SEVENTH FRAMEWORK PROGRAMME
Research Infrastructures

INFRA-2012-2.3.1 – Third Implementation Phase of the European High Performance Computing (HPC) service PRACE

PRACE-3IP

PRACE Third Implementation Phase Project

Grant Agreement Number: RI-312763

D8.3.4
Technical lessons learnt from the implementation of the joint PCP for PRACE-3IP

Final

Version: 1.5
Author(s): Stephen Booth, EPCC
Date: 10.01.2018

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: RI-312763**	
	Project Title: PRACE Third Implementation Phase Project	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: **D8.3.4**	
	Deliverable Nature: <Report>	
	Deliverable Level: PU	**Contractual Date of Delivery:** 31 / 12 / 2017
		Actual Date of Delivery: 15 / 01 / 2018
	EC Project Officer: Leonardo Flores Añover	

* - 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

Document	**Title: Technical lessons learnt from the implementation of the joint PCP for PRACE-3IP**	
	ID: **D8.3.4**	
	Version: 1.5	**Status:** ***Final***
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2010	
	File(s): D8.3.4.docx	
Authorship	**Written by:**	Stephen Booth, EPCC
	Contributors:	Fabio Affinito, CINECA Eric Boyer, GENCI Carlo Cavazzoni, CINECA Pekka Manninen, CSC Dirk Pleiter, JUELICH Philippe Segers, GENCI
	Reviewed by:	Stéphane Requena, GENCI Florian Berberich, JUELICH
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	19/10/2017	Draft	Initial outline
0.2	20/10/2017	Draft	
0.3	27/10/2017	Draft	PCP section
0.4	10/11/2017	Draft	General lessons
0.5	28/11/2017	Draft	Technical background
0.6	11/12/2017	Draft	E4 sections

0.7	11/12/2017	Draft	ATOS-BULL sections
0.8	12/12/2017	Draft	User experiences
0.9	14/12/2017	Draft	Formatting
1.0	14/12/2017	Draft	User experience section
1.1	16/12/2017	Draft	Commits & corrections
1.2	22/12/2017	Draft	Editorial updates
1.3	09/01/2017	Pre-Final	Updates based on internal review reports
1.4	10/01/2017	Pre-Final	Error in Table 19 fixed
1.5	10/01/2018		Last corrections

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, Joint Pre Commercial Procurement, Energy Efficiency

Table of Contents

List of Figures

List of Tables

References and Applicable Documents

[1] ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems DARPA TR-2008-13

[2] “Mediterranean-stile job scheduler for supercomputers - do less when it's too hot!”, A. Borghesi, C. Conficoni, M. Lombardi et Al, 2015 International Conference on High Performance Computing & Simulation, HPCS 2015, Amsterdam, Netherlands, July 20-24, 2015.

[3] D7.7 Performance and energy metrics on PCP systems, PRACE-4IP Deliverable

[4] http://www.eocoe.eu/events/scientific-applications-towards-exascale

[5] The bi-annually published Top500 list (https://www.top500.org) comprises the world’s 500 fastest HPC systems in terms of throughput of floating-point operations while executing the High Performance Linpack benchmark.

List of Acronyms and Abbreviations

AAA	Authorization, Authentication, Accounting.
ACF	Advanced Computing Facility
ADP	Average Dissipated Power
AISBL	Association International Sans But Lucratif (legal form of the PRACE-RI)
AMD	Advanced Micro Devices
APGAS	Asynchronous PGAS (language)
API	Application Programming Interface
APML	Advanced Platform Management Link (AMD)
ASIC	Application-Specific Integrated Circuit
ATI	Array Technologies Incorporated (AMD)
BAdW	Bayerischen Akademie der Wissenschaften (Germany)
BCO	Benchmark Code Owner
BLAS	Basic Linear Algebra Subprograms
BSC	Barcelona Supercomputing Center (Spain)
CAF	Co-Array Fortran
CAL	Compute Abstraction Layer
CCE	Cray Compiler Environment
ccNUMA	cache coherent NUMA
CEA	Commissariat à l'énergie atomique et aux énergies alternatives
CGS	Classical Gram-Schmidt
CGSr	Classical Gram-Schmidt with re-orthogonalisation
CINECA	Consorzio Interuniversitario, the largest Italian computing centre (Italy)
CINES	Centre Informatique National de l'Enseignement Supérieur (represented in PRACE by GENCI, France)
CLE	Cray Linux Environment
CoE	Center of Execellence
CPU	Central Processing Unit
CSC	Finnish IT Centre for Science (Finland)
CSCS	The Swiss National Supercomputing Centre (represented in PRACE by ETHZ, Switzerland)
CSR	Compressed Sparse Row (for a sparse matrix)
CUDA	Compute Unified Device Architecture (NVIDIA)
DARPA	Defense Advanced Research Projects Agency
DDN	DataDirect Networks
DDR	Double Data Rate
DEISA	Distributed European Infrastructure for Supercomputing Applications. EU project by leading national HPC centres.
DGEMM	Double precision General Matrix Multiply
DIMM	Dual Inline Memory Module
DKRZ	Deutsches Klimarechenzentrum
DMA	Direct Memory Access
DNA	DeoxyriboNucleic Acid
DP	Double Precision, usually 64-bit floating point numbers
DRAM	Dynamic Random Access memory
EC	European Community
EEA	European Economic Area
EESI	European Exascale Software Initiative

EoCoE	Energy oriented Centre of Excellence
EoI	Expression of Interest
EP	Efficient Performance, e.g., Nehalem-EP (Intel)
EPCC	Edinburg Parallel Computing Centre (represented in PRACE by EPSRC, United Kingdom)
EPSRC	The Engineering and Physical Sciences Research Council (United Kingdom)
eQPACE	extended QPACE, name of the FZJ WP8 prototype
ETHZ	Eidgenössische Technische Hochschule Zuerich, ETH Zurich (Switzerland)
ESFRI	European Strategy Forum on Research Infrastructures; created roadmap for pan-European Research Infrastructure.
EX	Expandable, e.g., Nehalem-EX (Intel)
FC	Fiber Channel
FFT	Fast Fourier Transform
FHPCA	FPGA HPC Alliance
FP	Floating-Point
FPGA	Field Programmable Gate Array
FPU	Floating-Point Unit
FZJ	Forschungszentrum Jülich (Germany)
GASNet	Global Address Space Networking
GB	Giga (= 2^{30} ~ 10^9) Bytes (= 8 bits), also GByte
Gb/s	Giga (= 10^9) bits per second, also Gbit/s
GB/s	Giga (= 10^9) Bytes (= 8 bits) per second, also GByte/s
GCS	Gauss Centre for Supercomputing (Germany)
GDDR	Graphic Double Data Rate memory
GÉANT	Collaboration between National Research and Education Networks to build a multi-gigabit pan-European network, managed by DANTE. GÉANT2 is the follow-up as of 2004.
GENCI	Grand Equipement National de Calcul Intensif (France)
GFlop/s	Giga (= 10^9) Floating point operations (usually in 64-bit, i.e. DP) per second, also GF/s
GHz	Giga (= 10^9) Hertz, frequency =10^9 periods or clock cycles per second
GigE	Gigabit Ethernet, also GbE
GLSL	OpenGL Shading Language
GNU	GNU's not Unix, a free OS
GPGPU	General Purpose GPU
GPU	Graphic Processing Unit
GS	Gram-Schmidt
GWU	George Washington University, Washington, D.C. (USA)
HBA	Host Bus Adapter
HCA	Host Channel Adapter
HCE	Harwest Compiling Environment (Ylichron)
HDD	Hard Disk Drive
HE	High Efficiency
HET	High Performance Computing in Europe Taskforce. Taskforce by representatives from European HPC community to shape the European HPC Research Infrastructure. Produced the scientific case and valuable groundwork for the PRACE project.
HMM	Hidden Markov Model
HMPP	Hybrid Multi-core Parallel Programming (CAPS enterprise)

HP	Hewlett-Packard
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HPCC	HPC Challenge benchmark, http://icl.cs.utk.edu/hpcc/
HPCS	High Productivity Computing System (a DARPA program)
HPL	High Performance LINPACK
HT	HyperTransport channel (AMD)
HWA	HardWare accelerator
IB	InfiniBand
IBA	IB Architecture
IBM	Formerly known as International Business Machines
ICE	(SGI)
IDRIS	Institut du Développement et des Ressources en Informatique Scientifique (represented in PRACE by GENCI, France)
IEEE	Institute of Electrical and Electronic Engineers
IESP	International Exascale Project
IL	Intermediate Language
IMB	Intel MPI Benchmark
I/O	Input/Output
IOR	Interleaved Or Random
IPMI	Intelligent Platform Management Interface
ISC	International Supercomputing Conference; European equivalent to the US based SC0x conference. Held annually in Germany.
IWC	Inbound Write Controller
JSC	Jülich Supercomputing Centre (FZJ, Germany)
KB	Kilo (= 2^{10} ~10^3) Bytes (= 8 bits), also KByte
KTH	Kungliga Tekniska Högskolan (represented in PRACE by SNIC, Sweden)
LBE	Lattice Boltzmann Equation
LINPACK	Software library for Linear Algebra
LLC	
LLNL	Laurence Livermore National Laboratory, Livermore, California (USA)
LQCD	Lattice QCD
LRZ	Leibniz Supercomputing Centre (Garching, Germany)
LS	Local Store memory (in a Cell processor)
MB	Mega (= 2^{20} ~ 10^6) Bytes (= 8 bits), also MByte
MB/s	Mega (= 10^6) Bytes (= 8 bits) per second, also MByte/s
MDT	MetaData Target
MFC	Memory Flow Controller
MFlop/s	Mega (= 10^6) Floating point operations (usually in 64-bit, i.e. DP) per second, also MF/s
MGS	Modified Gram-Schmidt
MHz	Mega (= 10^6) Hertz, frequency =10^6 periods or clock cycles per second
MIPS	Originally Microprocessor without Interlocked Pipeline Stages; a RISC processor architecture developed by MIPS Technology
MKL	Math Kernel Library (Intel)
ML	Maximum Likelihood
Mop/s	Mega (= 10^6) operations per second (usually integer or logic operations)
MoU	Memorandum of Understanding.
MPI	Message Passing Interface
MPP	Massively Parallel Processing (or Processor)

MPT	Message Passing Toolkit
MRAM	Magnetoresistive RAM
MTAP	Multi-Threaded Array Processor (ClearSpead-Petapath)
mxm	DP matrix-by-matrix multiplication mod2am of the EuroBen kernels
NAS	Network-Attached Storage
NCF	Netherlands Computing Facilities (Netherlands)
NDA	Non-Disclosure Agreement. Typically signed between vendors and customers working together on products prior to their general availability or announcement.
NoC	Network-on-a-Chip
NFS	Network File System
NIC	Network Interface Controller
NUMA	Non-Uniform Memory Access or Architecture
OpenCL	Open Computing Language
OpenGL	Open Graphic Library
Open MP	Open Multi-Processing
ORNL	Oak Ridge National Laboratory
OS	Operating System
OSS	Object Storage Server
OST	Object Storage Target
PCIe	Peripheral Component Interconnect express, also PCI-Express
PCI-X	Peripheral Component Interconnect eXtended
PGAS	Partitioned Global Address Space
PGI	Portland Group, Inc.
pNFS	Parallel Network File System
POSIX	Portable OS Interface for Unix
PPE	PowerPC Processor Element (in a Cell processor)
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PSNC	Poznan Supercomputing and Networking Centre (Poland)
QCD	Quantum Chromodynamics
QCDOC	Quantum Chromodynamics On a Chip
QDR	Quad Data Rate
QPACE	QCD Parallel Computing on the Cell
QR	QR method or algorithm: a procedure in linear algebra to compute the eigenvalues and eigenvectors of a matrix
RAM	Random Access Memory
RDMA	Remote Data Memory Access
RISC	Reduce Instruction Set Computer
RNG	Random Number Generator
RPM	Revolution per Minute
SAN	Storage Area Network
SARA	Stichting Academisch Rekencentrum Amsterdam (Netherlands)
SAS	Serial Attached SCSI
SATA	Serial Advanced Technology Attachment (bus)
SDK	Software Development Kit
SGEMM	Single precision General Matrix Multiply, subroutine in the BLAS
SGI	Silicon Graphics, Inc.
SHMEM	Share Memory access library (Cray)
SIMD	Single Instruction Multiple Data
SM	Streaming Multiprocessor, also Subnet Manager
SME	Small and Medium Sized Enterprise

SMP	Symmetric MultiProcessing
SNIC	Swedish National Infrastructure for Computing (Sweden)
SP	Single Precision, usually 32-bit floating point numbers
SPE	Synergistic Processing Element (core of Cell processor)
SPH	Smoothed Particle Hydrodynamics
SPU	Synergistic Processor Unit (in each SPE)
SSD	Solid State Disk or Drive
STFC	Science and Technology Facilities Council (represented in PRACE by EPSRC, United Kingdom)
STRATOS	PRACE advisory group for STRAtegic TechnOlogieS
STT	Spin-Torque-Transfer
SURFsara	Dutch national High Performance Computing & e-Science Support Center
TARA	Traffic Aware Routing Algorithm
TB	Tera (= 240 ~ 1012) Bytes (= 8 bits), also TByte
TCO	Total Cost of Ownership. Includes the costs (personnel, power, cooling, maintenance, ...) in addition to the purchase cost of a system.
TDP	Thermal Design Power
TFlop/s	Tera (= 1012) Floating-point operations (usually in 64-bit, i.e. DP) per second, also TF/s
Tier-0	Denotes the apex of a conceptual pyramid of HPC systems. In this context the Supercomputing Research Infrastructure would host the Tier-0 systems; national or topical HPC centres would constitute Tier-1
UFM	Unified Fabric Manager (Voltaire)
UNICORE	Uniform Interface to Computing Resources. Grid software for seamless access to distributed resources.
UPC	Unified Parallel C
UV	Ultra Violet (SGI)
VHDL	VHSIC (Very-High Speed Integrated Circuit) Hardware Description Language

Executive Summary

Energy efficiency is one of the key challenges affecting many parts of the IT industry including High Performance Computing (HPC). The aim of the PRACE joint Pre Commercial Procurement (PCP), between CINECA (Italy), CSC (Finland), EPCC (UK), GENCI (France) and JSC (Germany) was to explore, for the first time in Europe in the field of HPC, to what extent pre-commercial procurement could be used to improve the energy efficiency of general purpose supercomputers capable of running real PRACE production workloads. This is a very challenging aim. The existing HPC marketplace already takes energy efficiency fairly seriously (at least for the largest systems) and many of the requirements for energy efficiency are the same as requirements for high performance. As a result, there were no major inefficiencies in current offerings that the project could address in order to obtain an easy win. In addition, the key technologies that consume most of the energy such as processors and memory are manufactured by large down-stream suppliers with a wide user base and very large R&D costs making them unlikely to be directly influenced by a project of this kind with an overall budget of 9 Mio. EUR, so the project was more targeting a leverage effect.

This project has demonstrated that it is possible to achieve better energy efficiency by relaxing the requirement for a general purpose supercomputer. Better energy efficiency is possible but comes at a cost of either making the system harder to use or only suitable for a smaller range of applications. Usually this is by exploiting some form of accelerator technology. The project explored a number of such architectures including large numbers of low-power cores (Intel KNL processor), GPGPU acceleration and data-flow architectures implemented using FPGAs. The KNL based solution was the most general purpose system. The GPGPU system gave a very good energy efficiency when the application codes supported the GPGPU properly. This is quite an interesting result; though GPGPUs will not be suitable for all types of problem many major application codes now have good support for GPGPUs. The data-flow/FPGA solution demonstrated the potential for even greater energy efficiency in some cases though the effort required to utilise this technology implies code re-writing sections, so such an approach may be better suited to dedicated usages, rather than general purpose systems.

Energy efficient whole HPC solutions requires also energy efficient power supplies and cooling systems. High end HPC system vendors typically include these technologies as part of their standard offering however the rest of the IT industry is still dominated by air-cooled solutions which could limit the potential bidders to procurements that require them. The PRACE PCP has demonstrated that a PCP type mechanism can be used to fund the development of additional products with state-of-the-art power and cooling solutions increasing competition in this area. As matter of fact, vendor roadmaps have been influenced by this PCP as reported in “Impact on vendor roadmap” of Atos-Bull, Maxeler and E4, respectively in 4.5, 5.5 and 6.5.

The project has also demonstrated the potential of energy aware scheduling and whole-system power capping that might be used to allow a system to adjust its operation to fit within local infrastructure power and cooling constraints. This will be also very important tools in a future exascale context were monitoring will be even more complex and critical.

To allow applications to be optimized with respect to their energy use it is essential to have high resolution energy profiling tools that will allow application developers to investigate the energy usage of their applications including sufficient time resolution to distinguish the relative energy usage of different parts of the application. The project has developed a number of powerful tools and instrumentation to allow this to happen, such tools are now expected to become part of the software stacks of the different vendors involved into the PRACE PCP.

1 Introduction

This document collates the technical lessons learned from the PRACE PCP. We start with a review of the aims of the PRACE PCP and the overall technical landscape, as it is relevant to energy efficiency. Lessons are then drawn from each of the pilot systems delivered in the final phase of the PCP. Feedback and experiences from early users of the machines is analysed. Where additional lessons could be drawn, selected systems from earlier stages of the PCP are also considered as well as general lessons that are not specific to any particular system.

2 Overview of the PRACE PCP

The PRACE-3IP project performed a Pre-Commercial-Procurement (PCP). A PCP is primarily targeted towards the procurement of R&D services, but allows for a part of the budget to be spent on the procurement of equipment for demonstrating and testing the R&D results. The PRACE PCP was setup with the following technical goals: to procure/develop highly energy efficient HPC systems capable of general use, i.e. able to run real applications, be operated within a conventional HPC computing centre but nevertheless achieve very high total-system energy efficiency. In addition to the technical goals the PCP was also intended to develop the HPC vendor eco-system within the EEA and as such it is expected to result in commercially viable products.

The technical criteria were chosen as they reflect one of the most challenging issues facing the HPC market sector at the moment. It has been well recognised since at least 2008[1] that total energy consumption will be one of the major factors that could inhibit the adoption of Exascale systems. This will also be a major problem for smaller scale HPC systems. Though smaller HPC installations typically only are a fraction of the size of world-leading systems they will also be installed in facilities with only a fraction of the available power. Power consumption is already a significant part of the total cost of ownership of a HPC system. Current HPC procurements frequently consider energy efficiency as part of the evaluation process. The relative weighting of energy efficiency in procurements is expected to increase over time. Energy efficiency is also a major concern of the wider IT industry. At one end of the IT spectrum mobile computing requires high energy efficiency to maximise battery life. At the other end of the spectrum hyper-scale data centres as used by public cloud providers have many technologies in common with the HPC sector and share many of the same power and cooling issues.

The PRACE PCP did not place any restrictions on how energy efficiency was to be addressed. The systems were to be evaluated on total-system energy efficiency so vendors were free to address any aspects of the total system design.

Aligning the goals of the PCP with one of the major issues facing the HPC market sector has a number of results.

- First of all, this ensures that the research and development undertaken as part of the PCP will be commercially relevant. A PCP mechanism could be used to encourage vendors to develop new products to serve new user communities with specialised requirements. In that case it would be beneficial for the R&D activity to be run as co-design to bring together the domain knowledge of the end user with the technical knowledge of the vendor. In this case the vendors should have easily understood the requirements so there was no need for any co-design involving the procuring entities. However, this close alignment with a major issue facing a competitive industry such as IT does make it extremely challenging to make significant technical breakthroughs.

- Much of the energy budget of an HPC system goes towards components such as processors, memory and networks. The scale of the industry sectors that manufacture these is so large that we cannot expect that a € 9M PCP to have any significant direct influence over them but we could with some confidence expect them to be addressing energy efficiency in some form. In addition any easily exploitable route to energy efficiency specific to HPC will probably already have been explored by some of the existing HPC vendors.
- This does not prevent the PCP from acting as a mechanism to evolve the vendor ecosystem. Vendors could either have used the PCP as a way to develop new products in order to compete on equal terms with existing major players or to facilitate the development of more novel approaches that have not been adopted elsewhere. However, we have to accept that novel approaches with significant improvements in energy efficiency will probably come at some cost, for example systems that are harder to program or harder to use for some reason. As a result of this any vendor bidding for the PCP had to make an assessment of the market viability of these costs relative to the energy efficiency benefits when designing their solution so we would not expect the same levels of innovation as might be seen in a purely academic research project.

3 Technical Background

To be able to increase the energy efficiency of HPC it is necessary to either identify some inefficiency in current implementations that may be addressed or to entirely replace one of the current technologies used with a replacement technology that has a significantly energy cost. A disruptive change in the technologies used by the HPC industry would require significantly more funding than is available within this PCP so it is important to understand the current technical landscape and where inefficiencies might exist that could be exploited to improve overall energy efficiency.

To a first approximation the significant contributors to energy consumption in HPC compute nodes are CPU/logic, memory and data-movement. The CPU consumes the majority of the power. There is some scope within current CMOS technology to optimise for power efficiency. The performance of a processor is roughly proportional to the clock-speed and the degree of parallelism supported by the processor. However, increasing the clock-speed requires an increase in supply voltage and the dynamic power requirements of the processor are roughly proportional to the square of the supply voltage so increasing the clock-speed has a disproportionate negative impact on energy efficiency. Many modern processors support dynamic changes in clock speed allowing the processor to switch between high performance and low energy modes. This is particularly effective for codes that are memory rather than CPU bound where there is less performance advantage to a high clock-speed. Highly energy efficient system designs therefore tend to be characterised by relatively modest clock frequencies and high levels parallelism. In practice the number of applications that can effectively utilise very large numbers of low performance nodes is limited so most recent designs use a combination of node-based, thread-based and instruction-level (SIMD/vector) parallelism. SIMD/vector architectures may also be more energy efficient for HPC work-loads because a higher proportion of the circuitry is used for implementing floating point pipelines, and hence the power budget is dedicated to floating point calculations than is the case for general purpose processor designs. On the other hand the investment needed to bring a new CPU (or even a customised version of an existing one) to market is huge, significantly greater than the research component of the PCP so at most the PCP could only influence the

selection of components from existing available devices rather than result in any form of novel low energy processor design.

