UNIVERSITY OF LIVERPOOL

Liverpool HEP - Site Report

June 2010

John Bland, Robert Fay

Staff Status

No changes to technical staff since last year:

Two full time HEP system administrators

- John Bland, Robert Fay

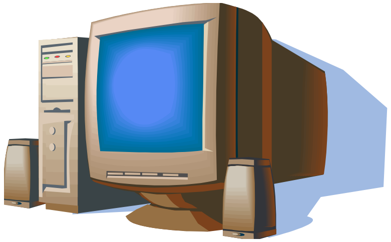
One full time Grid administrator (0.75FTE)

- Steve Jones, started September 2008

Not enough!

Mike Houlden has retired and David Hutchcroft has taken over as the academic in charge of computing.

Current Hardware - Users



Desktops

- ~100 Desktops: Scientific Linux 4.3, Windows XP, Legacy systems
- Minimum spec of 3GHz ‘P4’, 1GB RAM + TFT Monitor
- Hardware upgrades this Summer, SL5.x, Windows 7

Laptops

- ~60 Laptops: Mixed architecture, Windows+VM, MacOS+VM, Netbooks

Printers

- Samsung and Brother desktop printers
- Various OKI and HP model group printers
- Recently replaced aging HP LaserJet 4200 with HP LaserJet P4015X

Current Hardware – ‘Tier 3’ Batch

‘Tier3’ Batch Farm

- Software repository (0.5TB), storage (3TB scratch, 13TB bulk)
- ‘medium32’, ‘short’ queues consist of 40 32bit SL4 (3GHz P4, 1GB/core)
- ‘medium64’, ‘short64’ queues consist of 9 64bit SL5 nodes (2xL5420, 2GB/core)
- 2 of the 9 SL5 nodes can also be used interactively
- 5 older interactive nodes (dual 32bit Xeon 2.4GHz, 2GB/core)
- Using Torque/PBS/Maui+Fairshares
- Used for general, short analysis jobs
- Grid jobs now also run opportunistically on this cluster

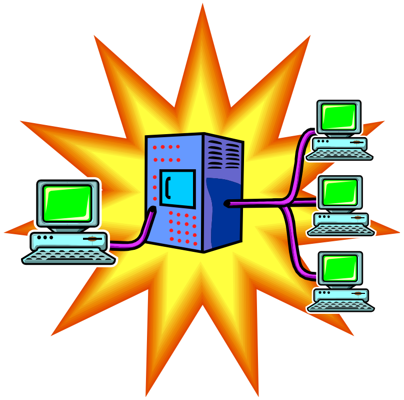
Current Hardware – Servers

- ~40 core servers (HEP+Tier2)
- ~60 Gigabit switches
- 1 High density Force10 switch
- Console access via KVMoIP (when it works)

LCG Servers

- SE 8-core Xeon 2.66GHz, 10GB RAM, Raid 10 array
  - Unstable under SL4, crashes triggered by mysqldumps
  - Temporarily replaced with alternate hardware
  - Testing shows it appears to be stable under SL5
- CEs, SE, UI, MON all SL4, gLite 3.1
- BDII SL5, gLite 3.2
- VMware Server being used for some servers and for testing
  - MON, BDII, CE+Torque for 64-bit cluster, CREAM CE, all VMs

