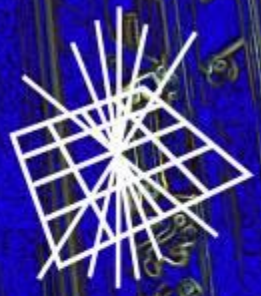


GridPP

UK Computing for Particle Physics

RAL Site Report

HEPiX Spring 2015 – Oxford

23-27 March 2015

Martin Bly, STFC-RAL

Outline

- Intro
- Hardware/Tapes
- Software/systems
- Networking
- JASMIN

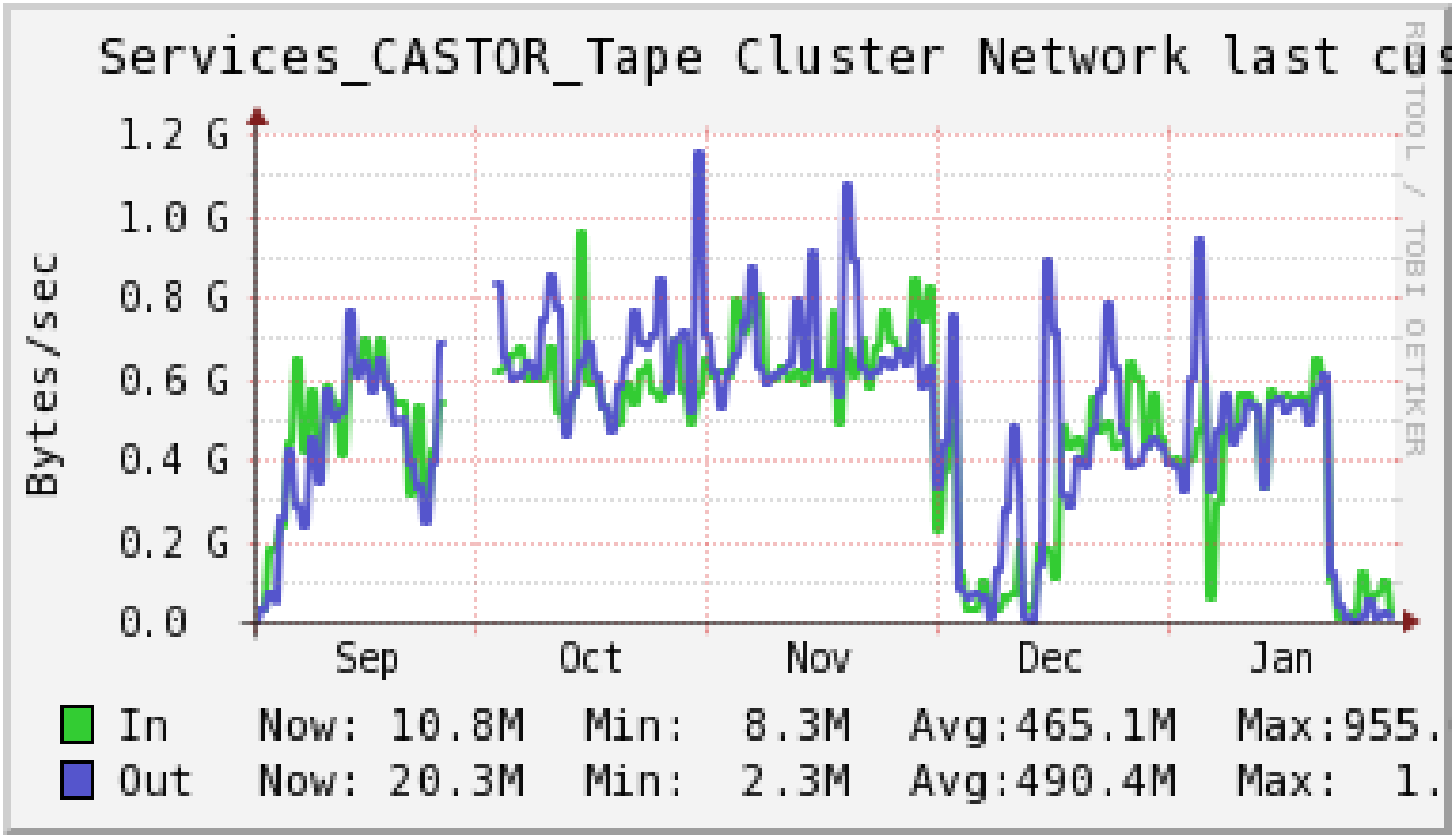
Rutherford Appleton Lab

- 15 miles south of Oxford on Harwell Campus
- Run by STFC
- Multi-discipline centre supporting university and industrial research in big facilities: Neutron Science, Lasers, Space Science, Computing
- Hosts UK LHC Tier1 Facility

