LIFE (OF BIG STORAGE) IN THE FAST LANE

I. Andrian[†], R. Passuello, I. Gregori, M. Del Bianco, Elettra Sincrotrone Trieste

Abstract

Parkinson's law: *work expands so as to fill the time available for its completion.*

Corollary applied to computers: *Data expands to fill the space available for storage.*

As time goes by, your storage will become too small. And too slow. And the pace is accelerating. What can we do about it? This paper presents Elettra Sincrotrone Trieste's experience in managing scientific data generated by its two lightsources, using state-of-the-art technology and tools, taking them to the limit just to discover some shortfalls and weaknesses.

FACING THE PROBLEM

Lightsources like Elettra, a third-generation synchrotron radiation facility, or FERMI (acronym for Free Electron laser Radiation for Multidisciplinary Investigations), the new seeded free electron laser (FEL) accelerator operating next to Elettra, are using a lot of detectors which produce etherogeneous data in form of streams of floating point numbers or raw images. Analysis like Computed Tomography (CT) scans are being used more and more frequently in many beamlines which are striving to increase the quality of their experiments on samples. This increase in quality usually can be performed by having better radiation light characteristics, by increasing the number of images taken per second and, last but not least, by using bigger detectors. From the historical "kilopixel" sensors, we are now in the era of various megapixel (MP) detectors, usually ranging from 1MP to 13MP, operating at higher frequencies compared to the past times (e.g., upgrading from 5 to 10Hz or even 120Hz, now entering the Khz range). In time-resolved studies (4DCT), several tens of datasets can be collected in sequence, yielding TB of data to be stored and managed (10 TB/day with Elettra and up to 100 TB/day with the planned Elettra 2.0). Advanced algorithms and processes are being developed to handle this huge amount of acquired data before it is even stored and used for further analysis [DiamondTC]: however, even reducing the amount of data, the result can be in the order or TB/day per active beamline which means that 1PB of data per year is nothing less than reality right now.

However, it must be stressed that not all these data are here to stay. Every investigation needs to be analysed and, eventually, part of the data will be deleted because useless or redundant. Lossless data compression algorithms can highly reduce the size of images.

The amount of storage needed for these kind of jobs is not the only problem: as anticipated, the sampling frequency is increasing which, also considering the bigger size of data, brings to the high throughput requirements of the storage system that will handle the data itself.

Huge, fast and... cheap, of course! These are the easy requirements for the storage systems at any lightsource facilities these days.

EVOLUTION OF STORAGE AT ELETTRA

Prehistoric era: minicomputers

Elettra was built in the early nineties, starting operating with its first beam in 1993. At that time, the main storage facility for scientific data (as well as the technology for data analysis) was based on a number of DEC AlphaServer 2100 and VAXstation machines running OpenVMS and, lately, Tru64 Unix (Fig. 1). When the use of personal computers became common, these were widely used as storage systems at the beamlines against the DEC servers; as a side effect, this led to a very etherogeneous situation with no centralised standard of access for the data. After a number of years this anarchic situation came to an end: a more organised approach was mandatory, and the needs for a high performance, reliable, centralised storage were born.

Figure 1: DEC AlphaServer 2100

Middle age: SANs

When the hype of data storage anarchy was gone leaving only the dark face of the trend, the new Storage Area Network products looked like the cure for any disease. EMC[2] entered the Elettra datacenter with a CX4-240 machine and everybody was happy with its assistance, internal redundancy and scalability. After a bit of use, however, some important limits of the system became clear to the operators: even if rock solid, so that a good lifetime of about ten years could be foreseen, the performances were not as needed. Expandibility was one key factor of the system, however it came at a cost: every original spare part, even if similar or even identical (i.e.,

[†] ivan.andrian@elettra.eu

rebranded) to many other products on the unbranded market, costed twice. When it came the time of a SAN expansion, right after the end of the included support period, the estimated cost of support renewal and new disks was absurdly high. A quick market research suggested that it was cheaper to buy a new complete solution based on Commodity Off The Shelf (COTS) hardware.

A new hope: DFS on COTS

At the beginning of the new millennium the server area of the PC market was quickly evolving in technology and power, while the prices were going down. Linux was becoming more stable and powerful, gaining many features ready for the enterprises, in particular for the research centres where the IT departments have always been interested in state-of-the-art technology. The evolution of the storage at Elettra passed through dedicated fileservers based on the x86_64 architecture with a lot of disks onboard, managed by dedicated RAID controllers and usually exporting the volumes via NFS to the data acquisition workstations at the beamlines. Pretty soon the increasing number of such servers was causing problems in terms of mainteinability, not counting the issues when the volumes needed to be expanded, operation that could often lead to downtimes due to physical installation of new hardware. At that time, however, open source Distributed Filesystems (DFS) where production ready.

