HPC Status and Perspectives in Europe

Piero Vicini

Scuola Dottorato Nazionale Tecnologie della Ricerca nelle scienze fisiche e astrofisiche

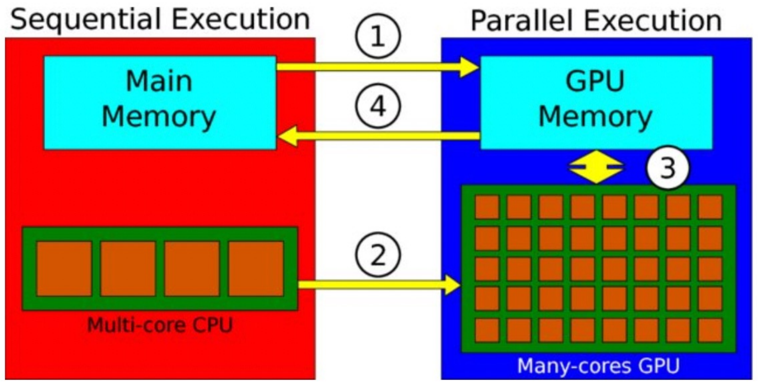
Feb. 2025

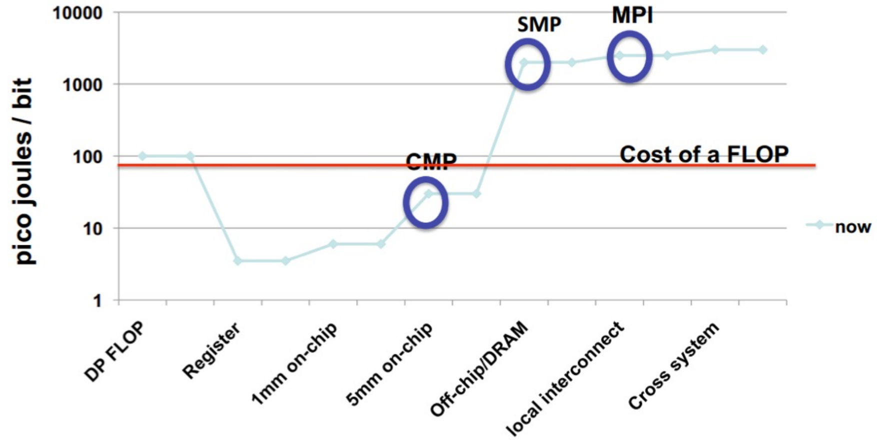
Introduction

- A general and all inclusive talk on HPC is quite impossible
  - Too many technologies, too many “pillars", too many (and divergent) needings and solutions
  - I’m not an all-round expert
- So this talk is a personal and incomplete review of few topics of interest to me and (hopefully) to the scientific computing community @INFN
  - Exascale HPC characteristics and challenges
  - Status of technological main components: CPU,GPU,DPU,FPGA,…
  - HPC infrastructure and R&I in Europe: status and funding opportunities
  - Post Exascale challenges
- More details and insights in the next talks

HPC is alive and still fighting with us...

- Traditional use:
  - to solve complex problems at large scale faster and more cost-effectively
  - fostering scientific discoveries and innovative technological advances
    - simulations and modelling for product development, new materials, weather&climate, energy analysis, aerospace, oil&gas,...
    - healthcare: drug development, drug analysis, real-time personal medicine...
    - fundamental science: LQCD, Montecarlo for HEP, material simulation, complex systems, neuromodeling...
- New fields of application pumping up its use
  - IoT & Big Data analysis
  - LLM training and operating
  - Hybrid Quantum computing (HPC as the interface to real/classic world)
  - Quantum ML ...
- But a number of open issues
  - maximize the computational efficiency
    - does not exist a “one fits all” computing architecture. Any application or class of applications has different peculiarities (fp vs int, compute bounded vs memory bounded...)
    - different data precisions: FP64/32 for HPC, FP16/8/4 for AI workloads
    - the current winner model “CPU+Accelerator” (GPU, FPGA, ASIP,..) is scalable and sustainable?
  - learn to exploit the “chiplet” approach
  - power&density i.e. minimize the ratio W/OPs while increase system density
    - this is true from embedded to exascale...
  - needs for low latency, high bandwidth, high throughput
    - memory architecture
    - interconnect network
  - not to mention
    - software and programming model, resiliency, economic sutainability...





Exascale challenges

Frontier US DOE Exascale supercomputer @ORNL (number 1 in top500 List)

- Power consumption
  - Currently 21 MW rating for Frontier system
  - GPU acceleration as critical efficiency factor, CPUs lagging behind in efficiency
  - Going is getting harder …
- Speed and energy of data movement
  - Interconnect networks, on-node I/O buses
  - Steady progress, yet outmatched by compute capabilities
  - This also includes I/O systems
- Fault tolerance
  - Improvements in component reliability & integration has delayed onset of this problem
  - New storage technologies help with checkpointing
- Billion-way parallelism
  - Nearly there: Frontier has 5.2×10^8 stream processors
  - Address problem by hierarchical parallelism: O(10K) nodes, O(100) CPU cores each, O(100K) streaming nodes
  - BUT: scaling general applications in an efficient way is still a major challenge!

ORNL

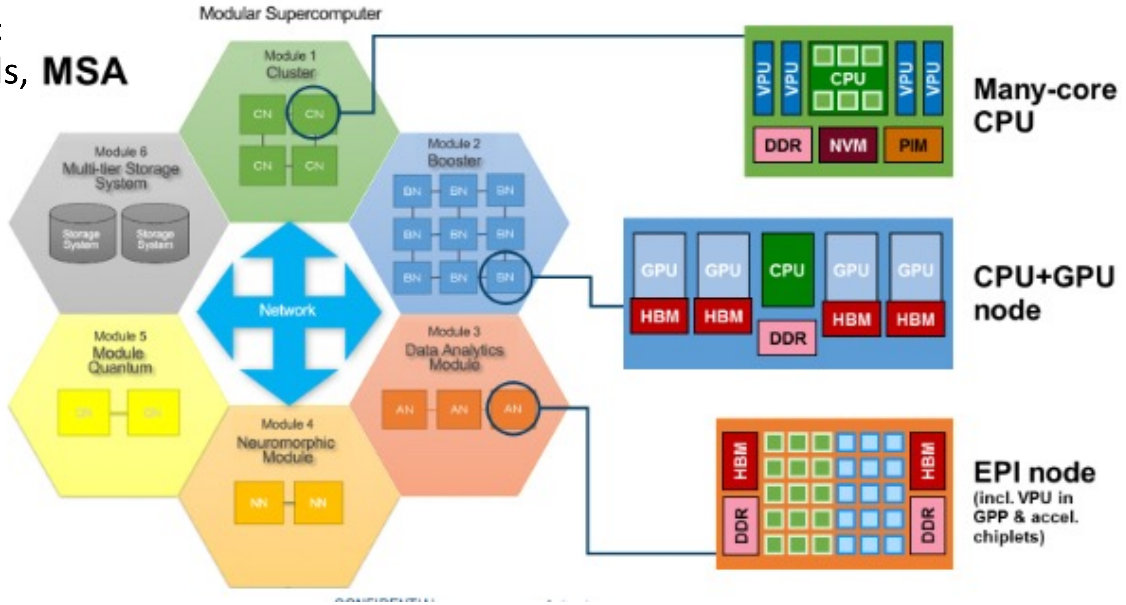
US DoE „Frontier“ System
9284 AMD EPYC 7453s CPUs, 594,176 cores
37136 AMD Instinct MI250x GPUs, 8,169,920 compute units (64 stream processors & 4 matrix processors each)
HPE Slingshot interconnect with 4×200 Gb/s per node
1.2 Exaflop/s for HPL code

Configuration for November 2023 Top500 List

J. Carretero talk @HIPEAC 2024

Heterogeneity, new players and convergency

- Modular heterogeneous system (Low level Computer continuum...)
  - The answer to the un-availability of “one fits all” architecture
  - Aggregation of different modules specialized for different computational tasks
  - It can be valid at any scale
    - MSA Modular Supercomputer Architecture
  - Basic ingredients CPU, computational accelerators (GPU, DPU, ASIP), networks, programmable components for the implementation of accelerators for specific computational tasks (FPGA/ASIC for ML or data analytics), programming models, integrated OS, real time schedulers, storage
  - Target is Data Center, on premises, on cloud or mixed approach
  - New entries to face emerging computing applications
    - AI booster
    - Quantum booster
    - Neuromorphic booster

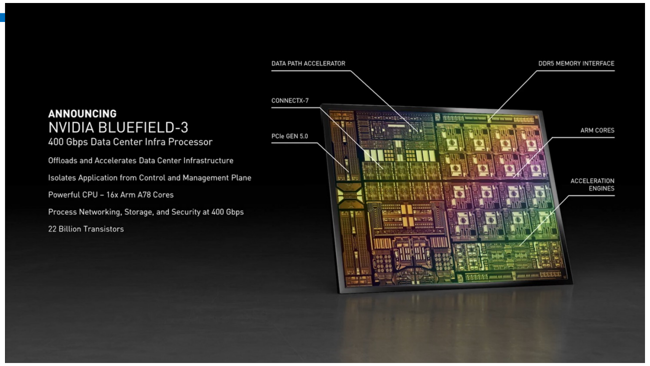


- Open issues:
  - Network is critical and does not exist one-for-all network architecture
  - Interfaces to individual computational modules
  - Different modules have different technological maturity, different complexity, different peculiarities of the data and computing task, different characteristics of the interfaces (type and timing) etc...
  - No clear solution for orchestration and programming model
  - Today lack of real and feasible application use cases (i.e Quantum...)

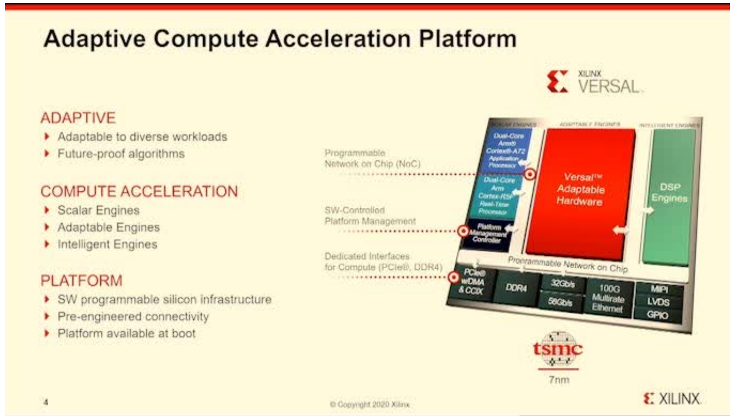
We are still far away, and more research is needed...

Big players become heterogenous and convergent

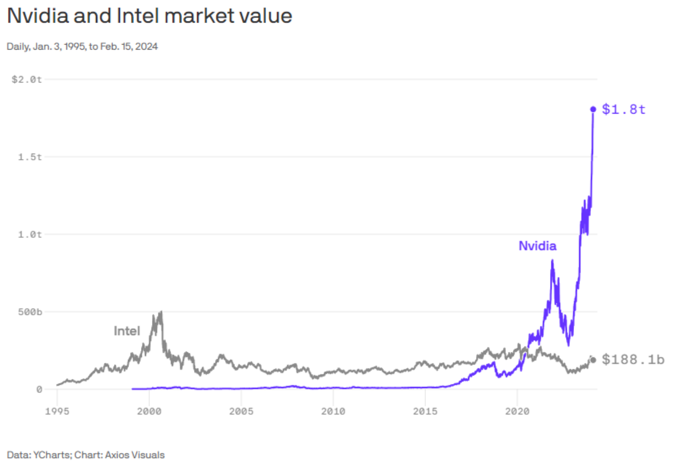
- 2016: INTEL(CPU) buys ALTERA(FPGA) for 17 B$!!!
  - Aiming to design tightly integrated CPU+FPGA devices at die-level
  - Up to now, not a big success.
  - Today DPU/IPU architecture is slowly emerging
  - INTEL oneAPI release: a programming framework for heterogenous architecture based on DPC++ (Data Parallel C++) to implement (partial) SYCL support (https://oneapi.io/)
  - No more Ponte Vecchio (again???) Welcome Gaudi…



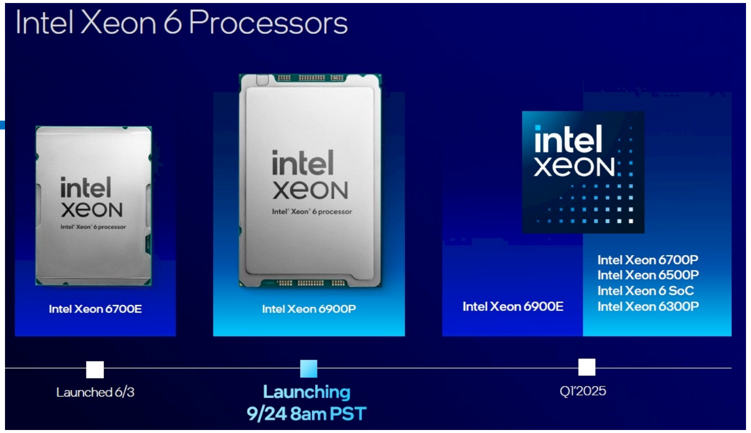
- 2020: NVIDIA (GPU) buys Mellanox (Infiniband Network for HPC) with 7 B$
  - GPU and Network integration to build in-house scalable mesh of GPU cluster interconnected via NVLink
  - status ongoing
  - DPU SoC heterogenous (ex BlueField) for task computing acceleration
  - *Issue: low number of competitors of network providers*



- 2022: AMD (CPU) buys Xilinx (FPGA) per 35 B$ con
  - Same goal of INTEL/Altera
  - Integrate CPU, GPU, FPGA SoC for HPC and ML inference tasks)
  - *Issue: no more independent competitors of FPGA*
- More recently
  - Quantum Computing large and small players
  - Impressive hype and huge interest in Europe (public and private funding)
- AI&ML gains visibility and request many many many resources →
  - NVidia reaches N*100*billions and starts skyrocketing and swinging....
  - *Issue: GPU cost becomes unsustainable*
- Low Power CPU:
  - rumours about big players consortium to invest in ARM to avoid "single owner" model



Basic ingredients: CPU



INTEL XEON: the perfect example

- CPU "issues" limiting its scalability
  - "small" number of cores due to the architectural model (shared mem)
  - Limited clock speed (from many years…)
  - Cores rich in features but with "low" performance: good for average user not for HPC at large scale.
  - Huge memory banks (caches) →
    - less transistors for computing
  - High ratio power/performance (TDP)
  - High cost/performance

Addressing Unique Workload Requirements

P-core			Workloads	E-core		
	Modeling and simulation	CAE	HPC			
			Web & microservices	Cloud-native	Consumer digital services	Application DevOps
CRM, ERP	Big data	In-memory analytics	Database & analytics	Unstructured databases	Scale-out analytics	
Generative AI	Deep learning Machine learning	Inference	AI			
HCI	Virtualization	Storage	Infrastructure & storage	Storage		
	CDN	Media & gaming	Networking	Network microservices	Cloud-native CDN	5G core
	Video	Edge analytics	Edge	Virtual protection relay		

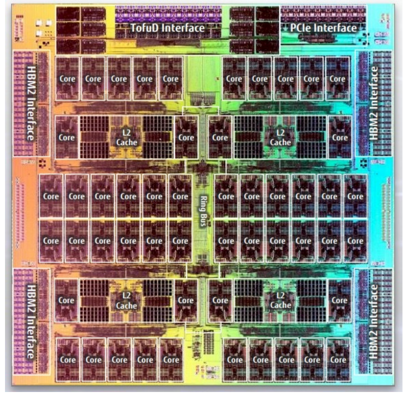
Xeon 6 P-Core	Clock Speed	Cores / Threads	L3 Cache	TDP (Watts)	Max Sockets	1K Tray Unit Price	Raw Clocks	$ / Raw Clock	Rel Perf	$ / Rel Perf
6900 Series						"Granite Rapids"				
6980P	2.0 GHz	128 / 256	504	500	2	$24,980	256 GHz	$97.58	62.00	$402.88
6979P	2.1 GHz	120 / 240	504	500	2	$24,590	252 GHz	$97.58	61.04	$402.88
6972P	2.4 GHz	96 / 192	480	500	2	$16,059	230.4 GHz	$69.70	55.80	$287.77
6952P	2.1 GHz	96 / 192	480	400	2	$14,051	201.6 GHz	$69.70	48.83	$287.77
6960P	2.7 GHz	72 / 144	432	500	2	$8,029	194.4 GHz	$41.30	47.08	$170.53

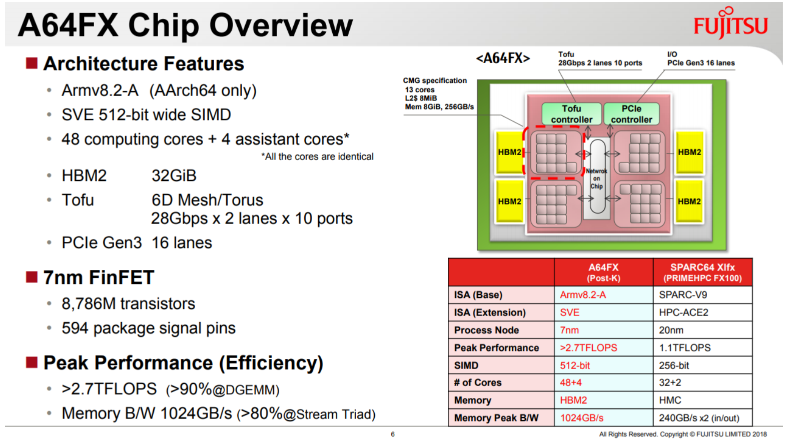
https://www.intel.com/content/www/us/en/products/details/processors/xeon/xeon6-product-brief.html

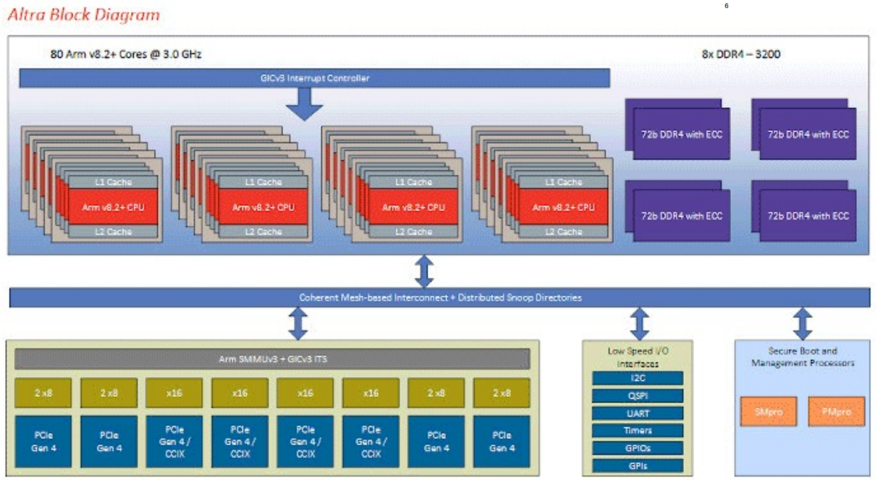
CPU low power (ARM)

Beyond the x86 mainstream: ARM

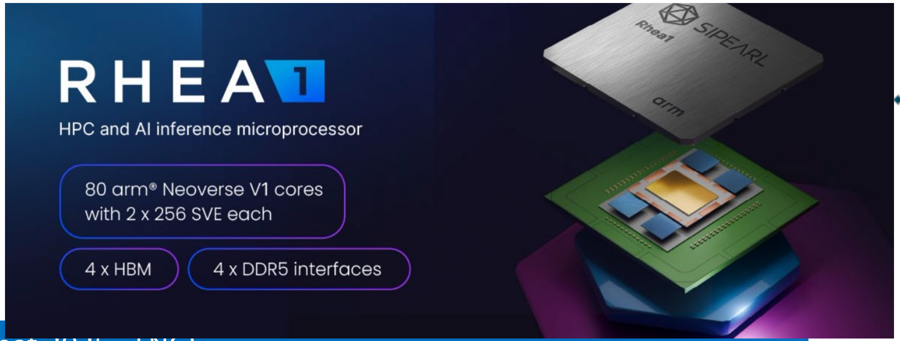
- The ex-European embedded chips maker with innovative business model: license not products
- A pletora of 32b/64b archs characterized by high ratio watt/ops, used on server and userver, embeddded, FPGA
  - low power → high number of cores
  - Limited cost per core
- In the past many CPU integration experiments (cavium, amcc) and EU projects: Mont Blanc, EuroServer, ExaNeSt...
- Current products based on ARM-64
  - APPLE Mx (agreement APPLE-ARM until 2040…)
  - AMPERE: (startup US per server ARM-based) multicore → 64-80
  - Fujitsu A64Fx in HPC system FUGAKU
  - Nvidia GRACE CPU
- SiPearl RHEA GPP (EU funded through EPI consortium): first samples in 2025?

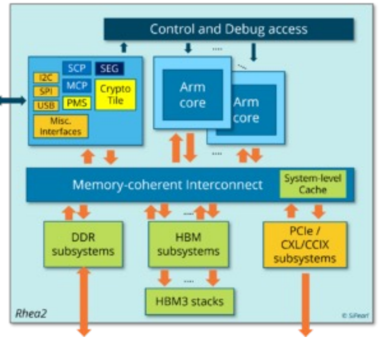






Integer Performance	Ampere Altra	AMD Epyc 7742	AMD Epyc 7702	Intel 8280 SP	Intel Xeon SP 8276	Intel Xeon SP 6238R
Cores	80	64	64	28	28	28
Clock Speed	3.3 GHz	2.25 GHz	2.0 GHz	2.7 GHz	2.2 GHz	2.2 GHz
SPECrate 2017 Integer Base	-	667	593	342	296	287
SPECrate 2017 Integer Base (GCC)	579	557	495	260	225	218
Performance/CPU	290	278	248	130	112	109
Performance/Core	3.62	4.35	3.87	4.64	4.02	3.90
Watts	205	225	200	205	165	165
Performance/Watt	1.41	1.24	1.24	0.63	0.68	0.66
Watts/Core	2.56	3.52	3.13	7.32	5.89	5.89
CPU Price	$5,800	$6,950	$6,450	$10,009	$8,719	$2,612
Price/Performance	$20.03	$24.96	$26.05	$77.02	$77.52	$23.95

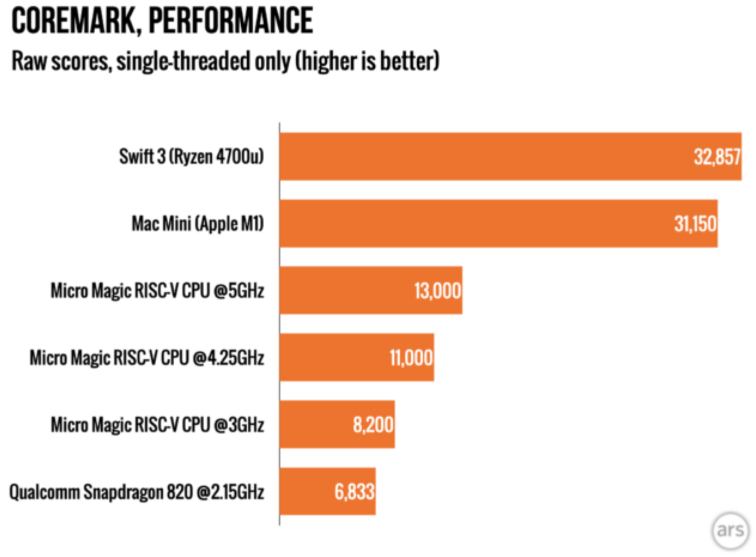


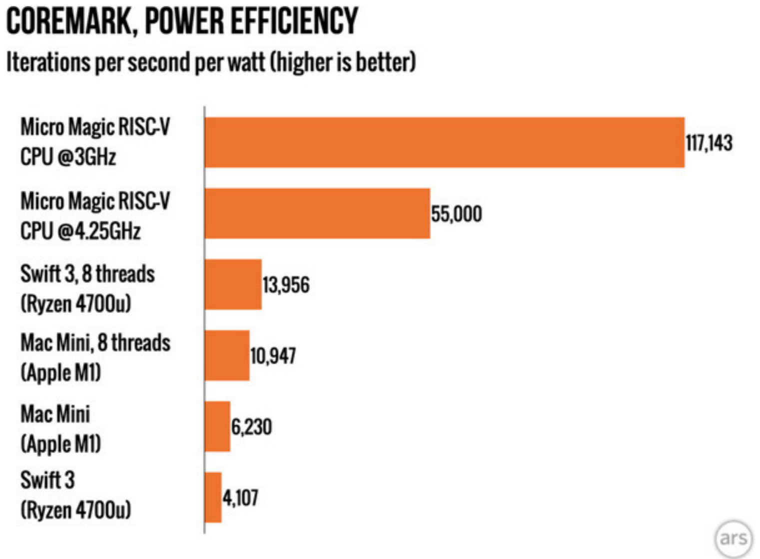


CPU low power (RISC-V)

Beyond the x86 mainstream: RISC-V

- RISC-V is a open source/license free ISA (Instruction Set Architecture)
  - Designed to reduce HW complexity and power consumption and to enhance programmability and computing efficency
- A long history (Berkeley 1981→)
- Today 5th gen supported by RISC foundation (*https://riscv.org/)*
  - 2K+ PARTNERS, tra cui IBM, Intel, Google, Samsung, Nvidia...
- CPU, many-cores acceleratori, ML, uControllori, HPC,..
- Reference platform for next gen EU chip developments
  - EPI, DARE, EPAC (R-V accelerators)…

