CCIN2P3 contribution to WLCG

Fabio Hernandez
fabio@in2p3.fr

COS/ESC 2009

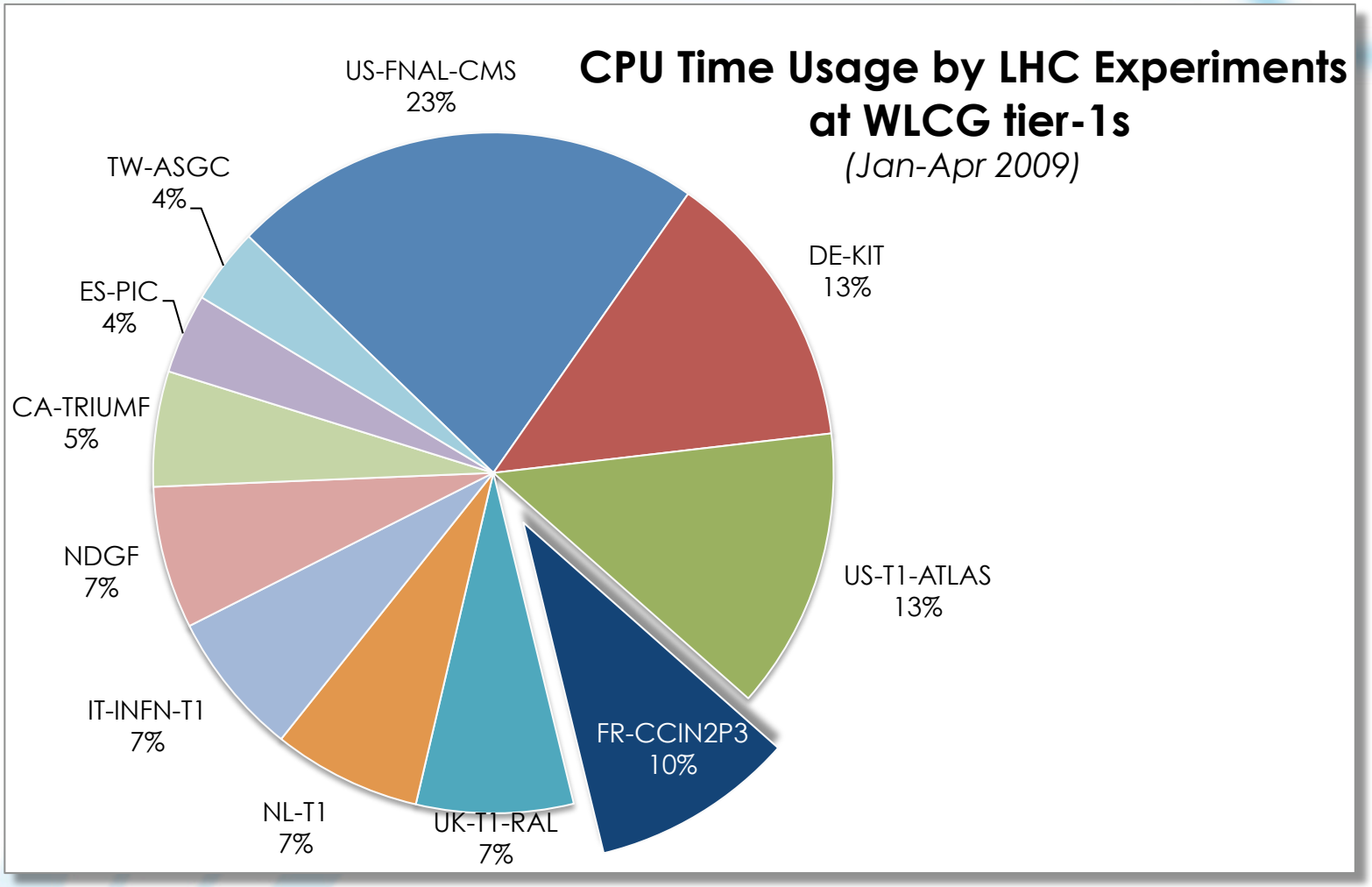
Lyon, June 15th-16th, 2009

Contents

- CCIN2P3 in the context of WLCG
- Site overview
- Main activities
  - Data exchange
  - Data storage
  - On-site data processing
- Other grid-related activities
- Perspectives
- Conclusions
- Questions & comments

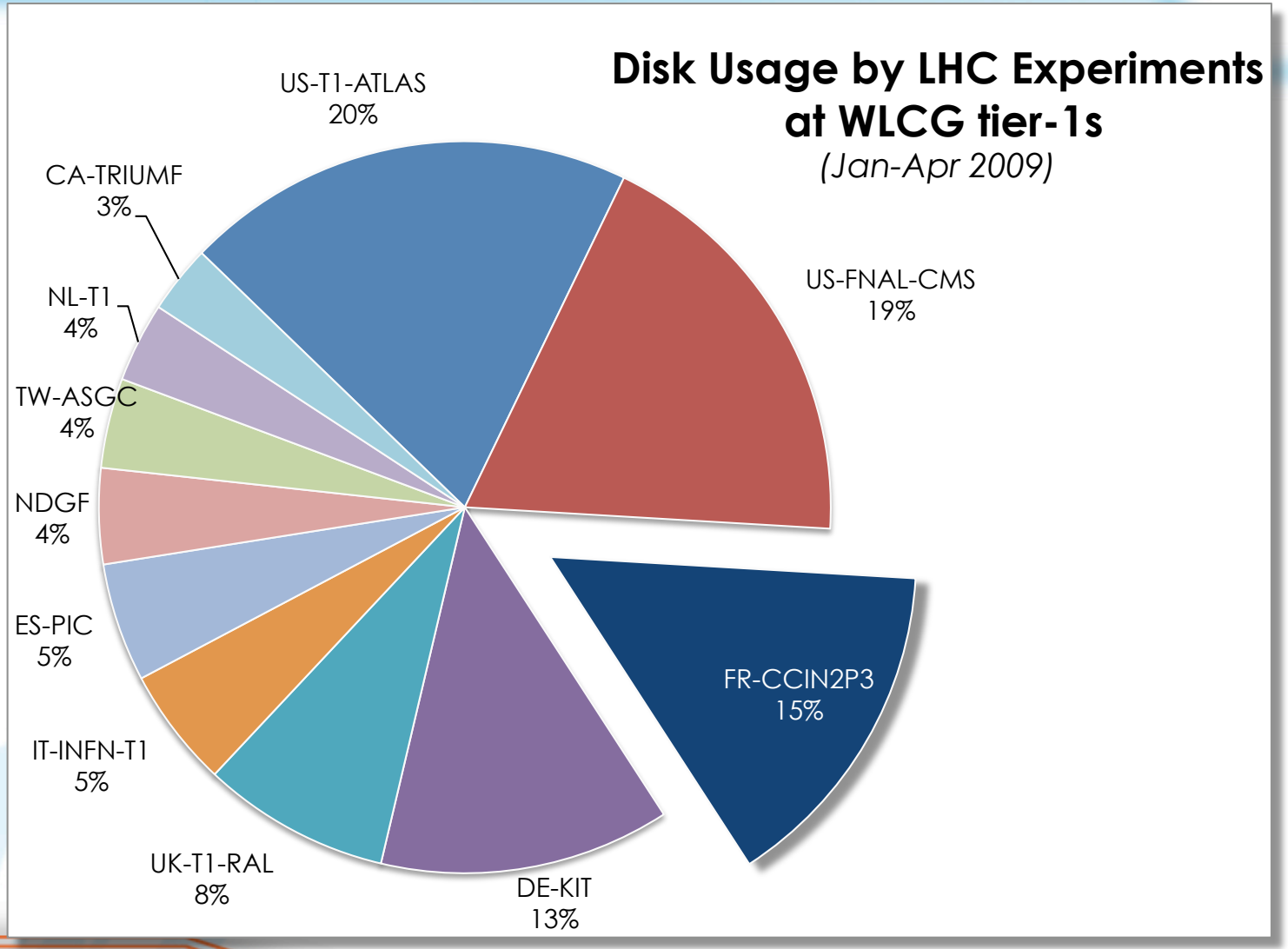
CCIN2P3 in the context of WLCG

Contribution: CPU



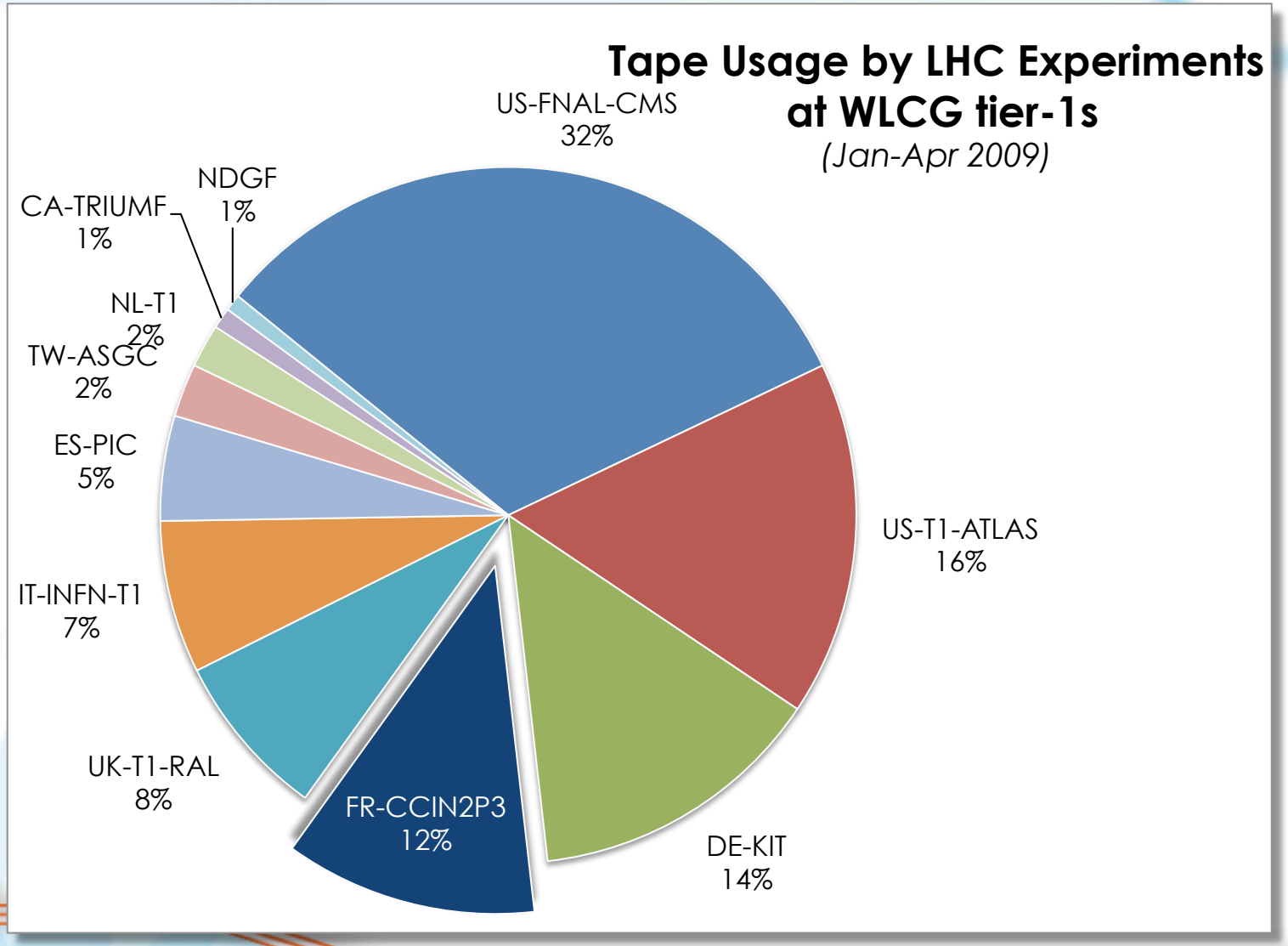
Source: WLCG CERN & Tier-1 Monthly Accounting Reports

