# Performance of Applications on Nurion Utilizing MVAPICH2-X

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

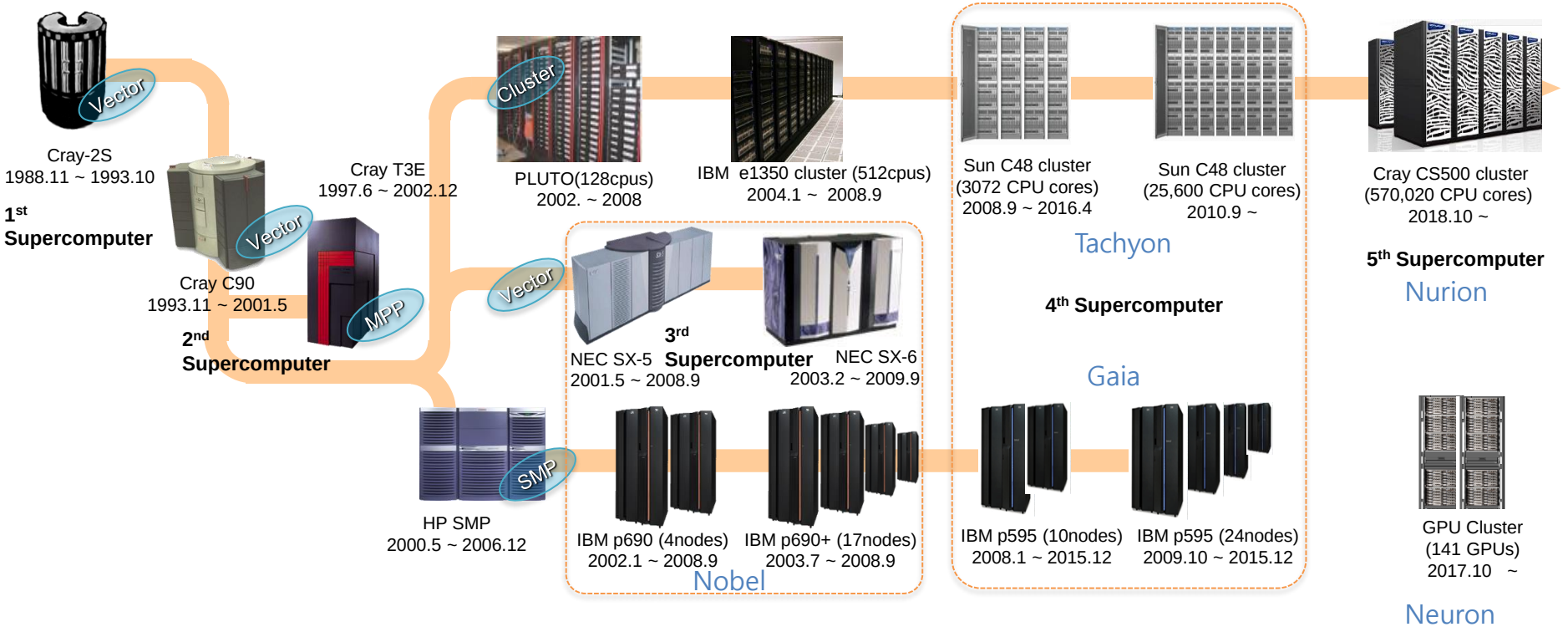
8th Annual MVAPICH User Group Meeting (MUG'20)

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

# Introduction to KISTI-5 Supercomputer, Nurion

# KISTI Supercomputing Center

- The National Supercomputing Center in Korea
- Provide computational resources and its support to R&D communities in Korea
- Nurion (KISTI-5): CPU system, Neuron: GPU system



# KISTI-5 Procurement & Deployment

| | |
|---|---|
| ’15.06 | Data Center Building constructed |
| ’15.07.07 | Approved from Preliminary feasibility study |
| ’16.03 | RFI and BMT Code release |
| ’16.10~12 | 1st Bidding |
| ’17.02~05 | 2nd Bidding |
| ’17.06~07 | Cray Inc. won the bid and Tech/Price Negotiation |
| ’17.08 | Contract finalized (49M USD) |
| ’17.11 | Pilot system(16nodes) delivered |
| ’17.12~’18.4 | Main system delivery and deployment |
| ’18.05~09 | BMT, Functional and Stability Test |
| ’18.07~10 | Early Access on Pilot system |
| ’18.10~11 | Main system Beta service |
| ’18.12~ | Production |

KISTI Facility PUE 1.35

Cray CS500 25.7PFlops

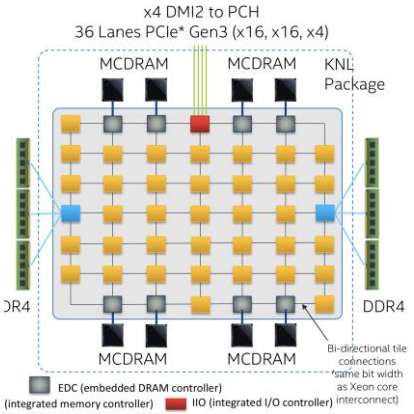
# KISTI-5 Compute Nodes

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

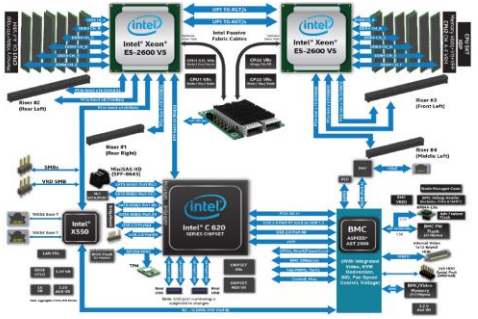
- 1x Xeon Phi KNL 7250, 68Cores 1.4GHz, AVX512
- 3TFlops Peak, ~0.2 Bytes/Flops,
- 96GB (6x16GB) DDR4-2400 6 channel RAM,
- 16GB HBM (460GB/s)
- 1x 100Gbps OPA HFI, 1x On-board GigE Port



