Advanced Virtualization Techniques for High Performance Cloud Cyberinfrastructure

Andrew J. Younge

Doctoral Proposal
Indiana University

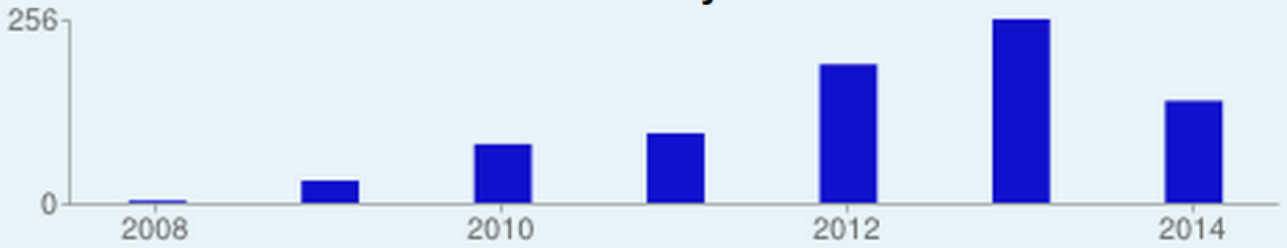
whoami

- PhD Candidate at Indiana University
  - Advisor: Dr. Geoffrey C. Fox
  - Persistent Systems Fellowship via SOIC
  - @ IU since early 2010
  - Worked on the FutureGrid Project
- Previously at Rochester Institute of Technology
  - B.S. & M.S. in Computer Science in 2008, 2010
- > dozen publications
  - Involved in Distributed Systems since 2006 (UMD)
- Visiting Researcher at USC/ISI East (2012 & 2013)
- Google summer code w/ UC/ANL (2011)

http://ajyounge.com

Citation indices

	All	Since 2009
Citations	818	811
h-index	10	10
i10-index	10	10



Outline

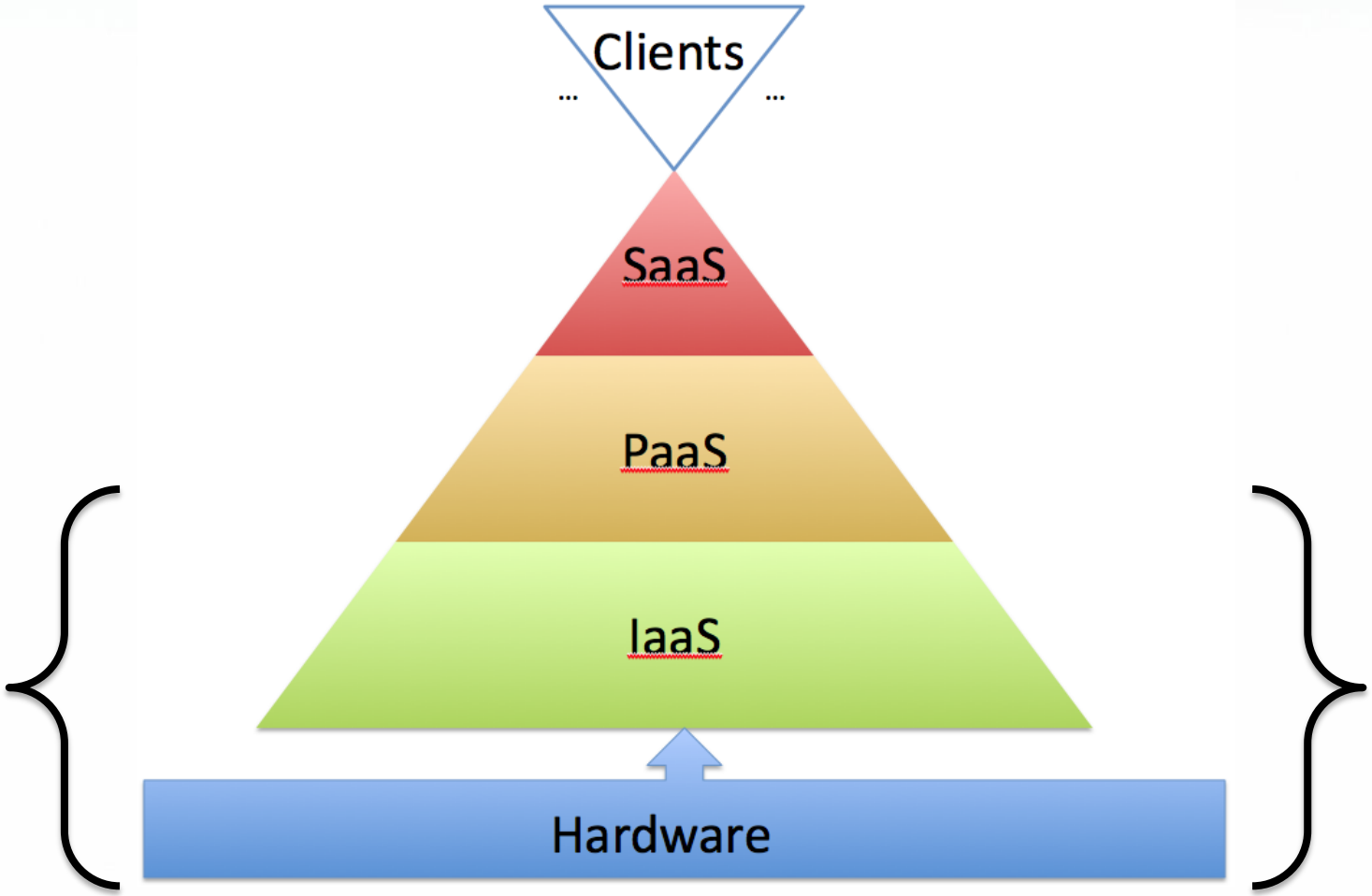
- Overview
- Motivation
- Hypervisor evaluation
- GPU Passthrough
- Comparing GPUs in Cloud
  - Multiple hypervisors
  - Different architectures
- InfiniBand SR-IOV integration
- Real-world application use cases
- Future Work

