MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

# High-Performance MPI Library with SR-IOV and SLURM for Virtualized InfiniBand Clusters

Talk at OpenFabrics Workshop (April 2016)

by

**Dhabaleswar K. (DK) Panda**

The Ohio State University

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

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

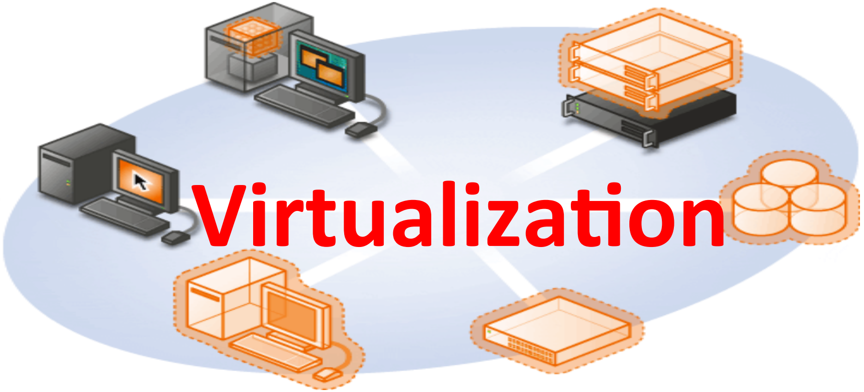
**Xiaoyi Lu**

The Ohio State University

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

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

# Cloud Computing and Virtualization


- Cloud Computing focuses on maximizing the effectiveness of the shared resources
- Virtualization is the key technology for resource sharing in the Cloud
- Widely adopted in industry computing environment
- IDC Forecasts Worldwide Public IT Cloud Services Spending to Reach Nearly $108 Billion by 2017 (Courtesy: http://www.idc.com/getdoc.jsp?containerId=prUS24298013)

# 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
- 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

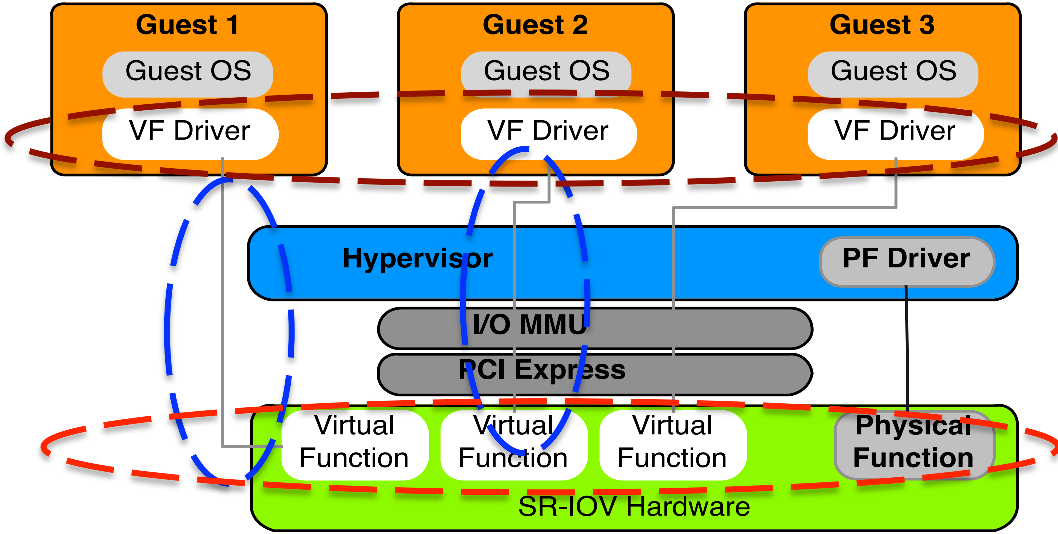
http://www.chameleoncloud.org/

Computation Institute TACC iCAIR The Ohio State University UTSA NSF

# 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)
- VFs are designed based on the existing non-virtualized PFs, no need for driver change
- Each VF can be dedicated to a single VM through PCI pass-through
- Work with 10/40 GigE and InfiniBand



# Building HPC Cloud with SR-IOV and InfiniBand

- High-Performance Computing (HPC) has adopted advanced interconnects and protocols
  - InfiniBand
  - 10 Gigabit Ethernet/iWARP
  - RDMA over Converged Enhanced Ethernet (RoCE)
- Very Good Performance
  - Low latency (few micro seconds)
  - High Bandwidth (100 Gb/s with EDR InfiniBand)
  - Low CPU overhead (5-10%)
- OpenFabrics software stack with IB, iWARP and RoCE interfaces are driving HPC systems
- How to Build HPC Cloud with SR-IOV and InfiniBand for delivering optimal performance?

# Overview of the MVAPICH2 Project

- 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 2,550 organizations in 79 countries**
  - **More than 360,000 (> 0.36 million) 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)

# MVAPICH2 Architecture

## High Performance Parallel Programming Models

- Message Passing Interface (MPI)
- PGAS (UPC, OpenSHMEM, CAF, UPC++)
- Hybrid --- MPI + X (MPI + PGAS + OpenMP/Cilk)

## High Performance and Scalable Communication Runtime

### Diverse APIs and Mechanisms

- Point-to-point Primitives
- Collectives Algorithms
- Job Startup
- Energy-Awareness
- Remote Memory Access
- I/O and File Systems
- Fault Tolerance
- Virtualization
- Active Messages
- Introspection & Analysis

