Australia-ATLAS
Running a site in the grid outback

Tim Dyce

tjdyce@unimelb.edu.au

The Australia-ATLAS site at Melbourne

The site was created in order to service the needs of our local high energy physics users at Melbourne & Sydney

ASGC is our TIER1

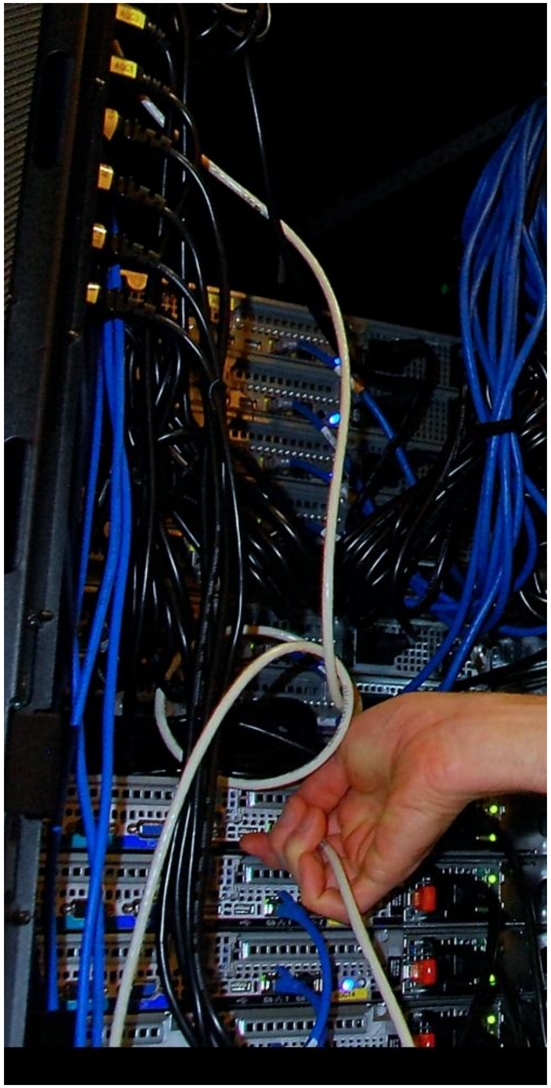
Hardware

TIER2 Hardware

- 80 Processors - 150 kSI2K
- 120TB Storage in disk via DPM
- We held off purchasing in 2008
- A large purchase later this year (at least doubling resources)

TIER3/TIER2.5 Hardware

- 16 Processors in a PBS queue
  - Used for short run jobs and testing potential TIER2 jobs
- A 5TB NAS serving a home area to the User Interface (UI)
- NAS backed up to tape via Tivoli Storage Manager
- 15TB of TIER2 storage reserved for local users



Virtualization

Virtualization

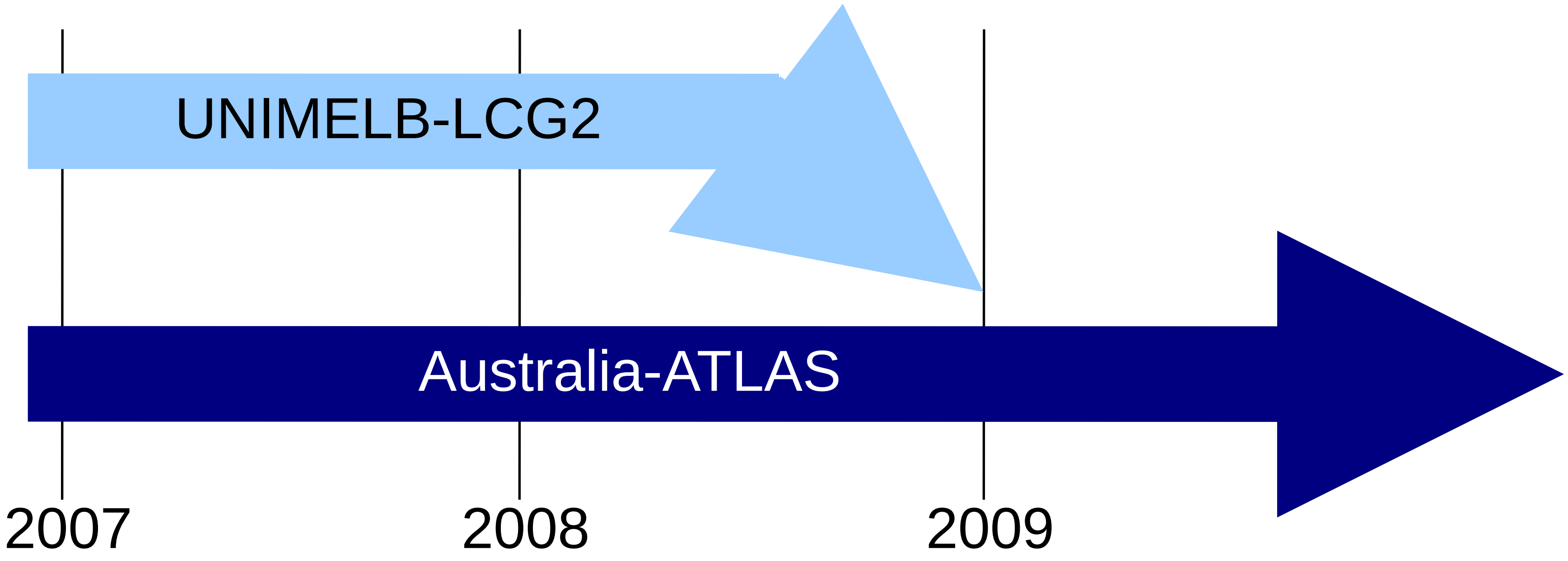
- We have used Xen to virtualize deployment and management components
- Virtual servers for:
    - PXE/TFTP server, Kickstart, Cfengine
    - DNS & DHCP
    - Repository mirrors
        - SL
        - GLite
        - Jpackage
        - Local extras
    - Syslog-ng (All nodes and servers log to a central syslog-ng server)
    - Ganglia
    - Backup (All SE/MON databases dumped twice daily)
- All virtual servers backup to tape via Tivoli Storage Manager

