SEVENTH FRAMEWORK PROGRAMME

Research Infrastructures

FP7-ICT-2011-7

DEEP

Dynamical Exascale Entry Platform

Grant Agreement Number: 287530

D3.8

DEEP Architecture Evaluation Document

Approved

Version:	2.0
Author(s):	H.-Ch. Hoppe (Intel), P. Arts (Eurotech), U. Bruening (UniHD), M.Ott (BADW-LRZ)
Contributors:	E.Suarez (JUELICH)
Date:	02.10.2015

Project and Deliverable Information Sheet

<table>
<tr><td rowspan="7">DEEP Project</td><td colspan="2">Project Ref. №: 287530</td></tr>
<tr><td colspan="2">Project Title: Dynamical Exascale Entry Platform</td></tr>
<tr><td colspan="2">Project Web Site: http://www.deep-project.eu</td></tr>
<tr><td colspan="2">Deliverable ID: D3.8</td></tr>
<tr><td colspan="2">Deliverable Nature: Report</td></tr>
<tr><td rowspan="2">Deliverable Level:
PU*</td><td>Contractual Date of Delivery:
31/08/2015</td></tr>
<tr><td>Actual Date of Delivery:
31/08/2015</td></tr>
<tr><td colspan="2">EC Project Officer: Luis Carlos Busquets Pérez</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 Architecture Evaluation Document</td></tr>
<tr><td colspan="2">ID: D3.8</td></tr>
<tr><td>Version: 2.0</td><td>Status: Approved</td></tr>
<tr><td colspan="2">Available at: http://www.deep-project.eu</td></tr>
<tr><td colspan="2">Software Tool: Microsoft Word</td></tr>
<tr><td colspan="2">File(s): DEEP_D3.8_Architecture_evaluation_document_v2.0-ECapproved</td></tr>
<tr><td rowspan="4">Authorship</td><td>Written by:</td><td>H.-Ch. Hoppe (Intel), P. Arts (Eurotech), U. Bruening (UniHD), M.Ott (BADW-LRZ)</td></tr>
<tr><td>Contributors:</td><td>E.Suarez (JUELICH)</td></tr>
<tr><td>Reviewed by:</td><td>I.Schmitz (ParTec)</td></tr>
<tr><td>Approved by:</td><td>PMT/BoP</td></tr>
</table>

Document Status Sheet

Version	Date	Status	Comments
1.0	31/August/2015	Final	EC submission
2.0	02/October/2015	Approved	EC approved

Document Keywords

Keywords:	DEEP, Exascale, Cluster-Booster, Evaluation

Table of contents

Executive Summary

In its 45 months of execution, the DEEP project has defined the DEEP System architecture, which for the first time implements the Cluster-Booster concept, followed by two actual implementations – the DEEP Booster with 384 Booster Nodes, and a smaller ASIC Evaluator with 32 Booster Nodes. The former is a fully integrated system designed and manufactured by Eurotech, using cold plate based liquid cooling and serving to validate the Cluster-Booster concept for the DEEP applications. The latter is an experimental design using immersive cooling and introducing the ASIC implementation of the EXTOLL interconnect.

This document revisits the major architecture and system design decisions and evaluates them with hindsight and in light of the results of the project. It identifies lessons learnt and discusses consequent improvements to the architecture and both systems that would be possible in the next couple of years. Even taking into account the significant delays encountered, DEEP has achieved significant technical results, and most decisions are vindicated at the time of writing.

The document concludes with a discussion of DEEP's lasting legacy.

1 Introduction

This Deliverable evaluates the key decisions taken in the definition of the DEEP architecture (as documented in Deliverable D3.1) and in the design of the DEEP system (as documented in Deliverables D3.4, D3.5, and D3.6 for the Eurotech Booster and D3.7 for the UniHD ASIC Evaluator). In addition, it also discusses the lessons learnt in implementing these systems, taking into account the experience gained in bringing up and operating both machines. Since the DEEP Cluster is based on trusted and proven technology, it is not covered here.

With a view towards the future, D3.8 examines ways to improve the DEEP architecture and design with technology that has been released or announced for release at the time of writing. While this does not cover the complete path to Exascale class systems, it does highlight critical issues and decisions to be taken for the next generation of Cluster-Booster systems, and contains recommendations that are actionable. The companion Deliverable D9.2 does present a wider and longer-term view of the architecture and technology evolution, and complements this with detailed projections of the DEEP system performance characteristics.

In the final section, the Deliverable evaluates the sustained legacy of the DEEP project, with a focus on the system architecture and the all-important SW interfaces implemented in this project. A discussion of DEEP's relation to the DEEP-ER project, which can be considered a younger sibling, concludes the document.

2 Assessment of the DEEP Prototype Architecture, Design and Implementation

In its 45 month term, the DEEP project has defined a system architecture that incorporates the key ideas of the Cluster-Booster concept [1], and that could actually be implemented in prototype form by the DEEP partners. This architecture was defined in Deliverable D3.1 at month 6, and has served as the basis for both the Eurotech systems (the 384-node Booster at Juelich and the 16-node Energy Efficiency Evaluator at LRZ) and the University of Heidelberg (UniHD) ASIC Evaluator. The DEEP architecture did use a common, off-the shelf (COTS) Cluster component; its innovative elements have been the DEEP Booster (including the EXTOLL interconnect) and the Booster Interface, which bridges between the Booster network and the InfiniBand-based Cluster fabric. Section 2.1 examines the key decisions related to the latter two sub-assemblies in light of the experience of designing, building and operating the two prototype implementations mentioned above. It does not consider the Cluster architecture.

Section 2.2 discusses the lessons learnt designing, manufacturing, bringing up and operating the Eurotech DEEP Booster. The original system design was documented in Deliverables D3.4 and D3.5, while D3.6 contains the description of the actual 384-node system installed in Juelich. The discussion here focuses on the latter system.

The ASIC implementation of the EXTOLL NICs performed outside of the DEEP project has experienced various setbacks and significant delays, with fully functional but performance-constrained samples becoming available only in the third year of DEEP. Therefore, the Eurotech DEEP implementation had by necessity to rely on an FPGA implementation of EXTOLL. To assess the impact of the performance gains finally delivered by the TOURMALET EXTOLL implementation in early 2015, an additional DEEP prototype to be designed and built by UniHD was ratified by the external project reviewers. This "ASIC Evaluator" is described in Deliverable D3.7, and section 2.3 examines its design and implementation.

