

# Monte Cimone: Paving the Road for the First Generation of RISC-V High-Performance Computers

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

**Abstract**—The new open and royalty-free RISC-V ISA is attracting interest across the whole computing continuum, from microcontrollers to supercomputers. High-performance RISC-V processors and accelerators have been announced, but RISC-V-based HPC systems will need a holistic co-design effort, spanning memory, storage hierarchy interconnects and full software stack. In this paper, we describe Monte Cimone, a fully-operational multi-blade computer prototype and hardware-software test-bed based on U740, a double-precision capable multi-core, 64-bit RISC-V SoC. Monte Cimone does not aim to achieve strong floating-point performance, but it was built with the purpose of “priming the pipe” and exploring the challenges of integrating a multi-node RISC-V cluster capable of providing an HPC production stack including interconnect, storage and power monitoring infrastructure on RISC-V hardware. We present the results of our hardware/software integration effort, which demonstrate a remarkable level of software and hardware readiness and maturity - showing that the first generation of RISC-V HPC machines may not be so far in the future.

## I. INTRODUCTION

The strategic role of High Performance Computing (HPC) systems is widely acknowledged in many fields, from weather forecasting to drug design. With the pervasive digitalization of our society, high performance computers fuel the most disruptive mega-trends, from the deployment of artificial intelligence (AI) at scale (e.g. for training large machine learning models) to industrial internet-of-things (IoT) applications (e.g. for creating and maintaining digital twins). Thus, HPC systems are today strategic assets not only for academia and industry, but also as for public institutions and governments [24].

The key challenge in designing HPC systems today and in the foreseeable future is increasing compute efficiency, to meet the rapidly growing performance demand (10x every four years) within a constant or modestly increasing power

budget, while facing the slow-down of Moore’s Law. To exacerbate the efficiency challenge, while integrated circuits technology is still delivering device density increases (albeit as a slower pace), power consumption does not scale down at the same rate. Hence power density grows and it is increasingly difficult to meet thermal design power specifications without compromising performance. Disruptive technologies, such as quantum or optical computing may bring long-term relief in some specific application areas, but there is no silver bullet in sight.

To tackle the efficiency issue, academia and industry are aggressively pursuing architectural innovation and co-design strategies to develop HPC systems that mitigate the efficiency limitations of programmable architectures through various forms of specialization and domain-specific adaptation. Instruction Set Architectures (ISAs) have to evolve rapidly to sustain architectural evolution and domain adaptation, and the advent of the RISC-V open, royalty-free and extensible ISA has been a major step toward accelerating innovation in this area. An additional advantage of RISC-V with respect to the dominant proprietary ISAs (x86 and ARM) is that it is maintained by a global non-for-profit foundation with members across the world, ensuring a high degree of neutrality with respect to geopolitical tensions and their technology downfalls.

Currently, high-performance 64bit (RV64) RISC-V processors and accelerator chips are being designed, promising prototypes are demonstrated in numerous publications [21] and products are announced at a fast cadence [6], [17]. It is thus reasonable to expect that high-performance chips based on RISC-V will be available as production silicon within the next couple of years. However, building a HPC system requires significantly more than just high-performance chips. Many think that the RISC-V software stack and system platform are extremely immature, and will need several additional years of development effort before full applications could be run, benchmarked and optimized on a RISC-V-based HPC system. Our goal is dispel this overly conservative notion.

The main contribution of this work is to present Monte Cimone, the first physical prototype and test-bed of a complete RISC-V (RV64) compute cluster, integrating not only all the key hardware elements besides processors, namely main memory, non-volatile storage and interconnect, but also a complete

A. Bartolini, E. Parisi, F. Beneventi, F. Barchi, A. Acquaviva, and L. Benini are with the Department of Electrical, Electronic and Information Engineering “Guglielmo Marconi”, University of Bologna, 40136 Bologna, Italy (e-mail: a.bartolini@unibo.it; emanuele.parisi@unibo.it; francesco.beneventi@unibo.it; francesco.barchi@unibo.it; andrea.acquaviva@unibo.it; luca.benini@unibo.it).

F. Ficarelli and D. Cesarini are with the Department of SuperComputing Applications and Innovation, CINECA, 40033 Casalecchio di Reno (BO), Italy (e-mail: d.cesarini@cineca.it; f.ficarelli@cineca.it).

D. Gregori, F. Magugliani, D. Cicala, and C. Gianfreda are with E4 Computer Engineering, 42019 Scandiano (RE), Italy (e-mail: daniele.gregori@e4company.com; fabrizio.magugliani@e4company.com; marco.cicala@e4company.com; cosimo.gianfreda@e4company.com).

software environment for HPC, as well as a full-featured system monitoring infrastructure. Further, we demonstrate that it is possible to run real-life HPC applications on Monte Cimone today. Even though achieving strong double precision performance will be possible only with upcoming high-performance chips, we achieved the following milestones:

