CIUDAD REAL 2023

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Presente y futuro de la computación… y el papel de las FPGAs

(An Open Computing Era)

22 de Septiembre de 2023

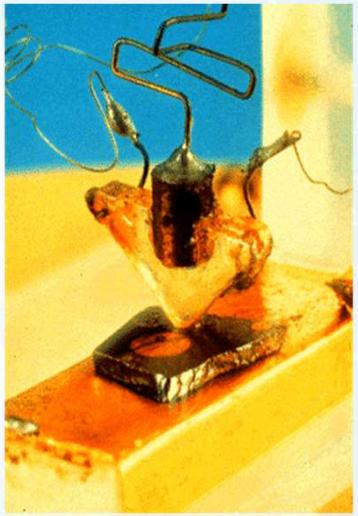
Teresa Cervero, Ph.D.
teresa.cervero@bsc.es

Un poco de historia…

Bell Labs, 1947 — **1958** — **1971** — **1990**

Transistor

- DEC PDP-1 (1957)
- IBM 7090 (1960)

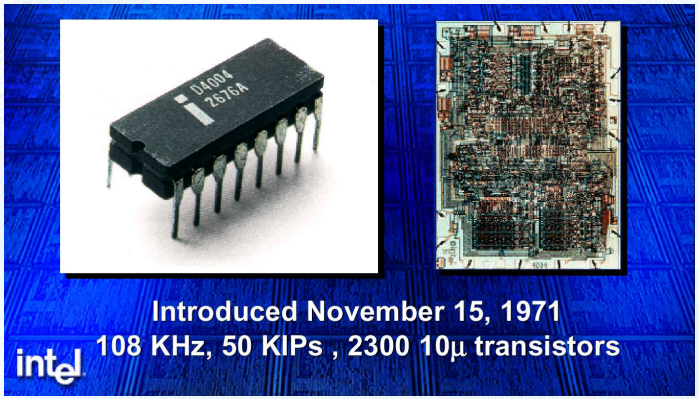


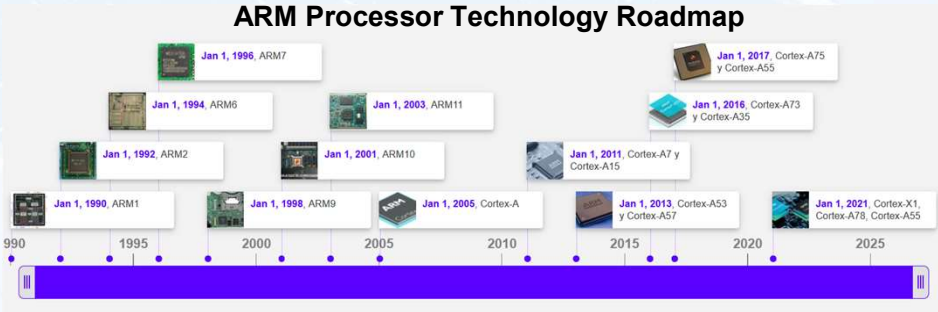
Integrated circuit

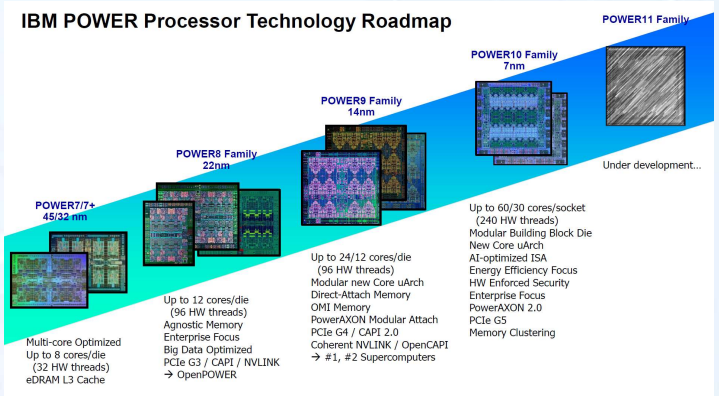
- IBM System 360 (1965)
- DEC PDP-8 (1965)

Microprocessor

- Intel 4004







NVIDIA A100 detailed specs

	Peak Performance
Transistor Count	54 billion
Die Size	826 mm^2
FP64 CUDA Cores	3,456
FP32 CUDA Cores	6,912
Tensor Cores	432
Streaming Multiprocessors	108
FP64	9.7 teraFLOPS
FP64 Tensor Core	19.5 teraFLOPS
FP32	19.5 teraFLOPS
TF32 Tensor Core	156 teraFLOPS | 312 teraFLOPS*
BFLOAT16 Tensor Core	312 teraFLOPS | 624 teraFLOPS*
FP16 Tensor Core	312 teraFLOPS | 624 teraFLOPS*
INT8 Tensor Core	624 TOPS | 1,248 TOPS*
INT4 Tensor Core	1,248 TOPS | 2,496 TOPS*
GPU Memory	40 GB (1555 GB/s BW), 40MB L2 cache
Interconnect	NVLink 600 GB/s PCIe Gen4 64 GB/s
Multi-Instance GPUs	Various Instance sizes with up to 7MIGs @5GB
Form Factor	4/8 SXM GPUs in HGX A100
Max Power	400W (SXM)

* Includes Sparsity

Supercomputadores europeos

- LUMI

375 petaflops Sustained performance
550 petaflops Peak performance

- LEONARDO

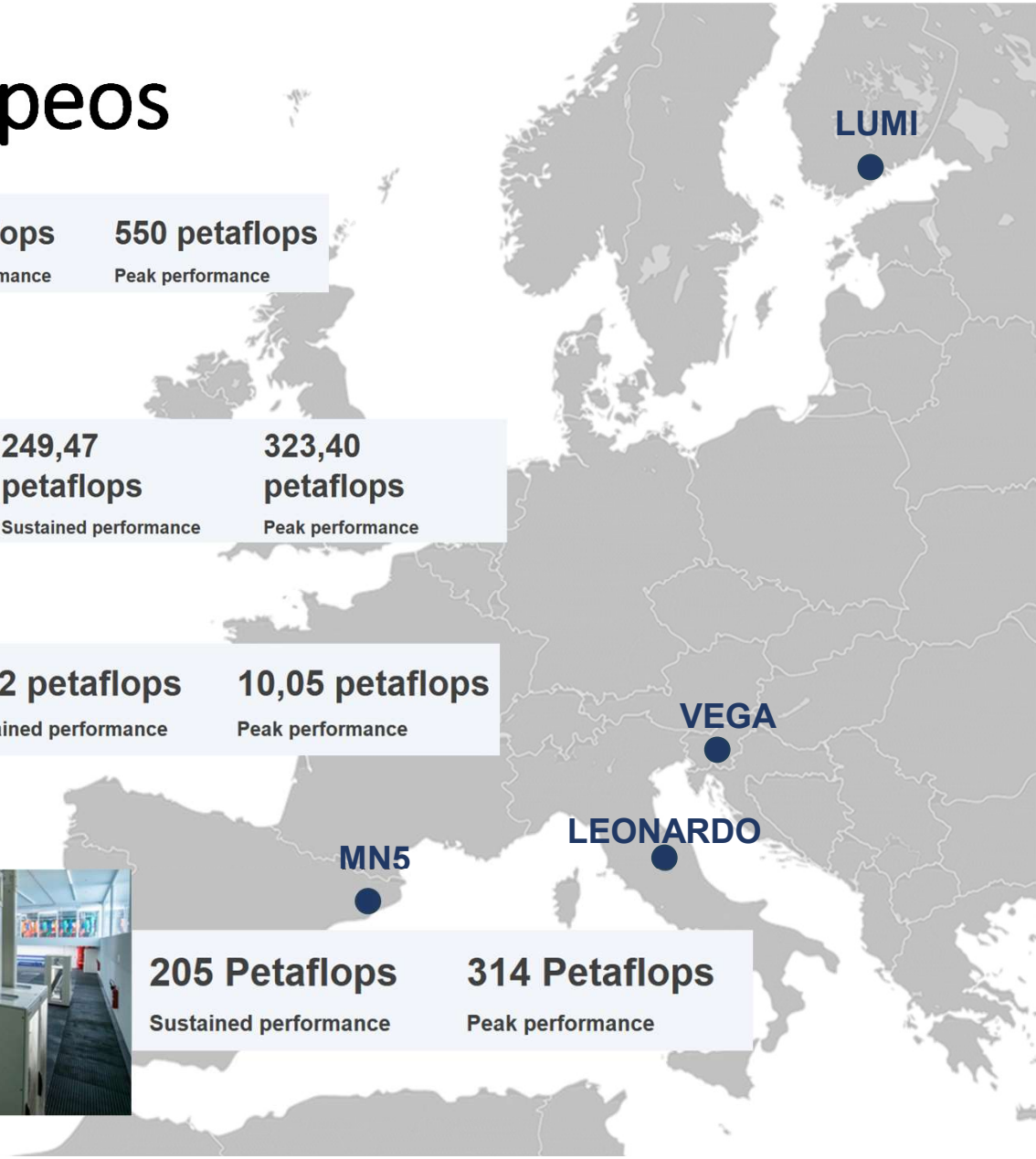
249,47 petaflops Sustained performance
323,40 petaflops Peak performance

- VEGA

6,92 petaflops Sustained performance
10,05 petaflops Peak performance

- MARENOSTRUM5

205 Petaflops Sustained performance
314 Petaflops Peak performance



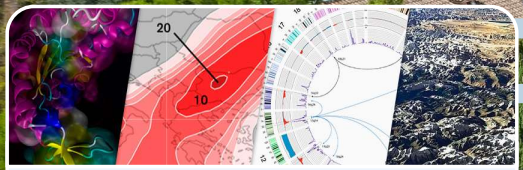
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Top 10 Supercomputadores, Junio 2023

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016
4	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos EuroHPC/CINECA Italy	1,824,768	238.70	304.47	7,404
5	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
6	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438
7	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93.01	125.44	15,371
8	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States	761,856	70.87	93.75	2,589
9	**Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	555,520	63.46	79.22	2,646
10	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000, NUDT National Super Computer Center in Guangzhou China	4,981,760	61.44	100.68	18,482

BSC-CNS objectives

Supercomputing services
to Spanish and EU researchers



R&D in Computer, Life, Earth
and Engineering Sciences

PhD programme, Tech
transfer, public engagement

BSC-CNS is a consortium that includes

Spanish Government	60%	Gobierno de España · Ministerio de Ciencia, Innovación y Universidades
Catalan Government	30%	Generalitat de Catalunya Departament d'Empresa i Coneixement
Univ. Politècnica de Catalunya (UPC)	10%	Universitat Politècnica de Catalunya BarcelonaTech

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MareNostrum 5

Total peak performance: **314** Pflops

PRACE

Access: prace-ri.eu/hpc_acces

RES
RED ESPAÑOLA DE SUPERCOMPUTACIÓN

Access: bsc.es/res-intranet

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MareNostrum 1
2004 – 42,3 Tflops
1st Europe / 4th World
New technologies

MareNostrum 2
2006 – 94,2 Tflops
1st Europe / 5th World
New technologies

MareNostrum 3
2012 – 1,1 Pflops
12th Europe / 36th World

MareNostrum 4
2017 – 11,1 Pflops
2nd Europe / 13th World
New technologies

Visión BSC sobre el Futuro del HPC en Europa

MontBlanc @ BSC
2014

Visión BSC sobre el Futuro del HPC en Europa

MontBlanc @ BSC
2014

Visión BSC sobre el Futuro del HPC en Europa

MontBlanc @ BSC 2014

1 B€ + 6 yrs

Fugaku #1 Top500 @ Riken for $1B over 7 years (CAPEX)

Visión BSC sobre el Futuro del HPC en Europa

European
HPC Accelerators
& Eventually GPP

EPI @ BSC

BUILD vs. buy

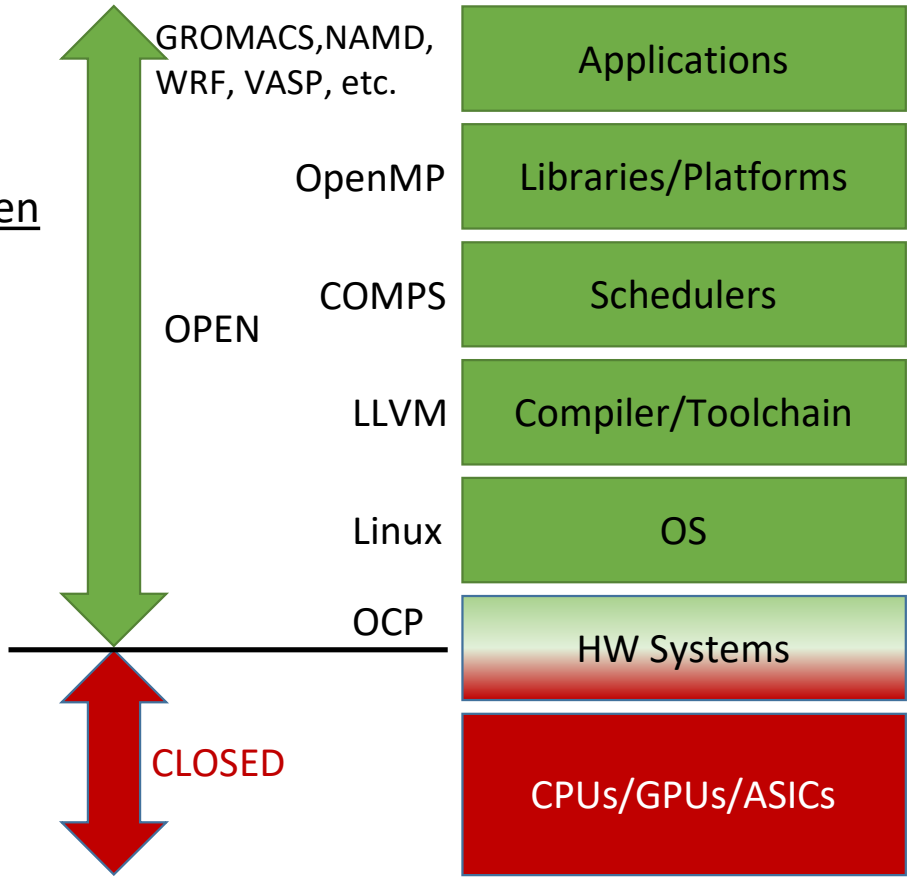
MareNostrum6, …

European Supercomputers @ Top500

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

HPC Hoy en día

- **Europe lidera la definición de un ecosistema de SW abierto en HPC**
  - **Linux:** estándar OS de facto
- Software landscape (Cloud-edge-IoT): beneficio de open source
  - Open source provides:
    - A common platform, specification and interface
    - Accelerates building new functionality by leveraging existing components
    - Lowers the entry barrier for others to contribute new components
    - Crowd-sources solutions for small and larger problems
- **Y…. ¿Qué pasa con el HW?**

