Computing Systems for Theoretical Physics

Sebastiano Fabio Schifano

[1] University of Ferrara and INFN-Ferrara

Young Investigators Symposium 2008

October 13-15 2008, Oak Ridge National Laboratory

Background

Massively parallel systems optimized for scientific applications.

Project Activities:

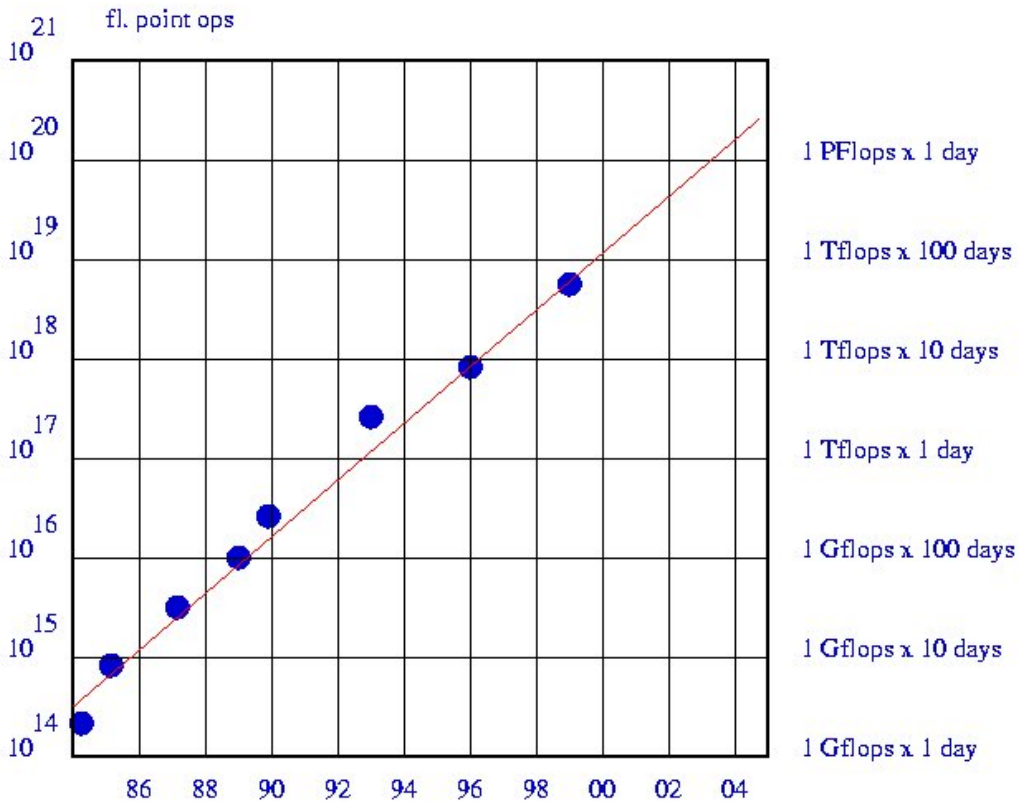
- past:
  Custom parallel machines for LQCD applications
- present:
  JANUS project a FPGA-based system for Spin-Glass simulations
- present/near future:
  QPACE project a petaflops scalable parallel system

Current Research interests:

- Reconfigurable systems (JANUS)
- Development and programming of commodity-based parallel systems optimized for scientific applications (QPACE, . . .)

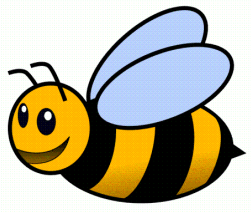
Quantum Chromodynamics on the Lattice (LQCD)

QCD describes the behaviour of hadrons (proton, neutron, ...)
$\Rightarrow$ important ingredient to understand high-energy experiments



