Ohio Supercomputer Center

An OH·TECH Consortium Member

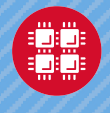
Overview of OSU INAM Deployment at Ohio Supercomputer Center and Live Demo

Heechang Na, hna@osc.edu

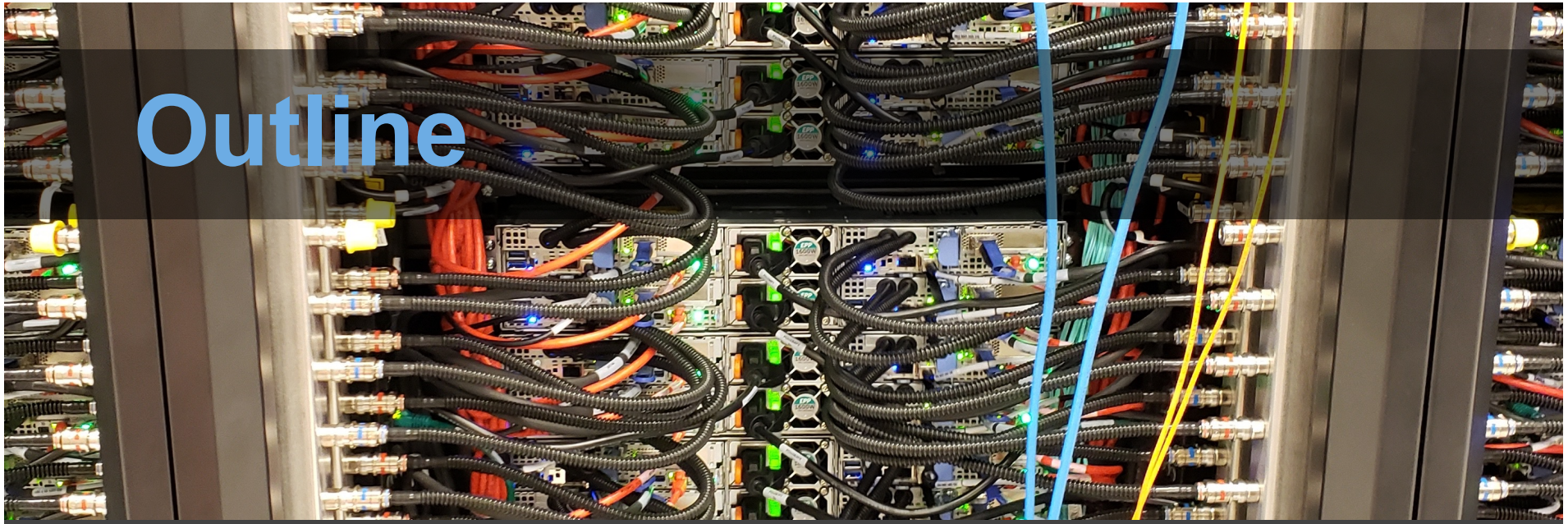
Pouya Kousha, kousha.2@buckeyemail.osu.edu

Trey Dockendorf, tdockendorf@osc.edu

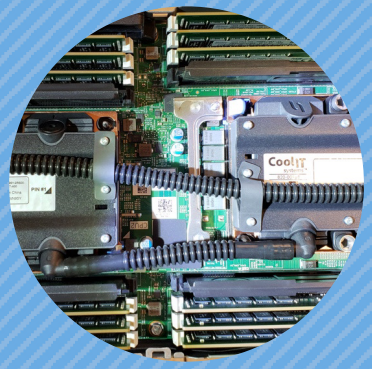
Karen Tomko, ktomko@osc.edu



Outline



- Overview of OSC's Systems & Fabric
- OSU INAM
- INAM at OSC
- Demo



Overview of OSC's Systems and Fabric

"To err is human, but to really foul things up you need a computer." – Paul Ehrlich

System Status (Aug 2021)

COMPUTE	Owens	Pitzer	Pitzer Expansion
Date	2016	2018	2020
Cost	$7 million	$3.4 million	$4.3 million
Theoretical Perf.	~1.6PF	~1.3PF	~2.6 PF
Nodes	824	260	398
CPU Cores	23,392 Broadwell	10,560 Skylake	19,104 Cascade Lake
RAM	~120 TB	~ 70.6 TB	~ 93.7 TB
GPUs	160 NVIDIA P100	64 NVIDIA V100	102 NVIDIA V100

