D4.1
Pilot Modular Architecture Description

Document Properties

Contract Number	101033975
Contractual Deadline	M15 – 31/03/2023
Dissemination Level	Public
Nature	Report
Edited by :	Maike Gilliot (CEA)
Authors	Maike Gilliot, (CEA), Romain Fihue (CEA), Philippe Gregoire (CEA)
Reviewers	Eric Boyer (GENCI), Sergio Saponara (Univ. of Pisa), Andrea Bartolini (UNIBO)
Date	28/03/2023
Keywords	Architecture definition, sizing, interconnect, storage
Status	Final
Release	1.0


This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101033975. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Germany, Italy, Greece, United Kingdom, Czech Republic, Croatia.

History of Changes

Release	Date	Author, Organization	Description of Changes
0.1	16/12/2022	M.Gilliot, CEA	Structure, outline and table of contents
0.2	08/03/2023	R. Fihue, CEA	Technical outline & details
0.3	18/03/2023	M.Gilliot, CEA	Adding and detailing more elements, preparing a version for internal review
0.4	27/03/2023	M.Gilliot, CEA	Version including all comments and corrections from the reviewers
1.0	28/03/2023	M.Gilliot, CEA	Final Version for submission after final comments have been included

Table of Contents

List of Figures

List of Tables

1 Executive Summary

The goal of the EUPEX Pilot is to integrate the different hardware components (provided by WP4) and software components (provided by WP5) into a single pilot system as to show the interoperability and the maturity of these components for future HPC systems, targeting exascale-size machines. This deliverable D4.1 depicts the suggested architecture for the EUPEX Pilot, focusing on the hardware components: its compute nodes, the interconnect and the memory and storage related aspects.

The definition of the compute nodes themselves is _not_ in the scope of this deliverable: the design criteria and choices for the GPU and the CPU blade are detailed in D4.2. Some parameters of this architecture have already been defined during the proposals phase, and others depend on SiPearl Rhea chip and cannot be influenced neither. For the remaining parameters, input from WP3 and WP5 was collected in a co-design approach.

This deliverable describes the structured dialogue with WP3 for understanding their requirements. It details the different parameters we examined and the suggested architecture for the pilot. The elements of this deliverable will feed into D6.2, which is due on M24 and will refine – if need be – the pilot architecture and define the annex components of the pilot system.

2 Introduction

The EUPEX Pilot will integrate the different hardware components (provided by WP4) and software components (provided by WP5) into a single pilot system. The goal is to show the interoperability and the maturity of these components for future larger HPC systems, with exascale-size machines as target. Some parameters of this architecture are already set:

- Some have already been defined during the proposals phase (such as the Modular Computing Architecture (MSA for short))
- Some depend on the SiPearl Rhea chip
- Some are defined by the GPP or GPU blades, which are part of T4.2 and presented in D4.2

For the remaining parameters, input from WP3 and WP5 was collected in a co-design approach. This deliverable is a first step towards the full architecture definition of the pilot, focusing on the requirements and the needs coming from the applications. It focuses on the hardware components: its compute nodes, the interconnect and the memory and storage-related aspects.

This deliverable D4.1 will serve as input for D6.2, (due in M24), which will refine – if need be – the pilot architecture and also define the annex components of the pilot system. D6.2 will also check the financial feasibility of the pilot and make sure that the pilot and its annex systems will remain within the planned budget of the EUPEX project. As such, D4.1 is an important building block for the overall architectural definition to come.

In Chapter 3, we will describe our co-design approach with WP3, before detailing the predefined parameters in Section 4.1. Section 4.2 then presents the suggested architecture for the pilot. Chapter 5 closes this deliverable by indicating the next steps regarding the overall design of the pilot.

3 Co-design approach

Despite the delay in building and acquiring the Pilot, the partners have agreed to keep this first architecture definition document in M15. As pointed out, these elements will feed into D6.2, which will describe the full Pilot system, including its annex components. D6.2 will also consider budget considerations. The architecture suggested in this deliverable is based on the elements at hand. There are also some open questions regarding the GPU blades, as not all decisions have been taken yet. This deliverable will therefore remain agnostic regarding the different options for the GPU.

Again, additional constraints that could arise in the upcoming months will be taken into consideration by deliverable D6.2. The main contribution of D4.1 is to present a first architecture based on the constraints and expectations from the software side: the applications (gathered in WP3) as well as the EUPEX software tools and stack components (gathered in WP5).

