November 2008

CC-IN2P3

Status and Prospects

FJPPL / FKPPL Workshops

dapnia
cea
saclay

CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

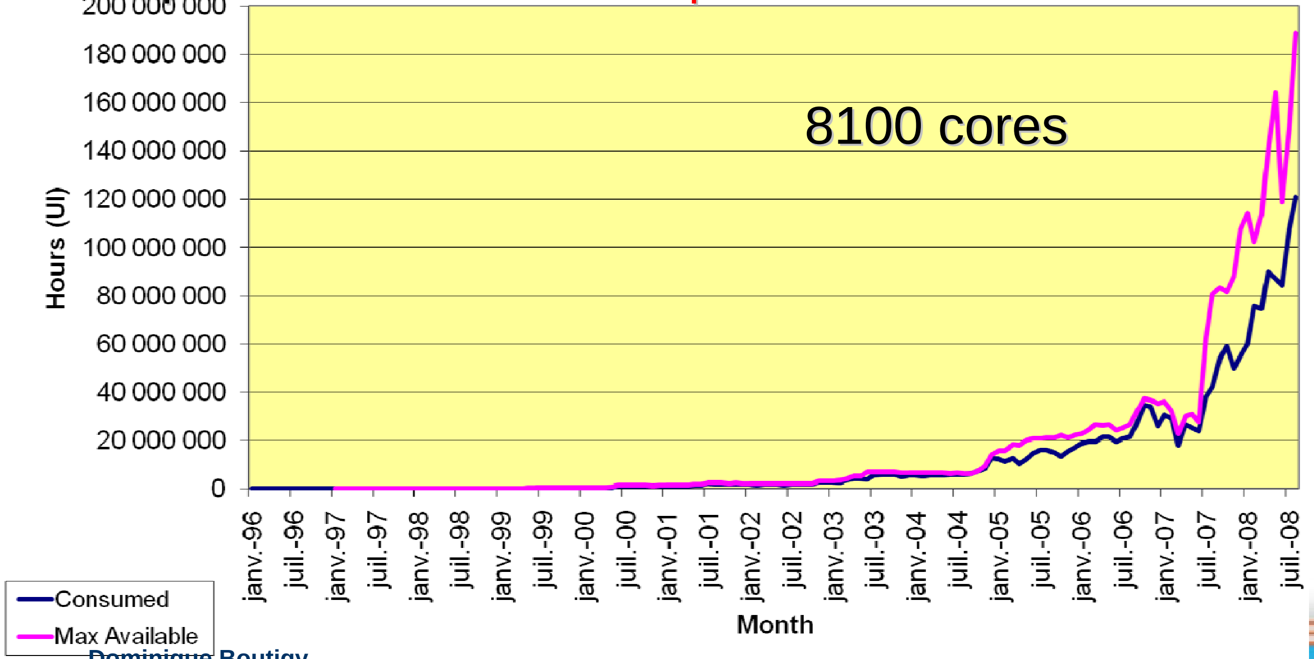
CPU

Computing farm:

Top 500

- ❑ 265 IBM dual CPU – dual core Operon -1.2 MSpecHEP
- ❑ 479 DELL dual CPU / quad core INTEL 5345 @ 2.33 GHz – 6.3 MSpecHEP
- ❑ 401 DELL dual CPU / quad core INTEL 5450 @ 3.0 GHz – 6.5 MSpecHEP