Contribution: disk



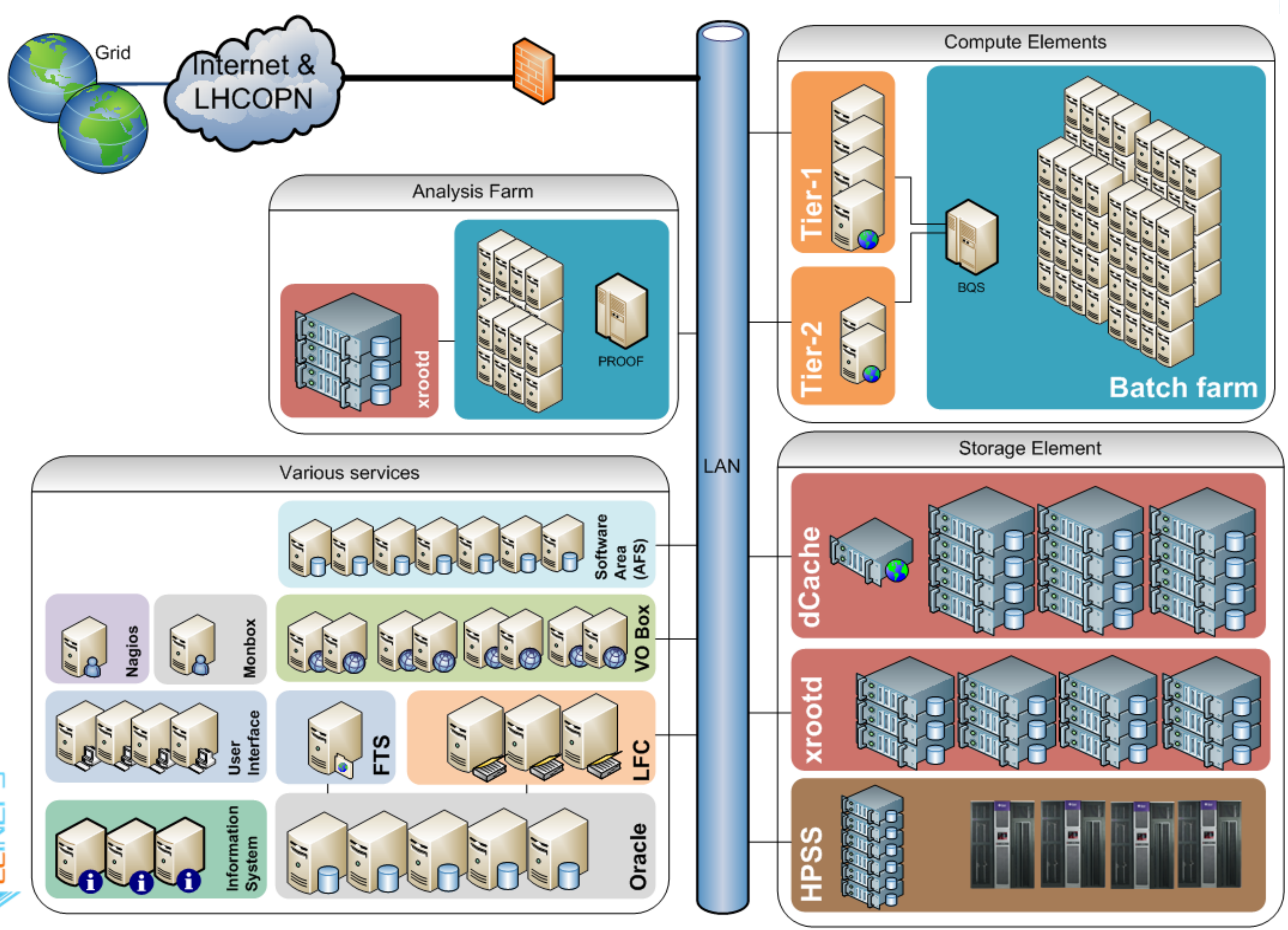
Source: WLCG CERN & Tier-1 Monthly Accounting Reports

Contribution: tape



Source: WLCG CERN & Tier-1 Monthly Accounting Reports

Site overview

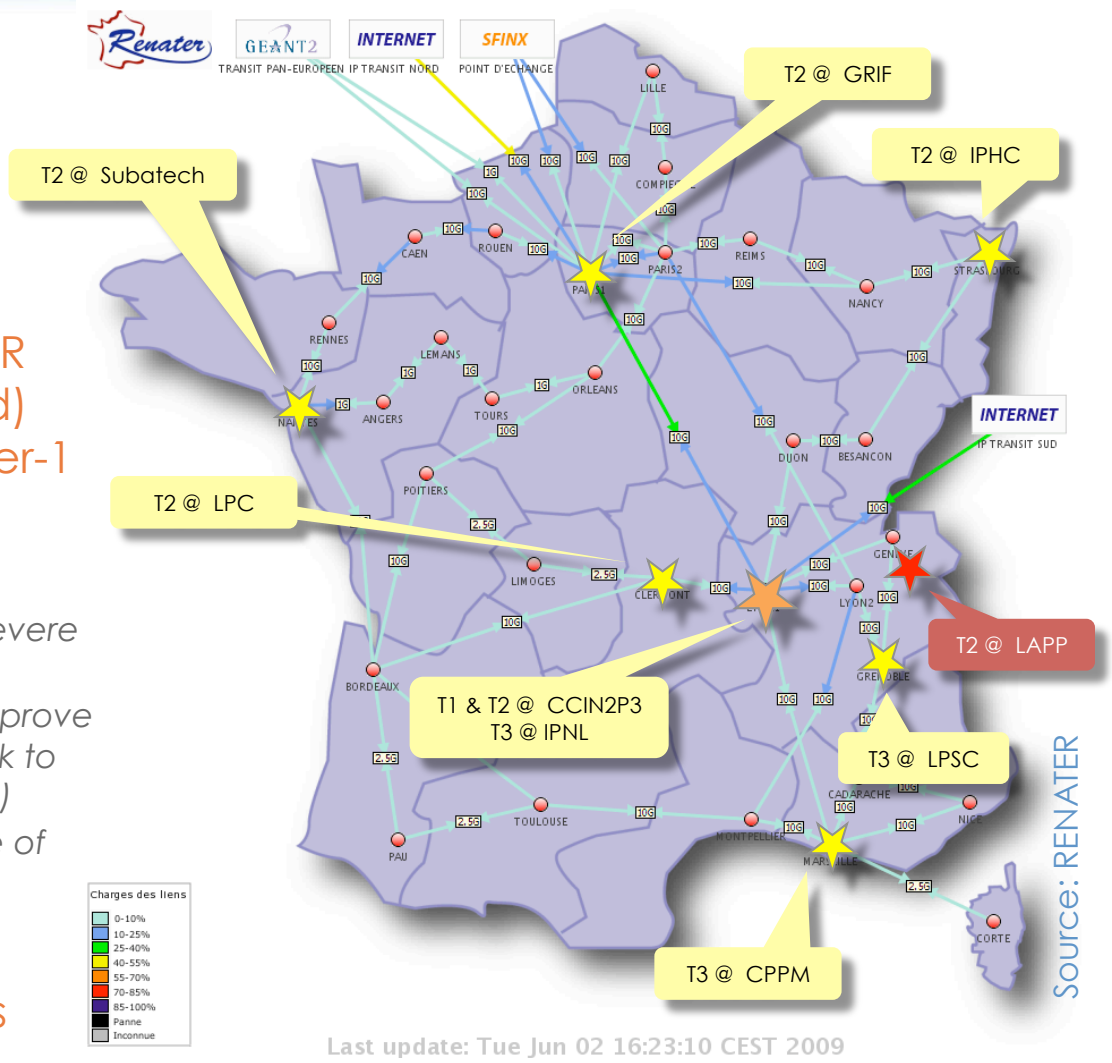


Author: Fabio Hernandez
Last Updated: 2009-06-09

Data Exchange

Connectivity

- Tier-0 and tier-1s
  - LHCOPN links (10 Gbps):
    - *CCIN2P3 ↔ CERN*
    - *CCIN2P3 ↔ KIT ↔ CERN*
- Domestic tier-2s and tier-3s
  - Towards 10 Gbps links to RENATER backbone (dedicated or shared) for exchanging LHC data with tier-1 in Lyon
  - Exception: T2 @ LAPP
    - *Currently shared 1 Gbps presenting severe service reliability problems*
    - *On going actions in 2 phases: 1) to improve reliability and 2) to look for a direct link to national backbone (more bandwidth)*
    - *This continues to be critical at the eve of LHC data taking*
- Foreign tier-2s and tier-3s
  - Link to GEANT routers at 10 Gbps

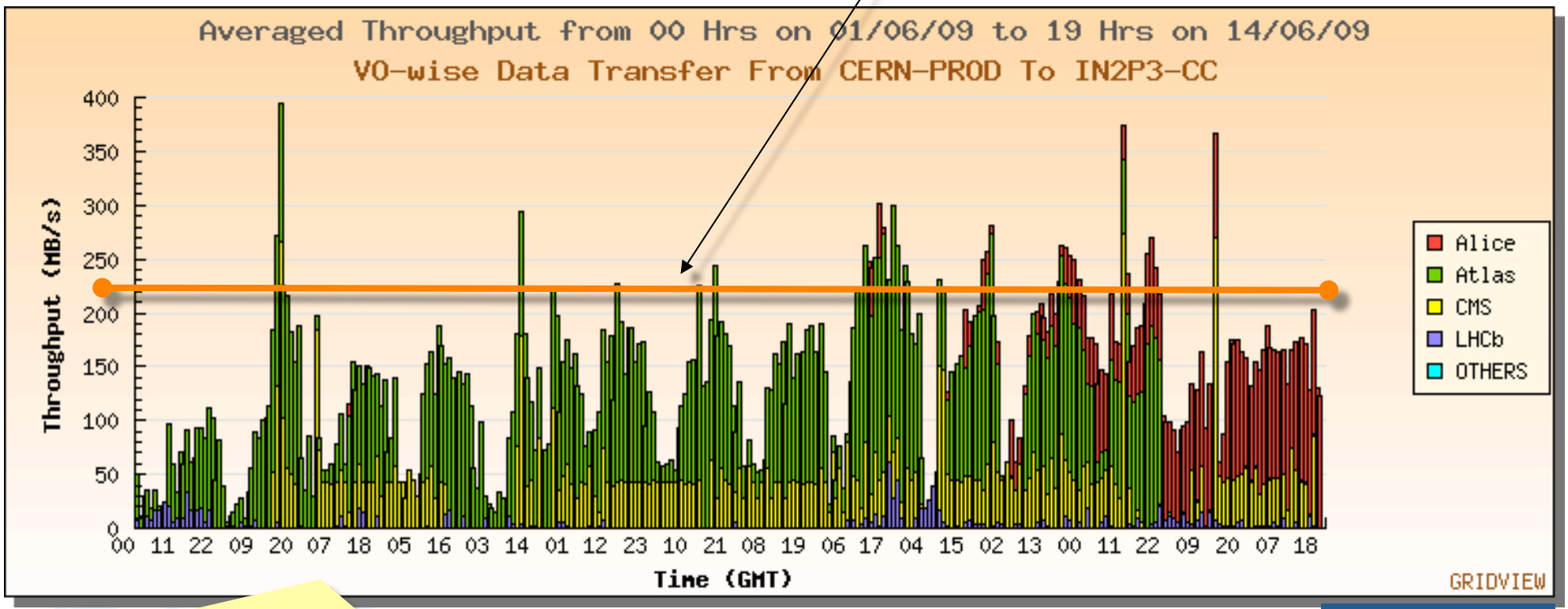


Data transfer

- Tier-0 → CCIN2P3

Nominal target: 225 MB/s

- ALICE: 6 MB/s
- ATLAS: 109 MB/s
- CMS: 100 MB/s
- LHCb: 10 MB/s



June 1-14, 2009

Sustained rates up to 250 MB/sec were already demonstrated during CCRC'08 over several days.

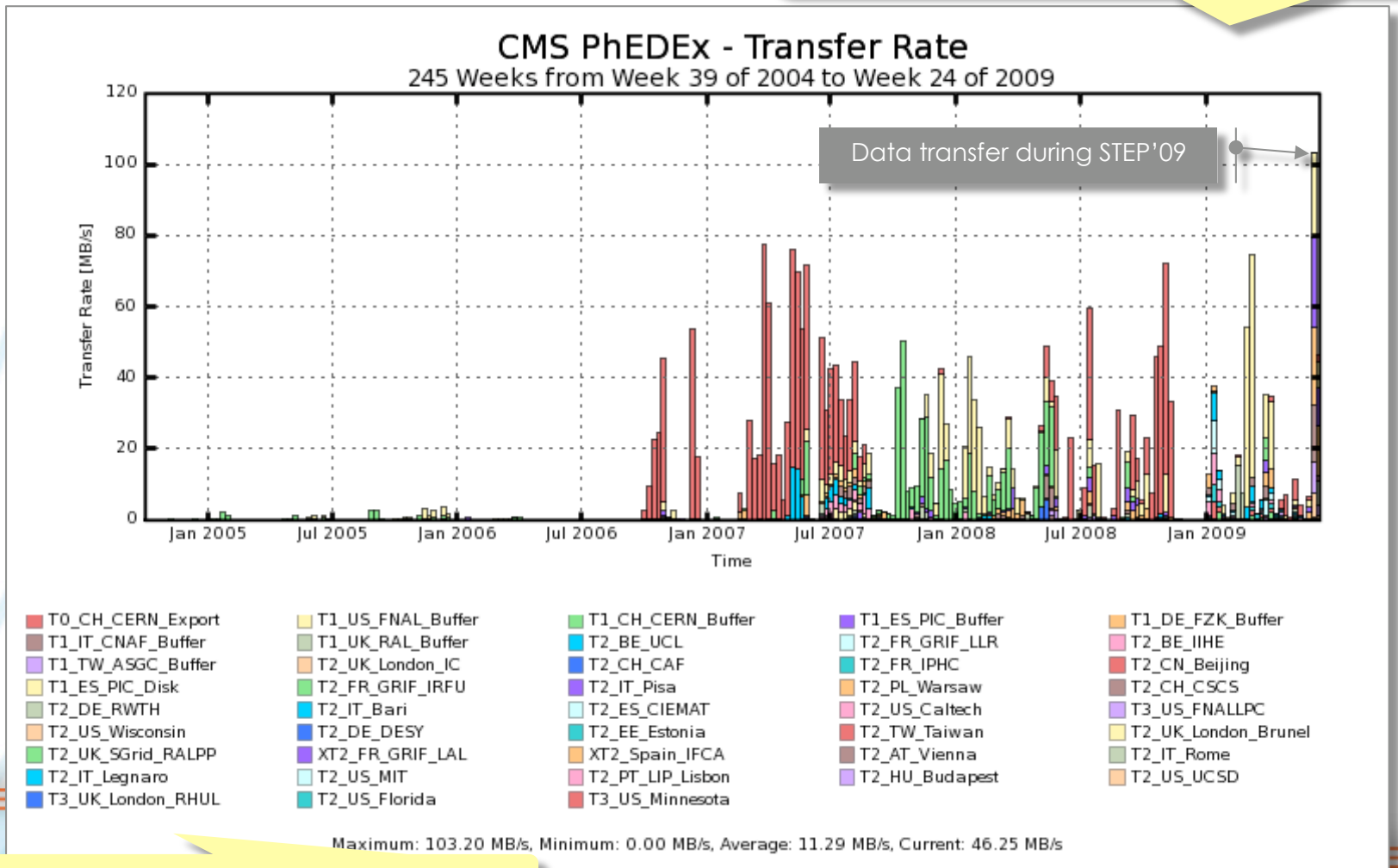
Source: Gridview http://gridview.cern.ch

Data transfer: example of CMS

- Other sites → CCIN2P3

Weekly-averaged rates.