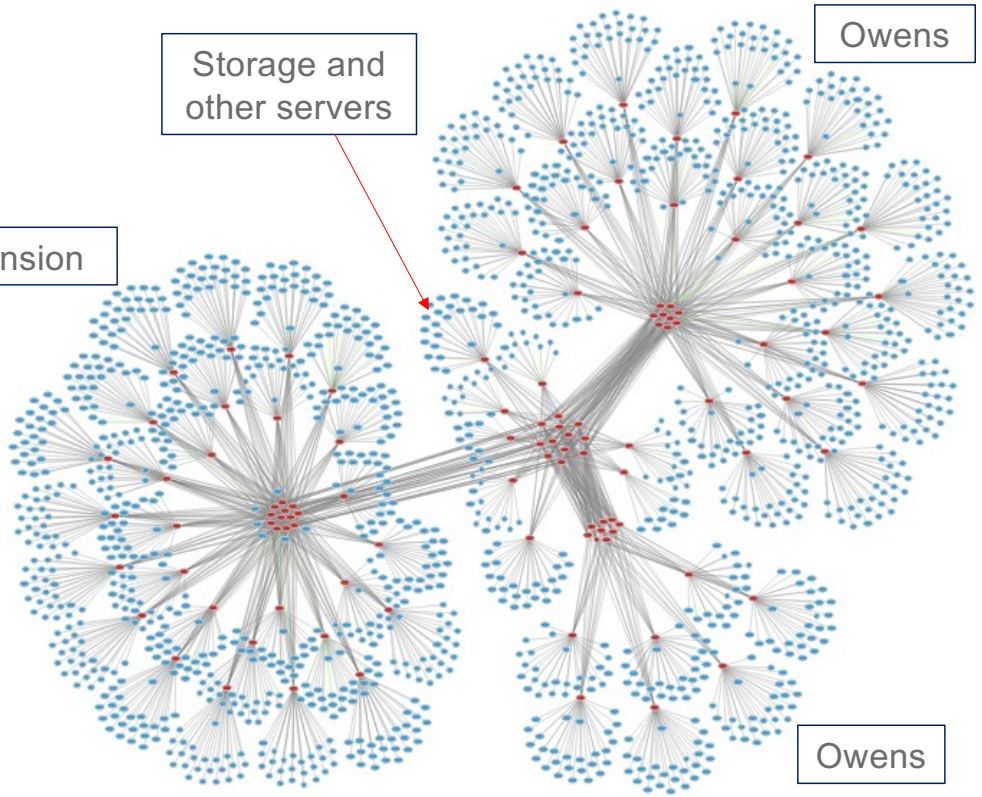
STORAGE	NetApp	DDN	IBM	Tape Library
Year Acquired	2016	2016	2020	2016-2018
Cost	$0.4 million	$2.9 million	$1.7 million	$0.6 million
Capacity	0.8 PB	4.8 PB	8.6 PB	10+ PB

Not your lab's fabric

OSC has a single integrated IB fabric

- Fabric Size: 109 IB switches, 1,544 hosts (1,482 compute nodes),
- Currently 2 compute clusters, 3 generations of hardware
- RDMA access to 2 generations of GPFS filesystems
- Multiple generations of InfiniBand (CX-4/CX-5 EDR)
- Different switch sizes and topologies for each cluster
- Mellanox UFM and routing chains for the complex topology

OSC's Fabric Topology



- Level 3: 12 EDR spine switches
- 6 EDR islands
- Level 2: 38 switches
- Level 1: 59 switches

