SEVENTH FRAMEWORK PROGRAMME

FP7-ICT-2013-10

DEEP-ER

DEEP Extended Reach

Grant Agreement Number: 610476

D7.1

Report on performance extrapolation based on design decisions

Approved

Version: 2.0

Author(s): N.Eicker (JUELICH)

Contributor(s): H.Ch. Hoppe (Intel), J.Gimenez (BSC), E.Suarez (JUELICH), I.Zacharov (Eurotech)

Date: 09.12.2015

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: D7.1</td></tr>
<tr><td colspan="2">Deliverable Nature: Report</td></tr>
<tr><td>Deliverable Level:
PU*</td><td>Contractual Date of Delivery:
30 / September / 2015</td></tr>
<tr><td>Actual Date of Delivery:
30 / September / 2015</td></tr>
<tr><td colspan="2">EC Project Officer: Panagiotis Tsarchopoulos</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="5">Document</td><td colspan="2">Title: Report on performance extrapolation based on design decisions</td></tr>
<tr><td colspan="2">ID: D7.1</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="3">File(s): DEEP-ER_D7.1_Report_performance_extrapolation_based_on_design_decisions_v2.0-ECapproved</td></tr>
<tr><td rowspan="4">Authorship</td><td>Written by:</td><td>N.Eicker (JUELICH)</td></tr>
<tr><td>Contributors:</td><td>H.Ch. Hoppe (Intel), J.Gimenez (BSC), E.Suarez (JUELICH), I.Zacharov (Eurotech)</td></tr>
<tr><td>Reviewed by:</td><td>S. Narasimhamurthy (Seagate), E.Suarez (JUELICH)</td></tr>
<tr><td>Approved by:</td><td>BoP/PMT</td></tr>
</table>

Document Status Sheet

Version	Date	Status	Comments
1.0	30/September/2015	Final	EC submission
2.0	09/December/2015	Approved	EC approved

Document Keywords

Keywords:	DEEP-ER, HPC, Exascale, design, extrapolation, modelling, Dimemas

Table of Contents

List of Figures

List of Tables

Executive Summary

Eurotech's Aurora Blade architecture has been chosen for the implementation of the DEEP-ER Prototype at the M18 review. In addition the core components of the DEEP-ER Prototype (processors, interconnect, memory) have been selected, and the methods to integrate them into the prototype system were fixed before M24. This Deliverable investigates the design decisions in light of their effect on system scalability, performance, and energy efficiency, extrapolating application results on the DEEP system.

The results of this investigation are clear – the DEEP-ER Prototype system design will bring substantial improvements in compute & interconnect speeds, available memory and memory bandwidth (due to the introduction of on-package memory). The use of PCI Express generation 3 (16 lanes) to the NVM devices and the EXTOLL TOURMALET NICs does not constitute a bottleneck, and the first performance predictions for DEEP/DEEP-ER applications show a better balance of the DEEP-ER system compared to its DEEP predecessor. These predictions are based on conservative assumptions, in particular regarding the end-to-end communication performance of the selected interconnect, and we are confident that the actual DEEP-ER prototype will in fact perform better than the current results do indicate.

1 Introduction

An important objective of the DEEP-ER project is the design and construction of a system prototype that implements the Cluster-Booster architecture using the latest processor technology, a uniform EXTOLL interconnect, and integrating non-volatile memory to add a level to the memory hierarchy that supports scalable I/O and resiliency mechanisms. The concrete physical realisation of the DEEP-ER Prototype will use Eurotech's Aurora Blade architecture, as decided in the interim review at M18.

The core components of the DEEP-ER Aurora Blade Prototype (processors, interconnect, memory) have been selected and the overall design of the platform and the compute, memory and backplane boards has already been fixed. These design decisions are described in detail in Deliverable D8.1. The present deliverable (D7.1) investigates the main design decisions taken and assesses their effects on system scalability, energy efficiency, and expected application performance.

Section 2 discusses the most relevant design aspects of the DEEP-ER Aurora Blade Prototype. Based on them, an analysis of their relevant consequences – when thinking about realising the DEEP-ER ideas at large scale and about the future development of the concept – is performed. Section 3 compares the Aurora Blade Prototype with the DEEP Booster and describes how the DEEP-ER Prototype design was defined taking the co-design feedback gathered within DEEP and DEEP-ER into account. Additionally, DEEP-ER profits from advanced HPC technology that was not available at the time the DEEP Booster was designed. Building on the results obtained within the DEEP project, Section 4 analyses three scientific applications: the MAXW-DGTD code from DEEP-ER partner Inria, and two DEEP applications, one from EPFL for brain simulation and a CFD code from CERFACS. The first two applications are part of the DEEP portfolio and have been thoroughly analysed in the DEEP project using the performance tools Extrae/Paraver from BSC, and modelled with BSC's Dimemas simulator. Taking into account the performance advantage provided by the improved many-core processor and interconnect used in DEEP-ER, the expected efficiency and scaling of the applications on the DEEP-ER platform has been predicted. Finally, Section 0 summarises the conclusions of the analysis performed for D7.1.

2 Main design decisions

The design of DEEP-ER Booster – as in the DEEP "Cluster-Booster" concept – follows the Eurotech Aurora Blade Architecture form factor and design rules with the specific addition of a unique dual-socket KNL blade. "Knights Landing" (KNL) is the code name for the second generation Intel Xeon Phi processor; the current Intel Xeon Phi version (" Knights Corner" or KNC) is used in the DEEP Booster. Usage of the KNL is one of the project requirements; other requirements are summarized in Table 1.

Table 1. DEEP-ER Project Requirements.

Item	**Value/Qty**[1]	**Aurora Blades**
Bootable KNL nodes	3+ TFlop/s DP per CPU	2 KNL nodes per Aurora blade, 3.5 TFlop/s per KNL theoretical peak performance
DDR4 Memory	96 GB 120 GB/s	Up to 384 GB DDR4 memory per KNL socket, using 6 DDR4 channels per KNL socket
Network bandwidth	>100 Gb/s	Integrate EXTOLL TOURMALET fabric; Connected to the KNL blades via PCIe gen3 x16
NVM usage	Every node	Integrate Intel DC P3700 NVMe SSD device (400 GB/board)
NAM infrastructure	≥2 devices	Integrate NAM prototype into the peripherals design
Commercial viability	RC	Aurora Blade Architecture is developed as a general purpose commercially available machine
No COTS boards	RC	The KNL Blade design is a unique contribution from Eurotech to the DEEP-ER project
Focus on density and liquid cooling	RC	Aurora: 18 Blades/7U chassis (max density)

An intensive co-design effort has been invested in the DEEP-ER project to accommodate the system software and applications requirements in the design of the DEEP-ER Prototype. The seven DEEP-ER applications are important codes for science and industry (see D6.1). Within the project they are being optimised and tuned for the "Cluster-Booster" architecture, and specifically for the I/O and resiliency extensions developed in DEEP-ER. With respect to DEEP, larger modelling significance can be achieved in the DEEP-ER project thanks to the increase in available processor performance above 3TFlop/s, supported by the introduction of on-package memory with very high bandwidth. The analysis of the applications showed that other parameters of the architecture must be increased as well. Specifically, at least 96 GB of memory is required for each KNL processor with memory access bandwidth of >100 GB/s. Furthermore, early application development and benchmarking has emphasized the need for a balance between the processor floating-point performance and interconnect speed, resulting in the bandwidth requirement of at least 100 Gbit/s per link.

[1] RC stands for Recommendation by the M12 review. The third column summarizes how the Aurora Blade Architecture satisfies the respective requirements.

The feasibility study for the DEEP-ER Prototype did evaluate a number of different architectures that could fulfil the project requirements. In particular, the “Brick” Architecture based on PCIe addin card form factor boards was compared with several alternatives. The resulting conclusion endorsed by the EC reviewers and adopted by the project calls for design and delivery of the Aurora Blade Prototype. Deliverable D8.1 describes the design of the Aurora Blade architecture in detail.

The KNL-based Aurora Blade Architecture addresses the DEEP-ER project requirements as shown in the right column of Table 1. To summarize, in a 19” chassis up to 18 boards with dimensions of 500x175 mm connect to a PCIe backplane, which provides connectivity between computing and peripheral resources. The backplane is designed following the 100 Gbits/s rules to carry the PCI Express gen-3 signals as previously indicated. The peripheral cards are adjacent to the KNL boards or placed in a card cage -- also called “Rootcard” – on top of the KNL boards, all connected through the backplane. The Rootcard also contains the power control infrastructure for all boards, the Ethernet switch for a 1GigE management and monitoring network, and a front-end management node for boot control and monitoring.

A shown in Figure 1, it has been decided to integrate the NVMe storage in the slots adjacent to the KNL boards, while the EXTOLL TOURMALET cards are integrated in the Rootcard on top of the chassis. Therefore, the height of each chassis is 7U (4U for KNL board plus 3U TOURMALET board). With this implementation, 18 KNL nodes will fit in 7U (310 mm) and two chassis back-to-back in a rack. The resulting density is 5 KNL/U, which is 2 times higher than the Intel reference design. All boards are liquid cooled with Aurora Liquid Cooled Technology (see section 2.4).

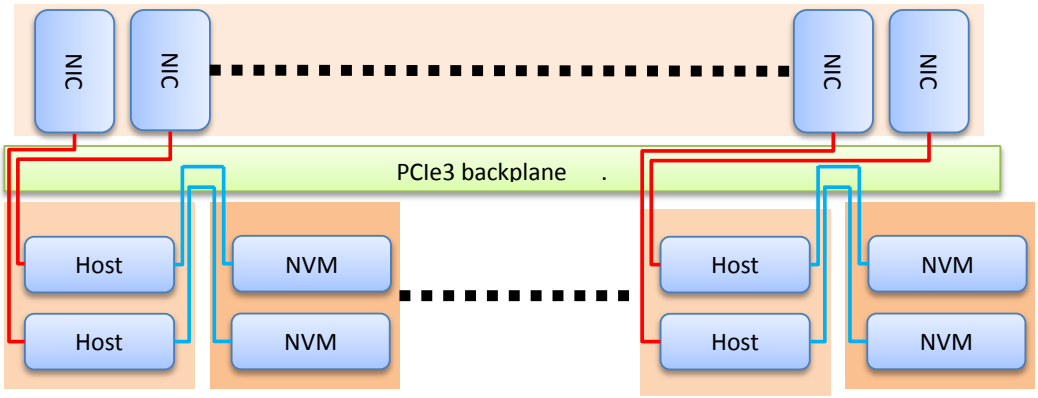

Figure 1: DEEP-ER Booster concept based on the Aurora Blade Architecture.

The Aurora Blade Architecture is also used to design general purpose blade servers with Intel® Xeon® processors and GPU cards. The addition of the KNL boards to the family of blade servers may open up commercial opportunities in this specialized space. In particular, as the KNL and general purpose processor boards are interchangeable in the blade chassis, the same architecture can integrate both the Cluster and Booster parts of a “Cluster-Booster” system).

Therefore, the decision to adopt the Aurora Blade Architecture by the DEEP-ER project and its endorsement by the European Commission opens up the possibility of wider use of European developed technology beyond the immediate needs of the DEEP-ER project.

2.1 KNL CPU and Blade design

The main characteristics of the KNL processor (next generation Intel Xeon Phi) are given in Table 2 from recently published material [1]. The KNL design specifications lead to the theoretical efficiency of ≥10 GFlops/W, the highest in the industry at this time.

Table 2. KNL publicly available design characteristics.

Characteristic	Qty	Comment
Double Precision performance [TFlop/s]	≥ 3	Peak DP performance
Number of cores per socket	≤ 72	Cores use 2D Mesh interconnect
On package memory MCDRAM [GB]	8 or 16	On-package high speed memory for increased bandwidth; STREAMS Triad numbers
Access speed to MCDRAM [GB/s]	≥ 400	
Support for off-package RAM [GB]	384	Enables applications with large memory footprint
Access speed to DDR4 RAM [GB/s]	120	6 channels to DDR4 memory
TDP [W]	~ 250	Similar to KNC TDP
Processor operation	Self-booting	Using a server-class chipset
Processor ISA	X86-64	With 512 bit vector extensions (AVX-512)
PCI Express links, generation 3	36	2 x 16 plus 1 x 4

KNL, the second generation Intel Xeon Phi processor works together with the server-class “Wellsburg” chipset. In contrast, the first Intel Xeon Phi generation (KNC) as used in the DEEP Project exists only as an accelerator card and needs an additional Intel Xeon processor for its operation. The requirement of the DEEP-ER project for a self-booting processor has been addressed by the Eurotech KNL board design using the Wellsburg chipset. Intel provided the reference design (customer reference board or CRB) and detailed design review support for this project. The following are the main design characteristics of the Aurora KNL board:

- The Aurora Blade Architecture size boards can host two KNL reference design instances. Therefore, each Aurora KNL blade provides two compute nodes (two KNL sockets, chipsets, memory etc.) to the system.
- An Aurora Blade has sufficient PCB real estate to provide 6 channels of DDR4 memory per KNL socket with soldered down devices, the use of which is necessary for applying the Eurotech liquid cooling technology to the boards. The kind of memory selected is the main difference between the Aurora Blade design and the Intel reference board.
- A Board Management Controller (BMC) manages each KNL node, which is responsible for the initial power up and for monitoring of the working node. This same BMC concept is used in the general purpose Aurora designs and also in the DEEP project. Eurotech extends this concept to DEEP-ER with partial reuse of firmware developed for the DEEP project. There are hence two BMCs per Aurora KNL blade.
- The Aurora KNL board also hosts a 1GigE Ethernet control infrastructure with connections to the front panel and to the backplane. In regular operations, the 1GigE

backplane connections are used for management and monitoring, combining the BMC and KNL physical Ethernet networks.