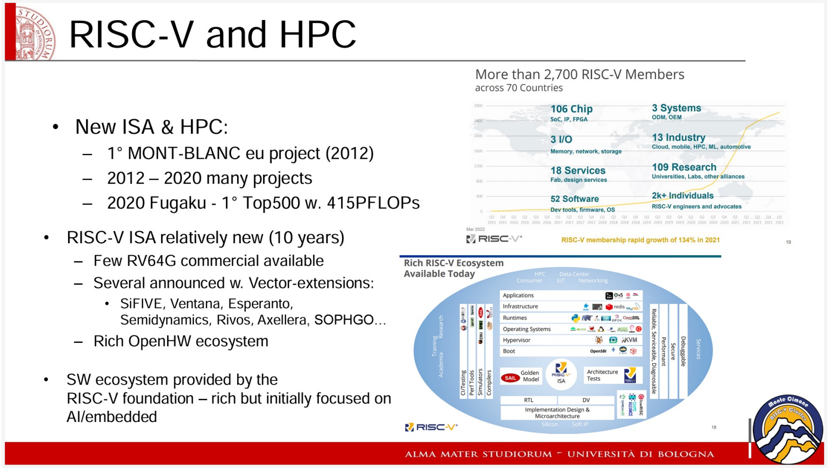




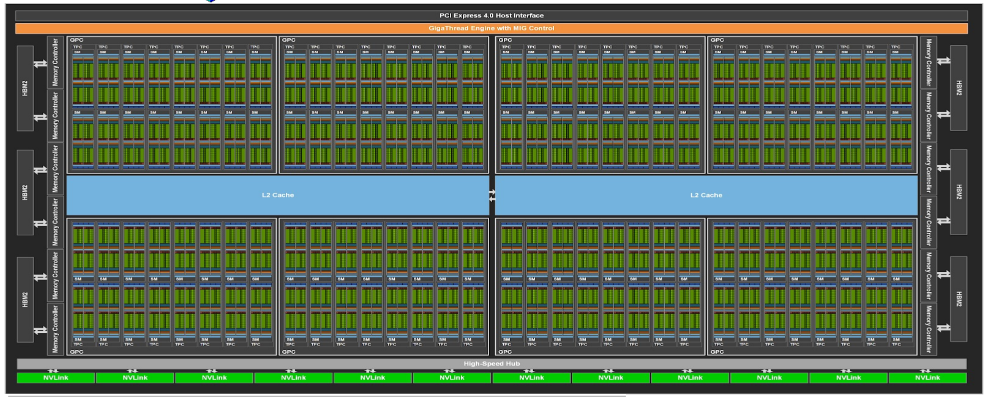
Disruptive **Technology**

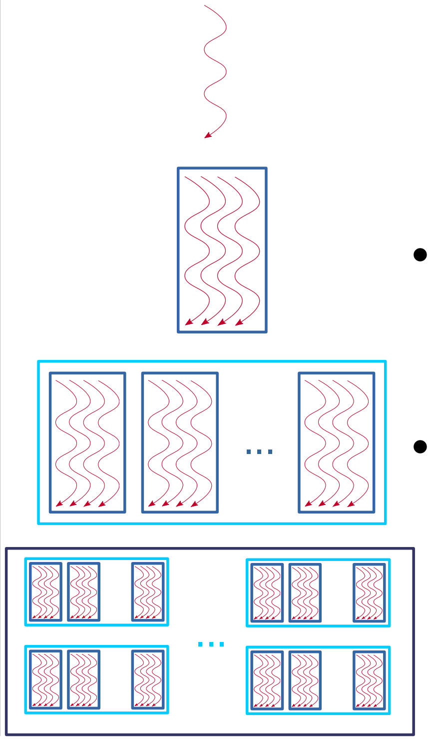
Barriers	Legacy ISA	RISC-V ISA
Complexity	1500+ base instructions Incremental ISA	47 base instructions Modular ISA
Design freedom	$$$ – Limited	Free – Unlimited
License and Royalty fees	$$$	Free
Design ecosystem	Moderate	Growing rapidly. Numerous extensions, open & proprietary cores
Software ecosystem	Extensive	Growing rapidly

RISC-V



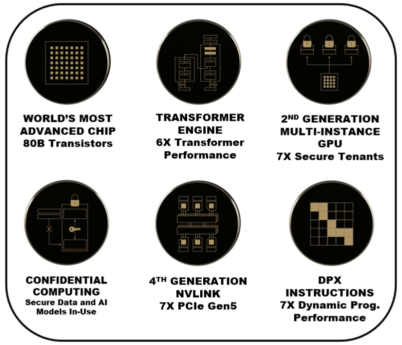
GPU





- Originally specialized processors for graphics
- GPUs are highly multithreaded and make intensive use of parallelism to achieve high performance (many SIMD instructions)
  - execution of many threads (up to 10^3...) in parallel distributed over many elementary cores (10^3)
  - no cache needed to mask memory access latency → a lot of computing, less memory
  - Use of “large” (10^2 bit) and “fast” (N∗Ghz per bit line) graphic memories
- Lots of state-of-the-art technoloy: Standard (DirectX, OpenGL, OpenCL) or proprietary (NVidia Compute Unified Device Architecture (CUDA)) programming languages
- Evolution towards scalable systems optimized (also) for AI
  - Extreme scale DGX systems essentially dedicated to the efficient training of deep networks
  - CPU+GPU integration (Grace Hopper)

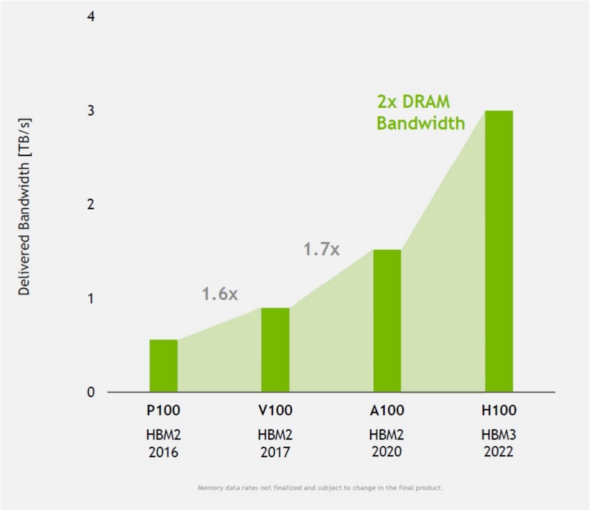
GPU Hopper: state-of-the-art techs



The **full implementation** of the GH100 GPU includes the following units:

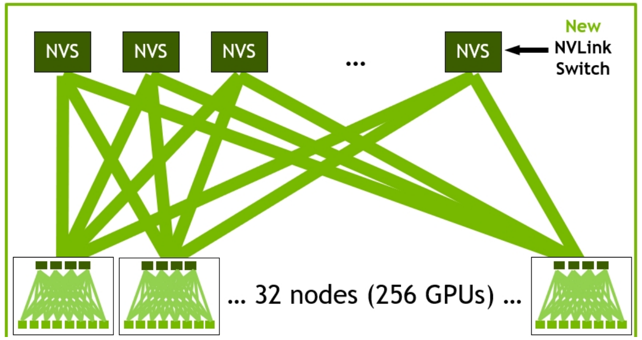
- 8 GPCs, 72 TPCs (9 TPCs/GPC), 2 SMs/TPC, 144 SMs per full GPU
- 128 FP32 CUDA Cores per SM, 18432 FP32 CUDA Cores per full GPU
- 4 Fourth-Generation Tensor Cores per SM, 576 per full GPU
- 6 HBM3 or HBM2e stacks, 12 512-bit Memory Controllers
- 60 MB L2 Cache
- Fourth-Generation NVLink and PCIe Gen 5

High dense SM with 4 TC/SM (576 in total



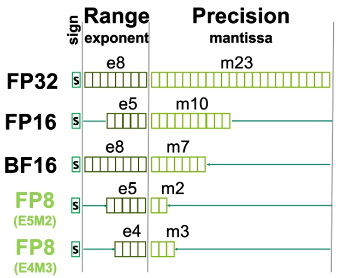
HBM3 memory: 3TB/s

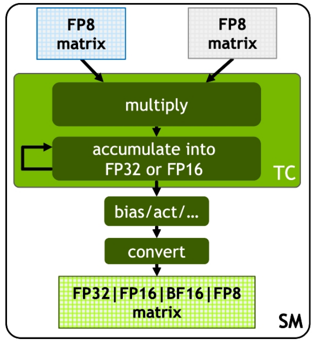
DGX H100 256 SuperPOD



NVLink configurable network, low latency high bandwidth, for
GPU-to-GPU direct access (total of 900GB/s...) + NVLink Switch

Tensor core supporting FPx (32/16/8)

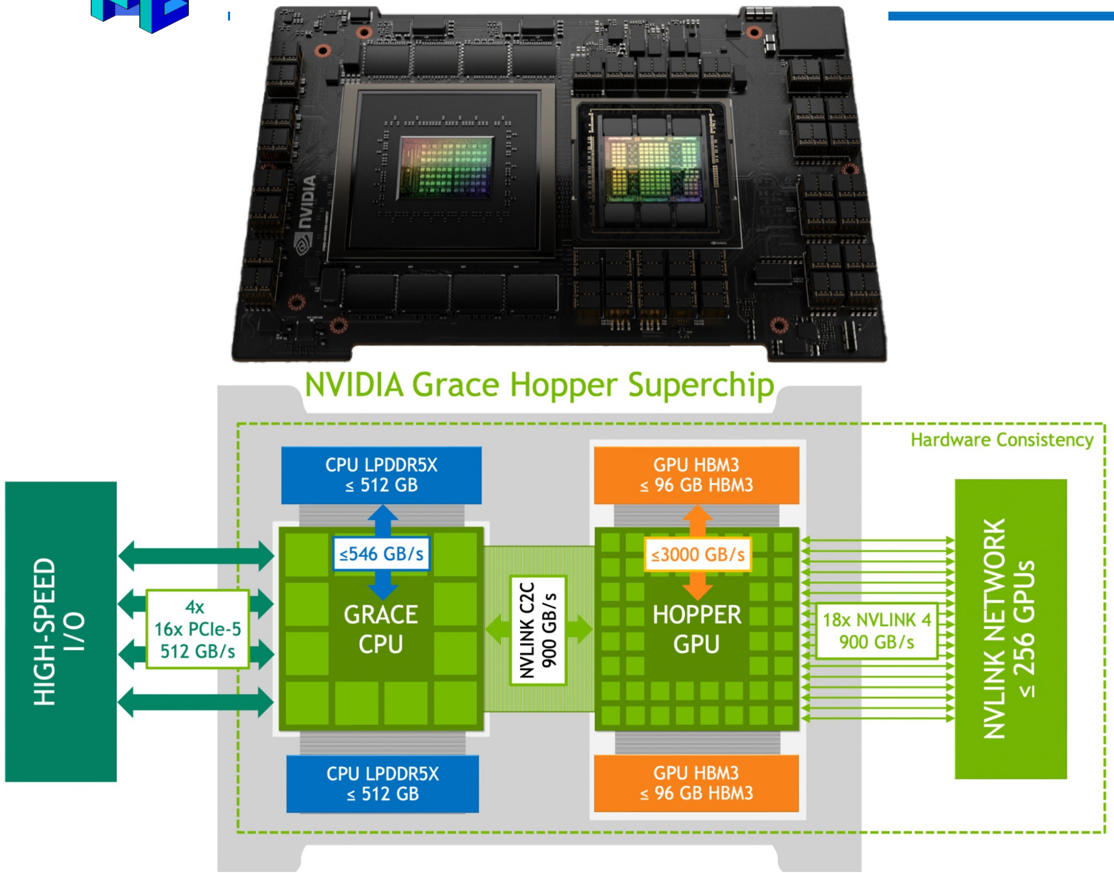




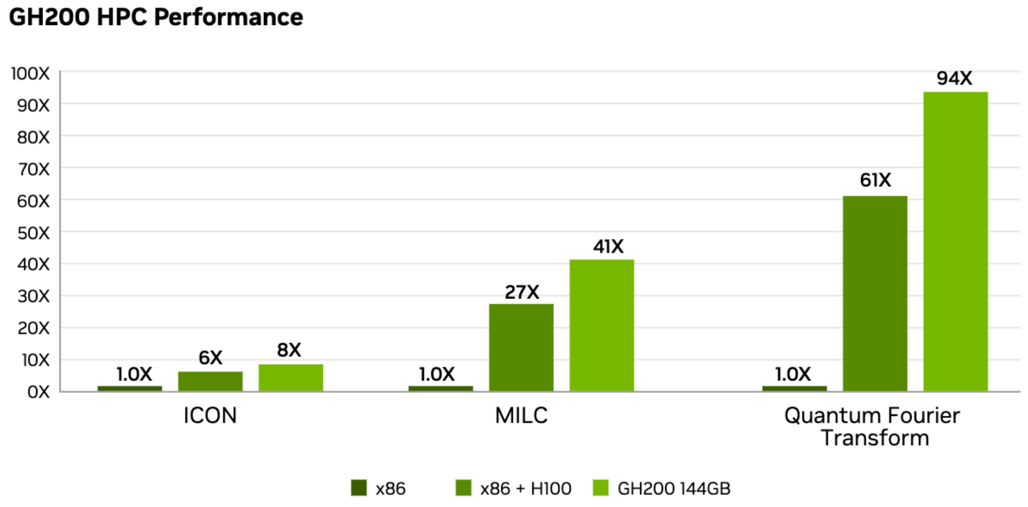
	NVIDIA H100 SXM5	NVIDIA H100 PCIe
Peak FP64	33.5 TFLOPS	25.6 TFLOPS
Peak FP64 Tensor Core	66.9 TFLOPS	51.2 TFLOPS
Peak FP32	66.9 TFLOPS	51.2 TFLOPS
Peak FP16	133.8 TFLOPS	102.4 TFLOPS
Peak BF16	133.8 TFLOPS	102.4 TFLOPS
Peak TF32 Tensor Core	494.7 TFLOPS | 989.4 TFLOPS[1]	378 TFLOPS | 756 TFLOPS[1]
Peak FP16 Tensor Core	989.4 TFLOPS | 1978.9 TFLOPS[1]	756 TFLOPS | 1513 TFLOPS[1]
Peak BF16 Tensor Core	989.4 TFLOPS | 1978.9 TFLOPS[1]	756 TFLOPS | 1513 TFLOPS[2]
Peak FP8 Tensor Core	1978.9 TFLOPS | 3957.8 TFLOPS[1]	1513 TFLOPS | 3026 TFLOPS[1]
Peak INT8 Tensor Core	1978.9 TOPS | 3957.8 TOPS[1]	1513 TOPS | 3026 TOPS[1]

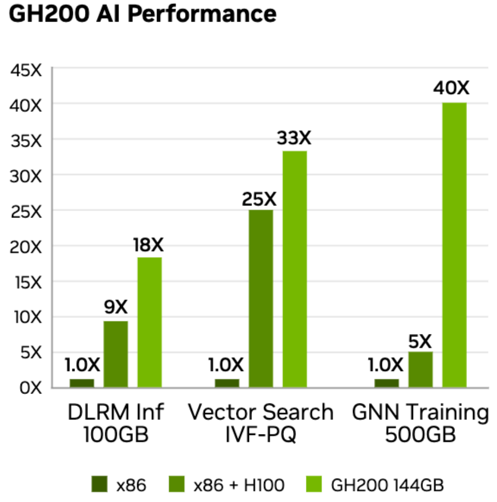
https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper

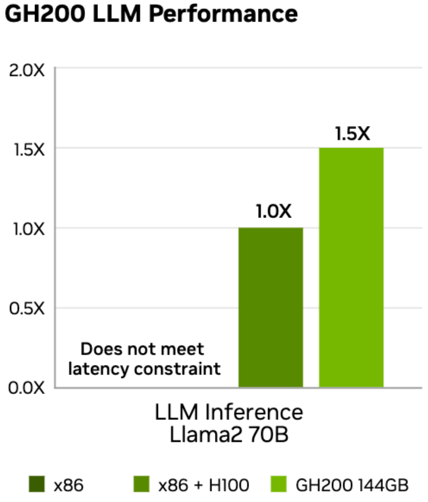
GPU and CPU integration, the "new" wave: GH200 Grace Hopper Superchip



Feature	Description
Grace CPU cores (number)	Up to 72 cores
CPU LPDDR5X bandwidth (GB/s)	Up to 500GB/s
GPU HBM bandwidth (GB/s)	4TB/s HBM3 4.9TB/s HBM3e
NVLink-C2C bandwidth (GB/s)	900GB/s total, 450GB/s per direction
CPU LPDDR5X capacity (GB)	Up to 480GB
GPU HBM capacity (GB)	96GB HBM3 144GB HBM3e
PCIe Gen 5 Lanes	64x





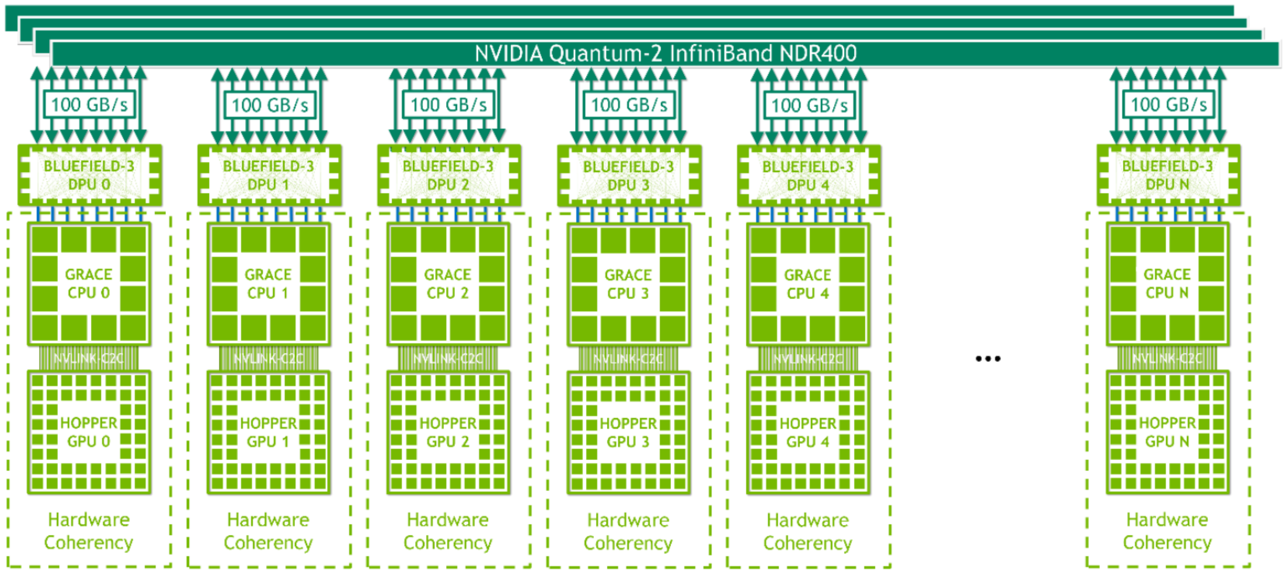


https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

Scaling with Grace Hopper and Grace Blackwell (next gen)

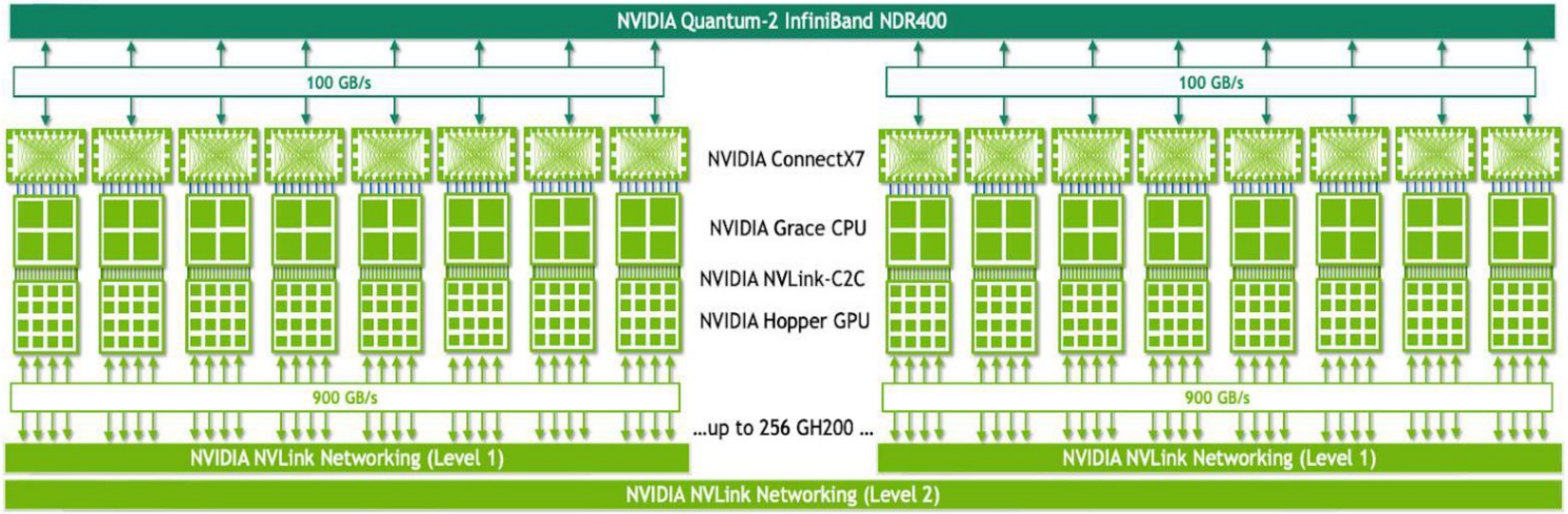
Allows for extreme scalability (mainly for AI)

https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper

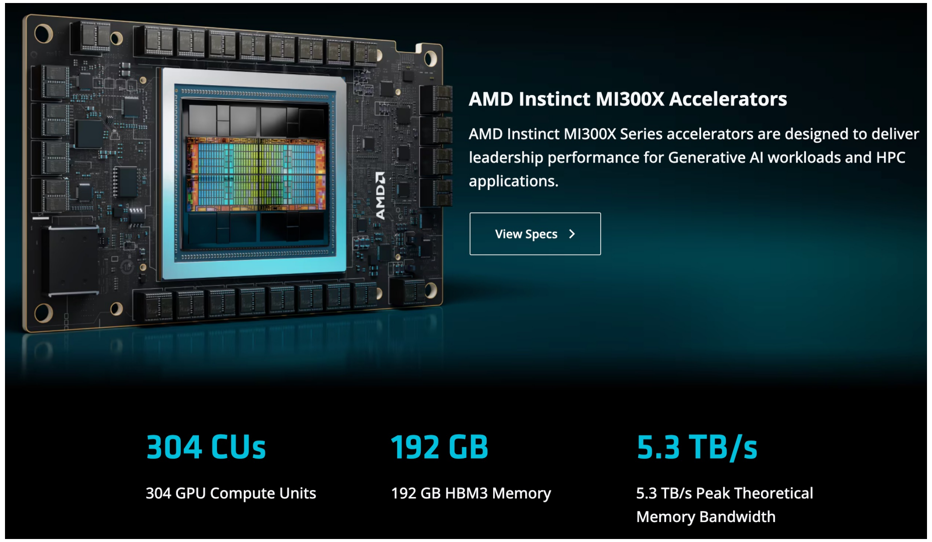


MGX200: Grace Hopper Superchip system with DPU+InfiniBand networking for scale-out ML and HPC workload

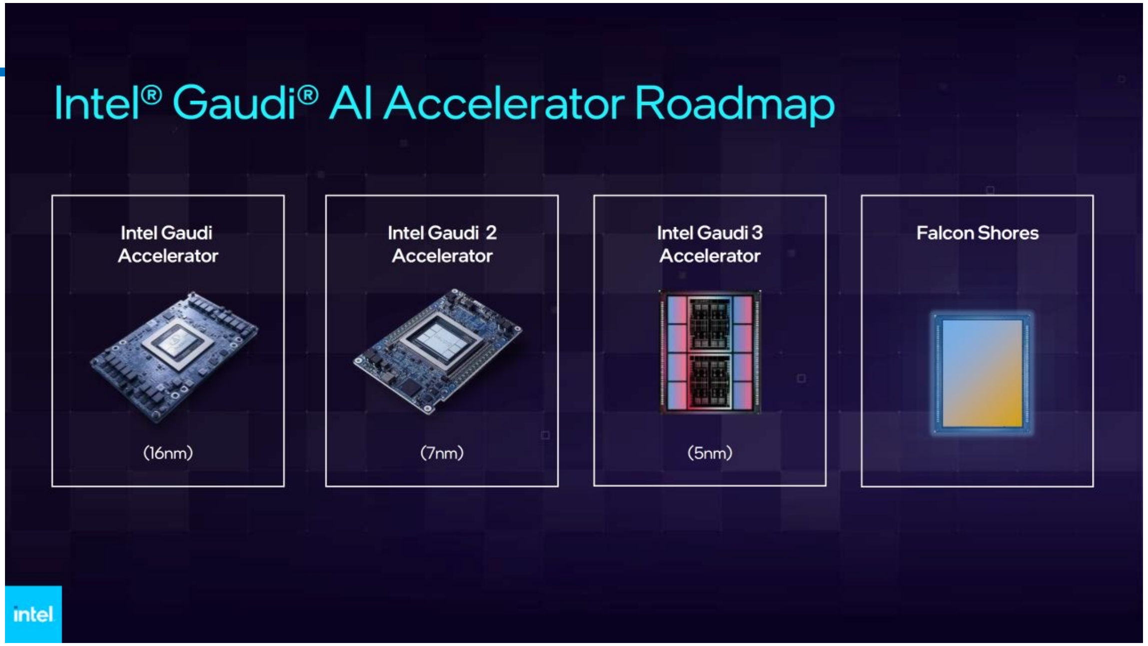
NVIDIA GH200 NVL32 with NVLink Switch System for strong-scaling giant ML workload



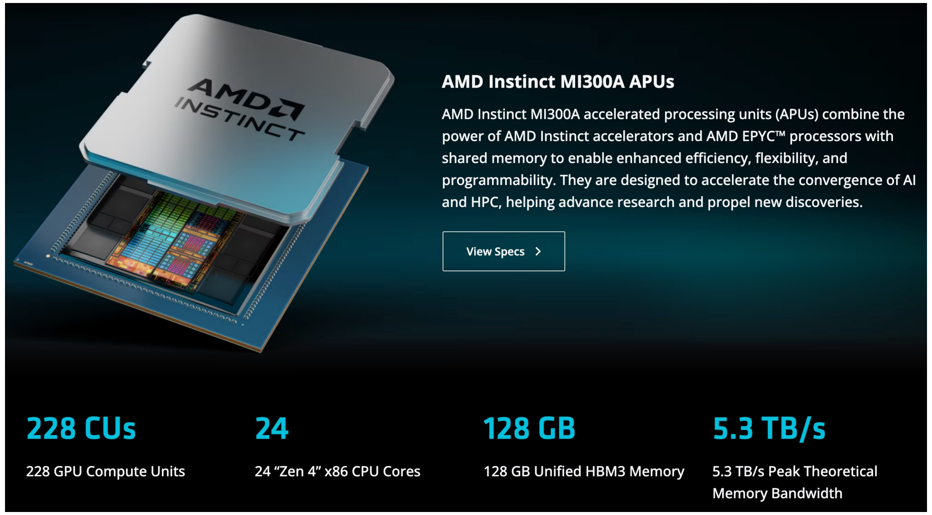
NVIDIA competition



AMD MI300X acc: Hopper competitor



Intel GaudiX??? Dejavu of Ponte Vecchio???



