Network Measurement and Monitoring Activities

OSG Site-admin meeting
Aug 6-7, 2009
Rich Carlson
rcarlson@internet2.edu

INTERNET2®

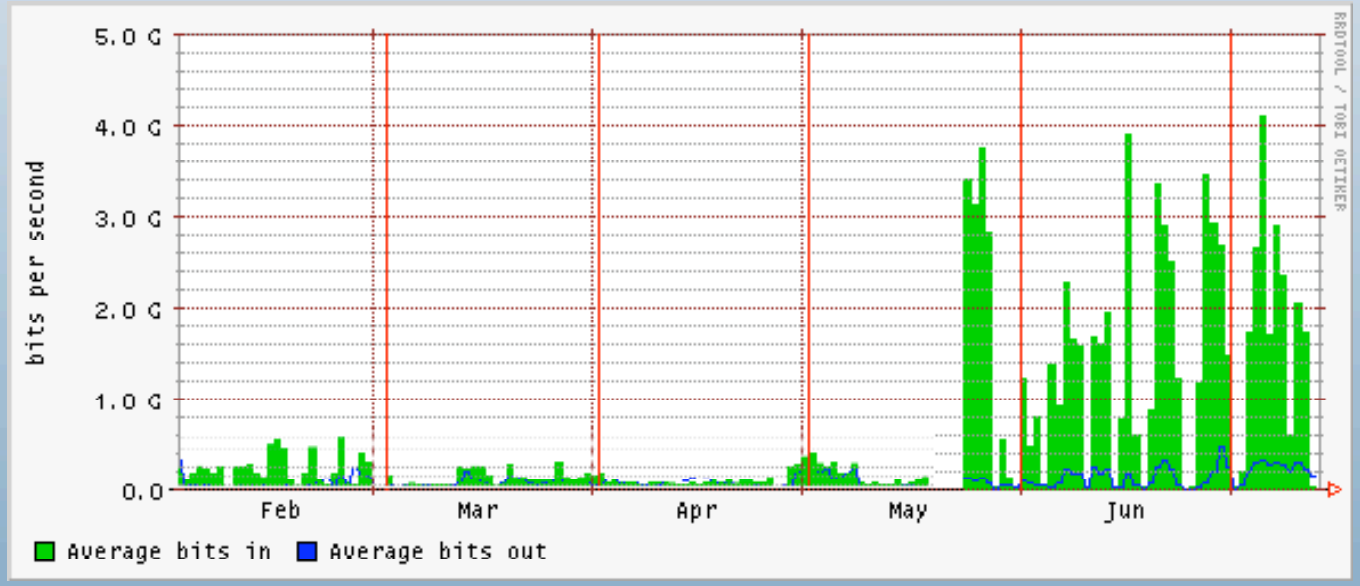
Network Management 101

- There are 3 major functions that every OSG site-admin needs
  - Monitoring – verifying the health and operation of the internal network
  - Measuring – verifying the capacity and utilization of the peer networks
  - Advanced diagnostics – identifying what has failed when a fault is found

INTERNET®

Network Monitoring

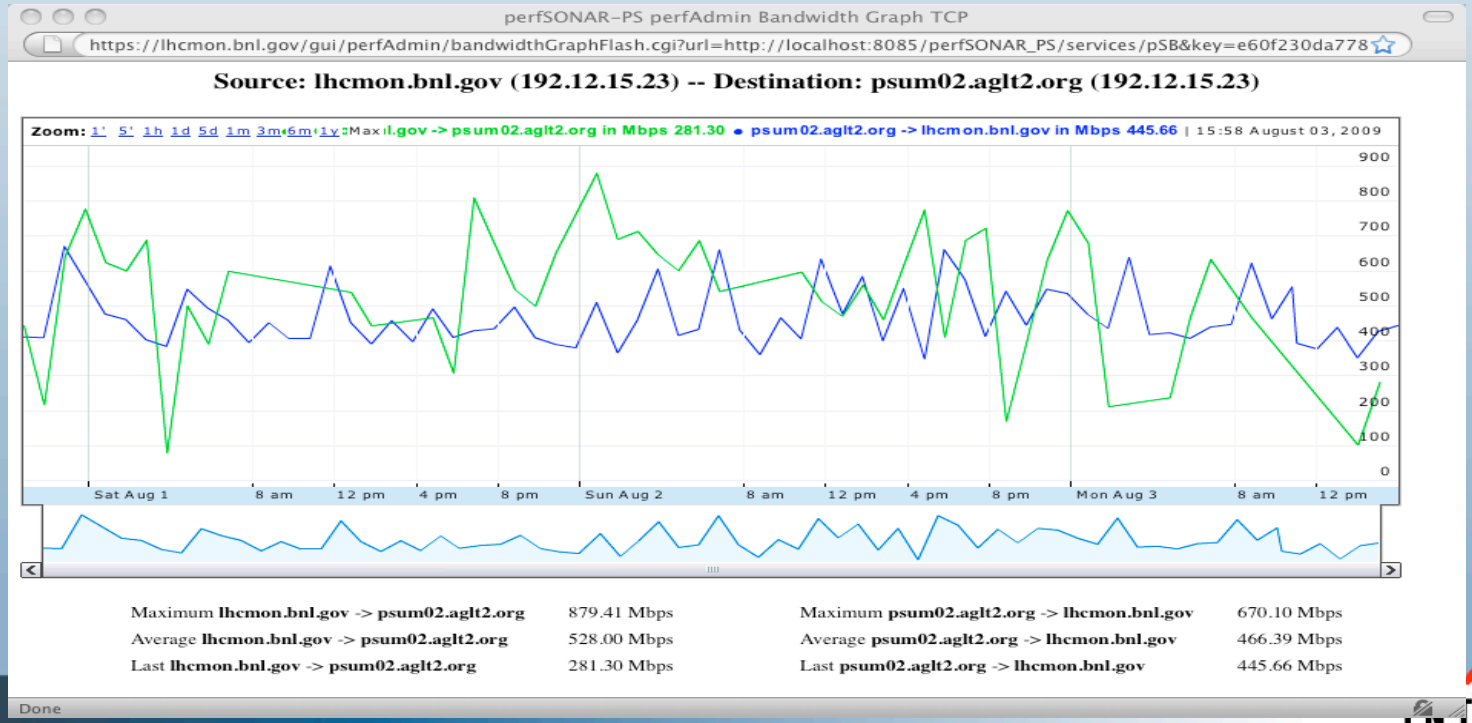
- Collect statistics from routers and switches
- Display results in graphical form



INTERNET®

Network Measurement

- Verify performance levels between peer sites



Network Diagnostic

- Collect/Analyze data to respond to faults
- Detailed results can point to real problems

PNNL sending to:	Seattle	Salt Lake City	Washington
Throughput	966 Mbps	930 Mbps	328 Mbps
Min RTT	6 msec	23 msec	76 msec
Loss	0.0094%	0.0045%	.0049%
Reordering	6.04%	5.5%	5.15%

INTERNET2®

perfSONAR Overview

- A set of high level services for managing multi-domain measurement/monitoring infrastructures
- International community of developers
  - Implementing Open Grid Forum (OGF) Network Measurement (NM-WG) recommendations
- Multiple sets of interoperable software

INTERNET®

perfSONARized tools

- Implement 1 or more perfSONAR services
  - Registration, Data Storage, …
- Existing tools include
  - Cacti
  - BWCTL
  - NDT
  - NPAD
  - OWAMP
  - PingER

INTERNET2®

Measurement/Monitoring Tools

- Client / Server based tools
  - Measure throughput
  - Measure delay & loss
  - Analyze infrastructure, switch/router configuration, and host configuration