Currently the main memory in all commercial HPC systems utilises some form of Dynamic Random Access Memory (DRAM). DRAM memory cells only retain their information for a limited period of time before the charge in the capacitor leaks away losing the information. Therefore DRAM cells need to be refreshed periodically by the memory controller. DRAM is a particularly cost effective technology with respect to storage capacity; the storage cells are extremely simple as each bit of storage only requires a single transistor and a capacitor to implement. This means that DRAM requires significantly fewer manufacturing steps than processor logic and can be manufactured at low cost. However, this cost advantage is removed if the DRAM cells are manufactured on the same wafer as the processor or any other complex logic so the memory controller is an external device (usually part of the processor). Though not the largest energy cost in a HPC system DRAM does consume a significant part of the overall energy budget most of this being data-movement between the DRAM chips and the processor itself. The simple constructions of DRAM chips means that the memory interfaces use fairly simple electrical interfaces, which are not particularly energy efficient. One significant technological change within the timescale of the PCP is the introduction of 3D stacked memory. Though still a form of DRAM the packaging is radically different. Previously DRAM chips were packaged as individual devices mounted on memory DIMMs and connected to the processor through the motherboard. With stacked memory a 3D stack of memory chips is constructed. The majority of the thickness of the silicon wafer is removed during manufacture to allow a large number of electrical connections to be manufactured through the body of the chip giving a very high degree of connectivity within the stack and therefore supporting greater parallelism and high bandwidth throughput. These through-chip connections are called Through-Silicon-Vias (TSVs). In additional this memory stack is installed in-package very close to the processor or memory controller. Though the primary motivation for 3D stacked DRAM is increased performance the much shorter communication distances may result in improved energy efficiency compared with conventional DRAM. One technology that follows this approach is HBM (High Bandwidth Memory), which is standardised by JDEC since 2013.[1] An alternative architecture (used by the Micron Hybrid Memory Cube) is to use a separate memory controller (again closely coupled with the DRAM stack) connected to the processor using high-speed-serial communication links. This architecture is also more energy efficient than convention DRAM and more flexible than directly connected HBM but at the cost of additional memory controller devices.

Data movement is one of the major consumers of energy in HPC systems. All current interconnects use the same underlying technologies. Though there are many different network products available the lowest level implementations (the physical transport layer of the OSI networking model) are essentially the same for all products. Over short distances electrical connections over copper cables are used, these are implemented over high speed serial links. The cost (both manufacturing costs and energy costs) of this technology are roughly proportional to the length of the connection. Over longer distances optical connections over fibre optic cables are used. In this case the costs (again both in terms of manufacturing and energy costs) are concentrated in the optical transceivers and are largely independent of distance so optical communications are the preferred technology over longer distances. The fundamental design principles of HPC networks are therefore very similar when optimising for manufacturing or energy costs. At most the optimal cross-over distance between electrical and optical communications might be different in an energy optimised design.

[1] https://www.jedec.org/document_search?search_api_views_fulltext=jesd235

The long term trend in interconnect technologies has been for the optimal electrical/optical cross-over distance to become shorter over time. There are significant R&D investments in silicon-photonics which have the potential to eliminate the copper cables entirely and connect fibre directly to silicon. This is a promising and potentially disruptive technology however, it is still in the research phase with some fundamental issues still to be resolved so it is not easy to predict a timescale when it might cross over into commercial products. When and if it does become available it may open up opportunities for innovative energy efficient designs.

Minimising the use of inter-node communication is a key part of parallel application performance the same optimisations used to improve performance and parallel scaling of applications are also those needed to improve the energy efficiency of the applications communications. Therefore we don't expect any major energy inefficiencies in the way that major HPC applications structure their inter-node communications.

Depending on the application there may be greater scope for energy efficiency improvements in data movement within memory systems. HPC applications are typically written to perform well in a memory hierarchy containing layers of cache memory. In this case good performance relies on ensuring a high proportion of cache hits rather than minimising the total amount of memory traffic. Techniques such as pipelining and pre-fetching allow the performance cost of memory accesses to be hidden by overlapping them with other operations but this does nothing to minimise the energy costs.

When using GPUs there is an additional time and energy cost in moving data between CPU and GPU memory. However, as this impacts performance as well as energy use well optimised GPU codes will already have been re-written to minimise these data movements. This is another reason why GPU systems may exhibit good energy efficiency.

Data-flow computers (as represented in the PCP by the Maxeler system) also attempt to minimise data movement. The aim of a data-flow system is to use re-configurable hardware to implement key computational kernels as a single pass through the data. Any data elements that are re-used within the kernel are retained in internal buffers until no longer needed.

The three pilot systems deployed by the PCP demonstrated three different responses to this technical landscape:

- The Intel KNL processor is an example of a many-core design using a large number of relatively simple CPU cores supplied with wide SIMD vector units. These therefore support a very high degree of parallelism with a large part of the available circuitry implementing the floating-point operations needed by HPC applications. This together with the 3D-stacked memory had the potential to improve energy efficiency while staying within the standard programming model of existing HPC systems to preserve the general-purpose nature of the solution.
- The NVIDIA P100 GPU accelerator uses 3D-stacked HBM2 and also uses wide vector instructions and high core counts internally. Using a GPGPU effectively requires a large application software investment. However the growing number of improving programming models has gained significant traction in recent years and large numbers of important HPC applications already support GPGPUs to some extent and there are a large number of existing application software engineers with the necessary skills.
- Using FPGAs to build application specific accelerators gives great flexibility to match circuit implementation to the requirements of the application. Constructing these accelerators as data-flow computers removes unnecessary memory traffic. However as with GPGPUs this requires a large application software investment and data-flow accelerators have not achieved the same market penetration as GPUs.

4 Lessons learned from the Atos-Bull KNL pilot system

4.1 Description of the system

4.1.1 *System description*

The Atos-Bull pilot system hosted by GENCI at CINES in Montpellier (France) has been deployed on Friday 14 April 2017. Hardware installation including cabling, powering and pipe connections was then finalized on 20 April 2017. The pilot system's hardware architecture is based on a standard Bull Sequana X1000 cell improved with specific PCP developments providing water cooled power supplies becoming the first installation with 100% water cooled racks.

For this PCP, Atos-Bull made the choice to use the last generation of Intel Xeon Manycore ("Knights Landing ") architecture (also known as Intel Xeon Phi) and especially the Intel 68-core 7250 and Mellanox InfiniBand EDR interconnect.

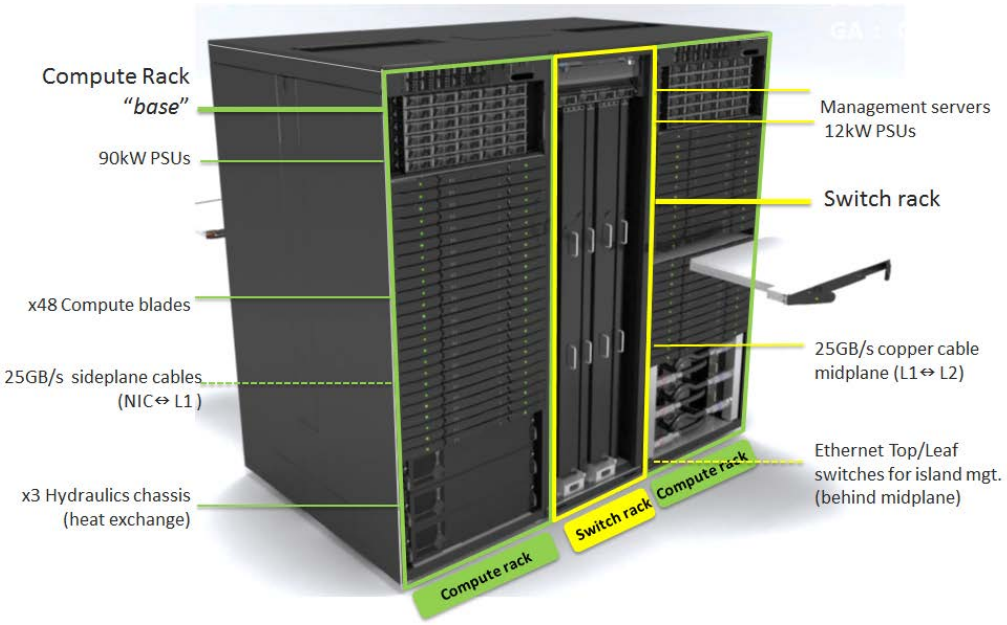

Figure 1: Sequana Cell view

A complete Sequana cell has been installed equipped with 56 Atos-Bull Sequana X1210 Intel Xeon Knights Landing (KNL) blades, providing a total of 168 compute nodes.

Each KNL blade contains three KNL nodes, and each compute node is equipped with:

- 1x Intel Xeon-Phi Knights Landing 7250 16GB HBM MCDRAM 215W,
- 6x 16GB@2400MT/s DDR4 DIMMs,
- 1x 240GB 2.5" 7mm SATA3 SSD,
- 1x InfiniBand 4x EDR mezzanine board to connect one 100Gb/s link to the first level of fat-tree in switch rack,
- 1x HDEEM (High Definition Energy Efficiency Monitoring) FPGA to provide accurate and high frequency power measurement. This component is key for the PCP project.

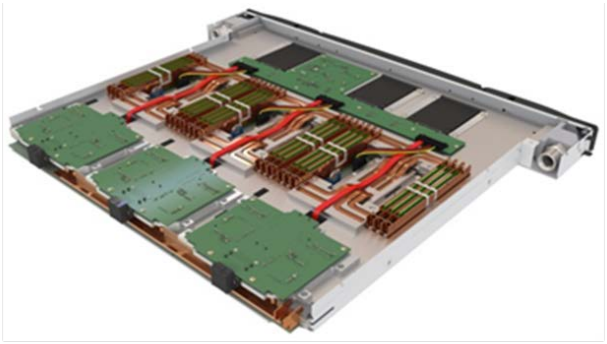

Figure 2: Sequana KNL blade with 3 KNL nodes

The high-speed interconnect is designed through a Mellanox InfiniBand EDR Fat Tree topology with a 2:1 blocking factor.

Storage facilities access is granted through direct connections to the InfiniBand fabric and existing data management solutions relying on Lustre or PanFS file systems and accessed through routers.

On the pilot system, MooseFS has been deployed as parallel and highly performing file system. It spreads data over all local disk of each compute nodes (chunk servers), which are visible to the user as one virtual disk. MooseFS is compliant and acts like any other Unix-like file system supporting:

- Hierarchical structure: Files and Folders;
- File attributes;
- Special files: Pipes, Sockets, Block and Character devices;
- Symbolic and Hard links;
- Security attributes and ACLs.

With this solution, it was demonstrated an IO throughput of 30 GB/s in both write and read using IOR benchmark.

4.1.2 *Software environment*

The Bull HPC software solution includes core components that provide tools, libraries and APIs to fulfil expectations of each kind of user:

- Administrators needing to install, configure and monitor all the physical and logical components of the solution to ensure maximum availability and performance,
- Developers needing tools to develop, and then manage code, analyse and tune applications;
- End-users needing efficient submission mechanisms and highly performing communications libraries.

The Bull Super Computer Suite version 5 (SCS 5) has been designed to address these different needs with simplicity and efficiency, maximizing stability of operational conditions and applications performance. Its main goals are to offer a set of autonomous components that can be used all together to create a complete solution, and to provide:

- a high-performance software environment for the supercomputer;
- an easy installation and modular update paths;
- the integration of hardware add-ons;
- quick security fixes;
- the support for several development environments;

- and an all-in-one solution for validating the system architectures whatever the size of the computer.

Based on selected foundation components, the hierarchical approach for managing thousands of equipment has been engineered and implemented to offer a scalable and easy-to-use environment. This solution is a completely new approach offering high level of resiliency and flexibility, from the installation to the day-to-day operations.

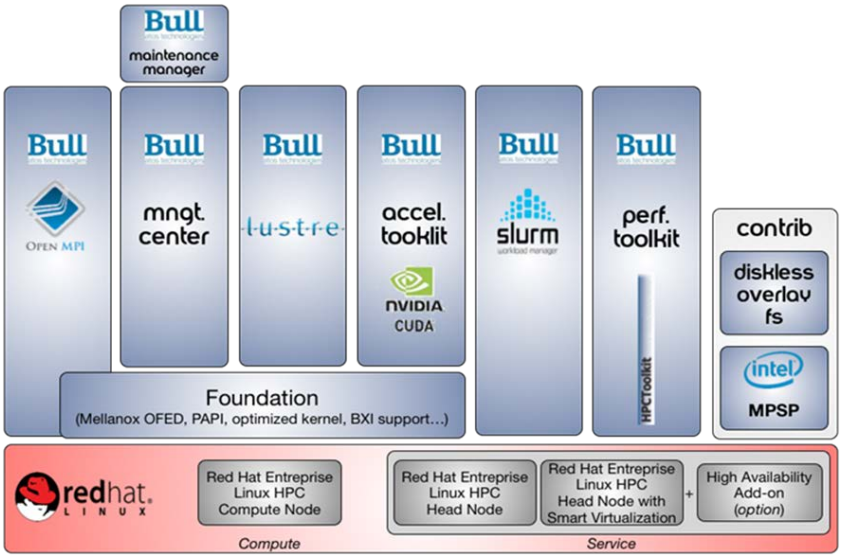

Figure 3: SCS5 Components

The Bull SCS 5 is based on a set of components that aim to provide software solution for different purposes on a HPC system. All those components are validated and built to perform at their maximum on a Red Hat Enterprise Linux HPC operating system.

- Bull Foundation
- Bull Management Center
- Bull Maintenance Manager
- Bull Lustre
- Bull SLURM
- Bull OpenMPI
- Bull Performance Toolkit
- Bull Accelerator Environment (NVIDIA CUDA)

The Bull SCS5 release 1 has been installed on the Pilot PCP. This release doesn't include any prototype of the tools developed during the PCP. However, such tools will be integrated into the next release of the solution.

4.1.3 *Energy efficiency aspects of the design*

Technology related

The Atos-Bull Sequana concept: The Sequana integration has a Power Usage Effectiveness (PUE) ratio of very close to 1 and energy consumption 10 times lower than the previous generation of supercomputers. With Sequana, 100 percent of the components - including compute nodes, power supplies and switches - are cooled using an enhanced version of the patented Atos-Bull Direct Liquid Cooling (DLC) technology. The second generation DLC solution is a proven cooling solution that minimizes a system's global energy consumption by using warm water at up to 40° C. This is an enhanced version of the proven DLC technology

in the Bullx DLC cabinets used with B700 series and already deployed at many large HPC sites, including DKRZ in Germany and Météo France.

The Sequana platform is an energy-aware system that integrates fine-grain energy sensors and a new generation of the High Definition Energy Efficiency Monitoring (HDEEM) technology to facilitate energy optimization.

Through these innovations, the Atos-Bull Sequana X1000 provides the following features:

- The lowest TCO and carbon footprint of any regular supercomputer available today;
- Enables power-aware scheduling to ensure that applications run as cost effectively as possible without compromising performance.

DLC minimizes the total energy consumption of a system because cooling is achieved with inlet water as warm as 40° C. Sequana X1000 is designed to support very powerful and energy-hungry nodes. In combination with peripherals placed in racks with water-cooled doors, this means that the proposed solution can extract more than 100 percent of the heat generated by the solution and do so with year-round free cooling. As an option, the DLC solution can be used to cool machine room air.

Sequana X1000 DLC solution is sized to evacuate the heat generated in compute nodes and BXI or EDR/HDR switching components, even with the most extreme configurations. Uniquely, it also use direct liquid cooling on the power supply units (PSUs); this capability enablingto extract the final 10 percent of heat generated by the system to water.

Water cooled Power Supply Units (PSU): In the initial Bull Sequana design, PSU were air-cooled and represented around 8% of total cell dissipation. During PCP phase II, a new PSU has been developed for PSU heat direct capture. Bull has first selected one partner "Brightloop Converters", a French SME specialized in high efficiency power converters, for the co-development of this PSU and the power shelves compatible with Sequana.

The objective is the full integration of two types of power shelves cooled by hot water in Sequana. Shelves will deliver up to 15,000 Watt of power at high performance capability and precise consumption measurement falling within the latest standards of the Green500[2].

The validation tests of the 12 kW sWitch Power Shelf Module (WPSM) and 15 kW Compute Power Shelf Module (CPSM) are shared by Bull and its partner. The main focus for Bull in Phase III is the integration in Sequana and the validation of final solution.

The two shelves (12 kW and 15 kW) are based on the assembly of 3kW individual convertors, thermally drained to a cold plate via the use of internal heat spreaders. Each 3 kW module contains its own fuse, PFC (Power Factor Corrector) stage, and a LLC resonant converter[3], to transform the input 230 VAC or 400 VDC to the main 54 V output.

These modules are mounted on a cold plate, and a backplane is used to interconnect them together and to the output connectors. An internal view of the 15 kW shelf is shown below. The 12 kW shelf is based on the same concept, except that it hosts only 4 modules, and that the external connectors are not located at the same place (and are different for the 54 V output).

[2] http://www.green500.org

[3] Resonant converters are a type of electric power converter that contains a network of inductors (L) and capacitors (C). LLC resonant converts are resonant converters with a specific topology and based on two inductors and one capacitor.

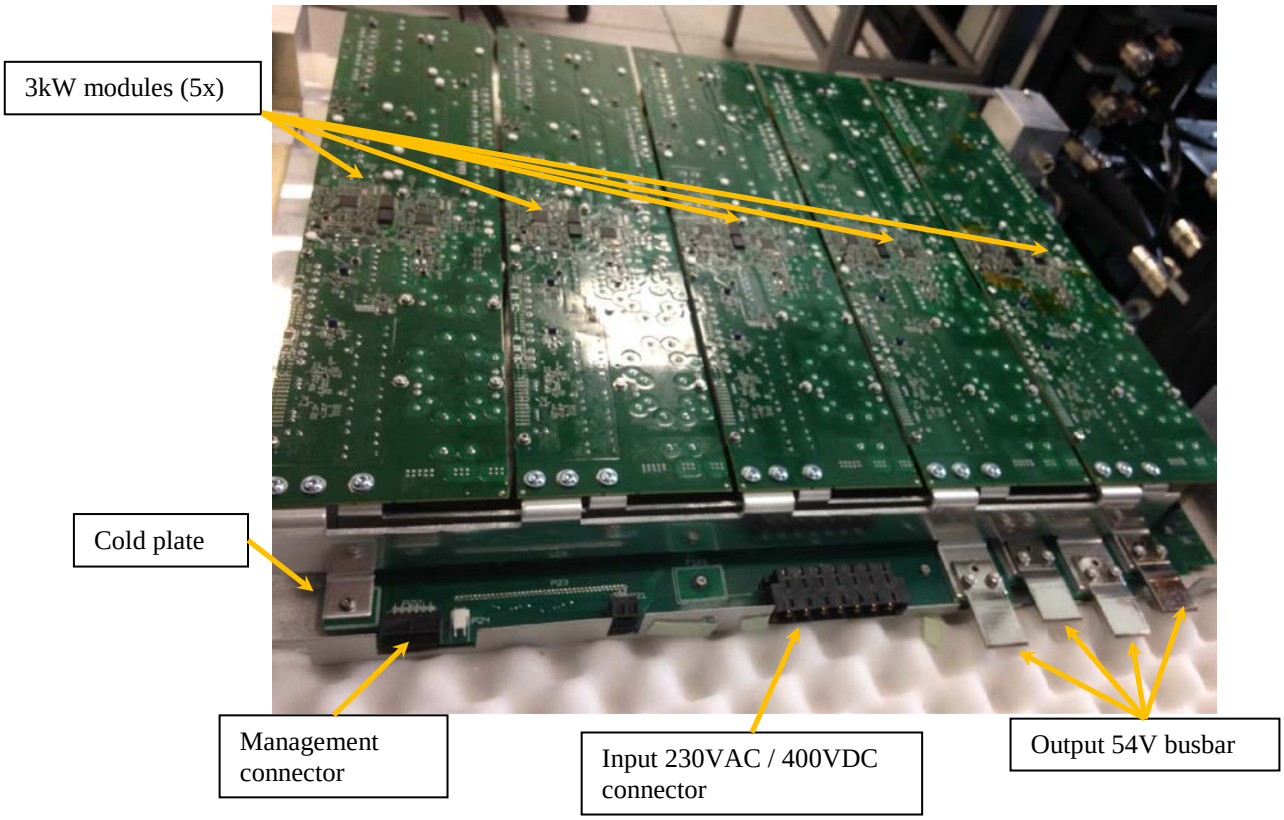

Figure 4: Internal view of the 15kW shelf

In each module, a dedicated DSP is used to control the primary side of the 3 kW module (PFC stage), and a second one is used for the secondary side (LLC stage). They are in charge of the regulation of the stage they are associated with. All DSP modules are managed by an additional DSP located on the shelf backplane. This last DSP (management DSP) is responsible for the management of the shelf, including power ON/OFF operation, monitoring, current sharing, and firmware upgrade. It communicates with the DSPs of the individual modules through a CAN bus, and with the rack power controller (PMC) through an I2C bus.

The required efficiency is higher than 94% at full load, which makes the shelf 80-Plus platinum certified.

The FAST_PROCHOT_N feature is handled at the shelf level by the management DSP, which asserts this signal low when the output power of the shelf exceeds 130% of its rated load. This is done very quickly (less than 100µs) so that the CPU can react fast enough (by reducing their drawn power) to avoid the PSU entering over power protection and shutting down.

To allow support for a local power backup source (ultra-capacitor modules) in case of short outages (up to 300 ms at full load, or 800 ms at half load), the shelf is required to restart in less than 150 ms upon input voltage recovery. Good load sharing and error handling is needed during this phase, so that no PSU exceeds its rated power, or enter any protection, while taking into account the tolerance in start time between all PSU of a same output rail (up to twenty 3 kW modules, e.g. 4 shelves).

Input power is monitored, at the shelf level, by the management DSP. It is based on the output power reported by dedicated sensors in each 3 kW module, and calculated using a pre-set look-up table to include the shelf efficiency. This one is characterized for different input voltages and temperatures conditions. The reported value is then filtered over a 100ms period of time. The obtained precision is required to be better than +/-3%, for loads ranging from 50% to 100% of full load.

