SEVENTH FRAMEWORK PROGRAMME

FP7-ICT-2013-10

DEEP-ER

DEEP Extended Reach

Grant Agreement Number: 610476

D1.3

Progress report at month 12

Approved

Version:	2.0
Author(s):	E.Suarez (JUELICH)
Contributor(s):	S.Eisenreich (BADW-LRZ), N.Eicker (JUELICH), H.Ch.Hoppe (Intel), V.Beltran (BSC), D.Alvarez (JUELICH))
Date:	21.10.2014

Project and Deliverable Information Sheet

<table>
<tr><td rowspan="7">DEEP-ER Project</td><td colspan="2">Project Ref. №: 610476</td></tr>
<tr><td colspan="2">Project Title: DEEP Extended Reach</td></tr>
<tr><td colspan="2">Project Web Site: http://www.deep-er.eu</td></tr>
<tr><td colspan="2">Deliverable ID: D1.3</td></tr>
<tr><td colspan="2">Deliverable Nature: Report</td></tr>
<tr><td>Deliverable Level:
CO *</td><td>Contractual Date of Delivery:
30 / September / 2014</td></tr>
<tr><td>Actual Date of Delivery:
30 / September / 2014</td></tr>
<tr><td colspan="3">EC Project Officer: Panagiotis Tsarchopoulos</td></tr>
</table>

* - The dissemination level are indicated as follows: **PU** – Public, **PP** – Restricted to other participants (including the Commission Services), **RE** – Restricted to a group specified by the consortium (including the Commission Services). **CO** – Confidential, only for members of the consortium (including the Commission Services).

Document Control Sheet

<table>
<tr><td rowspan="6">Document</td><td colspan="2">Title: Progress report at month 12</td></tr>
<tr><td colspan="2">ID: D1.3</td></tr>
<tr><td>Version: 2.0</td><td>Status: Approved</td></tr>
<tr><td colspan="2">Available at: Publishable part at: http://www.deep-er.eu</td></tr>
<tr><td colspan="2">Software Tool: Microsoft Word</td></tr>
<tr><td colspan="2">File(s):DEEP-ER_D1.3_Periodic_progress_report_M12_v2.0-ECapproved</td></tr>
<tr><td rowspan="4">Authorship</td><td>Written by:</td><td>E.Suarez (JUELICH)</td></tr>
<tr><td>Contributors:</td><td>S.Eisenreich (BADW-LRZ), N.Eicker (JUELICH), H.Ch.Hoppe (Intel), V.Beltran (BSC), D.Alvarez (JUELICH))</td></tr>
<tr><td>Reviewed by:</td><td>J.Morillo (BADW-LRZ), H.Ch.Hoppe (Intel). I.Schmitz (ParTec)</td></tr>
<tr><td>Approved by:</td><td>BoP/PMT</td></tr>
</table>

Document Status Sheet

Version	Date	Status	Comments
1.0	30/September/2014	Final	EC submission
2.0	21/October/2014	Approved	EC approved

Document Keywords

Keywords:	DEEP-ER, HPC, Exascale, progress report, month 12

Table of Contents

Project and Deliverable Information Sheet 1
Document Control Sheet 1
Document Status Sheet 2
Document Keywords 3
Table of Contents 4
List of Figures 5
Executive Summary 6
1 Publishable summary 8
1.1 Project objectives 8
1.2 Work performed and main results 11
1.3 Expected final results 20
Annex A 21
A.1 Listing of dissemination activities 21
List of Acronyms and Abbreviations 24

List of Figures

Figure 1: Participants to the DEEP-ER‘s kick-off meeting, Jülich, 2nd October 2013. 12

Figure 2: Group picture taken during the joint Workshop of the European Exascale Projects, 18th March 2004, Edinburgh (UK).....14

Figure 3: High-level view of the DEEP-ER Prototype. BN=Booster Node; CN=Cluster Node; NVM=Non-Volatile Memory; NAM=Network Attached Memory 16

Figure 4: Sketch of DEEP-ER I/O software layers. 17

Figure 5: Sketch of DEEP-ER resiliency layers.....18

Executive Summary

The DEEP – Extended Reach (DEEP-ER) project started on 1st October 2013 and will last three years. The project addresses two significant Exascale challenges: the growing gap between I/O bandwidth and compute speed, and the need to significantly improve system resiliency. DEEP-ER extends the Cluster-Booster Architecture first realised in the DEEP project by a highly scalable I/O system. Additionally, an efficient mechanism to recover application tasks that fail due to hardware errors will be implemented. The project will build a hardware prototype including new memory technologies to provide increased performance and power efficiency. As a result, I/O parts of HPC codes will run faster and scale up better. Furthermore, HPC applications will be able to profit from checkpoint and task restart on large systems reducing overhead seen today. To demonstrate it a set of seven applications with high societal impact are ported to the DEEP-ER prototype and make use of the I/O and resiliency capabilities available therein.

This report describes the objectives, work performed, resources used, and results achieved during the **first twelve months** of the DEEP-ER project. The main achievements in the reporting period are enumerated below:

- The project's organisation structure established, setting up all its bodies and selecting their members.
- Procedure to control the quality of reports and deliverables defined.
- Procedure for dissemination of the DEEP-ER results defined.
- Co-design effort initiated: application requirements with respect to hardware and software systems gathered, discussed, and fed into the corresponding developments.
- Overall DEEP-ER system architecture defined taking into account the application requirements.
- Detailed analysis of interconnect choices and proposal of a preferred and a fall back interconnect
- First prototypes of Non-Volatile Memory devices available and under test.
- Hybrid memory cube device (prototype for Network Attached Memory) available and under test.
- Design of the I/O software layer, taking into account the application requirements, completed: functionality to be provided by each of the three I/O APIs involved in the project –BeeGFS[1], SIONlib, and E10– identified; interfaces between them defined.
- Design of the resiliency software layer completed taking into account the application requirements: first sketch of abstraction layer below the user-level checkpoint/restart software finished; interplay between application-based checkpoint/restart and task-based resiliency agreed.
- Interface between I/O software and application-based checkpoint/restart software defined.
- Initial list of benchmarks to evaluate the I/O and resiliency software developments identified. First tests already implemented in the JuBE environment.
- Training workshop for application and middleware software developers organised.

[1] Formerly called FhGFS.