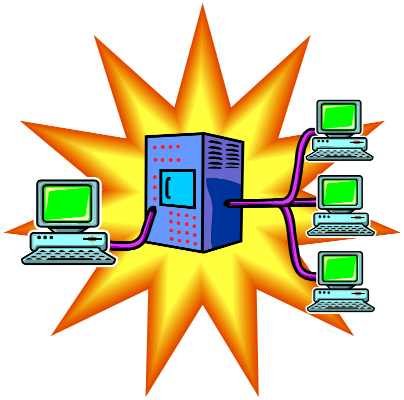
Current Hardware – Nodes



MAP2 Cluster

- Still going!
- Originally 24 rack (960 node) (Dell PowerEdge 650) cluster
- Nodes are 3GHz P4, 1-1.5GB RAM, 120GB disk – 5.32 HEPSPEC06
- 2 racks (80 nodes) shared with other departments
- 18 racks (~700 nodes) primarily for LCG jobs
- 1 rack (40 nodes) for general purpose local batch processing
- 3 racks retired (Dell nodes replaced with other hardware)
- Each rack has two 24 port gigabit switches, 2Gbit/s uplink
- All racks connected into VLANs via Force10 managed switch/router
    - 1 VLAN/rack, all traffic Layer3
- Still repairing/retiring broken nodes on a weekly basis
- But…

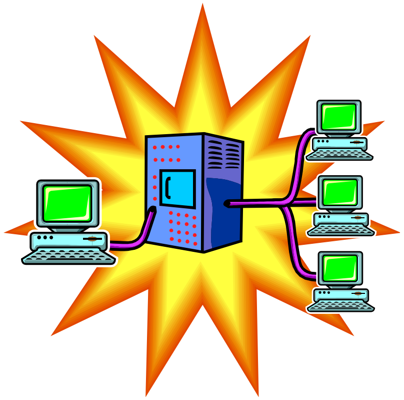
Current Hardware – Nodes



MAP2 Cluster (continued)

- Its days are hopefully numbered
- Internal agreement to fund replacement from energy savings
- Proposed replacement will be 72 E5620 CPUs (288 cores) or equivalent
- Power consumption will go from up to ~140kW to ~10.5kW (peak)