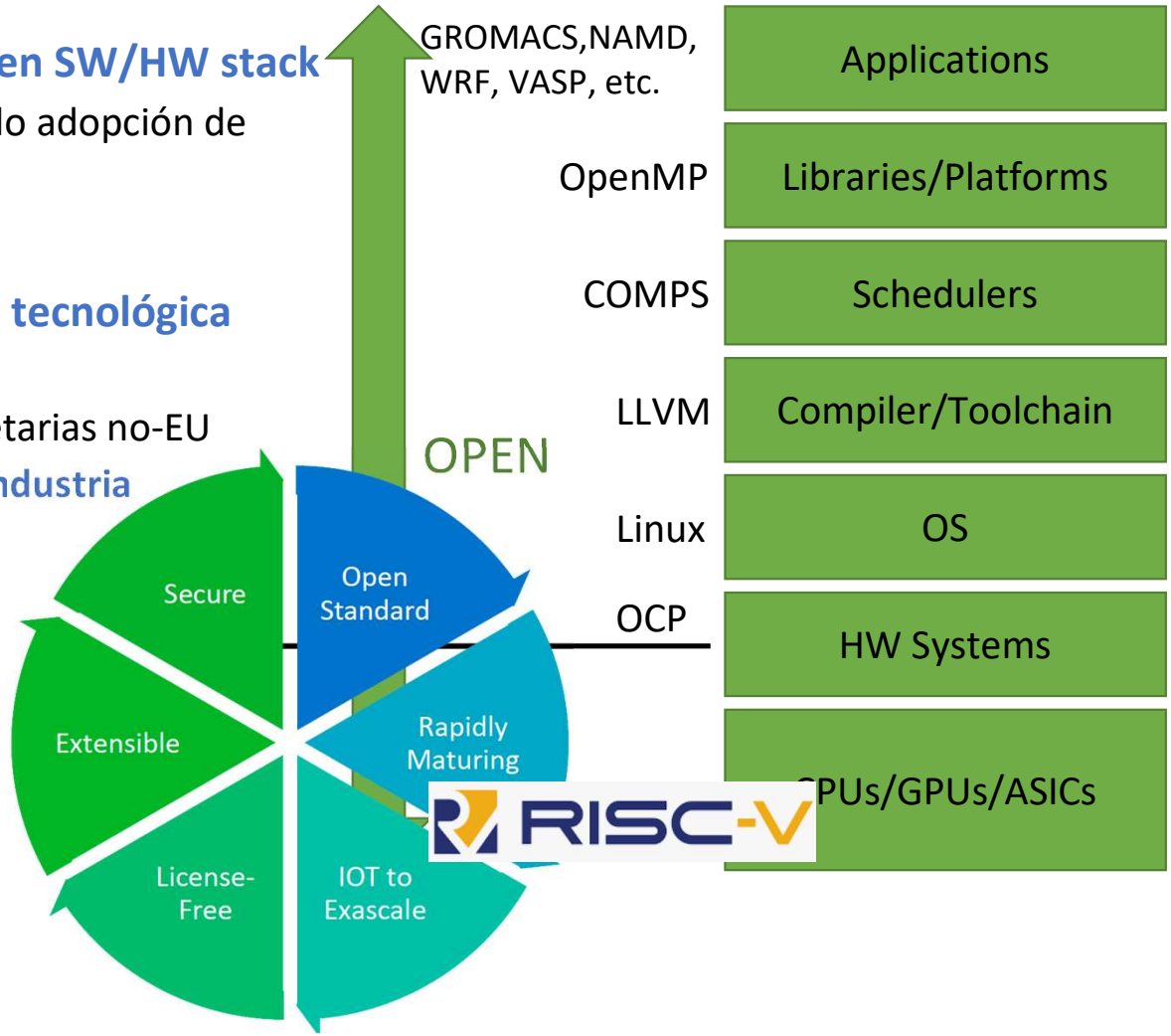


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

HPC Mañana

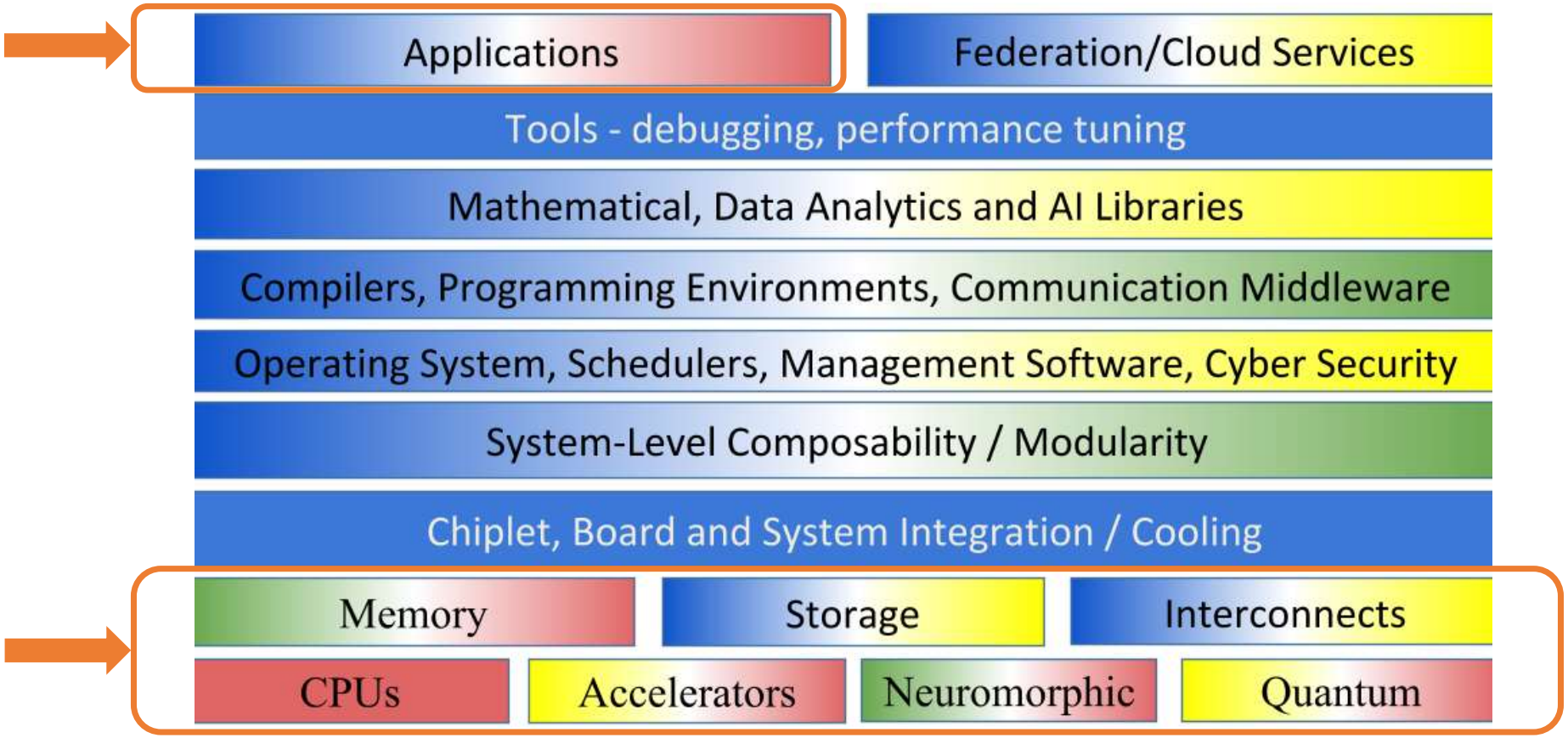
- Europe puede liderar el camino para **open SW/HW stack**
  - HW open-stack: **Inmadurez** = SW cuando adopción de Linux

- Europe puede conseguir **independencia tecnológica**
  - RISC-V = alternativa a soluciones propietarias no-EU
  - RISC-V puede **unificar y construir una industria microelectrósica en Europa.**



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

European HPC Stack TRL



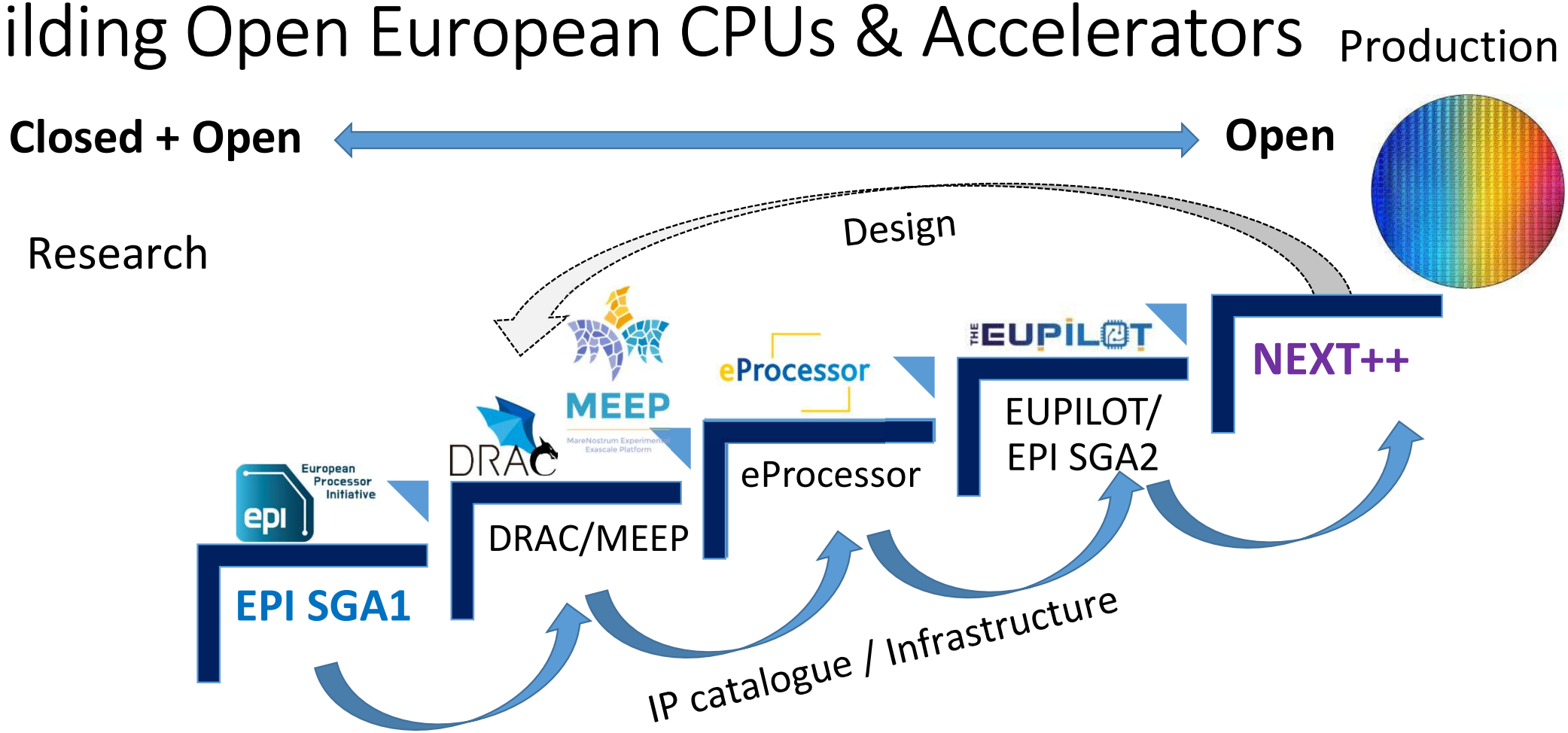
Blue (TRL 8-9) / Green (TRL 5-7) / Yellow (TRL 3-4) / Red (TRL 0-2)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

RISC-V@BSC

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Building Open European CPUs & Accelerators



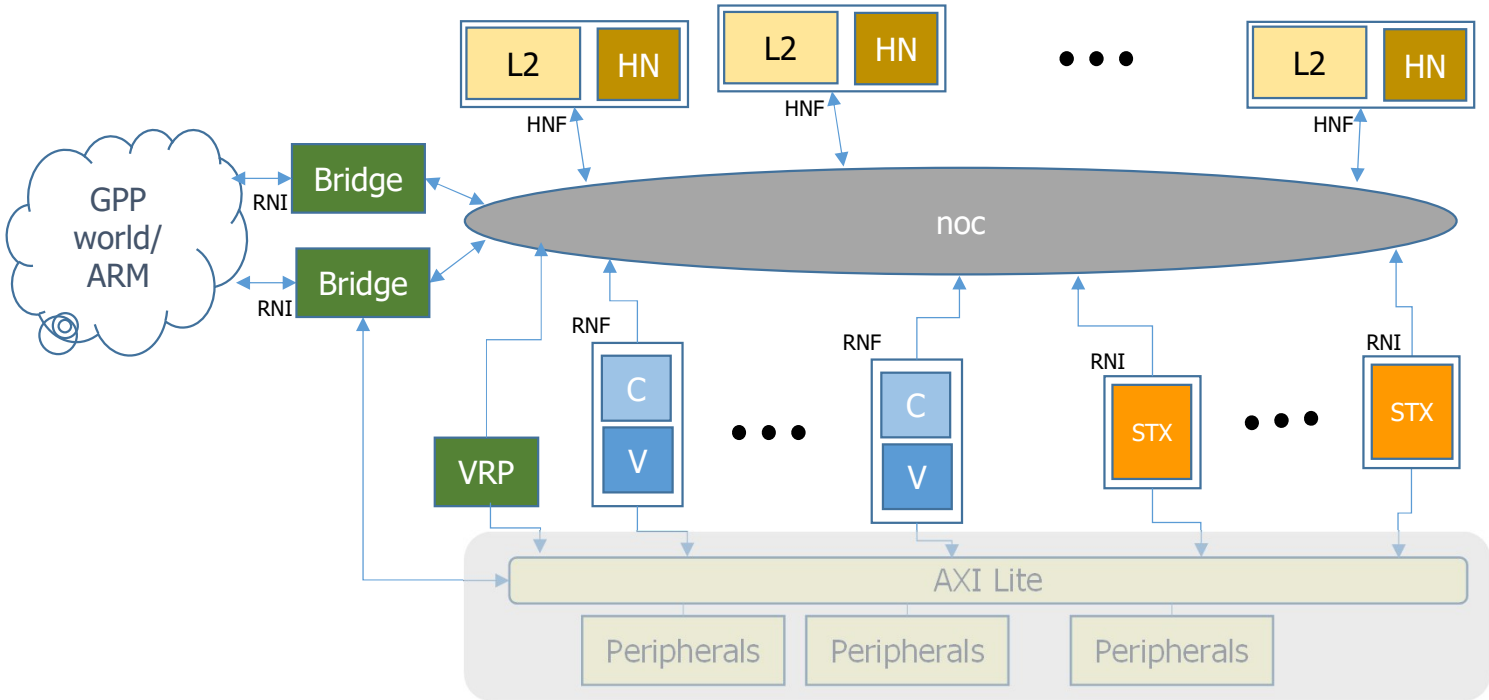
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EPAC architecture

