Comput Sci Res Dev (2010) 25: 149–154
DOI 10.1007/s00450-010-0122-4

SPECIAL ISSUE PAPER

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

H. Baier · H. Boettiger · M. Drochner · N. Eicker · U. Fischer · Z. Fodor · A. Frommer · C. Gomez · G. Goldrian · S. Heybrock · D. Hierl · M. Hüsken · T. Huth · B. Krill · J. Lauritsen · T. Lippert · T. Maurer · B. Mendl · N. Meyer · A. Nobile · I. Ouda · M. Pivanti · D. Pleiter · M. Ries · A. Schäfer · H. Schick · F. Schifano · H. Simma · S. Solbrig · T. Streuer · K.-H. Sulanke · R. Tripiccione · J.-S. Vogt · T. Wettig · F. Winter

Published online: 30 July 2010

Abstract QPACE is a novel massively parallel architecture optimized for lattice QCD simulations. Each node comprises an IBM PowerXCell 8i processor. The nodes are interconnected by a custom 3-dimensional torus network implemented on an FPGA. The architecture was systematically optimized with respect to power consumption. This put QPACE in the number one spot on the Green500 List published in November 2009. In this paper we give an overview of the architecture and highlight the steps taken to improve power efficiency.

Keywords Parallel architectures · Special-purpose and application-based systems

H. Baier · H. Boettiger · U. Fischer · G. Goldrian · T. Huth · B. Krill · J. Lauritsen · M. Ries · H. Schick · J.-S. Vogt
IBM Deutschland Research & Development GmbH, 71032 Böblingen, Germany

M. Drochner · N. Eicker · T. Lippert
Research Center Jülich, 52425 Jülich, Germany

N. Eicker · Z. Fodor · A. Frommer · M. Hüsken · T. Lippert
University of Wuppertal, 42119 Wuppertal, Germany

C. Gomez
IBM La Gaude, Le Plan du Bois, La Gaude 06610, France

S. Heybrock · D. Hierl · T. Maurer · B. Mendl · N. Meyer · A. Nobile · A. Schäfer · S. Solbrig · T. Streuer · T. Wettig · F. Winter
Department of Physics, University of Regensburg, 93040 Regensburg, Germany

I. Ouda
IBM Rochester, 3605 HWY 52 N, Rochester, MN 55901-1407, USA

M. Pivanti · F. Schifano · R. Tripiccione
University of Ferrara, 44100 Ferrara, Italy

D. Pleiter (✉) · H. Simma · K.-H. Sulanke · F. Winter
Deutsches Elektronen Synchrotron (DESY), 15738 Zeuthen, Germany
e-mail: dirk.pleiter@desy.de

1 Introduction

Numerical simulations of lattice Quantum Chromodynamics (QCD) require a huge amount of computational resources. To carry out such calculations highly scalable capability computers providing TFlops (and soon PFlops) of compute power are required. For a long time the need for compute cycles has been the driving factor for groups in this field to develop, build and deploy application-optimized HPC systems at affordable cost.

Since a significant and increasing fraction of the total cost of ownership is due to operational costs, in particular the electricity bill, power efficiency has become an important design goal. Furthermore, keeping power consumption low is a prerequisite for a high integration density, which typically reduces system costs. In QPACE several design features contributed to a reduction of the power consumption.

2 Application requirements

In lattice QCD applications typically the largest fraction of the compute resources is used for calculating the solution of a linear equation of type $Mx = b$. Because M is a huge but sparse matrix iterative Krylov-space methods are used. This computational task can be split in a (small) set of micro-tasks like matrix-vector multiplications, vector-vector operations and global reductions.

The single-node performance depends on efficient utilization of the floating-point pipelines and, far more importantly, on the efficient management of the input and output data. On current processor architectures including the PowerXCell 8i the bandwidth to the main memory is small compared to the available processing power for the relevant kernels. Good performance can only be achieved by algorithms and implementations which make optimal use of the on-chip memory, like the Local Store in case of the PowerXCell 8i processor.

The requirements of the application concerning the inter-node interconnect are relatively moderate. To parallelize lattice QCD applications a homogeneous domain decomposition is used. Communication with nearest-neighbor nodes arranged in a 3- or 4-dimensional torus is usually sufficient. The bandwidth and latency requirements depend on the problem size, and for a scalable architecture these hardware parameters have to be chosen such that also in the case of smallest local problem size good performance can be achieved. In practice meeting the latency requirements in the range $\lesssim 1$ µs is the bigger technological challenge.

