CCLRC
Rutherford Appleton Laboratory

Preparations for the LCG Tier1 Centre at RAL

LCG Workshop @ CERN

23/24 March 2004

CCLRC
Rutherford Appleton Laboratory

Outline

- Overview of RAL
- Hardware
- Infrastructure
- Staff
- LCG

Martin Bly
RAL Tier1/A Centre

CCLRC
Rutherford Appleton Laboratory

RAL Tier1/A

- Multi-experiment site used by LHC experements, BaBar collaboration, Dzero, H1, Minos, Zeus, QCD, theory users
- Committed to provide resources for BaBar (disk, CPU) – the`A' in `Tier1/A'.
- CPU/disk/tape resources grown over time for some years – heterogeneous cluster
- Key - Flexibility

Martin Bly
RAL Tier1/A Centre

Systems

- Main farm has most of CPU hardware
  - Job submission: OpenPBS
  - Scheduling: MAUI
- Disk storage available to batch workers via NFS
- AFS on all systems
- Installation and Update: Kickstart, YUM
- Monitoring: SURE, Gangila, Nagios, Yumit

Present Hardware - CPUs

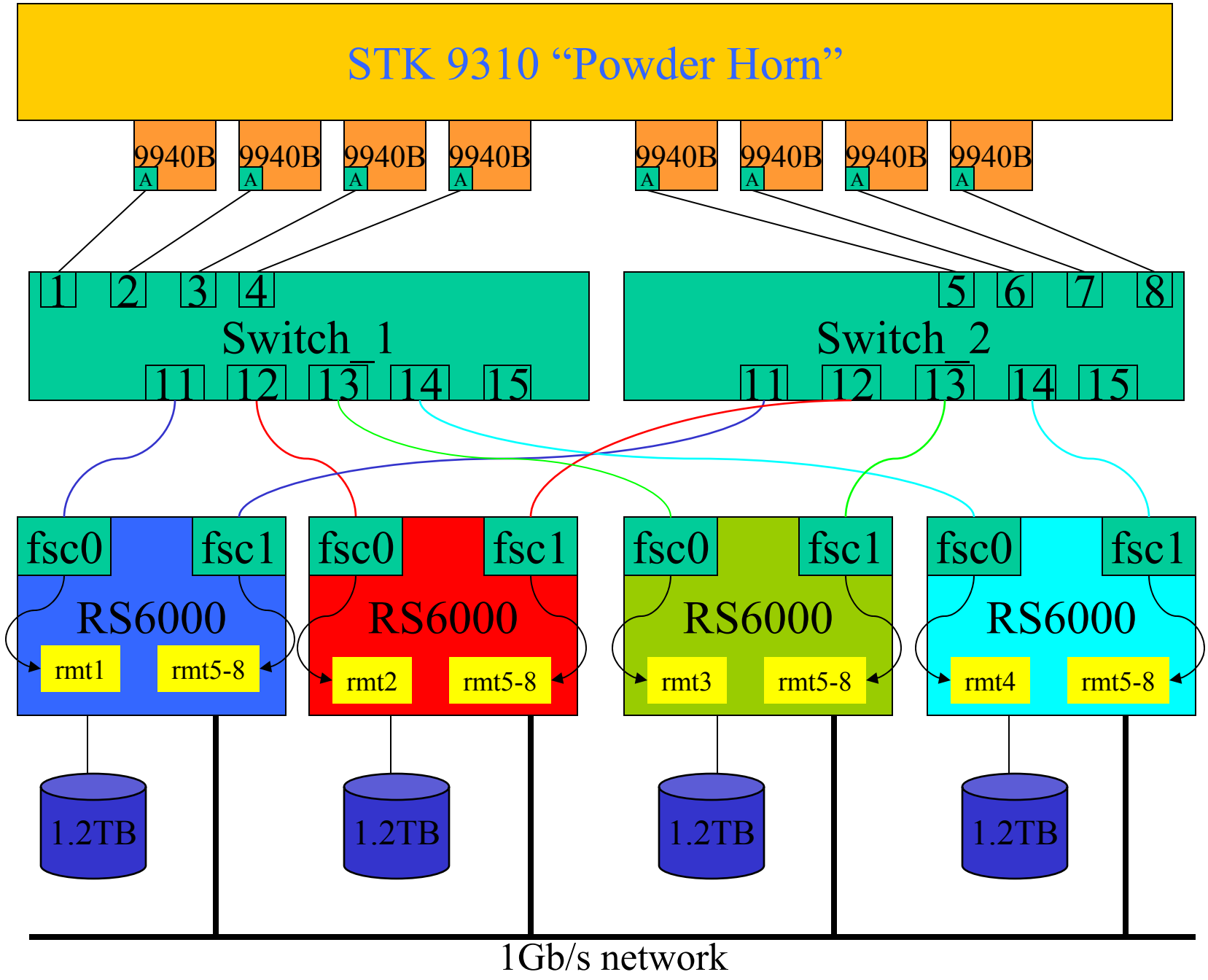
- Linux
  - 80 dual CPU 2.66GHz/533Mhz P4 HT Xeon rack mounted systems, 2GB memory, 30GB disk, 2x1Gb/s o/b NIC
  - 156 dual CPU 1.4GHz/133MHz P3 rack mounted systems, 1GB memory, 40GB disk, 2x100Gb/s o/b NIC
  - 47 dual CPU 1.0GHz/133MHz P3 tower systems, 1GB memory, 40GB disk, 100Gb/s o/b NIC
  - 40 dual CPU 600MHz/100MHz P3 tower systems, 512MB memory, 12GB disk, 100Gb/s NIC
  - 29 dual CPU 450MHz/100MHz P2 tower systems, 256MB memory, 12GB disk, 100Mb/s NIC – to be retired soon
  - Various tower and rack boxes for testbed systems
- Solaris (Babar - being phased out)
  - E4500 - 6 CPUs, 4GB memory
  - E3500 - 4CPUs, 1GB memory
  - 4 x E420 (4CPUs each, 1GB memory)

