CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

DSM

IN2P3 INSTITUT NATIONAL DE PHYSIQUE NUCLÉAIRE ET DE PHYSIQUE DES PARTICULES

LCG-France Tier-1

Status and Plans

Fabio Hernandez
IN2P3/CNRS Computing Centre - Lyon
fabio@in2p3.fr

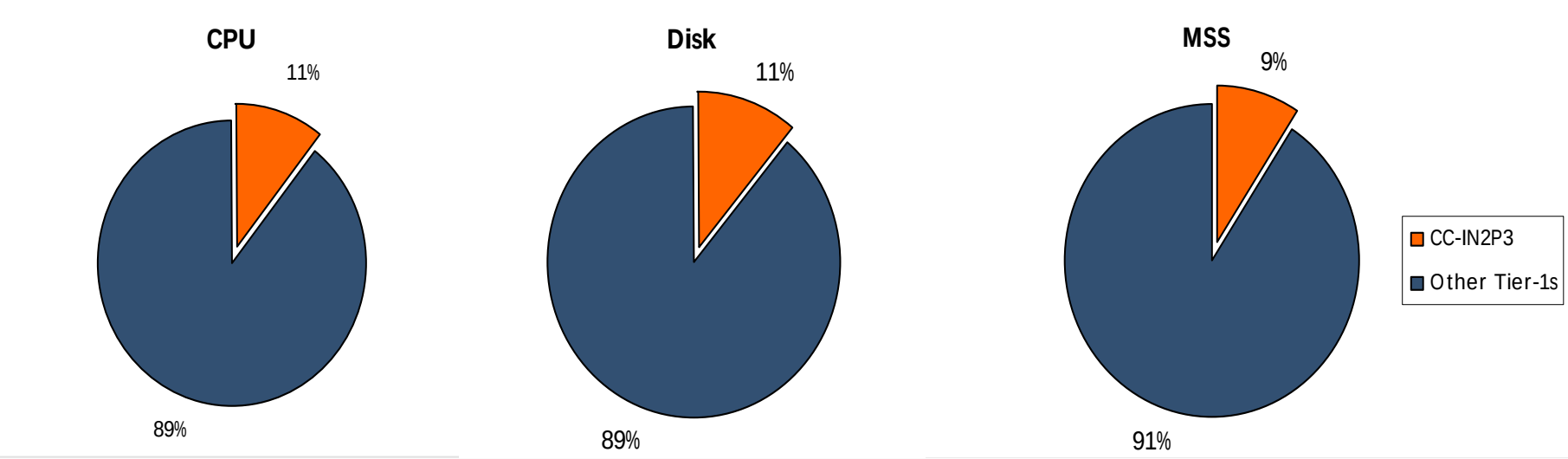
2nd LCG-France Workshop
Clermont, March 13th-14th 2007

Contents

- Activities and contribution during 2006
- Plans for 2007
- Conclusions
- Questions

Contribution

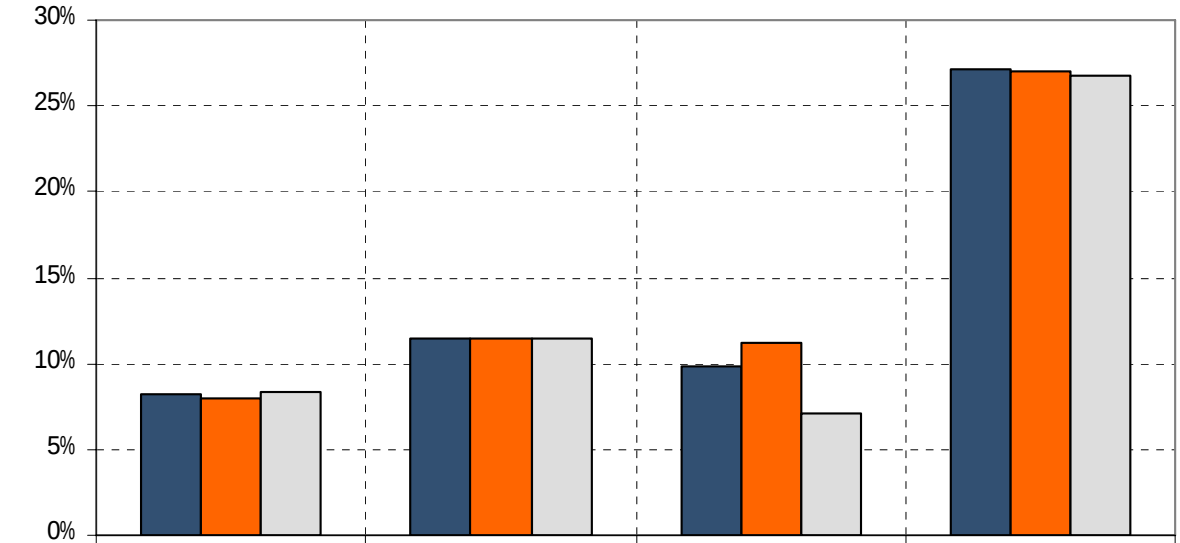
- Revised planned contribution of LCG-France Tier-1
  - % of required resources for all tier-1s in 2008 (experiment's requirements as of March 2007)



Source: Comparison of New Requirements with Current Pledges – 24/10/2006

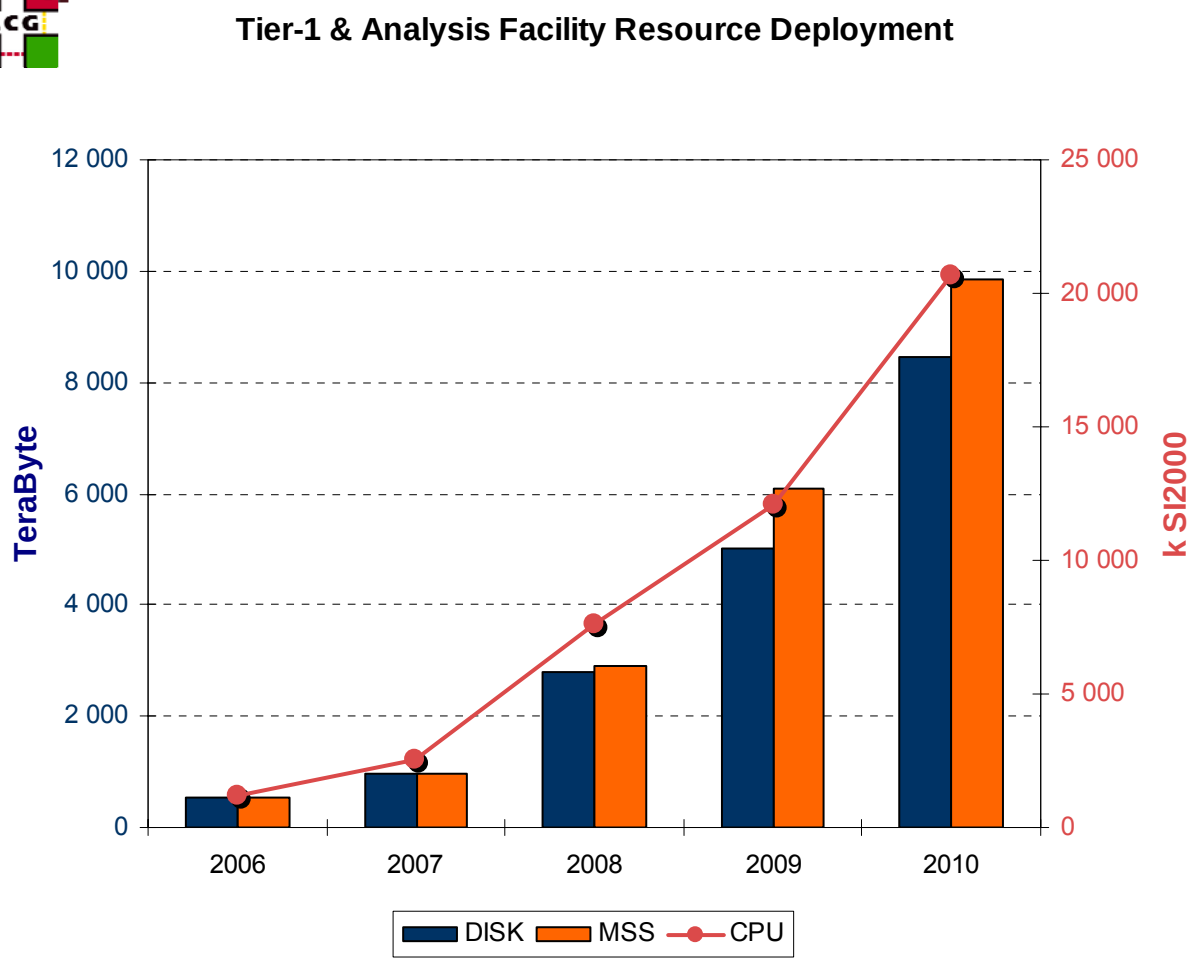
Contribution (cont.)

- Revised planned contribution of LCG-France tier-1
  - % of required resources in all tier-1s in 2008



	Alice	Atlas	CMS	LHCb
CPU	8%	11%	10%	27%
Disk	8%	11%	11%	27%
MSS	8%	11%	7%	27%

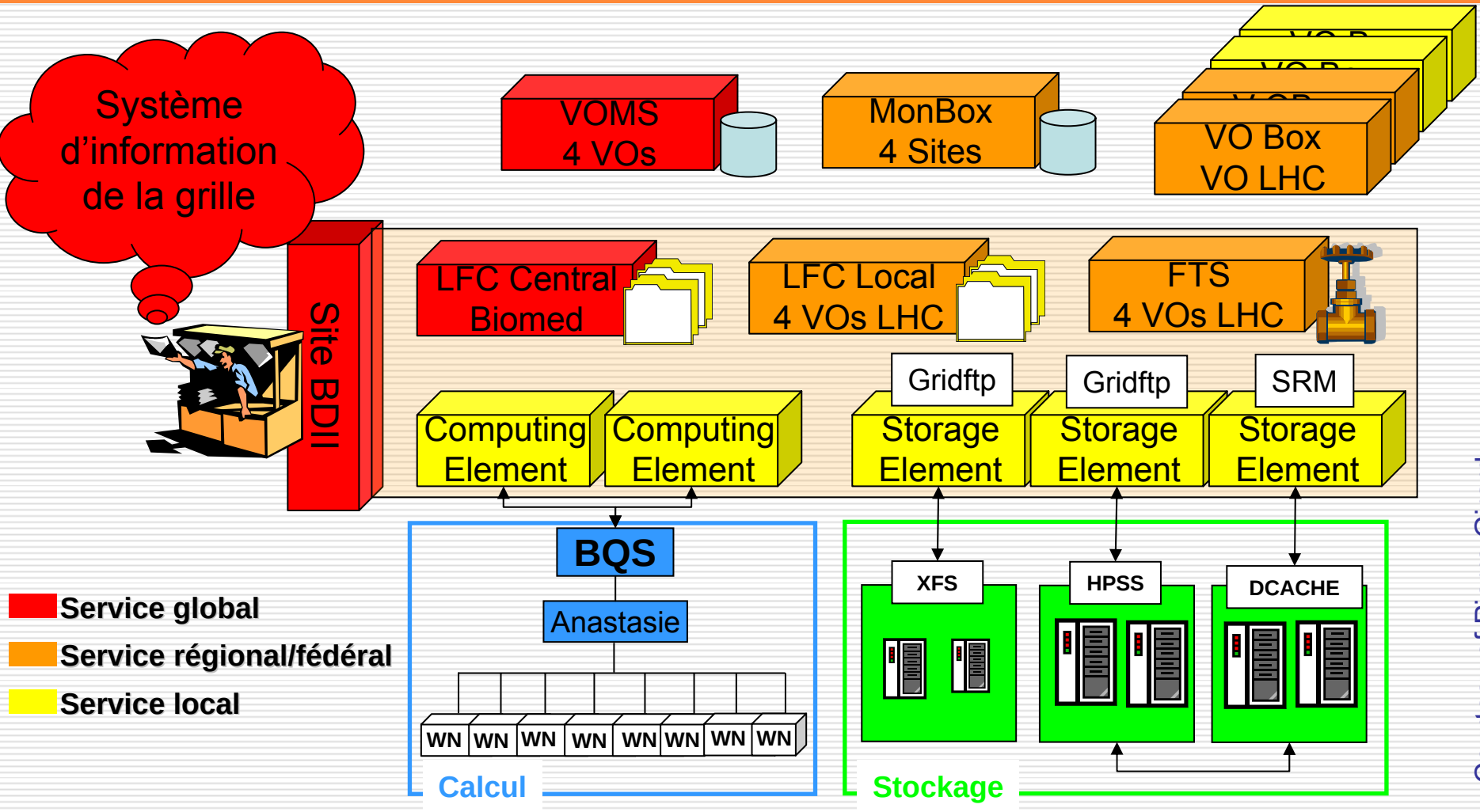
Planned Evolution



Increase rate over the period 2006-2010:

CPU: x 17
DISK: x 16
MSS: x 18

Site overview (current status)



