E-Infrastructures
H2020-EINFRA-2014-2015

EINFRA-4-2014: Pan-European High Performance Computing Infrastructure and Services

PRACE-4IP

PRACE Fourth Implementation Phase Project

Grant Agreement Number: EINFRA-653838

D6.7
Final report on the HPC Eco-system prototypes operation

Final

Version: 1.2
Author(s): Fabio Affinito, CINECA
Date: 09.01.2018

Project and Deliverable Information Sheet

PRACE Project	**Project Ref. №: EINFRA-653838**	
	Project Title: Final report on the HPC Eco-system prototypes operation	
	Project Web Site: http://www.prace-project.eu	
	Deliverable ID: < **D6.7**>	
	Deliverable Nature: < Report >	
	Dissemination Level: PU *	**Contractual Date of Delivery:** 31 / December / 2017
		Actual Date of Delivery: 15 / January / 2018
	EC Project Officer: Leonardo Flores Añover	

* - The dissemination level are indicated as follows: **PU** – Public, **CO** – Confidential, only for members of the consortium (including the Commission Services) **CL** – Classified, as referred to in Commission Decision 2991/844/EC.

Document Control Sheet

Document	**Title: Final report on the HPC Eco-system prototypes operation**	
	ID: D6.7	
	Version: <1.2>	**Status:** ***Final***
	Available at: http://www.prace-project.eu	
	Software Tool: Microsoft Word 2010	
	File(s): D6.7.docx	
Authorship	**Written by:**	Fabio Affinito, CINECA
	Contributors:	Olivier Rouchon, CINES Alexander Strube, JSC
	Reviewed by:	Brian Vinter, UCPH Florian Berberich, JUELICH
	Approved by:	MB/TB

Document Status Sheet

Version	Date	Status	Comments
0.1	14/November/2017	Draft	Initial draft
0.2	29/November/2017	Draft	Added first content on E4 prototype
0.3	12/December/2017	Draft	Added content about Maxeler and Atos-Bull prototype
1.0	13/December/2017	Draft	General revision
1.1	8/January/2018	Draft	Version including the reviewers comments
1.2	9/January/2018	Final	Final version for MB/TB approval

Document Keywords

Keywords:	PRACE, HPC, Research Infrastructure, HPC Eco-system, operation, PCP, prototypes

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

aisbl	Association International Sans But Lucratif (legal form of the PRACE-RI)
BCO	Benchmark Code Owner
CoE	Center of Excellence
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture (NVIDIA)
DARPA	Defense Advanced Research Projects Agency
DEISA	Distributed European Infrastructure for Supercomputing Applications EU project by leading national HPC centres
DoA	Description of Action (formerly known as DoW)
EC	European Commission
EESI	European Exascale Software Initiative
EoI	Expression of Interest
ESFRI	European Strategy Forum on Research Infrastructures
GiB	Giga (= 2^{30} ~ 10^9) Bytes (= 8 bits), also GByte or GB
Gb/s	Giga (= 10^9) bits per second, also Gbit/s
GB/s	Giga (= 10^9) Bytes (= 8 bits) per second, also GByte/s
GÉANT	Collaboration between National Research and Education Networks to build a multi-gigabit pan-European network. The current EC-funded project as of 2015 is GN4.
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
GPU	Graphic Processing Unit
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
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HPL	High Performance LINPACK
ISC	International Supercomputing Conference; European equivalent to the US based SCxx conference. Held annually in Germany.
KB	Kilo (= 2^{10} ~10^3) Bytes (= 8 bits), also KByte
LINPACK	Software library for Linear Algebra
MB	Management Board (highest decision making body of the project)
MiB	Mega (= 2^{20} ~ 10^6) Bytes (= 8 bits), also Mbyte or MB
MB/s	Mega (= 10^6) Bytes (= 8 bits) per second, also MByte/s
MFlop/s	Mega (= 10^6) Floating point operations (usually in 64-bit, i.e. DP) per second, also MF/s
MooC	Massively open online Course
MoU	Memorandum of Understanding.
MPI	Message Passing Interface

