SEVENTH FRAMEWORK PROGRAMME

FP7-ICT-2013-10

DEEP-ER

DEEP Extended Reach

Grant Agreement Number: 610476

D3.6

DEEP-ER Architecture Review

Approved

Version: 2.0

Author(s): H.Ch.Hoppe (Intel)

Contributor(s): M.Cintra (Intel), N.Eicker (JUELICH), M.Nüssle (UHEI), J.Schmidt (UHEI), I.Zacharov (EUROTECH)

Date: 04.05.2017

Project and Deliverable Information Sheet

<table>
<tr><td rowspan="7">DEEP-ER Project</td><td colspan="2">Project Ref. №: 610476</td></tr>
<tr><td colspan="2">Project Title: DEEP Extended Reach</td></tr>
<tr><td colspan="2">Project Web Site: http://www.deep-er.eu</td></tr>
<tr><td colspan="2">Deliverable ID: D3.6</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 / March / 2017</td></tr>
<tr><td>Actual Date of Delivery:
31 / March / 2017</td></tr>
<tr><td colspan="2">EC Project Officer: Juan Pelegrín</td></tr>
</table>

* - The dissemination level are indicated as follows: **PU** – Public, **PP** – Restricted to other participants (including the Commission Services), **RE** – Restricted to a group specified by the consortium (including the Commission Services). **CO** – Confidential, only for members of the consortium (including the Commission Services).

Document Control Sheet

<table>
<tr><td rowspan="6">Document</td><td colspan="2">Title: DEEP-ER Architecture Review</td></tr>
<tr><td colspan="2">ID: D3.6</td></tr>
<tr><td>Version: 2.0</td><td>Status: Approved</td></tr>
<tr><td colspan="2">Available at: http://www.deep-er.eu</td></tr>
<tr><td colspan="2">Software Tool: Microsoft Word</td></tr>
<tr><td colspan="2">File(s): DEEP-ER_D3.6_Architecture_Review_v2.0-ECapproved</td></tr>
<tr><td rowspan="4">Authorship</td><td>Written by:</td><td>H.Ch.Hoppe (Intel)</td></tr>
<tr><td>Contributors:</td><td>M.Cintra (Intel), N.Eicker (JUELICH), M.Nüssle (UHEI), J.Schmidt (UHEI), I.Zacharov (EUROTECH)</td></tr>
<tr><td>Reviewed by:</td><td>J.Morillo (BSC), E.Suarez (PMT)</td></tr>
<tr><td>Approved by:</td><td>BoP/PMT</td></tr>
</table>

Document Status Sheet

Version	Date	Status	Comments
1.0	31/March/2017	Final	EC submission
2.0	04/May/2017	Approved	EC approved

Document Keywords

Keywords:	DEEP-ER, HPC, Exascale, architecture, storage, interconnects

Table of Contents

Project and Deliverable Information Sheet 1
Document Control Sheet 1
Document Status Sheet 2
Document Keywords 3
Table of Contents 4
List of Figures 5
List of Tables 5
Executive Summary 6
1 Introduction 7
2 DEEP-ER Architecture and Design Decisions 8
2.1 CPU and Booster Node 9
2.2 System and Interconnect 10
2.3 Aurora Blade Chassis 13
3 DEEP-ER Architecture and Design Review 18
3.1 CPU and Booster Node 18
3.2 System and Interconnect 19
3.3 Aurora Blade Chassis 23
4 Conclusion 26
5 References 27
List of Acronyms and Abbreviations 29

List of Figures

Figure 1: DEEP-ER Top-Level Booster and Node Architecture. 9
Figure 2: Intel DC P 3700 PCIe attached SSD Card. 12
Figure 3: NAM architecture. 13
Figure 4: Intel S7200AP board hosting one KNL socket and six DDR4 DIMMs. 13
Figure 5: EXTOLL TOURMALET NIC as used for the DEEP-ER SDV and Booster....19
Figure 6: End-to-end MPI bandwidth (above) and latency (below) on the DEEP-ER SDV, using ParaStation MPI and the OSU benchmark. 20
Figure 7: IOZONE sequential read/write performance for Intel DC P 3700 device attached via PCIe ×4 compared to SATA SSD....21
Figure 8: NAM Prototype board (left) and integration of two NAM boards into the DEEP-ER SDV (right)....22

List of Tables

Table 1: Summary of the main DEEP-ER Co-Design Decisions and their implementation in the Eurotech Aurora Blade based DEEP-ER Booster. 8
Table 2: Storage interface comparison. 11

Executive Summary

The DEEP-ER system architecture was first specified in Deliverable D3.1 [1], and additional detail plus sketches for actual implementations of the architecture were discussed in Deliverable D3.2 [2]. Through the project, the plan for implementing the DEEP-ER Prototype did evolve, with a key decision in the interim month 18 review to commit to Eurotech's Aurora Blade architecture, which was reconfirmed at the review at month 24. The specification of that architecture became available in D8.1 [3], and the prototype system design, implementation, manufacturing and integration was performed in Work Package 8.

This Deliverable reviews the fundamental architecture decisions as documented in D3.1 and D3.2, plus the relevant high-level aspects of the Eurotech design as described in D8.1. It will not go very deep into Aurora blade design details. Within this scope, the document reflects the architecture decisions by the results achieved with the test systems, the Software Development System (SDV) and the Aurora Blade DEEP-ER prototype. Wherever possible, actual application results are used to assess the actual achievements and compare to the targets set during the co-design process.

1 Introduction

This Deliverable reviews the design decisions taken during the DEEP-ER project (as described in Deliverables D3.1, D3.2 and D8.1) in light of the achieved results in system performance, scalability, and energy efficiency, using actual application results from WP6 where possible. Scalability predictions are contained in Deliverable D7.2 [4].

Driven by a tight co-design collaboration with the system and application SW experts in the project, the DEEP-ER architecture was designed to bring substantial improvements in compute and interconnect speeds, available memory capacity and memory bandwidth, and very importantly fast local storage-class memory resulting in substantially improved end-to-end I/O performance and reduced checkpointing/restart overheads. In addition, the concept of fast, network attached memory was realized by the NAM prototype, which also offers compute functionality very close to high-performance memory.

The next section contains a concise definition of the key architectural decisions and features. Section 3 includes the actual detailed review of these, with references to discussions of future trends and developments in Deliverable D7.2. The document closes with a summary section, a list of references and a glossary.

2 DEEP-ER Architecture and Design Decisions

The DEEP-ER project did reach the fundamental system architecture and design decisions in a close co-design collaboration between the technology providers, system and application software developers and system operators in the consortium. This was an iterative process, with a first architecture detailed in Deliverable D3.1, which was further refined into an initial preferred system design in Deliverable D3.2 (“Brick Architecture”), which itself was duly transformed into the Eurotech Aurora Blade architecture as finally implemented (Deliverables D8.2 [5] and D8.3 [6]). The ongoing co-design collaboration did ensure that the system and application software requirements were met for each of these iterations.

This Deliverable looks at the architecture decisions that did shape the final DEEP-ER system and its components. It does not discuss the previous iterations – detailed material on why changes were required are contained in the respective Deliverables.

In a nutshell, the crucial decisions are summarized in Table 1. The following subsections rephrase the crucial decisions to provide a basis for the assessment contained in Section 3.

Item	**Value**	**Aurora Blade Architecture**
Bootable KNL nodes	~3 TFlop/s DP per CPU	One KNL CPU per node, equipped with full 16 GByte of fast, on-package memory (MCDRAM).
DDR4 Memory	96 GByte 120 GByte/s	Use all six memory channels to provide 96 GByte DDR4 memory via ULP DIMMs; theoretical capacity is 384 GByte.
Intra-node connection	PCIe gen 3 ×16	Each node is connected to peripherals on Root Card using two PCIe gen3 ×16 links.
Network bandwidth	>100 Gbit/s link bandwidth	EXTOLL TOURMALET NIC providing six links with 100 Gbit/s bandwidth each and attached via PCIe gen3 ×16.
On-node NVM	400 GByte NVM devices using NVMe protocol	One Intel® DC P3700 device attached to each node using PCIe gen3 ×4.
NAM infrastructure	≥ 2 devices attached to DEEP-ER system	Integrate NAM prototype into the peripherals design.
Commercial viability	DEEP-ER technology to be commercially viable	Aurora Blade Architecture is developed as a general purpose commercially available machine, and uses commercial components.
Density and energy efficiency	System to have better density & energy efficiency than commercial off-the shelf systems	Aurora blade architecture delivers superior density* and full direct liquid cooling, improving energy efficiency.

Table 1: Summary of the main DEEP-ER Co-Design Decisions and their implementation in the Eurotech Aurora Blade based DEEP-ER Booster.

* This was meant relative to a naive implementation of a KNL node as a 1U 19“ server.

