REliable Power and time-ConstraInts-aware Predictive management of heterogeneous Exascale systems

WP4 Architecture Level and Middleware Support

D4.2 Prototype Deployments

http://www.recipe-project.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 801137

Grant Agreement No.: 801137
Deliverable: D4.2 Prototype Deployments

Project Start Date: 01/05/2018 **Duration**: 36 months
Coordinator: *Politecnico di Milano, Italy*

Deliverable No:	D4.2
WP No:	4
WP Leader:	UPV
Due date:	30/04/2019
Delivery date:	30/04/2019

Dissemination Level:

PU	Public Use	X
PP	Restricted to other programme 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 SUMMARY INFORMATION

Project title:	**REliable Power and time-ConstraInts-aware Predictive management of heterogeneous Exascale systems**
Short project name:	RECIPE
Project No:	801137
Call Identifier:	H2020-FETHPC-2017
Thematic Priority:	Future and Emerging Technologies
Type of Action:	Research and Innovation Action
Start date of the project:	01/05/2018
Duration of the project:	36 months
Project website:	http://www.recipe-project.eu

D4.2 Prototype Deployments

Work Package:	WP4 Architecture Level and Middleware Support
Deliverable number:	D4.2
Deliverable title:	Prototype Deployments
Due date:	30/04/2019
Actual submission date:	30/04/2019
Editor:	R. Tornero
Authors:	R. Tornero, C. Hernandez, A. Cilardo, F. Parricelli, G. Massari
Dissemination Level:	PU
No. pages:	24
Authorized (date):	30/04/2019
Responsible person:	W. Fornaciari
Status:	Final

Revision history:

Version	**Date**	**Author**	**Comment**
v.0.1	01/04/2019	R. Tornero	Create skeleton
v.0.2	04/04/2019	R. Tornero	Add first outline
v.0.3	10/04/2019	R. Tornero	Add partner contributions
v.0.4	15/04/2019	R. Tornero	Add first complete version draft
v.1.0	19/04/2019	C. Hernandez	Release version for internal review
v.1.1	25/04/2019	R. Tornero	Add internal reviewer's comments

Quality Control:

	Who	Date
Checked by internal reviewer	BSC	24/04/2018
Checked by WP Leader	UPV	25/04/2018
Checked by Project Technical Manager	G. Agosta	30/04/2019
Checked by Project Coordinator	W. Fornaciari	30/04/2019

COPYRIGHT

ACKNOWLEDGEMENTS

RECIPE is a project that has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 801137. Please see http://www.recipe-project.eu for more information.

The partners in the project are Politecnico di Milano (POLIMI), Universitat Politècnica de València (UPV), Centro Regionale Information Communication Technology scrl (CeRICT), École Polytechnique Fédèrale de Lausanne (EPFL), Barcelona Supercomputing Center (BSC), Poznan Supercomputing and Networking Center (PSNC), IBT Solutions S.r.l. (IBTS), Centre Hospitalier Universitaire Vaudois (CHUV). The content of this document is the result of extensive discussions within the RECIPE ©Consortium as a whole.

DISCLAIMER

The content of the publication herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services. The information contained in this document is provided by the copyright holders ”as is” and any express or implied warranties, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose are disclaimed. In no event shall the members of the RECIPE collaboration, including the copyright holders, or the European Commission be liable for any direct, indirect, incidental, special, exemplary, or consequential damages (including, but not limited to, procurement of substitute goods or services; loss of use, data, or profits; or business interruption) however caused and on any theory of liability, whether in contract, strict liability, or tort (including negligence or otherwise) arising in any way out of the use of the information contained in this document, even if advised of the possibility of such damage.

Contents

1 Introduction 10

2 Hardware and Software Components 11

2.1 Hardware Components 11

2.2 Software Components 20

3 RECIPE Prototype 20

4 Conclusions 22

Executive Summary

This report documents the prototype implemented for RECIPE. In particular it describes and justifies from a technical point of view all the hardware and software components the prototype is composed of. Also, it presents the current view of the prototype at month 10 (M10), with a concrete hardware and software materialization, and how it will evolve across the different scheduled project phases: from M10 to month 18 (M18) and from M18 to moth 36 (M36). From M10 to M18 the prototype will keep split at UPV and CERICT partners premises. They will maintain the prototype and provide concurrent system access to the consortium members. During this period the project partners will refine the prototype strategy for the third phase (M18-M36) and opt for a unified or split system, depending on a cost/benefit analysis and priority of the technical objectives. In case the consortium decide to build a single prototype, the two parts will be put together at one of the facilities of a project partner and they will be interconnected by Infiniband® technology.

Acronyms

AI Artificial Intelligence.

AMAP As Much As Possible.

API Application Programming Interface.

CoWoS Chip-on-Wafer-on-Substrate.

CPU Central Processing Unit.

CUDA Compute Unified Device Architecture.

DDR Double Data Rate.

DIMM Dual Inline Memory Module.

EDR Enhanced Data Rate.

FHFL Full-Height Full-Length.

FLOPs Floating Point Operations per Second.

FPGA Field Programmable Gate Array.

GB GigaByte.

Gb GigaBit.

GbE Gigabit Ethernet.

GBps GigaByte per Second.

Gbps GigaBit per Second.

GPU Graphics Processing Unit.

GTps GigaTransfer per Second.

HBM High Bandwidth Memory.

HLS High Level Synthesis.

HPC High Performance Computing.

HT Hyper Threading.

I/O Input/Output.

IBTA InfiniBand Trade Association.

LP Low Profile.

LTS Long-Term Support.

MMI64 64-bit Multi-Message-Interface.

NIC Network Interface Card.

NVMe Non-Volatile Memory Express.

NVRAM Non-Volatile Random-Access Memory.

OPA Omni Path Architecture.

PCIe Peripheral Component Interconnect Express.

QDR Quad Data Rate.

QoS Quality-of-Service.