- Analyse the structure and main characteristics of the DEEP-ER applications and measure their performance in a standard cluster for later reference. First code optimisations already implemented.
- Dissemination of project goals and status in various workshops and conferences, amongst others the Supercomputing Conference (SC13) and the International Supercomputing Conference (ISC'14).
- Co-organisation and participation on the "European Exascale Projects Workshop 2014", jointly with the EU-projects DEEP, CRESTA, Mont-Blanc (1 and 2), EPiGRAM, EXA2CT, and NUMEXAS (Edinburgh, March 2014).

1 Publishable summary

The DEEP-ER project tackles two important Exascale challenges. Firstly, the increasing gap in the growth rate of compute power with respect to the amount and performance of memory and storage available in HPC systems. Secondly, the high failure rates expected in Exascale systems as a consequence of the increased number of components and the need to take their performance and energy efficiency to the limits. To tackle those issues, DEEP-ER will extend the heterogeneous Cluster-Booster Architecture implemented by the DEEP[2] project with additional I/O and resiliency functionalities.

DEEP-ER targets a seamless integration of a high-performance I/O subsystem into the Cluster-Booster Architecture. New memory technologies will be used to provide a multi-level I/O infrastructure capable of supporting data-intensive applications. Additionally, an efficient and user-friendly resiliency concept combining user-level checkpoints with transparent task-based application restart will be developed, to enable applications coping with the higher failure rates expected in Exascale systems.

DEEP-ER's I/O and resiliency concepts will be evaluated using seven HPC applications from fields that have proven the need for Exascale resources. These applications will be ported and optimised to demonstrate the usability, performance and resiliency of the DEEP-ER Prototype. Systems that use the DEEP-ER results will be able to run more applications increasing scientific throughput, and the loss of computational work through system failures will be substantially reduced.

1.1 Project objectives

The specific objectives of the DEEP-ER project and first results are:

1. Address two main Exascale challenges: I/O and resiliency. DEEP-ER will extend the DEEP Architecture by: i) a highly scalable, efficient and easy-to-use parallel I/O system; ii) providing a combination of low-overhead user-level checkpoint/restart and automatic task recovery.

   → The design of the DEEP-ER I/O software layer has been completed taking the application and resiliency software requirements into account (documented in deliverable D4.1). The interfaces between the three I/O APIs – BeeGFS, SIONlib, and Exascale10 (E10)– have been defined and documented in D4.2.

   → The resiliency software layer has been designed (documented in D5.1) taking the application requirements into account. First sketch of abstraction layer below the user-level checkpoint/restart software finished.

   → The interplay between application-based checkpoint/restart and task-based resiliency agreed.

2. Develop a prototype system of the extended DEEP Architecture that leverages advances in hardware components (Intel's second generation Intel® Xeon Phi™ processors, high-speed interconnects and non-volatile memory devices) to further improve the performance and efficiency of the DEEP-ER Prototype and realise the

[2] www.deep-project.eu

novel I/O system and resiliency improvements. This prototype will allow proving the viability of the concept for 500 Petaflop-class of supercomputers.

→ An initial DEEP-ER hardware architecture has been defined as documented in D3.1. This architecture takes into account the requirements of applications, of the I/O subsystem and of the distributed checkpoint/restart system. It is based on a two-level hierarchy – a “Brick” combines multiple CPU, NVM and interconnect devices using PCI Express, and the Bricks communicate with each other over a high-performance interconnect. Details are given in D3.2.

Detailed investigations into the implementation of that architecture and the design of its components and the associated risks have led to the adoption of a simplified architecture based on a compute board and PCI Express attached NVM and NIC devices. The changed architecture allows reaching all the goals of the project. The design of its components is discussed in D3.2.

3. Explore the potential of new storage technologies (non-volatile and network attached memory) for use in HPC systems, with a focus on parallel I/O and system resiliency by integrating them with the DEEP-ER Prototype.

   → The suitable non-volatile memory (NVM) technology to be used on the Booster Nodes has been chosen (see D3.1). Two NVM devices implementing the NVM Express interface have been installed at Jülich and are currently under test. First results from performance measurements are available. It is proposed that the DEEP-ER Prototype will use a product implementation of this technology, as explained in D3.2.

   → The architecture and functionality of the network-attached memory (NAM) have been discussed and summarised in a first proposal (see D3.1). A first design of a NAM based on Hybrid Memory Cube is ready, and the associated Hybrid Memory Cube controller is nearing completion.

4. Develop a highly scalable, efficient and user-friendly parallel I/O system tailored to HPC applications. The system will exploit innovative hardware features, optimise I/O routes to maximise data reuse, and expose a user friendly interface to applications. Its design will meet the requirements of traditional, simulation-based as well as emerging data-intensive HPC applications.

   → Discussions between the software developers within WP4 have served to define the interfaces between the three software packages involved in the implementation of the parallel I/O system: the Fraunhofer file system (BeeGFS), the parallel I/O library SIONlib, and the Exascale10 (E10) software stack.

   → The design of the DEEP-ER I/O system has been completed taking into account the outcome of the discussions with the experts from WP3 –to guarantee that the hardware provides the needed functionality– and with WP5 and WP6 –to gather all their requirements on the I/O infrastructure–.

   → The functionalities that each of the three I/O software packages must provide to the project, the interplay between them, and their interfaces have been described in deliverables D4.1 and D4.2.

→ A list of benchmarks to be used for the evaluation of the DEEP-ER I/O software has been identified. The integration of some of those benchmarks into the JuBE benchmark environment has already started.

5. Develop a unified user-level system that significantly reduces the checkpointing overhead by exploiting multiple levels of storage and new memory technologies. Extend the DEEP programming model to combine automatic re-execution of failed tasks and recovery of long-running tasks from multi-level checkpoints, and introduce easy-to-use annotations to control checkpointing.

   → Discussions between the software developers within WP5 took place to define the overall resiliency concept, based on user-level checkpoint/restart and task-based failure recovery. The role to be played by each of the two components and the interaction between them has been determined.

   → The functionality to be provided by the underlying resiliency abstraction layer has been specified and a first sketch of this layer has been prepared.

   → A co-design discussion has been driven to feed into the development of the resiliency software the requirements from the WP6 application developers, to inform WP4 of the needed I/O support, and to study the usability for multi-level checkpointing of the hardware developed by WP3.

   → A failure model has been designed and is currently being implemented. The goal is to optimise policies that determine for each application the frequency, redundancy level and storage-location of each checkpoint.

   → Deliverable D5.1, containing the design details, has been submitted on time.

