MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

Analyzing the Capabilities of Your System Using OSU Microbenchmarks

A Tutorial at MUG'24

Presented by

Hari Subramoni, Aamir Shafi, and Akshay Paniraja Guptha

The MVAPICH Team

The Ohio State University

http://mvapich.cse.ohio-state.edu/

OSU Micro Benchmarks v7.4

- New features since MUG'23
  - **Add support for RCCL benchmarks.**
    - **Pt2pt, Collective**
  - **Add new benchmarks for persistent collectives.**
  - **Add new benchmarks to measure network congestion.**
    - **osu_bw_fan_in, osu_bw_fan_out**
  - **Add support for custom percentile values to evaluate benchmark performance.**
  - **Add support to log validation failures.**
  - **Add new collective benchmarks**
    - **osu_reduce_scatter_block, osu_ireduce_scatter_block**

OMB Releases since MUG'23

- OSU Micro Benchmarks v7.3 (10/30/2023)
- OSU Micro Benchmarks v7.4 (04/26/2024)

OMB New Features

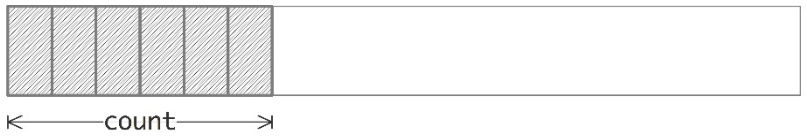
- Support for New Benchmarks
  - Benchmarks to measure network congestion
  - RCCL point-to-point and collective benchmarks
  - Benchmarks for Persistent collectives
  - osu_reduce_scatter_block, osu_ireduce_scatter_block
- Feature Enhancements
  - Support to log validation failure results
  - Support percentile values to evaluate benchmark performance

Using Derived Data Types (DDT) in OMB

OMB benchmarks now support derived data types enabled using '-D' option.

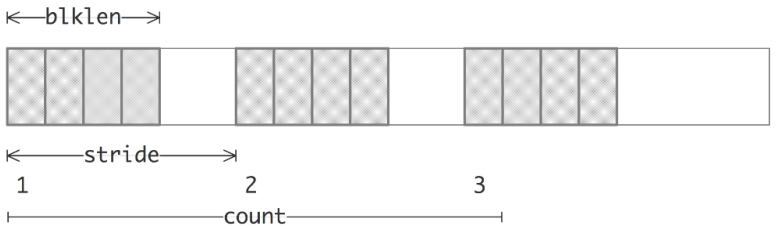
Contiguous

- -Dcont
- E.g: ./osu_allgather -Dcont



Vector

- -Dvect:[stride]:[block_length]
- E.g: ./osu_allgather –Dvect:6:4
  - Stride: 6
  - Block_length: 4



Indexed

- -Dindx:[ddt file path]

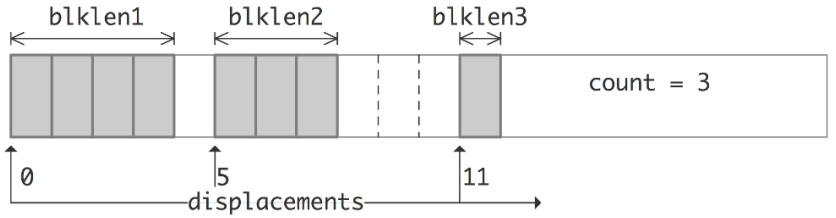


Image Source: **https://www.codingame.com/playgrounds/47058/have-fun-with-mpi-in-c/derived-datatypes.**

Sample Output/Input for DDT Support

- ./osu_allgather -Dvect:4:2

```
# OSU MPI Allgather Latency Test v7.2
# Datatype: MPI_CHAR.
# Size       Avg Latency(us)     Transmit Size
1                         1.10                 0
2                         1.09                 0
4                         1.45                 2
8                         1.52                 4
16                        1.57                 8
32                        2.06                16
64                        2.30                32
128                       2.32                64
256                       2.93               128
512                                          256
1024                                         512
2048                      8.30              1024
4096                     14.51              2048
8192                     27.02              4096
16384                    52.06              8192
32768                   117.76             16384
65536                   229.63             32768
131072                  439.85             65536
262144                  838.70            131072
524288                 1665.82            262144
1048576                3193.05            524288
```

Actual number of bytes transferred

- Indexed DDT parameters can be configured in a file as shown below.
- ./osu_allgather -Dindx:$OMB_HOME/c/util/ddt_sample.txt

```
#This is a comment
#Values must be number of elements.
#Displacement, Block Length
2, 10
12, 5
20, 4
```

Sample indexed DDT config file.