2.1 DEEP Architecture

The key design decisions that define the DEEP architecture can be summarized as follows:

I. *Booster Node CPU:* It was decided to use the Intel Xeon Phi series of co-processors to serve as the Booster Node CPU; these co-processors provide up to 61 general-purpose, x86-architecture cores with a shared and fully coherent L2 cache, and are implemented as add-in cards with integrated GDDR5 main memory of up to 16 GByte. One Booster Node (BN) contains a single Xeon Phi co-processor.

II. *Booster Node controlling entity*: The Intel Xeon Phi co-processors require a PCI Express connection to an x86 system to boot and operate; an Intel-supplied SW component (MPSS) running on that "host" system controls the co-processor and also provides IP communication with it. **Decision**: the Booster Interface nodes (BICs) shall serve as host systems, run the MPSS stack and control the Xeon Phi co-processors of its attached Booster Nodes through the Booster interconnect. A ratio of 16 BNs to be attached to a single BIC was foreseen.

III. *Booster Node control network*: The PCI Express communication required by the above has to be routed from the BICs to the BNs and on to the Xeon Phi co-processor, using its PCI Express link. Options include using the Booster interconnect for this, or

relying on an Ethernet-based management and RAS network. Since the EXTOLL network is able to tunnel PCI Express communication, and the NIC can drive a PCI Express link to the Xeon Phi, the **decision** was taken to route the Xeon Phi boot/control information from the BIC via an EXTOLL NIC and the EXTOLL fabric to the BNs, where a NIC connects directly to the Xeon Phi co-processor.

IV. *Booster Interconnect*: The above was a secondary reason to select EXTOLL as the Booster interconnect. The primary reasons for this **decision** were its projected performance (up to 120 Gbit/s per link), its scalability (since it is direct switched) and the ability to build arbitrary size 3D tori with six of its seven links, leaving an additional link for connection to the Booster interface.

V. *Use of EXTOLL FPGA implementation*: To enable the project to proceed before the ASIC implementation of EXTOLL was ready (the development of which did happen outside of the project), the decision was reached to first use NICs implemented on a top-of-the-line FPGA, and switch to the ASIC NICs in time for building the final DEEP system (D3.5 at the time).

VI. *Direct liquid cooling*: To enable efficient cooling and tight packaging, the **decision** was reached to use direct liquid cooling for the DEEP Booster components (BNs and BICs). Eurotech's proven Aurora technology, which uses precision cold plates that match the height profile of the boards and components to be cooled, was selected. It also offered the option of hot-plugging liquid-cooled boards, which was seen as a bonus for the bring-up phase and for expedited maintenance of the completed system.

VII. *Backplane*: Each BN has six EXTOLL links that implement the 3D torus, and a subset of them uses a seventh link to connect to the associated BIC. Each link would use up to 12 lanes, each implemented as a high-speed differential pair (going up to 10 Gbit/s). To enable scalable and dense construction of the final DEEP system, and to reduce the reliability issues posed by using cables, the **decision** was reached to route all EXTOLL links via a backplane, and only use cables for connections between Backplanes. Initial evaluation did show that Backplanes could support up to 32 Booster Nodes, given that the chassis width could be increased to 23".

VIII. *Booster Node integration*: Initial design of the boards for the Booster Nodes did show that it would be possible to combine two BNs into a single board of the Eurotech Aurora blade form factor (referred to as a Booster Node Card or BNC). Since this would lead to a reduced number of components, increased packaging density and enable the re-use of mechanical Aurora components, it was **decided** to go that route.

IX. *RAS Plane*: The integration of a powerful RAS Plane has been an integral part of the DEEP project since its early planning stages. This has shaped the DEEP architecture by the decision to include additional sensors for power and temperature in the BN and BIC (over and above the sensors provided by Xeon Phi and by standard Xeon boards), to extend the management controllers on the BNs to provide high-frequency access to these sensors, and to introduce a separate RAS plane interconnect, which would be implemented using Ethernet technology. This RAS plane would also be of large value during system debugging and bring-up.

X. *Booster Interface*: The Cluster–Booster protocol (CBP) allows high speed communication between the Cluster and the Booster. The BIC architecture includes the bridging function between the InfiniBand FDR network on the Cluster and the EXTOLL network on the Booster. The bridging function is implemented by PCI switch on the BIC with an InfiniBand NIC, EXTOLL NIC and BIC CPU attached.

In addition to these architecture decisions, D3.1 also prescribed the use of "evaluators" that would test and validate the functionality and characteristics of all novel (and therefore risky) component technologies and of the integration of them into the DEEP system:

XI. *Role of Evaluators*: A sequence of prototype systems (christened "Evaluator systems") were defined for the basic EXTOLL NIC and link functionality/performance (EXTOLL Evaluator), the integration of an EXTOLL NIC with a Xeon Phi co-processor via PCI Express (Interconnect Evaluator), and for the BIC in its role as a network bridge and a control agent for the Xeon Phi co-processors in the BNs (BIC evaluator). The Interconnect Evaluator would be updated once a first implementation of the BNC became available and serve to drive functional and signal integrity tests of this prototype.

At the end of the project and in retrospect, these decisions are evaluated and judged as follows:

I. The selection of the (then novel) Intel Xeon Phi co-processor for the BN has worked out well – initial concerns about availability did not come true, and since the Xeon Phi operating system is based on Linux and as such available as open source, the required integration EXTOLL drivers could be performed within the project. Availability of detailed technical expertise on Xeon Phi through partner Intel has also been material to solve technical issues arising in various integration stages.

II. The decision to combine the network bridging and Xeon Phi control tasks in the BIC has led to a very high level of integration for the Eurotech DEEP system. It has, however, contributed to delays in the final BIC design and implementation, which could partly be offset by the introduction of the Pseudo-BIC concept (see section 2.2). From today's perspective, the decision still stands, as the problems encountered could be solved.

III. This decision was instrumental to limit the effort required in porting the highly complex booting and control communication layer to a different network formulation. It has also enabled DEEP to profit from the rapid development of the MPSS stack, in particular with regards to LDAP support. As discussed in section 2.2, the flow-down requirements of this approach towards the BIC CPU subsystem did cause delays. However, these could have been avoided by early testing of CPU subsystem alternatives. In toto, the benefits of the decision do outweigh its repercussions.

