Monte Cimone v2: HPC RISC-V Cluster Evaluation and Optimization

Emanuele Venieri[1], Simone Manoni[1], Gabriele Ceccolini[1], Giacomo Madella[1], Federico Ficarelli[2], Daniele Gregori[3] , Andrea Acquaviva[1], Luca Benini[1,4], Andrea Bartolini[1]

[1]*University of Bologna, Italy*, [2]*CINECA, Italy*, [3]*E4 Company S.p.A.*, [4]*ETH Zürich Switzerland*

emanuele.venieri2@unibo.it

Outline

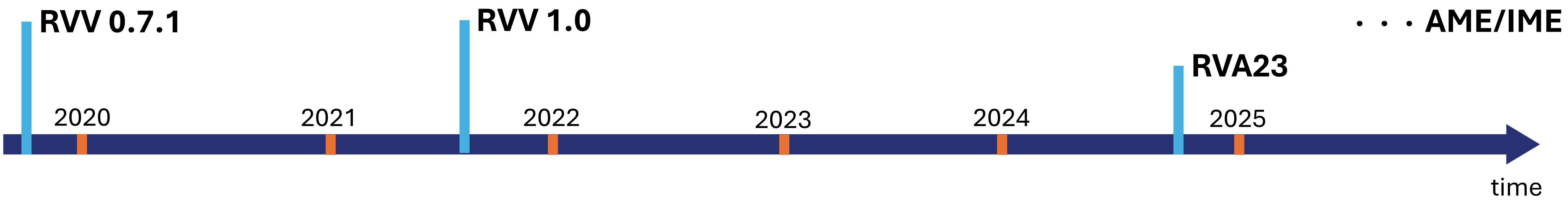
- RISC-V & HPC
- Monte Cimone evolution
- Performance evaluation
- BLAS vectorization
- Results

RISC-V & HPC

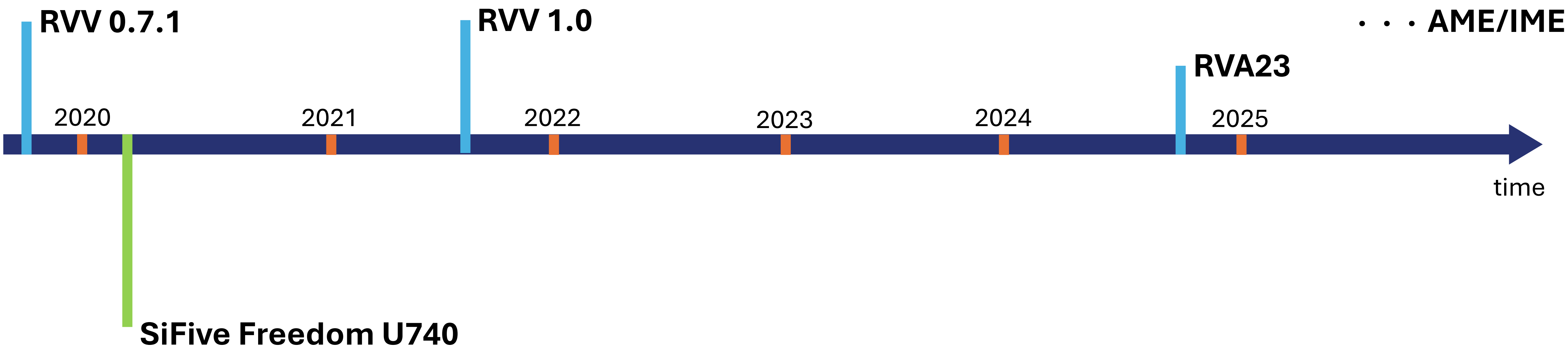
- HPC compute node requirements:
  - High core count + vector/matrix extensions
  - Large amount of system memory
  - PCIe support (accelerators, network cards, fast storage)
  - Software: OS, job schedulers, monitoring systems, packages and libraries

- RISC-V is mature in embedded domain
- How to foster RISC-V ISA uptake in HPC?

RISC-V HPC & Monte Cimone timeline



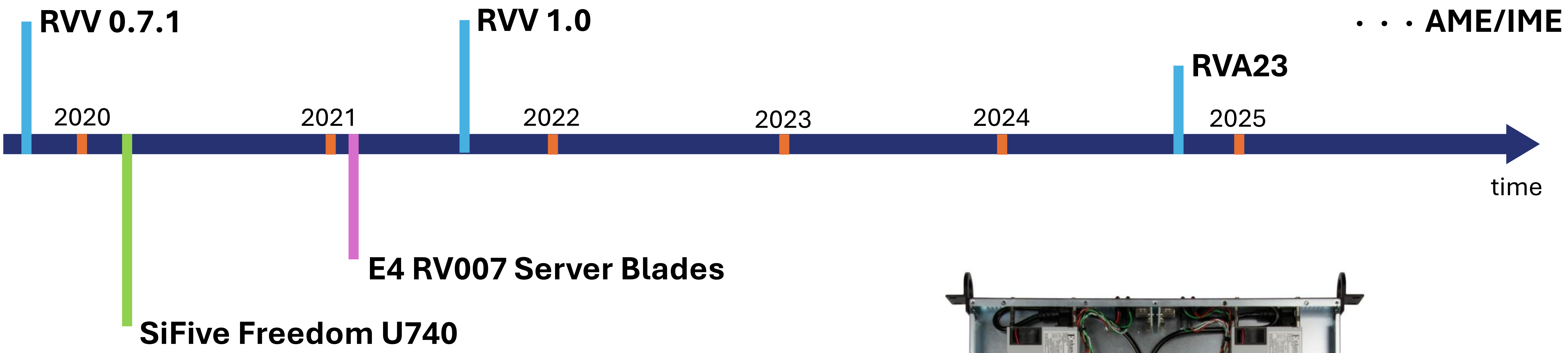
RISC-V HPC & Monte Cimone timeline



SiFive Freedom U740

- First RISC-V PC
- 4 core system RV64GC
- 16 GB DDR4
- PCIe Gen 3
- Linux ready

RISC-V HPC & Monte Cimone timeline



E4 RV007 Server Blades

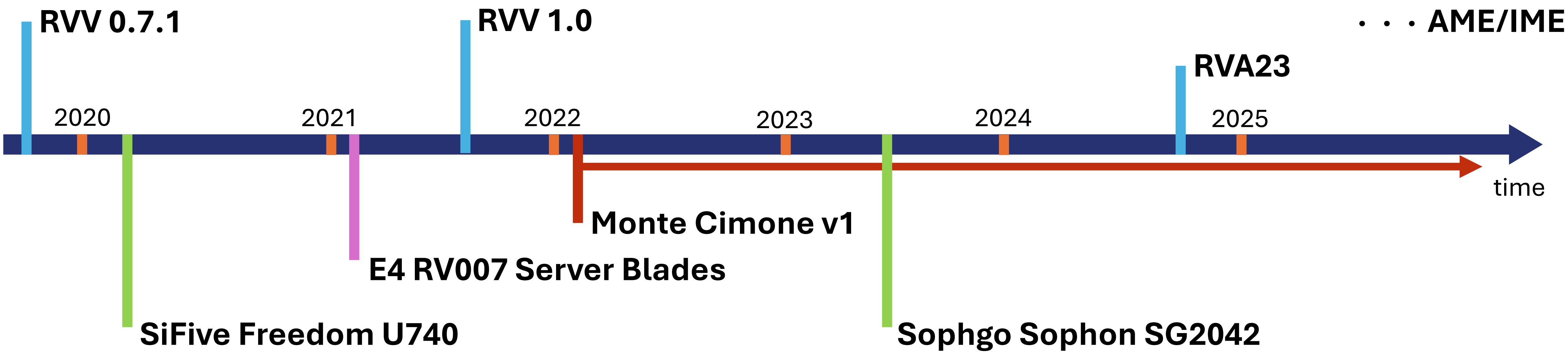
- Two SiFive Freedom U740
- Server form factor

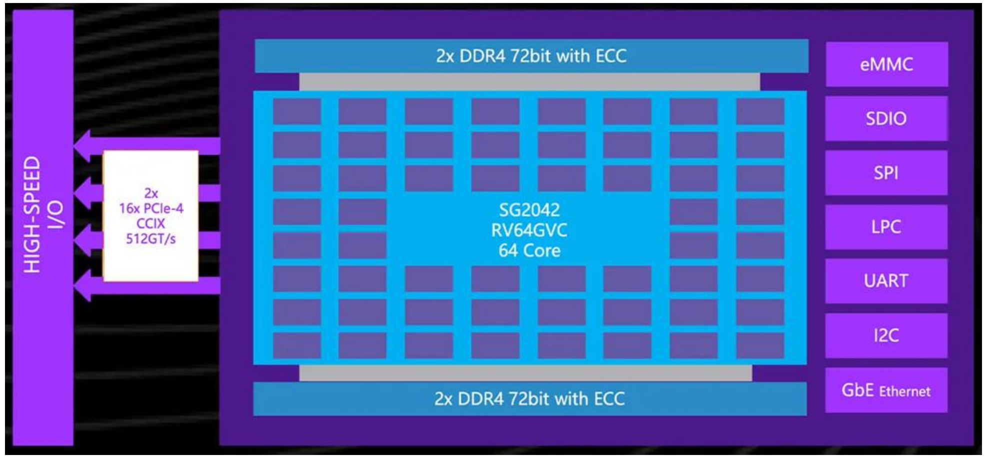
RISC-V HPC & Monte Cimone timeline

Monte Cimone v1

- First RISC-V cluster
- Four E4 RV007 blades
- 32 cores in total + 128 GB distributed system memory
- Scope: infrastructure + software validation

RISC-V HPC & Monte Cimone timeline

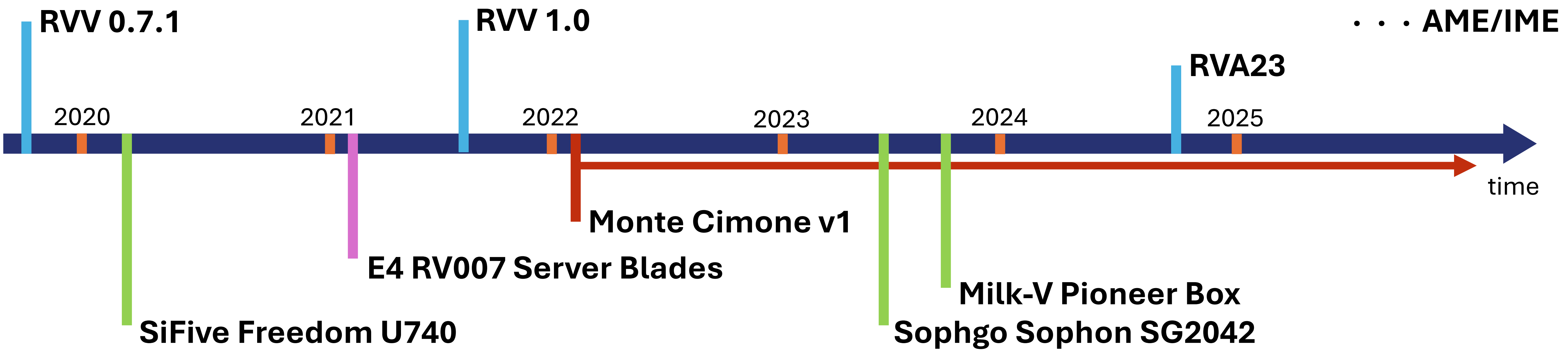




Sophgo Sophon SG2042

- First server class platform
- 64 cores RV64GC_0p7V
- 3-levels cache system
- ECC DDR4 memory
- 32 PCIe Gen 4 lanes

