SEVENTH FRAMEWORK PROGRAMME

FP7-ICT-2013-10

DEEP-ER

DEEP Extended Reach

Grant Agreement Number: 610476

D3.5

DEEP-ER prototype installation

Approved

Version: 2.0

Author(s): J. Kreutz (JUELICH)

Contributor(s): Hans-Christian Hoppe (Intel), P.Niessen (ParTec)

Date: 04.05.2017

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: D3.5</td></tr>
<tr><td colspan="2">Deliverable Nature: Report</td></tr>
<tr><td>Deliverable Level:
PU</td><td>Contractual Date of Delivery:
31 / October / 2016</td></tr>
<tr><td>Actual Date of Delivery:
31 / January / 2017 -(Delayed due to later availability of DEEP-ER prototype (see D8.3))
31 / March /2017 – Updated before final review to reflect last results</td></tr>
<tr><td colspan="2">EC Project Officer: Juan Pelegrín</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: DEEP-ER prototype installation</td></tr>
<tr><td colspan="2">ID: D3.5</td></tr>
<tr><td>Version: 2.0</td><td>Status: Approved</td></tr>
<tr><td colspan="2">Available 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_D3.5_Prototype_Installation_v2.0-ECapproved</td></tr>
<tr><td rowspan="4">Authorship</td><td>Written by:</td><td>J. Kreutz (JUELICH)</td></tr>
<tr><td>Contributors:</td><td>Hans-Christian Hoppe (Intel), P.Niessen (ParTec)</td></tr>
<tr><td>Reviewed by:</td><td>A. Emerson (CINECA), 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/January/2017	Final	EC submission
1.1	27/March/2017	Final	Updated before final review at M42, to reflect last results
2.0	04/May/2017	Approved	EC approved

Document Keywords

Keywords:	DEEP-ER, HPC, Exascale, Installation, prototype, infrastructure

Table of Contents

List of Figures

List of Tables

Executive Summary

The development and installation of a DEEP-ER prototype system is one of the main goals within the DEEP-ER project. The prototype system has been designed to extend the Cluster-Booster architecture of the former Dynamical Exascale Entry Platform (DEEP) project by adding I/O and memory capacity as well as resilience functionality. The system allows the potential of new storage technologies to be explored in order to provide increased performance and power efficiency. Additionally, compared to the DEEP system, it takes advantage of progress in the fields of many-core CPUs and high-performance networks.

The DEEP-ER Prototype system consists of a Cluster and a Booster part complemented by non-volatile and network attached memory (NAM). It uses a uniform high-speed interconnect across the Cluster, Booster and the memory devices. In addition to the hardware installation activities, software installation and configuration of the prototype system has to be performed in order to bring the system into operation and to provide it to the project.

The Cluster part of the DEEP-ER Prototype is installed and fully available to the project, and a small number of KNL nodes are attached to it via the EXTOLL network to enable application development in WP6.

At the time of writing, the installation of the Booster part is still work in progress; initial bring-up, configuration and tests of the first Booster chassis were promising, yet a coolant liquid leak caused by rapid chemical corrosion has stopped these activities. Detailed analysis of the root cause for the leakage has been performed .

This document is an updated verson of deliverable 3.5 to reflect the status of the DEEP-ER prototype installation at the end of the project.

1 Introduction

The DEEP-ER project investigates innovative concepts to provide fast and highly scalable parallel I/O and increased resilience functionality to a Cluster-Booster system based on new storage technologies. The DEEP-ER prototype demonstrates the benefits of these concepts and technologies in actual use for the seven applications in WP6. Figure 1 illustrates the different components included in the prototype system. The two main parts are the Cluster and the Booster Nodes to which the non-volatile memory cards are attached. The Booster is using the Eurotech Aurora concept of direct liquid cooled components, while the Cluster Nodes are air-cooled to keep costs and risks down. The different cooling technologies require the use of two distinct racks. The NAM devices and file servers are air-cooled as well, so they are integrated in the Cluster rack. The switchless EXTOLL network avoids the need for additional rack space hosting network switches, and the EXTOLL NICs will be directly cabled.