NDA	Non-Disclosure Agreement. Typically signed between vendors and customers working together on products prior to their general availability or announcement.
PA	Preparatory Access (to PRACE resources)
PATC	PRACE Advanced Training Centres
PCP	Pre-Commercial Procurement
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PRACE 2	The upcoming next phase of the PRACE Research Infrastructure following the initial five year period.
PRIDE	Project Information and Dissemination Event
RI	Research Infrastructure
TB	Technical Board (group of Work Package leaders)
TB	Tera (= 240 ~ 1012) Bytes (= 8 bits), also TByte
TCO	Total Cost of Ownership. Includes recurring costs (e.g. personnel, power, cooling, maintenance) in addition to the purchase cost.
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
UNICORE	Uniform Interface to Computing Resources. Grid software for seamless access to distributed resources.

List of Project Partner Acronyms

BADW-LRZ	Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, Germany (3rd Party to GCS)
BILKENT	Bilkent University, Turkey (3rd Party to UYBHM)
BSC	Barcelona Supercomputing Center - Centro Nacional de Supercomputacion, Spain
CaSToRC	Computation-based Science and Technology Research Center, Cyprus
CCSAS	Computing Centre of the Slovak Academy of Sciences, Slovakia
CEA	Commissariat à l'Energie Atomique et aux Energies Alternatives, France (3 rd Party to GENCI)
CESGA	Fundacion Publica Gallega Centro Tecnológico de Supercomputación de Galicia, Spain, (3rd Party to BSC)
CINECA	CINECA Consorzio Interuniversitario, Italy
CINES	Centre Informatique National de l'Enseignement Supérieur, France (3 rd Party to GENCI)
CNRS	Centre National de la Recherche Scientifique, France (3 rd Party to GENCI)
CSC	CSC Scientific Computing Ltd., Finland
CSIC	Spanish Council for Scientific Research (3rd Party to BSC)
CYFRONET	Academic Computing Centre CYFRONET AGH, Poland (3rd party to PNSC)
EPCC	EPCC at The University of Edinburgh, UK
ETHZurich (CSCS)	Eidgenössische Technische Hochschule Zürich – CSCS, Switzerland

FIS	FACULTY OF INFORMATION STUDIES, Slovenia (3rd Party to ULFME)
GCS	Gauss Centre for Supercomputing e.V.
GENCI	Grand Equipement National de Calcul Intensiv, France
GRNET	Greek Research and Technology Network, Greece
INRIA	Institut National de Recherche en Informatique et Automatique, France (3 rd Party to GENCI)
IST	Instituto Superior Técnico, Portugal (3rd Party to UC-LCA)
IUCC	INTER UNIVERSITY COMPUTATION CENTRE, Israel
JKU	Institut fuer Graphische und Parallele Datenverarbeitung der Johannes Kepler Universitaet Linz, Austria
JUELICH	Forschungszentrum Juelich GmbH, Germany
KTH	Royal Institute of Technology, Sweden (3 rd Party to SNIC)
LiU	Linkoping University, Sweden (3 rd Party to SNIC)
NCSA	NATIONAL CENTRE FOR SUPERCOMPUTING APPLICATIONS, Bulgaria
NIIF	National Information Infrastructure Development Institute, Hungary
NTNU	The Norwegian University of Science and Technology, Norway (3rd Party to SIGMA)
NUI-Galway	National University of Ireland Galway, Ireland
PRACE	Partnership for Advanced Computing in Europe aisbl, Belgium
PSNC	Poznan Supercomputing and Networking Center, Poland
RISCSW	RISC Software GmbH
RZG	Max Planck Gesellschaft zur Förderung der Wissenschaften e.V., Germany (3 rd Party to GCS)
SIGMA2	UNINETT Sigma2 AS, Norway
SNIC	Swedish National Infrastructure for Computing (within the Swedish Science Council), Sweden
STFC	Science and Technology Facilities Council, UK (3rd Party to EPSRC)
SURFsara	Dutch national high-performance computing and e-Science support center, part of the SURF cooperative, Netherlands
UC-LCA	Universidade de Coimbra, Labotatório de Computação Avançada, Portugal
UCPH	Københavns Universitet, Denmark
UHEM	Istanbul Technical University, Ayazaga Campus, Turkey
UiO	University of Oslo, Norway (3rd Party to SIGMA)
ULFME	UNIVERZA V LJUBLJANI, Slovenia
UmU	Umea University, Sweden (3 rd Party to SNIC)
UnivEvora	Universidade de Évora, Portugal (3rd Party to UC-LCA)
UPC	Universitat Politècnica de Catalunya, Spain (3rd Party to BSC)
UPM/CeSViMa	Madrid Supercomputing and Visualization Center, Spain (3rd Party to BSC)
USTUTT-HLRS	Universitaet Stuttgart – HLRS, Germany (3rd Party to GCS)
VSB-TUO	VYSOKA SKOLA BANSKA - TECHNICKA UNIVERZITA OSTRAVA, Czech Republic
WCNS	Politechnika Wroclawska, Poland (3rd party to PNSC)