Enabling Plotting Support in OMB

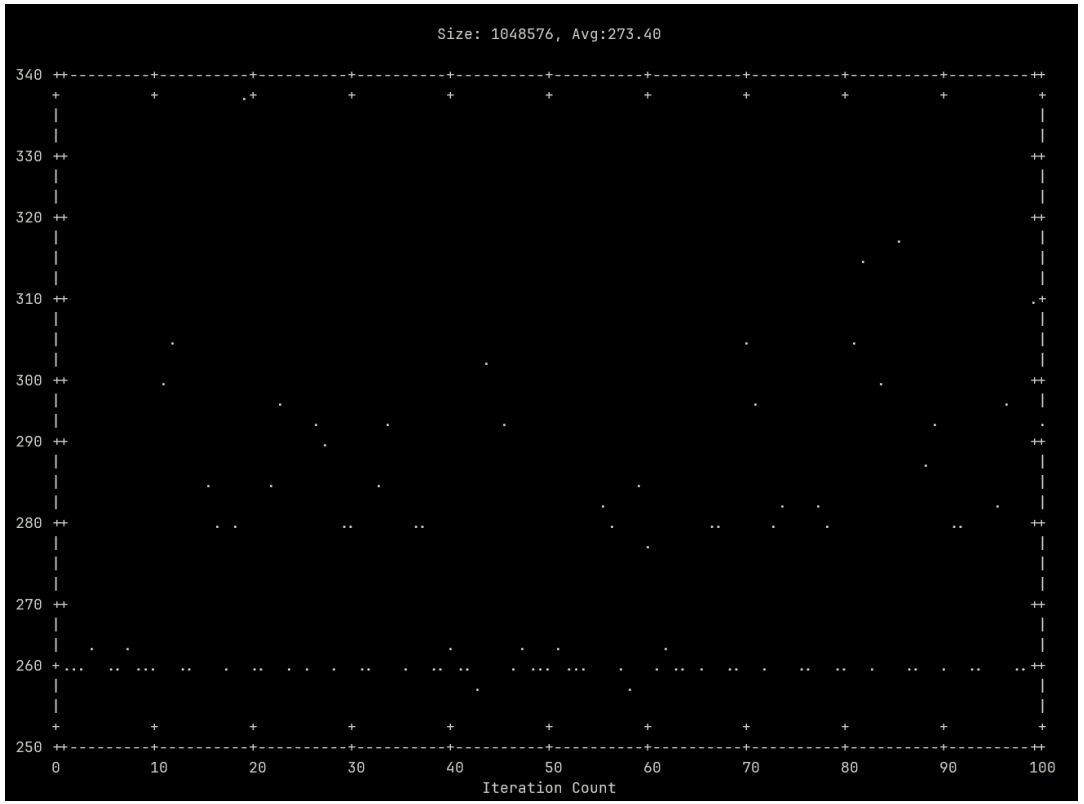
- Graphs of latency/bandwidth across iterations can now be plotted directly from OMB.
- Depends on ‘gnuplot’ to plot graphs.
  - If not in PATH, configure with --with-gnuplot=<path to gnuplot install dir>
- Depends on ‘convert’ to get output in pdf format.
  - If not in PATH, configure with --with-convert=<path to ImageMagick install dir> .
- Support enabled with -G, --graph [tty,png,pdf]

