# Introduction to KISTI-5 Supercomputer, *NURION*

Supercomputing Infrastructure Center, KISTI

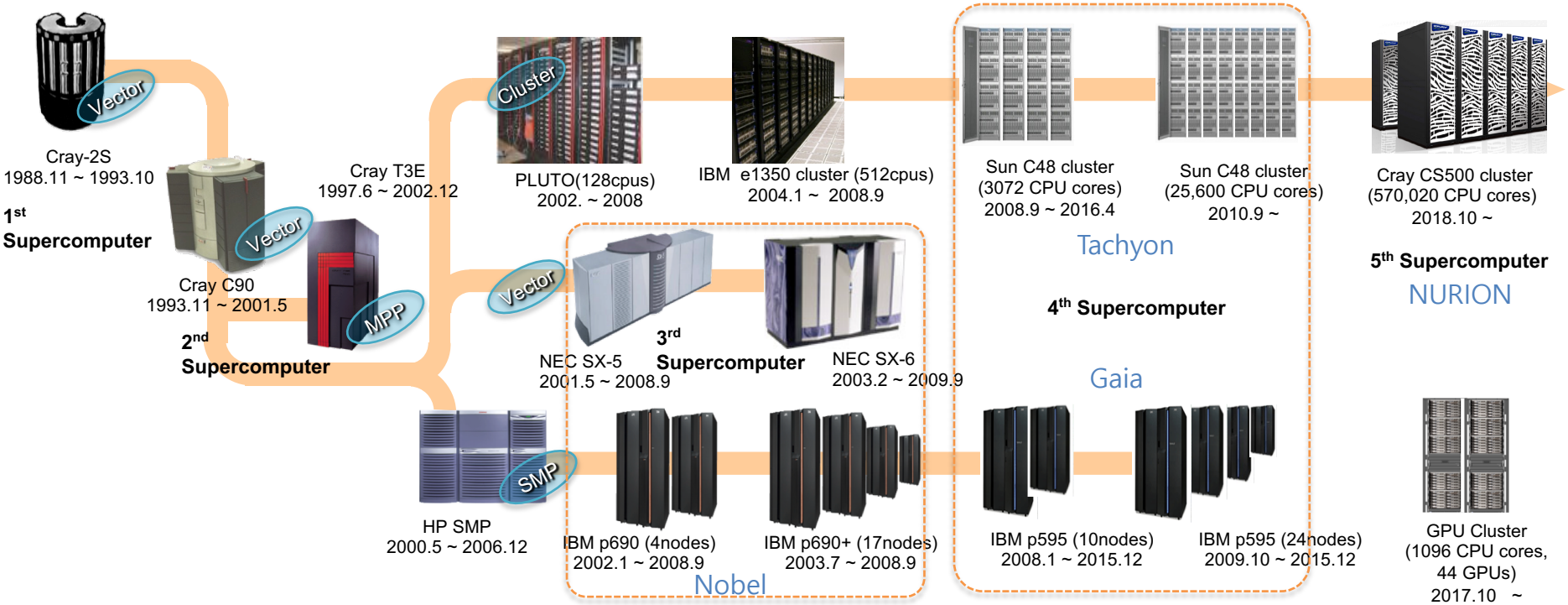
Taeyoung Hong

tyhong@kisti.re.kr

June 20 2019

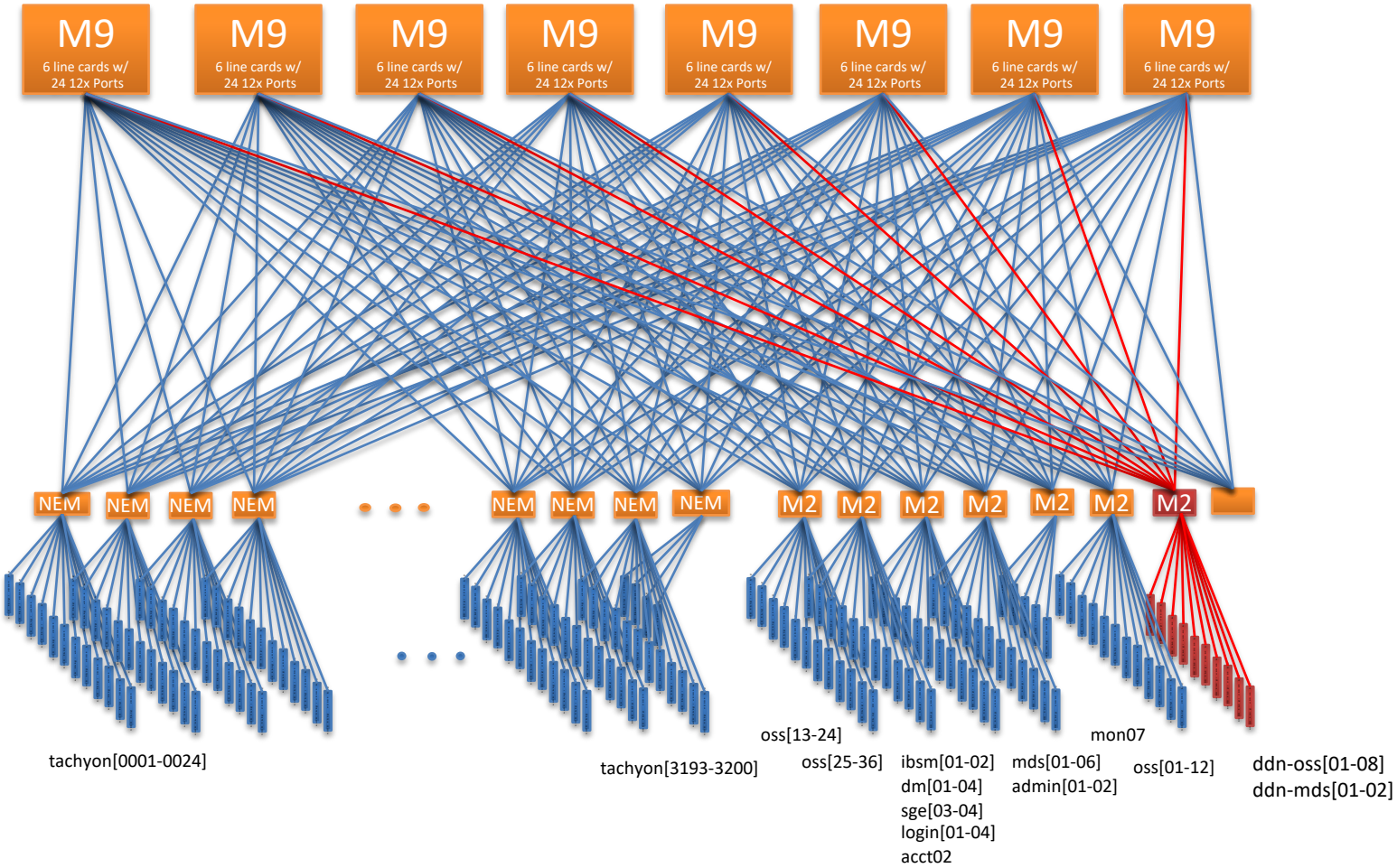
# KISTI Supercomputing Center

- The National Supercomputing Center in Korea
- Provide computational resources and its support to R&D communities in Korea
- In 2018, ~800 Projects, ~2,500 Users, ~200 SCI publications (KISTI-4)



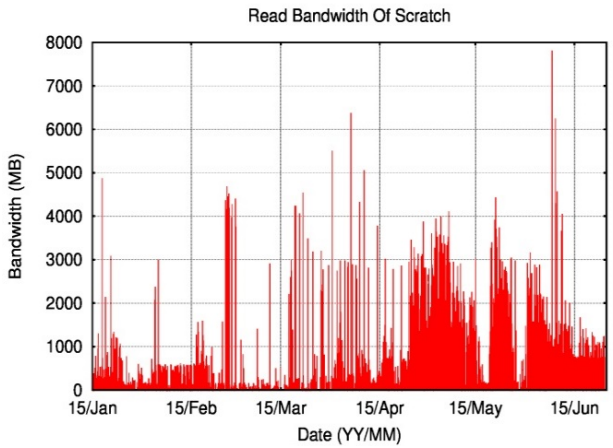
# KISTI-4 Infiniband QDR

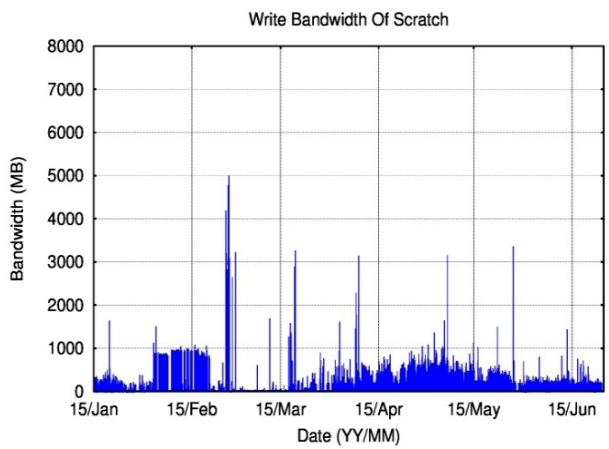
- KISTI-4 system's interconnect consists of 2 layer non-blocking fat-tree IB QDR.
- 1st layer has 134 48p QDR switches and 2nd layer has 8 432p Director switches.
- During 8-years production, ratio of replaced cables : >50% (BER: $10^{-12}$)

