Ohio Supercomputer Center

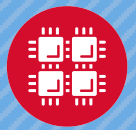
An OH·TECH Consortium Member

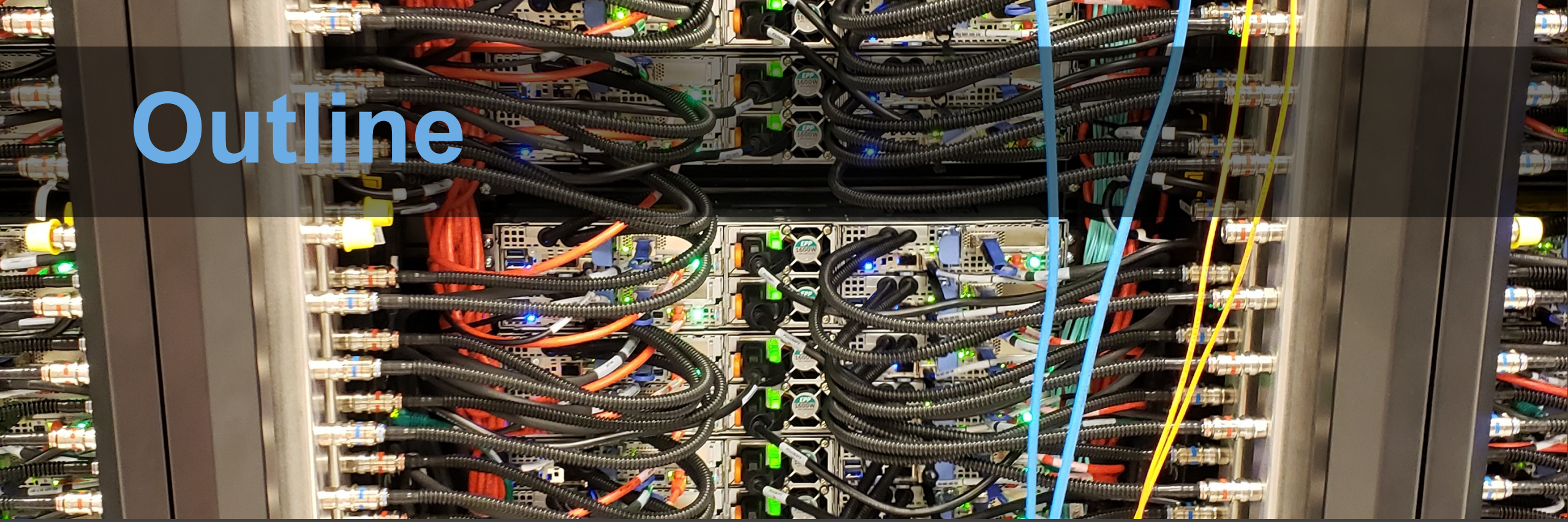
OSU INAM at Ohio Supercomputer Center

Karen Tomko, ktomko@osc.edu

Heechang Na, hna@osc.edu

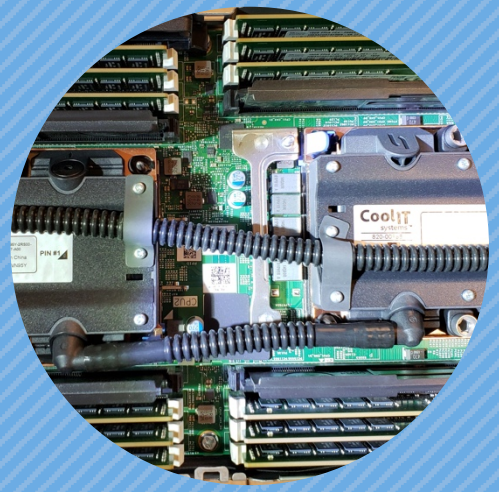
Trey Dockendorf, tdockendorf@osc.edu





Outline

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



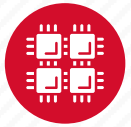
Overview of OSC's Systems and Fabric

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

System Status (Aug 2020)

COMPUTE	Ruby	Owens	Pitzer	Pitzer Expansion
Date	2014	2016	2018	2020
Cost	$1.5 million	$7 million	$3.35 million	$3.8 million
Theoretical Perf.	~144 TF	~1.6PF	~1.3PF	~2.6 PF
Nodes	240	824	260	398
CPU Cores	4800	23,392	10,560	19,104
RAM	~15.3 TB	~120 TB	~ 70.6 TB	~ 93.7 TB
GPUs	20 NVIDIA K40	160 NVIDIA P100	64 NVIDIA V100	102 NVIDIA V100

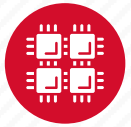
STORAGE	NetApp	DDN	IBM	Tape Library
Capacity	0.8 PB	4.8 PB	8.6 PB	10+ PB