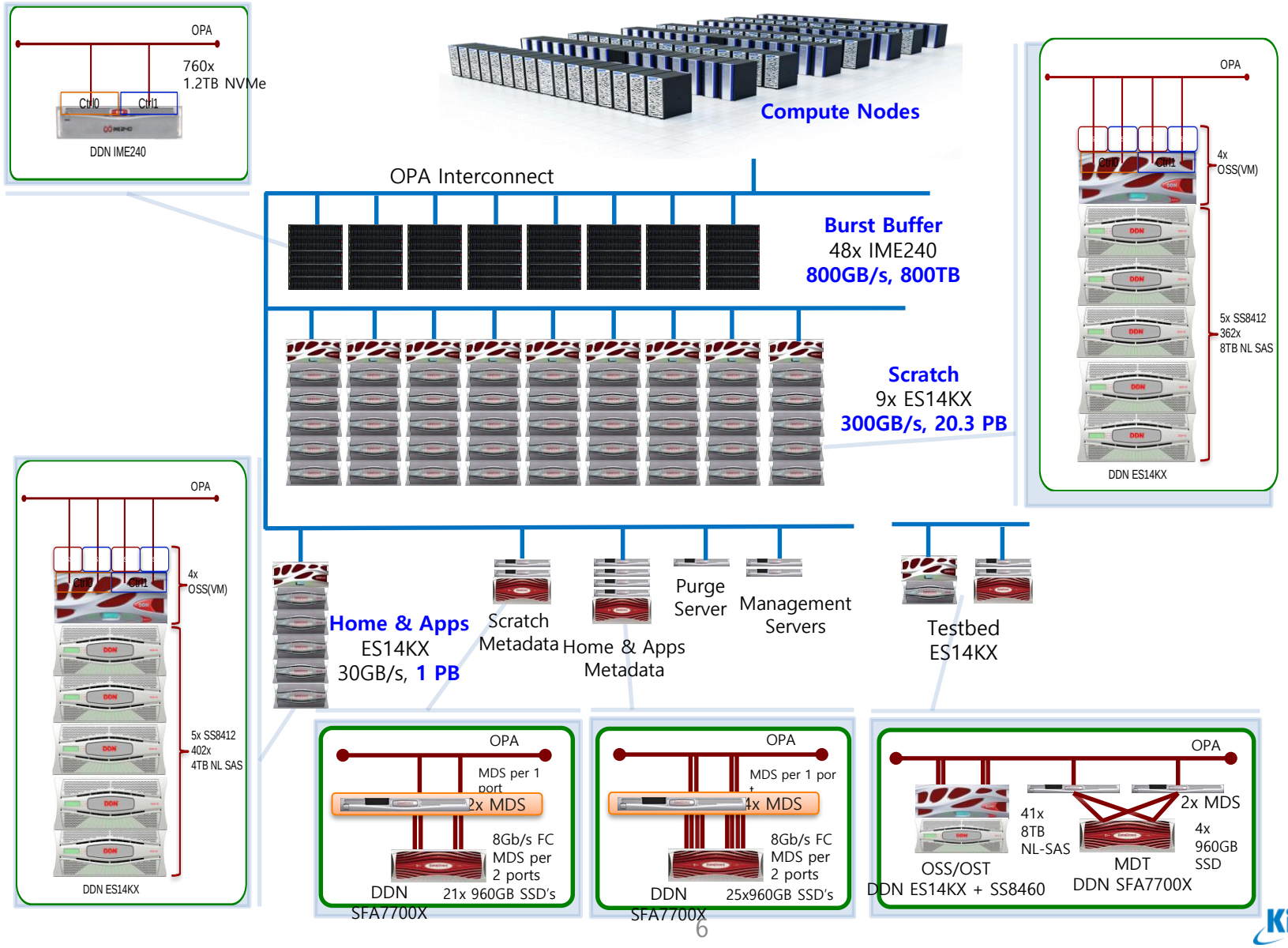

## CPU-only nodes

132 Skylake Computing modules, 4 Racks, 0.4PF

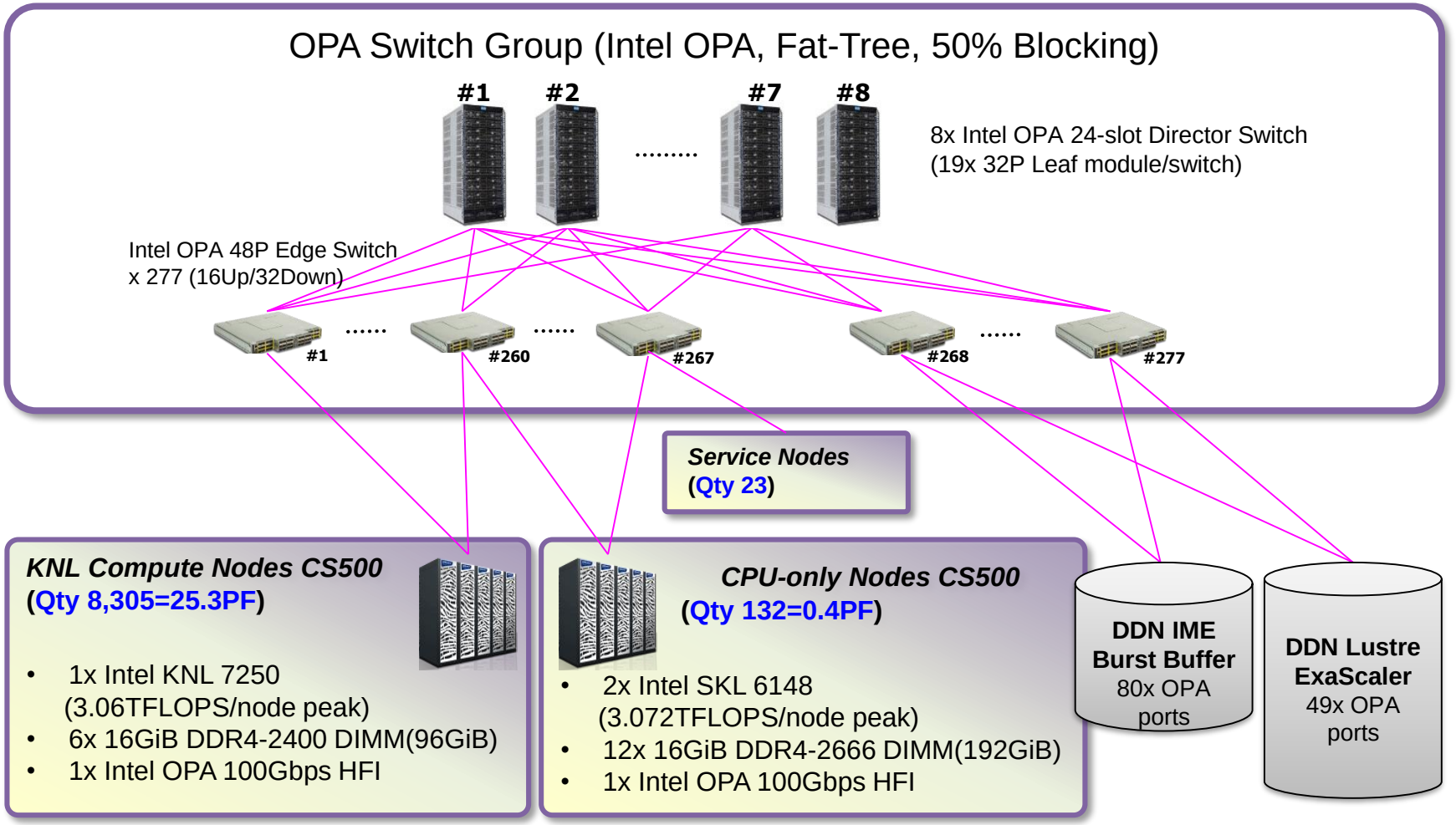
- 2x Xeon SKX 6148 CPUs, 2.4GHz, AVX512
- 192GB (12x 16GB) DDR4-2666 RAM
- 1x Single-port 100Gbps OPA HFI card
- 1x On-board GigE (RJ45) port



