Paving the Road towards Pre-Exascale Supercomputing

Dirk Brömmel, Ulrich Detert, Stephan Graf, Thomas Lippert, Boris Orth, Dirk Pleiter, Michael Stephan, and Estela Suarez

Institute for Advanced Simulation, Jülich Supercomputing Centre,
Forschungszentrum Jülich, 52425 Jülich, Germany
E-mail: th.lippert@fz-juelich.de

Supercomputing at scale has become the decisive challenge for users, providers and vendors of leading supercomputer systems. On next-generation systems, approaching exascale by the end of the decade, we will be confronted with millions of cores, and the need of massive parallelism. Beyond aggregating ever larger compute performance also the ability to hold and efficiently process drastically increasing amounts of data will be key to enable future leading research facilities for computational science. We report in this article on the evolving supercomputing infrastructure at Jülich Supercomputing Centre (JSC), research and development activities on future HPC technologies and architectures as well as on the computational science research and collaboration with science areas which will require exascale supercomputing in the future.

1 Introduction

In 2005 the Jülich Supercomputing Centre (JSC) started it's dual system strategy to most efficiently serve the application portfolio of the users of the Jülich Research Centre, the John von Neumann Institute for Computing in Germany and since mid of 2010 the Partnership for Advanced Computing in Europe (PRACE). Via the German Gauß Centre for Supercomputing in 2009 a first milestone was reached with the installation of the IBM Blue Gene/P system named JUGENE as highly scaling system (294,912 cores) and JUROPA (25,000 Intel Nehalem CPU cores) as highly flexible, general-purpose cluster system.

This dual system strategy has been carried forward end of 2012 with the installation of a new highly scaling, 28 rack IBM Blue Gene/Q system named JUQUEEN entering the TOP500 list at rank 7 world wide and as #1 in Europe (see Sec. 2). With it's 458,752 cores more than 1.5 million hardware threads can be executed concurrently by a single application. Several applications have been proven to scale to this extent and are now members in the "JSC High-Q Club" (see Sec. 2.1). The compute system is flanked by a new GPFS storage system providing for the first time full end-to-end data integrity and a maximum I/O bandwidth of 200 GByte/s. To enable future architectures where non-volatile storage devices are integrated into the supercomputer to provide even higher bandwidth and significantly higher access rates, JSC collaborated with IBM on an active storage subsystem attached to JUQUEEN (see Sec. 3).

To prepare the next step of replacing the general-purpose system JUROPA by a system approach of 2 PFlop/s peak performance, a new test cluster called JUROPA-3 has been installed. The work on future architectures where such clusters are coupled to a booster comprising tightly coupled many-core devices is continued in the new DEEP-ER project (see Sec. 5) which extends the ongoing Dynamical Exascale Entry Platform Extended Reach (DEEP) project. A new exascale lab, the NVIDIA Application Lab, focusses on another type of many-core architectures, namely GPUs (see Sec. 6).

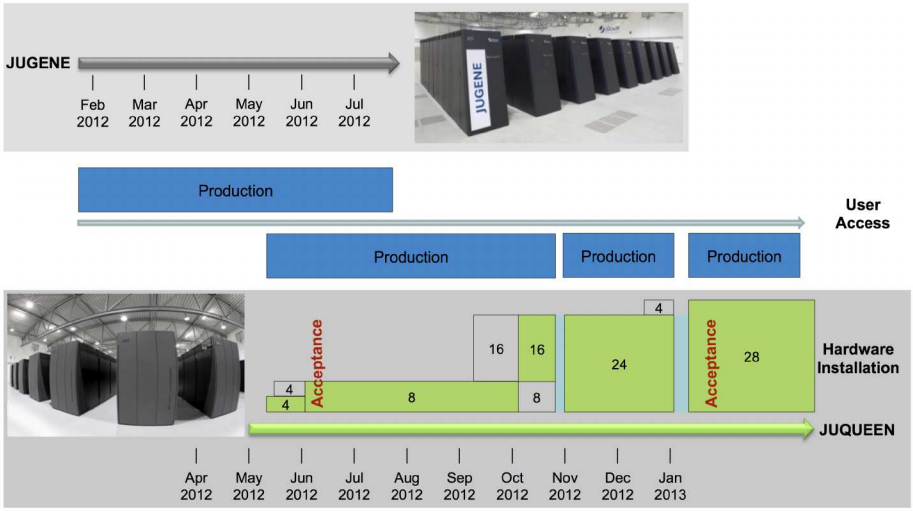

Figure 1. Timeline of the transition from JUGENE to JUQUEEN in 2012/13.

In October 2013 the Human Brain Project (HBP) – one of two large-scale initiatives selected out of six candidates to receive significant funding through the EU's new Future and Emerging Technologies (FET) Flagship Programme – started a 2.5-year ramp-up phase. The task of the JSC in this project will be to develop, deploy and operate the main HBP supercomputer in Jülich. Sec. 7 gives an overview of the project structure and its goals.

