Grant Agreement:	604102	Project Title:	Human Brain Project

Document Title:	Assessment of "Test series" progress (Phase 3 of PCP) (Evaluation in Month 30) and definition of further process for the Big Data pre-exascale system
Document Filename:	SP7 D7.7.7 (RUP M41) APPROVED 20170919 PUBLIC
Deliverable Number:	D7.7.7
Deliverable Type:	Report
Work Package(s):	WP 7.1
Dissemination Level:	PU = Public
Planned Delivery Date:	28 Feb 2017 (RUP M41)
Actual Delivery Date:	Submitted: 11 May 2017 (RUP M43), Resubmitted: 28 Aug 2017 (RUP M47), Approved 19 Sep 2017 (RUP M48)

Authors:	Dirk PLEITER, (JUELICH, P17), WP7.1
Compiling Editors:	Anna LÜHRS, (JUELICH, P17), T7.6.1, T7.7.1, SP Manager Boris ORTH, (JUELICH, P17), T7.6.1, T7.7.1, SP Manager
Contributors:	Juan Hernando VIEITES, (EPFL, P1), WP7.1
Coordinator Review:	
Editorial Review:	Guy WILLIS, Colin MCKINNON (EPFL, P1), Editorial Services

Abstract:	This Deliverable provides a final report on the Pre-Commercial Procurement (PCP), with a focus on the third and final phase. We provide an overview of the overall process, document the outcomes and provide lessons learned. We also provide a plan for further actions regarding dissemination and maximising impact. Finally, we provide an outlook on the future path towards exascale computing.
Keywords:	High Performance Computing, Pre-Commercial Procurement (PCP)

Table of Contents

List of Figures and Tables

1. Overview of the pre-commercial procurement

The Pre-Commercial Procurement (PCP) within the Human Brain Project (HBP) was started with the goal of enabling solutions for future pre-exascale systems, which are optimised for HBP applications. A key aspect was to facilitate interactive use of future supercomputers. A PCP is a relatively new procurement instrument promoted by the European Commission, which can be used for procuring research and development services. In this context, it was considered a suitable instrument, as it aims for innovative solutions with a high technical readiness level, with products becoming available at the end or soon after the end of the PCP. Another advantage is the competitive setup of the procedure, which aims for multiple solutions in order to avoid vendor lock-in.

The goal of the HBP PCP was not to enable designs of a next generation of supercomputers, which would have been out-of-scope given a budget of EUR 2.6 million and a duration of 30 months. Rather, the strategy was to invite companies with an existing roadmap towards pre-exascale systems to address specific aspects, which would enhance their solutions and make them more suitable for the needs of the HBP. The following technical goals were defined as important elements for enabling interactivity:

- **Dense memory integration**: HPC architectures suitable for the HBP were envisaged to require a significantly larger ratio of memory capacity to computing capabilities, compared to state-of-the-art supercomputers. For instance, more complex work flows including interactive data analysis and visualisation require the ability to hold significantly larger amounts of data within the system. The targeted capacity was expected to be realisable only by integration of storage devices based on dense memory technologies.
- **Scalable visualisation capabilities**: Enhanced visualisation capabilities are key to enabling interactivity. This concerns both the ability to process large data volumes, typically generated by simulators and stored in dense memory tiers, as well as optimisation for small latencies. Communication between simulation and visualisation was thus required to be minimised, which could only be achieved by tightly coupling the visualisation with the simulation within the HPC system.
- **Dynamic resource management**: Enabling interactive usage of future HPC systems requires more flexibility in managing resources. In the PCP, usage scenarios were included where initially all resources were provided to scalable simulations, which had to be partially released at a later point in time, to make them available for data analytics and visualisation pipelines.

A PCP is organised in three phases: solution design, prototype development and pre-commercial small scale product. After each of the first two phases, the contractors for that phase were invited to submit a bid for the following phase. The initial bids for being admitted to the PCP, as well as the later bids for Phases II and III, were all evaluated by an Assessment Committee that was installed by the Procuring Entity, i.e. JUELICH. This committee not only included experts from JUELICH (Forschungszentrum Jülich, Germany) itself, but also from other relevant HBP Partners, namely BSC (Barcelona Supercomputing Center, Spain), CINECA (Consorzio Interuniversitario del Nord Est italiano per il Calcolo Automatico, Italy), CSCS (Centro Svizzero di Calcolo Scientifico, Switzerland), EPFL (Ecole Polytechnique Fédérale de Lausanne, Switzerland), RWTH (Rheinisch-Westfälische Technische Hochschule Aachen, Germany) and UPM (Universidad Politécnica de Madrid, Spain).

The original goal was to have in these three phases five, three and finally two competitors respectively. Framework and Phase I contracts were awarded to the following companies or consortia (in alphabetical order):

- Bull
- Cray
- Dell, Extoll, ParTec
- Eurotech
- IBM, NVIDIA

Unfortunately, Bull and Eurotech did not sign these contracts. As a consequence, Phase I was executed with three contractors only. After a review of their proposals for Phase II, all of them were admitted to this next phase. They all submitted a bid for Phase III. Based on scores assigned by the Assessment Committee, contracts for Phase III were awarded to Cray as well as to IBM and NVIDIA.

To facilitate smooth transitions between the phases, keeping the transition period short is important, but also a challenge. In this PCP, only 2 weeks were foreseen between the submission of bids for the next phase and the start of that phase. During this short time, the bids had to be evaluated by the Assessment Committee and the contracts for the next phase signed. In this PCP, all transitions between phases were accomplished within the foreseen timeframe.

To test the usability of the solutions, the contractors were requested to deploy pilot systems at Jülich Supercomputing Centre (JSC) during Phase III. This allowed "system prototype demonstration in operational environment", which corresponds to a technology readiness level of 7. The prototypes had to be accessible for at least 6 months and physically available at the final location for at least 3 months before the end of the PCP. Towards the end of the PCP, JUELICH decided to continue to operate the systems for 2 more years, in order to keep them available as part of the HBP High Performance Analytics and Computing (HPAC) Platform.

During the execution of the PCP, the HBP decided to make surplus funds available to this PCP. This funding was used to extend the PCP contracts and allow the Phase III contractors to enhance the pilot systems through additional equipment and R&D efforts, in order to make them more suitable for data-intensive applications of the HBP.

After the end of the PCP, the Phase III contractors agreed to join a wrap-up workshop to present and discuss the outcomes of the PCP with members of the HBP. About 30 people attended this workshop at JSC in March 2017.

Together with Cray, it is planned to continue some specific research and development activities, in particular in the area of dense memory integration, until the end of 2017.

An overview on the timeline of the PCP is shown in Table 1.

Table 1: Timeline of the HBP PCP.

Date	PCP Event
18 Dec 2013	Open Dialogue Event in Brussels
30 Apr 2014	Publication of call for tenders
30 Aug 2014	Start of Phase I
6-7 Nov 2014	Monitoring visits
02 Feb 2015	Start of Phase II
01 Apr 2015	Application training event for Phase II contractors
16 & 24 Apr 2015	Monitoring visits
31 Jul 2015	Start of Phase III
25-26 Nov 2015	Monitoring visits
08 Jul 2016	Deployment of Cray pilot system completed
16 Sep 2016	Variation extending the PCP signed
27 Sep 2016	Deployment of IBM-NVIDIA pilot system completed
31 Jan 2017	End of the PCP
15-16 Mar 2017	Wrap-up workshop

2. Pilot systems

In this section, we describe the architecture of the delivered pilot systems in preparation of the documentation of the research and development in the next section.

2.1 JULIA

The pilot system delivered by Cray is based on the following components:

- 60 compute nodes, each with one Intel Xeon Phi 7230 processor (Knights Landing), 16 GByte of MCDRAM and 96 GByte DDR4 memory. Half of the nodes are equipped with Intel DC S3500 SSDs.
- Four visualisation nodes, each with four K40 GPUs.
- Four DataWarp nodes each with two Intel DC P3600 SSDs. Each of the PCIe-attached SSDs has a capacity of 1.5 TByte.
- Login and management nodes.

All nodes are interconnected using the new Intel OmniPath[1] (OPA) technology in a pruned fat-tree topology as shown in Figure 1.

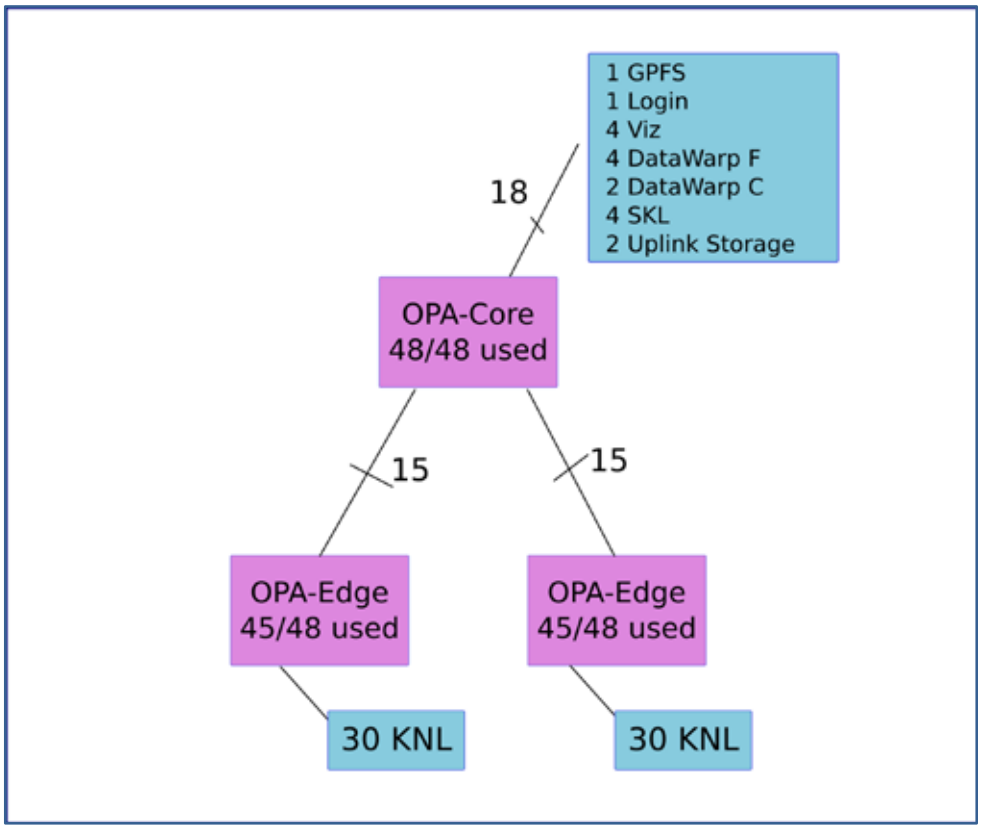

Figure 1: OPA network topology of JULIA

After the end of the PCP, the system was extended by two servers equipped with engineering samples of Intel Xeon processors of the Skylake generation. They are planned to be replaced later this year by more final versions of the processors and to be additionally equipped with DIMMs based on 3D-Xpoint technology. With continued support from Cray, this system will thus continue to serve as a very valuable test-bed for upcoming technologies.

