ITALIAN TIERS HYBRID INFRASTRUCTURES FOR LARGE SCALE CMS DATA HANDLING AND CHALLENGE OPERATIONS

D. Bonacorsi[#], INFN-CNAF, Bologna, Italy

(on behalf of: INFN-CNAF Tier-1 staff, INFN Tiers community and the CMS experiment)

Abstract

The CMS experiment is travelling its path towards the real LHC data handling by building and testing its Computing Model through daily experience on production-quality operations as well as in challenges of increasing complexity. The capability to simultaneously address both these complex tasks on a regional basis, e.g. within INFN, relies on the quality of the developed tools and related know-how, and on their capability to manage switches between testbed-like and production-like infrastructures, to profit from the configuration flexibility of a unique robust data replication system, to adapt to evolving scenarios in distributed data access and analysis. The work done in INFN in the operations of Tier-1 and Tier-2 centres within event simulation, data distribution and data analysis activities, in daily production-like activities as well as within LCG Service Challenges, are here presented and discussed.

INTRODUCTION

The CMS computing system relies on a distributed infrastructure of Grid resources, services and toolkits, whose building blocks are provided by the Worldwide LHC Computing Grid (WLCG). CMS builds application layers able to interface with several different Grid flavours (LCG-2, Grid-3, EGEE, NorduGrid, OSG). A WLCG-enabled hierarchy of computing tiers is depicted in the CMS computing model [1], and their role, required functionality and responsibilities are specified.

ROLE AND FUNCTIONS OF CMS TIERS

The functions of a Tier-1 for CMS are: *i)* scheduled data-processing operations; *ii)* data archiving (custody of raw+reco and subsequently produced data); *iii)* disk storage management (fast cache to MSS, buffer for data transfer); *iv)* data distribution (data import/export from/to any CMS Tier-0/1/2/n; *v)* analysis support (proficient data access via CMS+WLCG services). The nominal resources for an average CMS Tier-1 in 2008 are: an incoming/outgoing transfer capacity of 7.2/3.5 Gb/s respectively; a computing power of 2.5 M-SI2k (with scheduled reprocessing requiring twice the CPUs needed by analysis tasks); a disk capacity of 1.2 PB (~85% for analysis data serving); a MSS providing a 2.8 PB capacity.

The functions of a Tier-2 for CMS are: i) fast and detailed Monte Carlo event production; ii) data processing for physics analysis (including late stage analysis requiring very fast data access); iii) data processing for calibration and alignment tasks and for detector studies. The nominal resources for an average CMS Tier-2 in 2008 are: a WAN access at 1 Gb/s (at least); a computing power of 900 k-SI2k; a disk capacity of 200 TB.

The INFN Tier-1

The Italian Tier-1 is a multi-experiment computing centre located at CNAF, Bologna. It offers computing facilities for the INFN HEP community, through dynamic share of access to resources to all involved experiments.

Currently, the overall Tier-1 computing power of ~2500 CPUs is available to all experiments, with a fair-share on a monthly time-window. The LCG 2.7.0 middleware is deployed (Quattor 1.1.0 [2] is used to manage nodes) and the LSF 6.1 scheduler [3] is used (with fully-certified LCG interfacing). Already now, >97% of all CMS jobs exploit Tier-1 resources not locally but through the official Grid layer (see Figure 1).

The set-up of storage resources at CNAF is driven by requirements of LHC data processing at a Tier-1, i.e. simultaneous access of ~PBs of data from ~1000 nodes at high rate. The focus is on robust, load-balanced, redundant solutions to grant proficient and stable data access to distributed users. The disk storage capacity is divided into 4 NAS systems (60 TB), and 3 SAN systems (375 TB). The disk access patterns are derived from specific experiments use-cases. Currently CMS uses disk space to host and serve simulated data to the CMS analysis community, and uses tapes as a reliable, affordable means for storing large volumes of data, hence addressing the Tier-1 data custodial responsibility as required in the CMS computing and analysis model. At CNAF a Castor-1 HSM system operates a Stk L180 tape library (18 TB) and a Stk 5500 tape library (240 TB on six IBM LTO-2 drives and 136 TB on two 9940b drives).