Total power available: ~14 MSpecHEP

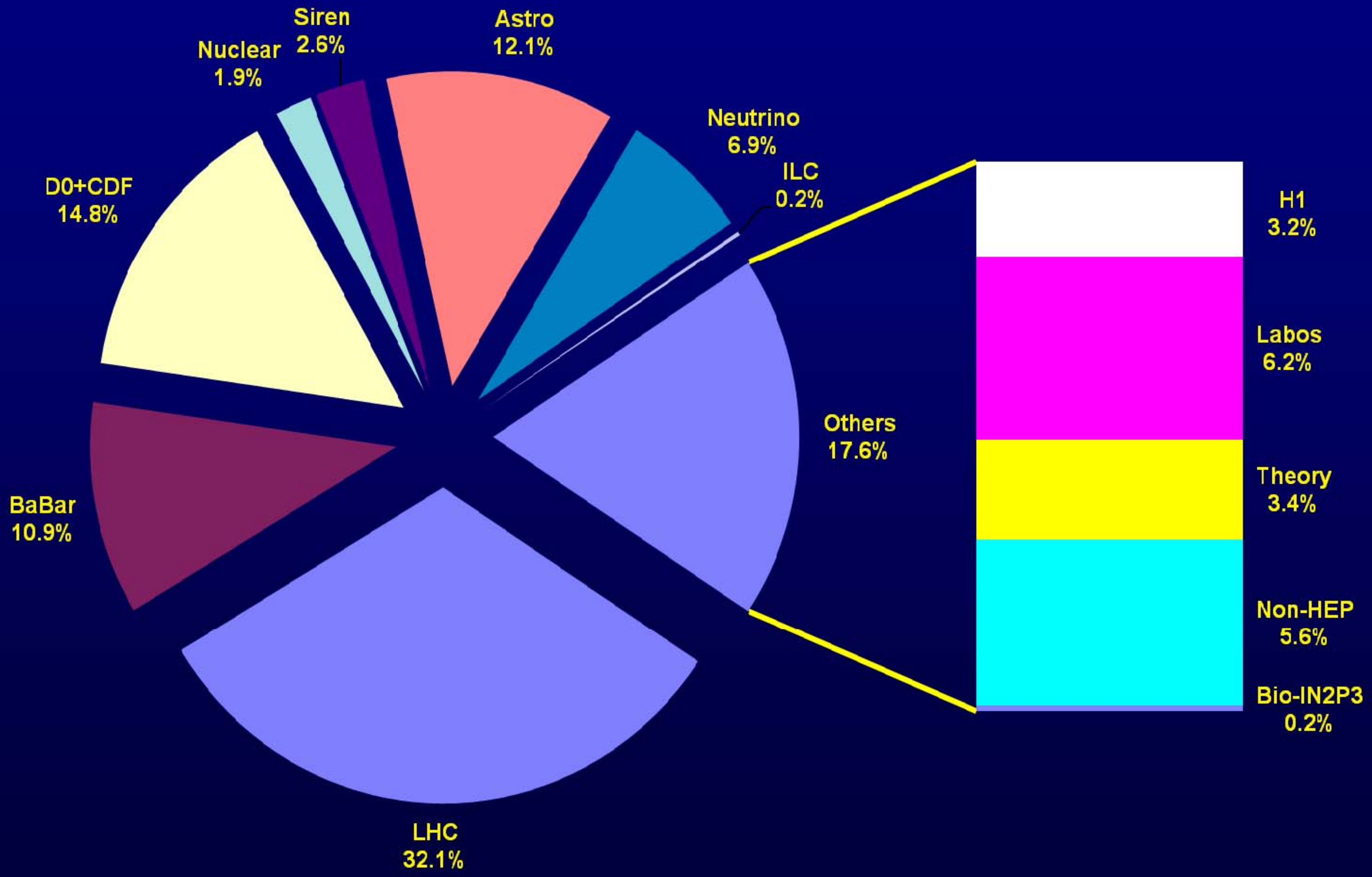


8100 cores

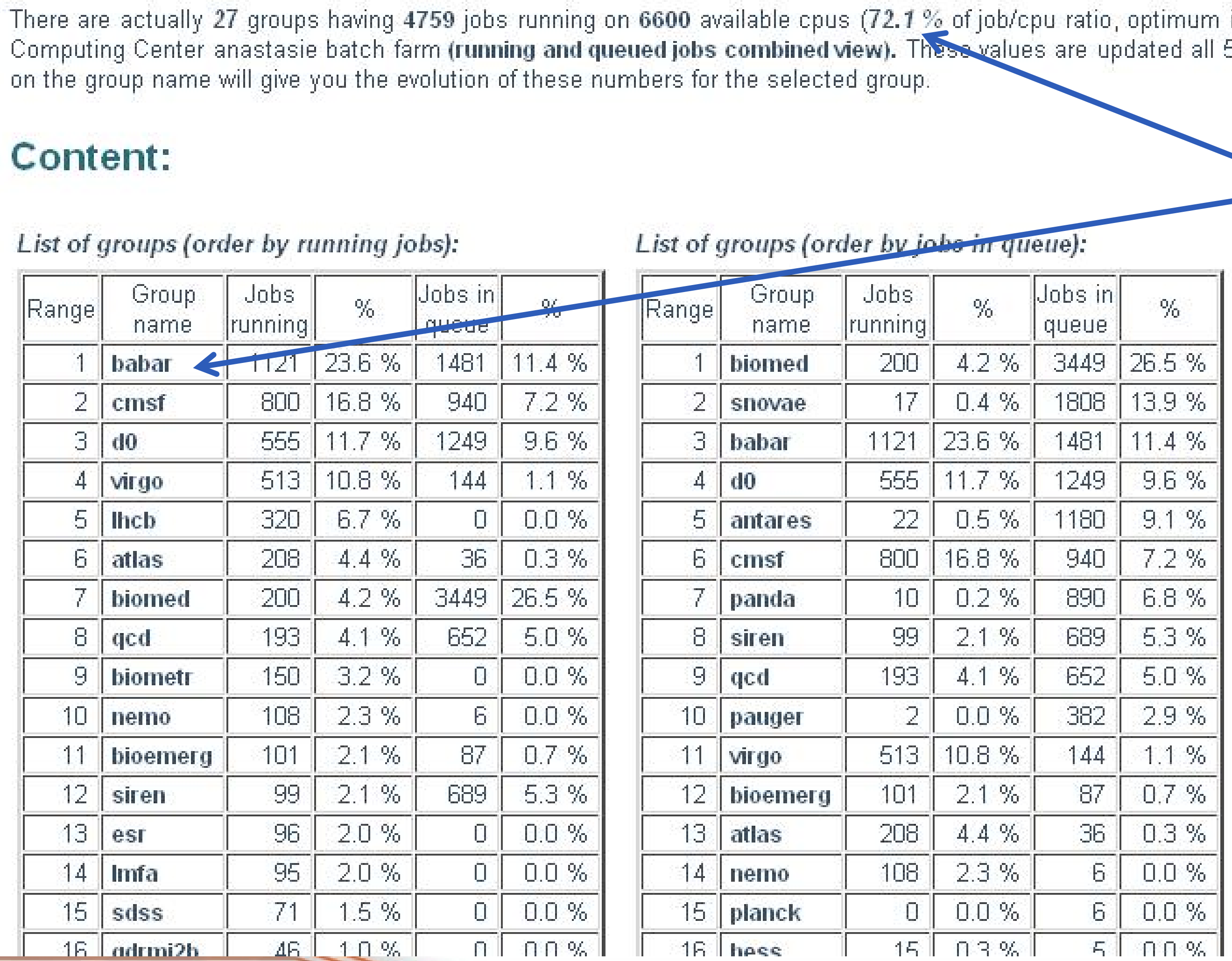
SpecHEP ~ SpecInt2000 x normalization

Batch worker lifetime is now < 3 years

CPU Consumption in 2008



The LHC glitch effect…



There are actually **27** groups having **4759** jobs running on **6600** available cpus (**72.1** % of job/cpu ratio, optimum is Computing Center anastasie batch farm **(running and queued jobs combined view).** These values are updated all 5 on the group name will give you the evolution of these numbers for the selected group.

Content:

List of groups (order by running jobs):

Range	Group name	Jobs running	%	Jobs in queue	%
1	**babar**	1121	23.6 %	1481	11.4 %
2	**cmsf**	800	16.8 %	940	7.2 %
3	**d0**	555	11.7 %	1249	9.6 %
4	**virgo**	513	10.8 %	144	1.1 %
5	**lhcb**	320	6.7 %	0	0.0 %
6	**atlas**	208	4.4 %	36	0.3 %
7	**biomed**	200	4.2 %	3449	26.5 %
