Performance of MPI Communications on Nurion Utilizing MVAPICH2-X with XPMEM

Minsik Kim, Ph.D.
Supercomputing Infrastructure Center, KISTI

The International Conference for
High Performance Computing, Networking,
Storage, and Analysis (SC'19), OSU Booth

KISTI Korea Institute of Science and Technology Information
www.kisti.re.kr

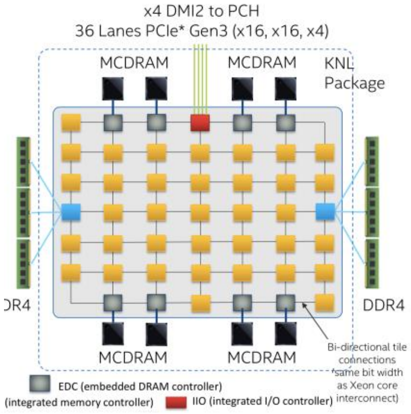
Introduction to KISTI-5 Supercomputer, Nurion

KISTI-5 Compute Nodes

The Largest KNL/OPA based commodity cluster System
Rpeak 25.7PFlops, Rmax 13.9PFlops

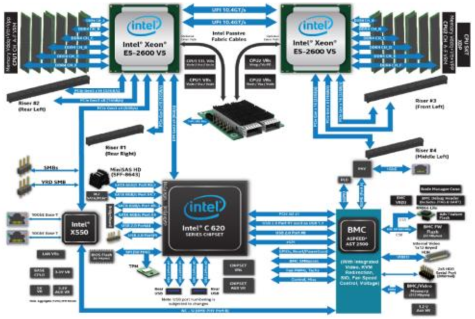
Compute nodes — 8,305 KNL Computing modules, 116 Racks, 25.3PF

- 1x Xeon Phi KNL 7250, 68Cores 1.4GHz, AVX512
- 3TFlops Peak, ~0.2 Bytes/Flops,
- 96GB (6x16GB) DDR4-2400 6 channel RAM,
- 16GB HBM (460GB/s)
- 1x 100Gbps OPA HFI, 1x On-board GigE Port

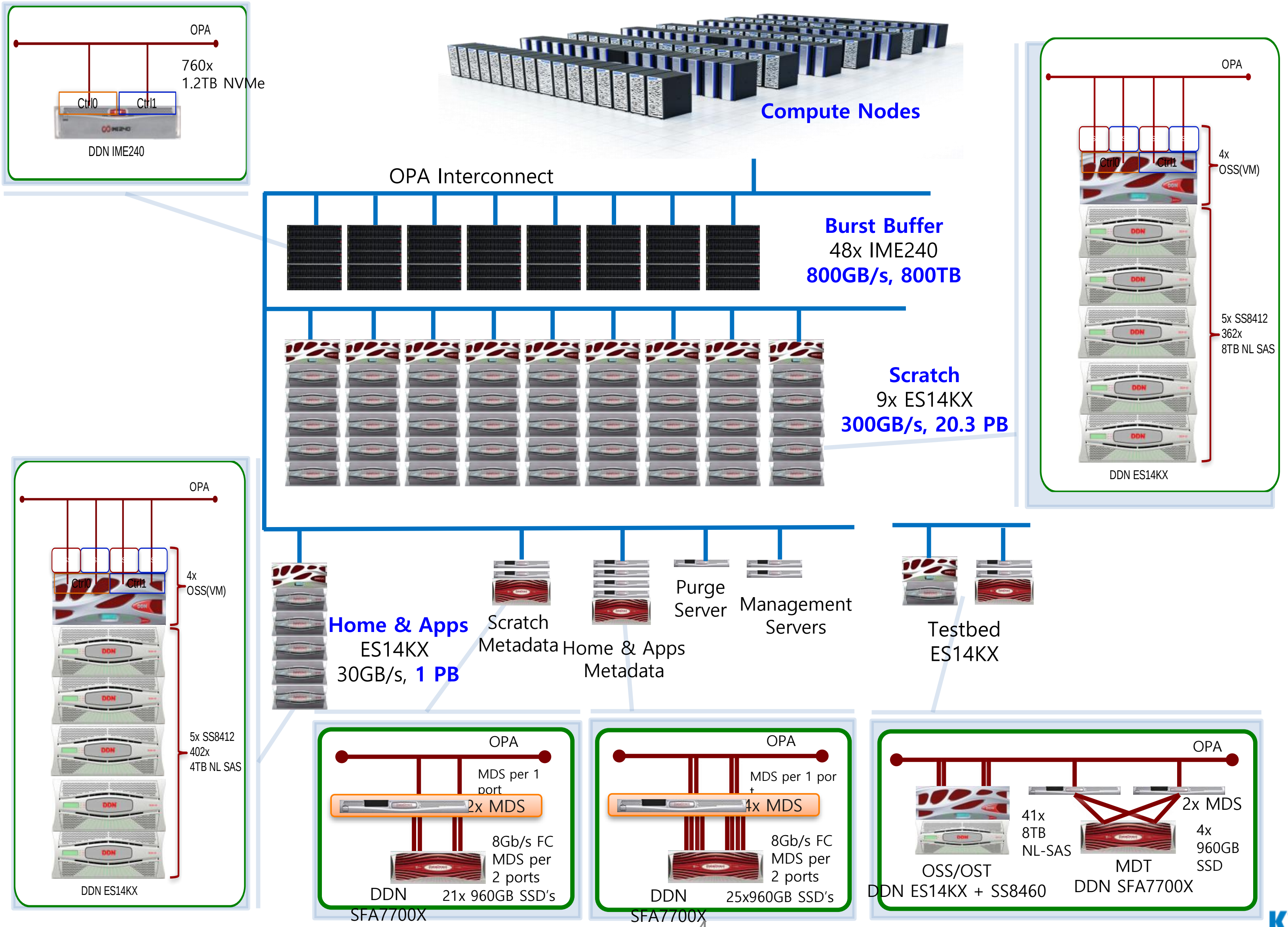

x4 DMI2 to PCH
36 Lanes PCIe* Gen3 (x16, x16, x4)
MCDRAM
MCDRAM
KNL Package
DR4
DDR4
MCDRAM
MCDRAM
Bi-directional tile connections same bit width as Xeon core interconnect)
EDC (embedded DRAM controller)
(integrated memory controller)
IIO (integrated I/O controller)

