August 10th 2010, OSG Site Admin Workshop - Network Performance

Jason Zurawski, Internet2

# NDT

# Agenda

- Tutorial Agenda:
  - Network Performance Primer - Why Should We Care? (**15 Mins**)
  - Getting the Tools (**10 Mins**)
  - Use of the BWCTL Server and Client (**30 Mins**)
  - Use of the OWAMP Server and Client (**30 Mins**)
  - Use of the NDT Server and Client (**30 Mins**)
  - BREAK (**15 mins**)
  - Diagnostics vs Regular Monitoring (**30 Mins**)
  - Network Performance Exercises (**1 hr 30 Mins**)

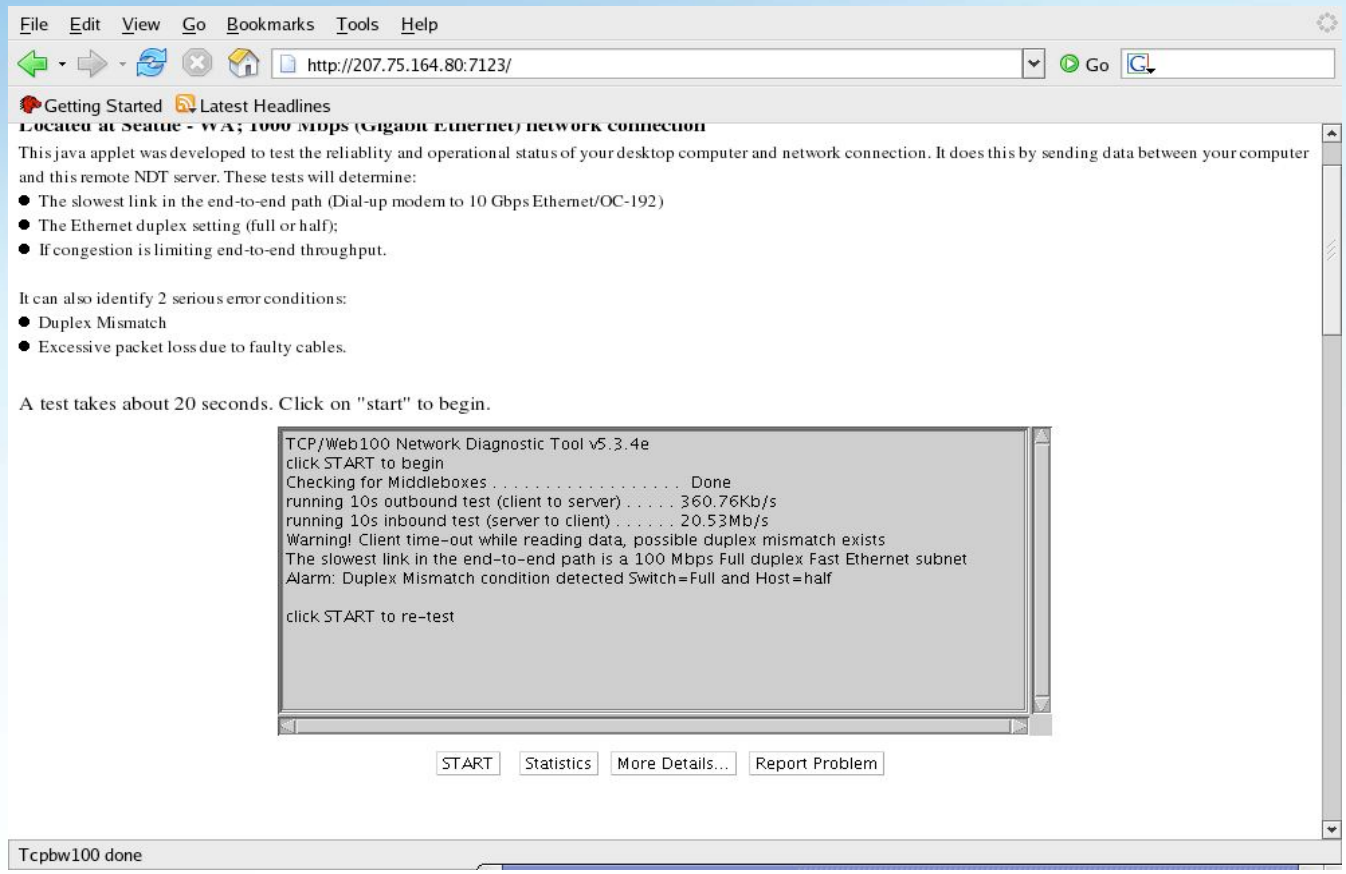
# Live Test

- MLab (Commodity Networking)
  - http://ndt.iupui.donar.measurement-lab.org:7123/
- Internet2 (R&E Networking)
  - http://ndt.atla.net.internet2.edu:7123/

