D4.1

Advanced Technologies Monitor

Kaveh Haghighi Mood, Lubomir Riha, Ondrej Vysocky, Julio Gutiérrez Moreno, Elisabetta Boella, Augustin Degomme, Luigi Genovese, Mattia Paladino, and Loris Lucido

Due date of deliverable: 31/12/2023 (month 12)
Actual submission date: 28/12/2023
Final version: 28/12/2023

Lead beneficiary: E4 (participant number 13)
Dissemination level: PU - Public

Document information

Project acronym:	MAX
Project full title:	Materials Design at the Exascale
Research Action Project type:	Centres of Excellence for HPC applications
EC Grant agreement no.:	101093374
Project starting / end date:	01/01/2023 (month 1) / 31/12/2026 (month 48)
Website:	www.max-centre.eu
Deliverable No.:	D 4.1

Authors: Kaveh Haghighi Mood, Lubomir Riha, Ondrej Vysocky, Elisabetta Boella, Augustin Degomme, Julio Gutiérrez Moreno, Loris Lucido.

To be cited as: L. Riha et al., (2023): Advanced Technologies Monitor. Deliverable D4.1 of the HORIZON-EUROHPC-JU-2021-COE-01 project MAX (final version as of 28/12/2023). EC grant agreement no: 101093374, E4, E4 Computer Engineering SpA.

Disclaimer:

Versioning and contribution history:

Version	Date	Author	Note
0.1	3.11.2023	Kaveh Haghighi Mood	Initial version based on working document.
0.2	9.11.2023	Lubomir Riha	Edited IT4I contribution and access to platforms section.
0.3	13.11.2023	Ondrej Vysocky	Added EUPEX Alpha Pilot platform and platforms for energy efficiency analysis
0.4	27.11.2023	Elisabetta Boella	Added AMD and Intel roadmap
0.5	28.11.2023	Lubomir Riha, Augustin Degomme	Added NVidia and Sipearl roadmap
0.6	28.11.2023	Kaveh Haghighi Mood	Added code property description
0.7	30.11.2023	Kaveh Haghighi Mood, Lubomir Riha	Version ready for internal review
1.0	15.12.2023	Kaveh Haghighi Mood, Lubomir Riha	Final version for submission

D4.1 Advanced Technologies Monitor

Content

1. Executive Summary

This report introduces the efforts of WP4, which concentrate on identifying interesting technologies coming up in the time-frame of the MAX CoE that can be used as co-design platforms, advanced hardware platforms, and systems suitable for energy-efficiency evaluation. These platforms will be primarily utilised for benchmarking purposes via application kernels extracted from MAX codes but in some cases entire applications will be used.

At first we present the workload evaluation of MAX codes and mapping to features of the advanced HW platforms that suits them most. Defining the expected workload for each application and its modules allows us to concentrate on evaluating the most suitable platforms for a given code. For example, if one wants to optimise a memory-bound kernel, the most promising approach is to focus on processors or accelerators with new and more powerful memory technologies, such as DDR5, HBM (2, 3, 3e, …) that provide higher memory bandwidth.

Next we present new and non-traditional hardware platforms provided by all consortium members that can be utilised for code performance evaluation, benchmarking, and co-design. A summary of all platforms is presented in Table 3.6. Access mechanisms to these platforms are described in Section 4. Each partner who provided the platforms also defined instructions for accessing the systems and described the restrictions of their usage.

Additionally, WP4 also aims at evaluating the energy consumption and energy efficiency of different HW platforms when running MAX codes or their kernels. Table 3.7 presents a list of platforms suitable for this study, as they enable us to measure energy consumption and adjust hardware parameters that affect the power usage of a processor or accelerator for a given workload. These hardware tuning actions are challenging to carry out on production systems due to several constraints. In the table we have only included platforms on which we can execute the desired actions. Important fact is that we also have access to production systems that allow us to perform these actions at scale.

The final section details the roadmaps of the primary technology developers as well as upcoming platforms that have recently been released but are not yet available to any of the MAX partners.

In summary, this document serves as a starting point for all MAX3 partners who are interested in exploring the advanced hardware platforms for their codes, mini-apps or kernels in finding the new paths for code optimization and porting in order to achieve higher performance.

2. Introduction

Preparing MAX codes for future HPC hardware and analysing the suitability, scaling, and performance for upcoming extreme-scale architectures are essential components of the WP4. In this deliverable, we look at interesting technologies that are going to be advantageous and accessible in the time frame of the project. The technology roadmaps of different vendors to identify technologies with the potential to benefit the MAX application use cases are explored.

3. Available technologies

Silicon technology has extended far beyond the realm of traditional computers and has found applications and optimised for a wide range of industries and applications. For the MAX project, we aim at technologies that have the potential to improve the performance of MAX codes. In the following sections first, we briefly review the performance characteristics of MAX codes then we list available resources that are or will be provided by MAX partners.

3.1 MAX codes important kernels and performance characteristics

MAX codes have diverse kernels with different performance characteristics, as reported in deliverables 4.5 and 4.6 [1,2] of the second phase of the MAX project. MAX codes employ diverse types of algorithms. All codes contain memory, compute and latency bound kernels. For large use cases network latency and bandwidth is the usual bottleneck. The diversity of kernel characteristics makes many upcoming technologies interesting to monitor.

State of the art hardware can improve the performance of MAX codes on different fronts. Novel processor architectures such as Power 10, ARM v9, and RISC-V can improve energy efficiency and compute-bound kernels like GEMM. Powerful new GPUs like the upcoming AMD MI300 and Nvidia H100 can be great for both compute and memory-bound kernels. Innovative memory technologies such as DDR5, LPDDR5, or HBM (2,2e,3 or 3e) can enhance the efficiency of memory-bound kernels. More exotic devices might be more suitable for other bottlenecks. For example, FPGAs can be configured to be used for latency-bound kernels in Modular

Supercomputing Architectures (MSAs) or DPUs to boost IO performance. Last but not least, innovative ways of combining these technologies have the potential to ease problems like host-device bottleneck issues we observe on conventional GPU workloads.

In the following, we look closer at the performance characteristics of individual MAX codes.

3.1.1 Quantum ESPRESSO

Quantum ESPRESSO is a collection of codes and libraries for electronic structure calculations. using plane wave basis and pseudopotentials. The main engine is *pw.x*, which calculates energy and force. The code works on many functions that are operated alternatively in reciprocal space and direct space. For this reason, the most computationally intensive kernel is *FFTXlib*, which performs this transformation on the functions. Another relevant functionality is the iterative solution of eigenvalue problems and linear systems on linear spaces of up to several thousand elements. Other minor kernels such as *calbec*, *sum_band*, and others operate on the wave functions with similar complexity. See table 3.1 for the list of important kernels.

Kernel	**Functionality**	**Characteristic**	**Percentage of WCT**
FFTXlib	performs 3D Fourier transforms with reciprocal space data restricted within a selected cutoff radius. it can operate on a bath of many functions	Memory Bound Communication Latency Bound	30 to 60 depending on system size, parallelisation scheme, heterogenous or homogenous
cegterg/regterg	performs iterative diagonalisation with Davidson method	Compute bound On GPU limited by the memory of a single device. Distributed on homogeneous machines.	10-15 GPU 20-40 Distributed on CPU
calbec	Performs dgemm/zgemm	Compute Bound	10-20

Sum Band	Computes output charge density	Memory Bound Communication Latency Bound	5 -10
mix_rho	mixes input and output densities	Memory Bound	5

Table 3.1. Important kernels and performance characteristics of Quantum ESPRESSO.

3.1.2 YAMBO

The Yambo code, implementing Green's function methods and focused on excited state properties, is a modular code composed of various kernels. Memory-bound kernels, such as "*X irredux*" and "*GW*," account for 70-80 percent of the total wall-time (WT), while compute-bound kernels like "*X redux*" and "*BSE solver*" constitute the majority of the remaining time. In the case of large computations, the network bandwidth becomes a bottleneck only when handling a substantial number of MPI tasks. Generally, typical runs of Yambo are characterised as being memory-bound (see Table 3.2).

Kernel	**functionality**	**characteristic**	**Percentage of WT**
Dipoles	Dipoles matrix elements computation via LA operations	Memory bound	5-10
X irredux	Computation of Xo (irreducible response function) (FFT operations)	Memory bound	30-50
X redux	Inversion of Dyson equation to compute X (reducible response function) (LA operations)	Compute bound	5-20
HF	Computation of Exchange Self Energy	Memory bound	5-10

Kernel	functionality	characteristic	Percentage of WT
	(LA operations)		
GW	Computation of Correlation Self Energy (FFT and LA operations)	Memory bound	20-50
BSE	Construction of BSE kernel (LA operations)	Memory bound	10-20
BSE solver	Diagonalization of BSE kernel (LA operations)	Compute bound Memory bound	10-20

Table 3.2. Important kernels and performance characteristics of Yambo.

3.1.3 BigDFT

BigDFT is a density functional theory (DFT) code describing the electrons in a material, expanded in a Daubechies wavelet basis set. The code uses pseudopotentials and employs self-consistent direct minimization or Davidson diagonalization to determine the energy minimum. The most significant kernels of the code are related to data manipulation and emerge as techniques of signal processing (FFT, convolution), consequently, the overall characteristic is memory bound (see Table 3.3).

Kernel	functionality	characteristic	Percentage of WCT
Convolutions	Application of Hamiltonian in wavelet basis	Memory Bound (data processing). OpenCL GPU-ported. Released as a mini-app (libconv library)	Depending on systems, from >90% for a k-point calculator up to <10% for linear-scaling calculation of large systems
Poisson Solver	Solution of the Poisson Equation in wavelet basis	Memory Bound (FFT) Ported in SYCL, CUDA. Released as a mini-app (Fock program in Psolver library)	<5% for semilocal functional >70% for hybrid functional calculations
Fermi Operator	Sparse Matrix Linear	Memory Bound	Between 20% and 50%

Expansion	Algebra	(computations distributed over the compute nodes)	of the walltime for linear-scaling calculations, unused in cubic scaling

Table 3.3. Important kernels and performance characteristics of BigDFT.

3.1.4 Siesta

Siesta is a DFT code using localised pseudo-atomic orbitals as basis set. Hence, the Hamiltonian and overlap matrices (H,S) are sparse, and they can be generated in O(N) operations. Charge densities and potentials are represented on a real-space mesh. The solver stage (in which most of the time is typically spent) is implemented in several ways: 1. (Dense) diagonalization. 2. Direct computation of the density matrix from H and S (possible with the lower-complexity PEXSI algorithm and other Fermi-Operator-Expansion schemes). 3. Minimization of a special functional, also with linear scaling (but not as robust for systems without a sizable gap). Diagonalization is accelerated on GPUs through the use of the ELPA library (stand-alone or offered through the ELSI library of solvers). Most code operations are memory-bandwidth bound, except possibly the diagonalization when using optimised libraries. The operations on sparse matrices require less memory but have irregular memory-access patterns. See table 3.4 for the list of important kernels.