Courtesy of Pierre Girard

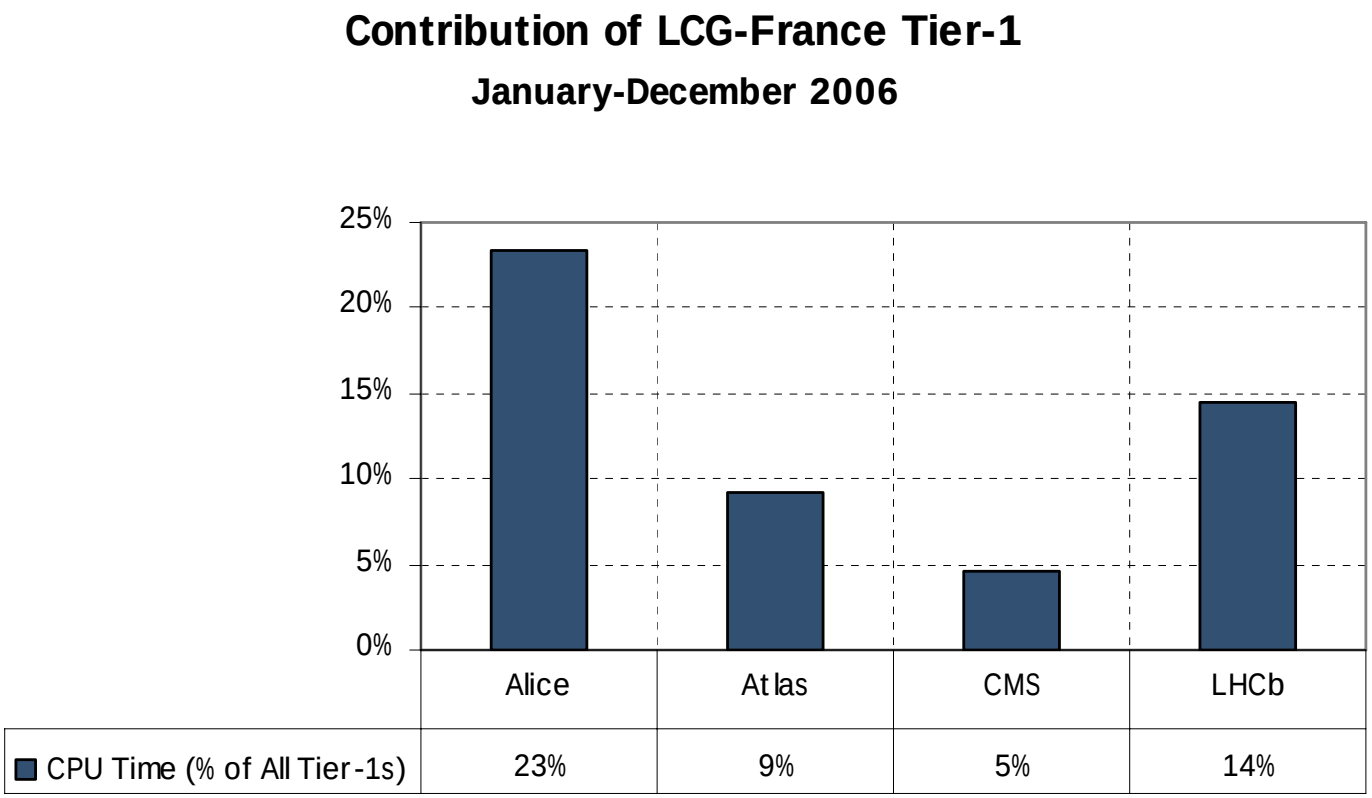
Site overview (cont.)

- Operating also several grid services for non-LHC VOs

Grid Service	alice	atlas	cms	lhcb	auvergrid	biomed	calice	cdf	dteam	dzero	egeode	embrace	esr	hone	ilc	ops	virgo
CE	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓
dCache/SRM SE	✓	✓	✓	✓					✓							✓	
Classic SE	✓	✓	✓	✓		✓			✓	✓	✓		✓	✓	✓	✓	✓
Local LFC	✓	✓	✓	✓													
VO Box	✓	✓	✓	✓				✓									
FTS	✓	✓	✓	✓													
Central LFC						✓											
RLS/RMC						✓											
VOMS					✓	✓					✓	✓					

Contribution in 2006

- CPU time contributed by the LCG-France tier-1 in 2006
  - % of CPU time (grid and non-grid) used by the experiments in all the tier-1s



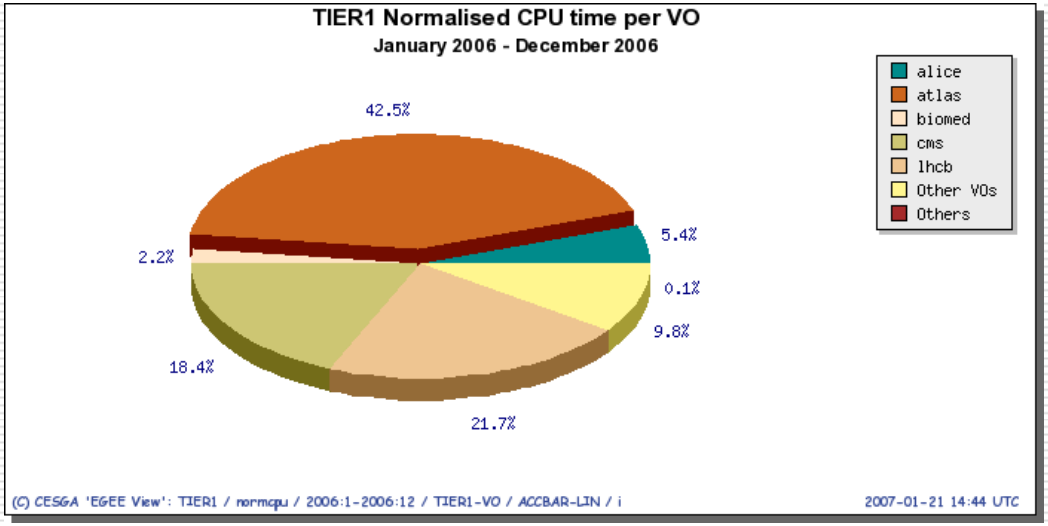
Source: WLCG Accounting Report Tier-1 Centres + CERN

The CC-IN2P3 contribution to the global effort in 2006 was 10% of the total CPU used by the 4 experiments in all the tier-1s.

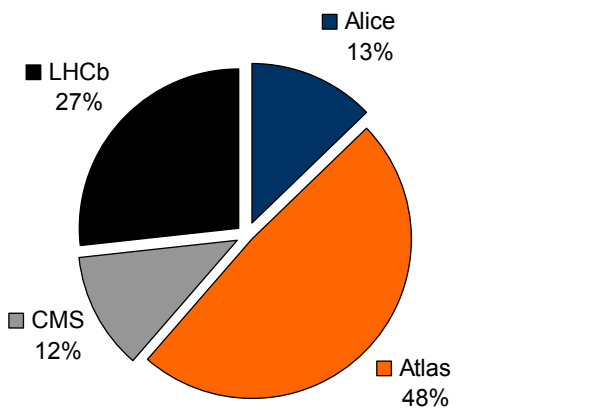
Contribution in 2006 (cont.)

- CPU utilisation by LHC experiments at all the tier-1s and at CC-IN2P3

All Tier-1s
(*does not include non-grid usage of some sites*)



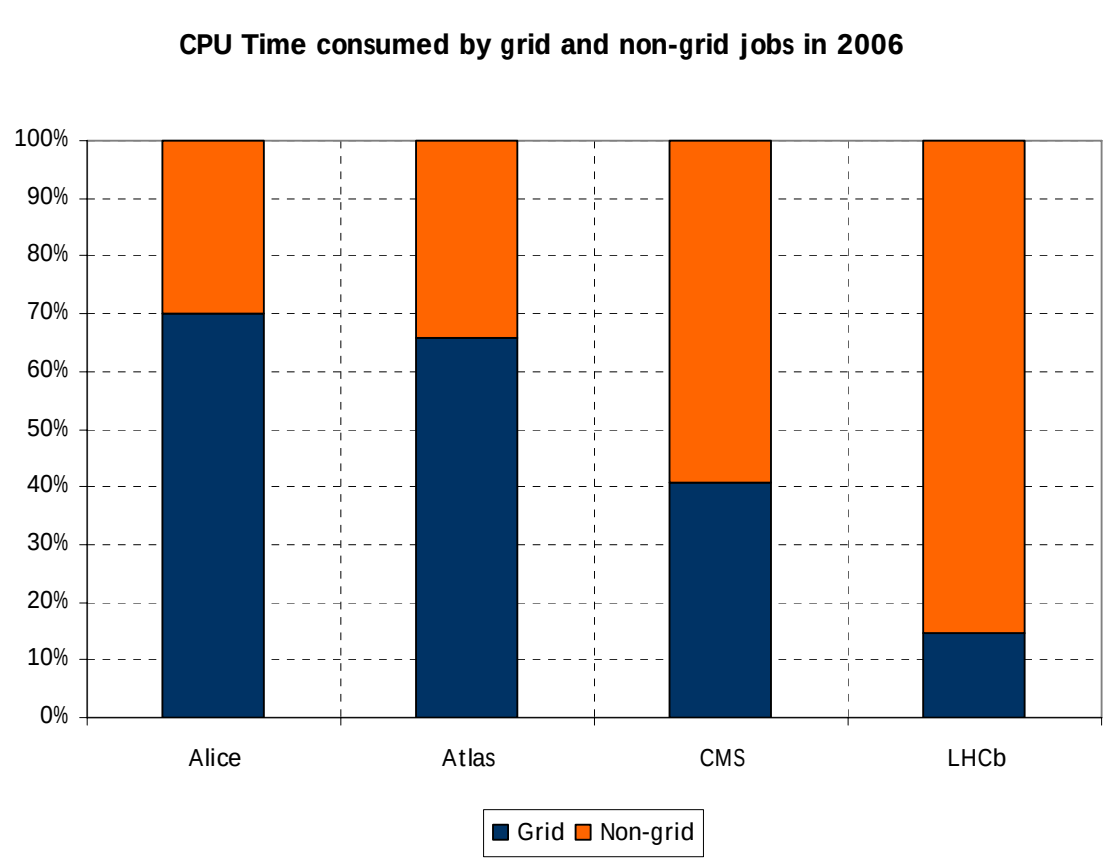
CC-IN2P3 (grid and non-grid)



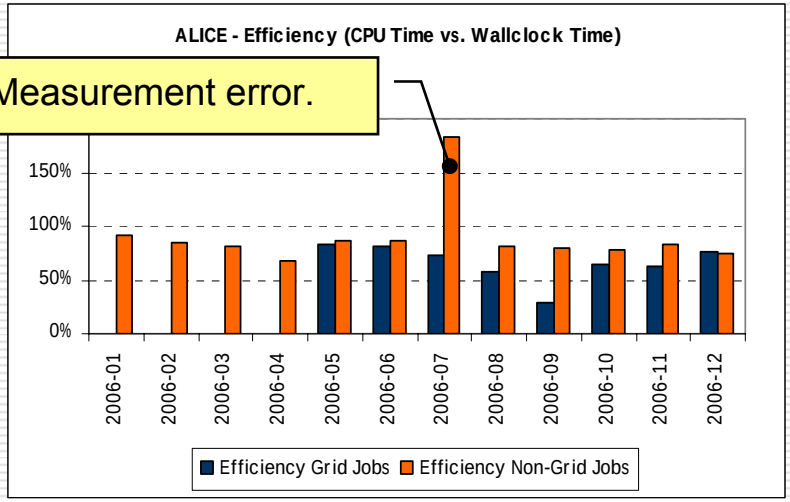
Source: http://www3.egee.cesga.es/gridsite/accounting/CESGA/tier1_view.html

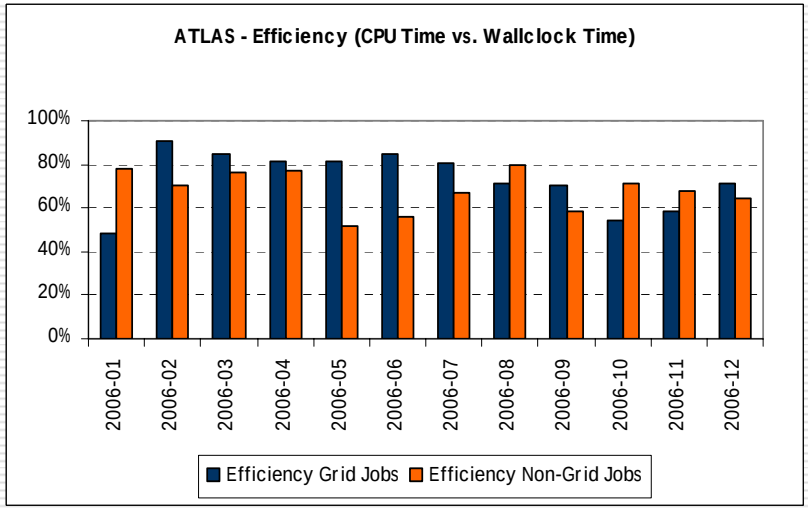
Grid vs. non-grid usage

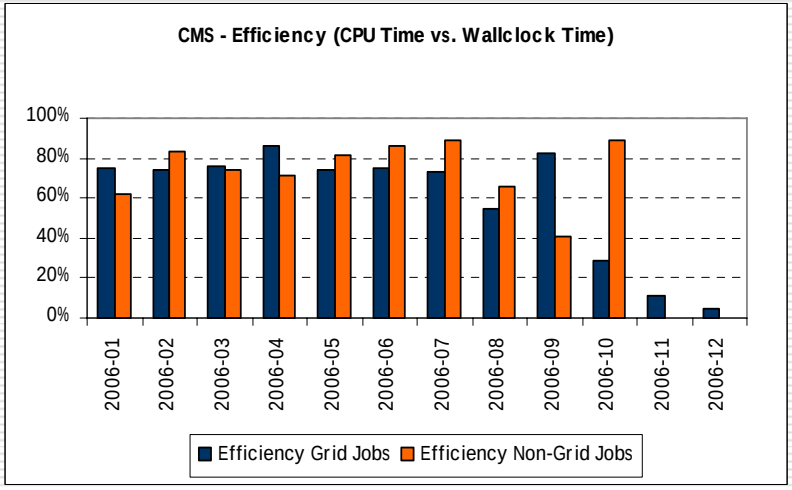
- Site usage (grid vs. non-grid) greatly varies from one experiment to another
  - Both in terms of consumed capacity and number of jobs

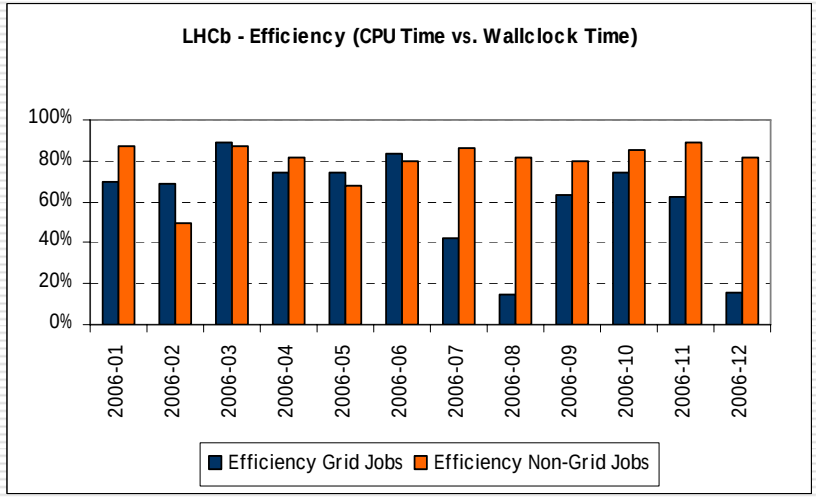


Efficiency (CPU time vs. wallclock)