A first presentation related to the Pilot architecture with WP3 and WP5 took place in April 2022, giving rise to some first requirements from WP5. However, it turned out that it is easier for the application owners to comment on a given architecture than to make architectural suggestions “out of the blue”. We therefore changed the approach: WP4 came up with a first suggestion in October 2022, assessed by WP3 and WP5. The input from WP3 and from WP5 allowed us to refine the initial suggestion. We also used the All-Hands meeting in Zagreb (13 and 14 of Feb. 2023) to refine the suggested Pilot architecture. In particular, some storage-related elements have been adapted to better suite the needs of T5.4 (dedicated to the EUPEX storage architecture).

The overall timeline of the pilot definition is sketched in the table below:

Where	When	Action
Special WP3/WP4/ WP5 call	20/04/2022	Presentation of the “set parameters” of the architecture and of the open points for discussion: kicking off the co-design discussion with WP3 and WP5: collecting first input and first elements from WP5
Special WP4 call	05/10/2022	Presentation of the first architecture elements to all WP4 partners
	5-12/10/2022	WP4 partners provide feedback in writing
WP4 call	19/10/2022	Present V01 of the overall architecture
WP3 call	14/10/2022	Present architecture to WP5 and collect their input
WP 5 call	04/112022	Present architecture to WP5 and collect their input
WP4 call	Nov 2022	Share outline of D4.1
	Dec 2022	Compile new Version taking into account the input from WP3 and WP5: V02 shared with all partners
All-hands meeting	13+14/02/2023	Discussion on D4.1 and the suggested architecture V02: review of the storage related element (for better integration of the IO-SEA software stack)
	17/03/2023	D4.1 ready for internal review

TABLE 1: OVERALL TIMELINE

4 Pilot Architecture

4.1 Pre-defined parameters

4.1.1 A modular supercomputing architecture

As expressed in the General Agreement, the EUPEX consortium aims to design, build and test a pilot system using the **Modular Supercomputing Architecture (MSA)** pattern. The pilot is the first system that will integrate the first generation of the **European Processor Initiative (EPI[1])**, code name RHEA, in combination with additional out-of-chip acceleration technologies (eg. GPGPU) in two different modules:

- A **GPP** module with CPUs nodes
- A **GPU** module with GPUs nodes

In addition, to validate the optimal interconnect topology to connect nodes within individual modules, the pilot must evaluate three valid approaches for communications between modules (cf. Figure 1):

1. Single federated topology:
   All the compute nodes share a single interconnection network

2. Gateway-linked distinct topologies:
   Distinct interconnection networks with gateways connected to each and every network

3. Federated distinct topologies:
   Distinct *interconnection networks with gateways connected to a single external network*

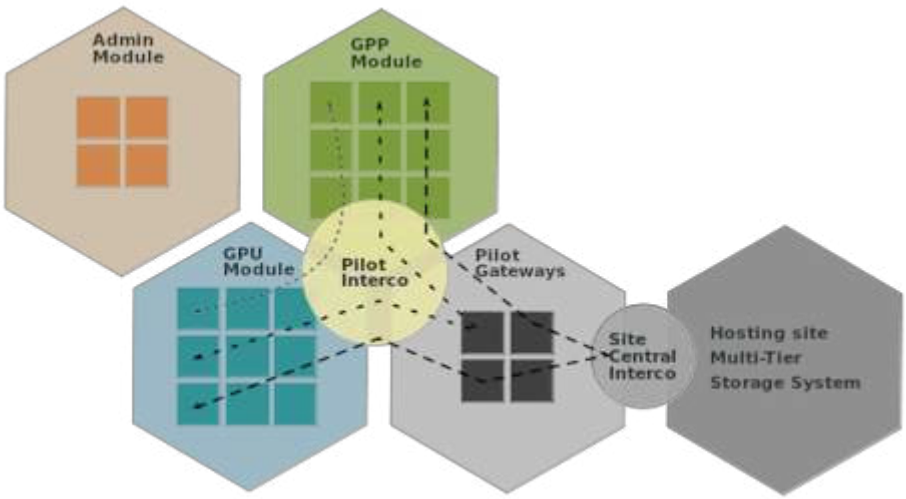

FIGURE 1: MODULAR SUPERCOMPUTING ARCHITECTURE: OVERVIEW

For the EUPEX Pilot we will select the first option: A unique network, to which we will apply different communication pattern to evaluate benefits and drawbacks of all three approaches.

