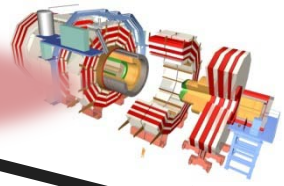
Organization of CMS and T2/T3 resources sharing at GRIF.

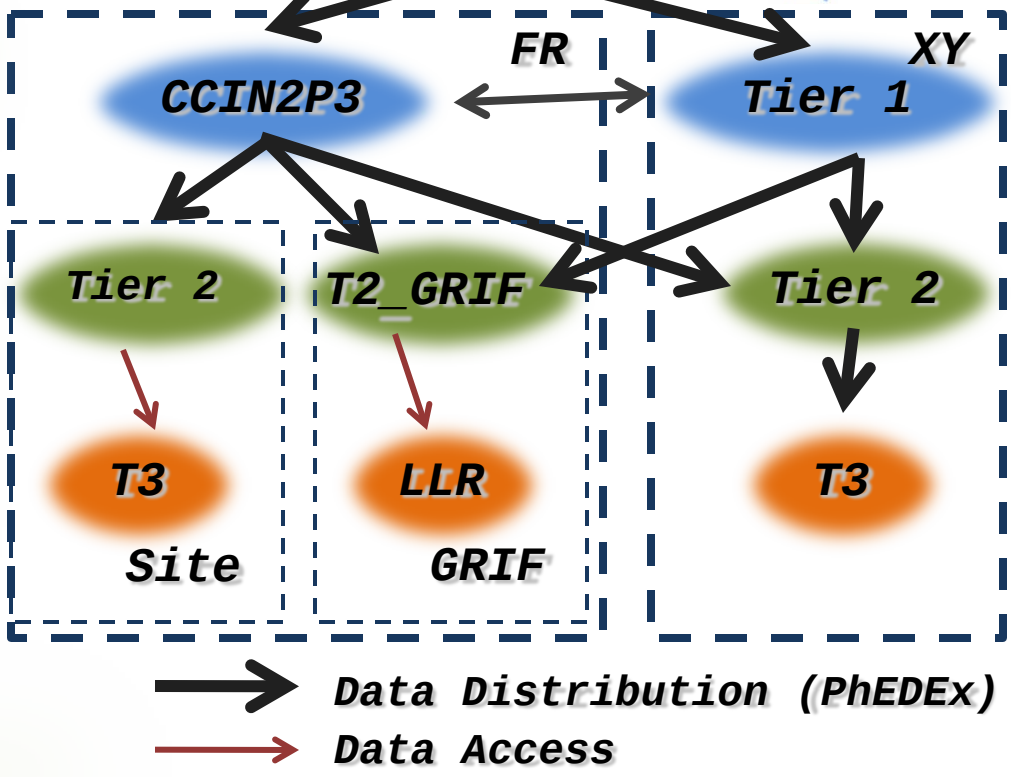
Christine Leroy (IRFU – CEA)
Pascale Hennion (LLR – Ecole Polytechnique)
Igor Semenjuk (LLR – Ecole Polytechnique)
Andrea Sartirana (LLR – Ecole Polytechnique)

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 (Analysis format) data among T1s;
  - Data for analysis at T2s;
  - MonteCarlo upload from T2 to T1;
- ***Analysis*** takes place at ***T2s and T3s***
  - Resources for official analysis groups and local communities





