Le projet ANR petaQCD

- Why petaflops?
- Prospects in other countries
- Previous related ANR (QCDnext, PARA)
- The ANR petaQCD
- Getting money?

Why Petaflops?

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→$\pi\pi$: 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
  - $\pi\pi$, Kπ scatt. Length: 100 Tflops
  - Deuteron binding and other properties: 1 Pflops
- New Physics
- Numerical experiments

Σ > 1 Pflops
Several physics subjects
Define priorities

Why Petaflops?

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

0.7% Exp. | 2% ? | 26% | 2.5% | 2%

Non-lattice errors<4%

Vcb=(41.56±0.39±0.08)10^{-3}

m(top)= (170.9 ±1.8) GeV

$(f_B)^2$

f_B

USQCD 2007

ement	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	± 0.04	± 0.02

Need >100 Tflops

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

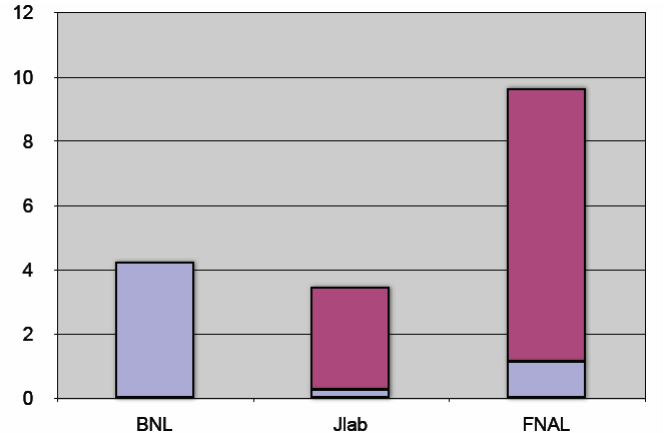
- Lattice founding organized and allocated on a national basis
- Available lattice computing will continue to expand in 2010 and beyond

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

Previous ANR: QCDnext

- 0.6 MEuros, 30/11/05-30/11/08
- 4 components (LPT, Inria, LAL, Dapnia/SphN)
- Purchase of 2 apeNEXT computers, installed in Rome
- 2 Postdocs (1 year)
- (20-50 kE for material), software/hardware activities

➢ examine critical parts in MILC and HMC codes

➢ IBMCell simulator, real time measurements on IBMCells

➢ ……

Previous ANR: PARA

- Work on HMC generator (common meetings with QCDnext)
- Evaluate gains
- Consider different platforms: Itanium cluster, BlueGene, IBMCell, GP-GPU
- Significant gain obtained on “classical” machines
- Report in preparation

The ANR project petaQCD

- A new ANR project, **PetaQCD**, has been submitted March 26, within the framework COSINUS-2008
  - Conception et Simulation (**Axe thématique : PetaScaling**)
- It gathers the following members:
  - LAL (G.Grosdidier coordinator)
  - **LPT (Orsay)**
  - **LPSC (Grenoble)**
  - **CEA/IRFU**
  - **INRIA Saclay**
  - **IRISA (Rennes)**
  - **PRISM (UVSQ)**
  - **CAPS-Entreprise (Rennes)**
  - **Kerlabs (Rennes)**

1.4ME
>50% salaries
0.4 ME for PME
0.06ME material

- Aim : conceive a petaflop machine optimized for LQCD with a maximum of 4000 processors
  - Optimize performance/price/consumption
  - Using standard (off-the-shelf) material

The ANR project petaQCD

- Expertise
  - Hardware and Software related to parallelism at several levels
  - Know-how in installing and operating large installations
  - Expertise on Algorithm for LQCD
  - Explore fault tolerances, strategies for error recovery

- Machines
  - IBM Cell (CCIN2P3) ➲ experimental
  - GP-GPU (IRFU, INRIA Rennes) ➲ experimental (Tesla/NVIDIA)
  - Multi-cores (DSM/CCRT)
  - BlueGene/P (IDRIS)

The ANR project petaQCD (contacts with IBM)

- After ANR submission, contacts with IBM
  - (through test-bed installation at CCIN2P3)
  - Propose to have a convention if ANR accepted
    - Interested also to collaborate in ANR not accepted
  - Already help to purchase few QS22 bladecenters

