Experiments with MPI benchmarks on Intel GPU cards

Kacper Kornet

Research Computing Services, University of Cambridge

Acknowledgments

Research Computng Services and Cambridge Open Zettascale Lab, University of Cambridge

Intel

Disclaimer

Photo by Kevan / CC BY

Hardware

Test node:

- 2 x Intel(R) Xeon(R) Platinum 8470Q (104 cores)
- 512GB of RAM
- 4 x Intel(R) Data Center GPU Max 1550 (128GB of HBM)
- 4 x HDR 200

Hardware

A100 node:

- 2 x AMD EPYC 7763 (64 cores)
- 1TB of RAM
- 4 x A100 (80GB of HBM)
- 2 x HDR 200

Software stack

	Intel	Nvidia
compiler	ifx 2023.1	
MPI	Intel MPI 2021.10.0	NVHPC 23.7

Benchmarks

- Intel MPI Benchmarks 2021.3
- modified OSU Micro-Benchmarks 7.2 (SYCL buffers)

Lower level software stack for Intel

- bundled libfabric (1.18.0-impi)
- fabric provider PSM3 v3.0-oneapi-ze built for IEFS OFA DELTA 11_5_0_0
- kernel rendezvous module with PVC support

GPU aware Intel MPI

Environment

- I_MPI_OFFLOAD=1
- I_MPI_OFFLOAD_RDMA=1
- I_MPI_OFFLOAD_IPC=1
- PSM3_GPUDIRECT=1
- PSM3_ONEAPI_ZE=1
- PSM3_RDMA=1

Counting GPU cards

- **Nvidia**: logical devices == physical devices
- **AMD MI200**: logical devices == 2x physical devices
- **Intel PVC**: chosen by user

Counting GPU cards

- **Nvidia**: logical devices == physical devices
- **AMD MI200**: logical devices == 2x physical devices
- **Intel PVC**: chosen by user

Counting GPU cards

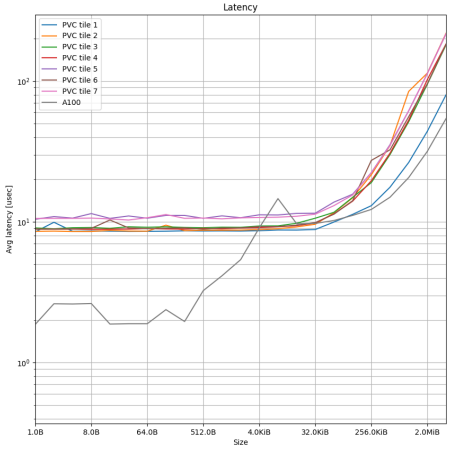
- **Nvidia**: logical devices == physical devices
- **AMD MI200**: logical devices == 2x physical devices
- **Intel PVC**: chosen by user

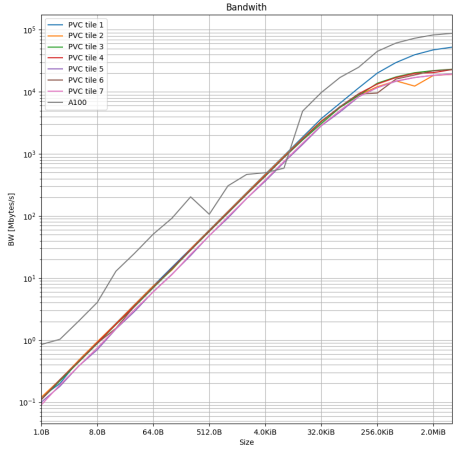
Hardware topology

```
[0] MPI startup(): ===== GPU topology on pvc-r-1 =====
[0] MPI startup(): NUMA nodes : 2
[0] MPI startup(): GPUs       : 4
[0] MPI startup(): Tiles      : 8
[0] MPI startup(): NUMA Id GPU Id         Tiles                          Ranks on this NUMA
[0] MPI startup(): 0       0,1            (0,1)(2,3)                     0,1
[0] MPI startup(): 1       2,3            (4,5)(6,7)
[0] MPI startup(): ===== GPU pinning on pvc-r-1 =====
[0] MPI startup(): Rank Pin tile
[0] MPI startup(): 0 {0}
[0] MPI startup(): 1 {1}
```

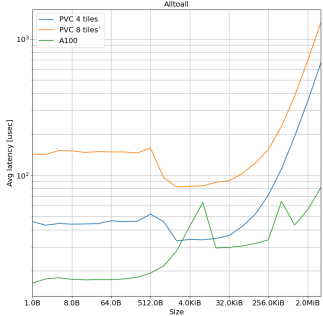
Environment variables: *I_MPI_OFFLOAD_TOPOLIB*, *I_MPI_OFFLOAD_CELL*, *I_MPI_OFFLOAD_DOMAIN_SIZE*, *I_MPI_OFFLOAD_DEVICES*, *I_MPI_OFFLOAD_CELL_LIST*, *I_MPI_OFFLOAD_DOMAIN*

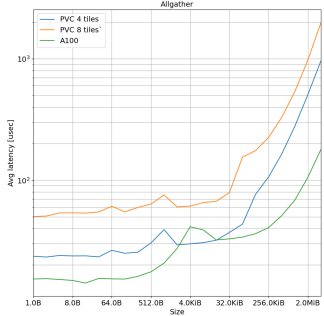
Latency and bandwidth

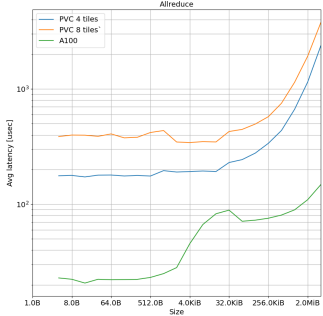




Collectives







Future plans

- multinode benchmarks
- benchmark applications
- benchmarks with MVAPICH
- finish extensions to OSU micro-benchmarks