Obtaining Good Performance from Computer Networks

Clive Davenhall
(clive@nesc.ac.uk)

13 May 2005

PPARC e-Science Summer School 2005

Transmission of Free Messages on Astronomical Subjects over the Transatlantic Cables

A very important concession has been made to the Smithsonian Institution by the Directors of the Associated Transatlantic Cable Companies, who have agreed to transmit gratuitously between Europe and the United States, a limited number of short messages on astronomical subjects. Under this arrangement two telegrams have already been received from the United States by the Astronomer Royal, who on his part has undertaken, at the request of Dr. Henry, Secretary of the Smithsonian Institution, to forward from Europe any message announcing an important astronomical discovery. The Directors of the Associated Companies have consented that ten messages, of ten words each, may be sent free over the cables annually. This liberal concession on the part of the Directors cannot be too highly appreciated by astronomers generally, and especially by the Fellows of this Society.

In conformity with this agreement the Astronomer Royal will be prepared to forward any important astronomical message, limited to *ten* words, which may be sent to him for this purpose from the principal European astronomers.

Monthly Notices of the Royal Astronomical Society,
April 1873.

Introduction to Computer Networks

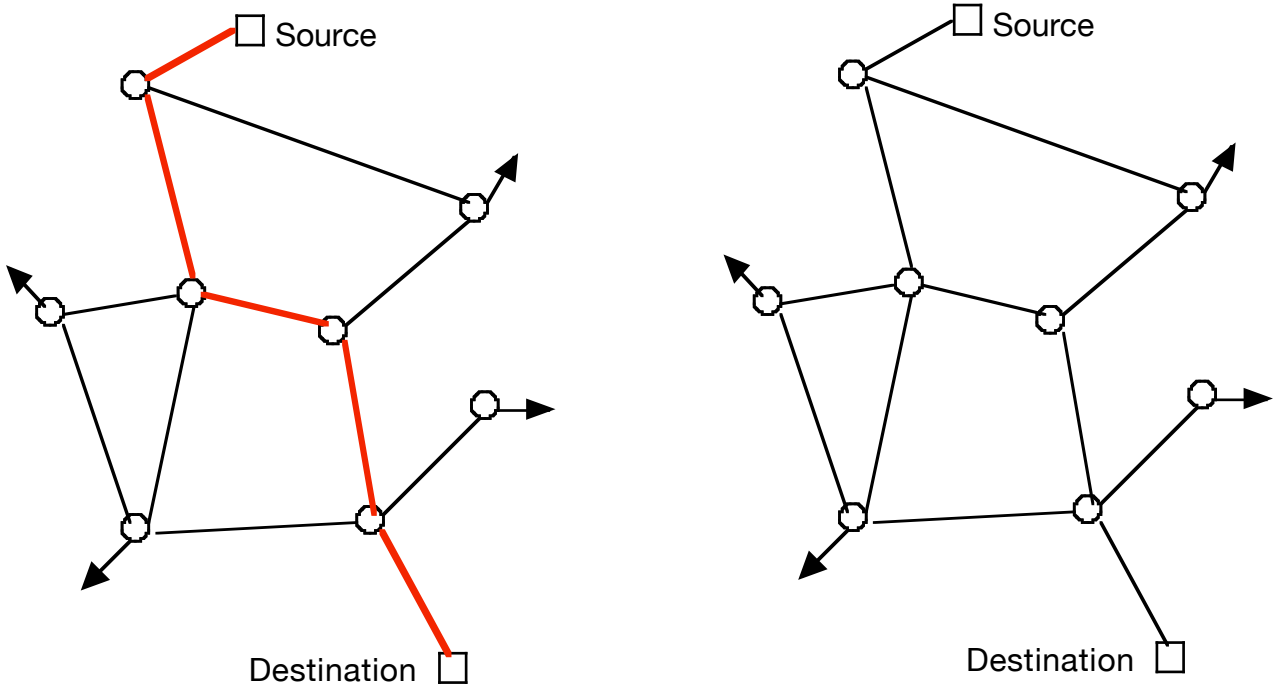
- What do we mean by ‘computer network:’

  ‘an *interconnected* collection of *autonomous* computers.’

- Need **protocols** for two machines to communicate; an agreement on how to interpret signals.

- Information divided into **packets** to facilitate checking for errors and resending lost information.

