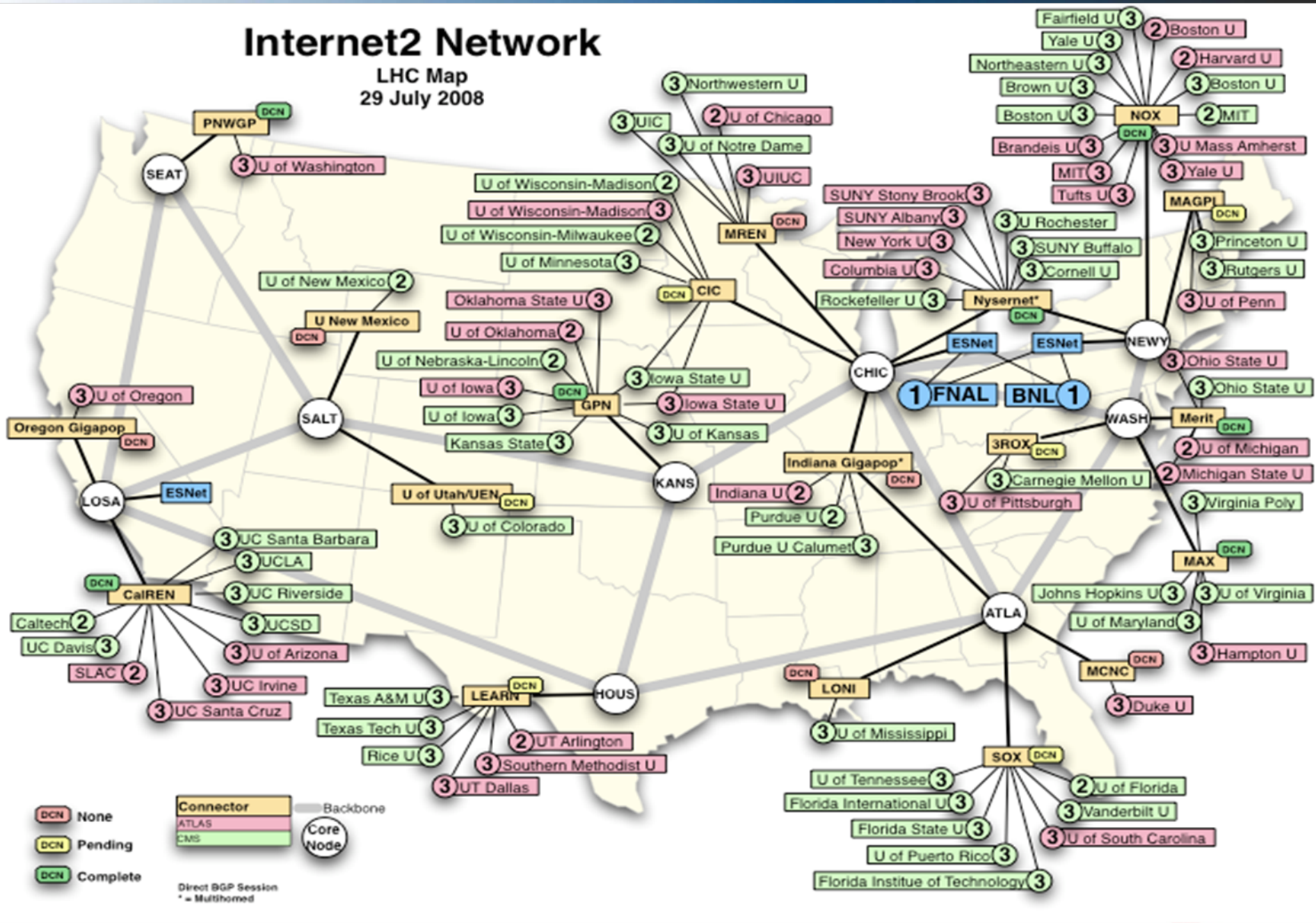
Tier3 Network Issues

Richard Carlson
May 19, 2009
rcarlson@internet2.edu


Internet2 Network
LHC Map
29 July 2008
PNWGP
SEAT
U of Washington
U of Oregon
Oregon Gigapop
LOSA
ESNet
UC Santa Barbara
UCLA
CalREN
UC Riverside
Caltech
UCSD
UC Davis
U of Arizona
SLAC
UC Irvine
UC Santa Cruz
U of New Mexico
U New Mexico
SALT
U of Utah/UEN
U of Colorado
Oklahoma State U
U of Oklahoma
U of Nebraska-Lincoln
U of Iowa
U of Iowa
Kansas State
GPN
Iowa State U
Iowa State U
U of Kansas
KANS
Northwestern U
U of Chicago
UIC
U of Notre Dame
UIUC
MREN
U of Wisconsin-Madison
U of Wisconsin-Madison
U of Wisconsin-Milwaukee
U of Minnesota
CIC
CHIC
Indiana Gigapop*
Indiana U
Purdue U
Purdue U Calumet
Fairfield U
Yale U
Northeastern U
Brown U
Boston U
Brandeis U
MIT
Tufts U
Boston U
Harvard U
Boston U
MIT
NOX
U Mass Amherst
Yale U
SUNY Stony Brook
SUNY Albany
New York U
Columbia U
Rockefeller U
U Rochester
SUNY Buffalo
Cornell U
Nysernet*
MAGPI
Princeton U
Rutgers U
U of Penn
ESNet
ESNet
NEWY
Ohio State U
Ohio State U
1 FNAL
BNL 1
WASH
Merit
U of Michigan
Michigan State U
Virginia Poly
3ROX
Carnegie Mellon U
U of Pittsburgh
MAX
Johns Hopkins U
U of Virginia
U of Maryland
Hampton U
ATLA
MCNC
Duke U
LONI
U of Mississippi
SOX
U of Tennessee
Florida International U
Florida State U
U of Puerto Rico
Florida Institue of Technology
U of Florida
Vanderbilt U
U of South Carolina
HOUS
LEARN
Texas A&M U
Texas Tech U
Rice U
UT Arlington
Southern Methodist U
UT Dallas
None
Pending
Complete
Connector
ATLAS
CMS
Backbone
Core Node
Direct BGP Session
* = Multihomed

Internet2 overview

- Member organization with a national backbone infrastructure
  - Campus & Regional network members
  - National and International peers
- Tiered connection model
  - Campus ←→ Regional ←→ 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!