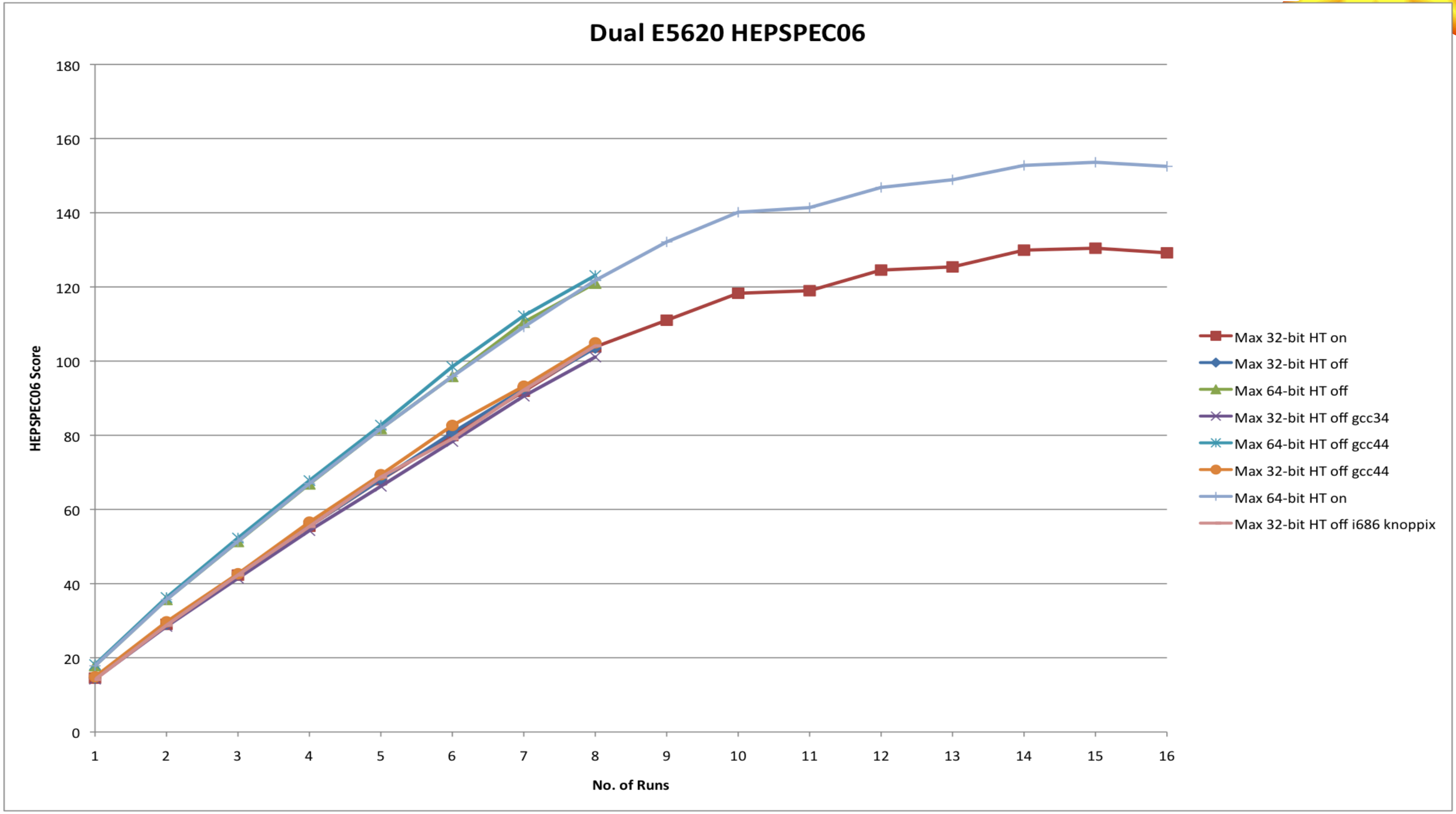
Current Hardware – Nodes



New cluster

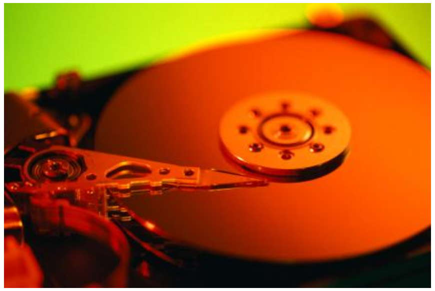
- Started with 7 dual L5420 nodes (56 cores with HEPSPEC06 8.01)
- With last round of GridPP funding added 7 of:
  - SuperMicro SYS-6026TT-TF quad-board 2U chassis
  - Dual 1400W redundant PSUs
  - 4 x SuperMicro X8DDT-F motherboards
  - 8 x Intel E5620 Xeon CPUs
  - 96GB RAM
  - 8 x 1TB Enterprise SATA drive
  - 224 cores total

Current Hardware – Nodes



UNIVERSITY OF LIVERPOOL

Storage

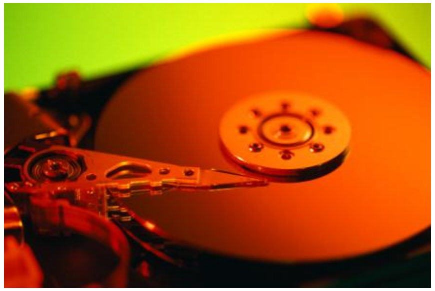


RAID

- Majority of file stores using RAID6. Few legacy RAID5+HS.
- Mix of 3ware and Areca SATA controllers
- Adaptec SAS controller for grid software.
- Scientific Linux 4.3, newer systems on SL5.x
- Arrays monitored with 3ware/Areca web software and email alerts
- Now tied in with Nagios as well
- Software RAID1 system disks on all new servers/nodes.

