MVAPICH

MPI, PGAS and Hybrid MPI+PGAS Library

HiBD

High-Performance Big Data

Building Efficient Clouds for HPC, Big Data, and Neuroscience Applications over SR-IOV-enabled InfiniBand Clusters

Talk at OFA Workshop 2018

by

Xiaoyi Lu

The Ohio State University

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

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

HPC Meets Cloud Computing

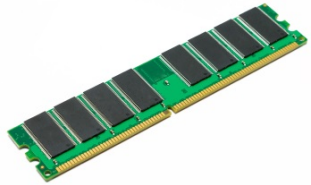
- Cloud Computing widely adopted in industry computing environment
- Cloud Computing provides high resource utilization and flexibility
- Virtualization is the key technology to enable Cloud Computing
- Intersect360 study shows cloud is the fastest growing class of HPC
- HPC Meets Cloud: The convergence of Cloud Computing and HPC

Drivers of Modern HPC Cluster and Cloud Architecture

Multi-/Many-core Processors

High Performance Interconnects – InfiniBand (with SR-IOV)
<1usec latency, 200Gbps Bandwidth>

SSDs, Object Storage Clusters



Large memory nodes
(Upto 2 TB)

- Multi-core/many-core technologies, Accelerators
- Large memory nodes
- Solid State Drives (SSDs), NVM, Parallel Filesystems, Object Storage Clusters
- Remote Direct Memory Access (RDMA)-enabled networking (InfiniBand and RoCE)
- Single Root I/O Virtualization (SR-IOV)

SDSC Comet

TACC Stampede

Microsoft Azure

amazon web services EC2

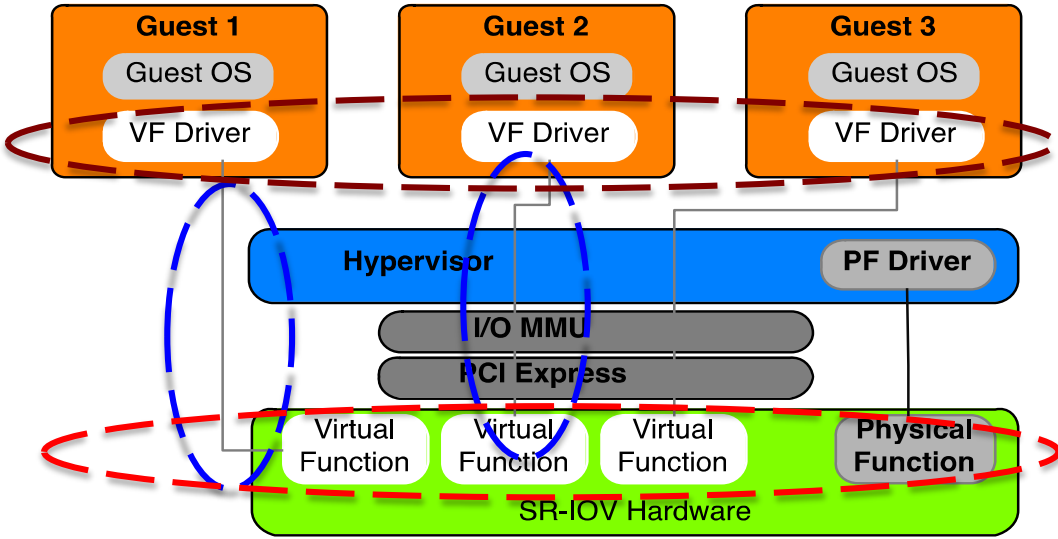
ORACLE Cloud

Chameleon Cloud

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



Broad Challenges of Building Efficient HPC Clouds

- Virtualization Support with Virtual Machines and Containers
  - KVM, Docker, Singularity, etc.
- Communication coordination among optimized communication channels on Clouds
  - SR-IOV, IVShmem, IPC-Shm, CMA, etc.
- Locality-aware communication
- 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 nodes (128-1024 cores)
  - Multiple end-points per node
- NUMA-aware communication for nested virtualization
- Integrated Support for GPGPUs and Accelerators
- Fault-tolerance/resiliency
  - Migration support with virtual machines
- QoS support for communication and I/O
- Support for Hybrid MPI+PGAS programming (MPI + OpenMP, MPI + UPC, MPI + OpenSHMEM, MPI+UPC++, CAF, …)
- Energy-Awareness
- Co-design with resource management and scheduling systems on Clouds
  - OpenStack, Slurm, etc.

Approaches to Build HPC Clouds

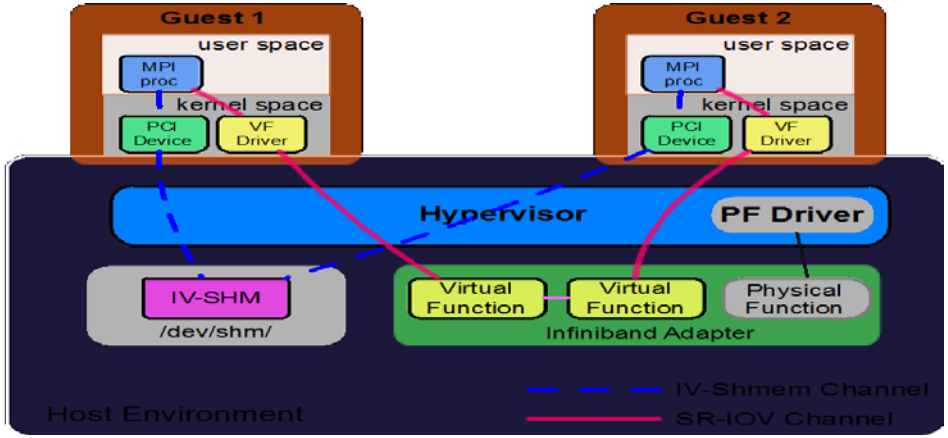
- MVAPICH2-Virt with SR-IOV and IVSHMEM
  - Standalone, OpenStack
- SR-IOV-enabled VM Migration Support in MVAPICH2
- MVAPICH2 with Containers (Docker and Singularity)
- MVAPICH2 with Nested Virtualization (Container over VM)
- MVAPICH2-Virt on SLURM
  - SLURM alone, SLURM + OpenStack
- Neuroscience Applications on HPC Clouds
- Big Data Libraries on Cloud
  - RDMA-Hadoop, OpenStack Swift

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-3.1), 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 2,875 organizations in 86 countries**
  - **More than 462,000 (> 0.46 million) downloads from the OSU site directl**
  - Empowering many TOP500 clusters (Nov '17 ranking)
    - **1st, 10,649,600-core (Sunway TaihuLight) at National Supercomputing Center in Wuxi, China**
    - 9th, 556,104 cores (Oakforest-PACS) in Japan
    - 12th, 368,928-core (Stampede2) at TACC
    - 17th, 241,108-core (Pleiades) at NASA
    - 48th, 76,032-core (Tsubame 2.5) at Tokyo Institute of Technology
  - 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

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
- **Support deployment with OpenStack**

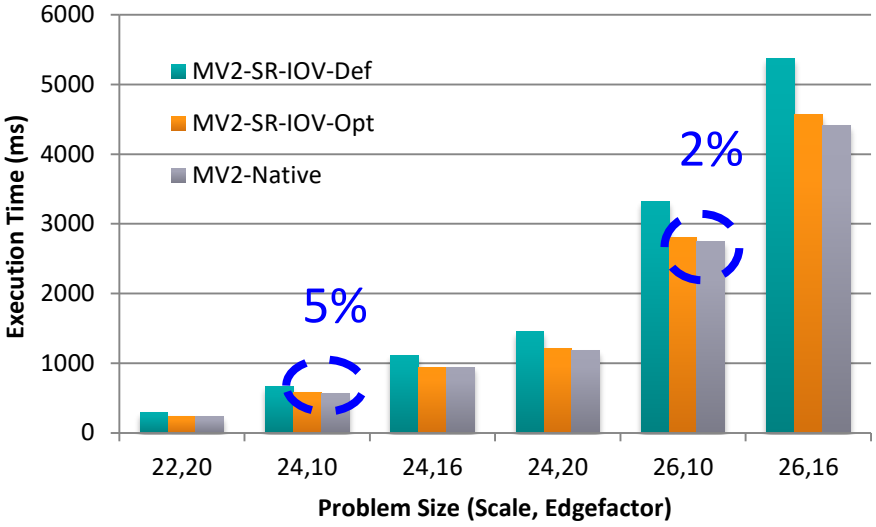


J. Zhang, X. Lu, J. Jose, R. Shi, D. K. Panda. Can Inter-VM Shmem Benefit MPI Applications on SR-IOV based Virtualized InfiniBand Clusters? Euro-Par, 2014

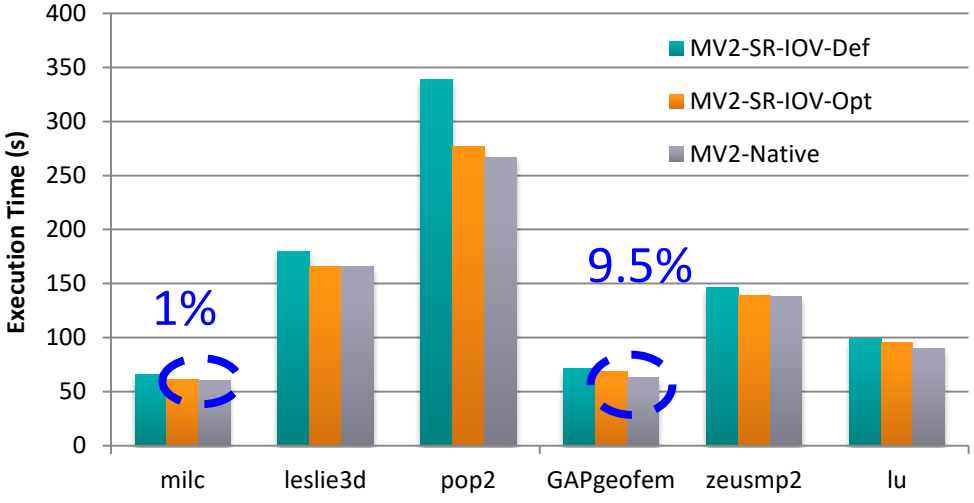
J. Zhang, X. Lu, J. Jose, R. Shi, M. Li, D. K. Panda. High Performance MPI Library over SR-IOV Enabled InfiniBand Clusters. HiPC, 2014

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

Approaches to Build HPC Clouds

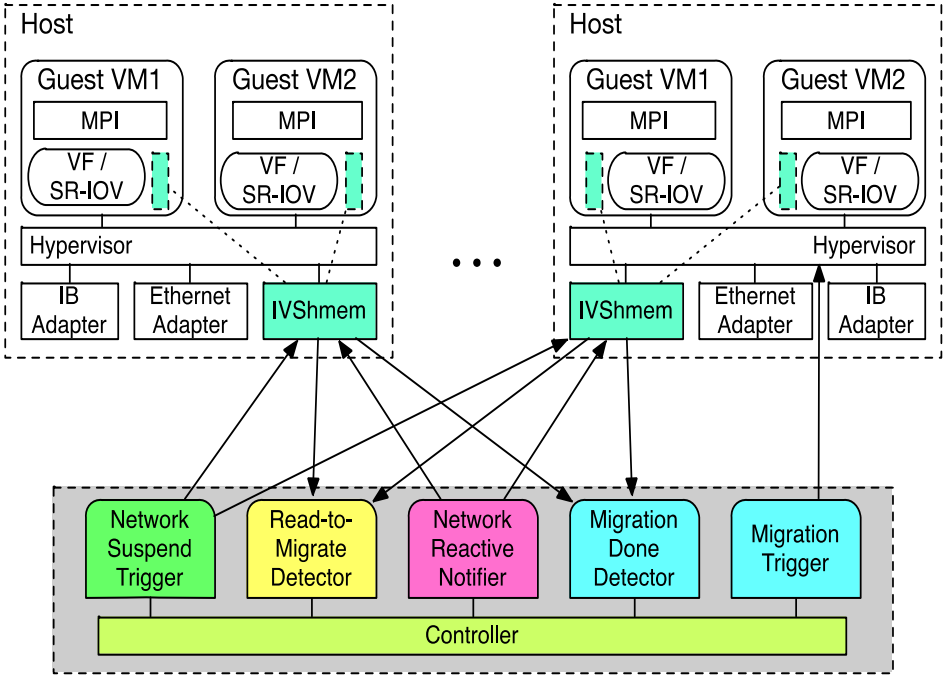
- MVAPICH2-Virt with SR-IOV and IVSHMEM
  - Standalone, OpenStack
- SR-IOV-enabled VM Migration Support in MVAPICH2
- MVAPICH2 with Containers (Docker and Singularity)
- MVAPICH2 with Nested Virtualization (Container over VM)
- MVAPICH2-Virt on SLURM
  - SLURM alone, SLURM + OpenStack
- Neuroscience Applications on HPC Clouds
- Big Data Libraries on Cloud
  - RDMA-Hadoop, OpenStack Swift