2.2 JURON

The pilot system delivered by IBM and NVIDIA comprises of a head node as well as 18 IBM S822LC servers (also known as Minsky). The latter have the following characteristics:

- Two IBM POWER8 processors with 256 GByte of DDR4 memory attached.
- Four NVIDIA P100 GPUs, each with 16 GByte of high-bandwidth HBM2 memory.
- Each group of one processor and two GPUs is interconnected through the new NVLink technology.
- HGST Ultrastar SN100 Series NVMe SSD with a capacity of 1.6 TByte.
- Mellanox ConnectX-4 Infiniband network adapters.

An overview on the node architecture (excluding SSD and network adapter) is shown in Figure 2.

[1] http://www.intel.com/content/www/us/en/high-performance-computing-fabrics/omni-path-architecture-fabric-overview.html

The nodes are interconnected by a Mellanox Infiniband EDR fabrics in a full fat-tree topology.

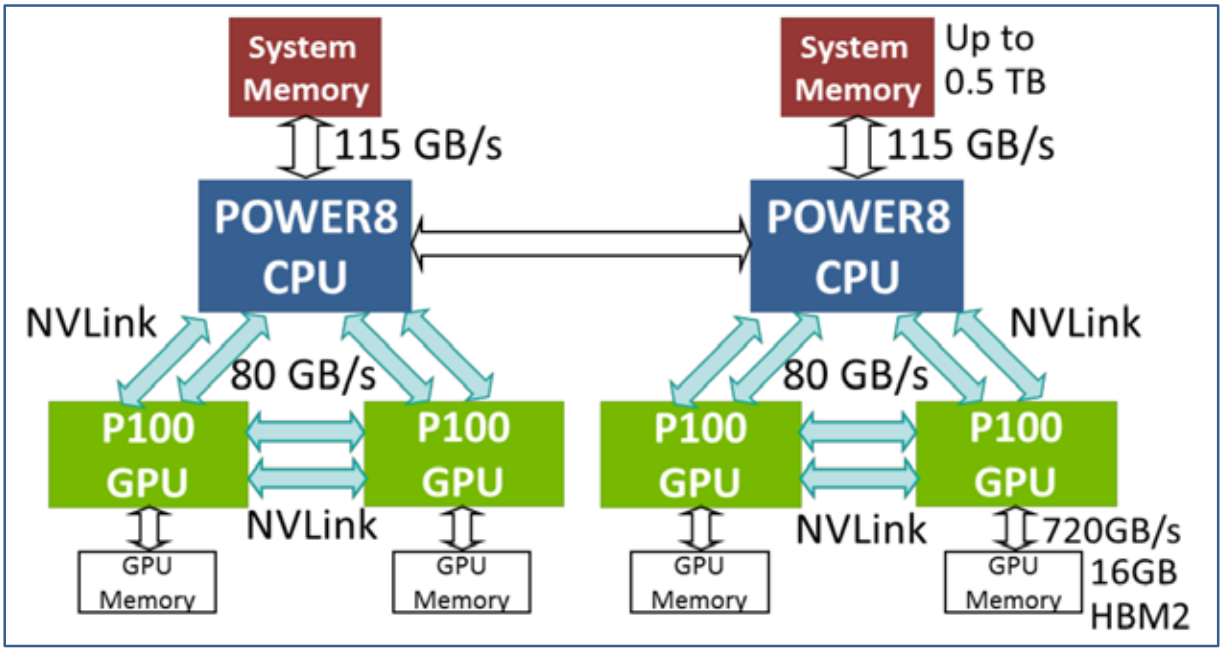

Figure 2: Schematic view on the JURON compute node architecture

2.3 Data centre integration and shared storage tier

Both pilot systems have been deployed in the main machine room of Jülich Supercomputing Centre (JSC). Each of the nodes used for interactive access to the system as well as the visualisation nodes are connected via at least one 10-GE Ethernet link. For accessing JSC's JUST main storage cluster, one node per pilot system is connected via two bonded 10-GE Ethernet links. These nodes act as GPFS servers for the GPFS clients within the pilot system.

As the 20 Gbit/s nominal bandwidth of the connection between each pilot system and a large capacity storage tier was considered insufficient for HBP applications running in production mode on those systems, an extension of the PCP was used to augment the setup with a shared GPFS Storage Server (GSS)[2] building block comprising two Lenovo x3650 GSS Servers. Each of these servers is connected to each of the pilot systems via one Infiniband EDR or one OPA link, resulting in a nominal bi-sectional bandwidth between the pilot systems and the shared storage tier of 400 Gbit/s per direction.[3]

2.4 Application benchmarks

For the purpose of ensuring usability of the pilot systems for brain simulations, the contractors were obliged to demonstrate execution of two established brain simulators, namely NEST and CoreNeuron. Cray, as well as IBM and NVIDIA, were able to demonstrate correct execution of the provided benchmarks on the pilot systems JULIA and JURON, respectively.

On JULIA, Cray compared the single node performance between a compute node with a single Xeon Phi 7230 processor and a node with two Intel Xeon E5-2680 v4 (Broadwell)

[2] https://www.ibm.com/support/knowledgecenter/SSYSP8_2.0.0/sts20_welcome.html
[3] This bandwidth cannot be saturated, as the internal bandwidth within this GSS building block is significantly smaller.

processors. The latter was observed to be significantly faster. This could be partially attributed to the lack of a fast scalar computational unit on Xeon Phi and the limited capabilities for out-of-order rescheduling of instructions. Both NEST and the extensively used MPI library functions contain many branches and scalar operations. On JURON, NEST could be built using XLC and successfully executed using the CPU cores only, as NEST does not support GPUs. NEST can use neither JULIA nor JURON efficiently. A new application, namely NestMC[4] (MC = multi-compartment), which was not available at the start of the PCP, has meanwhile been ported by the developers to both pilot systems, where good performance results have been obtained.

Also for CoreNeuron, Cray observed significantly better performance using two Intel Xeon E5-2680 v4 (Broadwell) processors than with a single Xeon Phi 7230 processor. The version of the application provided can neither efficiently exploit the on-package, high-bandwidth memory of the Xeon Phi nor its SIMD units. Significant re-factorisations of code were not within the scope of this PCP. As the provided version of CoreNeuron also did not support GPU acceleration, it could only be executed on the CPU cores of JURON, where it showed excellent scaling.

Figure 3: Pilot systems JULIA (left) and JURON (right). The racks in the middle host the shared storage.

2.5 Optimisation for data-intensive applications

During the execution of the PCP, the initial focus of the HBP with respect to HPC resources shifted from brain simulators only to include other, data-intensive, applications, such as those that occur in the context of brain image analyses required for the creation of high-resolution, 3-dimensional brain atlases. The applications' requirements overlap with those of brain simulators regarding the need for dense memory integration, which enables fast access to large data sets, as well as visualisation. These applications additionally drove the need for the integration of a fast, shared storage subsystem (see Section 2.3) and the deployment of additional software stacks.

On both JULIA as well as JURON, the following software stacks have been made available, using in most cases a version optimised for the given architecture:

[4] https://github.com/eth-cscs/nestmc-proto

- **caffe**: A Deep Learning framework originally developed at UC Berkeley[5]
- **cuDNN**: The GPU-accelerated Deep Neural Network library (cuDNN) from NVIDIA[6]
- **Lasagne**: A lightweight library to build and train neural networks in Theano[7]
- **OpenCV**: A library of programming functions mainly aimed at real-time computer vision[8]
- **TensorFlow**: An open source software library for machine learning[9]
- **Theano**: A numerical computation library for Python that is primarily developed by a machine learning group[10]

3. Research and development results

3.1 Distributed Shared Storage (DSS)

Within this PCP, IBM worked on a solution that allowed for an integration of distributed dense memory via the network, such that both compute and visualisation nodes could access it at a granularity of 1 Byte. The design team defined as additional requirements that the solution should (i) be based on (industry) standards for remote access, (ii) allow for hardware acceleration, and (iii) allow for dense memory being provided by host-attached NVMe drives.

The IBM's work builds on previous efforts in Direct Storage Access (DSA), which enabled "get" and "put" operations from and to node-local non-volatile memory, using the RDMA API from OpenFabrics. Within the PCP, IBM designed and implemented the Distributed Shared Storage-class-memory (DSS) architecture, which enables get and put operations from and to non-volatile memory devices distributed over multiple nodes. The dense memory thus becomes globally addressable. An overview on the full software stack is shown in Figure 4.

