OSU INAM: A Tool for Analyzing Communication in MPI Runtime and InfiniBand Network

H. Subramoni, A. A. Mathews, M. Arnold, J. Perkins,
X. Lu, S. Chakraborty, K. Hamidouche, and D. K. Panda

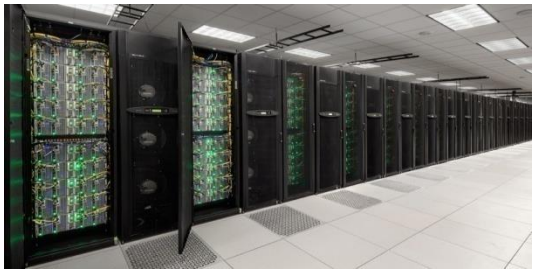
Department of Computer Science and Engineering
The Ohio State University

Outline

- Introduction
- Motivation
- Challenges & Contributions
- Features of OSU INAM & Demo
- Conclusions

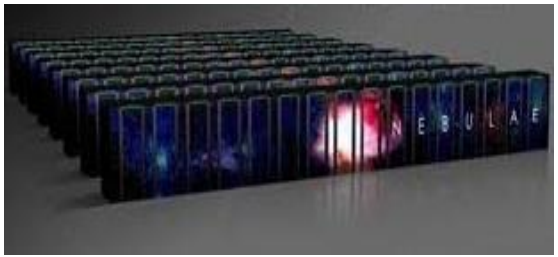
Current Trends in HPC

- Supercomputing systems scaling rapidly
  - Multi-core architectures and
  - High-performance interconnects
- InfiniBand is a popular HPC interconnect
  - 257 systems (51.4%) in Jun'15 Top500



Stampede@TACC

SuperMUC@LRZ



Nebulae@NSCS

OpenSM

- InfiniBand Subnet Manager (IBA Specifications)
- Part of OFED software package
  - Open Fabrics Enterprise Distribution
  - Open source software for RDMA and kernel bypass applications
  - Needed by the HPC community for applications which need low latency and high efficiency and fast I/O
- Scans, Initiates and Monitors the InfiniBand Fabric
- Performance Counters and Subnet Management Attributes
  - Not supported at VL granularity
- Subnet Manager (SM), Subnet Management Agent (SMA)
- At least one instance required per Subnet
- Usage of Virtual Lanes

Message Passing Interface

- Message Passing Interface (MPI) used by vast majority of HPC applications
- MPI 3.1 was approved on June 4, 2015
  - Specification is available from: http://mpi-forum.org/docs/mpi-3.1/mpi31-report.pdf
- MPI provides different communication primitives
  - Two-sided Point-to-point
  - One-sided (Remote Memory Access) Point-to-point
  - Collective (Blocking and Non-blocking)
- MPI_T based support for analyzing and understanding the MPI runtime

MPI Tools Interface

- **Introduced in MPI 3.0 standard to expose internals of MPI to tools and applications**
- **Generalized interface – no defined variables in the standard**
- **Variables can differ between**
  - MPI implementations
  - Compilations of same MPI library (production vs debug)
  - Executions of the same application/MPI library
  - There could be no variables provided
- **Two types of variables supported**
  - **Control Variables (CVARS)**
    - Typically used to configure and tune MPI internals
    - Environment variables, configuration parameters and toggles
  - **Performance Variables (PVARS)**
    - Insights into performance of an MPI library
    - Highly-implementation specific
    - Memory consumption, timing information, resource-usage, data transmission info.
    - Per-call basis or an entire MPI job
- **More about the interface: http://mpi-forum.org/docs/mpi-3.1/mpi31-report.pdf**

Outline

- Introduction
- Motivation
- Challenges & Contributions
- Features of OSU INAM & Demo
- Conclusions

Existing Monitoring Tools

- Nagios [Agent Based]
  - + Easy to Integrate & Configure
  - + Supports multiple interconnects
  - - No discovery process
  - - Involves more overhead
  - - No Layer 2, Switch Dependent
  - - Cannot classify traffic based on MPI primitives
- Ganglia [Agent Based]
  - + Portable and Scalable
  - + Distributed Modules provide higher sampling rates
  - + Supports multiple interconnects
  - - Use of Daemons (gmond) involves more overhead
  - - Metric measurements in compiled code
  - - Adding custom metrics can be a bit complicated
  - - Cannot classify traffic based on MPI primitives
- Fabric IT [Agent Less]
  - + Good Sampling Rates
  - + Agent less
  - + Integrated into the Subnet Manager
  - - Proprietary by Mellanox, Specific for IB
  - - Does not show communication patterns
  - - Does not show Link usage pertaining to a Job
  - - No long term data storage
  - - Cannot classify traffic based on MPI primitives