2 JUQUEEN

When IBM first announced Blue Gene/Q as the third generation of their Blue Gene supercomputing family in 2011, JSC decided to enter a collaboration with IBM which resulted in the installation of a prototype system. This allowed exploring the new architecture including its scalability before the development of the system was completed[1]. The first racks of a Blue Gene/Q production system arrived at Jülich in April 2012. Only a few weeks later the new system, called JUQUEEN, consisting of 8 racks could be made accessible to users for early production runs delivering a peak performance of 1.6 PFlop/s, 60 % more than JUGENE. The first 8 racks were fulfilled through the Helmholtz programme "Supercomputing".

JUGENE was an at that time still operational 72 racks Blue Gene/P system with 1 PFlop/s peak performance. The system had been very reliable and stable over several years such that a user job utilisation of over 90 % could be achieved. The system was nevertheless shut down at the end of July 2012 and dismounted to free space for the final JUQUEEN system consisting of 28 racks. To minimise the impact on the users and to limit the downtime to a minimum, the transition (see Fig. 1) took place in several steps, starting with 8 racks, then switching to a 16 rack system, 24 racks, and finally 28 racks in February 2013. After the system went into production the job utilisation immediately reached 70 % and is now at a level >90 %. The additional 20 racks were fulfilled through the project

"Peta GCS" of the German Gauß Centre for Supercomputing.

This translates into a more than five-fold increase of compute performance on Blue Gene available for scientific computing in Jülich, Germany and Europe. A large fraction of it is now available for users of the Gauss Centre for Supercomputing and PRACE.

2.1 The High-Q Club

As scaling up processing pipeline performance by increasing the clock frequency has reached its limits, performance boosts can only be achieved by scaling-out and increasing parallelism at different levels. From the Blue Gene/P to Blue Gene/Q the width of the SIMD vector instruction pipelines doubled and the number of threads per core as well as the number of cores per node increased by a factor 4.

To help our users in migrating their codes and leveraging the performance on the new architecture, we organised a first porting and tuning workshop in Spring 2013[2]. Following this workshop and to promote the idea of exascale capability computing, we have established the High-Q Club[3], a showcase for codes able to utilise the entire 28-rack Blue Gene/Q machine at JSC. The club members comprise a collection of the highest scaling codes on JUQUEEN, through which we intend to encourage other developers to invest in tuning and scaling their codes. We want our users to show that they are capable of using all 458,752 cores, and for example more than 1 million concurrent threads on JUQUEEN.

The diversity of members of the High-Q Club establishes that it is possible to scale real applications to the complete JUQUEEN using a variety of programming languages and parallelisation models, demonstrating individual approaches to reach that goal. High-Q status thus marks an important milestone in application development towards future HPC systems that envisage even higher core counts.

To qualify for membership, developers should submit evidence of the scalability of their codes across all available cores. While we currently do not set a strict minimum efficiency, we do expect the codes to profit from additional cores with an increase in speed. The benchmark used should also be as close as possible to a production scenario: trivial kernels or libraries will not be accepted.

At the time of writing this article, the members of the High-Q Club were:

dynQCD
dynQCD is a code for simulations in the field of Lattice Quantum Chromodynamics and can be used for different fermion actions. The code is developed at the University of Wuppertal and the Simulation Laboratory for Nuclear and Particle Physics at JSC and is written in C. The code features hybrid-parallelisation using POSIX threads. Inter-process communication on Blue Gene/Q is done via the low-level System's Programming Interface (SPI) by-passing higher-level communication libraries like MPI.

Gysela
Gysela is a **GY**rokinetic **SE**mi-**LA**grangian code for plasma turbulence simulations developed at CEA Cadarache. It is, e.g., used in simulations of the electrostatic branch of the ion temperature gradient turbulence in tokamak plasmas. Gysela is written in Fortran90 and C and uses MPI, OpenMP and POSIX threads for parallelisation.

JuSPIC
JuSPIC[4] is the Jülich Scalable Particle-in-Cell code for fully relativistic plasma simulations in electromagnetic fields. The non-linear interaction between the fields and the plasma is described by the relativistic Vlasov equation and Maxwell's equations. JuSPIC is developed by the Simulation Laboratory for Plasma Physics at JSC in Fortran using MPI and OpenMP. Since it is also used to test new architectures and programming models, there also is a version that uses the StarSs model.

MP2C
The Massively Parallel Multi-Particle Collision Dynamics code (MP2C)[5] implements a hybrid representation of solvated particles in a fluid to simulate soft matter physics and mesoscopic hydrodynamics. It is developed by the Simulation Laboratory for Molecular Systems at JSC and is written in Fortran. The parallelisation is based on MPI while scalable I/O on the full JUQUEEN is achieved by using SIONlib[6].

$\mu\varphi$ (muPhi)
$\mu\varphi$ (muPhi)[7] combines two packages to model and simulate water flow and solute transport in porous media. It can be used for the prediction and control of groundwater production, the assessment of water contamination and becomes more and more important for flood and climate prediction. Among other parts, the code make use the iterative solver template library (ISTL) developed at the University of Heidelberg within framework of the DUNE project. $\mu\varphi$ is written in C++ using MPI for parallelisation. Like MP2C it relies on SIONlib for scalable parallel I/O.

