Grid Computing Operations for the CMS Experiment at the GRIF T2.

LLR – E.Poly., Palaiseau:
Pascale Hennion
Paulo Mora de Freitas
Igor Semenjuk
Philippe Busson
Andrea Sartirana

IRFU – CEA, Saclay:
Christine Leroy
Frederic Schaer
Pierrick Micout

LAL – U.Paris-Sud, Orsay:
Michel Jouvin
Guillaume Philippon

LPNHE – U.Paris-7, Paris:
Victor Mendoza
Intro CMS Computing Model

Designed to fulfill the requirements for storage, processing and analysis of data produced by CMS experiment.

- Rely on a distributed infrastructure of Grid resources, services and toolkits whose building blocks are provided by Worldwide LHC Computing Grid [WLCG]
  - WLCG: Computing resources available for LHC experiments. Different MiddleWare implementations: EGEE, NorduGrid, OSG, etc;
- The CMS VO provides the application layer
  - Data Bookkeeping and Location, Data Transfer and Placement, Distributed Analysis and Production Tools, etc.;
- Resources are organized in a tier-ed hierarchical structure:
  - T0 at CERN, 7 national level T1’s, ~50 regional level T2’s.

References:
- CMS computing model document (CERN-LHCC-2004-035)
- CMS C-TDR released (CERN-LHCC-2005-023)
Intro

Tiers & Data Flow

- Data are **collected from online, stored and reconstructed at T0**
  - Information on existing data stored in central DBS at CERN;

- Data **re-reco and filtered in AOD at T1s**
  - according to Ph requests;

- Data **distribution managed by PhEDEx**.
  - RAW/RECO from T0 to T1s;
  - AODs (Analysis format) data among T1s;
  - Data for analysis at T2s;
  - MonteCarlo upload from T2 to T1;

- **Analysis and MC** takes place at T2s (and T3’s)
  - Resources for official analysis groups and local communities.

---

Tier 0

Tier 1

Tier 2

Site

GRIF

LLR-T3

T2_GRIF

Tier 2

XY

FR

GRIF

Data Distribution (PhEDEx)

Data Access

---

09/03/2010

ISGC - Academia Sinica, Taipei, Taiwan
**Intro**

**Tier-2’s Role**

- **Average T2 from Comp. Technical Design Report (CMS C-TDR)**
  - 0.9MS12k, 200TB disk, 1Gb/s WAN;

- **Resources for MC Production**
  - 50% of computing power devoted to simulation;
  - ~20TB for MC data storage;

- **Resources organized Analysis**
  - 40% of comp. power to support each Analysis Group activity;
  - ~30-50TB centrally managed (Analysis Operation)
    - Primary datasets/skims, global interest MonteCarlo samples;
  - ~30-50TB for each supp. analysis group
    - Importing data relevant for analysis;
    - Skims and private productions;

- **Resources local Analysis**
  - 10% of computing power can be reserved to local communities;
  - 30-60TB storage devoted to local usage;
  - ~1TB for each supported user.
Intro CMS GRID Applications

CMS builds its own **app layer** above the GRID MW:

- Data Transfer and Placement Service: **PhEDEx**
  - Distribute data to sites selecting sources;
  - Interfaced with **FTS**;
  - Central brain and local agents at sites;

- Data Bookkeeping and location: **DBS**
  - Global: all *metadata of all official collaboration data*;
  - Also DBS dedicated to *analysis groups and to local communities*;

- Distributed Analysis and Prod tools: **CRAB, ProdAgent**
  - Integrated with DBS and PhEDEx;
  - Fit CMS “data driven” model: *jobs go where data are*;

- Condition Database: frontier
  - *Squid proxy* at each site;

CMS also has its own **support infrastructure**:

- **Contacts at each site** (this is part of the model)
  - Managing CMS *specific apps*;
  - Connecting sites with the central CMS teams;

- **Savannah** tickets with squads of experts
  - “Interfaced” with **GGUS**.
GRIF

Grille au service de la Recherche en Ile de France

Introducing GRIF...

GRIF is a distributed T2 built out of 6 sub-sites in the region of Ile de France.

- CEA, Saclay;
  - 1700 cores;
  - 650 TB.

- U.Paris-Sud, Orsay;
  - 400 cores;
  - 35 TB.

- E.Poly., Palaiseau;
  - 850 cores;
  - 300 TB.

- U.Paris-7, Paris;
  - 250 cores;
  - 165 TB.

- U.Paris-Sud, Orsay;
  - 1800 cores;
  - 240 TB.

- U.Paris-7, Paris;
  - 60 cores;
  - 32 TB.
Introducing GRIF...