[5] http://caffe.berkeleyvision.org/
[6] https://developer.nvidia.com/cudnn/
[7] https://github.com/Lasagne/Lasagne/
[8] http://opencv.org/
[9] https://www.tensorflow.org/
[10] https://github.com/Theano/Theano/

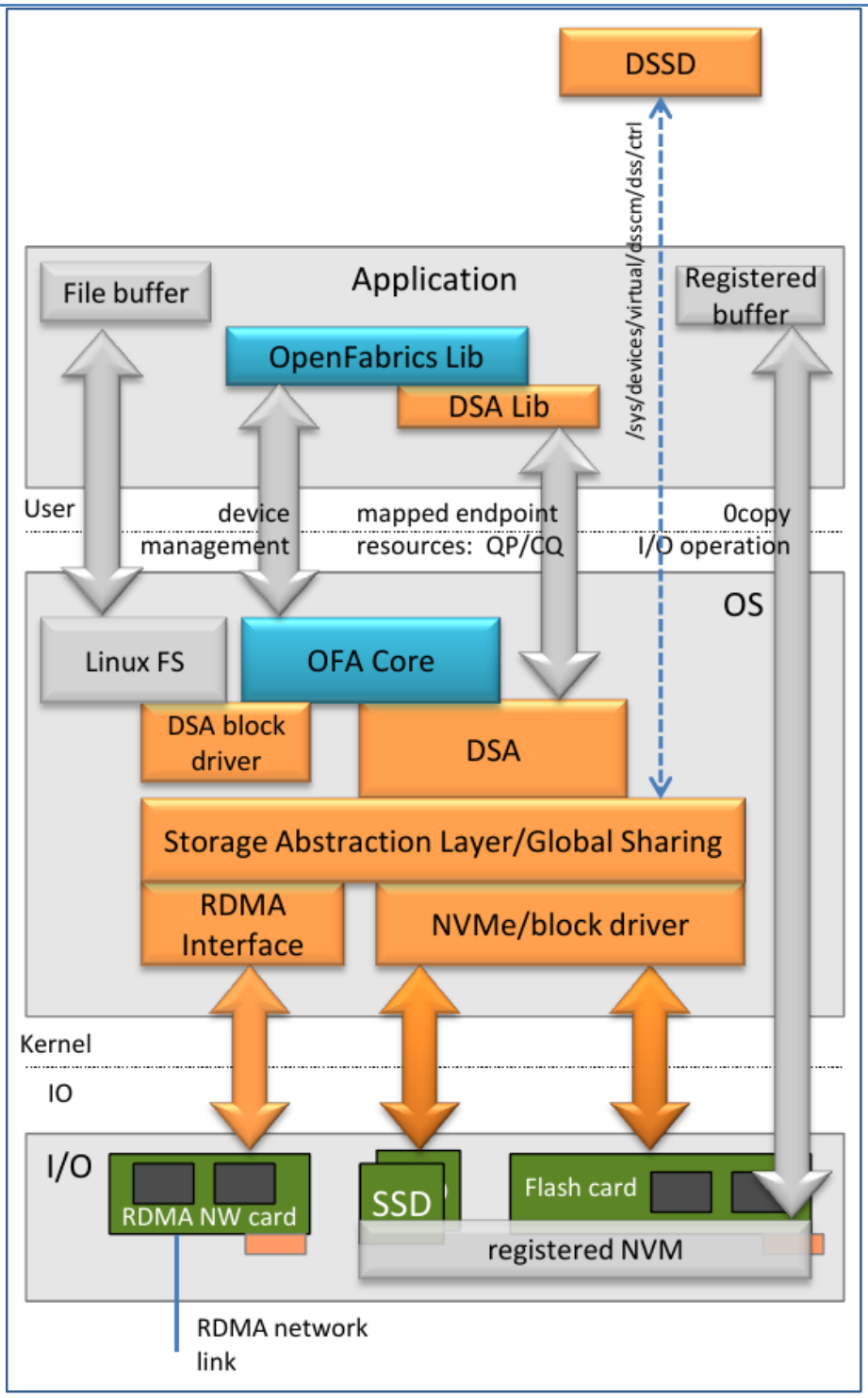

Figure 4: DSS software stack architecture.

The solution takes advantage of the RDMA API, allowing the kernel to be bypassed, which is important for reducing latencies. This API adds only about 5 µs to the latency for native local access to the dense memory. The DSS protocol for remote access adds another 10 µs. This allows high IOPs rates to be maintained, even for remote access. On JURON, the pilot system delivered by IBM and NVIDIA, it was shown that the Mellanox Infiniband EDR links could be fully saturated.

IBM plans to make this solution open source, under a dual license (BSD and GPLv2).

For further details see Section 9.1.

3.2 Scalable visualisation on POWER

IBM and NVIDIA are pursuing a future HPC roadmap based on servers with POWER processors and multiple graphics processing units (GPU) per processor socket. An HPC architecture based on a massive number of GPUs is ideal for visualisation. As the nodes of the proposed HPC architecture (see Section 2.2) also comprise a significant amount of dense memory, which is globally addressable due to DSS, two different modes of visualisation are possible:

- *In-situ* visualisation: data remains within the node and is used for local rendering pipelines. The data is typically kept in volatile memory.
- In-transit visualisation: data remains within the system, but may have to travel through the high-performance network to reach the nodes where rendering pipelines are implemented. The (transient) data is typically buffered in a dense memory tier.

To enable visualisation on the proposed architecture, it was necessary to enable the required software stacks on the POWER architecture, in particular OpenGL. This was a focus of R&D activities within the PCP. Towards the end of the PCP, the following could be demonstrated:

- RTNeuron, a complex visualisation software stack used within the HBP and one of the benchmarks that had been provided to the PCP contractors, could successfully be executed on the deployed pilot system.
- A 10% higher rendering performance could be achieved on nodes with IBM POWER8 processors and NVIDIA P100 GPUs attached via the new NVLink technology, compared to the same type of GPUs attached to x86 processors attached via PCIe.

Shortly after the end of the PCP, the software stacks were released by the vendor as products.

Below, we show results using a recent version of RTNeuron on JURON in the weak scaling case, for cylinders (Figure 5) and meshes (Figure 6). The problem size was chosen so that each GPU processed six mini-columns for cylinders and four in the case of meshes. Each of the mini-columns comprised 100 neuron cells. Results for the strong scaling for cylinders are shown in Figure 7, Figure 8 Figure 9 and for meshes in Figure 10, Figure 11 and Figure 12. For the strong scaling case, two different types of parallel rendering algorithms are shown, namely direct send (DBDirec) and sort-first (Static2).

NOTE: in Figures 5-12 below, as a function of the number of GPUs, the $Q_{0.50}$ quantile is plotted as the horizontal line inside each box, while the bottom of each box shows the $Q_{0.25}$ quantile and the top of each box the $Q_{0.75}$ quantile.

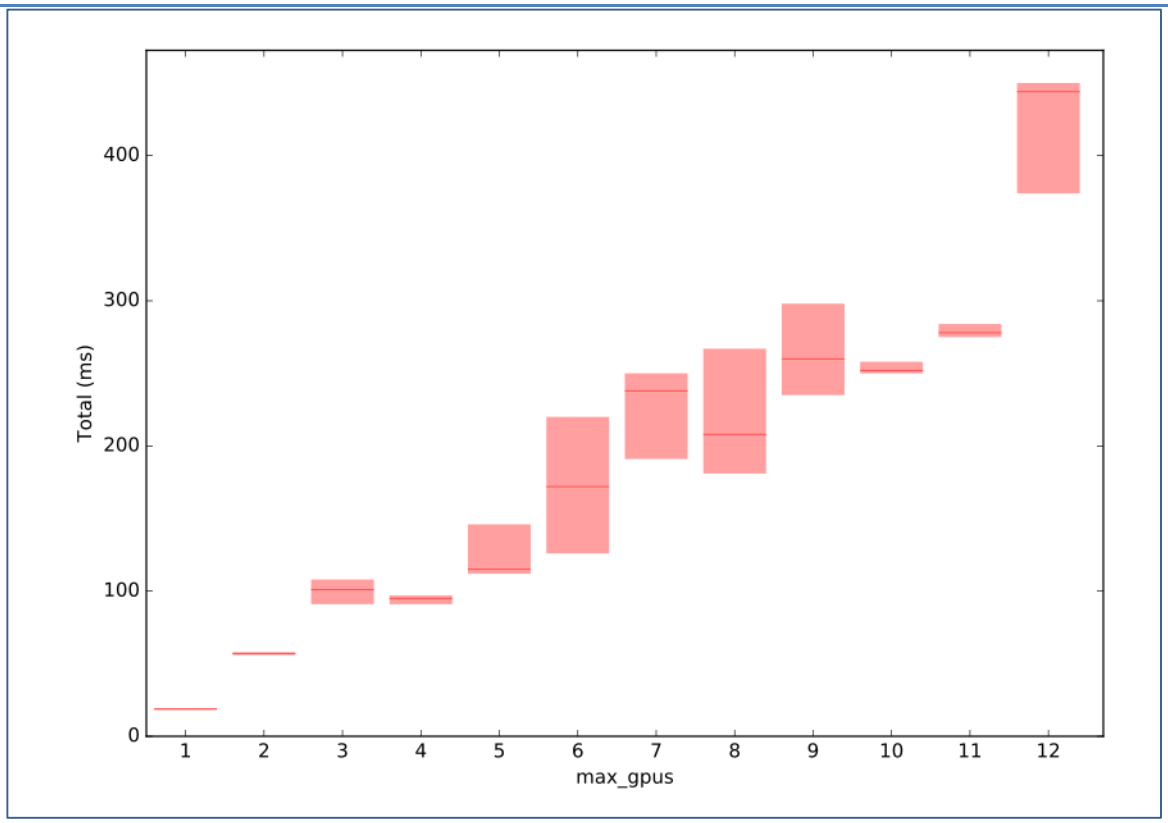

Figure 5: Weak scaling of RTNeuron on JURON using pseudo-cylinders.