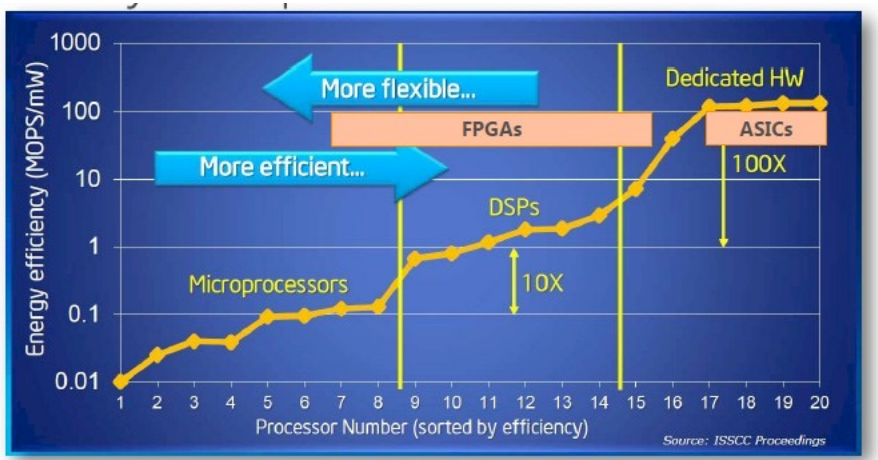
AMD MI300A AIPU: Grace Hopper competitor

AMD+INTEL+Google: Open standard SW stack for accelerators: CUDA competitor

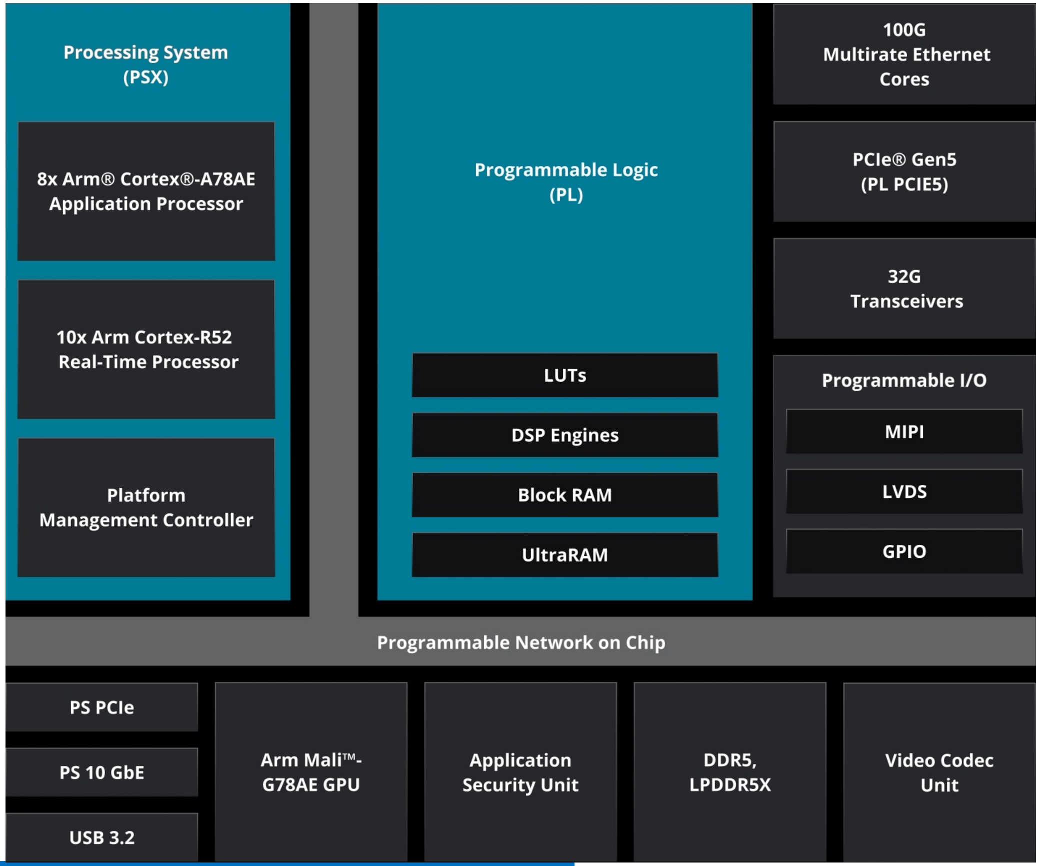
FPGA

FPGA: programmable device characterized by flexibility, reconfigurability, power efficiency, short time-to-market

- from original PAL/GAL to current Btransistors SoC
- Example: XCV80 HBM Versal AMD
  - → 7nm FinFET node several 10xB transistors
  - → Multiple (2->8) ARM Cores (A72) + R5F Real TimeCores @1.5GHz
  - → 128 transceivers 32-56-(112) Gbps chip-to-chip or via backplane interconnect
  - → 3M system logic cells (up to 1GHz)
  - → Up to 32GB in-package HBM DRAM (~820GB/s) e 500Mb memory
  - → Many industrial standards: ETH100g o 200g, PCIExpress gen3/4/5 x16...
  - → 38 TOPs (22 TeraMACs) DSP computing performance
  - → IP specialized for per ML inference



Source: Bob Broderson, Berkeley Wireless group



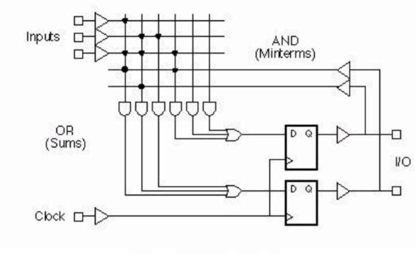
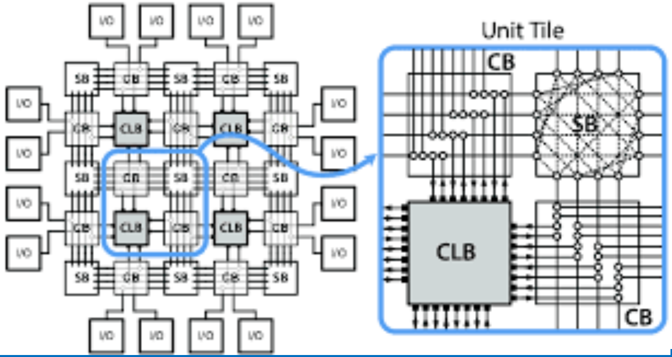


Figure 3 PAL Architecture



FPGA Programming

- A bit of history
  - originally hand synthesis...
  - then CAD tools (Schematic and synthesizers) to map high-level design to a hardware netlist
  - with increasing complexity → abstracting using HDL Hardware Description Language (VHDL, Verilog) + compilers + synthesis
- more abstraction needed
  - Introduction of HLS (High Level Synthesis) from C++ (or similar)
  - “Task parallelism" model to exploit the intrinsic parallelism of the FPGA
  - Development tools Xilinx VITIS HLS, INTEL oneAPI,...
- Thanks to these high-level programming tools → FPGA “democratization”
  - short time-to-design
  - A “simple” application specialist can design FPGA and exploit them
  - New high-level tools for automatic design and mapping of scalable parallel applications
    - OmPSS-FPGA (BSC) for programming MPI-based multiple FPGA systems
    - HLS4ML (CERN) automatic generation of FPGA firmaware for ML computing tasks
    - INFN APEIRON Framework

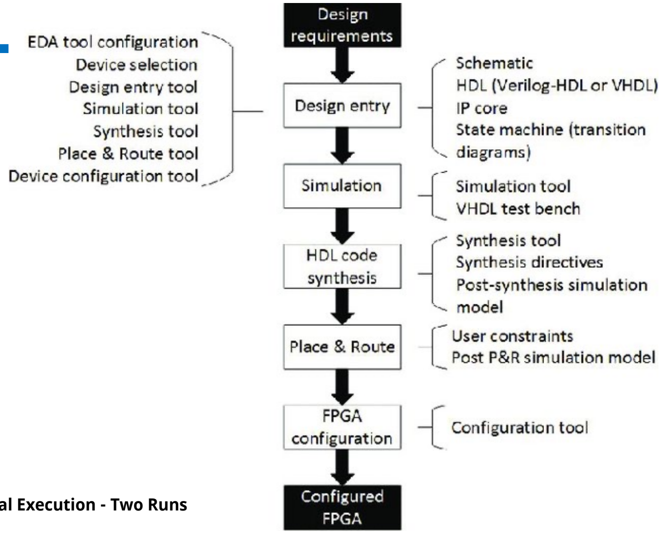


Figure 5: **Sequential Execution - Two Runs**

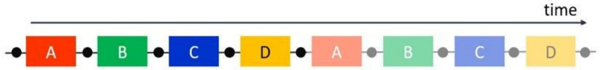


Figure 6: **Task Parallelism within a Run**

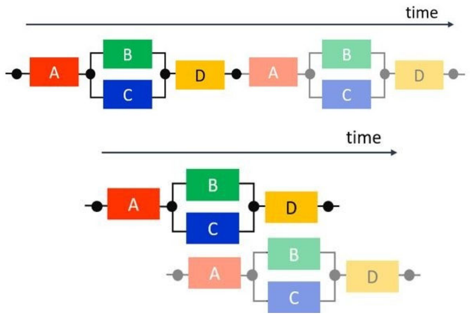
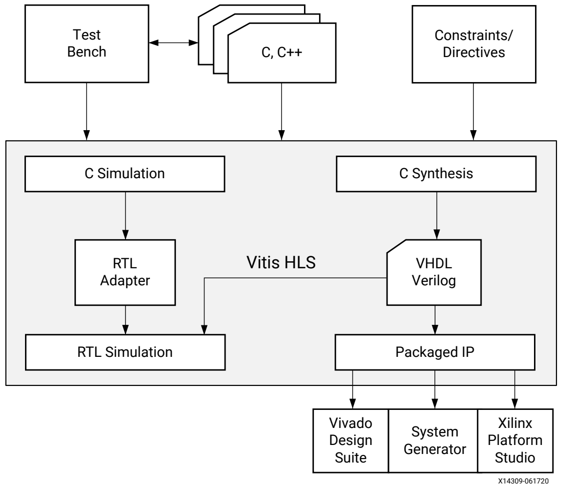
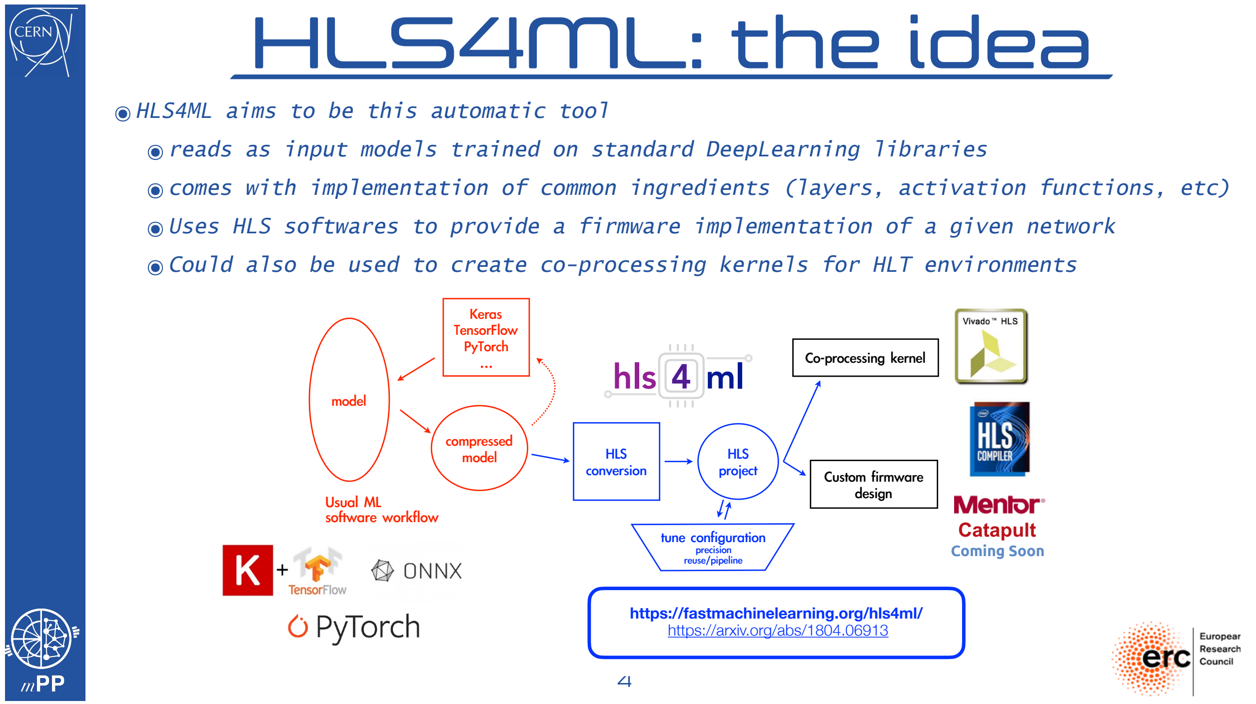
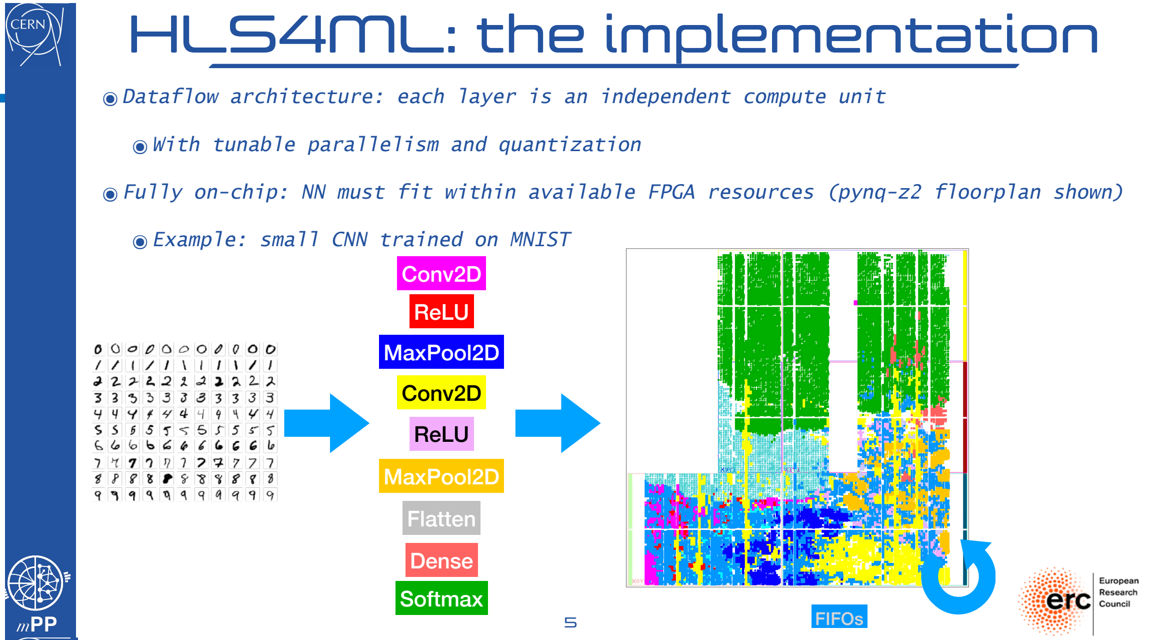


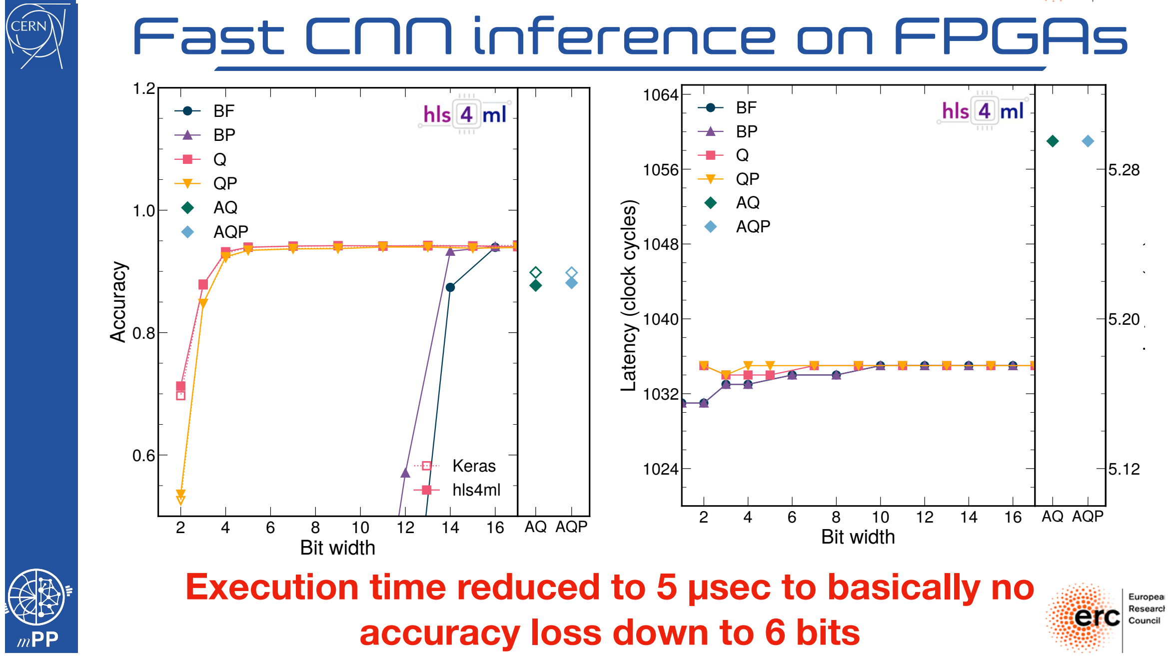
Figure 18: **Vitis HLS Design Flow**



FPGA Programming (per ML): H







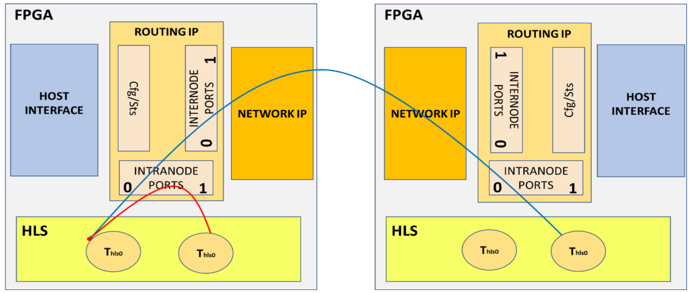
- Supporto (a vari livelli di maturita') per
  - pruning
  - compression
  - quantizzazione
  - parallelizzazione
  - Graph Nets
  - "Knowledge distillation" (teacher-student model)

M. Pierini (WS AI INFN 2022) https://agenda.infn.it/event/29907/contributions/163448/attachments/90265/121584/AI%40INFN.pdf

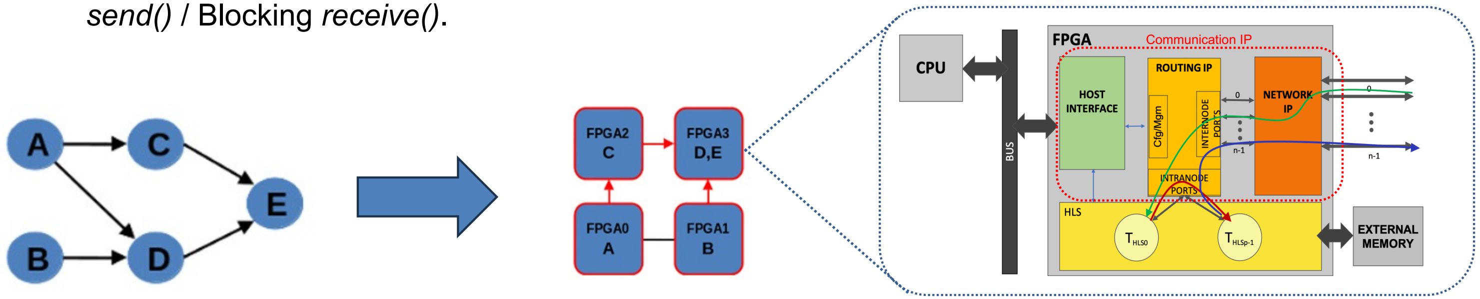
APEIRON programming framework

APEIRON enables the scaling of Xilinx Vitis® High Level Synthesis applications on multiple FPGAs interconnected by the INFN communication IP

- Enables the mapping the dataflow graph of the application on the distributed FPGA system and offering runtime support for the execution.
- Allows users, with no (or little) experience in hardware design tools, to develop their applications on such distributed FPGA-based platforms:
  - ➢ Tasks are implemented in C++ using High Level Synthesis tools (Xilinx Vitis®).
  - ➢ Lightweight C++ communication API: Non-blocking *send()* / Blocking *receive()*.



INFN communication IP
- Direct network for FPGA accelerators
- Dimension Order (DOR) routing policy
- Virtual Cut Through switching technique
- Implemented in VHDL as a Xilinx Vitis RTL kernel
- Low-latency communication between HLS processing tasks:
  - ➢ Intra-node latency < 400ns for message sizes up to 1kB
  - ➢ Inter-node latency < 1μs for message sizes up to 1kB



Host Interface IP: Interface the FPGA logic with the host through the system bus.
- Xilinx® XDMA PCIe Gen3

Routing IP: Routing of intra-node and inter-node messages between processing tasks on FPGA.

Network IP: Network channels and Application-dependent I/O
- Custom APElink 20/40 Gbps
- UDP/IP over 1/10/25/40 GbE

HLS Kernels: user defined processing tasks

https://apegate.roma1.infn.it/?page_id=1328

→ Many more details in Francesca Lo Cicero talk

EPJ Web of Conferences **295**, 11002 (2024)
https://doi.org/10.1051/epjconf/202429511002

Are FPGA suitable for HPC?

