OSU INAM: A Profiling and Visualization Tool for Scalable and In-Depth Analysis HPC Clusters

Pouya Kousha

PhD student @ The Ohio State University

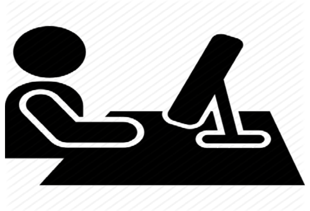
Advisor: Prof. DK Panda

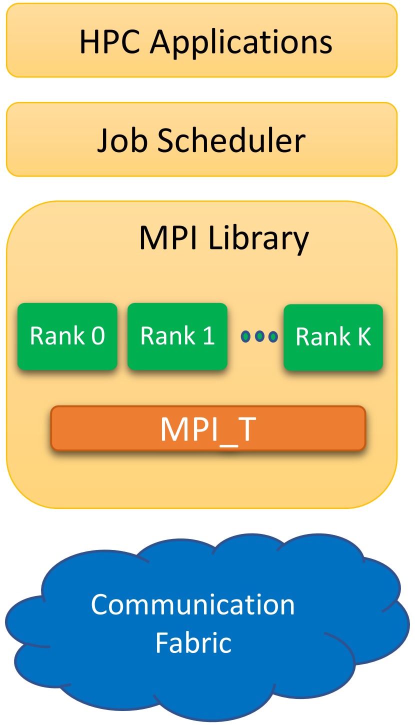
Overview

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

Profiling Tools Perspective and Challenges

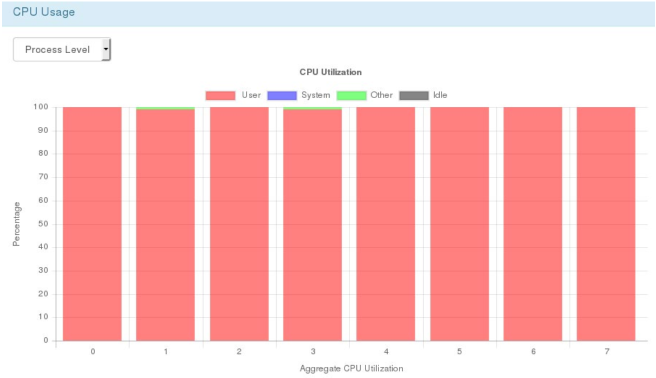
- There are 30+ profiling for HPC systems
- System level vs User level
  - User level novelty
- Different types of Users have different needs
  - HPC administrators
  - HPC Software developers
  - Domain scientists
- Different HPC layers to profile
  - How to correlate them?

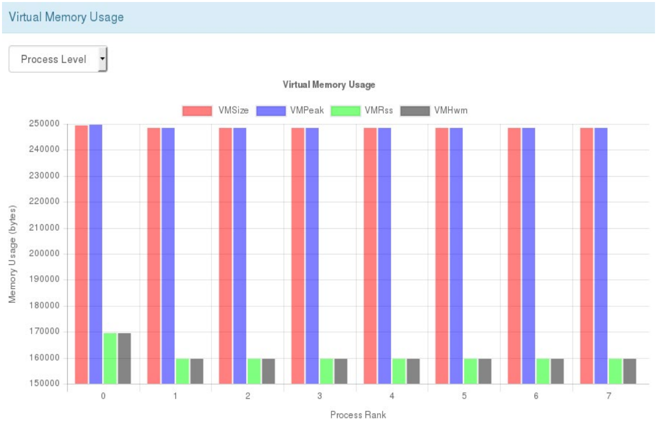




Use Case: domain scientist

- You are a domain scientist running your application
  - Expecting getting better results, you get performance degradation
  - Where is the source of degradation in HPC system?
- How can a domain scientist get a holistic view of the HPC ecosystem?
  - Integration with job scheduler, MPI library, and fabric
  - In-depth performance monitoring
- High productivity tools perspective for HPC users
- Capable to reuse the stored data from OSU INAM daemon