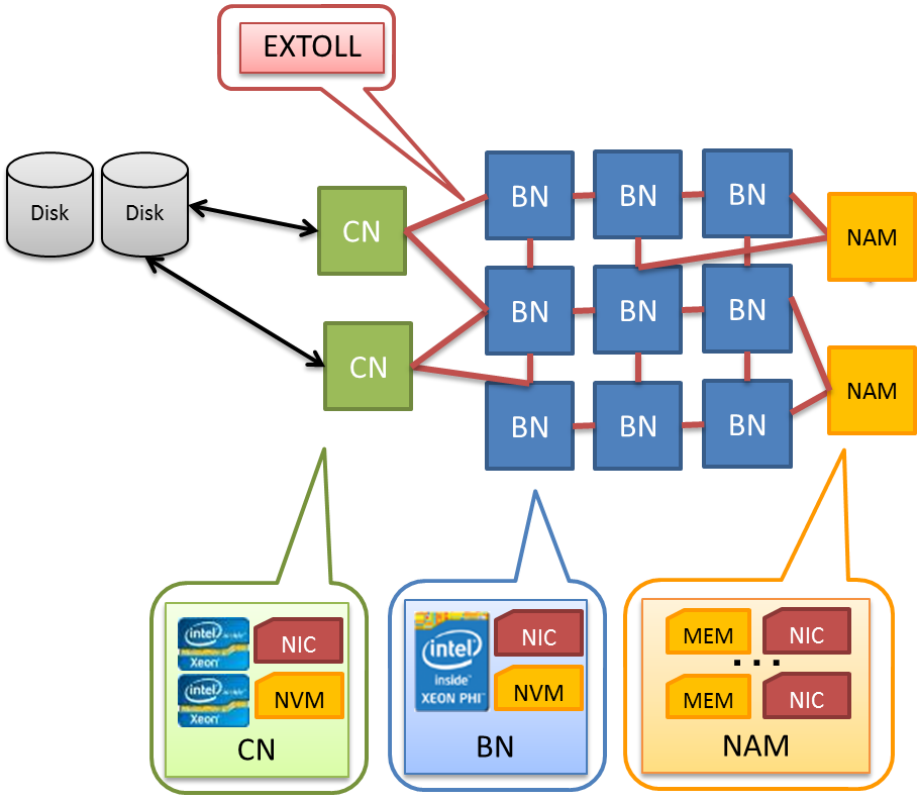


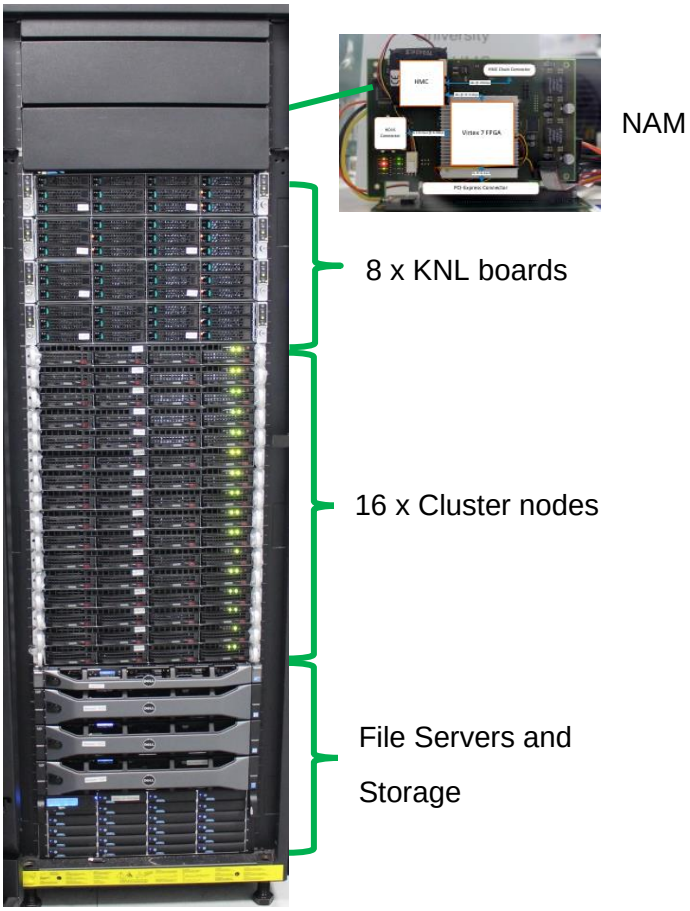
Figure 1: Schematic overview of the DEEP-ER prototype system components

This document describes the installation process and status of the DEEP-ER Prototype system. The software development vehicle (SDV) constituting the Cluster part of the system is described in section 2. Section 3 explains the infrastructure preparation and environment configuration to integrate the liquid cooled Booster part of the system at the Jülich Supercomputing Centre. Details of the Booster hardware installation including the power and water connections are covered in section 4. The software installation and system configuration necessary to bring the system up and to make it available for the application developers is described in section 5. It discusses the system bring-up, configuration and testing performed with the first Booster chassis, which was stopped by the occurrence of a serious coolant liquid leak.

2 DEEP-ER Cluster / Software Development Vehicle (SDV)

The DEEP-ER Booster was designed to achieve highest energy efficiency and density by leveraging Eurotech's custom hardware technology [4]. To enable system and application software developers to port and implement their codes before the DEEP-ER Booster would become available towards the end of the project, a software development vehicle (SDV) using off-the-shelf technology was configured, installed and made available to the project.

This SDV consists of 16 dual-processor Intel® Xeon® nodes complemented by 8 pre-release nodes equipped with the second generation of Intel® Xeon Phi™ processors (Knights Landing, also being used in the definitive DEEP-ER Booster), a local storage system and an EXTOLL based interconnect. Table 1 and Figure 2 present details of the SDV.



16 Cluster nodes	2 x Intel® Xeon CPU E5-2680 v3 (Haswell generation)
8 KNL nodes	1 x Intel® Xeon Phi 7210 on Intel® Server Board S7200AP (Adams Pass) pre-release boards [6]
Non-volatile memory	Per node: 1 x Intel® DC P3700 SSD 400 GB
Network attached memory	Per rack: 2 x EXTOLL FPGA + Hybrid Memory Cube (HMC)
High-speed interconnect	EXTOLL TOURMALET NICs

Table 1: Hardware details for the DEEP-ER Software Development Vehicle (SDV)

For the installation of the SDV rack there were no particular difficulties since all of the components are air-cooled and the waste heat can be removed by the existing room air-

conditioner. Furthermore, the rack provides standard power connections and internal power distribution bars to which the SDV components can be connected. The 16 Intel Xeon nodes in the SDV were chosen and configured in the way necessary to fulfil the requirements for the Cluster Nodes of the final DEEP-ER Prototype system with an EXTOLL TOURMALET card and an Intel NVMe card to extend the node with non-volatile memory. Both PCI Express cards are identical to the ones attached to the DEEP-ER Booster Nodes. Hence the 16 Xeon nodes do form the Cluster part in the DEEP-ER Prototype system.

3 DEEP-ER Booster Rack Installation

The specific requirements for the DEEP-ER Booster installation on the JUELICH infrastructure (floor space, power supply and cooling) are presented in Table 2.

Rack dimensions and weight	
Size of rack (foot print)	800mm x 760mm, 42U high
Weight of assembled system	Up to 900kg
Power consumption	
Per chassis	7900W
Full rack	31600W
Power voltage	48V (DC)
Liquid Cooling	
Inlet temperature range	18°C – 42°C
Ambient temperature and humidity	18°C – 25°C, 20% - 60%
Filter size in cooling loop	50µm
Water quality	Following AESHRAE guidelines
Required additives	Corrosion inhibitor and biocide

Table 2: DEEP-ER Booster installation requirements

Several actions at the installation site were taken to meet the installation requirements. These are described in the following subsections.

3.1 Floor space

Due to the heavy weight of the DEEP-ER Booster rack, a support frame for the floating floor was needed. It was implemented under the floor plates and covers an area of 1.80m × 1.80m. There is enough space to put both the DEEP-ER Booster and the SDV rack on the area of support. The distance between the two racks is therefore quite small, and the EXTOLL cable length for the connection between the Cluster and the Booster parts can be kept below 1.5m which is the maximum length of EXTOLL copper cables. An additional restriction for rack positioning comes from the need for statutory escape ways in the installation room. There has to be an escape way present with a width of at least 90cm. Figure 3 shows the floor plan of the installation room which fulfils the escape way requirements. The racks "06" and "07" form the DEEP-ER Prototype system with "06" being the SDV and "07" being the DEEP-ER Booster.

