Heavy Quarks on Fast Computers

Hubert Simma

NIC DESY

6.1.2010

Plan:

- ❑ Lattice QCD Challenges
- ❑ Machine Developments
- ❑ Theoretical Performance Analysis

Lattice QCD

Discretization on the Lattice provides framework for

- ✔ rigorous regularization of QFT
- ✔ non-perturbative results (renormalization, matrix elements)
- ✔ "ab initio" computation from QCD Lagrangian
- ✔ numerical evaluation of Path Integral by Monte Carlo method

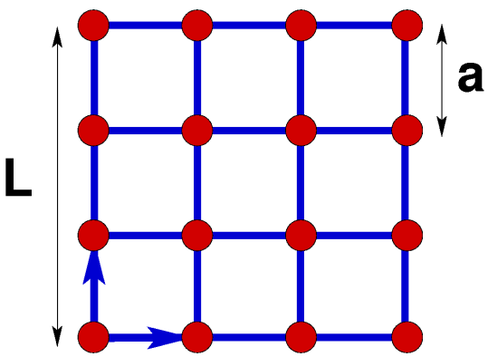
$$\frac{1}{N}\sum_{U} O(U) \;\longrightarrow\; \langle O\rangle = \int D[U]O(U)\cdot e^{-S(U)}$$

with N gauge-field configurations $U(t,x)$ distributed according to

$$P(U) \sim e^{-S_g(U)} \cdot detM(U)$$

Lattice QCD

Challenges (theoretical and technical)



✘ all observables in Euclidian space

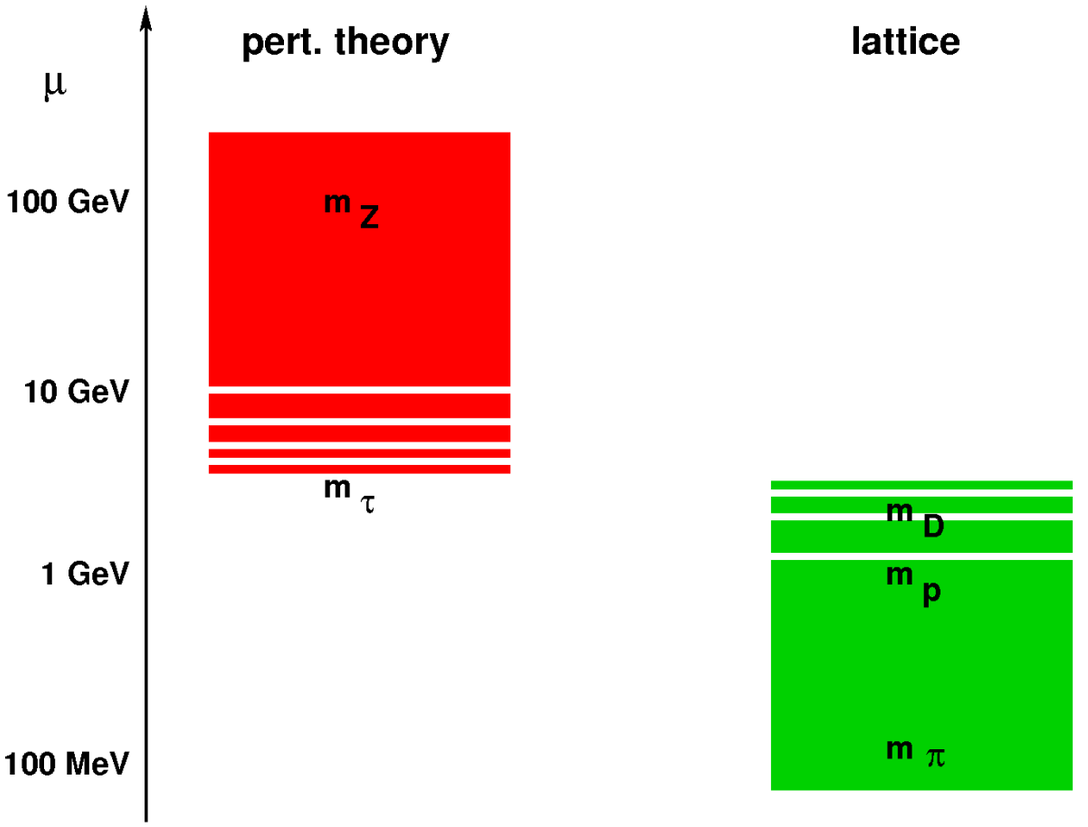
✘ explicitly broken (chiral and space-time) symmetries

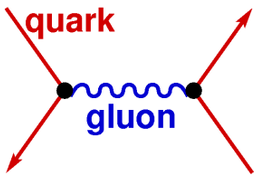
✘ extrapolation to continuum limit $a \longrightarrow 0$

✘ extrapolation to physical values of light-quark masses

✘ limited physical volume L (isolation of hadronic ground states)