NEST
NEST[8] is a simulator for spiking neural network models that focus on the dynamics, size and structure of neural systems rather than on the exact morphology of individual neurons. It includes over 25 neuron and 10 synapse models and uses a hybrid parallelisation scheme to perform the computations of the neuron and synapse dynamics. The code is a development by the NEST initiative in C++, combining MPI and OpenMP.

PEPC
The **P**retty **E**fficient **P**arallel **C**oulomb[9] solver is used for N-body simulations and was developed within the Simulation Laboratory for Plasma Physics at JSC. PEPC is not restricted to a specific force law or physical problem and can thus be applied to different problems, e.g., beam-plasma interaction, vortex dynamics, gravitational interaction or molecular dynamics simulations. The code is written in Fortran 2003 and C and is parallelised using MPI as well as OpenMP or POSIX threads.

PMG+PFASST
PMG+PFASST[10] combines a parallel multigrid solver with a time parallel approximation scheme to solve ODEs with linear stiff terms. The two parts have been developed at the Lawrence Berkeley National Lab (PFASST) and the University of Wuppertal (PMG) and have been coupled to one application by developers from the cross-sectional team Mathematical Methods and Algorithms at JSC and Universita della Svizzera Italiana.

PMG+PFASST is written in Fortran 2003 and C. Parallelisation is implemented using MPI and POSIX threads.

Terra-Neo
Terra-Neo[11] is used for modelling earth mantle dynamics. The development team is built from members of Ludwig-Maximilians-Universität München, Universität Erlangen-Nürnberg, Regionales Rechenzentrum Erlangen and Technische Universität München. The interdisciplinary team involves geophysics and algorithmic experts as well as software and hardware experts to enable the high performance and scalability on future HPC architectures. The Terra-Neo framework is implemented in C++ and Fortran using MPI and OpenMP for parallelisation.

waLBerla
waLBerla[12] (widely applicable Lattice Boltzmann solver from Erlangen) is a computational fluid dynamics application. Originally, the waLBerla framework has been centred around the Lattice Boltzmann method for the simulation of fluid scenarios but in the meantime evolved to a code that is also suitable for a wide range of applications based on structured grids. It is developed in C++ and uses MPI and OpenMP. Accelerator devices are supported using CUDA and OpenCL.

2.2 GPFS Storage Cluster – JUST

For the storage back-end it is hard to keep pace with the performance increase of the compute systems. In parallel to the upgrade of the Blue Gene/P system JUGENE to the Blue Gene/Q system JUQUEEN the storage cluster JUST3 has been replaced by a new storage system JUST4. The main goal was to achieve a significant bandwidth improvement from about 60 GByte/s to approximately 200 GByte/s. This could only be achieved by a significant increase of the number of disks which was not straightforward. With both the number of disks and the capacity per disk increasing standard RAID technology would not have provided sufficient protection against failures resulting in data loss. With the old storage systems we observed a disk failure rate of 2-3 disks per week and rebuild times in the order of 12 hours during which performance could degrade significantly. For this reason, JSC was the first to install IBM's GPFS Storage Servers (GSS).

Instead of RAID controllers GSS uses GPFS Native RAID (GNR), which is a software implementation of storage RAID technologies within GPFS. One key feature of GNR is the Declustered RAID concept where user data is stored in strips which are grouped in RAID arrays. The strips are distributed over multiple disks such that even if multiple disks breaks no data loss occurs. GNR supports 2- and 3-fault-tolerant Reed-Solomon codes and 3-way and 4-way replication. Whenever any of the RAID arrays becomes critical, i.e. when no further disk failure could be tolerated, GPFS will maximise rebuild priority accepting performance degradations. Once this critical state is left, rebuild priorities are lowered to improve performance. Tests have shown that the critical phase of the rebuild lasts only for a couple of minutes in contrast to 12 hours for classical RAID controllers.

Another key feature of GNR is the support of end-to-end checksums. These checksums are created and verified at the client which is writing and reading the data, respectively, and sent over the network together with the data. This significantly reduces the risk of silent

errors, i.e. errors which are not detected by the storage system itself.

Since September 2013 the GSS storage cluster JUST4 is in production and used to host two file systems, one scratch file system with 3.2 PBytes and one data file system with 1.9 PBytes dedicated to special, data intense projects.

3 Blue Gene Active Storage

Typical HPC architectures continue to move away from what is called Amdahl's rule for a balanced I/O performance: one bit of I/O per second for each instruction per second. This growing performance gap is becoming more critical as there is a growing number of applications processing big data volumes. Higher performance in terms of bandwidth could be achieved by scaling-out to an even larger number of spinning disks. This strategy is not only limited by cost but also by power and would not be affordable in an exascale context. Current projections aim for an I/O bandwidth of up to 60 TBytes/s[13].

