SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

FP7-ICT-2011-7

DEEP

Dynamical Exascale Entry Platform

Grant Agreement Number: 287530

D3.6
DEEP Booster Upscale

Approved

Version: 2.0
Author(s): A.Somma (Eurotech)
Date: 02.10.2015
# Project and Deliverable Information Sheet

<table>
<thead>
<tr>
<th>DEEP Project</th>
<th>Project Ref. №</th>
<th>287530</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project Title:</strong></td>
<td>Dynamical Exascale Entry Platform</td>
<td></td>
</tr>
<tr>
<td><strong>Project Web Site:</strong></td>
<td><a href="http://www.deep-project.eu">http://www.deep-project.eu</a></td>
<td></td>
</tr>
<tr>
<td><strong>Deliverable ID:</strong></td>
<td>D3.6</td>
<td></td>
</tr>
<tr>
<td><strong>Deliverable Nature:</strong></td>
<td>Prototype</td>
<td></td>
</tr>
<tr>
<td><strong>Deliverable Level:</strong></td>
<td>PU*</td>
<td></td>
</tr>
<tr>
<td><strong>Contractual Date of Delivery:</strong></td>
<td>30/06/2015</td>
<td></td>
</tr>
<tr>
<td><strong>Actual Date of Delivery:</strong></td>
<td>30/06/2015</td>
<td></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:</th>
<th>DEEP Booster Upscale</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ID:</strong></td>
<td>D3.6</td>
<td></td>
</tr>
<tr>
<td><strong>Version:</strong></td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td><strong>Status:</strong></td>
<td>Approved</td>
<td></td>
</tr>
<tr>
<td><strong>Available at:</strong></td>
<td><a href="http://www.deep-project.eu">http://www.deep-project.eu</a></td>
<td></td>
</tr>
<tr>
<td><strong>Software Tool:</strong></td>
<td>Microsoft Word</td>
<td></td>
</tr>
<tr>
<td><strong>File(s):</strong></td>
<td>DEEP_D3.6_Booster_Upscale_v2.0-ECapproved</td>
<td></td>
</tr>
</tbody>
</table>

**Authorship**

- **Written by:** A. Somma (Eurotech)
- **Contributors:** P. Arts (Eurotech)
- **Reviewed by:** F. Geier (ParTec), N. Eicker (JUELICH), E. Suarez (JUELICH)
- **Approved by:** PMT/BoP
## 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>10/October/2015</td>
<td>Approved</td>
<td>EC approved</td>
</tr>
</tbody>
</table>
Document Keywords

| Keywords: | DEEP, Exascale, Booster |

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 .............................................................. 2
Document Control Sheet ............................................................................................... 2
Document Status Sheet ................................................................................................. 3
Document Keywords ...................................................................................................... 4
Table of contents ........................................................................................................... 5
List of Figures ................................................................................................................ 6
Executive Summary ....................................................................................................... 7
1 Introduction .................................................................................................................. 8
2 Description of the rack ................................................................................................. 10
3 Boards ......................................................................................................................... 13
  3.1 BIC (Booster Interface Card) ................................................................................... 14
    3.1.1 BIC Architecture ............................................................................................... 14
    3.1.2 BIC Firmware Development ............................................................................ 15
    3.1.3 BIC Assembly .................................................................................................... 16
    3.1.4 JUNO .................................................................................................................. 16
  3.2 BNC (Booster Node Card) ....................................................................................... 17
  3.3 Backplanes ............................................................................................................. 17
4 Other elements in the rack ......................................................................................... 17
5 Third party components .............................................................................................. 20
6 References and Applicable Documents ...................................................................... 21
7 List of Acronyms and Abbreviations ......................................................................... 22
8 Annex 1: JUNO board ................................................................................................. 25
List of Figures