Infrastructure for Energy monitoring and optimization

During this project, different prototypes of energy efficiency oriented tools and products have been developed. On the hardware side, Atos-Bull Sequana blades provide accurate power consumption measures at a high frequency embedding the High Definition Energy Efficiency Monitoring (**HDEEM**) technology developed by Atos-Bull.

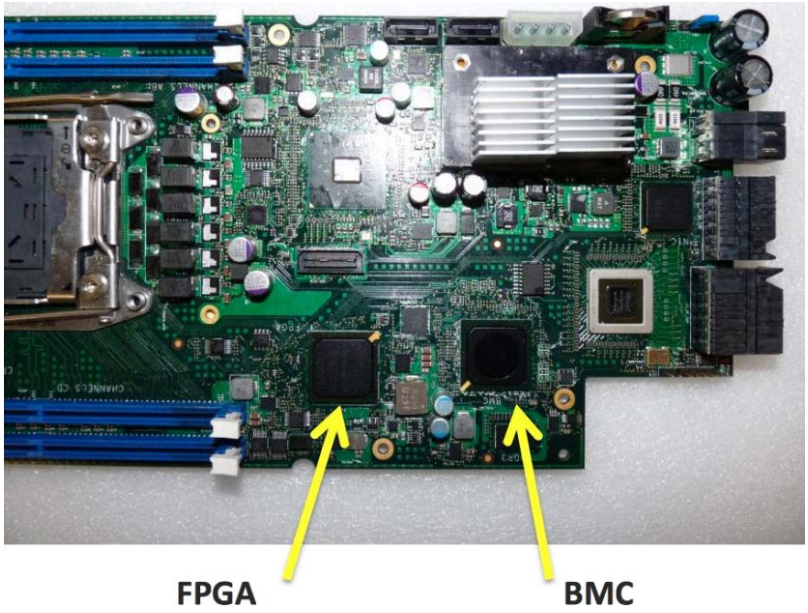

Figure 5: HDEEM board

More precisely, HDEEM provides

- a **sampling rate** up to:
  - 1 kHz for global power including sockets, DRAM, SSD and on-board,
  - 100 Hz for voltage regulators,
- And with an **high accuracy** with 2-5% of uncertainty after calibration,
  - 2% for blades,
  - 5% for VR,
- Different **collection modes** are available,
  - Out-of-band using BMC (4 Hz),
  - In-band through HDEEM API (eg SLURM).

One challenge of this PCP project was to develop tools able to collect HDEEM results as well as metrics coming from other components of the Sequana cells (switches for instance) and provide consolidated views of either jobs power consumption (for users) or the full system power consumption (for system administrators). As a result, the following components have been designed and developed within the PCP:

- Bull Energy Optimizer (**BEO**): Energy measurement tool that provides information on the whole system, which is non-intrusive, and provides100 Hz sampling and 3-5% precision,
- **HDEEViz**: Power consumption visualization tool,
- Energy oriented **SLURM** plugin based on adaptive scheduling,
- Bull Dynamic Power Optimizer (**BDPO**): Energy optimization tools able to dynamically adapt runtime parameters to save energy without significant impact on the execution time.

Description of the BEO component

The energy measurement sub-system that is integrated into the Pilot System relies on a new product that will be available in the Bull SCS 5 software suite. This product, Bull Energy Optimizer (BEO), is a software product dedicated to power management for HPC clusters.

Bull Energy Optimizer was originally referred as "Power Manager" module in the technical offer of phase III.

The various use cases that BEO aims at addressing can be grouped into four broad categories:

- Collecting data related to power consumption;
- Supporting diagnosis activities such as understanding the mechanisms that leads to a given power consumption level, or being able to identify where power is actually consumed;
- Predicting the behaviour of the system from the analysis of statistic data, depending on the run-time system configuration and application deployment options;
- Prescribing configuration changes according to power management policies, based on power consumption models.

This version of BEO addresses the two first use cases, it focuses on the descriptive aspects: being able to carefully monitor power and energy consumption, alerting when metrics thresholds are exceeded.

In future versions, BEO will include Predicting and Prescriptive actions. BEO will also include dynamic power optimizations features, where run-time environment can be automatically changed according to specific power consumption targets & power management policies.

BEO main features are:

- Providing power and energy consumption information related to any subset of the cluster. Such a subset is defined as a managed container that can group Compute Nodes, Switches, Chassis, Racks, Islets, and Cluster;
- Providing energy consumption related to any set of SLURM jobs;
- Providing alerting functionality based on thresholds definition on a given set of metrics *(ex: power limit for a set of hardware)*;
- Providing the ability to report energy costs, based on a configured description of power costs;
- Providing data as time series for a set of metrics;

These features are delivered as a Command Line Interface (CLI). Next releases will contain a REST API as well as a Graphical User Interface (GUI).

Additionally, it is possible to use the graphite Web interface to obtain a visual representation of the metrics via graphite/carbon. The following figure shows an example of the evolution of power metrics for a compute node:

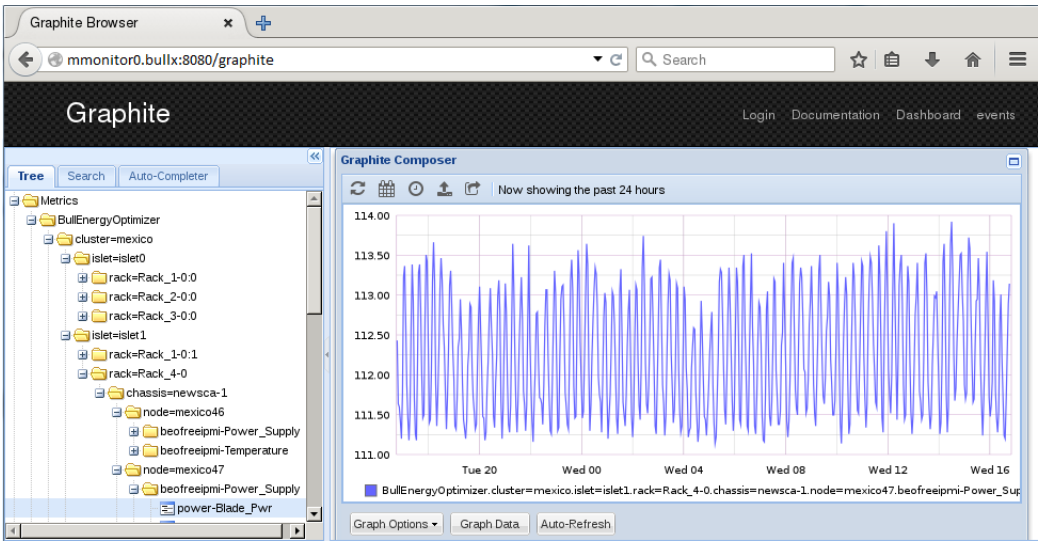

Figure 6: Evolution of power metrics for a compute node

Today, it is possible to dynamically optimize the performances of an application through the software prototype named Bull Dynamic Performance Optimizer (BDPO).

Bull Energy Optimizer manages the information related to the power consumed in a supercomputer. It relies on metrics managed by graphite.

The infrastructure allowing the collection and the storage of the metrics is available via the SCS5 Metrics component of Bull SCS 5. Only a complementary configuration is necessary to process new power-related metrics.

The following figure shows the architecture of the metrics framework:

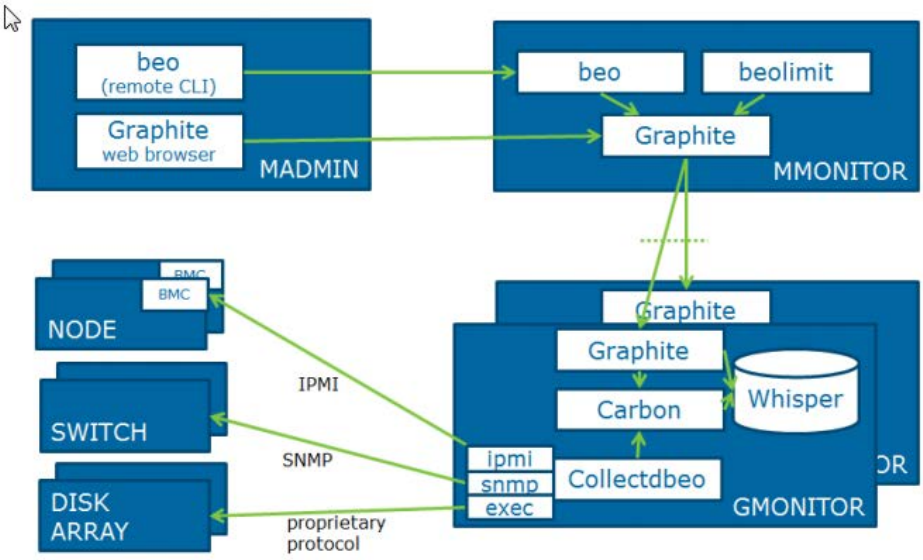

Figure 7: Architecture of the metrics framework

Power-related metrics are collected on Island management nodes using a dedicated collectdbeo service for the equipment in its scope.

Metrics data are stored locally in whisper database files.

The top management nodes do not store any data, but consolidates information on-demand from the different island management nodes.

Bull Energy Optimizer needs the following services to operate:

- On Top management nodes:

    - httpd;
    - postgresql.
- On Island management nodes:
    - httpd;
    - carbon-cache-a;
    - collectdbeo (installed by Bull Energy Optimizer).

Description of the BDPO component

BDPO objective is to complete BEO (Bull Energy Optimizer) software with the capability to dynamically adjust the software and hardware resources' runtime settings for energy efficiency, based on the identification of the phases of the executed application. By doing so, BDPO aims at optimizing the energy consumption associated with the execution of an application, without degrading the performances of the latter.

To summarize the main objectives of Atos-Bull were the following:

- Target real HPC applications, thus not limited to simple benchmarks, and running on large clusters
- No, or limited, requirement for a preliminary knowledge of the application:
    - Limit extensive off-line profiling.
- No code annotation;
- No code modification;
    - Works with all kind of applications (MPI, OpenMPI, Mixed MPI/OpenMP, …)
- No, or limited, performance degradation:
    - Energy reduction not done with the cost of severe performance degradation;
    - Keep execution time under control.
- Multi-platform (mainly Xeon and Xeon Phi architectures) (GPU not targeted yet)

BDPO has two main features: Profiler and Optimizer.

- The Profiler feature can be used to carefully follow different metrics and study applications behaviour. However, its main goal is to provide enough information for the Optimizer feature to know how and when change resources' configuration according to the different applications' phases.
- The Optimizer is acting on CPU frequencies.
    - BDPO is able to follow three different metrics:
    - CPU: Hardware performance counters (libpfm+perf_event interface):
        - Compute intensity: IPC (instructions/cycle);
        - Memory intensity: out-of-core memory traffic (on Haswell architecture only).
    - File system: Lustre statistics:
        - Number of read and written bytes.
    - All the collected metrics, as well as the events (crossed thresholds, decisions taken by the tool), are dumped into log files. This allows to study BDPO behaviour offline after a run of the application.

Description of the HDDEViz component

HDEEViz stands for "High Definition Energy Efficiency VIsualiZation" and aims to provide a visualization framework for users to access power consumption profiling of their application. Such profiling is allowed by HDEEM (High Definition Energy Efficiency Monitoring) library at a sampling of 1 kHz.

An example of the interface is shown on the next figure:

Figure 8: Example of energy visualization

Priority was given to the non-intrusive property for HDEEViz, as well as portability and ergonomic.

- The current visualization tool chosen is Grafana: this choice is motivated by its portability and ease of use. Grafana is accessible on every web browser and offers ergonomic use to create graphics. This choice is reinforced by the “hands on” session in EoCoE workshop, where users had good comments on Grafana interface.
- InfluxDB is the current data basis used to store the energy consumption data. This is motivated by InfluxDB portability and maintainability.
- HDEEViz automates the trigger of power consumption data by using HDEEM library. It then synchronizes the energy consumption data and fills a time-series database (currently InfluxDB). It then provides a direct access for users to the Grafana dashboard result.

On the user side, three steps are needed for users to access to power consumption graphics:

1. Add to your job script the HDEEViz module load and call to the tool,
2. Connect to Grafana server through a Web Browser,
3. Visualize

HDEEViz requires SLURM resources job management software. The srun launcher is used and is currently the only one available.

However, HDEEViz does not requires any SLURM specific configuration. As HDEEViz uses HDEEM library, one must ensure that HDEEM is installed and running on all compute nodes. (Three functions used: startHdeem, clearHdeem, printHdeem). HDEEViz also needs python to be installed on the compute nodes with a list of modules that must be available.

Installation of HDEEViz then only consists in an archive of files that must be on a directory on a shared file system where you want to install HDEEViz. In the context of PCP, HDEEViz is installed on a management node that is dedicated to energy software only.

The supercomputer must have a Web browser available for users able to connect to node where Grafana is running. It is advisable to have client-server software allowing fast rendering like X2Go. This implies that the supercomputer must be equipped of a visualization node either.

Specific development on the SLURM resource manager

A first version of power adaptive scheduling was already pushed into SLURM open-source in the latest 15.08 version. This scheduling is based on centralized mechanism to dynamically reduce the number of resources available for users and low down the power of nodes when launching new jobs. The two main mechanisms used to get this are the dynamic CPU frequency scaling, to decrease the power of nodes, and shutdown of nodes, to reduce the number of resources. This technique allows defining in advance time windows with specific power limits.

In addition, the energy fair sharing technique allows scheduling jobs depending on users' past energy usage. Based on classic fair sharing, this scheduling technique adds the past jobs energy consumption to CPU-time on the calculation of priority.

In this project, the focus was put on a new algorithm based on adaptive power management. Instead of fixing the power limit for the execution of jobs, the power adaptive mechanism adapts the power consumption limit during the execution.

The adaptation algorithm for power/RAPL plugin is at the level of sockets, so the adaption is at the fine-grained level. RAPL plugin power adaption is based on the step by step decrease or increase algorithm, so the real power consumption should not be updated immediately, but in the short span of time it should closely follow the application power consumption. The rate of increase, decrease and which criteria (threshold value) to increase or decrease are based on the configuration information, and it depends on the cluster and user behaviour. The behaviour of the powercaping on one socket depending on the real usage of power usage is graphically visualized in the following figure.

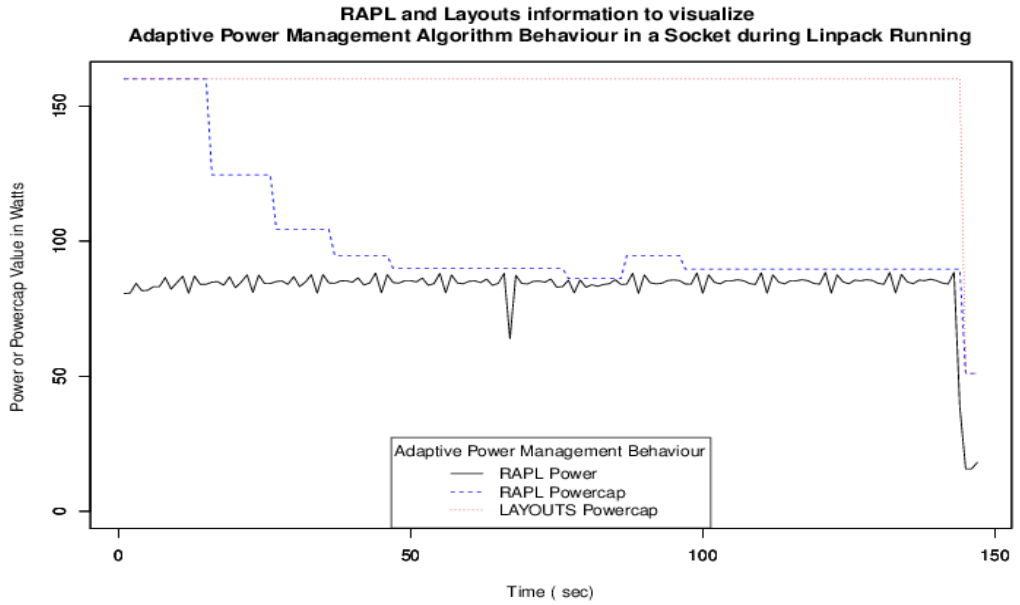

Figure 9: Adaptive Power Management behaviour during Linpack

Since power is allocated close to the application behaviour then the unutilized power of one job should be given to another job to improve system utilization and should reduce waiting time of jobs. In the previous figure, the area between red and blue dotted lines is the unutilized power of the currently running job that should be given to other jobs, which are running in the system or waiting in the queue to reduce job waiting time and improve system utilization rate.

4.2 Suitability for general purpose HPC

4.2.1 *Ease of code development and porting*

Both the porting and optimization exercise helped to evaluate and to foresee potential software and programmability issues introduced by the new processing architecture considered. The programmability and in particular the ease of programming is one of the key of the adoption of the technology by end users, keeping the "Ninja Gap" under control.

For some authors, in many ways, the hardware industry's shift toward parallelism has occurred much faster than the abilities of the software and systems designers to react. Technology provider such as Atos-Bull knows how to build complex multi-core processors or SoC, and we must build them to keep Moore's Law rolling along. But there is still issues on how to program them efficiently — both in terms of software development time and in terms of getting the best power-performance outcomes from them. Furthermore, the shift toward on-chip accelerators offers even greater programmability challenges. Finally, there are a host of programmability concerns that emanate from the basic goal of elevating power to a first-class design constraint alongside performance. For example, from a power perspective, information on the relative criticality of different communication or computation operations may be very useful, but current programming models offer few abstractions or constructs to help programmers manage this. This is clearly a strong challenges that technology provider are facing on today.

Atos-Bull made the choice of the Intel Xeon manycore architecture because of the promising FLOP/s per Watt ratio, but also because its architecture is derived for the x86 architecture. This is a great advantage of such architecture, in terms of knowledge, support and ecosystem.

All the codes have been ported without specific effort. Of course, some issues have been experienced, but not directly linked with the ISA itself. All the porting recipes have been detailed in the Atos main deliverables for Phase III of this project, and all source, makefiles, running scripts have been provided to the assessment committee.

To extract the maximum performance of the Intel KNL, one can start with two fundamental considerations: *scaling* in one hand and *vectorization & memory usage* in the other hand; therefore, the developers have to consider for optimizing their applications:

- **Hybrid programming**, mixing MPI and OpenMP. MPI has to be of course privileged for inter-socket communication whereas OpenMP (or threads) has been used intra-socket. The balance between the number of MPI tasks and the number of threads per tasks is highly depending on the OpenMP scalability of each application. For its part, the number of MPI threads per socket is limited by the footprint of the replicated data set and for the footprint of the MPI process itself. Question of the use of multithreading is also a key component of the performance, helping to mask the latency of memory accesses.
- **Vectorization**, through AVX-512 instruction set. Each KNL cores support two FMA 512-bits SIMD vector engine, leading to a core performance of 32 operations per cycle in double precision (64-bits). To exploit this computational power, code must be vectorised. Auto-vectorization done by the compiler itself is generally not sufficient; developers must re-write some data structures to enable vectorization, add also some compilation hints inserting compiler's pragmas or use intrinsic instructions. The codes could also benefit from vectorization using highly optimised library like mathematical core or kernel libraries.
- **Optimize memory usage and especially the MCDRAM** (fast memory). In order to obtain decent performance, it is supposed that most of the usable data fits this fast

memory. It requires working on the data memory prefetching and data blocking to improve the spatial and temporal data locality.

4.2.2 *Energy monitoring/prediction*

As already detailed in section 4.1.3 several innovative software and hardware component have been designed to allow both system administrators and end-users to monitor and profile power consumption on the Pilot System:

- **Bull Energy Optimizer (BEO)**, which provides information on the whole system at 100 Hz sampling rate and with 3-5% precision.
- **HDEEViz** which provides a visualization framework to end-users to access power consumption profiling of their application. Such profiling is allowed by HDEEM (High Definition Energy Efficiency Monitoring) library at a sampling of 1 kHz.
- **Energy oriented SLURM plugin** based on adaptive scheduling, offering first power consumption monitoring at the job level and the basic framework for the Dynamic Resource Reconfiguration.
- **Bull Dynamic Power Optimizer (BDPO)**, with the capability to dynamically adjust the software and hardware resources' runtime settings for energy efficiency, based on the identification of the phases of the executed application.

4.2.3 *System usability*

After a stabilization period longer than expected, it was possible to organize beginning of October 2017 a workshop with the EoCoE CoE during which a selected panel of users had access to the Pilot system and some tools prototypes. The system usability was good enough to have a smooth and useful workshop for both users and Atos-Bull experts.

- Regarding the PCP developments:

Water Cooled PSU are installed since end of October, since that date the system stays stable so it doesn't impact the system usability.

- Software developments:

Some of the tools like Slurm energy plugins or some features of BEO, HDEEViz are ready to be used by all PCP users. However, BDPO, and some BEO feature are too much oriented for system administrators in their prototype versions. This is clearly something that will be improved in their next releases.

4.3 Impact on energy efficiency

The following table summarizes the initial performance projections proposed at the end of phase II by Atos-Bull for the 0,5 PFLOP/s system deployed at CINES premise.

Applications	TTS and ETS projections 0.5 PFLOP/s Pilot System			
	number of nodes required per copy	number of copies	Time-to-solution [s]	Energy-to-solution [kW.h]
LINPACK	168	1	**1 908**	**38.7**
NEMO	76	2	**1 942**	**28.0**
BQCD	168	1	**2 311**	**37.5**

QE	67	2	**3 200**	**36.4**
SpecFEM3D	168	1	**15 500**	**297.5**

Table 1: TTS and ETS projections for 0.5 PFLOP/s Bull Pilot System

These numbers have been obtained considering the downsizing of the systems to the different vendors.

In terms of single performance, the original assumptions (inputs) of the model presented in the previous phases of the project were not modified.