Execute Live Migration with SR-IOV Device

```
[root@sandy1:migration]$
[root@sandy1:migration]$ssh sandy3-vm1 lspci
root@sandy3-vm1's password:
00:00.0 Host bridge: Intel Corporation 440FX - 82441FX PMC [Natoma] (rev 02)
00:01.0 ISA bridge: Intel Corporation 82371SB PIIX3 ISA [Natoma/Triton II]
00:01.1 IDE interface: Intel Corporation 82371SB PIIX3 IDE [Natoma/Triton II]
00:01.2 USB controller: Intel Corporation 82371SB PIIX3 USB [Natoma/Triton II] (rev 01)
00:01.3 Bridge: Intel Corporation 82371AB/EB/MB PIIX4 ACPI (rev 03)
00:02.0 VGA compatible controller: Cirrus Logic GD 5446
00:03.0 Ethernet controller: Red Hat, Inc Virtio network device
00:04.0 Infiniband controller: Mellanox Technologies MT27700 Family [ConnectX-4 Virtual Function]
00:05.0 Unclassified device [00ff]: Red Hat, Inc Virtio memory balloon
[root@sandy1:migration]$
[root@sandy1:migration]$
[root@sandy1:migration]$
[root@sandy1:migration]$
[root@sandy1:migration]$
[root@sandy1:migration]$virsh migrate --live --rdma-pin-all --migrateuri rdma://sandy3-ib sandy1-vm1 qemu://sandy3-ib/system
error: Requested operation is not valid: domain has assigned non-USB host devices

[root@sandy1:migration]$
```

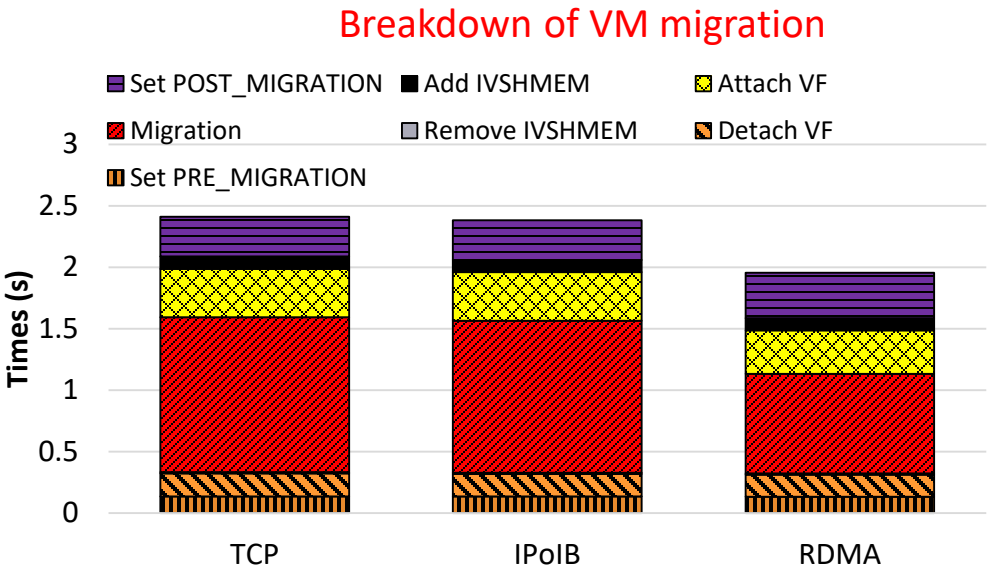
High Performance SR-IOV enabled VM Migration Support in MVAPICH2

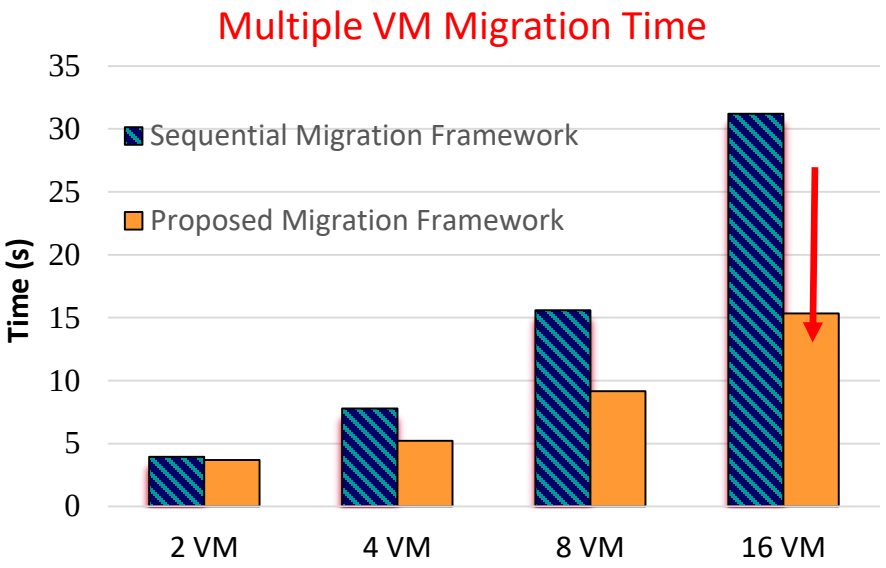


- Migration with SR-IOV device has to handle the challenges of detachment/re-attachment of virtualized IB device and IB connection
- Consist of SR-IOV enabled IB Cluster and External Migration Controller
- Multiple parallel libraries to notify MPI applications during migration (detach/reattach SR-IOV/IVShmem, migrate VMs, migration status)
- Handle the IB connection suspending and reactivating
- Propose Progress engine (PE) and migration thread based (MT) design to optimize VM migration and MPI application performance

J. Zhang, X. Lu, D. K. Panda. High-Performance Virtual Machine Migration Framework for MPI Applications on SR-IOV enabled InfiniBand Clusters. IPDPS, 2017

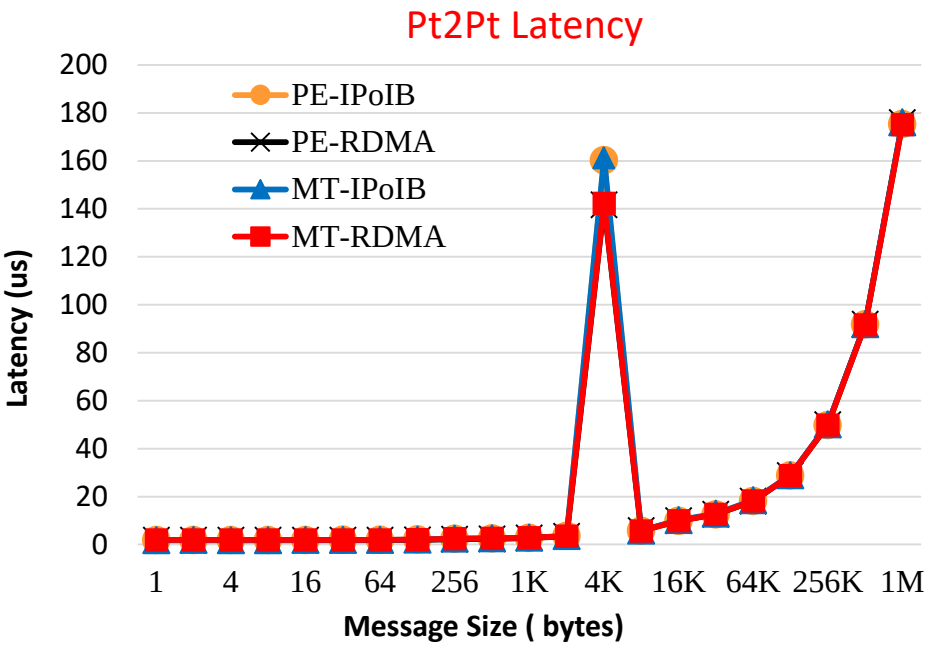
Performance Evaluation of VM Migration Framework

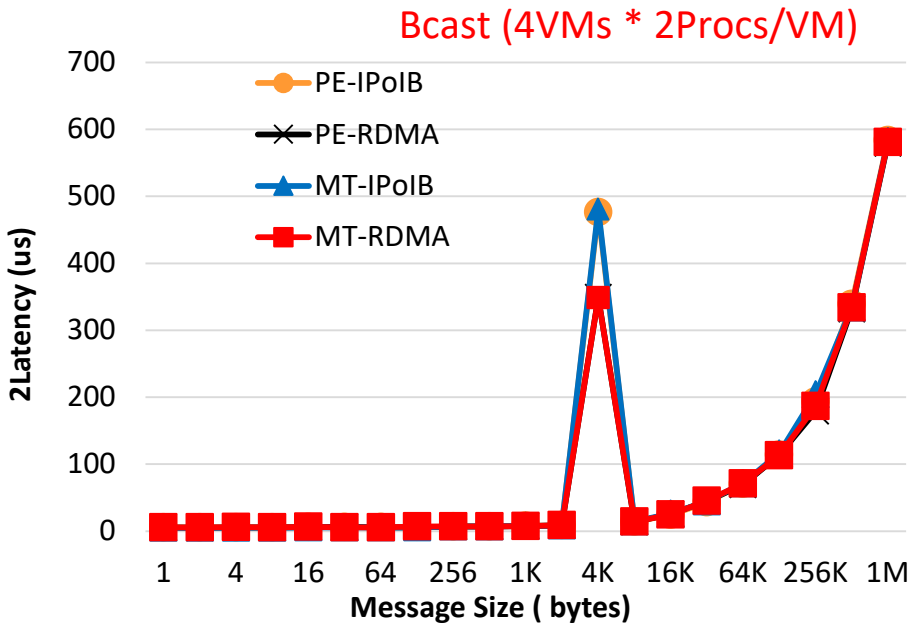




- Compared with the TCP, the RDMA scheme reduces the total migration time by 20%
- Total time is dominated by `Migration' time; Times on other steps are similar across different schemes
- Proposed migration framework could reduce up to 51% migration time

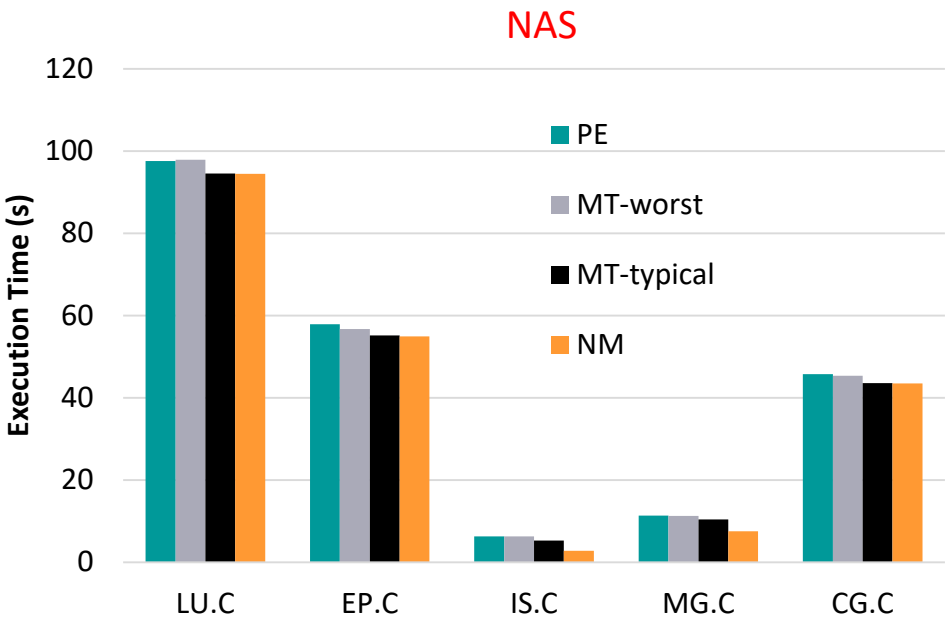
Performance Evaluation of VM Migration Framework