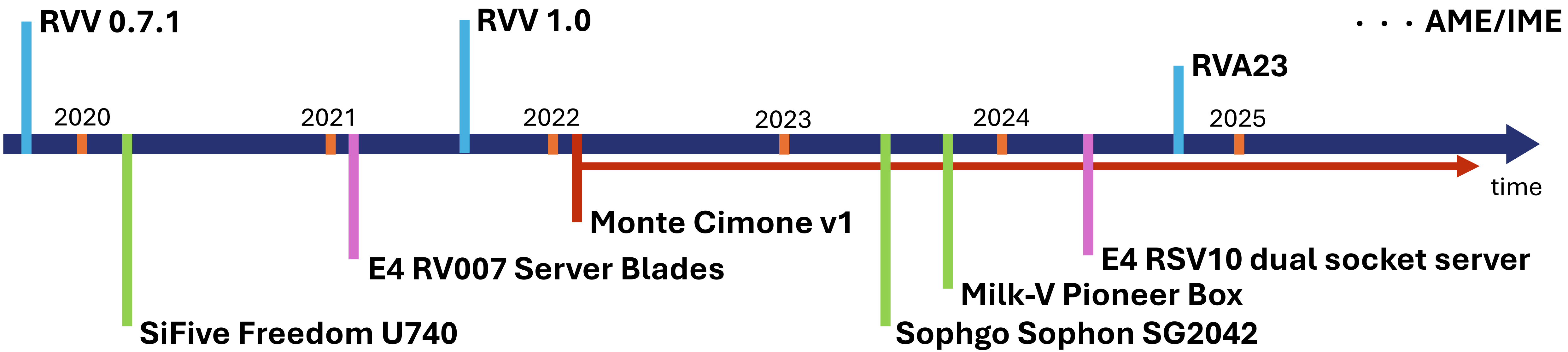
RISC-V HPC & Monte Cimone timeline



Milk-V Pioneer Box

- Single socket SG2042
- 128 GB RAM

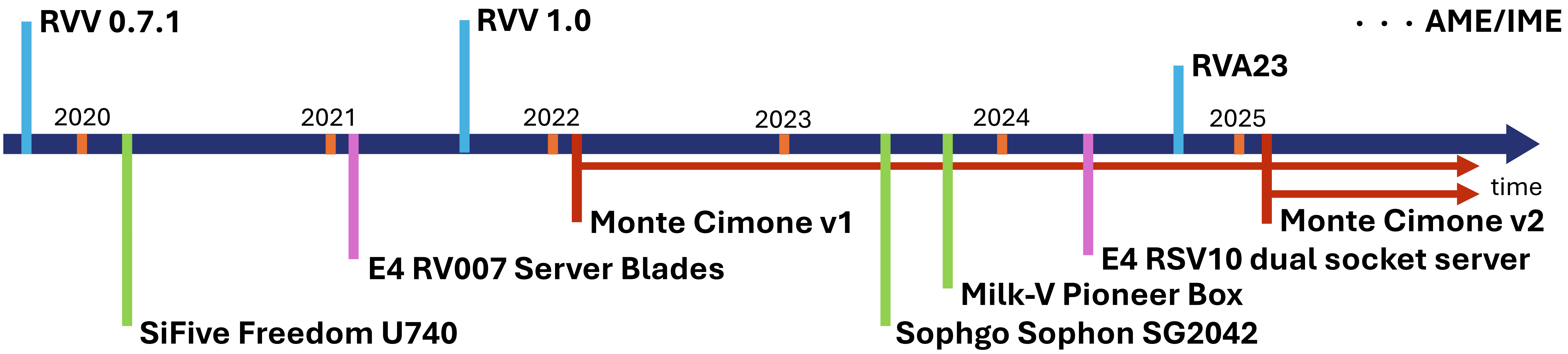
RISC-V HPC & Monte Cimone timeline



E4 RSV10 dual socket server

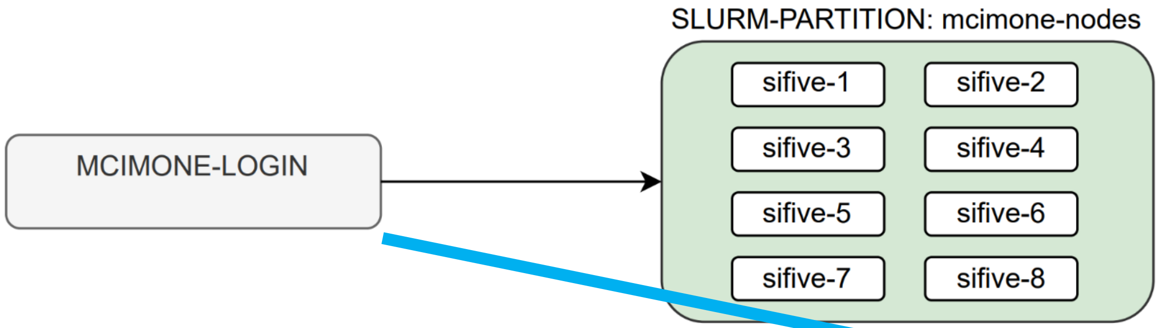
- Dual socket SG2042
- 256 GB RAM

RISC-V HPC & Monte Cimone timeline



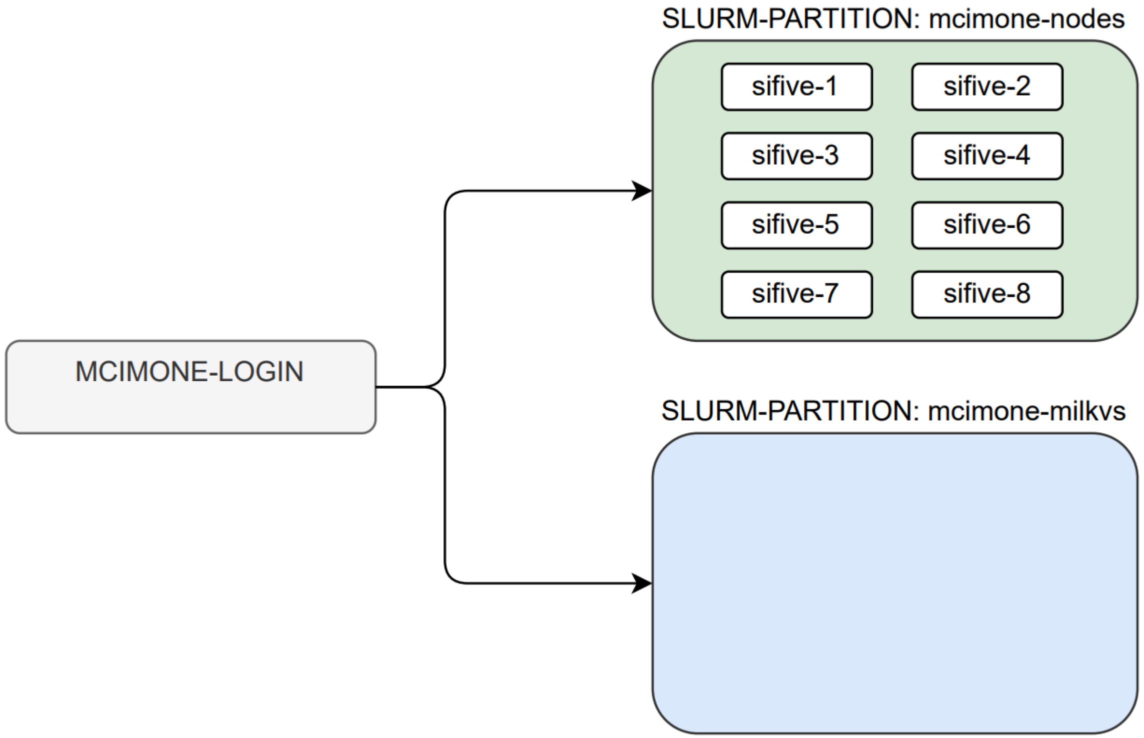
Monte Cimone v2 bring up

Starting from Monte Cimone v1



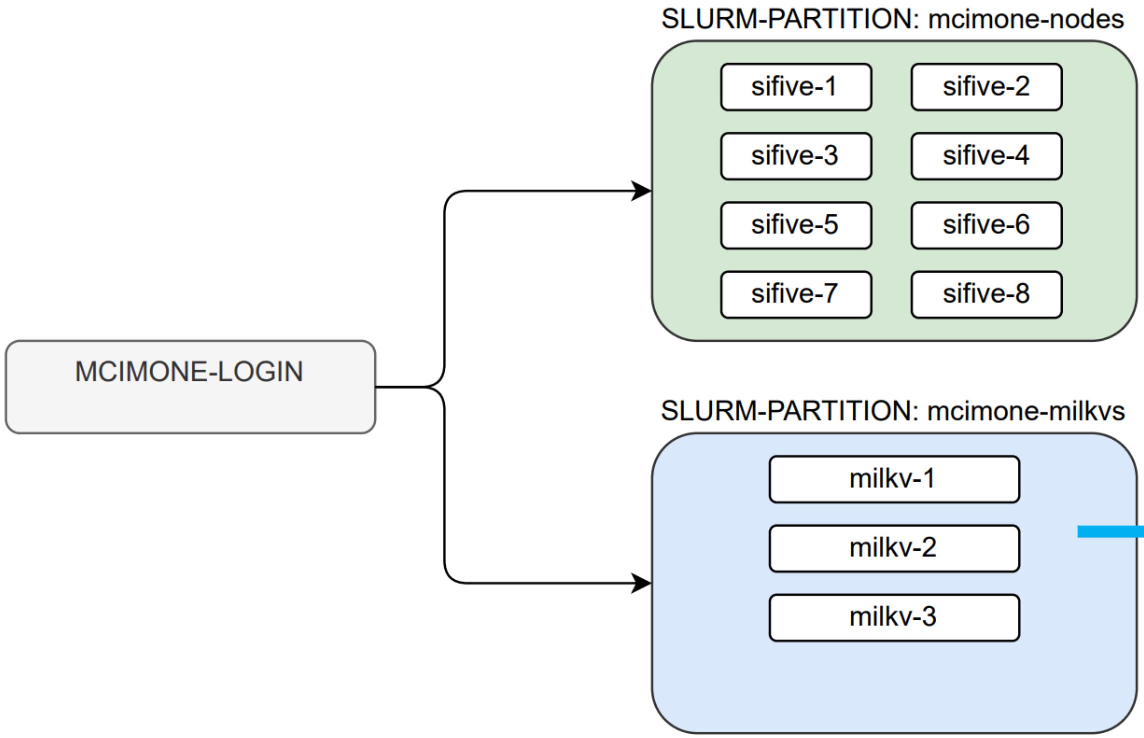
Monte Cimone v2 bring up

- New SLURM partition



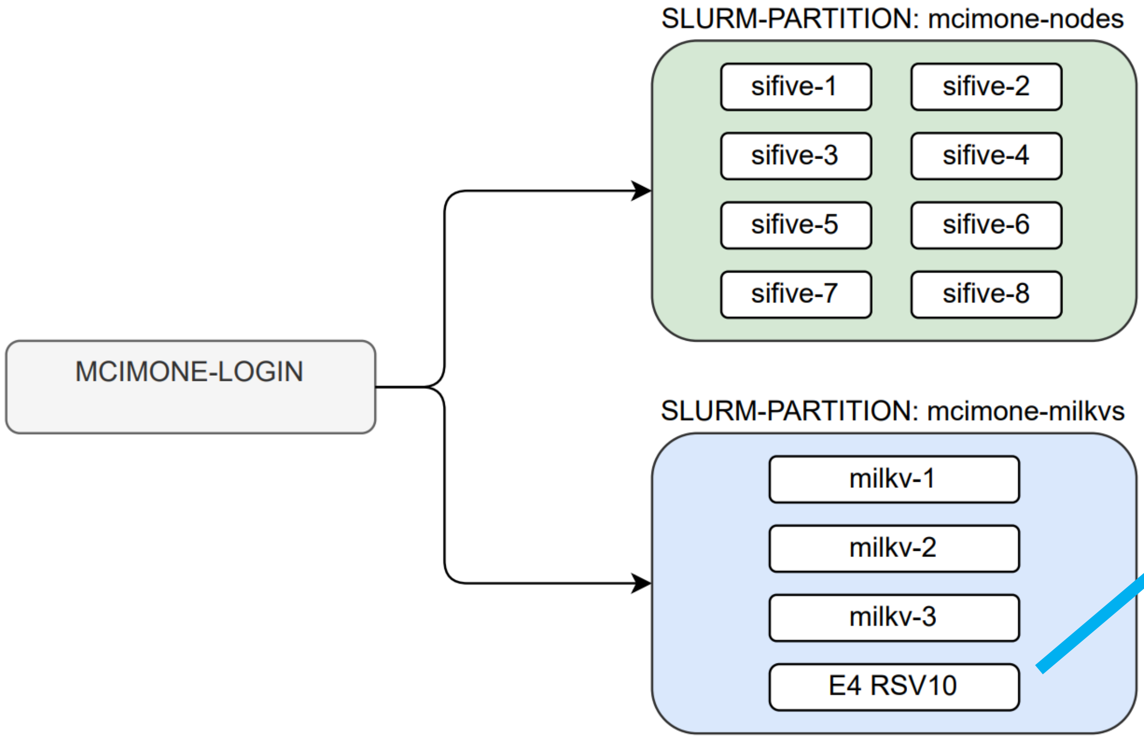
Monte Cimone v2 bring up

- New SLURM partition
  - 3 Milk-V Pioneer Box



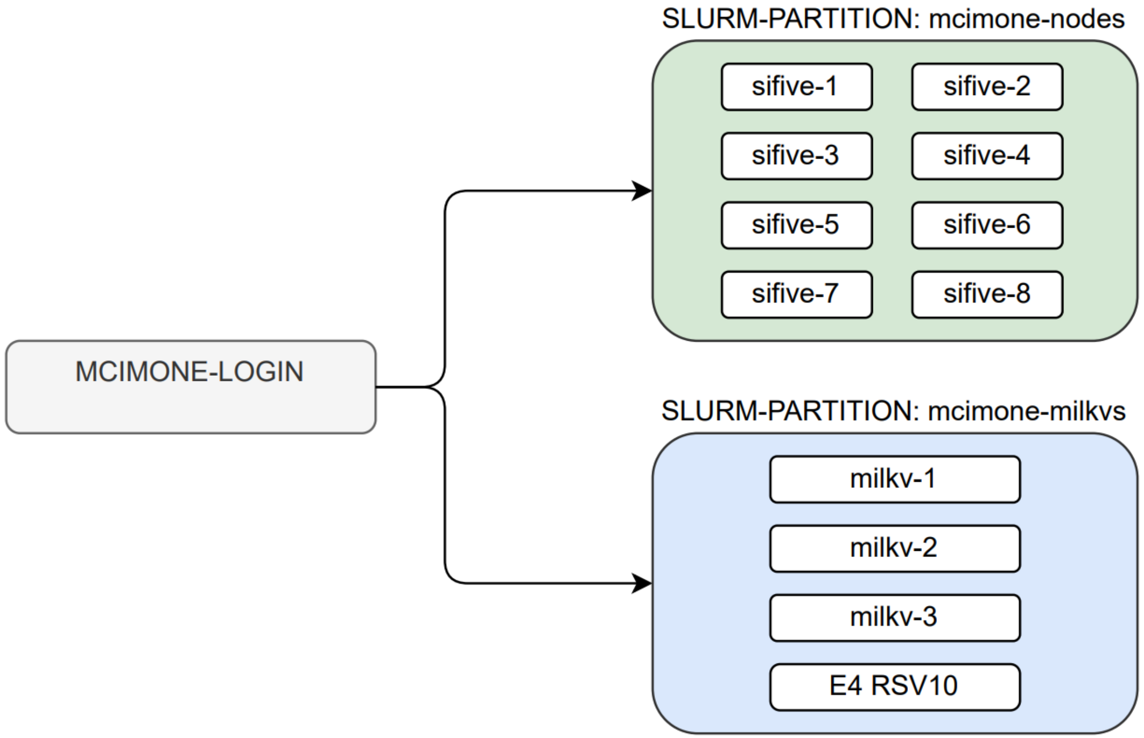
Monte Cimone v2 bring up

- New SLURM partition
  - 3 Milk-V Pioneer Box
  - 1 E4 RSV10 Dual Socket RISC-V Server



