

# Tier3 Network Issues

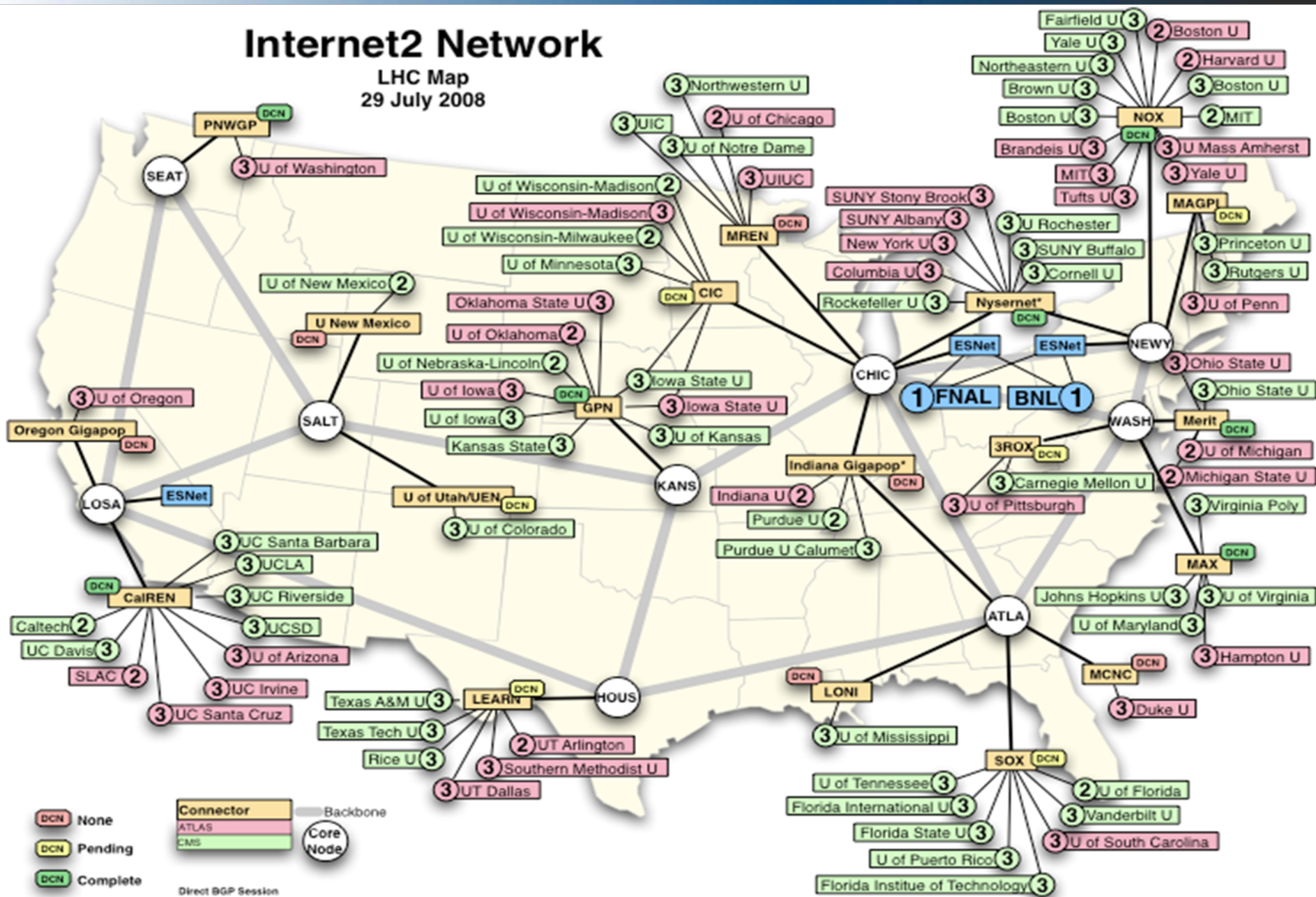
Richard Carlson

May 19, 2009

[rcarlson@internet2.edu](mailto:rcarlson@internet2.edu)

# Internet2 Network

LHC Map  
29 July 2008



# Internet2 overview

- Member organization with a national backbone infrastructure
  - Campus & Regional network members
  - National and International peers
- Tiered connection model
  - Campus  $\leftrightarrow$  Regional  $\leftrightarrow$  Backbone
  - Shared IP and Circuit based infrastructure
- Assistance with technical and non-technical problems

# Basic Premise

- Application's performance should meet your expectations!
- If they don't you should complain!
- However, you must complain effectively!