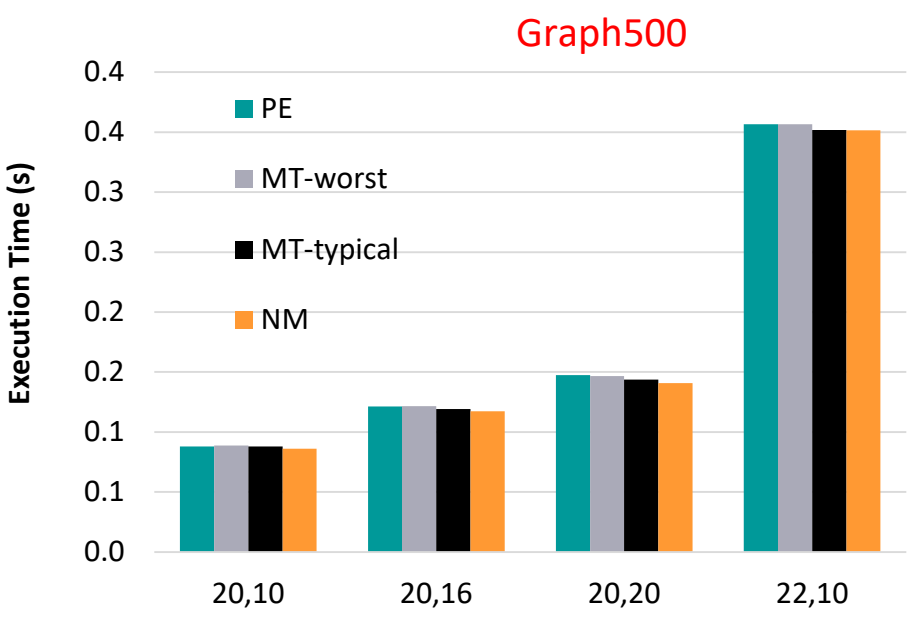




- Migrate a VM from one machine to another while benchmark is running inside
- Proposed MT-based designs perform slightly worse than PE-based designs because of lock/unlock
- No benefit from MT because of NO computation involved

Performance Evaluation of VM Migration Framework



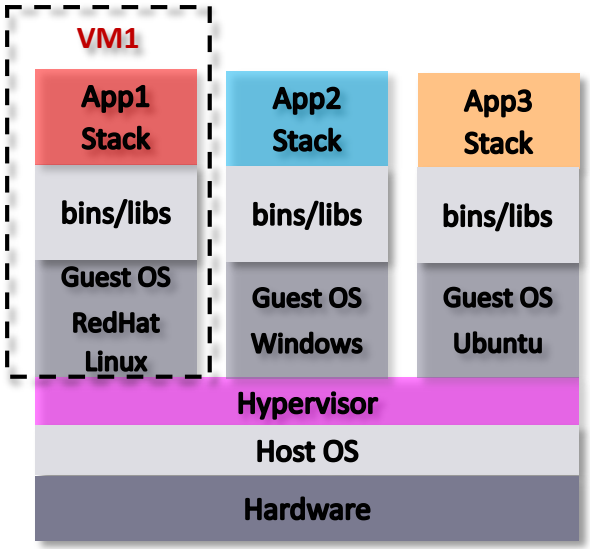


- 8 VMs in total and 1 VM carries out migration during application running
- Compared with NM, MT- worst and PE incur some overhead compared with NM
- MT-typical allows migration to be completely overlapped with computation

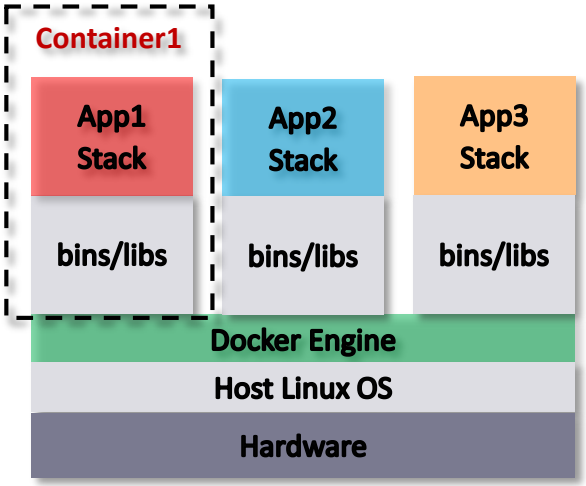
Approaches to Build HPC Clouds

- MVAPICH2-Virt with SR-IOV and IVSHMEM
  - Standalone, OpenStack
- SR-IOV-enabled VM Migration Support in MVAPICH2
- MVAPICH2 with Containers (Docker and Singularity)
- MVAPICH2 with Nested Virtualization (Container over VM)
- MVAPICH2-Virt on SLURM
  - SLURM alone, SLURM + OpenStack
- Neuroscience Applications on HPC Clouds
- Big Data Libraries on Cloud
  - RDMA-Hadoop, OpenStack Swift

Overview of Containers-based Virtualization



Hypervisor-based Virtualization

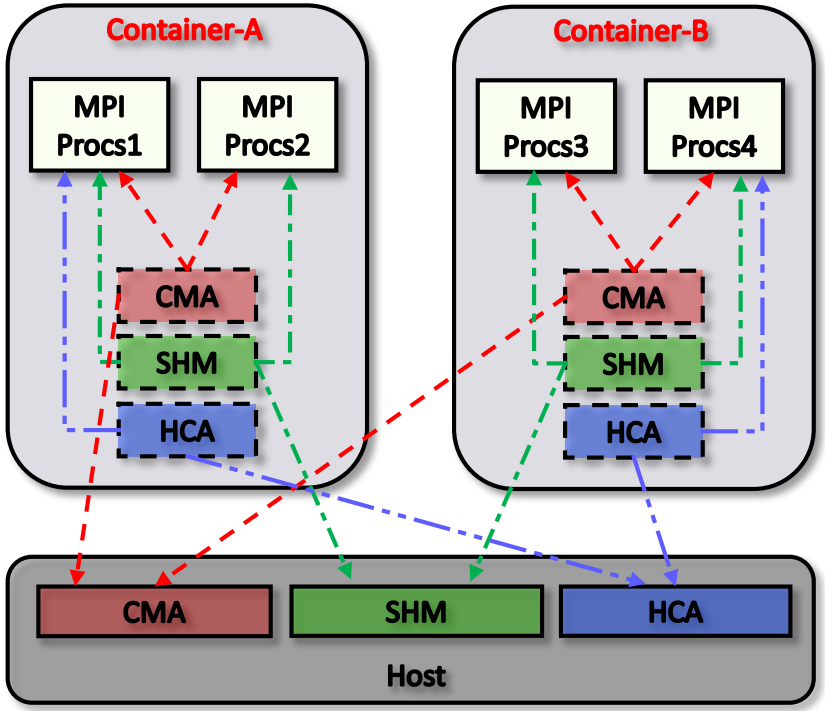


Container-based Virtualization

- Container-based technologies (e.g., Docker) provide lightweight virtualization solutions
- Container-based virtualization – share host kernel by containers

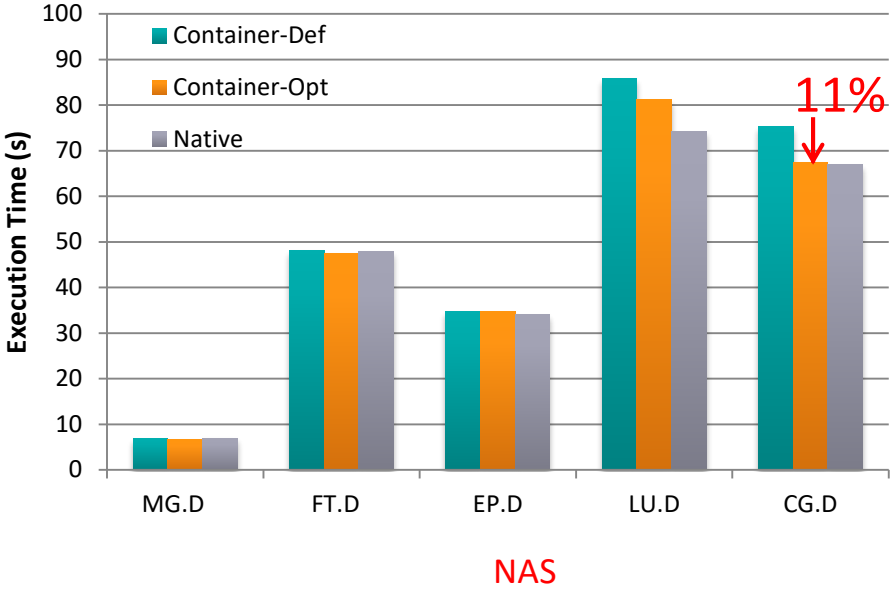
Containers-based Design: Issues, Challenges, and Approaches

- 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

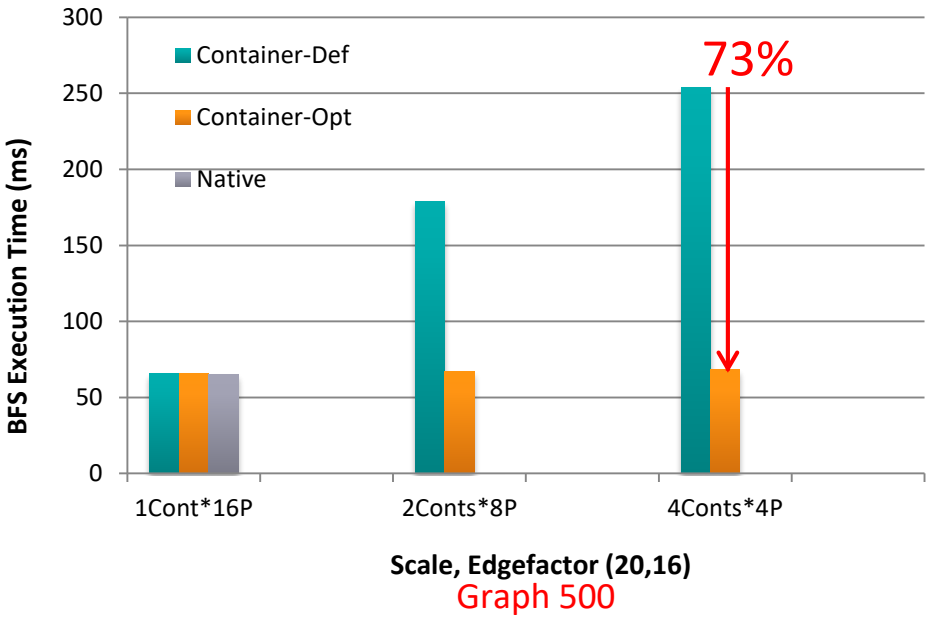


J. Zhang, X. Lu, D. K. Panda. High Performance MPI Library for Container-based HPC Cloud on InfiniBand Clusters. ICPP, 2016

Application-Level Performance on Docker with MVAPICH2



NAS

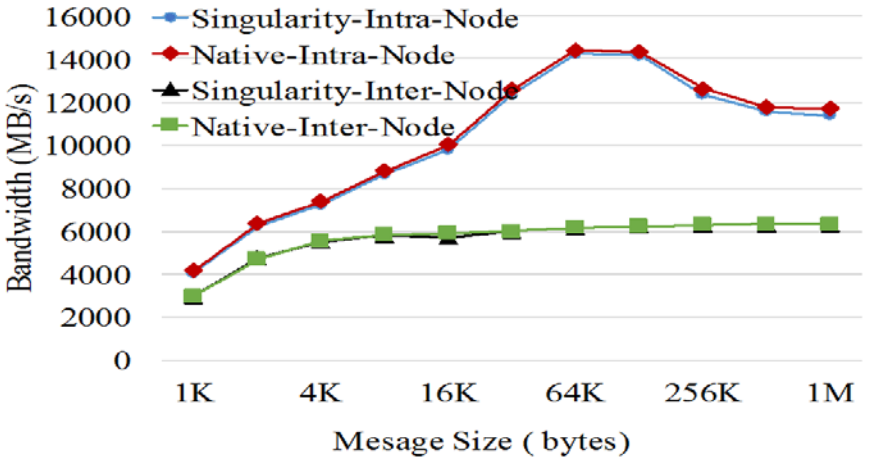


Graph 500

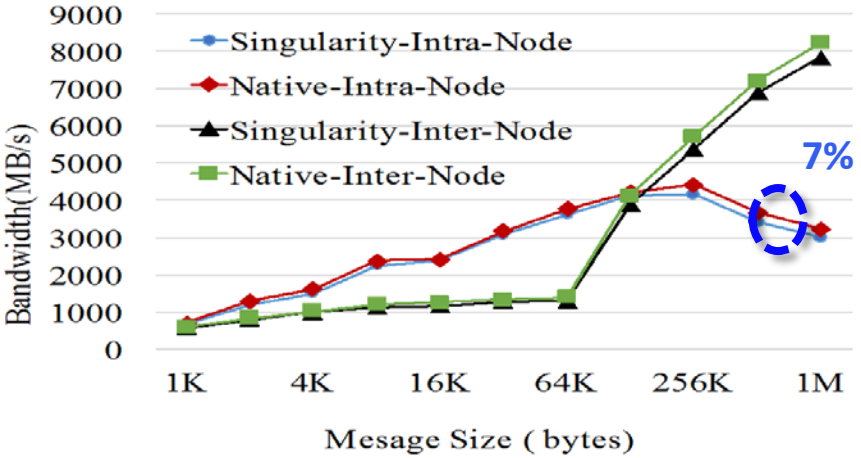
- 64 Containers across 16 nodes, pining 4 Cores per Container
- Compared to Container-Def, up to 11% and 73% of execution time reduction for NAS and Graph 500
- Compared to Native, less than 9 % and 5% overhead for NAS and Graph 500

