QPACE: Power-efficient parallel architecture based on IBM PowerXCell 8i

Dirk Pleiter

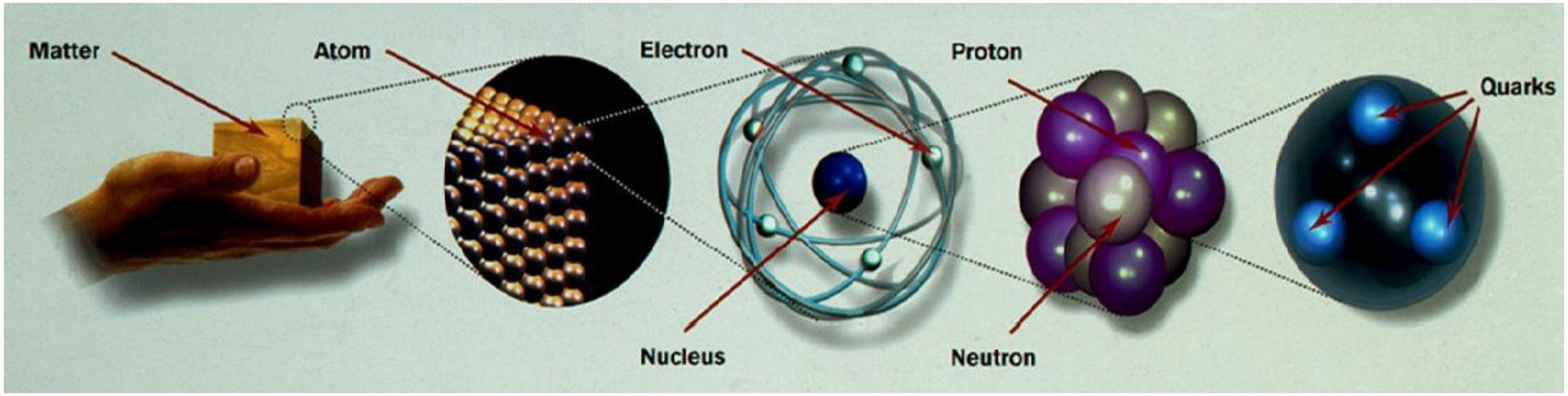
DESY (Zeuthen site)

17 September 2010, EnA-HPC, Hamburg

QPACE Collaboration / Credits

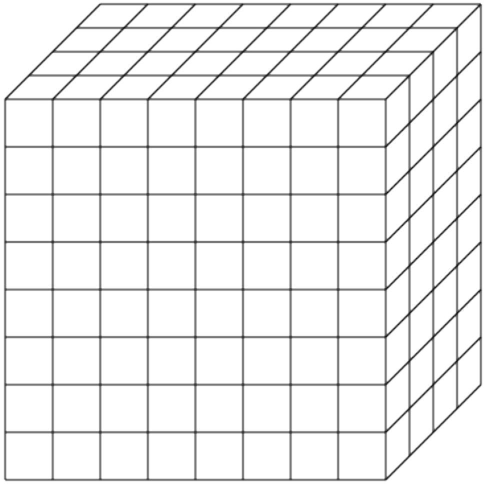
- Academic partners
  - U Regensburg: S. Heybrock, D. Hierl, T. Maurer, B. Mendl, N. Meyer, A. Nobile, A. Schäfer, S. Solbrig, T. Streuer, T. Wettig, F. Winter
  - U Wuppertal: Z. Fodor, A. Frommer, M. Hüsken
  - U Ferrara: M. Pivanti, F. Schifano, R. Tripiccione
  - U Milano: H. Simma
  - DESY Zeuthen: D.P., K.-H. Sulanke
  - Research Lab Jülich: M. Drochner, N. Eicker, T. Lippert
- Industrial partner: IBM (Böblingen, Rochester, La Gaude)
  H. Baier, H. Boettiger, A. Castellane, J.-F. Fauh, U. Fischer, G. Goldrian, C. Gomez, T. Huth, B. Krill, J. Lauritsen, J. McFadden, I. Ouda, M. Ries, H.J. Schick, J.-S. Vogt
- Main funding: DFG (SFB TR55), IBM
- Special partners: Eurotech (I), Knürr (D), Axe (I), Zollner (D)