- The "time-to-solution" (TTS) projections have been performed from scalability runs performed on Haswell and Broadwell systems. First projections have considered Intel Manycores technology as well as NVIDIA GPU Tesla Volta (V100) one. Projections have been done aggressively on Intel Knights Landing technology considered as a minimum and lowest performance achievable by NVIDIA V100.
- The "energy-to-solution" (ETS) metrics has been projected and estimated in the early phase of the project with a first, simplified model and cross-checked between phase II and phase III of the project with 'power estimator tool' provided by the Atos R&D team. Recent measurements made on some site show that the accuracy of this tool is within a 5 to 6.5% error margin.

4.3.1 *Final results for HPL*

HPL performance is relatively predictable, especially at small and medium scale (~PFLOP/S), both in TTS and ETS. For HPL, key parameters are

- The interconnect network: even if HPL looks not memory bound, performance of HPL is driven by communications especially at the end of the runs and/or for small matrices.
- Size of the matrix to solve (N parameter). Larger is N, better is the performance. It is generally recommend committing between 80-90% of the memory available per node. Considering 96 GB of RAM per node, 90% is highly recommended.

According to end of phase report provided by Atos-Bull, for the HPL test, 40% of the nodes memory was used. The matrix size is 930720 and the block size is 336, which is the optimal block size for KNL 7250. 64 threads were set per MPI process, as it is advised to let 4 processes free of computation in the context of KNL 7250. HPL threads use the KNL 7250 physical cores.

TTS (sec)	ETS (MJ)	ETS (kWh)	Nodes	Tasks per Node	Tasks	Threads
1910	121.5	**33.61**	168	1	168	64

Table 2: HPL large test case results

TFLOP/s	Nodes (kWh)	GFlops/Watts (Nodes)
281,37	32.2	8,74

Table 3: HPL GFlops per Watts (nodes)

4.3.2 *Final results for NEMO*

The NEMO small test case uses 960 steps while the large test case uses 21120 steps.

Nemo is not using OpenMP, therefore this simplifies the execution environment. In this version, IOs are deactivated.

It is generally advised to distribute the MPI tasks in a round robin fashion to execute NEMO faster. However, this behaviour is not observed on PCP supercomputer and increases the execution time if processes are not distributed in a block fashion.

The variables "`I_MPI_DAPL_UD_SEND_BUFFER_NUM`" and "`I_MPI_DAPL_UD_RECV_BUFFER_NUM`" are usually set to 8192 for large run jobs while Intel advises the value "ntasks*4 +16". In the context of NEMO, setting the Intel advised value provided better execution time.

The variable "`I_MPI_ADJUST_ALLREDUCE=4`" also contributes to reduce NEMO execution time.

TTS (hh:mm:ss)	TTS (sec)	ETS (MJ)	ETS(kWh)	Nodes	Tasks per Node	Tasks	OMP
00:04:52	292	11,5	3,194	150	64	9600	#

Table 4: NEMO small test case BULL results

TTS (hh:mm:ss)	TTS (sec)	ETS (MJ)	ETS(kWh)	Nodes	Tasks per Node	Tasks	OMP
1:31:55	5515	221,7	61,58	150	64	9600	#

Table 5: NEMO large test case BULL results

The following figure presents the node 1048 power consumption on the first 255 seconds:

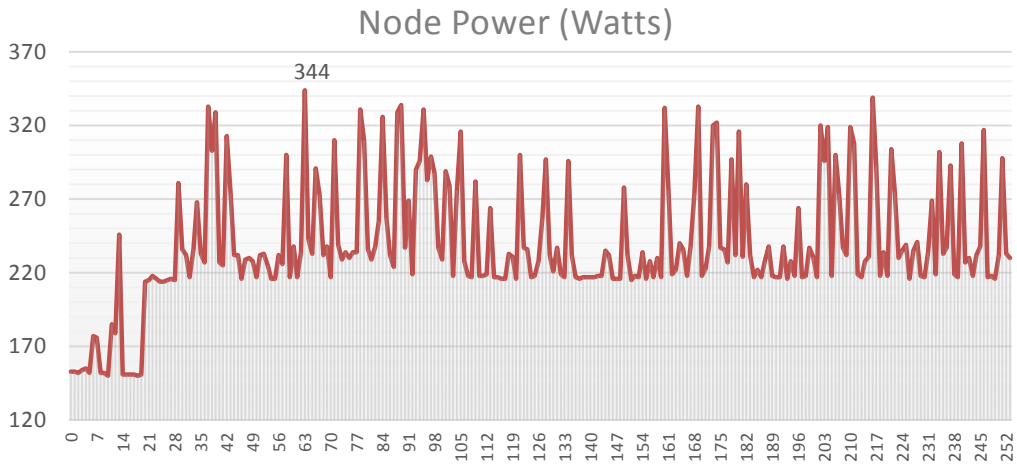

Figure 10: Node 1048 power evolution

On the figure above, the initialization phase (reading input data) is noticeable. However, the computation phase is dense as node power consumption goes until 320-340 Watts. The energy consumption profile for computation phase is the same for all the NEMO execution.

NEMO, as announced in BULL proposal was the most challenging to port and optimize on KNL.

Current work leads to a TTS equal to 5515 seconds for one copy running on 150 nodes, with an ETS of 51.58 KWatt.h (~40.2 KW). Initial projection leads to 1942 seconds for 2 copies of

76 nodes each. It represents a degradation factor of x5.6 which is not acceptable, even if performance projection was very aggressive on this benchmark.

Preliminary conclusion is that Intel Knight Landing is not the best target for NEMO. The code is poorly vectorized and mainly memory bound. Moreover the code is not hybrid (OpenMP) and optimization was not possible in the time frame of this project. It will require a huge effort in terms of code modification and optimization for a small benefit expected on the KNL. Standard CPU looks clearly more suitable for such kind of code. Current Intel Xeon Skylake will certainly improve the performance of this application. Following the model, 200 nodes with 2 sockets of Intel Xeon Gold 6130 allow to reach the target of 1942 seconds.

4.3.3 *Final results for SPECFEM3D*

Due to the downsizing of the pilot system, it has been agreed to review the SpecFEM3D large test case configuration for some problem with the memory footprint. After some analysis and internal discussions, it was agreed to run the SpecFEM3D with parameter NEX_XI set to 864 and with 2 copies of 81 nodes each on the Pilot System. This configuration ensures to maximize the load of the system (162 nodes total out of the 168 available) as well as the memory footprint (74-80% of the 96 GB available on each compute) node.

Small test case

TTS (hh:mm:ss)	TTS (sec)	ETS (MJ)	ETS(kWh)	Nodes	Tasks per Node	Tasks	OMP
00:00:45	45	0,0081	0,002	1	16	16	2

Table 6: SPECFEM3D mesher small test case results

TTS (hh:mm:ss)	TTS (sec)	ETS (MJ)	ETS(kWh)	Nodes	Tasks per Node	Tasks	OMP
00 :23 :10	1390	0,3329	0,092	1	16	16	2

Table 7: SPECFEM3D solver small test case results

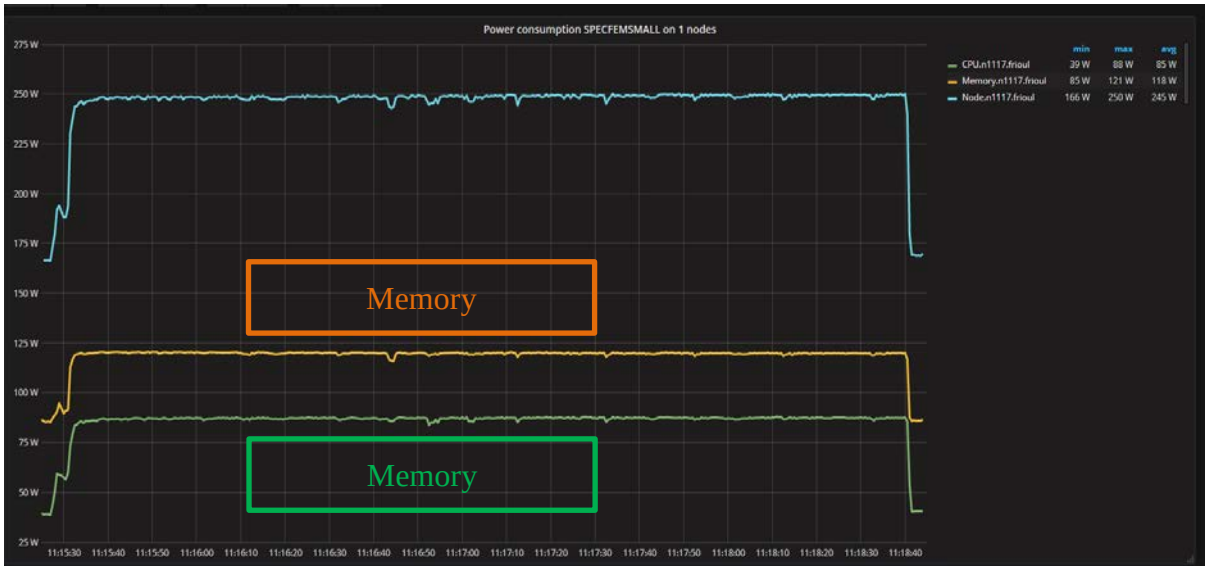

Figure 11: SPECFEM3D small test case on 1 node

Figure above shows the energy consumption (in Watts) for node n1108. Atos-Bull executed the small test case provided by PRACE for specfem3D. This test shows that energy

consumption is 34.8% due to the CPU, while 48.26% is due to the Memory energy consumption (note the typo in the picture: the green curve refers to the power consumed by the CPU).

16.98% of the energy consumption is due to other components on the node.

Figure 12: SPECFEM3D small test case on 16 nodes

On this figure one can see all the nodes used to execute the SPECFEM3D small test case. In this case, by taking into account all the nodes, 54.37% of the energy consumption is due to the memory, while 19.71% is due to the CPU.

Clearly, node 1133 has higher energy consumption than the other nodes; this is clearly visible on the CPU consumption. The gap observed on nodes consumption for n1133 comes from higher energy consumption at the CPU level. The node 1133 is the node where the sbatch is executed, but the gap might be explained by additional operations executed by process 0 for SPECFEM3D.

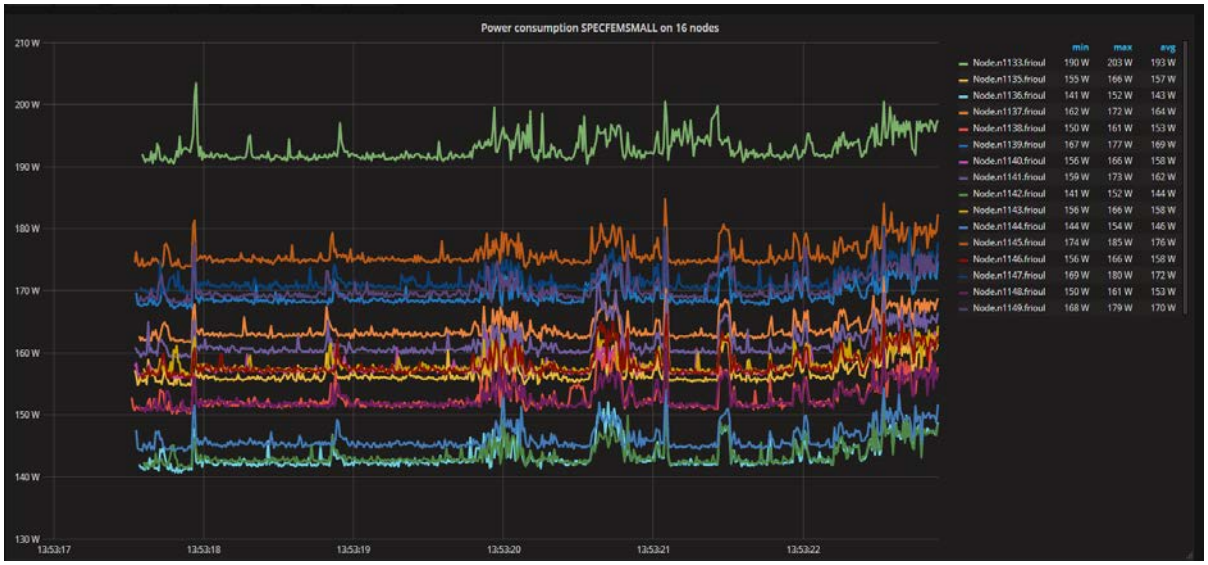

Figure 13: zoom of specfem3D small test case on 16 nodes

By zooming the previous figure as the computational part was relatively flat at this scale. On this figure, only the nodes energy consumption is shown.

All nodes have the same energy consumption line, except the node n1133, with some energy consumption peak between 13:53:20 and 13:53:22.

Large test case

TTS (hh:mm:ss)	TTS (sec)	ETS (MJ)	ETS(kWh)	Nodes	Tasks per Node	Tasks	OMP
00:04:10	250	4 ,4	1 ,22	81	48	7776	2

Table 8: SPECFEM3D mesher large test case results for 1 copy

TTS (hh:mm:ss)	TTS (sec)	ETS (MJ)	ETS(kWh)	Nodes	Tasks per Node	Tasks	OMP
00:30:03	1996	48,9	13,58	81	48	7776	2

Table 9: SPECFEM3D solver large test case results for 1 copy

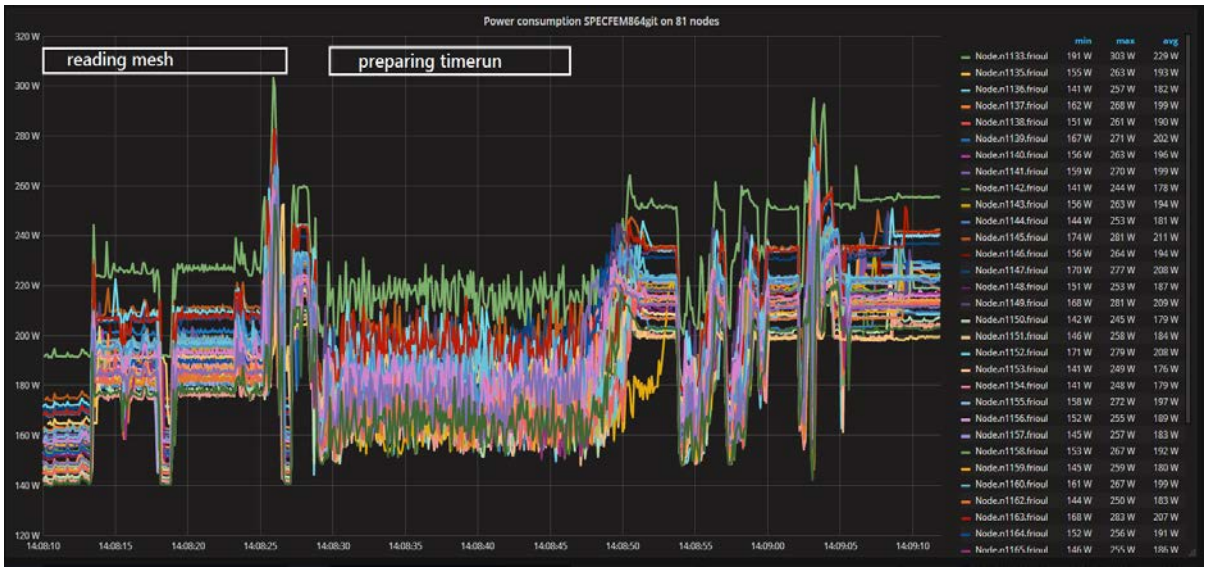

Figure 14: Specfem3D large test case, zoom on 1 copy

On the figure above, one can observe at the beginning of SPECFEM3D two initialization phases: read of mesh input, which is not regular, followed by the run preparation.

The computation part is very smooth and regular, it consumes between 200-230 Watts for each nodes.

The energy consumption differences for each node do not come from memory and CPU energy consumption but from other components.

4.3.4 *Final results for BQCD*

TTS (hh:mm:ss)	TTS (sec)	ETS (MJ)	ETS(kWh)	Nodes	Tasks per Node	Tasks	OMP
00:01:19	79	0,1519	0,042	8	64	512	#

Table 10: BQCD small test case BULL results

TTS (hh:mm:ss)	ETS (MJ)	ETS(kWh)	Nodes	Tasks per Node	Tasks	OMP
04:18:52	505.6	140.28	128	64	8192	#

Table 11: BQCD large test case BULL results

The BQCD large test case suffered from several problems:

- The number of nodes is not sufficient to run such test case. Ideally, using 256 nodes would perform much better.
- Using OpenMP threads freezes the BQCD application. For this reason, OpenMP were not used, even though it should drastically improve BQCD performances.
- BQCD suffers from an output statement overflow when writing the output in bqcd.571.u files.
- Finally, establishing the correlation between BQCD input parameters of older version with latest (5.1.0) is very difficult: some parameters are not enough documented.

BQCD is suffering from the same symptom from NEMO. Revised projections estimated TTS to 2311 seconds on 168 nodes. Current measurements show a TTS around 15532 seconds which is not acceptable at all. The standard model used for estimate the performance looks not accurate to project performance from standard Xeon processors to manycore ones. It is due to the poor performance of single KNL core (low frequency) and difficulty to take into account MCDRAM constrain and specifics in the model. BQCD will certainly perform much better on standard Xeon core.

4.3.5 *Final results for Quantum Espresso*

Quantum Espresso was not executed by using the last Intel compilers version 2018 update 0, instead the version 2017 update 2 was preferred, as QE runs faster with this one than with the version 2018. Note that Intel compilers version 2017 update 0 is not advised as bug has been reported by using this version. Still, Intel provides a workaround if you want to use Intel compiler update 0.

Quantum Espresso execution is strongly dependent of the application parameters itself, as well as MPI execution parameters.

The small test case does not require any specific settings as it runs in only 24 seconds.

TTS (hh:mm:ss)	TTS (sec)	ETS (MJ)	ETS(kWh)	Nodes	Tasks per Node	Tasks	OMP
00 :00 :24	24	0,0091	2,530	2	32	64	#

Table 12: Quantum Espresso small test case BULL results

In order to decrease Quantum Espresso execution time, one must retrieve at least the following parameters:

Name	Computed/Input	Value (large test case)
N(1)	Computed	216
k-points	Computed	26
Kohn-Sham states	Computed	326

Table 13: Quantum Espresso parameters value for large test case

The parameters above were used to fix the number of MPI tasks and the input parameters `ntg`, `ndiag`, `npool` and `nk` that help to decrease execution time.

One copy of Quantum ESPRESSO on 72 nodes was executed, as 72 is a divider of 216. The `nk` parameter is fixed to 26, which is the value of k-points. Finally, the number of tasks is the multiplication of 72 and 26, which is 1872.

`ntg` was set to 12, as 12 is the biggest common divider of 216 and 26.

Finally, setting 2 for `npool` and `ndiag` provided the best results.

`Ndiag` is an input parameter that controls the tasks involved in diagonalization part. While it is advised to use `ndiag`=1 on multicores CPU, it performed better when set at 2. `ndiag` controls SCALAPAK linear algebra parallelization.

The `npool` parameter activates k-point parallelization if set with –`ntg` and `ndiag`.

Round robin fashion tasks pining on processors is as efficient as a block distribution. This is done by using `-distribution=block:fcyclic` or `-distribution=block:block` SLURM parameter.

Using the OpenMP considerably slows down the execution. Therefore, in what follows, OpenMP was not used. This behaviour is probably due to the MPI execution environment as Quantum ESPRESSO is very sensitive to its environment. The variable(s) or compilation option origin of OpenMP hang, was not determined.

Using carefully the `ntg`, nk, `ndiag` and `npool` parameters can improve drastically the execution time, reducing the initial run of two hours and 30 minutes to just 35 minutes.

TTS (hh:mm:ss)	TTS (sec)	ETS (MJ)	ETS(kWh)	Nodes	Tasks per Node	Tasks	OMP	Copy
2:23:20	8600	52,2	14,50	32	8	256	#	1
1:52:42	6762	94,5	26,25	72	26	1872	#	1
00:35:40	2140	33,7	9,36	72	26	1872	#	1

Table 14: Quantum Espresso large test case Atos-Bull results

	Jobid 8017 (2:23:20)			Jobid 8201 (1:52:42)		
Info	nodes	Eth. Switches	IB. Switches	nodes	Eth. Switches	IB. Switches
	n[1060-1091]	esw[102,104,118,120,122], ewmlm0, ewmtm[0-1]	isw[102,104,118,120,122,200,204,208,214,218,222]	n[1060-1083,1132-1155,1168-1191]	esw[102,110,114,118,120,122] ewmlm0, ewmtm[0-1]	isw[102,110,114,118,120,122,200,204,208,214,218,222]
Energy	Nodes		Switches	Nodes		Switches
	47.9 MJ		4.3 MJ	88.1 MJ		6.4 MJ

Table 15: Quantum Espresso energy consumed for job 8017 and 8201

The run 8201 is executed in 1h:52m:42 sec versus almost 2h30 for job 8017. Even though the job 8201 is considerably faster, adding CPUs and switches to improve the performance is not necessarily interesting in this case, as decrease the execution time from approximately 30 min does not overcome the energy consumption of nodes and switches involved.

	Jobid 8256 (00:35:40)			Jobid 8201 (1:52:42)		
Info	nodes	Eth. Switches	IB. Switches	nodes	Eth. Switches	IB. Switches
	n[1060-1083,1132-1155,1168-1191]	esw[102,110,114,118,120,122], ewmlm0, ewmtm[0-1]	isw[102,110,114,118,120,122,200,204,208,214,218,222]	n[1060-1083,1132-1155,1168-1191]	esw[102,110,114,118,120,122] ewmlm0, ewmtm[0-1]	isw[102,110,114,118,120,122,200,204,208,214,218,222]
Energy	Nodes		Switches	Nodes		Switches
	31.6 MJ		2.0 MJ	88.1 MJ		6.4 MJ

Table 16: Quantum Espresso energy consumed for job 8256 and 8201

Regarding the jobs 8201 and 8256, the same number of nodes and switches were used§. However, correctly setting the QE input parameters (`ndiag`, `npool`, `ntg` and nk) improves drastically QE performances. In this context, compared to the previous job 8017, the execution time is sufficiently decreased to overcome the nodes and switches consumption.

