HEPiX 2007 Storage Day

HEPiX FSWG Report

- most sites have responded to query
- HPSS – CASTOR – dCache - Lustre
- CEA, LLNL, CERN, FZK: 1+ PB disk
- considered for further investigations:
  - HPSS, CASTOR, dCache
  - Lustre, AFS, NFS
- Lustre has 5.5PB, but not deployed!

Object based storage

- AFS-OSD (RZG) T10 OK, works
  - alternative to MR-AFS
- Lustre (CEA) scales well, security?
  - easy config, troubleshooting is problematic
- Panasas (BNL) will be phased out
  - expensive, diminished support quality, crashes while Nessus scans, upgrade troubles
- Panasas (JLAB): will be replaced (ZFS)
- Panasas (DESY): mostly OK, high cost

Cluster filesystems

- GPFS (NERSC): manageability issues
  - metadata on FC, data on SATA
  - fault resilient fabric is needed
  - stale NFS handles, OOM, corruptions
- GPFS (CNAF): works fine
  - 110 TB testbed, repackaged to deploy
  - CASTOR reads 1.3G/s, writes 0.85G/s
  - GPFS reads 1.5G/s, writes 1.2 G/s
  - will migrate all NFS to GPFS

GPFS/HPSS integration

- long development project @NERSC
- DMAPI needs to be extended
- namespace consistency problems
- release planned 2007Q3
- backup functionality planned 2007Q4
- plans to extend HPSS storage to 60TB

ZFS vs XFS

- Thumpers are very popular
- ZFS has checksumming...
- XFS is on par with ZFS for writes
- XFS wins in reads
- ZFS wins in metadata operations
- ZFS is OK w/dCache, not OK w/Lustre
- Thumpers have a SpoF, the CPU/memory controller

Thumpers everywhere...

- IN2P3: ~800 TB
- DESY: ~170 TB
- ZFS is still very young
- manageable filesystems are 1-2 TB
- used as “foundation filesystem” for HSM disk layer
- ...but requires Solaris 10

DPM

- SRM v2.2 basic tests OK
- xrootd plugin prototype being tested
- CASTOR-DPM common RFIO: 2007Q4
- Perl API
- GSI/Kerberos 5, POSIX ACLs
- VOMS integration
- 90 production sites, 100 Vos
- plans: SRMv3, 64-bit, encryption, accounting/quotas, LEMON integration

dCache

- SRM v2.2 almost OK
- interface to several HSM systems
- fully automated code-to-product chain
- “dCache in 10 minutes” - easy config
- future plans include NFS v4.1 support
  - access to namespace
- Java5 support in 1.8.0

SRM

- permanent files
- permanent space
- space reservation
- permission functions
- directory management
- data transfer
- file access protocol negotiation
- relative paths

Silent Corruptions

- talk relatively well received
- sites have similar issues
- fsprobe tool was requested by many
  - move from swrep to SLC4
- good news: WD firmware upgrade campaign seems to have reduced the number of Type III corruptions observed

HEPiX 2007 Benchmark Day

SPECint2000 vs SPECint2006

- Si2K: difficult to find for new CPUs
- Si2006: difficult to find for older CPUs
- AMD/Intel rate higher in Si2006
- 64-bit clean is usually better perf.
- multiple cores scale horizontally
- fluctuations: Si2K is not a good indicator for HEP code performance (based on ROOT/SUSY/CMS_sw)

SPECint2000 hurdles

- latest published numbers are
  - done with modern gcc
  - only one benchmark per CPU
- gap increasing (up to 10%)
  - SPECint2000 numbers
  - our own measured numbers
- run our own, share results
- require vendors to run our code

Quad-core Systems

- clock race has ended
- core race has started
- soon insuff. HW to “feed” all cores
- good scaling compared to dual-core
- simple threading is not believed to be enough, parallel paradigm required (MPI, OpenMP)