IV. The technical problems encountered by the development of the EXTOLL ASIC implementation have impacted the DEEP development schedule as well as the interconnect performance of the Eurotech DEEP Booster, which had to use a FPGA implementation of the EXTOLL NIC. The advantages of the EXTOLL fabric wrt. scalability have been demonstrated, on the other hand, and the performance effect that the TOURMALET implementation of EXTOLL brings to a DEEP system could be evaluated via the ASIC Evaluator platform, and projections could be made for the larger DEEP system.

    The alternative choice of an InfiniBand interconnect would have posed severe problems in itself, not the least of it being the need to integrate a very substantial number of switches into the Booster, thereby reducing efficiency and packaging density and creating technical risks. In addition, the aggregate bandwidth of the then available InfiniBand solutions would not have exceeded that of the EXTOLL FPGA implementation used.

One of the key problems encountered during DEEP is that the project was unable to provide material assistance to partner University of Heidelberg or exert significant influence when the ASIC development first encountered financial and technical problems. This was due to the fact that the ASIC development was carried out entirely outside of DEEP, and that the project did not have the budget to make a substantial difference anyway.

V. This decision can actually be considered to have saved the hardware part of the project: proceeding with the FPGA-based design did enable WP3 to "cut metal" early on and produce BNC (and BIC) prototypes for testing, identification of hardware bugs and design of corrections/improvements. Waiting for availability of the final ASIC specification would have made the design and manufacturing of a fully integrated DEEP Booster impossible, even within the two project extensions.

VI. This decision has led to a "bump in the road", with reliability issues in the DEEP Cluster installation requiring design changes and introduction of additional safeguards into the rack and RAS system of the DEEP Booster. However, the demonstrated efficiency results fully vindicate this decision, and the above mentioned improvements do constitute valuable progress in the state of the art for direct liquid cooling.

VII. The Backplane concept has indeed led to a compact system and rack design, with the number of cable connections within the 23" rack kept to a manageable level. It has also enabled the integration of the power lines towards the BNC, the Ethernet RAS plane and control lines from the BIC to the BNs. Liquid connectors pass through holes in the Backplane, enabling the stacking of electrical and liquid connections for operational safety. This decision is therefore fully vindicated.

VIII. The integration of two BNs in a single Aurora form factor blade did lead to a very compact board design for the BNC proper. One side-effect was that full functionality and performance was achieved with the third version of the BNC board, the earlier versions showing different signal integrity problems. While this has led to considerable delays in getting the final DEEP BNCs ready, the advantages of the selected, compact design do prevail; it has to be kept in mind that the lane speed design point of 10 Gbit/s did pose a steep learning curve for the available design and manufacturing chain.

IX. The final implementation allows for a comprehensive overview of the operating conditions of each individual hardware component. It gathers sensor data wherever available and stores it persistently in a scalable database. Where individual components did not expose sufficient sensor data, we have added additional external sensors to accommodate such shortcoming. All sensor data is sampled at high frequency and high accuracy and can be easily retrieved from the database via a uniform interface as raw data. Additionally, tools are available to visualise the data according to user specifications and requirements. Yet, the RAS plane not only provides easy access to an unprecedented wealth of sensor data, but also utilises this wealth to protect the system: all sensors are permanently monitored for out-of-specification conditions. In case of critical sensor readings, email notifications will be triggered and, if necessary, the corresponding hardware components shut down.

Therefore, the decisions taken at the architecture stage have worked out well.

X. The Cluster–Booster protocol was implemented using memory mapped IO, by mapping back all the memory of all the KNCs to the virtual memory of the BIC CPU. This requirement was missed during the feasibility of the hardware design and as a

consequence not taken into account when defining the CPU and the firmware. But due to the tests done on the Interconnect Evaluator in an early stage and the implementation work done by WP4, the hardware impact became clear and was shared with WP3. Future projects should use evaluators as early in the project as possible. The impact of complex functions like bridging between protocols can only be overseen when experimental implementations are tested with evaluators that re-assemble the final architecture as close as possible.

XI. The concept of requiring "evaluators" that demonstrate the feasibility of integrating novel components in innovative ways has fully paid off. It did provide early "ground truth" that positively impacted the system design and enabled system SW work to proceed without having functional BICs and BNCs. While the evaluators did lack the level of integration of the final solutions, they could be made available much earlier (akin to the benefits of "rapid prototyping" in SW development) and with much less financial outlay. Key design decisions (such as the integration of Xeon Phi with EXTOLL) could be validated early on, taking substantial risks off the table. Any problems were also noticed early on, with time to investigate the best mitigation. Future system projects should embrace this concept with a vengeance.

2.2 Eurotech DEEP Booster

Parts of the design of the Eurotech Booster system were already documented in Deliverable D3.1 at month 6; the full detailed design of the BNC and BIC boards, the Backplane is specified in D3.5 at month 36 and D3.6 at month 43. The design of the Booster rack, including improvements in cooling and power distribution, is discussed in D3.4 at month 29.

Important design decisions, steps in the development, and lessons learnt in manufacturing, integrating and bringing up of the Eurotech DEEP Booster do include:

I. *BNC integration of EXTOLL NIC (FPGA implementation) with its high-speed links*: the BNC board as designed by Eurotech integrates two EXTOLL NICs and routes six of their high-speed links to the Backplane connector. Each link consists of 8 lanes with a per-lane bandwidth of 10 Gbit/s. This design was chosen as a compromise: the FPGA NIC implementation can only drive up to 8 lanes (the ASIC would be able to drive 12 lanes), and the PCI Express generation 2 bandwidth to the Intel Xeon Phi co-processor is limited to 8 GByte/s. The two NICs are connected to each other with a LVDS link which carries a seventh EXTOLL link. The Altera Stratix V FPGA product was selected due to its high number of SERDES devices, high number of logic elements and PCI Express hardcoded-blocks, which would lead a best price/performance for the project. Eurotech's had experience with FPGAs from several vendors but had prior good experience with integration of this type of FPGA into their Aurora line.