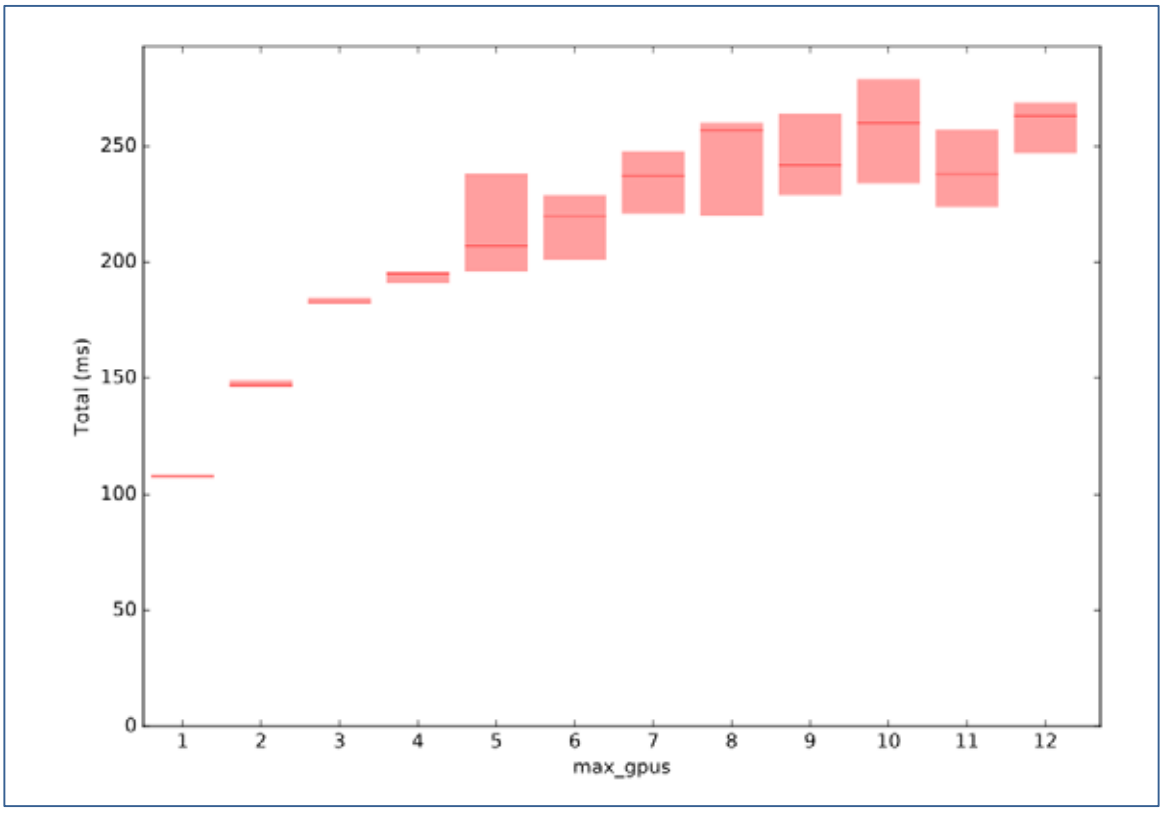

Figure 6: Weak scaling of RTNeuron on JURON using meshes.

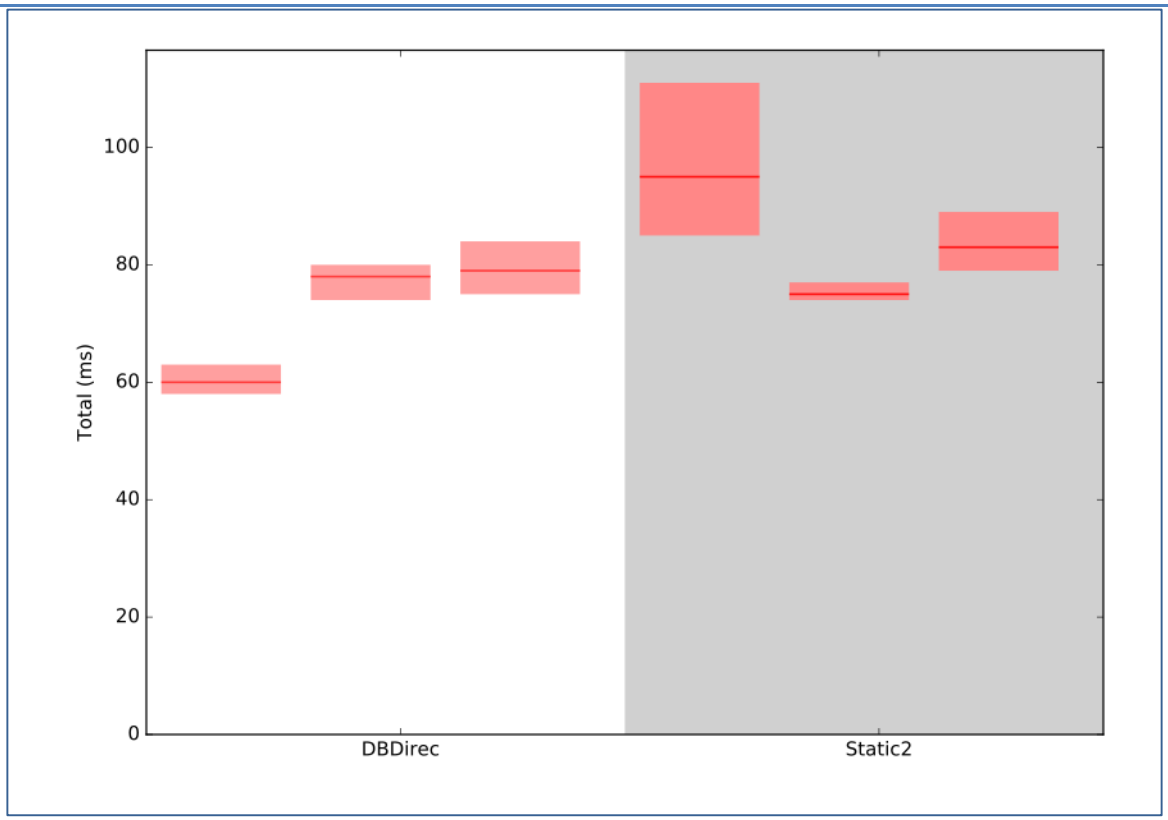

Figure 7: Strong scaling of RTNeuron on JURON for 1,000 neurons using cylinders.

NOTE: Figures 7-12 show results using 1, 2, or 3 JURON nodes (left to right), using all 4 GPUs available per node. Left half show results using parallel rendering algorithms direct send (DBDirec), right half those using sort-first (Static2).

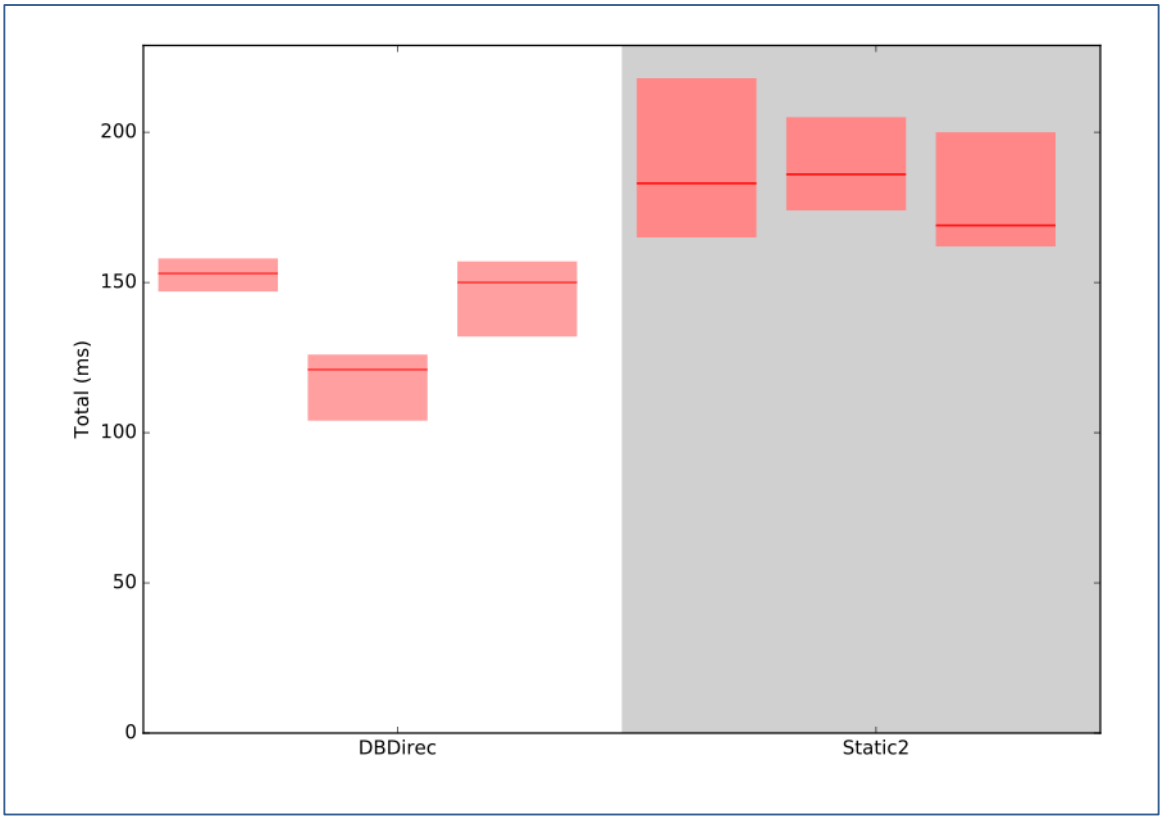

Figure 8: Strong scaling of RTNeuron on JURON for 2,000 neurons using cylinders.