- **Objective: Develop** & demonstrate **fully European** processor IPs based on the **RISC-V ISA.** Build on existing EU IP, leverage EU background and vision
- Provide a very **low power** and high computing throughput **accelerator for HPC & Emerging -> Automotive**

Main Contributors:

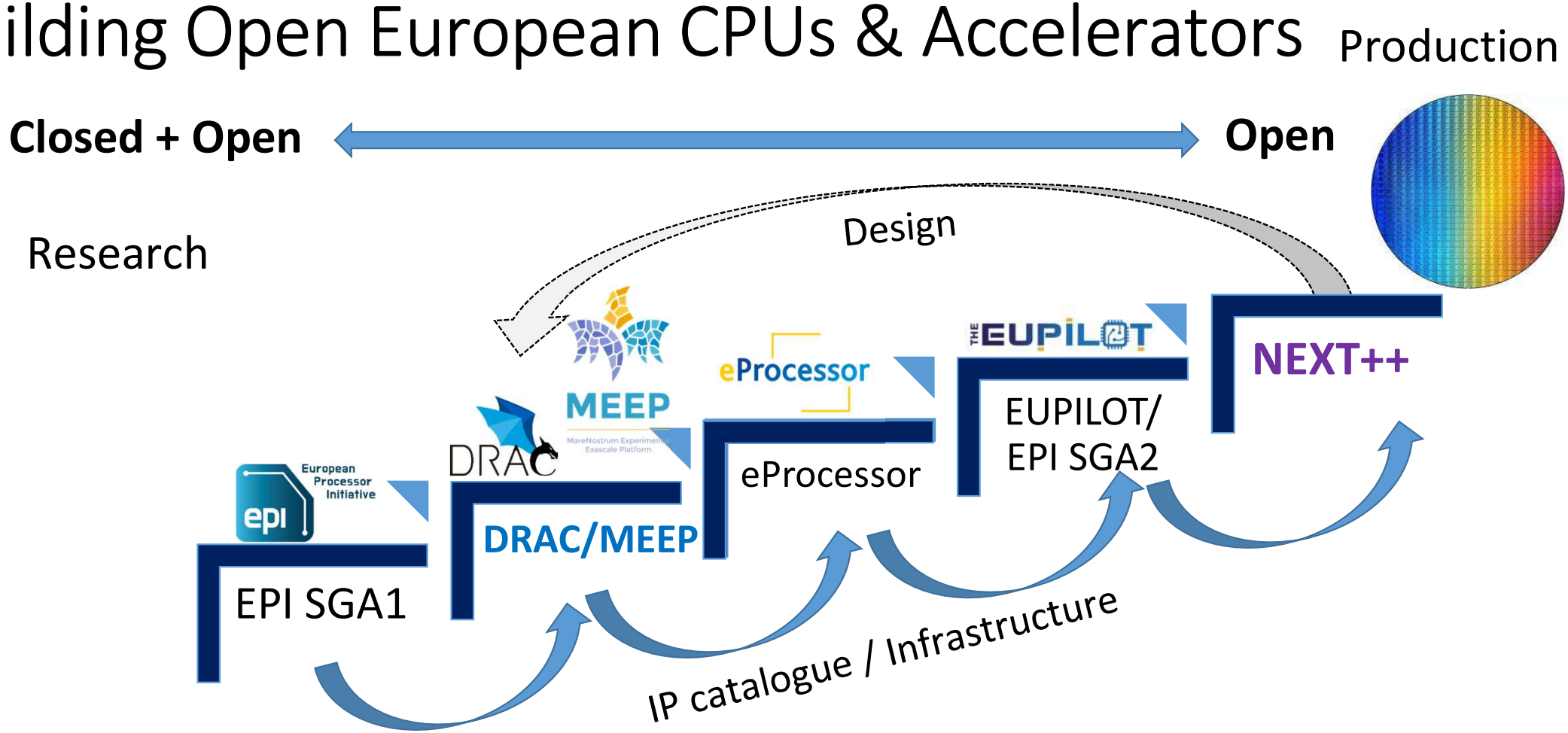
- **RISC-V in the Tile:**
  - **8 Vector cores, 8 STX, 1 VRP**
- **Vector Core :**
  - **C: RISC-V Scalar core**
  - **V: RISC-V Vector core: 8 lanes**
- **STX**
  - **RISC-V NTX (AI/ML/DL) + Stencil**
- **VRP**
  - **RISCV + Extended precision FPU**



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

UNIBO | IST | CEA | Chalmers | FORTH | Semidynamics | BSC +FER | ETH + FhG | EXTOLL | ETH

Building Open European CPUs & Accelerators

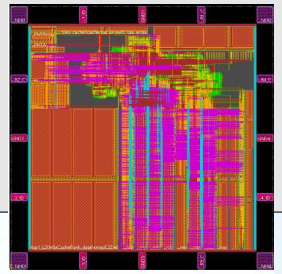


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

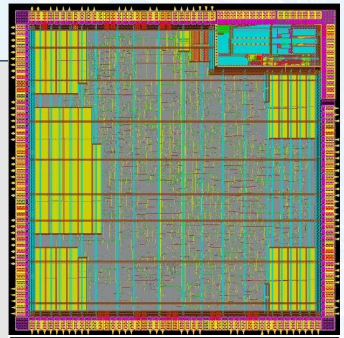
Lagarto RISC-V Tapeouts (DRAC)

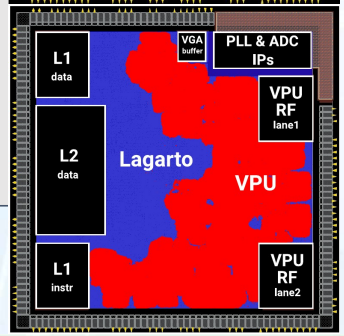
2019 | 2020 | 2021 | 2022 | 2023

Lagarto (May'19)
- Lagarto Hun 5-stage in-order
- 150MHz (external)
- TSMC 65nm
- 2.5mm2

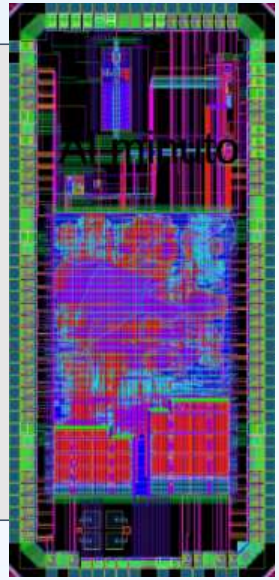


DVINO (Apr'21)
- Lagarto Hun in-order
- VPU
- PLL 600 MHz
- SDRAM mem cont
- HyperRAM
- VGA
- ADC
- TSMC 65nm
- 8mm2

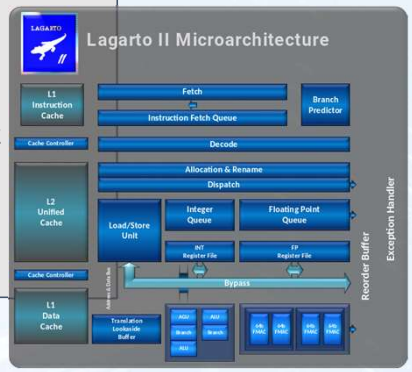




Sargantana (Feb'22)
- Sargantana 7-stage in-order
- PLL 1.5 GHz
- Custom extensions
- SDRAM
- Prototype analog IPs: SerDes 8GHz
- GF 22nm
- Area: 2.9mm2

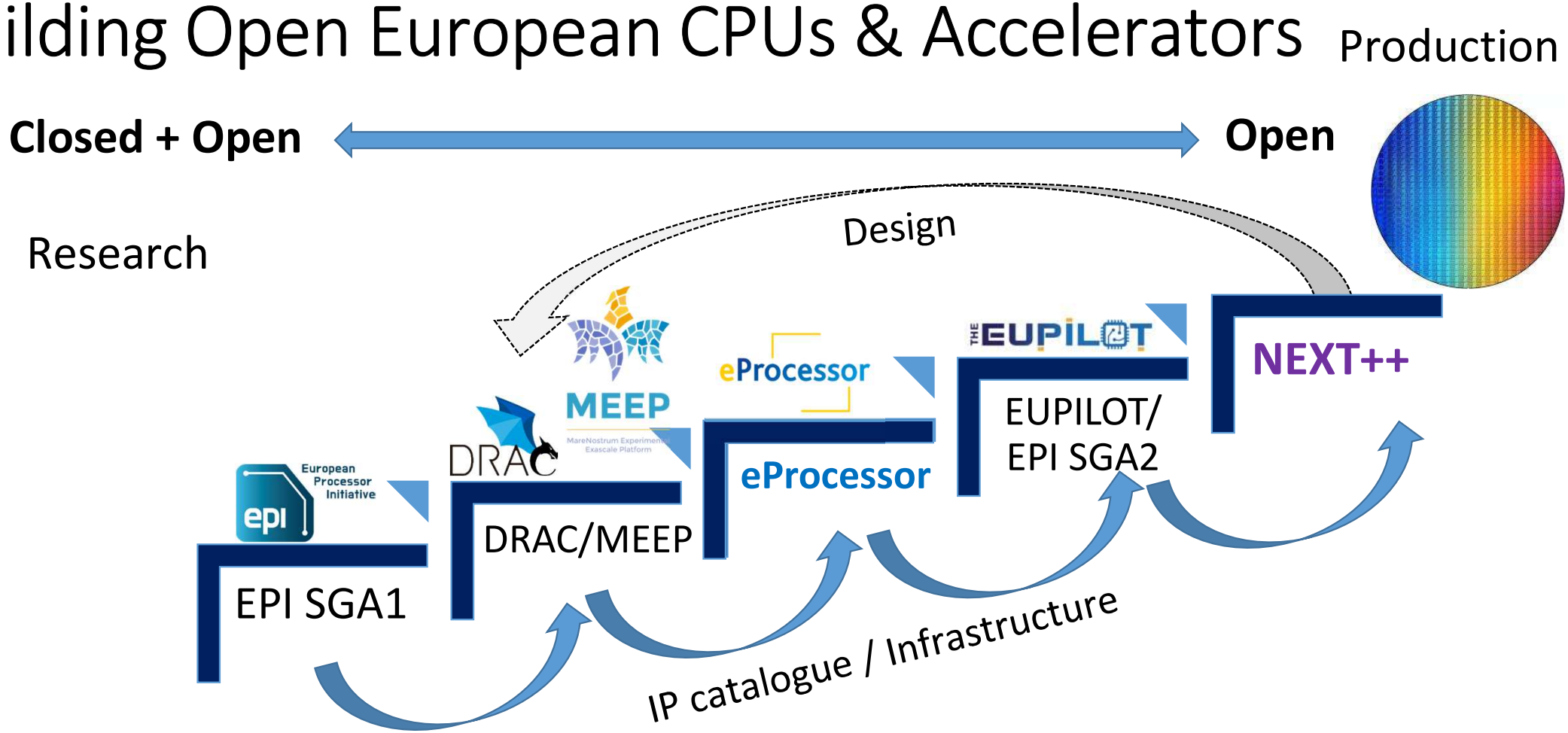


Kameleon (Dec'22)
- Lagarto Ka ooo
- PLL 1.5 GHz
- Automotive Acc
- Crypto Acc
- Genomic Acc
- PICOS Acc
- SerDes 8 GHz
- GF 22nm
- Area: 9mm2



DVino → Sargantana → Largato Ka → Largato Ox

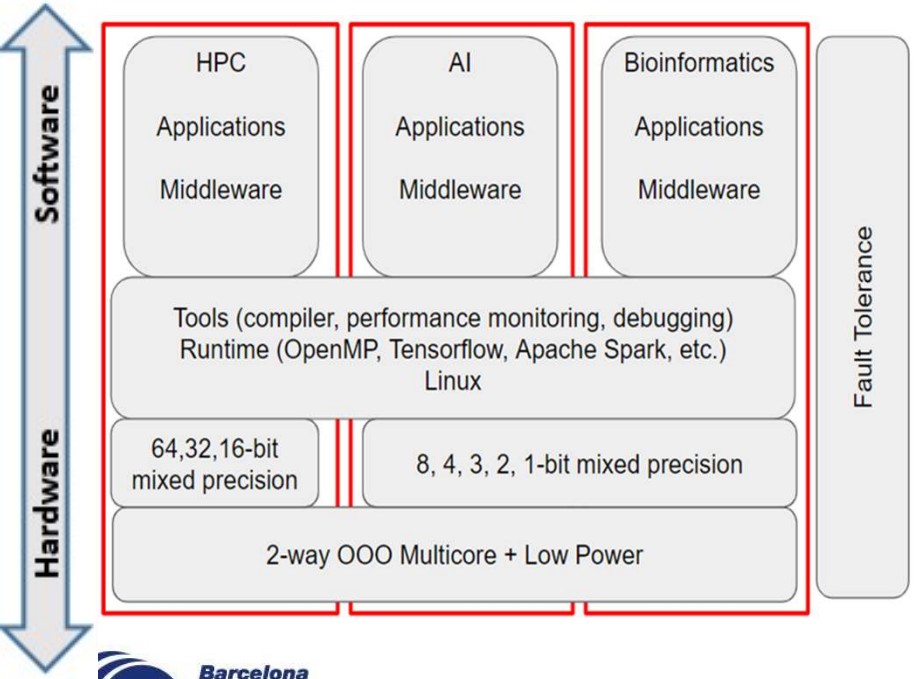
Building Open European CPUs & Accelerators