INTERNET®

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

INTERNET®

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

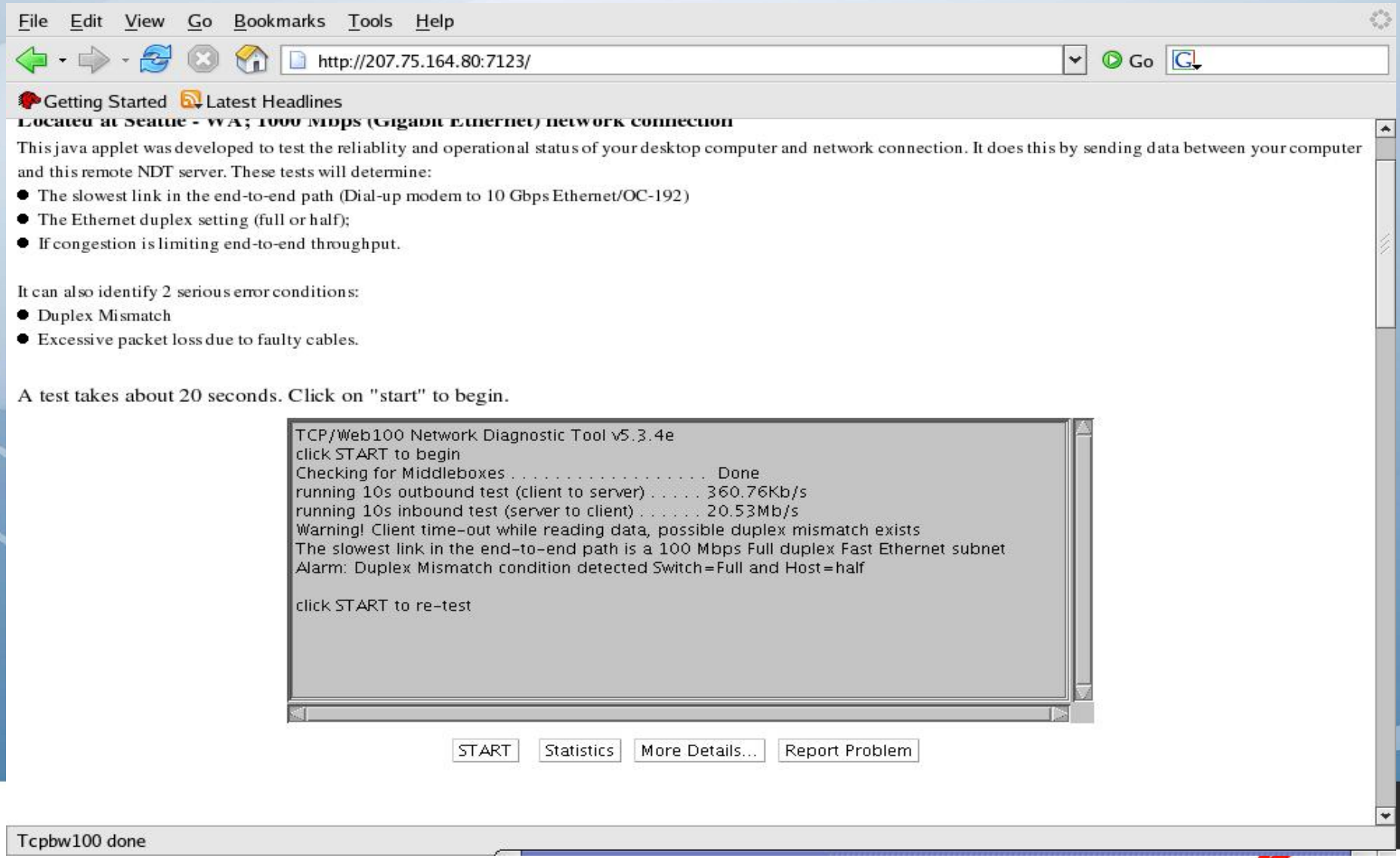
NPAD/pathdiag

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

