EUROfusion

Benchmarks and validation of the Pitagora HPC system

Serhiy Mochalskyy

Sixth IFERC workshop on the usage of GPU based system for fusion applications

June 24, 2025

Advanced Computing Hub Garching
Max-Planck-Institut für Plasmaphysik
Boltzmannstr. 2, D-85748 Garching, Germany

Pitagora CPU and GPU partitions

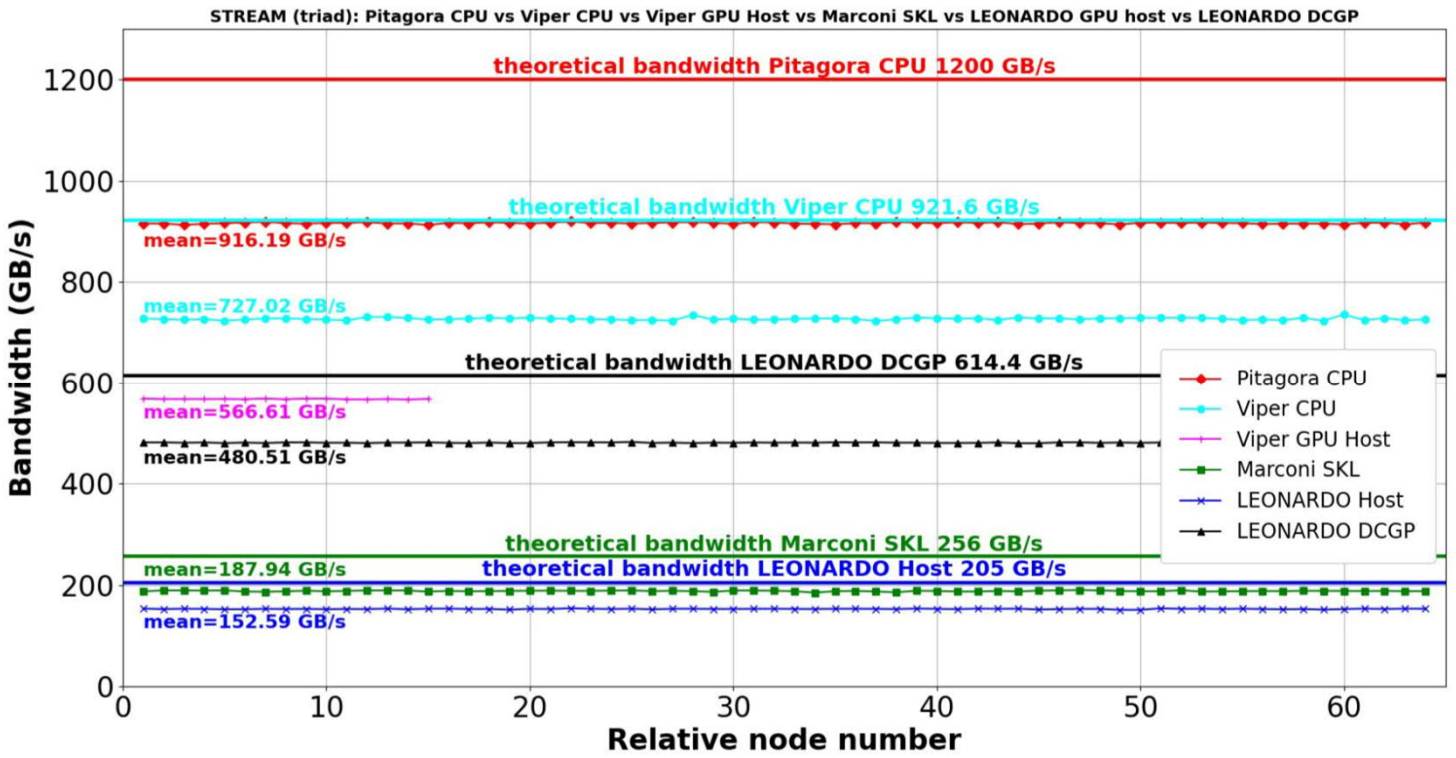
GPU

- 7 racks
- 28.2 PFlops (Rmax)
- 168 Compute nodes
- 2x Intel Emerald Rapids 32c
- 512 GB DDR5 6400 MT/s
- 4x NVIDIA H100 SXM 94GB HBM2e
- 2.3x performance over A100
- 4x NDR200 adapters (200 Gb/s each)

CPU

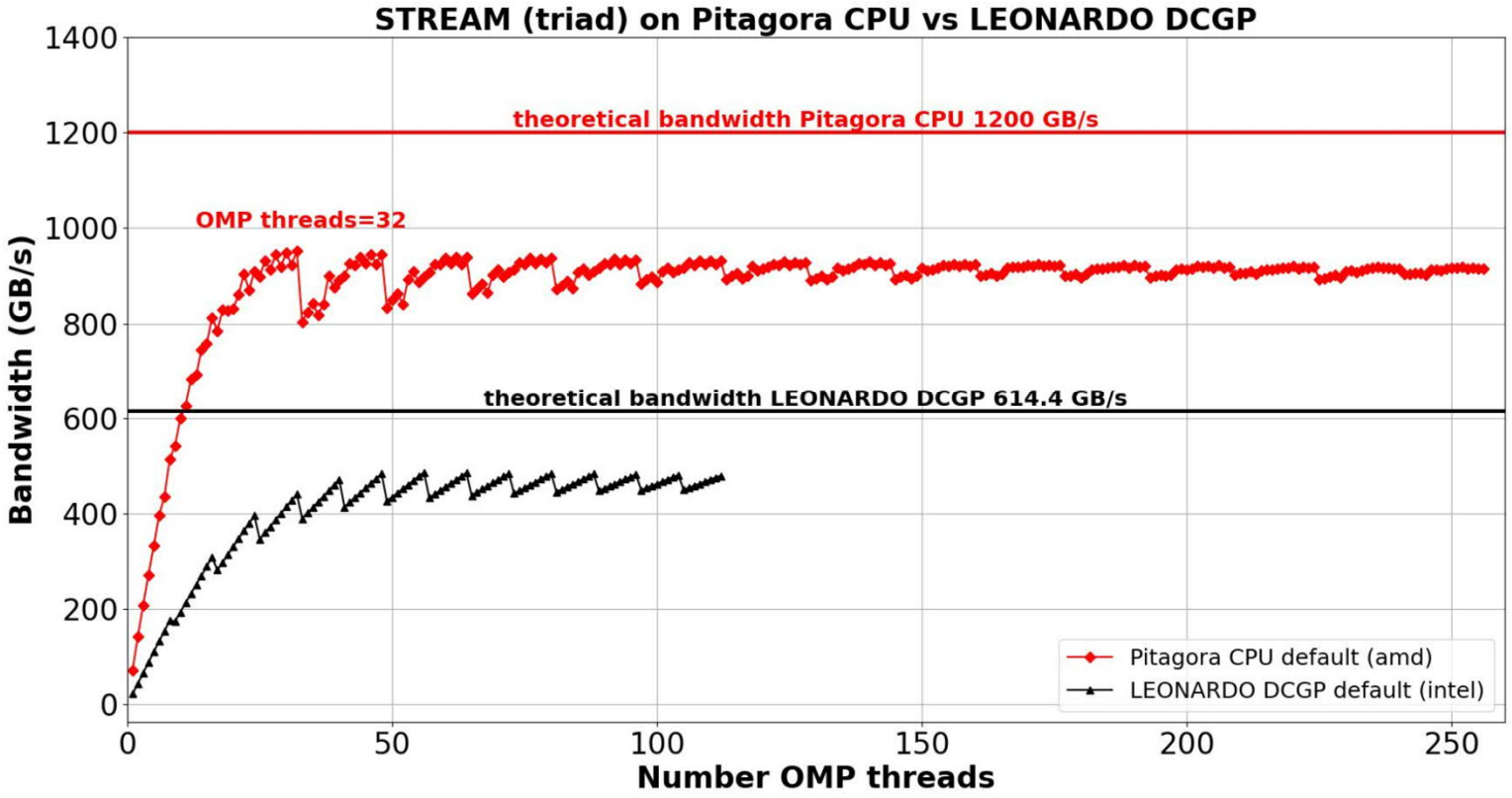
- 14 racks
- 17 PFlops (Rmax)
- 1008 Compute nodes
- 2x AMD Turin 128c (Zen5) 2.3 GHz
- 768 GB DDR5 6400 MT/s

