CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

DSM

IN2P3
Institut National de Physique Nucléaire
et de Physique des Particules

LCG-France Tier-1 and Analysis Facility

Overview

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

CMS Tier-1 tour
Lyon, November 30th 2007

Contents

- Introduction to LCG-France
- Introduction to CC-IN2P3
- Resources for LHC experiments with focus on CMS
  - budget, plan and pledges
  - contribution
- Current and future work
- Conclusions
- Questions

LCG-France project

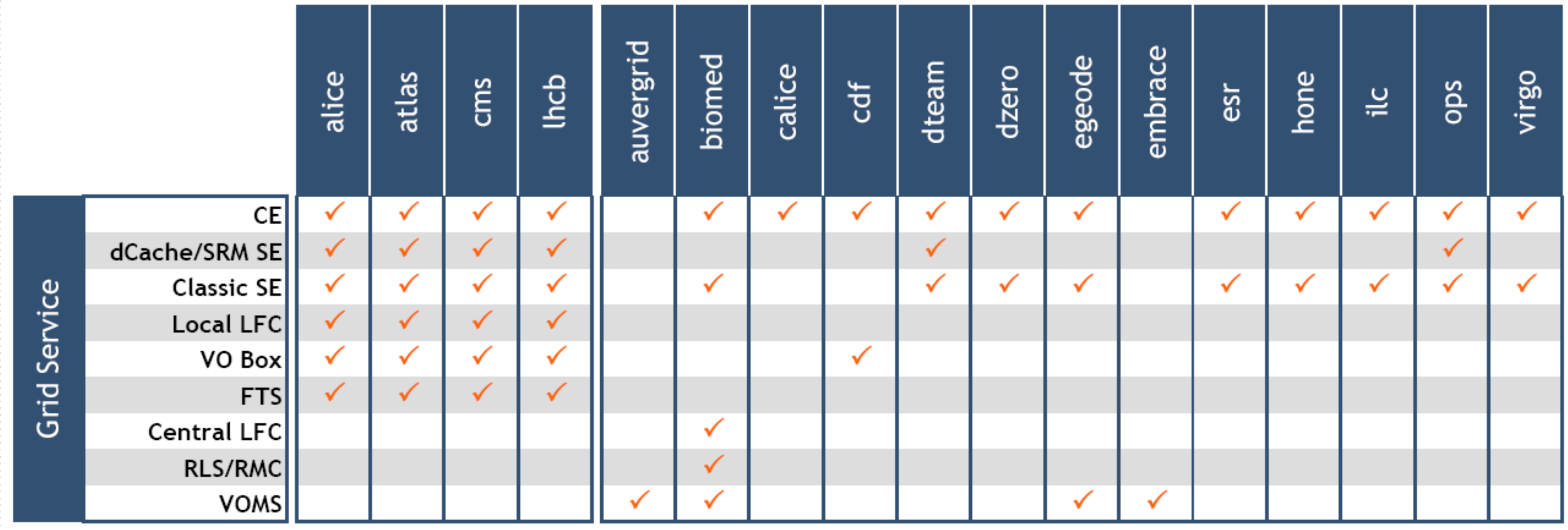
- Goals
  - Setup, develop and maintain a LCG Tier-1 and an Analysis Facility at CC-IN2P3
  - Promote the creation of Tier-2/Tier-3 french sites and coordinate their integration into the WLCG collaboration
- Funding
  - national funding for Tier-1 and AF
  - Tier-2s and tier-3s are funded by universities, local/regional governments, hosting laboratories, …
- Organization
  - Started in June 2004
  - Scientific and technical leaders appointed for a 4-year term, management board (executive) and overview boards in place since then
- Leverage on EGEE operations organization
  - Co-location of the tier-1 and the EGEE regional operations centre
- More information
  - Project web site: http://lcg.in2p3.fr
  - Project charter: https://edms.in2p3.fr/document/I-003682

Introduction to CC-IN2P3

- Data repository and processing facility **shared** by several experiments
    - Operates a WLCG tier-1, a tier-2 and an Analysis Facility for the 4 LHC experiments
- The main compute farm used by both grid and local users
    - Grid middleware is "just another" interface for using our services
- Data storage infrastructure (disk and mass storage) accessible to all jobs running in the site
    - Although not all storage spaces have a gridified interface

Introduction to CC-IN2P3 (cont.)

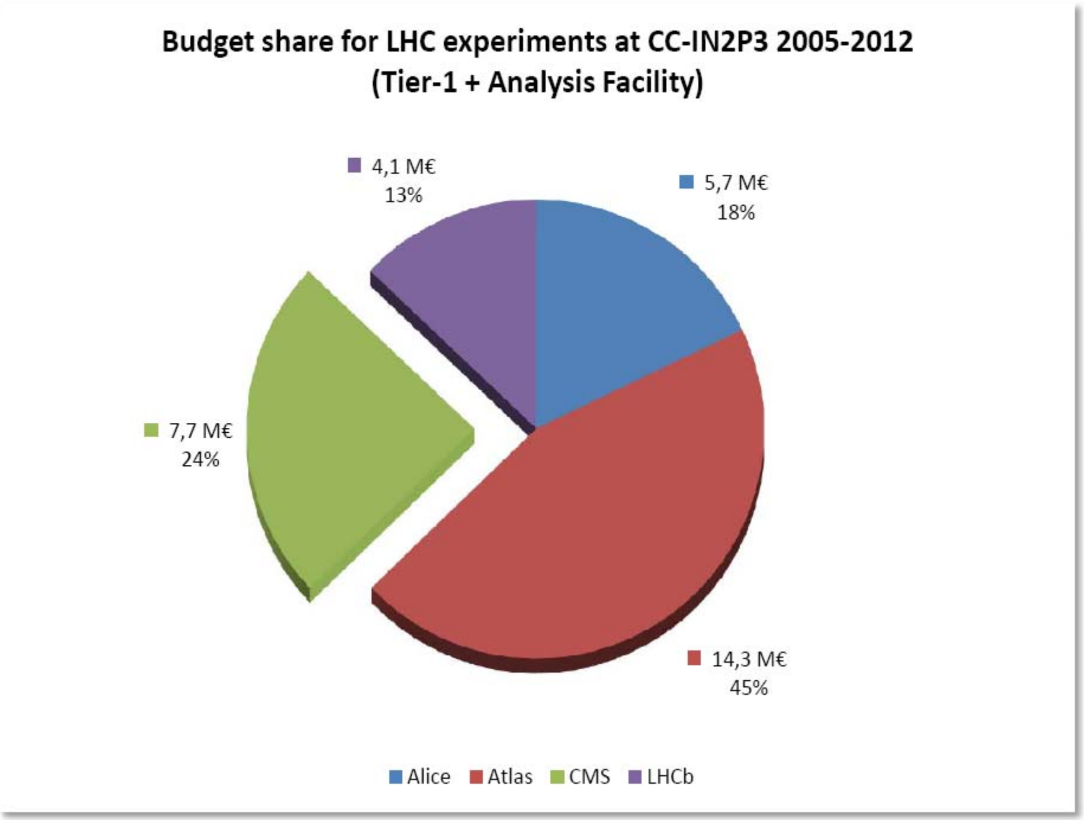
- Also operating grid services for non-LHC VOs



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

Budget: all LHC Experiments

- Equipment and running costs for all LHC experiments at CC-IN2P3 (2005-2012)
  - Total required: 31,9 M€
    - the refurbishment of current machine room and the construction of a second one are NOT included
- Budget requested on a pluriannual basis
  - Approval on a yearly basis
  - Impact on hardware procurement process