8	**qcd**	193	4.1 %	652	5.0 %
9	**biometr**	150	3.2 %	0	0.0 %
10	**nemo**	108	2.3 %	6	0.0 %
11	**bioemerg**	101	2.1 %	87	0.7 %
12	**siren**	99	2.1 %	689	5.3 %
13	**esr**	96	2.0 %	0	0.0 %
14	**lmfa**	95	2.0 %	0	0.0 %
15	**sdss**	71	1.5 %	0	0.0 %
16	**adrmi2b**	46	1.0 %	0	0.0 %

List of groups (order by jobs in queue):

Range	Group name	Jobs running	%	Jobs in queue	%
1	**biomed**	200	4.2 %	3449	26.5 %
2	**snovae**	17	0.4 %	1808	13.9 %
3	**babar**	1121	23.6 %	1481	11.4 %
4	**d0**	555	11.7 %	1249	9.6 %
5	**antares**	22	0.5 %	1180	9.1 %
6	**cmsf**	800	16.8 %	940	7.2 %
7	**panda**	10	0.2 %	890	6.8 %
8	**siren**	99	2.1 %	689	5.3 %
9	**qcd**	193	4.1 %	652	5.0 %
10	**pauger**	2	0.0 %	382	2.9 %
11	**virgo**	513	10.8 %	144	1.1 %
12	**bioemerg**	101	2.1 %	87	0.7 %
13	**atlas**	208	4.4 %	36	0.3 %
14	**nemo**	108	2.3 %	6	0.0 %
15	**planck**	0	0.0 %	6	0.0 %
16	**hess**	15	0.3 %	5	0.0 %

Notice !