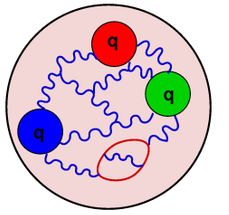
Large Scale Differences



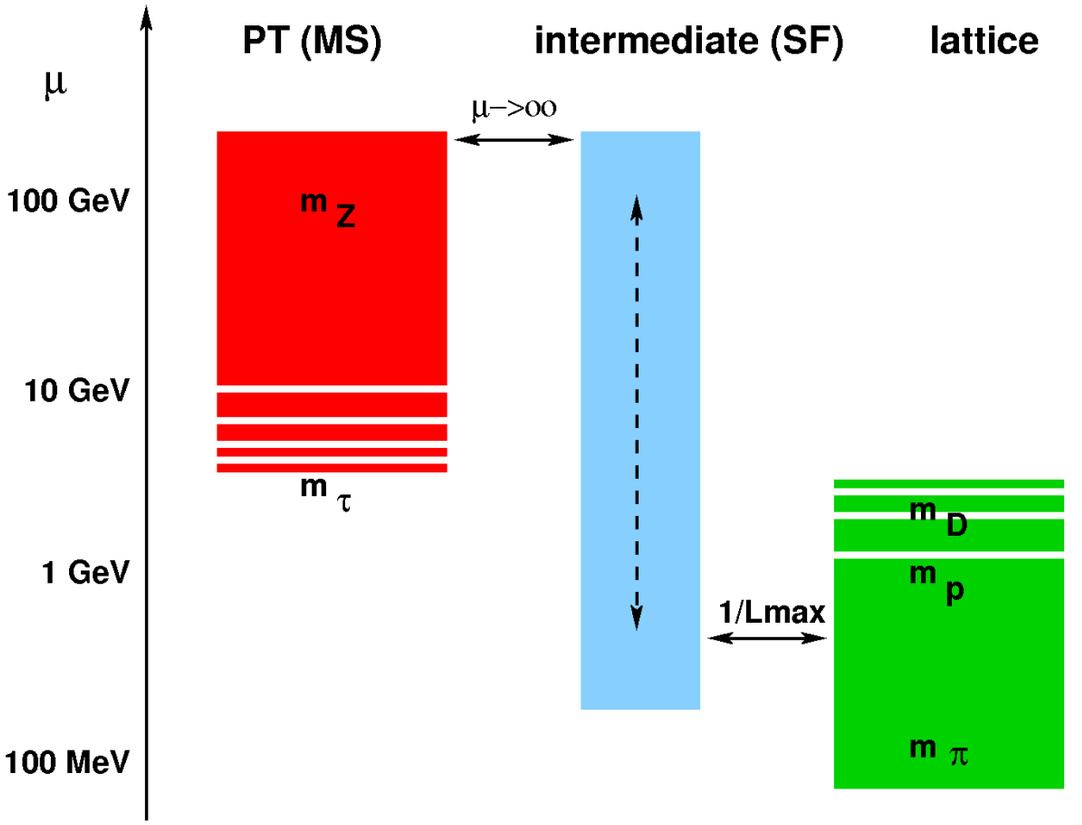


$$a^{-1} \gg \mu_{\rm PT} \gg m_{\rm hadr} \gg L^{-1}$$

$(0.05\ fm)^{-1}$ $(3\ fm)^{-1}$



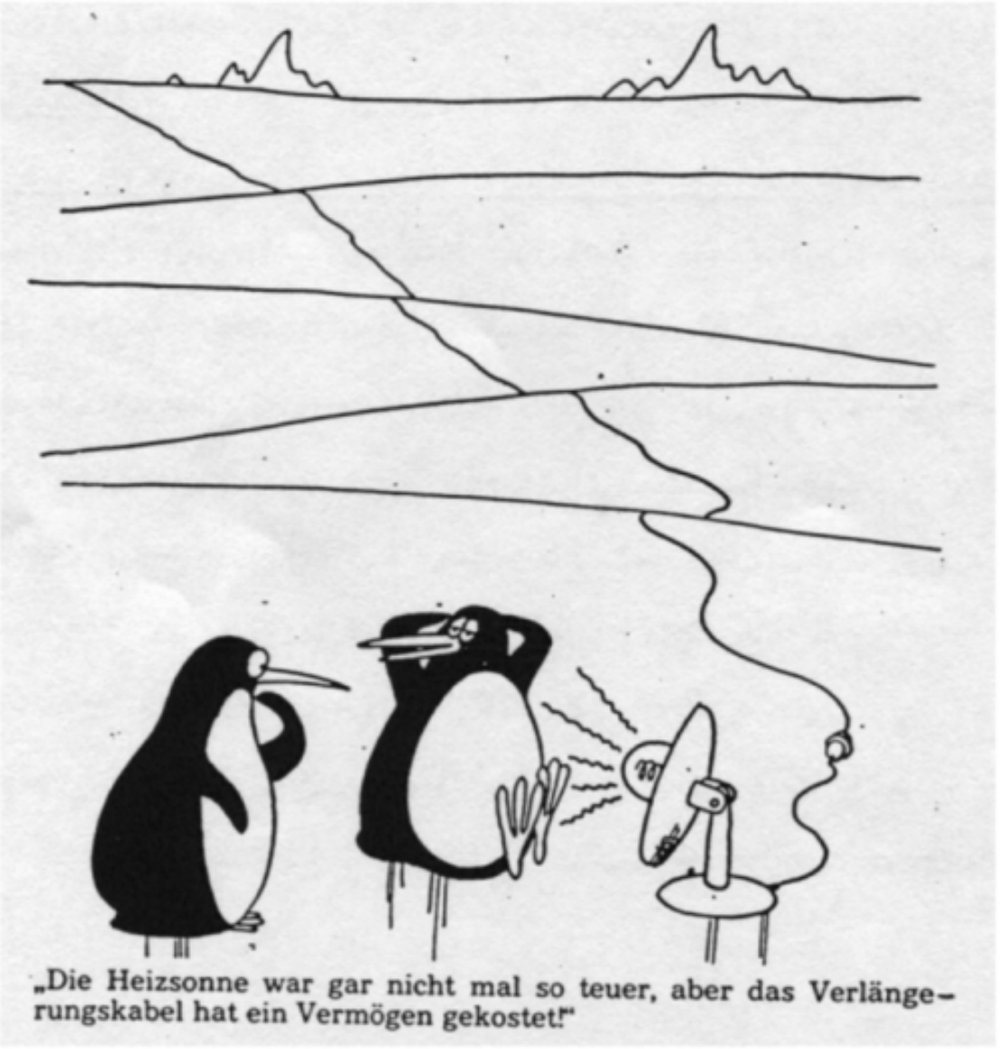
Renormalization Schemes



$$\underbrace{\begin{pmatrix} \Lambda_{\rm QCD} \\ \hat{M}_{\rm light} \\ M_{\rm s} \\ M_{\rm c} \end{pmatrix}}_{\text{RGI parameters}} \quad \mathcal{L}_{\rm QCD}(g_0, m_0) \quad \Longleftrightarrow \quad \underbrace{\begin{pmatrix} F_\pi \\ m_\pi \\ m_{\rm K} \\ m_{\rm D} \end{pmatrix}}_{\text{Hadronic observables}}$$

B-Physics?

