



# IN2P3 Lyon Tier-1 : Avancement et Perspectives

**Fabio Hernandez / Pierre Girard**

fabio@in2p3.fr / pierre.girard@in2p3.fr

LCG France 2010

Marseille, June 24th-25th, 2010



# ▶ Introduction



- Transition with Fabio started early in May 2010
  - I was unfortunately out of service before
  - But, in few time, Fabio provided me with a lot of (precious) information
  - Time needed to assimilate
  - Expect to be operationnal by the automn 2010



# ▶ Contents



- CCIN2P3 in the context of WLCG
- Site overview
- Main activities
  - Data exchange
  - Data storage
  - On-site data processing
- WLCG service metrics
- Current concerns
- Perspectives
- Conclusions
- Questions & Comments

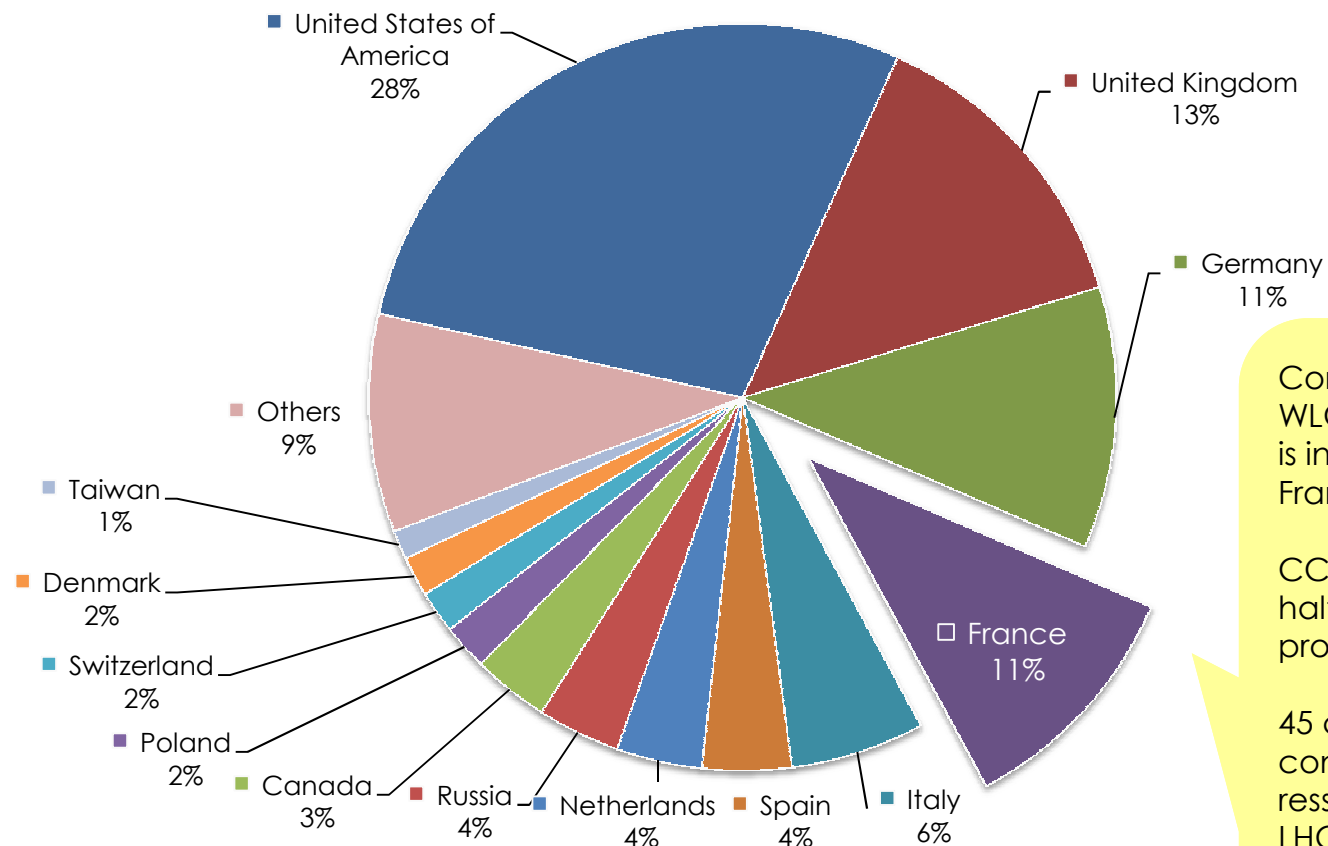
# **CCIN2P3 in the context of WLCG**

# ▶ Contribution per country



## WLCG -- CPU contribution per country

Normalised CPU time (HEP-SPEC06)  
All LHC experiments -- Jan 2009 - May 2010



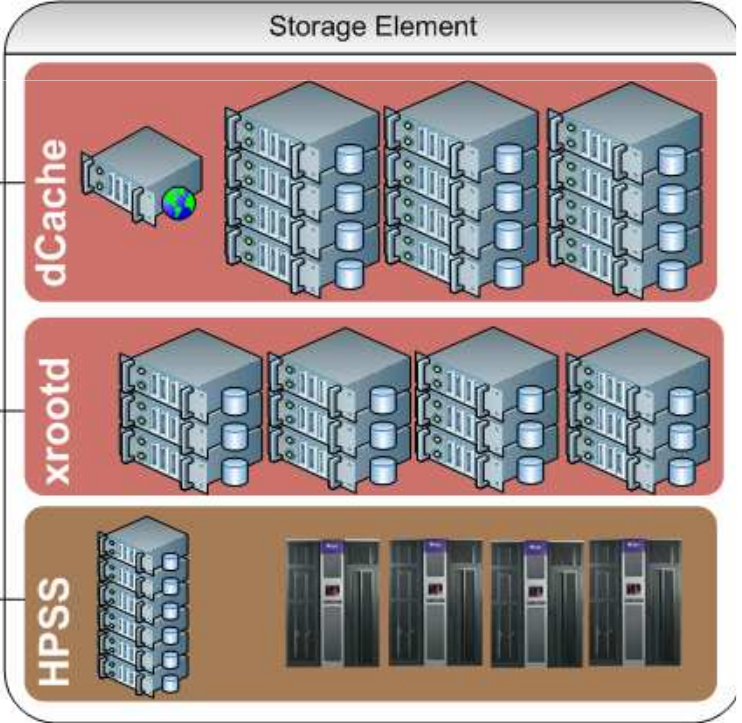
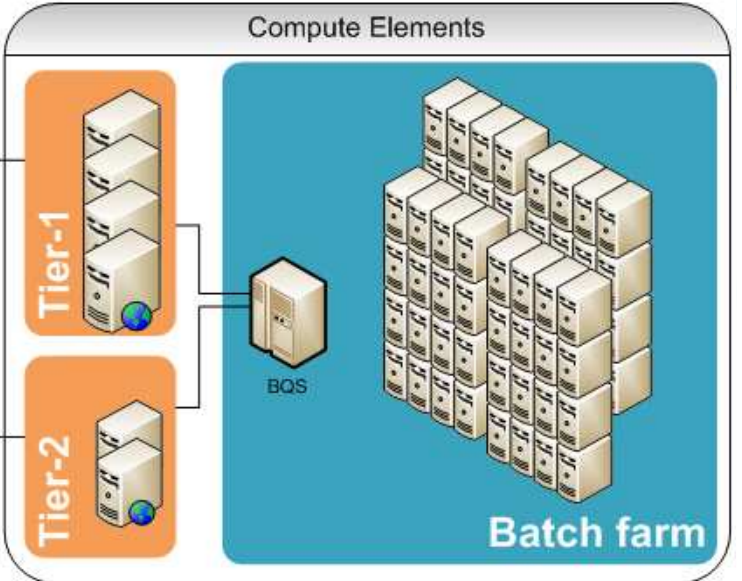
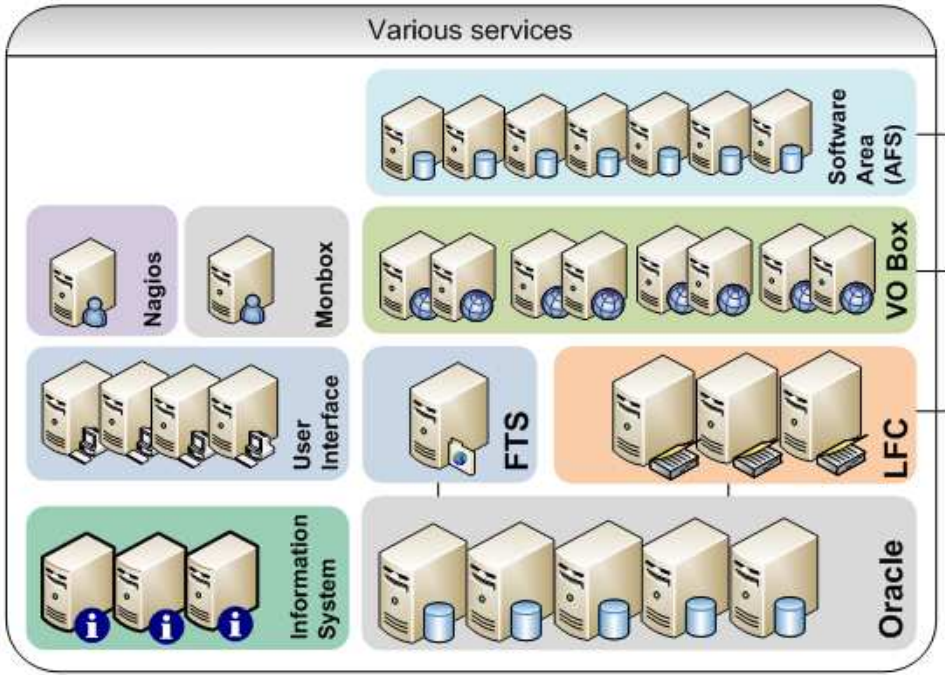
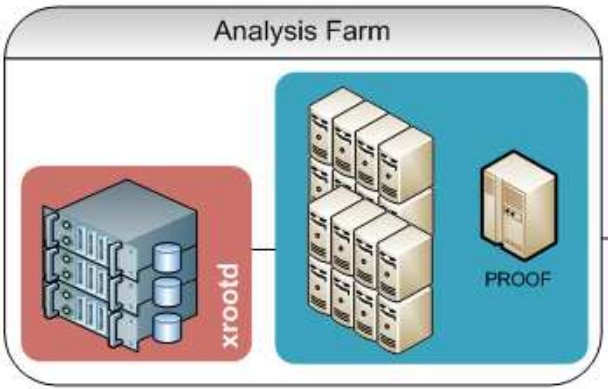
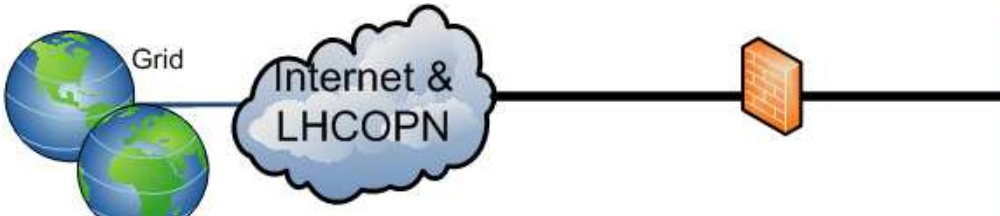
Contribution to the WLCG collaboration is in line with LCG-France's target.

