SEVENTH FRAMEWORK PROGRAMME

FP7-ICT-2013-10

DEEP-ER

DEEP Extended Reach

Grant Agreement Number: 610476

D1.4

Midterm management report at month 18

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:	09.12.2015

Project and Deliverable Information Sheet

<table>
<tr><td rowspan="6">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: D1.4</td></tr>
<tr><td colspan="2">Deliverable Nature: Report</td></tr>
<tr><td rowspan="2">Deliverable Level:
CO*
(the present document contains only the public part of the deliverable)</td><td>Contractual Date of Delivery:
31 / March / 2015</td></tr>
<tr><td>Actual Date of Delivery:
31 / March / 2015</td></tr>
<tr><td colspan="2">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: Midterm management report at month 18</td></tr>
<tr><td colspan="2">ID: D1.4</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_D1.4_Midterm_management_report_M18_v2.0-ECapproved-PublishablePart</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>S.Höfler-Thierfeldt (JUELICH), J.Kreutz(JUELICH), 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	31/March/2015	Final	EC submission
2.0	09/December/2015	Approved	EC approved

Document Keywords

Keywords:	DEEP-ER, HPC, Exascale, midterm report, month 18

Table of Contents

List of Figures

List of Tables

No table of figures entries found.

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 **months 12 to 18** of the DEEP-ER project. The main achievements in the reporting period are enumerated below:

- Co-design effort further enhanced: discussions between hardware, software, and application developers to assure a coherent development that addresses all requirements.
- Detailed feasibility study by partner Eurotech on the “Aurora Blade” architecture, including design for a KNL-based node board. Commitment by partner Eurotech to produce this design, expressed at the interim review, and decision for DEEP-ER to follow this route.
- First results of I/O benchmarks and application mock-ups running on Non-Volatile Memory devices.
- Functionality Hybrid Memory Cube (HMC) controller (required for Network Attached Memory (NAM) prototype) established. The NAM architecture and a first version of the functional specification have been completed.
- Development of the DEEP-ER I/O software stack – containing BeeGFS, SIONlib, and E10 – ongoing, taking into account the requirements from resiliency software and applications. Implementation of node-local I/O caches for BeeGFS.
- Implementation of resiliency software layer progressing well. Bi-weekly discussions between I/O and resiliency software tasks take place to define interfaces and guarantee a coherent development of the full software stack. Models for the performance impact of resiliency measures implemented.
- Benchmarks integrated in JUBE environment and tests performed: identification of parameters required by BeeGFS to achieve high-speed in metadata handling. Integration of full applications ongoing.
- Application adaptations and improvements ongoing: optimisation to take benefit from Intel® Xeon Phi™, code partition between Cluster and Booster parts of the DEEP architecture, integration with the I/O and resiliency tools developed in DEEP-ER.
- Dissemination of project goals and status in various workshops and conferences, amongst others the Supercomputing Conference (SC14).

- Coordination and co-organisation of joint dissemination activities with the rest of European Exascale Projects, i.e. for the joint booth at SC14 and ISC 2015, and the industry-focussed Exascale workshop at the PRACEdays15.

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 these issues, DEEP-ER will extend the heterogeneous Cluster-Booster Architecture implemented by the DEEP[1] project by 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 the results already achieved, which contribute to their fulfilment, 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 I/O and resiliency software and between application-based checkpoint/restart and task-based resiliency software has been agreed.

   → Implementation of I/O and resiliency software ongoing. Bi-weekly cross-Work Package discussions take place to guarantee a coherent development.

[1] www.deep-project.eu

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

   → Initially, it was planned to build the DEEP-ER Prototype based on Eurotech's Aurora HiVe concept, as defined and documented as the "Brick architecture" in Deliverables D3.1 and D3.2. Detailed investigations into the technical and market risks associated with developing a PCIe add-in card form factor KNL board (self-bootable and with the amount of memory required for the project), led to the project's proposition to adopt a simplified architecture based on an air-cooled, readily available compute board and PCI Express attached NVM and NIC devices. However, this proposal was considered too conservative by the external reviewers.