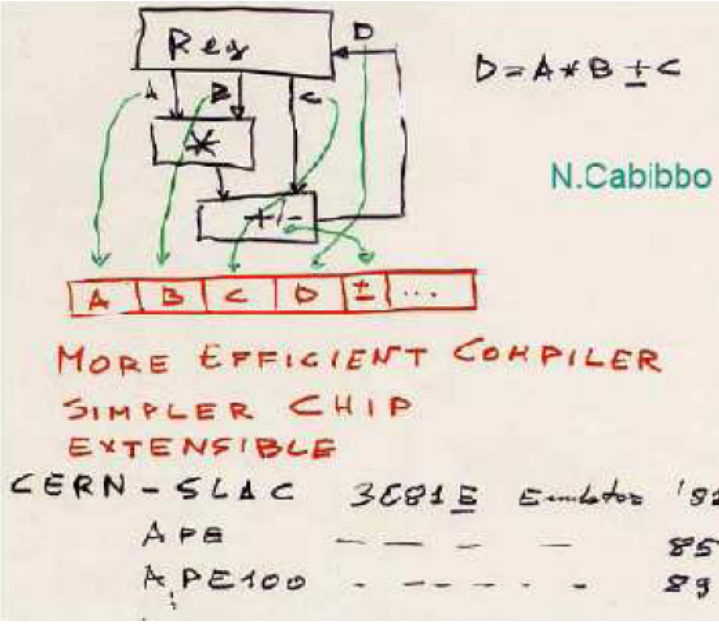
Petaflops performance is needed around 2010

Array Processor Experiment (APE)

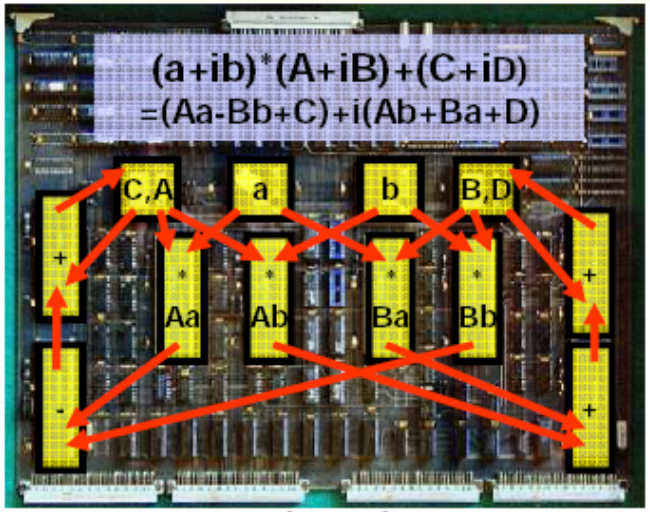


Started in the mid 80s at INFN to provide computing resources for LQCD simulations.

- low-power processor
- efficient complex FP arithmetics:
  $a \times b + c, \quad a, b, c \in \mathbb{C}$
- accurate balance between computation and IO access (memory, communication):
  $W(N)/P = I(N, m)/B$
- memory and network interface directly coupled to the processor
- nearest neighbour communications (3D torus)