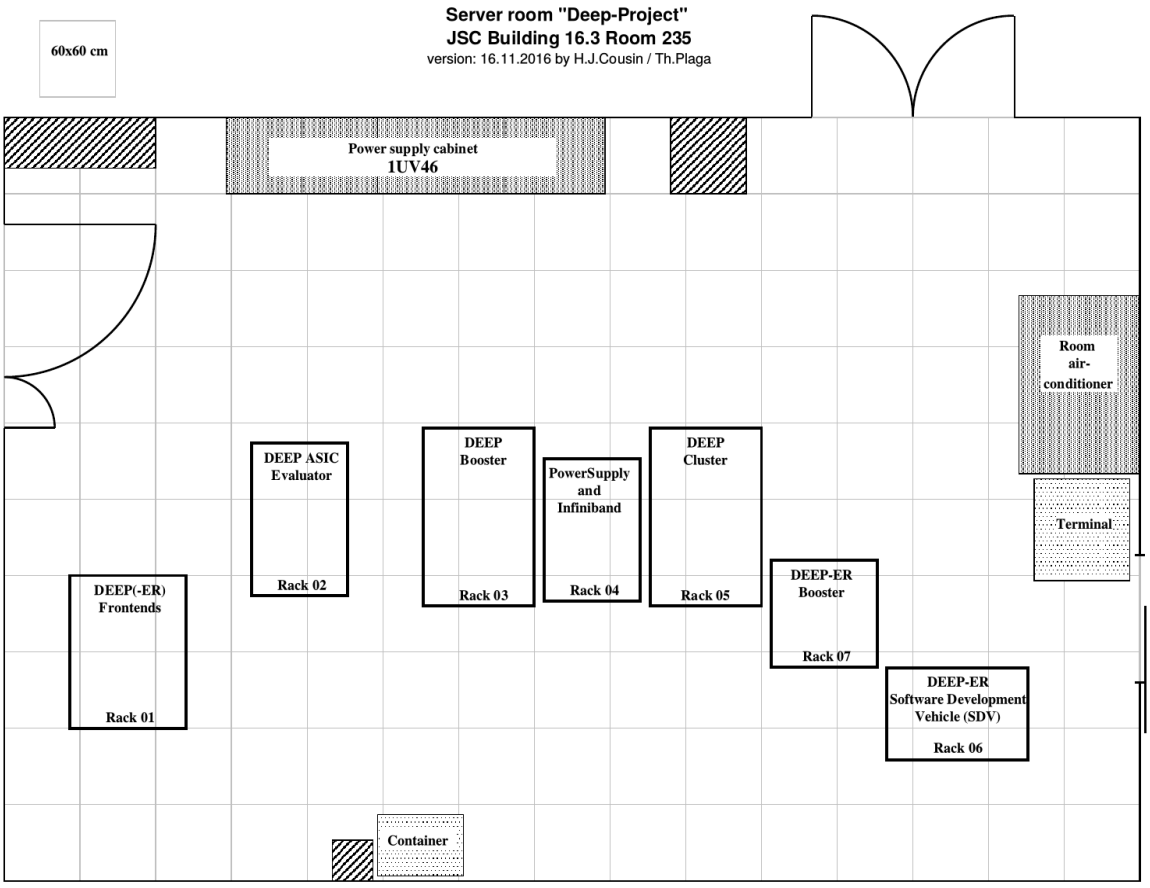


Figure 3: Floorplan of the DEEP and DEEP-ER installation room

The two racks are placed orthogonally with regards to the position of the EXTOLL network devices: for the SDV, the EXTOLL TOURMALET cards are located on the back side of the rack (left side in the floor plan), for the Booster chassis the connections relevant for the Cluster Booster interconnect are located on the right side. The positions of the two racks as illustrated in Figure 3 help to reduce the distance between the connections and hence to allow us to comply with the maximum EXTOLL cable length.

3.2 Power Supply

There are four spare power shelves from the DEEP system available that deliver 48V DC and can be re-used to supply the four DEEP-ER Booster chassis with power. Each of the power shelves hosts four rectifiers providing a maximum current of 3kW each. With the expected maximum power consumption of less than 8kW per chassis a 3 plus 1 redundancy can be implemented. This means that a chassis will stay online even if one of the corresponding rectifiers fails.

For remote power control, two additional Web-relay cards (see Figure 4) were added to the main power supply cabinet (shown in Figure 3) and connected to the rectifiers in the power shelves. Each Web-relay card exposes 10 relays of which 8 are in use (2 x 4 rectifiers). This means each Web-relay card can control the power of two chassis. The 10 relays on the cards can be operated independently, so the power of the single chassis can be controlled autonomously.

Figure 4: Web-relay card used for remote power control

Special power connectors are used for the power connection of the chassis. They can be plugged and unplugged very easily and avoid touching screwed cable connections to physically control the power connections. This is of particular importance for a very dense compute system like the DEEP-ER Booster where certain components like the power connection of a backplane cannot be reached easily. As shown in Figure 5 the connectors are mounted to the rack to guarantee easy access and safe fixation.

Figure 5: Power connectors used for the power supply of the Booster rack

3.3 Liquid Cooling

For the liquid cooling of the DEEP-ER Booster rack the system was attached to the existing cooling loop used by the DEEP Booster, since:

- Both systems have almost the same cooling requirements in terms of water quality and inlet temperatures;
- There is sufficient cooling capacity left to cover the waste heat of both systems;
- Implementing a separate cooling loop for the DEEP-ER Booster would, in addition to placing more pipes in the floating floor, have required adding two more heat exchangers to the existing loops, one for the connection to the outside dry cooler and one for the connection to the central cold water supply of the campus. Due to limited space in the engineering room, this would have been very challenging.

An existing diversion in the loop is used to implement a new branch line for the additional rack of the DEEP-ER Booster. Along with the pipes and valves forming the new branch line, a flow meter was integrated to check the flow rate for the DEEP-ER Booster rack. Furthermore, the existing filter in the main Booster cooling loop was replaced by a fine-grained 50 µm device. Figure 6 shows the overall layout of the cooling infrastructure in place for the DEEP and DEEP-ER hardware. In addition to the DEEP and DEEP-ER Booster racks the so called ASIC Evaluator from the DEEP project is connected to the Booster cooling loop whereas the DEEP Cluster part is using a separate loop. Both cooling loops are connected to an outside dry cooler and the central cold water supply available on the campus via heat exchangers.

