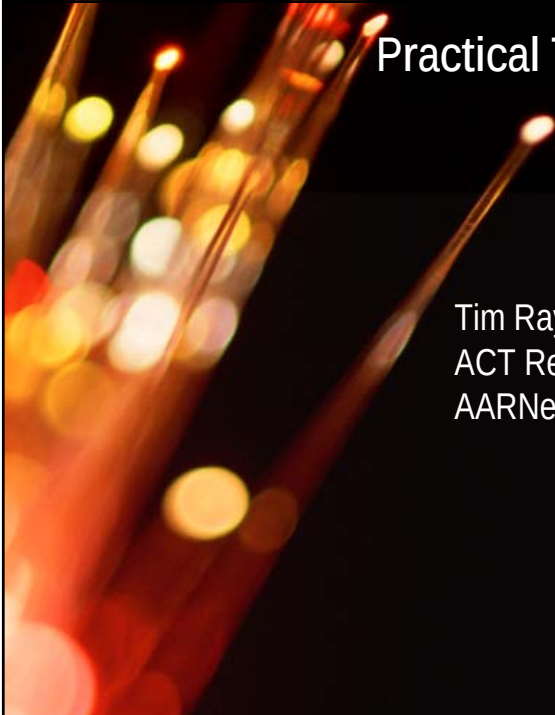




Practical TCP Performance Testing

Beyond 1Gbps
on the AARNet Network

Tim Rayner
ACT Region Network Manager
AARNet

QUESTnet 2008
Gold Coast, July 9th 2008



		AARNet Copyright 2008
	Contents	
		▪ New 10G offerings from AARNet ▪ Throughput questions ▪ The project •Equipment choice •Results •Incremental improvements

New offerings from AARNet

- 10 Gbps cpe connections to the AARNet Network
- 10 Gbps optical circuits
- ...will be covered in other AARNet presentations



3

Throughput questions

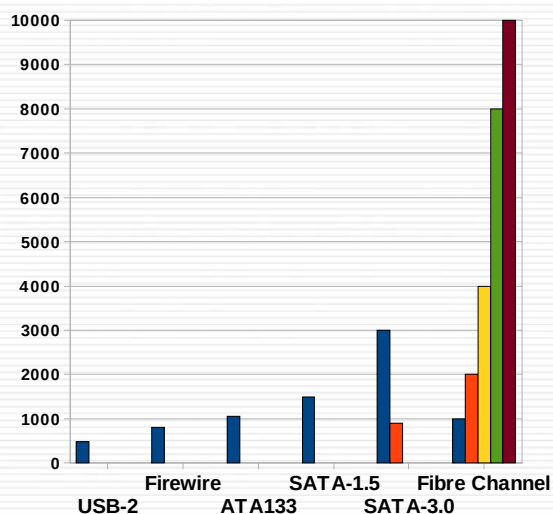
- We're getting 260 Mbps from our 1Gbps optical circuit
 - Is that good ?
- I'm only getting 20 Mbps from the AARNet mirror
 - I thought we had a Gigabit connection ?
- As AARNet makes even faster services available, we're expecting more questions.

Project Aims

- Install a 10Gbps service
- Purchase capable test gear
 - Match real world environments
- Measure throughput
 - Within same room
 - Within same city
 - Across the country – Canberra to Perth
- Wanted to experience some hurdles toward 10Gbps
 - Hopefully overcome some of them.
 - Better equipped to answer customer questions.