# Realistic Expectations

- What the ATLAS physicists needs to define
  - How large is a dataset
  - How long should it take to move this dataset
  - How often will this dataset be renewed
- This data can be turned into network infrastructure requirements

# Data movement over REN networks

Link Spd	Byts/hour	Fair Share	xfer 1 TB
100 Mbps	45 GB/h	34 GB/h	28 hours
1 Gbps	450 GB/h	120 GB/h	8 hours
10 Gbps	4.5 TB/h	1 TB/h	1 hour

# Basic Connectivity Tests

- Ping
  - Confirms that remote host is 'up'
  - Some network operators block these packets
- Traceroute
  - Identifies the routers along the path
  - Same blocking problem as above
  - Routers treat TR packets with lower priority

# Advanced user 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



# 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

# NPAD/pathdiag

- A new tool from researchers at Pittsburgh Supercomputer Center
- Finds problems that affect long network paths
- Identifies host tuning and network infrastructure problems

# NDT/NPAD user interface

- Web100 based servers (requires patched Linux kernel)
- Web-based JAVA applet allows testing from any browser
- Command-line client allows testing from remote login shell – Client installed in OSG VDT

# Initial NDT testing shows Duplex Mismatch at one end

File Edit View Go Bookmarks Tools Help

http://207.75.164.80:7123/

Getting Started Latest Headlines

Located at Seattle - WA; 1000 Mbps (Gigabit Ethernet) network connection

This java applet was developed to test the reliability and operational status of your desktop computer and network connection. It does this by sending data between your computer and this remote NDT server. These tests will determine:

- The slowest link in the end-to-end path (Dial-up modem to 10 Gbps Ethernet/OC-192)
- The Ethernet duplex setting (full or half);
- If congestion is limiting end-to-end throughput.

It can also identify 2 serious error conditions:

- Duplex Mismatch
- Excessive packet loss due to faulty cables.

A test takes about 20 seconds. Click on "start" to begin.

```
TCP/Web100 Network Diagnostic Tool v5.3.4e
click START to begin
Checking for Middleboxes . . . . . Done
running 10s outbound test (client to server) . . . . . 360.76Kb/s
running 10s inbound test (server to client) . . . . . 20.53Mb/s
Warning! Client time-out while reading data, possible duplex mismatch exists
The slowest link in the end-to-end path is a 100 Mbps Full duplex Fast Ethernet subnet
Alarm: Duplex Mismatch condition detected Switch=Full and Host=half

click START to re-test
```

START Statistics More Details... Report Problem

Tcpcb100 done

# NPAD Sample results

**Test conditions**

Tester: (none) (207.75.164.80) [?]  
Target: (none) (24.15.178.61) [?]  
Logfile base name: c-24-15-178-61.hsd1.il.comcast.net:2007-01-18-23:15:48 [?]  
This report is based on a 7 Mb/s target application data rate [?]  
This report is based on a 22 ms Round-Trip-Time (RTT) to the target application [?]  
The Round Trip Time for this path section is 21.048524 ms.  
The Maximum Segment Size for this path section is 1460 Bytes. [?]

**Target host TCP configuration test: Warning!** [?]  
Warning: TCP connection is not using RFC1323 timestamps. [?]  
Diagnosis: The target (client) is not properly configured. [?]  
Warnings reflect problems that might not affect target end-to-end performance. [?]  
> See TCP tuning instructions at <http://www.psc.edu/networking/projects/tcp tune/> [?]

**Path measurements** [?]

**Data rate test: Pass!** [?]  
Pass data rate check: maximum data rate was 8.969226 Mb/s [?]

**Loss rate test: Pass!** [?]  
Pass: measured loss rate 0.035214% (2839 packets between loss events). [?]  
FYI: To get 7 Mb/s with a 1460 byte MSS on a 22 ms path the total end-to-end loss budget is 0.282486% (354 packets between losses). [?]

**Suggestions for alternate tests**  
FYI: This path may even pass with a more strenuous application: [?]  
Try rate=7 Mb/s, rtt=62 ms  
Try rate=8 Mb/s, rtt=48 ms  
Or if you can raise the MTU: [?]  
Try rate=7 Mb/s, rtt=383 ms, mtu=9000 bytes  
Try rate=8 Mb/s, rtt=299 ms, mtu=9000 bytes

**Network buffering test: Pass!** [?]  
Pass: The network bottleneck has sufficient buffering (queue space) in routers and switches. [?]  
Measured queue size, Pkts: 36 Bytes: 52560 [?]  
This corresponds to a 48.333600 ms drain time. [?]  
To get 7 Mb/s with on a 22 ms path, you need 19250 bytes of buffer space. [?]