   It took a total of three revisions to get the BNC board running reliably for the lane speeds produced by the EXTOLL FPGA implementation (6 Gbit/s per lane, 4 lanes per link). Issues to be resolved included the FPGA power supply, and the routing of high-speed lanes both with regards to each other and relative to ground and power planes in the PCB. Clearly, the design and manufacturing did require a steep learning curve due to the signal rates involved and the very high requirements of the FPGA NIC on quality of the power lines and its highly variable power rating. The problems could be analysed and corrected by a very close collaboration between Eurotech,

University of Heidelberg and Intel, and the final design does work reliably at the designed performance level.

University of Heidelberg had previously used the Xilinx Virtex line of FPGAs for their EXTOLL NIC implementations. For the BNC, they had to extend their implementation by a PCI Express root complex (to connect to the Xeon Phi), to switch from the older 64-bit FPGA RTL architecture to the new 128-bit architecture and port the implementation over to the Altera toolchain and the Altera-supplied IP blocks. The latter task took longer than expected, due to many small differences in the behaviour of SERDES and PCI Express IP blocks between Xilinx and Altera. The delay caused by this did impact the test & validation of the first BNC design.

In retrospect, the ambitious dense BNC board design does look like the right direction to take, since it enables tight packaging and reduces the number of separate components in a DEEP system. The decision to move to a different FPGA supplier for the EXTOLL NIC is more problematic, since it delayed the critical test & validation of the first BNC design. This effect was not foreseen, since the effort of porting between FPGA suppliers had been severely underestimated.

II. *BNC integration of the Intel Xeon Phi co-processor: Intel Xeon Phi is available form factors:* initially, it was planned to use the more compact "dense form factor" (DFF) version of Xeon Phi for the DEEP Booster. The standard PCI Express add-in card form factor versions were to be used for the initial tests only, and the BNC card was designed to accommodate both form factors using specific riser cards. After the delay in testing & validation of the first BNC designs (see above), it was decided to stay with the high-end 7120X PCI Express add-in form factor SKU of Xeon Phi to eliminate the effort and risk associated with producing and qualifying a new set of riser cards.

In retrospect, this design decision looks sound – the DFF versions would not have resulted in any energy or packaging improvements. In the DEEP Booster, a fraction of the riser cards for the 7120X Xeon Phi cards did show contact problems which could be fixed by manually intervention. It is not clear whether these could have been avoided with a DFF-based design.

III. *BNC bring-up and tests*: as discussed in I above, the Altera Stratix V version EXTOLL implementation was delayed. The initial tests and validation of the A0 BNC design started only after this implementation was available, and they did show problems with power quality for the FPGA. With hindsight, this problem could have been found and fixed earlier, using a dummy payload for the Altera FPGAs. The lesson learn here is quite clear: push for early testing of hardware features and do not wait for everything to be in place wrt. firmware.

The later issues found and corrected were related to signal quality over the high-speed links. These did require parts of the EXTOLL stack to run and could not have been significantly pulled in. Also, the cross-organisation team at that time had gathered experience, and the validation and debugging worked well.

IV. *BNC board management infrastructure and sensors:* the BNC board used a combination of a full-fledged board management controller (BMC) and a small µ-controller to provide full management functionality when powered up and avoid drawing significant power in the "suspended" state. The board also contained Ethernet devices and switches to implement the RAS plane and offer Ethernet from both a front-panel connector and through Backplane links.

One of the main benefits of developing own hardware within the project was to be able to accommodate specific requirements that were not available of the shelf. However, this also implies that pre-existing solutions will have to be adapted. This is obvious true for the Xeon Phi cards that would need to be remotely booted and controlled. For other components, however, this was realised during the project execution. For example, the development of the firmware for the Emulex Pilot-III BMC on the BNC required much more effort than originally anticipated. Although the BMC itself is technically mature and the AMI MegaRAC framework for developing the firmware well established, both were designed to be used for a different purpose: a service processor for a server system. In the DEEP project, however, the BMC is also used to provide network access to the FPGA flash memory, configure the Ethernet switch on the BNC, to start/stop/reset the Xeon Phi cards, to acquire, process, and push sensor data to the central database, etc. Implementing these features took longer than expected as developing software for embedded systems by itself is not as straight forward as for desktop machines and in this particular case also required modification of notoriously poorly documented low-level kernel drivers. Furthermore, there is a chicken-and-egg problem with the development of firmware for custom-made hardware and production tests of the hardware: without hardware, the firmware cannot be tested and vice-versa. This causes additional delays, as both cannot be developed independently from each other but require interlock.

Consequently, the writing and debugging the firmware for the two management devices did require a large effort, and partner BADW-LRZ took the initiative to carry out most of the firmware development in the last two years of DEEP. All designed management functions could be implemented this way, yet in toto, this did require significantly more effort than planned in the DoW.

During bring-up of the first DEEP chassis, it was noticed that the Ethernet links over the Backplane did not work properly. Eurotech and BADW-LRZ worked together to identify the root cause, and devised a HW fix that was retroactively applied to approx. 2 dozen BNCs and phased into BNC production. The Ethernet implementation was tested using the first debugging Backplane design (see VIII below), and the bug was triggered by the different electrical characteristics of the final Backplane and mechanical problems obstructing full insertion of the BIC into the Backplane. Such problems are very hard to avoid in ambitious HW projects, and the resolution was quick and effective. A lesson to be learnt here is to be conservative in planning the time required to bring up a complex HW artefact.

V. *BIC integration of EXTOLL NIC (FPGA implementation) and PCI Express switch*: the initial BIC design was completed at the same time as the original BNC design; after this, focus of the testing and validation switched to the BNC, with a Pseudo-BIC setup (see VII below) used as a Booster interface. Testing of the BIC design did start only after the issues with the BNC design were addressed and a working design finished. Unfortunately, some serious issues were immediately found with the FPGA and PCI Express switch integration, and significant time & effort had to be spent to analyse and correct them. Areas of improvement needed were the power quality for the FPGA and PCIe switch, the configuration of the PCIe switch (which did require a very steep learning curve), and the PCI Express lines to the FPGA and InfiniBand HCA.

While all issues were in the end identified, analysed and solved, it would have been much better if the testing of the BIC design had happened earlier and in parallel with BNC test & validation. None of the problems did require a working BNC or even

EXTOLL FPGA implementation. The lesson here is that one should strive to test all HW artefacts as early as possible.