Building blocks of matter

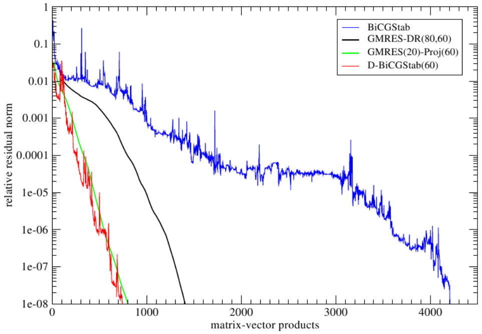


- Quarks are the constituents of matter which strongly interact exchanging gluons
- Particular phenomena:
  - Confinement
  - Asymptotic freedom (Nobel Prize 2004)
- Theory of strong interactions = **Quantum Chromodynamics (QCD)**

Observables from first principles



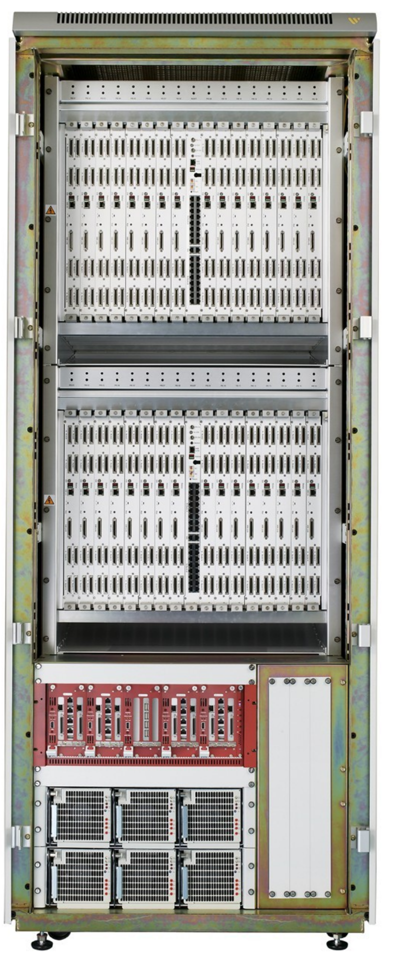
- Discretize theory on finite, **4-dimensional lattice**: Formulation of QCD called **Lattice QCD** → enables **numerical simulations**
- Problem includes inversion of **huge, but sparse linear equation** of type $M x = b$
  - Typical dimension of *M:* 10^7-10^9



- Standard algorithms: iterative, Krylovspace methods (e.g. conjugate gradient)

Computing requirements

- Application performance signature
  - Floating-point intensive (typically memory bound)
  - Equal load distribution, simple flow control
  - Homogeneous communication patterns
- Resource requirements
  - Progress in this field is largely limited by available compute power
  - Lattice QCD community aims for **O(1–3) PFlops/s sustained** beyond 2010