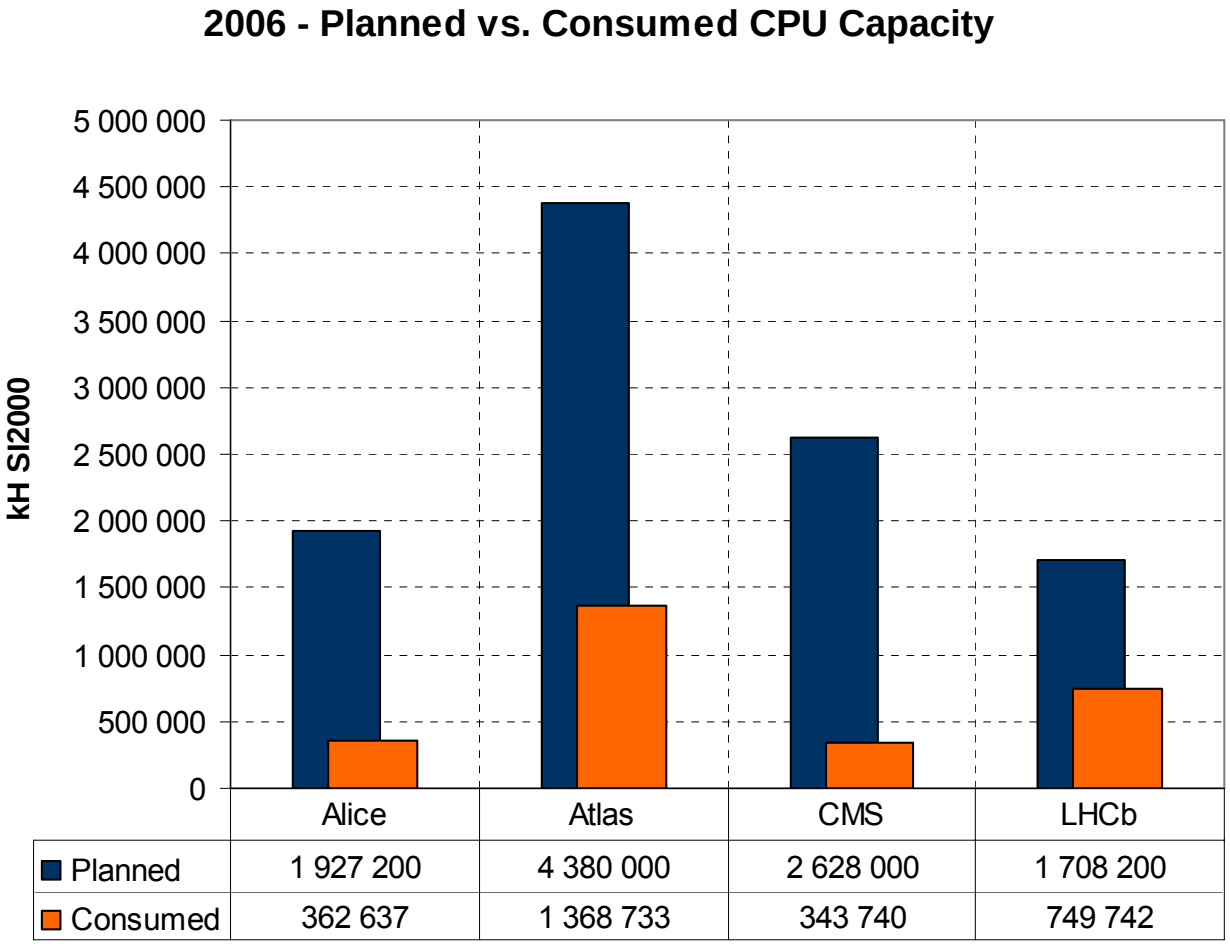








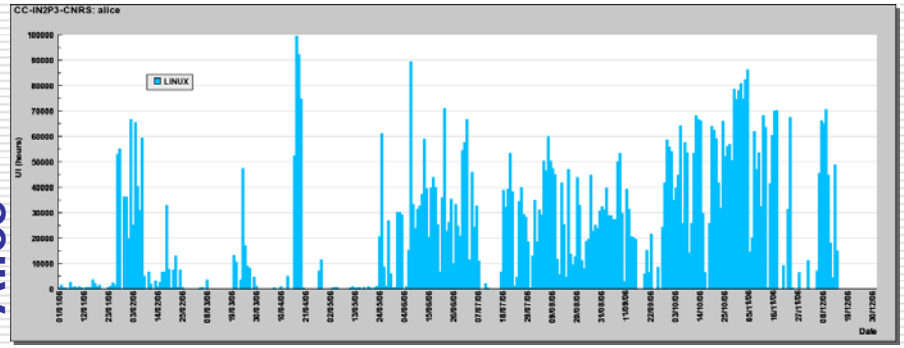
CPU planned vs. actual consumption



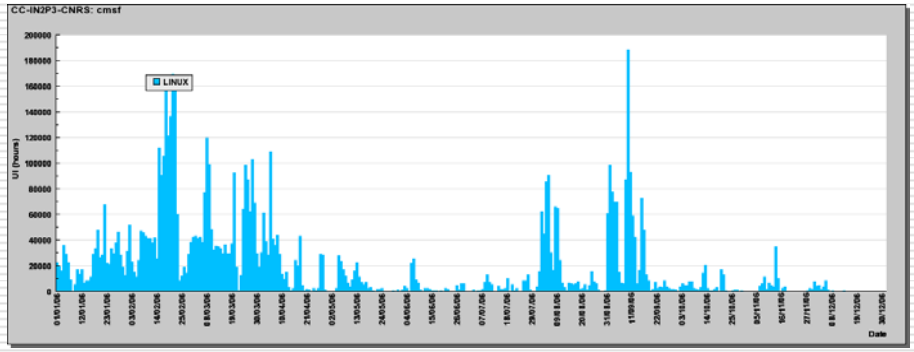
	Alice	Atlas	CMS	LHCb
Planned	1 927 200	4 380 000	2 628 000	1 708 200
Consumed	362 637	1 368 733	343 740	749 742

Observed Experiment Activity at the Site

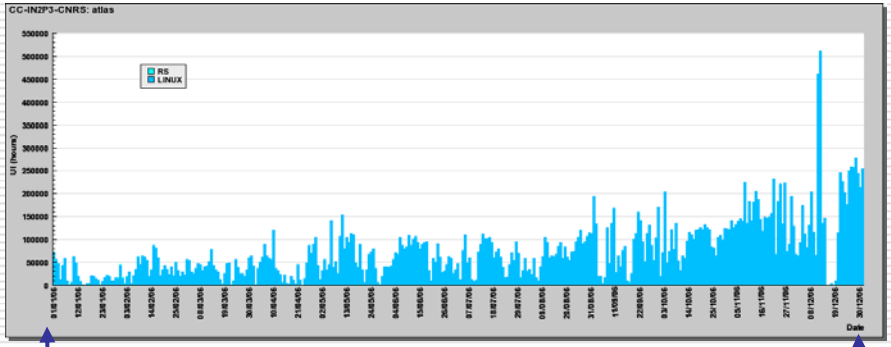
- LHC experiments CPU activity vs. time
  - NOTE: Y axis scale is not the same in all plots



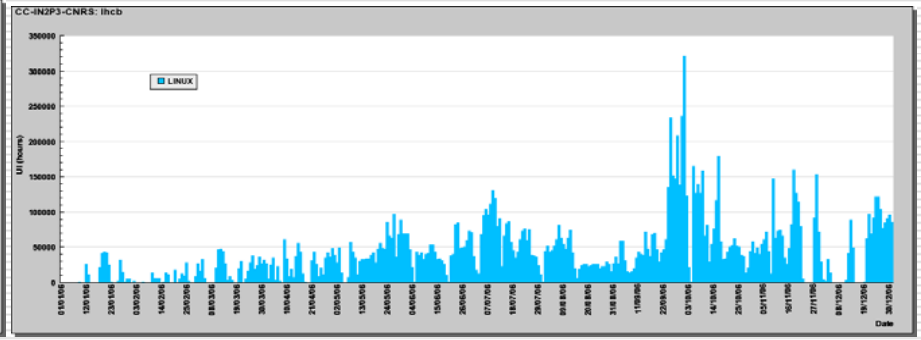
Alice



CMS



Atlas

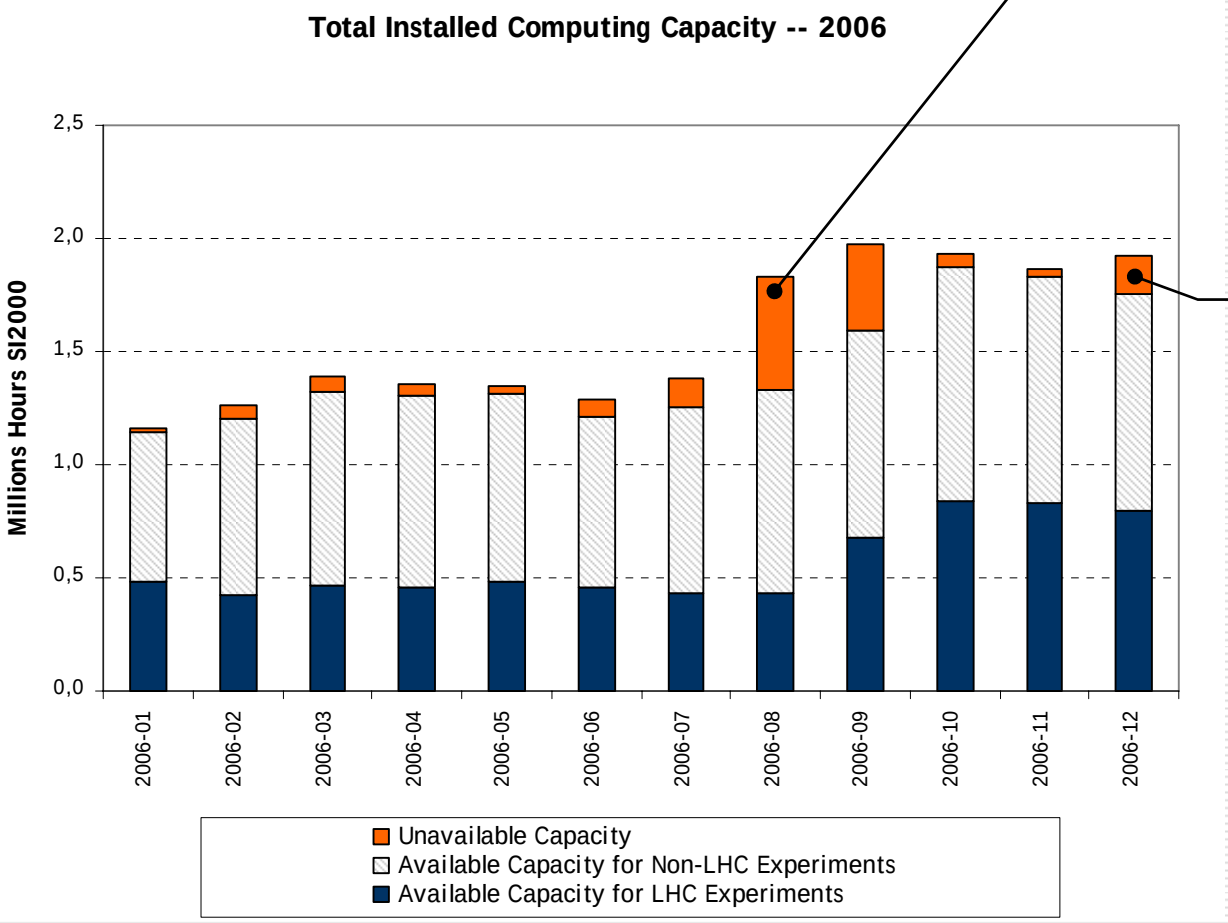


LHCb

Jan 2006

Dec 2006

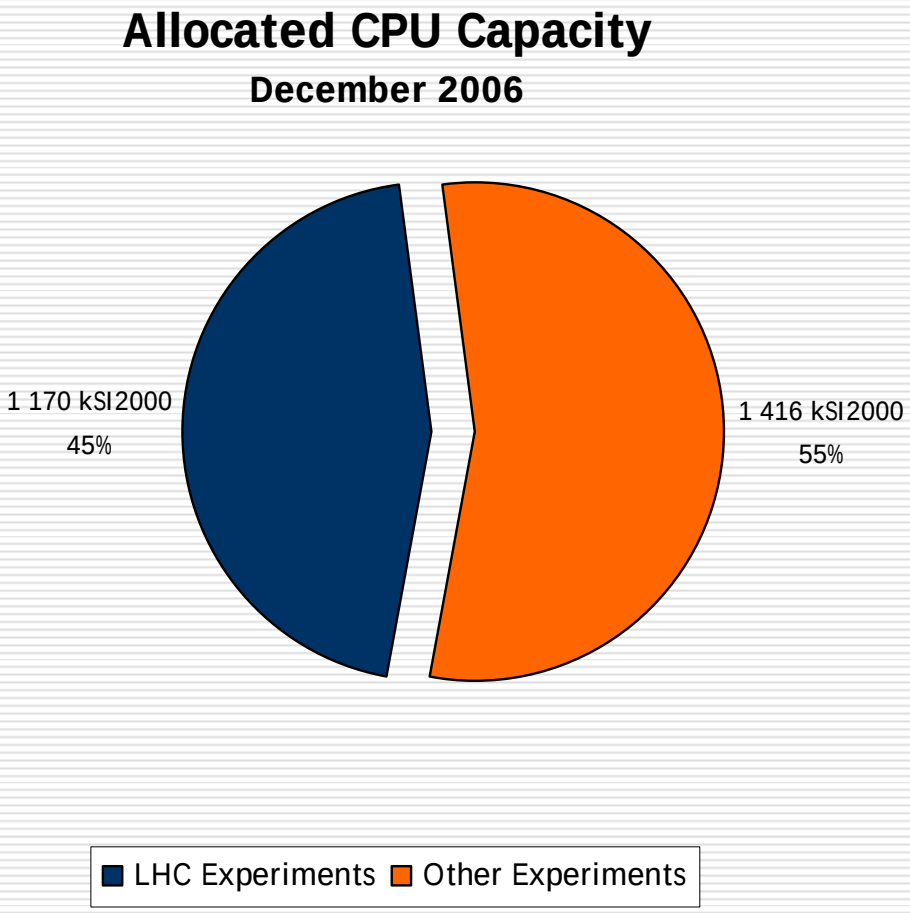
Delivered CPU capacity



Several service interruptions in August and September due to incidents with the cooling or power infrastructure