UNIVERSITY OF LIVERPOOL

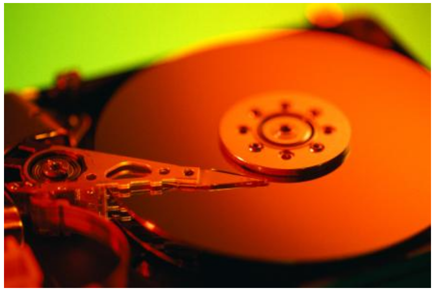
Storage



File stores

- 13TB general purpose ‘hepstore’ for bulk storage
- 3TB scratch area for batch output (RAID10)
- 2.5TB high performance SAS array for grid/local software
- Sundry servers for user/server backups, group storage etc
- 270TB RAID6 for LCG storage element (Tier2 + Tier3 storage/access via RFIO/GridFTP)

Storage - Grid



- Now running head node + 12 DPM pool nodes, ~270TB of storage
  - This is combined LCG and local data
  - Using DPM as ‘cluster filesystem’ for Tier2 and Tier3
  - Local ATLASLIVERPOOLDISK space token
- Still watching Lustre-based SEs

UNIVERSITY OF LIVERPOOL

Joining Clusters

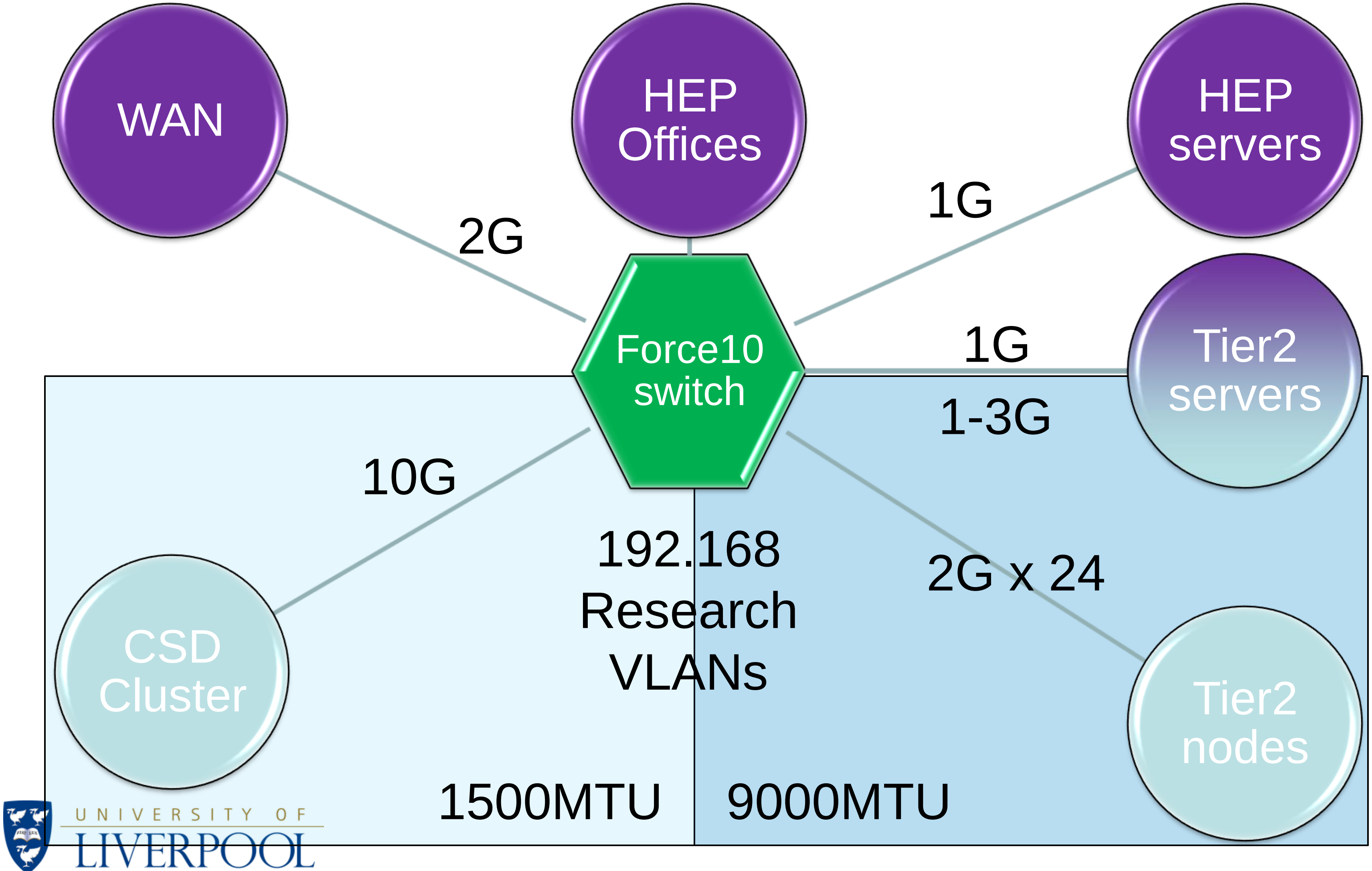
- High performance central Liverpool Computing Services Department (CSD) cluster
- Physics has a share of the CPU time
- Decided to use it as a second cluster for the Liverpool Tier2
  - Extra cost was second CE node (£2k)
  - Plus line rental for 10Gb fibre between machine rooms
- Liverpool HEP attained NGS Associate status
  - Can take NGS-submitted jobs from traditional CSD serial job users
  - Sharing load across both clusters more efficiently
- Compute cluster in CSD, Service/Storage nodes in Physics