The Billion Euro question…

- Technology is aligned in terms of silicon process and density. Just a ½ of hardware peak clock frequency
- A pletora of tools to exploit programmability
- Very low ratio power/performance compared to CPU or GPU
- Many years of projects, study, protos, work for accelerators and networks showing (at low TRL) promising behaviour…

- But
  - Peak performance not (yet) comparable with CPU/GPU due to the peculiarities of architecture
  - Costs/performance is again too high for (almost) any application
  - Proven really effective in a small number of specific area (ML at reduced precision, streaming computing, SmartNIC) and for very few scientific applications
  - programming is easier than before but still not within everyone's reach
- SO, the answer is "NI"
  - Last gen FPGA are impressive in terms of capabilities and resources and are good for prototype/evaluate/debug architectures
  - proven succesful for specific applications (ML at reduced precision, streaming computing, robotics)
  - valuable and cost effective for study new architecture and new features

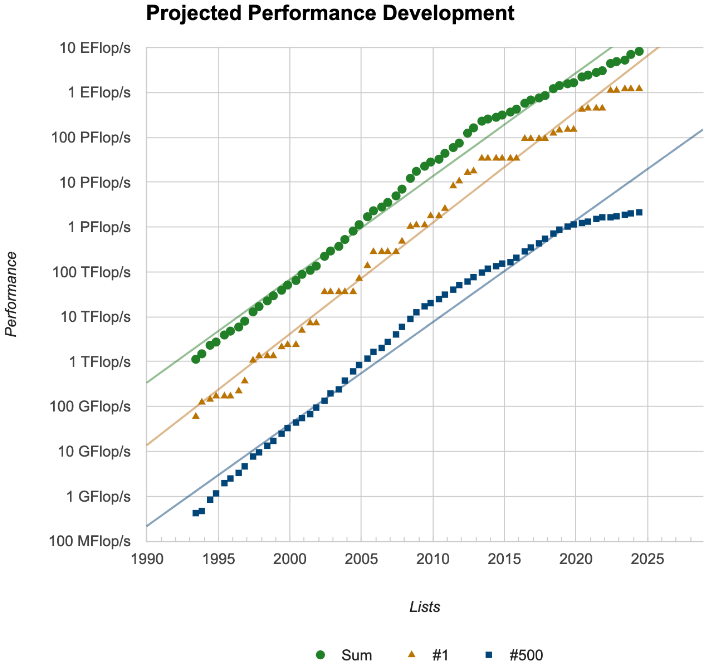
→ Additional research is again needed, on innovative architecture optimized for FPGA, tools, suitable applications, exploitation of heterogeneity…

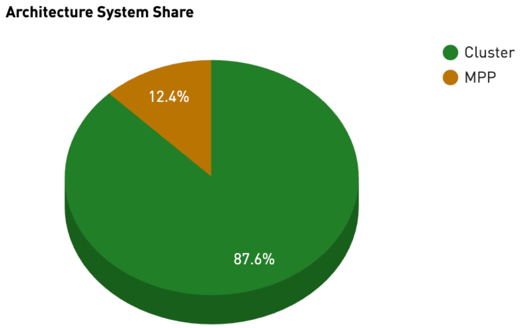
TOP500 (6/2024)

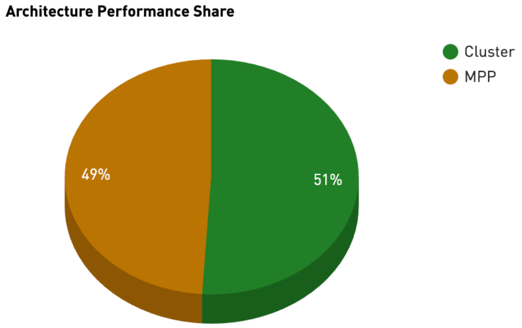
Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,206.00	1,714.81	22,786
2	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
3	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	
4	**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
5	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107

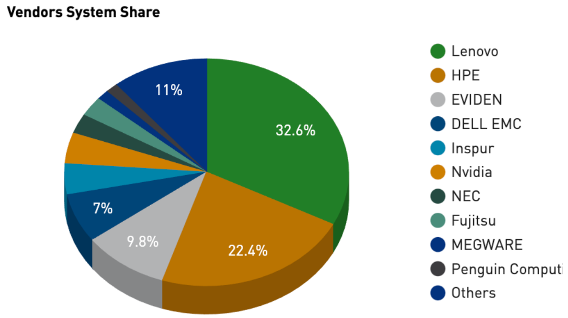
Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
6	**Alps** - HPE Cray EX254n, NVIDIA Grace 72C 3.1GHz, NVIDIA GH200 Superchip, Slingshot-11, HPE Swiss National Supercomputing Centre (CSCS) Switzerland	1,305,600	270.00	353.75	5,194
7	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, EVIDEN EuroHPC/CINECA Italy	1,824,768	241.20	306.31	7,494
8	**MareNostrum 5 ACC** - BullSequana XH3000, Xeon Platinum 8460Y+ 32C 2.3GHz, NVIDIA H100 64GB, Infiniband NDR, EVIDEN EuroHPC/BSC Spain	663,040	175.30	249.44	4,159
9	**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
10	**Eos NVIDIA DGX SuperPOD** - NVIDIA DGX H100, Xeon Platinum 8480C 56C 3.8GHz, NVIDIA H100, Infiniband NDR400, Nvidia NVIDIA Corporation United States	485,888	121.40	188.65	

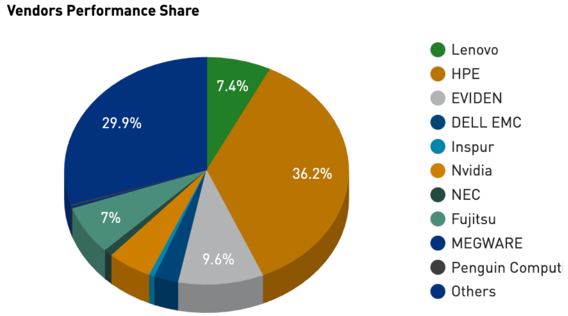
TOP500 (6/2024)

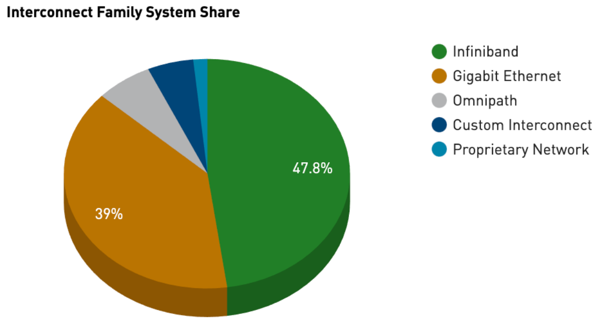


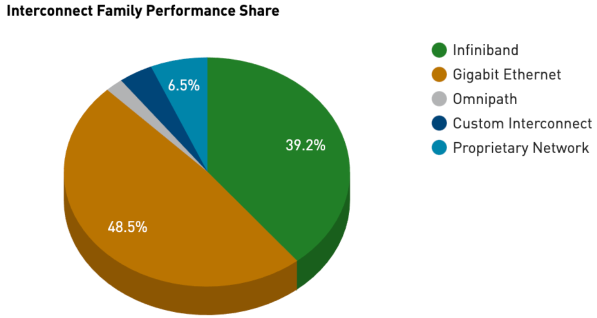


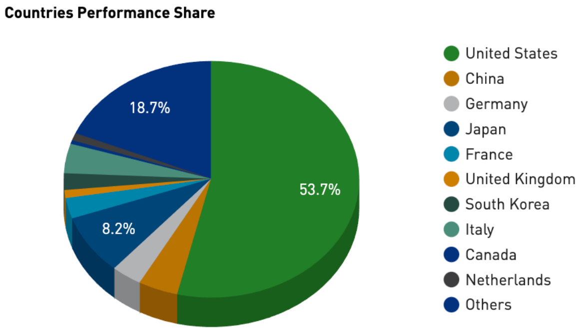












European competitiveness report (i.e. rapporto Draghi)

Public spending on R&I in Europe lacks scale and is insufficiently focused on breakthrough innovation.

In the US, the vast majority of public R&I spending is carried out at the federal level.

In the EU, governments overall spend a similar amount to the US on R&I as a share of GDP, but only one tenth of spending takes place at the EU level.

......

The EU's key instrument to support radically new technologies at low readiness levels – the European Innovation Council's (EIC) Pathfinder instrument – has a budget of EUR 256 million for 2024, compared with USD 4.1 billion for US Defence Advanced Research Projects Agency (DARPA) and USD 2 billion for the other "ARPA" agencies.

.....

Lack of intra-EU coordination affects the wider innovation ecosystem as well.

Most Member States cannot achieve the necessary scale to deliver worldleading research and technological infrastructures, in turn constraining R&I capacity. By contrast, the examples of CERN and the European High-Performance Computing Joint Undertaking (EuroHPC) showcase the importance of coordination when developing large R&I infrastructure projects.

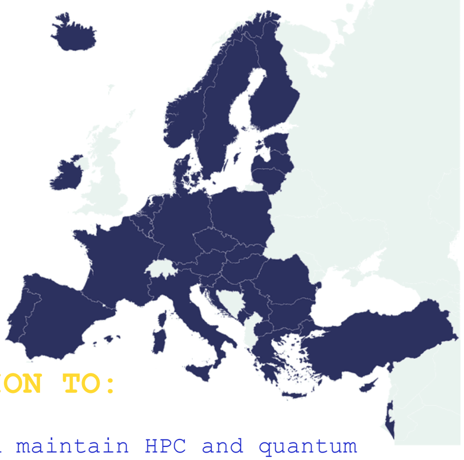
EU funding for HPC and beyond: EuroHPC

- European effort to sustain and push forward the HPC in Europe
- At the beginning a couple of main pillars:
  - infrastructure (funds for HPC systems) and technological research
- Today convergent initiatives: HPC, AI and Quantum Computing

A QUICK EUROHPC RECAP…

WE ARE:

- An EU body and funding entity
- Existing since 2018 and autonomous since 2020
- Based in Luxembourg
- Governed by a Board composed of the European Commission, 34 Participating States and 3 Private Members



WITH A BUDGET COMING FROM 3 EU FUNDING PROGRAMMES:

- Digital Europe Programme: EUR 1.98B
- Horizon Europe Programme: EUR 900M
- Connecting Europe Facility: EUR 200M

- EU contributions are matched by national contributions

WITH A MISSION TO:

i. Buy, build and maintain HPC and quantum infrastructure in Europe
ii. Fund innovative R&I projects, to develop European skills, applications, software amd hardware and foster a European supply chain
iii. Provide access to HPC and Quantum Users across Europe and support the development of skills

EuroHPC Mission

Develop, deploy, extend and maintain in the Union a world leading federated, secure and hyper-connected supercomputing, quantum computing, service and data infrastructure ecosystem; support the production of innovative and competitive supercomputing systems based on a supply chain that will ensure components, technologies and knowledge limiting the risk of disruptions and the development of a wide range of applications optimised for these systems; and, widen the use of this supercomputing infrastructure to a large number of public and private users, and support the development of key skills for European science and industry.

EuroHPC Objectives

A Federated supercomputing infrastructure

- *Exascale and post exascale supercomputers*
- *Mid-range supercomputers*
- *Industrial-grade supercomputers*
- *Quantum Computers*

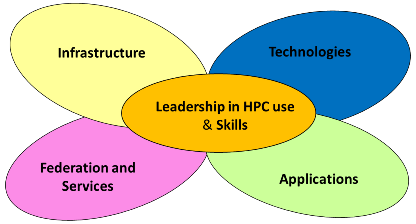
Technologies & Applications for the HPC ecosystem

- *R&D on new computing technologies and architectures and their integration in supercomputing systems*
- *Advanced industrial, scientific and public sector applications*

Leadership in use and Skills

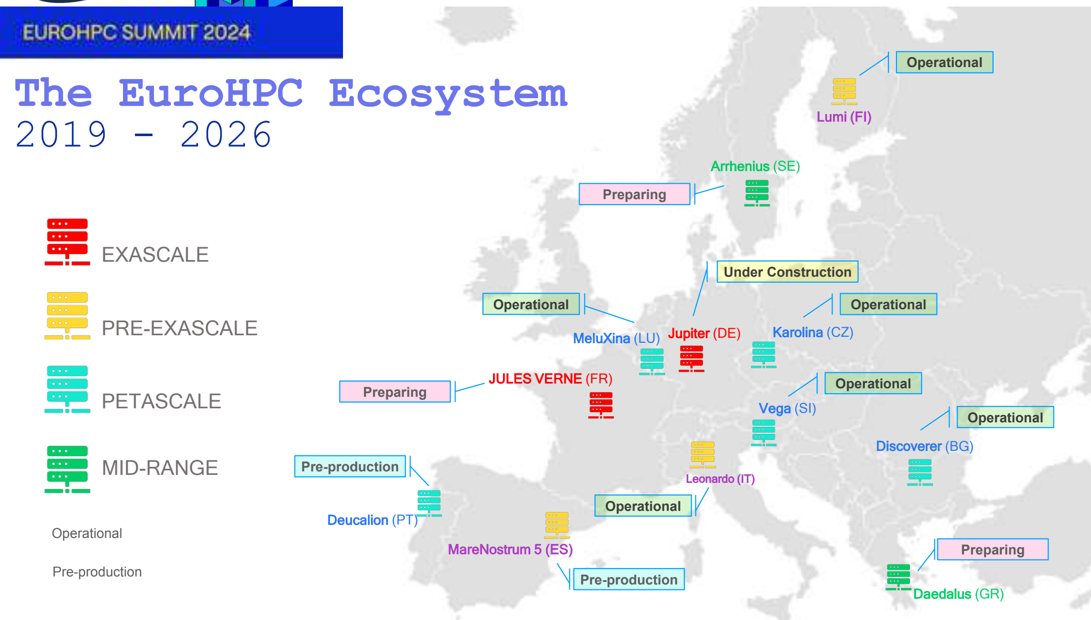
- *Wide use primarily for civilian applications – incl. for EU strategic initiatives (e.g. Destination Earth, personalised health, crisis management, etc.)*
- *HPC Skills, Education, Training*

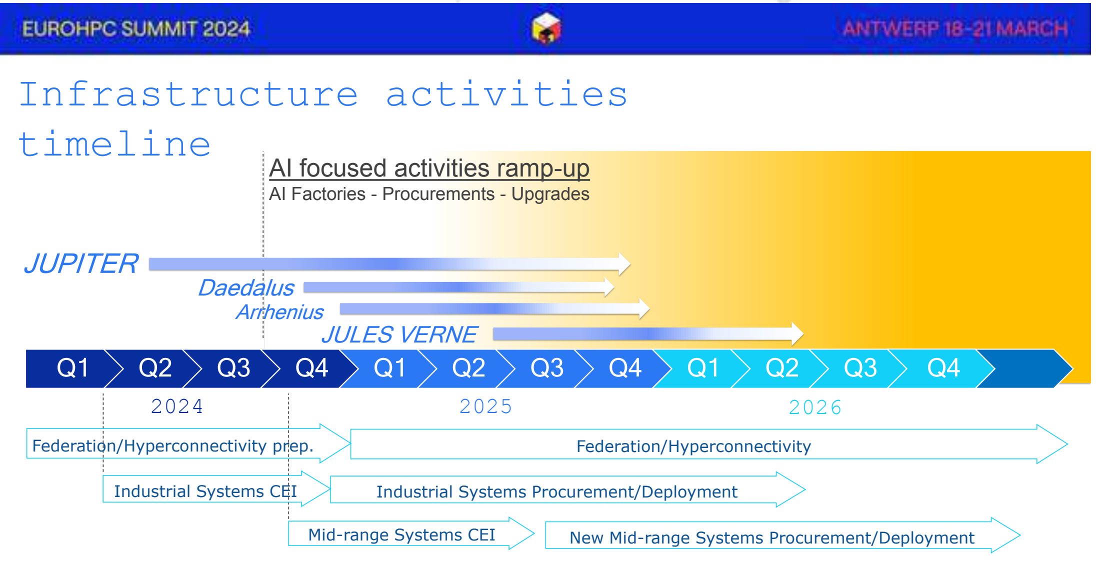
The 5 Pillars of activity



European Commission

EuroHPC: funding infrastructure





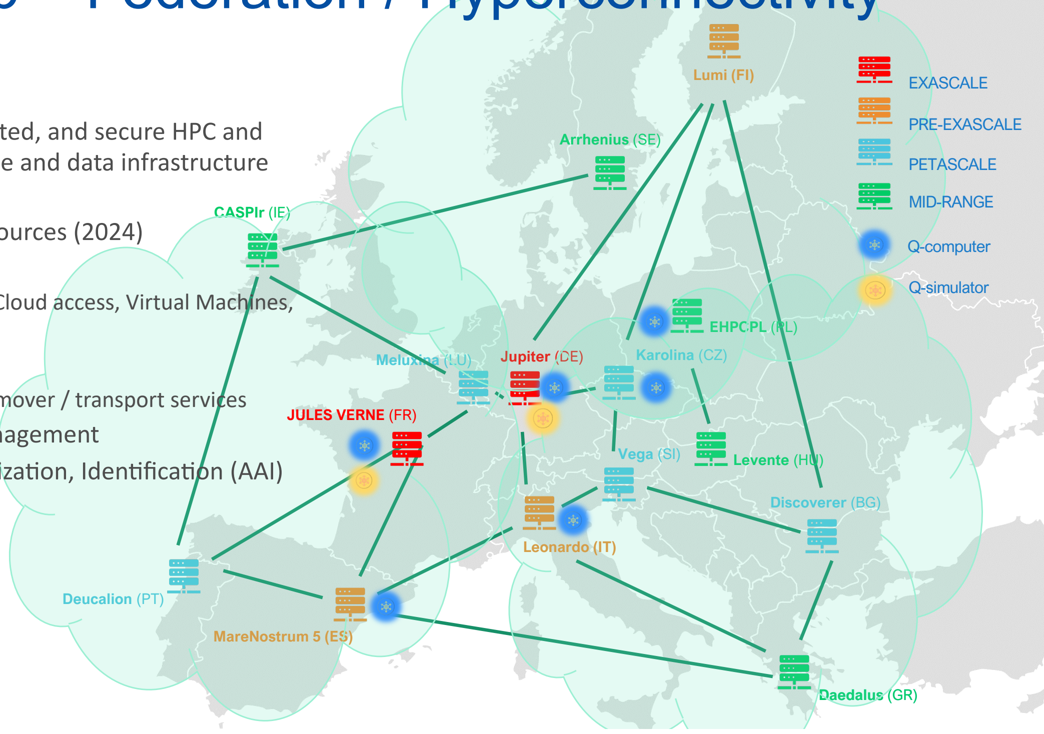
5 PETASCALE
- Vega (Slovenia)
- Karolina (CZ)
- ⍰ Discoverer BG) → under upgrade
- Meluxina (Lux)
- Deucalion (PT)

3 PRE-EXASCALE
- LUMI (Finland)
- Leonardo (Italy)→ LISA upgrade for AI (28ME)
- MareNostrum 5 (Spain)

1+1 EXASCALE
- Jupiter (DE) assembling
- JulesVerne (FR) planned

Coming up – Federation / Hyperconnectivity

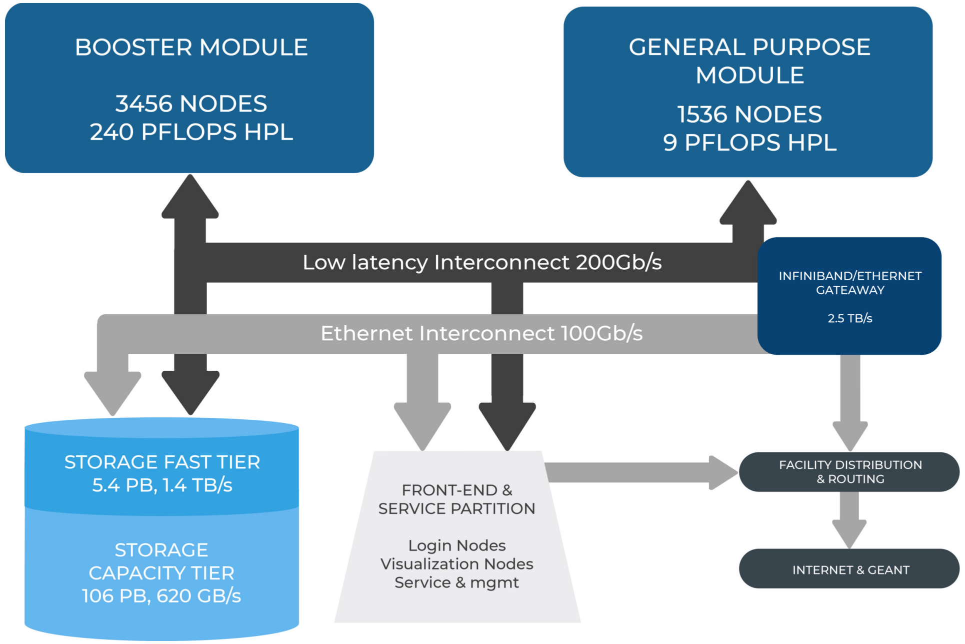
- A hyper-connected, federated, and secure HPC and quantum computing service and data infrastructure ecosystem
- Federation of EuroHPC resources (2024)
  - ✓ Computing services
    - Interactive Computing, Cloud access, Virtual Machines, Containers
  - ✓ Data services
    - Data repositories, Data mover / transport services
  - ✓ User and Resource management
  - ✓ Authentication, Authorization, Identification (AAI)
- Hyperconnectivity (2025)



LEONARDO

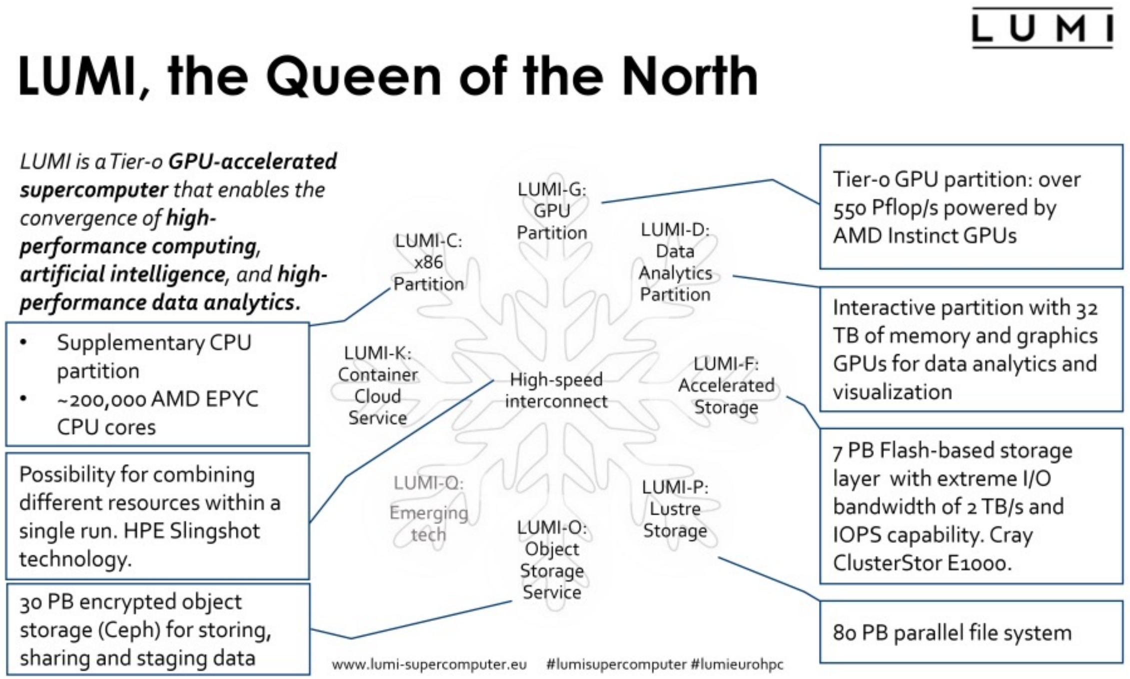
Leonardo (CINECA)

- **3456 computing nodes,** each equipped with **four NVidia A100 SXM6 64GB GPUs** (240PFlops)
- A **Data Centric module** aiming to **satisfy a broader range of applications**. Its 1536 nodes are equipped with two Intel Sapphire Rapids CPUs (9 PFlops of sustained performance).
- All the nodes are interconnected through an **Nvidia Mellanox** network @200gbps
- **BullSequana** XH2000 mechanics
- Funded by JU (120ME) and MIUR (120ME)



LUMI: Large Unified Modern Infrastructure

- Pan-European pre-exascale supercomputers in CSC's data center in Kajaani, Finland.
- Consortium: Finland, Belgium, the Czech Republic, Denmark, Estonia, Iceland, the Netherlands, Norway, Poland, Sweden, and Switzerland.
- Tech:
  - The LUMI is based on an **HPE Cray EX supercomputer**.
  - The LUMI-G GPU partition: 2978 nodes for a total of 11912 **AMD GPUs**.
  - Fast Cray **Slingshot** interconnect of 200 Gbit/s per node
  - The Linpack performance of LUMI-G is **380 Pflop/s**.
  - LUMI-C CPU-only partition with 64-core 3rd-generation AMD EPYC™ CPUs (**over 262 000 CPU cores**) and 256-1024 GB per node.
  - **400m2** of space, which is about the size of two tennis courts. The weight of the system is nearly **150 000 kilograms** (150 metric tons).
- Total budget **144ME**

