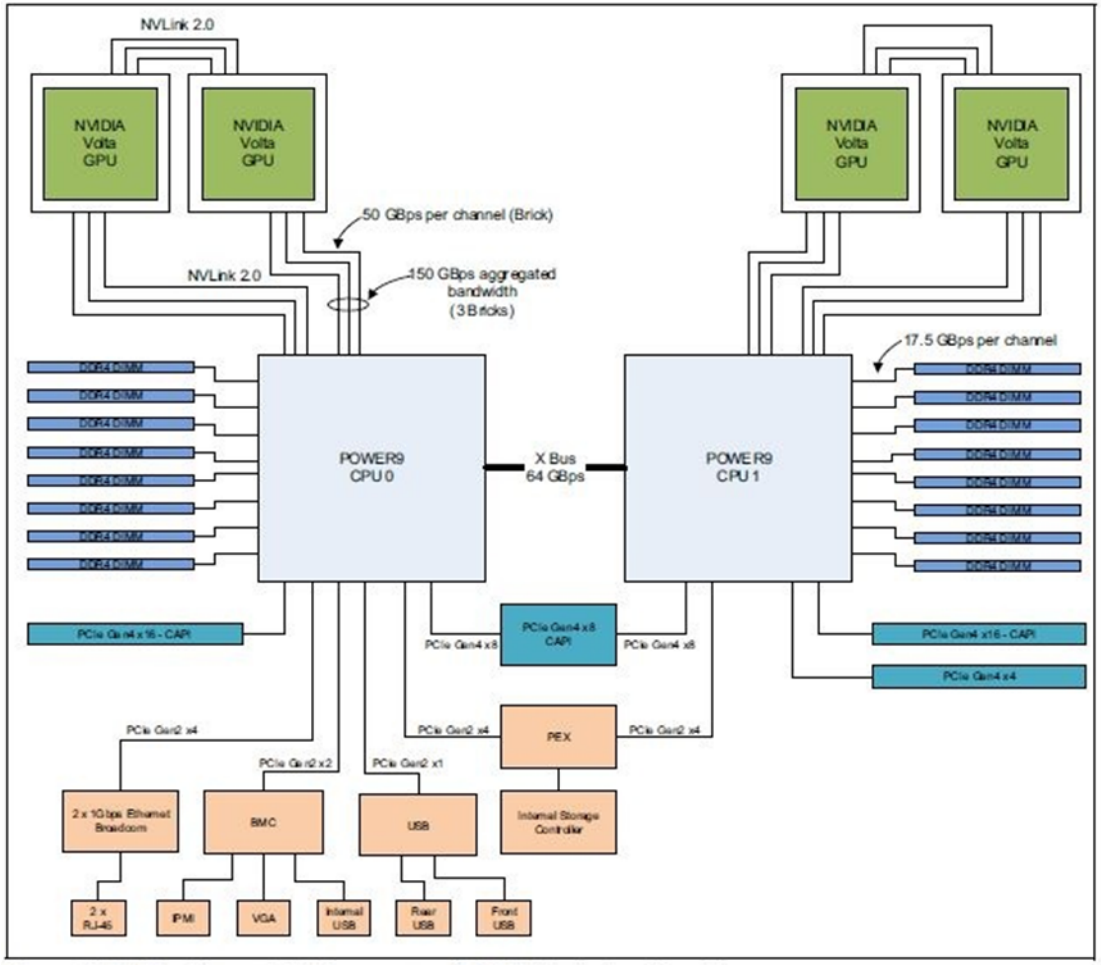
Benchmarks and validation of the Marconi100 HPC system

Serhiy Mochalskyy

IFERC Workshop
December 16th, 2020

High Level Support Team
Max-Planck-Institut für Plasmaphysik
Boltzmannstr. 2, D-85748 Garching, Germany

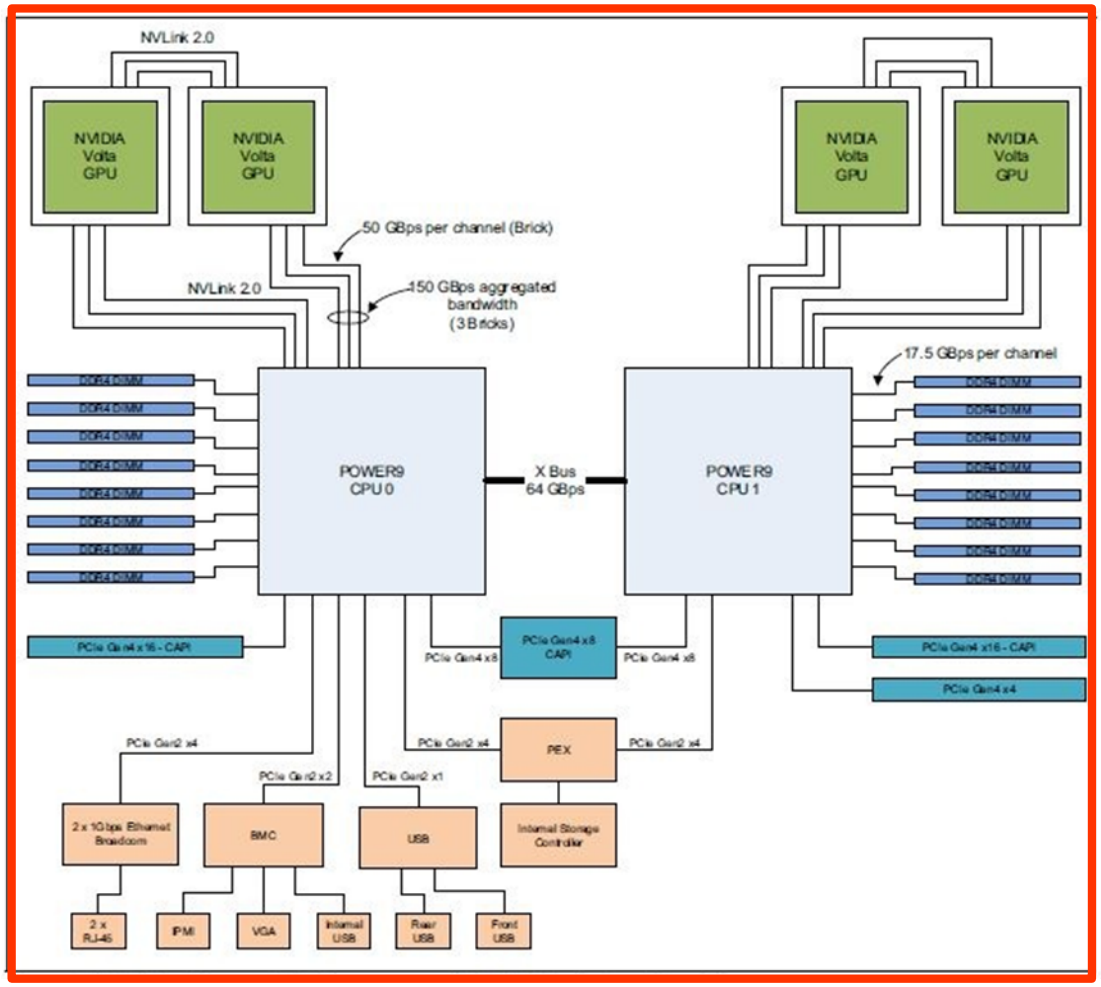
Marconi100 Power9 node (AC922) architecture



- 1 node
- 2 IBM Sockets (CPUs) / node
- 4 NVIDIA Volta V100 GPUs / node
- 16 cores / CPU
- 4 threads / core

- node – node: Mellanox IB EDR DragonFly+
- CPU – CPU: X Bus
- CPU – GPU: NVLink
- GPU – GPU: NVLink
- DDR4 – CPU: PCIe

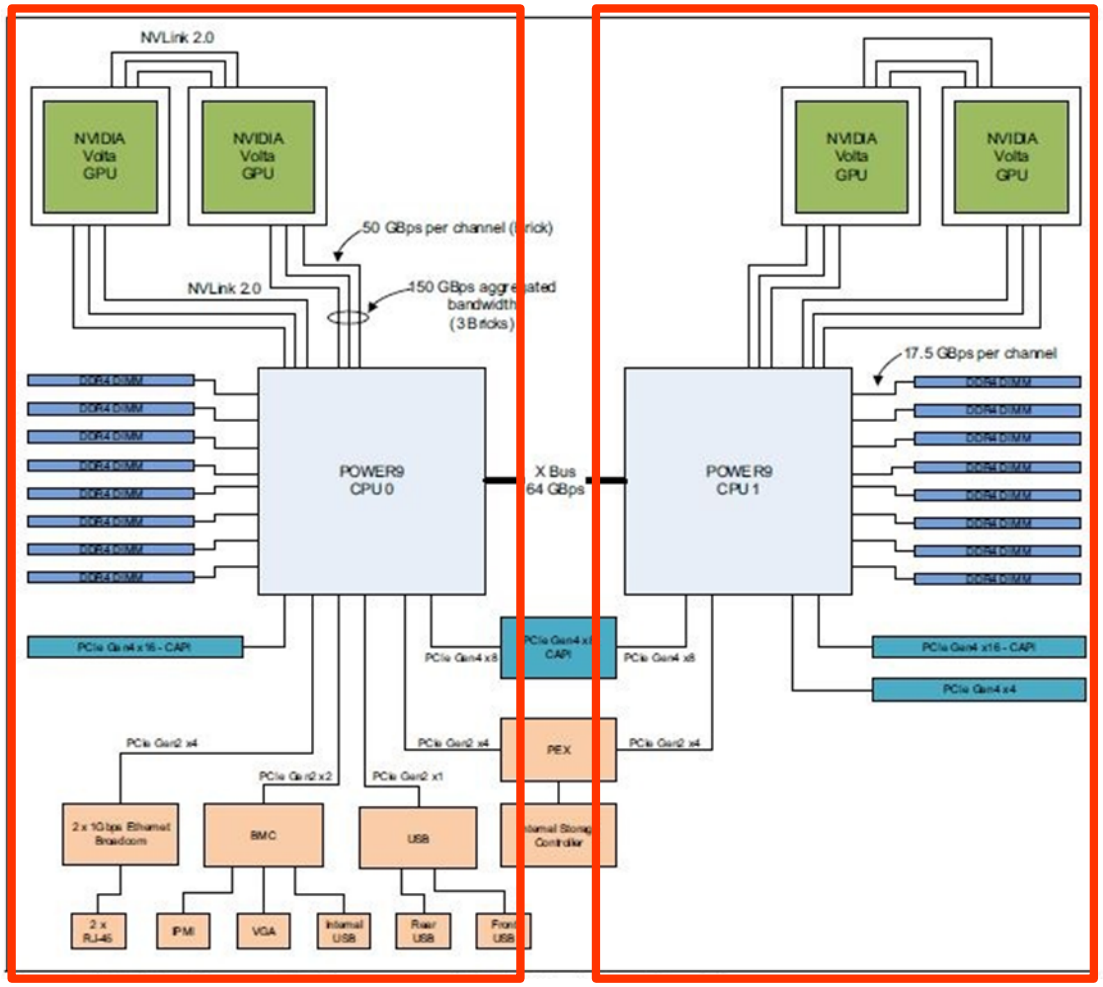
Marconi100 Power9 node (AC922) architecture



- **1 node**
- **2 IBM Sockets (CPUs) / node**
- **4 NVIDIA Volta V100 GPUs / node**
- **16 cores / CPU**
- **4 threads / core**

- **node – node: Mellanox IB EDR DragonFly+**
- **CPU – CPU: X Bus**
- **CPU – GPU: NVLink**
- **GPU – GPU: NVLink**
- **DDR4 – CPU: PCIe**