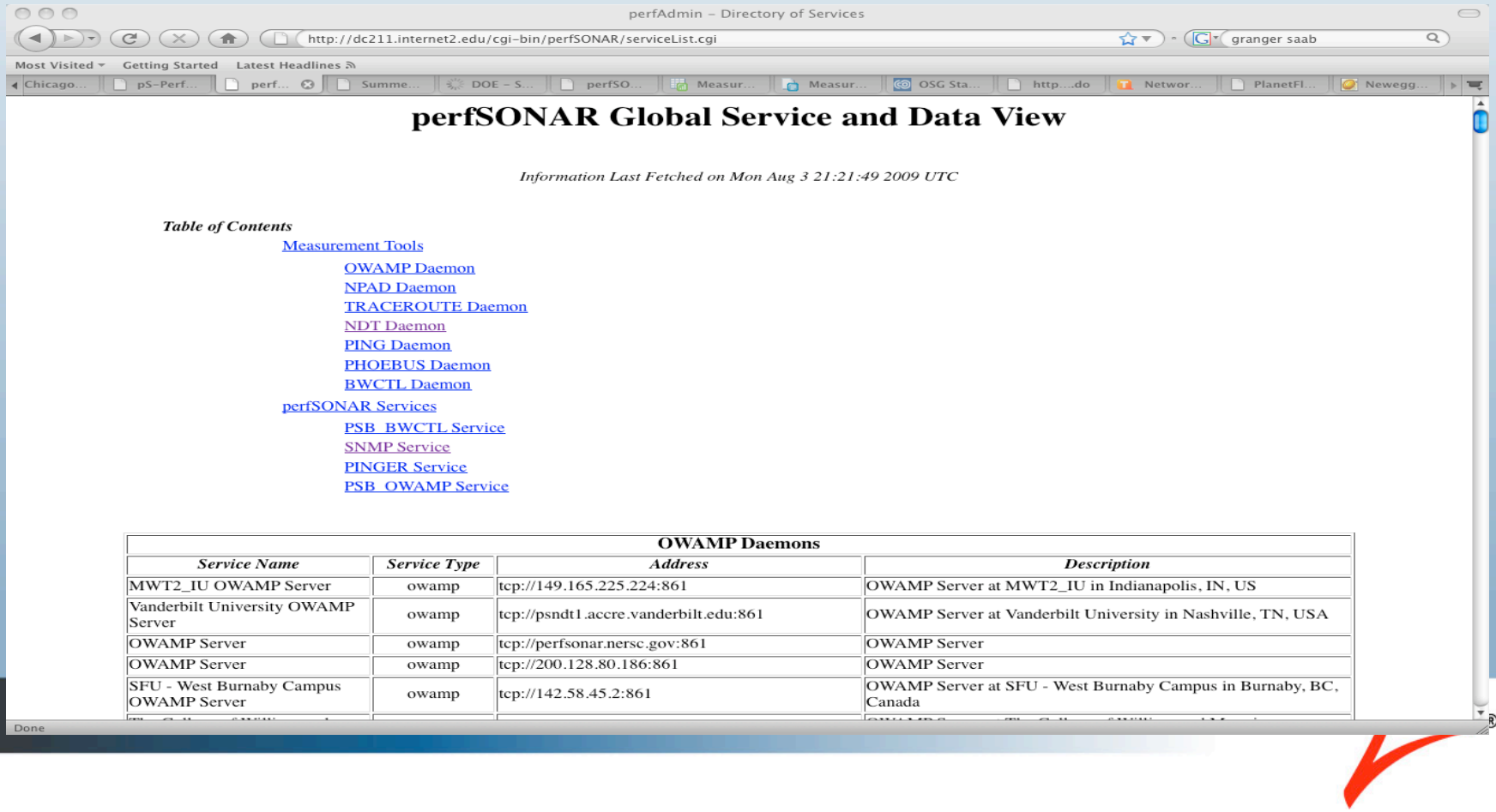
INTERNET®

Server Deployment options

- Source tar files – build your own
- OS Distribution packages (yum, apt)
- Bootable CD

Finding a deployed server

- http://www.perfsonar.net/activeServices



Client Deployment options

- Source tar files – build your own
- OS distro packages
- Java based clients using web browser
- Client executables distributed via VDT
  - Allows testing from any CE/SE
  - Command syntax and output will be covered later in this talk

Monitoring Tools

- BWCTL – Bandwidth Control
  - Allows single person operation over wide area testing environment
  - Runs NLANR ‘iperf’ program
- OWAMP – One way Delay Measurement
  - Advanced ‘ping’ command
  - Allows single person operation over wide area testing environment

INTERNET®

Sample BWCTL results

rcarlson@triton:~/ndt — ssh — 116×33

```
-bash-3.2$ bwctl -i2 -t20 -s lhcmon.bnl.gov
bwctl: Unable to contact a local bwctld: Spawning local tool controller
bwctl: NuttcpAvailable(): We were unable to verify that nuttcp is working. Likely you do not have it installed. exit
 status: 1: output: exec(nuttcp): No such file or directory
bwctl: Couldn't initialize tool "nuttcp". Disabling it.
bwctl: Using tool: iperf
bwctl: 28 seconds until test results available

RECEIVER START
bwctl: exec_line: iperf -B 207.75.164.104 -s -f a -m -p 5001 -t 20 -i 2
bwctl: start_tool: 3458324284.538605
------------------------------------------------------------
Server listening on TCP port 5001
Binding to local address 207.75.164.104
TCP window size: 85.3 KByte (default)
------------------------------------------------------------
[ 12] local 207.75.164.104 port 5001 connected with 192.12.15.23 port 5001
[ 12]  0.0- 2.0 sec  19.5 MBytes  81.8 Mbits/sec
[ 12]  2.0- 4.0 sec  22.4 MBytes  94.0 Mbits/sec
[ 12]  4.0- 6.0 sec  22.4 MBytes  94.0 Mbits/sec
[ 12]  6.0- 8.0 sec  22.4 MBytes  94.0 Mbits/sec
[ 12]  8.0-10.0 sec  21.4 MBytes  89.7 Mbits/sec
[ 12] 10.0-12.0 sec  19.0 MBytes  79.8 Mbits/sec
[ 12] 12.0-14.0 sec  21.6 MBytes  90.6 Mbits/sec
[ 12] 14.0-16.0 sec  22.4 MBytes  94.1 Mbits/sec
[ 12] 16.0-18.0 sec  19.6 MBytes  82.4 Mbits/sec
[ 12] 18.0-20.0 sec  20.2 MBytes  84.6 Mbits/sec
[ 12]  0.0-20.1 sec    213 MBytes  88.5 Mbits/sec
[ 12] MSS size 1448 bytes (MTU 1500 bytes, ethernet)
bwctl: stop_exec: 3458324308.780934

RECEIVER END
-bash-3.2$
```

INTERNET®

Using BWCTL: commands

bwctl -i2 -t20 -c bwctl.losa.net.internet2.edu

bwctl -i2 -t20 -s bwctl.newy.net.internet2.edu

- bwctl = name of program
- i2 = report intermediate results every 2 seconds
- t20 – run test for 20 seconds
- s name = remote end will send data to you
- c name = you will send data to the remote host