# NDT User Interface

- Web-based JAVA applet allows testing from any browser
  - One Click testing
  - Option to dig deep into available results
  - Send report of results to network administrators
- Command-line client allows testing from remote login shell
  - Same options available
  - Client software can be build independent of server software

# NDT Results


File Edit View Go Bookmarks Tools Help
http://207.75.164.80:7123/
Go
Getting Started
Latest Headlines
This java applet was developed to test the reliablity 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
Tcpbw100 done

# Motivation for Work

- Measure performance ***to users desktop***
  - Lots of tools to measure performance to a nearby server
  - Also 'plugable' hardware to measure everything up to the network cable
  - Want something to accurately show ***what the user is seeing***
- Develop "single shot" diagnostic tool that doesn't use historical data
- Combine numerous Web100 variables to analyze connection
- Develop network signatures for 'typical' network problems
  - Based on heuristics and experience
  - Lots of problems have a ***smoking gun*** pattern, e.g. duplex mismatch, bad cable, etc.

# How It works

- Simple bi-directional test to gather end to end data
  - Test from client to server, and the reverse
  - Gets the 'upload' and 'download' directions
- Gather multiple data variables from server
  - Via Web100, also some derived metrics (packet inter arrival times)
- Compare measured performance to analytical values
  - How fast ***should*** a connection be given the observations of the host and network
- Translate network values into plain text messages
- Geared toward campus area network

# Web100 Project

- Joint PSC/NCAR project funded by NSF
- Develop a ***system mib***, similar to data that is exposed via SNMP
- 'First step' to gather TCP data
  - Kernel Instrument Set (KIS)
- Requires patched Linux kernel
- Geared toward wide area network performance
- Goal is to automate tuning to improve application performance
- Patches available for ***vanilla*** kernels (e.g. non vendor modified)

# Web Based Performance Tool

- Operates on Any client with a Java enabled Web browser
  - No additional client software needs to be installed
  - No additional configuration required

- What it can do:
  - State if Sender, Receiver, or Network is operating properly
  - Provide accurate application tuning info
  - Suggest changes to improve performance
- What it can't do
  - Tell you where in the network the problem is
  - Tell you how other servers perform
  - Tell you how other clients will perform

# 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
- 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

# NDT Testing – Normal Operation

home-ndt - SecureCRT

File Edit View Options Transfer Script Tools Window Help

```
[root@home-ndt rcarlson]#
[root@home-ndt rcarlson]#
[root@home-ndt rcarlson]#
[root@home-ndt rcarlson]#
[root@home-ndt rcarlson]#
[root@home-ndt rcarlson]#
[root@home-ndt rcarlson]#
[root@home-ndt rcarlson]#
[root@home-ndt rcarlson]#
[root@home-ndt rcarlson]# web100clt -l triton
Testing network path for configuration and performance problems
Checking for Middleboxes . . . . . Done
running 10s outbound test (client to server) . . . . . 86.29 Mb/s
running 10s inbound test (server to client) . . . . . 94.31 Mb/s
The slowest link in the end-to-end path is a 100 Mbps Full duplex Fast Ethernet subnet

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

Web100 reports the Round trip time = 11.06 msec;the Packet size = 1448 Bytes; and
No packet loss was observed.
This connection is receiver limited 99.76% 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=7, Client=2
The theoretical network limit is 998.70 Mbps
The NDT server has a 8192 KByte buffer which limits the throughput to 5785.57 Mbps
Your PC/Workstation has a 128 KByte buffer which limits the throughput to 90.40 Mbps
The network based flow control limits the throughput to 91.88 Mbps

Client Data reports link is ´  5´, Client Acks report link is ´  5´
Server Data reports link is ´  5´, Server Acks report link is ´  5´
Packet size is preserved End-to-End
Server IP addresses are preserved End-to-End
Client IP addresses are preserved End-to-End
[root@home-ndt rcarlson]#
```