- Each KNL node has access to a mSATA SSD, which can be used to boot the system and/or for local storage on the node. In the DEEP-ER project, it is foreseen that the KNL nodes will be booted from the 1GigE control network over the backplane, amongst other available booting choices.
- The KNL PCIe ports of each socket are connected to two PCIe gen 3 x16 lines on the backplane. One port is connected to PCIe infrastructure routing to the adjacent slot in the node card cage, the other port is routed to the PCIe slot in the peripherals card cage (Rootcard).
- The KNL board will also be able to accommodate a future KNL-F variant with integrated Intel® Omni-Path support on the CPU package. The Omni-Path connectors will be routed to the front panel of the card. The Omi Path NIC uses two PCIe gen 3 x16 ports, and the remaining KNL PCIe ×4 port will be connected to the backplane and can be used for the storage infrastructure if required.

An impression of the proposed Aurora KNL board schematics is shown in Figure 2.

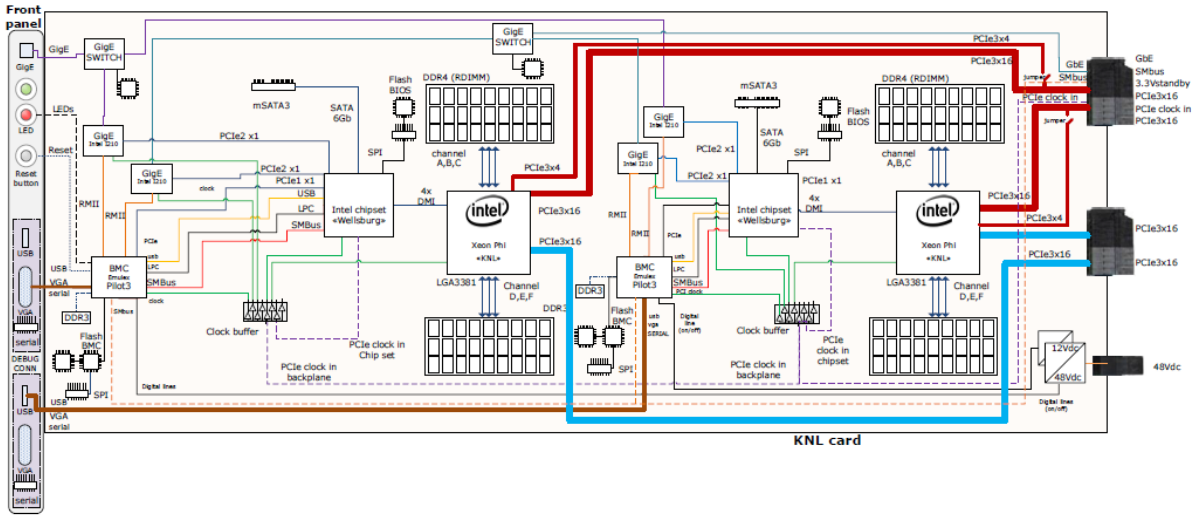

Figure 2. Aurora KNL blade schematics[2].

The Aurora Processor board is a unique development for each processor generation (Intel "Tock"), since each new micro-architecture usually defines a new socket. The "Tick", i.e. a feature shrink in the process technology can utilize the same processor board design. Aurora design rules decouple the development of the processor boards from the development of the peripheral devices, which follow their own development cycle. The PCI-Express backplane facilitates this decoupling. Current 3rd generation PCI Express will continue to be state-of-the art until at least 2018, when we expect PCI Express of the 4th generation to be adopted or some other interconnect technology to become mainstream. In this time window, we expect the Aurora infrastructure to host up to 3 Intel "Ticks/Tocks" and several generations of Network interconnect devices. Similarly, several generations of the PCIe storage devices can be integrated in the Aurora machine, following the independent PCIe Storage development cycle.

[2] KNL PCIe port1 (fat red line) goes to the upper PCIe backplane connector, routing to the peripherals card cage; port2 (fat blue line) goes to backplane routing to the adjacent slot.

Intel has announced that the third product generation of Xeon Phi will be called “Knights Landing” or KNH, and is scheduled for approx. 2018 availability. KNH will be a self-booting processor, and the investments in the Aurora Blade architecture could be leveraged for that processor with the design of a KNH blade.

2.2 NVM devices

After a survey of available options for bringing non-volatile memory (NVM) close to the DEEP-ER compute nodes, it was decided to use SSD replacement devices that implement the NVM Express (NVMe) protocol and are attached via PCIe links rather than SATA connections. Currently, Intel has a range of products available, and the project has picked the DC P3700 series [2] as the initial NVM device. This product uses 4 PCIe generation 3 lanes and provides between 400 GB and 2 TB of data capacity. Physical form factor is that of a half-height PCIe add-in card, which fits nicely into the Aurora Blade architecture.

Raw performance characteristics of the 400 GB version selected are 2.8 GB/s read and 2 GB/s write bandwidth, and a read and write latency of 20 µs. Measurements with synthetic I/O benchmarks and models for some DEEP/DEEP-ER applications have shown substantial performance increases versus best-of-breed conventional SSD storage.

The integration of NVMe storage devices is based on the Expansion blade designed by Eurotech outside of the DEEP-ER project. The board extends the (proprietary) backplane PCIe connections from adjacent slot to standard PCIe connectors suitable for 3rd party PCIe cards and provides the necessary 12V DC power and control infrastructure.

Each KNL port2 (see Figure 2) from the adjacent board is routed through the backplane to one of the ports on the expansion card next to it. The expansion card maintains signal quality through the re-timers to feed riser cards that provide standard PCIe connectors. The default width of the PCIe connector is x16; narrower cards are naturally accommodated.

The mechanical design for the integration of Intel DC P3700 NVMe storage cards is shown in Figure 3. The width of two cards together is 138 mm, less than the available space of 175 mm. The front plate of DC P3700 does not include any functionality and will be removed.

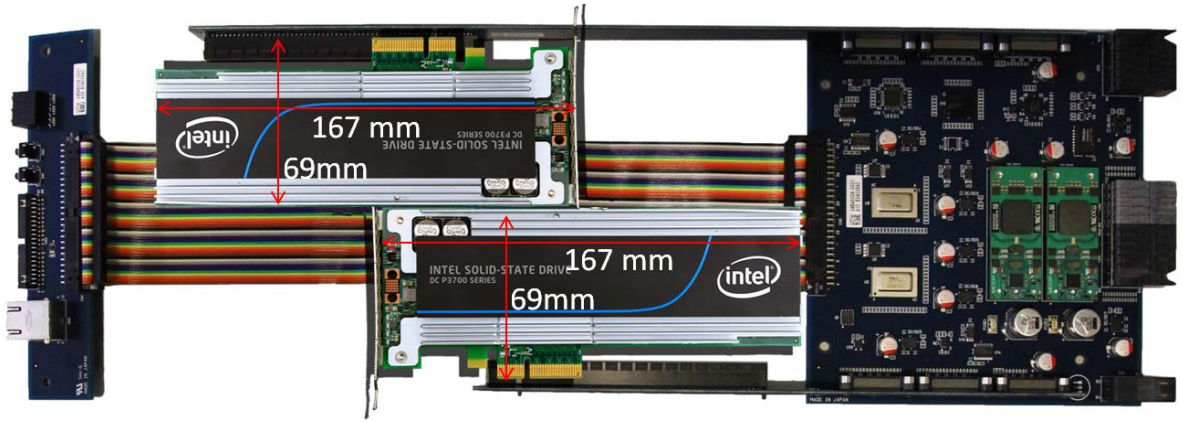

Figure 3: Intel DC P3700 NVMe cards mounted on Aurora extension blade.

An Aurora liquid cooled heat spreader covers all cards in this assembly in a single construction that also maintains the rigidity (see Figure 7).

In the DEEP-ER Prototype, the chassis is populated with 9 KNL blades (18 Booster nodes) and 9 NVMe blades (18 cards DC P3700), for a total of 18 blade positions.

Intel's recent announcement [3] (together with Micron) of the 3D XPoint™ technology does show the next steps in the NVM roadmap. Intel® Optane™ technology will consist of NVMe devices (connected via PCIe) with further reduced latency and substantially increased bandwidth in the 2016 timeframe and of NVM devices that are compatible with the DDR4 form factor and electrics (Intel® DIMM) in the future.

The Aurora Expansion Blade concept can accommodate Optane and other future NVM storage technologies. Depending on the exact timing of availability, it might even be possible to intercept the DEEP-ER prototype with this technology.

2.3 EXTOLL TOURMALET Interconnect

The network chosen by the DEEP-ER project is based on the EXTOLL TOURMALET technology, which combines the NIC and Router functionality in a single ASIC. The DEEP Booster uses the EXTOLL protocol implemented in an FPGA, and the ASIC implementation is capable of significantly higher speeds. The main characteristics of EXTOLL TOURMALET are summarized in Table 3.

Table 3. Main characteristics of the EXTOLL TOURMALET card.

Characteristics	Qty	Comments
Bandwidth per link [Gbit/s]	100	60Gbit/s, 100Gbit/s under qualification
Number of links (ports)	7	7th port for I/O at 60 Gbit/s
Switch Latency [ns]	60	@ 750MHz core clock frequency
Topologies supported with 6 links		Topology independent: 3D Torus, Xbar is, Fat Tree, etc.
TDP [W]	25	Max power when all links are active
MPI half round trip latency [ns]	600	2 switch layers included
MPI message rate [Mio messages]	100	Dependent on CPU performance and core count

The card with ASIC chip is shown in Figure 4. It is a PCIe generation 3 card with a x16 wide connector to the host and 6 copper connectors for the network.

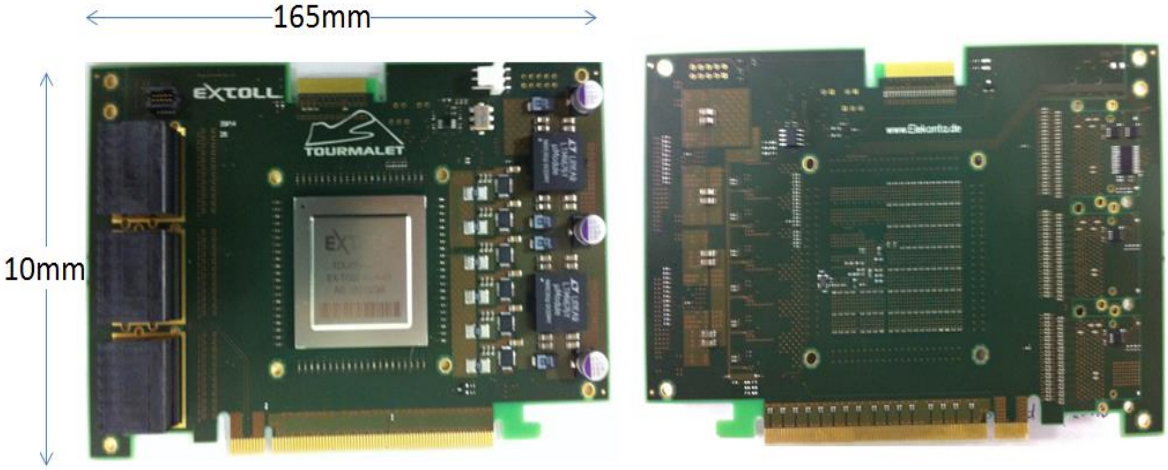

Figure 4. EXTOLL TOURMALET card with 6 connectors for Torus network topology. The 7th link is on top.

