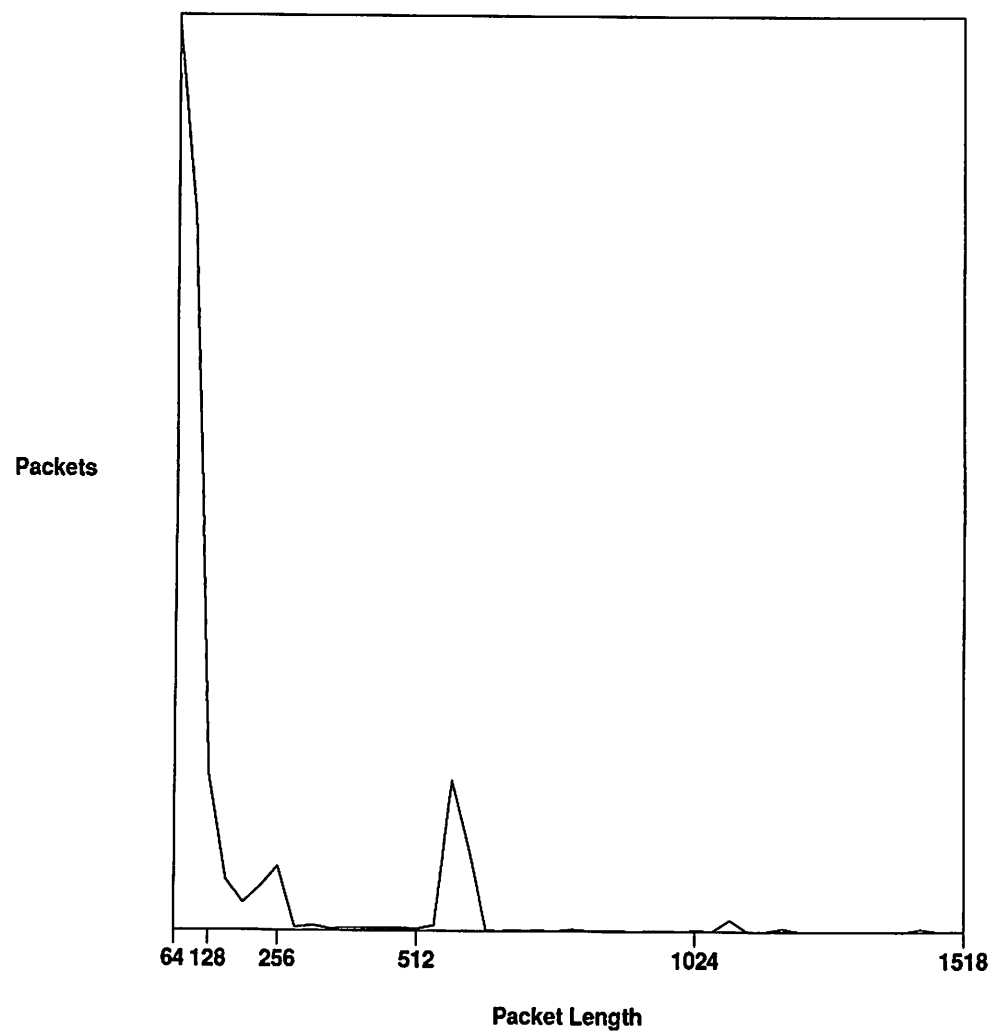
Life on a real world network:

"Normal" conditions

- NFS servers
- named
- NNTP
- SMTP
- PC clusters
- FTP
- terminal servers

Life on a real world network:

Packet length distribution on Harvard backbone



Life on a real world network:

Potential zaps

Network management
- standards
  - SNMP, CMOT
- proprietary

Documentation
- Fit for human consumption?

User interface
- How expensive a guru is needed?

Reachibality
- Can it be managed over the network?
- How easy is it to crash the router so that it requires manual intervention?
- How easy is it overload the router so that the processor does not respond to commands on the serial line?

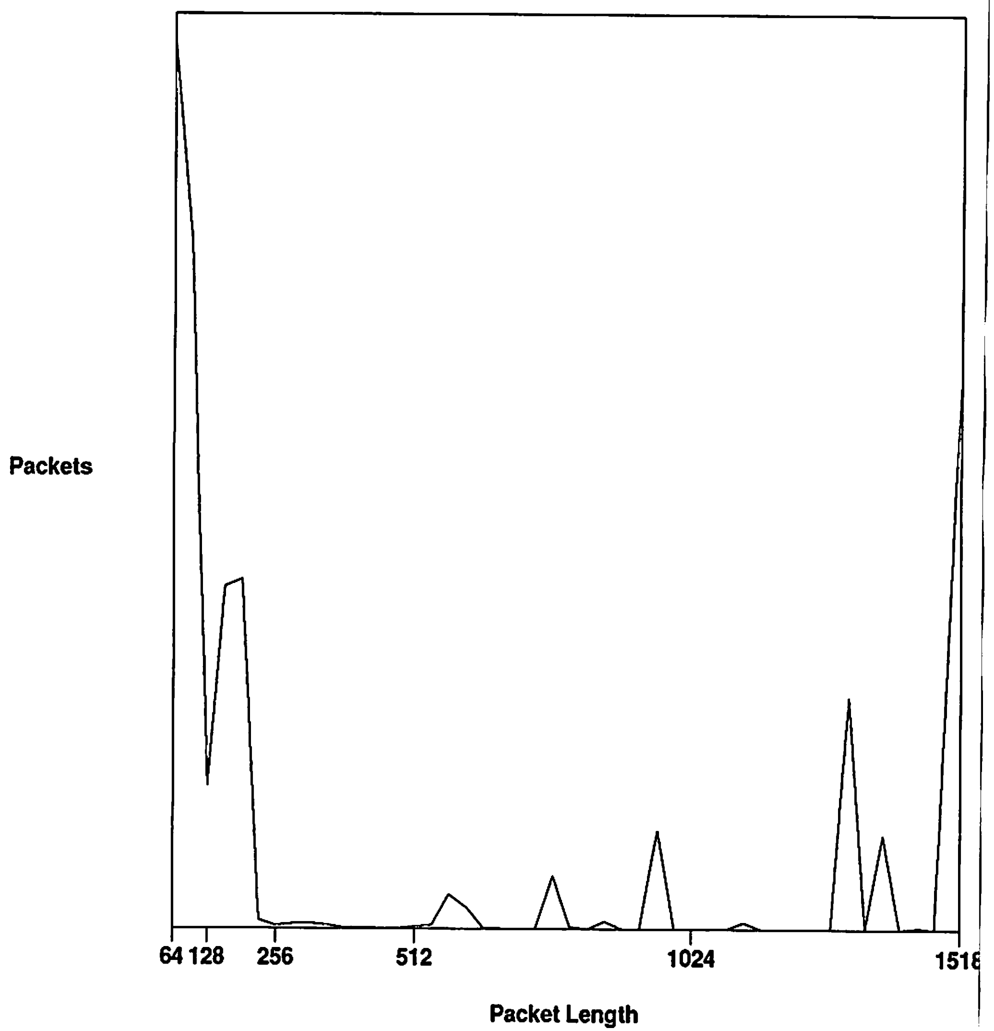
Life on a real world network:

Security

- What access controls on router?
- What sorts of filtering can be done on traffic?
  - On source of traffic.
  - On destination of traffic.
  - On protocol type?

Life on a real world network:

Packet length distribution on Harvard subnet



Life on a real world network:

Routing protocols

- Standards
  - RIP
    Routing Information Protocol
  - EGP
    Exterior Gateway Protocol
  - HELLO

  - OSPFIGP (OSPF)
    Open Shortest-Path-First
    Interior Gateway Protocol
  - BGP
    Border Gateway Protocol

- Proprietary
  - IGRP
    Interior Gateway Routing Protocol
    (cisco)

Testing, how to simulate real world.

- Can't do a very good job of simulating the "real world".

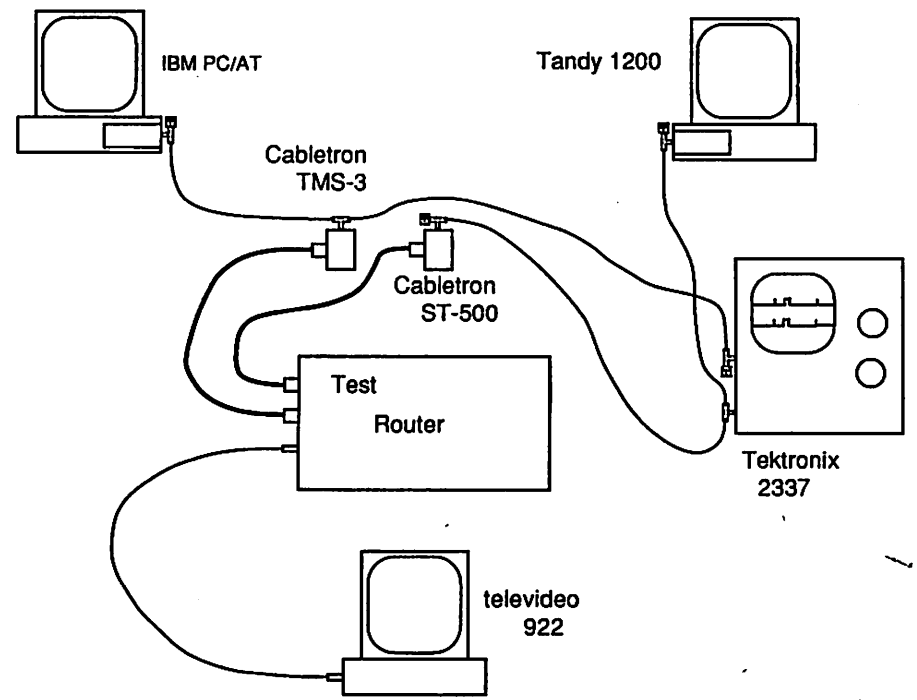
- Easy to check simple things.
  - Idle state.
  - Delay through router.
  - Effects of various filtering options.
  - Accuracy of counters.
  - Reaction to error packets.

