

Supporting SR-IOV and IVSHMEM in MVAPICH2 on Slurm: Challenges and Benefits

Slurm User Group Meeting 15, Sep '15

Xiaoyi Lu, Jie Zhang, Sourav Chakraborty, Hari Subramoni,
Mark Arnold, Jonathan Perkins, Dhabaleswar K. Panda

The Ohio State University

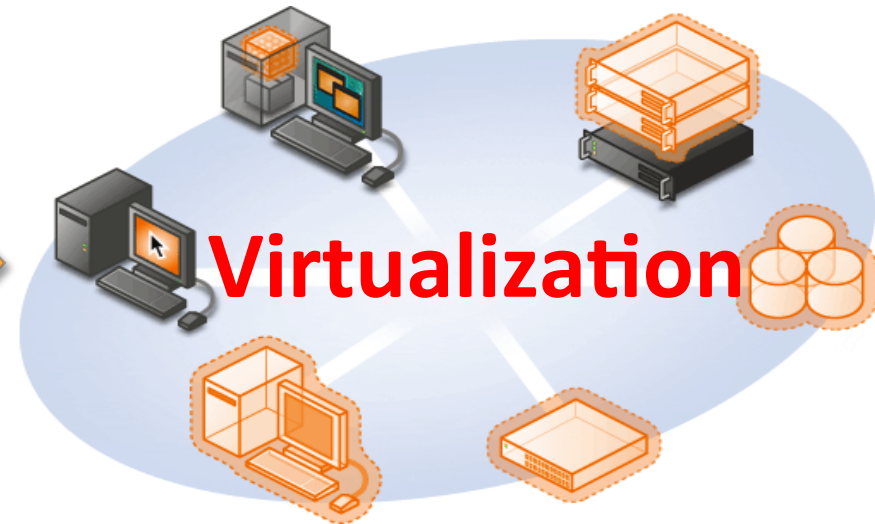
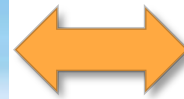
E-mail: {luxi, zhanjie, chakrabs, subramon, arnoldm, perkinjo, panda}
@cse.ohio-state.edu



Outline

- HPC Cloud with SR-IOV and InfiniBand
- MVAPICH2-Virt with SR-IOV and IVSHEM
 - Standalone, OpenStack
- MVAPICH2-Virt on Slurm
- Performance Benefits
- Conclusion

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
 - 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

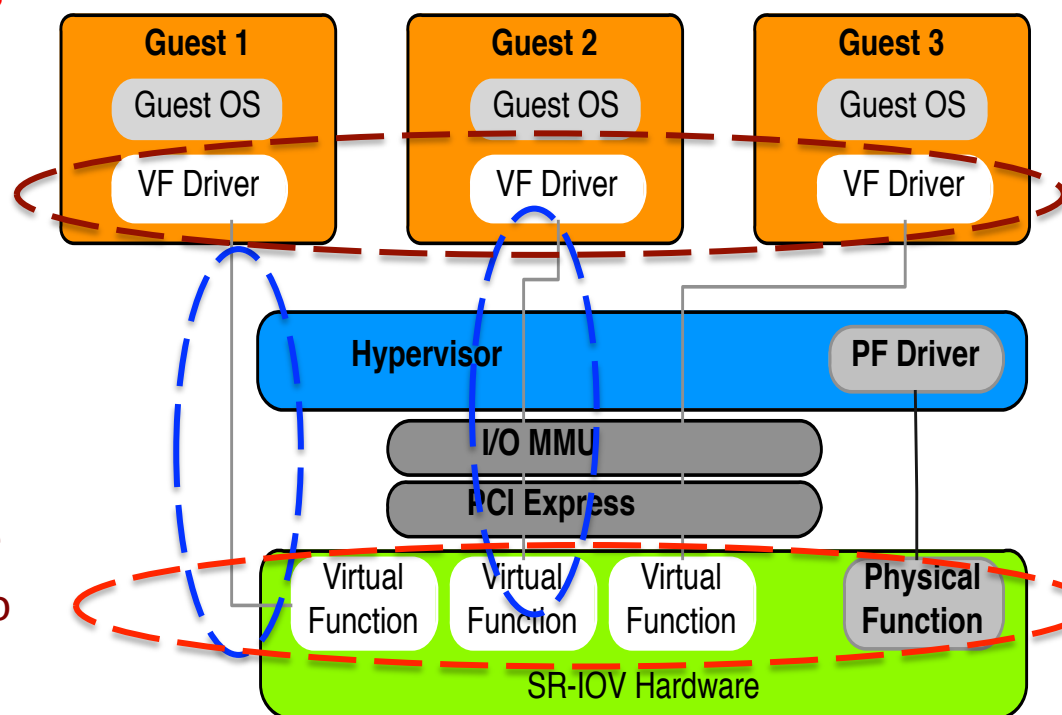


<http://www.chameleoncloud.org/>



Single Root I/O Virtualization (SR-IOV)

