DEEP

Dynamical Exascale Entry Platform

Grant Agreement Number: 287530

D6.5
Installation of scale-up Booster and ASIC Evaluator

Approved

Version: 2.0
Author(s): J. Kreutz, JUELICH
Date: 02.10.2015
Project and Deliverable Information Sheet

DEEP Project

<table>
<thead>
<tr>
<th>Project Ref. №:</th>
<th>287530</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Title:</td>
<td>Dynamical Exascale Entry Platform</td>
</tr>
<tr>
<td>Project Web Site:</td>
<td><a href="http://www.deep-project.eu">http://www.deep-project.eu</a></td>
</tr>
<tr>
<td>Deliverable ID:</td>
<td>D6.5</td>
</tr>
<tr>
<td>Deliverable Nature:</td>
<td>Report</td>
</tr>
<tr>
<td>Deliverable Level:</td>
<td>PU*</td>
</tr>
<tr>
<td>Contractual Date of Delivery:</td>
<td>30/June/2015</td>
</tr>
<tr>
<td>Actual Date of Delivery:</td>
<td>30/June/2015</td>
</tr>
<tr>
<td>EC Project Officer:</td>
<td>Luis Carlos Busquets Pérez</td>
</tr>
</tbody>
</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>
<thead>
<tr>
<th>Document</th>
<th>Title: Installation of scale-up Booster and ASIC Evaluator</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID:</td>
<td>D6.5</td>
</tr>
<tr>
<td>Version:</td>
<td>2.0</td>
</tr>
<tr>
<td>Status:</td>
<td>Approved</td>
</tr>
<tr>
<td>Available at:</td>
<td><a href="http://www.deep-project.eu">http://www.deep-project.eu</a></td>
</tr>
<tr>
<td>Software Tool:</td>
<td>Microsoft Word</td>
</tr>
<tr>
<td>File(s):</td>
<td>DEEP_D6.5_Installation_scaledup_Booster_andASICEvaluator_v2.0-ECapproved</td>
</tr>
</tbody>
</table>

Authorship

| Written by: | J. Kreutz, JUELICH |
| Contributors: | |
| Reviewed by: | G. Staffelbach (CERFACS), I. Schmitz (ParTec) |
| Approved by: | BoP/PMT |

Document Status Sheet

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Status</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>30/June/2015</td>
<td>Final</td>
<td>EC submission</td>
</tr>
<tr>
<td>2.0</td>
<td>02/October/2015</td>
<td>Approved</td>
<td>EC approved</td>
</tr>
</tbody>
</table>
### Document Keywords

| Keywords                  | DEEP, HPC, Exascale, prototype installation |

---

### Copyright notices

© 2011 -2015 DEEP Consortium Partners. All rights reserved. This document is a project document of the DEEP project. All contents are reserved by default and may not be disclosed to third parties without the written consent of the DEEP partners, except as mandated by the European Commission contract 287530 for reviewing and dissemination purposes. All trademarks and other rights on third party products mentioned in this document are acknowledged as own by the respective holders.
Table of Contents

Project and Deliverable Information Sheet ................................................................. i
Document Control Sheet ............................................................................................... i
Document Status Sheet ............................................................................................... i
Document Keywords .................................................................................................... ii
Table of Contents......................................................................................................... iii
List of Figures .............................................................................................................. iv
List of Tables ................................................................................................................ iv
Executive Summary ...................................................................................................... 5
References and Applicable Documents ......................................................................... 21
List of Acronyms and Abbreviations ............................................................................ 22
List of Acronyms and Abbreviations ............................................................................ 22
List of Figures

Figure 1: DEEP Cluster and Booster components ................................................................. 6
Figure 2: DEEP ASIC Evaluator components ........................................................................ 7
Figure 3: Final cooling concept for the DEEP System .............................................................. 9
Figure 4: Remote power control for the DEEP ASIC Evaluator via Web-GUI .......................... 9
Figure 5: Anderson Power connectors used for backplane power supply ............................. 10
Figure 6: Improved isolation for backplane connections ........................................................ 10
Figure 7: EXTOLL 3D torus topology for the DEEP Booster .................................................. 11
Figure 8: Molex cables implementing the 3D EXTOLL torus network ................................ 12
Figure 9: Fully assembled DEEP Booster with network cabling ............................................ 12
Figure 10: BNC and EXTOLL integration into the existing networks ................................. 14
Figure 11: Exemplary output of the EXTOLL link test tool ..................................................... 19
Figure 12: USB cabling between Juno and BIC boards to update ARM7 firmware ............... 20
Figure 13: Graphical User Interface of Altera tool for updating FGPGA firmware image on BICs .... 20
Figure 14: JTag cabling using USB blaster to perform FPGA update of the BIC ................. 21

