RAL Site Report

John Gordon
HEPiX/HEPNT
Catania
17th April 2002

Outline

- Recent hardware changes
  - CPU
  - Disk
  - Tape
  - Network & Wireless
  - Videoconferencing
- Recent service changes
  - TierA Centre for BaBar
  - EDG Testbed1
- Issues

Recent hardware changes

- CPU
- Disk
- Tape
- Network

CPU

- Recently started buying 1u racked dual cpu boxes
  - 14 dual 1GHz for EDG Testbed late 2001
  - 156 dual 1.4GHz PIII in March 2002
- Large increase on existing 250 cpus
  - Speeds from 450MHz to 1GHz

CPU

- Recent purchase from Compusys
- Result of competitive EU Tender
- Chose PIII when we defined the spec last October
  - Worried about P4 compiler support and performance
- Kickstarted into existing infrastructure so running within days

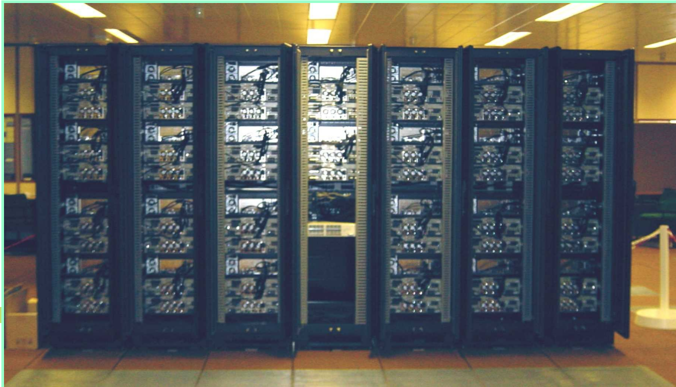
COMPUSYS
COMPUSYS
COMPUSYS
COMPUSYS

CPU

- 312 1.4GHz Pentium III Tualatin cpus
- 1GB of memory/dual
- Internal 40GB Maxtor Viper disk
- Tyan S2518 Serverworks LE Motherboard
- 100Mbit Ethernet NIC
  - Connected to switch with multiple Gb uplinks

Recent Tender for Disk

- Delivered March 2002 so no running experience yet
- Compusys the supplier
- Chose SCSI/IDE RAID solution
  - IDE disk in RAID controller
  - SCSI connection to Host
  - No modification to host operating system
  - Quicker to replace host, reconfigure, add extra arrays
- 26 rack-mounted Linux servers with 52 RAID Arrays
  - = ~ 50,000GB raw, 45TB RAID5



APC
APC

RAID Arrays & Servers

RAID

RAID Controller	Zero-D X-3I-R
Drives Per Controller	12
Drive Size/Manufacturer	80GB Maxtor Viper
Speed	7200rpm

Servers

Processors	2x1.266GHz
Motherboard	Tyan S2518 Serverworks LE
Memory	1GB ECC PC133 SDRAM
NIC	Intel pro/1000T

Performance

- Benchmarked several of the offered solutions (IOZONE shown)
  - Can’t disclose benchmark results of other suppliers
- Zero-D RAID controller showed the best performance,
  - particularly in sequential read where we have the most pressing requirement.
- Infotrends 6300 also offers good performance.
- The remaining systems offer only tolerable or poor performance
- Benchmarking was most illuminating regarding the suppliers knowledge of the equipment, technical expertise, ability to cope under pressure and ability to provide support on their product.
  - This information was also fed into the tender evaluation.

Benchmark Results

Single Array

SINGLE THREAD SEQUENTIAL READ (KB/s) at Varying Record Size		Read
1K	50867	61481
8K	63324	59265
16K	63272	60965
32K	62788	61568

THROUGHPUT TEST (aggregate KB/S) at Fixed Record Size (32K)		Read
1 Thread	62788	61568
2 Thread	58583	40974
4 Thread	55986	43347
8 Thread	53729	43735
16 Thread	51800	37641
32 Thread	49310	30953

Stride Throughput Test 32K Read 15 Skip		Read
1 Thread		6187
2 Threads		6701
4 Threads		10706
8 Threads		15012
16 Threads		18033
32 Threads		17799
Random IO	29501	5926

RAID0 across 2 RAID5 Arrays

Single Thread

Record Size	Write	Read
1K	149609	122886
8K	156983	124887
16K	156030	129952
32K	151636	129589

Throughput test (aggregate KB/s) at Fixed Record Size (32K)

Threads	Write	Read
1	151636	129589
2	127627	107764
4	120266	118642
8	115657	108699
16	109554	98601
32	94960	93125

Stride Test

CMS Data read under Objectivity mimic 1000MB file

32K read. 15 Record Skip

Threads	Read
1	6944
2	7415
4	12913
8	19423
16	26528
32	31189

1 Stride Reader (30 Record Skip) 6958

Random I/O test 1000MB file 32K read

	Write	Read
1 Random I/O	86767	5690

IDE RAID

- dual 1GHz PIII system
- two 3ware Escalade 7810 controllers (64bit/66MHz PCI)
- 8 x 100GB maxtor disks per controller 7 per RAID 5 set 1 hot spare giving 1.1TB usable filespace in total
- Bonnie performance (kernel 2.4.17) MB/s

RAID level	Block write	Block Read
0	91	82
5	11	95

- Good RAID5 read performance but not good write performance

Disk Services

- Still need to investigate how best to run new disk servers
- RAID0 vs RAID5 vs RAIDn
- Filesystems?
- Will probably try different setups for data files/databases/scratch

Tape

- Single STK Powderhorn Robot
- 5632 slots – not full
- 5 IBM 3590 drives (10GB)
- 5 STK 9940 drives (60GB)
- Currently: 3160x3590 + 864x9940 = 81TB
- If full of 9940s = 330TB

Network

- Nortel Gbit infrastructure for whole site
- 3xSummit 7i for HEP services
- WAN 622Mb to local WAN, 2.5GB UK backbone
- 2.5GB to RAL summer 2002 when backbone goes 10Gb
- 100s of cpus access 10s of disk servers (Gb) needs:-
  - Very big switch as interconnect – OR
  - Clustering of disk servers and cpu clusters

Wireless Networking

- Installing wireless access ports in conference rooms
- DHCP gives IP numbers on separate class C ‘visitors network’
  - Outside firewall
- Staff and other approved people can use PPTP to enter RAL site from wireless

Video Conferencing

- Conference Room ISDN
- Personal VRVS
  - UK reflector at RAL
  - Ad-hoc use by bigger meeting
- H.323
  - Regular use in UK HEP.
  - Most groups have Zydacron room-based systems or Polycom ViaVideo personal systems
  - No production MCUs yet, using several test services

Other Services

I have concentrated on changes, we still continue to run

- 2 Compaq AlphaSC systems with Quadrics switches
- AMD-based Beowulf clusters
- Sun cpu and disk for BaBar
- IBM batch farm and fibre-based disk server for CDF

Recent service changes

- TierA Centre for BaBar
- EDG Testbed1
- More about this on Friday

Issues

- Network layout cpu vs disk
  - Big switch with high-end backplane vs localised use
- Disk Reliability
  - Many problems with IBM 75GB IDE
  - Got a batch of 60 replaced by supplier
- AFS
  - How long can we get away with AFS?
  - What platform should we use for OpenAFS?