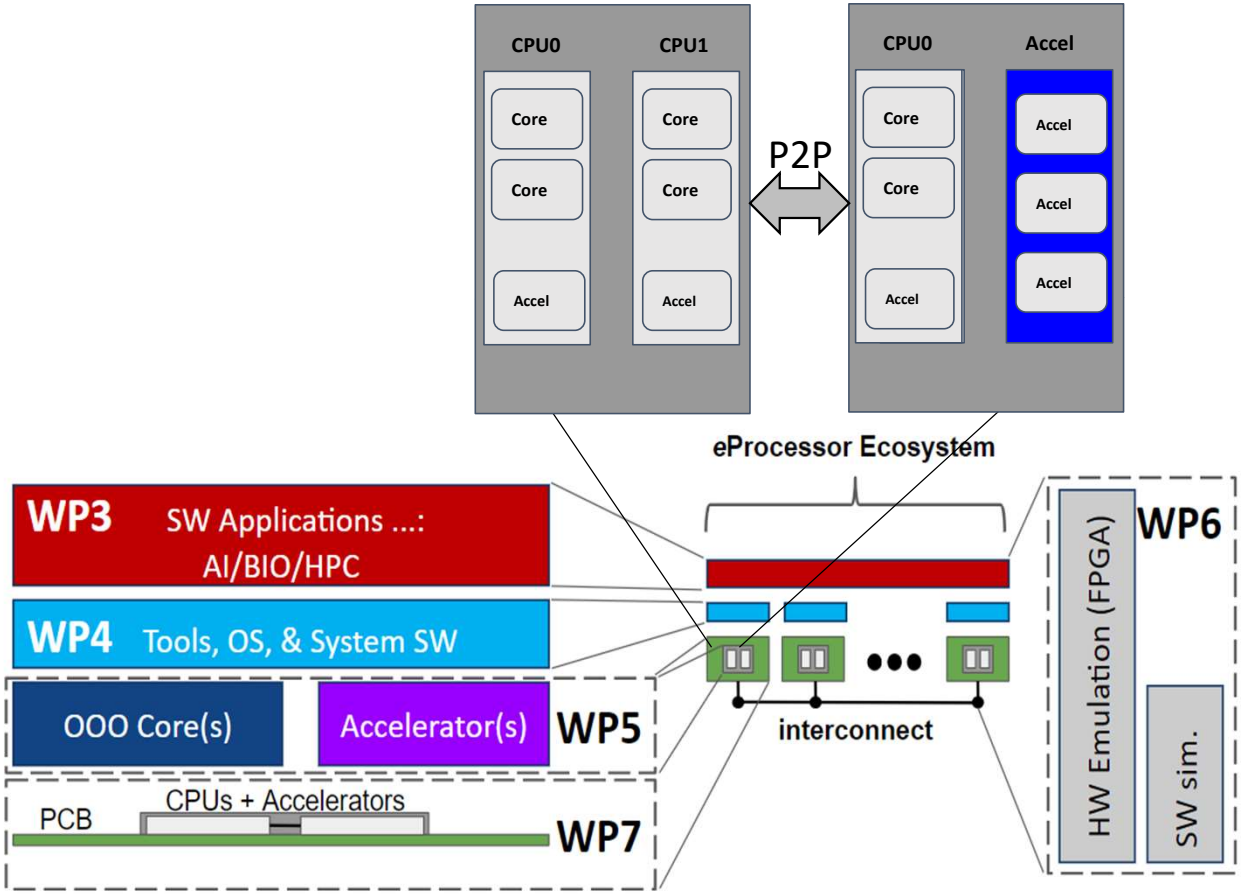


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

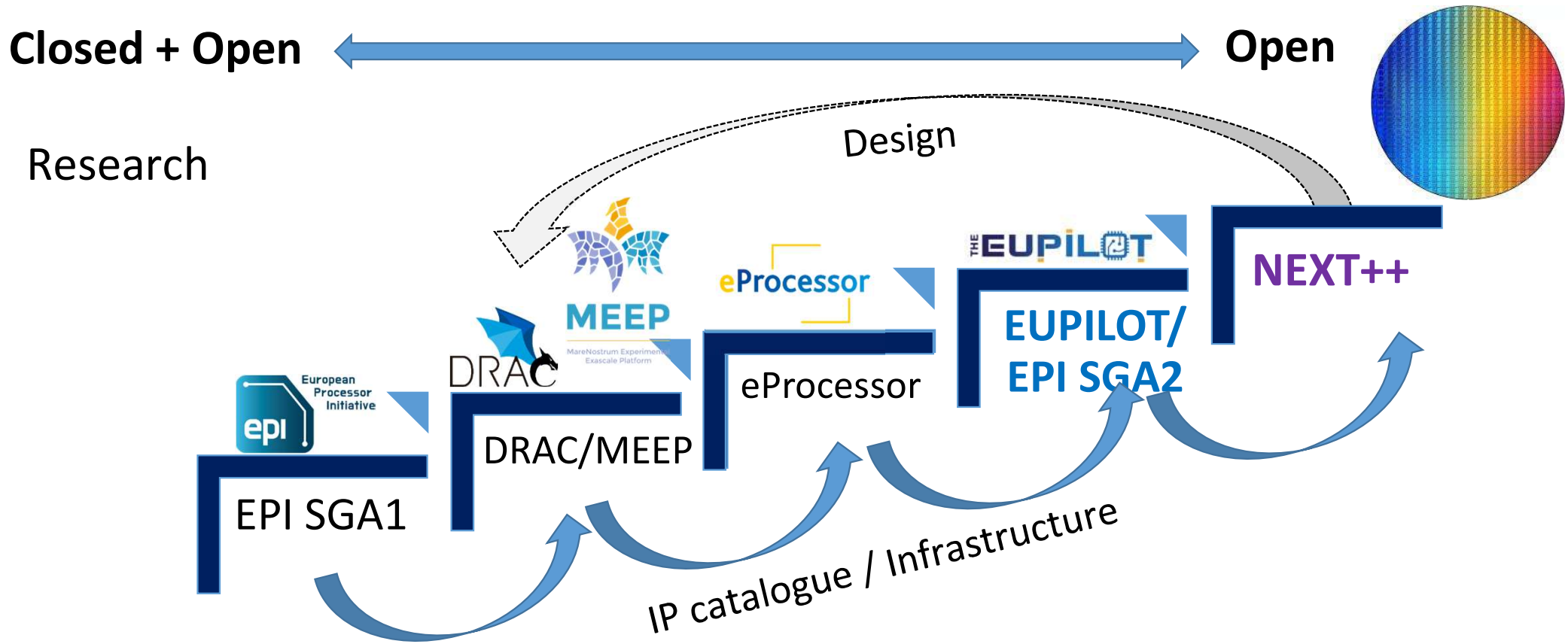
eProcessor

- **Objective: HW/SW co-design** for improved application performance and system energy efficiency
- 1er **Open source** high performance **OoO 64-bit RISC-V platform**





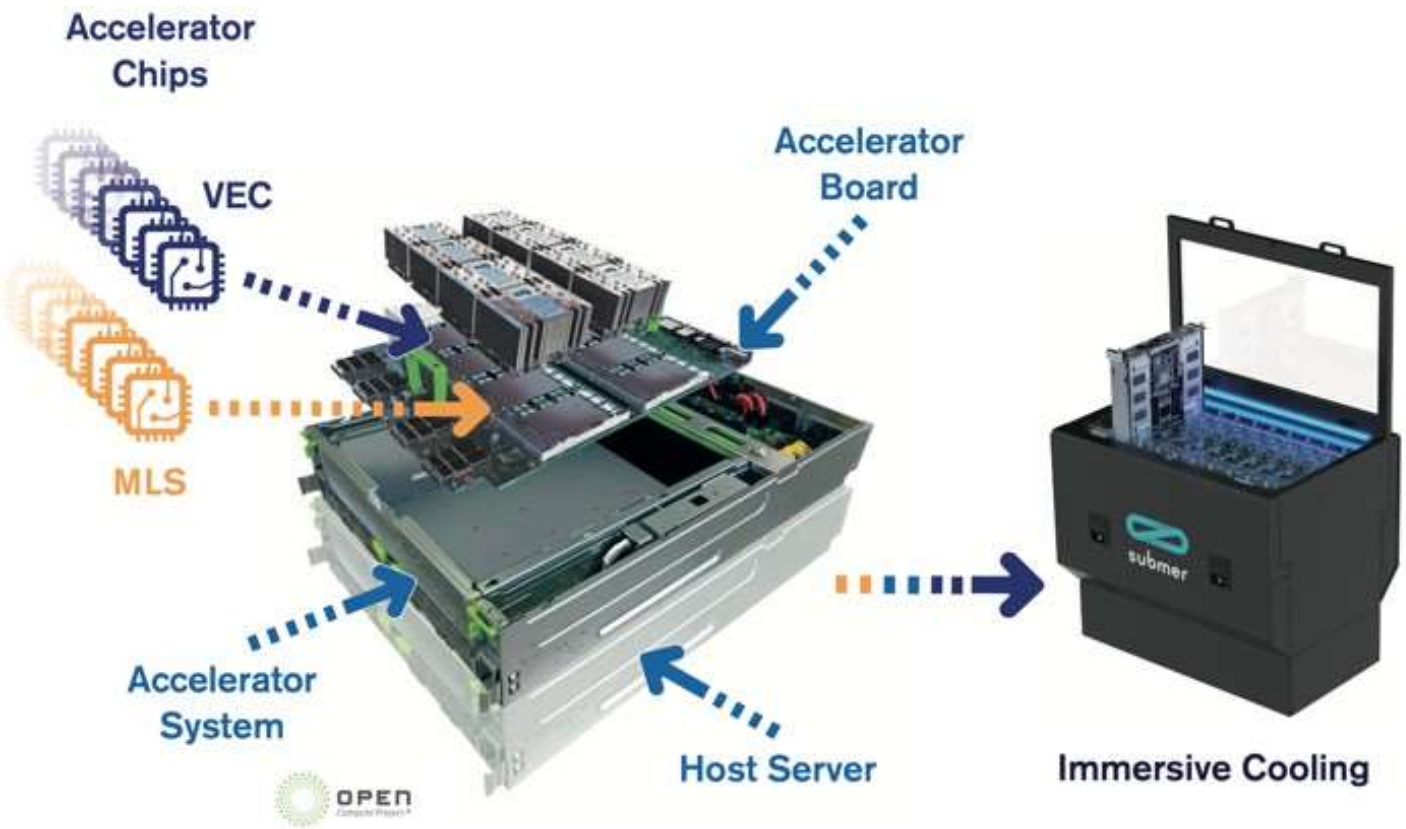
Building Open European CPUs & Accelerators



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EUPilot System Overview

- **Objective: Develop**, build & validate **1st EU-based pre-exascale Acc system** for HPC and ML/AI
- Open source tech, RISC-V, aceleradores especializados (VEC, MLS). Refrigeración por inmersión

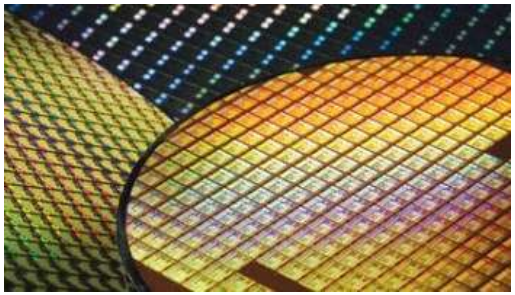


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MEEP
(MareNostrum Experimental Exascale Platform)

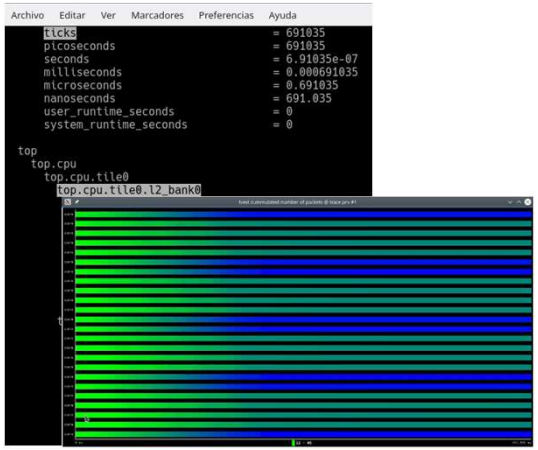
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Fabricación



