

Monte Cimone

Paving the Road for the First Generation of RISC-V High-Performance Computers

Federico Ficarelli, Andrea Bartolini, Emanuele Parisi, Francesco Beneventi, Francesco Barchi, Daniele Gregori, Fabrizio Magugliani, Marco Cicala, Cosimo Gianfreda, Daniele Cesarini, Andrea Acquaviva, Luca Benini

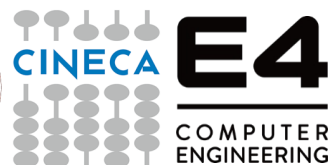


Question:

How mature is the RISC-V ecosystem? Is the **RISC-V ecosystem** mature enough to **build HPC production clusters**?

This work:

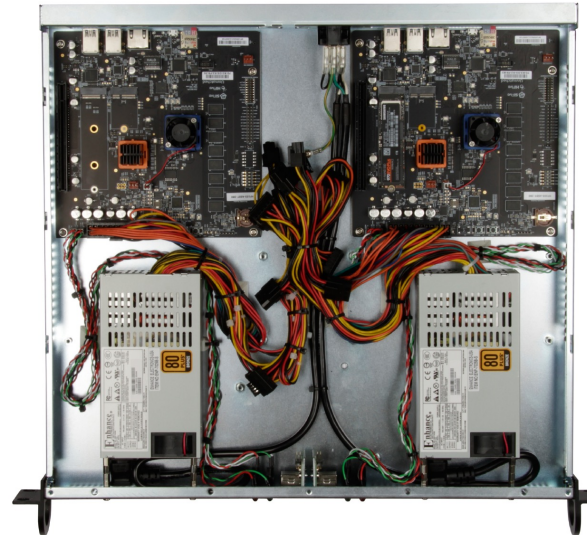
We designed and built **Monte Cimone**, the **first physical prototype** and test-bed of a **complete RISC-V (RV64) compute cluster** integrating **compute, interconnect, a complete software stack for HPC** and a **full-featured system monitoring infrastructure**.



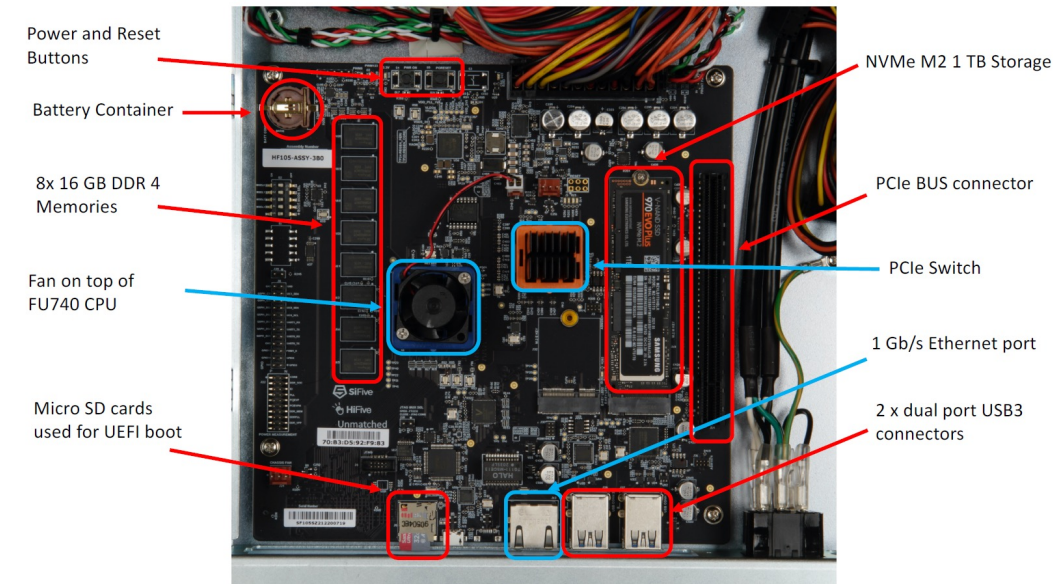
Monte Cimone: Hardware



E4 RV007 blade prototype



SiFive HiFive Unmatched board

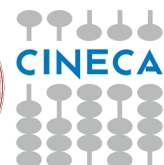
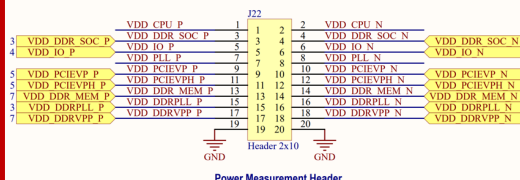


