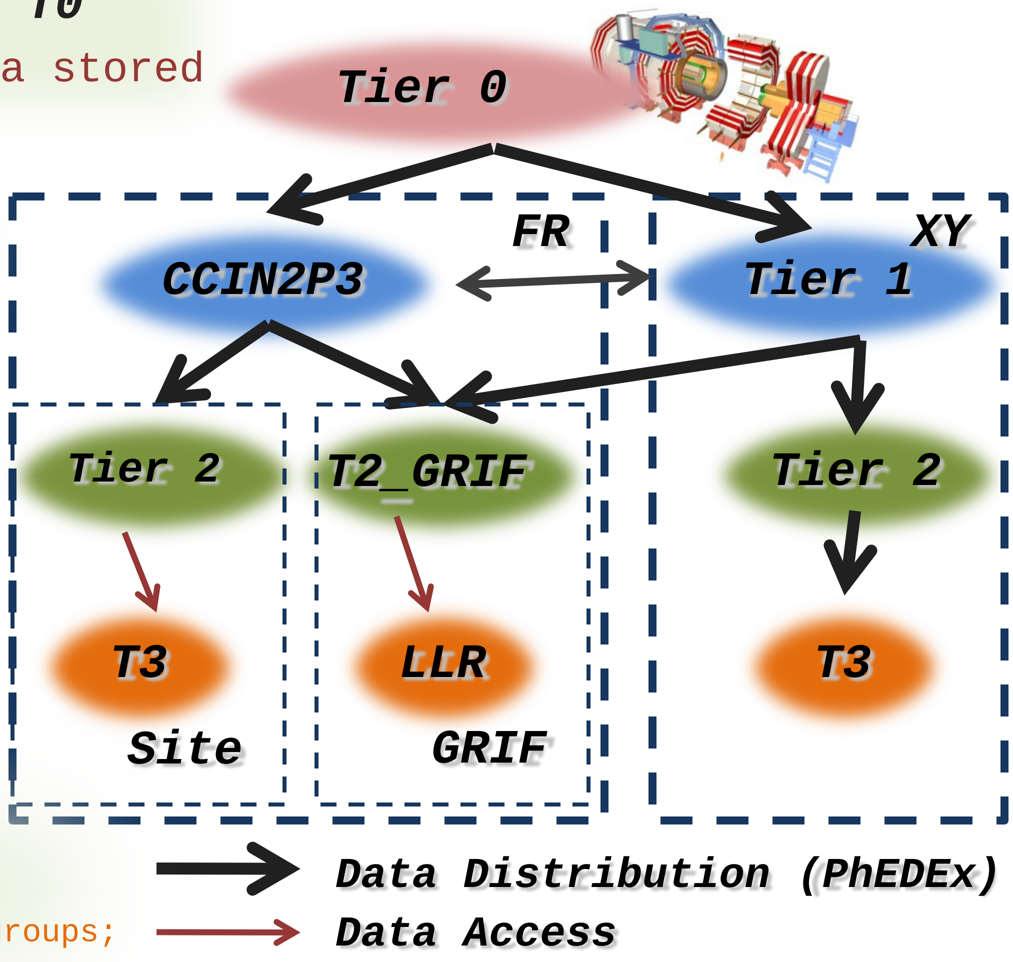
Link between Tier 3 and Tier 2

Hints for discussion on the basis of LLR T3 example.

Andrea Sartirana (LLR – Ecole Polytechnique).

Introduction Tiers

- 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 among T1s;
  - Data for analysis at T2s;
  - MC upload from T2 to T1;
- ***Analysis*** takes place at ***T2s and T3s***
  - T2 official analysis groups
    - DBS instances dedicated to ph groups;
  - T3 local communities
    - Local DBS instances.



