Testing and evaluating storage technology to build a distributed Tier1 for SuperB in Italy

S Pardi 5, A Fella 1,10, F Bianchi 3, V Ciaschini 4, M Corvo 1,2, D Delprete 5,13, A Di Simone 6, G Donvito 7, F Giacomini 4, A Gianoli 8, S Longo 2, S Luitz 9, E Luppi 3, M Manzali 8, A Perez 10, M Rama 11, G Russo 5, B Santeramo 7, R Stroili 2 and L Tomassetti 8

1 CNRS, Orsay, France
2 INFN and Università di Padova, Padova, Italy
3 INFN and Università di Torino, Torino, Italy
4 CNAF, Bologna, Italy
5 INFN and Università di Napoli, Napoli, Italy
6 INFN and Università di Roma Tor Vergata, Tor Vergata, Italy
7 INFN and Università di Bari, Bari, Italy
8 INFN and Università di Ferrara, Ferrara, Italy
9 SLAC, Menlo Park, USA
10 INFN and Università di Pisa, Pisa, Italy
11 LNF, Frascati, Italy
13 GARR, Roma, Italy

Abstract. The SuperB asymmetric energy e⁺e⁻ collider and detector to be built at the newly founded Nicola Cabibbo Lab will provide a uniquely sensitive probe of New Physics in the flavor sector of the Standard Model. Studying minute effects in the heavy quark and heavy lepton sectors requires a data sample of 75 ab⁻¹ and a luminosity target of 10³⁶ cm⁻² s⁻¹. This luminosity translate in the requirement of storing more than 50 PByte of additional data each year, making SuperB an interesting challenge to the data management infrastructure, both at site level as at Wide Area Network level. A new Tier1, distributed among 3 or 4 sites in the south of Italy, is planned as part of the SuperB computing infrastructure. Data storage is a relevant topic whose development affects the way to configure and setup storage infrastructure both in local computing cluster and in a distributed paradigm. In this work we report the test on the software for data distribution and data replica focusing on the experiences made with Hadoop and GlusterFS.

1. Introduction

The SuperB[1] project is an international collaboration that aims to build a new asymmetric e⁺e⁻ collider and the accompanying detector to measure the effects of physics beyond the Standard Model. SuperB plans to improve 100 times the luminosity of a current B factory by adopting a well-proven approach already in use in synchrotron radiation facilities, together with a novel “crabbed waist” design for the interaction region.
The collider will be built in Italy at the Tor-Vergata Campus by the Cabibbolab consortium [2], led by the Italian National Research Institute for Nuclear Physics (INFN). The consortium includes members from many countries geographically distributed worldwide.

The data analysis process will take advantage from the distributed computing technologies and will involve several partners institutions. In this framework the Italian cloud will be responsible of an important part of data storage and computation resource management due the central role played by the INFN in term of financial support and scientific engagement.

According with this strategic view, a new supercomputing infrastructure has been financed by the Italian Ministry within the ReCaS project[3]. This will foresee a federation of 4 data-centers distributed in south Italy starting from the already existing resources by an empowering program. Referring to the computational resource classification of the LHC experiments like Atlas or CMS, we expect that the ReCaS infrastructure will provide Tier0/Tier1-like services. In particular it will provide the facilities to store all the data produced by the SuperB collider and will provide protocols to share them among all physicists of the collaboration.

In this work we present the on-going activities and the first achievements in evaluating data-storage technologies. We focus on two main aspect i.e. storage at site level and storage plan for the geographical management of large storage area.

The rest of the paper is organized as follows. In section 2 we describe the SuperB requirements in order to show the scale of the problem, in section 3 we introduce the Data Storage R&D program defined in the SuperB community. In sections 4 and 5 we discuss about our activities with GlusterFS and Hadoop in Naples and Bari, presenting the current setup and the first cases-study. In section 6 we introduce the program for a geographical deployment of the studied file systems in the SuperB site and the current implementation. Further steps are discussed in conclusions.

2. The Requirement of SuperB

The SuperB collider will produce an unprecedented amount of data due the dramatic increase of the luminosity with respect to the past and present experiences of Babar and LHC experiments. During the definition of the technical design, a big effort has been spent to predict the scale of the CPU and Data Storage requirements in the long term. A first estimate (see fig.1) has been performed by extrapolating the numbers coming from the past experience. Additional correction factors take into account future foreseen improvements in skimming algorithms [4].

![Fig.1 – First draft of Storage Resource Estimates with an hypothetical start in 2016](image-url)
A large part of the produced data will be hosted and managed in south of Italy. To accomplish this goal, the ReCaS infrastructure define a spending plan including hardware and infrastructure elements i.e. racks, cooling systems and power cabinet.

The data Storage R&D program has become a crucial aspect in defining the acquisition specs; in the next paragraph we will show the technologies in evaluation at present time, the features and the strategy of resource exploitation.