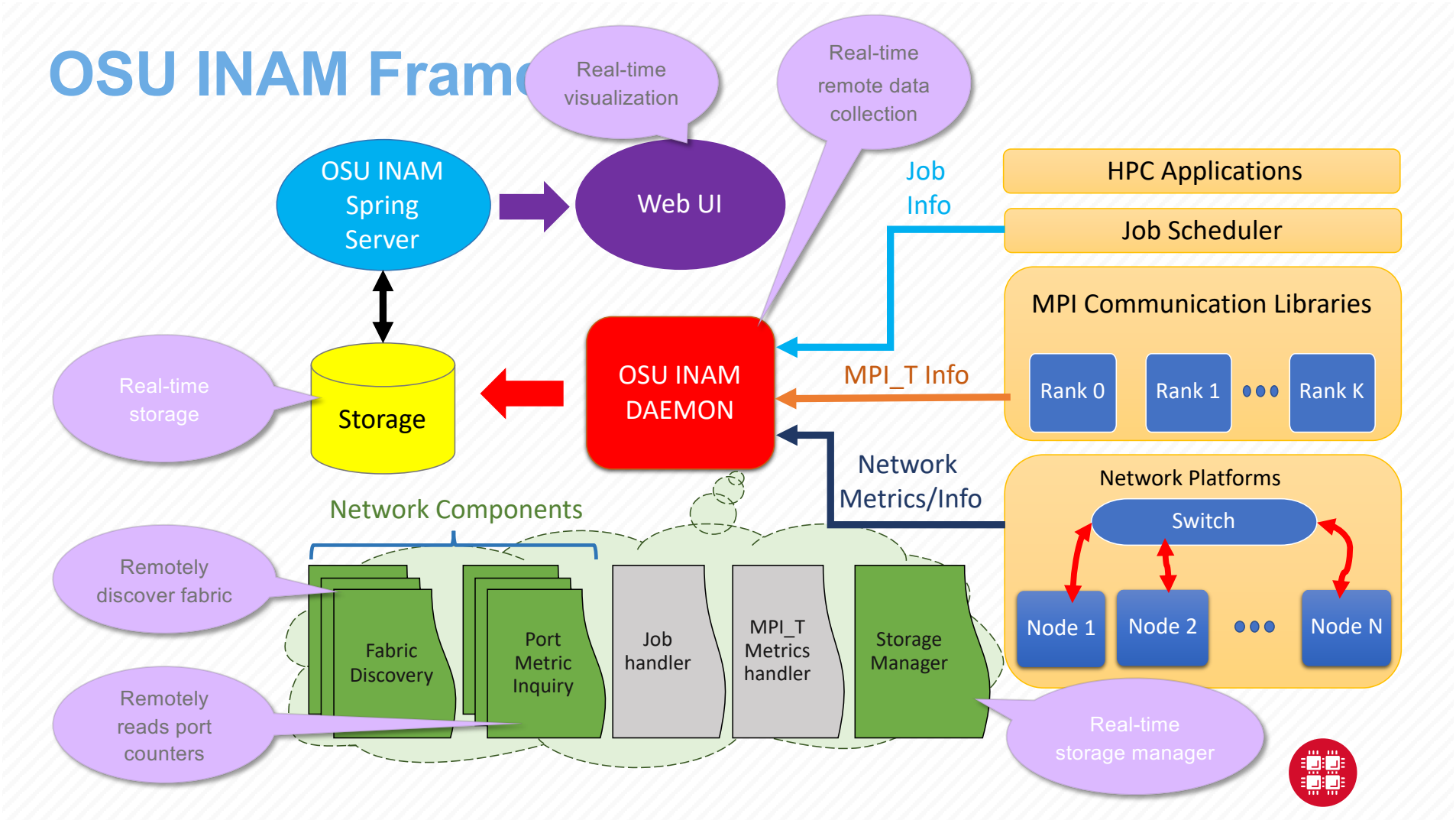
OSU INAM

"Alone we can do so little; together we can do so much." – Helen Keller

Overview of OSU InfiniBand Network Analysis and Monitoring (INAM) Tool

- **A network monitoring and analysis tool that is capable of analyzing traffic on the InfiniBand network with inputs from the MPI runtime**
  - http://mvapich.cse.ohio-state.edu/tools/osu-inam/
- **Monitors IB clusters in real time by querying various subnet management entities and gathering input from the MPI runtimes**
- **Capability to analyze and profile node-level, job-level and process-level activities for MPI communication**
  - Point-to-Point, Collectives and RMA
- **Ability to filter data based on type of counters using "drop down" list**
- **Remotely monitor various metrics of MPI processes at user specified granularity**
- **"Job Page" to display jobs in ascending/descending order of various performance metrics in conjunction with MVAPICH2-X**
- **Visualize the data transfer happening in a "live" or "historical" fashion for entire network, job or set of nodes**
- **Sub-second port query and fabric discovery in less than 10 mins for ~2,000 nodes**

- **OSU INAM 0.9.6 released**
  - Support for PBS and SLURM job scheduler as config time
  - Ability to gather and display Lustre I/O for MPI jobs
  - Enable emulation mode to allow users to test OSU INAM tool in a sandbox environment without actual deployment
  - Generate email notifications to alert users when user defined events occur
  - Support to display node-/job-level CPU, Virtual Memory, and Communication Buffer utilization information for historical jobs
  - Support to handle multiple job schedulers on the same fabric
  - Support to collect and visualize MPI_T based performance data
  - Support for MOFED 4.5, 4.6, 4.7, and 5.0
  - Support for adding user-defined labels for switches to allow better readability and usability
  - Support authentication for accessing the OSU INAM webpage
  - Optimized webpage rendering and database fetch/purge capabilities
  - Support to view connection information at port level granularity for each switch
  - Support to search switches with name and lid in historical switches page
  - Support to view information about Non-MPI jobs in live node page

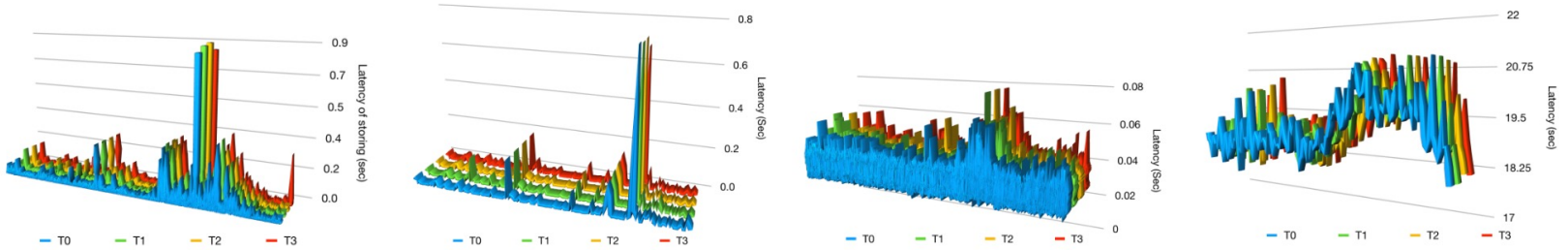

OSU INAM Frame
Real-time visualization
Real-time remote data collection
OSU INAM Spring Server
Web UI
Job Info
HPC Applications
Job Scheduler
MPI Communication Libraries
Rank 0
Rank 1
Rank K
Real-time storage
Storage
OSU INAM DAEMON
MPI_T Info
Network Metrics/Info
Network Platforms
Switch
Node 1
Node 2
Node N
Network Components
Remotely discover fabric
Fabric Discovery
Port Metric Inquiry
Job handler
MPI_T Metrics handler
Storage Manager
Remotely reads port counters
Real-time storage manager

Best Practice for OSC INAM and Thread Load balancing

- What is the proper allocation of number of thread based on number of CPU cores for each module inside OSU INAM Daemon?

Cluster size	fabric discovery	performance port inquiry	MPI_T and job thread	Purge thread
< 500	2	2+	1	2
500< size <1000	4	8+	1	2
> 1000 (OSC)	**8**	**16+**	**2**	**2**

- Load Balancing of Threads in Port Inquiry



(a) Timing of write phase for each thread for fabric discovery (b) Timing of write phase for each thread for port inquiry (c) Timing of query phase for each thread for port inquiry (d) Timing of query phase for each thread for fabric discovery

INAM at OSC

"How wonderful that we have met with a paradox. Now we have some hope of making progress." – Niels Bohr

INAM Project Collaboration

Central Question:

Can a high performance and scalable tool be designed which is capable of analyzing and correlating the communication on the fabric with behavior of HPC/Big Data applications through tight integration with the communication runtime and the job scheduler?

