The MVAPICH2 Project: Heading Towards New Horizons in Energy-Awareness, Virtualization and Network/Job-Level Introspection

Talk at OSC/OH-TECH Booth (SC ‘15)

by

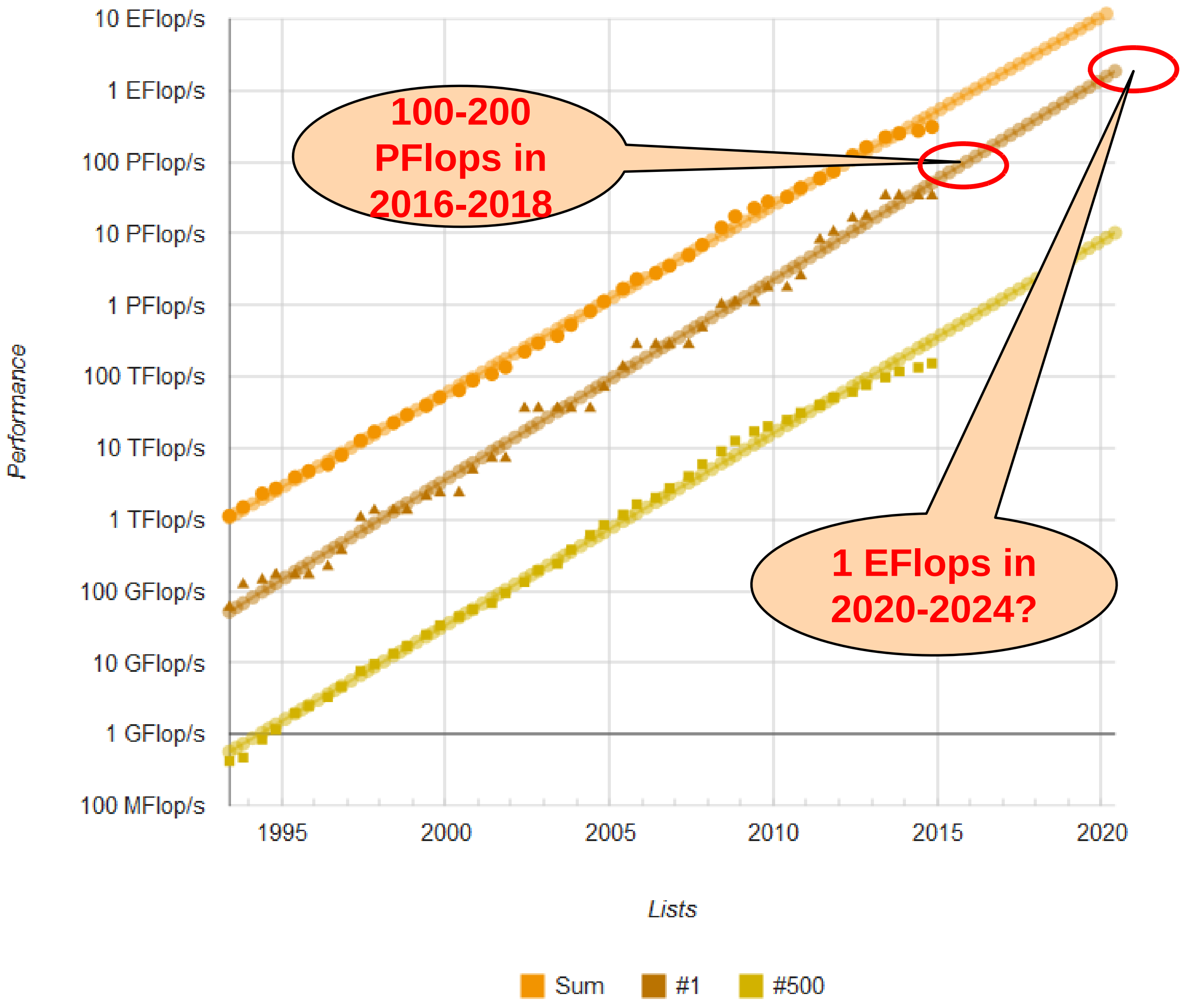
Dhabaleswar K. (DK) Panda

The Ohio State University

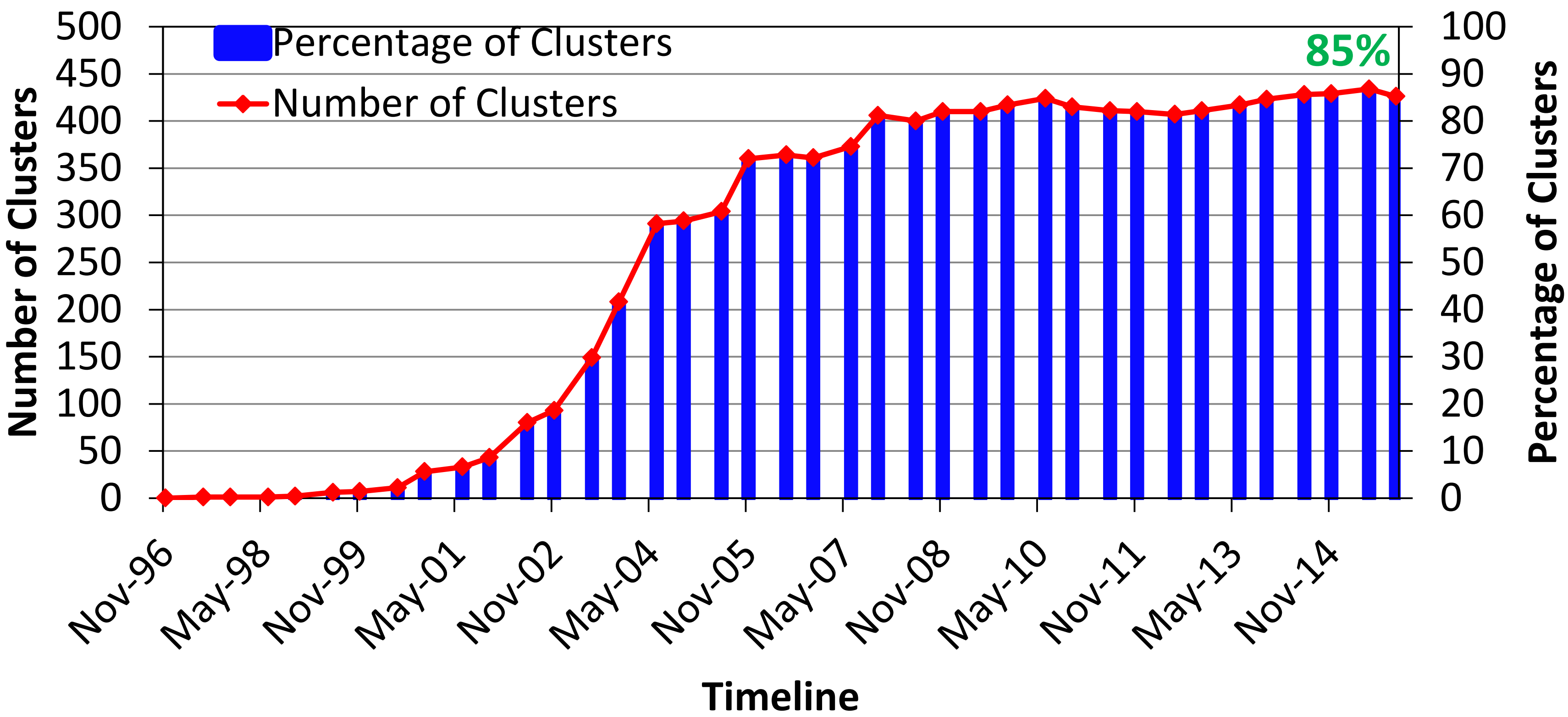
E-mail: panda@cse.ohio-state.edu

http://www.cse.ohio-state.edu/~panda

High-End Computing (HEC): PetaFlop to ExaFlop



Trends for Commodity Computing Clusters in the Top 500 List (http://www.top500.org)



Drivers of Modern HPC Cluster Architectures

Multi-core Processors

High Performance Interconnects - InfiniBand <1usec latency, 100Gbps Bandwidth>

Accelerators / Coprocessors high compute density, high performance/watt >1 TFlop DP on a chip

SSD, NVMe-SSD, NVRAM

- Multi-core/many-core technologies
- Remote Direct Memory Access (RDMA)-enabled networking (InfiniBand and RoCE)
- Solid State Drives (SSDs), Non-Volatile Random-Access Memory (NVRAM), NVMe-SSD
- Accelerators (NVIDIA GPGPUs and Intel Xeon Phi)

Tianhe – 2

Titan

Stampede

Tianhe – 1A

Large-scale InfiniBand Installations

- 235 IB Clusters (47%) in the Nov' 2015 Top500 list (http://www.top500.org)
- Installations in the Top 50 (21 systems):

462,462 cores (Stampede) at TACC (10th)	76,032 cores (Tsubame 2.5) at Japan/GSIC (25th)
185,344 cores (Pleiades) at NASA/Ames (13th)	194,616 cores (Cascade) at PNNL (27th)
72,800 cores Cray CS-Storm in US (15th)	76,032 cores (Makman-2) at Saudi Aramco (32nd)
72,800 cores Cray CS-Storm in US (16th)	110,400 cores (Pangea) in France (33rd)
265,440 cores SGI ICE at Tulip Trading Australia (17th)	37,120 cores (Lomonosov-2) at Russia/MSU (35th)
124,200 cores (Topaz) SGI ICE at ERDC DSRC in US (18th)	57,600 cores (SwiftLucy) in US (37th)
72,000 cores (HPC2) in Italy (19th)	55,728 cores (Prometheus) at Poland/Cyfronet (38th)
152,692 cores (Thunder) at AFRL/USA (21st)	50,544 cores (Occigen) at France/GENCI-CINES (43rd)
147,456 cores (SuperMUC) in Germany (22nd)	76,896 cores (Salomon) SGI ICE in Czech Republic (47th)
86,016 cores (SuperMUC Phase 2) in Germany (24th)	**and many more!**

Designing High-Performance Middleware for HPC: Challenges

Application Kernels/Applications