Executive Summary

This deliverable reports about the installation and the operation of the three prototypes delivered at the end of the PRACE Pre-Commercial Procurement (PCP). In the plans of the PRACE-4IP projects, the release of such prototypes was expected during the summer 2017. Delays in the procedure reduced the testing to the period from September to the end of November 2017. Due to this delay, the only feasible activity in terms of operation was to support to the installation on the premises and, more important, to make the prototypes available to the users. All the available time, thus, was spent in order to make it possible for PRACE users to access the resources as soon as possible, and to facilitate the operation of the systems. In this deliverable we explain how the systems were deployed on the chosen premises and how they were made accessible to the users. Then, we will briefly describe the plans for the future of the prototypes and their utilisation.

1 Introduction

This deliverable describes the work performed on the operation for the three pilot systems (prototypes) delivered in the Phase III of the PRACE Pre-Commercial Procurement. The target of the PCP was to perform R&D on a “Whole system design for energy efficient HPC”. ”. This deliverable describes exclusively the work made on the PCP prototypes during the extension of the PRACE-4IP project. In particular, this work consisted in the support to the installation and deployment of the hardware provided at the end of the PCP. The PCP lasted for 4 years, at the end of which 3 contractors (i.e. ATOS-Bull, E4 and Maxeler) were selected with three different design architectures. At the end of the PCP the three contractors delivered three pilot systems that were installed respectively:

- ATOS-Bull system in CINES (Montpellier, France)
- E4 system in CINECA (Bologna, Italy)
- Maxeler system in JSC (Julich, Germany)

According to the original plans, the installation of the three prototypes was scheduled at the beginning of the summer of 2017 (June-July). Due to several delays, the installation of all the three system was shifted to the following months, between September (ATOS-Bull) and November (E4 and Maxeler). As a consequence of this delay, the only feasible activity of a operational nature, within the time-frame of the PRACE-4IP project, was mainly to support the installation and the grant of access to the PRACE users. It should be also considered that, according to the PCP contract, the property of the systems, and hence their management, belonged to the three contractors until the date of the 30.11.2017, when the final documents of the PCP were delivered. For this reason, for a long time, any kind of operation activity was necessarily performed jointly between the ATOS-Bull/E4/Maxeler teams and the PRACE compute-center personnel.

In this document we describe the features of the three prototypes, the support to the installation on the three premises, performed in collaboration with the staff of the system manufacturers, and, finally, how we granted access to the users. We also briefly describe the plans for the next future of the three prototypes, their integration and their exploitation by the users’ community.

2 Description of the prototypes

In this section we describe three prototypes and the support to the operations provided by the PRACE project. In particular, we describe the installation phase, the availability to the users and the plans for the exploitation in the next future.

2.1 Atos-Bull prototype (CINES)

The ATOS/Bull bid, as part of the third phase of PRACE-3IP PCP, was to provide a Sequana based Pilot System to be hosted in the CINES datacenter in Montpellier (France), in order to port and optimize PRACE applications, develop prototypes of Energy efficiency oriented tools, integrate all prototypes in a "production-like" system, and eventually validate these tools with PRACE applications.