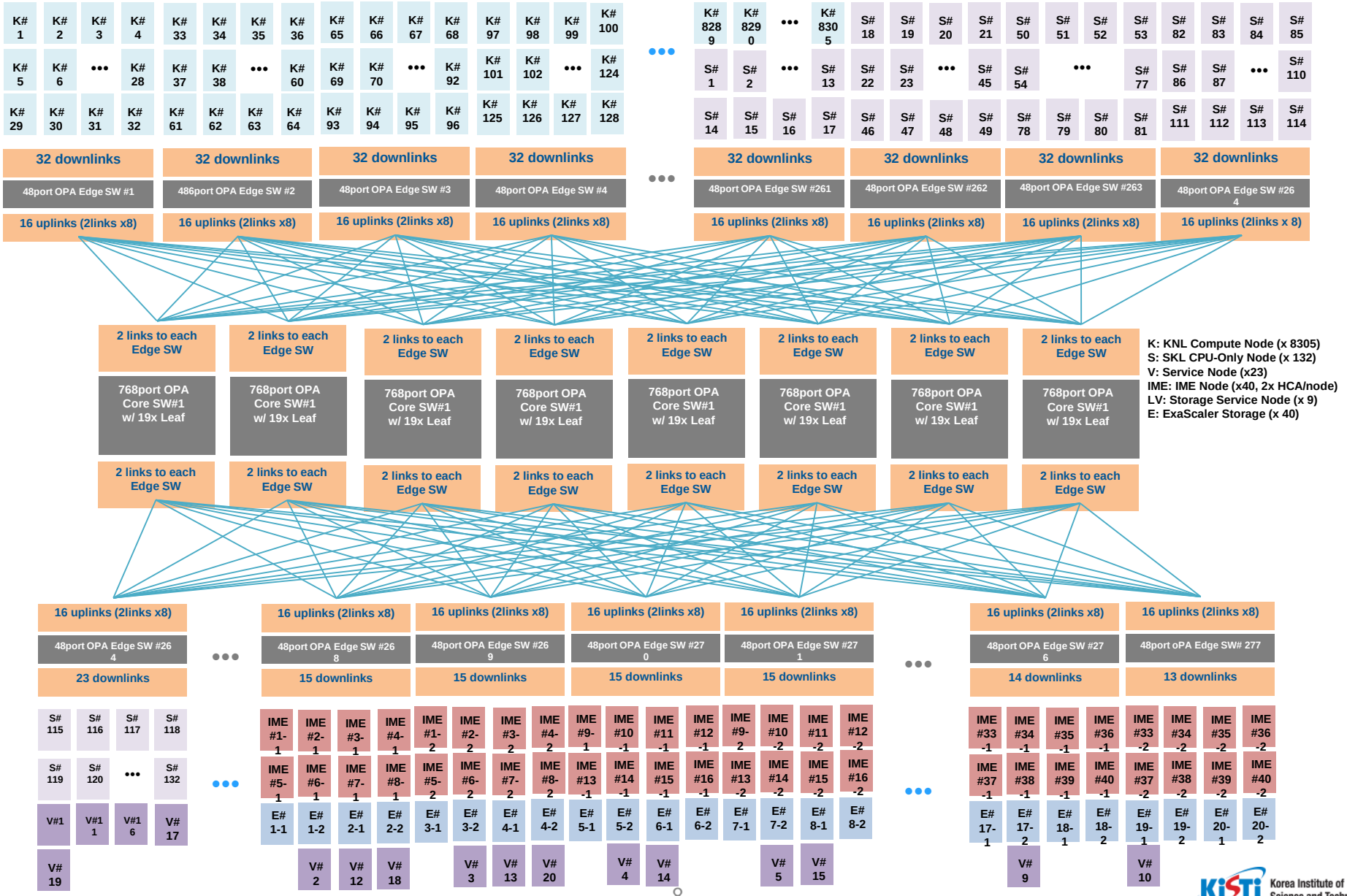
# KISTI-5 Storage System



# KISTI-5 OPA Interconnect

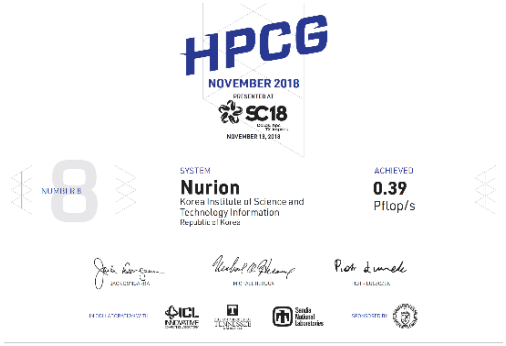


# 2:1 Blocking OPA Interconnect



# Benchmark Performance Result

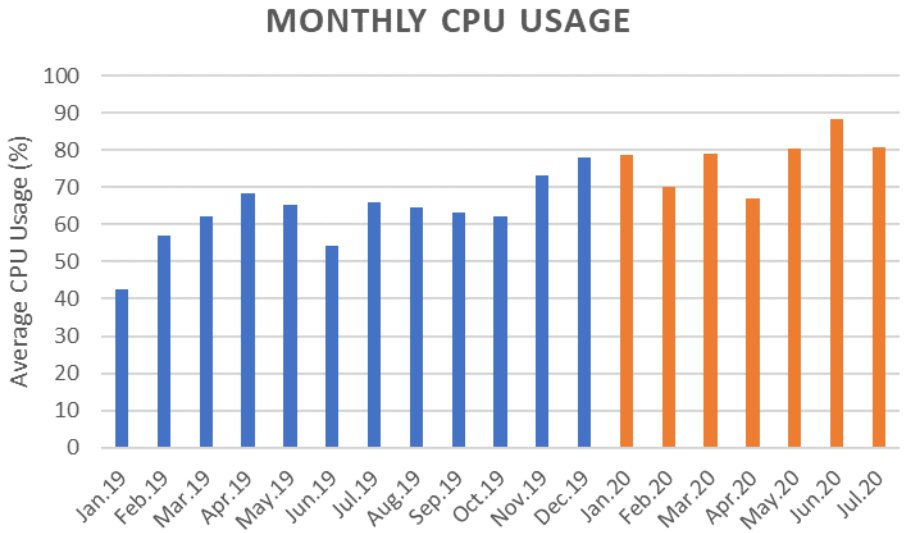
| Category | Features | # of nodes | Score | World Ranking |
|---|---|---|---|---|
| HPL | Large-scale Dense Matrix Computation<br>Used for Top500 | 8,174(KNL)<br>+ 122(SKX) | 13.93PF | 17th (Jun 2020) |
| HPCG | Large-scale Sparse Matrix Computation<br>Similar to normal user applications | 8,250(KNL) | 0.39PF | 10th (Jun 2020) |
| Graph500 | Breadth-First Search,<br>Single-Source Shortest Paths | 1,024(KNL) | 1,456GTEPS<br>337GTEPS | 7th (Jun 2020)<br>3rd (Jun 2020) |
| IO500 | Various IO Workloads | 2,048(KNL) | 282.45 | 6th (Jun 2020) |