The EXTOLL TOURMALET card is a very recent development from EXTOLL GmbH. Currently, stepping A2 of the TOURMALET ASIC is available on the ref C PCB and is

capable of 8 Gbit/s on the PCIe link side (PCIe gen3 x16 interface). All the ASIC digital functions are working as expected at the full internal clock speed of 750 MHz , as well as all software and drivers. The A2 stepping showed significant link speed improvements in the analogue high speed SERDES module, but there are some issues in the stability of 8 Gbit/s transfer rates which are currently under evaluation. The 5 Gbit/s transfer speed has been fully verified. It is planned to tune the links with the added CTLE driver functions and a stable 8 Gbit/s version is expected in Q1 2016 or earlier, depending on the measurement results. Compared to the original specification which was presented in the project description, only the link bandwidth was reduced from 120 Gbit/s to 60 Gbit/s in the 5 Gbit/s version, and will achieve 100 Gbit/s once the 8 Gbit/s lane speed can be stabilised. All other features, like latency and message rate are already available.

The advantage of EXTOLL TOURMALET is its switchless design, which integrates the switch in the NIC and allows constructing arbitrary large 3D torus topologies without additional external switches. The availability of an additional FPGA implementations of the EXTOLL link and functions allows to add additional experimental resources to the network, e.g. the NAM. Here the re-programmability of the FPGA allows testing new functions in the network attached memory while being compatible with the fast ASIC links.

For the next generation EXTOLL ASIC an own SERDES development is planned in order to avoid the problems and time delays of the external IP which was purchased for TOURMALET. This ASIC will improve the host interface to gen4 and the link speed. The SERDES function is expected to work with 20 Gbit/s per lane for the EXTOLL links.

Eurotech integrates the EXTOLL TOURMALET card in the Aurora Rootcard, on the upper part of the chassis. The TOURMALET card will have its own heat spreader with passive cooling connection to a cold plate, which is part of the Rootcard construction.

The Rootcard provides control and housing for 36 PCIe gen 3 x16 slots. The routing of the PCIe slots is shown in Figure 5.

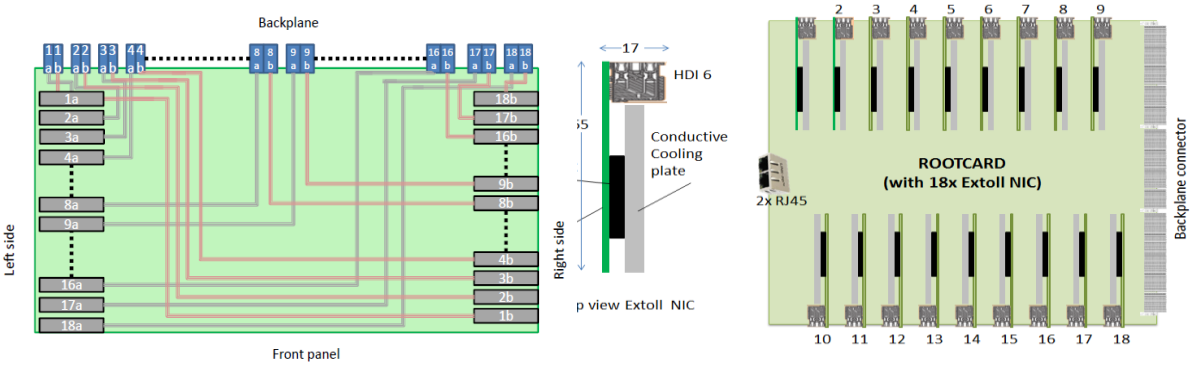

Figure 5. *Left:* Routing of the PCIe slots on the Aurora Root peripherals card cage.[3] *Right:* Position of TOURMALET.

Each KNL blade connects to a left slot (KNL 1) and to a right slot (KNL 2) of the Rootcard. The NVMe cards are integrated with the Expansion blade that has no PCIe connection to the Rootcard. The corresponding slots are not connected.

All PCIe slots are powered independently and control maintains the power even when the node is down/not connected. This feature is essential for providing the routing function of the

[3] There are 18 EXTOLL TOURMALET cards to provide NICs for 18 KNL nodes in the chassis.

EXTOLL TOURMALET cards, since these must be up and running independent of the state of the node they are connect to.

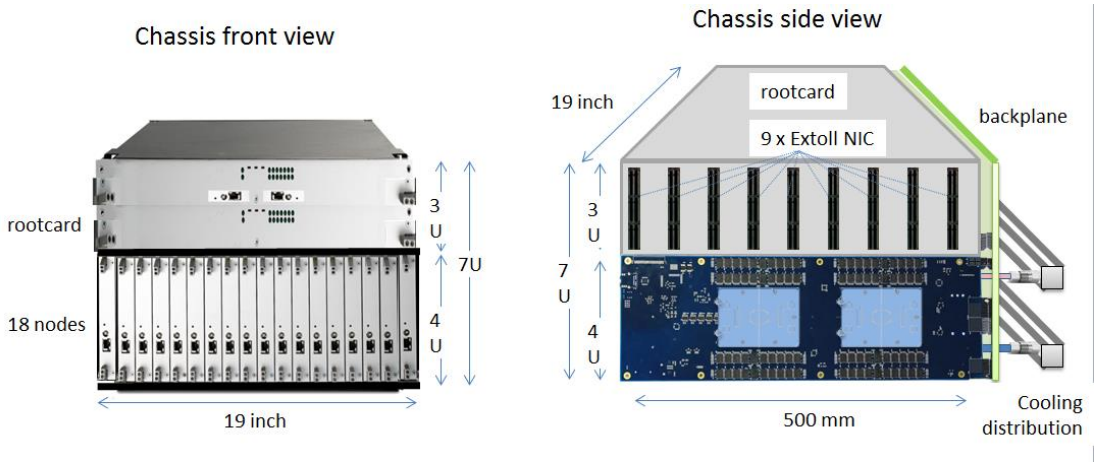

Figure 6. Front view (left) and side view (right) of the assembled chassis.

The DEEP-ER project has selected a 3D Torus network topology for the Booster part of the Prototype. The cabling of the network extends from the EXTOLL TOURMALET cards on the two sides of the chassis. There are 54 ports to be cabled on each side, i.e. 27 cables are needed. For the DEEP-ER Prototype with few chassis (up to single rack sizes) this cabling complexity is acceptable. With larger number of chassis (nodes) a special design must be adopted to bring the explosion in the number of interconnect cables under control.

For larger systems, the Aurora design rules with the PCIe backplane can be used to create an intra-chassis interconnect structure. PCIe links from each node here lead to a special multi-NIC to be designed on its own PCB plane with EXTOLL TOURMALET ASICs. One particular design point could be:

- Each chassis with fully loaded 36 KNL nodes (18 cards) is (part of) a 3D Torus with 4×3×3 topology.
- There are 216 ports to be connected. All internal connections can be provided on the new NIC PCB. Therefore, only 66 external links connect one chassis to the others.

2.4 Water cooling and integration

The Aurora Blade Architecture utilizes Eurotech water cooling technology. This consists of a heat spreader covering all active electronics on a board. The heat is removed by water flowing through the heat spreader. For the DEEP-ER project Eurotech implements an evolution of the technology used for the DEEP Booster. This is already tested and utilized in some other Eurotech products (e.g. Aurora Hive). An impression is given in Figure 7.

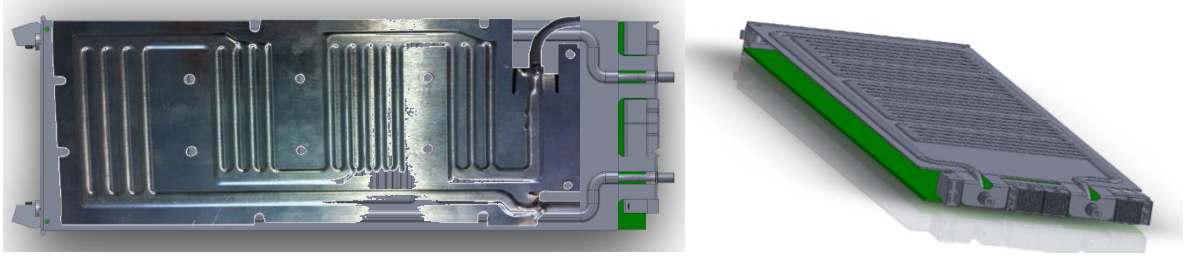

Figure 7: Impresion of the Heat Spreader and water distribution structure to eliminate the heat.

This new cooling technology makes the heat spreader and water connections thinner and uses less material, which leads to reduced weight of the installation. Therefore, there are 18 boards per 19” chassis in the present design. With the previous cooling technology, which was also used in the DEEP project, only 16 boards could be accommodated in an Aurora Blade machine.

The DEEP-ER project is an example of integration of three developments: state of the art processors (KNL), EXTOLL NIC and NVMe Storage. Each of these can be independently upgraded following the evolution of the respective technology, while the other parts can stay the same. Additionally, the exact system configuration, e.g. the amount of computing cards and storage cards, can be adapted to the specific requirements of the user and the network technology can be also freely chosen, at least as long as the corresponding NICs come in a PCIe-card configuration.

The Aurora line from Eurotech is mounted in 48U, double sided racks. In the exact system configuration chosen for the DEEP-ER Prototype (one NVMe per KNL node and a 3U Rootcard for the interconnect), each chassis has a 7U height and hosts 18 KNL devices (see Figure 7). This means, 6 chassis fit in each side of the rack, achieving a total density of 216 KNLs/48U-rack, i.e. about 648 TFlop/rack.

If NVMe is not used, the Aurora Blade system can be also populated, which would double the amount of KNL devices per chassis. Additionally, optimisation on the interconnect integration would allow for reduction of the Rootcard to the Aurora standard size of 2U. In this case, 8 chassis would fit in each side of the rack, achieving an unprecedented density of up to 576KNL/48U-rack. This is possible only with liquid cooling, since the heat dissipation may grow to 150 kW/Rack. A benefit of high density is the option to use short copper links for high-speed connections. Usage of copper saves energy which otherwise must be used for optical conversions and costs.

Yet, the number of network cables grows to unmanageable proportions when the machine grows in size. While the scalability in terms of chassis and racks is linear, the interconnect must be re-designed especially for large machine sizes. This challenge has been considered in Section 2.3.

3 Design assessment

The design of the DEEP-ER Prototype reflects the lessons learned from the DEEP System. For DEEP several compromises had to be made due to technical constraints:

In DEEP we were forced to use two different fabrics for the Cluster and the Booster sides of the system. While EXTOLL's software stack was not mature enough to be used in the Cluster (e.g. support for parallel filesystems was required) InfiniBand was not able to support the remote-boot functionality needed to run KNC processors quasi stand-alone. Since applications still required exchanging data between their Cluster and Booster parts, gateways had to bridge between these two fabrics. On a physical level this was realised by the so-called Booster Interface Cards while on a logical level the highly efficient Cluster Booster Protocol reduced the overhead to the bare minimum. Nevertheless, it is clear that the Booster Interface introduced a bottleneck in the overall design.

DEEP-ER addresses this shortcoming by having a common fabric for the Cluster and the Booster parts of the system. On the one hand the next generation Xeon Phi processor (KNL) will be self-booting, thus, the constraints on choosing the fabric are significantly relaxed. On the other hand EXTOLL and its software stack grew mature over the years. In fact it provides native support for the BeeGFS parallel filesystem in the meantime.

For the DEEP-ER System an EXTOLL implementation based on the TOURMALET ASIC was chosen. This implementation provides significantly higher bandwidth compared to the FPGA implementation used in the DEEP Booster. This holds for both sides of the NIC, the PCIe bus connecting to the processor (TOURMALET supports PCIe gen3 x16 compared to PCIe gen2 x8 in the DEEP Booster) and for the links (each TOURMALET link provides 100 Gbit/s bandwidth, while in the DEEP Booster links were limited to 32 Gbit/s).

Nevertheless, KNL provides a total of two x16 PCIe connections. Thus the current DEEP-ER design will waste half of the available external bandwidth provided by KNL. This decision was made in order to reduce the technical complexity of the design, for cost reasons and last but not least to enable the inclusion of non-volatile memory. Today NVM is attached via PCIe, thus, additional links are used for this purpose. Nevertheless, the modular concept of the Aurora blade design allows increasing injection bandwidth into the fabric in future versions of a Booster system, when either NVM is not used or could be attached via standard memory channels, as hinted at by Intel's 3D XPoint technology announcement.

A significant constraint from the application's point of view was the very limited memory capacity available on KNC. For a total of ~1 TFlop/s at most 16 GB of memory were available. This improves significantly on KNL given t that up to six DDR4 channels are supported with a maximum buildout of 384 GB. DEEP-ER decided to populate each of them with 16 GB of memory leading to a total of 96 GB at about the same speed as the GDDR5 memory of KNC. Since KNL is expected to provide ~3 TFlop/s of compute performance, this decision enables application developers to double the amount of data per Flop. Of course, at the same time the bandwidth per Flop to external memory is reduced due to the larger compute capabilities of KNL. Intel addresses this challenge by introducing 16 GB of on-package MCDRAM memory with ≥ 4× DDR4 bandwidth to be either used as a cache or fully controlled by software. Thus, we expect KNL and thus DEEP-ER to be better balanced from memory capacity and bandwidth point of view than KNC and DEEP.