6. Analyse the requirements of HPC codes carefully selected to represent the needs of future Exascale applications with regards to I/O and resiliency, guide the design of the DEEP-ER hardware and software components, optimise these applications for the extended DEEP Architecture and use them to evaluate the DEEP-ER Prototype. Selected applications cover the fields of Health, Earthquake Physics, Radio Astronomy, Oil Exploration, Space Weather, Quantum Physics, and Superconductivity.

   → A questionnaire was prepared and sent to the application developers to gather the relevant information and application requirements needed in WP3, WP4, and WP5.

   → A co-design meeting took place in January 2014 in Jülich to discuss on the answers provided by the application developers to the questionnaire, clarifying open issues.

   → The co-design discussion continued at the application review meetings in February and September 2014, during the project F2F meetings at Kaiserslautern (Germany), and Havant (UK), respectively.

   → The structure of the applications has been analysed and first performance and scaling tests have taken place. Optimisations have been implemented in some of the applications to address performance issues discovered during the previously mentioned measurements.

7. Demonstrate and validate the benefits of the extended DEEP Architecture and its first implementation (the DEEP-ER Prototype) with the DEEP-ER pilot applications and for applications that exploit generic multi-scale, adaptive grid and long-range force parallelisation models.

1.2 Work performed and main results

During its **first twelve months**, the DEEP-ER project has achieved almost all the milestones foreseen for the present period by the Description of Work (DoW). The only exception is MS5, for the reasons described below:

- MS1: Management structure and bodies established during the kick-off-meeting
- MS2: Quality control and dissemination plan defined, as documented in deliverables D1.1 and D2.1, both submitted in month 3. Project website (www.deep-er.er) online and periodically updated to include new events and job offers.
- MS3: Hardware architecture and software specifications defined. Results are documented in deliverables D3.1, D4.1, and D5.1, submitted in month 6.
- MS4: Hardware component design (and associated risks) investigated in depth. This has led to a revision of the architecture specification and is documented in D3.2, which was submitted on time.
- MS5: The Booster CPU evaluator planned for M12 is not yet available due to a delay in the availability of the second generation Intel® Xeon Phi™ (code-named Knights Landing - KNL). This is the consequence of delays in the KNL hardware development, which are out of the control of the DEEP-ER project. Immediate impact of the delay on the DEEP-ER project is limited, since the software from WP4, WP5, and WP6 can continue on standard clusters (such as the DEEP Cluster) and on the Knights Corner (KNC) platforms available to the project partners. KNL specific platform features (like the new AVX-512 instruction set extension and the on-package high bandwidth memory) can be tried out using available SW and HW evaluators.
- MS6: The structure of the applications has been analysed, as documented in D6.1.

Management, legal and administrative tasks

The DEEP-ER consortium consists of 13 Partners and three additional Third Party Partners. The management of this relatively large consortium requires legal and administrative regulations. The Consortium Agreement, signed by all DEEP-ER Partners, is the legal frame that describes the rights and responsibilities of all Partners in the consortium. With it in place, the Grant Agreement with the European Commission was signed on 3rd September 2013. About three weeks later, on the 1st October 2013, the DEEP-ER project officially started.

Figure 1: Participants to the DEEP-ER's kick-off meeting, Jülich, 2nd October 2013.

DEEP-ER's effective kick-off was the face-to-face meeting of the whole consortium that took place in Jülich (Germany) on the 2nd October 2013 (see Figure 1). This meeting gave the DEEP-ER members the opportunity to meet each other in person and identify the people with whom they will work most closely in the next three years. The motivation and goals of the project were presented; the project management structure with all its bodies was established; the procedure for quality control of the deliverables was introduced; and the tasks to be performed in all Work Packages (WPs) during the first six months of the project were discussed, assigning responsibilities and contact persons for each subject. During the last part of the meeting parallel sessions on hardware, software, and application topics were organised to allow for more detailed internal discussions. Consortium face-to-face meetings are organised on a regular basis every six months. The location is always at the site of a different project partner to spread the organisational workload over the whole consortium. The second consortium meeting took place on the 25th – 26th February 2014 in Kaiserslautern (Germany), and the third one on 3rd – 4th September 2014 in Havant (UK). In both meetings the status of the project and of the deliverables due at each time were discussed. The face-to-face meetings were also the scenario of detailed review of the application status, and of co-design discussions between WP3, WP4, WP5 and WP6. Furthermore, monthly teleconferences of the Team of Work Package leaders (ToW) and bi-weekly teleconference of the Design and Development Group (DDG) were organised to periodically discuss the overall progress of the project.

One of the main goals for the Project Management Team in the first months of the project was to trigger and foster an early start of work and discussions on all the Work Packages (WPs). For this purpose, additionally to the meetings mentioned above, four face-to-face meetings of the DDG were organised. The DDG is the board of technical experts that takes the design decisions in the DEEP-ER project. The first one took place on the 11th November 2013 in Mannheim (Germany) and served for hardware and software experts in WP3, WP4, and WP5 to discuss and exchange ideas. Overlaps and inconsistencies between the design plans of the three work packages were identified. Out of this discussion a first overall design concept was sketched, which was then further developed through internal WP discussions

and through the regular DDG teleconferences. The remaining three DDG face-to-face meetings took place in Jülich (Germany). The second on the 23rd-24th January 2014 included hardware, software and also application experts. In advance to that meeting a questionnaire to the application developers had been prepared to identify all their requirements and feed them into the hardware and software designs. During the meeting, the answers to the questionnaires were discussed and open questions were clarified. The third meeting, focused on the hardware design and the impact of the KNL delay on the project, took place on 2nd June 2014. The fourth and final one, on 21st August 2014, served for WP4 and WP5 to discuss the I/O functionality needed by the application-based checkpoint/restart software, in particular for buddy-checkpoints, and the nature of the interface provided by WP4 for that purpose.

To guarantee the quality of Deliverables and Reports an internal review process has been established. One or two people are selected from each partner as internal reviewers. Before its submission to the European Commission, each Deliverable is reviewed by one internal reviewer plus one member of the PMT. Reasonable internal deadlines for this reviewing process have been set, to be on time with the submission deadline given by the European Commission. All Deliverables foreseen for this reporting period on the DoW were timely submitted to the European Commission after having passed through the mandatory DEEP-ER internal review process.

Dissemination, training and outreach

The implementation of the innovative DEEP-ER I/O and resiliency concepts constitutes a challenge that will require the development of new techniques and tools never tested before. Access to the know-how achieved in this process shall not remain limited to the group of people directly involved in the project, but must be made available for a wider community. For this reason, WP2 in DEEP-ER is entirely devoted to the dissemination of the knowledge accumulated along the project's duration, and to train the users on its application.