   → Various alternative architectures have been studied in response to the reviewers' recommendations to fully address energy efficiency, density, and cooling. In conclusion to this analysis, in the interim review at month 18 of the project it was decided to go for Eurotech's Aurora Blade architecture.

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.

   → 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 and connected via PCI Express have been installed at Juelich, and detailed performance measurements with I/O benchmarks and application mock-ups are available. The DEEP-ER Prototype will use the product implementation of this technology, as explained in D3.2, or the successor product generation, which will bring further performance improvements.

   → Architecture and principal design of the NAM have been defined, and a first version of the functional specification was completed and is being discussed with the system software and application work packages. The NAM will use partner UHEI's HMC controller implementation.

   → UHEI's HMC controller implementation has been completed and functionally validated. Reliability issued noticed during validation could be tracked to early silicon versions of Micron's HMC devices and the FPGA used – they will not apply to the NAM prototype, which is based on a later stepping.

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.

→ 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 – the Fraunhofer file system (BeeGFS), the parallel I/O library SIONlib, and the Exascale10 (E10) software stack – must provide to the project, the interplay between them, and their interfaces have been described in deliverables D4.1 and D4.2.

→ In BeeGFS two new functionalities have been implemented: cache domain handling and user-level stripe-size definition. The cache domain will be executed on the node-level NVM devices and can be executed synchronous or asynchronously. The synchronous version is already available and tested, the asynchronous is under implementation.

→ The needed preparatory phase for the further development of SIONlib is completed. Amongst other improvements, its communication layers have been refactored to eliminate code replications and increase the overall modularity and manageability of the library.

→ Partner Seagate has integrated into E10 the new BeeGFS functionalities for cache handling and user-level strip-size definition.

→ A list of benchmarks to be used for the evaluation of the DEEP-ER I/O software has been identified (see D4.3) and they have been integrated into the JUBE benchmark environment. They will be used to regularly monitor the overall I/O performance to measure the impact of the various developments done in the DEEP-ER project. This activity has resulted in the identification of the right BeeGFS parameters to achieve optimal performance for metadata handling.

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.

   → In a co-design effort, the overall resiliency software stack has been defined (see D5.1) taking into account the requirements from the WP6 application developers, the WP3 hardware capabilities, and the I/O functionality required from/provided by from WP4.

   → The role to be played by the user-level and the task-based resiliency functionalities, and the interfaces between them have been defined.

   → The Scalable Checkpoint/Restart (SCR) library is being adapted to the needs of the DEEP-ER project including an API (abstraction layer) for the application users to apply SCR in their codes (see D5.2).

   → OmpSs adaptations for task-based resiliency are under implementation.

→ In close collaboration with WP3, an event Monte Carlo failure model has been designed and implemented. Goal is to optimise policies that determine for each application the frequency, redundancy level and storage-location of each checkpoint. A closed formula has been developed. Results from failure model and closed formula are aligned.

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.

   → The application requirements – in terms of hardware capabilities, I/O and resiliency functionalities – have been gathered through a questionnaire. DDG teleconferences and face-to-face meetings are used for further co-design discussions, as applications evolve with time through code optimisations and implementation of new functionalities.

   → The structure of the applications has been analysed, performance and scaling tests have taken place.

   → Various improvements are being implemented in the application. Important topics in the code optimisations are: vectorisation, optimising communication strategies and/or numbering schemes, improving I/O, implementing checkpointing, etc.

   → SIONlib and OmpSs are being integrated in several applications.

   → In collaboration with WP3 and WP4, mock-ups from the applications are being prepared to use them for I/O benchmarking. Mock-ups from the space weather and seismic applications are already available.