VI. *BIC selection of CPU subsystem*: The initial design relied on using a COM Express CPU board as the CPU subsystem. Since the COM Express interface is standardized, it was assumed that a choice of COM Express implementations would be available for use in the BIC. After this design decision, the GDDR5 RAM capacity in the Xeon Phi cards went up to 16 GByte, and the Cluster-Booster protocol introduced optimizations that did require the full Xeon Phi RAM to be mapped into the BIC PCI Express address space. As a result, a suitable COM Express board had to support a PCIe address range of 512 GByte and memory mapped I/O space of more than 256 GByte. Common x86 desktop CPUs from Intel or AMD with available BIOS versions did not fulfil these requirements, and no COM Express board could be found that relied on a server-class CPU with sufficient PCIe address range and MMIO space.

Faced with this impasse, Eurotech quickly adapted a liquid-cooled Intel Xeon board they had under development to be used for an updated BIC design. This "Juno" board connects to the PCI Express switch on the BIC board via PCI Express ribbon cables, and with an adapted BIOS produced by a Eurotech affiliate, does fulfil the address space requirements. Physically, the Juno board attaches to the BIC board, shares the liquid cooling connection, and extends the thickness of the BIC. To adapt to this change, it was decided to change the spacing of chassis in the rack, limiting the number of chassis to six per rack, and the size of the final DEEP Booster to 12×32 = 384 Booster nodes.

Some changes were necessary to the Ethernet switch (BIC) and management device (BNC) firmware to preserve the remote management functions.

In retrospect, Eurotech has reacted very quickly to a potential showstopper, and they came up with a well-integrated and workable solution.

A lesson to be learnt here is that extreme caution has to be taken to assess the effects of seemingly innocuous changes in design and requirements.

VII. *Introduction of Pseudo-BIC systems and HDI6 adaptor*: To speed up the test & validation of the BNC design, standard Intel Xeon servers were equipped with InfiniBand and EXTOLL Galibier NICs, enabling them to stand in for a BIC and run the Cluster-Booster protocol as well as the Xeon Phi management software. These servers were christened "Pseudo-BIC" and did serve the project extremely well.

The Pseudo-BICs could be connected to the initial "debugging" Backplane (see VIII below) using standard EXTOLL cables; for use with the final Backplane, a special adapter was developed and produced by University of Heidelberg. This "HDI6 adaptor" has a Samtec HDI6 connector for plugging in an EXTOLL cable and a Molex connector for plugging into the BIC slot of the Backplane. To enable optical EXTOLL cables to be used, it also includes a 3.3 V power supply.

The Pseudo-BICs were used for the test and validation of the BNC designs, for the "Proto-Booster" (see X below), and for the bring-up of the first production Booster chassis. They also serve as BICs in the ASIC Evaluator (see section 2.3).

VIII. *Backplane design, tests and bring-up*: Eurotech did design and produce a number of samples of a small, "debugging" Backplane that can host four BNCs (equivalent to eight Booster nodes) and provides multiple connectors for debugging, including

sockets for Ethernet and EXTOLL connections. This early Backplane was instrumental for testing and validation of the BNC design.

The Backplane was initially designed for up to 10 Gbit/s per lane and eight lanes per EXTOLL link. Molex connectors of 100 Ω impedance were chosen, and the complete Backplane (and BIC/BNC) designs were based on this assumption. Unfortunately, the supplier did pull the 100 Ω line of products, and Eurotech was forced to use 85 Ω connectors instead. This did require a full signal integrity re-evaluation, which luckily did not identify relevant signalling problems. However, the project did lose time because of this change, and this clearly shows the critical dependence from suppliers outside the project. It also shows that important components for HPC systems are not being developed anymore in Europe.

The final Backplane design hosts eight BNCs and one BIC (two Backplanes make up one Chassis). No debugging connectors could be included due to space constraints. The transition from the "debugging" Backplane to the final version was straightforward, with only two issues occurring: manufacturing problems with the Molex connectors resulting in few bent pins and signalling problems with the Ethernet links across the Backplane. The latter did require a deep analysis and could be fixed by a HW patch applied to the BNC boards.

IX. *Rack, cooling and power distribution design*: Eurotech's original Aurora architecture was based on power distribution bars that supply 48V directly to all chassis. It also used quick disconnect connectors sourced from aerospace suppliers that should enable hot plugging of blades without posing a risk of liquid leakage. The initial rack design for the DEEP Booster also adopted these two aspects.

Direct liquid cooling was one of the main design points of the DEEP System. To this end, the cooling concept developed in the project works reliably and efficiently, allowing inlet water temperatures of up to 50°C for stable system operation and hence allowing for free cooling year-round in any climate. Although direct liquid cooling requires additional effort during installation as it integrates tighter with the building infrastructure, its operation is as trouble-free as air-cooling when following the existing best-practice guidelines from the American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE).

A severe excursion during operation of the DEEP Cluster at Juelich did force a re-evaluation of the rack, cooling and power distribution deign. A liquid leak did short out some of the per-chassis voltage converters with resultant damage to several chassis. To prevent such an excursion from happening again, Juelich required that the 48 V power has to be switchable per chassis, that changes be applied to the liquid connectors that make an accumulation of liquid near the voltage converter unlikely, and that additional leakage sensors be integrated into the chassis. These changes were applied to the DEEP Cluster and to the DEEP Booster, with Ethernet-controlled relays being used to control the per-chassis power supplies. A control authority was introduced to the RAS system to automatically shut off power to chassis in case of power use excursions or overheating of the system.

The above is a classic example of learning from operational experience; the project did act very quickly and materially improved the design of the DEEP system to prevent a reoccurrence of liquid leakage base excursions and damage.

X. *Booster evaluator and system bring-up*: To provide the system SW developers with an early development platform, an additional evaluator system was defined and produced:

the “Proto-Booster” uses one “debugging” Backplane and up to four patched early BNC cards. Physically placed at University of Heidelberg, it is coupled to a Pseudo-BIC and made available to DEEP system SW developers over the Internet. This evaluator has enabled WP4 to develop the Cluster-Booster protocol and the highly tuned global MPI for the DEEP system well before the definitive Eurotech Booster hardware became available.