Middleware

Programming Models

MPI, PGAS (UPC, Global Arrays, OpenSHMEM), CUDA, OpenMP, OpenACC, Cilk, Hadoop (MapReduce), Spark (RDD, DAG), etc.

Communication Library or Runtime for Programming Models

- Point-to-point Communication (two-sided and one-sided
- Collective Communication
- Energy-Awareness
- Synchronization and Locks
- I/O and File Systems
- Fault Tolerance

Networking Tech. (InfiniBand, 40/100GigE, Aries, and OmniPath)

Multi/Many-core Architectures

Accelerators (NVIDIA and MIC)

Storage Tech. (HDD, SSD, and NVMe-SSD)

Co-Design Opportunities and Challenges across Various Layers

Performance

Scalability

Fault-Resilience

Broad Challenges in Designing Communication Libraries for (MPI+X) at Exascale

- Scalability for million to billion processors
  - Support for highly-efficient inter-node and intra-node communication (both two-sided and one-sided)
- Scalable Collective communication
  - Offload
  - Non-blocking
  - Topology-aware
- Balancing intra-node and inter-node communication for next generation multi-core (128-1024 cores/node)
  - Multiple end-points per node
- Support for efficient multi-threading
- Integrated Support for GPGPUs and Accelerators
- Fault-tolerance/resiliency
- QoS support for communication and I/O
- Support for Hybrid MPI+PGAS programming (MPI + OpenMP, MPI + UPC, MPI + OpenSHMEM, CAF, …)
- Virtualization
- Energy-Awareness
- Integrated Network Management

MVAPICH2 Software

- High Performance open-source MPI Library for InfiniBand, 10-40Gig/iWARP, and RDMA over Converged Enhanced Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.0), Available since 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
  - **Used by more than 2,475 organizations in 76 countries**
  - **More than 307,000 downloads from the OSU site directly**
  - Empowering many TOP500 clusters (Nov '15 ranking)
    - 10th ranked 519,640-core cluster (Stampede) at TACC
    - 13th ranked 185,344-core cluster (Pleiades) at NASA
    - 25th ranked 76,032-core cluster (Tsubame 2.5) at Tokyo Institute of Technology and many others
  - 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) ->
  - Stampede at TACC (10th in Nov'15, 519,640 cores, 5.168 Plops)