Note the spiky nature of the transfers.
Average rate does not reflect reality.

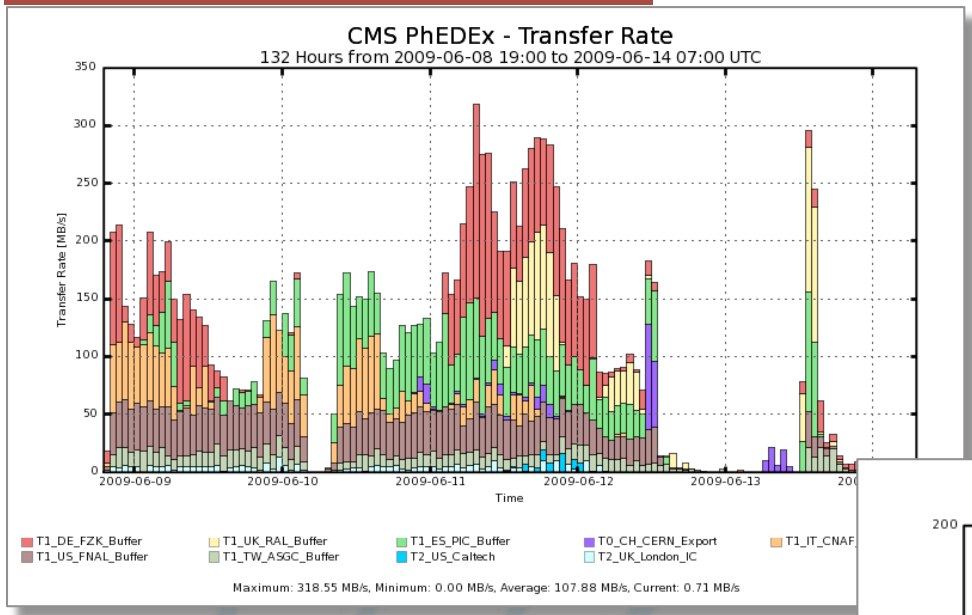


Source: CMS Phedex http://cmsweb.cern.ch/phedex/

Daily exchange of data with dozens of sites

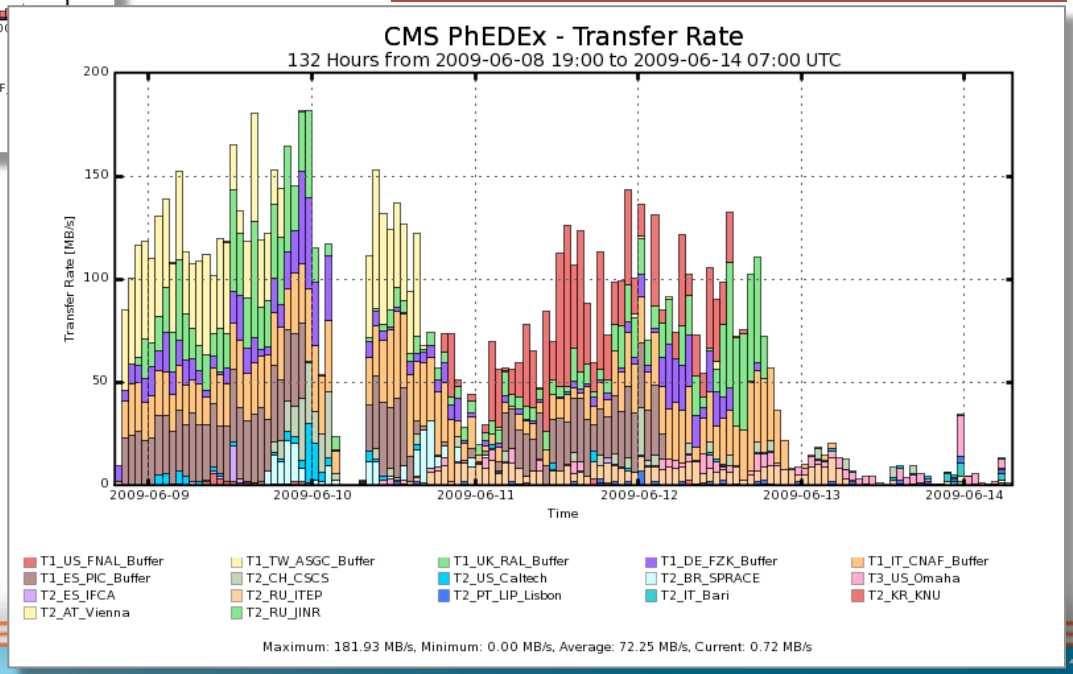
Data transfer: CMS

CMS sites → CCIN2P3



STEP'09

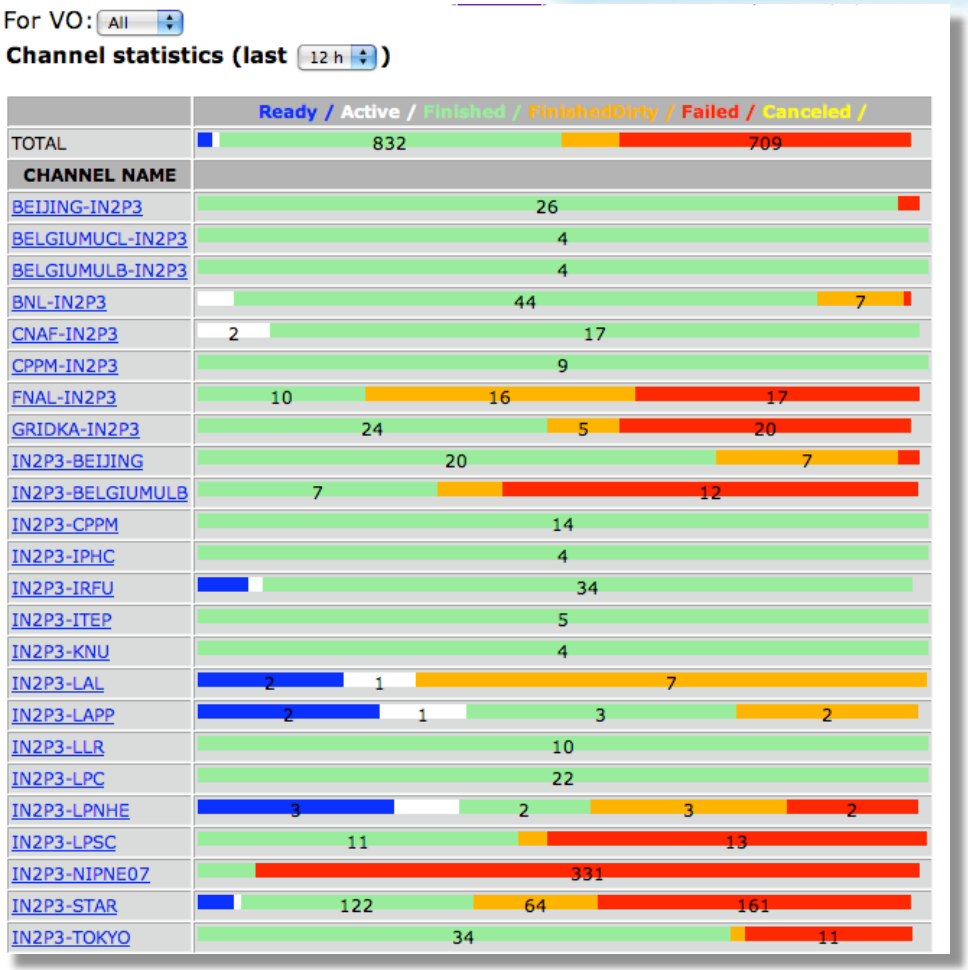
CCIN2P3 → CMS sites



Source: CMS Phedex http://cmsweb.cern.ch/phedex/

File Transfer Service (FTS)

- In charge of scheduling data transfers for importing data from other T1 and for importing/ exporting data from/to tier-2s
- Stable configuration
  - 4 machines for handling the load
  - 1 additional standby virtual machine
  - Proved sufficient during CCRC'08. Since then we upgraded the hardware and deployed the latest stable version on SL4 64bits.
- In-house developed tool for real-time monitoring of requests
  - Ongoing work to improving it by collecting and displaying historical information

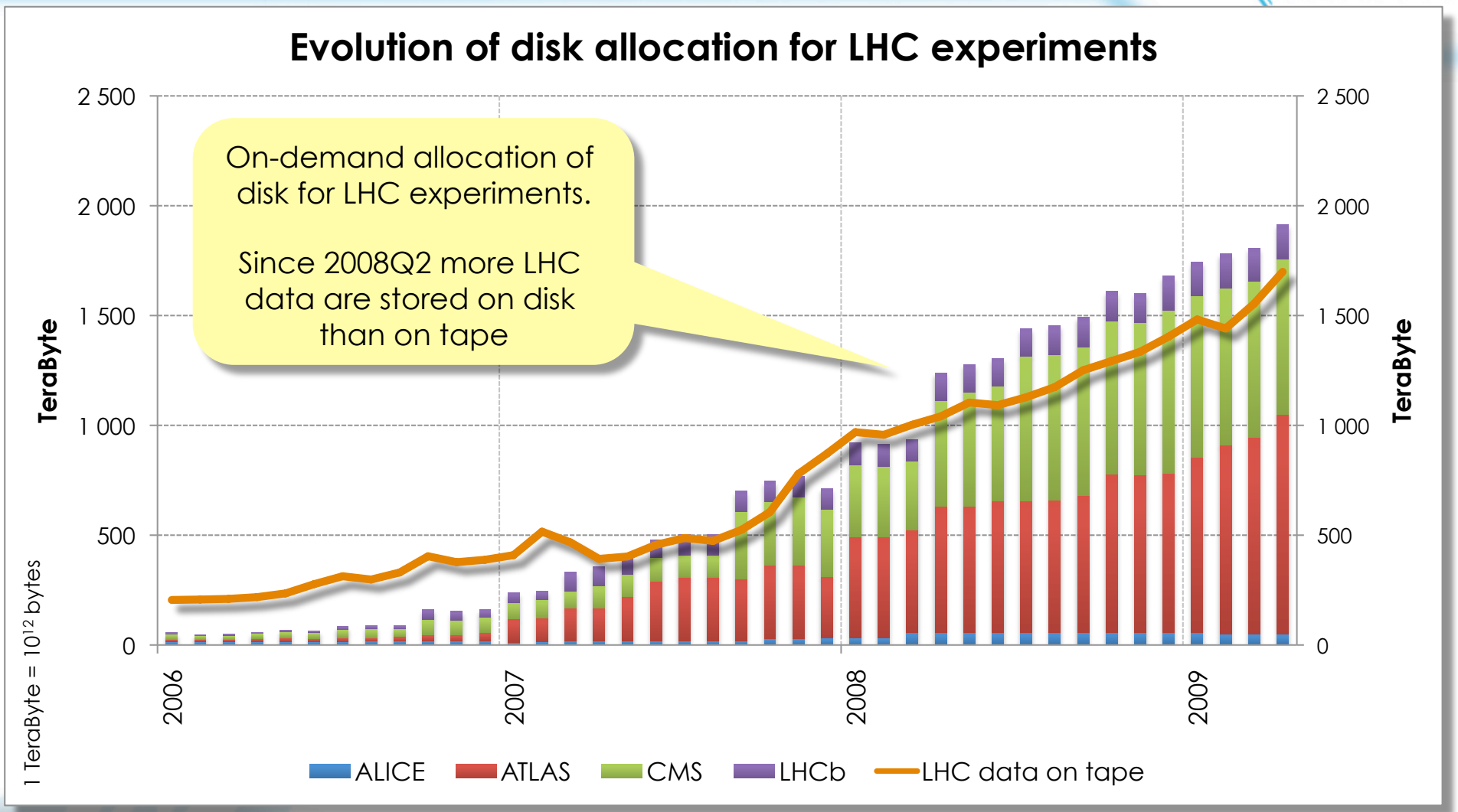

For VO: All

Channel statistics (last 12 h)