  E.g: ./osu_allgather -Gtty,png

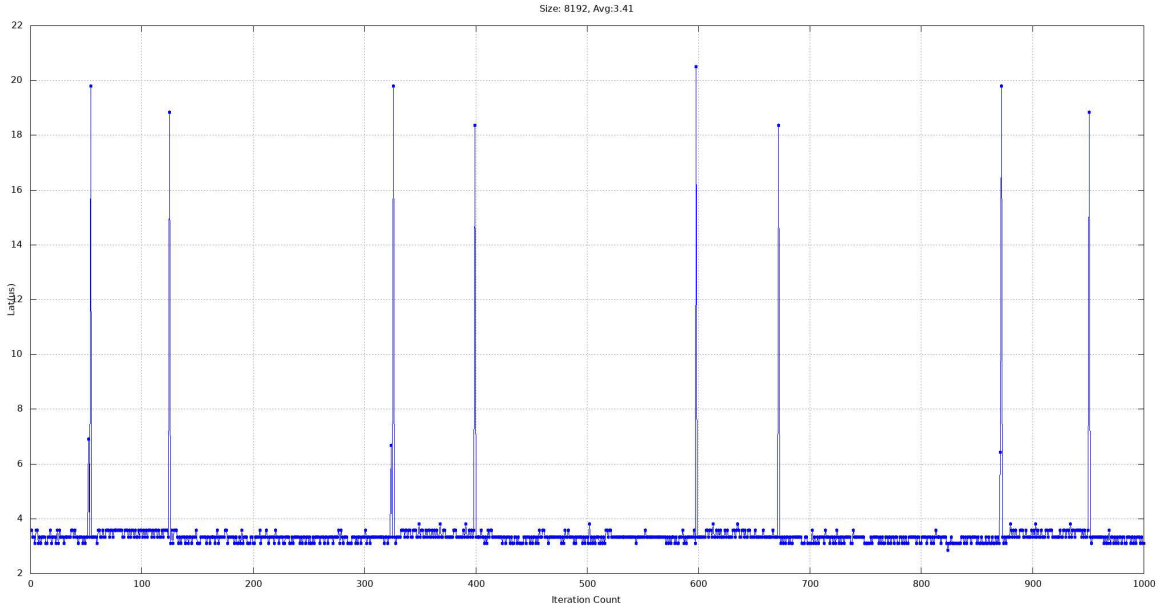
  ./osu_allgather -Gpng

Sample Plot Outputs from OMB

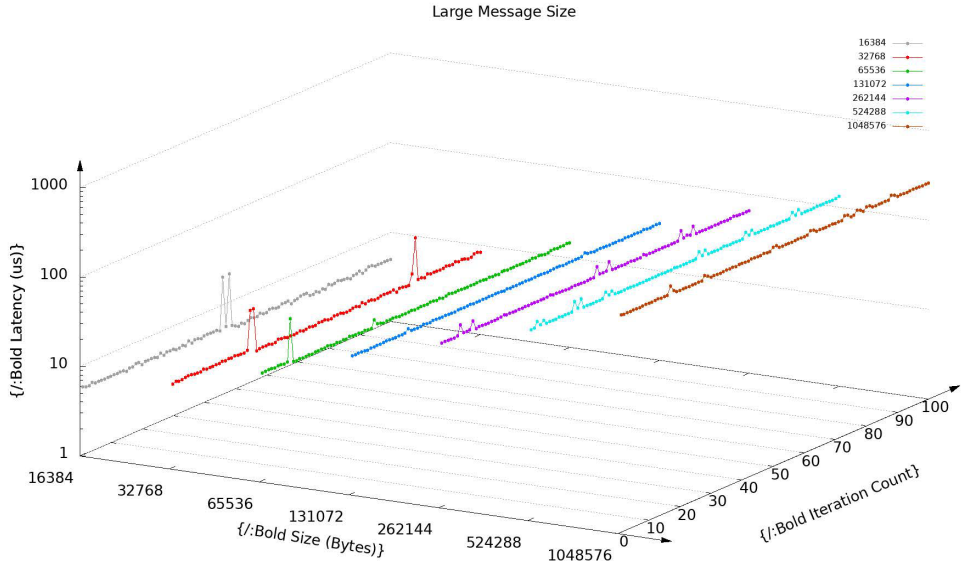
- **Terminal plot** – basic plot with necessary information
- **png, pdf** – detailed plots with 3D plots for smaller and large message sizes.