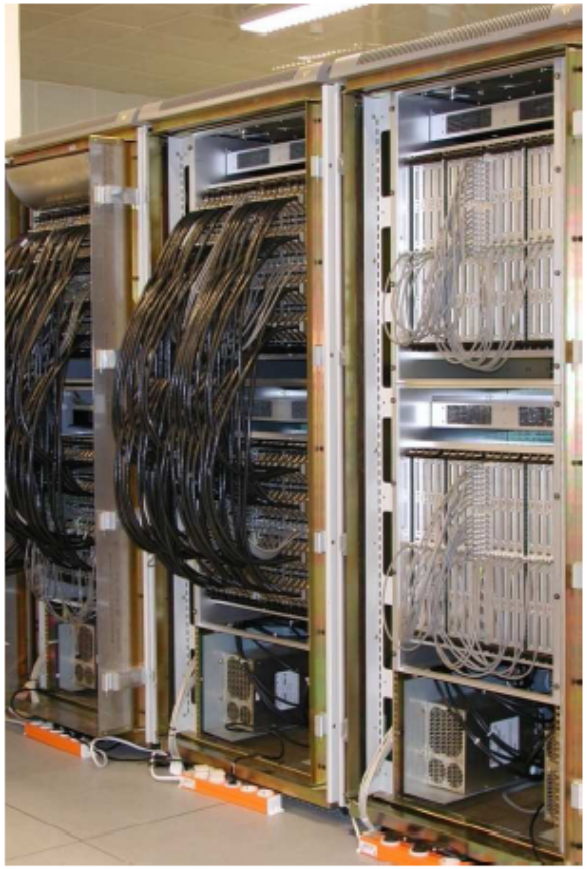


From APE1 to apeNEXT

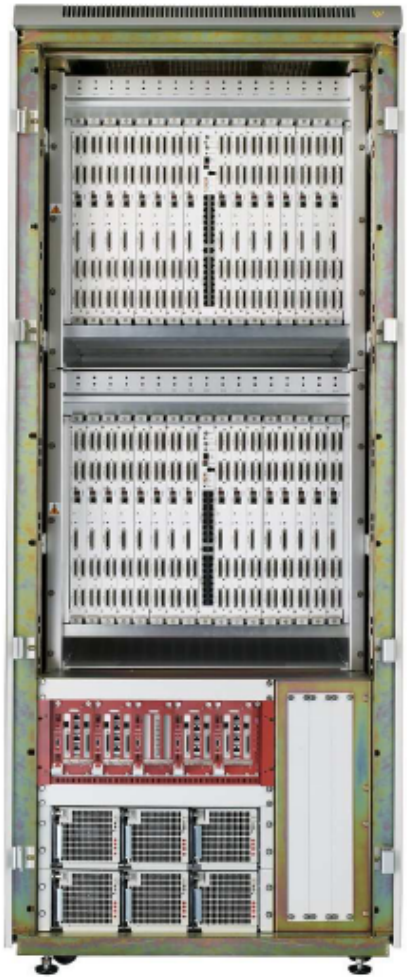


APE1 (1988) 1GF

APE100 (1992) 25GF, SP, REAL



APEmille (1999) 128GF, SP, Complex



apeNEXT (2004) 800GF, DP, Complex

History of LQCD Machines

	EU	US	Japan
1990 - 1995	APE100 300 Gflops	Columbia2	
1995 - 2000	APEmille 2 Tflops	QCDSP	**CP-PACS**
2000 - 2005	apeNEXT 15 Tflops	QCDOC	**PACS-CS**
2005 -	—	**BG/L,P,Q**	—

- Since 1996 APE has become an international European collaboration (France, Germany, and Italy)
- $>$ 25 of the 100 most relevant publications on LQCD since 1995 have been obtained APE machines (20 out of the 79 most cited articles, Spires)

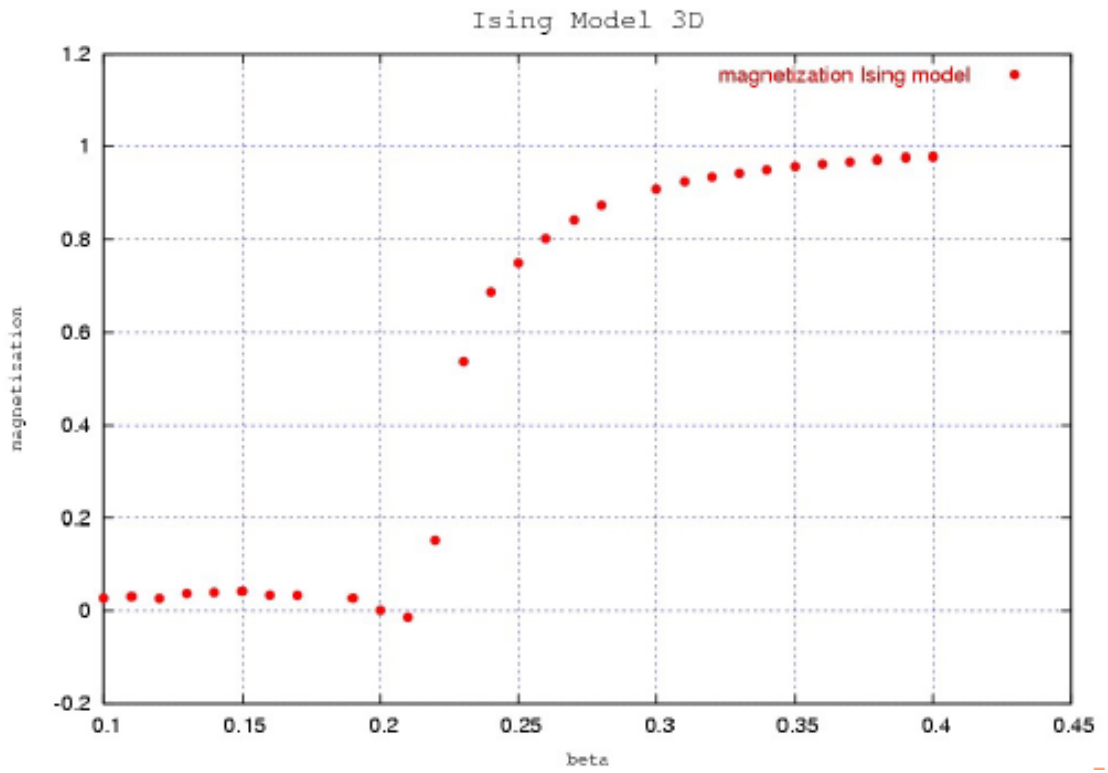
Today 15 Tflops apeNEXT are installed in Europe (DP, $\varepsilon \approx 40\%$)