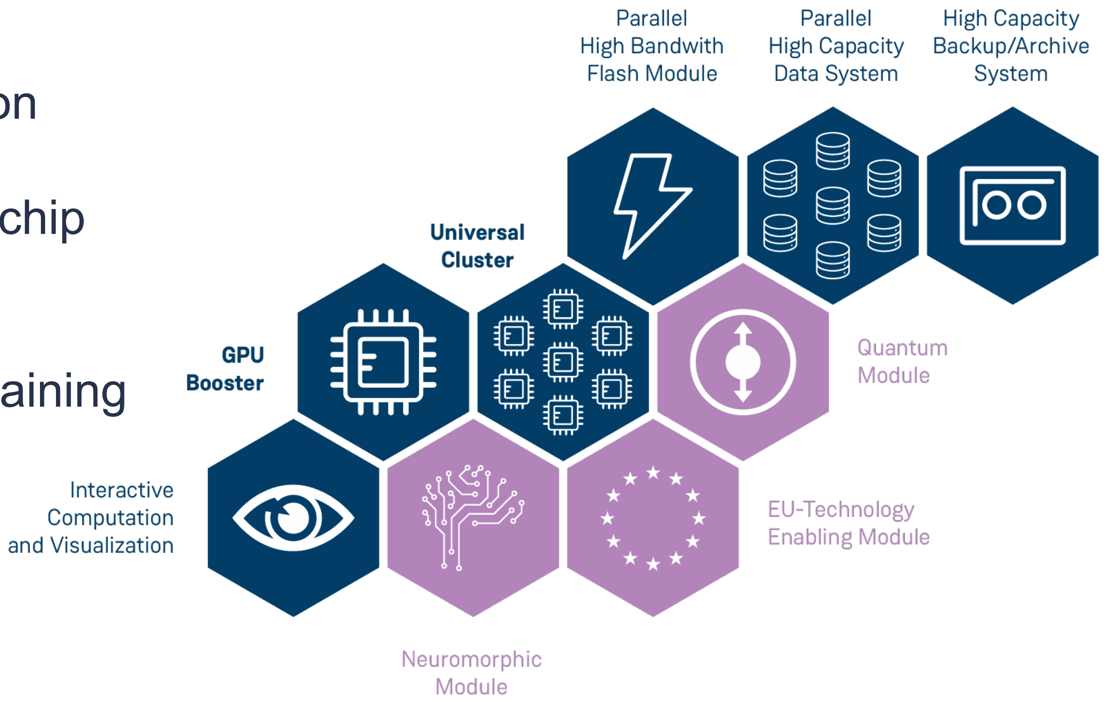


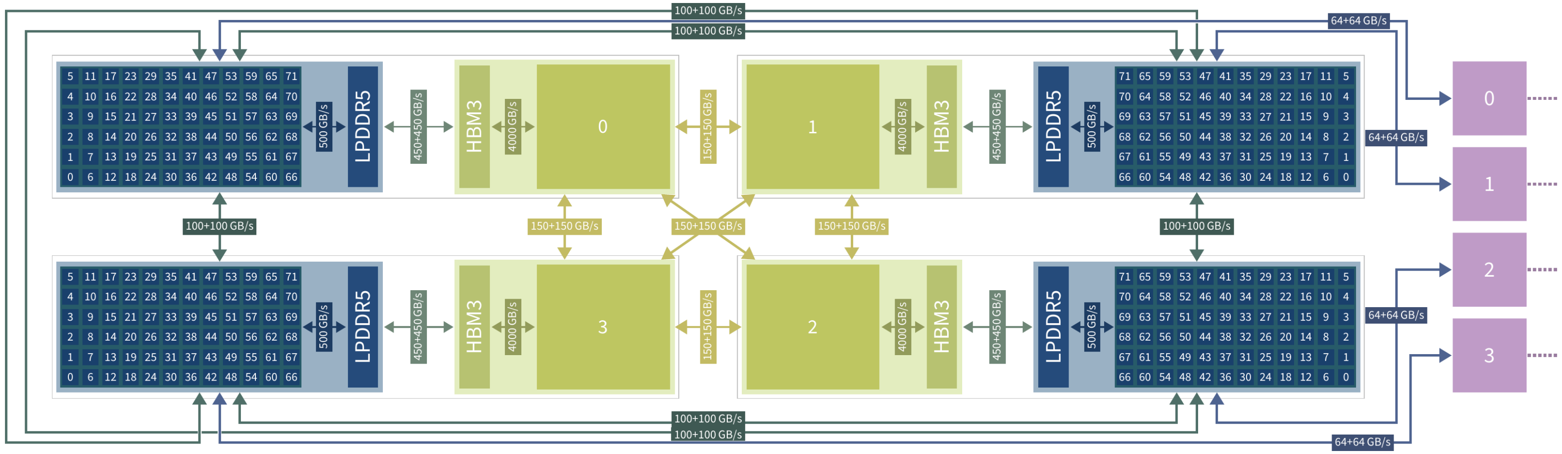
https://docs.lumi-supercomputer.eu/

JUPITER: Joint Undertaking Pioneer for Innovative and Transformative Exascale Research

First EU funded Exascale HPC system under construction @Julich

- Made of 24.000 **NVIDIA GH200** Grace Hopper Superchip
- 25 BullSequana **XH3000** racks interconnected by NVIDIA **Quantum-2 InfiniBand** networking.
- Over **70 Exaflops** for 8-bit calculations (common in training AI models,
- A partition made of **European tech** (RHEA from EPI)





Alice Recoque (FR)

- The second Exascale HPC system funded by EU
- JULES VERNE consortium (FR+NL)
- Up to now, selected the hosting center @CEA near Paris…
- Declared fully operational in 2026. Most likely 2027
- MSA architecture with (one) native European technology:SiPearl's ARM-based Rhea-2 chip, succeded to Rhea-1 chip in Jupiter.
- More (hopefully) to come...

Jules Verne : The French led Exascale project

A French/NL consortium
- GENCI (FR) *Hosting Entity*
- CEA (FR) *Hosting Site*
- SURF (NL) as member of consortium

Full TCO over 5 years: 542 M€
(50% EuroHPC, 50% consortium)

Goal: Deploy a world-class Exascale supercomputer, based on European hardware and software technologies, addressing European major societal and scientific challenges via the convergence at scale of numerical simulations, massive data analysis and artificial intelligence.

Access to EuroHPC platforms

EUROHPC SUMMIT 2024 — ANTWERP 18-21 MARC

Calls for preparatory activities

BENCHMARK ACCESS CALL

- For scaling tests & benchmarks
- Fixed amount of allocation for 2 or 3 months
- Continuously open with monthly cut-offs
- Results and access to system: 2 weeks from cut-off date

DEVELOPMENT ACCESS CALL

- For code and algorithm development
- Fixed amount of allocation for 6 or 12 months
- Continuously open with monthly cut-offs
- Results and access to system: 2 weeks from cut-off date

Calls for production activities

REGULAR ACCESS CALL

- For projects that require large-scale HPC resources
- Allocation duration: for 12 months
- Continuously open with 2 cut-offs per year
- Peer-review process duration: 4 months

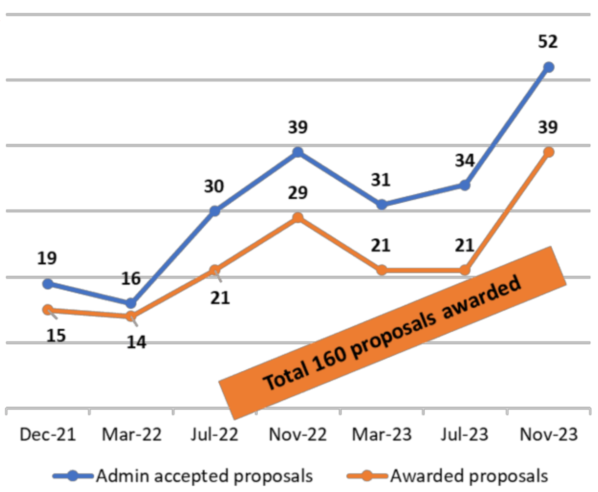
EXTREME SCALE ACCESS CALL

- For high-impact, high-gain projects requiring extremely large-scale HPC resources
- Allocation duration: for 12 months
- Continuously open with 2 cut-offs per year
- Peer review process duration: 6 months

AI AND DATA INTENSIVE APPLICATIONS ACCESS CALL

- For projects intending to perform artificial intelligence and data-intensive activities
- Fixed allocation for 12 months on first-arrived-first served basis
- Bimonthly cut-offs
- Peer-review process duration: 1 month

Administratively accepted vs awarded proposals - all cut-offs



Total 22,179,899 node hours awarded

No. of awarded proposals vs Administratively accepted proposals

Cut-offs	Proposal nos.	
	Admin accepted	Awarded
Dec-22	36	26
May-23	17	15

* October 2023 cut-off still under evaluations

Total 41,914,156 node hours awarded

- Proposal submission via the Peer-Review Platform available at https://pracecalls.eu
- **High success rate...**

EuroHPC: merging infrastructures HPC+QC

EUROHPC QUANTUM INITIATIVES

QUANTUM COMPUTERS

- Four procurements already launc
:
  - EuroQCS-Poland, located in Poland
  - Euro-Q-Exa, located in Germany
  - EuroQCS-France, located in France
  - LUMI-Q, located in Czechia

Each QC will be integrated into an existing supercomputer in Europe

⟨HPC|QS⟩

2 quantum simulators under development, to be integrated in:

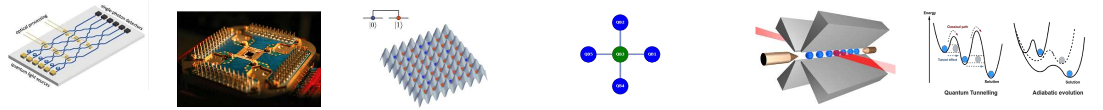
- Joliot Curie (France)
- JUWELS (Germany)

COMING NEXT

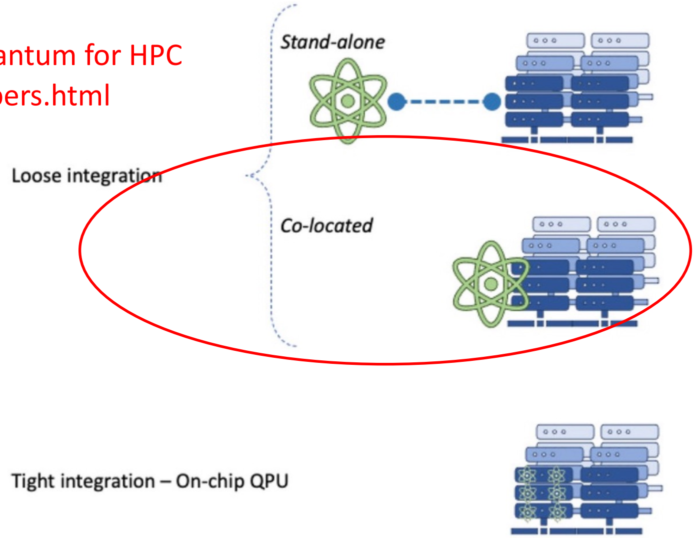
- Finalising the ongoing procurements of the quantum computers
- Calls for further quantum computers
- Development of HPC-Quantum technologies
- Development of Hybrid algorithms and applications
- Establishment of Quantum Excellence Centres
- quantum with 3rd countries
- Enabling Universal Access and Integration of Quantum Resources, to facilitate access and foster innovation

DIVERSITY IN QUANTUM TECHNOLOGIES

EuroQCS-France	Euro-Q-Exa (Germany)	EuroQCS-Italy	Lumi-Q (Czechia)	EuroQCS-Poland	EuroQCS-Spain
Photonic quantum computer	Superconducting qubits	Neutral atoms	Superconducting qubits with a star-shaped topology	Trapped ions	Quantum annealer



ETP4HPC WhitePaper: QC|HPC Quantum for HPC
https://www.etp4hpc.eu/white-papers.html



EuroHPC Actions:

- 2021→HPCQS preparatory initiative
  - a couple of Quantum simulators (up 100 qbits)
  - feasibiity analysis for HPC-Quantum integration
- 2022→ Calls for 6 QC-systems deployments as a booster of Pre&Exascale European systems in a “Co-located” way
  - Technology chosen to get diversity as higher as possible
  - 10-20ME per site
- 2024→ deployments start....

EuroHPC: merging infrastructures HPC+QC in Italy

EuroQCS @CINECA booster of Leonardo

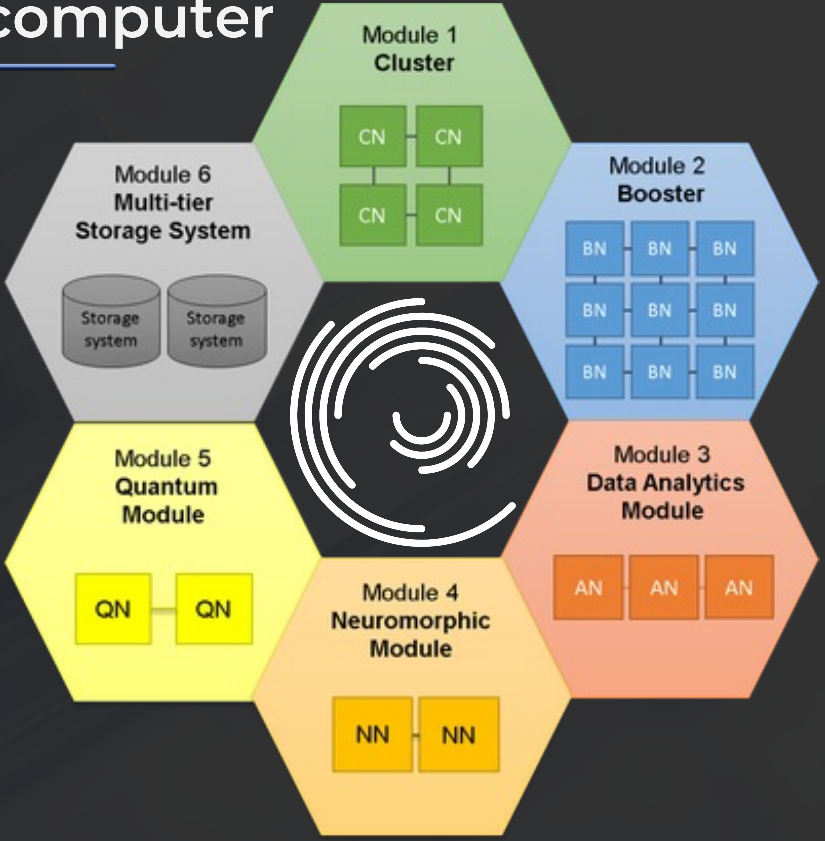
PRESS RELEASE | 1 August 2024 | European High-Performance Computing Joint Undertaking | 3 min read

EuroHPC JU Launches the Procurement for a New Quantum Computer in Italy

The European High Performance Computing Joint Undertaking (EuroHPC JU) has launched a call for tender for the installation of EuroQCS-Italy, a new EuroHPC quantum computer to be integrated into the EuroHPC pre-exascale system Leonardo.

Leonardo: A Modular Supercomputer

- First half 2023: Leonardo
  - Fourth most powerful supercomputer in the World
  - 255+ petaflops (peak performance)
  - Modular Supercomputing Architecture (MSA)
- End 2024: Quantum Module
  - 200 qubits Neutral Atoms Quantum Simulator (analog QC)
- End 2025: QM Improvement
  - Enabling digital and mixed analog/digital mode
- End 2026: QM Improvement 2
  - 500 qubits digital/analog QC
- Future Improvements…



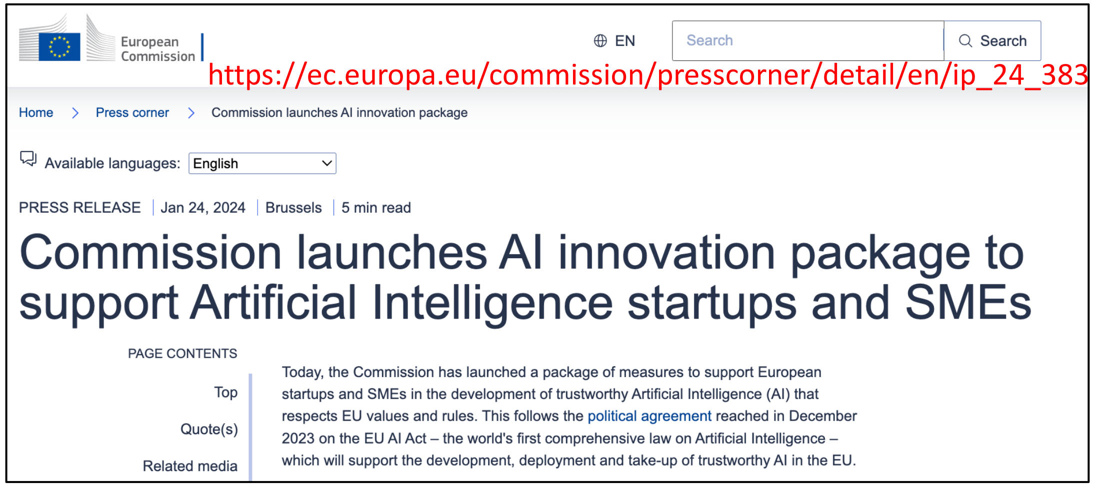
D. Ottaviani talk @HIPEAC 2024

CINECA QUANTUM COMPUTING LAB

EU strategies for AI…

- *First step*: the 2023 EU AI Act
- *Second step*: **the AI innovation package** made of
  - **amendment to EuroHPC to set up AI Factories**
  - **decision to establish an AI Office** supporting the forthcoming AI Act.
  - **EU AI Start-Up and Innovation Communication**:
    - initiatives to strengthen EU's generative AI talent
    - encourage public and private investments in AI start-ups and scale-ups (EIC accelerator & InvestEU)
    - **development and deployment of Common European Data Spaces**, made available to the AI community
    - **GenAI4EU initiative**, support the development of novel use cases and emerging applications in public sector and Europe's 14 industrial ecosystems
      - robotics, health, biotech, manufacturing, mobility, climate and virtual worlds.

- *last week* @AI Action Summit in Paris new announcement:
  - InvestAI initiative to mobilise €200 billion of investment in EU
    - AI gigafactories (100 000 last-generation AI chips)
    - AI Research council....


European Commission | EN | Search | Search

https://ec.europa.eu/commission/presscorner/detail/en/ip_24_383

Home > Press corner > Commission launches AI innovation package

Available languages: English

PRESS RELEASE | Jan 24, 2024 | Brussels | 5 min read

Commission launches AI innovation package to support Artificial Intelligence startups and SMEs

PAGE CONTENTS

Top

Quote(s)

Related media

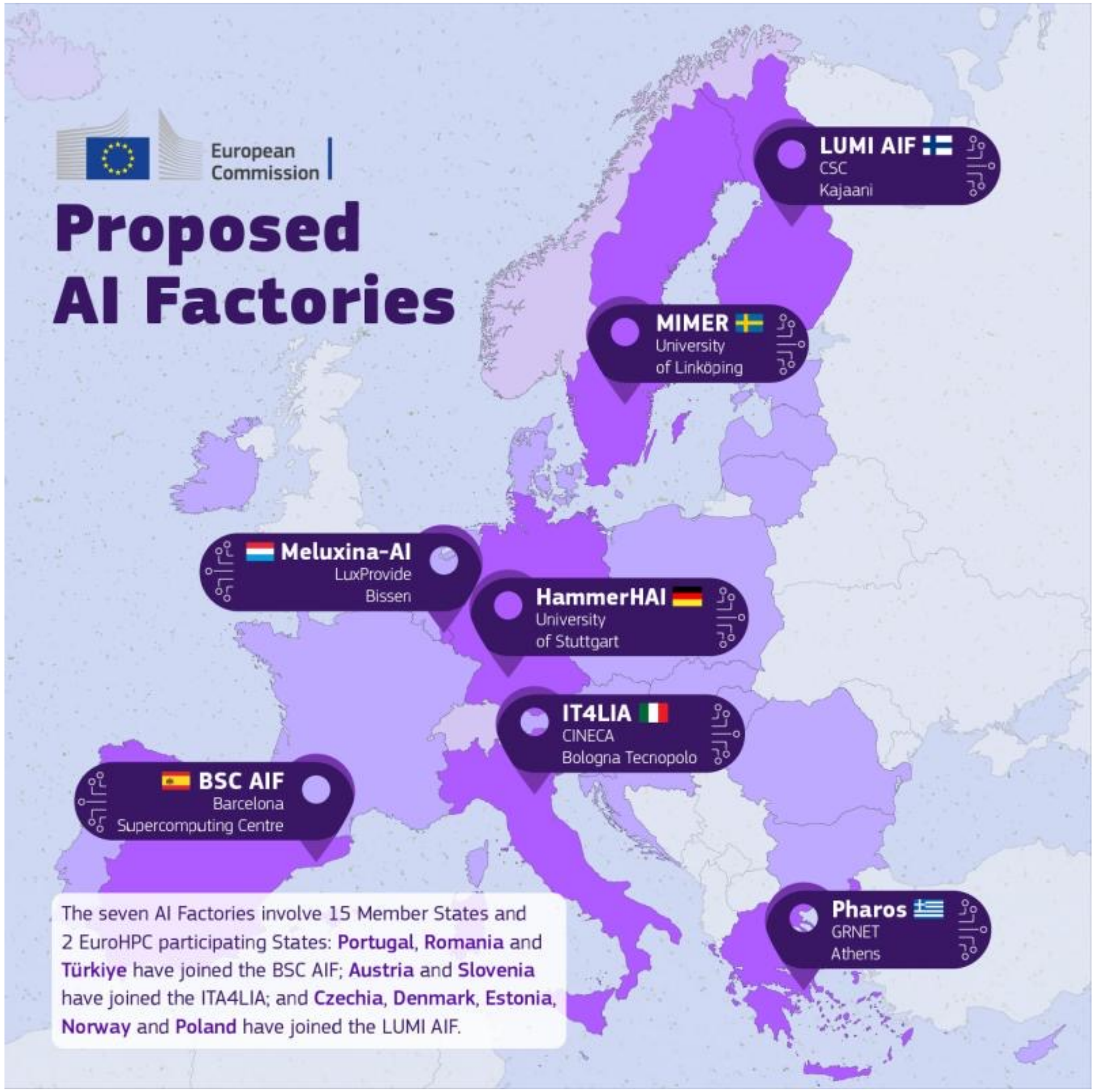
Today, the Commission has launched a package of measures to support European startups and SMEs in the development of trustworthy Artificial Intelligence (AI) that respects EU values and rules. This follows the political agreement reached in December 2023 on the EU AI Act – the world's first comprehensive law on Artificial Intelligence – which will support the development, deployment and take-up of trustworthy AI in the EU.

https://ec.europa.eu/commission/presscorner/detail/en/ip_25_467

Commission President Ursula **von der Leyen** said: *"AI will improve our healthcare, spur our research and innovation and boost our competitiveness. We want AI to be a force for good and for growth. We are doing this through our own European approach – based on openness, cooperation and excellent talent. But our approach still needs to be supercharged. This is why, together with our Member States and with our partners, we will mobilise unprecedented capital through InvestAI for European AI gigafactories. This unique public-private partnership, akin to a CERN for AI, will enable all our scientists and companies – not just the biggest - to develop the most advanced very large models needed to make Europe an AI continent."*

EuroHPC AI factories funds

- Target tier0 (maybe 1?) EU computing centers (i.e CINECA, JULICH,…)
- Two different calls
  - EOI for procurement of advanced experimental AI Factories (AI-01)
  - EOI for acquisition of AI supercomp. OR upgrade current EuroHPC HPC supercomputers with AI optimised booster (AI-02)
- Total budget: minimum of 800ME BUT co-funded by national governments at 50%
  - 400ME for AI-02 in 2024
  - 180ME for AI-01
  - 15 ME for 3 years operating cost
- → Approved systems
- Next steps "gigafactory"…

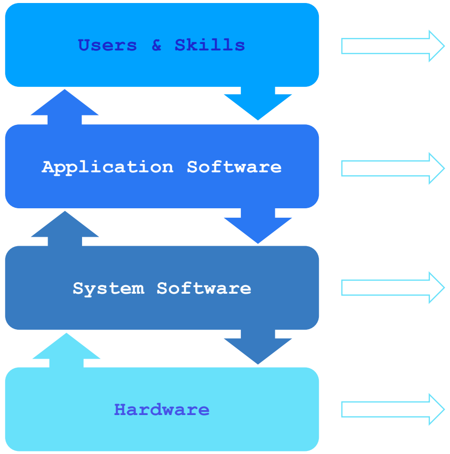


EuroHPC Research & Innovation

R&D on new computing technologies and architectures included their integration in supercomputing systems
Advanced industrial, scientific and public sector applications

RESEARCH & INNOVATION

Currently over 40 ongoing or concluded projects in a range of domains and contributing to European digital autonomy

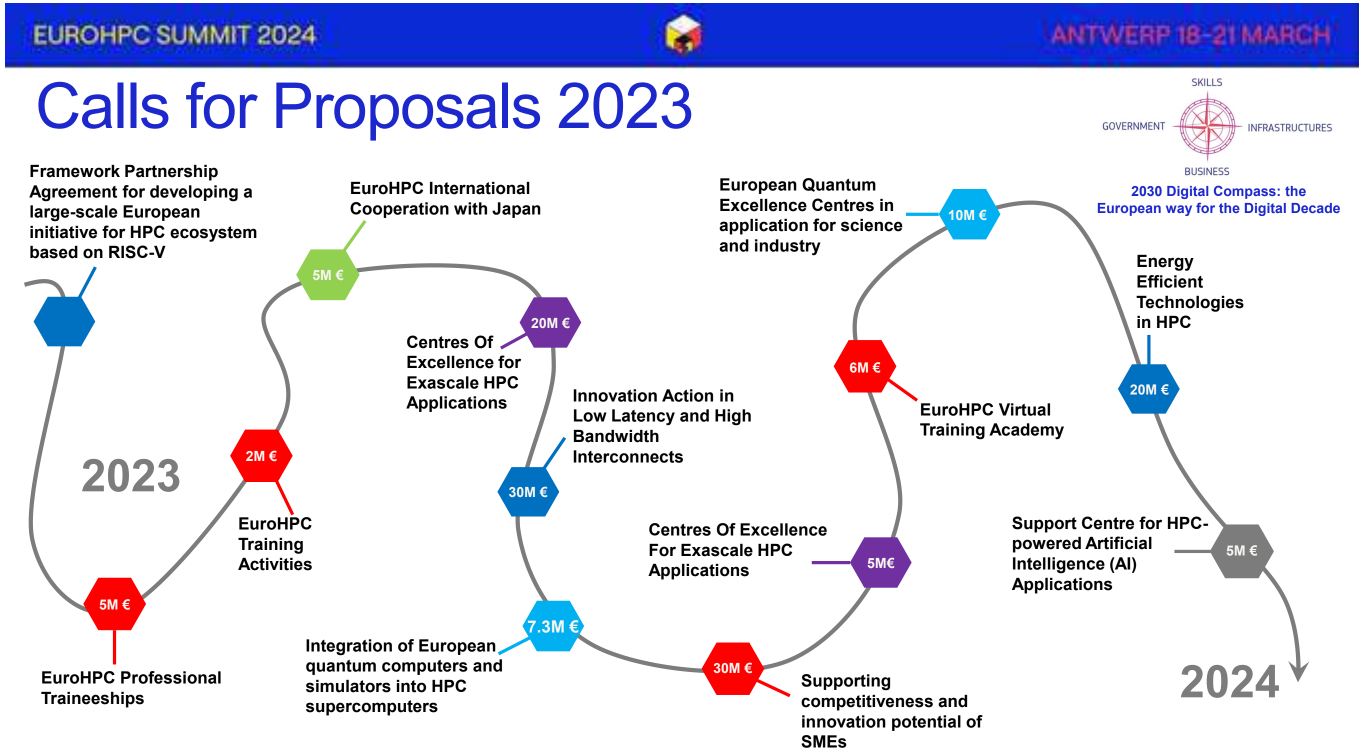


Several initiatives are in place or upcoming to support users, such as
and to harmonise the European HPC skills base, including **EuroCC2**, **Epicure**, **EUMaster4HPC** and the future **HPC VirtualTraining Academy**