Figure 1: Diagram of the DEEP architecture ........................................................................................................ 9
Figure 2: The DEEP Booster ................................................................................................................................. 11
Figure 3: The DEEP Booster Rack (Mechanical Drawing) .................................................................................... 12
Figure 4: BIC (Booster Interface Card) ................................................................................................................ 13
Figure 5: BNC (Booster Node Card) ..................................................................................................................... 13
Figure 6: Left half-Backplane .............................................................................................................................. 14
Figure 7: Right- and left-half backplane inserted in the rack ............................................................................... 14
Figure 8: BIC block diagram ............................................................................................................................... 15
Figure 9: JUNO card .............................................................................................................................................. 16
Figure 10: Connection of the JUNO card to the BIC main board ....................................................................... 17
Figure 11: 3-way valves ....................................................................................................................................... 18
Figure 12: De-airing valves ................................................................................................................................. 18
Figure 13: Inlet/Outlet tubes ............................................................................................................................... 19
Figure 14: Anderson connectors ......................................................................................................................... 19
Figure 15: 3D-Torus cables .................................................................................................................................. 20
Figure 16: CPU card schematic .......................................................................................................................... 25
Figure 17: Assembled BIC with JUNO mounted on the top of the cold plate ..................................................... 26
Executive Summary

One important outcome of the DEEP project is the “DEEP System”, a specifically developed prototype, which is used in the project for the demonstration of the Cluster-Booster architecture, in which it is based. The DEEP System is constituted by two main parts:

- The **Cluster**: an out-of-the shelf Intel Xeon Cluster with InfiniBand interconnect.
- The **Booster**: a cluster of Intel Xeon Phi (“Knights Corner” product line) with EXTOL interconnect.

Both parts of the system are based on Eurotech’s water cooled Aurora technology.

This deliverable describes the DEEP Booster, including its main components and the way they are integrated to provide the functionality needed in DEEP. The hardware for the DEEP Booster has been installed at JSC and a large part of the software bring-up has been already completed.
1 Introduction

While the Cluster subsystem of the DEEP System is based on production-level commodity hardware, the Booster is a custom design that was developed and implemented in Work Package 3. The overall DEEP Architecture consisting of a Cluster, a Booster and the bridging Booster Interface is sketched in Figure 1.

The final DEEP System consists of one cabinet holding an Intel® Xeon®-based Cluster (128 nodes) and one cabinet for the Intel® Xeon Phi™-based Booster (with 384 co-processors of the “Knights Corner” product line). An additional rack that is placed between the two cabinets holds power supplies and interconnect equipment.

Successful integration of the prototype hardware with the environmental and operational infrastructure at the installation site (Jülich) is critical to be able to conduct a meaningful evaluation of the DEEP Architecture as a possible blueprint for future generations of energy-efficient HPC systems.

Among others, the DEEP System will be used for further experiments on hot-water cooling and energy-efficient computing above and beyond the results that have been obtained with existing hot-water cooling experimentation platforms at BADW-LRZ.

The Booster is consisting of 192 Booster Node Cards (BNC) integrated into a total of 12 Booster chassis. The BNCs each hold two Intel Xeon Phi cards and two EXTOLL network interface controllers (NIC), the latter implemented by means of Altera Stratix V FPGAs, forming logically two Booster Nodes (BNs). The BNCs are interconnected via a custom designed backplane, which also carries management networks and power connections. Components on the BNCs are cooled by custom-made, high-precision coldplates, with liquid coolant connections passing through the backplane. Thermal validation shows that a fully loaded DEEP Booster can be cooled with treated water of up to 50 °C, enabling the use of free cooling for ≥ 98% of the days at the installation site at Jülich, Germany.

The BICs (Booster Interface Card) use a JUNO board¹ with a standard x86 CPU, up to two InfiniBand adapter plug-in cards and a single EXTOLL NIC integrated in an onboard FPGA. An onboard PCI Express switch connects these components for peer-to-peer EXTOLL↔InfiniBand communication. The BIC actually used in the DEEP Booster utilises a single InfiniBand adapter.

¹For a more comprehensive description of the JUNO board, see “Annex 1: JUNO board”.

8
Figure 1: Diagram of the DEEP architecture

As part of WP6 the DEEP infrastructure has been designed and implemented at Jülich as described in D6.4 (See [2]).

The Cluster subsystem has been integrated with the infrastructure and put into production use as reported in Deliverable D6.3. The lessons learned from this integration and the initial operation of the DEEP Cluster have been taken into account in the design of the Booster subsystem rack. With the results of Tasks 3.1 and 3.2, the rack dimensions for a Booster holding up to 192 BNCs, 24 BICs, and 24 half backplanes in 12 chassis were defined.