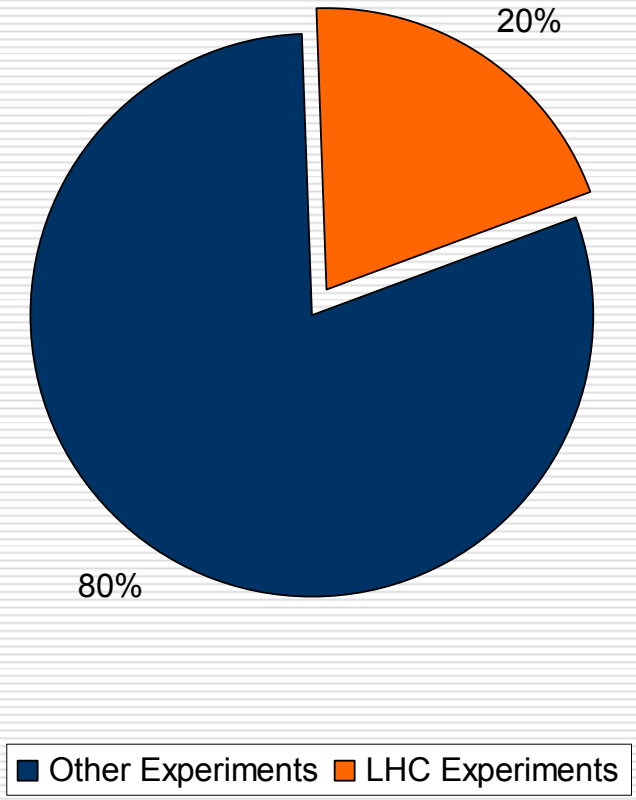
4 days-long scheduled complete shutdown of the site for replacing some central electric and cooling equipement

CPU capacity - allocation



CPU capacity - consumption

- CPU time consumed by LHC experiments
  - % of consumed CPU time by all experiments at CC-IN2P3

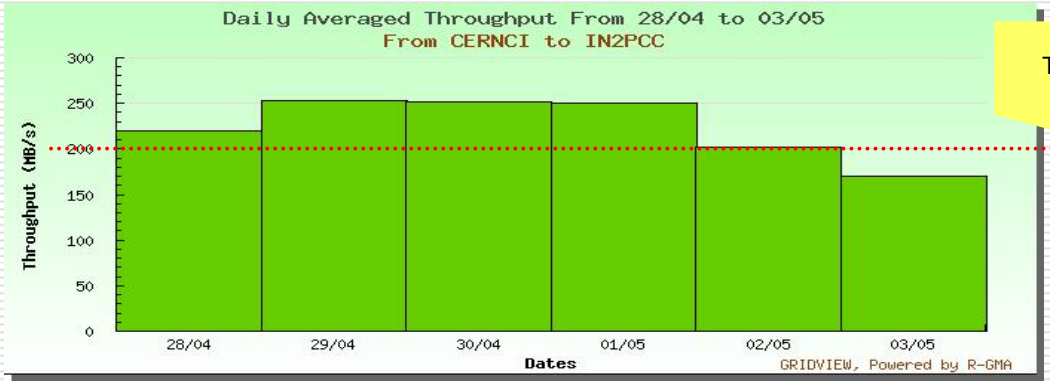


Delivered Storage

- Disk storage capacity
  - Delivered 34% (180 TB out of 520 TB planned)
  - More on this later
- Tape storage capacity
  - Installed capacity (as planned) of 535 TB (of which 73% was actually used)

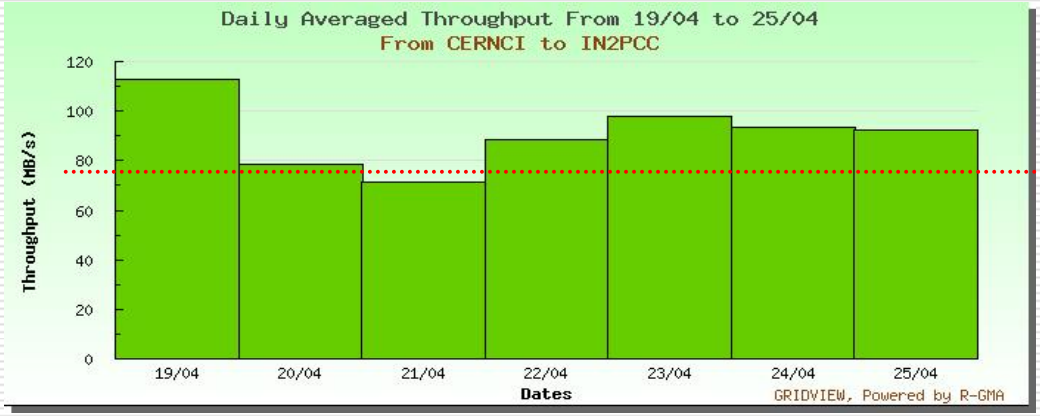
Data transfer exercises

- CERN → CC-IN2P3 (disk)
  - April 2006



Target: 200 MB/sec

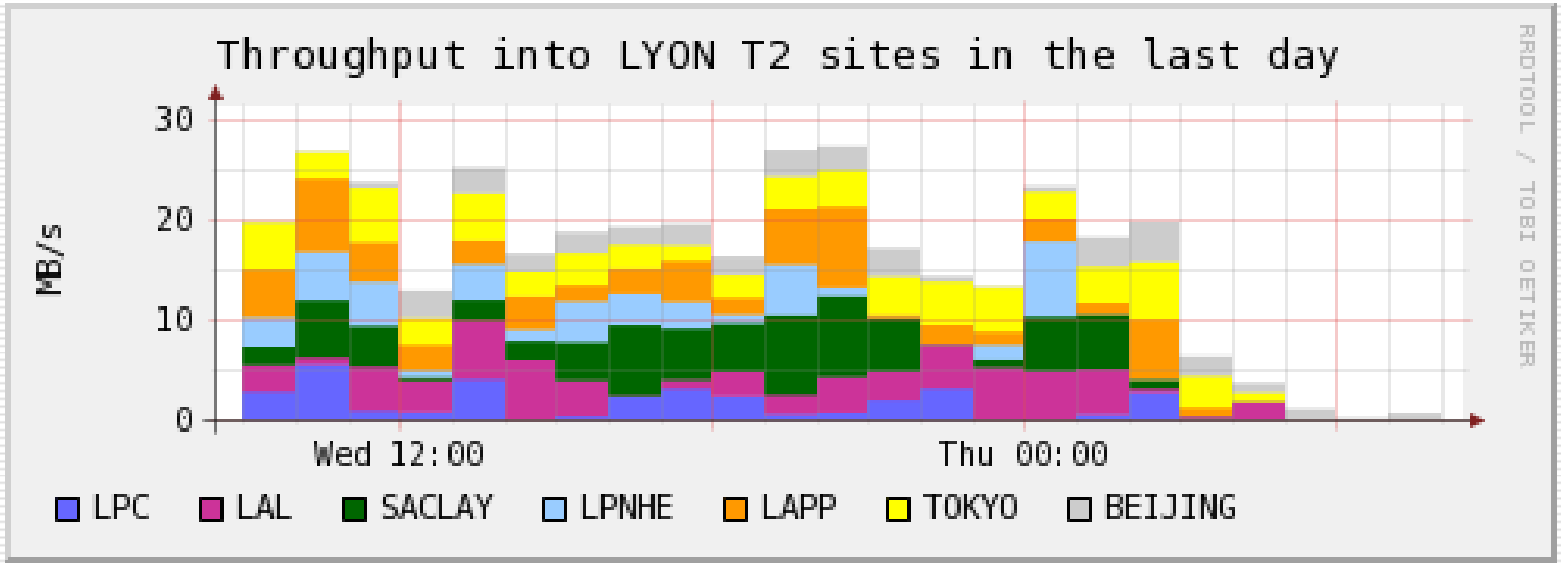
- CERN → CC-IN2P3 (MSS)
  - April 2006



Target: 75 MB/sec

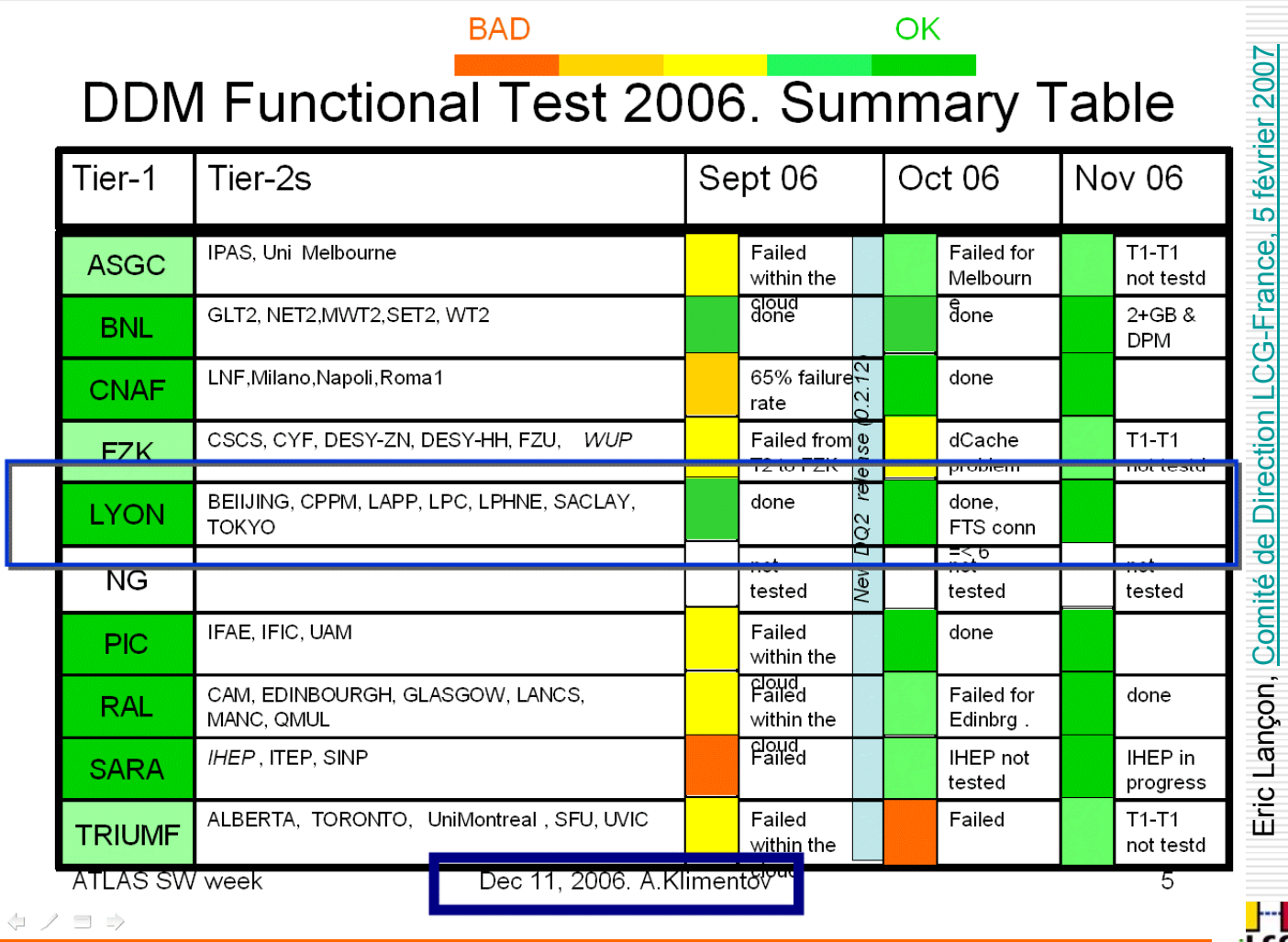
Data transfer exercises (cont.)

- ATLAS: data transfer tests from Tier-1 to linked Tier-2s
  - July 7th 2006



Data transfer exercises (cont.)

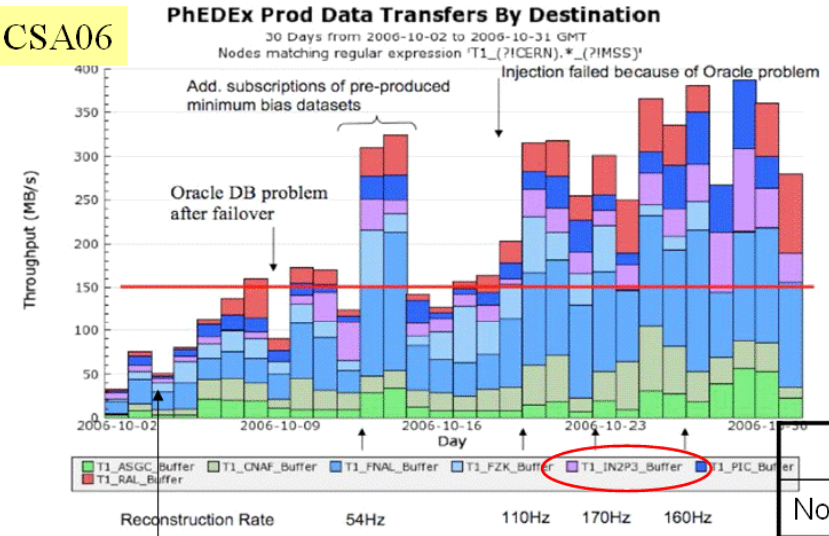
- ATLAS

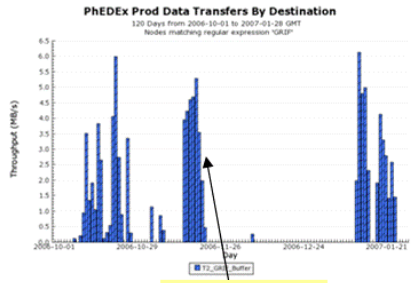


Data transfer exercises (cont.)