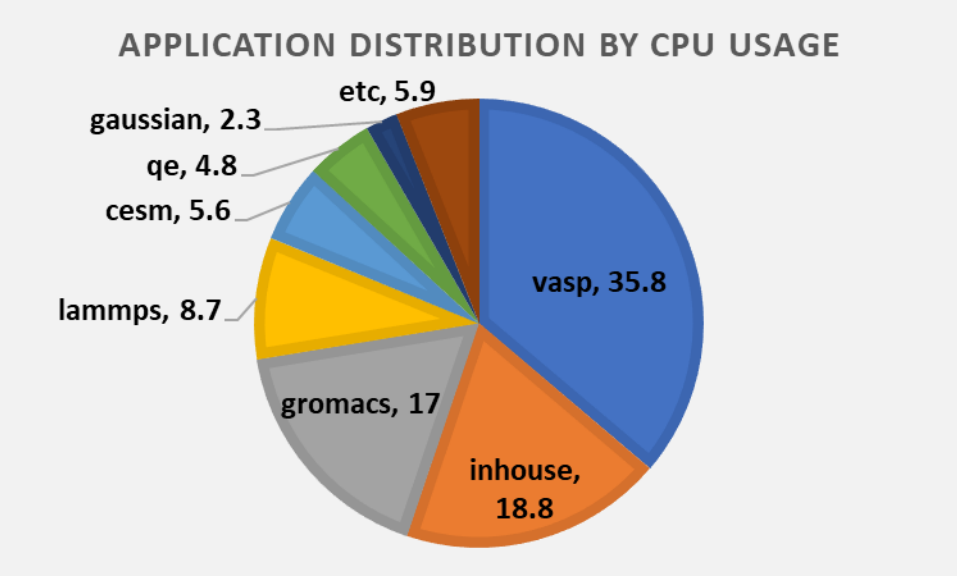





# Nurion Production Status

- CPU usage are getting higher since system production started
- 63.1% in 2019, 77.7% in 2020, 88.3% in June 2020
- Large jobs requires more waiting time
- Schedular policy changed in July 2020 for large scale jobs
- Time to prepare the next supercomputer KISTI-6
- VASP, GROMACS, and LAMMPS have been widely used in Nurion



# Necessity of MPI Intra-Node Communication Optimization

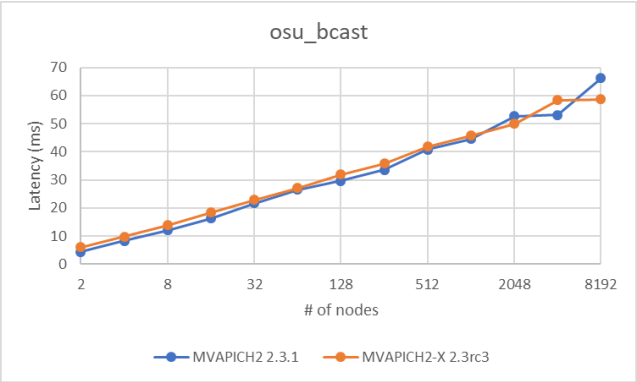
- HPC Benchmarks
  - HPL and HPCG
  - OpenMP + MPI implementation
- HPC Applications
  - VASP, GROMACS, and LAMMPS
  - MPI only or OpenMP(2-3 threads)+MPI implementation
  - Nurion compute node: Manycore processor, Intel Xeon Phi KNL 7250 (68 cores)
  - MPI Intra-node communication should be optimized
- MVAPICH2-X with XPMEM on Nurion
  - Interconnect: Intel OPA
  - Intra-node Communication Optimization
    - XPMEM (Cross-Process Memory Mapping)
    - Linux kernel module
    - Enables a process to map the memory of another process into its virtual address space
  - Optimized Asynchronous Progress

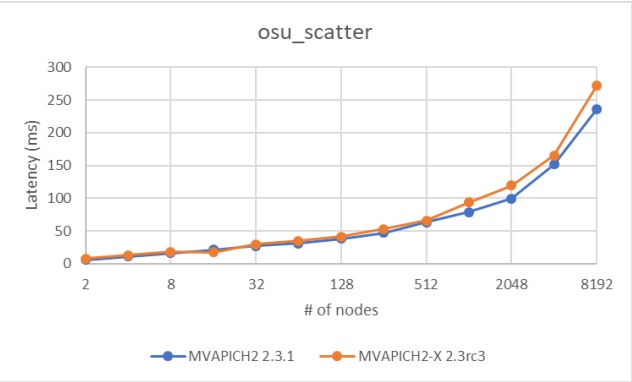
## Performance Test on Nurion Supercomputer

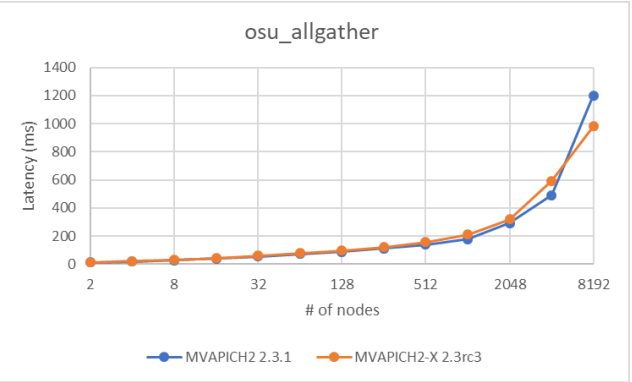
# Performance of Applications on Nurion Supercomputer

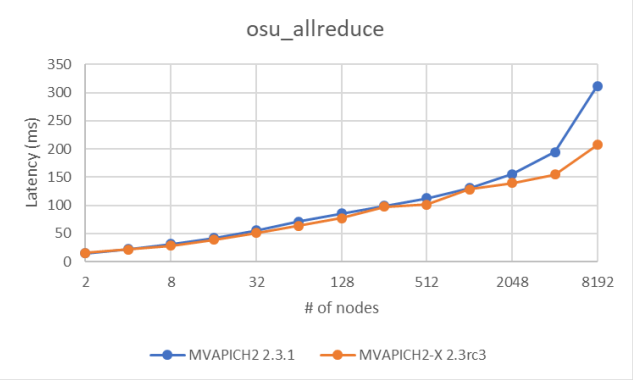
- Performance improvements of the following configurations
  - MPI only implementation
  - OpenMP + MPI hybrid implementation in intra-node level
- Benchmarks
  - OSU Micro-Benchmarks (OMB)
  - NAS Parallel Benchmarks (NPB)
- Applications
  - Direct Numerical Simulation – Turbulent Boundary Layer (DNS-TBL)
  - Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS)
  - Vienna Ab initio Simulation Package (VASP)
- Experimental environment
  - Intel Fortran/C++ compiler 19.0.5
  - MVAPICH2 2.3.1
  - MVAPICH2-X 2.3rc3
    - MV2_IBA_EAGER_THRESHOLD=200000
    - MV2_OPTIMIZED_ASYNC_PROGRESS=0

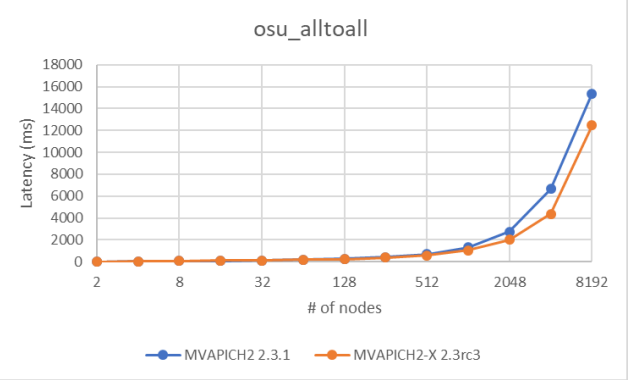
# OMB: Inter-node communication (PPN=1)