### Support for Modern Networking Technology (InfiniBand, iWARP, RoCE, OmniPath)

**Transport Protocols:** RC, XRC, UD, DC

**Modern Features:** UMR, ODP*, SR-IOV, Multi Rail

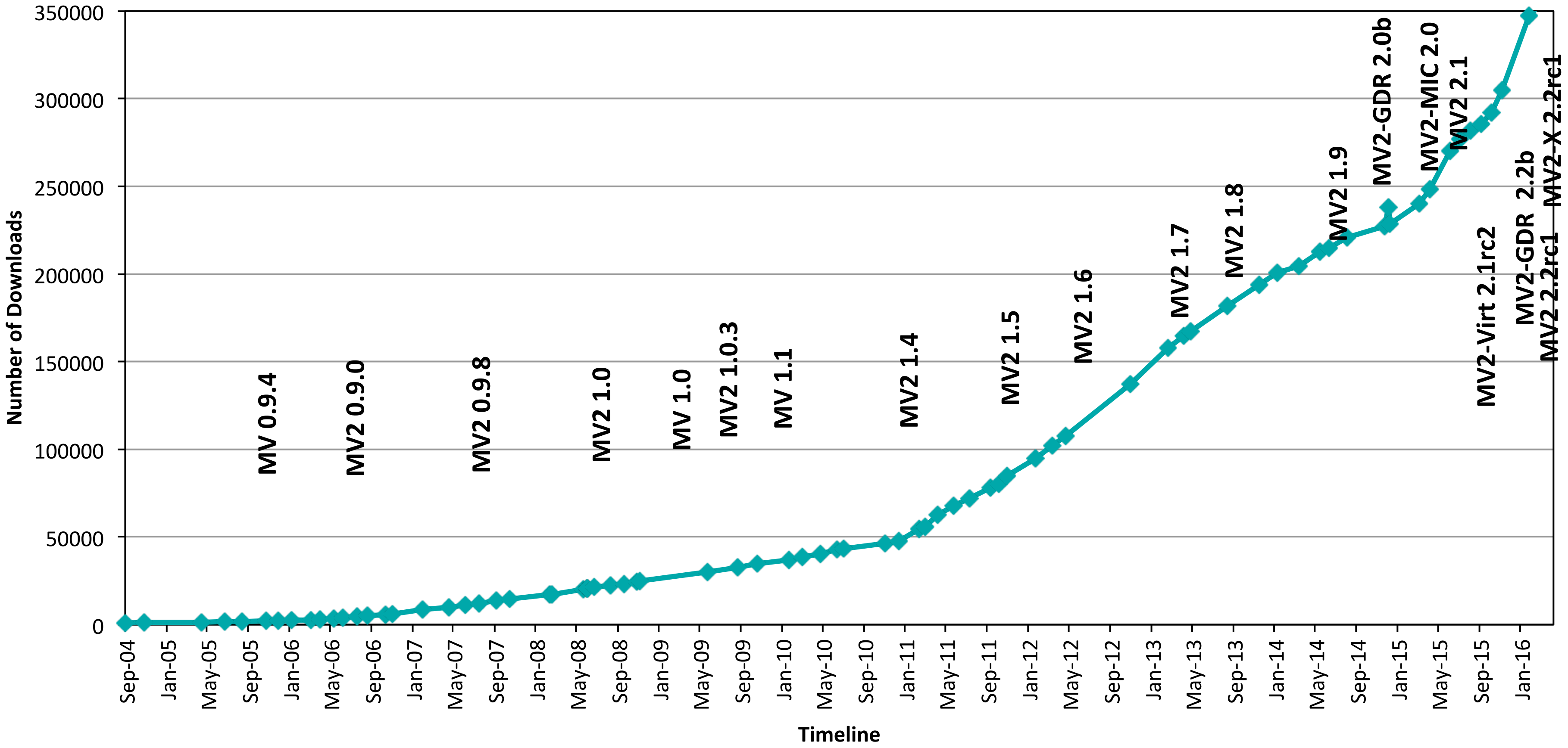
### Support for Modern Multi-/Many-core Architectures (Intel-Xeon, OpenPower, Xeon-Phi (MIC, KNL*), NVIDIA GPGPU)

**Transport Mechanisms:** Shared Memory, CMA, IVSHMEM

**Modern Features:** MCDRAM*, NVLink*, CAPI*

* Upcoming

# MVAPICH/MVAPICH2 Release Timeline and Downloads



# MVAPICH2 Software Family

| Requirements | MVAPICH2 Library to use |
|---|---|
| MPI with IB, iWARP and RoCE | MVAPICH2 |
| Advanced MPI, OSU INAM, PGAS and MPI+PGAS with IB and RoCE | MVAPICH2-X |
| MPI with IB & GPU | MVAPICH2-GDR |
| MPI with IB & MIC | MVAPICH2-MIC |
| HPC Cloud with MPI & IB | MVAPICH2-Virt |
| Energy-aware MPI with IB, iWARP and RoCE | MVAPICH2-EA |

# Three Designs

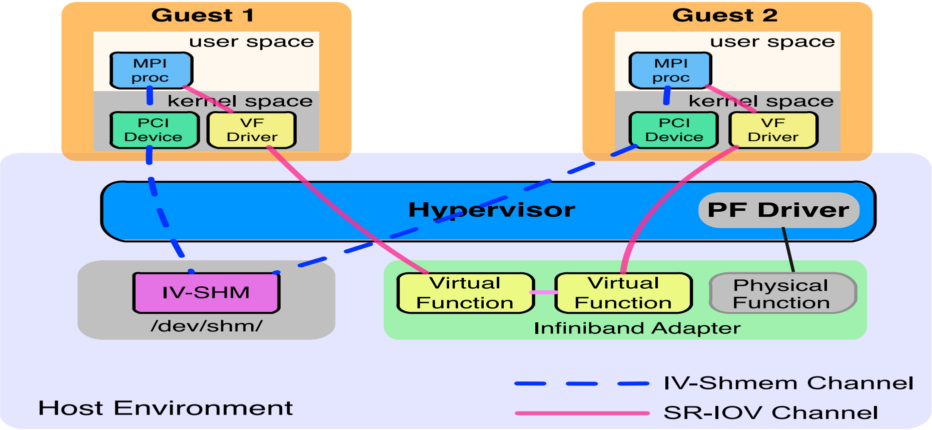
- MVAPICH2-Virt with SR-IOV and IVSHMEM
  - Standalone, OpenStack
- MVAPICH2-Virt on SLURM
- MVAPICH2 with Containers

# MVAPICH2-Virt 2.1

- Major Features and Enhancements
  - Based on MVAPICH2 2.1
  - Support for efficient MPI communication over SR-IOV enabled InfiniBand networks
  - High-performance and locality-aware MPI communication with IVSHMEM
  - Support for auto-detection of IVSHMEM device in virtual machines
  - Automatic communication channel selection among SR-IOV, IVSHMEM, and CMA/LiMIC2
  - Support for integration with OpenStack
  - Support for easy configuration through runtime parameters
  - Tested with
    - Mellanox InfiniBand adapters (ConnectX-3 (56Gbps))
    - OpenStack Juno