4x E4 RV007 1U Custom Server Blades:

- 2x SiFive U740 SoC with 4x U740 RV64GCB cores
- 16GB of **DDR4**
- 1TB node-local **NVME** storage
- **PCIe Gen 3** expansion card w/**InfiniBand HCAs** (2x Mellanox ConnectX-4 FDR)
- **Ethernet + IB parallel networks**

SiFive U740 SoC w. 7 separated power rails:

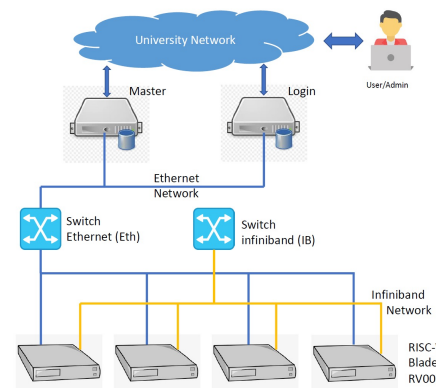
- Core complex, IOs, PLLs, DDR subsystem and PCIe.
- Board implements distinct shunt resistors



Monte Cimone: Software Stack

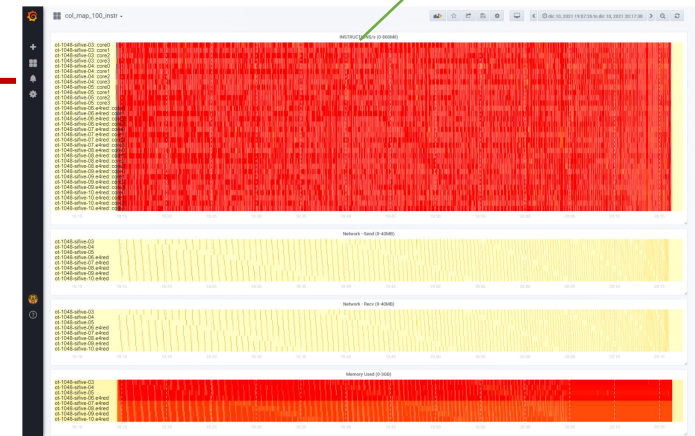
- SLURM job scheduler, NFS filesystem, Nagios
- User-space deployed via **Spack** package manager
- Upstream and custom **toolchains**
- **Scientific libraries**
- Industry-standard **HPC benchmarks and applications** (e.g.: quantumESPRESSO, OpenFOAM)
- The **ExaMon** datacenter automation and monitoring framework

Package	Version
gcc	10.3.0
openmpi	4.1.1
openblas	0.3.18
fftw	3.3.10
netlib-lapack	3.9.1
netlib-scalapack	2.1.0
hpl	2.3
stream	5.10
quantumESPRESSO	6.8



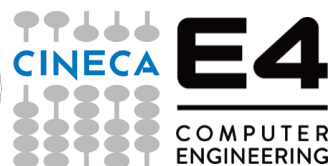
Traditional architecture: x86 login node + x86 master node running daemons

Live power draw, thermal profiles, loads



Same admin tools used in production on Tier0 machines @ Cineca: ExaMon + Nagios allows complete observability.