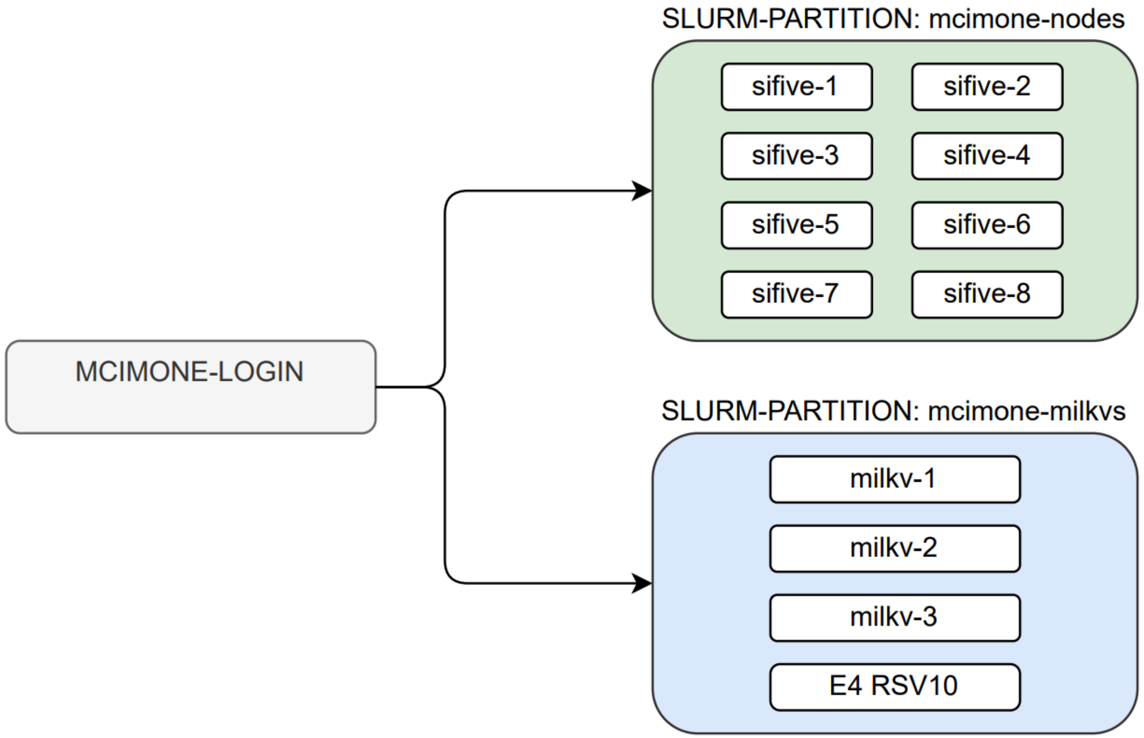
Cluster environment

- SLURM



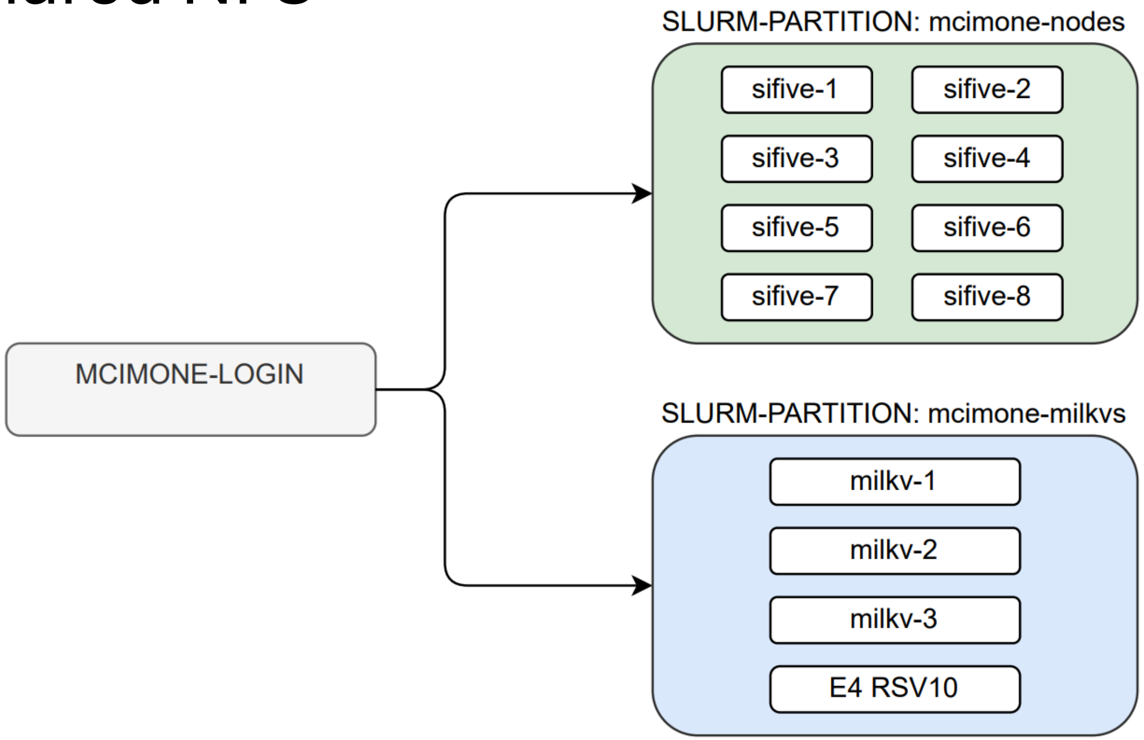
Cluster environment

- SLURM
- Examon



Cluster environment

- SLURM
- Examon
- Shared NFS



Cluster environment

- SLURM
- Examon
- Shared NFS
- Spack + tools

```
----------------------------------------- /opt/share/spack/modules/linux-fedora38-riscv64 ---------------------------
autoconf/2.69/gcc-13.1.1-qhn4yrq                          libxml2/2.9.12/gcc-13.1.1-hsxijdm
automake/1.16.3/gcc-13.1.1-7nxi645                        libxml2/2.9.12/gcc-13.2.1-z3zbxqs
berkeley-db/18.1.40/gcc-13.1.1-lxhzsx5                    m4/1.4.19/gcc-13.1.1-i4vpilc
berkeley-db/18.1.40/gcc-13.2.1-hp5nxny                    m4/1.4.19/gcc-13.2.1-yxpwh6s
bzip2/1.0.8/gcc-13.1.1-ng7qlnb                            ncurses/6.2/gcc-13.1.1-7zxllwi
bzip2/1.0.8/gcc-13.2.1-5ftm54j                            ncurses/6.2/gcc-13.2.1-tf3ke2q
cmake/3.21.4/gcc-13.1.1-miqhn23                           netlib-lapack/3.9.1/gcc-13.1.1-6ogqvqr
cmake/3.21.4/gcc-13.2.1-gmkgtyz                           netlib-scalapack/2.1.0/gcc-13.1.1-openblas-openmpi-jk64vjy
diffutils/3.8/gcc-13.1.1-rstfsbg                          numactl/2.0.14/gcc-13.1.1-iusd2n5
diffutils/3.8/gcc-13.2.1-6ymjvui                          openblas/0.3.24-xuantie/gcc-13.1.1-vzf2ar6
expat/2.4.1/gcc-13.2.1-psiydy2                            openmpi/4.1.1/gcc-13.1.1-dagtgno
fftw/3.3.10/gcc-13.1.1-openmpi-h7dvbht                    openssh/8.7p1/gcc-13.1.1-bc4zgjl
gdbm/1.19/gcc-13.1.1-vvnxvcx                              openssl/1.1.1l/gcc-13.1.1-jbfpudk
gdbm/1.19/gcc-13.2.1-poile56                              openssl/1.1.1l/gcc-13.2.1-qhgnwzo
gettext/0.21/gcc-13.2.1-q4ysmpw                           pcre/8.44/gcc-13.2.1-bqb5mu5
gnuconfig/2021-08-14/gcc-13.1.1-ukvdjeo                   perl-data-dumper/2.173/gcc-13.2.1-l2mlwep
gnuconfig/2021-08-14/gcc-13.2.1-6e7nzx6                   perl/5.34.0/gcc-13.1.1-44jcn7m
gromacs/2021.3/gcc-13.1.1-openmp-openblas-openmpi-cppmmcj perl/5.34.0/gcc-13.2.1-ayklsrw
hpl/2.3/gcc-13.1.1-netlib-openmpi-2kp2xmy                 pkgconf/1.8.0/gcc-13.1.1-d5tydj3
hpl/2.3/gcc-13.1.1-openblas-openmpi-l65lo5i               pkgconf/1.8.0/gcc-13.2.1-slpwl57
hpl/2.3/gcc-13.1.1-openmp-openblas-openmpi-mlxx6me        pmix/4.1.2/gcc-13.1.1-jbftrd2
hwloc/2.6.0/gcc-13.1.1-gjnxly5                            python/3.8.12/gcc-13.2.1-jvhkedj
hwloc/2.6.0/gcc-13.2.1-vdbjczp                            readline/8.1/gcc-13.1.1-k3w2s2n
libbsd/0.11.3/gcc-13.2.1-zhfraqo                          readline/8.1/gcc-13.2.1-hlqg4yv
libedit/3.1-20210216/gcc-13.1.1-2xstcc7                   slurm/22.05.6/gcc-13.1.1-56tbvst
libedit/3.1-20210216/gcc-13.2.1-wyvqcir                   sqlite/3.36.0/gcc-13.2.1-ynn4tq4
libevent/2.1.12/gcc-13.1.1-v6fvwbk                        swig/4.0.2/gcc-13.2.1-4n7y3a7
libffi/3.3/gcc-13.2.1-ng3lwp7                             tar/1.34/gcc-13.2.1-xceklou
libiconv/1.16/gcc-13.1.1-5h2j62z                          util-linux-uuid/2.36.2/gcc-13.2.1-htybqgp
libiconv/1.16/gcc-13.2.1-pl6wjmc                          util-macros/1.19.3/gcc-13.1.1-iog7smq
libmd/1.0.3/gcc-13.2.1-6c34tq7                            util-macros/1.19.3/gcc-13.2.1-r2mutbi
libpciaccess/0.16/gcc-13.1.1-74m3dlt                      xz/5.2.5/gcc-13.1.1-hdi3zss
libpciaccess/0.16/gcc-13.2.1-4qgzygo                      xz/5.2.5/gcc-13.2.1-vmurcyz
libsigsegv/2.13/gcc-13.1.1-ershbg6                        z3/4.8.9/gcc-13.2.1-5mrlxk3
libsigsegv/2.13/gcc-13.2.1-lxya2dy                        zlib/1.2.11/gcc-13.1.1-oomzb7m
libtool/2.4.6/gcc-13.1.1-4ez6fpq                          zlib/1.2.11/gcc-13.2.1-usf3zjp
libtool/2.4.6/gcc-13.2.1-3t47ydt
```

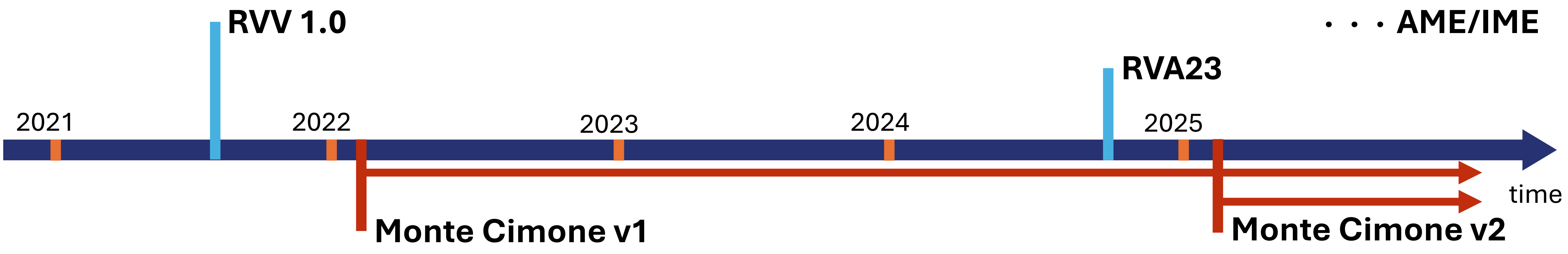
Cluster environment

- SLURM
- Examon
- Shared NFS
- Spack + tools
  - Added compilers:
    - Xuantie GNU Toolchain (vendor GCC w. RVV 0.7.1 support)
    - GCC 14 (first version w. xtheadvector)

Cluster environment

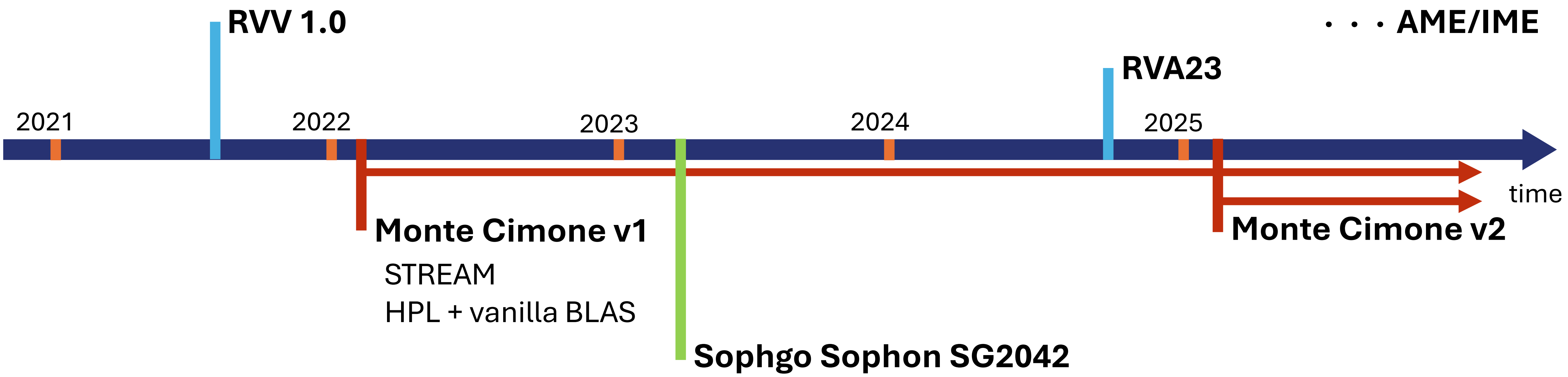
- SLURM
- Examon
- Shared NFS
- Spack + tools
  - Added compilers:
    - Xuantie GNU Toolchain (vendor GCC w. RVV 0.7.1 support)
    - GCC 14 (first version w. xtheadvector)
  - Added BLAS libraries:
    - OpenBLAS (vanilla no RVV, vendor optimized)
    - BLIS with proposed optimized kernels