# 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

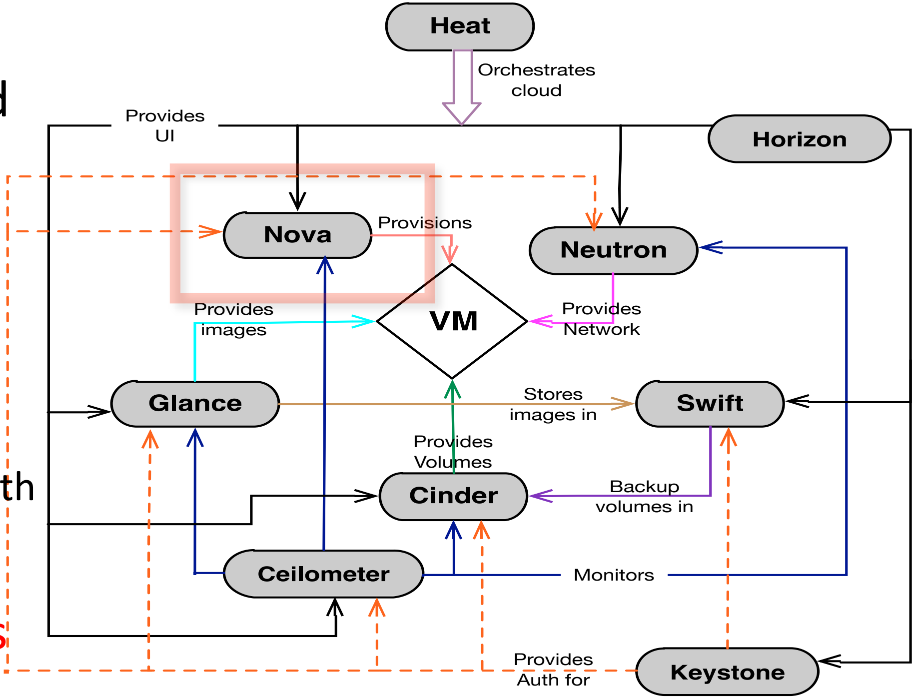



J. Zhang, X. Lu, J. Jose, R. Shi, D. K. Panda. Can Inter-VM Shmem Benefit MPI Applications on SR-IOV based Virtualized InfiniBand Clusters? **Euro-Par**, 2014

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

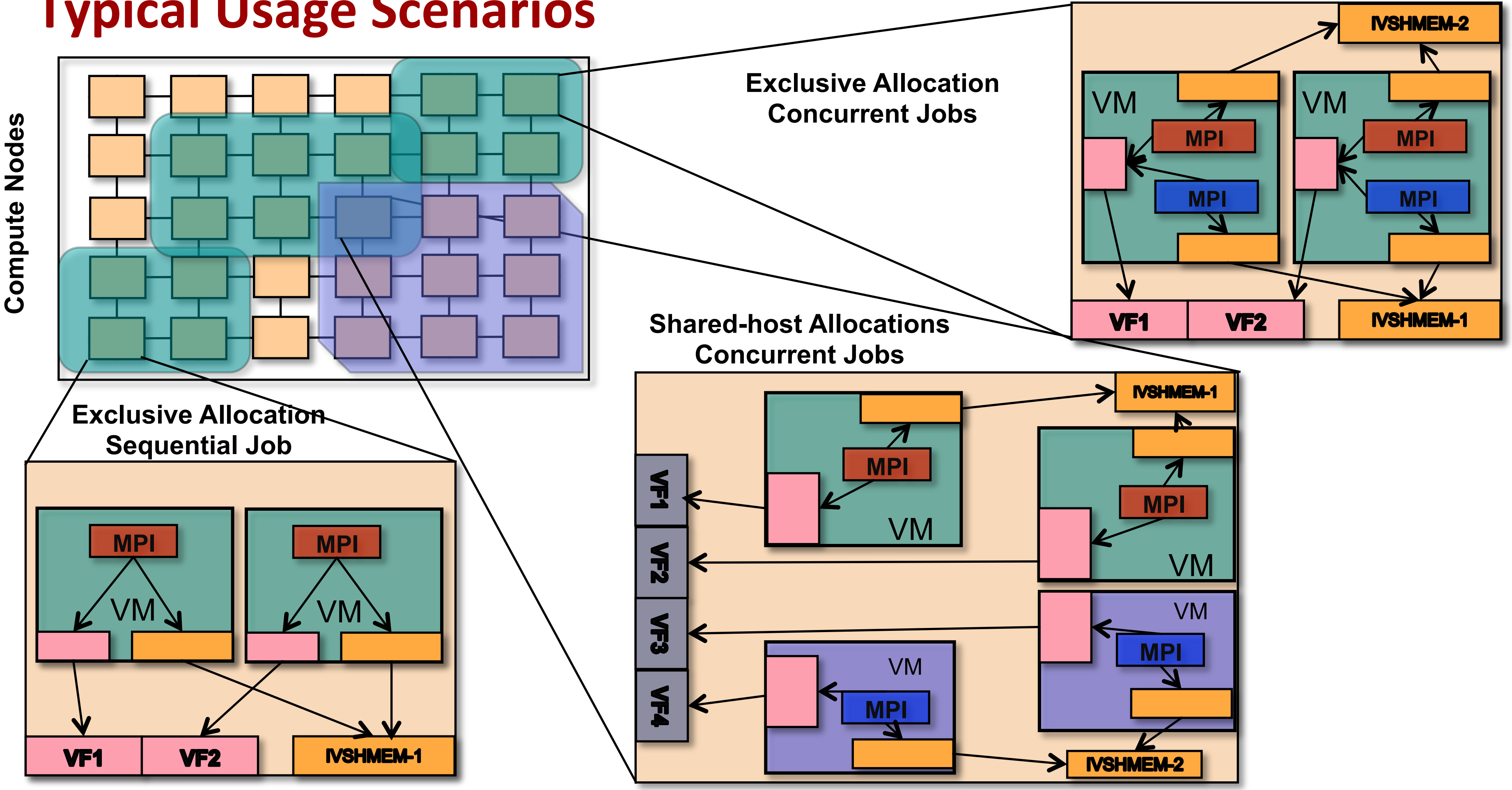
# Three Designs

- MVAPICH2-Virt with SR-IOV and IVSHMEM
  - Standalone, OpenStack
- MVAPICH2-Virt on SLURM
- MVAPICH2 with Containers

# Can HPC Clouds be built with MVAPICH2-Virt on SLURM?

- SLURM is one of the most popular open-source solutions to manage huge amounts of machines in HPC clusters.
- How to build a SLURM-based HPC Cloud with near native performance for MPI applications over SR-IOV enabled InfiniBand HPC clusters?
- What are the requirements on SLURM to support SR-IOV and IVSHMEM provided in HPC Clouds?
- How much performance benefit can be achieved on MPI primitive operations and applications in "MVAPICH2-Virt on SLURM"-based HPC clouds?