The network path passed all tests! [?]

**Tester validation: Pass!** [?]

Done

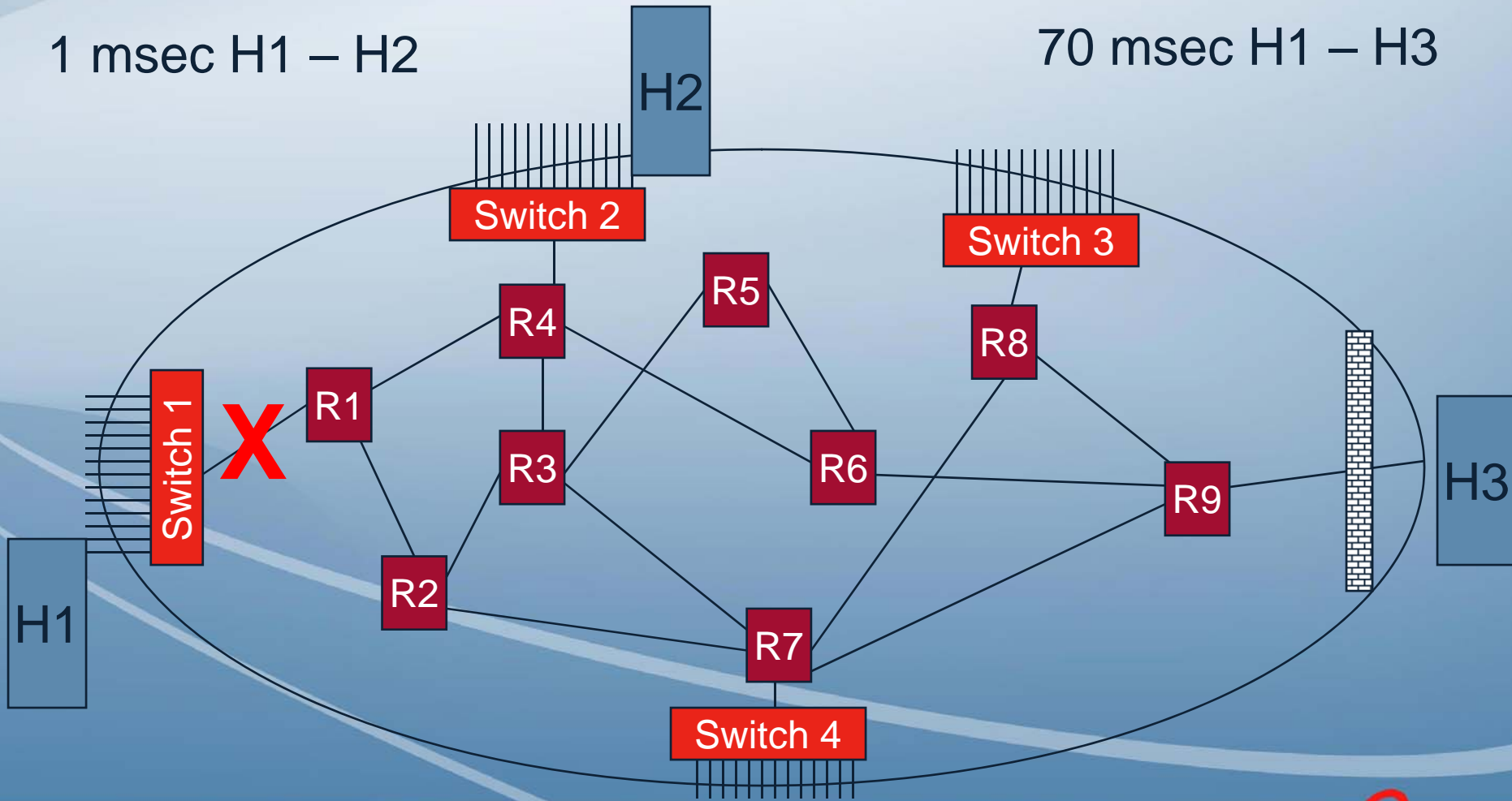
# Long Path Problem

- E2E application performance is dependant on distance between hosts
- Full size frame time at 100 Mbps
  - Frame = 1500 Bytes
  - Time = 0.12 msec
  - In flight for 1 msec RTT = 8 packets
  - In flight for 70 msec RTT = 583 packets