3rd party testing: command

bwctl -i2 -t20 -c bwctl.salt.net.internet2.edu -s bwctl.atla.net.internet2.edu

- User can run a test between 2 remote hosts

INTERNET®

OWping Results

rcarlson@nms-rlat:~ — ssh — 116×34

```
[rcarlson@nms-rlat ~]$ ping -c3 owamp.losa.net.internet2.edu
PING eth-1.nms-rlat.losa.net.internet2.edu (64.57.17.162) 56(84) bytes of data.
64 bytes from nms-rlat.losa.net.internet2.edu (64.57.17.162): icmp_seq=1 ttl=60 time=56.6 ms
64 bytes from nms-rlat.losa.net.internet2.edu (64.57.17.162): icmp_seq=2 ttl=60 time=56.7 ms
64 bytes from nms-rlat.losa.net.internet2.edu (64.57.17.162): icmp_seq=3 ttl=60 time=56.6 ms

--- eth-1.nms-rlat.losa.net.internet2.edu ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2005ms
rtt min/avg/max/mdev = 56.692/56.697/56.705/0.275 ms
[rcarlson@nms-rlat ~]$ owping owamp.losa.net.internet2.edu
Approximately 13.2 seconds until results available

--- owping statistics from [2001:468:2:12::17:34]:35621 to [owamp.losa.net.internet2.edu]:55664 ---
SID:    00170162ce21d9957fb85e80c7f6958a
first:  2009-08-03T21:48:06.786
last:   2009-08-03T21:48:15.918
100 sent, 0 lost (0.000%), 0 duplicates
one-way delay min/median/max = 28.1/28.2/28.3 ms, (err=3.13 ms)
one-way jitter = 0.1 ms (P95-P50)
TTL not reported
no reordering


--- owping statistics from [owamp.losa.net.internet2.edu]:34879 to [2001:468:2:12::17:34]:35622 ---
SID:    00170034ce21d9959d6ffc11d84cfeac
first:  2009-08-03T21:48:06.907
last:   2009-08-03T21:48:17.704
100 sent, 0 lost (0.000%), 0 duplicates
one-way delay min/median/max = 28.5/28.6/28.6 ms, (err=3.13 ms)
one-way jitter = 0 ms (P95-P50)
TTL not reported
no reordering

[rcarlson@nms-rlat ~]$
```

Using OWPING

- owping owamp.losa.net.internet2.edu
  - owping = program name
  - owamp… = name of server
- Output results
  - Separate statistics for both directions
  - Number of packets sent and lost
  - One-way delay statistics min/median/max
  - Number of IP hops in path
  - Number of packets that arrives out-of-order

INTERNET®

Advanced Diagnostic tools

- Existing NDT tool
  - Allows users to test network path for a limited number of common problems
- Existing NPAD tool
  - Allows users to test local network infrastructure while simulating a long path

INTERNET®

Network Diagnostic Tool (NDT)

- Measure performance to users desktop
- Identify real problems for real users
  - Network infrastructure is the problem
  - Host tuning issues are the problem
- Make tool simple to use and understand
- Make tool useful for users and network administrators

INTERNET®

NDT Client test

rcarlson@nms-rexp:~ — ssh — 116×35

```
[rcarlson@nms-rexp ~]$ web100clt -4 -n ndt.kans.net.internet2.edu -l
Testing network path for configuration and performance problems  --  Using IPv4 address
Checking for Middleboxes . . . . . . . . . . . . . . . . . . .  Done
checking for firewalls . . . . . . . . . . . . . . . . . . . .  Done
running 10s outbound test (client to server) . . . . .  983.18 Mb/s
running 10s inbound test (server to client) . . . . . . 977.73 Mb/s
The slowest link in the end-to-end path is a 1.0 Gbps Gigabit Ethernet subnet
Information: Other network traffic is congesting the link
Server 'ndt.kans.net.internet2.edu' is not behind a firewall. [Connection to the ephemeral port was successful]
Client is not behind a firewall. [Connection to the ephemeral port was successful]

        ------  Web100 Detailed Analysis  ------

Web100 reports the Round trip time = 47.32 msec;the Packet size = 8948 Bytes; and
No packet loss was observed.
This connection is receiver limited 1.63% of the time.
This connection is network limited 98.08% of the time.

    Web100 reports TCP negotiated the optional Performance Settings to:
RFC 2018 Selective Acknowledgment: ON
RFC 896 Nagle Algorithm: ON
RFC 3168 Explicit Congestion Notification: OFF
RFC 1323 Time Stamping: ON
RFC 1323 Window Scaling: ON; Scaling Factors - Server=10, Client=10
The theoretical network limit is 154.82 Mbps
The NDT server has a 32768 KByte buffer which limits the throughput to 5410.20 Mbps
Your PC/Workstation has a 16673 KByte buffer which limits the throughput to 2752.82 Mbps
The network based flow control limits the throughput to 1732.74 Mbps

Client Data reports link is '  7', Client Acks report link is '  7'
Server Data reports link is '  7', Server Acks report link is '  7'
Packet size is preserved End-to-End
Server IP addresses are preserved End-to-End
Client IP addresses are preserved End-to-End
[rcarlson@nms-rexp ~]$
```

Finding Results of Interest

- Duplex Mismatch
  - This is a serious error and nothing will work right. Reported on *main* page, on *Statistics* page, and **mismatch:** on *More Details* page
- Packet Arrival Order
  - Inferred value based on TCP operation. Reported on *Statistics* page, (with loss statistics) and **order:** value on *More Details* page

INTERNET®

Finding Results of Interest

- Packet Loss Rates
  - Calculated value based on TCP operation. Reported on *Statistics* page, (with out-of-order statistics) and **loss:** value on *More Details* page
- Path Bottleneck Capacity
  - Measured value based on TCP operation. Reported on *main* page