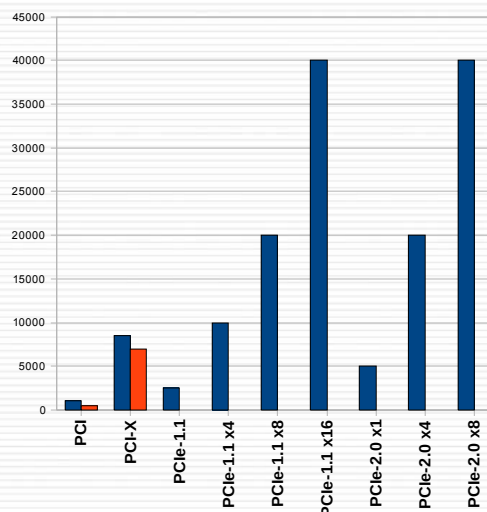
Speed Limitations – disk & storage

- Disk Bus speeds
- Disks slower than I/O Buses
- Actual Throughput short of Max speed
- Disk transfers won't saturate 10Gbps Network
- More to learn about practical SAN speeds



Speed Limitations – Server Bus

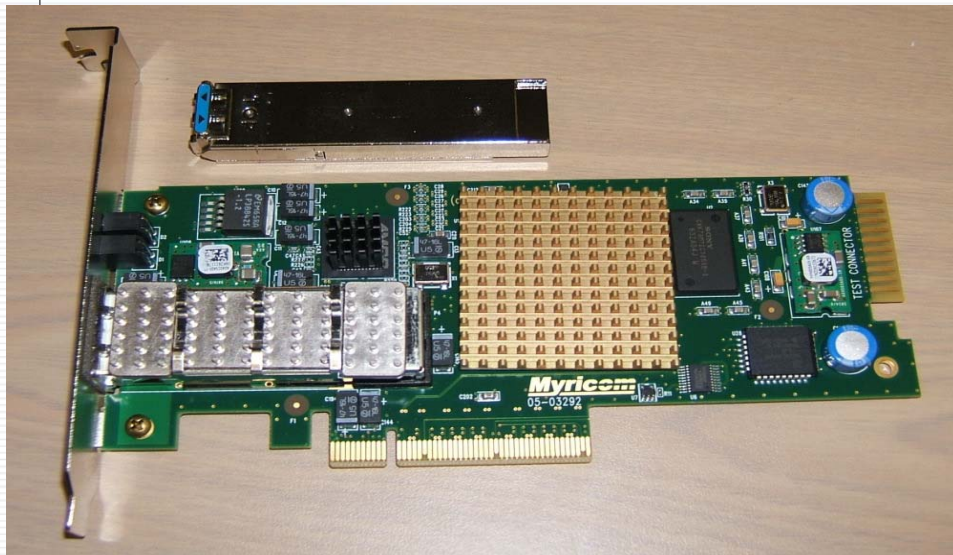
- PCI : 32bit x 33 MHz
- PCI-X : 64bit x 133 MHz
- PCI Express (PCIe)
 - V1.1
 - n channels x 2.5 Gbps
 - V2.0
 - Newly available
 - n channels x 5.0 Gbps



Equipment choice

- Network Card (must be PCIe1.1 x8)
 - Timeframe: end of 2007
 - Only available Intel card was PCI-X 64x166
 - Not close enough
 - Around since 2003, 2004
 - Recent PCIe2.0 x8 2x10G Card available
 - Myricom card (myri10ge)
 - PCIe1.1 x8
 - Supports MyriNet and Ethernet
 - Published benchmarks to 9.8 Gbps
 - Drivers in standard linux kernels
 - Versatile pluggable XFP optics

Myricom Network Card



10G Module choices

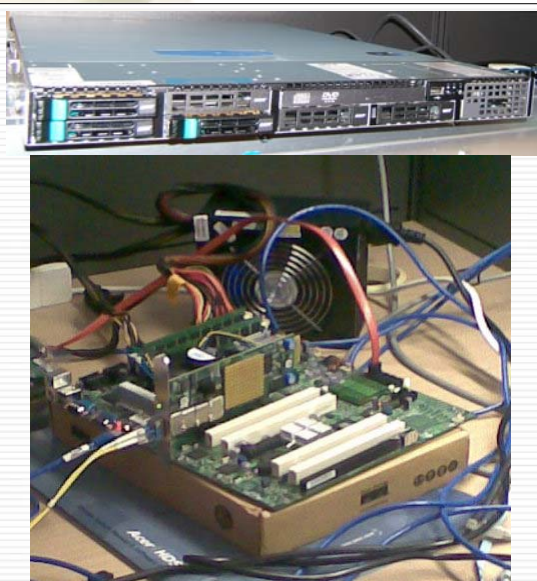


- Fibre
 - XenPak
 - X2
 - XFP
 - SFP+
- Copper
 - CX4
 - Twin-axial
 - UTP (Cat6+) - when ?