Terminal output of iterations(X) vs latency(Y) (-Gtty)



PNG/PDF output of iterations(X) vs latency(Y) (-Gpng,pdf)



PNG/PDF 3D output across message sizes (-Gpng,pdf)

Enabling PAPI Support in OMB

- OMB now supports Performance Application Programming Interface(PAPI) used for collecting performance counter information from various hardware and software components.
- Configured with --enable-papi --with-papi=<PAPI install path>
- -P, --papi [EVENTS]:[PATH] Enable PAPI support
  - [EVENTS] //Comma separated list of PAPI events
  - [PATH] //PAPI output file path

Using PAPI with OMB

- E.g: ./osu_allreduce -PPAPI_L1_DCM,PAPI_TLB_DM,PAPI_FML_INS:papi.out

```
Size: 1
>>================================================================>>
PAPI Event Name                Rank:0            Rank:1
PAPI_L1_DCM                     14433             13555
PAPI_TLB_DM                     13560             11195
PAPI_FML_INS                     2000              2000
##================================================================##

Size: 2
>>================================================================>>
PAPI Event Name                Rank:0            Rank:1
PAPI_L1_DCM                     14304             13204
PAPI_TLB_DM                     13726             12322
PAPI_FML_INS                     2000              2000
##================================================================##

Size: 4
>>================================================================>>
PAPI Event Name                Rank:0            Rank:1
PAPI_L1_DCM                     14743             14561
PAPI_TLB_DM                     13521             12737
PAPI_FML_INS                     2000              2000
##================================================================##
```

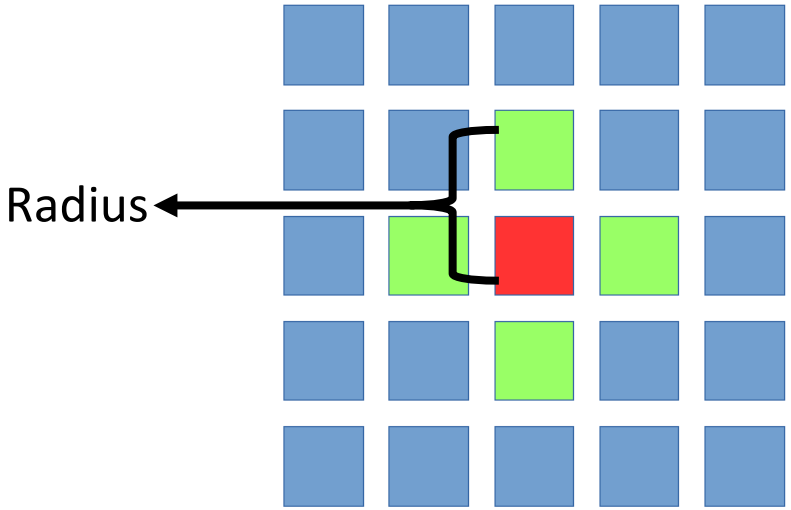
Sample PAPI output file(papi.out).

Support for Neighborhood Collectives in OMB

Cartesian

- -N cart:<num of dimensions:radius>
- E.g: ./osu_neighbor_allgather -N cart:2:1

Cartesian Neighborhood



Graph

- -N graph:<adjacency graph file>

Graph Neighborhood

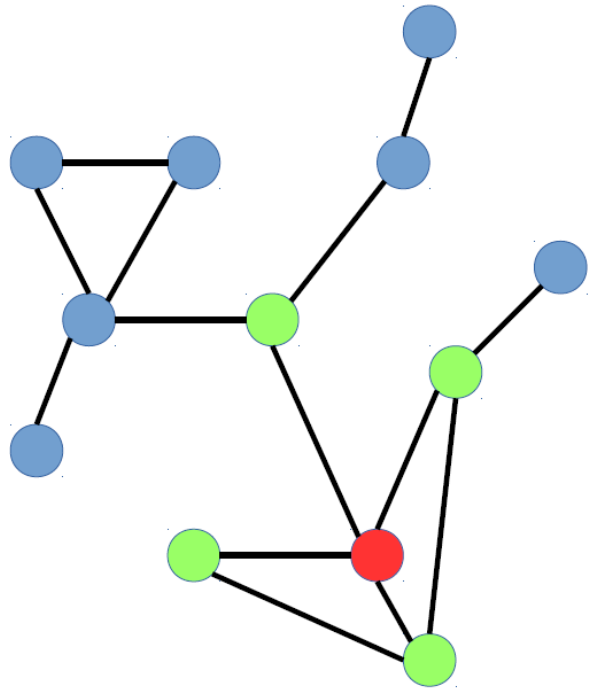


Image Source: **https://cvw.cac.cornell.edu/mpiadvtopics/process-topologies/neighborhood-collectives.**