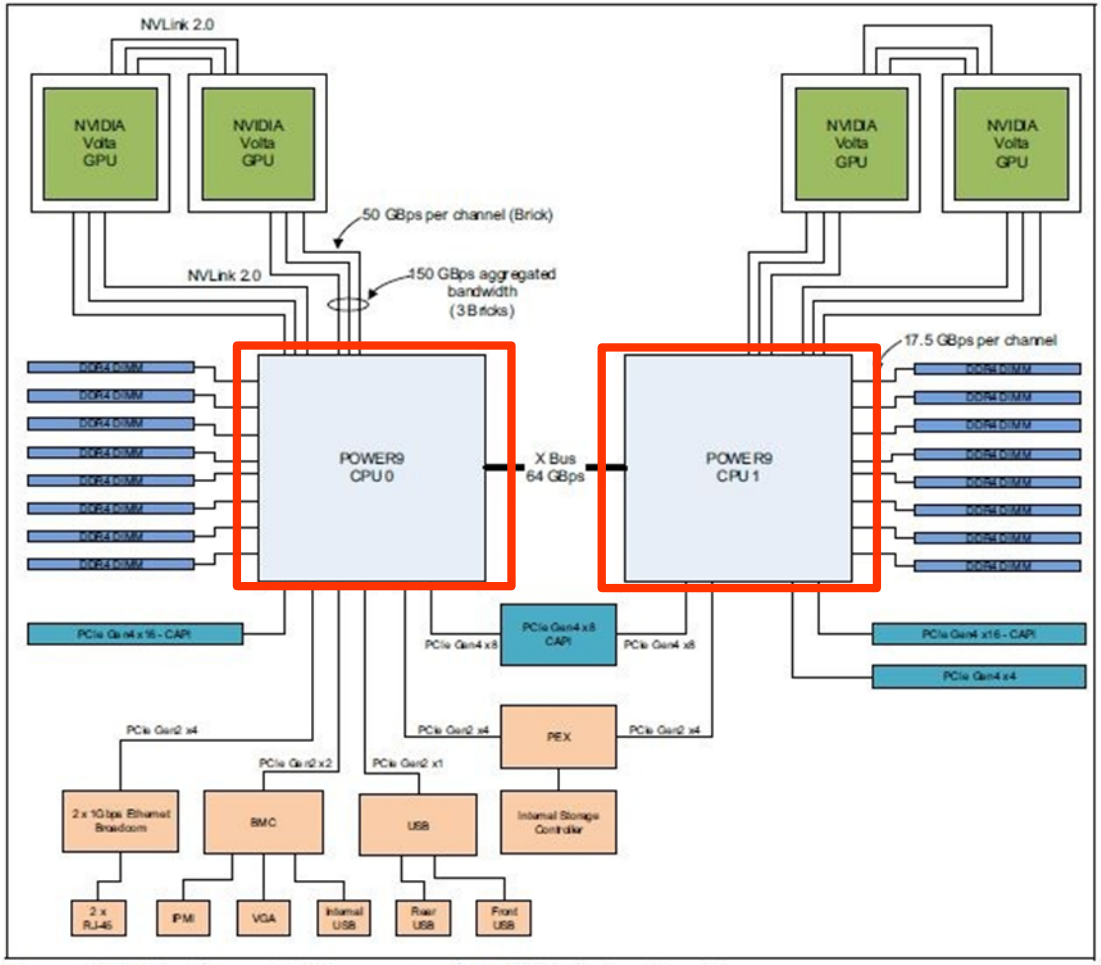
Marconi100 Power9 node (AC922) architecture



- 1 node
- 2 IBM Sockets (CPUs) / node
- 4 NVIDIA Volta V100 GPUs / node
- 16 cores / CPU
- 4 threads / core

- node – node: Mellanox IB EDR DragonFly+
- CPU – CPU: X Bus
- CPU – GPU: NVLink
- GPU – GPU: NVLink
- DDR4 – CPU: PCIe

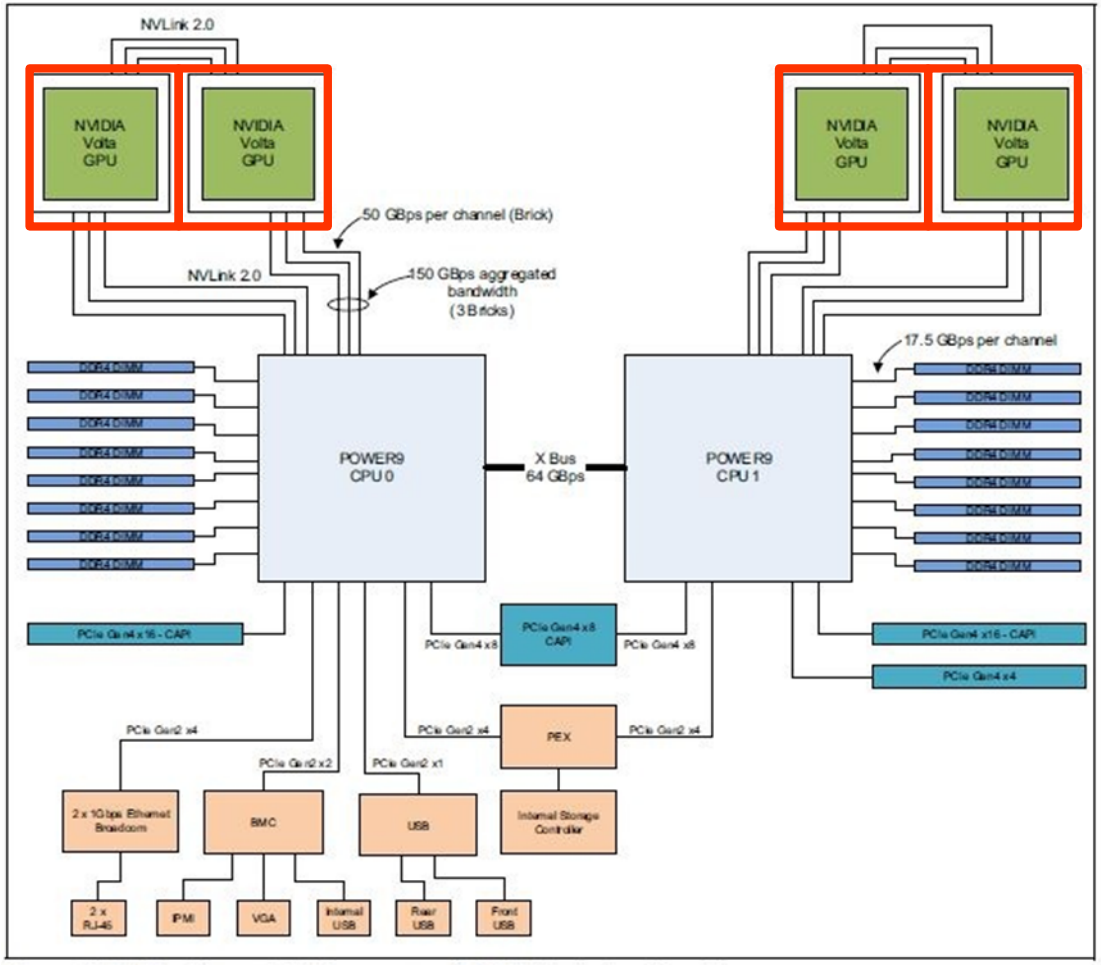
Marconi100 Power9 node (AC922) architecture



- 1 node
- 2 IBM Sockets (CPUs) / node
- 4 NVIDIA Volta V100 GPUs / node
- 16 cores / CPU
- 4 threads / core

- node – node: Mellanox IB EDR DragonFly+
- CPU – CPU: X Bus
- CPU – GPU: NVLink
- GPU – GPU: NVLink
- DDR4 – CPU: PCIe

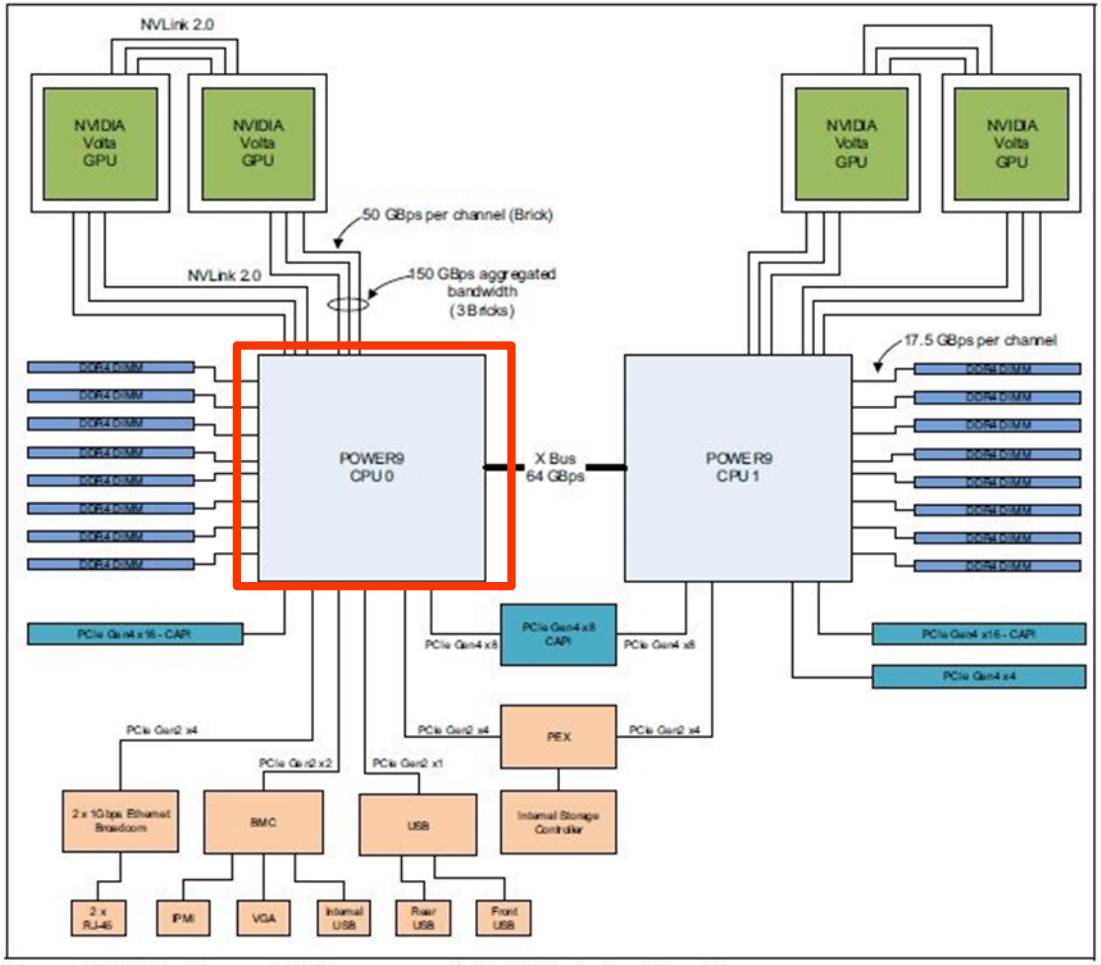
Marconi100 Power9 node (AC922) architecture



- 1 node
- 2 IBM Sockets (CPUs) / node
- 4 NVIDIA Volta V100 GPUs / node; 2 / Socket
- 16 cores / CPU
- 4 threads / core

- node – node: Mellanox IB EDR DragonFly+
- CPU – CPU: X Bus
- CPU – GPU: NVLink
- GPU – GPU: NVLink
- DDR4 – CPU: PCIe

Marconi100 Power9 node (AC922) architecture

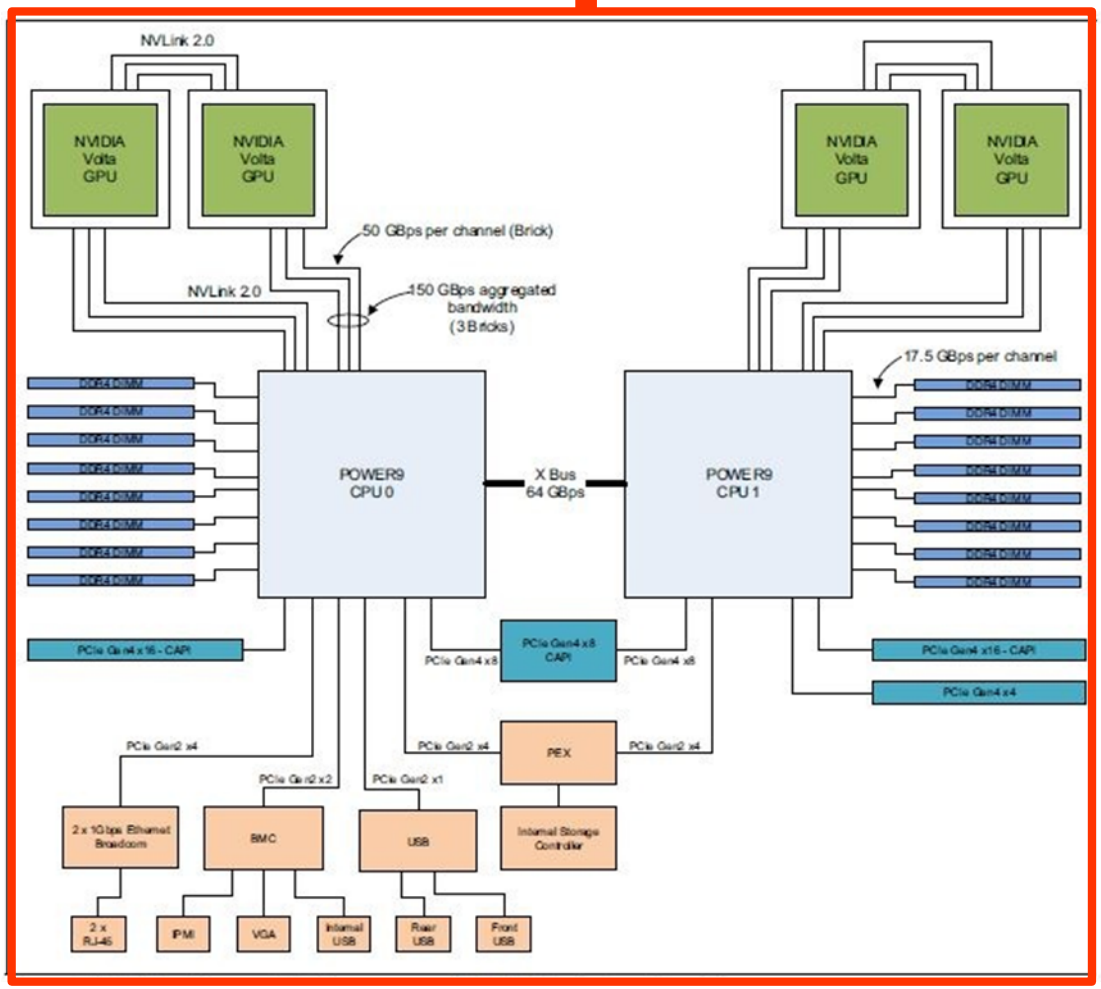


- 1 node
- 2 IBM Sockets (CPUs) / node
- 4 NVIDIA Volta V100 GPUs / node
- 16 cores / CPU
- 4 threads / core

