

PSI Site Report

(Spring HEPIX 2008)



Heiner Billich

Paul Scherrer Institut, Switzerland

PSI Site Report - Topics

- LCG Tier-3 Cluster for CMS
- Collaboration Application – TWiki, Jira and iDok
- HPC – Move to Blades
- HPC – GPU Computing
- Scientific Linux support at PSI
- Puppet for automated system configuration
- GPFS at SLS (synchrotron) beamlines
- Backup: Introduce disk cache and LTO4

LCG Tier-3 Cluster for CMS

- PSI is building a LCG Tier-3 for the CMS groups at ETHZ, PSI and University of Zurich
- Storage will be managed by dCache
- Coupled to CMS Grid Storage Infrastructure (PhEDEx)
- Currently no intention to couple job queues to Grid
- System size

Year	Compute cores	Storage / TB
2008	64	75
2009	184	250

- Storage using Sun's ZFS file system with RAID/Z (HW:X4500)
- Will use Intel based quad-core processors

Collaboration Applications – TWiki, Jira and iDok

building blocks for applications to support collaboration, projects and documentation at PSI.

- mix and match as you need
- TWiki – collaboration, documentation, mostly unstructured
- Jira (www.atlassian.com) - bug/issue/problem-tracker, workflow.
- iDok – document management system, structured, hierarchical.
see next slide
- our contribution: Customize, integrate, interfaces and iDok

Collaboration Applications – iDok

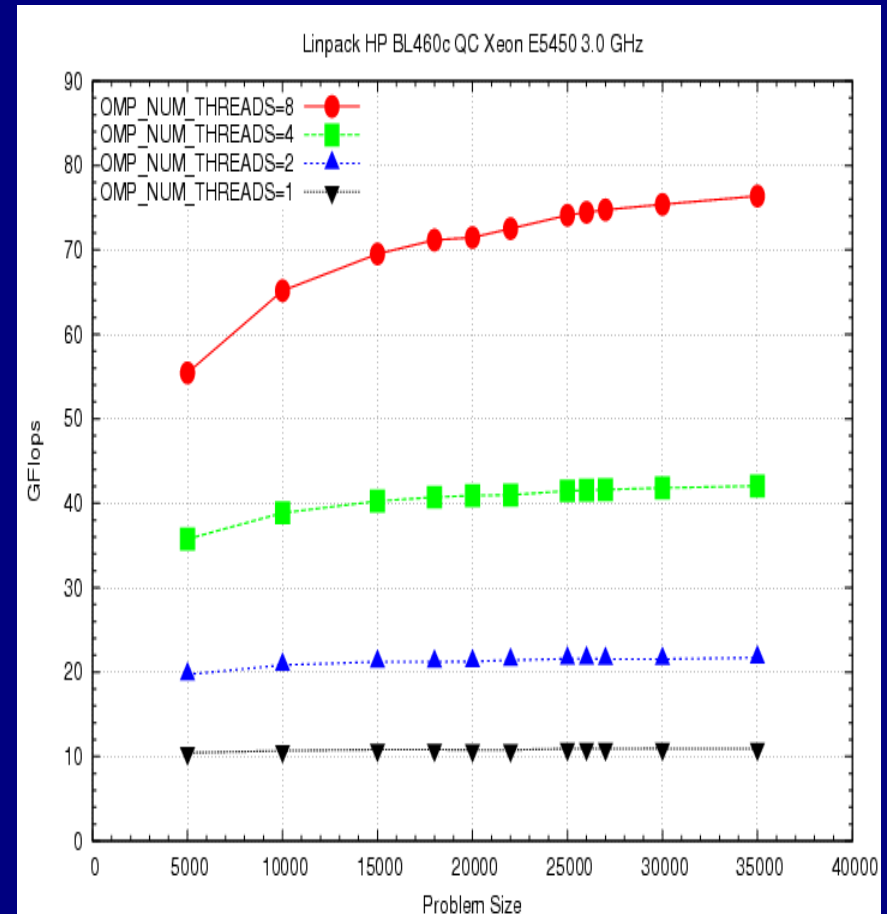
An open source document management system based on subversion