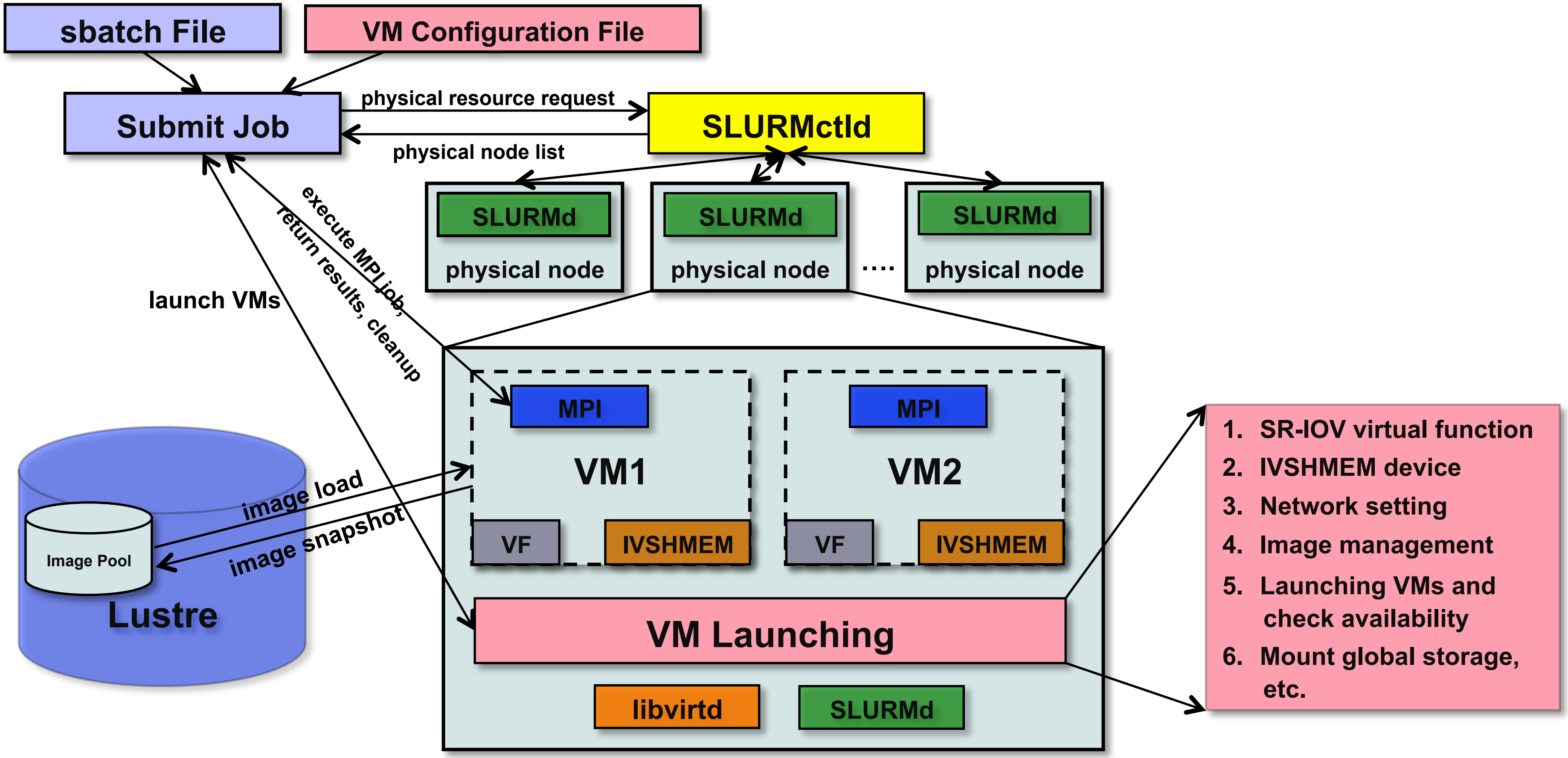
# Typical Usage Scenarios



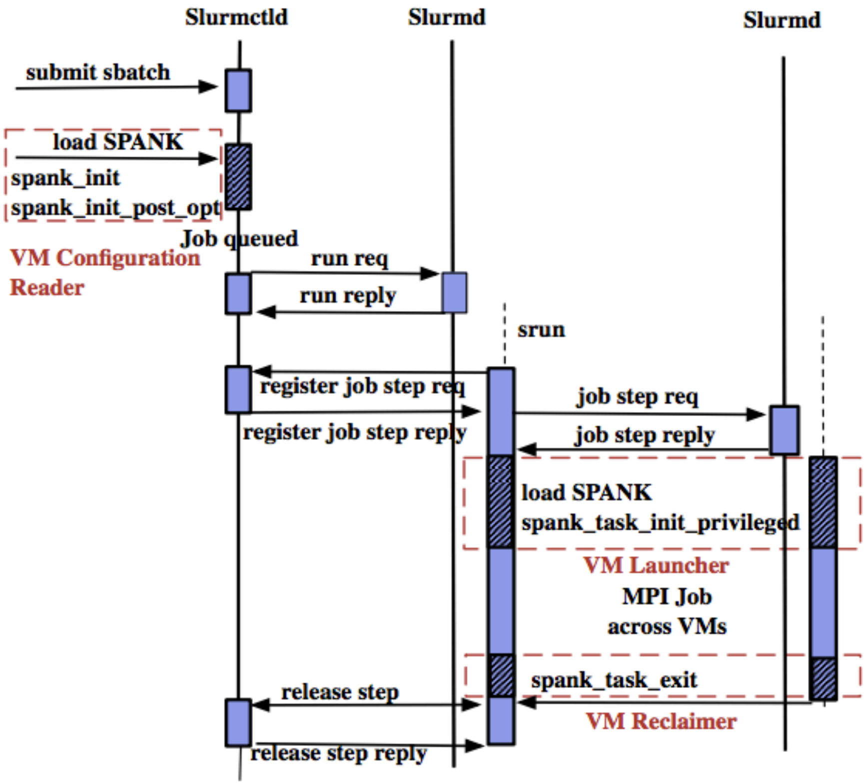
# Need for Supporting SR-IOV and IVSHMEM in SLURM

- Requirement of managing and isolating virtualized resources of SR-IOV and IVSHMEM
- Such kind of management and isolation is hard to be achieved by MPI library alone, but much easier with SLURM
- Efficient running MPI applications on HPC Clouds needs SLURM to support managing SR-IOV and IVSHMEM
  - Can critical HPC resources be efficiently shared among users by extending SLURM with support for SR-IOV and IVSHMEM based virtualization?
  - Can SR-IOV and IVSHMEM enabled SLURM and MPI library provide bare-metal performance for end applications on HPC Clouds?

