

Getting the most out of your network connection

Further work Testing at 10Gbps
on the AARnet network

Tim Rayner
ACT Regional Network Manager
AARNet

QUESTnet 2009

Gold Coast, July 9th 2009



AARNet Copyright 2009

Contents

- ➔ ▪ The End-to-End Performance Problem
- AARNet member examples
- Limiting factors
- 10 Gbps Test Kit
- Measurement tools – Iperf and NDT
- Test Results
- Future work

End to End performance

- The symptom:
 - I'm not getting the throughput I expect
 - Issues increase with latency
- AARNet provides long, thick pipes: 1Gbps, 10Gbps optical PtP & A3
 - Generally un-congested – people understand congestion & are forgiving!
- Customer examples:
 - 300 km 1Gbps optical circuit -> achieve 150 Mbps
 - Group in one Uni accessing a research centre at another – 2Mbps
 - 20ms RTT 1Gbps optical circuit apps are slower than 4ms 34 Mbps uWave.
 - CSU – first 10Gbps customer circuit Wagga – Bathurst – test shortly.
 - 1 Mbps across optical circuit

Survey: Real World Best Performance ?

- What's the best performance you've received for real-world tasks ? –
 - 1. < 20 Mbps
 - 2. 20 – 80 Mbps
 - 3. 80 – 250 Mbps
 - 4. 250 Mbps – 900 Mbps
 - 5. 900 Mbps – 1.5 Gbps
 - 6. > 1.5 Gbps
 - 7. Unknown

Survey: Latency

- What is the maximum RTT latency you experience between campuses of your institution ?
 - 1. < 1 ms
 - 2. 1 – 5 ms
 - 3. 5 – 20 ms
 - 4. 20 – 80 ms
 - 5. 80 – 300 ms
 - 6. > 300 ms

Limiting Factors – non Network

- System BUS speeds:
 - Standard PCI 33 MHz, 32 bit: 500 – 1000 Mbps
 - PCI-X 133MHz, 64 bit: 6 – 8 Gbps
 - PCI-Express1.0 2.5 Gbps x8 channels: 20 Gbps
 - PCI-Express2.0 5.0 Gbps x8 channels: 40 Gbps
- Storage Speeds:
 - SATA: 1.5 Gbps, 3.0 Gbps
 - Individual Disks significantly slower
 - Fibre Channel SAN: 1,2,4,8 Gbps
- Conclusion: We need memory – memory transfers across PCI-e x8 to approach 10 Gbps

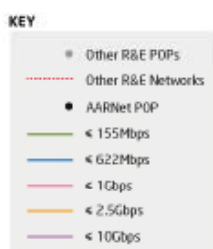
Limiting Factors: Network

- Host based:
 - TCP stack tuning – good for linux kernel $\geq 2.6.18$
 - TCP Window scaling – Max. window size \geq Bandwidth x RTT
 - Linux: net.ipv4.tcp_rmem & net.ipv4.tcp_wmem in /etc/sysctl.conf
 - Set last parameter to large size (eg. 150 Mbytes)
 - Fast enough CPU
- Network based:
 - MTU: 9000 Byte Jumbos, rather than standard 1500 Bytes
 - Tracepath – like traceroute, but reports Path MTU to each hop
 - Expect busier CPUs at 1500 Bytes. 3.7Gbps vs. 5.6Gbps UDP iperf
 - Make sure every step has jumbos enabled – physical i/f and vlans
 - Testing to uHawaii – 9000 Bytes one direction, 1500 other direction

Test Locations:

- Canberra & Perth : 45ms RTT

AARNet National Network



Test Locations:

- Canberra to uHawaii - RTT: 99ms

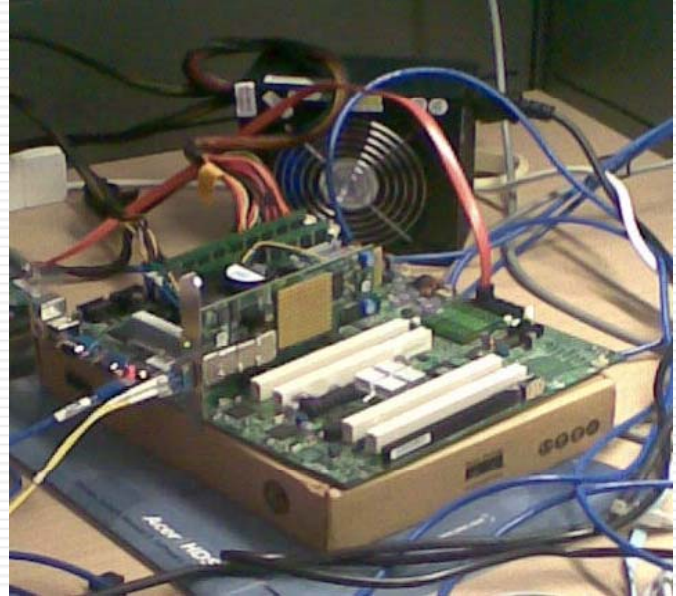
AARNet International Network



Test Kit: details

- Last year:
 - 1RU Acer 2.0GHz Intel Xeon servers with CentOS5
 - 3 Gbps with standard kernel (2.6.18)
 - 6 Gbps with custom kernel 2.6.24
 - Updated myri10ge drivers improved performance & stopped hangs
 - 3.16GHz desktop E8500 approached 10 Gbps
 - Got fastest available Xeon at 3.16 GHz – gave 9.8 Gbps without being cpu limited.
 - Servers were expensive, and not very portable
 - Exclusively used the myricom cards with LR singlemode XFP optics

Test kit: Last year Acer R520 & DIY system



Test Kit: cards – myricom XFP & intel dual SFP+



Test Kit: New server details

- Tried new intel core-i7 processors when newly released
 - Best performance yet with the slower 2.67 GHz.
 - \$500 processor getting wirespeed – faster version was 4 x price
 - System price around \$1600, plus card at \$695 + \$900 US
- Buy as small a system as possible with same processor
 - System ordered
 - Delays expected – “If you change to processor ... we can deliver”
 - In the end delivered with new technology 2.13 GHz Xeon
 - Client results disappointed a little. – 8, 7.5, 6.5 Gbps as sender
 - Performed well as server – up to wirespeed

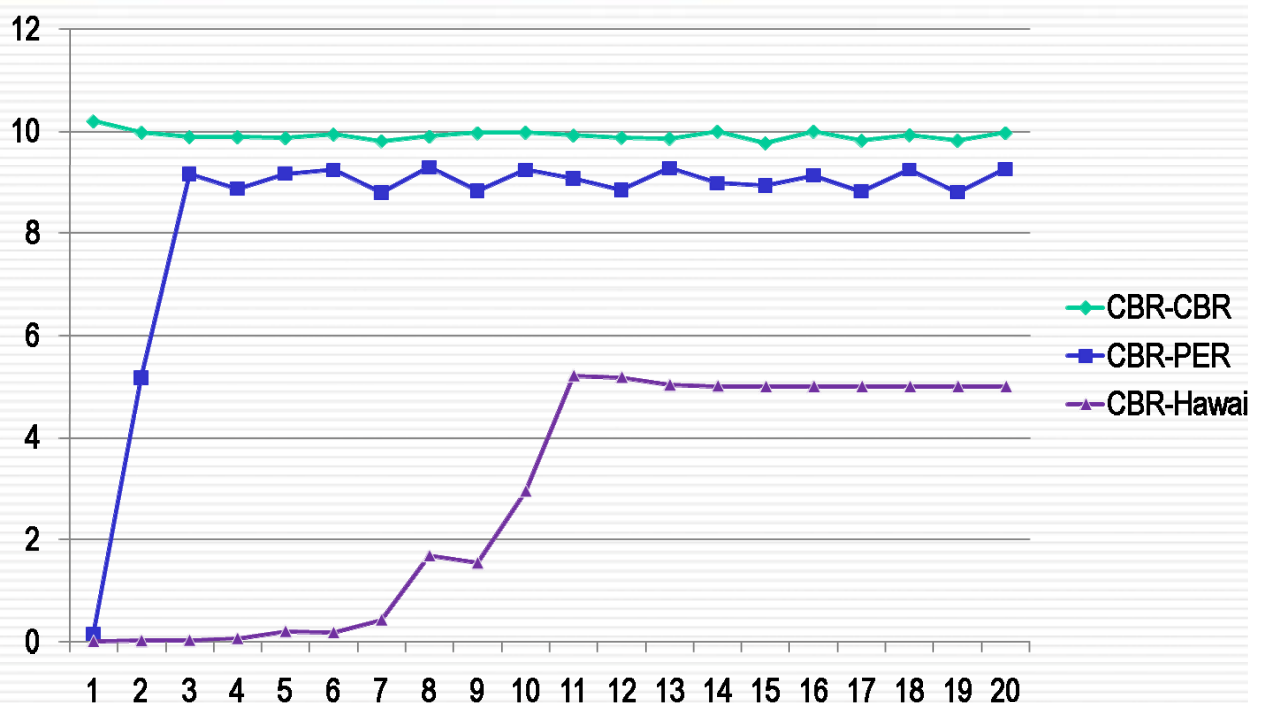
Test Kit: New servers: core i7 & small 1RU