Singularity Performance on Different Processor Architectures

BW on Haswell

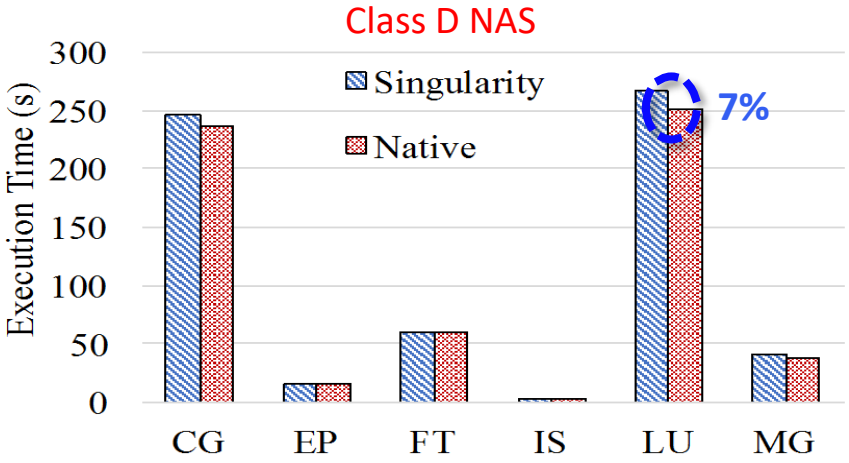


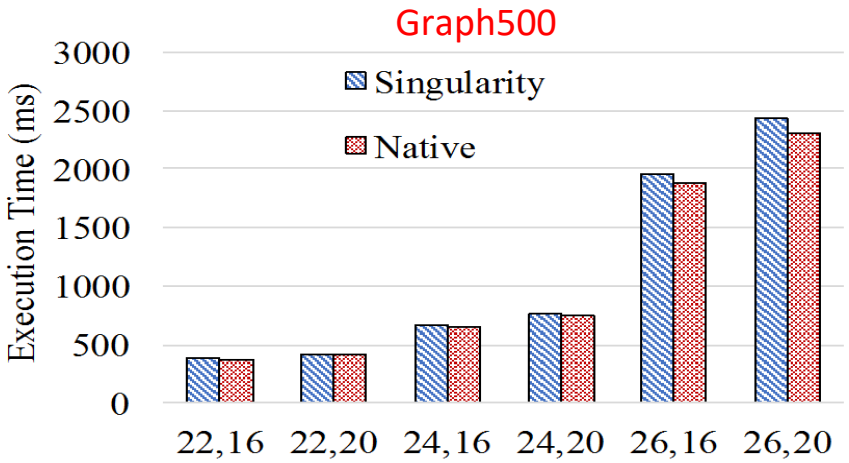
BW on KNL



- MPI point-to-point Bandwidth
- On both Haswell and KNL, less than 7% overhead for Singularity solution
- Worse intra-node performance than Haswell because low CPU frequency, complex cluster mode, and cost maintaining cache coherence
- KNL - Inter-node performs better than intra-node case after around 256 Kbytes, as Omni-Path interconnect outperforms shared memory-based transfer for large message size

Singularity Performance on Haswell with InfiniBand





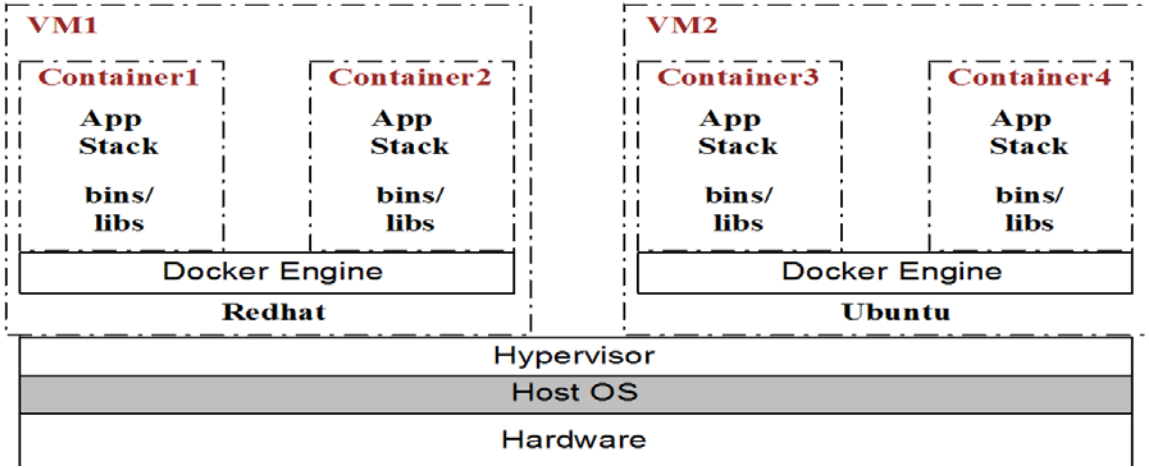
- 512 processors across 32 Haswell nodes
- Singularity delivers near-native performance, less than 7% overhead on Haswell with InfiniBand

J. Zhang, X. Lu, D. K. Panda. Is Singularity-based Container Technology Ready for Running MPI Applications on HPC Clouds? UCC 2017. (Best Student Paper Award)

Approaches to Build HPC Clouds

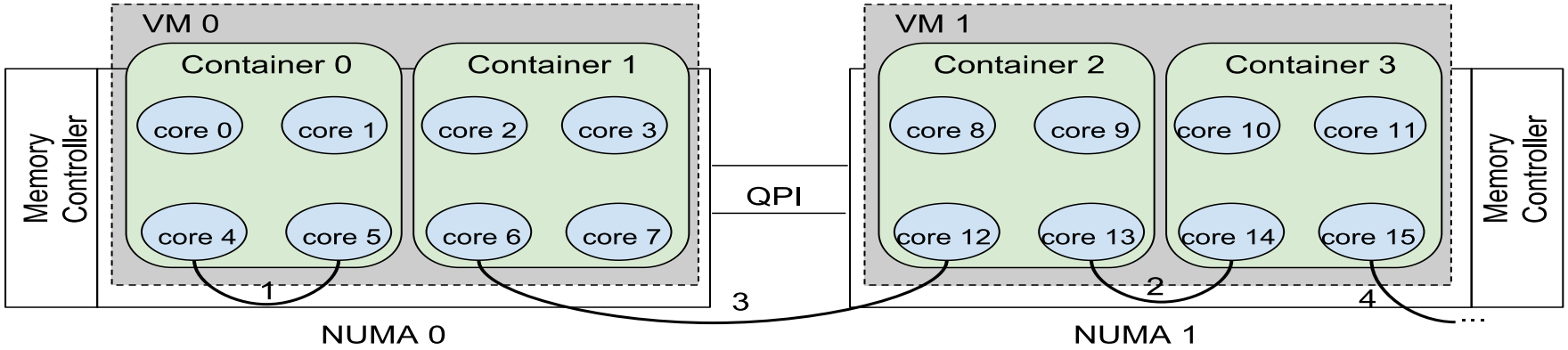
- MVAPICH2-Virt with SR-IOV and IVSHMEM
  - Standalone, OpenStack
- SR-IOV-enabled VM Migration Support in MVAPICH2
- MVAPICH2 with Containers (Docker and Singularity)
- MVAPICH2 with Nested Virtualization (Container over VM)
- MVAPICH2-Virt on SLURM
  - SLURM alone, SLURM + OpenStack
- Neuroscience Applications on HPC Clouds
- Big Data Libraries on Cloud
  - RDMA-Hadoop, OpenStack Swift

Nested Virtualization: Containers over Virtual Machines



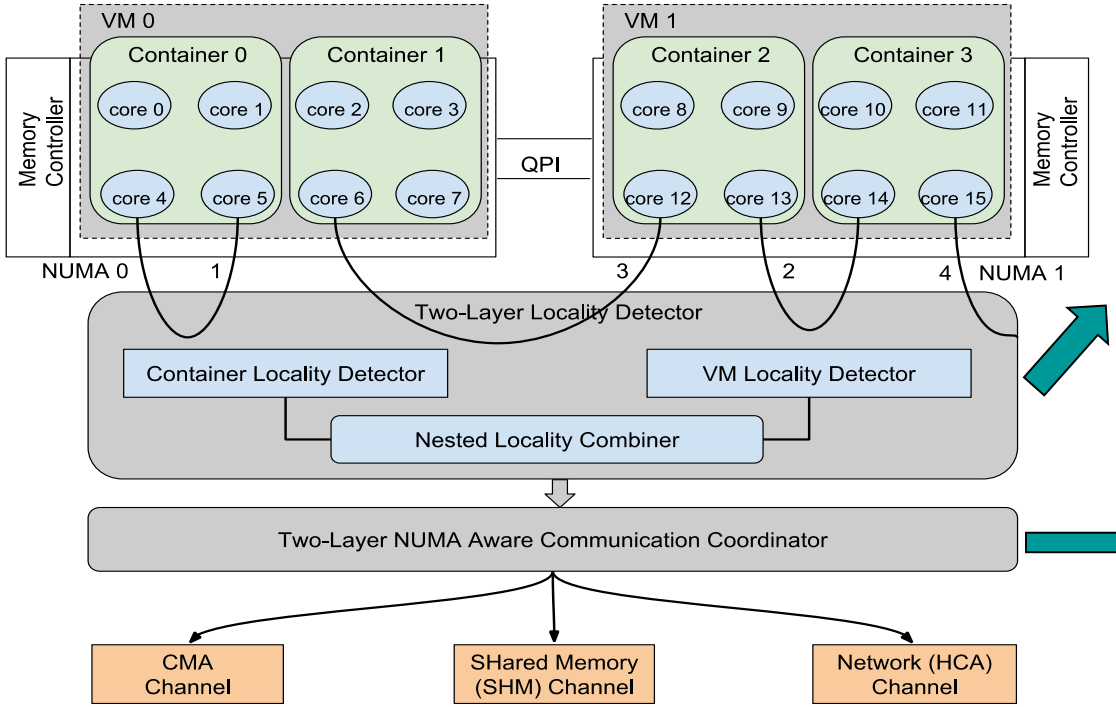
- Useful for live migration, sandbox application, legacy system integration, software deployment, etc.
- Performance issues because of the redundant call stacks (two-layer virtualization) and isolated physical resources

Multiple Communication Paths in Nested Virtualization



- Different VM placements introduce multiple communication paths on container level
  1. Intra-VM Intra-Container (across core 4 and core 5)
  2. Intra-VM Inter-Container (across core 13 and core 14)
  3. Inter-VM Inter-Container (across core 6 and core 12)
  4. Inter-Node Inter-Container (across core 15 and the core on remote node)

Overview of Proposed Design in MVAPICH2

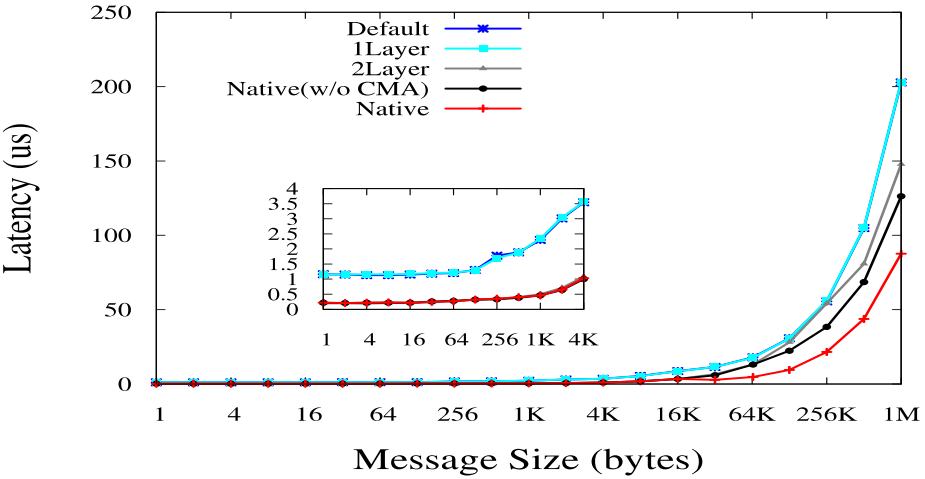


Two-Layer Locality Detector: Dynamically detecting MPI processes in the co-resident containers inside one VM as well as the ones in the co-resident VMs