The PhEDEx [4,5,6] system offers a large-scale, reliable and scalable dataset replication, and it is currently in use by CMS as official data distribution management system. PhEDEx addresses HEP use-cases via a push-pull negotiation: it grants replication, data safety, tape migration/stage via a layered architecture and agnosticism on the underlying transfer mechanisms, and ensuring reliability and replication robustness. The Transfer Management Database is used as blackboard (with Oracle backend) for software agents' communication. PhEDEx is in production use for over a year for CMS. The Storage Resource Management (SRM) provides a standard interface to storage systems.

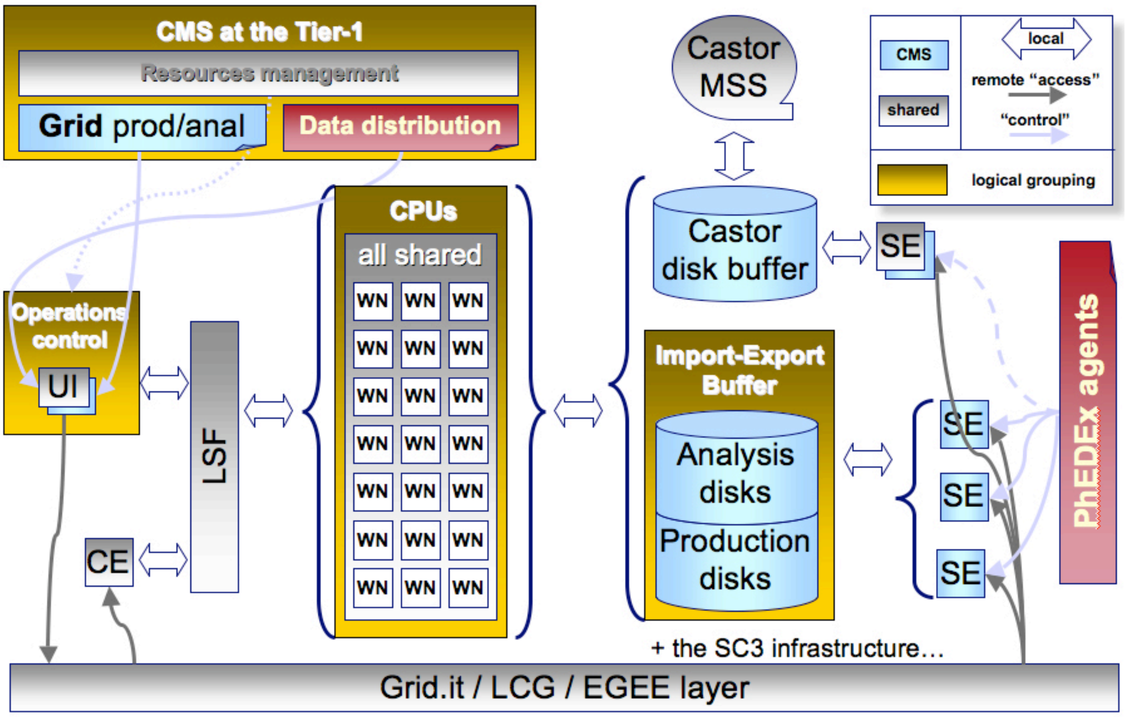

Figure 1 – Set-up of CMS resources and management at INFN-CNAF Tier-1.

The INFN Tier-1 plays an active role in the PhEDEx topology, and demonstrated to be able to import ~1 TB/day from the Tier-0, as well as to sustain e.g. multi-Tier simultaneous inbound/outbound traffic to/from CNAF and bidirectional T1-T1 data movement.

PhEDEx is deployed and used in INFN at CNAF Tier-1 and in 4 Tier-2 sites. The focus in PhEDEx operations at INFN is now towards a robust and stable 24/7 level of service.

INFN Tier-2 sites for CMS

There are proposals for four INFN centres to act as Tier-2 centers for the CMS experiment: Bari, Legnaro-Padova, Pisa, Rome. Some details on each site are given below.