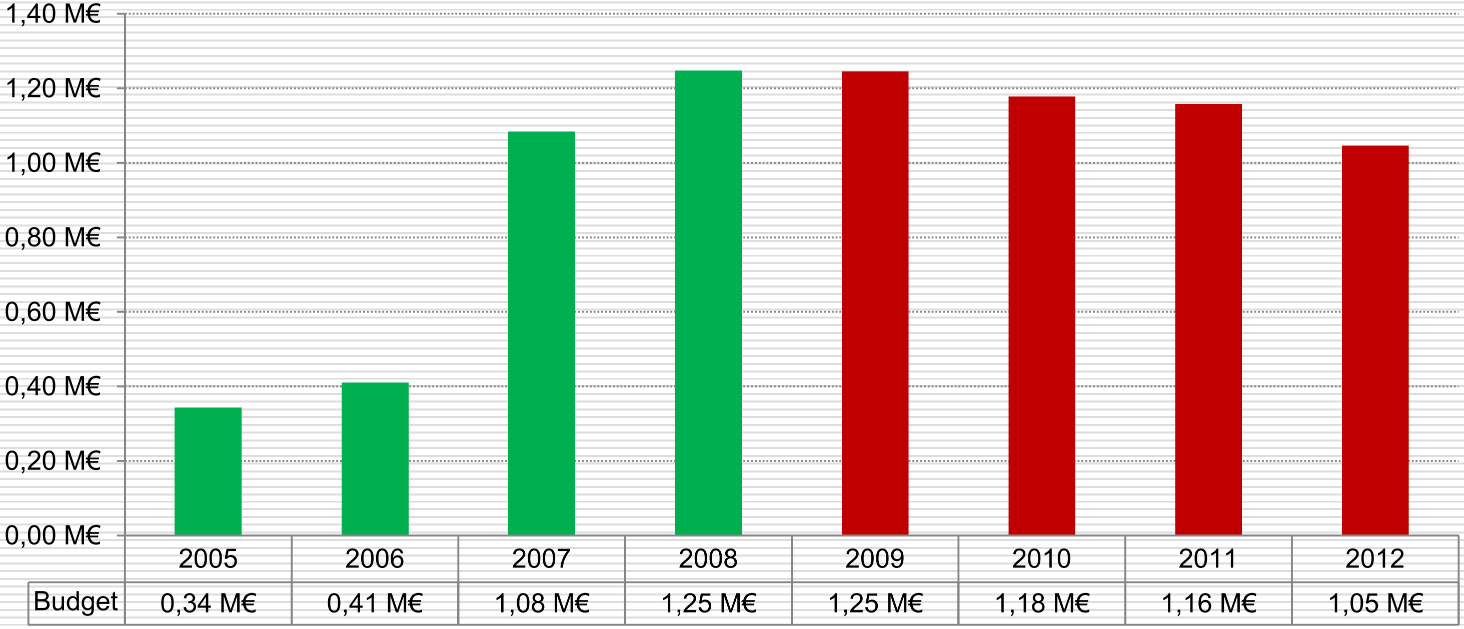


Budget: CMS

- Equipment and running costs for CMS needs (2005-2012)
  - 7,7 M€

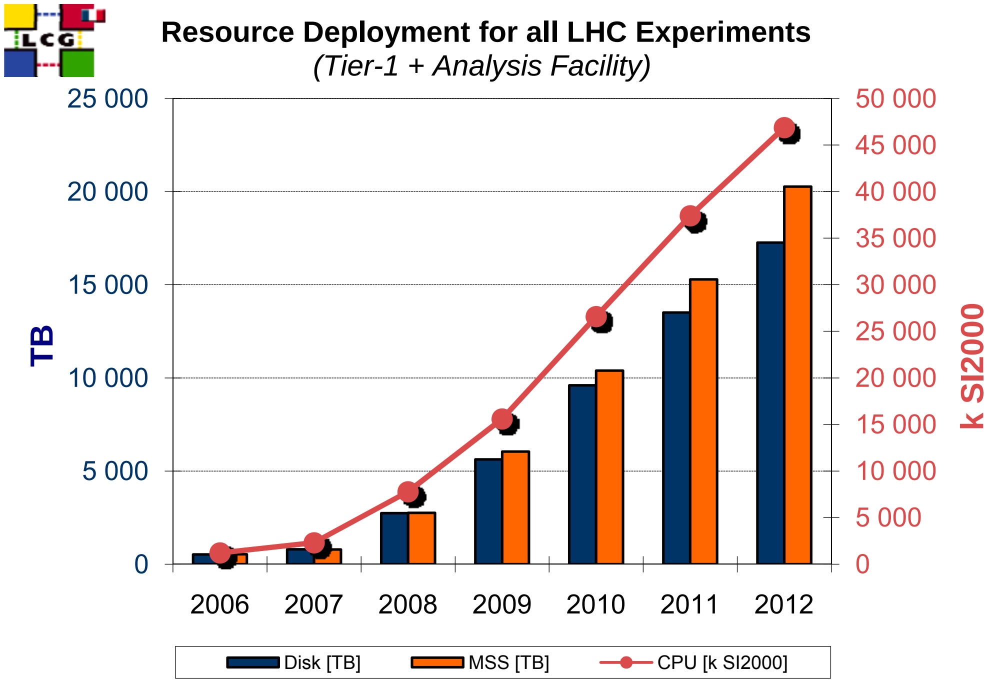
Approved and Requested budget for CMS

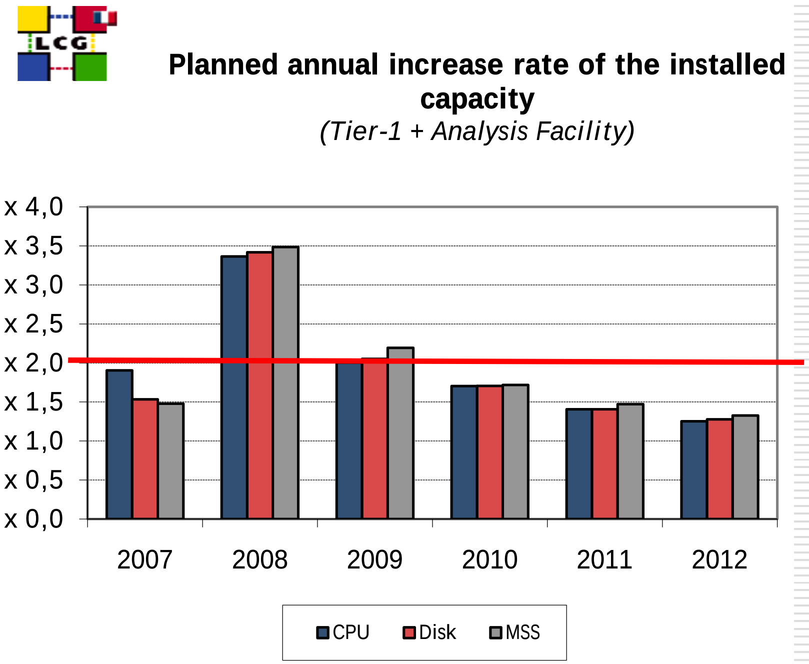
(tier-1 + analysis facility)



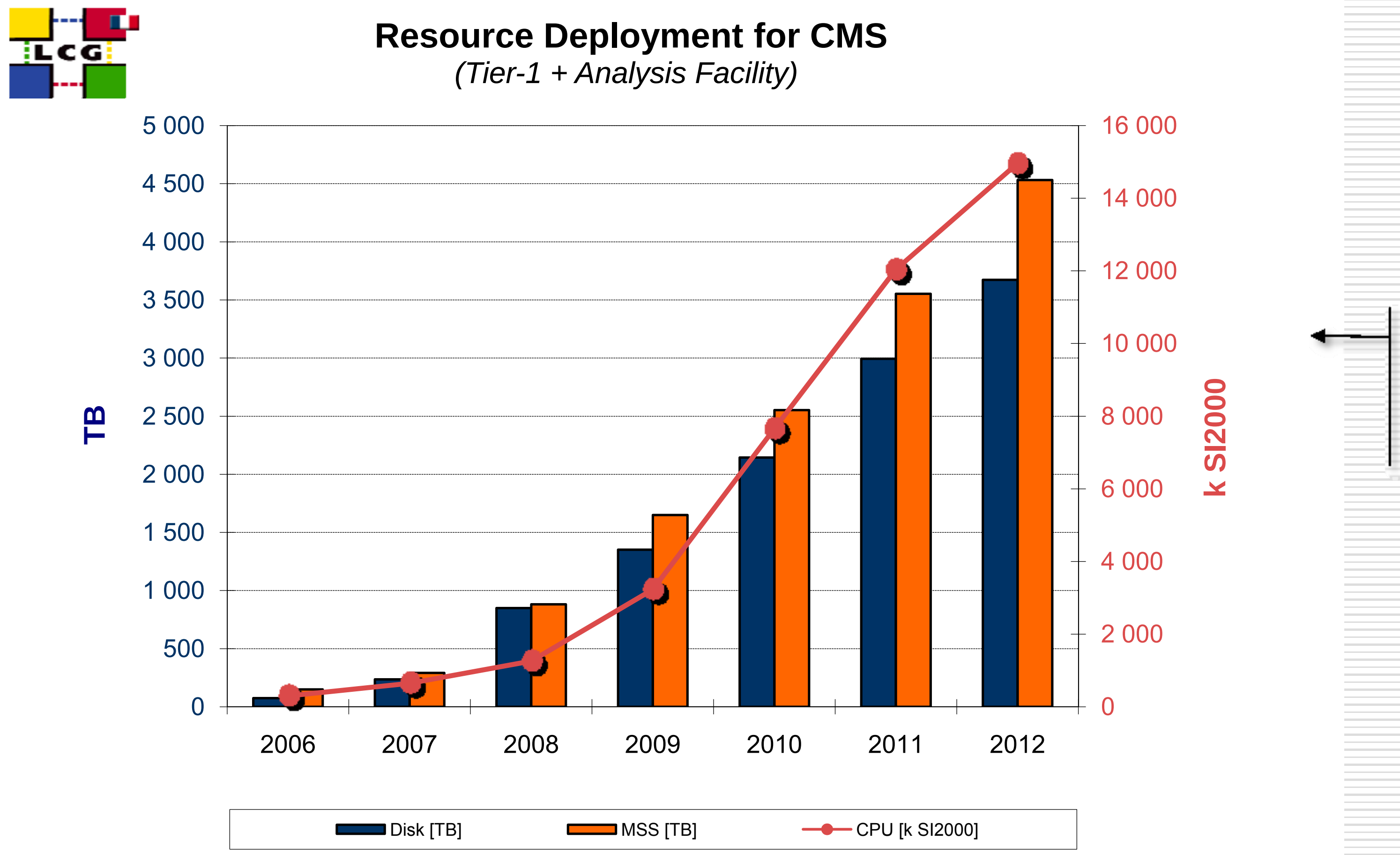
	2005	2006	2007	2008	2009	2010	2011	2012
Budget	0,34 M€	0,41 M€	1,08 M€	1,25 M€	1,25 M€	1,18 M€	1,16 M€	1,05 M€