Circuit-Switching and Packet-Switching

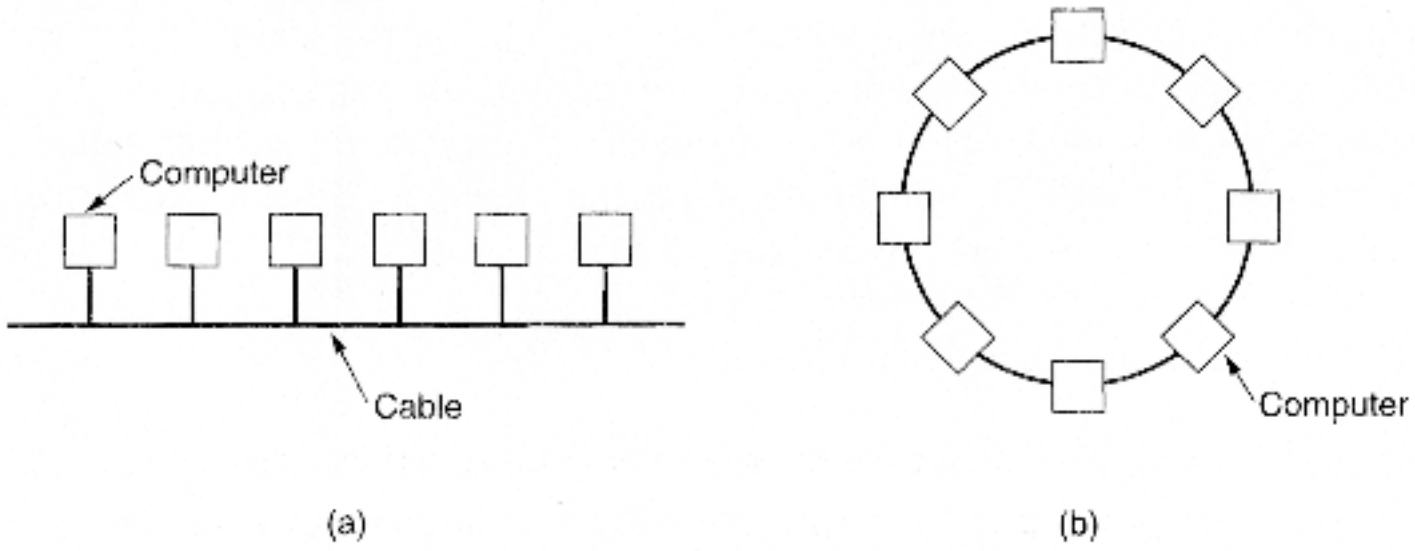


Types of Network

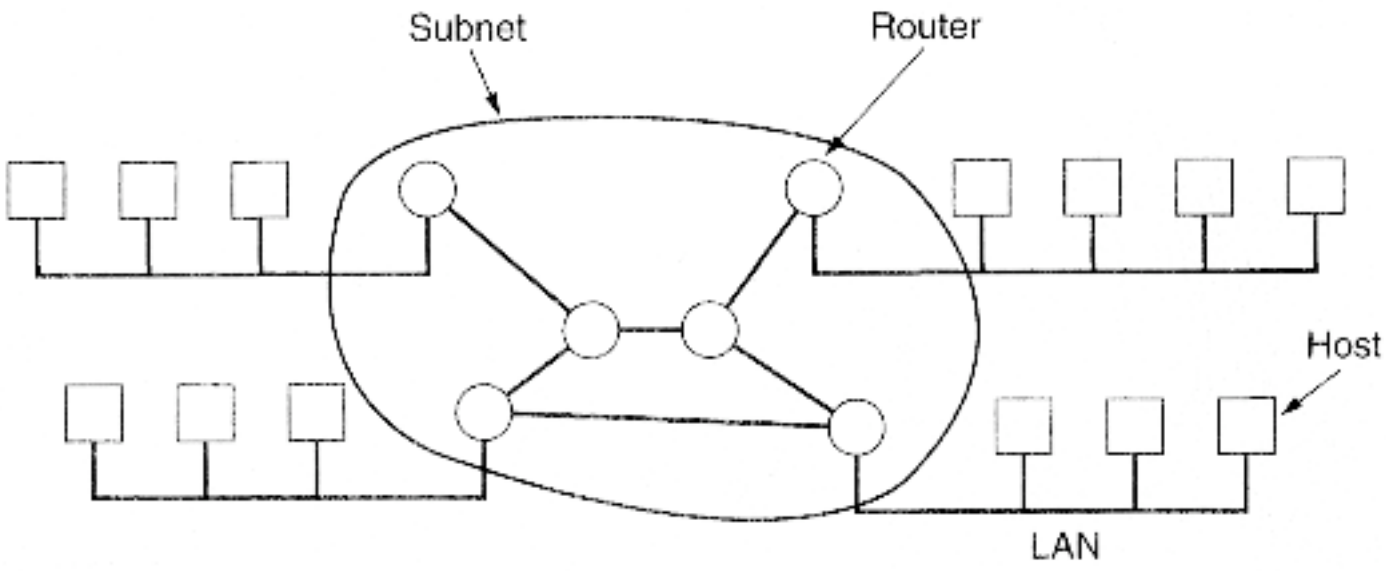
- Two properties important: transmission technology and scale.