Monte Cimone v2 benchmark

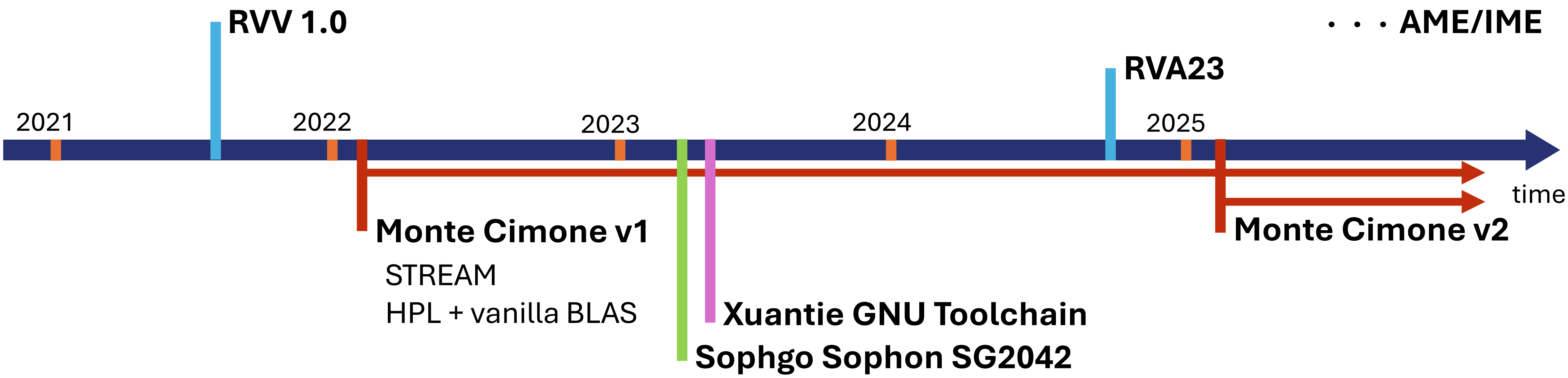


STREAM
HPL + vanilla BLAS

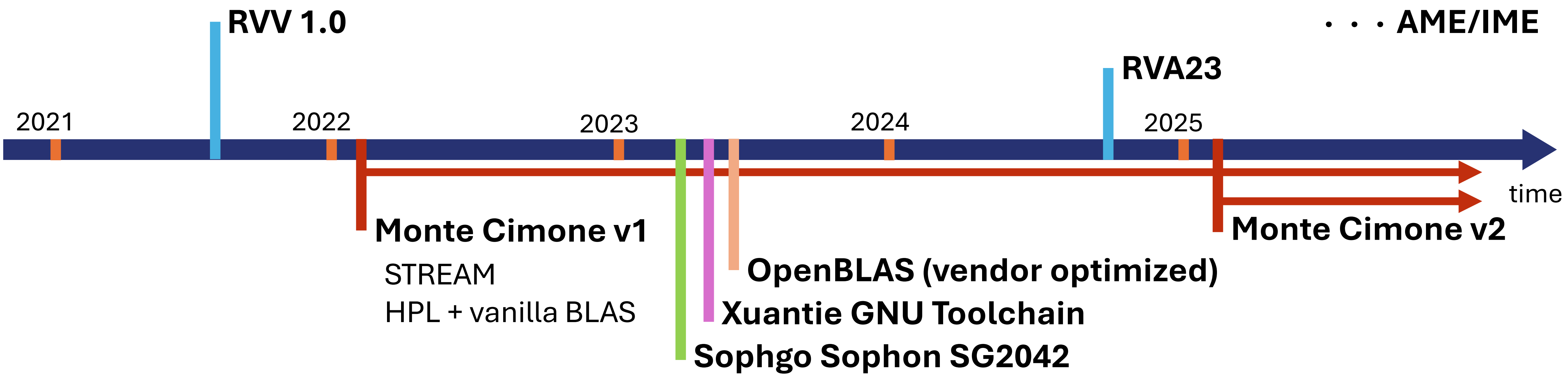
Monte Cimone v2 benchmark



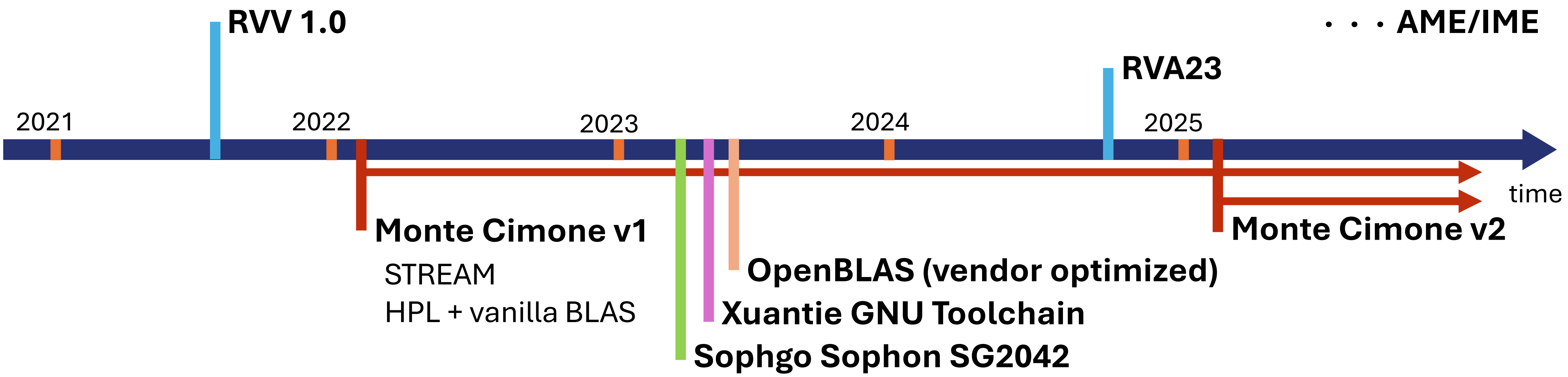
Monte Cimone v2 benchmark



Monte Cimone v2 benchmark

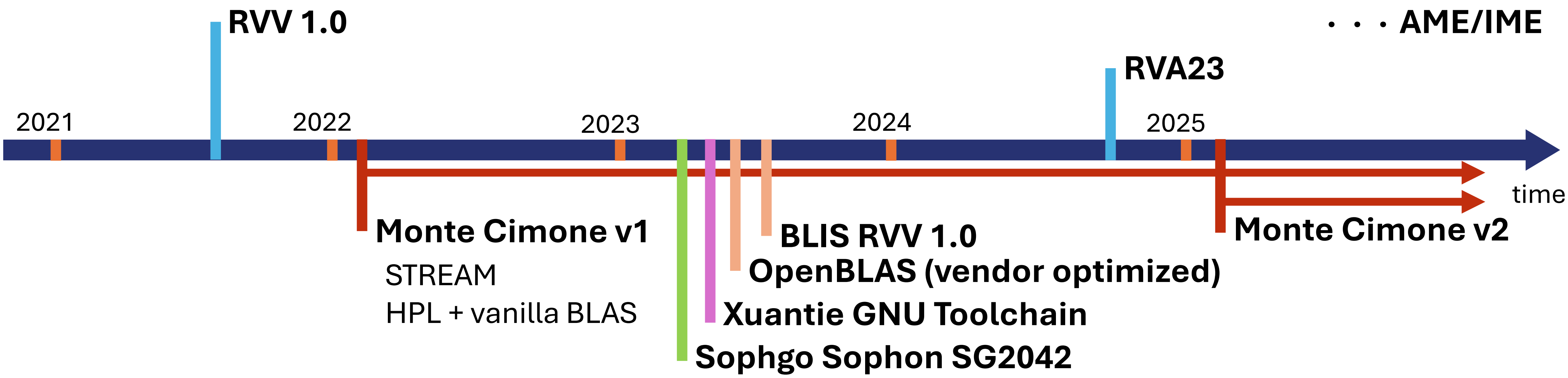


Monte Cimone v2 benchmark



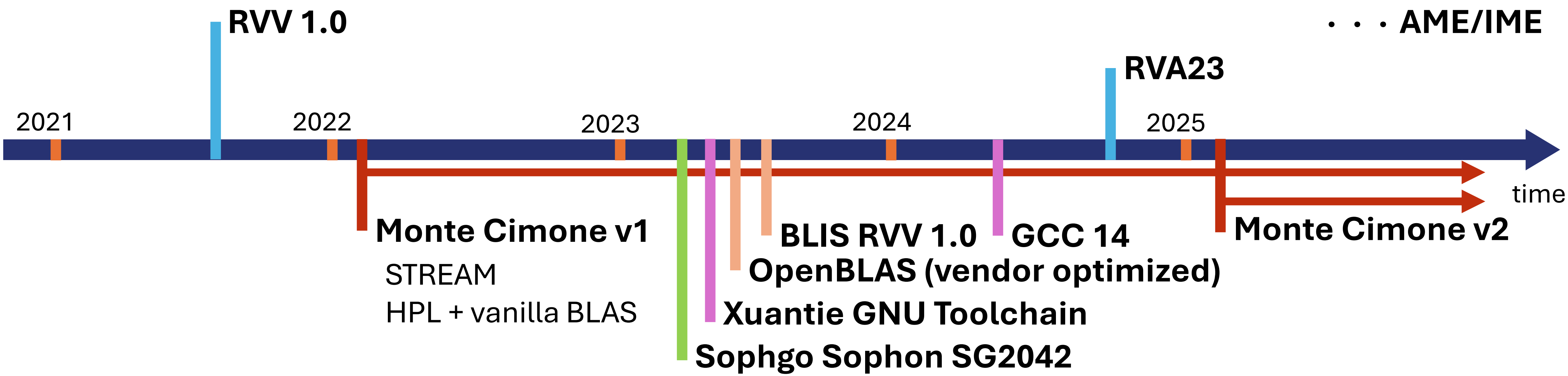
- **Performance assessment of new hardware:**
  - STREAM
  - HPL + OpenBLAS w/ Xuantie GNU Toolchain

Monte Cimone v2 benchmark



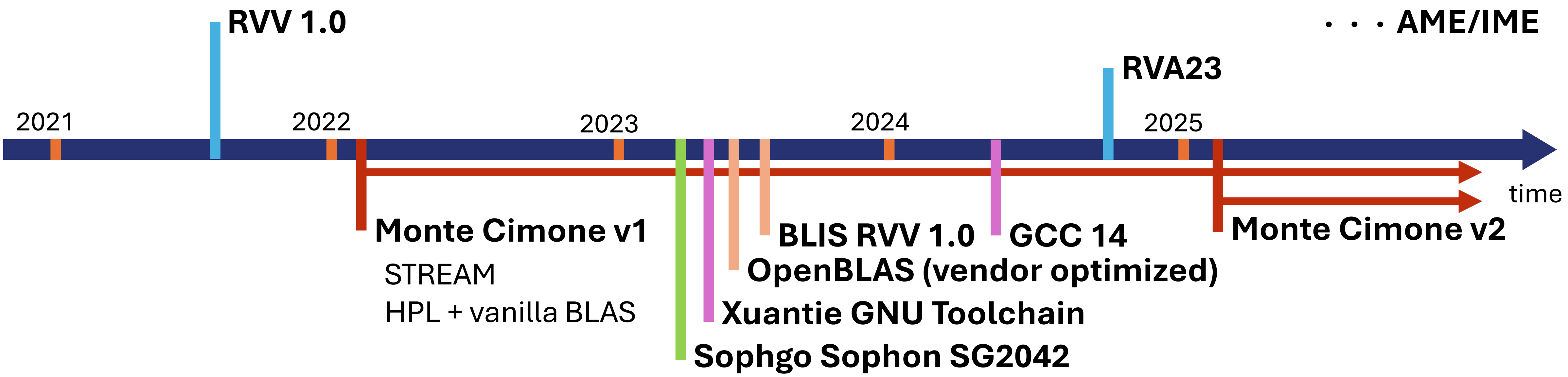
- **Performance assessment of new hardware:**
  - STREAM
  - HPL + OpenBLAS w/ Xuantie GNU Toolchain

Monte Cimone v2 benchmark



- **Performance assessment of new hardware:**
  - STREAM
  - HPL + OpenBLAS w/ Xuantie GNU Toolchain

Monte Cimone v2 benchmark

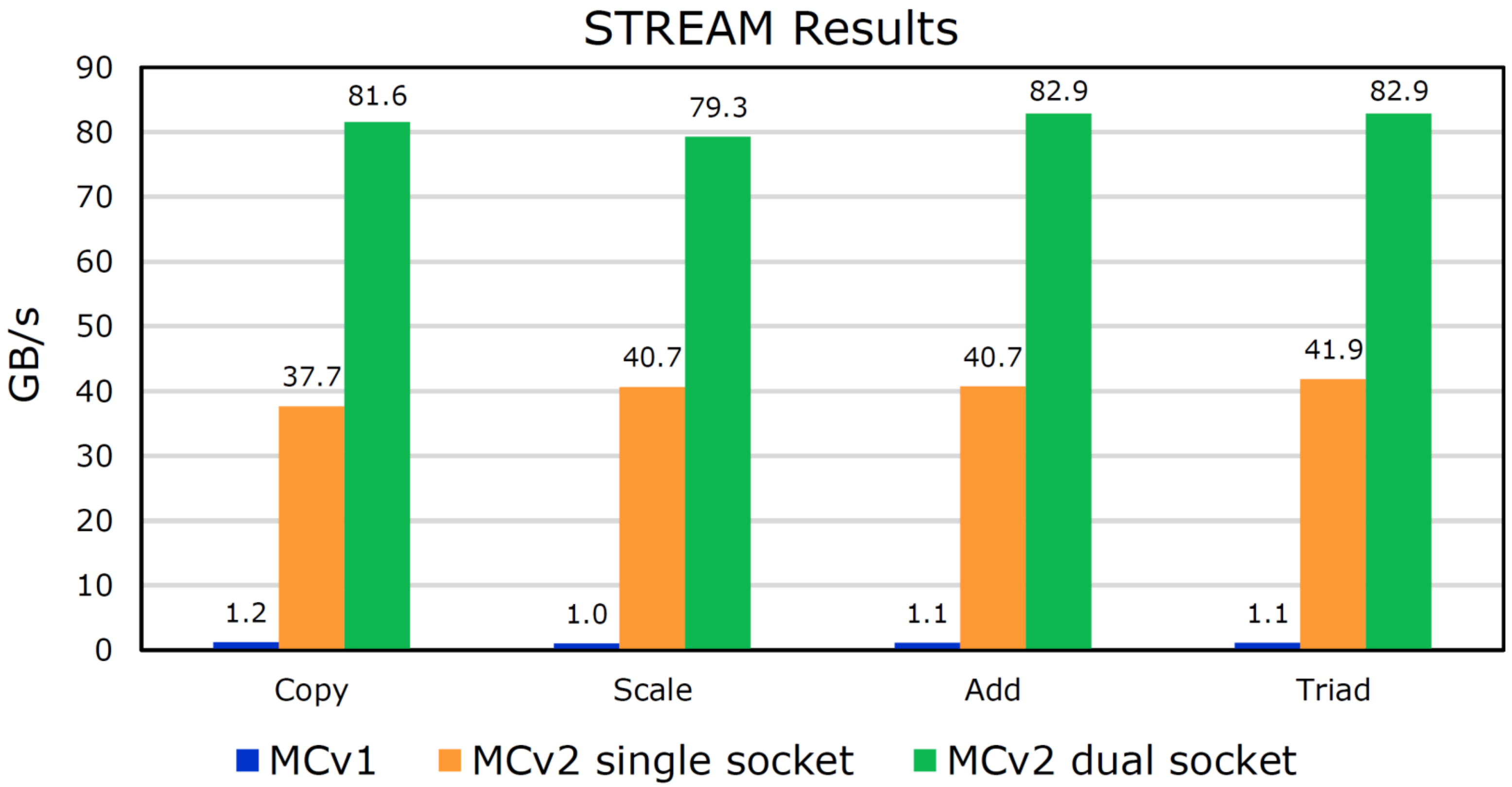


- **Performance assessment of new hardware:**
  - STREAM
  - HPL + OpenBLAS w/ Xuantie GNU Toolchain
- **BLIS library porting and optimization**