Stream benchmark on Pitagora CPU



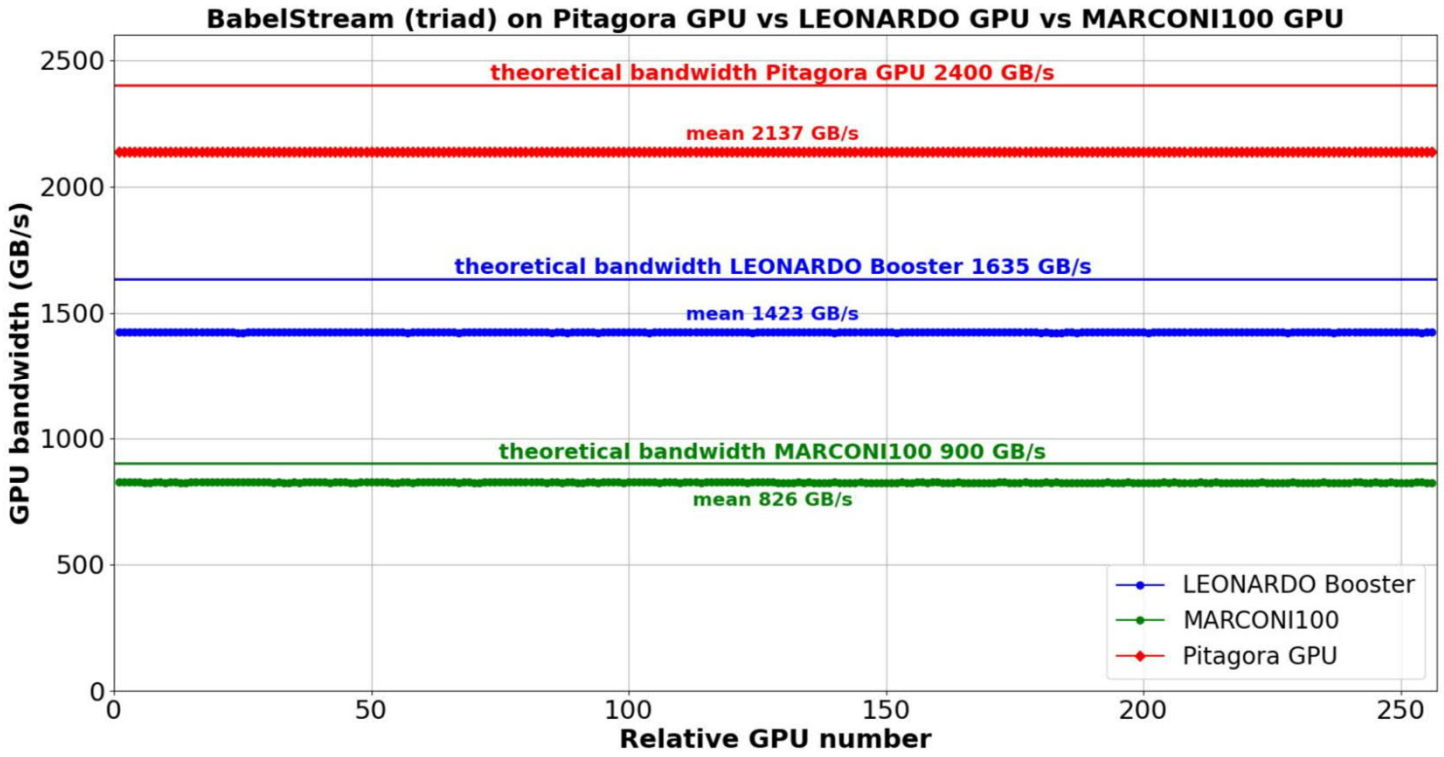
- **Pitagora CPU (mean): 916 GB/s from 1200 GB/s theoretical value (76%).**
- **Viper-CPU (mean): 727 GB/s from 921.6 GB/s theoretical value (79%).**
- **Viper-GPU Host (mean): 567 GB/s.**
- **LEONARDO DCGP (mean): 481 GB/s from 614.4 GB/s theoretical value (78%).**
- **MARCONI SKL (mean): 188 GB/s from 255.94 GB/s theoretical value (73%).**
- **LEONARDO Booster Host (mean): 153 GB/s from 205 GB/s theoretical value (75%).**

Pitagora CPU: STREAM on a single node



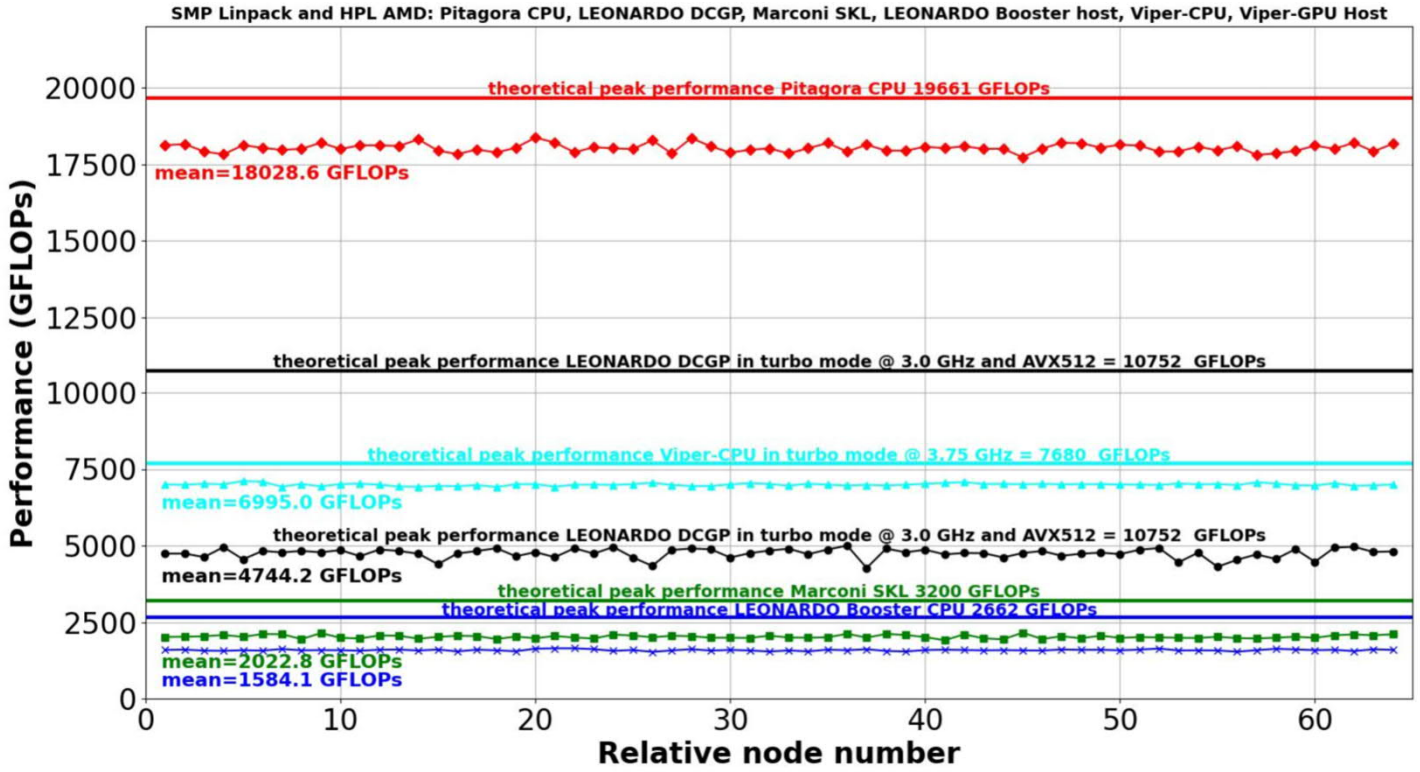
- **Pitagora CPU (mean): 916 GB/s from 1200 GB/s theoretical value (76%).**
- **Pitagora CPU: saturation with 32 OpenMP threads.**
- **LEONARDO DCGP (mean): 481 GB/s from 614.4 GB/s theoretical value (78%).**
- **LEONARDO DCGP: saturation with 48 OpenMP threads.**