CCIN2P3 contributes half of the CPU provided by France

45 countries contributed CPU resources to the 4 LHC experiments in this period

Source: EGEE Accounting Portal  
[https://www3.egee.cesga.es/gridsite/accounting/CESGA/country\\_view.html](https://www3.egee.cesga.es/gridsite/accounting/CESGA/country_view.html)

# Site overview



Author: Fabio Hernandez  
Last Updated: 2009-06-09



# ▶ LCG-France budget cut effects



- In 2010, we faced a cut of 40% in the equipment budget
  - Resulted in a reduction of the pledged capacity from 10% to 20%, in accordance with each experiment priorities
- The equipment purchased in 2009 contributed to minimize the negative impact of this year cut in the experiments activities
  - We foresee for 2011 to come back to the requested level of equipment budget
  - The growth plan of the site and therefore the future pledges to WLCG are based on this assumption



# Data Exchange



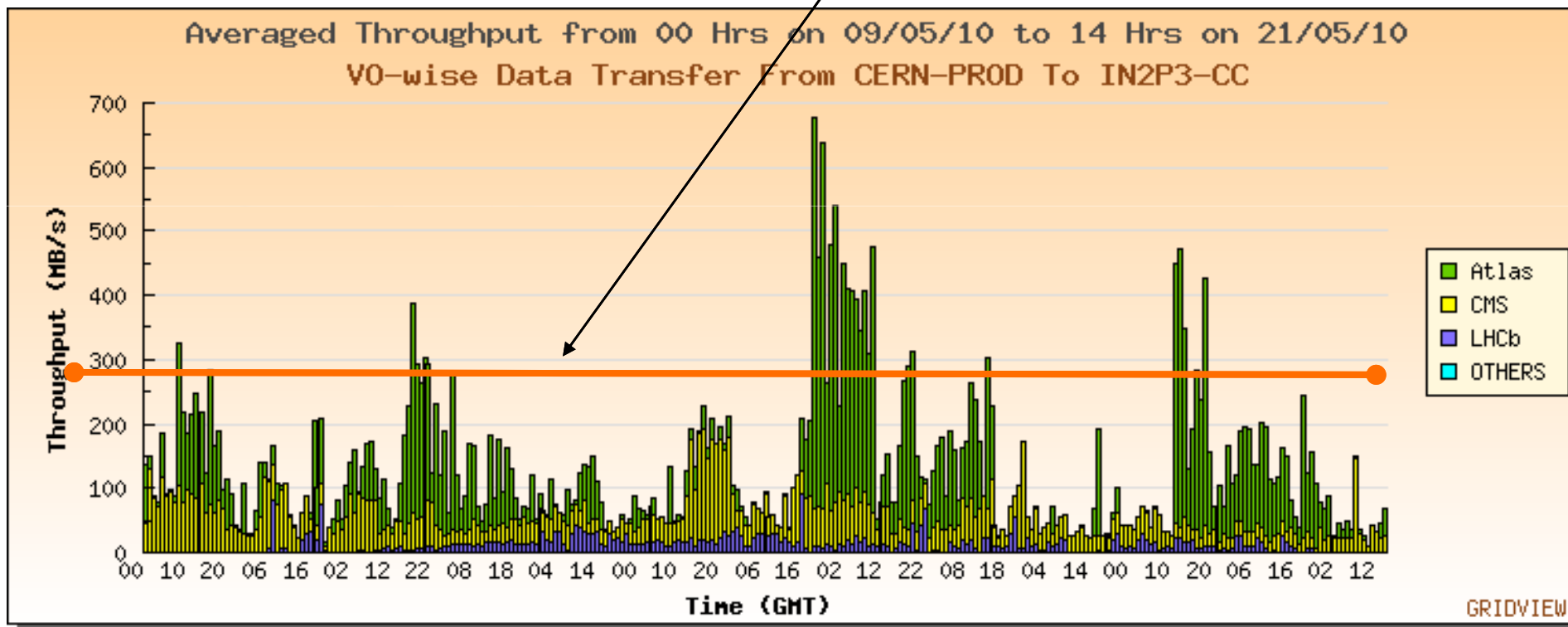
# Data import

(collisions data-taking period)

Nominal target: ~270 MB/s

- ALICE: ~10 MB/s
- ATLAS: ~100 MB/s
- CMS: ~150 MB/s
- LHCb: ~10 MB/s

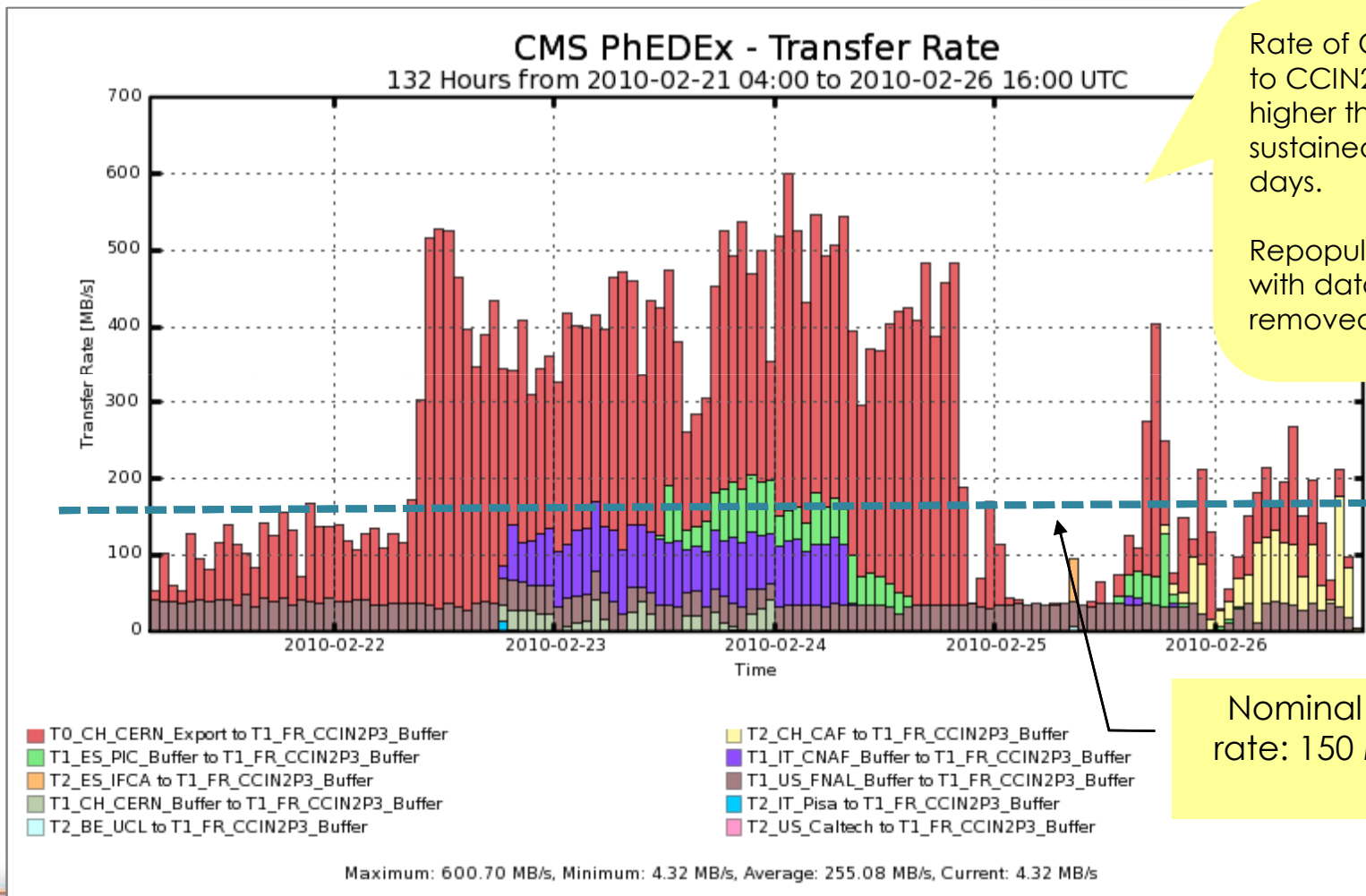
- Tier-0 → CCIN2P3



May 9-21, 2010

Source: Gridview <http://gridview.cern.ch>

# ▶ Data import (cont.)



Rate of CMS data import to CCIN2P3 significantly higher than nominal rate, sustained over several days.

Repopulation of the site with data accidentally removed in Nov. 2009

Nominal target rate: 150 MB/sec

Source: CMS PhEDEx  
<http://cmsweb.cern.ch/phedex>

February 21-26, 2010

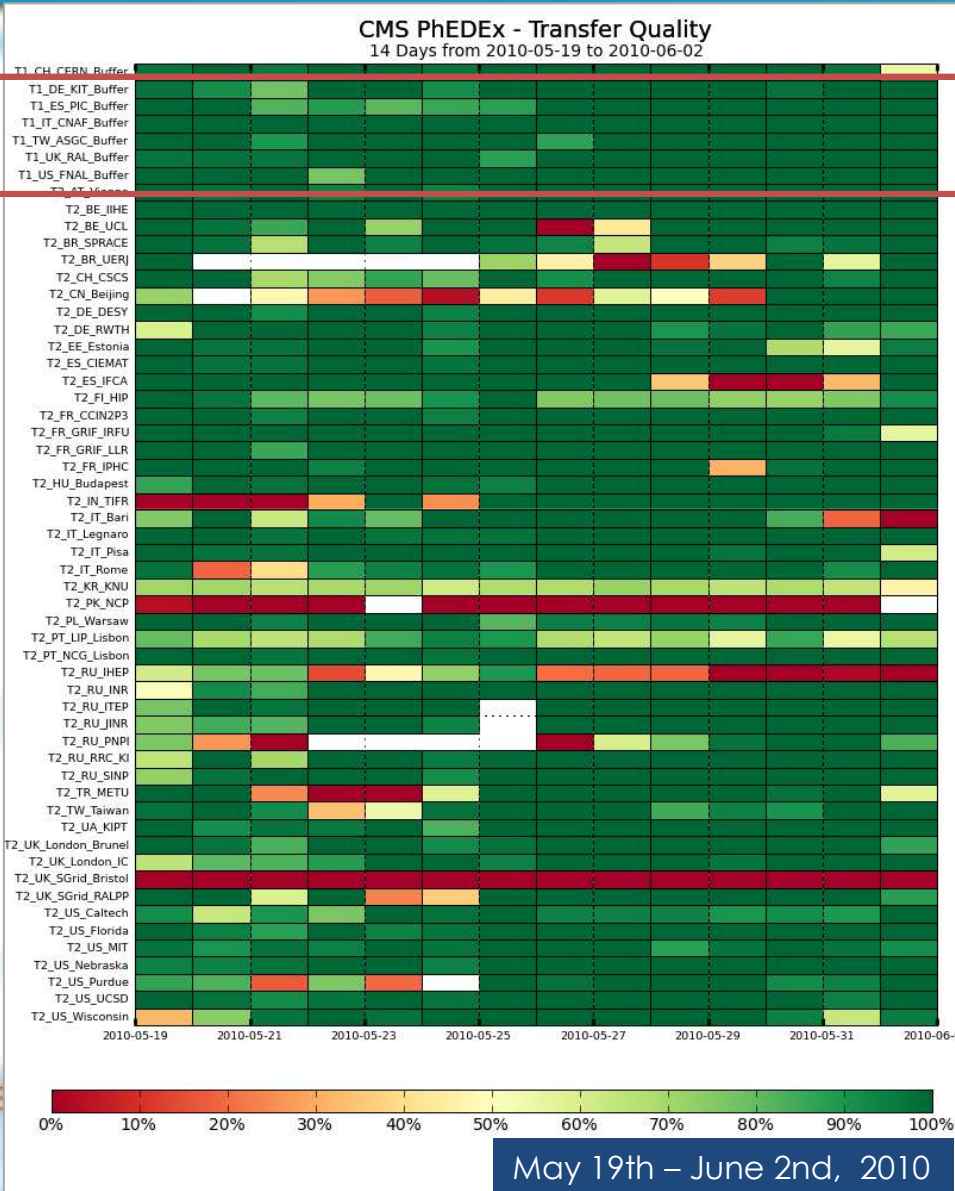


# Data export

(the power and the complexity of the grid)



CMS tier-1s



Transfer quality for CMS data export from CCIN2P3 to other sites, as measured by the experiment.

CCIN2P3 exchange data with **50+ sites** all over the world.

The quality of every single channel is routinely monitored and human interventions by site experts are triggered when needed.