STREAM Results

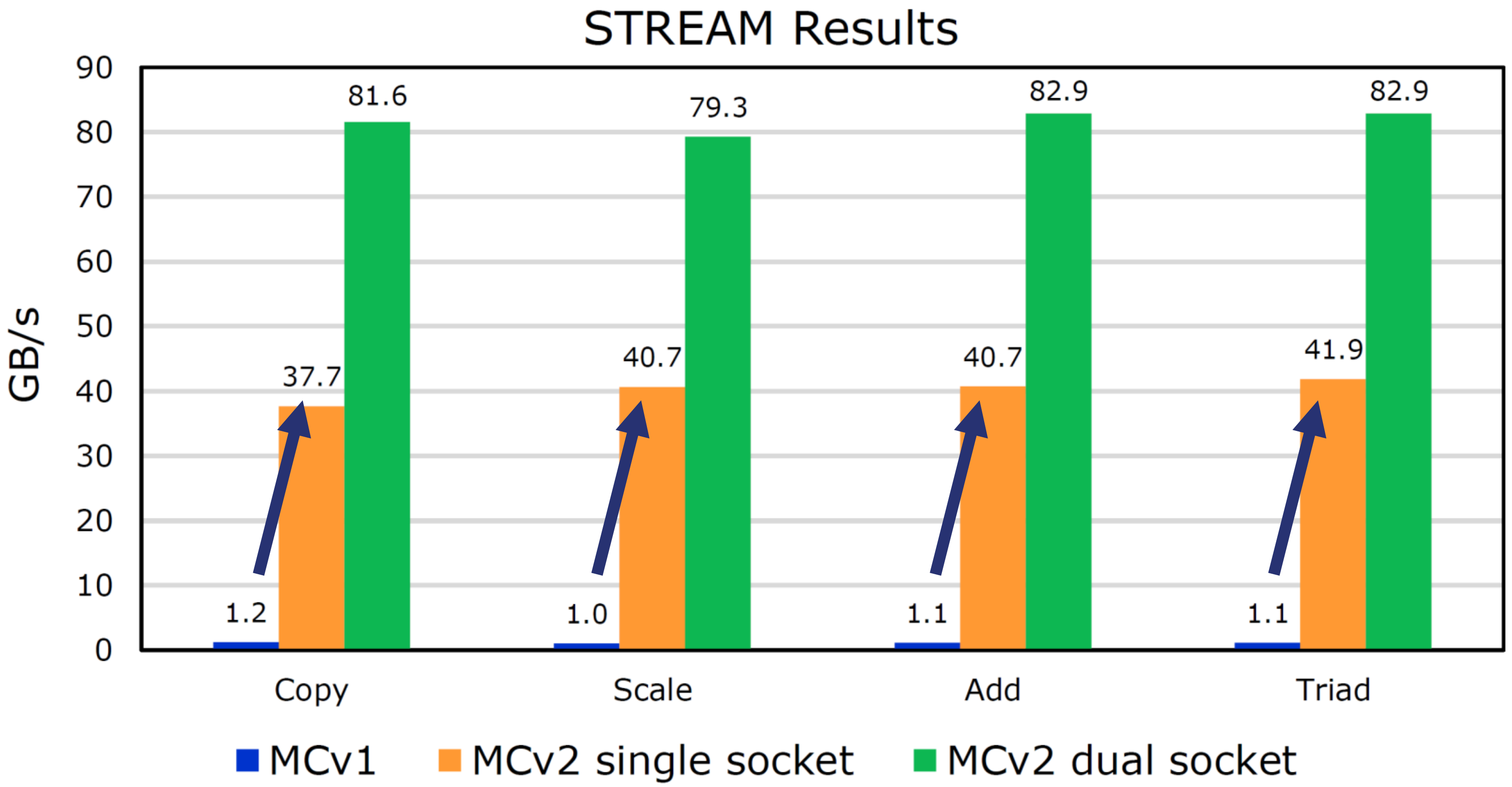
- **Monte Cimone v1 node:**
  - Max: 1.2 GB/s

- **Milk-V Pioneer Box:**
  - 64 threads
  - 4 memory channels
  - Max: 41.9 GB/s



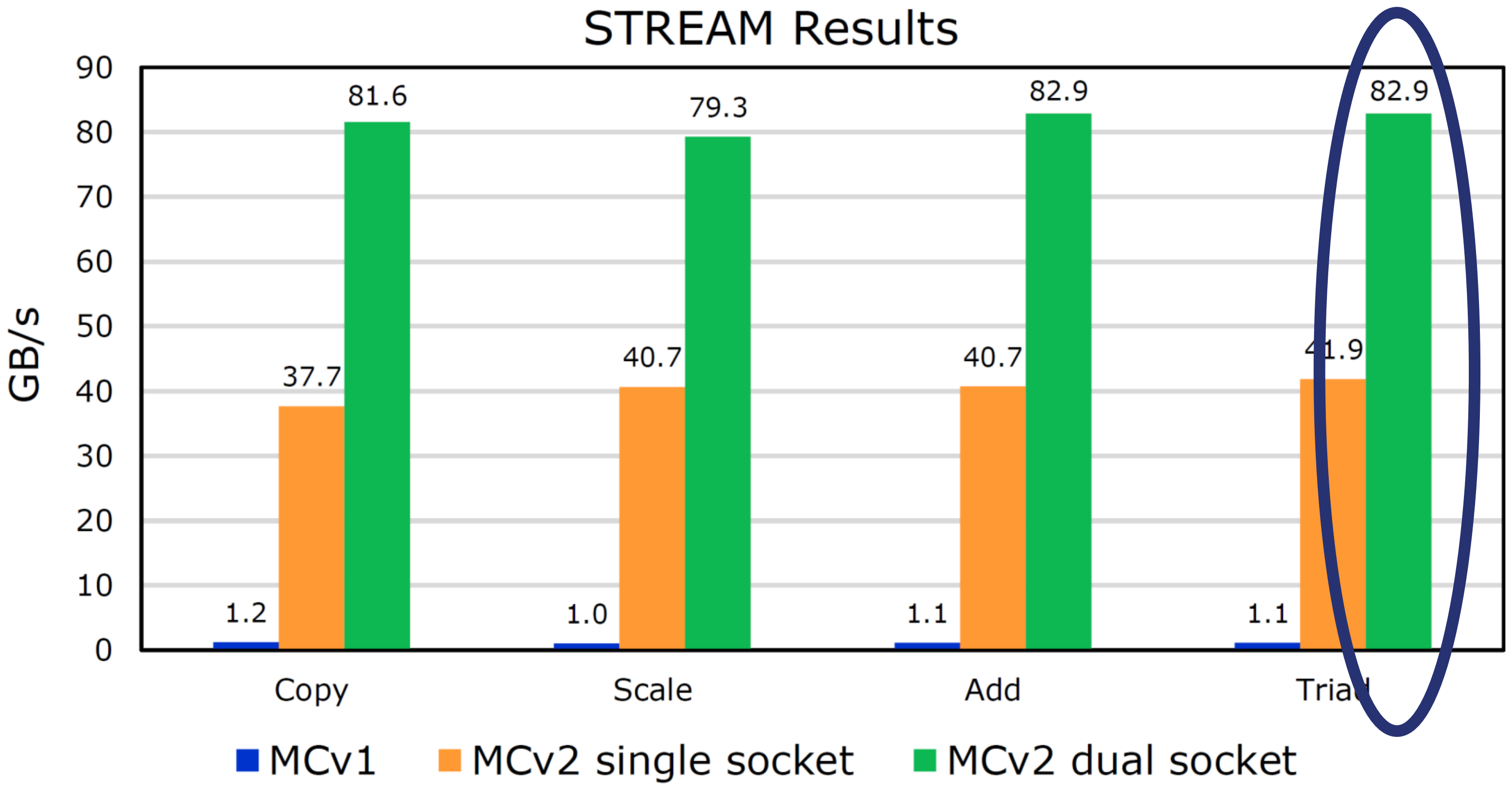
STREAM Results

- **Monte Cimone v1 node:**
  - Max: 1.2 GB/s
- **Milk-V Pioneer Box:**
  - 64 threads
  - 4 memory channels
  - Max: 41.9 GB/s



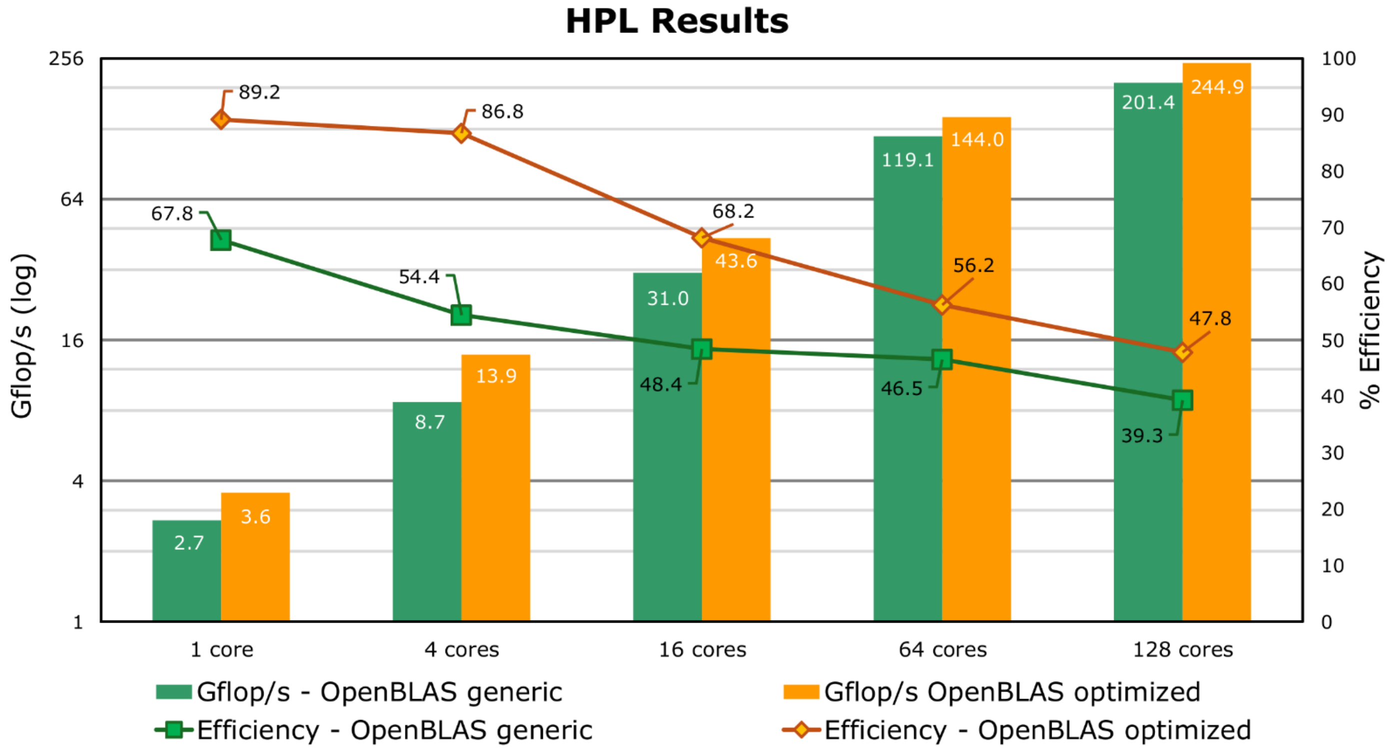
STREAM Results

- **Monte Cimone v1 node:**
  - Max: 1.2 GB/s
- **Milk-V Pioneer Box:**
  - 64 threads
  - 4 memory channels
  - Max: 41.9 GB/s
- **E4 RSV10 dual socket:**
  - 64 threads
  - 8 memory channels
  - Max: 82.9 GB/s



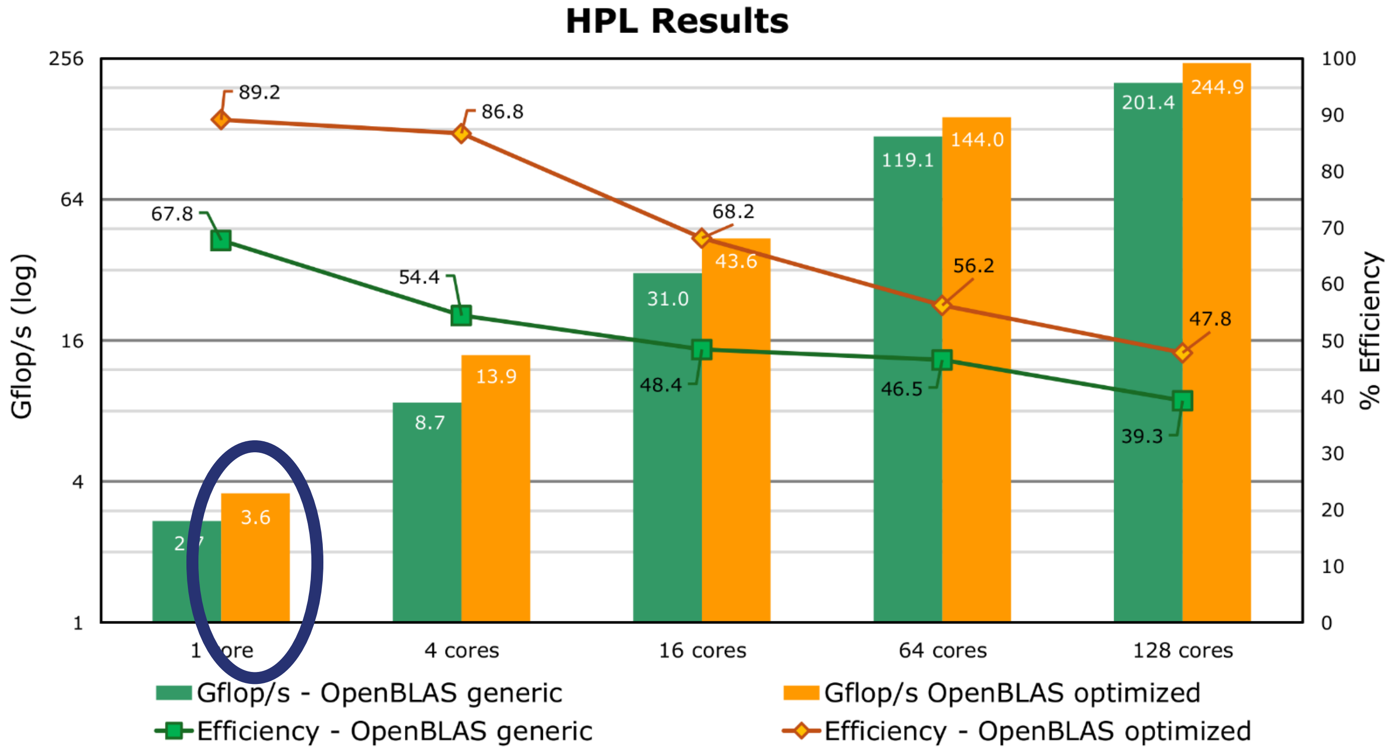
HPL Single Node Results

- **Reference results of SG2042:**
  - HPL + vanilla OpenBLAS
  - HPL + optimized OpenBLAS