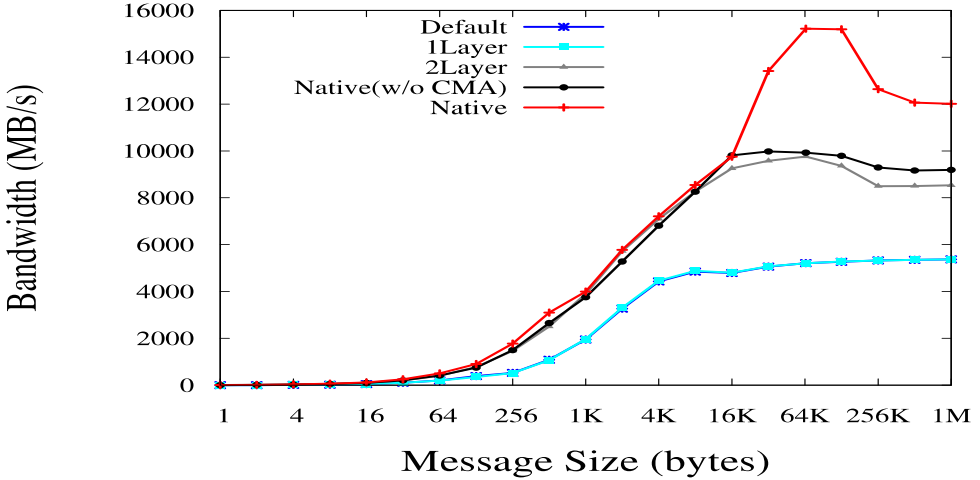
Two-Layer NUMA Aware Communication Coordinator: Leverage nested locality info, NUMA architecture info and message to select appropriate communication channel

J. Zhang, X. Lu, D. K. Panda. Designing Locality and NUMA Aware MPI Runtime for Nested Virtualization based HPC Cloud with SR-IOV Enabled InfiniBand, VEE, 2017

Inter-VM Inter-Container Pt2Pt (Intra-Socket)



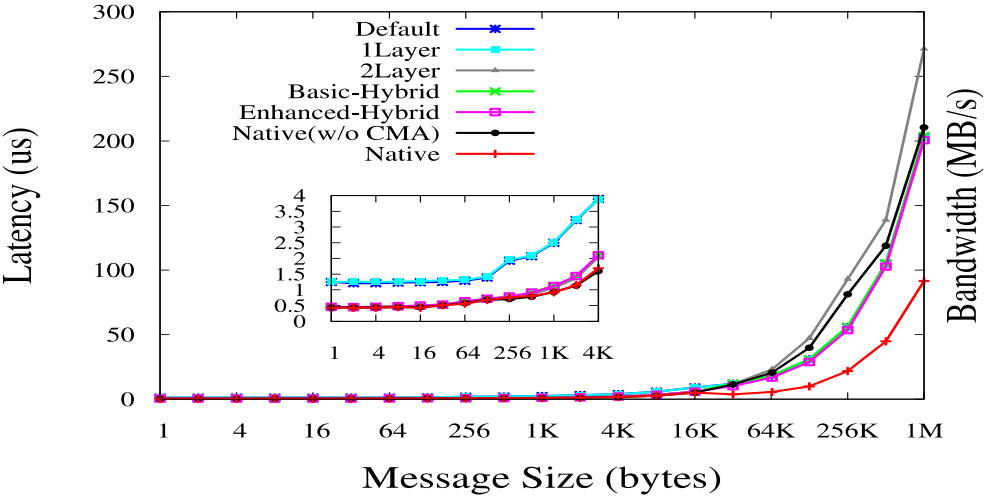
Latency



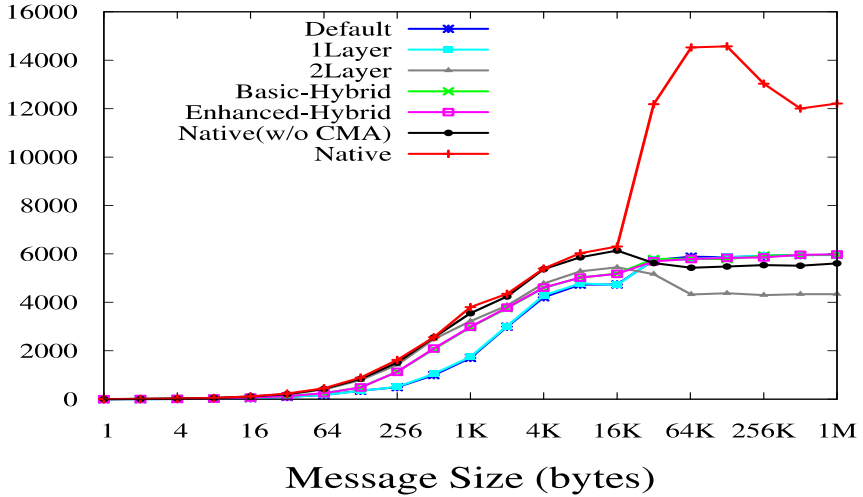
BW

- 1Layer has similar performance to the Default
- Compared with 1Layer, 2Layer delivers up to 84% and 184% improvement for latency and BW

Inter-VM Inter-Container Pt2Pt (Inter-Socket)



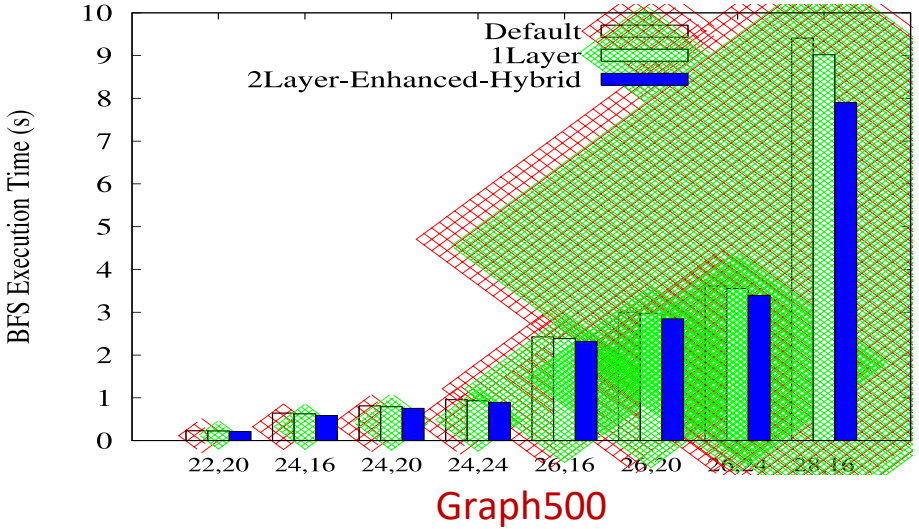
Latency



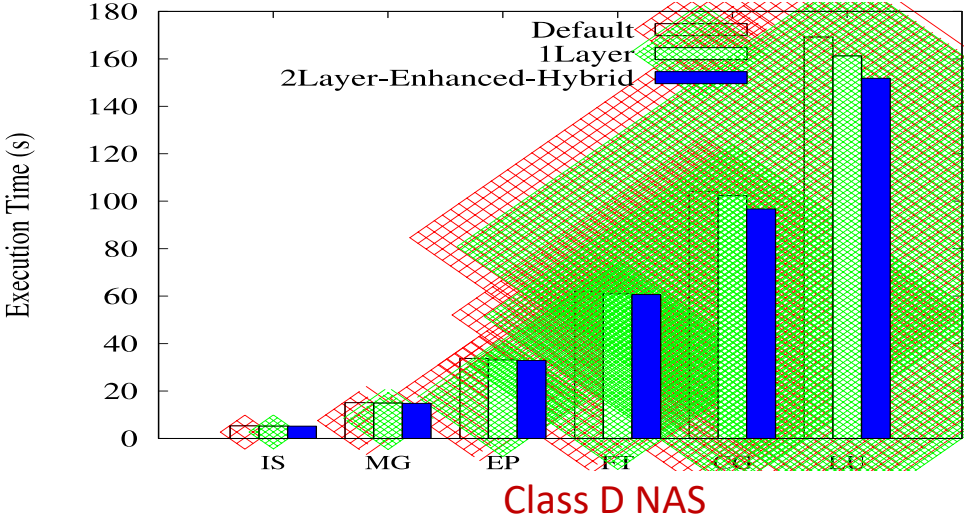
BW

- 1-Layer has similar performance to the Default
- 2-Layer has near-native performance for small msg, but clear overhead on large msg
- Compared to 2-Layer, Hybrid design brings up to 42% and 25% improvement for latency and BW, respectively

Application-level Evaluations



Graph500



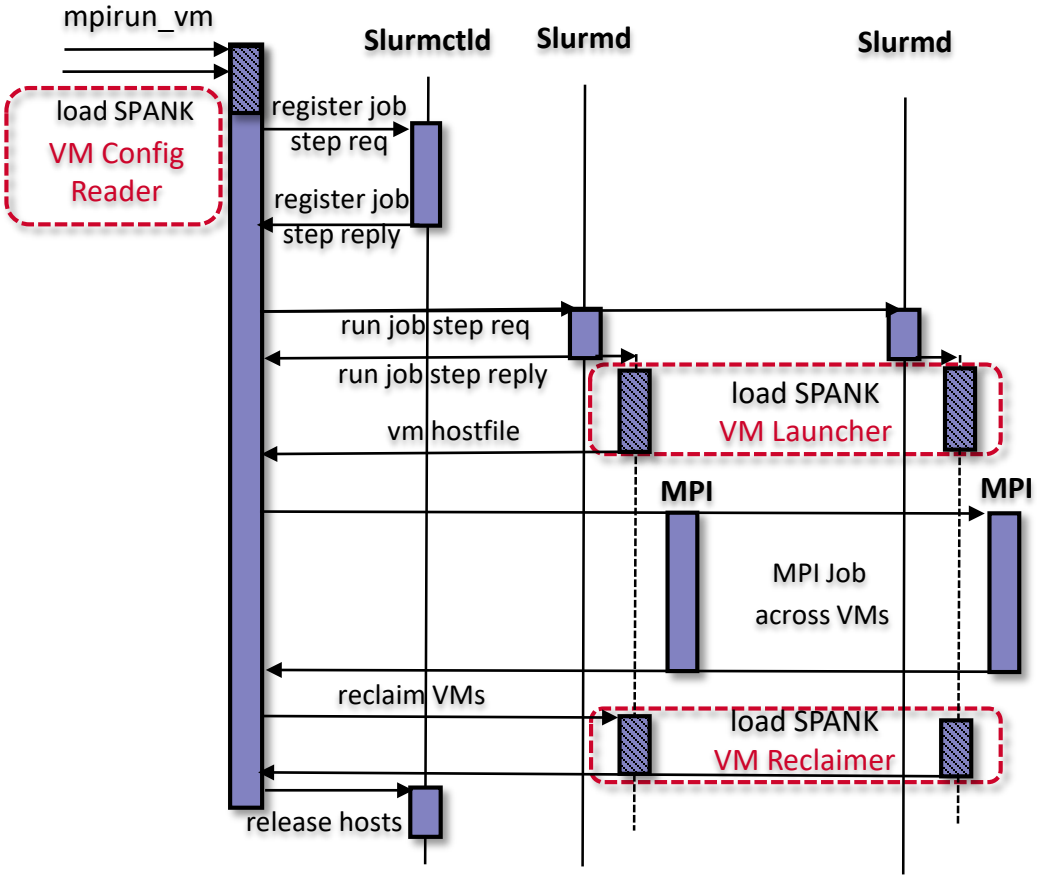
Class D NAS

- 256 processes across 64 containers on 16 nodes
- Compared with Default, enhanced-hybrid design reduces up to 16% (28,16) and 10% (LU) of execution time for Graph 500 and NAS, respectively
- Compared with the 1Layer case, enhanced-hybrid design also brings up to 12% (28,16) and 6% (LU) performance benefit

