



Wir schaffen Wissen - heute für morgen

Paul Scherrer Institut

Derek Feichtinger

PSI Site Report for HEPiX 2011

1. Storage for fast detectors
2. GPFS at PSI
3. Central virtualization service
4. PSI archive migration
5. PSI Scientific Linux + Puppet configuration management
6. CMS Tier-3 cluster
7. PSI AFS overview

PSI develops new 2d photon counting X-ray detectors to be deployed at the Swiss Light Source (SLS).

Current detectors: compressed rates of up to 72MB/s
→ New detectors: uncompressed rates of **3+ GB/s**: (40fold increase)

Started to use HPC style storage for beamlines.

- DDN S2A9900 couplet + 5 enclosures (300 SATA 1 TB disks)
- Infiniband QDR network
- GPFS
- Can sustain 4+ GB/s random (4 MB block) IO on this config
- In order to fully profit from this
 - Trying to get users to move to HDF5 based file formats and reasonable file sizes (> 50 MB), away from Mio of small files
 - Currently data sets still typically consist of 10000s of 100 kB - 20 MB files
 - Parallelization of codes where applicable

- GPFS was chosen at PSI in 2006 as the best suited at this time
- very mature and stable. We made good experiences over the last 4 years
- Commercial. Requires licenses from IBM
- GPFS-3.2.1 in production on SL4 and SL5, GPFS-3.3 in testing for SL6
- Beamline environment
 - Using CTDB for achieving HA for NFS and CIFS based access for beamline clients
 - Flexible [GPFS multicluster setup](#)
 - multiple separate GPFS environments make use of the same consolidated storage infrastructure
 - client environments can be isolated by VLAN techniques (all can import assigned storage from central GPFS, but none can see each other)
 - DDN S2A9000 couplet, 5 enclosures, 300 1TB SATA disks
- Local PSI HPC environment
 - DDN S2A9000 singlet, Sun StorageTek 6140 (old). Total size: > 350 TB and growing (+ additional controller)
 - used for HOME and SCRATCH

VMwareESXi v4.1 on NetApp Metrocluster 3240 storage

- Management through VMware vCenter
- Features
 - Snapshots integrated with underlying storage
 - Automatic failover
 - NetApp well integrated with vCenter management console
 - Live migration of virtual machines (very useful for upgrading the underlying server infrastructure)
- Archiving via SnapVault onto a NetApp 2040. Some volumes further archived to tape
- Using 4 clusters of 12 servers each
- Currently running 130 VMs.
- In the process of virtualizing all suitable services and freeing up / getting rid of old HW.

- Needed a new archive system solution
 - Old system EMC DiskXtender reaching end of life
 - New high resolution detectors at PSI require larger archive.
 - We had about **1 PB** at end the end of 2010

Challenge:

- migrating 780 TB of old LTO2/LTO3 tapes to new LTO4 tapes while keeping up production services and access to data

EMC DiskXtender for Unix/Linux **to** Oracle/Sun SAM-FS/QFS V4.6

REQUIREMENTS

- good scalability
- flexible and secure access (ideally real file system), supporting ftps, nfs4, ...
- reasonable migration path inflicting minimal impact on service/users for migration of all old LTO2/3 DX data tapes to LTO4

Choice fell on: Oracle/Sun SAM-FS / QFS

- Good market penetration (~3.000+ installations)
- Available on Solaris/x86 (not just SUN Sparc)
- Open format for customer flexibility and guarantee for long-term accessibility
- Moving towards object based filesystems, billion of files

Migration Partner: *HMK Computer Technologies GmbH*, Kronberg -
Germany

large experience and toolset for migration from DiskXtender to SAM-FS /
QFS

SAM-HSM Migration Suite from HMK Computer Technologies

- Fast initial conversion of meta data
 - _Production-ready SAF-FS/QFS file system after one weekend
- Able to include old unmigrated tapes in native DX format
 - _need not operate old system in parallel!
- Offers two migration methods to be used concurrently
 - _Fast *Bulk Migration* for whole tapes in a constant background process
 - _Direct access and migration of files on old tapes, when requested by a client

Migration:

Start metadata migration	Jan. 11 2010
Start productive operation	Jan. 13
Start bulk data tape migration	Jan. 14
	→ migration of up to 7TB of DX tape data per day
End of migration:	May 9 2010

Server SUN X4600M2 4*QuadCore CPU 2.7GHz, 16GB RAM
2x4Gb FC Dual Port HBA
OS: Solaris 10

Storage SUN STK6140 FC, 2xFC-Contr., 2x1GB Cache, 48x300GB FC-AL HDD 15k
(among these 8.7 TB for Diskcache, 1.6 TB small file disk archive)