What is Cloud Computing?

- "Computing may someday be organized as a public utility just as the telephone system is a public utility... The computer utility could become the basis of a new and important industry."
  - John McCarthy, 1961

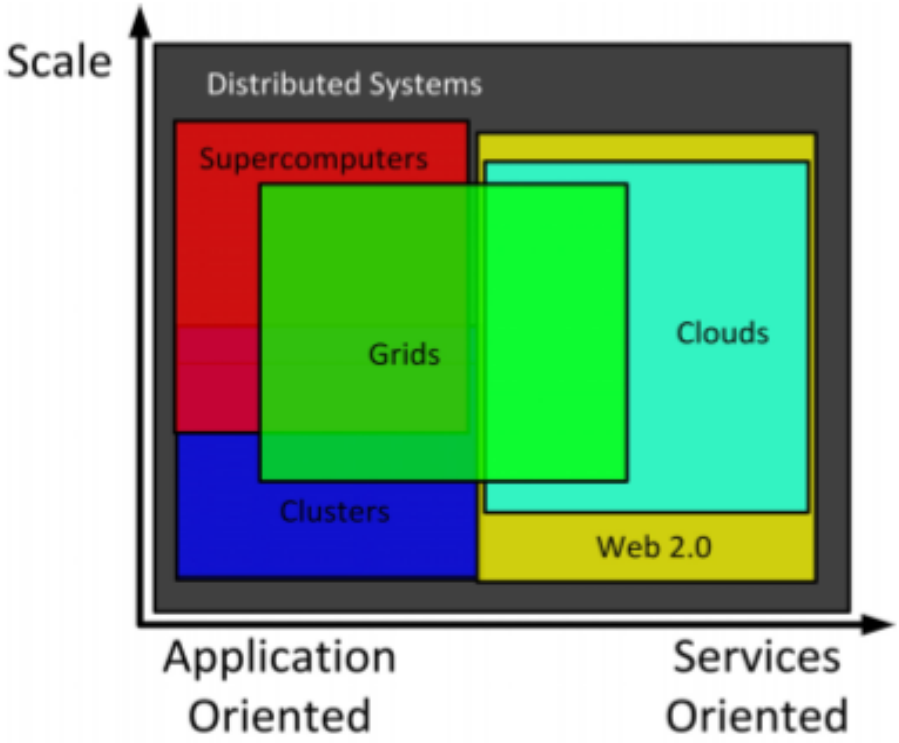
- "Cloud computing is a large-scale distributed computing paradigm that is driven by economies of scale, in which a pool of abstracted, virtualized, dynamically scalable, managed computing power, storage, platforms, and services are delivered on demand to external customers over the Internet."
  - Ian Foster, 2008

*-as-a-Service



Cloud Infrastructure

- Distributed Systems encompasses a wide variety of technologies.
- Per Foster, Grid computing spans most areas and is becoming more mature.
- Clouds are an emerging technology, providing many of the same features as Grids without many of the potential pitfalls.