BabelStream benchmark on Pitagora GPU



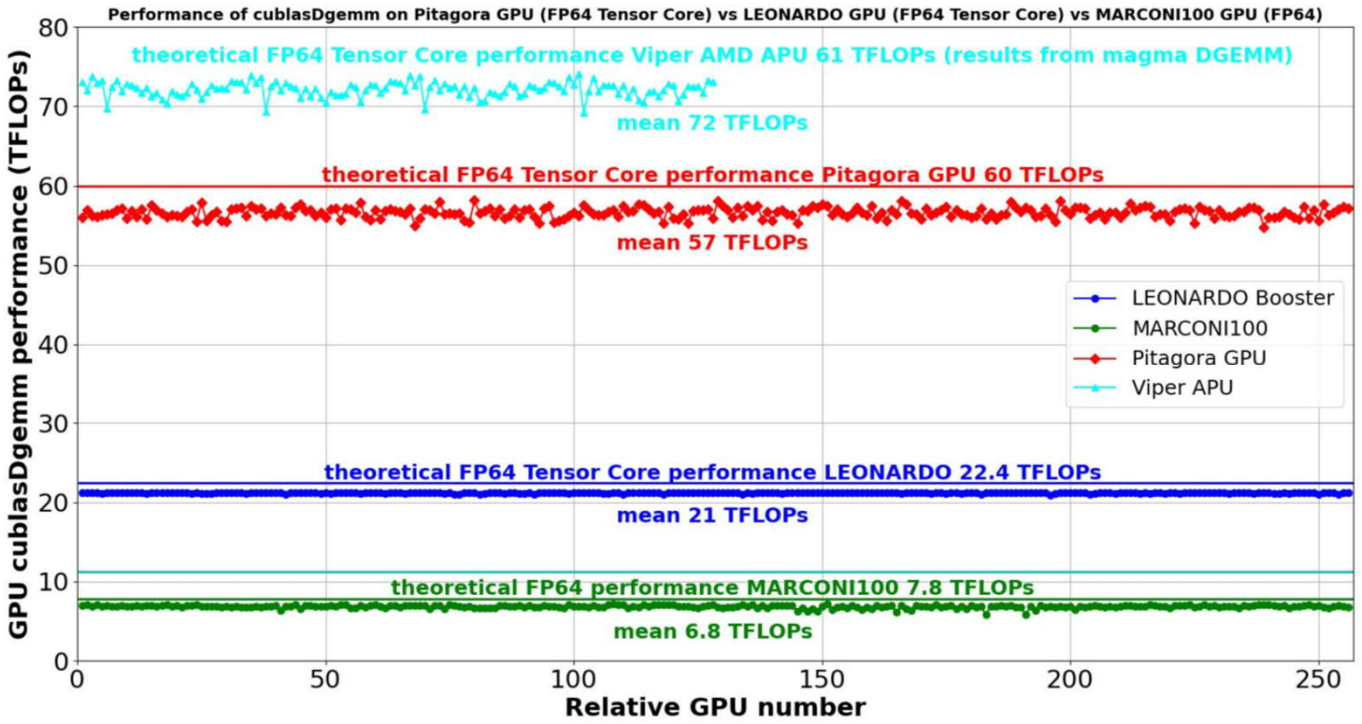
- All GPUs provide high, stable and symmetric bandwidth close to the theoretical value.
- No difference between GPUs on different nodes or GPUs inside one node.
- Pitagora GPU (mean): 2137 GB/s from 2400 GB/s theoretical value (89%).
- LEONARDO Booster (mean): 1423.5 GB/s from 1635 GB/s theoretical value (87%).
- MARCONI100 (mean): 845 GB/s from 900 GB/s theoretical value (94%).

Pitagora CPU: performance stability test



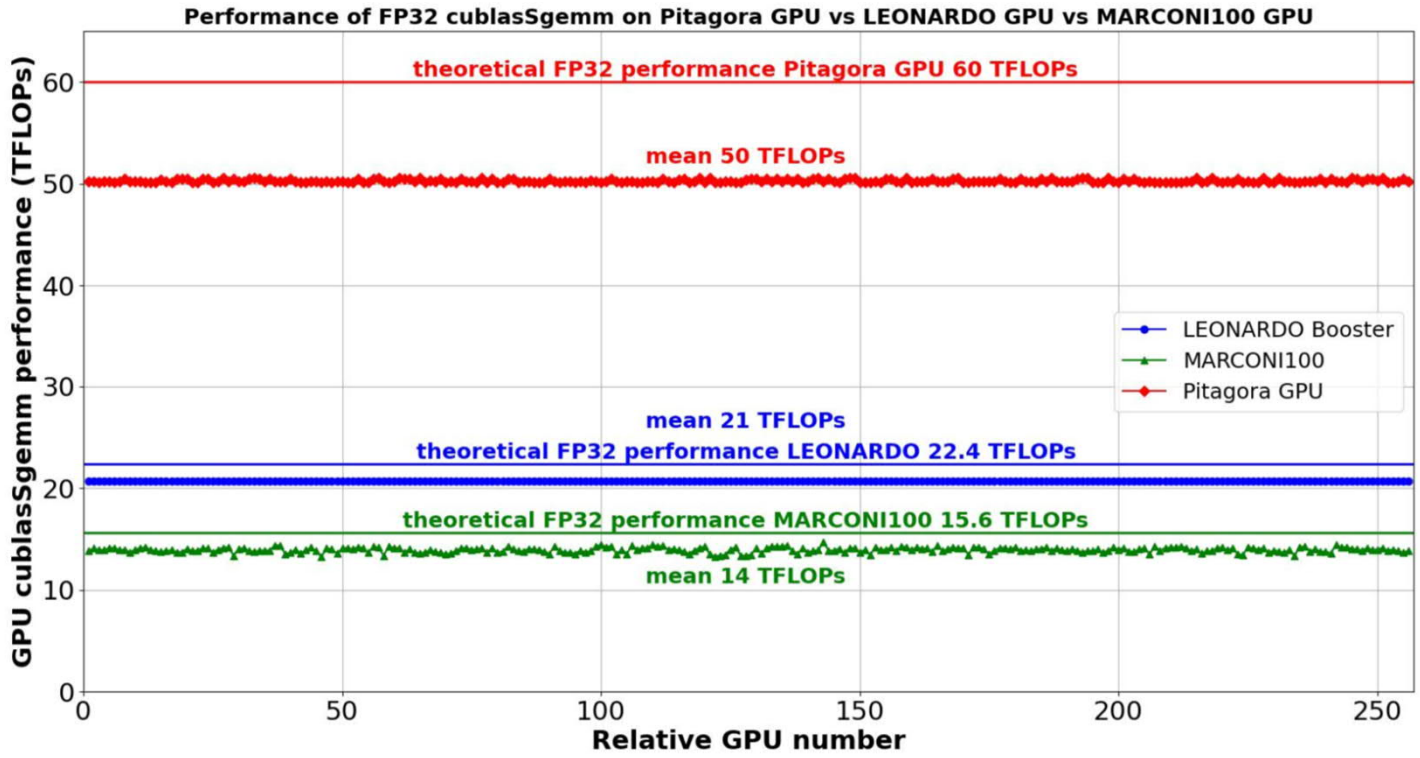
- All nodes provide high, stable and symmetric performance close to the theoretical value.
- Pitagora CPU (mean): 18029 GFLOPs from 19661 GFLOPs theoretical value (92%).
- Viper-CPU (mean): 6995 GFLOPs from 7680 GFLOPs theoretical value (91%).
- LEONARDO DCGP (mean): 4744 GFLOPs from 10752 GFLOPs theoretical value (44%).
- MARCONI100 Host (mean): 2023 GFLOPs from 3200 GFLOPs theoretical value (63%).
- LEONARDO Booster Host (mean): 1584 GFLOPs from 2662 GFLOPs theoretical value (60%).

DGEMM (cublasDgemm) benchmark on Pitagora GPU



- All GPUs provide high, stable and symmetric performance close to the theoretical value.
- No difference between GPUs on different nodes or GPUs inside one node.
- Viper-GPU FP64 (mean): 72 TFLOPs per APU from 61 TFLOPs theoretical value (118%).
- Pitagora-GPU FP64 (mean): 57 TFLOPs per GPU from 60 TFLOPs theoretical value (95%).
- LEONARDO FP64 (mean): 21 TFLOPs per GPU from 22.4 TFLOPs theoretical value (94%).
- MARCONI100 FP64 (mean): 6.8 TFLOPs per GPU from 7.8 TFLOPs theoretical value (87%).

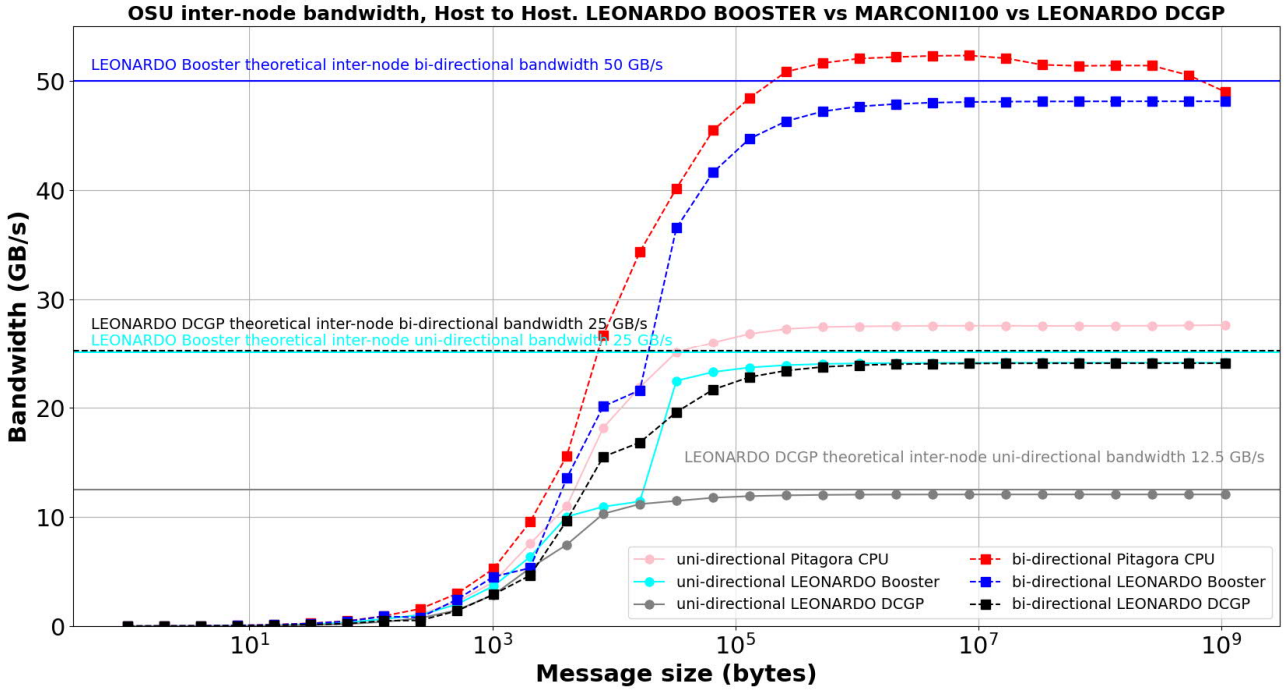
SGEMM (cublasSgemm) benchmark on Pitagora GPU



- All GPUs provide high, stable and symmetric performance close to the theoretical value.
- No difference between GPUs on different nodes or GPUs inside one node.
- Pitagora-GPU FP32 (mean): 50 TFLOPs per GPU from 60 TFLOPs theoretical value (83%).
- LEONARDO FP32 (mean): 21 TFLOPs per GPU from 22.4 TFLOPs theoretical value (94%).
- MARCONI100 FP32 (mean): 14 TFLOPs per GPU from 15.6 TFLOPs theoretical value (90%).