### 3. The Data Storage R&D

Data storage is a relevant topic whose development affects the way to configure and setup storage infrastructure both in local computing cluster and in a distributed paradigm.

The evolution of file systems is one of the most prominent aspects of data storage field. Today we have a large set of file system solutions i.e. academic open-source, large scale file system under GPL license or commercial products. The large experience in HEP community provides important feedback on the most relevant features to take into account.

The other main topic of investigation is the future of the data storage in the multi/many-core era. Technological trends push to a computational model empowered by cluster of nodes with a large amount of computational cores. The number of application executions have increased together with the requested data I/O. CPU I/O starvation became a probable scenario; in the last 10 years the CPU flop/s vs DISK I/O per second (IOPS) is increased dramatically. This configuration can introduce a bottleneck in connection between computing and storage. Scalability and reliability issues should be solved.

The data storage evolution is also driven by the foreseen upgrade in network infrastructure; the work is currently in progress; the model moves to a Nx10Gbit/s connection server-side.

The new technologies could invalidate the current cluster and site setup, especially in according to the evolving data model. Current configurations appears not adequate to take advantage from the future hardware and need to be reviewed and reconsider connection between storage and computing.

The Storage R&D program of the SuperB collaboration is focusing to file systems evaluations. These activities have the following main goals:

- Provide feedback to the SuperB computing centers about the storage technologies that could be used to support the storage requirements of the experiment
- Study storage technologies that allows to build a unique distributed computing center that could setup a complete fail-over solution for SuperB data processing.

The storage technologies that are under testing at the present time are: GlusterFS[5], Hadoop-FS[6], NFSv4.1 [7]and EOS/XRootd[8]

The focus of those test are on:

- Fail-over characteristics
- Performance using data analysis applications
- Scalability while increasing the amount of storage served and the number of active clients
In the next sections we will show the first evaluation setup using GlusteFS and Hadoop.

4 GlusterFS in a 10Gbit/s farm

GlusterFS is an open source Network Area Storage (NAS) solution, capable to scale up several Petabytes. It provide a standard posix compliant interface, that implements all the functionality of a standard file system. Gluster works as a supervisor on existing posix file system (i.e. Ext4), introducing several functionalities for data distribution and native data replication among the nodes (Brick nodes in the Gluster naming). The GlusterFS does not provide any head-node, but uses a daemon running in each Disk node and manages automatically replication and data distribution through the posix extended attributes.

Starting from the local experiences in monitoring and managing network over cluster for distributed system [9][10][11] we tested the GlusterFS in Naples. We setup a 10Gbit/s cluster with the following hardware: 12 Server PowerEdge R510 (2 CPU quad-core Intel® Xeon), 32 Gigabyte Ram 1066 Mhz per node, 4 HD –Raid0 500 GB 7200 Rpm and a Broadcom NetXtreme I 57711 10GbE NIC.

The setup provides a GlusterFS area of 4.5TB raw, configured as a distributed file system with replica factor 3 for a total of 1.5TB. We tested the main characteristics of GlusterFS using SuperB analysis Job over simulated data.

In this framework, we made an experimental estimation of the bandwidth required for each core in processing SuperB analysis job on the GlusterFS. By running a single analysis job, we obtained a peak of 59 MB/s in reading with 13 MB/s of average.

As shown in the graph (fig.3) the analysis of the network usage shows that GlusterFS can scale over a 10Gbit/s in a small/medium cluster. In fact in our 12 nodes cluster, a full CPU load in the server farm (8 jobs per node) produces a network traffic lower than the maximum available bandwidth. In particular we measured 333MB/s as bandwidth peak that is 1/3 of the maximum 10Gbit/s network performance.
In order to test the impact of GlusterFS on computation tasks, we have measured the execution time of a single analysis job on a single node and then we have increased the number of jobs up to 24 in a single server in order to test both performances and file system reliability. To improve the significance each test has been repeated 1000 times.

| File System | 1 Proc. | 2 Proc. | 4 Proc. | 8 Proc. | 16 Proc. | 24 Proc. |
|-------------|---------|---------|---------|---------|----------|----------|
| GlusterFS   | 115     | 116     | 114     | 135     | 249      | 421      |

Table 2 - The table shows the time spent for the computation of 1, 2, 4 to 24 jobs on the same machine.

In Table 2 we show the average time spent for the computation of 1 up to 24 independent jobs on the same machine. From the reliability point of view the test has been successful, no faults occurred.

As regarding the performances, the results in Table 2 show that GlusterFS starts to affect the computation performance when all the available cores resulted fully exploited. This overhead is due to the file system management. More precisely in Fig. 3 we show the CPU usage by the Gluster Demon during the jobs execution. The graph show an increasing activities of the file system demon in correspondence of the increase on jobs number, however the maximum load not overcome the use of a single core. It suggests to dedicate a core for node in order to limit the impact on performances in terms of jobs/s rate, and then limits the job slots of each server at 7.