Introduction

Tier 2’s

- Resources for ***MC Production***
  - ***50% of computing*** power devoted to simulation;
  - ~20TB for MC data storage
    - ~5MB/s[*] upload rate to reg. T1;
- Resources ***organized Analysis***
  - ***40% of computing*** power devoted to Physics groups activity;
  - ***~30TB centrally managed storage***
    - Primary datasets/skims, global interest MC samples;
  - ***~30TB***[*]for each DPG/POG/PAG supported
    - Needed for ***host data*** (real or simulated) relevant for analysis, ***store "private production" and results***;

T2’s are "public" resources in the CMS Comp. Model.

- ***0.9MSI2k*** of computing power (corresponding to several ***100s of batch slots***);
- ***200TB Disk*** Storage;
- ***1Gb WAN*** Network.

Nominal T2 *(from CTDR).*

In the real world resources can considerably ***vary from case to case.***

- Resources ***opportunistic/local Analysis***
  - ***10% of computing*** power can be reserved to local communities;
  - ***~1TB*** for each supported user.

[*] From link commissioning metrics.

Introduction Tier 3's

- T3 ***may mean*** many *different things*
  - Some are just ***fractions of a T2***
    - Prioritized/reserved usage of Comp resources;
    - Storage space;
  - Some are ***real individual resources***
    - Local institutes clusters/farms;
    - Some are as big as T2's;
  - Many are a mix of the two;
- Resources for ***local Analysis*** groups
  - Real requirements came from the local community
    - All that a is needed by the end-user to setup his/her analysis;
    - A way to perform urgent tasks: prio-ritized/exclusive access to resources;
- ***Opportunistic MC*** resources.

A Tier 3 is a "private" resource.

- There are ***no requirements*** for Tier 3 resources
- Tier 3 ***do not play any official role*** in the CMS computing system
- Tier 3 are part of CMS Computing system: they may have ***PhEDEx node***, can be included in the ***SAM/JobRobot*** infrastructure, etc.

Recently CMS made a survey "in order to try to understand the range and diversity of what CMS is calling a Tier 3" (D.Colling) [*].

[*]http://indico.cern.ch/getFile.py/access?contribId=21&sessionId=0&resId=1&materialId=slides&confId=56278

Introduction T3 in France

- From CRB survey

T2 fraction.