Running Neighborhood Collectives in OMB

- ./osu_neighbor_allgather -N cart:2:1

```
Dimensions size = 4 4
Time took to create topology graph:52.01 us.

# OSU MPI Neighborhood Allgather Latency Test v7.2
# Datatype: MPI_CHAR.
# Size       Avg Latency(us)
1                       3.95
2                       4.00
4                       4.05
8                       4.10
16                      4.11
32                      4.87
64                      5.43
128                     7.10
256                     7.53
512                     5.96
1024                    7.88
2048                   12.89
4096                   12.45
8192                   24.95
16384                  74.66
32768                  37.60
65536                  61.97
131072                103.27
262144                280.24
524288                703.98
1048576              2491.98
```

- ./osu_neighbor_allgather -N graph:$OMB_HOME/c/util/nhbrhd_graph.adj

```
#This is a comment
#All values are ranks of the process
#Source, Destination
2, 0
0, 1
1, 2
2, 3
1,3
0,3
```

Sample adjacency graph file.

Using MPI Data Types with OMB

- OMB now supports the following MPI datatypes,
  - MPI_CHAR
  - MPI_FLOAT
  - MPI_INT
- MPI Data Type can be set using '-T' option.
  - -T<all,mpi_char,mpi_int,mpi_float>
  - E.g: ./osu_allgather –Tmpi_int

- ./osu_allgather -Tall -m :64

```
# OSU MPI Allgather Latency Test v7.2
# Datatype: MPI_CHAR.
# Size       Avg Latency(us)
1                       3.86
2                       2.83
4                       4.20
8                       4.71
16                      5.28
32                      6.49
64                      8.63
# Datatype: MPI_INT.
# Size       Avg Latency(us)
4                       3.91
8                       4.33
16                      5.09
32                      6.38
64                      8.39
# Datatype: MPI_FLOAT.
# Size       Avg Latency(us)
4                       3.79
8                       4.33
16                      4.86
32                      6.11
64                      8.29
```

Support percentile values to evaluate benchmark performance

- Benchmarks have been extended to support the following additional metrics :

  "-z" Outputs P99, P90, P50 percentiles"

  "-z<1-99,1-99,1-99..>" Comma seperated percentile range

- ./osu_allreduce -z80,90,98

```
# OSU MPI Allreduce Latency Test v7.4
# Datatype: MPI_INT.
# Size       Avg Latency(us)  P80 Tail Lat(us)  P90 Tail Lat(us)  P98 Tail Lat(us)
4                       1.82              1.91              2.15              2.98
8                       1.88              1.91              2.15              2.98
16                      1.88              1.91              2.03              2.98
32                      1.85              1.91              2.15              3.34
64                      1.87              1.91              2.03              2.86
128                     2.12              2.15              2.38              3.10
256                     2.35              2.74              2.86              3.34
```