After M36, the first two half chassis with the final Backplane and BNC versions became available at Juelich, followed by the EEE at Garching. They were initially tested using two Pseudo-BICs. The bring-up did focus on establishing the RAS plane (Ethernet via front panel) and the EXTOLL connections between BNCs and BICs. Changes had to be made to the Ethernet switch and device management firmware on the BNCs, and also the EXTOLL FPGA firmware was updated to correct bugs and signal quality issues with the LVDS links that connect the two NICs on a single BNC. A major area of work was the creation of the torus network topology, since the EXTOLL auto-configuration mechanism did not work. On the EEE, the issues with the Ethernet links over the Backplane were analysed and HW patched for the BNCs were developed.

Once the updated BIC design was ready, all twelve Booster chassis were installed at Juelich, and the test and bring-up of the full system started in earnest. Significant time was spent to exhaustively test the EXTOLL links, and to bring the BIC network bridging up to speed. For the latter, a tricky issue with the PCI Express connection to the InfiniBand HCA had to be solved, which resulted in spurious device removal/insertion events.

The booting and control of Xeon Phis could be established after additional tweaks to the EXTOLL drivers, and due to operational requirements, a new MPSS version with LDAP support was brought up.

At the time of writing, the Booster can finally be used to run applications. One remaining system issue under analysis concerns the tunnelling of IP communication over EXTOLL. Turning this feature on leads to severe instability, with the suspected reason being problems in the EXTOLL driver or in its interaction with the Linux kernel that runs on the Booster nodes. The work-around is to use a slower IP over MPSS setup – impact to application performance is expected to be minor, since IP is only used for process startup and control, and all MPI communication uses EXTOLL directly.

In retrospect, it is clear that the Proto-Booster has been instrumental to enable progress in the project during HW analysis and debugging. This has been particularly important due to the earlier delays in EXTOLL implementation having pushed out the HW tests. It also clearly corroborates the importance of having technology evaluators as soon as possible.

The system bring-up at Juelich and Garching has taken significant time, with a number of issues coming up that were not caught in earlier versions of the hardware, were caused by updates to the SW stack triggered by operational requirements, or relate to the scale of the system. Collaboration between partners has been very good, and constant progress is being made to sort through the issues and solve them.

2.3 EXTOLL DEEP ASIC Evaluator

The ASIC Evaluator (AE) was proposed as a mitigation of the mounting delays in the EXTOLL ASIC design, test and manufacturing. The main reason for these was entirely out of the sphere of DEEP, being the SERDES IP, which was bought from an external IP provider and had many problems which need to be fixed with mask changes. At the month 24 review, the creation of a smaller scale system prototype with the TOURMALET ASIC implementation of EXTOLL then planned for early 2015 was ratified by the reviewers. Christened "ASIC Evaluator", this system would demonstrate DEEP Booster functionality and performance achievable with the faster ASIC NICs. By necessity, the level of integration had to be lower than for the Eurotech Booster, and capability for continued 24×7 operation after the end of the project was not assured; yet; consensus was reached that this approach would enable the project to predict the achievable performance of a full-scale DEEP system with the latest EXTOLL NIC technology, and that such findings would be an important part of the DEEP results and justify the effort to create the ASIC Evaluator. In the month 36 review, the scale was fixed at 64 Booster nodes that would form a 4×4×3 torus.

The ASIC Evaluator was designed, built and tested by partner University of Heidelberg, with EXTOLL GmbH and Megware assisting. The design is documented in Deliverable D3.7 at month 43.

Important design decisions, steps in the development, and lessons learnt in manufacturing, integrating and bringing up of the ASIC Evaluator do include:

I. *Immersive cooling solution*: The implementation required a different cooling solution of the Booster part, because the cold plate technology from Eurotech was too expensive to modify for only a small amount of nodes. Therefore UniHD decided to evaluate a new cooling approach for the AE Booster. The goal was to study a dense assembly which might be useful for an Exascale system in the future.

   For the AE Booster a two-phase immersion cooling technique with Novec 649 fluid from 3M was implemented. All electronic components of the Booster are immersed into the fluid, which is inert, non-conducting and not harmful for the environment. The global warming factor of Novec 649 is equal to 1. In collaboration with 3M and Wieland AG, a chassis has been developed which fits into 19" racks and can be cooled with hot water of about 35˚ C inlet temperature und outlet temperatures around 49˚ C to 50˚ C. Novec 649 has a boiling temperature of 49˚ C. Using the two-phase approach avoids using pumps for the fluid because the vapour bubbles increase the convection and can remove the heat from all components in a very efficient manner.

   It should be noted that boiling enhancement structures for the most power consuming components like the Intel Xeon Phi CPU and the ASIC are required to achieve small enough bubbles which then can efficiently remove the heat from the components.

   The immersion caused trouble during development, because some of the materials have shown to be incompatible with Novec 648, e.g. the silicone material for making cable feedthrough gas tight showed degradation over time and there is no alternative material in the moment.

   The whole task of immersion cooling showed to have a steep learning curve and required a lot of interdisciplinary know-how, e.g. "plumbing", thermodynamic physics, mechanical design and chemical compatibility studies.

II. *Booster Node design and packaging*: For the AE Booster, the Booster nodes consist of a dense form factor 7120D Intel Xeon Phi card (DFF) and the PCIe version of TOURMALET NIC (with the EXTOLL ASIC). Both are connected via 16 PCIe lanes between the PCIe ×16 interface of the DFF-KNC and the PCIe ×16 Interface of the TOURMALET board. .A Dense Backplane (DBP) structure has been developed to hold eight Booster Nodes (i.e. eight pairs of one KNC and one NIC each) Four of such DBPs can be assembled into one chassis of the AE Booster. The interconnection network is mostly connected inside one chassis using short length rigid-flex cables, forming a 4x4x2 3D Torus. Using cables and not a multilayer backplane allows to interconnect the ASIC boards on top of the 8 node structure and at the same time leaving space for the Novec vapour passing between the cables to the heat exchanger pipe at the top of the chassis.

   This task seems to be a simple and straightforward approach, but it showed that there had been a lot of pitfalls, which cost precious project time and effort.