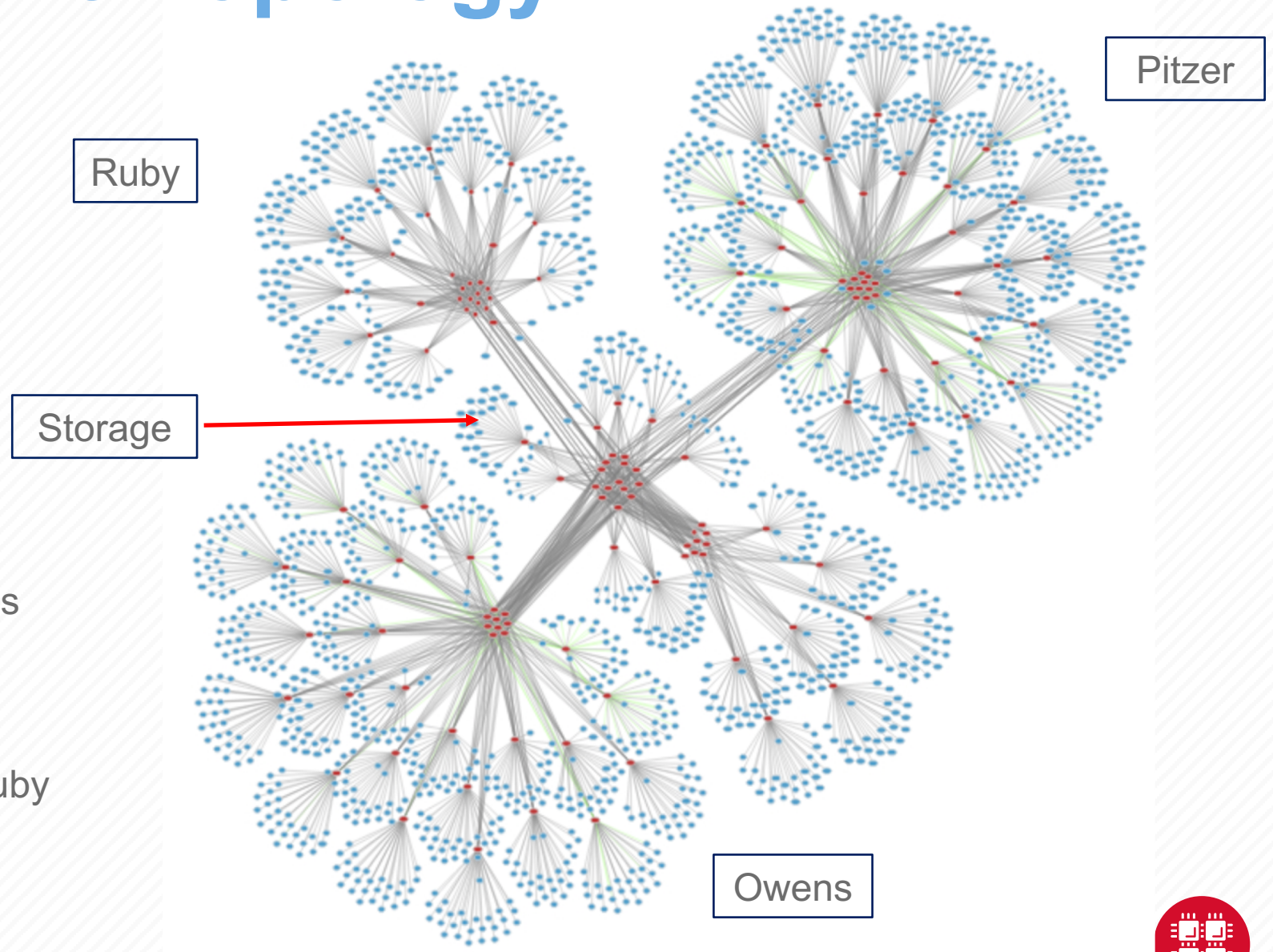
Not your lab's fabric

OSC has a single integrated IB fabric

- Fabric Size: 139 IB switches, 1722 compute nodes,
- Currently 3 compute clusters, 4 generations of hardware
- RDMA access to 2 generations of GPFS filesystems
- Multiple generations of InfiniBand (FDR, 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: 40 switches
- Level 1: 59 switches
- Legacy FDR/FDR10 island (Ruby + infrastructure servers)

INAM at OSC

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

FAMII 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:

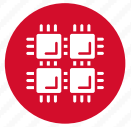
OSU: Pouya Kousha, Nick Sarkauskas, Kamal Sankar, Bharath Ramesh, Mansa Kedia, Aamir Shafi, Hari Subramoni, DK Panda

OSC: Trey Dockendorf, Heechang Na, Karen Tomko

Status:

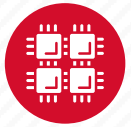
- INAM has been running at OSC on production systems for more than a year
- Iterative test and development cycle between OSC/OSU

Thank you to the National Science Foundation for supporting this project NSF OAC-1664137