The Bari computing centre is a prototypal Tier-2 centre for the CMS and ALICE experiments. Currently the Bari centre hosts a CPU capacity of ~20 k-SI2k, a disk capacity of ~6 TB and a 100 Mb/s network connection to GARR. Bari mainly contributes to CMS computing activities by developing and testing software tools addressed to data analysis in a distributed environment, like the CMS validation tool. Bari operates PhEDEx using LFC as LCG POOL file catalogue, and performs tests on DPM, FTS and storage technologies. The architectural choice in Bari is driven by the requirement of disentangling production activities from testbed-like activities: site resources are organized and operated in two independent farm environments, a production one and a test one, the latter being equipped with a gLite CE that shares the worker nodes with the production farm. The Bari computing centre is recently experiencing a very high farm uptime (>95% in 2005). The CMS staff on computing is less than 2 FTE. Bari also played an active role as a Tier-2 centre for CMS in LCG SC3.

The Legnaro-Padova computing centre is a prototypal Tier-2 centre for the CMS experiment. Currently the Legnaro-Padova centre hosts a CPU capacity of >200 k-SI2k, a disk capacity of ~20 TB and a Gbit network connection to GARR. Legnaro-Padova is an important site in INFN-Grid and EGEE/LCG [7,8]. The staff consists of 1.5 FTE on system administration, 0.5 FTE on Grid-interface and operations and 0.5 FTE on CMS activities, and all resources are set-up as a unique farm. Despite the site lacks people on installation and maintainance of CMS software and services, Legnaro-Padova is able to grant stable Grid access to the site resources, and recently had an active involvement as a Tier-2 centre in the LCG SC3 throughput phase for CMS, operating PhEDEx-driven transfers.

The Pisa computing centre is prototypal Tier-2 centre for the CMS experiment. Currently the Pisa centre hosts a CPU capacity of >100 k-SI2k, a disk capacity of ~10 TB and a Gbit network connection to GARR. Pisa contributes to Monte Carlo production on LCG for CMS, and runs PhEDEx for data harvesting, injection and movement. A staff of ~1 FTE for CMS operates Pisa resources on a modular farm with different Computing Elements, each managing his own pool of worker nodes. Pisa did not participated to SC3 for CMS, and its experience in driving crucial services for CMS is still growing.