Pitagora-CPU inter-node network bandwidth

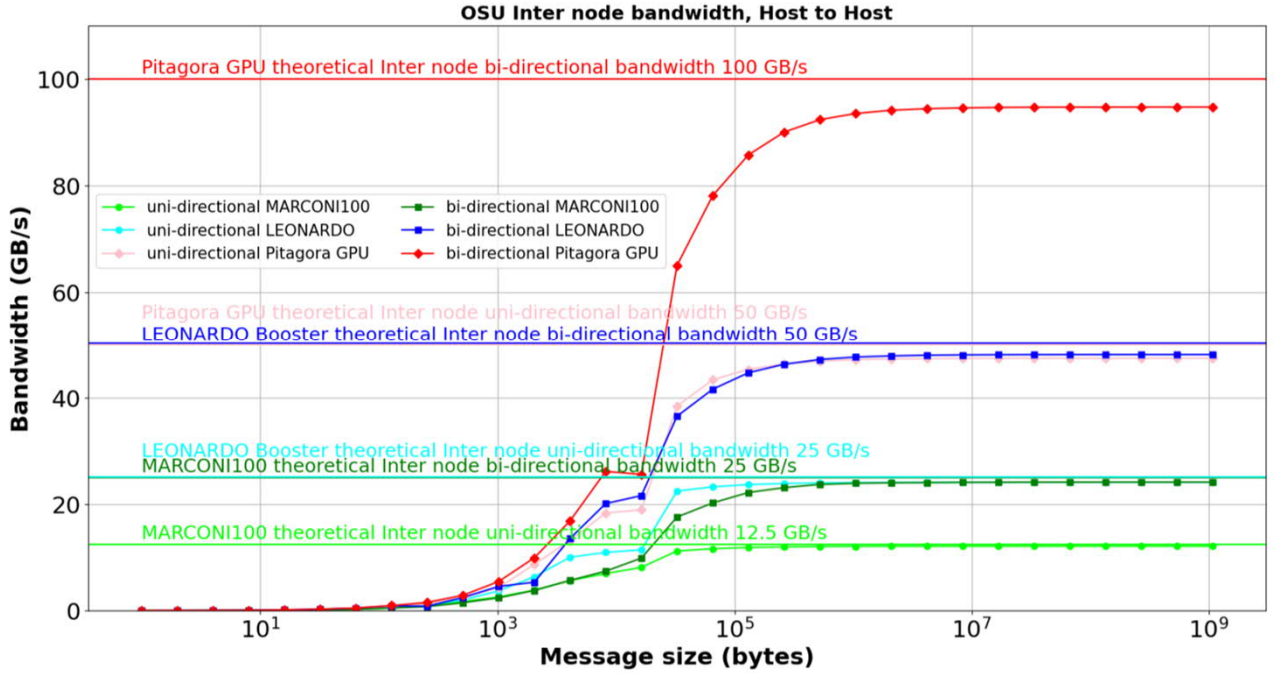
using osu_bw and osu_bibw benchmarks from OSU microbenchmark



- **Stable and high bandwidth for uni- and bi-directional data transfer.**
- **Pitagora-CPU: bi-directional bandwidth ~52 GB/s.**
- **Pitagora-CPU: uni-directional bandwidth ~27 GB/s.**
- **LEONARDO: bi-directional bandwidth ~49 GB/s from 50 GB/s of the theoretical value (98%).**
- **LEONARDO: uni-directional bandwidth ~24 GB/s from 25 GB/s of the theoretical value (96%).**
- **LEONARDO DCGP: bi-directional bandwidth ~24.2 GB/s from 25 GB/s of the theoretical value (97%).**
- **LEONARDO DCGP : uni-directional bandwidth ~12.1 GB/s from 12.5 GB/s of the theoretical value (99%).**

Pitagora-GPU inter-node network bandwidth

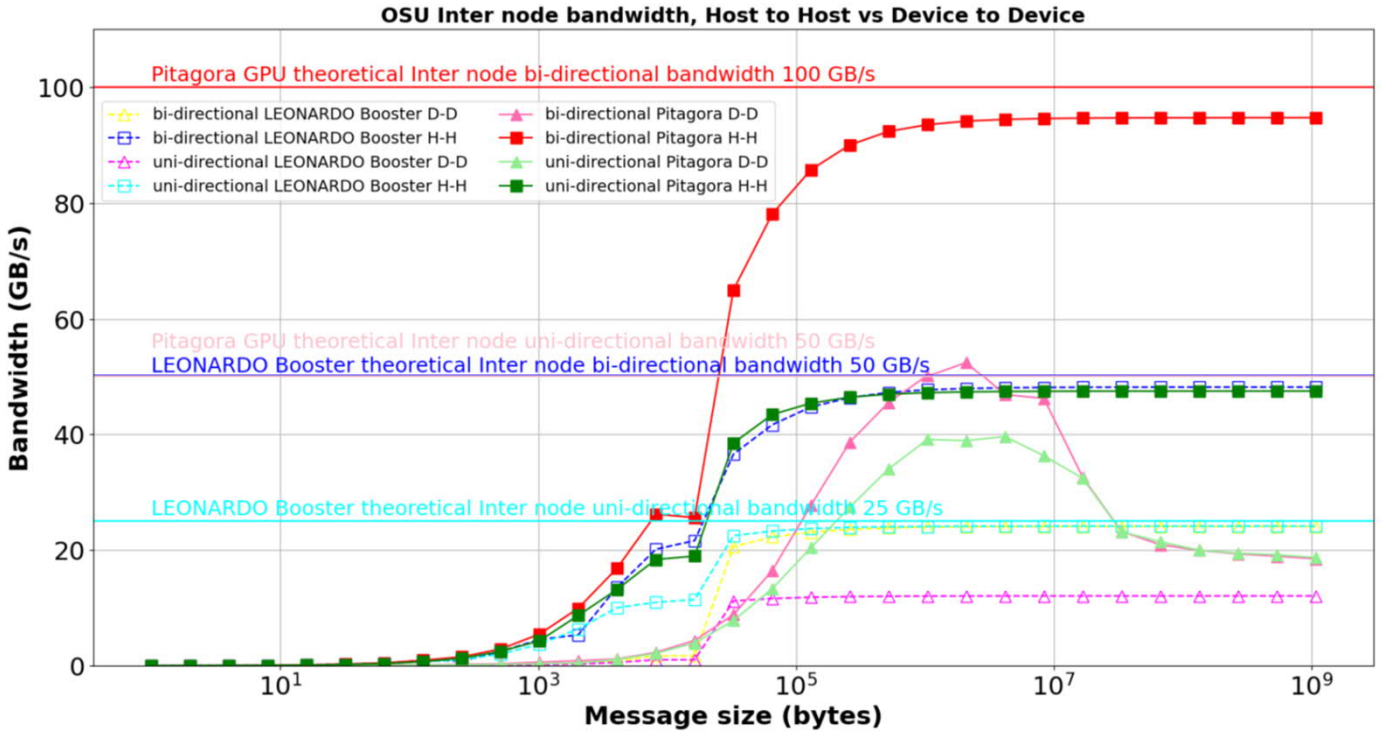
using osu_bw and osu_bibw benchmarks from OSU microbenchmark



- **Stable and high bandwidth for uni- and bi-directional data transfer.**
- **Pitagora-GPU: bi-directional bandwidth ~97 GB/s from 100 GB/s of the theoretical value (97%).**
- **Pitagora-GPU: uni-directional bandwidth ~49 GB/s from 50 GB/s of the theoretical value (98%).**
- **LEONARDO Booster: bi-directional bandwidth ~49 GB/s from 50 GB/s of the theoretical value (98%).**
- **LEONARDO Booster: uni-directional bandwidth ~24 GB/s from 25 GB/s of the theoretical value (96%).**
- **MARCONI100: bi-directional bandwidth ~24.2 GB/s from 25 GB/s of the theoretical value (97%).**
- **MARCONI100: uni-directional bandwidth ~12.1 GB/s from 12.5 GB/s of the theoretical value (99%).**