The meet these application requirements our strategy was to use one of the most powerful commodity processors available at project start and suitable for our application and to interconnect many such processors by an application-optimized custom network.

3 Architecture overview

The main building block of the QPACE architecture is the node card. The most important components on a node card are an IBM PowerXCell 8i processor, 4 GBytes of DDR2 memory, a Xilinx Virtex-5 FPGA and 6 10-GbE PHYs (see Fig. 1). 32 of these node cards can be connected to a single backplane together with 2 root cards. The root cards are used for management and control of the node cards. The node cards are water-cooled, and there is no need for air being able to flow from one side of the rack to the other. This allowed us to mount two rows of backplanes in the middle of the rack and to insert node cards from both the rack's front and back side. A rack houses 8 backplanes, i.e., the maximum number of node cards per rack is 256 (see Fig. 2).

A particular feature of the QPACE architecture is the use of an FPGA to implement a custom I/O fabric. This Network Processor (NWP) is directly connected to the PowerXCell 8i processor via two 8-bit wide (full-duplex) bi-directional links with a bandwidth of 4.6 GBytes/sec per direction. Also connected to the FPGA are six (full-duplex) bi-directional links to the 10-Gigabit Medium Independent Interfaces (XGMII) of the 10-GbE PHYs. The 6 external links are used to connect the nodes in a 3-dimensional torus and have a bandwidth of 1 GBytes/sec each. Additionally there are a number of lower-speed links including a Gigabit Ethernet link used for booting the node card and user I/O.

The main performance parameters of a QPACE rack are summarized in Table 1. For more technical details on the architecture see Ref. [1].

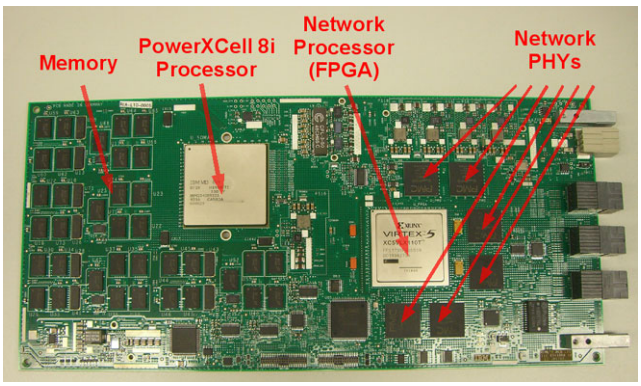

Fig. 1 QPACE node card

Fig. 2 A QPACE rack during assembly in the IBM Böblingen lab. Visible are 4 units with 32 node cards each. The vertical copper pipes on the left side are part of the cooling system

Table 1 Key parameters of a QPACE rack

Peak performance (double precision)	26 TFlops
Rack size (w × d × h)	$80 \times 120 \times 250$ cm^3
Weight	$\approx$ 1,000 kg
Maximum power consumption	$<$ 35 kWatts
Typical power consumption	21–27 kWatts

4 Power reduction

Several strategies have been employed to reduce the power consumption of the QPACE architecture:

- Wherever possible, power-efficient components have been used, and the number of power-consuming components has been reduced to a minimum.
- Some relevant voltages have been tuned to a minimal value.
- A water cooling system has been developed.

Most of the power of the system is consumed by the CPU. Here we use the most efficient processor in terms of power consumption per floating-point processing power available at project start. Systems based on the IBM PowerXCell 8i processor are dominating the Green500 List since June 2008 [2]. But also for other components we were able to identify power-efficient solutions, e.g., power supplies with an efficiency $\geq 89\%$.

An example for the reduction of power-consuming components are the memories. The amount of memory was chosen such that the machine can hold applications of the anticipated problem size and that the processor's memory interface bandwidth can be saturated. Like for many other HPC applications also for lattice QCD the memory bandwidth is typically a performance bottleneck.

For the QPACE node cards a further reduction of power consumption could be achieved by voltage tuning. The largest effect is obtained from reducing the processor's core voltage. This voltage is tuned individually for each node card. The automated tuning process is implemented in the service processor following this procedure:

1. The service processor starts the PowerXCell 8i processor with a certain (reduced) voltage setting for its core and array voltages.
2. On the processor a synthetic benchmark application is started which stresses all relevant functional units of the processor.
3. When execution terminated successfully the service processor will reduce the voltage setting and go back to step 1.
4. The service processor stores the minimal voltage for which this particular processor still operates correctly into the VPD (vital product data) memory.

