ISTITUTO NAZIONALE DI FISICA NUCLEARE

Sezione di Pisa

INFN/CCR-09/03
July 6, 2009

CCR-26/2008/P

INFN-PISA NETWORK AND STORAGE SCENARIO FOR LHC TIER 2 AND GRID DATA CENTER

Silvia Arezzini[1], Alberto Ciampa[1], Tommaso Boccali[1], Enrico Mazzoni[1]

[1)]*INFN-Sezione di Pisa, Largo Bruno Pontecorvo, 3, I-56127 Pisa, Italy*

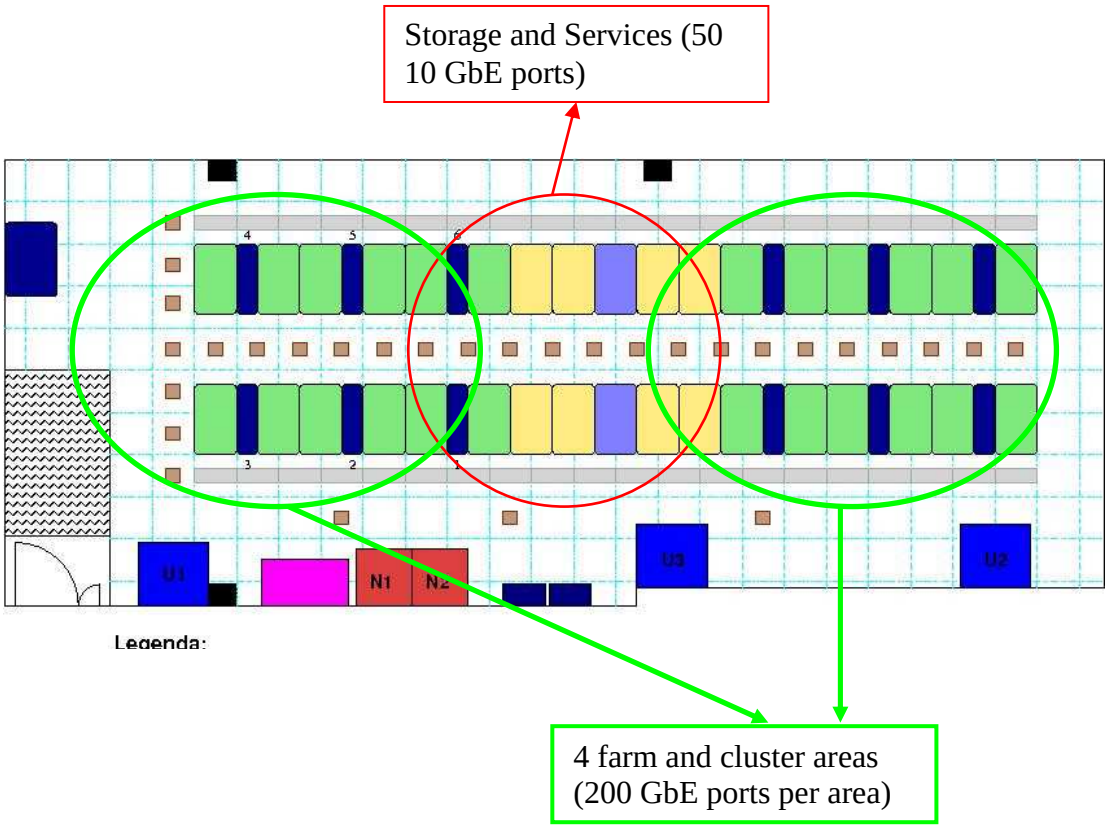
Abstract

Based on the specifications provided by the CMS Experiment, the Network and Storage requirements for a Tier 2 site (T2) are analyzed. A proposal for a possible solution is presented and discussed, targeted to the Pisa CMS T2 site. In the analysis, the current situation of the Pisa Grid Data Center is taken as a starting point, and the future growing path is taken into account for the next couple of years. Two different final architectures are presented.