- We are setting up our new CREAM-CE as a virtual server
- Xen 1500 MTU restrictions may prevent us virtualizing some services...

Other grid sites at Melbourne

Our legacy site: UNIMELB–LCG2

- This site was initially set-up as a test-bed site for grid activities at Melbourne
- Maintaining two sites was an unnecessary duplication of effort
- These resources are being re-integrated as a CE within Australia-ATLAS



Network

Network is the largest challenge to ATLAS grid computing in Australia

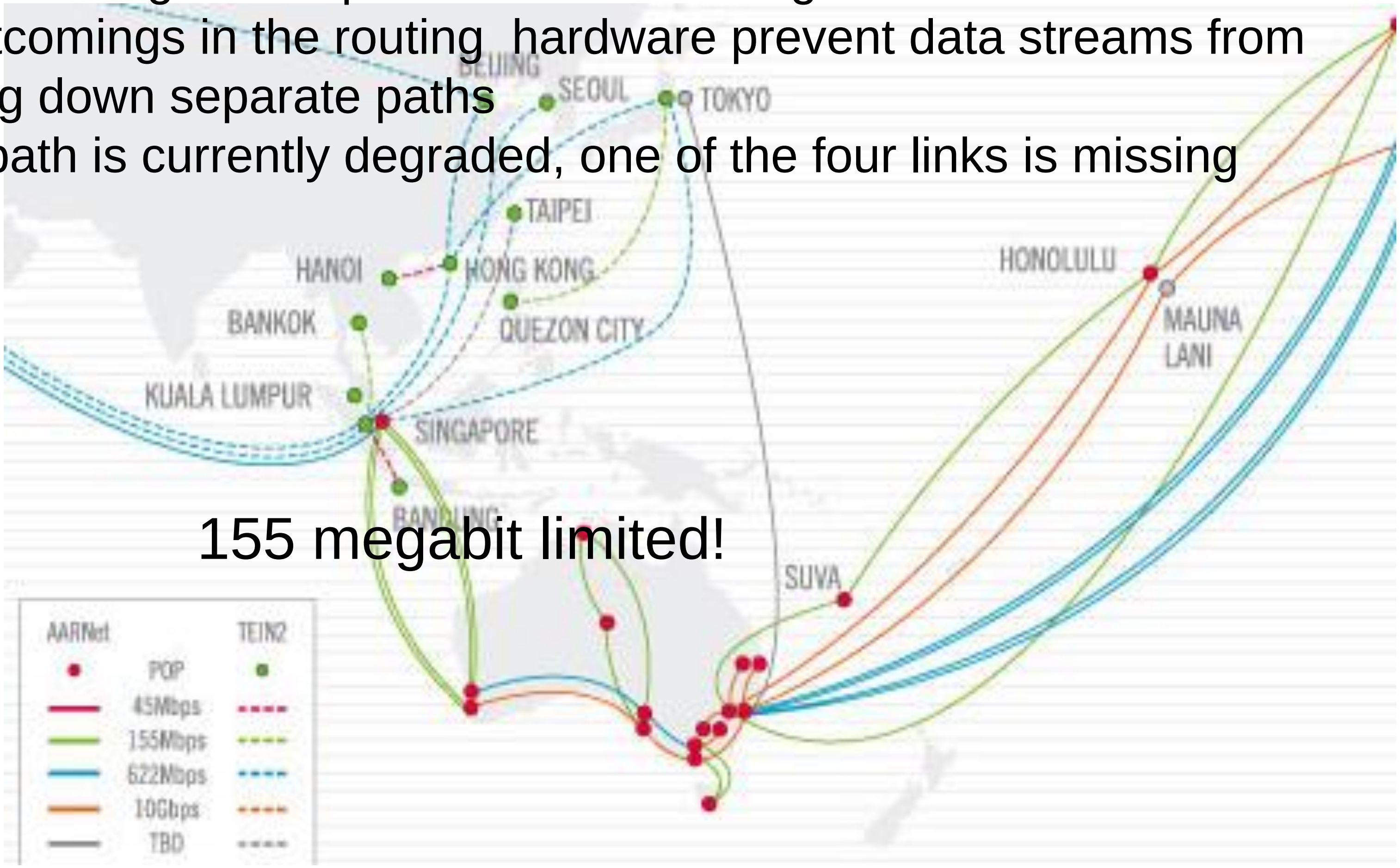
- We need sufficient speed transfer the large volumes of data required by the ATLAS project ~20MBytes/sec
- We need reliability