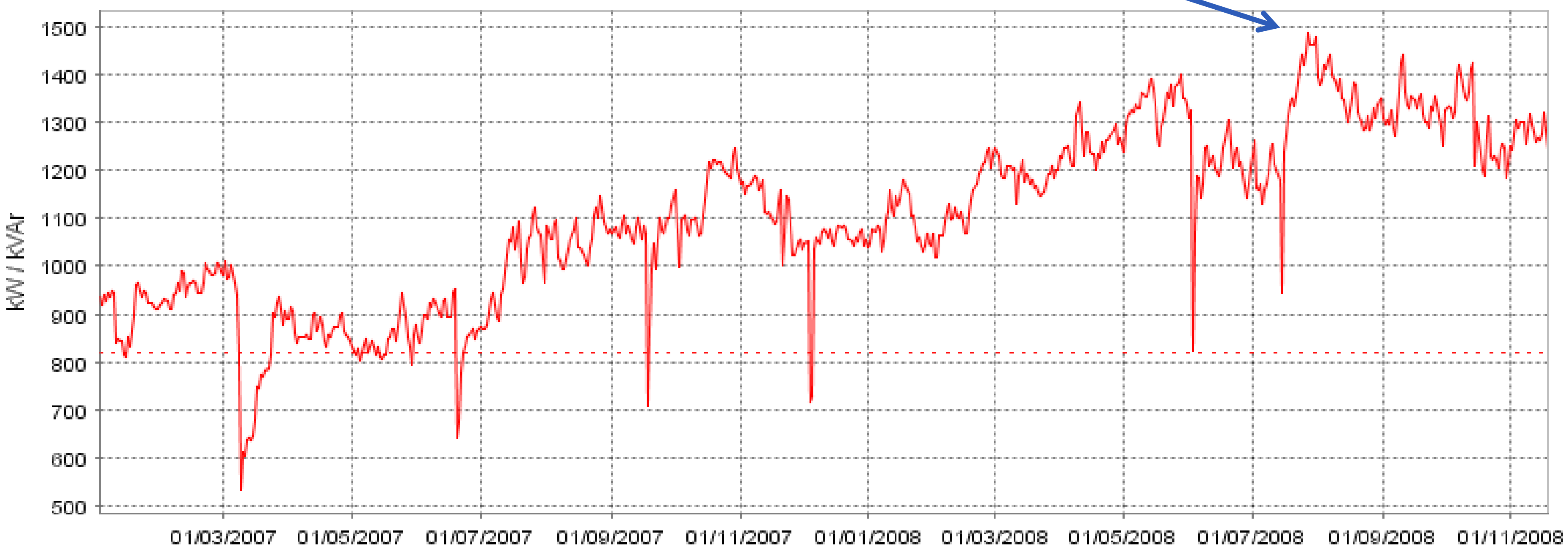
Storage

4 main systems:

- ❑ IBM DS8300: 250 TBytes – High performance SAN system : AFS + several other services
- ❑ SUN X4500 : 2.3 Pbytes + 3.6 Pbytes – DAS: dcache, xrootd
- ❑ IBM / DDN DCS9550
  - ➢ 640 TBytes : GPFS
  - ➢ 480 Tbytes : HPSS cache
- ❑ PILLAR Axiom: 42 TByte for Oracle databases

Electrical Power

1.5 MW reached during the hottest days this past summer



Infrastructure constraints

CC-IN2P3 survived thanks to a rental chiller unit connected to a rental electrical transformer

Current limits are coming from:

- ❑ Maximum allowed electrical power in cooling units: ~750 kW
  - ➔Settled by environmental constraints
- ❑ No real limitation on the power source itself (>10 MW)

The limit on the cooling power is very strong

Need at least 1.5 year to renegotiate

Infrastructure upgrade

New transformers : 3 x 1 MW ➔ will double electrical capacity

Old chiller units working with underground water will be decommissioned

Install a new chiller unit

➔ Total of 3 x 250 kW units working with air exchangers

We will have to survive within this cooling budget until we get a new building

Cost: 0.5 M€

How to survive ?

Improve the cooling efficiency

Air cooling ➔ Water cooling

New tender for CPU includes procurement and installation of water cooled racks

Front

Rear

Rear door heat exchanger

IBM i-dataplex system

252 servers – Intel 5430 LV – 2.66 GHz – 50 W

52 kW total – ~No heat outside the racks

As efficient as closed water cooled rack + blades, but cheaper

How to survive ? (2)

We already have 1 42 U rack hosted in Montpellier (300 km from Lyon)

The Montpellier center (CINES) will be able to host up to 6 racks (10 ?)