Project Team:

OSU: Pouya Kousha, Aamir Shafi, Hari Subramoni, DK Panda

OSC: Trey Dockendorf, Heechang Na, Karen Tomko

Status:

- INAM has been running at OSC on production systems for more than two years
- Iterative test and development cycle between OSC/OSU

Thanks to the Nations Science Foundation for supporting this project
NSF OAC-1664137: Started July 2017 Ended June 2020

Configuring for OSC

- Integration with the resource manager
  - INAM configured for Slurm
    - Alpha-numeric job names
    - Previously with Torque/MOAB
- Data collection parameters
  - Collection rate
    - 30 sec intervals for fabric counters
    - 30 second intervals for polling batch servers
  - Job history retained for 1 week
    - DB uses ~70GB of disk space
- MVAPICH2-X integration
  - Config file replicated on filesystem available to compute nodes

Impact of OSU/OSC collaboration on INAM

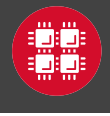
- Performance
- 15x reduction in fabric discovery time
- Caching of Rendered Fabric Diagram
    - Time reduced from ~2 minutes to just a few seconds
- Database Optimizations
    - Identified DB tuning parameters
        - E.g. batch insertions, indexing, sharding
    - Improved Fault-tolerance of Database
        - Automatic restart of MySQL service
- Installation and Configuration
    - Focus: make it easier to automate deployment of INAM
    - Single RPM with all components
    - Best practice and DB calculator
- User interface refinements and suggestions
    - Search by LID or destination port no.
    - Adding MV2-X data to historical plot
    - Identified various bugs
        - e.g. Correct unit displayed on a graph

More info:

- http://mvapich.cse.ohio-state.edu/tools/osu-inam/
- **Pearc 21 paper:** INAM: Cross-stack Profiling and Analysis of Communication in MPI-based Applications, *P. Kousha, K. Ram, M. Kedia, H. Subramoni, A. Jain, A. Shafi, DK Panda, T. Dockendorf, H. Na and K. Tomko. Practice and Experience in Advanced Research Computing 2021, Jul 2021*

INAM Demo

- Heechang
  - Network View for OSC clusters
- Pouya
  - Features with MVAPICH2-X



OH·TECH

Ohio Technology Consortium

A Division of the Ohio Department of Higher Education

info@osc.edu

osc.edu

twitter.com/osc

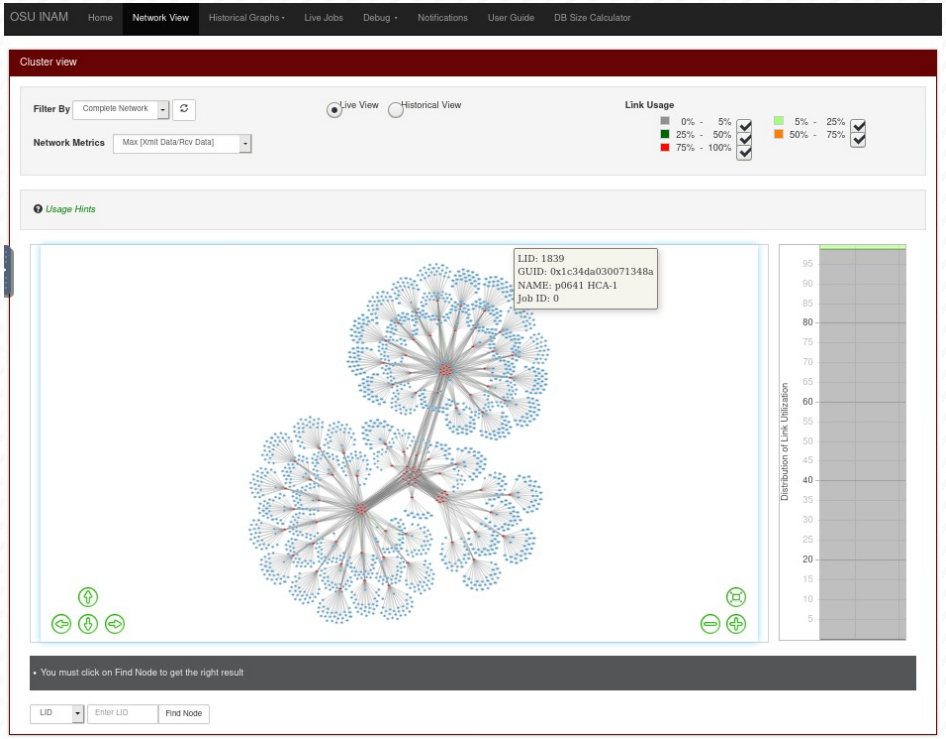
oh-tech.org/blog

facebook.com/ohiosupercomputercenter

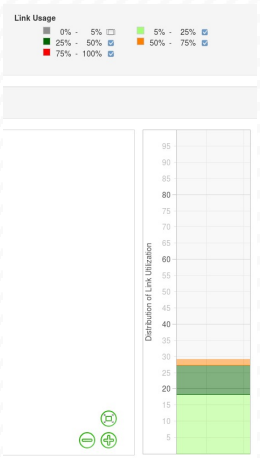
linkedin.com/company/ohio-supercomputer-center