From "Cloud Computing and Grid Computing 360-Degree Compared" in 2008

HPC or Cloud?

HPC

- Fast, tightly coupled systems
- Performance is paramount
- Massively parallel applications
- MPI for distributed memory computation & communication
  - Require advanced interconnects
- Leverage accelerator cards or co-processors (new)

Cloud

- Built on commodity PC components
- User experience is paramount
- Scalability and concurrency are key to success
- Big Data applications to handle the Data Deluge
  - 4th Paradigm
  - Long tail of science
- Leverage virtualization

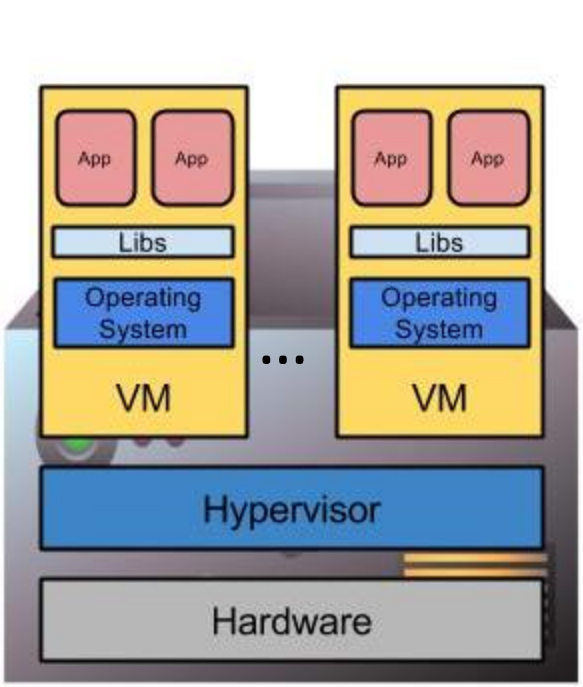
Hypothesis: Combine the performance of HPC with usability of Clouds to support mid-tier scientific computation

High Performance Cloud Infrastructure

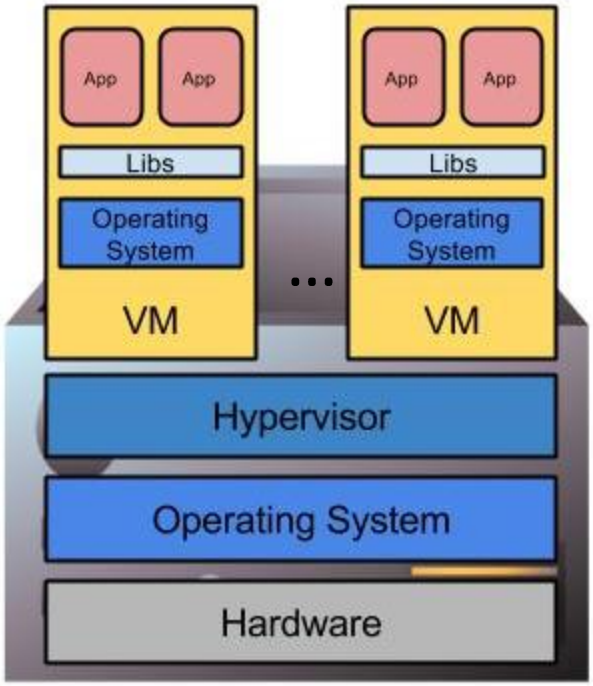
- Evaluate hypervisors against bare-metal hardware
  - Classify overhead when/where it exists
  - Find best hypervisors, configurations, & practices
- Enable accelerators & GPUs
  - GPU Passthrough of Nvidia Tesla GPUs
  - Evaluate performance & overhead
- Explore VM interconnects
  - InfiniBand SR-IOV & Passthrough
  - Tuning mechanisms for improved performance
- Apply research to Cloud Infrastructure in OpenStack
- Scale real-world applications on FutureGrid hardware
  - Molecular dynamics
  - Earthquake simulation