4 Performance extrapolation

In the previous section, we have discussed the main design decisions taken in DEEP-ER and their consequences/benefits in terms of performance, density, energy efficiency, etc. Naturally, this approach provides little information on whether scientific applications can actually leverage this technology and reap the benefits. Embracing true co-design, DEEP-ER is bringing along a total of seven real-world scientific applications, which will be ported to and optimized for the Aurora Blade prototype and its underlying enhanced Cluster-Booster concept. An obvious question is therefore, how these applications would scale to higher core counts and how their performance would benefit from the architecture. A model-based approach is used throughout the project to estimate the application scalability for different systems and architectures. In this section, we describe the most relevant studies performed until now, which leverage also the approach and results obtained within the DEEP project and the applications used therein.

4.1 Method description

The estimates for the scalability are based on the BSC parallel efficiency model [4]. In summary it is a multiplicative model that characterizes application performance using different factors and that can be applied to understand and predict the application scalability. The factors are measured as a value between 0 and 1, the higher the better. An efficiency of 0.85 indicates that 15% of the corresponding resources are wasted. This value can be considered a boundary between good and bad performance.

The parallel efficiency ($\eta_{\parallel}$) represents the percentage of time spent on the computation (useful work) with respect to the total execution time. The parallel efficiency can be decomposed into three main factors:

- Load balance (LB), measures the efficiency losses due to differences on the computing time between processes. If some processes take more time in the computation, the other processes would have to wait for them in the synchronization of the MPI calls for instance.
- Transfer, measures the reductions on the efficiency due to the need to transfer data between processes. If the application is dominated by communications, the transfer efficiency would be low.
- Serialisation (µLB), measures the efficiency losses due to dependencies during the execution. This factor also reflects load imbalances that can be compensated along time. If on the even iterations half of the processes do more work than the other half but on the odd iterations the behaviour is just the opposite, the load balance would report a good efficiency while the serialisation would reflect the loss of efficiency.

The load balance can be directly measured from an instrumented execution. In order to compute the transfer and serialisation factors, it is necessary to isolate the application execution from the network characteristics. This might be done using the Dimemas simulator [4] to predict the behaviour running on an instantaneous network.

Using the efficiencies at different core counts on a relatively small scale up to thousand processes, the estimates are computed assuming that the loss of efficiency when increasing the scale will follow Amdahl's law. By default, the Amdahl model would reflect the contention/serialisation on a given resource

$$Amdahl_{fit} = \frac{metric_0}{f_{metric} + (1 - f_{metric}) * P}$$

Where metric is the efficiency that we are modelling and *P* the number of processors for a given run. If the collected traces allow identifying an underlying physical phenomenon that follows a more specific law for a given efficiency factor, the parameter *P* in the previous formula can be substituted by the corresponding/approximate function based on the number of processes.

4.2 Modelling applications

Three applications have been analysed and are reported in this deliverable. One code comes DEEP-ER, and two are part of the DEEP application portfolio:

- **MAXW-DGTD**, the simulation of the impact of magnetic fields on human tissues from DEEP-ER partner Inria. Traces for this application were obtained in the DEEP

Cluster, which is a cluster of 128 Intel Xeon processors from the Sandy Bridge generation.

- **CoreBluron**, the brain simulation from EPFL. Traces from this application were obtained within the DEEP project on JSC's BlueGene/Q system JUQUEEN within the scope of the DEEP project. For this application scalability analysis of the parallel efficiency has been done, to estimate the results expected on very large platforms.
- **AVBP**, the CFD code from CERFACS. Traces have been obtained on JSC's BlueGene/Q system JUQUEEN within the scope of the DEEP project. Here the measurements are strong scaling. For this application scalability analysis of the parallel efficiency has been done, to estimate the results expected on very large platforms.

It is important to notice that the test case and size of the input data was selected suited to the platform used for obtaining the traces. Using input data adapted to the DEEP-ER Prototype should improve the results there, especially in the case of strong-scaling experiments (AVBP). Such measurements will be done once the DEEP-ER SDV and later the Prototype are available.

The estimates for the scalability are based on the BSC parallel efficiency model described in Section 4.1. DEEP-ER uses the second generation of Intel Xeon Phi (codenamed KNL). Differences seen when running applications on different platforms come not only from the improvement on raw performance but also from a larger amount of available local memory and better I/O thanks to the usage of NVMe. Since KNL was not yet available, we used the Dimemas simulator with an estimate of the relative performance with respect to the architecture on which traces were obtained.

The estimated performance on the DEEP[4] and DEEP-ER Booster is calculated based on the ratio between the peak performance of the processors in which the application traces were obtained and the expected performance from the KNL's processors. Additionally, a correction factor is applied to account for performance loses when using OpenMP, since both KNC and KNL rely more on multithreading than a general purpose processor, such an Intel Xeon. Since we at this time do not know exactly the performance that KNL can achieve for the applications used, we picked as baseline its predecessor (KNC), and we accounted for a range of ×2 and ×4 improvement in performance. In other words, KNC has a peak performance of 1.2 TFlop/s and Intel has announced a peak KNL performance of 3 TFlop/s, which suggests a factor of three in CPU performance. To be on the safe side, simulations were run with factors of 2 and 4.

Improvements in network performance has also been taken into account. The network on the DEEP Booster is an implementation of the EXTOLL protocol on FPGAs with an internode latency of 15 µs, while the DEEP-ER Booster will be built with the ASIC version EXTOLL TOURMALET, with an inter-node latency of ~2 µs. Since we have to take into account the end-to-end latency for MPI messages, the simulation runs were performed with a latency value of 4 µs. Depending on optimizations in the MPI SW stack, this value may turn out to be too high.

[4] At the time of writting the DEEP Booster has been installed but is not yet running stable, so that thorough application measurements on the platform could not be performed yet. Therefore, the data shown in this Deliverable on the DEEP Booster is not measured but modelled, using an equivalent method as for the DEEP-ER Prototype.

The parallel efficiency of the different applications has been modelled and analysed for three different configurations:

- **DEEP Booster:** KNC processors and the FGPA implementation of EXTOLL, with 15 µs latency and 1.2 GB/s bandwidth.
- **DEEP-ER Prototype (2×)**: assuming KNL to provide 2× the performance of KNC, and using the EXTOLL ASIC with 4 µs latency and 8 GB/s bandwidth.
- **DEEP-ER Prototype (4×):** assuming KNL providing 4× the performance of KNC, and using the EXTOLL ASIC with 4 µs latency and 8 GB/s bandwidth.

This study allows us to evaluate the impact of the upgrade from KNC to KNL and from the FGPA to the ASIC implementation of EXTOLL. Please note that the factors applied to the application traces to model the behaviour on the DEEP-ER Prototype (for the last two bullets) are calculated individually for each application. The specific factors applied to the traces depend both on the platform where the traces have been obtained and on some configuration factors used in the tracing run, such as the amount of processes that were running on each node.

4.2.1 MAXW-DGTD: simulation of electromagnetic fields interacting with human tissue, from partner Inria

In this case, traces were created running on the DEEP Cluster (Intel Xeon processors of the "Sandy Bridge" generation). The traces were obtained from a pure MPI run (no OpenMP), with one MPI rank per processor, i.e. 16 ranks per node. When modelling the DEEP-ER Booster we are considering running one rank per 8 cores of the KNL node, i.e. 9 MPI ranks per KNL on the assumption of 72 cores per KNL. This has been taken into account when calculating the CPU ratios.

Figure 8 shows the general view of the application run. In blue we have computational regions, orange are collective communications and yellow lines indicate actual MPI messages. The application first loads the mesh. This process is very time-consuming compared to the actual computation of the solver. The application developers do plan to rebuild it to improve its efficiency. The computation of the solver corresponds to the yellow column at the end (it should be blue, yet the yellow communication lines obscure the computation in this view). In the present deliverable the analysis will focus on a representative region of the application, which corresponds to the five iterations of the main loop and is displayed in Figure 9.

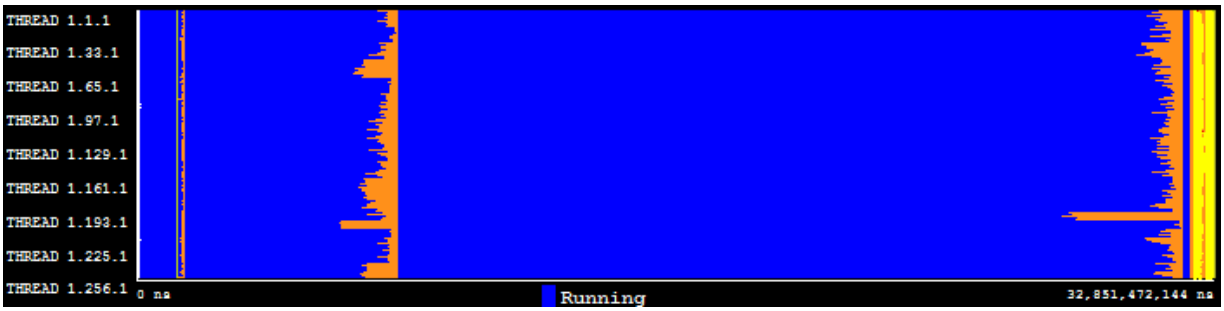

Figure 8: General view of the MAXW-DGTD application run.

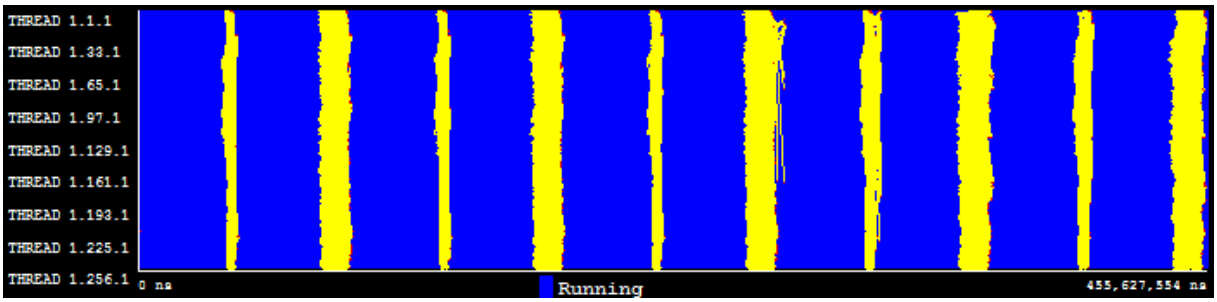

Figure 9: Zoom in to the region under analysis.

In order to provide a better understanding of the behaviour of the application we performed a cluster analysis, a commonly used explorative data mining technique for classification of data. We applied the cluster analysis to detect different trends in the computation regions of the code. In this study, we looked at the trends of the different “CPU bursts” (or simply “bursts”) of the application. We define a “CPU burst” as the region in a parallel application between calls to the parallel runtime (like calls to MPI or entering/exiting an OpenMP region). Then we apply the DBSCAN clustering algorithm [5]. As a result, we obtain the different groups of bursts according to the pair of performance counters used; in this case we chose the “Instructions Completed” and “Instructions Per Cycle” (IPC) metrics so as to reflect the different performance achieved by the different bursts along the application execution.

In Figure 10 and Figure 11 we can see the result of the cluster analysis for a run of 64 and 512 MPI processes, respectively. Three clusters have been identified. Cluster 1 and cluster 2 are pretty similar in IPC but the difference in the duration of the computation regions make us think that they belong to different parts of the workload. Then there is a third cluster, with lower number of instructions and higher variability in IPC. Comparing the two images we can see how clusters are kept along with their IPC but there is a difference in the number of instructions completed showing that the workload is running in strong-scaling mode, since number of instructions per burst decreases as we increase the number of processes.

Figure 12 (for a 64 MPI processes run) shows exactly how each cluster matches to each region of the code. Thanks to “useful duration view” at the top we can identify three different parts in the code (in this view, darker colours represent higher IPC values). Each iteration has a first long burst of computation (dark blue) corresponding to the computation of the electrical field. Then, there is some MPI communication (green because of lower IPC), followed by the computation of the magnetic field (again in blue, although slightly lighter). Finally there is the Fourier Transform computation, lying between some non-blocking receives and sends of the magnetic field computation, causing a lower IPC, so it is shown in green. The three regions can also be identified in the cluster view at the bottom of Figure 12. Cluster 1 (green) corresponds to the electrical field, cluster 2 (yellow) to the magnetic field and cluster 3 (red) to the Fourier Transform. Figure 13 shows the useful duration and the clusters for a 512 MPI processes run, using the same time scale as Figure 12. You can see that the same sequence of clusters occurs, but variations in compute times do increase.