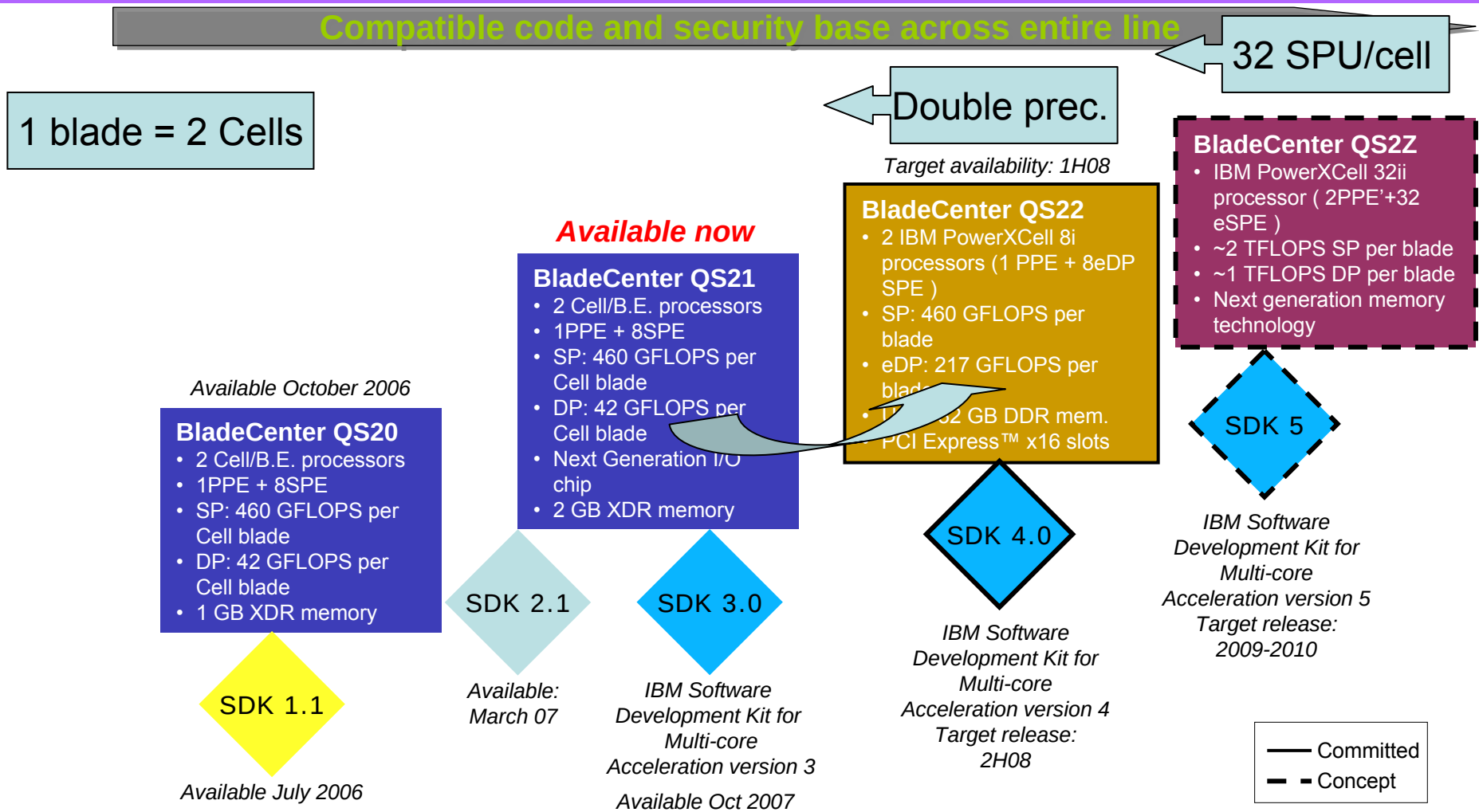
- Guided tour organized
  - At the Watson Research Center (NYC) 27 to 29 may
    - G. Grosdidier
  - Questions/answers started in view of (eventually) best matching our needs to (future) products

The ANR project petaQCD (connexion with QPACE?)