- At current we have two research network paths:
    - 620 megabit via Singapore
    - 1.2 gigabit via the U.S.

Network: Singapore path

The Singapore path

- The 620 megabit is split over four 155 megabit links
- Shortcomings in the routing hardware prevent data streams from passing down separate paths
- The path is currently degraded, one of the four links is missing

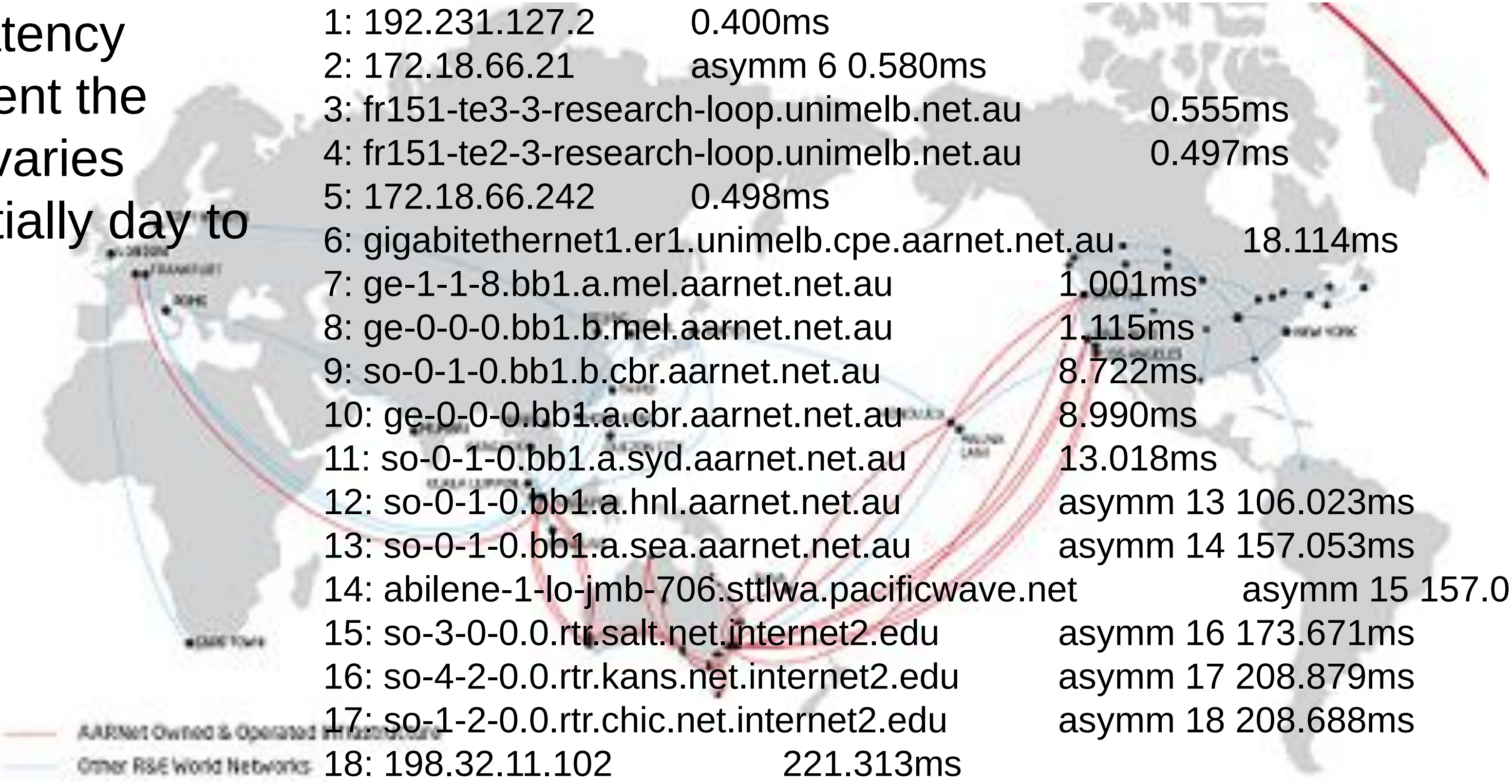


155 megabit limited!

Network: The U.S. path

The U.S. Path

- This path is long!
- High latency
- At current the latency varies substantially day to day



```
[root@agh5 ~]# tracepath w-fts.grid.sinica.edu.tw
1: agh5.atlas.unimelb.edu.au                    0.078ms pmtu 1500
1: 192.231.127.2          0.400ms
2: 172.18.66.21           asymm 6 0.580ms
3: fr151-te3-3-research-loop.unimelb.net.au               0.555ms
4: fr151-te2-3-research-loop.unimelb.net.au               0.497ms
5: 172.18.66.242          0.498ms
6: gigabitethernet1.er1.unimelb.cpe.aarnet.net.au              18.114ms
7: ge-1-1-8.bb1.a.mel.aarnet.net.au               1.001ms
8: ge-0-0-0.bb1.b.mel.aarnet.net.au               1.115ms
9: so-0-1-0.bb1.b.cbr.aarnet.net.au               8.722ms
10: ge-0-0-0.bb1.a.cbr.aarnet.net.au              8.990ms
11: so-0-1-0.bb1.a.syd.aarnet.net.au              13.018ms
12: so-0-1-0.bb1.a.hnl.aarnet.net.au              asymm 13 106.023ms
13: so-0-1-0.bb1.a.sea.aarnet.net.au              asymm 14 157.053ms
14: abilene-1-lo-jmb-706.sttlwa.pacificwave.net              asymm 15 157.071ms
15: so-3-0-0.0.rtr.salt.net.internet2.edu         asymm 16 173.671ms
16: so-4-2-0.0.rtr.kans.net.internet2.edu         asymm 17 208.879ms
17: so-1-2-0.0.rtr.chic.net.internet2.edu         asymm 18 208.688ms
18: 198.32.11.102               221.313ms
