Next Generation Networking and Peering in the USA

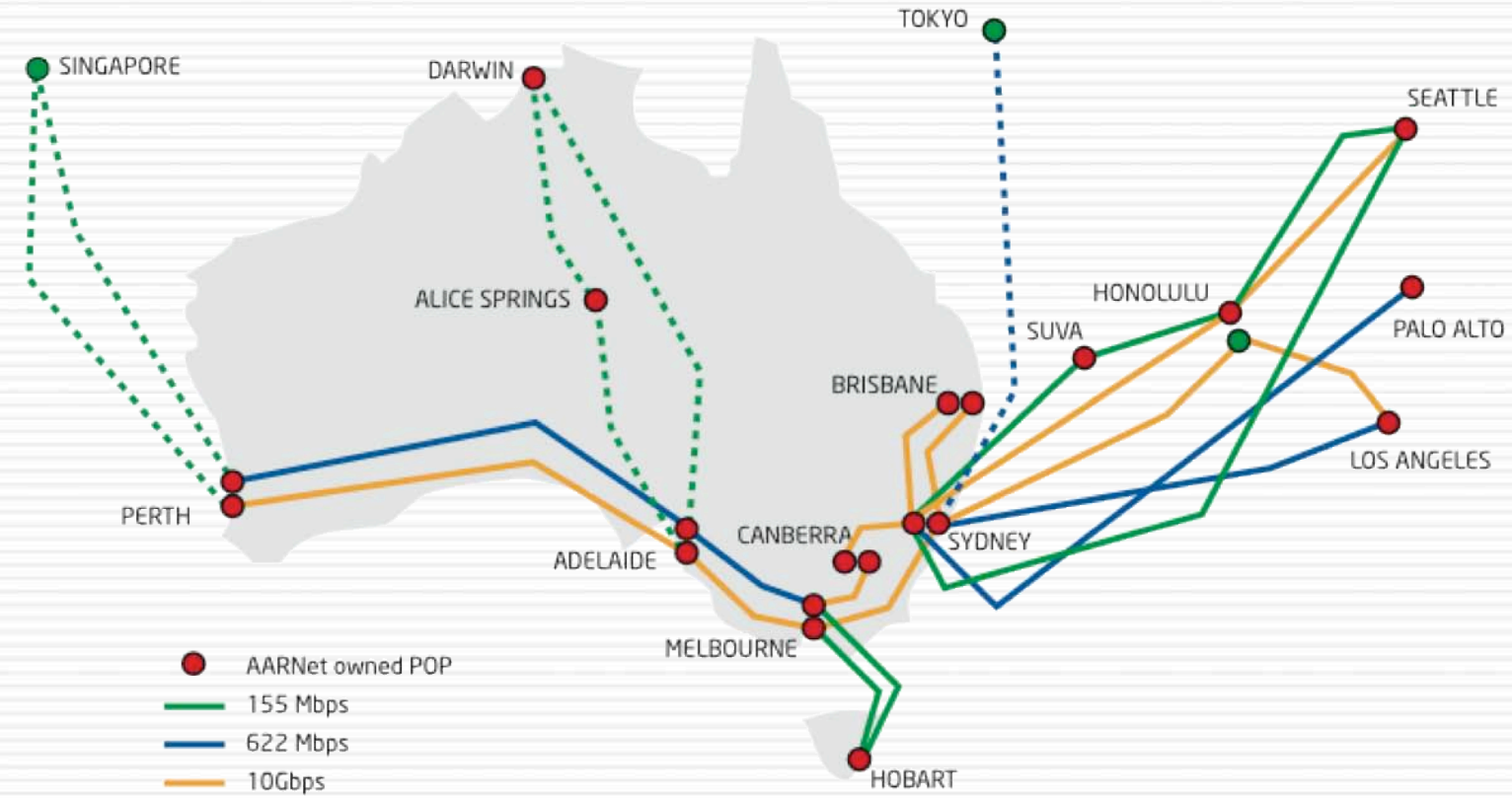
Equinix Peering Forum
Sydney - 26 July 2005