A Big Bench to accomodate all the data

Experiments with distributed filesystems at Elettra were performed in the past. The initial choice of IBM's General Parallel FileSystem on Linux (GPFS, now rebranded as IBM Spectrum Scale [GPFS]) was promising, with only minor stability problems when performing maintenance operations on problematic volumes. GPFS is always been a commercial product, but at the time it was offered at no cost for research facilities. Suddenly, IBM changed this license and the costs were not sustainable for our laboratory; it was then decided to move to an Open Source DFS. After some tests the choice went to Gluster [gluster], a promising technology initially developed by Gluster Inc. and then bought by Red Hat. Gluster exports one or more underlying filesystems to a cluster; both spanned and replicated (or mixed) configurations can be made, providing expandibility and redundancy. The simplified specifications of the Elettra *"Bigbench"* storage cluster commissioned in 2011 were the following:

- 7 Supermicro 4U servers;
- 24 x 3TB SAS disks per server (72TB raw per node);
- 2-copies mirrored volumes between the servers to cover possible failures or maintenance downtimes;
- total of 504 TB raw, 252 TB net;
- RAID 6 with dedicated controller on every server for in-node disk failure protection;
- LVM + XFS (lately, ZFS) as underlying gluster brick structure;
- dedicated 10Gbps network between the cluster nodes, 1Gbps network to the client workstations;
- glusterFS accessed by some XEN virtual machines used as frontend peers for the data acquisition machines to which the volumes were exported via NFS.

Before moving the servers into production some tests were performed on the machines in order to get some performance benchmarks. Considering the RAID controller, LVM and XFS, the *iozone* tool gave results around 1GB/s both in reading and writing, not considering the controller cache. However, performances were affected by the pile of layers, in particular Gluster and, obviously, by the network itself with its theoretical limit of 10Gb/s.

Analysing the whole architectural setup as depicted in fig. 2, it was decided to get rid of two layers: LVM and the RAID controller. This was made possible at first by moving to a different filesystem for the bricks: from XFS to ZFS. The performances of the latter, however, were negatively affected by the RAID hardware connecting the disks.

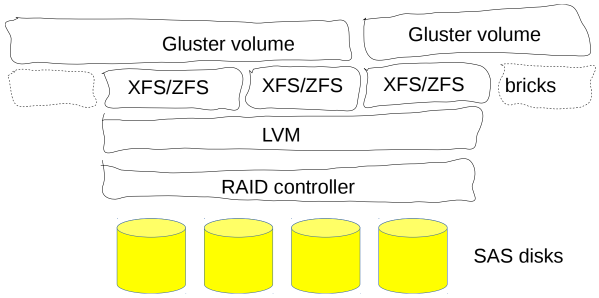

Figure 2: Bigbench architecture v1

Tests done with disks directly attached to a simple HBA demonstrated that the ZFS performances were at least at the same level of the RAID+XFS. Added to this, ZFS can act as an LVM with its concept of pools and volumes, so this layer became useless. The resulting configuration, used for additional nodes added to the cluster, is presented in fig. 3.

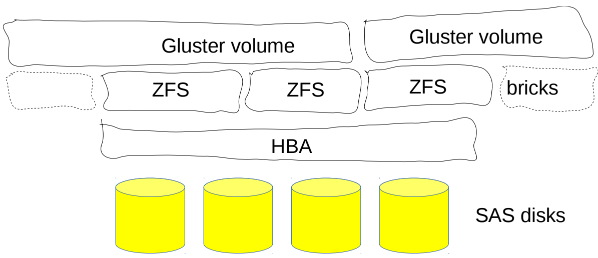

Figure 3: Bigbench architecture v2

This solution resulted to be successful and was able to accommodate the needs, in terms of space and speed, of most of the beamlines at the time. However, some strange behaviours were noticed by some beamlines in particular job configurations. The performances dropped down dramatically from time to time: the bottleneck was eventually recognised in Gluster itself. In fact, the Gluster architecture suffers of some limitations, in particular due to its userspace nature for GlusterFS and very poor performance when managing a great number of small files (due to its metadata in the filesystem-only feature). Modular expansion can be performed permitting horizontal scaling, with the constraints, for example, to keep the duality of servers if mirror volumes are to be implemented. These drawbacks, together with the introduction of new detectors with higher storage speed needs, brought the IT department to the decision of a new change in technology.