[1] https://www.european-processor-initiative.eu/project/epi/

To make sure that this design assessment is the best direction to take for the upcoming Exascale systems, an iterative process will be followed with WP3 and WP5 to make sure that even corner-case are correctly evaluated (wrt. application or software needs and hardware constraints).

4.1.2 RHEA processor and CPU node constraints

With the current information from partners, the CPU nodes and the RHEA processors set some constraints on the pilot architecture design:

- Each rack can host 38 slots (interconnect network switches or blade).blades)
- The OpenSequana blades are equipped for direct liquid cooling[2]
- Each OpenSequana blade will host three distinct GPP nodes
- Each GPP node will be equipped with two RHEA sockets
- Each socket will provide:
    - Four channels of DDR5 (5600 MT/s) in addition to embedded HBM memory
    - Four stacks of HBM2e. As of today, it provides a total of 64GB of memory at 1.8TB/s
    - Two PCIe Gen5 x16 slots
    - Additional PCIe Gen5 x16 slots can be provided by downgrading inter-socket links (CCIX links, consequences to be determined):
        - 1 less CCIX link among the 4 links allows two more slots (one per socket)
        - 2 less CCIX link among the 4 links allows four more slots (two per socket)

4.1.3 GPU and GPU node constraints

With the current information from partners, the GPU nodes, RHEA processors and GPU technologies evaluated set some constraints on the overall design:

- Each rack can host 38 flexible slots (interconnect network switches or blades)
- Each OpenSequana blade will host a single GPU node
- Each GPU node will be equipped with two RHEA sockets
- Each socket will provide:
    - Four channels of DDR5 (5600 MT/s) in addition to embedded HBM memory
    - Two PCIe Gen5 x16 slots
    - Additional PCIe Gen5 x16 slots can be provided by downgrading inter-socket links (CCIX links, consequences to be determined):
        - 1 less CCIX link among the 4 links brings two more slots (one per socket)
        - 2 less CCIX link among the 4 links brings four more slots (two per socket)
- All GPU boards considered so far require four PCIe Gen5 16x slot

[2] Details on the cooling technolgies within the OpenSequana blades are detailed in D4.2

4.1.4 Co-design output

As a reminder, for the CPU node, the WP3 defined the following co-design parameters:

CPU Node	Codesign outputs
DDR memory	≥ 512GB (64GB DDR5 DIMM per channel at least) - note that the HBM memory is fixed for the Rhea chip
Number of sockets	2
Network bandwidth per node	≥ 200 Gbit/s
Number of NICs per node	≥ 2 (≥1 NIC per socket)
Local disk	Not required for the applications

TABLE 2: CO-DESIGN OUTPUTS FOR THE CPU NODE

For the GPU node, the WP3 defined the following co-design parameters:

GPU Node	Codesign outputs
Number of GPUs	4
DDR Memory	≥ 512GB (64GB DDR5 DIMM per channel at least) - note that the HBM memory is fixed for the Rhea chip
Number of sockets	2
Network bandwidth per node	400 Gbit/s
Number of NICs per node	4
Local disk	No required for the applications

TABLE 3: CO-DESIGN OUTPUTS FOR THE GPU NODE

In the co-design phase, T5.4 also expressed some requirements about the IO-stack to be deployed and evaluated on the pilot. The pilot will integrate the IO software stack developed within the IO-SEA project[3]. The targeted IO stack will leverage NVMe devices that are present on the compute node and/or the gateway nodes. The pilot allows testing and to assessing the different options thoroughly. This means that gateway nodes and compute nodes should be equipped with NVMe devices, and some CPU load could also be executed on the gateway nodes.

Given the GPU node constraints, especially on the available PCIe lanes, it seems as of today that only GPP and gateway nodes can be equipped with such drives.

As part of this co-design effort, another need emerged for the ability to collect power metrics for the different hardware components in order to allow for a thorough analysis of energy consumption. This is the base for optimisation of any kind in a second step by WP5 energy tools (such as MERIC or the Regale software infrastructure[4,5]). The co-design effort on these topics is still ongoing. The outcome will feed directly into the blade design (covered by D4.2).

[3] https://iosea-project.eu/

[4] https://code.it4i.cz/vys0053/meric

[5] https://regale-project.eu/

4.1.5 Physical constraints

The EUPEX Pilot will be hosted in two ATOS's XH3000 racks. Each rack can host 38 flexible slots (interconnect network switches or blade), so in total 76 OpenSequana blade slots will be available for the GPP and GPU blades and the interconnect network switches.

An additional standard rack will be available to host commodity hardware like management or gateways nodes.