A 10 Gb/s link will be available soon between the 2 centers

If it is not yet enough we will have to rent some space in a private computer hosting company ➔ very expensive (~500 k€ / year for 10 racks)

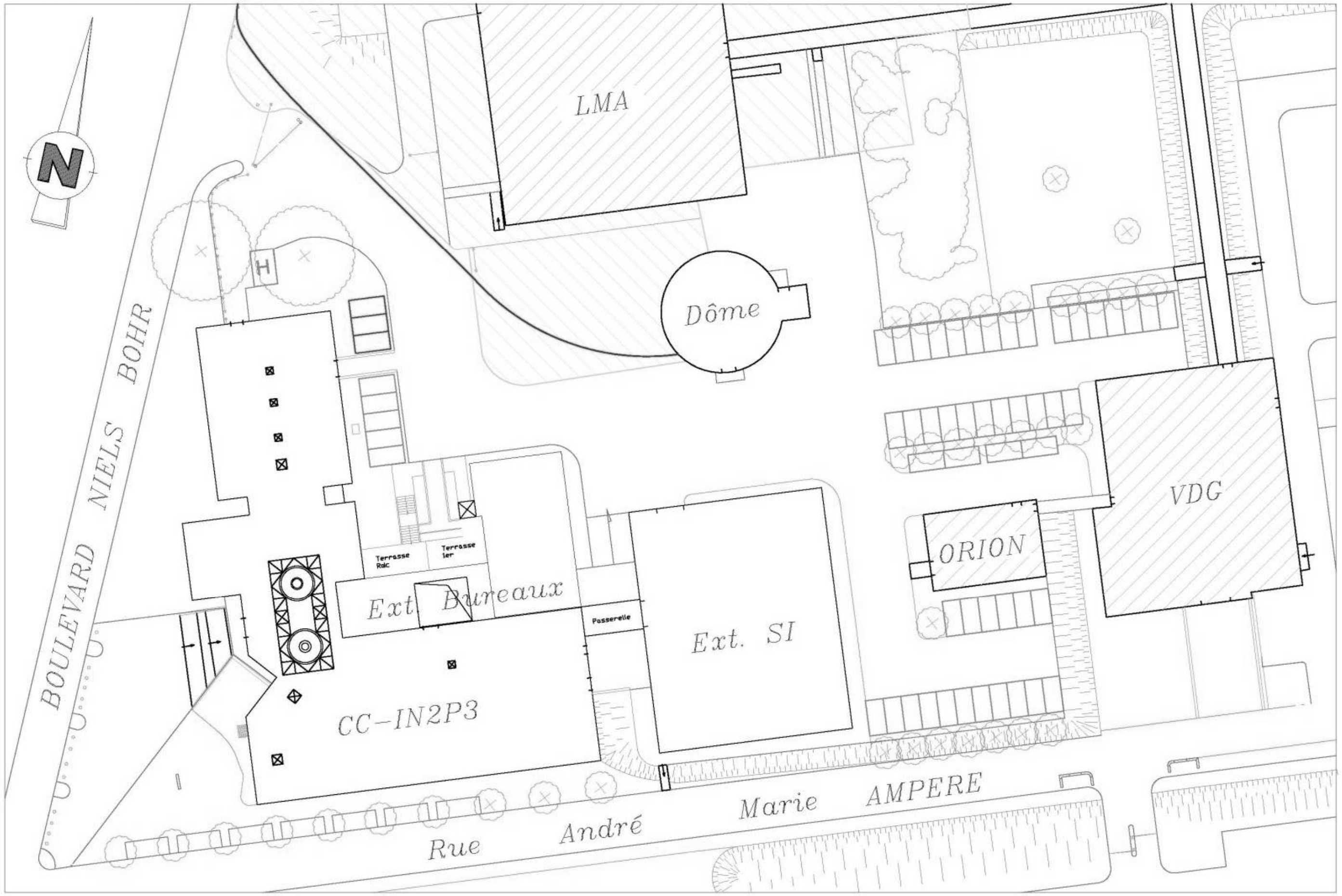
New building

Centre de calcul de l'IN2P3

Plan Masse - Esquisse n°7

Mise à jour le 02/03/07

Format A3 : Echelle 1/500



New building

- 800 m^2 computer room
- 800 m^2 infrastructure room underneath
  - Services: electricity, UPS, cooling etc.
- Will start with limited power (due to limited budget) but will increase up to at least 3.5 MW
- Target for start of operation: end of 2010

