Journal of Physics: Conference Series **219** (2010) 062042 doi:10.1088/1742-6596/219/6/062042

Commissioning of a StoRM based Data Management System for ATLAS at INFN sites

A. Brunengo[4], C. Ciocca[1], M. Corosu[4], M. Pistolese[5], F. Prelz[5], L. Rinaldi[3], E. Ronchieri[3], V. Sapunenko[3], A. Andreazza[2], S. Barberis[5], G. Carlino[6], A. Cavalli[3], S. Dal Pra[3], L. Dell'Agnello[3], D. Gregori[3], B. Martelli[3], L. Perini[2], A. Prosperini[3], P. Ricci[3], D. Vitlacil[3]

[1] CNAF/INFN and University of Bologna, Italy
[2] INFN and University of Milano, Italy
[3] CNAF/INFN, Italy
[4] INFN Genova, Itay
[5] INFN Milano, Italy
[6] INFN Napoli, Italy

E-mail: claudia.ciocca@bo.infn.it, massimo.pistolese@mi.infn.it

Abstract.

In the framework of WLCG, Tier-1s need to manage large volumes of data ranging in the PB scale. Moreover they need to be able to transfer data, from CERN and with the other centres (both Tier-1s and Tier-2s) with a sustained throughput of the order of hundreds of MB/s over the WAN offering at the same time a fast and reliable access also to the computing farm. In order to cope with these challenging requirements, at INFN Tier-1 we have adopted a storage model based on StoRM/GPFS/TSM for the D1T0 and D1T1 Storage Classes and on CASTOR for the D0T1. In this paper we present the results of the commissioning tests of this system for the ATLAS experiment reproducing the real production case with a full matrix transfer from the Tier-0 and with all the other involved centres. Noticeably also the new approach of direct file access from farm to data is covered showing positive results. GPFS/StoRM has also been sucessfully deployed, configured and commissioned as storage solution for an ATLAS INFN Tier-2, specifically the one of Milano. The results are shown and discussed in this paper together with the ones obtained for the Tier-1.

1. Introduction

The Large Hadron Collider (LHC) [1] will produce about 15 petabytes of data annually which will be distributed around the globe. Thousands of scientists will access and analyse this data thanks to the LHC Computing Grid, a distributed computing and data storage infrastructure [2].

The primary event processing occurs at CERN Tier-0 [3]. The RAW data is archived at CERN and copied to the Tier-1s around the world. These facilities archive the RAW data, provide the reprocessing capacity, provide access to various processed versions and allow scheduled analysis of the processed data by physics analysis groups. Derived datasets produced by the physics groups are copied to the Tier-2s for further analysis. Tier-2 facilities also provide the simulation capacity for the experiment, with the simulated data housed at Tier-1s. In addition, Tier-2 centres will provide analysis facilities.

© 2010 IOP Publishing Ltd

Data flow and main activities with distributed data of the ATLAS experiment are represented in Figure 1.