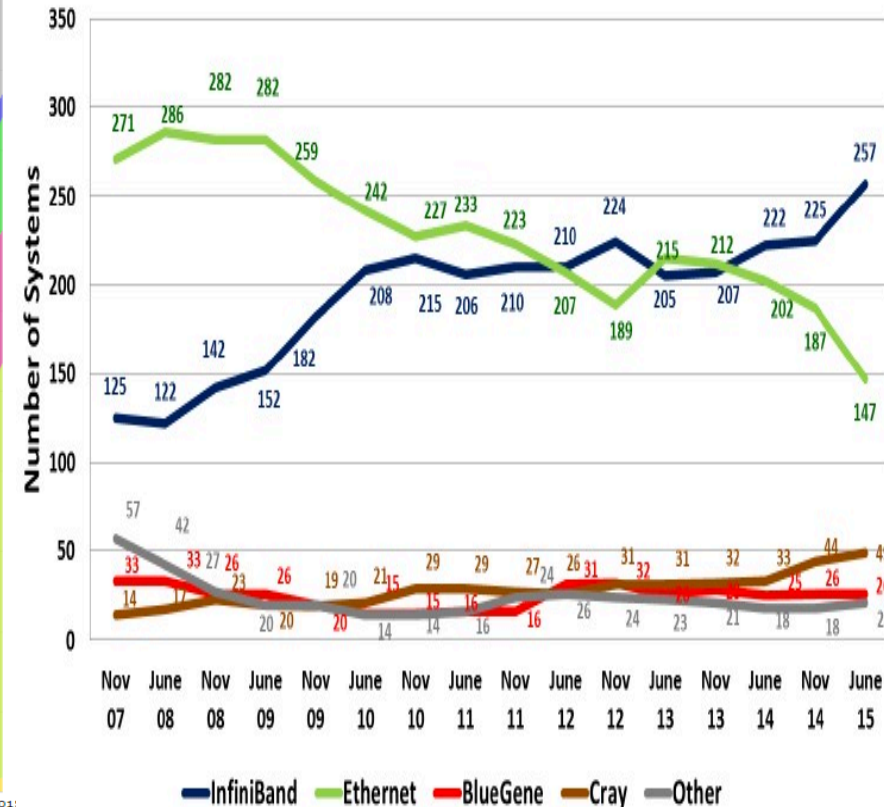
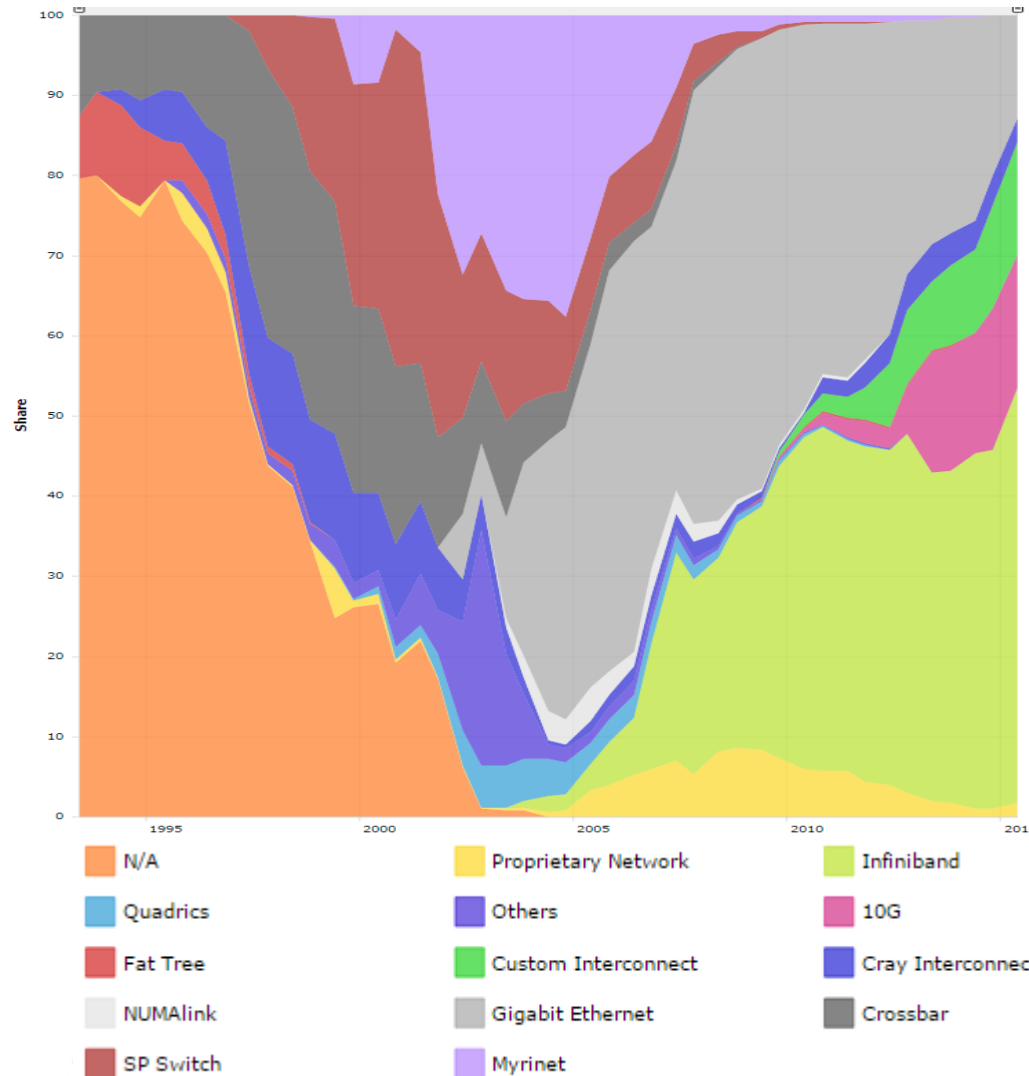
- Single Root I/O Virtualization (SR-IOV) is providing new opportunities to design HPC cloud with very little low overhead
 - Allows a single physical device, or a Physical Function (PF), to present itself as multiple virtual devices, or Virtual Functions (VFs)
 - Each VF can be dedicated to a single 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



Trends of Networking Technologies in TOP500 Systems

Percentage share of InfiniBand is steadily increasing

Interconnect Family – Systems Share



Courtesy:

<http://top500.org>

<http://www.theplatform.net/2015/07/20/ethernet-will-have-to-work-harder-to-win-hpc/>

Large-scale InfiniBand Installations

- 259 IB Clusters (51%) in the June 2015 Top500 list
(<http://www.top500.org>)
- Installations in the Top 50 (24 systems):

519,640 cores (Stampede) at TACC (8th)	76,032 cores (Tsubame 2.5) at Japan/GSIC (22 nd)
185,344 cores (Pleiades) at NASA/Ames (11 th)	194,616 cores (Cascade) at PNNL (25 th)
72,800 cores Cray CS-Storm in US (13 th)	76,032 cores (Makman-2) at Saudi Aramco (28 th)
72,800 cores Cray CS-Storm in US (14 th)	110,400 cores (Pangea) in France (29 th)
265,440 cores SGI ICE at Tulip Trading Australia (15 th)	37,120 cores (Lomonosov-2) at Russia/MSU (31 st)
124,200 cores (Topaz) SGI ICE at ERDC DSRC in US (16 th)	57,600 cores (SwiftLucy) in US (33 rd)
72,000 cores (HPC2) in Italy (17 th)	50,544 cores (Occigen) at France/GENCI-CINES (36 th)
115,668 cores (Thunder) at AFRL/USA (19 th)	76,896 cores (Salomon) SGI ICE in Czech Republic (40 th)
147,456 cores (SuperMUC) in Germany (20 th)	73,584 cores (Spirit) at AFRL/USA (42 nd)
86,016 cores (SuperMUC Phase 2) in Germany (21 st)	and many more!