The EuroHPC **Centres of Excellence**, focused on applications for specific sectors and upcoming **Quantum Excellence Centres**

The **SEA** projects are working to equip future Exascale systems with strong technologies, the recently closed **energy efficiency call** to develop energy-aware resource management technologies for HPC, and the upcoming **quantum middleware** initiative

EPI SGA2 which has led to the creation of the RHEA processor, which will be used in JUPITER
EuroHPC is also preparing initiatives for **RISC-V**

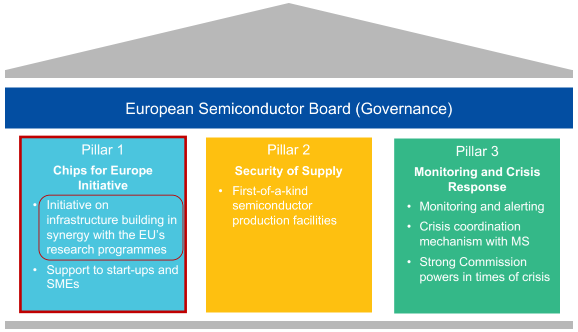


Example of Technological R&D: chip

EU motivations:

- Global semiconductor market value is 620B$ in 2024. **EU 20% of total market but only 10% of sales…**
- **CHIPS ACT** - create large innovation capacity and a resilient and dynamic semiconductor ecosystem
- **GenAI** has a strong demand for accelerators supporting training and inference of AI workloads
- No EU process technology in top500 → **close the gap**
- A couple of iniziatives
  - **CHIPS JU**
  - **EuroHPC initiatives**

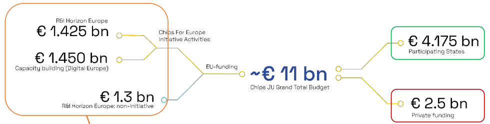
THE 3 PILLARS OF THE CHIPS ACT

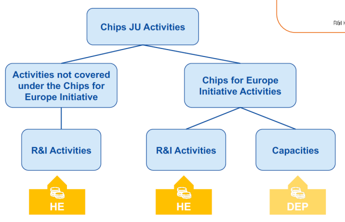


Jari Kinaret – 20 March 2024

EUROPEAN PARTNERSHIP

CHIPS JU





EU Contribution to Chips JU (excl	Chips for Europe Initiative	Non-Initiative	Total
R&I (Horizon Europe)	1.425	1.300	2.725
Cap building (DEP)	1.450	n.a.	1.450
Total	**2.875**	**1.300**	**4.175**

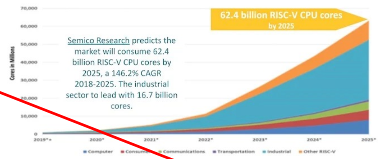
Jari Kinaret – 20 March 2024

EUROPEAN PARTNERSHIP

Technological Sovereignty

- **European Processor Initiative (EPI):**
  - EUPilot, EUPEX pilots integrating EU technology
  - Rhea (SiPearl) – powering JUPITER
  - EPAC - first European RISC-V acceleration
- **RISC-V – FPA (DARE)**
  - High-end general purpose processors
  - AI accelerators

Rapid RISC-V growth led by industrial



CHIPS AND COMPUTING: RISC-V

- **ECSEL heritage** - The ECSEL portfolio covers a variety of RISC-V aspects at a **project task level**
  - **Scope: architecture** extensions (e.g., accelerators, co-processors); (Low Power/High Performance) **microarchitectures** (e.g., implementations of the architectures);(Low Power) HW **realizations** (e.g., FD-SOI – By MEANS of?); **SW support** for RISC-V: System SW and tools for design, verification, testing, etc.
  - **ECSEL projects with RISC-V** tasks - **OCEAN 12 (2017-1-IA), CPS4EU (2018-1-IA), VALU3S (2019-2-RIA), FRACTAL (2019-2-SP2), Energy ECS (2020-1-IA), StorAIge (2020-1-IA), DAIS (2020-2-RIA) – most of these have also addressed AI.**
- **KDT JU/Chips JU RISC-V strategy – focused and linked actions**
  - Recommendations and **Roadmap for European Sovereignty** in Open Source Hardware, Software, and RISC-V Technologies , 2022
  - **SRIA update on RISC-V, 2022**
  - **Automotive RISC-V roadmap** – The Road towards a High-Performance Automotive RISC-V Reference Platform , 2023, updated 2024
- **KDT JU/Chips JU RISC-V calls**
- Call 2021-1-IA-Focus-Topic-1-Development of **open sources RISC-V building blocks**
  - **Project TRISTAN, 47 Partners,** *Total cost: € 54,371,711.93; Max HE Funding €15,597,798.00; National Funding: €13,603,678.17*
- Call 2022-1-IA Topic 3: Focus topic on Design of **Customisable and Domain Specific Open-source RISC-V** Processors (IA)
  - **Project ISOLDE, 39 Partners,** *Total cost: € 39,410,109.71; Max HE Funding €11,582,733.37; National Funding: € 11.451.467,64*

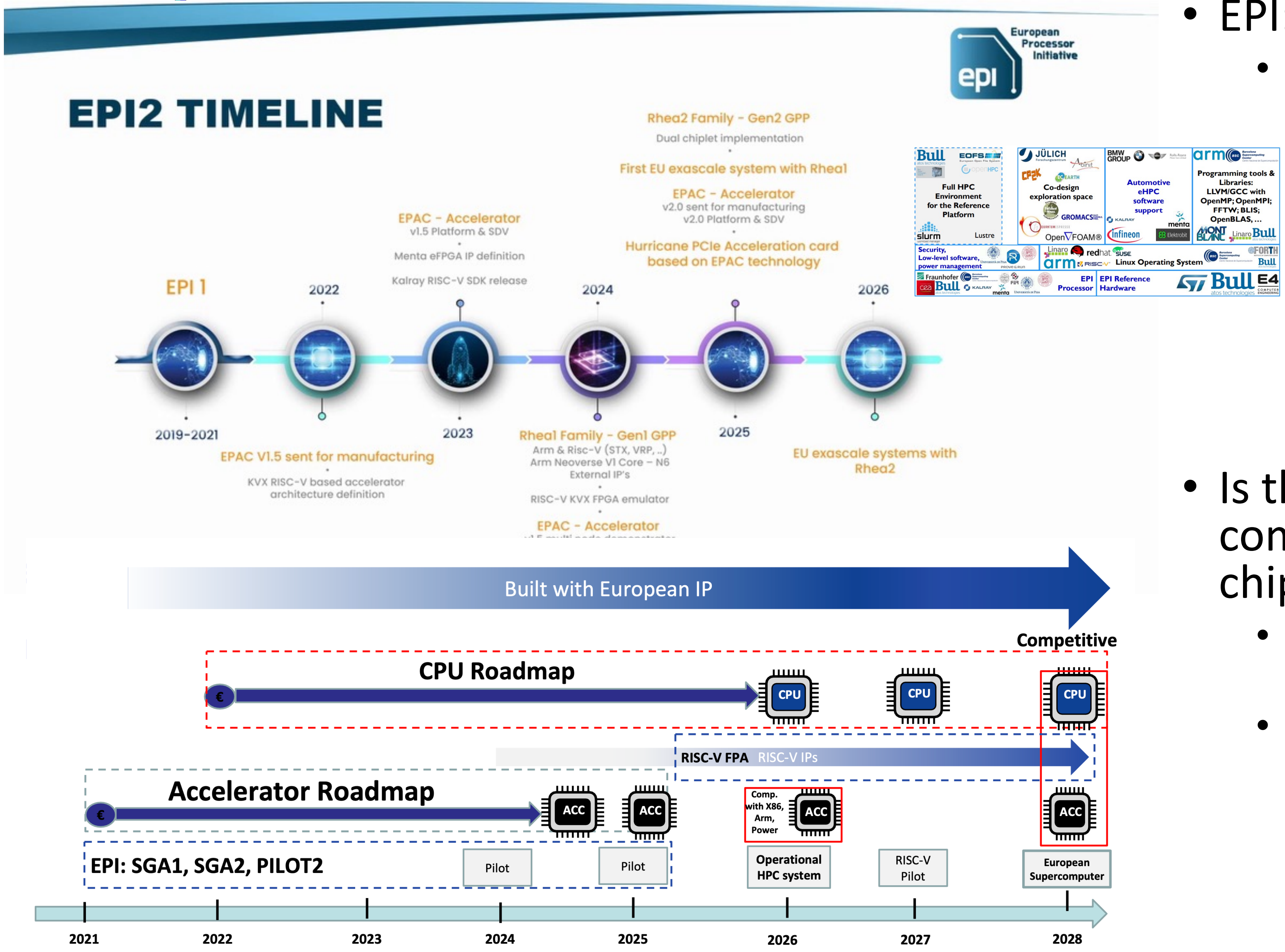
Chips JU investment in RISC-V so far (2 projects contracted):
Total Cost: ***€ 95M***, HE Funding: ***€ 27M***, National Funding: ***€ 25M***, Private in-kind: ***€ 43M***

- **Now Open!** Call 2024-1-IA Topic 2: Focus topic on **High Performance RISC-V Automotive Processors** supporting the vehicle of the future -
  - **Expected:** *Max 70 partners, Approx. cost: **€ 60-80M**; Max HE Funding **€ 20M**; Max National Funding: **€ 20M***

EUROPEAN PARTNERSHIP

Is there enough money to support the "internal" competition????

EuroHPC Chips: EPI (European Processor Initiative)

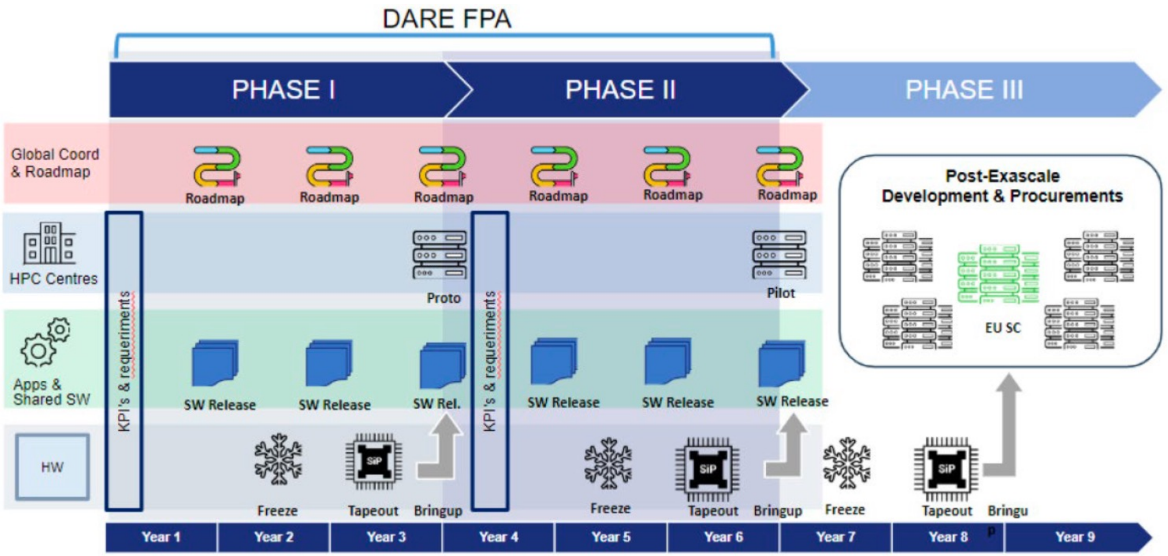


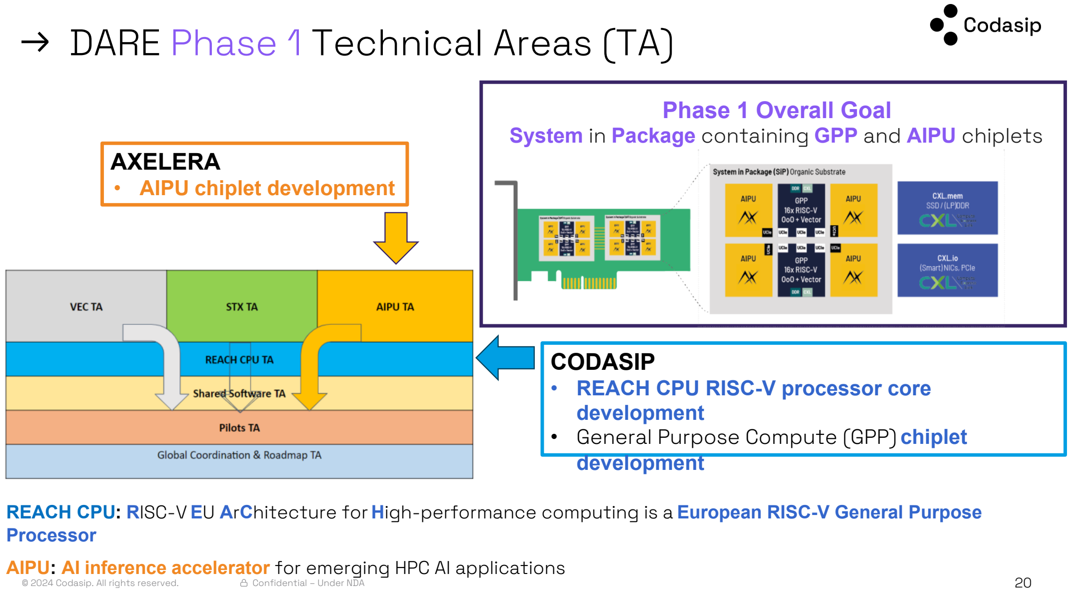
- EPI: European Processor Initiative
  - Public/private funds budget of 1.5BEuro
    - multi phase project: from ARM to RISC-V per CPU and accelerators
    - At the beginning academic and industrial R&D (SGA1 and SGA2 ~200+200MEuro). Then technology transfer (SiPearl)
    - First product, RHEA GPP ARM-based, (almost…) ready in 2024 202
- Is the effort big enough for real competition with US/JAPAN chipmakers?
  - **Probably not** for a TRL 8-9 product in mass production
  - **BUT it's needed** to guarantee technological control and to allow the development of new ideas, architectures, hardware and software, and new application fields

EU chips targeting EU HPC: DARE

DARE (EuroHPC FPA) consortium *aims to establish a clear path for software and hardware development in Europe, leveraging early access to **RISC-V** hardware emulation and simulation, with the goal of deploying the developed technologies in EuroHPC systems*.

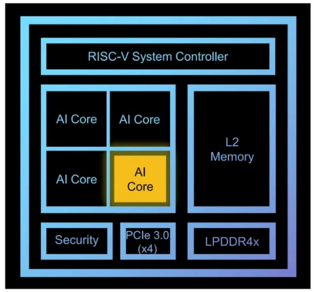
- Started Q1/2024**,** is a 10 years initiative divided in 3 subsequent phases:
  - Phase 1 Design&Proto: Prototype development @ 7nm process node
  - Phase 2 Pilot: Medium scale Pilot development @4nm process node
  - Phase 3 Production: …..
- Several synergic Technical Areas (TA): GPP (CPU), Accelerators (Vec, AI), SystemSoftware&Applications, Pilot Integration
- Industrial & academic consortium. TA leaded by industrial partners



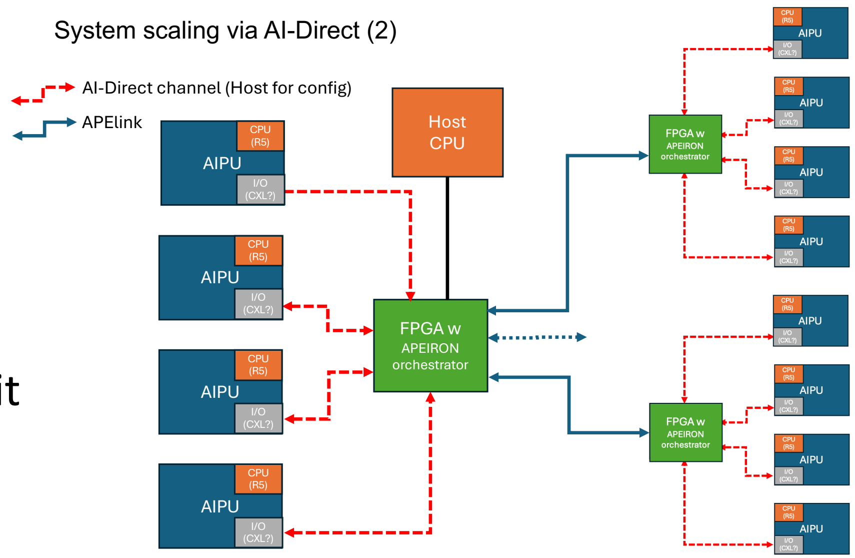


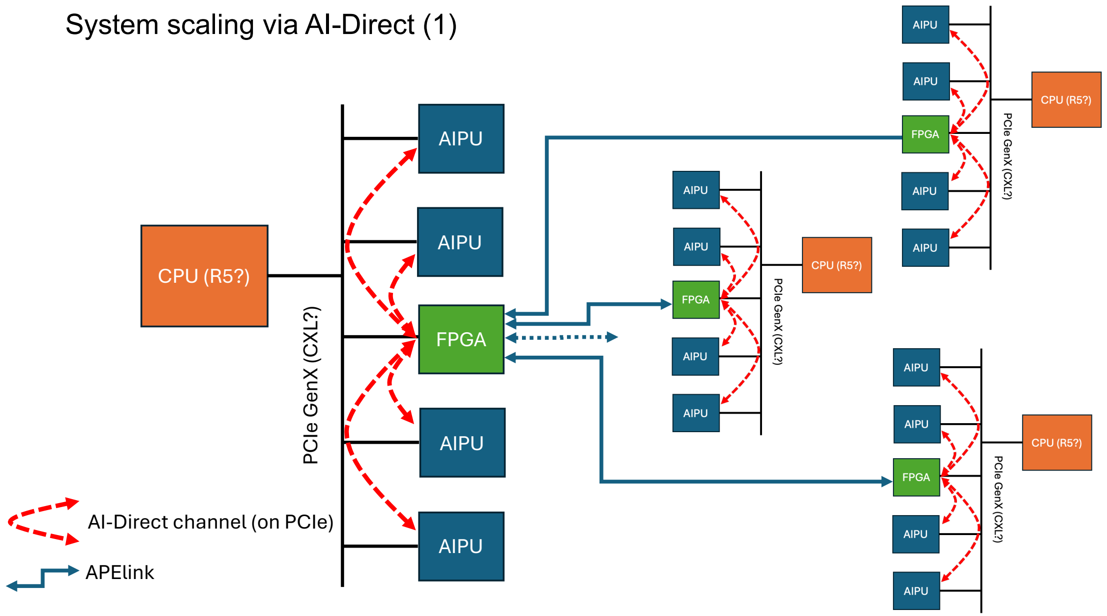
INFN (in ICSC) in DARE

- INFN (APE group) will contribute in DARE as an **affiliated partner of ICSC.**
  - **Forced by** due to the new restrictive EuroHPC funding rules…
- Different areas of development targeting same scientific/technological problems:
  - hardware IP (on FPGA) and its companion system software (linux device driver, user library) enabling the deployment of large scale NN models over multiple AIPU accelerators to boost performance of applications like AI-accelerated HPC and Generative AI.
- Three main pillars:
  - AI-direct engine
    - Specialized HW to provide high throughput/low latency access (on PCIe and/or custom direct channel) between different AIPU
  - APEiron-based orchestration
  - Scalable applications to benchmark distributed parallel solutions
    - NEST: brain-inspired neural network scalable simulator (brain modelling at large scale, HBP flagship)
    - RAIDER: High Energy Physics ML-based applications for particle tracking, identification and calorimeter clustering



AXELERA AIPU
AI Processing Unit



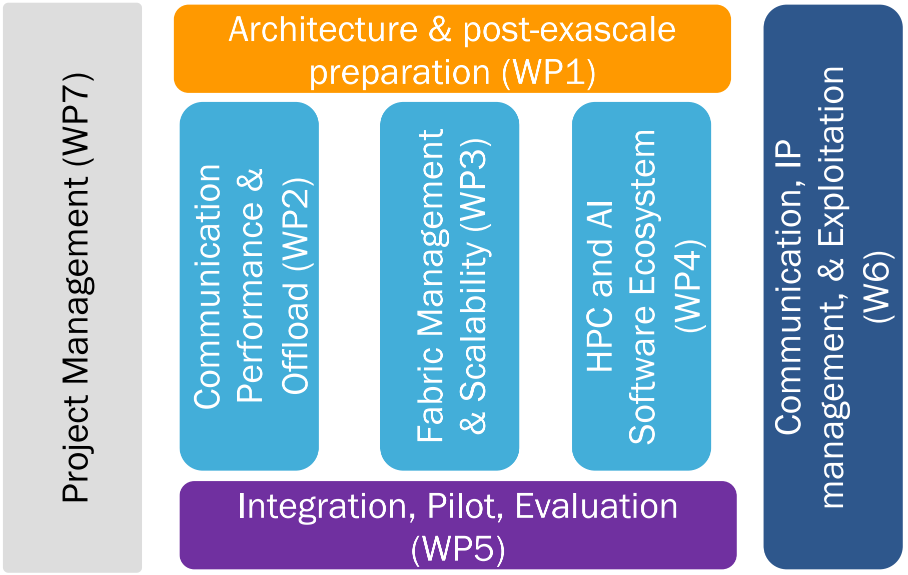


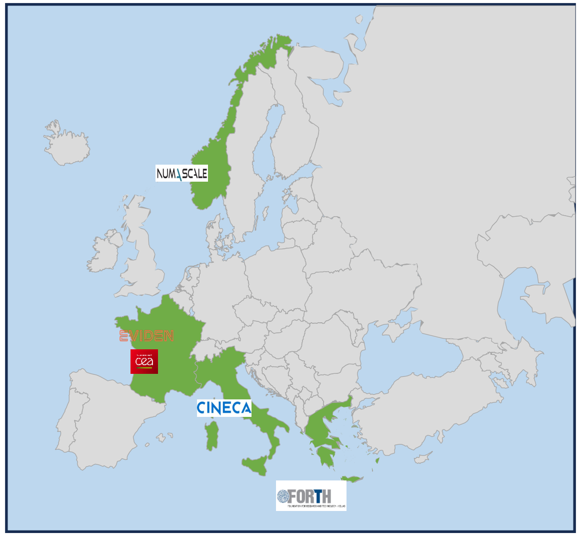
EU interconnect targeting EU HPC: NET4EXA

NET4EXA (Network for EXAscale systems**)** aims to develop a next-generation high-speed interconnect for **HPC and AI** systems, building on the success of the **BXI European HPC Interconnect** and the advancements made through research in the **RED-SEA** project and other previous European RIA initiatives.