2.1 CPU and Booster Node

One of the most fundamental architecture decisions of the project was that the Cluster and Booster nodes would be composed from components (CPU, storage-class memory, NIC) integrated using a capable existing intra-node interconnect technology, rather than tightly integrating them on a single node board. Figure 1 shows the according top-level architecture of the DEEP-ER Booster and the DEEP-ER Booster node.

The two main rationales driving this decision was the reduction of risks to design, build and integrate the DEEP-ER system (since partners can focus on the various components and leverage available technology), and the flexibility to upgrade and change components over time, which strengthens the market viability of DEEP-ER system and technology.

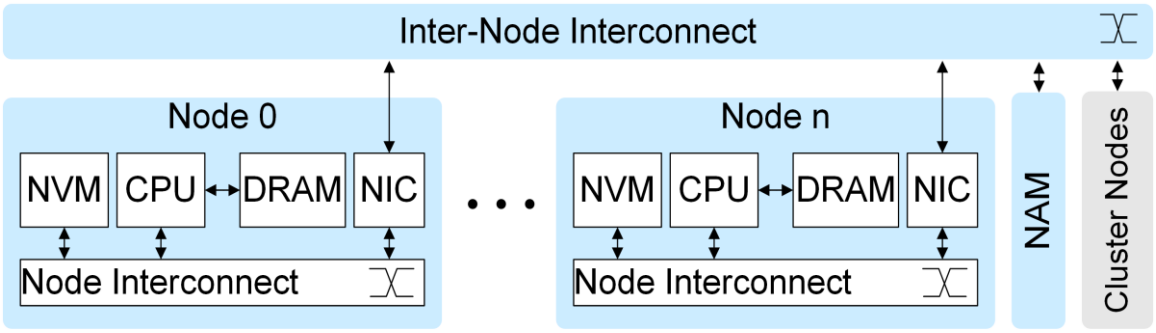

Figure 1: DEEP-ER Top-Level Booster and Node Architecture.

2.1.1 Booster Node CPU

The DEEP project had used the first generation Intel® Xeon Phi™ many-core CPU for the Booster, which depending on the version provides approx. 60 cores and a fixed amount of 16 GByte of GDDR5 main memory. The CPU cores offer the AVX instruction set extension for 256-Bit vector instructions, and a peak performance in the vicinity of 1 TFlop/s. Experience with the DEEP applications did show that the increase in core count and the addition of 256 bit vectors gave a nice performance boost compared to conventional Intel® Xeon® CPUs, yet the amount of available memory was seen as constraining, and the relatively low clock speed of the cores could compromise bandwidth to/from a modern interconnect. In addition, this version of Intel Xeon Phi was only available on a PCIe add-on card, and it required an Intel Xeon processor to boot and operate.

Compared to the above, the second generation of Intel Xeon Phi (code name “Knights Landing” or KNL) [7] did offer significant improvements: the CPU is now self-booting†, it supports six DDR4 memory channels with a total memory capacity of 384 GByte, and it introduces fast on-package memory (MCDRAM) which delivers 4× the DDR4 bandwidth (120 GByte/s vs. 480 GByte/s). The compute core count is now a maximum of 72, cores support out-of-order execution, clock speeds are higher, and the vector width has been increased to 512 bits with the AVX-512 instruction set, with very significant improvements in gather/scatter operations. This all leads to a peak performance per CPU of approx. 3 TFlop/s.

In contrast to Intel® Xeon®, KNL does not support multi-CPU nodes with shared memory between the CPUs. The thermal design power of KNL is similar to the one of the previous

† A leveraged-boot version on a PCIe addon card will be released; this version will only have limited memory, comparable to the first Intel Xeon Phi generation.

generation, with some additional power being consumed by the platform controller hub (PCH) and the DDR4 memory. KNL fully supports the Linux OS, with modern distributions (such as CentOS 7.2 and later) including all required kernel extensions and drivers.

In light of the improvements in peak performance, the high memory bandwidth delivered by MCDRAM, the flexibility of a general-purpose CPU and the incremental changes in the programming model, the project decided to use KNL as the Booster CPU.

2.1.2 Intra-Node Interconnect

The intra-node interconnect in DEEP-ER had to support the bandwidths and latencies required by applications for inter-node communication and storage access, as provided by the evolving fabric and storage device technology. For InfiniBand EDR and EXTOLL TOURMALET, link speeds of 100 Gbit/s were announced at the beginning of the DEEP-ER project, and this has did the minimum bar for the intra-node interconnect.

In addition, the availability of NICs and of storage devices supporting the selected interconnect had to be taken into account. For NICs, this clearly meant PCI Express, and for storage devices, the established SATA and SAS interfaces were being supplanted by PCIe attached storage that provided far better performance. This lead to the decision to select PCIe, with only its third generation delivering the required bandwidth for a ×16 connection (approx. 120 Gbit/s compared to approx. 64 Gbit/s for generation 2). This led to the decision to adopt PCIe gen3, and require connections of 16 lanes to the NICs and of a suitable width to the storage devices.

The technical risk of this decision was seen as manageable, since PCIe generation 3 was quickly becoming the industry standard in 2012, with new desktop and server platforms providing PCIe generation 3 lanes and a wide variety of PCIe add-in cards available that take advantage of it. The whole PCIe component ecosystem (retimers, switches, connectors, etc.) also quickly tooled up to support generation 3.

2.2 System and Interconnect

2.2.1 Inter-Node Interconnect

EXTOLL TOURMALET was selected relatively early in the project as the high-speed inter-node interconnect solution for DEEP-ER. The TOURMALET ASIC implements a full “network-on-a-chip” including a PCIe gen3 ×16 interface, communication engines for fast sending and receiving of packets, remote memory access and remote address space mapping as well as an integrated switch with up to seven external bi-directional links of 100 Gbit/s per direction of raw bandwidth. From this building block direct topologies can be built, with one device in each node of the fabric and cables connecting these together. Regular topologies like meshes and tori are a natural choice, but also irregular and hybrid topologies are supported by the router, which was exploited within the DEEP-ER project to link together several regular topologies and form a fabric of all of them.

Revision A3 of the TOURMALET ASIC became available for the project in time to be used for all machines. The A3 silicon revision with the respective board revisions enabled the EXTOLL network to run with full 100 Gbit/s per direction and link. As such, TOURMALET fulfilled the project’s requirements while also being a truly European technology.

The technology and especially the ASIC behind the network were entirely developed by the European SME EXTOLL GmbH. It is noteworthy that, while high-performance networking and HPC in general are of strategic relevance to the EU, development and production of the TOURMALET ASIC was financed completely privately, as public support for this kind of technology, unfortunately, seemed not to be available at the time.

2.2.2 Local Non-volatile Memory

Initial feedback from system and application SW developers quickly indicated that the use cases and requirements for the node-attached non-volatile memory would require a significant storage capacity, mainly driven by the BSC FWI oil & gas application and the need to store multiple checkpoints on each local device (for buddy checkpointing and for holding a sequence of checkpoints). This capacity had to be in the several hundreds of GByte class.

This and the planned scale of the full DEEP-ER system immediately ruled out the use of battery-backed RAM, early prototypes of emerging inherently non-volatile technology (like Intel 3D XPoint™, HP's announced resistive RAM, battery-backed HMC and the like). Instead, the project did decide to use the (then) latest NAND flash technology devices supporting the new NVMe protocol over PCIe generation 3. Table 2 shows a comparative study of storage interfaces).

	NVMe (PCIe gen 3.0 ×4)	**SATA Express (PCIe gen 3.0 ×1/×2)**	**SAS 12Gbit/s**	**SAS 6Gbit/s**	**SATA 6Gbit/s**
Port Bandwidth (GByte/s)	**4**	2	1.5	0.75	0.75
Streams per Port	**1**	1	1 - 2	1 - 2	1
Interface (GByte/s)	**8**	2	6	1.5	0.75
Duplex	**Full**	Full	Full	Full	Half
Host Controller	**No**	No	Yes	Yes	Yes
Added Latency (us)			25	25	22

Table 2: Storage interface comparison.

The clear winner of that comparison was the NVMe protocol over PCIe gen3, and after surveying the actual devices available in year 1 of the project, the Intel SSD DC P3700 device [8] with a capacity of 400 GByte was selected. This drive can deliver sequential read bandwidths of approx. 2.8 GByte/s and write bandwidths of 1.9 GByte/s. Figure 2 shows the device in its general configuration – the card is half-height and uses 4 lanes of PCIe gen3.