Emulación HW

Simulación



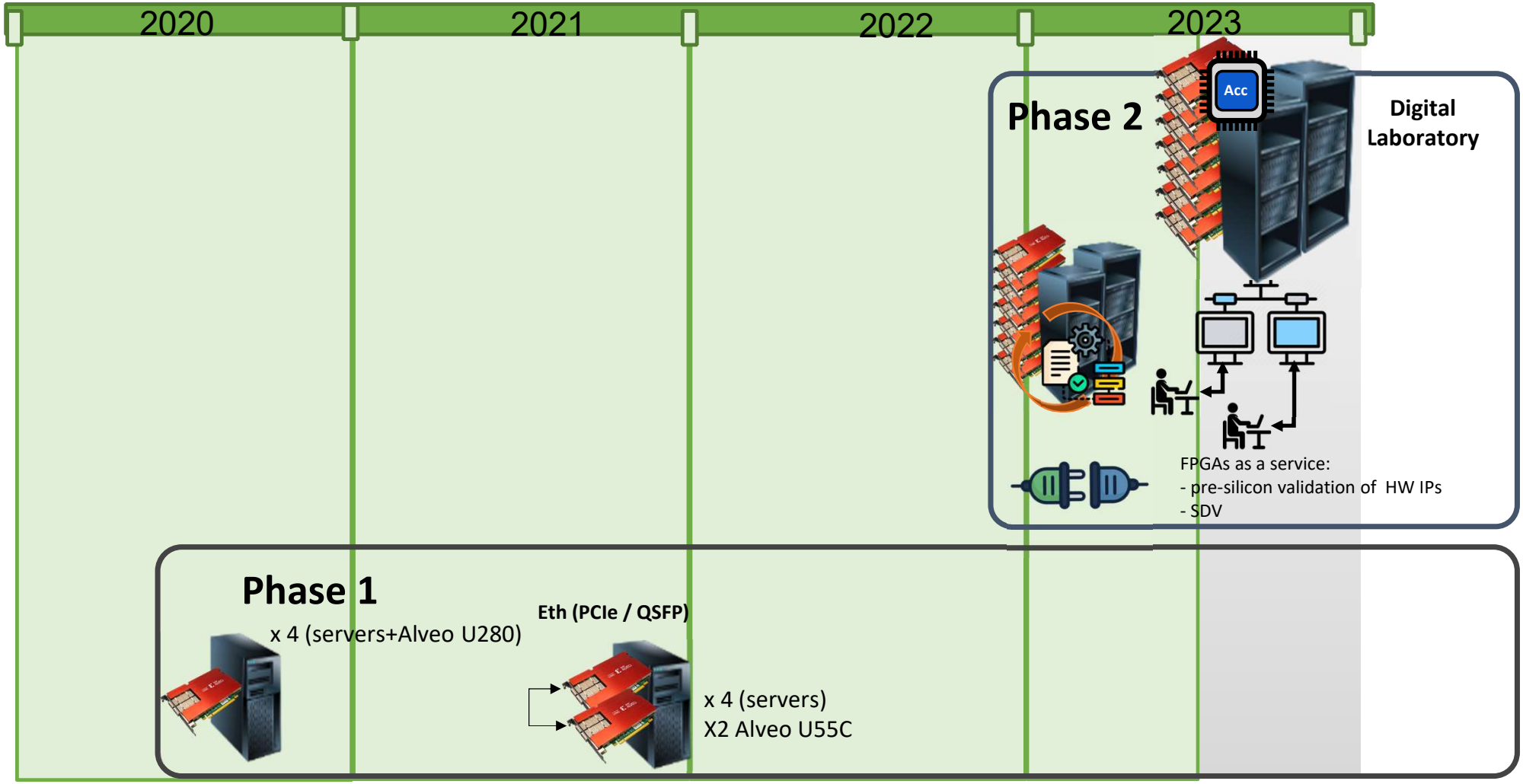
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MEEP… A Digital Laboratory

MEEP is a **flexible FPGA-based emulation platform** (infrastructure + tools) for HW/SW co-design exploration for Exascale Supercomputers and other hardware targets, based on European-developed IP. The project provides two functions:

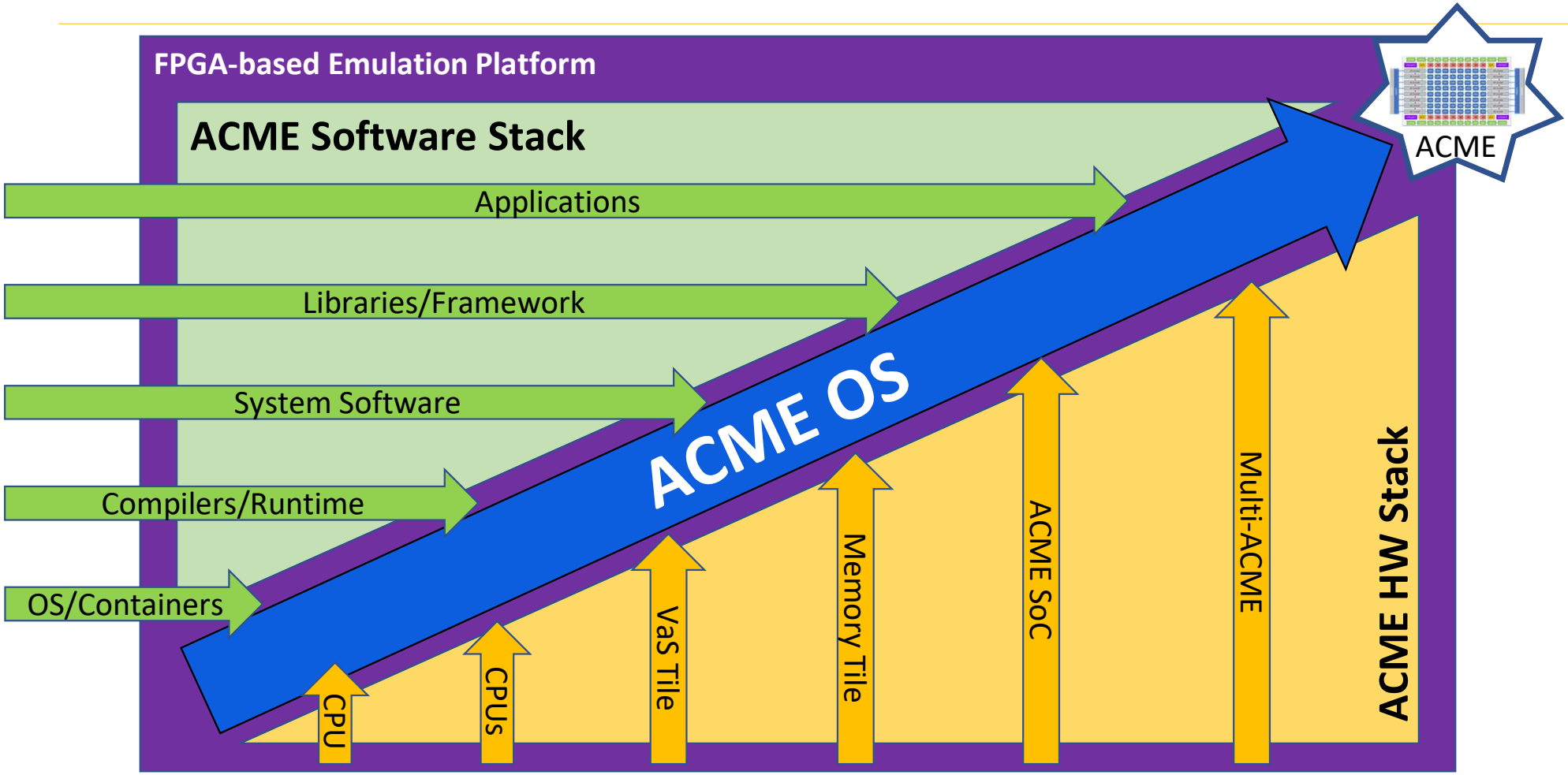
An evaluation platform of pre-silicon IP and ideas, at speed and scale.

A software development and experimentation platform to enable software readiness for new hardware.