Ready | ssh2: AES-128 | 39, 27 | 39 Rows, 105 Cols | VT100

perfSONAR powered

TERNET 2

# NDT Testing – Duplex Mismatch

home-ndt - SecureCRT

File Edit View Options Transfer Script Tools Window Help

Keymap Editor

```
[root@home-ndt rcarlson]#
[root@home-ndt rcarlson]# web100clt -l triton
Testing network path for configuration and performance problems
Checking for Middleboxes . . . . . . . . . . . . . . . . . . . . Done
running 10s outbound test (client to server) . . . . . 68.31 Mb/s
running 10s inbound test (server to client) . . . . . 278.98 kb/s
The slowest link in the end-to-end path is a 100 Mbps Half duplex Fast Ethernet subnet
Alarm: Duplex mismatch condition exists: Host set to Full and Switch set to Half duplex

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

Web100 reports the Round trip time = 0.00 msec;the Packet size = 1448 Bytes; and
There were 104 packets retransmitted, 104 duplicate acks received, and 156 SACK blocks received
Packets arrived out-of-order 40.00% of the time.
The connection stalled 51 times due to packet loss.
The connection was idle 10.71 seconds (107.10%) of the time.
This connection is network limited 99.96% of the time.
  Contact your local network administrator to report a network problem
  Contact your local network admin and report excessive packet reordering
Excessive packet loss is impacting your performance, check the auto-negotiate function on your local PC a
nd network switch

    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=7, Client=2
The theoretical network limit is 0.00 Mbps
The NDT server has a 8192 KByte buffer which limits the throughput to inf Mbps
Your PC/Workstation has a 63 KByte buffer which limits the throughput to inf Mbps
The network based flow control limits the throughput to inf Mbps

Client Data reports link is ´  5´, Client Acks report link is ´  5´
Server Data reports link is ´  4´, Server Acks report link is ´  4´
Packet size is preserved End-to-End
Server IP addresses are preserved End-to-End
Client IP addresses are preserved End-to-End
[root@home-ndt rcarlson]#
```

Display keymap editor dialog | ssh2: AES-128 | 39, 27 | 39 Rows, 105 Cols | VT100

# NDT Testing – Low Throughput

nmsx.internet2 - SecureCRT

File Edit View Options Transfer Script Tools Window Help

```
[rcarlson@nmsx-aami rcarlson]$
[rcarlson@nmsx-aami rcarlson]$
[rcarlson@nmsx-aami rcarlson]$
[rcarlson@nmsx-aami rcarlson]$
[rcarlson@nmsx-aami rcarlson]$ web100clt -l ndt-newyork
Testing network path for configuration and performance problems
Checking for Middleboxes . . . . . Done
running 10s outbound test (client to server) . . . . . 88.23 Mb/s
running 10s inbound test (server to client) . . . . . 13.78 Mb/s
The slowest link in the end-to-end path is a 100 Mbps Full duplex Fast Ethernet subnet
Information: The receive buffer should be 444 Kbytes to maximize throughput

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

Web100 reports the Round trip time = 36.35 msec;the Packet size = 1448 Bytes; and
No packet loss was observed.
This connection is receiver limited 97.49% of the time.
  Increasing the current receive buffer (62.50 KB) will improve performance
This connection is network limited 2.46% 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=9, Client=7
The theoretical network limit is 303.89 Mbps
The NDT server has a 32768 KByte buffer which limits the throughput to 7042.06 Mbps
Your PC/Workstation has a 62 KByte buffer which limits the throughput to 13.43 Mbps
The network based flow control limits the throughput to 13.67 Mbps

Client Data reports link is ´  5´, Client Acks report link is ´  5´
Server Data reports link is ´  8´, Server Acks report link is ´  4´
Packet size is preserved End-to-End
Server IP addresses are preserved End-to-End
Client IP addresses are preserved End-to-End
[rcarlson@nmsx-aami rcarlson]$
```