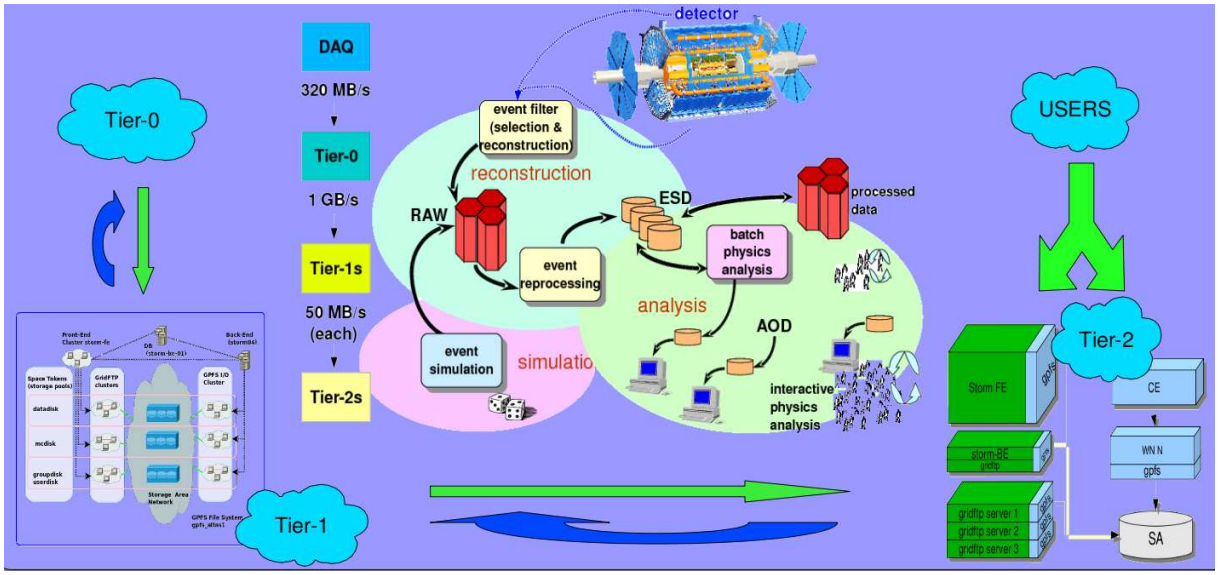


Figure 1. ATLAS data flow and activities in the LHC computing Grid.

High performance disk storage solutions based on parallel file systems are becoming increasingly important to fulfill the large I/O throughput required by High Energy Physics (HEP) applications.

Parallel file systems deployed on Storage Area Network (SAN) are capable to serve multi terabyte storage, offering a scalable, high performance system with reliability in case of disk failures, and with the possibility of great performance on multiple access on files from different nodes.

2. StoRM

StoRM [4] is a Grid Storage Resource Manager (SRM) for disk based storage systems which implements the SRM interface version 2.2. It is designed to support guaranteed space reservation and direct access (native POSIX I/O call) to the storage as well as other standard libraries (like RFIO). StoRM takes advantage from high performance parallel file systems like GPFS (from IBM) and Lustre, but any (distributed) file system with POSIX access is supported (e.g. XFS from SGI or ext3). A modular architecture decouples StoRM logic from the supported file system. It takes advantage of ACL support provided by the underlying file system to implement the security model.

StoRM has a multilayer architecture. The front end component exposes a web service interface where the user requests land to be processed an authorized. The back end is the main component: provides space tokens management, user redirection to the proper URL, and executes all SRM synchronous requests (e.g. creation of directories). A database is used to store SRM request data and the internal StoRM metadata. To remark that losing the database content only affects the ongoing operations.

While the front end component may be replicated with as many instances as needed, in the current version the back end is unique: this limitation (i.e. this potential single point of failure)

will be eliminated in a next version.

StoRM being a light diskspace and authorization manager does not require special hardware. Sites can change storage system without care about the SRM layer. It is simple, configurable and highly scalable, starting from a single machine up to a free scalable architecture satisfying a Tier-1 centre. It is efficient, provides high performance on SRM requests execution, and secure, with a layered security mechanism, VOMS based and highly configurable.

3. StoRM and GPFS

A cluster file system allows large numbers of disks attached to multiple storage servers to be configured as a single file system, providing transparent parallel access to storage devices while maintaining standard UNIX file system semantics, high speed file access to applications executing on multiple nodes of a cluster and high avalaibility.

StoRM takes advantage from aggregation functionalities provided by dedicated systems, such as parallel and cluster file systems. Such a file system allows to achieve complete redundancy without single point of failure increasing reliability and dynamic management of volumes (dynamic resize of file system, data migration between disks), all online, with a significant improvement of management flexibility.

StoRM and GPFS (General Parallel File System) [5] can represent a complete solution for space management, providing quick access to files, being each one of its elements just a GPFS client in the same cluster.

Data access is performed through Network Shared Disks (NSD) and GridFTP servers, respectively for LAN (Local Area Network) and WAN (Wide Area Network) access.

It is possibile to have as many GridFTP servers as needed to provide the required transfer throughput. Moreover the GridFTP servers can be partitioned per space token (i.e. per logical subset of the storage), so that the real traffic load can always be turned on different machines.

This system leverages the capabilities of a redundant, high throughput network and a highly available file system (such as GPFS) to sustain high data flows.

The GPFS file system allows direct access from the clients using the file protocol avoiding the need of any external protocol, such as *RFIO* or *Xrootd*.