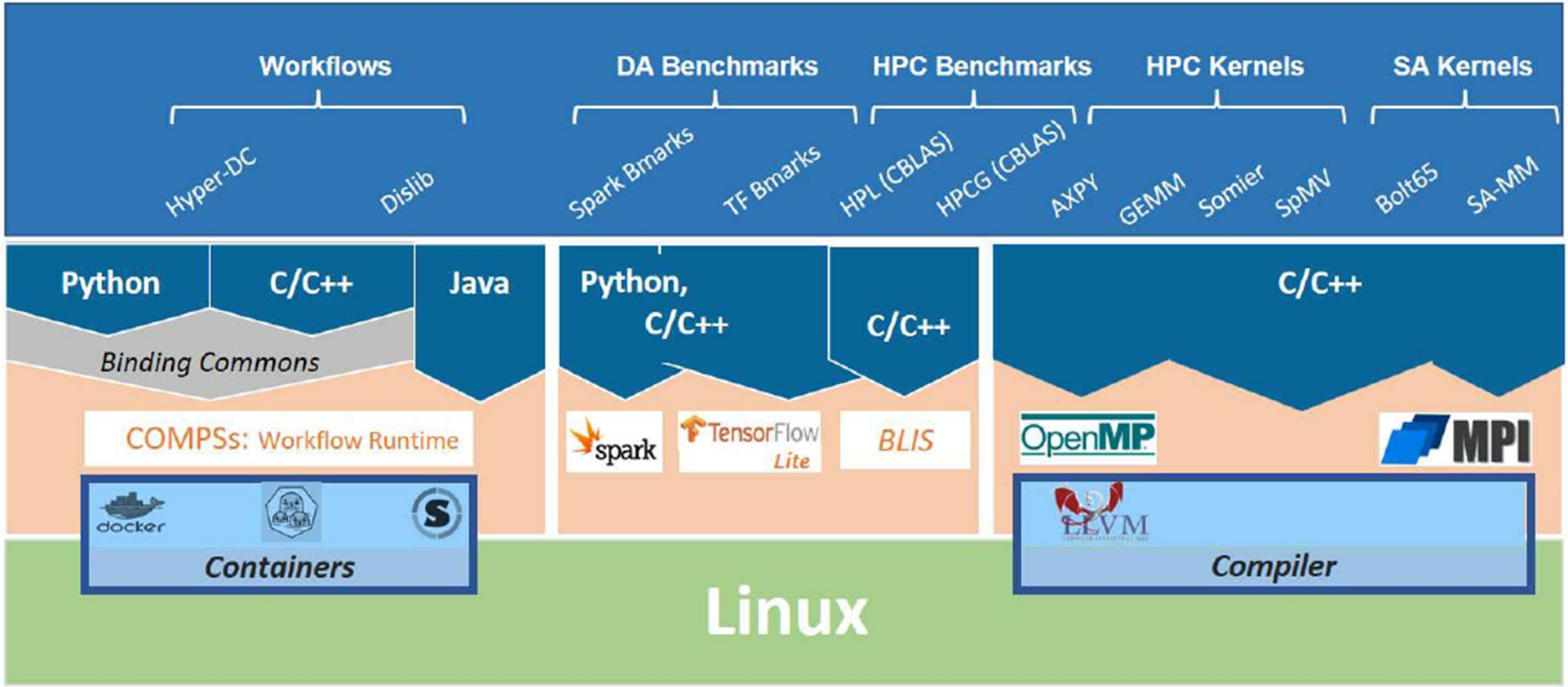


MEEP MareNostrum Experimental Exascale Platform

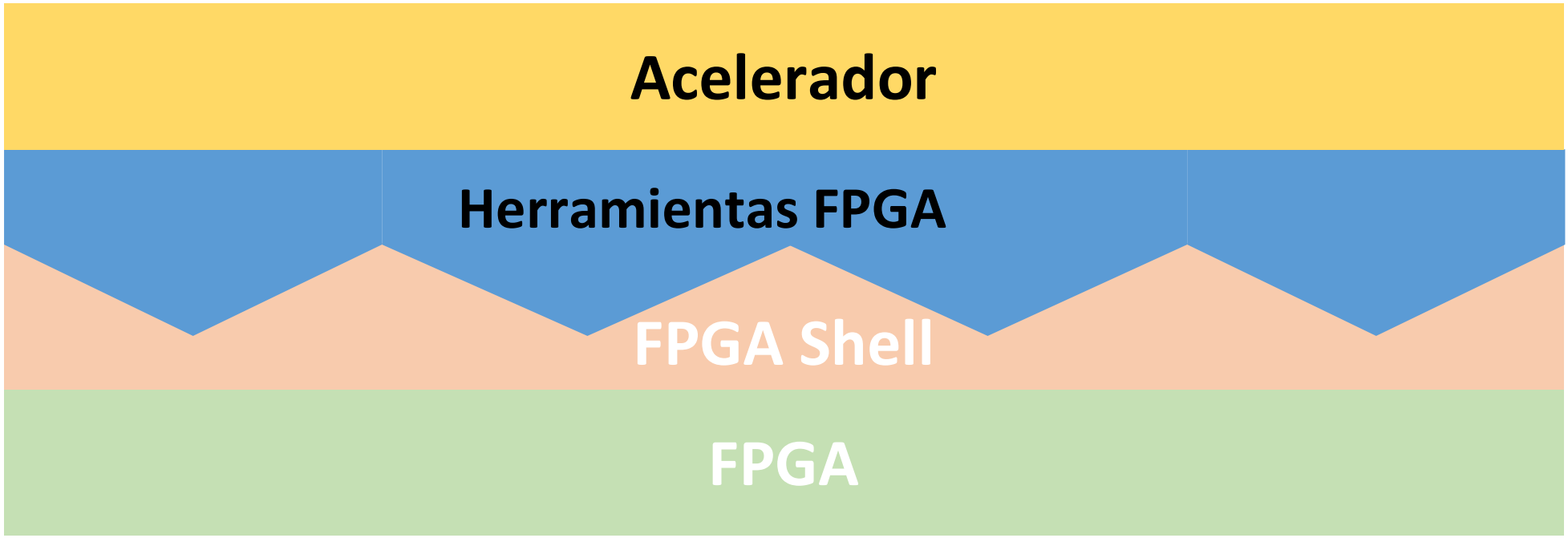
Visión general del proyecto



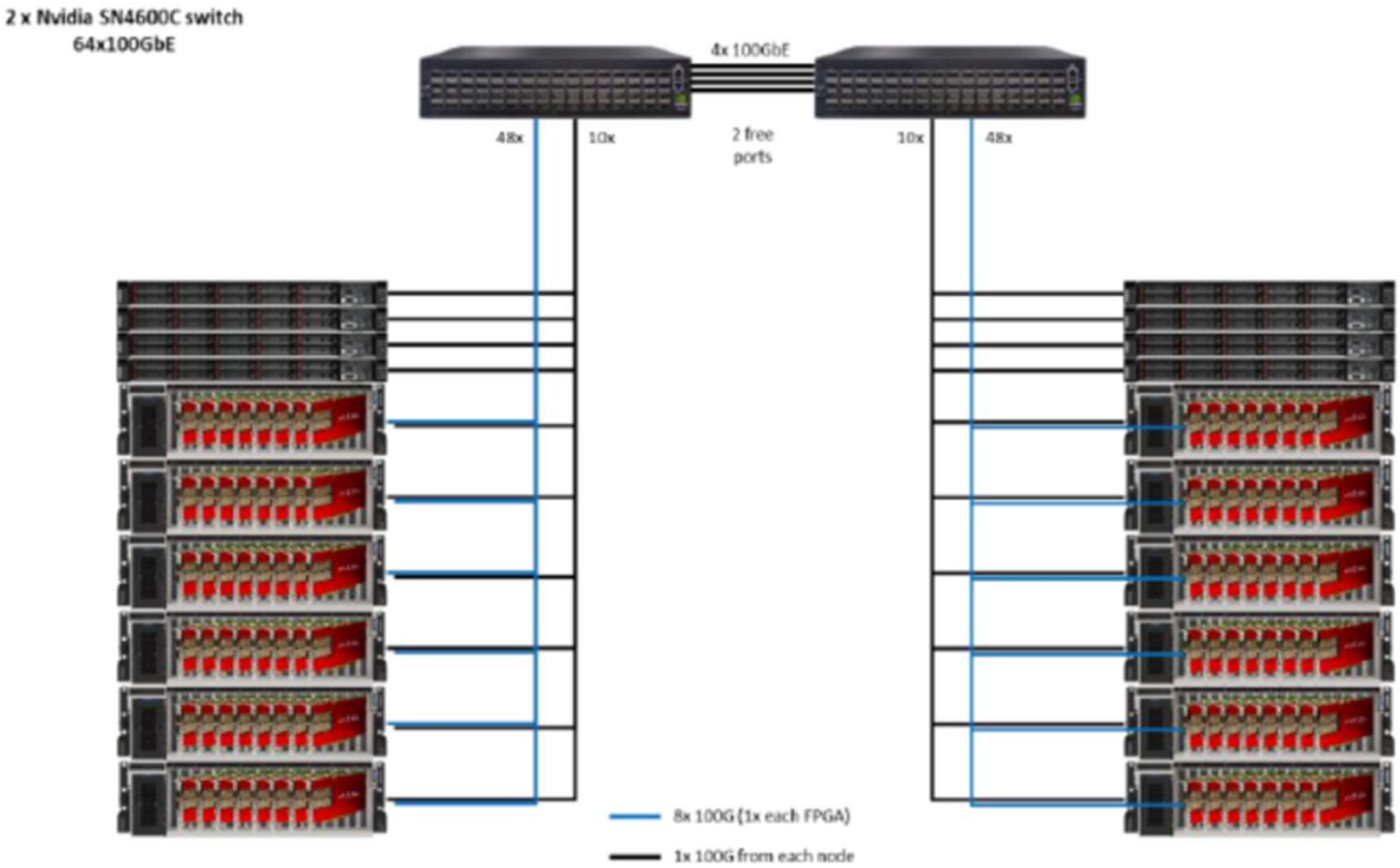
SW Stack


Workflows
DA Benchmarks
HPC Benchmarks
HPC Kernels
SA Kernels
Hyper-DC
Dislib
Spark Bmarks
TF Bmarks
HPL (CBLAS)
HPCG (CBLAS)
AXPY
GEMM
Somier
SpMV
Bolt65
SA-MM
Python
C/C++
Java
Python, C/C++
C/C++
C/C++
Binding Commons
COMPSs: Workflow Runtime
spark
TensorFlow Lite
BLIS
OpenMP
MPI
docker
Containers
LLVM
Compiler
Linux

HW Stack



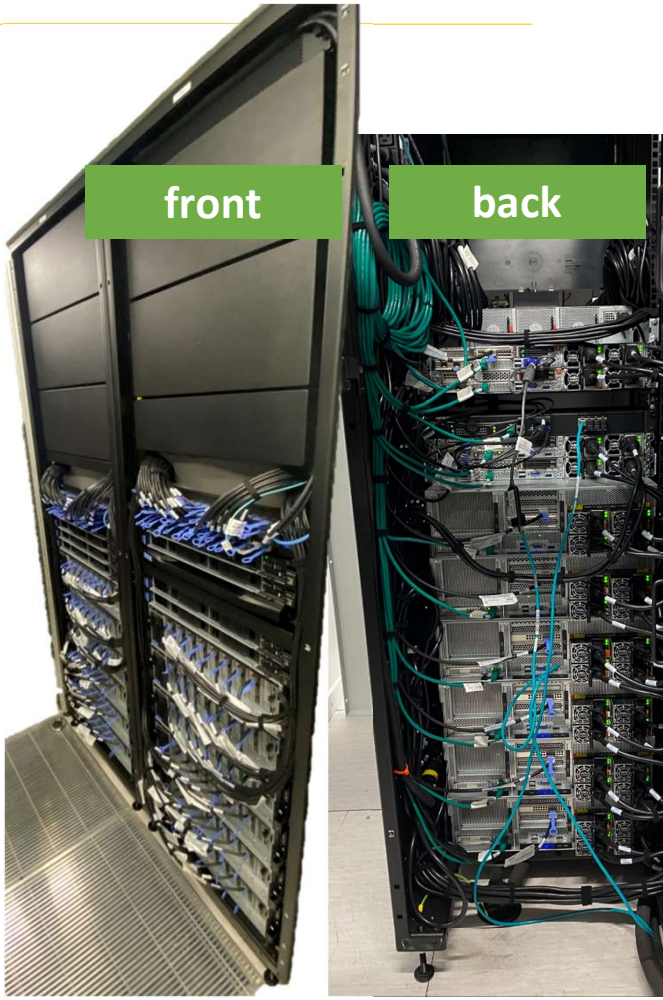
MEEP FPGA Cluster



FPGA

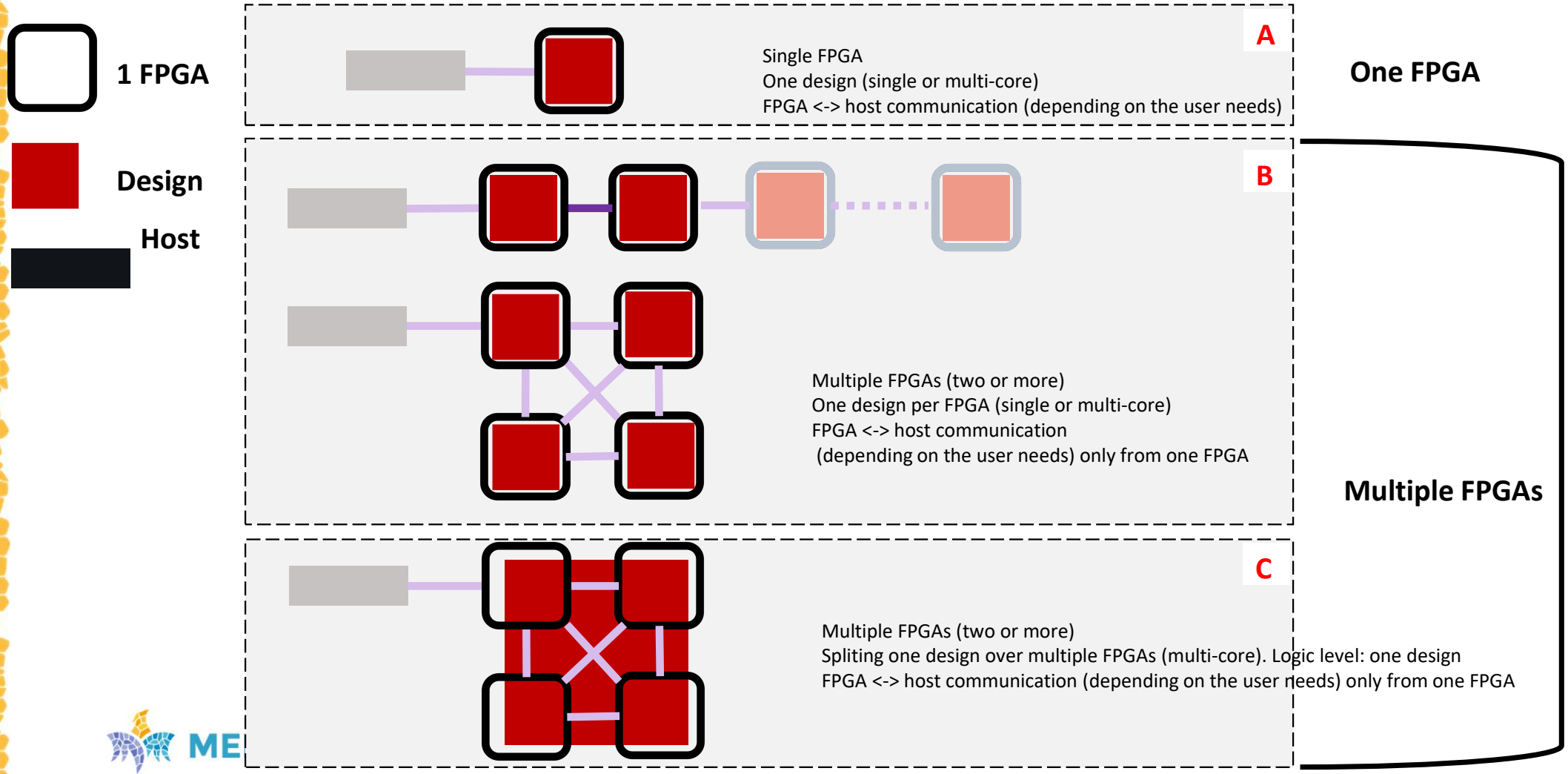
MEEP FPGA Cluster

```
[root@fpgahead1 profile.d]# ssh fpgan01
+-------------------------------------------------------------------------------------------------------------------------------------------------------------+
|    +------+        +------+        +------+        +------+        +------+        +------+        +------+        +------+        +------+        +------+    |
|    |swp18 |        |swp17 |        |swp16 |        |swp15 |        |swp14 |        |swp13 |        |swp12 |        |swp11 |        |swp10 |        |swp9  |    |
|    +------+        +------+        +------+        +------+        +------+        +------+        +------+        +------+        +------+        +------+    |
+--------^---------------^---------------^---------------^---------------^---------------^---------------^---------------^---------------^---------------^--------+
         |               |               |               |               |               |               |               |               |               |
+--------|-------+-------|-------+-------|-------+-------|-------+-------|-------+-------|-------+-------|-------+-------|-------+-------|-------+-------|--------+
|        v       |       v       |       v       |       v       |       v       |       v       |       v       |       v       |       v       |       v       |
|    +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |
|    |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |
|    |QSFP0 |    |   |QSFP0 |    |   |QSFP0 |    |   |QSFP0 |    |   |QSFP0 |    |   |QSFP0 |    |   |QSFP0 |    |   |QSFP0 |    |   |QSFP0 |    |   |QSFP0 |    |
|    |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |
|    +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |
|                |               |               |               |               |               |               |               |               |               |
|    +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |
|    |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |
|    |QSFP1 | <----> |QSFP1 |    |   |QSFP1 |  <----> |QSFP1 |    |   |QSFP1 | <----> |QSFP1 |    |   |QSFP1 | <----> |QSFP1 |    |   |QSFP1 | <----> |QSFP1 |    |
|    |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |   |      |    |
|    +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |   +------+    |
|  onic180s0f0   |  onic179s0f0  |  onic204s0f0  |  onic205s0f0  |  onic26s0f0   |  onic25s0f0   |  onic51s0f0   |  onic52s0f0   |  onic53s0f0   |  onic54s0f0   |
|                |               |               |               |               |               |               |               |               |               |
|     +-+        |    +-+        |    +-+        |    +-+        |    +-+        |    +-+        |    +-+        |    +-+        |    +-+        |    +-+        |
|     | |        |    | |        |    | |        |    | |        |    | |        |    | |        |    | |        |    | |        |    | |        |    | |        |
|     +-+        |    +-+        |    +-+        |    +-+        |    +-+        |    +-+        |    +-+        |    +-+        |    +-+        |    +-+        |
|  USB-UART-     | USB-UART-     | USB-UART-     | USB-UART-     | USB-UART-     | USB-UART-     | USB-UART-     | USB-UART-     | USB-UART-     | USB-UART-     |
| XFL1RIVW202O   |XFL1QTI2Z0CV   | XFL1L43EGBAE  | XFL1BZIEUP0P  | XFL105L3VVVU  |XFL1KEQBL4IM   | XFL1DF0P10SS  | XFL1VYH4S0UR  | XFL1DF0P10SR  | XFL1VYH4S0UP  |
+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+----------------+
```

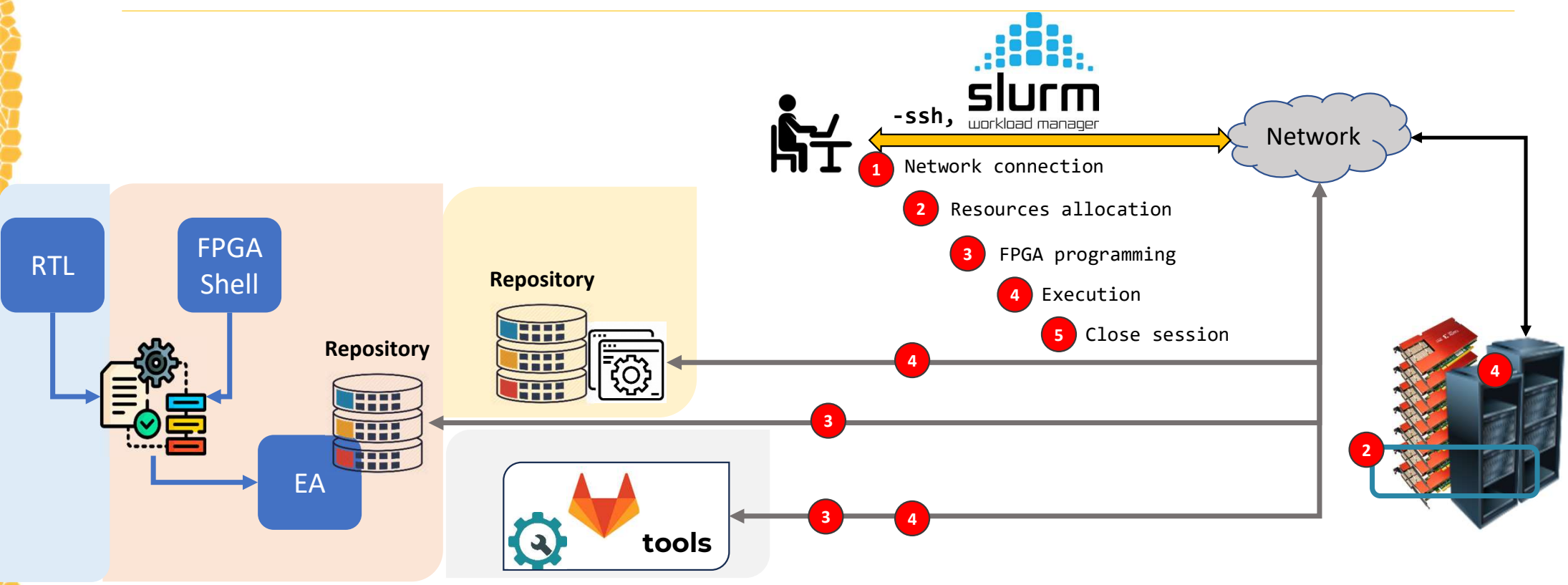


Total FPGA Power	
#FPGAs (Alveo U55c)	**96**
HBM2 Capacity	**1.5TB**
#LUT	**125M**
Max ASIC size (#cells)	**750M**