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
- Building HPC Cloud with SR-IOV and InfiniBand for delivering optimal performance

Outline

- HPC Cloud with SR-IOV and InfiniBand
- MVAPICH2-Virt with SR-IOV and IVSHMEM
 - Standalone, OpenStack
- MVAPICH2-Virt on Slurm
- Performance Benefits
- Conclusion

MVAPICH2 Software

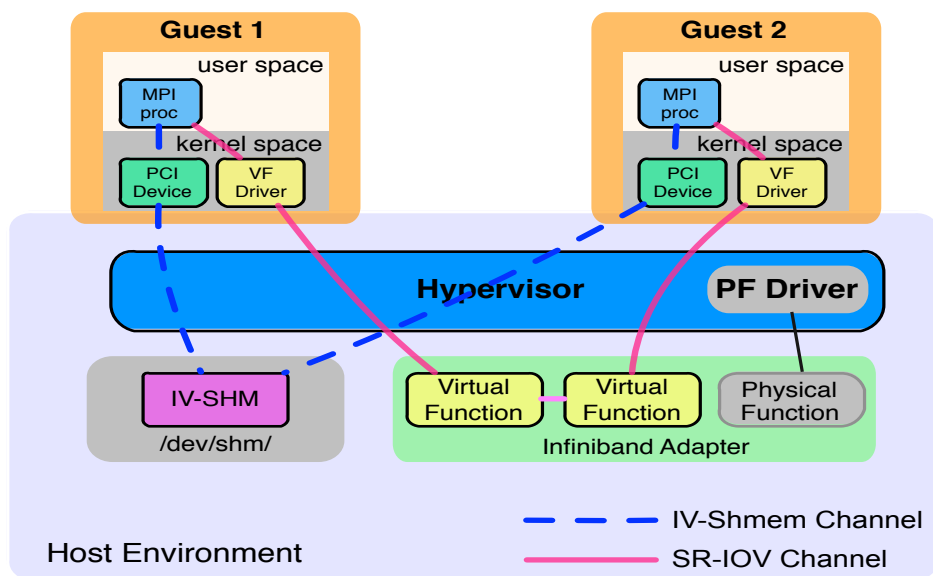
- **High Performance open-source MPI Library for InfiniBand, 10Gig/iWARP, and RoCE**
 - MVAPICH (MPI-1) , Available since 2002
 - MVAPICH2 (MPI-2.2, MPI-3.0 and MPI-3.1), Available since 2004
 - MVAPICH2-X (Advanced MPI + PGAS), Available since 2012
 - Support for GPGPUs (MVAPICH2-GDR), Available since 2014
 - Support for MIC (MVAPICH2-MIC), Available since 2014
 - **Support for Virtualization (MVAPICH2-Virt), Available since 2015**
 - Support for Energy-Aware MPI communications (MVAPICH2-EA), available since 2015
 - Used by more than 2,450 organizations in 76 countries
 - More than 285,000 downloads from the OSU site directly
 - Empowering many TOP500 clusters (Jun'15 ranking)
 - 8th ranked 519,640-core cluster (Stampede) at TACC
 - 11th ranked 185,344-core cluster (Pleiades) at NASA
 - 22nd ranked 76,032-core cluster (Tsubame 2.5) at Tokyo Institute of Technology and many others
 - Available with software stacks of many IB, HSE, and server vendors including 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 (8th in Jun'15, 462,462 cores, 5.168 Plops)

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

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

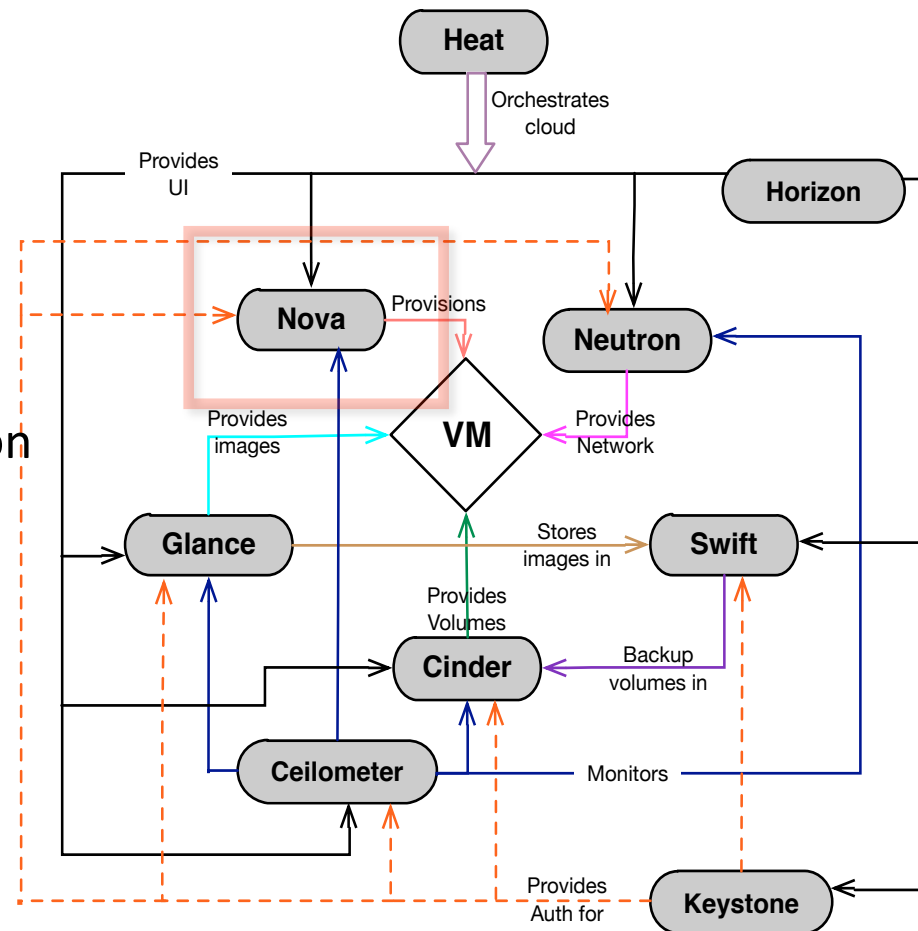


J. Zhang, X. Lu, J. Jose, R. Shi, D. K. Panda. Can Inter-VM Shmem Benefit MPI Applications on SR-IOV based Virtualized InfiniBand Clusters? **Euro-Par**, 2014.