19: asgc-startap.r0.chi.asgc.net            209.220ms
20: so-1-0-0.r1.tpe.asgc.net          428.191ms
21: coresw.tpe.asgc.net               asymm 20 428.674ms
22: w-fts02.grid.sinica.edu.tw              asymm 19 428.172ms
Resume: pmtu
```

High latency networks and the Middleware/LFC

Fragile and high latency network links are a challenge to the Grid Middleware and project production systems

Issues encountered in Australia:

- The BDII can have severe issues on non-ideal networks
- DDM processes using the LHC File Catalogue (LFC) are slow
- Data transfers can be very slow on high latency networks

BDII issues

Behaviour of the BDII

- The current BDII purges and re-inserts all entries into it's database during an update cycle
- The top-level BDII will have less entries or even be empty if it fails to contact a remote site during a cycle
- If a site takes too long to respond it can be entered into the database late in the cycle. Sometimes this leads to insufficient “breathe time” for the database, especially on a busy machine; making the BDII appear empty when queried

Tuning the BDII to make it work for Australia-ATLAS

- Time out values for the information provider feeding the top-level BDII needed to be increased by up to 5 minutes
- Breathe times for the top-level BDII need to be increased by up to 3 minutes

Issues remain..

- Top-level BDII rotation time is long
- Though rare; we still observed empty top-level BDII instances, or missing data from remote sites due to minor network glitches

The new and improved BDII

We are testing the new BDII (version 5-20) at Melbourne

- Rewritten in Python by Laurence Field (CERN/EGEE)
- No longer empties and rotates databases, just updates the difference into the running database instance

Significant stability and performance enhancements...