	Ready / Active / Finished / FinishedDirty / Failed / Canceled /
TOTAL	832 709
CHANNEL NAME	
BEIJING-IN2P3	26
BELGIUMUCL-IN2P3	4
BELGIUMULB-IN2P3	4
BNL-IN2P3	44 7
CNAF-IN2P3	2 17
CPPM-IN2P3	9
FNAL-IN2P3	10 16 17
GRIDKA-IN2P3	24 5 20
IN2P3-BEIJING	20 7
IN2P3-BELGIUMULB	7 12
IN2P3-CPPM	14
IN2P3-IPHC	4
IN2P3-IRFU	34
IN2P3-ITEP	5
IN2P3-KNU	4
IN2P3-LAL	2 1 7
IN2P3-LAPP	2 1 3 2
IN2P3-LLR	10
IN2P3-LPC	22
IN2P3-LPNHE	3 2 3 2
IN2P3-LPSC	11 13
IN2P3-NIPNE07	331
IN2P3-STAR	122 64 161
IN2P3-TOKYO	34 11

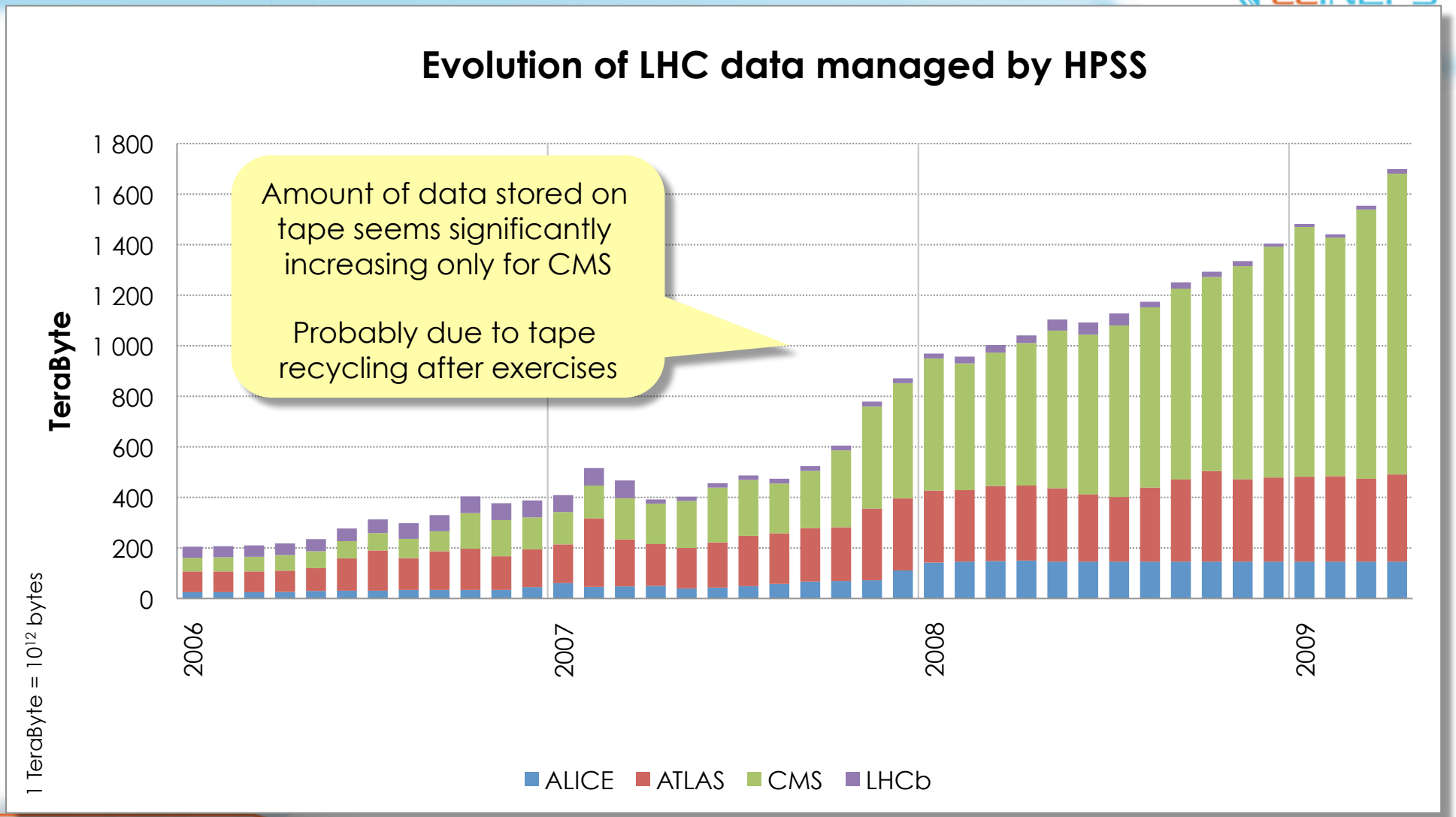
Data Storage

Disk-based storage



Source: WLCG CERN & Tier-1s accounting reports

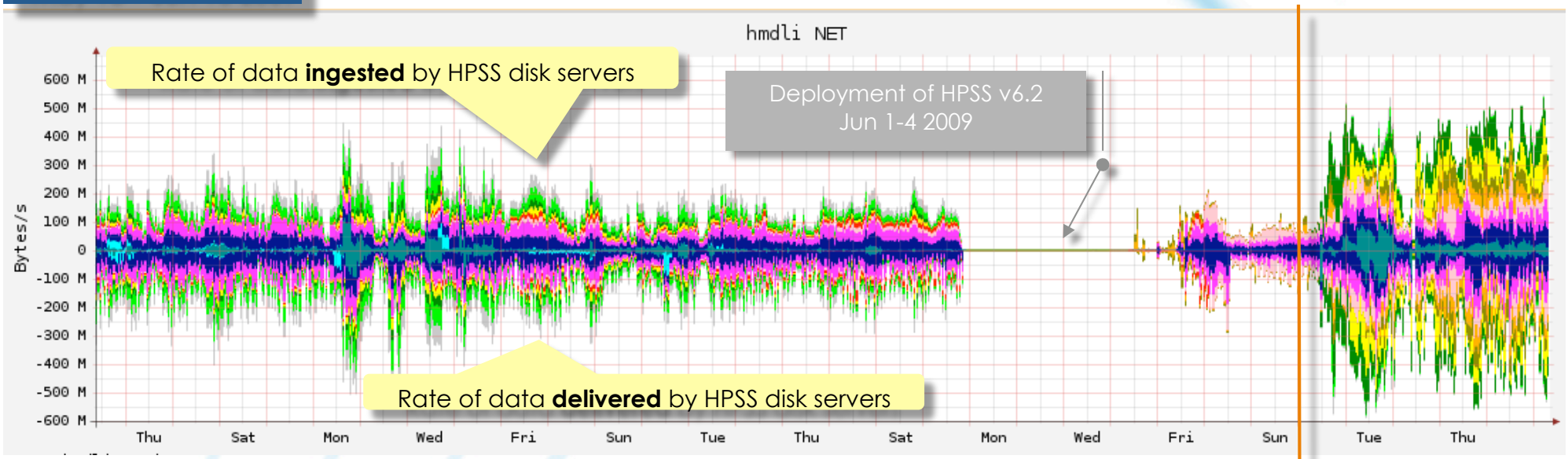
Tape-based storage

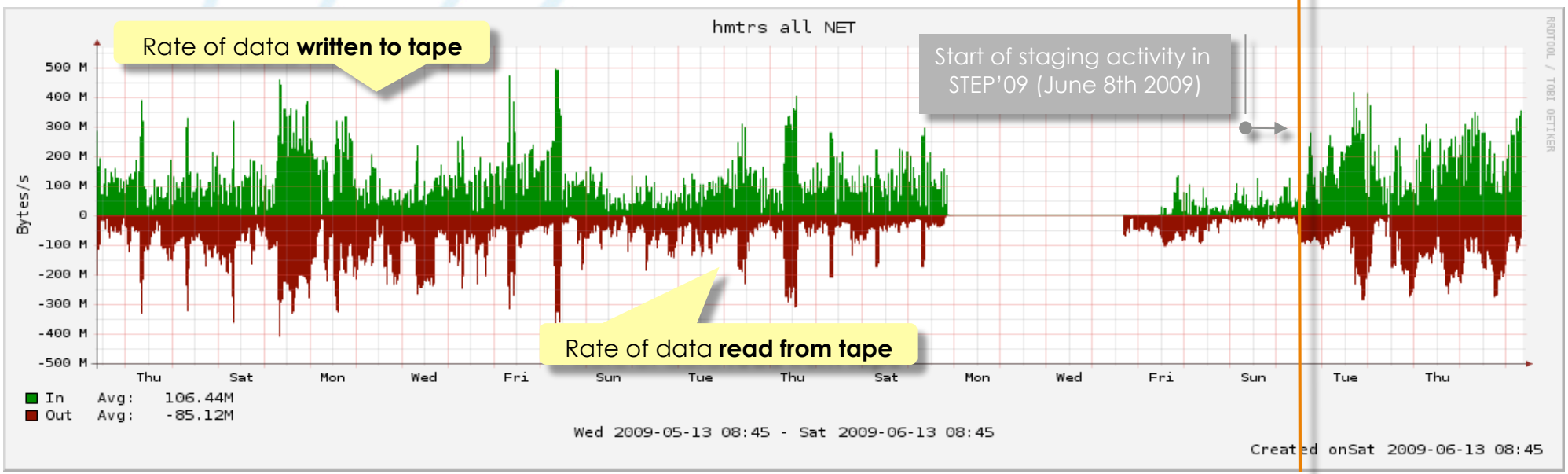


Source: WLCG CERN & Tier-1s accounting reports

Mass storage performance

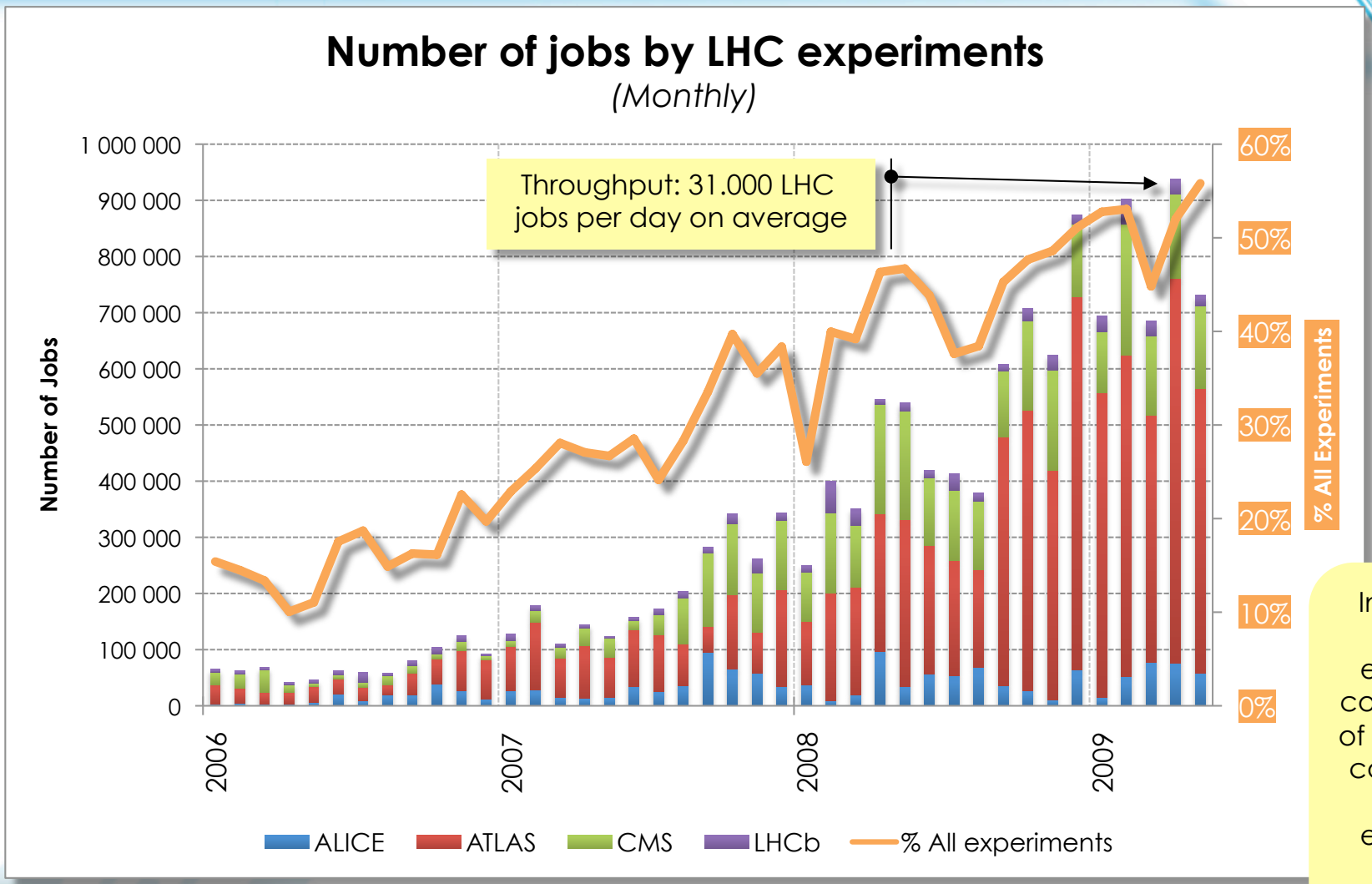
May 13 – Jun 13 2009





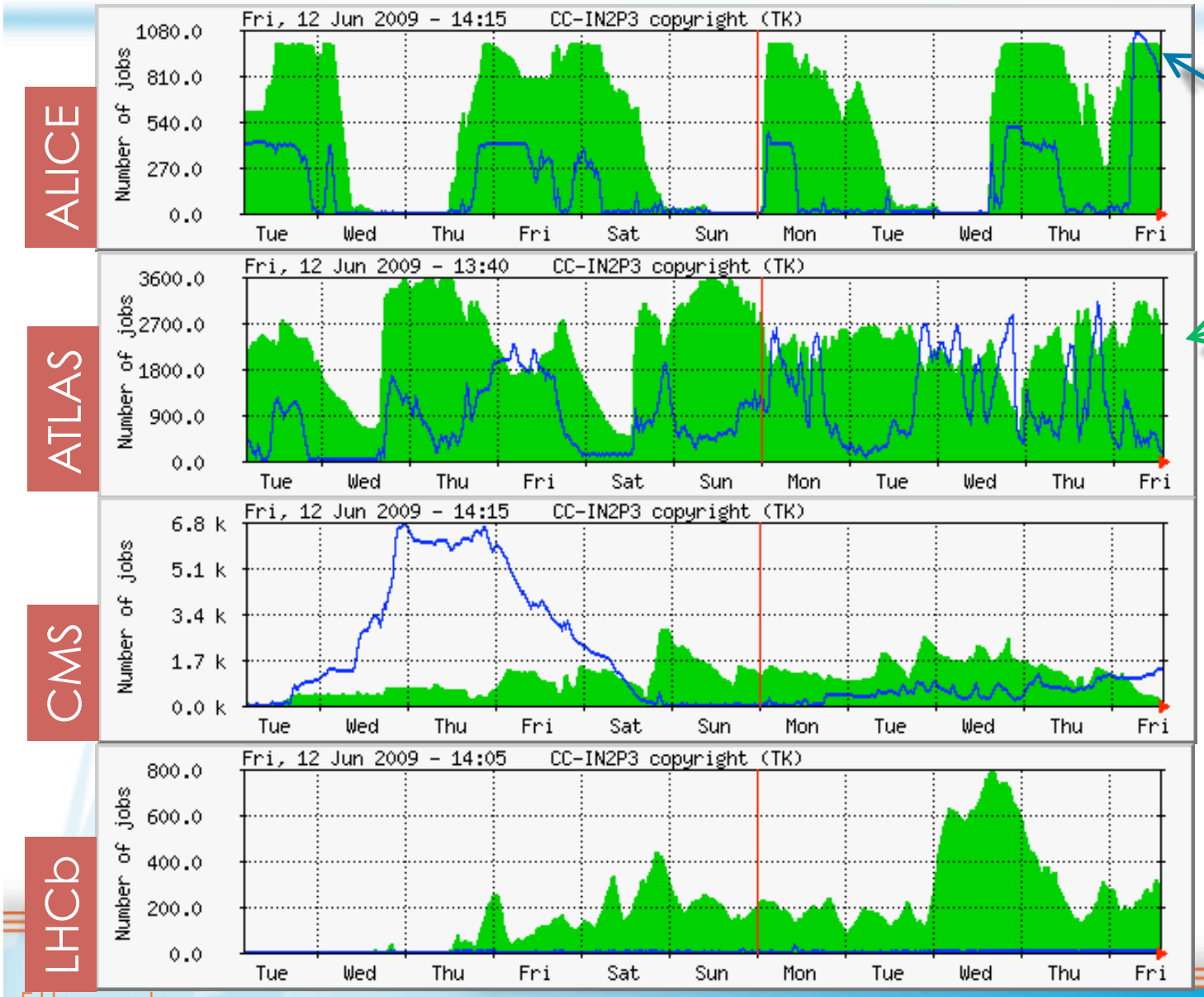
On-site LHC data processing

Batch workload



In April 2009, the 4 LHC experiments consumed 50% of the CPU time consumed by all the experiments served by CCIN2P3

Batch workload: STEP'09



Waiting jobs

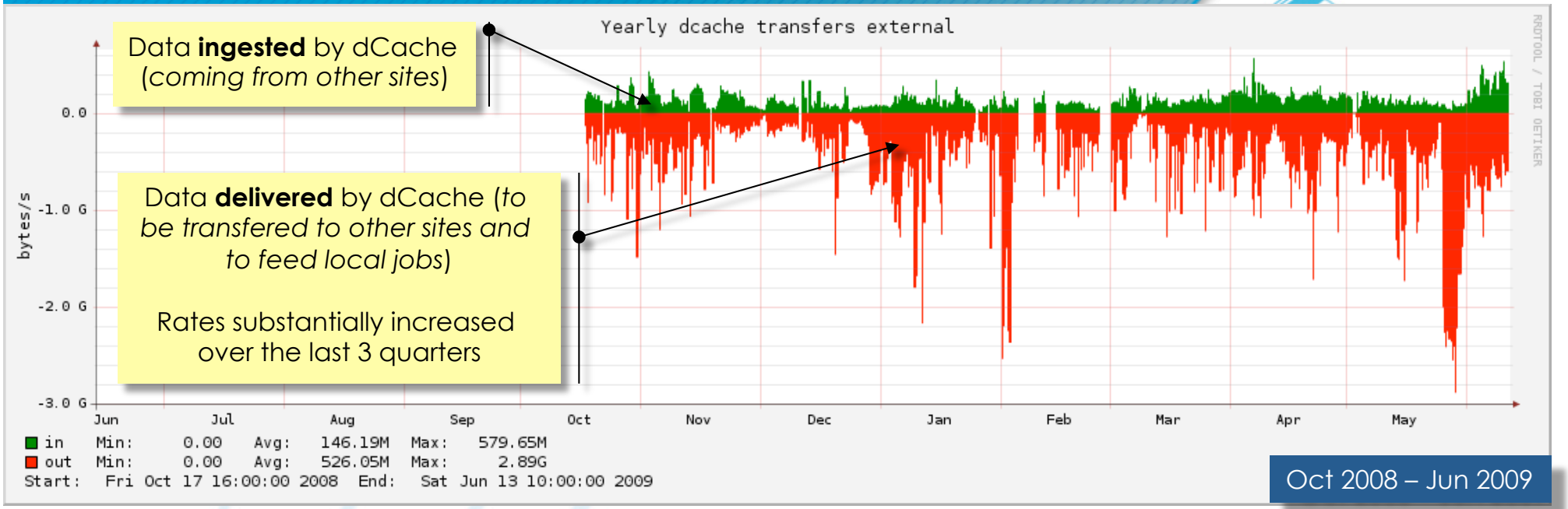
Running jobs

Average running jobs (June 2nd to June 12th 2009):