Overview of A Few Challenges being Addressed by MVAPICH2 Project for Exascale

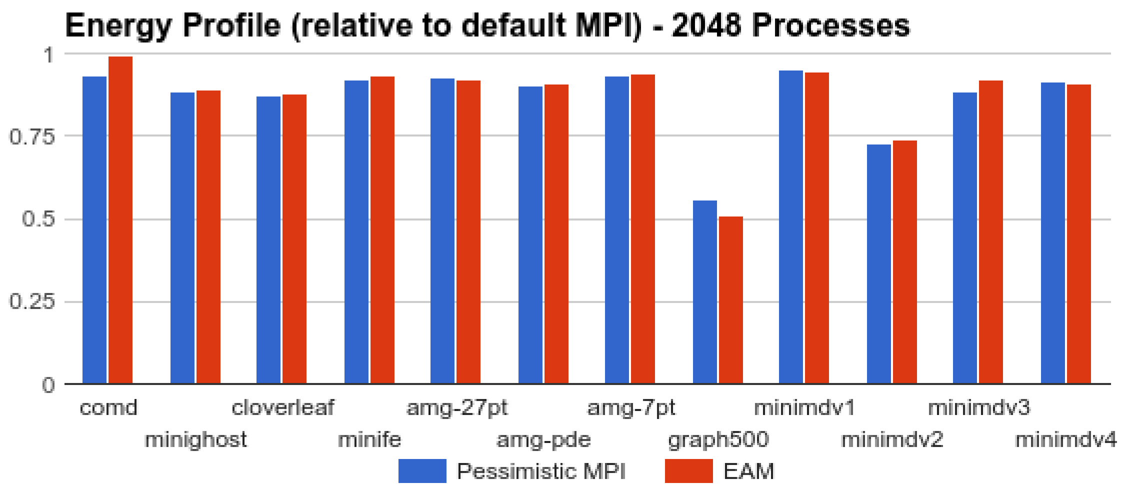
- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- MVAPICH2-Virt
  - Support for Basic SR-IOV
  - Locality-aware communication
  - Building HPC Cloud
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool

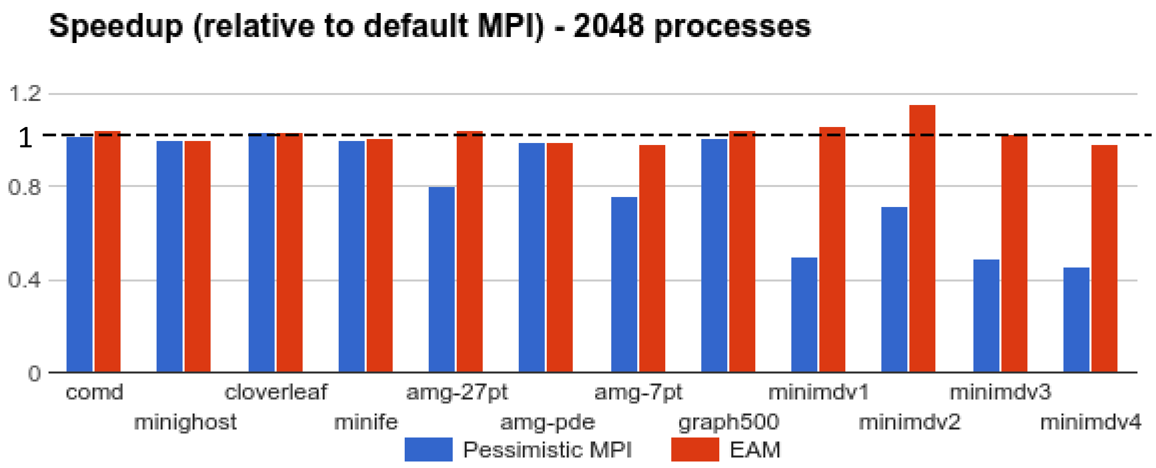
Energy-Aware MVAPICH2 Library and OSU Energy Management Tool (OEMT)

- MVAPICH2-EA (Energy-Aware) MPI Library
  - Production-ready Energy-Aware MPI Library
  - New Energy-Efficient communication protocols for pt-pt and collective operations
  - Intelligently apply the appropriate energy saving techniques
  - Application oblivious energy saving
  - Released 08/28/15
- OEMT
  - A library utility to measure energy consumption for MPI applications
  - Works with all MPI runtimes
  - PRELOAD option for precompiled applications
  - Does not require ROOT permission:
    - A safe kernel module to read only a subset of MSRs
- Available from: http://mvapich.cse.ohio-state.edu

MVAPICH2-EA: Application Oblivious Energy-Aware-MPI (EAM)

- An energy efficient runtime that provides energy savings without application knowledge
- A **white-box** approach
- **Automatically and transparently** use the best energy lever
- Provides **guarantees on maximum degradation** with 5-41% savings at <= 5% degradation
- Pessimistic MPI applies energy reduction lever to each MPI call





A Case for Application-Oblivious Energy-Efficient MPI Runtime A. Venkatesh , A. Vishnu , K. Hamidouche , N. Tallent , D. K. Panda , D. Kerbyson , and A. Hoise - Supercomputing '15, Nov 2015 , ***Best Student Paper Finalist, presented in the Technical Papers Program, Tuesday 3:30-4:00pm (Room 18CD)***

Overview of A Few Challenges being Addressed by MVAPICH2 Project for Exascale

- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- MVAPICH2-Virt
  - Support for Basic SR-IOV
  - Locality-aware communication
  - Building HPC Cloud
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool

HPC Cloud - Combining HPC with Cloud

- IDC expects that by 2017, HPC ecosystem revenue will jump to a record $30.2 billion. IDC foresees public clouds, and especially custom public clouds, supporting an increasing proportion of the aggregate HPC workload as these cloud facilities grow more capable and mature
  – Courtesy: http://www.idc.com/getdoc.jsp?containerId=247846
- Combining HPC with Cloud is still facing challenges because of the performance overhead associated virtualization support
  – Lower performance of virtualized I/O devices
- HPC Cloud Examples
  – **Amazon EC2 with Enhanced Networking**
    - Using Single Root I/O Virtualization (SR-IOV)
    - Higher performance (packets per second), lower latency, and lower jitter.
    - 10 GigE
  – **NSF Chameleon Cloud**