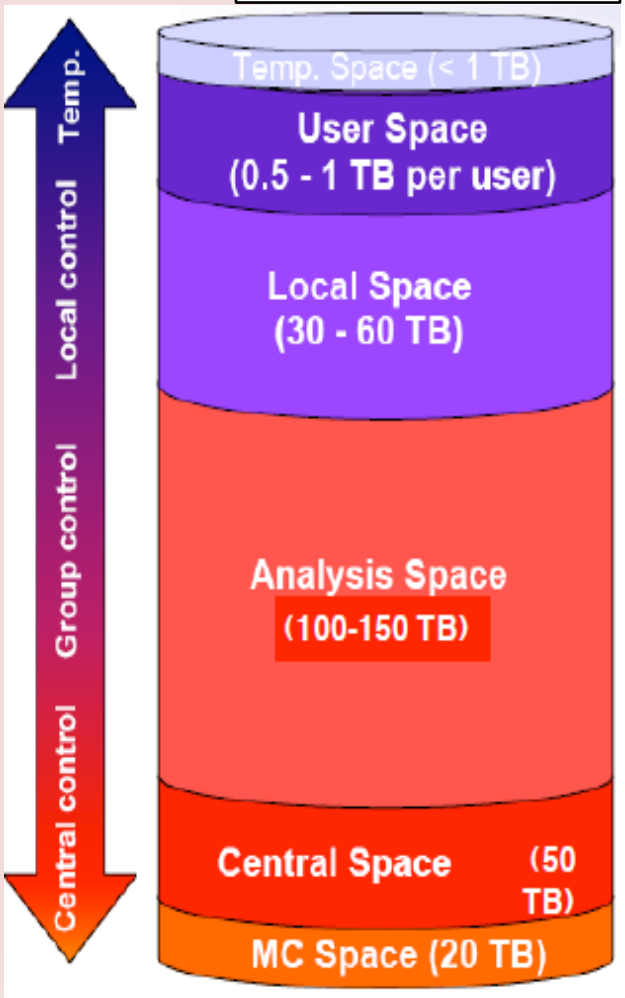
Introduction

Tier 2

T2:"public" resources

- ***Comp. Technical Design Report*:*0.9MSI2k*, *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 supp. DPG/POG/PAG activity;
  - ***~30-50TB centrally managed (AnaOps)***
    - Primary datasets/skims, global interest MC 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.

Average T2



Introduction

Tier 3

T3:"private" resources

- ***No requirements*** from Computing Model but fully ***embedded in CMS Computing***
  - Tier 3 do not play any official role and have no responsibilities;
  - They are part of the CMS Computing System: they can have PhEDEx nodes, can be included in DBS, in the SAM/JobRobot infrastructures, etc;
- ***Different*** in size and type
  - Some are ***just fractions of a T2:*** i.e. everything which is ***above the 20-30% fraction devoted to local communities*** according to Computing Model (see prev. slide)
    - Prioritized/reserved usage of Comp resources;
    - Storage space;
  - Some are **independent resources**
    - Local institutes clusters;
    - Real full GRID sites;

In most cases: a mix of the two things.

- Resources for ***local Analysis*** groups
  - Real ***requirements came from the local community***
  - All that is needed by the end-user to setup his/her analysis;
  - A mean to ***perform urgent tasks***;
- ***Opportunistic MC*** resources.

Introduction CMS GRID FW

- 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

In a nutshell

- 112 cores;
- 7 TB.

- 395 cores;
- 40 TB.

- 250 cores;
- 43 TB.

- 580 cores;
- 450 TB.

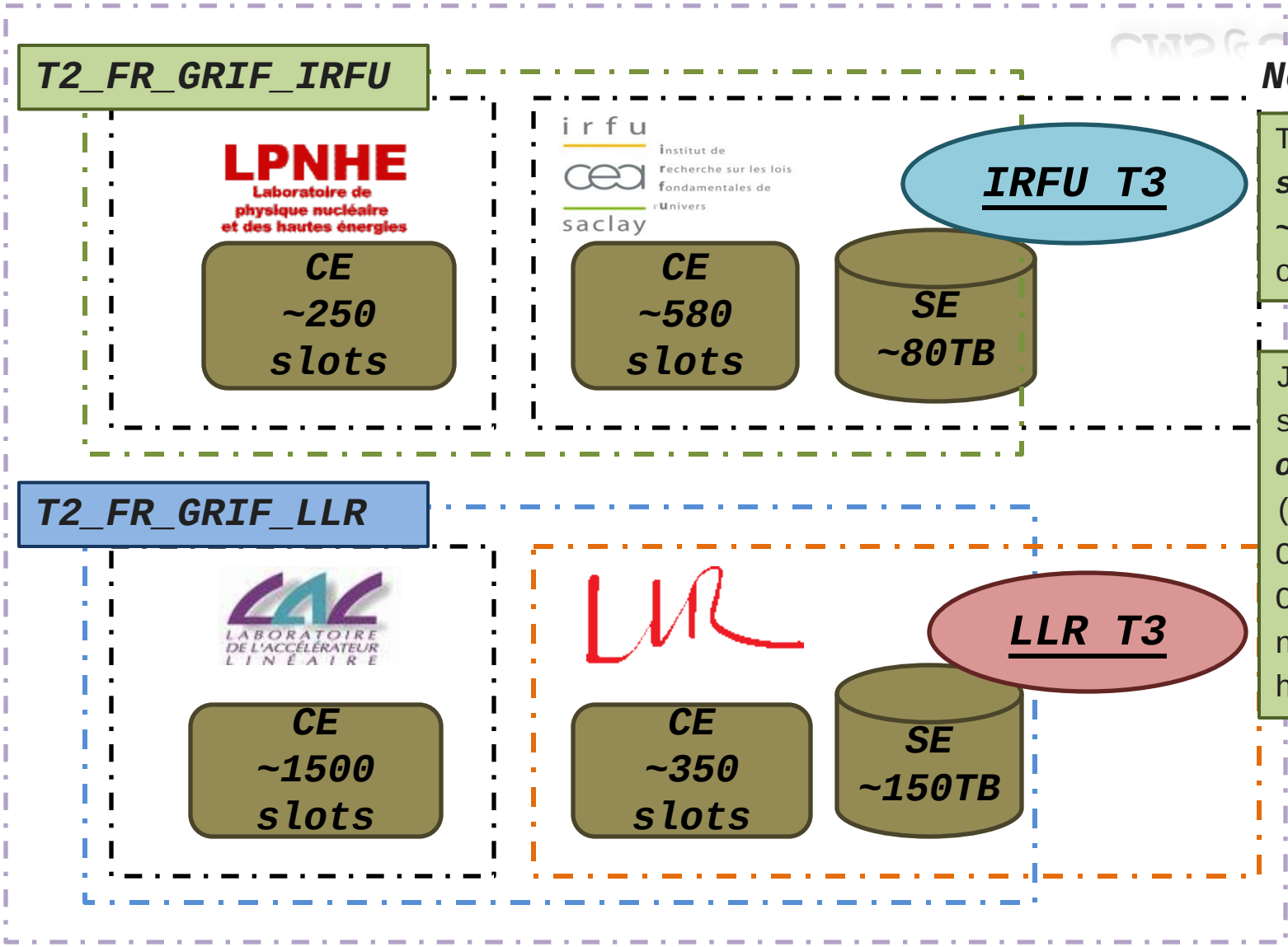
- 1750 cores;
- 105 TB.