Source: CMS PhEDEx  
<http://cmsweb.cern.ch/phedex>

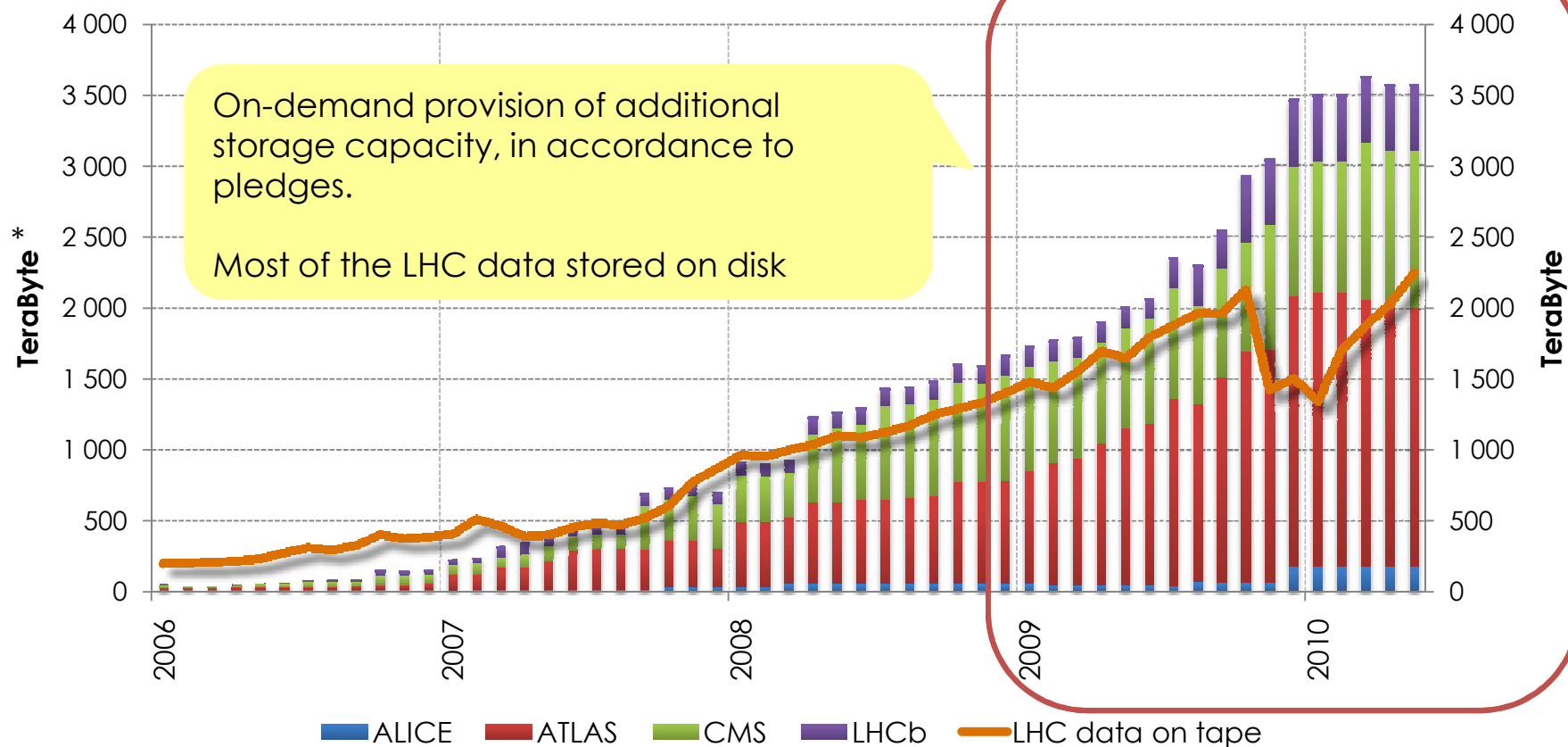
# Data Storage



# Data storage: disk & tape



## Evolution of disk allocation and tape usage All LHC experiments



\* 1 TeraByte =  $10^{12}$  bytes

# ▶ Storage service



- Significant efforts in 2009 on those components
  - Upgrade of HPSS deployed early June 2009
  - Introduction in production of TReqS in July 2009
  - Replacement of the dCache meta-data engine (a.k.a. Chimera) in September 2009 and version upgrade early 2010
- As a consequence, major improvements in the stability, performance and manageability of the whole chain were observed
  - Even if we still observe some glitches
- Regular campaigns of consistency checks of experiment-specific file catalogues and storage system's catalogues

## ▶ Storage service (cont.)



- Reconfiguration of HPSS for devoting some resources per experiment and to minimize interference among them
  - Introduction of 1 TB cartridges for large files and faster 120 GB cartridges for small files
- But, very low tape recall activity for LHC data since data taking started
  - The tape recall system has not yet been operated in real conditions



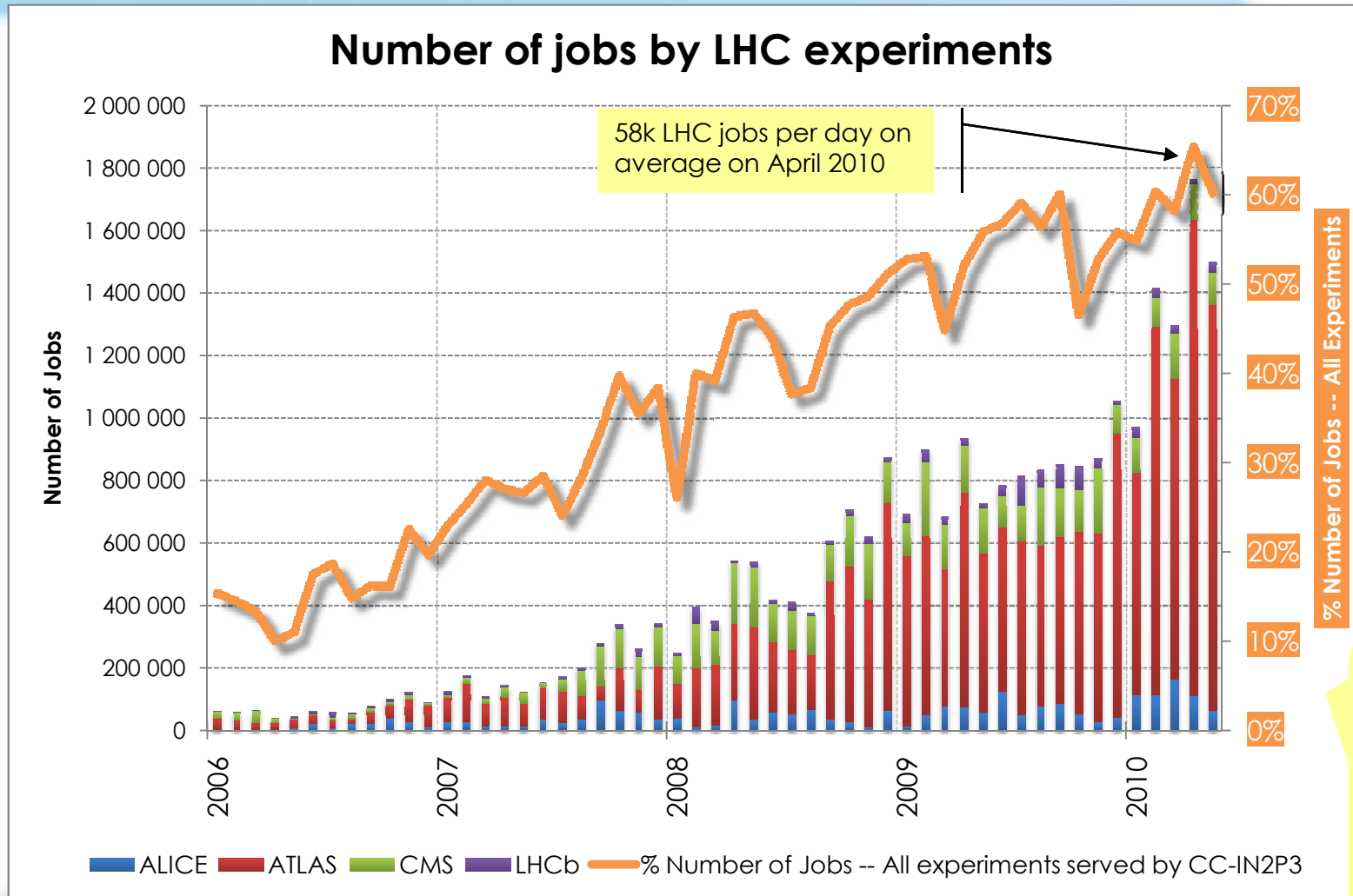
# On-site LHC data processing



# Batch workload: tier-1 & tier-2



## Number of jobs by LHC experiments



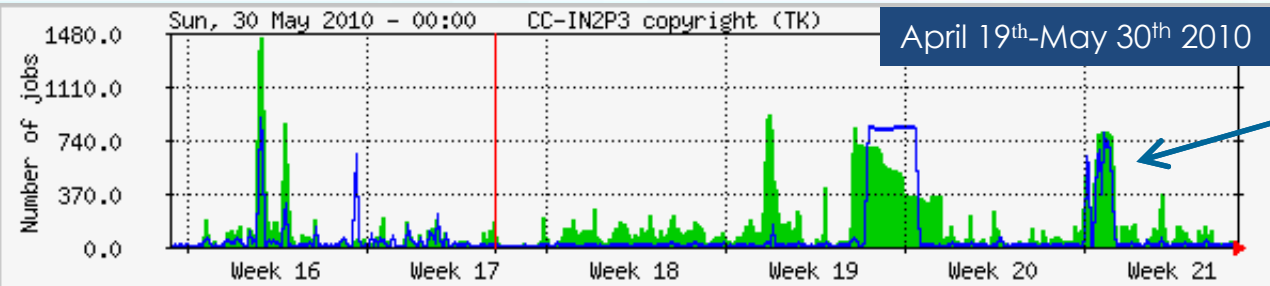
Sustained increase over the last 4+ years in both the job throughput and the share of the LHC experiments



# Batch workload: tier-1 & tier-2 (collisions data-taking period)

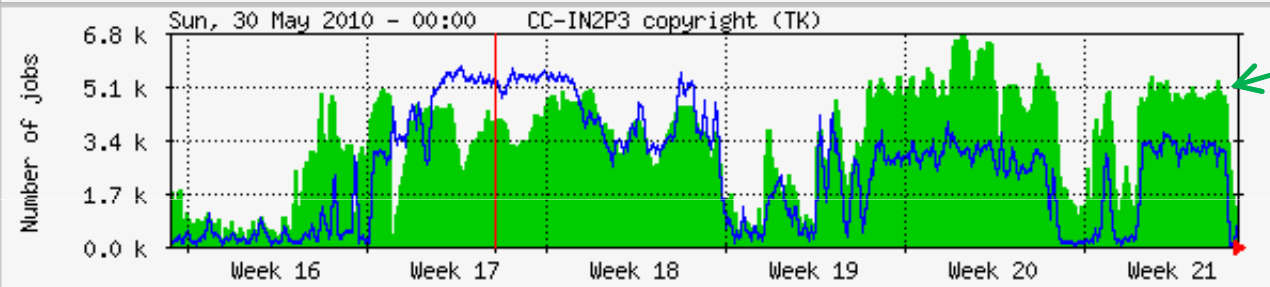


ALICE



Waiting jobs

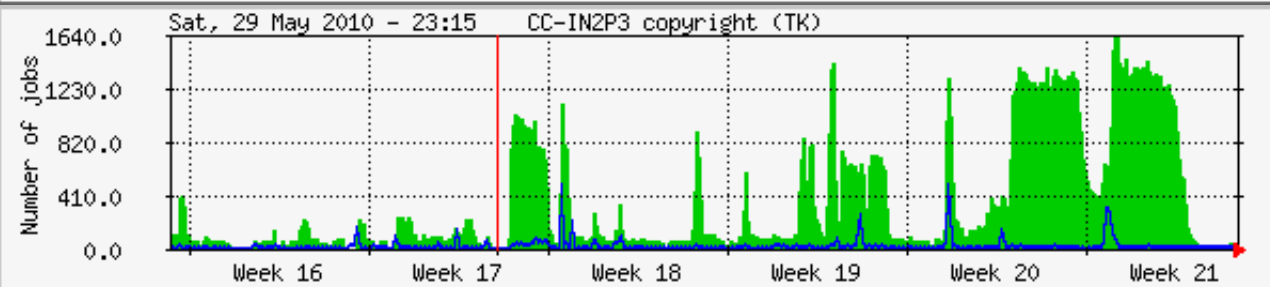
ATLAS



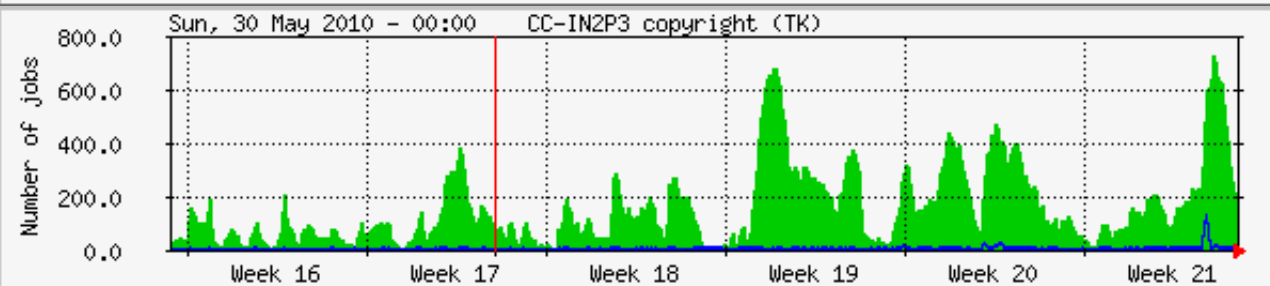
Running jobs

Average running jobs  
ALICE: 133  
ATLAS: 3318  
CMS: 333  
LHCb: 152

CMS



LHCb



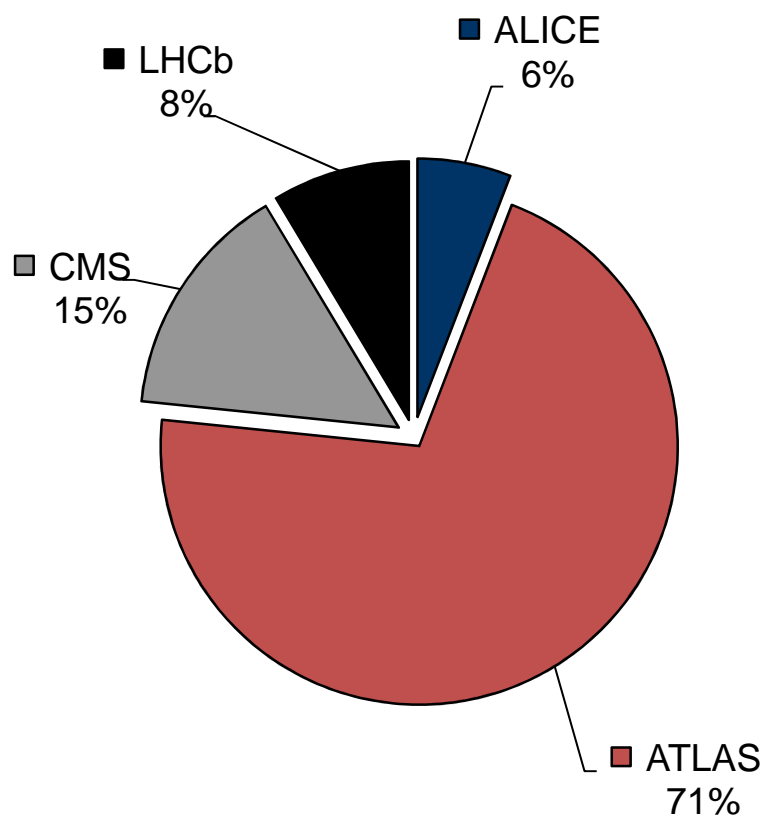
NOTE: Y axis scale is not the same on all plots