aarnet
Australia's Academic and Research Network

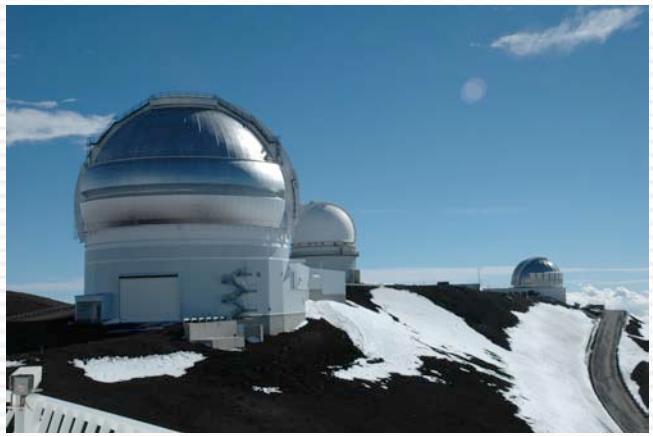
AARNet3 Network Highlights

- STM-64c (10Gbps) Backbone
- Dual STM-1 to NT & Tasmania
- Replacing Procket with Juniper M320
- Deploying DWDM from Adelaide to Brisbane
  - Providing multiple GigE to regional areas
  - Rolling our backbone onto our DWDM kit
- Multiple trans Pacific circuits
  - 2 x STM-64c for research and education
  - 2 x STM-4c (2 x 622Mbps) for commodity
  - 2 x STM-1
- Looking to expand footprint to Asia

AARNet3 Network



10G Trans Pacific



- Partnership with Southern Cross Cable Networks
- AUP - Research and Education only
- Dual STM-64c (OC192)
- Northern path to Seattle
  - Layer 3 routed
- Southern path to Los Angeles
  - Layer 1/2 - more later
- Catalyse Global Astronomy Initiative
  - Mauna Kea, Big Island

Why is a R&E network different?

- Peak demand can be driven by a single user driving a single application
- Interest in advanced services
  - IPv6
  - Voice, video, multicast
- Latency important to some but others more interested in bandwidth
- Need to build for peak demand
- So that means lots of “white space”
- But “Nature abhors a vacuum” …

IPv6

- Just plumbing but …
- Unlike GOSIP it needs to be taken seriously
- Important to Asia
- End-to-end principle
- First Australian IPv6 Summit
  - http://www.isoc-au.org.au/ipv6summit/
  - October 2005

Internet in China

- Internet users in China:
  from 6.8 Million to 78 Million within 6 months

Wireline	Dial Up	ISDN	Broadband
23.4M	45.0M	4.9M	9.8M

- Backbone: 2.5 - 10G DWDM
- International links: 20G
- Exchange Points: > 30G (BJ, SH, GZ)
- Last Mile: Ethernet, WLAN, ADSL, CTV, CDMA, ISDN, GPRS, Dial
- IP Addresses: 32M (1xA, 233xB, 146xC)

Source: J.P.Wu, APAN July 2004

How does AARNet support IPv6?

- AARNet3 is dual stack
- International transit/peers
- All customers can connect natively
- AARNet Migration Broker
  - http://broker.aarnet.net.au
  - Hexago appliance
  - Same as Freenet6
  - Open to anyone who can reach it via a domestic path

Divide Network Users into 3 Categories

- Cees de Laat classifies network users into 3 broad groups.
  1. Lightweight users, browsing, mailing, home use. Who need full Internet routing, one to many;
  2. Business applications, multicast, streaming, VPN's, mostly LAN. Who need VPN services and full Internet routing, several to several + uplink; and
  3. Scientific applications, distributed data processing, all sorts of grids. Who need very fat pipes, limited multiple Virtual Organizations, few to few, peer to peer.

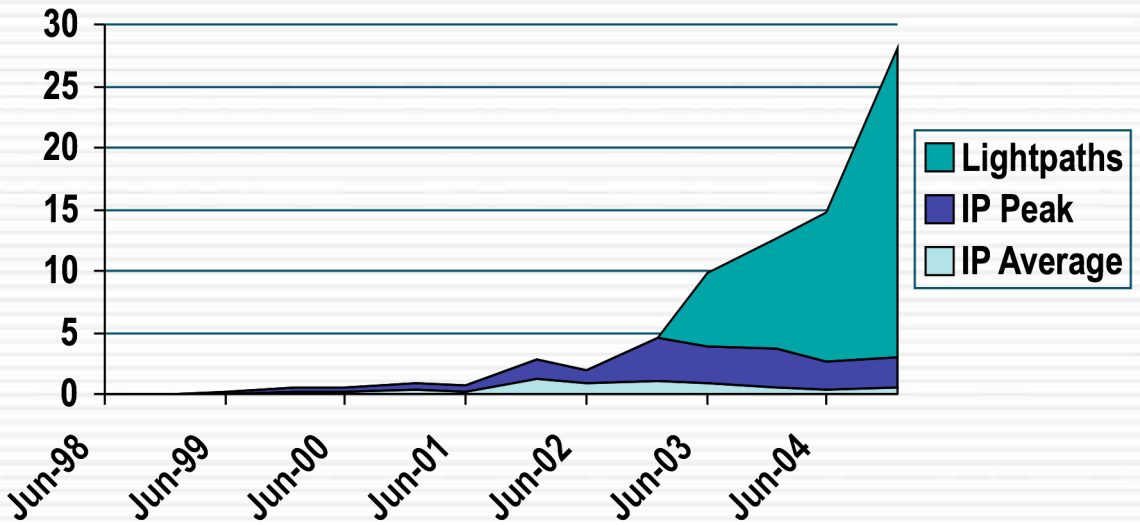
Who are these Type “3” Users?

