MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

The MVAPICH2 Project: Latest Developments and Plans Towards Exascale Computing

Presentation at Mellanox Theatre (SC '16)

by

Dhabaleswar K. (DK) Panda

The Ohio State University

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

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

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)

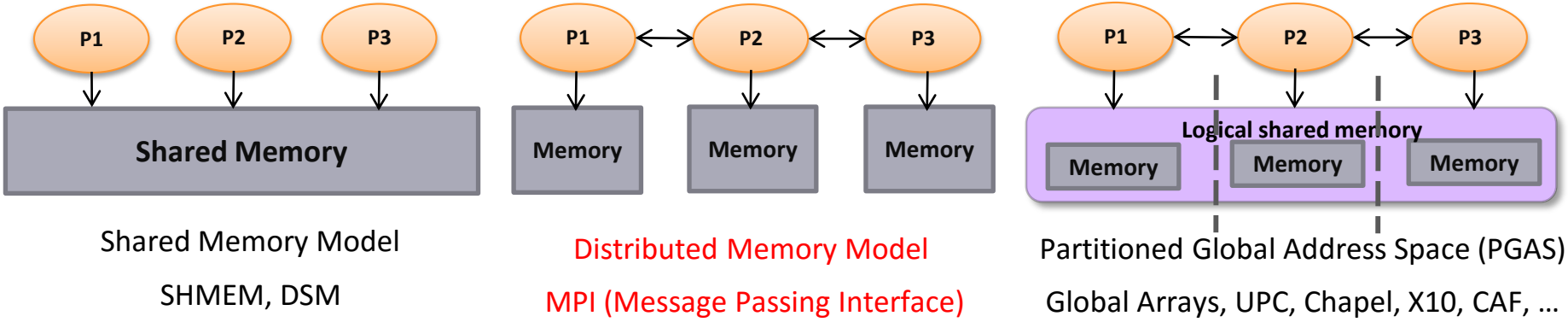
Tianhe – 2

Titan

Stampede

Tianhe – 1A

Parallel Programming Models Overview



Shared Memory Model

SHMEM, DSM

Distributed Memory Model

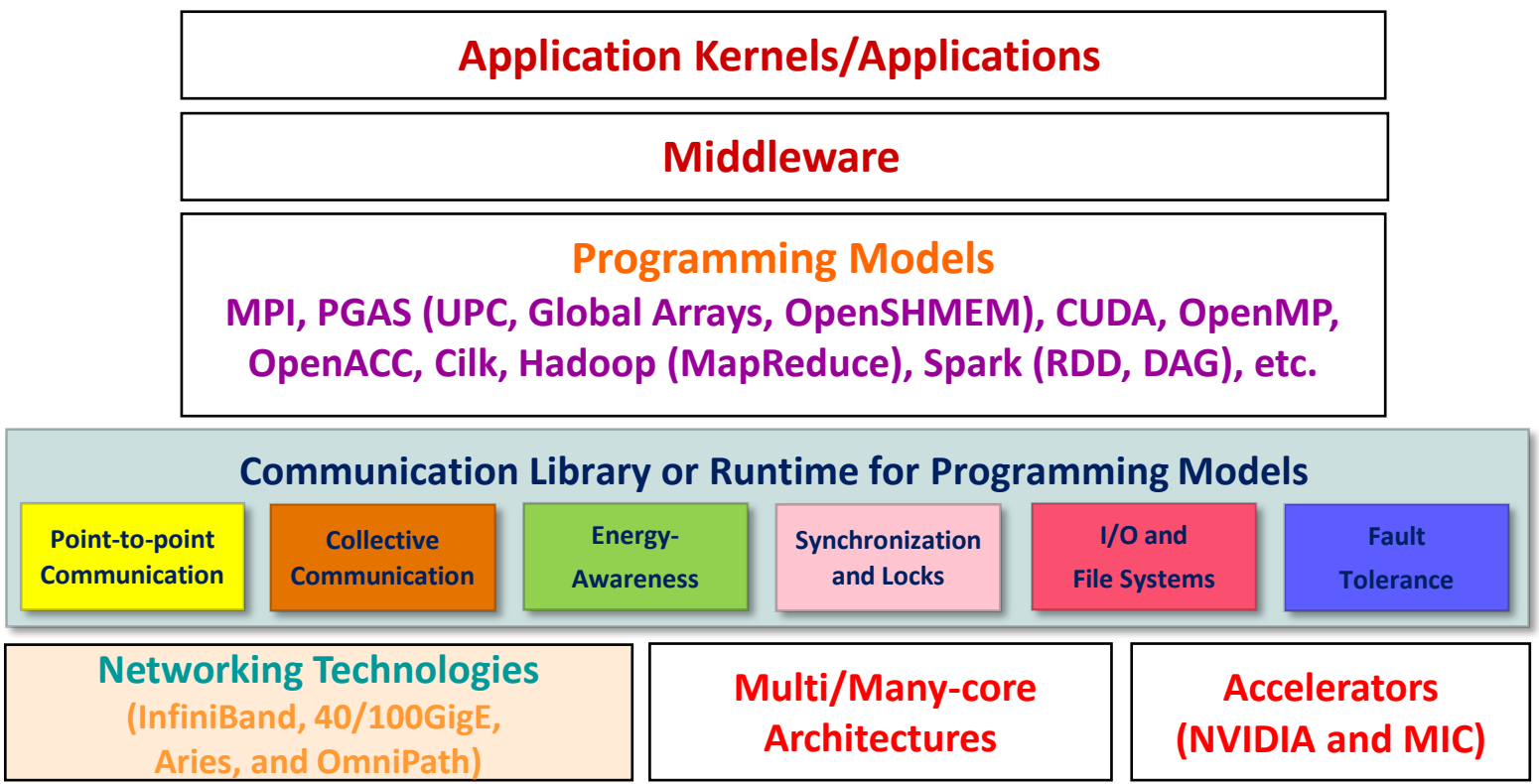
MPI (Message Passing Interface)

Partitioned Global Address Space (PGAS)

Global Arrays, UPC, Chapel, X10, CAF, …

- Programming models provide abstract machine models
- Models can be mapped on different types of systems
  - e.g. Distributed Shared Memory (DSM), MPI within a node, etc.
- PGAS models and Hybrid MPI+PGAS models are gradually receiving importance

Designing Communication Libraries for Multi-Petaflop and Exaflop Systems: Challenges



Co-Design Opportunities and Challenges across Various Layers

Performance

Scalability

Fault-Resilience

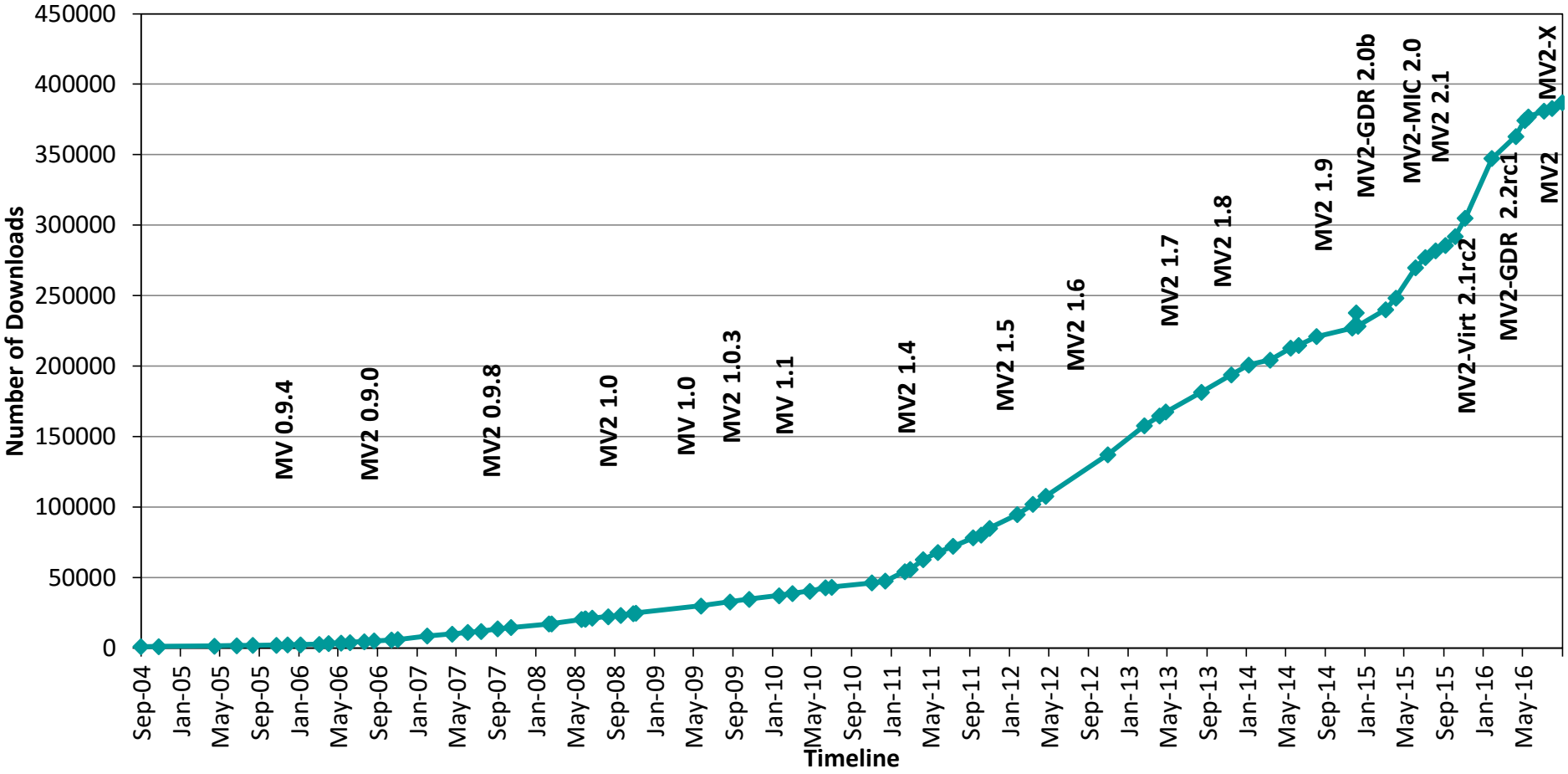
Designing (MPI+X) at Exascale

- 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

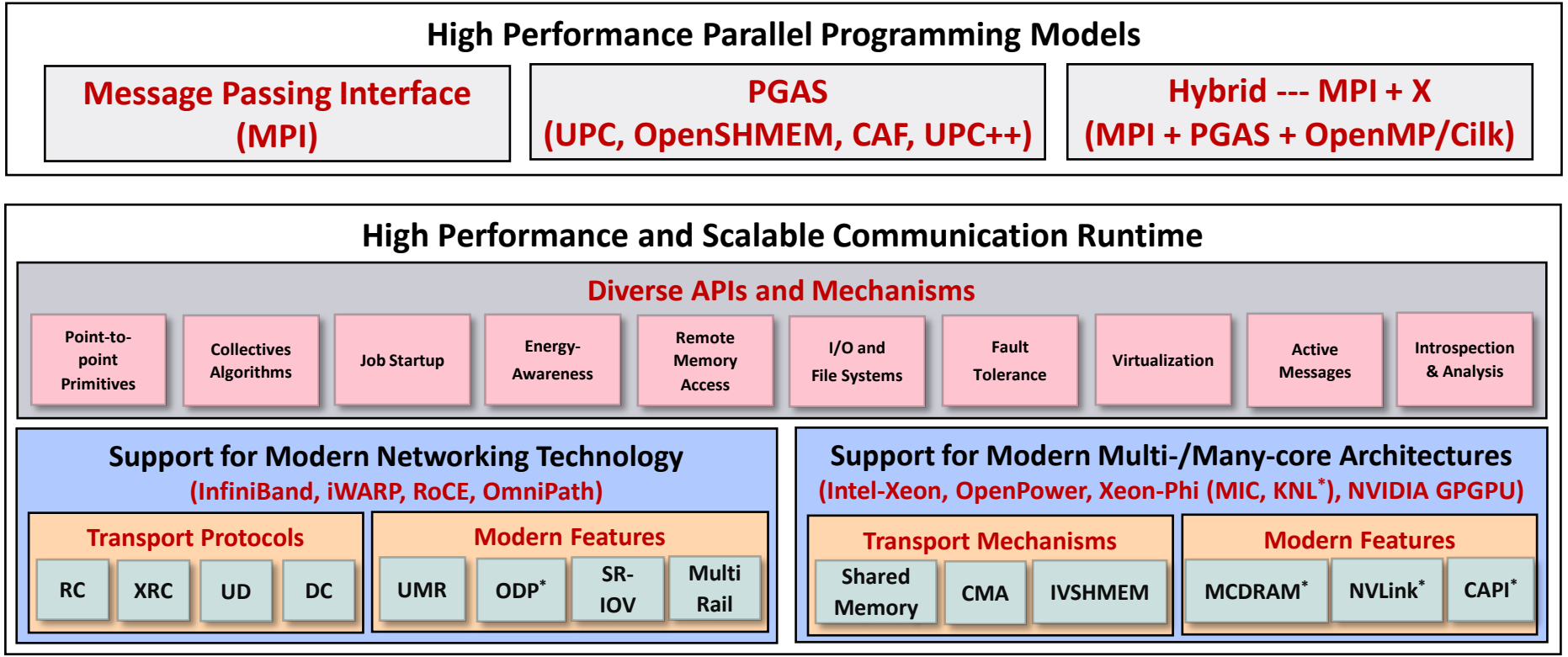
Overview of the MVAPICH2 Project

- 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 2,690 organizations in 83 countries**
  - **More than 402,000 (> 0.4 million) downloads from the OSU site directly**
  - Empowering many TOP500 clusters (Nov '16 ranking)
    - **1st ranked 10,649,640-core cluster (Sunway TaihuLight) at NSC, Wuxi, China**
    - 13th ranked 241,108-core cluster (Pleiades) at NASA
    - 17th ranked 519,640-core cluster (Stampede) at TACC
    - 40th 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) ->
    Sunway TaihuLight at NSC, Wuxi, China (1st in Nov'16, 10,649,640 cores, 93 PFlops)