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.

Outline

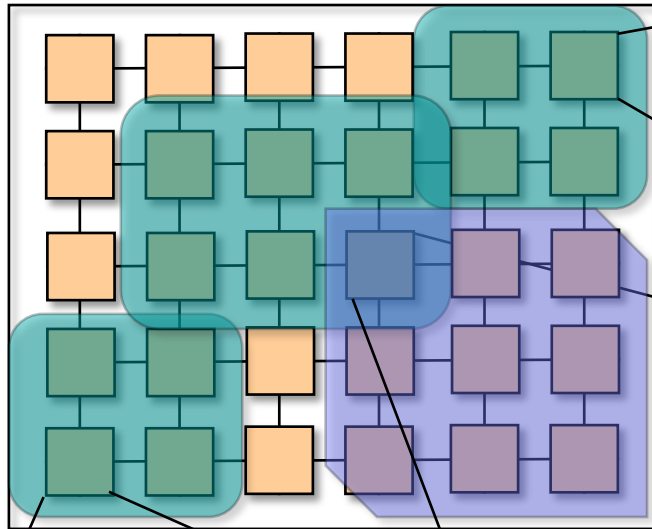
- HPC Cloud with SR-IOV and InfiniBand
- MVAPICH2-Virt with SR-IOV and IVSHMEM
 - Standalone, OpenStack
- MVAPICH2-Virt on Slurm
- Performance Benefits
- Conclusion

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

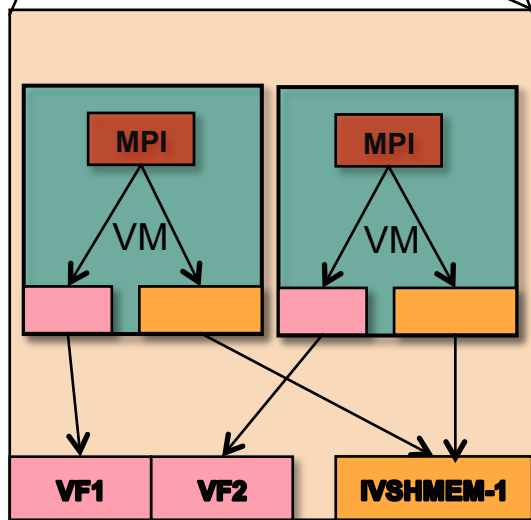
Compute Nodes



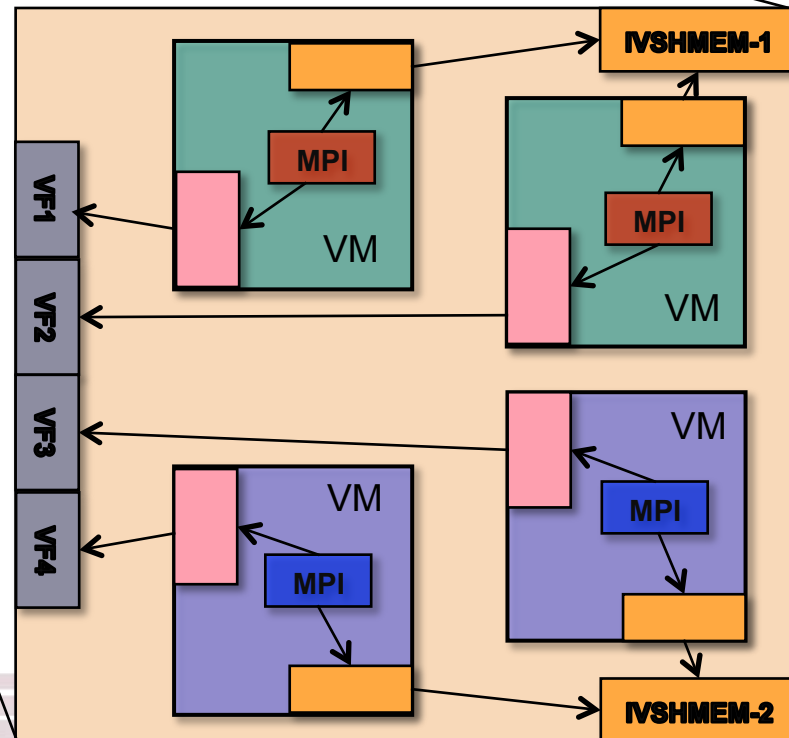
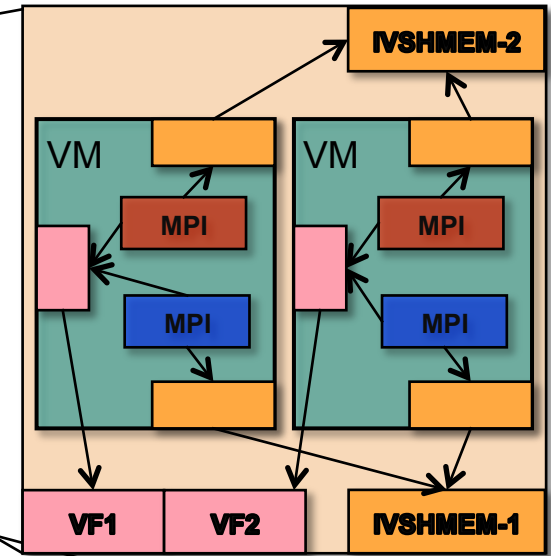
Exclusive Allocation
Concurrent Jobs