MVAPICH2 Software

- **High Performance open-source MPI Library for InfiniBand, 10Gig/iWARP, and RoCE**
  - MVAPICH (MPI-1) , Available since 2002
  - MVAPICH2 (MPI-2.2, MPI-3.0 and MPI-3.1), Available since 2004
  - MVAPICH2-X (Advanced MPI + PGAS), Available since 2012
  - Support for GPGPUs (MVAPICH2-GDR), Available since 2014
  - Support for MIC (MVAPICH2-MIC), Available since 2014
  - Support for Virtualization (MVAPICH2-Virt), Available since 2015
  - Support for Energy-Awareness (MVAPICH2-EA), Available since 2015
  - Used by more than 2,475 organizations in 76 countries
  - More than 298,000 downloads from the OSU site directly
  - Empowering many TOP500 clusters (Jun'15 ranking)
    - 8th ranked 519,640-core cluster (Stampede) at TACC
    - 11th ranked 185,344-core cluster (Pleiades) at NASA
    - 22nd ranked 76,032-core cluster (Tsubame 2.5) at Tokyo Institute of Technology and many others
  - Available with software stacks of many IB, HSE, and server vendors including Linux Distros (RedHat and SuSE)
  - http://mvapich.cse.ohio-state.edu
- **Empowering Top500 systems for over a decade**
  - System-X from Virginia Tech (3rd in Nov 2003, 2,200 processors, 12.25 TFlops) ->
  - Stampede at TACC (8th in Jun'15, 462,462 cores, 5.168 Plops)

MPI-T Support in MVAPICH2

- Initial focus on performance variables
- Variables to track different components
  - MPI library's internal memory usage
  - Unexpected receive queue
  - Registration cache
  - VBUF allocation
  - Shared-memory communication
  - Collective communication algorithms
  - IB channel packet transmission
  - Many more in progress..

Outline

- Introduction
- Motivation
- Challenges & Contributions
- Features of OSU INAM & Demo
- Conclusions

Broad Challenge

How can we design a tool that can analyze the communication traffic on the InfiniBand network with inputs from the MPI runtime

Contributions

- Design and develop OSU INAM
  - 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/
  - http://mvapich.cse.ohio-state.edu/userguide/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)
- Remotely monitor CPU utilization of MPI processes at user specified granularity
- Visualize the data transfer happening in a “live” or “historical” fashion for entire network, job or set of nodes

Outline

- Introduction
- Motivation
- Challenges & Contributions
- Features of OSU INAM & Demo
- Conclusions

Features of OSU INAM

- Analyze and profile network-level activities with many parameters (data and errors) at user specified granularity
- Capability to analyze and profile node-level, job-level and process-level activities for MPI communication (Point-to-Point, Collectives and RMA)
- Remotely monitor CPU utilization of MPI processes at user specified granularity
- Visualize the data transfer happening in a "live" fashion for
  - Entire Network - Live Network Level View
  - Particular Job - Live Job Level View
  - One or multiple Nodes - Live Node Level View
  - One or multiple Switches - Live Switch Level View
- Visualize data transfer that happened in the network for a time in the past for
  - Entire Network - Historical Network Level View
  - Particular Job - Historical Job Level View
  - One or multiple Nodes - Historical Node Level View

Switch Counters Supported by OSU INAM

- Xmit Data
  - Total number of data octets, divided by 4, transmitted on all VLs from the port
  - This includes all octets between (and not including) the start of packet delimiter and the VCRC, and may include packets containing errors
  - Excludes all link packets.
- Rcv Data
  - Total number of data octets, divided by 4, received on all VLs from the port
  - This includes all octets between (and not including) the start of packet delimiter and the VCRC, and may include packets containing errors
  - Excludes all link packets.
- Max [Xmit Data/Rcv Data]
  - Maximum of the two values above