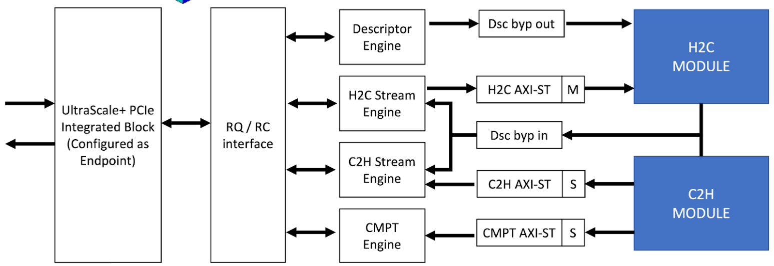
- EuroHPC Call: HORIZON-EUROHPC-JU-2023-INTER-02
  - Type of action: HORIZON-JU-IA HORIZON JU Innovation Actions (w/ TRL 8)
- Total costs : 71 126 351 €;
  - EU funding: 26 916 520,70 €;
  - + countries' funding
  - + in-kind contribution for industrial beneficiaries.
- Project Start date: Sep. 1st, 2024; Duration: 30 months

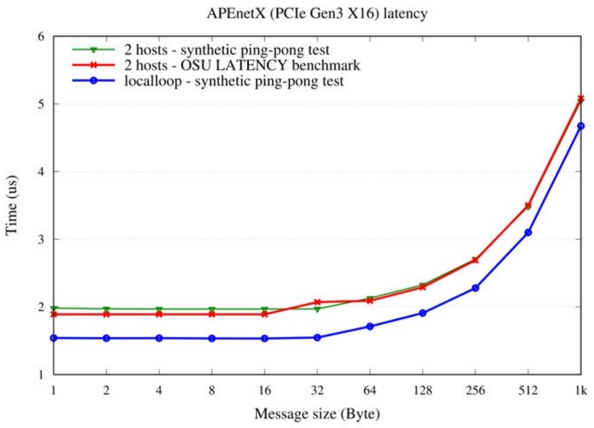
TYPE	NAME	Country
Large company	1 - BULL	FR
SMEs	2 - NUMASCALE AS	NO
	4 – Subco SCINTIL	FR
	4 – Subco Spearl	FR
	5 – Subco	IT
Large Datacenters & Research centers	4 - CEA	FR
	5 – CINECA	IT
Academic partners	3 – FORTH	GR
	5.1 CINECA – UNITRENTO	IT
	5.2 CINECA – UNIROMA1	IT
	5.3 CINECA - INFN	IT

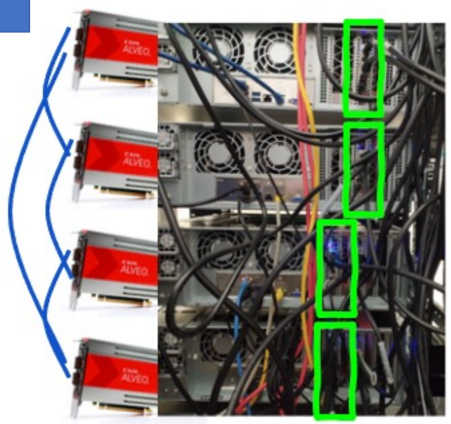


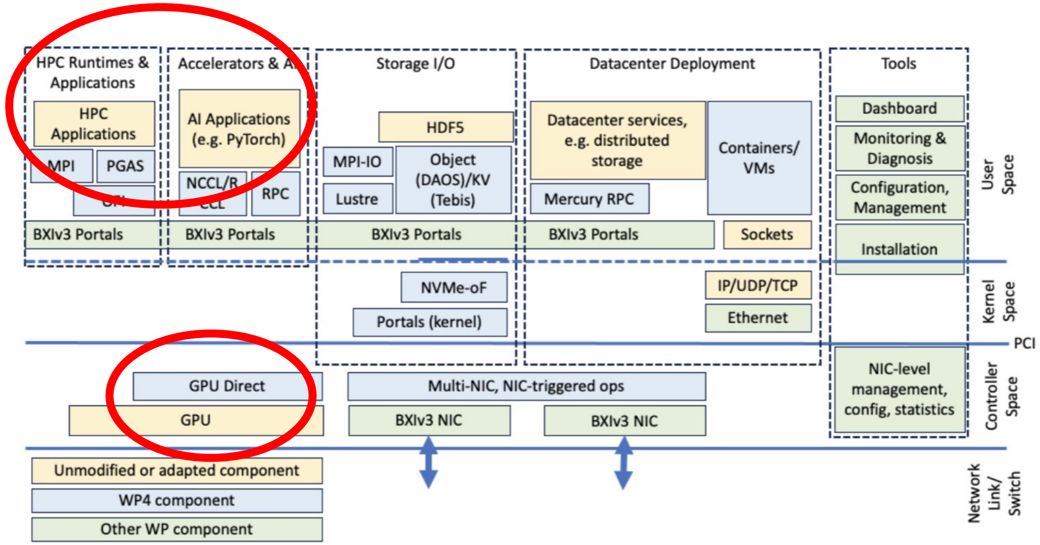


INFN (in CINECA) in NET4EXA









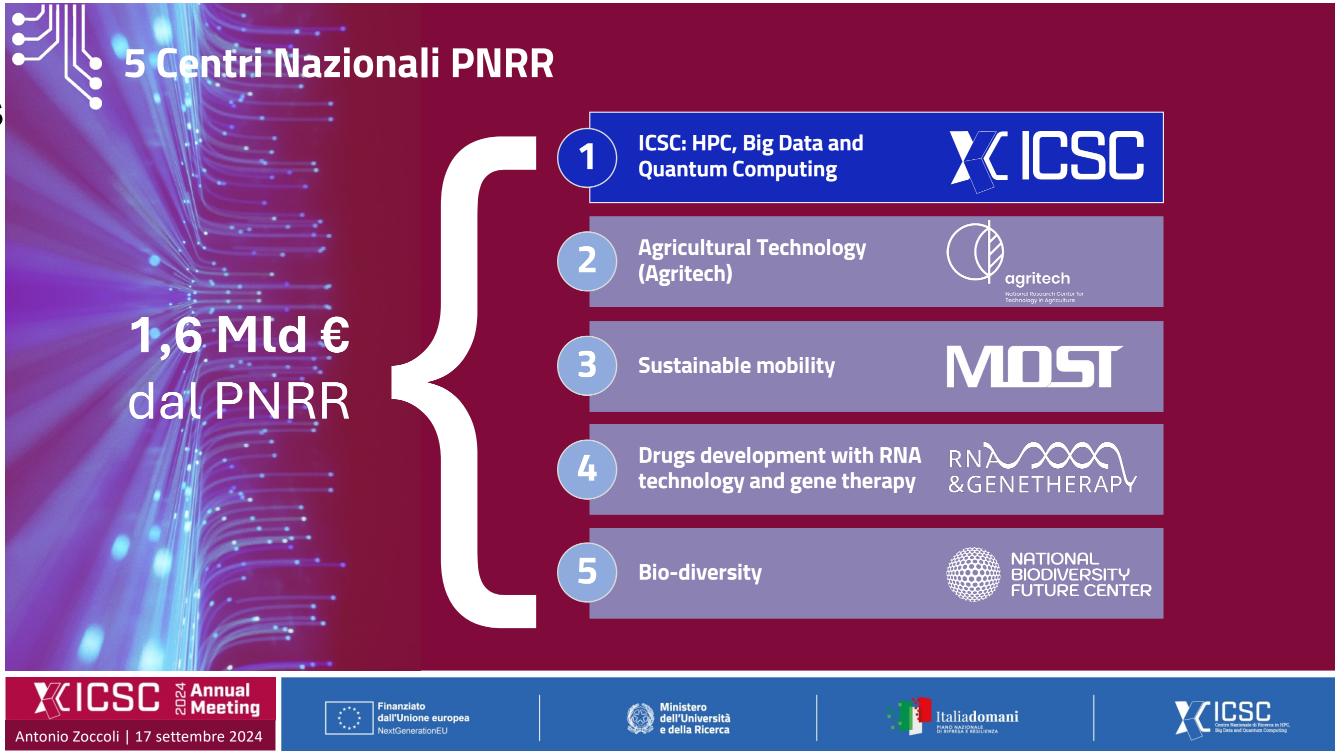
- INFN contribute in NET4EXA **as affiliated partner of CINECA**
- Leverage on previous project results:
  - RED-SEA, TEXTAROSSA, INFN APEnet
- Several areas of technical contribution
  - Integration of a medium scale (16-32) FPGA-based testbed
  - innovative mechanisms to enhance congestion control management (for BXIv4)
  - ON-NIC processing for task streaming computing,
  - prototyping new features supporting GPU triggered computing in BXIv4 via and for BXIv4network architecture, I
  - INFN key applications for benchmarking network architecture under design: NEST (Large scale brain simulation), RAIDER (HEP AI-oriented apps)
- INFN budget ~1.35 Meuro
  - 500 kEuro personnel, 850kEuro HW procurement
  - 50% co-funded by Italian Government

EU funds: NextGenEU (PNRR) and ICSC National Research Centre

ICSC National Research Centre for High Performance Computing, Big Data and Quantum Computing

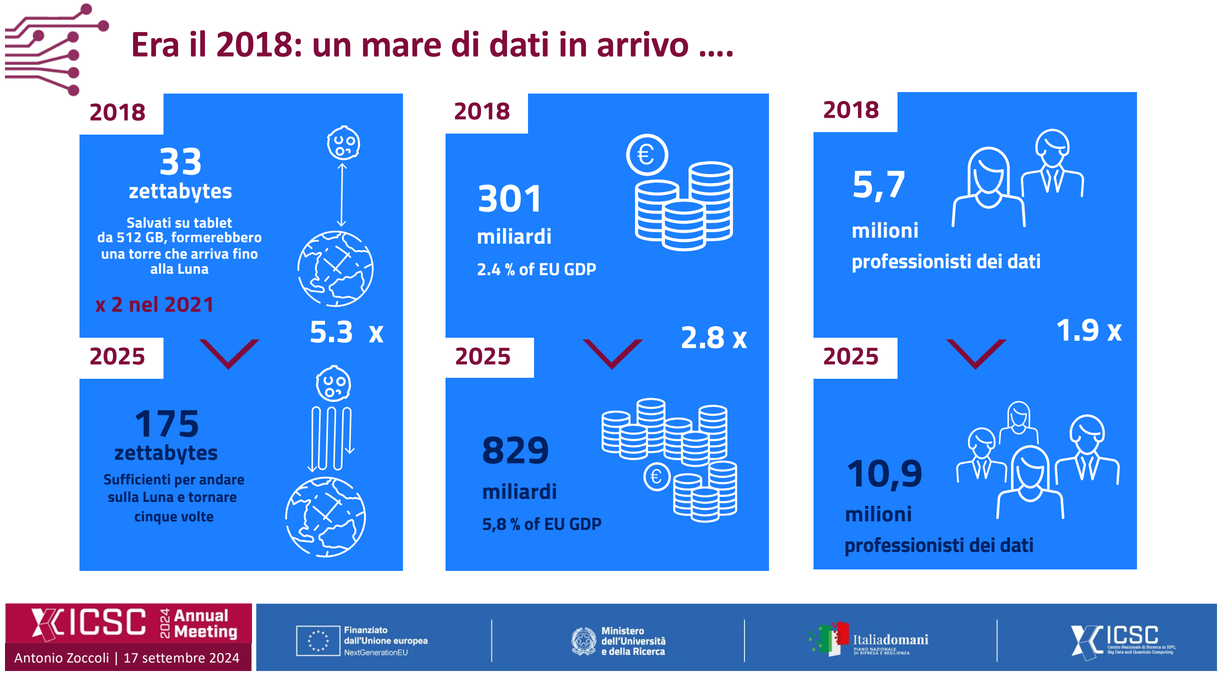
The Center conducts R&D, nationally and internationally, for ***innovation in high-performance computing, simulations, and big data analytics****. This aim is pursued through a* ***state-of-the-art infrastructure for high-performance computing and big data management****, which leverages existing resources and integrates emerging technologies, and an organization* ***and distribution of activities based on the Hub and Spoke model****.*

- One of the five National Centres established by the National Recovery and Resilience Plan (PNRR)
- Total Investment: 320 MEuro
- 51 Academic & industrial partners
- Lead by INFN with 51 Academic & industrial partners
- 11 thematic areas (Spokes)



ICSC motivations

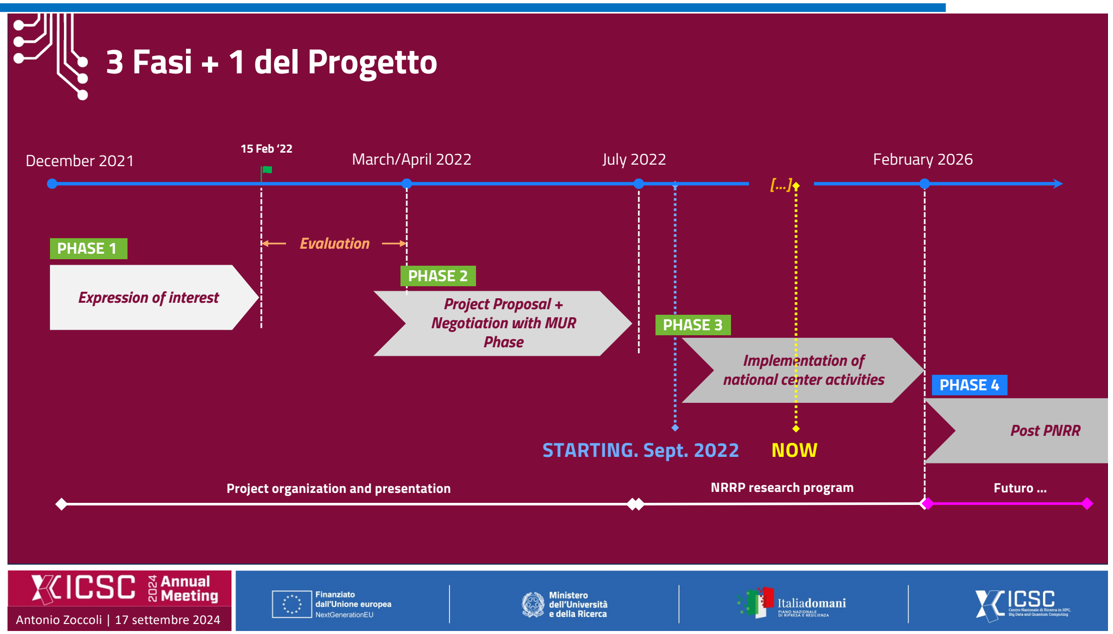
What's better than the boss's slides?

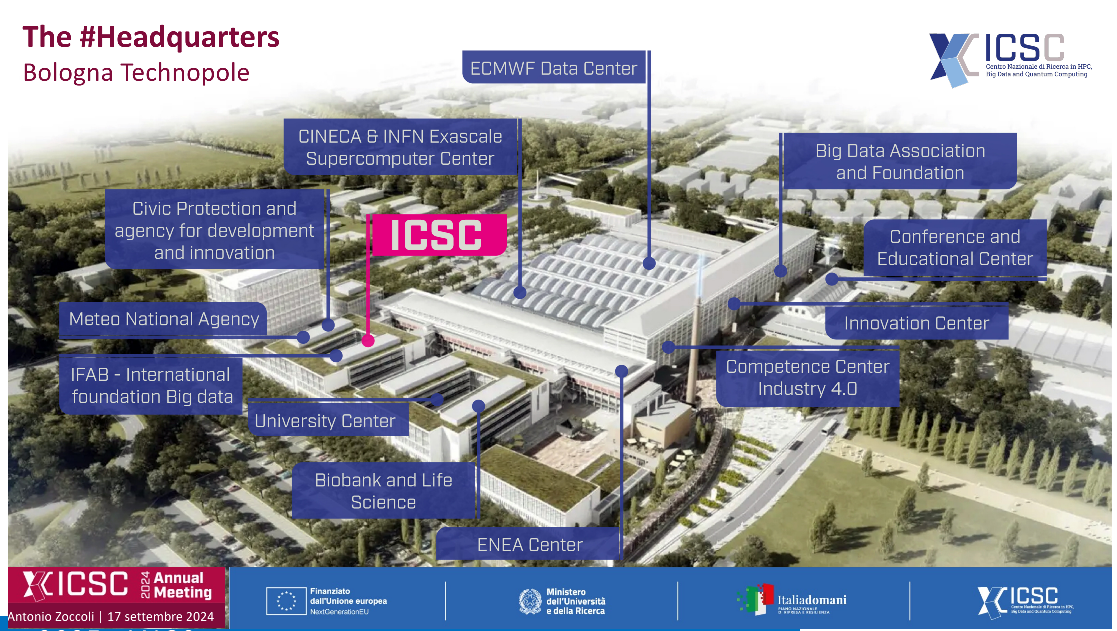


Why ICSC?

- Answer to the modern computing and data analytics challenges emerged by strategic sectors for the development of the country: i.e., simulations, computing, and high-performance data analysis
- Establish a National level hub
- Deploy a shared and open cloud/HPC infrastructure, representing a unique strategic asset for Italy and EU
- Promote the best interdisciplinary skills of science of engineering from basic research to computing sciences

ICSC organization



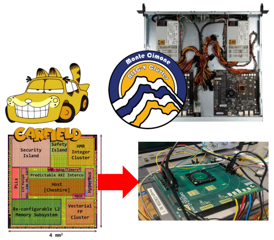


ICSC: highlights..

#Spoke 1 – Future Hpc & Big Data

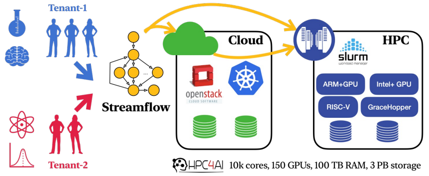
Budget dello Spoke	21.859.389 €
Personale Massa Critica	204
Personale reclutato	76
N. Pubblicazioni	244
N. Progetti innovazione	13

Living Lab «Hardware & Systems» HWS@UNIBO
- Integrazione di **MonteCimone, il primo cluster al mondo** RISC-V al mondo (in collaborazione con E4)
- Progettazione di processori per mercati specifici sviluppati a partire dalla piattaforma PULP/RISC-V : **CARFIELD** (automotive Intel16nm, 11/23), **ASTRAL** GF 12nm (spazio GlobalFoundries 12nm, 11/24, IG TASI)



Inaugurato a giugno 23

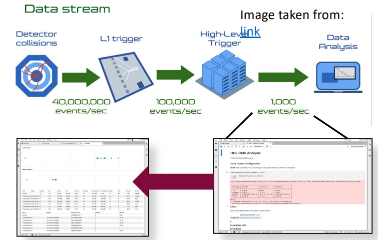
Living Lab «Software & Integration» SWI@UNITO
- Sviluppo (su MonteCimone) **della prima distribuzione completa al mondo di Pytorch** (Google+FB) per RISC-V – il software più utilizzato per AI/LLM. Oggi mainstream.
- Progettazione e sviluppo di **Streamflow** – workflow portabili per sistemi multicloud-HPC. EU innovation radar award. Utilizzato da 4 IG (ENI, Sogei, Unipol, iFAB), in valutazione per adozione: IBM, TIM, Astron, MiRRI EU ERIC



#Spoke 2 – Fundamental Research & Space Economy

Budget dello Spoke	18.939.814 €
Personale Massa Critica	193
Personale reclutato	67
N. Pubblicazioni	340
N. Progetti innovazione	10

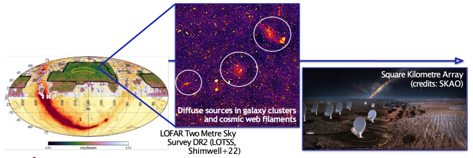
Analisi veloce su grandi basi di dati **(Petab+)**, con infrastruttura eterogenea **(Cloud + HPC + Grid)** e distribuita.
Disegnata e validata per la fisica a LHC, usabile anche in altri ambiti di ricerca (es. fisica medica) e industriali (es. immagini della space economy)



Taglio dei tempi di prototipizzazione in interattivo e accesso trasparente e ottimizzato a grandi basi di dati, anche remote: verrà resa disponibile a tutto ICSC.

#Spoke 3 – Astrophysics & Cosmos Observations

Budget dello Spoke	12.655.379 €
Personale Massa Critica	97
Personale reclutato	48
N. Pubblicazioni	130
N. Progetti innovazione	9



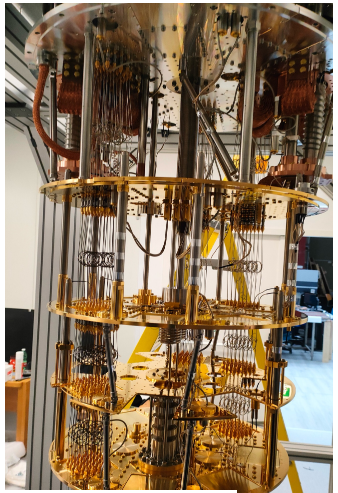
Big Data: processing e management
- A supporto dei grandi esperimenti: **SKA, EUCLID, CTA+, Fermi, LiteBird, LOFAR, …**
- Sviluppo di soluzioni rivoluzionarie di archiviazione, processamento e analisi di grandi volumi di dati **basate sul AI e capaci di sfruttare sistemi HPC stato dell'arte.**
- Sviluppo di soluzioni avanzate e di **strumenti di visualizzazione e analisi dati interattivi e collaborativi.**

Simulazioni Numeriche Exascale e oltre
- **Codici numerici astrofisici innovativi** capaci di sfruttare i più innovativi sistemi ibridi di calcolo massicciamente paralleli ed accelerati.
- **Algoritmi sofisticati, integrazione di soluzioni AI, altissima risoluzione** per problemi complessi in cosmologia, astrofisica e fisica dello spazio.

#Spoke 10 – Quantum Computing

Budget dello Spoke	30.578.631 €
Personale Massa Critica	181
Personale reclutato	62
N. Pubblicazioni	130
N. Progetti innovazione	11

Primo computer quantistico a semiconduttori a 24 qubit costruito in Italia



Inaugurato il 29 Maggio all'Università di Napoli

Prossimi obiettivi
- **40 qubits** entro Ottobre 2024
- Connessione con Cineca
- Accesso cloud

ICSC highlights: Terabit, workout gym for next gen HPCers...

The “HPC Bubbles”

- Progetto ICSC+ TERABIT+DARE: "HPC a tutte le scale”
- HPC Bubbles: disponibilità di risorse e servizi HPC Cloud-native, scalabili ai livelli IaaS, PaaS e SaaS.
  → integrazione tra rete, big data, cloud e risorse HPC.
- Comunicazione e federazione tra le HPC Bubbles

e altre infrastrutture HPC.

ICSC 2024 Annual Meeting
Antonio Zoccoli | 17 settembre 2024
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
ICSC

What are the HPC Bubbles, in practice?

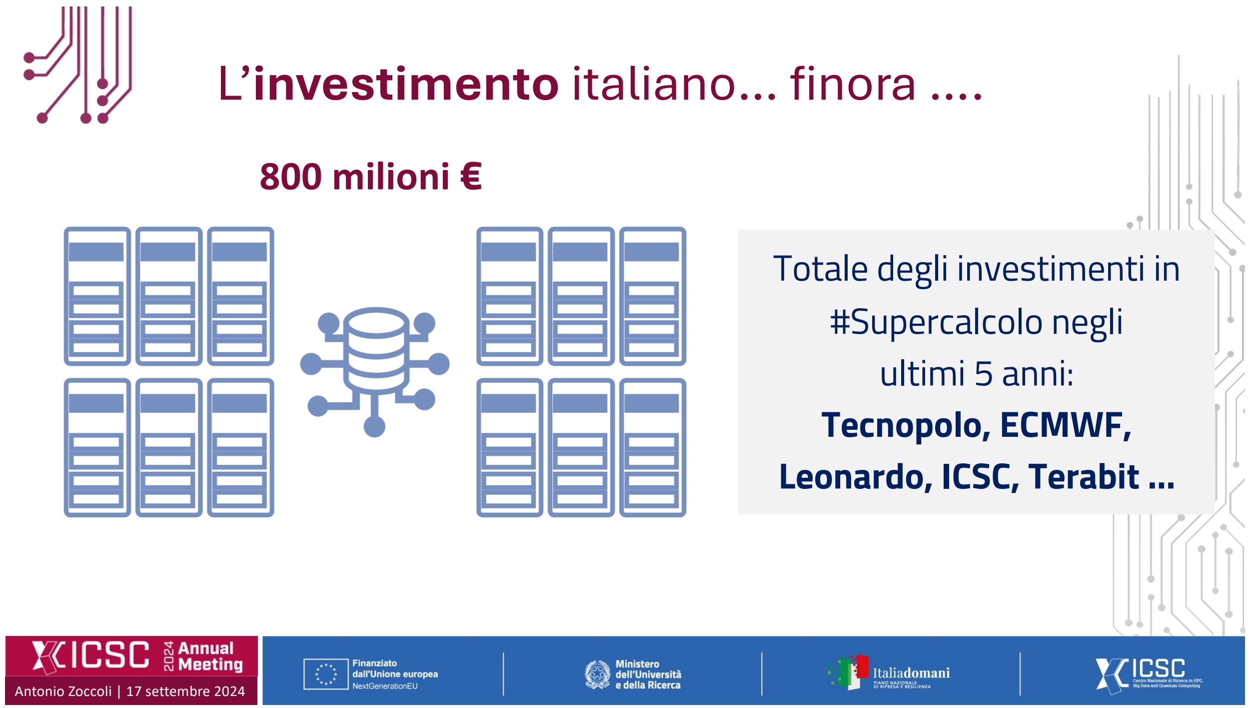
Tre tipi di HPC Bubbles:

- Cluster HPC modulari per l'IA, (8-16 nodi da 4 GPU NVIDIA H100) Globalmente, raggiungono la potenza di circa 3,8 PetaFLOP (FP64).
- Cluster HPC modulari generici ad alte prestazioni (8-16 nodi da 192 core CPU ad alto rendimento e 1,5 TB di RAM) Globalmente, forniscono circa 30.000 core di calcolo.
- Cluster HPC modulari basati su FPGA, (32 core CPU e 4 FPG) Globalmente, si tratta di 40 FPGA e 320 core.

Storage veloce e interconnessioni a bassa latenza, installate su più siti dell'infrastruttura cloud distribuita

Alcune HPC Bubbles sono presenti anche su zone cloud certificate ISO-INFN

ICSC: what next?



Post exascale

Post Exascale computing or (better) “beyond” exascale: Europe is a little bit behind…

The (post)-Exascale race, where are we?

KeyNote - Long-term Computing Vision 2024 ETP4HPC Conference

ECP EXASCALE COMPUTING PROJECT

ECP : DOE funded, NSF support (end dec. 2023)
+ Creation of 6 co-design centers
Still a challenge: Exascale ready app, sustainable software stack
New perspectives: Gen IA for Science, Trillion Parameter Consortium, 20B€

China initiatives:
– development of applications in preparation for the arrival of the Tianhe3 machine.