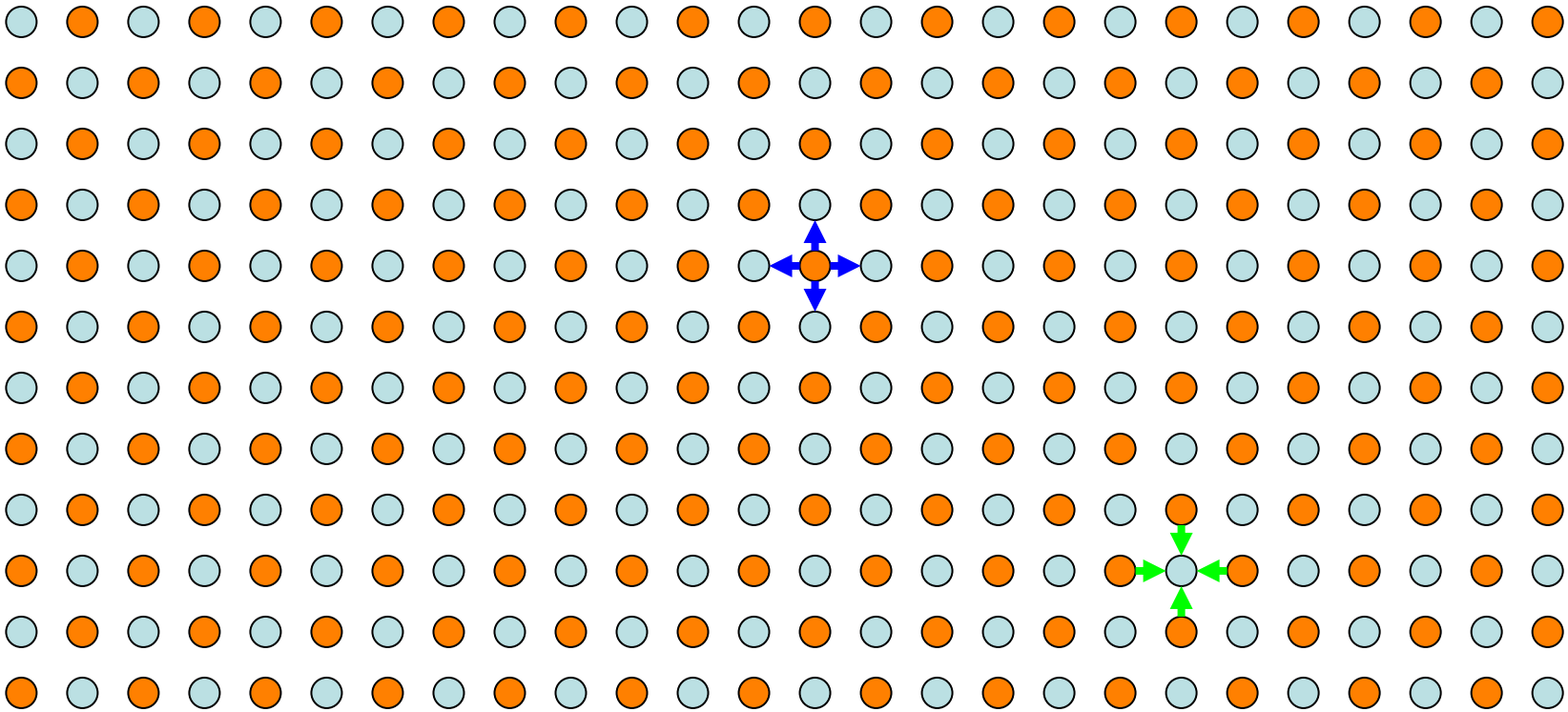
- QPACE: Qcd PArallel computing on CEll
  - Design a massively parallel QCD prototype
  - Key components: enhanced Cell processor, custom network processor
  - Financed by special grant (Univ. of Regensburg and Wuppertal) 3ME(?)
  - Strong contribution from IBM

- Timing
  - End 2008: small prototype
  - 2009: large prototype of 2 machines, 100 Tflops (peak) (~20 Tflops sustained)

The ANR project petaQCD (the IBMCell)