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

According to the DoW, two milestones were to be reached between **month 13 and month 18** of the DEEP-ER project. Unfortunately, hardware delays have prevented the consortium from fully reaching them, same as MS5, pending from last reporting period. Details are given below:

- **MS1 to MS4**, as well as **MS6** have been reached in the first year of the project.
- **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.

- **MS7** – "Cluster-Booster evaluator, NAM functional evaluator, and NVM evaluator available": as in MS5, the Cluster-Booster evaluator could not be deployed due to the delay in KNL. The HMC controller is a component of the NAM functional evaluator, but this prototype should also contain two KNLs, and is therefore not yet available. The NVM evaluator (two NVMe devices installed in workstations) is available at JUELICH already to the project since summer 2014. Its early availability has allowed doing benchmarking and obtaining early results showing very promising I/O performance.
- **MS8** – "Performance extrapolation based on design decisions": due to the need to re-visit the architectural design of the DEEP-ER Prototype after the review at M12, this milestone could not be reached. The responsible task (Tk7.1), was on hold until the review at M18, where the architecture was decided. Deliverable D7.1 is shifted to M24.

The DEEP-ER consortium will deploy as early as possible a Software Development Vehicle (SDV), including a small number of early KNL samples in Customer Reference Board form factor. The SDVs will provide the functionality of the Booster CPU evaluator, Cluster-Booster evaluator, and NVM evaluators.

Management, legal and administrative tasks

In the present reporting period the management activities focused on monitoring the progress of the project with regards to the achievement of all technical goals specified in the Description of Work (DoW) and the fulfilment of all commitments to the European Commission.

The Project Management Team organised the agenda for the first review meeting (at month 12) which took place on the 21st October 2014 in Brussels (Belgium). To fulfil the internal quality policies a rehearsal meeting one day before the review was conducted. As a result of the first review, the project has been evaluated as doing "good progress". Additionally, all deliverables submitted in the first year of the project were approved. All public approved deliverables have been uploaded to the project website. The comments from the reviewers and their recommendations concerning future work are addressed in section 2.2.

Addressing the reviewer recommendations at M12, an important management activity was focused on driving technical, confidential discussions with various HPC providers to identify possible architecture alternatives for the DEEP-ER Prototype. Additionally, the PMT also prepared agenda and slides for the interim review meeting at M18, focused on the prototype architecture, which took place on 26th March 2015 in Brussels. There, the need for a project extension was discussed. An amendment of the Description of Work will be prepared to adapt the project plan and its duration.

The financial statements from all partners were submitted to the NEF server after the end of the first project year. The financial data has been approved by the European Commission and the first interim payment has been further transferred to the partners.

Monthly teleconferences of the Team of Work Package leaders (ToW) were organised to periodically discuss the progress in all Work Packages (WPs). Deliverables D4.3, D1.4, and D5.2 have been 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[2]. 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. Following a reviewer's recommendation, a face-lift of the project website is under preparation. New design will be online in April 2015.

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. Updates 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 in the reporting period. Partners from the DEEP-ER consortium presented the project's concept in conferences and workshops, including one of the most important events in the HPC community: the Supercomputing Conference (SC), which took place in New Orleans (USA) in November 2014. There, a poster on the DEEP-ER I/O concept was presented at the Emerging Technologies Track.

At SC14 the DEEP-ER project co-organised a joint booth, together with other European Exascale Projects (EEP) – DEEP, Mont-Blanc (1 and 2), EPiGRAM, EXA2CT, NUMEXAS, and CRESTA). DEEP-ER had an own wall describing the architecture and main goals of the project, and the HMC controller (the core component of DEEP-ER's NAM), was displayed in the booth. Additionally, a DEEP-ER flyer was prepared and distributed at the EEP booth and at the booths of other project partners.

A further highlight of the DEEP-ER dissemination activities achieved in this reporting period is the project video, a joint effort between DEEP and DEEP-ER, which explains in an attractive manner for the general public, the importance of HPC for society, the impact of the two projects, and the value of collaboration. The video is accessible through the project website and has been selected as "visual of the week" by iSGTW and featured prominently on insideHPC.