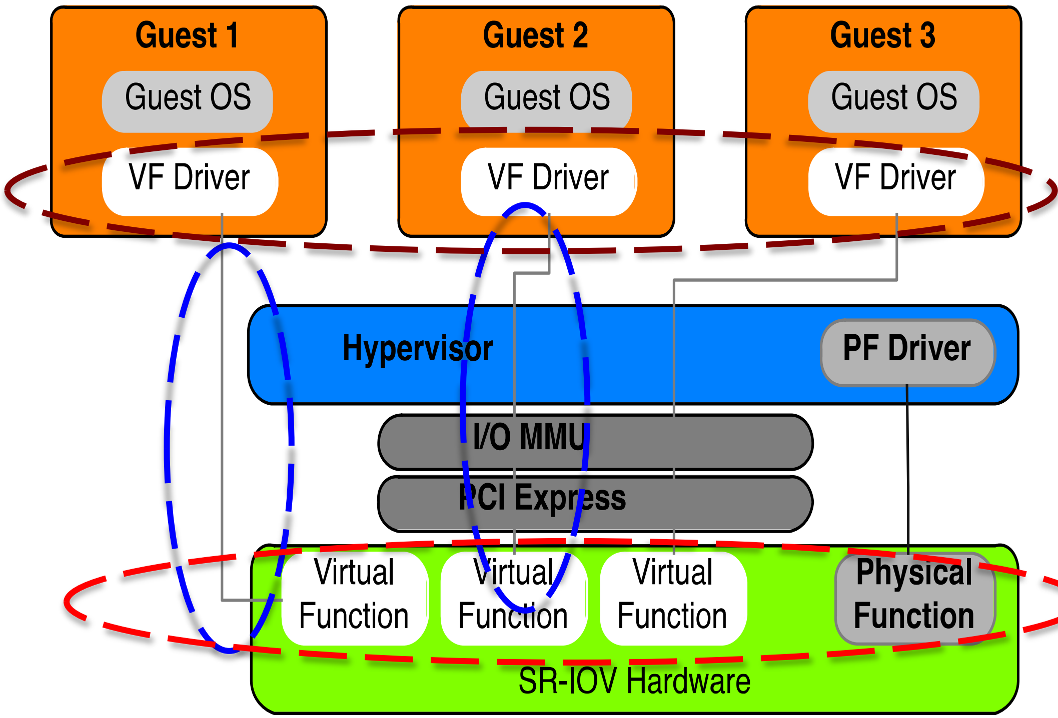
NSF Chameleon Cloud: A Powerful and Flexible Experimental Instrument

- Large-scale instrument
  - Targeting Big Data, Big Compute, Big Instrument research
  - ~650 nodes (~14,500 cores), 5 PB disk over two sites, 2 sites connected with 100G network
  - Virtualization technology (e.g., SR-IOV, accelerators), systems, networking (InfiniBand), infrastructure-level resource management, etc.
- Reconfigurable instrument
  - Bare metal reconfiguration, operated as single instrument, graduated approach for ease-of-use
- Connected instrument
  - Workload and Trace Archive
  - Partnerships with production clouds: CERN, OSDC, Rackspace, Google, and others
  - Partnerships with users
- Complementary instrument
  - Complementing GENI, Grid'5000, and other testbeds
- Sustainable instrument
  - Industry connections

http://www.chameleoncloud.org/

Single Root I/O Virtualization (SR-IOV)

- Single Root I/O Virtualization (SR-IOV) is providing new opportunities to design HPC cloud with very little low overhead
  - Allows a single physical device, or a Physical Function (PF), to present itself as multiple virtual devices, or Virtual Functions (VFs)
  - Each VF can be dedicated to a single VM through PCI pass-through
  - VFs are designed based on the existing non-virtualized PFs, no need for driver change
  - Work with 10/40 GigE and InfiniBand

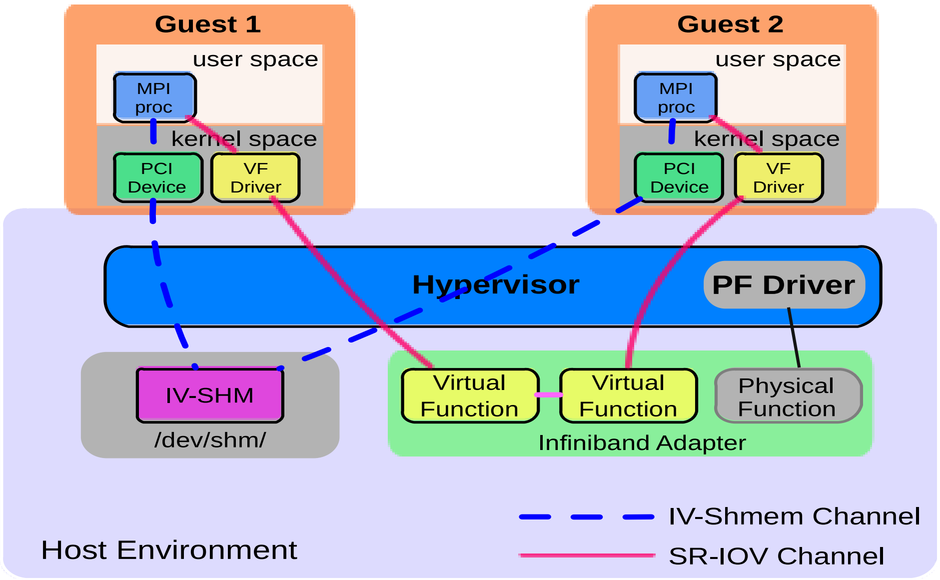


MVAPICH2-Virt: High-Performance MPI Library over SR-IOV capable InfiniBand Clusters

- Support for SR-IOV
  - Inter-node Inter-VM communication
- Locality-aware communication through IVSHMEM
  - Inter-VM Shared Memory (IVSHMEM) is a novel feature proposed for inter-VM communication, and offers shared memory backed communication for VMs within a given host
  - Intra-node Inter-VM communication
- Building efficient HPC Cloud
- Available publicly as MVAPICH2-Virt 2.1 Library

Overview of MVAPICH2-Virt with SR-IOV and IVSHMEM

- Redesign MVAPICH2 to make it virtual machine aware
  - SR-IOV shows near to native performance for inter-node point to point communication
  - IVSHMEM offers zero-copy access to data on shared memory of co-resident VMs
  - Locality Detector: maintains the locality information of co-resident virtual machines
  - Communication Coordinator: selects the communication channel (SR-IOV, IVSHMEM) adaptively