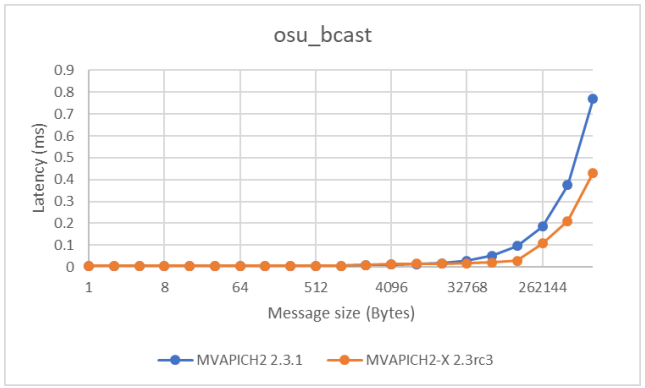



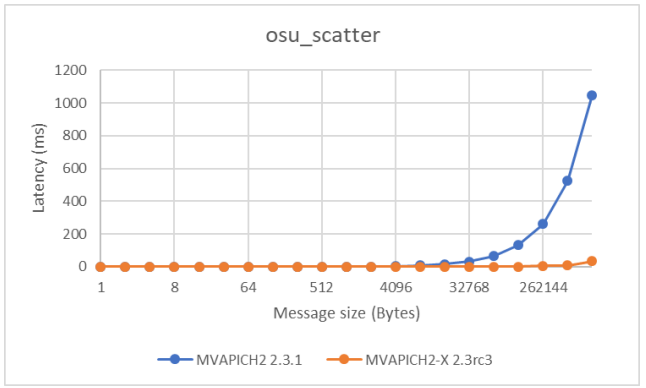



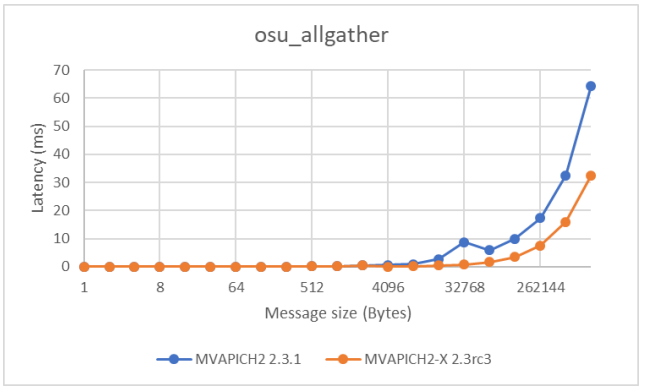



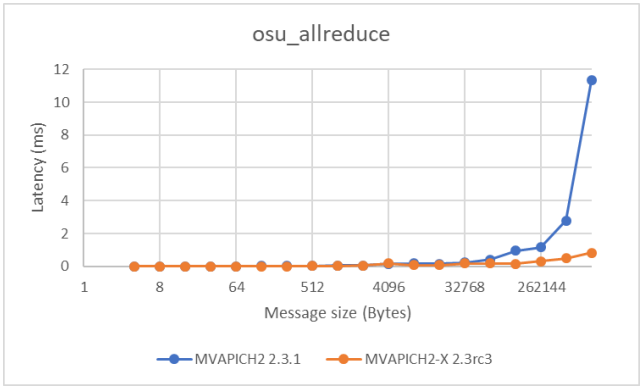

- Collective communications on 2 – 8192 nodes (Message size: 64 bytes)
- Better performance on MPI_Allreduce and MPI_Alltoall in large scale
- Similar performance on MPI_Bcast, MPI_Scatter, and MPI_Allgather

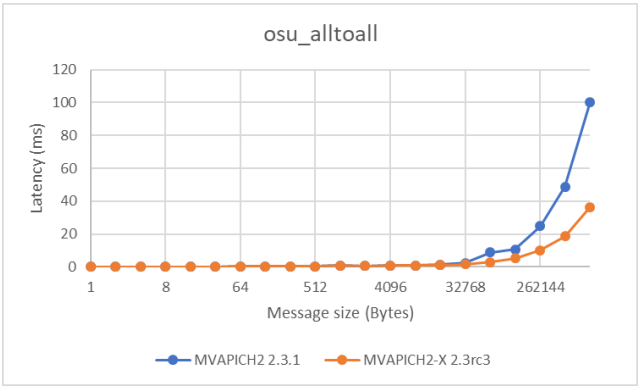
# OMB: Intra-node communication (PPN=64)



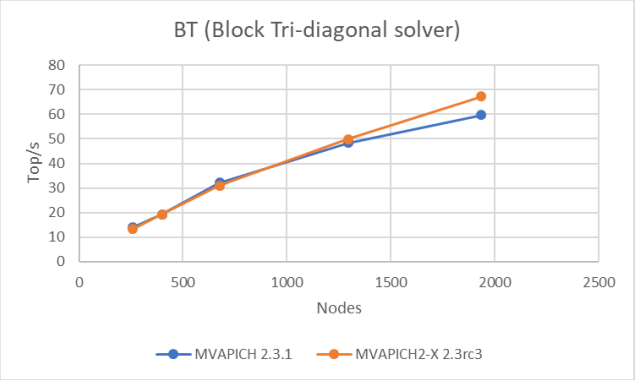



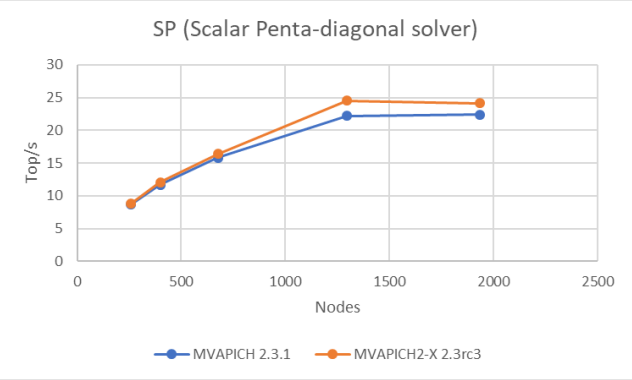



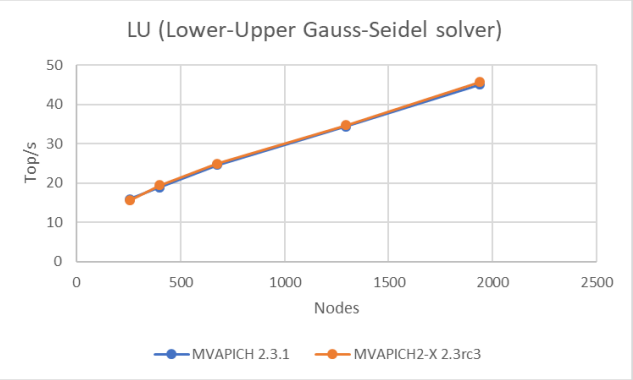




- Collective communications on single node (Message size: 64 bytes- 1MB)
- MVAPICH2-X better performance on every benchmarks
- Huge performance gain on **MPI_Scatter** and **MPI_Allreduce** with large message size