- Classification by transmission technology:
  - broadcast,
  - point-to-point.

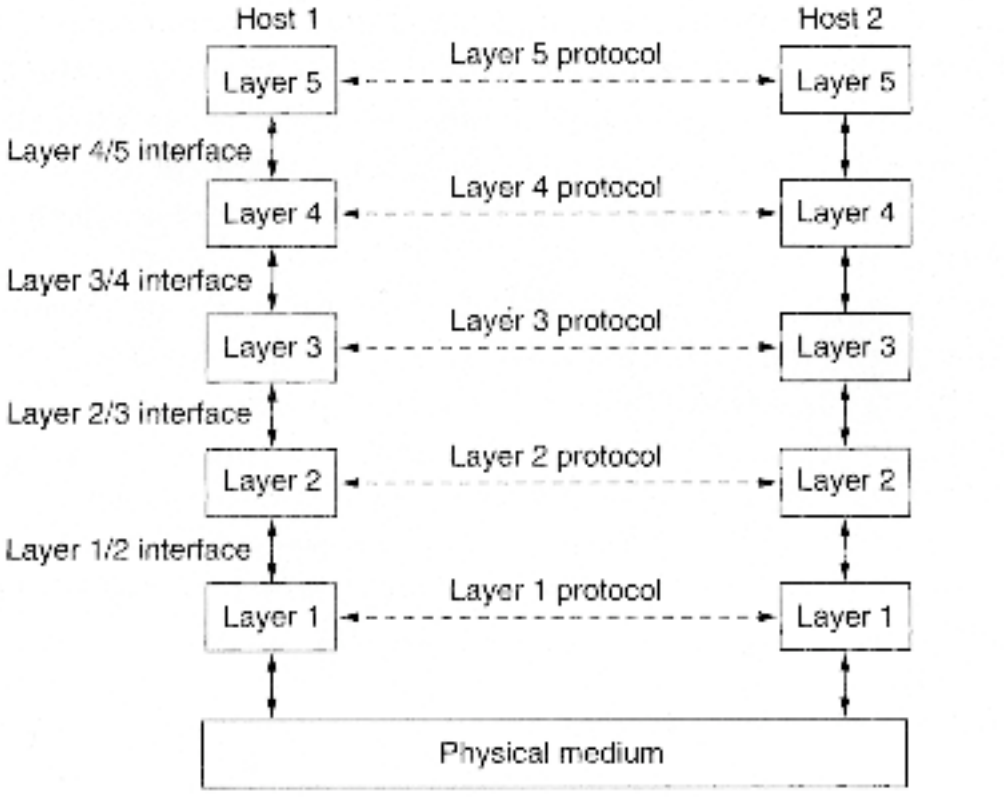
- Classification by scale:
  - LAN (Local Area Network),
  - MAN (Metropolitan Area Network),
  - WAN (Wide Area Network)