- node – node: Mellanox IB EDR DragonFly+
- CPU – CPU: X Bus
- CPU – GPU: NVLink
- GPU – GPU: NVLink
- DDR4 – CPU: PCIe

Marconi100 Power9 node (AC922) architecture

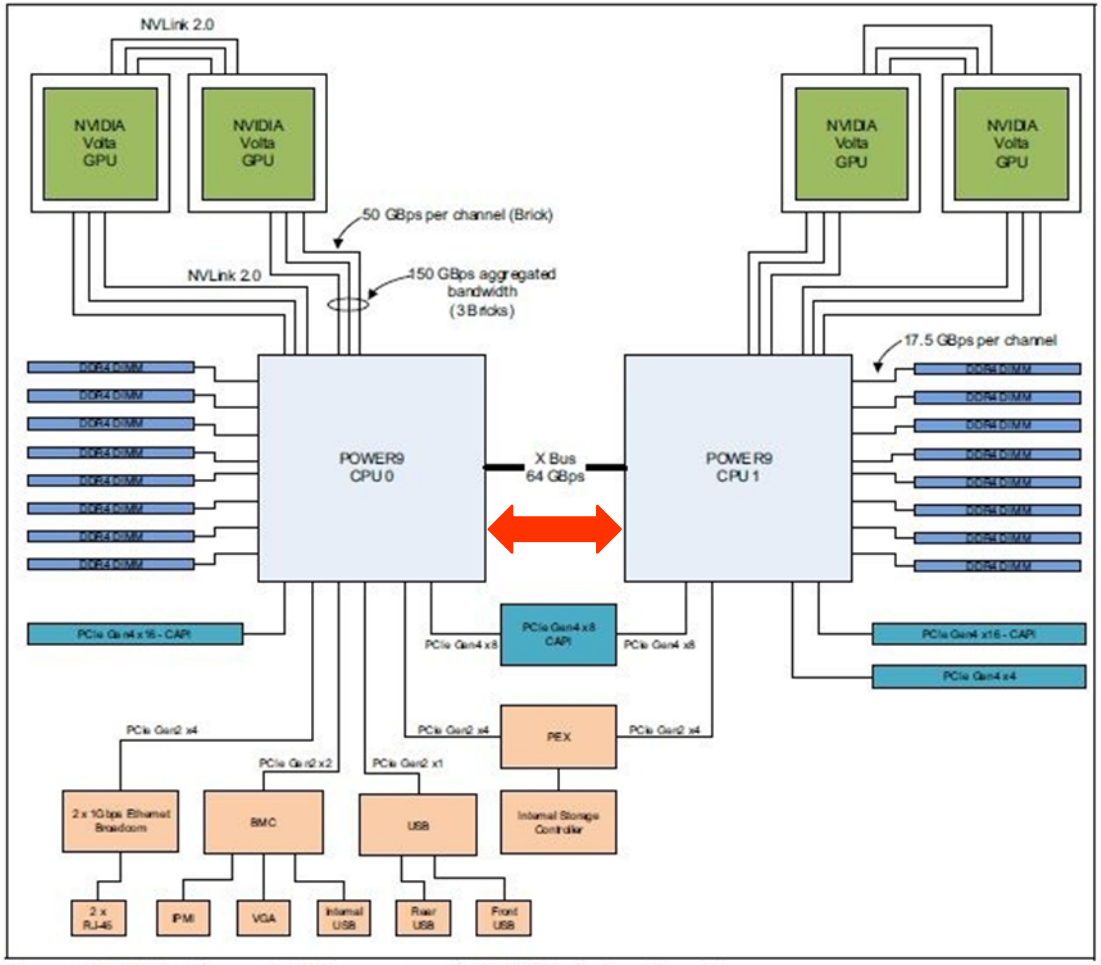
100 Gb/s bi-directional bandwidth



- 1 node
- 2 IBM Sockets (CPUs) / node
- 4 NVIDIA Volta V100 GPUs / node
- 16 cores / CPU
- 4 threads / core

- node – node: Mellanox IB EDR DragonFly+
- CPU – CPU: X Bus
- CPU – GPU: NVLink
- GPU – GPU: NVLink
- DDR4 – CPU: PCIe

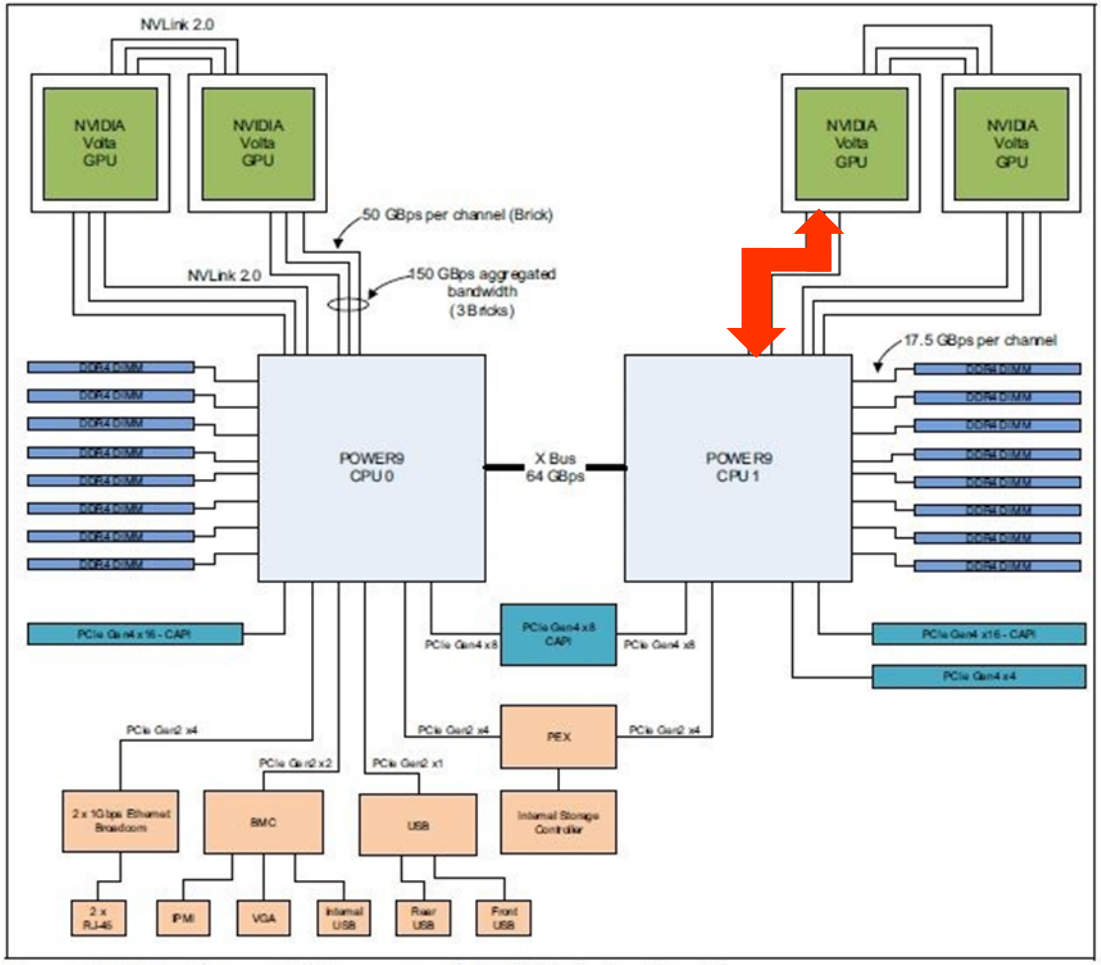
Marconi100 Power9 node (AC922) architecture



- 1 node
- 2 IBM Sockets (CPUs) / node
- 4 NVIDIA Volta V100 GPUs / node
- 16 cores / CPU
- 4 threads / core

- node – node: Mellanox IB EDR DragonFly+
- CPU – CPU: X Bus
- CPU – GPU: NVLink
- GPU – GPU: NVLink
- DDR4 – CPU: PCIe

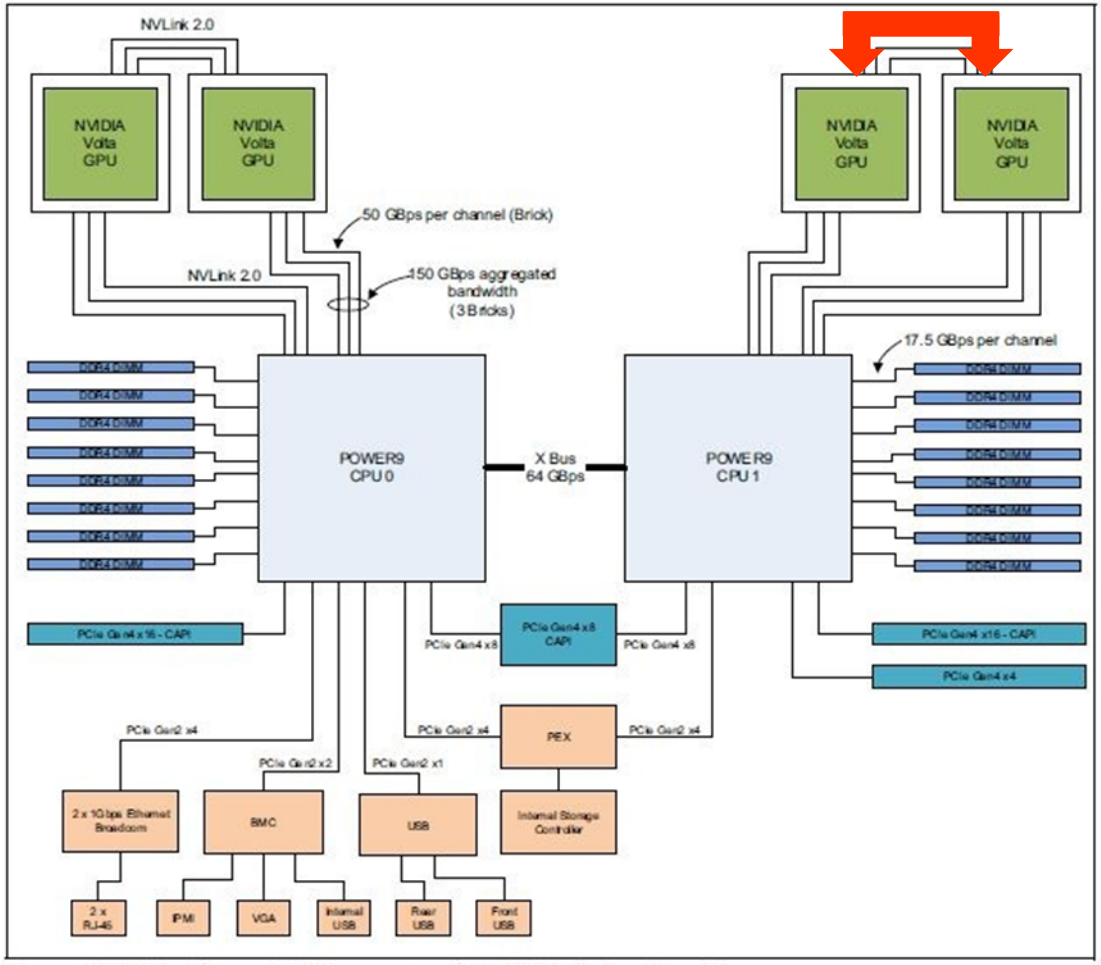
Marconi100 Power9 node (AC922) architecture



