HEPiX Report
Computing Seminar / After-C5

April 21st, 2006

Luca Canali, CERN

Outline

- Report from HEPiX Spring 2006 in Rome
- Topics covered
  - Databases
  - Network
  - Authentication
  - Storage

Database Setup and HW

- Reports from CERN, RAL and CNAF
  - Homogeneous deployments
- Oracle 10g RAC on Linux for DB services (Luca's talk)
  - Oracle 10g R2 (latest version)
  - Linux, Red Hat Enterprise (3 and 4)
  - Clusters of low-cost HW - typically dual CPU (Xeon) and 4GB
  - Storage built using low-cost storage arrays
    - Leverage Oracle ASM (volume manager/custom filesystem)
    - Large capacity and scalable performance
  - Backup: RMAN, Monitoring: OEM

Database Services

- Replication and 3D
  - Oracle Streams successfully tested between T0 and T1s
  - Frontier, complementary approach
  - Presenting T1 sites deploying in Q2 (service from Q3)
  - see also Dirk's talk
- CASTOR2
- DB services for LHC experiments and LCG
- Note: Oracle licensing for T1s was discussed

Authentication

- Reports from DESY, BNL, RAL and CERN
- Very important topic for grid deployment, but still not settled
  - Issues on the interoperability Unix/Windows
  - SSO, eases admin and users operations
    - But requires strong protocols
    - One Time Passwords useful but complex and not yet standard
  - Different solutions coexist, notably Kerberos and PKI
- Synchronization between PKI (Grid Certificates) and Kerberos
  - See Alberto’s talk

Network

- Reports from GARR and INFN
- High speed T0 – T1 connectivity
  - Global connectivity at 10Gbps already available
  - For example geant2, but other and similar networks exist
  - IPV6 already supported, can be good for grid (routing performance is increased)
- Measurements shown for 10Gbps transfer rate between CERN and CNAF
  - Some tuning results were presented

Storage

- Storage day
  - Tape and disk hardware, storage interconnects and protocols
  - File systems (local and distributed)
  - Storage Models (disk-to-tape migration)
  - The High Energy Data Pump - State of the Art in Hardware and Software
  - Backup technology

TAPE and DISK HW 1/2

- Reports from FNAL, DESY, CASPUR, CERN
- TAPE:
  - Pros: (a) very large capacity and (b) retention time
  - Cons: (c) specialized operation (hidden costs and unknown future), (d) peculiar performance characteristics
- DISK:
  - 'Storage in a box' solutions have low cost/TB
    - Good performance and reliability. Recent improvements also with RAID 6
  - E4 'fat' disk server (RAID 6, 14 TB raw, 2 CPUs, ~250 MBPS streaming)
  - Coming: newer generation with RAID 6 -> expected 700 MBPS

TAPE and DISK HW 2/2

- Many choices for disk based solutions (still open question)
  - Disk type: Fiber channel, SAS, SATA
  - Interconnect: Gb Eth, FC, 10Gbps Eth, IB
- Outlook and issues for disks:
  - HD technology: no significant improvements expected within 2-4y
    - Performance (throughput, IOPS and latency)
    - Cost/Capacity also flattening out
  - SAS will come (high end), SATA proven reliable (low end), FC no change
  - Error rate for RAID 5 not acceptable, RAID 6 is a solution
  - Interconnect is becoming a bottleneck: may need 10Gbps or IB
  - Object-Based Storage Devices: new but yet unproven architecture for storage scalability

Filesystems

- A few filesystems for Linux suitable for production
  - Ext3, XFS, JFS, ReiserFS.
  - XFS best choice for large files and streaming IO.
  - However Ext3 better choice when dealing with red hat.
  - Filesystems are quickly evolving
- WAN filesystems, Gpfs: 2Gbps throughput measured from CERN to INFN Bologna.
  - Multiple streams (40) used for performance
  - Note: GPFS requires trusted hosts, not a global solution
  - AFS on WAN: low performance

Disk-to-tape migration

- Disk Pool Management Systems either integrated with (Tape-) Mass Storage Systems or providing Interfaces to support a Storage Hierarchy.
- Three main players:
  - CASTOR2
    - DB-centric, feature rich, scalable and performing: 240 TB, 30 tape Drives,120 clients->2.2 GB incoming for 2days (see also Sebastien's talk).
  - dCache
  - HPSS
- Tape Storage Backend for Disk Pool Managers & Stagers
  - TSM

The High Energy Data Pump

- Software and HW infrastructure for high bandwidth data transfer between CERN and T1 is in place
  - Aggregated throughput figures close to 2GB/s
  - Stability and performance need further improvements
- HW:
  - Linux and storage in a box + external storage
  - Gbps Ethernet (some sites use bonding)
- Software:
  - Transport: http+mod_gridsite vs. – GridFTPv2. Similar performances but http allows encryption and solves FW issues
  - Service: FTS vs. RFT. Both solutionsnot fully mature, however FTS is more robust.

Backup Technology

- Survey of backup operations (David's talk):
  - TSM, Legato and home made solutions (mainly)
  - Daily incremental backup (average 500 GB/day, CERN 1TB)
  - Total size of retained backups: from 100 to 400 TB
  - TSM -> AIX (Linux planned), Legato -> Solaris
- Trend: backup volume increases, largely due to DBs growth
- Plan: backup with TSM over SAN (FC)
- AMANDA: low cost solution, good for small sites
  - Currently used at TRIUMF

Conclusions

- Topics covered
  - Databases
  - Network
  - Authentication
  - Storage
- Personal comments on general trends:
  - Almost all groups are pushing the limits of their technology area to get ready for LHC startup, typically deploying the latest software version or latest HW model.
  - Most architectural areas have already chosen the 'best solution', while other areas show directly competing software solutions. Any bets?