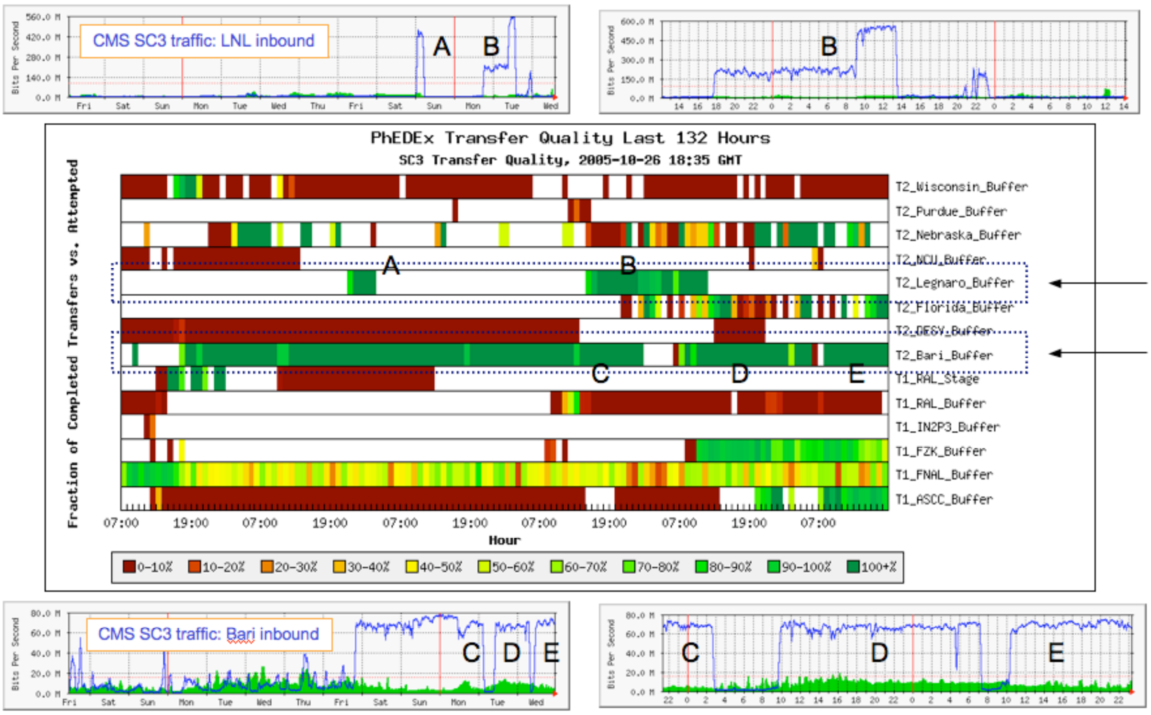

Figure 2 – Monitoring of throughput rates and transfer quality for PhEDEx-driven CMS transfers of data from INFN Tier-1 to INFN Tier-2's in LCG SC3.

The Rome computing centre is a prototypal Tier-2 centre for the CMS and ATLAS experiments. Currently the Rome centre hosts a CPU capacity of ~25 k-SI2k, a disk capacity of ~18 TB and a Gbit network connection to GARR. Rome is an important site in all steps of Monte Carlo event production for CMS. The staff consists of 1.5 FTE on system administration, ~1.5 on Grid support and ~1 FTE on CMS, and all resources are set-up as a unique farm. Despite it runs CMS software and services, Rome did not participated in SC3 for CMS and still needs to ramp-up through wider involvement in stable running crucial services for CMS.

INFN AND LCG/EGEE

The INFN Tiers are being operated in a WLCG-enabled world. The EGEE project offers an integration of current national/regional/thematic Grid efforts, in a seamless Grid infrastructure to support scientific research. The EGEE production Grid consists of ~200 sites in more than 40 countries, with a total computing capacity of 2x10^4 CPUs and 5 PB of storage space, managed through Regional Operation Centres (ROC), with GGUS as user support infrastructure. The operations of EU Grid infrastructure are guaranteed by EGEE SA1. INFN has a strong presence in EGEE/LCG: among the ~40 sites in INFN-Grid, 27 are already in the EGEE/LCG infrastructure (i.e. registered in GOCDB). INFN also has a strong participation to EGEE SA1, with the management of global Grid services via the two-levels ROC/CIC system, taking care o middleware releases and docs, control and certification of sites, user support and support to experiments for Grid integration.

INFN PARTICIPATION TO LCG SC3 FOR THE CMS EXPERIMENT

The general description of the CMS experience in LCG SC3 is given in [9]. INFN participated to SC3 for CMS with the involvement of CNAF Tier-1 and both Bari and Legnaro-Padova Tier-2 centres (see Figure 2).

The INFN-CNAF Tier-1 operated SC3 on a computing infrastructure which was mostly separate from the production infrastructure. Data movement was addressed via PhEDEx against a SRM/CASTOR1 system. In the throughput phase, the ability to switch fast to both *i)* different underlying transfer-mechanisms and *ii)* different infrastructure was put under challenge. In a pre-'service' phase, the Tier-1 collaborated with the LCG SC community to debug SRM, CASTOR2 at CERN and LHCOPN-related issues which prevented effective transfers to be operated. Then, in the 'service' phase CMS stably used the official SC3 infrastructure at CNAF. All SC transfers for CMS were triggered by PhEDEx, with a SC3 instance coexisting with the production one (but independent, so to avoid any possible destructive interference between usual daily production-like operations and time-limited challenge activities). A full use of the PhEDEx system was addressed at the Tier-1 for SC3, with the Castor MSS part deployed as well; MySQL was used as LCG POOL local file catalogue, and the existence of transferred data was published in the SC3 instance of the CMS PubDB system [10] (independent from the production instance, as for PhEDEx). About 10 TB of data were transferred to the Tier-1: all data were published and subsequently accessed by the ad-hoc SC3 job submission robot, based on the CMS CRAB tool [11,12]. During the SC3 exercise, ~15k jobs were submitted against data published at the INFN Tier-1.

The Bari computing centre participated to SC3 by running PhEDEx transfers against a SRM/dCache set-up with 1 admin node and 3 pool node. As for the Tier-1, the SC3 instances of both PhEDEx and PubDB in Bari worked independently from the production ones. Bari used LFC as the LCG POOL file catalogue. The CNAF-Bari bandwidth was frequently saturated with PhEDEx-driven transfers in SC3. During SC3, ~7 TB of data were transferred, inherently validated and published in Bari. The SC3 job submissions robot ran ~4k jobs against published data in Bari.