Spin-Glass

The Spin-glass is a statistic model proposed to study some behaviours of the most complex macroscopic systems, like:

- disordered magnetic materials
- real flow glasses (e.g. Notre-Dame windows)

Example: transition temperature of magnets beyond which they lose their magnetic state.

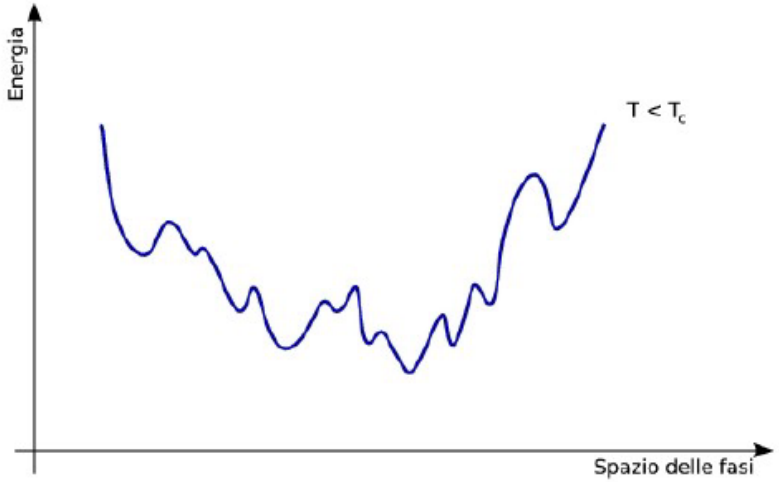


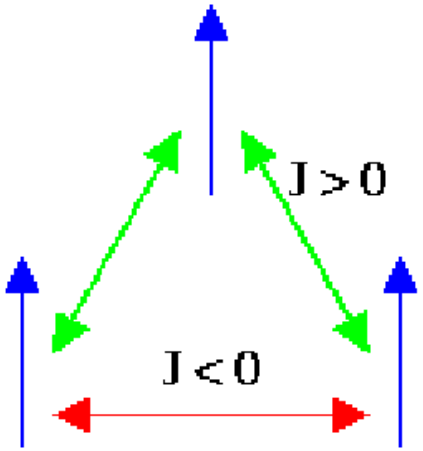
Spin-glass is a complex problem(1)

Averaging physics observables over all possible configuration S of the discrete 3D lattice of spins is extremely difficult due to **frustation** effects.

The energy function is defined as:

$$E(\{S\}) = -\Sigma_{\langle ij \rangle} J_{ij}\sigma_i\sigma_j, \quad \sigma_i, \sigma_j \in \{+1, -1\}, \quad J_{ij} \in \{+1, -1\}$$