The centre of the dissemination activities of DEEP-ER is its web site www.deep-er.eu[3]. The web page is updated regularly and referred to in all other materials (articles, press releases, brochures, presentations, etc.). It is used to publish general information about the project, current activities, training opportunities, job vacancies, publications, tutorials, success stories, and achievements of the project.

Two social media platforms have been chosen to disseminate DEEP-ER news amongst the HPC world and the general public: LinkedIn and Twitter. The already existing DEEP LinkedIn group has been extended to host also DEEP-ER. The strong link existing between both projects justifies the use of a single group. The same applies for Twitter. Ahead of the ISC'14 a DEEP/DEEP-ER joint twitter account was online. Posts are being regularly posted (at least at bi-weekly basis) and frequently re-posted by other Twitter users in the HPC community. The most recent Twitter posts are visible also at the main page of DEEP-ER's website.

Several dissemination activities have taken place since the start of the project. Partners from the DEEP-ER consortium presented the project's concept in conferences and workshops, including two of the most important events in the HPC community: the Supercomputing

[3] The domain www.deep-er-project.eu has been also reserved and leads to the same location.

Conference (SC) that took place in Denver (USA) in November 2013, and the International Supercomputing Conference (ISC'14) that took place in Leipzig (Germany) in June 2014.

On SC13 the DEEP project co-organised, together with the two European Exascale Projects (EEP) CRESTA and Mont-Blanc, a joint booth. There, a DEEP-ER flyer describing the goals and most important aspects of the project was distributed. Questions from the visitors were answered and, taking advantage of the strong link between DEEP and DEEP-ER, ideas from both projects were explained to the public. Additionally DEEP-ER extensions to the Cluster-Booster Architecture were mentioned during the BoF (Birds of a Feather) session organised by the three original European Exascale Projects.

On ISC'14, the DEEP-ER project co-organised and was present at the "European Exascale Projects" booth, where the project goals and current results were presented to the interested visitors. This time DEEP-ER had full presence at the booth, with an own wall describing the architecture and main goals of the project. Additionally, a flyer was prepared and distributed at the EEP booth and at the booths of other project partners. DEEP-ER was also presented in an EEP joint BoF session. Interviews with various HPC journalists (insideHPC, Scientific Computing World, etc.) took place during ISC'14 and helped, together with the frequent Twitter and LinkedIn posts, to spread the word on the focus and achievements of the DEEP-ER project. In addition, the DEEP-ER project was shown on the Intel booth, with focus on the architecture innovations and the applications.

Figure 2: Group picture taken during the joint Workshop of the European Exascale Projects, 18th March 2004, Edinburgh (UK)

DEEP-ER is actively participating in the preparation of further joint activities with the European Exascale Projects community. A joint workshop with CRESTA, DEEP, DEEP-ER, EESI-2, EPiGRAM, EXA2CT, Mont-Blanc, Mont-Blanc 2, and NUMEXAS took place in Edinburgh on 18th-19th March 2014. There an overview of the DEEP and DEEP-ER projects was given and key aspects of the programming environment and tools available on the platform were presented.

A list with all dissemination activities performed in the present reporting period is given in Annex A.1 of this report.

To foster cooperation activities with European industry and European R&D organisations, a liaison programme between the DEEP-ER project and relevant industrial and business partners will be established. For this purpose DEEP-ER is represented in the ETP4HPC and PROSPECT meetings, where key European HPC industry players participate, as a vehicle to promote the use of multi-Petascale to Exascale systems to industrial and academic users.

Training the community on how to use the software and hardware developed in DEEP-ER is also an important part of the project. The main goal of the training events in DEEP-ER is to teach the application developers participating in the project on how to use the software tools and programming environment that will run on the DEEP-ER Prototype and other intermediate hardware evaluators. DEEP-ER application and software developers participated on the first DEEP-ER Training Workshop that took place from 13th to 14th February 2014 in Barcelona (Spain). There, the OmpSs programming environment was introduced and an overview with best practice guides on the use of Intel® Xeon Phi™ processors (including Knights Landing – KNL) was given. The training was coupled to a VI-HPS training on performance analysis tools organised by the BSC on 10th–12th February 2014 in the same location. DEEP-ER application developers participated also on the VI-HPS training to learn how to use tools such as Extrae/Paraver and Scalasca to analyse the structure and bottlenecks in their codes. Both the VI-HPS and the DEEP-ER specific trainings included lectures and numerous hands-on sessions.

Technical Work

The technical work in DEEP-ER is grouped into three main topics: system hardware, system software (including I/O and resiliency software), and applications.

Overview

The DEEP-ER project designs and builds a second-generation prototype (see Figure 3) of the Cluster-Booster Architecture. In the DEEP-ER Prototype the second generation Intel® Xeon Phi™ processors (KNL) provides the compute power of the Booster Nodes, while the most recent Intel® Xeon® processors populate the Cluster Nodes. A uniform high-speed interconnect runs across Cluster and Booster, and network-attached memory (NAM) devices connected to it provide high-speed shared memory access. The Booster Nodes themselves also feature additional non-volatile memory (NVM) capabilities.

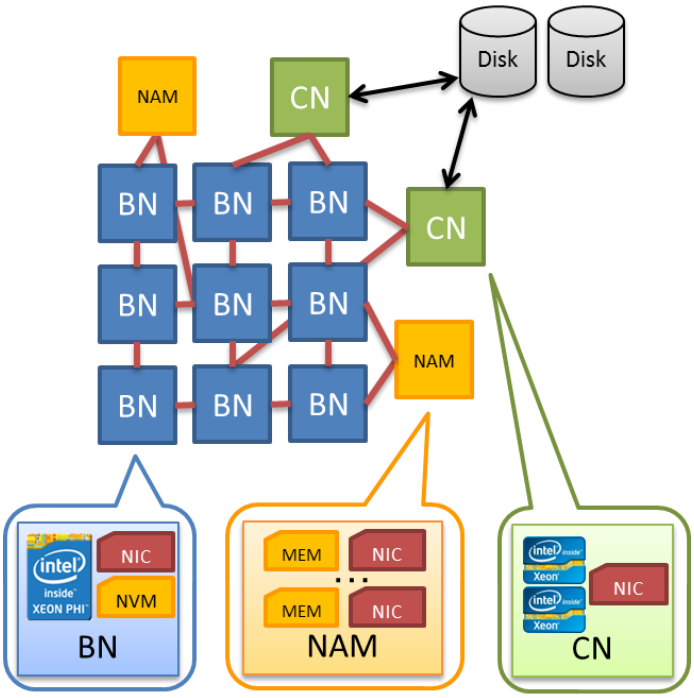

Figure 3: High-level view of the DEEP-ER Prototype. BN=Booster Node; CN=Cluster Node; NVM=Non-Volatile Memory; NAM=Network Attached Memory

The DEEP-ER multi-level I/O infrastructure has been designed to support data-intensive applications and multi-level checkpointing/restart techniques. The project will develop a scalable and efficient I/O software platform based on the Fraunhofer parallel file system (BeeGFS), the parallel I/O library SIONlib, and the I/O software package Exascale10 (E10). It aims to enable an efficient and transparent use of the underlying hardware and to provide all functionality required by applications for standard I/O and checkpointing.