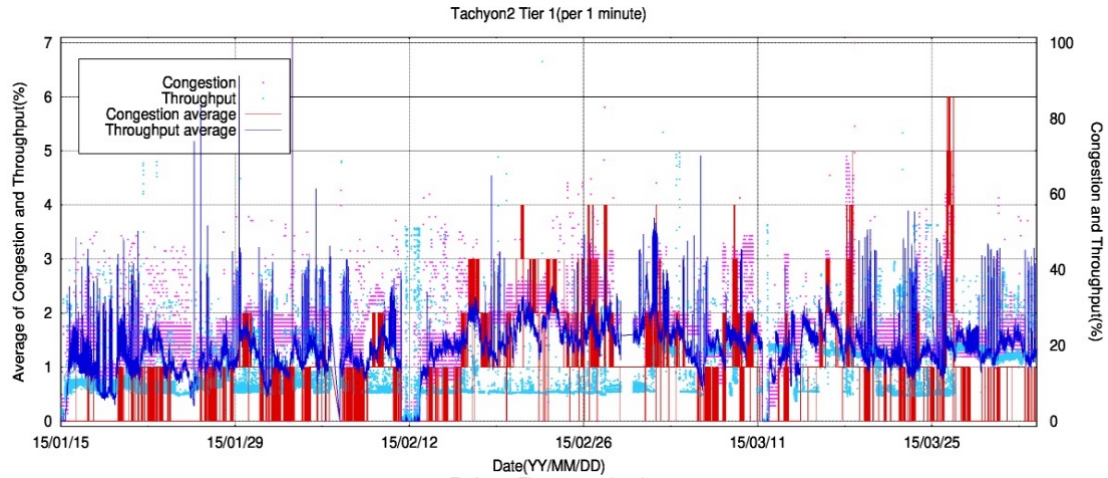


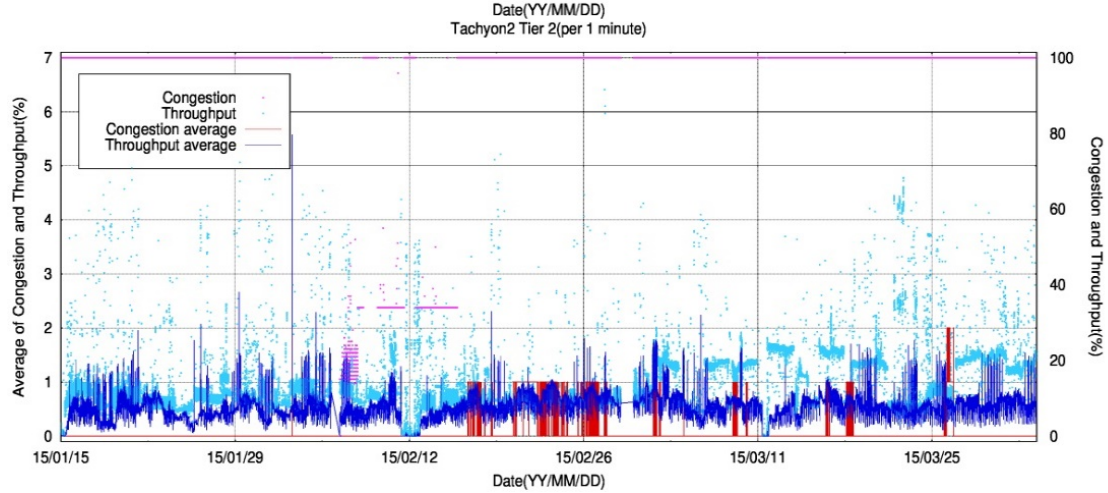
# KISTI-4 Infiniband QDR interconnect

- File IO usually use less than 20% of available Lustre Bandwidth
- Read IO throughput is normally larger than Write IO
- IB Interconnect  are spent only 1~6 % on average and 60% at peak
- Congestion is heavier on L1(compute node -> edge sw ) than on L2(edge sw -> core sw)

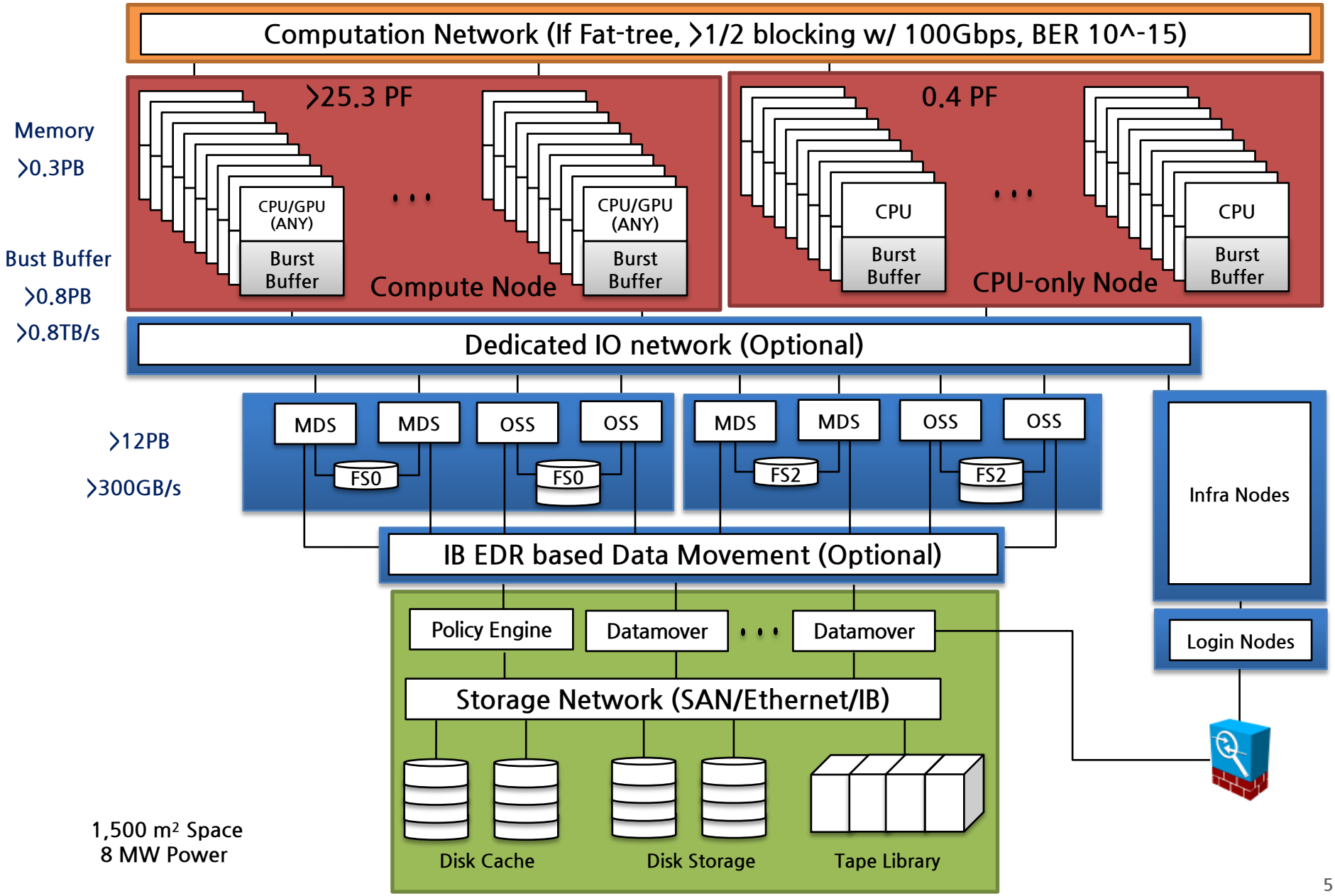


# KISTI-5's system requirements



# 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

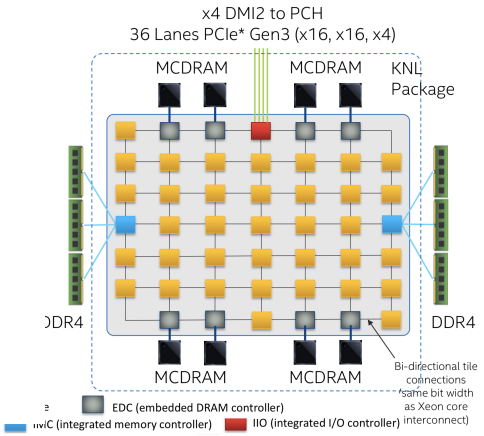
# KISTI-5 compute nodes

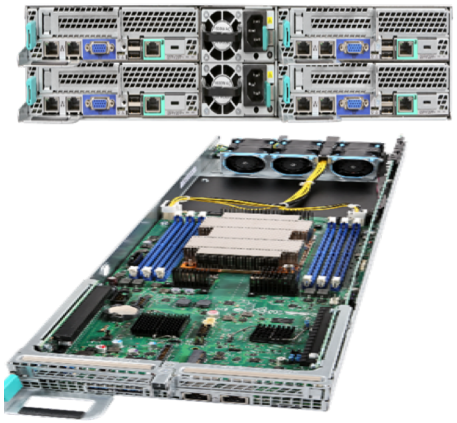
The Largest KNL/OPA based commodity cluster System
Rpeak 25.7PFlops, Rmax 13.9PFlops

## 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 6channel RAM,
- 16GB HBM (460GB/s)
- 1x 100Gbps OPA HFI, 1x On-board GigE Port