Interconnect constraints also sets some limitations of the pilot's architecture:

- Cables between nodes and level-1 interconnect network switches cannot go to another rack
- BXIv2 switches is equipped with 48 ports
- Interconnect technologies considered so far (BXIv2) supports "Fat Tree", "Pruned Fat Trees" and "DragonFly" topologies

4.2 Architectural decision points

The overall objective of this architecture is to be flexible enough to validate multiple architectures that are presented for the next exascale architectures. Even though this architecture should represent the next designs, it must be an object close to reality to be easily deployed and operated by WP6.

4.2.1 Interconnect

The interconnect is a known bottleneck, especially on GPU-based architectures. Therefore, our design maximizes the interconnect performance, especially the interconnect injection bandwidth.

Hence, with the constraints expressed before, the interconnect endpoints are defined as follow:

- GPP nodes will be equipped with two interconnect endpoints, one per PCIe slot
- GPU nodes will be equipped with four interconnect endpoints. The interconnect device connection is feasible but still to be completely defined (Task 4.2)
- Gateway node endpoints bandwidth should reach about 10% of the total compute node injection bandwidth on each module

A solution that satisfies these constraints is a two-level Fat Tree topology with:

- Six level 1 switches for the GPP module that connects 32 GPP nodes
  - All GPP node interconnect endpoints connected to distinct level-1 switches
    - First NIC on the first three L1-switches
    - Second NIC on the last L1-switches
  - 15 uplinks: 3 links to each L2-switch

    => 15/32 blocking factor (~½)
- Four level 1 switches for the GPU module
  - All GPU node interconnect endpoints connected to distinct level-1 switches
  - 15 uplinks: 3 links to each L2-switch

    => 15/32 blocking factor (~½)
- Two level 1 switches for the gateway nodes
  - 20 gateway endpoints per switch: 4 links to each L2-switch
  - Two endpoints per gateway
  - 20 uplinks: no blocking factor
- Five level 2 switches
  - 3 downlinks to GPP and GPU modules level-1 switches
  - 4 downlinks to gateway node switches

GPU Node	Codesign outputs
Number of L1 switchs	6 (GPP) + 4 (GPU) + 2 (Gateways) = 12
Number of L2 switchs	5
Blocking factor	15:32 for CPU & GPU Module; 1:1 for gateways
Available ports per L1 switch	1 (GPP/GPU) or 8 (Gateway)
Available ports per L2 switch	10

TABLE 4: OUTPUTS FOR THE INTERCONNECT TOPOLOGY

This topology is defined in the following figure:

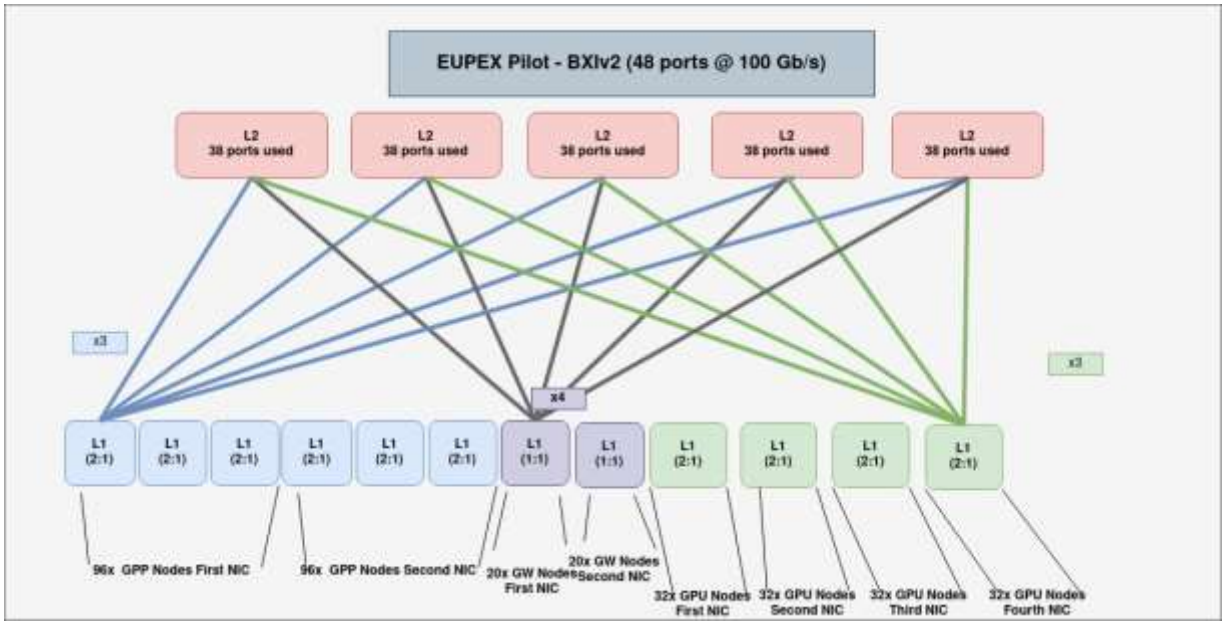

FIGURE 2: PILOT INTERCONNECT TOPOLOGY

4.2.2 Memory: type and size

The co-design output showed that applications may require more than 256GB of memory per socket. In fact, the requirement is on the quantity of memory per core. As of today's specifications of the RHEA1 and RHEA2 is very variable on the number of cores, 256GB of memory will most probably meet the requirement of a memory quantity per core between 2GB and 4GB. So, in total, each node must be equipped with 512GB of memory.

This requirement is the same for the GPP blades and the GPU blades, as they are equipped with the same CPUs. But another requirement has to be met that is induced by the presence of the GPU board. The total quantity of memory hosted on the GPU board (HBM stacks close to the GPU dies) must be lower than the quantity of memory on the host. So far, we are targeting 80GB to 128GB of memory per GPU. So with four GPUs, the total quantity of memory on the host must be close to 512GB. Given the RHEA processor constraints, this means that each GPU and GPP node must be equipped with 8 DIMM modules of 64GB each.

4.2.3 Gateway nodes characteristics

Gateway nodes must comply with requirements coming from WP6 for a correct integration into the existing computing center. Gateway parameters are defined as follows:

- Number of gateways

Indirectly defines the number of network endpoints. The main variable to reach the targeted bandwidth

- Number of endpoints per gateway

Depending on the design of the node, commodity hardware (cf. Figure 2) rarely proposes more than 2 PCIe X16 slots per CPU slot

The main constraint here is about latency between NICs. So two x16 slots per socket are the most pragmatic solution

- *Number of NVMe drives per gateway*

The number of NVMe drives should be comparable to the modules' interconnect bandwidth.

- Number of sockets

The gateway workload could be influenced by user-controlled workflows in addition to infrastructure workload (Site-central filesystems access)

GPU Node	Codesign outputs
Number of gateways	20
Number of BXI endpoints	2 mono-port cards
Number of HDR100 endpoints	2 dual-port cards
PCIe topology	1 BXI + 1HDR200 card on each CPU socket
DDR Memory	≥ 256GB
Number of sockets	2
Network bandwidth	200 Gbit/s per network
Number of NVMe drives	Enough drives to reach the total BXI bandwidth (~ 20 drives)
Local disk	HW-backed RAID for system installation

TABLE 5: OUTPUT FOR THE GATEWAY NODES

5 Conclusions and next steps

This deliverable describes the first version of the EUPEX Pilot architecture. It sets the scene for what our vision for the architecture is - as of today. This deliverable is a companion deliverable to D4.2, which provides more details and insight into the design choices for the GPP and the GPU blades (including aspects such as sizing, cooling, and others).

As pointed out, some elements are not fully available yet and may influence the architecture.

- The choice of the GPU
- As the porting effort in WP3 and WP5 progresses, more detailed requirements may emerge from the application side
- Budget constraints
- Availability of some technologies for this pilot

The elements provided by this deliverable will be taken up by D6.2 in order to define the overall Pilot architecture, including annex components for integrating the pilot into the existing Tier0 infrastructure at TGCC-CEA. As such, this deliverable is more a “starting point” than a “final presentation” of the EUPEX Pilot architecture definition.

6 Acronyms and Abbreviations

Term	Definition
BXI	Bull Interconnect (V2). 100GB/s link rate
CPU	Central Processing Unit
DDR5 DIMM	Dual inline memory module based on DDR5 technology
EPI	European Processor Initiative (EPI)
GPGPU	General Purpose GPU
GPP	General Purpose Processor
GPU	Graphical Processing Unit
HDR	Infiniband High Data Rate. 200GB/s link rate.
MSA	Modular Supercomputing Architecture
NIC	Network InterConnect
NVMe	Non-Volatile Memory express (protocol)
PCIe	Peripheral Component Interconnect Express

TABLE 6: ACRONYMS AND ABBREVIATIONS