Relaxing constraints on a Sofa

Object-based storage, in contrast to filesystem-based storage, was in development for years and many projects were looking promising. Around the end of 2015 the CEPH distributed storage system turned into being stable and some internal tests validated it for its adoption at Elettra. The new storage cluster, *Sofa*, was born with these specifications:

- 4 3U Supermicro servers (plus other 4 added during a first minor upgrade);
- 20 x 6TB 7200RPM SAS disks per server (120TB raw per node);
- journaling dedicated on 4 SSD per node (one SSD serves 5 HDD);
- CEPH configured in replica 2 between the OSDs (storage units), with copy constrained to be on different nodes;
- total of 480 TB raw (960TB after the upgrade);
- additional "caching" tier for lower CEPH volumes on 3 servers with:
  - 20 x 2.5" 600GB 15000RPM SAS disks per server
  - journaling on 4 x 800GB HGST Ultrastar SN100 NVMe
- 40Gbps dedicated network for the cluster nodes, 1Gbps,10Gbps and 40Gbps to the clients;
- volumes exported via RBD to KVM virtual machines (Proxmox), used as frontend peers for the data acquisition machines to which the volumes are exported via NFS or SMB.

This solution proved to be a big success in terms of flexibility in allocating storage to clients, horizontal scale-up capability, high iops and I/O throughput. Even if we haven't been able yet to extensively take the system to its I/O limits with data read and written from the clients, during a CEPH self-healing repair due to a number of off-line OSDs we were able to see I/O rates above 2.7GBps in combined reading and writing (Fig. 4). However, during an operation session of FERMI's most demanding endstation, an incoming throughput of more than 800MBps has been seen.

Still, we were able to face some issues not foreseen at the time of design. In particular, the choice of having the journaling of 5 HDD OSDs on one SSD was not a good choice for two reasons. We were confident that SSD were very reliable and their speed was optimal for the journaling role, as suggested by the CEPH documentation itself [CEPHSSD]. However, we didn't consider both the total writes limit of SSDs (e.g., a Kingston SE50S37 of 100GB is guaranteed with 310TB of Total Writes – TBW – at 3 Drive Writes per Day – DWPD) and that when an SSD fails it affects 5 OSDs. After only three months of operation we realised that the expected lifetime of our SSDs was one year at maximum – not a great deal in the long term for a critical system like this. In only 4 days we also experienced a sudden failure of half of the SSD installed on three servers due to a manufacturing defect that was present in one production lot. As a result, in that short period we had OSDs in two server being put offline at the same time because of their unavailable journals. The number and location of the remaining OSDs were too small to guarantee the operation, and the system went into a state of mulfunctioning. Fortunately, the design of CEPH is rock solid so that, just after replacing the SSDs and a bit of sysadmin's tricks to help the data relocation routines, the OSDs came back online and the storage was put into service again with no data loss.

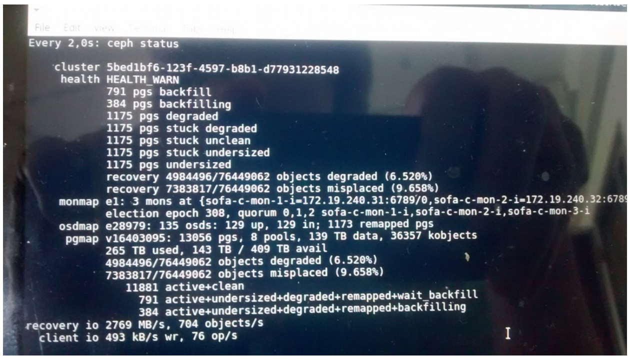

Figure 4: State of the CEPH cluster during a recovery operation

A bigger and better Sofa

We are now expanding the CEPH storage cluster, adding 8 new servers with 24 x 10TB SAS HDD (240TB per server), getting rid of all the SSDs and upgrading to the latest stable version of CEPH (Luminous), which brings Bluestore into stable state. Bluestore, the new data writing method, will guarantee higher performances even no SSD journaling. The total raw space will increase to ~ 3PB, wich means 1PB net with 3 replicas of the data (in different nodes). The volumes will be exported both in RBD and in CEPHFS.

ACKNOWLEDGEMENT

We would like to thank A. Curri which was the original designer of *Bigbench* and *Sofa*, and The Eagles for inspiring the title of this paper.

REFERENCES

[DiamondTC] Robert C. et al., A high-throughput system for high-quality tomographic reconstruction of large datasets at Diamond Light Source, 2015

[GPFS] https://www.ibm.com/us-en/marketplace/scale-out-file-and-object-storage

[gluster] http://www.gluster.com

[CEPHSSD] http://docs.ceph.com/docs/master/rados/configuration/osd-config-ref/?highlight=ssd