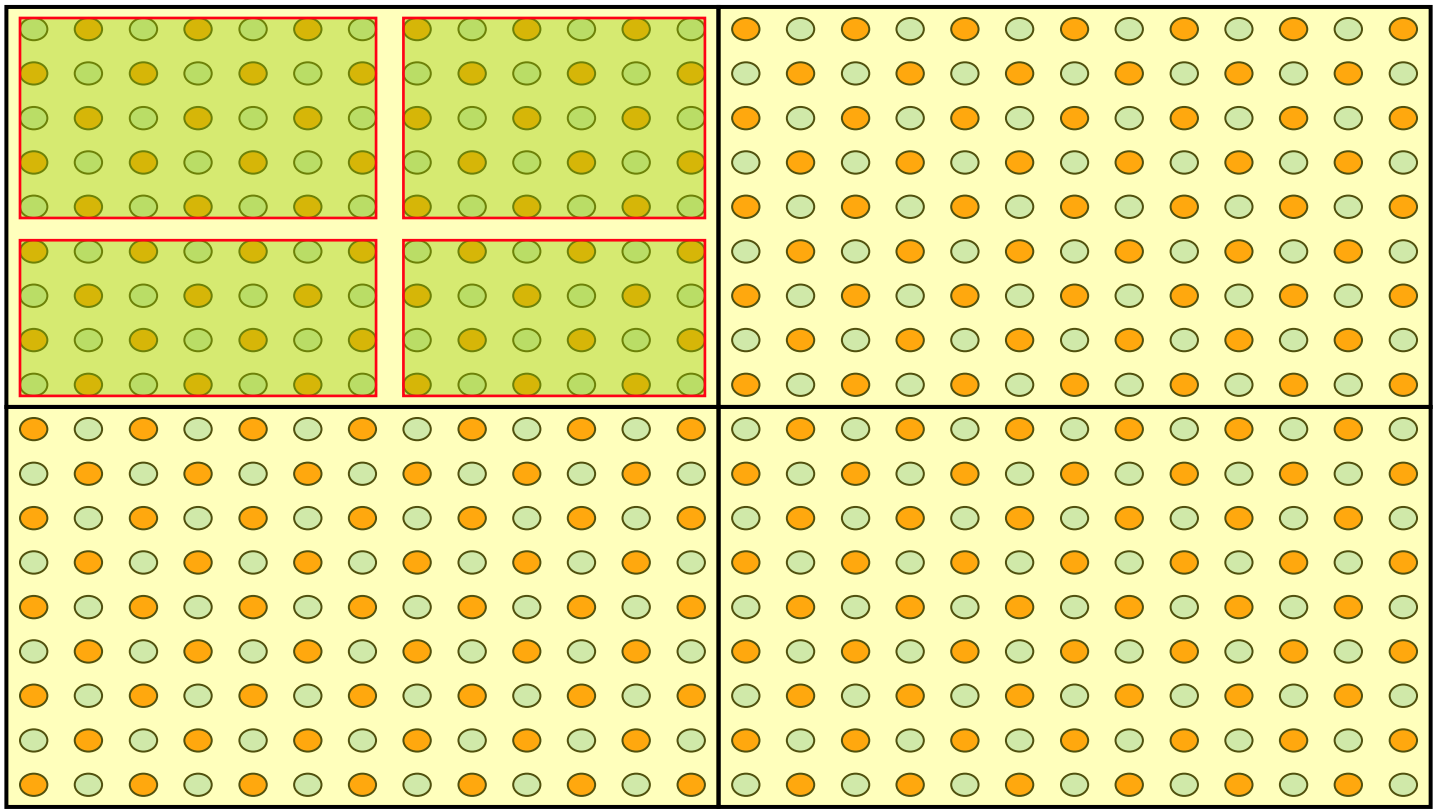


The ANR project petaQCD (the IBMCell)



Most of the computing time spent in inverting a large sparse matrix

The ANR project petaQCD (the IBMCell)



Put sub-arrays on computing units, find a compromise between time spent on a sub-array and the time spent to transfer data.