There are several opportunities which will allow to mitigate and overcome this problem. Firstly, new non-volatile memory technologies while limited in capacity (within an affordable budget) allow to realise much higher bandwidth plus dramatically higher access rates. The latter is an important feature for a large set of data-intensive applications. The advantages of high-capacity spinning disk storage systems and high-performance non-volatile memory technologies can be combined in a hierarchical storage architecture as it was explored by a PRACE prototype at JSC[14]. Secondly, active storage concepts provide an opportunity to reduce performance and energy costs of data movement. Active storage is an architectural concept where processing capabilities and storage are integrated.

Such a concept has been realised for JUQUEEN by Blue Gene Active Storage (BGAS)[15]. A BGAS node comprises a standard Blue Gene/Q I/O node as well as a PCIe card which integrates 2 TBytes of storage plus network ports towards external, large-capacity storage systems. Each of these nodes is connected by 2 network links to the Blue Gene/Q compute nodes, 6 network links to the neighbouring BGAS nodes within a 3-dimensional torus network and 2 10-GbE links to the external large capacity storage cluster JUST. The high bi-sectional bandwidth provided by this torus network enables fast movement of the data within the active storage system.

The various ways on how applications can efficiently use such a system are still to be explored. The architecture can, e.g., be used for post-processing data generated during large-scale simulations. In this case, the high bandwidth is exploited to write out huge amounts of data which is post-processed within the active storage system. Only a resulting, significantly reduced set of data is finally written to the external storage. Another option is to use such a system for multi-pass analysis, where massively parallel applications perform a large number of random read accesses to data sets loaded into the active storage system. Yet another option is the use of the fast storage to temporarily hold data which does not fit into the compute nodes main memory.

The BGAS system attached to JUQUEEN is the result of a cooperation with IBM in the framework of the Exascale Innovation Centre (EIC).

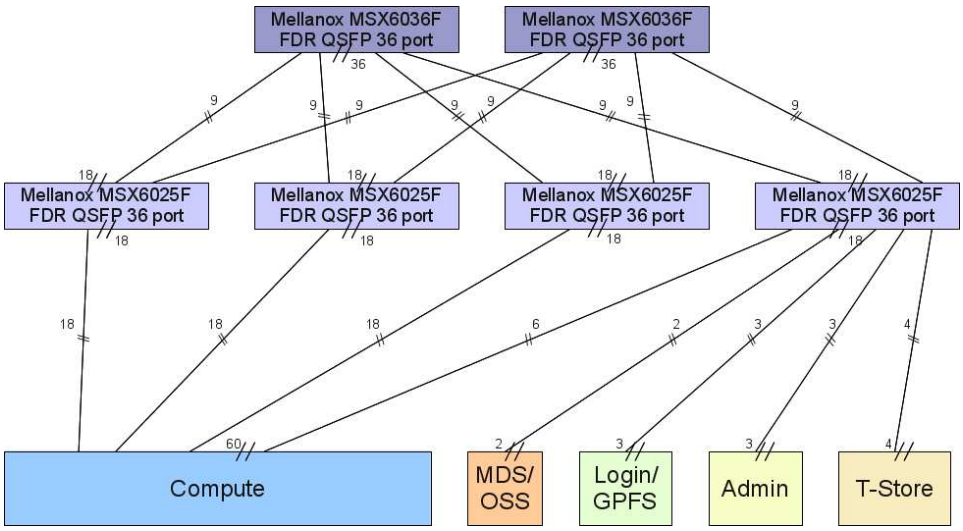

Figure 2. Fat-tree topology of FDR Infiniband fabric.

4 JUROPA-3

In order to prepare for a future replacement of the JUROPA HPC production system, a prototype system called JUROPA-3 has been installed at JSC in April 2013. JUROPA-3 is the outcome of a cooperation of JSC, ParTec Cluster Competence Center and T-Platforms, aiming at the development of solutions for fundamental questions in large-scale cluster computing. Topics like application check-point/restart, end-to-end data integrity, network topology, failure prediction and energy efficiency are addressed in this cooperation.

JUROPA-3 comprises 60 compute nodes, each equipped with 2 Intel Xeon E5-2650 CPUs (Sandy Bridge-EP) providing a total of 960 processor cores with a peak performance of 15.3 TFlops. In addition, 4 compute nodes are each enhanced with 2 NVIDIA Tesla K20X GPUs and another 4 nodes with 2 Intel Xeon Phi 5110P co-processors each. The co-processors amount to an extra performance of 18.5 TFlops peak. The compute nodes, the server nodes and the storage components of the cluster are connected by a 56 Gbit/s Infiniband network (FDR) with fat-tree topology providing high communication bandwidth for parallel applications and I/O (see Fig. 2).

JUROPA-3 uses the same hierarchical system structure with master nodes, administrative nodes and compute nodes like the current JUROPA production cluster. Several modifications and enhancements in JUROPA-3, however, are subject to further development and testing:

- Scientific Linux is used as the basic operating system combined with ParTec's ParaStation software for cluster management.
- The majority of compute nodes run in diskless operation mode.
- 8 compute nodes include additional local disks for checkpoint/restart development.