Figure 1: The Sequana cabinet

2.1.1 Description of the pilot system

This PCP-pilot system is made of one rack equipped with 56 Bull Sequana X1210 Intel® Xeon™ "Knights Landing" blades. Each blade includes three compute nodes:

- three identical mother boards (Intel® Groveport platform), one per Intel® Xeon-Phi™ node
- Identical mezzanine boards for IO connection (InfiniBand EDR), one per Intel Xeon-Phi node;
- one shared cold plate to cool all components in the blade (including the I/O);
- Storage is provided by HDD/SDD located in the ISMA server and shared through Ethernet integrated network.

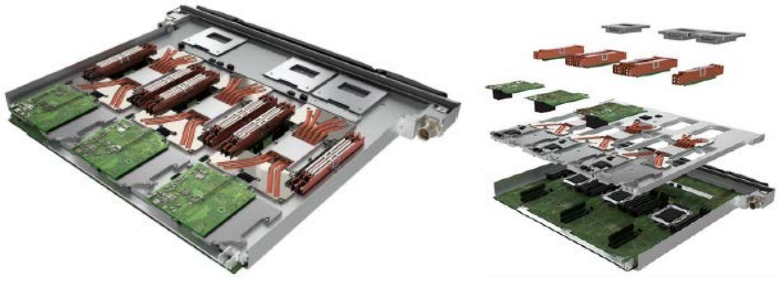

Figure 2: The Sequana blades

In total, the cluster provides a total of 168 compute nodes.

- one Intel® Xeon-Phi™ Knights Landing 7250 16GB GDDR5 215W Passive;
- six 16GB@2400MT/s DDR4 DIMMs;
- one onboard Integrated Management Controller (IBMC – Emulex Pilot3 SoC);
- thirty-two PCIe lanes routed to the mezzanine PCIe connectors to implement one or two interconnect channels, each supporting 100Gb/s, to be connected to L1 interconnect switch inside Bull Sequana X1000 switch cabinet. Only one EDR channel is used in the proposed solution;
- one FRU EEPROM for traceability;
- each node hosts an HDEEM FPGA to provide accurate and high frequency power measurement. This component is key for the PCP project.

In addition, new hardware technologies (i.e. water-cooled PSU prototypes) have been installed in the rack to have the first 100% water-cooled system in production.

2.1.2 Support to the installation on the premises

The installation of the prototype started in April when the hardware was delivered in the CINES premises. The main activities for CINES staff were to integrate the new environment in the existing infrastructure:

- Cooling and power (acquisition and deployment of sensors and other hardware for power consumption external measurement, development of tools to agregate data and generate statistics),
- Common services (network, email, DNS, LDAP, ticketing system, etc.)
- Security (firewall setup, IP filtering, etc.).

The deployment was completed by September, and the system was made available to project users for testing.

Over this period, the CINES Sysadmin and Support teams have received several requests for assistance:

- 14 tickets for application support : compilation issues, MPI errors, modules availability, etc.;
- 37 tickets for system administration: connection problems, workload manager configuration, nodes failures, energy measurement issues, etc.

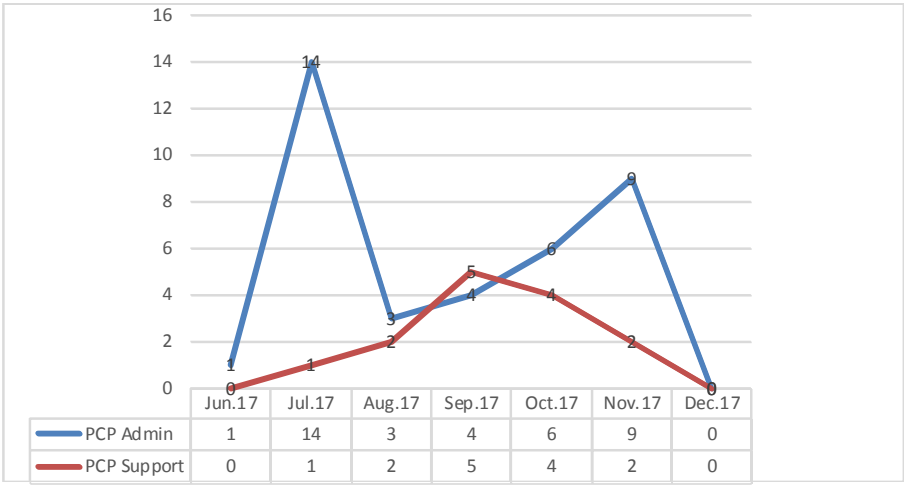

Figure 3: Number of tickets opened to the application and system support

The average time for the first response following the trouble ticket creation was less than twelve hours. 90% of the issues raised over the period have now been fixed and closed. Note that there is no specific Service Level Agreement in place, as the cluster is not yet considered as in production.