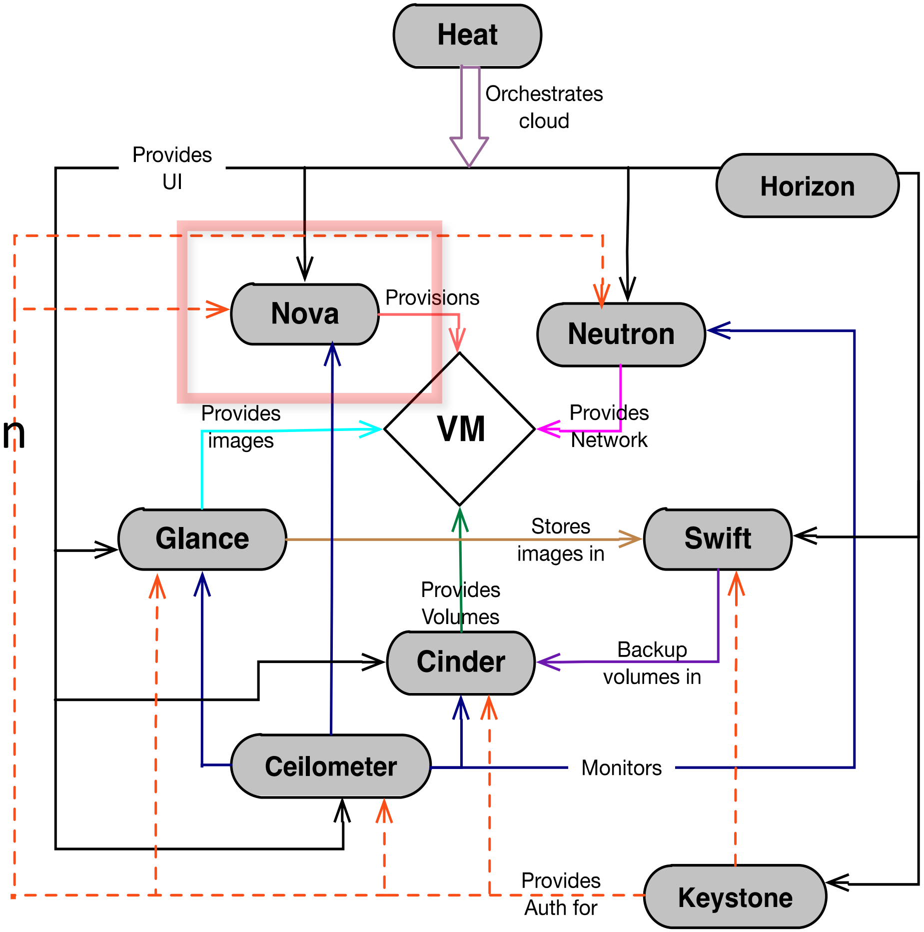


J. Zhang, X. Lu, J. Jose, R. Shi, D. K. Panda. Can Inter-VM Shmem Benefit MPI Applications on SR-IOV based Virtualized InfiniBand Clusters? **Euro-Par**, 2014.

J. Zhang, X. Lu, J. Jose, R. Shi, M. Li, D. K. Panda. High Performance MPI Library over SR-IOV Enabled InfiniBand Clusters. **HiPC**, 2014.

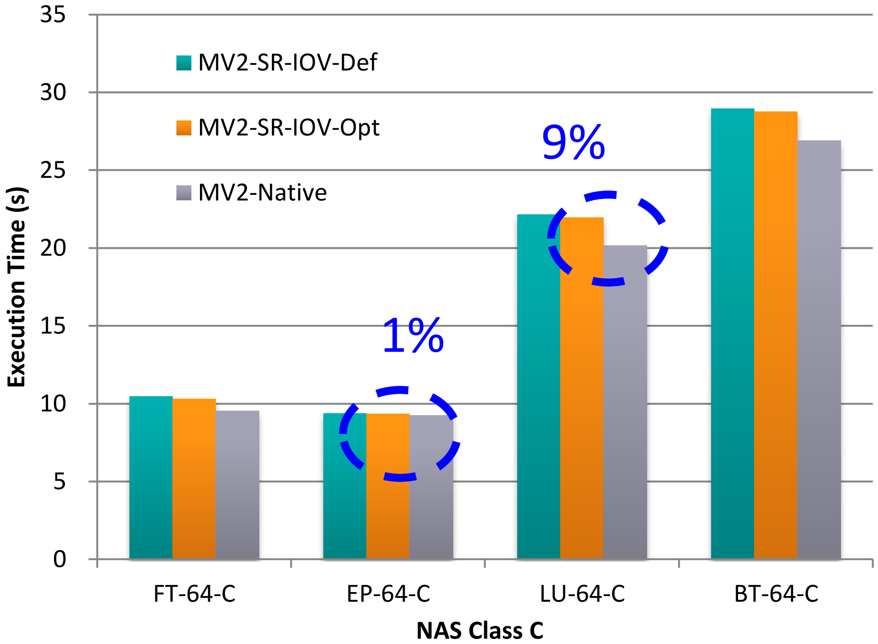
MVAPICH2-Virt with SR-IOV and IVSHMEM over OpenStack

- OpenStack is one of the most popular open-source solutions to build clouds and manage virtual machines
- Deployment with OpenStack
  - Supporting SR-IOV configuration
  - Supporting IVSHMEM configuration
  - Virtual Machine aware design of MVAPICH2 with SR-IOV
- An efficient approach to build HPC Clouds with MVAPICH2-Virt and OpenStack

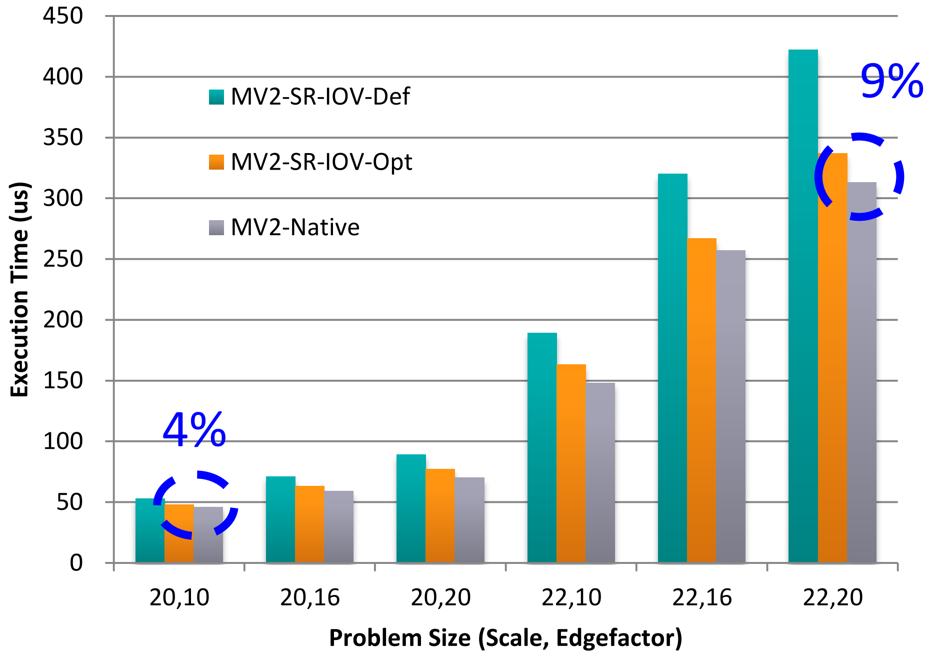


J. Zhang, X. Lu, M. Arnold, D. K. Panda. MVAPICH2 over OpenStack with SR-IOV: An Efficient Approach to Build HPC Clouds. **CCGrid**, 2015.