- Not too hard to simulate the pathological conditions.
  - High offered load.
  - Back to back packets.

- Much harder to test for table space related limits.
  - Routing table size.
  - Arp cache size.
  - Filtering list space.

We put together a setup that would do the easy tests, and punted on the harder ones.

Test system:



Packet source

- IBM PC/AT (old)

- NI5210 with 16KB of on-board memory

- Uses Intel 82586 Local Area Network Coprocessor
  - Uses buffer chaining design
    - Pointer to "next" next buffer control block
    - Can set up loop
      - A points to B, B points to A
  - Data put in on-board ram, no PC access needed

- Can send continuous stream of packets with an
  interpacket gap of 55usec
  (9.6usec legal minimum).

Packet counter

- Tandy 1200 (old)

- NI5210 with 16KB of on board memory

- Put chip into resource exhaustion mode
- Chip will count missed packets.

Hammer

Packet generating program.

```
hammer [-taddr] [-s#] [-c] [+#] [-n#] file[*#]
    [file ...]

 -taddr  Rewrite destination address of packet.

 -s#     Slow mode, use software loop, "#" as count.

 +#      Create a packet of "#" bytes length.

 -n#     Only rewrite address on "#" packets in loop.

 file    Name of data file containing packet.

 *#      Replicate file "n" times.
```

Packets are captured "ping" packets from BSD ping program. The different packet sizes are generated using the "packetsize" option to ping. Sizes used are 64, 128, 256, 1024 and 1518 bytes, with the size including the 4 byte crc.

Anvil

Packet counting program.

- Maintains cumulative counter.
- Maintains an average rate
  10 sec averageing period

Tests:

- **Idle State:**
  Count packets for 10 sec.

- **Delay:**
  `hammer -taddr -s100 packet/p64`
  Use scope to get time between end of input packet to start of output packet,
  3 packet sizes.

- **Raw rate:**
  `hammer -taddr -n1 +xx packet/p64`
  Adjust the length of the pad packet for the max throughput of the router.

- **Raw rate +25%:**
  `hammer -taddr -n1 +xx packet/p64`
  Adjust the length of the pad packet to offer a rate 25% faster than the rate determined above.

- **Max input rate:**
  `hammer -taddr packet/p64`
  Send packets to the router as fast as the test setup will allow.

Tests:

- **Back-to-back:**
  `hammer -taddr -s1000 packet/p64*n`
  Adjust "n" until router starts to drop packets.

- **Raw rate, filtering:**
  `hammer -taddr -n1 +xx packet/p64`
  Like "raw rate" but with router configured to do various types of filtering.

- **Raw rate, many routes:**
  `hammer -taddr -n1 +xx packet/p64`
  First send a set of packets containing RIP routing updates, then procede as with "raw rate".

Tests:

- Errors, crc:
  `hammer -taddr -s1000 -c packet/p64`
  Check router stats & check to see that
  all packets are dropped.

- Errors, runt:
  `hammer -taddr -s1000 +55`
  Check router stats & check to see that
  all packets are dropped.

- Errors, giant:
  `hammer -taddr -s1000 +1600`
  Check router stats & check to see that
  all packets are dropped.

- Counters:
  Record value on router counters.
  Restart anvil.
  Run one of above tests.
  Record value on router counters.

Types of problems found:

- Crashing routers.
  Heavy load
  Packet timing.
  ^C

- Dead cpu under load conditions.
  Can't disable bad port

- Eratic forwarding rates.
  Delay varies as much as 100ms
  Hurts round trip prediction software

- Shutting down interface improperly.
  Disable interface on non-fatal conditions
    runts on network
  "Keep alive" errors
    keep alive priority too low

Results:
Idle load.

- Much talk about routers loading networks with "keep alive" traffic.

- No tested router produced any significant load.

cisco	1 packet every 10 seconds
NSC	1 packet every 30 seconds
Proteon	1 packet every 3 seconds

- Non-tcp/ip protocols would add to the load.
- Routing packets add to the load.

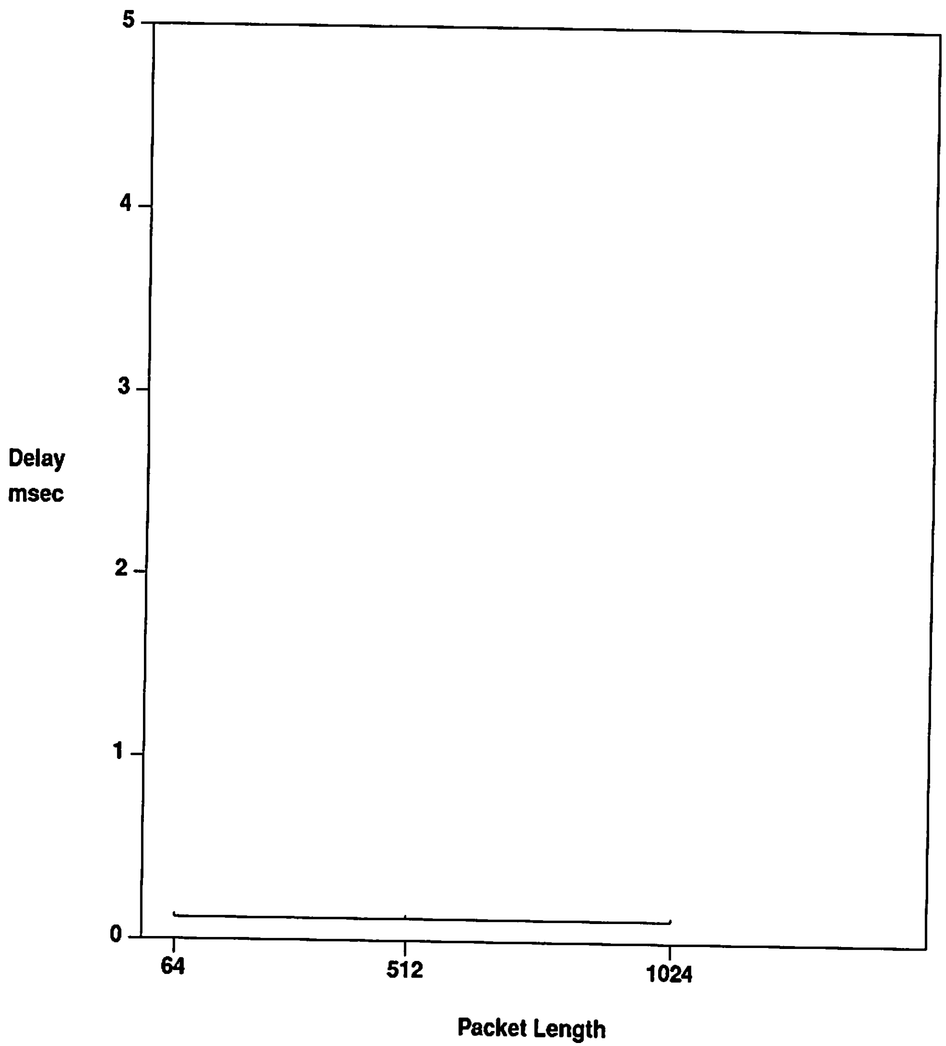
Results:
Delay

- TCP/IP uses round trip estimating to set the retransmission timer.
- Variability in the delay through a router would cause excess retransmissions or variations in timeout value.
- A large delay through a router would affect echo response time.

- The tested routers showed small transit delays that were mostly stable for a particular packet size.

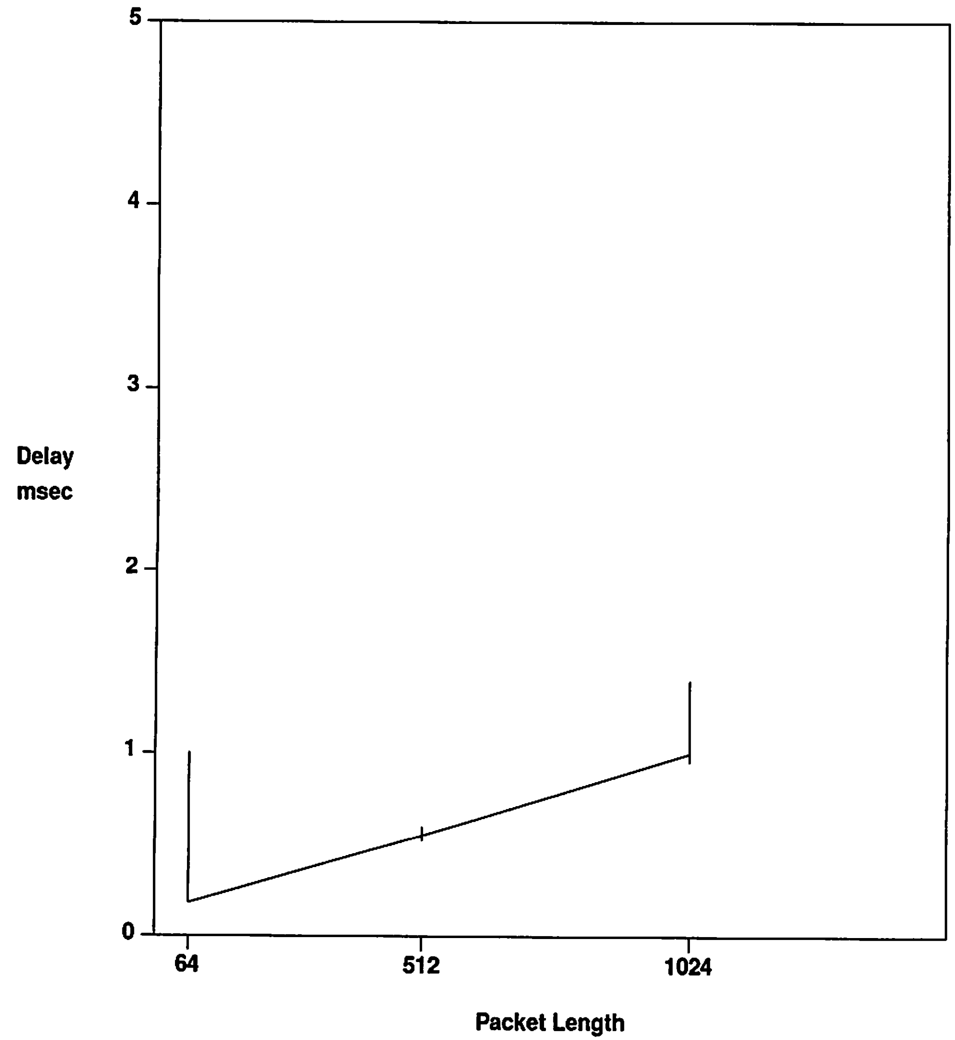
Delay:

cisco AGS within MCI card



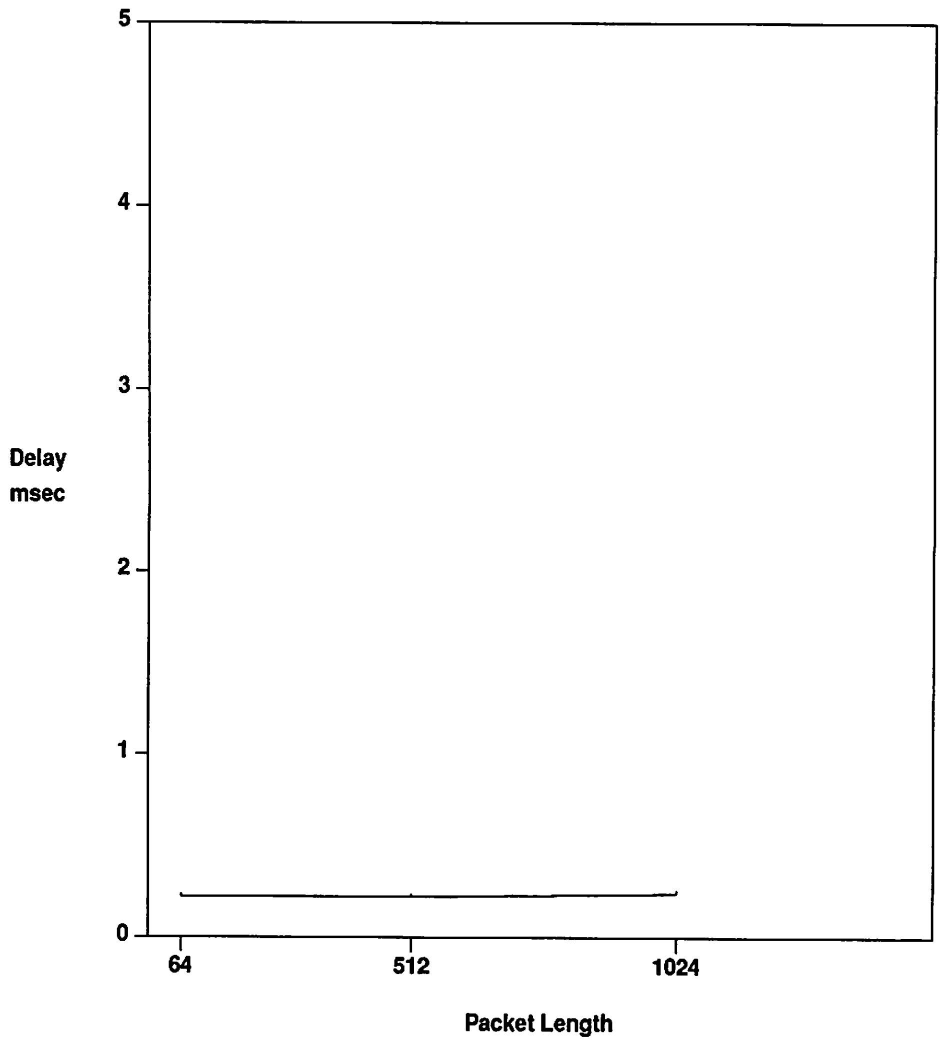
Delay:

cisco AGS between MCI cards



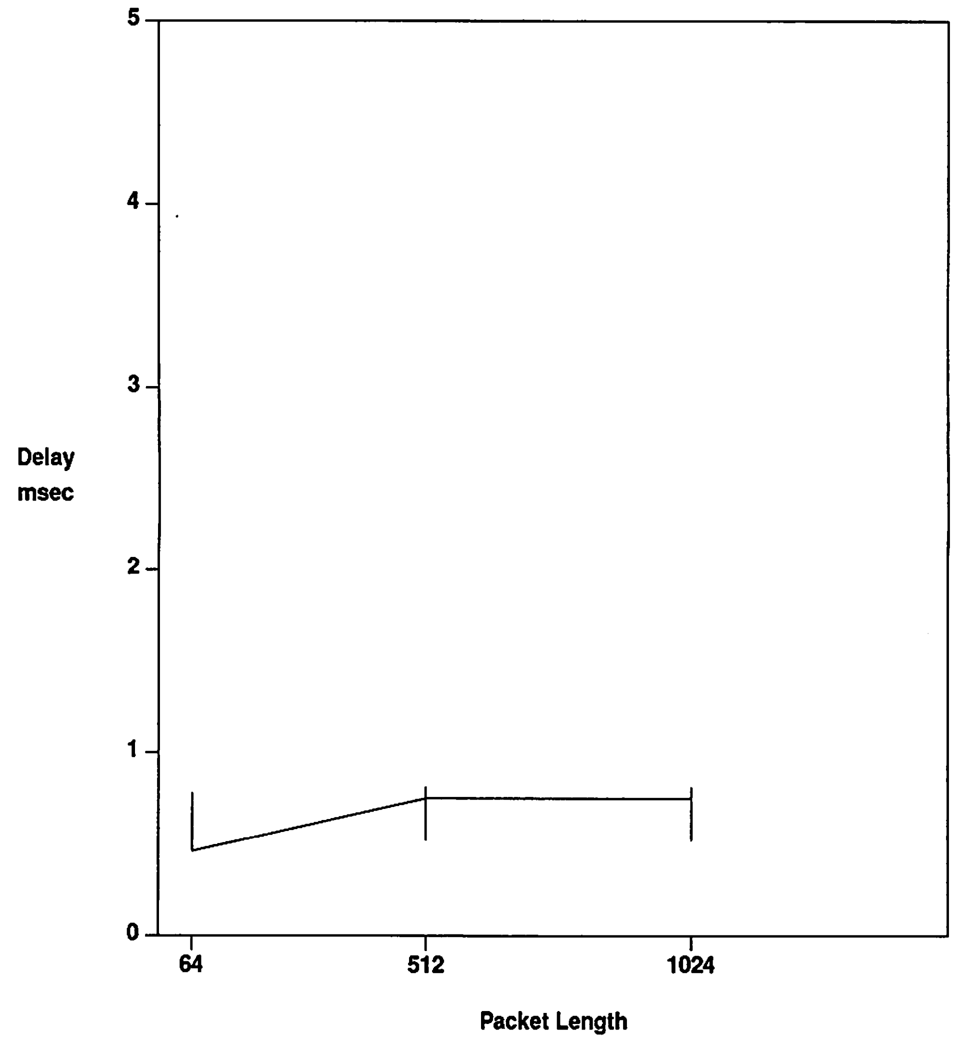
Delay:

NSC HYPERchannel-DX within NCET4 card



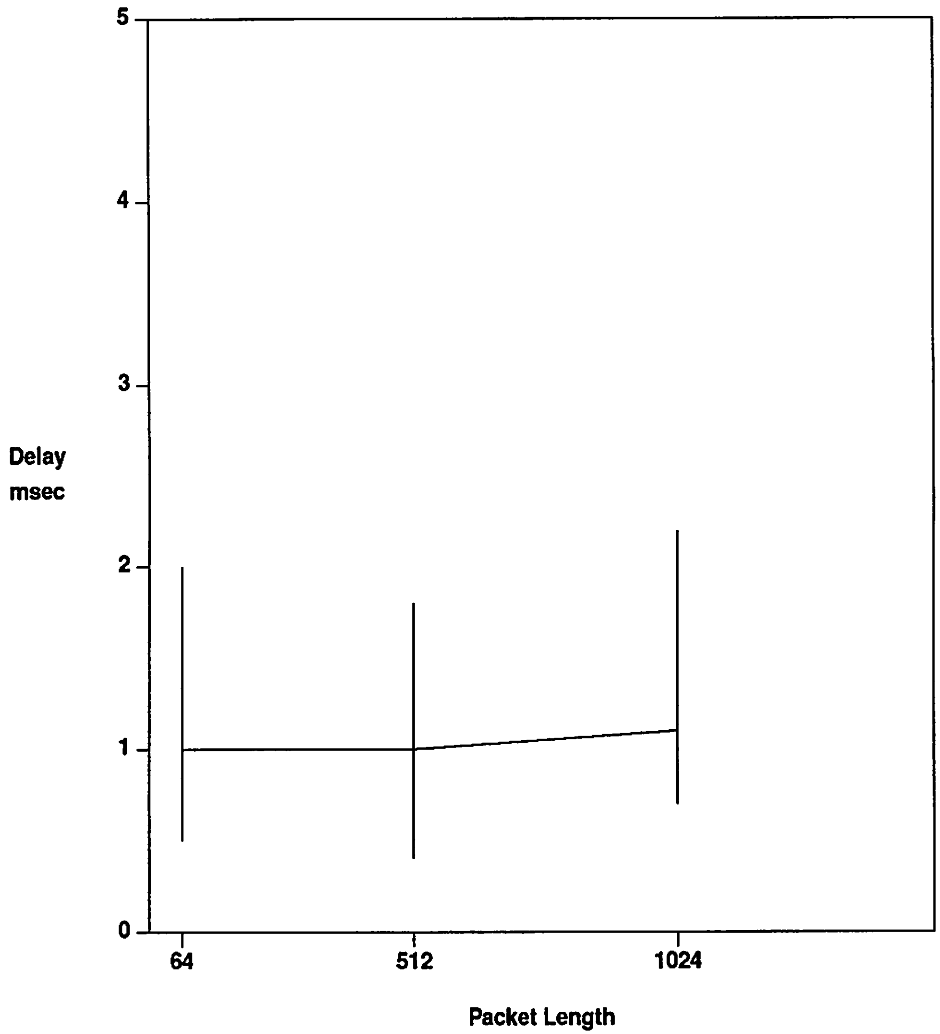
Delay:

NSC HYPERchannel-DX between NCET4 cards



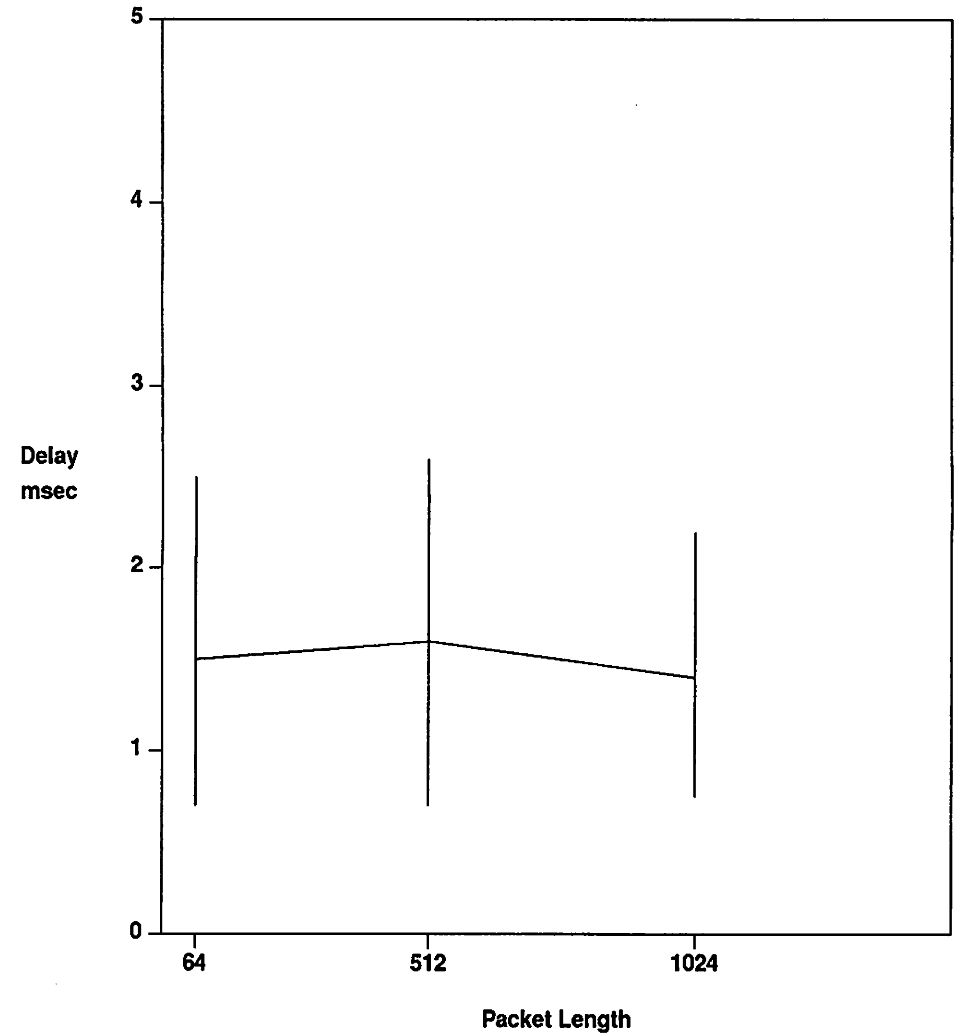
Delay:

Wellfleet - within interface board



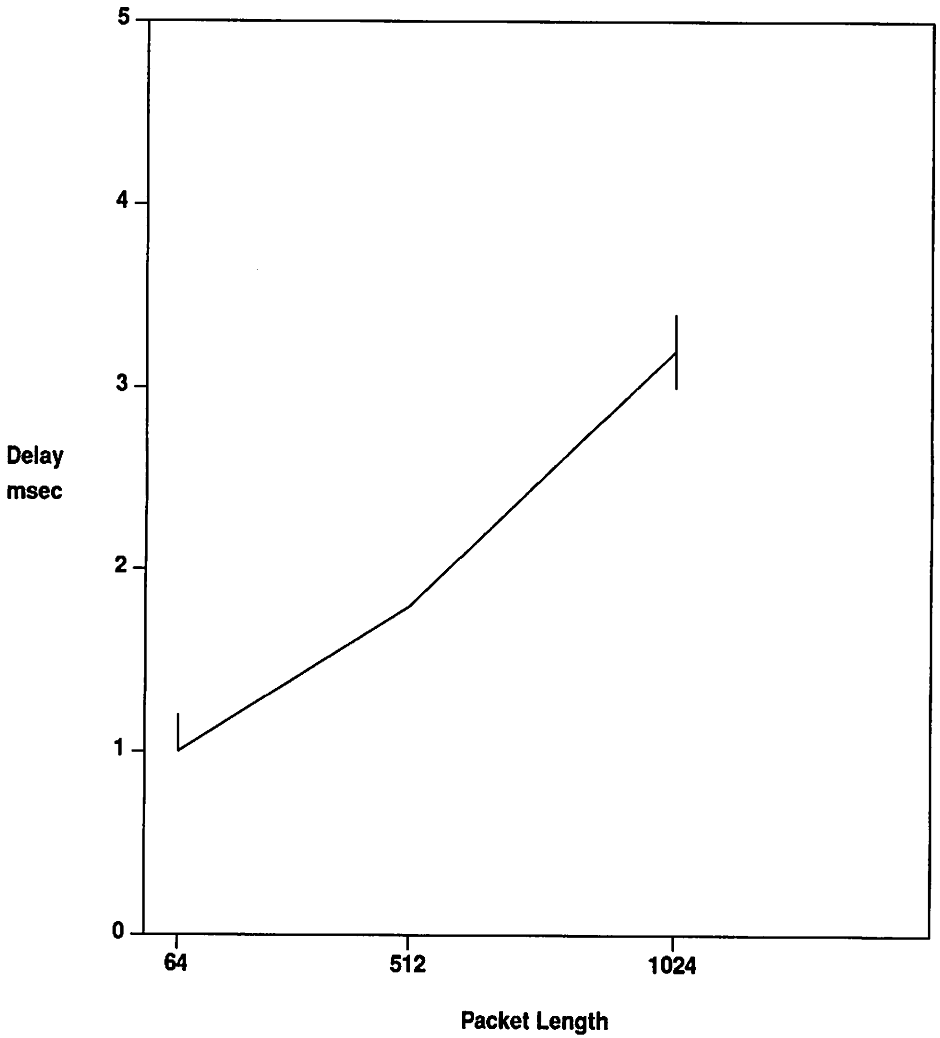
Delay:

Wellfleet between interface boards



Delay:

Proteon p4200



Results:

Max throughput

- Value measured was the maximum rate at which the router would forward packets without dropping.

- The packet source could not transmit packets with an interpacket gap of less than 55 usec where 9.6 usec is the "legal" minimum.
- Improved hardware is needed to adequately test some routers.
- Used calculated input rate and output counters to determine value.
- Output counter is dependant on the accuracy of the clock in the Tandy PC.
- Recorded output counts adjusted if greater than calculated input rate.
- Improved hardware required.
- Test setup could not determine if some small number of packets were dropped.
- Small numbers of dropped packets can have a large effect since system must wait for upper level protocol to timeout before retransmission.
- Improved test hardware could check for this.

Results:
Max throughput

- One router could not be tested because it disabled the interface when it saw runt packets on the ethernet even though the runt packets were not addressed to the router.
- The tested routers varied widely.
  Best were faster than the test equipment.
  Worst was still many times faster than observed 5 min average rates on Harvard netwoks.

Results:
Max throughput +25%

- An input load was generated that was 25% greater than the max rate determined above.

- To see the effect of small overloads.

- For most of the tested routers the throughput remained about the same as the maximum full throughput but more packets were dropped.
- One router could not be tested because it disabled the interface when it saw runt packets on the ethernet even though the runt packets were not addressed to the router.
- For one router the throughput was both greater at one packet length and less at other packet lengths.