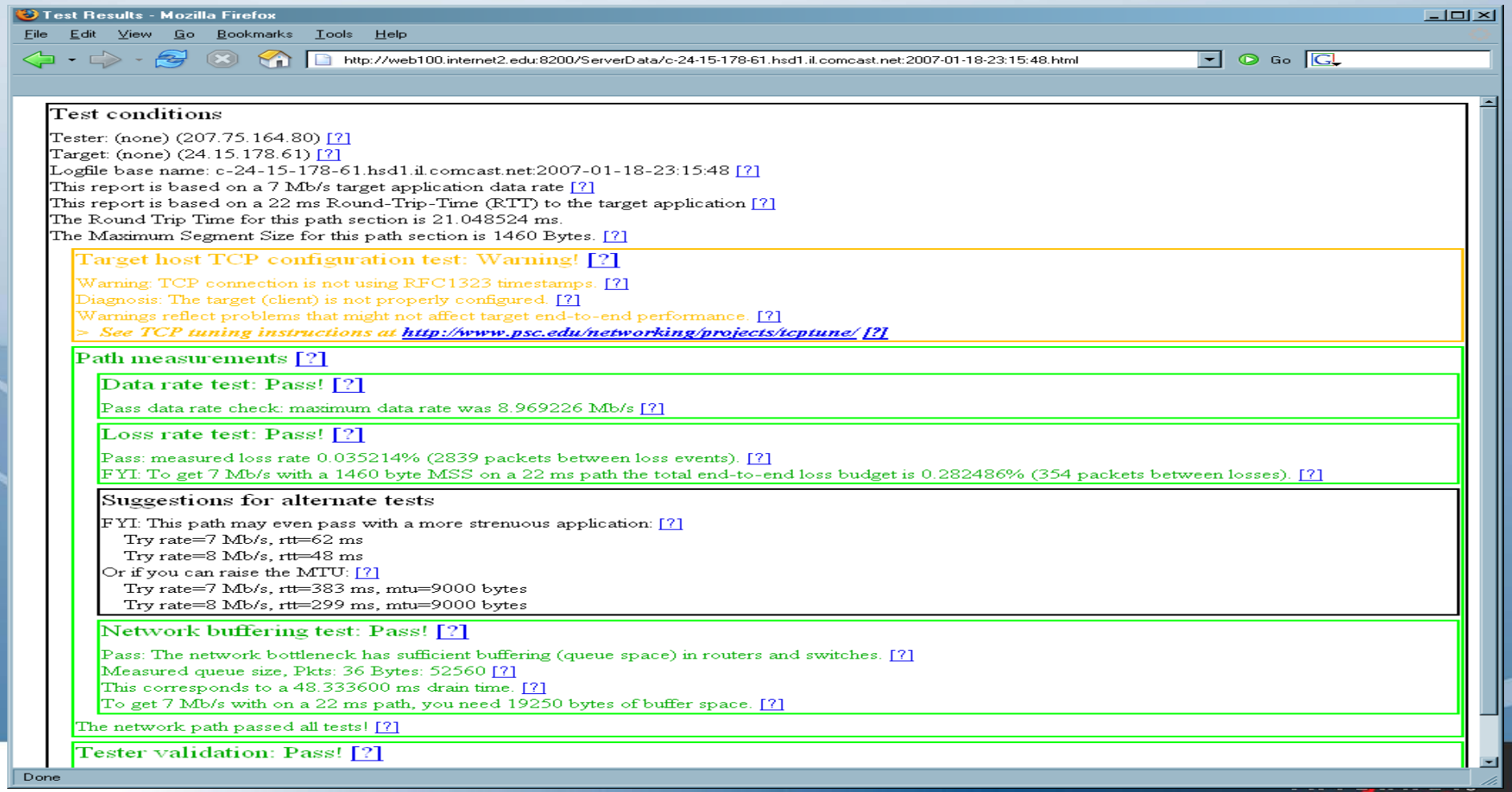
NDT/NPAD user interface

- 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



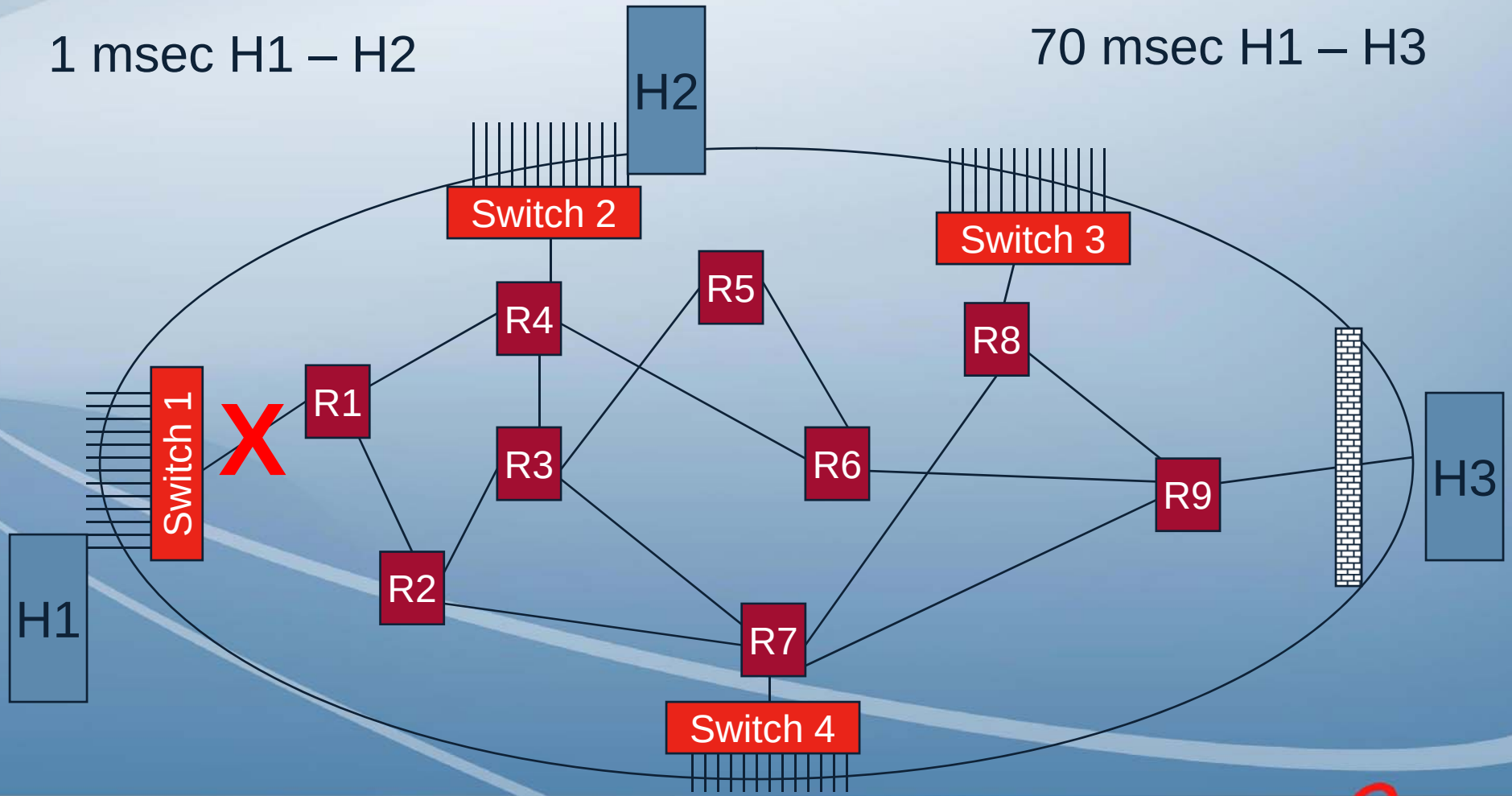
NPAD Sample results