- 1 node
- 2 IBM Sockets (CPUs) / node
- 4 NVIDIA Volta V100 GPUs / node
- 16 cores / CPU
- 4 threads / core

- node – node: Mellanox IB EDR DragonFly+
- CPU – CPU: X Bus
- CPU – GPU: NVLink
- GPU – GPU: NVLink
- DDR4 – CPU: PCIe

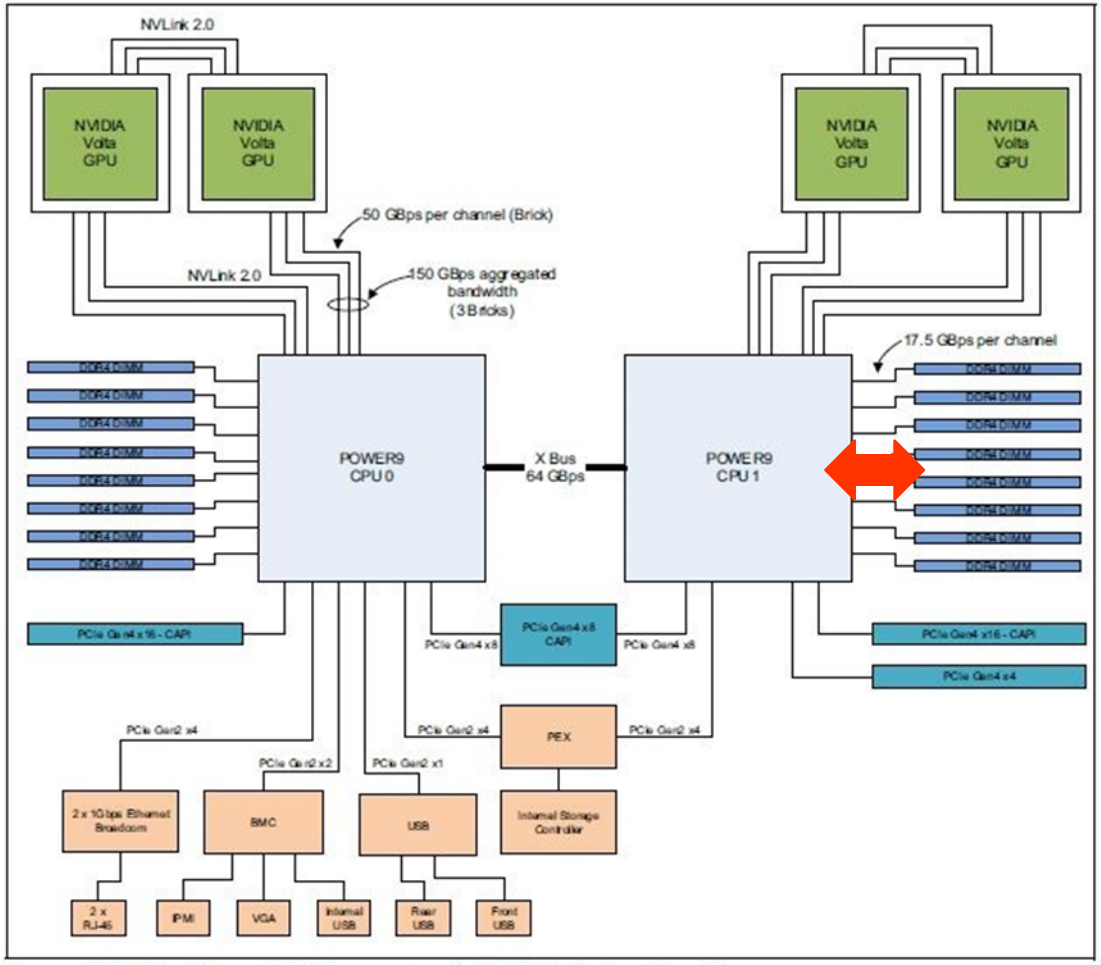
Marconi100 Power9 node (AC922) architecture



- 1 node
- 2 IBM Sockets (CPUs) / node
- 4 NVIDIA Volta V100 GPUs / node
- 16 cores / CPU
- 4 threads / core

- node – node: Mellanox IB EDR DragonFly+
- CPU – CPU: X Bus
- CPU – GPU: NVLink
- GPU – GPU: NVLink
- DDR4 – CPU: PCIe

Marconi100 Power9 node (AC922) architecture



- 1 node
- 2 IBM Sockets (CPUs) / node
- 4 NVIDIA Volta V100 GPUs / node
- 16 cores / CPU
- 4 threads / core

- node – node: Mellanox IB EDR DragonFly+
- CPU – CPU: X Bus
- CPU – GPU: NVLink
- GPU – GPU: NVLink
- DDR4 – CPU: PCIe

Marconi100 EUROfusion part

EUROfusion part – 80 nodes

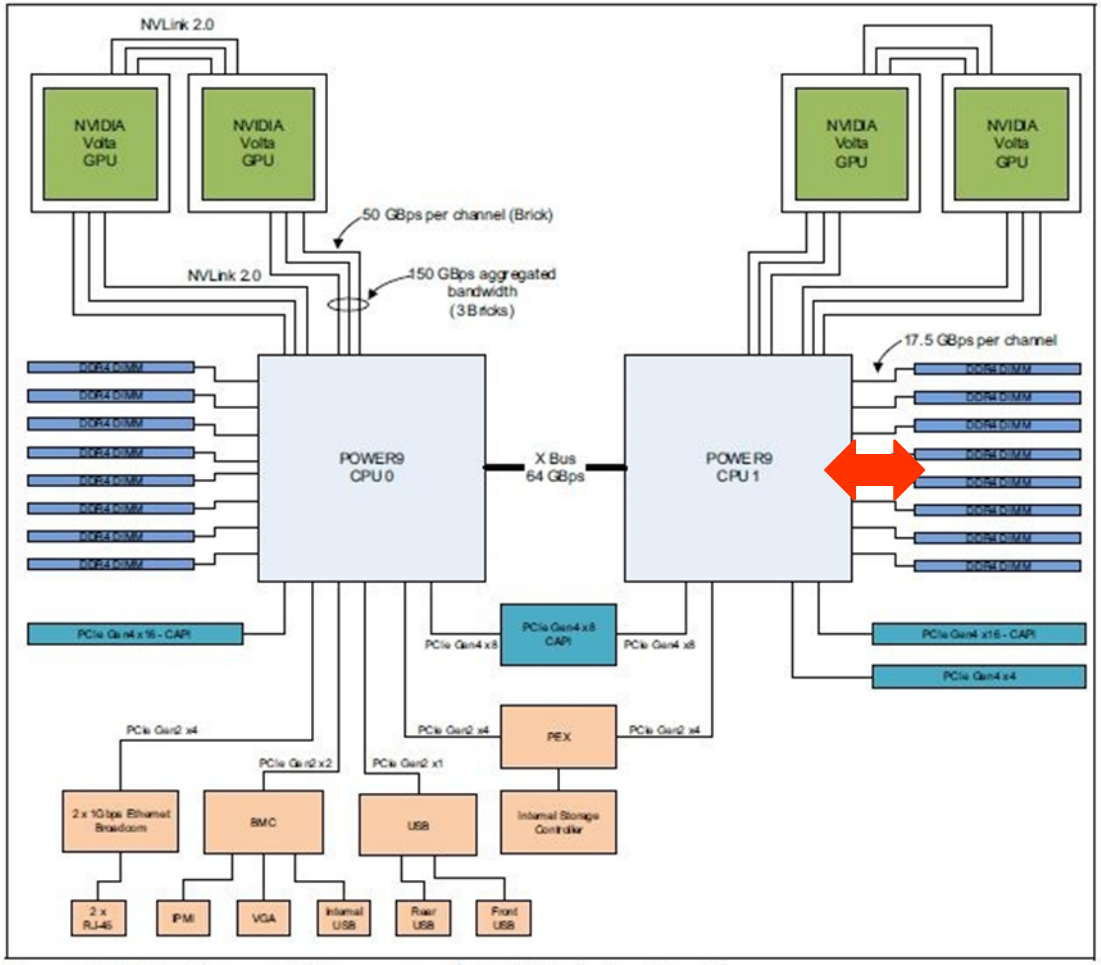
4 racks

r228n **r229n** **r231n** **r232n**

20 nodes per rack

Main memory bandwidth test

DDR4 – CPU: PCIe3



256 GB / node

128 GB / socket

Bandwidth:

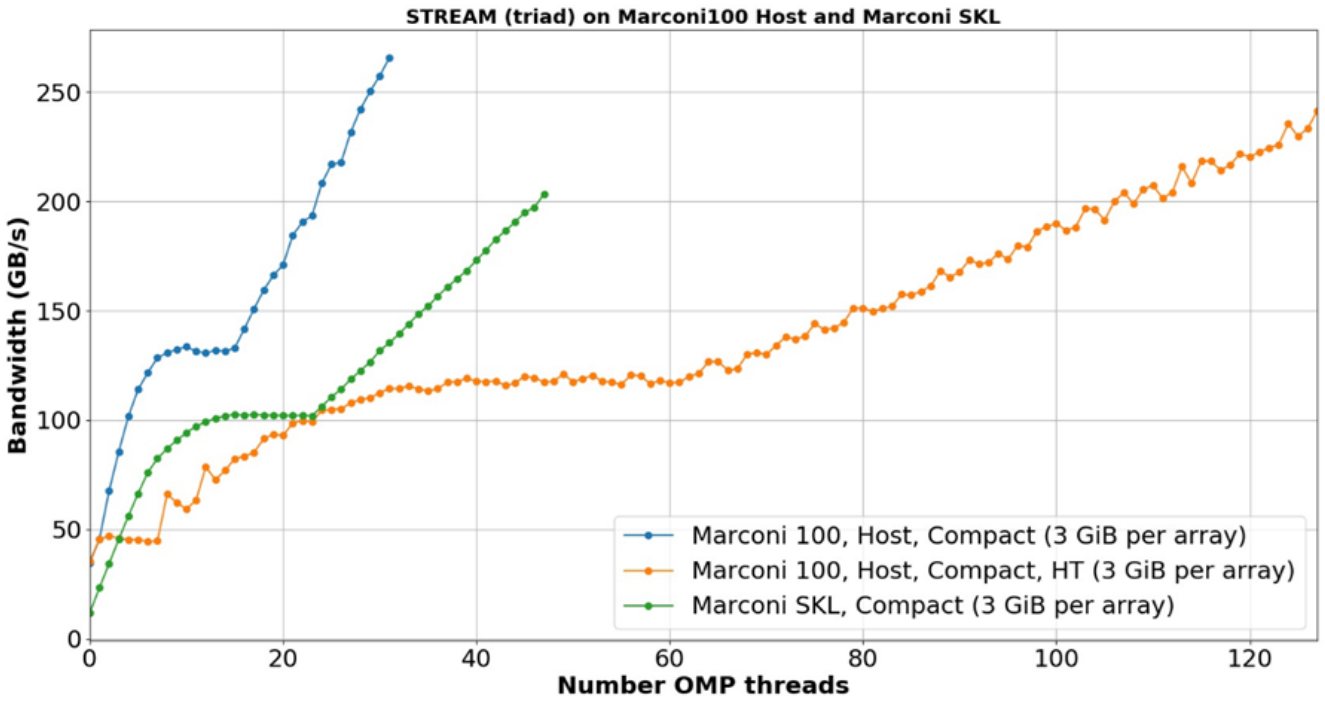
8 memory channels

17.5 GB/s / channel

140 GB/s / socket

280 GB/s / node

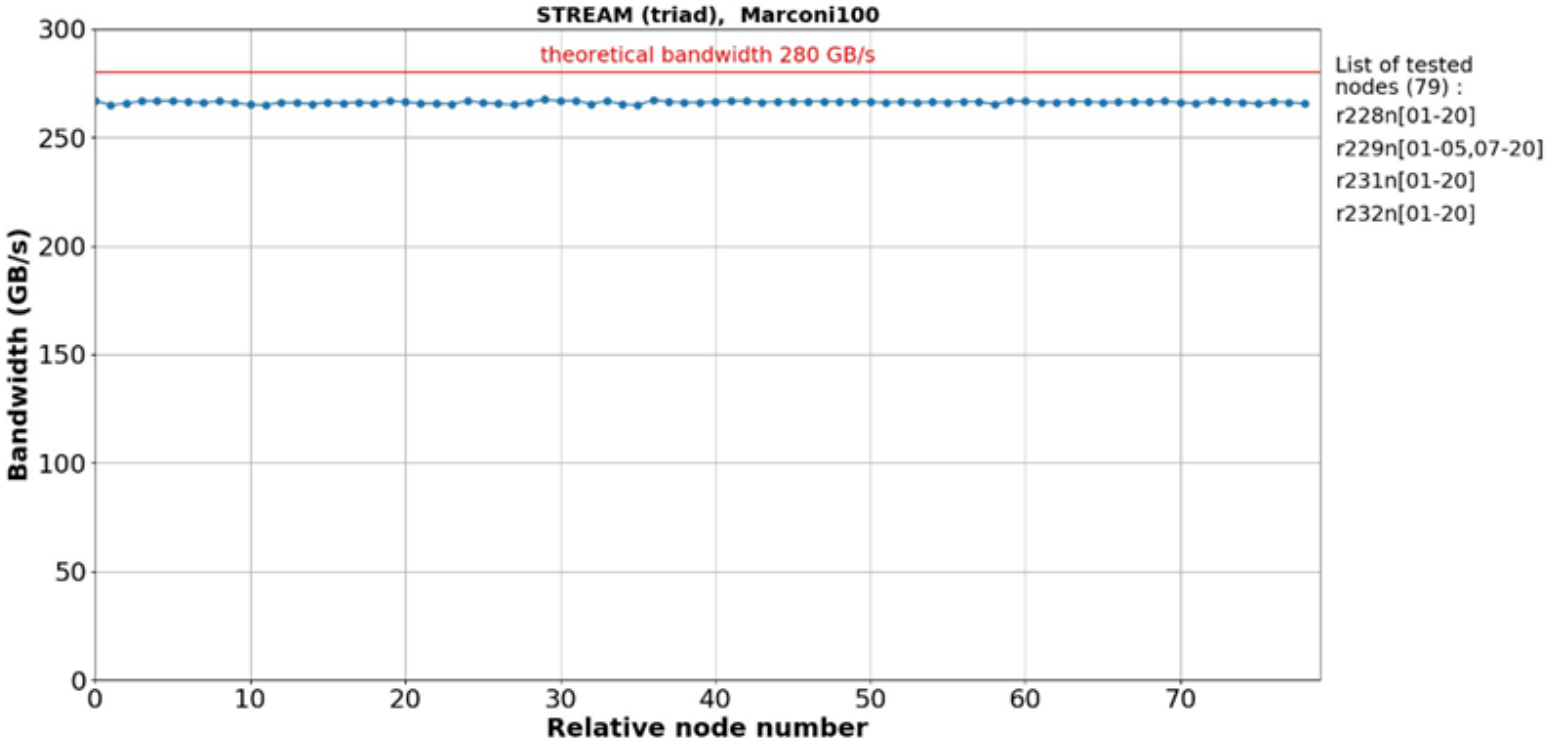
STREAM on a single node



- **Marconi SKL: 203 GB/s from 255.94 GB/s theoretical (80 %)**
- **Marconi100: 266 GB/s from 280 GB/s theoretical (95 %)**

April 28, 2020

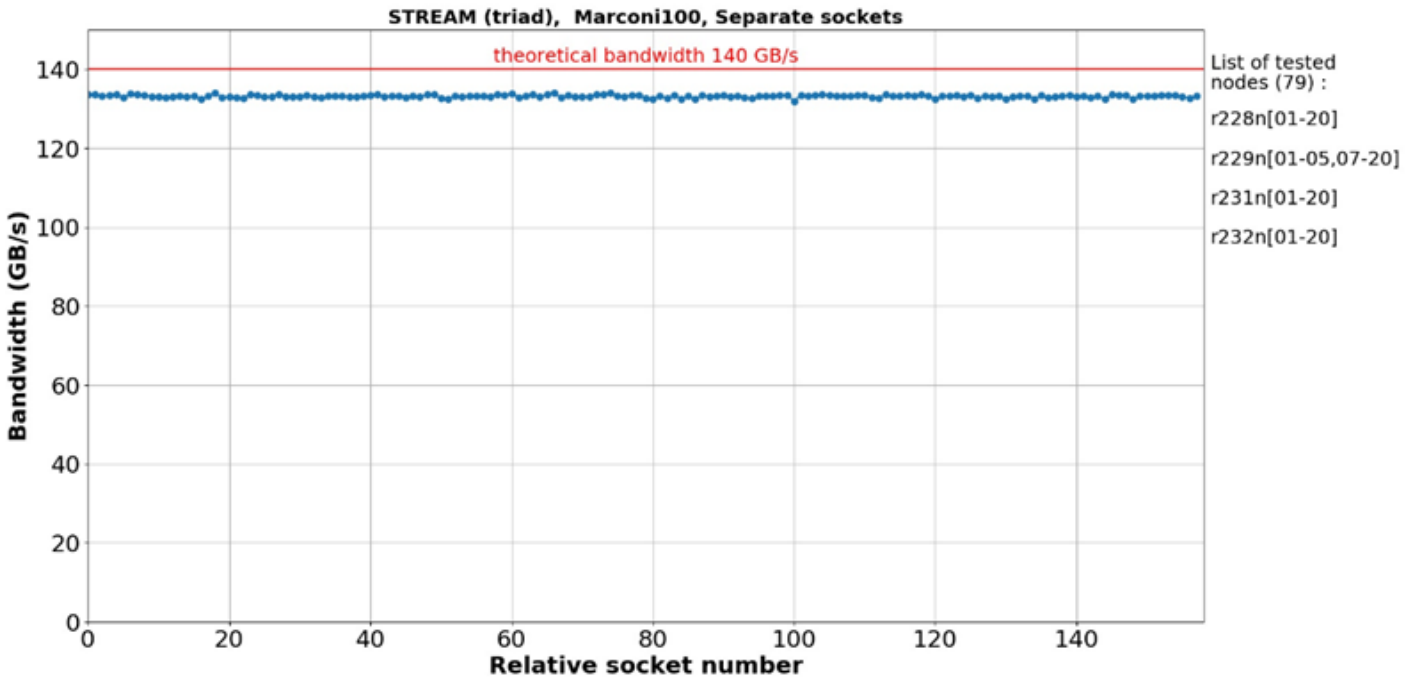
STREAM on the complete EUROfusion Marconi100 partition



- All nodes provide high and stable bandwidth close to the theoretical value.
- The average bandwidth is around 266 GB/s.

April 30, 2020

STREAM on the separate sockets of Marconi100 node

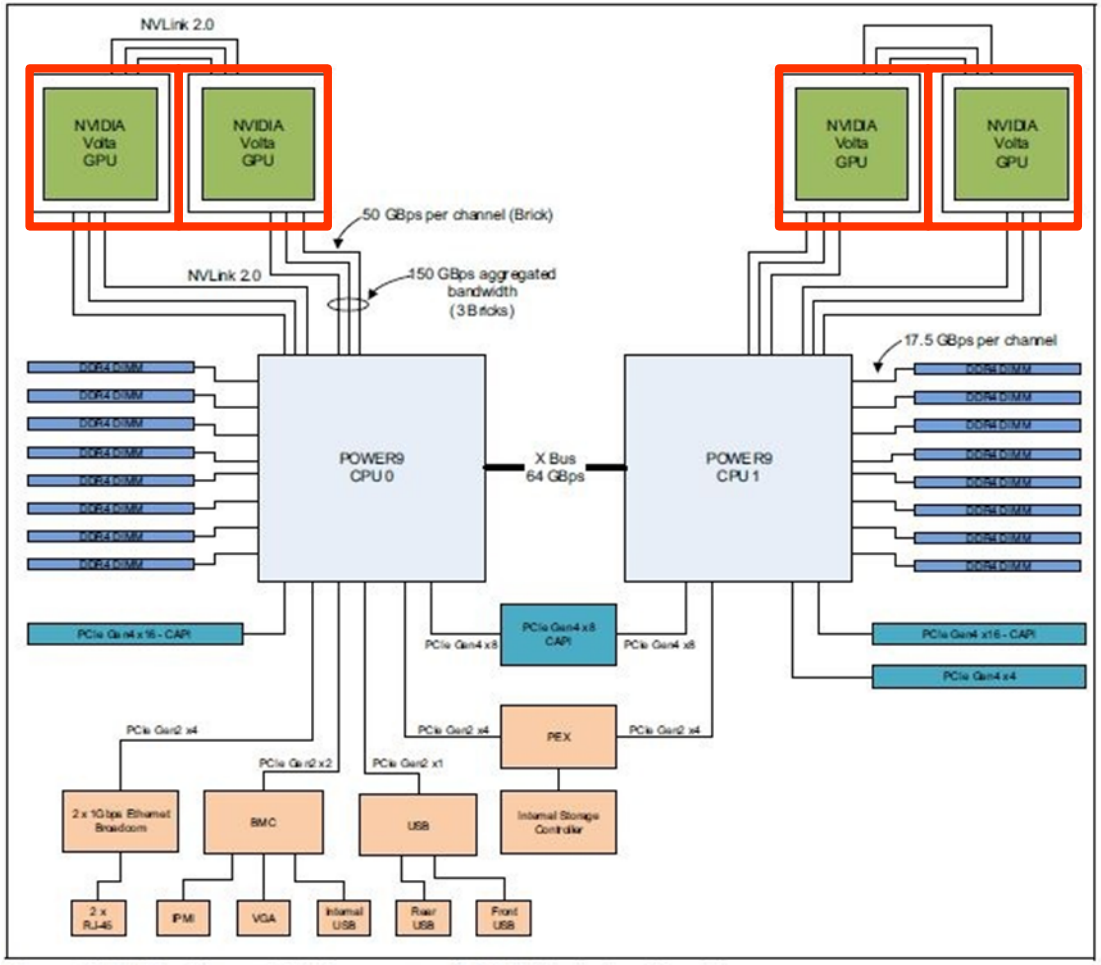


- All sockets provide high and stable bandwidth close to the theoretical value.
- No difference between two sockets inside one node was detected.
- The average bandwidth is around 133 GB/s.

April 30, 2020

GPU bandwidth and performance tests

4 NVIDIA Volta V100 GPUs / node; 2 / Socket



16 GB / GPU internal RAM

Bandwidth:

900 GB/s / card

(SKL = 255 GB/s)

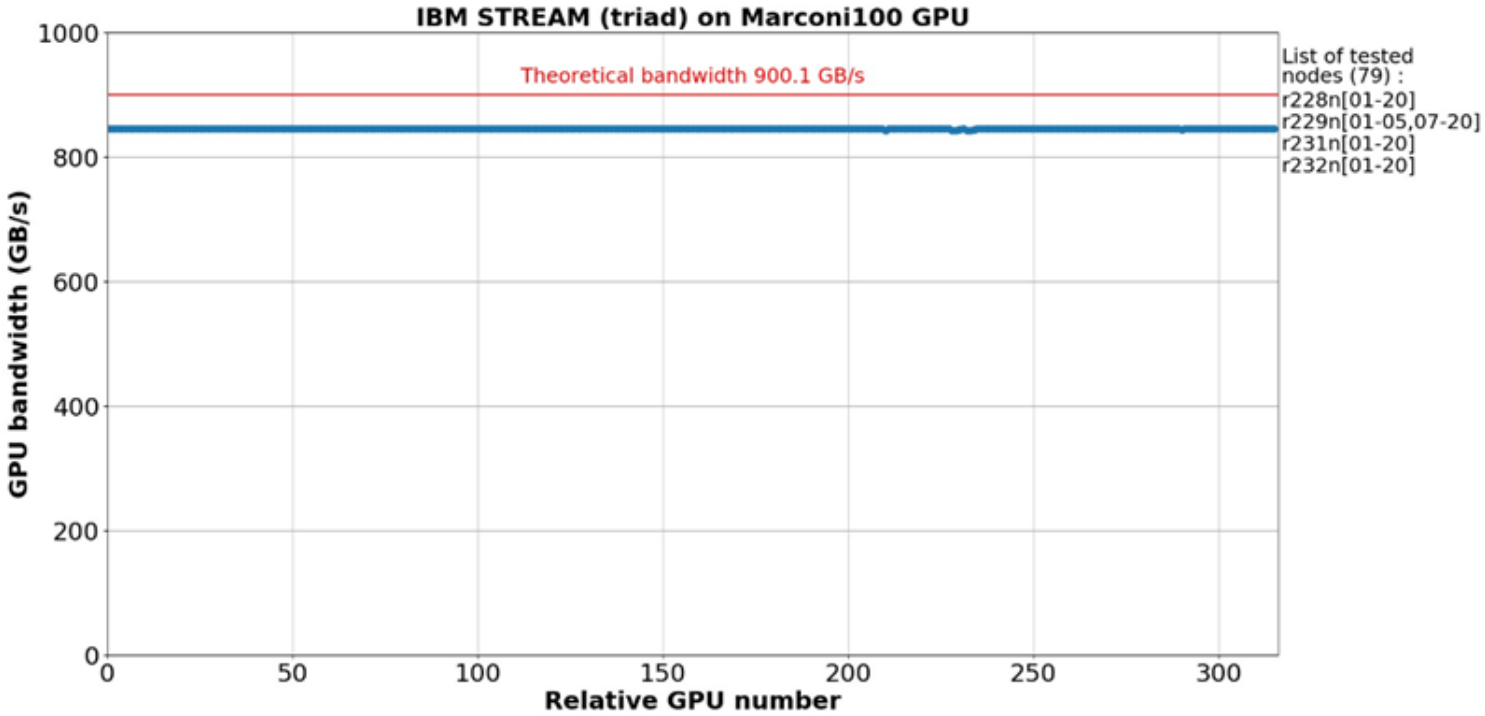
Peak performance:

7.8 Tflops

(SKL = 3.2 TFlop/s

Power9 = 0.8 TFlop/s)

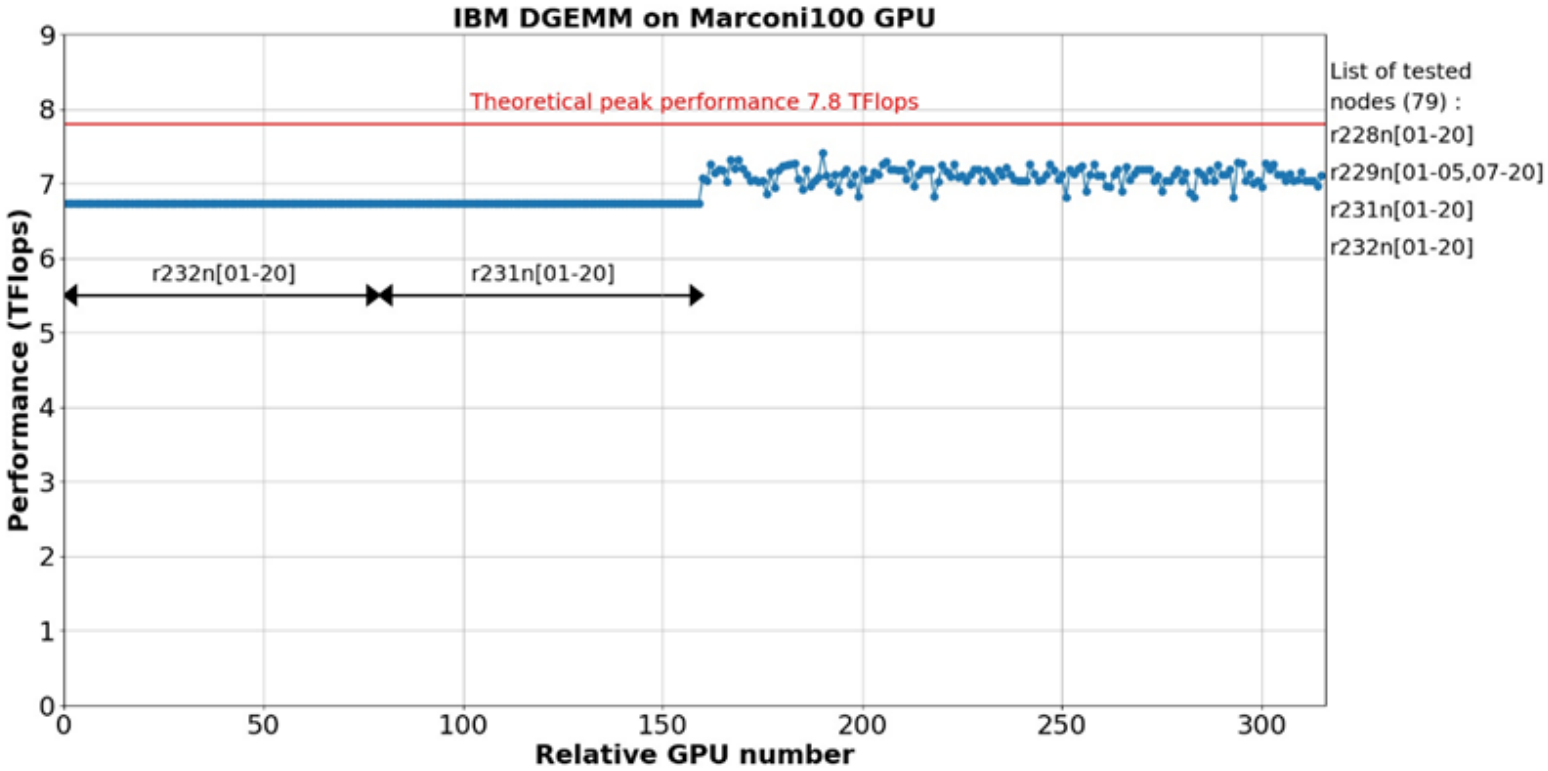
BabelStream benchmark on Marconi100 GPU



- All GPUs provide high, stable and symmetric bandwidth close to the theoretical value.
- No difference between two distinct GPUs on different nodes or two GPU cards inside one socket was detected.
- The average bandwidth is ~845 GB/s that is 94 % of the theoretical value.

April 30, 2020

DGEMM benchmark on Marconi100 GPU

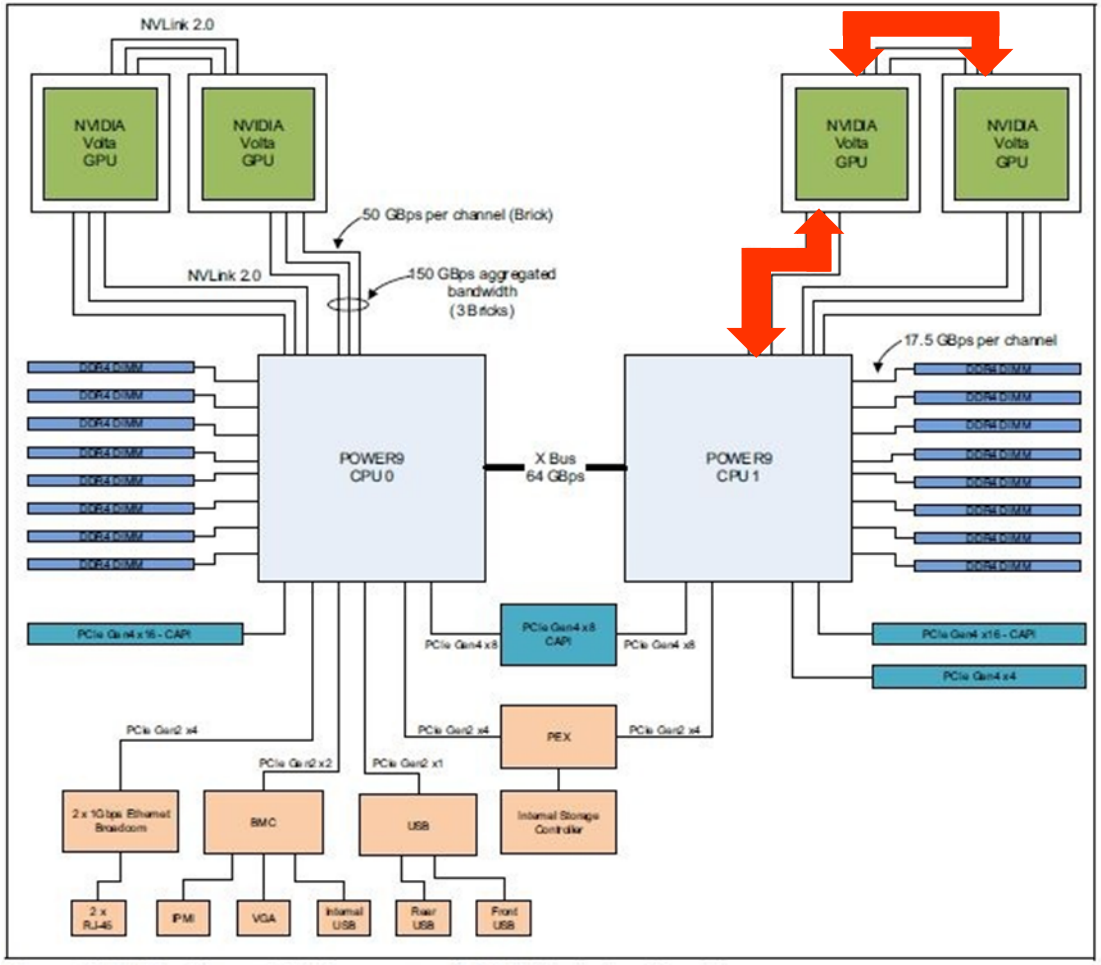


- In order to guarantee a more stable production the GPU frequency was manually set to 1250 MHz (target frequency 1312 MHz).
- A maximum performance of ~6,736 TFlops (~86 % of the theoretical peak performance — 7,800 TFlops), SKL = ~2 TFlops.

April 30, 2020

NVLink bandwidth tests

NVLink 2.0



Bandwidth:

3 memory channels

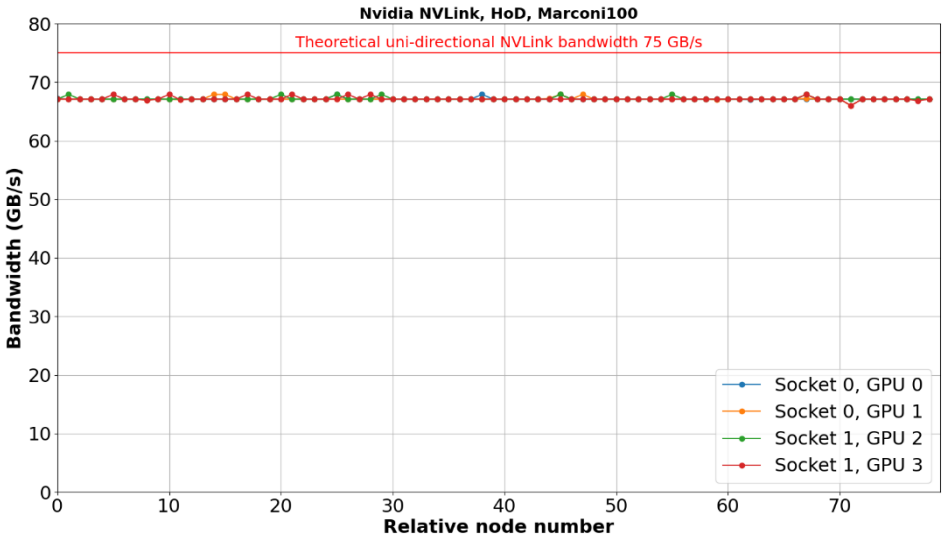
50 GB/s bi-directional / channel

150 GB/s bi-directional / connection

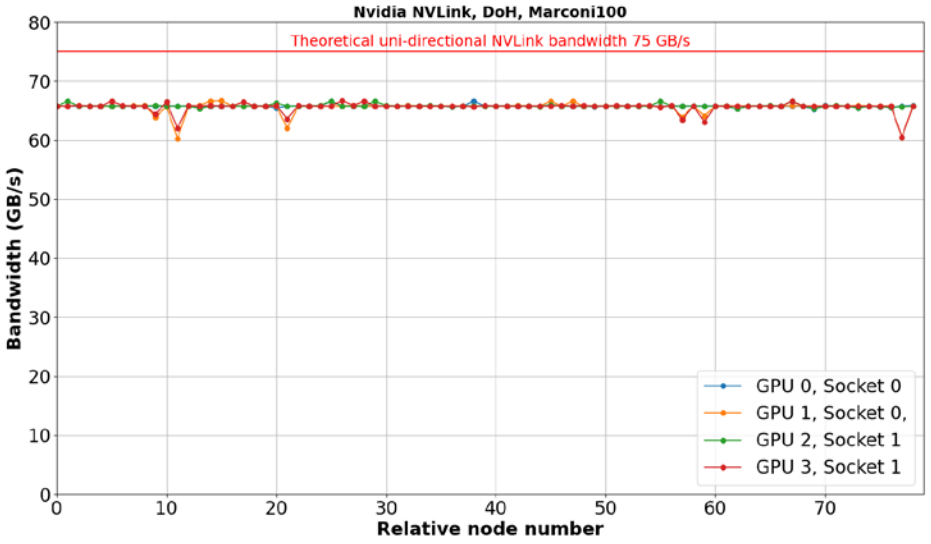
75 GB/s uni-directional / connection

NVLink benchmark on Marconi100 GPU

CPU to GPU



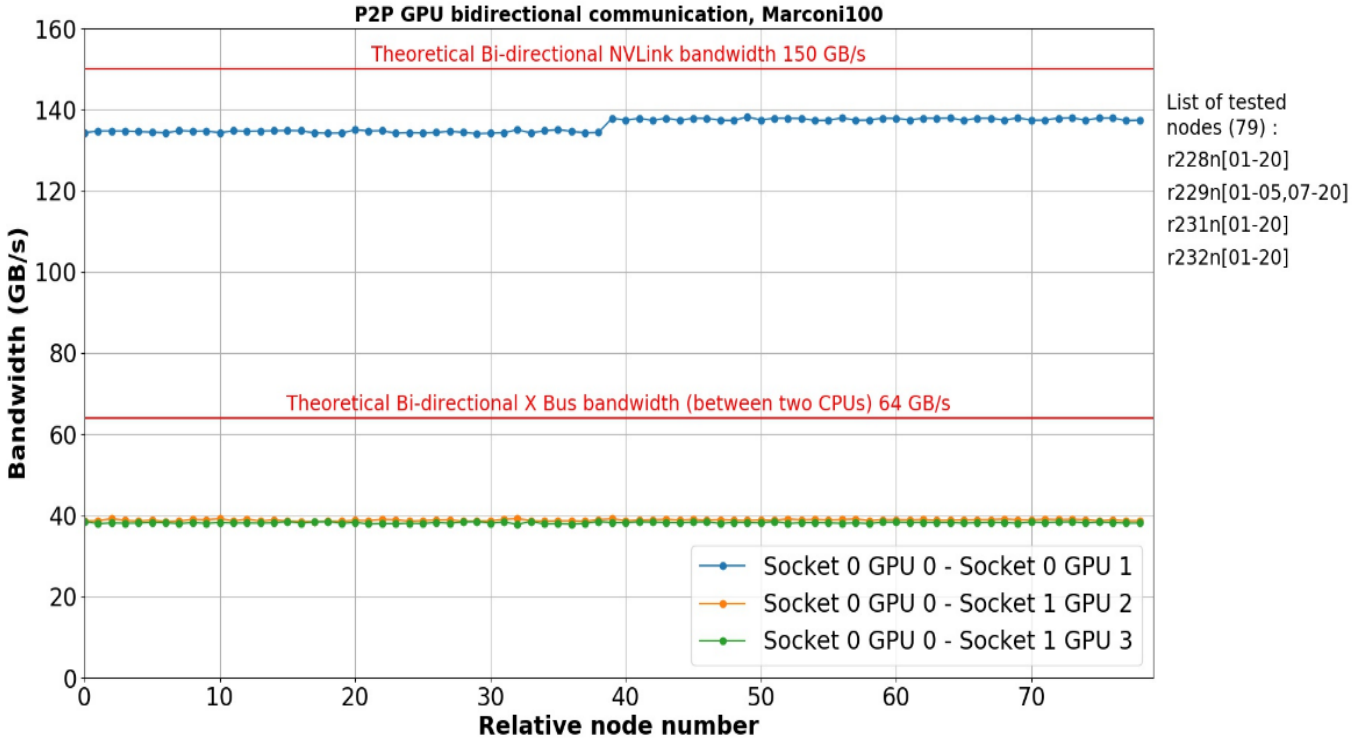
GPU to CPU



- **The results are stable and symmetric between each socket and its two directly connected GPU's.**
- **The same high average uni-directional bandwidth of ~67.2 GB/s was measured for all such CPU/GPUs pair combinations (~90 % of the theoretical value).**

May 6, 2020

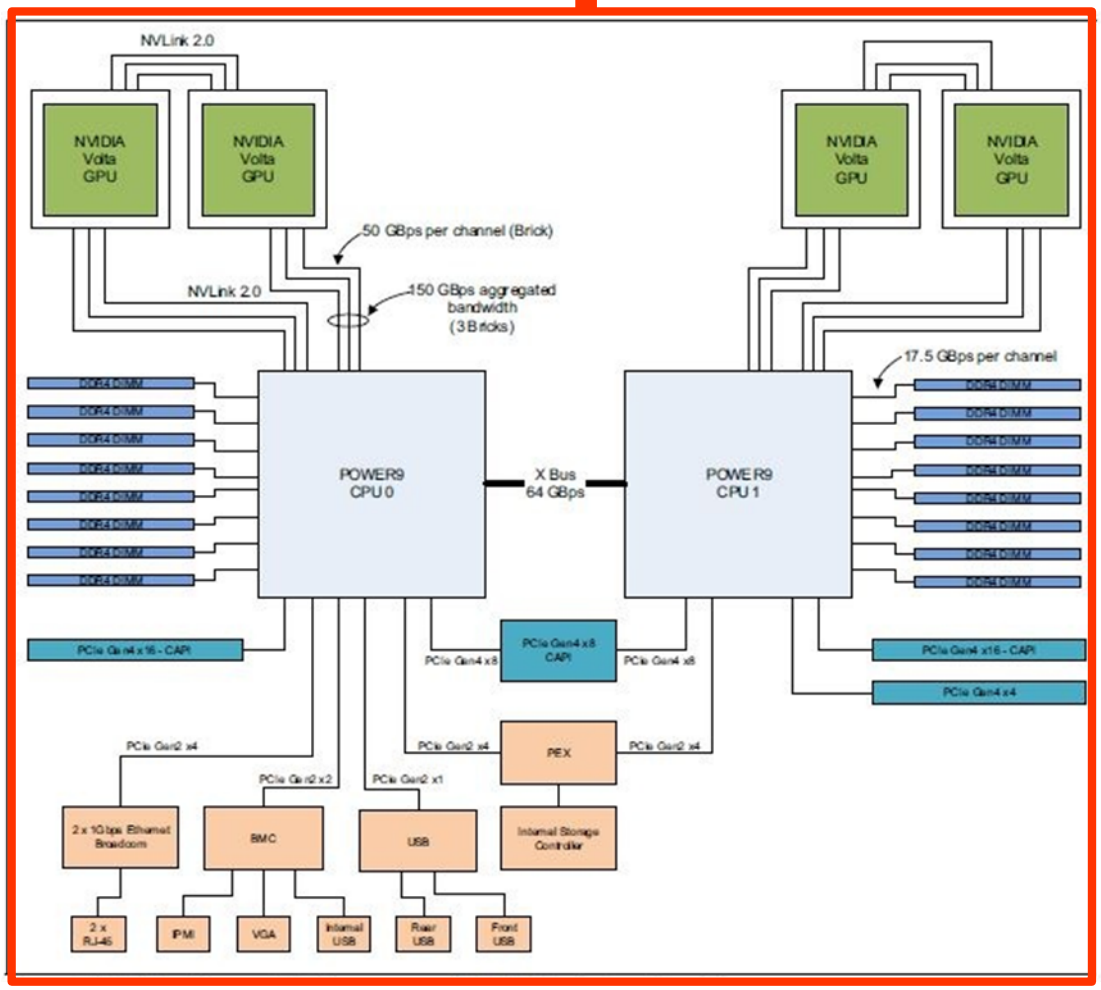
NVLink benchmark GPU to GPU



- **The results are stable and symmetric.**
- **The average intra socket bandwidth between two GPUs is ~136 GB/s that is ~90 % of the theoretical value.**
- **The average inter socket bandwidth between two GPUs is ~39 GB/s that is ~60 % of the theoretical value.**

May 1, 2020

Marconi100 inter node network tests



Inter node (node – node):

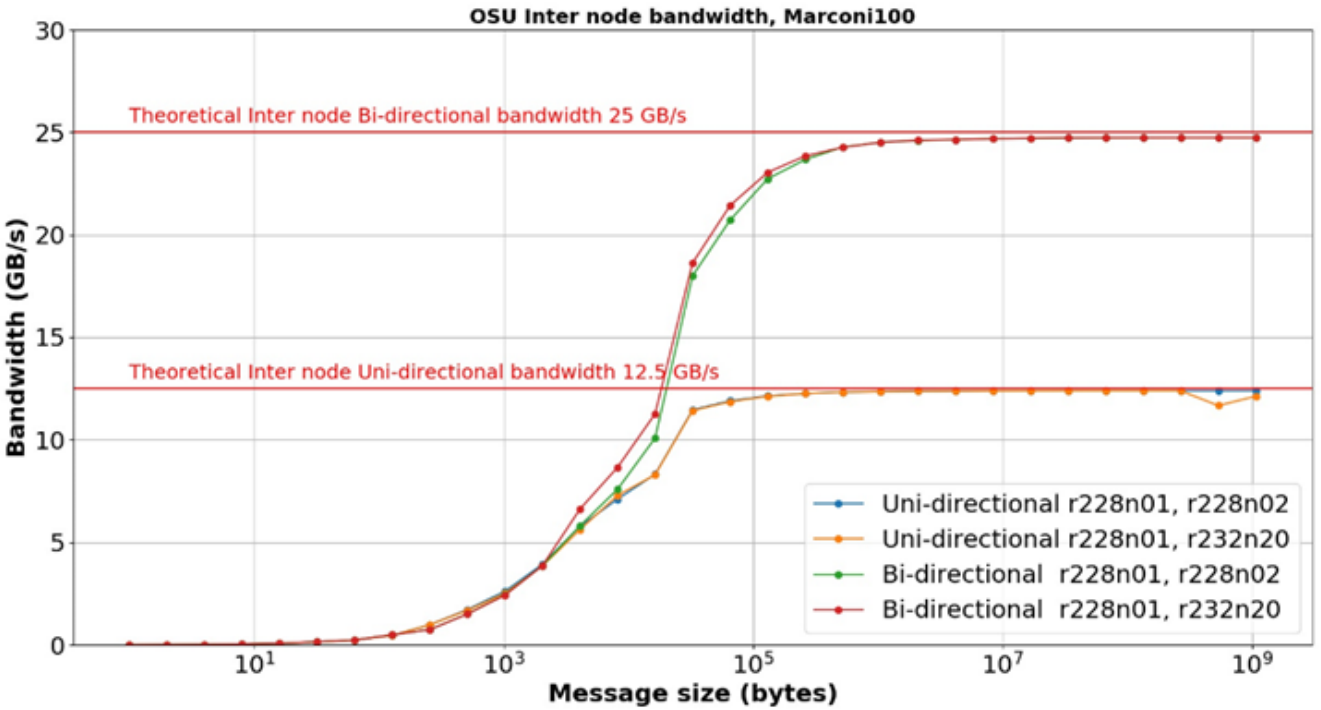
Mellanox IB EDR DragonFly+

Bandwidth:

100 Gb/s (12.5 GB/s) bi-directional bandwidth

Inter node network bandwidth

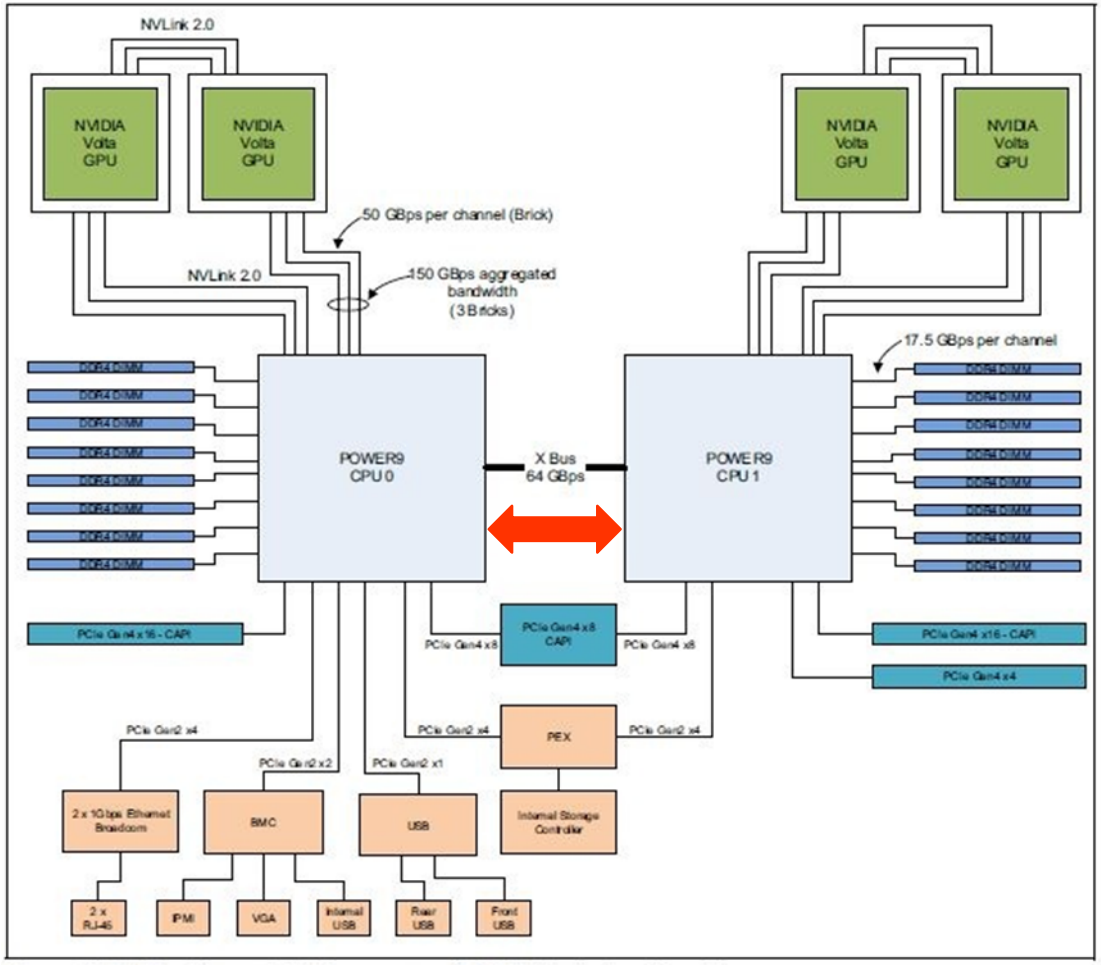
Ohio State University (OSU) microbenchmark