CPU-only nodes — 132 Skylake Computing modules, 4 Racks, 0.4PF

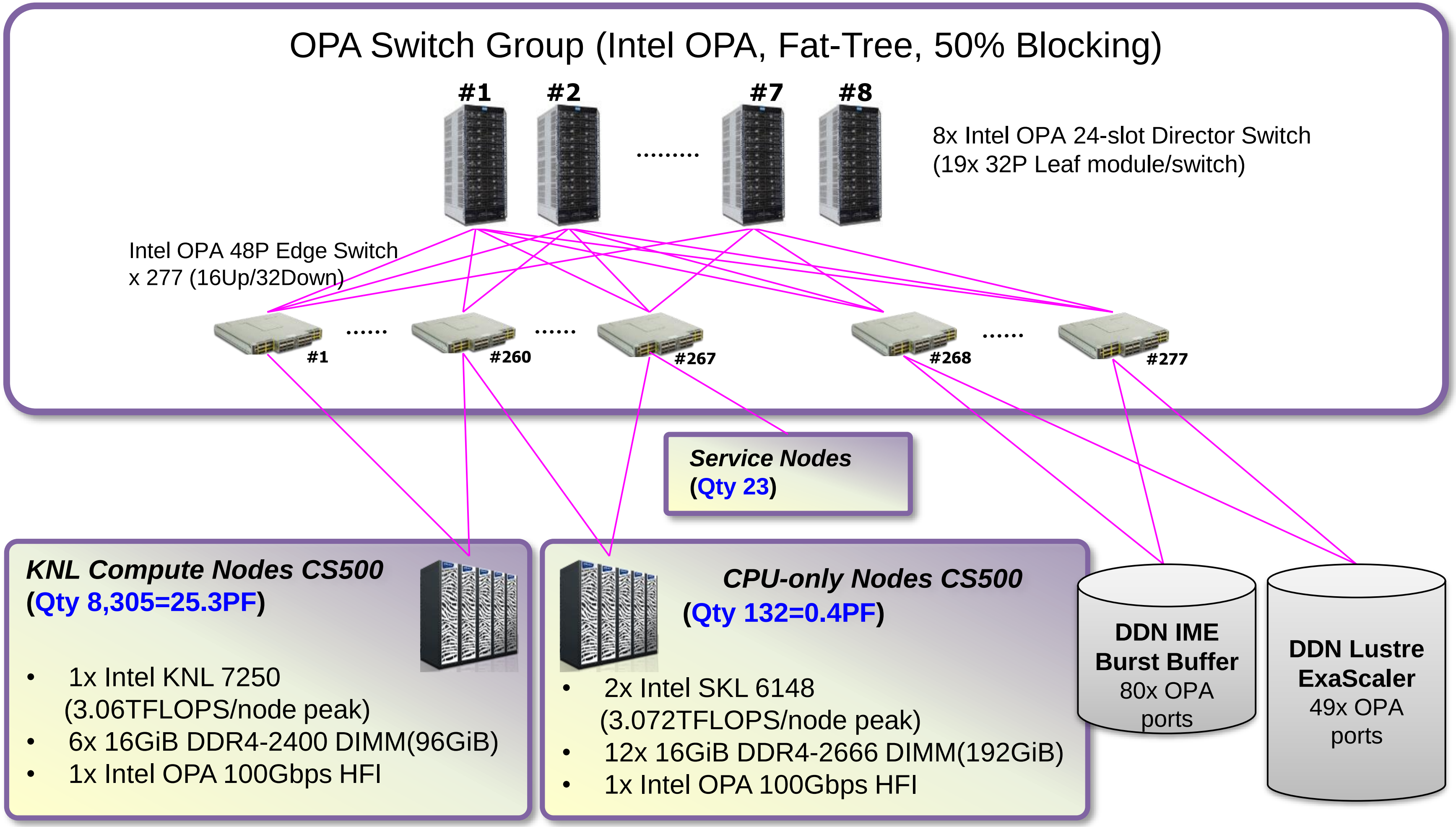
- 2x Xeon SKX 6148 CPUs, 2.4GHz, AVX512
- 192GB (12x 16GB) DDR4-2666 RAM
- 1x Single-port 100Gbps OPA HFI card
- 1x On-board GigE (RJ45) port



KISTI-5 Storage System



KISTI-5 OPA Interconnect



KNL Compute Nodes CS500
(Qty 8,305=25.3PF)

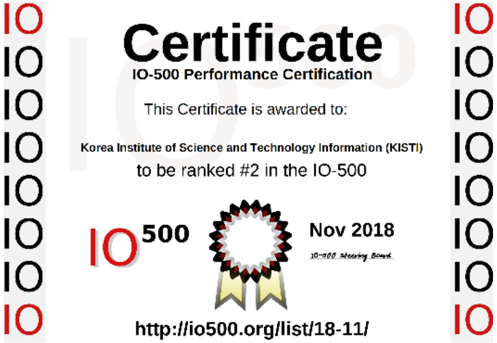
- 1x Intel KNL 7250 (3.06TFLOPS/node peak)
- 6x 16GiB DDR4-2400 DIMM(96GiB)
- 1x Intel OPA 100Gbps HFI

CPU-only Nodes CS500
(Qty 132=0.4PF)

- 2x Intel SKL 6148 (3.072TFLOPS/node peak)
- 12x 16GiB DDR4-2666 DIMM(192GiB)
- 1x Intel OPA 100Gbps HFI