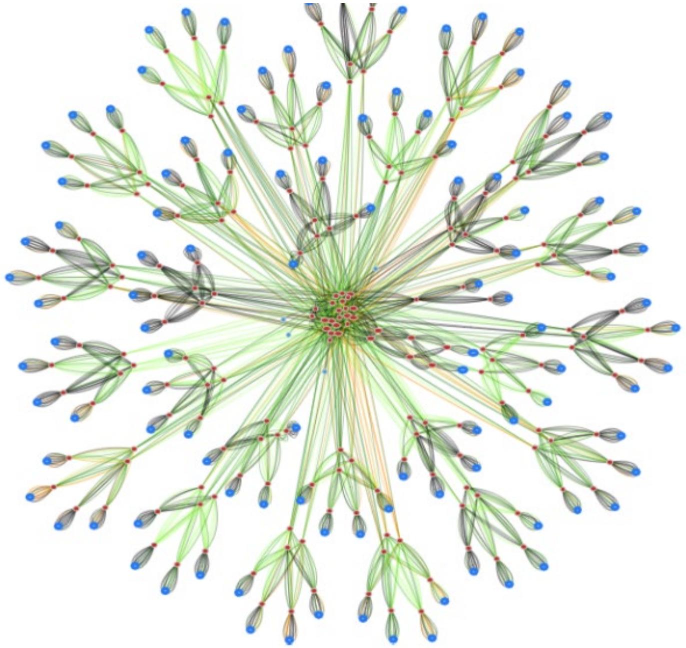




Overview of OSU INAM

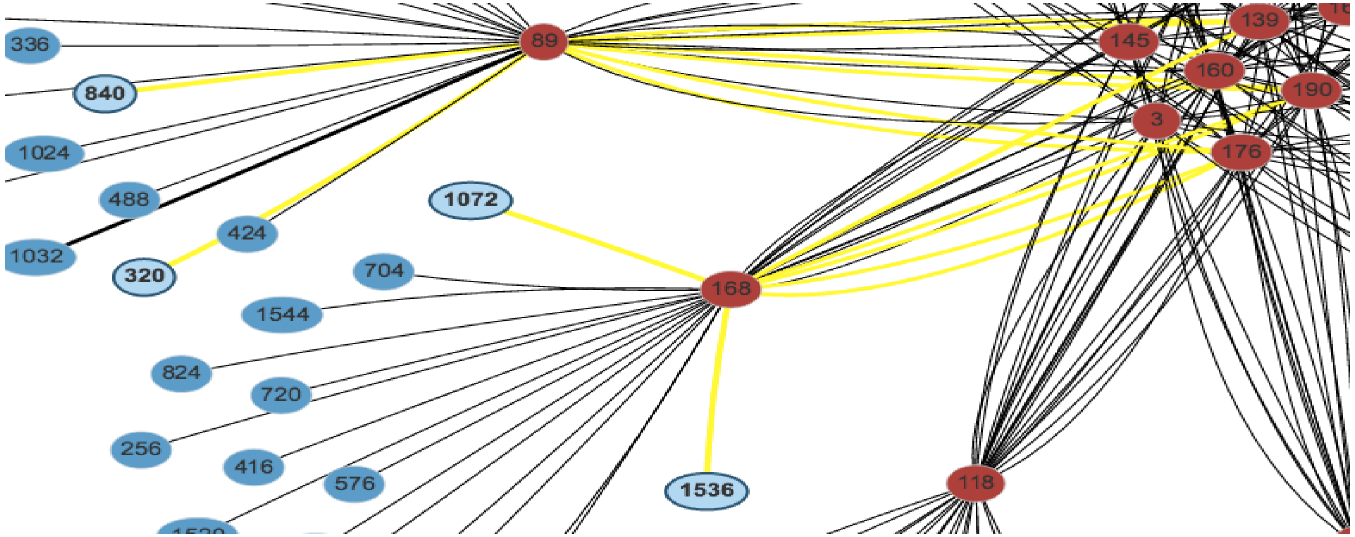
- A network monitoring and analysis tool that is capable of analyzing traffic on the InfiniBand network with inputs from the MPI runtime
- Remotely monitors IB clusters in real time by querying various subnet management entities and gathering input from the MPI runtimes
- Capability to analyze and profile 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
- Fabric discovery in less than 10 mins for ~2000 nodes
- Sub-second IB port inquiry for ~2000 nodes
  - Enhanced fault tolerance for database operations
    - Thanks to Trey Dockendorf @ OSC for the feedback
  - OpenMP-based multi-threaded designs to handle database purge, read, and insert operations simultaneously
  - Improved database purging time by using bulk deletes
  - Tune database timeouts to handle very long database operations
  - Improved debugging support by introducing several debugging levels