$$a^{-1} \gg m_B \ \ldots \ m_\pi \gg L^{-1}$$

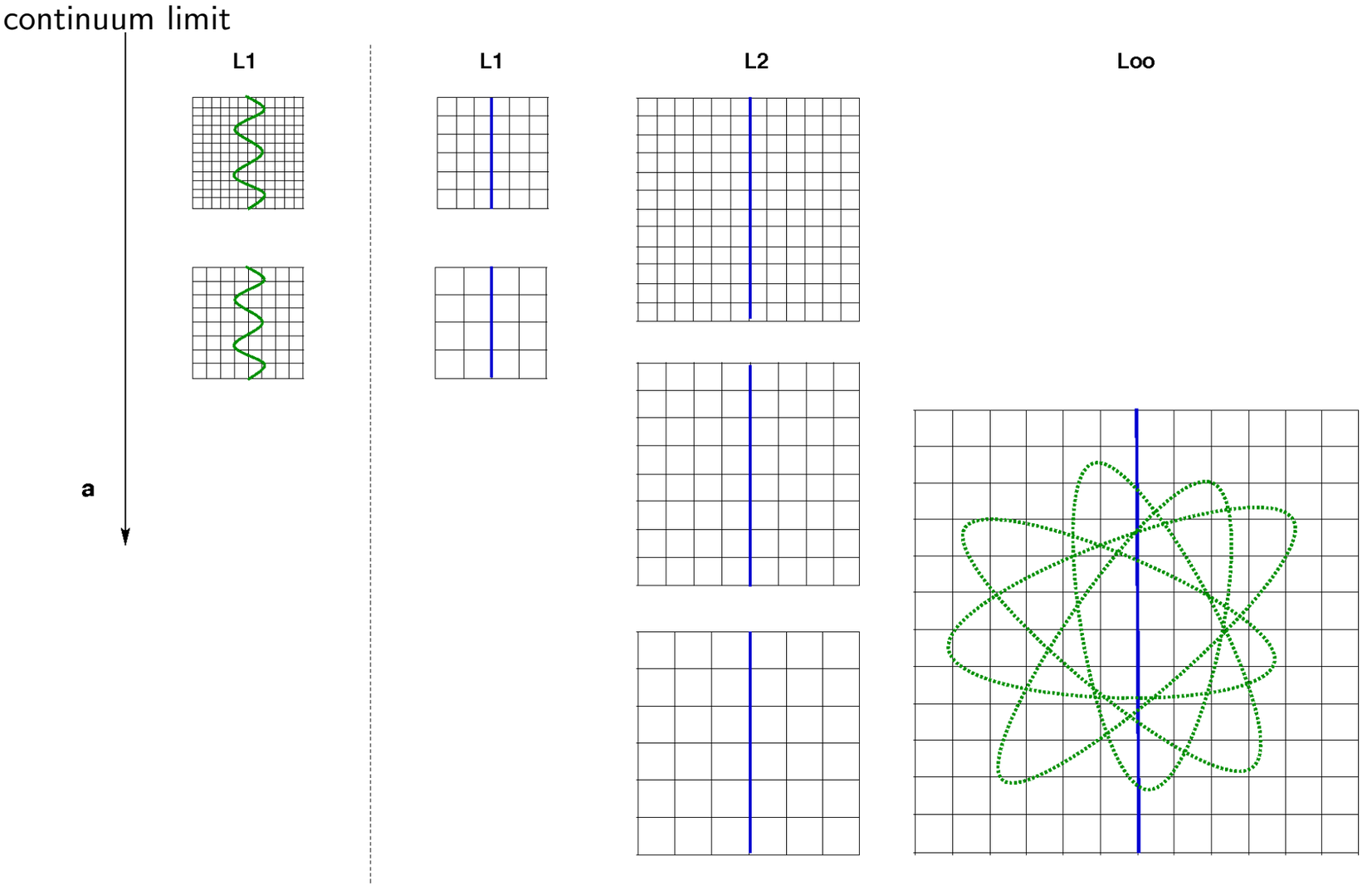


The heater wasn't so expensive, but the cable has cost a fortune!

Non-Perturbative HQET Strategy [J. Heitger, R. Sommer, ALPHA]

$\mathcal{L}_{\rm QCD}$

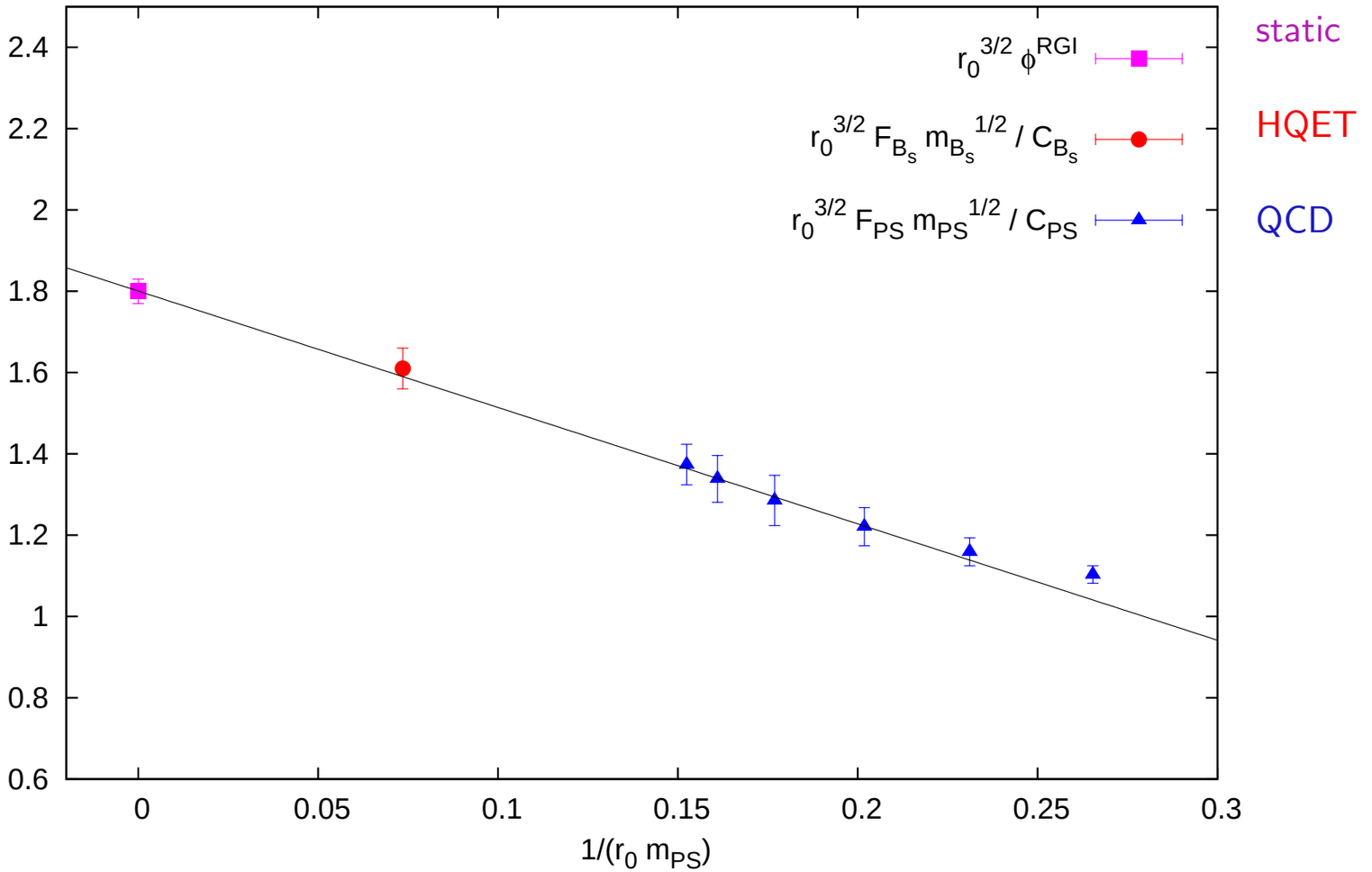
$\mathcal{L}_{\rm HQET} = \mathcal{L}_{\rm stat} - \omega_{\rm kin} O_{\rm kin} - \omega_{\rm spin} O_{\rm spin} + O(1/m_b^2)$



$\Phi_i(L, M, a) = A_{ij}(L, a) \cdot \omega_j(M, a) + B_i(L, a)$

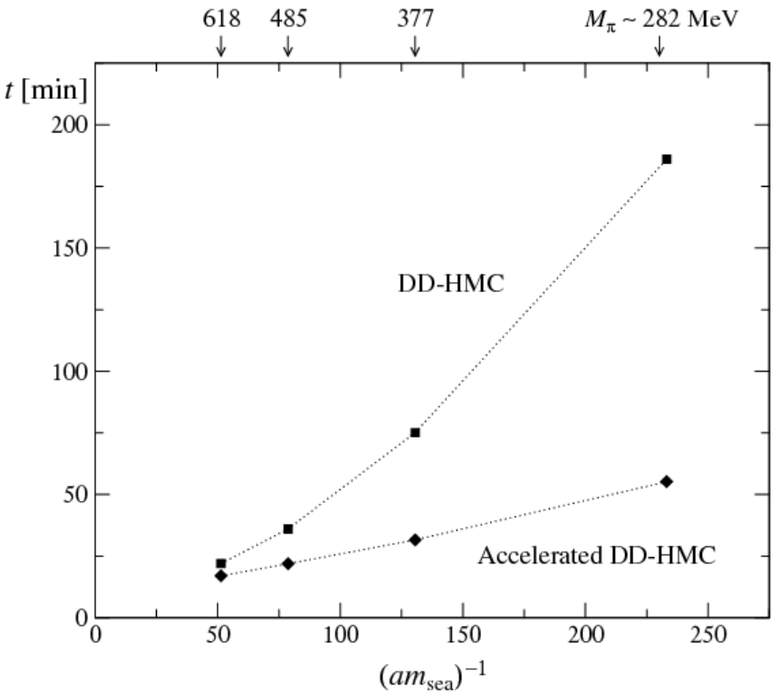
HQET Tests

F_{B_s} quenched



In progress: unquenched $N_f = 2$ . . .