List of Tables

Table 1: List of IP subnets for the DEEP System and the ASIC Evaluator ............................. 15
Table 2: Firmware versions for the BNC components .............................................................. 16
Table 3: Firmware versions for the Juno-BIC components ..................................................... 16
Executive Summary

Installing the DEEP System applying the novel Cluster Booster Architecture and including an appropriate software stack is one of the major goals of the DEEP project. The Booster part and the appending software stack were developed within the different work packages of the project whereas the DEEP Cluster part, already installed in Jülich at end of March 2013 (by month 16), is built on Eurotech’s Aurora platform [9], which is available as a commercial product. Both, the Cluster and the Booster, are designed for hot water liquid cooling to allow for energy efficient operation of the system and for investigation of energy efficiency aspects in HPC. Similar to the Cluster part, the installation of the DEEP Booster includes two main tasks, the hardware and the software installation which covers several actions to be taken during installation process.

- **Hardware installation:**
  The Booster Rack including water connections and power supply has to be integrated into the existing infrastructure in Jülich. The components for each chassis have to be assembled into the rack and to be connected to the existing Ethernet (Eth) and InfiniBand (IB) networks. In addition the cabling for the EXTOLL torus network has to be implemented inside the rack.

- **Software installation and configuration**
  Since the Booster Nodes are more complex in set-up the software installation is expected to be more elaborate compared to the Cluster Nodes. The installation process has to cover firmware installation for several of the Booster Node Card (BNC) components, e.g. the EXTOLL FPGA and the Baseboard Management Controller (BMC), and to coalesce all software parts developed in the remaining work packages, namely WP4, WP5 and WP7. This includes modified versions of ParaStation MPI and the PBS batch system, the Cluster-Booster protocol and the global RAS plane software used for system monitoring and energy efficiency investigation. Software installation affects most of the components within the DEEP System: the BNCs, the management servers and the Booster Interface Cards (BIC). In addition the integration into, and the interaction between the different networks (Eth, IB, EXTOLL) have to be thoroughly set-up and implemented. Systematic tests have to be included for the different stages in the installation process.

The installation of the DEEP Booster already started in autumn 2014 with two first half chassis (see also D6.4). The remaining chassis have to be integrated to achieve the fully assembled DEEP Booster with 192 Booster Node Cards (384 KNCs). Along with the scale-up DEEP Booster the scale-up version of the DEEP ASIC Evaluator has to be performed, which will host 64 additional Booster Nodes (KNCs) using EXTOLL Tourmalet cards for the interconnect and an innovative immersion cooling concept. The software setup on the ASIC Evaluator will be similar to the DEEP Booster relying on the software parts developed within the project. Details on the installation activities for both, the hardware and the software installation, of the two scale-up subsystems will be described within this document. It is meant to complement the D6.4 deliverable describing the installation and bring-up of the first Booster chassis in Jülich. Hence the focus will be on scale-up and the installation activities that have been performed during the last six month.
1 Introduction

The DEEP System integrates two main parts: a Cluster part and a Booster part. While the Cluster part will process the less scalable code parts, the Booster part with its fast EXTOLL network is meant to execute the highly scalable code parts of an application and to provide substantial speedup in execution time. Each Booster Node Card contains two logical Booster Nodes, each with one Intel® Xeon Phi™ (Knights Corner, KNC) and one FPGA implementing the EXTOLL network. Special Booster Interface Cards (BICs) will be used to manage the KNC cards within the Booster Nodes and to interconnect the Cluster and Booster parts by mapping the InfiniBand network used within the Cluster to the EXTOLL network of the Booster.

![Diagram of DEEP Cluster and Booster components](image)

Figure 1: DEEP Cluster and Booster components

A detailed description of the Booster components and the final DEEP System can be found in D3.5. The properties and the installation of the DEEP Cluster as well as the cooling concept for the final system allowing for energy efficient hot liquid cooling have already been described in D6.3. The D3.6 document focuses on the scale-up of the DEEP Booster.

With the last project review in January 2015 the decision was taken to perform a scale-up of the DEEP ASIC Evaluator to host 64 Booster Nodes (KNCs). At that time the installation of the DEEP ASIC Evaluator had already started and the rack including the Pseudo-BIC servers
Installation of scale-up Booster and ASIC Evaluator

(providing the BIC functionalities) as well as a small “Booster” with 4 KNCs was available. Nevertheless some modifications are necessary to implement a scale-up of this system.