Kernel	**functionality**	**characteristic**	**Percentage of WCT**
Phi On Mesh	Calculation of the basis function values on mesh.	Memory bound	Up to 5% of the computation time for small systems (<500 orbitals), negligible for larger systems.
Poisson	Solution of the poisson equation	Currently done with FFT Memory bound	Up to 15% on small systems, it decreases to less than 2% for large systems.
Rho of D	Calculation of the electron density on the grid from a sparse subset of the density	Memory bound	Up to 10% on small systems, it decreases to less than 1% for systems larger than

	matrix.		10.000 orbitals.
XC	Calculation of the exchange-correlation potential.	compute bound	Up to 30% of the computation time for small systems (<500 orbitals), depending on the type of XC functional used (LDA/GGA/VDW). Becomes less than 5% for systems larger than 10.000 orbitals.
Diagon	Diagonalization of the Hamiltonian	Compute-bound when using optimised libraries.	For smaller systems (less than 1.000 orbitals) it may take about 20% of the total time, but it easily reaches more than 90% of the total time with increasing system sizes (more than 20.000 orbitals)

Table 3.4. Important kernels and performance characteristics of Siesta.

3.1.5 Fleur

Depending on the properties/workflow calculated, FLEUR can show very different computational characteristics. In a standard DFT calculation, the diagonalization of the Hamiltonian is the computationally most relevant kernel followed by the setup of the matrices. These parts scale cubically with the system size. In the case of non-collinear simulations with spin-orbit coupling also the setup of the spin-off diagonal part of the Hamiltonian can become important. For simulations using hybrid functionals the construction of the non-local potential is by far the most computationally expensive operation. It scales with the fourth power of the system size and consists of several independent kernels. As listed in Table 3.5, Fleur is formed of kernels with different characteristics, but the code can be considered memory bound.

Kernel	**functionality**	**characteristic**	**Percentage of WCT**

diagonalization	Diagonalization of Hamiltonian	Handled by external libraries. Generalised eigenvalue solver. Often memory bound, for larger distributed systems communication bound	Typically 40-60% in standard DFT calculations
hsmt_nosph	Setup of non-spherical contributions to Hamiltonian	zherk/zgemm calls, mostly compute bound (non-square matrices)	Typically 10-30% in standard DFT calculations
hsmt_sph	Setup of spherical contributions to hamiltonian and overlaps	Self-written kernel. Memory bound	Typically 5-20% in standard DFT calculations
wavefproducts_INT	Projection of products of wavefunctions onto Mixed-Product basis (INT part) for hybrid functionals	Many 3D FFTs (relatively small). Usually memory bound	40-60% of hybrid functional calculations
wavefproducts_MT	Projection of products of wavefunctions onto Mixed-Product basis (MT part) for hybrid functionals	Self-written kernel. Memory bound	30-50% of hybrid functional calculations
sparse_matmul	Product of projections with Coulomb matrix for non-local hybrid functional potential	Combination of zgemm and self-written kernel.	About 10% of hybrid functional calculations

Table 3.5. Important kernels and performance characteristics of Fleur.

3.2 Currently available and planned resources

In this section, available or planned hardware resources to be used by MAX are documented. A list of platforms of interest and reasons for that is provided at the end of the section.

3.2.1 qIT4I@VSB

For the purpose of the evaluation of advanced hardware platforms, IT4I@VSB provides its Complementary systems. In addition, for work to be done in Task 4.3, it contributes by its production systems, Karolina and Barbora, which enables energy efficiency evaluation of MAX3 applications at scale.

Production systems used for energy efficiency evaluations

Karolina (EuroHPC petascale system)

CPU Partition based on AMD Zen2

- 720 nodes of 2x AMD 7h12 (64 cores)
- 256 GB RAM per node
- CPU HW parameters tuning enabled and power and energy monitoring in place

GPU Partition based on AMD Zen3 CPUs + NVIDIA A100 GPUs

- 72 nodes of 2x AMD 7763 (64 cores) + 8x NVIDIA A100
- 1024 GB RAM per node
- CPU and GPU HW parameters tuning enabled and power and energy monitoring in place

Barbora

A preferred system for energy-efficiency analysis and tuning due to HDEEM power monitoring system.

CPU partition based on Intel Cascade Lake

- 192 nodes of 2x Intel Cascade Lake 6240 (18 cores)
- 192GB RAM per node
- Atos HDEEM power monitoring system
- CPU HW parameters tuning enabled

Co-design, Advanced hardware platforms

Complementary system 1

Partition 1 - ARM (A64FX)

- 8 compute nodes with 1x Fujitsu A64FX CPU with 32 GB of HBM2 memory

Partition 2 - Intel (Ice Lake, NVDIMMs)

- 2x 3rd Gen Intel Xeon Gold 6338 CPU
- 8448 GB NVDIMM memory (16x 512GB NVDIMM persistent memory modules)
- 2x Bittware 520N-MX FPGA cards

Partition 3 - AMD (Milan, MI100 GPUs + Xilinx FPGAs)

- 2x AMD Milan 7513 CPU - 32 cores @ 2.6 GHz
- 4x AMD GPU accelerators MI 100 - interconnected with AMD Infinity Fabric™ Link for fast GPU to GPU communication
- server 1 has 2x FPGA Xilinx Alveo U250 Data Center Accelerator Card
- server 2 has 2x FPGA Xilinx Alveo U280 Data Center Accelerator Card

Partition 4 - Edge Server

- 1x x86_64 CPU Intel Xeon D-1587
- 1x CUDA programmable GPU NVIDIA Tesla T4

Partition 5 - FPGA Synthesis Server

- AMD EPYC 72F3, 8 cores @ 3.7 GHz nominal frequency

The energy and power monitoring system contains PDUs with outlet-level power monitoring capabilities. The diagram of the system is shown in Figure 3.1.

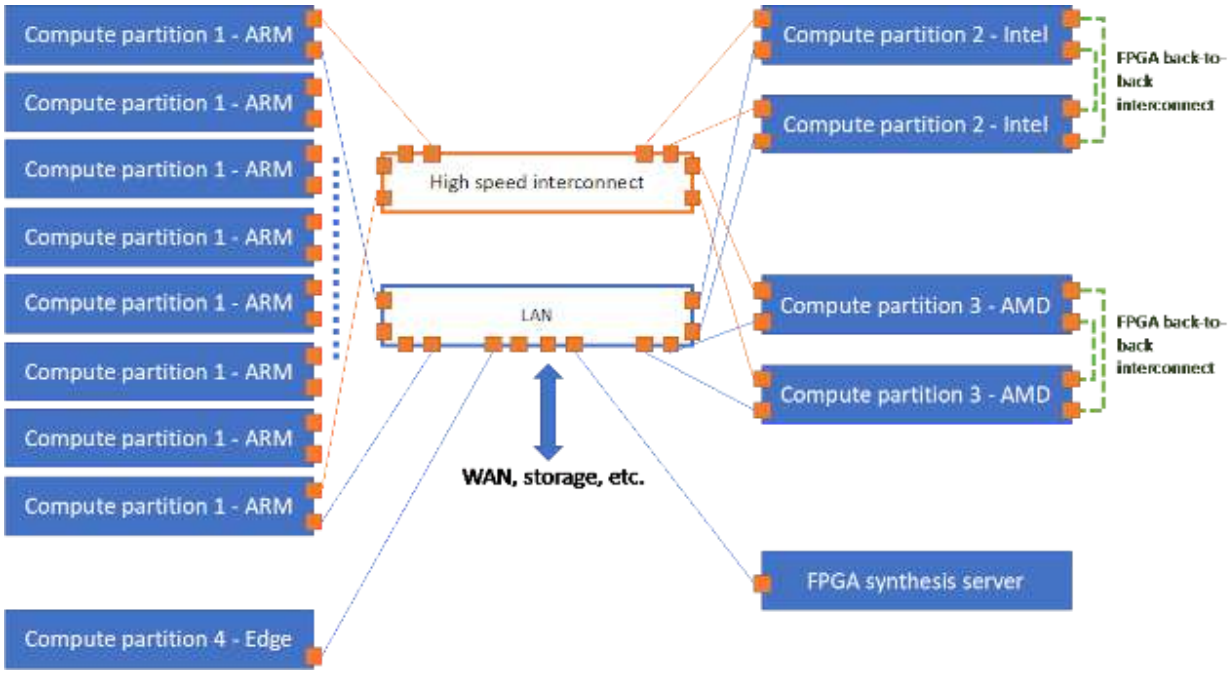

Fig. 3.1. Block diagram of the Complementary systems 1 at IT4I@VSB.

Complementary system 2

The system consists of the following logical components:

Partition 1 (ARM + Nvidia GPU + DPU)

- 1x Ampere Altra Q80-30 (80c, 3.0GHz)
- 512 GB RAM
- 2x Nvidia A30 GPU
- 2x NVIDIA BlueField-2 DPU

Partition 2 (Power architecture)

- 2x Power 10 CPU
- 512 GB of DDR4 memory - min. 350GB/s memory bandwidth per socket
- IBM community editions: IBM XL C/C++&Fortran/ESSL CE

Partition 3 (high-capacity L3 cache processor)

- 2x AMD Milan X

    - 64 cores
    - 750 MB L3 cache
- 256 GB RAM

Partition 4 (virtual GPU accelerated workstations)

- 2x AMD CPUs - 24 cores
- 512 GB of RAM
- 2x NVIDIA A40

Partition 5 (Intel SPR + HBM)

- 2x Intel Xeon CPU Max 9468
- 48 cores, 2 h-threads (currently active)
- 250 GB of RAM

The energy and power monitoring system contains PDUs with outlet-level power monitoring capabilities. The diagram of the system is shown in Figure 3.2.