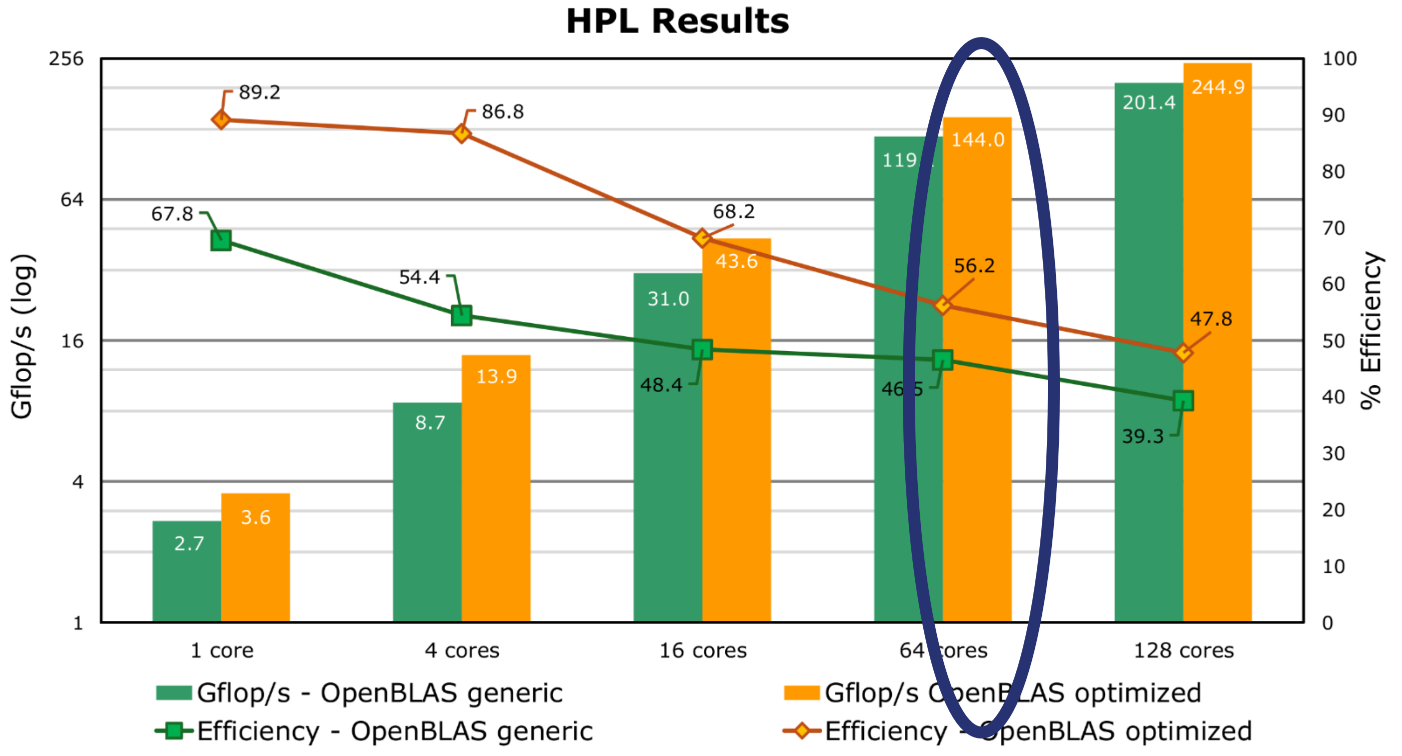
HPL Single Node Results

- **Reference results of SG2042:**
  - HPL + vanilla OpenBLAS
  - HPL + optimized OpenBLAS
- **Single nodes performance:**
  - 1 core:
    - 3.6 Gflop/s



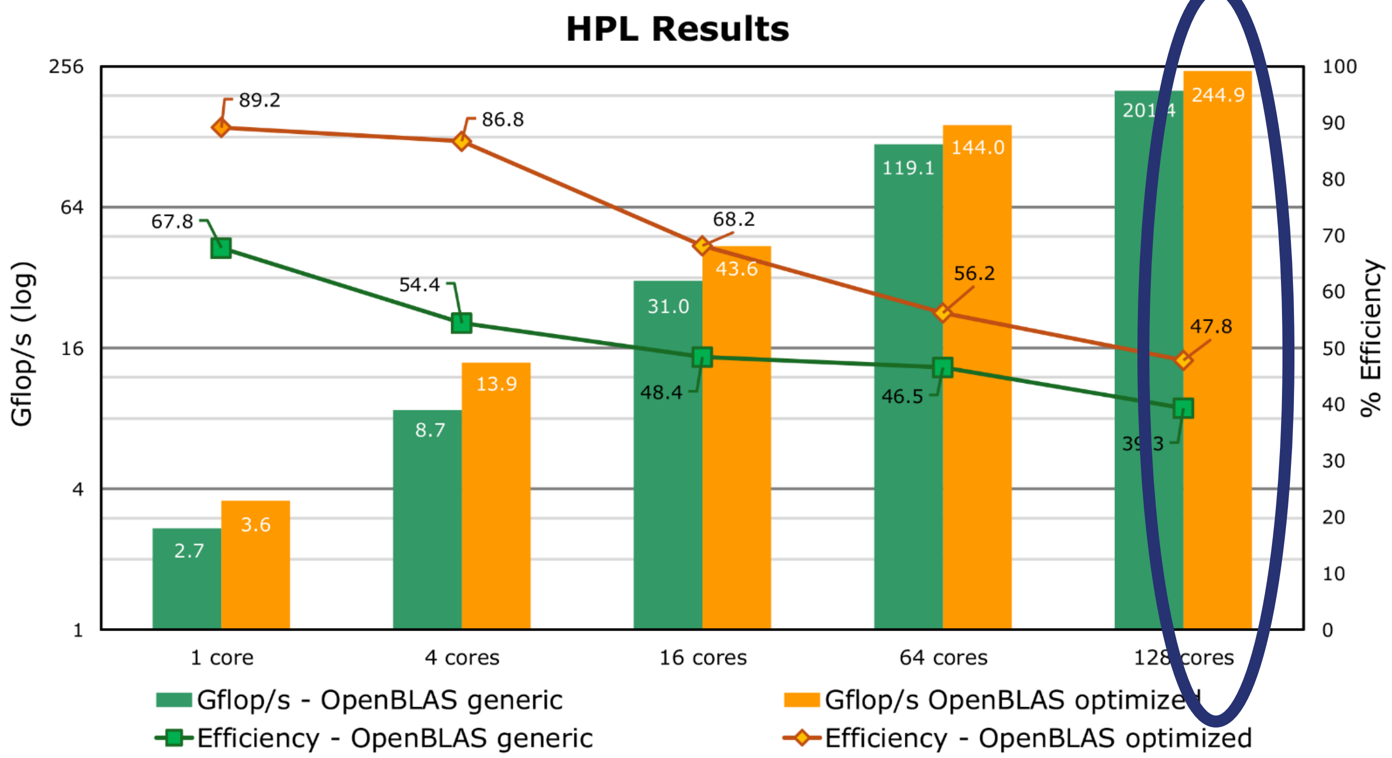
HPL Single Node Results

- **Reference results of SG2042:**
  - HPL + vanilla OpenBLAS
  - HPL + optimized OpenBLAS
- **Single nodes performance:**
  - 1 core:
    - 3.6 Gflop/s
  - 64 cores:
    - 144 Gflop/s



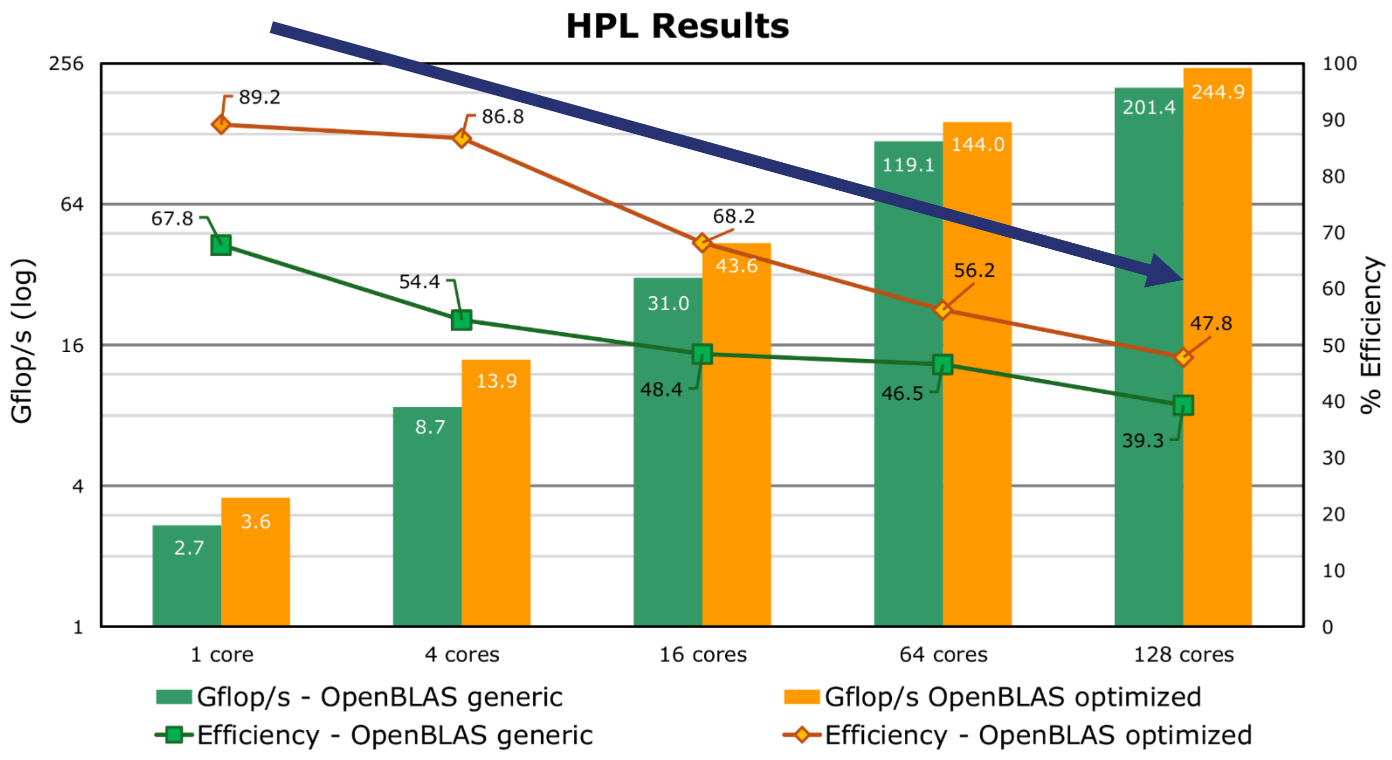
HPL Single Node Results

- **Reference results of SG2042:**
  - HPL + vanilla OpenBLAS
  - HPL + optimized OpenBLAS
- **Single nodes performance:**
  - 1 core:
    - 3.6 Gflop/s
  - 64 cores:
    - 144 Gflop/s
  - 128 cores:
    - 244.9 Gflop/s



HPL Single Node Results

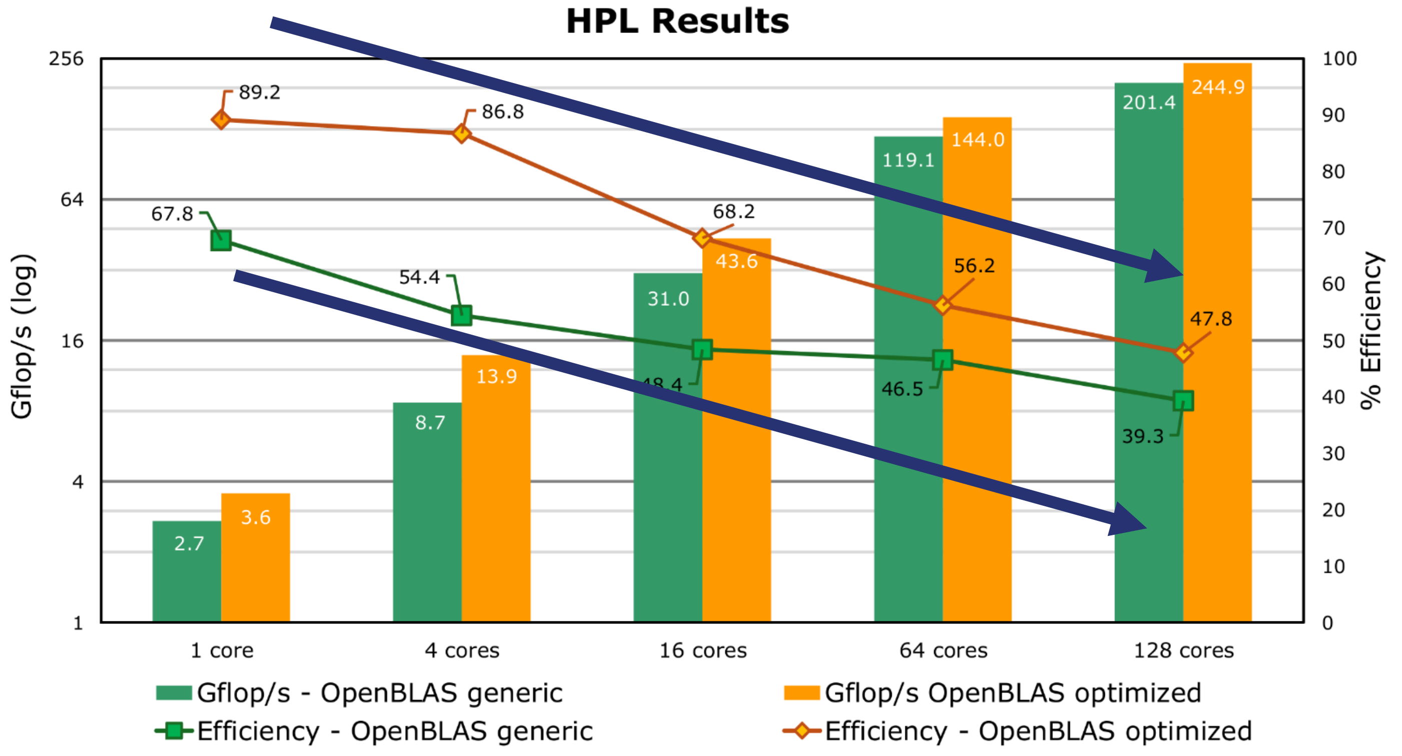
- **Reference results of SG2042:**
  - HPL + vanilla OpenBLAS
  - HPL + optimized OpenBLAS
- **Single nodes performance:**
  - 1 core:
    - 3.6 Gflop/s
  - 64 cores:
    - 144 Gflop/s
  - 128 cores:
    - 244.9 Gflop/s