- Same deployment tools used in production on Tier0 machines @ Cineca (e.g.: Spack already supported our linux-sifive-u74mc target)
- Upstream rv64 Ubuntu 20.04 image



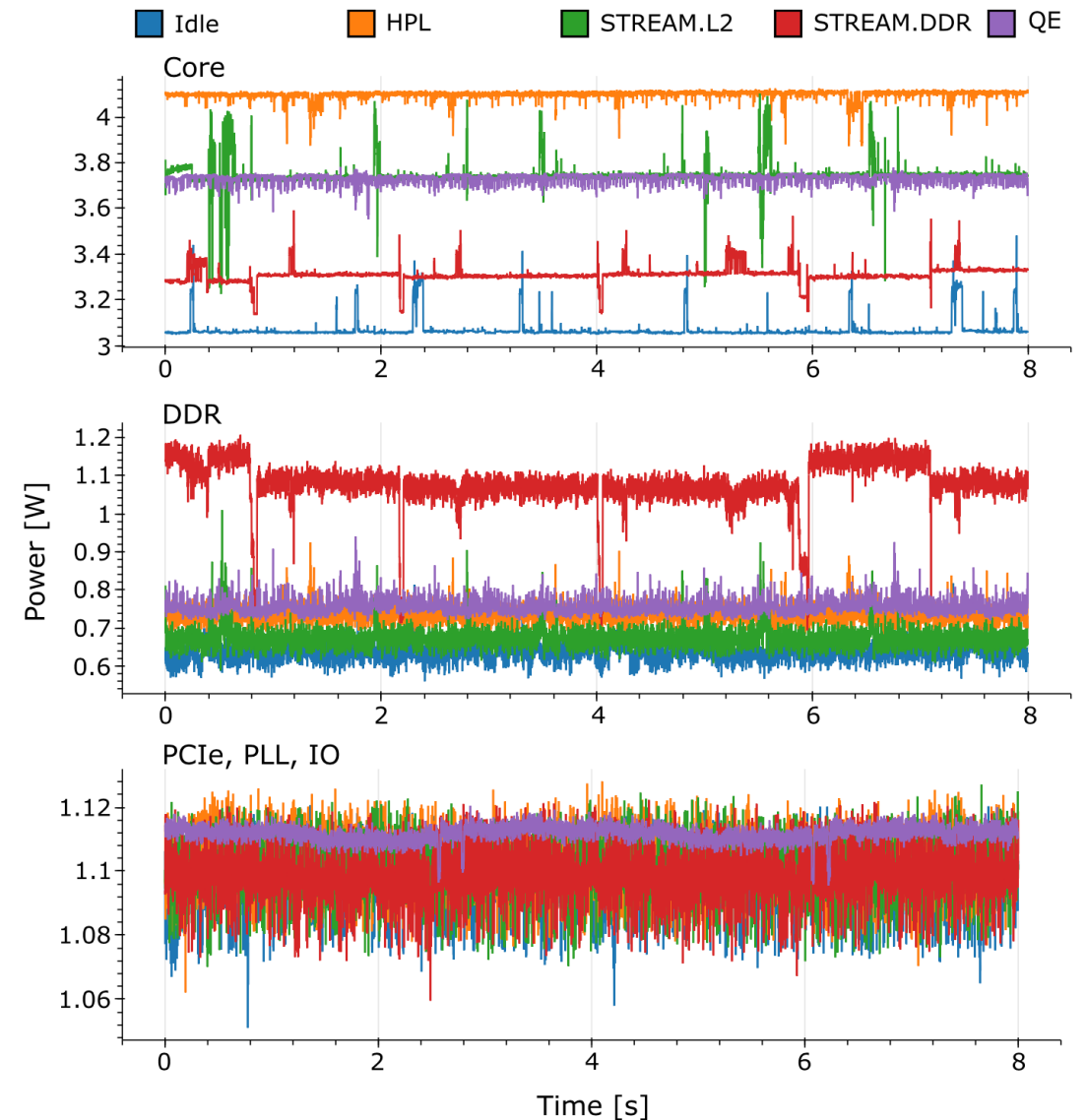
Power characterization

We extended the **ExaMon^[1] monitoring/data analysis framework**; extracted the **power profile** of the full cluster in a production-like setup during full-scale benchmarks:

Idle: 4.81W (64% core, 13% DDR, 23% PCI)