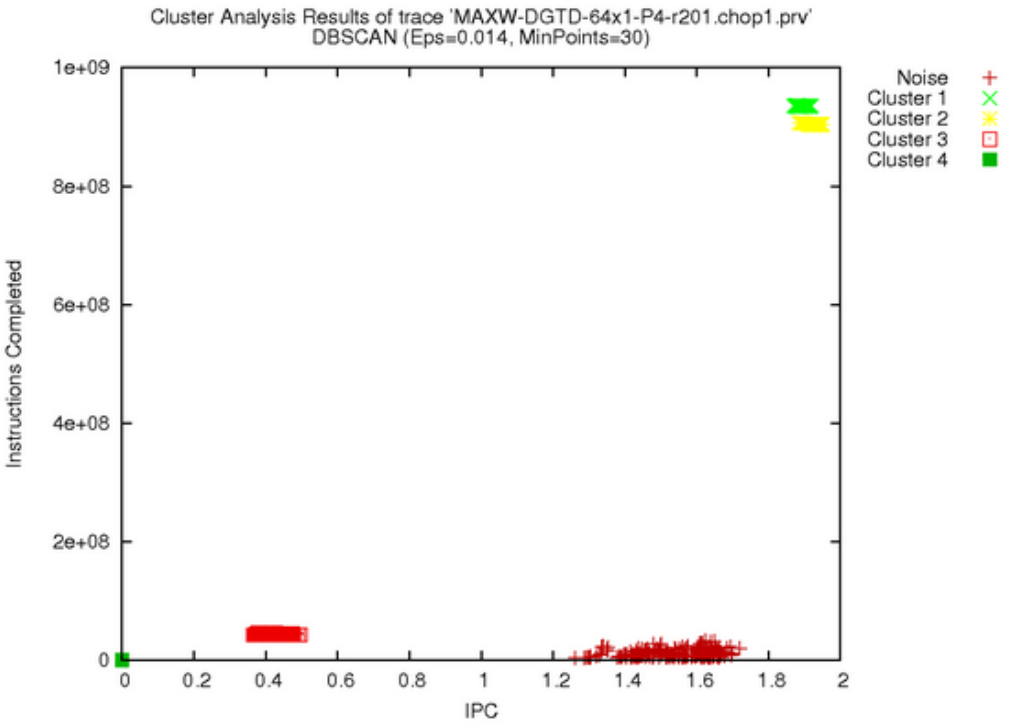

Figure 10: MAXW-DGTD: Cluster analysis results for 64 MPI processes.

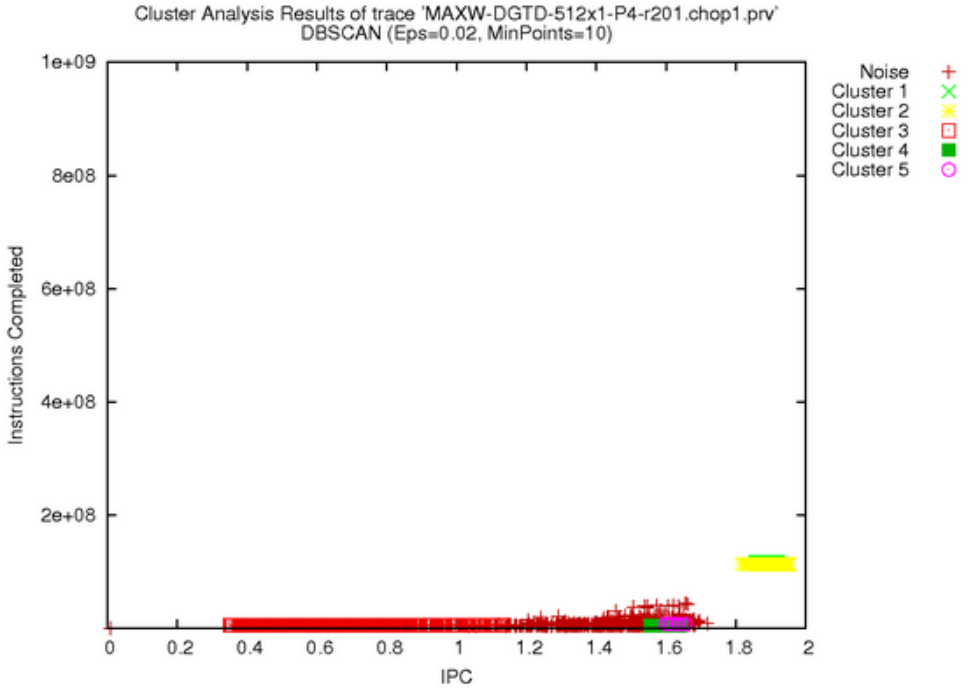

Figure 11: MAXW-DGTD: Cluster analysis results for 512 MPI processes.

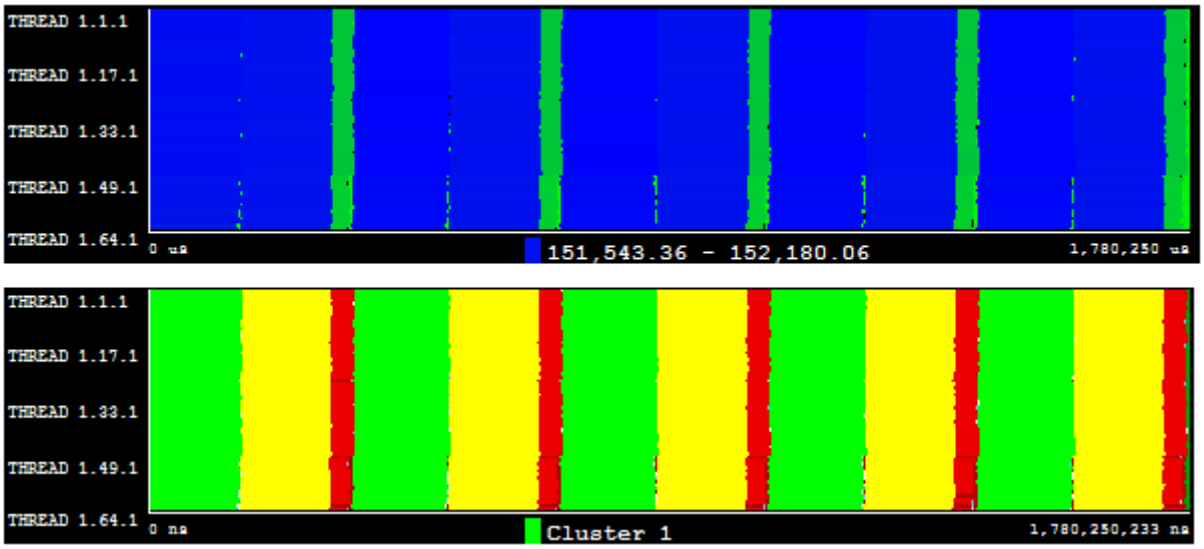

Figure 12: Useful duration view (top) and Cluster ID view (bottom) for 64 MPI processes.

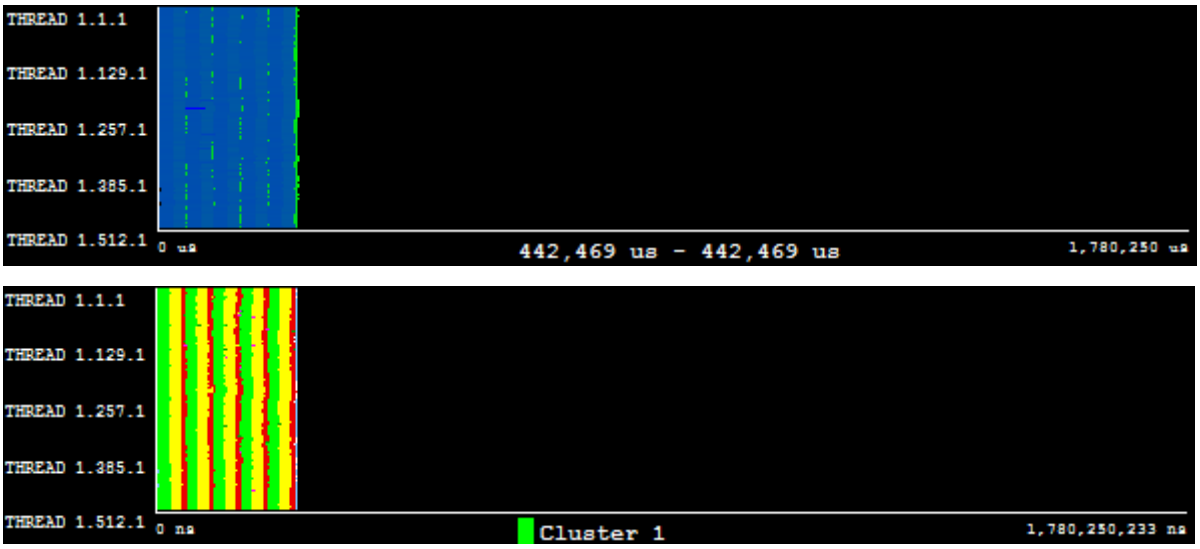

Figure 13: Useful duration view (top) and Cluster ID view (bottom) for 512 processes.

Figure 14 depicts the parallel efficiency of MAXW-DGTD running on the DEEP Cluster, showing a good parallel efficiency (close to 100%) up to using 512 processors, although it decreases slightly. The data shown here have been measured on the platform, but we do not have the estimates for a larger number of processors, yet.

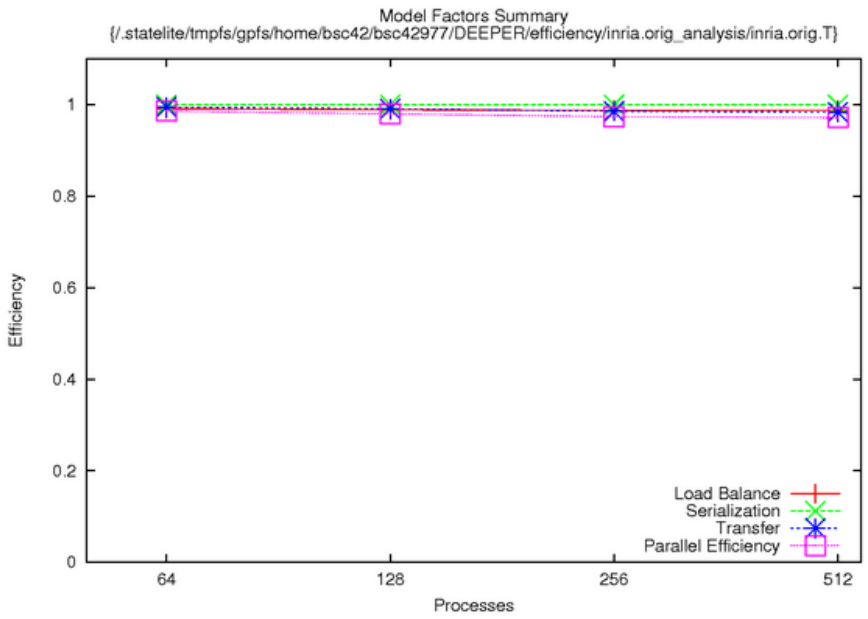

Figure 14: MAXG-DGTD: Parallel efficiency on the DEEP Cluster (Intel Sandy Bridge processors).

Figure 15 depicts the predicted parallel efficiency for running on the DEEP Booster. Parallel efficiency is really good even when using 512 MPI processes (~90%) but shows a trend of degrading linearly with the number of processes.

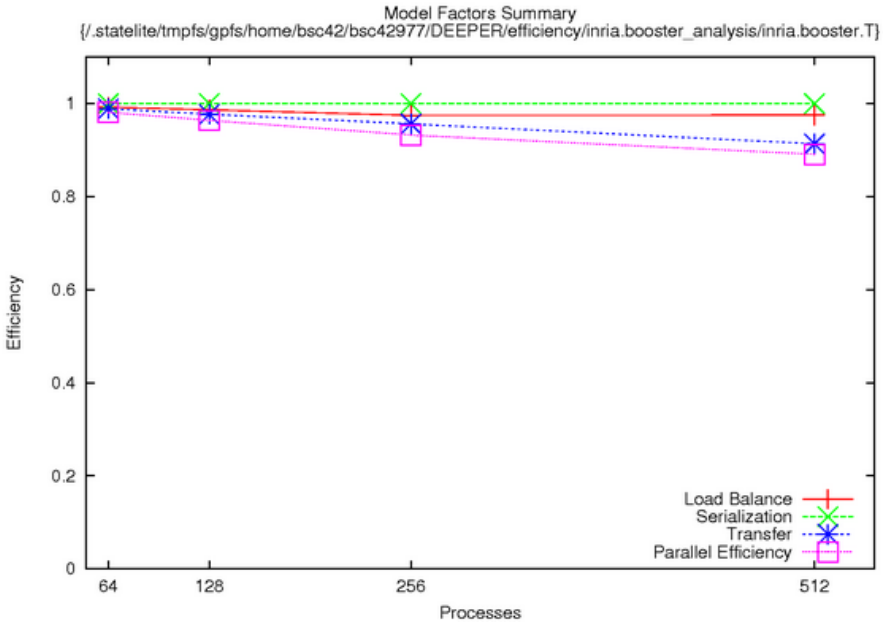

Figure 15: MAXG-DGTD: Parallel efficiency on the DEEP Booster (KNC with FPGA-EXTOLL).