QSFP Quad Small Form-factor Pluggable.

RDMA Remote Direct Memory Access.

SATA Serial Advanced Technology Attachment.

SDK Software Development Kit.

SDRAM Synchronous Dynamic Random-Access Memory.

SL Service Lane.

SME Small Medium Enterprise.

SRAM Static Random-Access Memory.

SSD Solid State Device.

TB TeraByte.

TBD To Be Discussed.

Tbps TeraBit per Second.

TDP Thermal Design Power.

UPI Ultra Path Interconnect.

UPS Uninterruptible Power Supply.

VL Virtual Lane.

W Watt.

1 Introduction

Exascale computing is expected to cover an increased range of application classes. In addition to traditional massively parallel ”number crunching” applications, new classes are emerging such as real-time High Performance Computing (HPC) and data-intensive scalable computing. Furthermore, Exascale computing is characterized by a ”democratisation” of HPC: to fully exploit the capabilities of Exascale-level facilities, HPC is moving towards enabling access to its resources to a wider range of new players, including Small Medium Enterprises (SMEs), through cloud-based approaches. Finally, the need for much higher energy efficiency is pushing towards deep heterogeneity which is going to widen the range of options for acceleration, moving from the traditional Central Processing Unit (CPU) only organization, to the Graphics Processing Unit (GPU) which currently dominates the Green500, to more complex options including different programmable accelerators and even (reconfigurable) hardware accelerators.

In this light, the current approach towards resource management, which is essentially limited to assigning to each application a set of physical nodes (both its cores and memories) at the job scheduling level, is not sufficient. First, it is going to waste resources, as each application may only employ a given set of accelerators and GPU cores. Thus, multiple applications, which might coexist on the same nodes, are actually allocated to different nodes, leading to increased costs for the end user and reduced utilization of all types of resources (cores, bandwidth, logical and physical memory, etc.). Second, with increasing amounts of computational resources, hardware failures are going to increase in frequency. Currently, predictable performance, which is needed for real-time HPC applications, is provided by assigning resources in an exclusive mode to the application, which limits the efficiency of the systems. At the same time, in the context of reliability, checkpointing techniques are useful, of course, but they only address the needs of non-real-time applications.

In the RECIPE project, we address these concerns by applying runtime resource management techniques leveraging resource virtualization and disaggregation to provide the ability to partition and allocate resources at a finer grain than usually provided by job schedulers – even partitioning accelerators – and predictive reliability leveraging hardware monitors to automatically assess the failure risk level of resources (against both short term/transient faults and long term faults due to thermal stress and aging), and assign them according to the resilience needs of the applications and to improve the predictability of the performance.

To demonstrate the validity of these core techniques, we employ a set of real-world applications, including weather forecasting, subsoil properties identification, and bio-medical big-data applications. We perform our demonstration on both industry-grade, pre-Exascale systems and on emerging deeply heterogeneous technologies. Thus, the RECIPE methodology is based on the development of a prototype infrastructure where heterogeneity is the main driver for the component selection process. For this end, *Task 4.2: Integration of the RECIPE prototype platform* (Task Leader: UPV; Participants: CERICT, POLIMI) addressed the integration of different state-of-the-art hardware resources to build a scaled deeply heterogeneous system, in which the different nodes are interconnected in a disaggregated way based on a high-performance fabric, by means of a number of interconnect switches. In particular, the prototype was planned to include general purpose resources like commodity processors targeted for the HPC market and manycore hardware (e.g. Intel® Xeon® processor and Intel® Phi), while specific resources should consist of GPU, Field Programmable Gate Arrays (FPGAs), as well as next-generation advanced smart Network Interface Cards (NICs).

Within the first phase of the project, the RECIPE partners carried out a careful analysis of the current technological trends, which led to the identification of key opportunities at the architectural level. These basic choices involve state-of-the-art server-grade CPUs as well as high-end GPU and FPGA cards targeted at the datacenter/HPC market. Concerning manycore technologies, the consortium discarded discontinued products, like Intel® Phi, which was originally mentioned in the RECIPE proposal, and those targeting non-HPC segments, e.g. the Mellanox® Bluefield manycore system, which illustrates the concept of Smart NICs. On the other hand, as planned, the project consolidated the integration of the FPGA-based infrastructure provided by the MANGO H2020 FETHPC project [7], offering a large-scale *multi-FPGA* setting that perfectly fits the purposes of the RECIPE project. This setup allows RECIPE to fully demonstrate advanced fault-tolerance, reliability and Quality-of-Service (QoS) capabilities through resource management at each level of the infrastructure.

This deliverable describes the prototype implemented in RECIPE as the outcome of Task 4.2. First, Section 2 describes and justifies from a technical point of view the hardware and software components selected to build the RECIPE system. Next, Section 3 presents the prototype built in RECIPE. Finally, Section 4 provides the conclusion of this deliverable.

2 Hardware and Software Components

As a result of the requirement and gap analysis carried out by the consortium in *Task 1.1: Requirements and Gap Analysis*, a listing of the minimum technology and software specifications was defined, which is summarized in Tables 1 and 2. Table 1 focuses on the specifications required for the general purpose hardware acquired for the project, while Table 2 contains the use case requirements that the GPU and FPGA accelerators should fulfill as part of the prototype. In the latter table, IBTS is not present since its use case is highly coupled to the PSNC application, thus relying on PSNC acceleration solution. Notice that both listings are provided in a high level of abstraction in such a way that they can be understood easily by non-technology experts.

In this section we describe and justify from a technical point of view the hardware and software components selected for building the RECIPE prototype. We focus on the hardware components first and the software elements later.

2.1 Hardware Components

The selection of hardware components was done in three steps. First, we created a list of components that fit the technology requirements resulting from the output of Task1.1. Second, we took into account market prices for the different options satisfying the former premise and third, we took the final selection with the aim of targeting a deeply heterogeneous prototype given the resulting options of the second step.