*Published by **SIS–Pubblicazioni***
Laboratori Nazionali di Frascati

1 PHYSICAL LAYOUT

Final configuration:



2 GENERAL SPECIFICATIONS

The data center will be mainly dedicated to operate as one of the national T2s of the CMS High Energy Physics experiment, under development at CERN in the framework of the LHC programme. We do not expect that the CMS community will load the full system continuously and, moreover, we intend to take the realization of the CMS-T2 system as a chance to set up a more general GRID-oriented data center for the entire scientific community referring to INFN-Pisa and in the national context.

CMS T2 general specifications:

- Global envinronment: GRID (LCG: LHC Computing Grid)
- Worker nodes and servers run under Linux
- Farm usage: no interactive, only jobs submitted through a queue system (LSF)
- From the storage point of view there are two kinds of job (all files order of 2-10 GB):
  - Read from 1 file and write 1 file (montecarlo production mode). We expect this case being CPU bound.
  - Read from order of 10-100 files and write on 1 file (analysis mode). This mode we expect to be I/O bound.
  - For Montecarlo mode 1 MB/sec of throughput per job will be enough
  - For analysis mode 5 MB/sec or more, per job, will be useful

- During the processing activity, the upload/download of big quantities of data (order of 200 TB) to/from outside (the appointed T1 center) will be performed periodically (we consider at least twice a year, probably more until the CMS calibrations are well understood)
- The storage will be accessed through two levels of server:
    - A level of disk storage, directly connected to the storage devices (for instance GPFS servers)
    - A level of SRM (storage resource manager, running proprietary software: dCache, DPM, StoRM or other) servers, connected to the worker node farms

3 FINAL DIMENSION

- Worker nodes: 2000 job slots (i.e. 2000 CPU cores, max 2000 concurrent jobs) is the minimum required, it is foreseen to scale up to 5000 cores
- Storage space: up to 1 PB
- Expected number of SRM storage: 10 to 30
- Expected number of disk servers: as many as necessary (depending on the FS architecture adopted)

4 REQUESTED AGGREGATE BANDWIDTHS

The goal of this paragraph is to provide the two typical aggregate bandwidths of the architecture:

- Worker-nodes ↔ storage and storage ↔ external

To do that we refer to the two main kinds of job we expect: Montecarlo (1 MB/s per job) and analysis (5MB/s per job or more). Taking into account a total amount of 2000 concurrent jobs of the two above mentioned types we obtain:

- Nodes ↔ Storage: 2 – 10 GB/sec (minimum) aggregate throughput

As far as the Storage ↔ external throughput is concerned, we have the following elements:

- CMS will require to transfer order of 200 TB at least twice a year, this leads to a “net mean” throughput of at least 0.12 Gb/sec
- We evaluate the transfer requirements of the other major users (notably the LHC ATLAS experiment T3) to be order of 20-30% with respect to the CMS one

These two elements sum up to 0.15 Gb/sec. Following the usual procedure of dimensioning a link bandwidth starting from a “net mean” throughput (the link is 10 times the mean throughput), we obtain:

- Storage ↔ External: 1.5 Gb/sec

In the final scheme we connect the “big Ethernet switch” with the Router using a 10 GbE link; this will have a negligible economic impact on the chosen architecture. The bandwidth of the geographical link is out of the scope of this document. About the external link we will have, over time, the bandwidth provided by the appointed Italian organization

(GARR).

We expect that more than 75% of the GRID farm will be used with jobs of the two above mentioned types for a considerable amount of time. This means that, when the growing of quad-core processors and 1U 4 processors server will allow to get close to the 5000 cores, we could expect at maximum to double the aggregate throughput requests to maintain an adequate CPU efficiency.

5 ARCHITECTURE CONSTRAINTS

- The portion including worker nodes and SRM servers has a fixed architecture decided by the LCG organization (specifically in our case the national part of it: INFN-GRID) and by the experiment (as far as the SRM is concerned)
- Regarding the portion including the disk servers and the storage devices (including the FS choice) we are free to find the most suitable architecture.