Special purpose machines

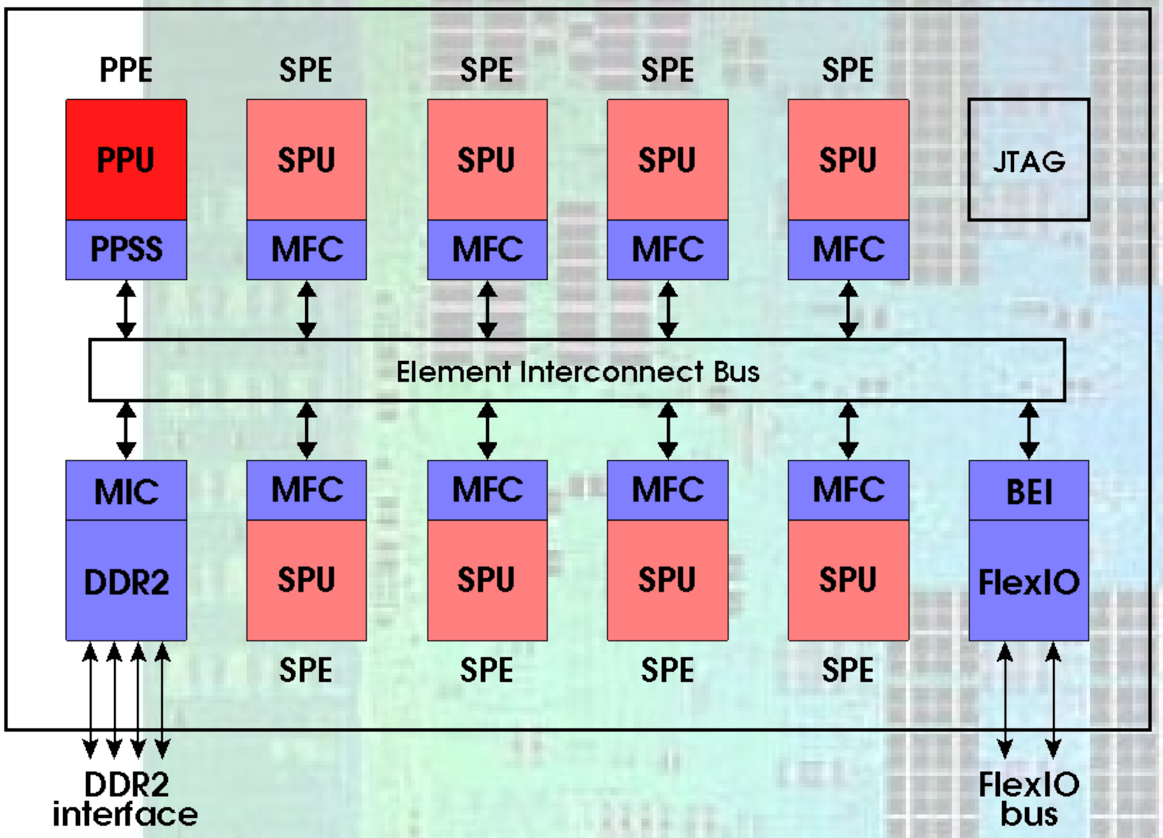
- Special purpose machines
  - ApeNEXT, QCDOC
  - SOC design
- Optimized commodity solutions
  - PC-Cluster based e.g. PACS-CS, Aurora
  - GPUs

QPACE Architecture

QPACE Architecture

- Goal: Scalable architecture optimized for lattice QCD calculations
- Concept:
  - Fast commodity processor = IBM PowerXCell 8i
  - Custom network → custom network processor
  - Custom system design
- Challenges:
  - High single node sustained performance
  - Scalability = high bandwidth, low latency network
  - Cost efficient system integration