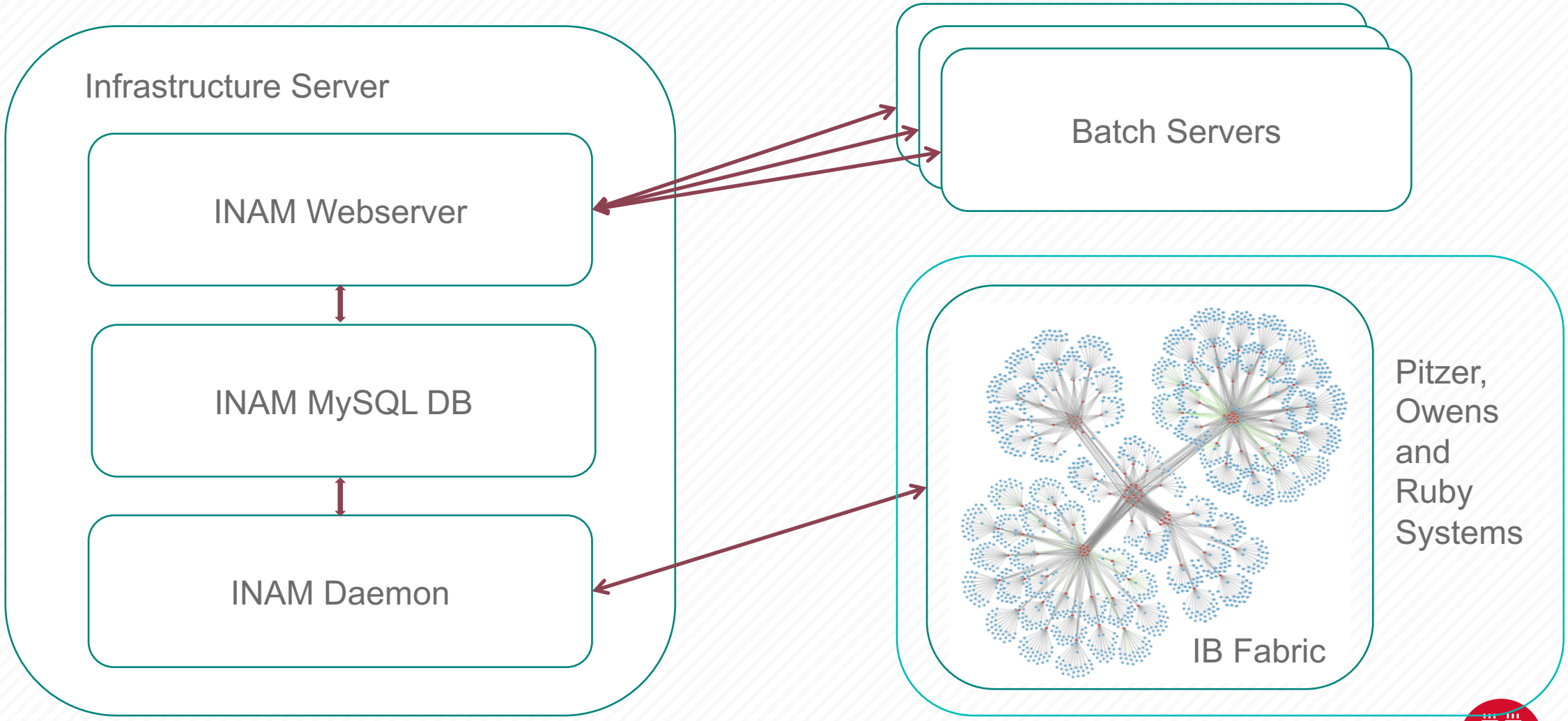


Overview of OSU INAM

- What is INAM
  - OSU InfiniBand Network Analysis and Monitoring (INAM) tool
  - http://mvapich.cse.ohio-state.edu/tools/osu-inam/
- Main features
  - Monitors IB clusters in real time
  - Ability to analyze and profile node-level, job-level and process-level activities for MPI communication
  - Visualize live or historical data transfer metrics
  - Network view or Job view
  - Filter by node, switch, link utilization

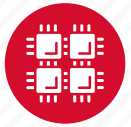


OSC INAM Deployment



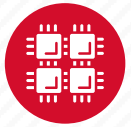
Configuring for OSC

- Integration with the resource manager
  - OSC currently uses two resource managers (migrating to Slurm)
  - INAM configured for Torque/MOAB
    - Separate batch server for each cluster
    - Alpha-numeric job names
- 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 ~56GB of disk space
- MVAPICH2-X integration
  - Config file replicated on filesystem available to compute nodes



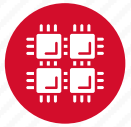
Impact of OSU/OSC collaboration on INAM (1)

- Performance
  - Multi-threaded fabric discovery
    - 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
  - Time for insertion operations reduced 2-4x
  - Improved Fault-tolerance of Database
    - Automatic restart of MySQL service



Impact of OSU/OSC collaboration on INAM (2)

- Installation and Configuration
  - Focus: make it easier to automate deployment of INAM
  - Simplified packaging
    - e.g. Single RPM with all components
  - Additional configuration items
    - e.g. Configurable path for MV2-X config file
- User interface refinements and suggestions
  - Focus: usability
  - 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

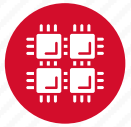


More info:

- http://mvapich.cse.ohio-state.edu/tools/osu-inam/

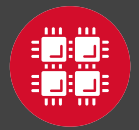
- **Pearc 20 paper:**

  Accelerated Real-time Network Monitoring and Profiling at Scale using OSU INAM, P. Kousha, S. D. Kamal Raj , H. Subramoni, DK Panda, H. Na, T. Dockendorf, and K. Tomko. Practice and Experience in Advanced Research Computing 2020, Jul 2020.