Benchmark Performance Result

Category	Features	# of nodes	Score	World Ranking
HPL	Large-scale Dense Matrix Computation Used for Top500	8,174(KNL) + 122(SKX)	13.93PF	15th (Jun 2019)
HPCG	Large-scale Sparse Matrix Computation Similar to normal user applications	8,250(KNL)	0.39PF	8th (Jun 2019)
Graph500	Breadth-First Search, Single-Source Shortest Paths	1,024(KNL)	1,456GTEPS 337GTEPS	10Th (Jun 2019) 3rd (Jun 2019)
IO500	Various IO Workloads	2,048(KNL)	160.67	5th (Jun 2019)



Certificate

IO-500 Performance Certification

This Certificate is awarded to:

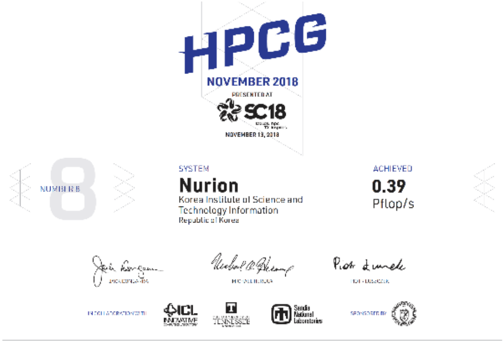
Korea Institute of Science and Technology Information (KISTI)

to be ranked #2 in the IO-500

Nov 2018

IO-500 Steering Board

http://io500.org/list/18-11/



HPCG

NOVEMBER 2018

PRESENTED AT SC18

NOVEMBER 13, 2018

NUMBER 8

SYSTEM

Nurion

Korea Institute of Science and Technology Information

Republic of Korea

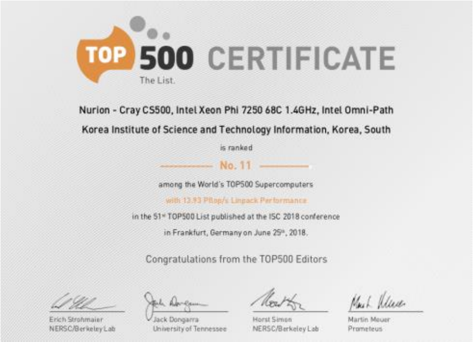
ACHIEVED

0.39 Pflop/s

JACK DONGARRA MICHAEL HEROUX PIOTR LUSZCZEK

IN COLLABORATION WITH

SPONSORED BY



TOP500 CERTIFICATE

The List.

Nurion - Cray CS500, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path

Korea Institute of Science and Technology Information, Korea, South

is ranked

No. 11

among the World's TOP500 Supercomputers

with 13.93 Pflop/s Linpack Performance

in the 51st TOP500 List published at the ISC 2018 conference

in Frankfurt, Germany on June 25th, 2018.

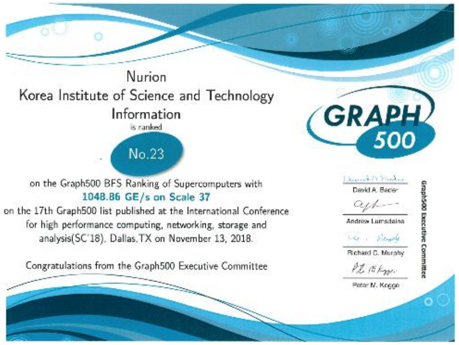
Congratulations from the TOP500 Editors

Erich Strohmaier NERSC/Berkeley Lab

Jack Dongarra University of Tennessee

Horst Simon NERSC/Berkeley Lab

Martin Meuer Prometeus



Nurion

Korea Institute of Science and Technology Information

is ranked

No.23

on the Graph500 BFS Ranking of Supercomputers with

1048.86 GE/s on Scale 37

on the 17th Graph500 list published at the International Conference for high performance computing, networking, storage and analysis(SC'18). Dallas,TX on November 13, 2018.