Joining Clusters

- CSD nodes
  - Dual quad-core AMD Opteron 2356 CPUs, 16GB RAM
  - HEPSPEC06 7.84
  - OS was SuSE10.3, moved to RHEL5 in February
  - Using tarball WN + extra software on NFS (no root access to nodes)
- Needed a high performance central software server
  - Using SAS 15K drives and 10G link
  - Capacity upgrade required for local systems anyway (ATLAS!)
  - Copes very well with ~800nodes apart from jobs that compile code
    - NFS overhead on file lookup is the major bottleneck
- Very close to going live once network troubles sorted
  - Relying on remote administration frustrating at times
  - CSD also short-staffed, struggling with hardware issues on the new cluster

UNIVERSITY OF LIVERPOOL

HEP Network topology



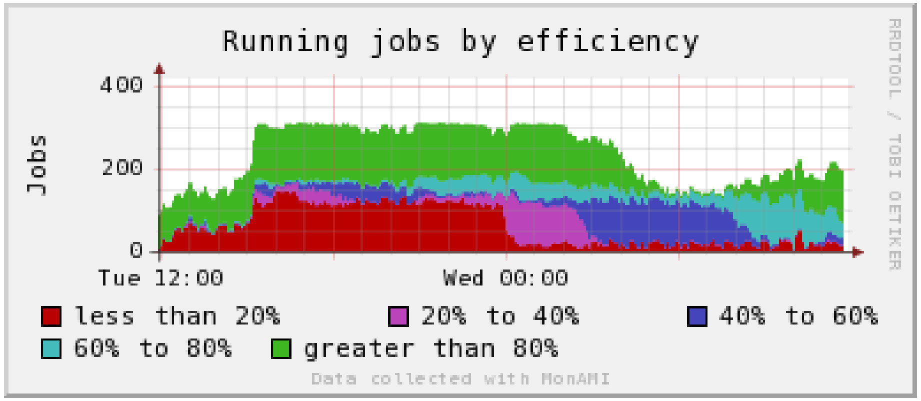
Configuration and deployment

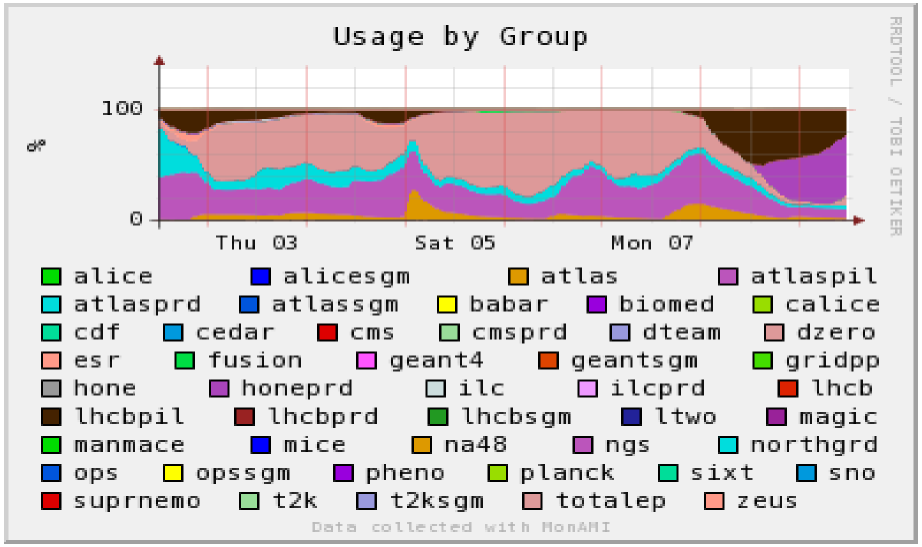
- Kickstart used for OS installation and basic post install
  - Previously used for desktops only
  - Now used with PXE boot for automated grid node installs
- Puppet used for post-kickstart node installation (glite-WN, YAIM etc)
  - Also used for keeping systems up to date and rolling out packages
  - And used on desktops for software and mount points
- Custom local testnode script to periodically check node health and software status
  - Nodes put offline/online automatically
- Keep local YUM repo mirrors, updated when required, no surprise updates (being careful of gLite generic repos)

Monitoring

- Ganglia on all worker nodes and servers
- Use monami with ganglia on CE, SE and pool nodes
  - Torque/Maui stats, DPM/MySQL stats, RFIO/GridFTP connections
- Nagios monitoring all servers and nodes
  - Continually increasing number of service checks
  - Increasing number of local scripts and hacks for alerts and ticketing
- Cacti used to monitor building switches
  - Throughput and error readings
- Ntop monitors core Force10 switch, but still unreliable
  - sFlowTrend tracks total throughput and biggest users, stable
- LanTopolog tracks MAC addresses and building network topology
- arpwatch monitors ARP traffic (changing IP/MAC address pairings).