INAM Demo

- Quick overview
- Features with MVAPICH2-X



OH·TECH

Ohio Technology Consortium

A Division of the **Ohio Department of Higher Education**

info@osc.edu

osc.edu

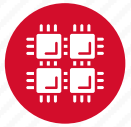
twitter.com/osc

oh-tech.org/blog

facebook.com/ohiosupercomputercenter

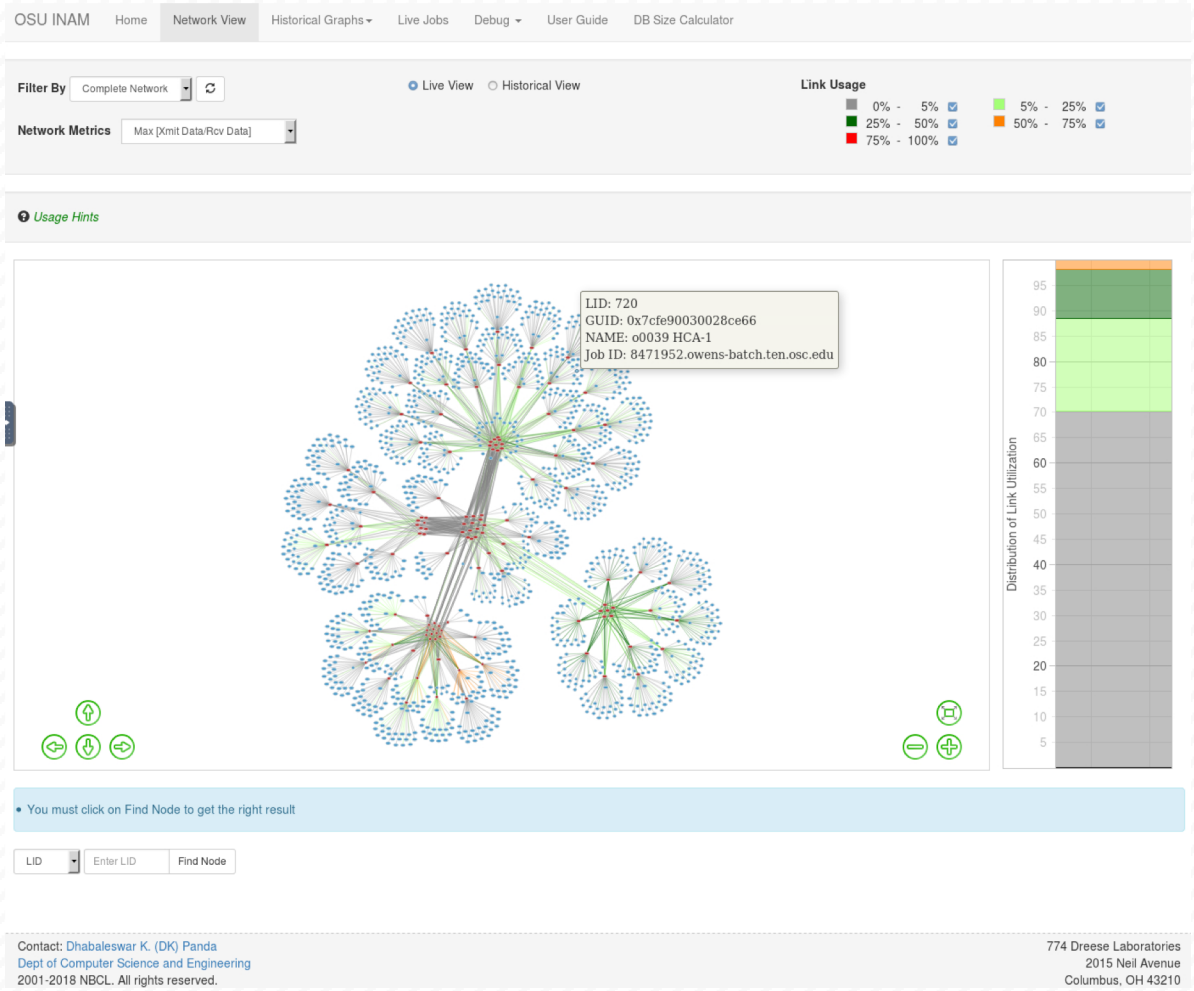
linkedin.com/company/ohio-supercomputer-center