Table 1: List of technology application requirements for general purpose hardware.

	BSC	PSNC	IBTS	CHUV
Compute intensive	x	x	x (2.6Ghz)	x
Parallelism	high	high/medium	medium	high
Multithreading	x	maybe		maybe
Multiprocessing	x	x	probably not	x
Number of threads and/or processes		<10k processes	about 4 threads	To Be Discussed (TBD)
Memory required	As Much As Possible (AMAP)		TBD	TBD
Hard disk technology	SSD	SSD, support for MPI I/O		SSD
Hard disk space	AMAP	depends on the scenario, <1TB	dominant part is PSNC output, if/when stored	<1TB
Operating System	Unix compatible	Unix compatible	Linux Ubuntu 16.04	Linux
Programming language	C, C++	Fortran, C++	C, C++	Python, C/C++
Libraries		(p)netcdf, hdf5, standard	lm, lpthread	Keras
Frameworks				Tensorflow

A study of hardware product manufacturers was conducted to determine which ones can meet these requirements and fit the prototype solution pursued by the consortium members in RECIPE. Manufacturers included Supermicro®, NVIDIA®, AMD®, Intel®, and Xilinx®, among others. After evaluating all constraints, two different Supermicro® servers were selected for the server part of the prototype, while NVIDIA® GPU and Intel® FPGA boards were selected for enabling different types of acceleration support. We decided to acquire Supermicro® servers based on the advise of some of the partners having a solid expertise and knowledge of the HPC market, suggesting that this company offers the best price/performance ratio for a large variety of products. The top manufacturers of FPGAs are Intel® and Xilinx®. As the preferred choice, we selected Intel®, because of the floating point requirements of the use case provided by BSC partner. Unfortunately Xilinx® does not offer FPGA solutions with hardened floating point functional units. However, Xilinx® devices will be possibly used for demonstrating specific features, e.g. High Bandwidth Memory (HBM2), in case the corresponding products by Intel® are not available on the market within the required timing. On the other hand, NVIDIA® GPUs were selected since they are the most powerful high performance architecture for accelerating HPC workloads nowadays and, according to the compatibility matrix provided by the vendor [17], they are the only device compatible with the servers chosen for the project as well.

Table 2: List of technology application requirements for the accelerator hardware.

	BSC	PSNC	CHUV
types of accelerators			
GPU	N/A	not yet	not yet
GPU family or model	N/A		
FPGA	x	N/A	maybe
FPGA family or model	Arria 10 10AX115	N/A	
GPU	N/A		
number of kernels	N/A		
number of threads/kernel	N/A		
Memory required	N/A		
Programming language	N/A	CUDA	tensor-GPU (CUDA)
Host-Device bandwidth (GB/s, Gb/s)			
FPGA			
Memory required	AMAP	N/A	N/A
Floating point arithmetic	yes (32-bits)	N/A	yes (TBD)
Fixed point arithmetic	No	N/A	yes
Computational model	Data-flow, pipelining	N/A	
Programming language	HLS	N/A	HLS
Host-Device bandwidth (GB/s, Gb/s)	AMAP	N/A	

In addition, existing portions of the source code of some use cases are already prepared to be run on CUDA, which is the target programming model for NVIDIA® GPU cards. Next, we describe every hardware component in the prototype. Some of these components use an innovative approach to memory design, Chip-on-Wafer-on-Substrate (CoWoS®) to integrate in a single package computing resources with HBM2 to boost memory bandwidth performance.

Supermicro® servers

The Supermicro® X11 GPU [18] is a GPU system conceived to suit the most demanding workloads requiring GPU co-processing capabilities, as it occurs with a wide range of HPC applications. It supports the second generation of Intel® Xeon® processors [4], Intel® Optane$^{\text{TM}}$DC Persistent Memory [3] and multiple options for hard disk drives offering the optimal foundation for maximum performance with the latest GPU technologies. The line of GPU systems supports 1, 2, 3, 4, 8, 16 or 20 GPUs across form factors ranging from 1U[1] to 10U systems. Namely, we adopted the SuperServer® SYS-1029GP-TR [19] and SYS-1029GQ-TNTR [20] as general purpose resources in the RECIPE prototype.

SuperServer® SYS-1029GQ-TNRT. The Figure 1 extracted from the Supermicro® Web-

[1] A rack unit (abbreviated U) is a unit of measure defined as 44.45mm and it is typically used as a measurement of the overall height of rack frames, which normally host these types of equipment.

site shows the internal top view of this server. Its main key features are listed below.

- 1U chassis form factor.
- Up to 4 NVIDIA® GPUs with active or passive cooling mechanism.
- Dual socket P (LGA 3647) compatible with the second generation of Intel® Xeon® Scalable processors (Cascade Lake/Skylake) for a maximum of 28 cores with Intel® Hyper Threading (HT) technology and three Intel® Ultra Path Interconnect (UPI) for point-to-point processor communication.
- CPU Thermal Design Power (TDP) up to 165W.
- Up to 3TB of Dual Data Rate (DDR4) Registered buffered or Load Reduced memory at 2933MHz.
- 4 PCI-Express 3.0 x16 Full Height Full Length (FHFL) slots and 2 PCI-Express 3.0 Low Profile (LP) slots providing 16 and 8 lanes respectively.
- 4 2.5" drive bays. Two of them are Hot-swap Non-Volatile Memory Express (NVMe) and the other two are Serial Advanced Technology Attachment (SATA).
- 2000W redundant Titanium level (96%) power supplies.