- How many times can you say “CERN”?
- Astronomers, eVLBI
- Synchrotron
- Music Master Class
- High Definition TV over IP
- Massive data transfers from experiments running 24x7

Lightpaths for Massive data transfers

- From CANARIE

A small number of users with large data transfer needs can use more bandwidth than all other users



User Controlled Light Paths

- Techno speak for end user created dedicated Gigabit Ethernets
- Could be across the campus or across the world
- Various organisations working on creating the point and shoot interface

What is the GLIF?

- Global Lambda Infrastructure Facility
  - http://www.glif.is
- International virtual organization that supports persistent data-intensive scientific research and middleware development
- Provides ability to create dedicated international point to point Gigabit Ethernet circuits for “short term” experiments
- AARNet is Australia’s participant

“Big Science” projects driving networks

Billion dollar globally funded projects

Massive data transfer needs

- Large Hadron Collider
  - Coming on-stream in 2007
  - Particle collisions generating terabytes/second of “raw” data at a single, central, well-connected site
  - Need to transfer data to global “tier 1” sites. A tier 1 site must have (currently) a 10Gbps path to CERN
  - Tier 1 sites need to ensure gigabit capacity to the Tier2 sites they serve
- Square Kilometre Array
  - Coming on-stream in 2010?
  - Greater data generator than LHC
  - Up to 125 sites at remote locations, data need to be brought together for correlation
  - Many logistic issues to be addressed

Large Hadron Collider (LHC)

CMS

ATLAS

Per experiment

40 million collisions per second

- After filtering, 100 collisions of interest per second
- A Megabyte of digitised information for each collision = recording rate of 100 Megabytes/sec
- 1 billion collisions recorded = 1 Petabyte/year

1 Megabyte (1MB)
A digital photo

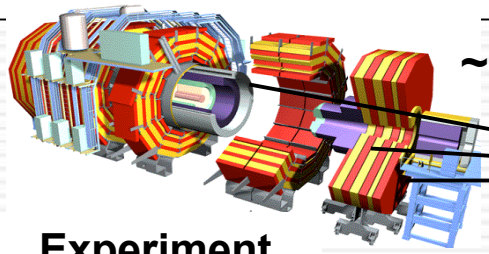
1 Gigabyte (1GB)
= 1000MB
A DVD movie

1 Terabyte (1TB)
= 1000GB
World annual book production

1 Petabyte (1PB)
= 1000TB
10% of the annual production by LHC experiments

1 Exabyte (1EB)
= 1000 PB
World annual information production

LHC Data Grid Hierarchy



Experiment

~PBytes/sec

Online System

CERN/Outside Resource Ratio ~1:2
Tier0/(Σ Tier1)/(Σ Tier2) ~1:1:1

~100-1500 MBytes/sec

Tier 0 +1

CERN Center
PBs of Disk;
Tape Robot

Tier 1

10-40 Gbps

IN2P3 Center

RAL Center

INFN Center

FNAL Center

...