- 350 cores;
- 254 TB.

- **6 sites confederated** in a single GRID T2
  - ~3500 cores;
  - ~900 TB DPM disk storage;
  - **GRID services**: CE, SRM,BDII,LFC, etc;
  - >16 VO's supported: LHC,ILC,etc;
- **Fast network** connections
  - 10Gb/s private inter-GRIF;
  - 4Gb/s VLAN to Lyon/CERN;
- **Redundancy and load sharing**
  - Replication of services leading to **high availability**;
  - **Shared configuration** (quattor);
  - **Shared manpower** for support/administration.

GRIF

CMS@GRIF



Notes:

These are ***total slots***. CMS has ***~25% fairshare*** over all GRIF

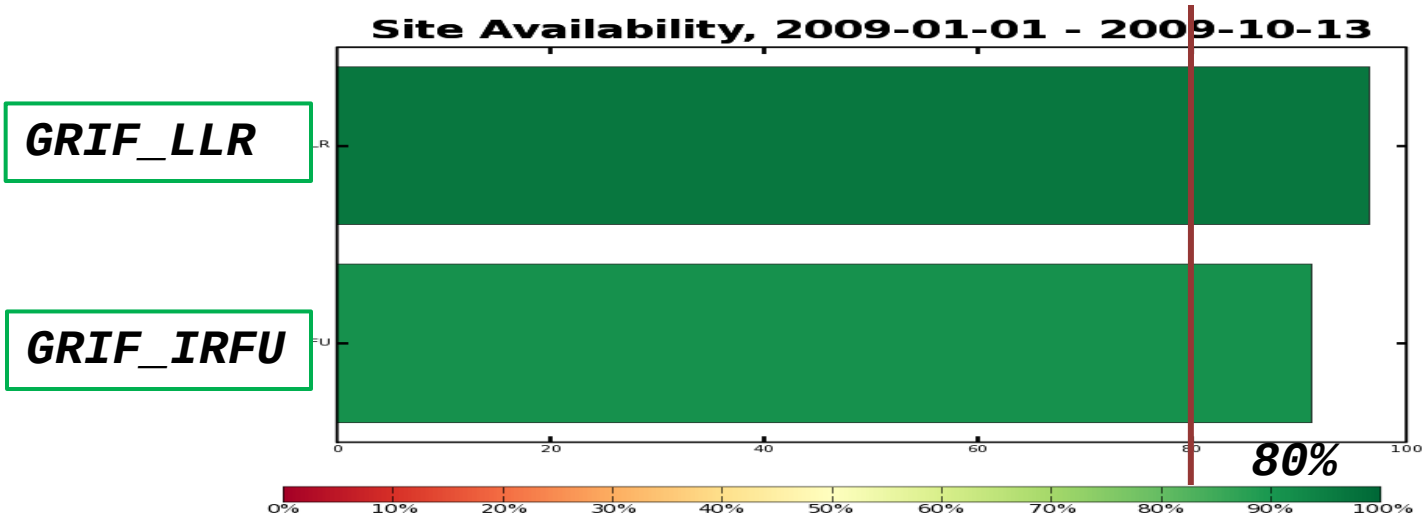
Jobs at each T2 site access ***only*** to ***1 SE*** (published as Close SE for CMS). This is needed as CMS has ***no LFC***.