- Decreasing efficiency with incremental core count

HPL Single Node Results

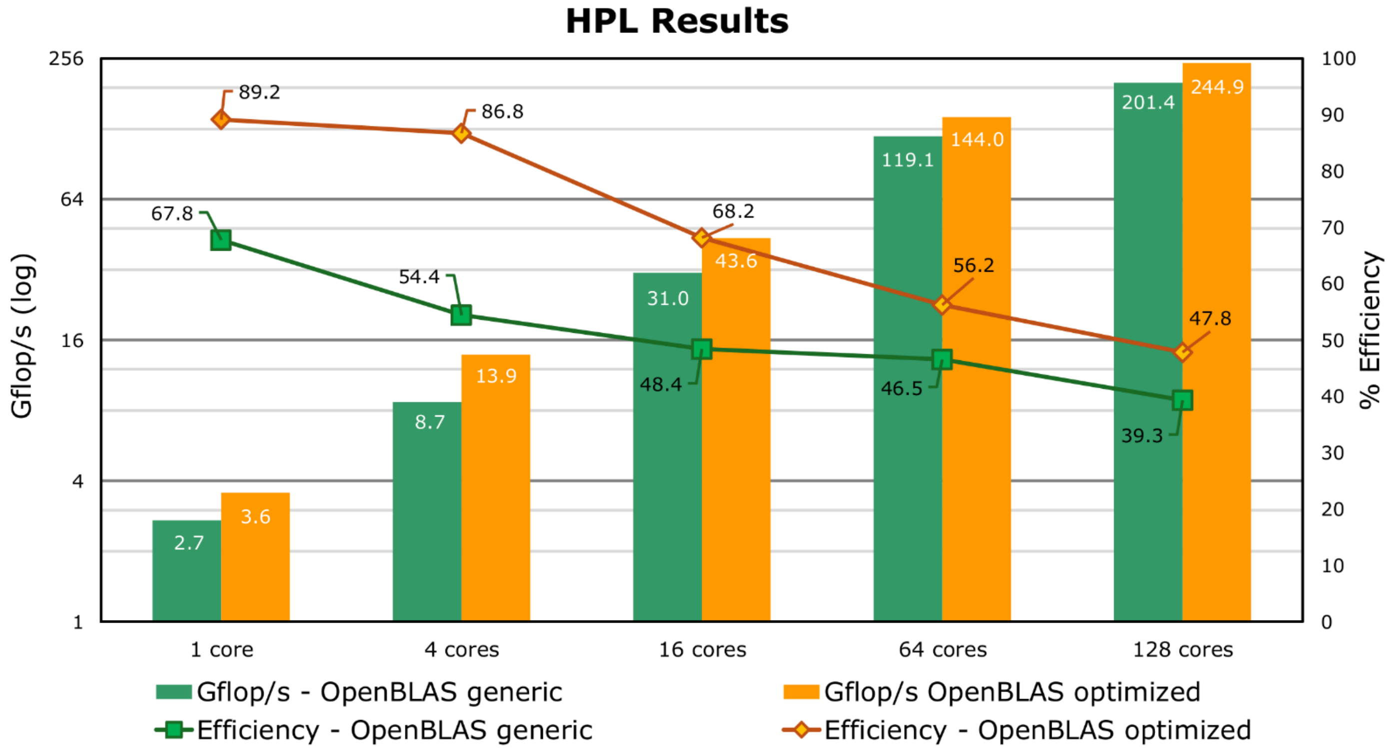
- **Reference results of SG2042:**
  - HPL + vanilla OpenBLAS
  - HPL + optimized OpenBLAS
- **Single nodes performance:**
  - 1 core:
    - 3.6 Gflop/s
  - 64 cores:
    - 144 Gflop/s
  - 128 cores:
    - 244.9 Gflop/s



- Decreasing efficiency with incremental core count
- Optimized OpenBLAS experience the same bottlenecks as vanilla ones

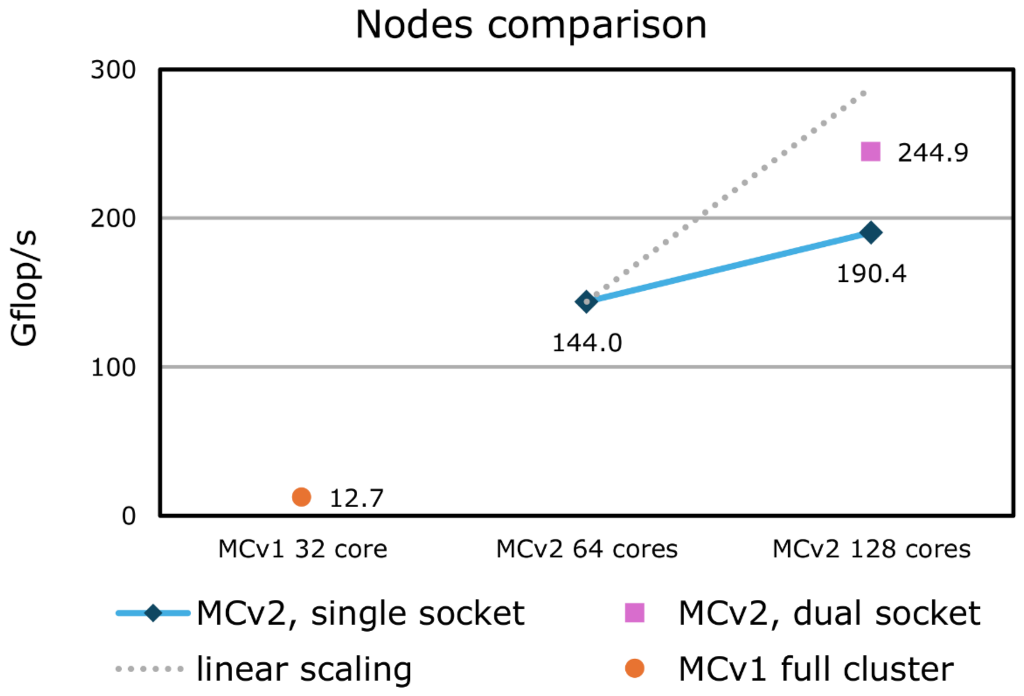
HPL Single Node Results

- **Reference results of SG2042:**
  - HPL + vanilla OpenBLAS
  - HPL + optimized OpenBLAS
- **Single nodes performance:**
  - 1 core:
    - 3.6 Gflop/s
  - 64 cores:
    - 144 Gflop/s
  - 128 cores:
    - 244.9 Gflop/s



- Decreasing efficiency with incremental core count
- Optimized OpenBLAS experience the same bottlenecks as vanilla ones
- The bottleneck was identified as the memory subsystem

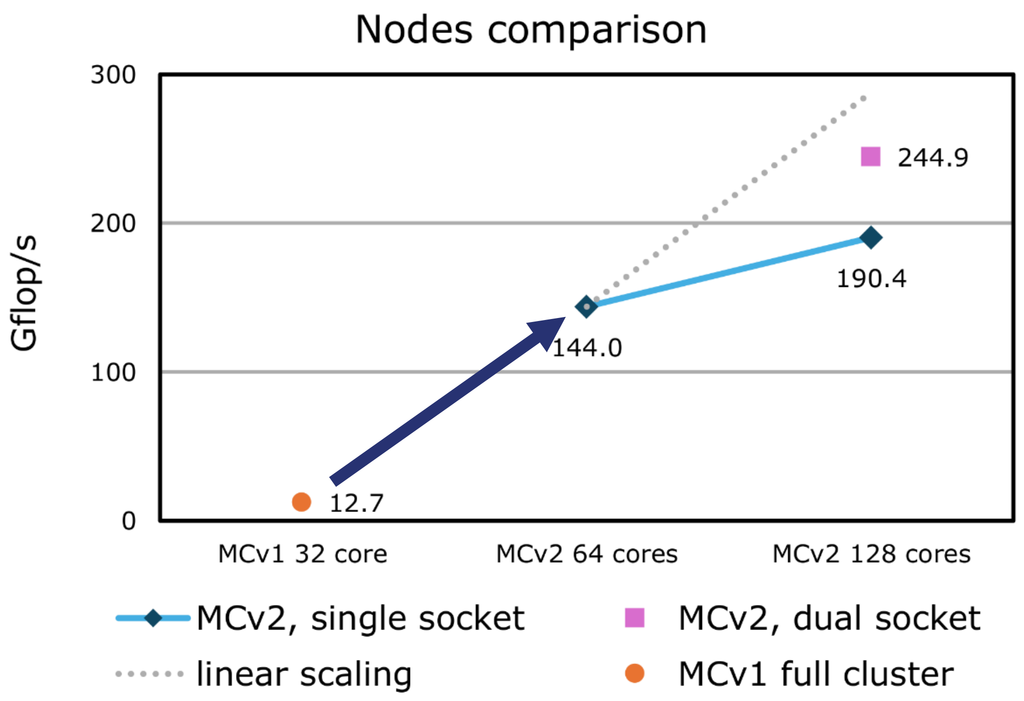
HPL Multi Node Results



HPL Multi Node Results

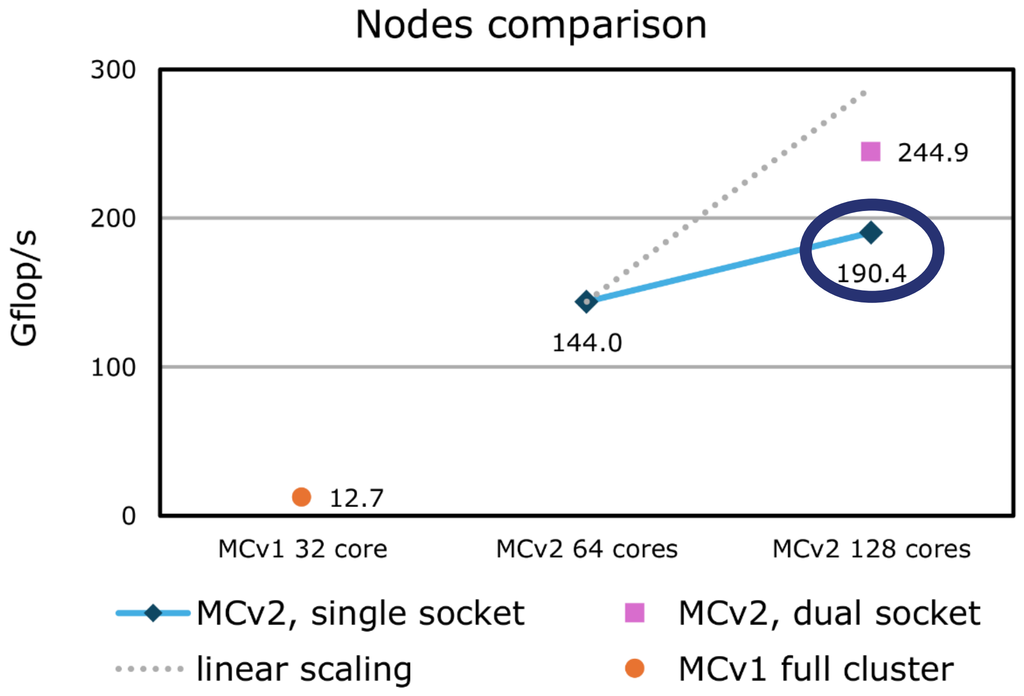
- **MCv1 full cluster:**
  - Linear scaling in multi node runs
  - HPL: 12.65 Gflop/s

- **MCv2 single node, single socket:**
  - 64 cores
  - HPL: 144.0 Gflop/s



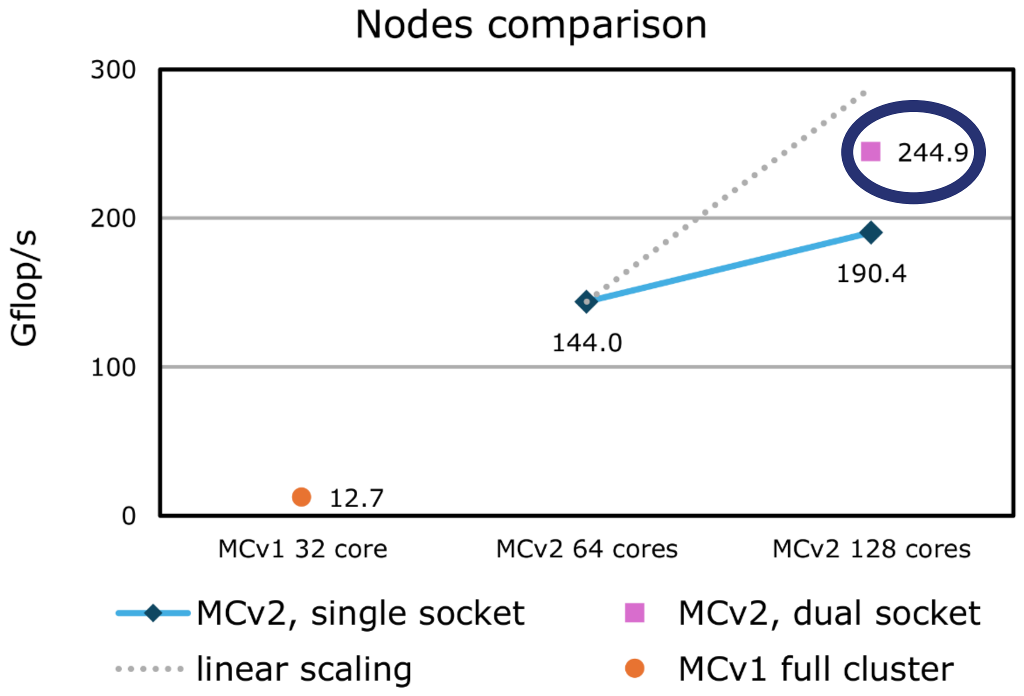
HPL Multi Node Results

- **MCv1 full cluster:**
  - Linear scaling in multi node runs
  - HPL: 12.65 Gflop/s
- **MCv2 single node, single socket:**
  - 64 cores
  - HPL: 144.0 Gflop/s
- **MCv2 dual node, single socket:**
  - 128 cores
  - Ethernet 1Gbit/s network
  - HPL: 190.4 Gflop/s