Results:
Flood input

- The packet source was set to produce packets as fast as it could.
- Simulates conditions like arp storms.
- The packet source is not as fast as a real ethernet. 55 usec gap vs 9.6 usec.

- For most of the tested routers the observed forwarding rates were about the same as the maximum full throughput.

- One router stopped passing packets for packet sizes greater than 250 bytes.

Results:
Filtering

- For security, filtering can be used to exclude specific nodes.
  Example: Exclude all traffic to or from a student computer other than SMTP.
- Filtering can be used to include only permitted nodes for accounting or security.
  Example: if billing per node, filtering would be setup to only pass those who were registered and had paid their bill.
- Filtering on protocol type allows exclusion of "dangerou protocols like tftp at campus boundary.

- Filtering capabilities ranged from quite limited, ip sourc destination pairs, to very extensive.
- Filtering has a negative effect on the throughput of mos of the tested routers.

Results:
Filtering; NSC filtering options

- Very extensive filtering functions.
- Filters can be cascaded.
- TCP/IP & DECNET.

- Filter parameters

Any
All packets will match.
Hardware source addr ok
Checks physical ethernet address against IP address.
IP datagram length
Checks length of IP datagram.
IP destination address
Checks the destination address of the IP packet.
IP source address
Checks the source address of the IP packet.
IP protocol
Checks the "protocol" field in the IP header.
e.g. ICMP, GGP, TCP, EGP, UDP, ISO-TP4
IP type of service
Checks the "type of service" field of the IP packet.
e.g. normal, priority, immediate, flash, etc
TCP source port
Checks the source port of the TCP packet.
e.g. echo, ftp, telnet, smtp, finger. etc
TCP destination port
Checks the destination port of the TCP packet.
UDP source port
Checks the source port of the UDP packet.
e.g. echo, time, nameserver, bootp, tftp, snmp, etc
UDP destination port
Checks the destination port of the UDP packet.
Gateway address
Checks the address of the next gateway that the packet would go to next.
Gateway address to
Check based on whether router knows route between two IP addresses.

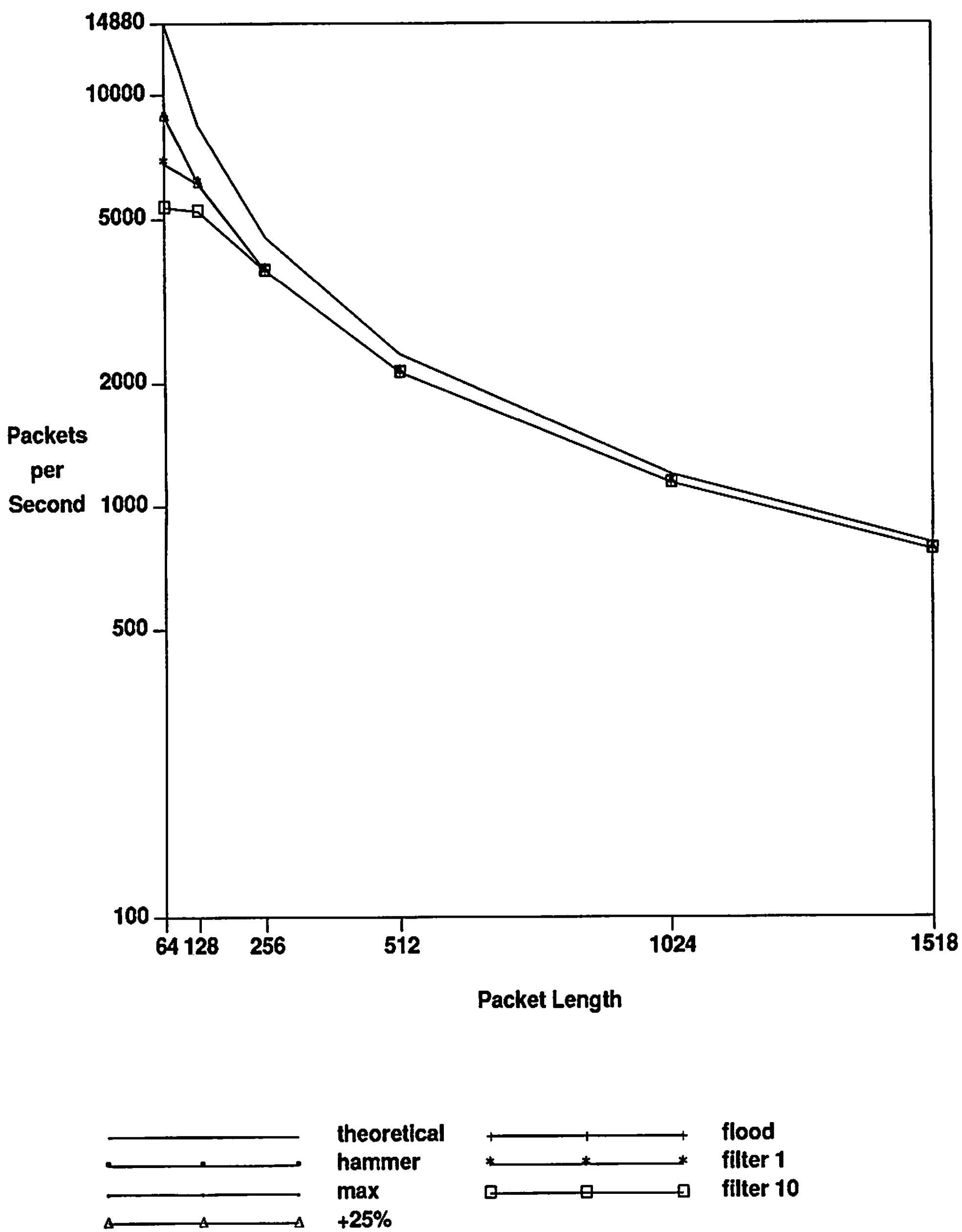
Results:
Filtering; NSC filtering options

- What does router do when pattern matched?

Accumulate statistics
Increments per source-destination pair counters.
Counter 1
Increment auxiliary counter #1 for pair.
Counter 2
Increment auxiliary counter #2 for pair.
Alarm
Generate console alarm message
ICMP unreachable
Send an ICMP unreachable message back to sender.
No ICMP unreachable
Cancel the sending of ICMP unreachable messages.
Route to
Re route the packet to an alternate host or gateway.
No route to
Cancel the route to function.
Copy to
Send a copy of the packet to a selected address.
No Action
Don't do anything.

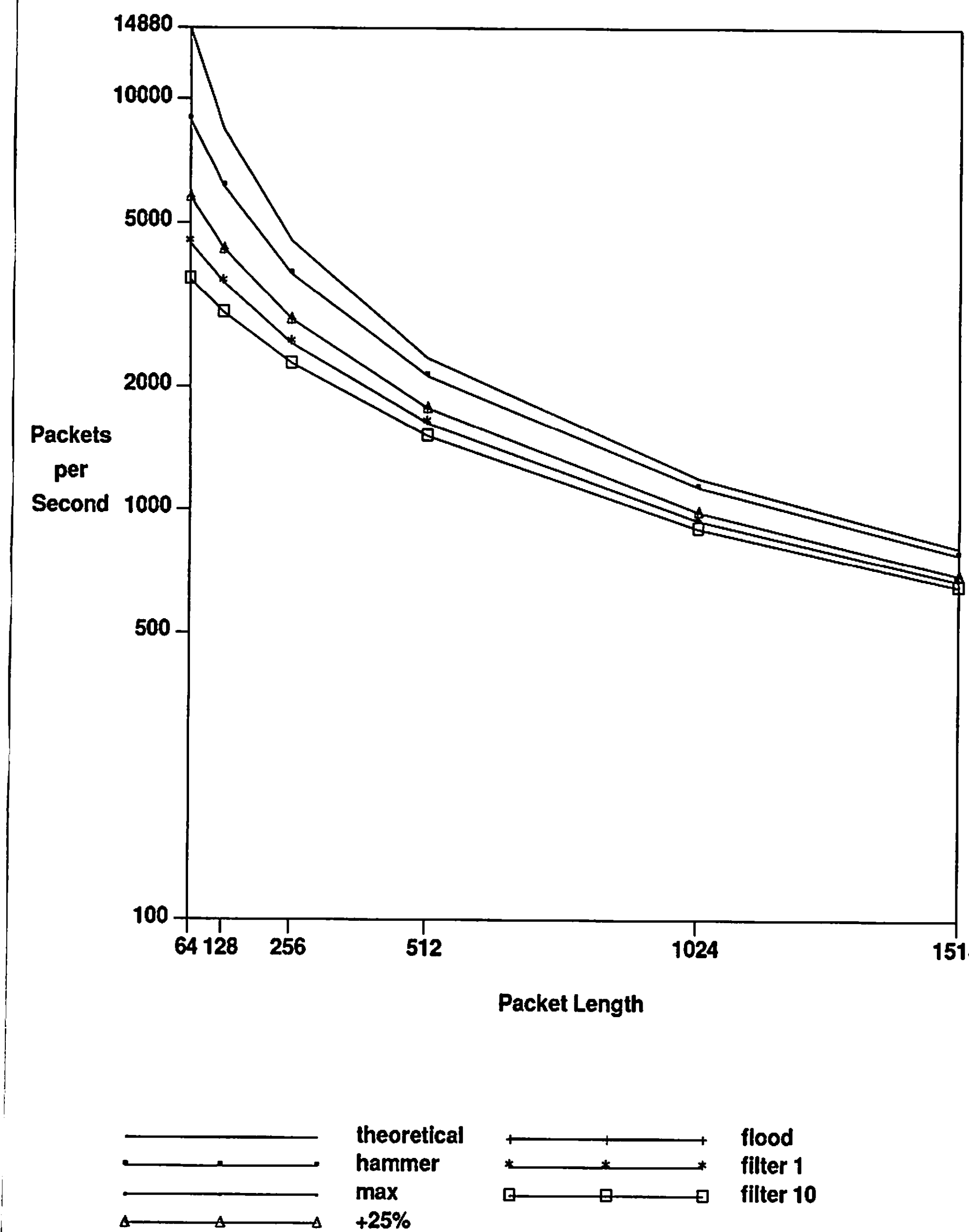
Performance:

cisco AGS within MCI card



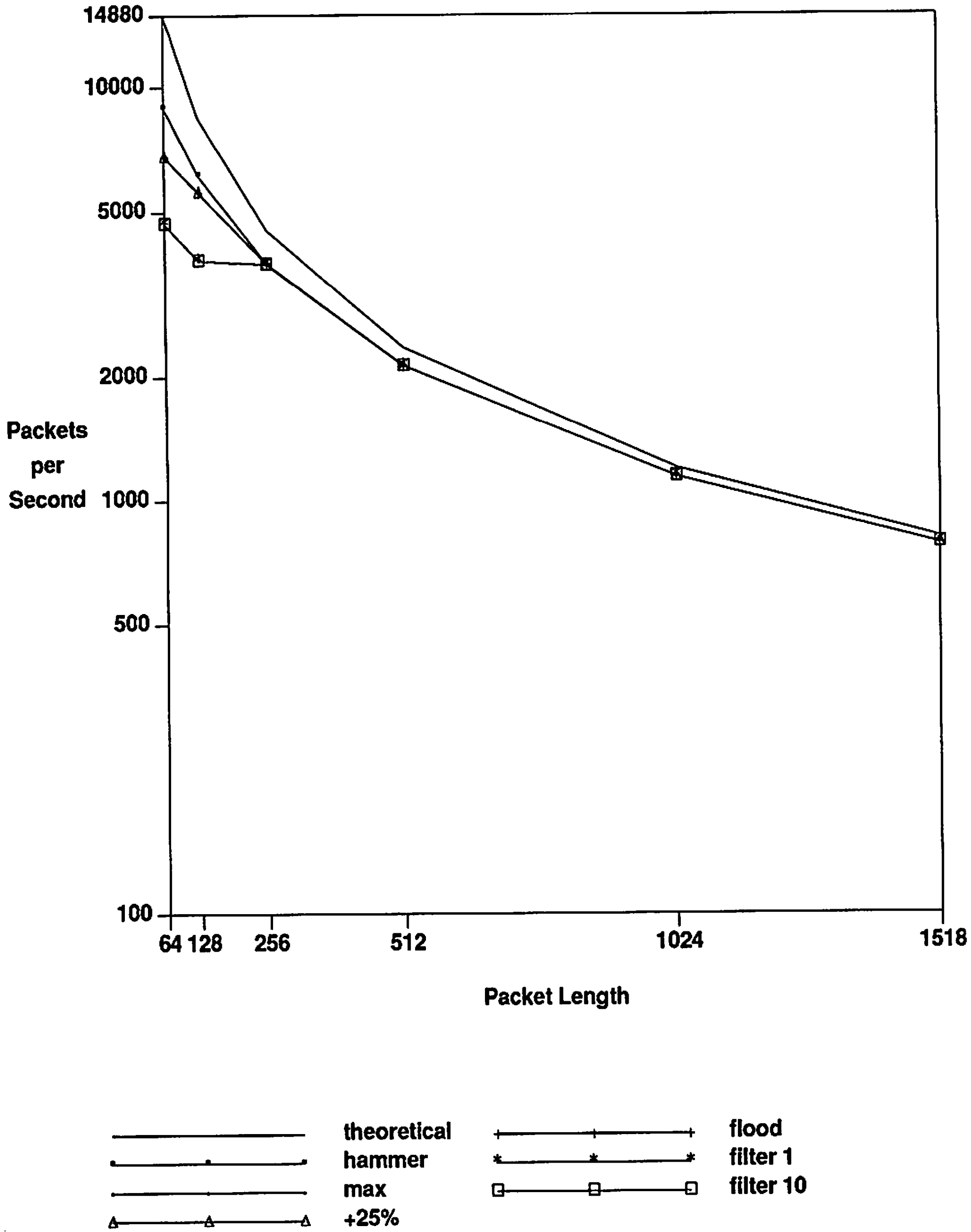
Performance:

cisco AGS between MCI cards



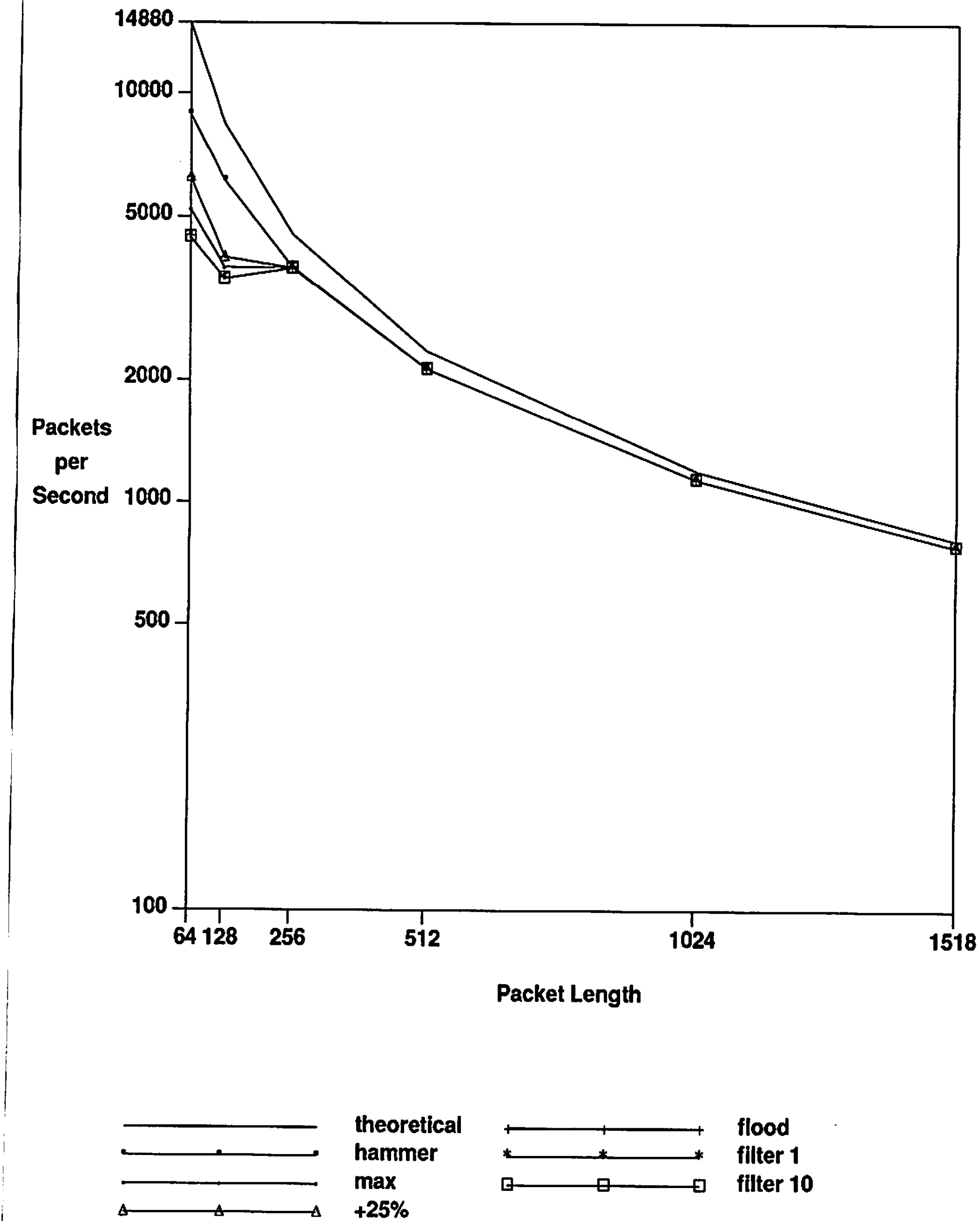
Performance:

NSC HYPERchannel-DX within NCET4 card



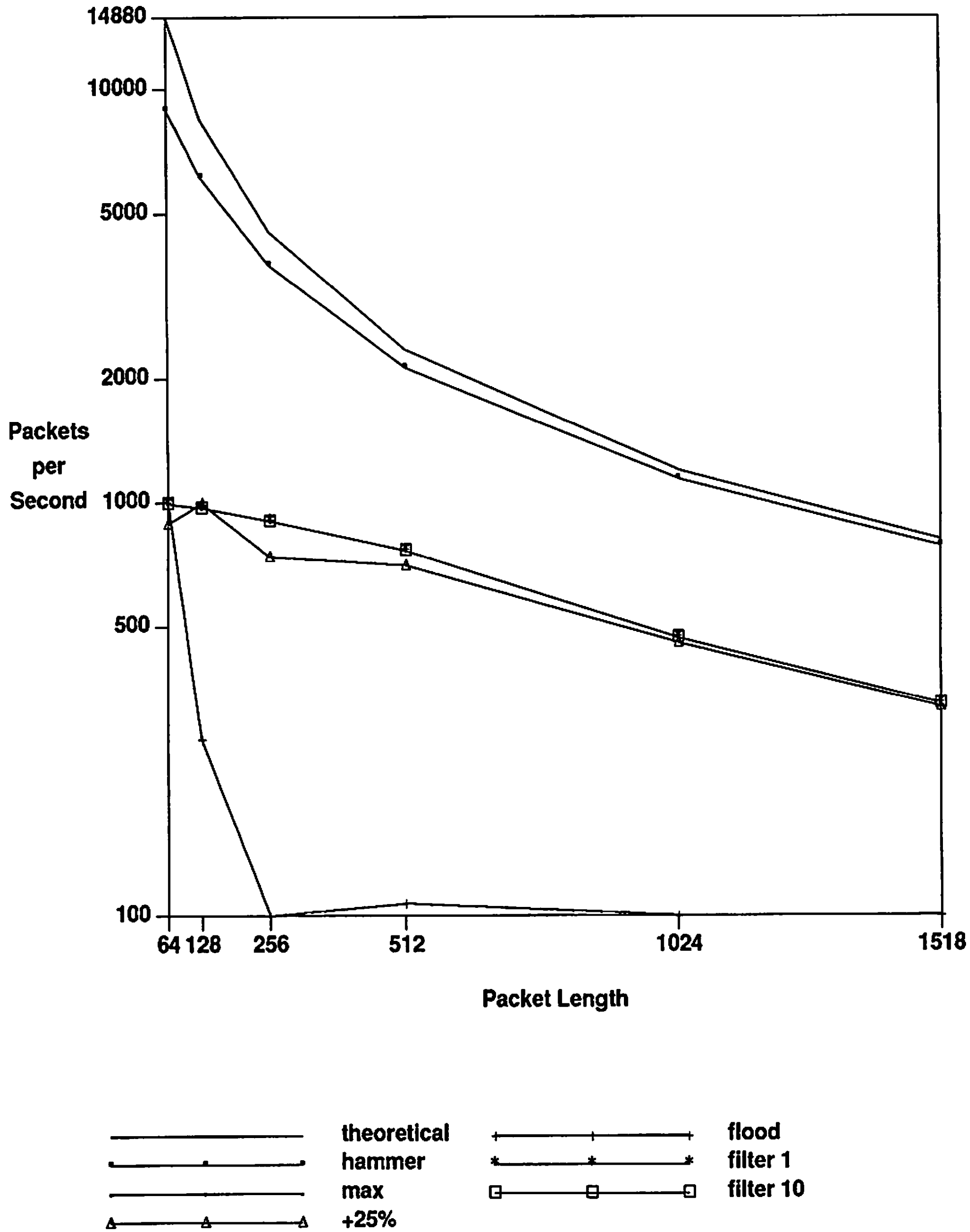
Performance:

NSC HYPERchannel-DX between NCET4 cards



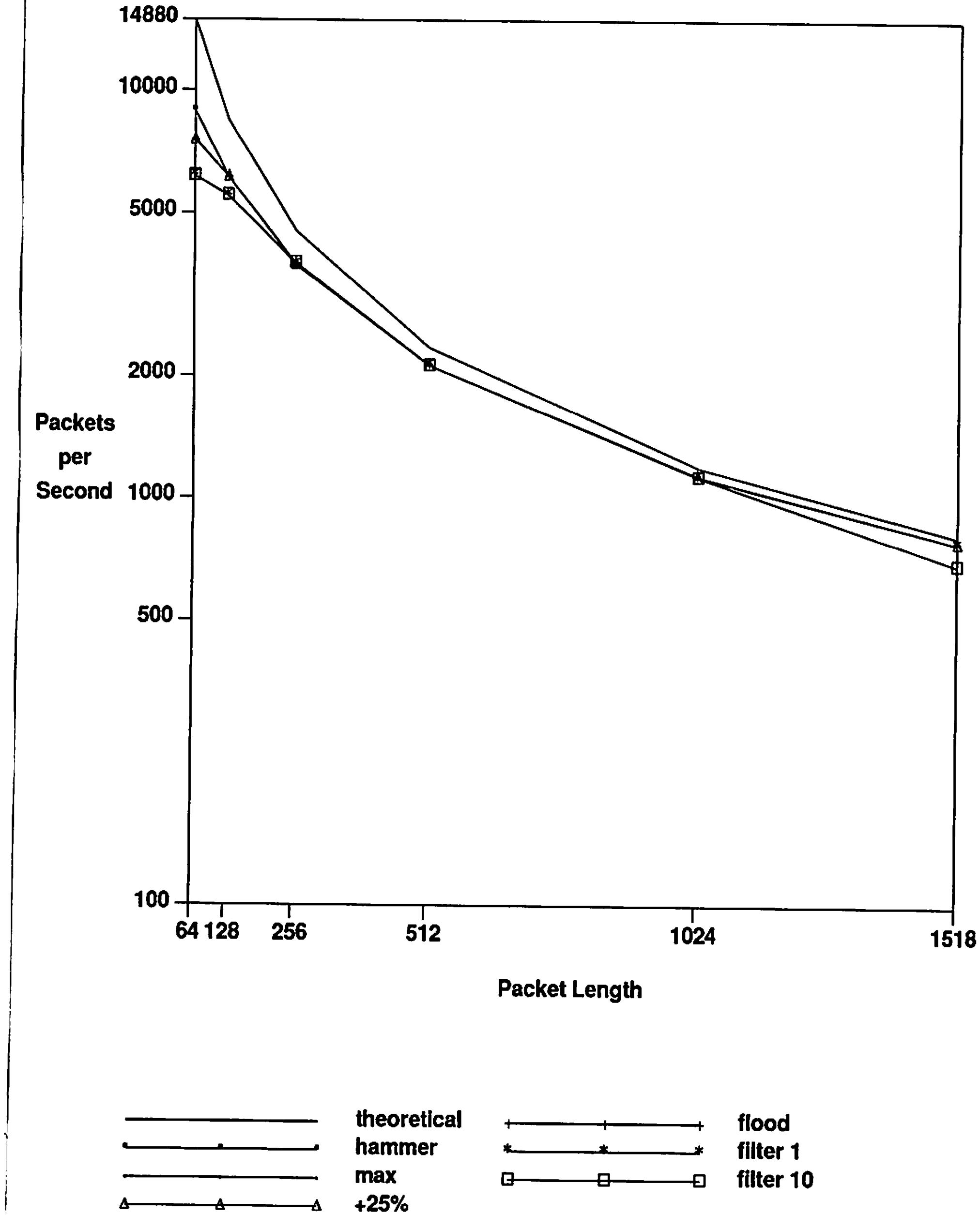
Performance:

Proteon p4200



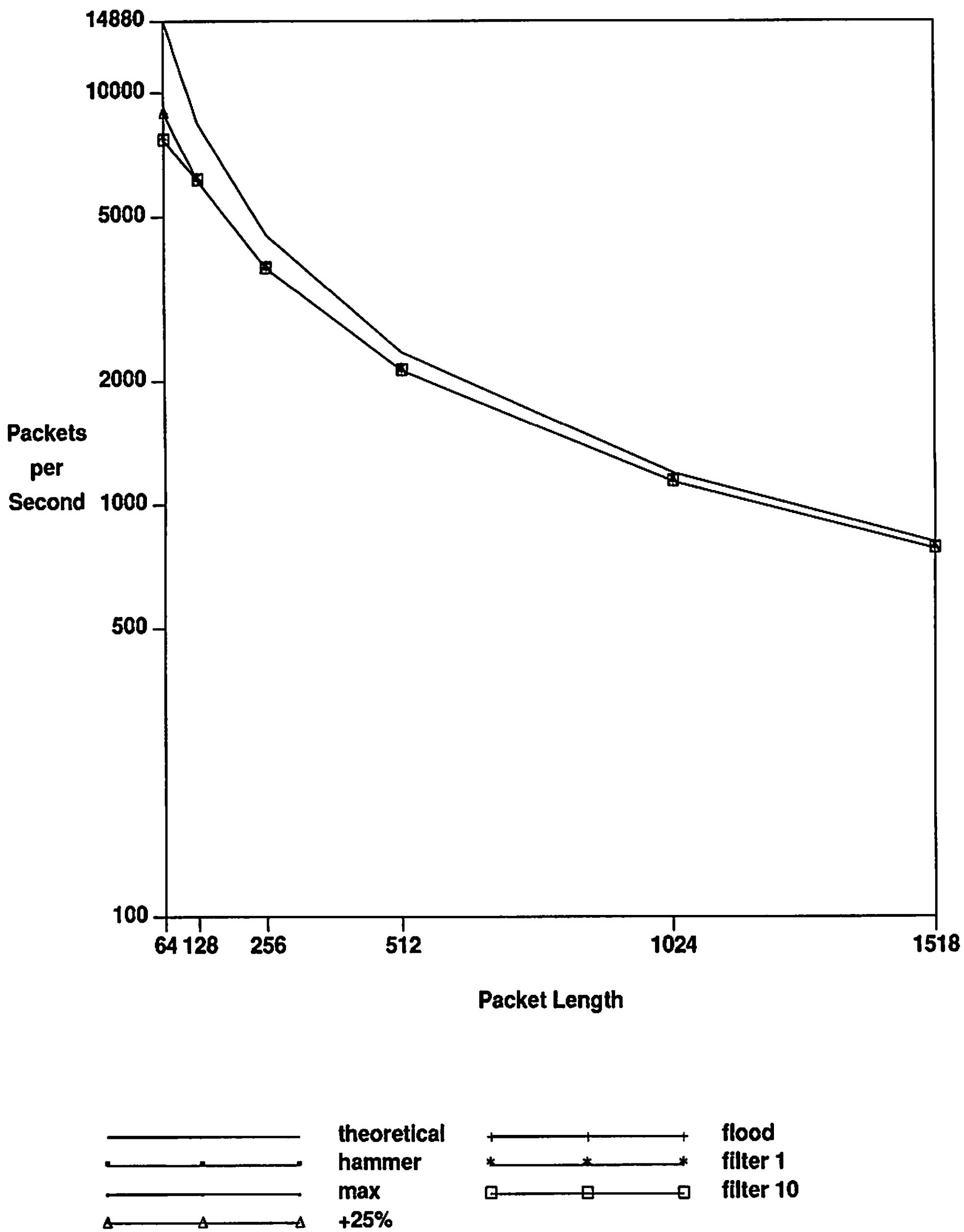
Performance:

Wellfleet Link Node - within interface card



Performance:

Wellfleet Link Node - between interface boards



Results:
back to back

- See how many "back to back" packets the router can take before overflowing internal buffers.