During a normal boot operation the service processor uses the voltage settings from the VPD and adds a guard band of 119 mV. Typically the tuning algorithm is executed only once during initial node card testing. But the procedure can be repeated at any moment during the lifetime of the node card.

Some additional power reduction is obtained by optimizing the memory voltage margins. As a result we could reduce also this voltage (by a fixed amount) while staying above the minimal voltage for which the memories are specified.

5 Cooling system

In the environments in which we are operating the machines, water cooling is beneficial in terms of power consumption, although the gain is difficult to quantify. First, the number of power-consuming fans inside the QPACE rack is reduced to a minimum needed to air-cool the Ethernet switches and the power supplies. Second, the amount of cold water required to cool the machine room is reduced because it is more efficiently used, thus reducing the overall system power budget.

For QPACE a new liquid-cooling concept has been developed and successfully realized. The node cards are packed into thermal boxes made of aluminium. These boxes act as a single big heat sink conducting the heat from the internal components to the surface. The thermal box is then connected to a cold plate as shown in Fig. 3. Water is pumped through channels inside the cold plate, thus moving the heat out of the system.

Our design addresses some of the typical disadvantages of liquid cooling, as there is no water flowing through the thermal box and a so-called dry connection is used to connect the node card to the cooling system. Therefore, after installation of the machine the water circuit does not have to be opened for almost any of the expected maintenance operations. This eliminates the need for expensive, self-healing valves often used in other systems.

There are two critical thermal interfaces which have to be kept under control. The first interface is between the electronic components on the node card and the thermal box. On the inside, the thermal box had to be carefully designed according to the height of the components and the amount of heat they generate. Three different types of thermal grease are used to optimize the thermal contact between these components and the aluminium, depending on component heat, gap sizes and mechanical tolerances. The second important thermal interface is between the thermal box and the cold plate. The heat has to be conducted through a rather small surface of about 40 cm^2. In order to avoid thermal grease, which would be rather inconvenient during maintenance, a suitable type of silicon oil has been selected to improve the

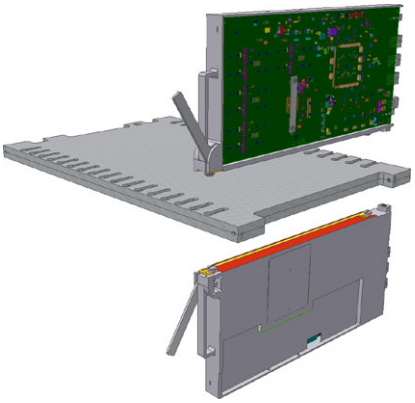

Fig. 3 (Color online) The cold plate together with node cards, one about to be attached from above, the other from below. The *red strip* indicates the thermal interface between node card and cold plate

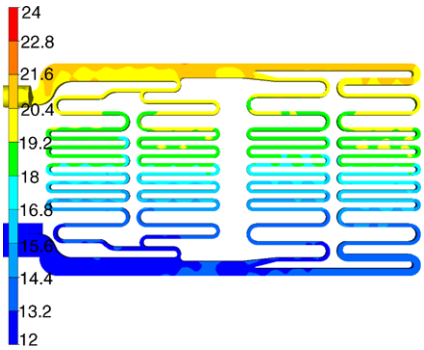

Fig. 4 Result of a simulation of the temperature distribution inside the water-conducting channels of the cold plate

thermal contact. Springs mounted in the thermal box additionally improve contact by pressing the node card onto the cold plate after the node card has been mounted.

The cold plate has been carefully designed in order to make sure that all 32 node cards connected to one cold plate are equally well cooled. In Fig. 4 we show the results for the temperature distribution within the cold plate from a simulation. The water is first distributed through a broad channel on one side of the cold plate (in this picture the bottom side). When the water meanders through the small channels it passes from one end of the node cards to the other end, slowly heating up. It is important to notice that there is only a temperature gradient from bottom to top, whereas from left to right the temperature is constant, i.e., all node cards are cooled equally well. The results of this thermal simulation have been confirmed by temperature measurements using the final system.

All cold plates are connected to a closed water circuit, the so-called secondary circuit. Heat exchangers separate this circuit from the building's primary circuit. We use two 75 kW Knürr CoolTrans systems, which are connected in parallel, to cool an installation of 4 QPACE racks (see Fig. 5).