- The V5 BDII copes much better when faced with large network latencies
- Start up and population time for the top-level BDII is now much lower, due to better forking during the search/update process.
- Eliminates previously seen “empty BDII” issues at Melbourne
- We are continuing to work with Laurence on the BDII
- Investigating using LDAP syncrepl to propagate changes between BDIIs

Issues with SAM GIS Sanity checks and the site BDII

- Currently the SAM BDII sanity & GSTAT checks look at the *createTimestamp* and *modifyTimestamp* fields in the site BDII. These are internal to the LDAP database, and do not necessarily change with every update under the new BDII
- This causes false “WARN: GIS has not updated in XXX minutes” SAM errors
    - The checks need to use the o=infosys section of the tree

DDM and LFC issues

Behaviour of Distributed Data Management

- Each dataset may consist of many files
- Every file which composes a dataset is registered into the LFC at Taiwan
- A replica entry for the dataset is entered into the ATLAS LFC at CERN

DDM LFC issues

- In the current version of DDM each file in a dataset is registered as individual operation
- Registering each file individually makes it more likely that a dataset registration process will fail
- This causes higher load on the LFC/Databases at Taiwan and CERN
- Causes additional delays at sites with high network latency

Australia-ATLAS local users

- Large network latencies make file registration and listing processes for our local users slow
    - lfc106.cern.ch: icmp_seq=5 ttl=42 time=322 ms
    - lfc.grid.sinica.edu.tw: icmp_seq=0 ttl=46 time=383 ms

LFC Bulk Registration

Bulk data registration in DDM

- The new version of DDM will uses bulk registration; registering all the files in the dataset into the LFC at once.
- Reduces time wasted in network transactions for each file
- Increases the chances of a dataset registration succeeding
- Makes local users happier when registering datasets

Australia-ATLAS local user listing issues remain..

- LFC bulk registration fixes DDM registration issues for our local users, but does not help with listing operations
- We have set-up daily cron jobs to dump a full listing of the LOCALGROUPDISK and ATLASDATADISK space tokens/sites
    - Allows users to search for files/datasets more easily (grep/less)
    - Does not help if they have specific queries, or are accessing datasets created on the same day
- Running a local LFC is not an option at current
- The users will just have to cope
    - We encourage them to script as much as possible

Data transfers: Tuning

Tuning the network to get the performance we need.

In order to get decent performance out of the network we need to do quite a lot of tuning

- Continually check and revise the best route path with our network provider, Australia's Academic Research Network (aarnet).
- Use 64 bit kernel disk servers to get around TCP window size limits within the kernel
- Tune TCP kernel parameters (YAIM's are very conservative for our network)
- Use jumbo frames (MTU 8000, still under way); delayed by routers on the path not being enabled for jumbo frames