Planned resource deployment





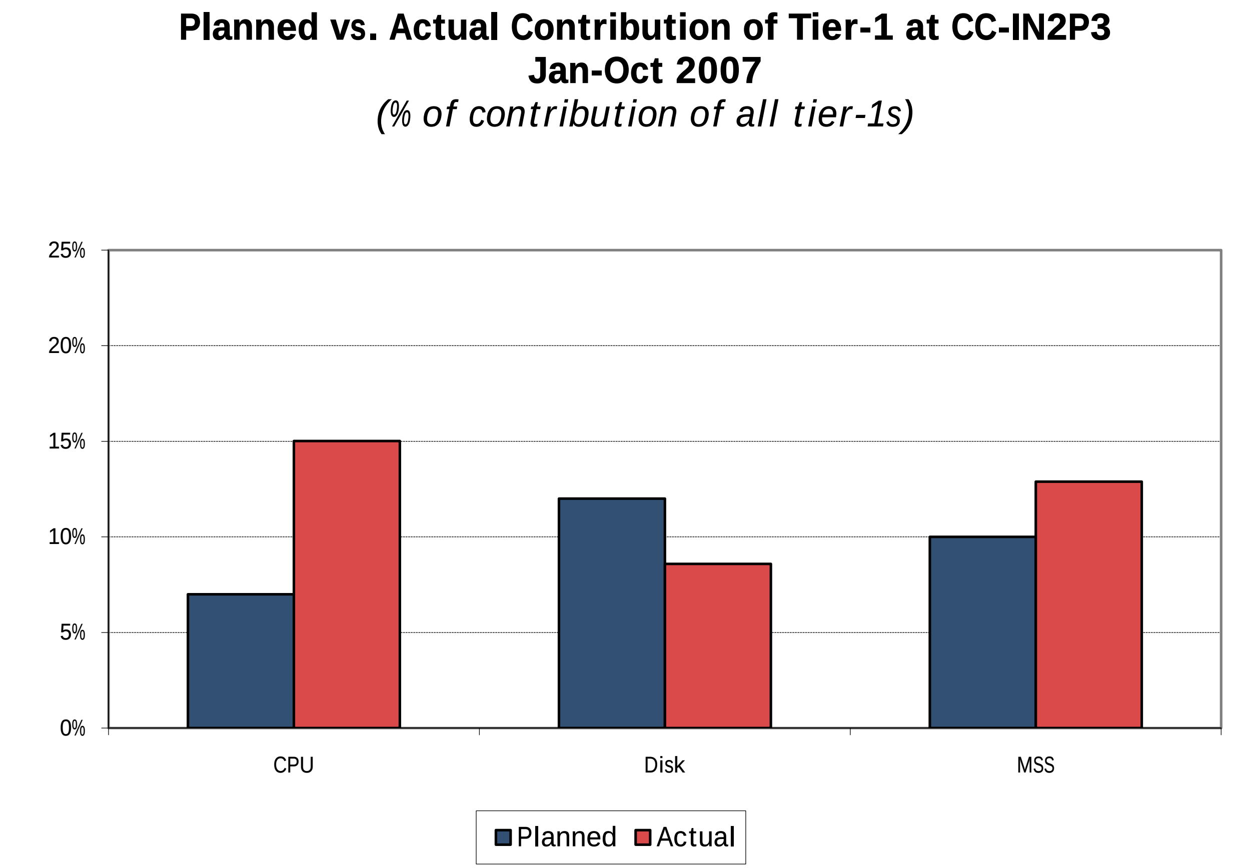
Planned resource deployment: CMS



A fraction of those resources are not pledged

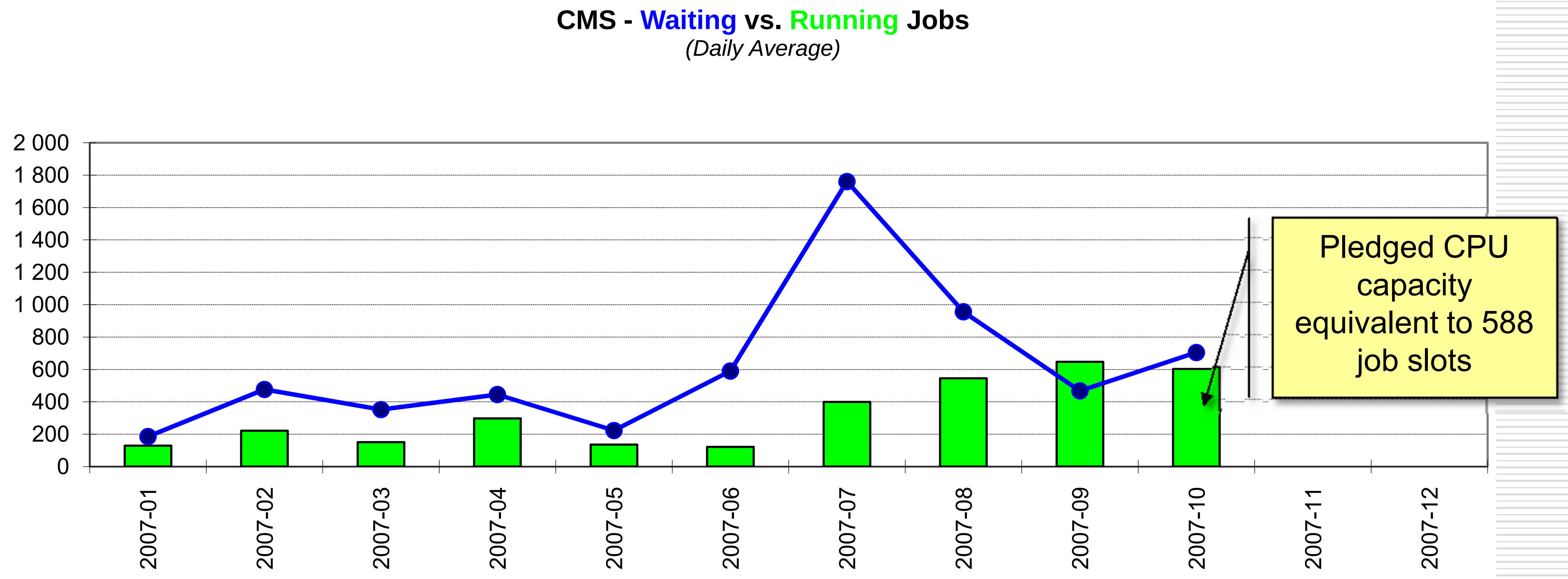
Planned vs. Actual Contribution

CMS



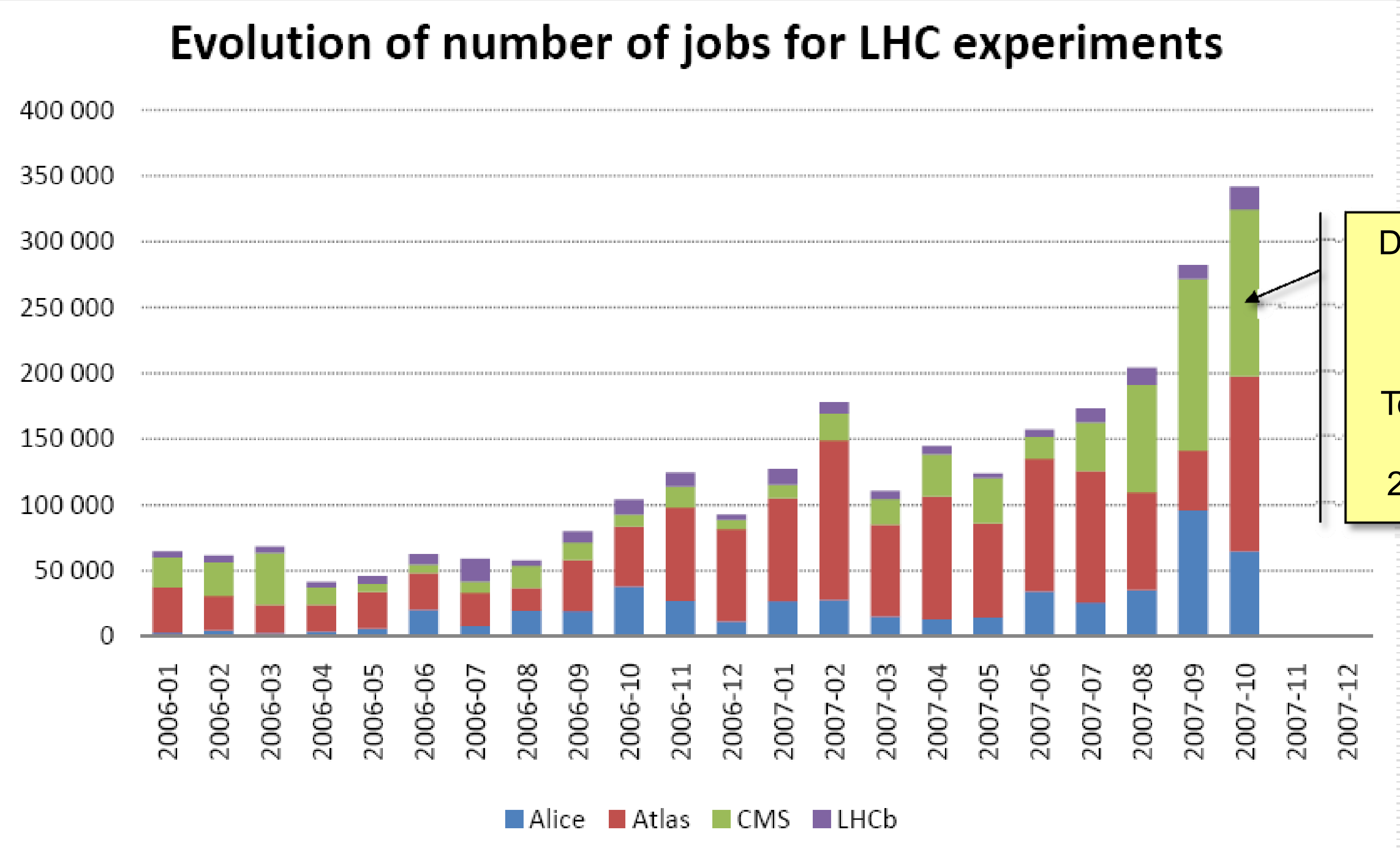
Source: http://lcg.web.cern.ch/LCG/MB/accounting/accounting_summaries.pdf