All hosts linking the file systems belong to a common, global controlled cluster. From a single manager is possible to check, add and remove nodes. It is also possible to add or remove dinamically disks to the file system while clients are still operating.

Each NSD nodes can serve all the shared disks, as long as they are physically connected: this allows to increase the bandwidth simply adding more servers. The maximum flexibility and scalability can be obtained using a Fibre Channel (FC) infrastructure to interconnect disk systems with NSD and GridFTP servers.

A fundamental feature offered by GPSF is the redundancy of the system: the unavailability of one (or even several) server only decreases the performance of the overall system.

A disadvantage of GPFS is the cache amount limited by the operating system. For frequent access to a large amount of files this can slow very much the I/O operations. Directories suffering for this issue (e.g. shared software areas) can be exported using the Cluster Network File System (CNFS), an highly scalable and clustered version of NFS leverages on GPFS.

4. StoRM at CNAF, the INFN Tier-1

4.1. Setup

At CNAF, the INFN Tier-1, the file system for ATLAS consists of three different storage pools, created over three EMC CX3-80 storage subsystems, served by 4 GridFTP and 6 GPFS server each. Every pool is composed by 24 LUNs of 8 TB each. The storage pools are dedicated to different space tokens. Every server (GridFTP and GPFS I/O) is connected by Fibre Channel at 4 Gb/s to the SAN and 1 Gb/s to the LAN. StoRM architecture is composed by four front end servers, scalable on the load expected, a dedicated database and a back end servers. The layout of the system is depicted in Figure 2.

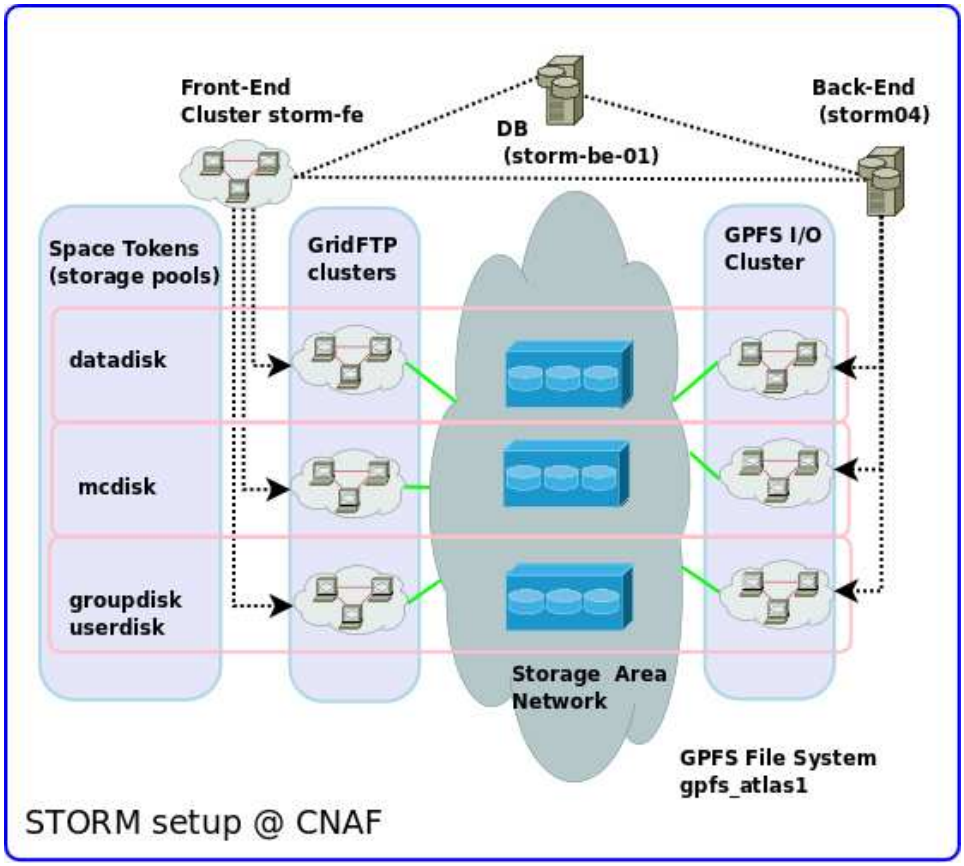


Figure 2. StoRM and GPFS setup at CNAF, the INFN Tier-1.