- 16 fat nodes possess an increased amount of main memory (256 GB and 128 GB, respectively) which allow for production runs of structural mechanics applications.

- A small storage area (2 server nodes, 60 TB disk space) is dedicated to the development and testing of future Lustre parallel file system versions and end-to-end data integrity enhancements.

- A GPFS gateway node with Infiniband HCA on one side and 4x10 GbE on the other side provides for connectivity between JUROPA-3 and JSC's file server JUST4. GPFS is mounted in addition to Lustre on all cluster nodes and serves as the main file system for home and scratch data.

- The SLURM resource manager will be integrated with ParaStation and be used for job management and control.

5 DEEP – Extended Reach (DEEP-ER)

Multiple challenges have to be overcome to make exascale computing possible by the end of the decade. Simply scaling up today's HPC concepts and technology will not be sufficient: the overall power consumption must be drastically reduced; applications will have to be modified to scale up and extract performance from systems with millions of cores; resiliency methods must be developed to deal with the reduced mean time to failure (MTTF); and additional layers in the memory hierarchy are needed to reduce the increasing gap between the growing compute performance and the limited bandwidth of both memory and storage.

Two of the above mentioned exascale challenges (concurrency and power consumption) are addressed in the DEEP project[16,17]. The novel DEEP architecture combines the high scalability of dedicated massively parallel systems with the ubiquity and cost effectiveness of commercial off-the-shelf HPC clusters. Its straightforward, scalable programming model allows applications to run their code parts with different scalability characteristics on the part of the system best suited to them. Furthermore, DEEP's direct warm water-cooling concept and the use of many-core processors reduces drastically the overall energy consumption.

The DEEP architecture will be substantially extended with resiliency, memory hierarchies and I/O functionalities in the new project "DEEP – Extended Reach" (DEEP-ER)[18]. The improvement in performance and power efficiency when using emerging memory technologies such as non-volatile memory (NVM) and memory attached to the network (NAM) will be explored. A prototype will be built and scientific applications will be ported to demonstrate the achieved improvements in scalability, parallel I/O efficiency, and system reliability.

In the DEEP-ER prototype (see Fig. 3) the computing power at the Booster Nodes (BN) will be provided by the second generation of Intel Xeon Phi. At the BNs, NVM devices represent one further level of memory hierarchy additional to the existent processor's main memory. NVM allows for much larger memory sizes than DRAM and can be globally accessible from any part of the DEEP-ER prototype. A uniform high-speed interconnect will run across Cluster and Booster. NAM nodes attached to it will provide access to a

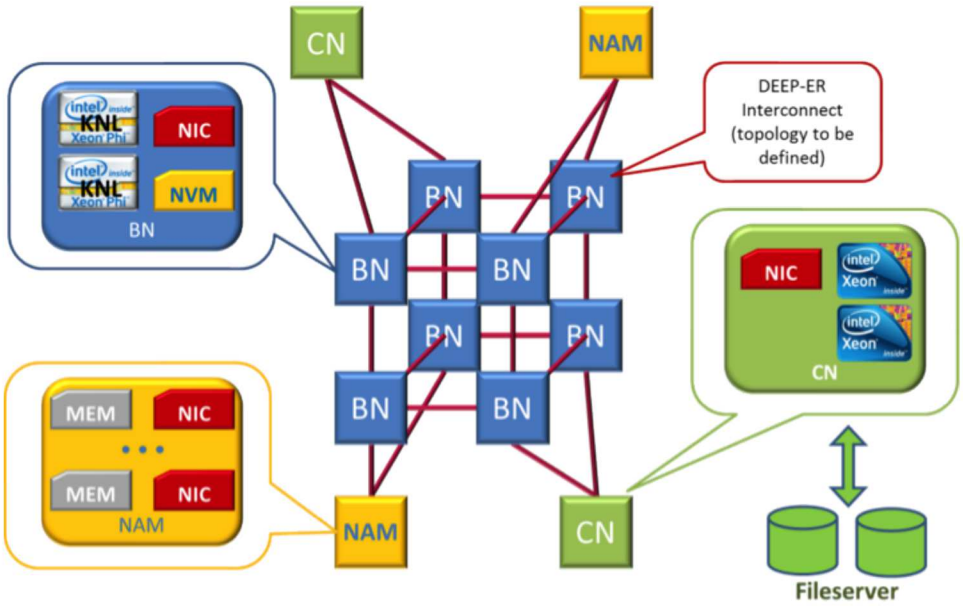

Figure 3. Hardware architecture of the DEEP-ER prototype.

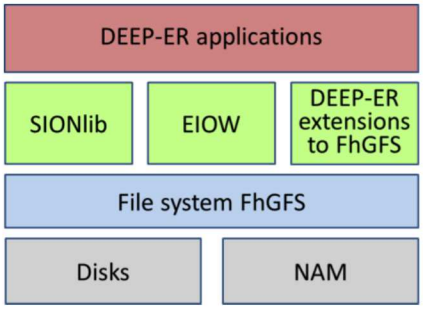

Figure 4. DEEP-ER I/O architecture.