Ready | ssh2: AES-128 | 37, 32 | 37 Rows, 115 Cols | VT100

# NDT Testing – Increase TCP Buffer Size

nmsx.internet2 - SecureCRT

File Edit View Options Transfer Script Tools Window Help

```
[rcarlson@nmsx-aami rcarlson]$
[rcarlson@nmsx-aami rcarlson]$
[rcarlson@nmsx-aami rcarlson]$
[rcarlson@nmsx-aami rcarlson]$ web100clt -l -b2097152 ndt-newyork
Testing network path for configuration and performance problems
Checking for Middleboxes . . . . . Done
running 10s outbound test (client to server) . . . . . 87.57 Mb/s
running 10s inbound test (server to client) . . . . . 84.40 Mb/s
The slowest link in the end-to-end path is a 100 Mbps Half duplex Fast Ethernet subnet
Information: Other network traffic is congesting the link

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

Web100 reports the Round trip time = 38.80 msec;the Packet size = 1448 Bytes; and
There were 585 packets retransmitted, 769 duplicate acks received, and 1354 SACK blocks received
Packets arrived out-of-order 3.87% of the time.
This connection is receiver limited 2.68% of the time.
This connection is network limited 97.28% of the time.
  Contact your local network administrator to report a network problem

    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=9, Client=7
The theoretical network limit is 52.98 Mbps
The NDT server has a 32768 KByte buffer which limits the throughput to 6598.79 Mbps
Your PC/Workstation has a 3070 KByte buffer which limits the throughput to 618.33 Mbps
The network based flow control limits the throughput to 257.42 Mbps

Client Data reports link is ´  5´, Client Acks report link is ´  5´
Server Data reports link is ´  8´, Server Acks report link is ´  5´
Packet size is preserved End-to-End
Server IP addresses are preserved End-to-End
Client IP addresses are preserved End-to-End
[rcarlson@nmsx-aami rcarlson]$
```

Ready | ssh2: AES-128 | 37, 32 | 37 Rows, 115 Cols | VT100

# Bottleneck Link Detection

- What is the slowest link in the end-to-end path?
  - Monitors packet arrival times using libpcap routine
    - Data and ACK packets
    - Is aware of packet sizes – used to calculate speed
  - Use TCP dynamics to create packet pairs
  - Quantize results into link type bins
    - Broad classification, e.g. "FastE"
    - No fractional or bonded links currently
- Example:
  - Consider the following setup
    - 1G network card on Host
    - 1G LAN
    - 100M (FastE) Wall Jack
  - NDT will report there is a slow link somewhere in the path. It can't tell you where, but something is limiting the test speed

# Duplex Mismatch Detection

- Duplex Mismatch:
  - Operation between a host and an interface are at different duplex modes (e.g. one half, one full)
  - Common in networks where auto negotiation is disabled, or faulty
  - Classic example of a “soft failure”, connectivity is present and speeds are poor
- Developed analytical model to describe how Ethernet responds
- Expanding model to describe UDP and TCP flows
- Develop practical detection algorithm
- Test models in LAN, MAN, and WAN environments

# Faulty Hardware or Link

- Detect non-congestive loss due to
  - Faulty NIC/switch interface
  - Bad Cat-5 cable
  - Dirty optical connector