Application-Level Performance (8 VM * 8 Core/VM)



NAS



Graph500

- Compared to Native, 1-9% overhead for NAS
- Compared to Native, 4-9% overhead for Graph500

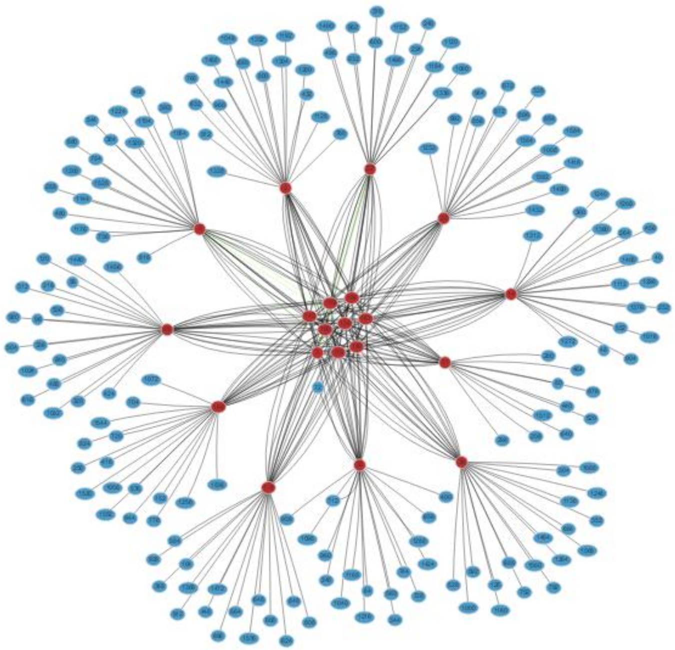
Overview of A Few Challenges being Addressed by MVAPICH2 Project for Exascale

- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- MVAPICH2-Virt
  - Support for Basic SR-IOV
  - Locality-aware communication
  - Building HPC Cloud
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool

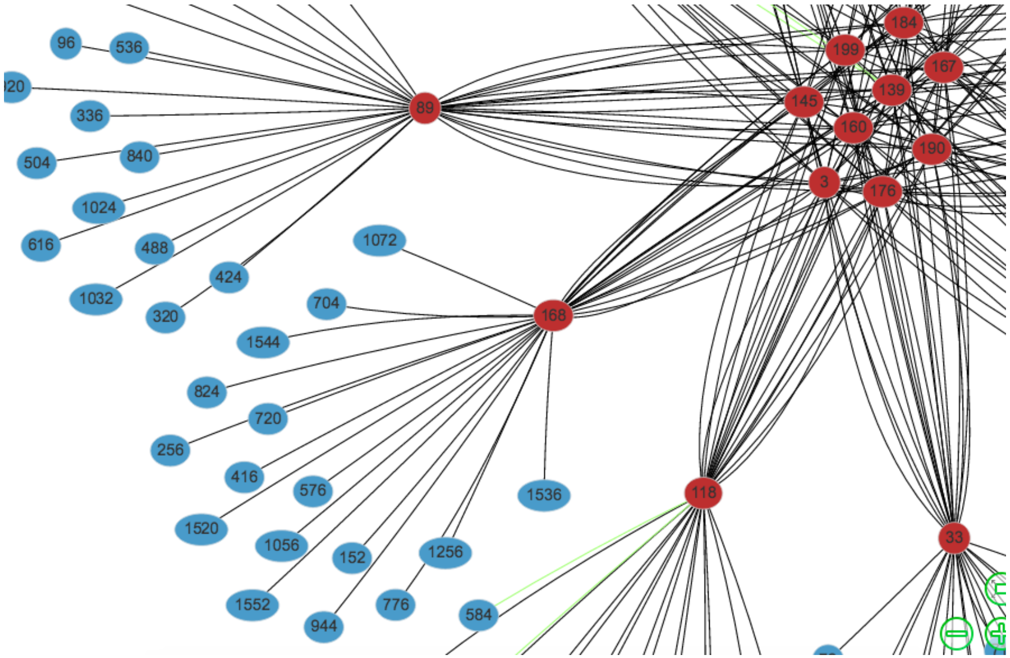
Overview of OSU INAM

- OSU INAM monitors IB clusters in real time by querying various subnet management entities in the network
- 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 (pt-to-pt, collectives and RMA)
  - Capability to profile and report the following parameters of MPI processes at node-level, job-level and process-level at user specified granularity in conjunction with MVAPICH2-X 2.2b
    - CPU utilization
    - Memory utilization
    - Inter-node communication buffer usage for RC transport
    - Inter-node communication buffer usage for UD transport
  - Improve network load time by clustering individual nodes
  - Introduce "Job Page" to display jobs in ascending/descending order of various performance metrics in conjunction with MVAPICH2-X 2.2b
  - Visualize the data transfer happening in a "live" fashion - Live View for Entire Network, Particular Job and One or multiple Nodes
  - Capability to visualize data transfer that happened in the network at a time duration in the past - Historical View for Entire Network, Particular Job and One or multiple Nodes

OSU InfiniBand Network Analysis Monitoring Tool (INAM) – Network Level View



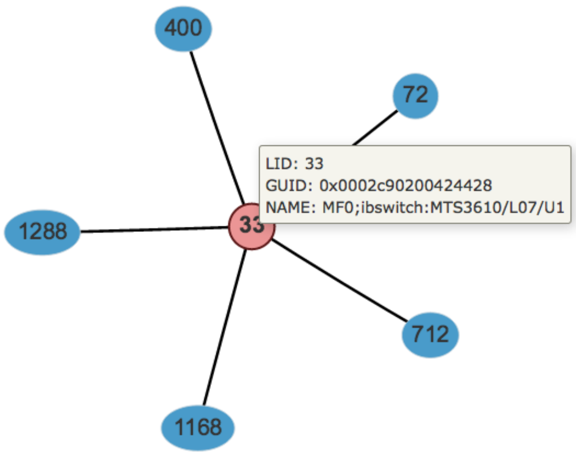
Full Network (152 nodes)