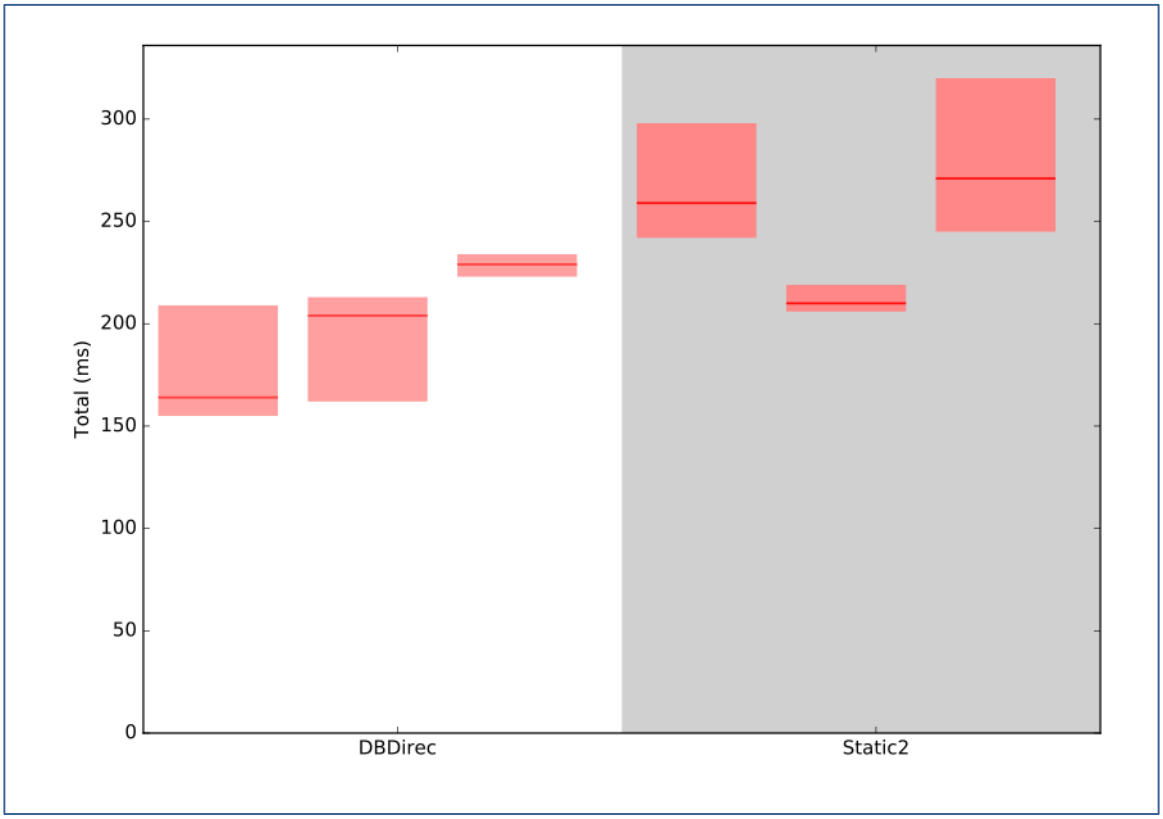

Figure 9: Strong scaling of RTNeuron on JURON for 3,000 neurons using cylinders.

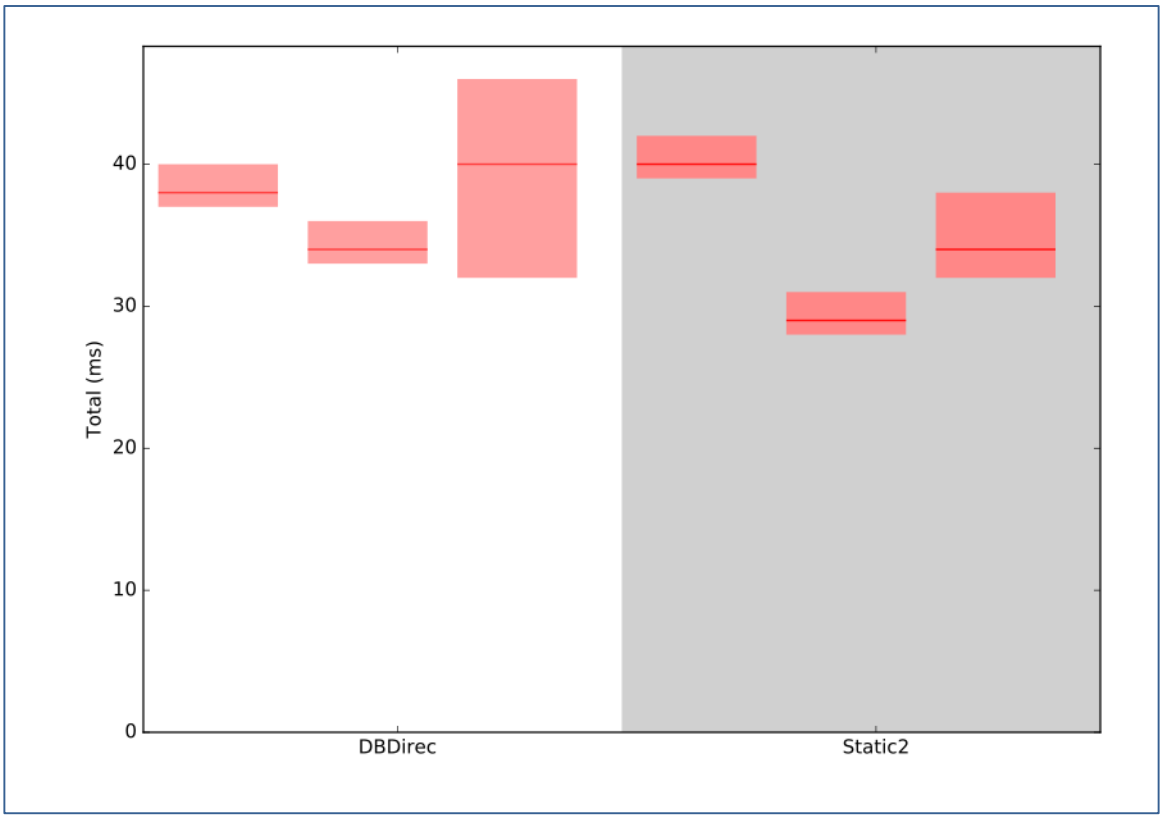

Figure 10: Strong scaling of RTNeuron on JURON for 200 neurons using meshes.

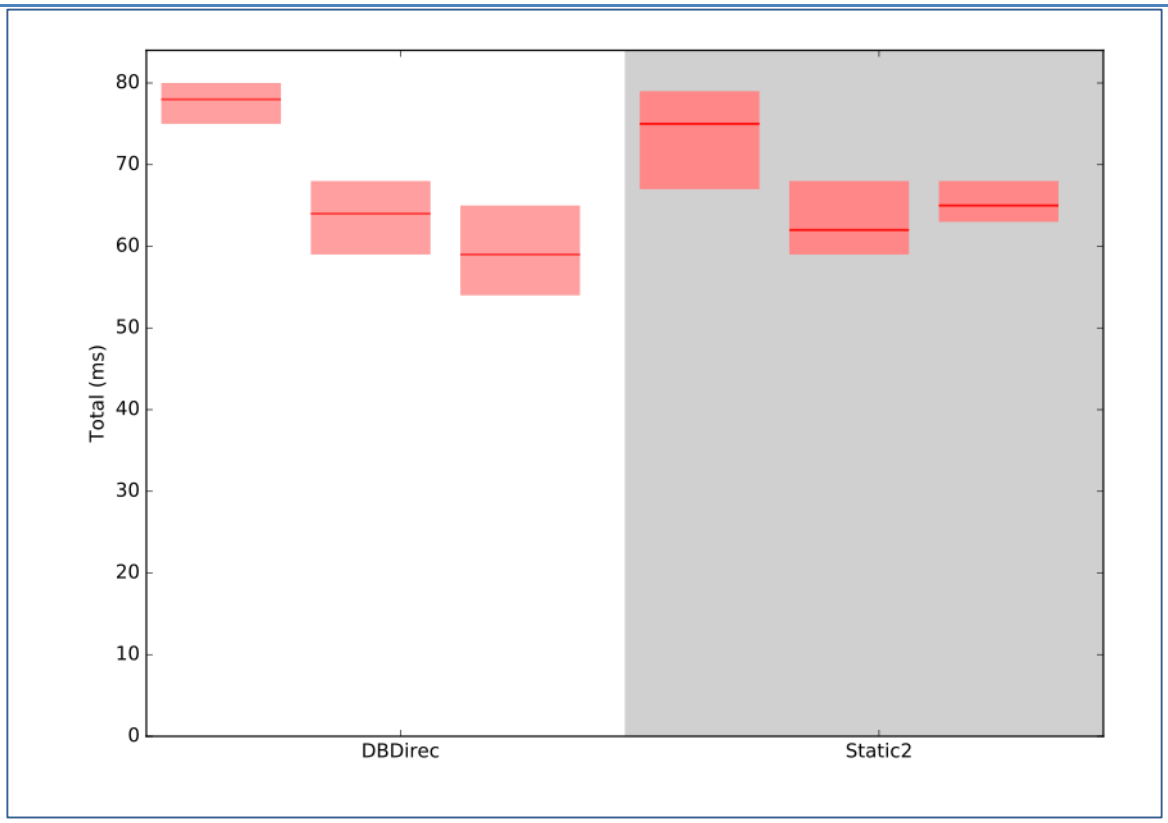

Figure 11: Strong scaling of RTNeuron on JURON for 500 neurons using meshes.