Approaches to Build HPC Clouds

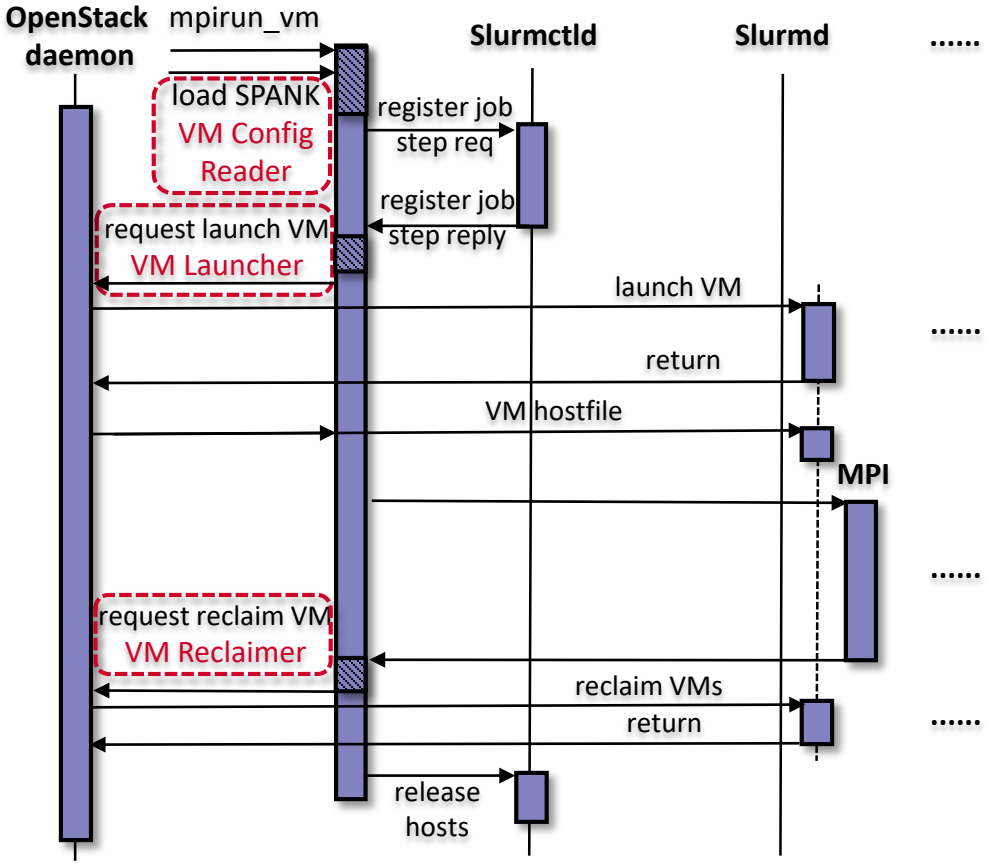
- MVAPICH2-Virt with SR-IOV and IVSHMEM
  - Standalone, OpenStack
- SR-IOV-enabled VM Migration Support in MVAPICH2
- MVAPICH2 with Containers (Docker and Singularity)
- MVAPICH2 with Nested Virtualization (Container over VM)
- MVAPICH2-Virt on SLURM
  - SLURM alone, SLURM + OpenStack
- Neuroscience Applications on HPC Clouds
- Big Data Libraries on Cloud
  - RDMA-Hadoop, OpenStack Swift

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

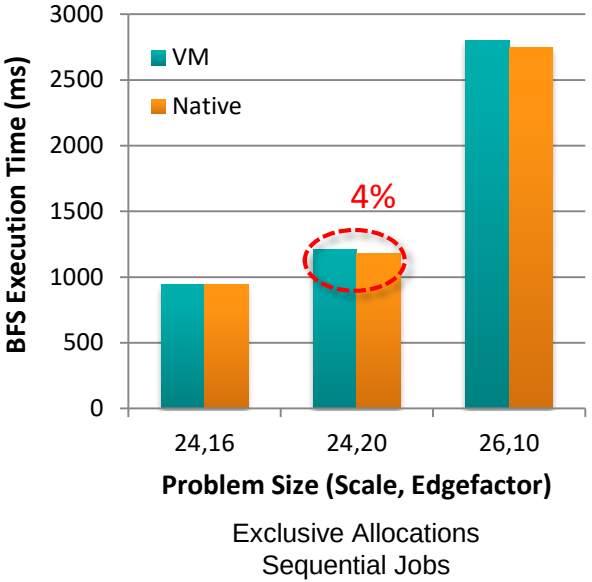
SLURM SPANK Plugin with OpenStack based Design



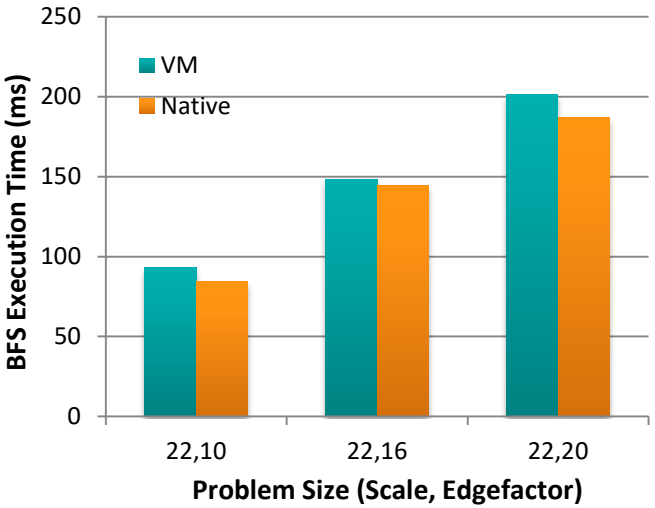
- VM Configuration Reader – VM options register
- VM Launcher, VM Reclaimer – Offload to underlying OpenStack infrastructure
  - PCI Whitelist to passthrough free VF to VM
  - Extend Nova to enable IVSHMEM when launching VM

J. Zhang, X. Lu, S. Chakraborty, D. K. Panda. SLURM-V: Extending SLURM for Building Efficient HPC Cloud with SR-IOV and IVShmem. Euro-Par, 2016

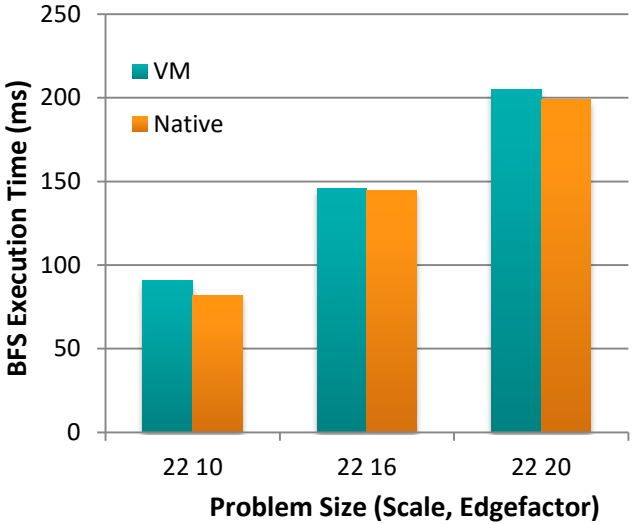
Application-Level Performance on Chameleon (Graph500)



Exclusive Allocations
Sequential Jobs



Shared-host Allocations
Concurrent Jobs



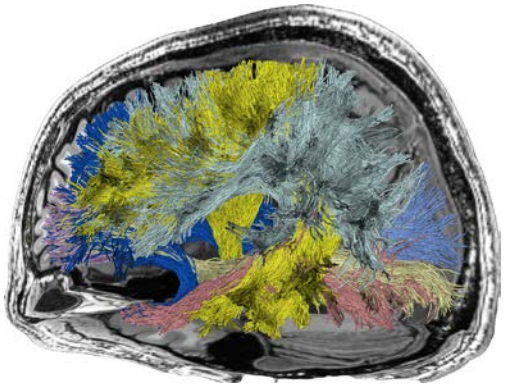
Exclusive Allocations
Concurrent Jobs

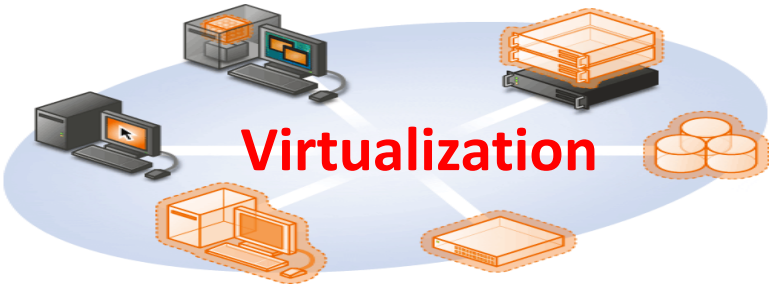
- 32 VMs across 8 nodes, 6 Core/VM
- EASJ - Compared to Native, less than 4% overhead with 128 Procs
- SACJ, EACJ – Also minor overhead, when running NAS as concurrent job with 64 Procs

Approaches to Build HPC Clouds

- MVAPICH2-Virt with SR-IOV and IVSHMEM
  - Standalone, OpenStack
- SR-IOV-enabled VM Migration Support in MVAPICH2
- MVAPICH2 with Containers (Docker and Singularity)
- MVAPICH2 with Nested Virtualization (Container over VM)
- MVAPICH2-Virt on SLURM
  - SLURM alone, SLURM + OpenStack
- Neuroscience Applications on HPC Clouds
- Big Data Libraries on Cloud
  - RDMA-Hadoop, OpenStack Swift

NeuroScience Meets HPC Cloud





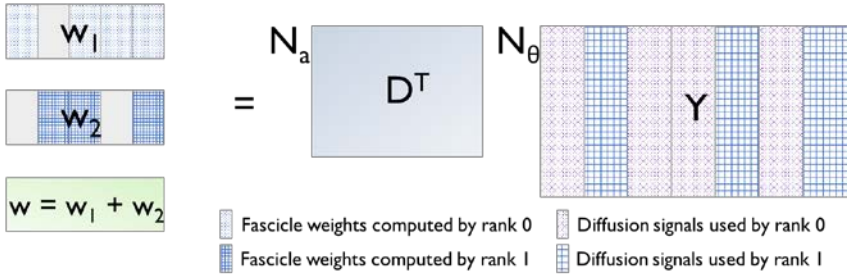
The Brain Connectome. *Illustration of a set of fascicles (white matter bundles) obtained by using a tractography algorithm. Fascicles are grouped together conforming white matter tracts (shown with different colors here) connecting different cortical areas of the human brain.* LiFE[1] (Linear Fascicle Evaluation) is an approach to predict diffusion measurements in brain connectomes.