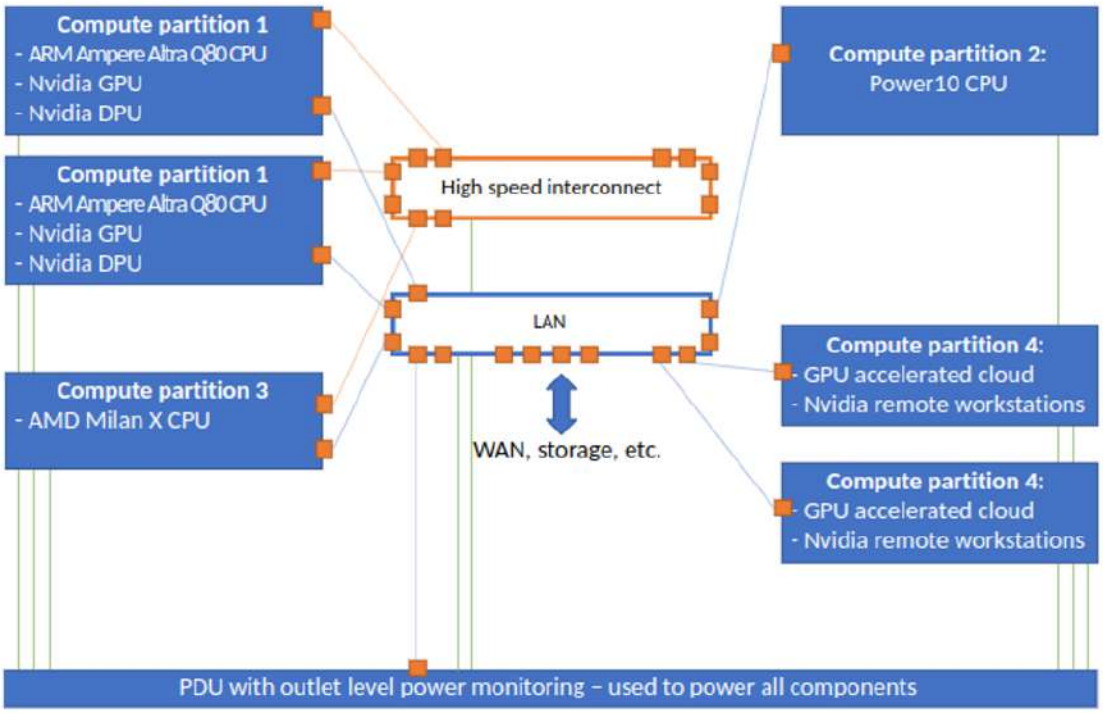

Fig. 3.2. Block diagram of the Complementary systems 2 at IT4I@VSB.

3.2.2 EUPEX Alfa Pilot Platform (EAP)

The European pilot for exascale (EUPEX) project plans to start an EUPEX Alfa Pilot Platform (EAP) to provide a development vehicle supporting Centers of Excellences and other European projects. The EAP will be a A64FX machine operated by CEA, which is a hardware sharing two key aspects with the future European processor SiPearl Rhea1: SVE instruction set and HBM memory.

- At this moment unknown amount of compute nodes with 1x Fujitsu A64FX CPU with 32 GB of HBM2 memory
- EUPEX software stack

3.2.3 E4

E4 provides 5 different partitions with different processors and accelerators.

Partition 1 – Intel (Xeon, GPU)

- 2 nodes are dual socket, 16 cores per socket with Intel Xeon Gold 6426Y processor
- 508 GB RAM
- On two nodes are, there one NVIDIA A2 GPU each

Partition 2 – AMD (Epyc)

- supermicro twinsquare: https://www.supermicro.com/en/Aplus/system/2U/2124/AS-2124BT-HNTR.cfm (8 nodes)
- each node is a dual-socket with 2x AMD EPYC 7313 16-Core Processor
- 256 GB RAM

Partition 3 – AMD (Epyc, Mi100 GPUs)

- 2 nodes dual socket with 2x AMD EPYC 7313 16-Core Processor
- 256 GB RAM
- 2x AMD Mi100 GPU each one

Partition 4 – Intel (Xeon)

- supermicro twinsquare: https://www.supermicro.com/en/products/system/2U/2029/SYS-2029BT-HNTR.cfm (4 nodes)
- each node is a dual socket Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz, 16-Core Processor
- 192 GB RAM

Partition 5 – ARM (Neoverse, GPUs)

- 2 nodes with ARM Neoverse-N1 80-Core Processor
- 1x NVIDIA V100 GPU each
- 500 GB RAM

3.2.4 Eviden / Atos

Seven different hardware configurations are available on the Spartan machine provided by Eviden.

AMD CPU partition

- ~160 nodes
- 2x AMD EPYC™ Milan 7763, 64-cores (https://www.amd.com/en/products/cpu/amd-epyc-7763)
- 256 GB RAM, DDR4 3200 MT/s
- HDR IB interconnect

AMD CPU partition

- ~20 nodes
- 2x AMD EPYC™ Genoa 9654, 96-cores (https://www.amd.com/en/products/cpu/amd-epyc-9654)
- 384/768 GB RAM, DDR5 4800 MT/s
- EDR IB interconnect

FUJITSU CPU partition

- 3 nodes

- 1x ARM A64FX 48-cores (FX700) (https://www.fujitsu.com/global/products/computing/servers/supercomputer/a64fx/)
- 32 GB RAM, HBM2
- EDR IB interconnect

INTEL CPU partition

- ~80 nodes
- 2x Intel® Xeon® Platinum 8480, 56-cores (https://www.intel.com/content/www/us/en/products/sku/231746/intel-xeon-platinum-8480-processor-105m-cache-2-00-ghz/specifications.html)
- 512 GB RAM, DDR5 4800 MT/s
- HDR IB interconnect

INTEL CPU partition

- < 5 node
- 2x Intel® Xeon® Platinum 9480, 56-cores (https://www.intel.com/content/www/us/en/products/sku/232592/intel-xeon-cpu-max-9480-processor-112-5m-cache-1-90-ghz/specifications.html)
- neover
- EDR IB interconnect

NVIDIA GPU partition

- 10 nodes
- 2x AMD EPYC™ Milan 7763, 64-cores (https://www.amd.com/en/products/cpu/amd-epyc-7763)
- 512 GB RAM, DDR4 3200 MT/s
- 4x Nvidia A100 40 GB (https://www.nvidia.com/en-us/data-center/a100/)
- HDR IB interconnect

AMD GPU partition

- <10 nodes
- 2x AMD EPYC™ Milan 7763, 64-cores (https://www.amd.com/en/products/cpu/amd-epyc-7763)
- 256 GB RAM, DDR4 3200 MT/s

- 4x AMD MI250 (8 GCD) 64 GB (https://www.amd.com/en/products/server-accelerators/instinct-mi250)
- EDR IB interconnect

ARM CPU partition

- 1 node
- 1x Ampere Altra Max Q8030, 80 cores
- 256 GB RAM, DDR4
- EDR IB interconnect

3.2.5 FZJ

Forschungszentrum Jülich provides production supercomputers as well as evaluation platform for MAX partners.

Production Supercomputers

JUWELS Cluster:

- 2271 standard compute nodes
  - 2× Intel Xeon Platinum 8168 CPU, 2× 24 cores, 2.7 GHz
  - 96 (12× 8) GB DDR4, 2666 MHz
  - InfiniBand EDR (Connect-X4)
- 240 large memory compute nodes
  - 2x Intel Xeon Platinum 8168 CPU, 2× 24 cores, 2.7 GHz
  - 192 (12× 16) GB DDR4, 2666 MHz
  - InfiniBand EDR (Connect-X4)
- 56 accelerated compute nodes
  - 2× Intel Xeon Gold 6148 CPU, 2× 20 cores, 2.4 GHz
  - 192 (12× 16) GB DDR4, 2666MHz
  - 2× InfiniBand EDR (Connect-X4)
  - 4× NVIDIA V100 GPU, 16 GB HBM
- Diskless

- Mellanox InfiniBand EDR fat-tree network with 2:1 pruning at leaf level and top-level HDR switches
- 250 GB/s network connection to JUST for storage access

JUWELS Booster:

- 936 compute nodes
  - 2× AMD EPYC Rome 7402 CPU, 2× 24 cores, 2.8 GHz
  - 512 GB DDR4, 3200 MHz
  - 4× NVIDIA A100 GPU, 40 GB HBM2e
  - 4× InfiniBand HDR (Connect-X6)
- Diskless
- Mellanox InfiniBand HDR DragonFly+ topology with 20 cells - 5 TB/s connection to Cluster
- 700 GB/s network connection to JUST for storage access

JUSUF:

- 144 standard compute nodes
  - 2× AMD EPYC 7742, 2× 64 cores, 2.25 GHz
  - 256 (16× 16) GB DDR4, 3200 MHz
  - InfiniBand HDR100 (Connect-X6)
  - local disk for operating system (1× 240 GB SSD)
  - 1 TB NVMe
- 61 accelerated compute nodes
  - 2× AMD EPYC 7742, 2× 64 cores, 2.25 GHz
  - 256 (16× 16) GB DDR4, 3200 MHz
  - InfiniBand HDR100 (Connect-X6)
  - local disk for operating system (1× 240 GB SSD)
  - 1 TB NVMe
  - 1× NVIDIA V100 GPU with 16 GB HBM2e
- Mellanox InfiniBand HDR full fat-tree network with HDR100 speed on the nodes and full HDR on the inter-switch level

- 100 Gbit/s network connection per login node and 40 Gbit/s network connection per compute node to JUST for storage access

JURECA Evaluation Platform

AMD MI200 nodes:

- 2 compute nodes
  - CPU: AMD EPYC 7443 processor (Milan); 2 sockets, 24 cores per socket, SMT-2 (total: 2×24×2 = 96 threads) in NPS-4 1 configuration (details for AMD EPYC 7443 on WikiChip)
  - Memory: 512 GiB DDR4-3200 RAM (of which at least 20 GB is taken by the system software stack, including the file system); 256 GB per socket; 8 memory channels per socket (2 channels per NUMA domain)
  - GPU: 4 × AMD MI250 GPUs, each with 128 GB memory; the GPUs are built as Multi Chip Modules (MCM) and because of that they are shown as 8 GPUs with 64 GB memory each.
  - Network: 1 × Mellanox HDR InfiniBand ConnectX 6 (100 Gbit/s), HCA (not yet final)
  - Details about the hardware can be found on Gigabyte's webpage.
  - Details about the node topology can be found in AMD's CDNA2 whitepaper.

Graphcore IPU-POD4:

The IPU-POD4 consists of two parts:

- an AMD EPYC based access server on which user applications are launched with
  - CPU: AMD EPYC 7413 (Milan); 2 sockets, 24 cores per socket, SMT-2 (total: 2×24×2 = 96 threads) in NPS-4 1 configuration (details for AMD EPYC 7413 on WikiChip)
  - Memory: 512 GiB DDR4-3200 RAM (of which at least 20 GB is taken by the system software stack, including the file system); 256 GB per socket; 8 memory channels per socket (2 channels per NUMA domain)

    - Network: 1 × Mellanox EDR InfiniBand ConnectX 5 (100 Gbit/s) to connect to other compute nodes and 1 × Mellanox 100 GigE ConnectX 5 to connect to the IPU-M2000
- a Graphcore IPU-M2000 which is connected directly to the access server with
    - IPUs: 4 × GC200 IPUs
        - More information from Graphcore

NVIDIA Arm HPC Dev Kit:

2x NVIDIA Arm HPC Dev Kits, each consisting of:

- an Ampere Altra Q80-30 CPU with 80 cores and 512 GB memory,
- 2 NVIDIA A100-PCIe-40-GB GPUs,
- 2 NVIDIA Mellanox BlueField2 DPUs (200 GbE)

More details on NVIDIA's documentation about the Dev Kit.