- The 6 GRIF sites are seen as a single T2 site from the point of view of the Grid
  - Single BDII entry;

- Summing up to a considerable amount of delivered services/resources
  - ~5000 slots (11 farm clusters);
  - ~1400 TB DPM disk storage (7 DPM instances);
  - GRIF services: CE, SRM, BDII, LFC, WMS, etc;

- More than 20 VOs supported
  - 4 LHC VO’s;
  - Non-LHC VO’s: ILC, babar, dzero, biomed, fusion, etc.

  - 8e+6 h
  - 5e+6 h

- FR-GRIF Normalised CPU time (kS/2K) per VO
  - LHC VOs (and Other VOs), March 2009 - February 2010
  - 37.1%
  - 22.8%
  - 26.9%
  - 3.6%
  - 9.5%
Introducing GRIF

Backbone of fast network connections
- 10Gb/s private network among the GRIF sites is being deployed (LPNHE and APC still to be connected);
- 5Gb/s VLAN to Lyon/CERN;

Seeking high availability by redundancy of services
- Most services are replicated in many instances (CE, SRM, etc.);
- Central services replicated in failover mode (e.g. BDII);

Centralized Nagios monitoring system
- Email notification for failures, problems, etc;

Sharing manpower and technical knowledge
- Configurations managed (and shared) by a central QUATTOR instance;
- Internal wiki and ticketing system;
- Monitoring shifts made by all GRIF administrators.
These are total slots. CMS has ~25% fairshare over all GRIF.
4 GRIF sub-sites support CMS. Grouped in 2 full CMS T2 sites

- From SiteDB: 320 pledged slots (800 kSI2K). 180TB disk;

Adapt the GRIF multisite layout to the CMS “data driven” computing model

- 2 T2 sites with a single SE per site;
- ATM we have a single entry in SiteDB. Soon moving to a completely split configuration;

4 CMS analysis groups supported: Higgs, E-gamma, Exotica, Heavy Ions

- T2_FR_GRIF_IRFU: Exotica + AnaOps managed storage;
- T2_FR_GRIF_LLR : Higgs + E-gamma + Heavy Ions;

2 squid/frontier servers. One for each site with possibility of inter-site failover

- Very stable service, almost no need of management;

One PhEDEx node for each site (SE).

- Sharing configuration and managed in an completely cooperative way.
Some difficulties brought by the multi-site setup

- Still *seeking for consistency* in the dashboard/SAM infrastructure;
- Some *problems multiplied*: 2 SE and 4 CE with different configuration/environment, 2 PhEDEx nodes with different performances, 4 sw areas, etc.

Profiting of redundancy
~205TB of hosted data
× 131TB requested by supported analysis groups;

× Consistent amount of user data. Which includes private productions made by Analysis Groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>Subscribed</th>
<th>Resident</th>
</tr>
</thead>
<tbody>
<tr>
<td>DataOps</td>
<td>1.29 TB</td>
<td>1.29 TB</td>
</tr>
<tr>
<td>FacOps</td>
<td>488.57 GB</td>
<td>488.57 GB</td>
</tr>
<tr>
<td>e-gamma_ecal</td>
<td>67.91 TB</td>
<td>67.91 TB</td>
</tr>
<tr>
<td>ewk</td>
<td>7.97 GB</td>
<td>7.97 GB</td>
</tr>
<tr>
<td>exotica</td>
<td>101.48 GB</td>
<td>101.48 GB</td>
</tr>
<tr>
<td>higgs</td>
<td>20.67 TB</td>
<td>20.67 TB</td>
</tr>
<tr>
<td>local</td>
<td>15.25 TB</td>
<td>15.25 TB</td>
</tr>
<tr>
<td>trigger</td>
<td>1.76 TB</td>
<td>1.76 TB</td>
</tr>
<tr>
<td>undefined</td>
<td>2.18 TB</td>
<td>2.18 TB</td>
</tr>
<tr>
<td></td>
<td>109.63 TB</td>
<td>109.63 TB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group</th>
<th>Subscribed</th>
<th>Resident</th>
</tr>
</thead>
<tbody>
<tr>
<td>AnalysisOps</td>
<td>6.04 TB</td>
<td>6.04 TB</td>
</tr>
<tr>
<td>DataOps</td>
<td>462.77 GB</td>
<td>462.77 GB</td>
</tr>
<tr>
<td>FacOps</td>
<td>488.57 GB</td>
<td>488.57 GB</td>
</tr>
<tr>
<td>e-gamma_ecal</td>
<td>5.95 TB</td>
<td>5.95 TB</td>
</tr>
<tr>
<td>exotica</td>
<td>9.03 TB</td>
<td>9.03 TB</td>
</tr>
<tr>
<td>undefined</td>
<td>1.22 TB</td>
<td>1.22 TB</td>
</tr>
<tr>
<td></td>
<td>23.16 TB</td>
<td>23.16 TB</td>
</tr>
</tbody>
</table>
Link commissioning: T1s and France

- IRFU <-> LLR connection over 10Gb/s GRIF private network;
- 5Gb/s LCG VLAN connection with Lyon;
- Connection among FR T2’s: IRFU->IPHC still missing;
- Working up/down links with all the CMS T1’s.
CMS@GRIF

Data Links...