Installation and configuration of the various components included in the software stack that has been developed within WP4 and WP5 as well as the RAS plane implementation designed in WP7 is of particular importance for the bring-up of the DEEP System and the scale-up DEEP ASIC Evaluator. Though the different software parts have already been tested individually within the single work packages using different prototype machines and evaluators, their interaction and the functionality of the software stack as a whole has not been fully verified. This requires a sophisticated approach for the installation process and the deployment of the scale-up systems including extensive tests.

The purpose of this document is to describe the installation process for the scale-up DEEP Booster and the DEEP ASIC Evaluator. Two main topics will be covered: the hardware installation starting with the integration of the revised Booster rack into the existing infrastructure as well as the integration of the Booster Nodes into the ASIC Evaluator and the software installation including system configuration. Functionality tests of the hardware components and the software stack will be presented as well. Section 2 summarises the preparative activities required for the scale-up of both systems. The hardware installation of the scale-up DEEP Booster and the ASIC Evaluator is focused in Section 3 and covers the assembly of the Booster chassis into the rack and the integration of the immersion cooled Booster Nodes into the ASIC Evaluator rack. Section 4 describes the software installation and the bring-up activities including different levels of firmware and software. Tests that have been performed during the installation process and their outcome will be presented in Section 5.
2 Preparative actions

Several preparative actions including the installation and integration of the Cluster part were already accomplished with the deliverables D6.1, D6.2, D6.3 and D6.4 before. Hence the installation activities described in this deliverable report started from those previous experiences.

2.1 Floor space, cooling and power supply

In addition to the DEEP System the DEEP ASIC evaluator requires liquid cooling as well, where inlet temperatures from 18°C to 35°C can be used. Hence it was positioned in the same installation room next to the DEEP Booster which allows for access to the same infrastructure in terms of power supply, cooling and network integration. The total weight of the rack is much lower than for the DEEP Booster, thus no additional support frame within the floating floor was needed.

Since some of the cooling components within the DEEP ASIC Evaluator are made of copper and the cold plates within the DEEP System are made of aluminium, a separate cooling loop for this subsystem would be the best option to avoid a material mix within the tubing. But the effort for the installation of an extra cooling loop for the DEEP ASIC Evaluator would be very high. One more pump and two more heat exchangers are needed for connecting it to the outside dry cooler as well as to the central cold water supply at JUELICH. In addition the expected lifetime for the DEEP ASIC Evaluator is rather short due to its prototype character. The risk of negative effects on the components caused by the material mix of copper and aluminium within the same loop are considered negligible in that time frame. Additionally the ranges of proper inlet temperature for both systems are pretty much the same. The DEEP ASIC Evaluator allows for inlet temperatures of up to 35°C. This is still enough for doing investigation on energy efficiency and warm water cooling on the DEEP Booster and to allow for using free cooling (almost) all over the year. Hence the decision was taken to attach this system to the DEEP Booster cooling loop instead. Of course, both subsystems have to get separate valves for decoupling from the cooling loop for maintenance purposes (i.e. stopping the cooling) on one of the subsystems while the remaining one is kept running. Aside to these modifications an overpressure valve and a loopback were added to the DEEP Booster cooling loop. They are needed to mitigate the high capacity of the pump included in that loop which was designed to work with a fully assembled Booster rack hosting 512 KNCs (the original planned Booster size). Having these components added the cooling loop can be used for an arbitrary number of chassis between 1 and 16 and has also enough capacity to include the DEEP ASIC Evaluator. Figure 3 illustrates the final DEEP cooling concept highlighting the modifications that were needed for the scale-up of the DEEP Booster and the DEEP ASIC Evaluator.

The WebRelay cards used for the Cluster part turned out to be very useful. They form the basis for the self-protection software to perform emergency power offs for certain components in case of critical over temperature. These WebRelay cards were also included for the 12 Booster chassis. Six additional WebRelay cards were installed: each of them controls the power for 2 Booster chassis. The DEEP ASIC Evaluator uses a separate power supply chain. The power distribution unit within the rack already includes remote power control features similar to the WebRelay cards for the DEEP System. Aside from remote power on/off current power consumption can be accessed on a website (Figure 4).
Figure 3: Final cooling concept for the DEEP System

Figure 4: Remote power control for the DEEP ASIC Evaluator via Web-GUI
To integrate all of the components of both scale-up subsystems in the different network fabrics (Eth, IB, EXTOLL) an IP and subnet scheme had to be planned. Further details on the IP setup including the different network types can be found in section 4.1. The necessary hardware in terms of Ethernet and InfiniBand switches had already been installed together with the DEEP Cluster, thus no additional procurement of hardware was needed here. In addition to the network layout the distribution and installation of firmware and software components had to be scheduled with respect to potential dependencies.