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

Figure 1: Joint EEP booth at SC14.

DEEP-ER is actively participating in the preparation of further joint activities with the European Exascale Projects (EEP) community. A joint workshop has been at the International Supercomputing Conference (ISC 2015), and will take place on July 16, 2015 in Frankfurt (Germany).

Furthermore, already several articles and publications on project contents have been submitted. The Annex of this report gives a list of presentations and publications. A list with all dissemination activities performed in the present reporting period is given in Annex A.1 of this report.

Regarding contacts with industry, a satellite event to the PRACEdays15 will take place on May 26, 2015 in Dublin (Ireland), organised in collaboration with the other EEP. DEEP-ER will highlight the aspects of the project most interesting for industrial users, and a flyer explaining how to get access and use them will be distributed.

Training the community on how to use the software and hardware developed in DEEP-ER is 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. A hands-on training event for the DEEP-ER application developers is under preparation and will take place in May 2015 in Barcelona (Spain).

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 2) 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.

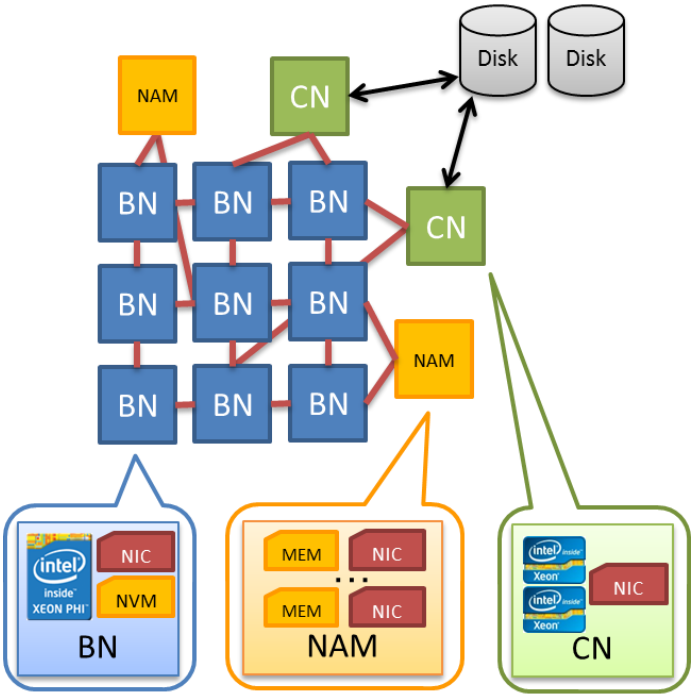

Figure 2: High-level view of the DEEP-ER Prototype. BN=Booster Node; CN=Cluster Node; NVM=Non-Volatile Memory; NAM=Network Attached Memory

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 BeeGFS parallel file system, 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

Deliverable D3.1 submitted at M6 did specify the innovative "Brick architecture" (named Aurora HiVE by Eurotech), compared it to more conventional approaches and proposed to adopt it for the DEEP-ER Prototype. However, detailed investigations into the technical and market risks associated with developing a commensurate, PCIe add-in card form factor KNL board that satisfies the project requirements (in particular with regards to DRAM memory) led

partner Eurotech to decide not to follow this path. As a consequence, Deliverable D3.2 proposed following a conventional approach based on air-cooled, readily available compute boards. This “Node-Board” architecture would not have fully addressed energy efficiency, cooling and scalability aspects, which were not among the project’s stated objectives. The M12 review rejected this approach, and the reviewers recommended re-visiting the architecture: Eurotech should re-consider the decision of developing a KNL board for the Brick, and the project as a whole should search for alternatives.