GRIF Tier 2 Overview

- ***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***;
- 3 CMS analysis groups supported: ***Higgs, E-gamma, Exotica***
  - T2_FR_GRIF_IRFU: exotica + AnaOps managed storage;
  - T2_FR_GRIF_LLR : Higgs + E-gamma;
- **2 squid/frontier servers**. One for each site with possibility of intersite 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;

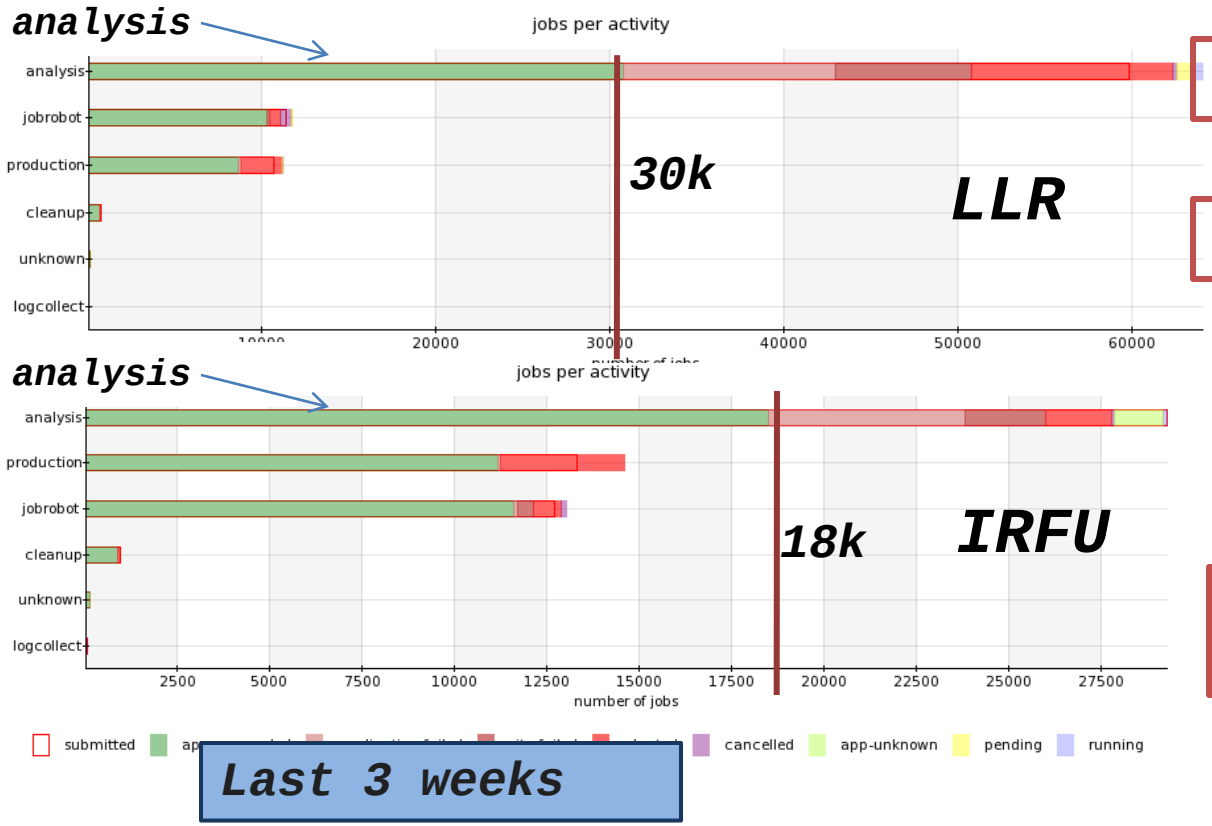
GRIF Tier 2 Availability

- ***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/cooperation: High availability and reliability***
  - ***High availability*** ranking for both WLCG and CMS monitoring;



GRIF Tier 2 Activity

- ***Rich ongoing CMS activity***
  - Up to 30-80kjobs/month;
  - Data stored for supported groups analysis
  - ~30TB User data stored in DPM.