FPGA cluster: escenarios de configuración



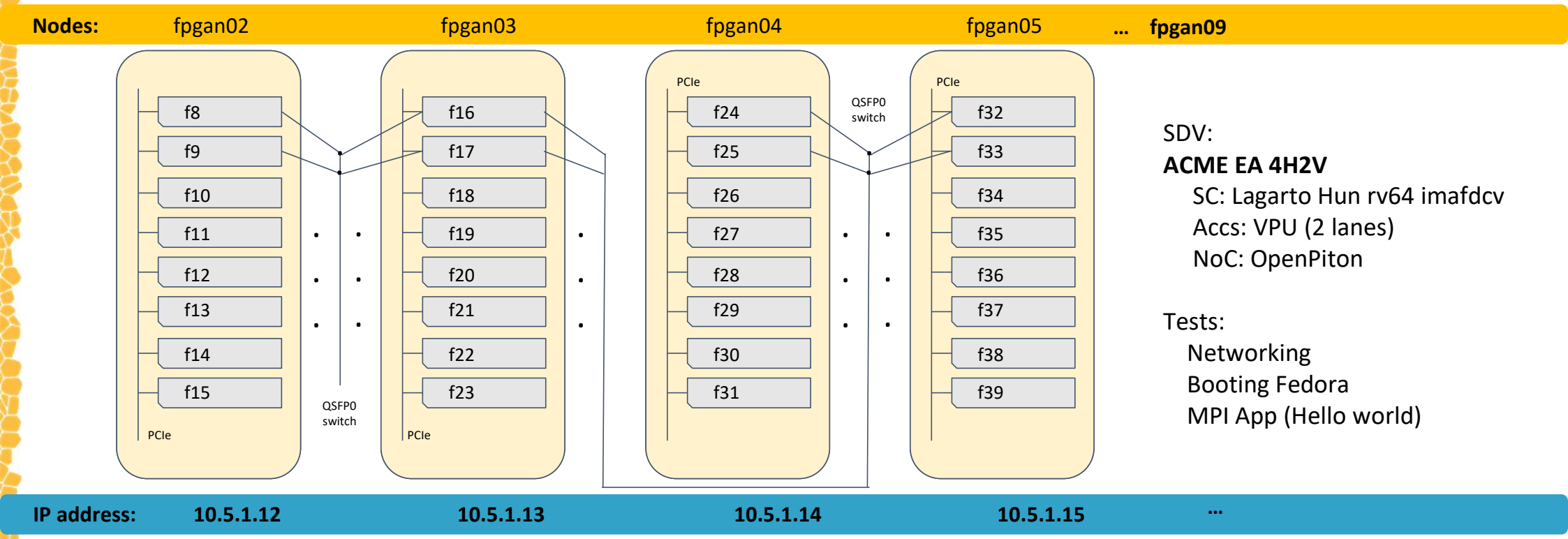
Acceso al Cluster



MEEP Demo: Network tests

Execution environment: MEEP system

Using nodes fpganode02: fpganode09 → FPGAs f8 to f71



SDV:

ACME EA 4H2V

- SC: Lagarto Hun rv64 imafdcv
- Accs: VPU (2 lanes)
- NoC: OpenPiton

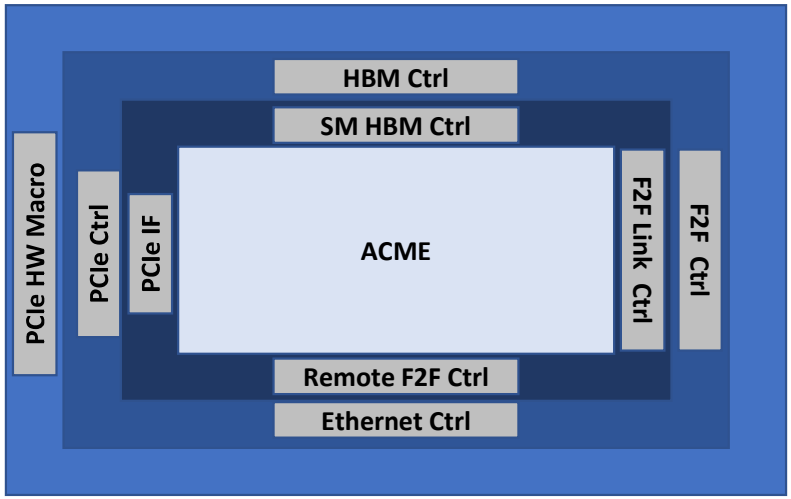
Tests:

- Networking
- Booting Fedora
- MPI App (Hello world)

MEEP MareNostrum Experimental Exascale Platform

RISC-V nodes on FPGAs: IP addresses from **10.5.1.151** to **10.5.1.246**

MEEP FPGA Shell



- FPGA Shell Interfaces:

Communication IPs	
PCIe	xdma / qdma
UART	
HBM	Multi-MC, Memory exploration
Remote F2F (10/100GbE)	10/100GbE
F2F link (Aurora)	Aurora with DMA
Others	InfoROM

- FPGA Shell features:
  - SW support (Drivers (ONIC), tool (images))
  - FPGA support (FPGA Shell generation)
  - Scripts & TCLs (e.g.,timing policies)
  - FPGA Flow (Gitlab CI/CD & Vivado *Non-Project* mode)

Open Source repo
https://github.com/MEEPproject/

MEEP MareNostrum Experimental Exascale Platform

Rapid FPGA Prototyping utilizando el Shell+herramientas

Config file (accelerator_def.csv)

```
EANAME=acme
PCIE,yes,pcie_axi,,PCIE_CLK,pcie_clk,pcie_rstn,dma,0,,02
DDR4,no,mem_axi
HBM,yes,mem_axi,AXI3-256,CLK0,0x0,mem_calib_complete,00
HBM,yes,ncmem_axi,AXI3-256,CLK0,0x0,mem_calib_complete,01
HBM,no,mcx_axi0,AXI3-256,CLK0,0x0,mem_calib_complete,04
HBM,no,mcx_axi1,AXI3-256,CLK0,0x0,mem_calib_complete,05
HBM,no,mcx_axi2,AXI3-256,CLK0,0x0,mem_calib_complete,06
HBM,no,mcx_axi3,AXI3-256,CLK0,0x0,mem_calib_complete,07
AURORA,no,raw,<name>
UART,yes,uart_axi,,CLK0,0x0,normal,uart_irq
ETHERNET,yes,eth_axi,AXI4LITE-64,CLK0,100Gb,eth_irq,qsfp1,hbm,29
BROM,no,sram_axi,,CLK0,0x0,initFile.mem
BRAM,no,sram_axi,,CLK0,0x0,none
CLK0,50000000,chipset_clk
CLK0,,eth_axi_aclk,eth_axi_arstn,LOW
CLK0,,mc_clk,mc_rstn,LOW
CLK1,100000000,vpu_clk
GPIO,5,pcie_gpio,0x00
ARST,LOW,ExtArstn
```

Make

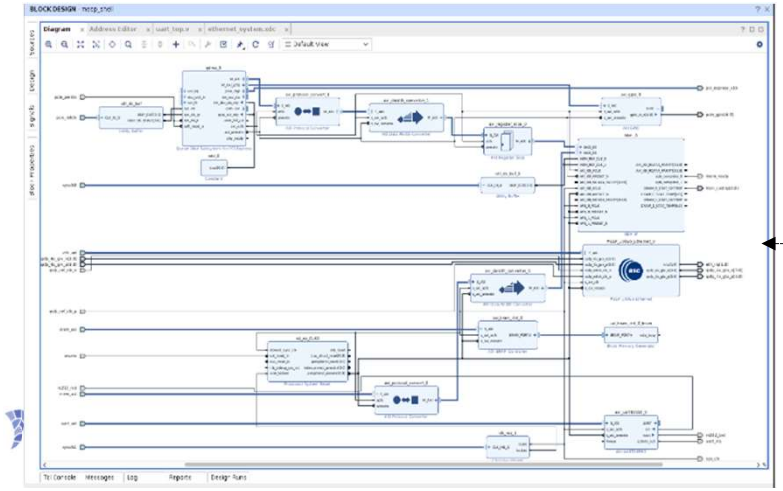
Transparent from user

MEEP Shell Scripts (**TCL**)
- shell_aurora
- shell_hbm
....

Vivado Project
(automatically generated)

bitstream

FPGA Tools



PCI (XDMA, QDMA)

SDRAM (HBM, DDR)

Ethernet (10/100 GBe)

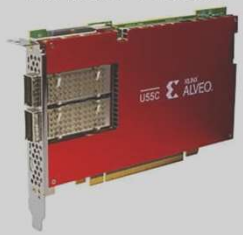
UART
Aurora-DMA

InfoROM

```
./build_pci_drivers.sh
./load_bitstream.sh
./load_fedora.sh
./fpga_test.sh
./boot_acme.sh
./boot_ariane.sh
....
```

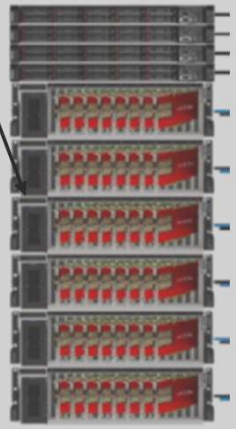
Supported Hardware

Alveo U55c



Alveo U280

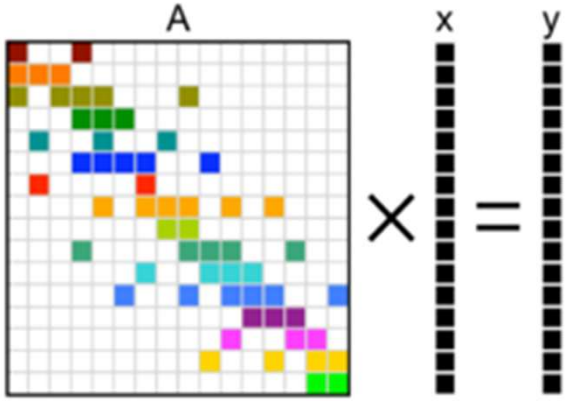
FPGA Cluster



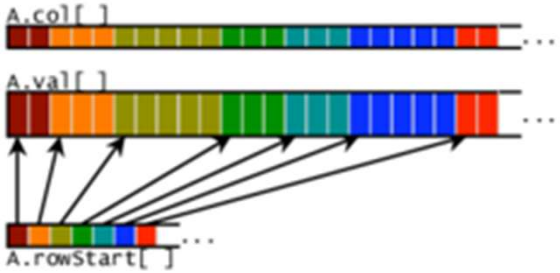
Aceleradores para sparse

Acelerador

- DGEMM → Multiplicación densa de matrices → Alta localidad espacial y temporal
- SpMV → Multiplicación sparse de matriz x vector → Dispersión de datos, poca localidad



(a)
algebra conceptualization

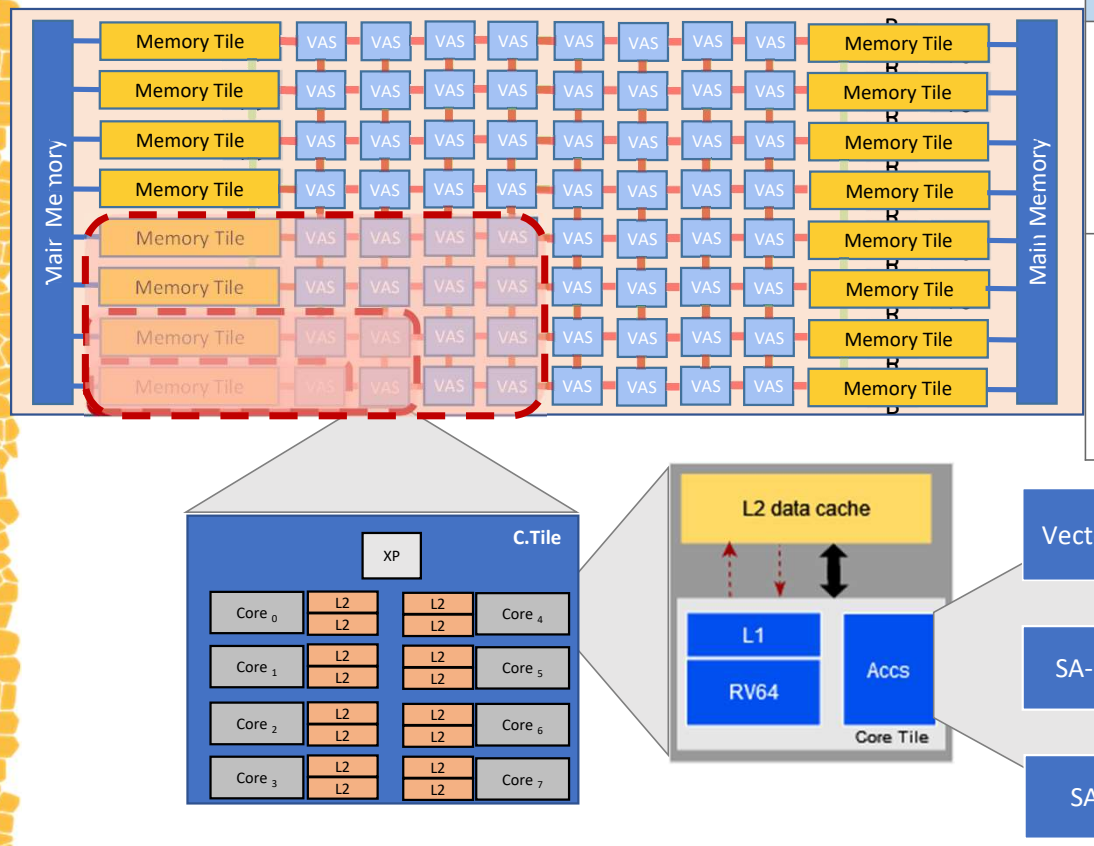


(b)
CSR data structure

```
for (r=0; r<A.rows; r++) {
  double y0 = 0.0;
  for (i=A.rowStart[r]; i<A.rowStart[r+1]; i++){
    y0 += A.val[i] * x[A.col[i]];
  }
  y[r] = y0;
}
```