Figure 16 and Figure 17 show the expected parallel efficiency for MAXW-DGTD in the DEEP-ER Prototype. The increase in CPU performance comes along with an improvement in the network interconnect, which brings an improvement in parallel efficiency since this is a latency bound application and latency has been reduced by almost four times with the new generation of EXTOLL network. With a CPU ratio of ×2, the parallel efficiency is higher than 95% (see Figure 16), which is a really good efficiency. However, with a CPU ratio of ×4 efficiency seems to degrade faster with 512 MPI processes or more and is similar to the efficiency of the DEEP Booster, but it is still very good (~90%), see Figure 17.

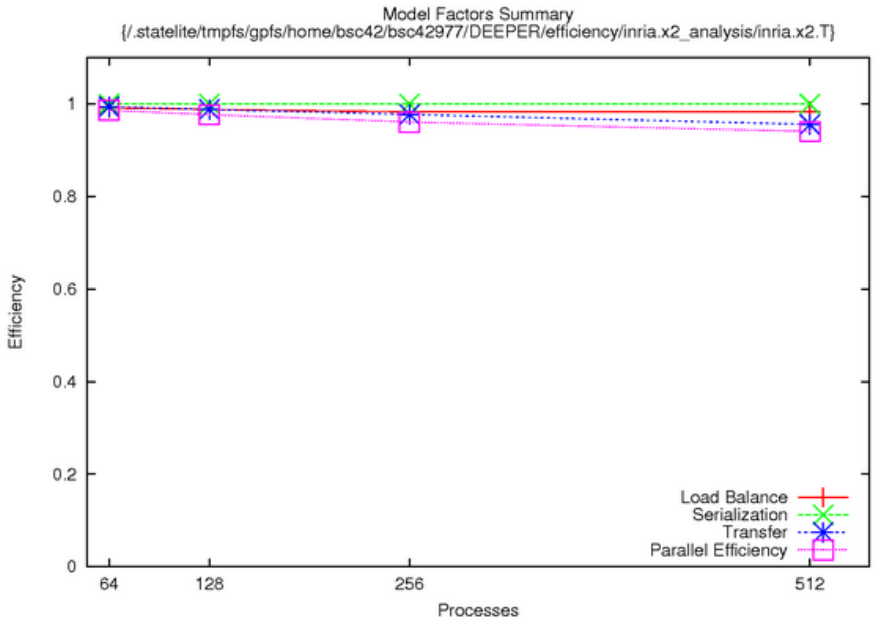

Figure 16: MAXW-DGTD; Parallel efficiency expected on the DEEP-ER Prototype (assuming KNL performance = KNC ×2)

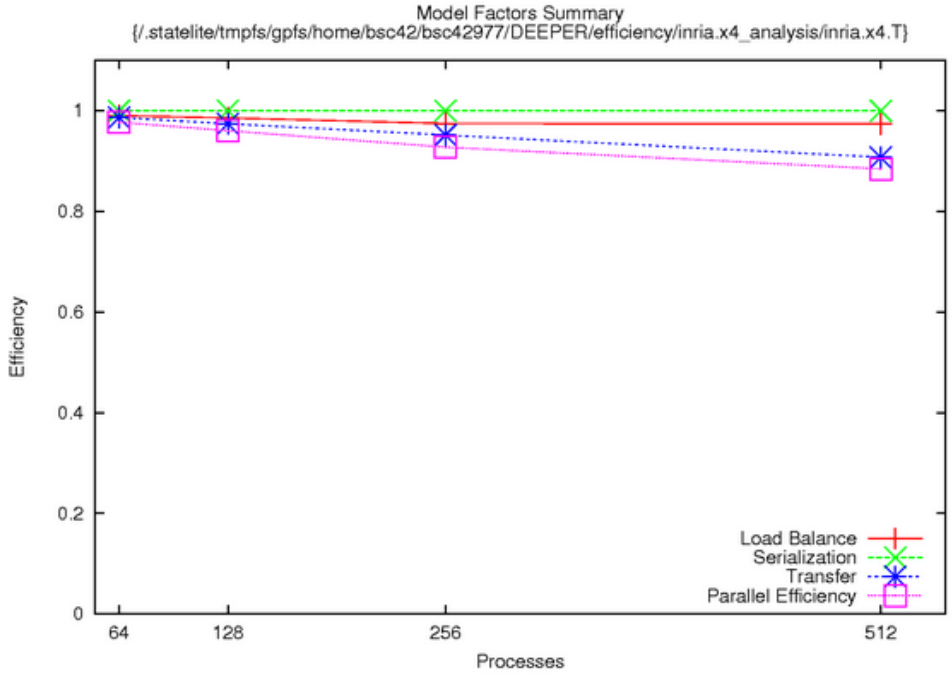

Figure 17: MAXW-DGTD; Parallel efficiency expected on the DEEP-ER Prototype (assuming KNL performance = KNC ×4)

4.2.2 CoreBluron: brain simulation application from EPFL

The traces for this application were obtained on JSC's BlueGene/Q system JUQUEEN within the scope of the DEEP project.

The Figure 18 depicts the estimates for the parallel efficiency of CoreBluron on the DEEP Booster. Figure 19 and Figure 20 give the estimates for the DEEP-ER Booster, assuming a KNL performance of ×2 and ×4, respectively. In all three cases estimations of the application efficiency when scaling up the systems has been performed taking into account the specific platform characteristics.

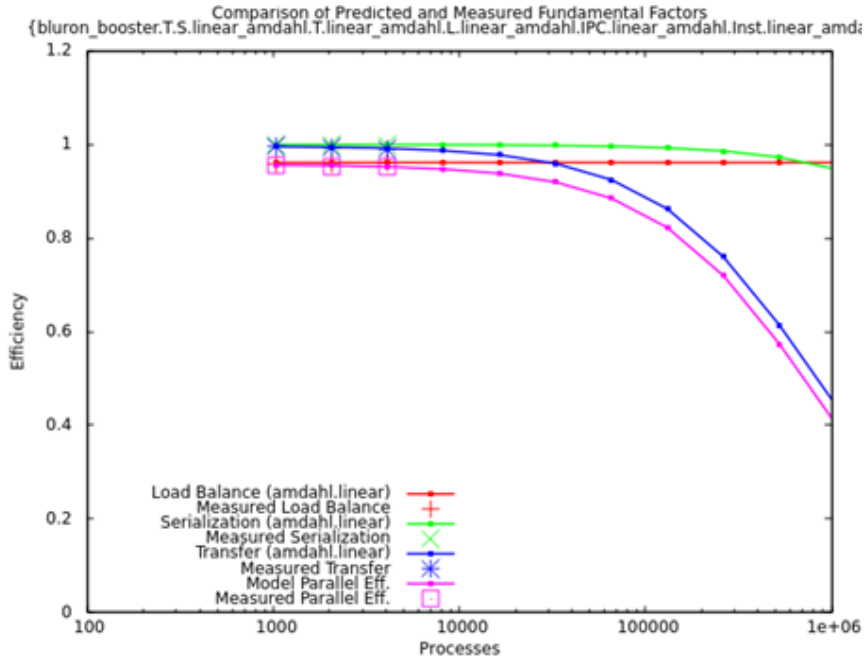

Figure 18: CoreBluron; Parallel efficiency estimated on the DEEP Booster (KNC with FPGA-EXTOLL).

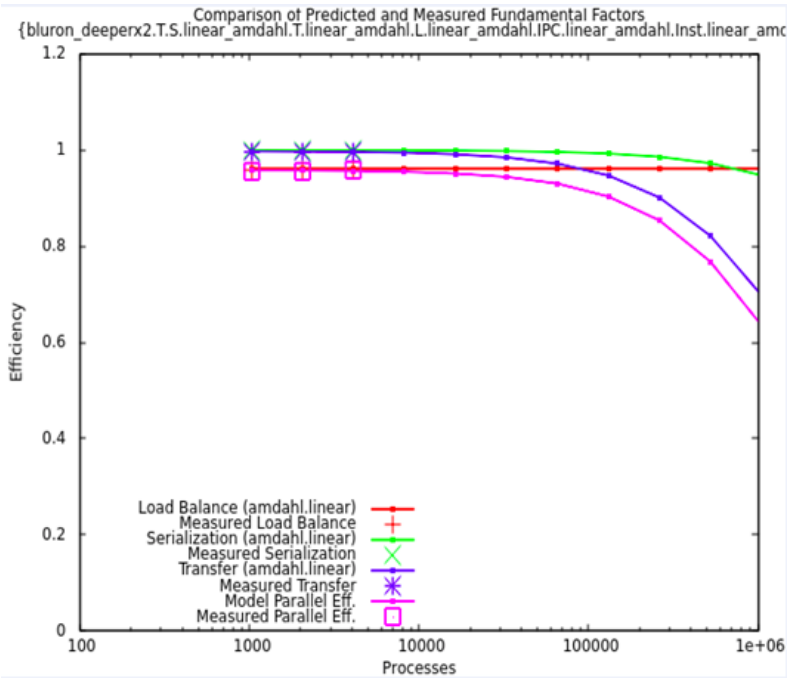

Figure 19: CoreBluron; Parallel efficiency estimated on the DEEP-ER Prototype (assumming KNL perfromance as ×2 KNC).

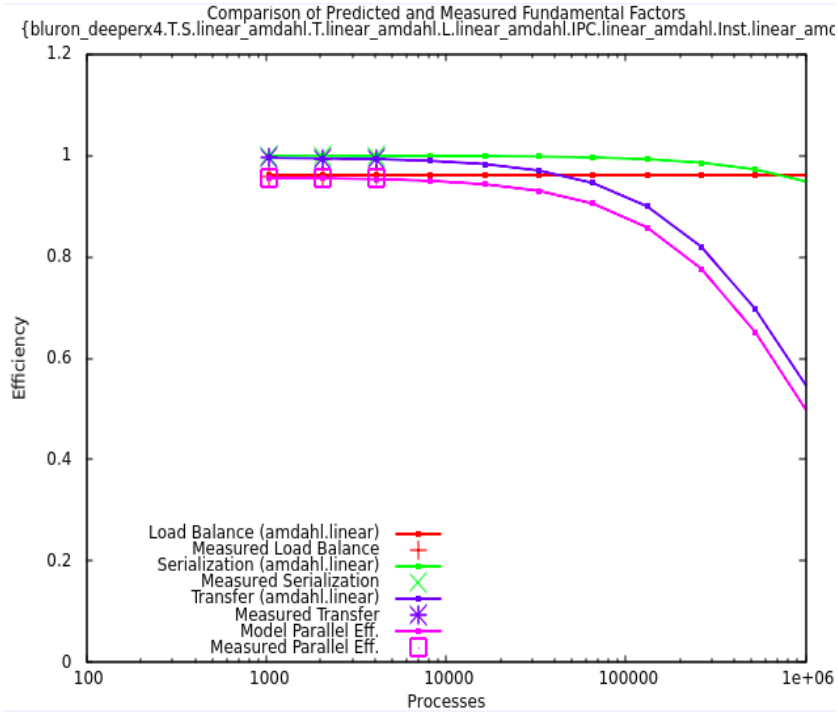

Figure 20: CoreBluron; Parallel efficiency estimated on the DEEP-ER Prototype (assumming KNL perfromance as ×4 KNC)

In all three cases, efficiency is very good (~95%) up to 64K nodes and after that point parallel efficiency drops. While load balance remains constant and serialization efficiency decreases only less than 10%, the transfer efficiency is the factor making the biggest impact on the drop in parallel efficiency.

The difference between the two scenarios (×2 and ×4) on DEEP-ER comes from faster computation when using a CPU ratio of ×4, which produces a reduction of the computation's weight while the transfer's weight remains the same.

When comparing the timeline of the duration of computing bursts for the 4096 runs on the 3 simulated scenarios, it was observed that computation time is strongly reduced.

For both CPU factors, the DEEP-ER Prototype shows appreciably higher efficiency than the DEEP Booster. From the analysis performed in DEEP, we do know that the end-to-end message latency is relevant for the application performance. Further improvements of the DEEP-ER system efficiency can come from a careful tuning of the MPI SW stack resulting in latencies of <4µs.