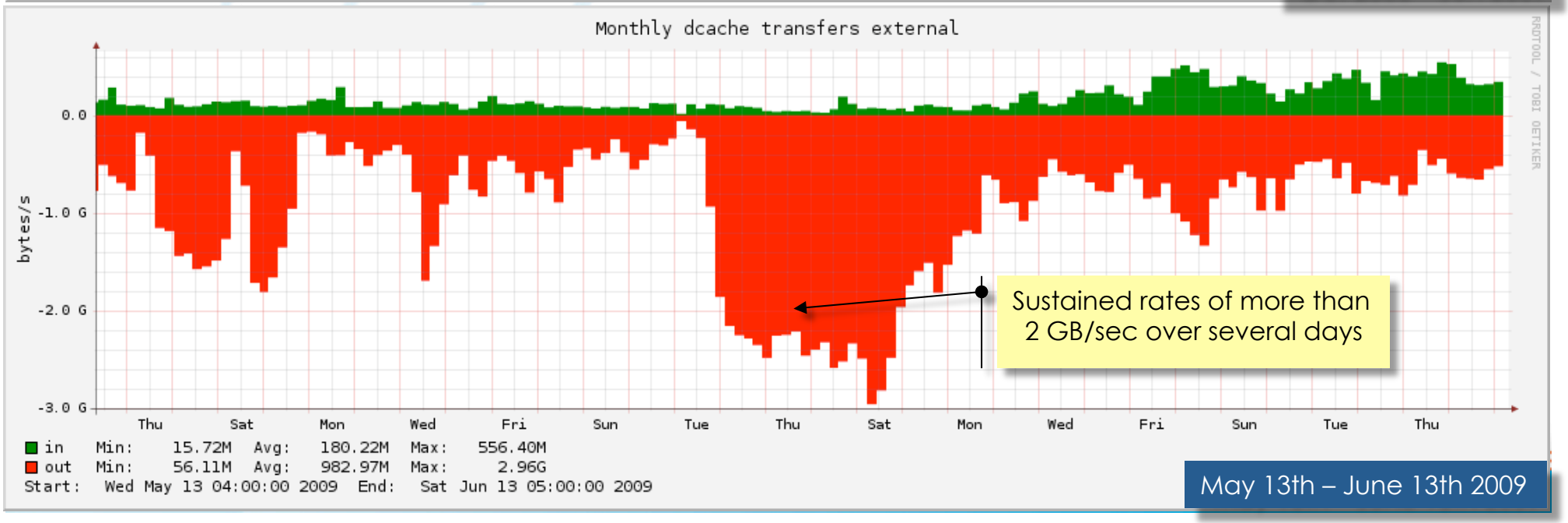
ALICE: 527
ATLAS: 2284
CMS: 1048
LHCb: 192

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

Serving local jobs



Oct 2008 – Jun 2009



May 13th – June 13th 2009

Other core services

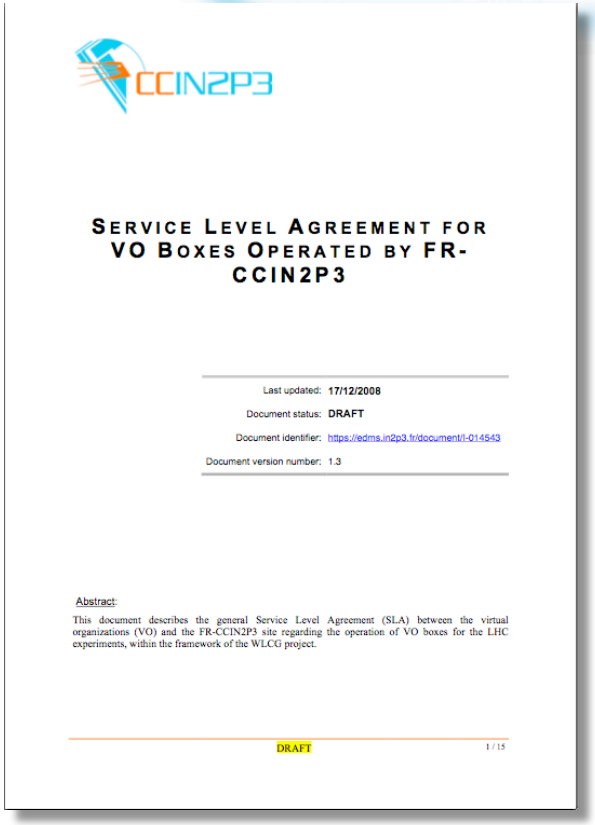
Other services

- File cataloguing: 2 instances of LFC, one for ATLAS and another for LHCb
  - Some stability problems were observed during CCRC'08 that were corrected since then by the developers. Now running stably in a redundant configuration.
  - Each instance is backed by an experiment-dedicated Oracle cluster

- Database replication by using Oracle streams
  - CERN → CCIN2P3
    - *ATLAS: replication of conditions data*
      - Recent tests conducted by ATLAS-France demonstrated that thousands of jobs can query directly the database for retrieving conditions data. Details in backup slides.
    - *LHCb: replication of file catalogue data base (LFC) and conditions data*
  - CCIN2P3 → CERN
    - *ATLAS: replication of AMI (Atlas Metadata Catalogue) backend. After some necessary modifications in the front-end application, replication is in production since early June*

Other services (cont.)