Shared-host Allocations
Concurrent Jobs

Exclusive Allocation
Sequential Job



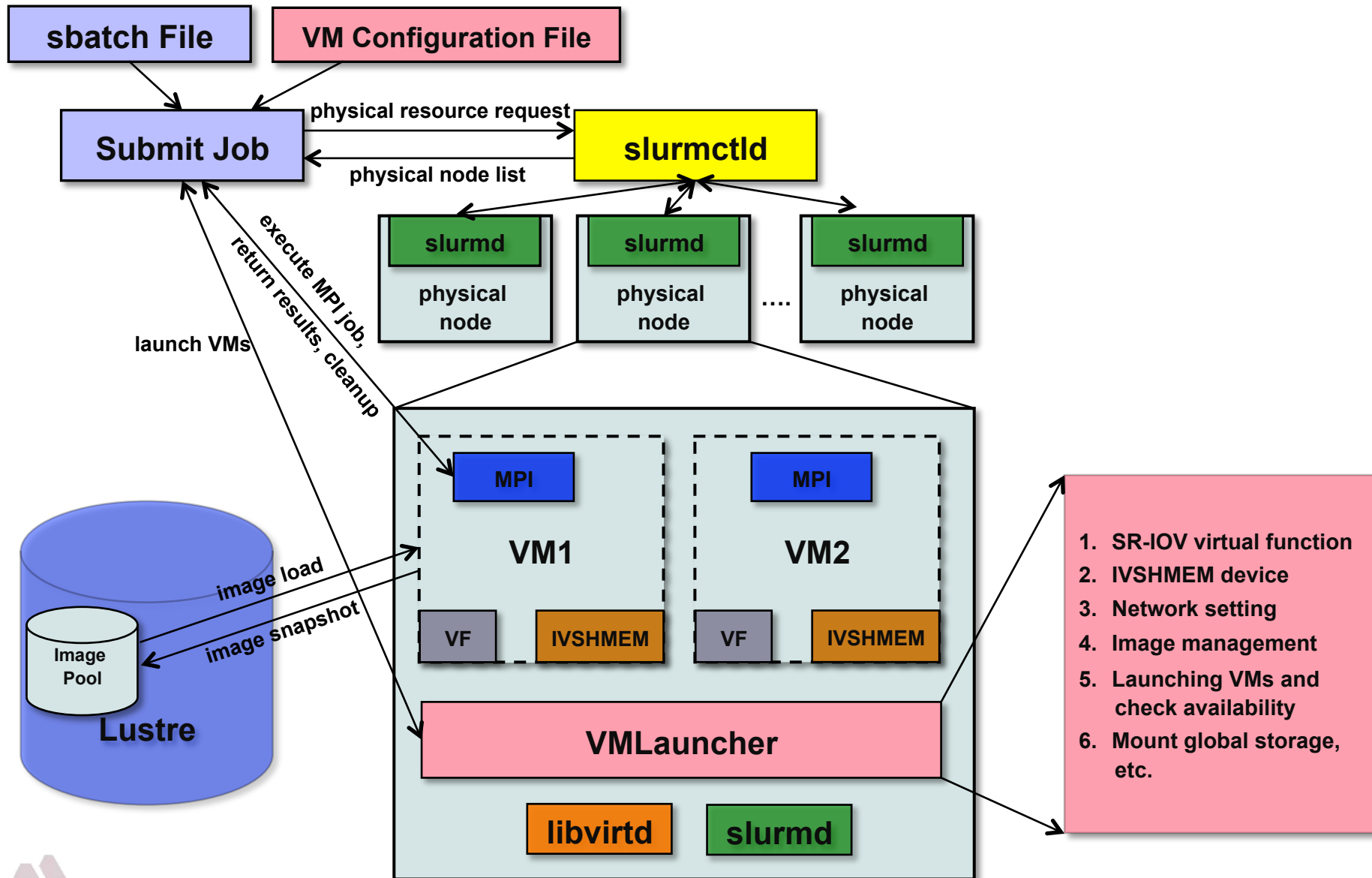
SLUG '15



Needs 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



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

Outline

- HPC Cloud with SR-IOV and InfiniBand
- MVAPICH2-Virt with SR-IOV and IVSHMEM
 - Standalone, OpenStack
- MVAPICH2-Virt on Slurm
- Performance Benefits
- Conclusion

Experimental Setup

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 ixgbev SR-IOV driver ^[2]	

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

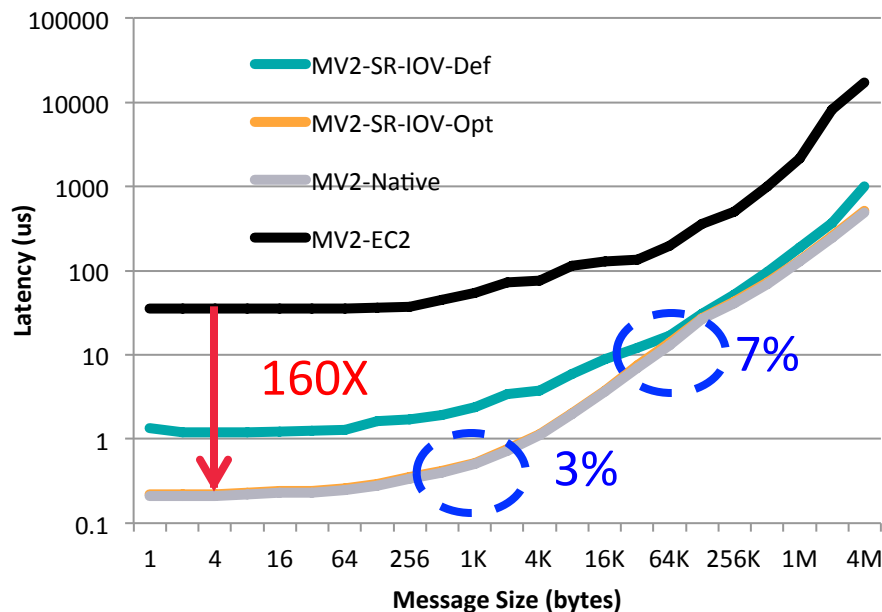
[2] **Nowlab Cloud is using InfiniBand FDR (56Gbps), while Amazon EC2 C3 instances are using 10 GigE. Both have SR-IOV support.**