Congratulations from the Graph500 Executive Committee

David A. Bader

Andrew Lumsdaine

Richard C. Murphy

Peter M. Kogge

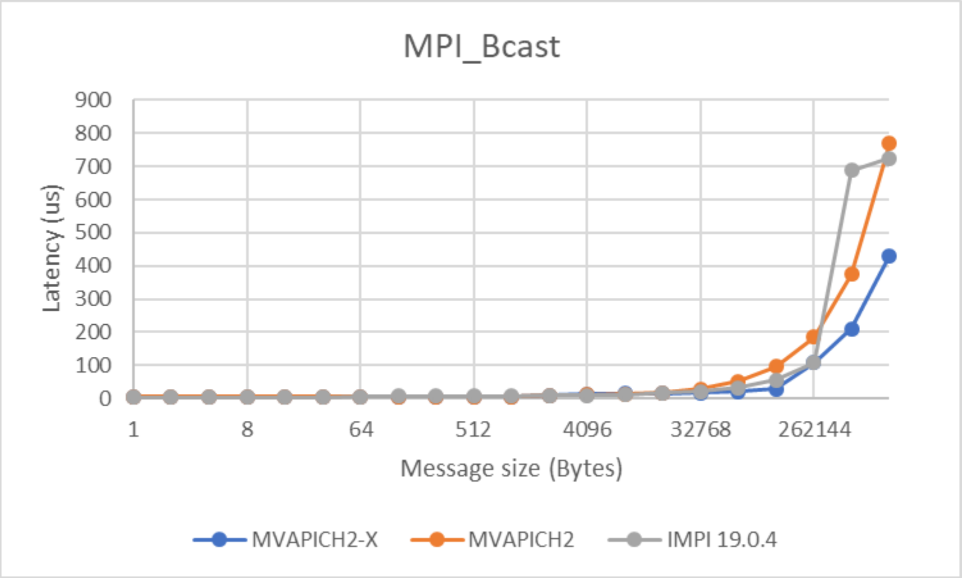
Graph500 Executive Committee

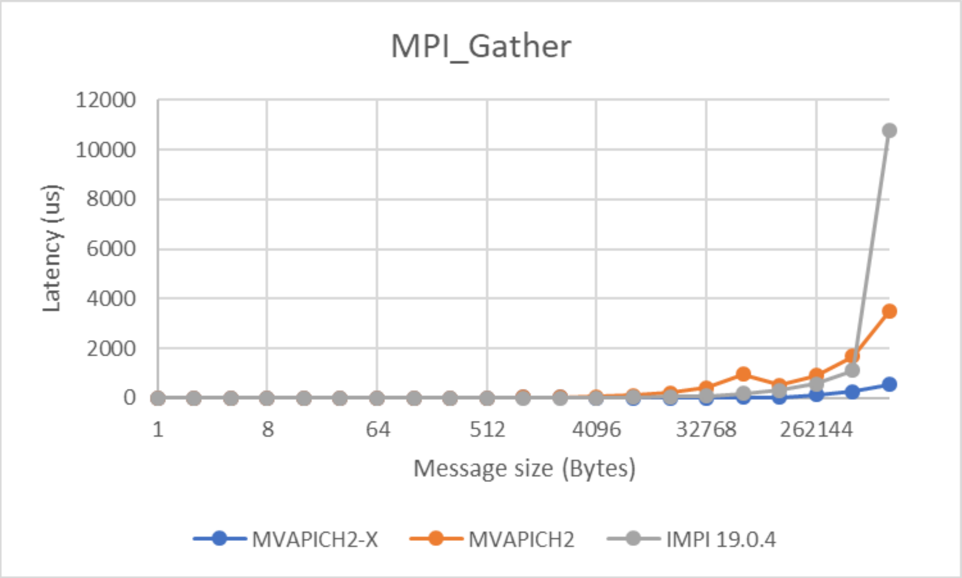
MVAPICH2-X with XPMEM on Nurion Supercomputer

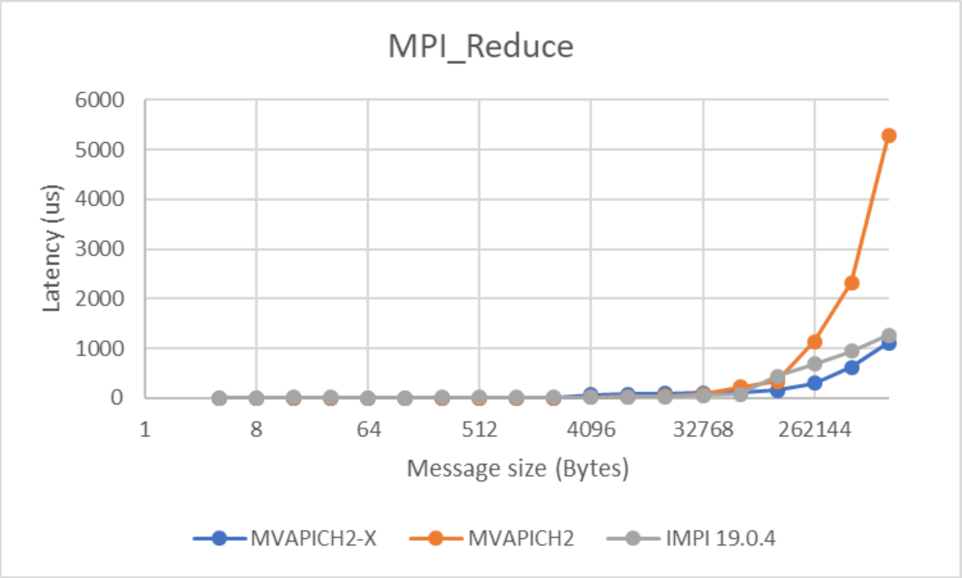
- Compute node
  - Manycore processor: Intel Xeon Phi KNL 7250 (68 cores)
  - MPI intra-node communication
- XPMEM (Cross-Process Memory Mapping)
  - Linux kernel module
  - Enables a process to map the memory of another process into its virtual address space
- OSU Micro-Benchmarks
  - Collective Communications: Latency
  - Point-to-point communications: Bandwidth
- MVAPICH2-X with XPMEM
  - GCC compiler 4.8.5 (Intel IFS 10.6)
  - Build issues on Intel compiler version
- Experimental environment
  - Intel MPI 19.0.4
  - MVAPICH 2.3
  - MVAPICH2-X 2.3rc2 with XPMEM

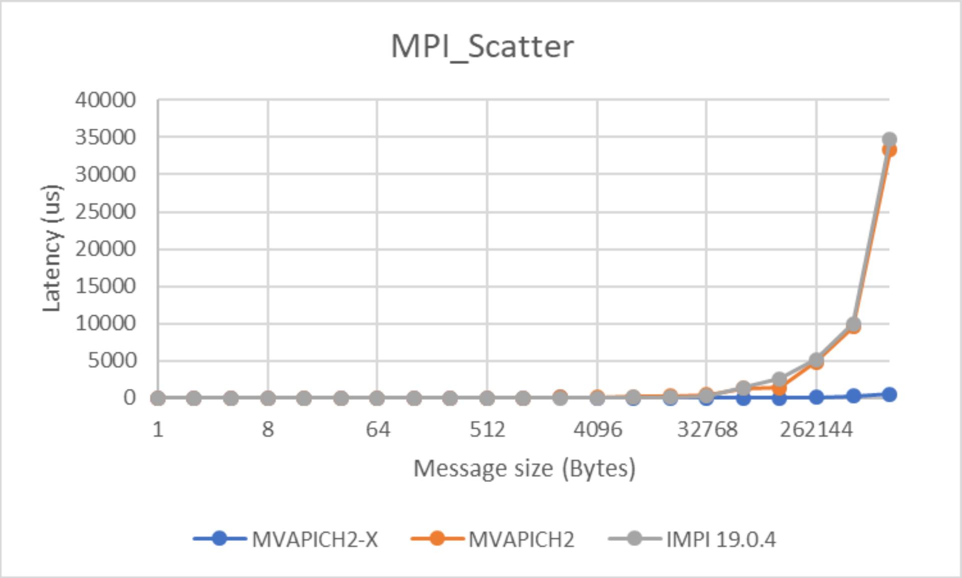
MVAPICH2-X with XPMEM: Single Node

Collective Communication on Single Node: Latency (PPN=64)