6 FURTHER CONSIDERATION

6.1 Growing path.

The growing path would be driven by the growing profile of the main user needs (there is a CMS time schedule that includes two global test campaigns per year before the start of the production); this would lead to an equilibrated growing of the center, developing and deploying concurrently adequate systems for the farms (worker nodes and GRID services), the storage (disc space) and the network (number of ports and requested throughput, both LAN and WAN). However in our case this is not completely true. Today we have already about 1200 CPU cores dedicated to worker nodes and we expect to reach 2000 cores during 2009; so the profile of our growing path is shaped on the dimension of the storage system available (a growing portion of the final amount necessary with respect to the number of CPU) and the consequent network required. This will drive the join projects: storage + network.

Our "major user" (LHC CMS experiment) will require the full configuration in 2009/2010, ready to start the production, although this will not be a 0/1 change of state; there will be a ramp to reach the full capacity spanning a couple of years. We spent the year 2007 to properly set up the farms and the GRID infrastructure (roughly 50% of the minimum final dimension), equipped with about 50TB of storage (not in the architecture herein described) and using the existing network expanding it; by the end of 2008 we have to set up and test under stress the complete system architecture with the near full storage configuration (however >50% with respect to the final) and a network able to provide the required data transfer aggregate throughput.

6.2 Software environment (as far as we know today).

Fixed and known:

- Linux: Scientific Linux CERN (derived from Red Hat).

- Kernel version now running is 2.6.9-67 64bit; the experiment will follow the kernel evolution with 2 years of delay.
- SRM environment (running on the SRM servers) will be decided by the experiment (now using dCache, may be in future StoRM). The numbers of the SRM servers is expected to reach 30 in the final configuration. Now we are running 13 dCache servers and they will grow accordingly to the number of jobs in production and the required throughput, provided an adequate storage system.

what we are free to decide:

- We think that a SAN architecture is better suited than a NAS one for our scenario, but this decision is up to us. We can also decide FS architecture and therefore the configuration and number of disk servers. In general we refer to GPFS as a distributed file system by which the whole storage disk space may be accessible by each disk server.
- The idea is that the SRM server layer takes care of balancing the communication among the different worker nodes to the SRM server links (storage job requests). On the other side the disk servers layer will have to balance the traffic on the different physical storage links (serving the storage job requests).

7 ARCHITECTURE DRIVERS

Two main drivers have been identified to lead the architecture definition:

- Manageability and robustness: limited amount of human intervention required, limitation imposed by the very few human resources that are and will be available.
- Self-balancing capacities: we are not able to provide a scheme or a model of the users' behaviour in terms of job distribution on the farm and file allocation on the storage devices. Hence we will have no chance to study and deploy efficient optimization strategies to balance the I/O (worker nodes ↔ storage) on the LAN.

These two drivers, along with the considerations exposed in the previous paragraph, lead to the following scheme, represented in the final two figures:

(cloud of worker nodes)
↕
(big ethernet switch)
↕
(SRM servers)
↕
(direct coupling / 10GbE switch)
↕
(disc servers)
↕
(big FC or IB switch)
↕
(storage devices)

Two solutions have been identified for the 1GbE / 10 GbE “big” switch and are both under a deeper investigation, they are based on the Foundry “Big Iron” (RX 32) and Force10 “E series” (E1200 chassis) architectures (Cisco and Extreme solutions have been examined too, but they didn’t match our specifications). After several months of technical and economical analysis, the Force 10 E1200 solution has been chosen and its procurement is under way.

As far as the storage system is concerned, the two following figures depict the general architecture, in two different scenarios:

1. the “storage devices switch” is inside the storage system: the storage has the required switch capabilities built in (for instance: EMC2 Symmetrix or Data Direct HPC configuration series).
2. the “storage devices switch” is external to the storage system and we have to set it up with a proper dimensioning and configuration (any SAN architecture): we will not need to require an up to 1PB scaling because in this case the solution will deploy several identical systems joined together by the GPFS structure. This approach is driven by the aggregate throughput requests.