Network congestion benchmarks

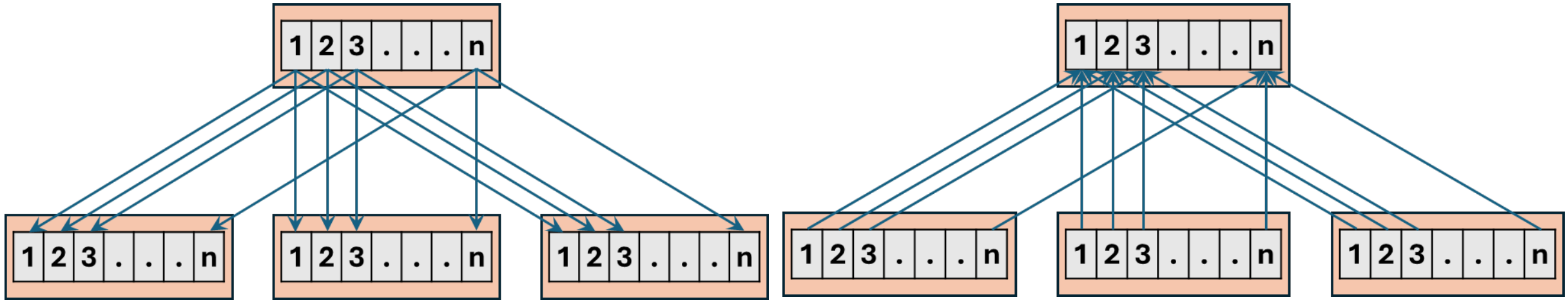
- Network Congestion bandwidth test evaluates the aggregate uni-directional bandwidth between multiple pairs of process across nodes.

- osu_bw_fan_out

- 1 sender node, n receiver nodes

- osu_bw_fan_in

- n sender nodes, 1 receiver node



Additional features in OMB

- MPI_IN_PLACE support
  - OMB not supports running benchmarks with MPI_IN_PLACE enabled by passing '-I' options.
  - E.g: ./osu_allgather --in-place
- MPI-4 session
  - Currently MPI-4 standards describes support for mpi://WORLD, mpi://SELF. Since most OMB benchmarks require more than one process, mpi://WORLD is set as default when running with MPI session support.
  - Enabled by passing '-I' option.
- Set root rank
  - OMB benchmarks support setting root rank for rooted collectives using '-k' option.
    - Fixed
      - Root rank is fixed for all iterations of the benchmark.

      E.g: ./osu_reduce -k fixed:1
    - Rotate
      - Root rank varies in a cyclic manner for each iteration on the benchmark.

      E.g: ./osu_reduce -k rotate

Additional features in OMB(cont.)

- Enabling RCCL support
  - Configure with

    "--enable-rcclomb --with-rccl=<path to RCCL>"
- Persistent collectives benchmarks
  - OMB auto-detects MPI-4 libraries at configure time and builds MPI-4 features and benchmarks by default.
  - MPI-4 support can also be enabled by configuring with "--enable-mpi4"
- Support to log validation failure results
  - Validation failures can now be saved into a file.
  - This feature can be enabled by using

    "-c log:<dir>"

OMB – Future Roadmap

- Support for new OpenSHMEM benchmarks to measure bandwidth.
  - osu_oshm_get_bw
  - osu_oshm_get_nb_bw
  - osu_oshm_put_bw
  - osu_oshm_put_nb_bw
- Support for Intel GPUs using oneAPI/SYCL.
- Add accelerator support for neighborhood collective benchmarks.
- Support for new partitioned pt2pt benchmarks.
  - osu_partitioned_latency
  - osu_partitioned_bw
  - osu_partitioned_bibw

Funding Acknowledgments

Funding Support by

Equipment Support by


Acknowledgments to all the Heroes (Past/Current Students and Staffs)

Current Students (Graduate)

- K. Al Attar (Ph.D.)
- N. Alnaasan (Ph.D.)
- Q. Anthony (Ph.D.)
- C.-C. Chun (Ph.D.)
- T. Chen
- N. Contini (Ph.D.)
- J. Hatef (Ph.D.)
- S. Lee (Ph.D.)
- B. Michalowicz (Ph.D.)
- B. Ramesh (Ph.D.)
- K. K. Suresh (Ph.D.)
- A. T. Tran (Ph.D.)
- A. Potlapally (Ph.D.)
- S. Xu (Ph.D.)
- L. Xu (Ph.D.)
- G. Kuncham (Ph.D.)
- J. Yao (P.hD.)
- J. Jani (M.S.)
- J. Queiser (M.S.)