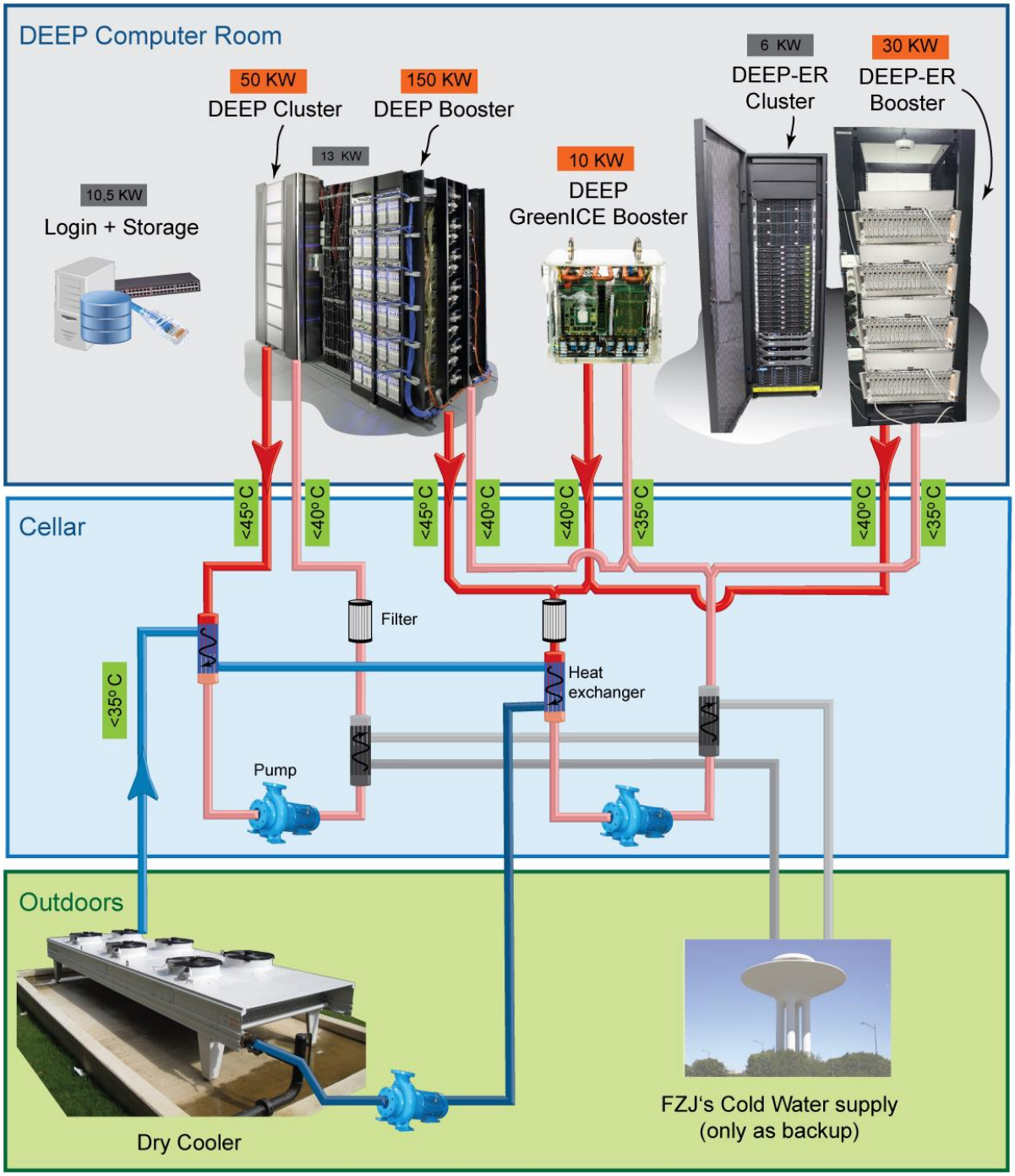


Figure 6: Schematics of the DEEP and DEEP-ER direct liquid cooling concept

The concept of having two cooling backends (an outside dry cooler and the central cold water supply), both of them capable of dissipating the waste heat of all liquid cooled systems, provides the highest flexibility for the inlet temperatures (they can be chosen in the range of 18 °C to about 35 °C.) In addition, the central cold water supply can serve as backup for the outside dry cooler on hot summer days or in case of technical problems (e.g. pump failure in the loop connected to the dry cooler).

3.3.1 Water Quality Analysis

As mentioned in the introduction a significant coolant leak occurred in the first chassis in December after about 3 weeks of continuous operation. The root cause of the leak was identified as pitting corrosion in three of the nodes (see also [5]).

Profound studies of the water quality in the cooling loop were done by frequently taking samples from the loop and comparing them with the results from the water quality analysis from December as well as synthetic samples of the coolant (before entering the loop). Additional analyses have been done at EUROTECH in Amaro. Table 3 provides a list of samples that have been taken to collect a reasonable amount of data for investigation. A visual timeline for the most important samples and future measurements can be found in D8.3.

It has to be mentioned that all measures performed introduce a potential deviation of the measured values of 10% with regards to the measurement process. This applies to the PH and conductivity values as well as to the determined chemicals and led us to focus on those parameters showing a reasonable deviation (much more than 10 %) from the given specs.

Date	Description
20.06.2016	Demineralised Water, JUELICH, outside loop
07.12.2017	Demineralised Water + 2% Protectogen C Aqua + 0.2% Mergal BIT 20X, JUELICH, from loop (before leak)
11.01.2017	Demineralised Water + 2% Protectogen C Aqua + 0.2% Mergal BIT 20X, JUELICH, from loop (after leak)
11.01.2017	Demineralised Water + 2% Protectogen C Aqua + 0.2% Mergal BIT 20X, JUELICH, outside loop
27.01.2017	Demineralised Water, JUELICH, from loop (for 1 day)
31.01.2017	Demineralised Water, EUROTECH, outside loop
03.02.2017	Demineralised Water + 2% Protectogen C Aqua, JUELICH, outside loop
03.02.2017	Demineralised Water + 2% Protectogen C Aqua, JUELICH, from loop (for 1 day)
21.02.2017	Demineralised Water + 2% Protectogen C Aqua, EUROTECH, outside loop
21.02.2017	Demineralised Water + 2% Protectogen C Aqua + 0.2% Acticide B40, EUROTECH, outside loop
01.03.2017	Demineralised Water + 2% Protectogen C Aqua, JUELICH, from loop (for 4 weeks)
16.03.2017	Demineralised Water + 2% Protectogen C Aqua, JUELICH, from loop (for 6 weeks)

Table 3: List of water quality samples that have been analysed in JUELICH and at EUROTECH

In a first step the water quality before the leakage was investigated. Therefore a sample that was taken from the cooling loop about 3 weeks before the leakage was studied and compared with a sample taken shortly after the leakage occurred. The results can be found in Table 4**Error! Reference source not found.**.