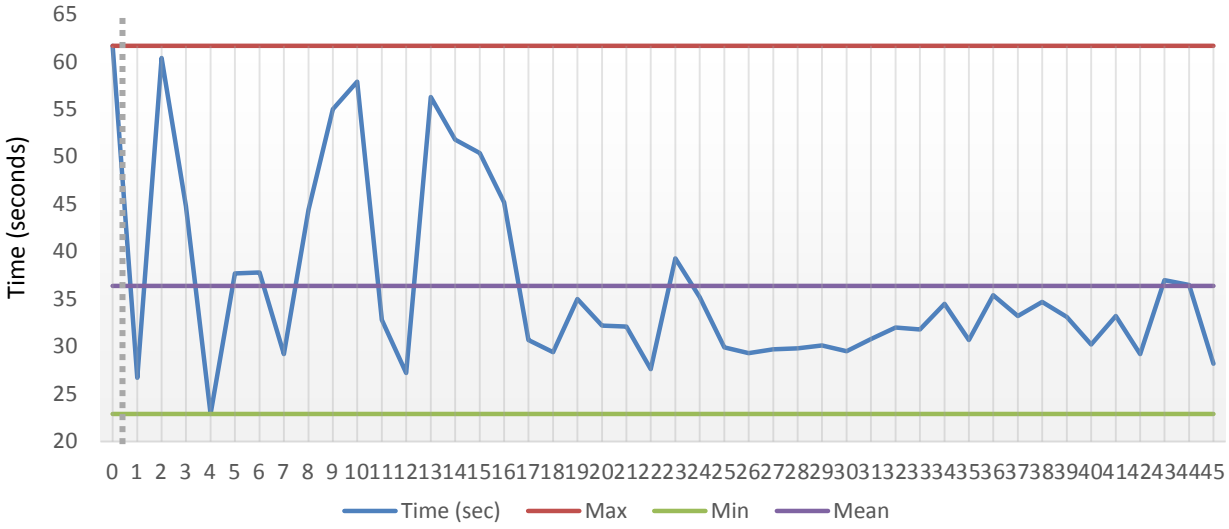

Figure 15 Time in second per interation in Quantum ESPRESSO

On QE large test case, the first iterations time execution is not regular while it tends to stabilize after the 17th iteration (on 45 iterations plus one initialization phase).

In the following figure, QE error convergence is presented with respect to the time execution per iteration. Though iterations execution time and energy consumption become more regular after the 17th iteration, it was not possible to conclude on a particular impact of iteration execution time and energy consumption with the convergence speed of QE error.

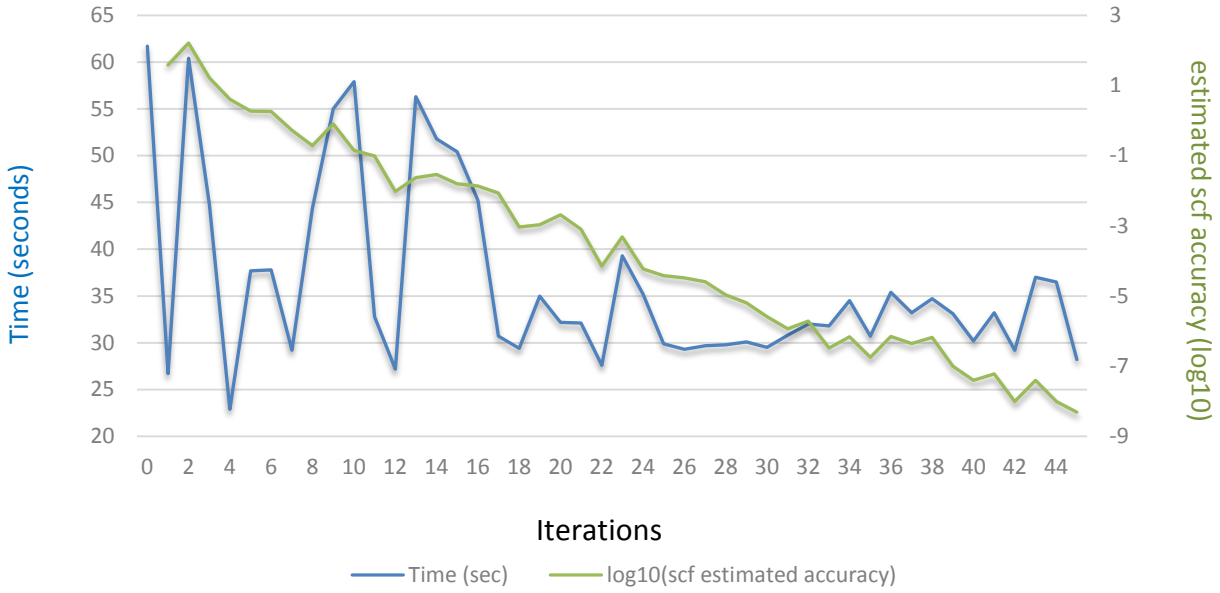

Figure 16: Time in second per iteration with respect to scf estimated accuracy (log10) for Quantum Espresso

Below, is presented the energy consumption of node 1105 in watt per second. Each line represents respectively the initialization phase followed by iteration 1, 2, until the 11th iteration.

The energy presented below is retrieved thanks to SLURM energy profiling. The energy is retrieve at 1Hz.

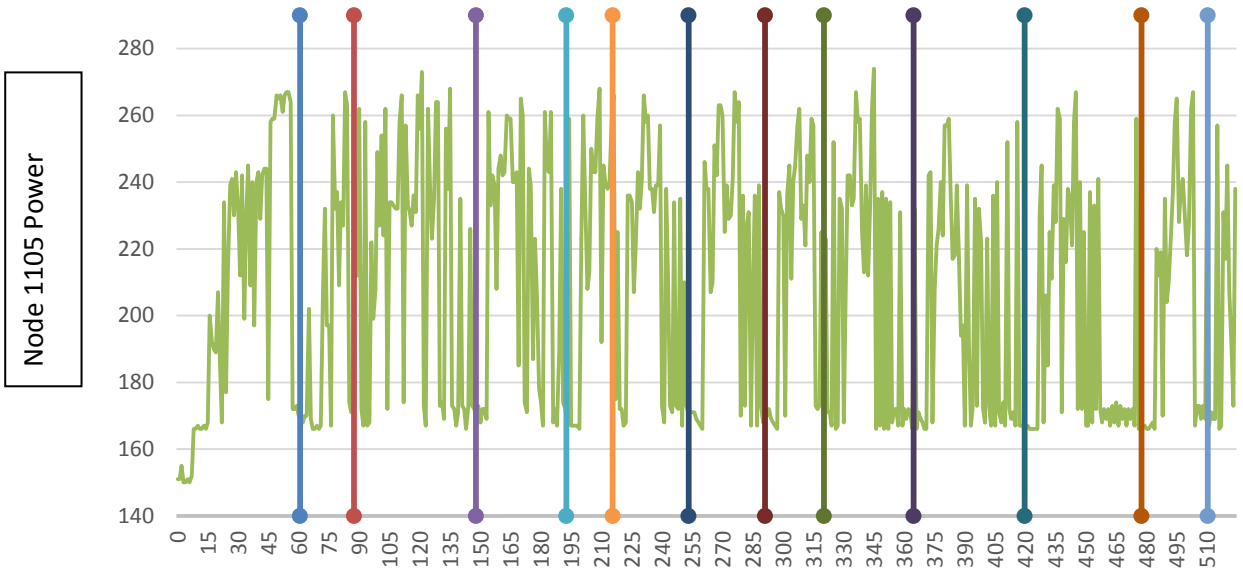

Figure 17: Node n1105 Power per second, from initialisation to iteration 11

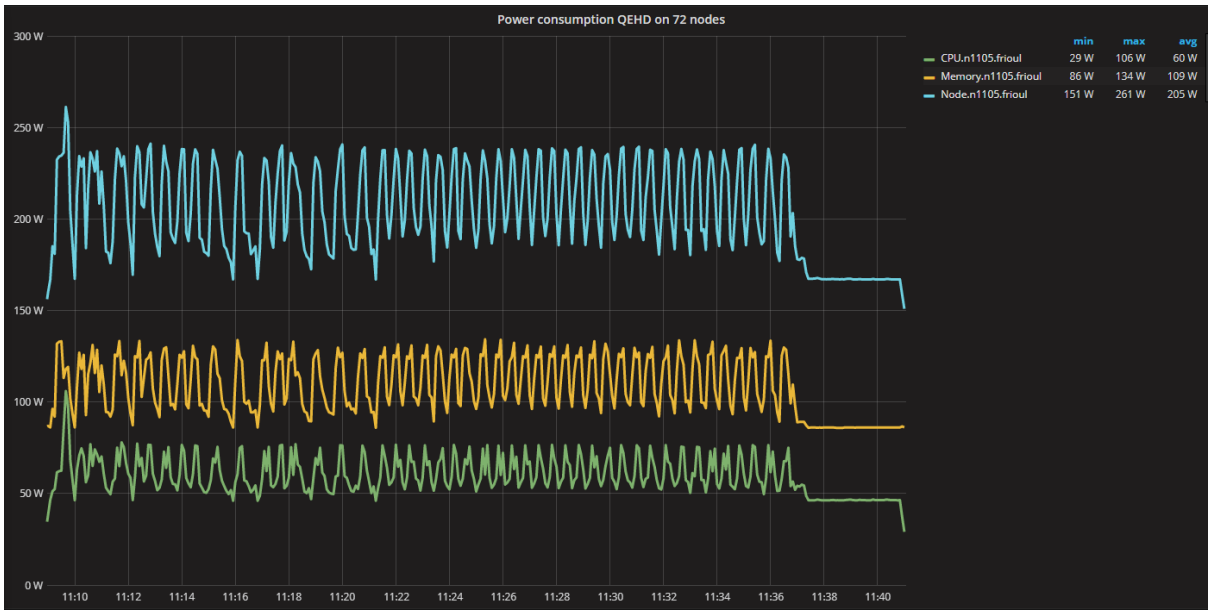

Figure 18: Node 1105 energy consumption for Quantum ESPRESSO large test case

Figure 18 shows the energy consumption for a particular node as retrieved by the HDEEM library (1 kHz). The node energy consumption values are very similar to the values retrieved by SLURM energy profiling (1 Hz).

However, this graphic (provided by HDEEVIZ) shows that memory is the dominant energy consumption point.

It is shown on the figure below the node 1116 and 1215. The node 1116 is the one that consumed the most energy during the QE large test case execution while n1215 is the one that consumed the least energy for the same run.

One can observe that the memory consumption and CPU consumption have approximately the same values for nodes 1116 (i.e. the greediest node for energy consumption) and n1215 (has the minimum energy consumption for QE execution).

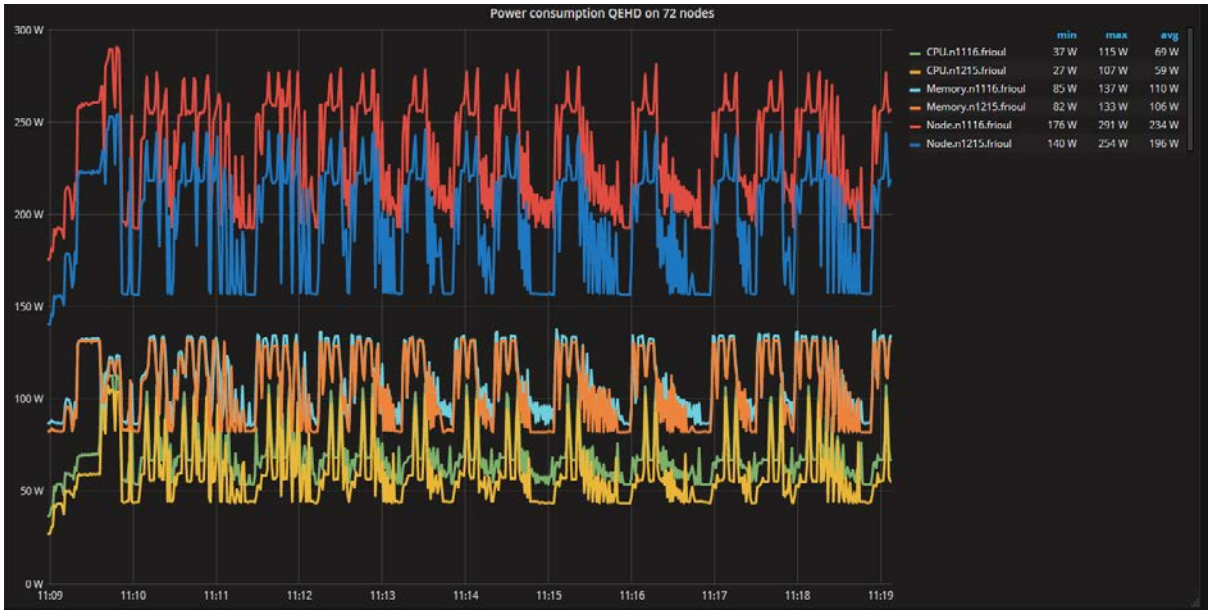

Figure 19: Nodes energy consumption (max and min) for Quantum ESPRESSO large test case

In the context of Quantum ESPRESSO, comparison was done on the execution time for a job without energy consumption measure with a job using HDEEViz and a job using SLURM energy profiling.

	No energy consumption		HDEEViz		Srun	
Energy (MJ)	**nodes**	**Switches**	**Nodes**	**Switches**	**Nodes**	**Switches**
	27.9	1.8	31.1	2.1	27.9	1.8
Energy (kWh)	**nodes**	**Switches**	**Nodes**	**Switches**	**Nodes**	**Switches**
	7.75	0.5	8.64	0.58	7.75	0.5
Time	00:31:35		00:36:25		00:31:39	
	1895		2185		1899	

Table 17: Comparisons of QE runs with and without energy consumption

The SLURM profile has 0 impact on the job execution time and energy consumption. However, it was observed that generating graphics with HDEEVIZ has a 5 minutes impact on the job execution. In these tests, it was considered that HDEEVIZ has an impact between 10 and 15% on the execution time for this test case.

4.4 Schedule and timing

During phases I and II of the PCP, the targeted architecture for the Atos Pilot System was a 1 PFLOP/s system based on GPU accelerators. For phase III it has finally been decided to go for smaller configuration based on KNL processors.

The Pilot System has been delivered on Friday 14 April 2017 at CINES in Montpellier (France). Hardware installations including cabling, powering and pipe connections were finalized on 20 April 2017.

The software installation was significantly longer than expected. Indeed, the PCP Pilot is sharing its Sequana infrastructure with another cluster named Frioul, which led to complexes installation issues.

Moreover, as for any system hosted in CINES, the PCP Pilot configuration had to follow strict security rules which was sometimes complex to setup with prototypes tools.

Even if several jobs have been executed earlier, the PCP pilot including first versions of BEO, HDEEViz and SLURM energy plugin became available for experimentations around mid-July.

4.5 Impact on Atos-Bull roadmap

Two main impacts have to be mentioned:

- **Water-cooled power supplies**: as a result of this PCP, Atos-Bull decided to introduce water cooled power supplies in the Atos Hardware catalogue. First customer shipments are planned in H1 2018.
- **Software developments:** the software tools developed during the PCP phase III (BEO, BDPO, HDEEVIZ, SLURM Energy saving plugins) will be part of Atos-Bull Supercomputer Suite 5 Release 2 (SCS5 R2) that will be release in Q1 2018.

4.6 Summary of lessons learned from the Atos-Bull KNL pilot system

The main learnt lessons during this project are the following:

- Applications execution/optimization on KNL architectures: thanks to the project, a lot of time was spent on the proposed applications (Nemo, BQCD, Quantum Espresso and SPECFEM3D) allowing Atos-Bull and the community to learn a lot on the way to run them on KNL based architecture
- During the project, design, production and installing of prototypes of water-cooled power supply were performed. The pilot system is the first system with a 100% water-cooled rack in a production like environment. This installation gave Atos-Bull the confirmation of both the stability and efficiency of their design
- SLURM adapted capabilities: the experiments made in Atos-Bull labs are promising but unfortunately they didn’t have time to test it on the pilot system
- BDPO: tested on system but KNL seems not to be the most appropriate architecture to take the best benefit of this technology
- BEO: has been installed and tested during the EoCoE workshop. First feedback were that this release of that tool was too much oriented for system administrators. Atos-Bull took that into account and plan that next releases will then include more features for users.
- HDEEVIZ: has been installed and tested during the EoCoE workshop. This first prototype didn’t take into account some security requirement of HPC Datacentres. Then some configuration/adaptation were performed manually. The next release should improve these aspects.

5 Lessons learned from the Maxeler data-flow pilot system

5.1 Description of the Maxeler data-flow pilot system

The pilot system by Maxeler is based on Maxeler's most recent generation of Data Flow Engines (DFE) cards, called MAX5. A DFE card comprises an FPGA, which is used to implement part of the application following a data-flow paradigm. An FPGA (Field-Programmable Gate Array) is an integrated circuit like a CPU, but unlike a CPU it is reconfigurable such that its behaviour can be customised. For each new application, which is executed on the pilot system, the DFE cards are reconfigured. Multiple DFE cards can be integrated into a single server, which also comprises standard CPUs. Given the small budget allocated to Maxeler during Phase III, only a relatively small pilot system could be delivered. It nevertheless serves the purpose of demonstrating that real-life HPC applications, as they are used on PRACE systems, can be implemented on this architecture. Furthermore, it provides users within PRACE with the opportunity to further explore this technology.

5.1.1 *System description*

The Maxeler pilot system is still in the process of being integrated into a larger testbed for reconfigurable and data-intensive computing as shown in Figure 20.

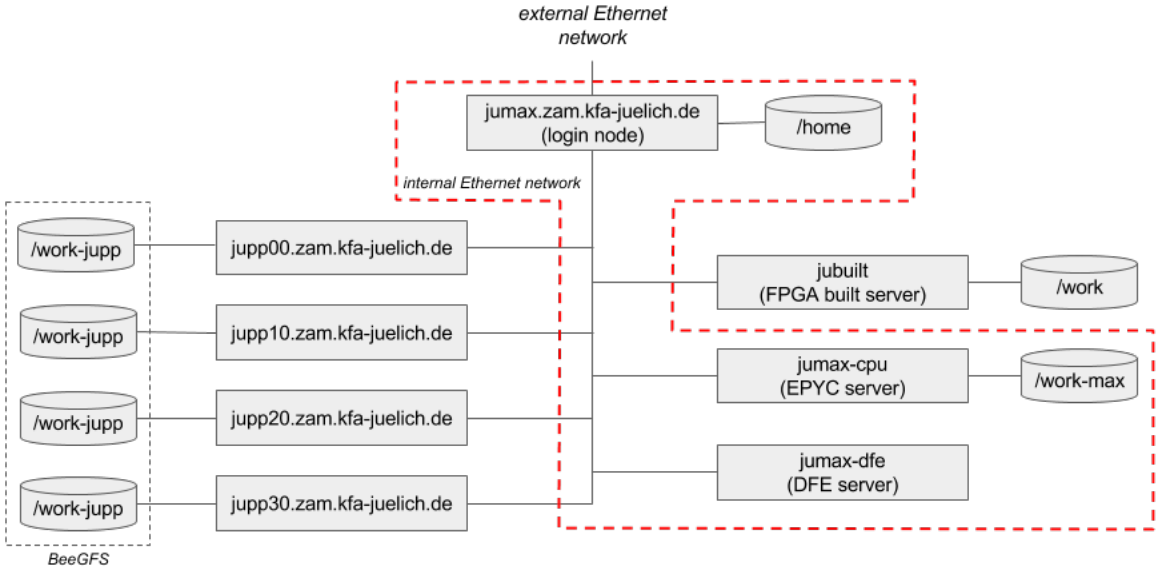

Figure 20: Overview on the planned testbed for reconfigurable and data-intensive computing with the integrated Maxeler Pilot System. The components of the latter are shown by the dashed line.

The pilot system delivered by Maxeler comprises the following components:

- MPC Node (jumax-dfe): A Maximum Performance Computing (MPC) server with MAX5 Data Flow Engine (DFE) cards
- CPU Node (jumax-cpu): A node with high-end CPUs and a very large memory capacity for executing the code part that are not ported to the DFEs; and
- Head Node (jumax): A node serving as a head node that acts as a gateway to the Pilot System and provides a few management services.

The MPC Node is a server with 8 MAX5 cards. MAX5 is a standard PCIe form factor card with an FPGA, which can communicate with a commodity CPU via PCIe as shown in Figure 21. A MAX5C card as shown in Figure 22 has the following hardware features:

- Xilinx VU9P FPGA
- 42 MByte of „fast memory“ (FMEM) integrated into the FPGA
- 48 GByte of „large memory“ (LMEM) attached to the FPGA
- Custom ARM-based control board

The DFE cards are interconnected by a proprietary network called MaxRing. The node features a dual-port Infiniband FDR card to connect it to the CPU Node.

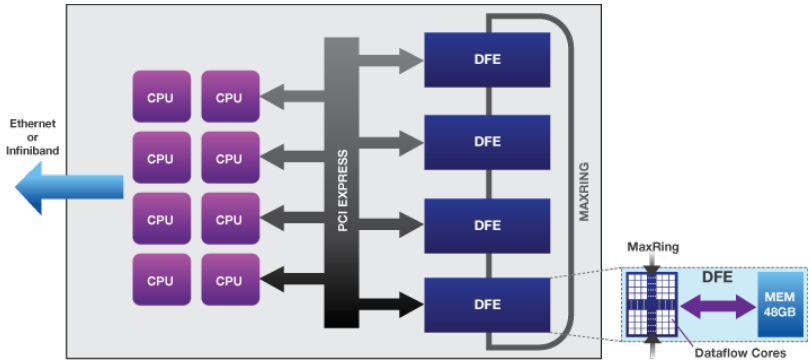

Figure 21: Maxeler MPC series node architecture (©Maxeler).

Figure 22: A Maxeler MAX5C card (©Maxeler).