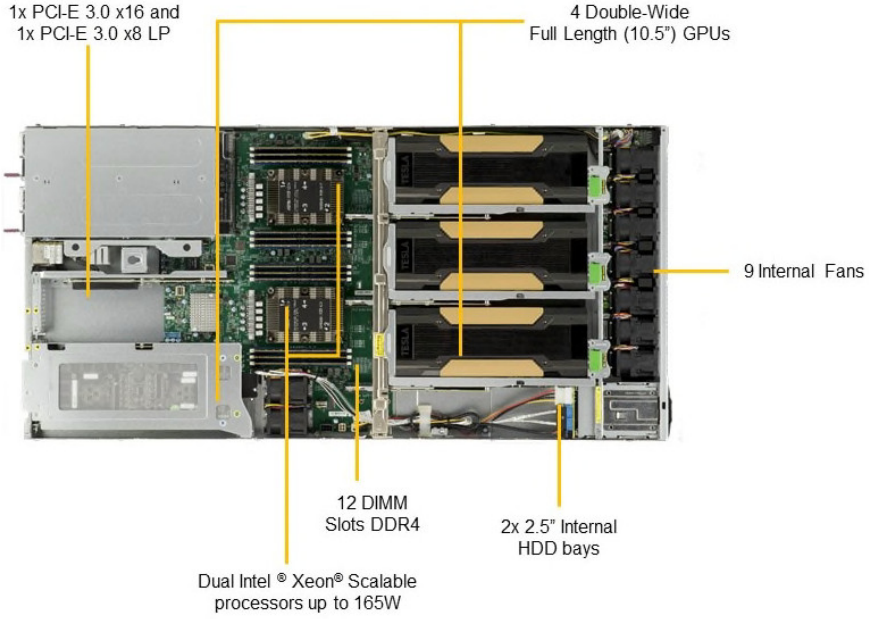

Figure 1: Supermicro® SuperServer® SYS-1029GQ-TNRT internal view.

SuperServer® SYS-1029GP-TR. The Figure 2 extracted from the Supermicro® Website shows a front view of this system. Although it is very similar to the previous one, there are minor differences that we highlight in the next listing.

- Up to three passive cooling NVIDIA® GPUs in a 1U rack mount chassis form factor.
- Up to 4TB DDR4 Registered buffered or Load Reduced memory at 2933MHz.
- 1600W redundant Platinum level (94%) power supplies.

Figure 2: Supermicro® SuperServer® SYS-1029GP-TR chassis

NVIDIA® Tesla® GPU boards

In words of NVIDIA® company "NVIDIA® Tesla® [12] constitutes the most advanced data center GPU platform in the world specifically conceived for accelerating nearly all demanding HPC and hyperscale workloads". For this project, the consortium decided to select two different models: the PCIe NVIDIA® Tesla® P100 [13] and NVLink™ V100 [14] GPUs. The decision of acquiring two different models was determined in order to evaluate the performance impact of NVIDIA Volta cores. The main specifications of the NVIDIA® Tesla® V100 and P100 GPUs are given in Table 3.

Table 3: NVIDIA® Tesla® P100 and V100 specifications.

	P100	V100
Architecture	Pascal	Volta
Tensor Cores	-	640
CUDA® Cores	3584	5120
Double-Precision Perfomance	4.7 teraFlops	7.8 teraFlops
Single-Precision Performance	9.3 teraFlops	15.7 teraFlops
PCIe x16 Interconnect Bandwidth	32GB/s	32GB/s
NVIDIA® NVLink™ Interconnect Bandwidth	-	300GB/s
CoWoS® HBM2 Stacked Memory Capacity	12GB	32GB
CoWoS® HBM2 Stacked Memory Bandwidth	549GB/s	900GB/s
Max Power Consumption	250W	300W

NVIDIA® NVLink™ [10] is a high bandwidth and power efficient communication interface for interconnecting GPUs among them, and the GPU and NVLink™-enabled CPUs as well. It can bring up to 70% more performance to an otherwise identically configured server by providing higher bandwidth, more interconnect links and improved scalability for multi-GPUs and multi-GPU/CPU system configurations.

Pascal architecture [11] enables a new computing platform built in a set of technological breakthroughs: I) a 16 nanometer FinFET fabrication technology delivering the fastest performance and best energy efficiency for very high computing demand workloads. II) NVIDIA® NVLink™ for maximum application scalability. III) Unified processor and data in a single package using an innovative approach to memory design. IV) New half-precision, 16-bit floating point instructions. V) Build a large number of CUDA® cores, offering over 5 TFLOPS of double precision performance.

Volta architecture [15] is targeted at Artificial Intelligence (AI). It is a new architecture composed of CUDA® and Tensor Cores, which incorporates 640 Tensor Cores for delivering over 125 TFLOPs of deep learning performance. It uses next generation NVIDIA® NVLink™ interconnect technology to deliver 2X throughput, compared to the previous generation of NVLink™.

HBM2-enabled FPGA card

In order to provide the most advanced solution for the RECIPE Prototype the consortium decided to acquire a 520N-MX HBM2-enabled device [1] announced by BittWare® company. The key features of this card are as follows.

- Intel® Stratix® 10 MX FPGA.
- 16GB integrated HBM2 memory.
- High bandwidth memory up to 512GB/s.
- Four QSFP28 cages supporting up to 100G per port.
- Two DIMMs supporting DDR4 SDRAM, QDR-II SRAM or Intel® Optane™ 3D-XPoint.
- Two OCuLink ports for direct expansion to NVMe SSD arrays.
- OpenCL Software Development Kit (SDK).

However, the 520N-MX card was expected to be commercialized originally in the first half of 2019 according to the vendor. Unfortunately, its commercialization has been postponed to the last semester of the year, generating some uncertainty regarding the compatibility with the RECIPE timings. Thus, in order to meet the planned timing for the integration of the FPGA card in the prototype, we also considered acquiring as an alternative to this solution a XUP-VVH [2] board from the same vendor, whose main features are listed below:

- Xilinx® Virtex® UltraScale+ VU35P/VU37P
- 8GB of integrated HBM2 up to 460GBps
- PCIe x16 interface supporting up to Gen3
- Four QSFP cages for 4x 40/100GbE or 16x 10/25GbE
- Up to 256GB DDR4
- UltraPort SlimSAS™ for serial expansion

or a Xilinx® Alveo A-U280-P32G-ES1-G board, whose main features are listed below:

- Xilinx® UltraScale+ device
- 8GB of integrated HBM2 up to 410GBps
- PCIe Gen4x8
- Two QSFP28 (100GbE)
- 32GB DDR4.