- We designed and set up the first RISC-V-based cluster containing eight computing nodes enclosed in four computing blades. Each computing node is based on the U740 SoC from SiFive and integrates four U74 RV64GCB application cores, running up to 1.2 GHz and 16GB of DDR4, 1 TB node-local NVME storage, and PCIe expansion cards. The cluster is connected to a login node and master node running the job scheduler, network file system and system management software.
- We ported and assessed the maturity of a HPC software stack composed of (i) SLURM job scheduler, NAS filesystem, LDAP server, Spack package manager (ii) compilers toolchains, scientific and communication libraries, (iii) a set of HPC benchmarks and applications, (iv) the ExaMon datacenter automation and monitoring framework.
- We measured the efficiency of the HPL benchmark and STREAM benchmark with the toolchain and libraries installed by the SPACK. We compared the attained results against the one obtained for other RISC ISA architectures used in the 1st and 2nd ranked Top500 supercomputers (namely, Summit and Fugaku). We build the HPL benchmark and Stream benchmark following the same approach for the Monte Cimone cluster on two SoA computing nodes, namely the Marconi100 [4](ppc64le, IBM Power9) computing node and the Armida [1] computing node (ARMv8a, Marvell ThunderX2) and compared the attained FPU utilization as a metric of efficiency against the one obtained by Monte Cimone while keeping the same benchmarking boundary conditions (e.g.: vanilla, unoptimized libraries and software stack deployed via a popular package manager). Results show that upstream HPL achieved 46.5% utilization on Monte Cimone, the Marconi100 [4] and Armida [1] compute nodes achieved 59.7% and 65.79% of their peak respectively. The Monte Cimone node achieves slightly lower FPU utilization but in the range with the state of the art. When running an unoptimized Stream benchmark, Monte Cimone obtained just the 15.5% of the peak bandwidth, while Marconi100 and Armida obtained an efficiency of 48.2% and 63.21% respectively, pointing to significant margins for improvement in application and software stack tuning to the hardware target.
- We extended the ExaMon monitoring framework [12] to monitor the Monte Cimone cluster. We characterised the power consumption of various applications executed on Monte Cimone. We reported a power consumption of 4.81W in idle, composed of 64% of core power, 13% related to DDR and 23% of related to PCI subsystem.

During CPU intensive benchmark run the SiFive Freedom U740 SoC we reported a power consumption of 5.935W, composed of 69% of core power, 14% related to DDR and 18% related to PCI subsystem. By profiling the power consumption of the core complex during the boot process we measured a 0.981W of leakage only power (32% of the Idle power) and measured 0.514W of power consumed by the operating system during idle (17% of the Idle power) and a remaining 1.577W of dynamic and clock tree power, accounting for the 51% of the core idle power. In addition to providing a detailed analysis of power consumption, ExaMon enabled us to detect and mitigate a thermal design issue in the cluster.

## II. RELATED WORKS

The most recent successful effort to introduce a new ISA to HPC has involved the ARM ISA. Bringing the Arm ISA and software ecosystem to HPC maturity has required almost a decade and several funding rounds: The Mont-Blanc EU project series started in 2011, leading to the first ARM-based HPC cluster deployed in 2015 [28], based on SoCs developed for the embedded computing market. Notably, since June 2020, Fugaku [29], the fastest supercomputer in TOP500, is based on ARM scalable vector extension (SVE) ISA, and achieves more than 400 PFLOPs. Further, high-performance ARM-based SIMD processors are being adopted in servers and datacenters worldwide. We observe that it took approximately a decade for ARM to become a strong player in these highly competitive markets, even though X86 is still by far the dominant architecture in HPC and cloud.

The RISC-V ISA has been conceived just a decade ago, thus clearly its market penetration is much smaller than the incumbent ARM and X86 ISAs. Today, only a few 64-bit RISC-V (RV64G ISA) SoCs are available commercially and none is in volume production for HPC or performance servers. Nevertheless, several high-performance RISC-V processors have been announced for high-performance general-purpose and accelerated computing markets [2], [8], [9]. In addition, a few research prototypes have been presented in the recent literature that demonstrate on silicon the technical feasibility and competitiveness of high-performance RISC-V computing engines [15], [30], [32]. Furthermore, the European Processor Initiative (EPI) launched in 2019 is funding a major research thrust to develop RISC-V based accelerators for HPC [3].

Among the RV64G chips available in low volumes on the market, for our cluster we chose the SiFive Freedom U740 SoC, featuring a 64-bit dual-issue, superscalar RISC-V U7 core complex configured with four U74 cores and one S7 core, an integrated high speed DDR4 memory controller, a root complex PCI Express Gen 3 x8 and standard peripherals. The availability of a main memory interface with reasonable performance and a PCIe root complex for connecting fast storage, IOs and accelerators, makes this SoC a good basis for exploring the deployment of RISC-V processors in a scalable cluster and working on the software stack. Still, it is apparent that the performance and number of cores in the SoC is not

sufficient to achieve performance comparable to mature ARM and X86 cores.

The maturity of the software ecosystem around RISC-V has been growing at a very fast rate. A reasonably complete snapshot of major software packages available for RISC-V is maintained by the RISC-V foundation [6]. While the list is not complete, due to the very fast growth of the RISC-V community of developers, it is clear that porting efforts so far have mainly focused on embedded and AI applications. A HPC special interest group (SIG) for RISC-V has been founded in 2019 [11]. However, to the best of our knowledge, the demonstration of a complete software stack and HPC applications running on real hardware on RISC-V nodes in a multi-blade cluster is still missing. The present work aims at filling this gap.