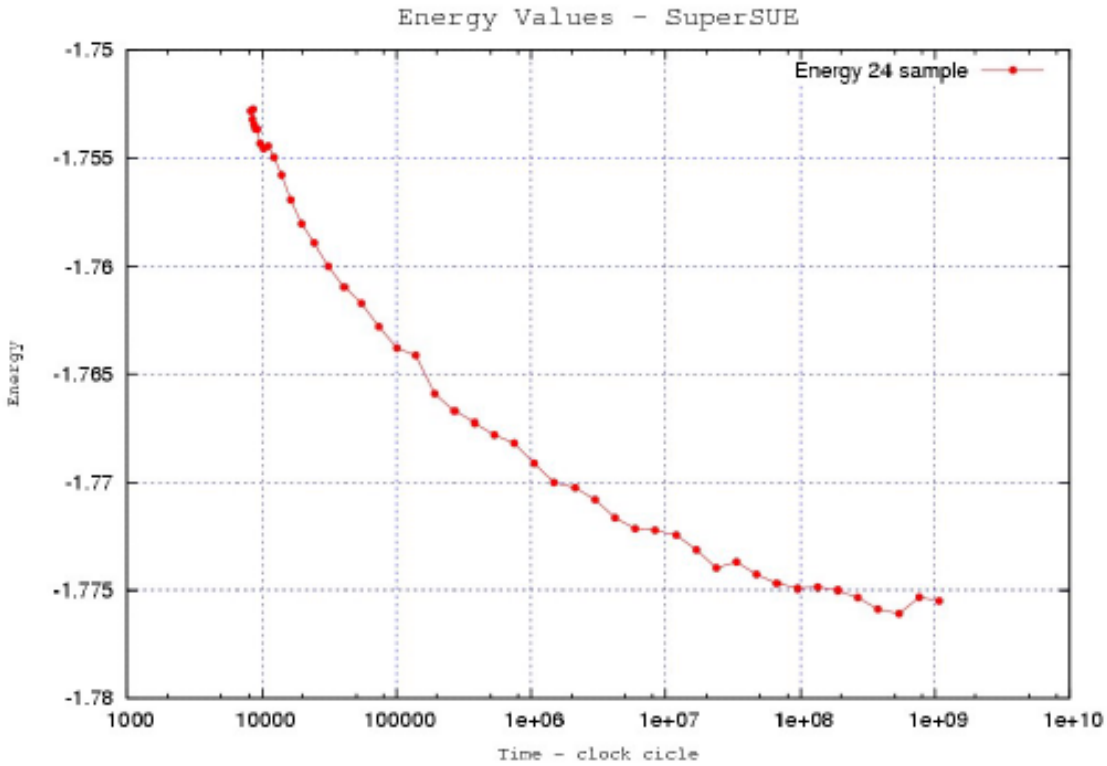




Spin-glass is a complex problem(2)

Bring a system of $48^3, 64^3, 80^3$ lattice points to thermal equilibrium:

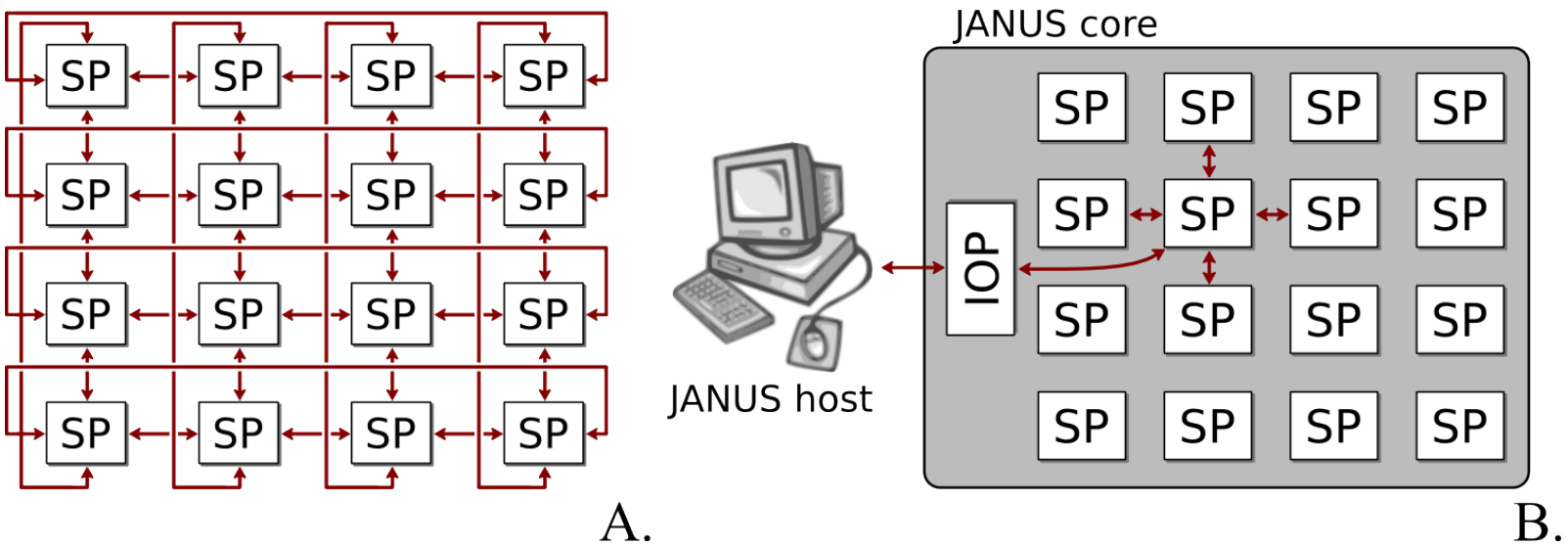
- the system must be followed over $10^{12} - 10^{13}$ MonteCarlo steps
- $\mathcal{O}(100)$ indipendent systems (replicas) have to be simulated