Transfert

CSA06





T2-GRIF

Figure 14: The rate of data transferred between the Tier-0 to the Tier-1 centers in MB per second.

	CC-IN2P3
Nominal (CSA) rate	25 MB/s
Last 30 Day average	23 MB/s
Last 15 Day average	34 MB/s
Outage (Days)	1
MSS used	YES

CCIN2P3: Pb serveurs de disk

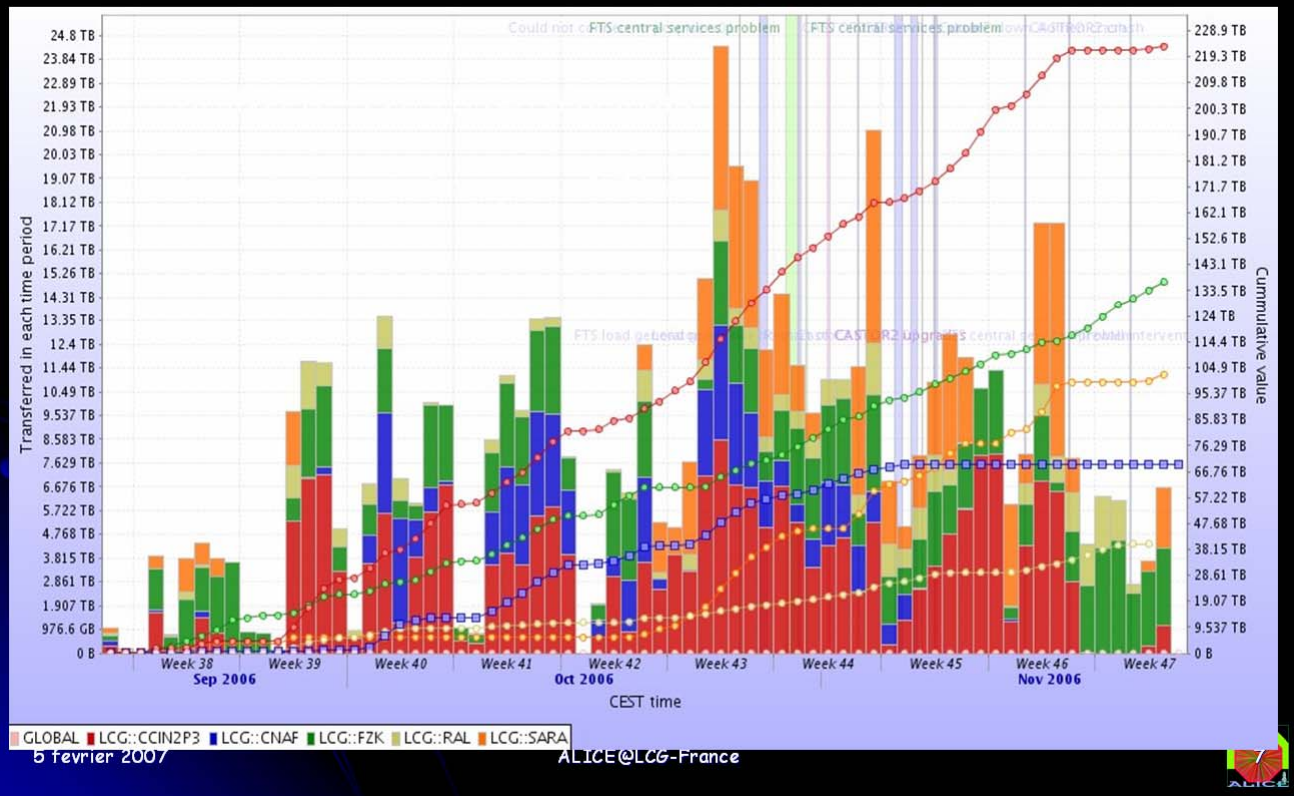
C. Charlot, Calcul CMS, LCG-DIR, fév 2007

Claude Charlot, Comité de Direction LCG-France, 5 février 2007

Data transfer exercises

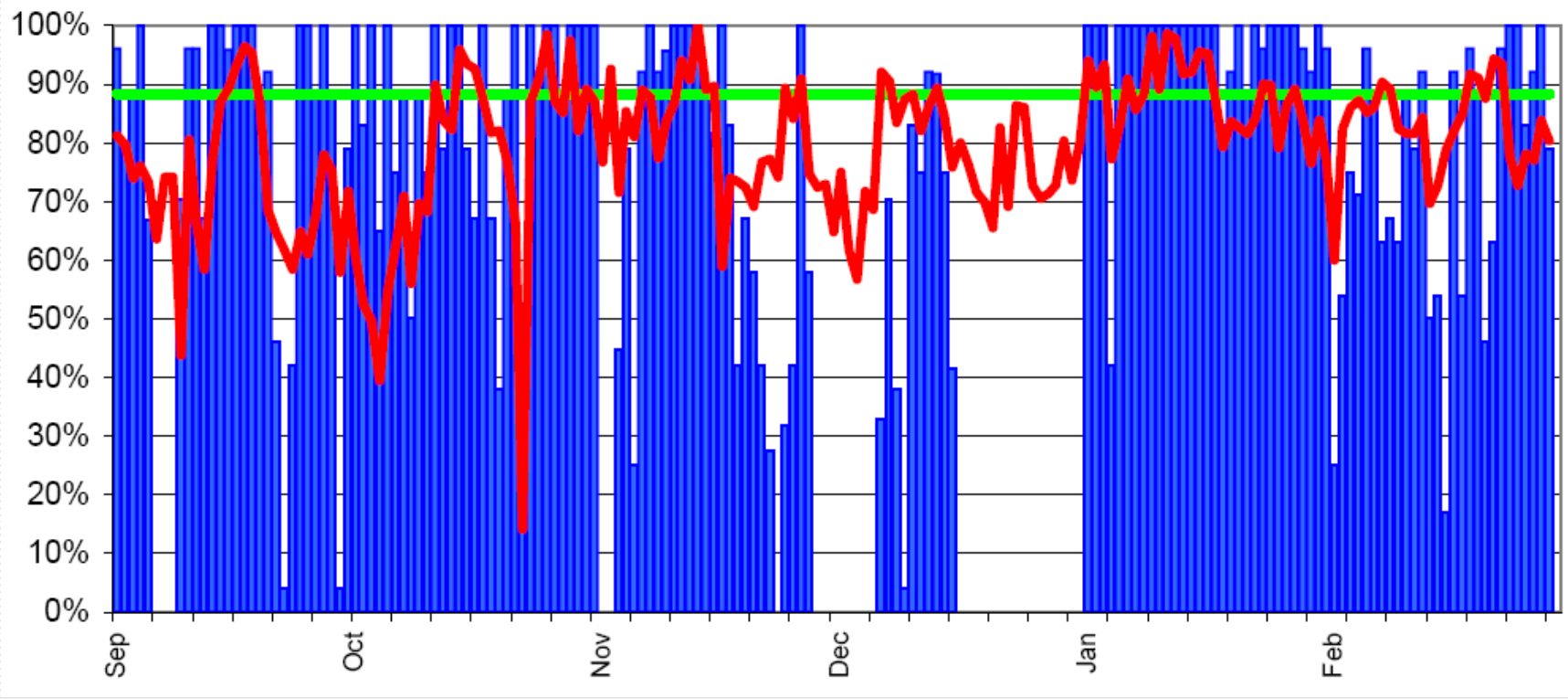
Ressources – stockage et transferts

- Sep-Nov 2006, test FTS, T0 → 5 T1s



Yves Schutz, Comité de Direction LCG-France, 5 février 2007

Site availability



IN2P3-CC av.reliability last 3 mths 65%

target (90% of MoU) 88%

last 3 month averages: all sites 83%

Source: http://lcg.web.cern.ch/LCG/MB/availability/site_reliability.pdf

Computing capacity increase in 2006

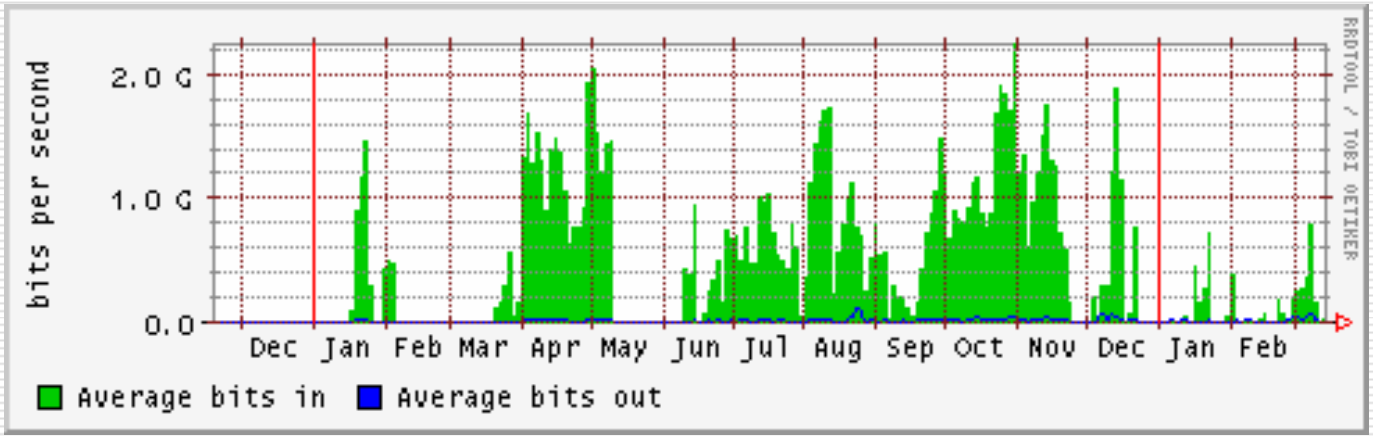
- CPU
  - +265 worker nodes (IBM, dual-processor dual-core AMD Opteron 275, 2.2 GHz, 2 GB/core, 290 GB internal disk)
  - Theoretical power: 1573 SI2000 per core
    - Total: 1,6 M SI2000
    - Observed power with typical applications is ~30% less than theoretical
- Disk storage
  - +400 TB of rack-mounted Sun Fire X4500 (aka Thumper)

Computing capacity increase in 2006 (cont.)

- Tape storage
  - Call for tender for a new cartridge library
  - Selected Sun/StorageTek SL8500
    - 10.000 slots (500 GB cartridges)
    - 30 T10000 drives
    - 10 LTO-3 drives
    - Will progressively replace the current one
  - Installation started: expected to be finished by end of April 2007

Computing capacity increase in 2006 (cont.)

- Databases
  - Reconfiguration of Oracle cluster
    - Extensible hardware architecture
  - +1 TB added to the dedicated SAN (2 TB total)
  - +3 front-end database servers (5 total)
    - 2 of them will share the load of the LHC experiments
- International connectivity
  - Dedicated link CC-IN2P3↔CERN 10 Gbps



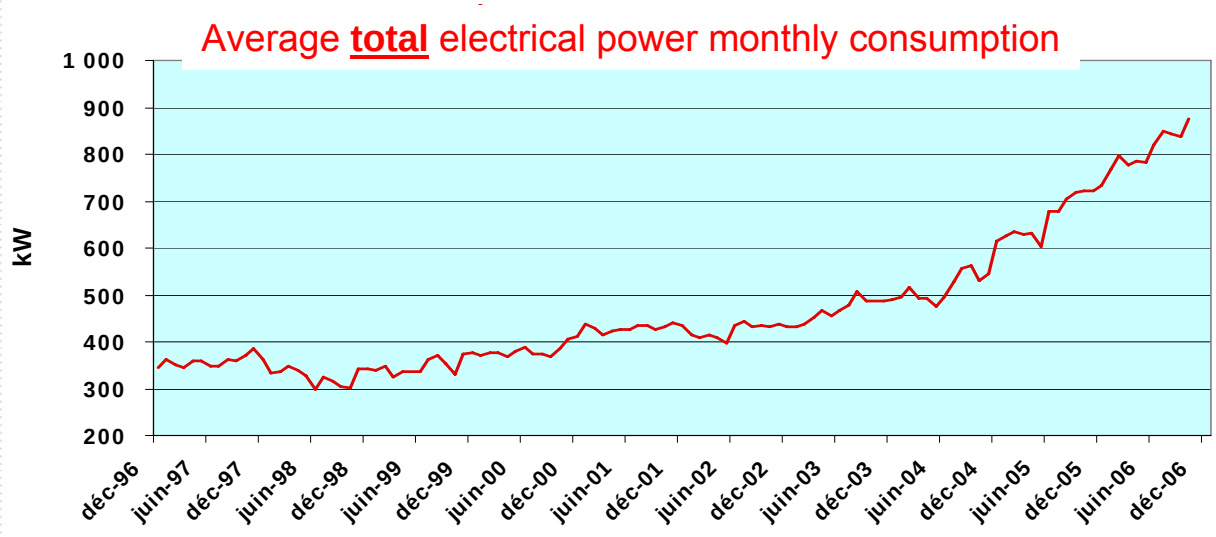
  - 2 x 1 Gbps links CC-IN2P3 ↔ Fermilab

Hardware procurement

- Procurement process (evaluation, publication, selection) is more or less under control
  - Delivery delays are not!
  - In 2006, we suffered delivery delays of several months for some equipment
- Procurement of equipment is an issue
  - Several constraints: space in the machine room, budget constraints, delivery delays, requested availability, …

Facility Upgrade

- Major effort for upgrading the electric and cooling infrastructure of the site
  - Currently reaching the limits of the installation
  - When the current works will be finished (April 2007)
    - from 500 kW to 1000 kW usable for computing equipment

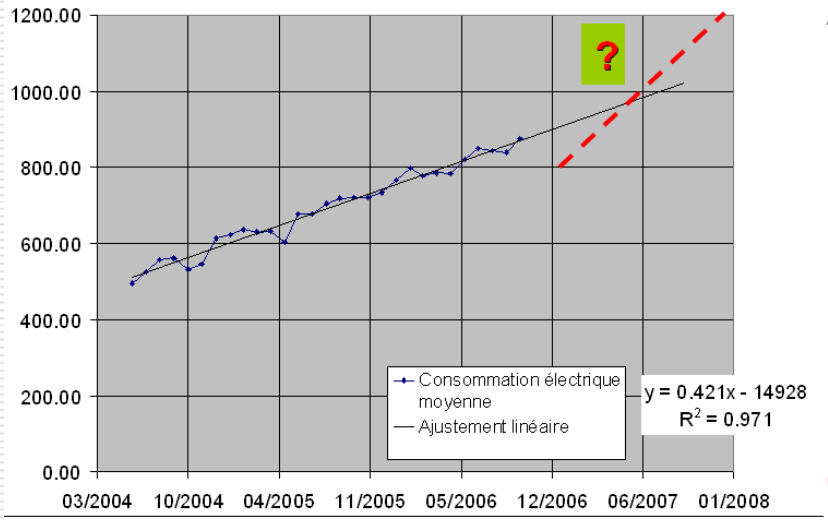


Courtesy of Dominique Boutigny

Facility Upgrade (cont.)