# NPB: NAS Parallel Benchmarks



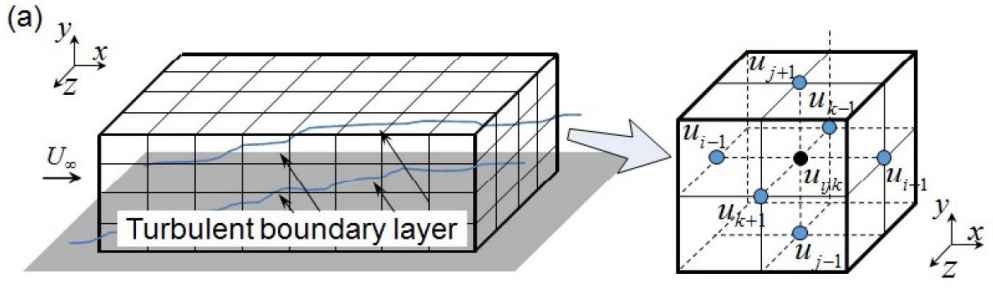




- Benchmarks are derived from computational fluid dynamics applications
- Experimental environment
  - Evaluation on 256-1936 nodes on normal queue (cache mode)
  - Problem Class = F (Grid size: $2560 \times 2560 \times 2560$), strong scaling
  - 1 OpenMP thread, PPN = 64, 16384-123904 MPI processes
- Performance evaluation
  - MVAPICH 2-X faster than MVAPICH 2 on BT (Block Tri-diagonal solver)
  - MVAPICH 2-X faster than MVAPICH 2 on SP (Scalar Penta-diagonal solver)
  - Similar performance on LU (Lower-Upper Gauss-Seidel solver)
  - MVAPICH 2-X shows better performance on the **large number of nodes** (1936 nodes)

# DNS-TBL

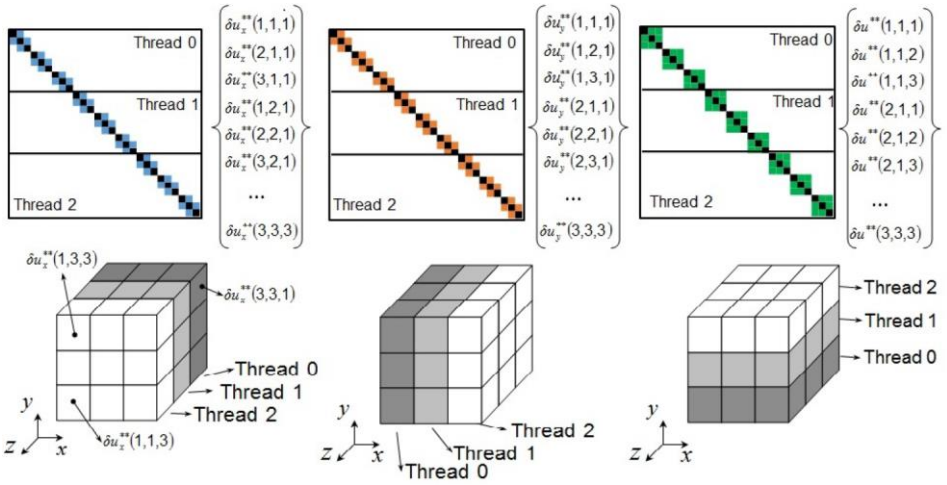
- Direct Numerical Simulation – Turbulent Boundary Layer*
- Velocity solver and pressure solver (Application of turbulent flow)
- Solve the continuity and incompressible **Navier-Stokes equation**
  - Second-order finite difference scheme on the 7-point stencil
  - Discretized into hepta-diagonal matrix in 3D, and broke into 3 tridiagonal matrices
  - Transformed pressure Poisson's equation into a single tridiagonal matrix by 2D FFT
- Optimized code for Intel® Xeon Phi™ Processor, Fortran, FFTW 3.3.7 library




$$I + \Delta t M_{11} = \left(I + \Delta t M_{11}^{x}\right) \times \left(I + \Delta t M_{11}^{y}\right) \times \left(I + \Delta t M_{11}^{z}\right) + O\left(\Delta t^{2}\right)$$

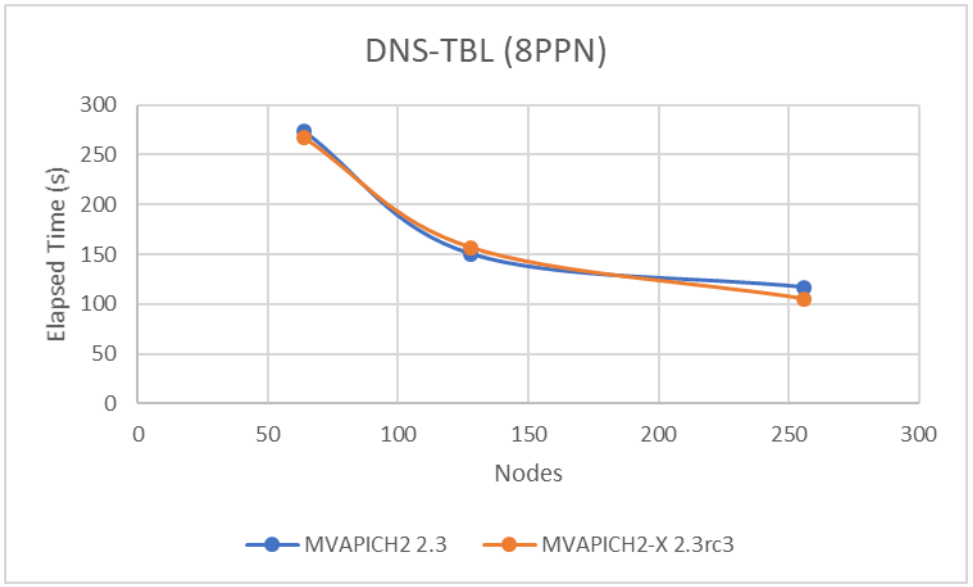
$$\left(I + \Delta t M_{11}^{x}\right) \times \left(I + \Delta t M_{11}^{y}\right) \times \left(I + \Delta t M_{11}^{z}\right) \delta u^{**} = RHS_{x}$$

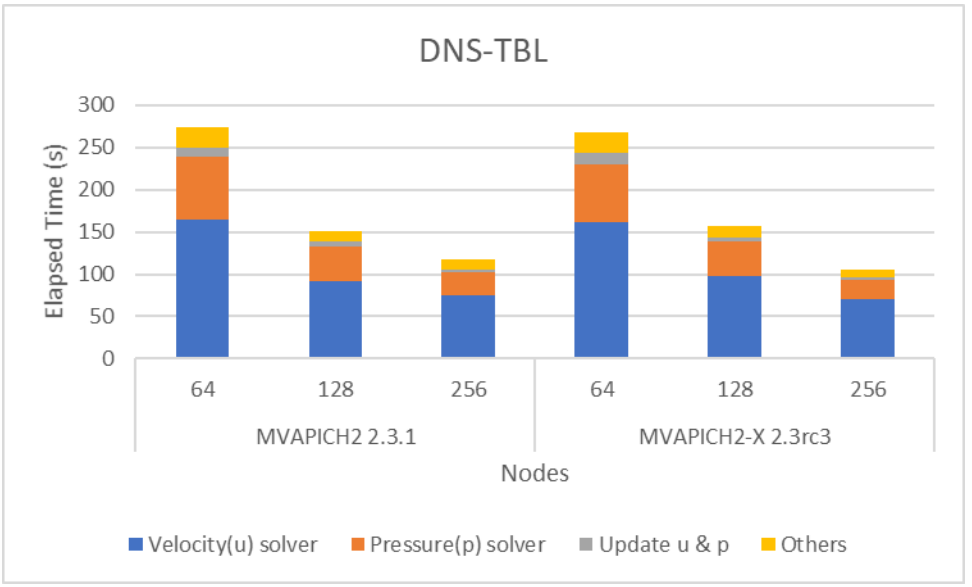
$$\left(I + \Delta t M_{11}^{x}\right) \delta u_{x}^{**} = RHS_{x} \qquad \left(I + \Delta t M_{11}^{y}\right) \delta u_{y}^{**} = \delta u_{x}^{**} \qquad \left(I + \Delta t M_{11}^{z}\right) \delta u^{**} = \delta u_{y}^{**}$$