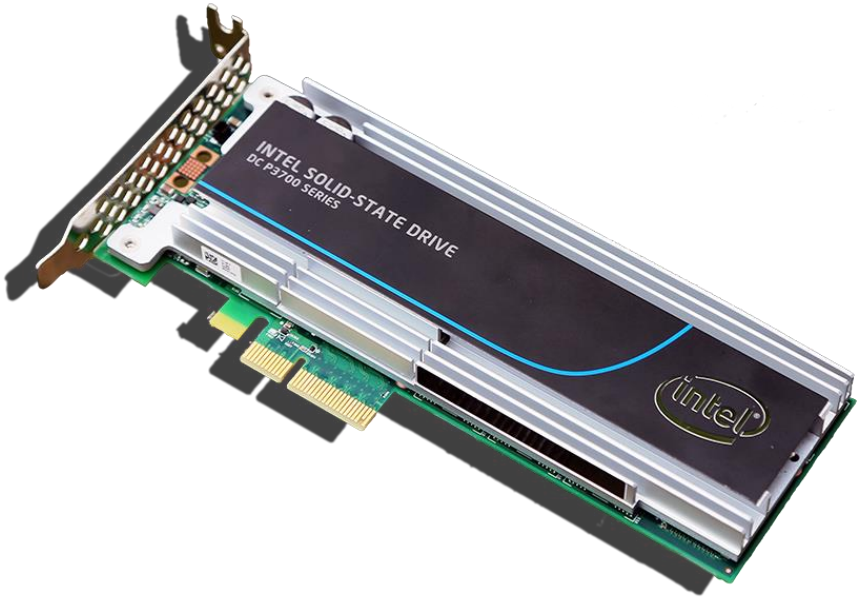

Figure 2: Intel DC P 3700 PCIe attached SSD Card.

2.2.3 Network Attached Memory

The architecture of the NAM has been discussed and fixed early in the project. A PCIe form factor PCB holds an FPGA and Hybrid Memory Cube (HMC) memory device and provides external interfaces to connect one or two EXTOLL links (with 12 lanes each) and PCIe gen3 x16 (for maintenance and reprogramming). Programmability of FPGAs allows for fast prototyping with short design turn-around times and which is extremely helpful for developing innovative hardware. This approach also avoids the extra hardware and manufacturing costs involved in conventional hardware re-spins.

As shown in Figure 3, the FPGA implements three different function blocks:

- EXTOLL fabric interface for one or two links
- NAM logic, implementing the RDMA functionality and any additional compute or data processing functions
- HMC controller

The selected FPGA device, a Xilinx Virtex7 690T, provides sufficient logic resources and high-speed I/O channels and was less expensive than current state of the art Xilinx Ultrascale FPGAs. In addition, UHEI had considerable design experience for similar PCB/FPGA combinations, lowering the risks in the design and bring-up of this new PCB.

HMC was chosen as it provides valuable advantages over DDR4 RAM. These are an increased bandwidth, a smaller I/O and PCB footprint, and increased energy efficiency. By the start of the project UHEI already had an early tested version of an HMC host controller, the firmware of which was released as Open Source, nullifying the cost to obtain a license and decreasing memory-link bring-up time. HMC also provides the ability to increase the capacity by daisy chaining additional devices.

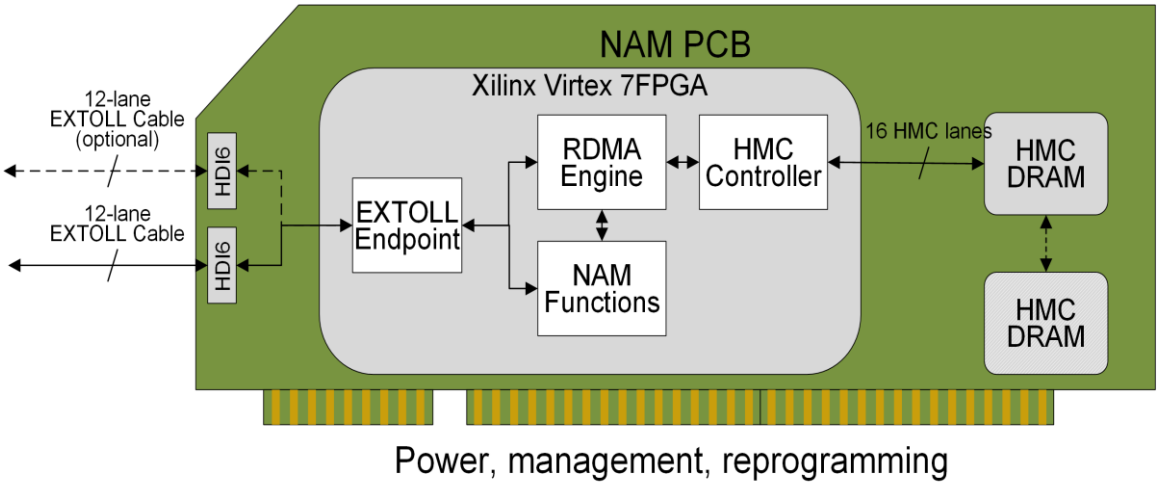

Figure 3: NAM architecture.

2.3 Aurora Blade Chassis

The name given by Eurotech to the final architecture for the DEEP-ER Booster is “Aurora Blade”. This section discusses the design rationale behind the Aurora Blade architecture, which consists of the KNL node board assembly with its use of ULP DIMMs, the Backplane, the Root Card and the specific Aurora direct liquid cooling technology.

2.3.1 Aurora Blade Node

The Aurora blade design is based on Intel’s S7200AP reference board [9] with a single Intel Xeon Phi (KNL) processor (Figure 4). The functionality of the S7200AP fully satisfies the design criteria for the DEEP-ER Booster node as captured in Table 1. Therefore, to optimize the development effort Eurotech integrated the S7200AP board into the Aurora chassis using three interface boards.

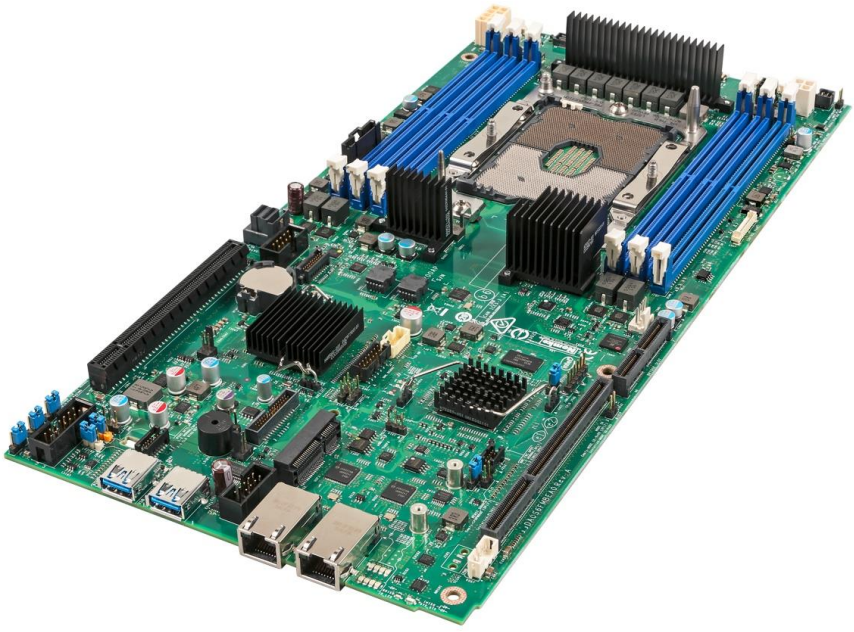

Figure 4: Intel S7200AP board hosting one KNL socket and six DDR4 DIMMs.

These interface boards were designed by Eurotech to connect the S7200AP to the Aurora infrastructure for power and high speed and management signalling, and also to adhere to the Aurora form factor. The form factor mandated the dimensions of 24.5×175×500 mm for each node with Molex connector connectivity to mate with the backplane. The three interface cards comprise:

- The DC/DC converter to provide 12V power from the 48V DC distribution (CROW). It contains sensor to measure the power and the accumulated energy consumed by the node. DEEP-ER Booster requirements contain a definition for such capability.
- The PCIe board for the Gigabit Ethernet management network and the PCIe signal shaping re-timers to define the interface to the backplane with the Molex connector (LARK)
- Two PCIe passive raiser boards that connect the S7200AP PCIe connectors to LARK (OWLS).

The CROW and LARK boards head and trail the S7200AP. The whole assembly is covered by the interposer for strength and to spread the heat from the electronics. An active cold-plate with water channels is firmly attached to the interposer for the water to absorb the heat.

To maintain the desired thickness of the assembly ultra-low profile (ULP) DIMMs have been selected for the memory subsystem. This is described in the next subsection.

The design of the Aurora KNL blade node was Eurotech's first 3D integration work. The design decisions were supported by a 3D model that accurately predicted how different parts would fit. During this phase of the design a criticality with the cabling of 1GigE network was identified[‡], for which a special stripped down cable had to be constructed that routed the 1 GigE signals from LARK to the front panel.