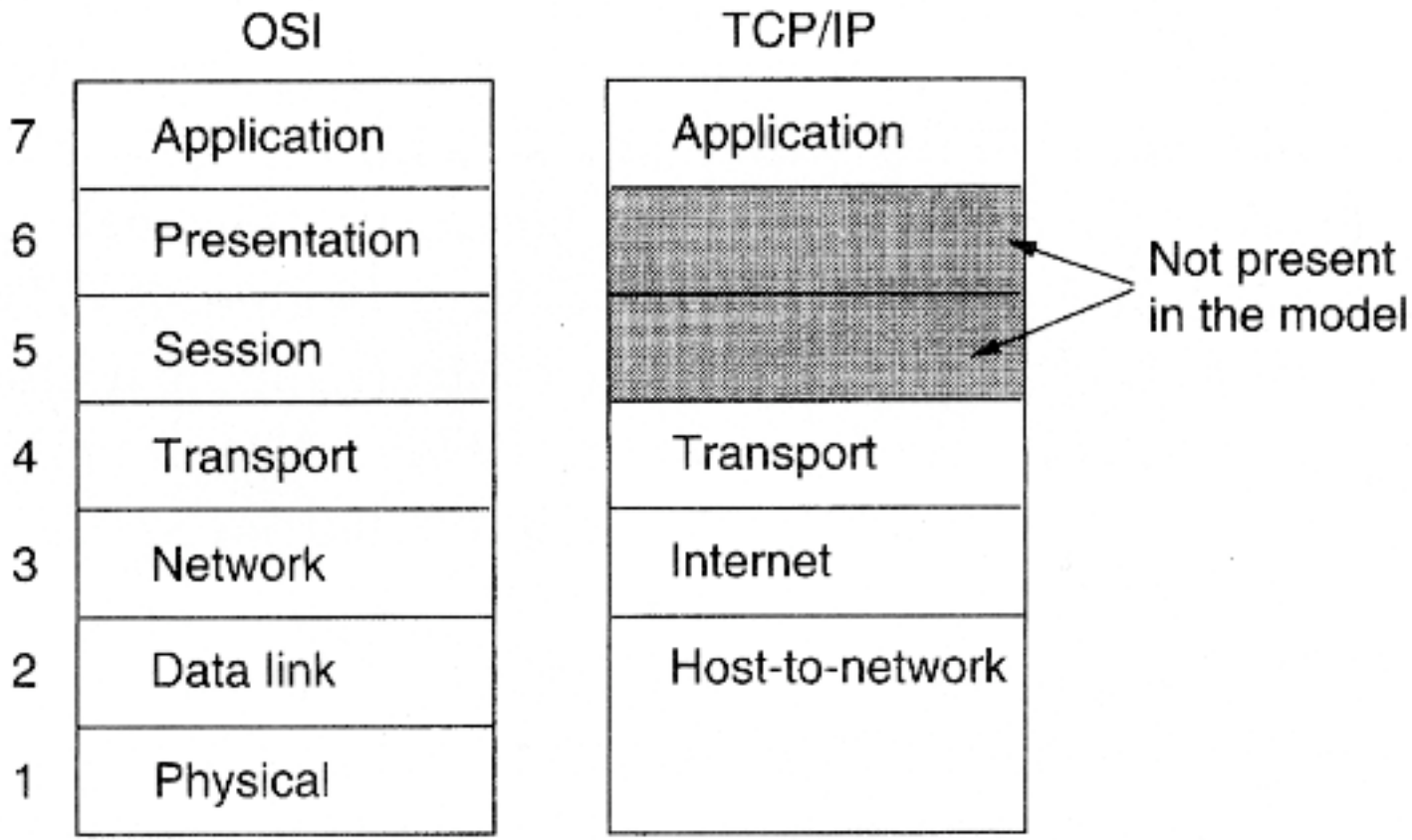
LAN topologies



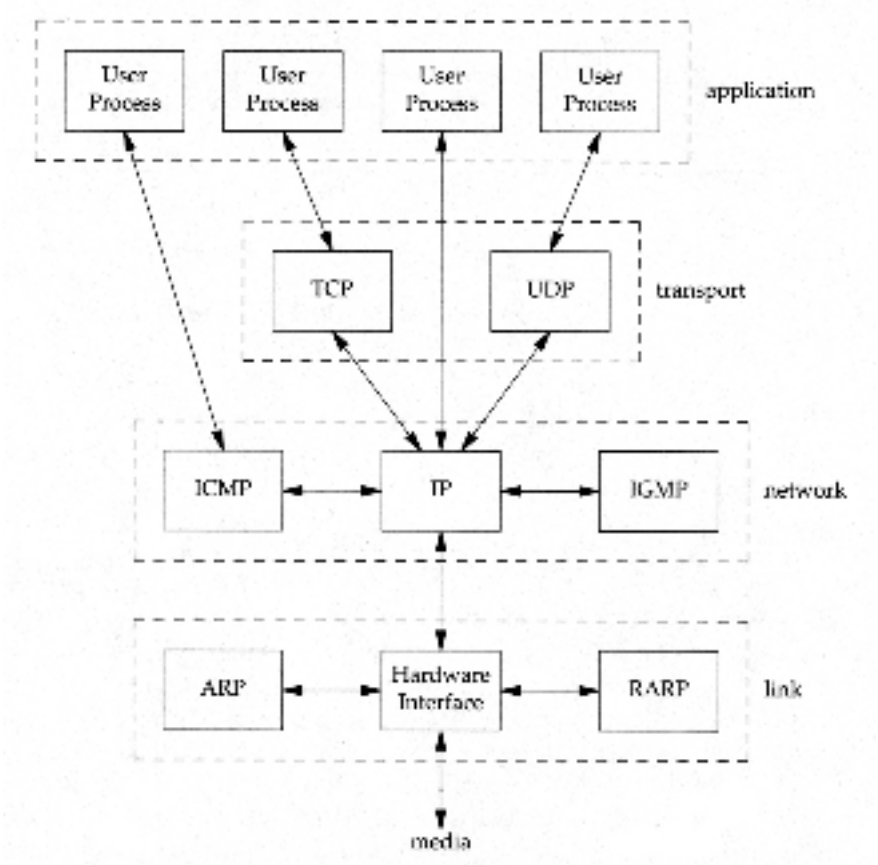
WAN topologies



Network protocols



OSI 7 layer model and TCP/IP



TCP/IP (Internet Protocol Suite)

Network Monitoring

- Types of measurement: passive or active?
- Measure at one point or many?
- Measure network or application performance?