Pitagora-GPU inter-node network bandwidth

using osu_bw and osu_bibw benchmarks from OSU microbenchmark

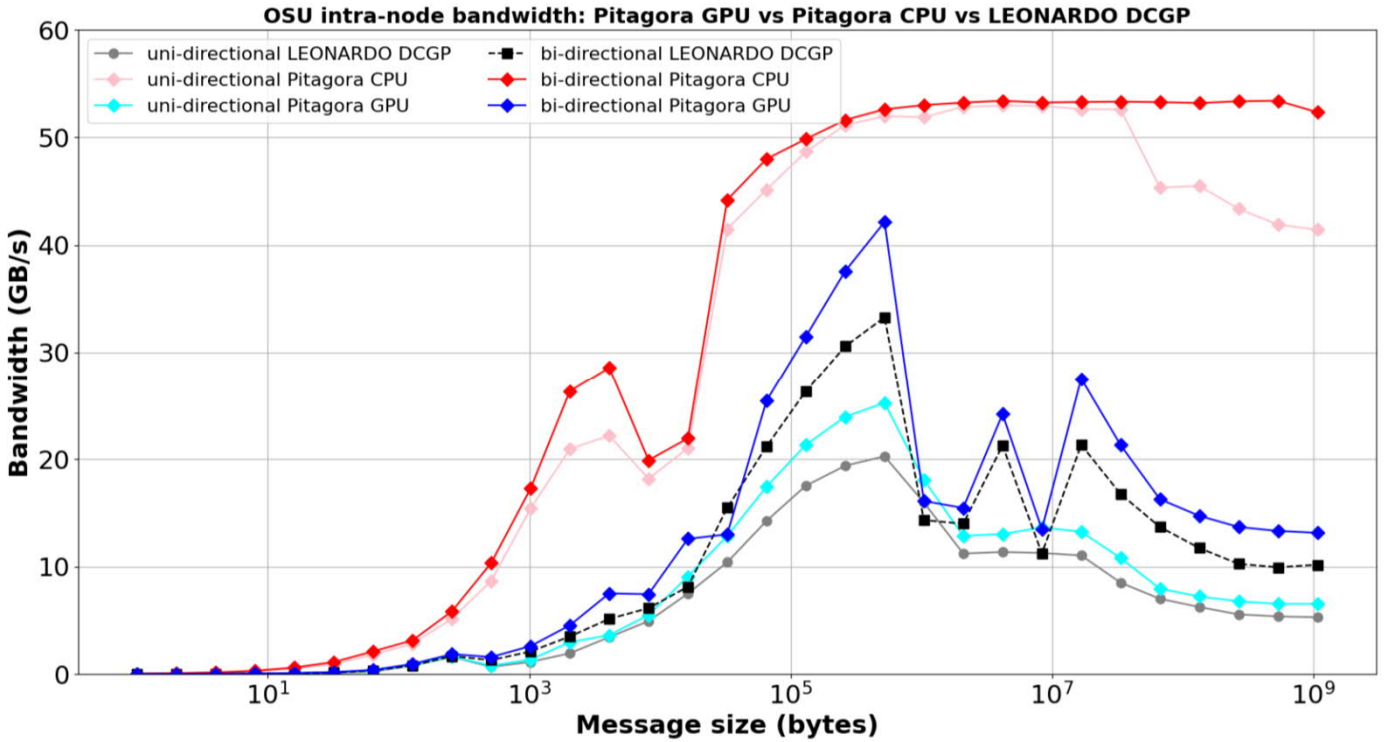


- **Pitagora-GPU H-H: bi-directional bandwidth ~97 GB/s. Pitagora-GPU D-D: bi-directional bandwidth ~52 GB/s.**
- **Pitagora-GPU H-H: uni-directional bandwidth ~49 GB/s. Pitagora-GPU D-D: uni-directional bandwidth ~40 GB/s.**
- **LEONARDO Booster H-H: bi-directional bandwidth ~49 GB/s. LEONARDO Booster D-D: bi-directional bandwidth ~24 GB/s.**
- **LEONARDO Booster H-H: uni-directional bandwidth ~24 GB/s. LEONARDO Booster D-D: uni-directional bandwidth ~12 GB/s.**
- **Pitagora-GPU D-D transfer, for both uni- and bi-directional data, behaves as if the transfer is performed through the CPU.**

Pitagora-GPU: intra-node bandwidth

Host to Host connection is three UPI links 3:

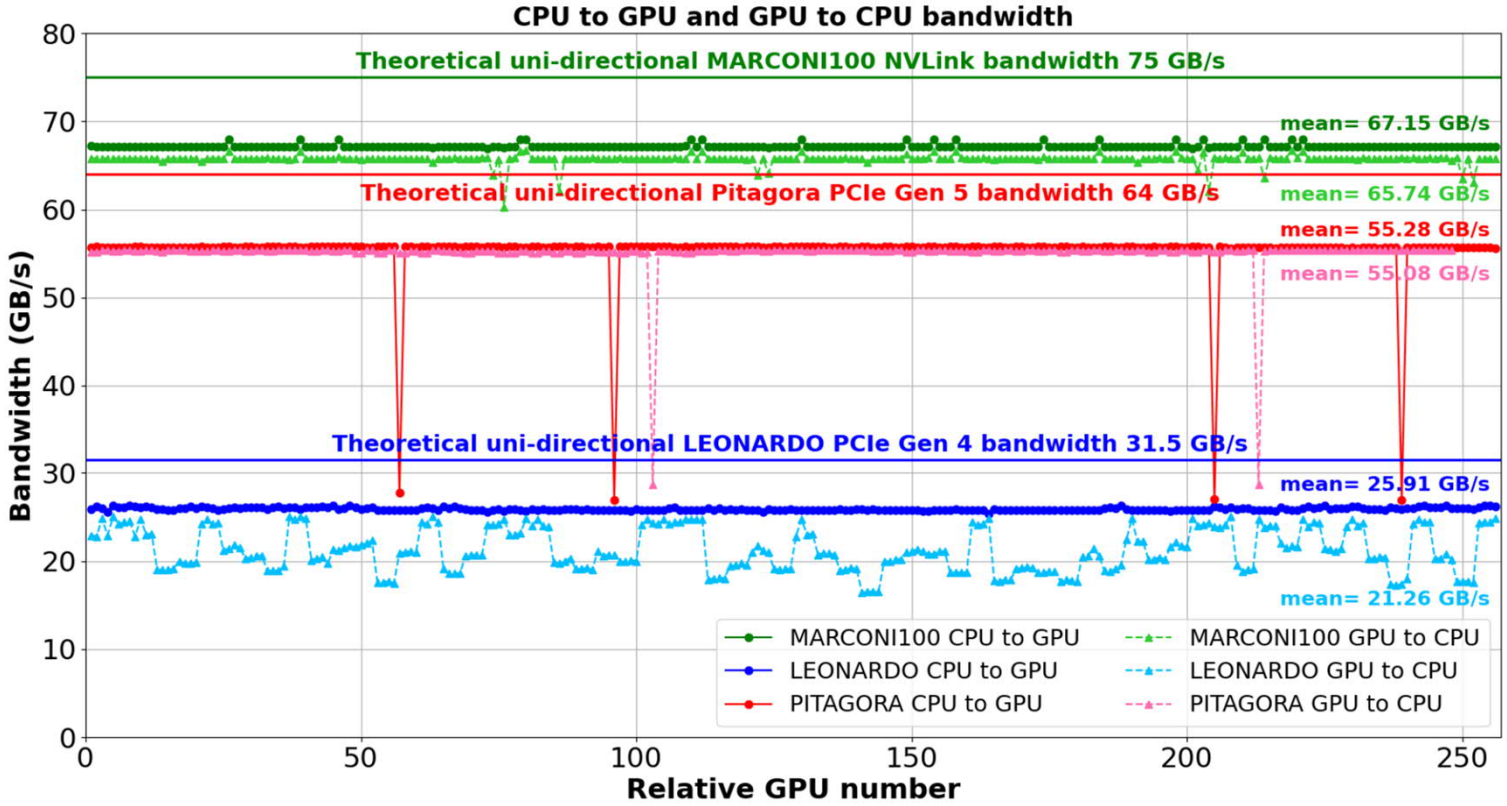
- **Bandwidth is still unclear from 62.4 GB/s to 120 GB/s uni-directional**



- **Bi-directional bandwidth is the same as uni-directional (~52 GB/s) — possible issue.**
- **The theoretical bandwidth is unclear.**

Pitagora-GPU: Host to Device connection

Host to Device connection is PCIe Gen 5:
- **128 GB/s bi-directional bandwidth**
- **64 GB/s uni-directional bandwidth**