2.3.2 Ultra-Low Profile DIMM

The S7200AP board utilizes DIMMs for the memory subsystem. While this is a standard arrangement with the vertical mounting of the DIMMs on the server board, previous designs of Eurotech did require memory soldered to the board. Soldered-down memory presents a flat surface amenable to heat extraction with Eurotech style cold-plate that covers the whole board.

The decision to use the S7200AP board presented the challenge to fit the board with vertically mounted DIMMs into the architecturally defined 24.5 mm thickness and provide heat extraction from the vertically oriented surfaces in contrast to previous flat Eurotech designs.

The Ultra-low profile (ULP) DIMMs with vertical dimension of 17.8 mm allowed maintaining the required thickness of the assembly. In the mounted position the ULP DIMMs in the socket are 20 mm above the board, which is the maximum acceptable height. The active cold-plate for the board was designed with a cut out at the place of the ULP DIMMs, such that the ULP DIMMS protrude through the cold-plate.

The reduced surface of the ULP DIMMs presented a challenge for the heat extraction. A vertical arrangement to cool the ULP DIMMs was patented as result of this work.

[‡] A standard 1GigE cable could not be passed towards the front panel due to lack of space available.

2.3.3 Aurora Blade Backplane

The need for the backplane comes from the principal decision to separate nodes from the peripherals. These two entities follow a different evolution path, as they are typically developed by different companies with different pace of generation change. A unifying principle is the use of the PCI Express connections to couple them into a functioning unit; the PCIe evolution has again a different pace (generation change every 5 years). As discussed in sections 2.1.2 and 3.1.2, PCIe gen3 was selected for the DEEP-ER project.

As another principle decision born from the DEEP project is to prefer PCB to cables, a fixed PCB arrangement is more robust in terms of stability and maintenance of the installation. For maximum flexibility the backplane should route all the PCIe gen3 ports available on the node with 16 lanes (PCIe gen3 ×16 at 8 Gbit/s per lane) to the peripheral cage.

The Aurora backplane was designed in a generic way with two PCIe gen3 x16 links routed to the peripherals cage and two PCIe gen3 ×16 routed to an adjacent slot. A chassis holds 18 nodes and 36 peripheral cards with standard connectors; this gives the maximum capacity of the peripheral cage in a 19” form factor with the cards arranged along the sides of the cage (see section 2.3.4 for description of the root card). For the DEEP-ER Booster with the S7200AP based nodes the adjacent slot routing was not used, since the KNL has a total of just 36 PCIe lanes.

The PCIe signals together with the Gigabit Ethernet network and I2C control signals occupy all pins in the 120-pin Molex connector that mates each node with the backplane. The backplane is a passive PCB with 2x18 vertical Molex mail connectors to connect the nodes and the horizontal Amphenol Xcede connector to attach to the root card on top of the chassis.

“High Speed” rules have been used for the backplane design. Nevertheless, during the testing of the backplane significant damping around the connector area for the PCIe gen3 signals was discovered (see section 3.3.3). Because of this the DEEP-ER Booster prototype had to run with PCIe gen2 signals (5 Gbit/s per lane) to the peripherals. This affected mainly the speed of the EXTOLL network interconnect, while the impact to NVMe storage bandwidth was smaller.

Eurotech is building a product based on the DEEP-ER Booster prototype where compliance with PCIe gen3 signalling is essential. Physical instrumentation and computer modelling revealed a dominating effect of the stubs on the damping around the connector area (see also section 3.3.3). The remedy is to re-manufacture the affected PCBs to remove the stubs at the PCB factory equipped for this procedure. Prior to that, Eurotech is building a test board to confirm the results of the findings and verify the manufacturing rules going forward.

2.3.4 Aurora blade Root Card

The “Root Card” is the name of the chassis controller and peripheral card cage. It is a massive board on top of the chassis that hosts all peripherals and all control functions. In detail, it has:

- BMC (Pilot-3) for the low level control of power up and monitoring of the chassis
- Qseven COM (Computer on Module) with Linux OS for high level monitoring functions
- Ethernet switch (EtS) for the management network distributed over the backplane to the nodes and connecting the COM and the root card BMC.

- CPLD for the power up and low level control functions
- 36 PCIe gen3 ×16 standard connectors with hot swap controller and other support electronics
- Re-timers for PCIe signal shaping for each of the 36 PCIe slots (18 for TOURMALET NICs, 18 for NVM devices).

The hierarchical control functions allow safe transition through each step of the power up sequence and operation conditions. This is the sequence of CPLD-BMC-EtS-COM in bringing up the control, before it is safe to apply power to the nodes. The peripheral card power is controlled by the root card. This is important for the EXTOLL network, since the TOURMALET cards perform the routing even when the KNL nodes themselves are switched off.

There have been three versions of the root card design (LYNXvS, LYNXvA and LYNXvB). The last version (LYNXvB) makes use of higher quality dielectric material in the manufacturing and introduces some routing improvements to make the card suitable for PCIe gen3 signals

2.3.5 Eurotech Direct Liquid Cooling

Eurotech applies the direct water cooling principle to all of its products. Aluminium heat sinks have been designed, which cover full boards following the skyline of the electronics. The first designs adopted a massive aluminium plate with manufacturing drilling out tiny channels for the water to pass through the plate for the heat absorption. This technique was refined in the DEEP project with optimization of the drilling.

For the DEEP-ER Booster cooling Eurotech transitioned to the 2nd generation designs where the heat sink that covers the board is a passive aluminium blade (the “interposer”), while the water passing active cooling part is manufactured separately in a roll-bond process. The flat side of the roll-bond is firmly attached to the interposer in the assembly step, making the cold-plate that fully covers the board.

The cold-plate made in this fashion is lighter and easier to produce. This type of the cold-plate was used previously in another project (QPACE-2), but for DEEP-ER this type of the cold-plate has been tested on much bigger node cards. In addition to the heat extraction function, the cold-plate serves as stability back-bone for the node card assembly, which consists of several boards (S2700AP, CROW, LARK, OWLS) as discussed in Subsection 2.3.1.

Another part of the design covers the cooling arrangement for the ULP DIMMs. The principle of a single interposer covering the whole board cannot apply to the DIMMs, since these are placed vertically and therefore require another solution. A patented arrangement extracts the heat from the ULP DIMMs by conduction to the active part of the cold plate.

The technology developed to cool the peripheral cards has also been patented by Eurotech. It concerns the layout of the heat sink that covers the peripheral cards. The Eurotech-designed heat sink couples thermally and mechanically to the root card interposer. In this way the peripheral cards can be cooled from the active cold-plate of the root card, while the peripherals do not need their own water delivery.

Special attention was given to the EXTOLL TOURMALET card, since it has several hot spots (the ASIC and the DC/DC converters). A special arrangement was designed for this card to take care of the heat transfer from multiple places.

3 DEEP-ER Architecture and Design Review

This section assesses the architecture and design decisions as laid out in section 2 at the end of the DEEP-ER project, having the advantage of 20-20 hindsight. It is organized in sub- and subsubsections corresponding to the structure in Section 2, with each subsubsection covering the results achieved, obstacles overcome and lessons learned in the project.

3.1 CPU and Booster Node

3.1.1 KNL CPU

At the time of its selection as the Booster CPU, KNL was still in active development, with no working samples available. Intel passed technical information on to partner Eurotech as the OEM responsible for the Booster node as it became available, and the project at large was kept informed about the progress. Application analysis and initial optimizations could proceed nevertheless, using the DEEP Booster as a surrogate – the programming model of KNL is in essence an evolution of that for the earlier Intel Xeon Phi version, and essential optimization steps are the same. Thus, while the “CPU evaluator” planned in the DoW was late in becoming available, useful SW work was done, and impact to the project was kept at bay.

First KNL silicon (A0 stepping) was made available for remote access late in 2015, and after the somewhat controversial M18 review, Eurotech started a close collaboration with Intel’s hardware experts on the design of what became the Aurora Booster node board. Early A0 silicon and pre-release “Adams Pass” board versions were made available to Eurotech, and the Aurora Booster Node design evolved to its final shape (as discussed in sections 2.3 and 3.3), supported by an Intel OEM engagement team (not funded by the project).

First KNL software development platforms were delivered to the project in April of 2016, to be quickly integrated into the existing DEEP-ER SDV and validated for use of EXTOLL TOURMALET and the DC P3700 storage device. To give sufficient time for integration, validation and actual use of the DEEP-ER system by applications, a project extension for six additional months was proposed and accepted.

The KNL CPU version selected (Intel Xeon Phi 7210) supports 64 cores and a full 16 GByte complement of fast on-package MCDRAM; peak performance is approx. 2.7 TFlop/s. Reason for picking this version and not the top-of-the-line SKU was to assure supply for Eurotech to build the DEEP-ER system in late summer of 2016.