Measurement tool: iperf

- Memory to memory transfers
- Lots of options – option order is important.
- Command line clients/servers for range of operating systems
 - Linux, windows, OSX & other unix.
- Supports TCP and UDP
- Don't set manual TCP windows – disables window auto-scaling
- Conventional wisdom: use UDP
 - Bypasses any TCP tuning/buffering issues. – congestion risk.
 - Need undocumented -l option to set jumbo datagram size
 - Don't do it! – tune your TCP stack for better results & side-step congestion.

10Gbps Test Results



Iperf examples: TCP

```

tcr@cpe-aarnet-es8:~
File Edit View Terminal Tabs Help
[tcr@cpe-aarnet-es8 ~]$ ./iperf-2.0.2/src/iperf -c 10.10.10.81 -i 1 -t 10
-----
Client connecting to 10.10.10.81, TCP port 5001
TCP window size: 27.4 KByte (default)
-----
[ 3] local 10.10.10.88 port 39247 connected with 10.10.10.81 port 5001
[ 3] 0.0- 1.0 sec 1.15 GBytes 9.90 Gbits/sec
[ 3] 1.0- 2.0 sec 1.15 GBytes 9.88 Gbits/sec
[ 3] 2.0- 3.0 sec 1.15 GBytes 9.88 Gbits/sec
[ 3] 3.0- 4.0 sec 1.15 GBytes 9.88 Gbits/sec
[ 3] 4.0- 5.0 sec 1.15 GBytes 9.88 Gbits/sec
[ 3] 5.0- 6.0 sec 1.15 GBytes 9.88 Gbits/sec
[ 3] 6.0- 7.0 sec 1.15 GBytes 9.88 Gbits/sec
[ 3] 7.0- 8.0 sec 1.15 GBytes 9.88 Gbits/sec
[ 3] 8.0- 9.0 sec 1.15 GBytes 9.88 Gbits/sec
[ 3] 0.0-10.0 sec 11.5 GBytes 9.88 Gbits/sec
[tcr@cpe-aarnet-es8 ~]$

tcr@sm-10g-tst:~
File Edit View Terminal Tabs Help
^C[tcr@sm-10g-tst ~]$ ./iperf-2.0.2/src/iperf -s
-----
Server listening on TCP port 5001
TCP window size: 85.3 KByte (default)
-----
[ 4] local 10.10.10.81 port 5001 connected with 10.10.10.88 port 39247
[ 4] 0.0-10.0 sec 11.5 GBytes 9.88 Gbits/sec