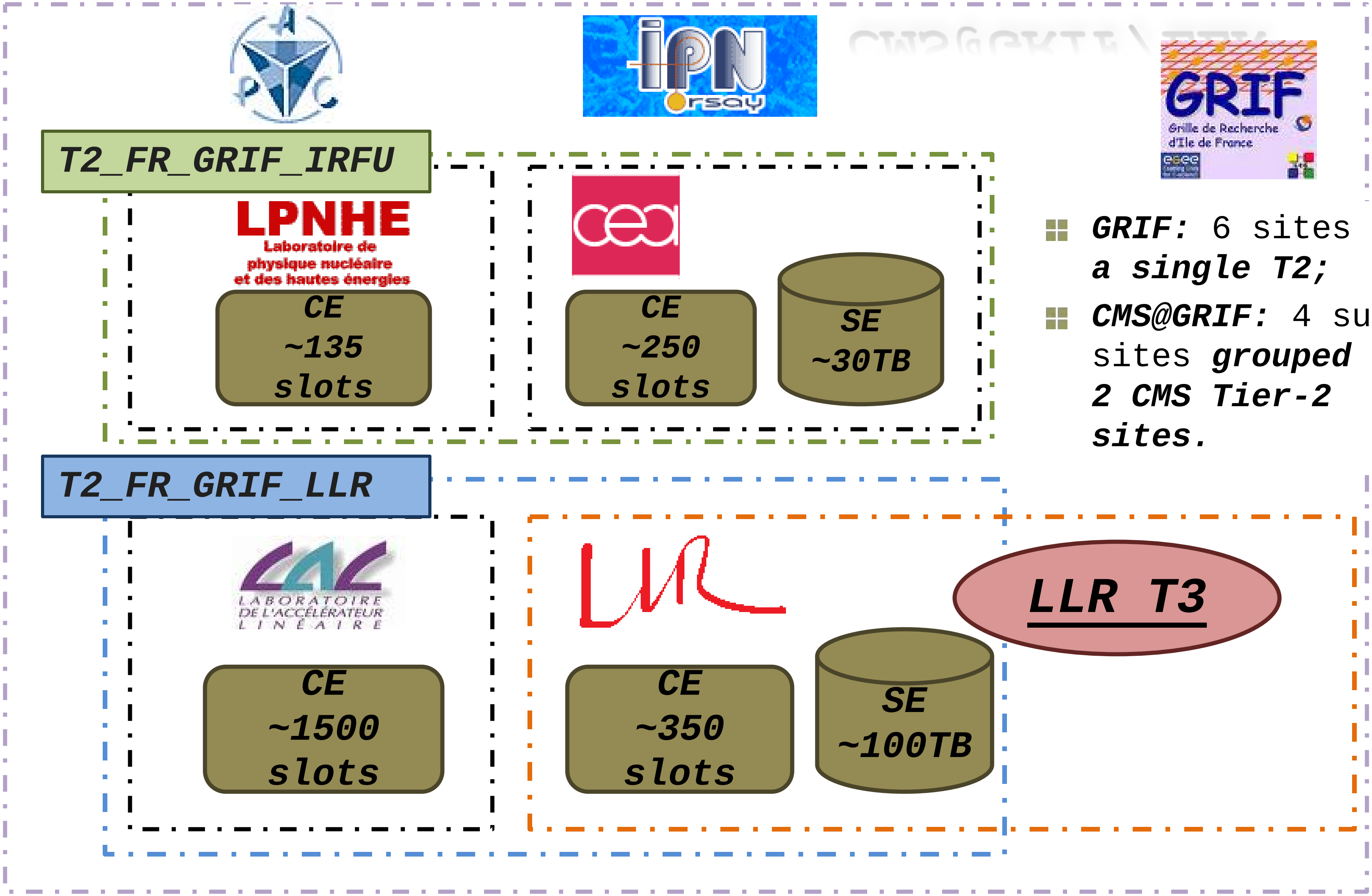
- ✖ ***GRIF_IRFU:*** *"Just going to take 20% of the T2 resources, however no mechanism for only allowing local users on to 20% of disk – possible desire for space tokens";*
- ✖ ***GRIF_LLR:*** *"Co-located with T2, still planning but expect around 10% of the T2 size. Will be a high priority part of the T2 farm. Storage will not be grid enabled.";*
- ✖ ***IPHC:*** *"~10% of T2 (sometimes more). Completely Grid enabled extension of the T2. Priorities handled through fair share. Situation evolving.";*

- ✖ ***T3_FR_IPNL:*** *"Not a T2 site. 172 cores (70% for CMS) and 50 TB of storage (24 To for CMS, declared as a Phedex node). Will increase the number of cores to around 400 in the coming weeks. Site is open to all CMS but local users have a higher priority 3 person (part time) support team (both system and user). Regular contact with other French T1 and T2s. Contributes to MC production.".*

Full CMS Site.

- This was 2009-03-04. Where are we now?

Introduction CMS@GRIF/LLR



T2_FR_GRIF_IRFU

LPNHE
Laboratoire de physique nucléaire et des hautes énergies

CE
~135
slots

CE
~250
slots

SE
~30TB

T2_FR_GRIF_LLR

LABORATOIRE DE L'ACCÉLÉRATEUR LINÉAIRE

CE
~1500
slots

CE
~350
slots

SE
~100TB

LLR T3

- ***GRIF:*** 6 sites ***as a single T2;***
- ***CMS@GRIF:*** 4 sub-sites ***grouped in 2 CMS Tier-2 sites.***

T3 Setup

Requirements

To start with, we made up a list of "user's requests"