Intel Sapphire Rapids HBM:

- QuantaGrid D54Q-2U
- 2x Intel (R) Xeon (R) CPU Max 9480 (Intel ark)
- 2x 64 GB HBM
- 16x 64 GB DDR5 4800 MT/s
- 4480KiB L1 cache
- 112MiB L2 cache
- 112MiB L3 cache
- 960GB INTEL SSDSC2KG96 SSD
- 1x BlueField-2 ConnectX-6 DPU @ EDR (100 Gbit/s)

Intel Sapphire Rapids + NVIDIA H100:

- 2 x Intel Xeon 'Sapphire Rapid' 8452Y Processor (x86, 64 bit): 36 Cores, 2,0 GHz, 67,5 MB L3 Cache, 300 Watt TDP
- 4 x NVIDIA H100 80 GB GPU

- 512 GB (16 x 32 GB) DDR5 RDIMM 4800 MHz ECC
- 2 x 960 GB NVMe PCIe4x4 U.2 (2,5'') 1 DWPD
- 2 x RJ-45 10 Gbase-T LAN Port via Intel X710-AT2 Controller
- 1 x RJ-45 Dedicated IPMI LAN Port
- 1x BlueField-2 ConnectX-6 DPU @ EDR (100 Gbit/s)

Jupiter evaluation platform:

- single GraceHopper superchip server
- 1x NVIDIA Grace™ with 72 Arm® Neoverse V2 cores
- 480GB LPDDRX embedded
- 96GB HBM3 GPU memory
- Internal Interconnect - NVIDIA® NVLink®-C2C 900GB/s

JuMax

The system has been developed and deployed as part of the PRACE PCP. It is a high-performance computing system that consists of an AMD EPYC CPU server and a Maxeler MPC-X dataflow node with 8 MAX5 Dataflow Engines (DFEs), see Figure 3.3.

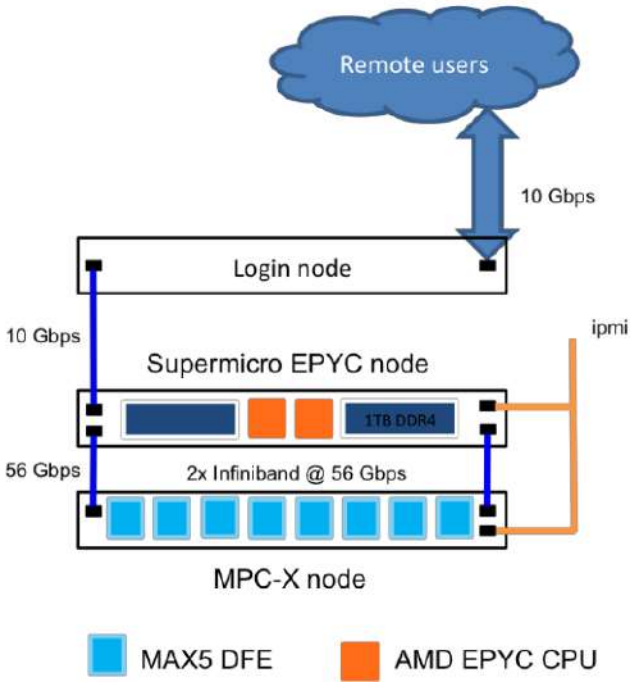

Fig. 3.3. JuMax system with JuMax-cpu servers and a Maxeler MPC-X dataflow node with 8 MAX5 Dataflow Engines (DFEs).

The high-performance compute node (JuMax-cpu) serves as the main host system for application processing and directly invokes DFEs via an InfiniBand connection to the MPC-X node. The machine uses two top-of-the-line CPUs from the AMP EPYC server-grade processor line. This is combined with a total of 1 TB of DDR4 memory.

The machine configuration is:

- CPU: 2 x AMD EPYC 7601 (32 cores / 64 threads per CPU @ 2.2 GHz)
- RAM: 1 TB DDR4 @ 2666 MHz
- Disks:
    - 2 x 240 GB Intel SSD (RAID1, 240 GB usable, used for OS)
    - 8 x 1.2 TB SSD Intel (RAID0, 8.7 TB usable, used for local data storage)

The Maxeler MPC-X 3000 node:

- integrates 8 FPGA-based MAX5 DFEs into a dense 1U dataflow system

- MAX5 DFEs are PCIe accelerator cards that combine a Xilinx UltraScale+ VU9P FPGA with 48 GB of DDR4 ECC memory

3.2.6 CINECA

Cineca mainly provides the production system platforms, however, they have several interesting features like NVDIMM memory on a large number of nodes (Galileo 100) or a specialised version of the Nvidia A100 GPU (Leonardo).

Galileo 100 (CPU)

- Nodes: 636 (+10 login nodes,+5 service nodes)
  Processors: 2x CPU x86 Intel Xeon Platinum 8276-8276L (24c, 2.4Ghz)
- Cores: 48 cores/node
- Accelerators: 2x GPU nVidia V100 PCIe3 with 32GB RAM on 36 Viz nodes
- RAM: 384GB (+ 3.0TB Optane on 180 fat nodes)
  Internal Network: Mellanox Infiniband 100GbE
- Interconnect: 100Gbit/s Infiniband network

Out of all compute nodes the most interesting for WP4 are:

- 180 data processing nodes ("fat nodes") 2TB SSD with **3TB Intel Optane**

NVidia DGX

- Model **:** DGX-A100
- Nodes**:** 3
- Processors: Dual AMD Rome 7742 CPU @ 2.25-3.4 GHz (128 cores)
- Accelerators: 8 x NVIDIA A100 GPUs per node (40GB) , NVlink 3.0
- RAM: 1TB/node
- Internal Network**:** Mellanox IB EDR fully connected topology
- Storage**:** 15 TB/node NVMe, 95 TB Gluster storage

Leonardo - Booster (GPU)

- Nodes: 3456 nodes (currently 512 is maximum)
- Processors: 1x Intel(R) Xeon(R) Platinum 8358 CPU @ 2.60GHz (32 cores)
- Accelerators: 4 x NVIDIA Ampere GPUs/node (64GB HBM2)
- Cores: 32 cores/node
- RAM: 512 GB DDR4 3200 MHz
- Network: NVIDIA Mellanox HDR DragonFly++ 200Gb/s
  2 x NVIDIA HDR 2×100 Gb/s cards GPU nodes
- Storage: 106PB Large capacity storage
  5.4 PB of High-performance storage

3.2.7 SiPearl

Also available internally:

- 2 nodes with 2x80 cores Ampere Altra 3.3GHz with 256GB RAM
- AWS Graviton 3/3e accesses,
- multiple Intel Xeon Gold 6254 (2x18cores HT @3.1GHz),
- AMD EPYC 74F3 (2x24cores HT @3.2GHz) nodes,
- future Rhea platforms when available

3.2.8 Summary of the selected novel resources for MaX codes

Table 3.6 summarises currently available or planned resources provided by MAX partners to the consortium. The list [3] is going to be actively updated if new technologies become available during the MAX project.

Partner	Machine	Advance and/or co-design HW	Reason
IT4I	Complementary systems	**CPU**: ● Fujitsu A64FX ● IBM POWER 10 ● Intel® Xeon® Max 9468	Memory/compute-bound kernels
		ARM CPU + GPU + DPU combination: ● ARM Ampere Altra Q80-30 CPU	Memory/compute-bound kernels

		• NVidia A30 GPU • NVidia BlueField 2 DPU	IO
		GPU + FPGA combination: • 4x AMD MI100 GPUs • 2x Xilinx Alveo U250 or U280 FPGAs	Latency bound kernels
		FPGA accelerators: • Bittware 520N-MX with Intel Stratix 10 MX FPGA with HBM2 memory • Xilinx Alveo U280 with HBM2 memory • Xilinx Alveo U250 with DDR4 memory	Latency bound kernels
		Memory technologies: • NVDIMM memory (Intel® Optane™ Persistent Memory) • HBM on A64FX CPU • HBM on Intel® Xeon® Max 9468 CPU • Large L3 on AMD EPYC 7773X Milan-X CPU	Kernels with large memory requirements Memory bound kernels
E4		**GPU accelerators**: • NVidia V100 • AMD MI100 **CPUs**: • ARM Neoverse-N1 80-Core Processor • Intel Sapphire Rapids • AMD Milan	Memory/compute bound kernels
Eviden / Atos	Spartan	**CPU**: • AMD Genoa (EPYC 9654) • A64FX • Intel	Memory/compute bound kernels
		Memory technologies: • DDR5 4800 MT/s on AMD Genoa • HBM2 (on A64FX)	Memory bound kernels
		GPU accelerators: • AMD MI250, • NVidia A100	Memory/compute bound kernels

FZJ	JuMax	**FPGA accelerators**: ● Maxeler MPC-X 3000	Latency bound kernels Dataflow friendly kernels
FZJ	JURECA DC (incl. Evaluation Platform)	Different HW partitions: **GPU accelerators**: ● AMD MI200, ● NVidia A100, ● NVidia H100	Memory/compute bound kernels
		CPUs: ● Intel Sapphire Rapids (Xeon 8524Y) ● ARM (Ampere Altra Q80-30) ● AMD (EPYC 7443)	Memory/compute bound kernels
		Intelligence Processing Unit (IPU): ● Graphcore IPU-POD4	AI/ML kernels
FZJ	JUPITER Cluster and Booster	**NVIDIA NVIDIA GH200** ● **CPU:** NVIDIA Grace CPU ● **GPU:** NVIDIA H200 ● NVLink-C2C: 900GB/s of coherent memory ● Up to 480GB of LPDDR5X	Memory/compute bound kernels kernels with host device bottlenecks
		Jupiter evaluation platform: ● single GraceHopper superchip server	Memory/compute bound kernels
		CPU Cluster ● SiPearl Rhea processor	Memory/compute bound kernels
CINECA	Galileo 100, Leonardo and Leonardo - Booster	**Memory technologies**: ● NVDIMM	Kernels with large memory requirements
		GPU accelerators: ● NVidia A100 (version with 64GB HBM2) ● Intel Xe-HPC Data Center GPU Max 1550	Memory/compute bound kernels
BSC	MareNostrum5	**CPUs**: ● Intel Sapphire Rapids, ● NVidia Grace	Memory/compute bound kernels

		GPU accelerators: ● NVidia H100	Memory/compute bound kernels
		CPU + GPU combination ● Intel Sapphire Rapids + H100	Memory/compute bound kernels
BSC	Software Development Vehicles (SDV) for RISC-V	Hardware and software tools for enabling long vector computation on RISC-V. For more information see RISC-V Vector Environment.	Memory/compute bound kernels
SiPearl	ARM	**CPU**: ● ARM AWS Graviton 3/3e (remote access)	Memory/compute-bound kernels
EUPEX	Alfa Pilot platform (EAP)	**CPU**: ● Fujitsu A64FX	EPI/EUPEX SW stack Memory bound kernels

Table 3.6. Summary table of the co-design and advanced platforms considered for WP4.