Porting of the system software to KNL was a very quick process, thanks to the support of standard Linux distributions and the full set of server functionality and interfaces provided by KNL and its PCH. Porting and optimization of the applications was likewise a rather smooth process. Both activities did make heavy use of the DEEP-ER SDV. Applications do show performance improvements compared to the DEEP Booster nodes of up to nearly a factor of 3×, the fast on-package memory helped to speed up memory-bound code parts, and the full 96 GByte of DRAM allowed realistic problem sizes to be run per node.

Looking back, the calculated risk of selecting a CPU that was under development has on one hand contributed to the need for a six-month extension, yet on the other hand did enable the project to use cutting-edge CPU technology and provide relevant per-node performance as well as memory capacity and flexibility for the DEEP-ER applications.

3.1.2 PCI Express Generation 3

The decision to select PCIe generation 3 was clearly vindicated by the measured MPI messaging performance with EXTOLL TOURMALET (see Section 3.2.1) and the achieved I/O bandwidths to the local storage device (see Section 3.2.2) on the DEEP-ER SDV. These performance figures could not have been achieved using PCIe gen2.

On the other hand, the substantial increase in operating frequencies brought by PCIe gen3 (60% compared to generation 2) does make the design of signal paths more challenging, and it requires tighter tolerances in manufacturing. This has impacted the development of the Aurora Blade architecture, which relies on PCIe gen3 ×16 connections across a backplane and a number of connectors. As discussed in section 3.3.4, the bring-up of gen3 speeds did not succeed in the term of DEEP-ER, and the DEEP-ER system will use PCIe gen2 speeds until Eurotech can address the technical issues, which have already been analysed. This is mainly a matter of time, with little apparent technical risks remaining.

In toto, there was no realistic alternative to using PCIe gen3 as the intra-node interconnect; in fact, first PCIe gen4 systems and devices are appearing on the market. The challenges in designing the long signal paths that blade systems do require are very real, and optical backplanes can be a solution in the future. For DEEP-ER, and for the foreseeable future, such optical backplanes do not exist in ready-to-use implementations, though.

3.2 System and Interconnect

3.2.1 EXTOLL TOURMALET Interconnect

The EXTOLL TOURMALET device was employed in its A3 revision on the C2 board revision, in all the different SDV nodes, fileserver nodes, KNL evaluation nodes and the final Aurora Blade Booster. TOURMALET A3 offers a fully functional PCIe gen3 link to the host, as well as 6 working 100Gbit/s links. While the ASIC offers a 7th link, electro-mechanical problems prevented them to being fully qualified for the project, so they were not used in DEEP-ER[§].

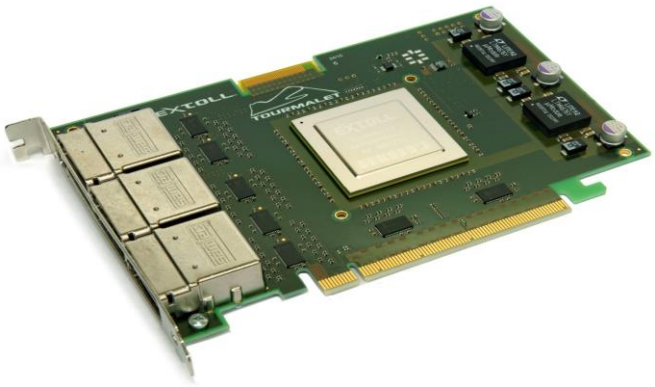

Figure 5: EXTOLL TOURMALET NIC as used for the DEEP-ER SDV and Booster.

[§] The size of a standard PCIe card bracket can only accommodate six EXTOLL links, although the connectors used are already the most compact cable connectors available.

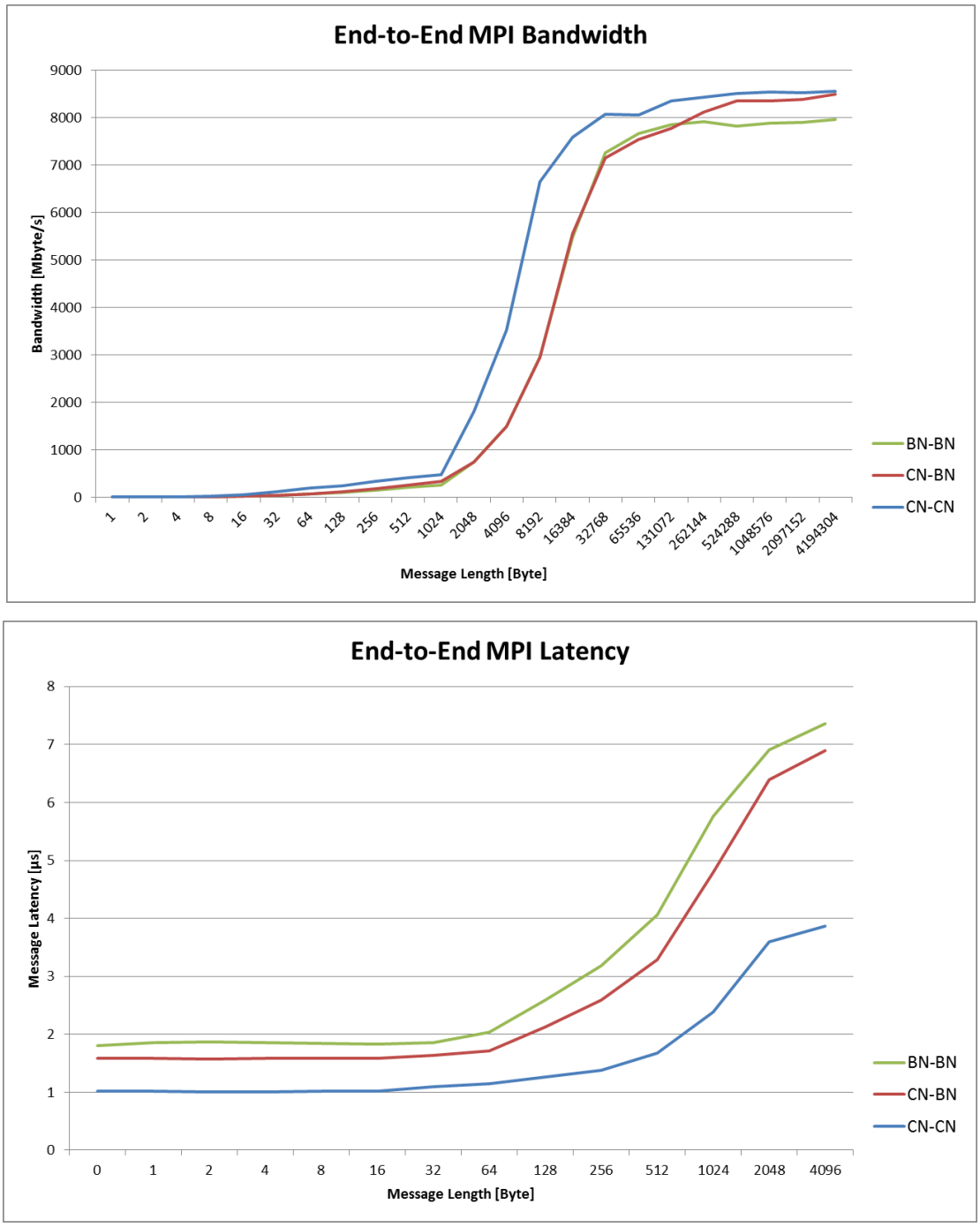

Figure 6: End-to-end MPI bandwidth (above) and latency (below) on the DEEP-ER SDV, using ParaStation MPI and the OSU benchmark.

The performance of the network was in general as expected (please refer to Figure 6). KNL nodes (BN) have a higher MPI latency compared to standard Xeon nodes (CN), which is caused by the lower single-thread performance of the Xeon Phi cores, with the memory subsystem being even more optimized for bandwidth instead of latency and the absence of a write combining buffer. Difference in maximum MPI bandwidth is smaller, since in both cases hardware functional units of the TOURMALET ASIC are used for remote put and get operations.

Within the project multiple improvements of the river and management software (EMP) could be developed and tested. This included major improvements in reliability and manageability,

which proved to be very useful when dealing with pre-production hardware, where frequent faults of different kinds occurred.

Also, the topology used in DEEP-ER, proved to be the most complex set-up of EXTOLL to date. While the 16 DEEP-ER SDV nodes for example were connected in a 4D Hypercube, this was combined with a 1D Torus (ring) of three nodes for the filesystem server nodes, a 2D Torus of 8 nodes for the initial KNL nodes and finally a hybrid mesh/torus for the actual DEEP-ER Booster. All of these sub-topologies were interconnected with each other. Implementing a correct routing set-up for this complex fabric in a general way proved to be a challenge, which could be solved and improves the overall features available from the EXTOLL software stack.