Server Choice

- Needed PCIe1.1 x 8 slots
- AARNet uses Acer servers
- Acer R520 met the Bus requirements
 - Tried with standard machines
 - Needed new ones with faster processors
- DIY machine
 - much cheaper
 - Didn't want an Ugly Duckling
 - Price increases with rack mount

Acer R520 Server & Friend



Operating system choice

- Used linux because of familiarity
- Didn't wish to display my windows ignorance
- Needed recent kernels for:
 - Device support ($\geq 2.6.18$)
 - Auto-tuning enhancements ($\geq 2.6.17$)
- Tried:
 - Fedora 9
 - RHEL5
 - CentOS5.1
 - Both 32 and 64 bit versions
- Tested with Interface MTU of 1500, but mostly 9000

Testing Software

- Must avoid disk bus speed limitations
 - Needs to perform memory to memory transfers
- Iperf
 - Well known
 - Simple and useful reporting options
 - Supports TCP and UDP
 - Uses more CPU
- Netperf
 - Used for published high-end benchmarks
 - More cumbersome, no intermediate reporting
 - Uses less CPU

Testing Software (cont.)

- Home-grown TCPtest
 - Courtesy of Greg Wickham's C++ skills
 - Idea for UDP in parallel with TCP
 - Would allow easy parallel RTT and jitter measurements
 - Haven't used it in the end
- Perl TCP client and server
 - Didn't get to this
 - Interested to view CPU effect.
- Don't use unmodified ssh/scp
 - Implements its own buffers – they get in the way with large RTT
- Tried some apache tests

Initial Results – Back2Back in lab

- 2 x Acer R520 servers
 - One Dual core 5110 Xeon@1.6 GHz, 1GB RAM
 - One Quad core E5405 Xeon@2.0 GHz, 1GB RAM
- Out of the box with Fedora 9
 - Booting problems with recent distros
 - Disappointing performance

Throughput (Gbps)	Test Program	Server	ServerCPU %	Client	Client CPU %age
2.43	iperf	2.0	163%	1.6	199%
2.72	iperf	1.6	185%	2.0	199%
3.06	netperf	2.0	29%	1.6	99%
3.91	netperf	1.6	57%	2.0	99%
2.46	netperf	1.6	loopback	1.6	
2.78	netperf	2.0	loopback	2.0	

Initial Results

- Receiving kernel errors about NMI interrupts
 - Occasional server hangs
- Performance limited by Client CPU
- Netperf performs better as a result
- Loopback interface results un-predictable
 - eg. later unexplained iperf jumps from 16 Gbps to 6 Gbps

Customising the OS

- Moved to CentOS5.1
 - Less “bleeding-edge” than Fedora 9
- Built a newer, custom kernel (2.6.24)
 - Up from 2.6.18 that ships with CentOS5
- Installed Newer myricom drivers
 - Up from 1.0.0 that ships with 2.6.18
 - Up from 1.3.2 that ships with 2.6.24
 - Latest version was 1.4.1 at the time
- Switched off unnecessary services
- Tried 64 bit OS install
- Aimed to later re-trace these steps

Customised OS results

- Interrupt warnings & system hangs stopped
- Significantly better results:
 - Acer to Acer: 7.7 & 8.8 Gbps with netperf
 - Acer to Acer: 6.6 & 7.8 Gbps with iperf
- Client CPU still limiting factor
- Introduced 3.16 GHz E8500 system
 - 9.8 & 7.5 Gbps to Acers with netperf
 - 8.9 & 7.5 Gbps to Acers with iperf
- Conclusion:
 - Need Faster CPUs