MVAPICH/MVAPICH2 Release Timeline and Downloads



MVAPICH2 Architecture



* Upcoming

MVAPICH2 Software Family

High-Performance Parallel Programming Libraries	
MVAPICH2	Support for InfiniBand, Omni-Path, Ethernet/iWARP, and RoCE
MVAPICH2-X	Advanced MPI features, OSU INAM, PGAS (OpenSHMEM, UPC, UPC++, and CAF), and MPI+PGAS programming models with unified communication runtime
MVAPICH2-GDR	Optimized MPI for clusters with NVIDIA GPUs
MVAPICH2-Virt	High-performance and scalable MPI for hypervisor and container based HPC cloud
MVAPICH2-EA	Energy aware and High-performance MPI
MVAPICH2-MIC	Optimized MPI for clusters with Intel KNC
Microbenchmarks	
OMB	Microbenchmarks suite to evaluate MPI and PGAS (OpenSHMEM, UPC, and UPC++) libraries for CPUs and GPUs
Tools	
OSU INAM	Network monitoring, profiling, and analysis for clusters with MPI and scheduler integration
OEMT	Utility to measure the energy consumption of MPI applications

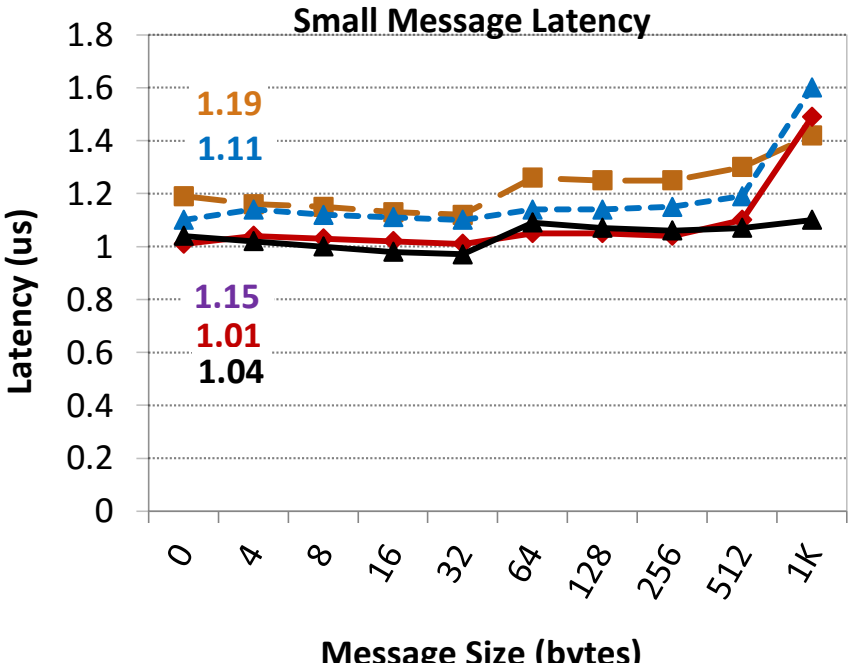
MVAPICH2 Distributions

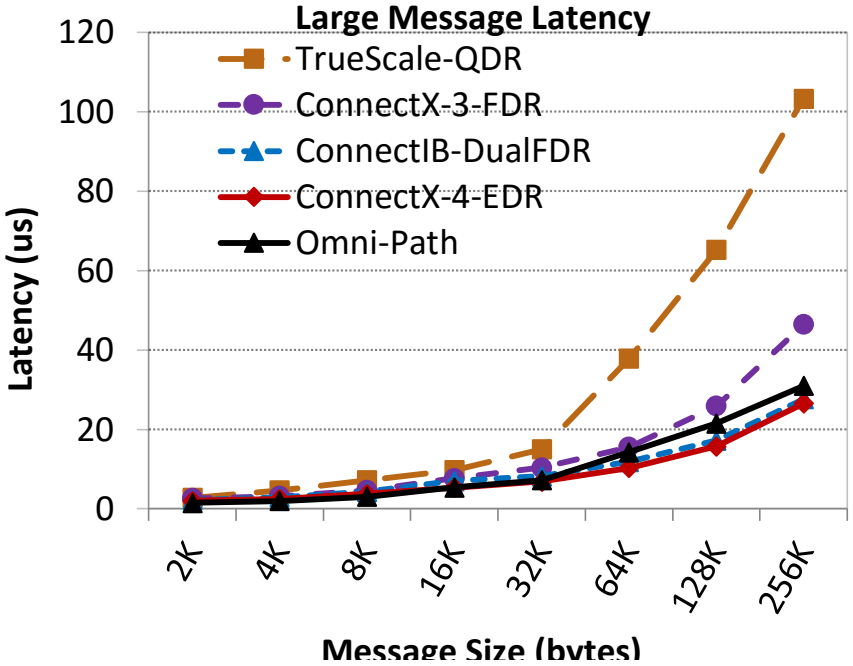
- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE
  - Basic GPU (CUDA-aware MPI) support
- MVAPICH2-X
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - with and without Open Stack
- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- OSU Micro-Benchmarks (OMB)
  - MPI (including CUDA-aware MPI), OpenSHMEM and UPC
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool
- MVAPIC2-MIC (KNL)
  - Optimized for IB clusters with Intel Xeon Phis
- MVAPICH2-GDR and Deep Learning (Will be presented on Thursday at 11:00am)

MVAPICH2 2.2 GA

- Released on 09/09/2016
- Major Features and Enhancements
  - **Support and optimization for OpenPower, KNL, Omni-Path , PSM2, Mellanox EDR**
  - Enhanced performance for MPI_Comm_split through new bitonic algorithm
  - Enable graceful fallback to Shared Memory if LiMIC2 or CMA transfer fails
  - Enable support for multiple MPI initializations
  - Remove verbs dependency when building the PSM and PSM2 channels
  - Allow processes to request MPI_THREAD_MULTIPLE when socket or NUMA node level affinity is specified
  - Collective tuning for Opal@LLNL, Bridges@PSC, and Stampede-1.5@TACC
  - Tuning and architecture detection for Intel Broadwell processors
  - Warn user to reconfigure library if rank type is not large enough to represent all ranks in job
  - Unify process affinity support in Gen2, PSM and PSM2 channels

One-way Latency: MPI over IB with MVAPICH2





TrueScale-QDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

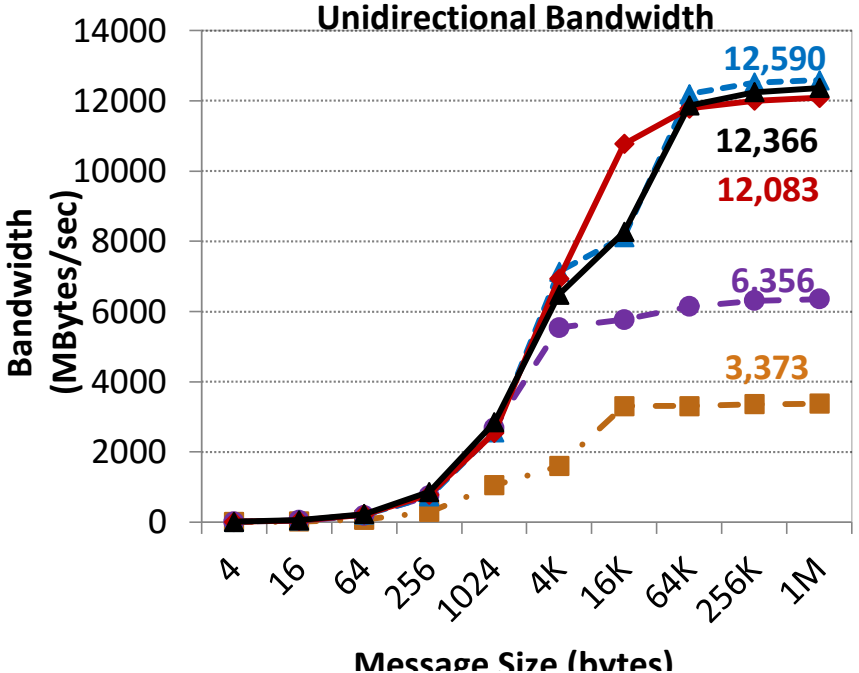
ConnectX-3-FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch

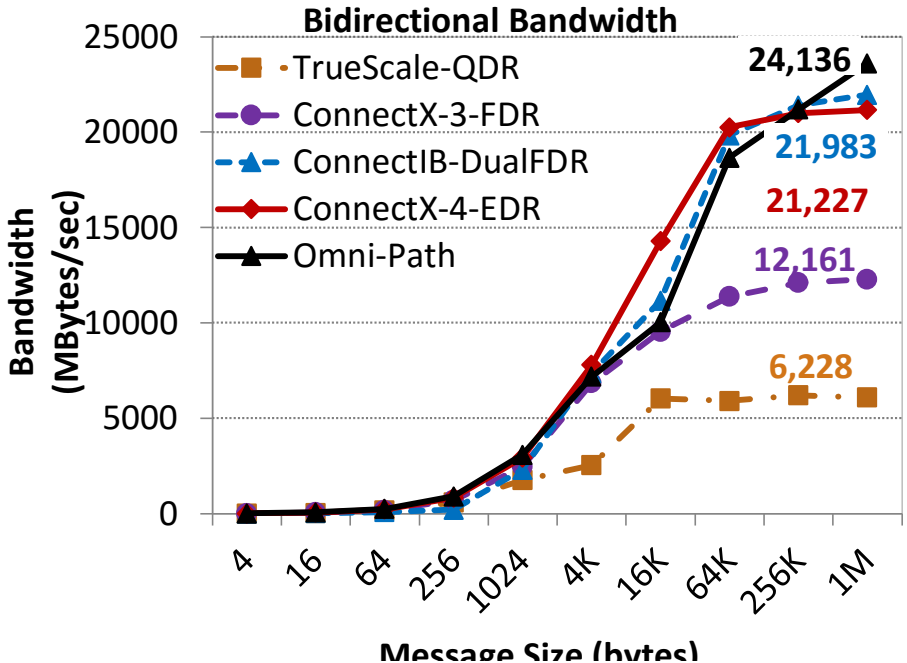
ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

ConnectX-4-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB Switch

Omni-Path - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with Omni-Path switch

Bandwidth: MPI over IB with MVAPICH2





TrueScale-QDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

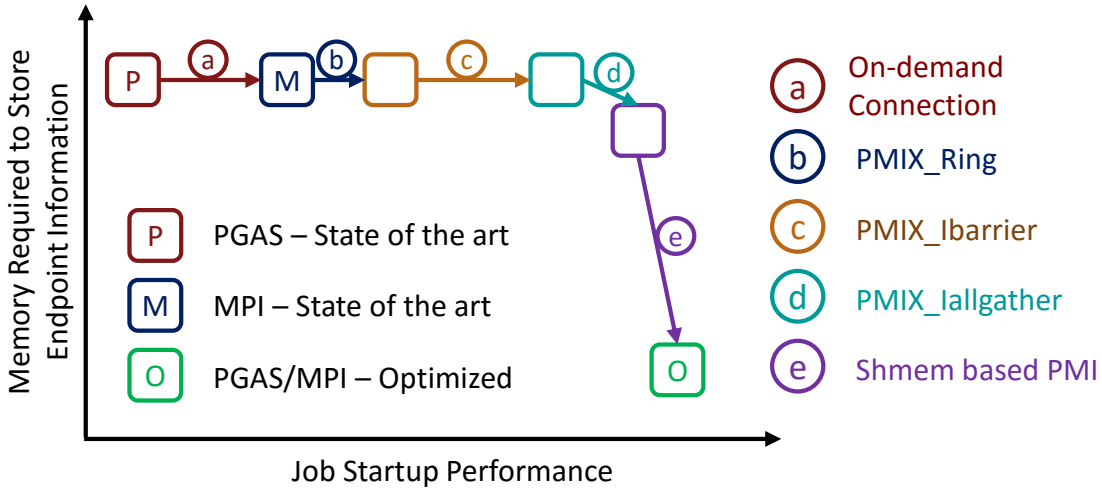
ConnectX-3-FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch

ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch

ConnectX-4-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 IB switch

Omni-Path - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with Omni-Path switch

Towards High Performance and Scalable Startup at Exascale



- Near-constant MPI and OpenSHMEM initialization time at any process count
- 10x and 30x improvement in startup time of MPI and OpenSHMEM respectively at 16,384 processes
- Memory consumption reduced for remote endpoint information by O(processes per node)
- 1GB Memory saved per node with 1M processes and 16 processes per node

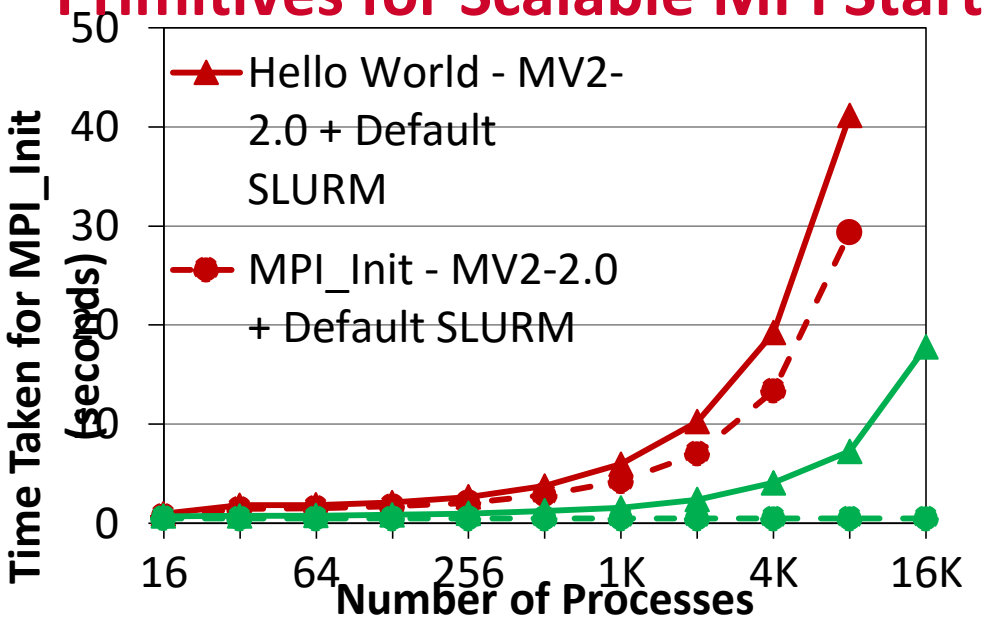
(a) **On-demand Connection Management for OpenSHMEM and OpenSHMEM+MPI.** S. Chakraborty, H. Subramoni, J. Perkins, A. A. Awan, and D K Panda, 20th International Workshop on High-level Parallel Programming Models and Supportive Environments (HIPS '15)

(b) **PMI Extensions for Scalable MPI Startup.** S. Chakraborty, H. Subramoni, A. Moody, J. Perkins, M. Arnold, and D K Panda, Proceedings of the 21st European MPI Users' Group Meeting (EuroMPI/Asia '14)

(c) (d) **Non-blocking PMI Extensions for Fast MPI Startup.** S. Chakraborty, H. Subramoni, A. Moody, A. Venkatesh, J. Perkins, and D K Panda, 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid '15)

(e) **SHMEMPMI – Shared Memory based PMI for Improved Performance and Scalability.** S. Chakraborty, H. Subramoni, J. Perkins, and D K Panda, 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid '16)

Non-blocking Process Management Interface (PMI) Primitives for Scalable MPI Startup



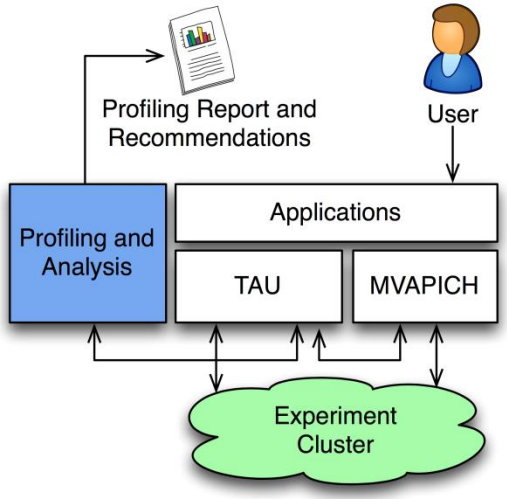
- Address exchange over PMI is the major bottleneck in job startup
- Non-blocking PMI exchange hides this cost by overlapping it with application initialization and computation
- New PMI operation PMIX_Allgather for improved symmetric data transfer
- Near-constant MPI_Init at any scale
- MPI_Init is 59 times faster at 8,192 processes (512 nodes)
- Hello World (MPI_Init + MPI_Finalize) takes 5.7 times less time at 8,192 processes

Available since MVAPICH2-2.1 and as patch for SLURM-15.08.8 and SLURM-16.05.1

More Details in Student Research Poster Presentation (5:15-7:00pm today)

Performance Engineering Applications using MVAPICH2 and TAU

- Enhance existing support for MPI_T in MVAPICH2 to expose a richer set of performance and control variables
- Get and display MPI Performance Variables (PVARs) made available by the runtime in TAU
- Control the runtime's behavior via MPI Control Variables (CVARs)
- Introduced support for new MPI_T based CVARs to MVAPICH2
  - MPIR_CVAR_MAX_INLINE_MSG_SZ, MPIR_CVAR_VBUF_POOL_SIZE, MPIR_CVAR_VBUF_SECONDARY_POOL_SIZE
- TAU enhanced with support for setting MPI_T CVARs in a non-interactive mode for uninstrumented applications



VBUF usage without CVAR based tuning as displayed by ParaProf

Name	MaxValue	MinValue	MeanValue	Std. Dev.	NumSamples	Total
mv2_total_vbuf_memory (Total amount of memory in bytes used for VBUFs)	3,313,056	3,313,056	3,313,056	0	1	3,313,056
mv2_ud_vbuf_allocated (Number of UD VBUFs allocated)	0	0	0	0	0	0
mv2_ud_vbuf_available (Number of UD VBUFs available)	0	0	0	0	0	0
mv2_ud_vbuf_freed (Number of UD VBUFs freed)	0	0	0	0	0	0
mv2_ud_vbuf_inuse (Number of UD VBUFs inuse)	0	0	0	0	0	0
mv2_ud_vbuf_max_use (Maximum number of UD VBUFs used)	0	0	0	0	0	0
mv2_vbuf_allocated (Number of VBUFs allocated)	320	320	320	0	1	320
mv2_vbuf_available (Number of VBUFs available)	255	255	255	0	1	255
mv2_vbuf_freed (Number of VBUFs freed)	25,545	25,545	25,545	0	1	25,545
mv2_vbuf_inuse (Number of VBUFs inuse)	65	65	65	0	1	65
mv2_vbuf_max_use (Maximum number of VBUFs used)	65	65	65	0	1	65
num_calloc_calls (Number of MPIT_calloc calls)	89	89	89	0	1	89

VBUF usage with CVAR based tuning as displayed by ParaProf

Name	MaxValue	MinValue	MeanValue	Std. Dev.	NumSamp...	Total
mv2_total_vbuf_memory (Total amount of memory in bytes used for VBUFs)	1,815,056	1,815,056	1,815,056	0	1	1,815,056
mv2_ud_vbuf_allocated (Number of UD VBUFs allocated)	0	0	0	0	0	0
mv2_ud_vbuf_available (Number of UD VBUFs available)	0	0	0	0	0	0
mv2_ud_vbuf_freed (Number of UD VBUFs freed)	0	0	0	0	0	0
mv2_ud_vbuf_inuse (Number of UD VBUFs inuse)	0	0	0	0	0	0
mv2_ud_vbuf_max_use (Maximum number of UD VBUFs used)	0	0	0	0	0	0
mv2_vbuf_allocated (Number of VBUFs allocated)	160	160	160	0	1	160
mv2_vbuf_available (Number of VBUFs available)	94	94	94	0	1	94
mv2_vbuf_freed (Number of VBUFs freed)	5,479	5,479	5,479	0	1	5,479
mv2_vbuf_inuse (Number of VBUFs inuse)	66	66	66	0	1	66

Dynamic and Adaptive Tag Matching

Challenge

Tag matching is a significant overhead for receivers

Existing Solutions are

- Static and do not adapt dynamically to communication pattern
- Do not consider memory overhead

Solution

A new tag matching design

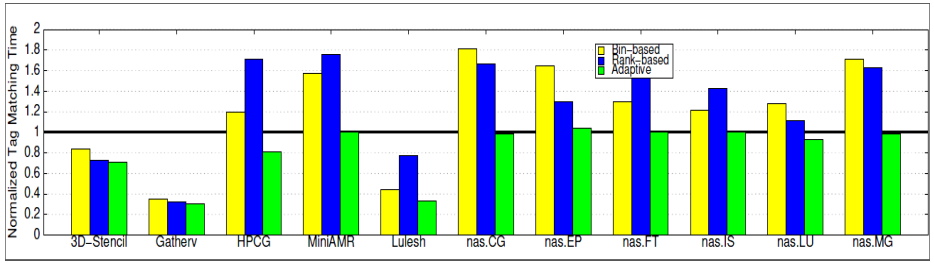
- Dynamically adapt to communication patterns
- Use different strategies for different ranks
- Decisions are based on the number of request object that must be traversed before hitting on the required one

Results

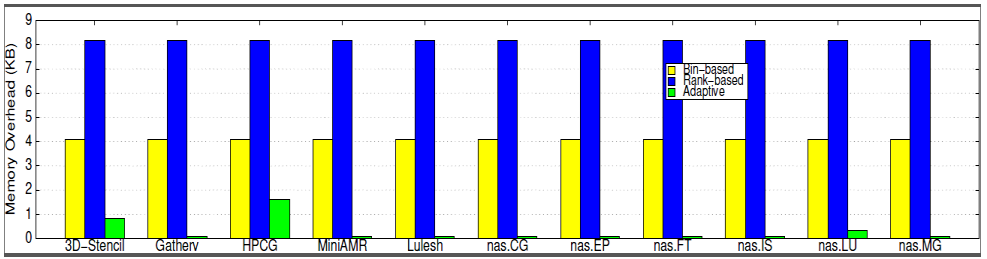
Better performance than other state-of-the art tag-matching schemes

Minimum memory consumption

Will be available in future MVAPICH2 releases



Normalized Total Tag Matching Time at 512 Processes
Normalized to Default (Lower is Better)



Normalized Memory Overhead per Process at 512 Processes
Compared to Default (Lower is Better)

Adaptive and Dynamic Design for MPI Tag Matching; M. Bayatpour, H. Subramoni, S. Chakraborty, and D. K. Panda; IEEE Cluster 2016. [Best Paper Nominee]