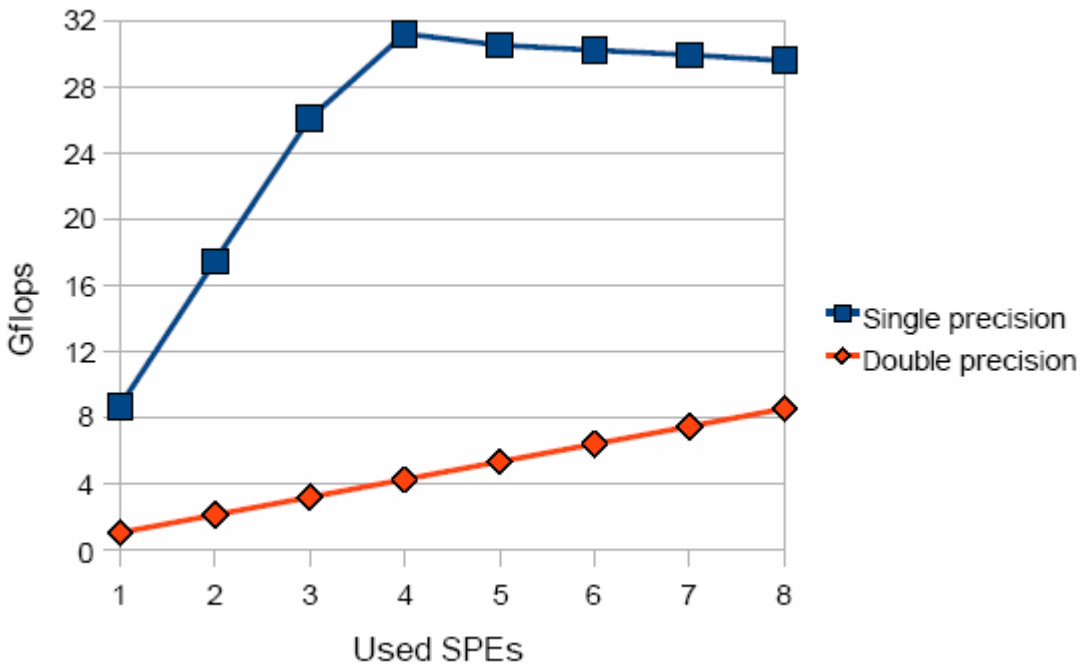
The ANR project petaQCD (the IBMCell)

- Tests have been driven on a QS20 CellBlade in Barcelona (thanks)
  - With a HMC kernel routine known to burn almost 90% of the overall HMC CPU time with large lattice sizes (Hopping_Matrix)
  - The test was running on SPU only (no data up/downloaded to/from PPU)
    - H_M routine was split into 2 almost equal parts *(Kseries & Lseries)*

- Both parts lead to about the same performance
  - Out of a loop of 4-5k cycles (3.2 GHz clock)
  - 416-464 Floating Point instructions are executed repeatedly
    - In a $8*10^7$ loop
    - Double Precision
  - And this required about 100-125 sec.

- Math leads to 0.9-1.0 Gflops/sec/SPU
  - which means a 20% CPU efficiency compared to QS20/21 DP performance expectations

The ANR project petaQCD (the IBMCell)



The computing power capability of the SPUs cannot be fully exploited:
- increase local SPU memory → spend more time on each SPU
- increase the data exchange rate between PPU anda SPU

The ANR project petaQCD (exercise with IBMCell)

- Lets assume that
  - A lattice made of $256*128^3$ sites
  - A cluster is built with 4096 Cell processors
- This means a partition of $32*16^3$ sites per Cell
  - Road map shows that in 2010, there will be 32 SPUs per Cell (32ii brand)
  - If one thinks that all site data must reside on SPU
    - This requires to host 16^3 sites on each SPU
    - And each site requires 3584B (2 Spi + 8 Mat + 8 HSp)
- However, currently, LS memory size is only 256kB
  - And a SPU can now hold only 16 sites (*double_buffer), K+Lseries = 115kB
    - the rest is for the program, plus other (more static) data
  - We have then to increase LS size by a 256 fold !
    - Meaning a size up to **32MB**/SPU

The ANR project petaQCD (exercise with IBMCell)

- Other possibility is to increase the speed for data transfer
- Some of these points (and others already discussed with IBM)
- Some numbers: (sustained)

  expect 2 Gflops/SPU (with current setup)