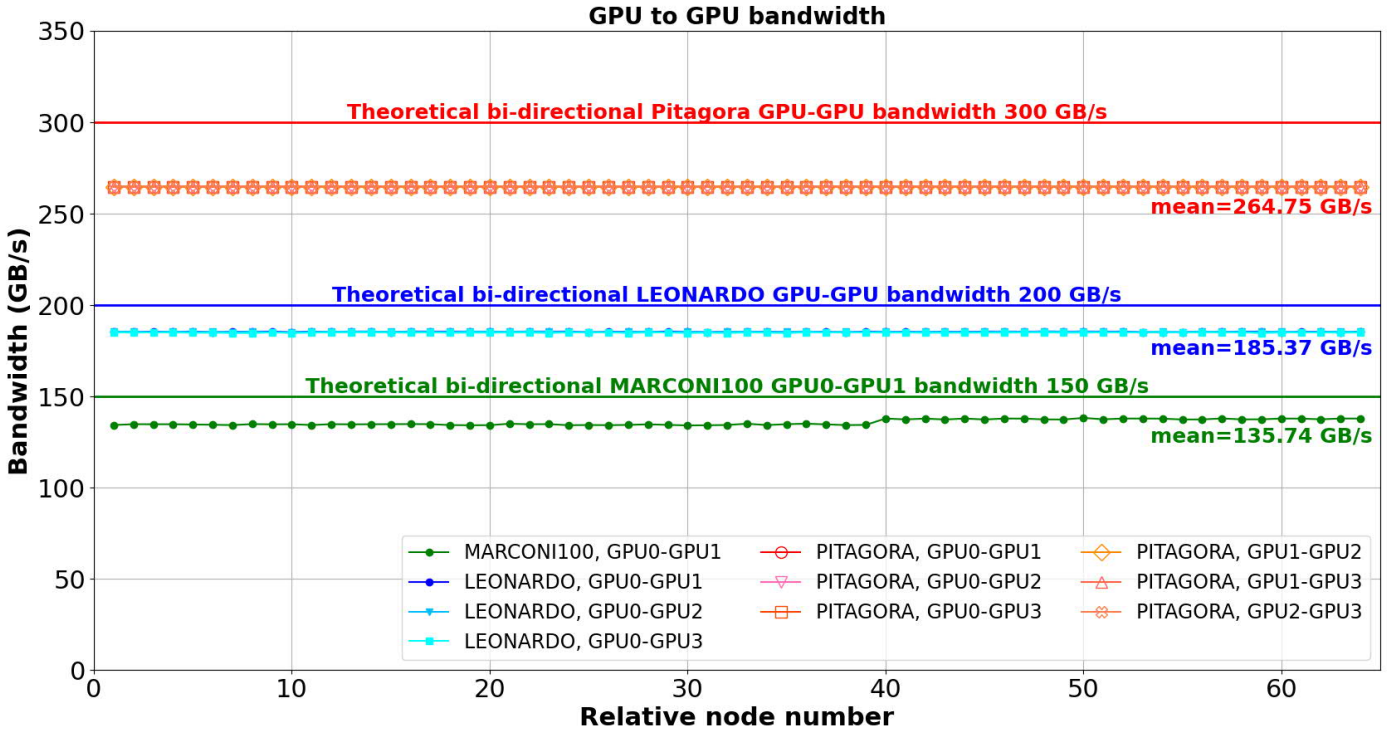


- **Bi-directional bandwidth is the same as uni-directional (~52 GB/s) — possible issue.**
- **Some GPUs (or connections) are slower compared to others — see figure.**

Pitagora-GPU: Device to Device connection

Device to Device connection is NVLink 4.0:

- **18 links (6 for each GPU pair) with 50 GB/s bi-directional bandwidth per link.**

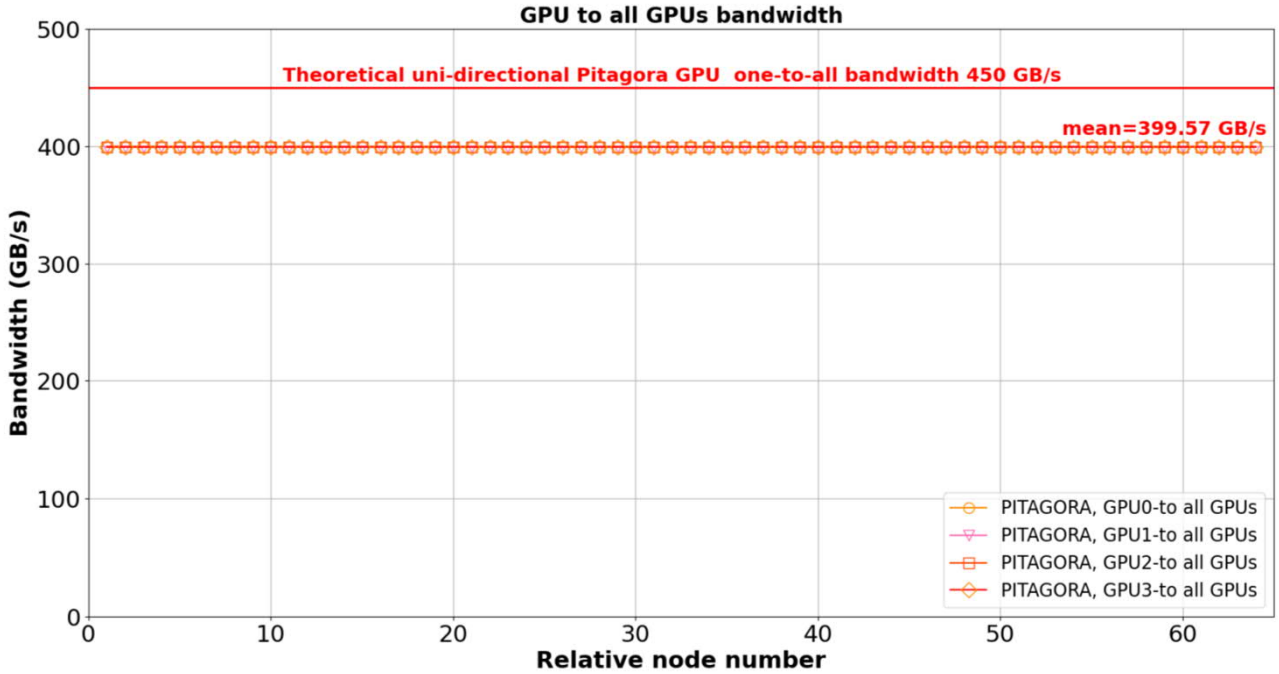


- **The results are stable and symmetric.**
- **Pitagora-GPU: the mean bi-directional bandwidth of all GPU pairs 264.75 GB/s from 300 GB/s of the theoretical value (88%): 6 NVLinks with 50 GB/s each.**
- **LEONARDO: the mean bi-directional bandwidth of all GPU pairs 185.5 GB/s from 200 GB/s of the theoretical value (93%): 4 NVLinks with 50 GB/s each.**
- **MARCONI100: the mean bi-directional bandwidth of ~136 GB/s from 150 GB/s of the theoretical value (90%).**

Pitagora-GPU: GPU to all GPUs connection

Device to Device connection is NVLink 4.0:

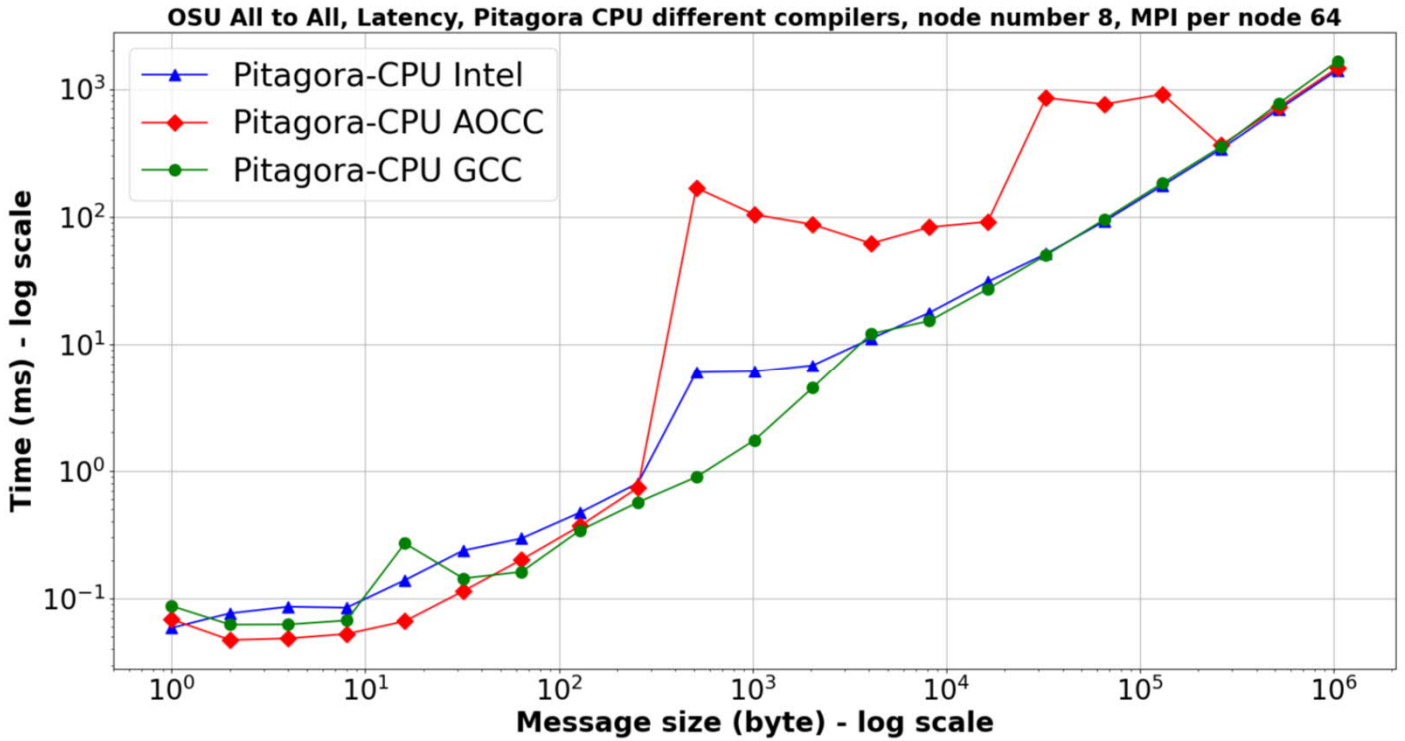
- **18 links (6 for each GPU pair) with 50 GB/s bi-directional bandwidth per link or 25 GB/s uni-directional.**