# CPU utilization



## CPU consumption by LHC experiments at CCIN2P3

Jan-May 2010



Utilisation by LHC experiments of the aggregated CPU capacity provided by CCIN2P3 (tier-1 and tier-2)

To be compared with the utilization in the same period in all WLCG sites:

ALICE	7%
ATLAS	70%
CMS	20%
LHCb	3%

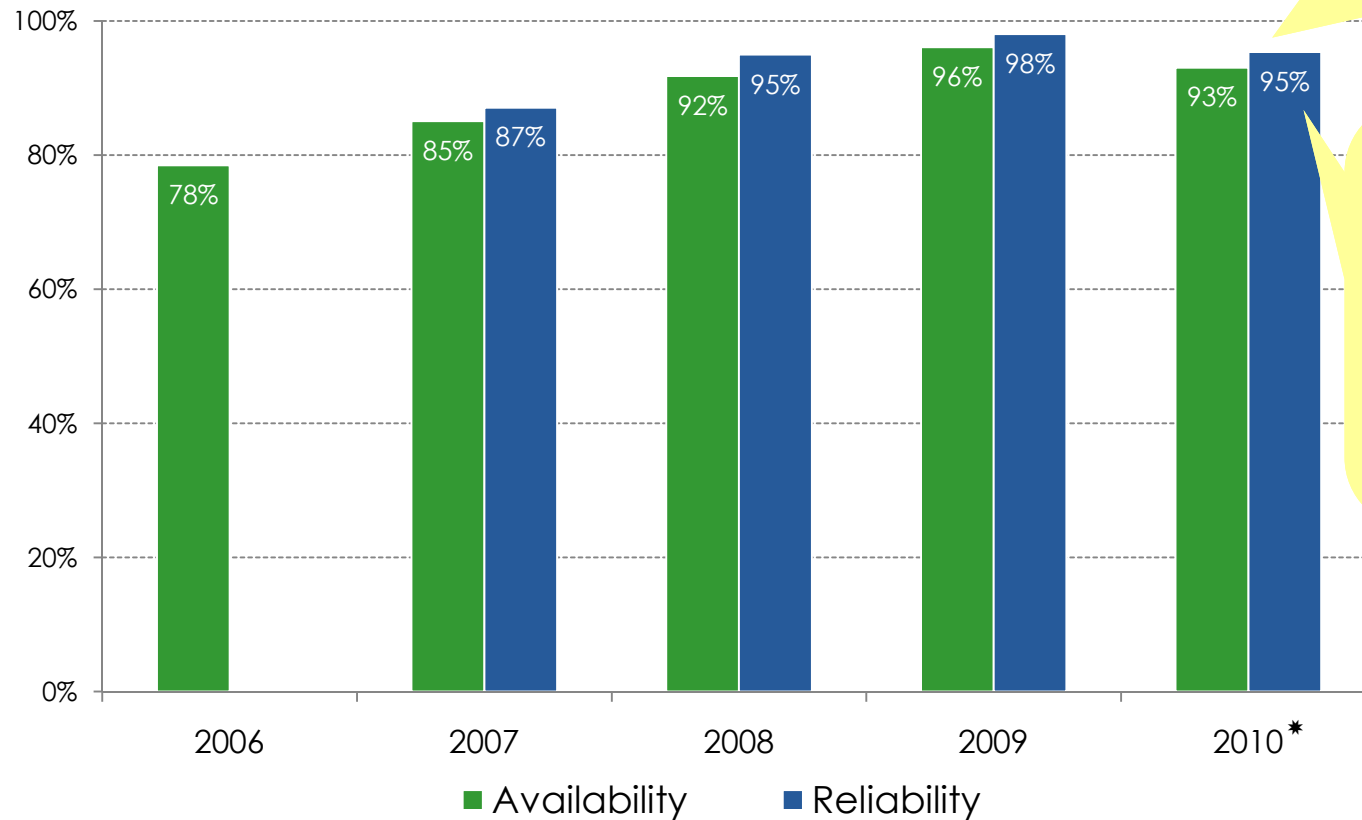
# **Service targets according to WLCG MoU**

# ▶ MoU targets: reliability

Room for improvement on the reliability of the site. We aim to come back to the levels observed in 2009.

10 incidents of various levels of severity since the beginning of 2010

**LCG-France tier-1: yearly evolution of availability and reliability (VO OPS)**



MoU target is a reliability of 98% of the time, integrated over the year.

95% reliability in 2010 means 133 hours during which at least one component was not fully functional

\* January-April 2010

Source: WLCG Reliability Report  
<http://lcg.web.cern.ch/LCG/reliability.htm>



# Alarm tickets

[01/06/2009 – 01/06/2010]



Experiment	Real	Test
ALICE		3
ATLAS	1*	2
CMS		2
LHCb		3

Response to the only real alarm ticket within the agreed limits according to the WLCG Memorandum of Understanding

Submitted: 2009-06-10 07h43  
Acknowledged by human: 2009-06-10 08h04  
Solved: 2009-06-10 10h09  
\* Ticket: [https://gus.fzk.de/ws/ticket\\_info.php?ticket=49392](https://gus.fzk.de/ws/ticket_info.php?ticket=49392)

Source: GGUS <https://gus.fzk.de>

# Concerns, Perspectives & Conclusions



# ▶ Current Concerns



- Connectivity
  - Exchange between LAPP and CC
    - Both in terms of bandwidth and reliability
  - Transfer exchanges with some sites
    - 4MB/s instead of ~35MB/s with BNL, but also with FNAL, TRIUMF, and PIC (10 MB/s)
    - Under investigation by both network team and storage team
      - Could be a problem of TCP stack implementation between Solaris and Scientific Linux
- Staff
  - End of EGEE
  - People moving
- Monitoring
  - Transition EGEE/EGI: Operational tools changed a lot over the last months
  - Local monitoring has shown to be not sufficient
  - Monitoring histograms are needed both by site and VOs to better understand how the infrastructure is used

# ▶ Perspectives



- Data storage
  - Improve the mechanisms for preventing unauthorized users to put excessive load on the mass storage system by chaotically recalling tape data
  - Introduce a mechanism for protecting the data against unintended removal by authenticated users
- Computing
  - Introduce GridEngine in production in a progressive and as transparent as possible way
- Service
  - Improve the reliability of the site
  - Finalize and exploit the platform that should allow for finer monitoring and analytics of LHC data processing activities
- End-user analysis
  - Promote the use and improve the usability of the national interactive analysis farm

# ▶ Conclusions



- The improvements implemented in the the last 12 months has led to a more stable and manageable site
  - In particular, regarding the storage service, event if there are still several areas that need improvement
- Data distribution and most data processing activities on the grid platform are understood and have been routinely exercised
  - Understanding the needs of the end-user analysis activity is the next problem to tackle
- The site is in good shape and able to face the ramp up of the LHC
- The contribution of LCG-France and CCIN2P3 to the LHC data processing activity should directly benefit IN2P3 and Irfu physicists

# ▶ Acknowledgements



- Fabio Hernandez (obviously)
- Several people provided direct input for this talk:
  - ALICE@CC-IN2P3: R. Vernet
  - ATLAS@CC-IN2P3: G.Rahal, C.Biscarat, E.Cogneras
  - CMS@CC-IN2P3: F.Fassi, D.Mercier
  - LHCb@CC-IN2P3: L.Arrabito
  - Batch service: S.Poulat
  - Storage services: B.Delaunay, A.Gomez, Y.Calas, L.Schwarz
  - Middleware deployment: P. Girard
  - LFC & FTS: D. Bouvet
  - Databases: P.E.Macchi
  - Network: J.Bernier



# Questions & Comments



# Backup slides

# ▶ Storage service



- Storage chain composed of 3 main components
  - dCache: disk-based system exposing gridified interfaces
    - *One instance serving the 4 LHC experiments*
  - HPSS: tape-based mass storage system, used as permanent storage back-end for dCache
    - *One instance for all the experiments served by CCIN2P3*
  - TReqS: mediator between dCache and HPSS for scheduling access to tapes, for optimization purposes
    - *To minimize tape mounts/dismounts and to optimize the sequence of files read within a single tape*

# ▶ Cross-experiment activities



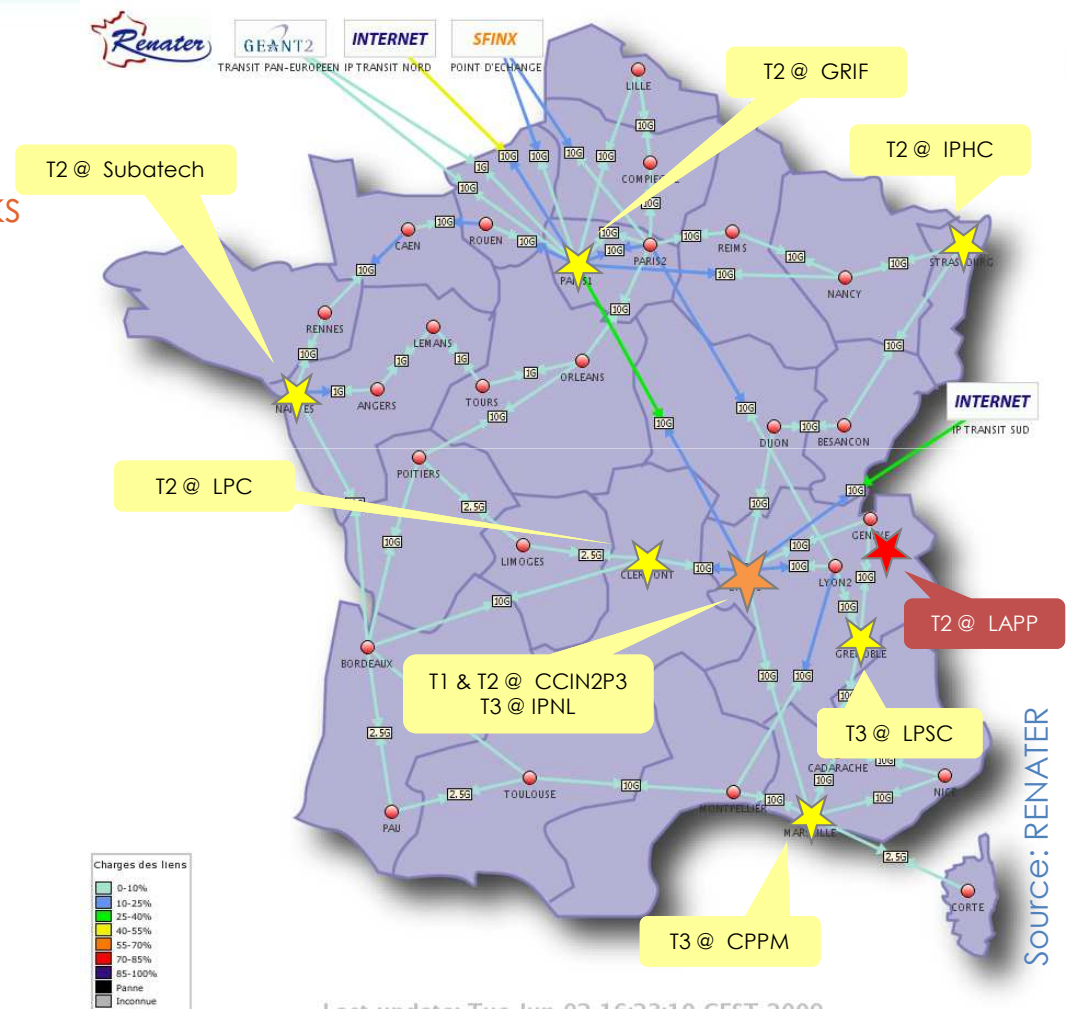
- Inter-site database replication
  - CERN → CCIN2P3
    - ATLAS: replication of conditions data
    - LHCb: replication of file catalogue data base (LFC) and conditions data
  - CCIN2P3 → CERN
    - ATLAS: CCIN2P3 provides the high-availability hardware and software infrastructure for the ATLAS central meta-data catalogue (AMI), developed and operated by LPSC Grenoble
      - Replication of the backend database to CERN for availability purposes
- Usual updates of the software stack and introduction of new middleware components
  - ScientificLinux v5, CREAM CEs, information system, LFC, FTS, VO boxes, gLExec & ARGUS, ...
- Interactive analysis farm
  - See Ghita's talk