Resources: CPU



Pledged CPU capacity equivalent to 588 job slots

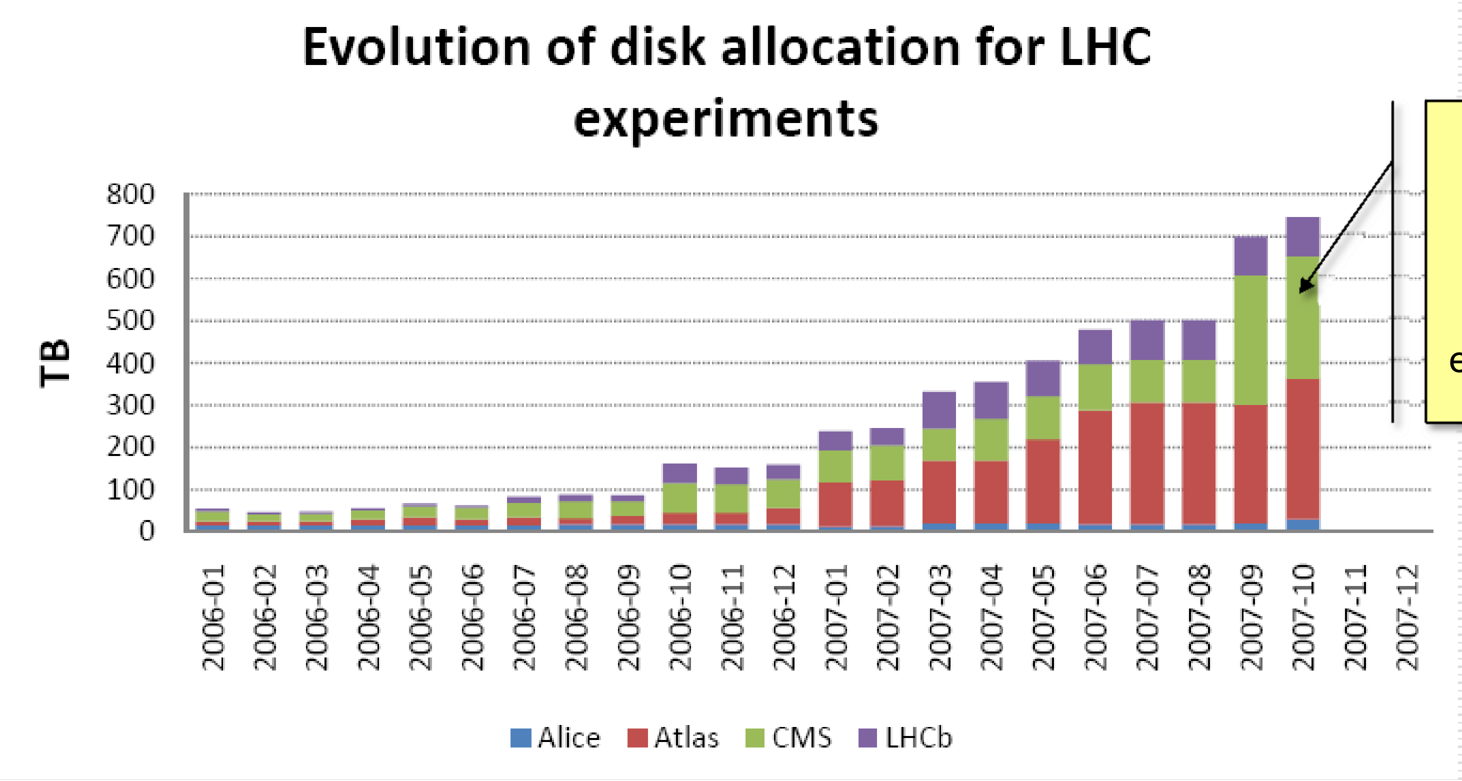
Resources: CPU



Daily throughput for CMS:
4000 jobs

Total throughput of the site:
28000 jobs/day

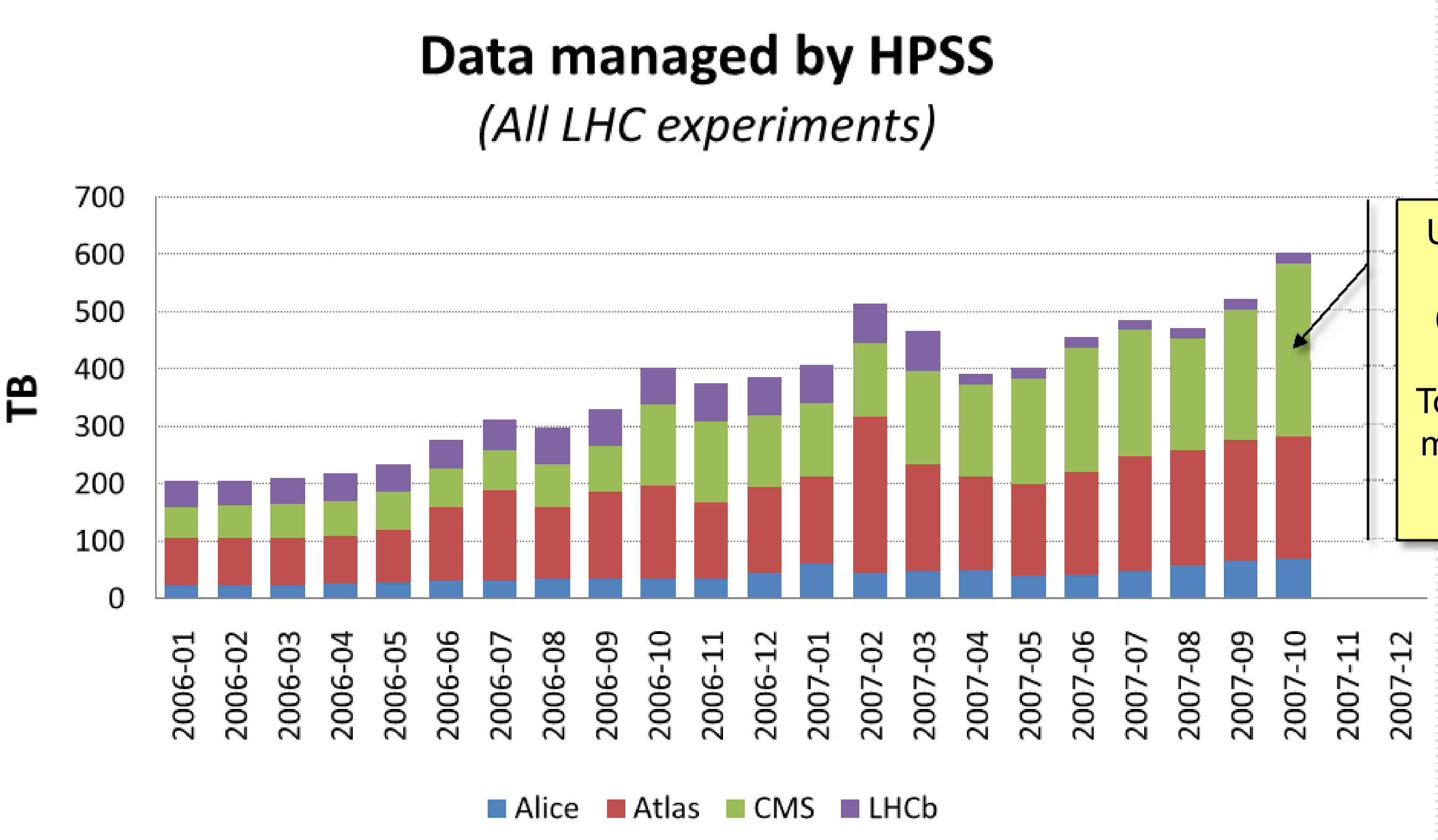
Resources: disk



Allocation: 292 TB (pledged: 237 TB)