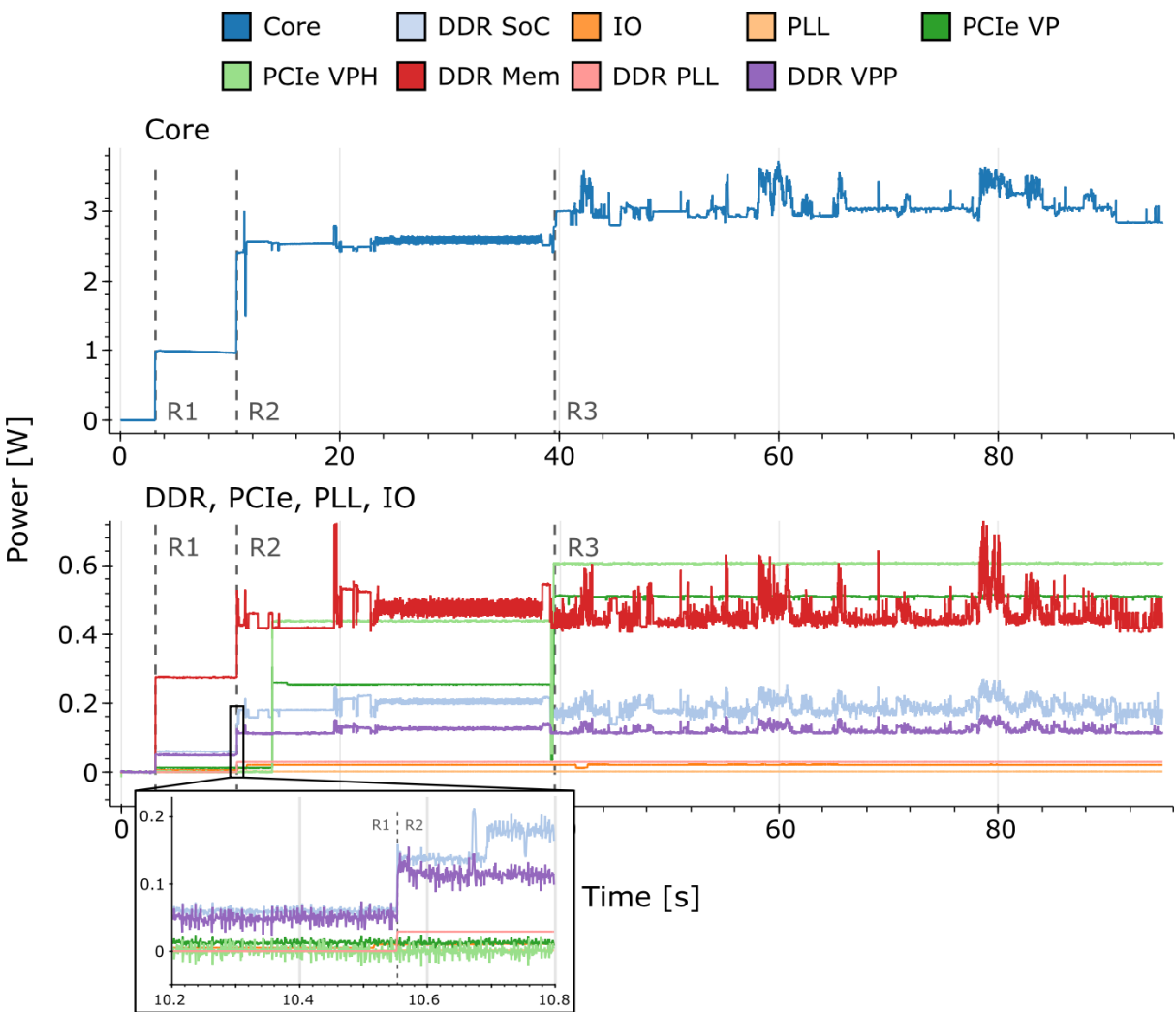
Load (HPL): 5.935W (69% core, 14% DDR, 18% PCI)

Line	Idle		HPL		STREAM.L2		STREAM.DDR		QE	
	[mW]	[%]	[mW]	[%]	[mW]	[%]	[mW]	[%]	[mW]	[%]
core	3075	64	4097	69	3714	68	3287	62	3825	67
ddr_soc	139	3	177	3	170	3	232	4	176	3
io	20	0	20	0	20	0	20	0	20	0
pll	1	0	1	0	1	0	1	0	1	0
pciev	521	11	527	9	524	10	522	10	530	9
pcievph	555	12	554	9	554	10	555	10	561	10
ddr_mem	404	8	440	7	401	7	592	11	434	8
ddr_pll	28	1	28	1	28	1	28	1	28	1
ddr_vpp	67	1	90	2	73	1	98	2	95	2
Total	4810	100	5935	100	5486	100	5336	100	5670	100



[1] ExaMon???