The CPU Node has the following hardware features:

- 2 AMD 7601 EPYC CPUs (2*32 cores), 2.7 GHz
- 1 TByte main memory
- 9.6 TB SSD (data storage), 2*240 GB SSD (system)
- 2x IB FDR ports for point-to-point connection to MPC Node
- 1x 10GE port for point-to-point connection to Head Node

The head node serves mainly as gateway to the pilot system. It has the following hardware features:

- AMD Opteron 6338P (12 cores), 2.3 GHz
- 64 GByte main memory
- 1 TB HDD
- 1x 10GE port for point-to-point connection to the CPU Node
- 1x 10GE port used as uplink to the data centre

5.1.2 *Software environment*

All nodes are operated using the current version of CentOS, i.e. CentOS version 7.4. This Linux distribution is commonly used for many HPC systems and thus allows providing users an environment to which they are used.

More specific for the pilot system are the development tools needed to use the MPC Node with its MAX5 DFE cards. The following tools are installed on Head Node and CPU Node:

- Maxeler's Eclipse-based integrated development environment (IDE) "MaxIDE";
- Maxeler's compiler "MaxCompiler"; and
- Xilinx's FPGA design suite Vivado.

Maxeler agreed to provide regular updates for their tools during the lifetime of the pilot system, i.e. during a longer time period after the PCP ended.

5.1.3 *Energy efficiency aspects of the design*

The strategy of Maxeler in this PCP was to focus on the demonstration of the usability of their technology for real-life HPC applications and thus to exploit the potential for significant improvement of energy efficiency. This requires a significant fraction of the application workloads, which were part of this PCP, to be ported to FPGAs using the Maxeler development tools and following a data-flow paradigm. It should be noted that today's HPC applications typically based on the notion that they run on a processor that executes a series of operations. To program such processors in scientific computing typically imperative programming languages like C, C++ or FORTAN are used. An application implemented in a data-flow programming model can be much easier and more efficiently be mapped onto FPGAs.

Data-flow programming models an application as a directed graph of the data flowing between operations. Let us consider the example of adding two arrays of floating-point numbers with two input arrays `A` and `B` and the output assigned to the array `C`. Expressed in an imperative language like C this may look as follows:

```
float A[N]; float B[N]; float C[N];
for (int i=0; i<N; i += 1)
  B[i] = A[i] + B[i];
```

Expressing the same numerical task in Maxeler's data-flow language MaxJ may look as follows:

```
DFEVar A = io.input(”A”, dfeFloat(8, 24));
DFEVar B = io.input(”B”, dfeFloat(8, 24));
DFEVar C = A + B;
io.output(”B” , B , dfeFloat(8, 24));
```

The following is worthwhile to notice:

- In this example the loop length `N` is not defined as the operation will end when the flow of data is stopped, i.e. the length of the arrays only needs to be known to the routine that is driving this small data-flow engine.
- The floating-point type is parametric with the first and second integer valued parameter defining the size of the exponent and mantissa, respectively. In the example shown above we have 8 exponent bits and 24 mantissa bits and thus a format that is equivalent to the IEEE single-precision floating-point format. However, also different choices would be possible.

The base technology needed to program and use FPGAs using a high-level programming language has been developed by Maxeler outside of this PCP. Nevertheless, significant

development efforts were required to facilitate implementation of HPC applications on their technology. These HPC applications differ from the applications, for which the Maxeler architecture is used so far, in the following aspects:

1. Among the data processing operations, which are performed for each of the applications, the floating point operations clearly dominate;
2. All provided applications are highly complex and comprise a large number of small kernels; and
3. Significant parts of the applications cannot be easily ported to the DFE cards.

The 1st aspect is important as floating point operations are relatively expensive in terms of consumed FPGA resources. While FPGAs can be used for double-precision, i.e. 64-bit, floating point arithmetics, the costs in terms of consumed FPGA resources are high. Using a lower precision or even fixed point arithmetics reduces the required amount of FPGA resources significantly and it thus becomes easier to fit a kernel into a single DFE. From a hardware perspective it is despite the use of high-end FPGAs for the pilot system therefore desirable to lower the precision for floating point operations or to replace floating point arithmetics with fixed-point arithmetics. Whether this is acceptable from an application perspective is a topic for further research.

As a consequence of the 3rd aspect, significant parts of the applications continue to run the standard CPUs integrated into the architecture of the pilot system, i.e. they do not get accelerated by the DFEs. To mitigate the risk of being affected by Amdahl's law, where overall application speed-ups cannot be realised due to the time spent in the non-accelerated part of the application, it is therefore desirable to overlap execution of kernels on the CPUs and the DFEs. This will also improve the overall utilisation of the provisioned hardware and improve energy efficiency as unused hardware typically continues to consume power.

To address the outlined challenges, Maxeler developed within the PCP the following components:

- A Value Profiling Library to support floating-point to fixed-point conversion;
- An Execution Analysis and Visualisation Tool, which enables detailed monitoring of an application with regards to its DFE and CPU parts, supporting the designer in achieving full overlapping of DFE and CPU execution; and
- A kernel merging tool, which performs resource sharing and optimisations on designs with multiple kernels.

5.2 Suitability for general purpose HPC

In the following table the suitability of the solution proposed for "general purpose" HPC by Maxeler is assessed. The term "general purpose" need to be defined. In this section we use it to refer to systems that can – compared to other solutions – be used efficiently for a wider range of applications without significant porting and optimisation efforts.

Ease of code development and porting	Compared to other methods for programming FPGAs the Maxeler technology makes code development for and porting of applications to FPGAs significantly easier. However, since applications used on today's PRACE architectures are implemented for fundamentally different architectures, significant efforts are needed to port these applications following a data-flow paradigm. The ease of code development is partially limited due to lack of available code for data-flow

	architectures as well as the education and knowledge of the application developers. Additionally, for complex kernels the implementation on an FPGA remains challenging due to the limited amount of available resources. Non-trivial optimisation efforts might be required to facilitate successful placement and routing during the last step of the development chain before creating a bitstream for the FPGA.
Energy monitoring/ prediction	The solution does not provide energy monitoring capabilities that reach beyond what is available on typical HPC systems deployed as of today. During the PCP Maxeler could show that their development tools can predict time-to-solution and power consumption and thus energy-to-solution with a relatively high accuracy.
System usability	The development tools including an IDE and the stable run-time environment contribute positively to the usability of the system. Significant efforts would be needed to reach a level of usability similar to current PRACE Tier-0 systems when deployed at scale. This concerns, e.g., integration in resource management and monitoring systems.
User experiences and feedback	As the pilot system became operational by the end of October 2017, the time for collecting sufficient reports on user experiences and other feedback has been too short.

Another aspect affecting usability of the proposed solution is the possible need to reduce the precision, at which floating point operations are being performed, in order to reduce the resources needed on the FPGA in order to facilitate the place of a kernel on the available FPGA. While for the applications considered within the PCP some comparison of results obtained using double-precision floating-point operations have been performed, a more careful validation and possible algorithmic impacts need to be further analysed.

5.3 Impact on energy efficiency

Due to the small size of the pilot system it was not possible to execute all the large workloads, which were foreseen as benchmarks for tracking improvement in energy efficiency and for which accordingly reference numbers had been produced at the start of the PCP, and due to time constraints it was not possible (at the time of the writing of this report) for the project to run all downsized workload comparable with the new size of the pilot system.

The typical power consumption of the CPU Node at full load is about 500W, while it continues to consume about 100 W while in idle state. The MPC Node draws on average 276 W while it is in idle state. The power consumption while executing kernels depends on the number of DFEs that are used and the loaded bitstream. For instance, for BQCD each DFE consumes 12 W. During the execution of the main performance critical part of BQCD, i.e. the execution of the conjugate gradient solver, which is fully implemented on the DFE, the Compute Node consumes 100 W and the MPC consumes up to 372 W.

The following table is based on measurements performed by Maxeler on the delivered pilot system using smaller workloads, which had been provided within this team for development and testing purposes:

Table 18: Comparison of TTS and ETS on the Maxeler pilot system.

Application	Without DFE Acceleration		With DFE Acceleration	
	TTS [s]	ETS [Wh]	TTS [s]	ETS [Wh]
BQCD[4]	8*507	562 (estimated)	~2500	406
NEMO[5]	773	107 (estimated)	1099	160
Quantum Espresso[6]	447	62	195	42
SpecFEM3D	-	-	158	21

In comparison with a state-of-the-art CPU technology, energy-to-solution (ETS) could be reduced by 40-50% for BQCD and Quantum Espresso. For the workload used for BQCD, about 65% of the time was spent in the kernel that was fully ported to the DFE. For more realistic workloads this fraction is typically in the range of 70-80%, i.e. energy savings could be higher for larger workloads.

For NEMO a deterioration of both time-to-solution (TTS) as well as ETS can be observed. Maxeler attributes this to inefficiencies for an intermediate software layer used to interface the NEMO Fortran code and the kernels running on the DFEs. Their developers claim that TTS could be reduced to about 470 s and a reduction of ETS by about a factor 2.

For SPECFEM3D no new reference figures could be produced at the time of writing this report.

5.4 Schedule and timing

While the R&D services had been performed according to schedule, the shipment of the pilot system was delayed. The delivery was originally scheduled for September 2017, but had to be postponed due to the late availability of the CPU Node, which is based on a brand new processor type from AMD. The hardware has eventually been delivered on 20.10.2017 and immediately installed the days thereafter. The physical integration and base installation, which allowed for early user access, could be completed by 27.10.2017.

5.5 Impact on Maxeler roadmap

This PCP enabled Maxeler to produce a proof-of-principle for the use of their technology for applications as they are used on high-end PRACE HPC systems. In this context a number of tools, technologies and technological improvements have been developed that help using the Maxeler technologies for such kind of applications. Internally, Maxeler could built-up experience with a number of relevant applications that share features with a large fraction of applications used on PRACE Tier-0 systems.

Based on the results from this PCP and the opportunities resulting from the deployment of the technologies at a leading European supercomputing centre, we expect the following impact on the vendor roadmap:

- Technical roadmap:

[4] Performing either 8 runs without DFE acceleration or running 8 copies of the problem on 8 CPU cores of the CPU Node and one DFE each.
[5] Using 8 copies, each using 8 cores on the CPU Node and one DFE.
[6] Single copy of the application, using a single DFE, only.

  - Further development and enhancement of tools required for porting scientific applications to the Maxeler architecture.
  - Optimisation of the hardware design for floating point intensive applications that also needs high memory bandwidth and would thus benefit from high-bandwidth memory technologies.
- Commercial roadmap:
  - Opening of new market opportunities for selling Maxeler hardware solutions to HPC data centres both in the academic as well as the commercial sphere.
  - Broadening of the customer basis for services based on Amazon EC2 F1 Instance.[7]

5.6 Lessons Learnt

- The solution proposed by Maxeler starts to make the use of FPGAs for real-life PRACE applications, i.e. scientific computing applications with a high ratio of floating point operations, a realistic option, although it still involves significant challenges and short-comings.
- Adopting a data-flow programming paradigm seems a very promising avenue to efficiently implement kernels of scientific applications on FPGAs. This requires, however, a significant redesign of the application or the relevant kernels. The return of such an investment may be very high in terms of reduced time-to-solution and significantly improved energy-to-solution.
- The required FPGA resources can be significantly reduced when using reduced precision of fixed point arithmetics. For complex floating point intensive kernels this may be mandatory as otherwise the kernel does not fit in the available FPGA. As the use of lower precision arithmetics can also help to improve performance on commodity hardware, this observation is yet another motivation for computational scientists to perform research on the consequences of using reduced precision or for the development or adaptation of mixed precision algorithms, where the bulk of floating point operations is done in reduced precision but the final result remains at the desired high precision level.[8] Comparison of results obtained using these mixed precision and standard double-precision floating-point operations, and especially their impacts on the precision of the final result is a promising way of research.

6 Lessons learned from the E4 Power-8+/Pascal pilot system

6.1 Description of the system

The E4 pilot system as well as the product line it originates is named D.A.V.I.D.E. (Development for an Added Value Infrastructure Designed in Europe). In what follow, we will use DAVIDE to refer to the E4 pilot system.

[7] Amazon's EC2 F1 Instance comprises compute nodes with FPGA boards that are very similar to Maxeler's MAX5 solution. Maxeler acts as a partner of Amazon (see: https://aws.amazon.com/ec2/instance-types/f1/partners/) that enables customers to use this hardware offering in the cloud.

[8] For a use of reduce precision arithmetics on GPUs for simulations of Lattice Quantum Chromodynamics (LQCD) see, e.g.: M.A. Clark et al., "Solving Lattice QCD systems of equations using mixed precision solvers on GPUs," Comput.Phys.Commun. 181 (2010) 1517-1528 (doi: 10.1016/j.cpc.2010.05.002).

DAVIDE is based on the OpenPOWER architecture (IBM POWER Processors with GPUs) and it has been built using generally available hardware components plus custom hardware and an innovative middleware system software, to maximize the exploitation of energy efficiency feature of the pilot system.

6.1.1 *System description*

DAVIDE is composed by 45 nodes connected with an Infiniband EDR 100 Gb/s networking (by Mellanox), with a total peak performance of 990 TFlops and an estimated power consumption of less than 2KW per node. Each node is a 2U OpenCompute (OCP) form factor and hosts two IBM POWER8 Processors with NVIDIA NVLink and four Tesla P100 data center GPUs, with the intra-node communication layout optimized for best performance.

Total number (racks)	3
Total number of nodes	45 (compute) + 2 (service & login nodes)
Compute node form factor	2 OU
SoC	2xPOWER8 NVlink
GPU	4xNVIDIA Tesla P100 HSMX2
Network	2xIB EDR, 2x 1GbE
Cooling	SoC and GPU with direct hot water
Heat exchanger	Liquid-liquid, redundant pumps
Max performance (per node)	22 TFLOPs (double precision), 44 TFLOPs single precision
Storage	1xSSD SATA
Power	DC power distribution

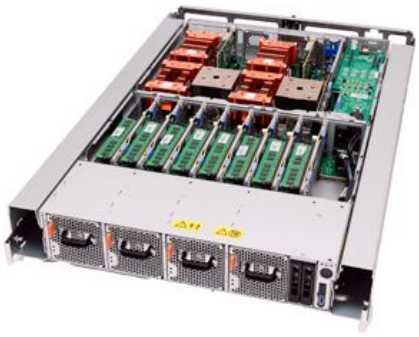

Figure 23: DAVIDE compute node

Compute node

- Derived from the IBM® POWER8 System S822LC (codename Minsky).
- 2 21” Open Rack Enclosure with integrated piping & power distribution.
- IBM Power8-based node in OCP form-factor, with leading edge features specifically engineered for HPC workloads.
- Two IBM POWER8 processors with NVlink and four NVIDIA Tesla P100 GPUs using a HSXM2 form factor
- Differently from Minsky, DAVIDE uses direct liquid cooling for CPUs and GPUs.
- Each compute node has a peak performance of 22 TFLOPS and a power consumption of less than 2kW.

Liquid cooling

- Direct hot-water cooling (about 27 °C) for the CPUs and GPUs.
- Extremely flexible and requiring minor modifications of the infrastructure.
- Each rack has an independent liquid-liquid or liquid/air heat exchanger unit with redundant pumps.
- The system has internal pumps on the GPUs. Each Rack has its CDU.
- The compute nodes are connected to the heat exchanger through pipes and a side bar for water distribution.

Compute accelerators

- The system is coupled with four NVIDIA Tesla P100 HSMX2 per node with NVLINK interconnect providing:
- 5.3 TFLOPS of double precision floating point (FP64) performance
- 10.6 TFLOPS of single precision (FP32) performance
- 21.2 TFLOPS of half-precision (FP16) performance

- A single link supports up to 40 GB/s of Bidirectional Bandwidth. The NVLink implementation in NVIDIA Tesla P100 supports up to four links, enabling ganged configurations with aggregate maximum bidirectional bandwidth of 160 GB/sec.

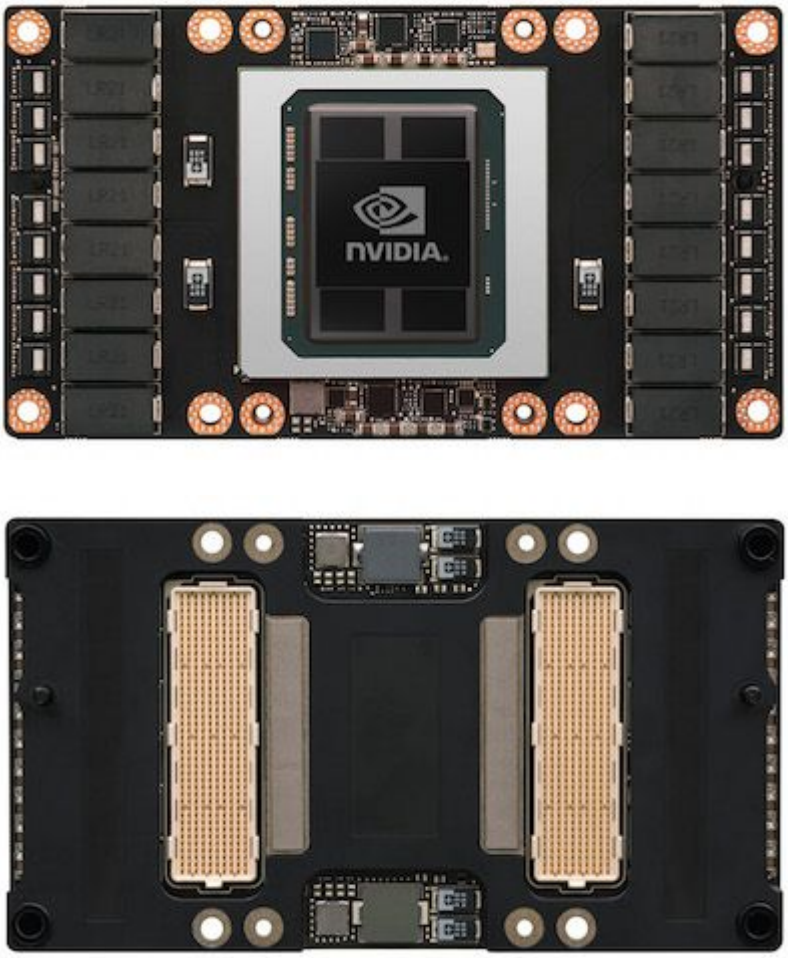

Figure 24**:** NVIDIA Tesla P100

6.1.2 *Energy efficiency aspects of the design*

Technology related

- GPUs and fast host processors reducing overall time to solution
- Direct hot water cooling
- SSD storage

Infrastructure for Energy monitoring and optimization

A key feature of DAVIDE is an innovative technology for measuring, monitoring and capping the power consumption of the node and of the whole system, through the collection of data from the relevant components (processors, memory, GPUs, fans) to further improve energy efficiency. The technology has been developed in collaboration with the University of Bologna.

Key features of this solution are the following:

- Off-the-shelf components: Most of the hardware components are off-the-shelf and thus the effort to design custom hardware components is avoided;

- High speed and accurate per-node power sensing synchronised among the nodes: The integrated hardware components allow measuring the power consumed by a node with a high accuracy. The measurements are performed with a high frequency to obtain accurate estimates of the consumed energy. Due to a good synchronisation of the real-time clocks of the involved components, it becomes possible to correlate the measurements performant at different places of the system.
- Data accessible out-of-band and without processor intervention: Power and energy are measured by (simple) dedicated hardware without the need for support by the processor. This means that the application performance is not affected by these measurements.
- Out-of-Band and synchronized fine grain performance sensing: The additional hardware is furthermore capable to read processor performance figures and thus allows for sampling also this additional information without support by the processor and thus without affecting application performance.
- Dedicated data-collection subsystem running on management nodes: The information sampled by the additional hardware are collected through a sub-system, which is dedicated to this task and thus does not affect the remaining part of the system.
- Predictive Power Aware job scheduler and power manager: The collected information is used by a Predictive Power Aware job scheduler and a power manager, which ensure that the power budget defined by the operators of the hosting data centre is not exceeded. Based on historical data a prediction of the power consumed by a job is made, which is waiting for being scheduled by the batch queueing system (here SLURM is used).

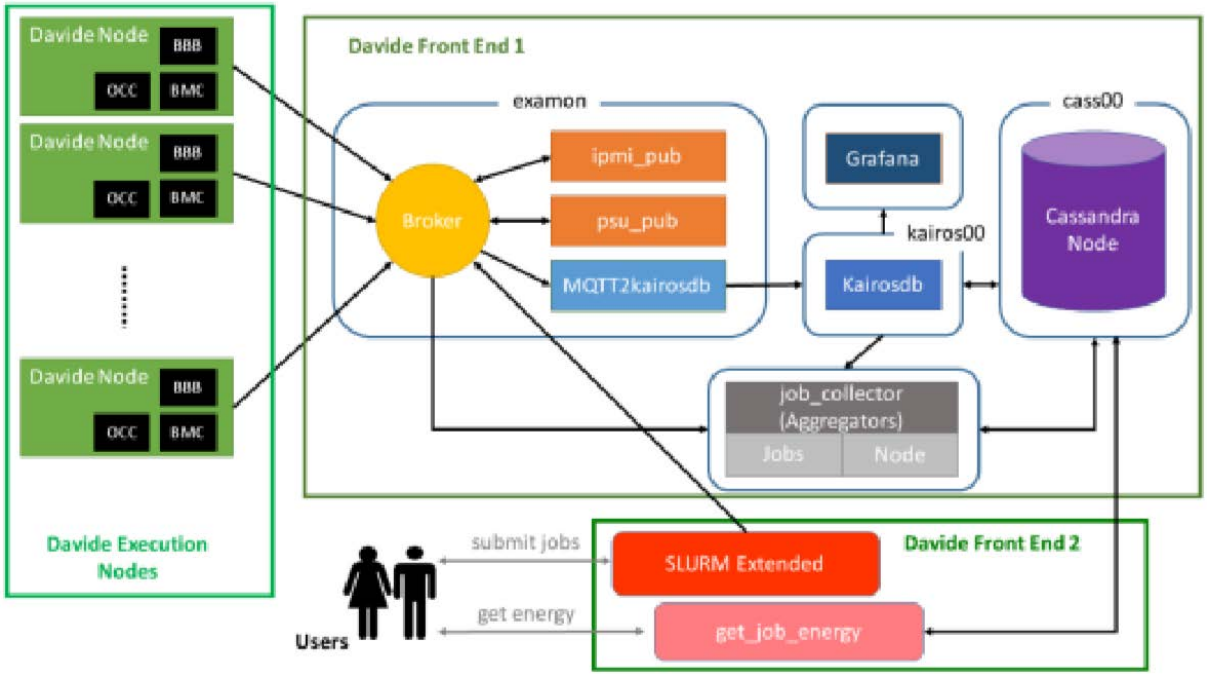

Figure 25: EXAMON: The general architecture of the monitoring framework

The general architecture of the monitoring framework installed on DAVIDE. is portrayed in Figure 25. The following paragraphs give a more detailed description of the components.