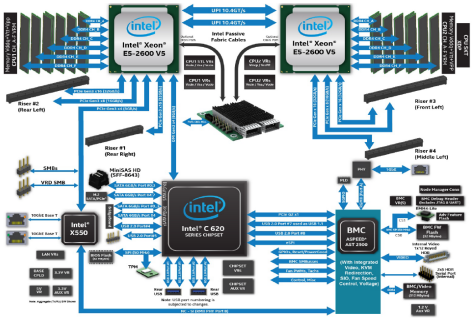


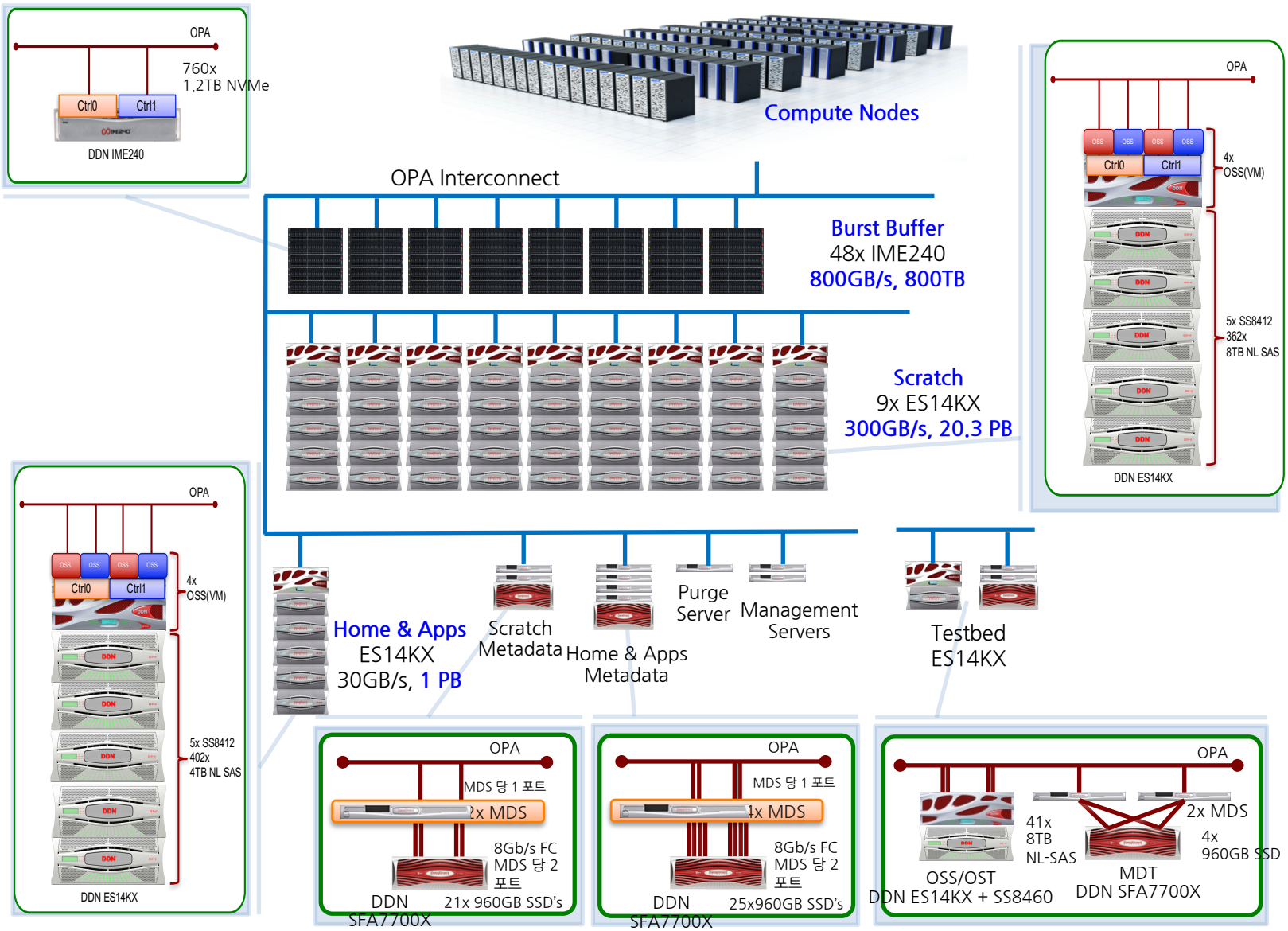
## CPU-only nodes

132 Skylake Computing modules, 4 Racks, 0.4PF

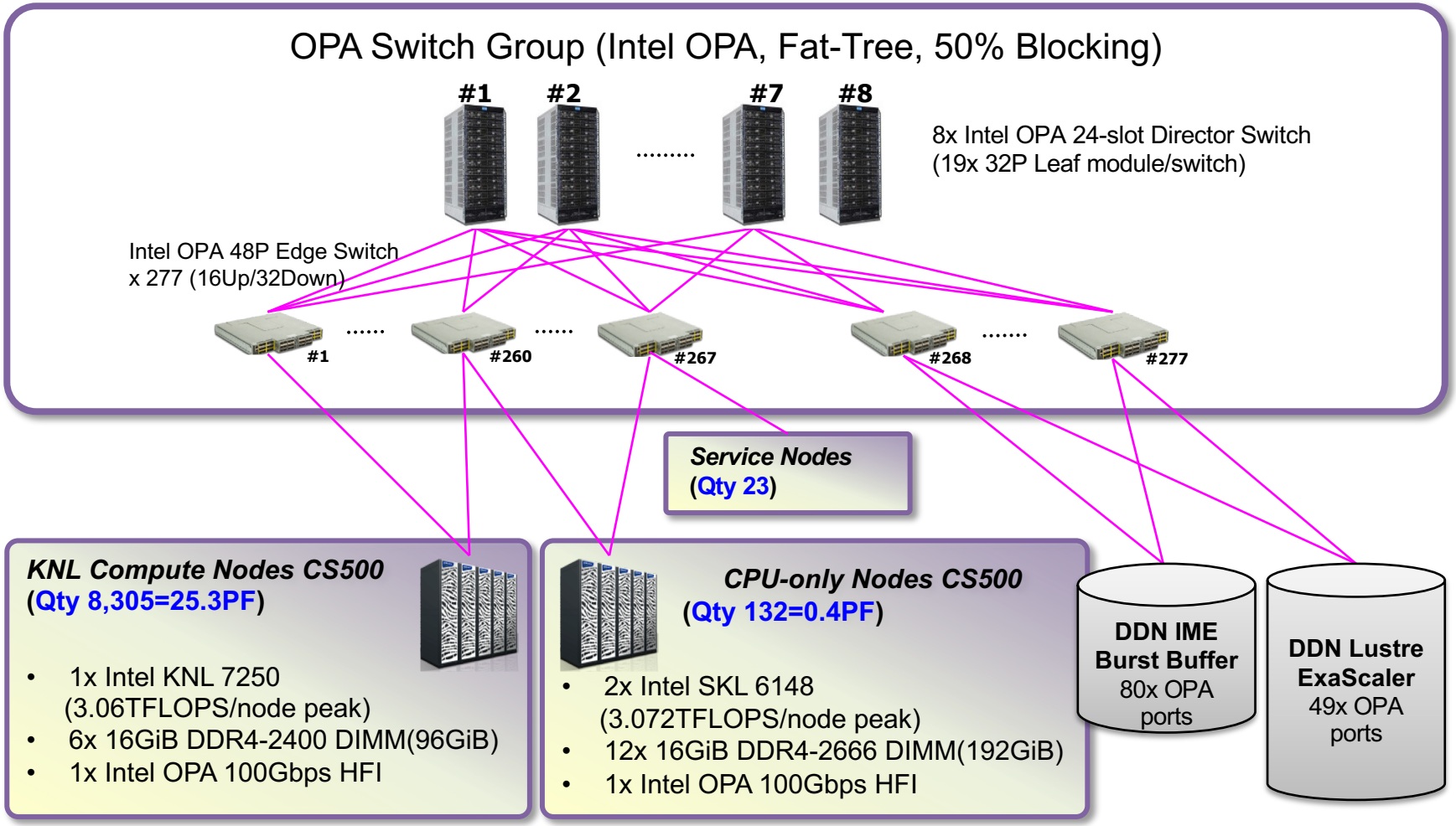
- 2x Xeon SKX 6148 CPUs, 2.4GHz, AVX512
- 192GB (12x 16GB) DDR4-2666 RAM
- 1x Single-port 100Gbps OPA HFI card
- 1x On-board GigE (RJ45) port