Last 3 weeks

T2_FR_GRIF_LLR Group Usage

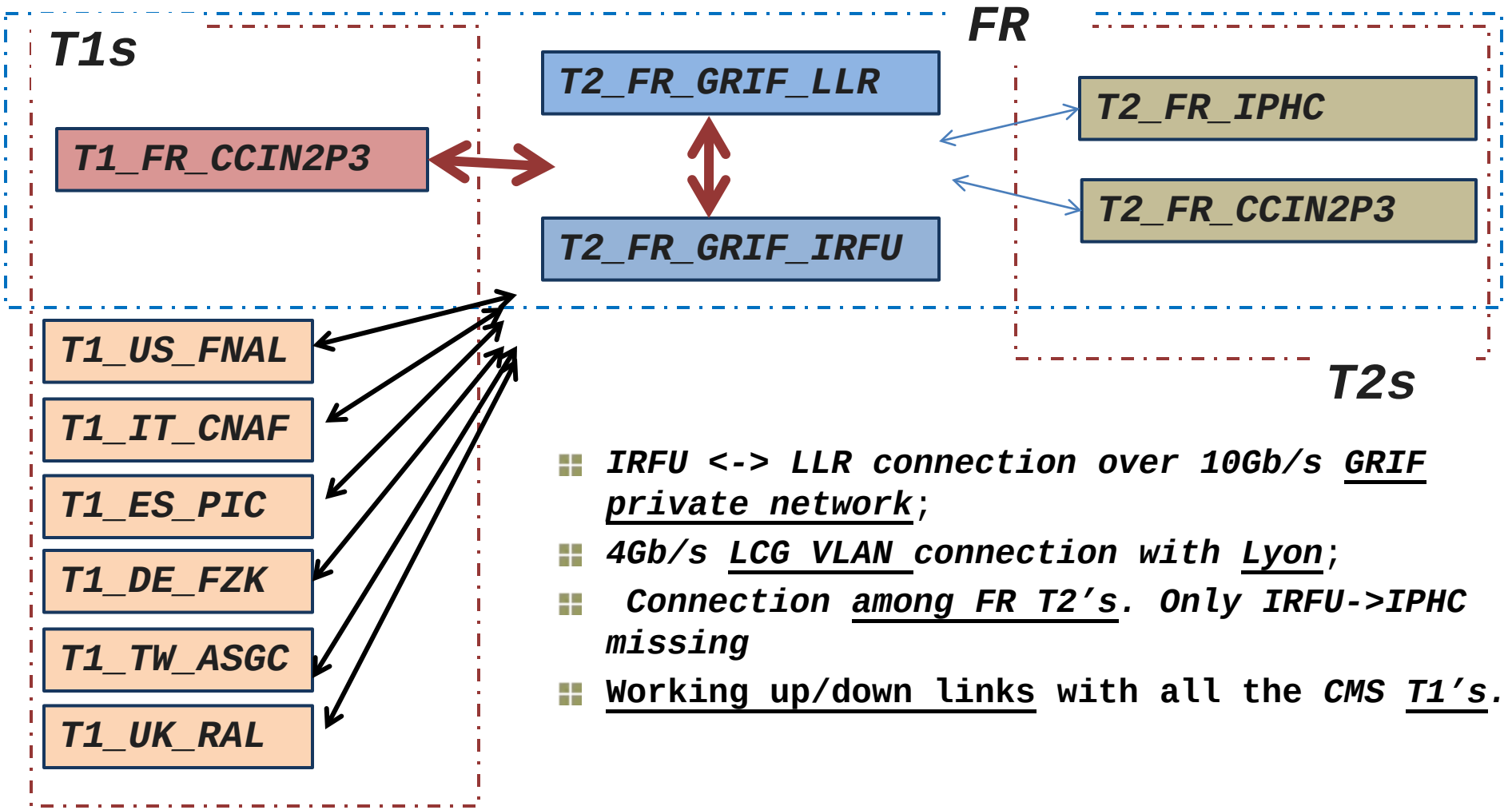
Group	Subscribed	Resident
DataOps	853.60 GB	853.60 GB
e-gamma_ecal	48.57 TB	48.50 TB
ewk	7.07 GB	7.07 GB
exotica	1004.51 MB	1004.51 MB
higgs	18.17 TB	18.17 TB
trigger	1.76 TB	1.76 TB
undefined	399.27 GB	326.38 GB
	69.73 TB	69.59 TB

T2_FR_GRIF_IRFU Group Usage

Group	Subscribed	Resident
AnalysisOps	17.86 TB	17.47 TB
exotica	2.74 TB	2.74 TB
undefined	2.91 TB	2.91 TB
	23.51 TB	23.13 TB