3.3 Platforms for energy efficiency evaluations

Table 3.7 lists currently available resources provided by MAX partners with the ability to measure energy consumption of the compute components or the whole blade. Moreover some of these platforms allow users to control hardware specific parameters with impact on power consumption. Some of these platforms will be selected for evaluation of the MAX codes energy efficiency, which will be reported in the D4.2 Report on energy consumption evaluation (delivery M18).

Partner	**Machine**	**Power monitoring**	**HW tuning**
IT4I	Barbora - CPU partition	HDEEM + Intel RAPL	Intel CPU
IT4I	Karolina - CPU partition Karolina - GPU partition	AMD RAPL AMD RAPL + NVML	AMD CPU AMD CPU + NVidia GPU
IT4I	Complementary systems	PDU Intel RAPL AMD RAPL	Intel CPU, AMD CPU,

		IBM OCC A64FX performance counters	IBM CPU, N/A
CINECA	G100 - CPU partition	Intel RAPL	Intel CPU
CINECA	Marconi	Intel RAPL	Intel CPU
CINECA	Leonardo - CPU Partition (under deployment) Leonardo - GPU Partition (under deployment)	Intel RAPL Intel RAPL + NVML	Intel CPU Intel CPU + NVidia GPU
IJS	Vega - CPU Partition Vega - GPU Partition	AMD RAPL, IPMI	AMD CPU AMD CPU
E4	All partitions	PDU	IntelCPU, AMD CPU, ARM CPU
CEA (EUPEX)	Irene A64FX	A64FX performance counters	N/A
FZJ	JURECA DC (Evaluation Platform partition)	Intel RAPL + NVML	Intel CPU + NVidia GPU

Table 3.7. Summary table of the platforms considered for WP4 for energy efficiency evaluation.

Intel, AMD, and IBM CPUs allow to specify power limit and target core frequency (respected if neither power nor thermal violated), moreover Intel CPUs provide interface to control boundaries of frequency of on-chip off-core subsystem (so-called uncore), e.g. L3 cache, or cores interconnect. ARM CPUs power management capabilities are vendor-specific, but in general one may expect that at least core frequency should be possible to control.
Similarly, NVIDIA and AMD hi-end server GPUs have an interface to specify the frequency of streaming multiprocessors, and power limit.

4. Strategies to access technologies and restrictions

MAX partners provide access to prototypes containing advanced technologies as described in the previous section. This section provides information on how MAX3 partners can get access to the resources.

4.1 IT4I

The access to the IT4I resources is provided using a multi-year project. The MAX3 project member access is organised by IT4I, and as of June 2023, it is valid for 3 years. Each MAX partner can have access to the complementary systems.

In addition to access to the complementary systems, the project also provides access and resources to the CPU partition of the Barbora system[1] used mainly in T4.3 for energy efficiency evaluation of the MAX3 codes, which contains a HDEEM power and energy monitoring system. The project also provides access to the CPU and GPU partitions of the Karolina machine[2]. This system is also in the scope of WP4 used for energy efficiency evaluation on the AMD (Rome and Milan) platforms and NVidia A100 GPUs. The resources are also provided to WP3 for their related activities.

Key information: To apply for an account, follow the procedure described in the IT4I cluster documentation[3] and use project **OPEN-28-69**.

Restrictions: Barbora, Karolina and complementary systems are open to all users of the IT4I infrastructure. There are no access restrictions for MAX3 partners. Any results published using the resources must include the following acknowledgment *"This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPS II) project "IT4Innovations excellence in science - LQ1602" and by the IT4Innovations infrastructure which is supported from the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254)."*

Contacts: Lubomir Riha <lubomir.riha@vsb.cz> and Ondrej Vysocky <ondrej.vysocky@vsb.cz>

4.2 E4

The access to the E4 resources is provided through a VPN access to E4 premises. In order to get access, contact E4.

[1] IT4Innovations Barbora machine: https://docs.it4i.cz/barbora/hardware-overview/
[2] IT4Innovations Karolina machine: https://docs.it4i.cz/karolina/hardware-overview/
[3] IT4Innovations documentation - Account creation process: https://docs.it4i.cz/general/obtaining-login-credentials/obtaining-login-credentials/

Key information: Submit an account request for the cluster via the E4 customer portal[4]

Restrictions: All systems are open to users of the MAX3 project, there are no access restrictions.

Contacts: Daniele Gregori <daniele.gregori@e4company.com>

4.3 Eviden / Atos

All Eviden systems are restricted to their internal users. Exceptions are possible, to ask for an exception, please contact Eviden.

Key information: Access provided on peer-to-peer cases. Contact Eviden.

Restrictions: To be clarified per case.

Contacts: Erwan RAFFIN <erwan.raffin@eviden.com>

4.4 FZJ

To access FZJ resources in general, one needs to apply for a computing time allocation or ask to join an existing project. MAX3 partners can apply for a test project anytime. For Larger compute time projects on production machines, one can apply twice a year.

In the scope of WP4, MAX3 partners can use the Exalab project for development and benchmarking. The Exalab project is not intended for production runs, as resources are limited. To request access to the Exalab project, please contact FZJ.

To connect to JSC machines, you can use SSH or Jupyter-JSC portal.
As mentioned before, for the Intel Sapphire Rapids HBM node, partners need approval from Intel for access.

[4] E4 customer portal: https://service.e4company.com/servicedesk/customer/portals

Key information: To connect to JSC machines, you can use SSH or Jupyter-JSC portal[5]. Use the **Exalab** project to get access to the resources/

Restrictions: Publishing of results obtained on some machines needs approval before the publication. For details and approval, contact FZJ.

Contacts: Andreas Herten <a.herten@fz-juelich.de> and Kaveh Haghighi Mood <k.haghighi.mood@fz-juelich.de>

4.5 CINECA

In order to get a personal username for the HPC systems in CINECA, you have to register to CINECA UserDB portal[6]. Moreover, you need to be associated with an active project, as a *Collaborator* or as a *Principal Investigator* (PI).
MAX3 partners can use the Max3_devel project for development. This project is not intended only for production runs, as resources are limited, but only for development and benchmarking runs.
CINECA clusters can be reached with SSH (Secure Shell) protocol. It is mandatory to use two-factor authentication (2FA).

Key information: New accounts must register at CINECA UserDB portal[7]. Use **Max3_devel** project to get resources.

Restrictions: No particular restrictions for MAX project members.

Contacts: Fabio Affinito <f.affinito@cineca.it>

4.6 BSC

BSC will offer competitive access to MareNostrum5 through the periodic calls of EuroHPC and the Spanish Supercomputing Network (Red Española de Supercomputación, RES) open for Spanish and European researchers.

[5] How to access JUWELS using SSH: https://apps.fz-juelich.de/jsc/hps/juwels/access.html#ssh-login
[6] CINECA UserDB portal: https://userdb.hpc.cineca.it/
[7] CINECA UserDB portal: https://userdb.hpc.cineca.it/

For RISC-V Vector Environment the access needs to be discussed directly with BSC. The software development vehicles (SDV) include a software emulator, RISC-V scalar mini-clusters and FPGA prototype of RISC-V long vector architecture. Direct access is provided upon collaboration agreement. For more information see RISC-V Vector Environment[8] description.

Key information: The new pre-exascale MareNostrum5 is currently being installed at the BSC premises. The general purpose partition (GPP) is expected to be available by the end of 2023 while the accelerated one (ACC) will start production in 2024. Access to MAX partners will be granted following a procedure similar to other EuroHPC machines.

Restrictions: All MareNostrum partitions will be open to users without any access or usage restrictions.

Contact: Julio Gutiérrez <julio.gutierrez@bsc.es>

4.7 EUPEX Alfa Pilot platform (EAP)

The EAP at this stage is designed to provide its users access to the EUPEX software stack, so the users may get familiar with the tools that will be available for the Rhea when the CPU will be available.

Key information: EUPEX EAP is expected to start operation in Q1 2024. Access to the system is granted upon acknowledged request to the EUPEX consortium.

Restrictions:
Exact list of conditions for EAP users will be provided at the beginning of the system start. As of now expected requirements consists of:

- Evaluate EUPEX software stack.
- Fill in a user-experience survey.
- Acknowledge the EUPEX EAP in any publication or presentation using results obtained with the support of the EUPEX EAP.

Contacts: <eap@eupex.eu>

[8] BSC RISC-V Vector Env.: https://repo.hca.bsc.es/gitlab/epi-public/risc-v-vector-simulation-environment

4.8 SiPearl

Access to SiPearl platforms is to their employees only.

Key information: No access for non-SiPearl personnel, but runs can be performed by SiPearl if necessary.

Restrictions: To be clarified per case.

Contacts: Augustin Degomme <augustin.degomme@sipearl.com>

4.9 Summary

Table 4.1 provides a summary of the essential information about the accesses to the advanced platforms identified under the WP4. We can conclude that, within the MAX3 consortium, we can offer both the code developers and HPC experts access to the cutting-edge HW platform enabling the exploitation of the new hardware technologies for MAX codes.

Partner	Access	Restrictions
IT4I	Available as of 1.6.2023 - OPEN-28-69 access project	None for MAX3 partners
E4	Request VPN access to E4	None for MAX3 partners
Eviden	Currently available for Eviden employees, for others case by case agreements	case by case agreements
FZJ	Available using test accounts and the Exalab project	Some resources need approval before publications
CINECA	Available for users associated to a valid project	None for MAX3 partners
BSC	Direct access to MareNostrum through EuroHPC calls. Access to RISC-V platforms upon collaboration agreement	None for MAX3 partners
EUPEX	Expected start Q1 2024, access granted upon acknowledged request.	Providing feedback to EUPEX project

SiPearl	Available for Sipearl employees only.	N/A

Table 4.1. Summary of the access mechanisms and restrictions for accessing the advanced platforms under MAX3.

5. The technology roadmaps

One of the aims of WP4 is to gather information on the near future HPC technologies, such that MAX applications will be ready to fully exploit the next-generation hardware. In this section, we are going to describe the publicly available roadmaps of the main actors in the HPC market, exploring different architectures and devices.