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



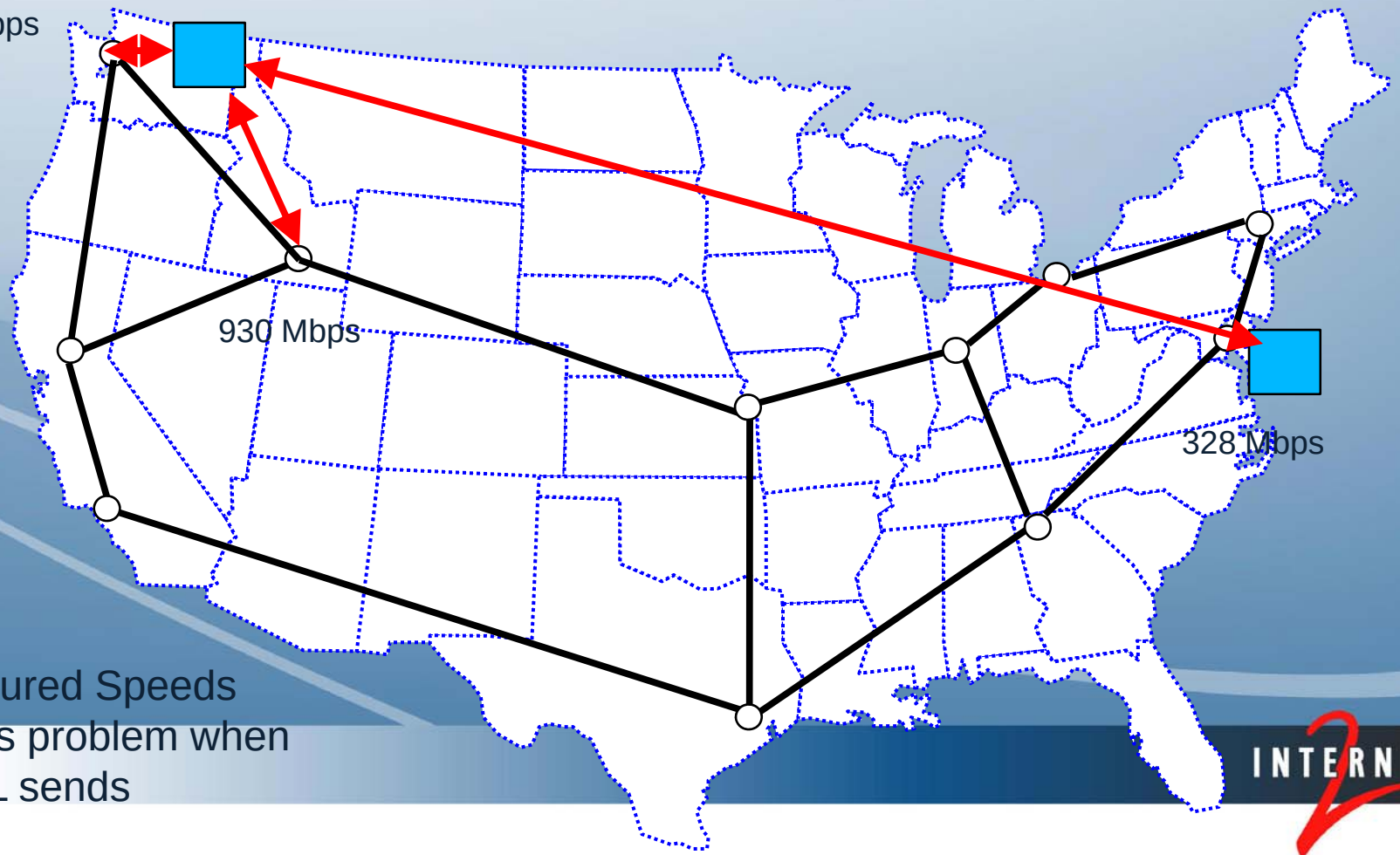
INTERNET2®

TCP Congestion Avoidance

- Cut number of packets by ½
- 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