Library two IBM TS3500 Tape Libraries (Type 3584)
with 2 x LTO3 Tape Drive, 3 x LTO4 Tape Drive
2876 Slots (2436 for the archive)

Tape Library Problems after adding a new S54 High Density Frame and changing Library operation to ALMS (Advanced Library Management System)

- **Exchange Media** SCSI Command not supported anymore by new HD Library
 - _in SAM-FS the use of this command is the default ⇒ complete showstopper
 - _Workaround implemented by HMK to get system working again
 - _**SUN case to fix remained open for months and never got solved (closed without providing a solution)**
- We notice some problems with library and drive operation in the new HD frame (dropped tapes. Happened four times over course of 1 year)

Scientific Linux PSI (SLP):

- SL Repos + RPMS built at PSI → SLP repos
- Basic Installation via PXE boot + customized kickstart
- Desktop and server configuration by **puppet** (SLP5)
 - Installation boot prompt arguments choose puppet profiles to use
 - Desktop update: Auto update procedure performed monthly by client-side cron

SLP systems:

- Desktop Hardware: Fujitsu-Siemens
- Server Hardware: HP, Sun, IBM, Dell
- Servers: 150 SLP5 (30 % virtual)
- Desktops: 700 SLP5, 100 SLP4
- Compute Nodes: 120 SLP5
- SLP6 release in June 2011

```

Paul Scherrer Institute
Network Boot and Installation Server
PSI

- Press <ENTER> to boot from the local hard disk
- Press [F2] to install Windows XP (32bit)
- Press [F3] to install PSI Scientific Linux 4 (32/64bit)
- Press [F4] to install PSI Scientific Linux 5 (32/64bit)
- Press [F8] to install PSI Scientific Linux Server (32/64bit)
- Press [F6] to boot a diskless client (Linux and Windows)

For a productive Linux system we recommend PSI Scientific Linux 5.

- Use the function keys listed below for more information.

[F1-Main] [F2-SL3] [F3-SL4] [F4-SL5] [F5-Help] [F6-Diskless] [F8-Server]

boot: s15432desktop hostid=tukan54-32:Desktop/Common/Unstable_
  
```

Fig. 1: PXE network boot screen

Puppet infrastructure:

- 3 Puppet servers for 700 clients (Desktops and Compute Nodes)
- puppet version 0.25.1
- puppet client scheduled by cron, staggered (6 clients / min).

Filesystem:

- OpenAFS 1.4.12, 1300 users

Printing:

- Cups server on Mac OS 10.6
- 200 Network printers

- joint Cluster for CMS physicists of ETHZ, PSI, and University of Zurich
- Compute nodes: 20 SUN x6270 blades with 2*Xeon 5560 (160 cores total)
- Grid Components: **SE (dcache), UIs, CMS-VOBox, BDII**
 - For easier maintenance, it was decided to deploy **no CE**. Access only to internal users via SGE batch system
 - Aim of the system anyhow on fast turnaround and latency than on keeping all CPUs constantly filled
 - User applications (e.g. CRAB job submitter) integrated with grid + local job submission
- Using PSI standard VMs for service virtualization
- 6 UIs
 - Also used for Interactive graphical work via NX

- dCache SE (v1.9.5):
 - 2 SUN x4150 head nodes for main services and DB/namespace
 - 5 SUN x4500 “Thumper” (48*500 GB) + 5 x4540 “Thor” (48*1TB) file servers / door nodes
 - x4540s have flash cards for OS, so no 1TB disks need to be wasted for OS
 - Solaris10, ZFS Raidz2 pools
 - using 4 * 1Gb/s bonded eth interfaces
 - OS installations through PXE using a virtual **Jumpstart** server, configuration through **puppet**
- Home directories via NFS on an x4500 with Solaris10/Raidz2
 - Currently use 100-150 GB home directory size, very much appreciated by users, 3.5 TB in total
 - daily **ZFS snapshots** mounted into user namespace (file recovery by users)
 - ZFS **incremental snapshot backups** allow efficient backups to another “Thumper” which also can serve as a fast replacement

PSI AFS Cell

- 1533 Accounts
- 3893 Volumes (882 without backup to tape)
- 103 TB total storage, 50 TB used
- 79 Mio files (38% < 1MB, 30% 1-8MB, only 5% > 1GB)

Servers

- 3 Database servers, 7 file-Servers
- SL5.1 with OpenAFS 1.4.11

Hardware

- 13 HP ProLiant DL360G5 servers
- 8 SUN STK 6140 (0.5 / 1TB SATA), RAID6
- 1 LSI ProFibre 4000R (400 GB FC disks), RAID5
- all Storage direct attached