- NFS servers can produce back to back packets.
- If one packet in a fragmented datagram is lost the whole datagram must be resent.
- This procedure can take forever.
- cisco has delay option to "fix" the problem in NSF systems.

- The packet source cannot produce actual back to back packets.
- The packet source was not sufficient to test the faster routers.

- The tested routers all accepted enough back to back packets for normal applications.
- One router performed much better under this test than under continious load. The design seems to be tuned for episodic conditions.

Results:
back to back

- Theoretical:
  64 byte - 140 packets
  256 byte - 59 packets
  1024 byte - 17 packets

- cisco between MCI cards:
  64 byte - 90 packets
  256 byte - 45 packets
  1024 byte - 15 packets

- cisco within MCI card:
  Device is too fast for test setup.

- NSC between NCET4 cards:
  64 byte - 22 packets
  256 byte - 57 packets
  1024 byte - 17 packets

- NSC within NCET4 card:
  Device is too fast for test setup.

- Proteon
  64 byte - 20 packets
  256 byte - 14 packets
  1024 byte - 6 packets

- Welfleet
  Device is too fast for test setup.

Results:
counters

- The accuracy of the packet counters in the routers was tested.

- The information from these counters could be vital to network monitoring.
- Traffic information could also be useful in redesigning a network as the usage pattern changes.

- The counters on all of the tested routers were accurate within the limits of the test setup.

Results:
errors

- Packets were generated with specific types of errors; runts, giants, no crc.
- Counters were checked to see if the error packets were registered.
- Output was checked to see that the error packets were discarded.

- Error packets can indicate problems on the network. For example runts can show colisions.

- All of the tested routers discarded the error packets.

- There were mixed results on the counters. Only one router had all of the error statistics that one would want.

Results:
errors

- cisco
  - bad crc — CRC error counter incremented.
  - runt — Runt error counter incremented.
  - giant — Giant error counter incremented.
- NSC
  - bad crc — CRC error counter not incremented. Alignment error counter incremented.
  - runt — No counter.
  - giant — No counter.
- Proteon
  - bad crc — CRC error counter incremented.
  - runt — No counter.
  - giant — No counter.
- Wellfleet
  - bad crc — CRC error counter incremented.
  - runt — No counter.
  - giant — Packet imcomplete counter incremented.

Summary:

- all:
  - Did what they were asked to do.
  - Have all basic ip functions.
  - Faster than observed Harvard traffic.
- cisco:
  - Fastest within interface board.
    - 2 ethernet ports per interface.
    - Slice processor on interface.
  - Fast between interface boards.
  - Lots of protocols.
  - Single CPU design.
- Network Systems Corp:
  - Fast within interface board.
    - 4 ethernet ports per interface.
  - Fast between interface boards.
  - IP only.
  - Master CPU, router CPU and intelligent interfaces.
  - Channel interfaces.
  - Very good filtering options.
  - Nice user interface.
  - Good documentation.

Summary:

- Proteon:
  - Fastest network interface (P80).
  - Token ring interfaces faster than ethernet.
  - Lots of protocols.
  - Single CPU design.
- Wellfleet
  - Fastest between interfaces.
  - CPU per ethernet port.
  - REQUIRES vt100 terminal.
  - Menu interface.
  - Very good documentation.

Documentation:

- The documentation supplied with the routers was reviewed.
- It should be easy to locate information in the documentation.
- It should be easy to understand the commands.

Documentation:

- cisco
  - Gateway System Manual
    - Chapters on functional topics.
    - Uses bold to show interaction.
    - Prose form command descriptions.
    - Tabs on chapters.
    - Good explanations of terms like subnetting.
    - Command reference.
    - Index.

Documentation:

- Network Systems
  - HYPERchannel-DX Nucleus Customer Reference Manual
    - Overview of HYPERchannel-DX system.
  - HYPERchannel-DX 16-Slot Chassis Reference Manual
    - How to change fans etc.
  - HYPERchannel-DX NCET4 Network Interface Customer Reference Manual
    - Description of ethernet interface system.
    - 4 pages of statistic register names.
    - Ethernet packet description.
  - HYPERchannel-DX NDIP1 Router Co-Processor Customer Reference Manual
    - User interface and command description.
    - Uses many fonts to show interaction.
    - Exhaustive description and explanation of commands and options.
    - Command summary.
    - Full MIB list and description.

  Each volume has its own index.

Documentation:

- Proteon
  - p4100/p4200 Router Software User's Guide
    - Uses changes in fonts to show interaction.
    - Examples of commands.
    - Page or pages for each command within each mode.
    - Chapters per interaction mode type.
    - A bit terse.
    - List of messages and their meanings.
    - Index.

- Vitalink
  - TransPATH Reference Manual
    - Configuration for all options.
    - Good graphics, responses set off visually.
    - Examples of menu screens.
    - Concise but clear text.
    - Listing of system messages.
    - Glossary.
    - Index.

Documentation:

- Wellfleet

  Series of volumes.

  Well written, clean and clear.

  Overview Guide
    - Overview of network designs.
    - Overview of Wellfleet router products.
    - Includes all environmental info on devices.
    - Good definition of performance terms.
    - Includes performance data on devices.
    - No index.

  Installation Guide
    - Rules and regulations for Telco connections.
    - Even includes Telco form for T1 installation
    - Mechinical installation of devices.
    - Configuring pc board jumpers.
    - No index.

  Configuration Guide
    - Configuration for all options.
    - Examples of control screens.
    - Full definiations of many techinical terms.
    - Much general information, like ethernet type fields.
    - Configuration site survey form.
    - No index.

User interface:

cisco

- command line
- 2 level access, with passwords
  - look only
  - "enable"
- full word commands, can abbreviate
- most changes done in "config" mode
  - non-interactive in config mode
  - errors reported at end
- commands in config mode take effect at end of config mode
- no useful way to set ALL of current config
  - must be done function at a time
  - reload of config "or" with current
- has "?" for option expansion
  - ?<CR>, does wrong thing
  - no "?" in config mode
- simple cold boot
  - asks for ip addresses & mask
- config can be downloaded with tftp
- uses RARP or BOOTP for startup
- with config memory
  - no requirement for config download
  - specific boot hosts can be selected
- can upload config to file with tftp
- one box can be boot host to another

User interface:

Network Systems

15 level access with seperate passwords if wanted
    "authority" levels
each command has an authority level
claimed to appear "similar" to UNIX™
    actually closer to MS-DOS
has 32K "filesystem" for startup and config files
use "ed" to create and modify these files
dir, run, type, rename, copy, erase, format filesystem
"display tasks" same as "ps"

simple "startup" file

```
SETNAME Harvard
DEFINE ADDRMASK 128.103.0.0 0xffffff00
ROUTE ADD 0.0.0.0 128.103.8.1
START IF EN61 128.103.1.229
START IF EN41 128.103.8.1
```

extensive help system
    does not allow abreviations for topics
very straightforward user interface
well thought through
uses xx.conf files for config of daemons
    gated.conf, snmpd.conf
full word commands, can abreviate
commands have immediate effect

User interface:

Proteon

command line
single level access control, password
DDT like
"talk" to 6 processes
    select by number not name
within process, select modes
    select by number not name
full word commands
good use of "?" to get possible inputs
    can be used at all stages of commands
config can be downloaded with tftp
can have local floppy
with config memory
    no requirement for config download
    specific boot hosts can be selected
must load initial config with tftp or from floppy
simple cold boot
    asks for boot device info
some commands require a reboot of router to
    take effect

User interface:

Vitalink

command line and menu screen
can mark specific commands as privledged
in command line functions, abreviated command words
    echo full command and options
local floppy for boot

Help lines:

cisco

24 hour, 1-800-553-2447

Network Systems

24 hour, region dependant number
call is to local SA, can call in national

Proteon

8am-8pm e?t, 1-508-898-3100

Vitalink

24 hour, 1-800-523-9550

Wellfleet

24 hour, 1-800-222-7611

Vendor addresses:

Advanced Computer Communications
720 Santa Barbara St.
Santa Barbara CA 93101
(800) 444-7854
fax (805) 962-8499

cisco Systems, Inc.
1350 Willow Road
Menlo Park, CA 94025
(800) 553-NETS
fax (415) 326-1989

Network Systems Corp.
7600 Moon Av. North
Minneapolis, MN 55428
(800) 328-9108
fax (612) 424-2853

Proteon Inc.
Two Technology Dr.
Westborough, MA 01581
(508) 898-2800
fax (508) 366-7930

Vitalink Communications Corp.
6607 Kaiser Dr.
Fremont, CA 94555
(415) 794-1100
fax - (415) 795-1085

Wellfleet Communications, Inc.
12 DeAngelo Dr.
Bedford, MA 01730
(617) 275-2400
fax - (617) 275-5001