4.2.3 AVBP: CFD code from CERFACS

In this case traces were obtained in JUQUEEN, which is a BlueGene/Q supercomputer from JUELICH.

Figure 21, Figure 22, and Figure 23 present the estimates of the parallel efficiency of AVBP using the DEEP Booster and different CPU ratios for the DEEP-ER Prototype.

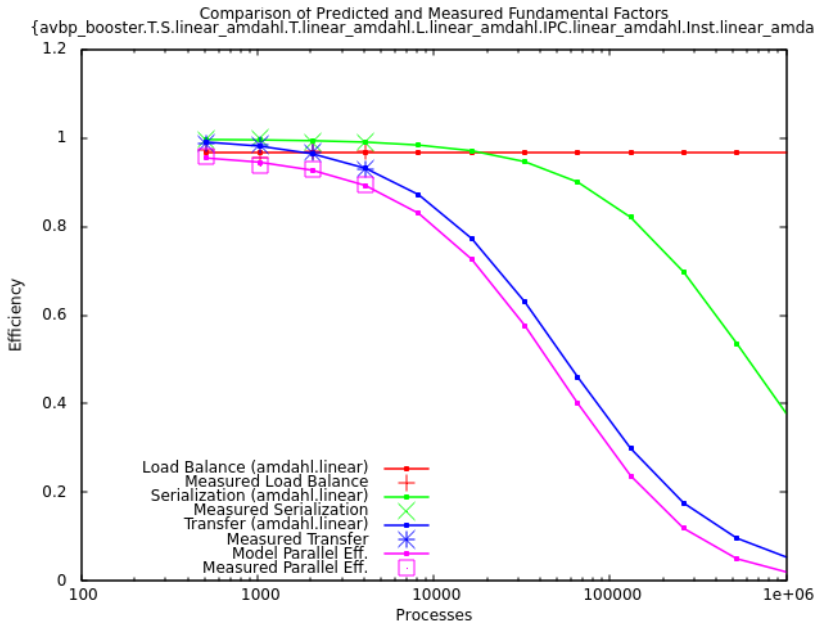

Figure 21: AVBP: Parallel efficiency estimated on the DEEP Booster (KNC with FPGA-EXTOLL).

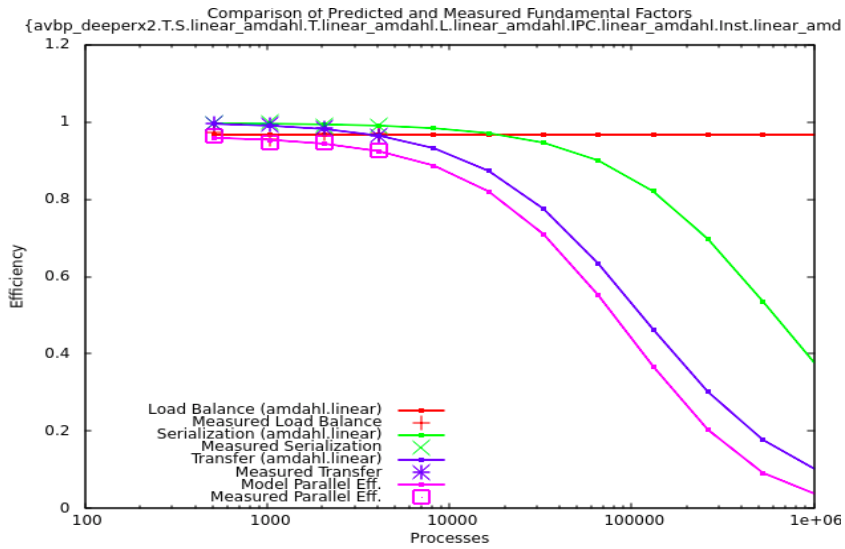

Figure 22: AVBP: Parallel efficiency estimated on the DEEP-ER Prototype (assumming KNL perfromance as ×2 KNC)

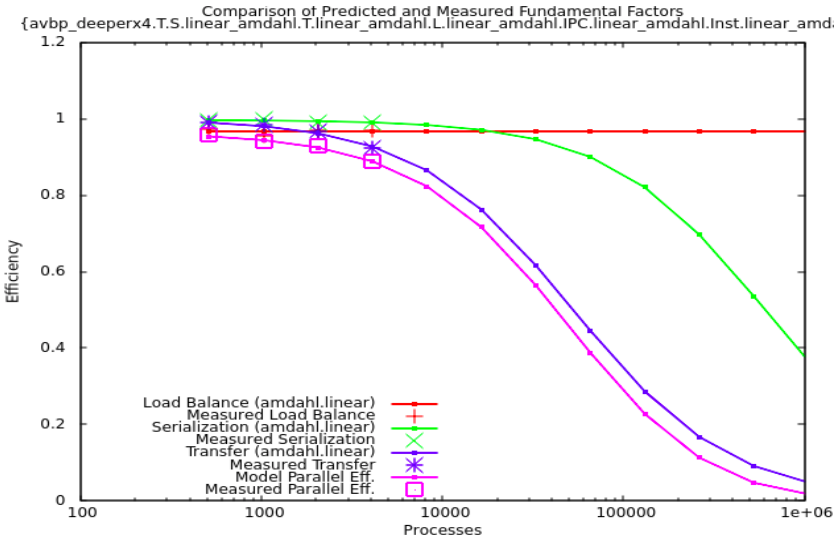

Figure 23: AVBP: Parallel efficiency estimated on the DEEP-ER Prototype (assumming KNL perfromance as ×4 KNC).

In this case parallel efficiency drops quickly and goes down up to less than 5% for 1 million processes in all three cases. Up to 16k cores the efficiency is good (around 80%) considering that it is using strong scaling. With 1 million processes the efficiency is 3.46%.

The AVBP experiments are done with strong scaling. The input data was adapted to run up to some hundred thousand cores in BlueGene/Q, but it is not large enough to run with the many-core approach of KNC/KNL, in which a very large number of MPI processes is required.

While load balance remains constant, both transfer and serialization efficiencies do not scale well with the number of processes. However, transfer efficiency is again the main responsible for the poor parallel efficiency. Thus, doubling the CPU ratio reduces even more the parallel efficiency, since computation's weight is reduced while transfer weight is maintained.

Finally, we compared the execution of the 3 simulated environments for the 1024 MPI ranks case using as reference the MPI calls. We can see that the time spent on MPI communication reduces when improving the network performance, same as the time spent in computation when using a faster CPU.

Also in this case, the parallel efficiency predicted for the DEEP-ER is higher than that of the DEEP Booster. The same argument – on the effect of improvements of the end-to-end MPI latency to bring it below the assumed 4 µs – made in Section 4.2.2 applies here.

5 Summary and Conclusion

The configuration of the DEEP-ER Prototype – based on the Aurora Blade architecture from Eurotech –, has been decided following a co-design approach in which the requirements of the software and application developers from the project were used as guidelines. Additionally, the lessons learned from the DEEP project have also been applied. The goal of this approach, in conjunction with the use of better processor, memory and network technologies, is to make the DEEP-ER Prototype a better machine than the DEEP Booster. A very important aspect is the potential for future commercial development and exploitation; here, the Aurora Blade design is much superior to the DEEP design, due to its energy efficiency, flexibility and adaptability to different portfolios and purposes.

Preliminary analysis of three HPC applications and their expected efficiency on the DEEP-ER Prototype have been shown. For two of the codes the results were additionally extrapolated to a much higher number of processes, i.e. to a much larger Aurora Blade system than the prototype realised within the project. With these studies the application's performance on a hypothetical production machine based on the DEEP-ER Prototype can be reproduced. In all cases, the predicted parallel efficiency of the DEEP-ER Prototype appreciably exceeds that of the DEEP Booster, even for quite conservative assumptions on the communication performance. The concrete results depend of course not only on the platform, but also on the specific application characteristics and its intrinsic scalability. One of the applications (CoreBluron) shows an >80% parallel efficiency for up to about 200.000 processes, decreasing to about 60% when running with 1 million processes. On the other hand, the parallel efficiency of the AVBP code decreases quite fast. However, it shall be stressed that in this case the test is done with strong scaling, and that the input data used for the analysis was suited for the BlueGene/Q architecture, rather than for the DEEP-ER platform. A use case suited for the DEEP-ER Prototype will be selected once the platform is available, and the experiments will be repeated.

The efficiency results for different processor performance (DEEP-ER Prototype ×2 and ×4) show that a better processor raises the application requirements with respect to the network, pointing to the communication-bound nature of the analysed applications. Both CoreBluron and AVBP show reduced transfer efficiency when the processor speed rises. Both results stress the importance of the end-to-end communication latency, which seems to have a larger influence on achieved efficiency than network bandwidth. Further work is needed to quantify levels of message latency that should be achieved by the DEEP-ER Prototype, which in turn might require work in the system software layers.

Next steps in WP7 are to perform a detailed analysis of the DEEP-ER applications. For this purpose, their traces will be obtained with the Extrae/Paraver tool from BSC, and the Dimemas modelling software will be applied. To make the extrapolation more reliable, measurements on the SDV (to address the EXTOLL network performance) and on the earliest available KNL-platforms (to address the KNL performance) will be done and included in further deliverables. Whenever possible, additional aspects, such as the impact of the I/O and resiliency functionalities developed within the project, will be taken into account.

6 References

1	Sodani, A.	Knights Landing (KNL): 2nd Generation Intel® Xeon Phi™ Processor, Presentation at Hot Chips: A Symposium on High Performance Chips, Cupertino, August 23-25, 2015	2015
2		Intel® Solid-State Drive DC P3700 Series, product specification available at http://www.intel.com/content/www/us/en/solid-state-drives/ssd-dc-p3700-spec.html	2015
3		Intel and Micron Produce Breakthrough Memory Technology, available at http://newsroom.intel.com/community/intel_newsroom/blog/2015/07/28/intel-and-micron-produce-breakthrough-memory-technology	2015
4	Rosas, Claudia; Giménez, Judit; Labarta, Jesus	Scalability prediction for fundamental performance factors, Supercomputing frontiers and innovations, vol. 1, no. 2, pp. 4–19, 2014.	2014
5	Ester, Martin; Kriegle, Hans-Peter; Sander, Jörg; Xu, Xiowei	A density-based algorithm for discovering clusters in large spatial databases with noise, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, 1996, pages 226–231, available at http://www.dbs.ifi.lmu.de/Publikationen/Papers/KDD-96.final.frame.pdf	1996

List of Acronyms and Abbreviations

A

AoS: Array of Structs
API: Application Programming Interface
ASIC: Application Specific Integrated Circuit, Integrated circuit customised for a particular use
ATOLL: Predecessor of EXTOLL
Aurora: The name of Eurotech's cluster systems
AVBP: A parallel CFD code for reactive unsteady flow simulations on hybrid grids developed by DEEP-partner CERFACS

B

BADW-LRZ: Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften. Computing Centre, Garching, Germany
BeeGFS: The Fraunhofer Parallel Cluster File System (previously acronym FhGFS). A high-performance parallel file system to be adapted to the extended DEEP Architecture and optimised for the DEEP-ER Prototype.
BIOS: Basic I/O system. Boot and system initialisation code run before the OS starts
BLAS: Basic Linear Algebra Subprograms
BLN: Brick local network. Used to locally connect the Brick modules
BMC: Board management controller. Used to physically monitor and manage a compute blade.
BN: Booster Node (functional entity); refers to a self-booting KNL board (Node board architecture) including the NVM and NIC devices connected by PCI Express or a Brick (Brick architecture).
BNC: Booster Node Card is a physical instantiation of the BN
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.
Brick Module: Smallest functional HW entity. Up to 6 modules are aggregated into a Brick
BSC: Barcelona Supercomputing Centre, Spain
BSCW: Basic Support for Cooperative Work, Software package developed by the Fraunhofer Society used to create a collaborative workspace for collaboration over the web

C

CAE: Computer Aided Engineering
CD: Corporate Design

CEPBA: European Centre for Parallelism, Barcelona, Spain

CERFACS: Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique, Toulouse, France. An application partner in the DEEP project bringing AVBP, a CFD application.

CFD: Computational Fluid Dynamics

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)

CI: Corporate Identity

CINECA: Consorzio Interuniversitario, Bologna, Italy

CN: Cluster Node (functional entity)

CNR: National Research Council, Italy

CNRS: Centre National de la Recherche Scientifique, Paris, France

Coordinator: The contractual partner of the European Commission (EC) in the project

CoreBluron: An application for brain simulation developed by EPFL, a partner in the DEEP project.

COTS: Commercial off the shelf.

CPU: Central Processing Unit

CRB: Customer Reference Board. An early version of a KNL board developed by Intel

CRESTA: Collaborative Research into Exascale Systemware Tools & Applications: EU-funded Exascale project.