Aim of the PetaQCD project

- Design a PetaFlop cluster dedicated to LQCD simulations
  - Lattice Quantum ChromoDynamics
    - The canonical 4D lattice size will be $256*128^3$
  - Specifically targeting the Hybrid Monte Carlo application family
    - Which exhibits some kind of 'evil parallelism'
    - We are usually working with the European Twisted Mass Collaboration implementation
- Trying to stay into reasonable limits
  - 2000 (4000 ?) nodes
    - each node may be a multi-core multi-processor
    - issues: volume, cooling, power consumption, price/TCO, …
  - 20 (40-50 ?) cabinets
    - issues: volume again, networking, price, …
  - using off-the-shelf components, wherever possible
- A smaller size mock-up cluster should be built by 2010
  - as a proof-of-concept
  - to establish clear performance baselines
- Grant allocated by ANR last July (probable kick-off Jan '09)
  - Nine French teams involved, from: CNRS, CEA, INRIA + 2 SMEs

Slides from Gilbert Grosdidier (LAL)

Analysis farm

- LHC experiments will need to have access to a farm optimized for analysis
  - Interactive work oriented
  - Fast access to data
  - Fast turn over
    - Able to re-run many times on the same set of data
- Should not interfere with production activities
- State of the reflection already well advanced in ALICE

  ➔ PROOF test bed at CERN
- A national analysis facility is being built in Germany
- CC-IN2P3 will host such a farm ➔ 1 engineer in charge of the project
  - Would like to work closely with SLAC on this project
  - CC-IN2P3 would like to setup something common for all experiments

Status of '*Coyote*' test-bed in CCIN2P3

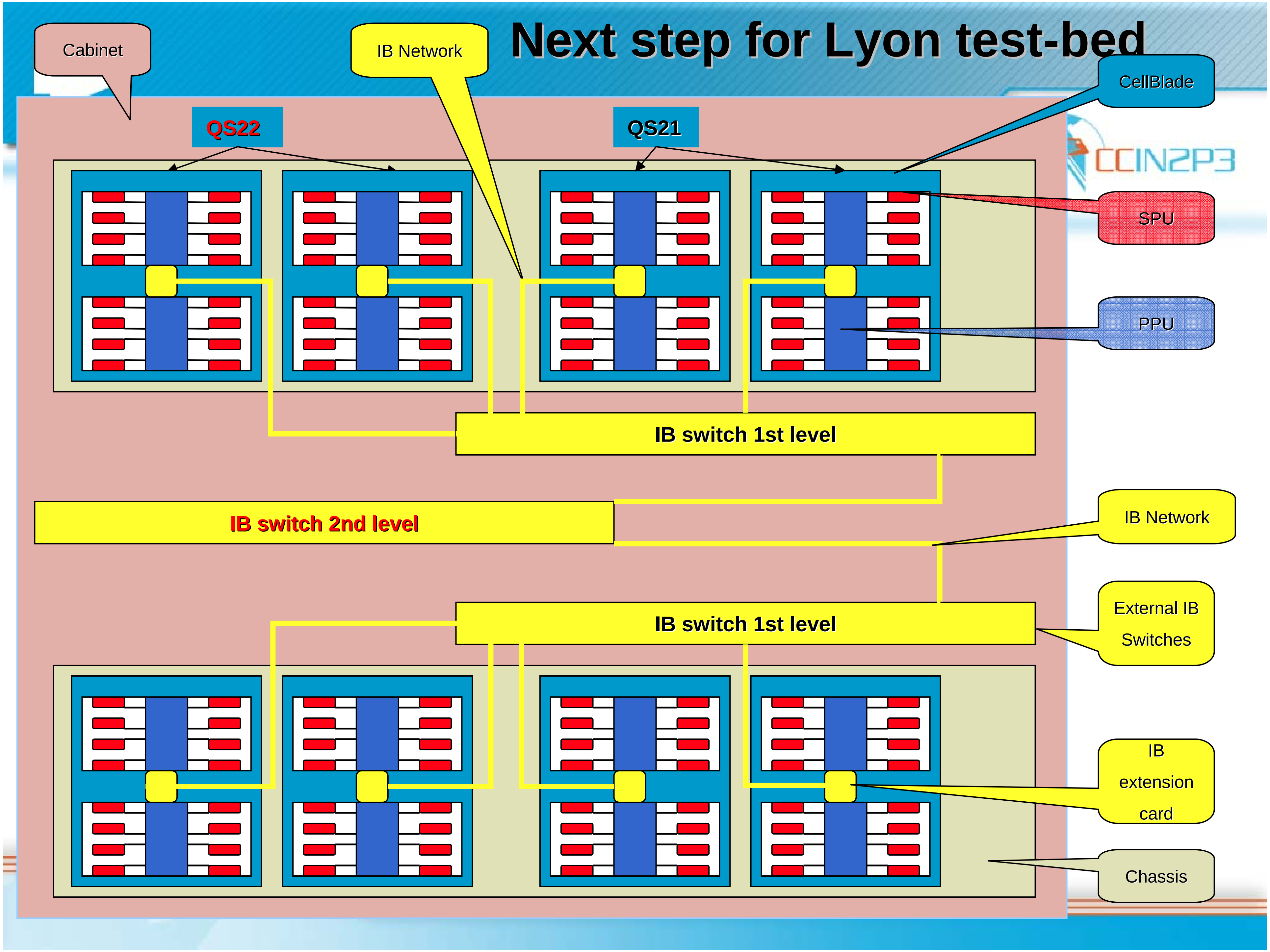
- IBM-Cell hardware installed mid-July (mostly funded by CC itself)
  - 8 CellBlades with IB: 4xQS21 + 4xQS22 (with Enhanced Double Precision)
  - 1 Voltaire InfiniBand switch
  - 1 x86_64 front-end for cross-development, and job submission
- RHEL 5.2 + OpenMPI 1.2.5 (native IB support) + OFED 1.2.5