Process Level Counters Supported by OSU INAM

- Xmit Data
  - Total number of bytes transmitted as part of the MPI application
- Rcv Data
  - Total number of bytes received as part of the MPI application
- Max [Xmit Data/Rcv Data]
  - Maximum of the two values above
- Point to Point Send
  - Total number of bytes transmitted as part of MPI point-to-point operations
- Point to Point Rcvd
  - Total number of bytes received as part of MPI point-to-point operations
- Max [Point to Point Sent/Rcvd]
  - Maximum of the two values above
- Coll Bytes Sent
  - Total number of bytes transmitted as part of MPI collective operations
- Coll Bytes Rcvd
  - Total number of bytes received as part of MPI collective operations
- Max [Coll Bytes Sent/Rcvd]
  - Maximum of the two values above
- RMA Bytes Sent
  - Total number of bytes transmitted as part of MPI RMA operations. Note that due to the nature of the RMA operations, bytes received for RMA operations cannot be counted

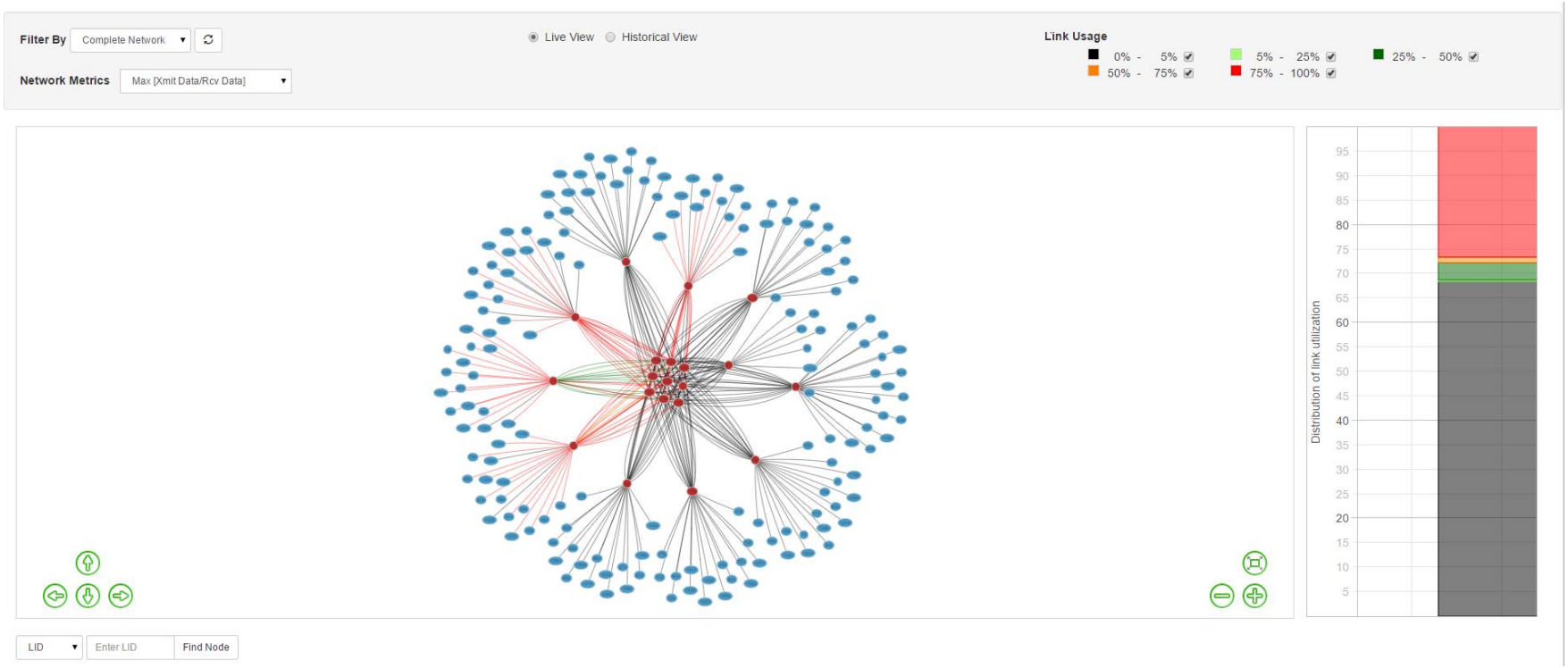