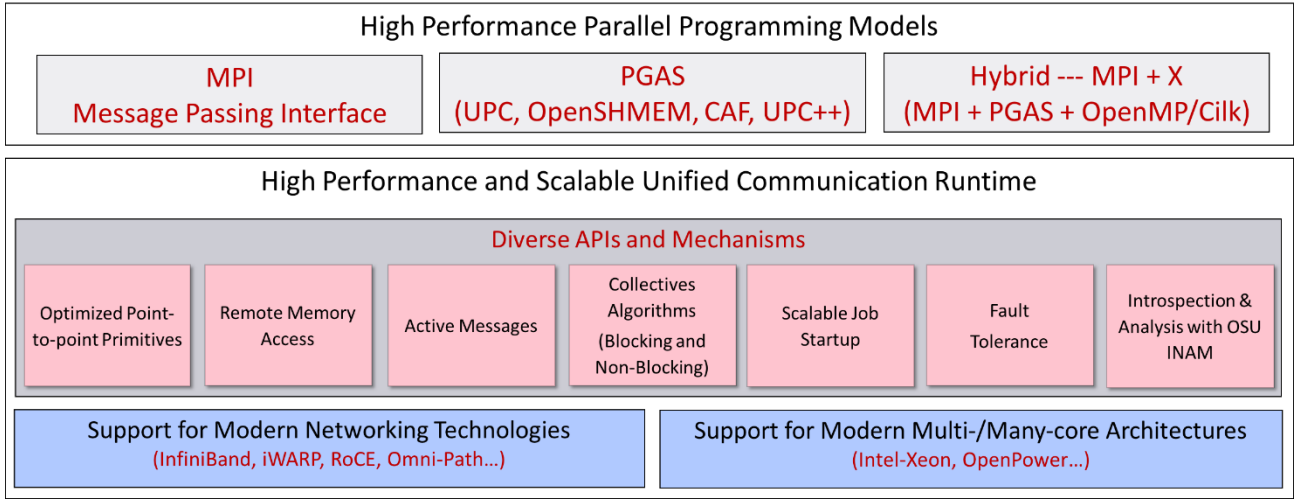
MVAPICH2 Distributions

- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE
  - Basic GPU (CUDA-aware MPI) support
- MVAPICH2-X
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - with and without Open Stack
- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- OSU Micro-Benchmarks (OMB)
  - MPI (including CUDA-aware MPI), OpenSHMEM and UPC
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool
- MVAPIC2-MIC (KNL)
  - Optimized for IB clusters with Intel Xeon Phis
- MVAPICH2-GDR and Deep Learning (Will be presented on Thursday at 11:00am)

MVAPICH2-X 2.2 GA

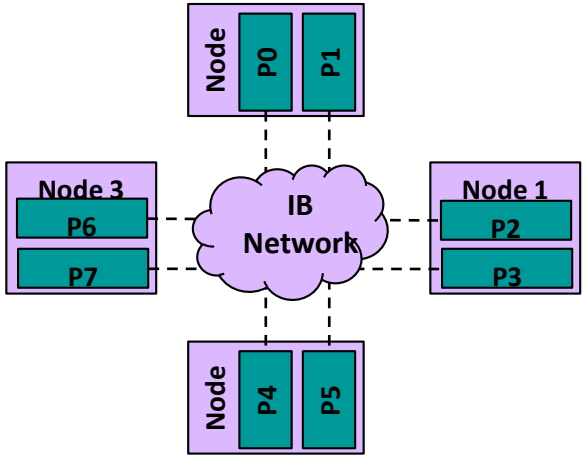
- Released on 09/09/2016
- Major Features and Enhancements
  - MPI Features
    - Based on MVAPICH2 2.2 GA (OFA-IB-CH3 interface)
    - Efficient support for Unified Memory Registration (UMR) and On Demand Paging (ODP) feature of Mellanox for point-to-point and RMA operations
    - Support for Intel Knights Landing and OpenPower architectures
  - UPC Features
    - Support for Intel Knights Landing and OpenPower architectures
  - UPC++ Features
    - Support for Intel Knights Landing and OpenPower architectures
  - OpenSHMEM Features
    - Support for Intel Knights Landing and OpenPower architectures
  - CAF Features
    - Support for Intel Knights Landing and OpenPower architectures
  - Hybrid Program Features
    - Support Intel Knights Landing and OpenPower architectures for hybrid MPI+PGAS applications
  - Unified Runtime Features
    - Based on MVAPICH2 2.2 GA (OFA-IB-CH3 interface). All the runtime features enabled by default in OFA-IB-CH3 and OFA-IB-RoCE

MVAPICH2-X for Hybrid MPI + PGAS Applications



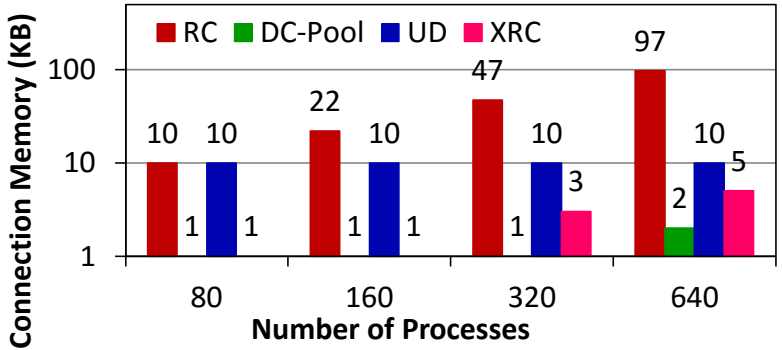
- Current Model – Separate Runtimes for OpenSHMEM/UPC/UPC++/CAF and MPI
  - Possible deadlock if both runtimes are not progressed
  - Consumes more network resource
- Unified communication runtime for MPI, UPC, UPC++, OpenSHMEM, CAF
  - Available with since 2012 (starting with MVAPICH2-X 1.9)
  - http://mvapich.cse.ohio-state.edu

Minimizing Memory Footprint by Direct Connect (DC) Transport

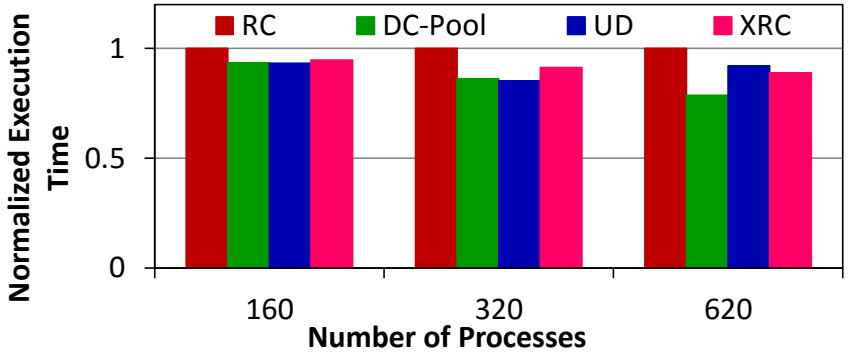


- Constant connection cost (*One QP for any peer*)
- Full Feature Set (RDMA, Atomics etc)
- Separate objects for send (DC Initiator) and receive (DC Target)
  - DC Target identified by "DCT Number"
  - Messages routed with (DCT Number, LID)
  - Requires same "DC Key" to enable communication
- Available since MVAPICH2-X 2.2a

Memory Footprint for Alltoall



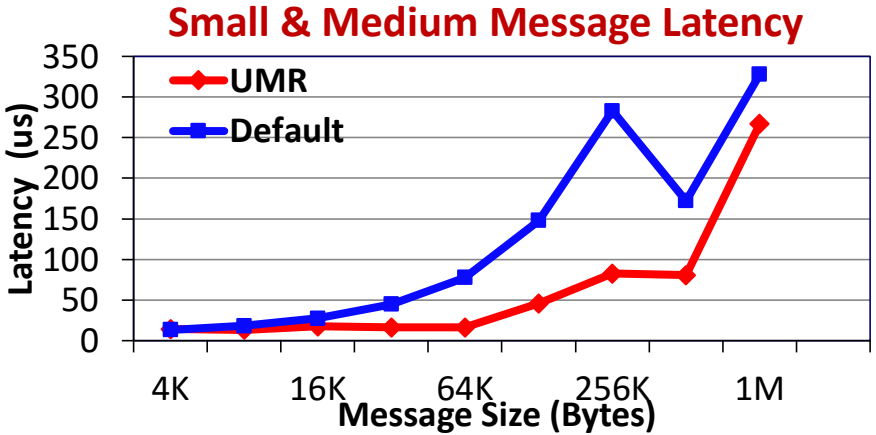
NAMD - Apoa1: Large data set

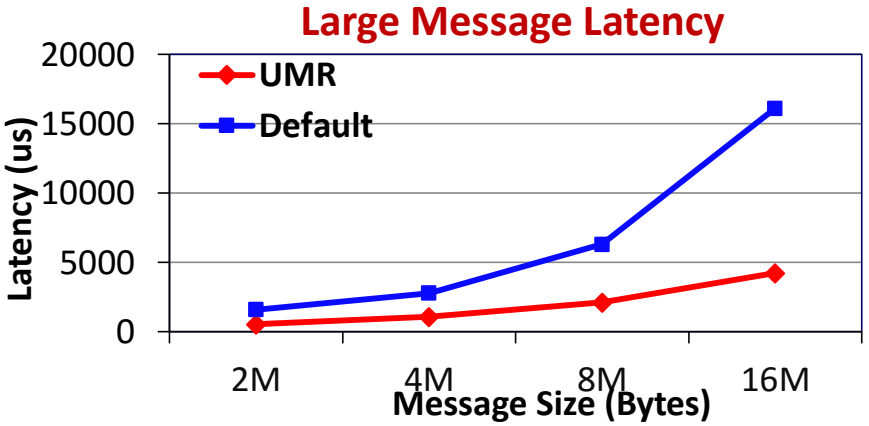


H. Subramoni, K. Hamidouche, A. Venkatesh, S. Chakraborty and D. K. Panda, Designing MPI Library with Dynamic Connected Transport (DCT) of InfiniBand : Early Experiences. IEEE International Supercomputing Conference (ISC '14)

User-mode Memory Registration (UMR)

- Introduced by Mellanox to support direct local and remote noncontiguous memory access
- Avoid packing at sender and unpacking at receiver
- Available in MVAPICH2-X 2.2b



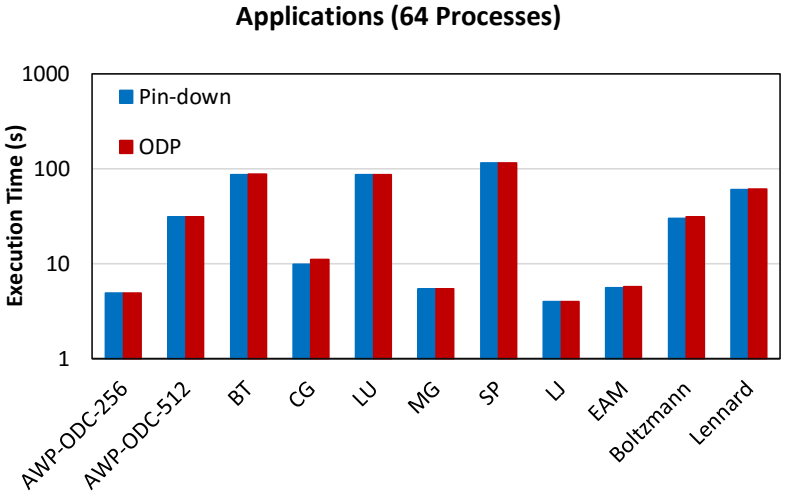


Connect-IB (54 Gbps): 2.8 GHz Dual Ten-core (IvyBridge) Intel PCI Gen3 with Mellanox IB FDR switch

M. Li, H. Subramoni, K. Hamidouche, X. Lu and D. K. Panda, "High Performance MPI Datatype Support with User-mode Memory Registration: Challenges, Designs and Benefits", CLUSTER, 2015

On-Demand Paging (ODP)

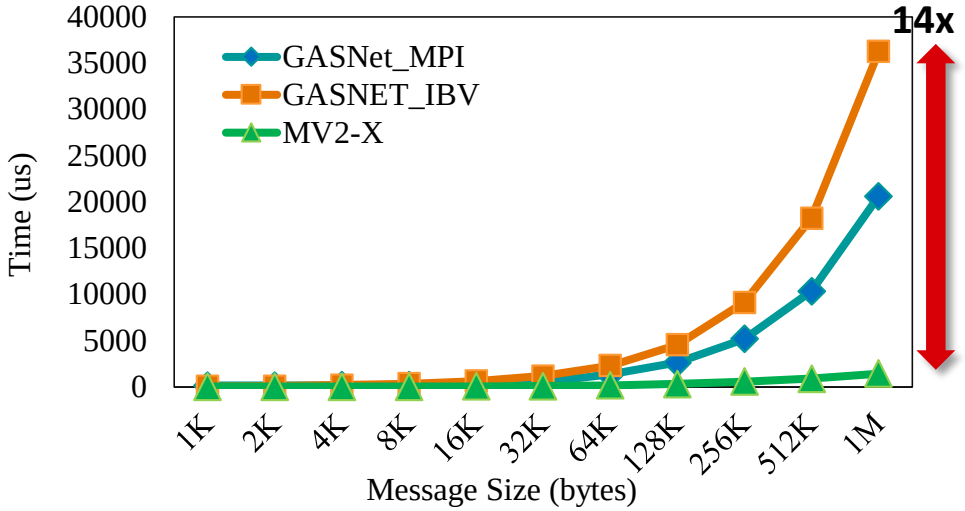
- Applications no longer need to pin down underlying physical pages
- Memory Region (MR) are **NEVER** pinned by the OS
  - Paged in by the HCA when needed
  - Paged out by the OS when reclaimed
- ODP can be divided into two classes
  - Explicit ODP
    - Applications still register memory buffers for communication, but this operation is used to define access control for IO rather than pin-down the pages
  - Implicit ODP
    - Applications are provided with a special memory key that represents their complete address space, does not need to register any virtual address range
- Advantages
  - Simplifies programming
  - Unlimited MR sizes
  - Physical memory optimization



