

CHEP 2010

Storage

summary

CHEP 2010

There was no “storage” track.

- “Software Engineering, Data Stores, and Databases” did not have much storage in it.
- Much more was in “Distributed Processing and Analysis” and “Grid and Cloud Middleware.”
- Some in “Computing Fabrics and Networking Technologies.”

Site-to-site

“SRM is the big success story of grid middleware’ – discuss.”

“FTS may have to be replaced—any-to-any traffic breaks it, and star channels are hard to manage.” Want to move toward zero configuration, with capacity a function of endpoints, not channels.

Some plugging of xrootd. Not universally liked.

Data management

Predictive data placement was pretty much a flop.

- Example: ATLAS users had said they would want heavy access to the RAW data, but in fact did not use it much.

Local access

A lot of “buzz” around NFSv4.1.

- Very favorable test vs dcap.
 - Moved 40% as many bytes, at 2x the BW, for a given amount of physics work.
- Looks good against xrootd.

Client will be in RHEL 6.1, and available in a few months in SLF5 as a tested but less supported option.

Local access

WebDAV for access (Gerard, PIC)

- Very firewall-friendly.
- dCache WebDAV must go through door, not be redirected to pool.
- ROOT client has trouble accessing a dCache WebDAV by http:

Disk storage

SSD: Price is still prohibitive for general use, and the write performance is not compelling.

At “File Stager” task, efficiencies

- SATA RAID0 > SATA RAID1 > Intel SSD > 1 HDD > cheap Kingston SSD
- And RAID0 cost \approx 1/3 Intel SSD, per percentage point of efficiency.

Disk storage

DESY-Z

- 5 Lustres, 300TB; 2 dcache, 970TB
- 55TB AFS, 9TB Panassas

Direct-attached storage lower cost,
≈ performance, simpler than SAN or NAS.

dCache vs. Lustre

Lustre: full POSIX, top single-client performer; no migration/rebalancing

dCache: HSM, many protocols; cannot modify parts of files

DESY-Z Finds dCache and Lustre very alike, with ~15 minor, 1-2 major incidents/PB/y.

Tape

CERN:

- Plenty of floor space.
- Low (40MB/s) per-drive performance.
3 tape marks/file and peak of file size histogram is 10^6 – 10^8 bytes!
 - “No” to a new tape format.

Removed tape-read privilege from most users.

Tape

PIC

Moving to 4 drives/mover with 10Gb network because of power & heat.

Setting discipline (xfer/client limit) to 1 raised average tape performance from 60-70MB/s to 100-110MB/s.

Having dCache delay writes when files are too few, too small, too new.

Tape

INFN T-1 at CNAF:

GEMSS \approx GPFS + TSM + StoRM + GridFTP

- Had to extend StoRM to include more of SRM functions, for tape.
- Had to do work to integrate GPFS+TSM
- LHCb first, other expt's one by one.

Miscellany

LHCb:

“Insane” environment variables such as \$PATH and \$LD_LIBRARY_PATH were causing ~ 40,000 stat() and access() syscalls on job startup.

Created “RFS” with FUSE which caches negative & positive answers, caches opened files in local RAMDISK.

Other tidbits

Roger Jones (Lancaster), on offline processing & monitoring: “We shouldn’t be punishing sites for things that aren’t their fault.”

David South (DESY): At end of experiment there is the best data and the worst funding.

Wojciech Lapka: WLCG-wide messaging middleware. Favors ActiveMQ.