Error Counters Supported by OSU INAM

- SymbolErrors
  - Total number of minor link errors detected on one or more physical lanes
- LinkRecovers
  - Total number of times the Port Training state machine has successfully completed the link error recovery process
- LinkDowned
  - Total number of times the Port Training state machine has failed the link error recovery process and downed the link
- RcvErrors
  - Total number of packets containing an error that were received on the port. These errors include:
    - Local physical errors
    - Malformed data packet errors
    - Malformed link packet errors
    - Packets discarded due to buffer overrun
- RcvRemotePhysErrors
  - Total number of packets marked with the EBP delimiter received on the port.
- RcvSwitchRelayErrors
  - Total number of packets received on the port that were discarded because they could not be forwarded by the switch relay

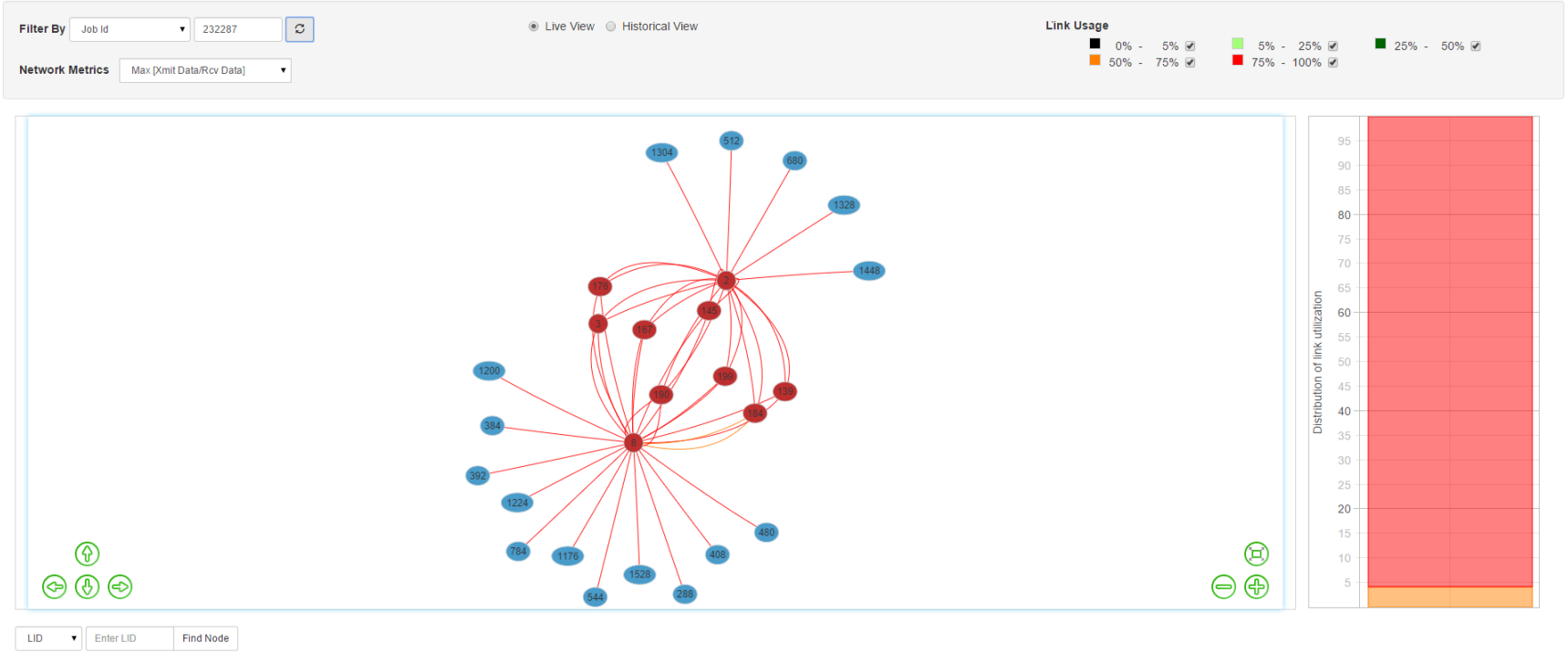
Error Counters Supported by OSU INAM (Cont)