Slides from Gilbert Grosdidier (LAL)

- First full OpenMPI test 4 weeks ago with
  - *hmc_tm* and also *invert* (HMC, 8^4 lattice)
  - only PPUs were used (no SPU yet in this test)
  - all 16 processors in use (2 PPUs per blade)
  - data messaging over IB and/or shared memory

  → tests show that results for *invert* are fully compatible with Intel Xeon results, w/ same or different partitioning values. Hence, OpenMPI/IB are validated.
- Standalone SPU DP performance (w/ embedded data, HMC kernel)
  - QS21: 1.0 GFlop/sec/SPU; QS22: 2.5 GFlop/sec/SPU
- Infiniband Network performance (2-way exchanges, w/ barrier)
  - 160 μsec for 2x100 kB messages (between 2 Blades, through OpenMPI)

Next step for Lyon test-bed



EGI / NGI

Discussion are going on, concerning the future National Grid ➔ CNRS Grid Institute

13 EGEE nodes are now running in France

Lyon (**Lyon University + CC-IN2P3** **)** is candidate to host the EGI.org headquarters

EGI.org in charge of the grid operation and other central functions (user support and training management, middleware certification and distribution,..) – Not in charge of middleware development

Would be hosted in a building close to CC-IN2P3 : 50 FTE

Regional Grid

Regional grids will be major components of the NGI

CC-IN2P3 recently launched an initiative to create an EGEE regional grid based on existing nodes in Annecy, Grenoble and Lyon

A regional VO has been created: *vo.rhone-alpes.idgrilles.fr*

Now in the process to attract regional users and to help them to port their applications on the Grid

Yonny Cardenas is in charge to coordinate the deployment of the Regional Grid

Some thoughts about the relationship with HPC

Traditionally computer farms and grids have been very disconnected from the HPC world in France (and probably elsewhere)