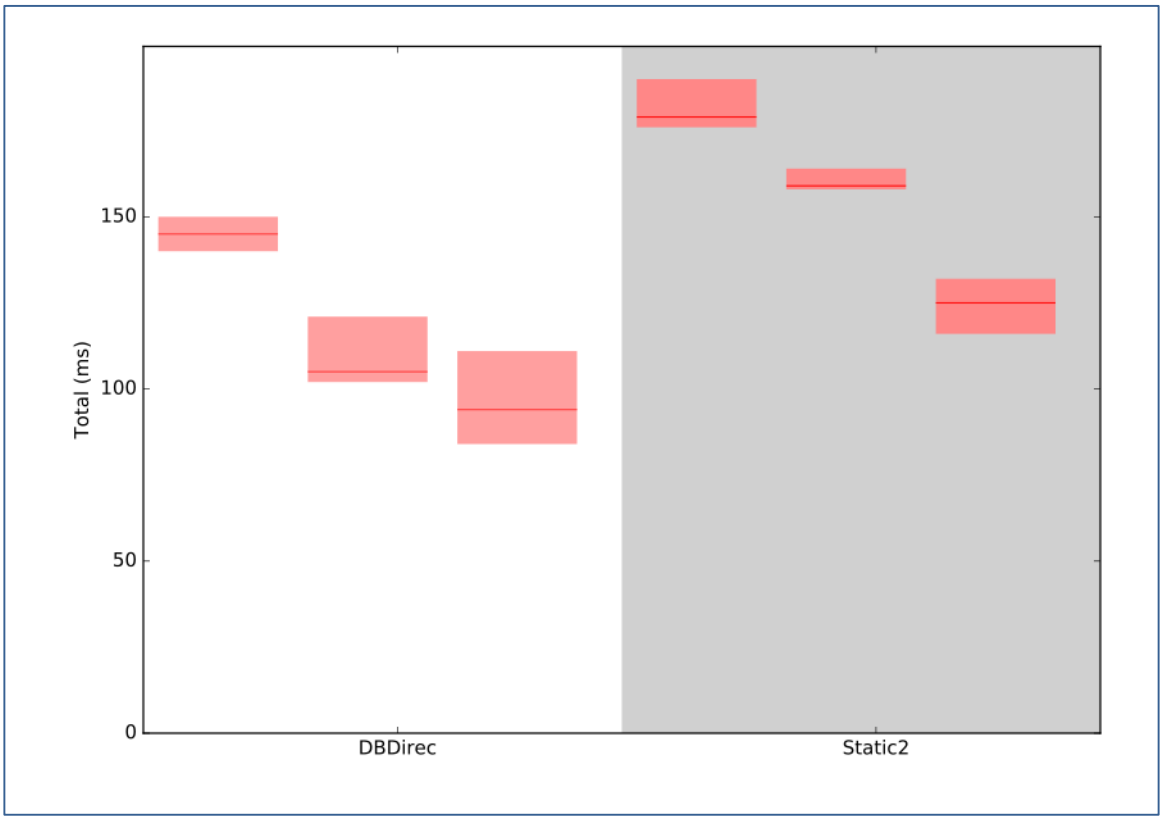

Figure 12: Strong scaling of RTNeuron on JURON for 1,000 neurons using meshes.

In neither the weak nor the strong scaling cases was the observed scaling of time-to-solution of the number of GPUs N_{GPU} as expected. In the former case, one would expect time-to-solution to remain constant, while in the latter, a decrease proportional to $1/N_{GPU}$.

It is, however, important to realise that RTNeuron does not allow strict control of the amount of the work that needs to be executed, and consequently the amount of work is not directly proportional to the number of neurons. This means, for instance, that in the case of a weak scaling analysis, when the amount of used hardware resources employed (here measured in units of number of GPUs) is doubled, the amount of work does not exactly double - only the circuit size does.

Strong scaling limitations were partially due to frustum culling and scene decompositions, which added an overhead that could be reduced by further parallelisation. View frustum culling required finding the intervals that need rendering in the triangle list from each neuron. Computing these intervals could be done either on the GPU, using CUDA, or on the CPU, using an octree. In the strong scaling case, the ratio of triangles (or any other primitive) rendered per OpenGL draw call had to be high enough to make sure that the GPU was kept busy all the time. RTNeuron allows two different geometric models to be used for neurons: meshes or pseudo-cylinders. For frustum culling, the amount of work was less in the case of pseudo-cylinders. We believe that this explains, at least partially, the significantly worse scaling behaviour observed in the case of pseudo-cylinders.

The rendering of the scene decomposition used for direct-send was expected to suffer from a similar scaling limitation. The spatial partition tried to balance the number of triangles per GPU, but it had to provide a total spatial order between all the pieces. This was accomplished by arranging them in a k-dimensional tree (k-d tree), which, however, did not guarantee that the number of neurons that intersected each leaf in the tree would be constant. In practice, a large fraction of the neurons was present in each leaf due to the shape of the neurons. Distributing the problem over a larger number of GPUs resulted in a larger number of smaller objects, which had to be rendered. This effect was most notable in the case of pseudo-cylinders. When using a larger number of GPUs, the number of cylinders per neuron became small in relation to the number of objects the CPU traversed in the scene graph. This resulted in the OpenGL driver and the CPU-GPU communication becoming the main bottlenecks.

In the weak scaling case, we observed an unfavourable scaling behaviour when using pseudo-cylinders (see Figure 5). When using meshes, which require more computations on the GPUs, we observe a plateau beyond 8 GPUs or 2 nodes (see Figure 6). For this setup, using 1 or 2 GPUs was more favourable for small problem sizes.

In the strong scaling case, we observed again an unfavourable scaling behaviour when using pseudo-cylinders, even when using a larger circuit (see Figure 7, Figure 8, and Figure 9). When using meshes and 1,000 neurons, we found a reasonably strong scaling behaviour (see Figure 12).

For further details provided by the contracting consortium of IBM and NVIDIA, see Section 9.2. Note that these results were obtained with an older version of RTNeuron, which was provided to the contractors at the start of the project.

3.3 Dynamic resource management integrated into LSF

The solution for dynamic resource management proposed, designed and implemented by IBM enabled resource reallocation for parallel applications with intermediate checkpoint restart. IBM's product Platform Load Sharing Facility (LSF)[11] was extended in the following respects:

- A new command for requesting more nodes was introduced,
- The capability to shrink and grow indefinitely was added, and
- The capability was added for a job, which did initially not request all slots, to request all remaining slots.

For checkpoint restart, a modified version of the Scalable Checkpoint Restart Library (SCR)[12] was used. SCR provided the application with an interface to write and read checkpoint files. To enable malleability, SCR needed to be modified to allow for applications to restart with a smaller or larger number of processes. The extensions included support for processes to retrieve checkpoint files created by other processes. Furthermore, support for spawning of processes as introduced with the MPI-2 standard was introduced.

The modifications to LSF will be integrated into a future product release, which is expected to become available later in 2017.

For further details see Section 9.3.

3.4 Enhancement of Apache Spark on JURON

During the Ramp-Up Phase of the HBP, a need for support of data analytics frameworks like Apache Spark was identified. The current version of Apache Spark could not, however, fully exploit the capabilities of hardware architectures comprising fast storage devices that can be accessed through a high-performance network supporting RDMA. To accelerate Spark, IBM provided Crail[13] on the pilot system, which is a distributed file system that can be accessed through an HDFS API as well as a dedicated I/O interface used by Spark plugins.

Within the PCP, the Crail Data Node was extended to facilitate efficient integration of Flash memory tiers. On the pilot system, IBM demonstrated a more than two-fold increase in performance for the TeraSort benchmark, using 16 nodes and a data set of 100 GByte.

The Crail data node integration will be released as open source at https://github.com/zrlio/crail-nvmf.

3.5 Integration of dense memory on JULIA

The initial goal of Cray was to demonstrate the integration of DataWarp nodes using the new non-volatile memory technology 3D-Xpoint, which has been developed by Intel and Micron. This has not yet been realised, due to the delayed availability of this technology. It was planned to use this technology in a version where it is attached to the memory bus of the processor. This would require Intel Xeon server processors of the next (Skylake) generation, which were not available until the end of the PCP.

Instead, Ceph[14] was used to provide access to the NVMe flash memory devices that were integrated in the DataWarp nodes of the pilot system JULIA. So far, Ceph has not been used

[11] http://www-03.ibm.com/systems/spectrum-computing/products/lsf/
[12] http://computation.llnl.gov/projects/scalable-checkpoint-restart-for-mpi
[13] https://crail.io
[14] http://ceph.com/

within HPC systems, but it provides a number of features which are relevant in this context, including metadata scalability and resilience based on mirroring and erasure encoding. However, identified performance limitations require further investigation and triggered the exploration of other solutions. As all considered solutions are open source, the configuration and setup, which are planned to be documented in a white paper, can be exploited by others.

JUELICH and Cray agreed to continue the work on dense memory integration after the end of the PCP. This will include the integration of memory and storage devices based on the aforementioned memory technology, 3D-Xpoint.

For further details see Section 9.4.

3.6 Scalable remote visualisation on JULIA

The scalable visualisation solution realised by Cray was based on visualisation servers with four GPUs of NVIDIA's Kepler generation. All servers were tightly integrated into the high-performance network. Using the non-volatile memory integrated in the DataWarp nodes as well as in the compute nodes allowed implementation of in-transit visualisation solutions, where data needs to be moved between nodes but remains within the HPC systems.

The capabilities of this solution were demonstrated using RTNeuron. A reasonable strong scaling and a good weak scaling behaviour could be demonstrated on up to four nodes.

During the evaluation period, the Virtual Reality & Immersive Visualization group from RWTH Aachen started to explore the use of the pilot system's compute nodes for *in-situ* visualisation, with good results.

For further details see Section 9.6.

4. Dissemination and impact

4.1 Dissemination activities during the implementation of the PCP

Pre-commercial procurement is a relatively new instrument, in particular in the HPC market. Both public procurers and solution providers are still in the process of collecting experience. Therefore, communication played an important role during the implementation of this PCP.

Before publishing the call for tender, communication was crucial to inform the market about the planned PCP. The public Open Dialogue event, which was organised in Brussels on 18 December 2013, was used to present a concrete snapshot of how we intended to organise the PCP, both in technical as well as legal terms. Interested suppliers were strongly encouraged to provide feedback. This feedback was taken into account in order to make participation to the PCP as attractive as possible, without compromising on the technical goals of the PCP.