3.2.2 Intel NVM Storage-Class Memory Devices

The Intel DC P3700 devices were introduced into the project first as pre-release samples attached to two Intel Xeon servers. NVMe drivers were at that time already integrated with Linux distributions, and the devices could be quickly configured and validated. When the DEEP-ER SDV was built up, the 16 Intel Xeon nodes were configured with these devices. The later added eight KNL nodes also received attached NVM devices attached via PCIe ×4 cables attached to a PCIe riser card, since the physical space available in the 1U Intel KNL chassis was taken up by the full height EXTOLL TOURMALET card. For the KNL nodes, Linux drivers were contained in the Linux distribution, and operating the devices was straightforward.

Measured performance of the devices were fully within expectations -- Figure 7 shows sequential read and write numbers for the IOZONE benchmarks, compared to a best-of-breed SATA-attached SSD. A much more detailed discussion of the delivered performance for applications can be found in Deliverables D3.3 [9] and D6.3 [10].

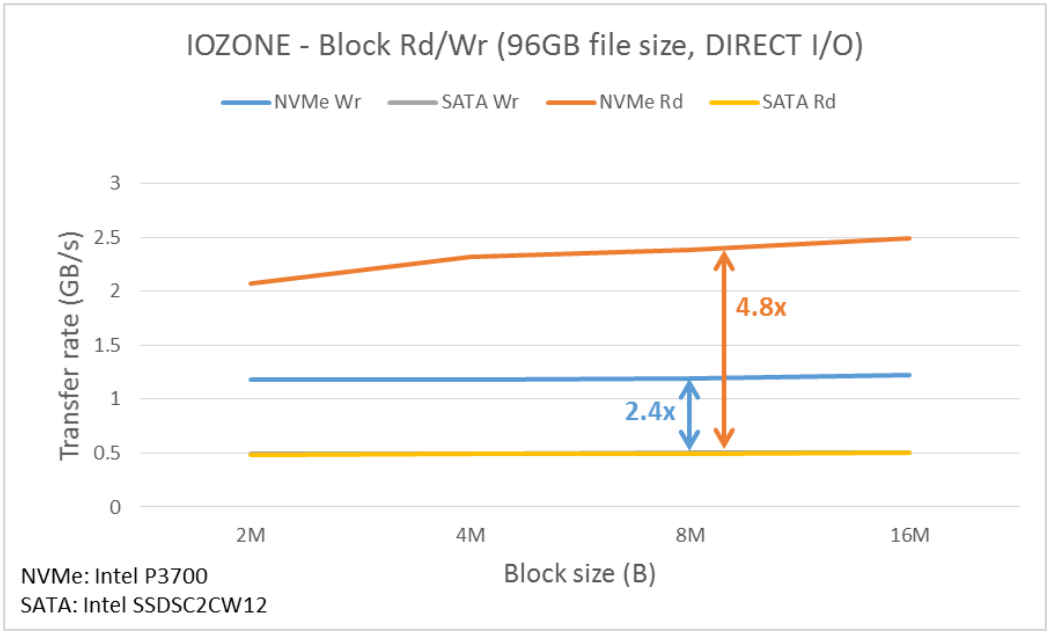

Figure 7: IOZONE sequential read/write performance for Intel DC P 3700 device attached via PCIe ×4 compared to SATA SSD[**].

[**] Please note that the read performance can certainly not be supported by a PCIe generation 2 ×4 link.

On the KNL nodes of the DEEP-ER SDV, an inconsistency in the FIO random read benchmarks for the NVMe device was recognized, which did fall short of the specified performance. This effect was not seen on the Intel Xeon nodes, and also not on the KNL nodes manufactured by Eurotech, leading to the conclusion that this would likely be related to using pre-release KNL boards in the SDV.

In toto, the selection of PCIe attachment and the NVMe protocol has worked out well, and the specific device did perform as expected. Being based on a standard interface, new generation of NVM devices (such as the Intel® Optane™ series [11]) can easily be used in future deployments of the DEEP-ER technology, keeping its I/O performance at the leading edge.

3.2.3 Network Attached Memory

The NAM prototype as integrated into the DEEP-ER prototype 9see D3.4 [12] for details) does achieve the design goals in full – it was possible to fit two EXTOLL links (100 Gbit/s each), the NAM functional unit and a HMC controller onto the selected FPGA. The NAM-specific functions provide basic memory registration and RDMA functionality for all nodes on the DEEP-ER prototype, as well as additional functionality for creating checkpoint parity data for up to 48 client nodes. Figure 8 shows the NAM node board and the installation of two NAMs in the DEEP-ER SDV.

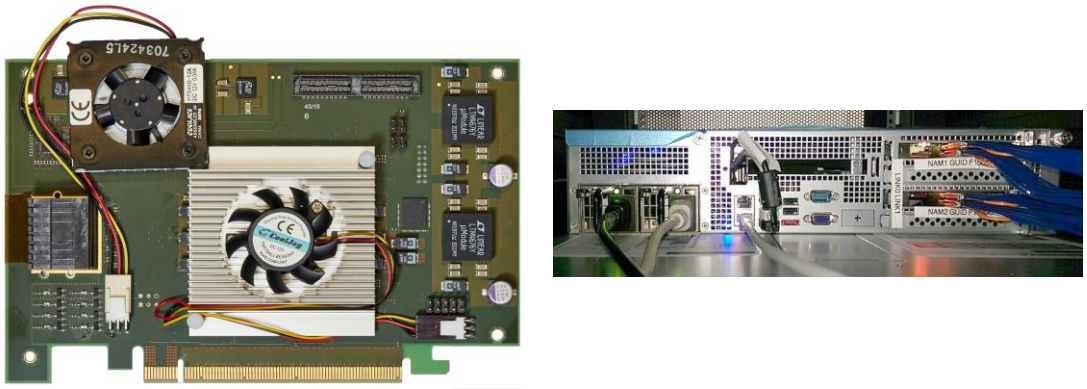

Figure 8: NAM Prototype board (left) and integration of two NAM boards into the DEEP-ER SDV (right).

Details about functionality and performance are given in an update to Deliverable D3.4 submitted at the end of the project. In its final form, the NAM can support 11 GByte/s write and 10.2 GByte/sec read bandwidth end-to-end from nodes on the EXTOLL fabric.

The checkpoint/restart use case has clearly validated the assumption that usage-specific functionality can indeed be integrated into a NAM in a way that delivers real benefit. Within the term of the project, it was not possible to explore farther reaching ideas, like for instance to offload collective MPI operations onto the NAM.

Regarding the choice of HMC memory technology, availability of larger capacity devices was delayed, and as a result, the NAM is using 2 GByte devices. The device daisy chaining functionality was declared as experimental and a corresponding connector was integrated onto the PCB. Unfortunately, chaining could not be tested due to time restrictions, and the NAM prototype in DEEP-ER is therefore limited to a total capacity of 2 GByte. While this is smaller than the project hoped for, it still allows to experiment with relevant use cases and allows validating the NAM concept. It should also be noted that the FPGA-based NAM

architecture can be adapted to use different memory technologies, should the evolution of HMC become stuck in the future, or should these other memory technologies offer substantial advantages.

An FPGA version of the EXTOLL link as implemented in the EXTOLL TOURMALET ASIC had to be developed for the NAM. Many of the existing EXTOLL link modules were redesigned to run at the four times wider data path of 512 bit, which was required to compensate for the lower operating frequency of FPGAs vs. ASICs. To keep up the total link bandwidth, more information must be processed internally in parallel. While the FPGA link itself is operational and fully verified, it still shows some limitations bandwidth-wise. EXTOLL related flow control features have been identified as one of the contributors and will require additional design action, which is out of the scope of the project.

The NAM checkpoint and restart architecture has been defined in close collaboration with WP4 and WP5 and is described in Deliverable D3.4.

3.3 Aurora Blade Chassis

3.3.1 Aurora Blade Node

The production of all 72 nodes for the DEEP-ER Booster (plus spares) has been completed. This entails the assembly of the S7200AP (including the processor and the memory) with the interface boards (CROW, LARK, etc.), all the necessary internal cabling and fixing the cold plate. All production tests have been run, which include the functionality testing to release the boards to further testing within the Aurora chassis.

The cooling capacity of the cold-plate has been tested using the LINPACK benchmark in a stand-alone setup. The results show that LINPACK runs at full performance for the inlet water temperature up to about T_{in}=45°C. The temperature of the processor cores is about 30-35°C higher than the inlet water temperature. Above 45°C and up to about 52°C the processor starts to throttle and reduce the run speed automatically in response to the increased temperature of the cores approaching 90°C. In all cases the benchmark completes successfully.

The power consumption is on average 335-340 W at T_{in}=25°C and 350-355 W at T_{in}=45°C.

The node has been tested in a test setup to address the PCIe end-point at gen3 ×16 correctly. The test setup included special instrumentation to mate with the Molex connector at the edge of the node going to the IDT re-timer board with the EXTOLL TOURMALET end-point.