- ***Interactive Usage:*** edit code, build application, run test jobs. Efficient processing of large number of root ntuples;
- ***Local batch Usage:*** *fast turnaround* job submission for testing/debugging analysis tasks. Possibility to *follow and debug jobs in real time*;
- ***Prioritized Farm Access:*** *Prioritized/privileged usage of Grid calculus resources* for local users;
- ***Data Access:*** *easy and convenient* access to storage to manipulate data files. *Fast* (prioritized?) *access to the T2 data;*
- ***User Data Management:*** space to *store user data*: results of analysis or private productions. *Tools for manage* these data, *share with other users*, etc (e.g. local DBS). *Safeguard* of unrecoverable private data;
- ***Local Data Areas:*** easily accessible (i.e. POSIX) storage areas for storing temporary files, logs, etc.. .

T3 Setup

Guidelines

Few Guidelines we would like to follow in setting up the T3:

- *Fulfill users requests/needs (of course!)*
  - The T3 should be, as much as possible, a user-tailored resource;
- *Avoid new load on the Tier 2 administrators*
  - They are already overwhelmed by the management of the T2;
  - The *Tier 2* is a “public” resource and *should not be penalized*;
  - *When possible, avoid adding new services/infrastructures* that have to be managed;
- *Avoid “violence” to the Tier 2*
  - “Hosting” T3 services inside the T2 requires adapting the configuration. We should try to *keep T2 configuration as standard as possible*;
  - Despite the previous point… sometimes dedicated services are the less demanding solution;
- *Clear deals on resources exploitation/management*
  - *What is T2 and what T3? Who manages what?*
  - Local CMS group may be required to manage some of the most CMS-specific services (e.g. local DBS).

Solutions Interactive Usage

Pool of User Interfaces

- should be sized to the effective needs of the local users
  - At the moment we have 2 CMS UI’s;
- Should be protected from misuse
  - E.g. interactive running of lengthy/heavy jobs;
- Ease of usage
  - Shared homes;
  - Single login;
  - Cluster-like organization (like “ccali”) [*];
- Xrootd/Proof cluster?
- Full access to DPM data
  - Still some problems (wrong libraries?) in accessing files on DPM by root.

Usage Details

- build code;
- run lightweight jobs;
- use root tools to access and analyze the output files;
- Root ntuple processing;
- access Grid resources.

Status

UI’s are already part of the Tier 2 services. We have to rationalize the Setup.

[*] Under study: single OS over more machines (Kerrighed-SSI kernel).

Solutions Local Batch Usage…

Usage Details

- local batch submission;
- Debug analysis tasks: limited number of heavy jobs;
- realtime access to the running jobs;
- Access to input data, space for storing outputs and logs.

Sol. 1:Cluster of dedicated WN's

- Few nodes (20-30 slots?) managed by a dedicated scheduler
  - Number of nodes may increase/decrease on demand;
- Optimized for debugging/short turnaround usage
  - Batch submission with local user;
  - Interactive access to the nodes (same login as the UI);
  - "real" WN's environment: realistic Grid-like test;

…

Solutions …Local Batch Usage…

….

- ✖ Shares with UI’s the same /home and /data areas
  - Easy to retrieve outputs and logs;
  - Easy to setup the CMSSW environment;
- ✖ Full access to DPM data.

Status

All the practical aspects (which scheduler, how to setup the node, etc..) are still under study. Note: this solution involves deployment and maintenance of extra-T2 services and should be discussed/negotiated with the T2 administrators.

Solutions …Local Batch Usage

Sol. 2:Enable privileged usage on the T2 farm.

- Local batch submission on the T2 CE
  - Also mounting the shared /home and /data areas on the nodes;
- (gsi)ssh user login on (some of the) the nodes.

Status

This solution has the advantage of not requiring extra services but may lead to a very weird configured T2 CE. For the moment it is excluded.