Power Characterization



Line	Idle		HPL		STREAM.L2		STREAM.DDR		QE		Boot	
	[mW]	[%]	[mW]	[%]	[mW]	[%]	[mW]	[%]	[mW]	[%]	R1	R2
core	3075	64	4097	69	3714	68	3287	62	3825	67	984	2561
ddr_soc	139	3	177	3	170	3	232	4	176	3	59	197
io	20	0	20	0	20	0	20	0	20	0	5	20
pll	1	0	1	0	1	0	1	0	1	0	0	2
pcievph	521	11	527	9	524	10	522	10	530	9	12	231
pcievph	555	12	554	9	554	10	555	10	561	10	1	395
ddr_mem	404	8	440	7	401	7	592	11	434	8	275	467
ddr_pll	28	1	28	1	28	1	28	1	28	1	0	29
ddr_vpp	67	1	90	2	73	1	98	2	95	2	49	122
Total	4810	100	5935	100	5486	100	5336	100	5670	100	1385	4024

Core complex @ boot process:

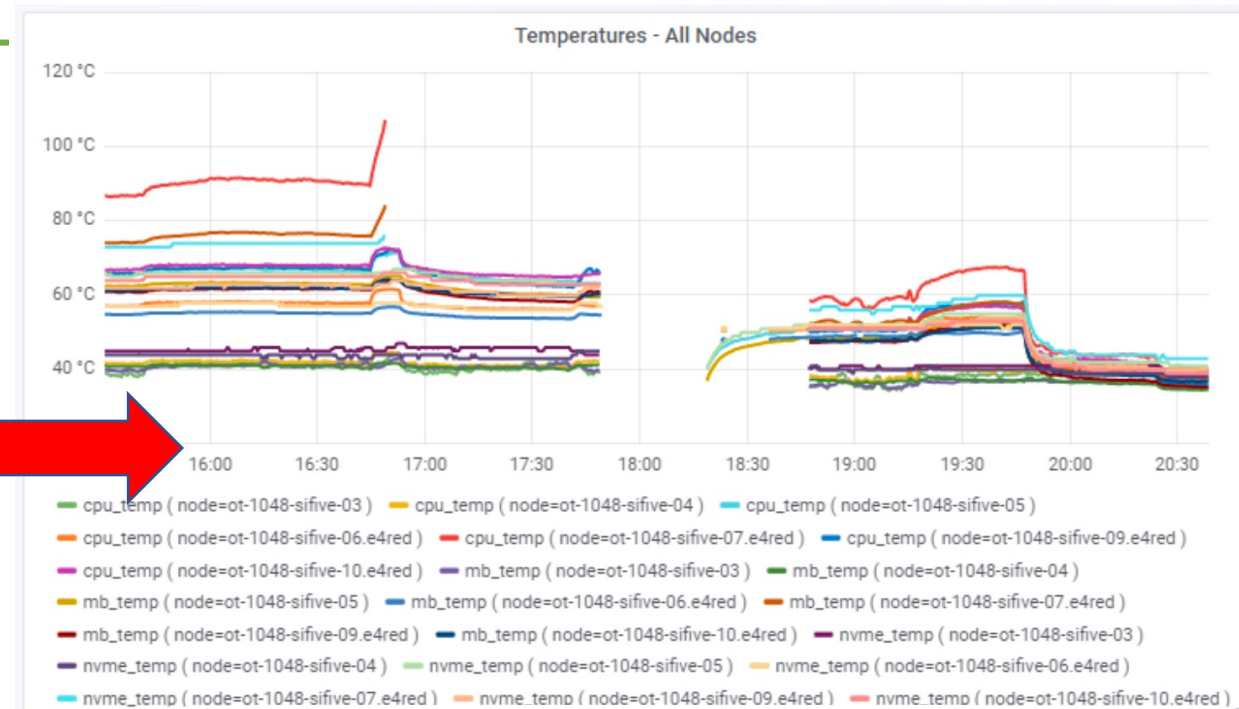
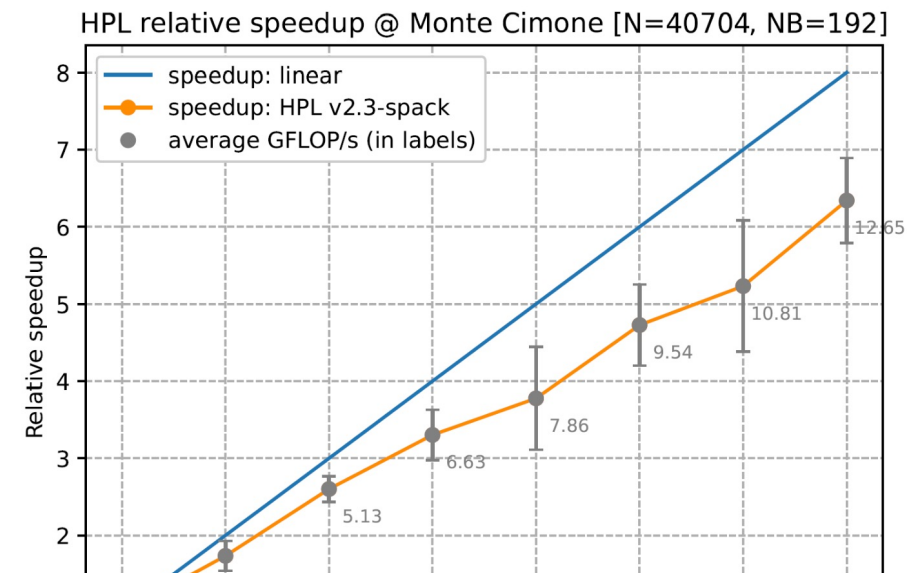
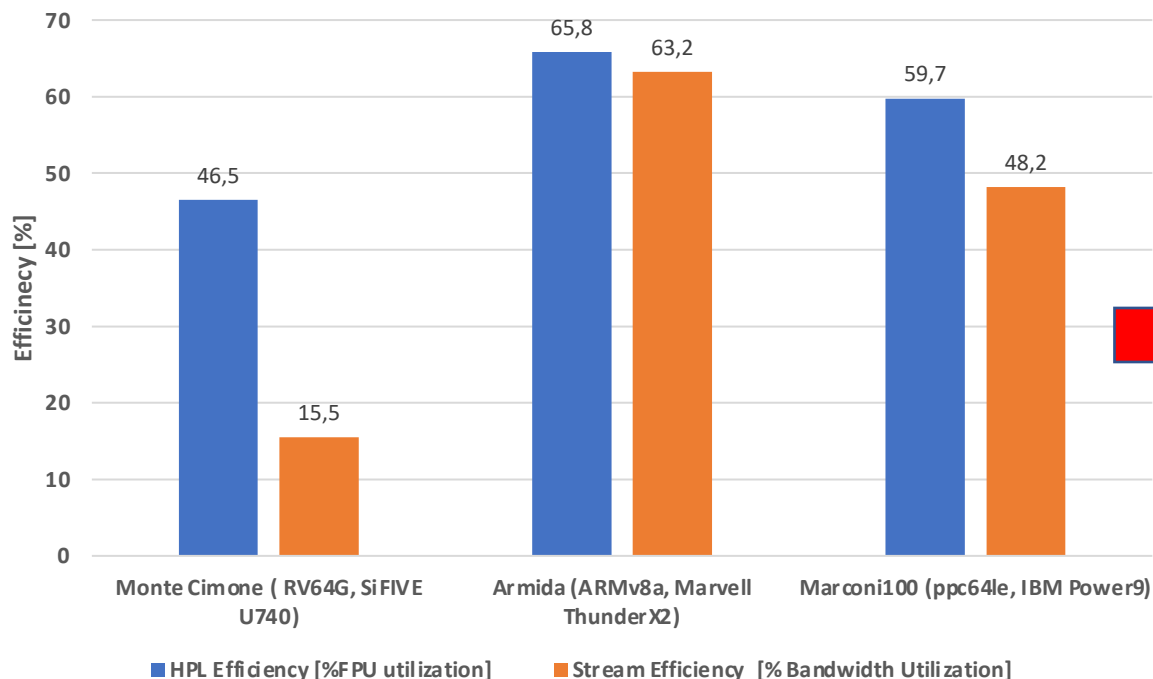
- 0.981W of leakage only power (32% of the idle power)
- 0.514W OS idle power (17% of the idle power)
- 1.577W of dynamic and clock tree power (51% of the idle power).