~10 Gbps

Tier 2

Tier2 Center

Tier 3

~2.5-10 Gbps

Institute

Institute

Institute

Physics data cache

1 to 10 Gbps

Workstations

Tier 4

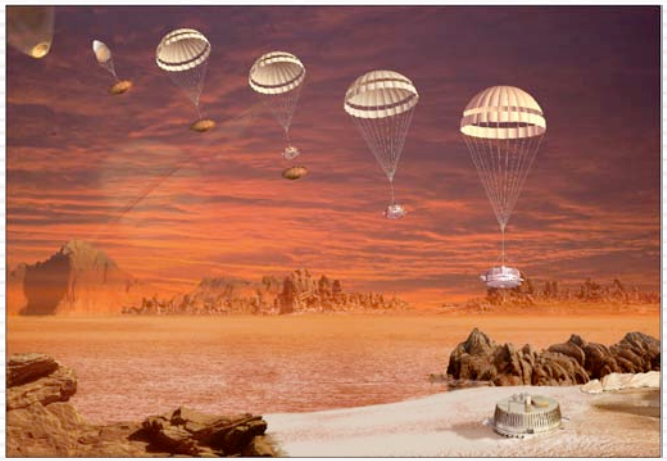
Tens of Petabytes by 2007-8.
An Exabyte ~5-7 Years later.

Source: H. Newman

High Definition TV over IP demo

- Joint AARNet & Research Channel demo for SC2004
  - Uncompressed signal
  - ~2million pixels per frame
  - 60 frames per second interleaved
  - using 1.5Gbps for each stream
- Used 10Gbps circuit from Canberra to Pittsburgh
- During the 30 hours of demonstration, 20 Terabytes of data were transmitted in each direction.
- No custom equipment involved, all off-the-shelf components

Huygens Space Probe



- Cassini spacecraft left Earth in October 1997 to travel to Saturn
- On Christmas Day 2004, the Huygens probe separated from Cassini
- Started it's descent through the dense atmosphere of Titan on 14 January 2005

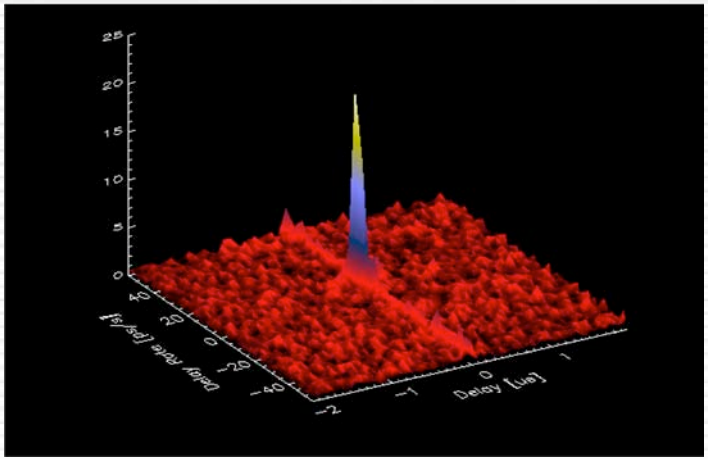
Tracking the Descent

- Very Long Baseline Interferometry (VLBI) is a technique where widely separated radio-telescopes observe the same region of the sky simultaneously to generate images of cosmic radio sources
- Using this technique 17 telescopes in Australia, China, Japan and the US were able to accurately position the probe to within a kilometre (Titan is ~1.5 billion kilometres from Earth)

Australian Contribution

- Created “dedicated” circuit
- The data from two of the Australian telescopes (Parkes [The Dish] & Mopra) was transferred via light plane to CSIRO Marsfield (Sydney)
- CeNTIE based fibre from CSIRO Marsfield to AARNet3 GigaPOP
- SXTransPORT 10G to Seattle
- “Lightpath” to Joint Institute for VLBI in Europe (JIVE) across CA*net4 and SURFnet optical infrastructure

Australian Contribution



VLBI Fringes

- The data was transferred at an average rate of 400Mbps
- 1Gbps path was available, TCP stack tuning important
- The data from these two telescopes were reformatted and correlated within hours of the end of the landing
- This early correlation allowed calibration of the data processor at JIVE, ready for the data from other telescopes to be added

And now for something completely different …

Peering and Transit in the USA

aarnet
Australia's Academic and Research Network

Schizophrenic AARNet

- Unlike many research and education networks AARNet also provides commodity Internet access
- AARNet connects to the commodity Internet via “Tier 1” providers on the US west coast and backhauls to Australia
- We think it provides more control but is it cheaper?

Rip off or not?

- The cost of a 10 gigabit circuit across the Atlantic between the USA east coast and Europe is US$18,000 per month
  - US$216,000 per annum
- The cost of a similar circuit between the USA west coast and Australia is US$2,500,000 per month
  - US$30,000,000 per annum!