Virtualization

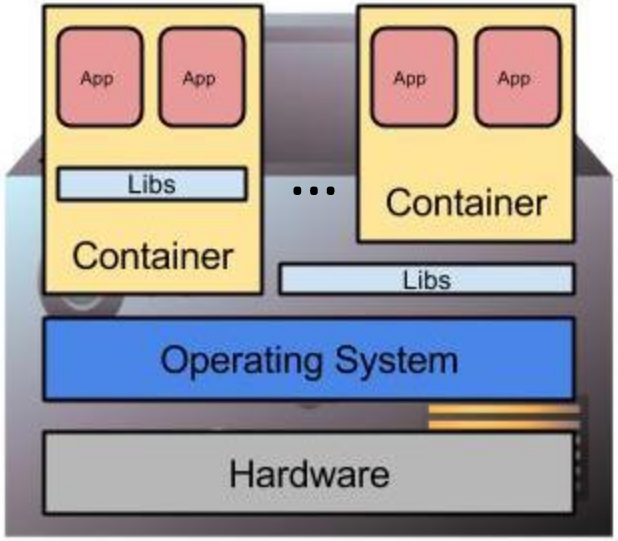
- Virtual Machine (VM) is a software implementation of a machine that executes as if it was running on a physical resource directly.
- Enables multiple operating systems & environments to run simultaneously on one physical machine.



Type 1 Hypervisor



Type 2 Hypervisor



Containers

Motivation

- Most “Cloud” deployments rely on virtualization.
  - Amazon EC2, GoGrid, Azure, Rackspace Cloud …
  - Nimbus, Eucalyptus, OpenNebula, OpenStack …
- Number of Virtualization tools or Hypervisors available today.
- Need to compare these hypervisors for use within the scientific computing community.

From: Andrew J. Younge, Robert Henschel, James T. Brown, Gregor von Laszewski, Judy Qiu, Geoffrey C. Fox, Analysis of Virtualization Technologies for High Performance Computing Environments, in Proceedings of the 4th International Conference on Cloud Computing (CLOUD 2011). 2011