Kerberos Unified authn against PSI Active Directory

But had to keep old Heimdal Kerberos up for special cases where users rely on possibility of renewing their tokens for longer periods (policy issue with AD krb)

Backup AFS volume dumps (vos dump) to disk cache

Copy to tape via EMC Networker

one full backup per month + daily incremental backup

15 TB Full backup size

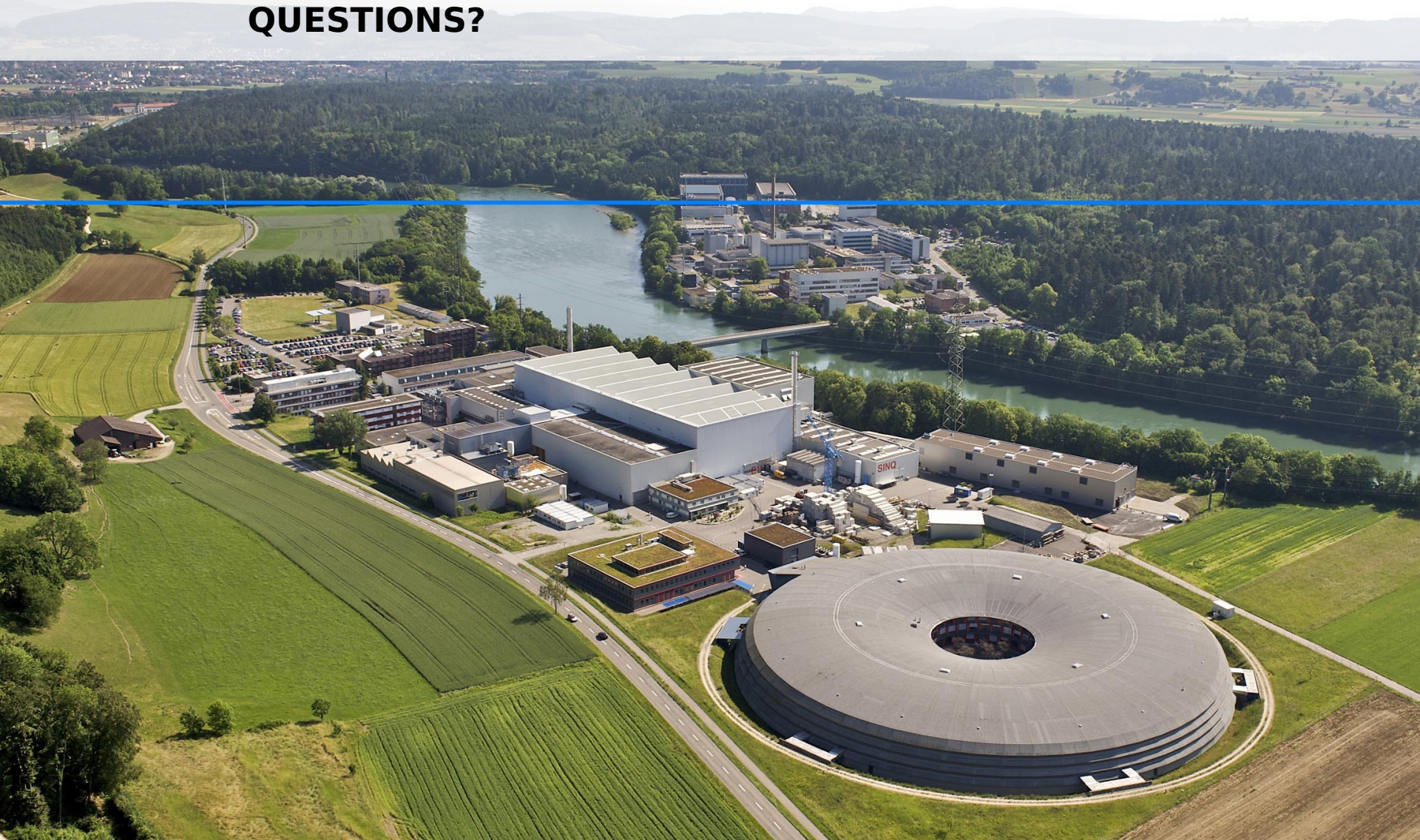
~4 TB sum of increm. backups per month

Hardware 2 Sun X4500 (48*500 GB) Servers for disk cache

Sun Solaris 10, ZFS/RAIDZ2 with file system compression

- <http://www.psi.ch/>
- Presentation: [DAQ_at_SLS](#), Heiner Billich, ESRFUP WP10 workshop 2011
- [PSI CMS Tier-3 Wiki](#)
- derek.feichtinger@psi.ch

Thank you for your attention!
QUESTIONS?



Thanks to The whole AIT team of PSI