In addition to libraries and tools for HPC application deployment, a production-ready HPC system must support fine-grain utilization, performance and power monitoring of the computing resources to enable efficient computing, power, thermal management and anomaly detection for reliability. Recently, several works have been proposed to extend the power monitoring attainable from the voltage regulator modules leveraging shunt resistors, current probes, and out-of-band telemetry [23]. In addition, Operational Data Analytics [26] (ODA) has been introduced focusing on monitoring and managing large scale HPC installations. In this area, vertical solutions encompassing all layers (from data gathering and storage to processing and analysis) have been proposed. Bautista et al. [13] describe an infrastructure for extreme-scale operational data collection, known as OMNI. In [12] Bartolini et al. describe ExaMon, an ODA infrastructure leveraging: i) Distributed sensing plugins (including node-level metrics, processing elements performance metrics, dedicated fine-grain power monitoring meters, facility data); ii) Scalable storage backends; iii) Visualization and analytics targetting anomaly detection and intrusion detection systems. Current ODA tools are available only for the dominant ARM and X86 environments. In this work we advance the state of the art demonstrating a fully-operational port of the Examon ODA infrastructure to the Monte Cimone RISC-V cluster.

### III. MONTE CIMONE HARDWARE

Monte Cimone is based on the SiFive Freedom U740 RISC-V SoC HiFive Unmatched board integrated in an HPC node form factor. The E4 RV007 blade prototype system, adopted as Monte Cimone building Block, is a dual-board platform server, with a form factor of 4.44 cm (1 RackUnit) high, 42.5 cm width, 40 cm deep. Two 250 W power supplies, one for each board (compute node), are installed inside the case. This allows to turn on every single compute node individually, see Figure 1, and makes the system ready with abundant power headroom for future expansions with hardware accelerators and PCIe Network Card connector.

The board follows the Industry Standard Mini-ITX with a size of 170 mm per 170 mm. It features one SiFive Freedom U740 SoC, 16 GB of 64-bit DDR4 memory operating up to

1866s MT/s and high-speed interconnects with PCIe Gen 3 x16 (but it's limited to x8 lanes), one Gigabit Ethernet, and four USB 3.2 Gen 1, see Figure 1.

In RV007 node the M.2 M-key expansion slot is occupied by a 1 TB NVME 2280 SSD Module storage device used by the Operating System. The Micro SD card is present and used for the UEFI Boot. Two buttons for reset and power up operations are available on top of the board and in front of the case.

The FU740-C000 is a Linux-capable SoC powered by SiFive's U74-MC, the first (to the best of our knowledge) commercially available superscalar heterogeneous multi-core RISC-V Core Complex. The FU740-C000 is compatible with all applicable RISC-V standards.

The U74-MC core complex is composed of four 64-bit U74 RISC-V (Application) cores. Each U74 core has a dual issue in-order execution pipeline, with a peak sustainable execution rate of two instructions per clock cycle. The U74 core supports Machine, Supervisor, and User privilege modes as well as standard Multiply, Single-Precision Floating Point, Double-Precision Floating Point, Atomic, and Compressed RISC-V extensions.

The SiFive Freedom board features a Microsemi VSC8541 chip to interconnect the SiFive Freedom U740 SoC with a single port gigabit Ethernet copper interface. Moreover, we equipped two of the compute nodes with an Infiniband Host Channel Adapter (HCA) widely used in large-scale HPC systems. We target an Infiniband FDR HCA (56Gbit/s) to leverage RDMA communications among different nodes to improve the network throughput and the communication latency. We used a Mellanox ConnectX-4 FDR HCA interconnect through the PCI-e interface available on the compute node. This HCA support x8 free PCIe Gen 3 lanes, which are currently supported by the vendor. The first experimental results show that the kernel is able to recognise the device driver and mount the kernel module to manage the Mellanox OFED stack. We are not able to use all the RDMA capabilities of the HCA due yet-to-be-pinpointed incompatibilities of the software stack and the kernel driver. Nevertheless we successfully run an IB ping test between two boards and between a board and an HPC server showing that full Infiniband support could be feasible. This is currently a feature under development.

In addition, the SiFive Freedom U740 SoC features 7 separated power rails including the core complex, IOs, PLLs, DDR subsystem and PCIe one. The HiFive Unmatched board implements separated shunt resistors in series with each of the SiFive U740 power rails as well as for the on-boards memory banks [10].

### IV. MONTE CIMONE SOFTWARE STACK

Since our goal was to build a software environment as close as possible to a production HPC cluster, we leveraged the Spack [19] package manager to deploy the full software stack and make it available to all system users via environment modules [18]. Actual Spack architecture and micro-architecture support, in the form of platform-specific toolchain flags, is provided by the archspec [16] module. Explicit

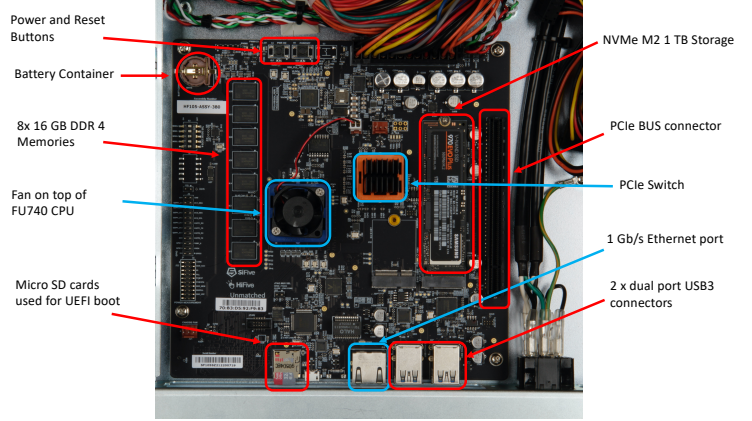


Fig. 1: The E4 RV007 Server Blade is based on a dual SiFive Freedom U740 SoC, the form factor is 4.44 cm (1 RackUnit) high, 42.5 cm width, 40 cm deep. The size of each RISC-V development board is 170 mm per 170 mm.