After launching the PCP, the obtained experience was shared with others to contribute to establishing best practices at the European level. Members of the team involved in the organisation of the PCP attended each "Annual Concertation Meeting of ongoing Innovative Procurement Projects" organised by the European Commission (EC). This event was often used to present details on the organisation of tender documents, as well on how to manage intellectual property rights. Discussions with participants who were in the process of preparing a PCP allowed transfer of knowledge and experience (e.g. by sharing the tender documents).

Throughout the PCP, the general public was informed about its progress by means of press releases:

- 01 October 2014: Press release informing about the awarding of framework agreement contracts[15]
- 01 October 2015: Press release informing about the start of the third and final phase of the PCP[16]
- 28 September 2016: Press release information about the installation of the pilot systems[17]

The pilot systems are prominently featured, with a dedicated webpage on the Forschungszentrum Jülich web portal.[18]

The HBP PCP was also presented at a number of HPC-related workshops and conferences:

- IDC User Forum, 27 October 2014, Stuttgart, Germany
- SC15 Exhibition, 10 November 2015, Austin, USA
- CEA-FZJ Workshop, 5 July 2016, Jülich, Germany
- Eastern Partnership E-Infrastructure Conference, 6-7 October 2016, Tbilisi, Georgia
- HBP Summit 2016, 12-15 October 2016, Florence, Italy
- Seminar at Tata Consulting Services, 28 November 2016, Mumbai, India
- Webinar on visualisation on JURON (Peter Messmer, NVIDIA), 21 February 2017.

4.2 Dissemination activities after completion of the PCP

The dissemination of results will continue after the end of the PCP. In particular, we plan to:

- Approach organisations that support PCP projects and offer to make our experience with PCP available. Candidates:
  - ZENIT GmbH, Mülheim an der Ruhr (http://www.zenit.de/)
  - KoWi, Verein zur Förderung der europäischen und internationalen wissenschaftlichen Zusammenarbeit e.V., Bonn (http://www.kowi.de/)
- Continue to attend the Annual Concertation Meetings organised by the EC
- Publish an article in the Spring 2017 edition of Inside (http://inside.hlrs.de/)
- Document PCP results in a leaflet, which will be used by JSC during upcoming SC and ISC exhibitions
- Work on getting articles published in media focusing on supercomputing, e.g., insideHPC (http://insidehpc.com/) or HPCwire (https://www.hpcwire.com/).

We explored options to publish our experience with implementing this PCP in a journal. However, no suitable journal on public procurement and innovation could be identified so

[15] http://www.fz-juelich.de/SharedDocs/Pressemitteilungen/UK/EN/2014/14-10-01HBP_PCP.html
[16] http://www.fz-juelich.de/SharedDocs/Pressemitteilungen/UK/EN/2015/15-10-01hbp-interaktiver-rechner.html
[17] http://www.fz-juelich.de/SharedDocs/Pressemitteilungen/UK/EN/2016/16-09-27hbp_pilotsysteme.html
[18] News on the installation of the systems were also posted on Twitter.

far. We will continue to explore Elsevier's journal Technovation[19], which could be a suitable candidate.

4.3 Maximising impact of PCP outcomes

After the end of the PCP, efforts have been started to maximise the impact of its outcomes. We have been working with the Phase III contractors to support the "productisation" of the outcomes of the research and development work performed within the PCP, and will continue these efforts. Furthermore, we have started to work on exploiting the results within the HBP and to promote their adoption.

4.3.1 Productisation and exploitation of PCP outcomes

During the HBP PCP wrap-up workshop, the following options for productisation and exploitation of the outcomes of the research and development work were identified:

- Research and development results on dynamic resource management obtained by IBM Ireland (IE) are being integrated into LSF and will become part of a next release of LSF that is planned to become available later in 2017.
- In the context of research on dynamic resource management, IBM Ireland implemented changes to the Scalable Checkpoint Restart (SCR) library. JUELICH will work with IBM Ireland on the integration of these changes into the main branch of SCR.
- NVIDIA DevTech experts in Switzerland and advanced rendering experts at the NVIDIA branch in Berlin (DE) worked on porting visualisation tools to OpenPOWER based HPC nodes. The results have already been integrated into the latest driver releases.
- At its EMEA Research Lab in Bristol (UK), Cray worked on enabling access to fast storage tiers integrated into HPC systems using the open-source storage solution Ceph. Results are planned to be published in a joint white paper and will thus be made available to other supercomputing centres.
- IBM developed a software layer that allows fast data transfer using addressable get/put operations from/to distributed NVMe devices. This software will be released open source. Technical details and evaluation results are planned to be published in a joint paper.
- The Crail data node integration by IBM has already been made available as open source material.

4.3.2 Exploitation and adoption of PCP outcomes within the HBP

Within the HBP, work has started to exploit and adopt the PCP outcomes.

The pilot systems will be available after the end of the PCP for two more years, i.e. until January 2019. The systems will be available to users from the HBP as part of the HPAC Platform. Specific work has been started to promote the uptake of these systems as well as the outcomes of the research and development work. One focus is the porting of brain atlas workflows to the new systems, with image registration focusing on JURON and deep learning frameworks for image processing focussing on JULIA. During a recently organised GPU Hackathon at JSC[20], three teams with a neuroscience background participated, which meanwhile have all started to use JURON.

[19] https://www.journals.elsevier.com/technovation/
[20] https://indico-jsc.fz-juelich.de/event/32/

The PCP pilot systems, including the deployed new technologies, have started to be actively used for research and development work within SGA1 of the HBP. This concerns in particular the work within Work Package WP7.2 on "Data-Intensive Supercomputing". Both systems provide an expanded memory, I/O hierarchy and a good environment for exploring different ways of accessing dense memory-based tiers.

5. Lessons learned from PCP implementation

In this section, we summarise the lessons learned from this PCP in the context of the HPC market.

5.1 PCP as an instrument favouring productisation and the relevance of vendor roadmaps

While many other research and development projects in the area of HPC that receive public funding are led by academic entities, the research and development work within a PCP is, by design, led by a commercial operator. At the same time, the public procurer remains in strong control, as it defines goals, objectives and requirements of solutions, which may become part of a follow-up procurement. This gives the involved commercial operators a strong incentive to organise the research and development work so that it leads to outcomes that are close to becoming products at the end of the PCP. As shown in Section 3, this PCP resulted in such outcomes, with some of them having already a clear path towards productisation.

It is, however, important to note that the design of a PCP is not a sufficient requirement to achieve such a situation. For this to happen, the following conditions are also important:

- The technical goals defined for a PCP must be aligned with the roadmaps of a sufficiently large number of commercial operators. One should not assume that a PCP will have a disruptive effect on their roadmaps.
- In particular, if the market is small, as it is the case for HPC, the public procurer should be in a position to create confidence that there is indeed a market for the solutions to be developed within the PCP. This means that the procurement roadmap of the public procurer (or buyers group) must be aligned with the goals of the PCP.

Both conditions were fulfilled in case of the HBP PCP. For several reasons, this PCP was designed so that the solutions would augment the relevant suppliers' roadmaps towards solutions for pre-exascale HPC systems. Furthermore, the HBP provided a strong vision for its future HPAC Platform and thus was able to explain convincingly how outcomes of the PCP could become part of a future HPC infrastructure.

5.2 Need to balance scope and available funding

When addressing the HPC market, one is confronted with the following specific challenges:

- The market for high-end HPC systems is relatively small, with a relatively small number of full-solution suppliers and a small number of potential customers, most of which are public procurers.
- The non-recurrent engineering (NRE) costs for many of the key technologies required for high-end HPC systems are extremely high and typically exceed by a significant margin the budgets for the PCPs executed so far in the field of HPC (including this PCP).

With the small number of potential bidders, it is important to avoid further reducing the number of interested bidders by defining technical goals whose realisation would result in NRE costs that significantly exceed the available budget. If the number of interested bidders is reduced, there is a high risk of the number of actual competitors in the PCP becoming too small.

This market situation did affect the HBP PCP. It was planned to award five contracts for Phase I, but only five bids were received. As these were all good enough, it was decided to award all bidders with a Phase I contract. However, two of them, Bull (DE) and Eurotech (IT), decided not to sign these contracts.

5.3 Relevance of attractive requirements for IPRs

Requirements concerning procurers' access to suppliers' intellectual property rights (IPRs) constitute another major obstacle which may discourage potential bidders. For many commercial operators in this field, IPRs are a crucial asset, which makes them keen to avoid any risk of losing control of them. Protection of a company's IPRs may be seen at risk if the procuring entity asks for the right to access "relevant background intellectual property (IP)". The latter refers to IP, which the contractor created independently of the PCP, but which the procurer needs to access in order to be able to use the foreground IP created within the PCP.

To mitigate this problem, it was decided within the HBP to differentiate between design specifications resulting from research and development services within the PCP, and design implementations. The rationale was that it would be easier to keep design specifications free of background IP than design implementations. For design specifications, the procuring entity requested an irrevocable, worldwide, free and non-exclusive license to use the foreground as well as the relevant background IP for any purposes. For implementation specifications, however, the procuring entity requested only an irrevocable, worldwide, free and non-exclusive license to use foreground IP only for the purpose of using this implementation non-commercially.

For the HBP PCP, we are not aware of any potential bidder that decided not to submit a bid because of our IP requirements. We thus believe that we provided sufficiently attractive IP environment in the given context.

5.4 Importance of a dialogue with the market

A dialogue with the market, before publication of the tender, has been identified as extremely important. On the one hand, it allows the market to be informed about the PCP's technical goals and procedures and, as a result, to attract more bidders. On the other hand, it provides an opportunity to receive feedback from companies, which can be used for finalising the tender documents. In the context of technical goals and requirements, this may – for example - result in a better understanding of whether the scope of the PCP and the funding provided are balanced. Information on procedural aspects are also important, as the PCP is a relatively new procurement instrument, and thus the procurer needs to address a lack of knowledge among the potential bidders. Furthermore, procedural aspects can have a crucial impact on whether or not commercial operators expect a given PCP to provide attractive conditions for providing commercial research and development services. The IP requirements discussed in the previous section provide a good example of this.