# Long Path Problem

1 msec H1 – H2

70 msec H1 – H3



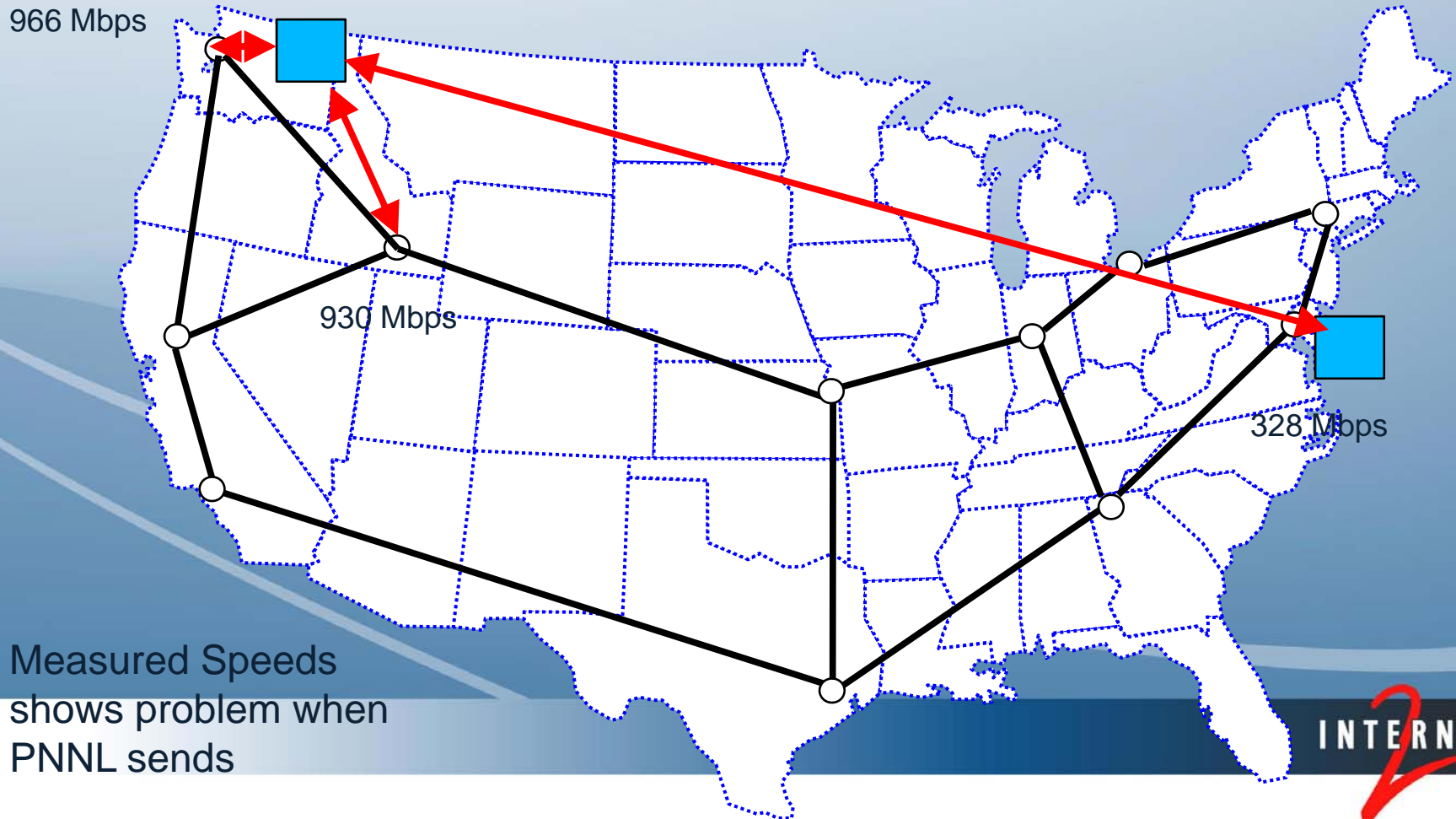
# TCP Congestion Avoidance

- Cut number of packets by  $\frac{1}{2}$
- Increase by 1 per RTT
  - LAN (RTT=1msec)
    - In flight changes to 4 packets
    - Time to increase back to 8 is 4msec
  - WAN (RTT = 70 msec)
    - In flight changes to 292 packets
    - Time to increase back to 583 is 20.4 seconds