5.1 AMD

CPUs

After the successful release of CPUs based on the Zen 4 microarchitecture between 2022 and 2023, which allowed AMD to gain a good portion of the server CPU market, the company is planning to launch the next generation CPUs based on the Zen 5 microarchitecture towards the end of 2024 [4]. This will be part of the EPYC series of processors and will be commercialised with the name Turin. Three variants of the processor will be developed: Turin based on Zen 5, Turin-X based on Zen 5 with 3D Vertical cache and Turin dense based on Zen 5c, which will be specific for the cloud.

Zen 5 is expected to improve performance and energy consumption with respect to Zen 4. In particular, it will feature from 10% to 15% more instructions per cycle than Zen 4 and a significant performance/Watt increase. Zen 5 L1 cache will be 48 kBytes (16 kBytes more than Zen 4 L1 cache). The size of Zen 5 L2 and L3 caches will remain as in Zen 4, but is expected to be larger in the case of Zen 5 with 3D Vertical cache (Turin-X processor). An important feature of Zen 5 is that it will support FP-512 capabilities [5, 6].

Turin will probably be released with 128 cores per node and each physical processor will support two logic units. The expected thermal design power (TDP) will be 500 W. Turin dense will have even more cores per node (up to 192 cores/node) [6]. The successor of Turin will be

launched in 2025 and will be based on the Zen 6 architecture, whose details are not known at the moment [6].

Accelerators

After more than one year from the release of Instinct MI2X0 (March 2022), AMD will launch the next-generation Accelerated Processing Unit (APU) this year. The product will be called Instinct MI300.

The MI300 series will be commercialised with two models: MI300A and MI300X. MI300A will combine CPU and GPU cores on the same chip package with high-bandwidth memory (HBM) for AI and HPC workloads. It will feature 24 Zen 4 CPU cores, a CDNA3 graphic engine, 128 GB of HBM shared by CPUs and GPU and Compute eXpress Link (CXL) interconnections [7, 8]. The presence of a unified pool of memory avoids redundant memory copies and will speed up performance. In particular, it is expected that this, together with the new supported maths format, will allow a 5x performance per watt increase compared to the previous accelerator generation [9]. MI300X will be a GPU-only product featuring 192 GB of HBM memory. It will be specially suited for AI and able to run large language models up to 80 billion parameters [8]. The MI300 series is part of AMD's strategy to expand its leadership in the data centre and AI markets and to directly compete with the NVIDIA Grace Hopper superchip.

The MI300 series will be followed by the MI400 series. However, its exact specifications and release date are not yet known. It will probably be an APU with Zen 5 CPUs and CDNA4 GPUs [10, 11].

AMD is also active on the FPGA landscape. The company is planning to launch a product called XDNA that integrates Xilinx AI engine and FPGA fabric technologies [12].

5.2 Intel

CPUs

The current HPC state-of-the-art for Intel CPU is represented by Sapphire Rapids, which was released in January 2023. Sapphire Rapids will be followed by Emerald Rapids, which is expected to be launched at the end of 2023. However, at the moment Intel has not revealed yet

whether an HBM version of Emerald Rapids more suitable for HPC will be commercialised [13]. Indeed the public Intel roadmap concerns mainly general datacenters and considers HPC as a subset of this.

The microarchitecture of Emerald Rapids, also known as Xeon Platinum 8580, is the new Raptor Cove core. The CPU will support more than 60 cores with 2 threads each. It will have a clock frequency of 2000 MHz. The size of L2 and L3 caches of Emerald Rapids will be 120 MB (same as Sapphire Rapids) and 300 MB (2.66 times larger than in Sapphire Rapids), respectively. Thus they will be 87.5% and 17.2% larger than the size of L2 and L3 caches of AMD EPYC 9554 (Genoa). The expected TDP will be 350 W and the overall performance-per-watt efficiency will be improved compared to Sapphire Rapids. Emerald Rapids will natively support DDR5-5600, instead of DDR-4800 present in Sapphire Rapids. As a consequence, the memory bandwidth available to the cores will be larger. There will also be 80 high speed PCIe 5.0 lanes for rapid connectivity [14] and CXL will be supported.

Emerald Rapids will be followed by Granite Rapids, whose release is planned for 2024. It will be the first Intel product to support higher bandwidth Multiple Combined Rank (MCR) DIMM memory. This feature allowed Granite Rapids to demonstrate 1.5 TB/second of memory bandwidth in demonstrative tests conducted by Intel on a dual socket system. Its design will be based on tiles with separate compute and I/O tiles (this is a major innovation with respect to Sapphire Rapids, where each tile is a complete system on chip) [15]. Furthermore the number of tiles is reduced with respect to Sapphire Rapids, passing from 4 to 2.

GPUs

The GPU counterpart of Sapphire Rapids is called Ponte Vecchio and was released together with the processor at the beginning of 2023 as part of the Intel Data Center GPU Max Series. The next product in the pipeline is Falcon Shores expected for 2025. Initially thought as an heterogeneous device composed of a CPU and GPU on a single chip, it has been converted to a pure GPU product. Falcon Shores will target HPC and AI workloads. The GPU will feature HBM3 which will provide up to 288 GB of capacity and 9.8 TB/s of bandwidth. CXL could be leveraged for CPU/GPU communications. According to Intel, a GPU-only product gives customers more flexibility. Indeed, Falcon Shores could be connected with different CPU types and the ratio CPU/GPU could be varied [17]. It is expected that Falcon Shores will show a 5x

performance-per-watt more and a 5x larger memory capacity with respect to Ponte Vecchio [18]. Falcon Shores, as Ponte Vecchio, will be fully supported through the Intel OneAPI software ecosystem [17].

5.3 NVIDIA

5.3.1 Recently released architecture (Nov 2023)

Currently the most important HW platform from NVidia with respect to EuroHPC is the Grace-Hopper [19] architecture which is used to build Europe's first exascale supercomputer Jupiter to be installed at Jülich Supercomputing Centre (JSC) [20]. This system will be installed 2024 and therefore will be operational within the timeframe of MAX3.

The NVIDIA Hopper GPU is the central component of the system, representing the latest release in NVIDIA's high-performance computing general-purpose GPUs. These powerful graphics processing units are integrated into the Grace-Hopper superchip, a unique combination of the latest-generation GPU and NVIDIA's first CPU, which is known as the GH200.

Each Booster node contains 4x GH200 superchips (four GPUs each closely attached to a partner CPU via NVLink Chip-to-Chip), see Figure 4.1 for more details. With 72 cores per Grace CPU, a node has a total of 288 CPU cores (Arm). In a node, all GPUs are connected via NVLink 4, all CPUs are connected via CPU NVLink connections.

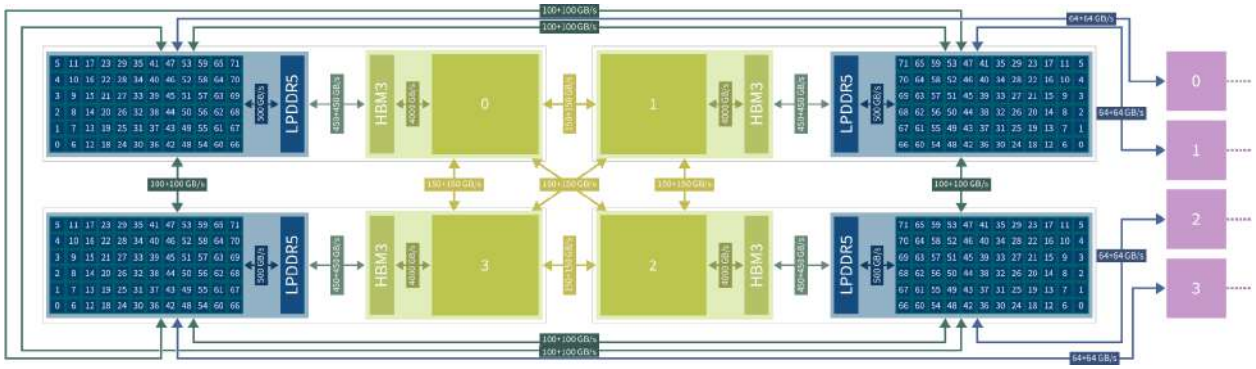

Fig. 5.1. Jupiter Booster node diagram based on NVidia Grace-Hopper architecture [taken from 20].

The Hopper H100 GPU installed in the system provides 96 GB HBM3 memory with 4 TB/s bandwidth from the GPU's multiprocessors. The H100 offers an increased number of multiprocessors, larger caches, new core architectures, and further advancements over previous

NVIDIA GPU generations. The documentation released by NVIDIA provides more information [21]. The new generation of the NVLink4 allows for transmitting data from one GPU to any other GPU in a node with 150GB/s per direction.

Each H100 GPU is connected to a Grace CPU. Grace is NVIDIA's first High Performance Computing (HPC) CPU, which utilises the Arm instruction set. The Grace CPU boasts 72 Neoverse V2 cores, each with SVE2 capability and four 128-bit functional units. With a bandwidth of 500 GB/s. The CPU SKU in the Jupiter has access to 120 GB of LPDDR5X memory. There are other SKUs with memory size up to 480 GB.

The superchip, see Figure 4.2, design's key feature is the tight integration between the CPU and GPU, which offers high bandwidth (450 GB/s per direction). Further details regarding Grace can be found in the NVIDIA documentation [22].

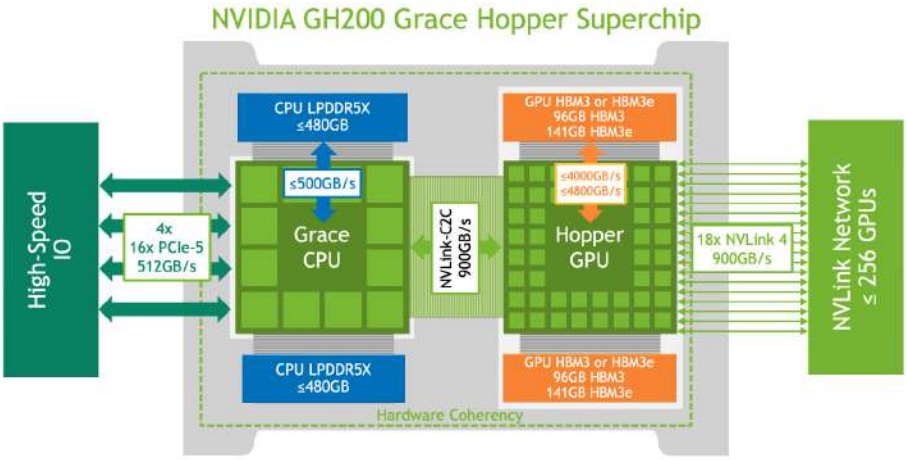

Fig. 5.2. NVIDIA Grace Hopper Superchip Architecture [taken from 22].