Algorithmic Challenges

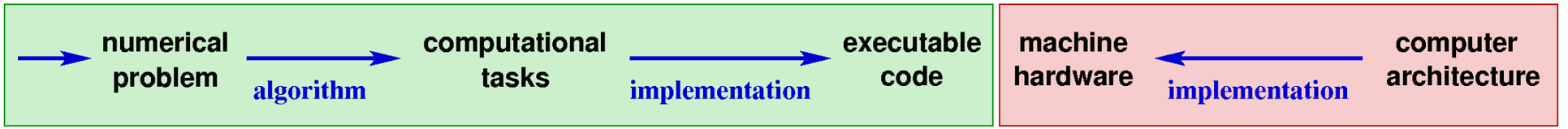


CPU cost per HMC trajectory: [M. Lüscher]

$$3 \text{ Tflops} \times \text{year} \cdot \left(\frac{L}{3\ fm}\right)^{5} \cdot \left(\frac{0.05\ fm}{a}\right)^{6} \cdot \left(\frac{20\ MeV}{m_q}\right)^{1} .$$

Towards continuum limit some observables (e.g. topological charge) can have serious critical slowing down (autocorrelations)!

Dedicated Machines



Idea: Focus on specific computational task to improve on

- Cost:

$$\begin{array}{ccccc} \text{development} & + & \text{production} & + & \text{operation} \\ \text{(politics, manpower)} & & \text{(market, technology)} & & \text{(RAS, kW in+out)} \end{array}$$

- Performance:

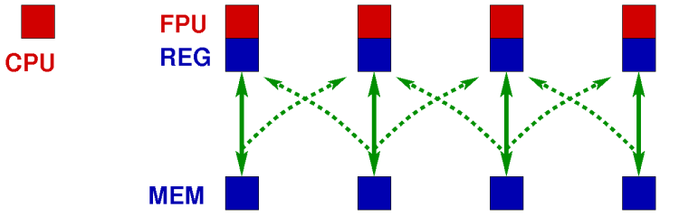
$$\begin{array}{ccccc} 1/T_{exe} & = & \text{work/cycle} & \times & f_{clk} \\ & & \text{(architecture)} & & \text{(technology)} \end{array}$$

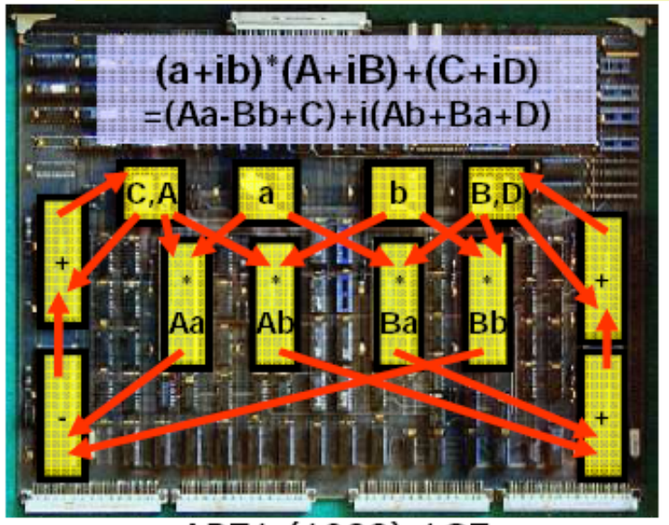
Examples of LQCD Machines:

- GF11 (IBM)
- QCDSP, QCDOC (Columbia U, IBM)
- CP-PACS (Tokio U, Hitachi)
- APE1, APE100, APEmille, apeNEXT (INFN, DESY, F)
- PC Clusters with tailored network (Budapest, . . .)

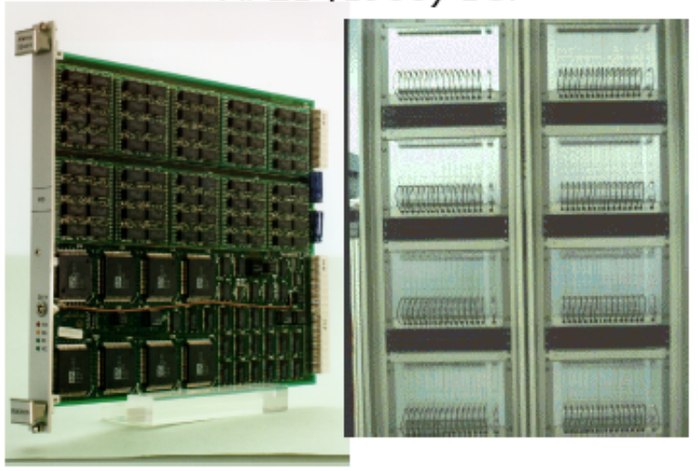
APE: Array Processor Experiment

- single logical CPU (SIMD)
- multiple FPUs with private memory
- 3d torus communication network