TABLE I: User-facing software stack deployed on Monte Cimone

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

support for the `linux-sifive-u74mc` target triple was already present (archspec version 0.1.3) and tested to be working without modifications. The user-facing software stack installed successfully via Spack (version 0.17.0) and presented to users is listed in Table I (transitive dependencies omitted for brevity). All of the nodes are running upstream Ubuntu 21.04 deployed from `riscv64` server images without modifications and mount a remote NFS.

#### A. System Software

We ported on Monte Cimone all the essential services needed for running HPC workloads in a production environment, namely NFS, LDAP and the SLURM job scheduler. Porting all the necessary software packages to RISC-V was relatively straightforward, and we can hence claim that there is no obstacle in exposing Monte Cimone as a computing resource in a HPC facility. However, full integration requires integrating Monte Cimone within a holistic monitoring framework. For that purpose we use the ExaMon framework [12].

In the next sub-section we describe its configuration and the measured metrics.

#### B. ExaMon Configuration

The typical configuration of ExaMon consists in installing plugins dedicated to data sampling, a broker for transport layer management and a database for storage. For Monte Cimone cluster both broker and database are installed in their basic configuration on a master node, while plugins have been specifically developed/adapted for this project and installed on the compute nodes. As a first step, we created a dedicated version of the `pmu_pub` [14] plugin to acquire the performance counters available in the Linux OS through the `perf_events` interface. In the current version of the Kernel the RISC-V architecture provides, through this interface, the fixed INSTRET and CYCLE counters. By default, the remaining programmable counters available on the hardware performance monitoring (HPM) unit of the U740 SoC [10] are disabled at boot time. We have therefore developed a patch for the bootloader (U-Boot) useful to enable and program all counters.

The counters are sampled for each core of the SoC in user-mode by the `pmu_pub` plugin at regular intervals (2Hz) and the values are sent to the MQTT transport layer. The data model adopted for this application follows the ExaMon specification and consists in the definition of the MQTT topic and payload as described in the Table II.

A second plugin has been installed and configured to collect operating system statistics, `stats_pub`. This plugin mainly accesses the `sysfs` and `procfs` filesystems to get useful metrics about system resources such as load, CPU usage, memory usage, network bandwidth and other metrics as described in Table III. In particular, the HiFive Unmatched board is equipped with three thermal sensors dedicated respectively to



TABLE II: ExaMon: Topic and payload formats

Plugin	Topic	Payload
pmu_pub	org/XXXXXX/cluster/XXXXXX/ node/<hostname>/plugin/ pmu_pub/chnl/data/core/ <id>/<metric_name>	<value>;<timestamp>
	org/XXXXXX/cluster/XXXXXX/ node/<hostname>/plugin/ dstat_pub/chnl/data/ <metric_name>	
stats_pub	org/XXXXXX/cluster/XXXXXX/ node/<hostname>/plugin/ dstat_pub/chnl/data/ <metric_name>	<value>;<timestamp>

TABLE III: Metrics collected by the stats\_pub plugin

Type	Metric
Load	load_avg.1m,load_avg.5m,load_avg.15m
I/O	io_total.read,io_total.writ
Processes	procs.run,procs.blk,procs.new
Memory	memory_usage.used,memory_usage.free,memory_usage.buff,
	memory_usage.cache
Disk	paging.in,paging.out
	dsk_total.read,dsk_total.writ
System	system.int,system.csw
CPU	total_cpu_usage usr,total_cpu_usage.sys,total_cpu_usage.idl,
	total_cpu_usage.wai,total_cpu_usage.stl
Network	net_total.recv,net_total.send
Temperatures	temperature.mb_temp,temperature.cpu_temp,
	temperature.nvme_temp

the SoC, the Motherboard and the NVME SSD. These sensors are available through the *hwmon sysfs* interface as shown in Table IV. This plugin samples data with a frequency of 0.2Hz.

Finally, the data collected for each board are available to be viewed and processed through the various interfaces provided by ExaMon. Through an instance of Grafana [12] connected to the database it is possible to visualize the trend of the metrics in real time, during the execution of the benchmark. The data can also be analyzed in batch mode using scripts and accessing the database through the dedicated RESTful API over HTTP.

## V. EXPERIMENTAL RESULTS

In this section, we report the characterisation of the Monte Cimone Cluster and of its software stack with the objective of assessing its maturity. In subsection V-A we focus on the software stack by compiling and running three different applications without manual optimisations. This lets us assess the available toolchains and libraries' capability to extract the application's performance given the new in HPC RISC-V ISA. We will then focus in subsection V-B one the power characterisation of one compute node. Finally, in subsection V-C we will describe cluster-level performance based on live dashboards extracted by ExaMon.