GRIF Tier 2 Data Links…

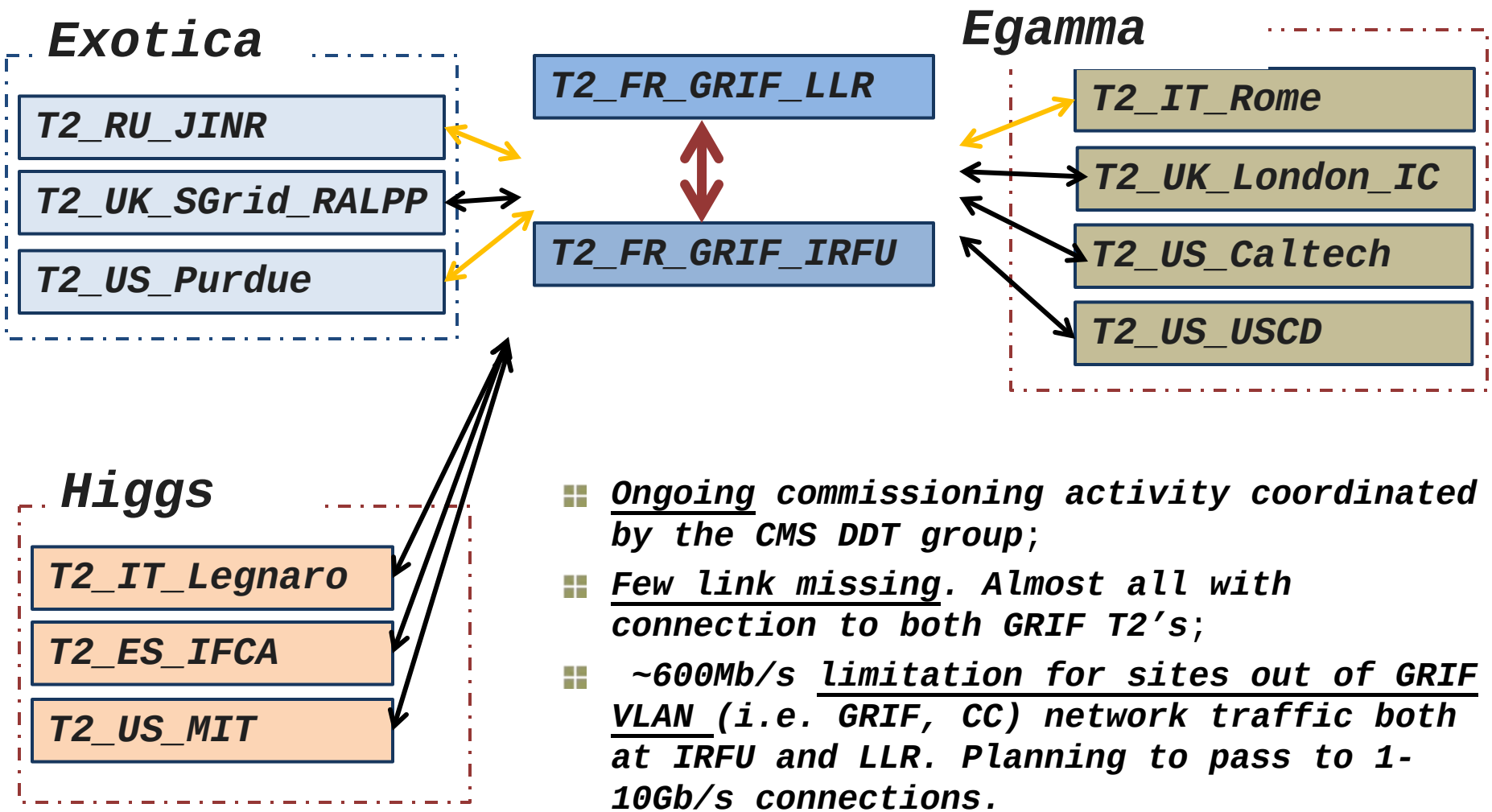
Link commissioning: T1s and France



- IRFU <-> LLR connection over 10Gb/s GRIF private network;
- 4Gb/s LCG VLAN connection with Lyon;
- Connection among FR T2's. Only IRFU->IPHC missing
- Working up/down links with all the CMS T1's.

GRIF Tier 2 Data Links…

Link commissioning: Analysis Group T2's



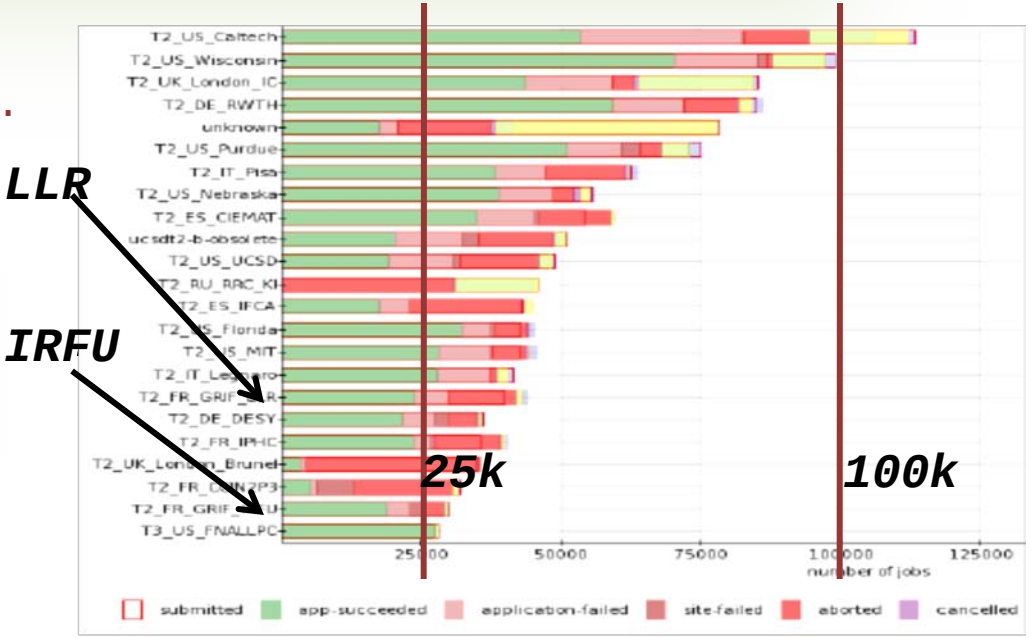
- *Ongoing commissioning activity coordinated by the CMS DDT group;*
- *Few link missing. Almost all with connection to both GRIF T2's;*
- *~600Mb/s limitation for sites out of GRIF VLAN (i.e. GRIF, CC) network traffic both at IRFU and LLR. Planning to pass to 1-10Gb/s connections.*