After the review, Eurotech started a detailed feasibility study on developing a KNL board for DEEP-ER, assisted by partner Intel. Since a board for the Brick architecture commensurate with the DEEP-ER requirements was seen as too risky, the feasibility study proposed an extension of the Aurora Blade architecture by a KNL board and an integration of the NVM and EXTOLL PCI Express cards. This approach does increase the available board area, substantially reducing the technical risk, preserves direct liquid cooling with an option for hot-water cooling, enables flexibility in mixing and matching of KNL and regular Xeon blades, and can support both EXTOLL and InifiniBand NICs.

Simultaneously and to address the reviewers request for alternatives, JUELICH initiated conversations with various vendors in search of those providing suitable KNL-based platforms, which would fulfil the project requirements.

In the M18 interim review, Eurotech did commit to develop the proposed Aurora Blade architecture, and it was decided to use it for the construction of the DEEP-ER Prototype. Eurotech plans to deploy the DEEP-ER Prototype in M33 (June 2016).

Furthermore, DEEP-ER did closely follow the progress of the EXTOLL ASIC development, which is done outside the project. Stepping A2 of the Tourmalet ASIC has become available during the present reporting period. Functionality at a speed of 8 Gbit/s per lane (with 12 lanes per link) was achieved and work is ongoing to fully validate the ASIC at that speed and to probe the performance envelope up to the full design speed of 10 Gbit/s. The currently achieved performance already comes very close to satisfying the project requirements of 100 Gbit/s per link.

The two NVMe devices available for the project have been used to perform I/O benchmarks and test the performance of application mock-ups. Focus of the work was on BSC’s seismic imaging application. The results show raw I/O performance significantly better than the one obtained with Sata-connected SSDs, and substantial application performance improvements through the use of local NVMe devices.

The NAM architecture and principal design is completed, using a Hybrid Memory Cube (HMC) as storage device. A first prototype of the HMC controller has been constructed and tested, with functionality being validated by reliability issues noticed. These were tracked to early silicon versions of Micron’s HMC devices and the FPGA used. A new version is being constructed with current silicon and will be used in the NAM prototype in the next months.

In addition to and before the DEEP-ER Prototype, the project will procure a small Software Development Vehicle (SDV). This system will be deployed as soon as early KNL samples (CRBs) are available to JUELICH, now expected in early summer 2015. It will also contain a number of Haswell-generation Intel Xeon nodes, NVMe devices and an EXTOLL A2 Tourmalet interconnect, providing all needed components for the development of the DEEP-ER system and application and showing similar performance characteristics to the DEEP-ER prototype.

System Software

On the software side, the reporting period focused on the implementation of the various components involved in the I/O and resiliency software stacks of the project. Regular discussions take place between the developers involved to guarantee a coherent and consistent global picture, where all the software components fit together. The overviews of the DEEP-ER I/O and resiliency software layers have been described in the DoW and are shown in Figure 3 and Figure 4, respectively.

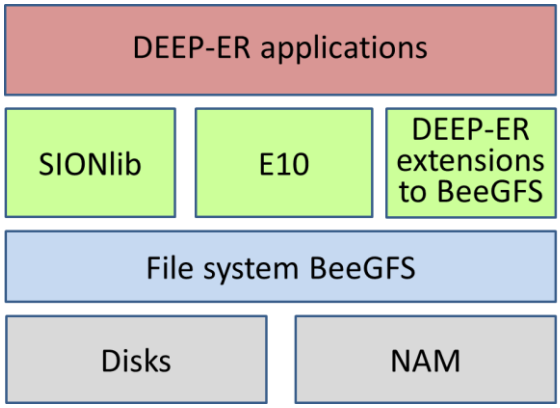

Figure 3: Sketch of DEEP-ER I/O software layers.

In particular, a close interrelation between BeeGFS, SIONlib, and the Scalable Checkcpoint/Restart library (SCR) has been established. All three components will cooperate to realise buddy checkpoints and the overall checkpointing functionality in an efficient way.

BeeGFS has implemented two new functionalities: a local cache layer in the file system and the capability to define the stripping size at user level. The former will make use of the node-local NVMe devices to reduce the frequency of I/O to the global file system and increase therefore the overall scalability of the I/O system. The latter will allow exploiting the knowledge of the user on the data layout in its application to more efficiently determine the physical distribution of the data in the storage.

