A Petaflop machine: why and how?

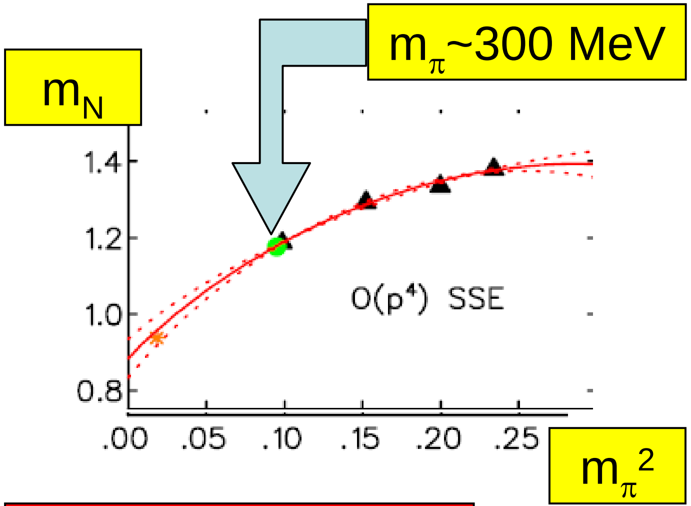
- Why petaflops?
- Prospects in other countries
- Model of a Petaflop machine
- Hardware activities?
- Software activities
- Error control/recovery
- Overall strategy

Why Petaflops? (in general)

http://theory.fnal.gov/theorybreakout2007/

- Fundamental param. (m_q, α_s,V_{ckm})
  - ➢ α_s,V_{ckm} already few % with 50 Tflops
  - ➢ K-K, B-B oscill. 100-500 Tflops (physical quarks)
  - ➢ K→ππ : 500 Tflops
- QCD thermodynamics: 100 Tflops
  - ➢ determine EoS
  - ➢ interpret experiments
- Hadronic physics
  - ➢ mπ~180 MeV, a~0.1F → 5% errors: 100 Tflops
  - ➢ quarks with phys. masses: 300 Tflops
  - ➢ ππ, Kπ scatt. Length: 100 Tflops
  - ➢ deuteron binding and other properties: 1 Pflops
- New Physics
- Numerical experiments