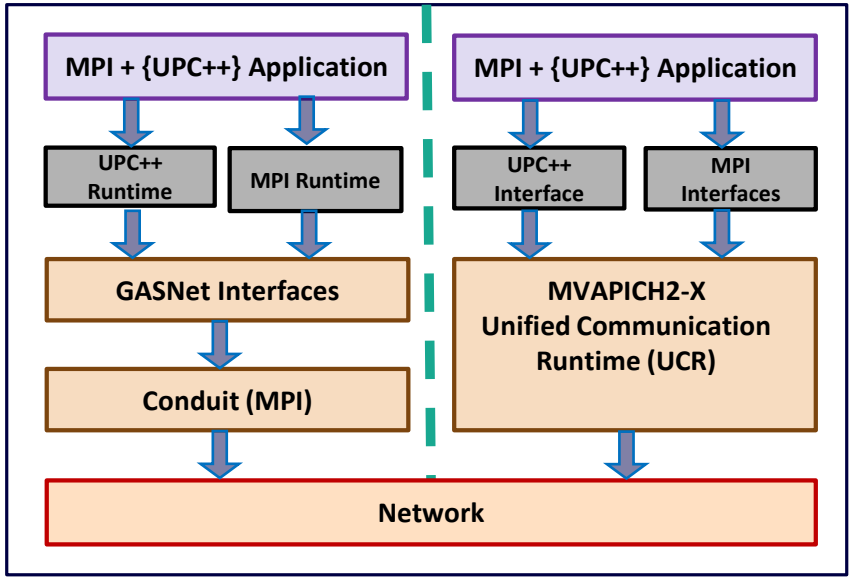
M. Li, K. Hamidouche, X. Lu, H. Subramoni, J. Zhang, and D. K. Panda, "Designing MPI Library with On-Demand Paging (ODP) of InfiniBand: Challenges and Benefits", SC 2016.

Wednesday 11/16/2016 @ 11:00 – 11:30 AM in Room 355-D

UPC++ Support in MVAPICH2-X



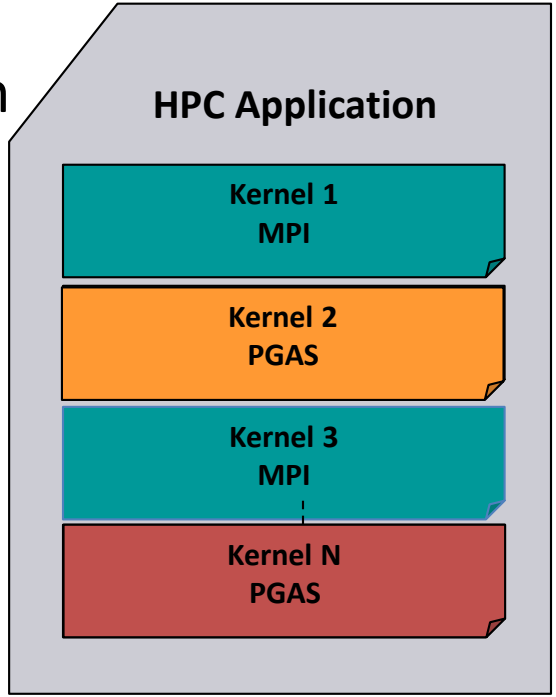
Inter-node Broadcast (64 nodes 1:ppn)



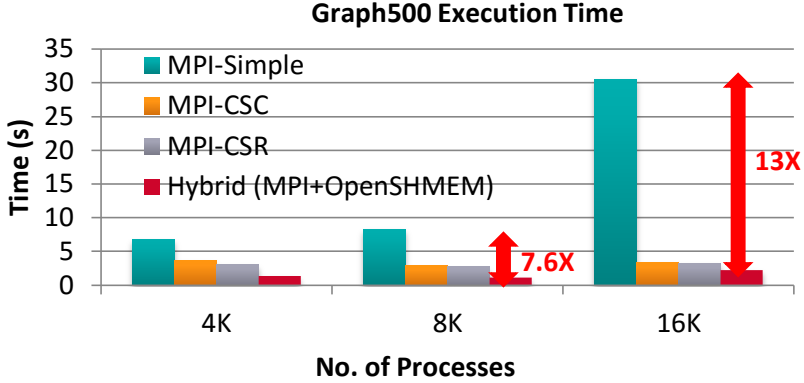
- Full and native support for hybrid MPI + UPC++ applications
- Better performance compared to IBV and MPI conduits
- OSU Micro-benchmarks (OMB) support for UPC++
- Available since MVAPICH2-X (2.2rc1)

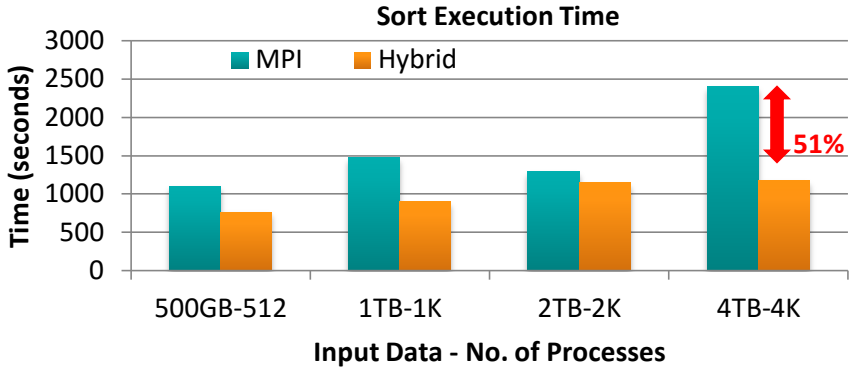
Hybrid (MPI+PGAS) Programming

- Application sub-kernels can be re-written in MPI/PGAS based on communication characteristics
- Benefits:
  - Best of Distributed Computing Model
  - Best of Shared Memory Computing Model



Application Level Performance with Graph500 and Sort





- Performance of Hybrid (MPI+ OpenSHMEM) Graph500 Design
  - 8,192 processes
    - **2.4X** improvement over MPI-CSR
    - **7.6X** improvement over MPI-Simple
  - 16,384 processes
    - **1.5X** improvement over MPI-CSR
    - **13X** improvement over MPI-Simple

- Performance of Hybrid (MPI+OpenSHMEM) Sort Application
  - 4,096 processes, 4 TB Input Size
    - MPI – 2408 sec; 0.16 TB/min
    - Hybrid – 1172 sec; 0.36 TB/min
    - **51%** improvement over MPI-design

J. Jose, K. Kandalla, S. Potluri, J. Zhang and D. K. Panda, Optimizing Collective Communication in OpenSHMEM, Int'l Conference on Partitioned Global Address Space Programming Models (PGAS '13), October 2013.

J. Jose, S. Potluri, K. Tomko and D. K. Panda, Designing Scalable Graph500 Benchmark with Hybrid MPI+OpenSHMEM Programming Models, International Supercomputing Conference (ISC'13), June 2013

J. Jose, K. Kandalla, M. Luo and D. K. Panda, Supporting Hybrid MPI and OpenSHMEM over InfiniBand: Design and Performance Evaluation, Int'l Conference on Parallel Processing (ICPP '12), September 2012

MVAPICH2 Distributions

- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE
  - Basic GPU (CUDA-aware MPI) support
- MVAPICH2-X
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - with and without Open Stack
- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- OSU Micro-Benchmarks (OMB)
  - MPI (including CUDA-aware MPI), OpenSHMEM and UPC
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool
- MVAPIC2-MIC (KNL)
  - Optimized for IB clusters with Intel Xeon Phis
- MVAPICH2-GDR and Deep Learning (Will be presented on Thursday at 11:00am)

Can HPC and Virtualization be Combined?

- Virtualization has many benefits
  - Fault-tolerance
  - Job migration
  - Compaction
- Have not been very popular in HPC due to overhead associated with Virtualization
- New SR-IOV (Single Root – IO Virtualization) support available with Mellanox InfiniBand adapters changes the field
- Enhanced MVAPICH2 support for SR-IOV
- MVAPICH2-Virt 2.1 (with and without OpenStack) is publicly available

J. Zhang, X. Lu, J. Jose, R. Shi and D. K. Panda, Can Inter-VM Shmem Benefit MPI Applications on SR-IOV based Virtualized InfiniBand Clusters? EuroPar'14

J. Zhang, X. Lu, J. Jose, M. Li, R. Shi and D.K. Panda, High Performance MPI Libray over SR-IOV enabled InfiniBand Clusters, HiPC'14

J. Zhang, X .Lu, M. Arnold and D. K. Panda, MVAPICH2 Over OpenStack with SR-IOV: an Efficient Approach to build HPC Clouds, CCGrid'15

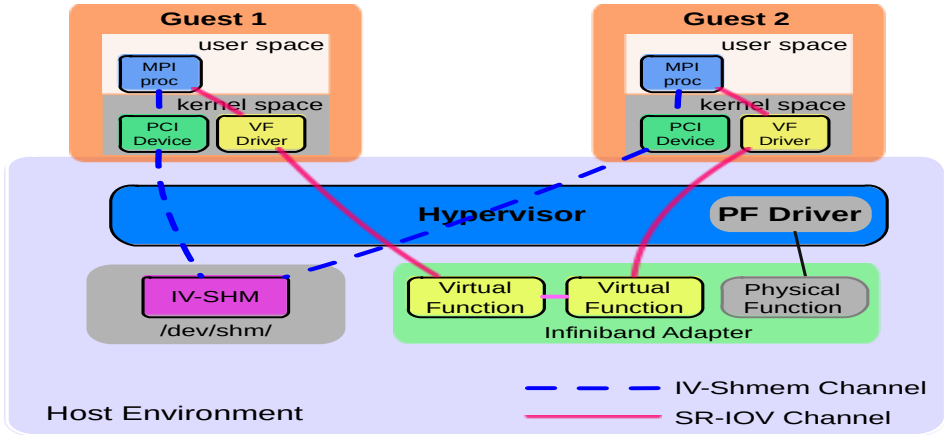
NSF Chameleon Cloud: A Powerful and Flexible Experimental Instrument

- 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
- 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/

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 shared memory based data access across co-resident VMs
  - Locality Detector: maintains the locality information of co-resident virtual machines
  - Communication Coordinator: selects the communication channel (SR-IOV, IVSHMEM) adaptively

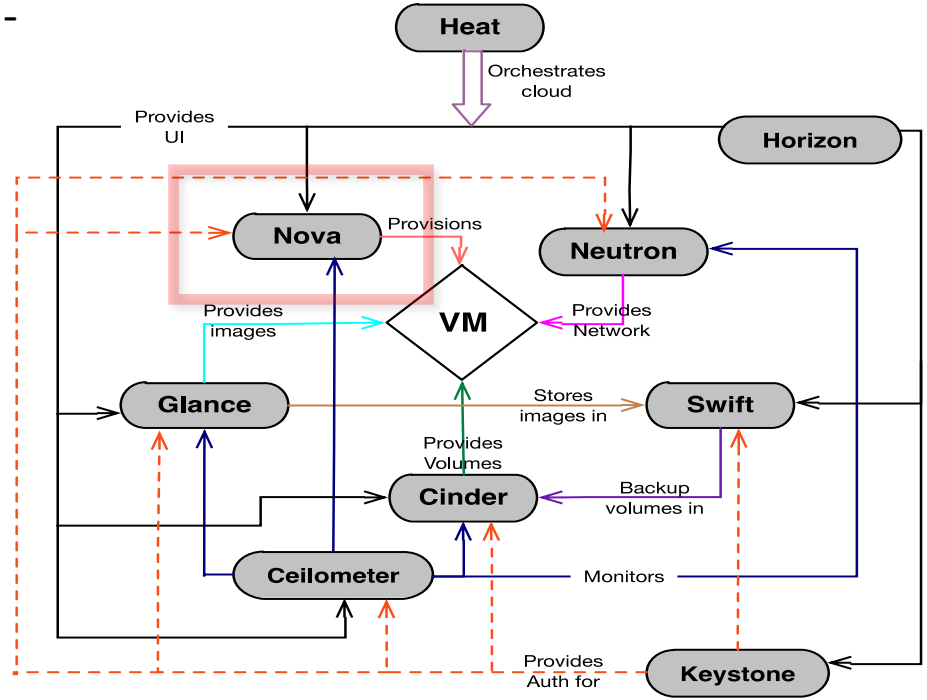


J. Zhang, X. Lu, J. Jose, R. Shi, D. K. Panda. Can Inter-VM Shmem Benefit MPI Applications on SR-IOV based Virtualized InfiniBand Clusters? **Euro-Par**, 2014