Related issue: Crab with local batch submission

- Crab functionality developed for the CAF (LSF and...)
  - Can we use it with, e.g. PBS? Will require development?
  - CRAB developers may not have the manpower. Should we contribute on our side?

Solutions Prio. Farm Access…

Usage Details

- High priority usage of a fraction of T2 Grid resources;
- Normal Grid submission with, as much as possible, dedicated/controlled resources (e.g. WMS);
- As much as possible, uniform to an usual Grid task.

Dedicated Tier 3 queue

- Can be setup to higher priority
  - Some FairShare translation of the required 20% of resources;
- On demand, it can also be pointed to some dedicated WN’s
  - Solution in sight of very urgent tasks;
- Not much load on the administrator side
  - Easy to setup (i.e. just a change in the configuration);
  - Resources can be increased/decreased on demand;
- 2 Possible solutions, on the user side, for accessing the queue.

Solutions …Prio. Farm Access…

Local LLR VO

- The T3 queue is mapped to a local Vo (e.g. vo.llr.in2p3.fr);
- This can be setup in Crab. Requires some user setup but not a big deal

Crab.cfg ex.

```
[EDG]
wms_service=https://grid25.lal.in2p3.fr:7443/glite_wms_wmproxy_server
ce_white_list = polgrid1.in2p3.fr
dont_check_proxy = 1                    #user has to take care of the proxy.
virtual_organization = vo.llr.in2p3.fr
```

- Relies only on “local” resources
  - Relies on GRIF voms servers and WMS’s;
- Not completely clean on the Grid point of view
  - E.g. files written on the storage with local VO’s permissions. Need to set by hand the ACL in the /store/user areas;

Solutions …Prio. Farm Access

Dedicated Role within CMS voms schema

- The T3 queue is mapped to some local Group/Role of the VO CMS;
- Should be discussed and agreed with the CMS VO managers;
- May be embedded in a more general French CMS Groups/Roles definition;
- CRAB setup is straightforward;
- Clean on the Grid POW.

Status

C.Charlot already took contacts with the CMS VO managers and discussed a possible setup of the French VOMS Roles/Groups. The solution with LLR VO has been tested. We just miss to make the desired changes in the CE configuration.

Solutions

Data Access

Usage Details

- Access to data for analysis;
- Prioritization wrt external users.

Access to the Tier 2 SE

- The Tier 3 relies on the Tier 2 SE for official data for analysis
  - No T3 dedicated PhEDEx node;
- Xrootd server can be used as cache disk for privileged access to data
  - ATM Xrood installed on DPM but without dedicated servers;
  - Not clear (at least to me…) how to enable T3 users to use xrootd within CMSSW.

Status

We need a thorough study of the storage access patters, taking into account Tier 2 as well as Tier 3 workflows. On the basis of this we may setup some prioritization mechanism for local users access to relevant data.

User Data Mgmt

Usage Details

- Store the analysis results;
- Handle, share, publish the user data;
- Safeguard of the unrecoverable data;

- Setting up tools for management and monitoring of the /store/user area
  - See e.g. http://polywww.in2p3.fr/~sartiran/monitoring/cmsmon.php;
  - User space usage may become an issue;
- Planning to install a local DBS for publishing the results within the local groups
  - Who is supposed to manage this?
- Studying the possibility to use the Lyon HPSS for safeguarding the user data
  - Not a full backup. Each user will select the data which he/she wants to safeguard. E.g. by copying them in a backup buffer on the local storage;
  - The actual transfer/replication to Lyon may be managed at administrator level.

Solutions

Local Data Areas

- We already have NFS mounted scratch areas on the UI's;
- We are thinking about moving to GPFS for all the locally mounted partitions
  - VO's SW area;
  - Users HOMES (on SAN for redundancy);
  - Scratch and data areas;
- Within this migration we will probably reorganize the filesystems and the mounting points
  - Uniform configuration;
  - Quotas;
  - Maybe some data filesystems also mounted on WN's.