CTC: Chief Technology Coordinator

D

DAG: Directed acyclic graph.

DDG: Design and Developer Group of the DEEP-ER project

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

DESCA: Comprehensive, modular consortium agreement for the Seventh Framework Programme (FP7)
DFG: Deutsche Forschungsgemeinschaft, German research organisation
DGTD: Discontinuous Galerkin – Time Domain solver
DMA: Direct Memory Access
DoW: Description of Work
DRAM: Dynamic Random Access Memory. Typically describes any form of high capacity volatile memory attached to a CPU

E

E10: Exascale 10. Parallel I/O software developed by a consortium of partners around the EOFS community. Partner Xyratex is responsible for the development needed for the DEEP-ER project.
EATC: European Altair Technology Conference
EC: European Commission
ECC: Error correction code. Corrects errors in storage and transmission systems by added redundancy
EC-GA: EC-Grant Agreement
ECL: ExaCluster Laboratory, A collaboration of Intel, ParTec and JUELICH to develop cluster management software for Exascale computing
EEP: European Exascale Projects
EESI: European Exascale Software Initiative (FP7)
EMEA: Europe, the Middle East and Africa, Regional designation used for government, marketing and business purposes
ENI: Italian Oil and Gas Company ENI, Italy
EOFS: European Open File Systems
EPFL: École Polytechnique Fédérale de Lausanne, Switzerland. An application partner in the DEEP project bringing CoreBluron, a brain simulation code.
EPiGRAM: Exascale ProGRAmming Models
eQPACE: European project to develop global communications for the QPACE architecture
ETP4HPC: European Technology Platform for High Performance Computing
EU: European Union
EUROPLANET: A European Research Infrastructure for Planetary Science (FP7)
Eurotech: Eurotech S.p.A., Amaro, Italy
EXA2CT: EXascale Algorithms and Advanced Computational Techniques
Exaflop: 10^{18} Floating point operations per second
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

FhGFS: Fraunhofer Global File system, a high-performance parallel I/O system to be adapted to the extended DEEP Architecture and optimised for the DEEP-ER Prototype

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

G

GCS: Gauss Centre for Supercomputing, The alliance of the three national supercomputing centres in Germany (Garching, Jülich and Stuttgart)
GEM: Geospace Environment Modelling
GMRES: Generalized Minimal RESidual method
GPU: Graphics Processing Unit
GREEN500 list: Provides rankings of the 500 top most energy-efficient supercomputers in the world
GridMonitor: Part of the ParaStationV5 cluster suite is a versatile system monitor for Linux-based compute cluster
GRS: German Research School for Simulation Sciences GmbH, Aachen and Juelich, Germany

H

H4H: Hybrid programming For Heterogeneous architectures (EU project)
H5hut: Library implementing several data models for particle-based simulations that encapsulates the complexity of parallel HDF5.
HDF5: Hierarchical Data Format: A set of file formats and libraries designed to store and organize large amounts of numerical data
Helmholtz Association: German research organisation
Healthchecker: Part of the ParaStationV5 cluster suite, A test suite to ensure the usability of compute and service nodes within a compute cluster
HMC: Hybrid Memory Cube
HMCC: Hybrid Memory Cube Consortium
HOPSA: HOlistic Performance System Analysis (EU-Russia FP7 project)
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
HTc: High critical temperature superconductors
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
ICT: Information and Communication Technologies
IEEE: Institute of Electrical and Electronics Engineers
IESP: International Exascale Software Project
INFN: Istituto Nazionale di Fisica Nucleare, Italy
Intel: Intel Germany GmbH Feldkirchen,
IP: Intellectual Property
IPC: Instructions Per Cycle
iPic3D: Programming code developed by the University of Leuven to simulate space weather
I/O: Input/Output. May describe the respective logical function of a computer system or a certain physical instantiation
IRST: Institute of Scientific and Technologic Research, Trento, Italy
ISC: International Supercomputing Conference, Yearly conference on supercomputing which has been held in Europe since 1986
ITEA-2: Strategic pan-European programme for advanced pre-competitive R&D in software for Software-intensive Systems and Services
ITWM: Institut für Techno- und Wirtschaftsförderung. An Institute of the Fraunhofer Society

J

JUBE: Jülich Benchmarking Environment
JUDGE: Juelich Dedicated GPU Environment: A cluster at the Juelich Supercomputing Centre
JUELICH: Forschungszentrum Jülich GmbH, Jülich, Germany

K

KNC: Knights Corner, Code name of a processor based on the MIC architecture. Its commercial name is Intel® Xeon PhiTM.
KNF: Knights Ferry, Intel first available processor based on the MIC
KNL: Knights Landing, second generation of Intel® Xeon PhiTM
KPI: Key Performance Indicator
KULeuven: Katholieke Universiteit Leuven, Belgium

L

LASA: Leuven Centre for Aero & Space Science, Technology and Applications, Brussels, Belgium
LGFMC: Lattice Green Function Monte-Carlo

LINPACK: Software library to perform numerical linear algebra calculations used as benchmark
LLNL: Lawrence Livermore National Laboratory
LMCC: Leuven Mathematical Modelling & Computational Science Centre, Belgium

M

MEW: Machine Evaluation Workshop
MIC: Intel Many Integrated Core architecture
MIC-OS: Operating System of the MIC architecture
MIUR: Ministry of Education, University and Research, Italy
MKL: Math Kernel Library
MMM@HPC: Project of Multiscale materials modelling on high performance computer architectures
Mont-Blanc: European scalable and power efficient HPC platform based on low-power embedded technology
Mont-Blanc 2: Follow-up project of Mont-Blanc
MPI: Message Passing Interface, API specification typically used in parallel programs that allows processes to communicate with one another by sending and receiving messages
MR-IOV: Multi-root I/O virtualisation. Standard to share a PCI Express endpoint between multiple hosts
MTBF: Mean Time Between Failures

N

NAM: Network Attached Memory, nodes connected by the DEEP-ER network to the DEEP-ER BN and CN providing shared memory buffers/caches, one of the extensions to the DEEP Architecture proposed by DEEP-ER
NAND Flash memory: Implementation of non-volatile memory used today for solid state disk.
NASA: National Aeronautics and Space Administration, Washington, USA
NetCDF: Network Common Data Form. A set of software libraries and data formats that support the creation, access, and sharing of array-oriented scientific data
NIC: Network Interface Card, Hardware component that connects a computer to a computer network
NSF: National Science Foundation, USA
NTB: Non-transparent bridge. A component required to connect PCI hierarchies
NUMA: Non-Uniform Memory Access
Numexas: NUMerical Methods and Tools for Key EXAScale Computing Challenges in Engineering and Applied Sciences
NVM: Non-Volatile Memory. Used to describe a physical technology or the use of such technology in a non-block-oriented way in a computer system
NVMe: Short form of NVM-Express
NVM-Express: An interface standard to attach NVM to a computer system. Based on PCI Express it also standardises high level HW interfaces like queues.

O

OEM: Original Equipment Manufacturer. Term used for a company that commercialises products out of components delivered by other companies.

OGS: Institute of Oceanography and Experimental Geophysics, Italy

OmpSs: BSC's Superscalar (Ss) for OpenMP

OpenMP: Open Multi-Processing, Application programming interface that support multiplatform shared memory multiprocessing

OS: Operating System

P

PA: Physical address space. Used on hardware level to access system components

ParaStation Consortium: Involved in research and development of solutions for high performance computing, especially for cluster computing

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

Paraver: Performance analysis tool developed by BSC

Paraview: Open Source multiple-platform application for interactive, scientific visualisation

ParTec: ParTec Cluster Competence Center GmbH, Munich, Germany

PATC: PRACE Advanced Training Centers

PC: Personal Computer

PCB: Printed circuit board.

PCH: Platform controller hub. Companion device to provide commodity peripherals to Intel® CPUs

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.

PCISIG: PCI special interest group. Industry association responsible for the development of the PCI/PCI Express standards

PCM: Phase change memory. A technology candidate for future non-volatile memories

PFlop/s: Petaflop, 10^{15} Floating point operations per second

PLL: Phase-locked loop. A control system that generates an output signal whose phase is related to the phase of an input signal. Used to demodulate a signal, recover a noisy signal, frequency synthesis and distribution of precisely timed clock pulses.

PLX: Provider of PCI Express system components

PM: Person Month or Project Manager of the DEEP project (depending on the context)

PMT: Project Management Team of the DEEP-ER project

POSIX: Portable Operating System Interface. A family of standards specified by the IEEE for maintaining compatibility between operating systems.
PPF: Poly-Phase Filter
PR: Public Relations
PRACE: Partnership for Advanced Computing in Europe (EU project, European HPC infrastructure)
PRACE-1IP: PRACE First Implementation Phase (EU project)
PRACE-2IP: PRACE Second Implementation Phase (EU project)
PRACE-3IP: PRACE Third Implementation Phase (EU project)
Project Coordinator: Leading scientist coordinating and representing the DEEP-ER project
PROSPECT: Promotion of Supercomputing Partnerships for European Competitiveness and Technology (registered association, Germany)

Q

QCD: Quantum Chromodynamics
QCDOC: Quantum ChromoDynamics On a Chip, special supercomputer developed by Universities of Edinburgh, Columbia and by IBM
QMC: Quantum Monte-Carlo
QPACE: QCD Parallel Computing Engine. Specialised supercomputer for QCD Parallel Computing

R

Rack: Compartment to mechanically assemble multiple chassis to form the final computer
RAID: Redundant Array of Independent Discs
RAM: Random-Access Memory
RDMA: Remote Direct Memory Access
RFI: Radio Frequency Interference
RMA: Remote Memory Access. Functional unit implemented in the EXTOLL NIC; used to implement highly optimized one-sided communication
RML: Risk management list used in the DEEP-ER project
R&D: Research and Development
RTD: Research and Technological Development

S

SC: International Conference for High Performance Computing, Networking, Storage, and Analysis, organised in the USA by the Association for Computing Machinery (ACM) and the IEEE Computer Society
Scalasca: Performance analysis tool developed by JUELICH and GRS
SCR: Scalable Checkpoint/Restart. A library from LLNL
SDV: Software Development Vehicle: a HW system to develop software in the time frame where the DEEP-ER Prototype is not yet available.
SEO: Search Engine Optimisation

SERDES Serializer/Deserializer functional block converting convert data between serial and parallel interfaces; used for data transmission over a single differential line.
SIMD: Single Instruction Multiple Data
SIONlib: Parallel I/O library developed by Forschungszentrum Jülich
SISSA: International School of Advanced Studies, Trieste, Italy
SKA: Square Kilometre Array
SM-Bus: Single-ended simple two-wire bus derived from I2C for the purpose of lightweight communication often used management of computer system components.
SME: Small and Medium Enterprises
SMFU: Shared Memory Functional Unit. Used by the EXTOLL network for mapping remote memory regions.
SoA: Struct of Arrays
SOTERIA: Project for the collection, organisation and the use of space physics data aimed at better understanding space weather (EU project)
SPR: Scientific Project Representative of the DEEP-ER Project
SRA: Strategic Research Agenda prepared by ETP4HPC
SR-IOV: Single Root I/O Virtualization: enables access to a PCI Express device from multiple virtualized guest operating systems.
SSD: Solid State Disk
StarSs: Generic programming environment developed by BSC
STRATOS: PRACE advisory group to foster development of HPC technologies in Europe
SW: Software
SWIFF: Space Weather Integrated Forecasting Framework, Leuven

T

TCO: Total Cost of Ownership
TDP: Thermal Design Power, a value describing the thermal limits of a computer system
TEXT: Towards Exaflop Applications (EU project)
TFlop/s: Teraflop, 10^{12} Floating point operations per second
Tier-0, Tier-1, …: Different classes of supercomputers ordered by their performance
TK: Task, Followed by a number, term to designate a task inside a work package of the DEEP-ER project
TLP: Transaction layer packet. Basic packet structure to transport transactions across a PCI Express infrastructure
ToW: Team of Work Package leaders within the DEEP-ER project
TP10: Third Party under special clause 10
TurboRVB: Quantum Monte Carlo Software for electronic structure calculations developed by SISSA

U

UHEI: University of Heidelberg, Germany
UREG: University of Regensburg, Germany
UPC: Universitat Politècnica de Catalunya. Barcelona, Spain

V

VMC: Variational Monte-Carlo
VELO: Functional unit and protocol implemented in an EXTOLL NIC; used to implement highly optimized two-sided communication
VI-HPS: Virtual Institute for High Productivity Supercomputing
VF: Virtual function. A functional element of a PCI endpoint
VTune: Commercial application for software performance analysis

W

WAN: Wide Area Network
WP: Work Package

X

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

Y

Z

ZITI Heidelberg: Institut für Technische Informatik Uni Heidelberg, Germany