OSU INAM Features



Comet@SDSC --- Clustered View

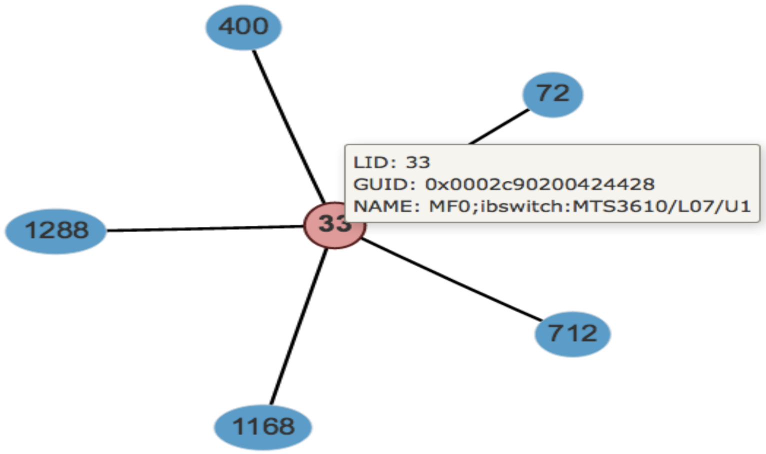
(1,879 nodes, 212 switches, 4,377 network links)



Finding Routes Between Nodes

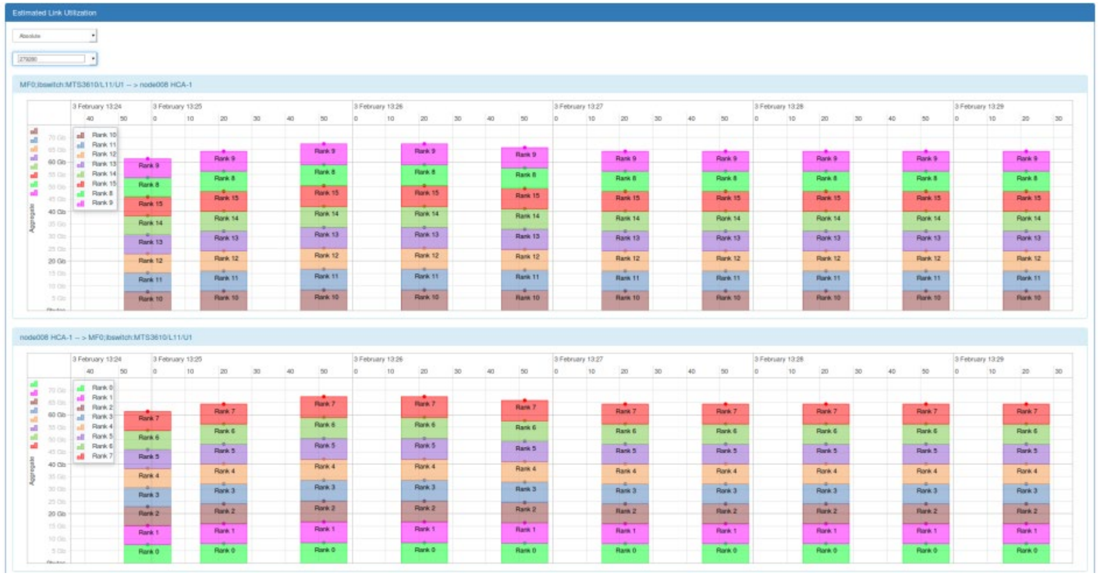
- Show network topology of large clusters
- Visualize job topology in the network
- Visualize traffic pattern on different links
- Quickly identify congested links/links in error state
- See the history unfold – play back historical state of the network

OSU INAM Features (Cont.)



LID: 33
GUID: 0x0002c90200424428
NAME: MF0;ibswitch:MTS3610/L07/U1

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