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

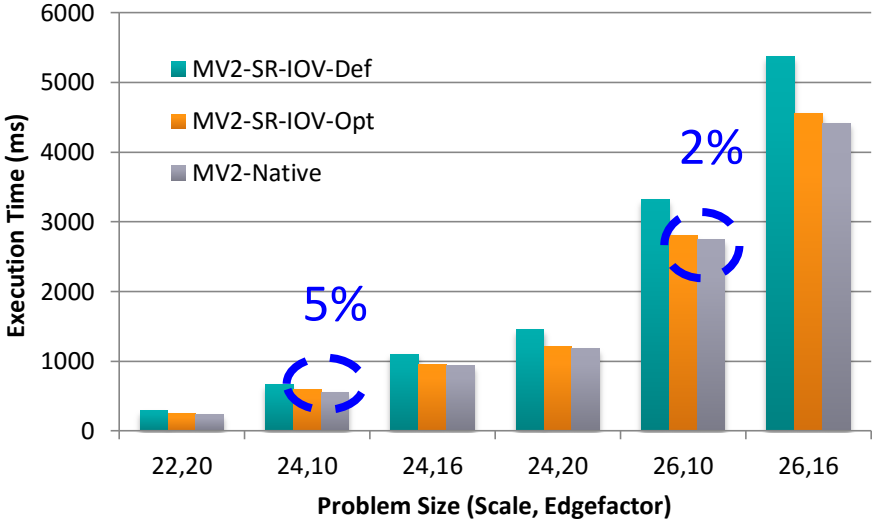
MVAPICH2-Virt with SR-IOV and IVSHMEM over OpenStack

- 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

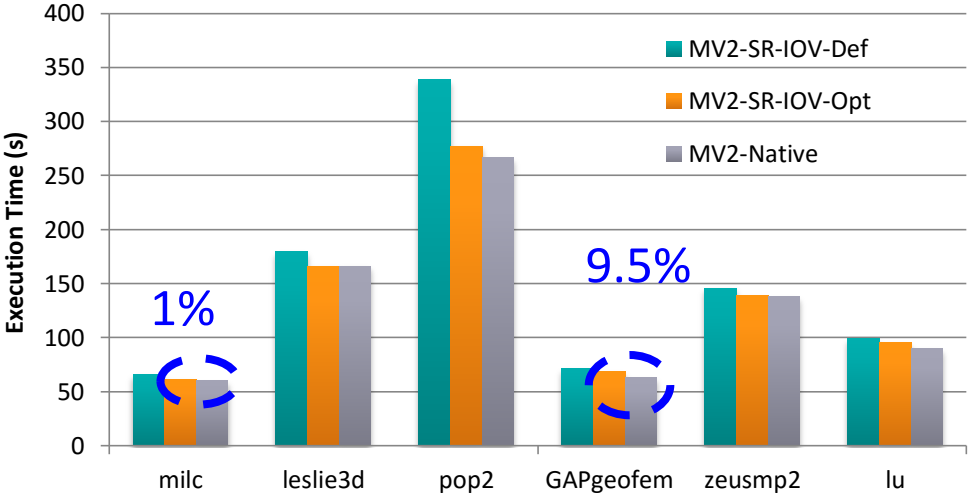


J. Zhang, X. Lu, M. Arnold, D. K. Panda. MVAPICH2 over OpenStack with SR-IOV: An Efficient Approach to Build HPC Clouds. **CCGrid**, 2015

Application-Level Performance on Chameleon



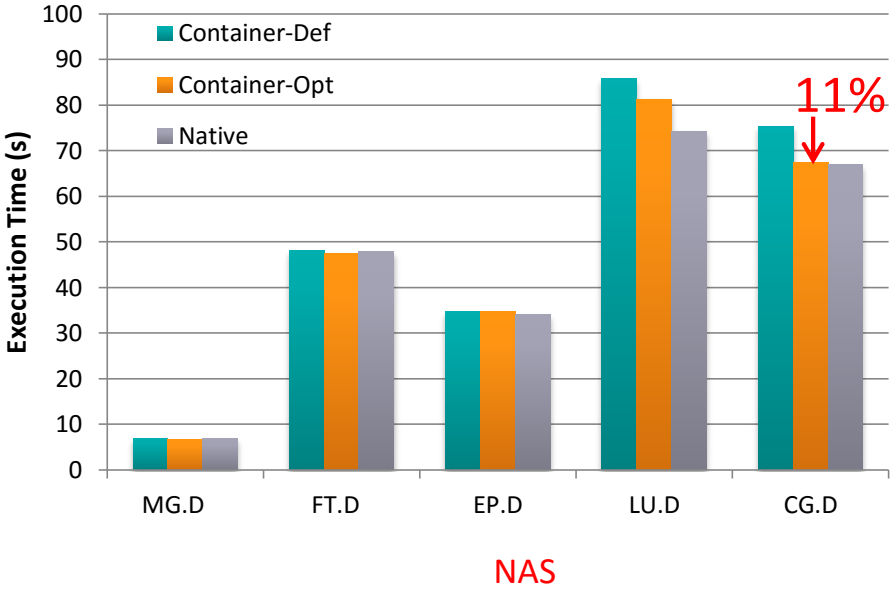
Graph500



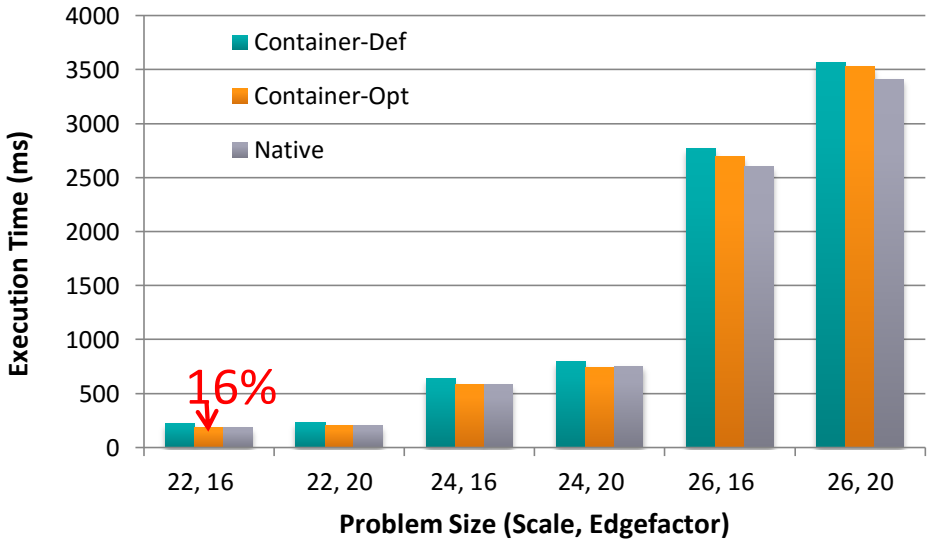
SPEC MPI2007

- 32 VMs, 6 Core/VM
- Compared to Native, 2-5% overhead for Graph500 with 128 Procs
- Compared to Native, 1-9.5% overhead for SPEC MPI2007 with 128 Procs

Containers Support: Application-Level Performance on Chameleon



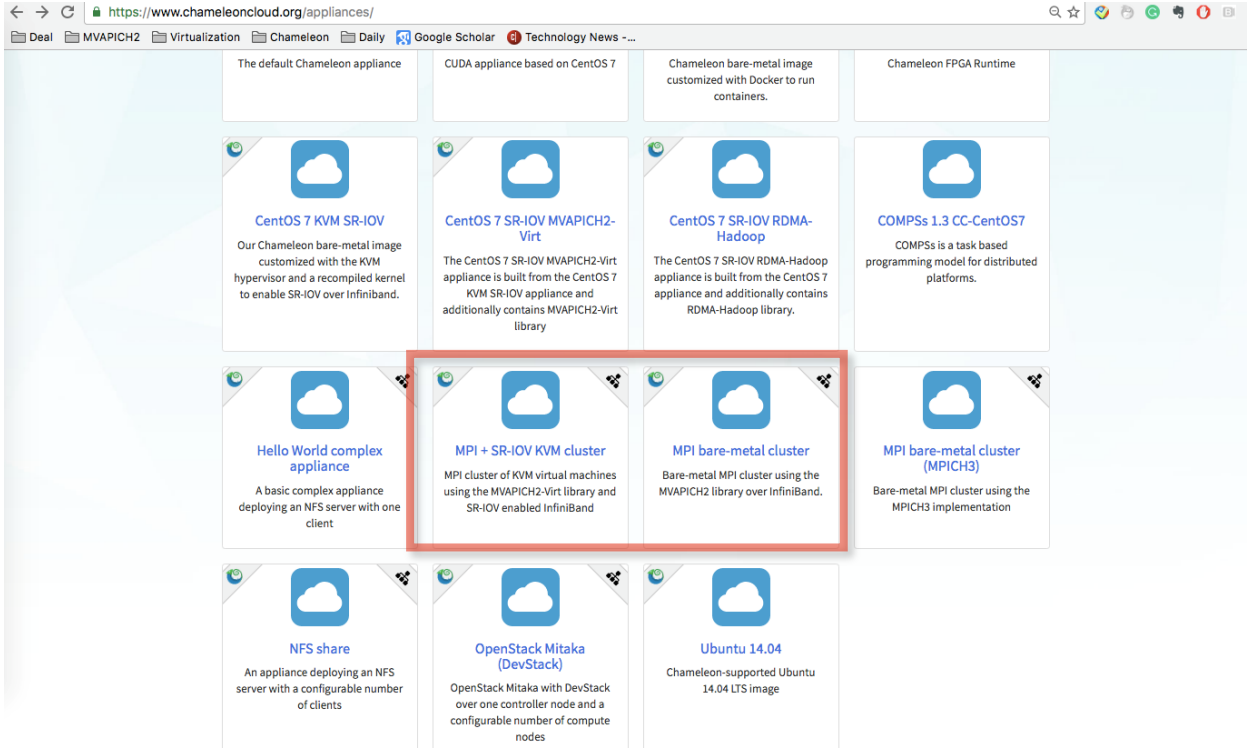
NAS



Graph 500

- 64 Containers across 16 nodes, pining 4 Cores per Container
- Compared to Container-Def, up to 11% and 16% of execution time reduction for NAS and Graph 500
- Compared to Native, less than 9 % and 4% overhead for NAS and Graph 500
- **Optimized Container support will be available with the upcoming release of MVAPICH2-Virt**

MPI Complex Appliances based on MVAPICH2 on Chameleon

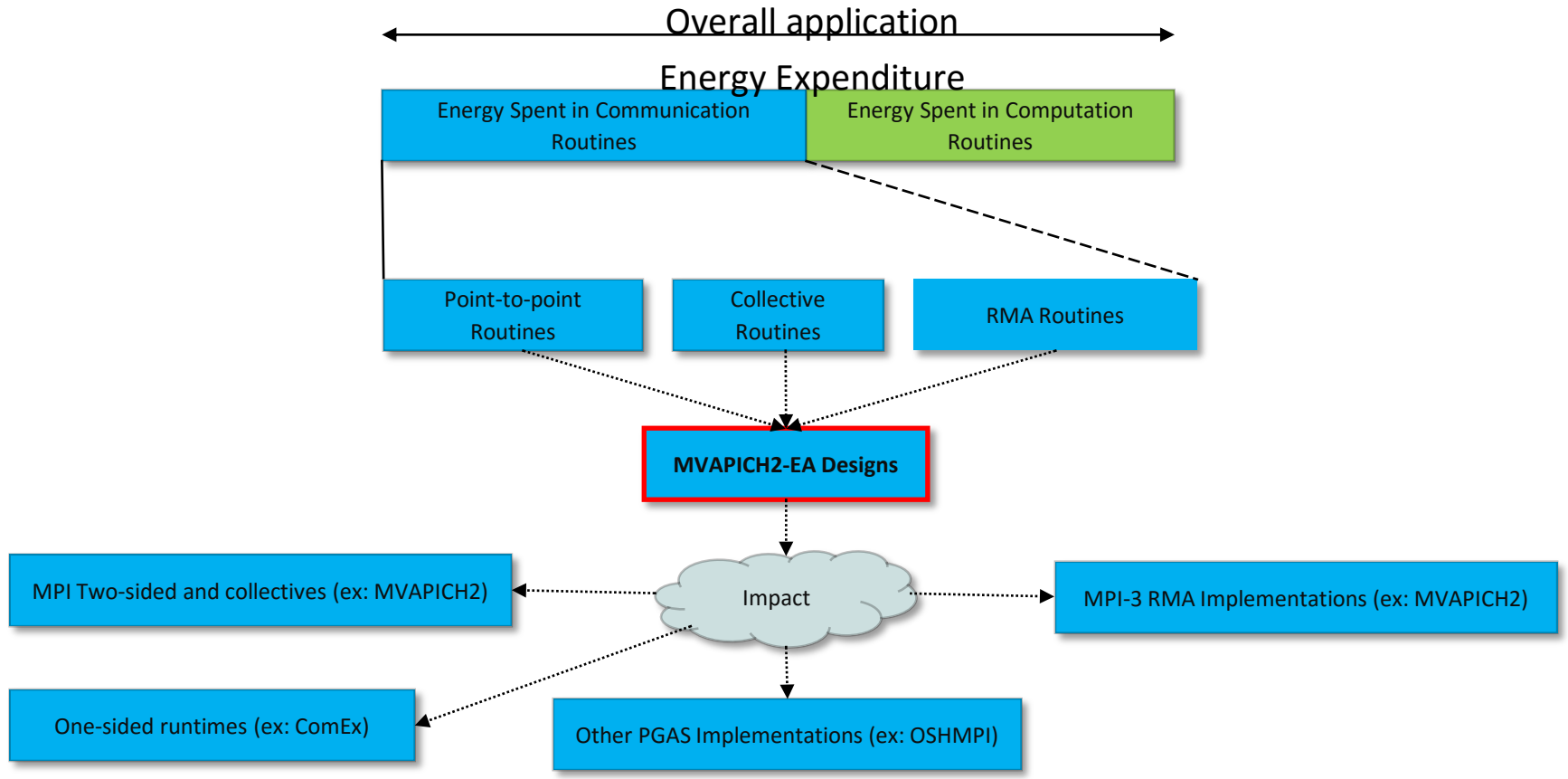


- Through these available appliances, users and researchers can easily deploy HPC clouds to perform experiments and run jobs with
  - High-Performance SR-IOV + InfiniBand
  - High-Performance MVAPICH2 Library with Virtualization Support
  - High-Performance Hadoop with RDMA-based Enhancements Support