# ▶ CCIN2P3 connectivity



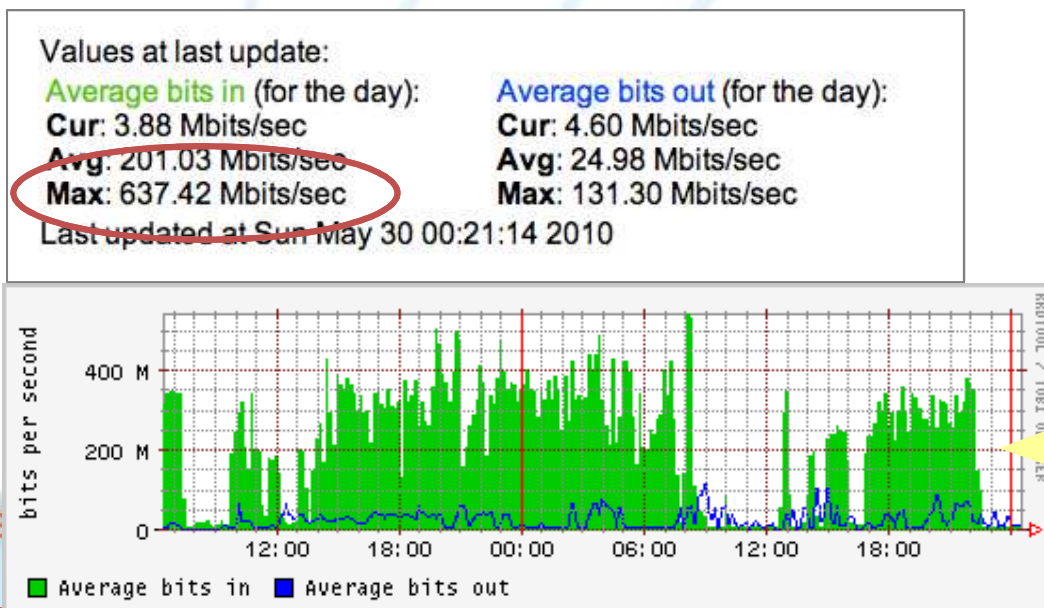
- Provided by RENATER & GEANT
- To tier-0 and tier-1s
  - Dedicated 10Gbps LHCO PN links
    - CCIN2P3 ↔ CERN
    - CCIN2P3 ↔ KIT ↔ CERN
- To foreign tier-2s and tier-3s
  - Connected to GEANT routers at 10Gbps
- To domestic tier-2s and tier-3s
  - All sites but one located near a RENATER point of presence
  - Direct connections to RENATER equipment or sharing a 10 Gbps link with other academic/research institutions in the same metropolitan/regional network



# ▶ CCIN2P3 connectivity (cont.)



- The lack of adequate connectivity (both in terms of bandwidth and reliability) to IN2P3-LAPP is a cause of concern
  - The available bandwidth may quickly become the limiting factor as the LHC ramps up and more data need to be processed
  - This issue is currently being followed by the direction of RENATER

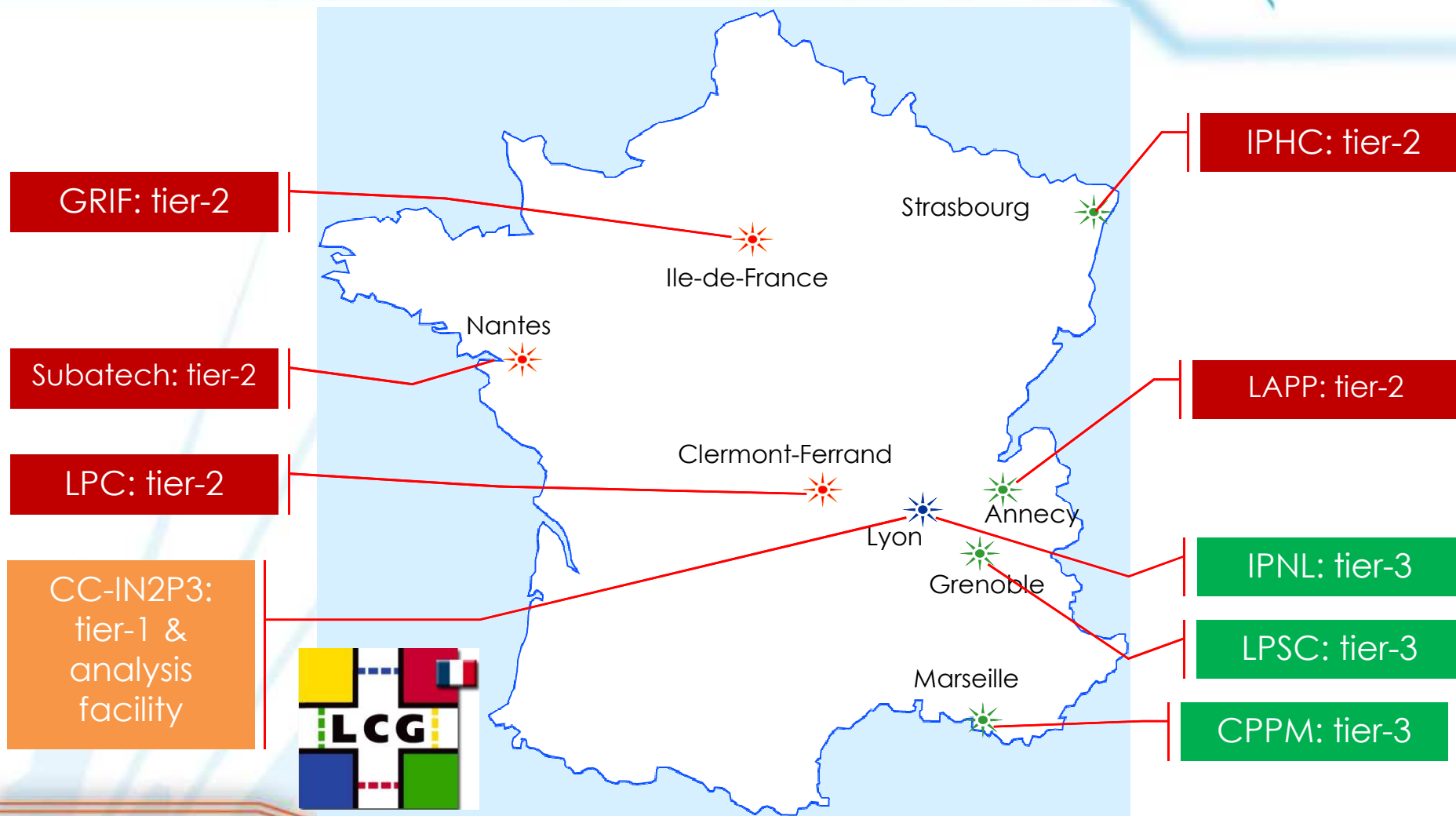


Averaged bandwidth used for data exchange between CC-IN2P3 and IN2P3-LAPP.

The available bandwidth of this link is 1000 Gbps

Source: <http://netstat.in2p3.fr/weathermap/graphiques/ann-nrd.html>

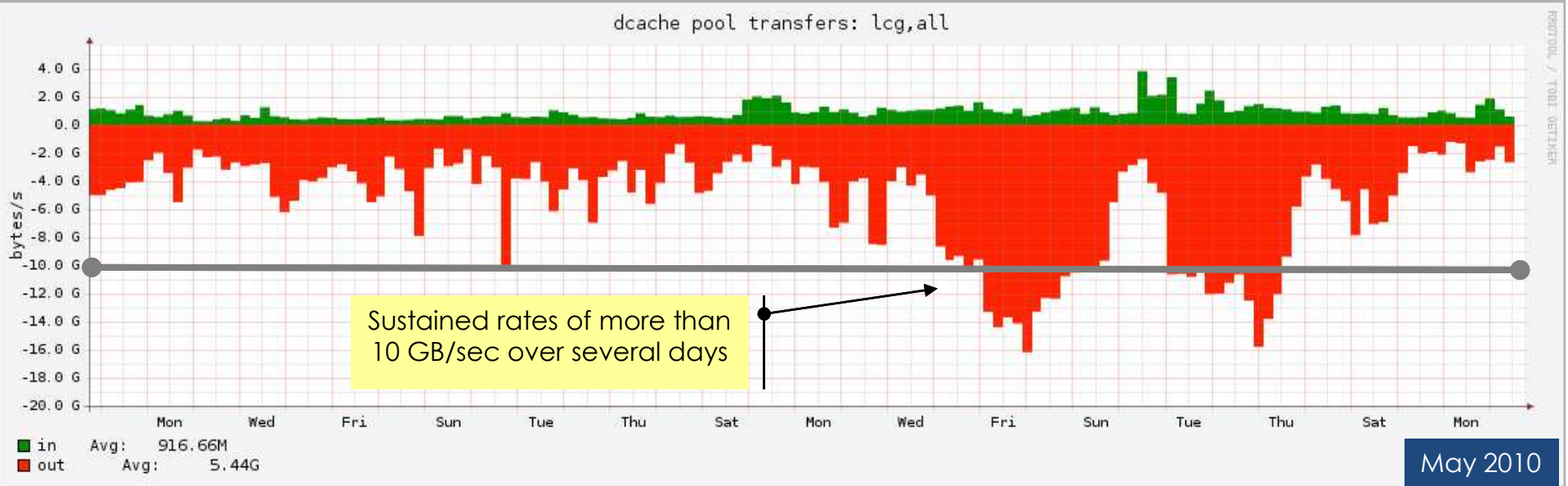
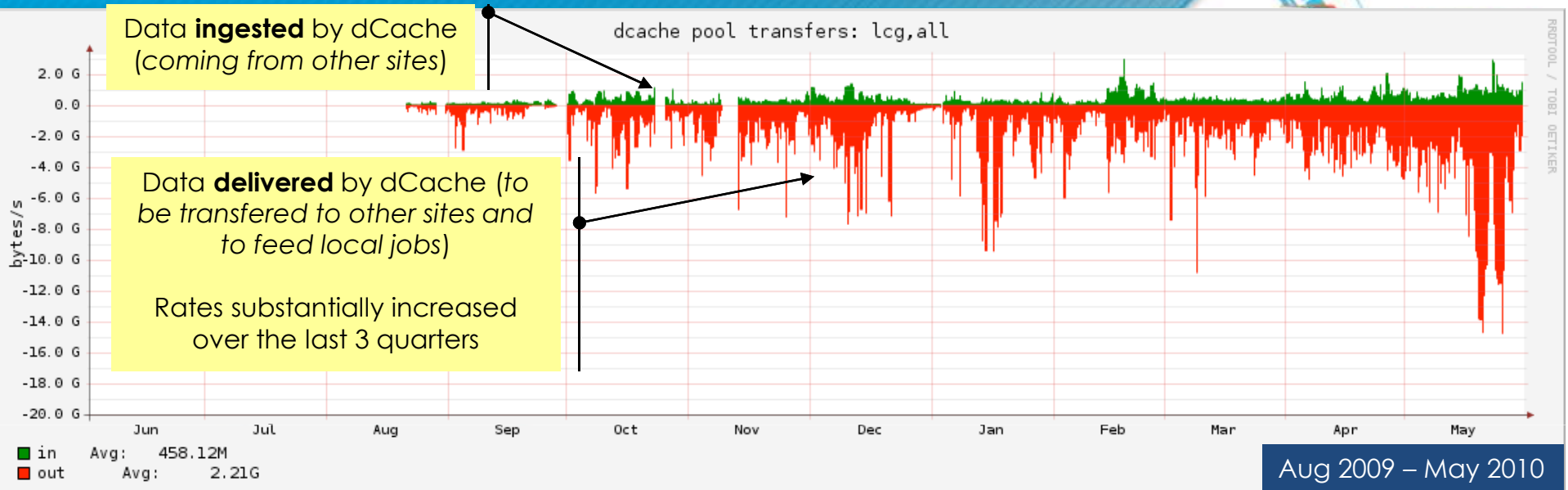
# ▶ LCG-France