Present Hardware - Disk

- 11 Linux rack mount servers providing ~40TB IDE disk
  - 11 dual 2.4GHz P4 HT Xeon servers with PCIx (1GB RAM), each with:
  - 2 Infortrend IFT-6300 arrays, each with:
  - 12 Maxtor 200GB Diamondmax Plus 9 drives per array, most configured as 11+1 spare in RAID 5 => ~2TB/array.
- 26 Linux rack mount servers providing ~44TB IDE disk
  - 26 dual 1.266GHz P3 servers (1GB RAM), each with:
  - 2 Accusys arrays, each with:
  - 12 Maxtor 80GB drives -1.7TB disk per server.
- 3 Linux tower servers providing ~4.8TB IDE disk
  - 3 Athlon MP 2000+ single processor tower servers, each with:
  - 1 x 3ware 7500-8 with 8 Maxtor DiamondMax Plus 9 as RAID5
- 2 Linux servers providing 300Gb SCSI RAID 5 (to be deployed).
- Solaris server with 4.5TB
- 3 x Ultra10 Solaris servers (being phased out)
- AFS Cell – 1.3TB, AIX +Transarc – migrate to Linux + OpenAFS server during 2004.

Present Hardware - Tape

- Storagetek 9310 Powderhorn tape silo
  - 6000 slots
  - 8 x STK 9940B drives (200GB 30MB/sec)
  - Potential 1.2PB of which 550TB allocated to HEP
- 4 IBM 610+ servers with two FC connections and 1Gb/s networking on PCI-X
  - 9940 drives FC connected via 2 switches for redundancy
  - SCSI raid 5 disk with hot spare for 1.2Tbytes cache space


STK 9310 “Powder Horn”
9940B
A
9940B
A
9940B
A
9940B
A
9940B
A
9940B
A
9940B
A
9940B
A
1
2
3
4
Switch_1
11
12
13
14
15
5
6
7
8
Switch_2
11
12
13
14
15
fsc0
fsc1
RS6000
rmt1
rmt5-8
fsc0
fsc1
RS6000
rmt2
rmt5-8
fsc0
fsc1
RS6000
rmt3
rmt5-8
fsc0
fsc1
RS6000
rmt4
rmt5-8
1.2TB
1.2TB
1.2TB
1.2TB
1Gb/s network

2004 Hardware - May

- 200+ dual CPU rack mount systems:
  - AMD Opteron or Intel P4 HT Xeon
  - 2 or 4GB memory
  - large local disk (120GB)
- 120+Tb configured RAID5 disk in Infortrend IDE/SCSI arrays with dual CPU servers to match
  - SATA disks, SCSI interface
- 200Gb tapes as required
- `Same again' later in 2004.