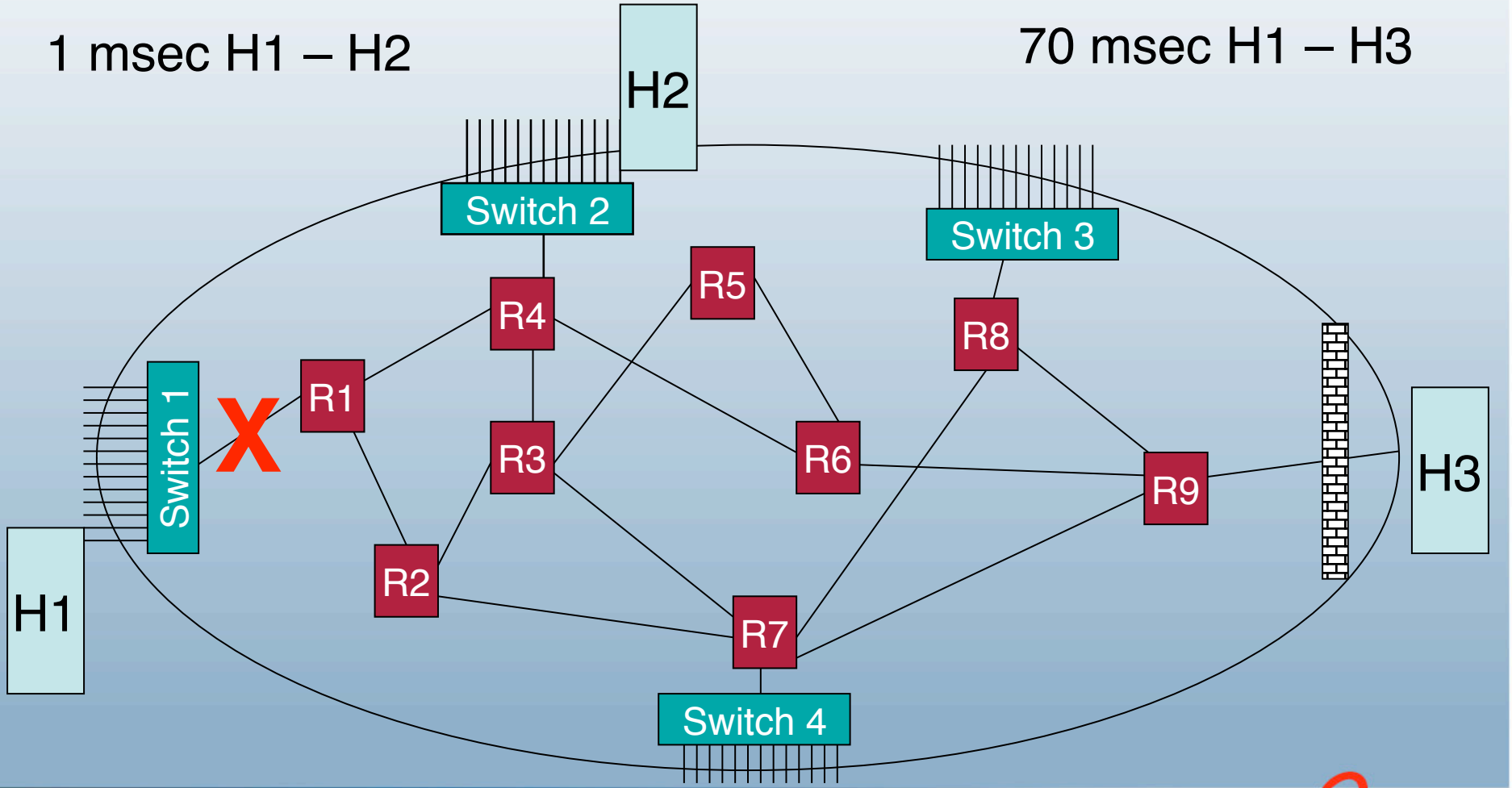
INTERNET®

NPAD/pathdiag

- A new tool from researchers at Pittsburgh Supercomputer Center
- Finds problems that affect long network paths
- Uses Web100-enhanced Linux based server
- Web based Java client

INTERNET®

Long Path Problem



INTERNET2®

Long Path Problem

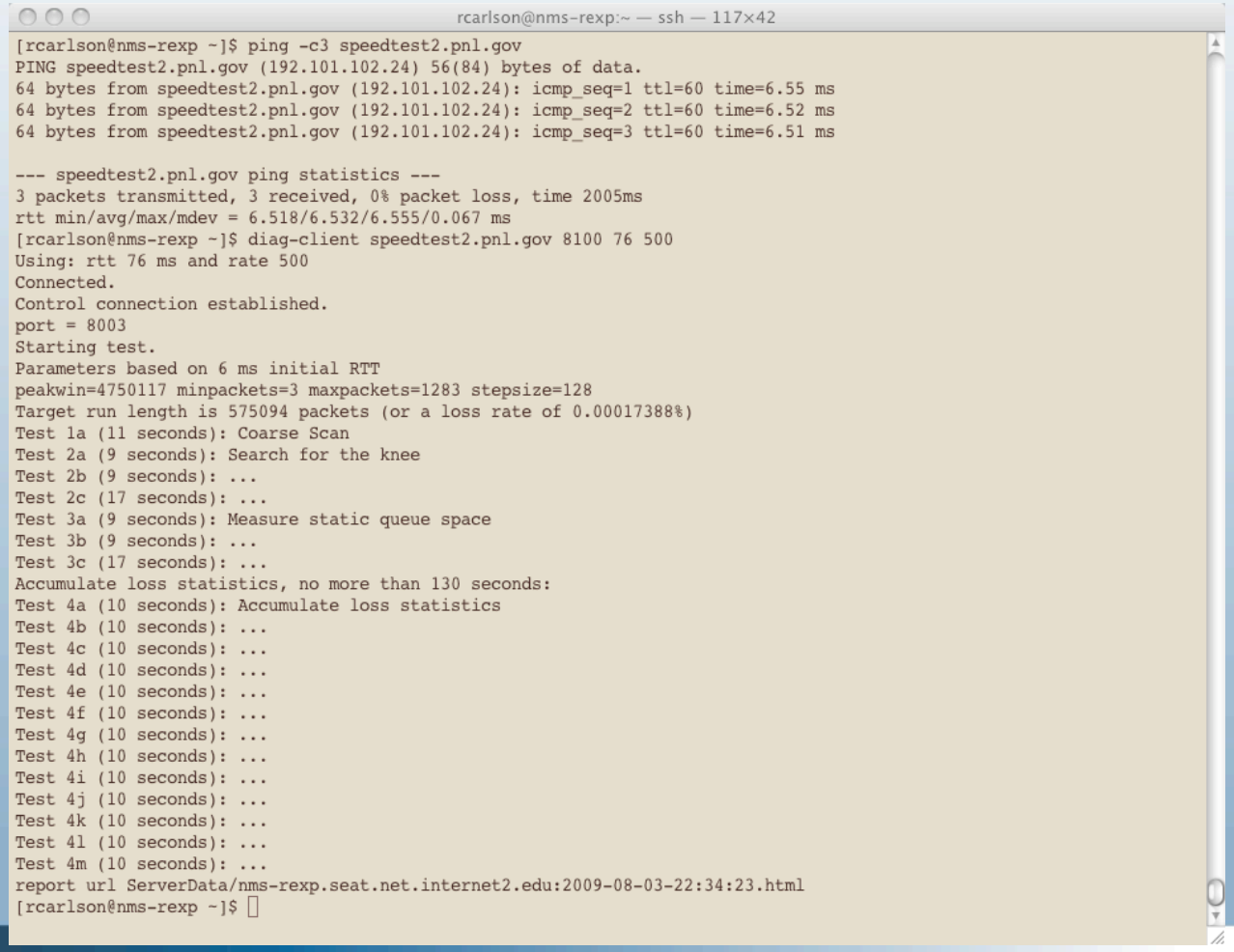
- E2E application performance is dependant on distance between hosts
- Full size frame time at 1000 Mbps
  - Frame = 1500 Bytes
  - Time = 12 usec
  - In flight for 1 msec RTT = 83 packets
  - In flight for 70 msec RTT = 5833 packets