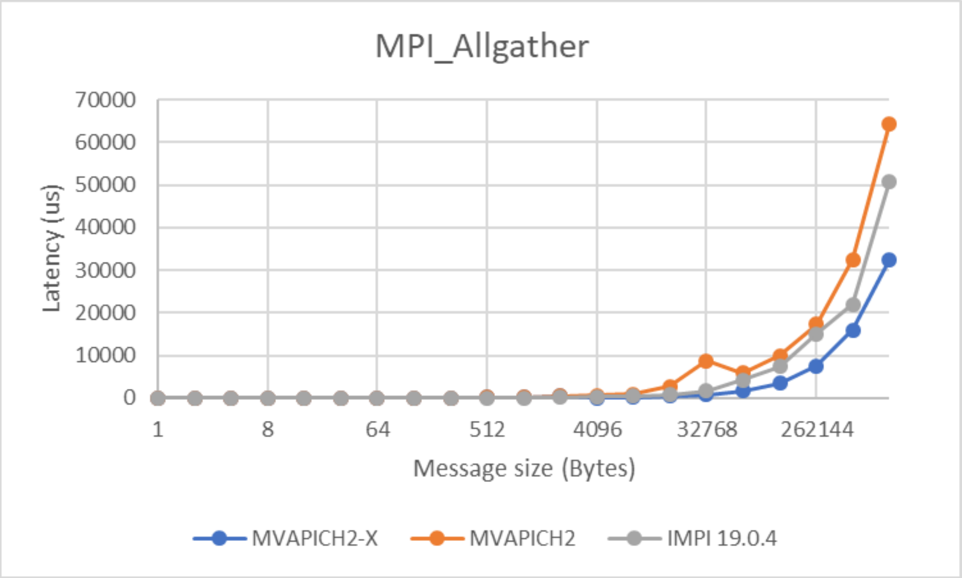


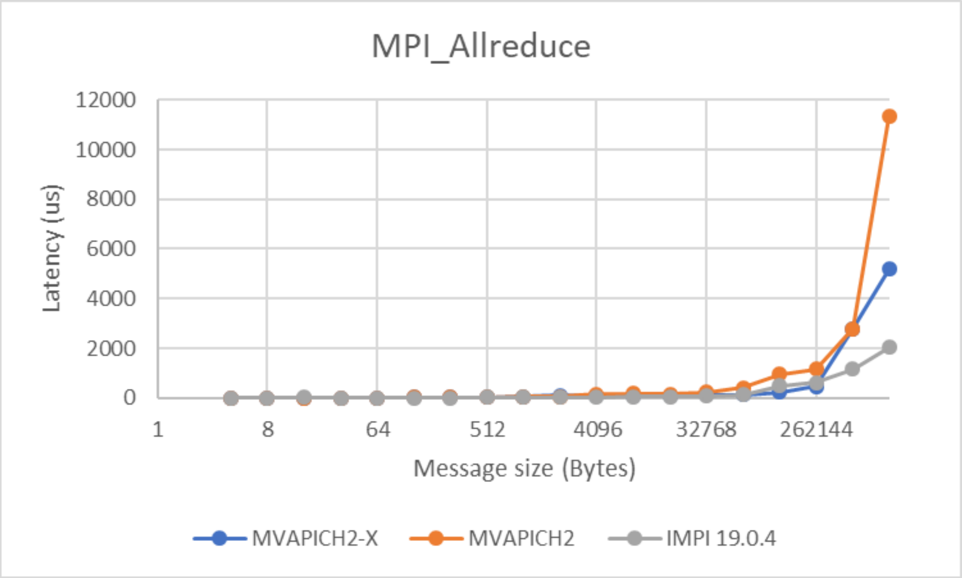




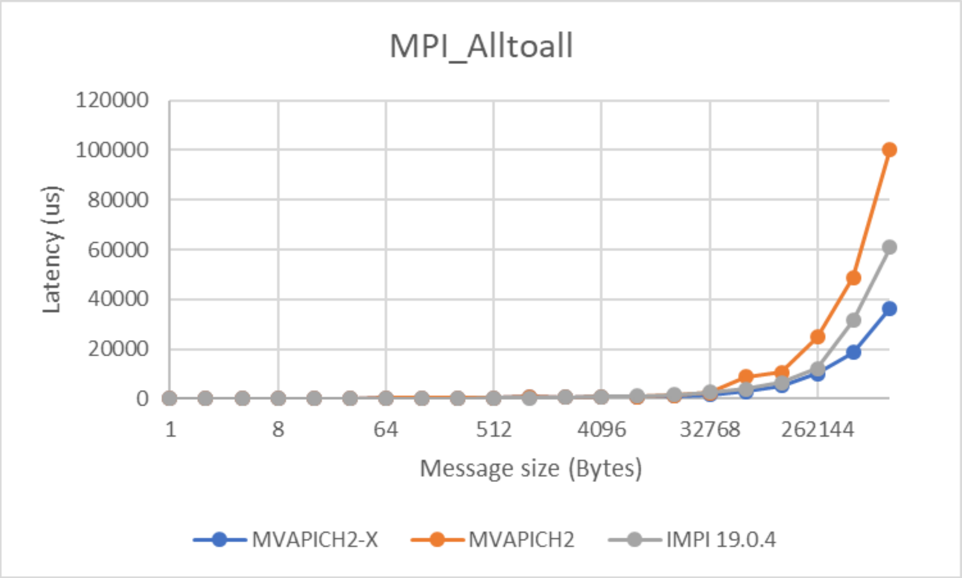


Collective Communication on Single Node: Latency (PPN=64)