The communication layers of SIONlib have been refactored to improve the modularity and manageability of the library. This preparatory phase is important to ease the integration of the functionality that will be implemented in SIONlib within DEEP-ER. For instance, SIONlib will be used to guarantee that checkpoints to buddy nodes will not generate a very large amount of small files but only a few large physical files, reducing the burden on the file system. An interface will be implemented in SIONlib to use the local cache functionality of BeeGFS.

The implementation of the E10 API for I/O has also progressed in the reporting period. For instance, it already integrates the local-cache and user-level data-stripping functionality form BeeGFS. First measurements performed on the DEEP Cluster show promising results.

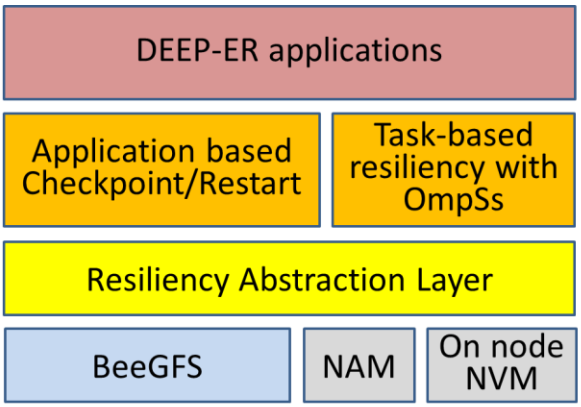

Figure 4: Sketch of DEEP-ER resiliency layers

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.

Additional to the strong cooperation with the I/O software developers, also the implementation in the pure failure recovery software packages has progressed. A first version of the resiliency abstraction layer has been implemented and described in D5.2. The abstraction layer adds to SCR specific functions that allow efficiently exploiting DEEP-ER's I/O functionality for checkpoint/restart applications. A meeting with Adam Moody, main developer of the SCR library, took place during SC14. The ideas behind DEEP-ER and the extensions planned were discussed and Mr. Moody expressed his interest in the future integration of DEEP-ERs developments in the main branch of SCR.

Progress has been also done regarding the task-based resiliency software. Its design is completed and the implementation is ongoing, performing the needed adaptions in OmpSs and its runtime to support the failure recovery functionality.

Beyond that, the extension of the ParaStation management daemon by an interface for querying resiliency-related status information from the MPI layer and thus also from the OmpSs runtime environment has been envisaged in detail.

Finally, a failure model has been implemented in the form of an event based Monte Carlo simulation, to determine the frequency and location of checkpoints required by a given application. A closed formula has been developed and served to validate the model.

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 have been implemented in JUBE and are run frequently on the DEEP Cluster to monitor the I/O performance, as the software is being developed and updates are installed. With this activity the parameters required in BeeGFS for an optimal performance of metadata handling have been determined. Work will continue with the integration of mock-ups, proxy- and full-fledged applications in JUBE.

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. Co-design discussions to gather more specific application requirements take place in the DDG and the consortium face-to-face meetings. Additionally, internal review meetings focused on the work done by WP6 take place during the consortium face-to-face meetings. The developers described their applications and presented the results that they had obtained since last meeting, 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.

In the reporting period the application developers have been performing modifications to adapt the codes to the DEEP-ER hardware and software architecture, as well as general code optimisations. Some examples are: code refactoring to implement the Cluster-Booster division, multithreading, improvement of vectorisation, integration of SIONlib for I/O and/or checkpointing, porting to Xeon Phi, integration and/or optimisation of OmpSs, etc.

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 functionality and an efficient infrastructure for failure 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 in the same time, thus increasing scientific throughput, and that 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 13 and 18** of the DEEP-ER project.