   The DBP was in principle a simple PCIe interconnect structure with only four PCB layers, two for power and two for transmit and receive PCIe lanes. A 100MHz reference clock distribution had to be added for 16 boards. The server power supply units (PSU) had also been integrated into the immersion cooling system which made the 12V connection from PSU to DBP much simpler, because any connection from outside the chassis needed to be gas-tight. In addition to the 12V supply, the DFF-KNC board required a 3.3V supply of about 1.5A per DFF-KNC. The PSUs did not support 3.3V and therefore a DC to DC converter was added to the DBP. This was a major problem because it turned out that this device emitted a 300MHz self-resonant pulse under special load conditions. These pulses disturbed the 100MHz reference clock at very seldom points in time. It required a large measurement effort to find this as the reason for PCIe bus retraining sequences which have been detected in the immersion chassis. Measurement was difficult, because the system was operated in a gas-tight mode and setting probes was difficult. A second release of the DBP fixed this error by changing the DC to DC converter type and is tested and verified.

III. *Interconnect cabling*: Nonetheless, a first chassis has been completed and tested. It contains one 4x4x2 3D Torus, which is only half of the planned AE Booster configuration. Connecting two such chassis as a 3D torus is a problem which shows that interconnect cabling structures are one of the major challenges for Exascale system. So far, we have used the HDI6 connector from Samtec, which is the densest connector in the market, but it is not dense enough to run 16 links with 12 lanes in bidirectional way between the two chassis. The available space at the backside of the chassis is not sufficient to run the 576 twinax cables for the Z dimension using available interconnect and connector technology. Here new ways of dense connectors and interfaces to electrical or optical links must be explored.

IV. *System bring-up*: At the time of writing the node to node latency and bandwidth are still being measured. Stable operation of the 32 nodes is work in progress, and the performance measurements will be done soon. The EXTOLL links are stable at 5Gbit/s per lane and the increase to 8Gbit/s is under production and test. PCIe interfaces are operational at PCIe generation 2 (which is the limit for KNC) with 5Gbit/s speed.

   Operation of the 2-phase immersion cooling chassis had been proven and will be further developed for the density and energy efficient cooling. This chassis is, beside

the shortcomings of a prototype, the densest accelerator based Booster without host nodes for KNCs.

3 Improving DEEP

The DEEP Deliverable D9.2 at month 45 contains an in-depth investigation of technology trends in HPC system architecture, processors and memory, as well as interconnects, and system software plus programing environments. It also charts the requirements wrt. energy efficiency and resiliency, and discusses how HPC architectures can evolve to provide Exascale levels of performance in a sustainable way.

This section discusses the changes and improvements to the DEEP architecture and the DEEP systems suggested by the experience gathered during the project that can be implemented in the next few years. It also details which of these are already incorporated in the sibling DEEP-ER project, which is planned to install the first next-generation prototype in around mid-2016.

This section does not discuss the need to address further Exascale challenges (such as scalable and efficient parallel I/O and resiliency) – measures to evolve the DEEP architecture to meet these have been defined in DEEP-ER.

3.1 DEEP Architecture

I. *Self-hosted Booster Nodes*: At the start of DEEP, all potential candidates for a Booster Node required an x86 control system attached by PCI Express for their operation. In addition, the prevalent programing model of these were the offload of rather small code parts to such a CPU, using it strictly in "accelerator" mode. Announcements by Intel (second generation Intel Xeon Phi), and first products by NVIDIA (Tegra X1) have changed the picture: It is now possible to use highly parallel and optimised stand-alone CPUs as Booster Nodes. Similar systems leveraging an ARM CPU and attached GPGPU will likely come from AMD and others.

   This presents a unique opportunity to simplify the DEEP architecture – the Booster Interface would no longer need to act as a control agent for the Booster Nodes. Instead, its sole function will be to bridge between the two dissimilar networks in the Cluster and the Booster.

   The second generation Intel Xeon Phi (code named Knights Landing, KNL) is interesting as a future Booster Node CPU, since it continues to provide the combination of general-purpose instruction set and highly tuned SIMD performance. The DEEP programing model could be used without changes, and the DEEP SW layers would need to be adapted, but not rewritten. With a performance to energy ratio of up to 10 GFlops/Watt (as claimed by Intel), such a choice could show a nice progression in energy efficiency. The DEEP-ER project had made that choice.

II. *Single, uniform network architecture*: As discussed in section 2.1, the state-of-the-art at the start of DEEP did pretty much require a combination of two very different network technologies. Since EXTOLL at that time was a novel, not yet proven approach, it would have been very risky to use it for the Cluster (which needs to run a wide variety of SW codes that rely on the interconnect), and relying on InfiniBand for the Booster would have restricted scalability and created substantial risk regarding remote control of Booster Nodes.

While the project has proven that an efficient Booster Interface can be built, and that network bridging can be made very efficient (CBP), this entity does complicate the architecture and implementation. In addition, the operational management of a combination of two networks joined at the hip will be complex and likely more than that of a single network. For this reason, opportunities to move the architecture to a single, highly scalable and efficient network should be investigated and taken up, in particular in light of the discussion under I. above.

Choices for future DEEP Architecture networks do include EXTOLL, EDA InfiniBand and Intel® OmniPath interconnect.

DEEP-ER has made the choice to use a uniform network, and is has selected EXTOLL TOURMALET.

III. *Evolution of EXTOLL network*: During the DEEP project, EXTOLL has made very significant progress: the earlier FPGA implementation was improved and hardened during the DEEP system development, and the new ASIC implementation (TOURMALET) provides performance that is competitive with both EDA InfiniBand and OmniPath.

Therefore, one straightforward evolution of the DEEP Architecture is to adopt EXTOLL as the interconnect for Cluster and Booster. Its capability to provide seven independent links per NIC enables the construction of a 3D Torus for the Booster Nodes, and the connection of that torus to a choice of Cluster topologies without requiring dedicated EXTOLL switches.

The future evolution of the EXTOLL technology will be a key factor here. Topics to consider include the integration of CPU and NIC with mechanisms that are more performant and efficient than PCI Express, the availability of EXTOLL switches (to accommodate general topologies), and the development of advanced cabling solutions. In addition, EXTOLL has to be properly integrated into state-of-the art system management solutions, and dynamic routing capabilities will be required to mitigate connection problems.

The DEEP-ER project has decided to use the EXTOLL TOURMALET technology.

IV. *Monitoring and RAS*: A very significant part of the DEEP Architecture is the sensor and actuator infrastructure combined with the scalable approach to collect, analyse, store and act on monitoring data embodied in the RAS plane concept.

The lessons learnt in DEEP will certainly improve the implementation of the sensors where necessary. The more important topics are a truly seamless integration of all compute and management CPUs (BMCs and μ-controllers) with the RAS plane, and ultimate scalability of the fabric used by the RAS plane. In addition, should the Booster Interface be eliminated following the discussion of I. and II., the RAS plane will need specific "concentrators" that would collect data from parts of the system.