MVAPICH2 Distributions

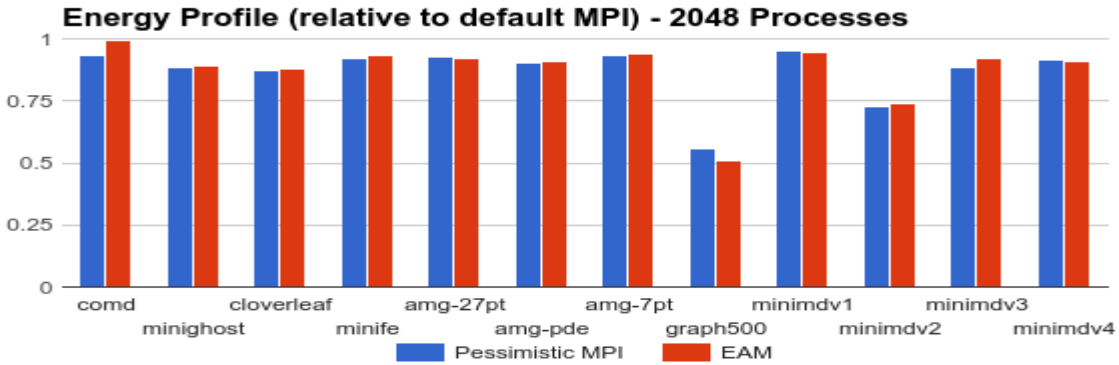
- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE
  - Basic GPU (CUDA-aware MPI) support
- MVAPICH2-X
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - with and without Open Stack
- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- OSU Micro-Benchmarks (OMB)
  - MPI (including CUDA-aware MPI), OpenSHMEM and UPC
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool
- MVAPIC2-MIC (KNL)
  - Optimized for IB clusters with Intel Xeon Phis
- MVAPICH2-GDR and Deep Learning (Will be presented on Thursday at 11:00am)

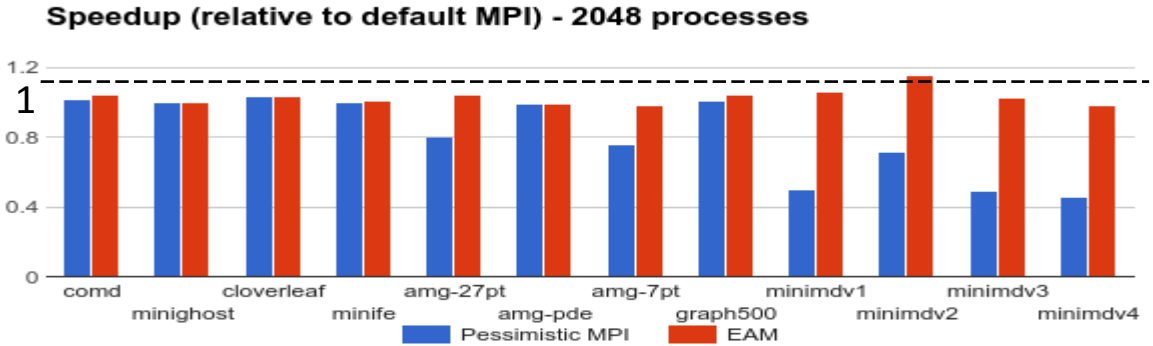
Designing Energy-Aware (EA) MPI Runtime



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

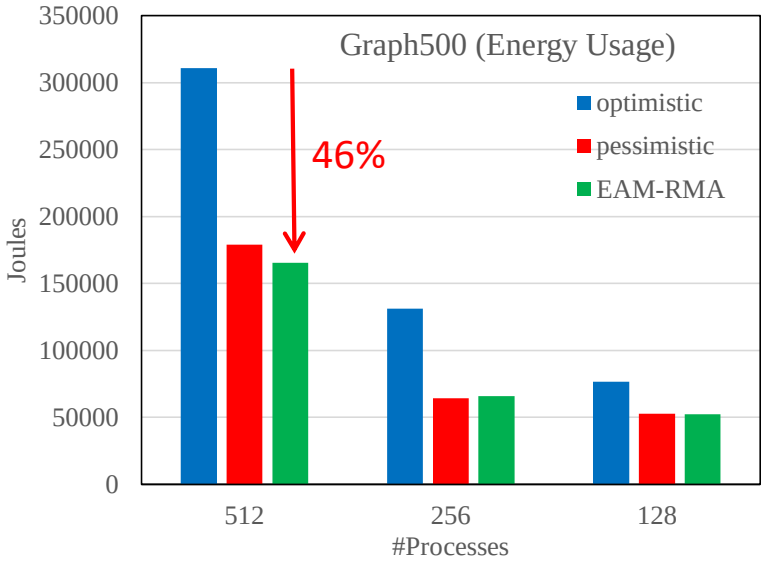
- An energy efficient runtime that provides energy savings without application knowledge
- Uses automatically and transparently the best energy lever
- Provides guarantees on maximum degradation with 5-41% savings at <= 5% degradation
- Pessimistic MPI applies energy reduction lever to each MPI call
- Released (together with OEMT) since Aug'15

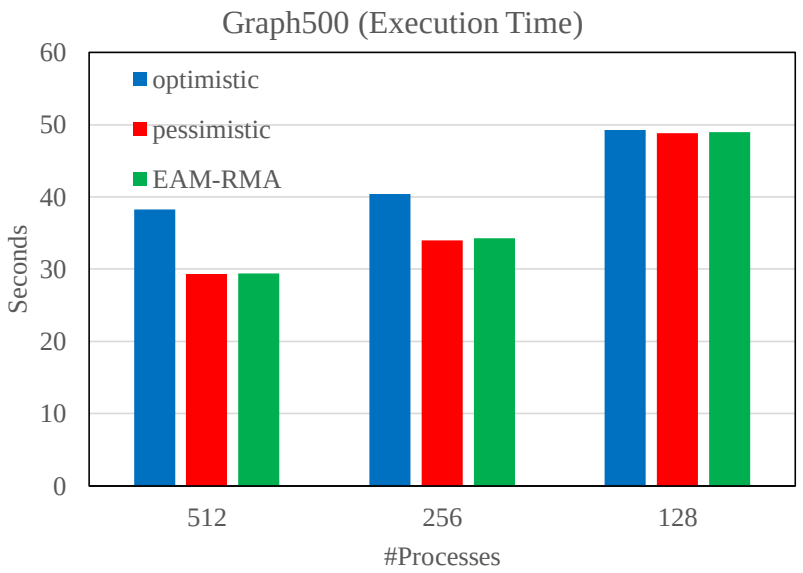




A Case for Application-Oblivious Energy-Efficient MPI Runtime A. Venkatesh, A. Vishnu, K. Hamidouche, N. Tallent, D. K. Panda, D. Kerbyson, and A. Hoise, Supercomputing '15, Nov 2015 *[Best Student Paper Finalist]*

MPI-3 RMA Energy Savings with Proxy-Applications





- MPI_Win_fence dominates application execution time in graph500
- Between 128 and 512 processes, EAM-RMA yields between 31% and 46% savings with no degradation in execution time in comparison with the default optimistic MPI runtime
- Solutions to be available in future releases of MVAPICH2-EA

MVAPICH2 Distributions

- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE
  - Basic GPU (CUDA-aware MPI) support
- MVAPICH2-X
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - with and without Open Stack
- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- OSU Micro-Benchmarks (OMB)
  - MPI (including CUDA-aware MPI), OpenSHMEM and UPC
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool
- MVAPIC2-MIC (KNL)
  - Optimized for IB clusters with Intel Xeon Phis
- MVAPICH2-GDR and Deep Learning (Will be presented on Thursday at 11:00am)

OSU Microbenchmarks

- Available since 2004
- Suite of microbenchmarks to study communication performance of various programming models
- Benchmarks available for the following programming models
  - Message Passing Interface (MPI)
  - Partitioned Global Address Space (PGAS)
    - Unified Parallel C (UPC)
    - Unified Parallel C++ (UPC++)
    - OpenSHMEM
- Benchmarks available for multiple accelerator based architectures
  - Compute Unified Device Architecture (CUDA)
  - OpenACC Application Program Interface
- Part of various national resource procurement suites like NERSC-8 / Trinity Benchmarks
- Please visit the following link for more information
  - http://mvapich.cse.ohio-state.edu/benchmarks/

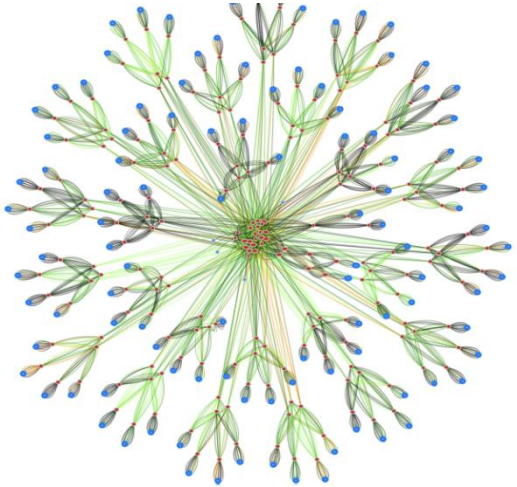
MVAPICH2 Distributions

- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE
  - Basic GPU (CUDA-aware MPI) support
- MVAPICH2-X
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - with and without Open Stack
- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- OSU Micro-Benchmarks (OMB)
  - MPI (including CUDA-aware MPI), OpenSHMEM and UPC
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool
- MVAPIC2-MIC (KNL)
  - Optimized for IB clusters with Intel Xeon Phis
- MVAPICH2-GDR and Deep Learning (Will be presented on Thursday at 11:00am)

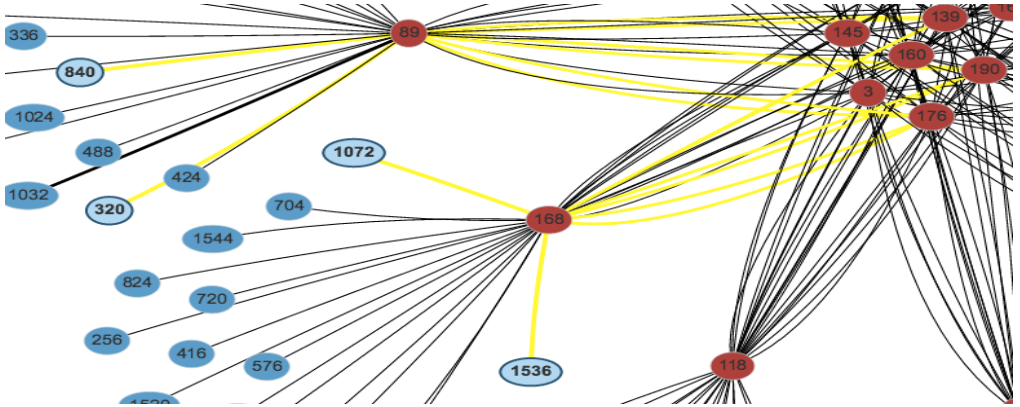
Overview of OSU INAM

- A network monitoring and analysis tool that is capable of analyzing traffic on the InfiniBand network with inputs from the MPI runtime
  - 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
- OSU INAM v0.9.1 released on 05/13/2016
- Significant enhancements to user interface to enable scaling to clusters with thousands of nodes
- Improve database insert times by using 'bulk inserts'
- Capability to look up list of nodes communicating through a network link
- Capability to classify data flowing over a network link at job level and process level granularity in conjunction with MVAPICH2-X 2.2rc1
- "Best practices " guidelines for deploying OSU INAM on different clusters
- 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