* J-H. Kang and Hoon Ryu, Acceleration of Turbulent Flow Simulations with Intel Xeon Phi(TM) Manycore Processors (IEEE CLUSTER 2017)

# DNS-TBL: Experiment Results

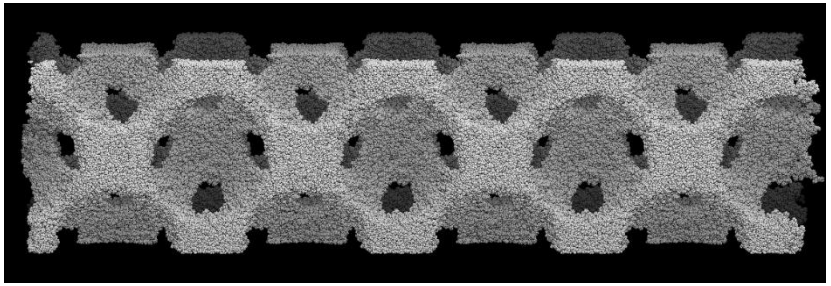


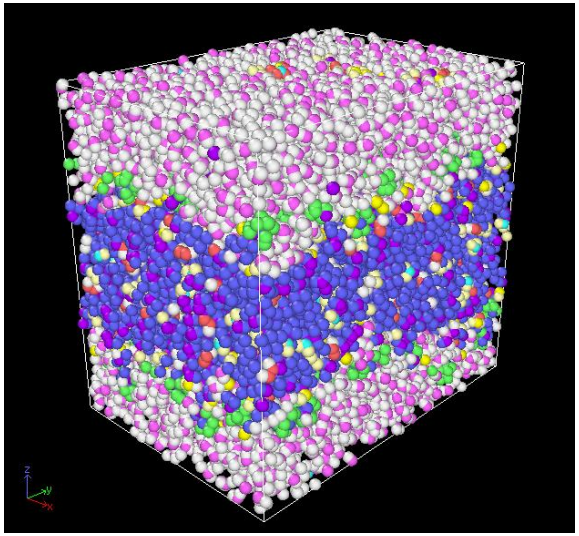




- Experimental environment
  - Evaluation on 64-256 nodes on normal queue (cache mode)
  - 8193 X 401 X 8193 grids, 800 Reynolds number, 10 time step, strong scaling
  - 8 OpenMP thread, PPN = 8, 512-2048 MPI processes
- Performance evaluation
  - MVAPICH2-X faster than MVAPICH2 on velocity and pressure solver
  - Similar performance results on update velocity and pressure and others
  - MVAPICH2-X shows better performance on the large number of nodes (256 nodes)

# LAMMPS

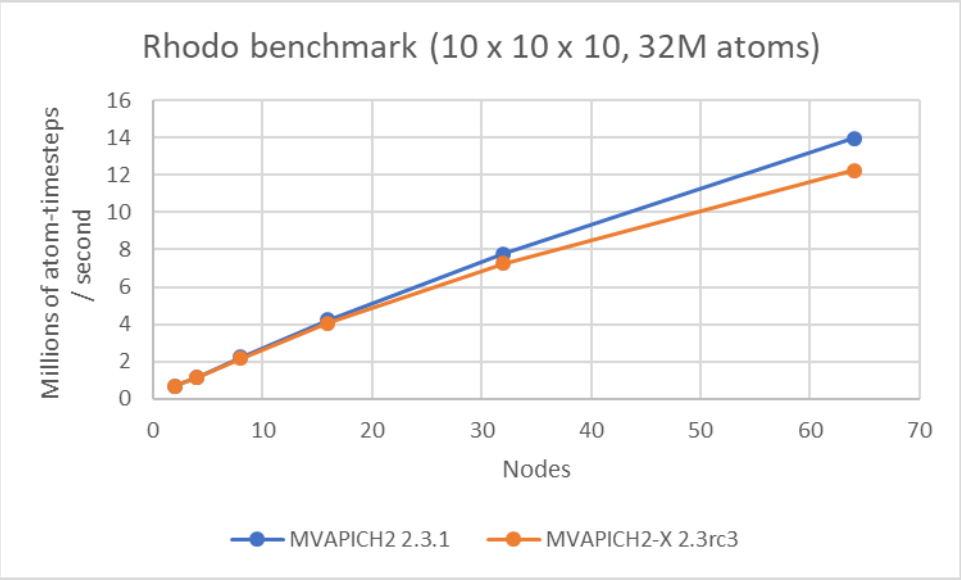
- Large-scale Atomic/Molecular Massively Parallel Simulator*
- Molecular dynamics simulator
- Rhodo benchmark: Rhodopsin protein in solvated lipid bilayer
- Pair, Bond, K-space, Neighbor, Output, Modify, and others
- C, Fixed version of 31Mar2017
  - Initialize by the neighbored 3D halo instead of ring communication pattern
  - Designed algorithms to have no difference on result values

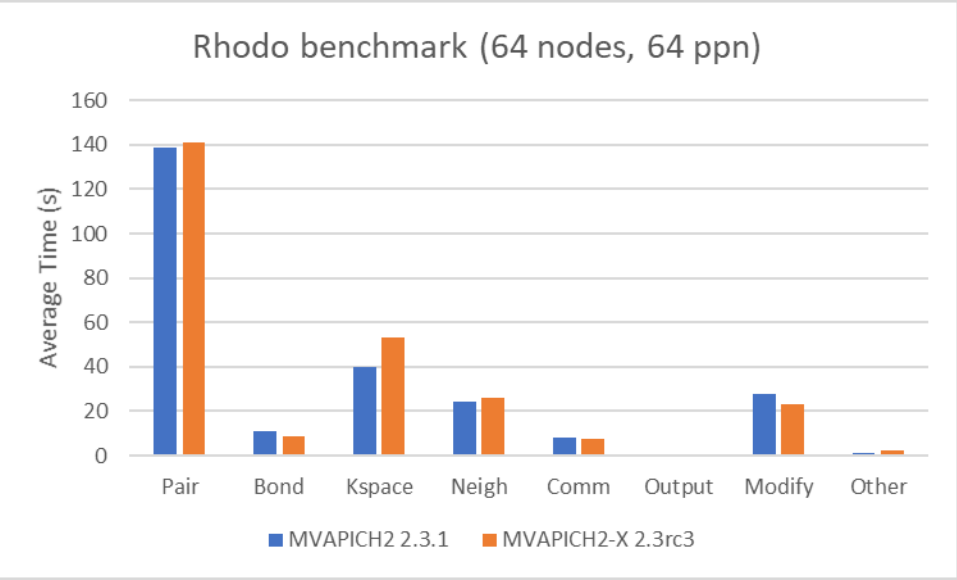




* S. J. Plimpton *et al.*, Particle-Mesh Ewald and rRESPA for Parallel Molecular Dynamics Simulations (SIAM Conference on Parallel Processing for Scientific Computing, 1997)