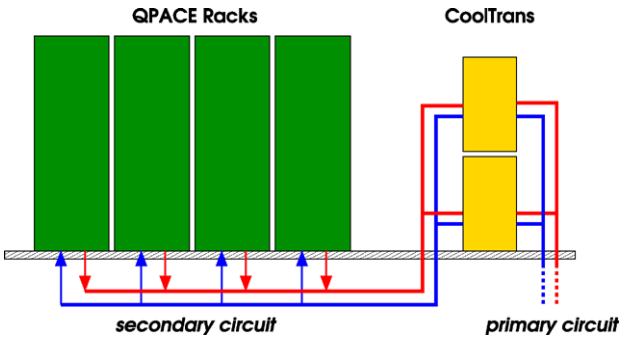

Fig. 5 Schematic view of the cooling circuit of a 4 rack QPACE installation with 2 CoolTrans heat exchanger modules to control flow and temperature within the secondary cooling circuit

How well the thermal interfaces designed and implemented for QPACE work can be quantified by considering the difference ΔT between the temperature of the water inside the cold plate and the temperature at the processor. Even at maximum load we measured $\Delta T \lesssim 30$–$40°\mathrm{C}$. The processor is specified for temperatures up to 95°C.[1] This means that using an inlet water temperature >30°C is a feasible option. We have tested our systems using a water temperature in the secondary circuit of up to 35°C while running a synthetic benchmark (described in the next section) which maximizes the node cards' power consumption. Typically the machines are operated using a temperature of about 18°C.

6 Power consumption

Within a full QPACE installation by far the biggest share of the power is consumed by the node cards. On the node card the processor is the largest consumer. Here the power consumption strongly depends on the utilization of the processor. As we will see later the amount of consumed power significantly changes when the Synergistic Processing Elements (SPU) are used.

The second largest share of power consumption is due to the (in-)efficiencies of the power supplies. The manufacturer of the power supplies used in QPACE specifies the typical efficiency at full load to be $\geq 89\%$.[2] The power consumption of other parts of the system is significantly smaller. The external heat exchangers consume at most 1.8 kW each, which corresponds to a maximum additional power consumption of 0.9 kW per QPACE rack or up to 3–4% of the typical power consumption (see Table 1). The Ethernet switches mounted in each of the QPACE racks consume up to 0.5 kW, i.e. up to 2% of the typical power consumption.

[1] The processor can even sustain higher temperatures without being damaged, but with no guarantee for correct functional behavior.

[2] This is consistent with the ratio between output and input power figures which we read from the power supplies during operation.

Table 2 Power consumption of a set of 32 node cards connected to one backplane with and without processor core voltage tuning enabled (as measured on PSU output)

Load	Default voltage	VMIN enabled
Linux	2.34 kW	2.18 kW
Application kernel	2.69 kW	2.35 kW
PowerLoad SPU	4.29 kW	3.96 kW

In Table 2 we show the power consumption of 32 node cards connected to one backplane. We compare 3 different loads:

- **Linux:** Node cards are booted to Linux but no user processes are running.
- **Application kernel:** Nodes execute a lattice QCD specific application kernel, a parallelized solver for Wilson-Clover fermions. The sustained performance of this application is about 20–25% of peak (depending on algorithmic parameters).
- **Powerload SPU:** This synthetic benchmark maximizes the processors power consumption.

For the (typical) application considered here we have a power consumption of about 80 W per node card. With synthetic benchmarks the power consumption can go up to 120–130 W per node.

In this table we also compare the measured power consumption with and without processor core voltage reduction enabled. For all three loads we see an improvement on the order of 10%.

To compare the power efficiency of the QPACE architecture with other architectures the HPL benchmark has been ported to QPACE [3]. This project took advantage of the fact that the Network Processor is reconfigurable. The machine can thus still be optimized for other applications. In this particular case an additional DMA engine has been implemented to improve main memory to main memory communication controlled by the PowerPC Processor Element. Based on this setup it has been possible to achieve a performance of 773.4 MFlops/W (double precision). This put QPACE in the number one spot on the Green500 List [2]. On our architecture the HPL benchmark runs almost 60 % more power efficiently than on the Chinese Nebulae computer, an Intel X5650 based system using NVidia Tesla C2050 GPUs as accelerators. Nebulae is the second-best system on the June 2010 Green500 List.

For our target applications the sustained performance is significantly less. The application kernel considered here runs at 40–50 GFlops per node (single precision), i.e., with voltage reduction enabled we have a performance of about 544–681 MFlops/W (single precision). It may be interesting to compare this with GPU-based systems, which are currently considered to be particularly power efficient. For a very similar application kernel M. Clark and collaborators obtain a sustained performance of 116.1 GFlops (single precision) on a single GeForce GTX 280 [4]. Let us assume that a host system with a single GPU consumes 250–300 W. In this case we get a performance of about 400–450 MFlops/W.