With this setup we spend about 12% of the total cores for the storage infrastructure. Notice that this percentage is comparable with the number of cores used in a classical Grid installation with storage element decouple by the computing cluster.

5 HDSF test in BARI

The HADOOP File System (HDFS) is a non-posix compliant file system written in Java, in provide a posix-like interface through a fuse kernel module. The files are distributed and replicated at block level over the node of the HDFS infrastructure, each block is 64MB; there are minimal possibilities to tune this value, e.g. one can increase the size to 128MB. HDFS provides a single Name Node that can serve multiple Data Nodes that compose the file system; it was created to manage large storage area and to support natively the map-reduce paradigm, implemented in the Hadoop middleware.

The test on HADOOP carried on at INFN-Bari is focused to the understanding the capabilities of Hadoop file-system to cope with any kind of hardware and software failures. In Figure 4 we showed the current cluster setup with the redundancy of the name node and the implementation of three rack, that is a storage partition following the HDFS naming.
With several tests that provide the switch off of the HDSF core services, we have verified that each node (both name node and data nodes) could fail with small or not service interruption.

At INFN-Bari, we are modifying the Hadoop replication policy in order to have a broader data distribution among different racks within a farm, and in order to provide the capability of understanding the rack topology.

This added feature could provide the capability of understanding the rack distribution among different farm or in general among different failures domain.

At the end of the work it is foreseen that this policy will allow the file-system to resist to a complete farm failure.

6 Distributed Scenarios

In order to complete the evaluation of the selected file systems, a distributed setup is needed to compare and validate the solution on geographic scale.

Right now we setup a test HDSF volume of 500GB of SuperB simulated data (fig 5). The current configuration provide a Name node in Bari with a replica in Naples, it manage several disk nodes distributed in the same sites. First tests of data access and disaster recovery with the switch off of disk nodes has been performed with success.

To do a full evaluation we plan to create two large volumes by using GlusterFS and HDSF between the sites of Napoli and Bari and test the performance over the GARR-X [10] Network that will provide 10Gbit/s interconnections by the end of 2012. The dark fibers will allow an easy upgrade to higher bandwidths, namely 40 and 100 Gbit/s, when the SuperB will start data taking.
7 Conclusion

First Tests done with GlusterFS and Hadoop File System showed interesting features for the SuperB purposes, in term of functionality, scalability and reliability. Both the file systems are been tested at first in local configuration by evaluating different aspects i.e. network usage, data replication strategy and the reliability to catastrophic events. Moreover the two solutions offer a multi-cluster configuration extensible over WAN with native auto replica between sites. A first geographic implementation of HDSF was implemented between Napoli and Bari. The next step will be setup a GlusterFS volume over WAN and then made several stress tests in order to compare the two technologies over the new 10Gbit/s geographic links.

References
[1] The SuperB Collaboration, SuperB Progress Report, Detector, http://arxiv.org/abs/1007.4241v1
[2] The Cabibbolab Home Page http://www.cabibbolab.it/
[3] The Pon ReCaS Project Home Page http://www.cabibbolab.it/
[4] S. Luitz “SuperB Resource Estimate Spreadsheet from Ferrara R&D WS” http://agenda.infn.it/materialDisplay.py?contribId=32&sessionId=1&materialId=0&confId=3868
[5] The Gluster Home Page http://www.gluster.org/
[6] The HDSF Home Page http://hadoop.apache.org/hdfs/
[7] NFS 4.1 http://www.pnfs.com/
[8] EOS Home Page http://eos.cern.ch/
[9] D. DelPrete, S. Pardi and G. Russo - “Evaluating new cluster setup on 10Gbit/s network to support the SuperB Computing Model” - In IEEE proceeding of CCP2011 June 21-24, 2011 Palinuro (SA) Italy - http://dx.doi.org/10.1109/CCP.2011.33 - ISBN: 978-0-7695-4528-8
[10] Francesco Palmieri, Silvio Pardi: “Towards a federated Metropolitan Area Grid environment: The SCoPE network-aware infrastructure.” Future Generation Comp. Syst. 26(8): 1241-1256 (2010). http://dx.doi.org/10.1016/j.future.2010.02.003 - ISSN: 0167-739X
[11] Catello Di Martino, Marcello Cinque, Domenico Cotroneo. “Assessing Time Coalescence Techniques for the Analysis of Supercomputer Logs.” In the proc. of the 42nd Annual IEEE/IFIP ICDSN (DSN 2012), June 25th-28th 2012, Boston, USA
[12] The GARR website with the GARR-X project http://www.garr.it/b/eng