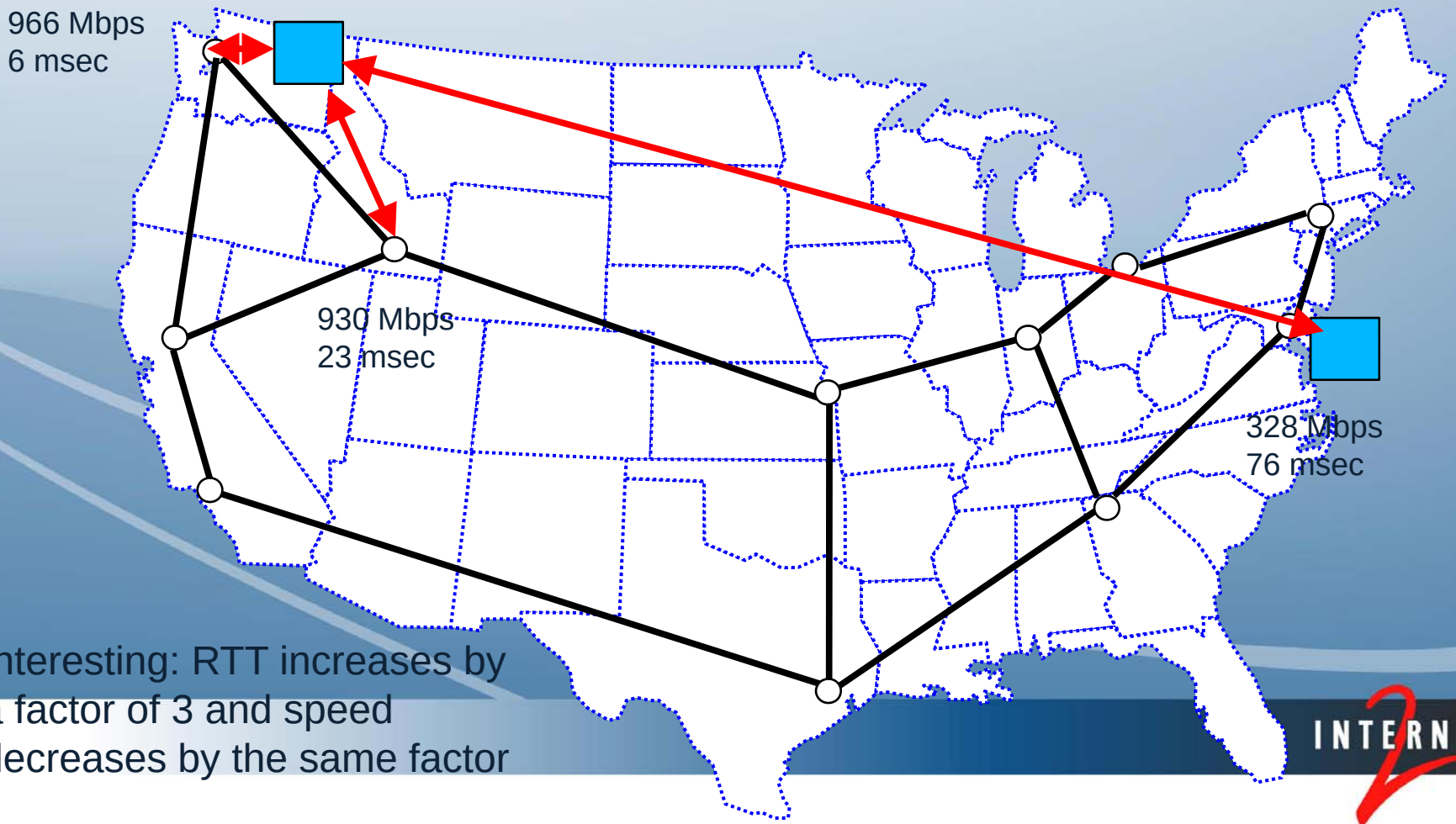
950+ Mbps from remote sites to PNNL



Measured Speeds shows problem when PNNL sends

PNNL Throughput Problem

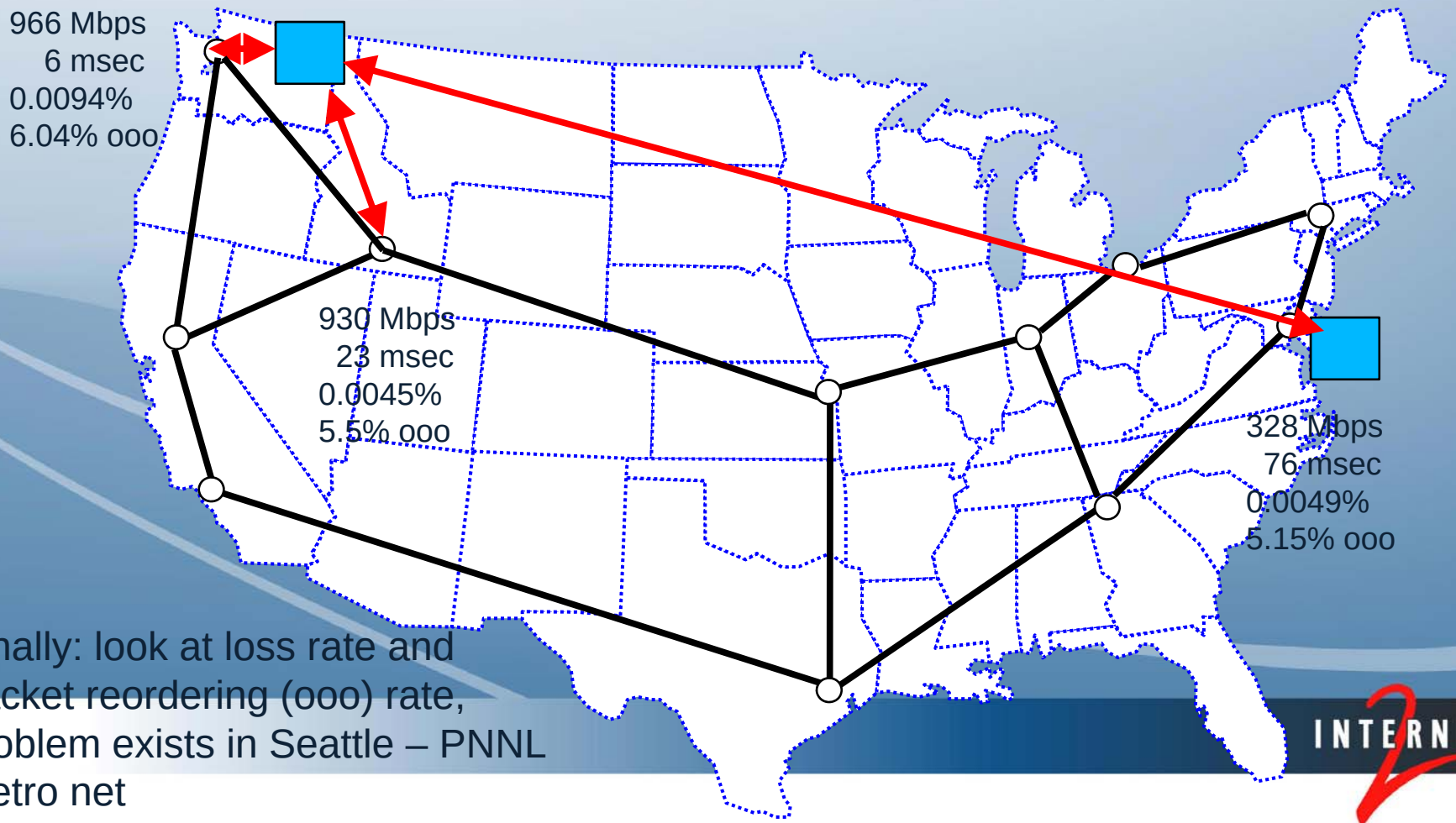
950+ Mbps from remote sites to PNNL