# Workflow of Running MPI Jobs with MVAPICH2-Virt on SLURM



# SLURM SPANK Plugin based Design




- VM Configuration Reader – Register all VM configuration options, set in the job control environment so that they are visible to all allocated nodes.
- VM Launcher – Setup VMs on each allocated nodes.
  - File based lock to detect occupied VF and exclusively allocate free VF
  - Assign a unique ID to each IVSHMEM and dynamically attach to each VM
- VM Reclaimer – Tear down VMs and reclaim resources

# Benefits of Plugin-based Designs for SLURM

- Coordination
  - With global information, SLURM plugin can manage SR-IOV and IVSHMEM resources easily for concurrent jobs and multiple users
- Performance
  - Faster coordination, SR-IOV and IVSHMEM aware resource scheduling, etc.
- Scalability
  - Taking advantage of the scalable architecture of SLURM
- Fault Tolerance
- Permission
- Security

# Performance Evaluation

| Cluster | Nowlab Cloud | | Amazon EC2 | |
|---|---|---|---|---|
| Instance | 4 Core/VM | 8 Core/VM | 4 Core/VM | 8 Core/VM |
| Platform | RHEL 6.5 Qemu+KVM HVM SLURM 14.11.8 | | Amazon Linux (EL6) Xen HVM C3.xlarge [1] Instance | Amazon Linux (EL6) Xen HVM C3.2xlarge [1] Instance |
| CPU | SandyBridge Intel(R) Xeon E5-2670 (2.6GHz) | | IvyBridge Intel(R) Xeon E5-2680v2 (2.8GHz) | |
| RAM | 6 GB | 12 GB | 7.5 GB | 15 GB |
| Interconnect | FDR (56Gbps) InfiniBand Mellanox ConnectX-3 with SR-IOV [2] | | 10 GigE with Intel ixgbevf SR-IOV driver [2] | |

[1] Amazon EC2 C3 instances: compute-optimized instances, providing customers with the highest performing processors, good for HPC workloads

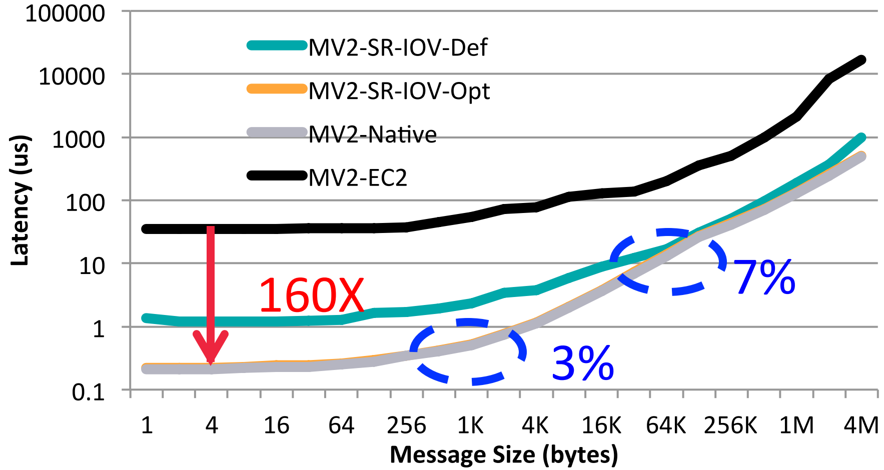
[2] Nowlab Cloud is using InfiniBand FDR (56Gbps), while Amazon EC2 C3 instances are using 10 GigE. Both have SR-IOV

# Experiments Carried Out

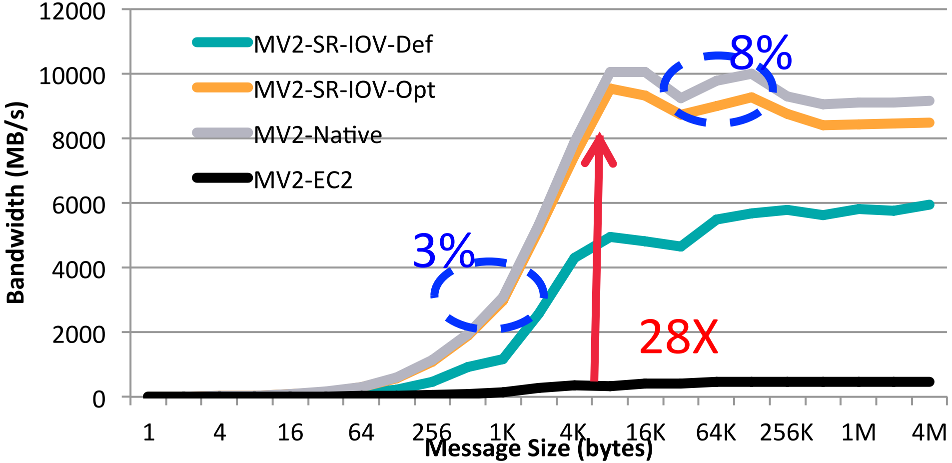
- Point-to-point
  - Two-sided and One-sided
  - Latency and Bandwidth
  - Intra-node and Inter-node [1]
- Applications
  - NAS and Graph500

[1] Amazon EC2 does not support users to explicitly allocate VMs in one physical node so far. We allocate multiple VMs in one logical group and compare the point-to-point performance for each pair of VMs. We see the VMs who have the lowest latency as located within one physical node (Intra-node), otherwise Inter-node.