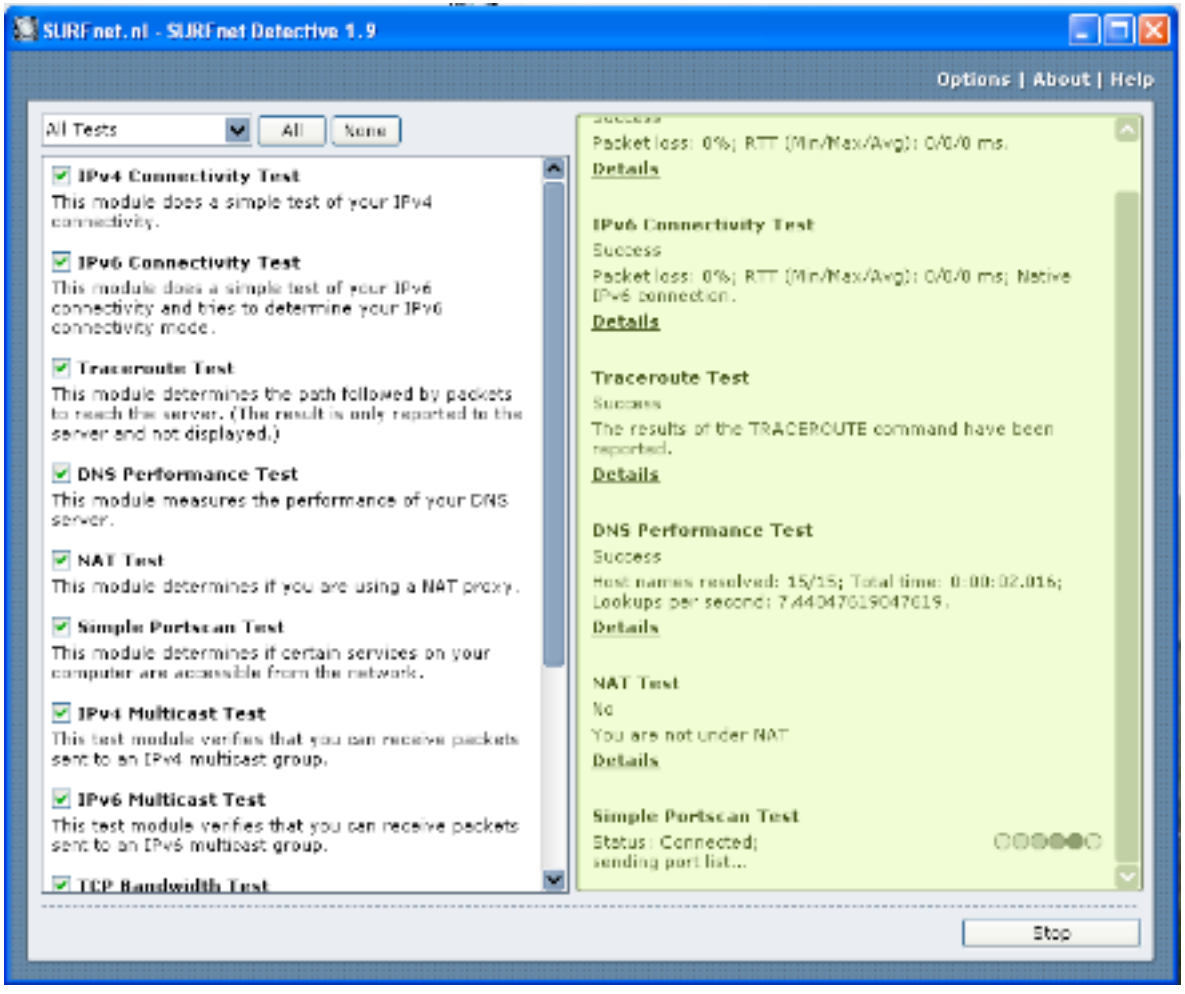
Common Metrics

- Latency.
- Packet loss.
- Throughput.
- Availability and reliability.