Infrastructure (2)

CC-IN2P3 average electrical power in kW



An important work is going on in order to upgrade the computer room

- Electrical distribution
- Cooling
- Uninterruptible Power Supply

➔ Up to ~1.6 MW of computing equipment + cooling (1 MW for computing equipment)

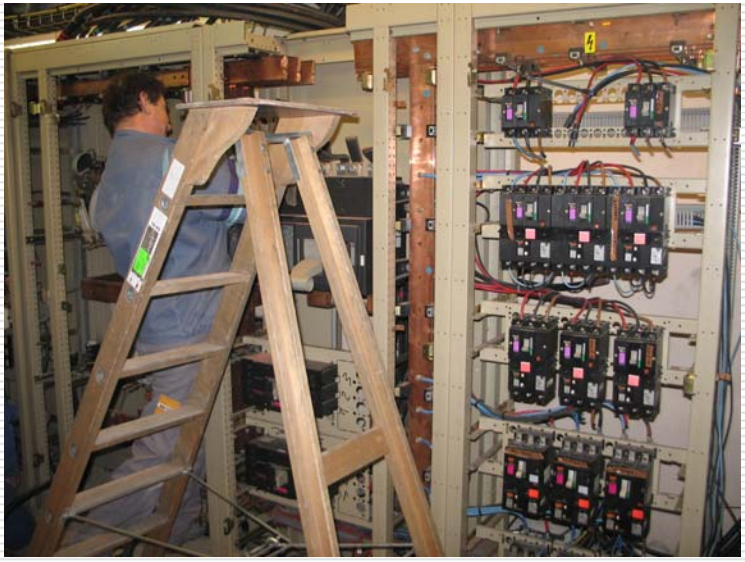
- The exponential increase of the computing resources has a significant impact on the computing centre infrastructure

Courtesy of Dominique Boutigny

Facility Upgrade (cont.)

- Scheduled 4 days-long complete shutdown of the site in December 2006 for replacing central electric equipment
  - Vital services (network equipment, mail servers, web servers, Oracle, FTS, LFCs, VOMS,…) were kept alive by ad hoc means)
    - Extensive use of virtual machines
  - Others services have been switched to partner sites
    - CIC Portal was hosted by CNAF during the shutdown and switched back to CC-IN2P3 afterwards
    - Failover procedure tested in real conditions

Facility U)



Courtesy of Jean-Louis Perrot

Site Operation

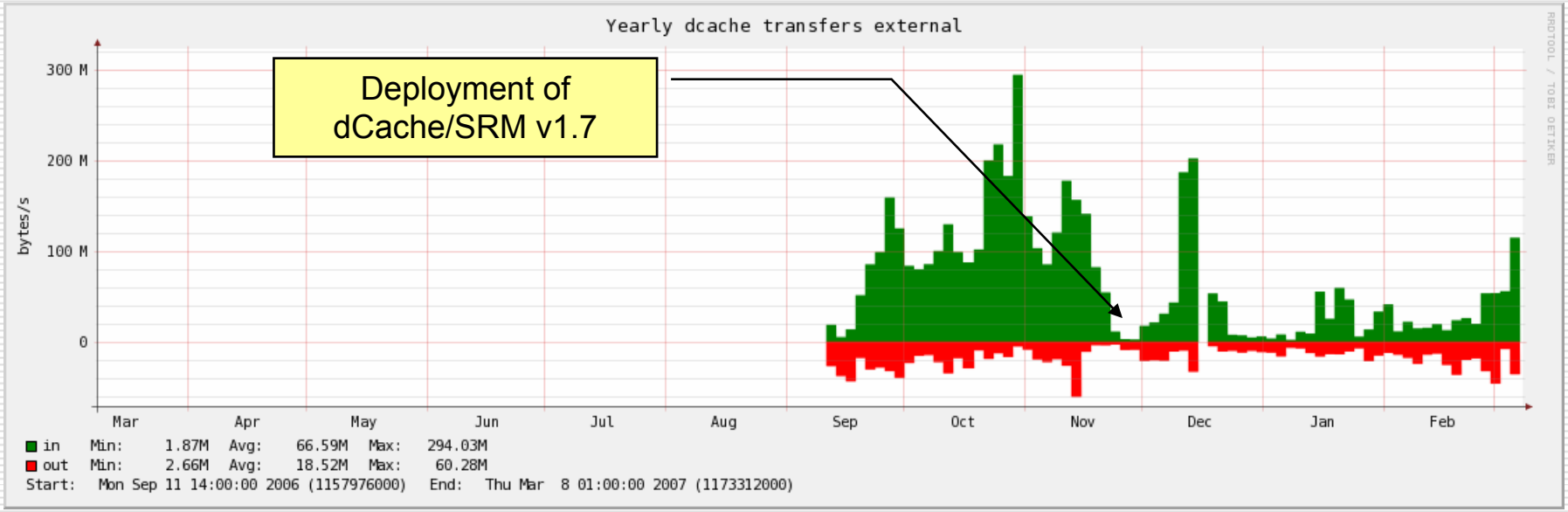
- Batch operations
  - Passing the LCG job requirements to the local batch scheduler is still necessary
    - Turnaround implemented to modify individual job requirements (memory and CPU) while it is in the BQS queue
      - *Set to less than 2 GB for LHCb and more than 2 GB for CMS (in some cases)*
  - Redefinition of maximum CPU time for some BQS queues to better fit the demand
  - Modification of the built-in BQS job monitoring mechanism to detect (and stop the execution of) pathological jobs
    - So not to block selected users while they do some testing (with pilot jobs, for instance)
  - Temporary solution for implementing priorities within the same VO based on the VOMS role
    - Tested with Atlas jobs. An equivalent solution will be put in place for CMS
  - Increase the usage of the BQS taging of jobs capability
    - For instance, for tagging the jobs requesting dCache so that when dCache (or HPSS) is not available, those jobs are not put in execution
    - Feature also used to regulate the execution of jobs with the same tag

Site Operation (cont.)

- Batch operations (cont.)
  - Improvements to BQS planned for 2007
    - Priority handling between jobs within the same VO and between grid and non-grid jobs
    - Associate the whole user's proxy to job information (in addition to just the proxy's subject) and other grid-related attributes of the job (i.e. grid name, grid job id, …)
    - Use the user's proxy as a criterion for scheduling
      - *For instance to prevent execution of a particular user's jobs*
  - Currently developing the BQS interface for gLite CREAM computing element
    - Expected to test it by the end of 2007Q1
    - Thanks to Massimo Sgaravatto for his support
    - Many difficulties encountered with the gLite CE interface (reported to the TCG on 01/11/2006)

Site Operation (cont.)

- Grid services operations
  - Storage Element
    - Stabilizing the SRM-based SE service since the deployment of dCache/SRM v1.7 has been extremely difficult
      - *Current service is not yet as stable as with previous release*



Traffic into and out of dCache since september 2006

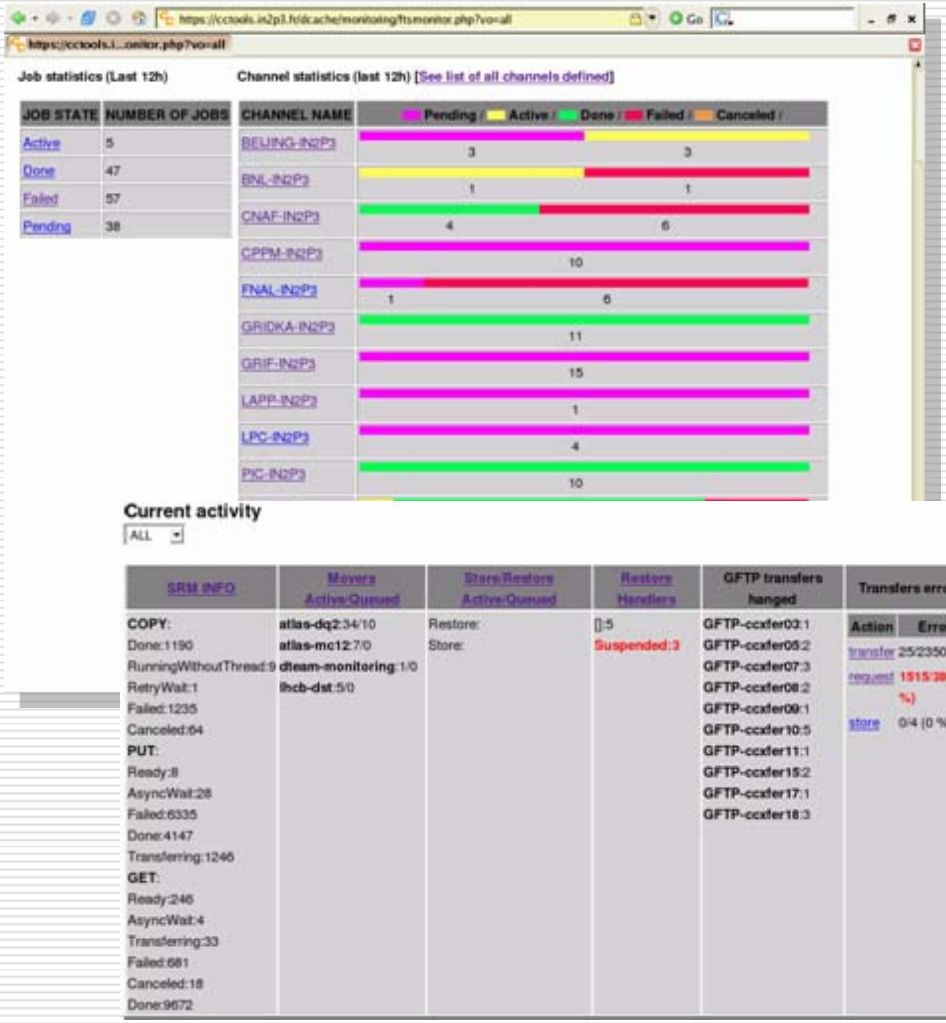
Site Operation (cont.)

- Grid services operations (cont.)
  - Storage Element (cont.)
    - Service unstability and unavailability severely impacted experiments during late november and december 2006
      - *In spite of the efforts deployed by the dCache/SRM developers for finding the roots of the problem*
    - Detailed report done by Lionel Schwarz during the dCache workshop in January 2007
    - IMHO, the real issue is how to test, in near real load conditions, a key component such as dCache/SRM before putting a new release in production?

Site Operation (cont.)

- Continuous effort to develop/adapt/deploy tools for easing the operations of the various grid services
  - Monitoring of FTS activity per channel, dCache activity and dCache errors
  - the ultimate goal is that the operations of the grid services be handled as the operations of the « traditional » services

Site Operation (cont.)



Job statistics (Last 12h)

JOB STATE	NUMBER OF JOBS
Active	5
Done	47
Failed	57
Pending	38

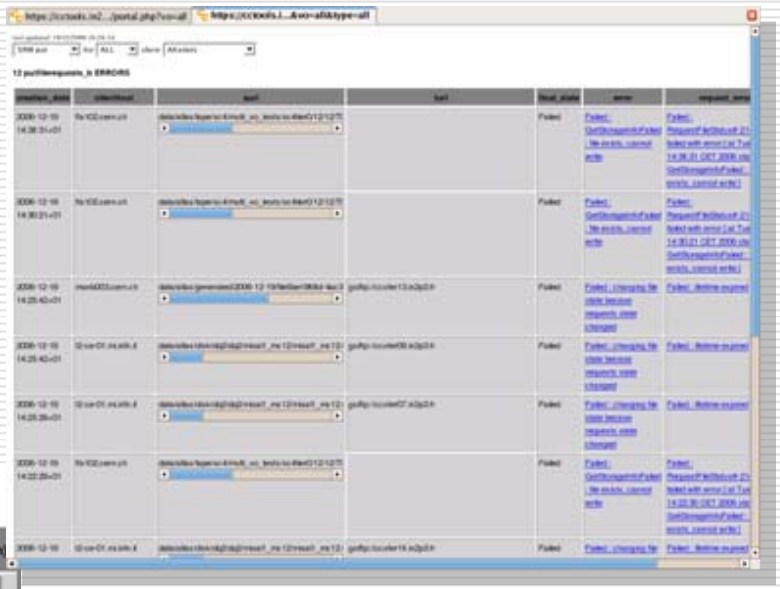
Channel statistics (last 12h) [See list of all channels defined]

CHANNEL NAME	Pending / Active / Done / Failed / Canceled /
BEIJING-IN2P3	3 3
BNL-IN2P3	1 1
CNAF-IN2P3	4 6
CPPM-IN2P3	10
FNAL-IN2P3	1 6
GRIDKA-IN2P3	11
GRIF-IN2P3	15
LAPP-IN2P3	1
LPC-IN2P3	4
PIC-IN2P3	10

Current activity

ALL

SRM INFO	Movers Active/Queued	Store/Restore Active/Queued	Restore Handlers	GFTP transfers hanged	Transfers errors (last 2h)
COPY: Done:1190 RunningWithoutThread:9 RetryWait:1 Failed:1235 Canceled:64 PUT: Ready:8 AsyncWait:28 Failed:6335 Done:4147 Transferring:1246 GET: Ready:246 AsyncWait:4 Transferring:33 Failed:681 Canceled:18 Done:9672	atlas-dq2:34/10 atlas-mc12:7/0 dteam-monitoring:1/0 lhcb-dst:5/0	Restore: Store:	[]:5 Suspended:3	GFTP-ccxfer03:1 GFTP-ccxfer05:2 GFTP-ccxfer07:3 GFTP-ccxfer08:2 GFTP-ccxfer09:1 GFTP-ccxfer10:5 GFTP-ccxfer11:1 GFTP-ccxfer15:2 GFTP-ccxfer17:1 GFTP-ccxfer18:3	Action / Errors/total: transfer 25/2350 (1 %); request 1515/3808 (39 %); store 0/4 (0 %)



Courtesy of Lionel Schwarz

Site Operation (cont.)