Total allocation (all supported experiments) : 1.2 PB

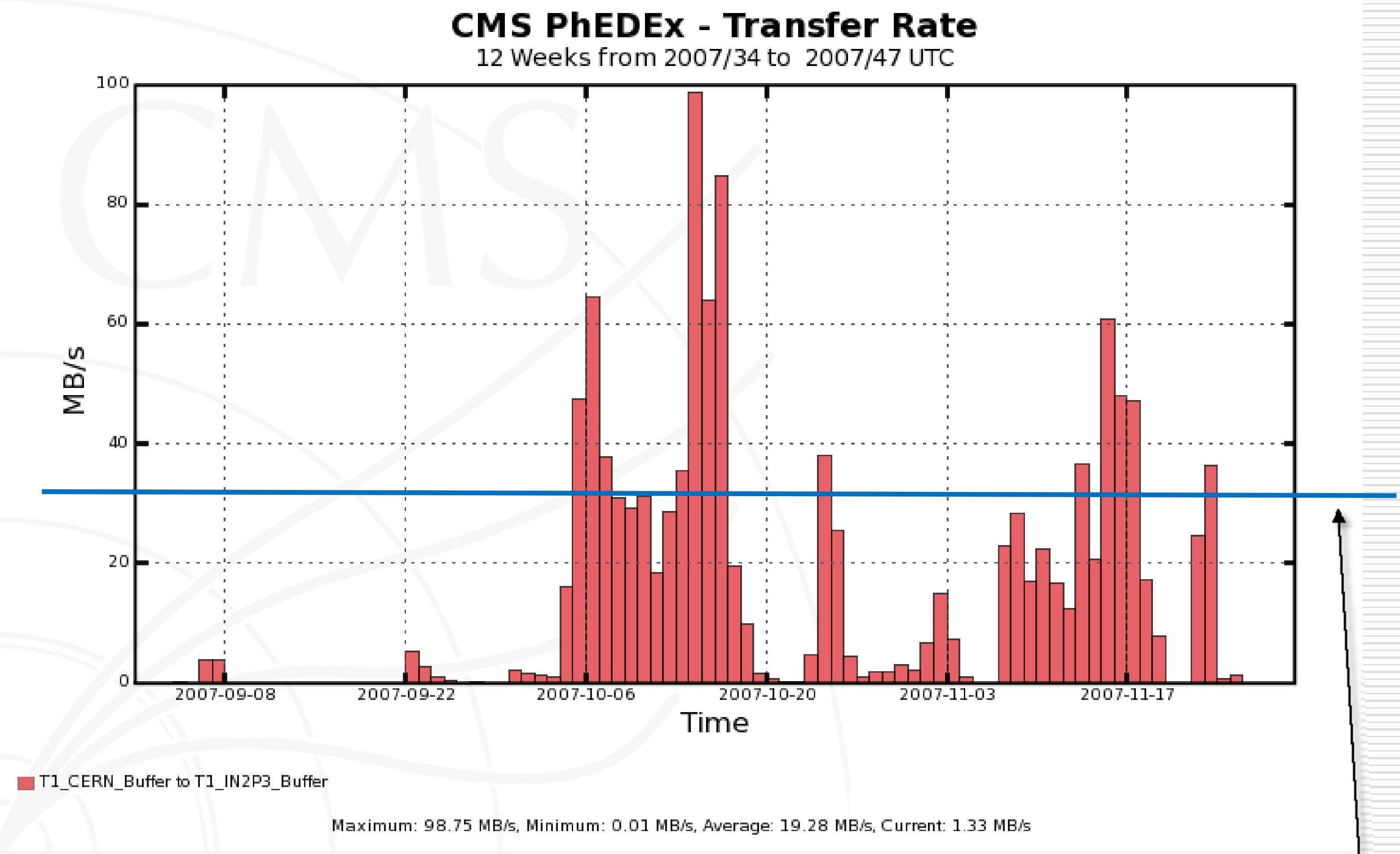
Resources: MSS



Utilization by CMS:
304 TB
(pledged: 292 TB)

Total amount of data managed by HPSS:
2.7 PB

Data transfer: CERN→CCIN2P3

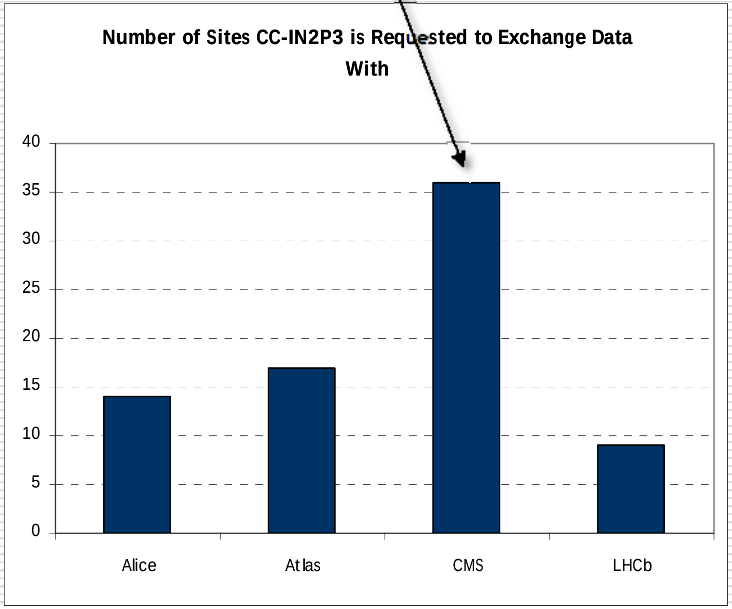


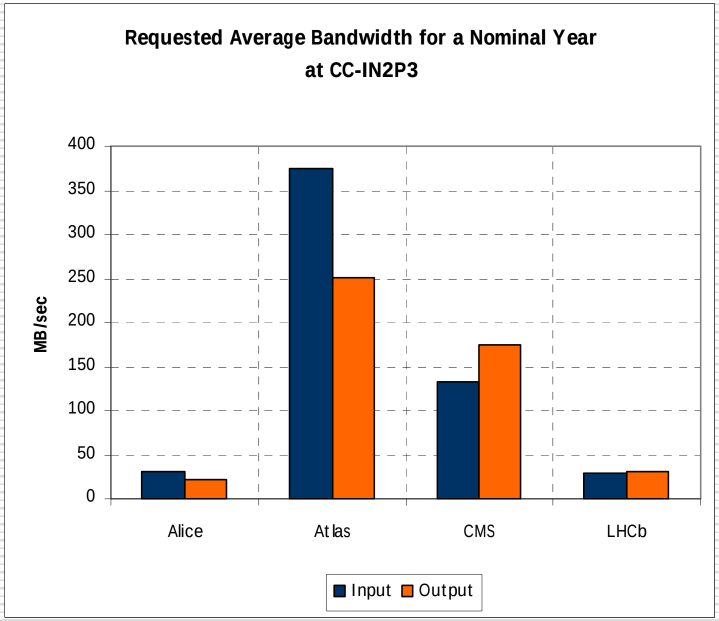
CERN → CCIN2P3 target rate for CMS:
32 MB/sec

WAN bandwidth requirements

The number of CMS sites we have to exchage data (and solve problems) with is worrying!!!

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: Megatable http://lcg.web.cern.ch/LCG/documents/Megatable240107.xls

Procurement

- We have to satisfy several constraints
  - Formal process is long
    - Call for tenders at the European level
  - Budget is approved on a yearly basis
    - « Final word » during last quarter each year
  - Limited machine room space available
    - Extensive in-situ tests to realistically identify the real characteristics of candidate hardware (when possible)
    - Optimize (computing power/m²) but also (computing power/€)
      - *Forecast of running costs performed at this stage*
  - Desired availability of computing equipment in operation by experiments
  - Delays in the deliveries of equipment

Procurement (cont.)