Further Results

- Reverted to 32 bit OS
 - Results with Acers were comparable to 64 bit
- Replicated testing through cisco 3750E switch at Layer2 and Layer3
- Looked at getting rack mount E8500 style systems
 - Opted for fastest available processor in the Acers
 - Eventually purchased 2x Acer R520 with 3.16 GHz X5460 Xeons
 - Each with 4 GB of RAM

New Faster Servers

- Results initially disappointing
 - No better than slower CPU servers
 - Servers reported only running @ 2.0 GHz
 - Solved: “cpuspeed” service was saving power, but killing results – switched it off.
- Results like we wanted...
 - Back2Back server results of 9.81 Gbps with iperf
 - CPUs were as low as 74%, server 85%
- We'd approached linespeed
 - With iperf (preferred tool)
 - Without CPU bottle-necks

Remaining Tests

- Test across fibre in Canberra
 - 16km of Fibre – AARNet office to ANU POP
 - Still < 1ms RTT
 - 10GBaseLR SM optics out of spec for this distance
 - Thankfully Just worked anyway
 - No significant differences
- Test between Canberra and Perth
 - Across AARNet 10G Backbone
 - Trialling with new 10G cpe equipment
 - RTT = 44.5ms
 - Cpe equipment configuration “under development”
 - Last Optics only arrived Friday last week!

Canberra to Perth

- Expected significant challenges
- Un-tweaked throughput
 - Up to 705 Mbps – compared with 9.8 Gbps Back2Back
- Tweaking the OS for large bandwidth-delay product can be quite involved
- Later linux kernels perform auto-tuning
- Auto-tuning has configurable max. buffer sizes of 4MB
`(/proc/sys/net/ipv4/tcp_wmem & tcp_rmem)`
 - $4M * 8 \text{ bits} / 0.045 \text{ sec} = 711 \text{ Mbps}$
 - Limiting factor for $RTT > 3ms$
- Make max buffer sizes bigger (100MB)