4.2. System performance

StoRM system at CNAF has been largely tested, largely used in production during ATLAS daily activities and stressed during challenges. Main features and results are:

- data transfer from external SRM endpoints to CNAF StoRM endpoint: throughput of 180 MB/s sustained smoothly for 14 hours;
- throughput test - maximum throughput reached with 15 parallel streams per transfer, 120 concurrent data transfers, file size of 100 MB: sustained rate of 350 MB/s for 8 hours with peak of 370 MB/s;
- file deletion: deletion rate of 0.8 TB every 60 seconds;
- concurrent GPFS operations: read/write throughput 750 MB/s, Figure 3 (left);

- GridFTP full operating conditions, max bandwidth on LAN 4 Gb/s: read/write throughput 500 MB/s, Figure 3 (right);
- test of backlog recovery obtained during CCRC 2008 phase II by ATLAS: 12 hours of data recovered in 2 hours, 200 MB/s sustained, Figure 4;
- ATLAS 10M file test on January 2009: 30 Hz sustained with peak of 40 Hz, Figure 5.

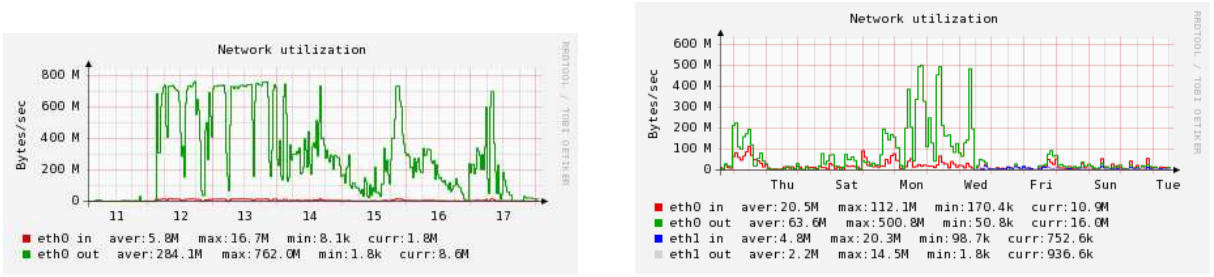
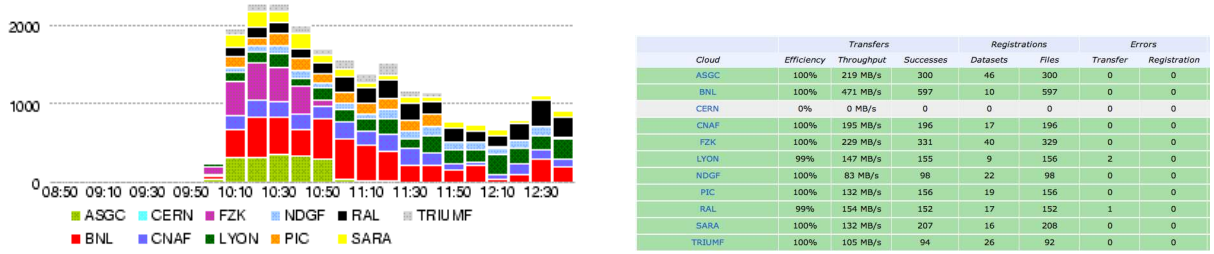


Figure 3. GPFS (left) and GridFTP (right) network utilization during ATLAS activities at INFN Tier-1.



	Transfers			Registrations		Errors	
Cloud	Efficiency	Throughput	Successes	Datasets	Files	Transfer	Registration
ASGC	100%	219 MB/s	300	46	300	0	0
BNL	100%	471 MB/s	597	10	597	0	0
CERN	0%	0 MB/s	0	0	0	0	0
CNAF	100%	195 MB/s	196	17	196	0	0
FZK	100%	229 MB/s	331	40	329	0	0
LYON	99%	147 MB/s	155	9	156	2	0
NDGF	100%	83 MB/s	98	22	98	0	0
PIC	100%	132 MB/s	156	19	156	0	0
RAL	99%	154 MB/s	152	17	152	1	0
SARA	100%	132 MB/s	207	16	208	0	0
TRIUMF	100%	105 MB/s	94	26	92	0	0

Figure 4. Throughput sustained by Tier-1 sites during stress test of backlog recovery during CCRC 2008.