Any CPU connects to the three neighbouring CPUs in a node through dedicated CPU NVLink (cNVLink) connections that offer bi-directional bandwidth of 100 GB/s. A CPU also has a PCIe Gen 5 connection towards its associated InfiniBand adapter (HCA). In the case of Jupiter, each node has four InfiniBand NDR HCAs that each offer 200 Gbit/s bandwidth.

5.3.2 NVidia's roadmap

For upcoming years NVidia is planning to release new products every year. In the scope of MAX we can therefore expect, in addition to GH100 and GH200 accelerators, the new Blackwell based GPU accelerators B100 and BH200 [24, 26]. While the first one to designed for x86 machines the second will be paired with ARM CPU as it is in case of GH200 used in Jupiter. Figure 4.3 is taken from the NVIDIA Investor Presentation in October 2023 [25].

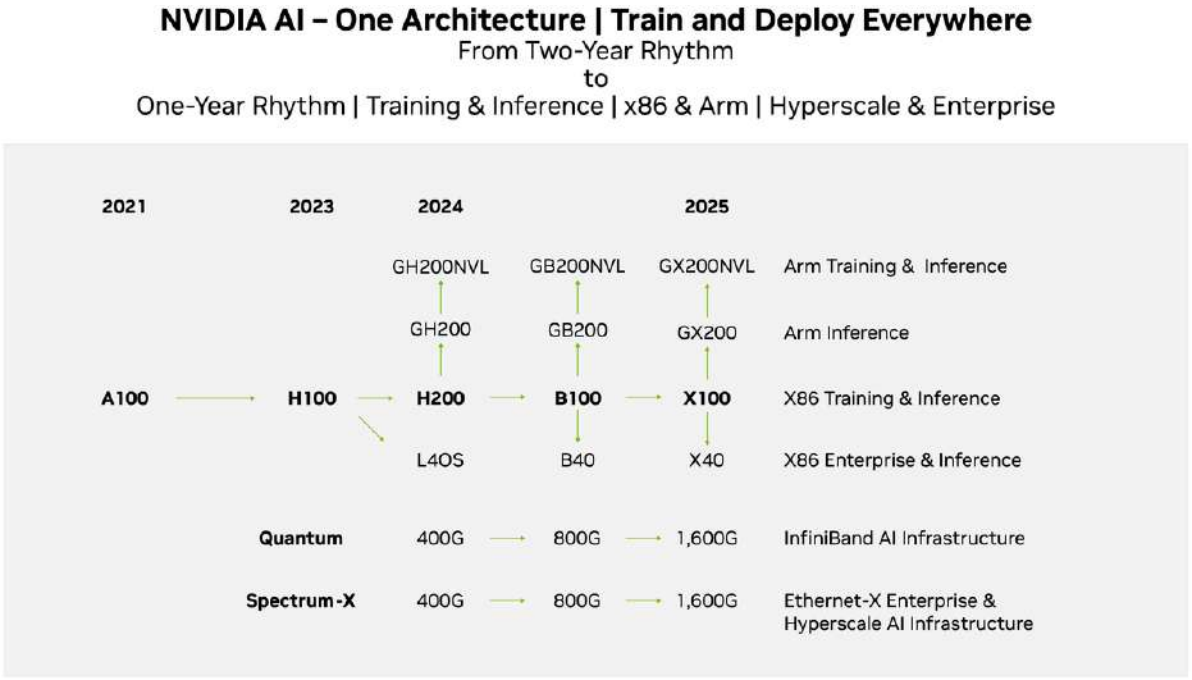

Fig. 5.3. NVIDIA Roadmap (taken from [24], Image credit: NVidia).

5.4 SiPearl

5.5 Rhea processor

SiPearl's first chip, Rhea, will be based on 72+ ARM Neoverse-V1 cores, each supporting ARM's Scalar Vector Extension (SVE) with a vector size of 256 bits. HBM2e memory will also be on board of the CPU, making it one of the fastest on the market. More technical details can be found in Figure 4.4.

Developed in partnership with the European Processor Initiative, it will also host prototypes of accelerators from various EPI partners, an important step for the European HPC community. It

will be part of the Jupiter Exascale CPU cluster system in Jülich, as announced in October, which will be the first system featuring an European processor. First chips should be available in 2024.

Core	- Arm Architecture Neoverse V1 cores - SVE 256 per core supporting 64/32/BF16 and int8 - Arm Virtualization Extensions
SoC	- Arm Mesh fabric - Advanced RAS support including Arm RAS extensions - Link protection for NOC and high-speed IO - ECC support for selected memory
Cache	- RAS supported for all Cache levels
Memory	ECC for memory and link protection for controllers - HBM2e - DDR-5
High Speed I/O	PCIe or CCIX/CXL: root and endpoint support
Other I/O	USB, GPIO, SPI, I²C...
Power Management	Power management block to optimize perf/watt across use cases and workloads.
Security Block Support	- Secure boot and secure upgrade - Crypto - True Random Number Generation

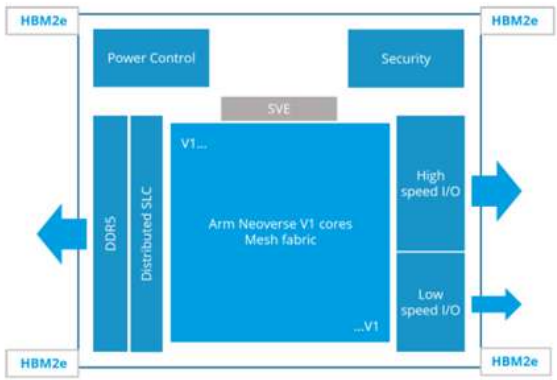

Fig. 5.4. SiPearl Rhea Processor characteristics [Image credit: SiPearl].

5.5.1 EUPEX

The European Pilot for Exascale (EUPEX) project is a cornerstone of the EuroHPC JU. Its goal is to gather and integrate European technologies from various domains, to create the first all-European platform for HPC: from system architecture and hardware, up to software and applications. The EUPEX prototype is designed to be open, scalable, and flexible thanks to an innovative modular architecture, and it adopts cutting-edge hardware specifically targeted at high-performance and energy-efficient computing, such as the first-generation Rhea processor developed in the EPI project.
Together with the hardware, the EUPEX project also features the creation of a suitable software ecosystem for the pilot, designed to take into account the modularity of the system and the needs of a set of key applications identified by the project.

In terms of hardware, EUPEX has a twofold objective:

- Design the hardware architecture of the heterogeneous modular Pilot platform
- Deploy and operate an Arm-based Pilot production-class platform using SiPearl Rhea CPU, as well as early test platforms, and make them available to the European scientific communities including MAX CoE

An overview of the planned system is shown in Figure 5.5.

Fig. 5.5. EUPEX pilot hardware system overview.

In terms of software, the objective of EUPEX is to provide a software ecosystem for the pilot based on European technologies. This includes a management software stack, an execution environment, some tools for tracing, optimization, monitoring and system status, and a multi-tier storage architecture (see Figure 2). Its design will not only take into consideration the needs of the key applications identified in EUPEX, but also those of the system operators for the management of large-scale Modular Supercomputing Architecture (MSA) systems.
An overview of the planned EUPEX software stack is shown in Figure 5.6.

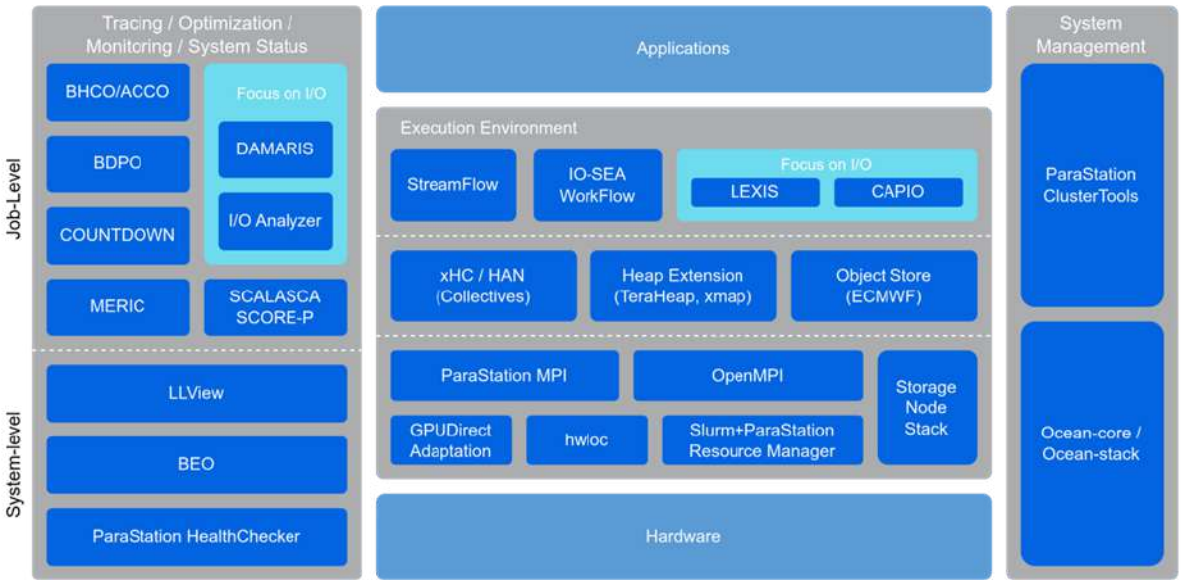

Fig. 5.6. EUPEX software stack overview.

5.6 RISC-V

Based on EuroHPC call the JU considers RISC-V as a key technology: "*The European Chips Act identified RISC-V as one of the next-generation technology where Europe should invest in order to preserve and strengthen its leadership in research and innovation as well as in equipment manufacturing, contributing to build and reinforce the Union's own capacity to innovate in the design, manufacturing and packaging of advanced, energy-efficient and secure chips, and turn them into manufactured products.*"[9] As such MAX will closely monitor RISC-V development in Europe.

5.6.1 EUPILOT VEC and MLS accelerators

EUPILOT is a pilot HPC ecosystem designed and manufactured entirely in Europe and is aimed at power-efficient exascale supercomputers. Within EUPILOT, two RISC-V-based accelerators created as part of the EPI project will be further developed into a full ecosystem (see Figure 4.5). **VEC** accelerator is designed to target traditional HPC applications, and **MLS** is optimised for Machine Learning and Stencil types of calculations.

VEC:

- Target at 1.5GHz
- 3.5 TFLOPs/unit
- supports RVV 1.0

MLS:

- Target at 1GHz
- 6.1 TFLOPs/EAS (8-bit)
- Supports SSR, SmallFloat, and more.