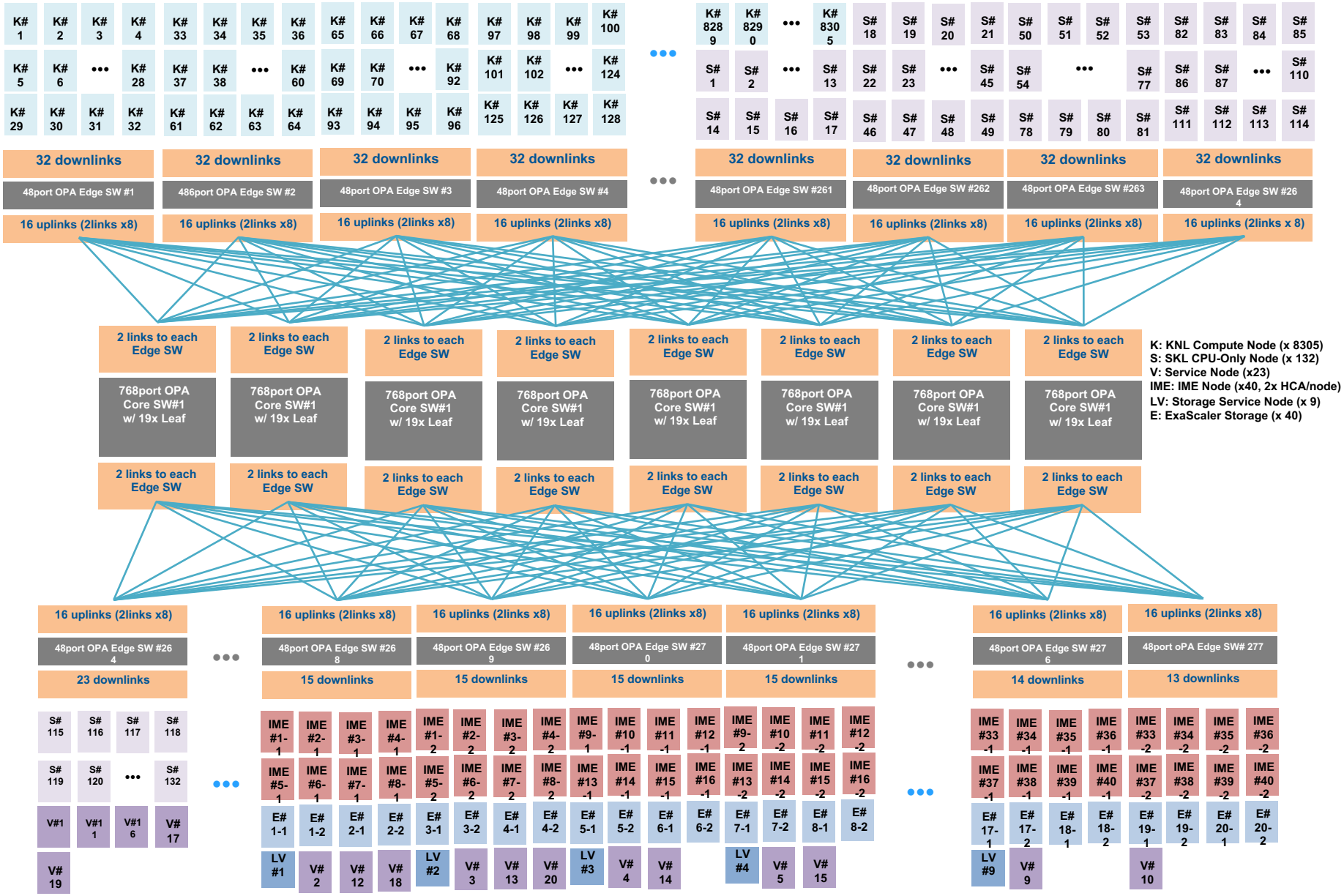
# KISTI-5 Storage System



# KISTI-5 OPA Interconnect



# 2:1 Blocking OPA Interconnect



# Hardware Spec

| Category | Sub-category | Compute Nodes | CPU-only Nodes |
|---|---|---|---|
| Model | System Model | Cray CS500 | Cray CS500 |
| | Number of Nodes | 8,305 | 132 |
| | Number of Racks | 116 | 2 |
| Performance | Peak Performance | 25.3PFLOPS | 0.4PFLOPS |
| CPU Type | Model | Intel Xeon Phi 7250 (KNL) | Intel Xeon 6148 (Skylake) |
| | Peak Performance | 3.0464TFLOPS | 1.536TFLOPS |
| | # of Cores per CPU | 68 | 20 |
| | # of CPUs per Node | 1 | 2 |
| | On-package Memory | 16GB, 490GB/s | - |
| HCA | Model | Intel 100HFA016LS | Intel 100HFA016LS |
| Main Memory (Off-package) | Model | 16GB DDR4-2400 | 8GB DDR4-2666 |
| | Configuration | 16GB x6, 6Ch per CPU | 8GB x12, 6Ch per CPU |
| | Size per Node(GB) | 96GB | 192GB |
| | Bandwidth | 115.2GB/s | 128GB/s |
| | Total Size | 778.5TB | 24.76TB |
| Local Disk | Capacity | N/A | 1TB |
| High Performance Interconnect | Topology | FatTree | |
| | Blocking Ratio | 50% Blocking | |
| | # of Switches | 274x 48-port OPA edge switches<br>8x 768-port OPA core switches | |
| | Bandwidth per Port | 12.3GB/sec | |
| | Bisectional Bandwidth | 27,060GB/sec | |
| Burst Buffer | # of Servers | DDN IME240 Servers 40ea | |
| | Configurations of Disks | 19x 1.2TB NVMe SSD, 2x 0.45TB NVMe SSD | |
| | Total Capacity | 0.8PB usable | |
| | Bandwidth | 20GB/sec per server, 0.8TB/s total | |
| Parallel File System | Scratch MDS | DDN SFA7700X 1ea, MDS 2ea | |
| | Home/Apps MDS | DDN SFA7700X 1ea, MDS 4ea | |
| | Scratch OSS | 4x OSS(embedded) per ES14KX<br>DDN ES14KX 9ea w/ 360x 8TB each | |
| | Home/Apps OSS | 4x OSS(embedded),<br>DDN ES14KX 1ea w/ 400x 4TB | |
| | Total Capacity | 20PB usable | |
| | Bandwidth | 0.3TB/s | |

# Software Spec

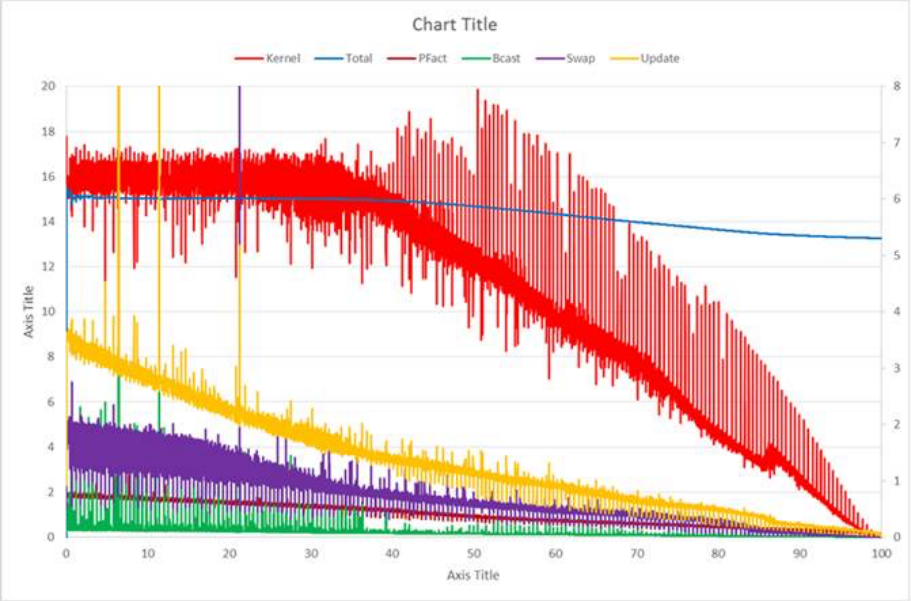
- For provisoning time speed-up, 132 SKX nodes are configured as relay servers used in rebooting on periodic maintenance window
- The ethernet network of the whole system is split into 16 VLANs to avoid MAC flooding due to the limit of the MAC entries of ethernet switches
- The delay of start-up and run time of MPI jobs due to PBS integration with MPI has been enhanced in several months.

**Scalability for 8400 diskless off-the-shelf Linux Cluster System**

| Specification | Products |
|---|---|
| Cluster Manager (Provisioing, Mgmt, OpenStack) | Bright Computing |
| Operating System | CentOS |
| Workload Manager | PBS Pro |
| Compilers | Intel, Cray PE, GNU |
| MPI Libraries | Intel MPI, Open MPI, MVAPICH2 |
| Interconnect Software | Intel OPA |
| Parallel File System | Lustre<br>DDN ES |
| Burst Buffer | IME |
| Debugger | Allinea DDT |
| Profiler | CrayPat, Intel Vtune |
| User Account Management | LDAP |
| Commercial Software | ABAQUS, ANSYS, LS-DYNA<br>NASTRAN, Gaussian |

# 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 | 11th (Jun 2018) |
| HPCG | Large-scale Sparse Matrix Computation<br>Similar to normal user applications | 8,250(KNL) | 0.39PF | 8th (Nov 2018) |
| Graph500 | Breadth-First Search,<br>Interconnect Performance | 1,024(KNL)<br>3,844(KNL) | 1,456GTEPS<br>1,048GTEPS | 10Th (Jun 2019)<br>23rd (Nov 2018) |
| IO500 | Various IO Workloads | 2,048(KNL) | 160.67 | 2nd (Nov 2018) |