The DEEP-ER project will re-use the RAS plane concept from DEEP as a baseline.

3.2 Eurotech DEEP Booster

I. *Booster Node integration*: The integration developed by Eurotech for the DEEP Booster Nodes can of course be applied to other types of accelerators (GPGPUs), and to leveraged-boot versions of the second generation Intel Xeon Phi. This would

however require a significant effort for the design-in of the TOURMALET ASIC, and qualification of all links to the full 10 Gbit/s per lane speed.

While the above is technically possible, a more impactful direction of development would be to leverage the competency in high-performance backplanes to disassociate the compute and network devices, enabling a more flexible mix-and-match of technologies. One example already taking shape is the Aurora Hive architecture, which implements a hierarchical system with nodes consisting of a small number of Xeon Phis or GPGPUs connected by PCI Express to each other and a NIC.

With a little farther view into the future, and considering the DEEP-ER project, the emergence of leveraged-boot Booster Nodes would enable the re-use of Backplane and integration technology already developed for server CPUs, leading to a DEEP System that would have Cluster and Booster Nodes, that can coexist in a Backplane, and associate with NICs and other peripheral devices via PCI Express. This is the direction the DEEP-ER project has adopted.

II. *Rack, cooling and power distribution*: Eurotech has already streamlined the cold plate technology compared to the one used for the DEEP BNCs. The energy and temperature measurements on the EEE clearly show that hot water cooling of 250 Watt CPUs is feasible, and that sufficient margin is left for all-year free cooling in most of Europe.

Further improvements will likely target management and operational aspects; enabling true hot pluggability of Booster Node blades that connect to a liquid distribution manifold, while ensuring leak tightness can reduce maintenance overhead, and the best control laws and policies for temperature and flow will emerge from operational experience.

III. *Management components and RAS*: The importance of ensuring that all components of a DEEP System can be properly managed and controlled was clear from the outset. Several compromises had to be made – f.i. the Intel Xeon Phi only provides a subset of the manageability functions and interfaces of an Intel Xeon, and the BIC design finally adopted lacks a BMC that can control all components.

Future systems, as argued in I. and II., will likely use stand-alone CPUs, with standard manageability provisions. As a result, the complex manageability firmware can be streamlined, and full control over all parts of the system established.

For the DEEP-ER project, the second generation Intel Xeon Phi does provide full server-style management functions, and a BMC that is compatible with Eurotech's firmware can and will be integrated.

3.3 EXTOLL DEEP ASIC Evaluator

I. *Immersive cooling system*: The immersive cooling concept and implementation (christened "GreenICE" by EXTOLL) will be applicable to other compute technology, for instance GPGPUs or compact x86 or ARM compute boards. Its main requirement is that the chosen technology supports a sufficiently fast PCI Express connection. A next step in the evolution of GreenICE can therefore be a redesigned basin for alternative accelerators.

The experience with the DEEP ASIC Evaluator has shown, that the long-term stability of an immersive cooling solution has to be carefully validated for the actual compute technology used, and that modifications to off-the-shelf parts might be needed.

II. *Scalability*: Currently, all NICs are connected by per-link cables using HDI6 Samtec connectors. This does work reasonably well within one basin (32 NICs), yet poses problems for the connections between basins. Focus of future development will be on addressing this challenge, with approaches including the use of bundled or multi-mode optical cables, the design of matching connectors, and the investigation of Backplane solutions applicable to the GreenICE design.

4 The DEEP Legacy

The DEEP System has for the first time implemented the Cluster-Booster architecture, proving that the key concept of dynamically associating different kinds of computing resources to best match workload needs can be implemented with state-of-the-art multi- and many-core technology, and that such a system has the potential to provide a superior combination of scalability and efficiency. It has opened up a new avenue towards affordable highly efficient and adaptable Exascale-class systems, merging the separate lines of massively parallel and commodity Cluster systems. The sibling project DEEP-ER is already carrying the flag further by integrating novel memory and storage concepts and providing scalable I/O and resiliency capabilities.

The direct-switched EXTOLL network could show its value in a HPC system of significant size and this proof of concept, combined with the level of performance of the new TOURMALET implementation will be instrumental in putting EXTOLL up as a credible, European alternative to switched networks such as InfiniBand.

With its unprecedented integration of sensors, the DEEP System delivers a wealth of voltage, current and temperature data for all system components at high frequency, and uses this data for good measure to optimise operating parameters and safeguard operation. This example will influence future HPC system designs and create opportunities for advanced monitoring data analysis and data-driven system management.

Eurotech is one of the world-wide pioneers of direct liquid cooling for HPC. The DEEP project did show that hot water cooling can be safely operated, is compatible with modern system technology and can indeed provide free cooling year-round. These results will most importantly shape the expectations of HPC customers, who now know that they can eliminate an important part of operating costs, and in turn materially increase the take-up of hot water, direct liquid cooling by future HPC systems.

The DEEP system software and programing model were carefully architected to be based on existing standards and product-quality solutions, and extend them where necessary to make the unique DEEP features available or enhance ease of programing. Supported by the application proof points, the resulting SW stack will certainly and substantially influence the direction of Exascale software architecture, with ParTec as a European industrial player in a key role.

Finally, managing a large system project and driving the co-design between applications experts, system SW developers and HW architects is no small task. JUELICH (for the whole project) and BADW-LRZ (for the critical energy efficiency area) have amply demonstrated their capability to do just that. This should set them up as prime partners for the next rounds of system-centric co-design projects. BSC has been instrumental for the success of the programing model co-design.

5 References and Applicable Documents

1	N. Eicker and T. Lippert	An accelerated Cluster Architecture for the Exascale. Gesellschaft für Informatik e.V., Parallel-Algorithmen und Rechnerstrukturen, 2011, vol. 28, pp. 110 – 119	2011

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

H

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 Deutschland GmbH, Munich, 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

JUELICH: 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

O

P

PCB: Printed Circuit Board

PCI: **P**eripheral **C**omponent **I**nterconnect. 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

Q

R

SMBus: System Management Bus

SW: Software

T

U

USB: Universal Serial Bus

V

W

WP: Work Package

X

Y

Z