# Point-to-Point Performance – Latency & Bandwidth (Intra-node)



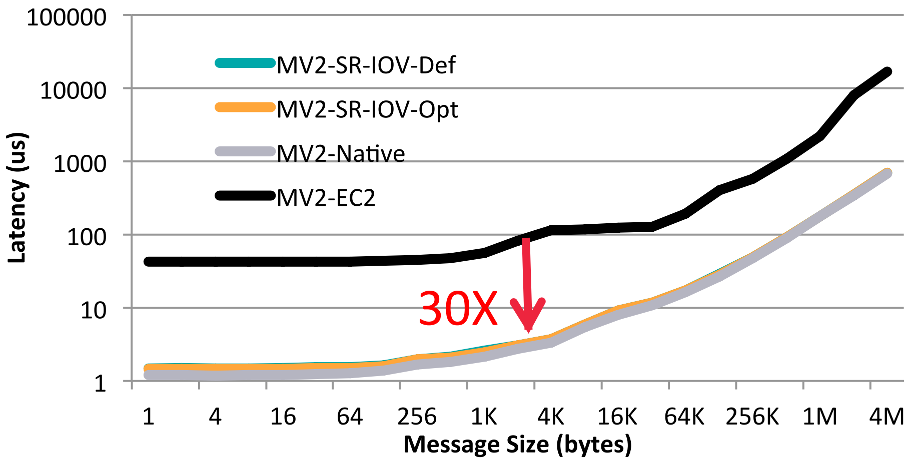

Intra-node Inter-VM pt2pt Latency



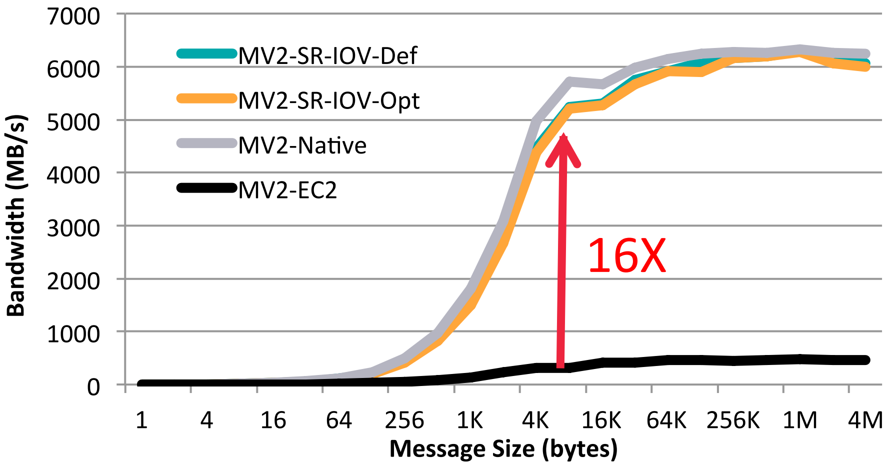

Intra-node Inter-VM pt2pt Bandwidth

- EC2 C3.2xlarge instances
- Compared to SR-IOV-Def, up to 84% and 158% performance improvement on Lat & BW
- Compared to Native, 3%-7% overhead for Lat, 3%-8% overhead for BW
- Compared to EC2, up to 160X and 28X performance speedup on Lat & BW

# Point-to-Point Performance – Latency & Bandwidth (Inter-node)



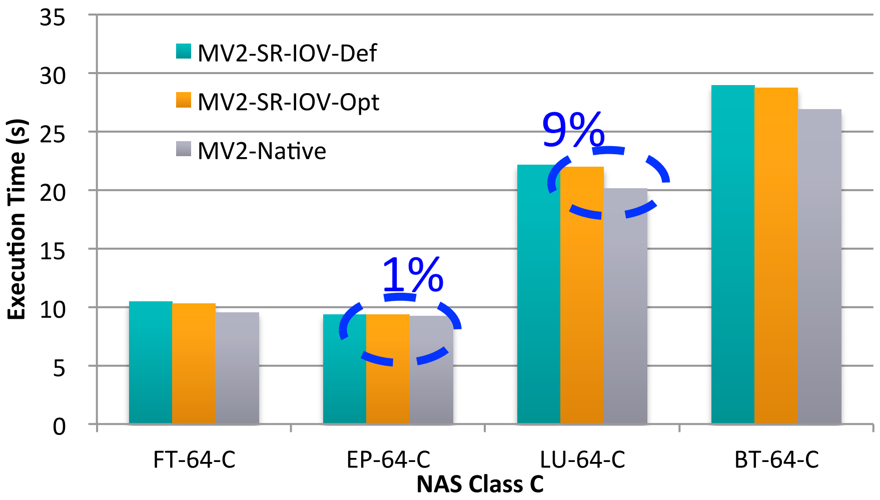

Inter-node Inter-VM pt2pt Latency



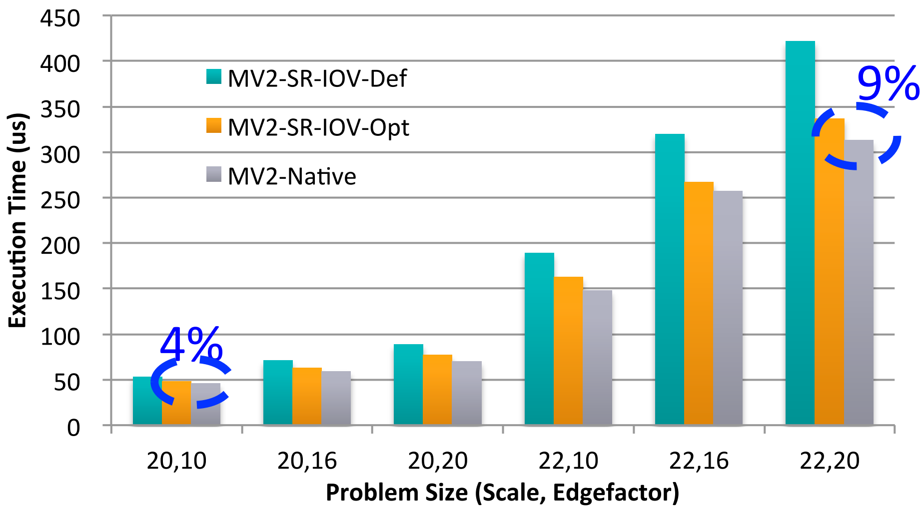

Inter-node Inter-VM pt2pt Bandwidth

- EC2 C3.2xlarge instances
- Similar performance with SR-IOV-Def
- Compared to Native, 2%-8% overhead on Lat & BW for 8KB+ messages
- Compared to EC2, up to 30X and 16X performance speedup on Lat & BW