INTERNET®

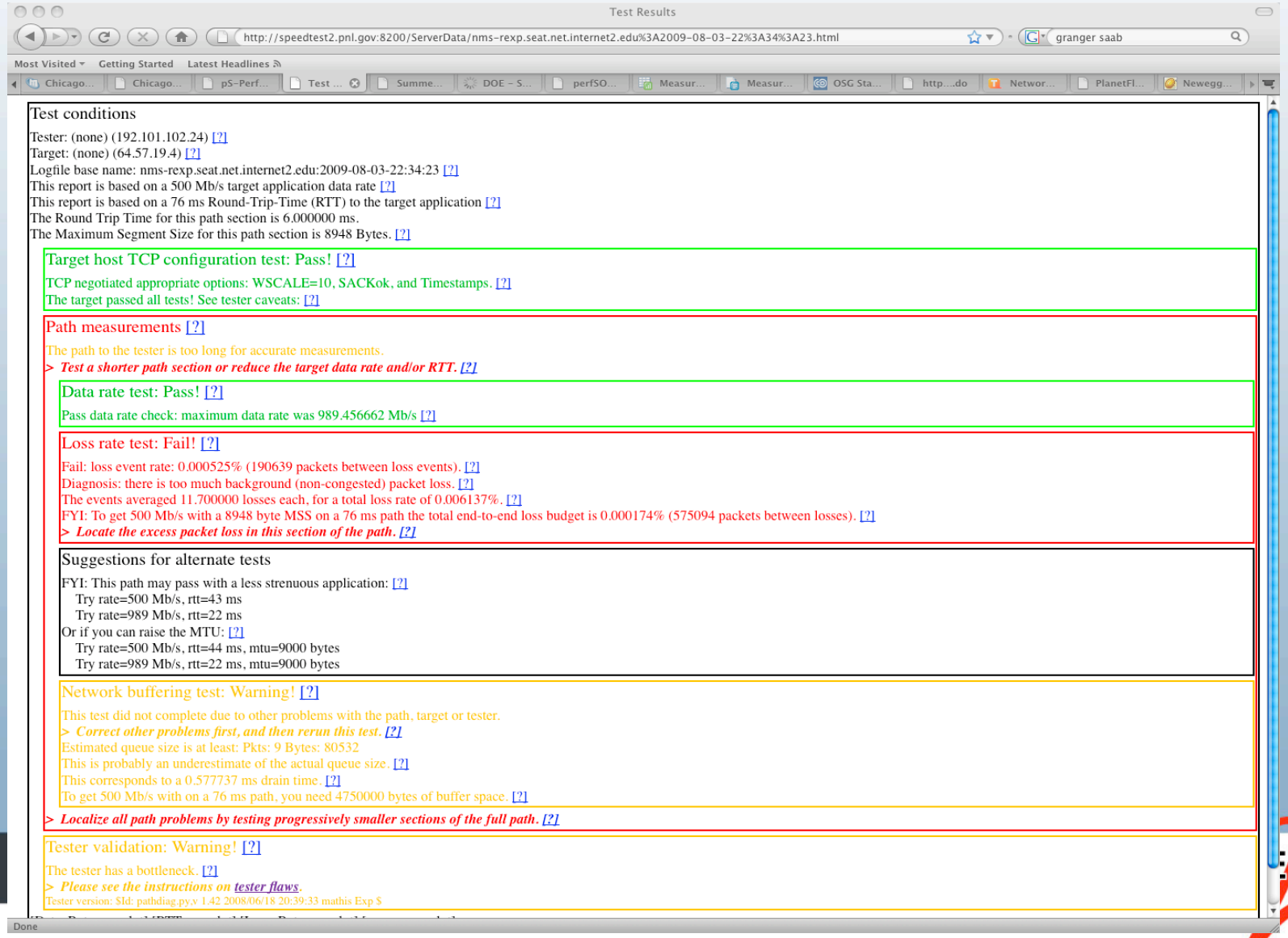
TCP Congestion Avoidance

- Cut number of packets by ½
- Increase by 1 per RTT
  - LAN (RTT=1msec)
    - In flight changes to 41 packets
    - Time to increase back to 83 is 42 msec
  - WAN (RTT = 70 msec)
    - In flight changes to 2917 packets
    - Time to increase back to 5833 is 3.4 minutes!

NPAD Client test



NPAD Sample results



Diagnostic Testing Results

- SC'04 – buffer problem on SGI server limited throughput to 20 Mbps
- PNNL – buffer limits in router causing TCP loss
- UMich – forwarding table overflowed some flows using CPU forwarding
- OU – Regional upgrade caused 100x slowdown
- MAX – Regional network uplink saturated
- I2 – Linux kernel mm penalizes applications that don't fill kernel buffers

INTERNET®

Putting it all together – SCP file transfer

Bob and Carol are collaborating on a project. Bob needs to send a copy of the data (50 MB) to Carol every ½ hour. Bob and Carol are 2,000 miles apart. How long should each transfer take?

- 5 minutes?
- 1 minute?
- 5 seconds?

INTERNET®

What should we expect?

Assumptions:

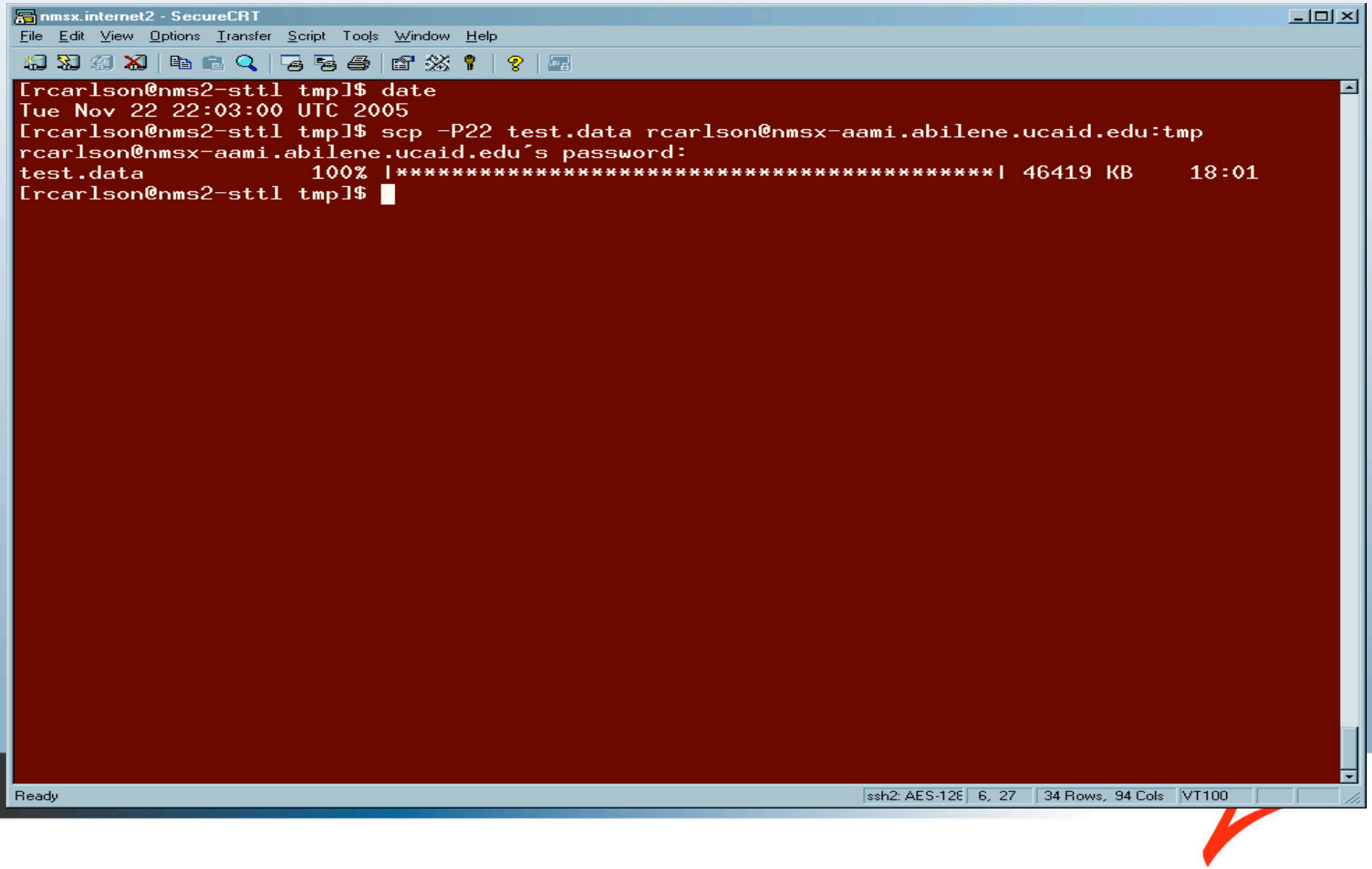
- 100 Mbps Fast Ethernet is the slowest link
- 50 msec round trip time

Bob & Carol calculate:

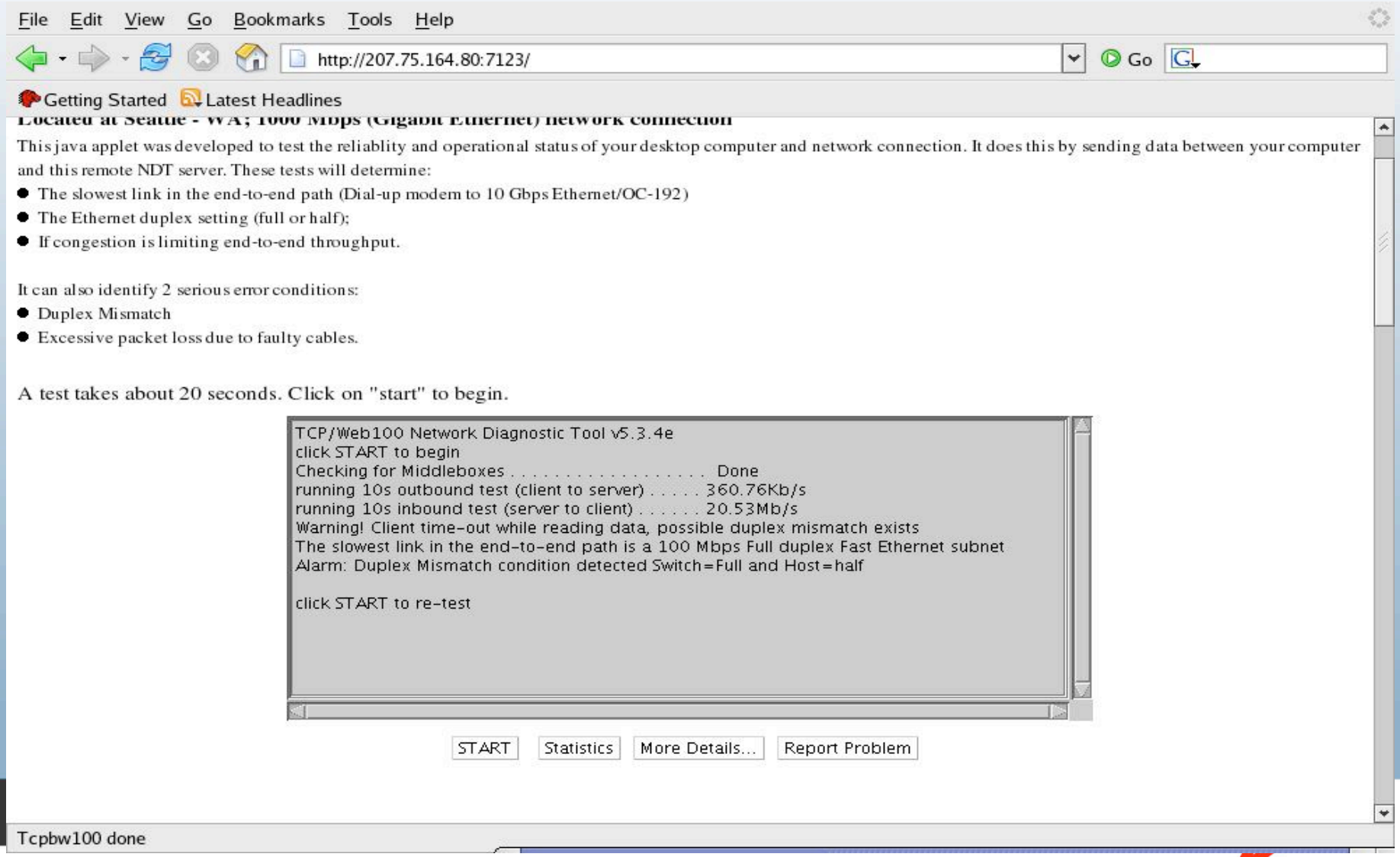
- 50 MB * 8 = 400 Mbits
- 400 Mb / 100 Mb/sec = 4 seconds