- **Easy and Fast Discovery is the key!**
- **Easy-to-use and High-Performance Technology is the key!**

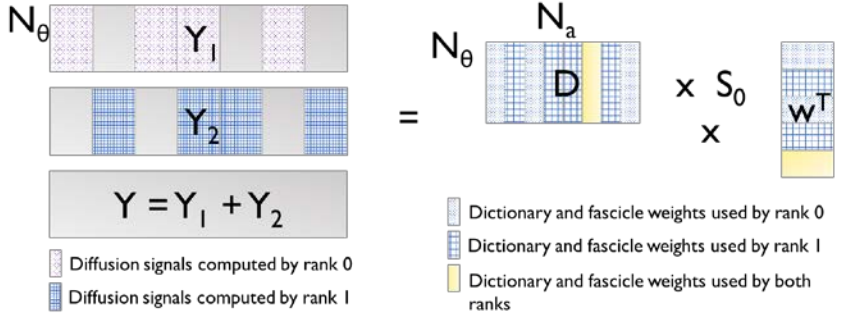
[1]**https://github.com/francopestilli/life**

MPI-based LiFE for Brain Health: Initial Design using MVAPICH2 MPI Library

- Identified computationally intensive tasks as the computations of matrix by vector products
  - $w = M^T y$ and $y = Mw$
- The computationally intensive functions have been parallelized using MPI by dividing the task among multiple MPI processes
- Implementation uses MVAPICH2[2], from OSU team



Computation of $w = M^T y$ using 2 MPI processes

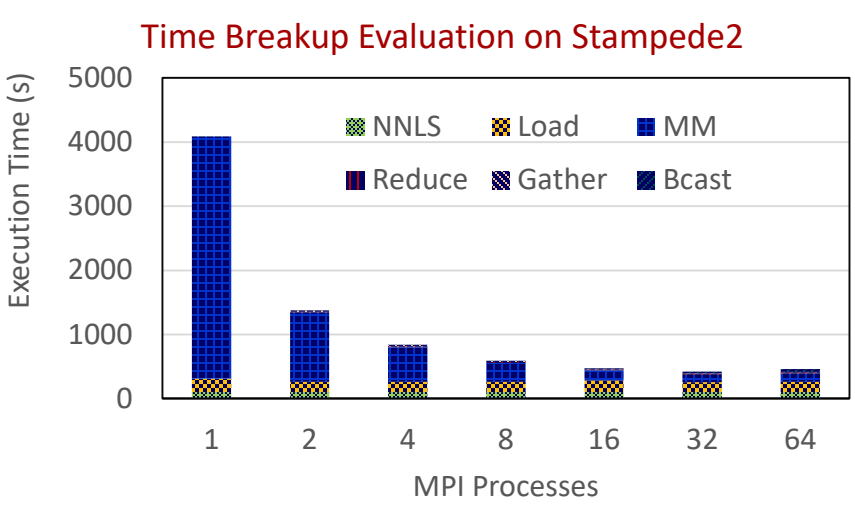


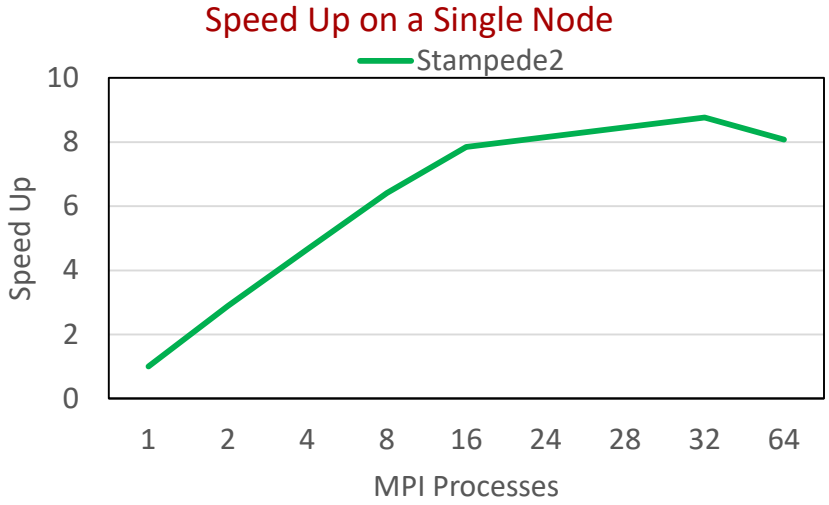
Computation of $y = Mw$ using 2 MPI processes

[1]https://github.com/francopestilli/life

[2]http://mvapich.cse.ohio-state.edu/

Design and Evaluation with MVAPICH2: Single Node with MPI on Intel Knights Landing (KNL)



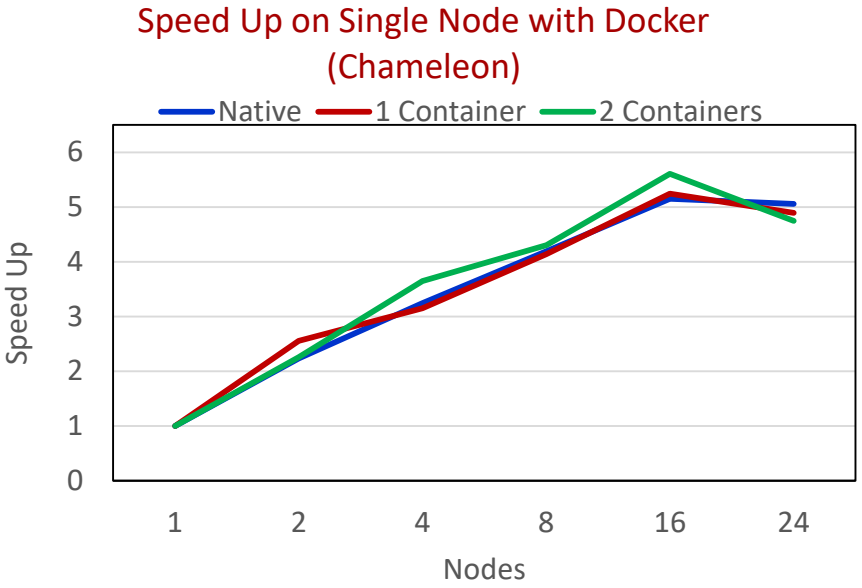


- Evaluation on TACC Stampede KNL (Intel Xeon Phi KNL CPUs, 68 cores, 96 GB memory per node)
- Up to 8.7x speed up

MPI-LiFE software is available from http://neurohpc.cse.ohio-state.edu

Docker-containerized version, Can run from laptop to clusters

Design and Evaluation of LiFE code with MVAPICH2-Virt+Docker



- Evaluation on Chameleon with Docker (Intel Haswell CPUs, 24 cores, 128 GB memory per node)
- Up to 5.5x speed up on Chameleon

Approaches to Build HPC Clouds

- MVAPICH2-Virt with SR-IOV and IVSHMEM
  - Standalone, OpenStack
- SR-IOV-enabled VM Migration Support in MVAPICH2
- MVAPICH2 with Containers (Docker and Singularity)
- MVAPICH2 with Nested Virtualization (Container over VM)
- MVAPICH2-Virt on SLURM
  - SLURM alone, SLURM + OpenStack
- Neuroscience Applications on HPC Clouds
- Big Data Libraries on Cloud
  - RDMA-Hadoop, OpenStack Swift

The High-Performance Big Data (HiBD) Project

- RDMA for Apache Spark
- RDMA for Apache Hadoop 2.x (RDMA-Hadoop-2.x)
  - Plugins for Apache, Hortonworks (HDP) and Cloudera (CDH) Hadoop distributions
- RDMA for Apache HBase
- RDMA for Memcached (RDMA-Memcached)
- RDMA for Apache Hadoop 1.x (RDMA-Hadoop)
- OSU HiBD-Benchmarks (OHB)
  - HDFS, Memcached, HBase, and Spark Micro-benchmarks
- http://hibd.cse.ohio-state.edu
- Users Base: 280 organizations from 34 countries
- More than 25,750 downloads from the project site

Available for InfiniBand and RoCE

Also run on Ethernet

Available for x86 and OpenPOWER

Upcoming Release will have support

For Singularity and Docker

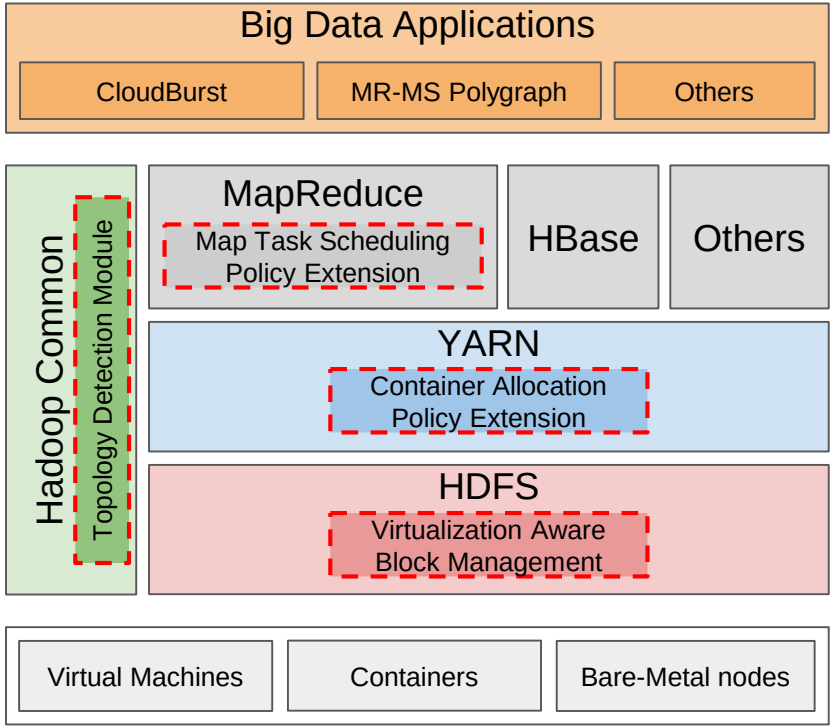
HiBD

High-Performance Big Data

Network Based Computing Laboratory

The Ohio State University

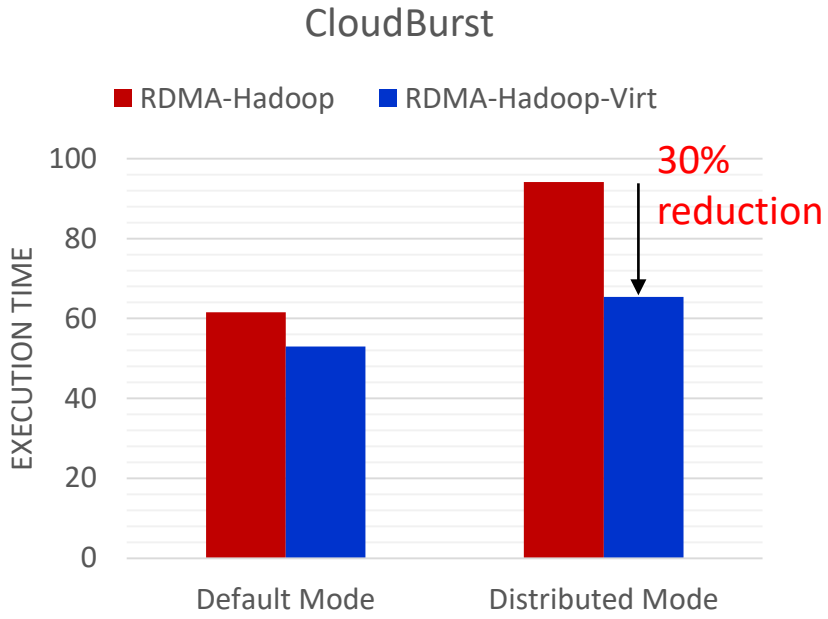
Overview of RDMA-Hadoop-Virt Architecture

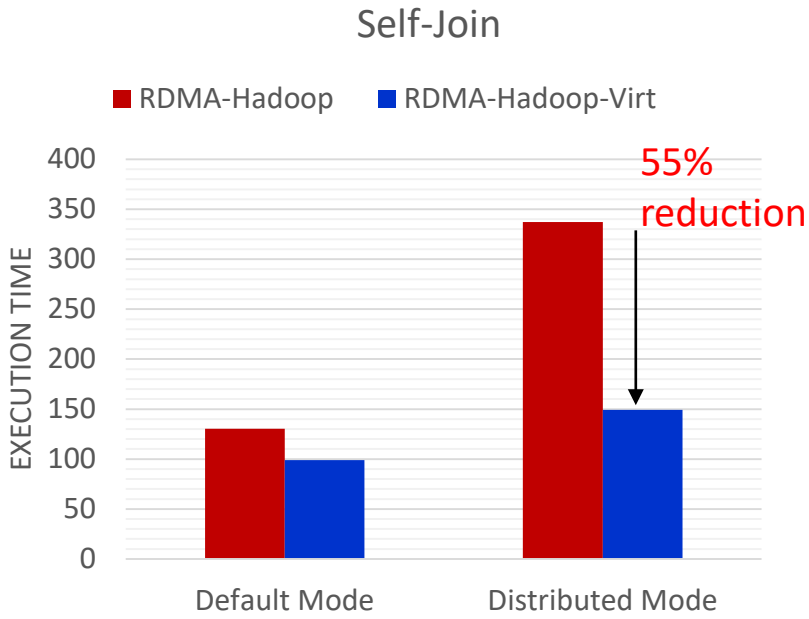


- Virtualization-aware modules in all the four main Hadoop components:
  - **HDFS**: Virtualization-aware Block Management to improve fault-tolerance
  - **YARN**: Extensions to Container Allocation Policy to reduce network traffic
  - **MapReduce**: Extensions to Map Task Scheduling Policy to reduce network traffic
  - **Hadoop Common**: Topology Detection Module for automatic topology detection
- Communications in HDFS, MapReduce, and RPC go through RDMA-based designs over SR-IOV enabled InfiniBand

S. Gugnani, X. Lu, D. K. Panda. Designing Virtualization-aware and Automatic Topology Detection Schemes for Accelerating Hadoop on SR-IOV-enabled Clouds. CloudCom, 2016.