Janus: A Spin Glass Engine

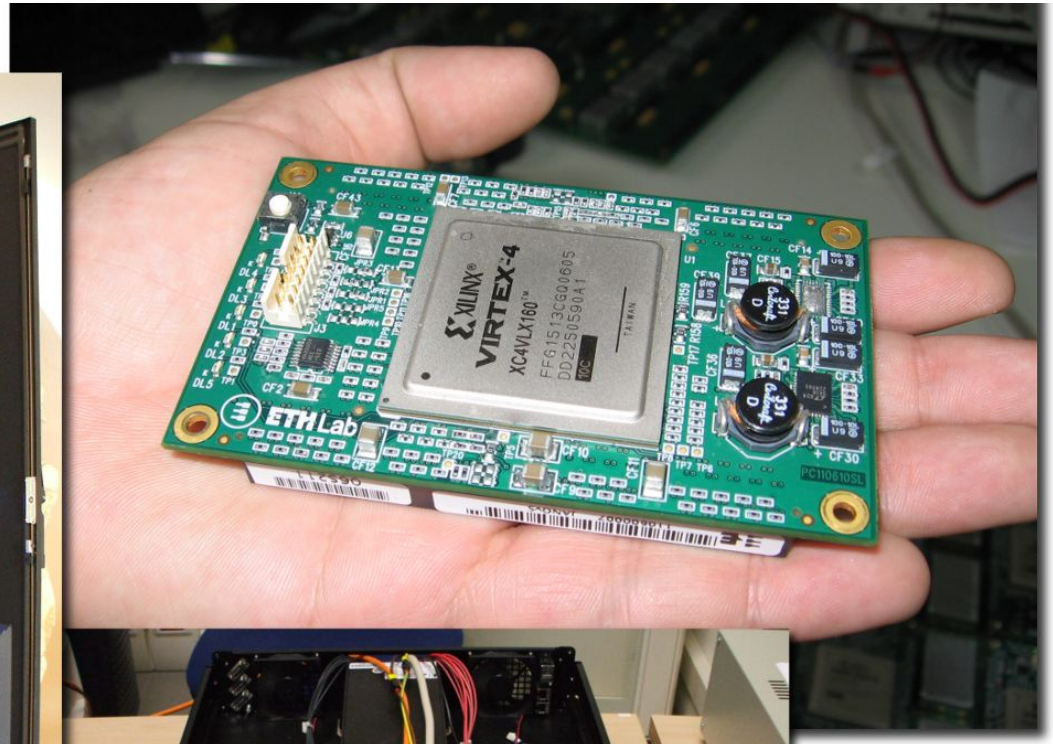
Architecture:

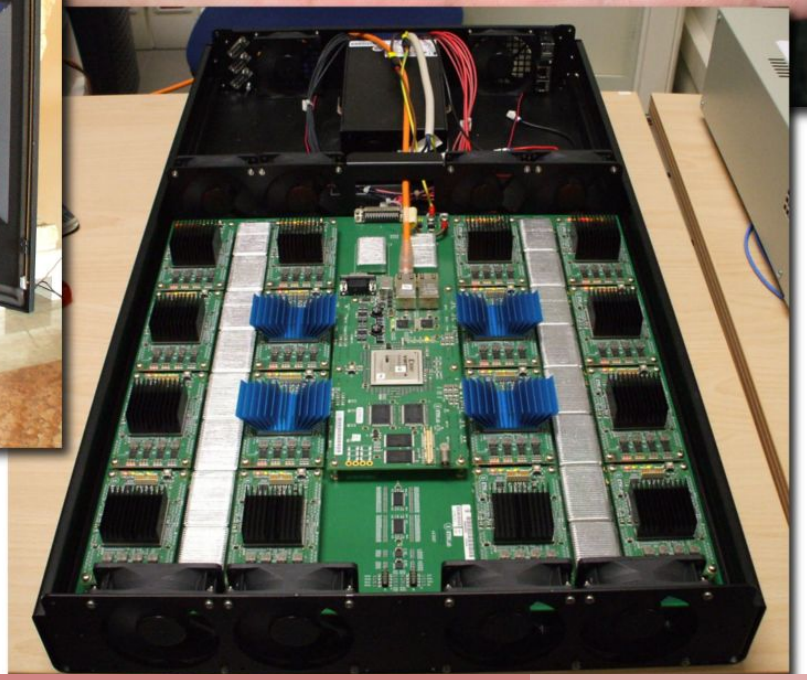
- a 2-D grid of 4 x 4 FPGA-based processors (SP's)
- data links among nearest neighbours on the grid
- one control processor on each board (IOP) with 2 gbit-ethernet