When testing the node in the chassis problems were discovered with the reset to the peripheral card situated in the root card. The firmware of the pre-production S7200AP boards gives a reset spike within 1 ms after processor boot. This reset spike interferes with the load of the re-timer parameters and leaves the connection in an undefined state. This condition has been cured by programming the CPLD on the root card to block the early resets to allow the re-timers to complete the reading. This reprogramming became possible with the larger CPLDs that were implemented in the revision B of the root card (LYNXvB).

Even after the reset spike problem was corrected, the end-points on the Root Card could not be operated from the nodes in full PCIe gen3 x16 mode. The problems and the suggested remedies are discussed in Section 3.3.3 (Aurora Blade Backplane).

3.3.2 Ultra-Low Profile DIMM

There was no separate temperature measurement or monitoring of the ULP DIMMs on the node. From the results of the LINPACK run, where all memory was used by the benchmark, it was concluded that the memory cooling is working correctly.

3.3.3 Aurora Blade Backplane

The backplane was designed to route PCIe gen3 signals to the Root Card. Overall, it is part of the chassis assembly in which all elements must function correctly to achieve the desired results. With all elements of the chassis in place (node(s), backplane, root card, end-points) it was discovered that PCIe gen3 operation meets difficulties. Most frequent symptoms relate to the situation when the node finished auto-negotiation with the end-point for speed and gen3 was established, the protocol is restarted before or after traffic proceeds. Thus, PCIe gen3 operation is not stable.

To understand the source of the difficulty the test setup has been extended such as to isolate each individual element in the PCIe signal chain. In this process the conclusion was reached that all connector areas degraded the signal and the noise accumulates from element to element on the signal path. However, the area around the Amphenol Xcede connector on the backplane seems to give largest contribution to the signal degradation. This conclusion was reached after building up a test setup, in which the backplane with its connector was replaced with a cable (×8 width), all other conditions (node, root card, end-point) being the same. In this special setup the gen3 signalling was working and stable. Following this finding computer simulations were run (with help of Amphenol) that models the connector area. The result of that simulation confirms that the signal degradation is caused mainly by the connector stubs on the PCB. Other factors modelled (such as the size and shape of the anti-pads) did not appreciably affect the signal to noise ratio. Since the stubs are present with all connectors, it can be concluded that all connector areas generate noise on the signal, but main contribution is from the Amphenol Xcede backplane connector area.

In summary, to complete the setup of the chassis for PCIe gen3 signalling the backplane production has to be redone with the stubs removed. This procedure is offered by some PCB vendors and is planned as part of the manufacturing procedure for the stable Eurotech Aurora product.

3.3.4 Aurora blade Root Card

The latest version of the Root Card (LYNXvB) has the potential to support all 36 peripheral cards at gen3 speeds. As has been the case with the previous version (LYNXvA), most functionality works and the card is usable.

However, several manufacturing problems with the root card batch have been established. In particular, in three of the five produced cards the corner re-timer does not work. This points to problems in the manufacturing process (same position on all three boards). A corrective action is in place to replace this re-timer chip.

Other problems with the initial LYNXvB have been overcome with local repairs. All these flaws will be eliminated with next revision of the root card, leading to a stable product.

3.3.5 Eurotech Direct Liquid Cooling

Eurotech direct liquid cooling has been put to test when installed in the 1st chassis at Juelich. The stress tests running single node LINPACK on all nodes simultaneously have completed successfully. However, after one month of operations a leak has developed in several cold plates. The reason has been identified as pitting corrosion. As compared to the previous generation cold plates, the new cold plates have the liquid cavities with thinner walls. Therefore the corrosion may give way to a leak faster.

Multiple water analysis and consideration to establish the working condition led to the definition of the operational rules that allow resuming safe operation of the DEEP-ER Booster prototype. It is necessary to monitor the conductivity of the water. The conductivity of the water is in direct proportion to the amount of the chemical for corrosion protection (CLARIANT Protectogen® C Acqua) in the water. If conductivity falls below threshold Protectogen has to be added to counteract corrosion. Conductivity can be monitored with periodic sampling of the water and with a special instrument that will be installed in the cooling loop.

4 Conclusion

From the more detailed review in the previous sections, it is clear that the basic decisions taken in the DEEP-ER project on the system architecture are vindicated by the results achieved. The principal node architecture, choice of intra-node interconnect, of storage devices and inter-node fabric did all result giving the expected functionality, capacity and performance, as demonstrated by the small DEEP-ER SDV system. The risk taken by selecting EXTOLL TOURMALET before it was available with full performance did pay off, with the NICs becoming available in time for building the SDV and DEEP-ER Booster.

The benefits of closely integrating high-performance storage-class memory with compute nodes has become abundantly clear in the application results, and the DEEP-ER system architecture does that in an efficient way.

The Aurora Blade system design by Eurotech did hit some rocky patches in manufacturing, bring-up and validation. The Aurora node board assembly itself is fully functional, and the Backplane and Root Card combination is working, albeit with the significant performance restriction of not supporting PCIe generation 3 at this time. It could be said that the flexibility of the Aurora Blade approach (which can support several different compute cards and almost any PCIe-based peripheral and is essential for Eurotech to obtain a sustainable market position) has raised the technical bar significantly (longer and more complex PCIe signal paths), leading to the unsatisfactory situation at present. Eurotech has spared no effort to analyse the root cause (see for instance Section 2.3.3), and a plan is in place to address the technical issues and provide full PCIe gen3 connectivity and performance after the end of the project.

The other unexpected technical challenge occurred with the new and very efficient implementation of Eurotech's direct liquid cooling scheme; as discussed in Section 3.3.5, rapid pitting corrosion did cause leaks in the first Aurora Blade installation at Juelich. Again, the root cause was investigated in depth (high basicity of the cooling loop water), and more precise rules for preparing the cooling water and additional measures for monitoring its key characteristics were put in place, allowing the re-installation of the Aurora Blade system at the end of the project.

The NAM concept was successfully implemented as a working prototype, achieving the project objectives in full. The slow progress in larger capacity HMC devices becoming available was surprising, yet it does not take anything away from the project results, since the NAM architecture can be adapted to drive other memory devices.

5 References

1	Authors	Title	Year
1	M.Kauschke, P.Arts, N.Eicker, M.Watson, H.-Ch.Hoppe, M.Cintra	DEEP-ER Deliverable D3.1: Architecture Specification	2014
2	H.-Ch.Hoppe, P.Arts, M.Cintra, N Eicker, M.Nuessle, J.Schmidt, T.Wettig	DEEP-ER Deliverable D3.2: Component Design	2014
3	P.Arts, I.Zacharov	DEEP-ER Deliverable D8.1: Design of Aurora Blade prototype for DEEP-ER	2016
4	J.Schmidt, H.Ch.Hoppe, J.Gimenez, V.Beltran, G.Congiu, N.Eicker, C.Clauss, A.Galonska	DEEP-ER Deliverable D7.2: Report on projections and improvements for the DEEP/DEEP-ER concept	2017
5	I.Zacharov, M.Rossi	DEEP-ER Deliverable D8.2: Components of Aurora Blade Prototype for DEEP-ER	2016
6	I.Zacharov, M.Rossi	DEEP-ER Deliverable D8.3: Aurora Blade Booster Prototype for DEEP-ER	2017
7	A.Sodani	Knights Landing (KNL): 2nd Generation Intel® Xeon Phi™ Processor, Hot Chips 27 Symposium (HCS), 2015 IEEE, URL: http://www.hotchips.org/wp-content/uploads/hc_archives/hc27/HC27.25-Tuesday-Epub/HC27.25.70-Processors-Epub/HC27.25.710-Knights-Landing-Sodani-Intel.pdf	2015
8	N/A	Intel® Solid State Drive DC P3700 Series – Product Specification, document number 330566-010US, URL: : http://www.intel.com/content/dam/www/public/us/en/documents/product-specifications/ssd-dc-p3700-spec.pdf, alternate URL: http://www.mouser.com/ds/2/612/ssd-dc-p3700-spec-769407.pdf	2015
9	N/A	Intel® Server Board S7200AP Product Brief, document number 334540-002, http://www.intel.com/content/dam/www/public/us/en/documents/	2016

		product-briefs/server-board-s7200ap-brief.pdf	
10	M.Cintra, H.-Ch.Hoppe	DEEP-ER Deliverable D3.3: Non-Volatile Memory (NVM) assessment	2016
11	A.Zitz, J.Morillo, R.Leger, S.Solbrig, S.Rodriguez, M.Petschow, A.Emerson, F.Affinito, G.Brietzke, J.Amaya, D.Gonzalez	DEEP-ER Deliverable D6.3: Final report on application experience	2017
12	N/A	Intel® Optane™ SSD DC P4800X Series Product Brief, Intel document number 335696-002, URL: http://www.intel.com/content/dam/www/public/us/en/documents/product-briefs/optane-ssd-dc-p4800x-brief.pdf	2017
13	J.Schmidt, U.Bruening	DEEP-ER Deliverable D3.4: Network Attached Memory (NAM)	2016

