MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

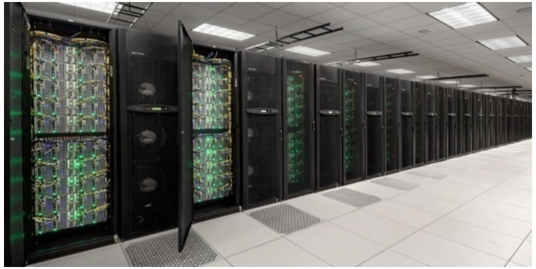
OSU INAM: Visualize Your HPC Network

Mark Arnold

Department of Computer Science and Engineering
The Ohio State University

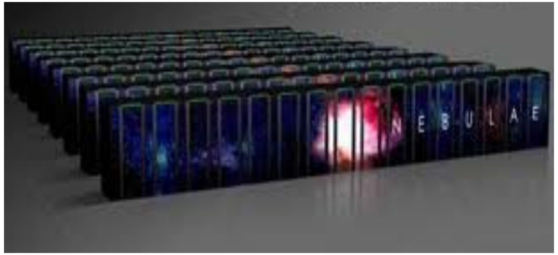
Current Trends in HPC

- Supercomputing systems scaling rapidly
  - Multi-core architectures and
  - High-performance interconnects
- InfiniBand is a popular HPC interconnect
  - 163 systems (32.6%) in Nov'17 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**

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

Overview of the MVAPICH2 Project

- High Performance open-source MPI Library for InfiniBand, Omni-Path, Ethernet/iWARP, and RDMA over Converged Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.0), Started in 2001, First version available in 2002
  - MVAPICH2-X (MPI + PGAS), Available since 2011
  - Support for GPGPUs (MVAPICH2-GDR) and MIC (MVAPICH2-MIC), Available since 2014
  - Support for Virtualization (MVAPICH2-Virt), Available since 2015
  - Support for Energy-Awareness (MVAPICH2-EA), Available since 2015
  - Support for InfiniBand Network Analysis and Monitoring (OSU INAM) since 2015
  - **Used by more than 2,825 organizations in 85 countries**
  - **More than 427,000 (> 0.4 million) downloads from the OSU site directly**
  - Empowering many TOP500 clusters (June '17 ranking)
    - **1st, 10,649,600-core (Sunway TaihuLight) at National Supercomputing Center in Wuxi, China**
    - 15th, 241,108-core (Pleiades) at NASA
    - 20th, 462,462-core (Stampede) at TACC
    - 44th, 74,520-core (Tsubame 2.5) at Tokyo Institute of Technology
  - Available with software stacks of many vendors and 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) ->
  - Sunway TaihuLight (1st in Jun'17, 10M cores, 100 PFlops)

16 Years & Going Strong!

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..