2.1.3 Availability to the users

The first months of operations were dedicated tot he ATOS-Bull R&D teams to install software and tools (Bull Energy Optimizer – BEO, Bull Dynamic Power Optimizer – BDPO, HDEEViz - power consumption visualization tool and an energy oriented plugin for SLURM based on adaptive scheduling), and validate the stability and readiness of the environement.

At the beginning of October, a workshop for the EoCoE community (http://www.eocoe.eu/events/scientific-applications-towards-exascale) was organized at CINES and with the pilot system. During the hands-on session, BEO, HDEEViz and SLURM energy plugin were presented and evaluated by a group of selected user. The feedback was rather positive, as the proposed tools were considered as useful to understand the power consumption behavior of their own applications.

The cluster is now available for PCP users, and there are currently 48 logins registered on the system, split in three different categories:

ATOS-Bull	CINES	PCP-Others
8	7	33

The first benchmarks (ALYA, Code_Saturne, CP2K, GADGET, GPAW, GROMACS, NAMD, NEMO, PFARM, QCD, Quantum Espresso, Specfem3D_Globe) have been executed as part of the WP7 PRACE-4IP extension program – accelerated benchmark suite - in November. Results will be integrated to the D7.7 deliverable (performance and energy metrics on PCP systems).

2.1.3 Future plans

Access to the prototype should be continued over 2018 through the Preparatory Access process. This will be confirmed by the PRACE PCP GOP (Group of Procurers)

2.2 E4 prototype (CINECA)

The prototype engineered by E4 is codenamed as D.A.V.I.D.E. (i.e. Development of an Added-Value Infrastructure Designed in Europe). It is based an OpenPower architecture design with compute nodes deriving from the IBM Minsky platforms. Differently from the original Minsky server, which is air cooled, in D.A.V.I.D.E. a better energy efficiency is reached by liquid cooling for CPUs, GPUs and memory.

2.2.1 Description of the pilot system

D.A.V.I.D.E. is composed by 45 nodes connected with an Infiniband EDR 100 Gb/s networking, with a peak performance of 990 TFlops and an estimated power consumption of less than 2kW per node. Each node is a 2 OU 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

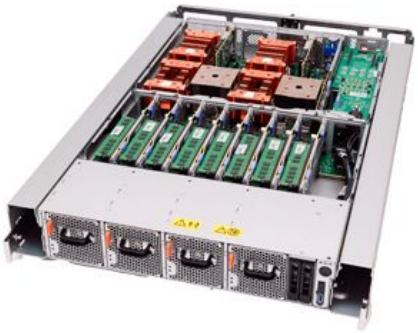

Figure 4: Schematic summary of the features of the D.A.V.I.D.E. prototype

The compute node is:

- Derived from the IBM® POWER8 System S822LC (codename Minsky).
- 2 OU 21” Open Rack Enclosure with integrated piping & power distribution.
- Power8-based node in OCP form-factor, with leading edge features specifically engineered for HPC workloads.
- Two IBM POWER8 with NVlink and four NVIDIA Tesla P100 HSXM2.
- 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.

The cooling is:

- 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.