- storage and retrieval of arbitrary documents.
- Client runs on linux and windows.
- secure access through established protocols like HTTP(S), WEBDAV(S) and CORBA
- fast searching of documents by full text content and by user-defined or intrinsic metadata.
- subversion + apache + lucene + java

 www.idok.ch

HPC – Move to Blades

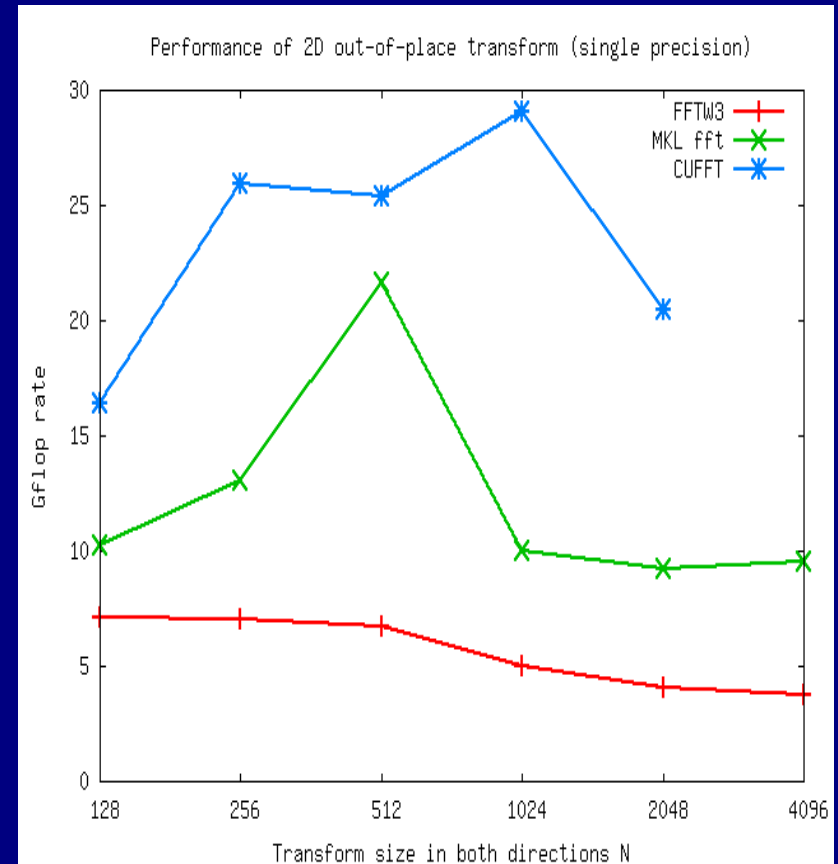
- We moved from 1U rack servers to blades (HP c7000/BL460c)
- From dual-core Opteron to dual- and quad-core Xeon
- admins like blades ...
- ... as long as they have enough „similar“ machines
- quad-core Xeons scale well
- HP BL460c dual-socket quad-core Xeon E5450 3.0 GHz, 16 GB RAM Linpack: 10GFlops/core
- now adding InfiniBand for storage and MPI – *comments welcome*



HPC – GPU Computing

Parallel Tomographic Reconstruction using Graphics Hardware

- Goal: Speed up the reconstruction algorithms to match the speed of the data acquisition process
- Implement numerical kernels on the GPU using the NVIDIA CUDA framework
- Performance results for 2D FFT (Intel 2.83 GHz Quad CPU vs NVIDIA 8800 GTS 512 GPU):
- work in progress



Scientific Linux, SL5, Puppet

■ Scientific Linux Support at PSI

- SL5 32/64bit supported since 2007-09
- SL4 end of support planned for 2009-09
- SL3 end of support planned for 2008-08

■ Move from cfengine to puppet

- both used for automated system configuration
- puppet as good or bad as cfengine
- puppet: ongoing development, open community, responsive mailing list

GPFS at SLS (Synchrotron) Beamlines

■ GPFS in production at six beamlines

(GPFS: General Parallel File System from IBM)

- Today we handle up to ~1TB data per day and beamline: measure, process, export to removable media.
- We observe up to 300-400MB/s sustained aggregate throughput per beamline in production.
- 120TB file system space in total at six „big“ beamlines. (all SATA disks, 4+2 Raid6)
- GPFS works and performs as expected, good support via mailing lists and web forum. Feels very mature and solid.