3 Installation of scale up systems

This installation phase covers the scale-up of two separate subsystems: the DEEP Booster and the DEEP ASIC Evaluator. Since the delivery date for both of them is the 30 June, the installation activities had to be performed in parallel, which introduced some challenges due to the fact that some of the project partners had to take actions on both systems.

3.1 Installation of DEEP Booster hardware

3.1.1 Installation of the Booster rack

In the first installation phase the DEEP Booster rack and two half chassis had already been installed in Jülich. But it turned out that the existing Booster rack could not be re-used for the scale-up of the DEEP Booster, because reducing the number of chassis from 16 to 12 and replacing the original planned Booster Interface Cards with a Juno-BIC version led to different chassis dimensions and position so the connections points, e.g. for the internal water distribution, did not fit anymore and a new rack had to be designed and delivered to Jülich. This enabled the chance to also apply some improvements regarding the rack design. Instead of enlarging a standard 19” rack as done with the first version, a completely new custom build 23” rack is being used. It provides better rigidity and uses additional support bars mounted to the chassis ground plate to prevent the chassis from bending. Improved connections and layout are used for the rack internal liquid distribution. The cold plates of the Juno-BICs are now connected to the same horizontal distribution bar as the BNCs they are hosting.

![Figure 5: Anderson Power connectors used for backplane power supply](image)

![Figure 6: Improved isolation for backplane connections](image)

Regarding the power supply the back plane connectors where revised to have a larger distance to the quick disconnectors of the liquid distribution and to provide a better isolation. Connectors were added between the backplanes and the fuses located on top of the rack to simplify mechanical decoupling from the power supply (see Figure 5 and Figure 6). The new rack was assembled and tested at Eurotech before shipment and the internal liquid distribution bars as well as the backplanes where already mounted on delivery to Jülich.
3.1.2 **Integration of EXTOLL cabling and chassis**

The first step after integrating the rack into the Jülich infrastructure was to perform the EXTOLL cabling. The EXTOLL 3D topology for the DEEP Booster is illustrated in Figure 7.

To achieve the target topology for the scale-up system a large amount of Molex cables had to be positioned building the connections of the later Booster Nodes in X, Y and Z direction. With respect to the large number of cables and the limited space inside the rack this procedure had to follow a certain procedure. Figure 8 shows the position and the route of the cable routes for the X, Y and Z directions. Labels were added for each connection to not mix up the different routes between the backplanes.
Since the general layout of the chassis did not change, the integration of the BNCs could be performed in the same way as with the first rack version which was already described in D6.4. In contrast to the first installation phase no more Pseudo-BIC servers were needed for control of the BNCs. Instead the liquid cooled Juno-BIC cards were available and could be integrated into the chassis as well. The fully assembled DEEP Booster now hosts 12 chassis including 24 Juno-BICs, 192 BNCs and 384 KNCs (Booster Nodes). Six of the chassis are located on the front side of the rack and 6 on the back side as shown in Figure 9. Once mounted to the chassis the Ethernet and InfiniBand cabling could be added for the BICs. The cabling for the so-called “spine switches” for the InfiniBand connection between the BICs and the Cluster part had to be arranged to achieve an overall balanced IB network. The BNCs use an Ethernet connections over backplane provide by the BICs and do not provide an InfiniBand interface.
3.2 Installation of DEEP ASIC Evaluator hardware

The immersion cooling technology used in the ASIC Evaluator poses some special environmental requirements. The JUELICH infrastructure experts have defined special measures to guarantee that use of the Novec 3M cooling liquid inside the DEEP ASIC Evaluator is safe:

- A drip pan had to be installed underneath the cubes to prevent the liquid from getting into the ground water in the unlikely case of a leakage
- Operation instructions and safety guidelines had to be provided to people working with the system taking into account that that installation room is now comparable with a laboratory environment, e.g. eating and drinking is not allowed when working on the system.