THE UNIVERSITY OF MELBOURNE

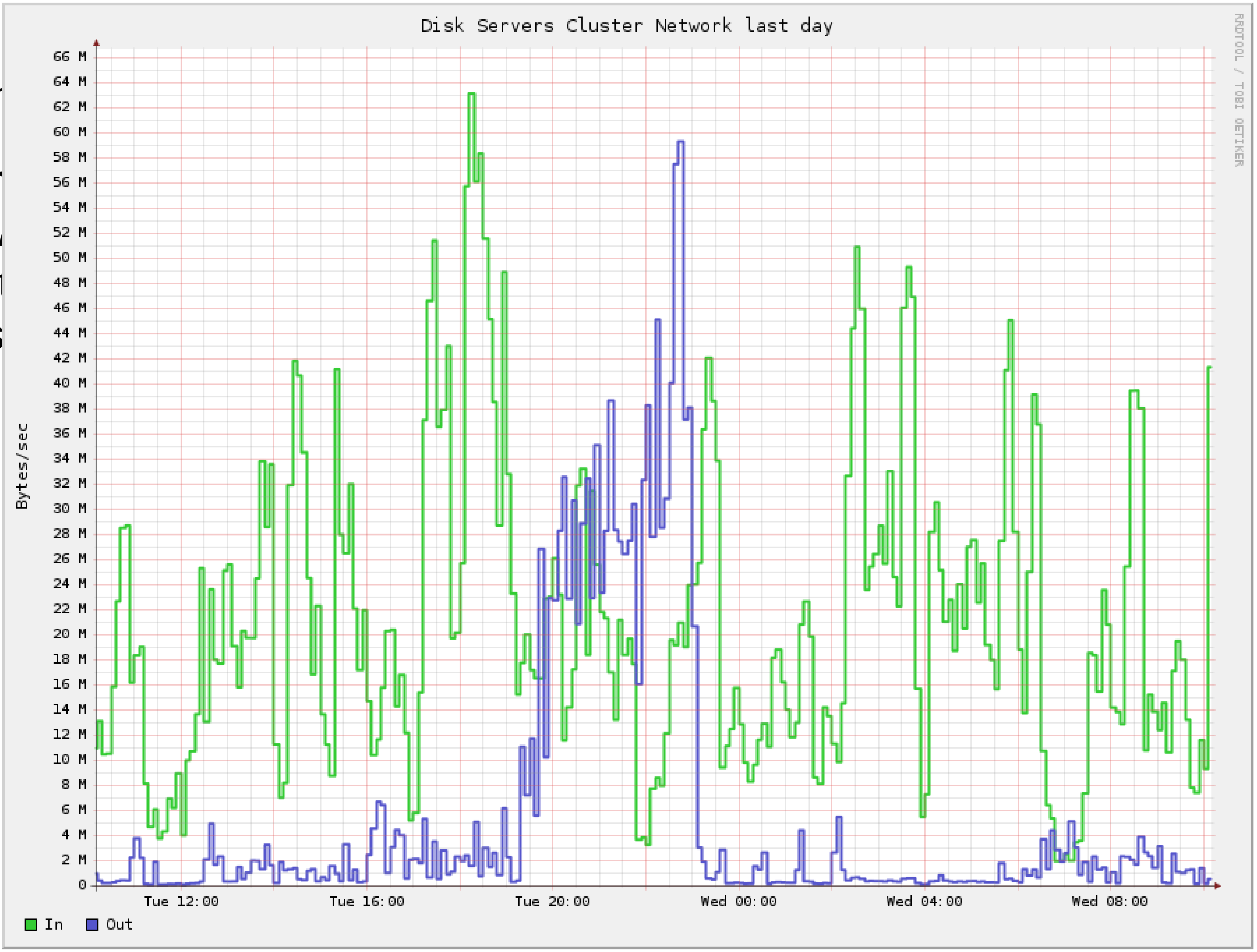
Data transfers: Testing the U.S. path

Speed

- Ensuring we have su

Testing the U.S. Path

- We have currently sv
- This allows us to test
as one of it's links has

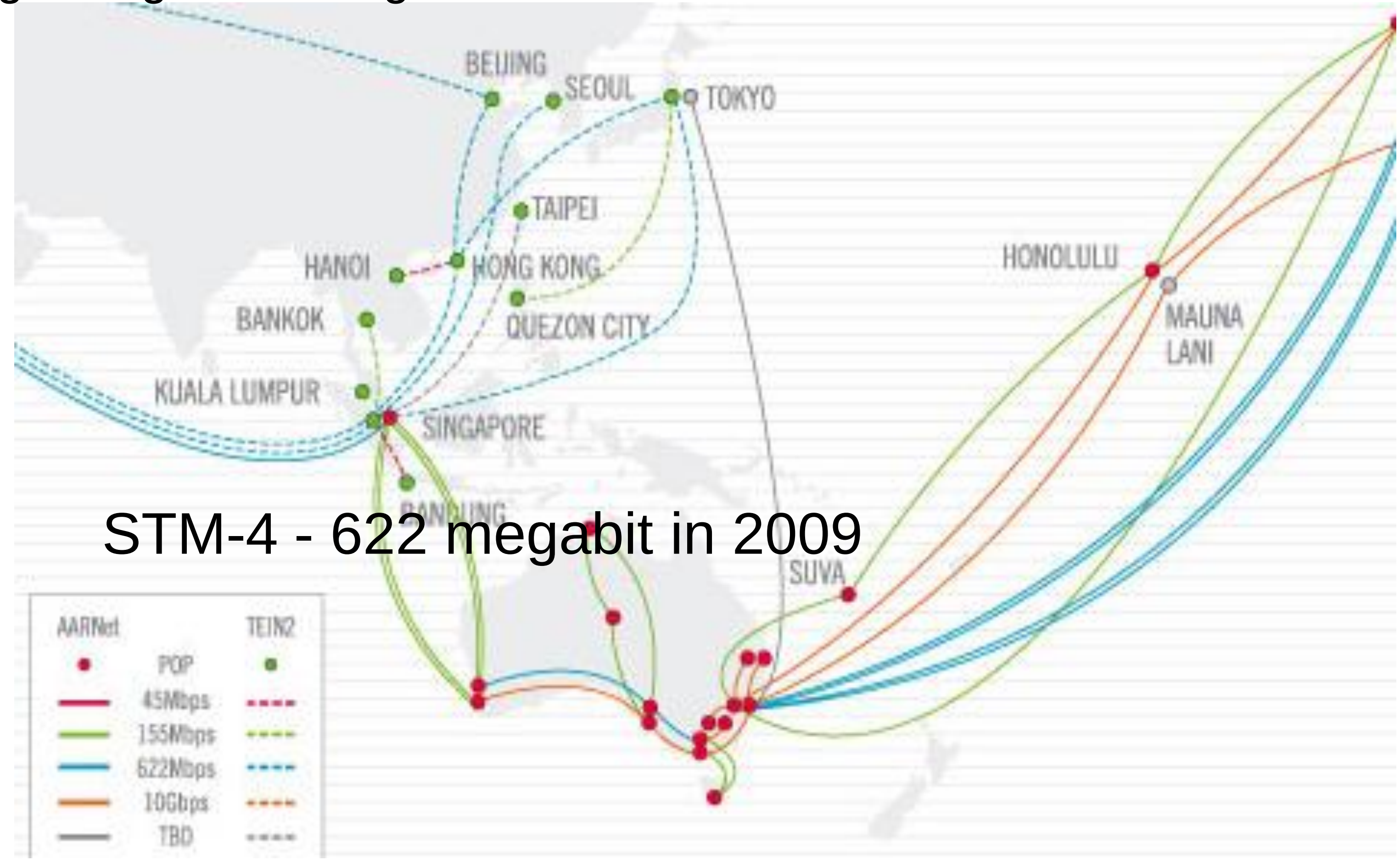


Incoming data rates as seen by disk servers during February 2009

~20MBytes/Sec is sustainable

Network: The light at the end of the tunnel

The hardware on one of the Singapore links is to be upgraded to STM-4, making a single 622 megabit link



STM-4 - 622 megabit in 2009

Network: The light at the end of the tunnel?

Singapore STM-4 upgrade solves the problem right?

It's a massive improvement, but not quite..

- Latency is still high; don't forget how far away Australia is!
- The network is still quite fragile, since we are relying on a single link over many hops
  - One idiot with a fishing trawler takes us off the air; sounds funny but it has happened!
- If we need to change paths, many changes need to be made; TCP windows etc.

Lots of work still to be done

- More thorough and continuous link testing (Iperf Sonar → Nagios)
- Test newly commissioned Singapore link
- We may be able to use the new PPC-1 link?
  - http://www.pipeinternational.com

THE UNIVERSITY OF MELBOURNE

Tim Dyce

tjdyce@unimelb.edu.au

Extra Slides...

Extra Slides...

Network: Tracepath for ASGC LFC

```
[root@agh5 ~]# tracepath lfc.grid.sinica.edu.tw
 1:  agh5.atlas.unimelb.edu.au (192.231.127.8)            0.101ms pmtu 1500
 1:  192.231.127.2 (192.231.127.2)                        0.461ms