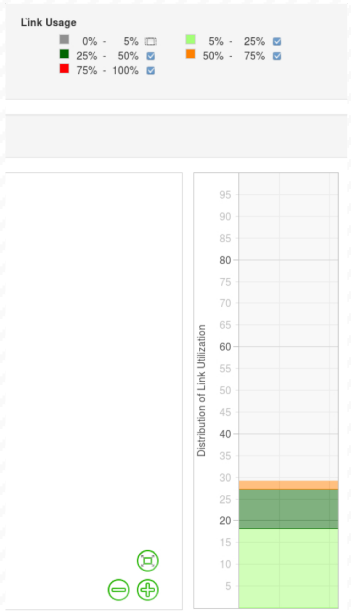
Backup Slides

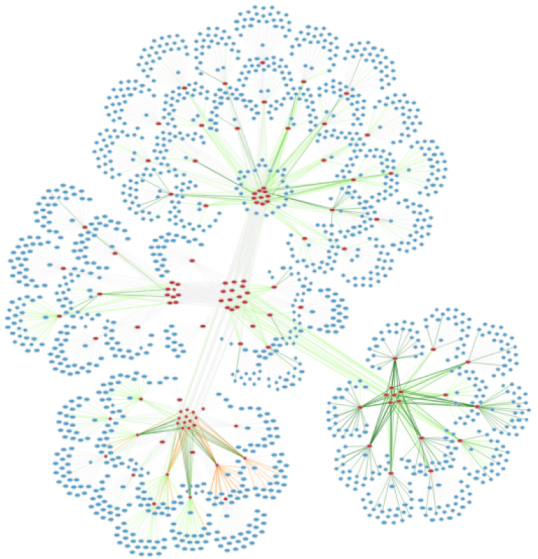


Network View

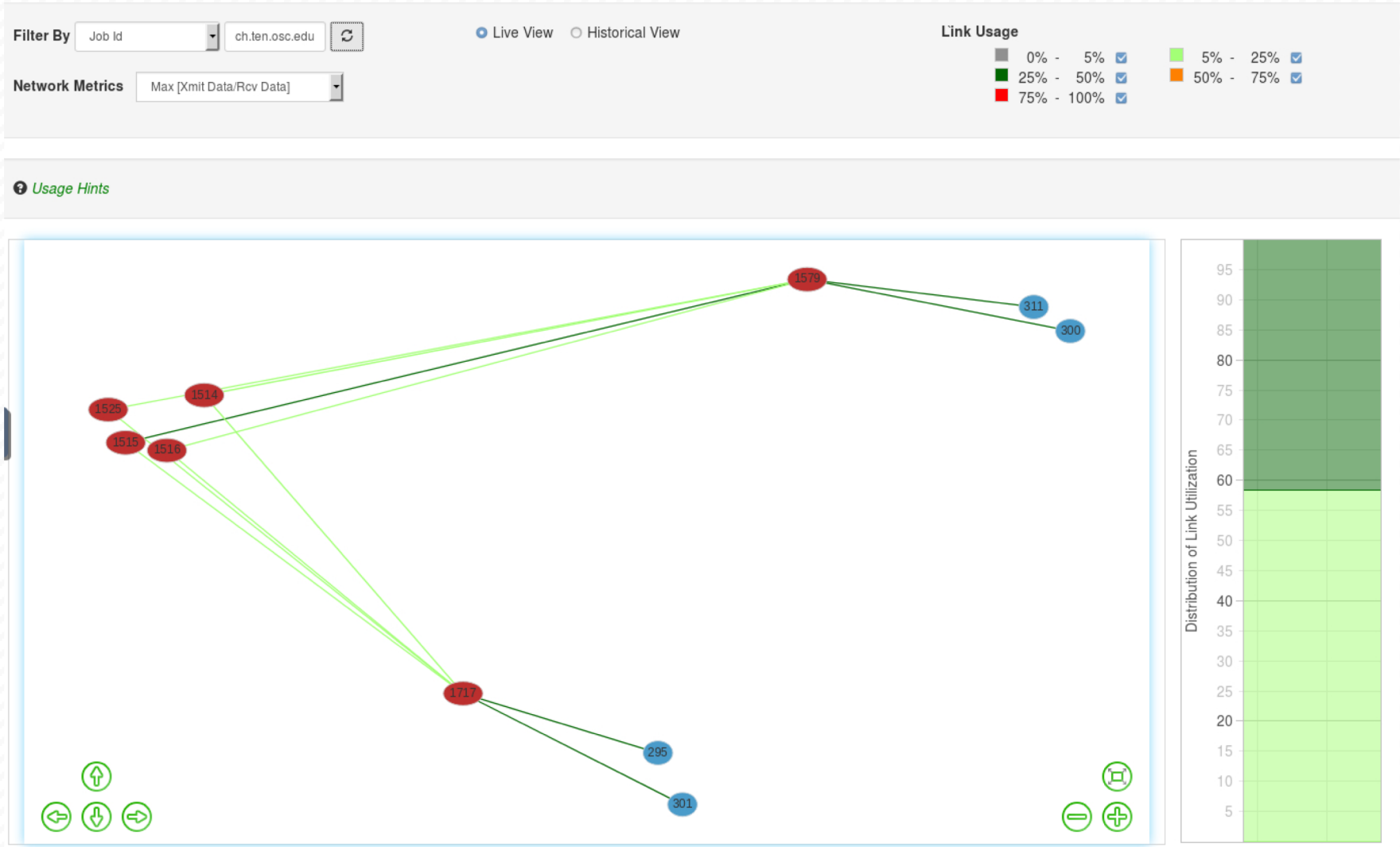
- Link utilization
  - Distribution
  - Link color in network graph
- Hover over node for details