- **Stable and high bandwidth for uni- and bi-directional data transfer.**
- **99 % percent of the theoretical value was reached for the bi-directional test and 100 % was measured for the uni-directional benchmark.**
- **No difference in the results for both '*close*' and '*remote*' node combinations.**

Marconi100 intra node network tests

CPU – CPU: X Bus

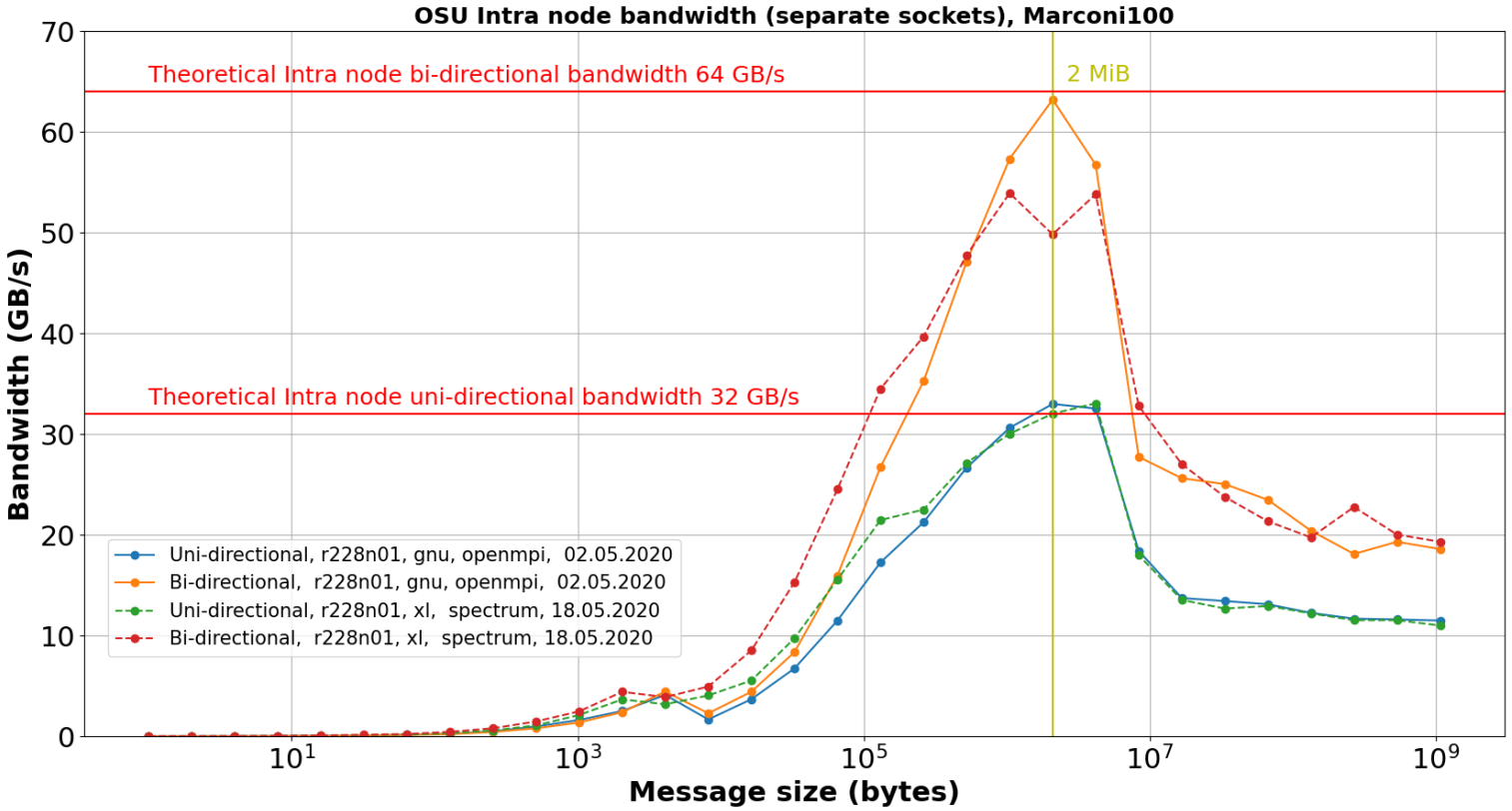


Bandwidth:

64 GB/s bi-directional data transfer

Intra node network bandwidth

OSU microbenchmark

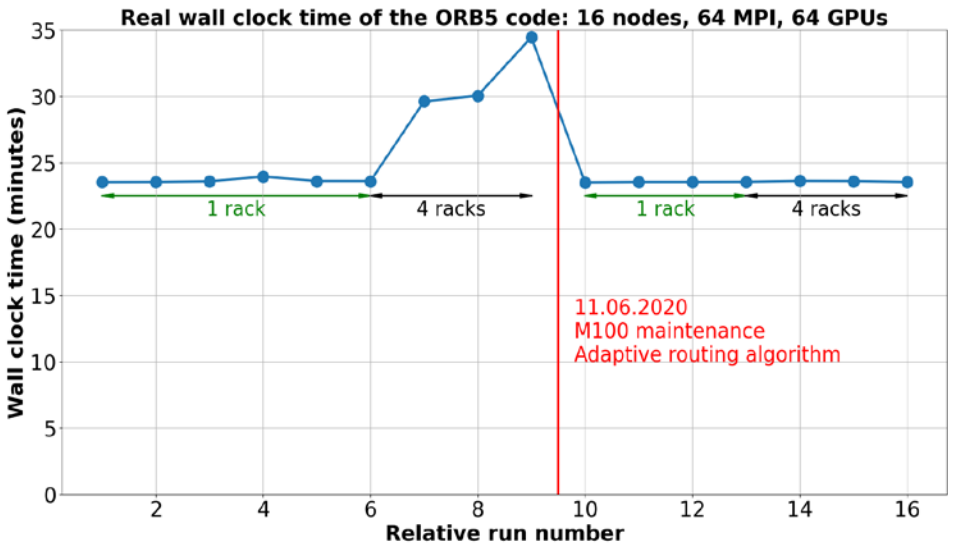


- **After reaching the maximum for a message size of 2 MiB the bandwidth drops for both uni- and bi-directional data transfer.**

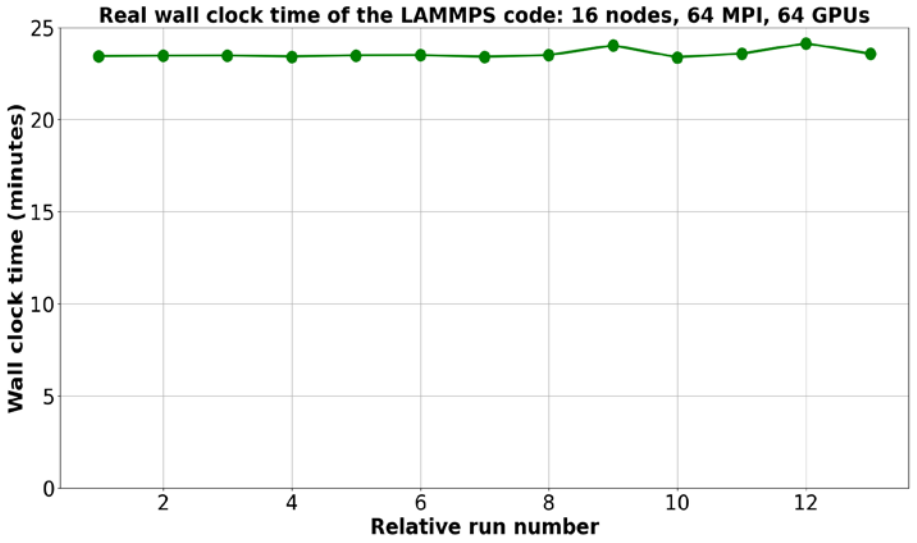
Real code stability performance

16 nodes, 64 MPI, 64 GPUs

ORB5



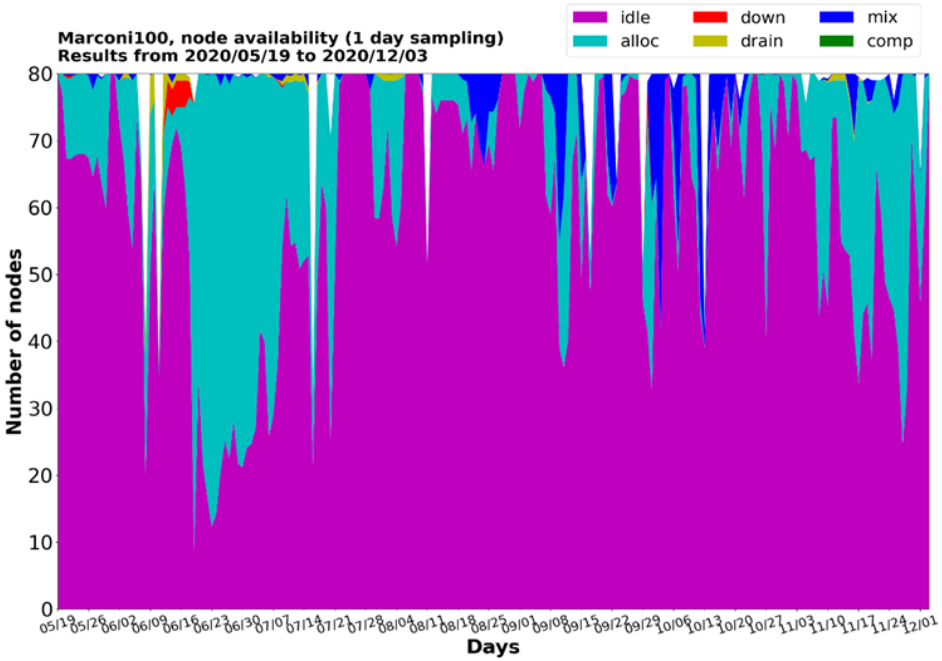
LAMMPS



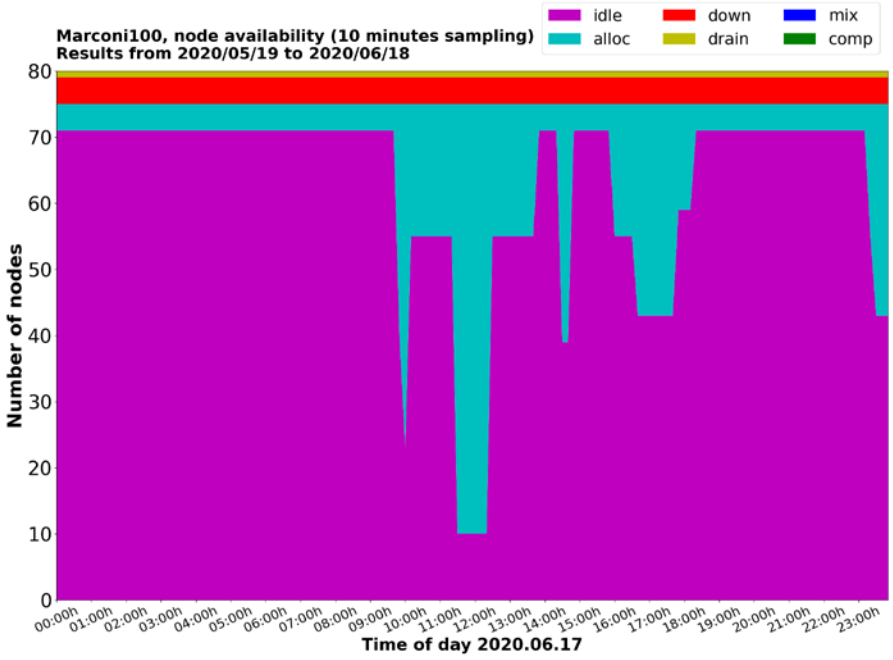
- The execution time is **stable** after M100 maintenance (UFM – Unified Fabric Management update: adaptive routing algorithm).
- Thanks to **Thomas Hayward-Schneider** and **Mariella Ippolito** for helping with the codes.

Marconi100 node usage

aggregates data over one day



high temporal resolution of 10 minutes



- **The partition uses less than 25 % of its capacity. Most of the time the nodes stay idle.**
- **Most of the time only few users launch their applications.**

December 3, 2020

Thank you for our attention!