The system is accelerated with a GPU system:

- The system is coupled with four NVIDIA Tesla P100 HSMX2 per node with NVLINK interconnect, to deliver performance for the most demanding compute applications, 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

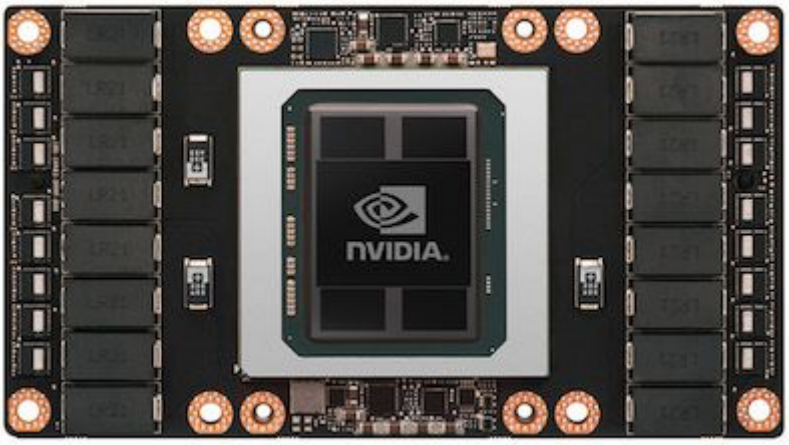

Figure 5: he NVIDIA Tesla P100 GPU integrated in the E4 prototype

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.

2.2.2 Support to the installation on the premises

The E4 Pilot System, named D.A.V.I.D.E. was assembled in the E4 facilities in Reggio Emilia. Thus, once the system was preliminarly tested, it was moved on the CINECA premises in the timeframe September-October 2017.

The operation support was provided by CINECA specialists who took contact with the E4 team since February 2017. During this timeframe, the movement and installation of the D.A.V.I.D.E. racks was accurately planned. During the planning phase, several joint meetings between the two teams and site inspections took place.

Together with the CINECA technical team, we agreed on the disposal of the 4 racks (1 for the Mellanox switches and login nodes, 3 for the compute nodes) on the datacentre. Also, the technical team had to perform an analysis of the power supply to the Pilot System. The cooling system of the pilot system took advantage of the already present system for the MARCONI (rear-cooled).

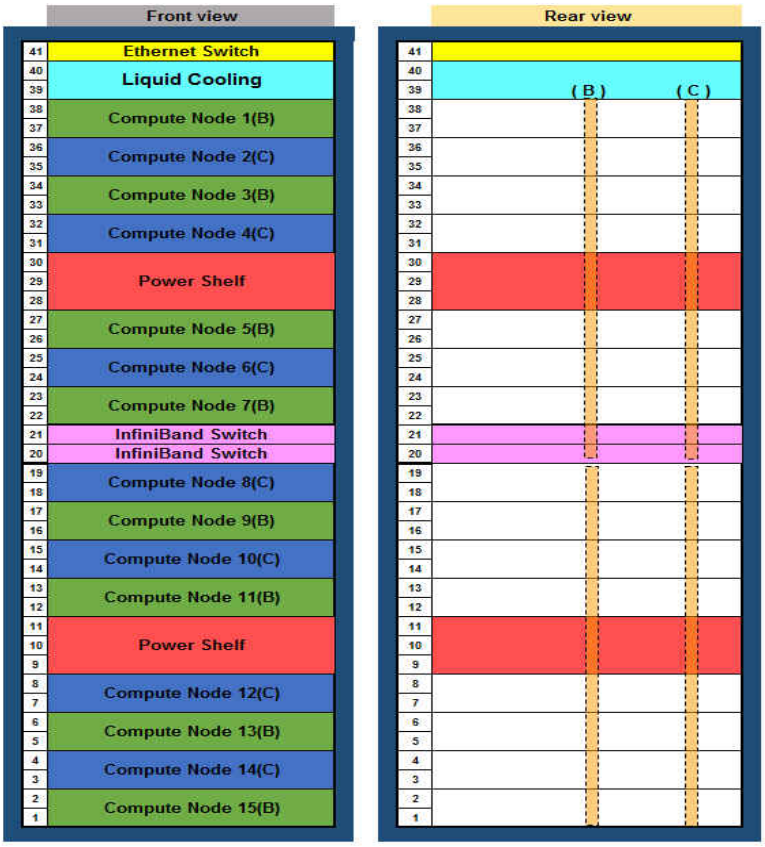

Figure 6: Layout of a D.A.V.I.D.E. rack in the final configuration.

