Prague Tier-2 storage experience

Ji Hork, Ji Chudoba, Tom Kouba
E-mail: koubat@fzu.cz
Institute of Physics, AS CR, Na Slovance 22, 182 21, Prague, Czech Republic

Abstract. Main computing and storage facilities for LHC computing in the Czech Republic are situated in Prague Tier-2 site. We participate in grid activities since the beginning of European Data Grid. The recent years were significant in the growth of our storage capacities. In this article we present the decisions in storage solutions, comparison of configurations of the hardware and performance tuning of the software. Our storage facilities vary from enterprise-level SAN infrastructure to in-house built cheap storage servers. We present a comparison of these solutions from the performance, configuration, management and monitoring point of view.

1. About Prague Tier-2
The Prague Tier-2 site was established in 2002 and it participated in the grid activities from the very beginning. First 34 computers were equipped with two Pentium III processors and 1GB RAM each and the first storage array contained 15 SCSI disks with total capacity of 1TB. In the middle of 2010 we have 336 computing nodes with 2630 CPU cores (with approx. computing power of 20500 HEPSpec units). The disk storage for grid usage is about 350TB available via DPM system [1] and 70TB available via NFS.

2. Overview of storage procurements
Our Tier-2 center usually buys new computing equipment every year in one big tender. Over the last 4 years we have obtained various storage solutions and currently we run the following big\(^1\) storage solutions:

| Year | Product         | Infrastructure | Type of disks | Number and size of disks |
|------|-----------------|----------------|---------------|--------------------------|
| 2007 | HP EVA 6000     | FC             | Fibre         | 64*0.5TB + 49*1TB        |
| 2009 | Overland Ultamus 4800 | FC             | SATA          | 144*2TB                  |
| 2010 | Nexsan Satabeast | FC             | SATA          | 126*2TB                  |

2.1. HP EVA 6000
This disk array was the first SAN solution installed in our infrastructure. The front-end machines are HP DL360s. The schema of this SAN is depicted in the figure 1. There are four main parts of this SAN:

- Front-end machines (ha1, ha2, ha3, sam3, storage5, goliasx98)

\(^1\) Regarding the year of the purchase
The front-ends serve for the following purposes:

- Virtual machine images (on host machines ha1, ha2, ha3) - using 1.3 TB
- Sam cache[7] for D0 experiment (host sam3) - using 6 TB
- NFS server with user home directories and experiment software areas (storage5) - using 18 TB
- DPM disk server (goliasx98) - using 27 TB

The rest of the raw capacity is used as the redundant space for ensuring reliability by the HP's VRAID5 technology.

We are very satisfied with this storage solution. It is pretty stable and reliable, we replaced only 6 disks over 3 years and we had one power failure in one of the two FC switches (neither of these problems caused downtime of the disk array). With 113 disks this means 494940 hours MTBF\(^2\) which is less than expected 1 million hours MTBF (estimated by the manufacturer), but it is still better value than for other (following) disk array procurements. The management interface runs on Windows server (part of the array itself, SMA-evickat in figure 1) and it is well-arranged and intuitive.

2.2. Overland Ultamus 4800

This system consists of 2 full boxes and one expander box. It was purchased in a tender in 2009 oriented on DPM storage for HEP experiments and tape storage for backup archives. After a good experience with reliability and performance of FC SAN we decided to go down this road again. This time the front-end machines are IBM x3650 servers. Two of them export the disk space for DPM and one is the front-end for the backup system. As another part of this SAN

\[^2\text{Mean time between failures}\]
Figure 2. Second SAN infrastructure

infrastructure (depicted in figure 2) we have smaller disk array Ultamus 1200 (13*2TB disks), which serves as a backup cache for the tape storage.

- Front-end servers for DPM system (dpmpool1, dpmpool2) - using 114 TB
- Front-end server for Legato backup system (backup) - using 6TB from ultamus1200 as a prestage cache and tape robot as the archiving medium
- Fibre channel switches (fcswR081, fcswR082)
- Disk array boxes (ultamus4800down, ultamus4800up, ultamus1200)

This array is not as reliable as HP EVA. We have had no data loss but there were more issues in its lifetime:

- first delivered box was not working
- there were few firmware upgrades needed (mainly for confusing false alarms)
- the management card loses network connection once in few months and complete reboot of the array is needed
- the management web interface is not comfortable (runs on port 9292, uses pop-up windows)

2.3. Nexsan Satabeast

The third SAN infrastructure in our computing center is built on Nexsan hardware. We have 3 storage boxes with 42 disks each and four front-end servers. Two of these servers are IBM x3650M2 and two are HP DL360 G6. Out of these four front-ends three work as DPM disk nodes and they publish 5/6 of the entire disk space. The rest is published via one NFS front-end.

- Front-end servers for DPM (dpmpool4, dpmpool5, dpmpool6) - using 170 TB
- Front-end server for NFS (storage7) - using 34 TB
- Fibre channel switches (fcswL021, fcswL022)
- Disk array boxes (Satabeast1, Satabeast2, Satabeast3)

The reliability of this storage system is satisfying. There were only two disk replaced during the ten months of the production period (this includes the first few weeks of burn-in tests).
3. How we measure storage performance

3.1. The main test

When deciding what storage should be bought we try to find a storage solution that gives as much disk space as possible for the least price possible but with reasonable performance. This golden rule is way too general. It is difficult to say what is the reasonable performance for our workload although we know what kind of I/O should we expect (mainly D0, Alice and Atlas jobs). It is also important to formalize the measurement properly so we can give the competing companies a clear instructions what they should measure for us.

We decided to use iozone\[2\] tool. After several tenders we chose iozone benchmark for its stability, reliability and portability. It is also widely used tool so some rough performance results can be found for several storage systems.