Past Students

- A. Awan (Ph.D.)
- A. Augustine (M.S.)
- P. Balaji (Ph.D.)
- M. Bayatpour (Ph.D.)
- R. Biswas (M.S.)
- S. Bhagvat (M.S.)
- A. Bhat (M.S.)
- D. Buntinas (Ph.D.)
- L. Chai (Ph.D.)
- B. Chandrasekharan (M.S.)
- S. Chakraborthy (Ph.D.)
- N. Dandapanthula (M.S.)
- V. Dhanraj (M.S.)
- C.-H. Chu (Ph.D.)
- T. Gangadharappa (M.S.)
- K. Gopalakrishnan (M.S.)
- R. Gulhane (M.S.)
- J. Hashmi (Ph.D.)
- M. Han (M.S.)
- W. Huang (Ph.D.)
- A. Jain (Ph.D.)
- W. Jiang (M.S.)
- J. Jose (Ph.D.)
- M. Kedia (M.S.)
- K. S. Khorassani (Ph.D.)
- S. Kini (M.S.)
- M. Koop (Ph.D.)
- P. Kousha (Ph.D.)
- K. Kulkarni (M.S.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- K. Kandalla (Ph.D.)
- M. Li (Ph.D.)
- P. Lai (M.S.)
- J. Liu (Ph.D.)
- M. Luo (Ph.D.)
- A. Mamidala (Ph.D.)
- G. Marsh (M.S.)
- V. Meshram (M.S.)
- A. Moody (M.S.)
- S. Naravula (Ph.D.)
- R. Noronha (Ph.D.)
- X. Ouyang (Ph.D.)
- S. Pai (M.S.)
- S. Potluri (Ph.D.)
- K. Raj (M.S.)
- R. Rajachandrasekar (Ph.D.)
- D. Shankar (Ph.D.)
- G. Santhanaraman (Ph.D.)
- N. Sarkauskas (B.S. and M.S)
- V. Sathu (M.S.)
- N. Senthil Kumar (M.S.)
- A. Singh (Ph.D.)
- J. Sridhar (M.S.)
- S. Srivastava (M.S.)
- S. Sur (Ph.D.)
- H. Subramoni (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- A. Vishnu (Ph.D.)
- J. Wu (Ph.D.)
- W. Yu (Ph.D.)
- J. Zhang (Ph.D.)
- Q. Zhou (Ph.D.)
- J. Sulewski (B.S.)

Past Post-Docs

- D. Banerjee
- X. Besseron
- M. S. Ghazimirsaeed
- H.-W. Jin
- J. Lin
- M. Luo
- E. Mancini
- K. Manian
- S. Marcarelli
- A. Ruhela
- J. Vienne
- H. Wang

Current Research Scientists

- A. Shafi

Current Research Specialist

- R. Motlagh

Current Faculty

- H. Subramoni

Current Software Engineers

- N. Pavuk
- N. Shineman
- M. Lieber
- A-. Guptha

Past Research Scientists

- K. Hamidouche
- S. Sur
- X. Lu
- M. Abduljabbar

Past Senior Research Associate

- J. Hashmi

Past Programmers

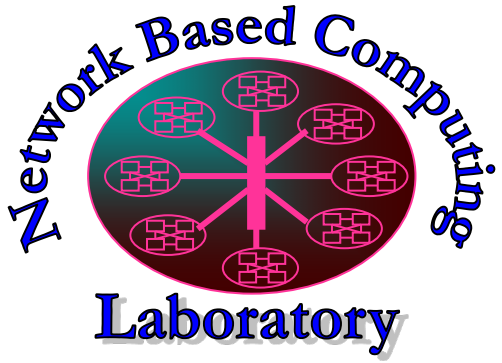
- A. Reifsteck
- D. Bureddy
- J. Perkins
- B. Seeds

Past Research Specialist

- M. Arnold
- J. Smith

Thank You!

panda@cse.ohio-state.edu, subramon@cse.ohio-state.edu



Follow us on

https://twitter.com/mvapich

Network-Based Computing Laboratory

http://nowlab.cse.ohio-state.edu/

MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

The High-Performance MPI/PGAS Project

http://mvapich.cse.ohio-state.edu/

HiBD

High-Performance Big Data

The High-Performance Big Data Project

http://hibd.cse.ohio-state.edu/

HiDL

High-Performance Deep Learning

The High-Performance Deep Learning Project

http://hidl.cse.ohio-state.edu/