HiBD
High-Performance
Big Data

MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

HiDL
High-Performance
Deep Learning

Understanding Communication Performance in HPC by using OSU INAM

Pouya Kousha

PhD student @ The Ohio State University

Advisor: Prof. DK Panda

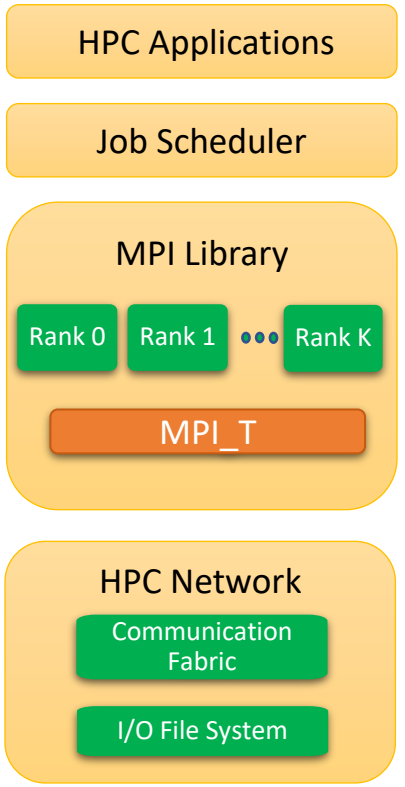
Overview

- Profiling tool challenges
- Usage case
- Overview of OSU INAM
- Current OSU INAM features
- Demo

Profiling Tools Perspective and Broad Challenges

- There are 30+ profiling tools for HPC systems
- System level vs User level
  - User level novelty
- Different set of users have different needs
  - HPC administrators
  - HPC Software developers
  - Domain scientists
- Different HPC layers to profile
  - How to correlate them and pinpoint the problem source?

Unified and holistic view for all users



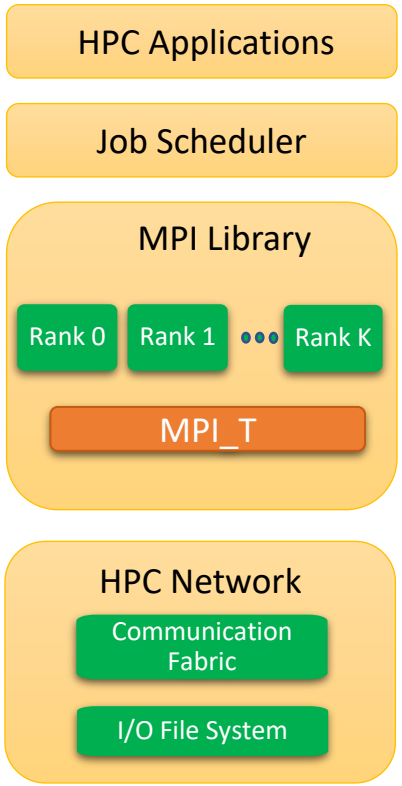
Summary of existing profiling tools and their capabilities

Tools	MPI Runtime		
	Applications	Network Fabric	Job scheduler
INAM*	✓	✓	✓
TAU	✓	✓	✗
HPCToolkit	✓	✗	✗
Intel Vtune	✓	✗	✗
IPM	✓	✗	✗
mpiP	✓	✗	✗
Intel ITAC	✓	✗	✗
ARM MAP	✓	✗	✗
HVProf	✓	✗	✗
PCP(used by XDMOD)	✗	✓	✓
Prometheus	✗	✓	✓
Mellanox FabricIT	✗	✓	✗
BoxFish	✗	✓	✗
LDMS	✗	✓	✗

* This design has been publicly released on 06/08/2020 and is available for free here https://mvapich.cse.ohio-state.edu/tools/osu-inam/

Profiling Tools Perspective and Broad Challenges

- Understanding the interaction between applications, MPI libraries, I/O and the communication fabric is challenging
  - Find root causes for performance degradation
  - Identify which layer is causing the possible issue
  - Understand the internal interaction and interplay of MPI library components and network level
  - Online profiling