Experiments Carried Out

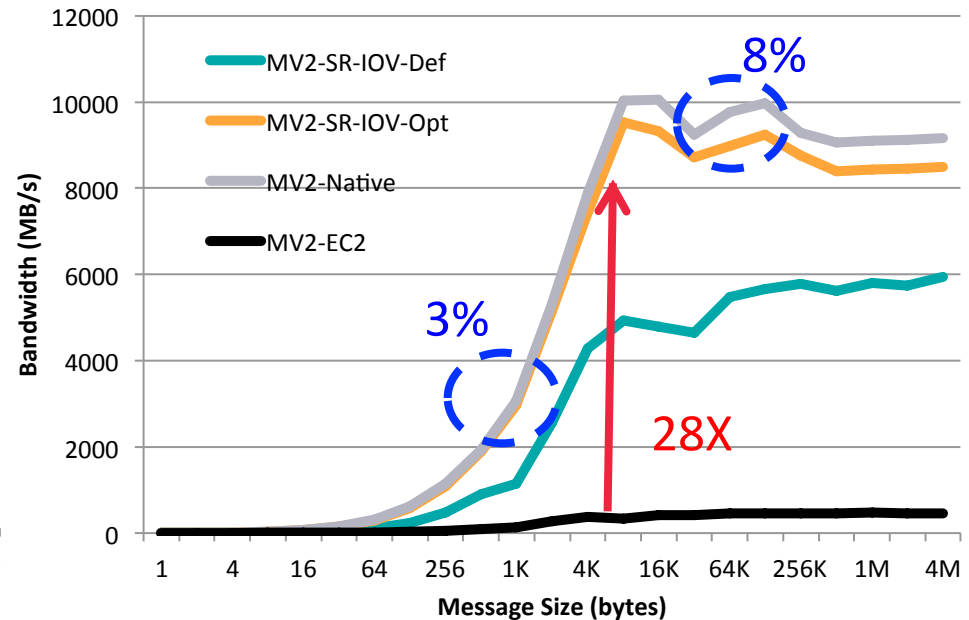
- Point-to-point
 - Two-sided and One-sided
 - Latency and Bandwidth
 - Intra-node and Inter-node ^[1]
- Applications
 - NAS and Graph500

[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)



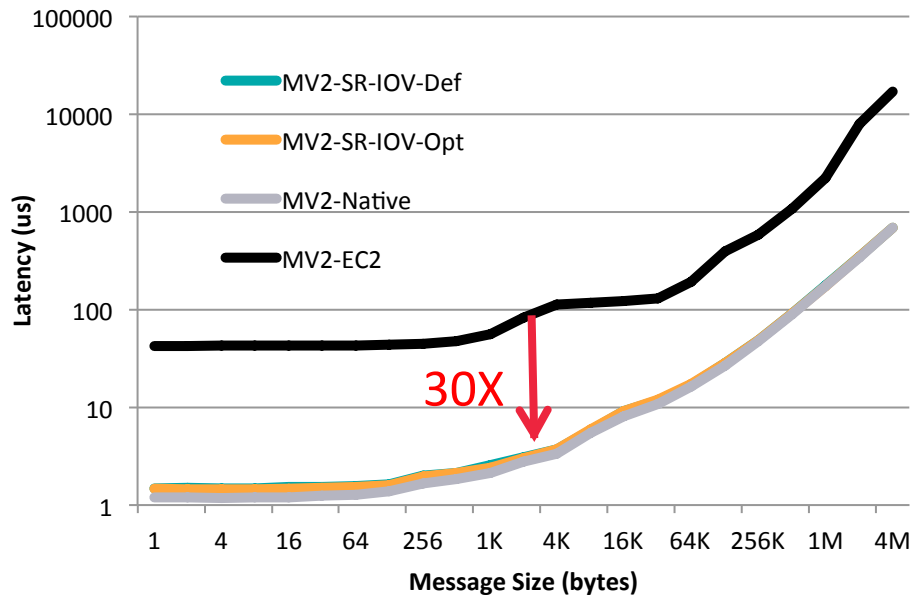
Intra-node Inter-VM pt2pt Latency



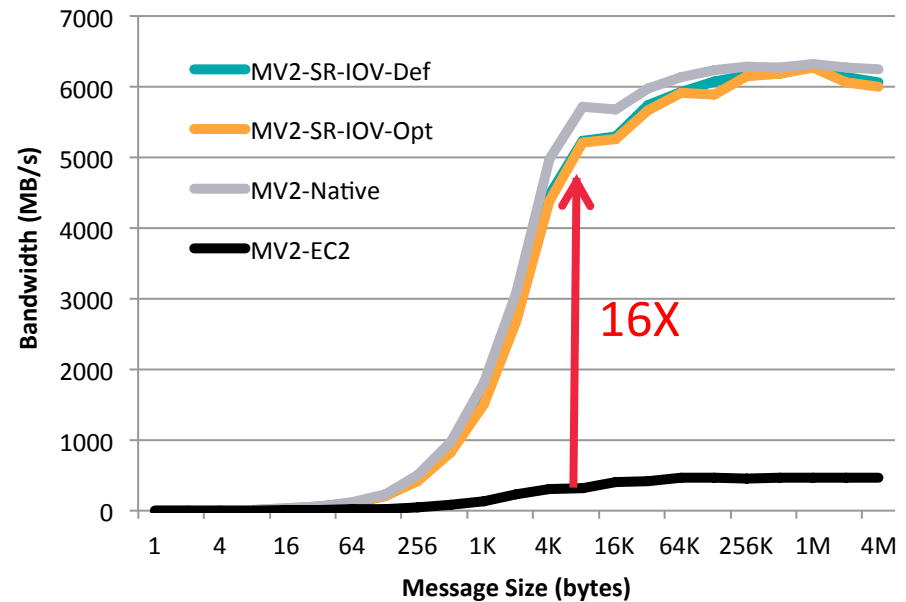
Intra-node Inter-VM pt2pt Bandwidth

- 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)