Simple Tools

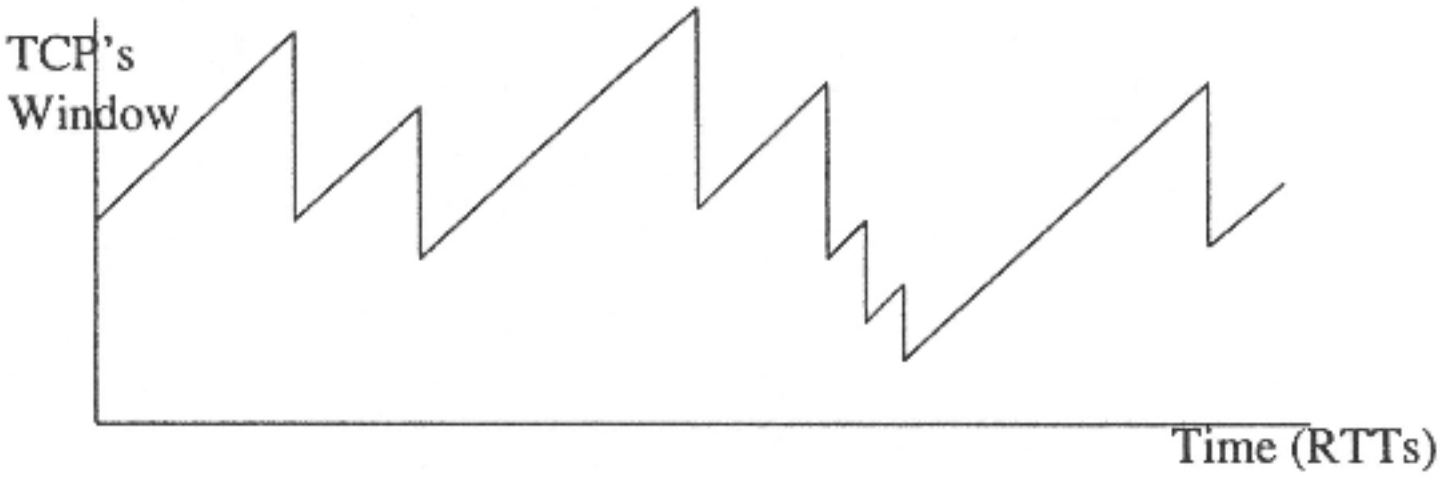
- ping
- traceroute
- pathchar
  http://www.caida.org/tools/utilities/others/pathchar/
- pchar
  http://www.kitchenlab.org/www/bmah/Software/pchar/
- iperf
  http://dast.nlanr.net/Projects/iperf/

```
pchar to aife.jb.man.ac.uk (130.88.24.56) using UDP/IPv4
Using raw socket input
Packet size increments from 32 to 1500 by 32
46 test(s) per repetition
32 repetition(s) per hop
Warning: target host did not respond to initial test.
0: 192.108.120.198 (grendel12.roe.ac.uk)
   Partial loss:      292 / 1472 (19%)
   Partial char:      rtt = 0.077017 ms, (b = 0.000165 ms/B), r2 = 0.999931
                      stddev rtt = 0.000254, stddev b = 0.000000
   Partial queueing:  avg = 0.000011 ms (67 bytes)
   Hop char:          rtt = 0.077017 ms, bw = 48562.749945 Kbps
   Hop queueing:      avg = 0.000011 ms (67 bytes)
1: 192.108.120.254 (teine.roe.ac.uk)
   Partial loss:      0 / 1472 (0%)

     ...

12: 194.66.25.38 (gw-jodrell.netnw.net.uk)
   Partial loss:      293 / 1472 (19%)
   Partial char:      rtt = 7.231458 ms, (b = 0.000766 ms/B), r2 = 0.999386
                      stddev rtt = 0.003526, stddev b = 0.000003
   Partial queueing:  avg = 0.000999 ms (66398 bytes)
   Hop char:          rtt = 0.175590 ms, bw = 30873.297757 Kbps
   Hop queueing:      avg = 0.000306 ms (1181 bytes)
13: 130.88.24.56 (aife.jb.man.ac.uk)
   Path length:       13 hops
   Path char:         rtt = 7.231458 ms r2 = 0.999386
   Path bottleneck:   30873.297757 Kbps
   Path pipe:         27907 bytes
   Path queueing:     average = 0.000999 ms (66398 bytes)
   Start time:        Wed Sep 22 16:00:09 2004
   End time:          Wed Sep 22 17:46:52 2004
```