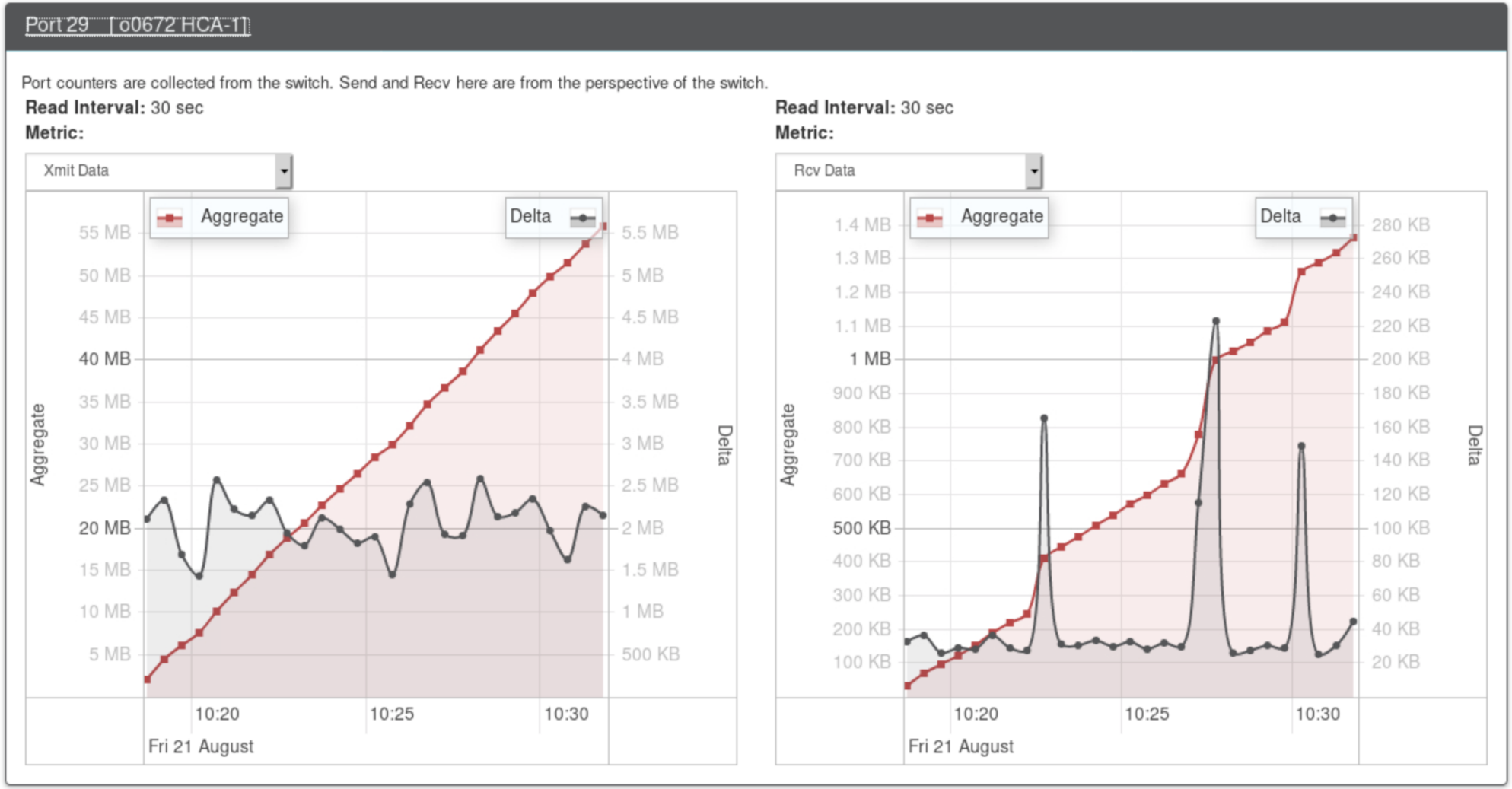




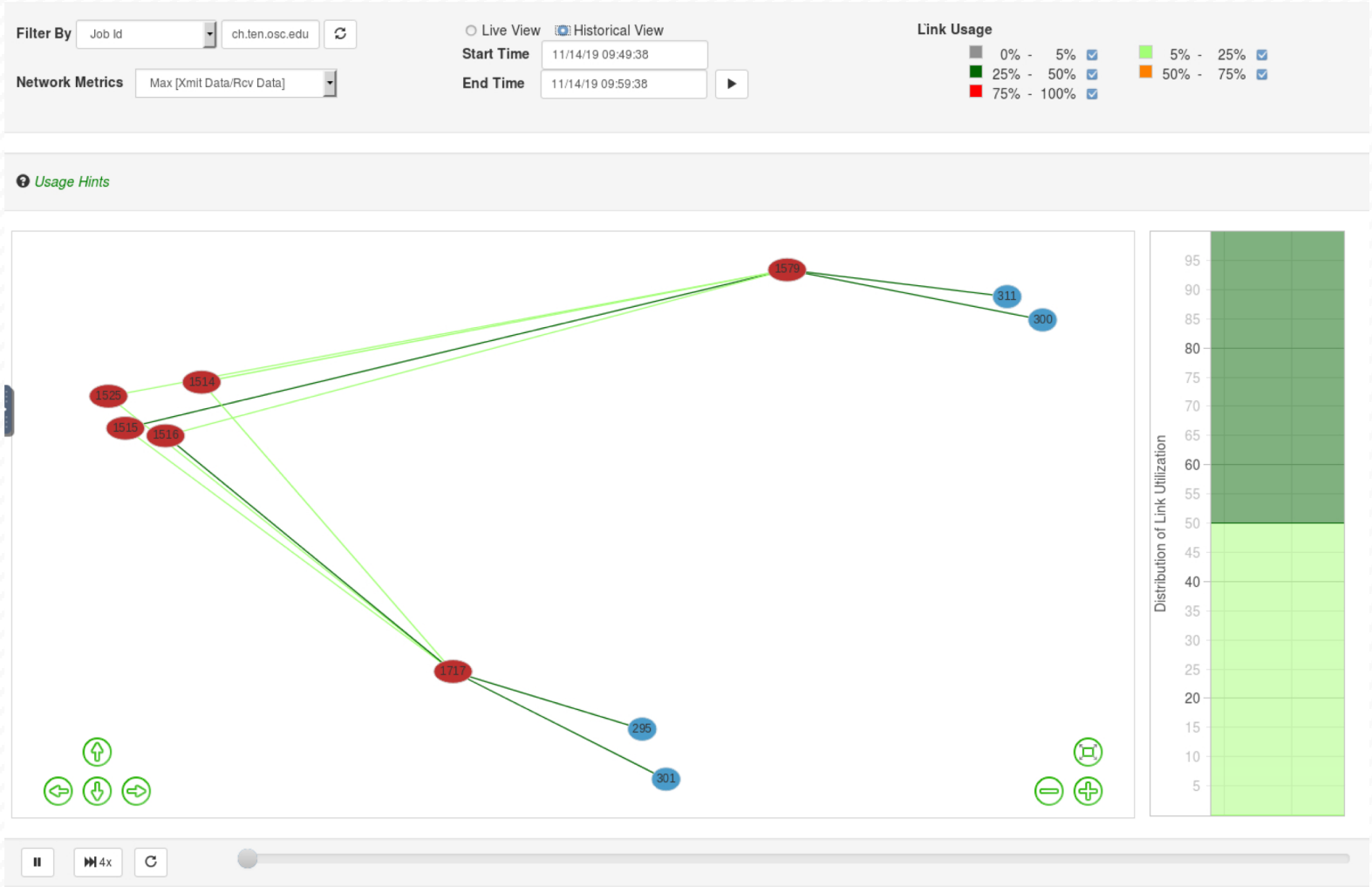
Live View by Job Id



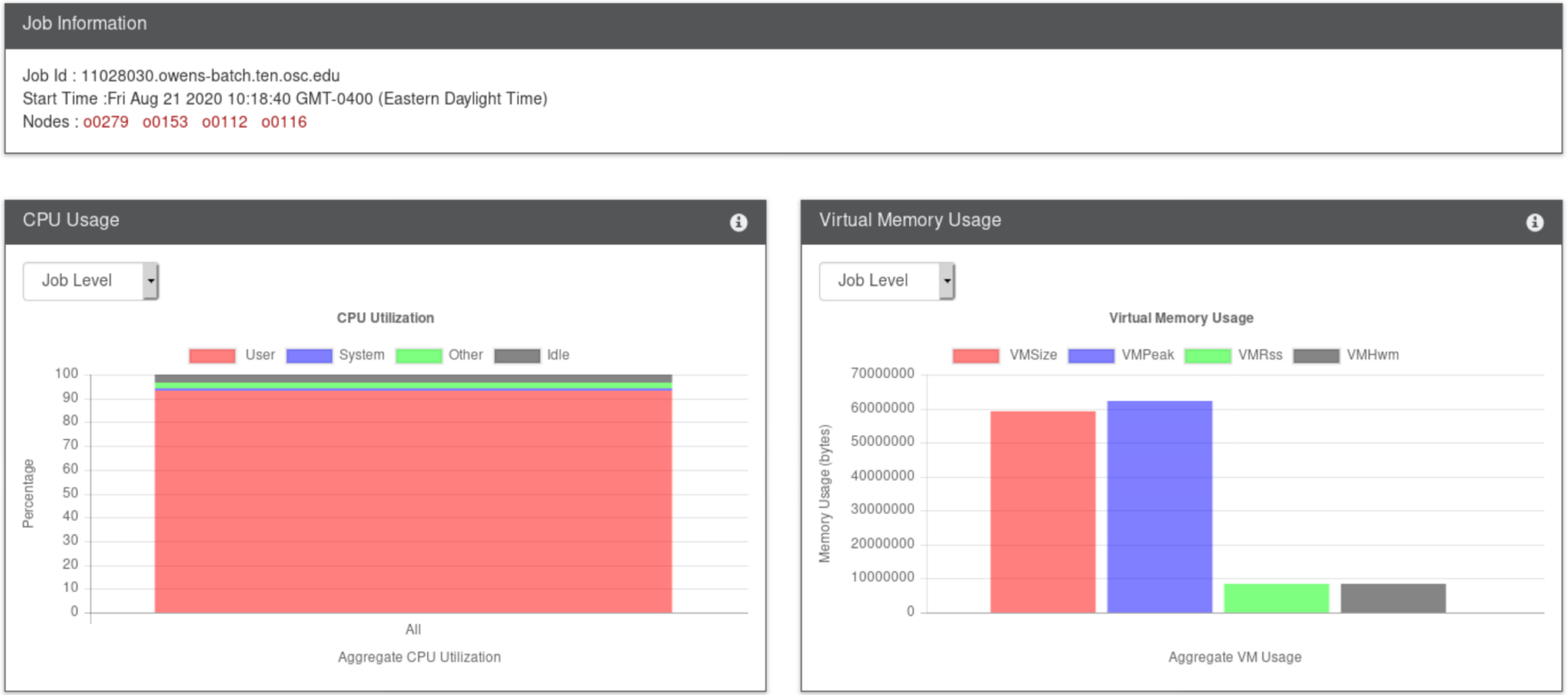
Node Info


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