Current Hypervisors

vmware®

Xen™

virtualiron®

KVM

Parallels™

OpenVZ
Linux Containers

Windows Server® 2008
Hyper-V™

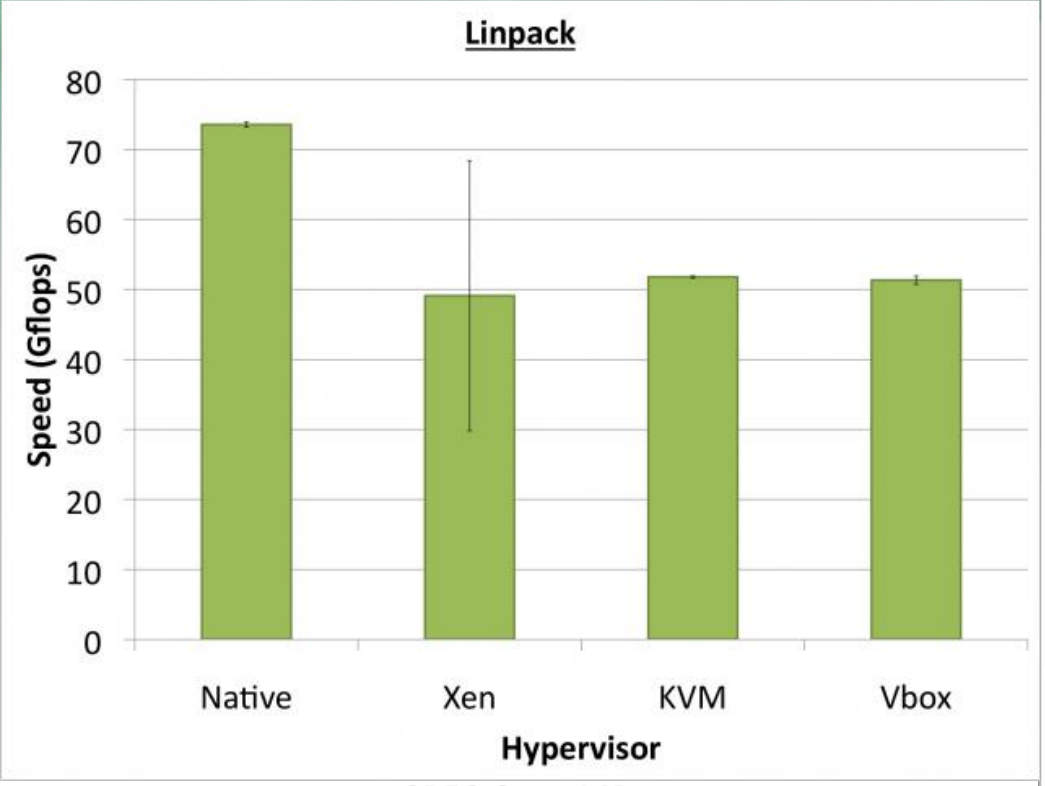
QEMU
open source processor emulator

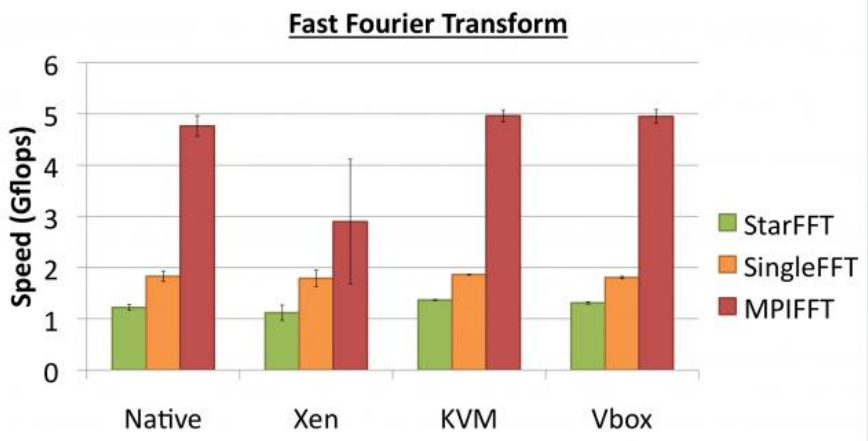
Future Grid

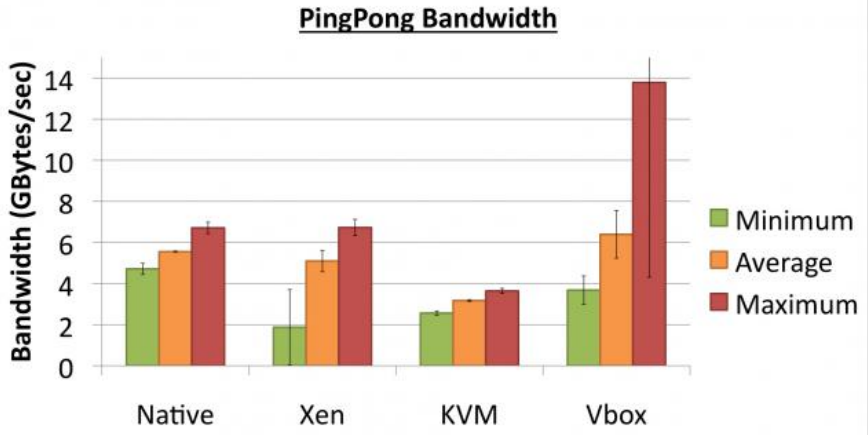
Features

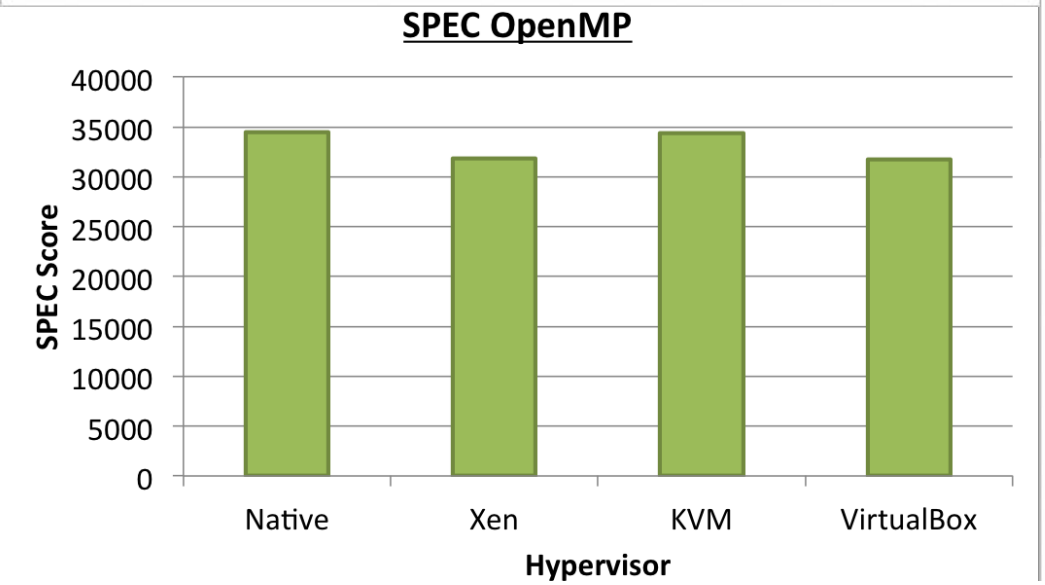
	Xen	KVM	VirtualBox	VMWare
Paravirtualization	Yes	No	No	No
Full Virtualization	Yes	Yes	Yes	Yes
Host CPU	X86, X86_64, IA64	X86, X86_64, IA64, PPC	X86, X86_64	X86, X86_64
Guest CPU	X86, X86_64, IA64	X86, X86_64, IA64, PPC	X86, X86_64	X86, X86_64
Host OS	Linux, Unix	Linux	Windows, Linux, Unix	Proprietary Unix
Guest OS	Linux, Windows, Unix	Linux, Windows, Unix	Linux, Windows, Unix	Linux, Windows, Unix
VT-x / AMD-v	Opt	Req	Opt	Opt
Supported Cores	128	16*	32	8
Supported Memory	4TB	4TB	16GB	64GB
3D Acceleration	Xen-GL	VMGL	Open-GL	Open-GL, DirectX
Licensing	GPL	GPL	GPL/Proprietary	Proprietary