(c)
CSR reference code

Architecture Outline



Performance limiter	Top-level Architecture	Micro-architecture	Components	Component features
Compute-bound	VAS Tile	Accelerator cores	Scalar core Co-processors - VPU, - SA-HEVC - SA-NN	• Specialized accelerators (VPU, SA-HEVC, SA-NN) • Common co-processor shell • ISA upgrade • Custom instructions • New interlane interconnect
Memory-bandwidth-bound	Memory Tile	MCPU MicroEngines	2-Level VRF MCPU Flexible L2/L3 Memory Ctrl.	• Streaming Vector Register File • Real & Virtualized vector length • Memory operation off-loading • Flexible memory hierarchy • Smart memory addressing • Sparse to dense transformation • Smart data management

MEEP MareNostrum Experimental Exascale Platform

Acelerador

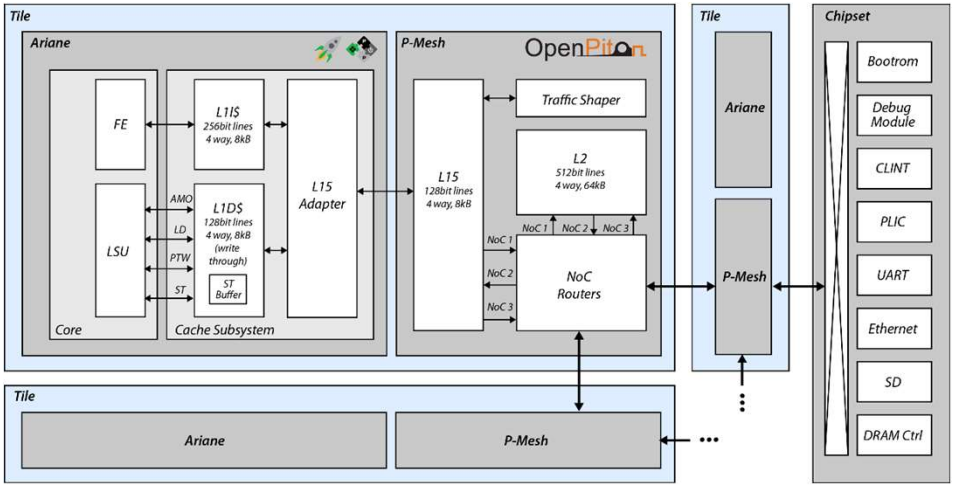
open source hardware

RTL → **Verificación** → **Validación**

OPENHW GROUP — PROVEN PROCESSOR IP

UVM

OpenPiton



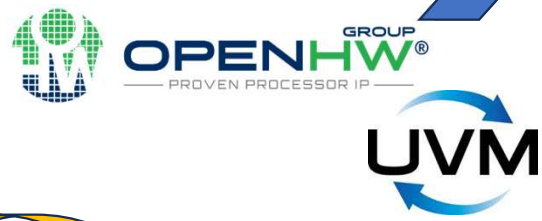
MEEP — MareNostrum Experimental Exascale Platform

https://github.com/PrincetonUniversity/openpiton

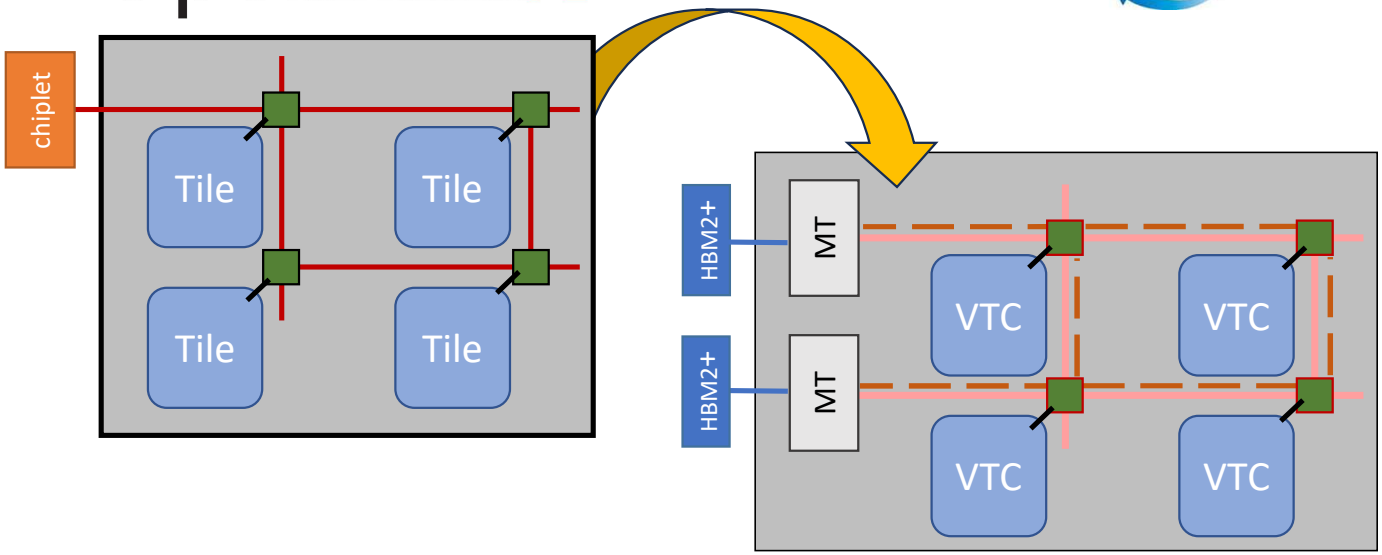
https://github.com/openhwgroup/

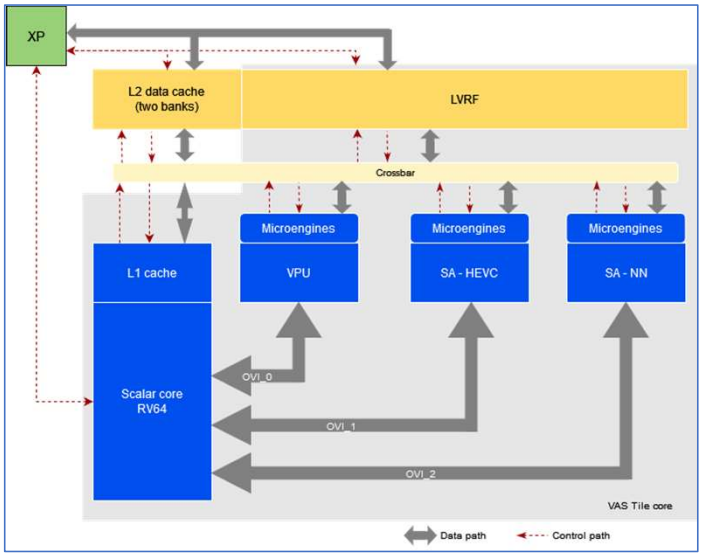
Acelerador

RTL → **Verificación** → **Validación**



OpenPiton

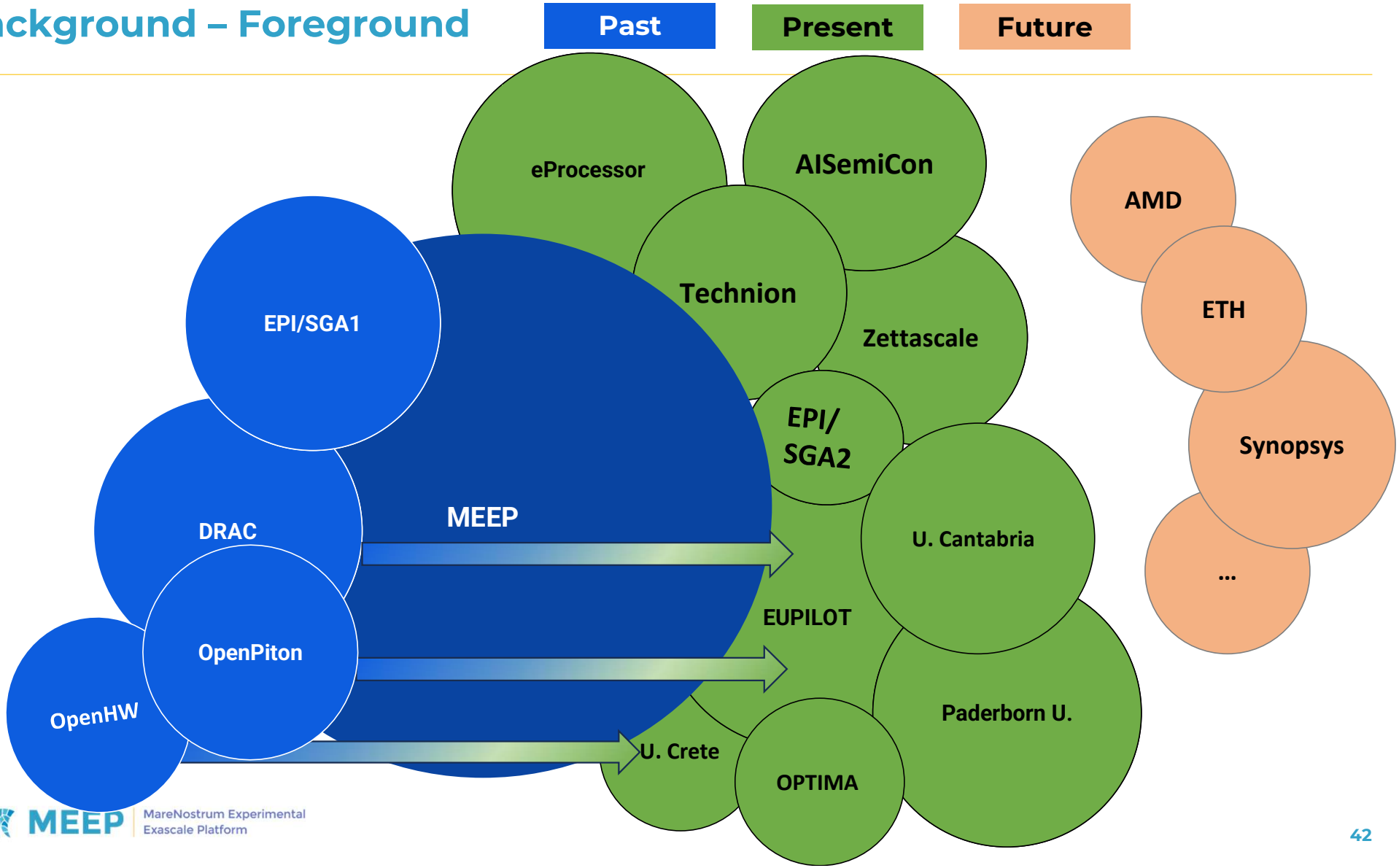




https://github.com/PrincetonUniversity/openpiton

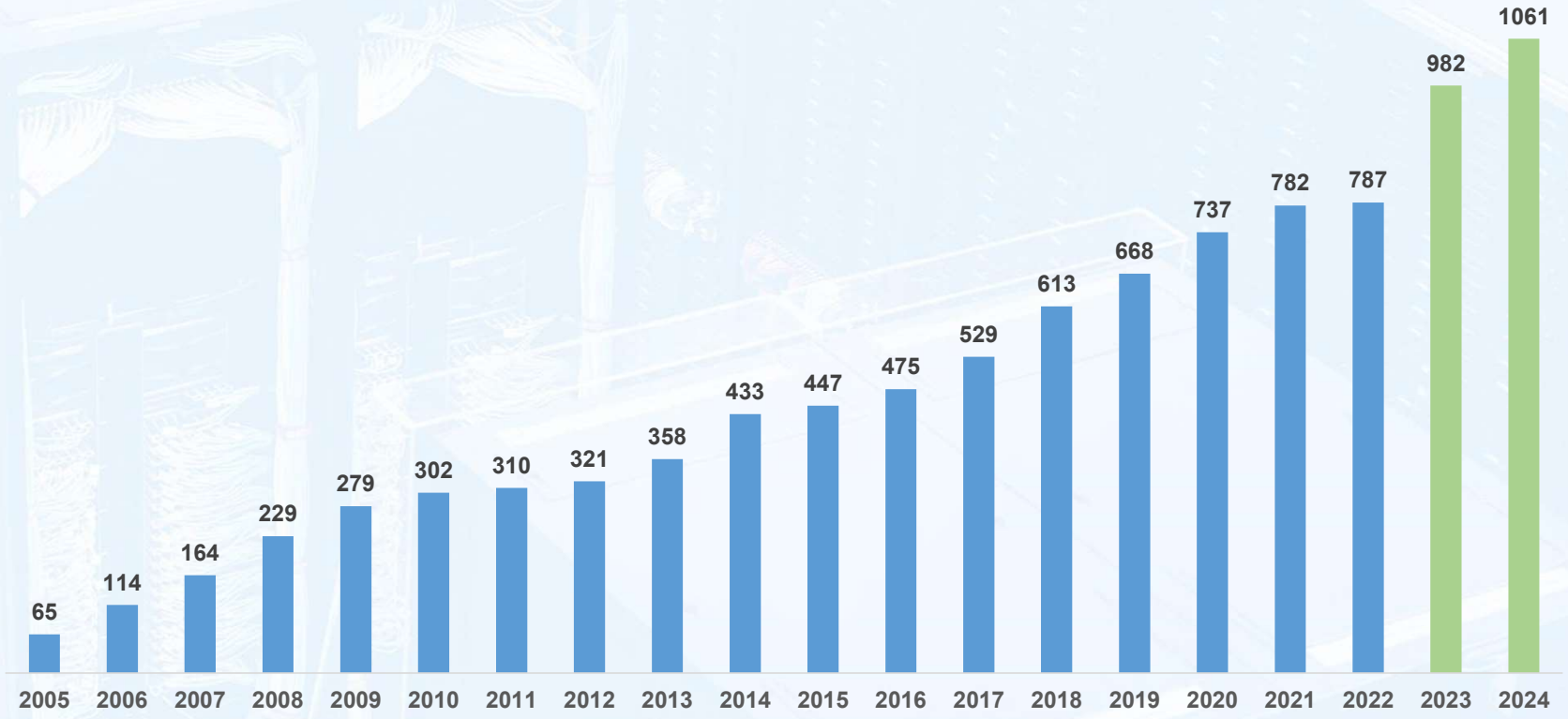
https://github.com/openhwgroup/

Background – Foreground



MEEP MareNostrum Experimental Exascale Platform

Evolución del personal BSC (2005-2024)



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

¡Gracias!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

teresa.cervero@bsc.es