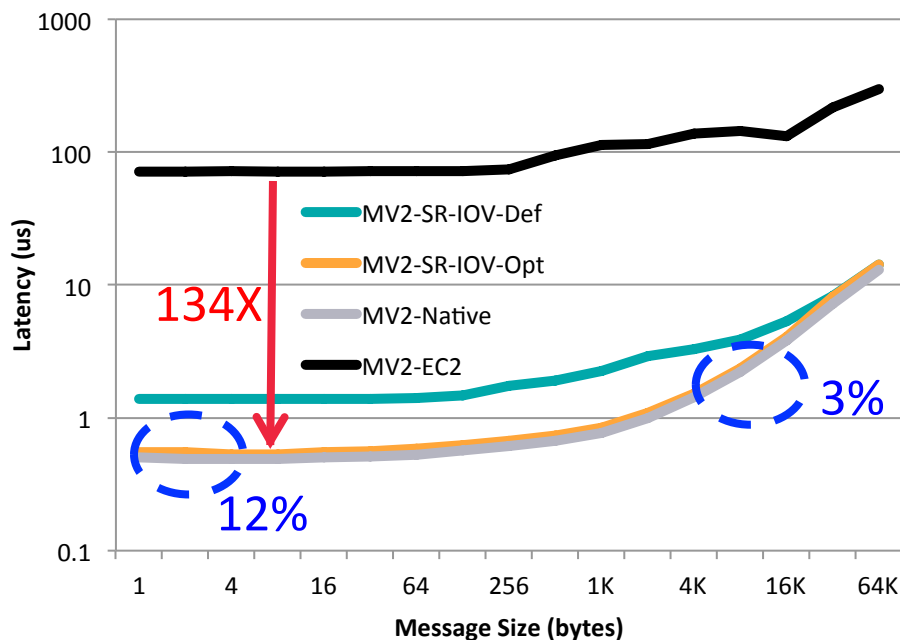
Inter-node Inter-VM pt2pt Latency



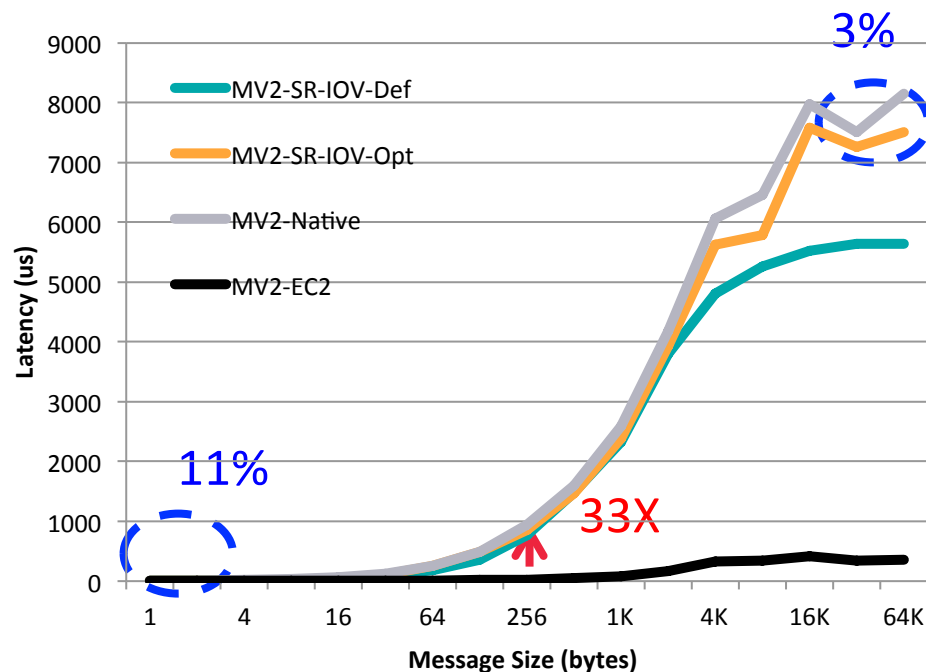
Inter-node Inter-VM pt2pt Bandwidth

- 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

One-Sided Put Performance – Latency & Bandwidth (Intra-node)



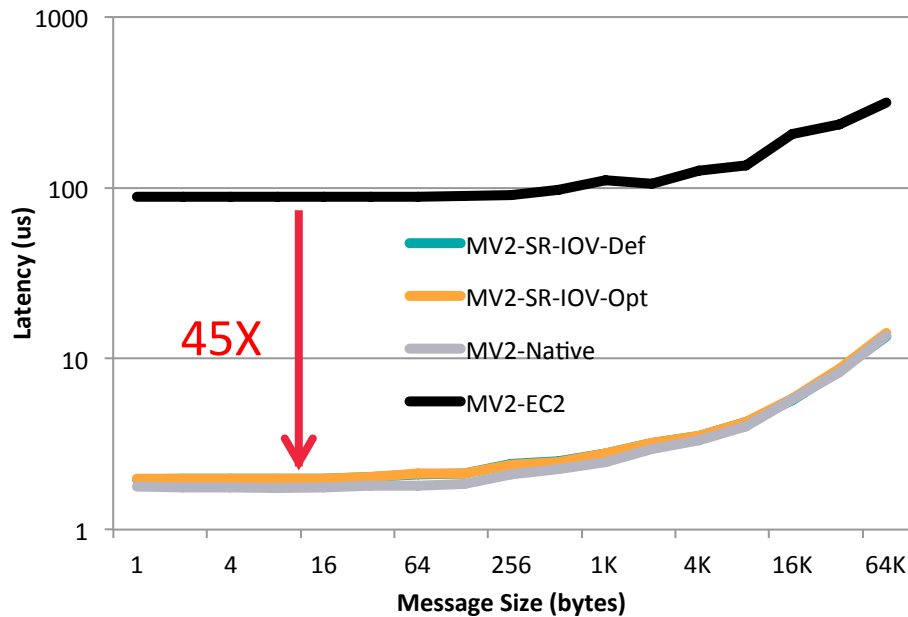
Intra-node Inter-VM One-Sided Put Latency



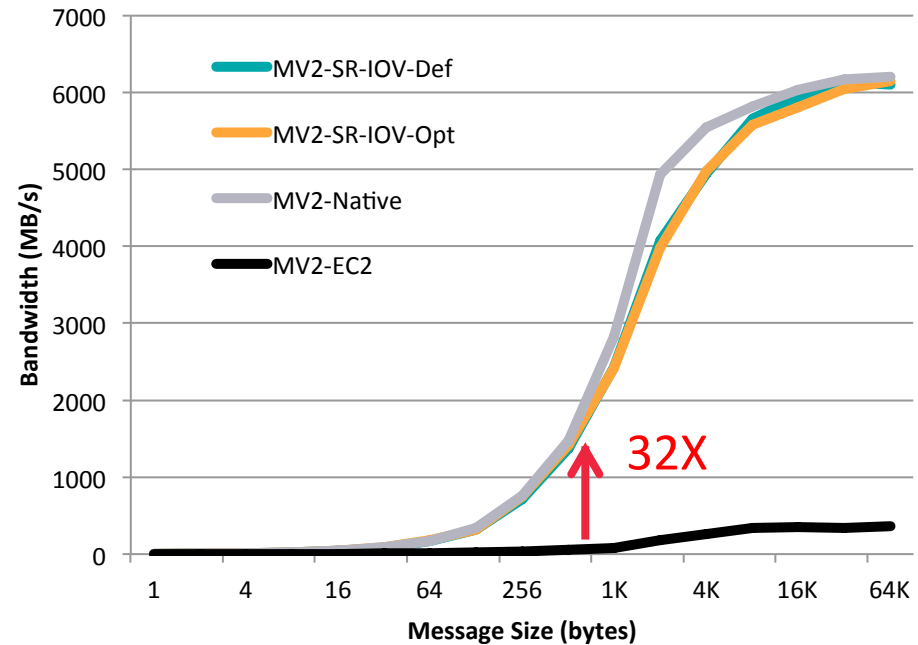
Intra-node Inter-VM One-Sided Put Bandwidth