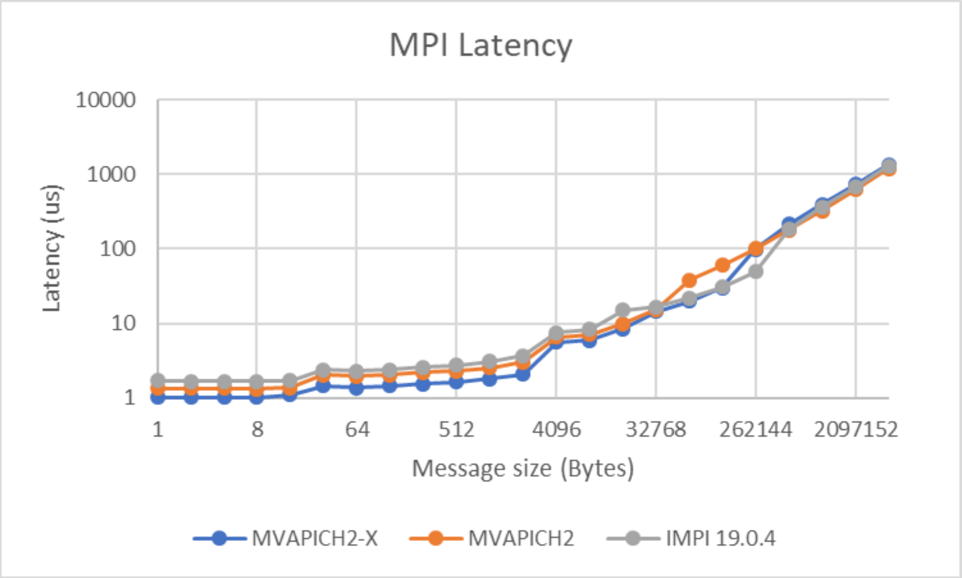


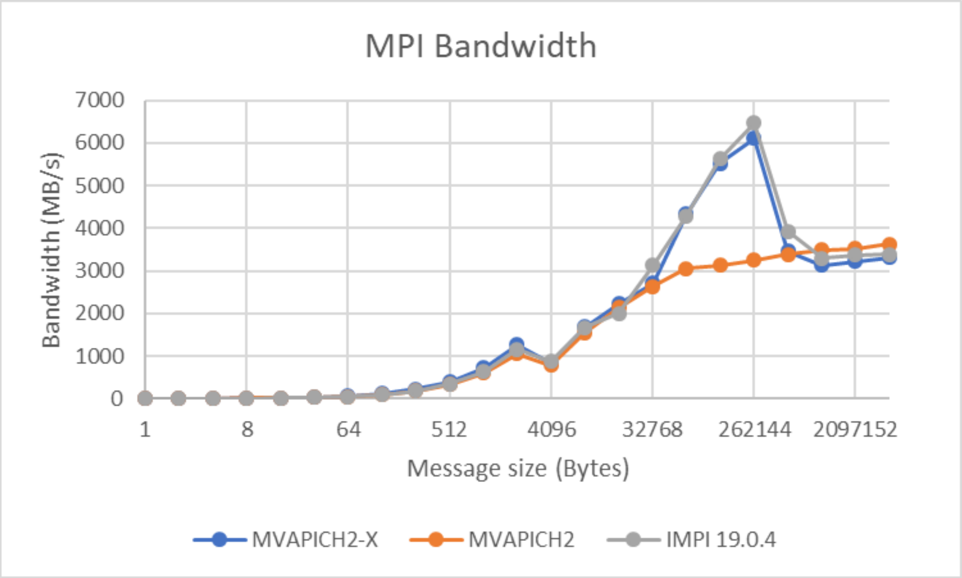


- Experimental environment
  - Single KNL node (cache mode)
  - Processes per node = 64
- Performance evaluation
  - MVAPICH2-X better performance on collective communications compared to MVAPICH2
  - Intel MPI 19.0.4 better performance on MPI_Allreduce