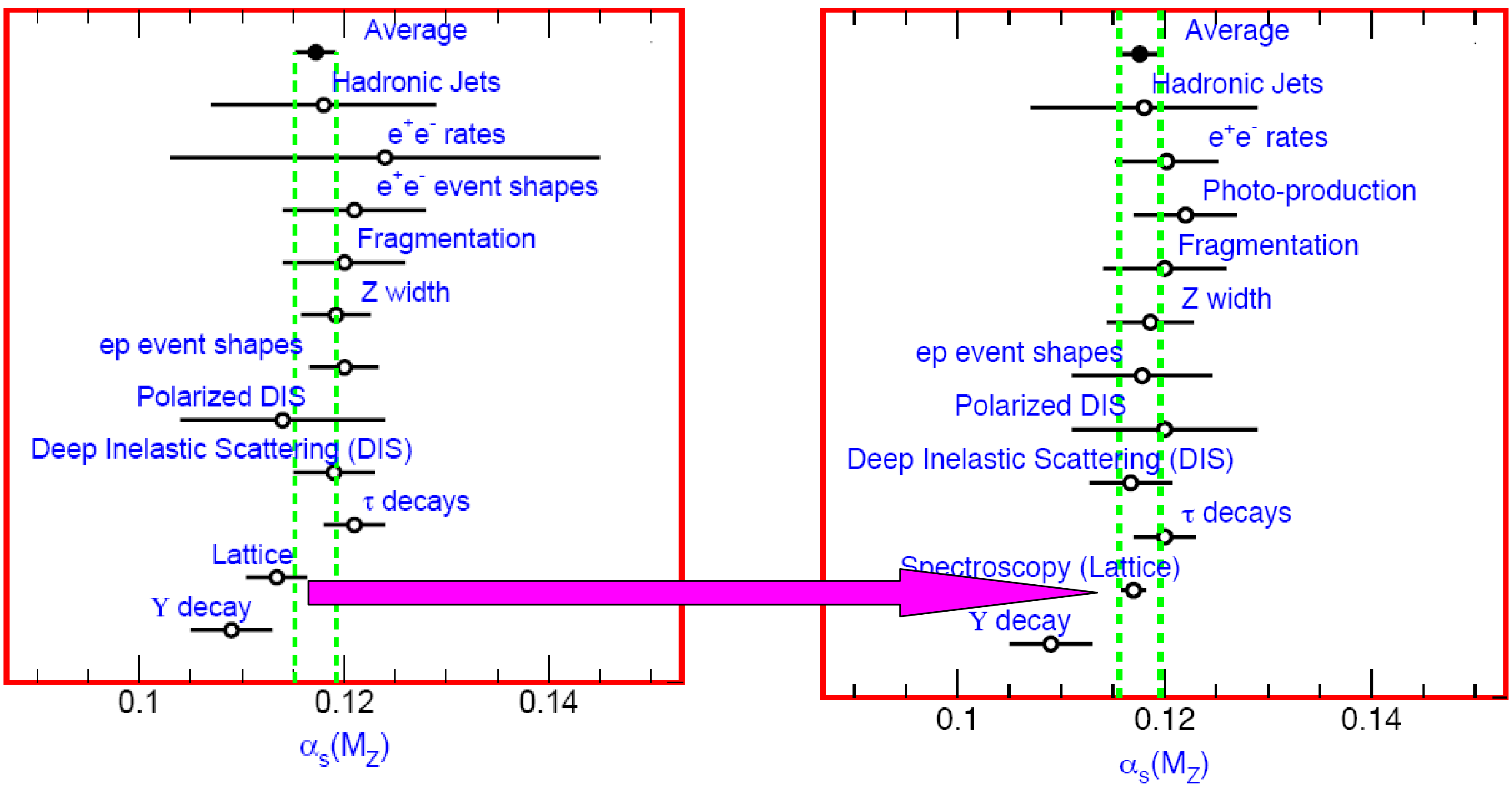
Sum > 1 Pflops
Several physics subjects
Define priorities



m_N = 963 ±12±8 MeV

arXiv:0803.3190, ETMC Coll.