The Power monitoring extensions consists of a set of agents running outside the computing components of the nodes, but tightly coupled with them. These agents monitor the power consumption of each computing node at the plug. From the hardware point of view, they are

composed of (i) power sensing components, namely the sensors for measuring current and voltage, and (ii) a dedicated embedded monitoring device, which sample, pre-process and send data via Message Queuing Telemetry Transport (MQTT) to the framework data collection backbone. As embedded monitoring device, an open hardware platform, namely the Beaglebone Black (BBB) has been used. The original design of the Power Distribution Board (PDU) has been modified to include these components and, at the same time, exploiting the already available hardware.

To measure the overall node power consumption, the current and voltage sensors are placed between the busbar and the DC-DC converters (the busbar and the DC-DC converters supply all the processing and electrical components within the node). The voltage is measured with a voltage divider connected to the P12V signal, while the current with a current mirror and a shunt resistor. The voltage divide, the current mirror and the shunt resistor are connected to the BBB via a low-pass filter. The same current mirror already used by the Board Management Controller (BMC) has been used to measure the overall node power consumption, but with a coarser time granularity, as fully compliant with the specifications.

The optimized software running on the BBB (namely the bb_pub daemon) exploits the built-in ADC to sampling data at 50 kS/s, which is 50x faster than best state-of-the-art systems.

D.A.V.I.D.E. features an out-of-band monitoring system of the performance of each node, leveraging the POWER8 characteristics. In the standard OpenPOWER system, the on-chip controller (OCC) has (250 µs) periodical access to a wide set of power and performance readings. These values include a wide set of per core and per processor performance counters and a per main component power consumption. In the standard firmware, these sensors readings can be retrieved by an agent connected to the Ethernet to the BMC by a IPMI raw command. These connections are called AMESTER in power system. In DAVIDE the standard firmware has been extended with a set of new AMESTER subcommands which are capable of reading the AMESTER sensors with faster speed. The improvements allow to read up to hundred AMESTER sensors in one IPMI command. To periodically gather these sensors with DAVIDE a python SW daemon has been created. Overall, in DAVIDE. 242 AMESTER sensors per node are read every 10 seconds.

The monitoring backend is called Examon. The main components of the Examon (a framework for data collection, storage and analysis for exascale clusters) framework are:

- Sensor Collectors: These are the low-level components having the task of reading the data from the several sensors scattered across the system and deliver them, in a standardized format, to the upper layer of the stack. We can distinguish collectors that have direct access to hardware resources like IPMI, AMESTER, BBB, and collectors that sample data from others applications as batch schedulers (i.e. SLURM).
- Communication layer: The framework is built around the MQTT protocol.
- Storage layer: Examon provides a mechanism to store metrics mainly for visualization and analysis of historical data. E4 use a distributed and scalable time series database (KairosDB) that is built on top of a NoSQL database (Apache Cassandra) as back-end.
- Applications Layer: The data gathered by the monitoring framework can serve multiple purpose, as presented in the application layer. Data can be visualized using web-based tools or, for example, machine learning techniques can be applied to build predictive models or online fault detection algorithms.

The job dispatcher installed in DAVIDE is an improved version of SLURM. The base dispatcher was extended in order to collect information regarding the job running on the system; the information is collected exploiting underlying SLURM APIs and then is sent to the Examon framework through MQTT protocol.

The power cap can be set specifying a corresponding value in SLURM configuration file. In the configuration file the system admins can set the power cap for each node through the following parameter: *PowerParameters=cap_watts=600* The system power cap enforced by the extended SLURM is then obtained by multiplying this value by the number of nodes in DAVIDE. The modified SLURM forbids the power to exceed a given budget; this is done by preventing the admission of new jobs if their power consumption would violate the constraint. In practice, the power consumption is seen as an additional resource (on top of the standard number of nodes, cores, etc.). The power prediction for a job needs to be made only once: when the scheduler considers a job that has already been held in a previous iteration its estimated power was already stored internally.

A key element of the power capping module is the need to estimate the power consumption of HPC applications before their execution, in order to decide whether a job would violate the constraint or not. The data collection framework allows the development and usage of power prediction models (using the methodology outlined in [2]).

The current power estimation module employs a machine learning approach to predict the power consumptions.

6.2 Suitability for general purpose HPC

6.2.1 *Ease of code development and porting*

DAVIDE is based on a heterogeneous architecture, and then the applications can exploit energy efficiency only if they are able to exploit the accelerators. Evaluating the complexity of the porting of an application to a heterogeneous environment is not the main objective of this PCP, but we can observe that in the case of DAVIDE only the applications with strong commitment of the developers toward the support of GPU (SPECFEM3D, QE and BQCD), have been able to exploit the pilot system. In the case of NEMO it was not possible for E4 to port it to GPU, even if they have tried several options and to liaise with the community of developers. Our understanding is that enabling a code for a GPU architecture require a deeper revision of the application.

Concerning code development, DAVIDE features all the tools available in a standard Linux HPC environment, with the addition of NVIDIA compiler suites (PGI as well), and IBM compiler suites and libraries. The feedback from users is that the IBM compiler suites are less mature for heterogeneous programming with respect of GNU or NVIDIA/PGI suites. The presence of multiple mathematical libraries for the same functions (ESSL, CUBLAS, BLAS/LAPACK), that require to be addressed/linked explicitly by the developers is also creating some difficulties in adapting the build system to take advantage of the best option, which may vary from case to case.

Energy efficiency features, do not have a direct impact on code development, but the energy and power reports available for each job and each system component can certainly help developers in application optimizations.

6.2.2 *Energy monitoring/prediction*

The energy monitoring framework, described above in detail, has been designed to be as open as possible and being portable to different architecture. In particular the prototype of PCP phase II and pilot system for PCP phase III have very different node architectures (ARM vs OpenPower), but the monitoring hardware and the software framework was the same. Then we can conclude that the monitoring framework E4 develop within the PCP can be used in a general purpose HPC system. What can change from one system to the other are the plugins

for the collection of sensors data. The power capping functionalities and the related energy prediction system, these are not specific to DAVIDE, since they are interfaced to the scheduler (SLURM in this case). The accuracy of this functionality depends on the quality of the energy measurement and the energy to solution accounted to the users' jobs.

6.2.3 *System usability*

DAVIDE is based on OpenPower architecture, NVIDIA GPU and Linux operating system, from this point of view the usability is quite the same of other system of the same class. The energy monitoring data can be accessed through standard Grafana web portal, where the users or the administrators could browse easily the data collected from the monitoring framework. At the level of the terminal, in order to get the energy to solution of a job the user has to issue a specific command. To exploit the power capping functionalities, the system administrators need to set specific parameters in the SLURM configuration file. The cooling system do not require specific datacentre adaptation, in the case of DAVIDE Cineca has connected the OCP racks manifold with the cold water loop already serving the KNL partition of CINECA' production system called Marconi.

6.3 Impact on energy efficiency

The energy efficiency of DAVIDE comes, largely, from the adoption of heterogeneous architecture, with 4 accelerators of latest generation (NVIDIA P100) per node. The monitoring framework, the middleware software E4 develop to implement scheduling policies and power capping, could help exploiting the characteristics of the architecture. The drawback is in the porting of applications. In fact, applications need to support the heterogeneous paradigm, and leverage CUDA accelerators programming languages. If an application is not enabled to run on GPU, the performance and efficiency of the architecture can be very poor, since the power consumption of the architecture is typically higher than a homogenous system but the floating point performances are typically lower.

In the case of the 4 PRACE PCP applications and HPL, the following consideration could be done.

HPL on DAVIDE exploits GPU efficiently since >90% FLOP is provided by the accelerators. The precompiled binary, provided by NVIDIA, is compatible with CUDA 8.0 and IBM Spectrum MPI. ESSSLSMP library is needed for that part of the benchmark not intensive enough to be offloaded to the GPU. Results in term of Flops/Watt for the air cooled intermediate pilot system and liquid cooled final pilot system are reported on Green500 list of June and November 2017, were DAVIDE was ranked high in the lists, being the highest OpenPower system.

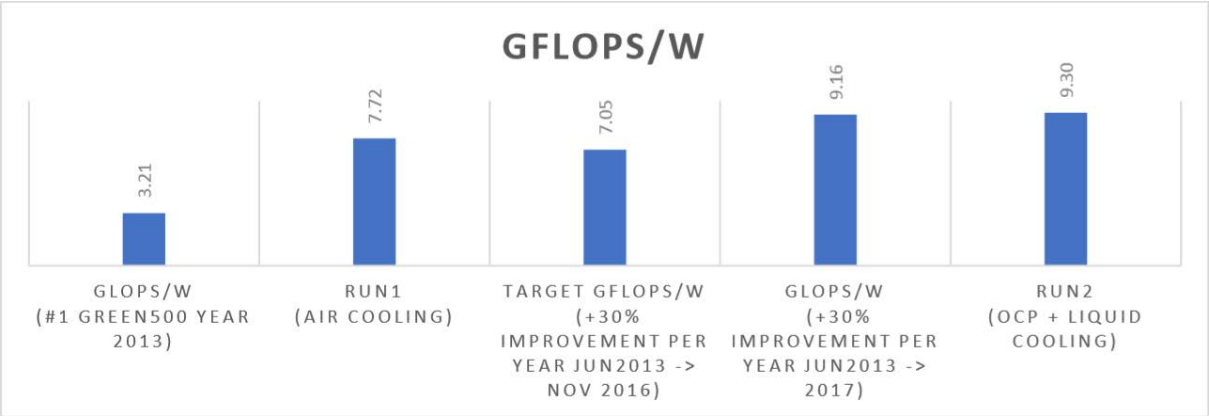

Figure 26: HPL energy efficiency of DAVIDE with different cooling systems as compared to the #1 system of the Green500 in 2013 (Eurora by Eurotech co-funded by PRACE as well)

From the figure above, it is to be noted that direct liquid cooling has an important role in improving the energy efficiency (7.72GFlops/Watt @ AIR vs 9.30GFlops/Watt @ direct water cooling). All the Linpack tests have been executed with an inlet/outlet cooling liquid temperature of 29/45°C.

Concerning Quantum ESPRESSO, a good energy efficiency performance was possible by the usage of the latest version 6.1 with the support for CUDA Fortran. It is designed to exploit CUDA IPC for intra-node GPU-to-GPU communication and CUDA-aware MPI inter-node GPU-GPU communication. The combination of both features on POWER8 + GPU architecture is extremely scalable. All GPU-accelerated Quantum ESPRESSO runs was performed using a version of the code compiled with PGI 17.10, Open MPI 1.10.2 and IBM ESSL 5.4. The CUDA FORTRAN runtime uses CUDA 8.0, Open MPI 1.10.2 is compiled with CUDA-aware support.

NEMO was not enabled to use GPU and the performance in term of energy to solution of this code are poor, since it run only in the host processors. E4 tried to use different NEMO versions with support to heterogeneous hardware, but it seems they were far from usable and satisfactory. On the other hand, the effort of re-factoring an application to run on GPU was too high to be done by the contractor within the PCP budget.

SPECFEM3D comes with a GPU-accelerated set of key routines with are extremely well optimized and maintained since long time. During the heaviest computation phases, there is no need to exchange data across MPI so there are no opportunities for SPECFEM3D to explore NVLink capabilities between GPU. The provided version (6.0.0) could compile neither the CPU nor the GPU path for the current system out-the-box. The CPU version of 7.0.0 performed better than that of 6.0.0 on the current system and minimal trivial changes were required for the input file to support the 7.0.0 GPU path.

The size of the reference dataset was too large to fit in the D.A.V.I.D.E main memory, and to solve this issue, a smaller revision of the input dataset was prepared.

BQCD can run on GPUs by employing the QUDA library. QUDA has a BQCD interface to its. QUDA is a library for performing calculations in lattice QCD on graphics processing units (GPUs), leveraging NVIDIA's CUDA platform. QUDA latest stable version is 0.8.0 and it supports interfaces to BQCD. It compiles without any problem using GCC 5.4.0 and OpenMPI 2.1.0. BQCD 5.1.0 is the latest public release available of the code.

In Table 19 we compare reference numbers for TTS and ETS provided by PRACE 3IP at the beginning of the PCP to the numbers, which E4 obtained on their pilot system DAVIDE For SPECFEM3D the size of the computed problem had to be reduced in order to make it fit on the pilot system. For this modified workload, however, no reference numbers are available.

Table 19: Comparison between reference and measured values for TTS / ETS on E4 pilot system

Application	TTS [s]		ETS [kWh]	
	Reference	Measured	Reference	Measured
BQCD	1584	16921	169.6	158.6
NEMO	1942	842670	364.5	1002.8
Quantum Espresso	3216	1063	342.6	13.7
SPECFEM3D	--	991	--	5.7

E4 found that performing an extrapolation of application performance from 1 to 100 PFlops is quite tricky and complex, especially because there are no exact performance models of complex applications like the ones under examination and because the ability of achieve large scale is heavily correlated with the type on science performed.

Regarding the four applications in the PRACE PCP benchmark suites, the following consideration could be done:

- Quantum ESPRESSO will benefit of a multi-petaflop machine only by allow scientists to run multiple copies of modest size calculations in parallel, increasing throughput (jobs per amount of time) rather than improve efficiency at very large scale. The motivation behind this usage model is mainly science driven, with material screening becoming more and more popular.
- NEMO needs to undergo huge software re-engineering to be able to exploit architecture with large vector units or massive parallel architecture. This can be done only by NEMO core developers which need to plan for a future that is quickly approaching. I/O is also a big bottleneck for weather and climate codes, something not to be underestimated.
- BQCD (and all QCD applications) will continue to scale for as long as more compute is available. There is a considerable amount of effort around Domain Scientific Library (DSL) like QUDA or GRID that are design to maximize the exploitation of target architectures. More compute will allow QCD collaborations to run large simulation and test more complex assumption about fundamental physics.
- SPECFEM3D will continue to exploit the high-memory bandwidth of future accelerators and all test done so far show scaling will continue for as long as bigger problems are considered.

In a GPU-accelerated architecture we expect the GPU accelerator achieve a performance leap of at least 2~3x within the same power envelope between now and 2020. Two systems will be deployed within few years in USA by the Department of Energy: SUMMIT (ORNL) and SIERRA (LLNL). Both systems, based on IBM POWER9 processors coupled through nVLINK2 to 4 to 6 NVIDIA Volta GPUs (two products already announced and present on the market), promise to hit ~100 and ~250 PFlops peak respectively by end 2018 / beginning 2019.

6.4 Schedule and timing

The whole system was fully assembled in air-cooled configuration in April 2017 at E4's facility in order to perform baseline performance, power and energy benchmarks. The system ranked #299 in TOP500 and #14 in GREEN500 in the June 2017 list. Between June and

September, the system has been converted to liquid cooling. Note that the supplier of the liquid cooling components has changed from CoolIT to Asatek, since CoolIT withdrew from delivering component for OpenPOWER system. This has caused approximately one month of delay in the timeline. The pilot system has been installed at CINECA in the September/October timeframe. The system is currently available to a select number of users for porting applications and profiling energy.

6.5 Impact on E4 roadmap

E4 believes that the market is willing to change technology paradigm and move in a more competitive scenario with silicon makers competing each other. Based on the experience of the PCP they are then committed to design a class of clusters with open hardware, sharing the design of the components with the community. The philosophy is the usage of existing components when available, and designs the missing one with a strong attention to the costs. An example, DAVIDE, besides being a pilot system, is the first of his kind in E4 products list, and is now available to other customers.

Thanks to the OCP design and modularity of the different components, E4 intend to offer the different technologies developed in the PCP, such as the monitoring framework to different type of architecture. E4 Computer Engineering has already received a number of RfI (Requests for Information) for DAVIDE. As an example, a leading Chinese system integrator (name not disclosed for competitive reasons) has initiated an in-depth evaluation of the technologies developed within the PRACE-3IP PCP project, with the objective of integrating these technologies in its own portfolio. Another example is the interest shown by IBM Corp for the Power Capping and monitoring technologies, and the assessment currently in progress to embed such technologies in its next generation products. With this approach, E4 will target many market segments from the single departmental computing facilities to the large research centre with multi petaflops.

6.6 Lessons Learnt

Porting applications to and heterogeneous architecture featuring GPUs, quite often require the involvement of the applications developers, and cannot be managed as porting task for the purpose of a benchmark. None of the applications included in the benchmark suite were properly suited for GPU, but in the timeframe of the PCP many developers had the opportunity to refactor the applications, and latest version of QE, BQCD and SPECFEM3D are enabled to use GPU. E4 have then used the most recent version. The lesson learnt in this case is that keeping the same version of the code for a period of three year turn out to be not possible, since code evolves, and older versions are not maintained neither ported to new HW. Keeping the same baseline in the benchmark suite can be done for synthetic well established benchmark (like HPL), but for full applications it would be better to re-baseline the application versions and reference values after the last phase of the PCP.

As a side effect of the problem in keeping the same baseline for the applications along all the PCP, the number of energy to solution and time to solution are less reliable, since other factors (change in the applications) could have contributed to the improvement of the applications performances and energy efficiency.

In general, the PCP goals and the rules on the R&D activities, have motivated E4 to liaise with researchers in the academia and applications developers, contributing to the improvement of the TRL on different components.

In the case of E4, the possibility to validate two cooling setup (air and direct liquid cooling), has demonstrated to have pros and cons. The comparison in the energy efficiency between the two cooling systems was useful to understand the advantages on the direct liquid cooling technology; on the other hand, this "double validation" introduced a significant delay in the deployment of the final pilot system at the CINECA premises. It has to be said that this was not the original plan by E4, but they were forced to start the deployment in their premises because of the delay in the decision about the hosting of the pilot systems. E4 at their premises did not have the possibility of cooling the pilot system with water, so that they did the first validation using air cooling only at their premises.

Considering the amount of budget for phase III and the critical aspect related to timing of pilot system delivery, probably a single site visit to the vendor premises is not enough to have the proper exchange of information with all stakeholder. This is especially true since in the last phase there is no selection and ranking of contractors, and the risks connected with the procurement process are small.

As a conclusion, for DAVIDE by E4 we may say that it reached the expected level of TRL for the pilot system, and it could become a valid option for HPC solutions, competing with the others already present in the market.

7 Lessons learned from the E4 ARM prototype system

In PCP phase II the prototype of E4 was designed with the idea of taking many different emerging technologies supported by large markets and glue them together to obtain the best compromise of price, performance and efficiency. This approach has demonstrated of having several drawbacks, especially due to the different TRL of the different components, and the cost of integration, especially if there is a dependency with propriety components like network and GPU drivers, requiring a commitment from the third-party suppliers.

7.1 Description of the E4 ARM prototype system

7.1.1 *System description*

E4 build three prototypes for phase II, with increasing complexity and integration. They are based on cost effective technologies, such as:

- DC power distribution through the rack, to increase reliability
- Liquid direct cooling high temperature (optional), to optimize energy efficiency
- Open standard, to remove vendor or ODM/OEM lock-in
- ARMv8 SoCs
- File System on Demand
- Energy-aware scheduler and programming interface
- Hardware tuning on application energy profile

The first prototype built for the purpose of Phase II is a cluster (codename Overkill) of 4 compute nodes and 1 frontend server. Each compute node is equipped with an X-gene1 SoC and a K40 GPU. The communication network is Infiniband FDR. This cluster does not have an embedded power monitoring, and to retrieve the values of energy-to-solution it has been connected to an external power meter. This initial prototype is limited in the computational

capability of the CPU (38.4 GFlops) and in the PCIe bus throughput, 6GB/S from host to device and 3 GB/s from device to host.

Figure 27: ARM+GPU compute node

The second prototype cluster (codename Tomberry) is the intermediate technology step to achieve the final prototype stage. Similarly, to Overkill is a 4 compute nodes cluster plus frontend server. The node is a dual socket Cavium ThunderX SoCs with a total of 96 cores and one k40 GPU. The communication network is Infiniband FDR. While the rack configuration is identical to Overkill with air cooled 2U servers, the energy monitoring is integrated into the server collected out of bound values.

The final prototype (Sluggish) has been installed in March 2016. The main difference is in the server density. The system is integrated at rack level, following the specification of the Openrack standard. Power conversion is centralized and distributed through a bus bar on the back of the rack. Fans are managed by a rack management module, which integrate also the BMC of each independent server. In addition, E4 has modified the rack to host the liquid cooling manifold and the pipes. The specifications of the compute nodes are identical to Tomberry ones, together with the network and storage setup. The power monitoring sensor is integrated in the power distribution board of the compute nodes and connected to an embedded out-of-bound monitoring system. Sluggish have a power efficiency and power management capability close to those expected by the pilot system.

7.1.2 *Software environment*

The R&D team was focused on having a GPU (Nvidia K40) and an Infiniband board (Mellanox ConnectX3) fully working. E4 initially tested both GPU and IB with Kernel 3.18 on a single socket Cavium ThunderX initial board: after some changes in Kernel sources and configurations they managed to have everything working properly. The Cavium ThunderX boards (on sluggish cluster), shipped with new additional features (like CPU power management), needed a newer version of Kernel (4.2) in order to implement these new features, so E4 had to restart work from scratch. Unfortunately, at this stage, both Mellanox and Cavium did not provide a sufficient effort and with a very small timeframe (few weeks) E4 R&D team was unable to fix the issues. E4 was able to setup a ThunderX cluster with Infiniband connectivity but without GPUs.