Related issue: FileMover

- Web tool for downloading a files on a local FS
  - User just need the LFN (or Dataset + Run) without caring about PFN and Source.
  - Installed and running on the LLR CMS UI's.

Summary/Conclusions

- We described the status of the ***setup of the LLR Tier 3 within the GRIF_LLR Tier 2***:
  - We started from ***list of requirements***:
    - They come from a ***discussion with C.Charlot, P.Busson, N.DeFilippis***, started ~1 month ago (but the project of an LLR T3 are much more longstanding);
    - Also discussed with LLR admins: P.Mora, P.Hennion, I.Semenjouk
    - They are based on the ***longstanding experience of analysis activity***, mostly ***at CCLyon,*** within CMS and other experiments;
  - We provided a, still non-definitive, ***list of possible solutions***:
    - Situation may change with the next iterations: new requirements may appear, or better insight on the existing ones;
    - The feasibility of some of these solution has not yet thoroughly investigated;
    - Some are alternative answer to the same problem.
- We wish to have ***feedback from the other sites***:
  - Are you facing the same issues?
  - Which is your roadmap to address them?
    - Different/Better solutions to the same problems?
  - ***Actual status?***

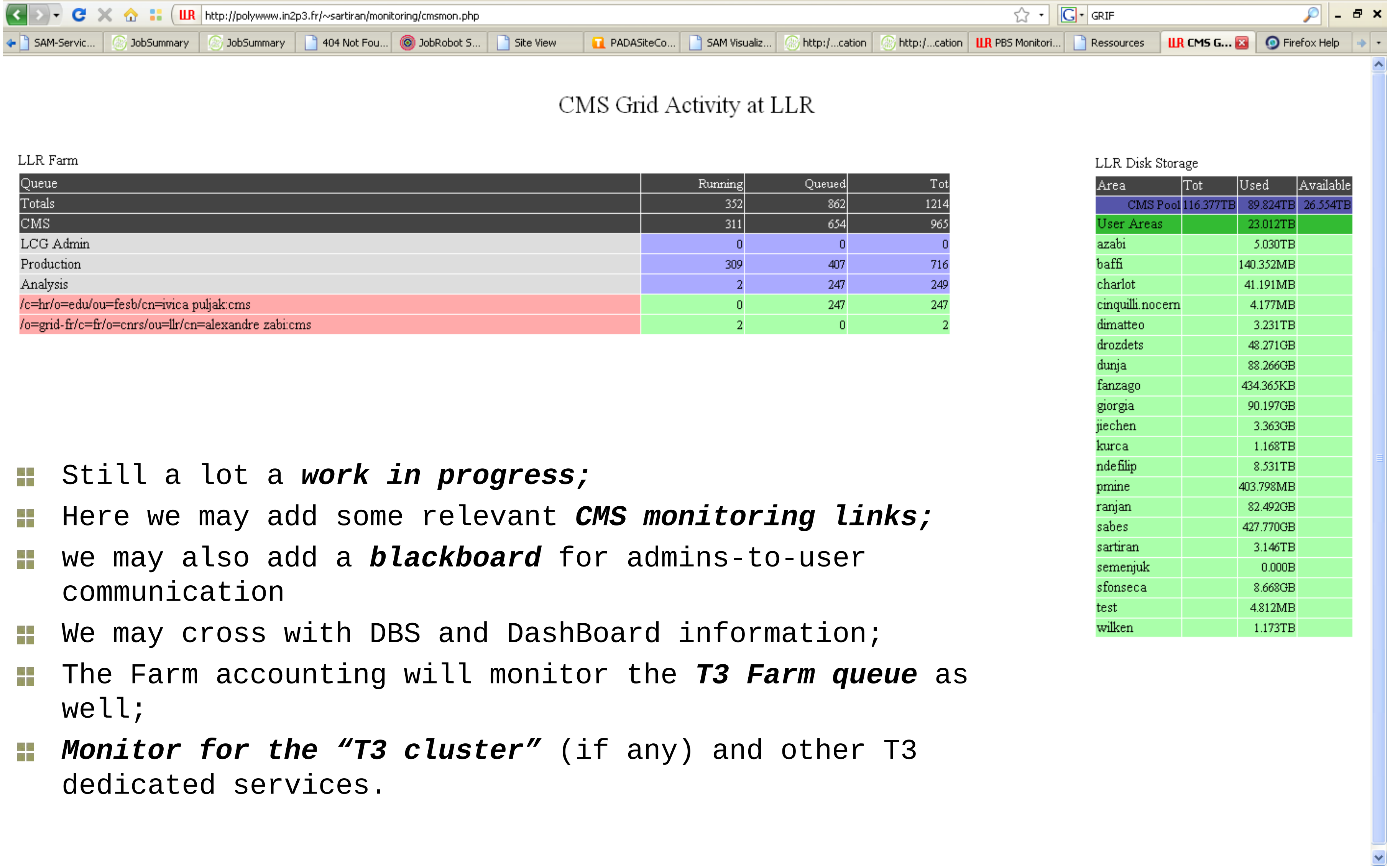
BackUp

CMS Computing

- Designed to ***fulfill the requirements for storage, processing and analysis*** of data produced by CMS experiment.
  - Rely on the services, toolkits and *distributed infrastructure* of the ***Worldwide LHC Computing Grid*** [WLCG]
    - WLCG : Computing resources available for LHC experiments. Different MW implementations: LCG-2, Grid-3, EGEE, NorduGrid, OSG;
  - ***Experiments*** should ***provide the application layer***