Since the PCP contract had not specifications about the provision of a storage infrastructure, the whole planning about the choice of the storage was in charge to the CINECA sysadmins. The final choice was to provide a NFS with a capacity of 100TB relying on the pre-existing CINECA facilities. On the near future, there is a plan to exploit a new storage system with BeeGFS technology.

In order to make a quick availability of the system to the users, two login nodes were made available through a public IP address.

2.2.3 Availability to the users

One of the most urgent activities, once the system was deployed in the CINECA facilities, was to make it available to the users. According to the schedules, the system was managed by the E4 team until the release of the deliverables of the Phase III of the PCP, i.e. 30/11/2017. However, CINECA and E4 staff agreed to open the system to the external users prior to this deadline. This permitted to test the system and, more important, to make the system accessible to the PRACE team.

Since, during this period of time, the system was not under any kind or shared directory access (i.e. LDAP), the creation of the accounts on this machine was synchronized between the staff of CINECA and E4.

Affiliation	No. of accounts
CINECA	11
E4	5
PRACE	12
Unibo	3
Others	14
Total	**45**

Table 1: Accounts created on D.A.V.I.D.E. during the pre-production phase, distributed by affiliation

At the end of November 2017, we count 45 active users on D.A.V.I.D.E. of. Most of them are PRACE users (12) and CINECA staff (11). Other accounts were granted to the E4 staff, University of Bologna staff, and experts affiliated to other institutions.

A module environment (v.3.2.10) is available at the time of the extension of this deliverable. The present structure of the modules is planned to change soon, with the management of the system moving from E4 to CINECA. The present version of the module environment only offers compilers (XL and Gnu) and production libraries (OpenMPI, Cuda, ESSL). This will increase with the porting of applications when time permits.

2.2.4 Future plans

Similarly to the other prototypes, this system will also be available to the PRACE users' through a Preparatory Phase approach. This will be formalized and delivered in the PRACE boards in the first months of 2018. This system will also serve as a working platform for porting the most interesting applications for the users' community in a Power+GPU architecture. This part of activity has been already started by the PRACE-4IP WP7. All the tools available on D.A.V.I.D.E. will also permit evaluation of the energy consumption of the submitted jobs, opening the perspectives on an energy awareness of the software engineering of the applications.

2.3 Maxeler prototype (JSC)

The Maxeler Pilot System hardware has 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.

2.1.3 Description of the prototype

The Maxeler Pilot System comprises 3 different nodes:

- **MPC Node**: A Maximum Performance Computing (MPC) server with MAX5 Data Flow Engine (DFE) cards
- **CPU Node**: 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**: A node serving as a head node that acts as a gateway to the Pilot System and provides a few management services.

Each component is described in more detail in the following sub-sections.

2.3.1.1 MPC node

Figure 7: A Maxeler M5C card

The MPC node is a server with 8 DFE cards. A MAX5C card as shown in Figure 7 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 propietary network called MaxRing. The node features a dual-port Infiniband FDR card to connect it to the CPU Node.

2.3.1.2 CPU node

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

2.3.1.3 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

2.3.1.4 Software components

The following development 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.

2.3.2 Support to the installation on the premises

The hardware was delivered on 20.10.2017. In the context of the installation of the pilot system JSC staff was involved in the following tasks:

- Physical installation of hardware in racks provided by JSC
- Connecting hardware electrically and power-on
- Connecting pilot system to network and network configuration
- Integration into JSC central user management system
- Setup wiki for hosting documentation (https://trac.version.fz-juelich.de/reconfigurable)

2.3.4 Availability to the users

Shortly after hardware availability the system was made accessible to users through a manual procedure. The users had to send their SSH keys and were added to a local NIS-based user database.

Meanwhile the Pilot System has been integrated with JSC's central user management system. New users can apply through JSC's central page for requesting accounts. The account application procedure includes a step where the applicant has to sign a usage agreement. Once the application has been approved, the account will be published internally through an LDAP service and a home directory is created on the Head Node. Once this step is completed, users can connect to the Head Node using an ssh key provided during the application procedure.

2.3.5 Future plans

The (relatively small) Maxeler Pilot System is planned to be integrated into a complex for reconfigurable computing and data analytics as shown in Figure 8. The other servers are already available at JSC and will provide the users with additional capabilities. In particular, a powerful server will be made available for building FPGA bitstreams, which is potentially a very CPU-intensive task that benefits from a built server with a significant amount of main memory.