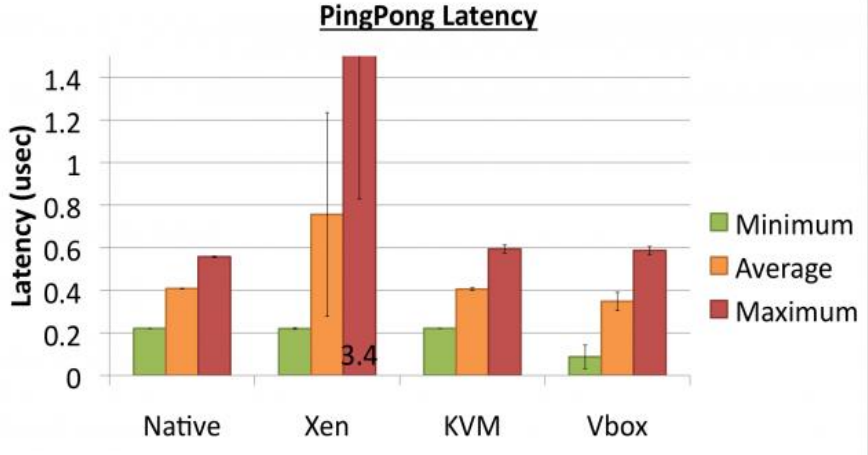
From: Andrew J. Younge, Robert Henschel, James T. Brown, Gregor von Laszewski, Judy Qiu, Geoffrey C. Fox, Analysis of Virtualization Technologies for High Performance Computing Environments, in Proceedings of the 4th International Conference on Cloud Computing (CLOUD 2011). 2011











From: Andrew J. Younge, Robert Henschel, James T. Brown, Gregor von Laszewski, Judy Qiu, Geoffrey C. Fox, Analysis of Virtualization Technologies for High Performance Computing Environments, in Proceedings of the 4th International Conference on Cloud Computing (CLOUD 2011). 2011

Virtualization in HPC

- Is Cloud Computing initially viable for scientific High Performance Computing?
  - Yes, some of the time
- Features: All T1 & T2 hypervisors are similar
- Performance: KVM is fastest across most benchmarks, VirtualBox close. Overall, we have found KVM to be the best hypervisor choice for HPC.
  - Xen's variability is more pronounced than other hypervisors

From: Andrew J. Younge, Robert Henschel, James T. Brown, Gregor von Laszewski, Judy Qiu, Geoffrey C. Fox, Analysis of Virtualization Technologies for High Performance Computing Environments, in Proceedings of the 4th International Conference on Cloud Computing (CLOUD 2011). 2011