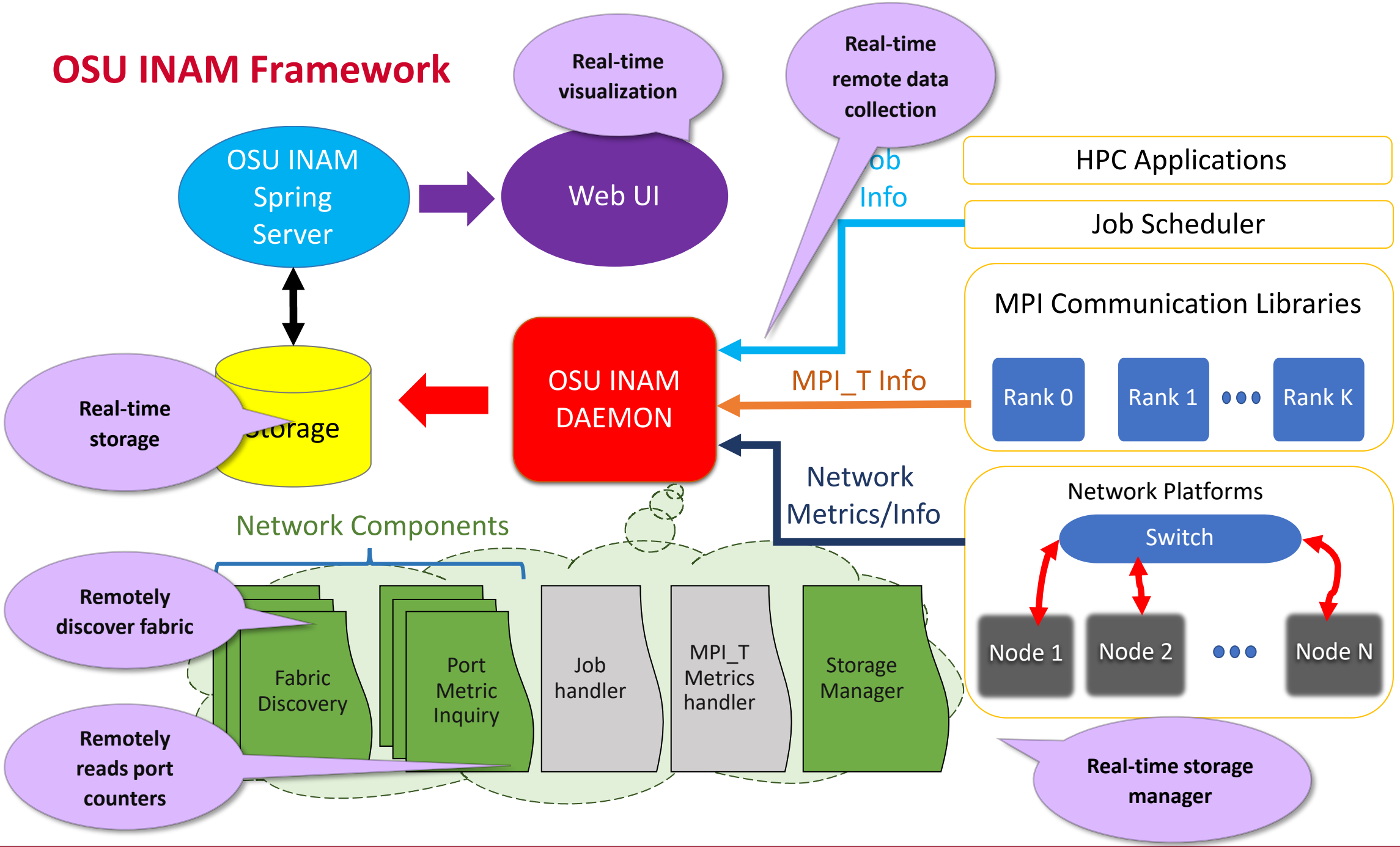
How can we design a tool that enables holistic, real-time, scalable and in-depth understanding of communication traffic through tight integration with the MPI runtime and job scheduler?

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**

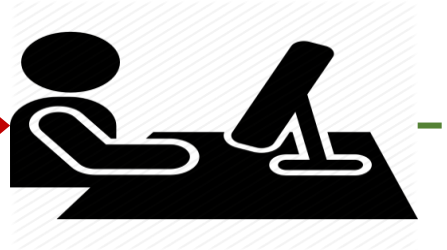
- **OSU INAM v1 released (11/10/2022)**
  - Support for MySQL and InfluxDB as database backends
  - Support for data loading progress bars on the UI for all charts
  - Enhanced database insertion using InfluxDB
  - Enhanced the UI APIs by making asynchronous calls for data loading
  - Support for continuous queries to improve visualization performance
  - Support for SLURM multi-cluster configuration
  - Significantly improved database query performance when using InfluxDB
  - Support for automatic data retention policy when using InfluxDB
  - 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 Framework



Flow of Using OSU INAM

Insight and Performance Recommendations



Tuning and Applying Recommendations

REMOTE OSU INAM

- Notifications and Feedback
- Real-time Visualization
- I/O and MPI Traffic Interaction
- Measurements
- Metric Collectors