SURFnet.nl - SURFnet Detective 1.9
Options | About | Help
All Tests
All
None
IPv4 Connectivity Test
This module does a simple test of your IPv4 connectivity.
IPv6 Connectivity Test
This module does a simple test of your IPv6 connectivity and tries to determine your IPv6 connectivity mode.
Traceroute Test
This module determines the path followed by packets to reach the server. (The result is only reported to the server and not displayed.)
DNS Performance Test
This module measures the performance of your DNS server.
NAT Test
This module determines if you are using a NAT proxy.
Simple Portscan Test
This module determines if certain services on your computer are accessible from the network.
IPv4 Multicast Test
This test module verifies that you can receive packets sent to an IPv4 multicast group.
IPv6 Multicast Test
This test module verifies that you can receive packets sent to an IPv6 multicast group.
TCP Bandwidth Test
Packet loss: 0%; RTT (Min/Max/Avg): 0/0/0 ms.
Details
IPv6 Connectivity Test
Success
Packet loss: 0%; RTT (Min/Max/Avg): 0/0/0 ms; Native IPv6 connection.
Details
Traceroute Test
Success
The results of the TRACEROUTE command have been reported.
Details
DNS Performance Test
Success
Host names resolved: 15/15; Total time: 0:00:02.016; Lookups per second: 7.440476190476191.
Details
NAT Test
No
You are not under NAT
Details
Simple Portscan Test
Status: Connected; sending port list...
Stop

TCP Congestion Control

- TCP adjust the rate at which it sends packets to avoid congestion.
- Behaviour is deliberately conservative or benign.
- Uses lost packets to control the transmission rate:
  - assumes packets are lost solely by being dropped from the queues of busy routers.
- Increases the transmission rate until it detects congestion.
- Drops it and starts working up again.
- AIMD: Additive Increase, Multiplicative Decrease.



Congestion control in TCP/IP using AIMD.

Round Trip Times

Path	Round trip time (msec)	Light travel time (msec)
LAN	1	-
Within the UK	<10	3.6
UK to Europe	15-30	6.1
UK to North America	~150	42.4

Default and suggested TCP parameters

Parameter	Default	Suggested
/proc/sys/net/ipv4/tcp_timestamps	1	1
/proc/sys/net/ipv4/tcp_window_scaling	1	1
/proc/sys/net/ipv4/tcp_sack	1	1
/proc/sys/net/core/wmem_max	65535	8388608
/proc/sys/net/core/rmem_max	65535	8388608
/proc/sys/net/ipv4/tcp_rmem	4096	4096
	87380	87380
	174760	4194304
/proc/sys/net/ipv4/tcp_wmem	4096	4096
	16384	65536
	131072	4194304
/proc/sys/net/ipv4/tcp_mem	48128	4194304
	48640	4194304
	49152	4194304

```
cat  /proc/sys/net/core/wmem_max

/sbin/sysctl  -w  net/core/wmem_max=8388608
```

