Monte Cimone: Exploring RISC-V High Performance Compute Cluster

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

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

Objective: 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, but also a complete software environment for HPC, as well as a full-featured system monitoring infrastructure. We demonstrate that it is possible to run real-life HPC applications on Monte Cimone today.

Monte Cimone Hardware Architecture:
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 E4 RISC-V SoC from SiFive and integrates:
- Four U74 RV64GC computing cores, running up to 2.1 GHz
- 32GB of DDR
- 1TB node-local NVMe storage
- PCIe expansion cards

Power Characterization:

<table>
<thead>
<tr>
<th>System</th>
<th>HPL Efficiency [%FPU utilization]</th>
<th>Stream Efficiency [%Bandwidth Utilization]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monte Cimone</td>
<td>46.5%</td>
<td>15.5%</td>
</tr>
<tr>
<td>Armida (ARMv8a, Marvell ThunderX2)</td>
<td>65.8%</td>
<td>63.2%</td>
</tr>
<tr>
<td>Marconi100 (ppc64le, IBM Power9)</td>
<td>59.7%</td>
<td>48.2%</td>
</tr>
</tbody>
</table>

We characterised the power consumption of various applications executed on Monte Cimone:
- Idle: 4.81W (64% of core power, 13% related to DDR and 23% related to PCIe subsystem)
- HPL 5.935W (69% of core power, 14% related to DDR and 18% related to PCIe subsystem)

HPL Benchmark:
- HPL peak theoretical value of 1.0 GFLOP/s/core, (from the micro-architecture specification)
- 4.0 GFLOP/s peak value for a single chip, the upstream HPL benchmark
- a sustained value of 1.86 ± 0.04 GFLOP/s on a single node (on a N=40704 and N=193 and a total runtime of 24105 ± 587 s)

HPLmultinode strong scaling w. 16Gb/s network:
- 39.5% of the entire machine's theoretical peak
- 85% of the extrapolated attainable peak in case of perfect linear scaling from the single-node case

QE Benchmark:
- LAMMPS test driver of the QuantumESPRESSO suite, compiled with OpenMPI
- performs a blocked (and optionally distributed) matrix diagonalisation for a 512² input matrix (benchmark representative of the full-scal application workload).
- 1.44 ± 0.05 GFLOP/s (36% of the theoretical FPU efficiency) on a single node over a total test duration of 37.40 ± 0.14 s.

STREAM Benchmark:
- SIFive U74 RV64C SoC peak DDR bandwidth 7760 MB/s. Possible causes to be investigate:
  1. 12 prefetcher capable of tracking up to eight streams, then why it is not hiding the DDR latency?
  2. Overall data size used by STREAM is currently limited by the RISC-V code model. The medany code model requires that every linked symbol resides within ± 2GB range from the pc register.
  3. 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.
  4. The upstream GCC 10.3.0 toolchain isn’t capable of emitting the 25a and 26b RISC-V bit manipulation standard extensions nor the underlying GNU as assembler (shipped with GNU Binutils 2.36.1). Experimentally supported on GCC 12 and Binutils 2.37.x.