Source: <http://lcg.in2p3.fr>



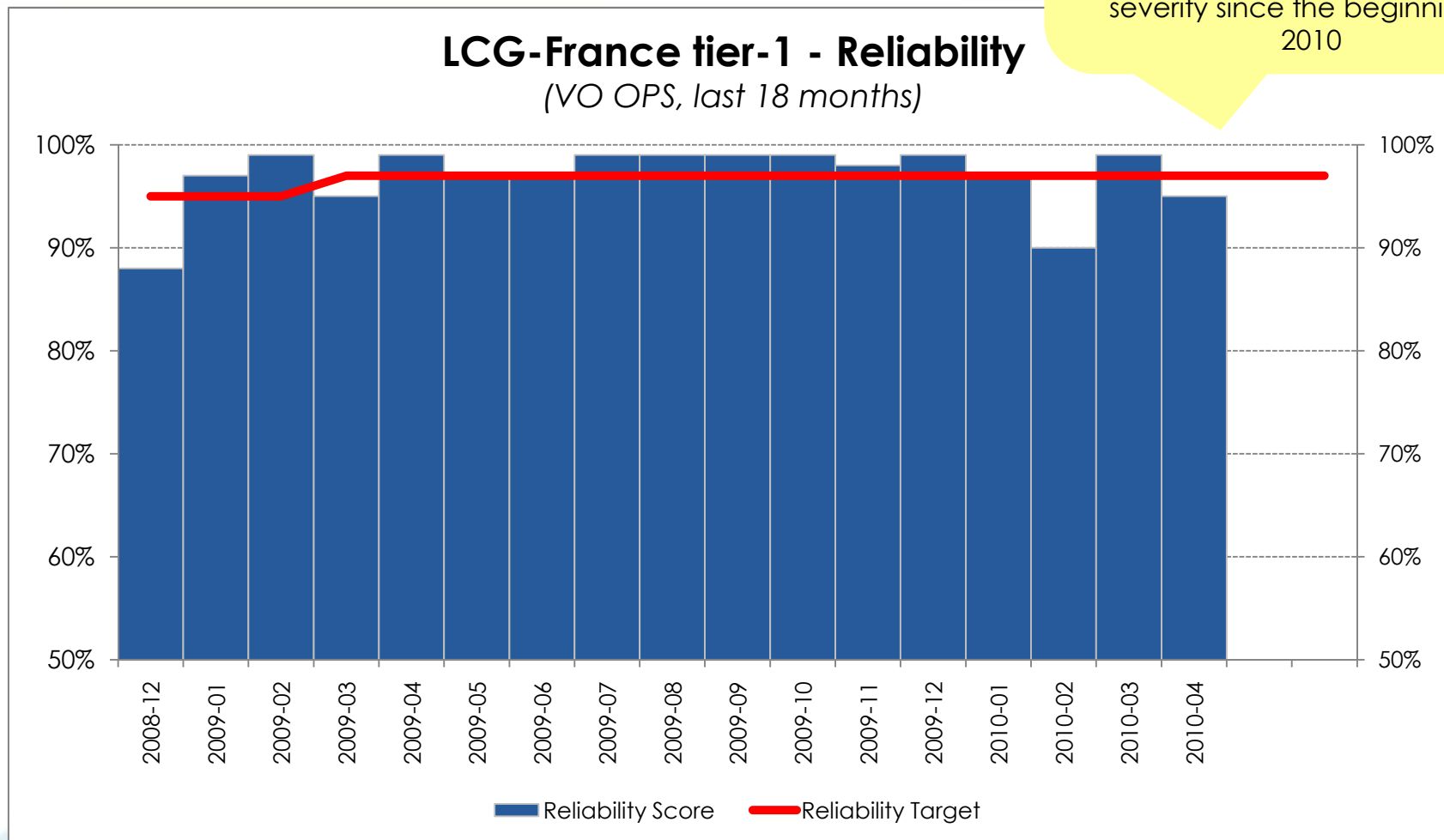
# dCache: serving local jobs



# ▶ MoU targets: reliability

Room for improvement on the reliability of the site. We aim to come back to the levels observed in 2009.

10 incidents of various levels of severity since the beginning of 2010

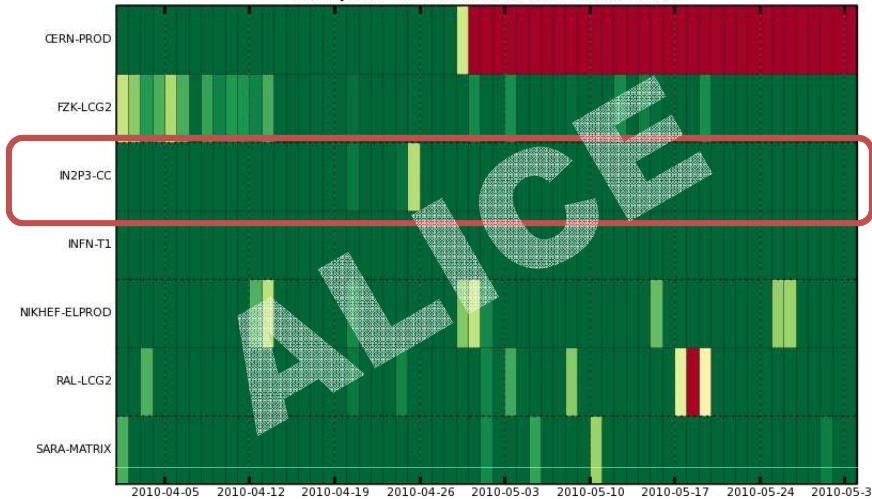


Source: WLCG Reliability Reports  
<http://lcg.web.cern.ch/LCG/reliability.htm>

# MoU targets: availability tier-1s (collisions data-taking period)

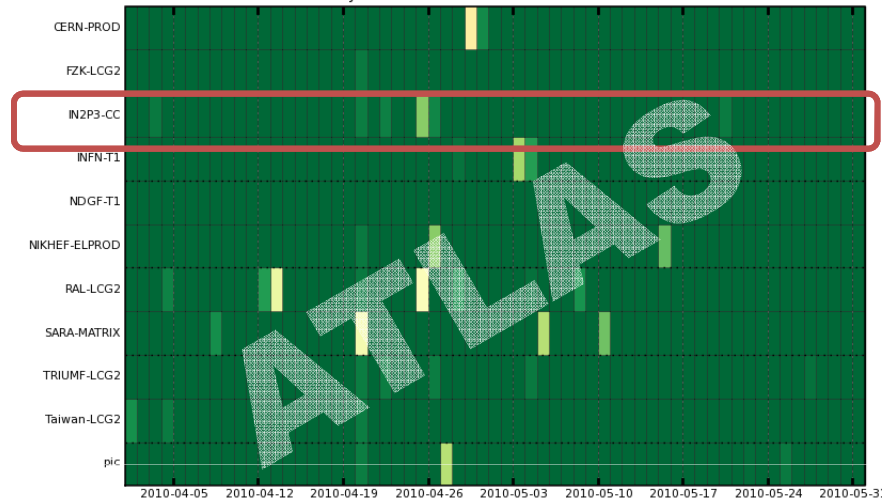
Site Availability using WLCG Availability (FCR critical)

61 Days from Week 13 of 2010 to Week 22 of 2010



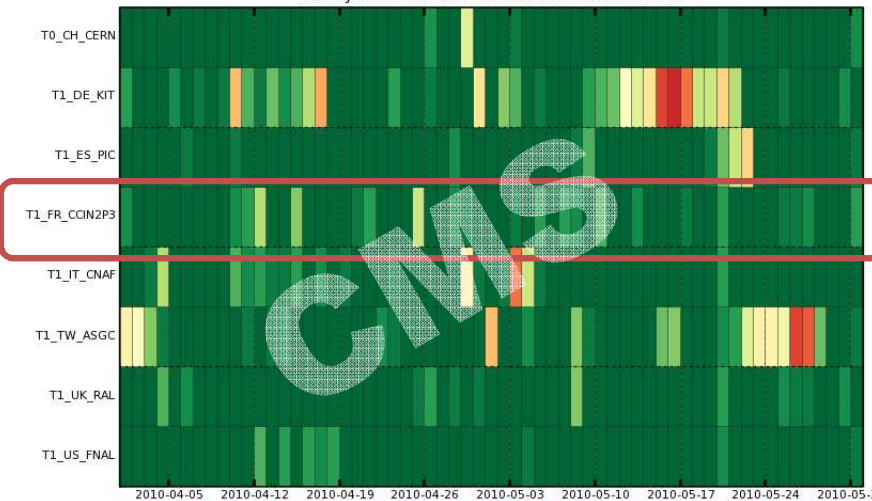
Site Availability using WLCG Availability (FCR critical)