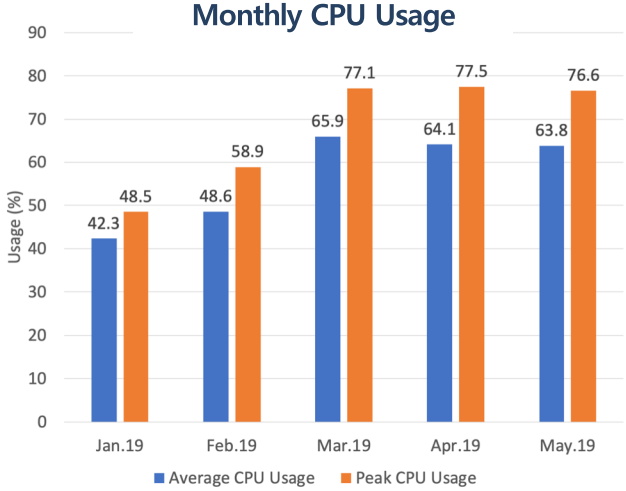


Source: Intel Dmitry Nemirov

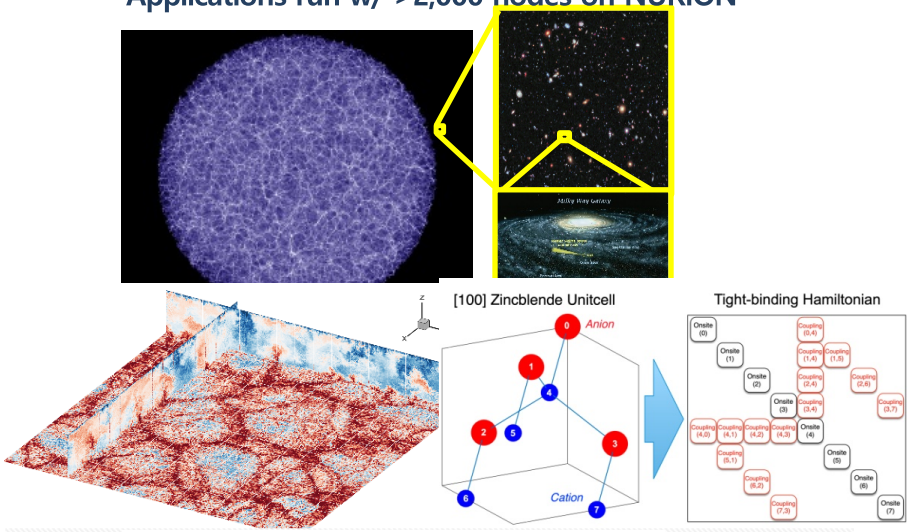
# NURION Production Status

- CPU usage are getting higher since system' production started in Dec 2018
- Large jobs with running >4,096 CPU cores has spent 68% of the CPU times of the whole system