These values largely exceed the required metrics for CNAF.

5. StoRM at Milano, an INFN Tier-2

Besides the above mentioned duties of a Tier-2, in Milano we have also to support local user activities on the farm. The chosen storage solution (GPFS and StoRM) offers an extensible, high available system.

5.1. Setup

In Figure 6 the current production configuration is depicted. We configured eight servers connected to the LAN at 2 Gb/s (bonding) and to three storage systems via Fibre Channel at 4 Gb/s. Currently no FC switch is present but this configuration will allow such future improvement. The disk space has been divided in three file systems, to support user directories, production space tokens, Grid software area.

To ensure a shared balanced software files access, all the servers have been promoted to CNFS servers, picking the files requested from an equal number of NSD.

A user interface pool has been created to give a transparent high available service: all these hosts share the same setup and in particular are configured as GPFS clients in a separated

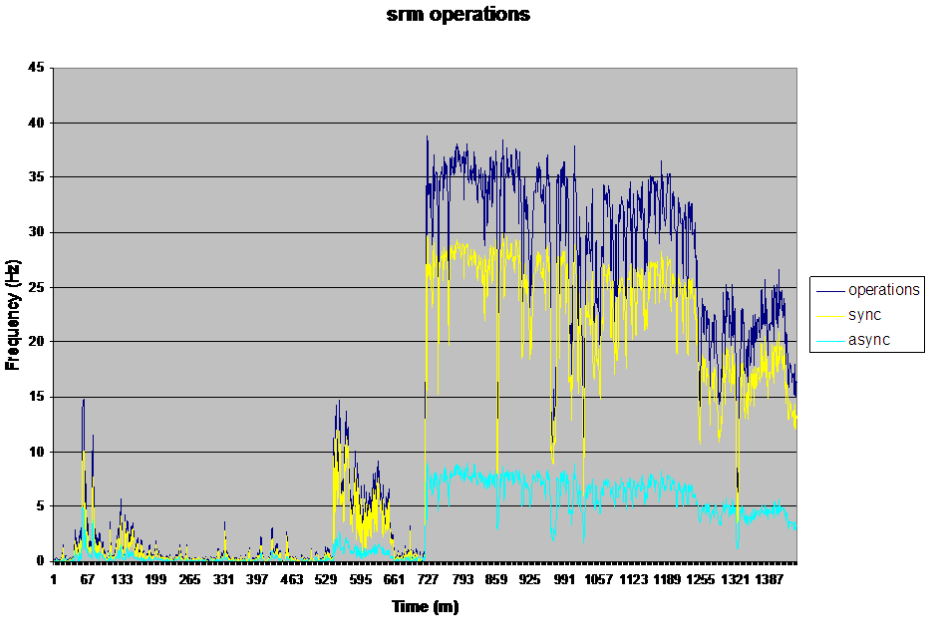


Figure 5. Frequency of SRM request on StoRM sustained during the 10M files stress test.

cluster. Local users can also submit their jobs via standard Grid tools to the batch system, if more computing power is needed. Moreover, users home directories are directly accessible via GPFS from the cluster of worker nodes, allowing the users to get job results without involving the complete Grid chain.

Our StoRM instance is composed by one front end and one back end. Four GridFTP servers cover the transfer requests. Since StoRM has been setup as a clustered installation, further machines could be easily added if needed.

5.2. System performance

Performance tests have been done with this configuration on GPFS and StoRM. Writing operations are limited to 800 MB/s on a storage system and to 600 MB/s for the others, due to the FC connections. First of all, we have measured the available network bandwidth, using netperf, to the disk servers obtaining a peak value of 412 MB/s, due to a not fully understood problem on some network interfaces (not able to achieve 1 Gb/s). Then we obtained the following results testing the GPFS system:

- reading from 8 head nodes 780 MB/s
- reading from 36 clients 680 MB/s
- writing from 8 head nodes 320 MB/s
- writing from 36 clients 150 MB/s
- writing from 16 clients 320 MB/s
- writing from 8 client 340 MB/s

We also verified the basic functionalities of StoRM (e.g. file access, put and get requests).