- XmtDiscards
  - Total number of outbound packets discarded by the port because the port is down or congested. Reasons for this include:
    - Output port is not in the active state
    - Packet length exceeded NeighborMTU
    - Switch Lifetime Limit exceeded
    - Switch HOQ Lifetime Limit exceeded This may also include packets discarded while in VLStalled State.
- XmtConstraintErrors
  - Total number of packets not transmitted from the switch physical port for the following reasons:
    - FilterRawOutbound is true and packet is raw
    - PartitionEnforcementOutbound is true and packet fails partition key check or IP version check
- RcvConstraintErrors
  - Total number of packets not received from the switch physical port for the following reasons:
    - FilterRawInbound is true and packet is raw
    - PartitionEnforcementInbound is true and packet fails partition key check or IP version check
- LinkIntegrityErrors
  - The number of time s that the count of local physical errors exceeded the threshold specified by LocalPhyErrors
- ExcBufOverrunErrors
  - The number of times that OverrunErrors consecutive flow control update periods occurred, each having at least one overrun error
- VL15Dropped
  - Number of incoming VL15 packets dropped due to resource limitations (e.g., lack of buffers) in the port

Live Network Level View



Live Job Level View



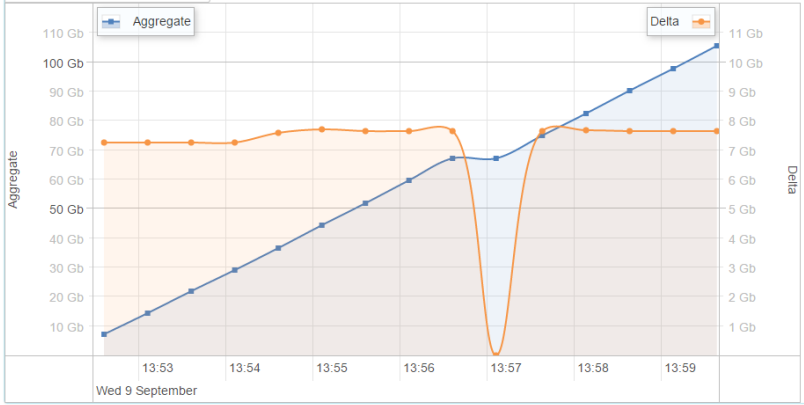
Live Node Level View

Rank112 [core 0]

Rank113 [core 1]

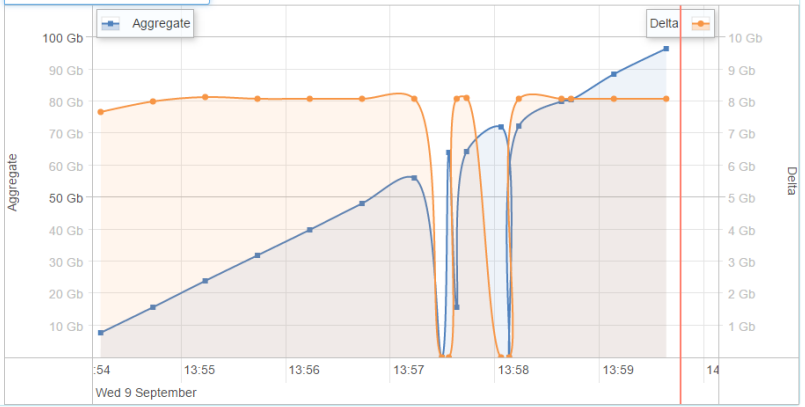
Data Counter:

Xmit Data



Data Counter:

Coll Bytes Sent



Rank114 [core 2]

Rank115 [core 3]

Rank116 [core 4]

Rank117 [core 5]

Rank118 [core 6]

Rank119 [core 7]

Live Node Level View (Cont.)