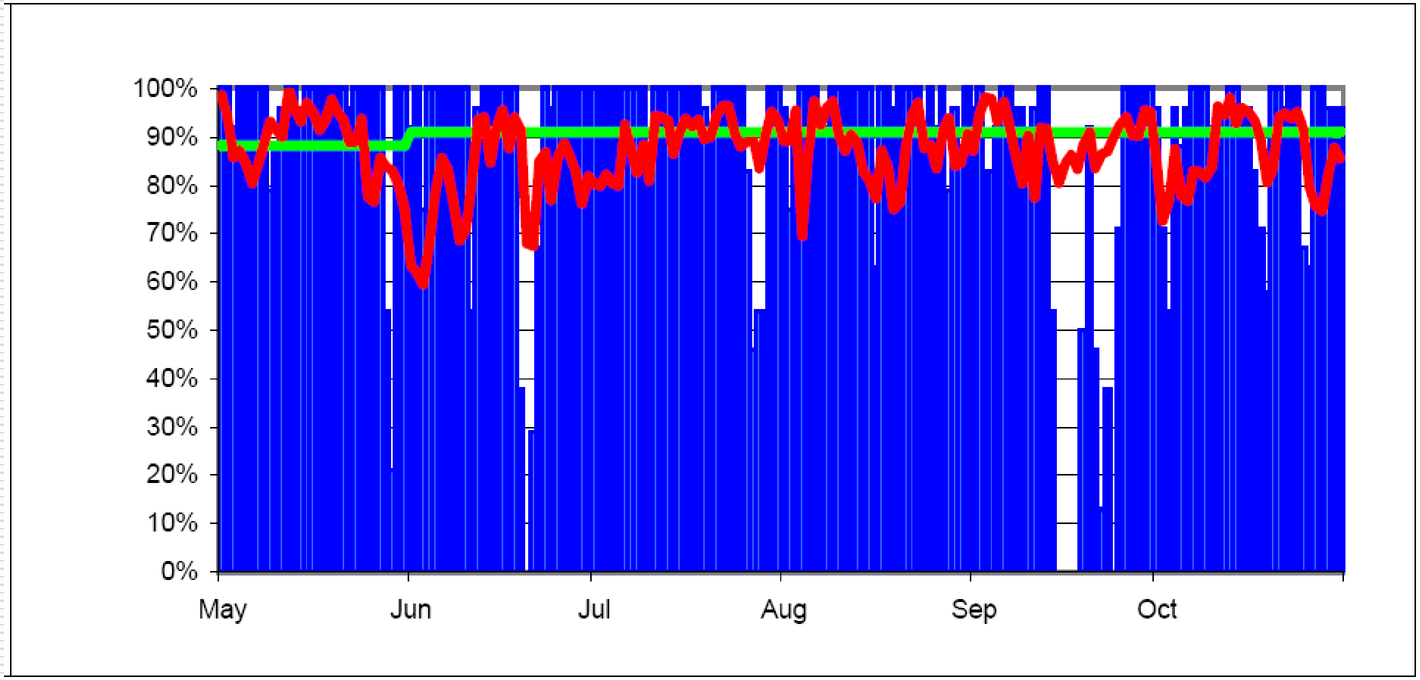
- Starting in 2007, we modified our procurement plan for LHC experiments
  - With budget for year N, purchase at least 40% of the required equipment for year N+1
  - Procurement process for remaining fraction triggered as soon as next year budget is known
- Our experience with this model for this year's procurement makes us confident that we will be in good shape to provide a significant amount of the pledged capacity by April 1st each year

2007: Capacity Increase

- Compute nodes
  - 479 worker nodes
    - DELL Intel Xeon 5345 @ 2.33 GHz, 8 cores, 2 CPUs, 16 GB RAM, 160 GB disk
    - 6.2 MSI2000
  - New bid ongoing
    - Expected ~400 equivalent machines
- Disk servers
  - 1.2 PB
    - Sun X4500 (Thumpers)
  - Order placed for additional 0.6 PB
- Mass storage
  - Cartridges for populating the SUN/STK SL8500 library
    - Both STK T10.000 and LTO4
  - Increase of the cartridge storage capacity: 2 PB

Site Availability

Site availability daily score: May – October 2007

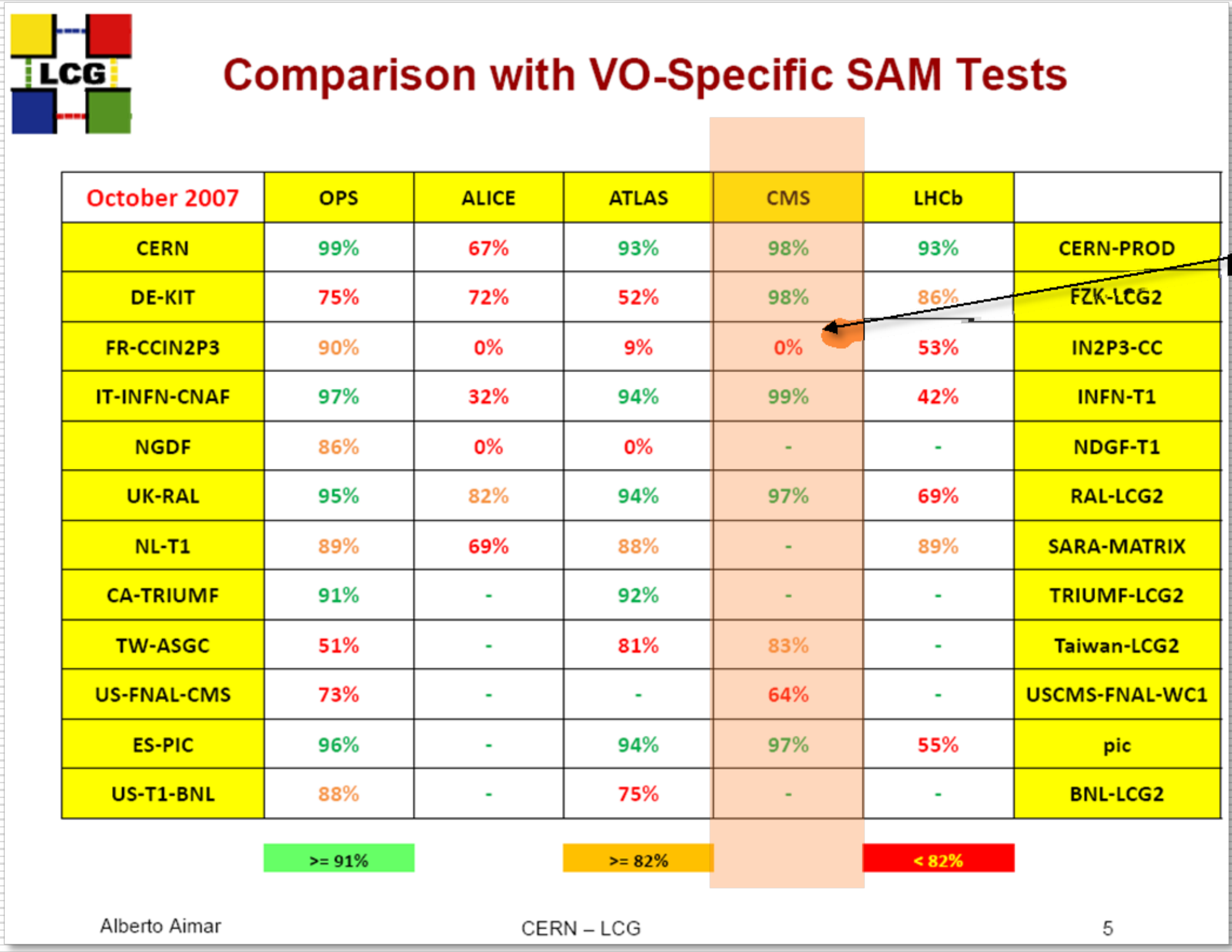


IN2P3-CC **av.reliability last 3 mths** **85%**

Below the target

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

VO-specific tests



Comparison with VO-Specific SAM Tests

October 2007	OPS	ALICE	ATLAS	CMS	LHCb	
CERN	99%	67%	93%	98%	93%	CERN-PROD
DE-KIT	75%	72%	52%	98%	86%	FZK-LCG2
FR-CCIN2P3	90%	0%	9%	0%	53%	IN2P3-CC
IT-INFN-CNAF	97%	32%	94%	99%	42%	INFN-T1
NGDF	86%	0%	0%	-	-	NDGF-T1
UK-RAL	95%	82%	94%	97%	69%	RAL-LCG2
NL-T1	89%	69%	88%	-	89%	SARA-MATRIX
CA-TRIUMF	91%	-	92%	-	-	TRIUMF-LCG2
TW-ASGC	51%	-	81%	83%	-	Taiwan-LCG2
US-FNAL-CMS	73%	-	-	64%	-	USCMS-FNAL-WC1
ES-PIC	96%	-	94%	97%	55%	pic
US-T1-BNL	88%	-	75%	-	-	BNL-LCG2

🟩 >= 91% 🟧 >= 82% 🟥 < 82%

Alberto Aimar CERN – LCG 5

We need to understand why our site is not perceived available by CMS, in spite of the number of jobs and data being constantly transfered to and from the site

Source: Alberto Aimar, Management Board meeting, 20th Nov. 2007, :http://indico.cern.ch/conferenceDisplay.py?confId=22187

Current work

- Top priority: **stability!!!**
- Continue the integration of the grid services to the standard operations
  - Including on-call service
  - Monitoring & alerting, progressively documenting procedures, identifying roles and levels of service, etc.
  - Strong interaction with people developing the grid operations portal (CIC)
- Consolidation of the services
  - Deploy for availability
    - Hardware redundancy
    - Usage of (real or virtual) stand-by machines
  - Including VO Boxes
- Assigning job priorities based on VOMS roles & groups
  - Interim solution in place

Current work (cont.)