Broad Challenge

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

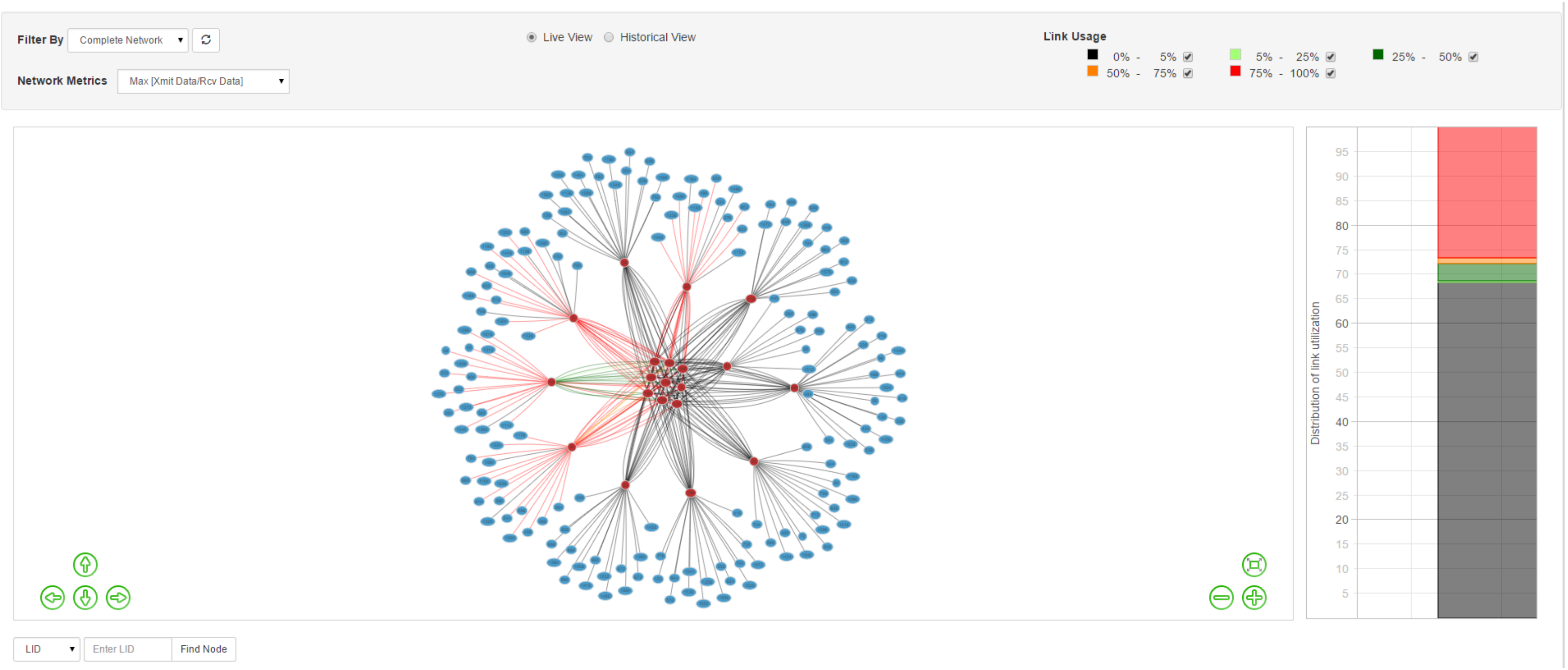
Contributions

- 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

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

Live Network Level View



Filter By Complete Network ▾

Network Metrics Max [Xmit Data/Rcv Data] ▾