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



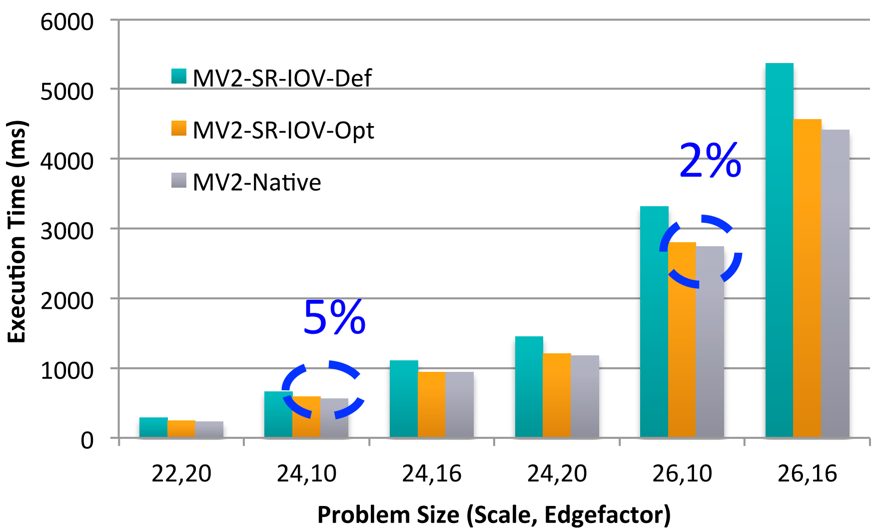

NAS



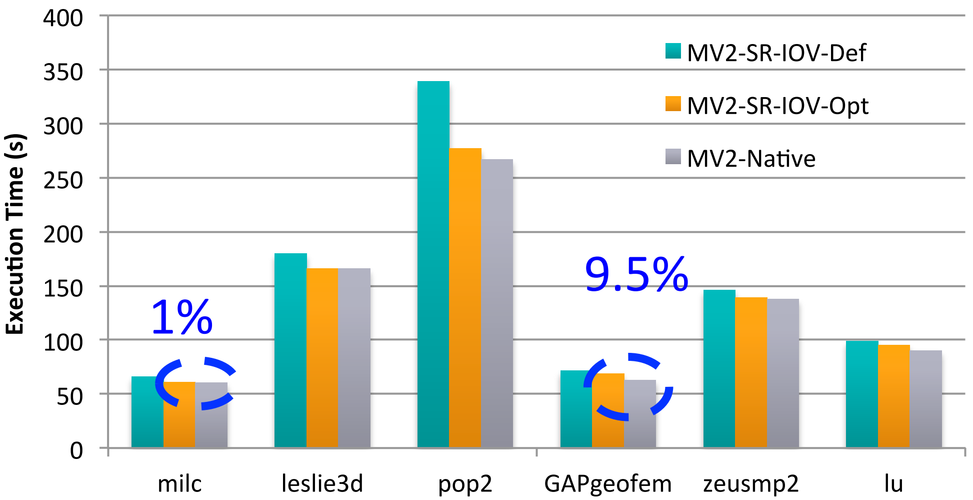

Graph500

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

# 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

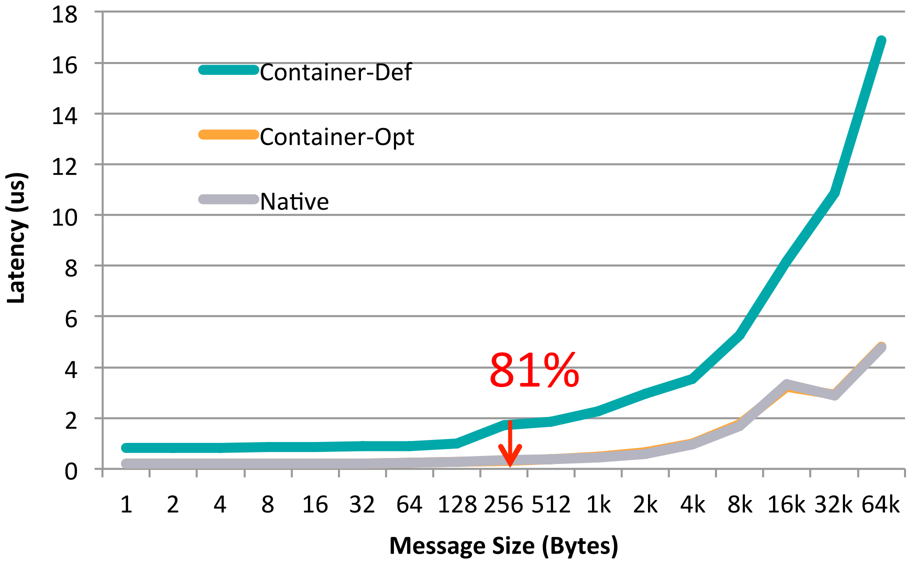
# Three Designs

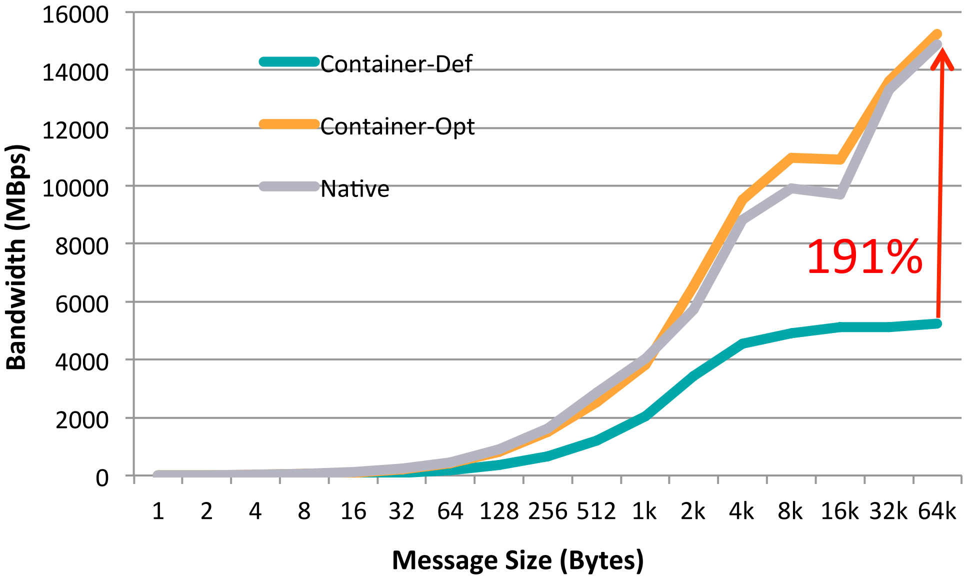
- MVAPICH2-Virt with SR-IOV and IVSHMEM
  - Standalone, OpenStack
- MVAPICH2-Virt on SLURM
- MVAPICH2 with Containers

# Containers-based Design: Issues, Challenges, and Approaches

- Container-based technologies (such as Docker) provide lightweight virtualization solutions
- What are the performance bottlenecks when running MPI applications on multiple containers per host in HPC cloud?
- Can we propose a new design to overcome the bottleneck on such container-based HPC cloud?
- Can optimized design deliver near-native performance for different container deployment scenarios?
- Locality-aware based design to enable CMA and Shared memory channels for MPI communication across co-resident containers

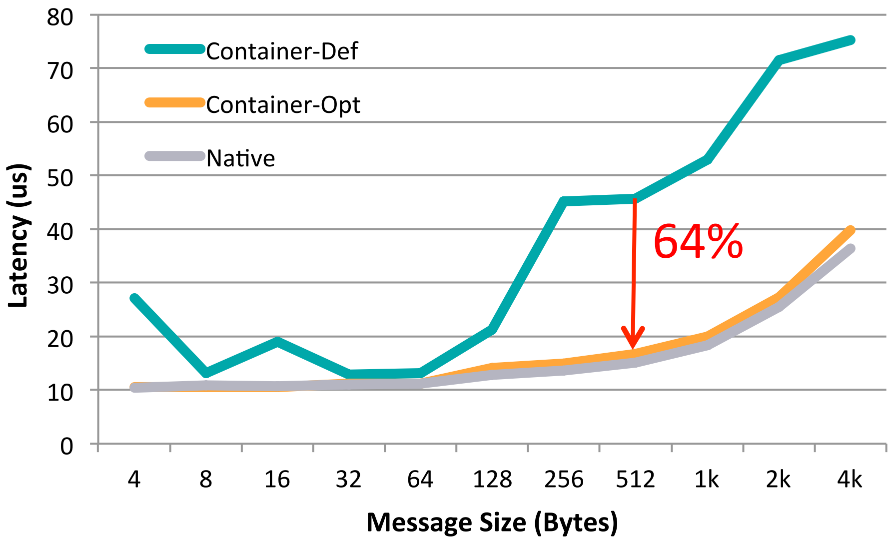
# Containers Support: MVAPICH2 Intra-node Inter-Container Point-to-Point Performance on Chameleon