MANGO Cluster

The MANGO cluster is based on the proFPGA prototyping system [16] commercialized by Prodesign Electronic GmbH company.

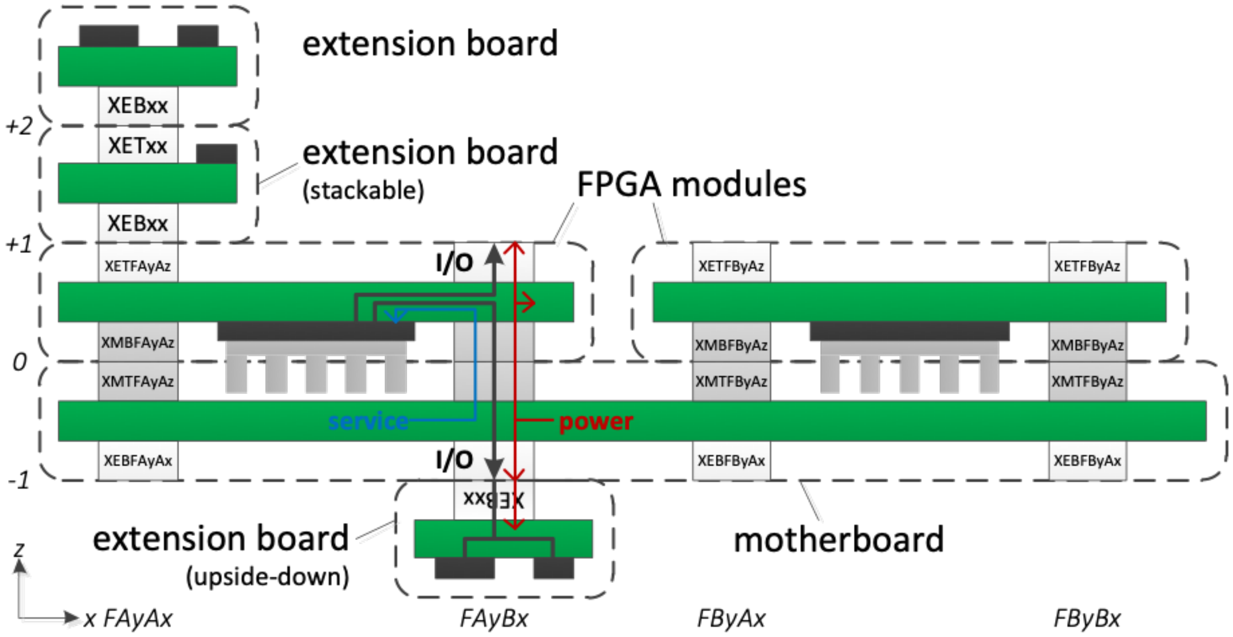

Figure 3: Modular hardware approach of the proFPGA prototyping system. FPGA module connectors (between motherboard and FPGA modules) are shown in grey. Extension board connectors (towards extension boards) are shown in white.

Motherboards provide the proFPGA system infrastructure. They offer mechanical fixture, power supply, Inter-Integrated Circuit (I2C) -based system management, clocking infrastructure, and 64-bit Multi Message Interface (MMI-64) communication for multiple FPGA modules. Motherboards have FPGA module connectors (carrying user I/O, power supply, service) on the top side and Extension board connectors (carrying user I/O, power supply) on the bottom side. The user I/O pins of top-side and bottom-side connectors are directly connected with each other, providing a transparent connection from the FPGA module on the top side to the extension board on the bottom side.

FPGA Modules contain the user design. They offer connectors to extension sites. Each FPGA module has access to MMI-64 communication from the motherboard. FPGA modules have 4

FPGA module connectors (user I/O, power supply, service) on the bottom side and Extension board connectors (user I/O, power supply) on the top side. Because the motherboard transparently converts the FPGA module connector into an Extension board connector, each FPGA module can access up to 4 extension boards on the bottom side. The FPGA module must be assembled with the FPGA facing downwards. The holes in the motherboard are intended for the FPGA heat sinks.

Extension Boards provide hardware functions to user designs inside the FPGA modules, e.g. SDRAM memory, user PCIe connection, debug access. One extension board occupies one or more extension board connectors of one FPGA module, giving the user design inside the FPGA module exclusive access to the extension board. The Extension board connectors of the FPGA module are located on the bottom side. Extension boards and interconnect boards on the bottom side of the motherboard must be assembled upside-down. Some extension boards (e.g. the user PCIe adapter) are stackable. Unused I/O pins from the FPGA module are mapped to a top-side connector, allowing further extension boards to be added.

Interconnects are special extension boards to connect I/O pins of different FPGA modules. They are available as boards and cables. Both Interconnect cables and boards connect two or more extension sites. Connections are either broadcast (e.g. the 4-way interconnect board) or point-to-point (e.g. all two-way interconnect boards and cables).

System Extension Boards may be extended by special hardware, such as motherboard PCIe adapter board or motherboard-to-motherboard connector cable. This hardware uses dedicated connectors on the motherboard.

Figure 4 shows the particular elements acquired to compose a basic MANGO cluster for the project. This particular cluster is composed of a dual motherboard, an Intel® Stratix® 10 FPGA module and a PCIe x8 Gen3, a 4GB DDR4 and a 128Gb NVRAM extension board. This setting will allow the implementation of checkpointing techniques at the FPGA level by using the plugged NVRAM extension board. In addition, the use cases requiring floating point units will take advantage of the hardened floating point units available in the Intel® Stratix® 10 FPGA.