PowerXCell 8i Processor

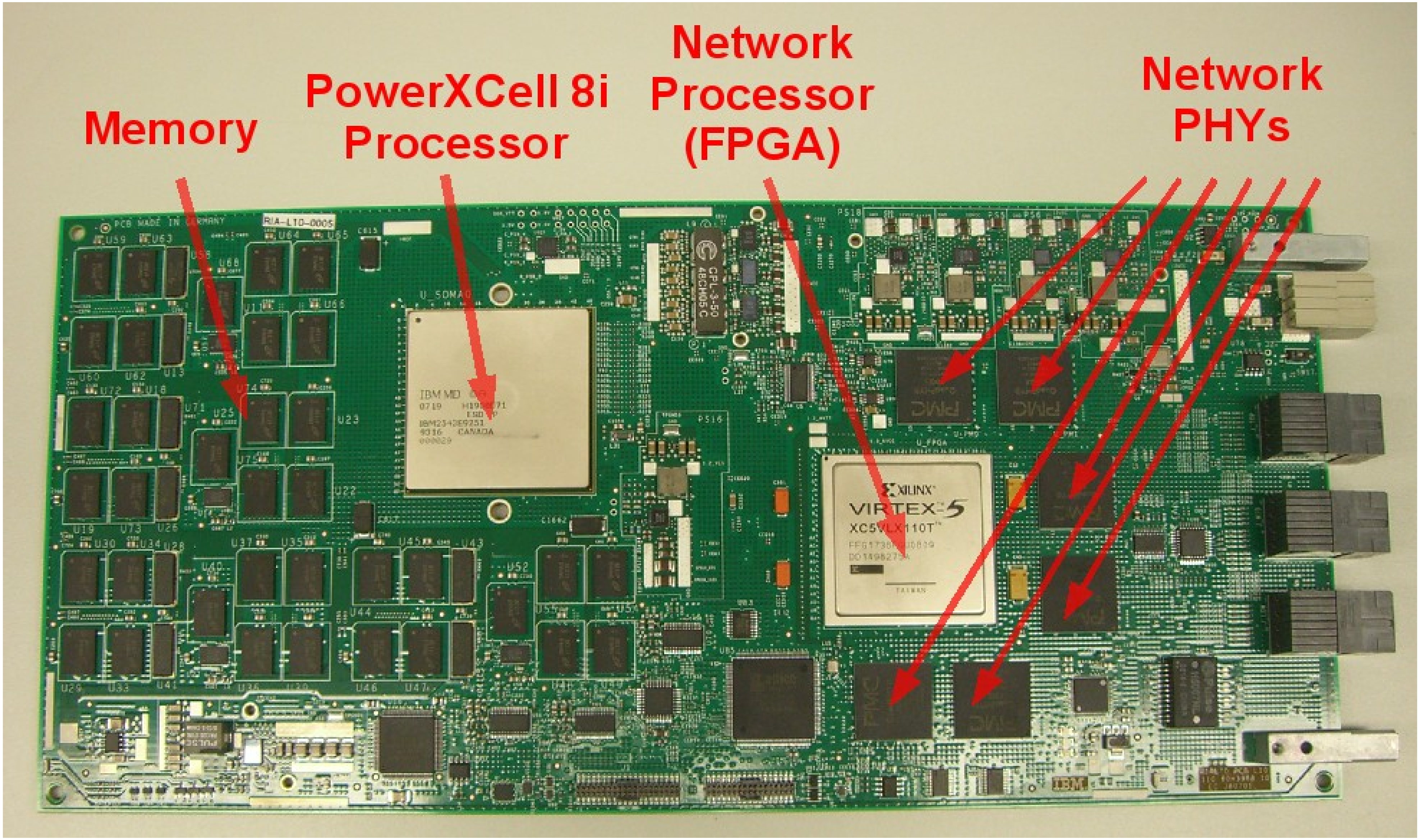


- 8 Synergistic Processing Elements
  - double precision, IEEE rounding
  - 12.8/25.6 GFlops double/single prec.
  - 256 kBytes local store
- DDR2 memory, 25 Gbyte/s
- Fast interconnect bus, 200 Gbyte/s
- Used for Roadrunner (rank 1 in top500 2008-09)

Network Processor / Torus Network

- Network Processor implements custom network
  - Hardware solution: FPGA
- I/O fabric managing data transfer via
  - 1x link to processor: 2.5 GByte/s, FlexIO
  - 6x links to torus network: 1 GBytes/s, 10 GbE
- Torus network features [F. Schifano, H. Simma]
  - Lean protocol, 2-sided communication model
    - Nearest neighbour communication only
  - SPU-SPU communication
  - Virtual Channel support

Node-card



Memory

PowerXCell 8i Processor

Network Processor (FPGA)

Network PHYs

QPACE rack

- Significant performance density due to liquid cooling system

26/52 TFlops/rack
(double/single precision)

- Other key rack parameter

Max. power consumption	< 35 kW
Typical power consumption	21-27 kW
Foot print	80 x 120 cm
Weight	O(1000) kg