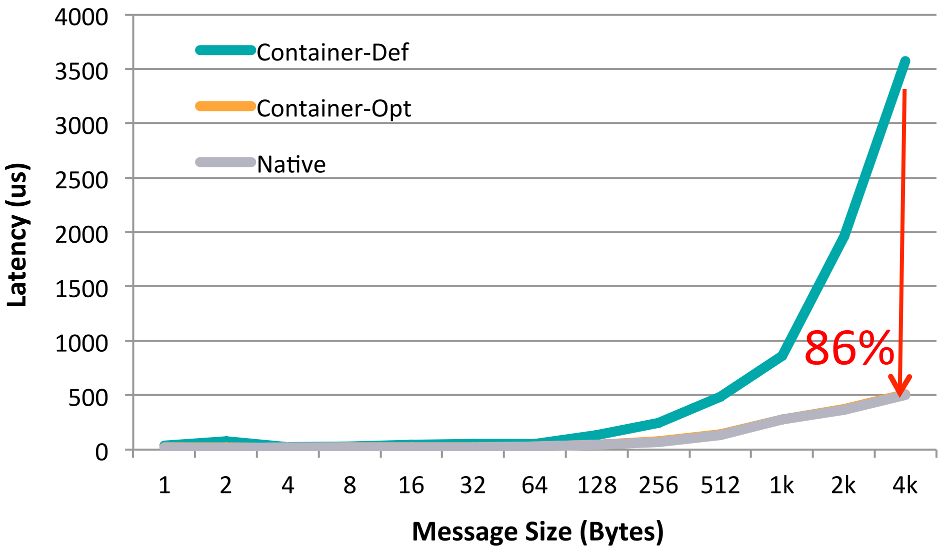




- Intra-Node Inter-Container
- Compared to Container-Def, up to 81% and 191% improvement on Latency and BW
- Compared to Native, minor overhead on Latency and BW

# Containers Support: MVAPICH2 Collective Performance on Chameleon



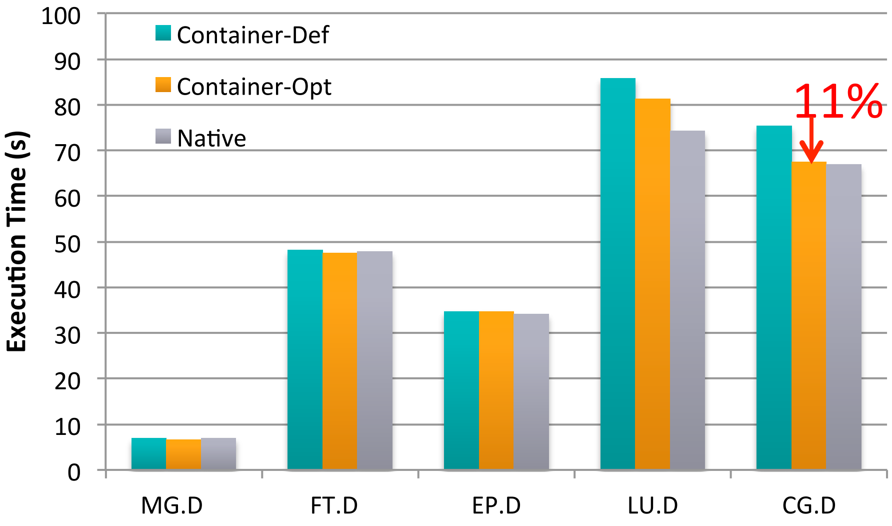

Allreduce



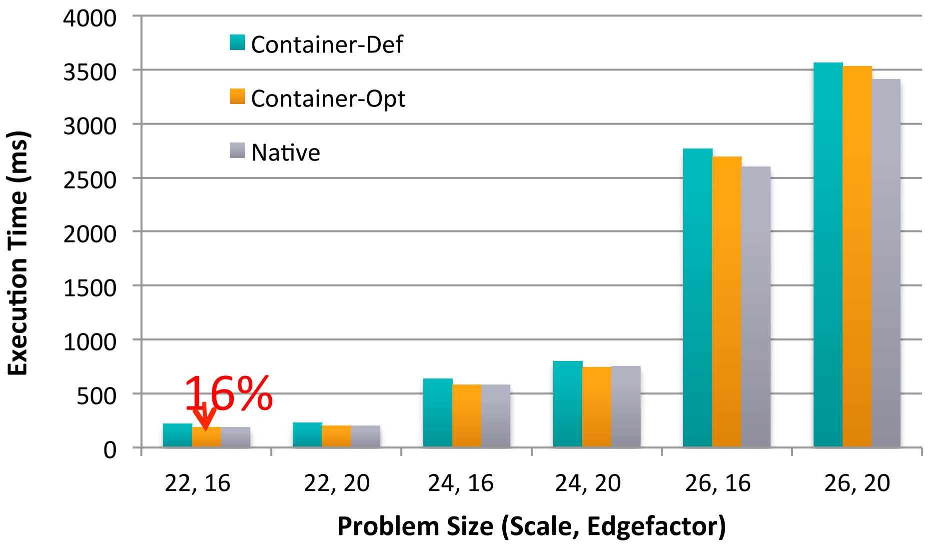

Allgather

- 64 Containers across 16 nodes, pinning 4 Cores per Container
- Compared to Container-Def, up to 64% and 86% improvement on Allreduce and Allgather
- Compared to Native, minor overhead on Allreduce and Allgather

# 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**

# Conclusions

- MVAPICH2-Virt with SR-IOV and IVSHMEM is an efficient approach to build HPC Clouds
  - Standalone
  - OpenStack
- Building HPC Clouds with MVAPICH2-Virt on SLURM is possible
- Containers-based design for MPAPICH2-Virt
- Very little overhead with virtualization, near native performance at application level
- Much better performance than Amazon EC2
- **MVAPICH2-Virt 2.1** is available for building HPC Clouds
  - SR-IOV, IVSHMEM, OpenStack
- Future releases for supporting running MPI jobs in VMs/Containers with SLURM

# Thank You!

panda@cse.ohio-state.edu, luxi@cse.ohio-state.edu

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

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