In the case of the HBP, a single Open Dialogue event was organised, at which the procuring entity presented its current thinking about the technical goals and the organisation of the procedure. Receiving feedback from potential bidders was considered to be a crucial aim

and the event agenda did foresee a significant amount of time for feedback[21]. Additionally, all participants and all other interested potential bidders were invited to provide written feedback within one month after the event. When time permits, a second Open Dialogue event could be useful during the preparation phase of a PCP. This was not an option for the HBP PCP, due to its tight time schedule.

5.5 Allow for co-design approaches

For the design of future HPC architectures and technologies, a co-design process which integrates feedback from potential users of the resulting product(s) has become an established approach.[22] On the one hand, this ensures that likely applications are analysed by engineers in order to develop the most efficient solutions. On the other hand, proposed architectures and technologies are reviewed by representatives of likely user communities to identify issues and opportunities to exploit.

A PCP intentionally creates a very formalised relationship between the procuring entity and the contractors, as the former has to ensure fair and equal treatment of the latter. This results in restrictions when it comes to implementing a co-design process. However, opportunities for at least partial realisation of such a process were identified during the HBP PCP:

- The contractors are free to engage with academic entities, which could provide the desired co-design input, as long as this does not involve the procuring entity or other organisations which are already strongly involved in the PCP, such as academic entities which are formal sub-contractors to the PCP. Potential bidders should be encouraged to consider seeking inputs from academic institutions which are not directly involved in the PCP, but which may use its outputs.
- After the start of Phase III of the PCP, no more comparative evaluation of the contractors was performed. This provided more freedom for interaction between the procuring entity and the contractors. To comply with the obligation of fair and equal treatment of all contractors, the procuring entity within the HBP PCP aimed to ensure a similar duration and level of interaction with each of the contractors, without insisting that the content of the interaction was the same in each case.[23]

5.6 Importance of monitoring and steering the process

It is crucial to track the progress of the work of the contractors and not just wait for end-of-phase reports and bids for the next phase. In case of the HBP PCP, monitoring visits were organised during each phase. The monitoring visits followed a formal scheme to ensure equal treatment of all contractors, which during Phase I and II were also competitors for entering the next phase. As this was no longer the case in Phase III, additional, less formal and more frequent channels for interaction were used during that final phase.

The interaction with the contractors was crucial to clarify open questions, provide feedback to the proposed approaches and to collect suggestions from them. The outcome of the monitoring visits was used to steer the process in the following ways:

[21] The event started on 10:30 am and finished at 4pm and did foresee 3 sessions of 30, 30 and 60 minutes, respectively, for questions and answers as well as discussion.
[22] See, e.g., R.F. Barrett et al., "On the Role of Co-design in High Performance Computing", Advances in Parallel Computing, 2013.
[23] This reflects the fact that also the solutions proposed by the different contractors are not the same.

- Information provided by the contractors was used to prepare the tender for the next phase.
- Monitoring visits were used to discuss changes to the process through variations to the contracts, and to explore whether all contractors could agree to such changes.

Interactions with the contractors also resulted in additional support activities like a joint training workshop, providing them with a deep dive into the applications benchmarks provided in this PCP.

5.7 Europe as an attractive place for commercial R&D work

By signing the framework contract of the HBP PCP, the contractors committed to performing at least 60% of the resulting research and development services within the EU Member States or countries formally associated to the European Commission's Framework Programme 7. All offers received for this PCP for all phases significantly exceeded this threshold. Both PCP Phase III contractors performed all their related research and development actvities in Europe.

This suggests that Europe is an attractive place for international companies to execute research and development work. The PCP framework probably also helped to increase the involvement of European subsidiaries of international companies headquartered outside Europe.

5.8 The PCP as an instrument for supporting SMEs

By allowing consortia to bid in a PCP, the risks for SMEs participating can be kept sufficiently low to attract their interest. This is a benefit for the procuring entity, as SMEs are often more willing to push for new solutions. A potential benefit for SMEs is the opportunity to demonstrate the capabilities and possibilities of their solutions in a specialised environment. The requirement to deliver a pilot system, as was the case in the HBP PCP, can thus be of particular interest for SMEs.

Within the HBP PCP, two small SMEs joined a consortium led by a large integrator to submit a joint bid. This consortium successfully participated in the first two phases of this PCP.

6. Path towards future exascale HBP infrastructure

During the execution of the HBP PCP, the original vision of interactive supercomputing changed with the inclusion of additional requirements related to data-intensive applications other than the original simulation goal. Within the HBP, there is an urgent need for interactive access to IT infrastructures that include scalable HPC systems for use cases such as:

- Enabling data curation and releases
- Interactive 3-dimensional data viewers
- Data analytics-based brain image analysis, e.g. for segmentation of cortical areas in the human brain
- Enrichment of the Human Brain Atlas
- Comparative analysis of experimental and simulated data
- Virtual anatomy and virtual imaging lab apps from CDP1

This change in requirements resulted in an adjustment of the architecture of the HPAC Platform and the new Fenix infrastructure. This infrastructure will bring together a federated data infrastructure and scalable compute resources with an aggregate compute capability of at least 50 PFlop/s. The federation is expected to involve five European supercomputing centres initially, namely BSC, CEA, CINECA, CSCS and JSC. Interactive use of this infrastructure will be facilitated through Interactive Compute Nodes (ICN), which will be deployed at all participating sites. From these ICNs, it will be possible to access data in the federated data infrastructure, to stage data to locally attached dense memory, thereby enabling extremely fast data access, and to exploit attached scalable compute resources and scalable visualisation capabilities.

The infrastructure, which is now in preparation, will be exploitable both by scalable brain simulators as well as by data analytics workflows, with process data generated by the simulators or coming from experimental sources like high-performance brain image scanners, electrophysiological experiments or neuromorphic compute facilities. Compared to the Platform which was available at the start of SGA1, this new one will provide an improvement in capacity and capabilities of several orders of magnitude during the next few years.

Technologies developed within the PCP will continue to play an important role in this infrastructure:

- Integration of dense memory continues to be necessary to meet the requirements for fast access to vast amounts of data. The solutions developed, implemented and evaluated within the PCP come with different performance characteristics and functional features. DSA/DSS will enable fast, fine-grained access to data, for applications such as different visualisation use cases. Approaches such as that based on Ceph will facilitate flexible solutions for software management storage tiers based on dense memory technologies.
- The results of the PCP enable tighter integration of visualisation and scalable HPC systems. Exploiting the dense memory facilitates in-transit visualisation concepts, where data producing pipelines and data-consuming visualisation pipelines are implemented on different parts of the system, while keeping the data within the system.
- The work on dynamic resource management will allow the attached scalable compute resources to be made more elastic.
- Apache Spark is a distributed processing framework which is increasingly used for large-scale data processing tasks, including some within the HBP, in particular the SP5 Neuroinformatics Platform. Optimised versions of Apache Spark or solutions like Crail, which allow acceleration of Apache Spark, will be part of the future HPAC Platform. It is planned to continue using the accelerated data access solutions developed within the PCP.

7. Recommendations received from reviewers

- *"Additional tests with dense memory integration are worthwhile and detailed brain simulation code optimisation for JURON and JULIA regarding GPU and KNL usage promises efficient use of these systems. This work can be performed in the additional two years the consortium and their HPC vendor partners offered on a voluntary basis for further development of the platforms and their environment."*

  JUELICH and Cray have agreed to continue the work on integration of dense memory using Ceph. The work focusses on using the DataWarp nodes for staging data, which is

used in brain image segmentation workflows, as well as for holding transient data for in-transit visualisation of brain simulation (NEST) output.

- *"SP7 should have the mandate to coordinate co-design activities with all the other SPs. The fact that the PCP was not designed with code optimisation in mind needs to be rectified for future projects. This will allow for a detailed performance assessment of the results. HBP as a whole would significantly benefit from a hardware-software co-design procedure. It is recommended to use mechanisms by which the tuning will feed back to the core code or exploiting so-called mini-apps that extract performance relevant kernels from the applications."*

  This recommendation has been taken into account for the SGA2 work plan. One of the key objectives of WP7.3 will be the development of concepts, numerical algorithms and software technology for simulation codes, focusing on their readiness for exascale architectures. Task T7.1.4 is responsible for sharing the created know-how with external stakeholders, in particular providers of exascale solutions. Mini-applications are considered as one option for facilitating this transfer of knowhow.

- *"In addition to above recommendation for closing D7.7.7, it was not clear whether strong or weak scaling was more important, as they affect the balance of the machine, memory vs. network, latency vs. bandwidth etc. It is important to assess the scalability requirements in the context of whether the neural networks, both neuromorphic and deep neural networks for image recognition, are fairly fixed in size or scale variably; such an effort should be part of the performance evaluation effort in SGA1."*

  For brain simulators, neuroscientists have defined weak scaling requirements to facilitate simulation of networks of maximum size. On current systems, network performance has not been identified as a limiting factor. Optimal balance of other hardware parameters, e.g. memory bandwidth versus throughput of arithmetic operations, differs significantly between the different simulators. A more systematic analysis and resulting annotation of use cases will be performed in SGA2 by Task T7.1.2.

8. Ethical issues

Supercomputers fall in the class of dual-use items that can be used for both civilian and military applications and/or can contribute to the proliferation of Weapons of Mass Destruction (WMD). They are thus subject to export control regulations.

This also concerns software developed on supercomputers, as it potentially falls under export control, according to Regulation (EC) No. 428/2009.[24] For instance, in item 4D001 of the appendix software is listed that is designed for the use of export-controlled equipment. Supercomputers do typically contain export-controlled equipment.

For this reason, all users of the HBP PCP pilot systems have to sign a usage agreement that includes the following clause: "I am fully aware that scientific results gained through the use of supercomputers may be liable to European export control regulations".

[24] http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32009R0428