The GNU compiler suite is available on every Linux distribution which support aarch64 (ARMv8 architecture). E4 selected Ubuntu 15.10 coming with GNU 4.9.2 as default. E4 additionally build GCC 5.2.0 in the system to be able to utilize latest enhancements inthe compiler to explicitly target Cavium ThunderX SoC (flags "*-march=armv8 -a -mtune=thunderx*"). ARM itself is directly involved in contributing into the GNU compiler code in order to generate better optimized code natively. PathScale EKOPath Compiler was the only commercial compiler available that support ARMv8 architecture and includes specific optimizations to target Cavium ThunderX SoC. Boh C (athcc) and FORTRAN (pathf90) also support OpenACC 2.0 extensions (flags "*acc -device=kepler*"). Documentation of supported OpenACC features was quite poor.

Numerical libraries for ARM were still in very early stage. ARM is actively working on porting basic numerical libraries (BLAS, LAPACK and FFT) and basic math routines via algorithmic enhancements (in, cos, tan, sinf, cosf, tanf, logf, expf, pow, ...). Parallel numerical libraries like ScaLAPACK were not yet directly in scope for ARM HPC team since their focus it the micro-architecture optimization. The work currently on-going is enablement of new ARM-based hardware and a direct engagement with key partners for application porting. On the other side NVIDIA GPU will continue to deliver optimized CUDA libraries cuBLAS and cuFF are the more relevant for on current Kepler and the future Pascal architecture.

7.1.3 *Energy efficiency aspects of the design*

One of the key aspects of the prototype is the out of band energy monitoring based on the beaglebone Black Board. BBB is an embedded IoT/smart sensor platform built around the TI Sitara AM3358BZCZ100, 1GHZ, processor, which features an ARM Cortex A8 processors. E4 used these key features of the BBB to implement specific monitoring services: (i) Analog-to-digital conversion and data processing, (ii) per-components power measurement, (iii) an MQTT data transfer front-end. The block diagram of the monitoring engine and its integration is represented in Figure 28: block diagram of the monitoring system Figure 28.

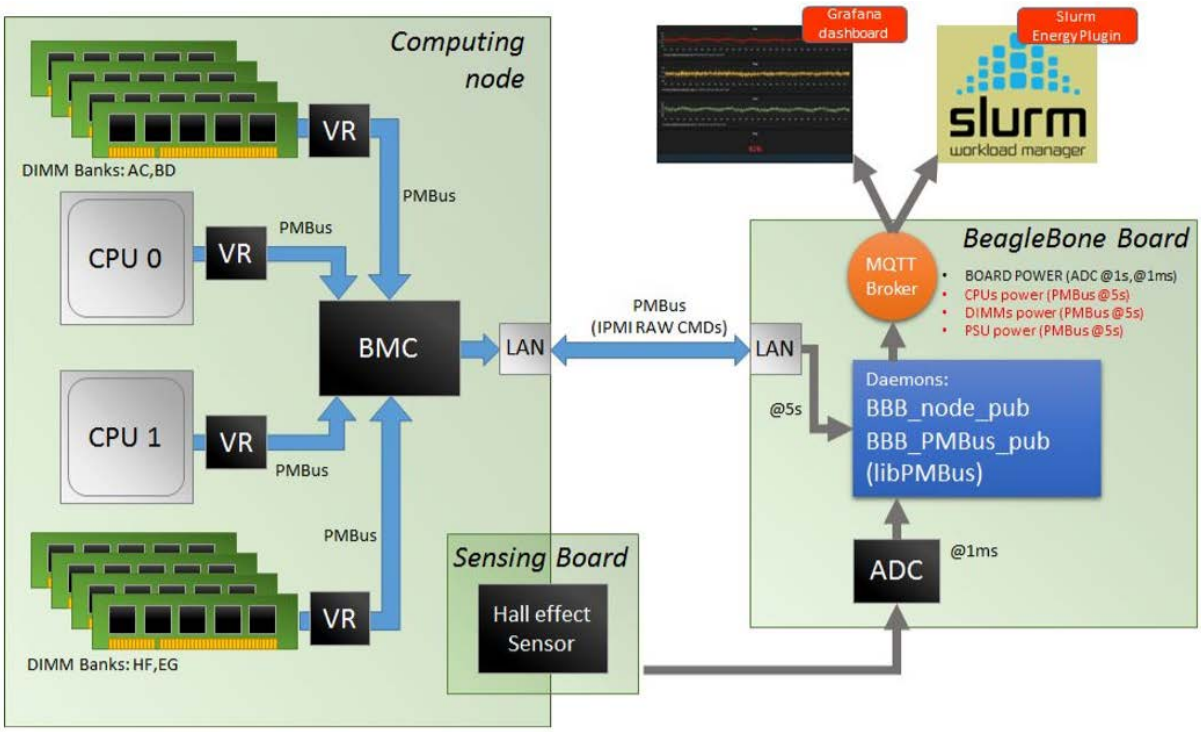

Figure 28: block diagram of the monitoring system

Other aspects of the design related to energy efficiency are the OCP with bus bar to distribute electricity, the direct liquid cooling of the CPU and GPU, the software for power management.

7.2 Suitability for general purpose HPC

The prototype for phase II was not meant for general purpose HPC, but as development vehicle to validate technologies.

7.2.1 *Ease of code development and porting*

The code development and porting on phase II prototype integrated by E4 require significant effort and commitment from components providers. Nevertheless is expected that the HPC systems based on ARM architecture will become more and more general purpose, with an increasing number of applications ported and development tools (see https://developer.arm.com/hpc/hpc-software/categories/Applications). It is remarkable that out of 9 applications already supported on ARM architectures, 2 (QE and BQCD) were investigated in phase II prototype by E4.

7.2.2 *Energy monitoring/prediction*

The energy monitoring system is probably one of the most interesting innovation of phase II prototype, based on a flexible out of band HW, capable of high frequency sampling of different components. Prediction feature was not yet included in the phase II prototype, but present in phase III prototype.

7.2.3 *System usability*

The prototype was installed at E4 premises and directly managed and used by E4 engineers. From the E4 deliverables we can deduce that the prototype was easy to use, but it is expected that support in the open source world of ARM based architecture is going to improve in the next year.

7.2.4 *User experiences and feedback*

No user had tested phase II prototype.

7.3 Impact on energy efficiency

The phase II prototype of E4 leverage ARM architecture (ThunderX), NVIDIA GPU (K40 & K80), OCP rack with bus bar and direct liquid cooling, out-of-band energy monitoring system, and energy optimization software. In term of energy two ThunderX sockets delivers 250GFlops of performance within a power consumption of 198Watts, thanks to the energy-efficiency APIs they can operate with only one core active (2.5Gflops) while consuming 72 W. In addition to the ThunderX power, the phase II system features a K40 Nvidia Card. An additional 80 W of power inclusive of the 256GBytes of DDR, SSD, IF card and board components, including the monitoring system. Indeed this value has been taken from real measurement on the sluggish cluster. This list does not contains the cooling cost accounted with the virtual datacenter model and leads to a cooling cost which consists of the 7.5% of the IT power (14% air and 86% hot-water cooling).

The phase II final prototype is capable of achieving a node energy efficiency of 3.27 GFlops/Watt which increases to the 3.7 GFlops/Watt when the power management APIs are used to statically shutdown unused ThunderX cores.

Considering this number and the cooling cost the extrapolated Energy-to-Solution (ETS) for the four target benchmarks for the different system configurations can be computed, with the assumption that the proposed 1Petaflop system will have the same TTS as the reference cases. These are computed by extracting the reference architectures flops (BQCD:840 TFlop/s, NEMO:1020 TFlop/s, QE:1050 TFlop/s, SPECFEM3D:815 TFlop/s). DC Energy is computed using the virtual data centre model. The computational engines (CPU and GPUs) are cooled with direct hot-water cooling (86% of total power) whereas the rest is cooled by air cooling (14%). This lead to an average PUE computed with the virtual data center of 1.075. ETS contains the sums of IT energy and DC Energy. The Phase II system thanks to the Energy Efficiency APIs is capable of obtaining a 30% of energy reduction for NEMO, similar performance for SpecFEM3D and BQCD and an energy loss of the 27% w.r.t. QE.

7.4 Schedule and timing

In PCP phase II, E4 produced 3 prototype systems (Overkill, Tomberry and Sluggish) of increasing complexity and integration. The prototypes were installed at E4 premises and shown to the PRACE PCP assessment committee during the site visit. The step by step incremental approach was good to reduce the risks, but the last one, with the final integration was delivered closed to the end of phase II, so there was not much time left for an exhaustive evaluation. On the other hand it was enough to understand, that a critical component (the thunder processor) was not at the right TRL to be integrated into the final pilot system for phase III.

7.5 Impact on E4 roadmap

E4 believe that the market is willing to change technology paradigm and move in a more competitive scenario with silicon makers competing each other. Based on the experience of the PCP they are then committed to design a class of clusters with open hardware, sharing the design of the components with the community. The philosophy is the usage of existing components when available, and designs the missing one with a strong attention to the costs. An example of flexibility to reduce the entry cost is the direct liquid cooling: it will be optional and installed only for those centre which have the cooling facilities capable to take advantage of it. With this approach, E4 will target many market segments from the single departmental computing facilities to the large research centre with multi petaflops.

8 User experiences and feedback

In order to get feedback from the larger community, the PCP pilot systems were opened in Autumn 2017 to a limited group of users from interested HPC projects. User experiences and feedback were collected from EoCoE and PRACE-4IP and are presented below. The plan is to open access to the systems for research purposes after the end of the project in the form of PRACE Preparatory Access, under supervision of PRACE.

8.1 EoCoE

The first week of October 2017, with the EoCoE CoE a two days workshop was organized on the PCP KNL pilot system.

People from Atos-Bull, CEA, CINECA, CINES, BSC, GENCI, IDRIS, Imperial College London, IRIT, IDRIS, JSC, Maison de la Simulation, gave talks during this application energy efficiency optimization oriented workshop.

Atos-Bull presented PCP architecture and the general concept of the energy software installed on the PCP supercomputer (BEO, HDEEViz and SLURM energy plugin).

Major effort was focused on the “hands on” session. For each software tools, a user manual was developed designed for the use of energy software and examples associated. The aim was to enable as fast and easy as possible usage of the energy software profiling tools available on PCP Pilot System.

During this workshop, users and developers from different area of interest had the opportunity to learn about the tools available on the Atos-Bull Pilot System and to perform hands-on experience with their codes. Users’ feedbacks were quite positive, as the proposed tools were considered as useful to understand the power consumption behaviour of their own applications. Some remarks and suggestions should be taken into account in the next release of these tools.

More information on EoCoE PCP workshop including presentation and recorded talks is available at [4]. This is an interesting first example of the exploitation that is expected to happen towards the users’ communities as outcome of the PCP.

8.2 PRACE-4IP

Within the PRACE-4IP H2020 project, a specific activity has explored the energy consumption of real-world application codes in the Atos-Bull KNL and E4 Power8+GPGPU systems with the aim to have comparable energy-to-solution results and suggestions on optimal run parameters. The codes investigated (e.g. ALYA, Code_Saturne, CP2K, GPAW, GROMACS, NAMD, PFARM, QCD, Quantum Espresso, SHOC and SPECFEM3D_Globe (already ported to accelerator) and GADGET and NEMO (newly ported)) are part of PRACE's accelerated Unified European Application Benchmark Suite[9].

Also, the HORSE+MaPHyS+PaStiX solver stack has been selected to be ported to Intel KNL, to represent “real case” of state of the art operational code, beyond usual benchmark codes. Focus here has been given to performing an energetic profiling of theses codes and studying the influence of several parameters driving the accuracy and numerical efficiency of the underlying simulations. PRACE-4IP work-package 7 activity on the PCP pilot systems are reported in details within D7.7 Performance and energy metrics on PCP systems [3].

8.2.1 *KNL pilot system*

Porting of the codes on the KNL platform was straightforward due to its similarities to x86 and no technical issues were encountered. Similarly, due to a very standard cluster environment (SLURM etc.), running of the codes was very easy from the user perspective. Achieving good performance (and thus energy consumption) was more of a mixed experience.

[9] http://www.prace-ri.eu/ueabs

Performance

So far, the KNLs have been configured to be exclusive on FLAT memory mode for the MCDRAM, which has limited the performance of some codes. Only codes that are able to fully utilize FLAT mode have achieved good performance and explore the actual limits energy consumption. In practice, many codes that have not been explicitly ported to KNLs, reach their optimal performance on KNLs when using the MCDRAM in CACHE mode. Energy results from codes that are not suited for FLAT mode can only be considered indicative of trends, not absolute limits.

Assuming a more flexible and/or heterogeneous configuration of the memory modes, all codes should reach good performance on the system.

Energy consumption

From the user perspective, it is very convenient to collect energy consumption results using the BEO provided in the system. BEO reports total energy consumptions during the job with a more fine-grained separation for the energy used in switches, disks and compute nodes.

It seems that the energy consumption is mostly a function of the total usage of processors / nodes, i.e. the number of nodes used times the runtime. For CP2K, initial results seem to suggest a slightly lower energy consumption for hybrid jobs (8 threads, 8 MPI tasks) compared to full MPI jobs (64 MPI tasks), especially when using tens of nodes. For most codes, the best energy-to-solution seems to be when using only a single KNL.

8.2.2 *Power8 + GPGPU pilot system*

Users report open source GCC compiler suite works fine as well NVIDIA/PGI compiler suites, or at least they are aligned with other heterogeneous solution available with x86 architectures. IBM suites works fine as well for the host processor, but creates some problems when dealing with the GPUs, it seems there are still too much bugs for large and complex applications. Moreover, the fixes are first released in the NVIDIA/PGI suites, and take some time to propagate in IBM compilers.

User reports that time to time the energy to solution reported for a job turn to be wrong (well sometime negative), this is true when the back-end of the monitoring framework is down or data are not available. In this case it would have been better if the command to query the monitoring framework returned an error code instead of a wrong value.

9 General lessons

In this section we consider general lessons that are either not specific to a single system or can be seen in multiple systems.

9.1 Impact of downstream component schedules

A recurring problem in this PCP was schedule delays and product roadmap changes from downstream technology providers. There is no indication that the period of the PCP was particularly volatile in this respect. It is common for schedules to slip and future product lines to change especially in the micro-electronics industry. The R&D component in a PCP means that it takes place over longer timescales and is more exposed to these changes than a conventional procurement. This is a risk that has to be born in mind when considering a PCP.

9.2 Interconnect

Though well within the scope of the PCP none of the participants chose to explore energy efficiency improvements related to the processor interconnect. This is unsurprising given the current state of interconnect technology. It is a highly specialised technological area where the key research and development takes place very far down the supply chain. Nor do we expect any serious energy inefficiencies in the way the HPC applications use the interconnect that might be addressed by better software solutions.

Until there is a disruptive change in networking technology (such as the commercial availability of silicon photonics) we see no evidence that mechanisms like the PCP will be effective in influencing the energy efficiency of HPC networking technologies.

9.3 File-systems

The energy efficiency of data-storage systems was within the scope of the PCP. The aim of the PCP was to support general purpose HPC work-loads. The current requirement of most HPC applications is to have the data-storage systems presented as Posix compliant parallel file-systems. None of the vendors chose to address this aspect of the system design in the early phases of the PCP. The file systems in the final pilot systems do reflect aspects of energy efficient design. For example the selection of Solid State Disks over conventional magnetic media and to host the file-system directly on the compute nodes rather than provision an additional set of file-system servers. However these are choices from within the space of conventional file-system design rather than new innovations which will result in new products. The evidence from this PCP therefore suggests that there is scope to influence storage system energy efficiency by setting appropriate procurement rules, but this should also be possible within a conventional procurement process and does not necessarily require the additional R&D support of a PCP.

9.4 Cooling systems

Cooling system efficiency is clearly a key component of total system energy efficiency. In particular it is well recognised in the industry that liquid cooling has significant advantages over air cooling in this respect. The relatively low heat carrying capacity of air means that significantly higher volumes of air need to be moved in order to achieve the desired level of cooling and therefore more energy needs to be expended in running the necessary fans than are required to run the pumps in an equivalent liquid cooled system. However air cooled systems are easier to host and operate so the majority of systems are still designed to be air-cooled with liquid cooling reserved for products designed for niche markets such as HPC, high density data-centres and at the low-end extreme performance gaming computers.

The PCP demonstrated a variety of research and development activities associated with cooling systems. The Bull product line is designed for the HPC space and was already largely water cooled. Some of the R&D effort within the PCP was used to increase the level of liquid cooling within this product line by extending it to the power supplies which previously had been air cooled. E4 also developed new liquid cooling solutions by modifying third-party compute nodes to convert them from air-cooled to liquid cooled.

Development of cooling systems seems to be particularly well suited to a PCP type mechanism. It is essentially mechanical design work that is clearly within the remit of the system vendor. Cooling systems are largely independent of micro-electronics, software and other technologies driven by the down-stream vendors. Cooling systems have to be designed

specifically for each product so there will be an additional cost associated with making a liquid cooled version of each product. Customers who wish to ensure that liquid cooling solutions are available can encourage this by helping to fund some of these development costs using a PCP type mechanism. This will be particularly useful if customers have specific requirements as part of a site energy strategy, such as wanting higher than normal inlet or outlet temperatures that make it easier to make use of waste heat. A PCP focused on cooling systems could then be used to enable some degree of co-design between system design and the design of the data-centre cooling infrastructure. Very innovative data-centre cooling infrastructure would run the risk of reducing the number of eligible vendors for subsequent procurements. It might be possible to mitigate this risk by also adding a small PCP phase to these procurements to help additional vendors develop compatible cooling solutions.

9.5 Energy aware scheduling

A number of energy aware scheduling and energy limiting software solutions were developed during the PCP, mostly as extensions to the SLURM job-manager. These systems attempted to predict the requirements of running applications and to keep them within a specified power budget by dynamically adjusting the clock frequency. Unlike existing dynamic frequency scaling where individual devices adjust their clock speeds these solutions adjust the clock-speed in a coordinated manner across all nodes used by a parallel job.

On conventional CPUs these software solutions showed a certain amount of potential however these failed to translate properly to the KNL based pilot system. As the KNL cores were relatively low powered there is less scope for dynamically changing the clock speed. In addition the high bandwidth 3D stacked memory means that memory and processor are in better balance. On a conventional processor running a memory bound problem there is more scope to reduce the clock speed (and reduce the energy consumption) without impacting time to solution much. Where the processor and memory is in better balance, any reduction in clock speed results in a proportionate increase in time to solution. The lesson here is that there is a risk associated with simultaneous innovations in both hardware and software.

9.6 Use of FPGAs

Within this PCP a new baseline for use of FPGAs for accelerating scientific applications could be established. A proof-of-principle for porting complex applications was made, which are widely used within PRACE and part of the PRACE benchmark suite. The results are encouraging in terms of potential improvements in term of energy efficiency. However, due to compromises in the numerical precision a firm conclusion in terms of reduction in energy-to-solution cannot be made. The research that was performed in this context, however, reemphasis the need for research on exploiting reduced precision arithmetics. Not only FPGAs, also other compute devices can be used at higher performance when using lower precision. This is likely to remain true in the near future as deep learning applications, which have become a major driver for compute device technologies, do not require high precision.[10]

[10] One such example is the introduction of so-called “tensor cores” in NVIDIA’s Pascal architecture, which are specialised on multiplying two 4x4 matrices using half-precision followed by an accumulation in single- or half-precision.

10 Conclusions

The PRACE PCP has been the first PCP that was executed in the area of HPC. It did address one of the most pressing challenges towards future exascale-level HPC, namely energy efficiency. Despite a relatively modest budget, with regard to R&D cost of activity in our field, we could demonstrate that promoting R&D by commercial operators by means of a PCP facilitates results that have impact on the roadmaps of the involved companies. New solutions have been developed and are being further enhanced for turning into products.

In the case of Maxeler the main achievement was not the development of a new technology, but the proof that it is possible to use a potentially extremely energy-efficient technology for real-life supercomputing applications. This could also allow HPC data centres to benefit from energy efficiency development performed for the wider customer basis of Maxeler, such as the one performed for services based on Amazon EC2 F1 Instance.

In the case of Atos-Bull the water cooled power supplies developed within the project will be introduce to their catalogue, with first customer shipments planned in H1 2018, and the software tools developed during phase III (BEO, BDPO, HDEEVIZ, SLURM Energy saving plugins) will be part of their Supercomputer Suite 5 Release 2 (SCS5 R2) that will be release in Q1 2018. PRACE partners and CoEs involved within this project already had already early access to it (through the CINES PCP-EoCoE workshop and 4IP-WP7 code enabling activity for instance).

In the case of E4, the DAVIDE development, besides being a pilot system, is the first of his kind in E4 products list, and is now available to other customers. This could be considered as a good example of R&D translation into product, which is one of the main goal of a Pre-Commercial Procurement.

Two out of three contractors in the final phase of this PCP are SMEs. By enabling them to deploy their technology in Tier-0 supercomputing centres creates significant visibility and allows these SMEs to grow. In the context of this PCP, E4 could for the first time realise a system that made it to the Top500 list [5]. A very positive effect of the PCP process to SMEs is that it allows them to have the phase of design funded by the project, which is something crucial for SMEs. In a “regular” R&D procurement the full cost of the design of the solution must be handled by the company, before competing without a guarantee of being selected and get a chance to cover this cost. This could be a high risk for a SME, avoided thanks to the PCP process. Another benefit that came directly from the multiple phase process is the possibility to modify the design from one phase to the next one, and even do withdraw without cost (which happen to one of our vendors between phase I and phase II). This limits the risk for the vendor of a binding commitment to provide a certain solution or a certain amount of performance that could become inaccessible along the line for exterior reasons, such as unforeseeable change in the roadmap of provider. Mitigating this contractual risk could allow vendors to take a lot more risk in their design, thus allowing more innovations. This possibility has been used, at different level, by all three vendors as it is reflected in the technology change between Phase II and Phase III reported within thin Deliverable.

Beside of that, a positive side effect of the PCP process in our field, is also the fact that both the Procurer Group and the vendors have gained a better view on the Total Cost of Ownership (TCO), buy being able to better understand the energy consumption of our systems. The methods developed within this PCP could be re-used in other regular procurement to get the better value for money. Finally, the PCP still remains a relatively new instrument. Various lessons have been learned in the context of the PRACE PCP as described in this report. They will allow to make even more successful use of this instrument in future PCPs.