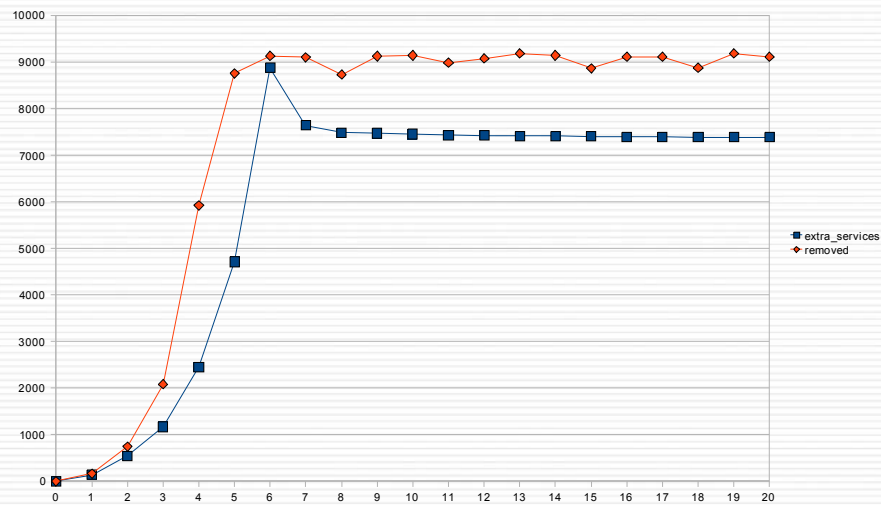
Canberra to Perth cont.

```
# echo "4096 87380 100000000" >/proc/sys/net/ipv4/tcp_wmem
# echo "4096 87380 100000000" >/proc/sys/net/ipv4/tcp_rmem

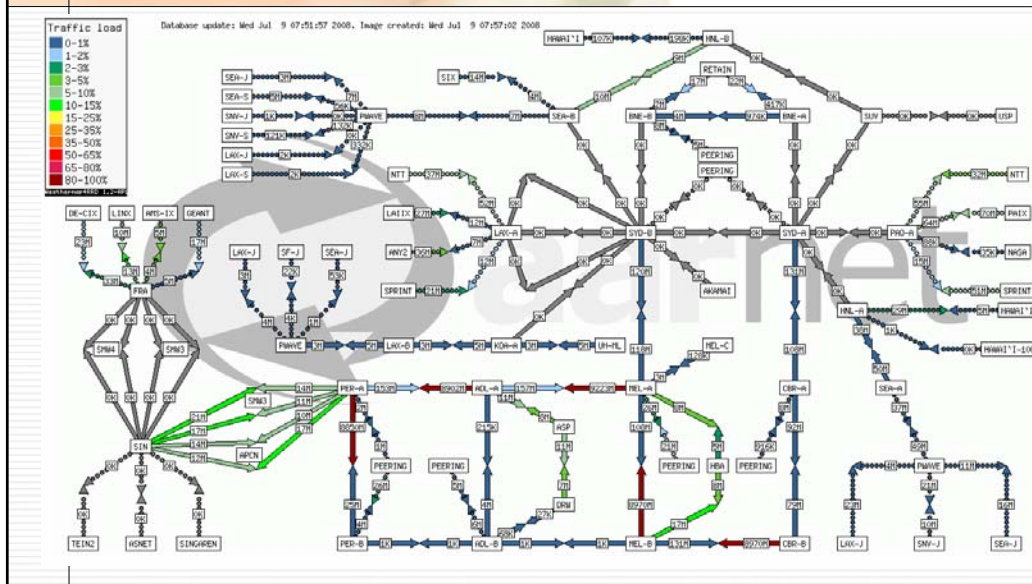
-----
Client connecting to 202.158.198.198, TCP port 5001
TCP window size: 85.3 KByte (default)

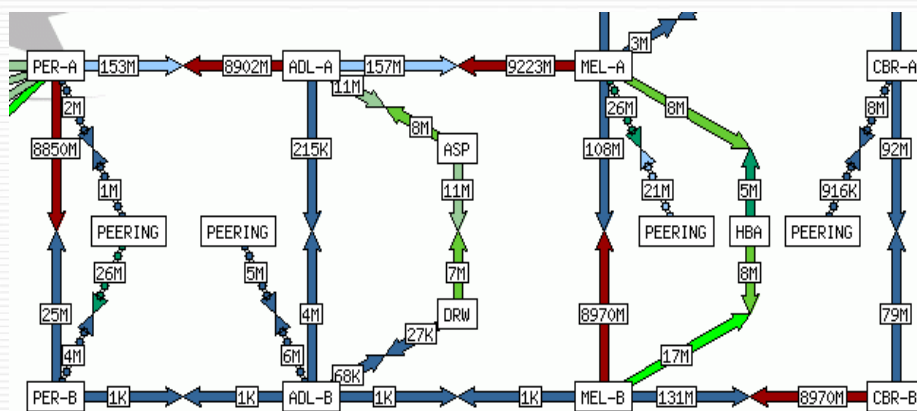
-----
[ 3] local 202.158.207.106 port 35681 connected with 202.158.198.198 port 5001
[ 3] 0.0- 1.0 sec  1.64 MBytes  13.8 Mb/s/sec
[ 3] 1.0- 2.0 sec  14.5 MBytes  122 Mb/s/sec
[ 3] 2.0- 3.0 sec  79.8 MBytes  669 Mb/s/sec
[ 3] 3.0- 4.0 sec  220 MBytes  1.85 Gb/s/sec
[ 3] 4.0- 5.0 sec  636 MBytes  5.34 Gb/s/sec
[ 3] 5.0- 6.0 sec  1.01 GBytes  8.68 Gb/s/sec
[ 3] 6.0- 7.0 sec  1000 MBytes  8.39 Gb/s/sec
[ 3] 7.0- 8.0 sec  998 MBytes  8.38 Gb/s/sec
[ 3] 8.0- 9.0 sec  1000 MBytes  8.39 Gb/s/sec
[ 3] 9.0-10.0 sec  1001 MBytes  8.40 Gb/s/sec
[ 3] 10.0-11.0 sec  999 MBytes  8.38 Gb/s/sec
[ 3] 11.0-12.0 sec  857 MBytes  7.19 Gb/s/sec
[ 3] 12.0-13.0 sec  1.10 GBytes  9.46 Gb/s/sec
```