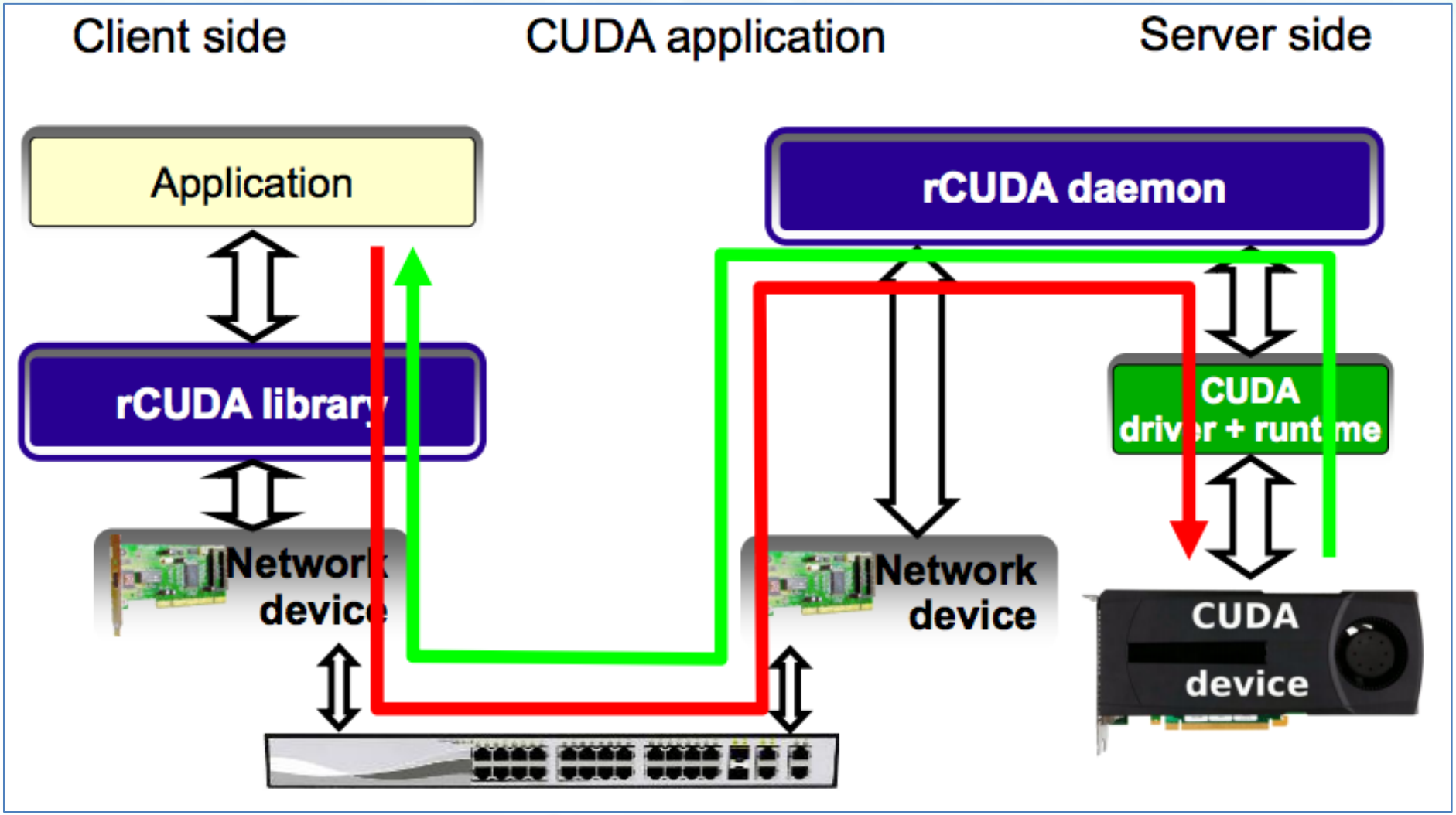
IaaS with HPC Hardware

- Providing near-native hypervisor performance cannot solve all challenges of supporting parallel computing in cloud infrastructure.
- Need to leverage HPC hardware
  - Accelerator cards
  - High speed, low latency I/O interconnects
  - Others…
- Need to characterize and minimize overhead wherever it exists

GPUs in Virtual Machines

- Need for GPUs on Clouds
  - GPUs are becoming commonplace in scientific computing
  - Great performance-per-watt
- Different competing methods for virtualizing GPUs
  - Remote API for CUDA calls
  - Direct GPU usage within VM
- Advantages and disadvantages to both solutions