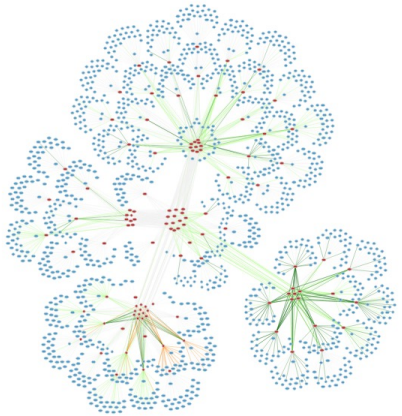
Backup Slides

Network View

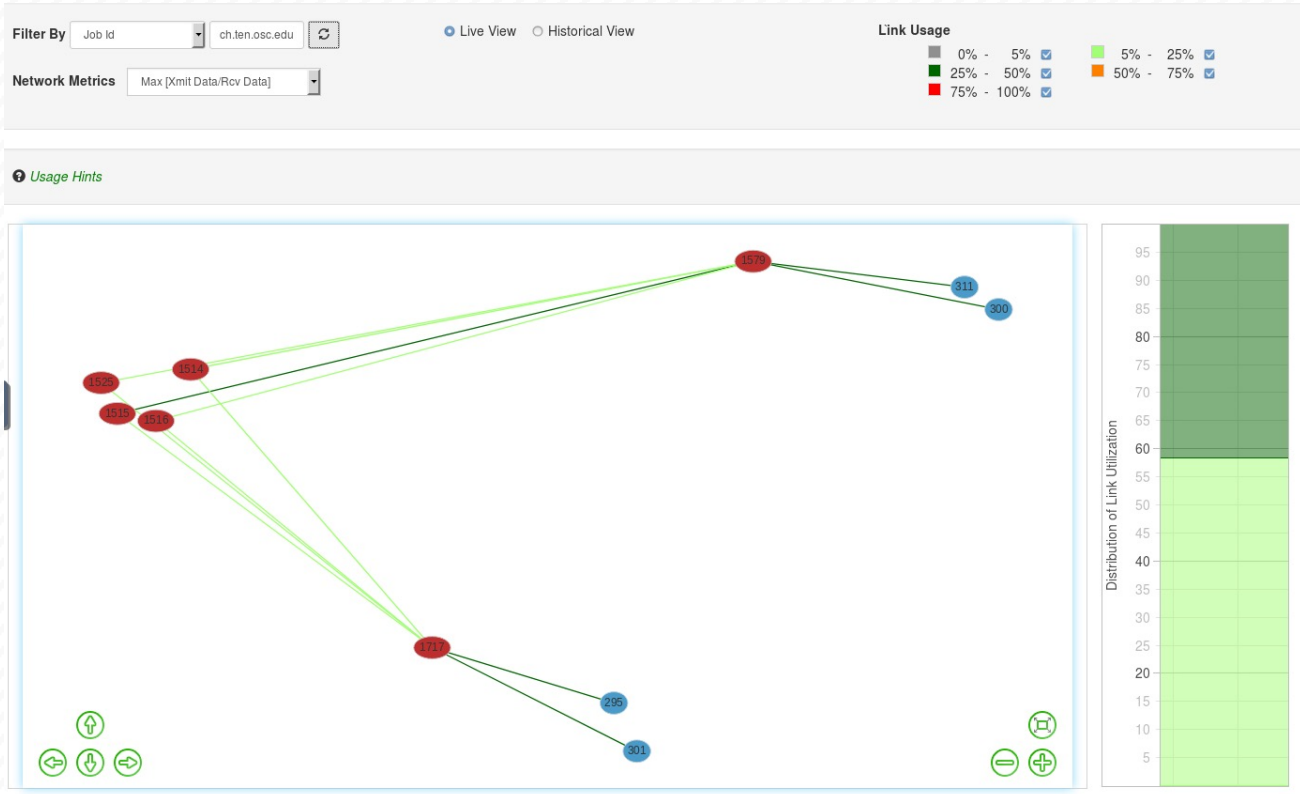


- Link utilization
  - Distribution
  - Link color in network graph
- Hover over node for details