GRIF Tier 2 — Oct'09 Exe

- ***First complete data-taking scale test of chaotic analysis***
  - nearly all users in the game for the first time;
  - Group's organization and association to sites finally defined and settled
- ***Important new functionality/requirements added and tested***
  - Inter CMS sharing tested (new priorityuser role);
  - New protocol for analysis results (and private production) placement and publication.
  - First "test" of Analysis

Operations in a real challenge.

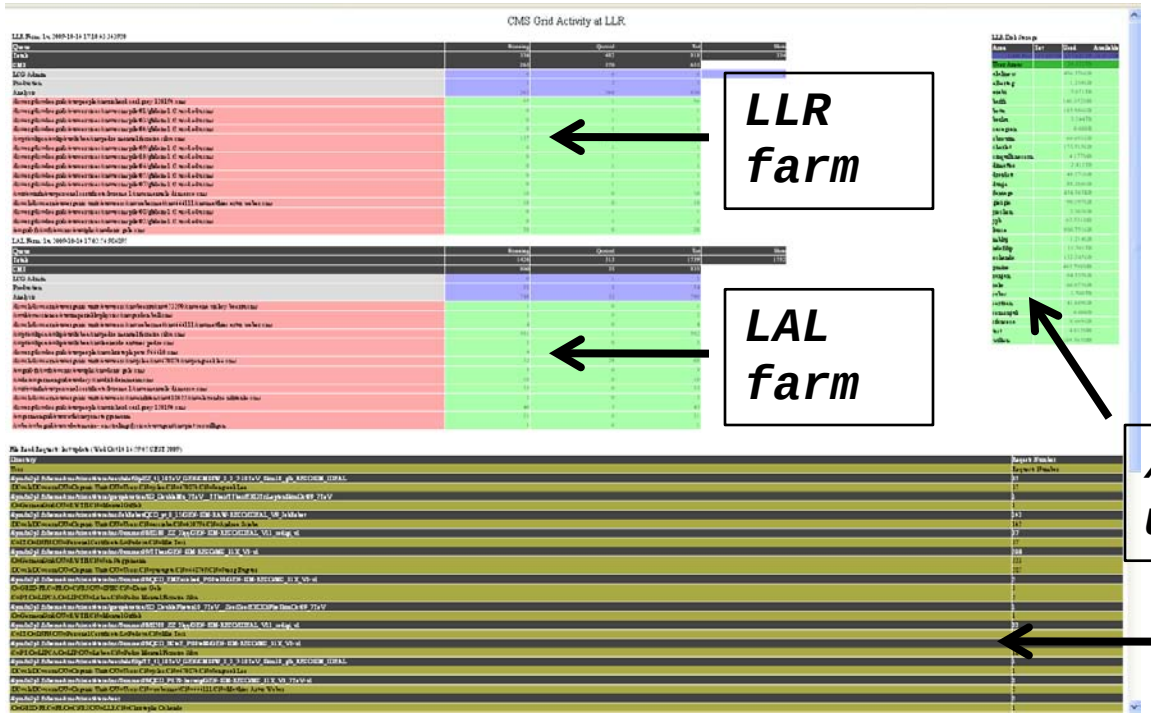
- ***So far…***
  - A lot of problems
  - 1 order of magnitude more users (not necessarily more load);
  - A lot of important feedback and information.



GRIF Tier 2

LLR CMS Mon

- Web page where user can ***monitor CMS activity*** at ***T2_FR_GRIF_LLR:***
  - Focused on local information not included in the CMS Mon infrastructure;
  - Useful both on admins and users side.



To Do

- ***More readable;***
- Add info on the ***/store/results /store/group***
- Integrate ***T3 info;***
- ***Push a bit users to look at it.***

GRIF Tier 3 Overview

- *Tier 3's are a **"private"** non-GRIF **fraction of the sub-sites***
  - ***CMS Physicists communities*** at LLR and IRFU decided to devote part of the T2 resources to support local activity;
  - Nominal fraction: LLR ~20% IRFU ~20%
- *Tools/services for fulfilling the **needs of local Physicists** in performing their analysis*
  - ***Interactive Usage***: UI's, Proof Clusters,etc;
  - ***Local batch cluster***;
  - ***Prioritized access*** to the T2 farm and resources;
  - ***Fast Access to T2 data***;
  - ***Management of user data***;
- ***Embedding into the GRIF** T2 environment*
  - Exploit as much as possible the services, the support and the configuration sharing of GRIF
  - Put up dedicated resources for all the services which cannot be embedded into the T2;
  - Clear deals on the sharing of computing and human resources between T2 and T3