INTERNET®

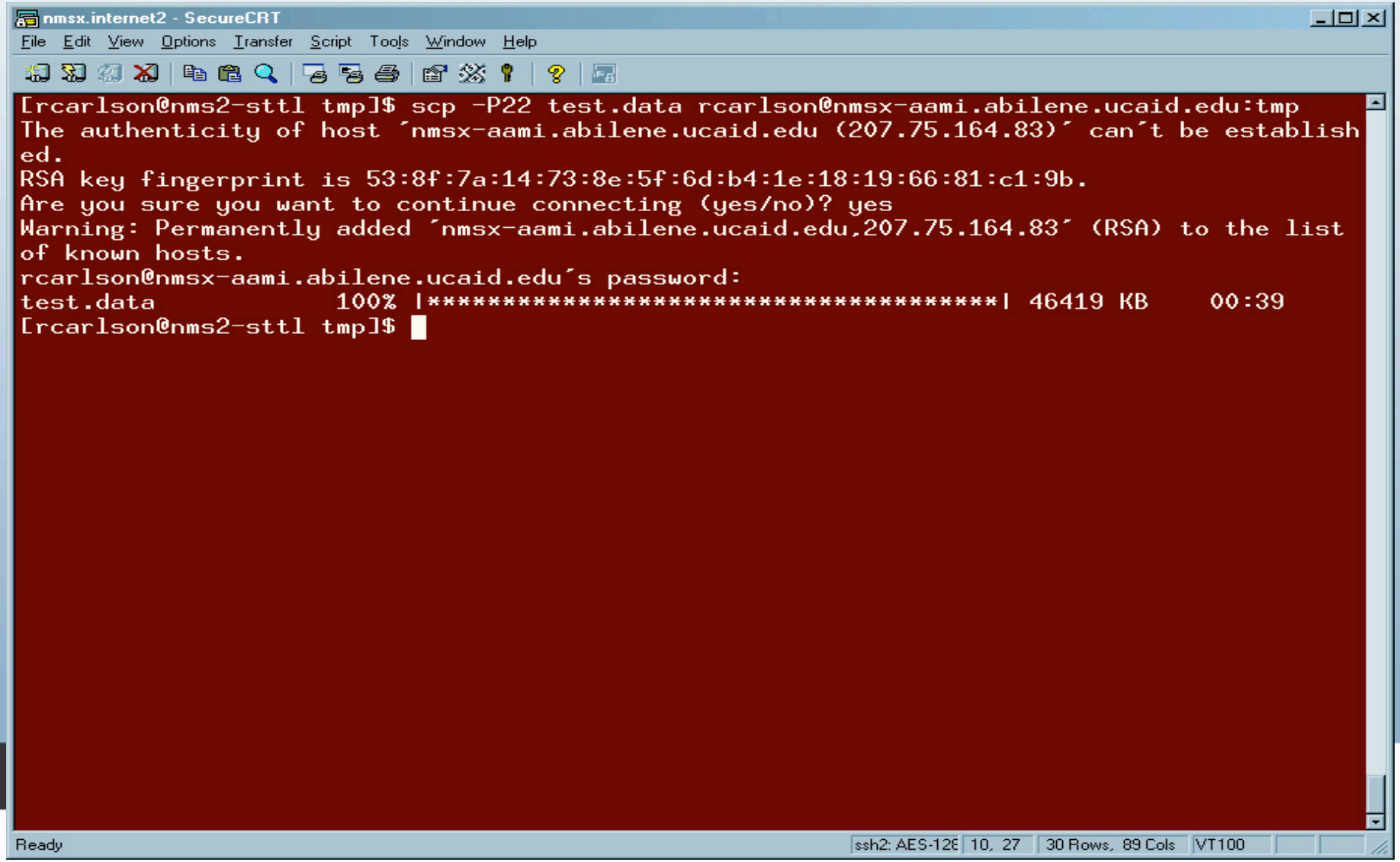
Initial SCP Test Results



Initial NDT testing shows Duplex Mismatch at one end



SCP results after Duplex Mismatch Corrected



Intermediate Results

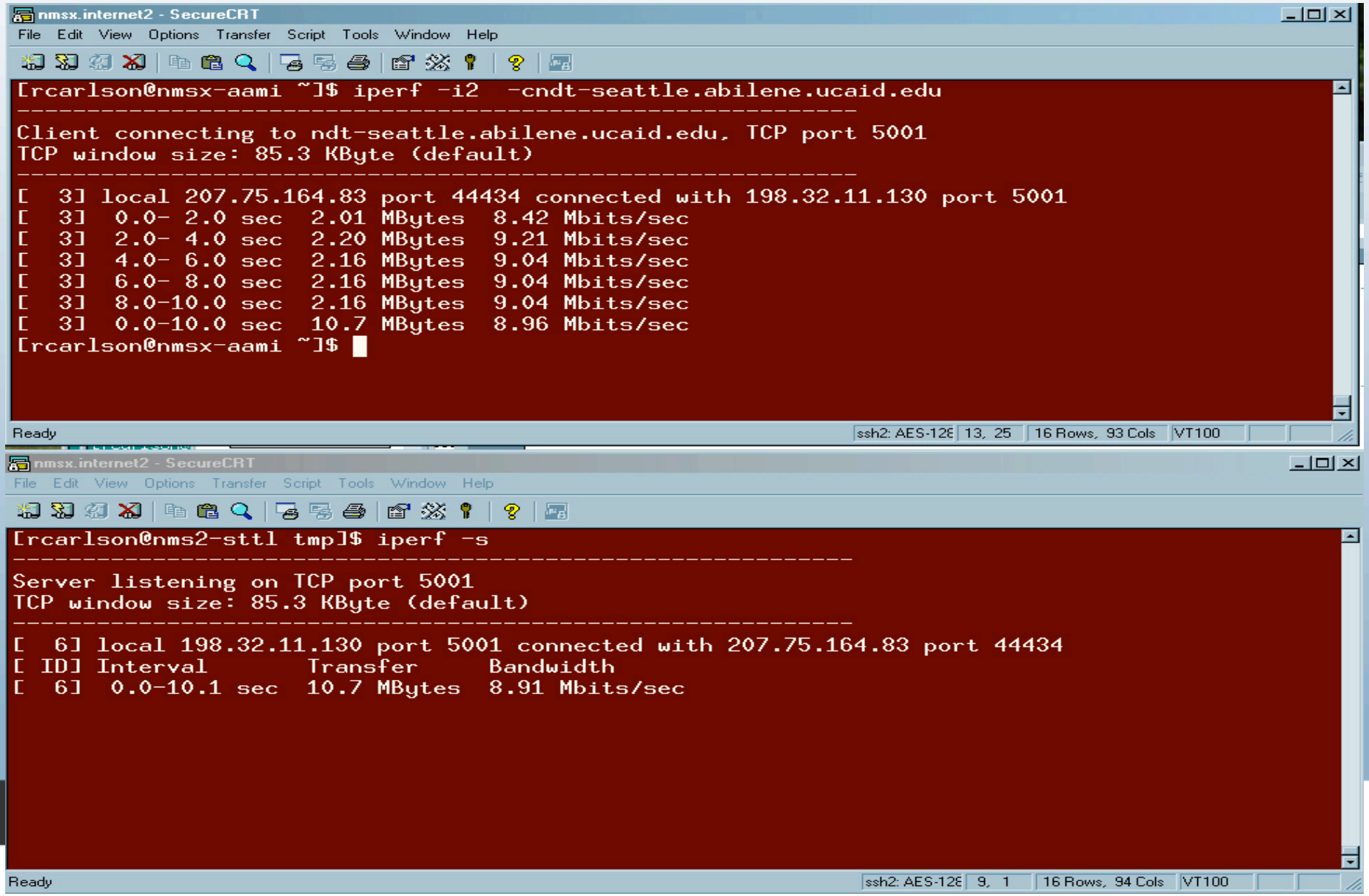
Time dropped from 18 minutes to 40 seconds.

But our calculations said it should take 4 seconds!

- 400 Mb / 40 sec = 10 Mbps
- Why are we limited to 10 Mbps?
- Are you satisfied with 1/10th of the possible performance?