The Legnaro-Padova computing centre participated to SC3 especially in the throughput phase, by running PhEDEx transfers against a SRM/DPM system with 1 SRM/DPM/DPNS server, 1 disk-server with two arrays in a DPM pool, and the DPM client on User Interface and worker nodes. As for CNAF and Bari, independent SC3-specific instances of both PhEDEx and PubDB were deployed. About 4 TB of data were transferred to Legnaro-Padova in SC3. Most time was later spent on debugging the access to DPM-hosted data from CMS applications due to the current rfio-dpm vs. rfio-castor incompatibility issue, and no publishing and job submission was hence possible at Legnaro-Padova within the scope of SC3.

In the job submission step, the job destination pattern shows a first (very draft) example of load balancing depending on data availability (i.e. publishing) at Tiers. This aspect will be further investigated in LCG SC4.

SUMMARY

The WLCG is ready to provide data management and workload management capabilities to the CMS experiment. The INFN Tiers must operate WLCG-enabled services to achieve CMS goals. LCG is approaching the 'regime' through several service challenges. The LCG SC3 experience emerged as a fruitful experience to ramp-up the operational know-how at the involved Tiers. The INFN Tier-1 centre is fast gaining operational experience in daily operations, and good fraction of Tier-2 centres is acting fast to become proficient sites for CMS.

The SC3 operations showed that the know-how of CMS manpower involved at sites is crucial to fulfil a required stability and quality of provided services. In particular at the Tier-2 centres, the number of involved people seems not to reflect the size of the physics community the site will aim to serve.

A level of hybridism* is required to set-up and drive a computing Tier for CMS. Running a Tier-1 for >dozen of experiments and running Tier-2's for 1-2 experiments, as well as drive them through both daily production-like operations and testbed-like service challenges is a complex task which requires hybrid approaches to the exploitation of both infrastructures and involved actors. In most cases it was seen that the operation of most crucial services at Tiers is currently not yet robust and stable enough and needs to be baby-sitted and guaranteed by just few multi-tasking people. If this on one hand helps to build a highly skilled Tiers community, on the other hand it raises issues on stability and reliability of crucial computing services which should not be underestimated in approaching the LHC first data taking, and that will hence be further stressed, investigated and evaluated in LCG SC4.

The INFN focus for the CMS experiment is now on preparing infrastructures for the delivery and the stable operation of robust services. Concerning the CMS Tiers, the INFN experience so far suggests the need to define and build a metrics for success, in order to quantify and evaluate the capability of Tiers to provide the required quality of services, in terms of "continuity" (e.g. long uptime, short recovery times), "efficiency" (e.g. capability to make CPUs proficiently available to users), "robustness" (in the delivery of the required services), "know-how sharing" (e.g. workload sharing, effectiveness of documentation and support).

REFERENCES

[1] CMS Computing Technical Design Report, CERN-LHCC-2005-023, June 2005

[2] Quattor, http://www.quattor.org/

[3] LSF, http://www.platform.com/

[4] http://cms-project-phedex.web.cern.ch/cms-project-phedex/.

[5] J. Rehn et al, "PhEDEx high-throughput data transfer management system", this conference

[6] T. Barrass et al, "Techniques for high-throughput, reliable transfer systems: breakdown of PhEDEx design", this conference

[7] EGEE project, http://public.eu-egee.org/.

[8] LCG project, http://www.cern.ch/lcg/.

[9] D. Bonacorsi et al, "CMS experience in LCG SC3", this conference

[10] PubDB, http://cmsdoc.cern.ch/swdev/viewcvs/viewcvs.cgi/OCTOPUS/PubDB/?cvsroot=OCTOPUS

[11] http://cmsdoc.cern.ch/cms/ccs/wm/www/Crab/.

[12] S. Lacaprara, "CRAB: a tool to enable CMS distributed analysis", this conference

* hybrid ['hI-br&d]: Latin hybrida: 1. an offspring of two animals or plants of different races, breeds, varieties, species, or genera; 2. a person whose background is a blend of two diverse cultures or traditions.