Tier1 Hardware

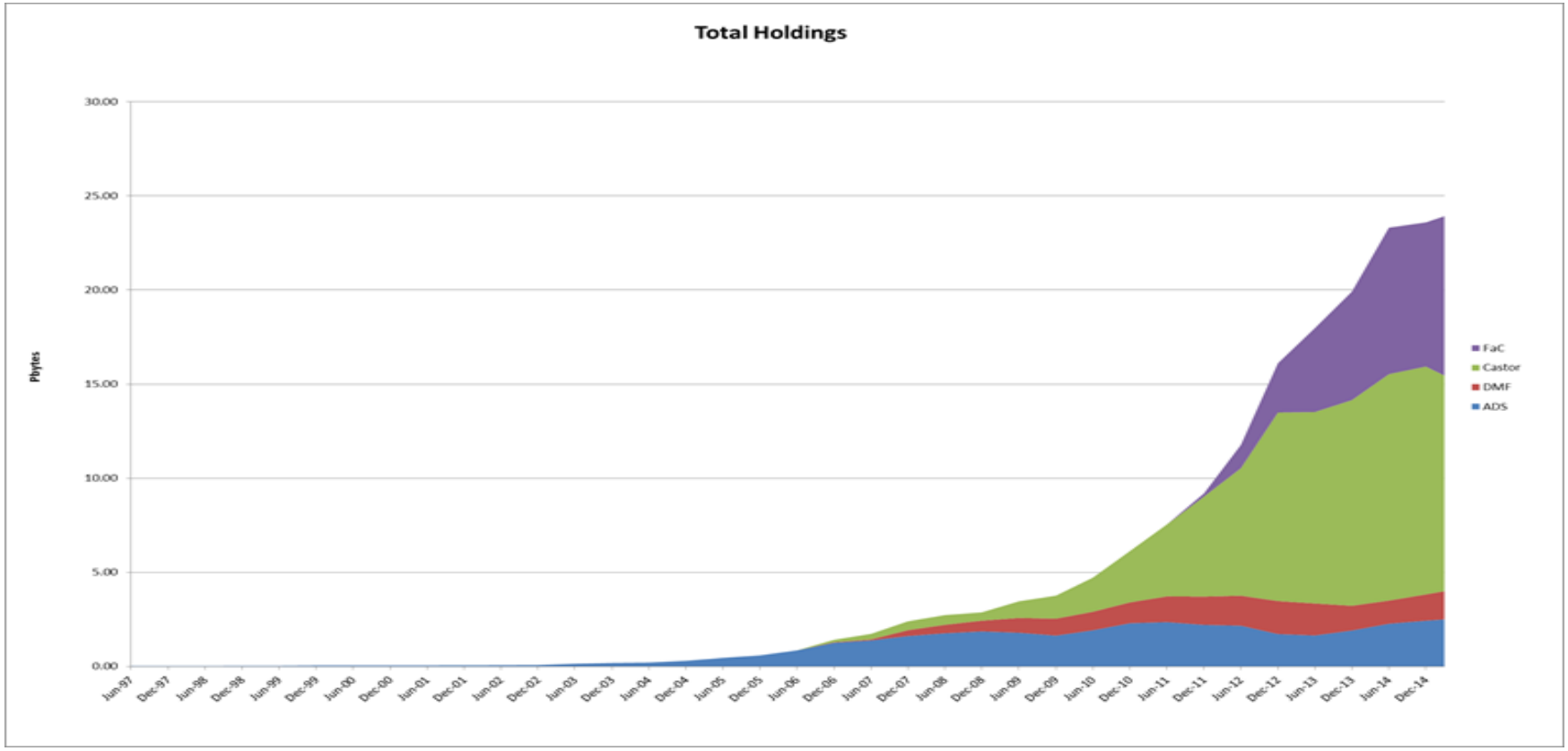
- CPU: ~119k HS06 (~10.6k cores)
  - FY 14/15: additional ~42kHS06 – delivered, in test
  - E5-2640v3 and E5-2650v2 (Fujitsu, Supermicro, 4 sleds/2U)
    - First WNs with 10GbE NICs
- Storage: ~13PB disk
  - TY 14/15: additional ~5.2PB – delivery and testing ongoing
  - Standard 'Castor' spec (SATA-in-a-box) +
    - SSD for CEPH journals, second CPU, 2 x 10GbE NICs
- Tape: 2 x 10k slot SL8500, 80+ drives
  - Migration from T10KA/B to T10KD tapes completed
    - Dedicated migration system
      - 4 stagers, 6 T10KB drives, 2 or 3 T10KD drives
    - Averaged 50 tapes copied per day
    - 3000 T10KA (1.2PB data) -> 160 T10KD
    - 3950 T10KB (CMS, 3.6PB data) -> 550 T10KD (3 months, no data loss)