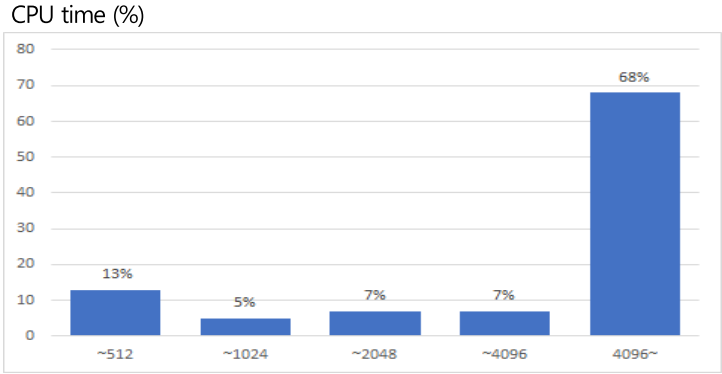
## Monthly CPU Usage



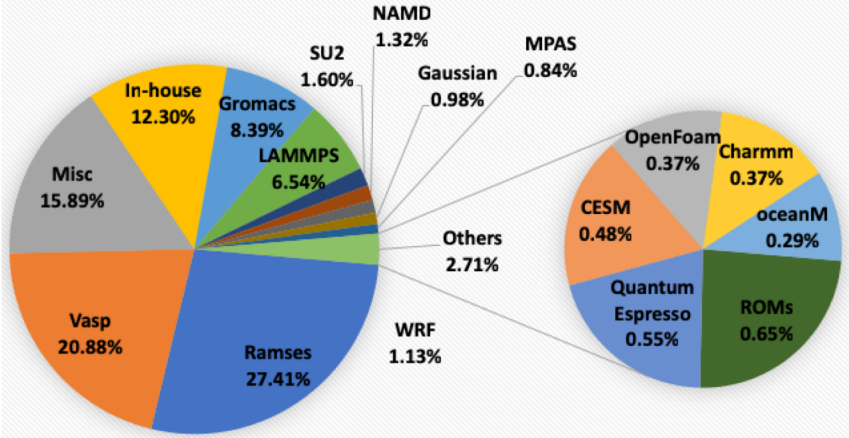

## Applications run w/ >2,000 nodes on NURION




## Users Jobs distribution by Job size (CPU cores)

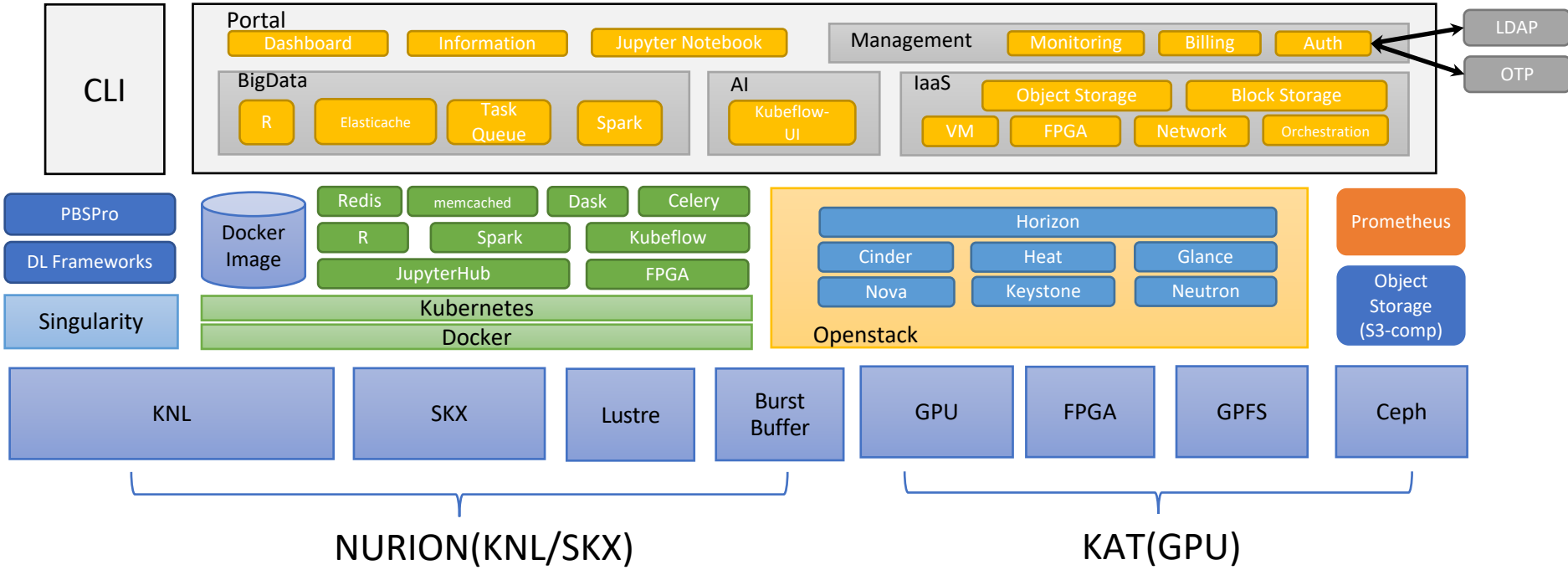



## Applications Distribution by CPU Usage



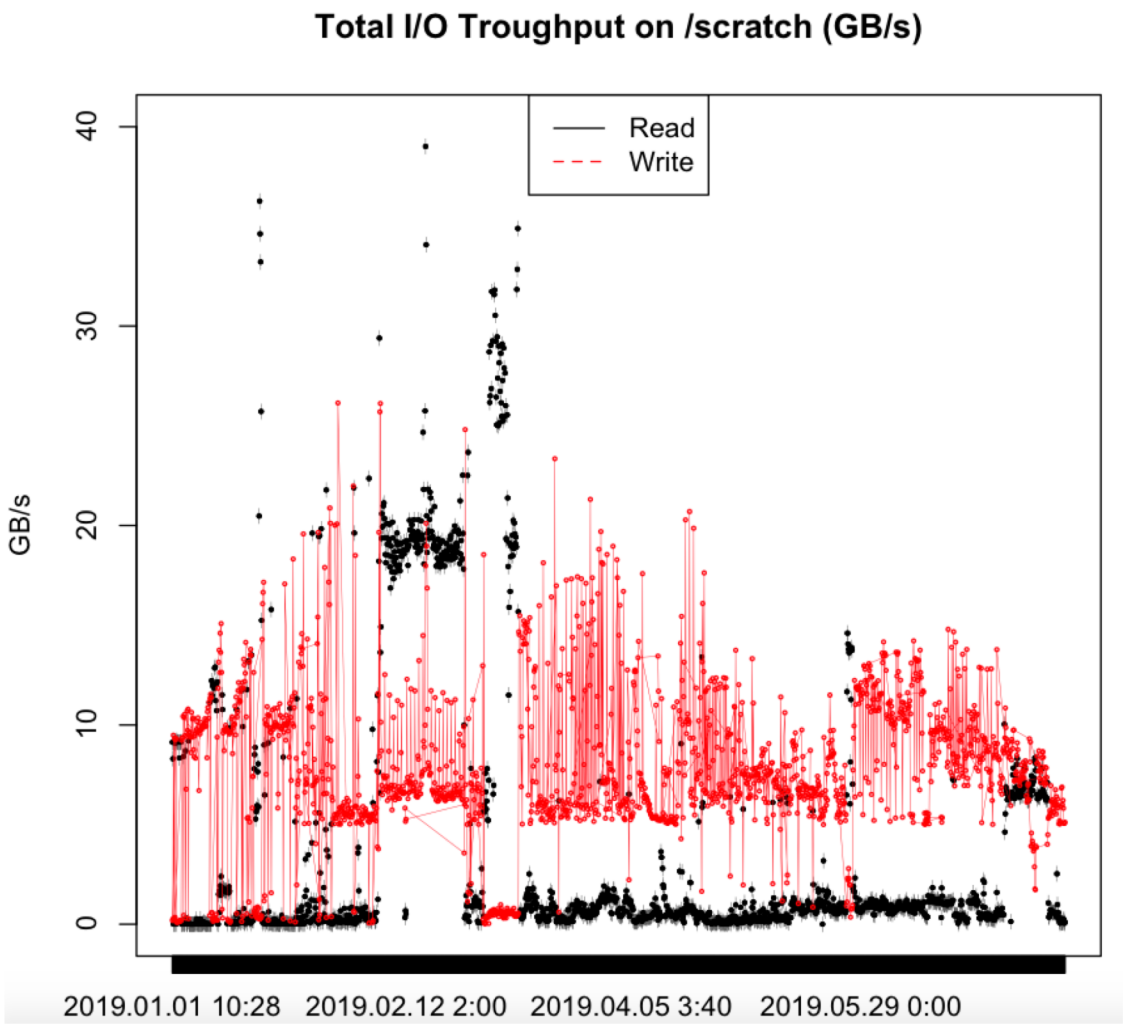
# Ongoing Work

- 5% of user jobs has used singularity image for ML software framework since Dec 2018
- We will launch more flexible user environments based on Container and OpenStack on the top of Bright computing and open-source
- Also Data analysis Platform composed of Spark, In-memory DB, and Jupyter Note on top of Docker and Kubernetes



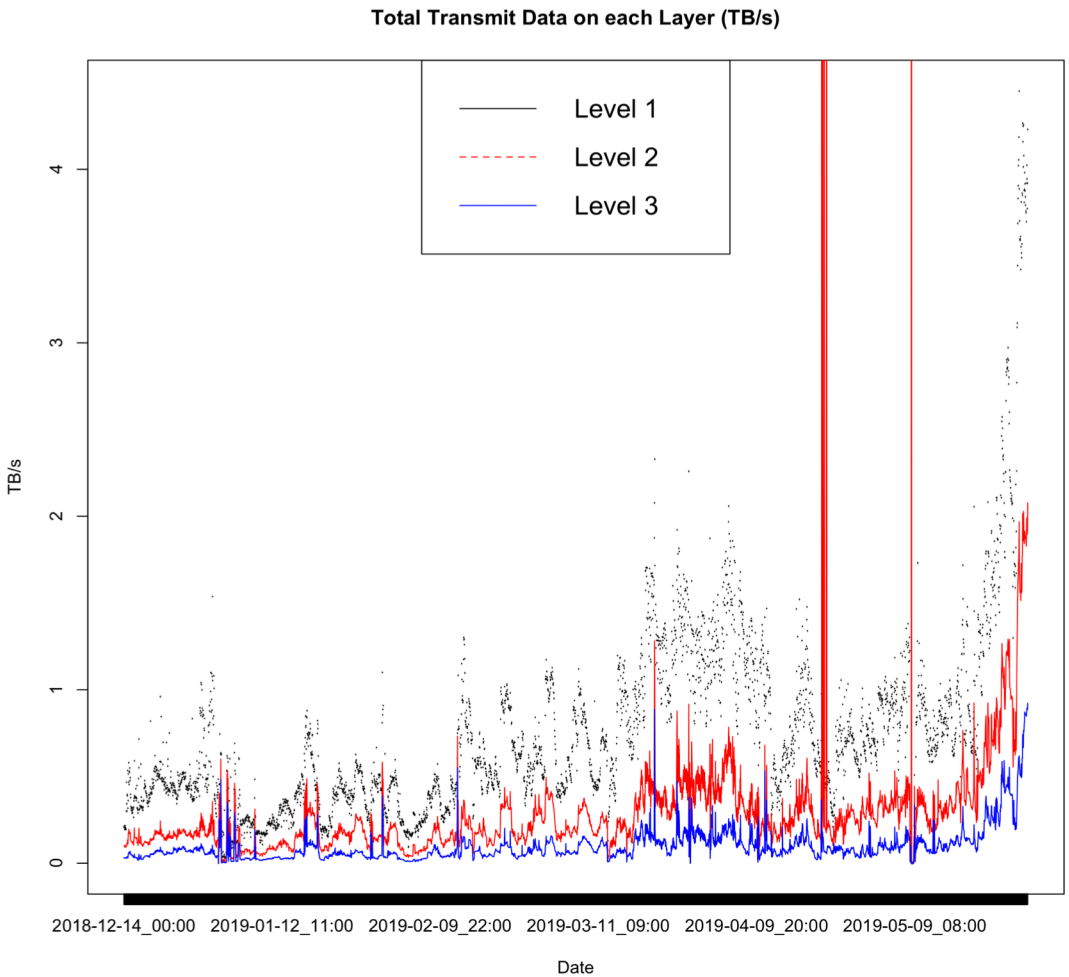
# File I/O throughput statistics

- Read/Write I/O usually use less than 10% of Lustre's peak bandwidth (300GB/s) and occupies 2~3% of real traffic (1~4TB/s) on OPA
- Write I/O is much more than Read I/O throughput in production



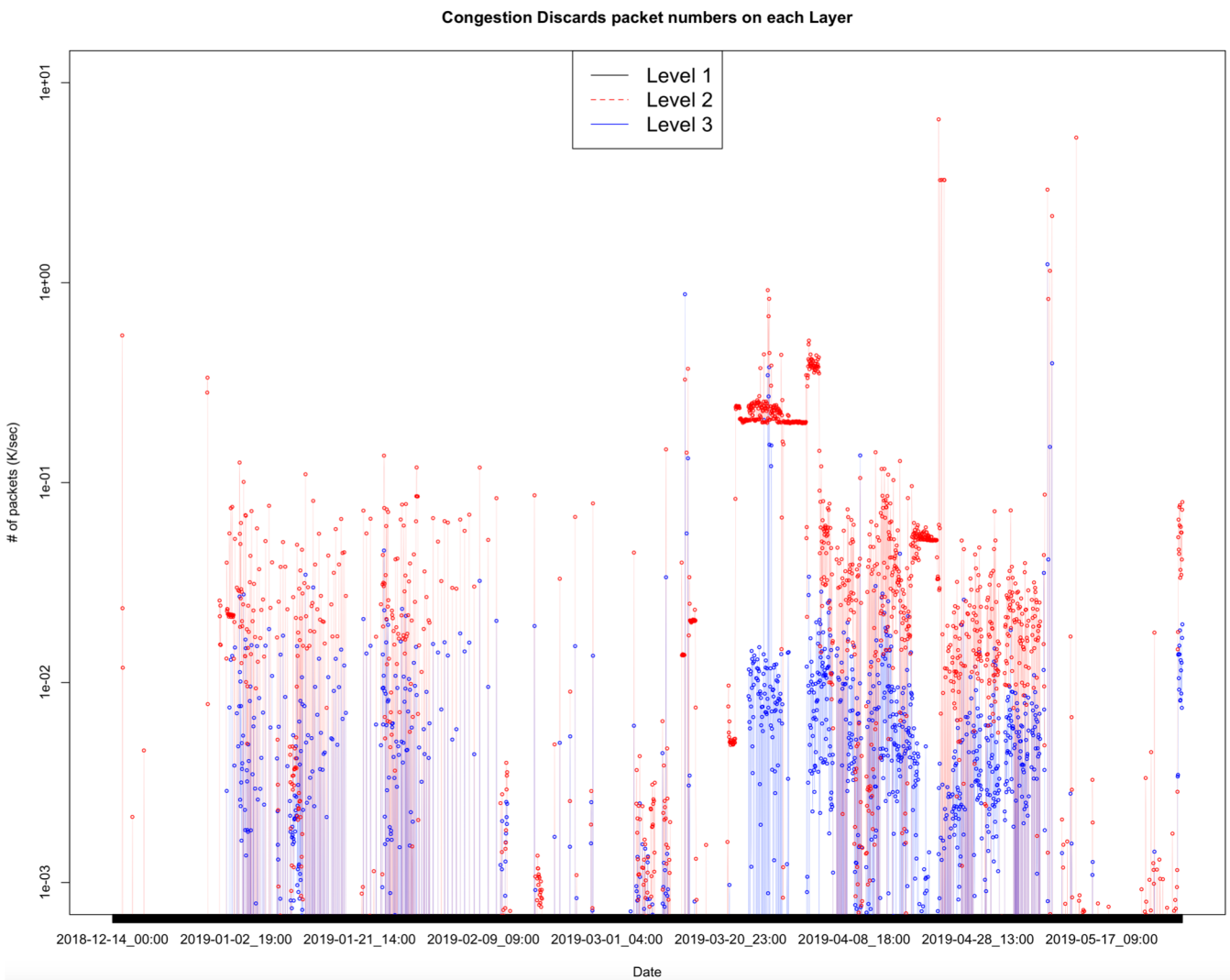
# Bandwidth statistics on each OPA Layer

- The graph below are smoothed out due to long measure cycle(1hour) and average of the spans
- L1 represents uplinks from compute nodes to edge switches, L2 edge to leaf, and  L3 leaf to spine.
- L1 data traffic is relatively low comparing with total available bandwidth (bisection bandwidth, 27TB/s)
- Traffic ratio of L1 : L2 : L3  is 1 : 0.37 : 0.13 on average for 5 months