We would like to establish a bridge between the 2 worlds

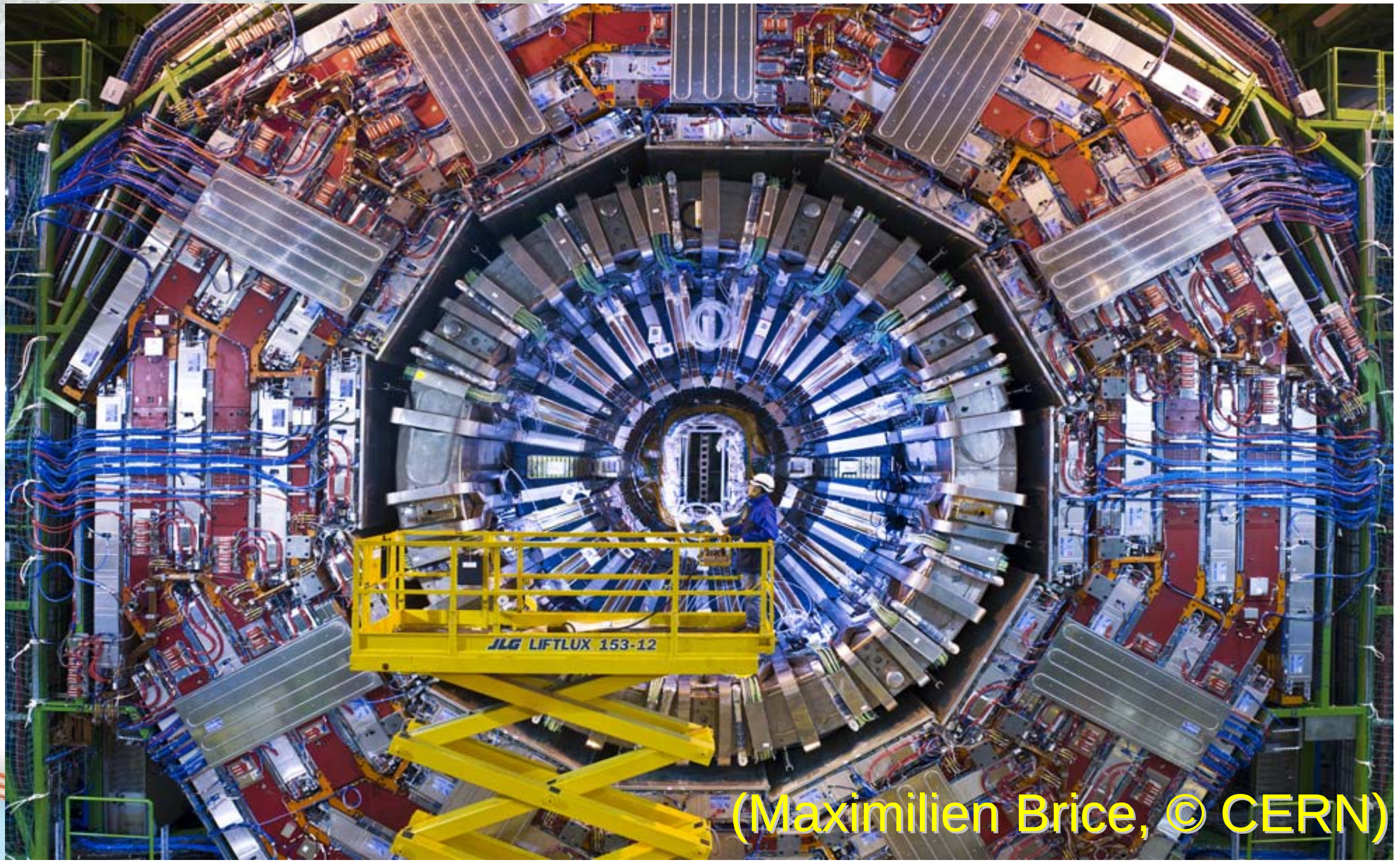
To be able to direct EGEE jobs toward parallel computers

➔ Will be a topic of work in the FKPPL framework

Grid storage protocols developed for HEP are probably suitable to store HPC outputs which will be post-processed on Grids or on smaller centers

Probably the same for SRB / i-RODS

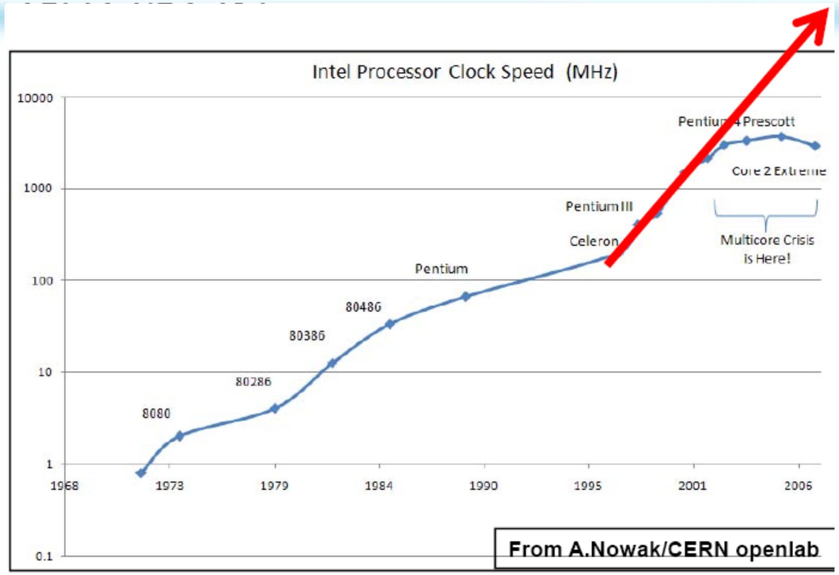
Mass of data



(Maximilien Brice, © CERN)

A remark on Computer architecture

Credit: Sverre Jarp (CERN)



From A.Nowak/CERN openlab

Evolution toward architecture with multi / many cores

Tomorrow PCs with 128, 512 or 1024 cores will be common

Buildings blocks for HEP applications

The current model with 1 job running on 1 core is going to fail very soon: Inefficient – It requests too much memory

This is a real challenge that the HEP community should face

Would that be a topic for FJPPL / FKPPL ?