During the next months it is planned to collaborate with Maxeler on providing training on the Pilot System with the goal of attracting more users within PRACE and beyond to explore the capabilities of the technology. This should lead to a broader range of applications, which can run on the Pilot System.

It is furthermore planned to continue together with Maxeler efforts on enhancing the application porting efforts, which have so far been executed within the PCP. The goal is to facilitate interactions with relevant user communities, e.g. the Lattice Quantum Chromodynamics (LQCD) community, and to create libraries that can exploit the DFEs, which can be used by

scientists without mandating in-depth knowledge of the technology. Similar efforts for hiding hardware complexity are, e.g., ongoing in the context of graphics processing units (GPU).[1]

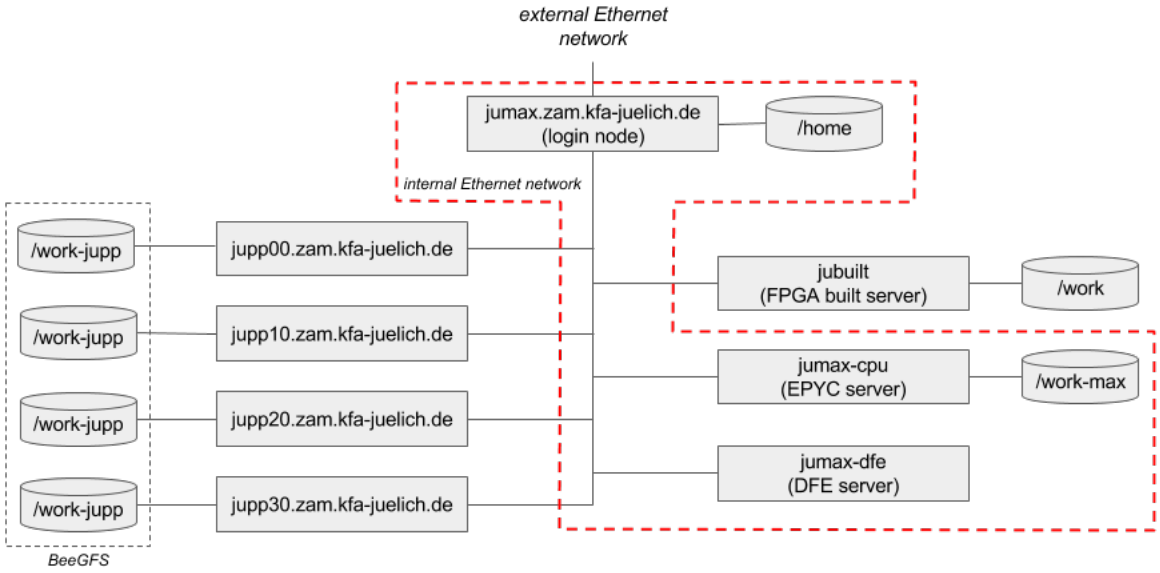

Figure 8: Overview on the planned testbed for reconfigurable and data-intensive computing with the integrated Maxeler Pilot System. The components of the latter are shown by the dashed line.

3 Conclusion

In this deliverable we showed the operation performed on the three prototypes delivered at the end of the PRACE Pre-Commercial Procurement. The work presented in this deliverable concerns the activities performed during the extension of the PRACE 4IP project exclusively on the PCP prototypes. Due to the delays in the deployment of such prototypes, we were able to perform only a part of the planned activities before the end of the project. In particular, we choose to focus on activities aimed to permit to open the systems to the production as soon as possible. Also, according to the legal constraint of the PCP, until the end of November 2017, the administration of the three prototypes was up to the technical teams of the three contractos. As a consequence, the possibility to operate on these systems was quite limited.

In this deliverable, we discussed the installation on the premises (CINES, CINECA and JSC) and the pre-production phase in which the accesses to PRACE users were granted.. We finally discussed the perspectives on the prototypes in the next months for a further integration and users' exploitation of the three pilot systems.

[1] See: https://lattice.github.io/quda/