Infiniband® Interconnect

In order to interconnect the different heterogeneous nodes integrated in the prototype, an Infiniband® network was adopted by the consortium, as planned originally in the proposal. High performance system interconnect technologies can be divided basically into three categories: Ethernet®, Infiniband®, and vendor specific interconnects, as the recently introduced Intel® Omni-Path Architecture (OPA) technology. Ethernet® is established as the dominant low level interconnect standard for mainstream commercial computing requirements and it has continued to evolve, driving specifications that reached performance levels of 400 Gbps in 2017. Infiniband® is designed for scalability, using a switched fabric network topology together with Remote Direct Memory access (RDMA) to reduce CPU overhead. It enables maintaining a performance and latency edge in comparison to Ethernet® for many HPC workloads. The IBTA roadmap has shown bandwidth for Infiniband® reaching 600 Gbps in 2017 and is intended to keep the rate of performance increase in line with system-level performance gains. Introduced in 2015, Intel's end-to-end OPA claims higher messaging rates and lower latency than Infiniband®,

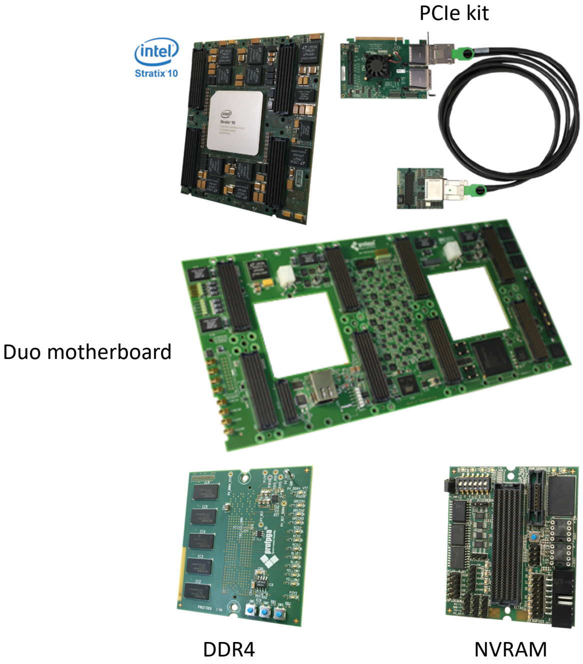

Figure 4: ProDesign proFPGA Duo motherboard, Stratix 10 FPGA, DDR4, NVRAM and PCIe Kit.

in addition to advanced features such as traffic flow optimization, packet integrity protection and dynamic lane scaling, but it is not a formal standard. Thus, Infiniband® technology was chosen as the HPC interconnect for the prototype due to its better performance with respect to Ethernet® and because, unlike Intel® OPA, it is a formal standard. In addition, Infiniband® provides greater level of flexibility than other standards to define Quality of Service (QoS) guarantees through the Virtual Lanes (VLs) and Service Levels (SLs) included in its specifications.

The adoption of Infiniband® in the RECIPE project will rely on the acquisition of two different components: a Mellanox® MCX555A-ECAT DDR network card [8] plus a Mellanox® MSB7890-ES2F EDR 1U 36 port QSFP28 UNMANAGED switch [9]. Although four of these network cards were acquired, the switch acquisition has been deferred to the point where a decision will be taken about the prototype location (either unified or split at UPV and CeRICT premises, as described in Section 3). The main reason is due to the fact that an Infiniband® switch is extremely expensive and it provides 36 ports which is largely enough to connect all different heterogeneous nodes acquired for the prototype. Thus, it is worth to evaluate carefully the possibility of acquiring only one of those, instead of two as proposed initially. Notice that this decision does not delay any project task because the nodes can be connected in a point-to-point way by using network cards enabling networking capabilities. The main features of the acquired network cards are listed below.

- Interopeability with Infiniband® switches up to Enhance Data Rate (EDR) communication signaling technique.
- IBTA Specification 1.3 compliant.
- Remote Direct Memory Access (RDMA) Send/Receive semantics.

- Hardware-based congestion control.
- Atomic operations.
- 8 virtual lanes.

Next, we list the key features of the selected Infiniband® switch.

- 19 inch rack mountable 1U chassis.
- 36 QSFP28 non-blocking ports with aggregate data throughput up to 7.2Tb/s (EDR).
- IBTA 1.21 and 1.3 compliant.
- 9 virtual lanes.
- Adaptive routing.
- Congestion control.

2.2 Software Components

The servers were installed with the required software to satisfy the application or use case requirements. In particular, the operating system installed was Linux Ubuntu 18.1 Long-Term Support (LTS), since it was the most up to date Linux Ubuntu version compatible with the Supermicro® servers. Initially, Linux Ubuntu 16.4 LTS was recommended by IBTS partner. Nevertheless, we finally agreed on installing a more updated version since there was no strong justification to avoid this. Also, we did not want to be out of date in an early stage of the project. On top of this operating system the common development toolsets for C, C++, Fortran, etc. as any other library specified in the application requirements were installed as well. For GPU development we decided to install CUDA 10, since it was the current stable version. Additionally, we installed the first version of the RECIPE software stack development kit, which is presented in Deliverable D2.1.