Power Optimisation and Efficiency

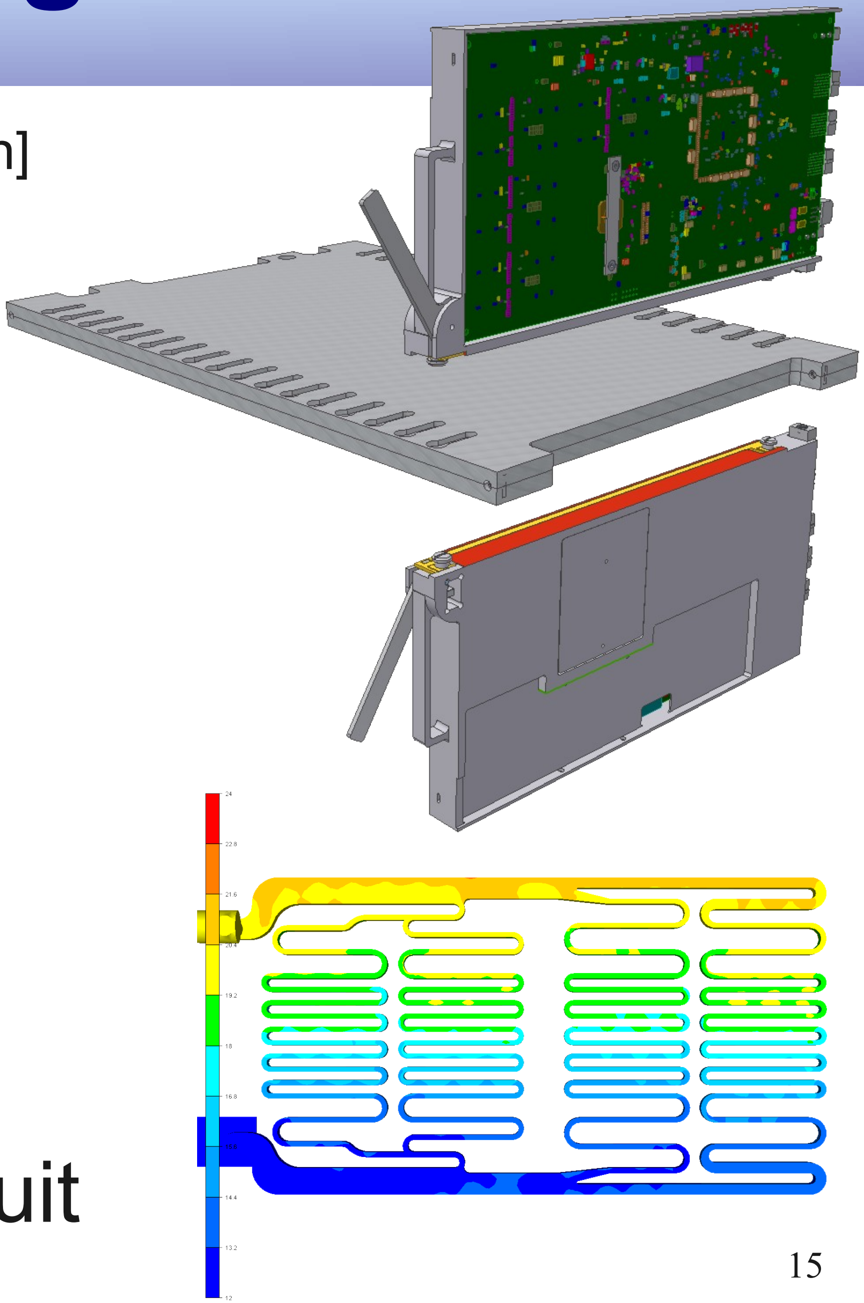
Power Reduction Strategies

- Selection of power efficient components
  - Processor: All Green500 #1 systems since June 2008 are based on PowerXCell 8i
  - Power Supplies: Efficiency ≥ 89%
  - Minimize amount of memory per node
- Voltage tuning
- New cooling system
- Not considered here: throttling of sub-systems
  - Special purpose systems can avoid unused sub-systems

Cooling

[G. Goldrian]

- Concept:
  - NC mounted in housing = heat conductor
  - Housing connected to liquid cooled cold plate
- Critical thermal interfaces
  - Processor – thermal box
  - Thermal box – cold plate
- Dry connection between node-card and cooling circuit