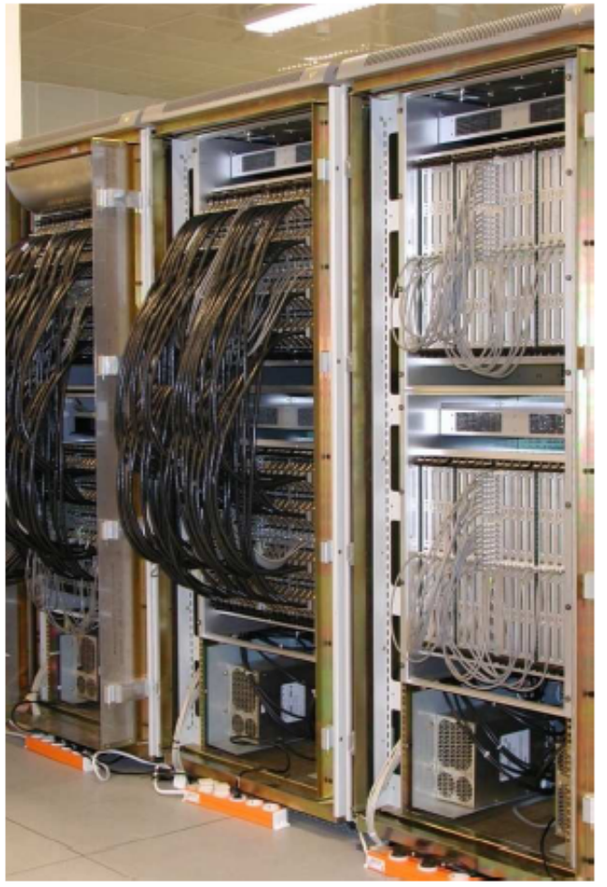




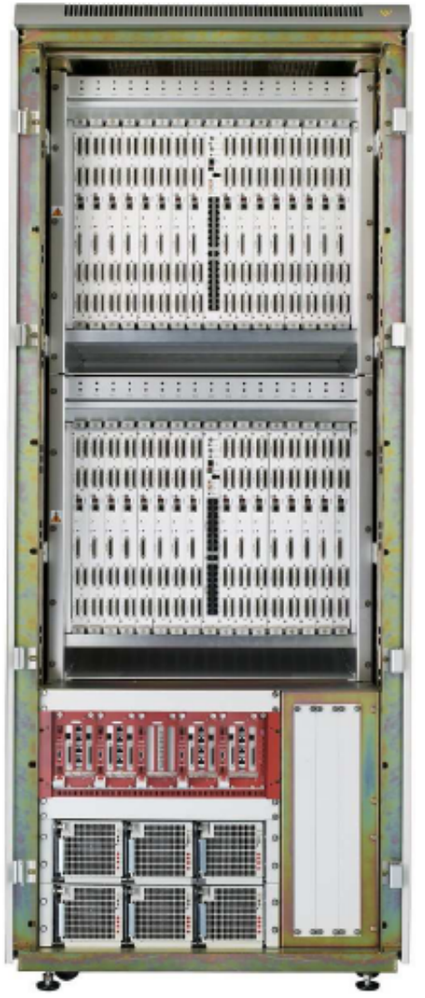
APE1 (1988) 1GF



APE100 (1992) 25GF, SP, REAL



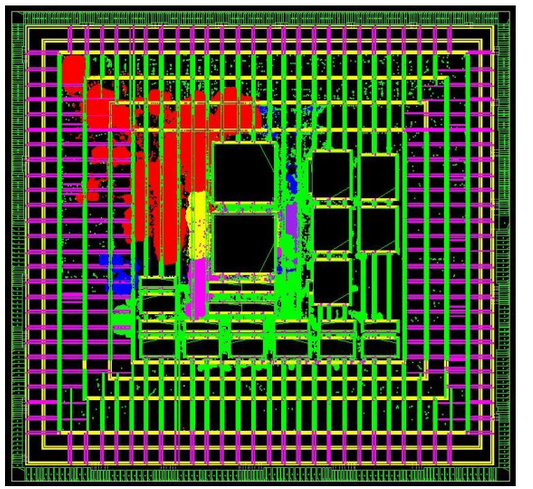
APEmille (1999) 128GF, SP, Complex



apeNEXT (2004) 800GF, DP, Complex

Evolution of APE Machines

Generation	APE100	APEmille	apeNEXT
bringup	1992	1999	2004
peak/board	0.4 Gflops	4 Gflops	12 Gflops
Architecture			
control	SIMD	SIMD	MIMD
synchronous	yes	yes	no
FP precision	32	32, 64	64
$a \times b + c$	$\mathbb{R}$	$\mathbb{C}, \mathbb{R}^2$	$\mathbb{C}, \mathbb{R}^2$
β_{FP}	2 SP/clk	8 SP/clk	8 DP/clk
β_{mem}/β_{FP}	1:4	1:4	1:4
β_{net}/β_{FP}	1:16	1:16	1:72
Technology			
ASICs	2	3	1
feature size	1.2 μ	0.5 μ	0.18 μ
f_{clk}	25 MHz	66 MHz	200 (135) MHz



Challenges of ASIC Design:

- Growing functionality
- Reduced architectural freedom
- Access to competitive chip technology at affordable price

Beyond APE

Basic Idea:

- ❒ **Fast commodity processor**
- ❒ **Tailored custom network (tightly coupled, simple)**
- ❒ **Custom node+system design**

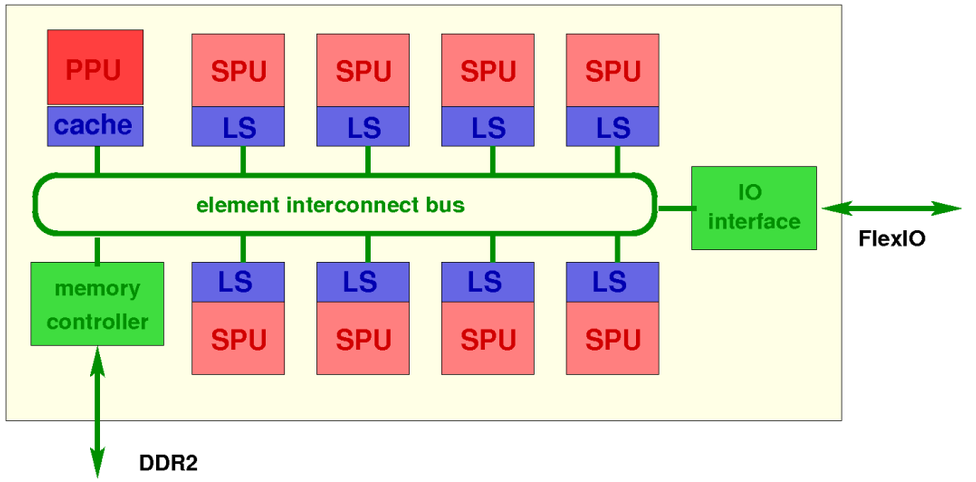
	APE	Cluster
	✘	✔
	✔	✘
	✔	✘

Challenges:

- High single-node performance ➜ multi-core processor
- Scalability ➜ fast processor interface and balanced torus network
- Cost efficiency ➜ integration and cooling

Cell BE Processor

- Innovative "Cell Broadband Engine" architecture
- Developed by Sony, Toshiba, IBM (Playstation 3)
- Enhanced version with DDR2 and DP: PowerXCell 8i (IBM blades, Roadrunner)



Key Features:

- PowerPC core for OS and control (PPU)
- 200 Gflops (SP peak) by 8 in-order cores (SPU) with SW-controlled private cache (LS)
- Integrated memory and IO interfaces
- Fast ring interconnect between cores and interfaces (EIB)

☛ A very fast APE board (without memory and communication network) on a single chip!

QPACE: Qcd PArallel computing on the Cell broadband Engine

Academic Partners:

- Uni Regensburg
- Uni Wuppertal
- Forschungszentrum Jülich
- DESY Zeuthen
- Uni Ferrara
- Uni Milano

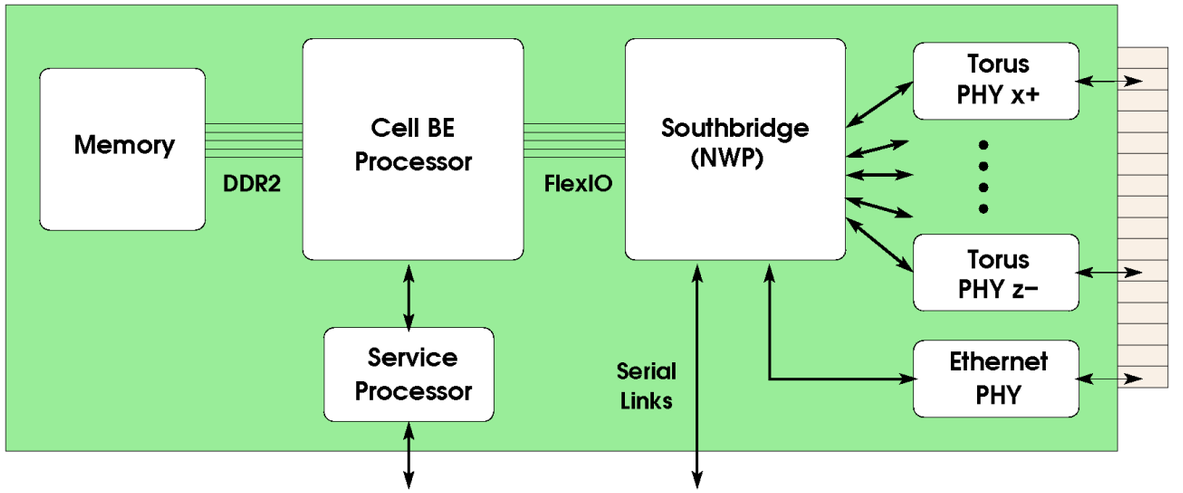
Industrial Partner: **IBM** (Böblingen, Rochester, La Gaude)