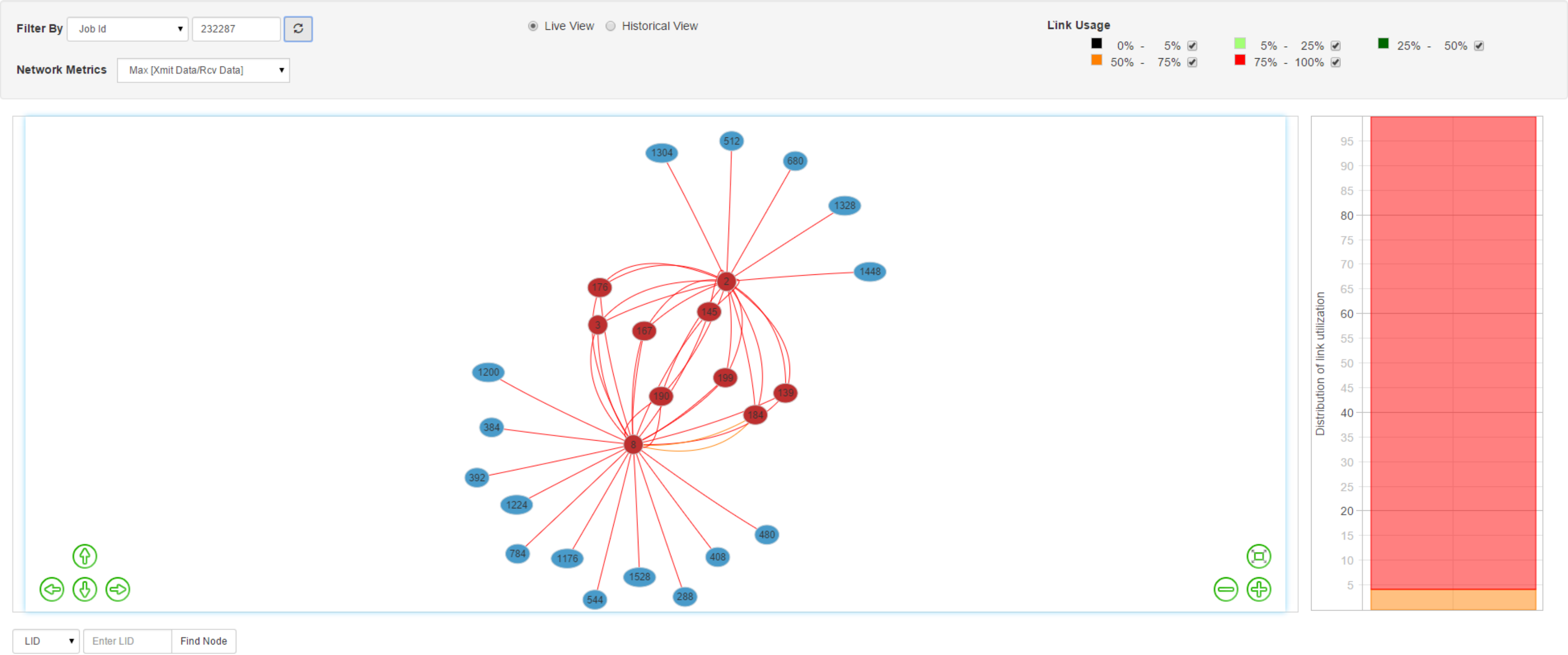
● Live View ○ Historical View

Link Usage

■ 0% - 5% ☑ 🟩 5% - 25% ☑ 🟩 25% - 50% ☑

🟧 50% - 75% ☑ 🟥 75% - 100% ☑

Live Job Level View



Filter By: Job Id ▾ | 232287

● Live View ○ Historical View

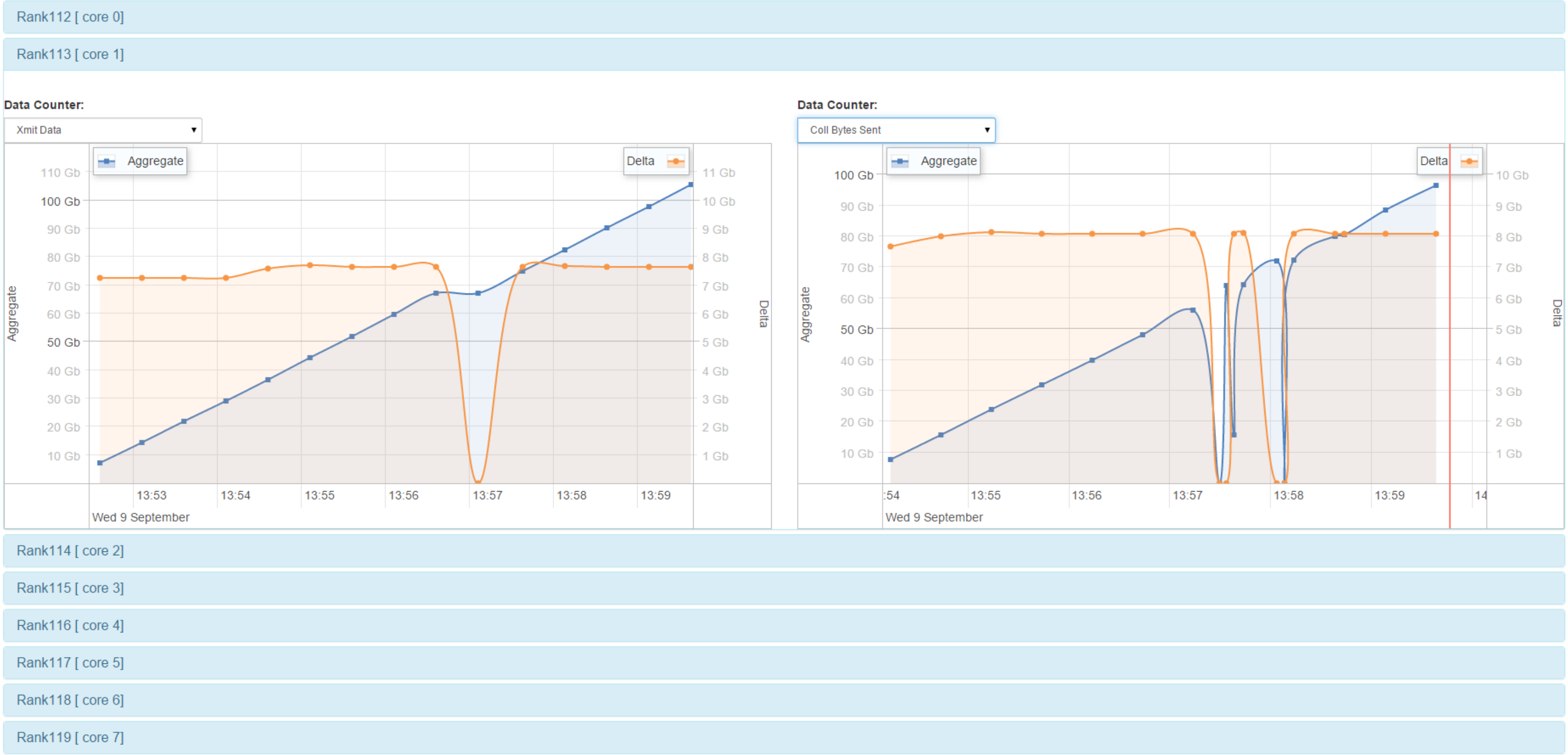
Network Metrics: Max [Xmit Data/Rcv Data] ▾

