CASPUR Site Report

Andrei Maslennikov

Lead - Systems

April 2007 - Hamburg

Summary

- Principal computing means
- State of infrastructure
- Works in progress
- Some other plans

Principal Computers – 2.45 Tflops

IBM SMP (1.2 Tflops):

- POWER-5 cluster (21 nodes, 168 p575 CPUs at 1.9 GHz, RAM: 400 GB)
- Communication subsystem: High Performance Switch (Federation) – low latency, 1 GB/sec
- OS: AIX 5.3 ML4

Opteron SMP (1.1 Tflops):

- 2 clusters with 224 cores in total, RAM: around 400 GB
- Communication subsystem: Infiniband (800 MB/sec), Qsnet (1 GB/sec)
- Plans to migrate to new generation of IB (1.6 GB/sec) e Qsnet (8 GB/sec) in 2007/8
- OS: mostly SuSE Linux, also tried the Microsoft CCS

NEC SX-6 (vector, 0.07 Tflops):

- 8 CPUs, RAM: 64 GB
- Seemingly few flops, but compared to the PWR5 one can get a speedup of 10 for some codes

HP SMP (0.08 Tflops):

- EV7 with 32 CPUs at 1.15 GHz, RAM: 64 GB, Tru64 5.1B+ - will soon be phased out

New platform: under discussion

- A working group was set up in February, goal is to find the most efficient match to our needs
- Will compare 5-6 variants, have to remain inside the budget

Infrastructure

Data Exchange Areas:

- NAS-1: AFS (home directories, software repositories, project areas) – 6TB (will soon be on R6)
- NAS-2: NFS (large files) – 30 TB R6

High Performance Storage:

- IBM cluster: GPFS – 900 MB/sec sustained – 12 TB, will be adding another 20-30 TB in summer
- Opteron clusters: PVFS2/Qsnet under test on one cluster, Lustre/IB is being prepared on another

Tape Systems:

- IBM 3584 Library: 8 drives (4xLTO2 + 4xLTO3), 250 slots – around 220 TB compressed
- 2007: upgrade of 2-4 drives to LTO4
- Used by backup applications for all areas (TSM and AFS native), and by the Staging System

Interconnects:

- Wirespeed bonded GigE outside FW for data exchange areas, mixed GigE/FE for the rest under FW
- 100+ Qlogic SAN ports (2/4 Gbit); have reached the E-port limit for a ring scheme, will consider a star

Authentication:

- Kerberos 5 (Heimdal), will be migrating to MIT as it fits better our way of doing U*X/Windows integration

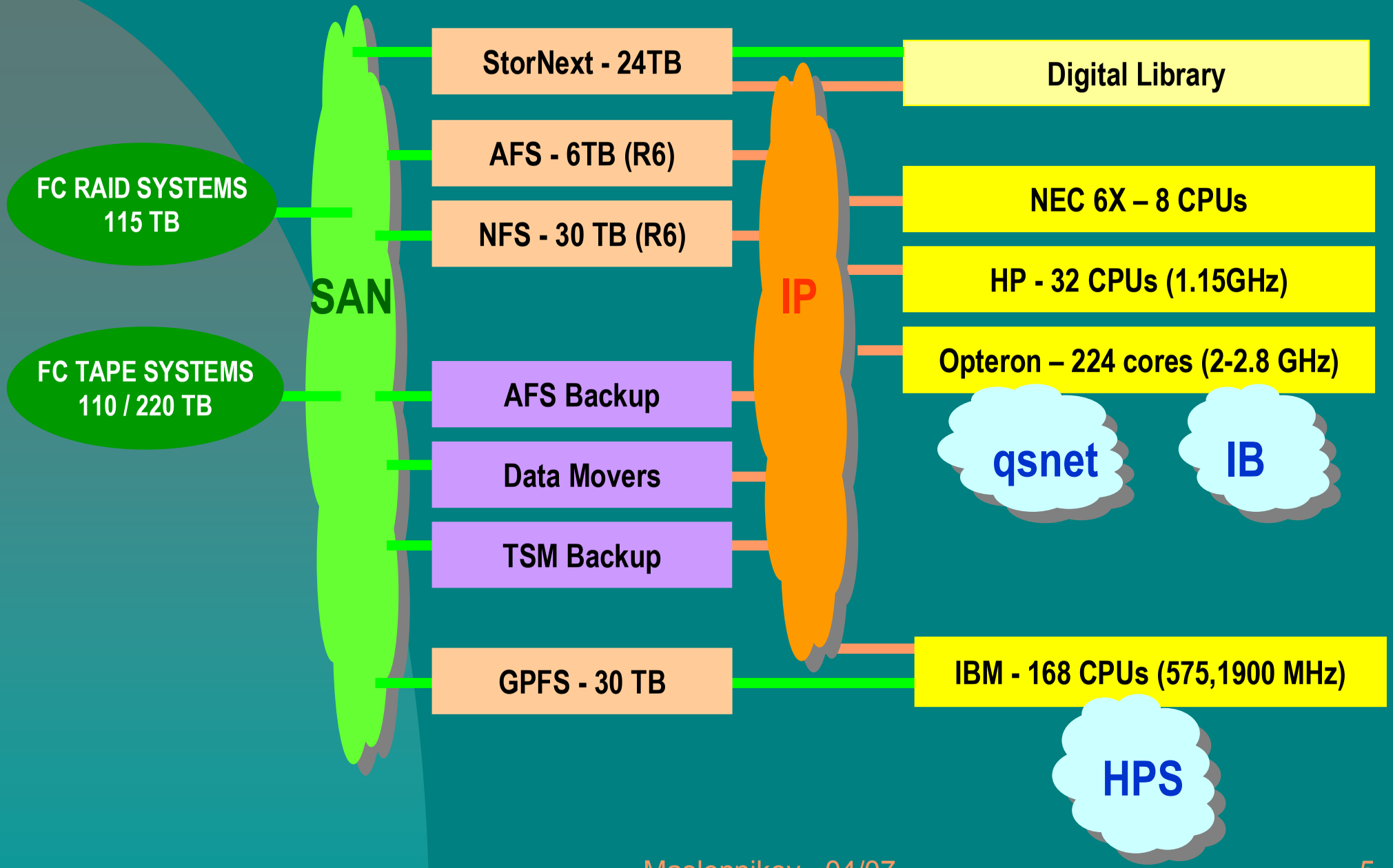
Load sharing:

- SGEEE (IBM,HP,NEC), PBS(Opteron)
- Currently under test on IBM cluster: Torque Scheduler + MOAB Workload Manager

Database:

- DB2 for user administartion and accounting; will soon be migrating to new hw base and DB2 v.9 Express C

CASPUR: principal resources in 2007



450 Million files

- Our Digital Library currently contains 450 Million small files in 12 TB (2 online copies)
- Must be set up on a cluster with a shared SAN filesystem for redundancy and load balancing
- Every two-three years the disk base has to be revisited, copy operations are “mastodontic” and take a lot of time: good moment to change the FS, if needed.
- For this service we have to use a commercial file system product with the official support

Epoch 1: Sistina GFS (now RedHat)

- Quite reliable, but support became virtual after Sistina was bought by RedHat
- Had to act very fast after several hangs due to the disk problems in the end of epoch
- Briefly tested Polyserve and StorNext, and preferred the former as it seemed to be more performant for small files

Epoch 2: Polyserve (now HP)

- Few problems for almost two years, but then started to suffer the spurious FC port fencing, sometimes in a “domino” sequence. Most probably this is a race condition under the ever increasing load, and Polyserve support was unable to solve it.

Epoch 3: ADIC StorNext (now Quantum)

- StorNext was in the meantime very well tested in our and other labs. Now migrating!

Virtual efforts

- Last year we decided to virtualise a series of not so important services like productivity web servers, management consoles, remote access gateway etc.

- We then studied 3 possible candidate solutions: MSVS, Vmware ESX and Xen. The first one did not support some essential hardware and we decided not to consider it. We could not discard neither Vmware nor Xen as both of them had good features.

- Xen is more performant, supports any hardware we might need and allows to save on the cost of host machines, but it is quite “spartan”. Vmware based installation is definitively more expensive but it provides excellent management, backup and monitoring tools.

  >>>> We hence decided to use a balanced combination of Xen and Vmware

- Currently decicated 3 powerful 4-core Intel machines with Areca R6 disk base to Xen and one large memory machine to Vmware, and are gradually populating them.

Other works in progress

Faster Data Link for NEC

- We have to ensure that the data produced on NEC SX-6 may be analysed on other clusters. However the native NEC implementation of NFS does not allow us to exceed 18 MB/sec on the Gigabit Ethernet line with Jumbo frames enabled and data transfers are sometimes are too slow. RFIO, tested with the help of CERN did not yield better performance.

- We then pointed at BBFTP, and this was a success: we now can operate at 50-70 MB/sec. In preparation: a tool for the users to let them initiate the data transfers to and from NEC with the help of this technology.

CASPUR Hadrware Verifier

- In 2007 we will be releasing a small but useful solution for burn-in testing of new hardware. A small image containing the test suite may be booted off USB or via network. The netboot variant includes a simple web interface with minimal bookkeeping means needed to keep track of multiple machines under test. The propotype had already been successfully tested.

Some other plans

Heimdal Escape Tool

- The CASPUR main Kerberos 5 service is based on Heimdal. We however maintain several other K5 realms, and they are all MIT based. In several occasions we have noted that MIT implementation has less integration problems with the Windows world; we also know that MIT version has more developers behind it, and hence potential problems are being addressed faster. So we would like to migrate from Heimdal to MIT.
- Heimdal provides a conversion tool for MIT database (i.e., you may easily migrate from MIT to Heimdal). But you cannot do it in the opposite directon… We will be investigating this and hope to be able to produce an appropriate solution.

ControlHost II

- Our historical product, ControlHost, was recently benchmarked on modern hardware base, and this was a success. This data handling package provides means for creation of complex distributed data environments it is still very actual.
- Encouraged by a series of exchanges with some experiment planners, we decided to revitalize this software, improve it and add several new features. Planning for this work has already started.