- VO boxes
  - Dedicated machines to run experiment-specific services
  - Currently operating 2 physical machines per experiment each with some built-in hardware redundancy
  - Experiment-specific software is expected to benefit of 2 machines to improve the availability of their services, in case of hardware failure
    - *See Service Level Agreement*



SERVICE LEVEL AGREEMENT FOR VO BOXES OPERATED BY FR-CCIN2P3

Last updated: 17/12/2008
Document status: DRAFT
Document identifier: https://edms.in2p3.fr/document/I-014543
Document version number: 1.3

Abstract:

This document describes the general Service Level Agreement (SLA) between the virtual organizations (VO) and the FR-CCIN2P3 site regarding the operation of VO boxes for the LHC experiments, within the framework of the WLCG project.

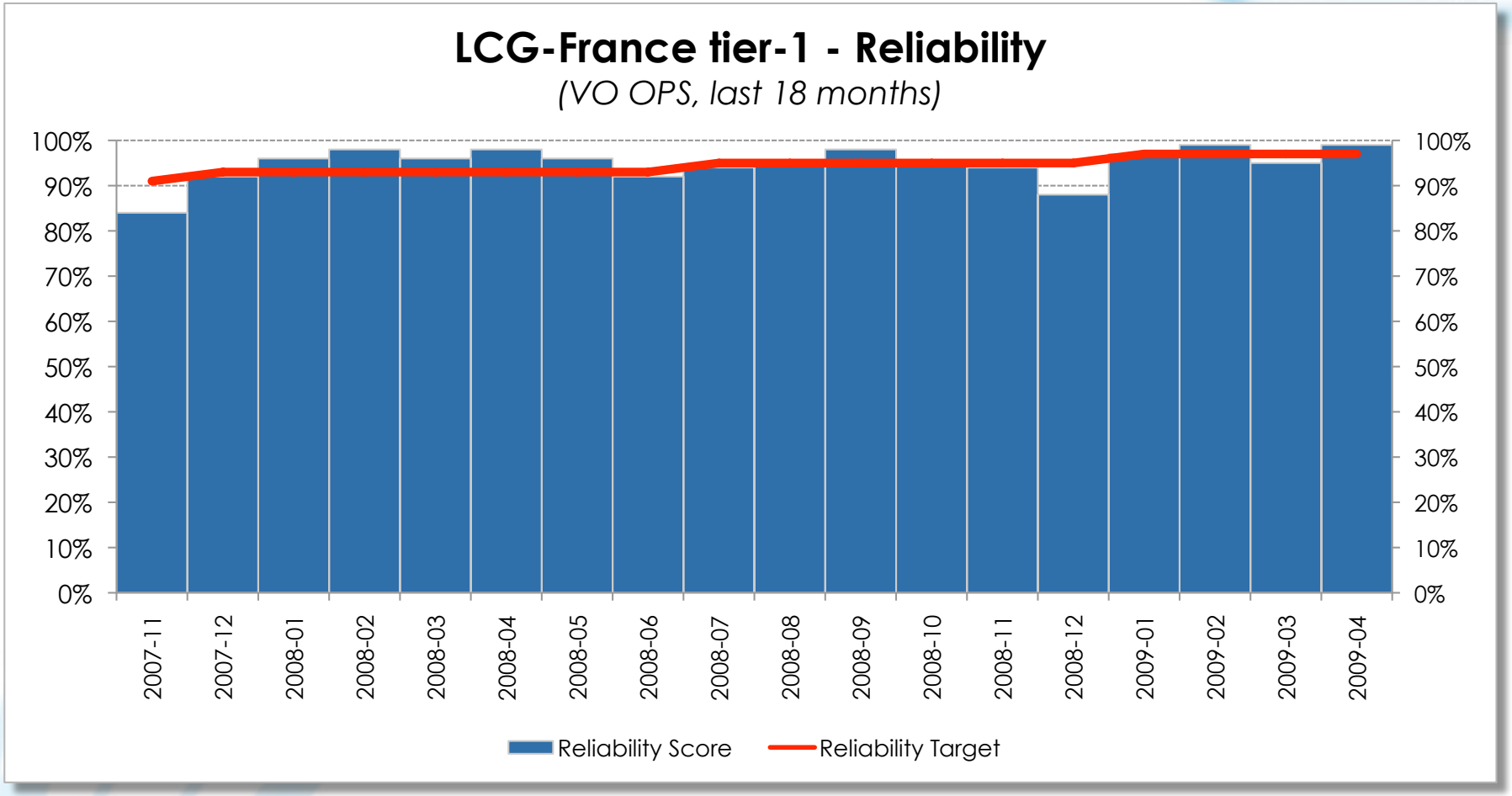
DRAFT 1/15

Source: https://edms.in2p3.fr/document/I-014543

- Regional TopBDII (information system)
- Regional MonBox (accounting)

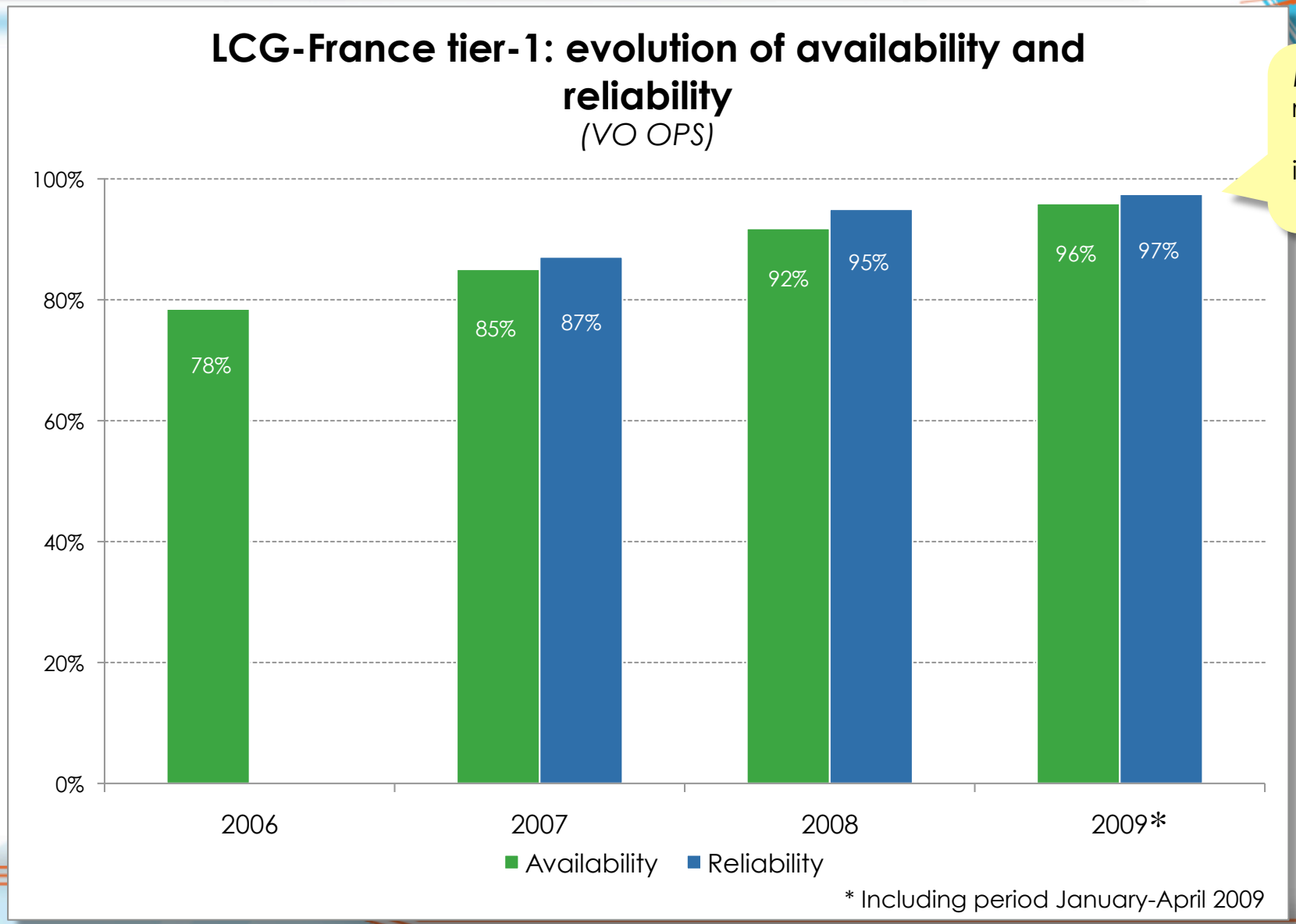
Service targets according to WLCG MoU

MoU targets: reliability



Source: WLCG reliability reports

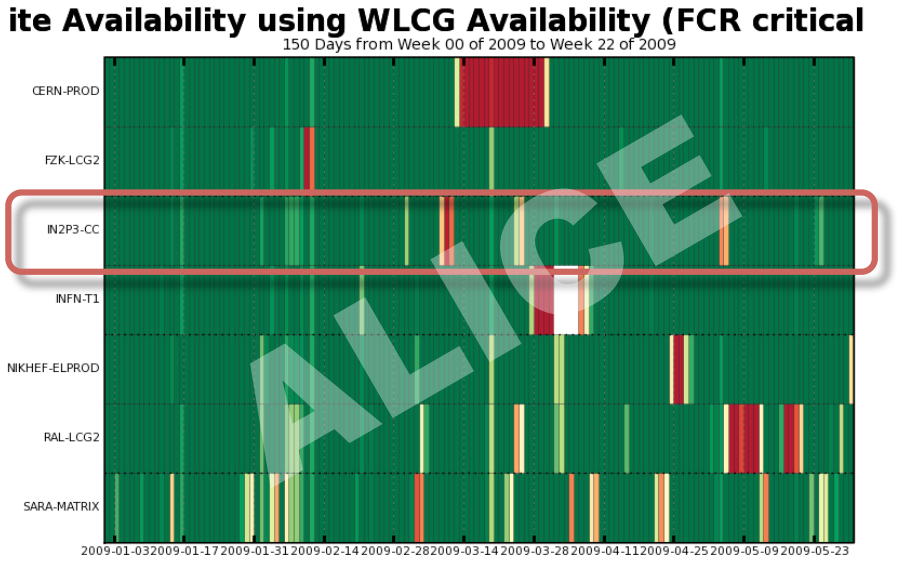
MoU targets: reliability

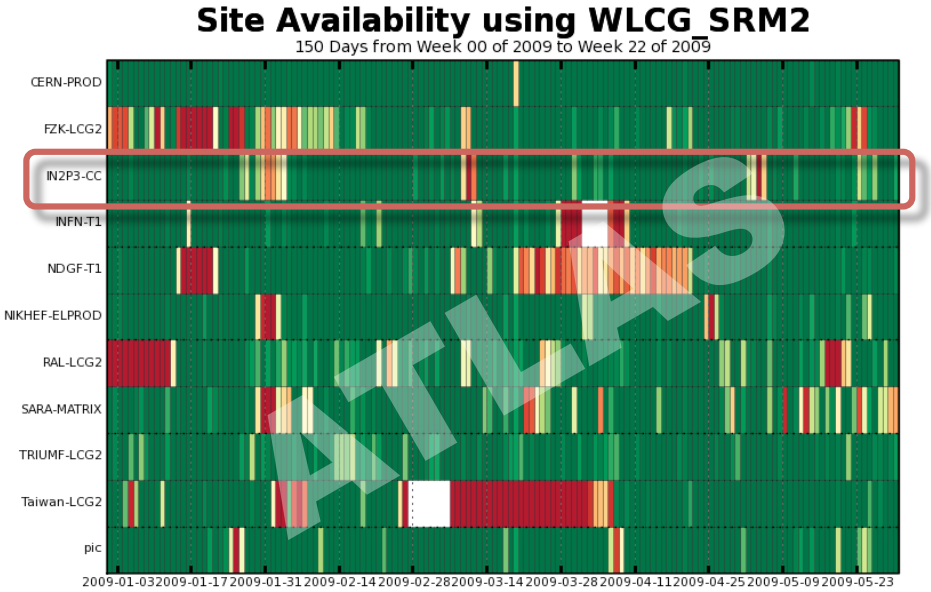


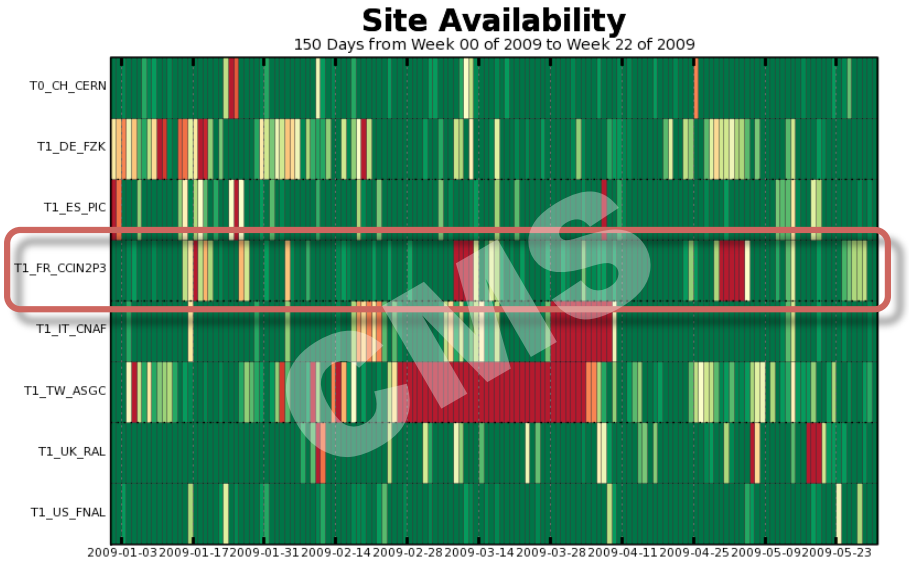
MoU target is a reliability of 98% of the time, integrated over the year