On top of this I/O infrastructure DEEP-ER will develop an efficient and user-friendly resiliency concept combining user-level checkpoints with transparent task-based application restart. OmpSs is used to identify application's individual tasks and their interdependencies. The OmpSs runtime will be extended in DEEP-ER in order to automatically re-start tasks in the case of transient hardware failures. In combination with a multi-level user-based checkpoint infrastructure to recover from non-transient hardware-errors, applications will be able to cope with the higher failure rates expected in Exascale systems. DEEP-ER's I/O and resiliency concepts will be evaluated using seven HPC applications from fields that have proven the need for Exascale resources.

System Hardware

On the hardware side, in the first twelve months of the project requirements from application and system level software have been gathered, architecture trade-offs with WP4, WP5 and WP6 have been discussed, and a design for the hardware architecture of the DEEP-ER Prototype has been derived.

In Deliverable D3.1, a novel two-level architecture was proposed: multiple CPU, NVM and NIC devices would be connected by PCI Express and form a "Brick"; multiple Bricks connected via the high-performance DEEP-ER fabric would then form the Booster. This architecture allows a tight integration of components using only standard interfaces, and thus provides flexibility and isolation from technical risks.

In the next step, the design of the principal components (Brick midplane board, CPU board, NVM device, NAM) was investigated in significant detail, and a thorough risk analysis was performed. Unfortunately, an internal evaluation of the involved risks lead Eurotech to the decision not to pursue the planned development of a custom, PCI Express form factor KNL CPU board. This will be mitigated by using conventional, server or workstation form factor KNL boards (from Intel or an OEM). NVM and NIC will still be connected via PCI Express, yet there is no use for the Brick midplane board anymore and therefore the DEEP-ER Prototype will follow a more conventional, air-cooled cluster approach. The requirements of the DEEP-ER applications will still be satisfied in full and all stated objectives of DEEP-ER are achievable. The impact of the architecture change is limited to density of the DEEP-ER Prototype, use of air instead of liquid cooling, and reduced flexibility to experiment in the future with associations of CPU:NVM:NIC other than 1:1:1.

A detailed investigation of interconnect fabrics available today or announced for end of H1/2015 at the latest has led to a proposal of a preferred choice (EXTOLL) and a fall back choice (Mellanox Infiniband). For the NVM device, a line of PCI Express connected NVM products with an NVM Express interface were selected. The NAM design has been completed, based on implementing the NAM-specific functionality in an FPGA and using a custom-designed Hybrid Memory Cube controller.

Full details on the hardware architecture for the DEEP-ER Prototype are given in Deliverables D3.1 and D3.2, while the component design is presented in D3.2. Both Deliverables were submitted to the EC in time.

System Software

On the software side, in the first twelve months of the DEEP-ER project the focus was on the definition of the details of the I/O and resiliency software environments and the interplay between the involved software components.

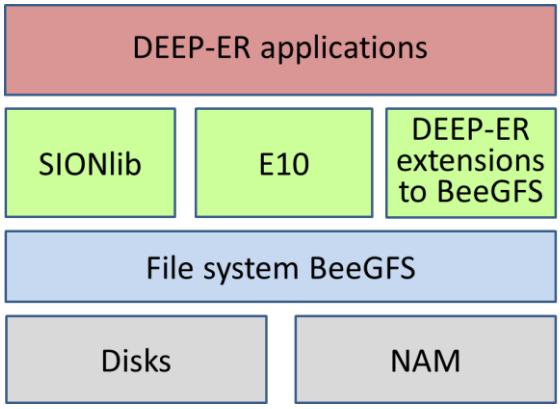

Figure 4: Sketch of DEEP-ER I/O software layers.

The overviews of the DEEP-ER I/O and resiliency software layers have been described in the DoW and are shown in Figure 4 and Figure 5, respectively. The different Work Packages involved in the development or directly affected by the characteristics of this software (DDG, WP4, WP5 and WP6) have engaged in numerous discussions to define the details of all the I/O and resiliency software layers. The requirements from the application developers from WP6 were gathered through a questionnaire and used by the developers in WP4 and WP5 to determine which features are needed in the final middleware.

Based on the results of the questionnaire the overall architecture of the DEEP-ER I/O system has been sketched in Deliverable D4.1. It will employ three different technologies that shall act in a complementary way within the project. As a global parallel file system BeeGFS will be utilised to provide the common basis of I/O operations in the context of DEEP-ER. On top of it, SIONlib will act as a compound-layer to allow the DEEP-ER applications to make use of global BeeGFS most efficiently with minimal effort of adaptation. This reflects the fact that currently almost all applications within the projects conduct task-local I/O. Last but not least E10 addresses the so-called small I/O problem. Here the ever-increasing level of parallelism results in a huge number of small I/O operations in order to achieve the global access to file data in Exascale parallel applications. Besides the actual applications a main consumer of these capabilities will be the multi-level checkpoint scheme to be developed by WP5.

The interfaces between the three mentioned I/O components and between them and the resiliency software from WP5 have been defined and are documented in D4.2. With this step, the overall design of the system software is completed and the implementation of the developments needed in each of the components has already started for some of them.