Cooling (2)

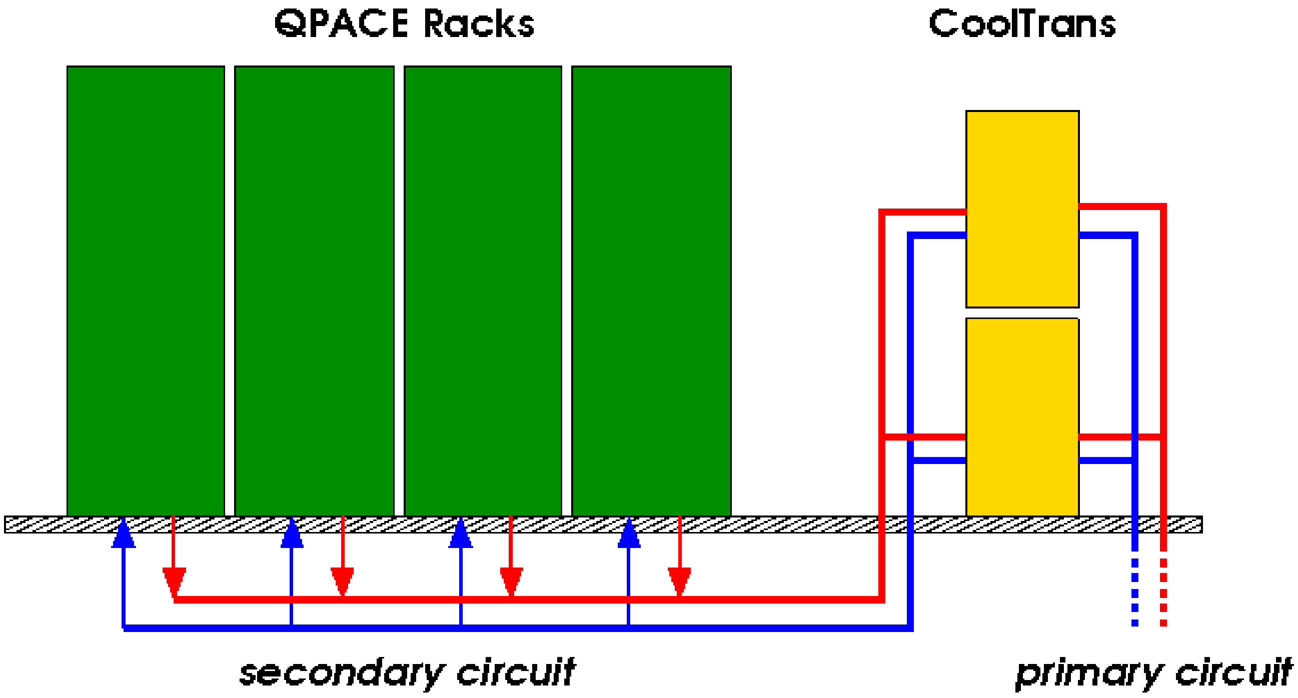
- Temperature limits

PowerXCell 8i	95 °C
Virtex-5	80 °C

- Consider ΔT = Difference in temperature at water inlet and processor
  At maximum load: $\Delta T \approx 35\text{-}40$ °C
- Water inlet temperature >30 °C feasible

Cooling (3)

- Cooling circuit of 4 rack installation:



- Power required for cooling:
  - CoolTrans water pumps
  - Power supply and Ethernet switch fans