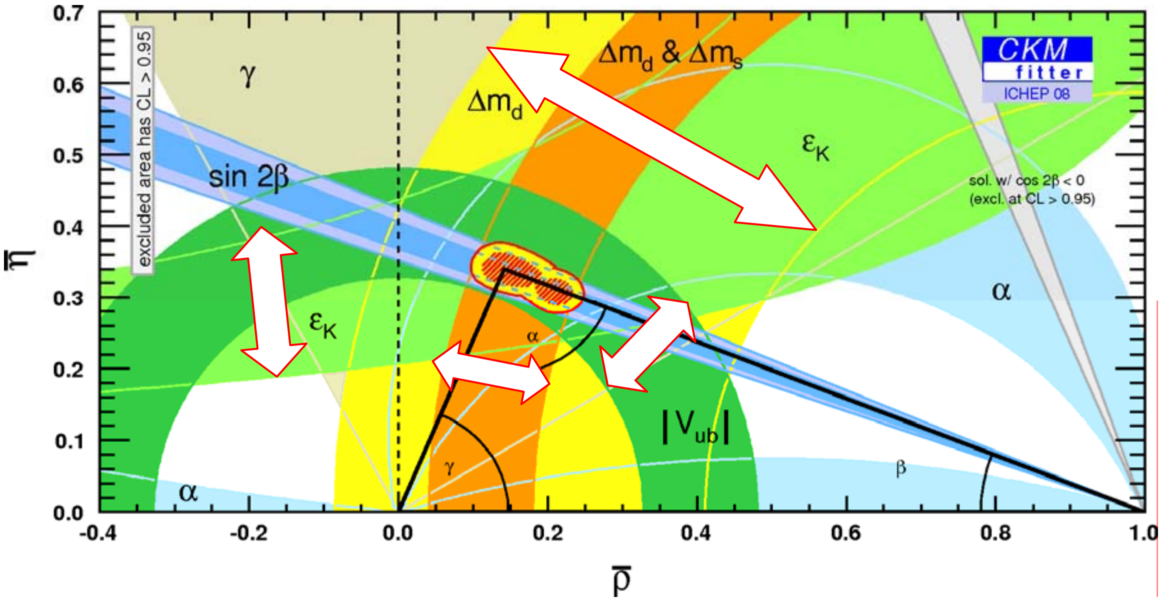
The strong coupling constant



PDG 2001

PDG 2005

Why Petaflops? (flavour physics)



Is there evidence for non-standard CP violation?

Increasing importance of LQCD

USQCD 2007

[illegible]	Quenched Estimate in 2000	Lattice Result Current	UTA Result Current	Lattice Errors 10. TF-Yr	Lattice Errors 50. TF-Yr
$\widehat{B}_K$	0.87 ± 0.15	0.77 ± 0.08	0.75 ± 0.09	± 0.05	± 0.03
$f_{B_s}\sqrt{\widehat{B}_{B_s}}$	262 ± 40 MeV	282 ± 21 MeV	261 ± 6 MeV	± 16 MeV	± 9 MeV
ξ	1.14 ± 0.07	1.23 ± 0.06	1.24 ± 0.08 **1.27±0.05**	± 0.04	± 0.02

<6MeV (2.5%)

O. Pène and P. Roudeau, PetaQCD (Orsay)

Why Petaflops? (specific example)

$$\Delta m_s = \frac{G_F^2}{6\pi^2}\eta_B m_{B_s} f_{B_s}^2 B_{B_s} m_W^2 S\left(\frac{m_t^2}{m_W^2}\right)\left|V_{ts} V_{tb}^*\right|^2$$

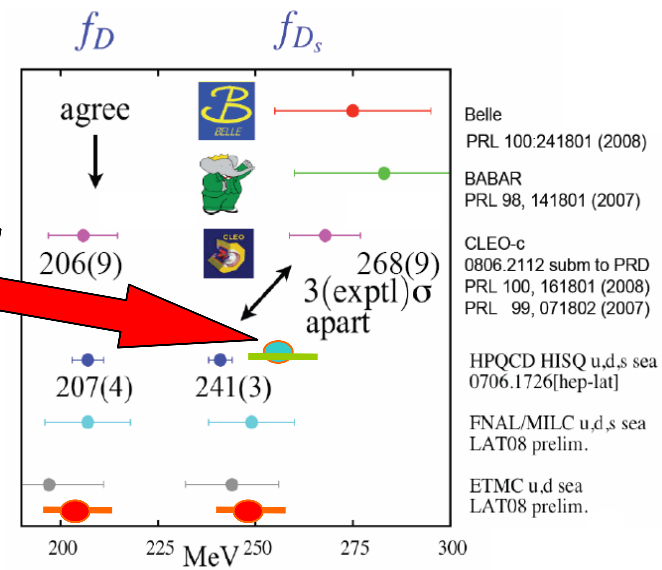
0.7% Exp. — 2% ? — 26% — 2.5% — 3.6%

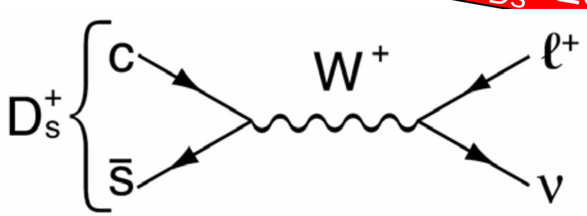
Non-lattice errors<5%

$Vcb=(41.49\pm0.48\pm0.58)10^{-3}$

m(top)= (170.9 ±1.8) GeV

New Cleo-c result
$f_{Ds}=259.5\pm7.3$ MeV





f_D = 205±7±7 MeV
f_{Ds}=248±3±8 MeV

arXiv:0810.3145
ETMC coll.