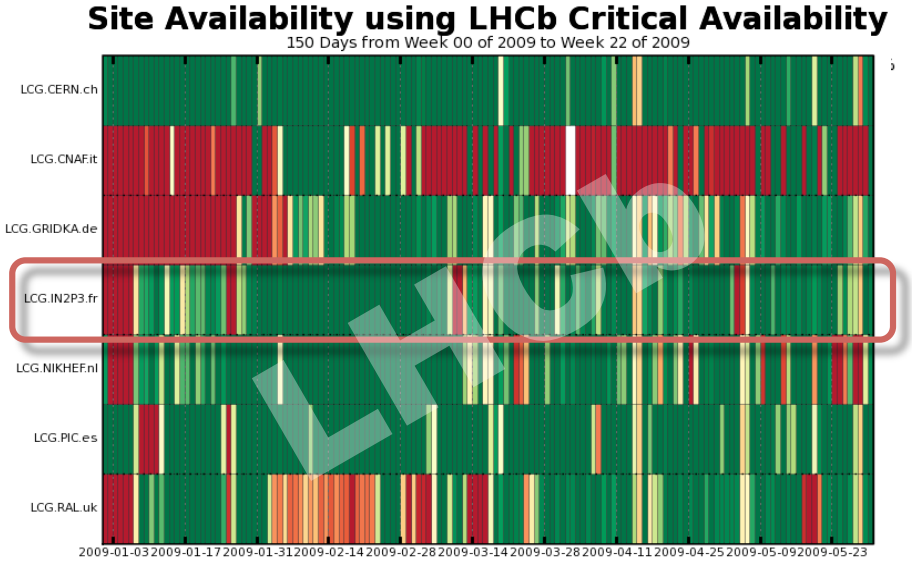
Source: WLCG reliability reports

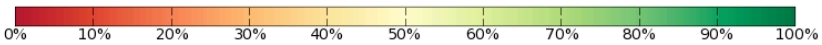
MoU targets: availability tier-1s

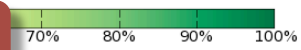












Period: Jan 1st – May 31st, 2009

Source: LHC experiments dashboard http://dashboard.cern.ch/

MoU targets: critical incident handling

- Few real alarms over the last 10 months
  - Reasonably good response time to them
- Details in Suzanne's talk

Transition to the national grid

National grid operations

- CCIN2P3 intends to continue playing a leading role in the operations of the national grid
  - Necessary for maintaining a good quality of service for all the experiments using the grid, in particular for the LHC experiments
  - Role of national grid operator for EGEE is currently completely fulfilled by CCIN2P3, but in the future this workload is expected to be shared by all the sites involved in and benefiting from the national grid
    - *Required tools are already available and grid operations portal is evolving in this direction to ease this task further*

Availability & Reliability – EGEE Regions

Weighted score over **15 EGEE sites** in France, out of which **10 are hosted by IN2P3** laboratories

GRIF federation, composed of 6 sites (5 IN2P3, 1 CEA/DSM/Irfu), counts for 1 EGEE site in this score.

Details in backup slides.

Jan 09

Region
CERN
France
UKI
Russia
AsiaPacific
SouthEasternE
CentralEurope
NorthernEurop
Italy
SouthWestern
GermanySwitz

Feb 09

Region
CERN
France
AsiaPacific
UKI
Italy
SouthEasternEurop
NorthernEurope
SouthWesternEurop
Russia
GermanySwitzerlan
CentralEurope

Mar 09

Region
UKI
CentralEurope
SouthWesternEurope
Russia
France
CERN
GermanySwitzerland
SouthEasternEurope
NorthernEurope
Italy
AsiaPacific

Apr 09

Region
France
SouthWesternEurope
UKI
NorthernEurope
SouthEasternEurope
GermanySwitzerland
CentralEurope
CERN
AsiaPacific
Italy
Russia

May 09

Region	Avail-ability	Reli-ability
France	96 %	98 %
UKI	94 %	96 %
AsiaPacific	94 %	95 %
SouthEasternEurope	93 %	94 %
CERN	93 %	93 %
CentralEurope	92 %	93 %
Italy	91 %	92 %
SouthWesternEurope	90 %	95 %
GermanySwitzerland	89 %	91 %
NorthernEurope	88 %	94 %
Russia	78 %	87 %

Source: https://edms.cern.ch/document/963325

Perspectives

- Consolidate the work being done with storage components
  - Finish integration of the scheduler of tape staging requests and dCache
    - *Fine tune the fair sharing, which could not be exercised during STEP'09*
  - More detailed monitoring and extraction of performance metrics, including the ones required by WLCG
  - Plan migration of Chimera as the dCache catalog
  - Stricter separation of storage spaces for tier-1 and tier-2 to avoid analysis tasks to interfere with tier-1 activities
- Focus on the analysis farm
  - See Yvan's talk, next

Conclusions

- Basic building blocks are in place and with reasonably good redundant configurations, wherever possible
- Improving monitoring is a permanent, and probably never ending, activity
- Pledged computing and storage capacity delivered on time and in accordance to schedule agreed in WLCG, in spite of the power and cooling limitations we have been come through
- STEP'09 exercise exposed the improvements in the site's infrastructure, in particular in the mass storage area
- Although we are confident that we will be able to cope with initial data taking, I'm concerned by the complexity of the infrastructure and the amount of human effort required to operate it

Acknowledgements

- I would like to thank the people from all the teams that worked very hard to reach this stage
  - There are really too many names to fit in one slide

- They would certainly like to have less meetings and more time to do real and interesting work, but the coordination needs of this project are very stringent
  - Both at the site level and globally

- The expertise and involvement of those people are undoubtely the key factor for this project

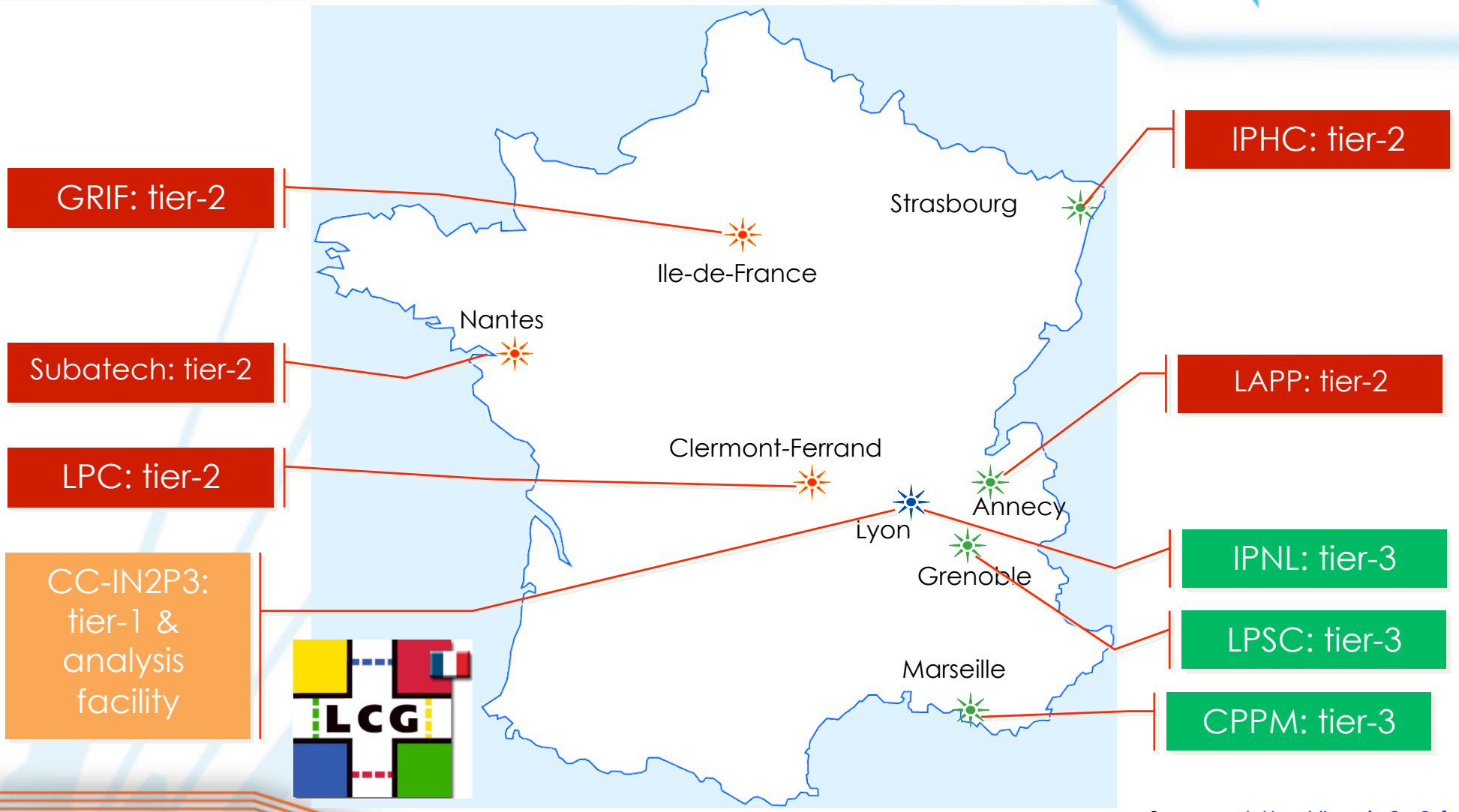
Just before finishing...

- A quote from ATLAS STEP'09 coordinator
  - "IN2P3 are having a storming finish with their shiny new HPSS »
    - *June 11th 2009, summarizing the ATLAS reprocessing activity in STEP'09*