CMS Migration



Tape Store

T1 Castor	files= 11880164	data= 11248 TB	Tapes (C/D)= 1932
Facilities	files= 1315890	data= 8463 TB	Tapes (C/D)= 1378
ADS	files= 300818	data= 2426 TB	Tapes(B)= 2126
DMF	files= 175353757	data= 1572 TB	Tapes (A/B)= 1955



Software and Systems I

- Batch: HTCondor
  - Very flexible, stable
- Storage: Castor
  - Updating next week to v2.1.14.-15
  - v2.1.15 pending
- Storage: CEPH
  - Evaluations continue for use of CEPH as replacement for Castor disk-only service
    - See talk by Alastair Dewhurst
- Databases:
  - Oracle On RHEL5, 11.2.0.3/4
  - Database replication for ATLAS 3D now using Oracle Goldengate
  - Refresh of hosts and storage in planning
    - More IOPs, faster SAN, more TB, distribution of data
  - MySQL, Postgres

Software and Systems II

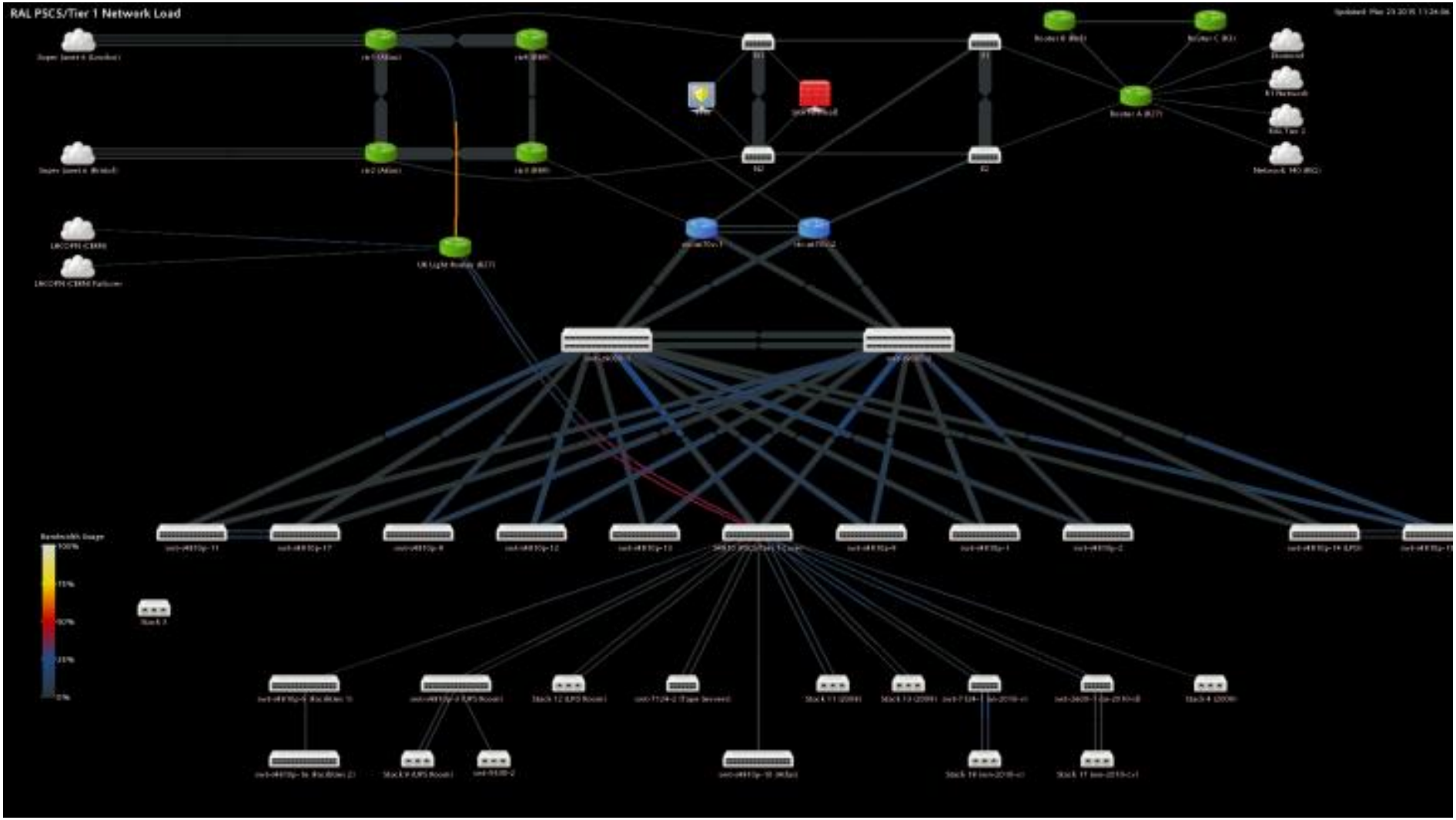
- ~120 production VMs running Grid production services
  - HyperV hypervisors, local storage
  - Shared storage cluster work ongoing
- Provisioning:
  - Quattor
    - Aquilon approved for production use
  - Talk by James Adams
- Logging:
  - Moving to Elastic Search infrastructure
- Monitoring:
  - Ganglia, Nagios, Cacti, Observium, ….
- Cloud:
  - 28 system cloud resource available for department testing
    - No production services allowed!
  - Talk by George Ryall