Evaluation with Applications

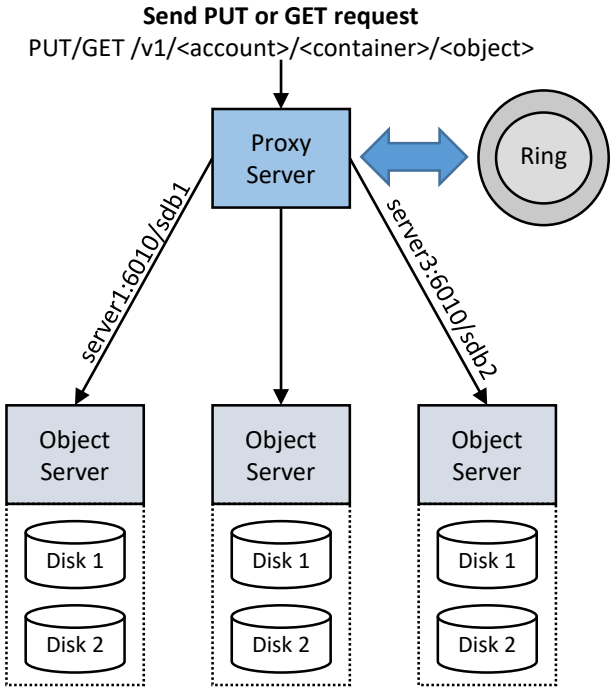




- 14% and 24% improvement with Default Mode for CloudBurst and Self-Join
- 30% and 55% improvement with Distributed Mode for CloudBurst and Self-Join

OpenStack Swift Overview

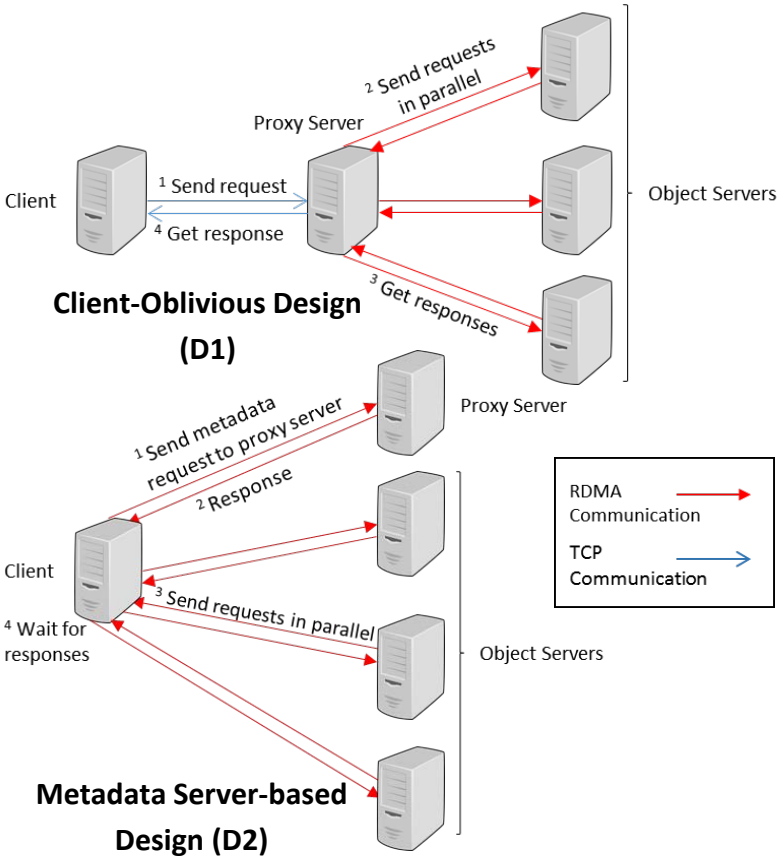
- Distributed Cloud-based Object Storage Service
- Deployed as part of OpenStack installation
- Can be deployed as standalone storage solution as well
- Worldwide data access via Internet
  - HTTP-based
- Architecture
  - Multiple Object Servers: To store data
  - Few Proxy Servers: Act as a proxy for all requests
  - Ring: Handles metadata
- Usage
  - Input/output source for Big Data applications (most common use case)
  - Software/Data backup
  - Storage of VM/Docker images
- Based on traditional TCP sockets communication



Swift Architecture

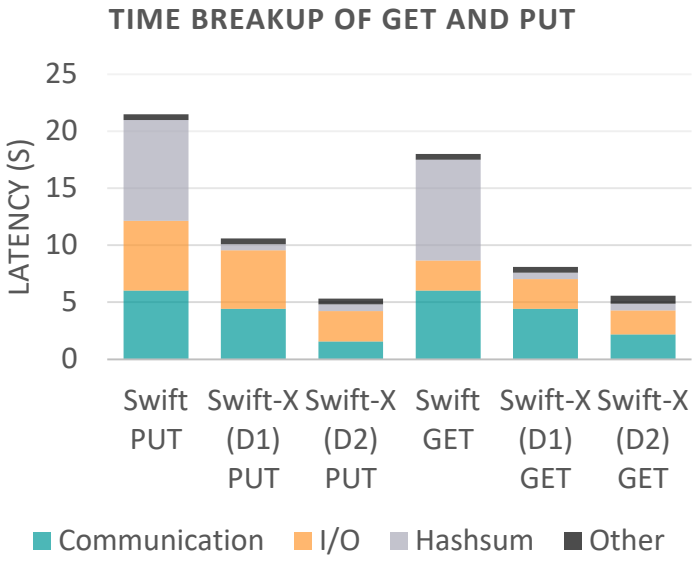
Swift-X: Accelerating OpenStack Swift with RDMA for Building Efficient HPC Clouds

- Challenges
  - Proxy server is a bottleneck for large scale deployments
  - Object upload/download operations network intensive
  - Can an RDMA-based approach benefit?
- Design
  - Re-designed Swift architecture for improved scalability and performance; Two proposed designs:
    - **Client-Oblivious Design**: No changes required on the client side
    - **Metadata Server-based Design**: Direct communication between client and object servers; bypass proxy server
  - RDMA-based communication framework for accelerating networking performance
  - High-performance I/O framework to provide maximum overlap between communication and I/O

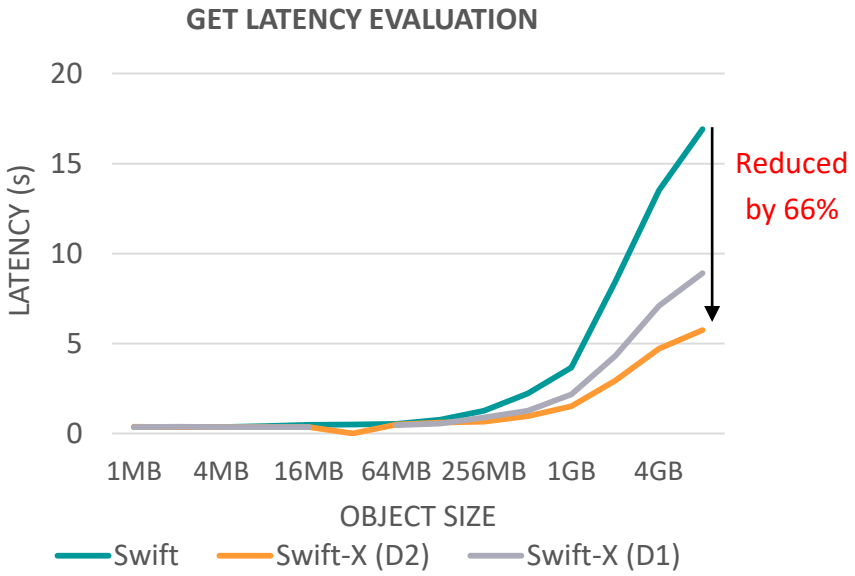


S. Gugnani, X. Lu, and D. K. Panda, Swift-X: Accelerating OpenStack Swift with RDMA for Building an Efficient HPC Cloud, accepted at CCGrid'17, May 2017

Swift-X: Accelerating OpenStack Swift with RDMA for Building Efficient HPC Clouds

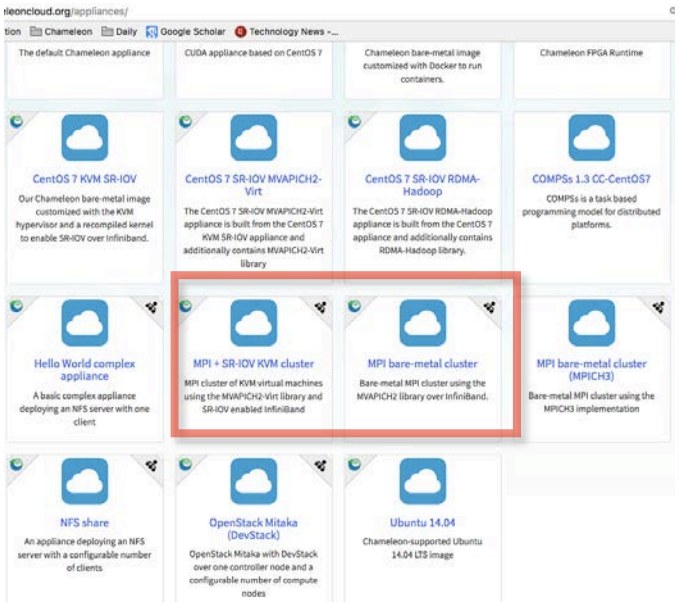


- Communication time reduced by up to 3.8x for PUT and up to 2.8x for GET



- Up to 66% reduction in GET latency

Available Appliances on Chameleon Cloud*



Appliance	Description
CentOS 7 KVM SR-IOV	Chameleon bare-metal image customized with the KVM hypervisor and a recompiled kernel to enable SR-IOV over InfiniBand. https://www.chameleoncloud.org/appliances/3/
MPI bare-metal cluster complex appliance (Based on Heat)	This appliance deploys an MPI cluster composed of bare metal instances using the MVAPICH2 library over InfiniBand. https://www.chameleoncloud.org/appliances/29/
MPI + SR-IOV KVM cluster (Based on Heat)	This appliance deploys an MPI cluster of KVM virtual machines using the MVAPICH2-Virt implementation and configured with SR-IOV for high-performance communication over InfiniBand. https://www.chameleoncloud.org/appliances/28/
CentOS 7 SR-IOV RDMA-Hadoop	The CentOS 7 SR-IOV RDMA-Hadoop appliance is built from the CentOS 7 appliance and additionally contains RDMA-Hadoop library with SR-IOV. https://www.chameleoncloud.org/appliances/17/

- 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 over bare-metal InfiniBand clusters
  - High-Performance MVAPICH2 Library with Virtualization Support over SR-IOV enabled KVM clusters
  - High-Performance Hadoop with RDMA-based Enhancements Support

[*] Only include appliances contributed by OSU NowLab

Conclusions

- MVAPICH2-Virt over SR-IOV-enabled InfiniBand is an efficient approach to build HPC Clouds
  - Standalone, OpenStack, Slurm, and Slurm + OpenStack
  - Support Virtual Machine Migration with SR-IOV InfiniBand devices
  - Support Virtual Machine, Container (Docker and Singularity), and Nested Virtualization
- Very little overhead with virtualization, near native performance at application level
- Much better performance than Amazon EC2
- **MVAPICH2-Virt** is available for building HPC Clouds
  - SR-IOV, IVSHMEM, Docker support, OpenStack
- Neuroscience applications can benefit from technologies on HPC clouds
- Big Data analytics stacks such as RDMA-Hadoop can benefit from cloud-aware designs
- Appliances for MVAPICH2-Virt and RDMA-Hadoop are available for building HPC Clouds
- SR-IOV/container support and appliances for other MVAPICH2 libraries (MVAPICH2-X, MVAPICH2-GDR, ...) and RDMA-Spark/Memcached

The 2nd International BoF on

Building Efficient Clouds for HPC, Big Data, and Deep Learning Middleware and Applications (HPC Cloud BoF)

HPC Cloud BoF 2018 will be held with ISC, Frankfurt, Germany, June, 2018

BoF Date: TBD, 2018

HPC Cloud BoF 2017 was held in conjunction with SC'17

http://sc17.supercomputing.org/presentation/?id=bof165&sess=sess357

One More Presentation

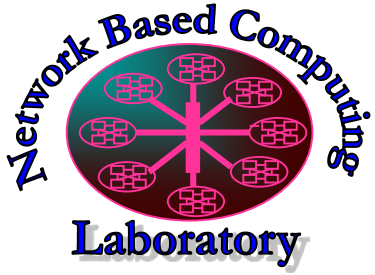
- Thursday (04/12/18) at 04:00 pm

High-Performance Big Data Analytics with RDMA over NVM and NVMe-SSD

Thank You!

luxi@cse.ohio-state.edu

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



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

MVAPICH/MVAPICH2
http://mvapich.cse.ohio-state.edu/

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