61 Days from Week 13 of 2010 to Week 22 of 2010



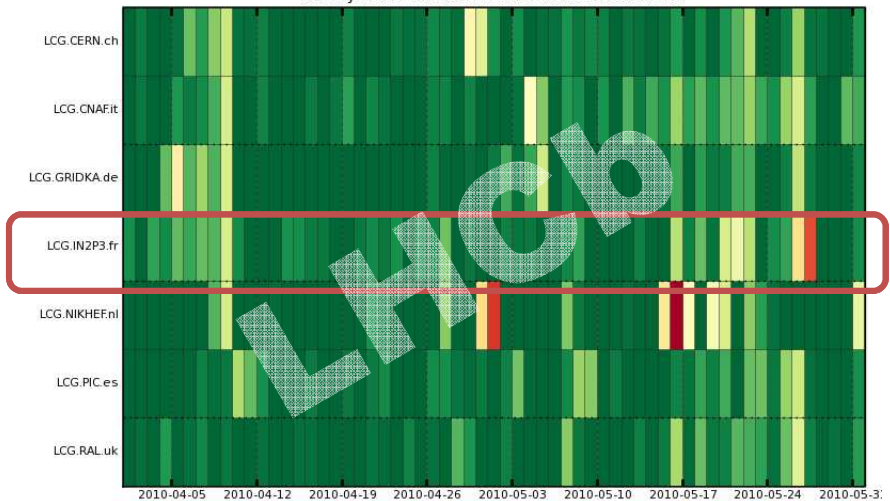
Site Availability

61 Days from Week 13 of 2010 to Week 22 of 2010



Site Availability using LHCb Critical Availability

61 Days from Week 13 of 2010 to Week 22 of 2010



Period: Apr 1<sup>st</sup> – Jun 1<sup>st</sup>, 2010

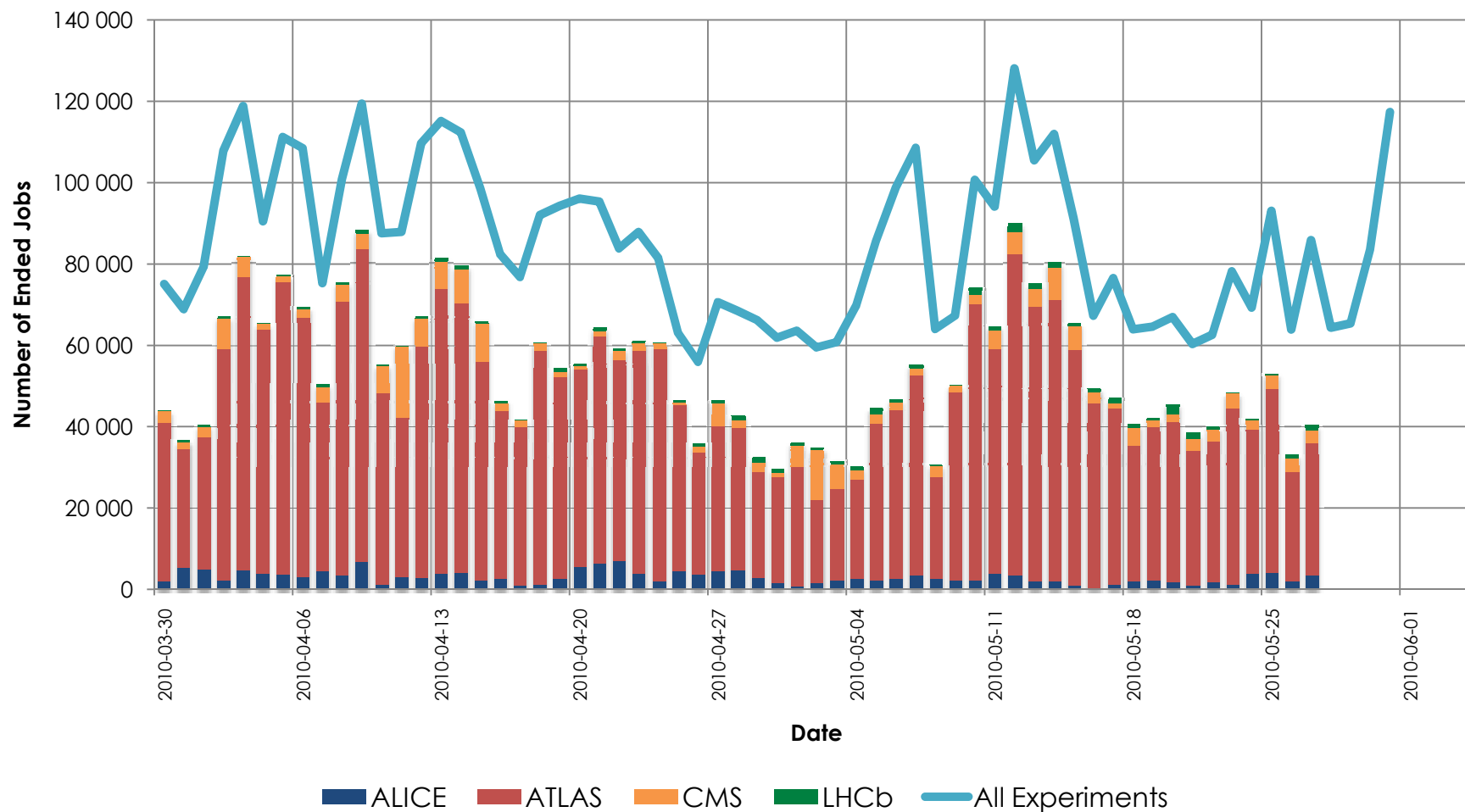
70% 80% 90% 100%

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

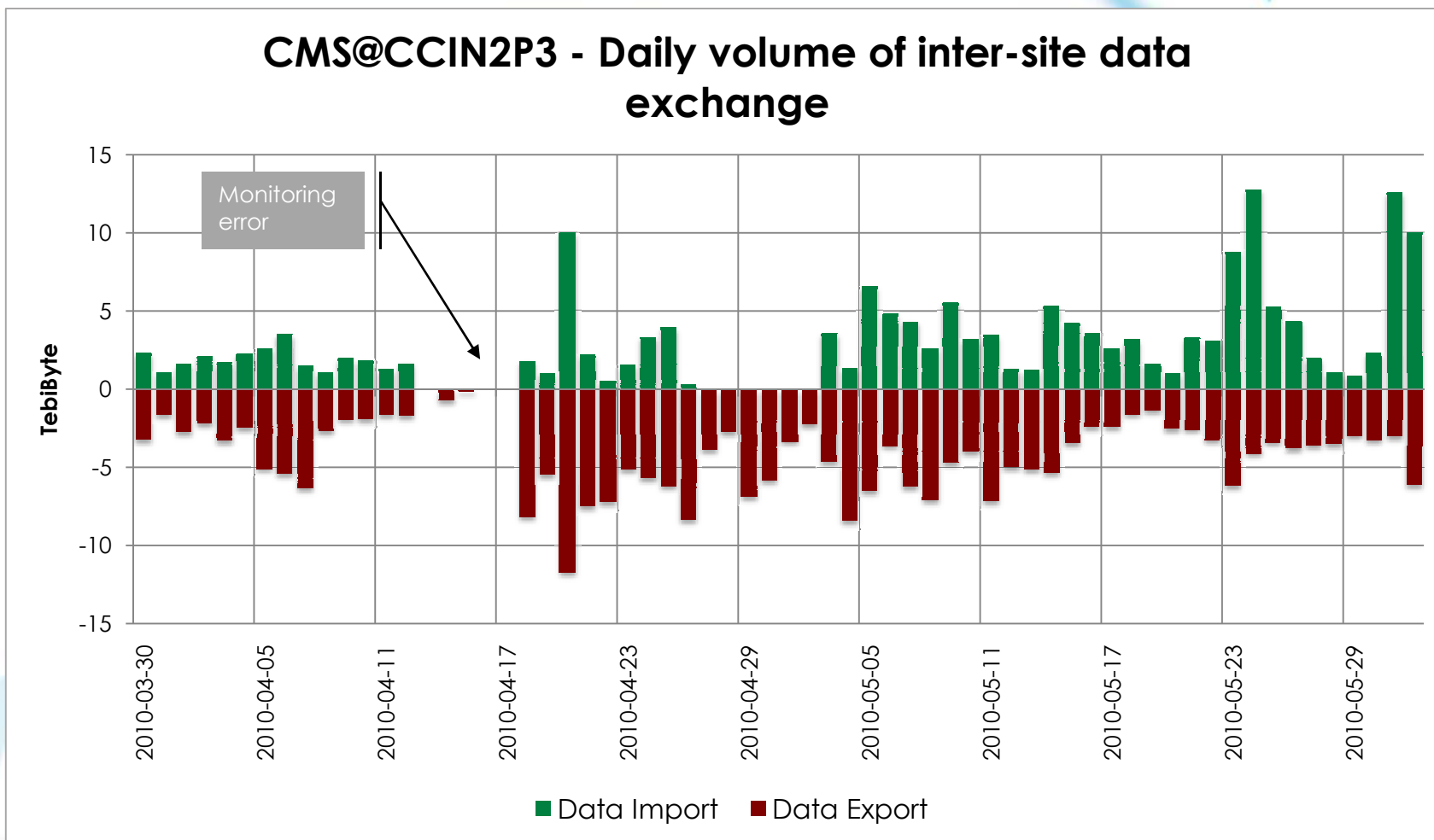
# ▶ Batch Activity -- LHC



## Daily Evolution of Batch Activity for LHC Experiments

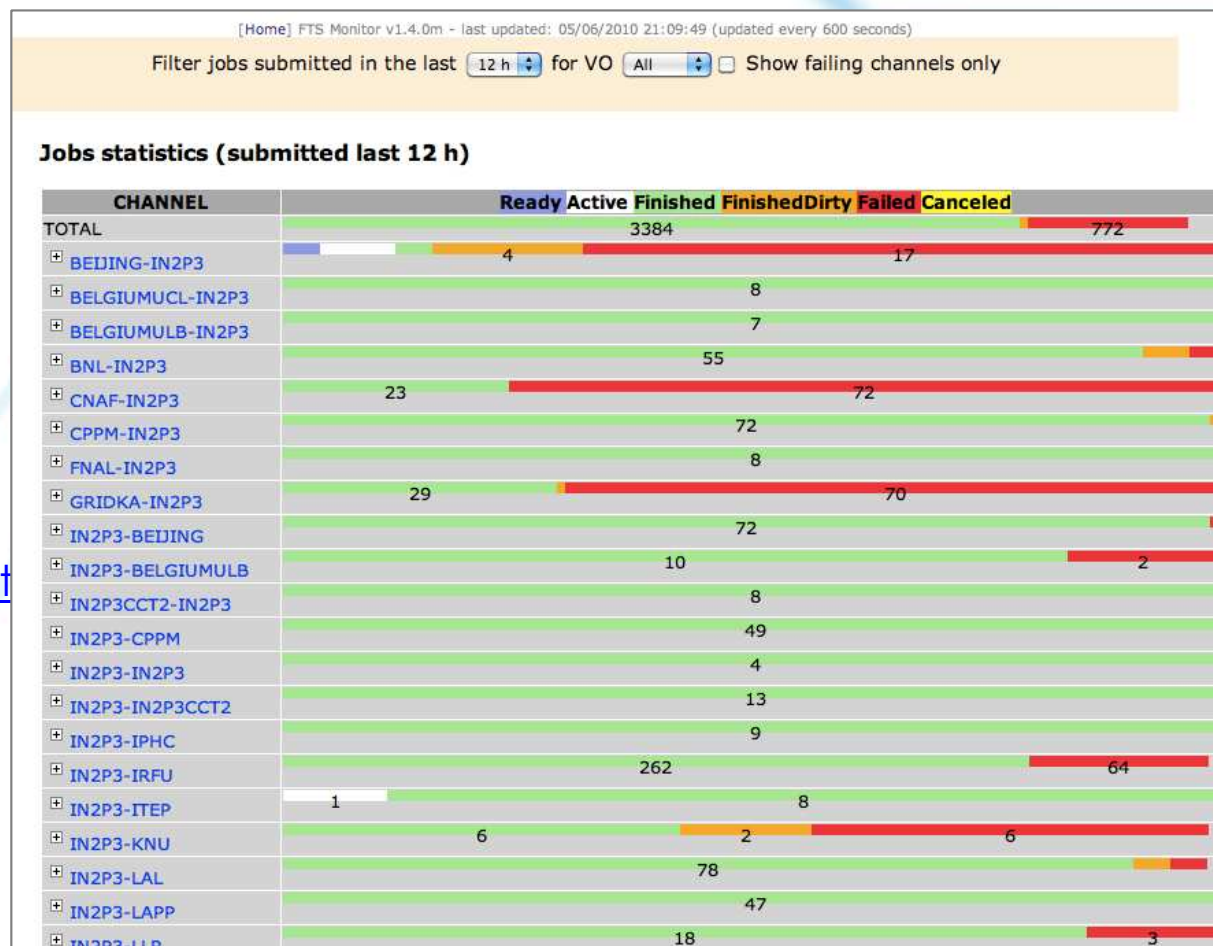


### CMS@CCIN2P3 - Daily volume of inter-site data exchange

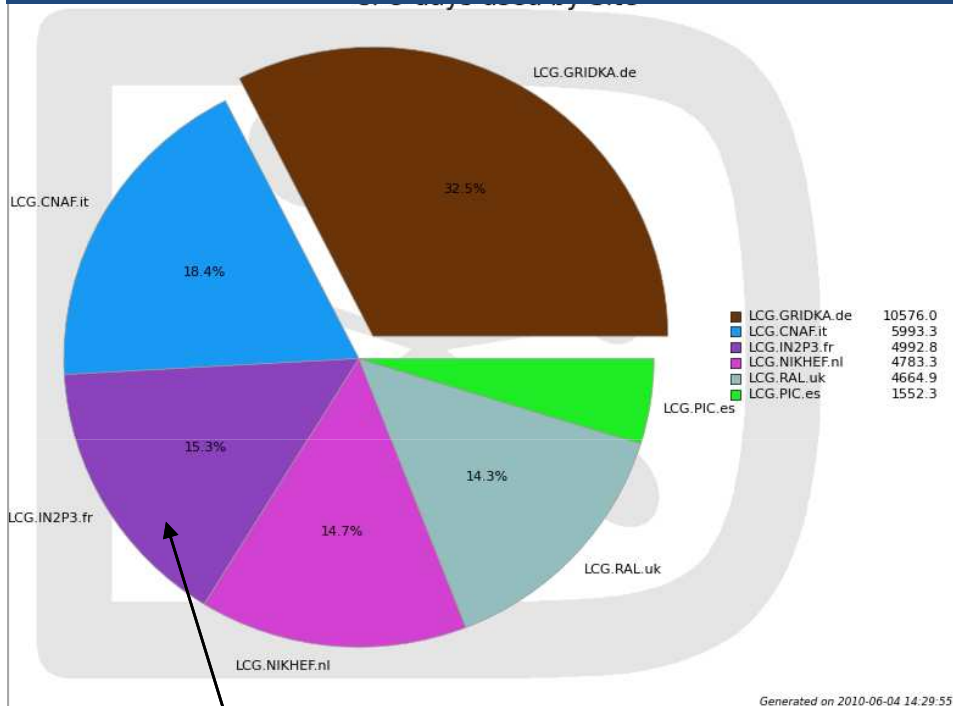




- CMS requires all the sites operating an FTS server to deploy ftsmonitor
  - Originally developed for our own purposes and now being used in several tier-1s
  - Details: <https://forge.in2p3.fr/projects/show/ftsmonitor>

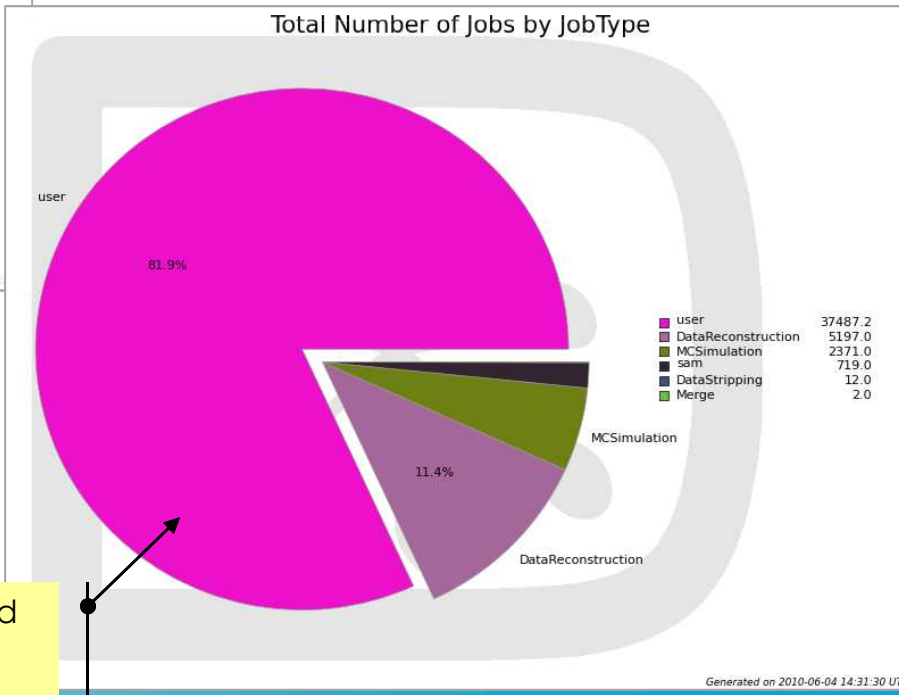


CPU time consumed at LHCb tier-1s  
 Period: Apr 1<sup>st</sup> – June 1<sup>st</sup> 2010



Relative contribution of CCIN2P3 lower than its share.