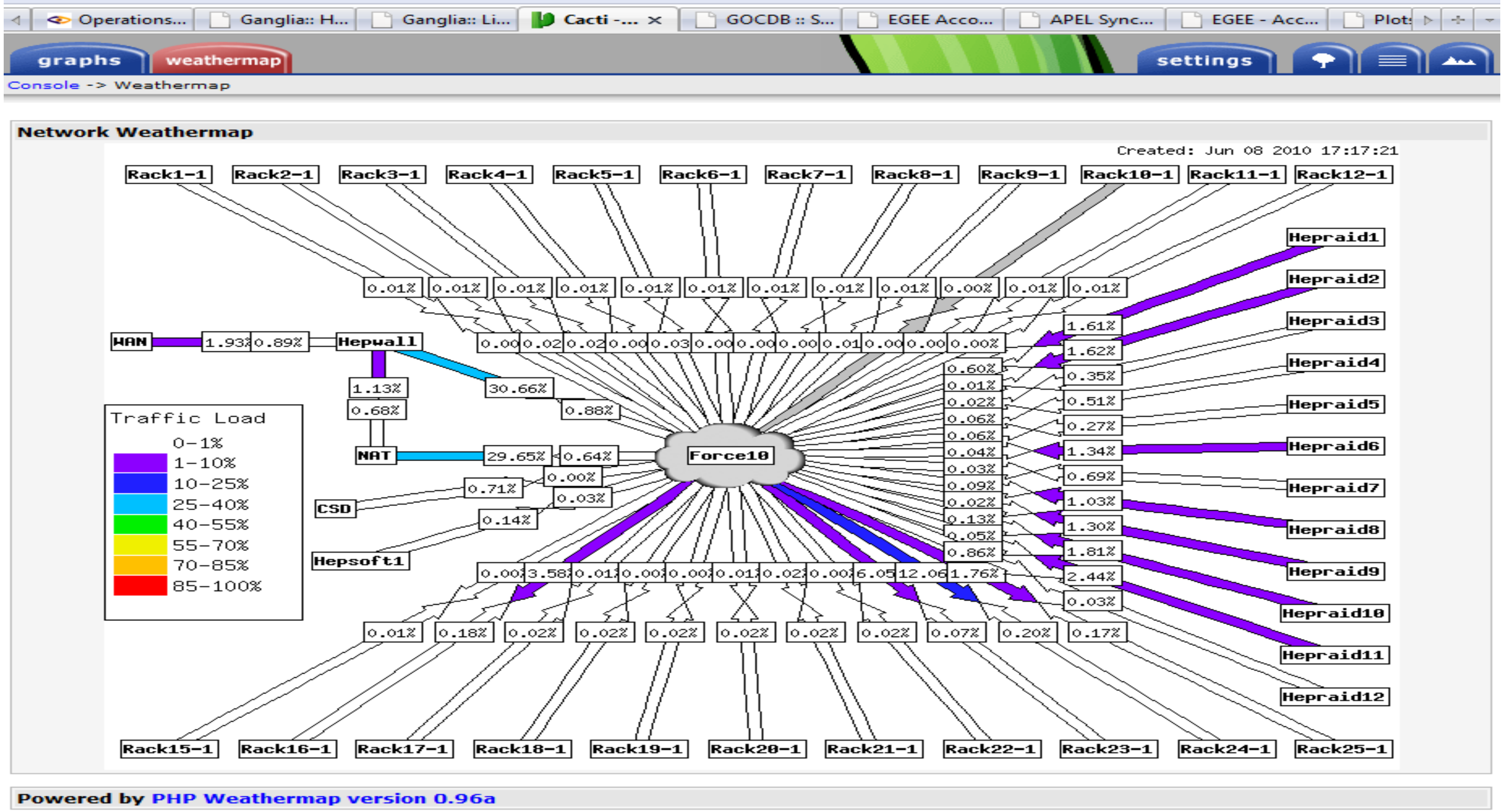
Monitoring - Monami





Monitoring - Cacti

- Cacti Weathermap



Network Weathermap

Created: Jun 08 2010 17:17:21

Traffic Load	
	0-1%
🟪	1-10%
🟦	10-25%
🟦	25-40%
🟩	40-55%
🟨	55-70%
🟧	70-85%
🟥	85-100%

Powered by PHP Weathermap version 0.96a

UNIVERSITY OF LIVERPOOL

Security

- Network security
  - University firewall filters off-campus traffic
  - Local HEP firewalls to filter on-campus traffic
  - Monitoring of LAN devices (and blocking of MAC addresses on switch)
  - Single SSH gateway, Denyhosts
  - Snort and BASE (need to refine rules to be useful, too many alerts)
- Physical security
  - Secure cluster room with swipe card access
  - Laptop cable locks (some laptops stolen from building in past)
  - Promoting use of encryption for sensitive data
  - Parts of HEP building publically accessible
- Logging
  - Server system logs backed up daily, stored for 1 year
  - Auditing logged MAC addresses to find rogue devices

Plans and Issues

- Replace MAP-2
  - Installation of new nodes shouldn’t be a problem
  - Getting rid of several hundred Dell PowerEdge 650s more of a challenge
  - Still need to think of a name for the new cluster
- Possibility of rewiring the network in the Physics building
  - Computing Services want to rewire and take over network management for offices
  - But there’s asbestos in the building so maybe they don’t
- IPv6
  - IANA pool of IPv4 addresses predicted to be exhausted by late 2010 / early 2011
  - Need to be ready to switch at some point…
  - Would remove any NAT issues!

Conclusion

- * New kit in, older kit soon (?) to be replaced
- We’re just about keeping on top of things with the resources we have

UNIVERSITY OF LIVERPOOL