INTERNET®

Default TCP window settings



Calculating the Window Size

Remember Bob found the round-trip time was 50 msec

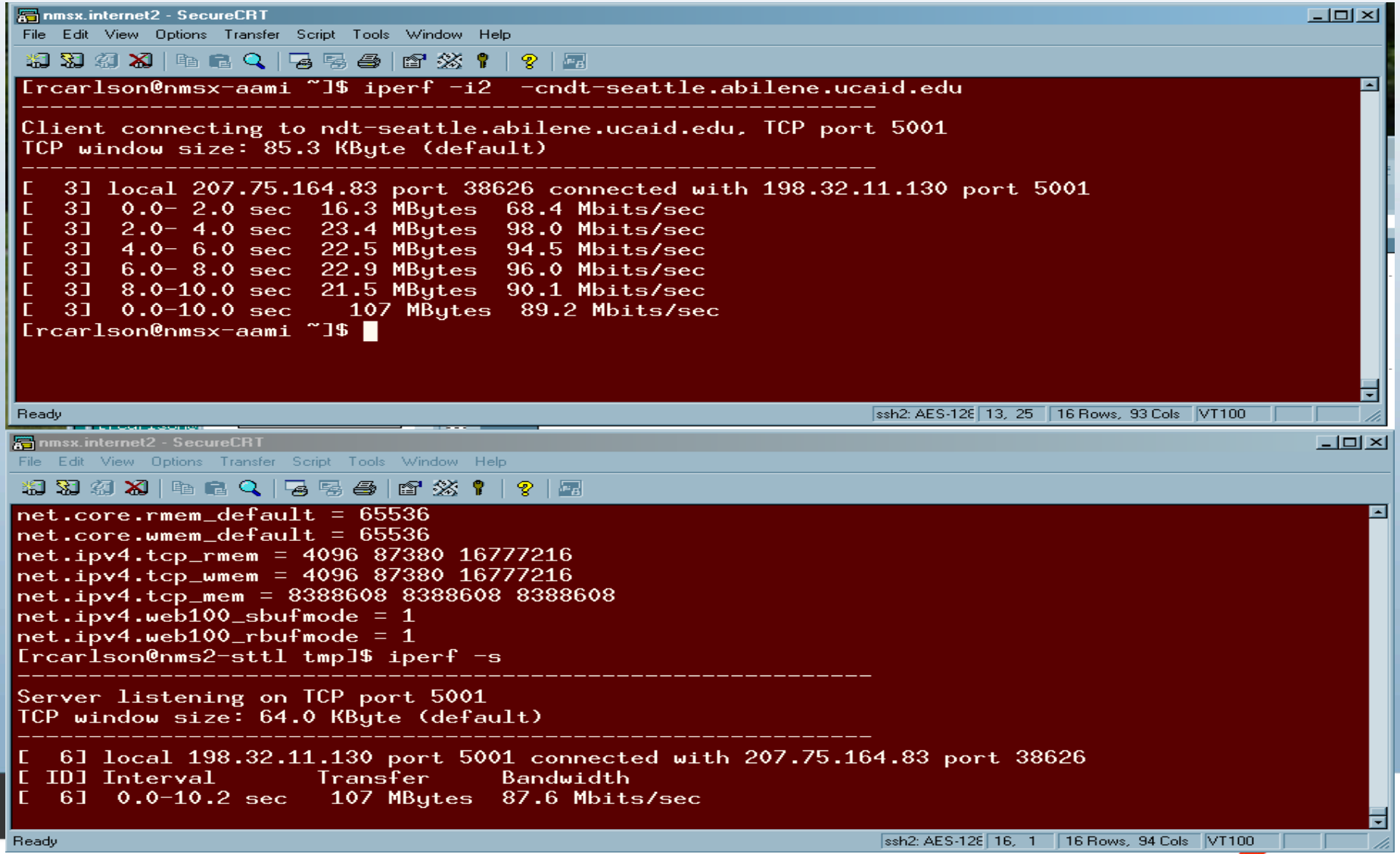
Calculate window size limit

- 85.3KB * 8 b/B = 698777 b
- 698777 b / .050 s = 13.98 Mbps

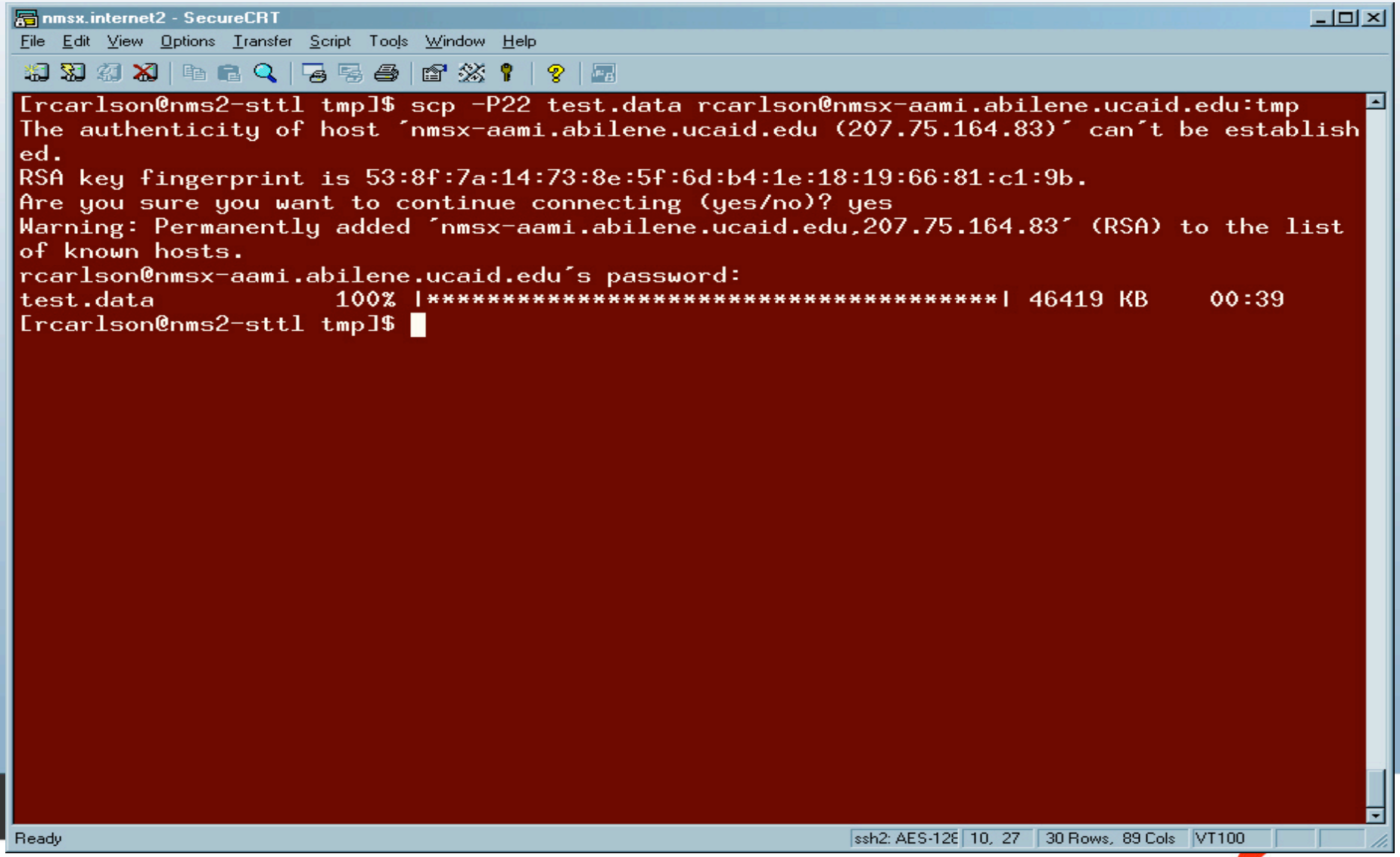
Calculate new window size

- (100 Mb/s * .050 s) / 8 b/B = 610.3 KB
- Use 1MB as a minimum

With TCP windows tuned



SCP results with tuned windows

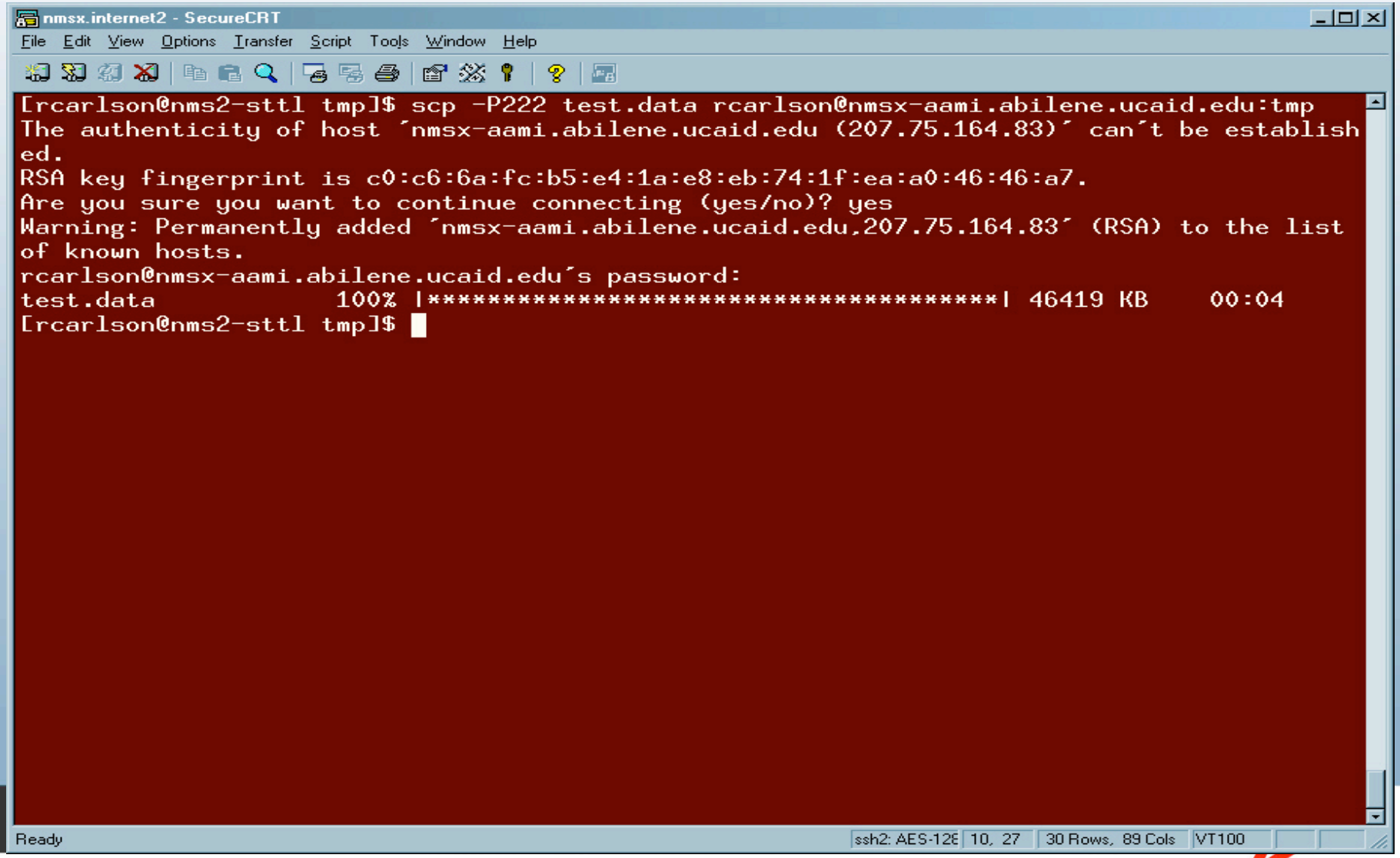


Intermediate Results

SCP still runs slower than expected

- Hint: SCP uses internal buffers
- Patch available from PSC

SCP Results with tuned SCP



Final Results

Fixed infrastructure problem

Fixed host configuration problem

Fixed Application configuration problem

- Achieved target time of 4 seconds to transfer 50 MB file over 2000 miles

INTERNET®

Conclusions

- Primary tools still useful
- Advanced tools are being developed
- Developing tools will make things even easier
- Demand 10 MB/s as the minimum acceptable throughput rate

INTERNET®