On the 18th of May Eurotech started the installation of the DEEP Booster rack in Jülich. This document describes the elements that have been delivered and that are now mounted in the DEEP Booster rack.
2 Description of the rack

The DEEP Booster rack holds a total of 12 chassis, 6 on the front and 6 on the back.

1 chassis is composed of 2 frames, so-called half chassis, (one left, one right) and each frame includes the following elements:

- 1 BIC (Booster Interface card) accompanied by a JUNO as described below. For the changes in the BIC and a more detailed description of the actual JUNO card refer to Annex 1 and Annex 2 respectively
- 8 BNCs (Booster Node Card)
- 1 backplane (which connects the BNCs to each other and to the BIC, and distributes power).

Figure 2 illustrates how the chassis are disposed in the rack (1 side).
Figure 2: The DEEP Booster
Figure 3 is a screenshot of the Mechanical Drawing of the rack.

After the delivery of the initial rack, the CPU of the BIC had to be exchanged to allow booting the KNCs within the Booster. The CPU on the BIC was changed from a COMe card to a server in PCI express form factor (see section 3.1.1). This required the delivery of a new rack with new dimensions. On top of that the Project (WP6) had a set of suggestions for the improvement of the mechanical design of the rack. Eurotech has installed a new rack, in order to amend the issues arisen with the structure of the rack previously installed in Jülich and to add the server CPU to the BIC. Main modifications of the Rack design are the following:

- Adoption of a 42U rack (the previous one was 48U)
- Increase the height of the chassis from 6U to 6U + 1 inch
- Introduction of a water collection plate on the bottom of the rack used in case of leakage
- The valves on the vertical distribution pipe have now a flat head connection so that the O-Rings are no longer damaged (the previous connection shape presented this risk)
- Introduction of a central column to prevent the bending of the shelves
- Adoption of a new front panel for the BIC so that the JUNO card can fit in as well (increase of the “L” shape support to 1 inch height)
- Electronic insulation of the +/- 48V connection
- Adoption of connection blocks for connecting the +/- 48V.

3 Boards

The DEEP Booster rack installed in Jülich includes:
- 24 BICs (Figure 4): On each BIC there is 1 JUNO processor board mounted on an Egeo carrier board (See Annex 2 for more details).
- 192 BNCs (Figure 5): On each BNC there are 2 KNC boards (total: 384 KNCs)
- 24 Backplanes: 12 left half-backplane and 12 right half-backplanes (Figure 6)

Both the BICs and the BNCs have coldplates mounted to allow the circulation of the liquid used to cool the system.

The BIC, BNC and Backplane are described in Deliverable D3.5 (See [2]). The following paragraphs describe the modifications and patches that have been introduced after the release of the document.
The BIC as presented in delivery D3.5 turned out to be unable to handle all 16 Xeon Phi processors located in the adjoined chassis-frame at the same time. The DEEP partners found out during the project that the selected COMe card and its Intel Core i7 CPU was incapable to map all memory of the connected Xeon Phi’s into the local PCIe address space as required by the remote-boot procedure and the Cluster-Booster protocol developed by WP4.

After thorough investigation, it was found that there is no CPU that comes in a COMe form factor that can support this amount of memory. A small server board providing the needed functionality is the Eurotech JUNO board. The CPU board is connected by a riser cable and interface board to the COMe slot of the BIC main board.

The JUNO card has a standard PCI connection to the BIC main board, while the Ethernet connection and USB connection are realised via the front panel of JUNO.
3.1.2 BIC Firmware Development

During power up the (self-bootable) JUNO board boots autonomously. After boot the board management controller (BMC) on JUNO controls the reset of the PCI express devices. When the reset line of the PCI Express switch is released, also PCI endpoint connected to the switch (InfiniBand HCAs and FPGA) come out of reset at the same time.

Booting of the DEEP Booster interface node consists of 6 stages.

1. When 12 VDC power is applied to the BIC, the ARM7 µ-controller and the JUNO board are powered up independently.
2. After booting the OS, the JUNO card controls the reset line of the Ethernet switches, PCI switch, FPGA and InfiniBand cards.
3. By releasing the reset lines, the BIC is powered up.
4. After starting the OS on the ARM7 the BNC nodes can be switched on and the BNCs are powered up.
5. When the BNCs are powered up the JUNO boots the Xeon Phi cards over EXTOLL.
6. The only firmware programmed to control and manage the BIC is that of the ARM7 µ-controller.