### A. Application performance

Considering the peak theoretical value of 1.0 GFLOP/s/core, inferred from the micro-architecture specification [10], leading to a 4.0 GFLOP/s peak value for a single chip, the upstream

TABLE IV: Sysfs entries for the temperature sensors

Sensor	Sysfs Files
nvme_temp	/sys/class/hwmon/hwmon0/temp1_input
mb_temp	/sys/class/hwmon/hwmon1/temp1_input
cpu_temp	/sys/class/hwmon/hwmon1/temp2_input

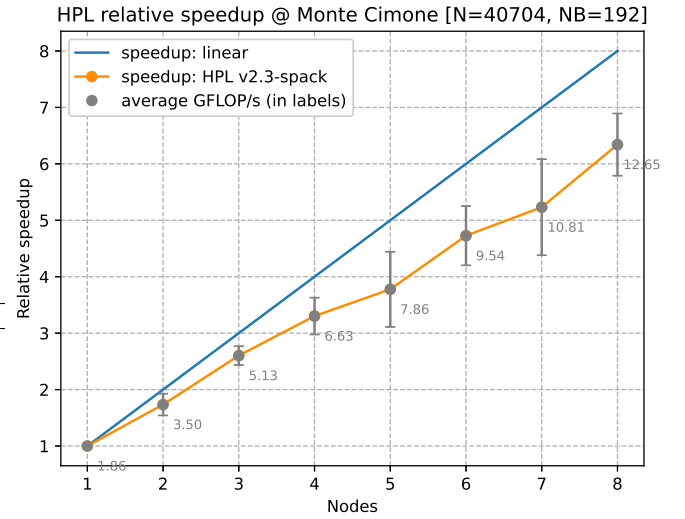


Fig. 2: HPL strong scaling tests on Monte Cimone. Average attained throughput values are shown in labels. Standard deviations are calculated on 10 repetitions.

HPL [27] benchmark (built on top of the software stack shown in Section IV) reached a sustained value of  $1.86 \pm 0.04$  GFLOP/s on a single node (on a  $N=40704$  and  $NB=192$  HPL configuration and a total runtime of  $24105 \pm 587$  s); this amounts to 46.5% of the theoretical peak, a result we deem to be promising considering the upstream, unmodified software stack used in this phase. The same experiment, run on both the Marconi100 [4] system at Cineca and the Armida [1] system at E4 using the same upstream software stack (and no vendor libraries) with the same MPI topology of 1 MPI task per physical core attained 59.7% and 65.79% of a single node's CPU-only theoretical peak respectively, a result that is comparable to what we observed on Monte Cimone. The same HPL configuration has been used to carry out a Monte Cimonefull-machine benchmark experiment leveraging the 1 Gb/s network currently available, reaching a sustained value of  $12.65 \pm 0.52$  GFLOP/s using all of the eight nodes (with a total runtime of  $3548 \pm 136$  s); this amounts to 39.5% of the entire machine's theoretical peak and to 85% of the extrapolated attainable peak in case of perfect linear scaling from the single-node case. Relative speedup obtained during the HPL strong scaling experiment are shown in Figure 2. Again, we consider these results to be promising and deserving both further optimization on the software side and tuning (or technology upgrade) on the interconnect side.

The STREAM [25] benchmark has been used to measure the attainable memory bandwidth on a single node. Out of the peak 7760 MB/s [10], a 4-thread experiment measured the values shown in Table V. Being the node a UMA system, no topology configuration had to be taken into account. We consider the results attained via upstream, unmodified STREAM unsatisfactory: the results on Monte Cimoneshow an attained bandwidth of no more than 15.5% of the available

TABLE V: STREAM, 4 threads

Test	STREAM.DDR	STREAM.L2
	1945.5 MiB [MB/s]	1.1 MiB [MB/s]
copy	1206 $\pm$ 3.26	7079 $\pm$ 2.11
scale	1025 $\pm$ 4.94	3558 $\pm$ 3.72
add	1124 $\pm$ 4.93	4380 $\pm$ 3.72
triad	1122 $\pm$ 5.63	4365 $\pm$ 3.56

peak bandwidth. The same experiment involving an upstream, unoptimized STREAM benchmark ran on both Marconi100 [4] and Armida [1] (using the same topology with 1 OpenMP thread per physical core) attained 48.2% and 63.21% of the peak bandwidth respectively, suggesting that a result higher than the lower quartile should be easily attained with little to no effort. This observation is worth of further experimentation, in particular:

(i) the L2 prefetcher provided by the micro-architecture [10], being able of tracking up to eight streams per core, should be perfectly capable of reducing the gap between the two experiments shown in Table V (DDR-bound and L2-bound) given the large degree of spatial and temporal locality shown by the STREAM memory access patterns. Further analysis is needed to understand how the prefetcher is currently operating and the modifications needed to leverage it properly; (ii) the overall data size used by STREAM is currently limited by the RISC-V code model. The *medany* code model used by RV64 requires that every linked symbol resides within a  $\pm 2\text{GiB}$  range from the *pc* register [10], [22]. Since the upstream, unmodified STREAM benchmark uses statically-sized data arrays in a single translation unit preventing the linker to perform *relaxed* relocations, their overall size cannot exceed 2 GiB. Further experiments on available workarounds for the absence of a *large code model* [31] and modifications to the STREAM source itself to overcome this limitation are needed; (iii) while the architecture provides both the Zba and Zbb RISC-V bit manipulation standard extensions [10], the upstream GCC 10.3.0 toolchain isn't capable of emitting them nor the underlying GNU *as* assembler (shipped with GNU Binutils 2.36.1) is able to properly assemble them. Experiments with the latest upstream GCC version (*minimal* support for bit manipulations code generation landed in GCC 12 [7]) and the upstream development version of GNU Binutils (patches already merged [5], expected to be shipped with GNU Binutils 2.37.x) are needed to assess its impact on current STREAM measurements.