The evaluated disk arrays can have very different size of the available storage per front-end server, which also means different storage size available via given network connection. So our requirements are compiled from different requests but still simple enough:

- Read speed at least 6.4MB/s per usable TB (as reported by iozone)
- Write speed at least 4.8MB/s per usable TB (as reported by iozone)
- Sufficient network connectivity on every front-end server. Server presenting -
  - less than 10 usable TB must have 2 * 1Gb
  - 10-20 usable TB must have 4 * 1Gb
  - more than 20 usable TB must have 10Gb NIC
- Health monitoring: SNMP (or Nagios sensor)
- Hot-plug disks

Results for the storage solutions already mentioned:
| Product     | Network         | Read performance | Write performance | Monitoring | Hot-plug |
|-------------|-----------------|------------------|-------------------|------------|----------|
| EVA 6000   | 4*2Gb + 2*1Gb   | not measured     | not measured      | SNMP       | Yes      |
| Ultamus 4800 | 2*10Gb         | 8.65MB/s per TB  | 4.48MB/s per TB   | SNMP       | Yes      |
| Satabeast  | 4*10Gb          | 9.42MB/s per TB  | 8.75MB/s per TB   | SNMP       | Yes      |

Notes:
- 2Gb means two 1Gb interfaces bonded together
- the performance is measured on the front-end servers

4. Bunch of small inexpensive 1U storage servers
When these critical requirements have been established and formalized in tender rules we realized we can try to meet such requirements with an in-house built storage server. Recently there have been good options to buy 1U chassis, simple motherboard with C7 or Atom CPU and SATA controller with enough (at least 4) SATA channels.

4.1. First try
As the first attempt in 2009 we purchased two instances of:
- VIA Epia SN 18000EG: 3GB RAM, 1x Via C7 1800MHz (only 32bit ⇒ not advisable for DPM)
- 4GB CF Card as the system disk
- 4x 1TB (WD Caviar Black) in RAID5 - /data and /var
- no hot-plug
- only one 1Gb NIC

The performance numbers are: 38.8MB/s per TB for reading and 24.5MB/s per TB for writing. This relative performance delivers much more data per terabyte than traditional big array with "high" throughput. It shows that this approach is very promising.

4.2. Second try
We were satisfied with the performance of the VIA solution but we wanted the new servers to support 64bit instruction set so we can install DPM from gLite 3.2 on it. So we purchased three servers with the following configuration:
- MSI IM-945GC, 2GB RAM, Intel Atom 230 (64bit ⇒ suitable for DPM)
  - the only MSI mini-ITX with Atom and 4 SATA ports available at the time of purchase
- 4GB CF Card as the system disk
- 4x 1.5TB (WD Caviar Green) in RAID5 - /data and /var
- 2x 1Gb NIC
- still no hot-plug (no suitable 1U chassis available)

Two of these servers work as NFS servers for experiment data (Atlas, Auger) and one as a DPM disk node. The performance is not as good as with VIA chip-set but still better than our tender requirements: 30.0MB/s per TB for reading and 8.18MB/s per TB for writing. The latter number is not so good because the controller chip 82801GB/GR/GH gives worse performance when writing to all 4 disks.

In this case, our request for 64bit platform with 4 SATA ports resulted in motherboard with poor controller so the performance went down comparing to the previous year. But it is still above our tender requirements.
4.3. Management challenges
There were no problems installing and managing these nodes. We simply put them into our existing infrastructure. We installed them from PXE with minimal kickstart, we configured them via CFengine[3]. We monitor them by nagios and its plugin check_linux_raid.pl

There were no problems with the performance of SW RAID on C7/Atom CPUs. There is one issue with new 2TB disks: they sometimes claim they use 0.5kB blocks internally but in fact they use 4kB blocks so it is important to set up the software raid correctly and to align its metadata properly. To ensure this, we use the following command:

```
mdadm --create /dev/md1 -e 1.0 --level=5 /dev/sd{a,b,c,d}1
```

4.4. Problems encountered
During the performance measurement and production run we encountered one big problem on the servers with the MSI chip-set. The performance was sometimes very low. We suspected bad sectors on disks or one completely bad disk. The problematic behavior can be seen in figure 4 (generated by zcat utility). It turned out that the problem is in the disks. The performance of the disk was significantly degraded (e.g. the reading performance was about 5kB/s) when more than one thread accessed the disk. This behavior was not easily predictable but it appeared more often when the I/O load was heavy. According our experience (and user discussion at [4]) the whole series is bad.

4.5. Conclusion and Future work
We are well satisfied with the reliability and performance of the hardware, but there are two obstacles that prevent us in broader deployment:

- Missing hot-plug support
- Missing remote management (BMC)

Recently there have been new products presented that should solve these issues. The first is a chassis with hot-plug support for SATA disks [5] and the second is a motherboard with better
I/O throughput and support for additional BMC card [6]. So we hope to address the mentioned issues this year and deploy this kind of storage solution more widely.

Acknowledgments
This work was supported by Academy of Sciences of the Czech Republic.

References
[1] Jean-Philippe Baud Disk Pool Manager https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm
[2] Don Capps, William D Norcott IOzone http://www.iozone.org/
[3] Mark Burgess CFengine http://www.cfengine.org/
[4] Western Digital WD community web http://community.wdc.com/t5/Desktop/WD15EADS-00P8B0-Really-slow-Or-am-I-just-crazy/td-p/1547
[5] mini-itx Travla TE-1160 http://www.mini-itx.com/store/?c=33#p3819
[6] mini-itx ASUS Hummingbird http://www.mini-itx.com/store/?c=65#hummingbird
[7] Fermi National Accelerator Laboratory The SAMGrid Project http://projects.fnal.gov/samgrid/WhatsSAM.html