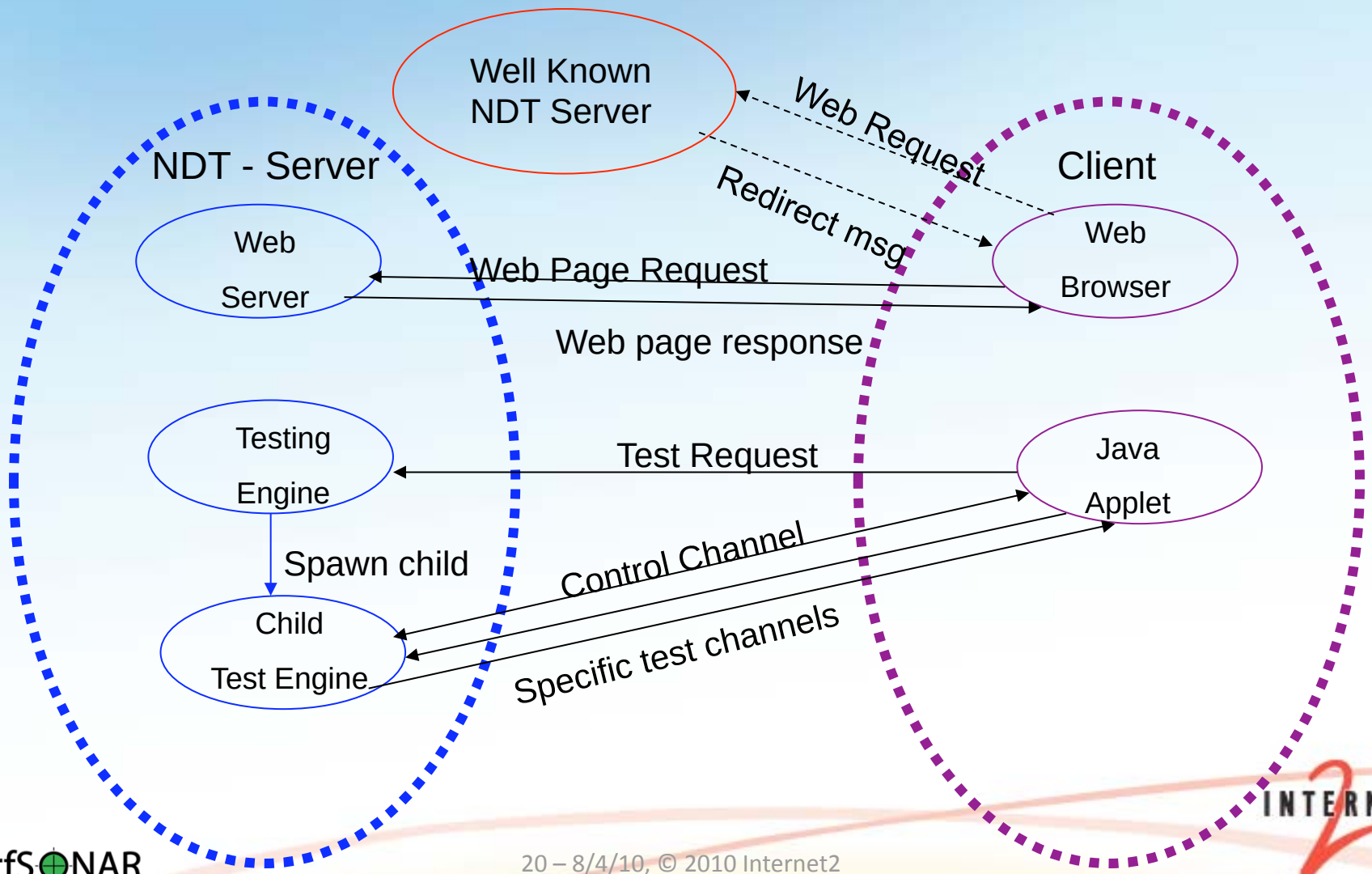
# Congestion Detection

- Shared network infrastructures will cause periodic congestion episodes
  - Detect/report when TCP throughput is limited by cross traffic
  - Detect/report when TCP throughput is limited by own traffic

# Additional Functions and Features

- Provide basic tuning information
- Features:
  - Basic configuration file
  - FIFO scheduling of tests, support for testing with simultaneous clients
  - Simple server discovery protocol
  - Logging of all test results on the server side
- Command line client support
- Other Clients can be developed against open Javascript API:
  - http://www.internet2.edu/performance/ndt/api.html
- Posted on Google Code:
  - http://code.google.com/p/ndt/

# Architecture



# Finding a Server – The Old Way

- Static List of servers – doesn't scale

E2Epi | NDT Servers - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://e2epi.internet2.edu/ndt/ndt-server-list.html

>web100srv(8)
>fakewww(8)

**E2E PIPES**
>piPEs Software
>>BWCTL
>>OWAMP
>>NDT
>> Thrulay
>Global PMP Directory
>Testing to Abilene
>>BWCTL
>>OWAMP
>Abilene Results

**WORKING GROUPS & COLLABORATIONS**
> Working Groups
> Collaborations

NDT Servers

**Abilene** - http://e2epi.internet2.edu/ndt/

| Location | Host | Interface | Online Stats |
|---|---|---|---|
| Indianapolis | http://ndt-indianapolis.abilene.ucaid.edu:7123 | GigE | http://ndt-indianapolis.abilene.ucaid.edu:7123/admin.html |
| New York | http://ndt-newyork.abilene.ucaid.edu:7123 | GigE | http://ndt-newyork.abilene.ucaid.edu:7123/admin.html |
| Seattle | http://ndt-seattle.abilene.ucaid.edu:7123 | GigE | http://ndt-seattle.abilene.ucaid.edu:7123/admin.html |
| Sunnyvale | http://ndt-sunnyvale.abilene.ucaid.edu:7123 | GigE | http://ndt-sunnyvale.abilene.ucaid.edu:7123/admin.html |
| Washington DC | http://ndt-washington.abilene.ucaid.edu:7123 | GigE | http://ndt-washington.abilene.ucaid.edu:7123/admin.html |

**PMP Equipment Used for Abilene:**

- Intel SCB2 motherboard,
- 2 × 1.266 GHz PIII, 512 KB L2 cache, 133MHz FSB,
- 2 ×512MB ECC registered RAM (one/slot to enable interleaving),
- 2 × SEAGATE 18GB SCSI (ST318406LC)
- Primary NIC: SysConnect Gigabit Ethernet SK-9843 SX
- Secondary NIC Inter Ethernet Pro 10/100+ (i82555)