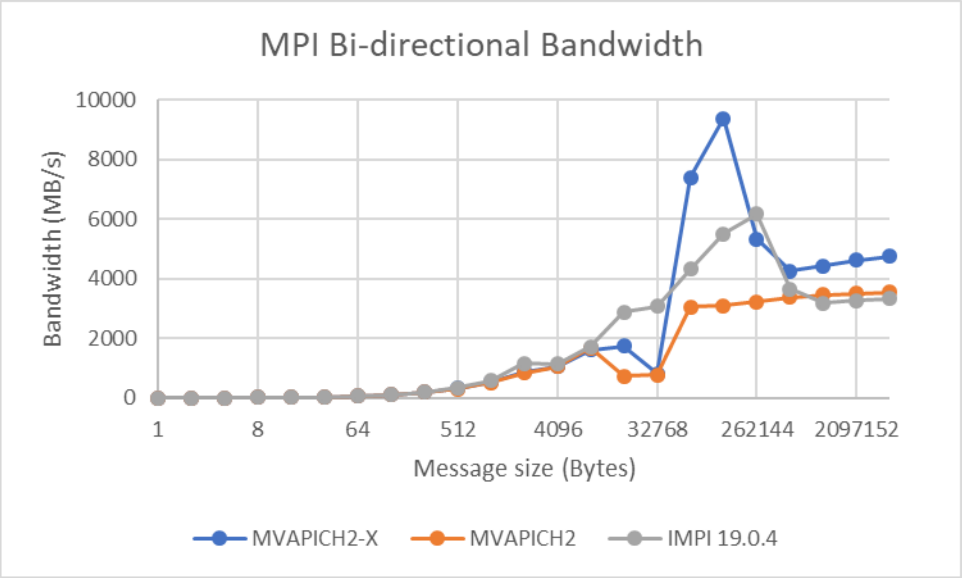


Point-to-Point Communication on Single Node: Bandwidth



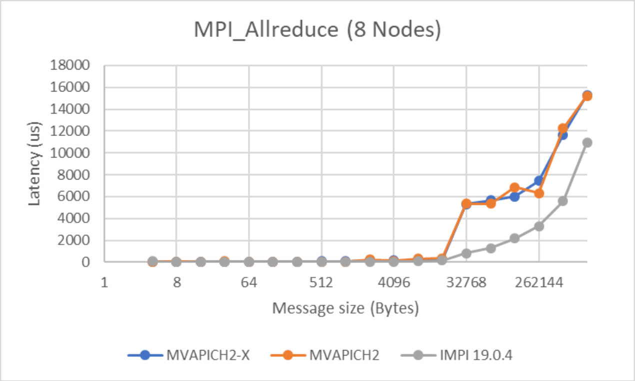


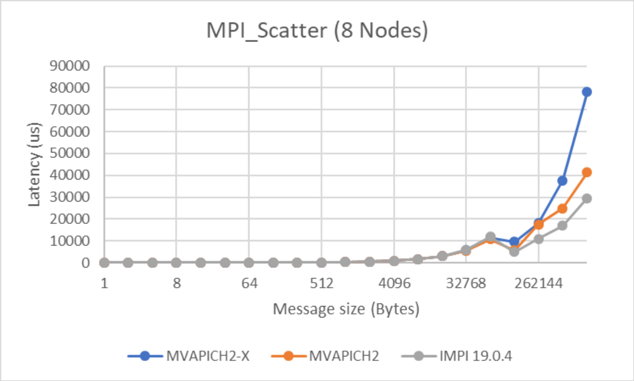
- Experimental environment
  - Single KNL node (cache mode)
- MPI Latency & Bandwidth
  - MVAPICH2-X better performance on 1B-256KB message size
  - MVAPICH2 better performance on 512KB-4MB message size
- MPI Bi-directional BW
  - MVAPICH2 better performance

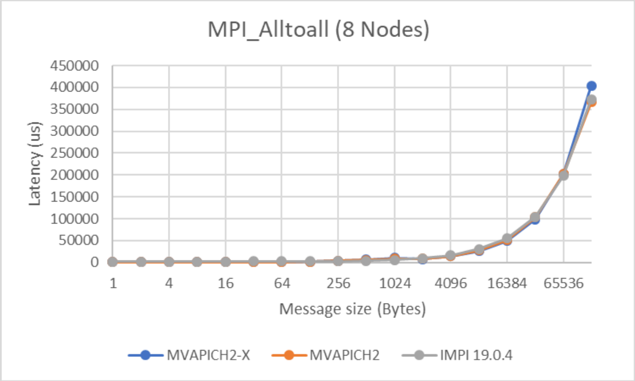


MVAPICH2-X with XPMEM: Multi Node

Collective Communication on Multi Node: Latency

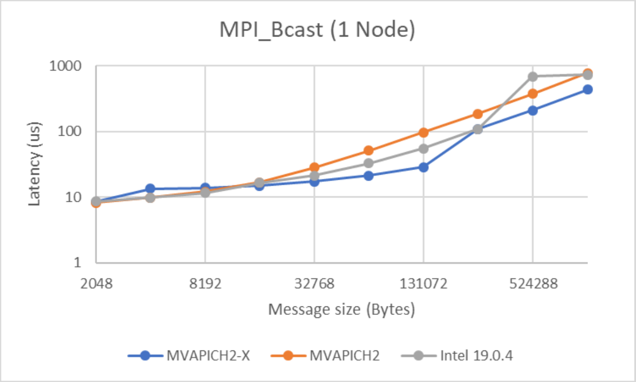


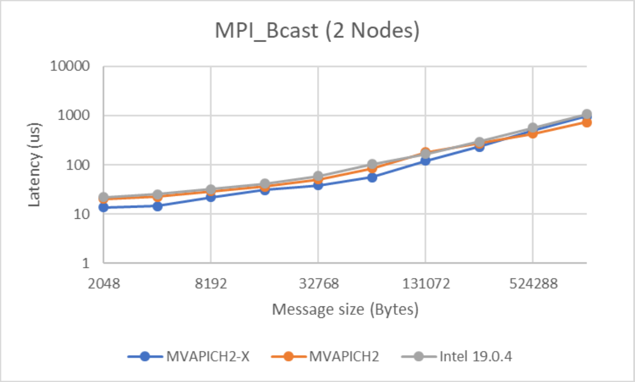


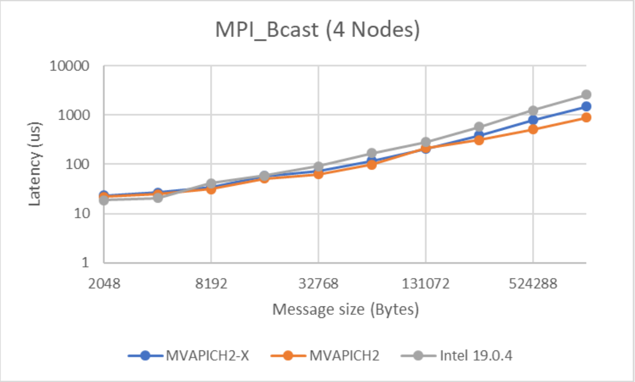


- Experimental environment
  - 8 KNL nodes (Nurion testbed, cache mode)
  - Processes per node = 64
- Performance evaluation
  - MPI_Allreduce: MVAPICH2-X better performance at specific range of the message size (2KB to 64KB)
  - MPI_Scatter: MVAPICH2 better performance
  - MPI_Alltoall: MVAPICH2-X slightly better performance, almost similar
  - In small message case, it could be better performance with threshold option (MV2_XPMEM_COLL_THRESHOLD)

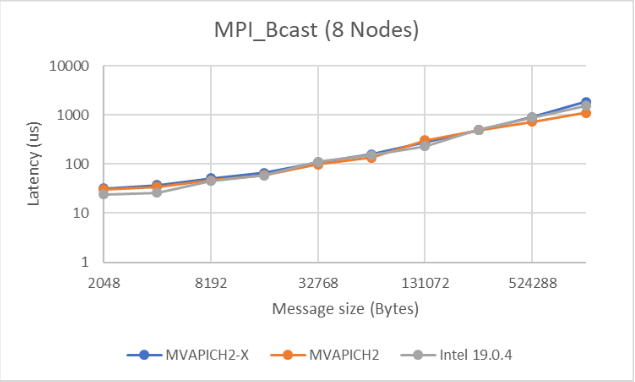
Collective Communication on Multi Node: Latency