```

Iperf Example: UDP

```

tcr@cpe-aarnet-es8:~
File Edit View Terminal Tabs Help
[tcr@cpe-aarnet-es8 ~]$ ./iperf-2.0.2/src/iperf -c 10.10.10.81 -b 6000M -l 8970 -i 1 -t 10
WARNING: option -b implies udp testing
-----
Client connecting to 10.10.10.81, UDP port 5001
Sending 8970 byte datagrams
UDP buffer size: 105 KByte (default)
-----
[ 3] local 10.10.10.88 port 50697 connected with 10.10.10.81 port 5001
[ 3] 0.0- 1.0 sec 778 MBytes 6.52 Gbits/sec
[ 3] 1.0- 2.0 sec 778 MBytes 6.52 Gbits/sec
[ 3] 2.0- 3.0 sec 778 MBytes 6.52 Gbits/sec
[ 3] 3.0- 4.0 sec 778 MBytes 6.52 Gbits/sec
[ 3] 4.0- 5.0 sec 778 MBytes 6.52 Gbits/sec
[ 3] 5.0- 6.0 sec 778 MBytes 6.52 Gbits/sec
[ 3] 6.0- 7.0 sec 778 MBytes 6.52 Gbits/sec
[ 3] 7.0- 8.0 sec 778 MBytes 6.52 Gbits/sec
[ 3] 8.0- 9.0 sec 778 MBytes 6.52 Gbits/sec
[ 3] 9.0-10.0 sec 778 MBytes 6.52 Gbits/sec
[ 3] 0.0-10.0 sec 7.59 GBytes 6.52 Gbits/sec
[ 3] Sent 909088 datagrams
[ 3] Server Report:
[ 3] 0.0-10.0 sec 6.50 GBytes 5.58 Gbits/sec 0.010 ms 131019/909087 (14%)
[ 3] 0.0-10.0 sec 1 datagrams received out-of-order
[tcr@cpe-aarnet-es8 ~]$

tcr@sm-10g-tst:~
File Edit View Terminal Tabs Help
[tcr@sm-10g-tst ~]$ ./iperf-2.0.2/src/iperf -s -u -l 8970
-----
Server listening on UDP port 5001
Receiving 8970 byte datagrams
UDP buffer size: 109 KByte (default)
-----
[ 3] local 10.10.10.81 port 5001 connected with 10.10.10.88 port 50697
[ 3] 0.0-10.0 sec 6.50 GBytes 5.58 Gbits/sec 0.011 ms 131019/909087 (14%)
[ 3] 0.0-10.0 sec 1 datagrams received out-of-order

```

Test Results

- Canberra to Perth
 - Proved last year over prototype A3 10Gbps customer cpe connection
 - Proved this year over NCN VPLS service
 - Canberra, Perth and Sydney servers on the same subnet
 - Rate limits tripped me up when trying to reproduce results

Test Results

- Trans-pacific
 - Did some iperf tests with uHawaii across 10Gbps SX-Transport North
 - 99ms RTT
 - Good Acer R520 in Canberra, myricom card
 - Less capable server in Hawaii, with myricom card
 - Consistent results > 5Gbps Canberra -> Hawaii
 - Highly variable results Hawaii – Canberra. 1.5 or 3 Gbps most times, but one time 8 Gbps
 - Results no better with netperf or nuttcp
 - CPU limited results can be highly variable
 - sometimes “in the groove” – 8 Gbps on one occasion

Measurement Tool: NDT

- Client side is simple - just needs a Java enabled browser
- Command-line client is available
- Server side more complicated – web100 linux kernel required
 - Packet sniffing used to measure packet arrival times
 - Confused by optimised network drivers – packets arrive together
- Attempts connections and 10 second each-way transfer
- Uses heuristics to report duplex issues, NAT, firewalling, constraining network connection
- Designed to help end users self-diagnose problems & provide evidence to a NOC team
- Heuristics can be inaccurate – eg Link speed, Middlebox MSS.

NDT Interface

AARNet Web100 based Network Diagnostic Tool (NDT)

Located at Canberra, ACT, Australia; 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 v5.5.4b
click START to begin

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

TCP/Web100 Network Diagnostic Tool v5.5.4b
click START to begin

** Starting test 1 of 1 **

Connected to: 202.158.221.8 -- Using IPv4 address

Checking for Middleboxes Done

checking for firewalls Done

running 10s outbound test (client-to-server [C2S]) 937.73Mb/s

running 10s inbound test (server-to-client [S2C]) 938.98Mb/s

The slowest link in the end-to-end path is a 10 Gbps 10 Gigabit Ethernet/OC-192 subne

click START to re-test

Applet started.