	Sample about 3 weeks before leakage	Sample shortly after leakage	*Recommended limits*
PH	9.1	9	*7 – 8 (6.5 - 8.5 tolerated)*
Conductivity [mS/cm]	2.6	2.6	*3.3-4.4 (when using Protectogen C Aqua)*
Chloride [mg/l]	182	19	*< 5*
Nitrate [mg/l]	76	37	*< 1*
Sodium [mg/l]	837	811	*< 20*
Sulphur [mg/l]	74	42	*No threshold available*
Sulphate [mg/l]	12,6	23	*< 10*

Table 4: Comparison of the two samples taken in JUELICH from the Booster loop before and after the leak.

As illustrated in deliverable D8.3 several values were not in the expected range at that time. Some of the chemical values such as sodium were way too high. This also applies to the PH value. With more than 9 it indicates a strong basic medium. Together with the high concentration of erosive components the liquid would indeed be able to cause severe corrosion as observed within the cold plates. Additionally, the low conductivity reveals that there was not enough corrosion inhibitor (Protectogen C Aqua) added to the loop as the given ratio of 1.5% - 2% (v/v) is expected to increase the conductivity to 3.3mS/cm – 4.4mS/cm when being added to demineralised water. This issue had already been discussed in December. The conclusion was to add more inhibitor with the next regular maintenance of the system, but could not be performed before the leakage.

While PH value, conductivity and the amount of sodium are almost unchanged before and after the leakage it is not yet clear why the amount of chloride decreases extensively in-between. Also it is not yet understood why the overall amount of sulphur atoms (including its chemical combinations) diminished whereas the amount of sulphate increases at the same time. The remaining parameters like total metals (irons, copper etc.) and turbidity did, however, meet the requirements for both samples. Only the amount of suspended solids – which is not included in the table - was slightly higher after the leakage, but this can be explained by particles entering the loop with the incidence.

Since it was not sure which factors have an impact on the chemical properties of the coolant and if materials in the loop are interacting with the cooling liquid the subsequent analyses were not only performed with samples taken from the cooling loop, but also applied to the coolant before entering the loop. Investigation was impeded by the complexity of the coolants chemical properties which is caused by using two additives: an inhibitor (Protectogen C Aqua from Clariant) and a biocide (Mergal BIT 20 X from Troy, which was suggested by EUROTECH as an alternative to the Nipacide 20 BIT from Clariant, that is unfortunately not available in Germany anymore). First the pure demineralised water was checked to be excluded as potential source of errors. Analyses showed that the demineralised water both in JUELICH and at EUROTECH fulfils all the specifications regarding water quality. To detect potential influences by the JUELICH Booster cooling loop itself (exposing a mixture of different materials) the analysis was repeated after inserting the demineralised water into the loop and taking a sample again after 1 day of circulation. Results are listed in Table 5.

	Demineralised Water EUROTECH	Demineralised Water JUELICH	Demineralised Water JUELICH (after 1 day of circulation in the loop)	*Recommended limits*
PH	6.12	7.4	7.9	*7 – 8 (6.5 - 8.5 tolerated)*
Conductivity [mS/cm]	0.00	0.00	0.03	*3.3-4.4 (when using Protectogen C Aqua)*
Chloride [mg/l]	< 0.1	< 0.1	2.71	*< 5*
Nitrate [mg/l]	0.1	< 0.1	0.5	*< 1*
Sodium [mg/l]	0.16	0.5	5.8	*< 20*
Sulphur [mg/l]		<0.01	1.6	*No threshold available*
Sulphate [mg/l]	0.4	<0.01	< 0.1	*< 10*

Table 5: Analyses of demineralised water in JUELICH and at EUROTECH.

Compared to the pure demineralised water (before entering the circuit) the sample taken from the loop after one day of circulation shows a slightly reduced water quality, but still fulfils the requirements. Although the loop was washed before starting the measurements it is possible that still some residuals (e.g. of chloride and sulphur) were present in the loop at that time. So it is hard to make any conclusion about potential sulphur or chloride sources within the loop yet.

After demonstrating that the demineralised water itself is fulfilling the water quality requirements the next step was to introduce the chemicals that have to be added for protection of the aluminium (inhibitor: Protectogen C Aqua) and to avoid formation of bacteria and algae (biocide: Mergal BIT 20 X). Table 6 compares the samples for pure demineralised water, demineralised water with inhibitor added and the complete mixture of demineralised water, inhibitor and biocide. None of these samples was circulated in the cooling loop.

	Demineralised Water	Demineralised Water + 2% Inhibitor	Demineralised Water + 2% Inhibitor + 0.2% Biocide	*Recommended limits*
PH	7.4	8.0	8.2	*7 – 8 (6.5 - 8.5 tolerated)*
Conductivity [mS/cm]	0.00	4.8	4.9	*3.3-4.4 (when using Protectogen C Aqua)*
Chloride [mg/l]	< 0.1	10.4	2.4	*< 5*
Nitrate [mg/l]	< 0.1	195	82	*< 1*
Sodium [mg/l]	0.5	1720	1810	*< 20*
Sulphur [mg/l]	<0.01	1.1	49	*No threshold available*
Sulphate [mg/l]	<0.01	0.1	0.19	< 10

Table 6: Impact of chemicals being added to demineralised (outside the loop).

Obviously both chemicals tend to impact the water quality. Conductivity is increased by adding the corrosion inhibitor. This is an expected behaviour due to the properties of the product. The vendor specifies a conductivity range of 3.3mS/cm – 4.4mS/cm when using a concentration of 1.5% - 2.0% v/v. Considering a concentration of 2% applied to the sample and the potential divergence of 10% for the measured values a conductivity of about 4.8mS/cm can safely be tolerated (hence marked yellow). Also the chloride is most likely being introduced by the inhibitor, although it is not evident, why the chloride value measured is much less in the full mixture (demineralised water, inhibitor and biocide). The situation is similar for the sodium and nitrate values. Since the inhibitor is among other chemical compounds built of sodium nitrate these two values consequentially raise compared to the pure demineralised water. Sulphur is clearly inserted by adding the biocide. It is not yet known, which kind of chemical compounds form out of the high amount of sulphur and there is now threshold available for the overall amount of sulphur (only sulphates and sulphides). However, the amount of sulphur seems to have only little impact on the sulphate rate which is still in the given range.