3 RECIPE Prototype

Figure 5 illustrates the final prototype built in the context of RECIPE. As it can be observed, it is composed of four heterogeneous nodes interconnected through an Infiniband® switch. Although the whole consortium was involved in the selection of hardware components during the three first months of the project, both the UPV and CeRICT partners took the responsibility of acquiring a half of the prototype each, as signed in the Grant Agreement. For the UPV partner, the acquisition procurement lasted about six months. Thus, the RECIPE prototype ordered by UPV was made available for the whole consortium in the ninth month (M9) of the project. For CERICT partner, the components acquisition took similar time and its prototype part was made ready for project partners in the tenth month of the project. From month 9 to 18 UPV and CERICT partners are maintaining their respective prototypes. This initial step will allow concurrency in the research and will maximize accessibility to the partners. In the last period

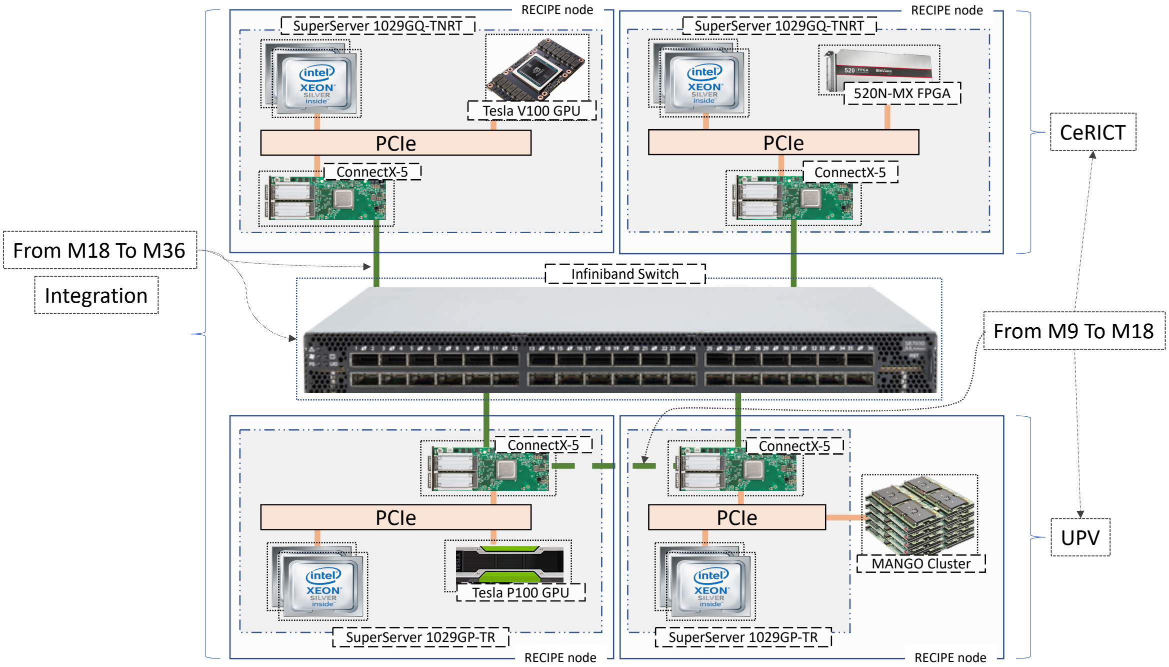

Figure 5: Whole RECIPE prototype showing its division in two parts from month 9 to 18 and possible integration in a single one from month 18 to 36 (to be decided).

of the project timeline (from month 19 to 36) the consortium will refine the prototype strategy and opt for a unified or a split prototype, depending on a cost/benefit analysis and the priority of the technical objectives. In case the two prototypes located at UPV and CeRICT premises will be put together, then the equipment will be hosted at one of the RECIPE partner facilities to compose a single prototype connected through a number of Infiniband® switches (in the Figure 5 only one switch is shown). For that goal, the Infiniband® switches will be acquired. However, as pointed out before, at this moment it is not clear to the consortium whether two Infiniband® switches will be acquired as planned initially or finally we will decide to acquire only one of them, since it does not compromise any research planned in the project work plan and the switch selected by the consortium contains enough ports for interconnecting the four servers without any problem. Up to the time of integrating the two prototypes by using that Infiniband® switch the two servers located at the different premises are connected in a point-to-point way across the network cards, as the slash green line shows in Figure 5.

The UPV prototype part (see the bottom part of the Figure 5) is composed of two Supermicro® SuperServer® 1029GP-TR servers, hosting up to four full-height double PCIe cards. Each server includes two Intel® Xeon® Silver 4112 processors [6] with 4 cores and 8 threads each, 8.25MB of L3 cache, 2.60GHz of processor base frequency and 9.6GT/s Intel® 2 UPI links for point-to-point communication between the two processors. In addition, the two servers have 32GB of system memory, a 480GB Solid State Drive (SSD) and one Infiniband® Mellanox® MCX55A-ECAT EDR network card each. The subsystem integrates in one of the two servers an NVIDIA® Tesla® P100 GPU computing processor with PCIe Gen3 x16 interface, 3584 CUDA® Cores, 12 GigaBytes of device memory based on HBM2 technology, 549 GBps of memory bandwidth, support for CUDA®, DirectCompute, OpenCL, and OpenACC APIs. In the other server a MANGO cluster of FPGAs is connected by PCIe Gen3 x8. This cluster consists of one Intel®

Stratix® 10 FPGA, 4GB of DDR4 memory and 128Gb of NVRAM, enabling the possibility of leveraging checkpointing mechanisms at FPGA level as well.