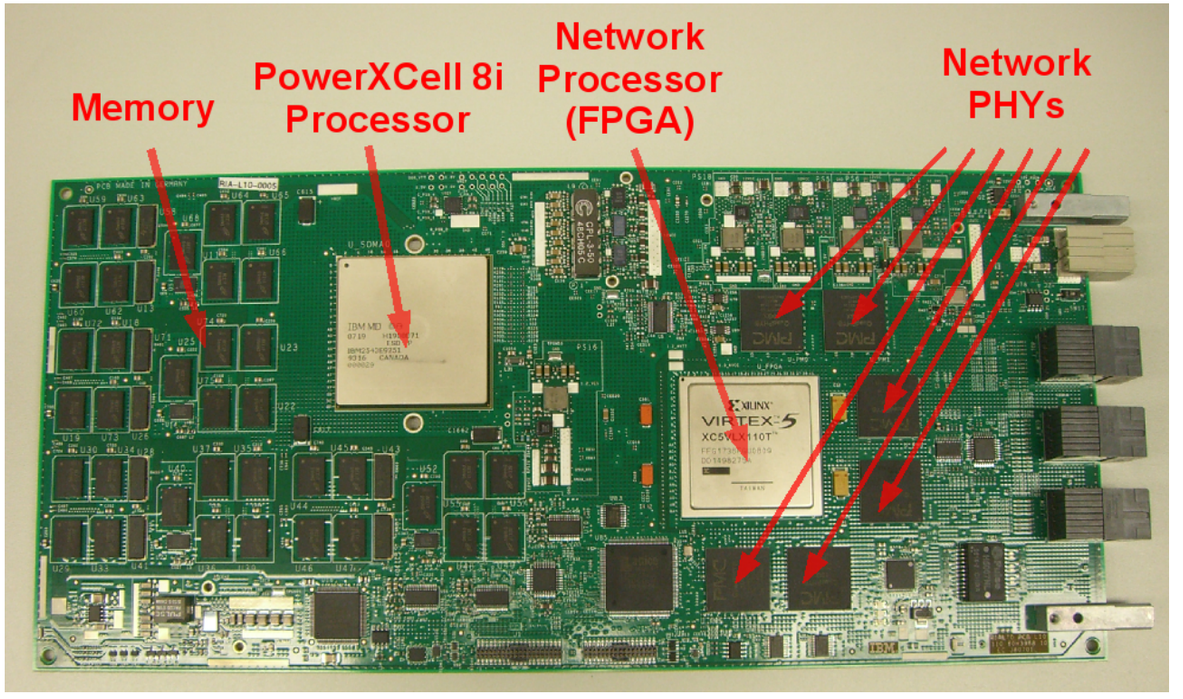
Support by: Eurotech (I), Knürr (D), Xilinx (US)

Main Funding: **DFG** (SFB TR55), **IBM**

Dec 2007	kick-off meeting
Apr 2008	prototype design completed
July 2008	start of prototype tests
Fall 2009	installation of $O(2000)$ nodes

QPACE Building Block: Node Card

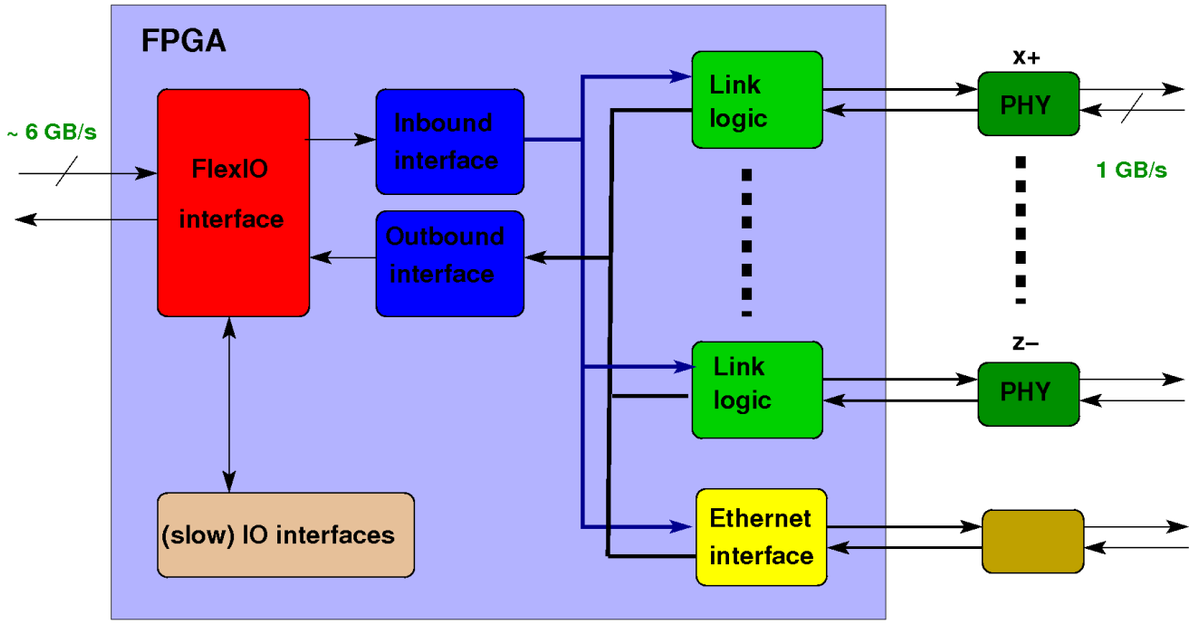




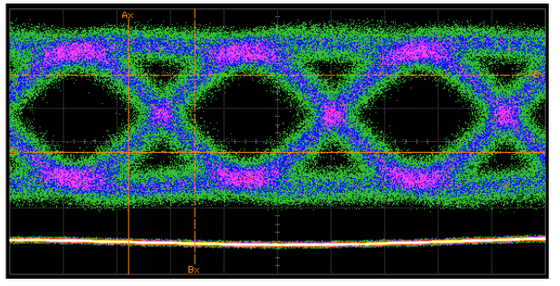
Network Processor (NWP)

Main Purpose: Route and control the data flow between Cell and 6 links of torus network