**Other Available Servers:**

| Location | Host |
|---|---|
| ANL | http://miranda.ctd.anl.gov:7123/ |
| Swiss Education and Research Network (Switzerland) | http://cemp1.switch.ch/network/performance/web100/tcpbw100.html |
| University of Michigan (Flint) | http://speedtest.umflint.edu/ |
| UCal Santa Cruz | http://nitro.ucsc.edu/ |
| Thomas Jefferson National Accelerator Facility | http://jlab4.jlab.org:7123/ |
| Stanford University | http://netspeed.stanford.edu/ |
| NSF (Arlington, VA) | http://ciseweb100.cise-nsf.gov:7123 |
| University of Hawaii (Honolulu) | http://farnsworth.uhnet.net:7123/ |
| St. Mary's College of Maryland (Maryland) | http://ndt.smcm.edu:7123/ |

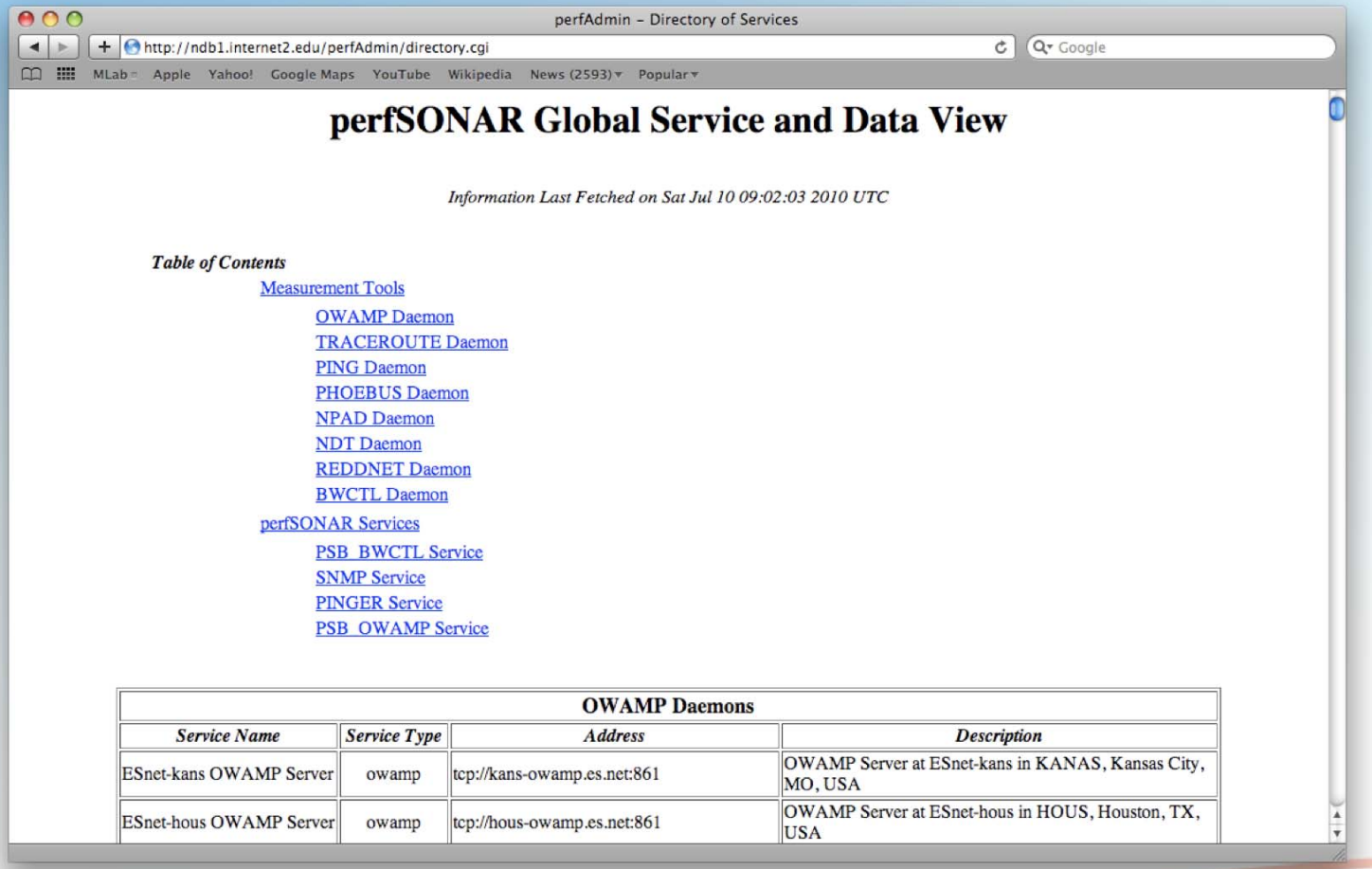
NDT 28 May, 2004

The original idea and implementation of the web-based testing server was designed and implemented by Tom Dunnigan from Oak Ridge National Laboratory. It has been extensively modified by Rich Carlson and changed to perform the current functions. This material is based on work supported [in part] by the Office of Science, U.S. Department of Energy under Contract W-31-109-ENG-38 and Argonne National Laboratory.

Done

# Finding a Server – The New Way

- perfSONAR Infrastructure – automatically search for instances