List of Acronyms and Abbreviations

A

AP: Adams Pass. Intel implementation of the KNL reference board
API: Application Programming Interface
ASIC: Application Specific Integrated Circuit, Integrated circuit customised for a particular use
Aurora: The name of Eurotech's cluster systems
AVX: Extensions to the x86 instruction set architecture for microprocessors from Intel and AMD covering SIMD/vector instructions up to 128 bit length
AVX512: Intel instruction set extension introducing 512 bit SIMD support and oher improvements; implemented by second generation Intel Xeon Phi CPUs (KNL) as used in the DEEP-ER Booster

B

BMC: Board management controller. Used to monitor and manage a compute blade.
BN: Booster Node (functional entity); refers to a self-booting KNL board (Node board architecture) including the NVM and NIC devices connected by PCI Express or a Brick (Brick Architecture).
BoP: Board of Partners for the DEEP-ER project
Brick: Modular entity forming a Booster Node in the Brick Architecture, composed of Host modules, NVMe and NIC devices all connected by an PCI Express switch.
Brick Architecture: Two-level hierarchical architecture for the DEEP-ER Booster, based on the "Brick" element as a Booster node.
BSC: Barcelona Supercomputing Centre, Spain

C

Chassis: Mechanical entity mounted in a rack. A chassis typically aggregates multiple mechanical sub-units (here: Bricks) through a chassis level infrastructure (e.g. backplane, power, cooling)
CN: Cluster Node (functional entity)
COM: Computer on-module: form factor for single-board computers
CPLD: Complex programmable logic device: used f.i. for controlling the power-on and pre-boot phases of computer systems.
CPU: Central Processing Unit
CROW: Eurotech interface board, part of the Aurora KNL node for DEEP-ER.

D

DDR-4: Interface standard to attach DRAM to a CPU
DEEP: Dynamical Exascale Entry Platform
DEEP-ER: DEEP Extended Reach: this project

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

DEEP-ER Global Network: High performance network connecting Bricks, CN, NAM and other global resources to form the DEEP-ER prototype system

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

DEEP-ER Network: High performance network connecting the DEEP-ER BN, CN and NAM; to be selected off the shelf at the start of DEEP-ER

DEEP-ER Prototype: Demonstrator system for the extended DEEP Architecture, based on second generation Intel® Xeon PhiTM CPUs, connecting BN and CN via a single, uniform network and introducing NVM and NAM resources for parallel I/O and multi-level checkpointing

DEEP Architecture: Functional architecture of DEEP (e.g. concept of an integrated Cluster Booster Architecture), to be extended in the DEEP-ER project

DEEP System: The prototype machine based on the DEEP Architecture developed and installed by the DEEP project

DoW: Description of Work

DRAM: Dynamic Random Access Memory. Typically describes any form of high capacity volatile memory attached to a CPU

E

EC: European Commission

EDR: Enhanced Data Rate: InfiniBand implementation that delivers 100Gbit/s per link; one of the candidates for the DEEP-ER inter-node interconnect

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

FDR: Full data rate: InfiniBand implementation that delivers 50 Gbit/s per link; used for the DEEP system

FIO: Flexible I/O Tester: set of synthetic benchmarks to measure I/O performance

FLOP: Floating point Operation

FP7: European Commission 7th Framework Programme.

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

FWI: Full Waveform Inversion: advanced technique to compute subsurface sound velocity fields from seismic experiment data.

H

HDR: High data rate: InfiniBand implementation by Mellanox that delivers 200 Gbit/s per link; announced in 2016.

HMC: Hybrid Memory Cube

Host Module: Self-booting computer board with an Intel KNL CPU, DDR4 memory and a PCI Express root complex. In the Brick Architecture, multiple host modules are connected to each other and NVMe and NIC devices by a PCI Express switch. In the Node Board architecture, a host module provides PCI Express slots to plug NVMe and NIC devices in.

HPC: High Performance Computing

HW: Hardware

Hybrid Memory Cube: Novel type of computer RAM that uses 3D packaging of multiple memory dies to increase memory capacity and number of data banks per device area. Technology is being developed by Micron Technology and backed by the Hybrid Memory Cube Consortium.

Hybrid Memory Cube Consortium: Industry association defining HMC interfaces and facilitating HMC Integration into a wide variety of systems. Includes Samsung, Micron Technology, Open-Silicon, ARM, IBM, SK-Hynix, Altera, and Xilinx.

I

I2C: Inter-Integrated Circuit bus. A low cost serial bus used to interconnect silicon devices. Typically used for status monitoring and configuration.

IB: InfiniBand

IBM: International Business Machines Corporation

ICT: Information and Communication Technologies

IEEE: Institute of Electrical and Electronics Engineers

InfiniBand: Computer-networking communications standard mainly used in high-performance computing; available implementations include FDR, EDR and HDR (announced in 2016).

Intel: Intel Germany GmbH Feldkirchen,

I/O: Input/Output. May describe the respective logical function of a computer system or a certain physical instantiation

J

JUELICH: Forschungszentrum Jülich GmbH, Jülich, Germany

K

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

L

LARK: Eurotech interface board, part of the Aurora KNL node for DEEP-ER.

LINPACK: Software library to perform numerical linear algebra calculations used as benchmark
LYNXvA: First version of the Aurora Blade Root Card.
LYNXvB: Second versiin of the Aurora Blade Root Card.
LYNXvS: First prototype of the Aurora Blade Root Card.

M

MCDRAM: Multi-Channel DRAM, a high bandwidth on package memory on KNL
MPI: Message Passing Interface, API specification typically used in parallel programs that allows processes to communicate with one another by sending and receiving messages

N

NAM: Network Attached Memory, nodes connected to the DEEP-ER network providing shared memory and special-purpose processing to the DEEP-ER DEEP-ER Booster and Cluster nodes.
NAND Flash: Implementation of non-volatile memory used today for solid state disks.
NIC: Network Interface Controller, Hardware component that connects a computer to a computer network
NVM: Non-Volatile Memory. Used to describe a physical technology or the use of such technology in a non-block-oriented way in a computer system
NVMe: Short form of NVM-Express
NVM Express: interface standard for attaching storage via PCie; also specifies high-level HW interfaces like queues.

O

OEM: Original Equipment Manufacturer. Term used for a company that commercialises products out of components delivered by other companies.
OS: Operating System
OWLS: Eurotech interface board, part of the Aurora KNL node for DEEP-ER.

P

PA: Physical address space. Used on hardware level to access system components
ParaStation MPI: Software for cluster management and control developed by ParTec
ParTec: ParTec Cluster Competence Center GmbH, Munich, Germany
PCB: Printed circuit board.
PCH: Platform controller hub: companion device to operate Intel CPUs and attach commodity peripherals
PCI: Peripheral Component Interconnect, Computer bus for attaching hardware devices in a computer
PCIe: Short form of PCI Express
PCI Express: Peripheral Component Interconnect Express started as an option for a physical layer of PCI using high-performance serial communication. It is

today's standard interface for communication with add-on cards and on-board devices, and makes inroads into coupling of host systems. PCI Express has taken over specifications of higher layers from the PCI baseline specification.

PMT: Project Management Team of the DEEP-ER project

Project Coordinator: Leading scientist coordinating and representing the DEEP-ER project

Q

QCD: Quantum Chromodynamics

QPACE: QCD Parallel Computing Engine. Specialised supercomputer for QCD Parallel Computing

QPACE-2: Successor system to QPACE using Intel Xeon Phi co-processors.

R

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

RAM: Random-Access Memory

RDMA: Remote Direct Memory Access

S

SAS: Serial attached SCSI: point-to-point serial interface for high-performance storage devices commonly used in server computers; covers speeds up to 1500 MByte/s.

SATA: Serial AT Attachment: point-to-point serial interface for storage devices commonly used in desktop and portable computers (with speeds up to 600 MByte/s); latest version covers up to 2000 Mbyte/s, competing with NVMe.

SDV: Software Development Vehicle: in DEEP-ER, an interim system with 16 Cluster and eight Booster nodes plus three file server nodes used for SW development and benchmarking in DEEP-ER.

SIMD: Single Instruction Multiple Data

SME: Small and Medium Enterprise

SSD: Solid State Disk

SW: Software

T

TFlop/s: Teraflop, 10^{12} Floating point operations per second

U

UHEI: University of Heidelberg, Germany

ULP: Ultra-low profile: new standard for DIMMs with a maximum height of 17.75 mm; ULP DIMMs are used on the Aurora Blade KNL nodes

W

WP: Work Package

X

x86: Family of instruction set architectures based on the Intel 8086 CPU