Due to the increased chemical complexity of the coolant when using the two additives and with regards to the current inlet temperature of 20°C (where it is quite unlikely to be faced with growth of bacteria and algae) we decided to start with a mixture of demineralised water and inhibitor (which is a mandatory requirement) and frequently check the water quality by taking samples from the loop. The following Table 7 illustrates the evolution of the critical parameters over time compared to pure demineralised water (that was proven to fulfil all the requirements). As with the previous tables the recommended limits are included in the last column.

	Demineralised Water	**Demineralised + 2% Protectogen**	**Demineralised + 2% Protectogen in the loop for 1 day**	**Demineralised + 2% Protectogen in the loop for 4 weeks**	**Demineralised + 2% Protectogen in the loop for 6 weeks**	***Recommended limits***
PH	7.4	8.0	8.0	8.0	7.9	*8*
Conductivity [mS/cm]	0.00	4.8	4.3	4.3	4.3	*4.4*
Chloride [mg/l]	0.1	10.4	9.4	1.4	1.74	*5*
Nitrate [mg/l]	0.1	195	279	302	231	*1*
Sodium [mg/l]	0.5	1720	1520	1530	1420	*20*
Sulphur [mg/l]	0	1.1	4.8	9.8	11.3	
Sulphate [mg/l]	0	0.1	0.1	0.9	0.9	*10*

Table 7: Behaviour of water quality in JUELICH Booster loop when using demineralised water plus inhibitor.

In the given time range of 6 weeks, from mixing and inserting the coolant to the last measurement, the PH and conductivity values are quite constant. This is an important conclusion since those two parameters are expected to be the most important once for indication of appropriate conditions in the cooling loop. Furthermore, the conductivity value can be taken as criteria for the right amount of inhibitor being present in the loop. The Protectogen C Aqua tends to be consumed over time. In the JUELICH environment this means that action is required if the conductivity might fall below 4.3mS/m. This would mean that some more inhibitor has to be added to ensure protection of the noble metals in the cooling loop. Along with the PH and conductivity values the chloride portion is below the recommended thresholds with the later samples, although it was about 2 times higher when adding the inhibitor. In contrast, the values for sodium (and nitrate) are still much too high, even after 6 weeks of circulation in the loop. The amount of sulphur even increases over time, part of it seems to appear as sulphates (also slightly increasing over time). With regards to the fact that no biocide was included this is something to further investigate. Possibly there is something inside the loop emitting sulphur over time. But more samples will be needed to prove this assumption. Another open question is the high amount of sodium (and nitrate) being inserted by the inhibitor. The values measured at EUROTECH when adding 2% v/v of Protectogen C Aqua to demineralised water are quite different (sodium: 630 mg/l, nitrate: 582 mg/l, chloride: 0.4 mg/l). According to the vendor (Clariant) the composition of the product can slightly change depending on the supplier for the product, but this fact cannot fully explain the large divergences. From the available samples there is no tendency crystallising yet. However, since the inhibitor, which seems to introduce those chemicals, is needed for protecting the noble metals from the demineralised water there is not much to be done about it.

Thus, the chemical behaviour of the coolant is not fully understood yet. Therefore, it is very important to continue monitoring the water quality within the cooling loop thoroughly. Samples will be taken bi-weekly to allow for quick interaction if needed and to prevent any (further) parameters to run out of scope. For continuous measures of the PH and conductivity values remotely accessible sensors are planned to be attached to the loop in near future.

After washing the loop the DEEP-ER Booster branch was closed to independently operate the DEEP part of the cooling loop in the meantime. Once the DEEP-ER branch is integrated again, the water quality has to be checked to see if the integration has any immediate impact. Using demineralised water with inhibitor only (and omitting the biocide) PH and conductivity values as well as most of the chemicals are found in the operation envelope of the EUROTECH cold plates. Hence, EUROTECH and JUELICH agreed that with careful monitoring of the water quality the DEEP-ER Booster rack can safely be taken into operation again under the current conditions, even though some questions are still open (e.g. the high amounts of sodium and sulphur in the loop). Once the parameters can be declared stable, biocide will be added again. Meanwhile EUROTECH has tested an alternative product from vendor Clariant: Acticide B 40. It is expected to be compatible with the inhibitor. First experiments have shown that this product has only little impact on the water quality and can most likely be used to meet the given requirements in the loop. For further details and the operation instructions for liquid cooled system see also D8.3 and D8.4.

3.4 Chassis Integration

The DEEP-ER Booster rack was delivered pre-assembled with four backplanes including power connectors and internal water distribution columns where the compute nodes can be easily attached using quick disconnectors.

First of all the rack internal water distribution columns for the chassis were connected to the new branch line in the cooling loop at the installation site. To do this, two manifolds were positioned in the floating floor exposing six connections with separate valves. Four of them are used for the Booster chassis; two of them are spare and can, for example, be used to bypass the rack by bridging the inlet and outlet manifolds. Before connecting the compute nodes to the quick disconnectors, the new branch line and the internal distribution columns had to be cleaned by filling the loop and using a small bypass bridging the internal distribution columns (see Figure 7).

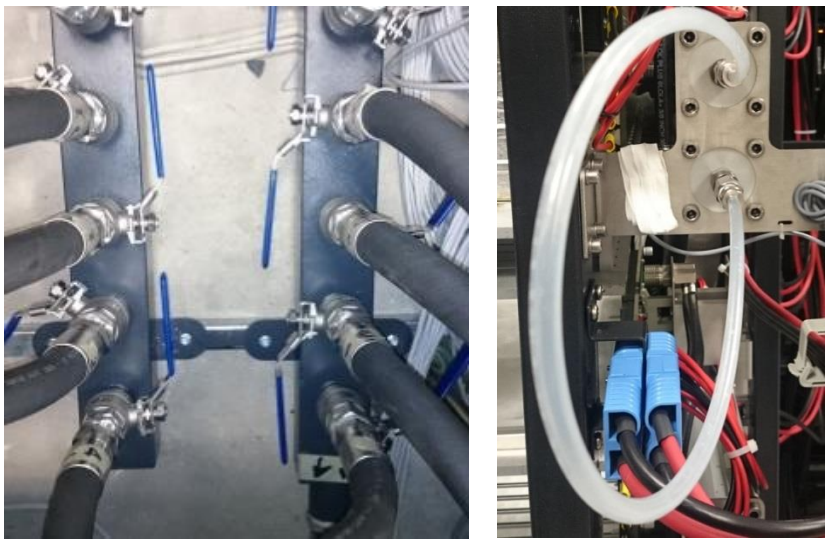


Figure 7: Manifolds in floating floor and bypass using the rack internal distribution columns

With the liquid cooling in place, the chassis frameworks were added. Each chassis holds one root card and 18 Booster Nodes equipped with an EXTOLL TOURMALET and an NVMe device. The PCIe devices are located on top of the root card behind a front cover. Figure 8 on page 18 shows the first chassis including Booster Nodes (BN) and root card. Reference [4] contains a full description of the Eurotech Aurora Booster system architecture and of the components.