7 Summary and conclusions

QPACE is an application optimized, massively parallel computer which is highly optimized with respect to power consumption. Several strategies have been applied to achieve this result. A leading position on the Green500 List proves the success of this strategy. A comparison with GPU-based systems also demonstrates the outstanding power efficiency of the QPACE architecture.

Our experience in this project shows that power efficiency has to be optimized in many places in order to achieve a significant overall improvement. This includes the selection of components using power consumption as one criterion, the stream-lining of the architecture, the optimization of the node design and the development of efficient cooling techniques.

In the summer of 2009 eight QPACE racks with an aggregate peak performance of 200 TFlops (double precision) have been deployed and are now used for scientific applications.

Acknowledgements We acknowledge the funding of the QPACE project provided by the Deutsche Forschungsgemeinschaft (DFG) in the framework of SFB/TR-55 and by IBM. We furthermore thank the following companies who contributed significantly to the project in financial and/or technical terms: Axe Motors (Italy), Eurotech (Italy), Knürr (Germany), Xilinx (USA) and Zollner (Germany).

References

1. Baier H et al (2009) QPACE: a QCD parallel computer based on cell processors. arXiv:0911.2174 [hep-lat]
2. http://www.green500.org (accessed on 01 July 2010)
3. Boettiger H, Krill B, Rinke S (2010) QPACE: energy-efficient high performance computing. In: 23th international conference on architecture of computing systems 2010 (ARCS '10) workshop proceedings
4. Clark MA, Babich R, Barros K, Brower RC, Rebbi C (2009) Solving lattice QCD systems of equations using mixed precision solvers on GPUs. arXiv:0911.3191 [hep-lat]

H. Baier was a senior engineer at the IBM Lab in Böblingen.

H. Boettiger is a researcher at the IBM Lab in Böblingen.

M. Drochner is a research associate at the Research Center Jülich, Germany.

N. Eicker is a research associate at the Research Center Jülich, Germany.

U. Fischer is a manager at the IBM Lab in Böblingen.

Z. Fodor is a professor of physics at the University of Wuppertal, Germany.

A. Frommer is a professor of mathematics at the University of Wuppertal, Germany.

C. Gomez is a senior engineer at the IBM Lab in La Gaude, France.

G. Goldrian is a Distinguished Engineer in the IBM Lab in Böblingen, Germany.

S. Heybrock is a PhD student in physics at the University of Regensburg, Germany.

D. Hierl a research associate at the University of Regensburg, Germany.

M. Hüsken is a research associate at the University of Wuppertal, Germany.

T. Huth is a researcher at the IBM Lab in Böblingen. Germany.

B. Krill is a research engineer at the IBM Lab in Böblingen. Germany.

J. Lauritsen is a researcher at the IBM Lab in Böblingen.

T. Lippert is the director of the Institute for Advanced Simulation at the Research Centre Jülich, head of the Jülich Supercomputing Centre, and a professor of physics at the University of Wuppertal, Germany. Germany.

T. Maurer is a PhD student in physics at the University of Regensburg, Germany.

B. Mendl is a PhD student in physics at the University of Regensburg, Germany.

N. Meyer is a PhD student in physics at the University of Regensburg, Germany.

A. Nobile is a research associate at the University of Regensburg, Germany.

I. Ouda is a senior engineer at the IBM Lab in Rochester, Minnesota.

M. Pivanti is a PhD student in computer science at the University of Ferrara, Italy.

D. Pleiter is a senior scientist at Deutsches Elektronen Synchrotron (DESY), Zeuthen, Germany.

M. Ries is a research engineer at the IBM Lab in Böblingen.

A. Schäfer is a professor of physics at the University of Regensburg, Germany.

H. Schick is a researcher at the IBM Lab in Böblingen.

F. Schifano is a senior scientist in computer science at the University of Ferrara, Italy.

H. Simma is a senior scientist at Deutsches Elektronen Synchrotron (DESY), Zeuthen, Germany.

S. Solbrig is a research associate at the University of Regensburg, Germany.

T. Streuer is a research associate at the University of Regensburg, Germany.

K.-H. Sulanke is an electronics engineer at Deutsches Elektronen Synchrotron (DESY), Zeuthen, Germany.

R. Tripiccione is a professor of physics at the University of Ferrara, Italy.

J.-S. Vogt is a research engineer at the IBM Lab in Böblingen.

T. Wettig is a professor of physics at the University of Regensburg, Germany.

F. Winter is a PhD student in physics at the University of Regensburg, Germany.