As of writing this document the DEEP ASIC Evaluator was not yet delivered to Jülich. Nevertheless, with the rack and the Pseudo-BIC servers already installed and integrated into the JUELICH infrastructure the installation process for the Booster Nodes (immersion cooled KNCs) should be straight forward. On arrival of the hardware the following steps have to be performed:

- Mechanically integrate the cubes into the rack (the position within the rack is highlighted in Figure 2)
- Connect the cubes to the Booster cooling loop and fill the basins with NOVEC fluid
- Do the power and network connection
- Activate the cooling and power on all components
- Integrate the components into monitoring and check temperatures

The software bring-up is expected to be very similar to the DEEP Booster which is described in the next section. Firmware upgrades have to be applied for certain components to enable the latest features being implemented by WP3 and WP7. Software bring-up also includes the installation of necessary drivers and software tools on the Juno-BICs/Pseudo-BIC servers to allow for remote boot of the KNCs and to access them via EXTOLL from the DEEP management nodes.

4 Software installation and system configuration

4.1 IP configuration and network integration

The DEEP Cluster already uses several networks for different aspects of cluster operation where the components of the DEEP Booster and the DEEP ASIC Evaluator had to be integrated into:

- 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 and the cluster hardware BMC management and can be used for power cycling nodes and SOL console access
- An IPoIB network (10.2.16.0/22) which handles the data exchange between the tasks of an MPI job as well as providing the access towards the BeeGFS file servers and the NFS traffic to forward the GPFS shares mounted on the frontends (master servers)
The Cluster Booster communication is implemented through the Booster Interface Cards (BICs). A limited number of 8 BNCs is connected to a single BIC through an EXTOLL network using the chassis backplane. The BIC serves as a bridge between the Cluster and the Booster network. To this end, both the management network and the InfiniBand network need to be routed to the BNCs. The actual Booster Nodes (KNCs) need to see the IPoIB network as well to mount the shares containing the user data. Figure 10 illustrates the Booster Node integration into the existing networks using the Booster Interface Cards. Each BIC includes an Ethernet switch to connect its 8 BNCs over the backplane as it is done for the EXTOLL network. Hence no additional cabling is needed for the BNCs.

![Figure 10: BNC and EXTOLL integration into the existing networks](image)

For the ASIC Evaluator the BNs (KNCs) are directly attached to an EXTOLL NIC using a customized backplane. No BNCs are used here. In addition the functionality of the BICs is implemented by using 4 Pseudo-BIC servers which take care of routing the Eth and IB networks to the EXTOLL connection for the KNCs. The remaining network configuration is the same as on the DEEP Booster.

Altogether the following components of the scale-up subsystems had to be integrated into the different IP subnets:

**DEEP Booster:**
- 24 x Juno-BICs:
  - 3 Eth (Server Chipset Juno, BMC of Juno, BIC (SAM7))
  - 1 x IPoverIB
  - 1 x IPoverEXTOLL
- 192 x BNCs:
  - 1 x Eth (for the BMC; over backplane)
- 384 x BNs (KNCs):
  - 1 x IPoverEXTOLL (over backplane)
Installation of scale-up Booster and ASIC Evaluator

DEEP ASIC Evaluator:

- 6 x Pseudo-BICs (the ones used before on the first two half Booster chassis had to be reconfigured when moving from the Booster to the ASIC Evaluator)
  - 2 Eth IPs (Server chipset, BMC)
  - 1 x IB connection
- 64 x BN (KNCs):
  - 1 x IPoverEXTOLL

For the integration of the BNs and BICs another subnet (10.2.20.0/22) was created in addition to the existing subnets already being used for the Cluster part. It covers the IPoverEXTOLL communication. All of the above mentioned components receive their IP address via DHCP either from the management node “deepm” or the Juno-BICs which provide the IP addresses for the BNs (KNCs). The subnets being used for the DEEP System and the DEEP ASIC Evaluator are listed in Table 1 and can be found in Figure 10.

<table>
<thead>
<tr>
<th>Subnet</th>
<th>Used for</th>
<th>Hosts included</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.2.8.0/22</td>
<td>ssh, batch system, client server communication</td>
<td>CNs, Cluster Root Cards (RCs), BICs, mngt servers</td>
</tr>
<tr>
<td>10.2.12.0/22</td>
<td>IPMI and administrative network traffic</td>
<td>SAM7 inside Juno-BICs and BMCs of: CNs, RCs, BNs, BICs, mngt servers</td>
</tr>
<tr>
<td>10.2.16.0/22</td>
<td>IPoverIB</td>
<td>CNs, BICs, Pseudo-BICs, mngt servers</td>
</tr>
<tr>
<td>20.2.20.0/22</td>
<td>IPoverEXTOLL</td>
<td>BICs, Pseudo-BICs, BNs</td>
</tr>
</tbody>
</table>

Table 1: List of IP subnets for the DEEP System and the ASIC Evaluator

In a first step the file systems on the Booster Nodes (KNC cards) are mounted via NFS which is supposed to provide sufficient performance. The home file systems on the KNCs will be the same as on the Cluster Nodes and are provided through a global parallel file system (GPFS). The required LDAP functionality is available within the MPSS version 3.4.3 that is running on the Booster Nodes (KNCs). Later tests with a BeeGFS client running on the MIC OS are planned.