OSU INAM Features



Comet@SDSC --- Clustered View

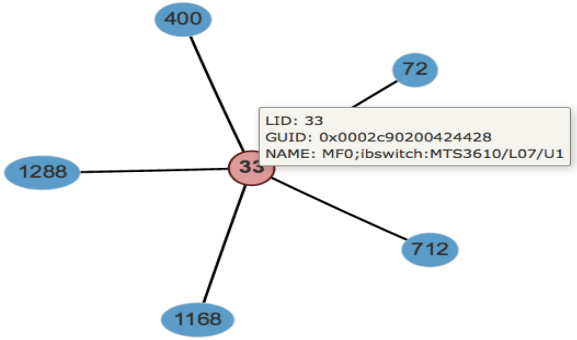


Finding Routes Between Nodes

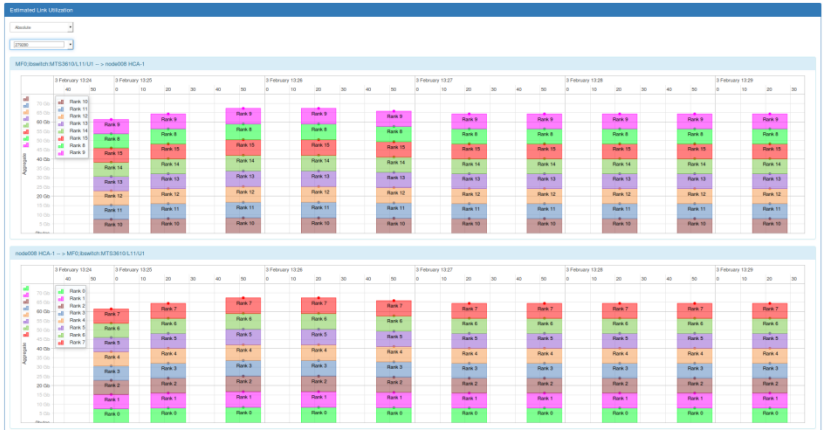
(1,879 nodes, 212 switches, 4,377 network links)

- 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 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 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

MVAPICH2 Distributions

- MVAPICH2
  - Basic MPI support for IB, iWARP and RoCE
  - Basic GPU (CUDA-aware MPI) support
- MVAPICH2-X
  - MPI, PGAS and Hybrid MPI+PGAS support for IB
- MVAPICH2-Virt
  - Optimized for HPC Clouds with IB and SR-IOV virtualization
  - with and without Open Stack
- MVAPICH2-EA
  - Energy Efficient Support for point-to-point and collective operations
  - Compatible with OSU Energy Monitoring Tool (OEMT-0.8)
- OSU Micro-Benchmarks (OMB)
  - MPI (including CUDA-aware MPI), OpenSHMEM and UPC
- OSU INAM
  - InfiniBand Network Analysis and Monitoring Tool
- MVAPIC2-MIC (KNL)
  - Optimized for IB clusters with Intel Xeon Phis
- MVAPICH2-GDR and Deep Learning (Will be presented on Thursday at 11:00am)

MVAPICH2-MIC Design for Clusters with IB and MIC

- Offload Mode
- Intranode Communication
  - Coprocessor-only Mode
  - Symmetric Mode
- Internode Communication
  - Coprocessors-only
  - Symmetric Mode
- Multi-MIC Node Configurations

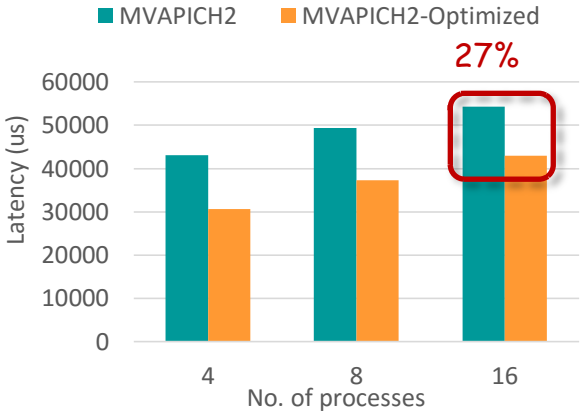
MVAPICH2 for Omni-Path and KNL

- MVAPICH2 has been supporting QLogic/PSM for many years with all different compute platforms
- Latest version supports
  - Omni-Path (derivative of QLogic IB)
    - Xeon family with Omni-Path
    - KNL with Omni-Path

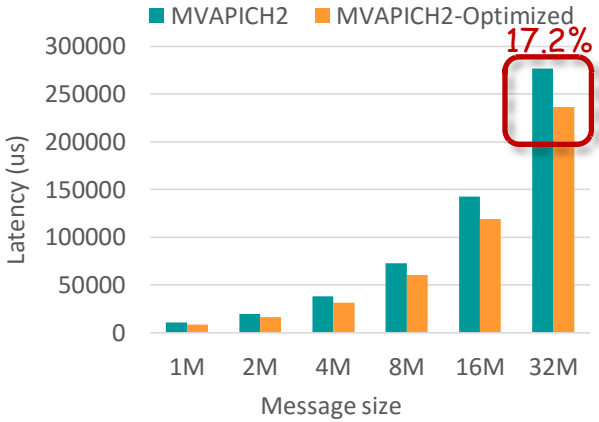
Enhanced Designs for KNL: MVAPICH2 Approach

- On-load approach
  - Takes advantage of the idle cores
  - Dynamically configurable
  - Takes advantage of highly multithreaded cores
  - Takes advantage of MCDRAM of KNL processors
- Applicable to other programming models such as PGAS, Task-based, etc.
- Provides portability, performance, and applicability to runtime as well as applications in a transparent manner

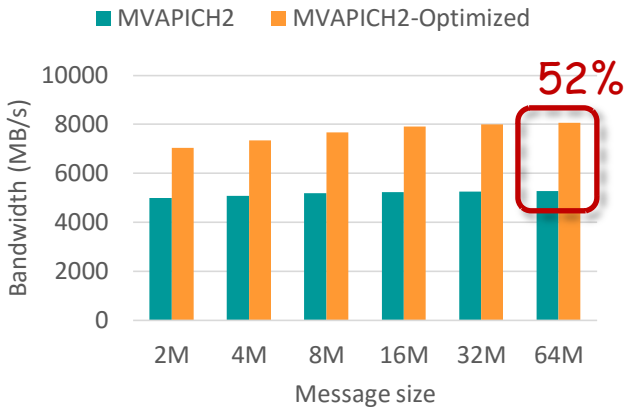
Performance Benefits of the Enhanced Designs



Intra-node Broadcast with 64MB Message



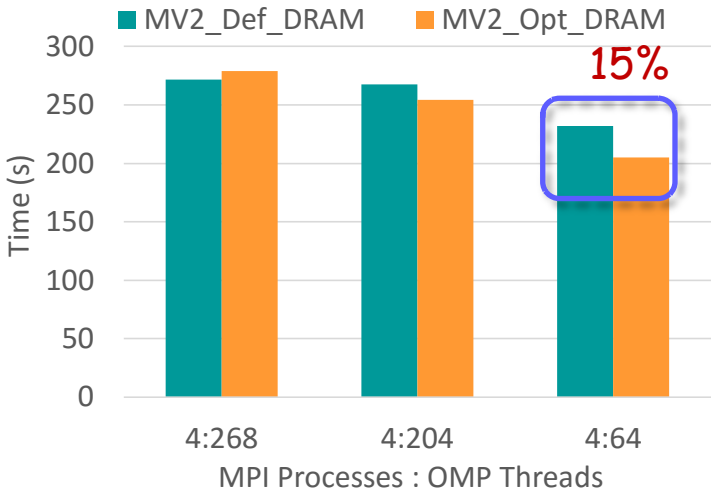
16-process Intra-node All-to-All



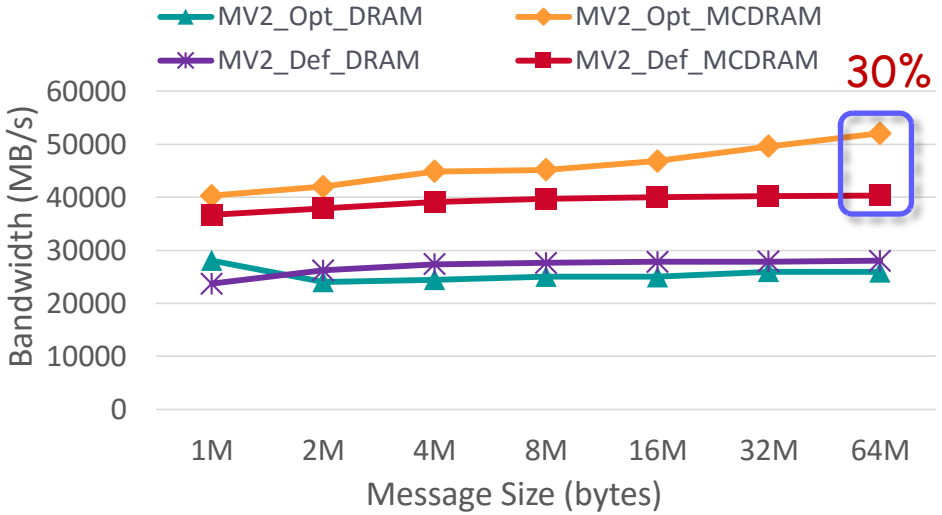
Very Large Message Bi-directional Bandwidth

- New designs to exploit high concurrency and MCDRAM of KNL
- Significant improvements for large message sizes
- Benefits seen in varying message size as well as varying MPI processes

Performance Benefits of the Enhanced Designs



CNTK: MLP Training Time using MNIST (BS:64)



Multi-Bandwidth using 32 MPI processes

- Benefits observed on training time of Multi-level Perceptron (MLP) model on MNIST dataset using CNTK Deep Learning Framework

Enhanced Designs will be available in upcoming MVAPICH2 releases

Applications-Level Tuning: Compilation of Best Practices

- MPI runtime has many parameters
- Tuning a set of parameters can help you to extract higher performance
- Compiled a list of such contributions through the MVAPICH Website
  - http://mvapich.cse.ohio-state.edu/best_practices/
- Initial list of applications
  - Amber
  - HoomDBlue
  - HPCG
  - Lulesh
  - MILC
  - Neuron
  - SMG2000
- Soliciting additional contributions, send your results to mvapich-help at cse.ohio-state.edu.
- We will link these results with credits to you.

MVAPICH2 – Plans for Exascale

- Performance and Memory scalability toward 1M cores
- Hybrid programming (MPI + OpenSHMEM, MPI + UPC, MPI + CAF …)
  - MPI + Task*
- Enhanced Optimization for GPU Support and Accelerators
- Taking advantage of advanced features of Mellanox InfiniBand
  - Switch-IB2 SHArP*
  - GID-based support*
- Enhanced communication schemes for upcoming architectures
  - Knights Landing with MCDRAM*
  - NVLINK*
  - CAPI*
- Extended topology-aware collectives
- Extended Energy-aware designs and Virtualization Support
- Extended Support for MPI Tools Interface (as in MPI 3.1)
- Extended Checkpoint-Restart and migration support with SCR
- Support for * features will be available in future MVAPICH2 Releases

Two More Presentations

- Wednesday (11/16/16) at 2:30pm

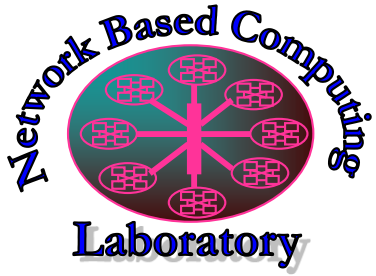
  High Performance Big Data (HiBD): Accelerating Hadoop, Spark and Memcached on Modern Clusters

- Thursday (11/17/16) at 11:00am

  MVAPICH2-GDR: Pushing the Frontier of HPC and Deep Learning

Thank You!

panda@cse.ohio-state.edu



Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

MVAPICH

The MVAPICH2 Project
http://mvapich.cse.ohio-state.edu/

HiBD
High-Performance
Big Data

The High-Performance Big Data Project
http://hibd.cse.ohio-state.edu/