The software running on the JUNO board for collecting high frequency sensor data and resiliency will be described in D7.5 “Final report on DEEP energy efficiency”. The Cluster-Booster protocol and the software allowing for the remote-boot of the Xeon Phi via the EXTOLL fabric are deliverables of WP4.

To protect the Booster system against overheating a safety system is developed that autonomously switches off the Booster when a too high temperature is detected. A daemon is
running on each JUNO to continuously check the temperature of the BNCs and BICs in the rack. When an over-temperature is detected the chassis is switched off by removing the 48 Vdc applied to the back plane.

### 3.1.3 BIC Assembly

During assembly the JUNO card, the ConnectX3 module and the hard disk are mounted on the BIC node card.

The JUNO card is a Eurotech standard Intel E3 mother board in a PCIe form factor. See figure below.

![Figure 9: JUNO card](image)

### 3.1.4 JUNO

The JUNO card is connected via a PCIe cable and a PCIe-to-COMe interface card to the COMe board-to-board connectors on the BIC main board. The PCIe-to-COMe interface card has been designed by Eurotech for this project. The PCIe riser cable is an off the shelf component: “Samtec PCIe Express® Jumpers”.

The BIC’s 2.5inch SSD is connected to the JUNO SATA slot.
Figure 10: Connection of the JUNO card to the BIC main board

The JUNO is cooled by mounting the card on the top side of the BIC coldplate. Its corresponding SSD is mounted on the top side of the cold plate as well.

Further details on the internal design of the JUNO board are given in the Annex of this document.

3.2 BNC (Booster Node Card)

The BNCs and the backplane are described in reference [2]. Analysis performed during the bring-up and test phase resulted in some patches and rework on the BNC after the release of D3.5. The main modifications to the design are related to improvements in the quality of the Ethernet connection between the BIC and BNC over the backplane. Filtering and DC decoupling was applied in the transmit and receive lines between the BIC and the BNC.

3.3 Backplanes

No modifications or Hardware patches have been introduced on the backplanes. The right half-backplane has been designed mirroring the design of the left half-backplane.

4 Other elements in the rack

Besides the elements described above, the rack includes:
1. **Thirteen (13) “3-way valves”**, situated on one side of the rack. The 3-way valves are used to switch on and off the cooling circuit. The front and back chassis on the same level share 1 inlet valve and 1 outlet valve. The 13th valve only supplies the inlet for the top front and back chassis. See Figure 11.

![Figure 11: 3-way valves](image)

2. **Two valves** used to remove the air/connection point for the de-airing system used in the datacenter (located on the top side of the rack). See Figure 12, red circles.

![Figure 12: De-airing valves](image)

3. **Two tubes**, 1 for water inlet and 1 for water outlet (located on the bottom side of the rack). See Figure 13, red circles.
4. **Forty eight (48) Anderson connectors** (Figure 14), situated on the left side of the rack (looking at the front side). The Anderson connectors are connected to the power supply to provide power to the system.

5. **Six hundred (600) 3D-Torus Cables** (Figure 15), used to connect all backplanes to each other via 3D Torus through X, Y and Z connections. The cables have different lengths, depending on the connection axis: 52 cm (X direction), 37 cm (Z) and 80, 52, 72 and 42 cm (Y). The number of cables delivered for each length is the following:
   - 37 cm $\rightarrow$ 96 pieces
   - 42 cm $\rightarrow$ 70 pieces
   - 52 cm $\rightarrow$ 166 pieces
   - 72 cm $\rightarrow$ 134 pieces
   - 80 cm $\rightarrow$ 134 pieces
5 Third party components

The DEEP Booster Rack includes several standard off the shelf components, not produced by Eurotech (“third party components”). You will find detailed specifications for each of them in the BSCW folder “third_party_components” provided with this document.