Type of LHCb jobs at CCIN2P3  
 Period: Apr 1<sup>st</sup> – June 1<sup>st</sup> 2010



Significant fraction of end user jobs

Source: <http://lhcbweb.pic.es/DIRAC>



# Computing Elements

(as of 02/06/2010)



Tier Level	CE hostname [.in2p3.fr]	CE Type	ALICE	ATLAS	CMS	LHCb
Tier-1	cclcgceli02	LCG	✓	✓		
	cclcgceli04	LCG			✓	✓
	cclcgceli07	LCG	✓	✓		
	cclcgceli08	LCG			✓	✓
	cccramceli01	CREAM	✓	✓	✓	✓
Tier-2	cclcgceli06	LCG	✓	✓	✓	✓
	cclcgceli09	LCG	✓	✓	✓	✓
	cccramceli03	CREAM	✓	✓	✓	✓

Source: <http://grid.in2p3.fr/index.php?chap=3&tit1=1&tit2=3>

# Experiment- specific activities

# ▶ ALICE



- Dedicated on-site liaison person (post-doc, 0.5 FTE) for ALICE activities joined the site in July 2009
  - Improved the interaction of the site and the experiment
- Introduction of stand-alone ALIEN xrootd-based storage element
  - The ALICE-specific security mechanisms necessary by this component needed to be adapted by the experiment to be compatible with the software platform used by CCIN2P3
    - *The experiment does not seem to be able to sustain this activity*
    - *On-going discussion between the site and the experiment for exploring the possible ways of making progress*
  - The tape staging capability is now available and being certified by the experiment

- Additional person (post-doc, 0.5 FTE) joined the ATLAS team late 2009
- Contribution to the regular campaigns of grid-based analysis
  - These exercises unveiled some inefficiencies in the way the experiment stored and traversed ROOT trees, which significantly increase the load on disk servers
  - Detailed feedback provided to the experiment by site experts contributed to a modification in the format of storing data for analysis
- Deployment of caching mechanisms for conditions data (FroNTier + Squid)
  - Important in particular for analysis activities at foreign associated tier-2s, mainly Beijing and Tokyo, to circumvent problems due to network latency

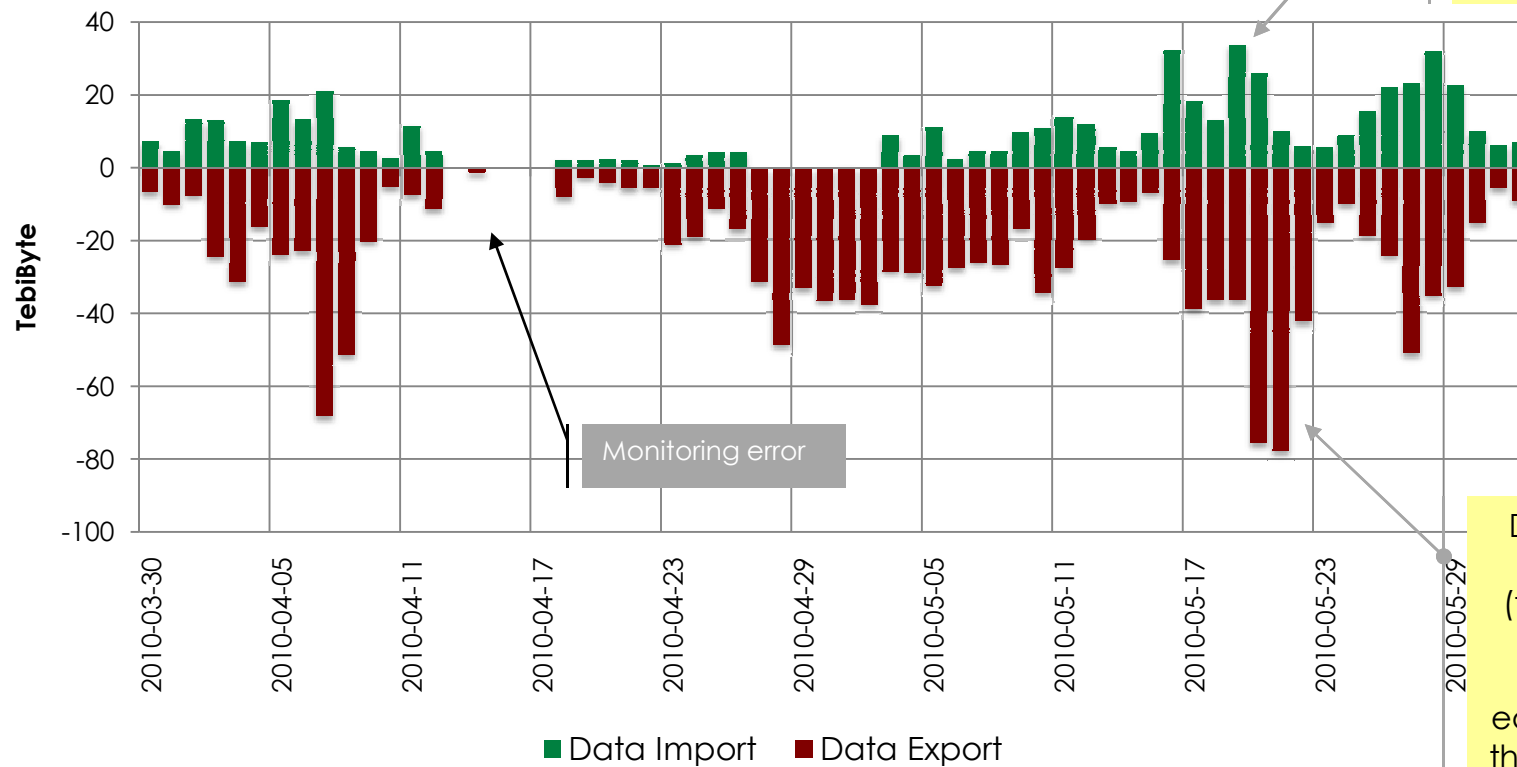
# ▶ ATLAS (cont.)



- Simulation
  - Most understood activity
    - Amounts to 50% of CPU activity on tier-1s ATLAS wide
  - Noticeable increase in the number of jobs
- Reprocessing
  - 2 campaigns since 7 TeV collisions: require CPU power and data exchange
    - Disk resident data: pre-staging not needed
    - High number of files to be transferred, among tier-1s and to tier-2s: 3 times more files transferred than during STEP'09
    - Even so, data distribution finished in 10 days, 4 times faster than foreseen by the computing model
  - Some problems observed (and fixed) transferring data from CCIN2P3 to some associated tier-2s. In addition, ongoing work to understand the measured slowness while transferring data with some ATLAS sites

# ▶ ATLAS (cont.)

## ATLAS@CCIN2P3 - Daily volume of inter-site data exchange



Data **imported** from other ATLAS sites (tier-1s and tier-2s)

Monitoring error

Data **exported** to other ATLAS sites (tier-1s and tier-2s)

This is roughly equivalent to 1/3 of the amount of data delivered for feeding local jobs



# ▶ ATLAS (cont.)



- Analysis
  - Observed a substantial increase (roughly x2) of user analysis jobs since 7 TeV collisions
    - *Both ATLAS wide and at CCIN2P3*
  - The batch farm parameters were tuned to improve the turn around of user analysis jobs
    - *At the expense of the throughput of simulation jobs*

# ▶ ATLAS (cont.)

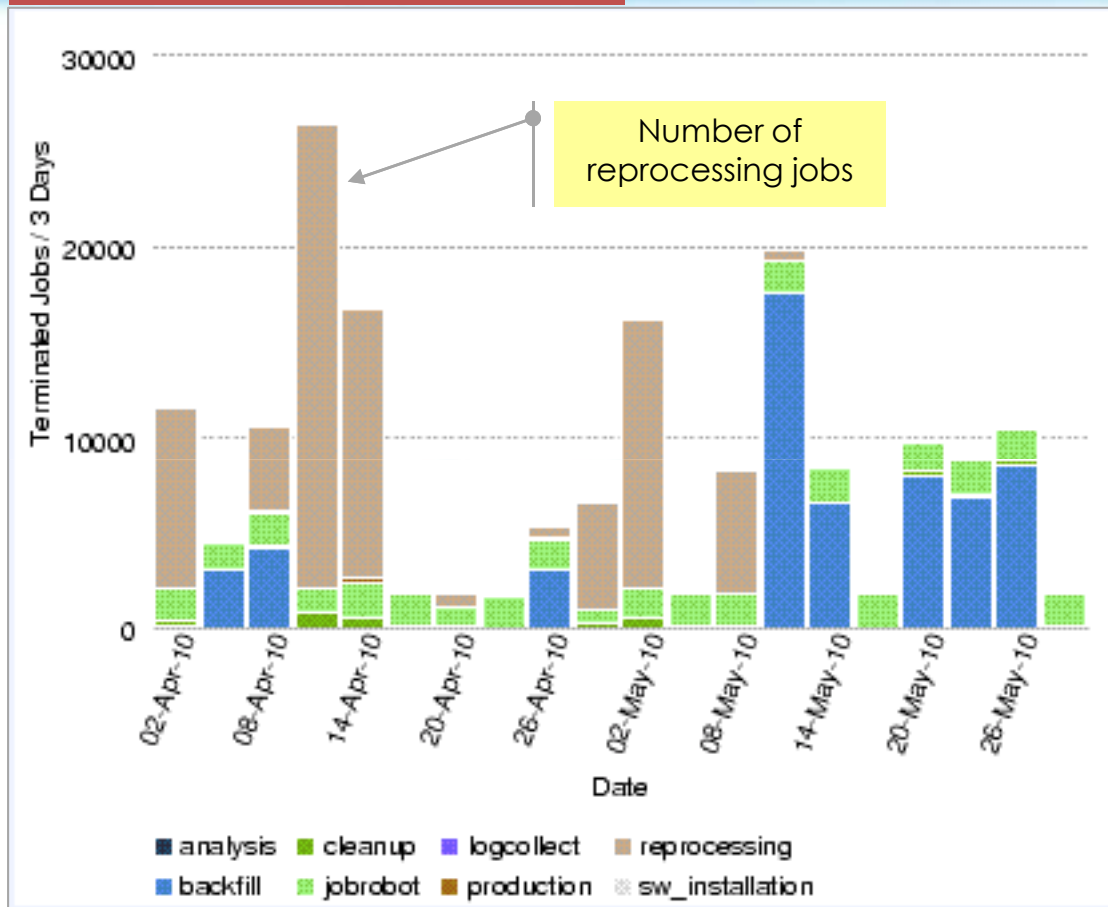


- Software area at tier-1
  - High workload on the software area effectively managed by automatically creating 3 replicas of each software release
    - *Sustained 5000+ simultaneous ATLAS jobs in execution over several weeks*
    - *5.8 million files, replicated 3 times*
  - The automated installation and replication chain is too fragile, difficult to debug and often requires human intervention, either from the experiment or by site experts
  - Ongoing work to further improve automation of the replication mechanism and shield ATLAS from site-specific failures

- A second CMS dedicated liaison person joined the site late 2009
  - Better share of the workload among the 2 CMS experts on site
- Complete revisit of the site configuration for the import/export of CMS data and improvements in the day-to-day operations of the data exchange activity
  - Increased stability of the site as measured by the experiment
- Implemented a stricter separation of storage spaces for the co-located CMS tier-1 and tier2
  - Prevent user analysis jobs to impact tier-1 activities, in particular reprocessing
- Improvements in the internal procedures for removing experiment's data, as a consequence of the data deletion incident late Nov 2009
  - Confusing instructions wrongly interpreted by the site's staff conducted to an unintended removal of 480 TB of custodial simulated data

# ▶ CMS (cont.)

## CMS Activities at CCIN2P3 tier-1



Period: Apr 1<sup>st</sup> – June 2<sup>nd</sup> 2010

Source: <http://dashb-cms-sam.cern.ch/dashboard/request.py/dailysummary>

- CCIN2P3 received 17% or RAW data from tier-0
  - Proportional to its share
- Tier-1: Relatively low reprocessing activity
  - Backfill and job robots amounts for more than half of the executed jobs
- Tier-2
  - Production: 33%
  - Analysis: 13%

- Experienced some difficulties accessing data by using authenticated protocols due to software bugs
  - Issues in both the file access layer and in experiment-packaged software, sometimes incompatible with components installed by sites
    - *Similar issues observed in other LHCb tier-1s using dCache*
  - Activation at tier-1 of an unauthenticated protocol while waiting for validation by the experiment that the observed issues with the secure ones are definitively corrected
- Reconfiguration of the batch farm's queues to better match LHCb needs
- Unexpected low activity, in any case lower than in other tier-1s, in spite of the availability of CPU and storage capacity
  - Currently investigating with LHCb computing experts how to make the site more attractive