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



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

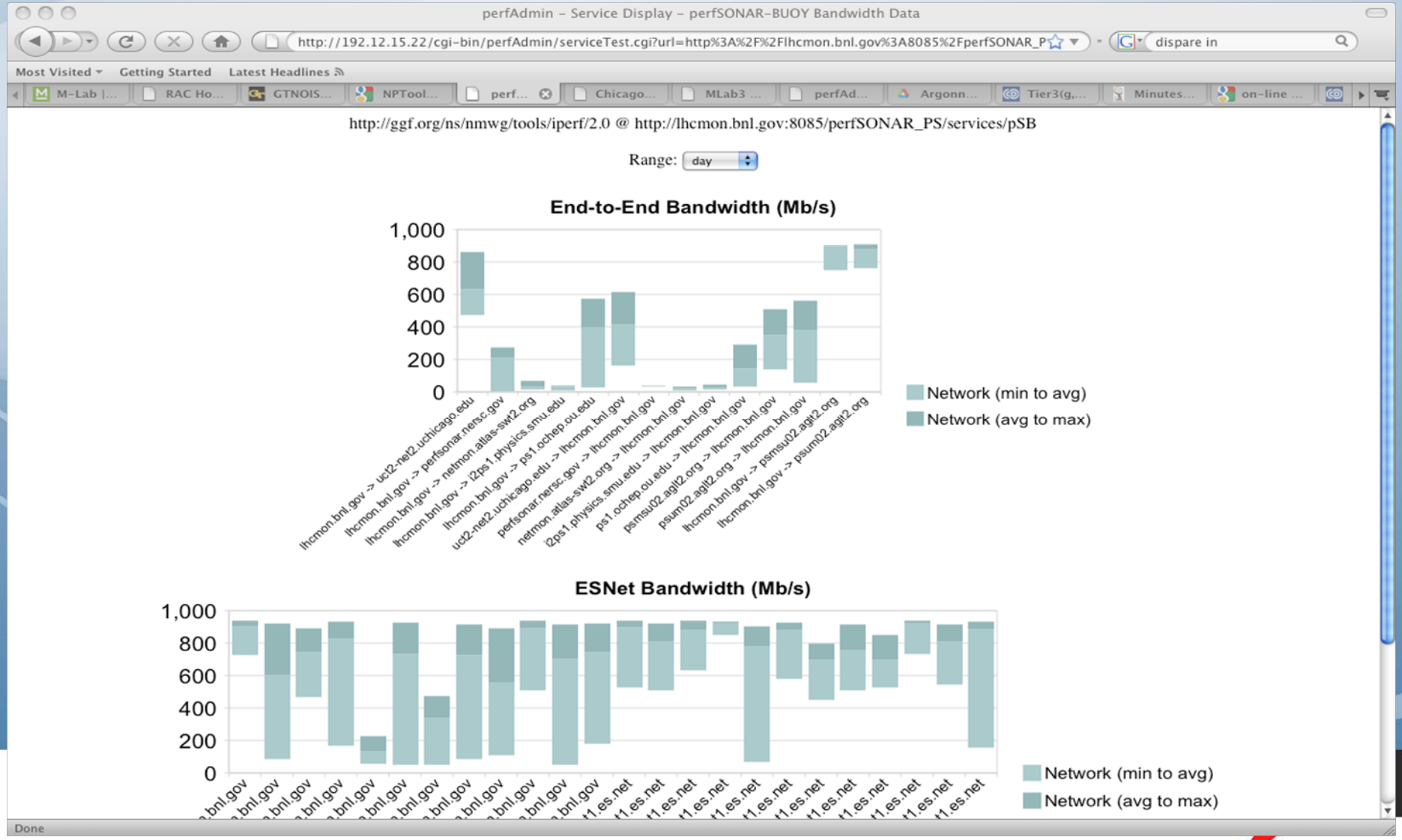
PerfSonar – Next Steps in Performance Monitoring