    - Data Bookkeeping/Placement, Distributed Analysis/Production Tools;
  - Computing resources are organized in a ***tier-ed hierarchical structure***:
    - Tier-0 (CREN): data from DAQ, real time RECO, custody on tape, distribution to T1's;
    - Tier-1's (7 national centers): 2nd custodial copy of data, re-reco, distribute data to T2's and simulated data;
    - **Tier-2's (regional centers, ~50) MC Production and user analysis;**
    - **Tier-3's (any other resources) end-user analysis.**

- CMS computing model document (CERN-LHCC-2004-035)
- CMS C-TDR (CERN-LHCC-2005-023)

BackUp CMS@LLR Monitoring



CMS Grid Activity at LLR

LLR Farm

Queue	Running	Queued	Tot
Totals	352	862	1214
CMS	311	654	965
LCG Admin	0	0	0
Production	309	407	716
Analysis	2	247	249
/c=hr/o=edu/ou=fesb/cn=ivica puljak:cms	0	247	247
/o=grid-fr/c=fr/o=cnrs/ou=llr/cn=alexandre zabi:cms	2	0	2

LLR Disk Storage

Area	Tot	Used	Available
CMS Pool	116.377TB	89.824TB	26.554TB
User Areas		23.012TB	
azabi		5.030TB	
baffi		140.352MB	
charlot		41.191MB	
cinquilli.nocern		4.177MB	
dimatteo		3.231TB	
drozdets		48.271GB	
dunja		88.266GB	
fanzago		434.365KB	
giorgia		90.197GB	
jiechen		3.363GB	
kurca		1.168TB	
ndefilip		8.531TB	
pmine		403.798MB	
ranjan		82.492GB	
sabes		427.770GB	
sartiran		3.146TB	
semenjuk		0.000B	
sfonseca		8.668GB	
test		4.812MB	
wilken		1.173TB	

- Still a lot a ***work in progress;***
- Here we may add some relevant ***CMS monitoring links;***
- we may also add a ***blackboard*** for admins-to-user communication
- We may cross with DBS and DashBoard information;
- The Farm accounting will monitor the ***T3 Farm queue*** as well;
- ***Monitor for the "T3 cluster"*** (if any) and other T3 dedicated services.