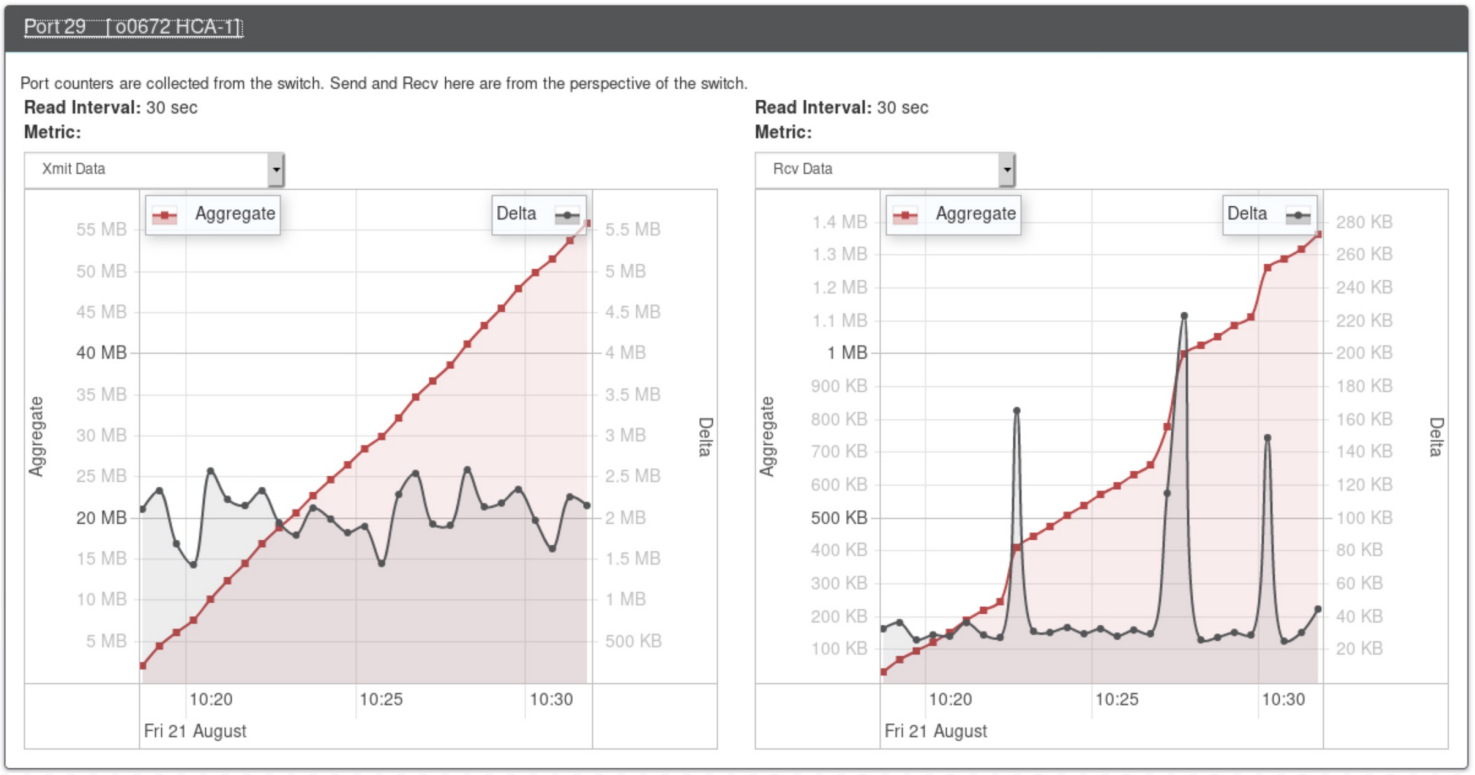




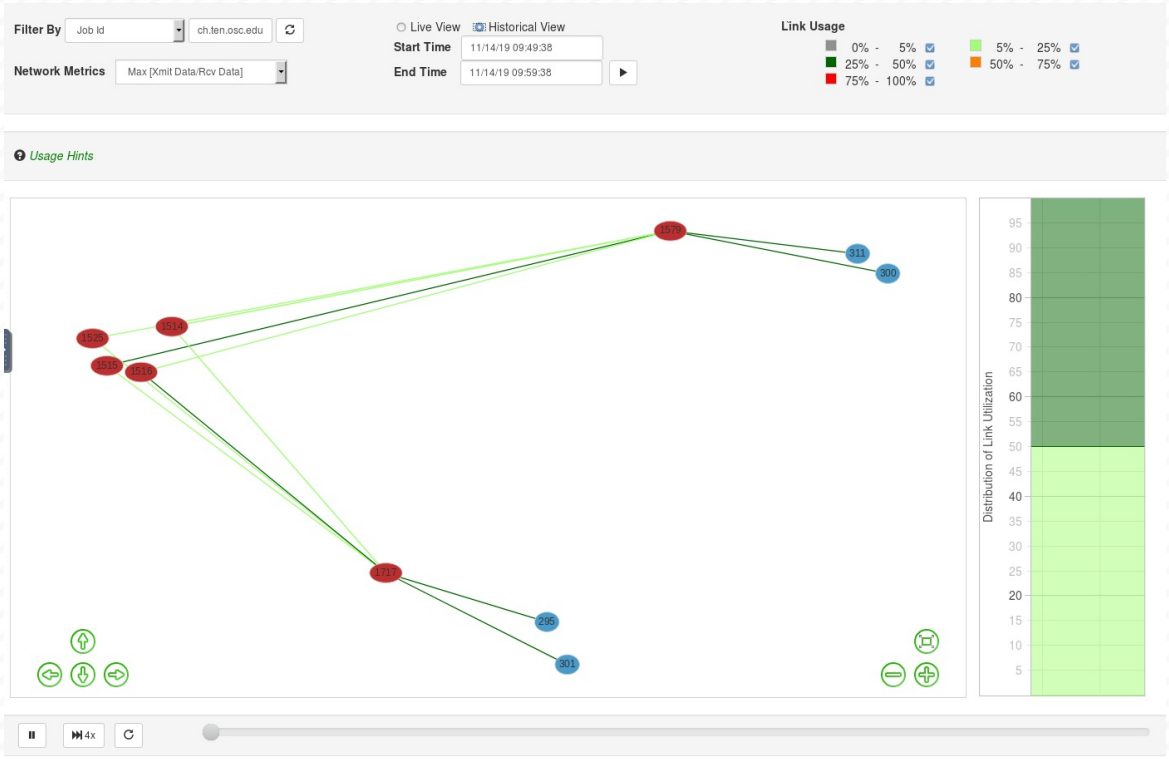
Live View by Job Id



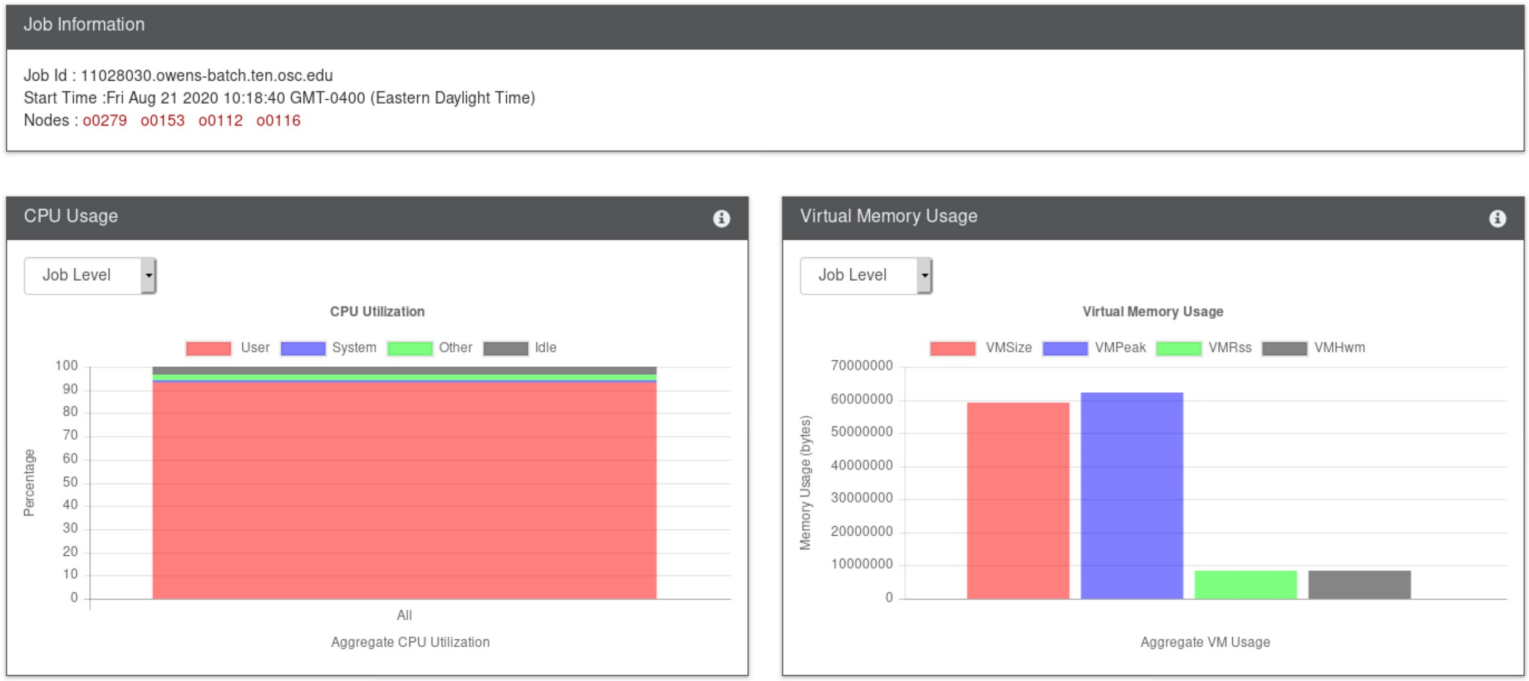
Node Info


Port 29 [0672 HCA-1]
Port counters are collected from the switch. Send and Recv here are from the perspective of the switch.
Read Interval: 30 sec
Metric:
Xmit Data
Aggregate
Delta
55 MB
50 MB
45 MB
40 MB
35 MB
30 MB
25 MB
20 MB
15 MB
10 MB
5 MB
5.5 MB
5 MB
4.5 MB
4 MB
3.5 MB
3 MB
2.5 MB
2 MB
1.5 MB
1 MB
500 KB
10:20
10:25
10:30
Fri 21 August
Read Interval: 30 sec
Metric:
Rcv Data
Aggregate
Delta
1.4 MB
1.3 MB
1.2 MB
1.1 MB
1 MB
900 KB
800 KB
700 KB
600 KB
500 KB
400 KB
300 KB
200 KB
100 KB
280 KB
260 KB
240 KB
220 KB
200 KB
180 KB
160 KB
140 KB
120 KB
100 KB
80 KB
60 KB
40 KB
20 KB
10:20
10:25
10:30
Fri 21 August