4.2 Firmware on the Booster Node Cards (BNCs)

Basic firmware versions for the BNC components were included on delivery. This counts for the KNC cards, the FPGAs and the BMCs. Since additional firmware updates (e.g. for the BMCs) were necessary after shipping the BNCs to Jülich, those updates had to be applied at installation site. An overview about the firmware versions for the BNC components can be found in Table 2. The BMC firmware update could be installed running a flash utility on one of the management nodes and using the backplane Ethernet connection of the BNCs which is connected to its BMC. The same approach was used for the FPGA firmware. A software tool called “bncctl” was provided by Eurotech which can be used to flash the FPGAs with a certain firmware image and for powering KNCs within a BNC. The KNC firmware provided with the delivery of the BNCs is not expected to require any update during administration of the BNCs. However, the KNC’s boot image for the MIC-OS had to be modified to change the root file system used by the MIC-OS. This is needed for configuration of the BNs, e.g. to allow for user management and to enable NFS mounts on boot of the KNC cards. The boot
image is loaded onto the cards by the MPSS software stack running on the BICs, whose setup is described in the next section.

<table>
<thead>
<tr>
<th>Component</th>
<th>Firmware version on delivery</th>
<th>Latest firmware version installed on site (as of writing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVR Microcontroller</td>
<td>V000106.elf</td>
<td>V000106.elf</td>
</tr>
<tr>
<td>CPLD</td>
<td>Diva_cpld_570v2.2.0.pof</td>
<td>Diva_cpld_570v2.2.0.pof</td>
</tr>
<tr>
<td>BMC</td>
<td>1.09</td>
<td>1.14</td>
</tr>
<tr>
<td>FPGA</td>
<td>717</td>
<td>743</td>
</tr>
</tbody>
</table>

Table 2: Firmware versions for the BNC components

4.3 Firmware and Software on the Juno-BICs

Two boards are included in each Juno-BIC: a Juno board equipped with an Intel® Xeon™ processor and the BIC board based on an ARM7 micro controller. For the latter one an update tool was provided by Eurotech to flash the ARM7 with a new firmware. It has to be executed on the Juno-OS and a USB connection between the Juno and the BIC boards is required. This can be implemented by using a standard USB cable plugged to the first USB port of the BIC board and the USB port of the Juno board. For updating the BMC on the Juno board the front panel Ethernet connection can be used with a flash utility running on one of the management servers. The update process is the same as for the BMCS on the Booster Node Cards. Regarding the FPGA update of the Juno-BICs a special procedure had to be followed using a USB blaster and a dedicated JTAG adapter being connected to the front panel connector of the BIC board. An Altera tool providing a graphical user interface could then be used to load the new bit file onto the FPGAs. Table 3 gives an overview over the firmware versions for the Juno-BIC components.

<table>
<thead>
<tr>
<th>Component</th>
<th>Firmware version on delivery</th>
<th>Latest firmware version installed on site (as of writing)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARM7 (on BIC board)</td>
<td>2.0.3</td>
<td>2.0.3</td>
</tr>
<tr>
<td>FPGA (on BIC board)</td>
<td>1.0.0.1</td>
<td>1.0.0.5</td>
</tr>
<tr>
<td>BMC (on Juno board)</td>
<td>1.10</td>
<td>2.00</td>
</tr>
<tr>
<td>CPLD (on Juno board)</td>
<td>A34p001b.pf</td>
<td>A34p001b.pf</td>
</tr>
<tr>
<td>BIOS RAM 16GB</td>
<td>BrickHSW.16.10.04.03</td>
<td>BrickHSW.16.10.04.03</td>
</tr>
<tr>
<td>InfiniBand</td>
<td>2.11.5100</td>
<td>2.32.5100</td>
</tr>
</tbody>
</table>

Table 3: Firmware versions for the Juno-BIC components

Figures illustrating the tools and the cabling needed to perform the required firmware update can be found in the annex to this document.