# LAMMPS: Experimental Results

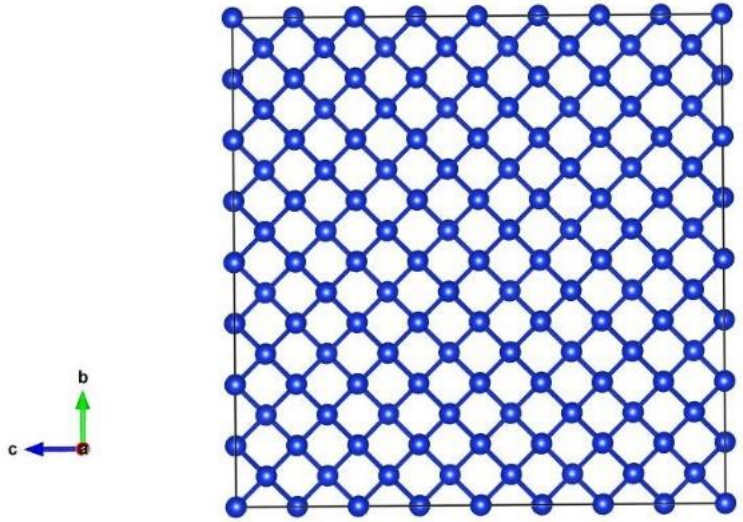



- Experimental environment
  - Evaluation on 2-64 nodes on normal queue (cache mode)
  - 10X10X10 with 32M atoms, strong scaling
  - 1 OpenMP thread, PPN = 64, 128-4096 MPI processes
- Performance evaluation
  - MVAPICH2-X slower than MVAPICH2
  - Performance degradation on K-space, Neighbor part, parameter tuning is required

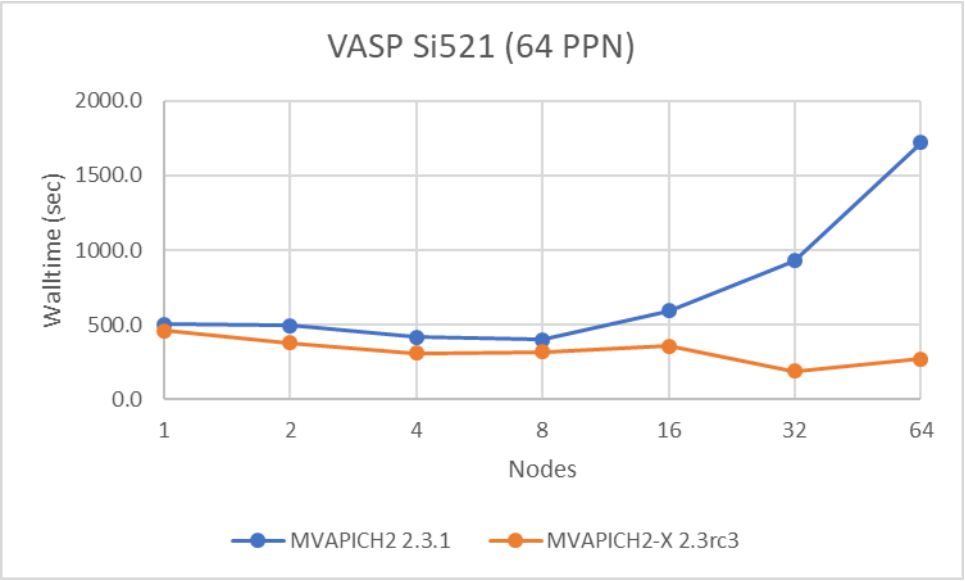
# VASP

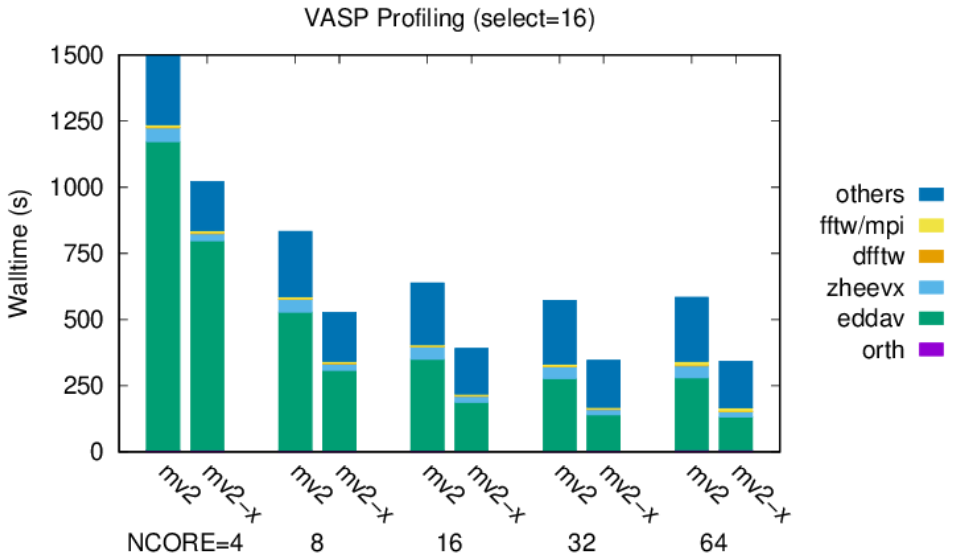
- Vienna Ab initio Simulation Package*
- Molecular dynamics simulator
- Projector augmented-wave(PAW) PBE: Si 05Jan2001
- VASP 5.4.4 version
- Ortho(Matrix multiplication), eddav(Matrix diagonalization), zheevx(sub-space diagonalization), dfftw, fftw/mpi, other
- NCORE variable: 4 - 64




* Guangyu Sun *et al.*, Performance of the Vienna ab initio simulation package (VASP) in chemical applications (Journal of Molecular Structure: THEOCHEM, Vol. 624, April 2003)

# VASP: Experimental Results




- Experimental environment
  - Evaluation on 1-64 nodes on normal queue (cache mode), strong scaling
  - 1 OpenMP thread, PPN = 64, 64-4096 MPI processes
- Performance evaluation
  - MVAPICH2-X shows better performance compared to MVAPICH2
  - Optimal point: 32 KNL nodes with NCORE=64
  - The performance difference increases as the number of nodes increases

# Conclusion & Future Plan

- Performance depends on the experimental configuration
  - Performance improvement on the most application
  - Performance improvement depending on the **communication pattern** for each application
- MVAPICH2-X parameter tuning for each application
  - LAMMPS, GROMACS, Graph500
- MVAPICH2-X on Mellanox Infiniband
  - Neuron (GPU system): Mellanox Infiniband
  - Core-direct based collective offload, SHARP-based collective offload
- MVAPICH2-GDR on Neuron

# THANK YOU