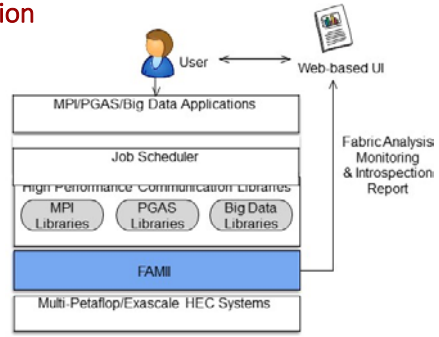
SI2-SSI (2020): FAMII: High-Performance and Scalable Fabric Analysis, Monitoring and Introspection Infrastructure for HPC and Big Data

PI: Dhabaleswar K. Panda, Co-Pis: Karen Tomko, Hari Subramoni, Heechang Na
Institutions: The Ohio State University, The Ohio Supercomputer Center

Award #: OAC-1664137

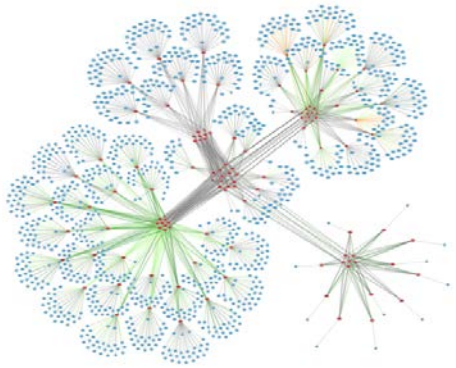
Vision

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/Deep Learning applications through tight integration with the communication runtime and the job scheduler?

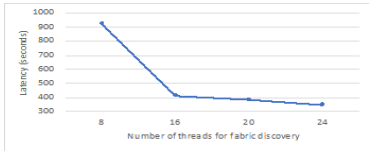


Enhanced Fabric Discovery and Port Metrics Inquiry

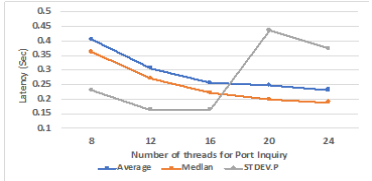
- Enhanced performance for fabric discovery using optimized OpenMP-based multi-threaded designs with 14x speedup
- Ability to gather InfiniBand performance counters at sub-second granularity for very large (>2,000 nodes) clusters



Network View of Ohio Supercomputer Center (OSC) with 3 heterogeneous clusters all connected to the same InfiniBand Fabric

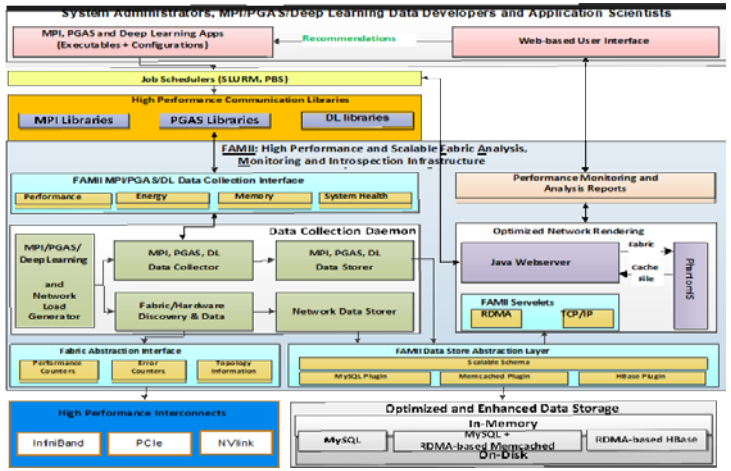


Impact of multi-threading on Fabric Discovery module on OSC cluster



Impact of multi-threading on Port Inquiry module on OSC cluster

Proposed Framework



The Proposed Performance Monitoring, Analysis, and Introspection Framework

High-Performance, Low Overhead, and Scalable GPU Profiling

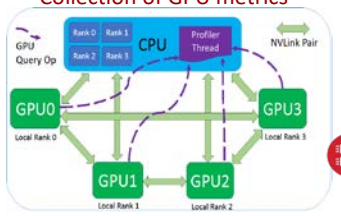
Each node will aggregate and send the GPU and PVAR metrics to OSU INAM

Startup: Each rank discovers the topology and updates shared region. Then, one rank per node setups and starts a profiler thread on CPU to profile all GPUs on the node once using GPUs.

Query: The profiler thread profile all enrolled GPUs based on user defined interval and send data to OSU INAM periodically

Exit: Once the ranks stop using device, profiler thread will perform one last read and send data then exit.

Collection of GPU metrics



Software Release, Community Engagement & Metrics

- A v0.9.5 release of OSU INAM has been made on Jan'20
  - http://mvapich.cse.ohio-state.edu/tools/osu-inam/
  - More than 600 downloads with support for PBS and SLURM
- This release has been installed at OSC and OSU to monitor clusters
- Tutorials at SC '19, ISC'19, HiPEAC '20, MUG'19
- Community Engagement with: NOAA, U. of Utah, CAE Services @ Germany, Pratt & Whitney, Ghent University @ Germany, and Cyfronet @ Poland

Research Publications

1. Designing a Profiling and Visualization Tool for Scalable and In-Depth Analysis of High-Performance GPU Clusters, P. Kousha, B. Ramesh, K. Kandadi Suresh, C. Chu, A. Jain, N. Sarkauskas, D. Panda, IEEE HiPC, Dec 2019
2. A. Ruhela, H. Subramoni, S. Chakraborty, M. Bayatpour, P. Kousha, and DK Panda, Efficient Design for MPI Asynchronous Progress without Dedicated Resources , Parallel Computing - Systems & Applications, Volume 85, July 2019.

Future Work

- Extend data collection daemon to further intra-node metrics, intra-node communication matrix, and power metrics
- Support to profile multiple MPI libraries through MPI_T interface
- Extend support for introspection of PGAS and DL applications

Supported by
OAC-1664137 & TG-NCR130002

Ohio Supercomputer Center
An OH·TECH Consortium Member

The Ohio State University

MVAPICH