2004 Hardware - later

- Finance recently approved for next three financial years – '04/'05/'06.
- Hardware spend for later in 2004 expected to be of same order +- as current round (~700GBP).
- Split depends on cpu/disk/tape requirements of experiments.
  - Priorities decided by GridPP experiment board.
- PPARC – long term commitment to Tier1/A centre at RAL for lifetime of LHC but no financial figures beyond next three years. On the financial roadmap of the future.

Purchasing / QA

- Strategy:
  - To specify what is required in some detail while leaving the vendors room to move
  - Use GCAT if possible to reduce the field to specialist outfits
  - Evaluate the offerings against detailed specs
  - Determine !/$ (bang/buck)
  - Pick the tender that provides the best overall offering – which is not necessarily the cheapest!

CCLRC Rutherford Appleton Laboratory

Infrastructure - Network

- Now:
  - CPUs on 100Mb/s links to 24 or 32 port switches with 1Gb/s uplinks to Tier1/A backbone switches.
  - Disk servers on 1Gb/s links to backbone
  - 1Gb/s link to site backbone (1Gb/s)
  - 2*1Gb/s link to the SuperJanet (SJ4) network via the Thames Valley Network (TVN) which itself operates at 2.5Gb/s
  - RAL hardware firewall (NetScreen 5400) capable of 4Gb/s bi-directionally to/from LAN/WAN
  - RAL is on SuperJanet development network
    - Separate 2.5Gb/s connection
    - MB-NG project

Infrastucture – Network (2)

- Future:
  - Local LAN can go to 10Gb/s as required
  - Partner institute in UKLight facility:
    - UKLight will connect UK research network to USA (StarLight) and Netherlands (NetherLight) via high capacity optical links
    - Onward connectivity to CERN
    - Multi-channel Gb/s controlled as if end-to-end path is a single switched lightpath or wavelength.
    - Access to UKLight as extended national development network alongside SJ4 service network
    - Integration of UKLight with LHC requirements

Infrastructure - Environment

- RAL facilities used to host several mainframes and there are 5 large machine rooms. Share with HPC factilities and RAL computing infrastructure.
- 450kW cooling plant based on two compressor/heat-exchanger units with third unit on standby.
- 600Amps in the 415V/3 phase supply is unused.

Infrastructure – Environment 2

- Don't expect major capacity problems for new equipment for Tier1/A in May (200-250 cpu units and 120+Tb disk).
- Some work may be needed for cooling and power distribution though no major problems expected for December 2004 disk/cpu.
- Tier1/A should achieve its steady state capacity of around 1000 CPUs purchased over 4 years by Dec 2005, as older equipment phased out.
- Ongoing survey of machine room environment. Better assessment of growth limits later this year.

Staff

- Now: 12.1 staff funded as part of GridPP project.
- September: increase to 16.5 (GridPP2)
- Additional effort will be mostly for staff to support the experiments and GRID deployment.

LCG activities

- LCG-0 March 2003
  - RB, SE, CE, 5xWN, BDII, MDS
- LCG-1 July/Aug 2003
  - RB, SE, CE, 5xWN, BDII,MDS
    - West Region MDS
    - Many problems with RB
    - Decommissioned Feb 2004 though MDS and BDII still running
- LCG-2 Jan/Feb 2004
  - RB, SE, CE, 73xWN, BDII, MDS, IC, MON
    - West Region MDS
    - Much more stable!
    - Real work, mostly from ALICE

Martin Bly
RAL Tier1/A Centre