START

Statistics

More Details...

Report

START

Statistics

More Details...

Report Problem

NDT details

The screenshot shows two windows from the NDT client. The 'Detailed Statistics' window on the left provides a comprehensive overview of the connection, including system details, network link information, and performance metrics. The 'Web100 Variables' window on the right displays a list of kernel variables related to the connection, such as MSS, RTT, and window sizes.

```

Detailed Statistics
WEB100 Enabled Statistics:
Checking for Middleboxes ..... Done
checking for firewalls ..... Done
running 10s outbound test (client-to-server [C2S]) ..... 937.73Mb/s
running 10s inbound test (server-to-client [S2C]) ..... 938.98Mb/s

----- Client System Details -----
OS data: Name = Linux, Architecture = i386, Version = 2.6.27.25-170.2.72.fc10.i686.PAE
Java data: Vendor = Sun Microsystems Inc., Version = 1.6.0_0

----- Web100 Detailed Analysis -----
10 Gbps 10 GigEthernet/OC-192 link found.
Link set to Full Duplex mode
No network congestion discovered.
Good network cable(s) found.
Normal duplex operation found.

Web100 reports the Round trip time = 3.62 msec; the Packet size = 1448 Bytes; and
No packet loss was observed.
S2C throughput test: Packet queuing detected: 0.12%
This connection is sender limited 99.92% of the time.
Increasing the NDT server's send buffer (127.0 KB) will improve performance

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

Server '202.158.221.8' is not behind a firewall. [Connection to the ephemeral port was succ
Client is probably behind a firewall. [Connection to the ephemeral port failed]
Information: Network Middlebox is modifying MSS variable
Server IP addresses are preserved End-to-End
Client IP addresses are preserved End-to-End

Web100 Kernel Variables:
Client: tcr-acr-It.aarnet.net.au/127.0.0.1
CurMSS: 1448
X_Rcvbuf: 262142
X_Sndbuf: 262142
AckPktsIn: 344955
AckPktsOut: 0
BytesRetrans: 0
CongAvoid: 0
CongestionOverCount: 0
CongestionSignals: 0
CountRTT: 344956
CurCwnd: 1216320
CurRTO: 206
CurRwinRcvd: 1550848
CurRwinSent: 20480
CurSsthresh: 2147483647
DSACKDups: 0
DataBytesIn: 0
DataBytesOut: 1184724192
DataPktsIn: 0
DataPktsOut: 812370
DupAcksIn: 0
ECNEnabled: 0
FastRetran: 0
MaxCwnd: 1216320
MaxMSS: 1448
MaxRTO: 207
MaxRTT: 16
MaxRwinRcvd: 1576512
MaxRwinSent: 20480
MaxSsthresh: 0
MinMSS: 1448
MinRTO: 201
MinRTT: 0
MinRwinRcvd: 5888
MinRwinSent: 17896
NagleEnabled: 1

```

Further work

- Get a tuned iperf server booting as a liveCD or thumbdrive
- Same for an NDT server – more difficult – kernel & Java work
- Explore newer OS – Fedora11 – squeeze more from hardware
- Explore options for iperf servers on customer cpe servers
- Explore a permanent production AARNet NDT server
- Do some high performance real-world storage tests.
 - New aarnet mirror will has 10 Gbps connection, and primed Flash cache. Preliminary aggregate 3.4 Gbps with F11 ISOs.
 - May bypass storage speed limitations.
- Assist customers experiencing end2end throughput issues
- We're keen to learn more & help - contact: noc@aarnet.edu.au.

Links

- IPERF – <http://iperf.sourceforge.net/>
- NDT - <http://e2epi.internet2.edu/ndt/>
- Internet2 Performance Workshops
 - Materials <http://www.internet2.edu/workshops/npw/materials.html>
- Search “TCP Performance Tuning” for your favorite OS

Questions ?

Tim Rayner
Tim.Rayner@aarnet.edu.au

QUESTnet 2009

Gold Coast, July 9th 2009