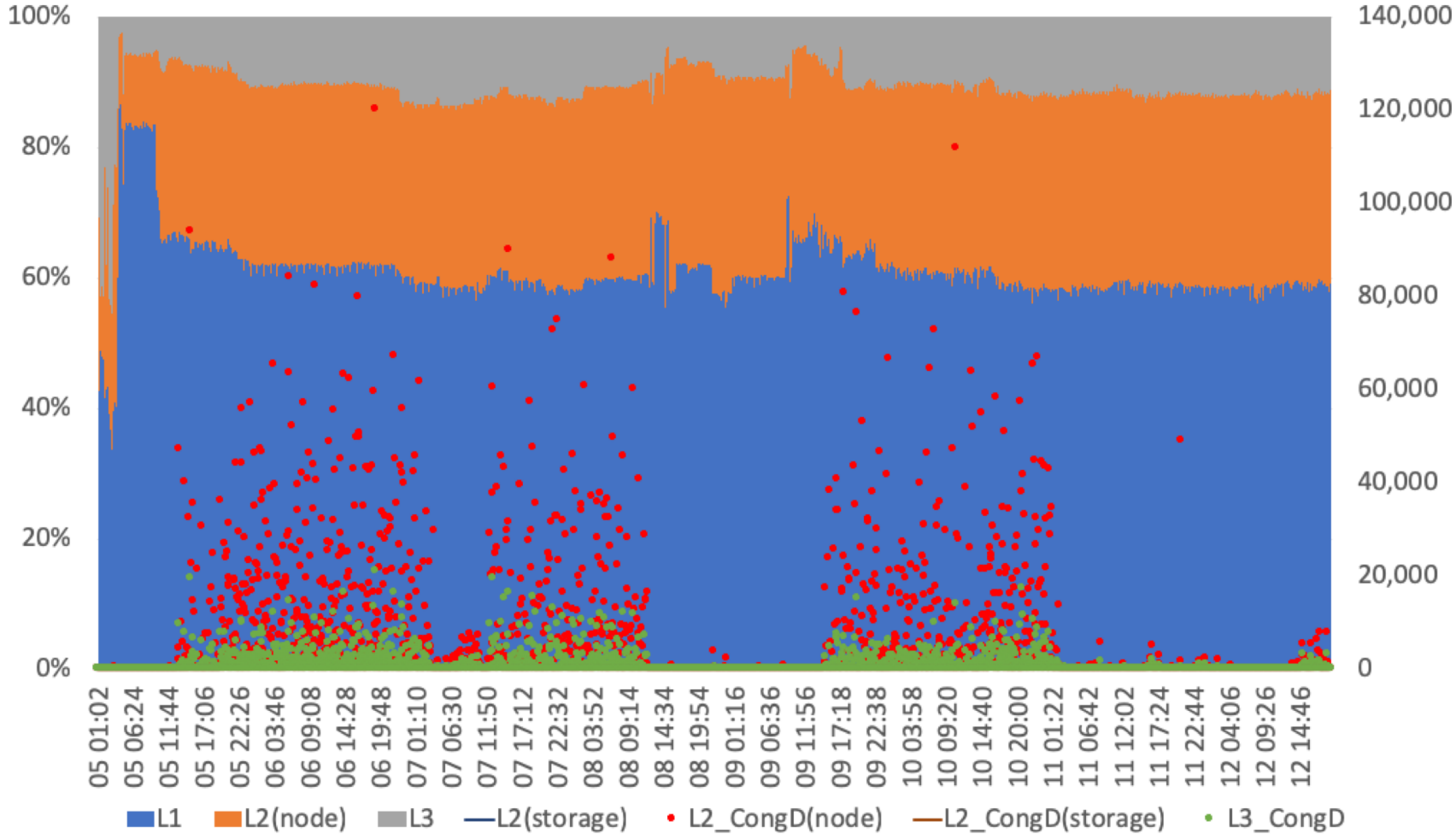
# Congestion Discards on each OPA Layer

- Packet discards by port congestion has been usually occurred in Level2(edge → leaf) and then Level3(leaf → spine) while L1(nodes → edge) has been normal



# Throughput and Congestion in detail

- Throughput and Packet discards by port congestion every 2min during 5days
- The similar pattern but congestion is more differentiated by layer of measuring



# Congestion Benchmark

- The interconnect shows relatively good performance in the network latency and bandwidth test under proxy load of congestion by GPCNET Benchmark

GPCNET benchmark

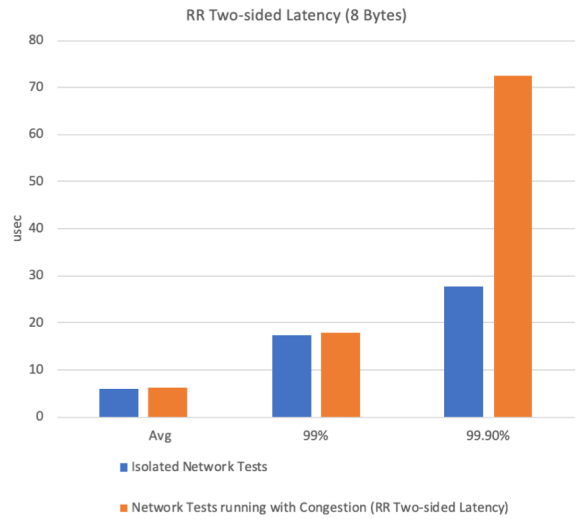
(NetworkLoad Tests v1.0)

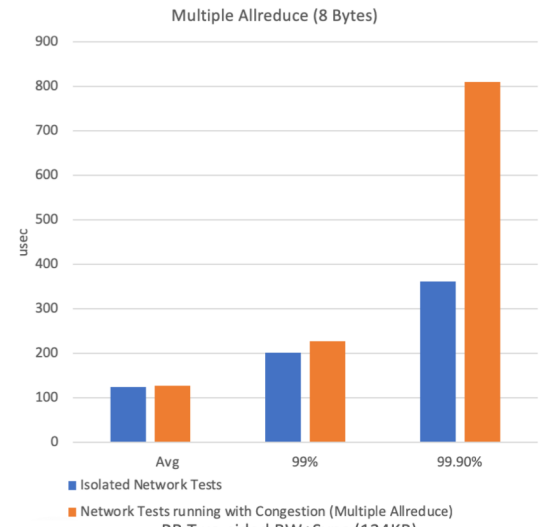
Test with 128,000 MPI ranks

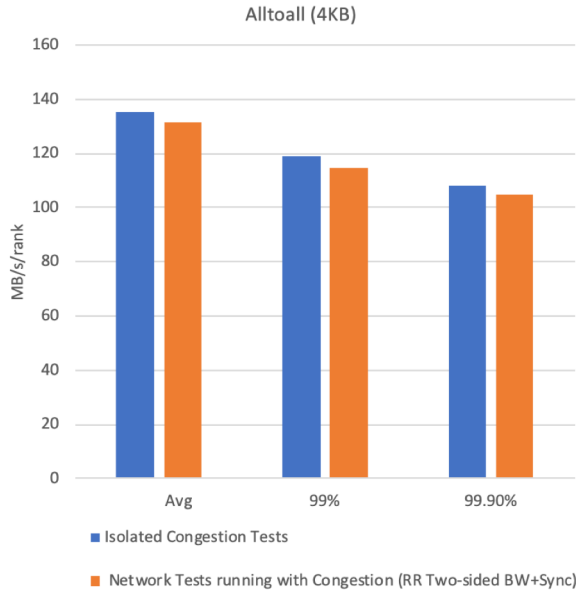
(16 ppn on 8,000 nodes)

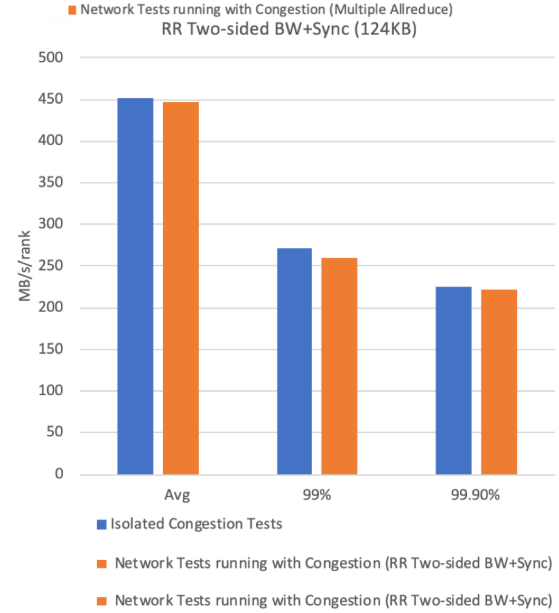
- 1,600 nodes running Network Tests
- 6,400 nodes running Congestion

  (1,600 nodes per congestor)