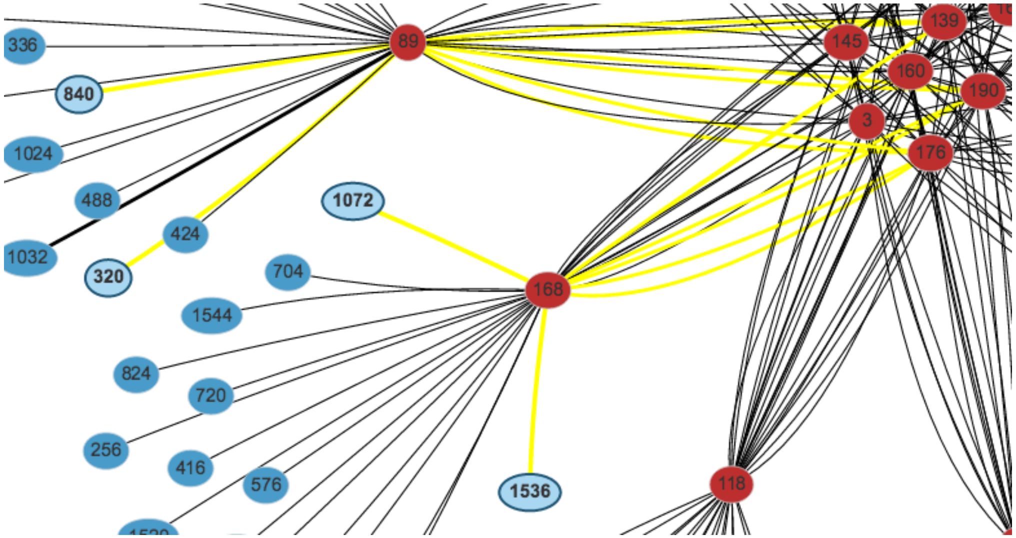
Zoomed-in View of the Network

- Show network topology of large clusters
- 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 Tool – Job and Node Level Views



Visualizing a Job (5 Nodes)



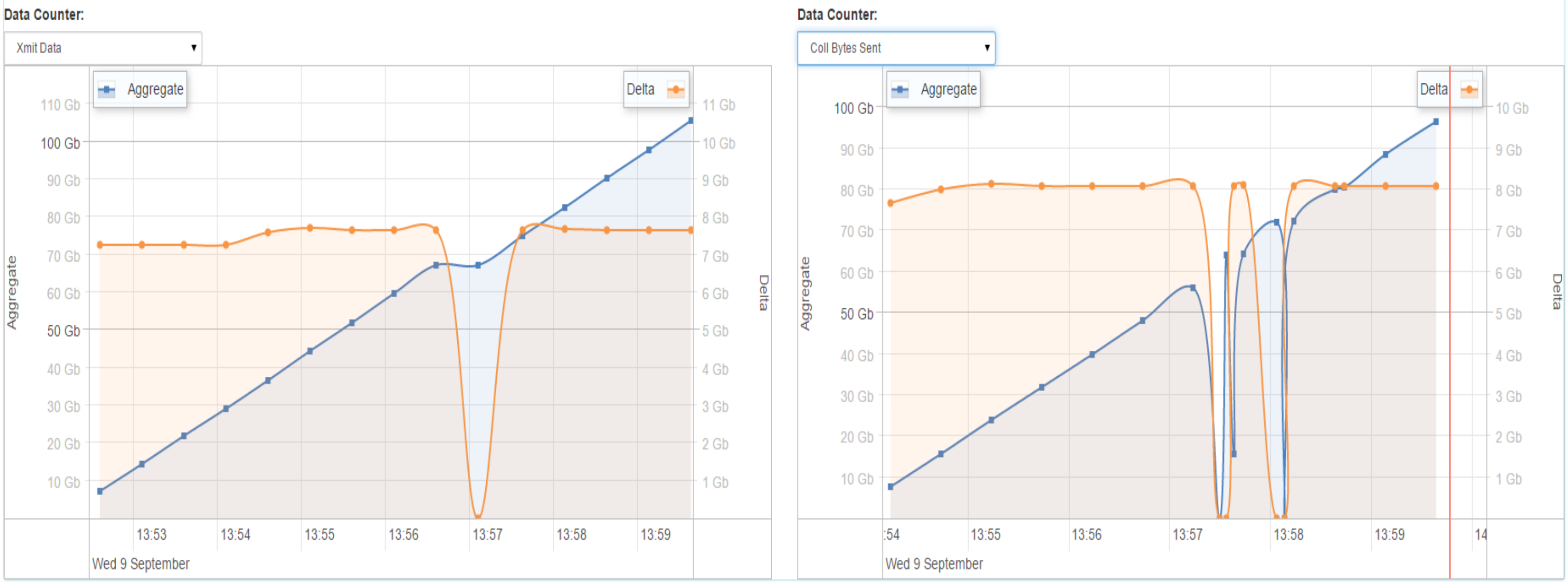
Finding Routes Between Nodes

- 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 provides details per process or per node
  - CPU 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

OSU INAM Tool – Live Node Level View

Rank112 [core 0]

Rank113 [core 1]



Rank114 [core 2]

Rank115 [core 3]