TCP Implementations

Tahoe

Reno

NewReno *

Sack *

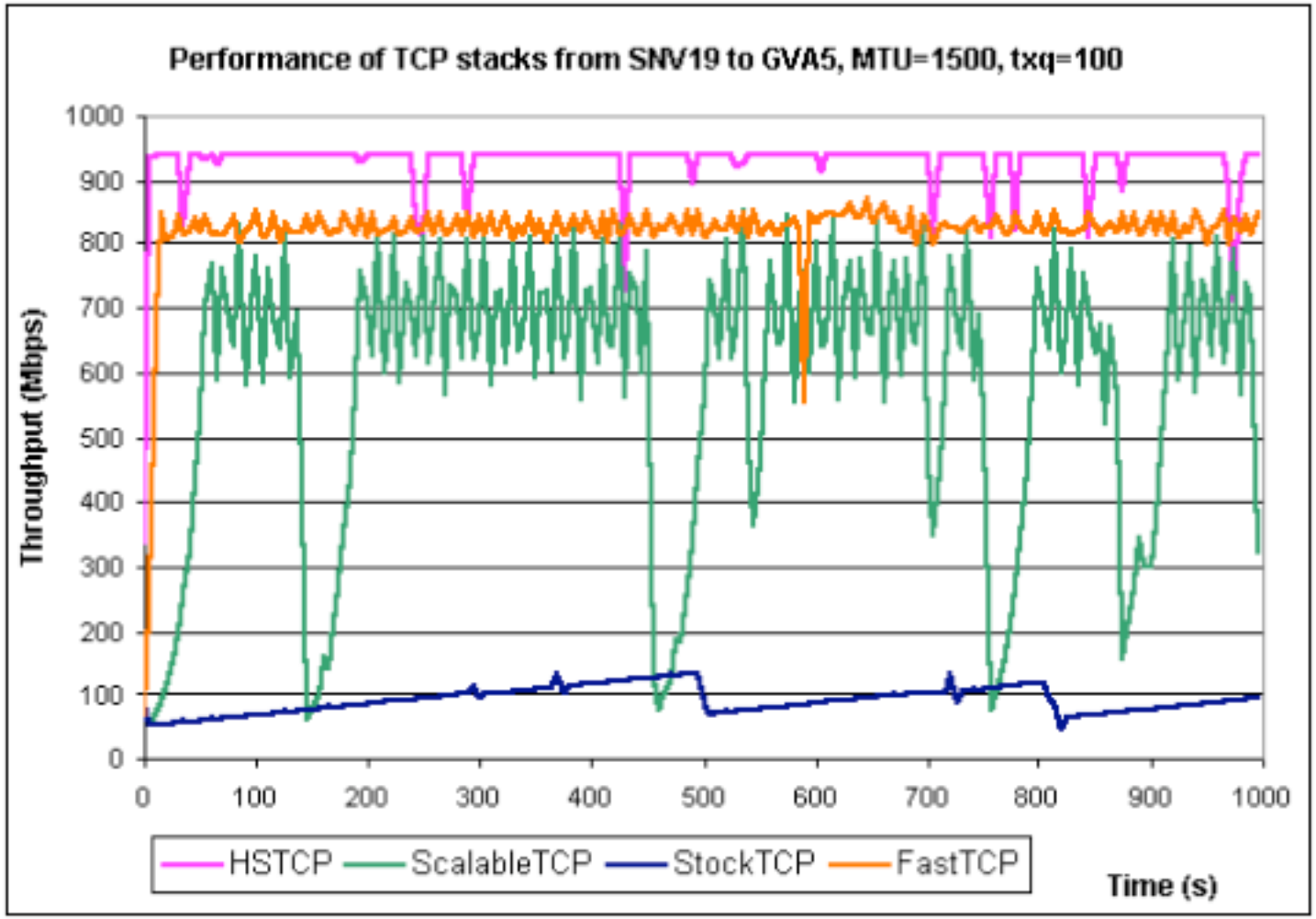
Fast TCP Implementations

- Scalable TCP (STCP)
  http://www-lce.eng.cam.ac.uk/~ctk21/scalable/

- High Speed TCP
  http://www.icir.org/floyd/hstcp.html

- Fast TCP
  http://netlab.caltech.edu/FAST/

- TCP Westwood+
  Available as part of the Linux kernel 2.6.6.


Performance of TCP stacks from SNV19 to GVA5, MTU=1500, txq=100
Throughput (Mbps)
1000
900
800
700
600
500
400
300
200
100
0
0
100
200
300
400
500
600
700
800
900
1000
HSTCP
ScalableTCP
StockTCP
FastTCP
Time (s)

Choosing Hardware

- The Network Interface Cards (NICs) should be well designed, usually with advanced PCI (Peripheral Component Interconnect).
- CPU-Memory bandwidth is important; data crosses the memory bus at least 3 times.
- Use motherboards with multiple 64 bit PCI-X buses to separate the data transfers. Also, a 32 bit 33 MHz bus is too slow for Gigabit transfer rates; a 64 bit 33 MHz bus will be more than 80% used.
- To sustain Gigabit transfer rates fast CPU rates are needed.

Likely Causes of Poor Performance

It is difficult to generalise, some of the likely causes of poor performance are, in decreasing order of likelihood:

- the end host hardware,
- the application software (as distinct from the TCP/IP stack) running on the end hosts,
- some local network limitation,
- a firewall,
- and only finally the WAN connecting the end hosts.

Firewall Rules

- Sometimes simply changing the order of firewall rules can lead to improved performance.
- Try to ensure that the rules are as few and simple as possible.
- Rules are processed sequentially, so the rules applying to high-volume traffic should be placed towards the start of the list.

Further Information

- A.S. Tanenbaum, 2002, *Computer Networks*, fourth edition, (Prentice Hall: Upper Saddle River, New Jersey).

- 'Networks for Non-Networkers' (NFNN2) workshop, to be run at NeSC, 20-21 June 2005,
  http://www.nesc.ac.uk/esi/events/533/

- The Web site for the first NFNN workshop, held at UCL last year,
  http://research-computing.ucl.ac.uk/NFNN.html

- *An Introduction to Computer Network Monitoring and Performance*, to be included with the course materials for NFNN2 and available separately on request (contact me: clive@nesc.ac.uk).