- 24 InfiniBand Host Channel Adapter Cards (IB HCA) (Intel AXX1FDRIBIOM FDR InfiniBand* ConnectX-3* I/O Module), one for each BIC.
- 12 Power Supply Unit shelves (PSU) (Eltek Valere: Smartpack Controller and 4 pieces of Flatpack 2HE) to provide 2x 100A@48V power to the two half chassis.
- 1 Remote I/O and Sensor Monitoring System (Poseidon2 4002) used for remote sensing of leakage and fire alarms.
- 2 Relay Board (Web Relay-10) used for power on/off of each of the PSU shelves.
- 6 Leakage Sensors (Produal VVN2 with LPH 10 monitor) located at the bottom of each chassis underneath the quick disconnect couplings and connected to the remote I/O and sensor monitoring system.
- 1 Fire Sensor (Optical smoke detector FDR26) located on the top of the Cluster – Booster racks and connected to the remote I/O and sensor monitoring system.
6 References and Applicable Documents

   https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d1458833/DEEP_D6.4_Installation_of_DEEP_System_v2.0_ECapproved.pdf

[2] DEEP_D3.5_Booster_v2.0_ECapproved.pdf
   https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d1458704/DEEP_D3.5_Booster_v2.0_ECapproved.pdf

[3] D3.1_Hardware_Concept_v2.0_ECapproved.pdf
   https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d892402/DEEP_D3.1_System_Hardware_Concept_v2.0_ECapproved.pdf

   https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d892430/DEEP_D3.2_Booster_Interconnect_Evaluation_Board_v2.0-ECapproved.pdf

   https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d1310757/DEEP_D3.3_Software_development_prototype—Proto-Booster_prototype_v2.0-ECapproved.pdf

   https://bscw.zam.kfa-juelich.de/bscw/bscw.cgi/d1458695/DEEP_D3.4_Booster_Rack_v2.0_ECapproved.pdf
7 List of Acronyms and Abbreviations

A

Aurora: The name of Eurotech’s cluster systems

B

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

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

COM Express: Computer-on-module (COM) form factor: highly integrated and compact PC that can be used like an integrated circuit component.

CPU: Central Processing Unit

D

DEEP: Dynamical Exascale Entry Platform: EU-FP7 Exascale Project led by Forschungszentrum Jülich

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

E

EC: European Commission

EGEO: Carrier board on the BIC

EU: European Union

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**

**HCA:** Host Channel Adapter

**HPC:** High Performance Computing

**HW:** Hardware

**I**

**IB:** InfiniBand

**InfiniBand:** Computer network communications link used in high-performance computing

**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).

**J**

**JÜLICH:** Forschungszentrum Jülich GmbH, Jülich, Germany

**JUNO:** Processor card on the BIC.

**K**

**KNC:** Knights Corner: Code name of a processor based on the MIC architecture

**L**

**M**

**N**

**NIC:** Network Interface Card: Hardware component that connects a computer to a computer network
PCB: Printed Circuit Board

PCI: Peripheral Component Interconnect. Standard for attaching components to a computer system. It covers mechanical, electrical and logical aspects.

PCI-Express: An implementation option of PCI using high speed serial links as physical interconnect layer.

PCIe: Same as PCI-Express

SMBus: System Management Bus

SW: Software

USB: Universal Serial Bus

WP: Work Package
Annex 1: JUNO board

The JUNO board is used in the Booster Interface to boot the 16 KNCs populating half a chassis, and to establish and maintain communication between them and the Cluster side of the DEEP System.

The CPU card main components are: an Intel Low Power E3-12xx v3 CPU, the BMC and the PCIe3x8.

The I/O (input/output) functions on the front panel of JUNO include 1 Gigabit Ethernet, 1 USB, 1 VGA and 1 Debugging port.

CPU card specifications

<table>
<thead>
<tr>
<th>CPU CARD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>E3-12xx v3</td>
</tr>
<tr>
<td>RAM</td>
<td>16 GB DDR3 (soldered memory)</td>
</tr>
<tr>
<td>Storage</td>
<td>1x 256 GB 2.5” SATA SSD</td>
</tr>
<tr>
<td>BMC</td>
<td>Emulex Pilot 3 SE-SM4310</td>
</tr>
</tbody>
</table>

The figure below shows how the components are placed on the board.

![CPU card schematic](image)

Figure 16: CPU card schematic
Figure 17: Assembled BIC with JUNO mounted on the top of the cold plate