To be confident that 2 results agree or differ requires effects >3 sigma … at least

LQCD in other countries

Country	Sustained Teraflop/s
Germany	10–15
Italy	5
Japan	14–18
United Kingdom	4–5
Unites States	
LQCD Project	9
National Centers	2
US Total	11

Feb. 2007

France ~0.6 (apeNEXT)
+BlueGene/P (2008): 3 (x2)
+CCIN2P3 (0.1)+CEA(0.02)

- Lattice founding organized and allocated on a national basis
- Available lattice computing will continue to expand in 2010 and beyond

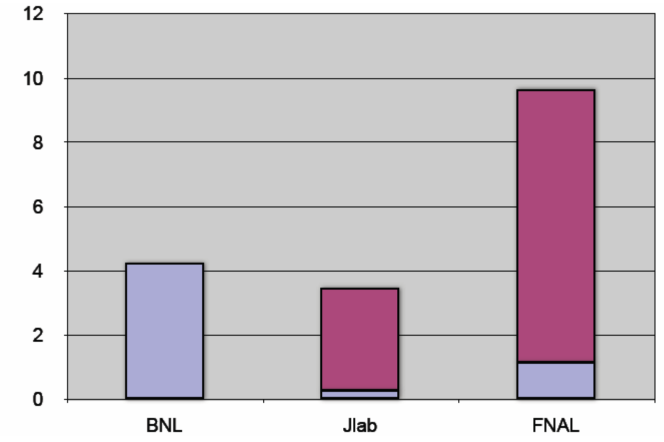
Not exact QCD yet

USQCD plans

For illustration: LQCD DoE project (2004 → 2009)

+ supercomputers

	Delivered teraflops for lattice QCD	
	Oak Ridge XT4	Argonne Blue Gene/P
2007	1.8	1.7
2008	3.75	3.75-7.5
2009	15	3.75-7.5



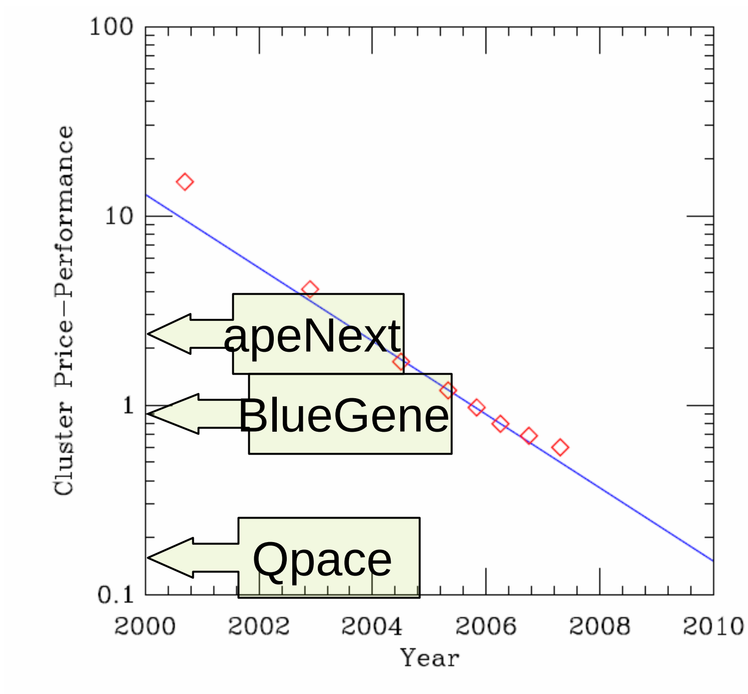
17.5 Tflps sustained
In 2009

Fiscal Year	Dedicated Hardware (Teraflop/s)	Leadership Class Machines (Teraflop/s)
2010	34	33
2011	61	52
2012	100	82
2013	161	131
2014	256	208