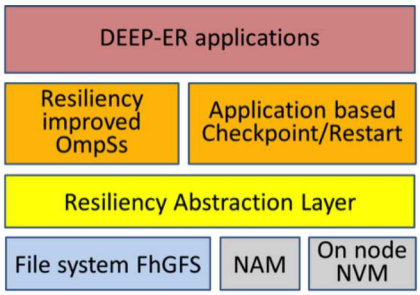

Figure 5. DEEP-ER Resiliency architecture.

large shared memory pool for all Cluster and Booster nodes. The memory will be part of the node address spaces and support fine-grain access.

The simplification of the hardware architecture and the newly introduced components open the door for new functionality in the software environment, in particular regarding I/O and resiliency (see Fig. 4 and 5).

5.1 I/O

For highly efficient access to the I/O devices, the Fraunhofer parallel file system (FhGFS)[19] will be extended and optimised, to allow for system-wide access to fast storage class mem-

ory devices, such as NVM or NAM. FhGFS will use these devices to build independent units of cache groups for subsets of nodes, allowing the nodes in one group to perform certain I/O operations without being influenced by ongoing I/O operations in other groups.

Additionally, efficient access to the storage subsystem at different levels of abstraction will be provided by APIs originating from FhGFS itself, the parallel I/O library SIONlib[6], and the software developed by Exascale I/O Workgroup EIOW[20], respectively. DEEP-ER will investigate their respective performance in comparative benchmarks.

Finally, to provide application developers convenient means to identify potential bottlenecks in the application I/O path, new storage management extensions will be integrated with the corresponding cluster management tools. New monitoring extensions (such as detailed live statistics or profiling capabilities) will be also added to the system. In this way the full potential of the DEEP-ER concept will be leveraged to match the storage access performance with the available compute power.

5.2 Resiliency

Employing the capabilities provided by the described I/O architecture and the characteristics of DEEP's programming model, a dual-approach resiliency concept (see Fig. 5) will be developed. It combines an application-based multi-level checkpoint/restart mechanism with a less intrusive scheme for task-controlled recovering from component failures.

The first approach will provide a common framework to store, identify, validate and reload checkpoints from multiple levels of the memory hierarchy. The performance of local NVM memory will be exploited for frequent (cheap/local) checkpoints and NAM and permanent parallel I/O storage for less frequent, (expensive/global) ones. A common resiliency abstraction layer will implement all common mechanisms required to write, locate, test and read checkpoints to the multi-levels of the checkpoint/restart subsystem, allowing for a transparent use of its functionality.

The second approach is based on OmpSs[21]. The OmpSs task-based programming model decomposes an application into stateless tasks arranged in a directed acyclic graph (DAG) representing the dependencies between these tasks. The OmpSs runtime system can detect a single task or a set of tasks failing because of a system fault, and it can then transparently re-execute these tasks from their respective beginnings. OmpSs' concept of stateless tasks was already extended in DEEP by identifying highly scalable code parts to be offloaded to the Booster as tasks, too. These coarse-grain tasks potentially run on many processors for a considerable amount of time. Re-executing them from the beginning for fault recovery would be a waste of resources.

Therefore, DEEP-ER combines the resiliency enhancements achieved by restarting single tasks with the possibility to write checkpoints within such parallel tasks. In case a coarse-grain task is being restarted by the task-controlled mechanism, checkpoints stored in DEEP-ER's multi-level system can be loaded to recover the most recent task state. Additional OmpSs annotations can be used by the application developer to identify the data required to be written at the checkpoint in a straightforward way. Furthermore, OmpSs might be used to enable a new class of checkpoints on a runtime-system-level, using the DAG to identify a synchronised state of the overall application.

Systems using the DEEP-ER developments will be able to run more applications increasing scientific throughput, and the loss of computational work through system failures

will be substantially reduced. Within the project, seven grand-challenge HPC applications will be optimised demonstrating the usability, performance and resiliency of the DEEP-ER Prototype. They will set the scale for the checkpoint/restart and I/O capabilities developed within the project. The DEEP-ER project started in October 2013 and will last three years. It is led by the Jülich Supercomputing Centre and brings together a total of 14 partners coming from both research and industry, including four PRACE supercomputing centres.

6 NVIDIA Application Lab at Jülich

Exploiting the performance of thousands of simple cores running at a low clock speed has the potential to significantly improve energy efficiency. This is, e.g., the case for applications where dense matrix operations dominate, like in the Linpack benchmark. This helped GPU accelerated systems to reach top positions in the Green500 list[22], the list of the most power-efficient HPC systems. To enable more applications to use GPUs is one of the key goals of the NVIDIA Application Lab in Jülich. This lab has been established in summer 2012. Since then it has built-up a broad application portfolio encompassing computational neuroscience, high-energy physics, radio astronomy, data analytics and others.