## perfSONAR Global Service and Data View

*Information Last Fetched on Sat Jul 10 09:02:03 2010 UTC*

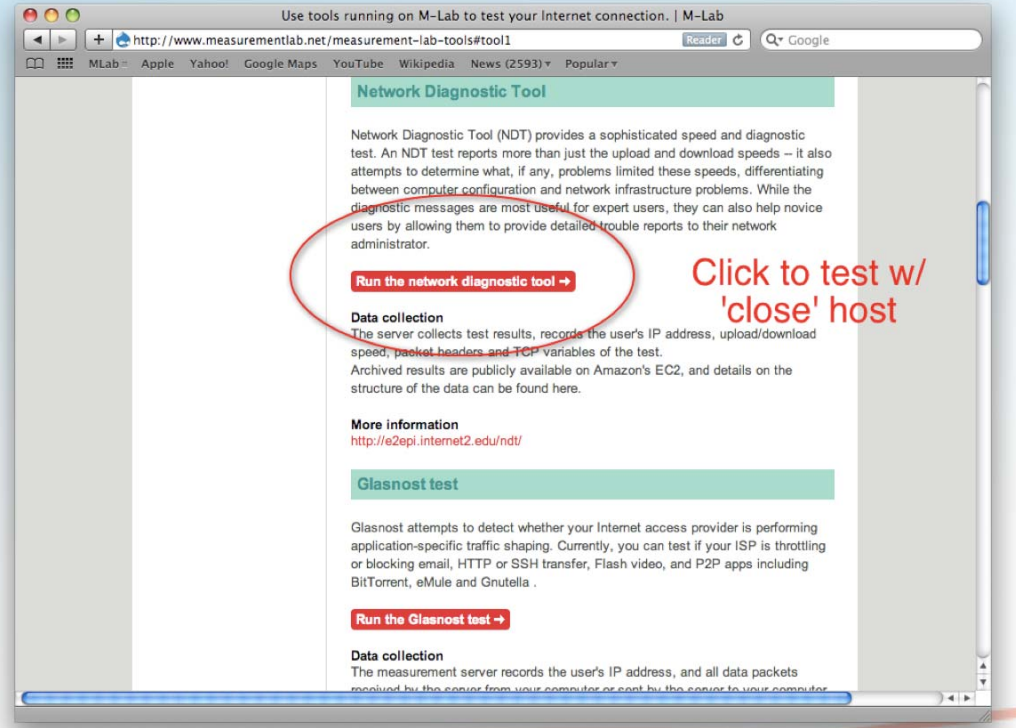
***Table of Contents***


- Measurement Tools
  - OWAMP Daemon
  - TRACEROUTE Daemon
  - PING Daemon
  - PHOEBUS Daemon
  - NPAD Daemon
  - NDT Daemon
  - REDDNET Daemon
  - BWCTL Daemon
- perfSONAR Services
  - PSB_BWCTL Service
  - SNMP Service
  - PINGER Service
  - PSB_OWAMP Service


| OWAMP Daemons | | | |
|---|---|---|---|
| ***Service Name*** | ***Service Type*** | ***Address*** | ***Description*** |
| ESnet-kans OWAMP Server | owamp | tcp://kans-owamp.es.net:861 | OWAMP Server at ESnet-kans in KANAS, Kansas City, MO, USA |
| ESnet-hous OWAMP Server | owamp | tcp://hous-owamp.es.net:861 | OWAMP Server at ESnet-hous in HOUS, Houston, TX, USA |

# Finding a Server – MLab

- Measurement Lab
  - Joint Project between several partners
  - More Info Here: http://www.measurementlab.net/
- Locate a 'close' NDT server using DONAR (http://donardns.org/)



# Interpreting Results

- Changing desktop effects performance
  - Lesson in why testing end-to-end is necessary
- Faulty Hardware identification
  - When is performance being effected by the environment or the equipment?