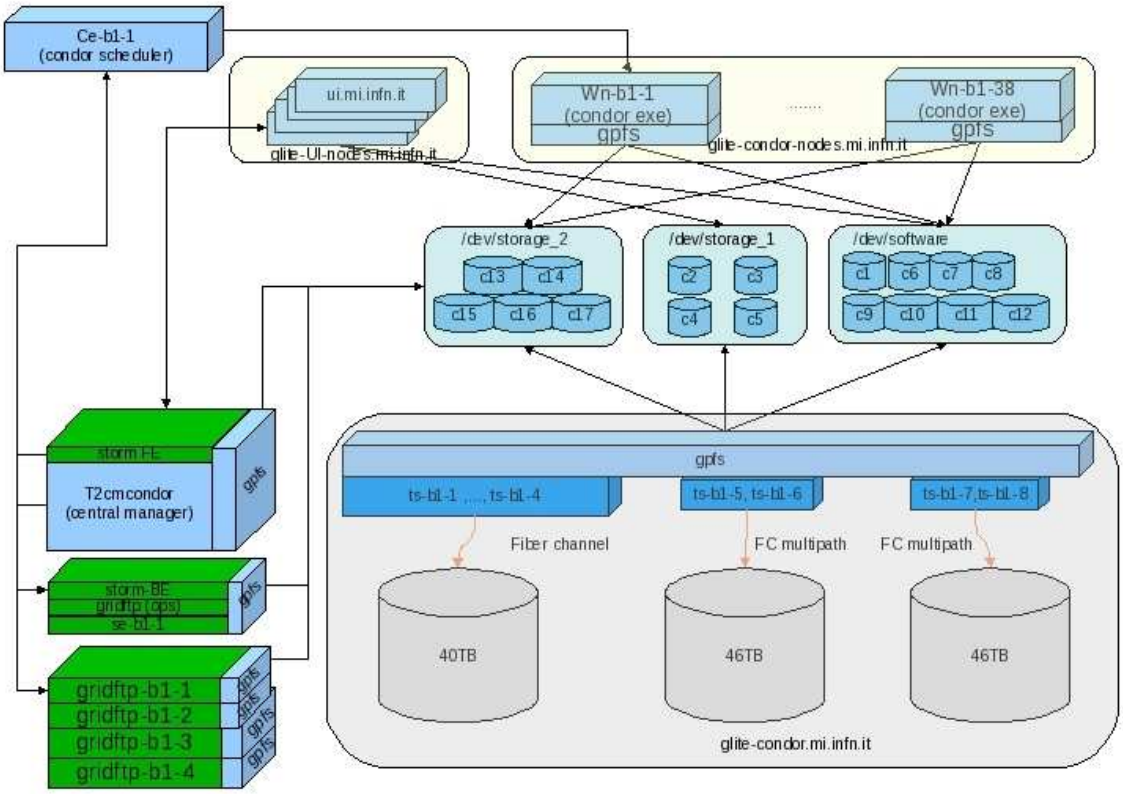


Figure 6. StoRM setup at Milano, INFN Tier-2.

Finally, we performed transfer tests from CNAF to Milano Tier-2, as in real operational scenario, being the maximum WAN bandwidth limited to 1 Gb/s. We obtained the following results with GridFTP in full operating conditions transferring 1 GB files:

- 1 server write 90 MB/s
- 3 servers simultaneous write 40 MB/s each hence saturating the WAN link

The results show that the system is able to fulfill the requested performance.

6. Future developements

In Milano site we are now in the process of acquiring two other storage systems, expanding the total space up to 150 TB with four additional NSD servers. These servers will be configured in the same GPFS cluster. Some NSD volumes will be then added to the previous file system. A further future improvement would be to install a FC switch to consolidate the FC infrastructure, enhancing the whole system bandwidth and balancing.

In general, two possible architectures of storage connection for a Tier-2 site are possible. They are represented in Figures 7.