Figure 8: DEEP-ER Booster chassis with 18 nodes and one root card

At the time of writing only the first chassis was installed in the rack. Current time line foresees the remaining components to arrive in JUELICH at 31st of March and to perform the integration of components and bring-up of the full system (72 compute nodes) during April. The installation procedure for the remaining 3 chassis will rely on the same activities described for the first chassis installation within this document and is expected to be straightforward.

3.5 Rack Sensors and Monitoring

Two leakage sensors (placed underneath the quick disconnectors of the chassis) and one smoke sensor were already included in the rack. Remote sensor access is possible through an associated sensor reader providing an SNMP interface. The reader is integrated in the rack and connected to the DEEP and DEEP-ER administration network (see section 4.1). The leakage and smoke sensors were added to a central monitoring system which also includes the infrastructure sensors at the installation site, e.g. pump states, temperatures and flow rates. The monitoring system is configured to trigger automatic power offs for affected components in case of alarms for critical sensors. This applies to the integrated rack sensors as well.

Figure 9: Rack mounted sensor reader and smoke detector

In addition to the infrastructure and rack sensors focusing on the liquid cooling status and potential emergency cases (e.g. formation of smoke), the computer hardware itself is also closely monitored. The status of the nodes is frequently checked by reading and evaluating IPMI sensors through the baseboard management controllers (BMC). A chassis protection system shipped with the hardware is running on the root card of each chassis. It observes the board temperatures and can power off single components or a full chassis in case of overheating.

4 Software Installation and System Configuration

To enable the DEEP-ER Prototype system to be put into operation, installation of software and validation of configuration is necessary. The DEEP-ER Booster Nodes were shipped with the latest firmware available for the underlying KNL S7200AP (code name “Adams Pass”) boards (see [6] for specification) and a recent CentOS version; several modifications were required to integrate the nodes into the JSC environment. This mainly affects network setup and the use of shared filesystems.

4.1 Network Setup

The DEEP system already uses several networks for different aspects of cluster operation into which the components of the DEEP-ER Prototype system could be integrated:

- A management network (10.2.8.0/22) which carries the common traffic between servers and clients (used for batch system, ssh access, etc.) and the ParaStation Management communication (e.g. MPI task start, monitoring and termination);
- An IPMI network (10.2.12.0/22) which carries the traffic between the master servers (front ends) and the cluster hardware BMC management and can be used for power cycling nodes, SOL console access and collection of IPMI sensor data;
- The EXTOLL network is used between the compute nodes as a fast interconnect for running HPC applications. The DEEP-ER system runs an EXTOLL topology physically separate from the one used in DEEP. Both EXTOLL setups are able to do IPoverEXTOLL in a shared network range (10.2.20.0/22). IPoverEXTOLL can be enabled or disabled on boot for both systems independently.

The BMCs are configured to use DHCP for getting an IP address in the IPMI network whereas the compute nodes and servers use fixed addresses in the management network.

4.2 Firmware and Operating System

The Cluster Nodes located in the SDV were shipped with an up-to-date BIOS and BMC firmware version, as were the Booster Node cards delivered by Eurotech. For the KNL S7200AP board the latest BIOS and BMC firmware versions available from Intel were pre-installed while the BMC firmware for the root cards was provided by Eurotech. Table 8 lists the currently installed BMC firmware and BIOS versions:

Cluster Node BIOS version	5.6
Cluster Node BMC firmware version	1.81
Booster Node BIOS version	S72C610.86B.01.01.0104.032220161509
Booster Node BMC firmware version	0.17
Root card BMC firmware version	0.02

Table 8: Firmware and BIOS versions used in the DEEP-ER Prototype System

While no firmware updates were required for taking the nodes into operation, some re-configuration had to be done for the BIOS parameters. These were needed for:

- Remote console and Serial over LAN (SOL);
- Turbo and power saving modes;
- PXE boot.

A system imaging tool is used to simplify the installation of system software and to manage the operating system itself. The OS image is shared between the Cluster and the Booster Nodes as well as the SDV KNL nodes to provide a consistent environment on all compute nodes. It is based on CentOS 7.2, but uses the kernel of the latest Intel® Xeon Phi™ Processor Software for Linux (XPPSL) which is part of the image. Table 9 lists the properties of the OS image used. The file servers for the local storage integrated in the SDV rack use a modified version of the OS image without the XPPSL software but the standard CentOS kernel is used. The root cards in the DEEP-ER Booster come with a pre-installed Ubuntu operating system (15.10 Wily Werewolf).

Operating system	CentOS 7.2 (64bit)
Kernel version	3.10.0-327.36.3.el7.xppsl_1.4.3.3482.x86_64
XPPSL version	1.4.3

Table 9: OS image properties used on the DEEP-ER compute nodes

To allow the SDV and the Booster Nodes to perform a PXE boot a TFTP server runs on one of the front end nodes. In case an updated image is available for a node it is automatically rolled out at boot, else the node will boot from its local disk. Shared filesystems are used in addition to the system imager to have the same data available on all nodes across the full system. For example, the user local filesystem (/usr/local) and the user homes are provided through the General Parallel Filesystem (GPFS) available at JSC. For this purpose a GPFS client is running on the front ends of the DEEP and DEEP-ER systems that export the filesystems via NFS. In addition, the nodes were integrated into the existing batch system to make them accessible to the users.

4.3 Memory and Storage Access

In addition to the local disk filesystems on the compute nodes and the globally available GPFS filesystems, the DEEP-ER Prototype system exposes different types of memory and storage that can be used by applications running on the system. These are:

- Non-volatile memory (NVMe):
  - Attached to each node as PCIe card
  - Size: 400 GB
  - Directly accessible or through BeeGFS on Demand

- Network attached memory (NAM):
  - Connected to the EXTOLL network
  - Size: 2 GB
  - Accessible via libNam API

- Local storage:
  - Integrated in the SDV rack (1 x RAID + 3 x fileservers)
  - Size: 144 TB
  - Accessible through BeeGFS (clients installed on the nodes)

The configuration of the different memory and storage types was implemented in collaboration with WP4. For further information about the memory and storage access as well as their use cases see also [2] and [3].