Plans

HEP +NP investment:
3.0 M$/year

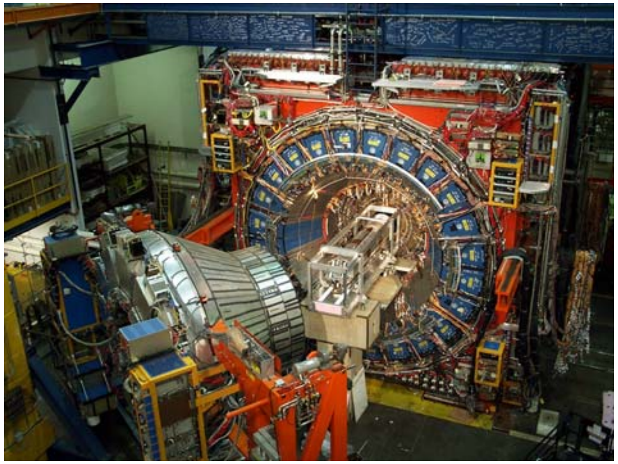
Price?



Price in M$/Tflop (sustained) (clusters) 1Euro = 1.56$

- Some prices:
  - apeNEXT: 0.75 MEuros/Tflop(peak)
  - BlueGene (CNRS): 0.12
  - « QPACE »: 0.02 ?
  - Can expect 1Pflops for ~10MEuros (2012) + operation

1 BaBar publication ~1M$



CDF+D0 RunII upgrade~30M$

Close to real QCD

- Large lattice size (5F), small spacing (a~0.04F)

  consider: 128^3 x 256

  Produce ~5000 trajectories / parameters setting/month

- This implies 1 Petaflops sustained

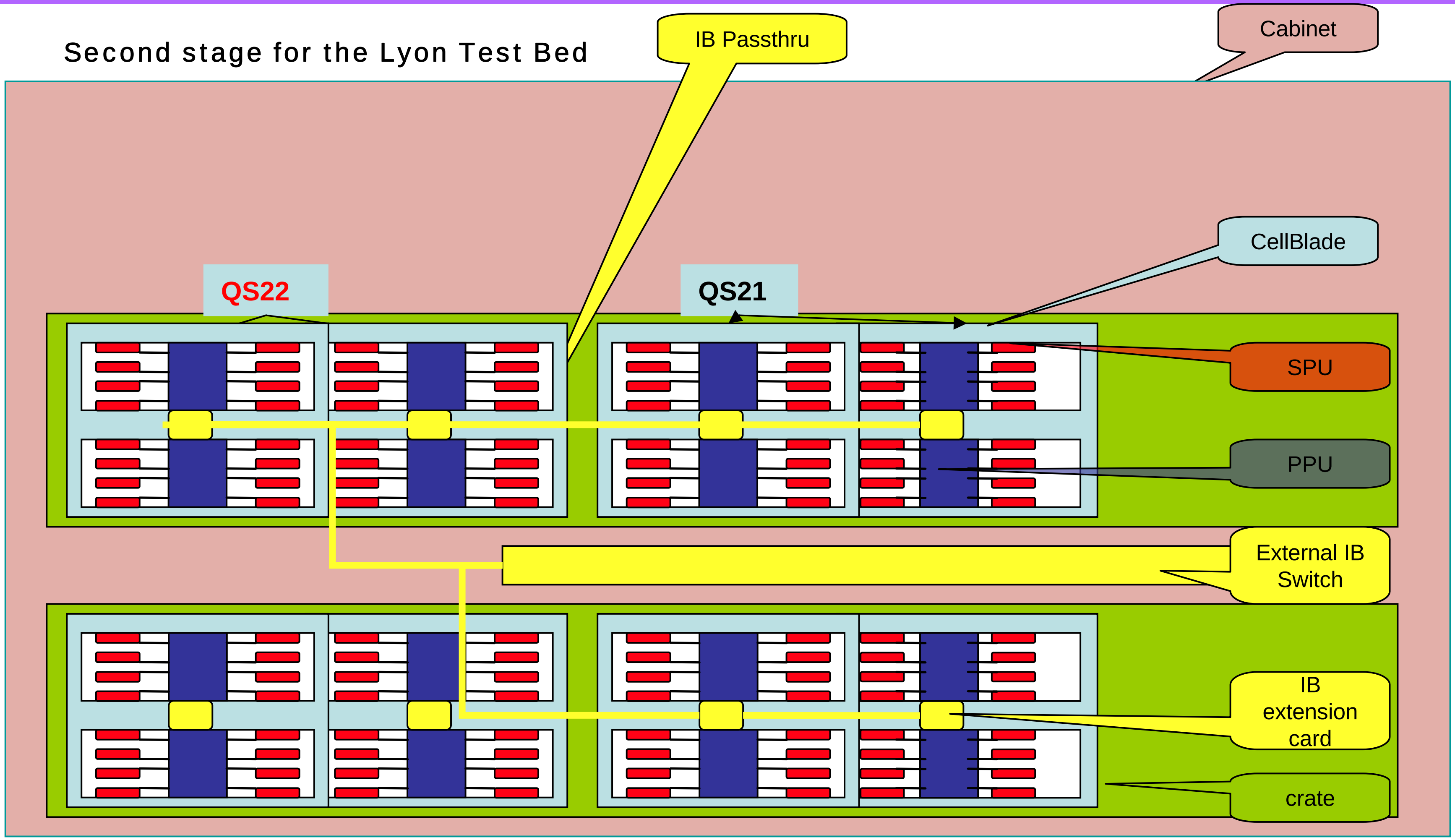
- ~few thousands computing units: 1Tflop peak/unit