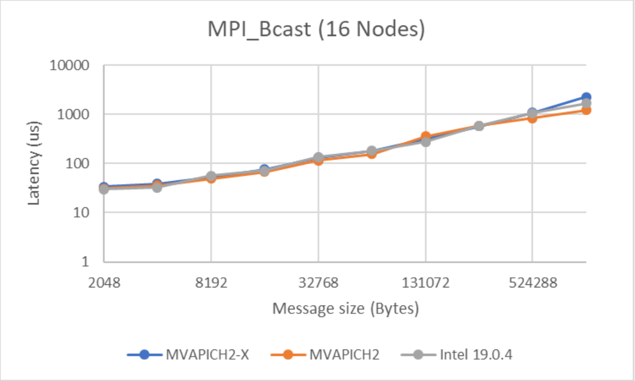




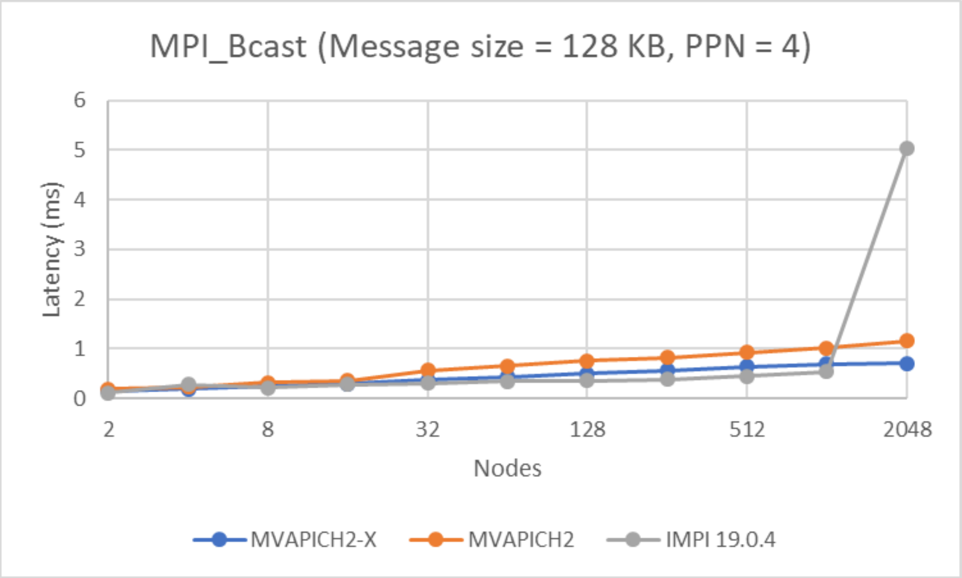


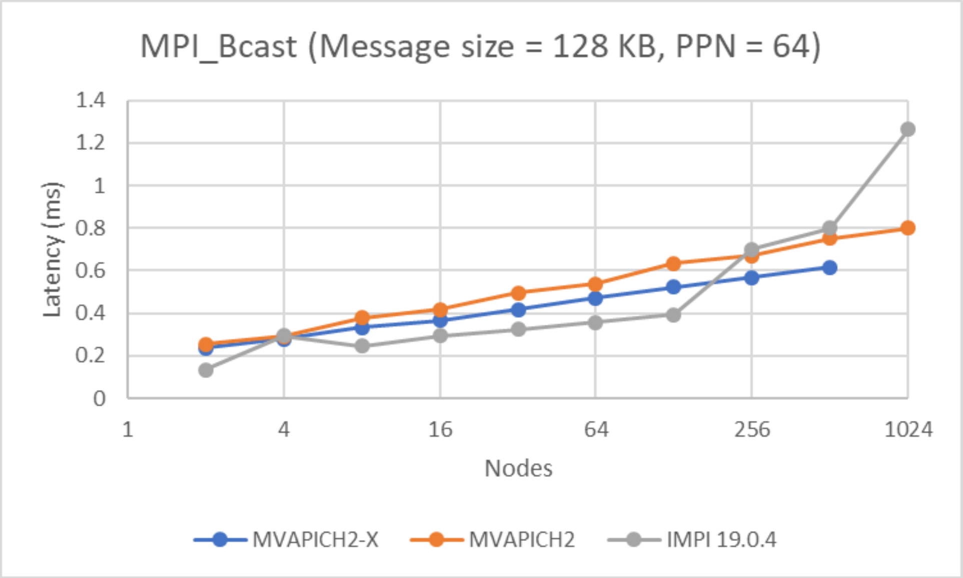
- Experimental environment
  - 1-16 KNL nodes (Nurion testbed, cache mode)
  - MPI_Bcast, processes per node = 64
- Performance evaluation
  - Single-node : MVAPICH2-X with XPMEM better performance if message size is larger than 8KB
  - Multi-node: MVAPICH2-X better performance at specific range of the message size
  - Similar performance results as the number of nodes increases





Collective Communication on Multi Node: Latency





- Experimental environment
  - 2-2048 KNL nodes (Normal queue, cache mode)
  - MPI_Bcast, message size = 128KB, processes per node = 4, 64
- Performance evaluation
  - MVAPICH2-X with XPMEM is faster than MVAPICH2 (PPN=4, 64)
  - MVAPICH2 and MVAPICH2-X with XPMEM are faster than Intel MPI 19.0.4 on 2048 nodes (PPN=4), 256-1024 nodes (PPN=64)
  - MVAPICH2-X with XPMEM has segmentation fault problem on 1024 nodes
  - Intel IFS version of Nurion is slightly different (Intel IFS 10.8)

Conclusion & Future Plan

- MVAPICH2-X with XPMEM
  - GCC compiler, Intel IFS 10.6 (Slightly different with version which is installed in Nurion)
  - Better performance at specific range of the message size on Nurion supercomputer
  - Another option for the large size message could improve the performance which is similar to the threshold option for the small size message (XPMEM_COLL_THRESHOLD)
- MVAPICH2-X installation
  - Intel compiler, Intel IFS 10.8
- Benchmark and application test (Presented at MUG'19)
  - Benchmark: HPCG, NPB
  - Application: DNS-TBL, GOTPM, LAMMPS
- Graph500 optimization with MVAPICH2-X
- Application optimization
  - WRF (Weather Research and Forecasting Model) application
  - Deep learning framework (TensorFlow, PyTorch, Caffe)

THANK YOU