Efficiency benchmarks

- Same boundary conditions: vanilla benchmarks deployed via Spack
- Compared **efficiency** (% of attainable peak) VS 2x production systems: **Marconi100@Cineca** (ppc64le), **Armida@E4** (aarch64)
- Compute: **HPL**, comparable behaviours (46.5% vs 59.7% and 65.79%)
- Bandwidth: **STREAM**, much more difficult (15.5% vs 48.2% and 63.21%)

Monte Cimone vs Armida vs Marconi100



Thermal runaway during HPL

Challenges so far

InfiniBand support is still **missing**.

- ✓ Linux device driver
- ✓ OFED stack
- ✓ IB ping
- ✓ IP over IB
- ✗ RDMA

We are working on it.

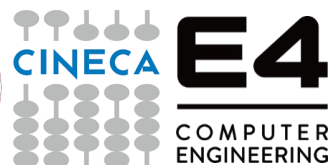
Access to **PMUs** via Linux kernel (e.g.: **perf**): done via SBI, major Linux distributions didn't support it out of the box, required kernel patches for OpenSBI.

Some **toolchains** not always up to speed when dealing with small, in-order cores:

- LLVM register allocator designed for large, OoO architectures
- LLVM still lacks proper support for some standard extensions (e.g.: Zfinx, Zdinx)
- GNU binutils lacking Zb*

We are working on it.

Lack of a **large code model** can be an issue for some kind of codes (e.g.: static allocators)



Production so far

As an **educational tool**: **2x courses at Università di Bologna**:

- Computer Architectures
- Laboratory of Big Data Architectures

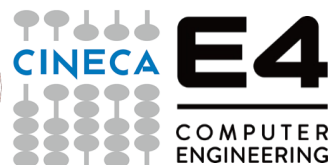
Introduced **~50 students** to **μarch profiling, HPC programming, distributed systems** right in a **RISC-V environment**.

Ported and ran production of **widespread HPC applications** (e.g.: **quantumESPRESSO, OpenFOAM**).

Several **research activities** currently ongoing on Cimone.

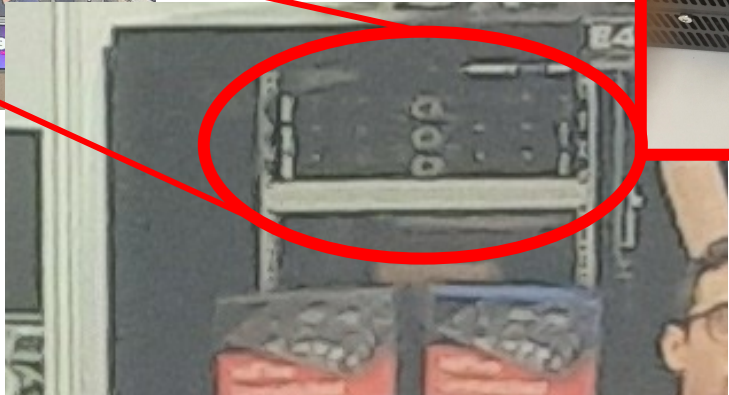
Is a **pilot system** in the  **EUPLEX** project.
European Pilot for Exascale

Access open to everyone interested (*no official contact point yet*, drop an email to: a.bartolini@unibo.it, f.ficarelli@cineca.it)





https://
2022



MPUTER



platform.
strong-

V-

HPC

- Home
- Technologies
- Sectors
- AI/ML/DL
- Exascale
- COVID-19
- Specials
- Resource Library
- Podcast
- Events
- Solution Channels
- About our Authors
- Link Policy

Marconi" department of the Università di Bologna, has contributed to Monte Cimone's system architecture definition, software stack development and integration within the Examon data-center automation environment.