- Grid services
  - Target availability of the tier-1 sites require that the grid services be designed and implemented with this goal in mind
    - Redundancy in the services must be possible without the need of current « gymnastics »
  - We need to improve the manageability of the grid services
    - Standard interfaces for administering, (remotely) controling, monitoring their activity and standard locations for logs and traces would help a lot in this direction

Alice

PDC06: conclusions

- En 2006 les ressources CPU fournies par LCG-France (T1 & T2s) sont à peu-près celles déclarées dans le MoU LCG
- Ces ressources sont insuffisantes
- Les ressources pour le stockage de données n'ont pas été utilisées du fait de l'absence de SE
- Les tests de transfert T0 → CC ont atteint les taux requis, mais la stabilité du service reste insatisfaisante
- Pas de tests de transfert CC ↔ T2s
- Depuis le début de l'année, le suivi des opérations au CC est problématique, en l'absence d'un contact sur place.

5 fevrier 2007 ALICE@LCG-France 10

Yves Schutz, Comité de Direction LCG-France, 5 février 2007

Conclusion provisoire

- Le Tier-1 influence l'efficacité des Tier-2 mais pas toujours
  - Problèmes récurrents de srm au CC
- Chaque Tier-2 a des problèmes spécifiques
- Il faut améliorer :
  - Le monitoring,
  - Plus de checks systématiques,
  - L'implication des sites,
  - Les relations avec les sites
- Cependant…
  - L'efficacité du nuage français est reconnu!

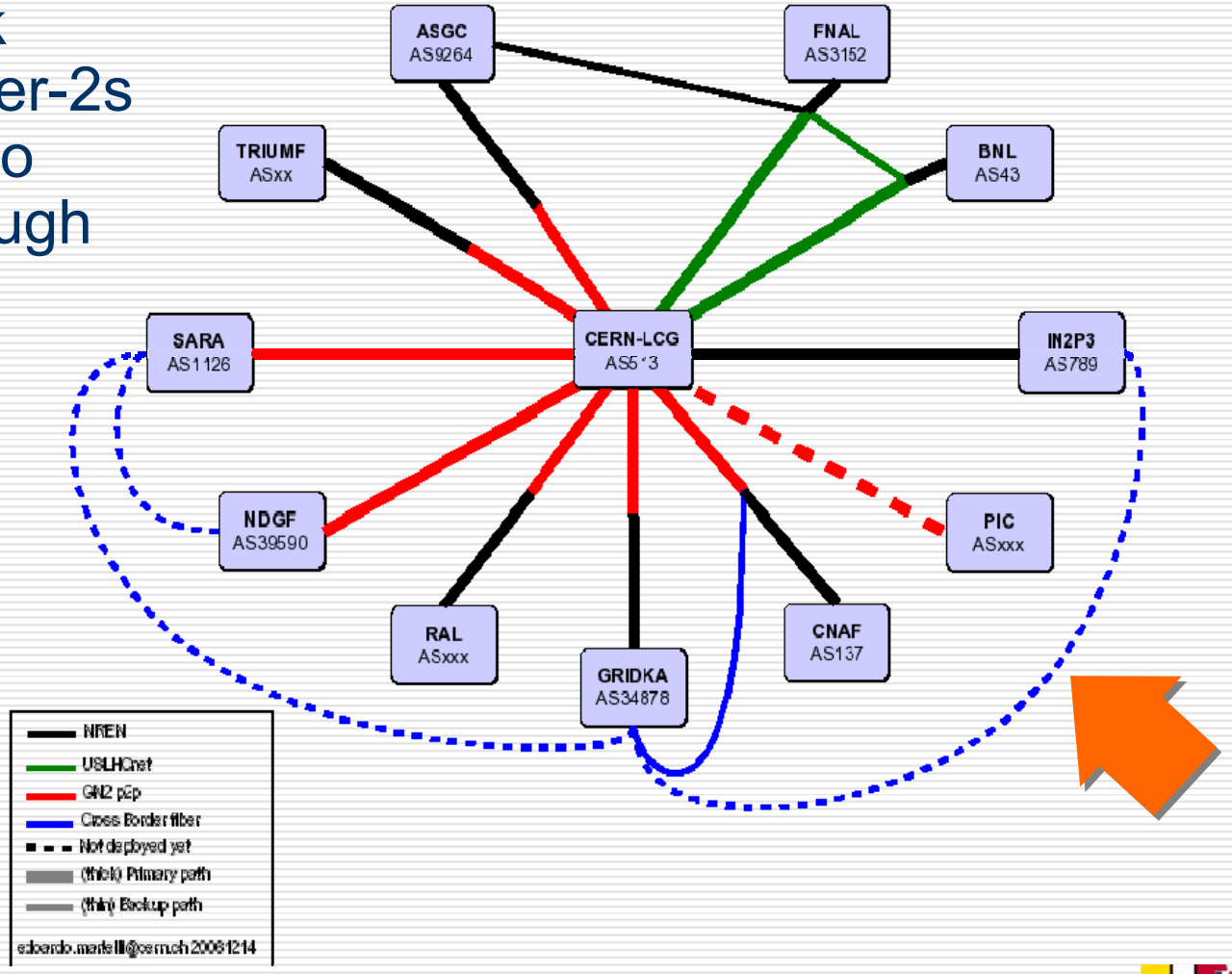
Plans for 2007

Facility Upgrade

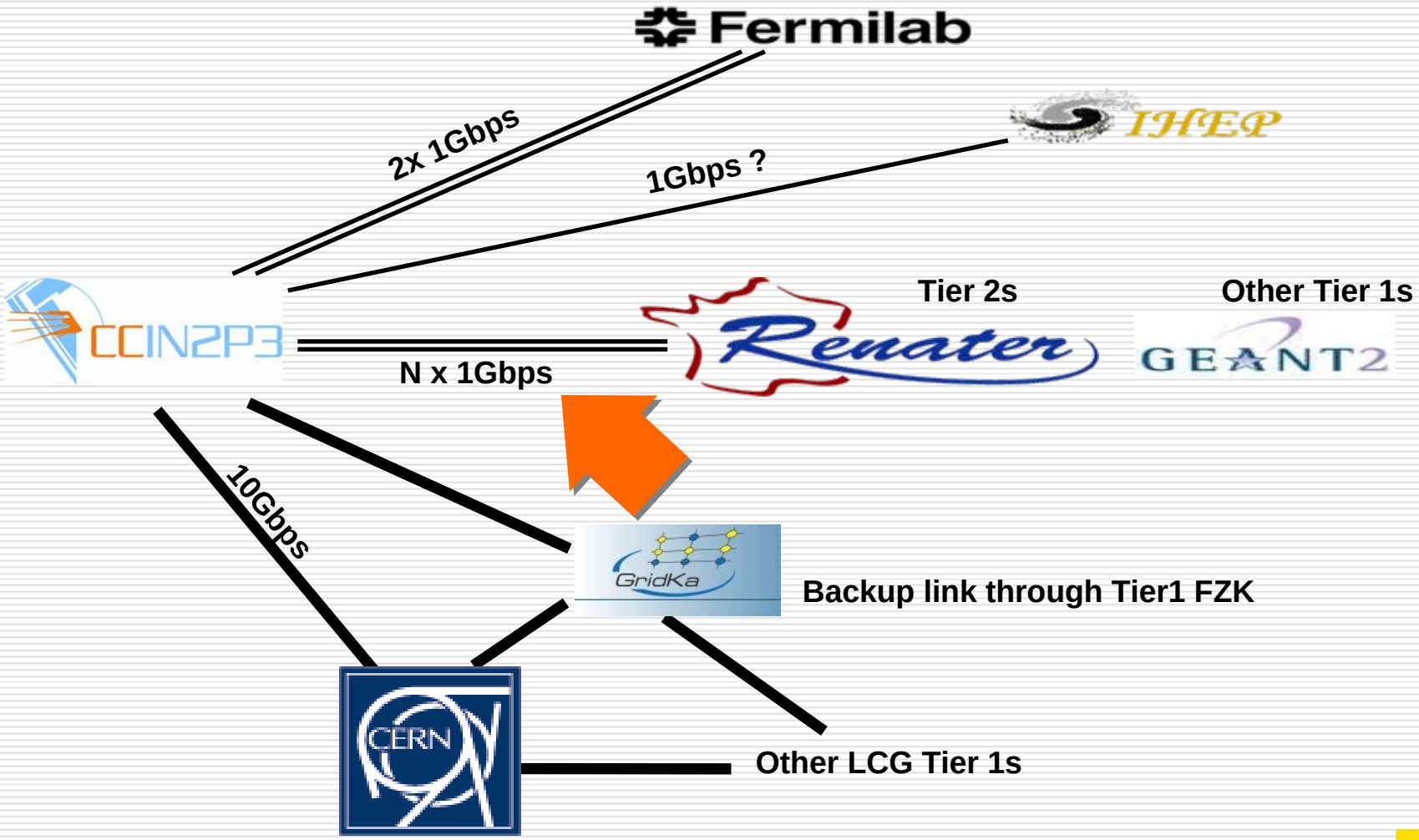
- Project scheduled to be finished by June 2007
  - 3 additional UPS
  - New diesel power generator
  - Additional power distribution equipment in the machine room
  - Additional cooling equipment
- Let's cross our fingers!

Connectivity

- Increase network bandwidth with tier-2s and backup link to other tier-1s through FZK



Connectivity (cont.)

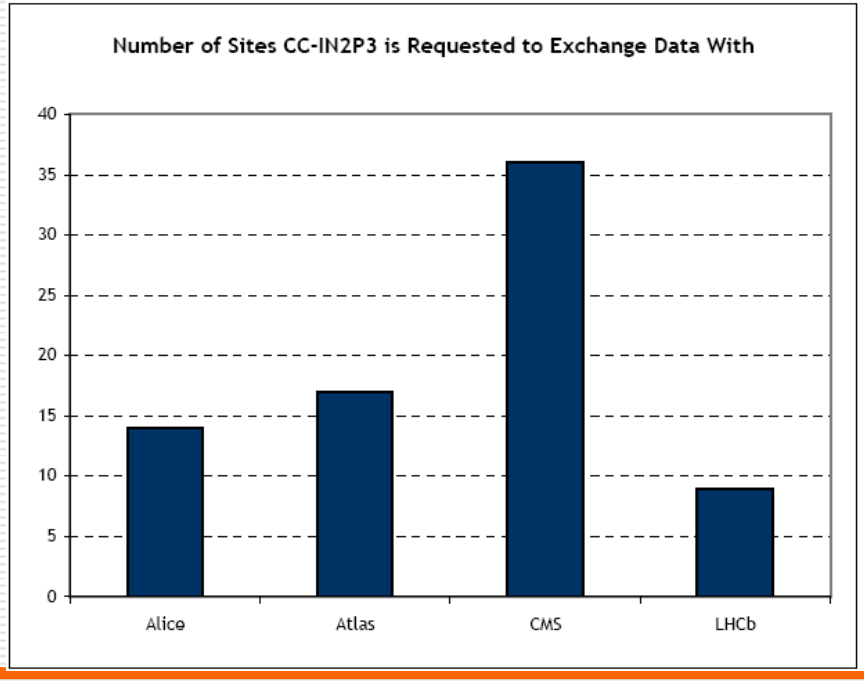


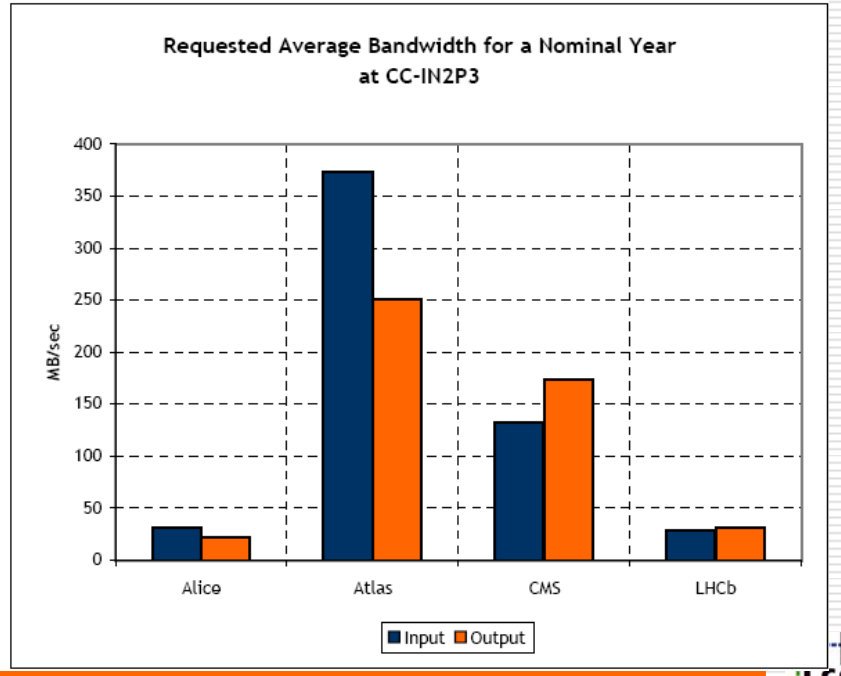
Courtesy of Jérôme Bernier

Network bandwidth requirements

- Summary by experiment

Experiment	Number of Sites	Input		Output	
		Average Bandwidth [MB/sec]	Peak Bandwidth [MB/sec]	Average Bandwidth [MB/sec]	Peak Bandwidth [MB/sec]
Alice	14	30,7	40,7	22,3	29,5
Atlas	17	373,8	522,4	251,6	359,8
CMS	36	132,7	132,7	174,2	404,2
LHCb	9	28,4	28,4	31,8	31,8
	Total	565,6	724,2	479,9	825,3



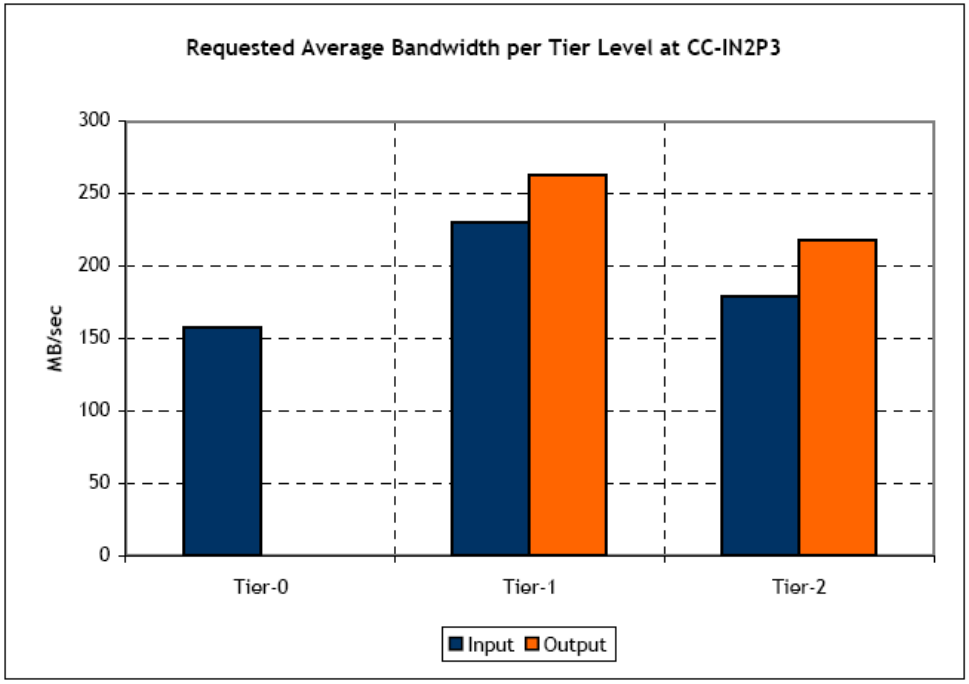


Source: https://edms.in2p3.fr/document/I-010099

Network bandwidth requirements (cont.)