In the first one it can be used a fully redundant connection from each disk server, equipped with dual head HBA (Host Bus Adapter), to each disk controller through two FC switches or some kind of redundant FC Fabric. That way every server can access every volume, allowing

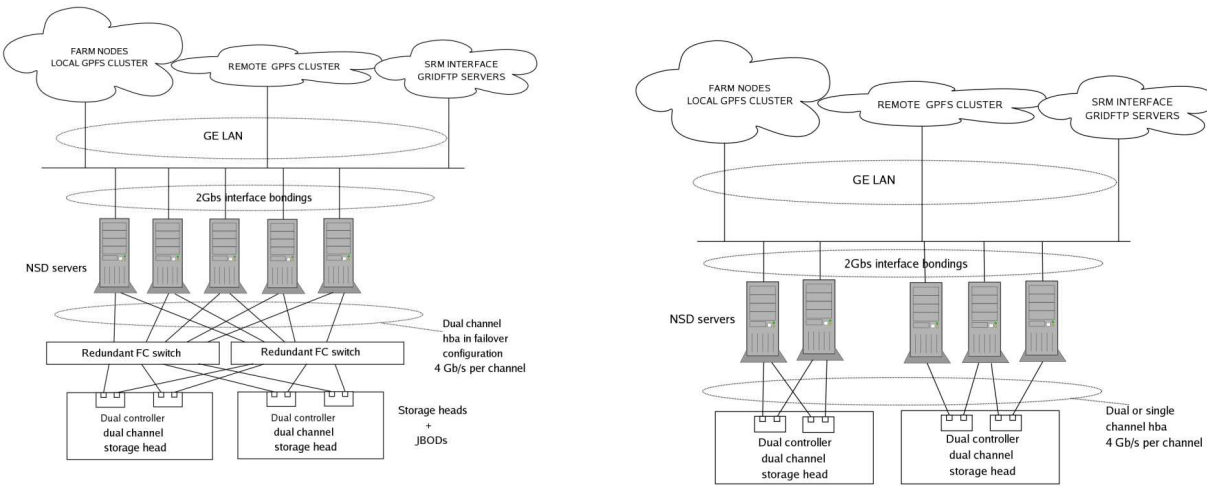


Figure 7. Two possible architectures of storage connection for a Tier-2 site

flexible management in assigning or reassigning NSD server associations, even without loss of service, in order to perform management activities as long as load rebalancing across the NSD servers. You can also perform data migration between different volumes (NSDs) inside the SAN without loading the production ethernet network. In addition you can benefit of GPFS features like the usage of multiple NSD server in failover configuration.

In the second one, every dual RAID controller is directly connected to two or three disk server, depending on the throughput required towards the storage behind a single RAID controller, realizing multiple small SAN islands. You can still take advantages of features described in the previous architecture, but only between volumes and servers that are directly connected. The reduced flexibility of this configuration is balanced by a lower cost for hardware (no FC switch is needed) and by a simplified management (no switch/zoning/filter configuration and maintenance).

In both architectures all disk servers are configured as NSD servers, connected to the LAN through 2 GE in bonding configuration (eventually through 10 GE connection) to maximize the throughput towards the storage. SRM and GridFTP servers acts as GPFS clients, eventually using different GE interfaces towards the NSD servers and the WAN. Farm nodes can be configured in the same GPFS cluster as the NSD servers, or in a different dedicated GPFS cluster. In the latter case, GPFS file system can be accessed through a remote cluster mount without loosing in performance.

6.1. Acknowledgments

Authors wishing to acknowledge assistance from colleagues of CNAF/INFN and of ATLAS collaboration involved in distributed computing operations, StoRM developers and special work by storage staff at CNAF and Milano sites.

References

[1] http://lhc.web.cern.ch/lhc/LHC-DesignReport.html
[2] I. Bird et al., "LHC computing Grid. Technical design report", CERN-LHCC-2005-024, Jun 2005;
[3] ATLAS Coll. (G. Duckeck et al.) "ATLAS computing: Technical design report", ATLAS-TRD-017, CERN-LHCC-2005-022, Jun 2005;
[4] http://storm.forge.cnaf.infn.it/
[5] http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.gpfs.doc/gpfsbooks.html