Commissioning T2’s LInks

Exotica

T2 RU JINR
T2 UK SGrid RALPP
T2 US Purdue

T2 FR GRIF LLR
T2 FR GRIF IRFU

Higgs

T2 IT Legnaro
T2 ES IFCA
T2 US MIT

Egamma

T2 IT Rome
T2 UK London IC
T2 US Caltech
T2 US USCD

Heavy Ions

T2 RU SINP
T3 US Vanderbilt

- Few link missing. Almost all with connection to both GRIF T2’s;
- ~600Mb/s limitation for sites out of GRIF VLAN (others than GRIF and CC) network traffic at IRFU (Planning to pass to 1-10Gb/s connections). 5Gb/s connection already available at LLR.
Data Imports

- 60MB/s daily peaks;
- ~100MB/s hourly peaks;
- Efficient import from many T1/T2 sources;
- Quality of IRFU imports should be improved.

GRIF Imports by source:
Rate vs Quality
Last year

CMS T2’s Rate vs Quality last year

60MB/s daily peaks;
~100MB/s hourly peaks;
Efficient import from many T1/T2 sources;
Quality of IRFU imports should be improved.
Site mostly focused on analysis

- ~50-100k jobs per month;
- Very active local community.

**Last Month activity**

- 98kjob
- 27kjob
- 13kjob

**Success percentage**

- 80%
- 30%

**CMS T2’s analysis jobs**

**Terminated GRIF analysis jobs**
Ramp up of activity with Data Taking and with the New Year

- Before we used to have few power-users
  - Great load by few well defined high-scale tasks;
  - Skilled users which do not need much support;
- Now number of users multiplied
  - This did not necessarily increased the load.
  - Smaller tasks;
  - More difficult to support an heterogeneous user community.
Main issue in supporting analysis turned out to be the file serving capacity

- The high load brought by many jobs reading files on the same server may lead the server performances to drop to 10% of the expected throughput.
  - DPM does not manage any queue for the rfio file requests, therefore 1000 requests turn into 1000 rfio processes running;

- In the short term we take care of distributing the files which are likely to be accessed together among the servers;

- In the longer term we may look for better storage access solutions
  - E.g. make some tests on the xrootd performances in such cases;

- Important to keep the equilibrium between storage and calculation power;

Some CMSSW vs DPM issues met passing from SL4 to SL5

- Similar problems already seen with versions of CMSSW on sl4. CMSSW is not much tested against different storage technologies (not against DPM, in particular)
  - Keeping an high level of warn to figure out when such issue turn up and come to a solution as fast as possible.
GRIF also supports the official MC production by CMS

- Much less activity than analysis
  - Order 5-10 smaller;
  - Burst of activity in late Summer. Then the site utilization dropped in Fall;
  - Of course we would prefer to have a constant load. But it is understandable that this may not be always the case;

- No particular issues met
  - MC is managed by a central team of skilled operators: issues are rare and immediately addressed.

Success percentage

Terminated GRIF MonteCarlo jobs

Last 6 months
09/09 – 02/10

118kjob

27kjob

6kjob
Summary

Longstanding experience in supporting **CMS activity** within the **framework of a distributed Tier-2**:

- **Shared management** of CMS specific services: PhEDEx, squid/frontier;
- Multi site setup brings **redundancy and sharing of manpower** leading to **high reliability/availability**;
- Ongoing **integration of the multi-site setup into the CMS framework**:
  - 2+2 site configuration works fine. **Moving to a full splitting** of the 2 CMS sub-sites;
  - Still some inconsistencies: downtimes accounting, etc.;

**CMS activity at GRIF** is **mostly focused on Analysis**:

- Important and stable activity **pushed by very active local communities**
  - Improved **data hosting capacity and links to relevant sites**;
  - Since last Summer **activity constantly increasing**;
- **Important issues** have been pointed out and roadmaps to their solution defined
  - **File serving capacity** is the key issue for analysis.