Canberra – Perth iperf throughput



Well utilised backbone

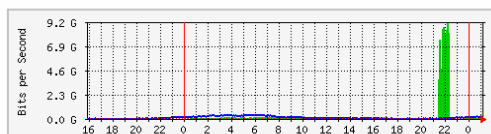




per-a-bb1 so-0/1/0 ADL-A-BB1 (STM64c APLAP101) 09-Jul-2008

Wed Jul 9 00:00:00 2008 Wed Jul 9 01:10:00 2008 +0000

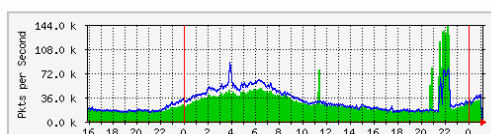
Bits per Second



Max	Inkbps	133448
Average	Inkbps	110865
Max	Outkbps	293501
Average	Outkbps	244927

Total	In	Mb	54047.10
Total	Out	Mb	119402.11

Pkts per Second

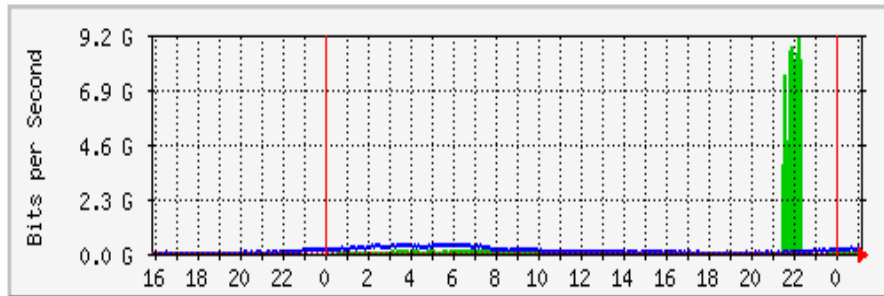


Max	Inpps	34424
Average	Inpps	29962
Max	Outpps	41348
Average	Outpps	34576

Total	In	Mp	116.86
Total	Out	Mp	134.85

Graph - zoomed

Bits per Second



Canberra to Perth Results

- Distance around 4000 km
- Network RTT is 44.5 ms
- Iperf single stream TCP performance
 - 6.94 Gbps over 60 seconds with iptables,cpuspeed
 - 8.98 Gbps over 10 minutes without
- Netperf performance
 - 8.59 Gbps over 10 minutes without

Conclusions

- Choose systems and network cards carefully
- Server CPU Speed seems a significant factor
- Auto-tuning in recent linux kernels made life with long RTT quite easy
- Nice to have real life experience with 10Gbps throughput
- We're ready to support customers with 10Gbps throughput issues on the AARNet network
- Might look towards 10G testing infrastructure in Seattle to test the SX-Transport research link across the pacific to the US

Links

- Performance tuning
 - <http://acs.lbl.gov/TCP-tuning/>
- Including recent linux kernels
 - <http://www.psc.edu/networking/projects/tcptune/>
- Iperf
 - <http://dast.nlanr.net/Projects/Iperf/>
- Netperf
 - <http://www.netperf.org/netperf/>
- Myricom Cards
 - <http://www.myri.com/Myri-10G/overview/>
 - <http://www.myri.com/scs/performance/>



Questions ?

Tim Rayner
Tim.Rayner@aarnet.edu.au

QUESTnet 2008

Gold Coast, July 9th 2008