 2:  172.18.66.21 (172.18.66.21)                         asymm  6   0.625ms
 3:  fr151-te3-3-research-loop.unimelb.net.au (172.19.1.162)   0.556ms
 4:  fr151-te2-3-research-loop.unimelb.net.au (172.19.1.161)   0.502ms
 5:  172.18.66.242 (172.18.66.242)                        0.528ms
 6:  gigabitethernet1.er1.unimelb.cpe.aarnet.net.au (202.158.200.249)  24.261ms
 7:  ge-1-1-8.bb1.a.mel.aarnet.net.au (202.158.200.245)     1.037ms
 8:  ge-0-0-0.bb1.b.mel.aarnet.net.au (202.158.194.182)     1.353ms
 9:  so-0-1-0.bb1.b.cbr.aarnet.net.au (202.158.194.30)     9.366ms
10:  ge-0-0-0.bb1.a.cbr.aarnet.net.au (202.158.194.201)     9.218ms
11:  so-0-1-0.bb1.a.syd.aarnet.net.au (202.158.194.42)     13.032ms
12:  so-0-1-0.bb1.a.hnl.aarnet.net.au (202.158.194.106)   asymm 13 105.946ms
13:  so-0-1-0.bb1.a.sea.aarnet.net.au (202.158.194.110)   asymm 14 157.110ms
14:  abilene-1-lo-jmb-706.sttlwa.pacificwave.net (207.231.240.8) asymm 15 157.299ms
15:  so-3-0-0.0.rtr.salt.net.internet2.edu (64.57.28.27)  asymm 16 173.579ms
16:  so-4-2-0.0.rtr.kans.net.internet2.edu (64.57.28.25)  asymm 17 198.166ms
17:  so-1-2-0.0.rtr.chic.net.internet2.edu (64.57.28.37)  asymm 18 208.701ms
18:  198.32.11.102 (198.32.11.102)                        209.573ms
19:  asgc-startap.r0.chi.asgc.net (117.103.111.146)       209.613ms
20:  so-1-0-0.r1.tpe.asgc.net (117.103.111.210)          383.509ms
21:  117.103.111.229 (117.103.111.229)                 asymm 20 384.892ms
22:  lfc.grid.sinica.edu.tw (117.103.103.48)           asymm 19 383.673ms reached
```