- Continous development of BQS
  - for coping with the expected load in the years ahead
  - for making it more grid-aware
    - Keep grid attributes in the job records
      - *Submitter identity, grid name, VO name, etc.*
    - Scheduling based on grid-related attributes
      - *VOMS roles/groups, grid identity, etc.*
    - Allow/deny job execution based on grid identity
  - Development of gLiteCE and CREAM compatible BQS-backed computing element

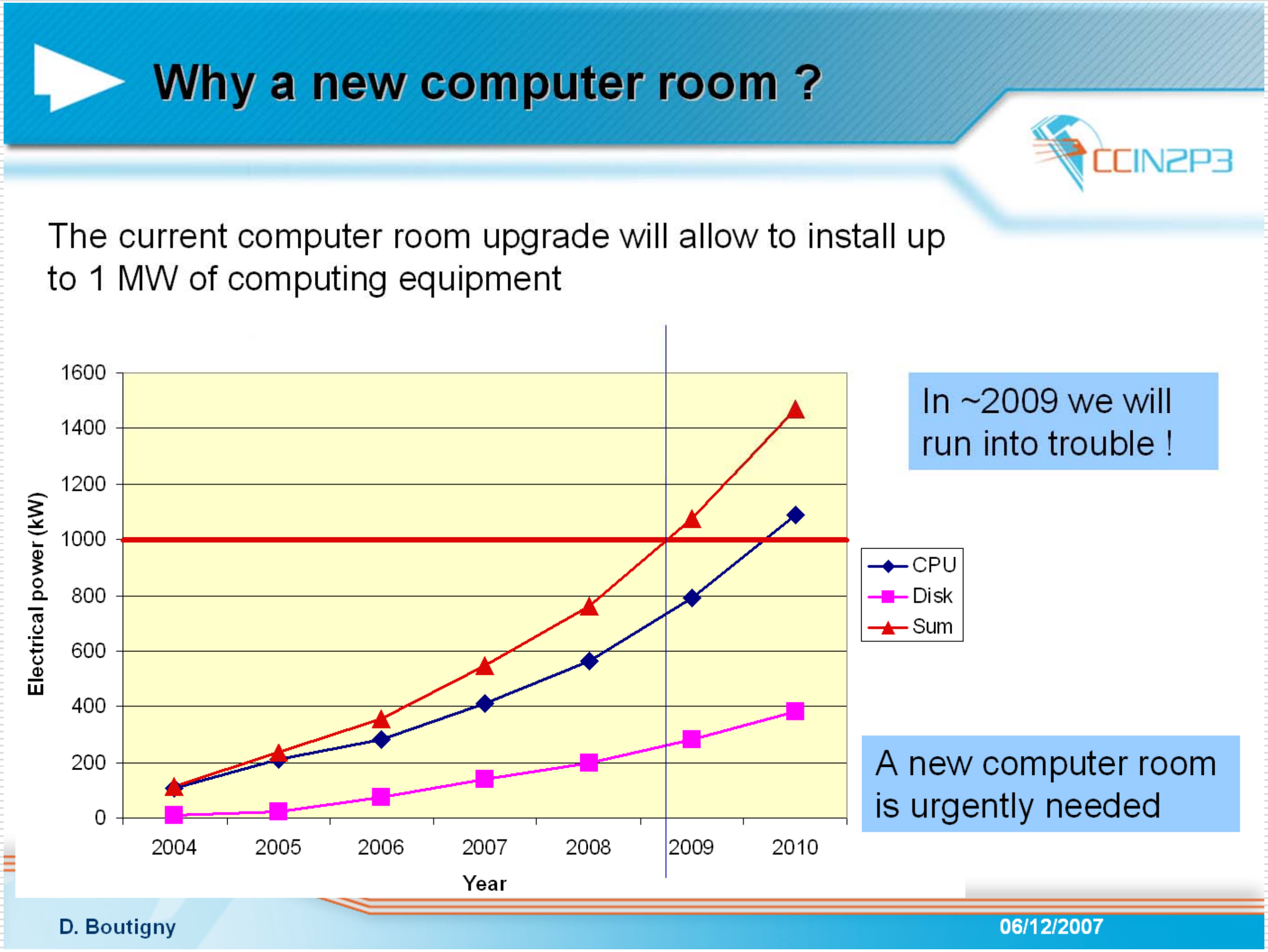
Current work (cont.)

- Storage services
  - Consolidating the AFS service
  - Continuous work for internal reconfiguration of HPSS according to the (known) needs of LHC experiments
  - Planned upgrade of the hardware for the core machines of dCache/SRM
- Trying to understand how the data will be accessed
  - What are the required rates for data transfers between MSS→disk→worker nodes and backwards…
  - ..for each one of the several kind of job (reconstruction, simulation, analysis, …)
- Job profiling
  - Studying the observed usage of memory and CPU time for LHC jobs
    - Memory requirements have a significant impact on budget and on the capacity of our site to efficiently exploit the purchased hardware
- Understanding how to build an Analysis Facility

Facility Upgrade

- Last July, a major effort for upgrading the electric and cooling infrastructure of the site was finished
  - From 500 kW to 1000 kW of electrical power usable for computing equipment
    - +600 kW for cooling
  - Improvements include
    - New diesel generator (880 kW, 72h autonomy)
    - 2 additional UPS (500 kW each)
    - Significant improvement of electrical distribution
    - 1 additional liquid cooler, pipe network for cool water, 7 additional chilled water units (for the machine room and for the UPS)
- Budget: more than 1,5 M€
  - 2+ years-long project
- People have done (almost heroic) efforts to maintain the site in (near normal) operating conditions

New building



Why a new computer room ?

The current computer room upgrade will allow to install up to 1 MW of computing equipment

In ~2009 we will run into trouble !

A new computer room is urgently needed

D. Boutigny

06/12/2007

Courtesy of Dominique Boutigny

New building (cont.)

- 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
- Encouraging signals recently received regarding funding, but the second machine room won't be available before 2010 as needed
  - Currently evaluating temporary solutions (offsite hosting of the equipment, shorten renewal intervals by leasing the hardware, use of modular transportable machine rooms to be installed in the parking lot)

New building (cont.)