- EC2 C3.2xlarge instances
- Compared to SR-IOV-Def, up to 63% and 42% performance improvement on Lat & BW
- Compared to Native, 3%-12% overhead for Lat, and 3%-11% overhead for BW
- Compared to EC2, up to 134X and 33X performance speedup on Lat & BW

One-Sided Put Performance – Latency & Bandwidth (Inter-node)



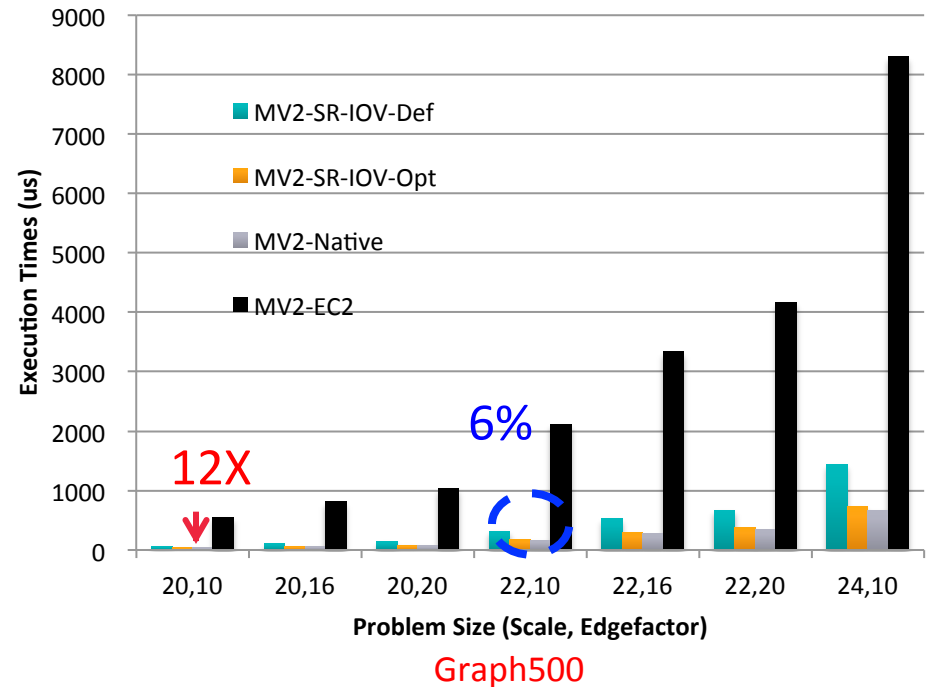
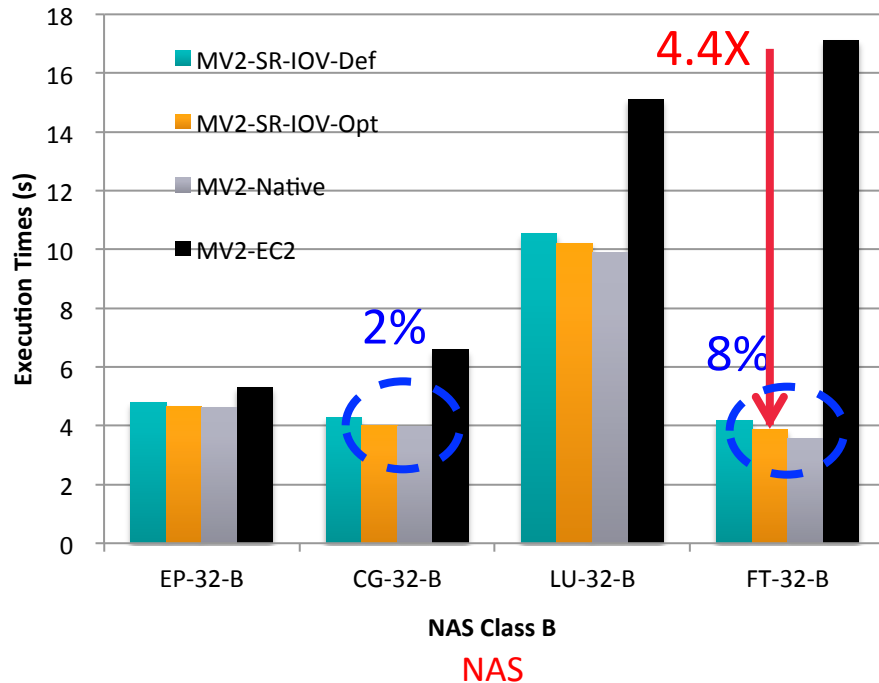
Inter-node Inter-VM One-Sided Put Latency



Inter-node Inter-VM One-Sided Put 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 45X and 32X performance speedup on Lat & BW

Application Performance (4 VM * 8 Core/VM)



- EC2 C3.2xlarge instances
- Compared to Native, **2%-8%** overhead for NAS, **around 6%** overhead for Graph500
- Compared to EC2, up to **4.4X** (FT) speedup for NAS, up to **12X** (20,10) speedup for Graph500

Outline

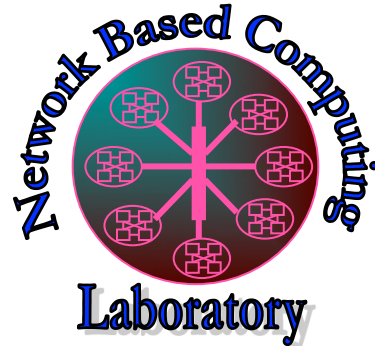
- HPC Cloud with SR-IOV and InfiniBand
- MVAPICH2 with SR-IOV
 - Standalone, OpenStack
- MVAPICH2-Virt on Slurm
- Performance Benefits
- Conclusion

Conclusion

- 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
- Performance numbers are promising
- Much better performance than Amazon EC2
- Near native performance at application level
- **MVAPICH2-Virt 2.1** is released!
 - SR-IOV, IVSHMEM, OpenStack
 - <http://mvapich.cse.ohio-state.edu/>
- Future releases for supporting running MPI jobs in VMs/Containers with Slurm

Thank You!

{luxi, zhanjie, chakrabs, subramon, arnoldm, perkinjo, panda}
@cse.ohio-state.edu



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



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