The Juno-BICs have to manage the network connectivity between the Cluster and the Booster Nodes (using EXTOLL on the one and InfiniBand/Ethernet on the other side) and to allow for managing the KNCs within the Booster Node Cards. They have to take over the routing functionality between the different networks described in section 4.1. Therefore the required
drivers and software tools have to be available on the operating system of these servers. For compatibility with the management servers and the Cluster Nodes the CentOS operating system is used on the Juno-BICs. The following software had to be installed and configured on top of CentOS:

- MPSS stack for managing the KNC cards in the BNCs
- EXTOLL drivers
- Cluster-Booster protocol (CBP) needed for communication between the Cluster and the Booster Nodes
- ParaStation MPI based on the CBP to provide a Global MPI being used on the Cluster as well as on the Booster Nodes
- IB OFED stack including the latest InfiniBand HCA firmware

Additional modifications had to be applied to the MPSS installation for integration into the installation environment, e.g. to use NFS shares and LDAP user management. An imaging system was set up to simplify distribution and maintenance of the CentOS for all 24 Juno-BICs.

### 4.4 System Monitoring

For system monitoring and investigation on energy efficiency focused in WP7 various sensors at different system levels were integrated with the installation of the DEEP System. A detailed description of the available sensors and the RAS-Framework can be found in D7.4. Aside from the energy efficiency aspect these sensors can also be used for system monitoring. The cooling infrastructure and the temperature of the various components are of special interest here. Therefore rack level sensors, a subset of infrastructure sensors reflecting the overall status of the system components (e.g. pump state) as well as the board temperatures of the compute elements were integrated into a central monitoring tool that can be accessed using a web user interface. It is able to provide an informative overview.

Corresponding clients were installed on the hardware to be monitored, e.g. the Cluster Root Cards (RCs) and the Juno-BICs. Both of them can also collect and provide temperature data for the Compute Nodes that are assigned to them. Components not capable to run the monitoring client software, but are accessible via network are frequently checked via “ping” command. This accounts for the BMCs and the ARM7 within the Juno-BIC cards. Infrastructure and rack level sensors can be accessed via SNMP which is supported by the monitoring software natively. Graphical presentation of collected data and generation of timelines are provided as well. A screenshot showing the monitoring web interface can be found in the D6.4 document. The following devices were included in the meantime and have been assigned to corresponding host groups (device types) within the monitoring software:

- Infrastructure
  - WebRelays
  - Power Modules
  - ClustSafe
  - Cooling loop components (pumps etc.)
  - Room temperature and humidity + leakage sensors
- DEEP Cluster
  - Root Cards (RCs)
o Cluster Nodes (CNs)

- DEEP Booster
  o Juno-BICs
  o Booster Node Cards (BNCs)
  o Booster Nodes (KNCs)

- DEEP ASIC Evaluator
  o Pseudo-BIC servers
  o Booster Nodes (BNs)

Though the monitoring system can inform the administrators about potential alarms immediate reaction on critical system states is mandatory and cannot be guaranteed by sending notifications. Hence safety system software was installed on the Cluster RCs and the Booster Juno-BICs. It frequently checks the temperatures of the compute nodes and is able to automatically power off certain hardware components in case of over temperature. Further implementations for automatic reactions, e.g. to stop the pump in case a leakage sensor is sending an alarm, will be considered.

5 System tests

Although most of the hardware components had been tested in the factory during production the bring-up phase of both subsystems required extensive testing. For the DEEP Booster first tests already started at Eurotech in Amaro. After shipment to Juelich and performing the integration into the infrastructure these tests were repeated. For each half chassis the Juno-BIC connected to local switch to check for the availability of

- The Juno board
- The ARM7 on the BIC board
- All 8 BNCs included in the half chassis

When finishing the final network cabling and configuration the first action was to test the (remote) access to all of the components via ssh. After fixing some issues with the Ethernet connection over backplane and the InfiniBand HCA on the Juno-BICs network access was possible to all Juno-BICs (chipset of Juno board, BMC of Juno board, ARM7 of BIC-board) and BNCs (BMC). In addition all BNCs and KNCs could be powered on/off remotely using the “bic_ctl” and “bncctl” tools provided by Eurotech. For the InfiniBand connections of the Juno-BICs a PingPong test was set up that confirmed the connections being in a good shape (0% packet loss). A special software tool was provided by UniHD to check the EXTOLL links for every half chassis. An exemplary output of this link test tool can be found in Figure 11.
Figure 11: Exemplary output of the EXTOLL link test tool

With the EXTOLL network configuration allowing to perform a remote boot of the KNCs the Cluster-Booster protocol (CBP) could be installed and tested. As of writing this document the tests were still ongoing. Further benchmarks and results will be performed and presented in the scope of WP4. The LDAP support was tested successfully on one of the BICs and the configuration could be distributed to the remaining BICs using the imaging tool. Additional tests for user login and data access to mounted and shared home directories will follow.