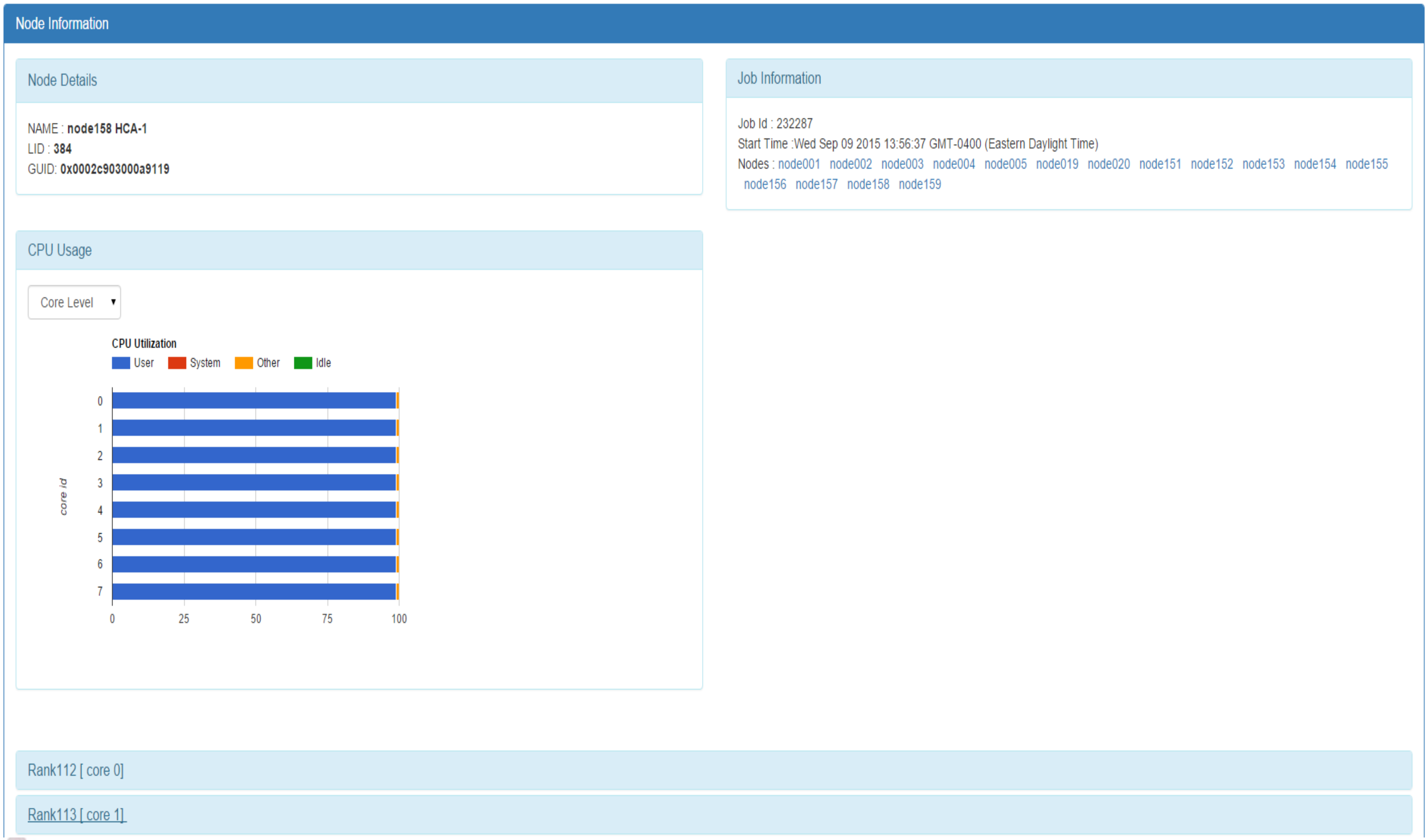
Rank116 [core 4]

Rank117 [core 5]

Rank118 [core 6]

Rank119 [core 7]

OSU INAM Tool – Live Node Level View (Cont.)



MVAPICH2 – Plans for Exascale

- Performance and Memory scalability toward 500K-1M cores
  - Dynamically Connected Transport (DCT) service with Connect-IB
- Hybrid programming (MPI + OpenSHMEM, MPI + UPC, MPI + CAF …)
  - Support for UPC++
- Enhanced Optimization for GPU Support and Accelerators
- Taking advantage of advanced features
  - User Mode Memory Registration (UMR)
  - On-demand Paging
- Enhanced Inter-node and Intra-node communication schemes for upcoming OmniPath enabled Knights Landing architectures
- Extended RMA support (as in MPI 3.0)
- Extended topology-aware collectives
- Energy-aware point-to-point (one-sided and two-sided) and collectives
- Extended Support for MPI Tools Interface (as in MPI 3.0)
- Extended Checkpoint-Restart and migration support with SCR
- Energy Awareness

One Additional Talk

- Thursday (10:00-10:30am)
  - How to Exploit MPI, PGAS and Hybrid MPI+PGAS Programming through MVAPICH2-X?

Funding Acknowledgments

Funding Support by

- Office of Science, U.S. Department of Energy
- NSF
- Mellanox Technologies
- Cisco Systems
- QLogic
- Cray – The Supercomputer Company
- intel
- Linux Networx
- NVIDIA
- NetApp
- Sun microsystems

Equipment Support by

- Mellanox Technologies
- intel
- AMD
- Sun microsystems
- APPRO
- advanced clustering technologies, inc.
- Microway – Technology you can count on
- NVIDIA
- QLogic

Personnel Acknowledgments

Current Students

- A. Augustine (M.S.)
- A. Awan (Ph.D.)
- S. Chakraborthy (Ph.D.)
- C.-H. Chu (Ph.D.)
- N. Islam (Ph.D.)
- M. Li (Ph.D.)
- K. Kulkarni (M.S.)
- M. Rahman (Ph.D.)
- D. Shankar (Ph.D.)
- A. Venkatesh (Ph.D.)
- J. Zhang (Ph.D.)

Current Research Scientists

- H. Subramoni
- X. Lu

Current Senior Research Associate

- K. Hamidouche

Current Post-Doc

- J. Lin
- D. Banerjee

Current Programmer

- J. Perkins

Current Research Specialist

- M. Arnold

Past Students

- P. Balaji (Ph.D.)
- S. Bhagvat (M.S.)
- A. Bhat (M.S.)
- D. Buntinas (Ph.D.)
- L. Chai (Ph.D.)
- B. Chandrasekharan (M.S.)
- N. Dandapanthula (M.S.)
- V. Dhanraj (M.S.)
- T. Gangadharappa (M.S.)
- K. Gopalakrishnan (M.S.)
- W. Huang (Ph.D.)
- W. Jiang (M.S.)
- J. Jose (Ph.D.)
- S. Kini (M.S.)
- M. Koop (Ph.D.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- K. Kandalla (Ph.D.)
- P. Lai (M.S.)
- J. Liu (Ph.D.)
- M. Luo (Ph.D.)
- A. Mamidala (Ph.D.)
- G. Marsh (M.S.)
- V. Meshram (M.S.)
- A. Moody (M.S.)
- S. Naravula (Ph.D.)
- R. Noronha (Ph.D.)
- X. Ouyang (Ph.D.)
- S. Pai (M.S.)
- S. Potluri (Ph.D.)
- R. Rajachandrasekar (Ph.D.)
- G. Santhanaraman (Ph.D.)
- A. Singh (Ph.D.)
- J. Sridhar (M.S.)
- S. Sur (Ph.D.)
- H. Subramoni (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- A. Vishnu (Ph.D.)
- J. Wu (Ph.D.)
- W. Yu (Ph.D.)

Past Post-Docs

- H. Wang
- X. Besseron
- H.-W. Jin
- M. Luo
- E. Mancini
- S. Marcarelli
- J. Vienne

Past Research Scientist

- S. Sur

Past Programmers

- D. Bureddy

Web Pointers

http://www.cse.ohio-state.edu/~panda

http://www.cse.ohio-state.edu/~subramon

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

MVAPICH Web Page

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

MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

panda@cse.ohio-state.edu

subramon@cse.ohio-state.edu