- **The results are stable and symmetric.**
- **Pitagora-GPU: the mean uni-directional bandwidth 399.57 GB/s from 450 GB/s of the theoretical value (89%).**

Pitagora-CPU: *all_to_all* latency communication test

using osu_alltoall benchmark from OSU microbenchmark

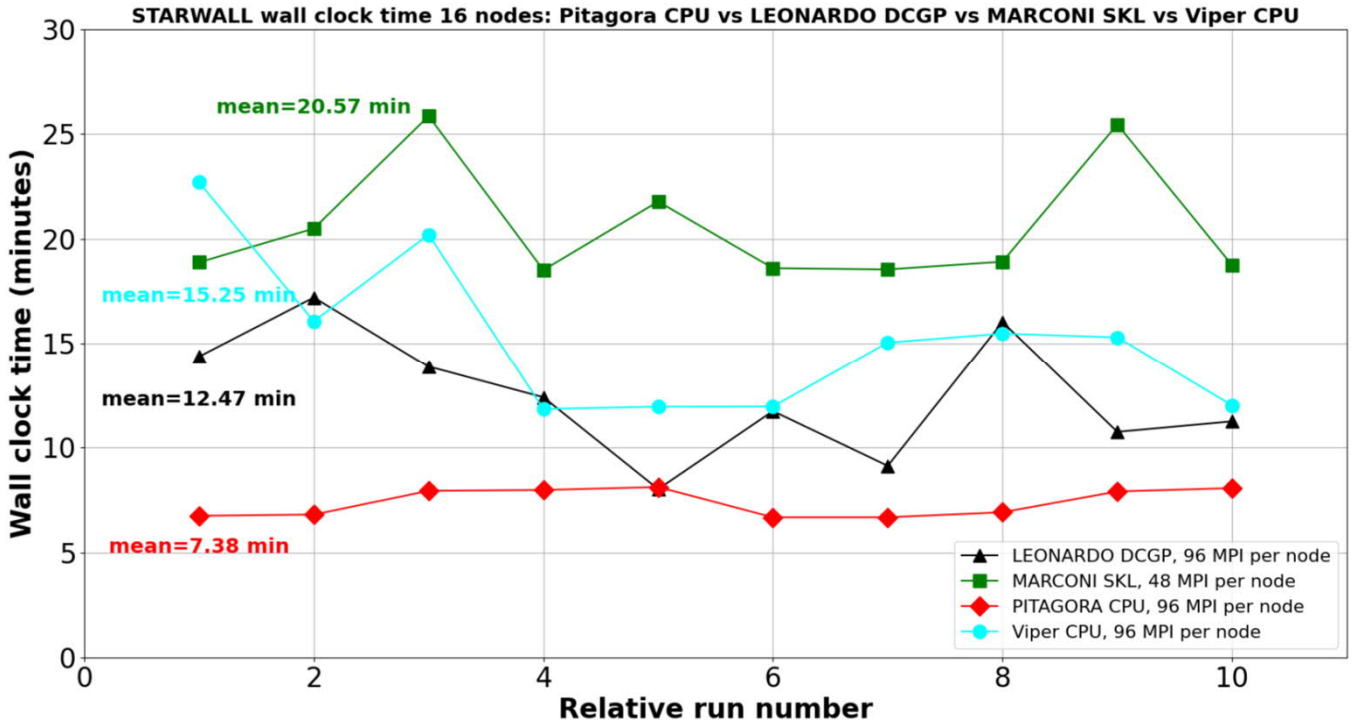


- **Average time to complete the *all-to-all* operation increases with message size, as expected.**
- **AOCC performs better for small messages (<32 B), but is slower for intermediate sizes (256 B and 256 kB).**
- **For large messages (>256 kB), all three compilers show similar performance.**

STARWALL performance

Pure MPI + ScaLAPACK

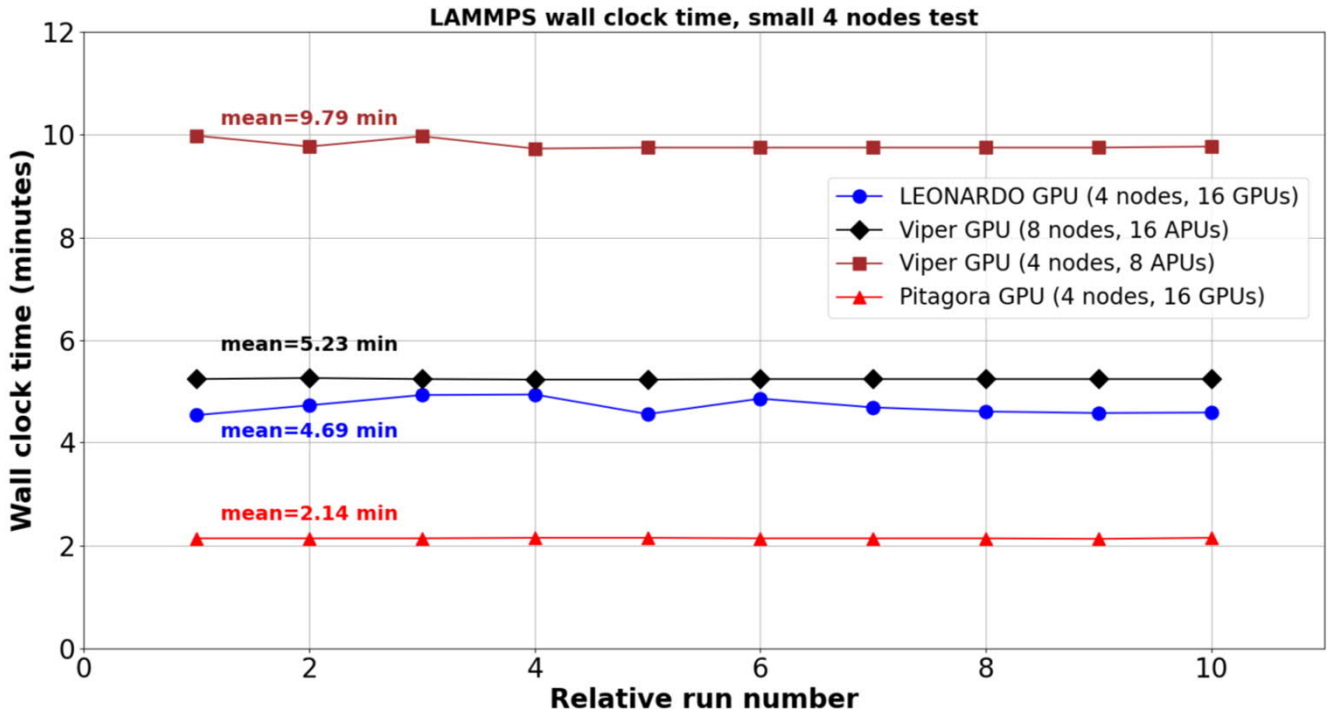
16 nodes, 96 MPI per node



- **The execution time fluctuates on all supercomputers.**
- **Pitagora-CPU delivers the fastest performance, despite using under half of the node (96 out of 256 cores).**
- **Many runs failed on Pitagora-CPU due to an issue with a ScaLAPACK library subroutine that is still under investigation.**

LAMMPS performance (small testcase)