Questions & Comments

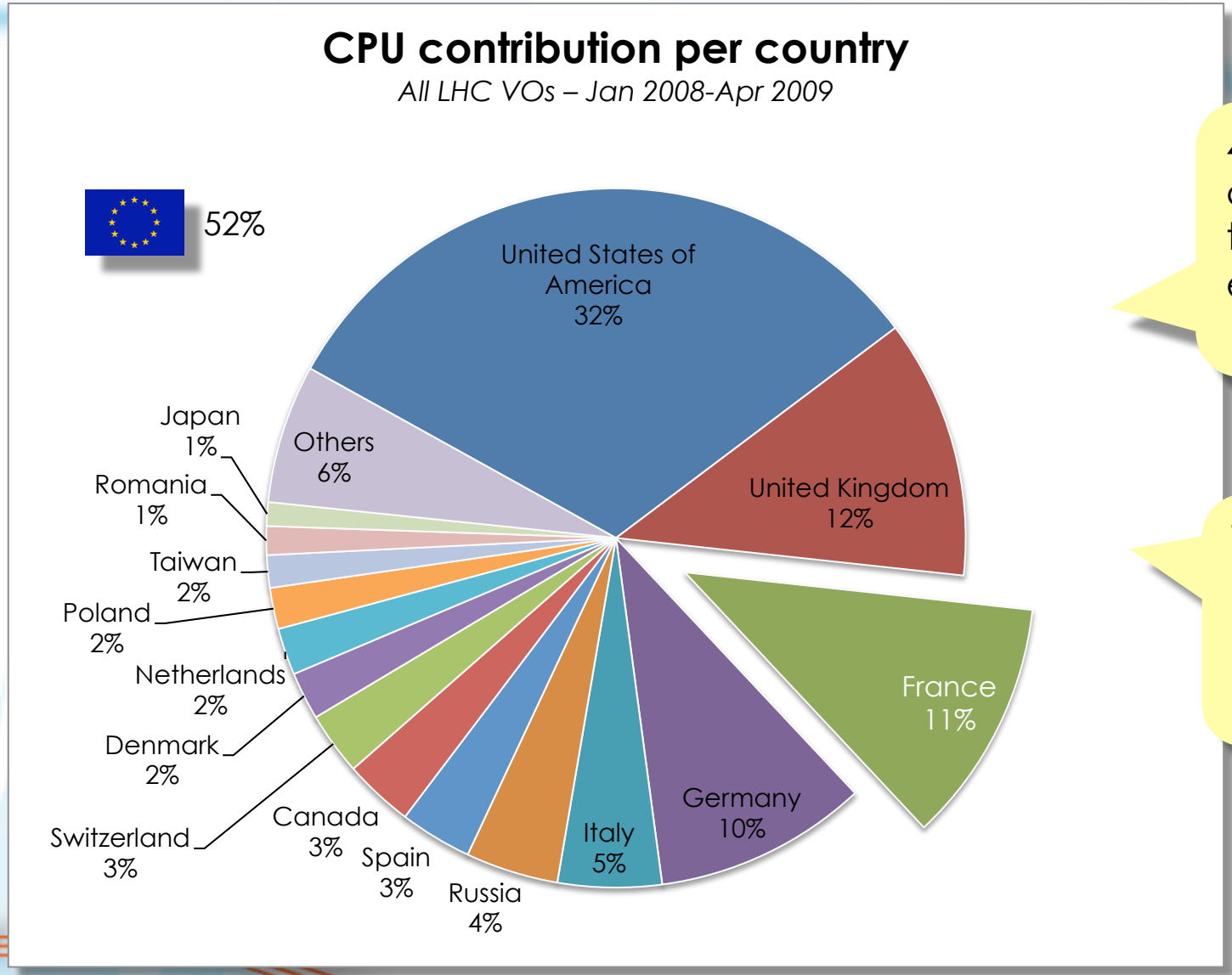
Backup slides

LCG-France



Source: http://lcg.in2p3.fr

LCG-France within WLCG



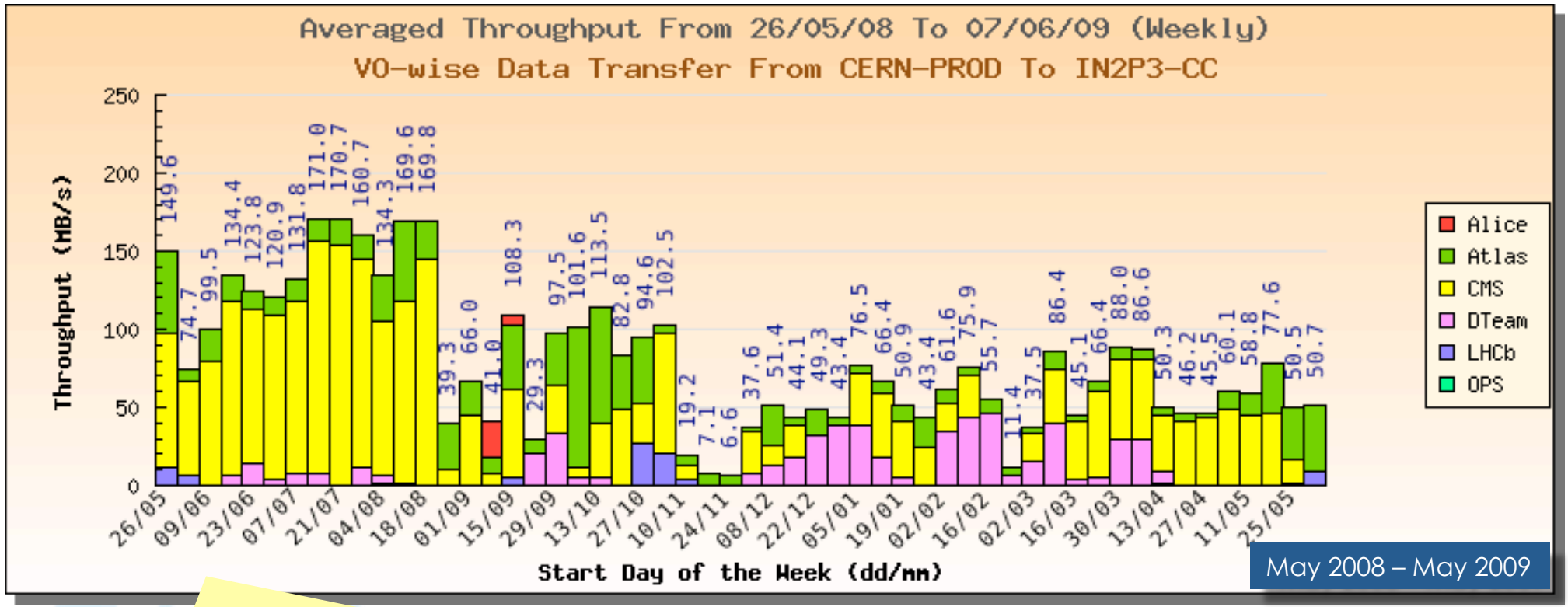
43 countries contribute CPU to the LHC experiments

Top 10 countries crontribute 85% of CPU

Source: EGEE Accounting Portal

Data transfer

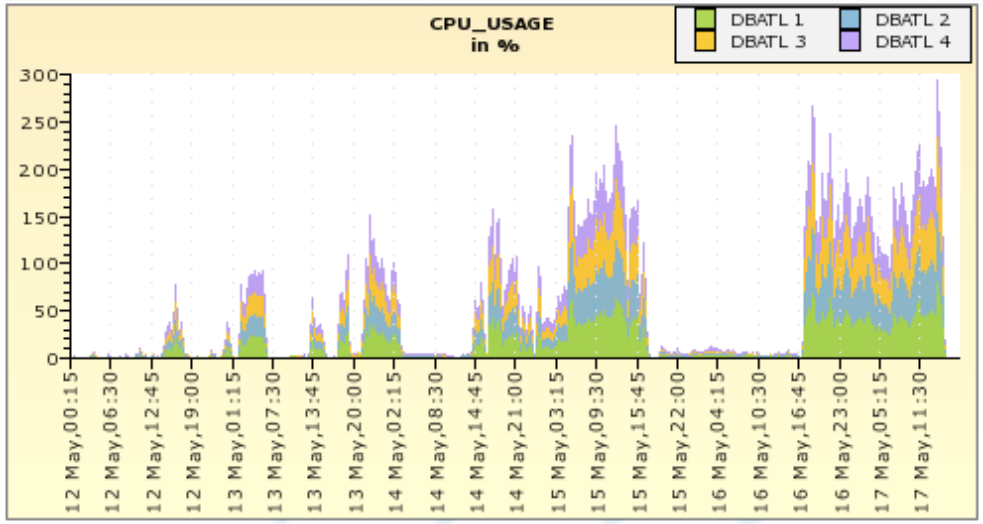
- Tier-0 → CCIN2P3



Sustained rates up to 250 MB/sec demonstrated during CCRC'08 over several days

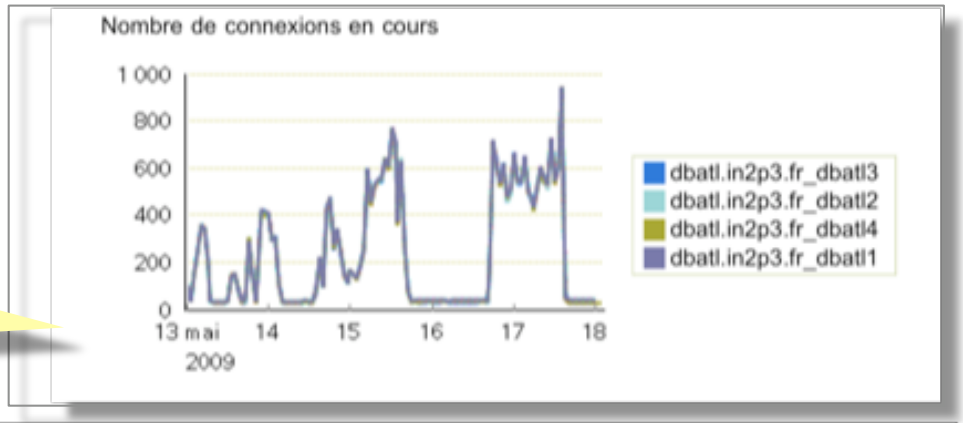
Source: Gridview, http://gridview.cern.ch

ATLAS: conditions database tests

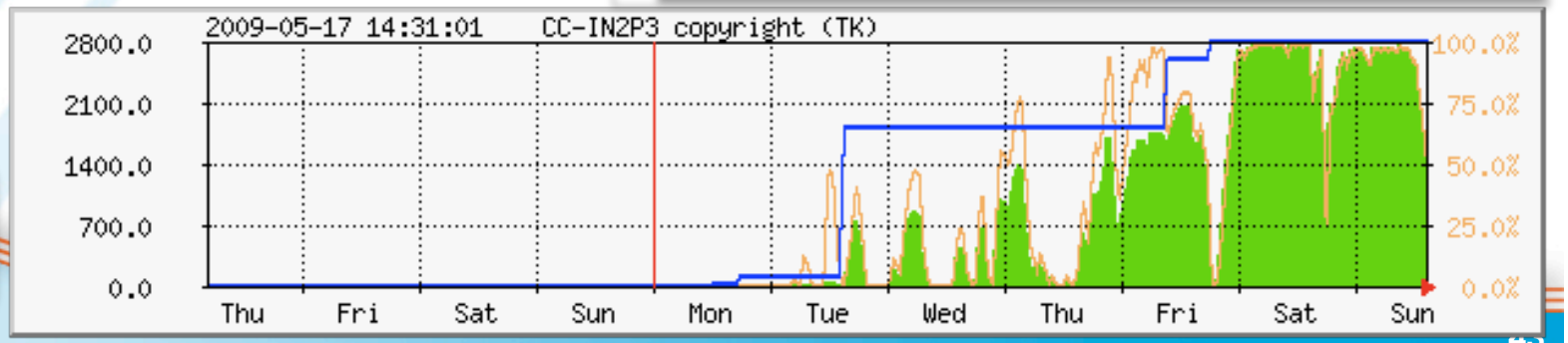


CPU load on the 4 Oracle servers. 200% on average.

Max 400% never reached.



Number of simultaneous connexions to the databases. To be compared to the number of jobs in execution, below.



Tests performed from May 12th to 17th 2009

CE configuration (as of 14/06/2009)

Tier Level	CE name	ALICE	ATLAS	CMS	LHCb
Tier-1	cclcgceli01	✓	✓		
	cclcgceli02	✓	✓		
	cclcgceli03			✓	✓
	cclcgceli04			✓	✓
Tier-2	cclcgceli05	✓	✓	✓	✓
	cclcgceli06	✓	✓	✓	✓

ATLAS & CMS use the sites for specific tasks (e.g. tier-1 for reprocessing, tier-2 for analysis and MC production). Acces is controled by using VOMS roles.

ALICE & LHCb use those sites indistinctly

EGEE France – availability & reliability

April 2009

Region	Site	Phy. CPU	Log. CPU	KSI2K	Avail-ability	Reli-ability	Availability History Jan-09	Feb-09	Mar-09
France (France)									
	AUVERGRID	42	42	75	98 %	98 %	98 %	100 %	98 %
	CGG-LCG2	80	80	49	91 %	91 %	46 %	73 %	96 %
	ESRF	16	16	43	86 %	86 %	83 %	89 %	98 %
	GRIF	3,338	2,180	3,908	100 %	100 %	99 %	100 %	92 %
	IBCP-GBIO	10	10	5	55 %	97 %	60 %	67 %	94 %
	IN2P3-CC	1,074	4,296	3,832	99 %	99 %	97 %	98 %	90 %
	IN2P3-CC-T2	1,074	4,296	3,832	99 %	99 %	96 %	97 %	89 %
	IN2P3-CPPM	358	358	537	98 %	99 %	99 %	97 %	95 %
	IN2P3-IPNL	452	440	656	98 %	99 %	98 %	96 %	98 %
	IN2P3-IRES	664	628	1,526	84 %	84 %	95 %	96 %	93 %
	IN2P3-LAPP	512	512	1,133	96 %	98 %	94 %	100 %	98 %
	IN2P3-LPC	448	448	802	84 %	99 %	94 %	93 %	99 %
	IN2P3-LPSC	120	112	43	71 %	96 %	97 %	99 %	94 %
	IN2P3-SUBATECH	275	380	803	97 %	97 %	99 %	96 %	99 %
	IPSL-IPGP-LCG2	34	34	41	96 %	96 %	94 %	100 %	96 %

Source: https://edms.cern.ch/document/963325

EGEE France – availability & reliability

Region	Site	Phy. CPU	Log. CPU	KSI2K	Avail-ability	Reli-ability	Availability History Feb-09	Mar-09	Apr-09
France (France)									
	AUVERGRID	42	42	75	90 %	90 %	100 %	98 %	98 %
	CGG-LCG2	80	80	49	84 %	97 %	73 %	96 %	91 %
	ESRF	16	16	43	100 %	100 %	89 %	98 %	86 %
	GRIF	3,336	3,122	5,637	98 %	99 %	100 %	92 %	100 %
	IBCP-GBIO	10	10	5	96 %	99 %	67 %	94 %	55 %
	IN2P3-CC	1,074	4,296	3,832	92 %	97 %	98 %	90 %	99 %
	IN2P3-CC-T2	1,074	4,296	3,832	91 %	97 %	97 %	89 %	99 %
	IN2P3-CPPM	358	358	537	99 %	99 %	97 %	95 %	98 %
	IN2P3-IPNL	452	440	656	98 %	99 %	96 %	98 %	98 %
	IN2P3-IRES	664	628	1,526	98 %	98 %	96 %	93 %	84 %
	IN2P3-LAPP	512	512	1,133	96 %	98 %	100 %	98 %	96 %
	IN2P3-LPC	448	448	802	96 %	96 %	93 %	99 %	84 %
	IN2P3-LPSC	120	112	43	100 %	100 %	99 %	94 %	71 %
	IN2P3-SUBATECH	275	380	803	99 %	99 %	96 %	99 %	97 %
	IPSL-IPGP-LCG2	34	34	41	96 %	96 %	100 %	96 %	96 %

May 2009

Source: https://edms.cern.ch/document/963325