  ~4 Gflops/SPU if some “plausible” improvements

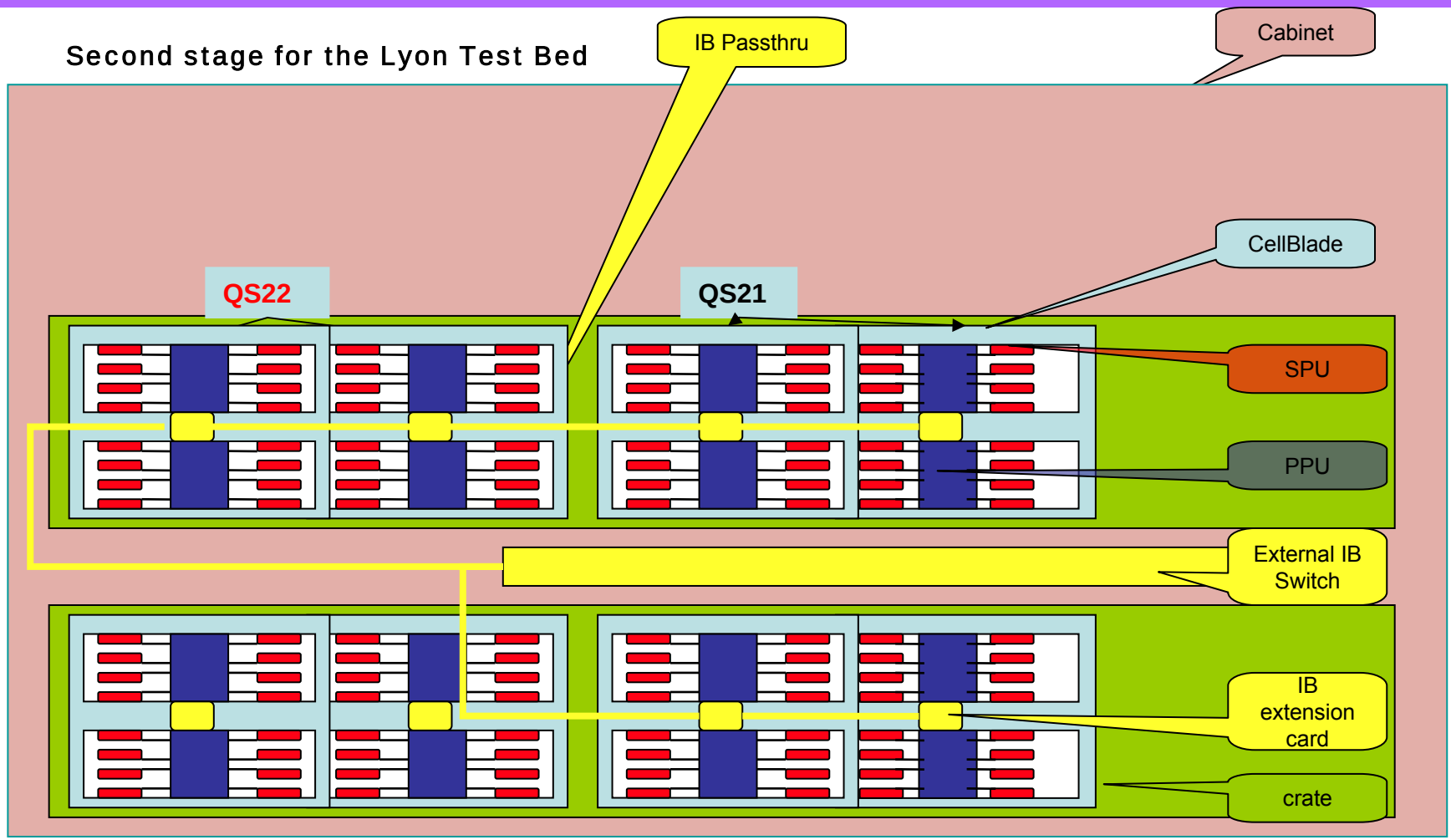
4Gflops x 32 x 4096 = 0.5 Pflops

What about communications between Cells?