Node Information

Node Details

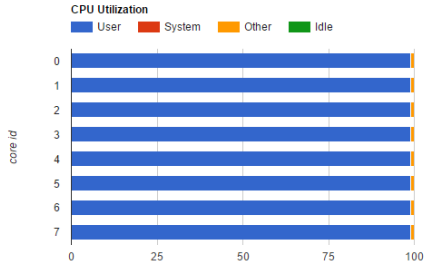
NAME : **node158 HCA-1**
LID : **384**
GUID: **0x0002c903000a9119**

Job Information

Job Id : 232287
Start Time :Wed Sep 09 2015 13:56:37 GMT-0400 (Eastern Daylight Time)
Nodes : node001 node002 node003 node004 node005 node019 node020 node151 node152 node153 node154 node155 node156 node157 node158 node159

CPU Usage

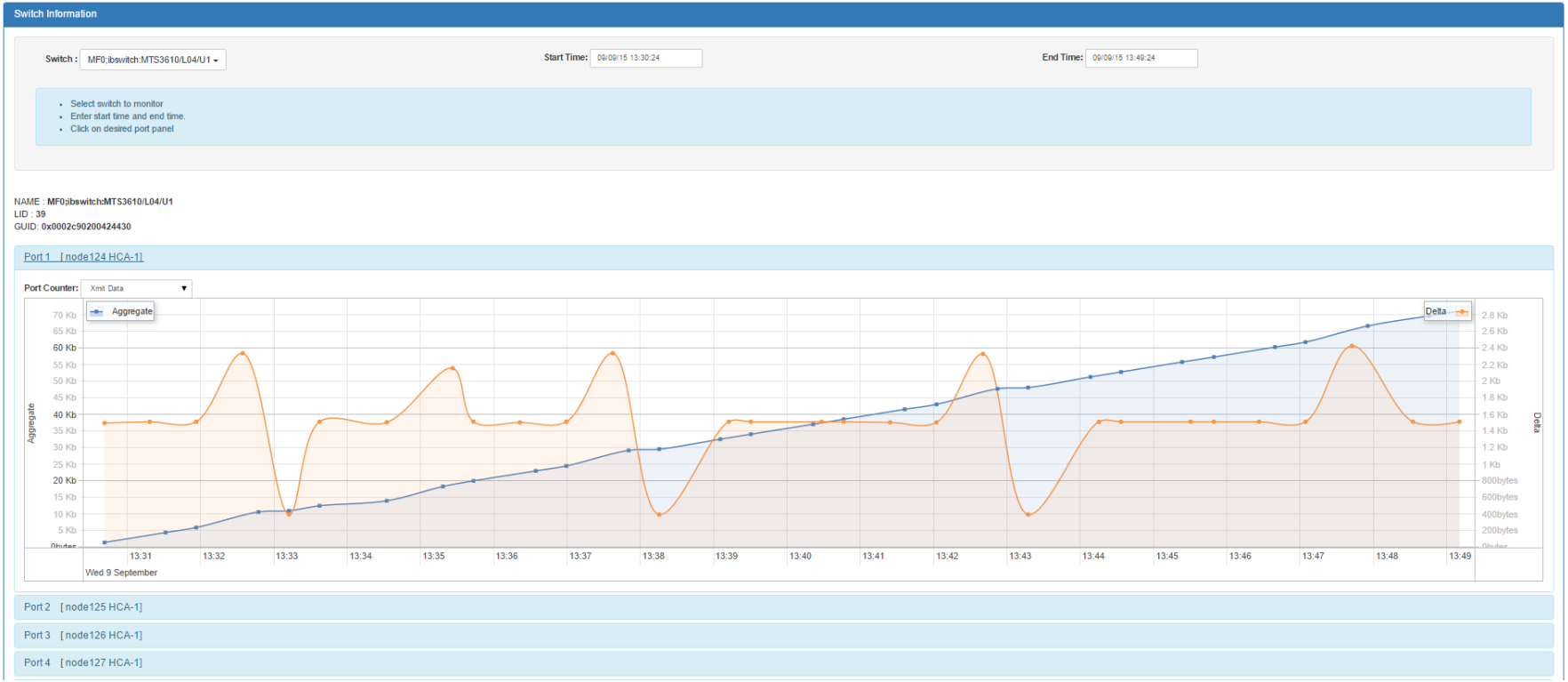
Core Level



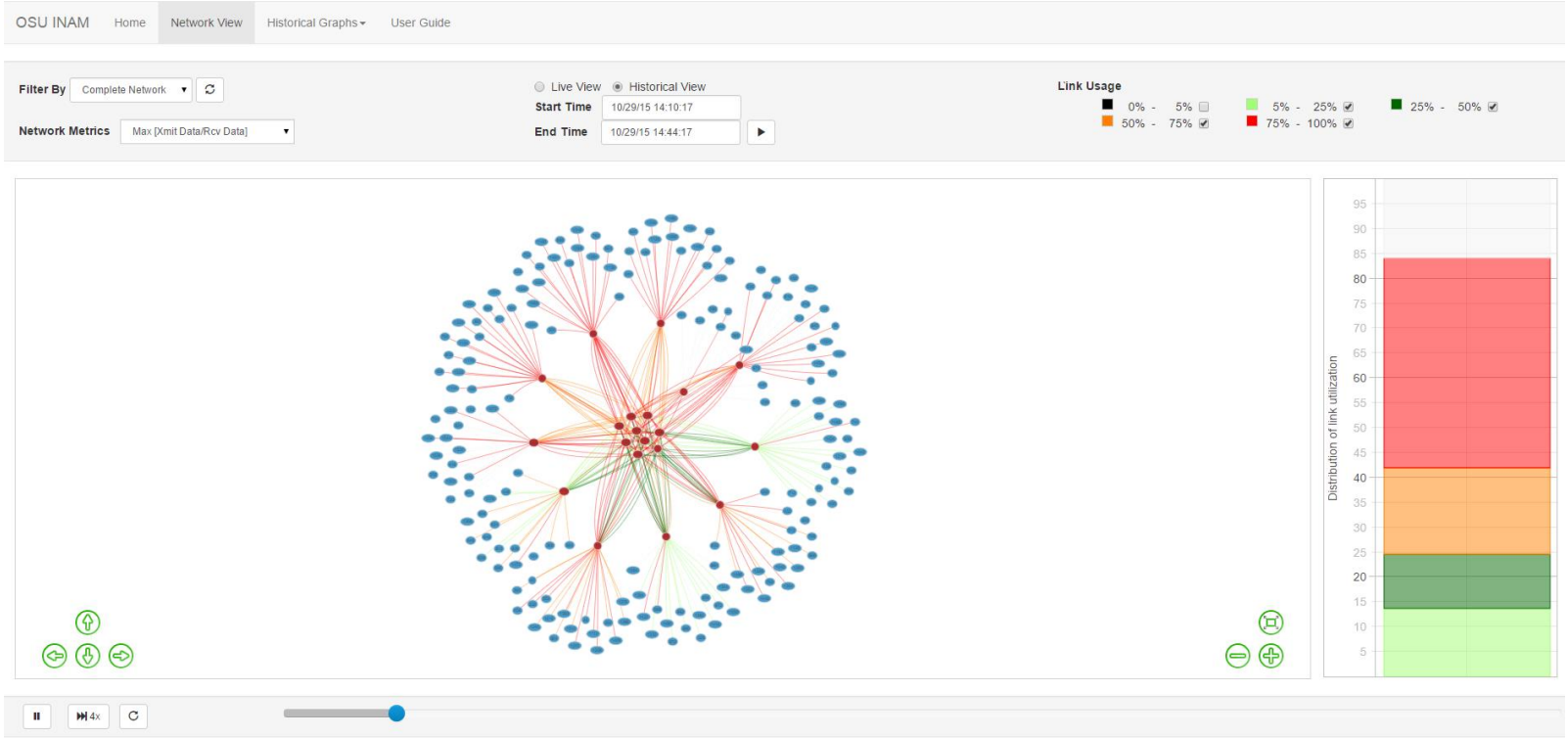
Rank112 [core 0]

Rank113 [core 1]