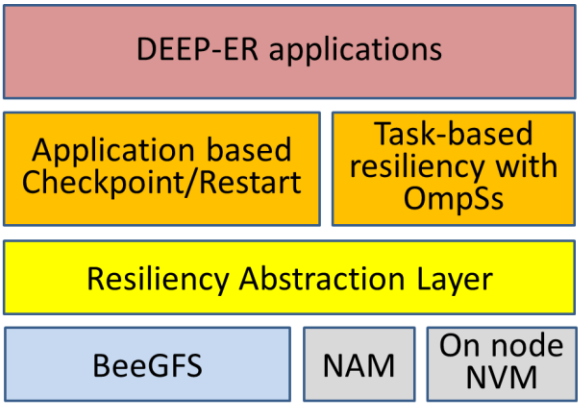

Figure 5: Sketch of DEEP-ER resiliency layers

The DEEP-ER resilience architecture is based on user-level checkpoint/restart techniques – which provide a high level of resiliency and are the most cost-effective in terms of I/O requirements– complemented with novel OmpSs task-based recovery techniques. With this combination, DEEP-ER develops new resiliency features to isolate partial failures of the system without requiring a full application restart, resulting in a more resilient, fine-grained and flexible architecture. ParaStationMPI will be extended to detect, isolate and clean up failures of MPI-offloaded tasks, which are then independently restarted, avoiding the need of a full application recover. Moreover, both user-based and task-based techniques will be adapted to take advantage of the advanced I/O technologies and storage hierarchy provided by WP4 and WP3, respectively. Furthermore, the interface and functionality to be provided by the I/O software has been agreed with WP4. The interface with WP4 happens mostly on the application-based checkpoint/restart part, while the task-based resiliency happens at a higher level and is rather decoupled from the rest.

Finally, a failure model is being developed to optimise polices that determine for each application the frequency, redundancy level and storage-location of each checkpoint. This model will take into account the probability and type of failure of the main hardware and software components, the performance of the user-based and task-based checkpoint/restart implementations on the DEEP-ER hardware architecture as well as the application specific

characteristics. Deliverable D5.1 describes the details of the resiliency concept that will be developed in the context of the DEEP-ER project.

The software developments in WP4 and WP5 are accompanied by benchmarking activities to document the progress in terms of performance and functionality. The Jülich Benchmarking Environment (JuBE) is used for this purpose. Benchmarks, proxy-applications, and full-fledged applications (from WP6) will be integrated into JuBE along the project duration. An initial list of benchmarks to be used in the project has been already defined. Several JuBE workshops have taken place to train the benchmarking team on the use of the tool. The first I/O benchmarks have been integrated in JuBE and first tests have been done on the DEEP Cluster to measure BeeGFS's baseline performance.

Applications

The application developers play a crucial role in the DEEP-ER project. Their work is two-folded: on the one hand they validate the work done by other technical work packages by porting their applications to the DEEP-ER prototype; on the other hand, their input drives the development and future of both hardware and software architectures.

In order to provide feedback to the project, application developers gave a first description of their application, with focus on current and future I/O behaviour. Darshan traces were collected, and results and application structures were analysed in the two face to face DDG meetings that took place in Mannheim and Jülich, respectively, where representative members from WP3, WP4 and WP5 were present. Further co-design discussions to gather more specific application requirements have taken place in the consortium face-to-face meetings.

Furthermore, application developers participated in the Barcelona training to familiarise themselves with the use of performance analysis tools, OmpSs, and Intel® Xeon Phi™. After that the developers started to analyse the structure of the applications, as well as their scalability and performance characteristics. This knowledge is critical to optimise the applications and later be able to take the right decisions on how to distribute the applications between Cluster and Booster in the DEEP-ER Prototype, and on how to make optimal use of the hardware and software functionality available.

Given the importance of the applications in the project, internal review meetings focused on the work done by WP6 took place during the consortium face-to-face meetings, both in Kaiserslautern and in Havant. There, the developers described their applications and presented the results that they had obtained until then, as well as the planned next steps. Other members of the consortium not involved in WP6 acted as internal reviewers and gave recommendations on the measures to be taken by each application team to achieve the results needed by the project.

Regarding the application development itself, in the first year of the DEEP-ER project the focus has been on analysing the structure of the applications, identifying their various phases, and measure the performance and scalability of each of these phases. With this information, first studies on the optimal distribution of the applications over the Cluster and Booster parts of the system have been done. The checkpointing strategies and I/O requirements from each application have been identified. The performance measurements have also served to identify potential bottlenecks, what lead to the implementation of code

improvements in several codes. All these activities have been documented in D6.1, submitted in month 12 simultaneously to the present progress report.

1.3 Expected final results

The DEEP-ER project will have installed the DEEP-ER Prototype in Jülich (Germany), containing the new generation of Intel® Xeon Phi™ processors, non-volatile memory in the Booster Nodes, as well as additional memory connected to the network. A complete software stack based on ParaStation MPI and OmpSs will run on the machine, providing parallel I/O functionalities and an efficient infrastructure for fail recovery via easy-to-use application interfaces.

Porting and optimising applications on the DEEP-ER Prototype will have demonstrated the scalability and performance of the I/O and resiliency tools developed within the project. The experience gathered will have served to demonstrate that systems using the DEEP-ER results will be able to run more applications increasing scientific throughput, and the loss of computational work through system failures will be substantially reduced.

Annex A

A.1 Listing of dissemination activities

This list reflects the disseminations activities performed **between months 1 and 12** of the DEEP-ER project.

1.3.1.1 Conferences, workshops, and meetings:

- **ScalPerf'13,** Bertinoro (Forli-Cesena), Italy, September 22-27, 2013.
  - B.Mohr (JUELICH), "Jülich – still – on the way to Exascale" (presentation)

- **European Research & Innovation Conference 2013" (ERIC 2103),** Nice, France, October 23, 2013:
  - N.Eicker (JUELICH): "The DEEP-ER Project - Extending the reach of the Cluster-Booster Architecture" (presentation)

- **HBP Summit, Lausanne,** Switzerland, October 9, 2013
  - N.Eicker (JUELICH): "DEEP and DEEP-ER – Booster for HPC" (presentation)

- **SC13,** Denver, USA, November 17-22, 2013:
  - N.Eicker (JUELICH), "DEEP and DEEP-ER: Innovative Cluster architecture for Intel® Xeon Phi™" (Intel Theatre presentation at the Intel booth, November 18, 2013)
  - N.Eicker (JUELICH), "The DEEP Project" (presentation at BoF session: "Building on the European Exascale Approach", November 19, 2013)
  - H.Ch.-Hoppe (Intel), E.Suarez (JUELICH), "DEEP and DEEP-ER – Innovative Cluster Architectures for Intel® Xeon Phi™" (repeated presentations at the Intel booth, November 18-21, 2013)
  - E.Suarez (JUELICH), DEEP and DEEP-ER presentation and discussion at the Panel "Emerging Technologies and Big Data (Euro-Centric)", November 21, 2013
  - Joint booth of the "European Exascale Projects", joining DEEP, CRESTA, and Mont-Blanc
  - DEEP-ER flyer distributed at the partners' booths
  - N.Eicker (JUELICH) interview to Computer World
  - E.Suarez (JUELICH) video interview on DEEP and DEEP-ER, as part of the "Discover Your Parallel Universe video project" series, November 21, 2013.