Regarding user applications, we carried out benchmarks for the quantumESPRESSO [20] suite, in particular using its LAX test driver, compiled with OpenMPI, that performs a blocked (and optionally distributed) matrix diagonalization as a benchmark representative of the full-scale application workload. For a  $512^2$  input matrix we obtained a value of  $1.44 \pm 0.05$  GFLOP/s (36% of the theoretical FPU efficiency) on a single node over a total test duration of  $37.40 \pm 0.14$  s.

## B. Power characterization

We characterised the system's power consumption under test, exploiting the set of nine power lines available on-boards with embedded shunt resistors for current monitoring.

Power consumption of a cluster node is characterised using a set of standard HPC benchmarks run on a single node with the maximum allowed parallelism. Additionally, we measured the system's power consumption in idle, when only normal OS services and daemons are running in the background to evaluate the impact of benchmark running on power consumption. Power measurement results are collected in Table VI. Figure 3 reports 8 seconds of power traces for each of the benchmark executed.

The power required by the system to run is comprised between 4.810 Watts, in idle, and 5.935 Watts when the most power-hungry computation is run. Most of the system consumption is due to the core subsystem, which absorbs 3.543 Watts on average, reaching a peak consumption of 4.097 Watts for CPU intensive benchmarks such as HPL. The results show two more main sources of power consumption. i) The PCIe subsystem consistently requires 1 Watt, roughly 20% of system consumption, even if nothing is attached to the HiFive Unmatched PCIe connector. ii) DDR4 memory requires between 0.638 Watts when the system is idle and 0.950 Watts when the STREAM benchmark is run with a data size sufficient to disrupt L2 data locality. In general, DDR memory subsystem power consumption sits between 12% and 18% of the overall. The PLL subsystem and the IO interfaces together stand below 1% of the overall consumption for the tested workloads.

Figure 4 reports 80 seconds of power traces measured during the boot process. It is interesting to note a region of power consumption ( $4s < t < 10s$ ) at which the core complex it is powered on, but PLL (reported in yellow) is not active yet, we call these regions, *R1*. The average power consumption of the core complex in that region is 0.984 Watt.

As soon as the PLL activates, the power consumption jumps to a value of 2.561 Watts (*R2*) which increases to the value of 3.082 Watts, comparable with idle power for  $t > 40s$  *R3*. These three regions are of interest as they allow us to estimate the three components of the power consumption, which are hard to extract from a commercial off-the-shelf device without complex laboratory equipment. As in region *R1*, only the power supply but no clock is applied to the core complex, which is consuming only leakage power, which accounts for 32% of the idle power. In region *R2*, the clock is propagated to the core complex, but the operating system is not yet loaded, memory is initialising, and boot-loader tasks are ongoing. This power consumption accounts mainly for the clock tree and core's dynamic power. In region *R3*, the operating system is executing, but no active workload is in execution.

We can thus conclude that the operating system power accounts for the gap between *R3* and Idle power (3.072 Watts) and *R2* power consumption (2.561 Watts), which is (0.514 Watts) the 17% of the idle power. Conversely the difference

TABLE VI: Power consumption

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

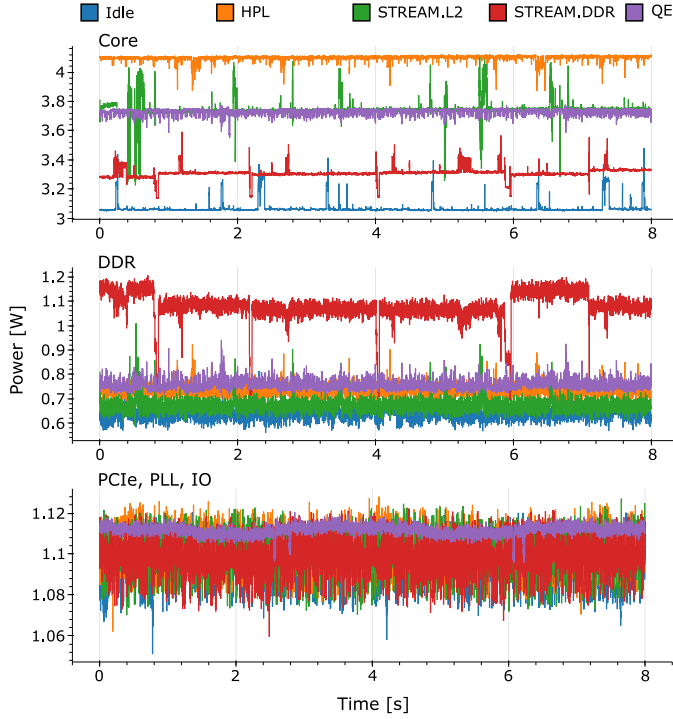


Fig. 3: Snapshot of the power consumption of the core (top), the DDR (middle) and the PCIe, PLL and IO subsystems (bottom). The traces are obtained observing power consumption for 8 seconds during benchmark run and averaging raw data using 1 ms windows.

between R2 and R1 accounts for the dynamic and clock tree power, which is 1.577 Watts equal to the 51% of the core idle power. By focusing to the DDR subsystem (ddr\_mem) we can make the similar consideration having in R1 0.275 Watts of leakage power, which is the 68% of their idle power the remaining part 32% is expected to be self-refresh and O.S. accesses for house keeping during O.S. idle period.