The first solution is more expensive than the second one. The extra cost is due to the performance guaranteed by the storage system itself, while in the second solution our aim and duty is to catch up the required performance.

Mainly due to budget limitations, the CMS experiment decided to follow the second direction that they are implementing using a series of separate storage appliances (not using a SAN architecture).

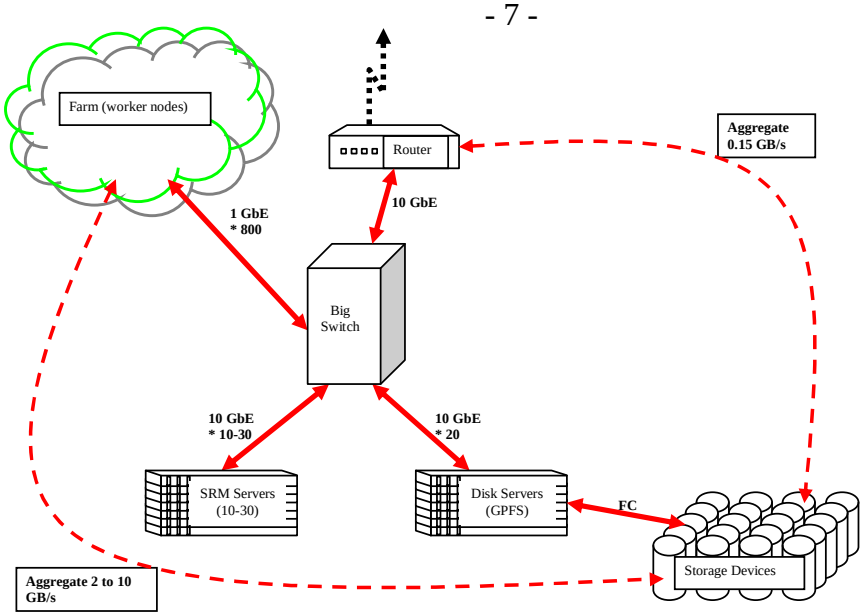


Fig.1 - Architecture based on storage system including internal switching capabilities (EMC2 Symmetrix or Data Direct HPC configuration).

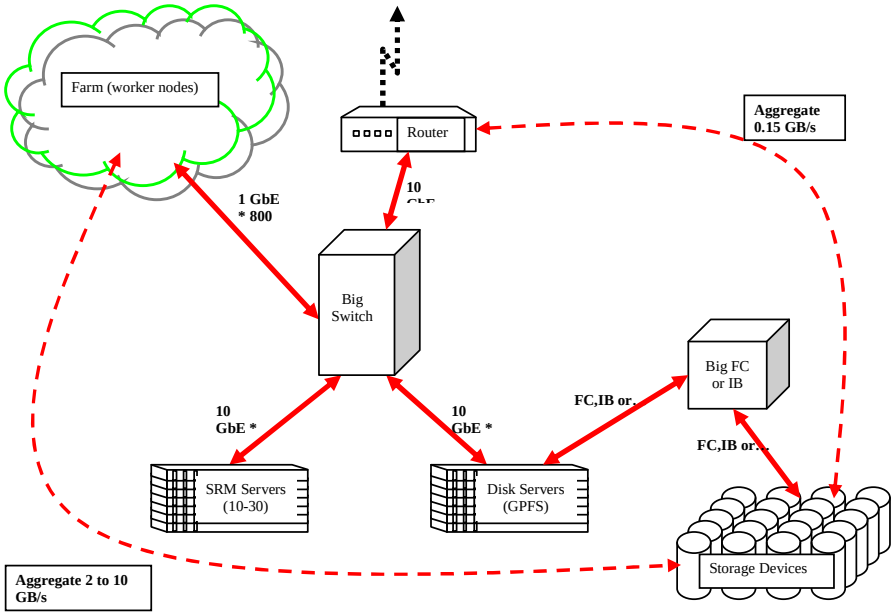


Fig. 2 - Architecture using standard SAN storage system.