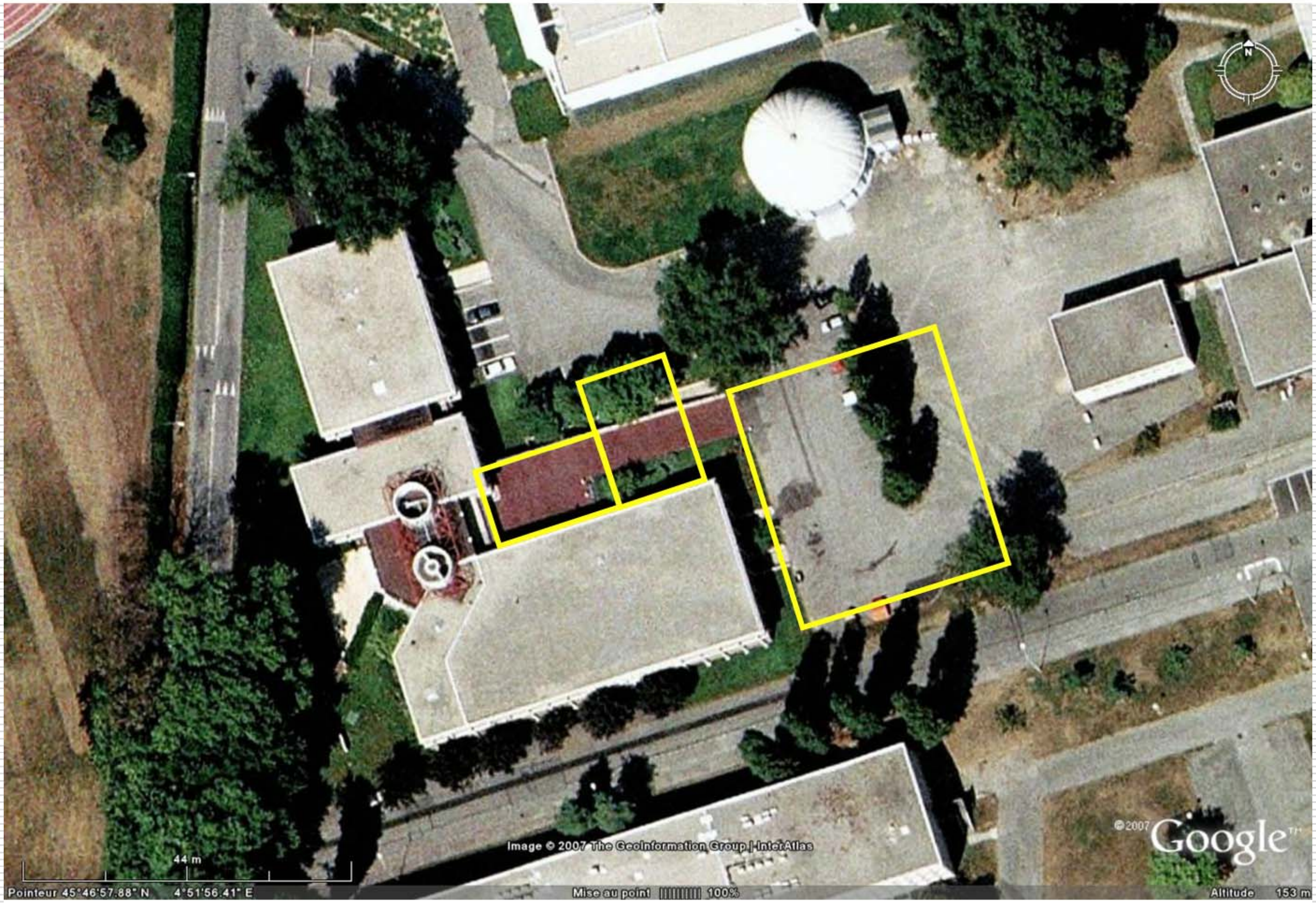
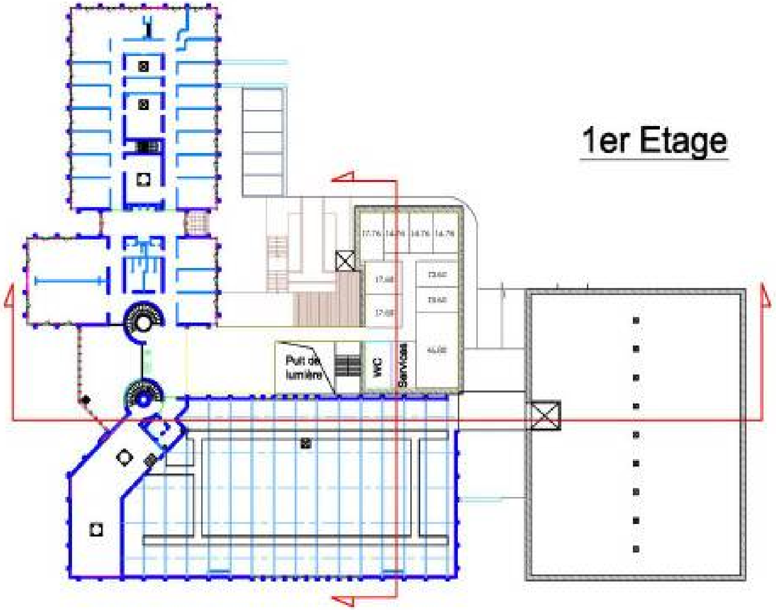


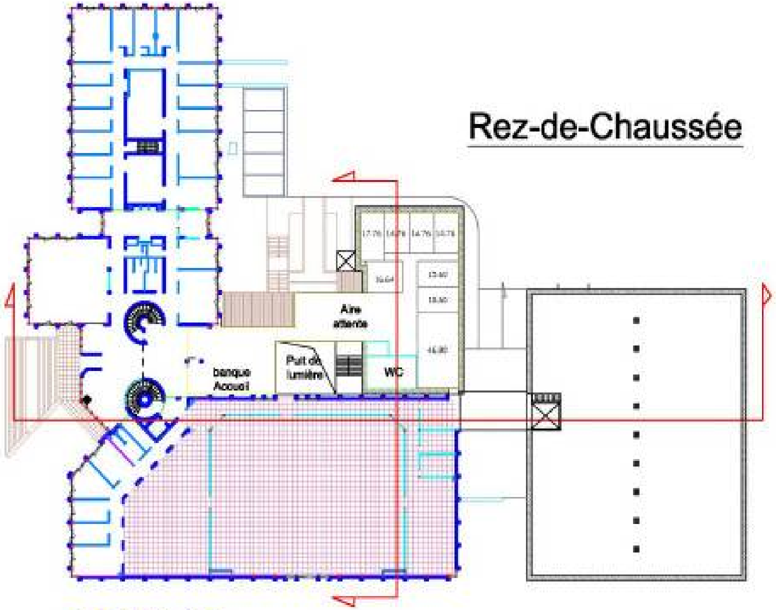
Image © 2007 The GeoInformation Group | InterAtlas
©2007 Google
44 m
Pointeur 45°46'57.88" N 4°51'56.41" E
Mise au point 100%
Altitude 153 m

Courtesy of Dominique Boutigny

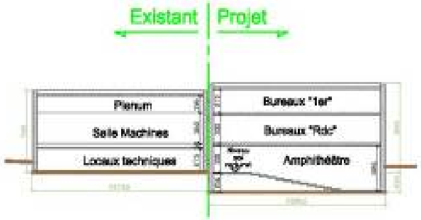
1er Etage



Rez-de-Chaussée

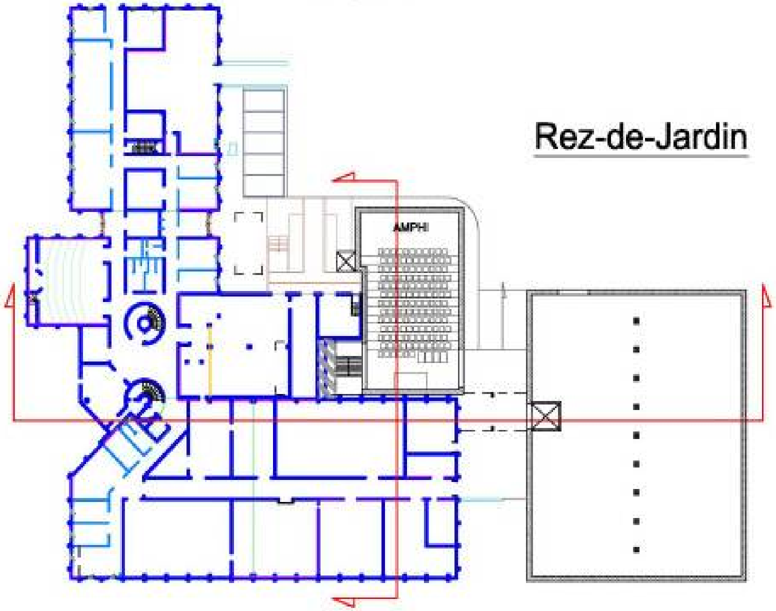


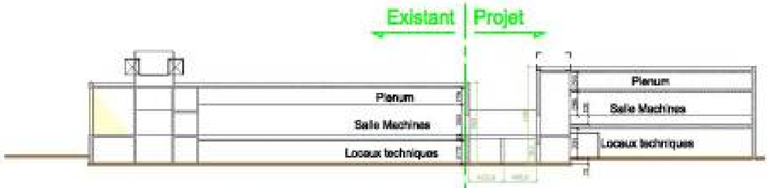
Quelques surfaces utiles en m²	
Local technique	850
Salle Informatique	845
Plenum	845
Passerelle	8,88 x 4,47 m.
Amphitéâtre (142 places)	245
Attente amphi.	85
Accueil (banque)	46
Terrasse rdc	38
Terrasse ascenseur rdc	8
Terrasse 1er	40
Puit de lumière	45
Espace de repos (1er étage)	110
SHOB	~ 5365 m²
SHON	~ 3820 m²
Emprise au sol (extension)	1477 m²



Coupe de principe (verticale)

Rez-de-Jardin





Coupe de principe (longitudinale)

Centre de calcul de l'IN2P3

27 Bvd du 11 Novembre 1918
69622 Villeurbanne Cedex

Esquisse n°7 - Rectificatif :

Salle informatique de plain pied
Amphithéâtre semi-enterré

Màj : 02/03/07

What’s next today

- In the coming presentations you will find the detailed status and plans of the
  - Storage infrastructure and data transfers
  - Grid services
  - Site operations
  - Network infrastructure

Questions

?

Fabien Wernli, 2006