### C. Temperature monitoring

We used the ExaMon monitoring subsystem to observe and report the cluster's activity to pinpoint inefficiencies

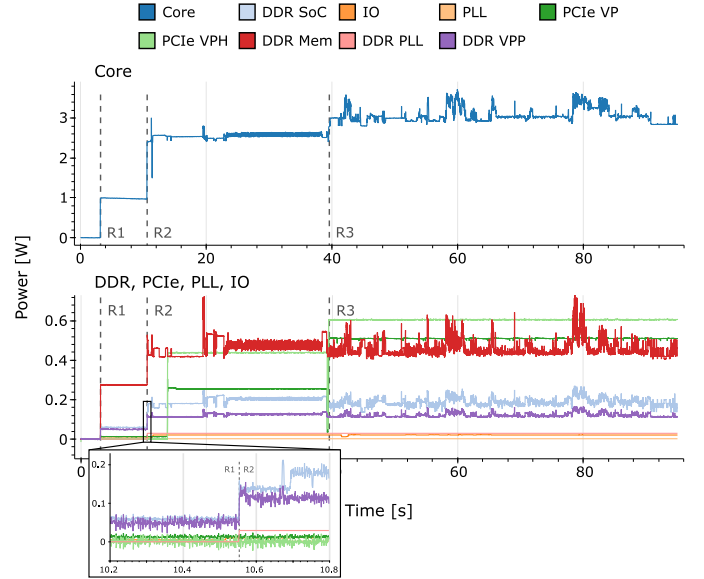


Fig. 4: Power consumption for the Core (top), DDR, PCIe, PLL and IO (bottom) subsystems during system boot. Boot phases: power-on (R1), bootloader (R2), O.S. boot (R3). The Figure also shows the detail of PLL activation.

and find opportunities for performance, power and thermal optimization. Figure 5 reports the nodes activity during the HPL run. From it we can identify the communication patterns, corresponding to a lower instruction count. We can expect to achieve higher performance once the RDMA will be supported over infiniband.

Figure 6 reports the temperature measured at the available sampling points over time. We can notice that during the first HPL runs, we encountered a thermal hazard on node 7, which reached 107°C and stopped executing. We noticed that the nodes in the centre blades were significantly hotter than the remaining ones by further inspection. This is an effect of the 1U case and the suboptimal airflow design that needs improvements to remove the heat generated by the PSUs. We addressed this issue by removing the lid and increasing the vertical spacing between the blades. This led to a significant

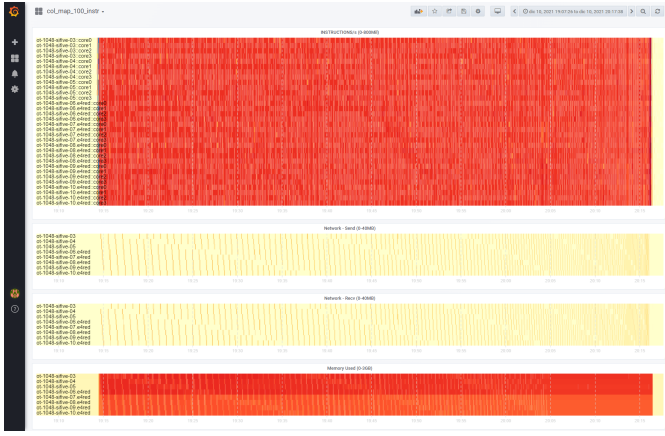


Fig. 5: ExaMon - HPL Heatmap: Instructions/s, Network traffic, Memory Usage

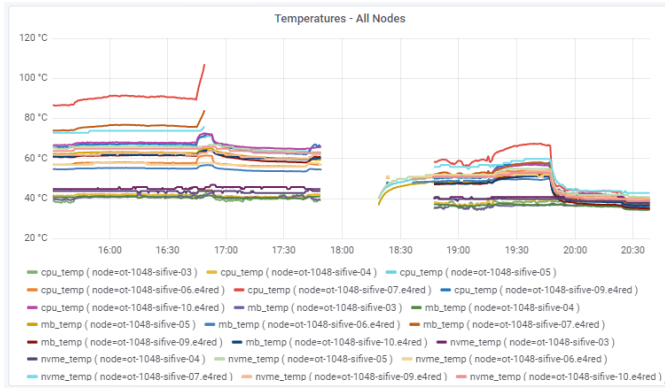


Fig. 6: Thermal runaway during HPL execution

reduction in the hotter node temperature, from 71°C to 39°C.

## VI. CONCLUSIONS AND FUTURE WORKS

In this manuscript we presented Monte Cimone : To the best of our knowledge, this is the first RISC-V cluster which is fully operational and supports a baseline HPC software stack, proving the maturity of the RISC-V ISA and the first generation of commercially available RISC-V components. We also evaluated the support for Infiniband network adapters which are recognised by the system, but are not yet capable to support RDMA communication.

We characterised in detail the power consumption of the SiFive Freedom U740 SoC for different workloads, measuring 4.81W in idle, with 64% due to core power (32% of leakage power, 51% dynamic and clock tree power and 17% by the O.S. workload), 13% related to DDR and 23% to the PCI subsystem. The power consumption increases to 5.935W under CPU intensive workloads. Furthermore, we ported the ExaMon ODA system on Monte Cimone and used it to detect thermal stability problems in the first configuration, which led to a thermal shutdown on the central node during the HPL run. We changed the enclosure to provide higher airflow to mitigate this issue.

Future work will focus on improving the software stack to achieve higher memory utilisation (i), to implement dynamic power and thermal management (ii), overcome the limitation in the Infiniband support (iv), extend Monte Cimone with PCIe RISC-V based accelerators (v).