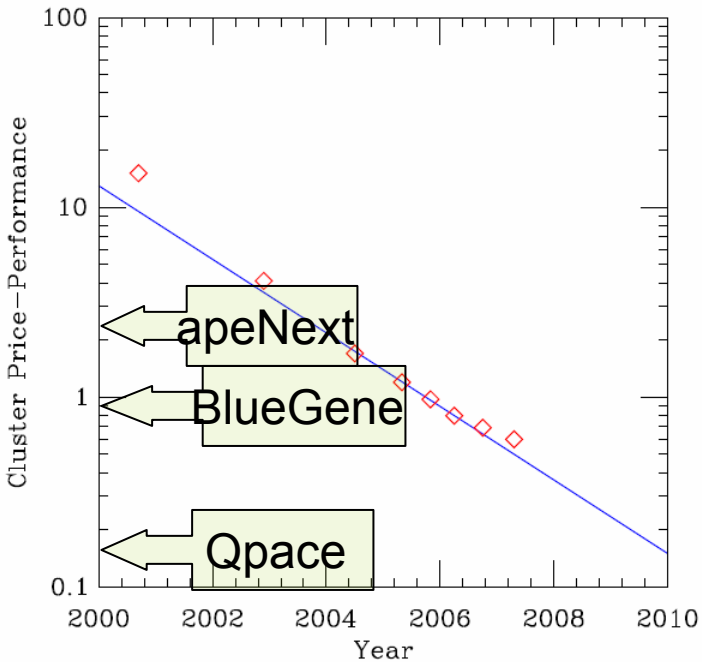
The ANR project petaQCD (testbed)

- First, a test bed is being built in CCIN2P3 @ Lyon
  - Currently: 4 QS21 CellBlades + a Cisco Infiniband switch
  - Next: 4 additional QS22 blades
- Evaluation will tackle mainly the data exchange performance
  - Network topology, bandwidth and latency issues (over Infiniband)
- Other issues
  - Number of processors, hence processing speed, of course
  - Power consumption, price & TCO, and reliability are also at stake
  - Checkpointing routine design (some facility with the blade)

The ANR project petaQCD (the IBMCell)



Getting Money?



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
- Negociate a constant level of financing by IN2P3 + IRFU (~1MEuros/year)
  - Machines for LQCD have to be considered as part of large expts related to this field
  - Define a strategy through the GDR to alert our authorities

BACKUP

The ANR project petaQCD

- **Several competences**
  - Phys. Théorique,
  - Phys. Expérimentale (nous entre autres !),
  - Informaticiens de métier
- Mais aussi
  - 7 labos publics (CNRS, CEA, INRIA)
  - 2 start-ups rennaises (rejetons de l'IRISA)
  - un centre de calcul "associé", le CCIN2P3 à Lyon
  - une entreprise "accompagnatrice", IBM-France ?
  - 2 centres de calcul de "référence", le CCRT et l'IDRIS
- Surtout : très ferme volonté de rester focalisés **strictement sur LQCD**
  - Important à souligner, car les informaticiens (français) ont habituellement mauvaise réputation à cet égard (trop théoriques)
  - Ceci est dans la droite ligne de 2 projets ANR finissants, donc pas de soucis
    - PARA, QCDNEXT
  - Notre cible : La famille de logiciels **HMC**
    - Pour Hybrid Monte Carlo (collab. ETMC)
- Le projet ANR lui-même vise la construction d'une **maquette** pour
  - démontrer la faisabilité
  - mesurer les performances

The ANR project petaQCD (the IBMCell)

Details of the BladeCenter® QS22

- **Core Electronics**
  - 2 PowerXCell 8i Processors w/ integrated DDR2 I/F (11S)
    - Clock rate: 3.2GHz
    - SP: 416 GFLOPS, DP: 217 GFLOPS per Blade
  - Up to 32GB DDR2 VLP DIMMs*
    - 667 & 800 MHz (Raw bandwidth: 667: 21.6GB/s, 800: 25.6GB/s)
  - 2 IBM Southbridge chips each supporting:
    - 2 PCI-E x16
    - 1 64b/100 MHz PCI-X
    - 1 DDR2 DIMM per IBM Southbridge**
    - 2 UART, SPI, JTAG, I²C
  - H8 Support processor (with IPMI)
  - Single wide blade form factor
- **Integrated features**
  - Socket for BC-H HS Daughter Card:
    - 2 ports IB x4 DDR and 10gE
  - Socket for: non-standard 2x PCI-E x16
  - Dual 1Gb Ethernet (BCM5704)
  - Serial/Console port, 4x USB on PCI
  - 8GB Flash Drive
- **Legacy IO connectors**
  - e.g. SAS connector card
- **Chassis shared features**
  - CD-ROM, Management module and Ethernet switch
  - Optional:
    - InfiniBand switch, SAS switch and 10gE switch
  - BC-H and BC-E support