- Summary by tier level

Tier-1	Number of Sites	Input		Output	
		Average Bandwidth [MB/sec]	Peak Bandwidth [MB/sec]	Average Bandwidth [MB/sec]	Peak Bandwidth [MB/sec]
Tier-0	1	157,0	157,0		
Tier-1	11	229,6	229,6	262,5	262,5
Tier-2	42	179,0	337,6	217,4	562,8
	Total	565,6	724,2	479,9	825,3



Source: https://edms.in2p3.fr/document/I-010099

Local Network Requirements

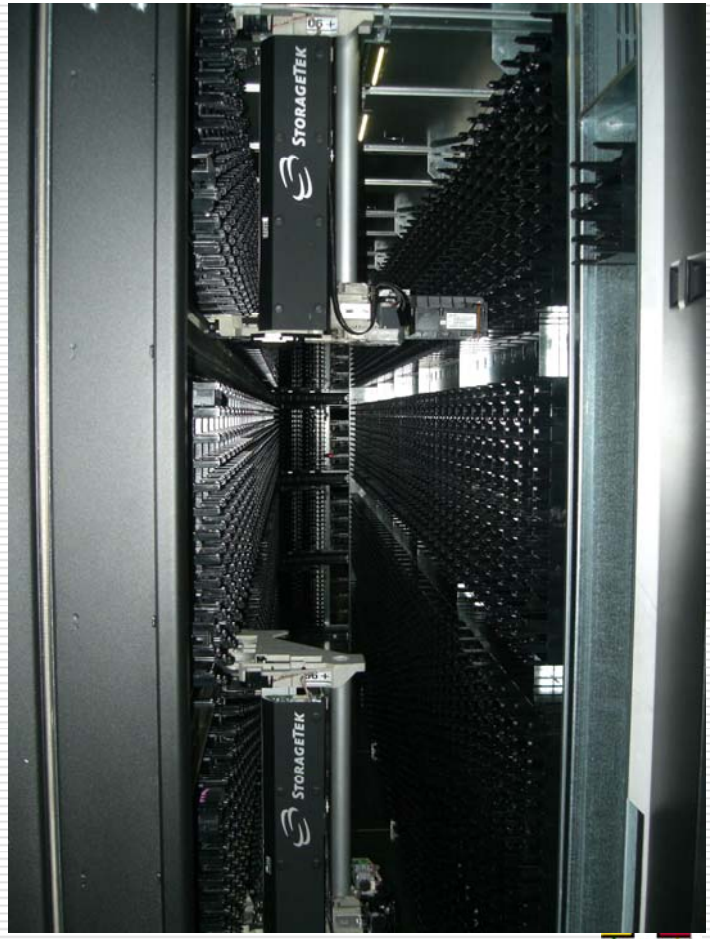
- We need to better understand how the data will be accessed by the jobs running in the site
  - Direct impact on the needs of the local network

Compute Capacity Increase

- On-going call for tenders for compute nodes and disk servers
  - +4,5 M SI2000
    - Non-LHC: 1 M SI2000
    - LHC
      - *Needs for 2007: 1,3 M SI2000*
      - *Provision for 2008: 2,2 M SI2000 (~40% of capacity required in 2008)*
  - +1200 TB (DAS)
    - LHC needs for 2007: 400 TB
    - LHC provision for 2008: 800 TB
  - +160 TB (SAN)

Compute Capacity Increase (cont.)

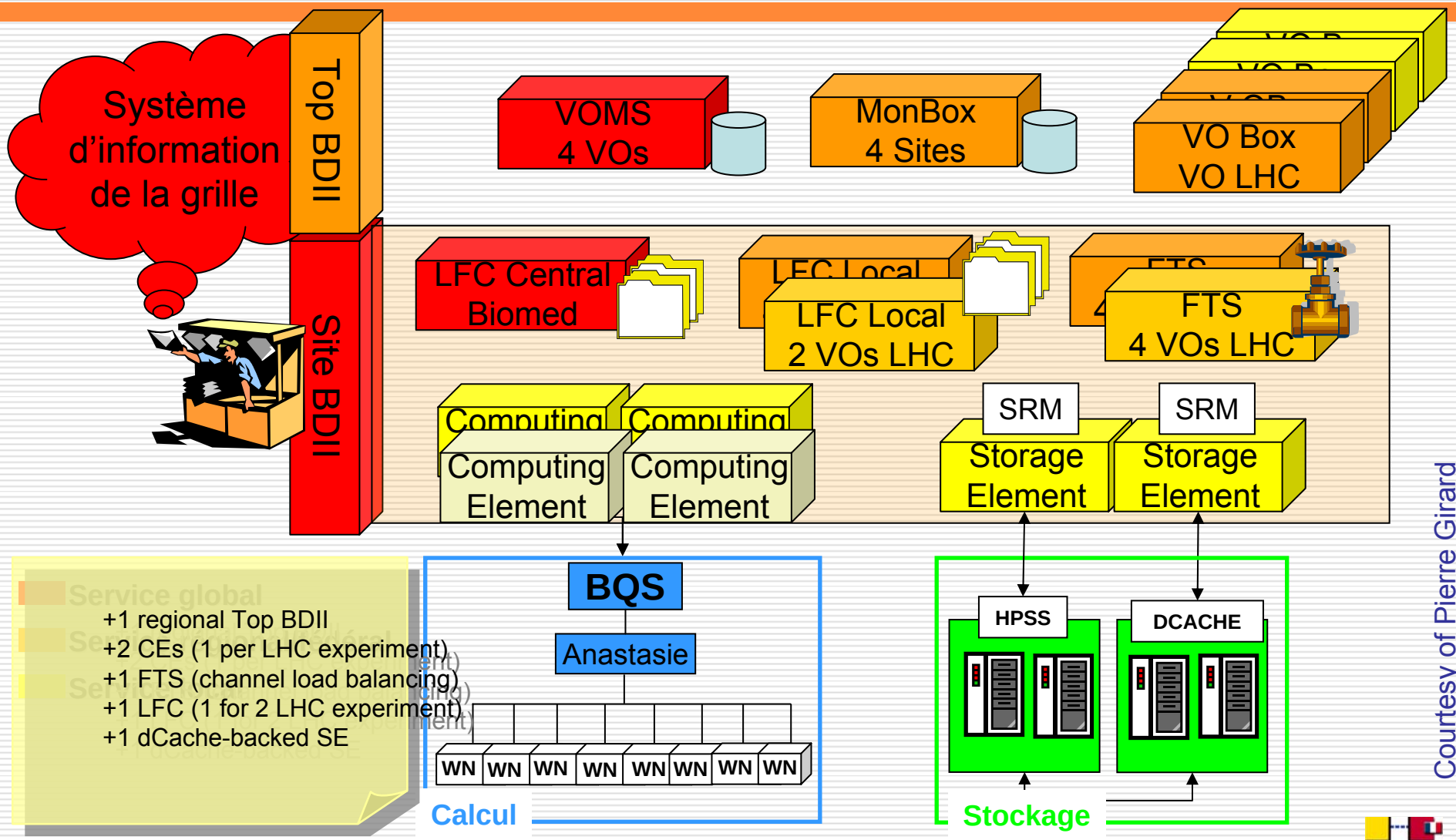
- Cartridge library
  - 10.000 slots, 30 drives, up to 5 PB



Grid Services

- Consolidate current grid services and integrate them into « normal » operations
  - Works towards the stability desired not only by the experiments but by the people operating the services at the site

Consolidation of grid services
(by the end of June 2007)



Courtesy of Pierre Girard

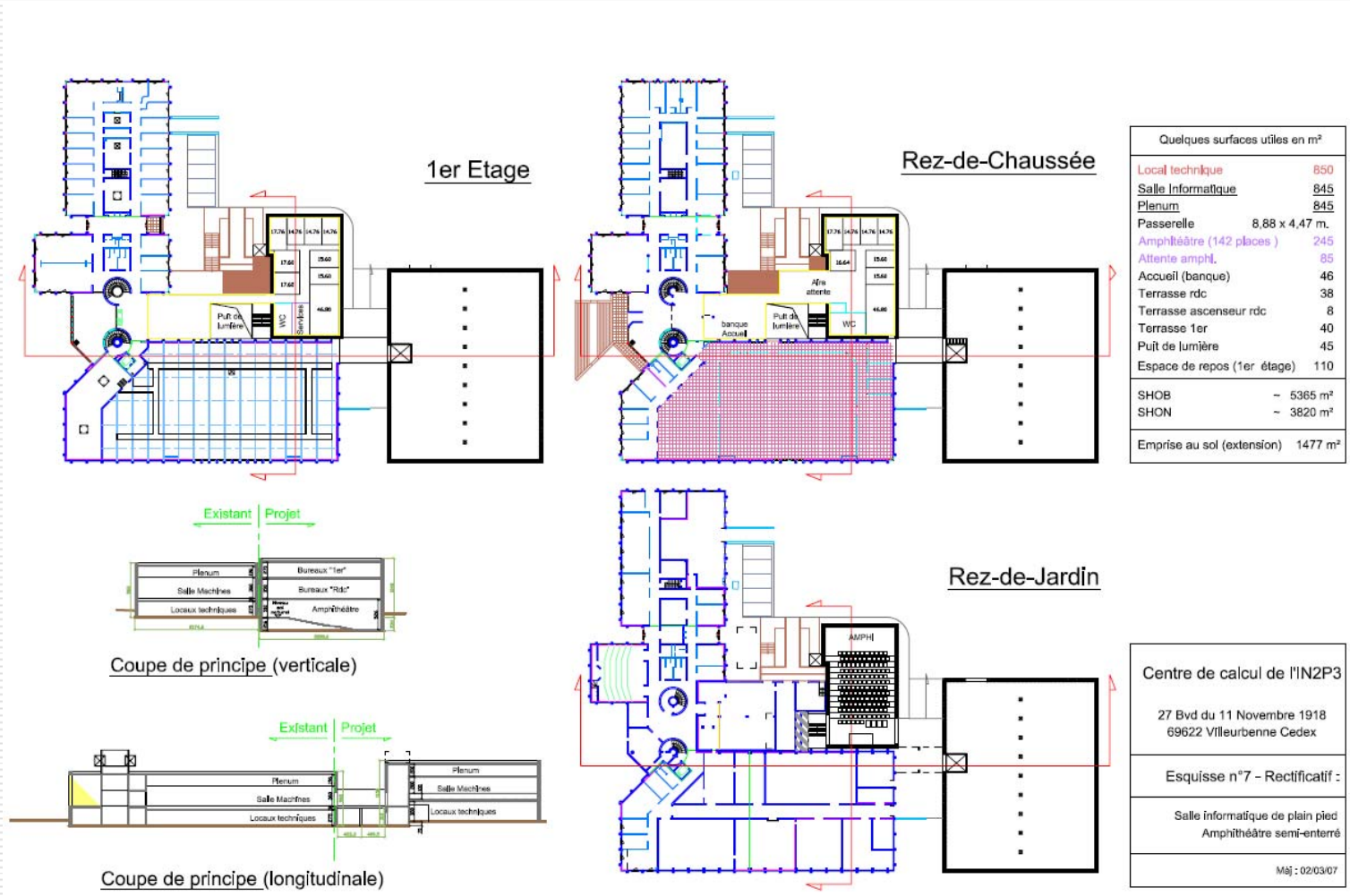
Analysis Facility

- We need to understand what it really means to design and operate an analysis facility
  - A big help from the experiments is required (also) in this area

New building

- On-going project for building an additional machine room
  - 800 m² floor space
  - Electric power for computing equipment: 1 MW at the beginning, with capacity for increasing up to 2,5 MW
- Offices: for around 30 additional people
- Meeting rooms, 140+ seats amphitheatre
- Target availability: mid 2009

New building (cont.)



Conclusions

- Ramp up plans of the site is rather aggressive
  - Several constraints don't really make our life easier
- Operating the grid services in their current status is complex and requires (highly competent and motivated) people
- On-site people dedicated for supporting the experiments are instrumental in optimising the utilisation of the site resources
- Don't understimate your infrastructure needs

Acknowledgments

- Thanks to the people that contributed material to this talk
  - This presentation would be even longer if I listed them all

More Information

- LCG-France website http://lcg.in2p3.fr
- LCG-France T2-T3 Technical coordination wiki page:http://lcg.in2p3.fr/wiki/index.php/T2-T3
- CC-IN2P3: http://cc.in2p3.fr

Questions

Fabien Wernli, 2006