DIY transit

- 2 Drop STM-4c IRU = US$11.28m
  - ~ US$63,000 per month
- Transit STM-1 commit x 4 ~ US$20k
- Backhaul x 2 ~ US$15k
- Colo and peering fabric ~ US$12k
- Routers ~ US$500k + US$2k p.m.
- Cost ~ US$250 per Mbps (at 500Mbps)

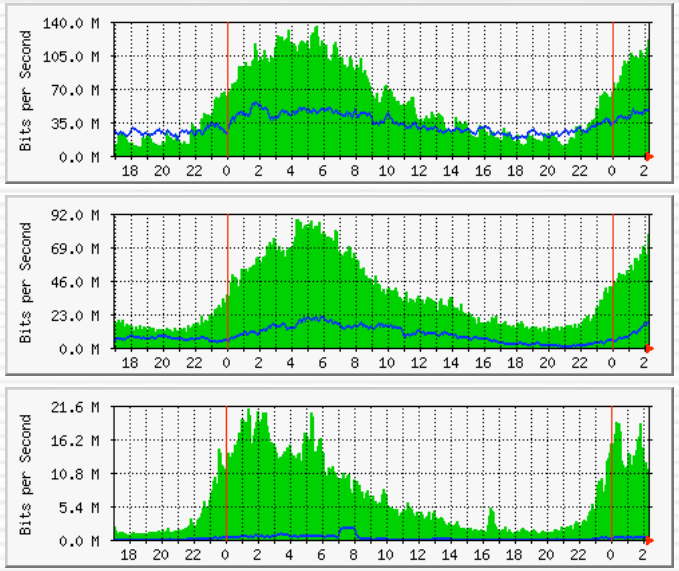
Looks good but …

- Plus
  - Control the quality of the transits
  - Marginal cost to take more transits low
  - Quality peering available but remember committed to 620Mbps transit anyway
- Minus
  - Asia/Pac content transits Pacific
  - Need to be able to fill the pipes
  - Hard to balance multiple transits
  - Need lots of cash to buy an IRU

Transit in Australia

- Maybe it's not so bad if …
- Australian transit provider has
  - A price close to DIY cost
  - Good links to Asia and Pacific region
  - Uses SCCN (fastest path) to USA
  - Has good transit relationships
  - Delivers Australia content locally
  - Provides the services you want

Peering in the USA



- Why bother at all? We're paying for 622M worth of transit anyway
- Better performance
- Ability to "do stuff", e.g. IPv6
- Slows the need to add more transit circuits when additional STM-4 pairs are added

Issues with Peering

- Large US ISPs want connections in multiple different time zones
  - Need presence on both US coasts
  - Hawai`i doesn't count :(
- Often also have minimum traffic volumes
- Less strict for IPv6, if supported
- Asian ISPs easier to deal with
  - Both out of market foreigners :)
  - Smaller European are probably OK too if you are on the east coast

How much does it cost?

- Hard to separate cost of transit from the cost of peering
- How do you apportion costs?
- But as an example consider locating a router at Equinix Los Angeles connected to our POP at Telehouse America only for the peering opportunities

What do we need?

- Link from Telehouse at 626 Wilshire Boulevard to Equinix 600 West 7th Street (about a block away)
- Collocation rack at Equinix and cross connect to a Gigabit Ethernet peering port
- Peering router at Equinix with at least two Gigabit Ethernet ports
- Additional router interface on the (Cisco 12404) router at Telehouse America

What do we need? (2)

	Capex	Opex per month
Metro fibre	US$ 750	US$ 500
Collocation rack	US$ 2,450	US$8,850
Router	US$15,000	US$ 450
Interface	US$ 9,000	US$ 200
Total to recover	US$27,200	US$10,000

So how much does peering cost?

- Assuming capex written off over 3 years we are spending US$10,750 per month for potentially 1Gbps
- Looks good but what if we can't fill the link?
- Need at least 250Mbps for it to break even against (expensive) transit
- Hard to see where an extra 250Mbps of peering traffic can come from out of 622Mbps

Where does AARNet peer?

- PAIX Palo Alto
- LAIIX/LAAP, Los Angeles
- Pacific Wave , Seattle
- Seattle Internet Exchange
- FIX West, NASA Ames
- WAIX, Perth
- TransACT, Canberra
- AARNet GigaPOP, Adelaide and UTS

What are we doing at Equinix, Sydney?

- Needed to build fibre between our POP at Rosebery and SCCN
- Built tail into Equinix since it was along the way
- Sell dark fibre between any of the locations
- Consider private peering