Hardware activities

- Ongoing: Coyote testbed at CCIN2P3
  GPU in Saclay and Rennes
- Possibility: benefit from LAL electronics department expertise

Coyote testbed for IBM Cell

Second stage for the Lyon Test Bed



The general architectural Scheme

- What should be kept from present QCD machines, including BlueGene, QPACE: A toric network of compute-nodes (not more than a few tousands)
- The nodes will have several computing units. It could be heterogeneous (CPU+GPU’s, IBM-CELL like), or homogeneous multicore (INTEL/Larrabee ?)
- The network should be « APE-like », but including technological progress (need of a technological watch).

Software activities

Some ideas we have in mind (see talks by Christine Eisebnbeis Denis Barthou,)

- An abstract language (example Fortress) to represent the main algorithms we use. This allows simpler and architecture independent manipulations.
- Automatic code generation tools, combining the abstract algorithm description and the basic architecture description;
- A « parameter space » of the possible codes in which we choose automatically and manually the « best » for a given architecture (compile time improvement)
- Use of standard or handmade profiling tools.
- Watch for better adapted algorithms, think about algorithmic improvements.

Error control / checkpointing

Larger systems have larger failure probabilities. This asks for a more systematic and automatic system of alert and of new start. At present our checkpointing is simple minded: saving the gauge configuration and the random numbers periodically and then and start from the last one in case of failure. One could think of some « daemons » launching automatically an alert for some symptoms. Find out the best periodicity of checkpoints. Is it possible to consider local restart (replacing on flight a node) ?

See the presentation by Pascal Gallard and/or Mathieu Ferré

Overall consistency of the project

Is this project well equilibrated ? What is missing ?

- **One obvious point has not been treated: hardware work on the network. Presumably impossible without a european collaboration.**
- **What is to be expected from the next steps of QPACE ?**
- **What are our italians colleagues projects ?**
- **What about the relationship with TOTAL, with IBM, ?**

BACKUP