Another focus of the Lab is the parallelisation of applications on multiple GPUs. Not only aggregation of more compute performance but also the need for more memory capacity are reasons for using multiple GPUs. For the application developers this means that beyond the significant amount of parallelism at device level, an additional level of parallelism needs to be managed. For an application of the Jülich Institute for Neuroscience and Medicine (INM-1) called JuBrain, several parallelisation strategies have been investigated and tested[23]. As part of an attempt to assemble a realistic, 3-dimensional model of the human brain, images of brain cuts have to be mapped onto each other. This image registration task involves repeated executions of a compute intensive kernel to calculate the mutual information metric. While low-resolution images can be processed using a single GPU, the available device memory is too small to hold high-resolution images.

JuBrain is only one of several data-intensive applications explored within the Lab. It could also be shown that density cluster analysis can be performed very efficiently on recent generations of GPUs[24]. Cluster analysis aims on the identification of regions of similar objects in a multi-dimensional data set. It is a standard method of data analytics which can, e.g., be applied to protein folding simulation data. A significant speed-up of this analysis enables interactive processing of large data sets.

7 Human Brain Project

In January 2013, the European Commission selected the Human Brain Project (HBP) as one of two large-scale initiatives out of six candidates to receive significant funding through the EU's new Future and Emerging Technologies (FET) Flagship Programme, starting October 2013. The 2.5-year ramp-up phase of the project (until March 2016), which is funded by the EU's 7th Framework Programme, will be followed by a – partially overlapping – operational phase under the upcoming next framework programme, Horizon 2020. Federating more than 80 European and international research institutions (with more partners joining the consortium later on through a Competitive Call Programme) under the lead of

Henry Markram from the Swiss Ecole Polytechnique Fédérale de Lausanne (EPFL), the HBP as a whole is planned to last ten years and estimated to cost one billion Euros.

The goal of the HBP is to collect all existing knowledge about the human brain and to reconstruct the brain, piece by piece, in multi-scale models and supercomputer-based simulations of these models. The resulting "virtual human brain" offers the prospect of a fundamentally new understanding of the brain and its diseases and of novel, brain-like computing technologies.

To reach this goal, the HBP will build a European research infrastructure consisting of six ICT (Information & Communication Technology) Platforms, dedicated respectively to Neuroinformatics, Medical Informatics, Brain Simulation, Neuromorphic Computing, Neurorobotics, and High-Performance Computing. Together, these platforms will make it possible to federate neuroscience data from around the world, to integrate the data in unifying models and simulations of the brain, to validate the results against empirical data, and to make them available to the scientific community. The resulting knowledge on the structure and connectivity of the brain will open up new perspectives for the development of "neuromorphic" computing systems incorporating unique characteristics of the brain such as energy-efficiency, fault-tolerance and the ability to learn. The HBP's models and simulations will enable neuroscientists to carry out *in silico* experiments on the virtual human brain that cannot be done *in vivo* for practical or ethical reasons.

Jülich plays a key role in the HBP as it contributes to the project a unique combination of expertise and infrastructure in the two relevant fields neuroscience and supercomputing. The increasingly close cooperation of the two research areas in Jülich becomes manifest in the joint activities of the Institute of Neuroscience and Medicine and the JSC within the Helmholtz Portfolio Theme Supercomputing and Modelling for the Human Brain and, in particular, within the new Simulation Lab Neuroscience. In fact, in the HBP, Jülich leads important subprojects in both neuroscience ("Strategic Human Brain Data") and supercomputing ("The High-Performance Computing Platform"). In addition, Jülich researchers are work package or task leaders in other subprojects (Brain Simulation, Neuroinformatics, Theory).

The task of the HBP's HPC Platform subproject is to build the supercomputing and data hard- and software infrastructure required to run cellular brain model simulations of the size of a full human brain, and to make this infrastructure available to the consortium and the wider community. Central element of the HPC Platform is the HBP Supercomputer, the project's main production system, which will be built in stages to arrive at the exascale capability needed for cellular simulations of the complete human brain towards the end of the decade. It will be the task of the JSC to develop, deploy and operate the HBP Supercomputer in Jülich as the future European HPC Facility for brain research. Jülich will work with HPC industry in the ramp-up phase to arrive at suitable, innovative HPC solutions meeting the specific requirements of the HBP (such as large memory and interactivity), and to lay the technological foundation for the subsequent procurement of a pre-exascale machine in the next phase of the project. The interactive supercomputing capabilities that will be developed for the HBP will be invaluable not only for neuroscience but also for a broad range of other applications in the life sciences and elsewhere (e.g., in civil security research). While the HBP is poised to become the main driver for the future development of high-performance computing at JSC, the breadth of applications involved and their requirements will warrant the usability of the HBP Supercomputer for many other fields, too.