- 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)

USATLAS Throughput Monitoring



Traceroute Visualizer

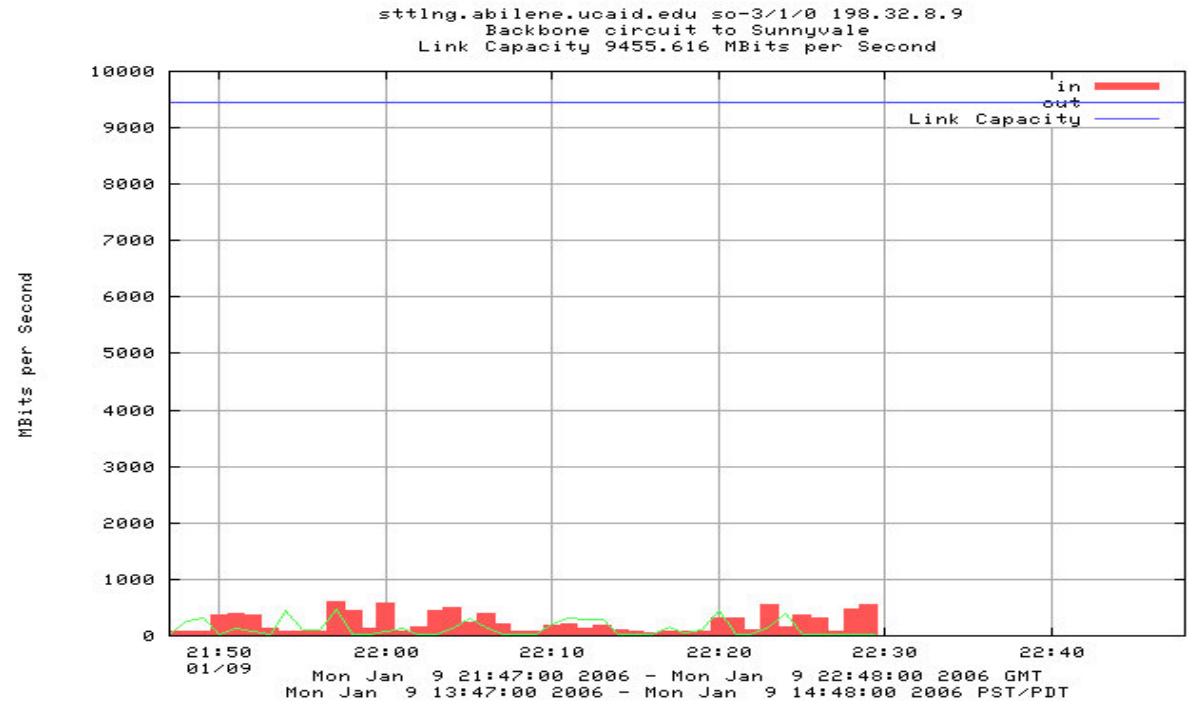
perfSONAR Traceroute Visualizer - Netscape

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop https://performance.es.net/cgi-bin/perfsonar-trace.cgi Print

Netscape Enter Search Terms Search Highlight Pop-Ups Blocked: 542 Form Fill Clear Browser History