# Example - PNNL Throughput Problem

950+ Mbps from remote sites to PNNL



# PNNL Throughput Problem

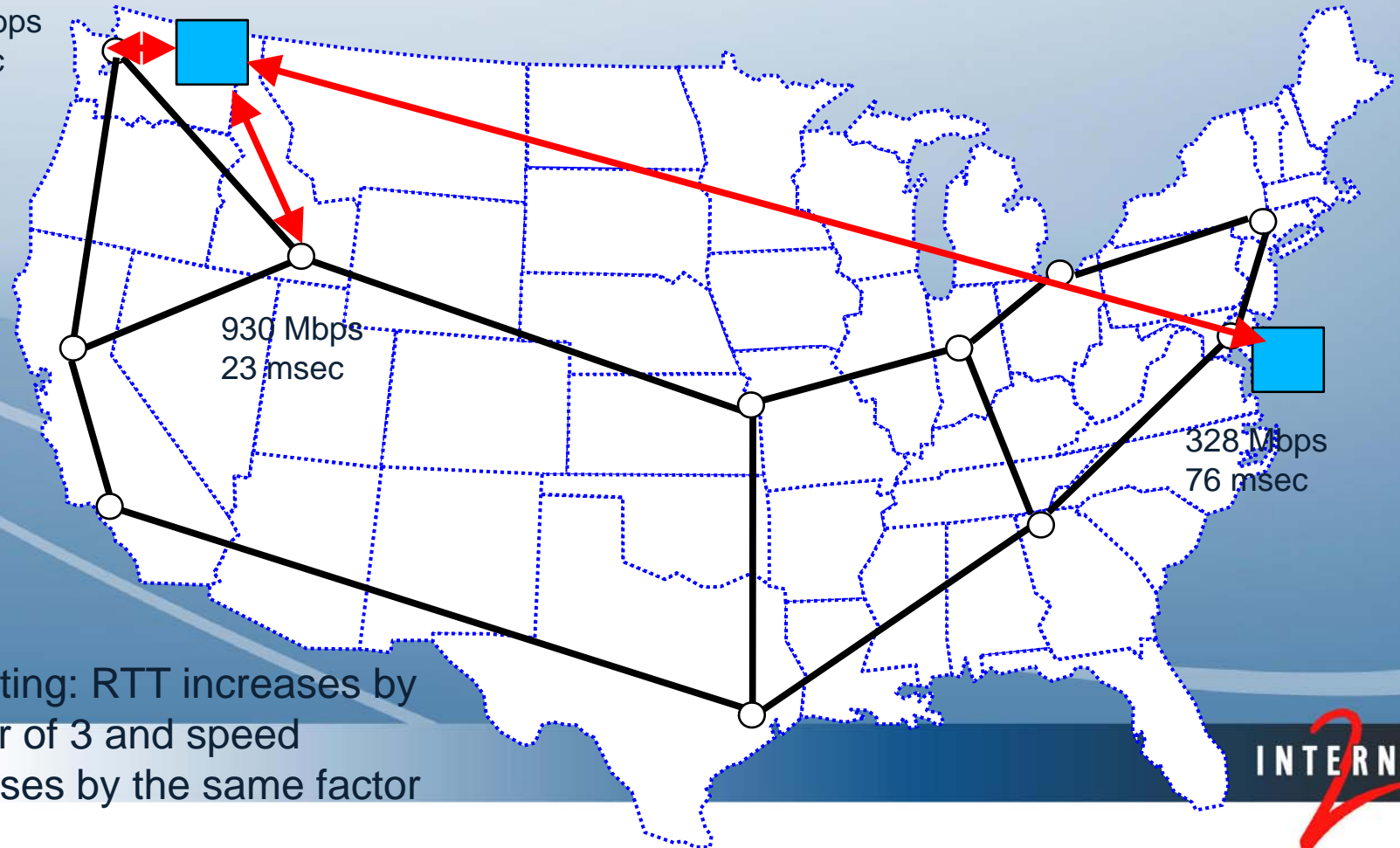
950+ Mbps from remote sites to PNNL

966 Mbps  
6 msec

930 Mbps  
23 msec

328 Mbps  
76 msec

Interesting: RTT increases by  
a factor of 3 and speed  
decreases by the same factor