Link Usage

■ 0% - 5% ☑	🟩 5% - 25% ☑	🟩 25% - 50% ☑
🟧 50% - 75% ☑	🟥 75% - 100% ☑	

LID ▾ | Enter LID | Find Node

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]

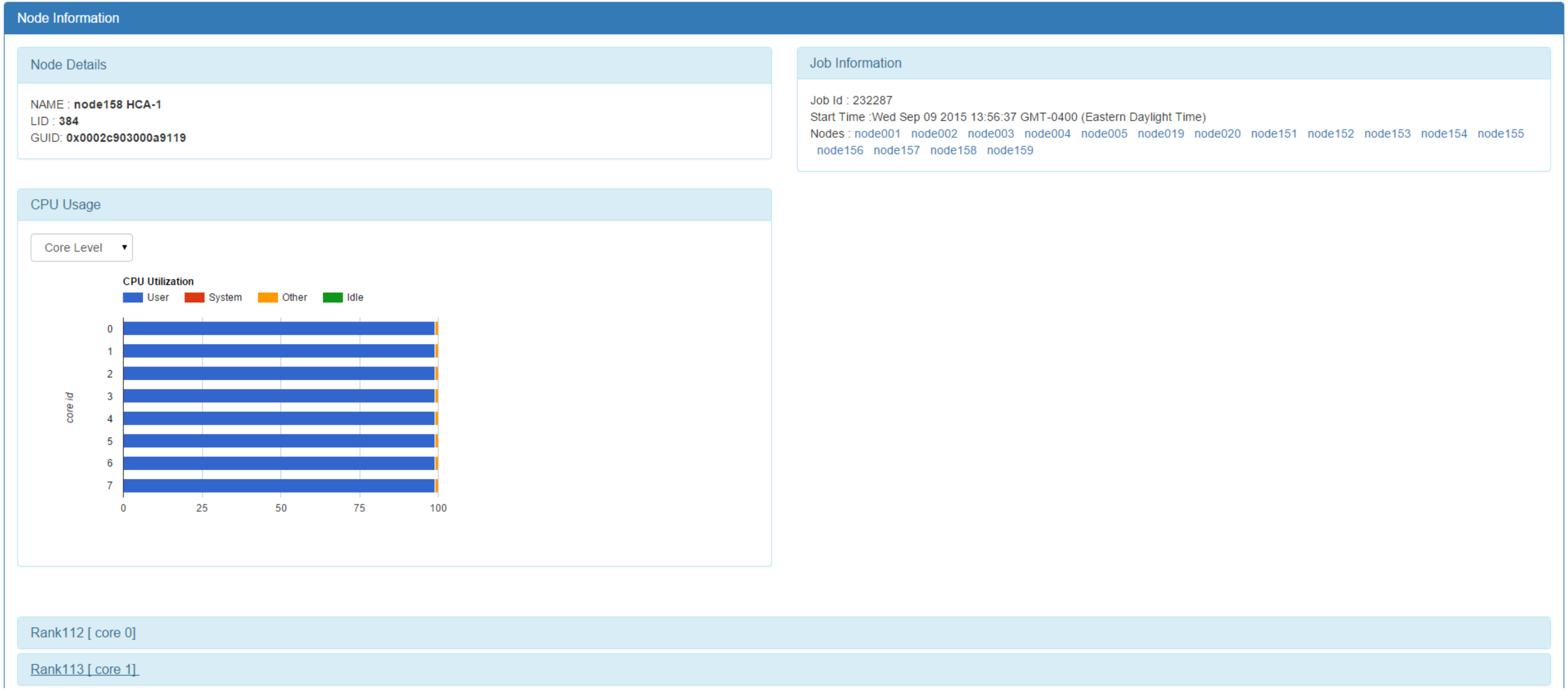
Rank116 [core 4]