❒ Field Programmable Gate Array (FPGA) Xilinx Virtex5 LX110T

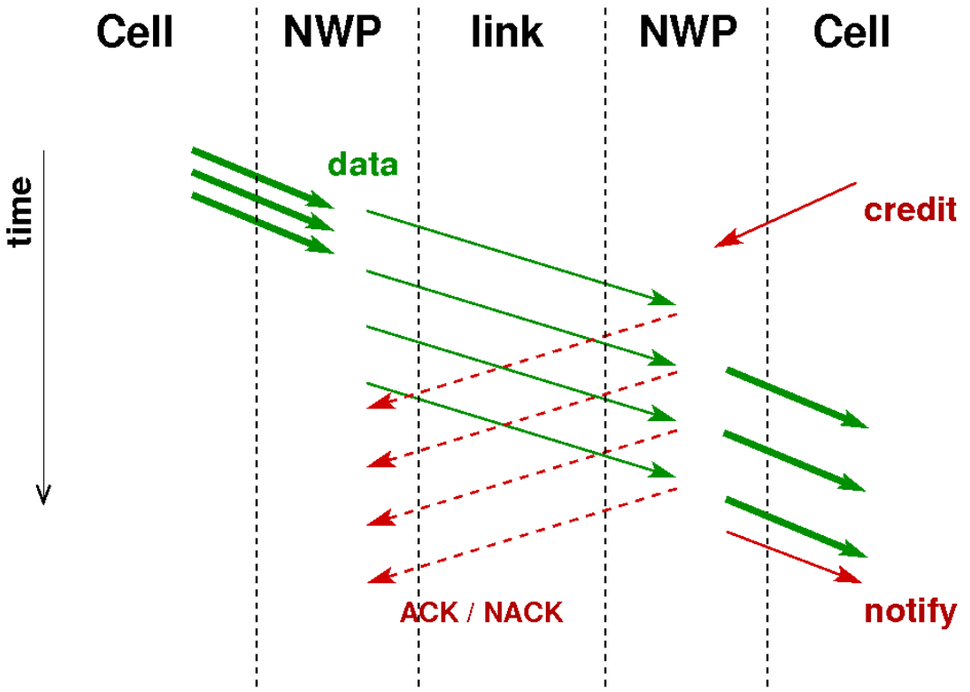


❒ External 10GbE/PCIe transceiver (PHY) PMC Sierra PM8358



$T_{eye} \approx 100$ ps

QPACE Communication Model



- 2-sided (separate send + receive)
- non-blocking (separate initiate + complete)
- data transport to/from main memory or LS
- nearest-neighbour connectivity
- light-weight link layer protocol
- multiple use of same link by 8 virtual channels

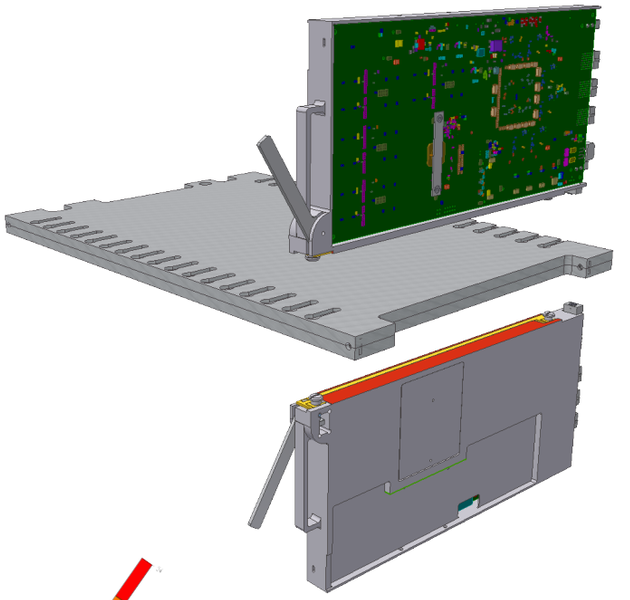
More details: `http://moby.mib.infn.it/~simma/tnw` [M. Pivanti, F. Schifano, H.S]

QPACE Cooling

[G. Goldrian, IBM]

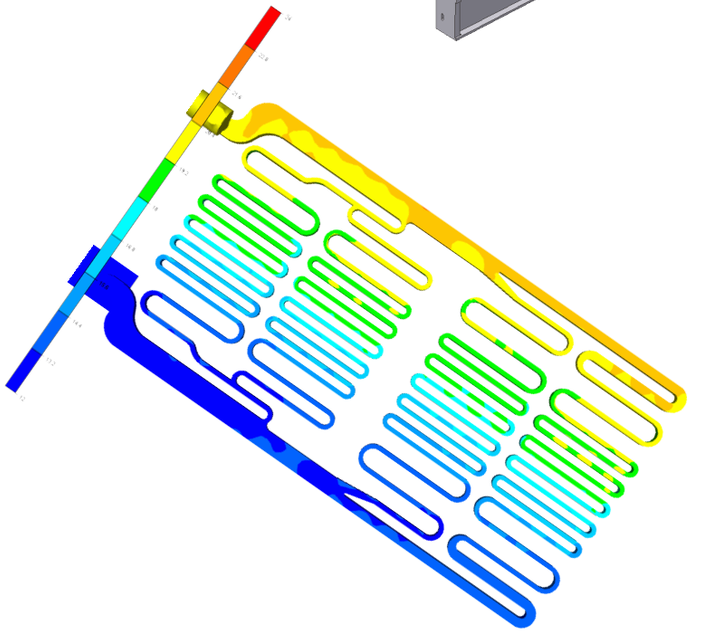
Concept:

- node housing ➜ heat conductor
- cold plate ➜ liquid cooling



Simulation:

- 10 l/min water at 12°
- load 4224 W



QPACE Cooling

[G. Goldrian, IBM]

Concept:

- node housing ➜ heat conductor
- cold plate ➜ liquid cooling

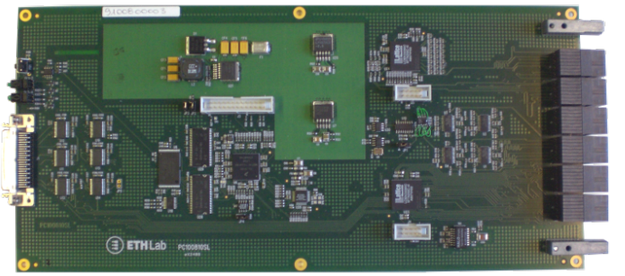
Simulation:

- 10 l/min water at 12°
- load 4224 W
- confirmed by tests . . .

Other QPACE Hardware Components

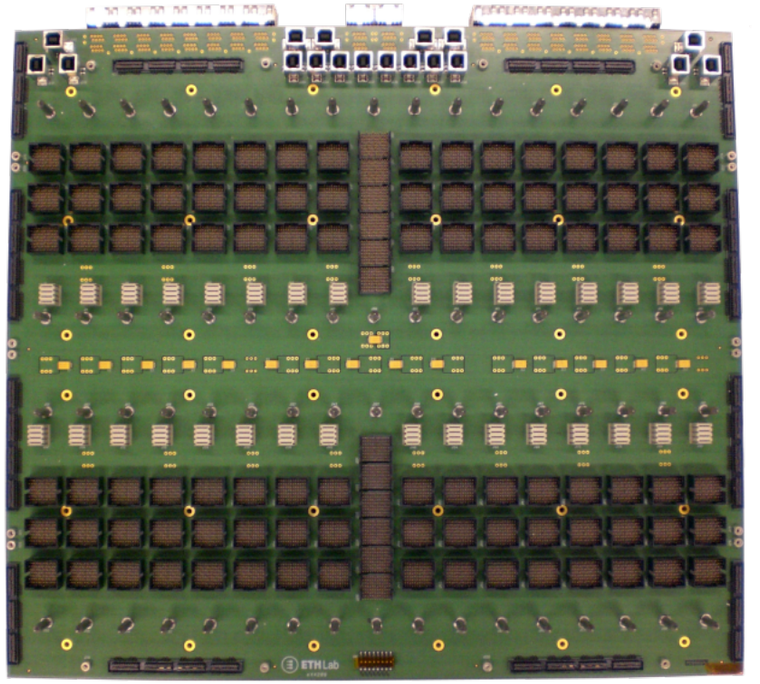
❒ Root Card

- Micro-processor with Ethernet interface
- Global signal tree (like APE)
- Serial links (configuration)
- Clock distribution



❒ Backplane

- High-speed signals and power distribution
- 22 layers
- 13'000 holes



QPACE Rack

Performance Density:

52 TFlops (SP) / rack

- Footprint: 80×120 cm
- Weight: O(1000) kg
- Power: O(29) kW

Power Efficiency:

➜ Number #1 in Green500 of Nov 2009

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)	TOP500 Rank*
1	722.98	Forschungszentrum Juelich (FZJ)	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	59.49	110
1	722.98	Universitaet Regensburg	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	59.49	111
1	722.98	Universitaet Wuppertal	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	59.49	112
4	458.33	DOE/NNSA/LANL	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Infiniband	276	29
4	458.33	IBM Poughkeepsie Benchmarking Center	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Infiniband	138	78
6	444.25	DOE/NNSA/LANL	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband	2345.5	2
7	428.91	National Astronomical Observatory of Japan	GRAPE-DR accelerator Cluster, Infiniband	51.2	445
8	379.24	National SuperComputer Center in Tianjin/NUDT	NUDT TH-1 Cluster, Xeon E5540/E5450, ATI Radeon HD 4870 2, Infiniband	1484.8	5
9	378.77	King Abdullah University of Science and Technology	Blue Gene/P Solution	504	18
9	378.77	EDF R&D	Blue Gene/P Solution	252	49

[www.green500.org]

Comparison

	apeNEXT	PowerXCell 8i	Intel Nehalem
bringup	2004	2007	2008
peak (DP)	12 Gflops/board	100 Gflops/chip	50 Gflops/chip
Technology			
feature size	180 nm	65 nm	45 nm
power	$O(10)$ W	$O(100)$ W	$O(80)$ W
f_{clk}	200 (135) MHz	3.2 GHz	2.8 GHz
Architecture			
control	SPMD	8 cores + PPU	4 cores
$a \times b + c$	$\mathbb{C}, \mathbb{R}^2$	$\mathbb{R}^4, \mathbb{R}^2$	2 × SSE3
β_{FP}	8 DP/clk	32 DP/clk	16 DP/clk
β_{mem}/β_{FP}	1:4	**1:32**	1:12
β_{net}/β_{FP}	1:72	1:140	1:140
cache	—	8×256 KB (LS)	8 MB (L3)

☛ Cell seems to have memory bottleneck for LQCD

Hardware Model

Devices for:

- control
- data storage
- data transport/processing

Parametrized by:

ISA, . . .

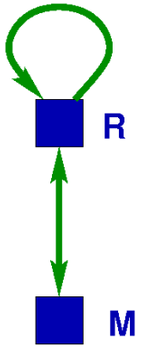
size: $0 \leq \sigma_i < \infty$

bandwidth: $\beta_{ij} < \infty$
latency: $\lambda_{ij} \geq 0$

Structure:

described by a "Hardware Architecture Graph" (HAG) with

- nodes = storage devices
- arcs = transport devices



Application Model

Computational Tasks:

- data set (input, output, temporary variables)
- data transport/processing tasks (assignments)

Quantified by:

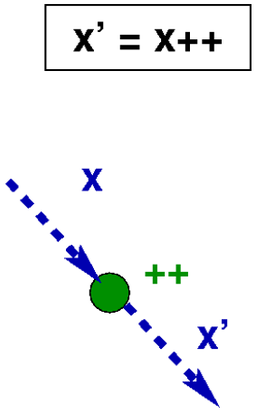
storage requirement: S_i

information exchange: I_{ij}

Data Dependencies:

described by a Directed Acyclic Graph (DAG) with

- arcs = RAW dependencies (variable lives)
- nodes = transport operations



Implementation

Main problems:

❒ **Code Selection**

transport operations (DAG) $\longrightarrow$ HW instructions (DAG')

❒ **Resource Allocation**

data set (variables) = arcs of DAG' $\longrightarrow$ storage devices = nodes of HAG

operations (instructions) = nodes of DAG' $\longrightarrow$ transport devices = arcs of HAG

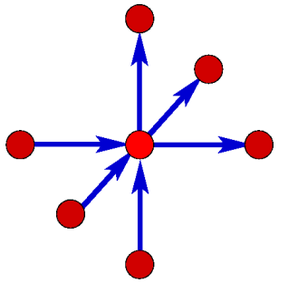
❒ **Scheduling**

☛ Allocation and scheduling are interrelated and NP-hard problems (to be tackled by algorithm, programmer, compiler, hardware)

Main Computational Task in LQCD: Wilson-Dirac Operator

Hopping term of $\not{D}\phi$:

$$\phi'(x) = \sum_{\mu=1}^{4} \{U(x,\mu)(1-\gamma_\mu)\phi(x+\hat{\mu}) + \cdots\}$$

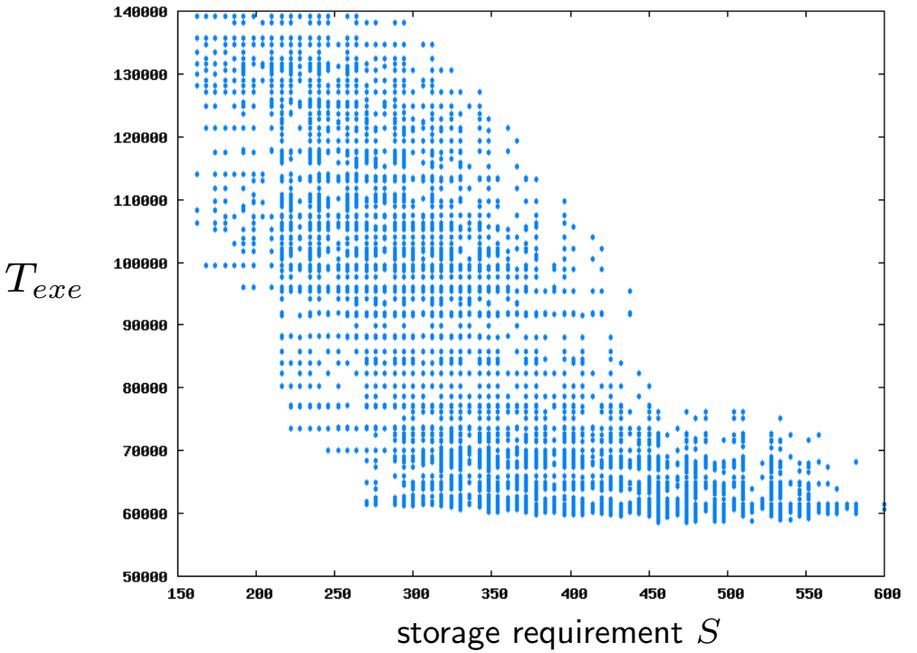


Naive analysis: (without data re-use)

- FP Computations: $I_{FP} = 1320$ FP operations $\quad (\rightarrow 840$ `muladd`$)$
- Memory Access: $I_{mem} = 9|\phi| + 8|U| = 360$ FP words

Performance Model: (minimal $I_{mem} = 192$ FP words)

- data re-use in a small cache
- prefetch to hide latencies
- 160'000 different schedules
- $T_{ij} = \lambda_{ij} + I_{ij}/\beta_{ij}$



Summary and Outlook

Progress is slow, but

- ❒ good perspectives to obtain accurate lattice results for B-physics
- ❒ dedicated machine developments can pay off
- ❒ implementation space for scientific computing problems is amazingly rich

Summary and Outlook

Progress is slow, but

- ❒ good perspectives to obtain accurate lattice results for B-physics
- ❒ dedicated machine developments can pay off
- ❒ implementation space for scientific computing problems is amazingly rich

All the best wishes and many thanks to Daniel!