Japan initiatives:
– FugakuNext prod. 2029 (xB$): co-design HW/SW/Apps
– Next Gen AI (x100M$), Quantum-HPC (140M$)
– (post)-Exascale as a Service (AWS/Rikean): from Fugaku to Virtual Fugaku

A strong effort in both hardware, software and applications/co-design

Exascale and Near-Exascale Leadership Systems (2020 to 2028)

Year Accepted	China	Europe	Japan	US	Other Countries*	Total Systems	Total Value
2020			1 near-exascale system ~$1.1B			1	$1.1B
2021	2 exascale ~$350M each	1 pre-exascale system ~$180M	--	1 pre-exascale system ~$200M	--	4	$1.1B
2022	1 exascale ~$350M	2 pre-exascale systems ~$390M total	--	1 exascale system ~$600M (2/3 accepted 2022)	--	4	$1.1B
2023	1 exascale system ~$350M	1 or 2 pre-exascale systems ~$150M each	1 near-exascale system ~$150M	Remaining 1/3 of Frontier system	--	4-5	~$1.0B
2024	1 exascale system ~$350M	1 exascale ~$350M, plus 1 exascale (or pre) system ~$200M	?	2 exascale system ~$600M	1 pre-exascale system ~$125M	5-6	~$1.6B
2025	1 or 2 exascale systems ~$300M each	2 or 3 exascale systems ~$350M each	1 exascale system ~$200M	1 or 2 exascale systems ~$350M each	1 near-exascale system ~$125M	6-9	$1.7B - $2.7B
2026	2 exascale systems ~$300M each	2 or 3 exascale systems ~$325M each	?	1 or 2 exascale systems ~$325M each	1 or 2 exascale systems ~$150M each	6-9	$1.7B - $2.5B
2027	2 exascale systems ~$275M each	2 or 3 exascale systems ~$300M	1 exascale system ~$150M	1 or 2 exascale systems ~$275M each	2 or 3 exascale systems ~$130M each	8-11	$1.8B - $2.5B
2028	2 exascale systems ~$250M each	2 or 3 exascale systems ~$275M	1 or 2 exascale systems ~$150M each	1 or 2 exascale systems ~$275M each	2 or 3 exascale systems ~$125M each	8-12	$1.7B - $2.6B
Total	12-13	14-19	5-6	8-12	7-10	47-61	$13.4B - $16.8B

* Includes S. Korea, Singapore, Australia, Russia, Canada, India, Israel, Saudi Arabia, etc.
Note: After 2023, many exascale systems will be 2-10 exascale.
Source: Hyperion Research, March 2024

Post exascale challenges (SW)

The digital continuum: open challenges

Unification of HPC Simulations/Big Data/AI towards a data-centric view

Moving, storing and processing data across the continuum: how to deal with the 3 Vs of Big Data?

- **Extreme Volume across the continuum**
  - Support the access and processing of “cold”, historical data and “hot”, real-time data + (virtually infinite) simulated data
- **Extreme Velocity across the continuum**
  - Unified real-time data processing (in situ/in transit, stream-based) in a common software ecosystem
  - Need disruptive reduction in data movement cost with new devices, packaging
    - All real-time data may not be storable in archives => real time training (bandwidth-oriented)
- **Extreme Variety across the continuum**
  - Unified data storage abstractions to enable distributed processing and analytics across the continuum
  - Interoperable data formats, "Semantic interoperability” through shared ontologies
- Digital Continuum is a multi-tenant and multi-owner environment.
  - Collected data used with multiple purposes
  - Computing Infrastructure is also shared

AI4Science / Science4AI

AI for science – towards HPC/AI hybridization

- End of Moore’s law, **develop hybrid approaches based on AI** to accelerate parts of scientific computing applications

=> 10 Exascale performance with x25 EDP+Data science pp and x42 HW improvements, some AI-based solvers can be sped up by 6 orders of magnitude, etc., weather forecast with Graphcast

- **Hybridization of HPC SW with AI :** physics-informed AI models for simulation codes, observational data reduction, digital twins.
- **Push forward a post-Exascale-ready SW stack embedding AI solutions** that answer the needs of the application communities

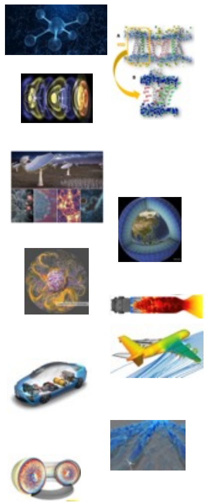
Software/application co-design

Key challenges

- How to get post-Exascale ready applications ?
- How to expand an application-driven SW stack ?
- How to make applications portable and sustainable at the post-Exascale era ?

International context

- Early-binding HW/SW/application co-design approach at Rikken (Japan)
- In the USA, DOE co-design centers were a key component of the Exascale Computing project (ECP)
- Inspired by ECP, co-design is central in the NumPEx project (FR)



Astronomy & Astrophysics
Earth System Models & Climate
Earth sciences & environment
Computational biology & Life science
Laboratory laser-plasma physics
High-energy particle physics
Quantum chemistry and materials
Digital health
Environmental & societal risks
Urban systems planing
Magnetically confined fusion plasma
Sustainable Transport & mobility
Energy production & transport

2/15/24

Take away messages

Software, the new frontier

Consolidating and accelerating the construction of a sovereign European **exascale software stack** (portable, interoperable, reproducible, sustainable)

Support and foster the developpement of disruptive Math & models

AI4Science – Science4AI

Push an **hybrid AI/HPC software stack**, to accelerate HPC and provide AI at scale

Support AI for Science, foster fully open AI use-cases/benchmarks, **not restricted to GenAI**

From edge to HPC system: the digital continuum

Coordinate efforts to share workflows, solutions and services for the convergence of HPC/Cloud/Edge

EaaS: **Exascale as a Service**, for Tier-0 European systems

Develop a data-everywhere, FAIR, ecosystem in Europe

Software/application co-design

HW/SW/application co-design to help the communities get prepared for post-Exascale

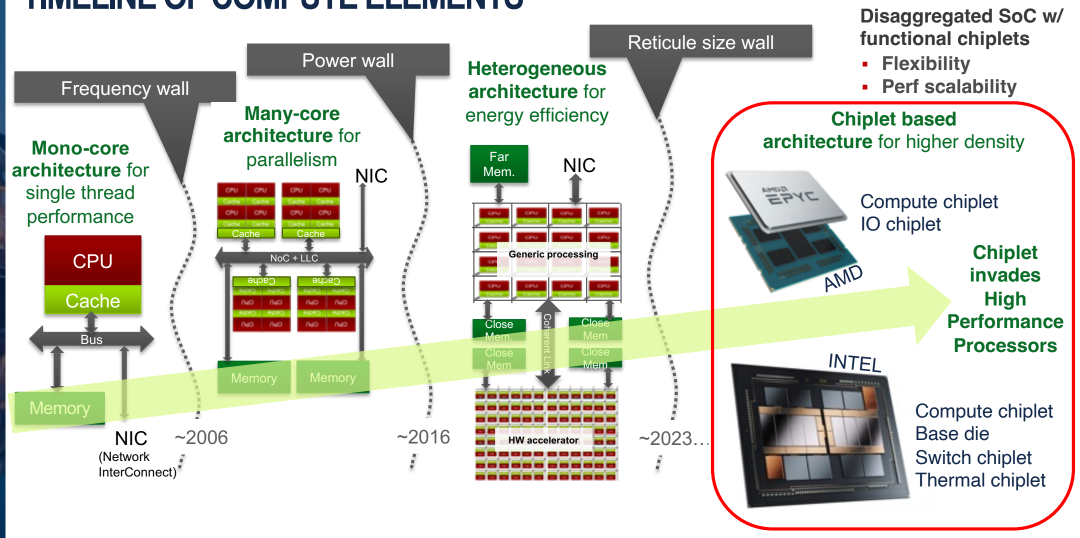
Foster the use/reuse of modular/interoperable and portable SW components

Push sustainable SW development model

Post exascale challenges (HW, disruptive tech, federation)

ETP4HPC SRA WG

TIMELINE OF COMPUTE ELEMENTS



ETP4HPC 43

Potential challenges for 2024-2028

Limit data transfer costs:

- **Limit the length of data transfers**
  - Unifying memory access of CPUs and accelerators
  - Chiplets on (photonic) interposers,
  - Programming models that can transparently benefit from heterogeneous computing elements.
- **Improved memory hierarchy**:
  - HBM (low latency, high bandwidth),
  - DRAM DIMMs (low latency, mid-size capacity),
  - NVM DIMMs (high capacity, persistency).

  will require SW support for placement of data.
- **Computation close to memory.**
  - Near-memory processing (HMC specs), Samsung (HBM PIM and AxDIMMs), Hynix (Computational Memory Solution--CMS), UPMEM (Data Processors - DPUs).

Improved Compute:

- **Efficient accelerators close to orchestrator**
  - Smooth integration of programming environments.
- **Efficient data types (variable precision)**
  - High precision to converge algorithms / avoid num. drift.
  - Mixed Precision aligned w compute needs
  - Lower precision (bfloat16, float8…) for AI.

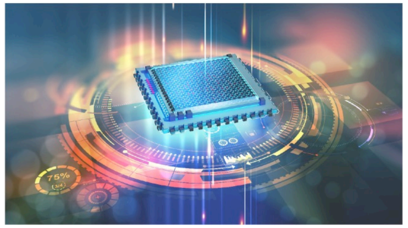
Other challenges:

- **Aggressive power saving techniques**
  - Dynamic power management
  - Resource management
  - Component monitors supporting Rack/system level power steering
- **Support new addressing schemes**
  - Byte addressing
  - Key-value (associative access)
  - Sparce matrices (gather – scatter)
  - HW supported multilevel indirect addressing

ETP4HPC 45

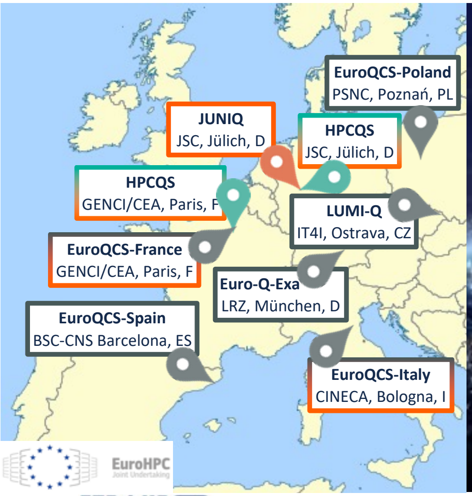
Integration of QCS in HPC Systems

- Integration of quantum computers and simulators (QCS) in HPC systems on a
  - system level: *loose* and *tight* models
  - programming level: full hardware-software stack
  - application level: optimisation, quantum chemistry and quantum ML
- Application-centric benchmarking
  - Test for the algorithm, the software stack and the technology
- Emulation of QCS with HPC systems
  - Ideal and realistic QCS
  - Designing, analysing and benchmarking QCS and quantum algorithms



59 ETP4HPC ETP4HPC Seminar 15/02/2024

EU Federated Hybrid HPC – QC Infrastructure



Developing and providing HPC – QC services for the benefit of all Europeans

Providing uniform, easy and affordable access to HPC – QC resources

Training of specialists in quantum computing and its potential applications

First of a kind in the world

61 ETP4HPC ETP4HPC Seminar 15/02/2024

Comments not conclusions..

- HPC is still the way to advance fundamental scientific and engineering research allowing to tackle simulation of larger and larger problems
- New paradigms and new approaches to large-scale computing have strongly emerged
  - Generative AI and LLMs (in the short term) and Quantum Computing (in the longer term) are fuelling a new level of growth
  - Also sustained through public funding and HPC cloudification
- There is no a clear winner one-for-all technology:
  - Too many expectations for AI/ML with the side effect to reach economic un-sustainability
  - QC exhibits low maturity (Tech and apps)
- So, the answer for current and next HPC systems is *convergence*: MSA architecture, where Classic HPC, QC and ML are tightly integrated
- BUT the scale (exascale and beyond), the complexity and the heterogeneity is making a nightmare the process of design, integrate and operate systems. It needs a lot of R&I to optimize
  - HW: ARM/RISC-V CPU low cost for power saving; innovative accelerators architecture for computing efficiency; interconnect networks for high throughput and low latency; storage architecture from “high speed and fast feed” to Data platform; fault tolerance,...
  - SW/APP: vertical software stacks for effective programmability; new applications & algorithms able to exploit the systems
  - Infrastructure: HPC as a service, new datacenter,...

Comments not conclusions..

- The added value is again human resources: researchers, technologists, computer architects, application experts, system managers...
- BUT, in general, we measure the growing scarcity of HPC experts due to:
  - Ageing of personnel, long training process of new staff, impressive rate of new technologies introduction
  - Misperception that any stuff related to HPC, including “human experts”, can be easily bought on the market...
- Academia, Research and IT industries, as users and technology providers, have to team to support this training process reversing the trend.
- National and International (EU) funding agencies must commit to funding not only HPC systems, but also this training process and R&D initiatives with a clear plan shared by the entire community
- INFN is on the right way, with its leading role in ICSC and contribution to R&I/R&D EU initiatives
  - govern them with a long-term view while avoiding the sloppy management of various short-term opportunities (GRID, PNRR...)

APE LAB TEAM

- 18 members (11 staff + 2 fixed-term + 5 PHD)
- 3 main research lines
  - HPC (system architectures, scalable networks, apps optimization)
  - Neuroscience (brain simulations, models, neuromorphic systems)
  - HEP Computing (Read-out systems, online trigger)
- Know-how
  - ASIC design, FPGA design, GPU programming and integration, network design, dense system integration, parallel programming and application coding (LQCD, neural networks, complex systems), system software, compilers and languages, data analysis, data processing, mathematical physics, theoretical models, statistics…
- National and International research network and industrial collaborations:
  - Grenoble Univ., Athena, FORTH, UPC, CINECA, CNR, Julich, LENS, Manchester Univ, UniMi, CERN, NVidia, EuroTech, E4, IceoTope, IDIBAPS, MonetDB, ATOS, EVIDEN, BULL, UCLM, UPV, CINI, ISS,...

https://apegate.roma1.infn.it/

https://twitter.com/APELab_INFN

BACKUP

Post exascale challenges at a glance

Exascale computing is a huge milestone, but the HPC community is already looking beyond it. Here are some of the key challenges facing post-exascale HPC:

1. The End of Moore's Law: Traditional processor scaling is slowing down. Finding new ways to increase performance, such as specialized hardware, new architectures, and more efficient algorithms, will be crucial.

2. Power Consumption: Exascale systems already consume massive amounts of power. Post-exascale systems will need to be even more energy-efficient to be sustainable and affordable. This will require innovations in power delivery, cooling, and hardware design.

3. Data Movement Bottlenecks: Moving data efficiently within the system is already a major challenge at exascale. As systems grow larger and more complex, this problem will only intensify. New interconnect technologies, memory hierarchies, and data management strategies will be needed.

4. Programming Complexity: Developing software for increasingly complex and heterogeneous systems will be a major challenge. New programming models, tools, and languages will be needed to simplify development and improve productivity.

5. Resilience and Fault Tolerance: With more components, post-exascale systems will be even more susceptible to failures. Developing robust fault tolerance mechanisms and ensuring system resilience will be critical.

6. Application Scalability: Not all applications can scale effectively to exascale and beyond. Developing algorithms and software that can take full advantage of these massive systems will be essential.

7. Quantum Computing Integration: Integrating quantum computers with classical HPC systems could unlock new possibilities, but it also presents significant challenges in terms of hardware, software, and algorithms.

8. AI and Machine Learning: AI and machine learning are becoming increasingly important in HPC, but they also pose challenges in terms of data management, model training, and integration with traditional HPC workflows.

9. Workforce Development: A skilled workforce is needed to design, build, program, and manage post-exascale systems. Addressing the skills gap through education and training will be crucial.

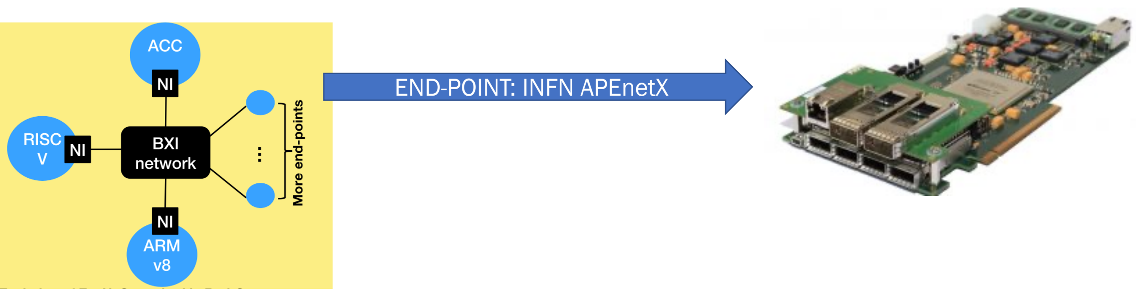
10. Ethical Considerations: As HPC systems become more powerful, it's important to consider the ethical implications of their use, such as potential biases in AI algorithms or the environmental impact of energy consumption.

EuroHPC JU R&D: progetto RED-SEA (Network) (A.Biagioni, P. Vi

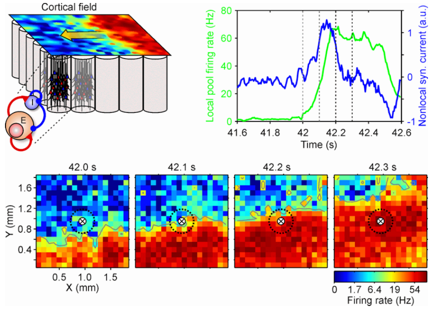
The four pillars of RED-SEA research

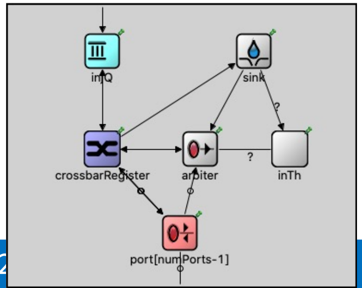
Architecture, co-design and performance	Optimizing the fit with the other EuroHPC projects and with the EPI processors	INFN
High-performance Ethernet	Development of a high-performance, low-latency, seamless bridge with Ethernet	Atos
Efficient Network Resource management	Including congestion management and Quality-of-Service targets while sharing the platform across application and users	Universitat Politècnica de València
Endpoint functions and reliability	End-to-end enhancements to network services - from programming models to reliability & security and to in-network compute	FORTH

Project start: 01/04/2021
Project duration: 36 months
Project budget: 8 M€ (INFN 700k€)



Integrazione della Network Interface (NI) con **RISC-V e ARMv8 cores (EPI) , piattaforma EU di HPC Network (Atos BXI)e con acceleratori FPGA e GPU**





- NEST (Spiking NN simulator) come benchmark e co-design application
  - Sviluppo di network IP per ottimizzazione Spiking NN simulator
- APEnet+ network simulators a larga scala
- Funzioni di network routing assistite da tecniche di ML

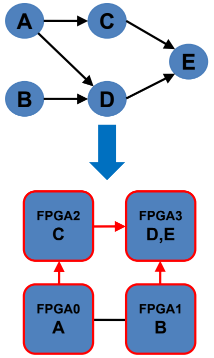
EuroHPC JU R&D: progetto TextaRossa (A. Lonardo, P. Vicini)

Obiettivi principali
- Energy Efficiency
- Sustained Performance delle applicazioni
- Integrazione di acceleratori riconfigurabili (FPGA)
- Sviluppo di IP
  - comunicazione, mixed precision AI, security, power monitoring,...
- Rilascio di nuove piattaforme (IDV)

11 partners from 5 countries:
ENEA, Fraunhofer, INRIA, ATOS, E4, BSC, PSNC, INFN, CNR, IN QUATTRO, CINI (Politecnico di Milano, Università di Torino, Università di Pisa),
LTP: Universitat Politecnica de Catalunya (UPC), Université de Bordeaux.

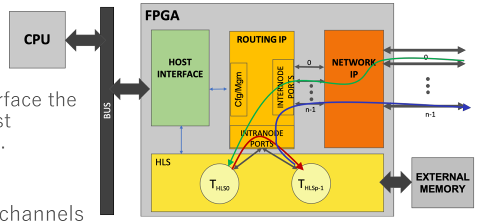
INFN Contribution to WP2/WP4: APEIRON

- Goal: offer hardware and software support for the execution on a system of **multiple interconnected FPGAs of applications developed according to a dataflow programming model**
- Map the directed graph of tasks on the distributed FPGA system and offer runtime support for the execution.
- Allow users with **no (or little) experience in hardware design tools** to develop their applications on such distributed FPGA-based platforms
  - Tasks are implemented **in C++ using High Level Synthesis tools (Vitis).**
  - Simple **Send/Receive** C++ communication API.



textarossa

INFN in WP2: IPs for low-latency FPGA commun.

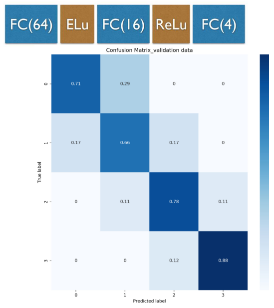


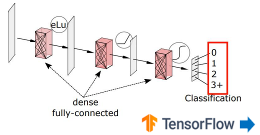
- **Host Interface IP:** Interface the FPGA logic with the host through the system bus.
  - PCI Express Gen3 → Gen4
- **Network IP:** Network channels and **Application-dependent I/O**
  - APElink 32 Gbps → 64/100 Gbps
  - UDP/IP over 10-25 GbE → 40/100 GbE
- **Routing IP**
  - Routing of intra-node and inter-node messages between processing tasks on FPGA.
- Implemented as incremental development on APEnet IPs over XILINX platforms.
- **Deliverable D2.5**
  - Intermediate database at M18
  - Deployed in the IDVs (WP5) at M30

textarossa 21

RAIDER Rings detection - Dense model on FPGA

Fully Connected
- **Input: 64 hits per event**
- **Architecture: 3 fully connected layers**
- **Output: 4 classes (0, 1, 2, 3+ rings per event)**
- **Qkeras, quantization aware training:**
  - ~75% average accuracy with low resource usage: LUT 14%, DSP 2%, BRAM 0% (VCU118)
- **Latency: 22 cycles @ 150MHz**
- **Initiation Interval (II): 8 cycles**



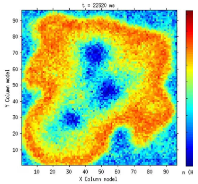


TensorFlow → hls4ml → VIVADO HLS → IP integration in LOTP+ infrastructure

https://fastmachinelearning.org/hls4ml/

Nest GPU (as NEST on GPU)

- The engine driving the neural simulations is the Nest GPU code which is C++ with CUDA extensions and is production-ready
- The Python script detailing the experimental protocol is ready – a 1000ms simulation of dynamics of one hemisphere of cortex of mouse brain with a realistic connectome inferred from data obtained with optical imaging methods on anesthetized mice – and will be run by the Nest GPU engine on the reference platform.
- As soon as the GPU-equipped is available, the simulation is ready to be benchmarked comparable with the same experiment on CPU-only engine (NEST).
- The specific KPI are:
  - Time-to-solution: Simulated-milliseconds-per-second
  - energy-to-solution: Synaptic UPdates per second (SUPs) per Watt



textarossa

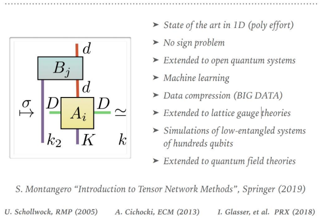
High Energy Physics high-level software tools

- For simulation, reconstruction (i.e. the transformation of detector signals to physics objects), data analysis
- Initial focus will be on the reconstruction software of the CMS experiment
  - Efforts are on-going to investigate parallelism and heterogeneous computing (CPU, GPU, possibly FPGA), based on TBB, CUDA, SYCL/OneAPI, Cupla/Alpaka, Vitis HLS, ...
  - Some solutions are already in production, but investigation continues
- We have identified two software components, for particle tracking and calorimeter clustering
- Two directions of work
  - Use of GPUs and FPGAs via SYCL
  - Remote offloading of computation to specialized nodes
- Activity just started, due to delays in recruiting

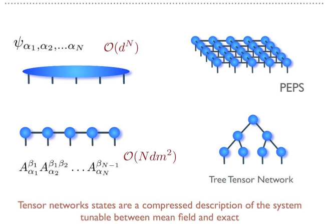
textarossa

Tensor Network Methods

TENSOR NETWORK ALGORITHMS



TENSOR NETWORKS STATES



Tensor network are state of the art methods for the simulation of many-body quantum systems, to understand complex quantum phenomena and to benchmark, verify and guide the developments of emerging quantum technologies (computers, simulations, sensors and communication).

Interpolation between mean field theory and exact description, faithful compression of the exponentially large many-body wave function.

textarossa

EU interconnect targeting EU HPC: NET4EXA

NET4EXA aims to develop a next-generation high-speed interconnect for HPC and AI systems, building on the success of the **BXI European HPC Interconnect** and the advancements made through research in the RED-SEA project and other previous European RIA initiatives.

Main outcomes & expected impact

- **A European-designed interconnect network solution** (hardware and software products):
  - The project will reduce reliance on non-European providers and promote technological sovereignty within Europe.
- **A Competitive Interconnect Network Solution with Key Differentiators**:
  - The developed solution will match or exceed the performance levels of competitors while offering unique features that set it apart in the market.
- **Mature technology for Exascale and Post-Exascale Clusters**:
  - The project focuses on developing a scalable and energy-efficient interconnect, including features for monitoring and controlling energy consumption, supporting both HPC and AI application use-cases, and suitable for Exascale and post-Exascale computing clusters, to ensure long-term viability and capability.
- **An interconnect network to facilitate datacenter internal and external communications**:
  - The solution will facilitate communications *within data centres* (both intra and inter-module) *as well as external communications* (with cloud infrastructures and inter-data centres).
- **A Skilled Workforce of Engineers and Researchers**:
  - The project will help build a highly qualified pool of engineers and researchers specialised in interconnect network technologies, capable of driving further innovations and providing consulting and support services.
- **Compatibility and Optimisation with European Processor and Accelerator Technologies**:
  - The interconnect network solution will be fully aligned with and optimised for European-developed processors and accelerators, enhancing integration and performance within European systems, and facilitating broader adoption.