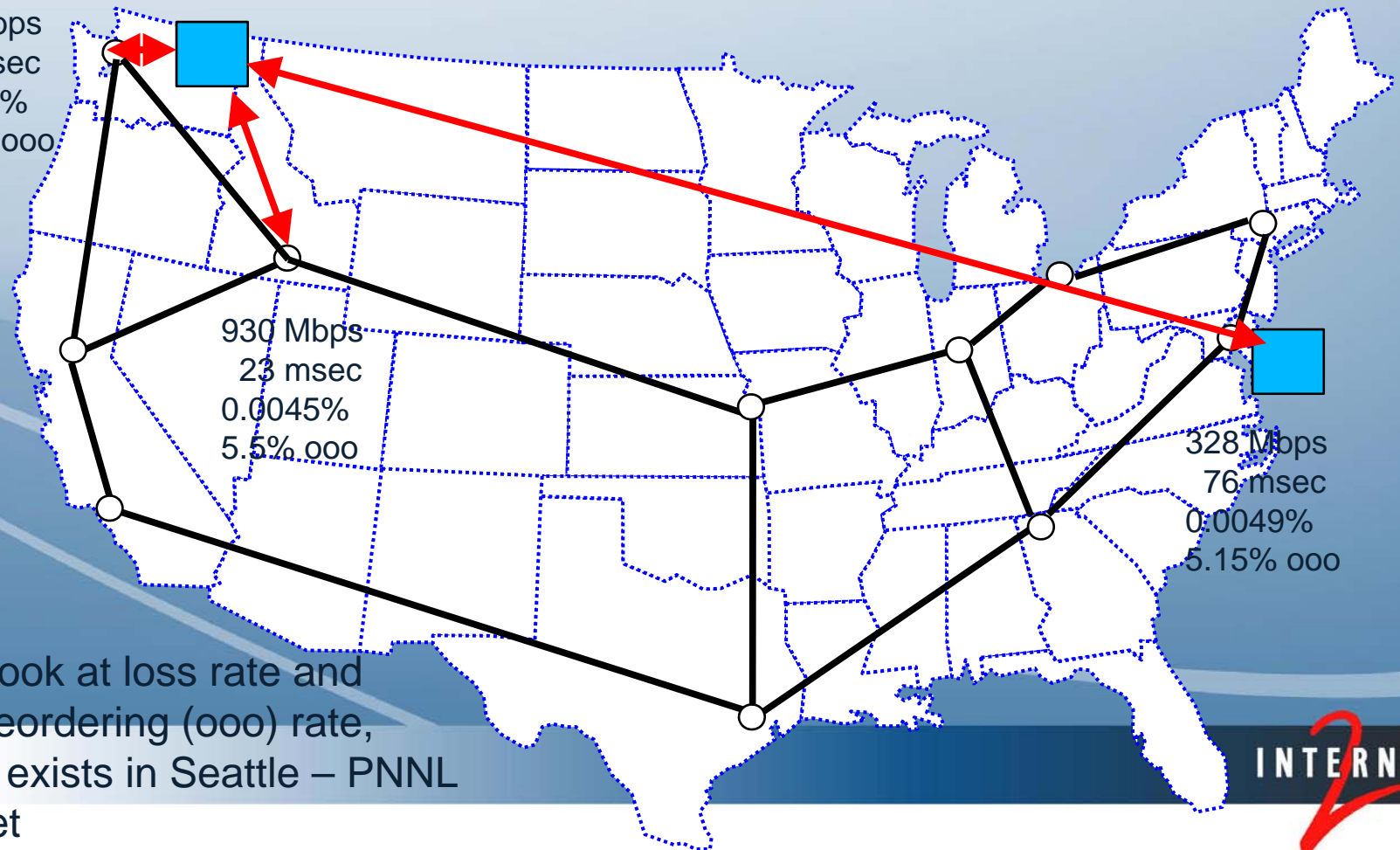
# PNNL Throughput Problem

950+ Mbps from remote sites to PNNL

966 Mbps  
6 msec  
0.0094%  
6.04% ooo

930 Mbps  
23 msec  
0.0045%  
5.5% ooo

328 Mbps  
76 msec  
0.0049%  
5.15% ooo



Finally: look at loss rate and packet reordering (ooo) rate, problem exists in Seattle – PNNL metro net

# Network Admin 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

# Under Active Development

- Emerging PerfSonar tool
  - Allows users to retrieve network path data from major national and international REN network

# PerfSonar – Next Steps in Performance Monitoring

- New Initiative involving multiple partners
  - ESnet (DOE labs)
  - GEANT (European Research and Education network)
  - Internet2 (staff and connectors)