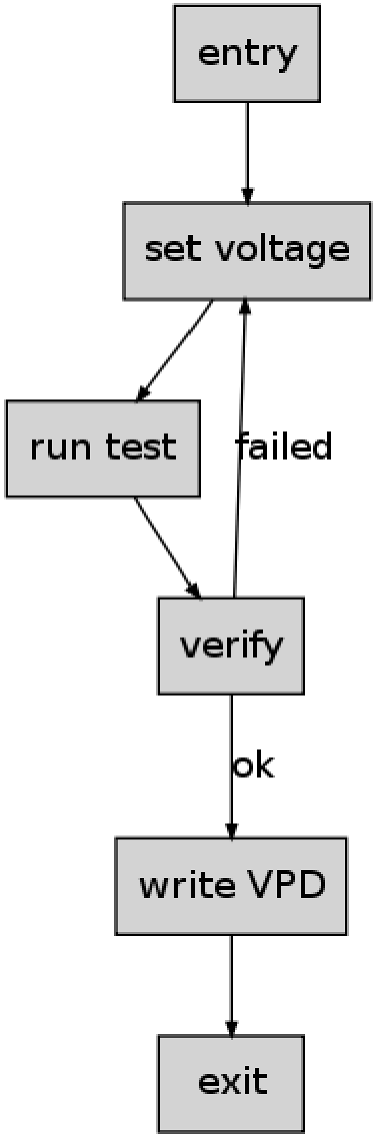
Voltage Tuning

[G. Goldrian, T. Huth, J.S. Vogt]

- CMOS gates power dissipation ~V^2
  → Voltage reduction is a promising strategy
- Suitable for HPC? Yes
  Voltage guard-bands → room for optimization
- Guard-bands too large for different reasons:
  - Lack of control of supply voltage level
    - Can be improved by better components/circuits
    - QPACE: reduction of memory voltages
  - Voltage limits differ significantly between different samples of the same component, e.g. processor

Processor Core Voltage Tuning

- VMIN Tuning algorithm:
  - Controlled by BMC
  - Test selected to stress all relevant functional units
- Tuning only performed once
  - Tuning results stored in VPD
  - Optional re-tuning during node live cycle
- After determination of cutting edge guard-band is added
  - Safety margin
  - Anticipate processor ageing



Typical Power Consumption

- Typical power consumption of a 4-rack (1024 nodes) installation:

Total	84-108 kW
Nodes	76-98 kW
Power supplies	8-10 kW
Heat exchanger	< 3.6 kW
Ethernet switches	< 2 kW

- Power consumption is completely dominated by power consumed by processing nodes
  - Strongly load dependent

Power Consumption Details

- Considered loads
    - Linux only
    - Application kernel
    - Synthetic benchmark Powerload SPU
- Measured power consumption per 32 nodes

Load	Default voltage	VMIN enabled
Linux	2.3 kW	2.2 kW
Application kernel	2.7 kW	2.4 kW
Powerload SPU	4.3 kW	4.0 kW

VMIN tuning gain: O(10%)

Green500

[H. Boettiger, B. Krill, S. Rinke]

- HPC communication requirements differ significantly from LQCD applications
  E.g. large messages, any-to-any communication
- Modified QPACE system configuration
  - Modified NWP performing inbound DMA read operations to fetch TX data
  - QPACE torus network API support added to OpenMPI BTL
- Latest result: 773.4 MFlops/W

44% increase compared to previous No. #1

Application Kernel and Other Architectures

- Power efficiency for key application kernel (parallelized version):
  40-50 GFLops (SP) / 73 W = 544-681 MFlops/W
- Comparison with GPU-based system
  - Same application on single GPU: 116.1 Gflops
  - Estimated power consumption: 250-300 W
    → 400-450 MFlops/W
- GPU-based HPL: 492.64 Mflops/W
  Dawning Nebulae at National Supercomputing Centre in Shenzhen

Summary & Conclusions

Summary & Conclusions

- QPACE is a new, scalable Lattice QCD machine based on the IBM PowerXCell 8i
- Design highlights:
  - FPGA directly attached to processor
  - LQCD optimized torus network
  - Novel, cost-efficient liquid cooling system
  - Very power efficient architecture
- Two installations with an aggregate performance of 200/400 TFlops (DP/SP)
  - Good sustained performance of O(20-30%) for key LQCD kernels → O(10-15) TFlops/rack (SP)

Summary & Conclusions

- Multiple power optimisation strategies:
  - Selection of components
  - Voltage optimisation
  - Cooling system
- Optimization of power efficiency at many places → significant overall improvement
- Successful interplay of application and technology driven HPC development also a model for future projects