Tapeouts are planned for the end of 2024, and accelerators should be ready in 2025.

[9] Cited from: New call for developing an HPC ecosystem based on RISC-V
https://eurohpc-ju.europa.eu/new-call-developing-hpc-ecosystem-based-risc-v-2023-02-01_en

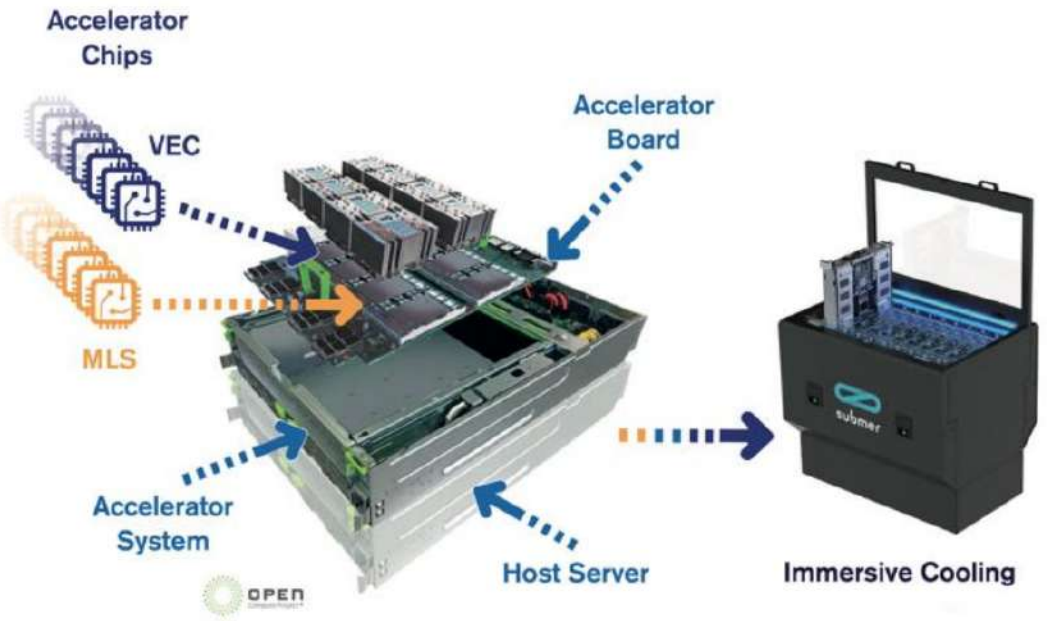

Fig. 5.7. EUPILOT accelerators [Image credit: EUPILOT website].

5.6.2 Driving the Convergence of AI and HPC Computing With Low Power RISC-V Solutions[10]

Esperanto Technologies has been at the forefront of RISC-V development since the company introduced the world's first fully RISC-V accelerator in 2021. With over 1,000 64b RISC-V cores operating at power levels as low as 15W for sparse AI workloads, Esperanto's ET-SoC-1 silicon is one of the few RISC-V accelerators shipping today.

Esperanto's advanced RISC-V architecture is being enhanced to accelerate the next generation of Generative AI large language models and High Performance Computing (HPC) workloads using the same silicon and software development environment. Building on the strengths of the first generation ET-SoC-1 and leveraging the RISC-V open standard, Esperanto's second and third generation processors are targeting Exascale systems with industry leading performance and memory bandwidth while keeping to very low levels of power consumption.

Esperanto's vision for a unified AI and HPC environment is based on the desires of enterprise and government customers that are looking to streamline the types of compute systems in data

[10] The following description has been adapted from a text directly provided by Esperanto Technologies

centres while taking advantage of similar processing tasks driving compute and memory bandwidth and capacity:

- **ET-SoC-1** (now shipping)
  - 7nm process node
  - Self-hosting capable
  - Up to 32 GB LPDDR4x DRAM
  - Supports Int 8, FP16 and FP32 data formats
  - Enabled by Esperanto's ML and General Purpose SDKs

- **ET-SoC-2L / 2LU** (under development)
  - 4nm process node
  - Self-hosting enabled
  - Up to 256 GB LPDDR5x DRAM
  - Expanded support for FP8, BF16 and FP64 data formats
  - DVFS and optional cache coherency features

- **ET-SoC-3C / 3LU** (under development)
  - 3nm process node
  - Self-hosting enabled
  - Up to six HBM4 stacked DRAM
  - Up to 9,600 GB/s memory bandwidth (peak)
  - Between 60W and 300W of power

6. Conclusions

This report introduces the efforts of WP4, which concentrate on identifying co-design platforms, advanced hardware platforms, and systems suitable for energy-efficiency evaluation. These platforms will be primarily utilised for benchmarking purposes via application kernels extracted from MAX codes.

In Section 3, we present the features of MAX codes relative to the advanced HW platforms. Defining the expected workload for each application and its modules allows us to concentrate on evaluating the most suitable platforms for a given code or kernel. For example, if we want to optimise a memory-bound kernel, the most promising approach is to focus on processors with new and more powerful memory technologies, such as DDR5 or HBM, that have higher memory bandwidth or bandwidth per CPU core.

This section also outlines the HW platforms of all consortium members that can be utilised for code performance evaluation, benchmarking, and co-design. A summary of all platforms is presented in Table 3.6.

Additionally, WP4 aims to evaluate the energy consumption and efficiency of different HW platforms. Table 3.7 presents a list of platforms suitable for this study, as they enable us to measure energy consumption and adjust hardware parameters that affect the power usage of a processor or accelerator. These procedures can be challenging to carry out on production systems due to several constraints. Accordingly, we have only included platforms on which we can execute the desired actions.

Access mechanisms to the platforms outlined above are described in Section 4. Each partner who provided the platforms outlined instructions for accessing the systems and described the restrictions.

Section 6 details the roadmaps of the primary technology developers, including upcoming platforms and those that have recently been released but are not yet available to any of the MAX partners.

In summary, this document also serves as a starting point for all MAX3 partners who are interested in exploring the advanced hardware platforms for their codes in finding the new paths for code optimization and porting.

7. References

[1] MaX Phase2 Deliverable 4.6, Final report on co-design activities : http://max-centre.eu/sites/default/files/D4.6_Final%20report%20on%20co-design%20activities.pdf

[2] MaX Phase2 Deliverable 4.5, Final report on code profiling and bottleneck identification: http://max-centre.eu/sites/default/files/D4.5_Final%20report%20on%20codes%20profiling%20and%20bottlenck%20identification.pdf

[3] MAX WP4 - platform table: https://docs.google.com/spreadsheets/d/11cvydnovr6f23PXrW-6E-qgOYaaW9JQX/edit#gid=605517971

[4] AMD Updated Epyc Roadmap: 5th Gen EPYC “Turin” Announced, Coming by End 2024 https://www.anandtech.com/show/17437/amd-updated-epyc-roadmap-until-2024-5th-gen-epyc-announced-coming-by-end-of-2024

[5] AMD Zen 5 Microarchitecture https://www.techpowerup.com/314231/amd-zen-5-microarchitecture-referenced-in-leaked-slides

[6] AMD Zen 5 & Zen 5C EPYC CPU https://wccftech.com/amd-zen-5-zen-5c-epyc-cpu-turin-16-ccd-128-cores-turin-dense-12-ccd-192-cores-turin-x-1-5-gb-cache/

[7] AMD Expands Leadership Data Center Portfolio https://www.amd.com/en/newsroom/press-releases/2023-6-13-amd-expands-leadership-data-center-portfolio-with-.html

[8] AMD Expands MI300 With GPU-Only Model https://www.tomshardware.com/news/amd-expands-mi300-with-gpu-only-model-eight-gpu-platform-with-15tb-of-hbm3

[9] AMD 2022-2024 GPU Roadmap Confirms Next-Gen GPUs & APUs by 2024 https://wccftech.com/amd-2022-2024-gpu-roadmap-confirms-next-gen-rdna-4-radeon-rx-8000-gpus-cdna-3-instinct-mi300-apus

[10] AMD Confirms Next-Gen Instinct MI400 Series AI Accelerators Already In The Works
https://wccftech.com/amd-confirms-next-gen-instinct-mi400-series-ai-accelerators-already-in-the-works/

[11] Instinct MI400 HPC APU Accelerator As Part of AMD's Instinct Roadmap
https://wccftech.com/lenovo-vp-confirms-instinct-mi400-hpc-apu-accelerator-as-part-of-amd-instinct-roadmap/

[12] AMD touts big datacenter, AI ambitions in CPU-GPU roadmap
https://www.theregister.com/2022/06/10/amd_chip_roadmap/

[13] Intel HPC Updates for ISC 2023
https://www.anandtech.com/show/18869/intel-hpc-update-isc-2023-falcon-shores-details-future-xpu-aurora-nearly-done

[14] 5th-Gen Emerald Rapids CPU Leak Shows 60 Cores, 420MB Cache
https://www.tomshardware.com/news/5th-gen-emerald-rapids-cpu-leak-shows-60-cores-420mb-cache

[15] Intel Updates Data Center Roadmap
https://www.anandtech.com/show/18797/intel-updates-data-center-roadmap-xeons-on-track-emerald-in-q423-sierra-forest-in-h124

[16] Intel Next Server GPU Will be Falcon Shores in 2025
https://www.anandtech.com/show/18756/intel-scraps-rialto-bridge-gpu-next-server-gpu-will-be-falcon-shores-in-2025

[17] Intel Reveals New Supercomputing Chip Roadmap
https://www.hpcwire.com/2023/05/22/intel-admits-gpu-mistakes-reveals-new-supercomputing-chip-roadmap/

[18] Intel Talks Falcon Shores Flup
https://www.tomshardware.com/news/intel-explains-falcon-shores-redefinition-shares-roadmap-and-first-details

[19] Grace-Hopper Architecture:
https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

[20] JUPITER Technical Overview: https://www.fz-juelich.de/en/ias/jsc/jupiter/tech

[21] NVIDIA Hopper Architecture In-Depth
https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/

[22] NVIDIA Grace Hopper Superchip Architecture Whitepaper
https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

[23] EUPILOT projects website
https://eupilot.eu/hardware/

[24] NVIDIA teases next-gen B100 Blackwell GPU
https://www.tomshardware.com/news/nvidia-may-move-to-yearly-gpu-architecture-releases

[25] NVIDIA Investor Presentation October 2023,
https://s201.q4cdn.com/141608511/files/doc_presentations/2023/Oct/01/ndr_presentation_oct_2023_final.pdf

[26] NVIDIA teases next-gen B100 Blackwell GPU
https://videocardz.com/newz/nvidia-teases-next-gen-b100-blackwell-gpu-performance-in-gpt-3-175b-large-language-model