A stress test called “power virus” was successfully performed on 8 BNCs to check the cooling configuration. No problems have been detected so far. Further investigation will be done running the stress test on the full machine as well as under-temperature checks to avoid the risk for potential condensation in case the inlet temperature drops too much.
6 Annex

The following additional figures illustrate the cabling and software tools that have been used to perform the firmware upgrades of the Juno-BIC hardware (FPGA and ARM7 on the BIC board).

Figure 12: USB cabling between Juno and BIC boards to update ARM7 firmware

Figure 13: Graphical User Interface of Altera tool for updating FPGA firmware image on BICs
Figure 14: JTag cabling using USB blaster to perform FPGA update of the BIC

References and Applicable Documents

[2] DEEP deliverable D3.6 – DEEP Booster upscale
[3] DEEP deliverable D3.7 – ASIC Evaluator
List of Acronyms and Abbreviations

A
ASIC: Application Specific Integrated Circuit: Integrated circuit customized for a particular use
ASIC evaluator: A platform consisting of 64 Booster nodes (KNCs) and EXTOLL Tourmalet cards, developed and used only in the DEEP project
Aurora: The name of Eurotech’s cluster systems

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
BIC: Booster Interface Card: Interface card to connect the Booster to the Cluster InfiniBand network
BMC: Baseboard Management Controller
BN: Booster Node (functional entity)
BNC: Booster Node Card: A physical instantiation of the BN
Booster System: Hardware subsystem of DEEP comprising of BNC, BIC and Intra-Booster network

C
CN: Cluster Node (functional entity)
CPLD: Complex Programmable Logic Device
CPU: Central Processing Unit

D
DEEP: Dynamical Exascale Entry Platform: EU-FP7 Exascale Project led by Forschungszentrum Juelich
DEEP Architecture: Functional architecture of DEEP (e.g. concept of an integrated Cluster Booster Architecture)
DEEP Booster: Booster part of the DEEP System
DEEP System: The production machine based on the DEEP Architecture developed and installed by the DEEP project
DHCP: Dynamic Host Configuration Protocol

E
EC: European Commission
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**

**FPGA:** Field-Programmable Gate Array: Integrated circuit to be configured by the customer or designer after manufacturing

**G**

**GPFS:** General Parallel File System by IBM

**H**

**HCA:** Host Channel Adapter

**HW:** Hardware

**I**

**IB:** InfiniBand

**Intel:** Intel GmbH Braunschweig, Germany

**Intel Xeon® Phi™:** Official product name of the Intel Many Core (MIC) architecture processors. The first available Intel Xeon® Phi™ product is code-named Knights Corner (KNC).

**I/O:** Input/Output

**IP:** Intellectual Property or Internet Protocol (depending on the context)

**IPMI:** Intelligent Platform Management Interface

**IPoEXTOLL:** Use the Internet Protocol on an EXOLL based network

**IPoIB:** Use the Internet Protocol on an InfiniBand based network

**J**

**JSC:** Juelich Supercomputing Centre

**JTAG:** Joint Test Action Group. Standard test access port and boundary-scan architecture

**JUELICH:** Forschungszentrum Jülich GmbH, Jülich, Germany

**K**

**KNC:** Knights Corner: Code name of a processor based on the MIC architecture. The commercial name of this product is Intel Xeon® Phi™.

**L**


**M**

**MIC-OS:** Operating System of the MIC architecture

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

**MPSS:** Intel many-core platform software stack. Software bundle to operate Xeon® Phi™ devices
**N**  
NIC: Network Interface Card: Hardware component that connects a computer to a computer network

**O**  
OS: Operating System

**P**  
ParaStationMPI: Software for cluster management and control developed by ParTec  
PCIe: PCI Express: Standard for peripheral interconnect, developed to replace the old standards PCI, improving their performance

**Q**

**R**  
RAS: Reliability, Availability and Serviceability  
RC: Root Card (functional entity within the Cluster part of the DEEP System)

**S**

**T**

**U**  
UniHD: University of Heidelberg, Germany

**V**

**W**  
WP: Work Package

**X**

**Y**

**Z**