- **HiPEAC Conference 2014,** Vienna, Austria, January 20-22, 2014
  - Presentation of the DEEP-ER project at the Eurotech booth

- **Internal JSC PoF Begutachtung, Juelich**, Germany, March 11, 2014
  - E.Suarez (JUELICH): "The DEEP and DEEP-ER projects: co-design aspects" (presentation)
  - E.Suarez (JUELICH): "The DEEP and DEEP-ER projects: The Cluster-Booster Architecture" (poster)

- **Joint Workshop of the European Exascale Projects, Edinburgh, March 18-19, 2014**
  - E.Suarez: "DEEP and DEEP-ER" (presentation)

- **G8 ECS internal workshop,** Kobe, K-computer facility, March 26-27, 2014
  - J.Labarta (BSC): “Behind DEEP and Mont-Blanc” (presentation)
- **Lustre User Group (LUG) Conference** (http://www.opensfs.org/lug14/), Miami, Florida, April 8-10, 2014.
  - Exascale10 work in DEEP-ER selected for presentation
- **International Supercomputing Conference ISC’14,** Leipzig, Germany, June 23-26, 2014:
  - Joint booth of the European Exascale Projects (EEP). Booth #833. Participant projects: DEEP, DEEP-ER, Mont-Blanc, CRESTA, EPiGRAM, EXA2CT, and NUMEXAS).
  - N.Eicker (JUELICH), “The DEEP and DEEP-ER projects” (presentation at the joint BoF of the European Exascale Projects)
  - E.Suarez (JUELICH), N.Eicker (JUELICH), P.Arts (Eurotech), J.Schmidt (UniHD), H.Ch.Hoppe (Intel), “DEEP & DEEP-ER Updates at ISC” (Video-Interview with “Inside HPC”). Online in: http://insidehpc.com/2014/07/video-deep-deep-er-project-updates-isc14/
  - E.Suarez (JUELICH), “Architectural Approaches to Energy Efficiency Exascale” (Interview with “Scientific Computing World”). Online in: http://www.scientific-computing.com/news/news_story.php?news_id=2527
  - E.Suarez (JUELICH), “DEEP-ER into Exascale” (presentation at the Intel booth)
  - E.Suarez (JUELICH), “DEEP and DEEP-ER” (presentation at Cross-Lab Workshop of Intel’s European Exascale Labs)
  - H.Ch.-Hoppe (Intel), Demonstration at the Intel booth, showing DEEP and DEEP-ER results, with special focus on the space weather application from KULeuven and the simulation of the effect of electromagnetic fields on human tissues (Inria).
  - DEEP and DEEP-ER fliers distributed at the EEP and the partners’ booths and on the attendees bag
  - DEEP video running at the booth of the European Exascale Projects
  - Th.Lippert (JUELICH) “Smart Acceleration for Clusters” (entry at ISC’14 Blog). Online at: http://www.isc-events.com/isc14/isc_blog/items/smart-acceleration-for-clusters.html
- **JSC-internal seminar with visit of Dr. Tjerk P. Straatsma, Oak Ridge Leadership Computing Facility, National Center for Computational Sciences, Oak Ridge National Laboratory,** Jülich, Germany, July 8 2014.
  - E.Suarez: “DEEP and DEEP-ER” (presentation)
- **International Conference on Parallel Processing (ICPP),** Minneapolis, USA, 9 – 12 September 2014.
  - S. Prabhakaran (GRS), “A Batch System with Fair Scheduling for Evolving Applications” (paper presented at
- **4th Brazil –France Workshop on High Performance Computing and Scientific Data Management**, Gramado, Brasil, 15 – 18 September 2014
  - R. Léger (INRIA): „A parallel Discontinuous Galerkin Time-Domain solver of Maxwell’s equations“

1.3.1.2 Publications, proceedings, press-releases, and newsletters:

- **Press release from the Forschungszentrum Juelich, 9.10.2013.** Published online: http://www.fz-juelich.de/SharedDocs/Pressemitteilungen/UK/DE/2013/13-10-09-DEEPER.html. Open Access: Yes
  - E.Suarez (JUELICH): “Mit DEEP-ER noch schneller zum Exascale-Rechner”.

- **Exascale -** Newsletter of Forschungszentrum Jülich on Supercomputing, Nr. 03/2013, p. 3. Available online: http://www.fz-juelich.de/SharedDocs/Downloads/PORTAL/EN/publications/exascale-newsletter/exascale_nl_03_2013.pdf?__blob=publicationFile. Open Access: Yes
  - Ch.Hohlfeld (JUELICH): “Faster and Safer with DEEP-ER”.

- **JSC News.** No. 217, November 2013, Published online: http://www.fz-juelich.de/ias/jsc/EN/News/Newsletter/newsletter_node.html. Open Access: Yes
  - E.Suarez (JUELICH): “Start of the Exascale Projects DEEP-ER and Mont-Blanc 2”

- **inSiDE; Innovative Supercomputing in Deutschland,** Published twice a year by The German National Supercomputing Centres HLRS, LRZ, JSC. Open Access: yes.
  - E.Suarez, N.Eicker (JUELICH): “Going DEEP-ER to Exascale” (submitted in march 2014)

- **International Innovation**
  - E.Suarez (JUELICH), “Extreme computing” (article on press). Online at: http://www.deep-er.eu/files/IntInnovation_DEEP_DEEP-ER.pdf

1.3.1.3 Participation at industry and business cooperation related events:

- PROSPECT General Assembly, Garching, Germany, October 18, 2013
- ETP4HPC Steering Board, Barcelona, Spain, November 4, 2013
- Joint booth of the “European Exascale Projects” at SC13, Denver, Colorado, November 17-22, 2014
- ETP4HPC Steering Board, München, Germany, December 9, 2013
- Booth at the HiPEAC Conference 2014, Vienna, Austria, January 20-22, 2014
- ETP4HPC Steering Board, Rome, Italy, January 24, 2014
- PROSPECT General Assembly, Barcelona, Spain, March 27, 2014

List of Acronyms and Abbreviations

A

B

BADW-LRZ: Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften. Computing Centre, Garching, Germany

BeeGFS: **The** Fraunhofer Parallel Cluster File System (previously acronym FhGFS). A high-performance parallel file system to be adapted to the extended DEEP Architecture and optimised for the DEEP-ER Prototype.