The prototype set up by CeRICT in RECIPE (see the top part of the Figure 5) is comprised of two independent servers, one extended with a GPU card and one to be extended with an FPGA card. Both cards used in the prototype offer high-end devices with top-of-the-line performance and support for high bandwidth memory. In addition, both servers are equipped with an Infiniband® adapter. In particular, the two servers are the Supermicro® SuperServer® 1029GQ-TNRT model, hosting up to four full-height double-width PCIe cards. Each server mounts two Intel® Xeon® Silver 4110 processors [5] with 8 cores and 16 threads each, 11 MegaBytes of L3 cache, 2.10 GHz of processor base frequency, 9.6GT/s Intel® UPI links for point-to-point processor communication. Furthermore, the two servers have each 64 GigaBytes of system memory, two redundant 480GB SSDs as well as one Infiniband® Mellanox® MCX55A-ECAT EDR network card. The GPU card plugged in one of the two servers is an NVTV100-32 device offering an NVIDIA® Tesla® V100 GPU computing processor with an NVLink^{TM} interface, 5120 CUDA® Cores, 32 GigaBytes of device memory based on HBM2 technology, 900 GBps of memory bandwidth, support for CUDA®, DirectCompute, OpenCL, and OpenACC APIs. The acquisition of the HBM2-enabled FPGA card has currently been deferred and is expected by September 2019. This timing was caused by the delayed commercialization of HBM2-enabled devices by Intel® (available on a 520N-MX board announced by BittWare®), which was originally planned for the first half of 2019 according to the vendor. Because of the current uncertainty regarding the maturity of HBM2-enabled devices, and in order to meet the planned timing for the integration of the FPGA card, CeRICT also considers acquiring as an alternative the BittWare®'s XUP-VVH board, providing a high-end Xilinx® UltraScale+ VU37P device with 8 GigaBytes of integrated HBM2 memory offering a bandwidth of 460 GBytes/s, a 3/4-Length PCIe form factor, Quad QSFP, and 256 GigaBytes of DDR4 memory, or a Xilinx® Alveo A-U280-P32G-ES1-G board, featuring a high-end Xilinx® UltraScale+ device, 8GB of integrated HBM2 up to 410GBps, PCIe Gen4x8, two QSFP28 (100GbE), 32GB DDR4. In any case, the main aim of the heterogeneous prototype being assembled by CeRICT is to provide RECIPE partners with access to top-of-the-line devices representing the GPU and FPGA computing paradigms, with the two types of devices both equipped with high-bandwidth memory in order to ensure consistent performance evaluation and comparisons. The two servers are hosted in a dedicated 42U server cabinet that has been purposely acquired, with each server being powered by a 3000 VA uninterruptible power supply (UPS) unit. An external PC has also been acquired by CeRICT to complete the equipment used for the purposes of RECIPE.

It should be pointed out that the prototypes acquired by UPV and CeRICT partners are slightly different with the aim of providing wider variety of options for demonstrating the main objectives the project is pursuing on through runtime resource management.

4 Conclusions

This deliverable reports on the heterogeneous HPC prototype built in RECIPE to fully demonstrate advanced fault-tolerance, reliability and Quality-of-Service (QoS) capabilities through resource management at each level of the system infrastructure.

Both the hardware and software components have been described and their acquisition for the RECIPE prototype has been justified from a technical point of view. In order to experiment with top-of-the-line FPGA devices, the consortium decided to defer the acquisition of an Intel® HBM2-enabled FPGA board to September 2019, when it is supposed to be commercialized, or otherwise opt for an HBM2-enabled Xilinx device as an alternative.

The assembled prototype system has been illustrated and described exposing how it has been acquired and how it will evolve across the different project phases from the initial setup to the final setting expected for the end of the project.

References

[1] BittWare 520N-MX. https://www.bittware.com/fpga/520n-mx/, Accessed April 2019.

[2] BittWare XUP-VVH. https://www.bittware.com/fpga/xup-vvh/, Accessed April 2019.

[3] Intel Optane Technology. https://www.intel.com/content/www/us/en/architecture-and-technology/intel-optane-technology.html?_ga=2.207337811.1759416065.1555062372-1290923783.1555062350, Accessed April 2019.

[4] Intel Xeon processors. https://ark.intel.com/content/www/us/en/ark/products/series/192283/2nd-generation-intel-xeon-scalable-processors.html, Accessed April 2019.

[5] Intel Xeon Silver 4110. https://ark.intel.com/content/www/us/en/ark/products/123547/intel-xeon-silver-4110-processor-11m-cache-2-10-ghz.html, Accessed April 2019.

[6] Intel Xeon Silver 4112. https://ark.intel.com/content/www/us/en/ark/products/123551/intel-xeon-silver-4112-processor-8-25m-cache-2-60-ghz.html, Accessed April 2019.

[7] MANGO: exploring Manycore Architectures for Next-GeneratiOn HPC systems, Accessed April 2019. http://www.mango-project.eu/overview.

[8] Mellanox ConnectX-5 IBA HCA. http://www.mellanox.com/page/products_dyn?product_family=258&mtag=connectx_5_vpi_card, Accessed April 2019.

[9] Mellanox Switch SB7800 Series. http://www.mellanox.com/page/products_dyn?product_family=225&mtag=sb7800, Accessed April 2019.

[10] NVIDIA NVLink. https://www.nvidia.com/en-us/data-center/nvlink/, Accessed April 2019.

[11] NVIDIA Pascal architecture. https://www.nvidia.com/en-us/data-center/pascal-gpu-architecture/, Accessed April 2019.

[12] NVIDIA Tesla. https://www.nvidia.com/en-us/data-center/tesla, Accessed April 2019.

[13] NVIDIA Tesla P100. https://www.nvidia.com/en-us/data-center/tesla-p100/, Accessed April 2019.

[14] NVIDIA Tesla V100. https://www.nvidia.com/en-us/data-center/tesla-v100/, Accessed April 2019.

[15] NVIDIA Volta architecture. https://www.nvidia.com/en-us/data-center/volta-gpu-architecture/, Accessed April 2019.

[16] ProDesign proFPGA system. https://www.profpga.com/products/systems-overview, Accessed April 2019.

[17] Supermicro compatible GPU list. https://www.supermicro.com/support/resources/GPU/, Accessed April 2019.

[18] Supermicro GPU solutions. https://www.supermicro.com/products/nfo/GPU_MIC.cfm, Accessed April 2019.

[19] Supermicro SuperServer 1029GP-TR. https://www.supermicro.com/products/system/1U/1029/SYS-1029GP-TR.cfm, Accessed April 2019.

[20] Supermicro SuperServer 1029GQ-TNRT. https://www.supermicro.com/products/system/1U/1029/SYS-1029GQ-TNRT.cfm, Accessed April 2019.