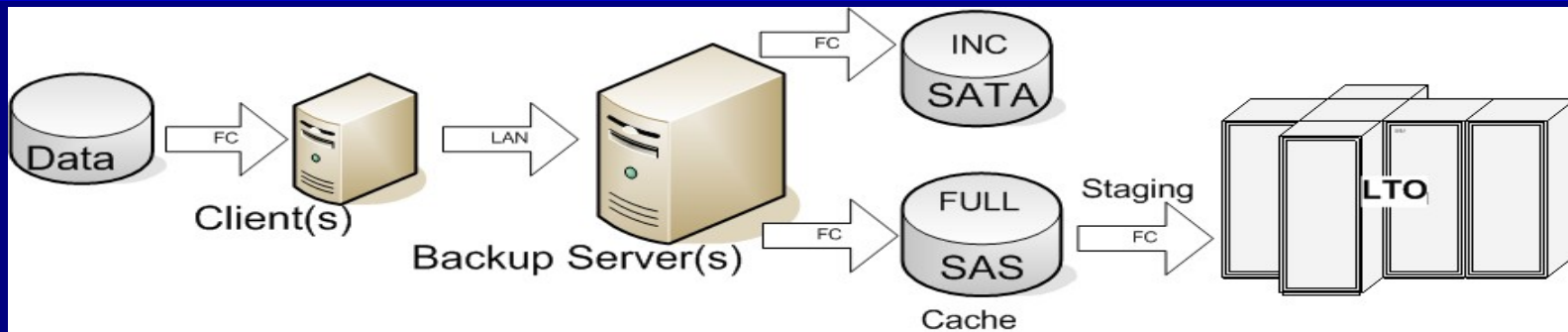
(Basic unit: Dual controller systems, 24x 500 GB Sata disks in four (4+2) Raid6 sets, one 4Gbit/s FC link per controller. Deploy 2 – 4 units per beamline. File- and compute-servers with direct 4Gbit/s FC connection.)

GPFS at SLS (Synchrotron) Beamlines

■ storage issues

- Raid controller HW reliability was an issue. Stable since ~3months
- Raid controller performance with full load still not o.k.
- read starves under heavy writes – *seems to be a general pattern*
- 15-20MB/s per 7200k Sata disk, reduced by 50% if purely random. 15k FC disks perform 2-2.5 times better. *Do you agree ?*
- Still searching for the „right“ storage benchmark to get reliable estimates for production performance. *What's your approach ?*

Backup: Use disk cache, LTO4



SATA: RAID6 (2*20TB)

SAS: RAID 50 (2*4TB)

EMC NetWorker 7.4(64bit) with Disk-Backup Option (for the staging process)

2 * IBM TapeLibrary 3584, LTO3 (80MB/s,400GB) LTO4 (120MB/s,800GB)

Clients: 120 Windows, Linux, UNIX

Windows 2003 Server (64-Bit) with

2 * HP ProLiant 380G5 with 2*Intel dual-core 3GHz CPU

2/4Gbit FC SAN, Gbit LAN to Clients

Backup: Experience

Good

- LTO streaming speed (120MB/s with LTO4)
- GBit LAN full used
- low cost & scalable with tape and SATA
- fast restores
- EMC Networker highly scriptable and customizable, stable and very scalable

Bad, Risks

- In-depth EMC Networker know-how required
- high INC amounts needs more SATA Storage
- good monitoring/alerting needs self scripting or additional product

Questions ?