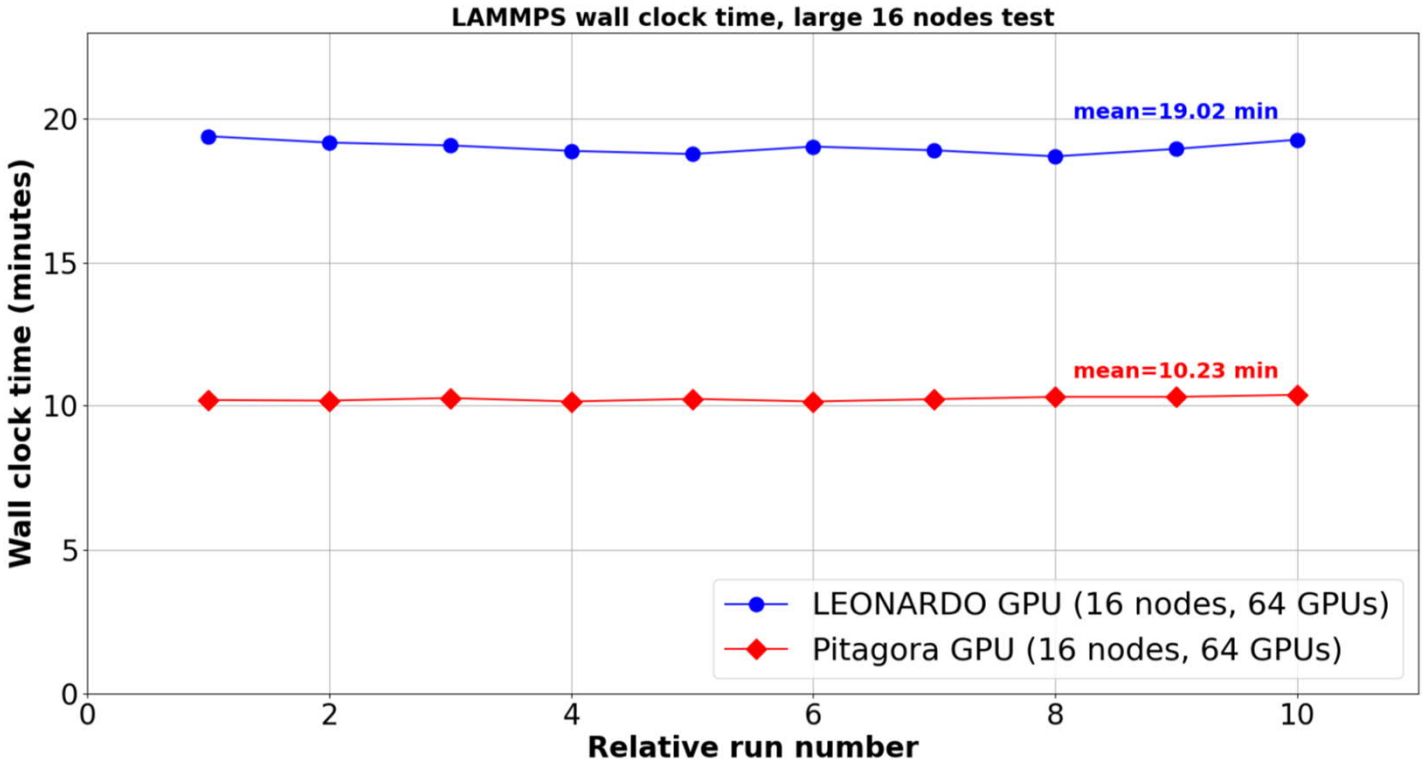
4 nodes, 16 MPIs, 16 GPUs



- The execution time is **stable** across supercomputers.
- On Pitagora, the code runs more than **twice as fast** compared to LEONARDO Booster and Viper-GPU.

LAMMPS performance (large testcase)

16 nodes, 64 MPIs, 64 GPUs

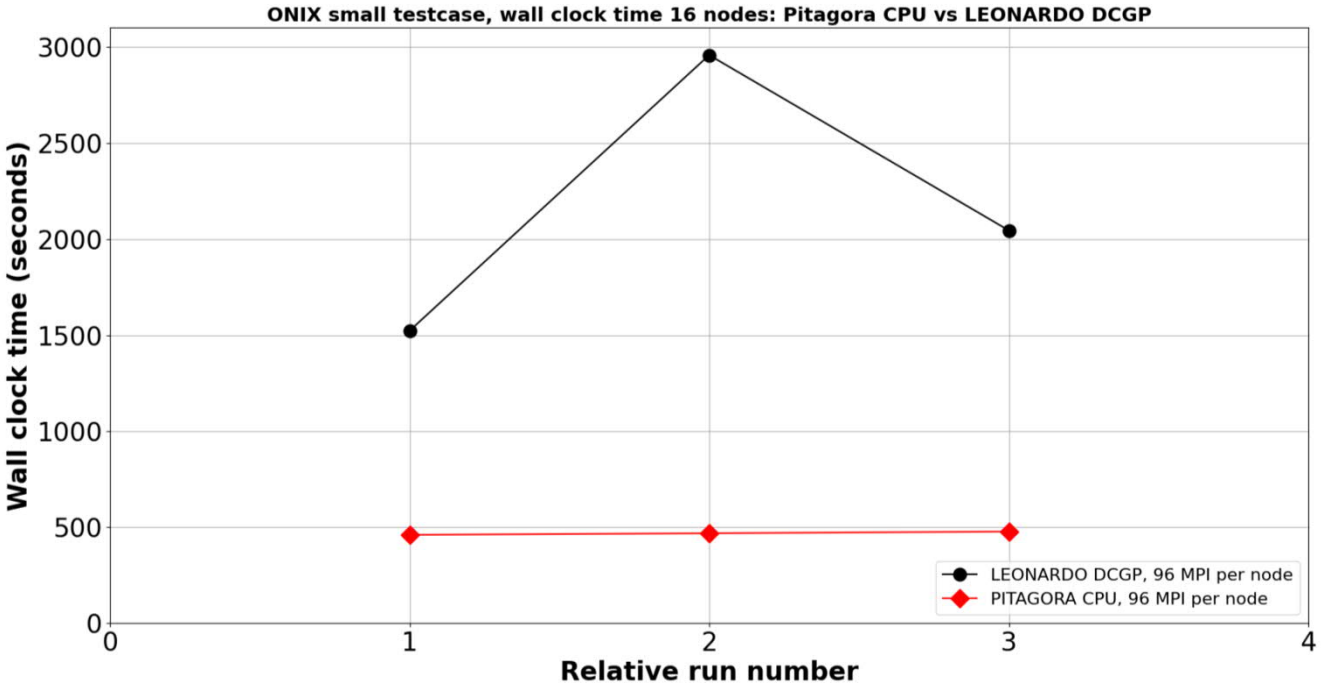


- **The execution time is stable across supercomputers.**
- **On Pitagora, the code runs almost twice as fast as on LEONARDO Booster.**

ONIX performance (small testcase)

Pure MPI

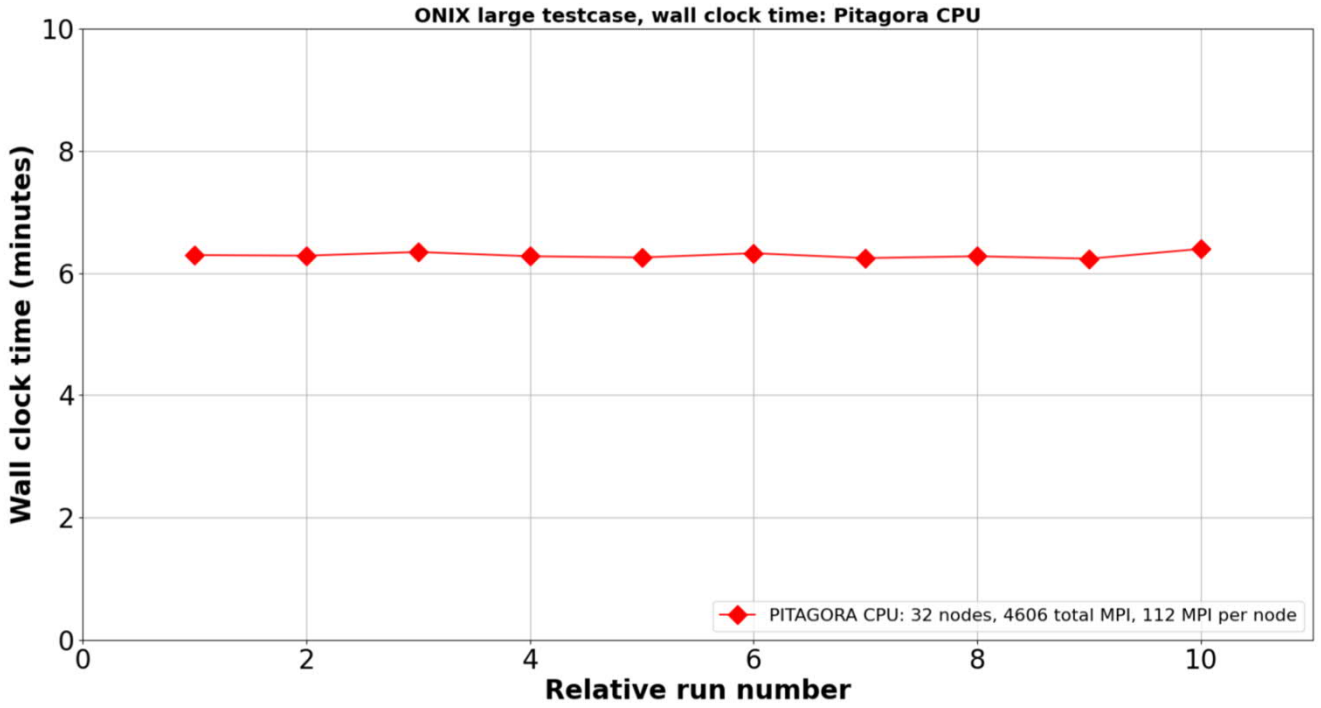
16 nodes, 96 MPI per node



- **The execution time is stable on all Pitagora-CPU.**
- **Pitagora-CPU delivers more than three times the performance of the LEONARDO DCGP partition, despite using less than half of the node (96 out of 256 cores).**

ONIX performance (large testcase)

32 nodes, 112 MPI per node



- The execution time is stable on all Pitagora-CPU.
- All runs completed successfully without any failed jobs.

Thank you for our attention!