Job Scheduler

- SLURM
- PBS Torque

HPC/DL Applications

- Co-designed Scientific Applications using Proposed MPI_T Extensions
- TensorFlow
- PyTorch
- MxNet
- Co-designed Horovod using Proposed MPI_T Extensions

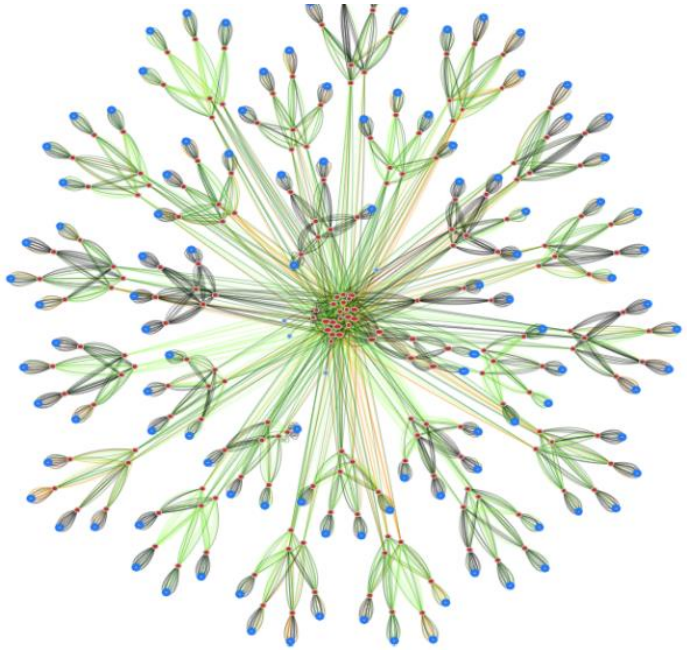
MPI Library with Support for Enhanced MPI_T-based Introspection

- Enhanced MPI_T Sessions
- Gathering I/O Traffic Information
- Enhanced PVAR Infrastructure
  - Counters
  - Buckets
  - Timers

HPC Network

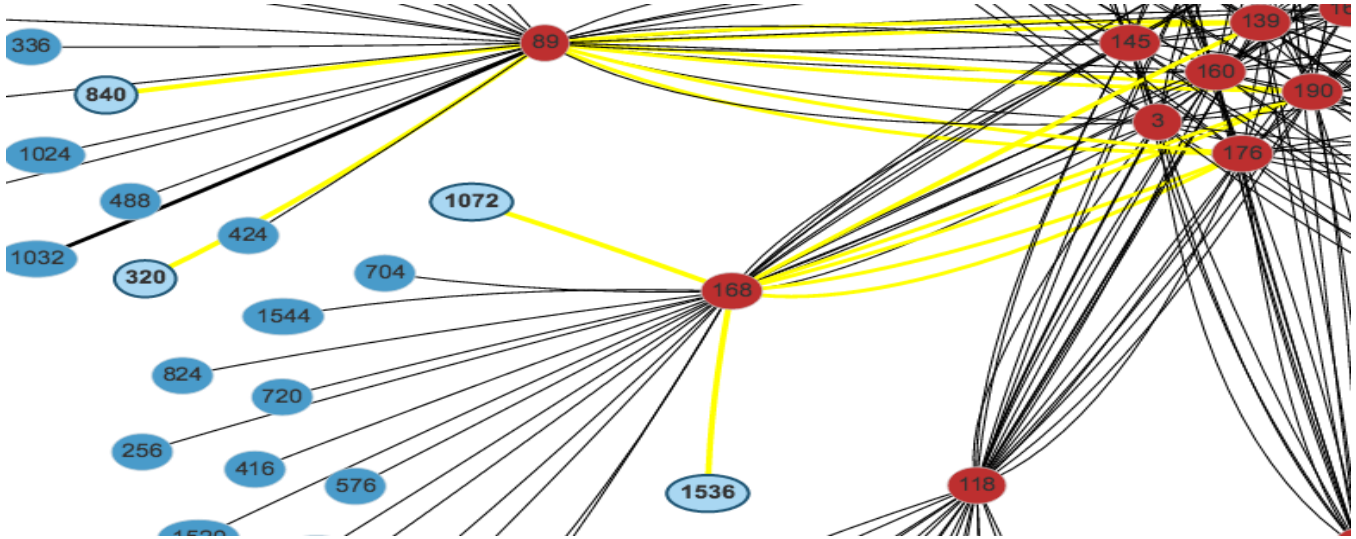
- InfiniBand
- Lustre, NFS
- CPU/Mem Usage