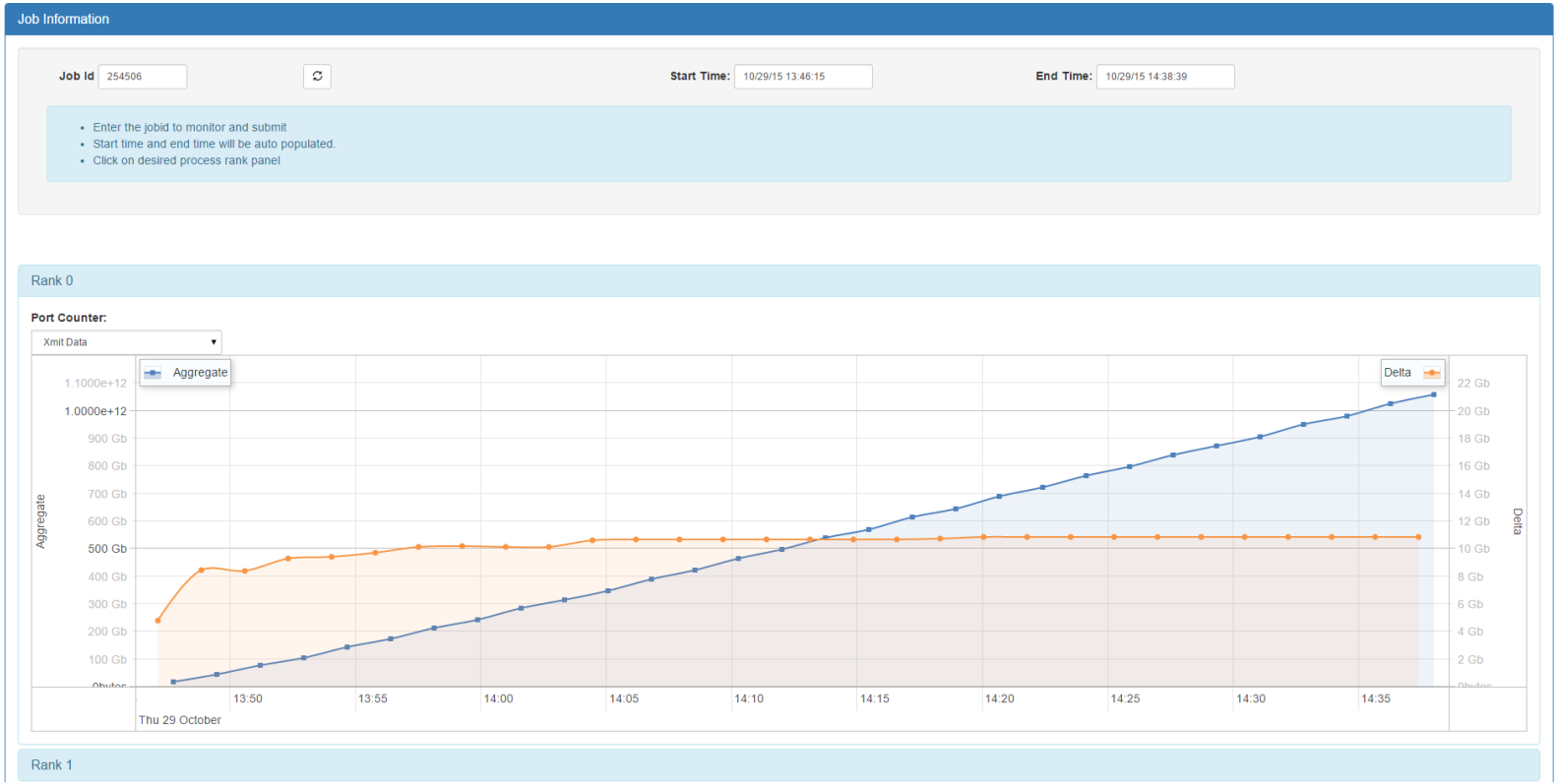
Live Switch Level View



Historical Network View



Historical Job View



Outline

- Introduction
- Motivation
- Challenges & Contributions
- Features of OSU INAM & Demo
- Conclusions

Conclusions

- Designed OSU INAM capable of analyzing the communication traffic on the InfiniBand network with inputs from the MPI runtime
- Major features of the OSU INAM tool include:
  - Analyze and profile network-level activities with many parameters (data and errors) at user specified granularity
  - Capability to analyze and profile node-level, job-level and process-level activities for MPI communication (Point-to-Point, Collectives and RMA)
  - Remotely monitor CPU utilization of MPI processes at user specified granularity
  - Visualize the data transfer happening in a "live" fashion for
    - Entire Network - Live Network Level View
    - Particular Job - Live Job Level View
    - One or multiple Nodes - Live Node Level View
    - One or multiple Switches - Live Switch Level View
- Capability to visualize data transfer that happened in the network at a time duration in the past for
  - Entire Network - Historical Network Level View
  - Particular Job - Historical Job Level View
  - One or multiple Nodes - Historical Node Level View

Thank you!

{subramon, augustal, arnoldm, perkinjo, luxi, chakrabs, hamidouc, panda}
@cse.ohio-state.edu

Network-Based Computing Laboratory

http://mvapich.cse.ohio-state.edu/