JANUS is a project carried out by BIFI, University of Madrid, Estremadura, Rome and Ferrara, and by Eurotech.

The Janus System





Janus vs PCs Performance(1)

- traditional architecture boost performance by processing several **replicas** of the system, mapped on the bits of their **long** data words (AMSC code)
- the same random number is shared by all replicas (introduce nasty but manageable correlations)
- however too-many replicas (> 256) become quickly useless and SMSC is required

System	spin update time (ps)	# Replicas
Janus core (16 FPGAs)	1	16
Intel Core Duo 2, 2.4 GHz	700 (AMSC)	128
Intel Core Duo 2, 2.4 GHz	3000 (SMSC)	64
IBM Cell 8-SPE $16 \leq L \leq 48$	38 - 30 (MIXED)	32
IBM Cell 8-SPE $L > 48$	98 - 100 (MIXED)	32

1. performance results on Cell are prelimary
2. Cell 8 SPE = 8 SPE on QS22

Janus vs PCs Performance(2)

Assuming to perform 10^{12} montecarlo steps on a lattice size of 64^3 we have:

	Janus	AMSC	SMSC
processor	1 SP	1 CPU	1 CPU
statistic	1 (16)	1 (128)	1 (4)
wall-clock time	50 days	770 years	25 years
energy	2,7 GJ	2,3 TJ	78,8 GJ
processor	256 SPs	2 CPUs	256 (64) CPUs
statistic	256	256	256
wall-clock time	50 days	770 years	25 years
energy	43 GJ	4,6 TJ	20 (5) TJ

Tipical number of replicas is 256, using more than 256 CPUs is physically not very interesting !

1 SP 40 W, 1 Janus-core 640 W, 1 PC 100 W

Goals for Future Systems (2010)

- high integration: 100 Tflops / m^3 (peak)
- power efficiency: 250 W / Tflops (peak)
- price / performance: $\approx$ 5 M€/ Pflops (peak)

Key elements:

- high FP-performance processor (400 Gflops)
- processor directly coupled to network by fast IO interface
- nearest-neighbours network, 3D torus

How to Reach the Goals ?

Current wisdom for the processor: custom design of processor ASIC no more **competitive**/needed:

- commodity processors have become efficient for QCD (SSE, cache-aware algorithms, ...)
- commodity processors allow scalable system architectures (power consumption, integrated memory and IO interface)

Possible alternatives are:

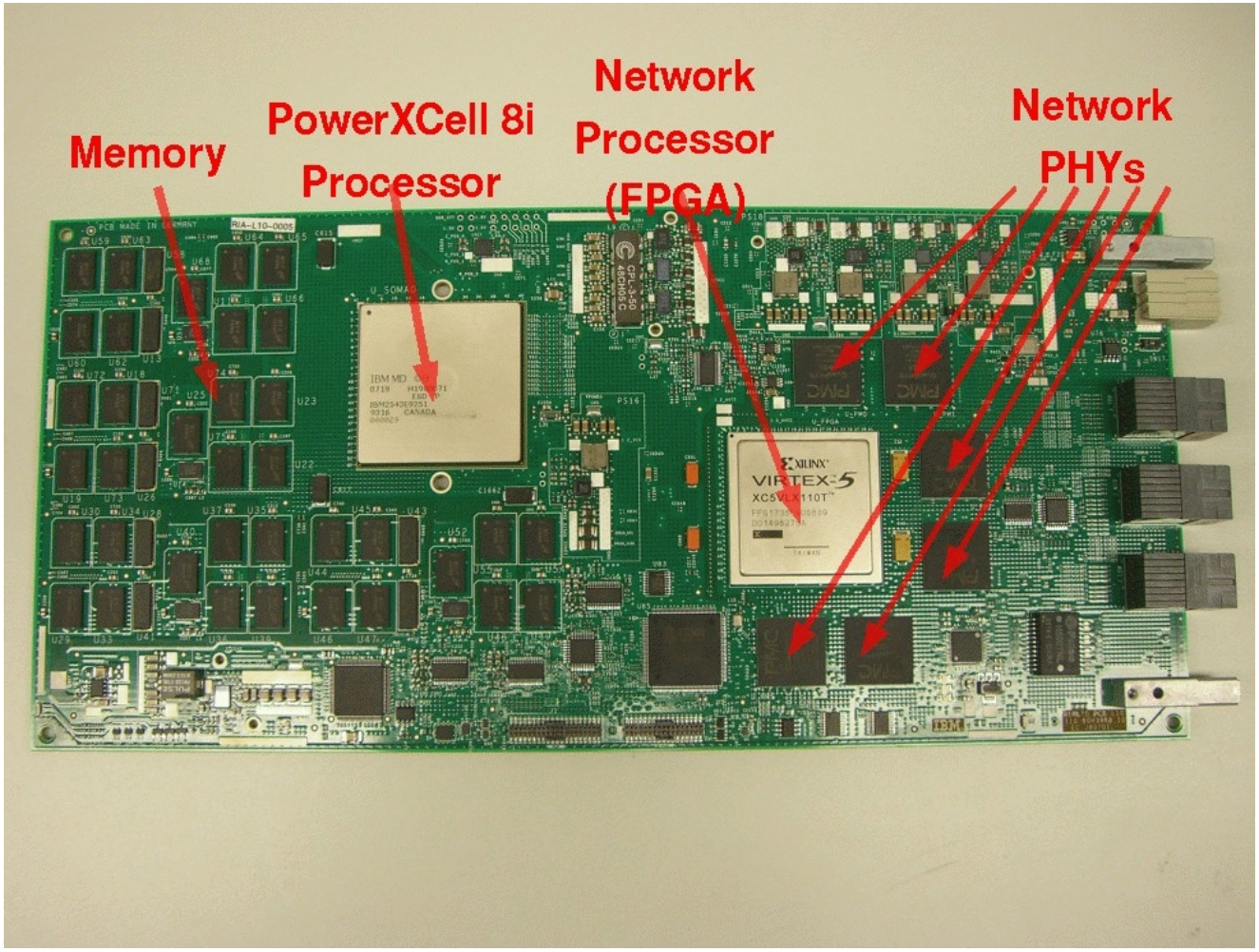
- use BlueGene[L,P,Q] systems . . . or . . .
- . . . do **better**(?) and interconnect commodity processors like
  - GPU
  - IBM Cell-BE processor
  - new generation of multi-core Intel processors

  through a custom network directly coupled to the processor.

QPACE: **Q**CD **PA**rallel computing on **CE**ll

- Design of a massively parallel QCD prototype (with suitability for other applications in mind)
- LQCD efficiency: **25**%
- Key components:
  - **Enhanced Cell BE processor** PowerXCell8i
  - **Custom network processor**: 3D-torus FPGA-based and 10-Gbe link technology, 1 GB/s per link, 1 μs latency
  - custom boards and system integration
- Timescale
  - End 2008/early 2009: small prototype running
  - Spring 2009: large prototype at $O(400)$ TFlops peak installed
- Secondary goals:
  - Gain experience with a multi-core CPU in a massively parallel environment
  - Ensure availability of strong-scaling architectures

QPACE node card



- University of Regensburg, Wuppertal, Ferrara, Milano and Padova
- DESY Zeuthen, Research Lab Jülich (FZJ)
- IBM Development Lab Böblingen

Summary and Conclusions

Spin Glass:

- commodity processors NOT provide the needed computing power
- **dedicated system** is the best approach
- results: 10.27 – 12.50 $ / Giga-ops, 8.75 Giga-ops / Watt

Lattice QCD:

- **past**: **dedicated systems** were the only viable choise
- **present**: the most used system is BlueGeneP
  - bunch of low-power simple processors
  - interconnected by a 3D-mesh torus
  - big effort on program porting
- **future**: LQCD ASIC design in not any more convenient/handy
  - use new generation of multi-core processor: low-power, large on-chip memory, high perfromance
  - interconnect them by a custom 3D torus network derived from experience of past Lattice-QCD machines (like APE)