1.3.1.1 Conferences, workshops, and meetings:

- **Institute for Cyber-Enabled Research at Michigan State University**, 5 November 2014.
  - A.Johnson (KULeuven): "Experience in DEEP/ER with porting iPic3D to MIC, focusing on appropriate use of SoA and AoS representations" (presentation)

- **Supercomputing Conference SC'14**, New Orleans, USA, November 17-20, 2014:
  - Joint booth of the European Exascale Projects (EEP). Booth #1039. Participant projects: DEEP, DEEP-ER, Mont-Blanc, CRESTA, EPiGRAM, and EXA2CT).
  - S.Breuner (FHG-ITWM), W.Frings (JUELICH), K.Thust (JUELICH), N.Eicker (JUELICH), G.Congiu (Seagate), S.Narasimhamurthy (Seagatet). „DEEP-ER I/O: Addressing Exascale I/O problems" (poster at the Emerging Technologies Track)
  - DEEP and DEEP-ER fliers distributed at the EEP and the partners' booths and on the attendees bag
  - Inria's application and its work within DEEP-ER has been presented at the Inria booth
  - E.Suarez (JUELICH): DEEP-ER project, presentation at the Intel booth
  - V.Beltran (BSC): OmpSs Collective Offload, presentation at the Intel booth
  - J.Romein (ASTRON): Correlating Radio Telescope Data for the Square Kilometre Array, presentation at the Intel booth
  - DEEP+DEEP-ER video running at the booth of the European Exascale Projects
  - Soft News for Website**:** DEEP/-ER go DEEP South: Announcement SC14 http://www.deep-er.eu/press-corner/news/51-deep-er-go-deep-south

- **JUELICH-JSC meeting (Visit C.Aubley),** Juelich, Germany, January 19, 2015:
  - E.Suarez (JUELICH), "DEEP and DEEP-ER" (presentation).

- **BDEC (Big Data and Extreme Scaling) Workshop,** Barcelona, Spain, January 28-30, 2015:
  - E.Suarez (JUELICH), "The DEEP (and DEEP-ER) projects" (presentation)

- **SUMA (Supermassive Computations in Theoretical Physics) Workshop**, 11 - 13 February 2015, Trento, Italy
  N. Eicker: "The European Supercomputer Projects DEEP & DEEP-ER"

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

- **insideHPC:** Experimenting with innovative memory technology, Online at: http://insidehpc.com/2014/10/interview-experimenting-deep-er-memory-technology/ (first published on DEEP-ER website)

- **insideHPC:** J.Schmidt (UHEI), Experimenting with DEEP-ER NAM Technology, Online at: http://insidehpc.com/2014/10/interview-experimenting-deep-er-memory-technology/

- **DEEP/-ER Video**, Online at: http://www.deep-er.eu/press-corner/news/52-video

- **insideHPC:** DEEP-ER project reaches for Exascale, Online at: http://insidehpc.com/2015/02/video-deep-er-project-reaches-for-exascale/

- **iSGTW:** DEEP-ER project reaches for Exascale (also featured as “visual of the week”), Oline at: http://www.isgtw.org/feed-item/video-deep-er-project-reaches-exascale

1.3.1.3 Participation at industry and business cooperation related events:

- ETP4HPC Steering Board, Teleconference, February 18, 2015
- ETP4HPC Steering Board, Barcelona, Spain, March 5, 2015
- ETP4HPC General Assembly, Barcelona, Spain, March 5, 2015
- PROSPECT General Assembly, Munich, Germany, March 25, 2015

List of Acronyms and Abbreviations

A

API: Application Programming Interface.

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

CRB: Customer Reference Board. An early version of a KNL board developed by Intel.

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
EEP: European Exascale Projects
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

QCD: Quantum Chromodynamics
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
SCR: Scalable Checkpoint/Restart library

SDV: Software Development Vehicle: a HW system to develop software in the time frame where the DEEP-ER Prototype is not yet available.
SEO: Search Engine Optimisation: the process of improving the visibility of a website or a web page in a search engine's results.
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 special 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