Networking

- Tier1 LAN
  - Mesh network transfer completed
  - Problem with X670v routers
    - Primary stalls routing when acting as master in master/slave pair
    - No failover – no failure of management layer interconnect
    - Extreme on the case
  - Router issue delaying progress with developments:
    - Phase 2: 40Gb/s redundant link T1 to RAL Site
    - Phase 3: move the firewall bypass and OPN links to x670v routers
      - Will provide 40Gb/s pipe to border
  - Small IPv6 network

- RAL LAN
  - Additional 40GbE capacity for core switching

- Site WAN
  - No changes


RAL PSCS/Tier 1 Network Load

JASMIN

- JASMIN = Joint Analysis System Meeting Infrastructure Needs
  - A “Super Data Cluster” not a “Super Computer”
  - Data movement and analysis.
  - Funded by NERC for all of NERC sciences.
  - Hosted at STFC RAL by the Research Infrastructure group of the SCD
- Satellite data, weather data, climate simulations from big HPC systems (“Archer”, MetOffce “Monsoon”, DKRZ.)
  - JASMIN Holds >60% of Data used by the latest IPCC report on Climate change
  - Largest data set 600TB

JASMIN Tech

- 16 PB useable (20PB raw)
  - ~ 3,200,000 DVD's = ~ 6km high tower of DVDs or > 36,000 years of MP3
  - Two largest Panasas 'realms' in the world (109 and 125 shelves).
  - Largest single site Panasas customer in the world (251 shelves)
  - 900TB useable (1.44PB raw) NetApp iSCSI/NFS for virtualisation + Dell Equallogic PS6210XS for high IOPS low latency iSCSI
- 4,000 CPU cores split dynamically between batch cluster and cloud/virtualisation (VMware vCloud Director and vCenter/vSphere)

JASMIN Net

- >3 Tb/s bandwidth (~3500 DVD's per minute)
- "hyper" converged network infrastructure
  - 10GbE + MPI low latency (~10uS) + iSCSI over same network fabric
    - No separate SAN or Infiniband
- Finalist for BCS UK industry Awards "Big Data Project of the Year" 2012 and 2014
- Managed with 2FTE, recruiting for a 3rd team member

Other stuff

- Electrical ‘shutdown’ for circuit testing
  - Completed in January
    - Phased circuit testing
    - Tier1 continued to operate with minimal loss of batch capacity, no loss of storage access
    - No issues major issues
      - a few wrongly identified circuits
      - one or two overload trips in batch racks
- Disposals
  - 2006/7/8 and older hardware being scrapped
    - Information security requirement to dispose of data securely
    - Disks ‘blanked’, SNs recorded, scrapped
    - Failed drives sent for secure disposal
- New telephone handsets being tested – VoiP
  - Call clarity said to be ‘stunning’

Questions?

- Used tapes, anyone?