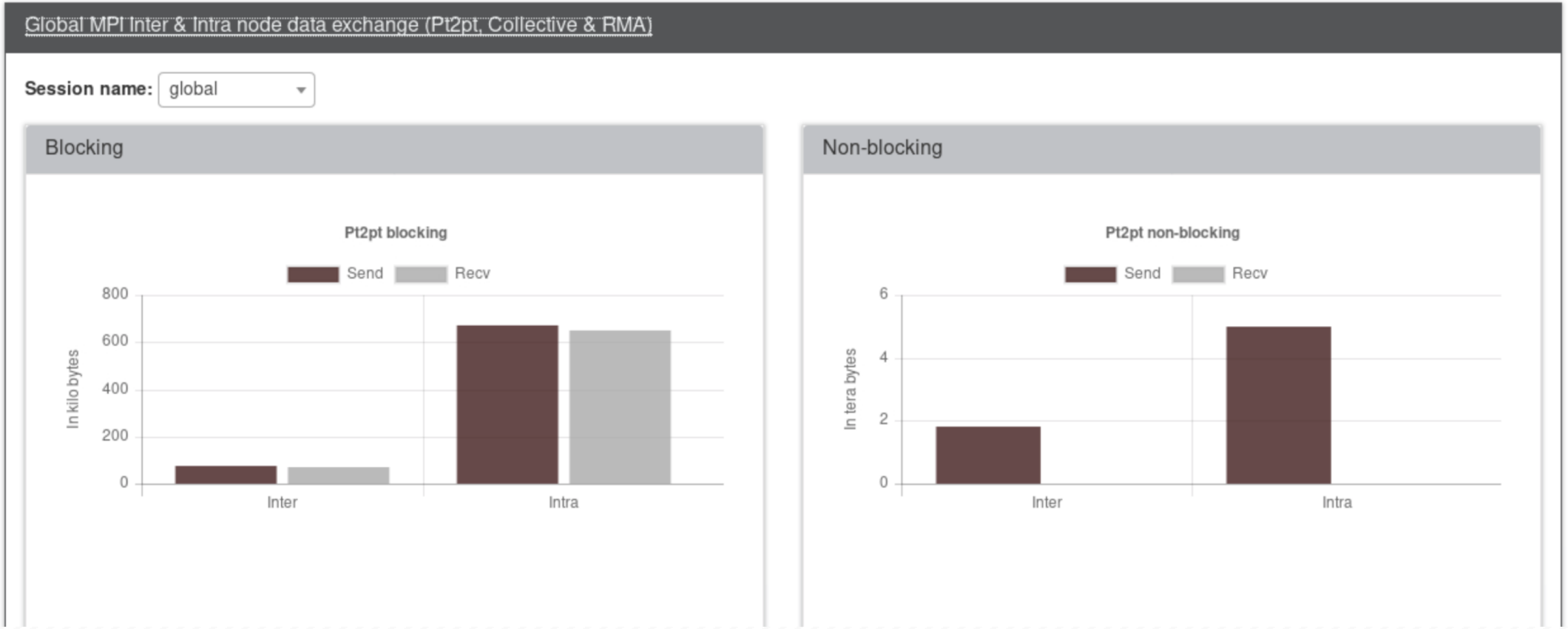
Historical View by Job Id



Features with mvapich2-x



Features with mvapich2-x



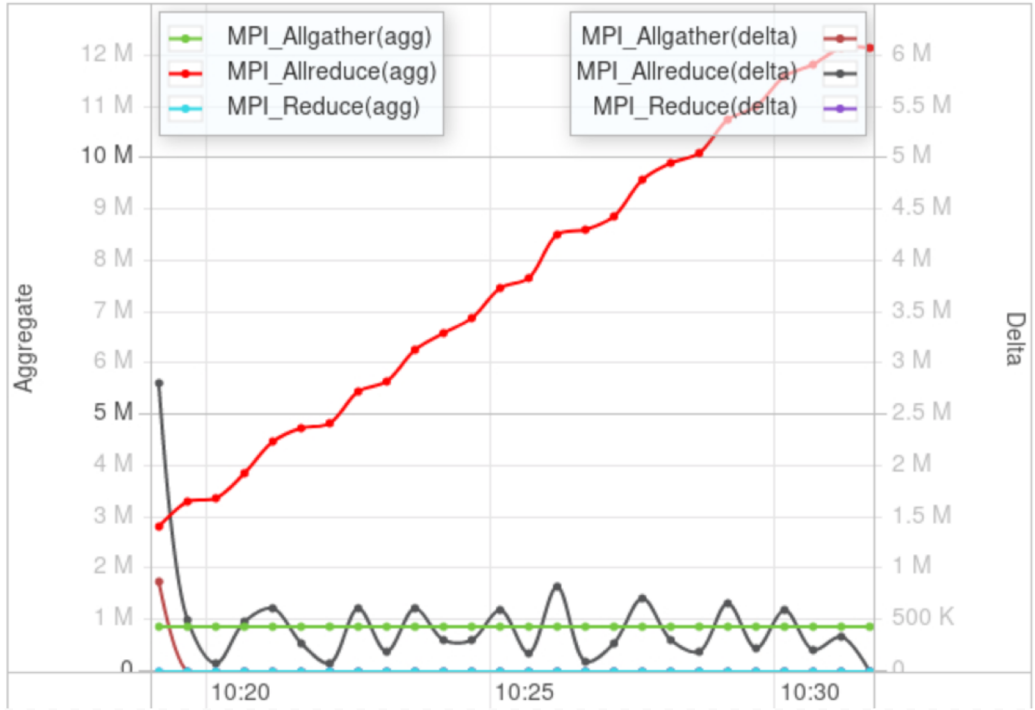
Features with mvapich2-x

MPI Primitives Information

Session name: global

MPI Primitives: usage over time

MPI Primitives: × Top3 **Metric:** Bytes sent



MPI Primitives: most used

Granularity: Job **Metric:** Bytes ... **Top:** 5

#	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})