4.4 First System Tests

The Cluster nodes located in the SDV have already been used by the application and software developers for a while and did operate in a production environment. This section therefore focuses on the tests regarding the DEEP-ER Booster hardware specifically the first Booster chassis installed at JUELICH at the end of 2016.

First, the network access was tested for the compute nodes and the root card. Having successfully checked the network connectivity for all components, the boot procedure via the

root card was tested using the appropriate software tools provided by Eurotech. All 18 nodes in the first chassis were able to successfully boot their OS images.

The next step was to run a high performance Linpack (HPL) benchmark on all Booster Nodes simultaneously to check if the cooling was working correctly and to identify potential performance issues on the nodes. To perform the test the Linpack implementation from the Intel micperf tool included with the XPPSL KNL software stack was used and the die and board temperatures measured. All 18 nodes finished the Linpack benchmark successfully showing the expected performance numbers (about 1.8 TFlop/s). None of the nodes showed any thermal events during the test and the maximum die and board temperatures did not reach any critical values.

Following the “burn-in” tests, it was planned to start with testing the PCI Express connectivity between the Booster KNL boards and the NVMe and TOURMALET NIC cards. The first version of the root card installed at JUELICH is known to have problems in this area, and expectation was to test a subset of Booster nodes only, deferring validation of the others to the installation of the second and improved root card version in January 2017.

A coolant liquid leak stopped the test activities. An in-depth Investigation of the causes for the leak is going on, and first findings are that the leakage occurs due to corrosion in the cold plates themselves (see details in [5]). Eurotech and JUELICH are closely collaborating to drill down to why chemical corrosion is occurring, and to find ways how to prevent this in the future. The chemical processes in the coolant water are complex, since two additives (for corrosion protection and for suppressing biological contamination) have to be added to the water, and since the cooling loop contains a mixture of metallic and non-metallic materials.

Once the corrosion problem is considered solved and the system can be put back into operation, the configuration and installation tests will continue:

- Check PCIe connections to the EXTOLL and NVMe cards for all KNL boards, using the second and improved version of the root card;
- Check NVMe read and write performance on the Booster and the Cluster Nodes;
- Setup EXTOLL topology and perform link tests for all connections (including the Cluster nodes and the interconnect between the two parts);
- Start testing applications using both parts of the system through the batch system.

As the remaining DEEP-ER Booster chassis are added, they will be subjected to the same sequence of configuration and tests as the first chassis, and the full 72-node EXTOLL topology will be built for the DEEP-ER Booster.

References and Applicable Documents

[1] http://www.deep-er.eu
[2] DEEP-ER Deliverable D4.4 – IO Software packages
[3] DEEP-ER Deliverable D5.3 – Resiliency Software
[4] DEEP-ER Deliverable D8.2 – Components of Aurora Blade prototype for DEEP-ER
[5] DEEP-ER Deliverable D8.3 – Aurora Blade Booster Prototype
[6] Technical Product Specification for Intel® Server Board S7200AP Family.
URL http://www.intelserveredge.com/assets/S7200AP_HNS7200AP_TPS_R1_0.pdf

List of Acronyms and Abbreviations

A

API: Application Programming Interface

ASIC: Application Specific Integrated Circuit, Integrated circuit customised for a particular use

Aurora: The name of Eurotech's cluster systems

B

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.

BIOS: Basic I/O system. Boot and system initialisation code run before the OS starts

BMC: Board management controller. Used to physically monitor and manage a compute blade.

BN: Booster Node (functional entity); refers to a self-booting KNL board (Node board architecture) including the NVM and NIC devices connected by PCI Express or a Brick (Brick architecture).

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

C

CINECA: Consorzio Interuniversitario, Bologna, Italy

CPU: Central Processing Unit

D

DEEP-ER: DEEP Extended Reach: this project

DEEP-ER Booster: Booster part of the DEEP-ER Prototype, consisting of all Booster Nodes and the NAM devices.

DEEP-ER Interconnect: High performance network connecting the Booster and Cluster nodes, the NAM and service nodes with each other to form the DEEP-ER Prototype.

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

Eurotech: Eurotech S.p.A., Amaro, Italy
Exascale: Computer systems or Applications, which are able to run with a performance above 10^{18} Floating point operations per second
EXTOLL: High speed interconnect technology for cluster computers developed by University of Heidelberg

F

G

GPFS: General Parallel File System (GPFS), a high-performance clustered file system developed by IBM

H

HMC: Hybrid Memory Cube
HPC: High Performance Computing
HW: Hardware

I

Intel: Intel Germany GmbH Feldkirchen,
I/O: Input/Output. May describe the respective logical function of a computer system or a certain physical instantiation
IPMI: Intelligent Platform Management Interface

J

JUELICH: Forschungszentrum Jülich GmbH, Jülich, Germany
JSC: Jülich Supercomputing Centre (belongs to JUELICH)

K

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

L

LAN: Local area network.
LINPACK: Software library to perform numerical linear algebra calculations used as benchmarks.

M

MCDRAM: Multi-Channel DRAM, a high bandwidth on package memory on KNL
MCE: Machine Check Exception

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

NFS: Network File Service, used to provide file access to remote hosts through a network

NIC: Network Interface Card, Hardware component that connects a computer to a computer network

NVM: Non-Volatile Memory. Used to describe a physical technology or the use of such technology in a non-block-oriented way in a computer system

NVMe: Short form of NVM-Express

NVM-Express: An interface standard to attach NVM to a computer system. Based on PCI Express it also standardises high level HW interfaces like queues.

O

OS: Operating System

P

ParaStation MPI: Software for cluster management and control developed by ParTec

ParTec: ParTec Cluster Competence Center GmbH, Munich, Germany

PCI: Peripheral Component Interconnect, Computer bus for attaching hardware devices in a computer

PCIe: Short form of PCI Express

PCI Express: Peripheral Component Interconnect Express started as an option for a physical layer of PCI using high-performance serial communication. It is today's standard interface for communication with add-on cards and on-board devices, and makes inroads into coupling of host systems. PCI Express has taken over specifications of higher layers from the PCI baseline specification.

PXE: Preboot eXecution Environment; used for network boot

Q

R

Rack: Compartment to mechanically assemble multiple chassis to form the final computer

RAID: Redundant Array of Independent Discs

S

SDV: Software Development Vehicle: a HW system to develop software in the time frame where the DEEP-ER Prototype is not yet available.
SNMP: Simple network management protocol
SOL: Serial Over Lan
SW: Software

T

TFlop/s: Teraflop, 10^{12} Floating point operations per second
TFTP: Trivial File Transfer Protocol

U

V

W

X

Y

Z