OSU INAM Features



Comet@SDSC --- Clustered View

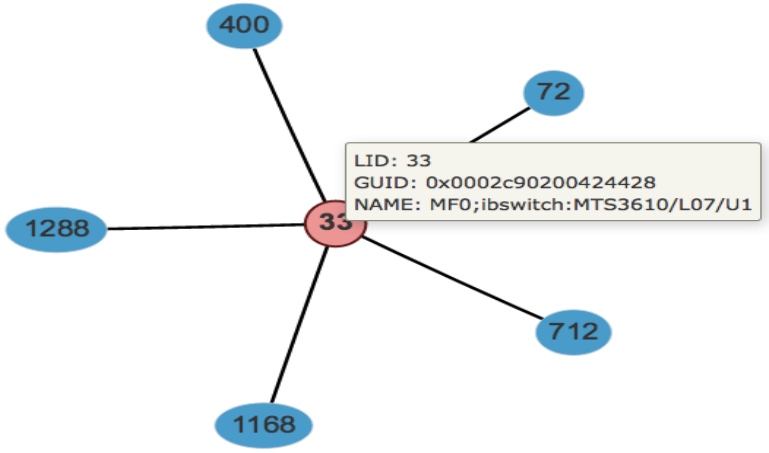
(1,879 nodes, 212 switches, 4,377 network links)



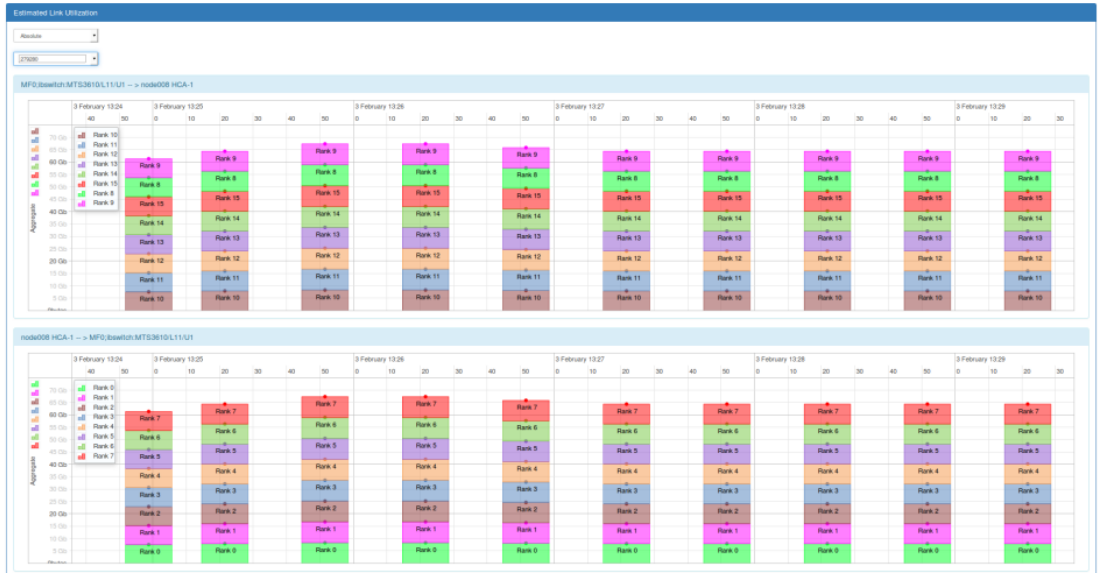
Finding Routes Between Nodes

- Show network topology of large clusters
- Visualize job topology in the network
- Visualize traffic pattern on different links
- Quickly identify congested links/links in error state
- See the history unfold – play back historical state of the network

OSU INAM Features (Cont.)



Visualizing a Job (5 Nodes)



Estimated Process Level Link Utilization

- Job level view
  - Show different network metrics (load, error, etc.) for any live job
  - Play back historical data for completed jobs to identify bottlenecks
- Node level view - details per process or per node
  - CPU and memory utilization for each rank/node
  - Bytes sent/received for MPI operations (pt-to-pt, collective, RMA)
  - Network metrics (e.g. XmitDiscard, RcvError) per rank/node

- Estimated Link Utilization view
  - Classify data flowing over a network link at different granularity in conjunction with MVAPICH2-X 2.2rc1
    - Job level and
    - Process level

More Details in Tutorial/Demo

Live Demo at OSC and OSU clusters