HPL Multi Node Results

- **MCv1 full cluster:**
  - Linear scaling in multi node runs
  - HPL: 12.65 Gflop/s
- **MCv2 single node, single socket:**
  - 64 cores
  - HPL: 144.0 Gflop/s



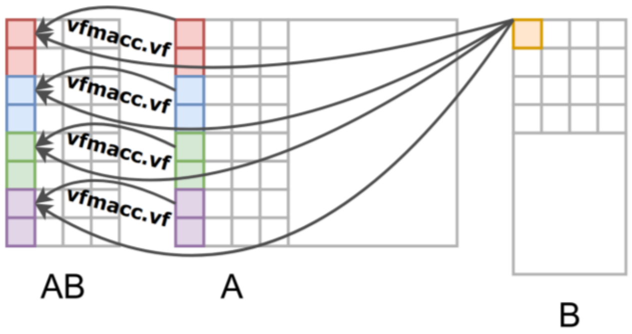
- **MCv2 dual node, single socket:**
  - 128 cores
  - Ethernet 1Gbit/s network
  - HPL: 190.4 Gflop/s
- **MCv2 single node, dual socket:**
  - 128 cores
  - On board communication
  - HPL: 244.9 Gflop/s

BLIS kernels optimization

- BLIS comes with RVV 1.0 generic micro-kernels written in assembly for DGEMM

BLIS kernels optimization

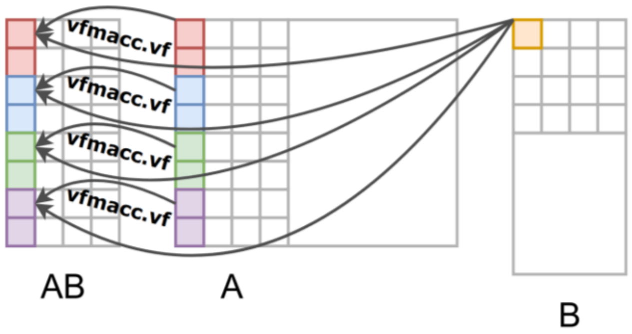
- BLIS comes with RVV 1.0 generic micro-kernels written in assembly for DGEMM
- Steps of our BLIS micro-kernel optimization:
  - Porting from RVV 1.0 to 0.7.1 to be compiled with GCC 14



BLIS vectorization strategy

BLIS kernels optimization

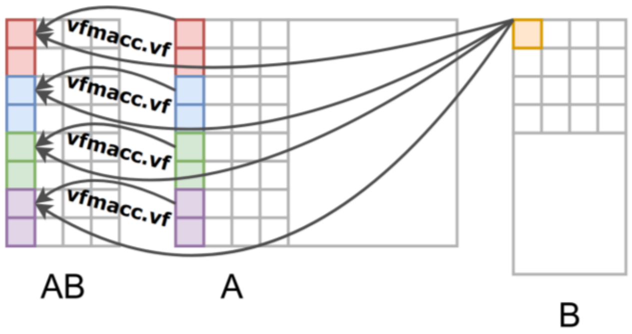
- BLIS comes with RVV 1.0 generic micro-kernels written in assembly for DGEMM
- Steps of our BLIS micro-kernel optimization:
  - Porting from RVV 1.0 to 0.7.1 to be compiled with GCC 14
  - First tests: 1/2 performance of vendor optimized OpenBLAS



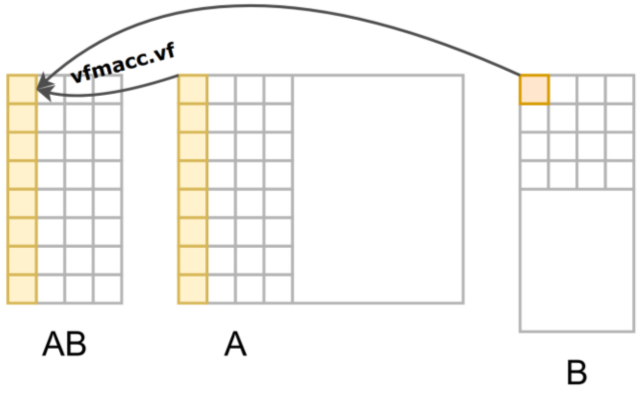
BLIS vectorization strategy

BLIS kernels optimization

- BLIS comes with RVV 1.0 generic micro-kernels written in assembly for DGEMM
- Steps of our BLIS micro-kernel optimization:
  - Porting from RVV 1.0 to 0.7.1 to be compiled with GCC 14
  - First tests: 1/2 performance of vendor optimized OpenBLAS
  - Optimization using register grouping



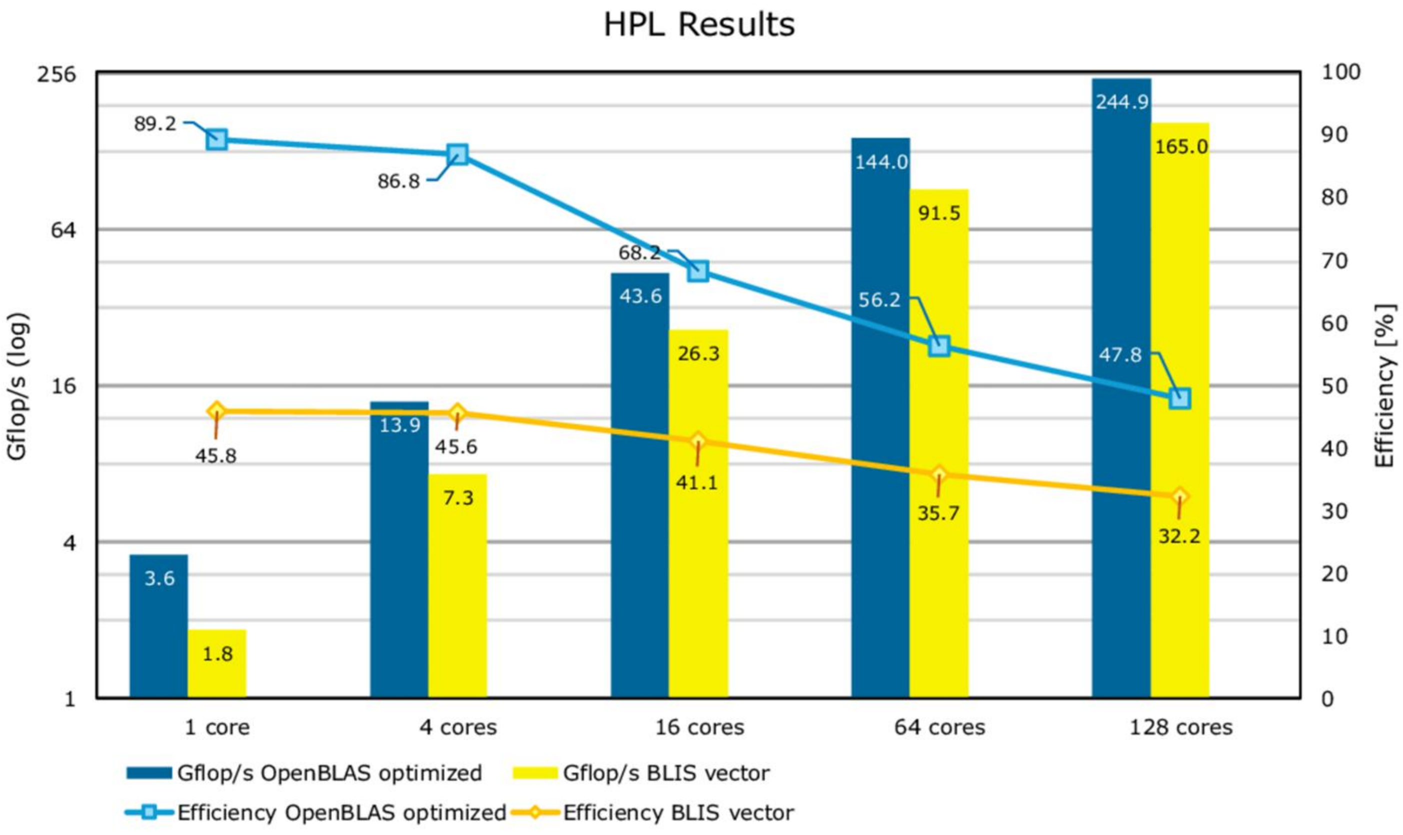
BLIS vectorization strategy



Proposed optimization strategy

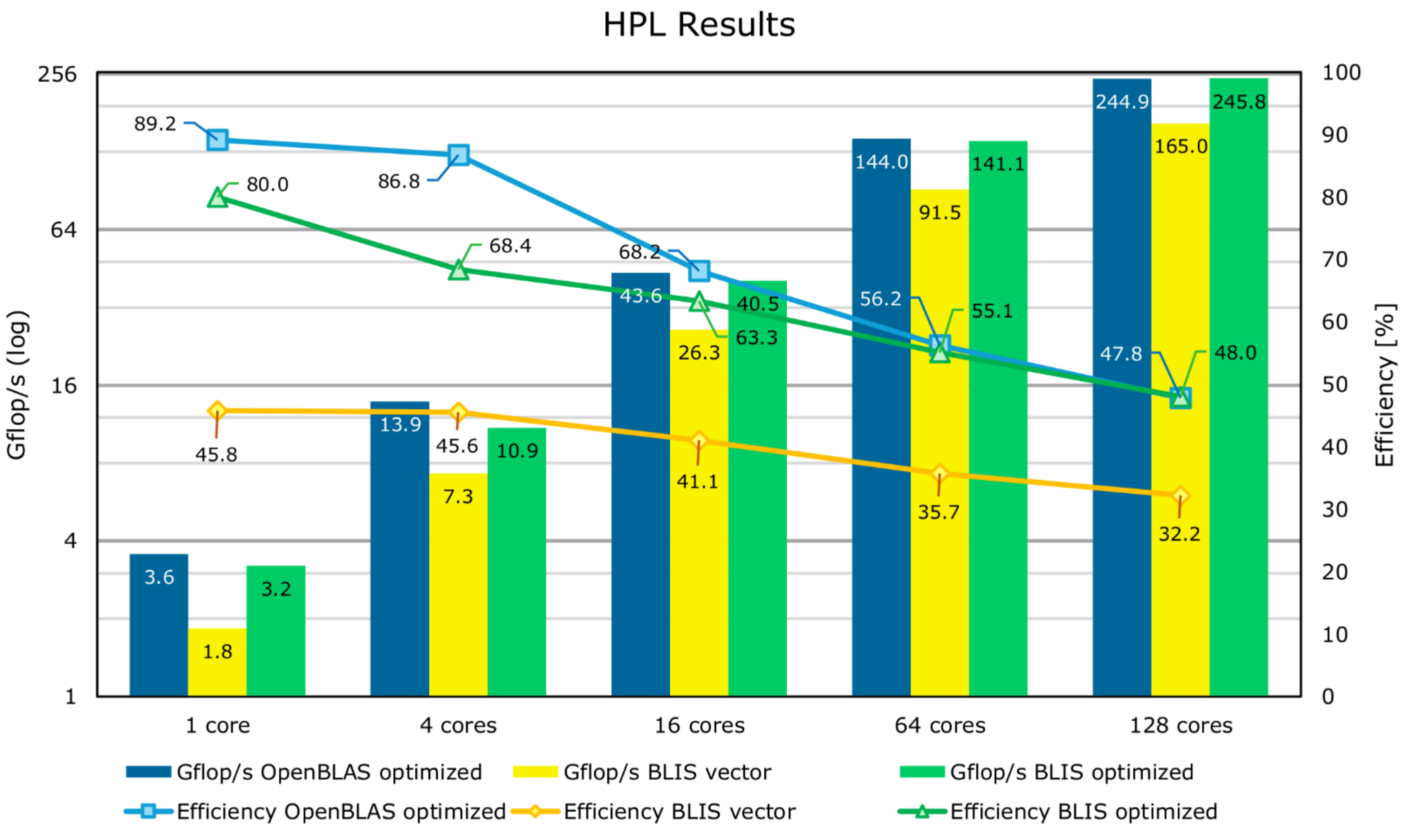
HPL BLIS Results

- The available vectorized BLIS micro-kernel performed poorly



HPL BLIS Results

- The available vectorized BLIS micro-kernel performed poorly
- **Boost in performance post optimization:**
  - 49% improvement w.r.t. original vectorized BLIS (dual socket and 128 core runs)
  - Performance on par with vendor OpenBLAS
  - Effectiveness of register grouping



Conclusions

- **MCv1 vs. MCv2: rapidly evolving RISC-V ecosystem**
  - In two years a performance improvement obtained in average in eight years (Top500)

Conclusions

- **MCv1 vs. MCv2: rapidly evolving RISC-V ecosystem**
  - In two years a performance improvement obtained in average in eight years (Top500)
- **BLIS library optimization**
  - +49% performance vs. original version
  - On par with vendor optimized BLAS library
  - Committed to the BLIS public repository

Conclusions

- **MCv1 vs. MCv2: rapidly evolving RISC-V ecosystem**
  - In two years a performance improvement obtained in average in eight years (Top500)
- **BLIS library optimization**
  - +49% performance vs. original version
  - On par with vendor optimized BLAS library
  - Committed to the BLIS public repository

This activity has been supported by the HE EU Graph-Massivizer (g.a. 101093202), DECICE (g.a. 101092582), and DARE (g.a. 101143421) projects, as well as the Italian Research Center on High Performance Computing, Big Data, and Quantum Computing.