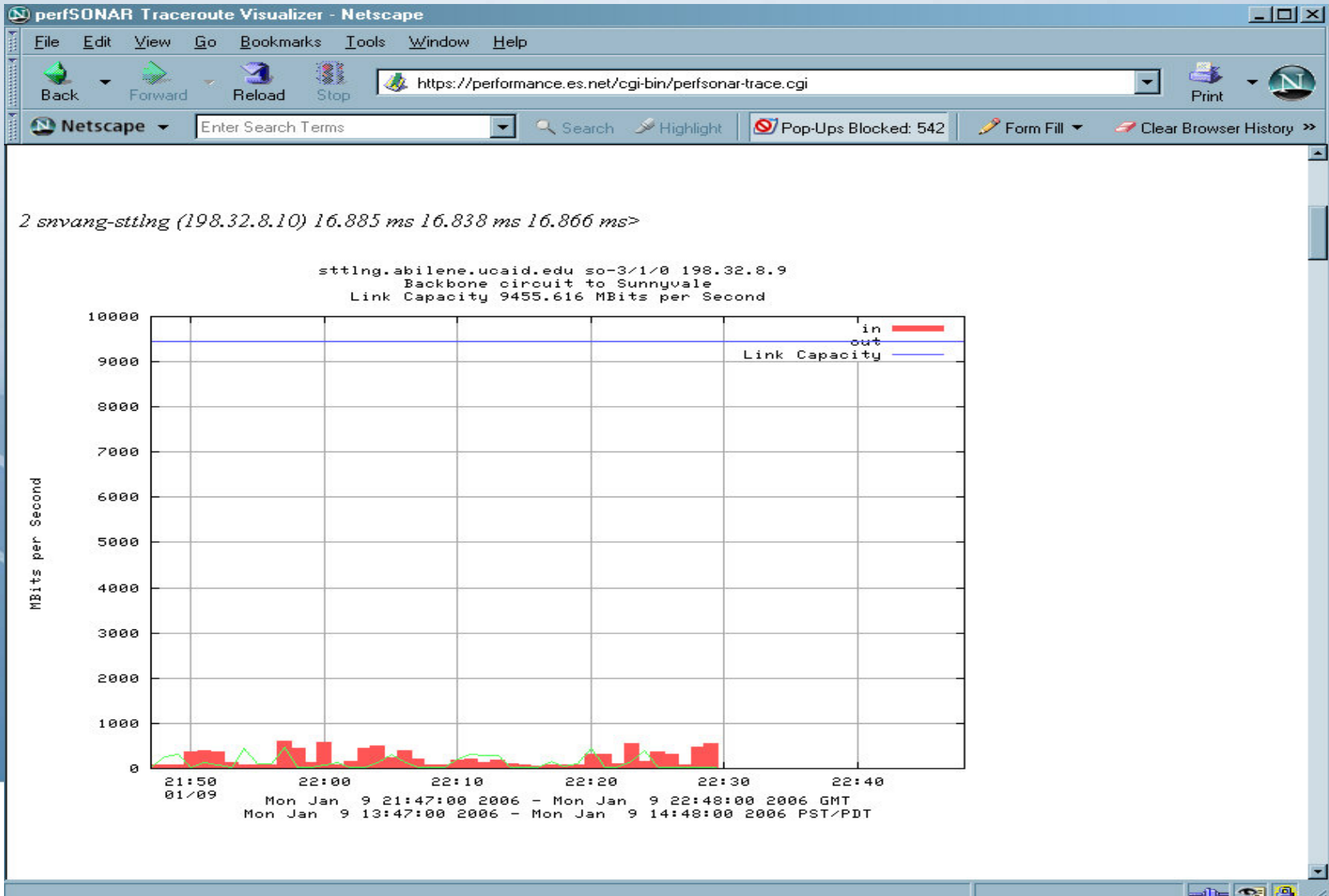
# PerfSONAR Services

- Measurement Archive (MA)
- Measurement Point (MP)
- Lookup Service (LS)
- Topology Service (TS)
- Authentication Service (AS)





# Traceroute Visualizer



# Finding a Server

- What? You don't have one running at your site?
- Install the Internet2  
Network Performance Toolkit  
Knoppix Disk

# PSC Tuning Page

**Enabling High Performance Data Transfers [PSC] - Netscape**

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop  Print

Netscape  Search Highlight Pop-Ups Blocked: 480 Form Fill Clear Browser History

## ADVANCED NETWORKING PITTSBURGH SUPERCOMPUTING CENTER

Users Partners Education Services Research News & Media Site Map | About PSC | Contacts

Home » [Advanced Networking](#) » [Research Projects](#) » Enabling High Performance Data Transfers

# Enabling High Performance Data Transfers

### System Specific Notes for System Administrators (and Privileged Users)

On this page: [Introduction](#), [Tutorial](#), [TCP Options](#), [Diagnostics](#), [Table](#), [Details](#),

**This (DRAFT) page is currently under active revision.** Please send any suggestions, additions or corrections to us at [nettime@ncne.org](mailto:nettime@ncne.org) so we can keep the information here as up-to-date as possible.