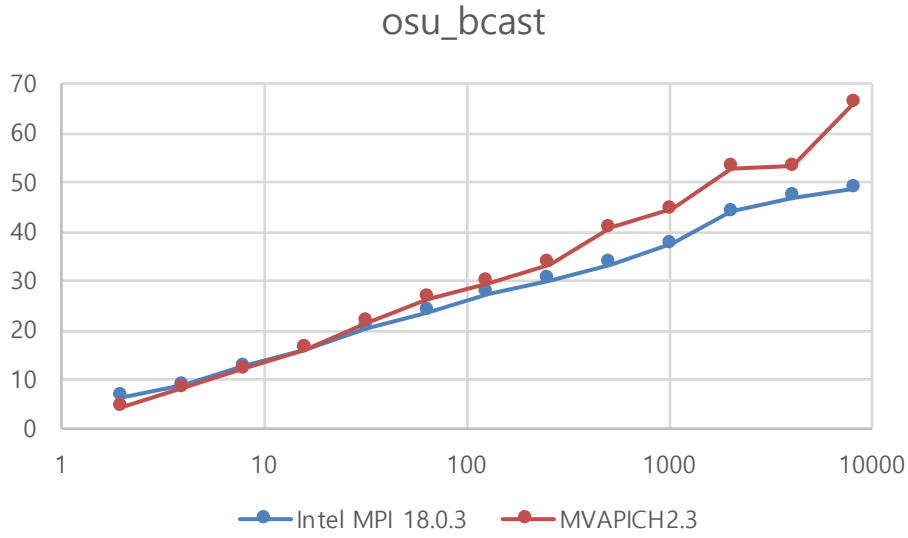



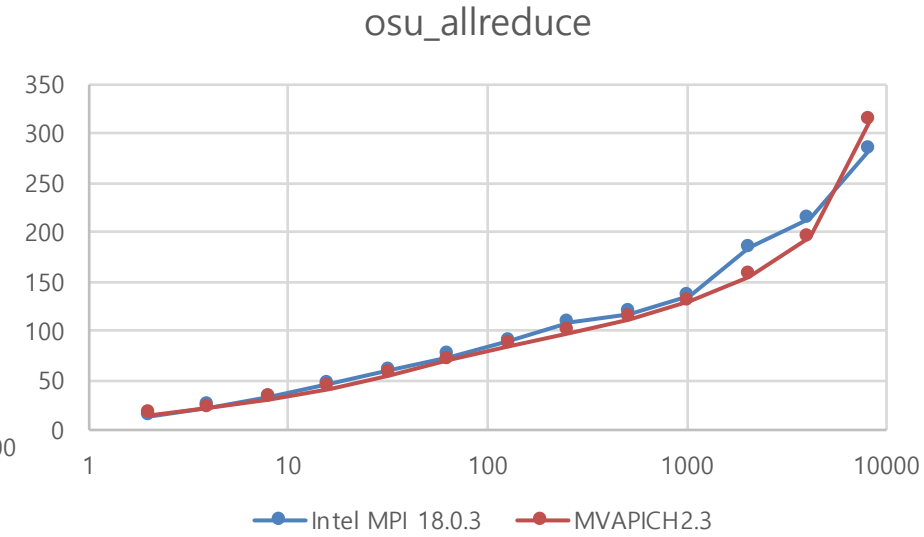



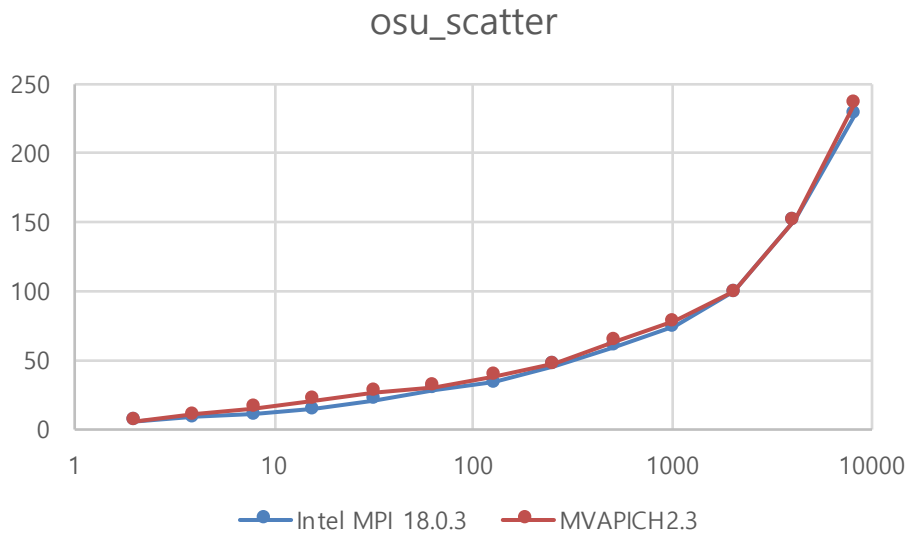



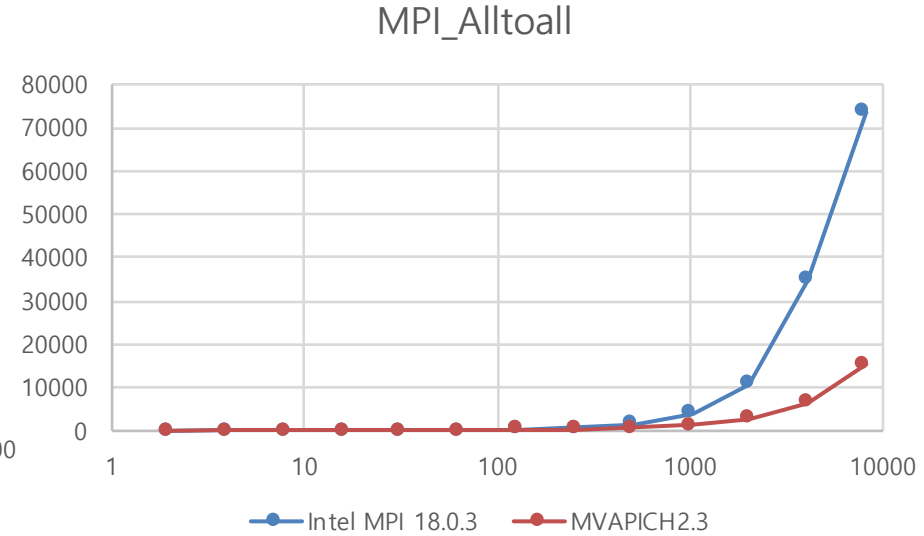
# IMB Benchmark

- The performances of all MPI collective communication shows similar shape as it scale
- The algorithm has to be properly chosen to get good collective performance over IntelMPI



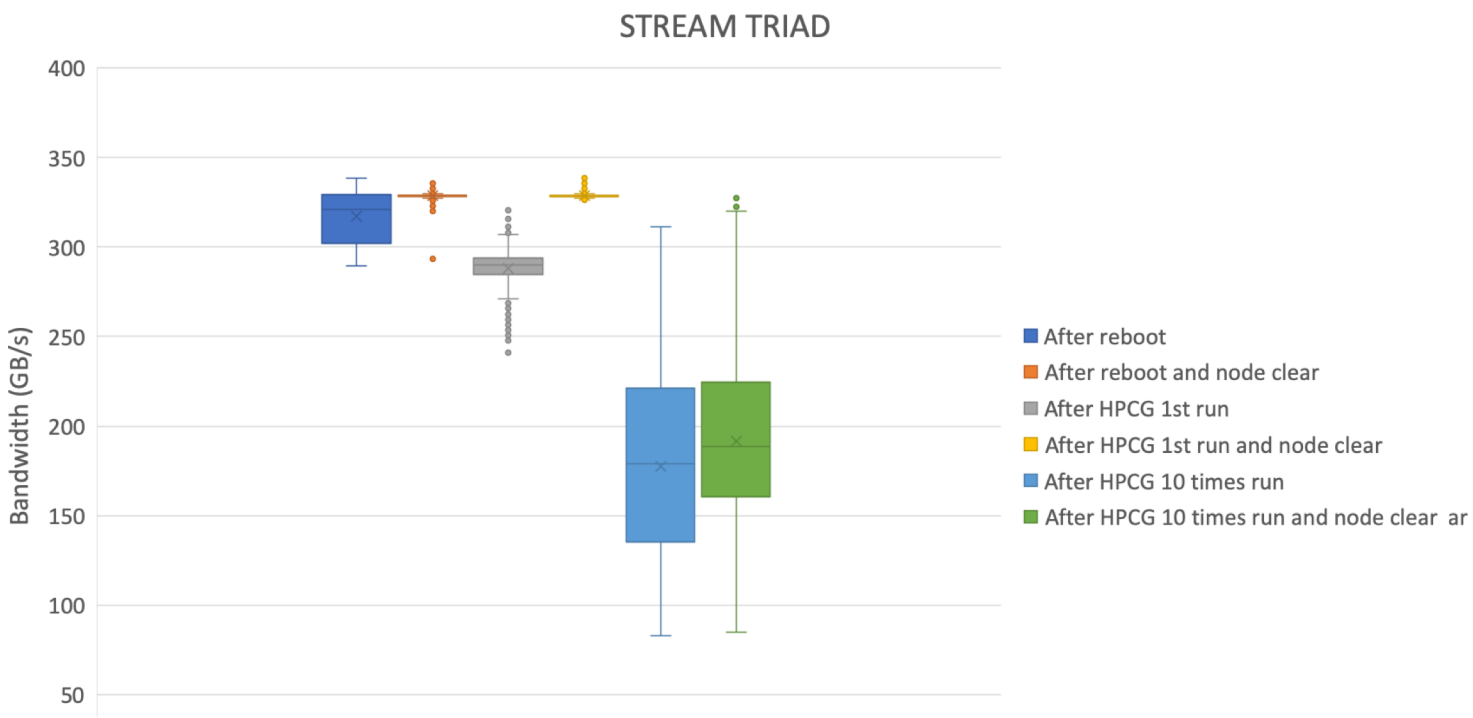



# Performance Degradation on KNL Cache mode

- In cache mode of MCDRAM, KNL's memory bandwidth performance has been degraded with the lapse of time after reboot
- Node clearing method such as Zone-sort, Drop-cache, etc are not enough
- The performance impact are different depending on applications types, HPCG gets slower but HPL has no difference



Thank you