## VII. ACKNOWLEDGMENTS

The study has been conducted in the context of the following projects: The european-project-initiative which 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 European PILOT project which 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. The REGALE project which 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 EPUX project which has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101033975. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Germany, Italy, Greece, United Kingdom, Czech Republic, Croatia.

## REFERENCES

- [1] The armida hpc system at e4.
- [2] Esperanto.ai. <https://esperanto.ai>.
- [3] European processor initiative. <https://www.european-processor-initiative.eu>.
- [4] The marconi100 hpc system at cineca.
- [5] RISC-v: Add support for Zbs instructions.
- [6] Risc-v exchange: Cores & socs. <https://riscv.org/exchange/cores-socs>.
- [7] RISC-v: Minimal support of bitmanip instructions.
- [8] Risc-v targets data centers. <https://semiengineering.com/risc-v-targets-data-center/>.
- [9] Semidynamics high bandwidth risc-v ip cores. <https://semidynamics.com>.
- [10] SiFive u74-MC core complex manual.
- [11] Special interest group: High-performance computing (hpc). <https://lists.riscv.org/g/sig-hpc>.
- [12] Andrea Bartolini, Francesco Beneventi, and et al. Paving the Way Toward Energy-Aware and Automated Datacentre. In *Proceedings of the 48th International Conference on Parallel Processing: Workshops, ICPP 2019*, pages 8:1–8:8, New York, NY, USA, 2019. ACM.
- [13] Elizabeth Bautista, Melissa Romanus, and et al. Collecting, monitoring, and analyzing facility and systems data at the national energy research scientific computing center. In *Proceedings of the 48th International Conference on Parallel Processing: Workshops, ICPP 2019*, 2019.
- [14] F. Beneventi, A. Bartolini, and et al. Continuous learning of hpc infrastructure models using big data analytics and in-memory processing tools. In *Proceedings of the Conference on Design, Automation & Test in Europe*, pages 1038–1043. European Design and Automation Association, 2017.

- [15] C. Chen Chen, X. Xiang, and et al. Xuantie-910: A commercial multi-core 12-stage pipeline out-of-order 64-bit high performance risc-v processor with vector extension : Industrial product. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, pages 52–64, 2020.
- [16] Massimiliano Culpo, Gregory Becker, and et al. archspec: A library for detecting, labeling, and reasoning about microarchitectures. In *2020 2nd International Workshop on Containers and New Orchestration Paradigms for Isolated Environments in HPC (CANOPIE-HPC)*, pages 45–52. IEEE.
- [17] Alexander Dörflinger, Mark Albers, and et al. A comparative survey of open-source application-class risc-v processor implementations. In *Proceedings of the 18th ACM International Conference on Computing Frontiers*, CF '21, page 12–20, 2021.
- [18] John L. Furlani. Modules : Providing a flexible user environment. 1991.
- [19] Todd Gamblin, Matthew LeGendre, and et al. The spack package manager: bringing order to HPC software chaos. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–12. ACM.
- [20] Paolo Giannozzi, Oscar Basergio, and et al. Quantum espresso toward the exascale. *The Journal of Chemical Physics*, 152(15):154105, 2020.
- [21] John L. Hennessy and David A. Patterson. A new golden age for computer architecture. *Communications of the ACM*, 62(2):48–60, 2019.
- [22] RISC-V International. RISC-v ABIs specification.
- [23] Antonio Libri, Andrea Bartolini, and Luca Benini. pAella: Edge-AI based Real-Time Malware Detection in Data Centers. *IEEE Internet of Things Journal*, pages 1–1, 2020.
- [24] Michael Malmes, Marcin Ostasz, Maïke Gilliot, Pascale Bernier-Bruna, Laurent Cargemel, Estela Suarez, Herbert Cornelius, Marc Duranton, Benny Koren, Pascale Rosse-Laurent, et al. *ETP4HPC's Strategic Research Agenda for High-Performance Computing in Europe 4*. PhD thesis, European Technology Platform for High-Performance Computing (ETP4HPC), 2020.
- [25] John D. McCalpin. Memory bandwidth and machine balance in current high performance computers. *IEEE Computer Society Technical Committee on Computer Architecture (TCCA) Newsletter*, pages 19–25, December 1995.
- [26] Alessio Netti, Woong Shin, and et al. A Conceptual Framework for HPC Operational Data Analytics. In *2021 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 596–603, September 2021.
- [27] Antoine Petitot, R. Whaley, and et al. Hpl – a portable implementation of the high-performance linpack benchmark for distributed-memory computers. 01 2008.
- [28] Nikola Rajovic, Alejandro Rico, Filippo Mantovani, and et al. The Mont-Blanc Prototype: An Alternative Approach for HPC Systems. pages 444–455, November 2016.
- [29] Mitsuhiro Sato, Yutaka Ishikawa, and et al. Co-design for A64FX manycore processor and "Fugaku". In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20, pages 1–15, November 2020.
- [30] Colin Schmidt, John Wright, and et al. 4.3 an eight-core 1.44ghz risc-v vector machine in 16nm finfet. In *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, volume 64, pages 58–60, 2021.
- [31] SiFive. RISC-v large code model software workaround.
- [32] Florian Zaruba, Fabian Schuiki, and Luca Benini. Manticore: A 4096-core risc-v chiplet architecture for ultraefficient floating-point computing. *IEEE Micro*, 41(2):36–42, 2021.