- **Introduction**
- **Tutorial**
  - Bandwidth\*Delay Products (BDP)
  - Buffers
  - Computing the BDP
- **High Performance Networking Options**
  - TCP Selective Acknowledgments (SACK, RFC2018)
  - Large Windows (RFC1323)
  - Maximum Buffer Sizes on the host

#### Advanced Networking

- [Research Projects](#)
- [Papers](#)
- [Collaborations](#)

**At the Speed of Light**  
The Three Rivers Optical Exchange connects the region to the nation's fastest networks.

#### Three Rivers Optical Exchange



PITTSBURGH  
SUPERCOMPUTING  
CENTER


# ESnet Tuning Page

Bulk Data Transfer over a WAN – Issues for Bulk Data Transfer over a WAN


http://fasterdata.es.net/ corei7 computers

Most Visited Getting Started Latest Headlines

GTNOIS... Newegg... 8 MB L3... Chicago... MLab3... perfAd... Argonn... Bulk ... Minutes... on-line ... 2009 A... AOPA O...

 U.S. DEPARTMENT OF **ENERGY** | Office of Science

## Guide to Bulk Data Transfer over a WAN



Version 1.0 - Last published May 18, 2009

<b>Bulk Data Transfer Tools</b>	<b>Issues for Bulk Data Transfer over a WAN</b>
<ul style="list-style-type: none"><li>Background</li><li>Throughput Requirements<ul style="list-style-type: none"><li>-- Bandwidth Chart</li></ul></li><li>Host Tuning Overview</li><li>Expected Throughput</li><li>File Transfer Tools</li><li>GridFTP Quick Start</li><li>Firewall Issues</li><li>Summary</li></ul>	<p>Many people think that data sets of 1 TeraByte are just too big to move across the WAN, and resort to sending DVDs or USB drives. This is no longer true. <b>Moving a TeraByte between most large research institutions in the US should only take around 8 hours.</b> This assumes an end-to-end path with a capacity of 1 Gbps or higher, and that only 1/3 of the capacity is used, leaving room for other users traffic. <a href="#">This chart</a> shows the bandwidth requirements for various data set sizes and times.</p> <p>If your network throughput is less than this, chances are that your hosts need tuning and/or you are using the wrong file transfer tools. The purpose of this site is to help you maximize your wide-area network bulk data transfer performance by tuning the TCP settings for your end hosts and by using file transfer tools that are designed to maximize network throughput.</p> <p>Historically, wide-area bulk data transfer has been plagued by poor performance for a variety of reasons. These include improper configuration of the sending and receiving hosts, software design issues, firewalls, and other factors. In most cases, however, large data sets can be moved long distances using today's networks with minimal effort.</p> <p>Most file transfer programs use the TCP protocol, and performance problems are often due to a <a href="#">TCP window</a> that is too small. The maximum congestion window is related to the amount of buffer space that the kernel allocates for each socket, and most operating systems by default limit this buffer space to a value that is too small for today's high-speed networks. To achieve maximum throughput, it is critical to use optimal TCP socket buffer sizes for the link you are using. This means you must use a file transfer tool that allows you to set the TCP buffer size, and that your end systems must be tuned to allow for large TCP socket buffers.</p> <p>Another common technique to speed up file transfers is to break the file into smaller pieces that are transferred in parallel. A number of tools include the option to do parallel transfers. If you have a large number of files to copy, you can do parallel transfers by copying several files at once (typically 4-5 is a good number to try). But in general it is more efficient to copy larger files than smaller files, so bundling multiple small files into a single larger file using tar or zip is also recommended.</p> <p><b>Following these steps will help ensure you are getting the best throughput possible.</b></p>
<b>Network Troubleshooting</b>	
<ul style="list-style-type: none"><li>Overview</li><li>Active perfSONAR Services</li><li>perfSONAR HowTo</li><li>Sample Network Issues</li></ul>	
<b>More Info</b>	
<ul style="list-style-type: none"><li>Supercomputer Center Data Transfers</li><li>Specific DOE Sites</li><li>TCP Tuning Details</li><li>News</li><li>Links</li></ul>	
<b>Talks</b>	
<ul style="list-style-type: none"><li>Short Tutorial</li></ul>	
<b>For Network Engineers</b>	
<ul style="list-style-type: none"><li>Network Design Hints</li><li>Cisco Config Hints</li><li>Forge10 Config Hints</li></ul>	

Done

# 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