GRIF Tier 3

IRFU

- ***Interactive Usage:***
  - The ***T2/T3 site has 1 UI*** at disposal of the local analysts;
  - ***Tar.gz UI*** version up to date (rsync) with the CMS sw;
  - ***UIs cluster*** configured for ***parallel interactive usage***:
    - Frontal machine distributing applications;
    - ***22*** 8-cores ***node-machines*** (16GB ram);
    - Dedicated ***GPFS storage*** (5TB). ***No dpm rfio*** access;
    - ***Proof cluster*** (tested and showed ***factor 10 improved performance*** in analysis tasks);
- ***Prioritized access to the T2 farm resources***
  - Currently ***~20% fraction*** of resources (fairshare implementation);
  - ***Matched certificates*** with 'IRFU' in the DN and mapped to higher priority user(s);
- ***Storage management for the user data***
  - Need to ***implement quotas*** for the local user into DPM storage: ***space tokens*** (not used by CMS at the moment);

GRIF Tier 3

LLR…

- ***Interactive Usage:***
  - ***3 UIs*** (2 SL4 + 1 SL5): ***single login***, ***shared homes*** (quotas enabled), ***sw area*** mounted, NFS ***data area*** mounted
  - CERN ***Virtual UI***: under testing by some users;

  To Do
  - ✘ Cluster-ize the Uis to assure load balancing;
  - ✘ Proof cluster (still thinking and waiting for feedback by other's experience);

- ***Local Batch System:***
  - Made for ***short turnaround testing of full jobs*** in a grid-like controlled environment;
  - Dedicated ***torque/maui scheduler*** (Not a grid CE);
  - Pool of ***UI nodes***: share the ***same users, homes and data area*** of the interactive UIs;
  - At the moment: 2 nodes (***2 sl4 slots and 8 sl5 slots***). ***Easy to increase***;

GRIF Tier 3 …LLR…

- ***…Local Batch System:***
  - Simple ***batch submission from*** the interactive ***UIs***;
  - ***Login on the worker nodes*** is possible with the same accounts as interactive UIs;
  - First usage by some “test user” during the Oct’09 Exe;

  To Do
  - ✘ Post-mortem of Oct ‘09 Exe. Collect user feedback: add new functionalities and increase functionality;
  - ✘ Would like to use crab submission: development needed;

- ***Prioritized access to T2 farm:***
  - ***Currently ~20% of the T2 resources*** (as fairshare fraction);
  - ***Prioritized queue*** for the VO ***vo.llr.in2p3.fr*** (LLR ***local users***);
  - New ***/cms/frcms VOMS group*** (admin: C.Charlot), which will be mapped into ***high priority user*** in order to guarantee fast access to CMS ***France community;***

GRIF Tier 3 …LLR…

- ***…Prioritized access to T2 farm:***
  - ***Submission*** with the prioritized groups/VO's is ***straightforward*** with the standard ***CMS tools*** (i.e. CRAB).

To Do

- ✘ Fix few ***technical problems with frcms*** at GRIF configuration level;
- ✘ ***Fairshare*** still need ***to be tuned***. In particular wrt other prioritized cms roles like lcg-admin, production, priorityuser, etc.

- ***Storage Access/Management:***
  - All T3 devices can ***access to all data on DPM*** T2 storage;
  - ***Locally mounted data area*** on UI and on the T3 local cluster;
  - CMS ***FileMover***;

GRIF Tier 3 …LLR

- ***…Storage Access/Management:***

To Do

- ***Backup on tape at Lyon*** of the ***user-produced data*** on DPM (these are not replicated): still checking feasability;
- Local ***CMS Data Bookkeeping Service:*** needed for locally publish and share user produced data;
- Study a way to guarantee the local users a ***prioritized access to DPM data.***

Summary

- ***Longstanding and fruitful experience*** at GRIF in supporting CMS Tier 2 activity:
  - Management of ***CMS specific services***: PhEDEx, squid/frontier;
  - Supporting ***analysis and production activity*** and data management:
    - ***High reliability***;
    - ***Important and stable activity***, also pushed by very active local communities;
  - Ongoing ***integration of the multi-site setup*** into the CMS framework:
    - 2+2 site config works fine;
    - Still some inconsistencies: downtimes accounting, etc…;
  - Important ***feedback from Oct'09 Exe***:
    - Good crash test;
    - Final setup of some functionalities in sight of Data Taking;
- Process of ***deployment and exploiting of Tier 3*** resources has ***started***:
  - LLR and IRFU already ***setup the big part of T3 services***
  - Already ***tested and proven to be useful*** during Oct'09 Exe
    - ***Waiting for feedback*** to improve resources and add new functionalities;