BN: Booster Node (functional entity)

BNC: Booster Node Card is a physical instantiation of the BN

BoP: Board of Partners for the DEEP-ER project

BSC: Barcelona Supercomputing Centre, Spain

BSCW: Basic Support for Cooperative Work, Software package developed by the Fraunhofer Society used to create a collaborative workspace for collaboration over the web

C

CINECA: Consorzio Interuniversitario, Bologna, Italy

CN: Cluster Node (functional entity)

Coordinator: The contractual partner of the European Commission (EC) in the project

CPU: Central Processing Unit

CRESTA: Collaborative Research into Exascale Systemware Tools & Applications: EU-funded Exascale project.

D

DDG: Design and Developer Group of the DEEP-ER project

DEEP: Dynamical Exascale Entry Platform

DEEP-ER: DEEP Extended Reach: this project

DEEP-ER Network: high performance network connecting the DEEP-ER BN, CN and NAM; to be selected off the shelf at the start of DEEP-ER

DEEP-ER Prototype: Demonstrator system for the extended DEEP Architecture, based on second generation Intel® Xeon Phi™ CPUs, connecting BN and CN via a single, uniform network and introducing NVM and NAM resources for parallel I/O and multi-level checkpointing

DEEP Architecture: Functional architecture of DEEP (e.g. concept of an integrated Cluster Booster Architecture), to be extended in the DEEP-ER project

DEEP System: The prototype machine based on the DEEP Architecture developed and installed by the DEEP project

E

E10: Exascale 10. Parallel I/O software developed by a consortium of partners around the EOFS community. Partner Xyratex is responsible for the development needed for the DEEP-ER project.
EC: European Commission
EC-GA: EC-Grant Agreement
EESI: European Exascale Software Initiative (FP7)
EOFS: European Open File System.
EU: European Union
Eurotech: Eurotech S.p.A., Amaro, Italy
Exaflop: 10^{18} Floating point operations per second
Exascale: Computer systems or Applications, which are able to run with a performance above 1018 Floating point operations per second
EXTOLL: High speed interconnect technology for cluster computers developed by University of Heidelberg
ETP4HPC: European Technology Platform for High Performance Computing.

F

FhGFS: Acronym previously used to refer to BeeGFS.
FLOP: Floating point Operation
FP7: European Commission 7th Framework Programme.
FPGA: Field-Programmable Gate Array, Integrated circuit to be configured by the customer or designer after manufacturing

G

GRS: German Research School for Simulation Sciences GmbH, Aachen and Juelich, Germany

H

H5hut: Library implementing several data models for particle-based simulations that encapsulates the complexity of parallel HDF5.
HDF5: Hierarchical Data Format: A set of file formats and libraries designed to store and organise large amounts of numerical data
HPC: High Performance Computing
HW: Hardware

I

ICT: Information and Communication Technologies
IEEE: Institute of Electrical and Electronics Engineers
Intel: Intel Germany GmbH Feldkirchen,
IP: Intellectual Property
iPic3D: Programming code developed by the University of Leuven to simulate space weather

ISC: International Supercomputing Conference, Yearly conference on supercomputing which has been held in Europe since 1986

J

JUBE: Jülich Benchmarking Environment

JUDGE: Juelich Dedicated GPU Environment: A cluster at the Juelich Supercomputing Centre

JUELICH: Forschungszentrum Jülich GmbH, Jülich, Germany

K

KNC: Knights Corner, Code name of a processor based on the MIC architecture. Its commercial name is Intel® Xeon Phi™.

KNL: Knights Landing, second generation of Intel® Xeon Phi™

KULeuven: Katholieke Universiteit Leuven, Belgium

L

M

MIC: Intel Many Integrated Core architecture

Mont-Blanc: European scalable and power efficient HPC platform based on low-power embedded technology

Mont-Blanc 2: Follow-up project of Mont-Blanc

MPI: Message Passing Interface, API specification typically used in parallel programs that allows processes to communicate with one another by sending and receiving messages

N

NAM: Network Attached Memory, nodes connected by the DEEP-ER network to the DEEP-ER BN and CN providing shared memory buffers/caches, one of the extensions to the DEEP Architecture proposed by DEEP-ER

NASA: National Aeronautics and Space Administration, Washington, USA

NetCDF: Network Common Data Form. A set of software libraries and data formats that support the creation, access, and sharing of array-oriented scientific data

NVM: Non-Volatile Memory

NVMe: NVM Express. Specification for accessing solid-state drives attached through the PCIe bus.

O

OEM: Original Equipment Manufacturer. Term used for a company that commercialises products out of components delivered by other companies.

OmpSs: BSC's Superscalar (Ss) for OpenMP

OpenMP: Open Multi-Processing, Application programming interface that support multiplatform shared memory multiprocessing
OS: Operating System

P

ParaStation Consortium: Involved in research and development of solutions for high performance computing, especially for cluster computing
ParaStationMPI: Software for cluster management and control developed by ParTec
Paraver: Performance analysis tool developed by BSC
Paraview: Open Source multiple-platform application for interactive, scientific visualisation
ParTec: ParTec Cluster Competence Center GmbH, Munich, Germany
PCI: Peripheral Component Interconnect, Computer bus for attaching hardware devices in a computer
PCIe: PCI Express, Standard for peripheral interconnect developed to replace the old standards PCI, improving their performance
PFlop/s: Petaflop, 10^{15} Floating point operations per second
PM: Person Month or Project Manager of the DEEP project (depending on the context)
PMT: Project Management Team of the DEEP-ER project
PRACE: Partnership for Advanced Computing in Europe (EU project, European HPC infrastructure)
PROSPECT: Promotion of Supercomputing Partnerships for European Competitiveness and Technology (registered association, Germany)

Q

QPACE: QCD Parallel Computing Engine. Specialised supercomputer for QCD Parallel Computing

R

R&D: Research and Development

S

SC: International Conference for High Performance Computing, Networking, Storage, and Analysis, organised in the USA by the Association for Computing Machinery (ACM) and the IEEE Computer Society
Scalasca: Performance analysis tool developed by JUELICH and GRS
SW: Software

T

TFlop/s: Teraflop, 1012 Floating point operations per second
ToW: Team of Work Package leaders within the DEEP-ER project

TP10: Third Party under Clause 10.

U

UHEI: University of Heidelberg, Germany
UREG: University of Regensburg, Germany

V

VI-HPS: Virtual Institute for High Productivity Supercomputing
VTune: Commercial application for software performance analysis

W

WP: Work Package

X

Y

Z