# Different Host – Same Switch Port

- Theme: Its important to test to the user's desktop to see what they are seeing. Changing hardware changes performance observations
  - E.g. tech support can't show up with a 'tuned' laptop to prove the network is functional – this doesn't help the user…
- Host 1: 10 Mbps NIC
  - Throughput 6.8/6.7 Mbps send/receive
  - RTT 20 ms
  - Retransmission/Timeouts 25/3
- Host 2: 100 Mbps NIC
  - Throughput 84/86 Mbps send/receive
  - RTT 10 ms
  - Retransmission/Timeouts 0/0
- Interpretation:
  - Ignore speed for a second
  - 70% utilization on the first vs 85%
  - Why are we seeing retransmissions and timeouts?

# LAN Testing

- The following is a test on our Lab LAN
  - 12 PCs
  - All connected to a Switch
  - 2 VLANs
  - Router linking VLANs
- All testing is VLAN to VLAN, e.g. crossing the router.
- Things to note:
  - 100MB Full Duplex unless noted
  - Look for correlations between RTT and Speed
  - Look at Loss rates
- Can you identify what may be suspect based on the observations?

# LAN Testing Results

**100 Mbps Full Duplex**

| Ave Rtt | %loss | Speed |
|---|---|---|
| 5.41 | 0.00 | 94.09 |
| 1.38 | 0.78 | 22.50 |
| 6.16 | 0.00 | 82.66 |
| 14.82 | 0.00 | 33.61 |

**10 Mbps**

| | | |
|---|---|---|
| 72.80 | 0.01 | 6.99 |
| 8.84 | 0.75 | 7.15 |

# LAN Testing Results

**100 Mbps Full Duplex**

| Ave Rtt | %loss | loss/sec | Speed | |
|---|---|---|---|---|
| 5.41 | 0.00 | 0.03 | 94.09 | Normal Operation |
| 1.38 | 0.78 | 15.11 | 22.50 | Bad Switch Interface |
| 6.16 | 0.00 | 0.03 | 82.66 | Reverse of Above… |
| 14.82 | 0.00 | 0.10 | 33.61 | Congestion |

**10 Mbps**

| | | | | |
|---|---|---|---|---|
| 72.80 | 0.01 | 0.03 | 6.99 | Normal Operation |
| 8.84 | 0.75 | 4.65 | 7.15 | Same Bad Interface |

# General Requirements – Support

- Source should compile for all modern *NIX
  - *BSD, Linux, OS X
  - configure/make/make install
- Web100 Patched Kernel
  - perfSONAR-PS Project also offers two alternatives:
    - pS Performance Toolkit (bootable ISO)
    - Pre-packaged kernel with Web100 for CentOS
- Other Software
  - Java SDK
  - Libpcap
- RPMs compiled specifically for CentOS 5.x
  - May work with other RPM based systems (Fedora, RHEL)

# Recommended Settings

- There are no settings or options for the Web based java applet.
  - It allows the user to run a fixed set of tests for a limited time period
- Test engine settings
  - Turn on admin view (-a option)
  - If multiple network interfaces exist use –i option to specify correct interface to monitor (ethx)
- Simple Web server (fakewww)
  - Use –l fn option to create log file
  - Could also use a 'real' web server like Apache

# Potential Risks

- Non-standard kernel required
  - Web100 patching may be difficult to apply to new kernels
  - Hard to keep up with vendor patching
  - GUI tools can be used to monitor other ports
  - Consider using pS Performance Toolkit enhancements if this scares you…
- Public servers generate trouble reports from remote users
  - Respond or ignore emails
- Test streams can trigger IDS alarms
  - Configure IDS to ignore NDT server

# Availability

- Main Page:
  - http://www.internet2.edu/performance/ndt
- Mailing lists:
  - ndt-users@internet2.edu
  - ndt-announce@internet2.edu

# Hands On

- Testing NDT:

# NDT

August 10th 2010, OSG Site Admin Workshop – Network Performance

Jason Zurawski – Internet2

For more information, visit http://www.internet2.edu/workshops/npw