2 snvang-sttlng (198.32.8.10) 16.885 ms 16.838 ms 16.866 ms>

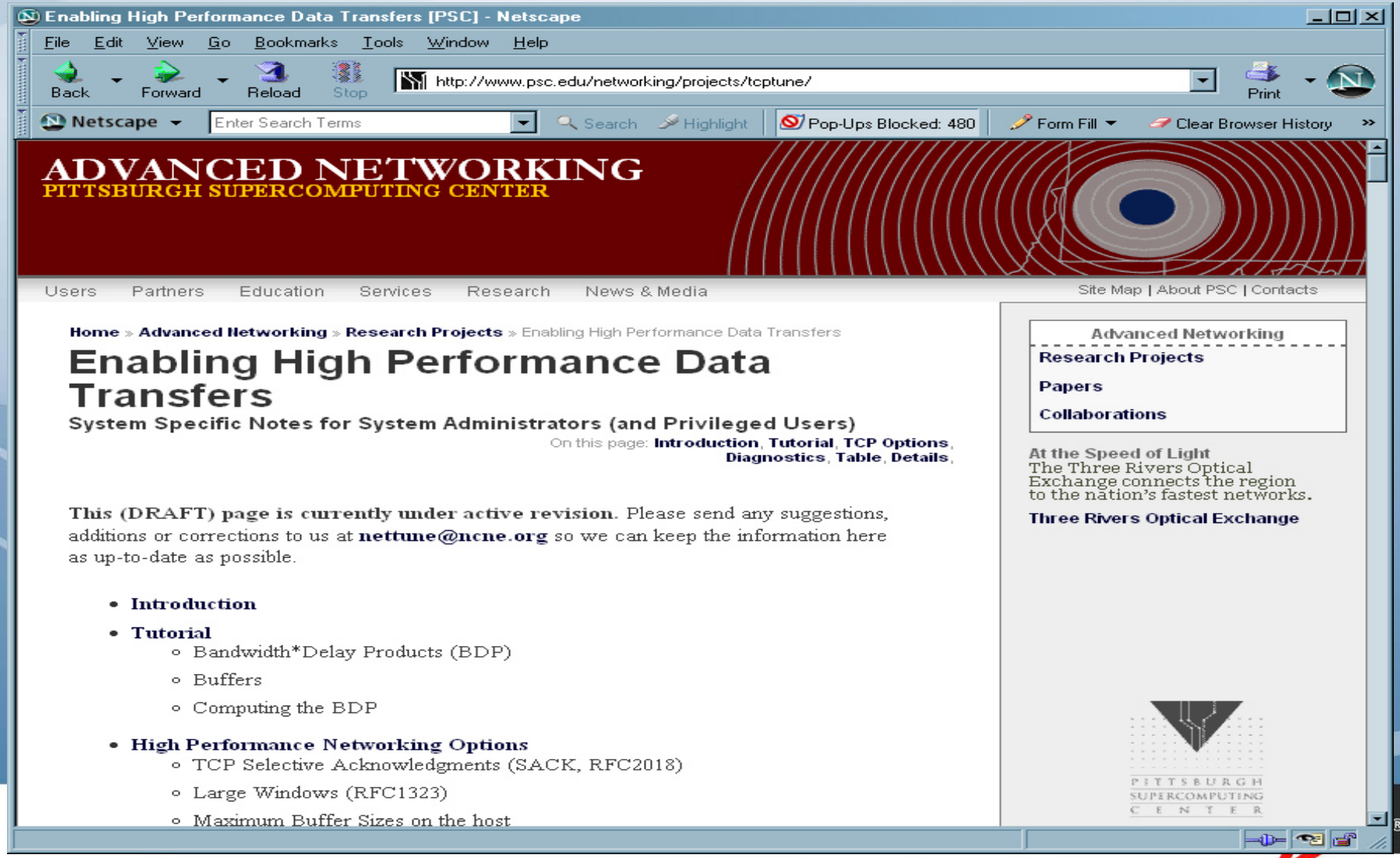


Finding a Server

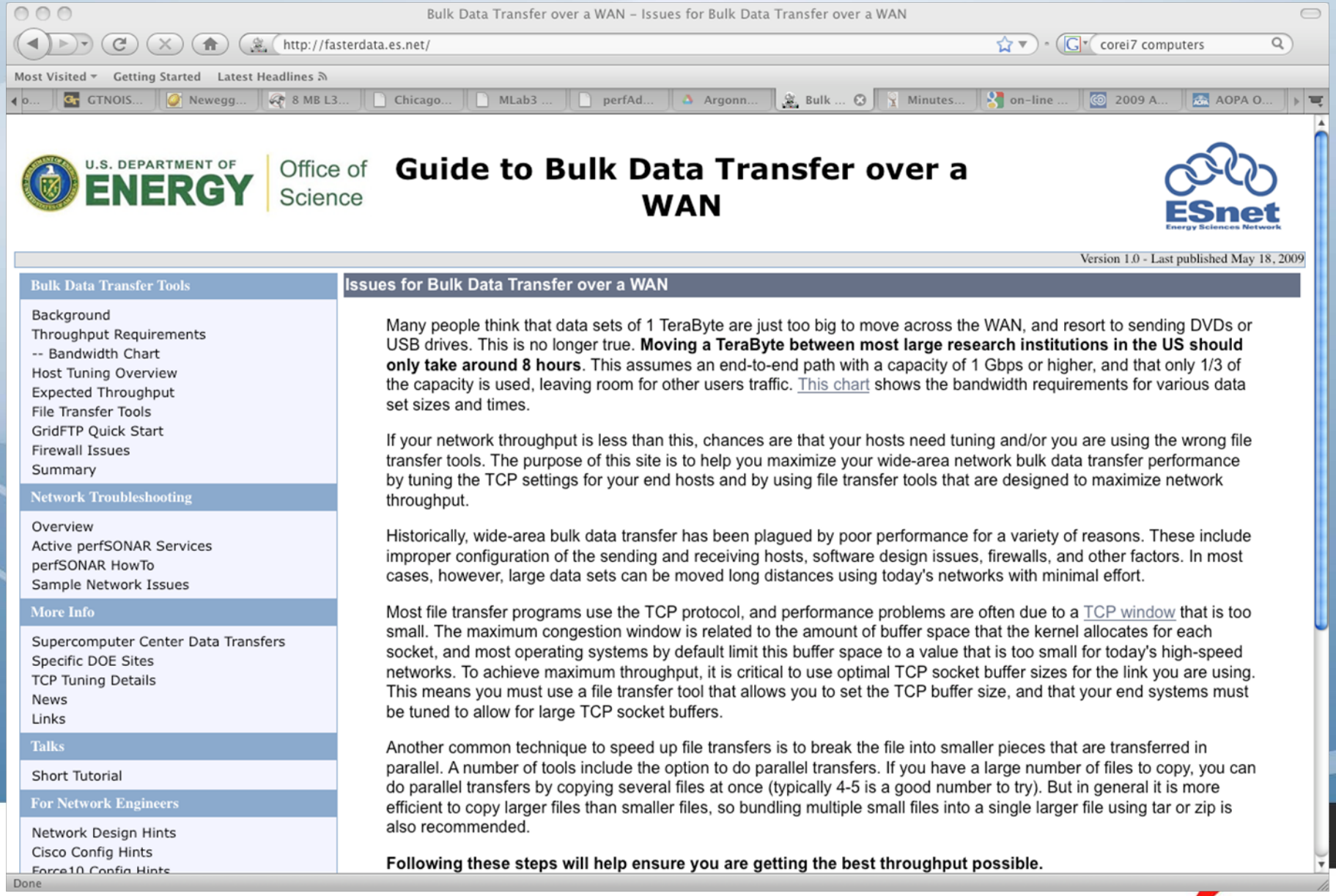
- What? You don't have one running at your site?

- Install the Internet2 `Network Performance Toolkit` Knoppix Disk

PSC Tuning Page



ESnet Tuning Page



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