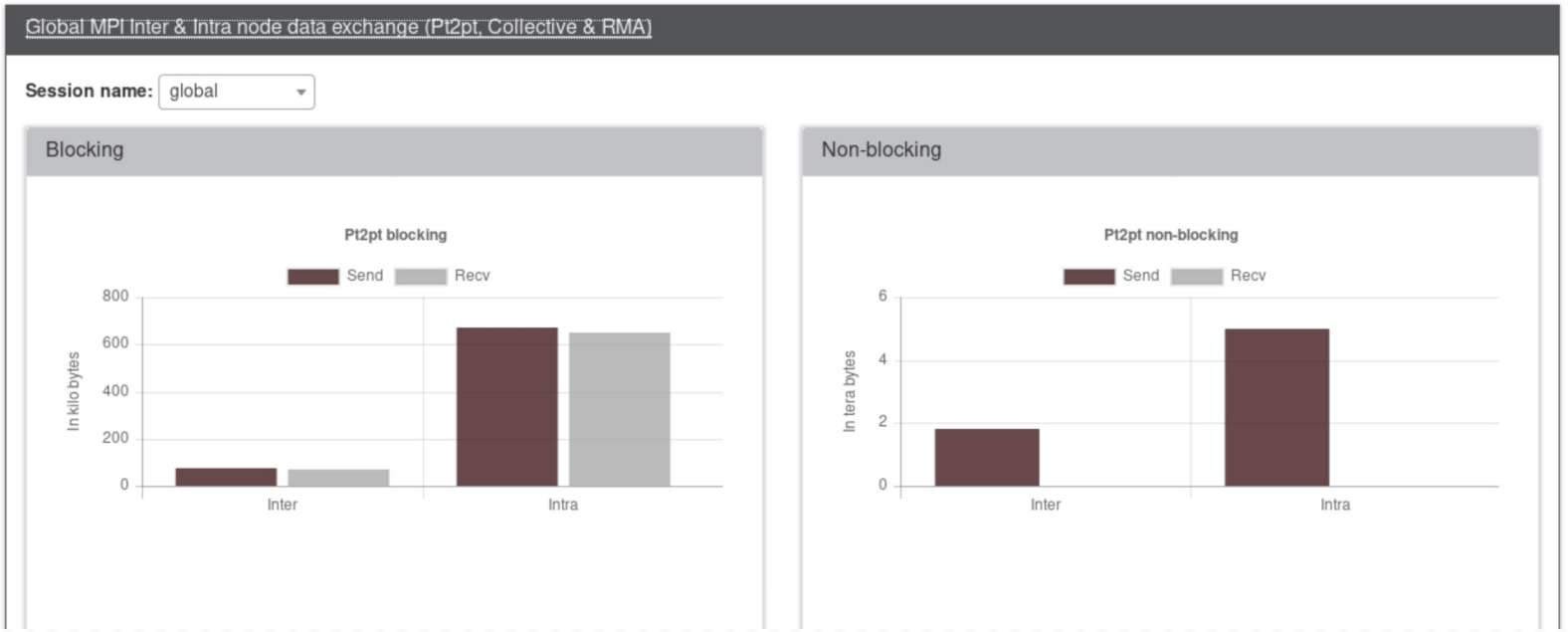
Historical View by Job Id



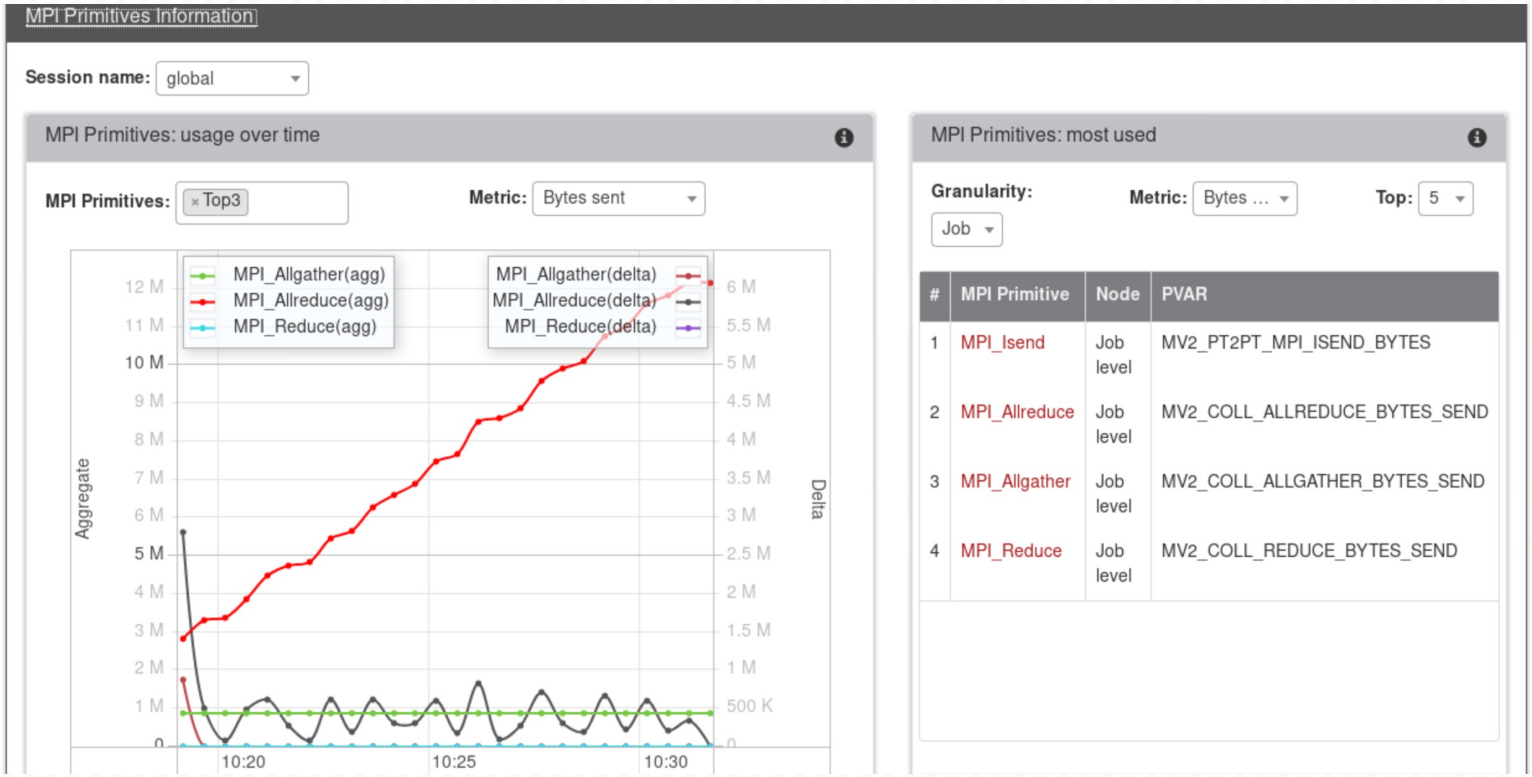
Features with mvapich2-x



Features with mvapich2-x



Features with mvapich2-x



#	MPI Primitive	Node	PVAR
1	MPI_Isend	Job level	MV2_PT2PT_MPI_ISEND_BYTES
2	MPI_Allreduce	Job level	MV2_COLL_ALLREDUCE_BYTES_SEND
3	MPI_Allgather	Job level	MV2_COLL_ALLGATHER_BYTES_SEND
4	MPI_Reduce	Job level	MV2_COLL_REDUCE_BYTES_SEND

Features with mvapich2-x

MPI_Allreduce

MPI_Allreduce - different Algorithms in MVAPICH

Rank	MPI Primitive	Node	PVAR	Value
1	MPI_Allreduce	Job level	MV2_COLL_ALLREDUCE_PT2PT_RD_BYTES_SEND	12.702M

Average time for nodes across msg size (in micro seconds)

Node	1B-512B	513B-2KB	2KB-8KB	8KB-64KB	64KB-1MB	>1MB
o0116 HCA-1	442.95us	0.00us	0.00us	0.00us	0.00us	0.00us
o0279 HCA-1	470.05us	0.00us	0.00us	0.00us	0.00us	0.00us
o0153 HCA-1	465.00us	0.00us	0.00us	0.00us	0.00us	0.00us
o0112 HCA-1	466.64us	0.00us	0.00us	0.00us	0.00us	0.00us

Legend:

K - Kilo (10^3) M - Mega (10^6) G - Giga (10^9) T - Tera (10^{12}) P - Peta (10^{15})