Besides the main HBP Supercomputer at Jülich the HBP's HPC Platform will consist of a smaller software development system at CSCS (Lugano, Switzerland), a system for molecular-level simulations at BSC (Barcelona, Spain), and a system for massive data analytics at CINECA (Bologna, Italy). During the ramp-up phase of the project, the HBP will negotiate with further PRACE Tier-0 institutions that have expressed their interest in adding in-kind support to the Platform. A high priority goal is to establish a PRACE community access programme, also to be negotiated in the ramp-up phase. This would allow access to the Tier-0 capability of the HPC Platform, reviewed by the HBP's International Access Board, via PRACE services.

Acknowledgments

The DEEP, DEEP-ER and Human Brain Project are partially funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement $\mathcal{N}^{\text{o}}$ 287530, 610476 and 284941. The Exascale Innovation Center is supported by the State of Nordrhein-Westfalen.

References

1. S. Alam *et al.*, *Early experiences with scientific applications on the IBM Blue Gene/Q supercomputer*, IBM Journal of Research and Development, Vol. 57, No. 1, 2013.
2. D. Brömmel, *First JUQUEEN Porting and Tuning Workshop*, Innovatives Supercomputing in Deutschland, Vol. 11, No. 1, 2013. `http://www.fz-juelich.de/ias/jsc/events/juqueenpt13`
3. High-Q Club website: `http://www.fz-juelich.de/ias/jsc/high-q-club`
4. JuSPIC website: `http://www.fz-juelich.de/ias/jsc/juspic`
5. J. Freche, W. Frings, G. Sutmann, *High Throughput Parallel-I/O using SIONlib for Mesoscopic Particles Dynamics Simulations on Massively Parallel Computers*, Proc. of Intern. Conf. ParCo, Page 423, 2010.
6. W. Frings, F. Wolf, V. Petkov, *Scalable Massively Parallel I/O to Task-Local File*, Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis (SC2009), Article No. 17, 2009.
7. O. Ippisch, M. Blatt, *Scalability test of $\mu\varphi$ and the parallel algebraic multigrid solver of dune-istl*, Jülich Blue Gene/P Extreme Scaling Workshop (Editors B. Mohr, W. Frings), Technical Report FZJ-JSC-IB-2011-02, April 2011.
8. M.-O. Gewaltig, M. Diesmann, *NEST (NEural Simulation Tool)*, Scholarpedia, Vol. 2, No. 4, Page 1430, 2007.
9. M. Winkel, R. Speck, H. Hübner, L. Arnold, R. Krause, P. Gibbon, *A massively parallel, multi-disciplinary Barnes-Hut tree code for extreme-scale N-body simulations*, Computer Physics Communications, Vol. 183, No. 4, Pages 880–889, 2012.
10. D. Ruprecht, R. Speck, M. Emmett, M. Bolten, R. Krause, *Extreme-scale space-time parallelism*, Poster, accepted for SC'13.
11. B. Gmeiner, H. Köstler, M. Stürmer, U. Rüde, *Parallel multigrid on hierarchical hybrid grids: a performance study on current high performance computing clusters*, Concurrency and Computation: Practice and Experience, 2012.

12. C. Godenschwager, F. Schornbaum, M. Bauer, H. Köstler, U. Rüde, *A Framework for Hybrid Parallel Flow Simulations with a Trillion Cells in Complex Geometries*, Proceedings of SC13: International Conference for High Performance Computing, Networking, Storage and Analysis, Article No. 35, 2013.
13. S. Ashby *et al.*, *The Opportunities and Challenges of Exascale Computing*, Office of Science, US Department of Energy, 2010.
14. S. El Sayed, S. Graf, M. Hennecke, D. Pleiter, G. Schwarz, H. Schick, M. Stephan, *Using GPFS to Manage NVRAM-based Storage Cache*, Lecture Notes in Computer Science, Volume 7905, 2013.
15. B. Fitch, *Blue Gene Active Storage*, Presentation at HEC FSIO 2010.
16. N. Eicker *et al.*, *Supercomputing at Scale – Architectural Evolution at Jülich Supercomputing Centre*, Proceedings of the NIC Symposium 2012, pages 7-10.
17. DEEP website: `http://www.deep-project.eu`
18. DEEP-ER website: `http://www.deep-er.eu`
19. S. Breuner, *The Fraunhofer Parallel File System*, 10th HLRS/hww Workshop on Scalable Global Parallel File Systems, 2011.
20. Braam, P., *The Exa-scale I/O initiative – EIOW*, Xyratex white paper, 2012.
21. A. Duran *et al.*, *OmpSs: A proposal for programming heterogeneous multi-core architectures*, Parallel Processing Letters, Vol. 21, No. 2, 2011.
22. Green500 website: `http://green500.org/`
23. A. Adinetz, M. Axer, M. Huysegoms, S. Köhnen, J. Kraus, D. Pleiter, *Computation of Mutual Information Metric for Image Registration on Multiple GPUs*, Euro-Par 2013 Proceedings.
24. A. Adinetz, J. Kraus, J. Meinke, D. Pleiter, *GPUMAFIA: Efficient Subspace Clustering with MAFIA on GPUs*, HeteroPar 2013 Proceedings.