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

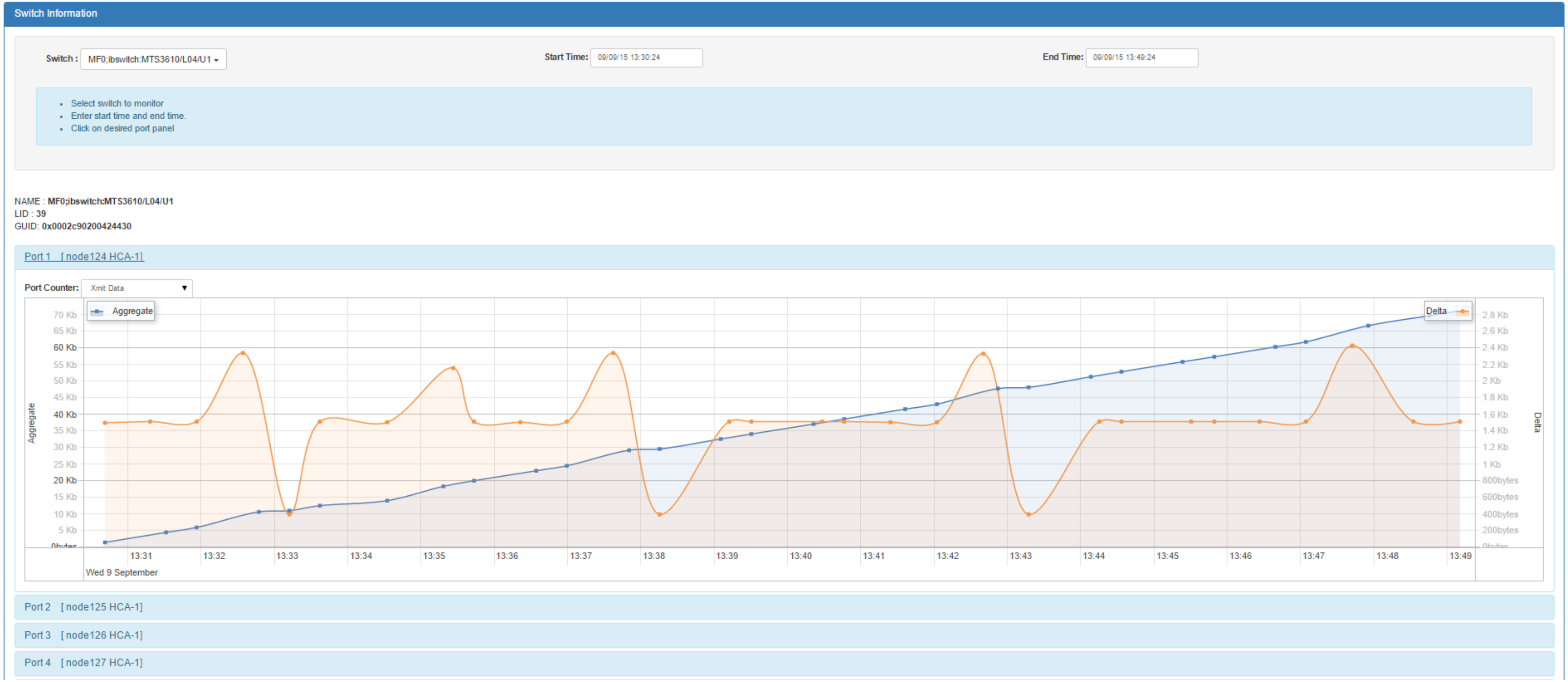
CPU Usage

Core Level ▾

Rank112 [core 0]

Rank113 [core 1]

Live Switch Level View



Switch Information

Switch : MF0;ibswitch:MTS3610/L04/U1

Start Time: 09/09/15 13:30:24

End Time: 09/09/15 13:49:24

- Select switch to monitor
- Enter start time and end time.
- Click on desired port panel

NAME : **MF0;ibswitch:MTS3610/L04/U1**
LID : **39**
GUID: **0x0002c90200424430**

Port 1 [node124 HCA-1]

Port Counter: Xmit Data

Port 2 [node125 HCA-1]

Port 3 [node126 HCA-1]

Port 4 [node127 HCA-1]

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