Network: Tracepath for CERN LFC

```
[root@agh5 ~]# tracepath prod-lfc-atlas-central.cern.ch
 1:  agh5.atlas.unimelb.edu.au (192.231.127.8)           0.108ms pmtu 1500
 1:  192.231.127.2 (192.231.127.2)                        0.417ms
 2:  172.18.66.21 (172.18.66.21)                        asymm  6   1.009ms
 3:  fr151-te3-3-research-loop.unimelb.net.au (172.19.1.162)   0.561ms
 4:  fr151-te2-3-research-loop.unimelb.net.au (172.19.1.161)   0.504ms
 5:  172.18.66.242 (172.18.66.242)                        0.548ms
 6:  gigabitethernet1.er1.unimelb.cpe.aarnet.net.au (202.158.200.249)   0.578ms
 7:  ge-1-1-8.bb1.a.mel.aarnet.net.au (202.158.200.245)     0.985ms
 8:  ge-0-0-0.bb1.b.mel.aarnet.net.au (202.158.194.182)     1.081ms
 9:  so-0-1-0.bb1.b.cbr.aarnet.net.au (202.158.194.30)      8.772ms
10:  ge-0-0-0.bb1.a.cbr.aarnet.net.au (202.158.194.201)     9.010ms
11:  so-0-1-0.bb1.a.syd.aarnet.net.au (202.158.194.42)     12.930ms
12:  so-0-1-0.bb1.a.hnl.aarnet.net.au (202.158.194.106)   asymm 13 105.980ms
13:  so-0-1-0.bb1.a.sea.aarnet.net.au (202.158.194.110)   asymm 14 166.863ms
14:  abilene-1-lo-jmb-706.sttlwa.pacificwave.net (207.231.240.8) asymm 15 157.144ms
15:  so-3-0-0.0.rtr.salt.net.internet2.edu (64.57.28.27)  asymm 16 173.620ms
16:  so-4-2-0.0.rtr.kans.net.internet2.edu (64.57.28.25)  asymm 17 198.150ms
17:  so-1-2-0.0.rtr.chic.net.internet2.edu (64.57.28.37)  asymm 18 208.776ms
18:  e513-e-rci76-2-te6.cern.ch (192.91.246.126)          321.688ms
19:  e513-e-rci76-1-ne1.cern.ch (192.65.184.53)           321.795ms
```

THE UNIVERSITY OF MELBOURNE

Network: Pings for CERN & CERN LFC's

```
[root@agh5 ~]# ping -c 5 lfc.grid.sinica.edu.tw
PING lfc.grid.sinica.edu.tw (117.103.103.48) 56(84) bytes of data.
64 bytes from lfc.grid.sinica.edu.tw (117.103.103.48): icmp_seq=0 ttl=46 time=383 ms
64 bytes from lfc.grid.sinica.edu.tw (117.103.103.48): icmp_seq=1 ttl=46 time=383 ms
64 bytes from lfc.grid.sinica.edu.tw (117.103.103.48): icmp_seq=2 ttl=46 time=383 ms
64 bytes from lfc.grid.sinica.edu.tw (117.103.103.48): icmp_seq=3 ttl=46 time=383 ms
64 bytes from lfc.grid.sinica.edu.tw (117.103.103.48): icmp_seq=4 ttl=46 time=383 ms

--- lfc.grid.sinica.edu.tw ping statistics ---
5 packets transmitted, 5 received, 0% packet loss, time 4005ms
rtt min/avg/max/mdev = 383.382/383.402/383.434/0.392 ms, pipe 2
[root@agh5 ~]# ping -c 5 prod-lfc-atlas-central.cern.ch
PING prod-lfc-atlas-central.cern.ch (128.142.173.141) 56(84) bytes of data.
64 bytes from lfc106.cern.ch (128.142.173.141): icmp_seq=0 ttl=42 time=324 ms
64 bytes from lfc106.cern.ch (128.142.173.141): icmp_seq=1 ttl=42 time=323 ms
64 bytes from lfc106.cern.ch (128.142.173.141): icmp_seq=2 ttl=42 time=323 ms
64 bytes from lfc106.cern.ch (128.142.173.141): icmp_seq=3 ttl=42 time=322 ms
64 bytes from lfc106.cern.ch (128.142.173.141): icmp_seq=4 ttl=42 time=323 ms

--- prod-lfc-atlas-central.cern.ch ping statistics ---
5 packets transmitted, 5 received, 0% packet loss, time 4006ms
rtt min/avg/max/mdev = 322.309/323.721/324.953/1.126 ms, pipe 2
```