CINECA, the leading Italian supercomputing center, has ported high-performance mathematical libraries (OpenBLAS, FFTW, Netlib-LAPACK, Netlib-scalAPACK) and scientific applications (HPL, Quantum Espresso) against the RISC-V ISA supporting E4 and Università di Bologna in the



3:25 PM · May 30, 2022

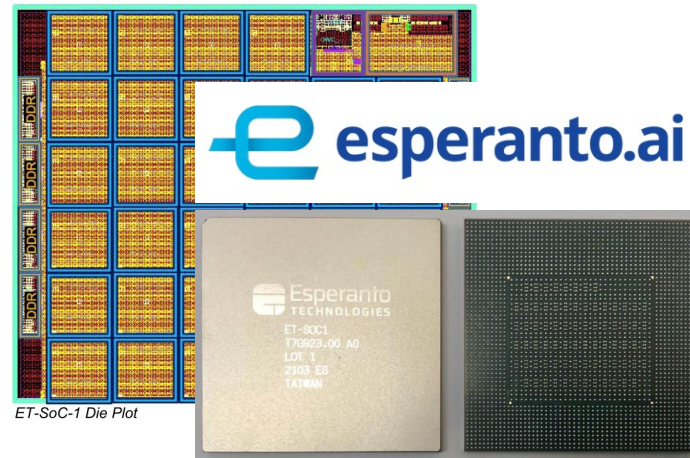


Next step: accelerated platform

Goal: Explore **RISC-V** accelerated HPC platforms in production.
Currently evaluating **several solutions**. Among those:



**PULP Platform energy
efficient accelerators^[1]**
[STX, Occamy, ...]



Esperanto Technologies ET-SoC-1^[2]



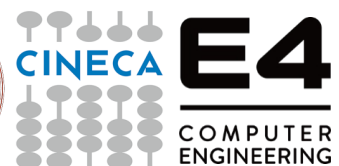
**Kalray MPPA Manycore
Accelerator^[3]**



[1] <https://pulp-platform.org>

[2] Accelerating ML Recommendation With Over 1,000 RISC-V/Tensor Processors on Esperanto's ET-SoC-1 Chip, David R. Ditzel, the Esperanto team, DOI: 10.1109/MM.2022.3140674

[3] Co-Design of the Kalray Manycore Accelerator for Edge Computing, Benoît Dupont de Dinechin, HiPEAC CSW Autumn 2021



Trivia: Why Monte Cimone?



Considering the spatial resolution of the human eye, the top of Mount Cimone is the geographical point from which the most Italian land surface can be seen.

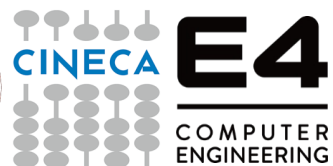
Tallest peak in the northern Apennines (2165m)

On clear days the summit is visible from the major cities in the area (Bologna, Firenze, Mantova, Modena, Reggio Emilia, Lucca, ...)



Conclusions: With **Monte Cimone**, the **first physical prototype** and test-bed of a **complete RISC-V (RV64) compute cluster**, *we demonstrated that it is possible to run real-life HPC applications on a RISC-V system today.*

Mission: Making high-performance **RISC-V processors and accelerators** ready for future **RISC-V-based HPC systems**.



Thanks.

The european-project-initiative has received funding from the European High Performance Computing Joint Undertaking (JU) under Framework Partnership Agreement No 800928 and Specific Grant Agreement No 101036168 (EPI SGA2). The JU receives support from the European Union's Horizon 2020 research and innovation programme and from Croatia, France, Germany, Greece, Italy, Netherlands, Portugal, Spain, Sweden, and Switzerland.

The EUPEX project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No.101034126. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Spain, Italy, Switzerland, Germany, France, Greece, Sweden, Croatia and Turkey.

This REGALE-project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 956560. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Greece, Germany, France, Spain, Austria, Italy.

The **Spoke Future HPC** of the Italian **National Center of HPC, Big Data e Quantum Computing** is funded by the **National Recovery and Resilience Plan (NRRP)**.