Front-end GPU API



Front-end API Limitations

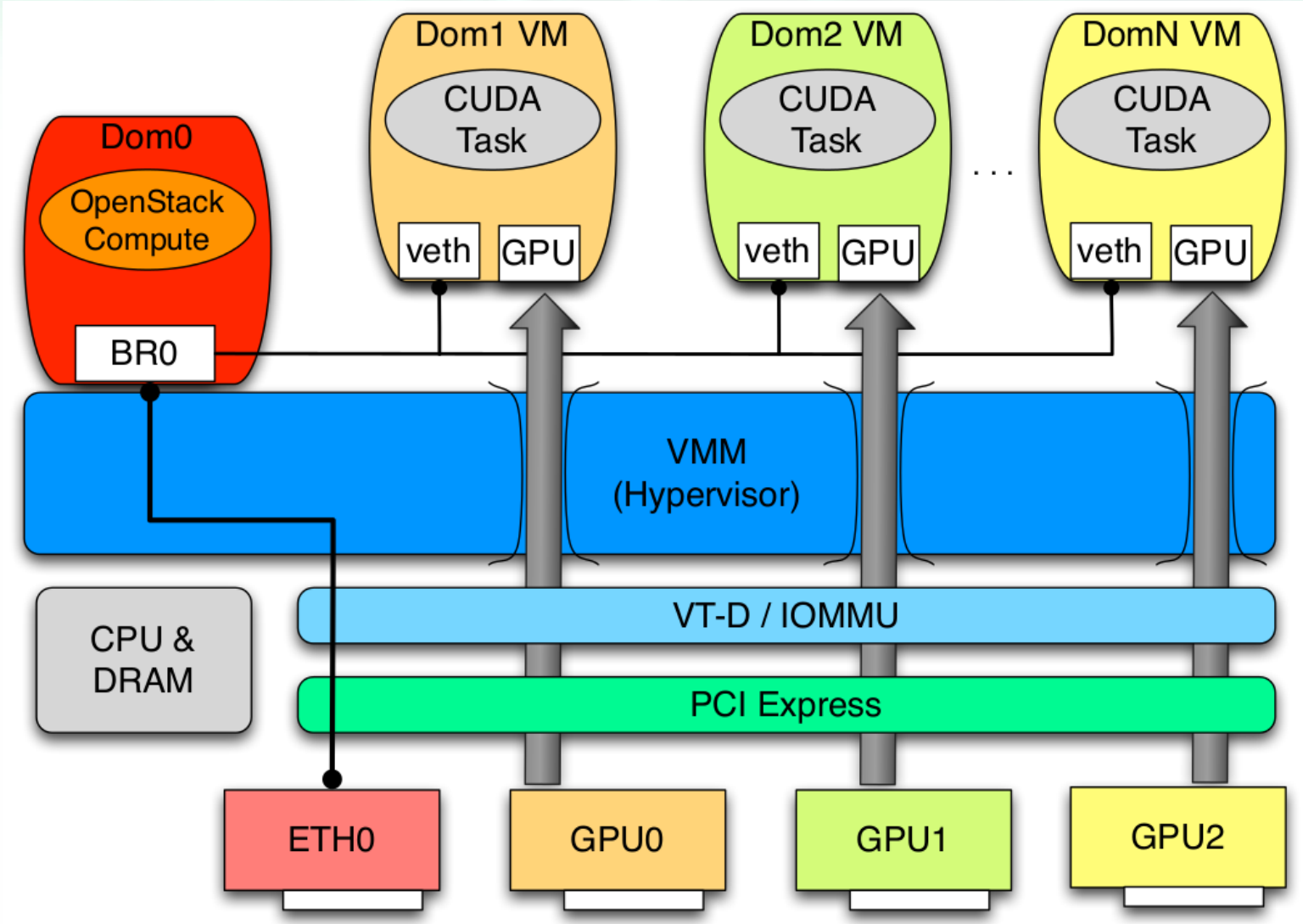
- Can use remote GPUs, but all data goes over the network
  - Can be very inefficient for applications with non-trivial memory movement
- Some implementations do not support CUDA extensions in C
  - Have to separate CPU and GPU code
  - Requires special decouple mechanism
  - Cannot directly drop in code with existing solutions.

From: Andrew J. Younge, John Paul Walters, Stephen Crago, Geoffrey C. Fox, Evaluating GPU Passthrough in Xen for High Performance Cloud Computing, in High-Performance Grid and Cloud Computing Workshop at the 28th IEEE International Parallel and Distributed Processing Symposium. 2014

Direct GPU Virtualization

- Allow VMs to directly access GPU hardware
- Enables CUDA and OpenCL code!
- Utilizes PCI Passthrough of device to guest VM
  - Uses hardware directed I/O virt (VT-d or IOMMU)
  - Provides direct isolation and security of device
  - Removes host overhead entirely
- Creates a direct 1-1 mapping between device and guest

Hardware Virtualization

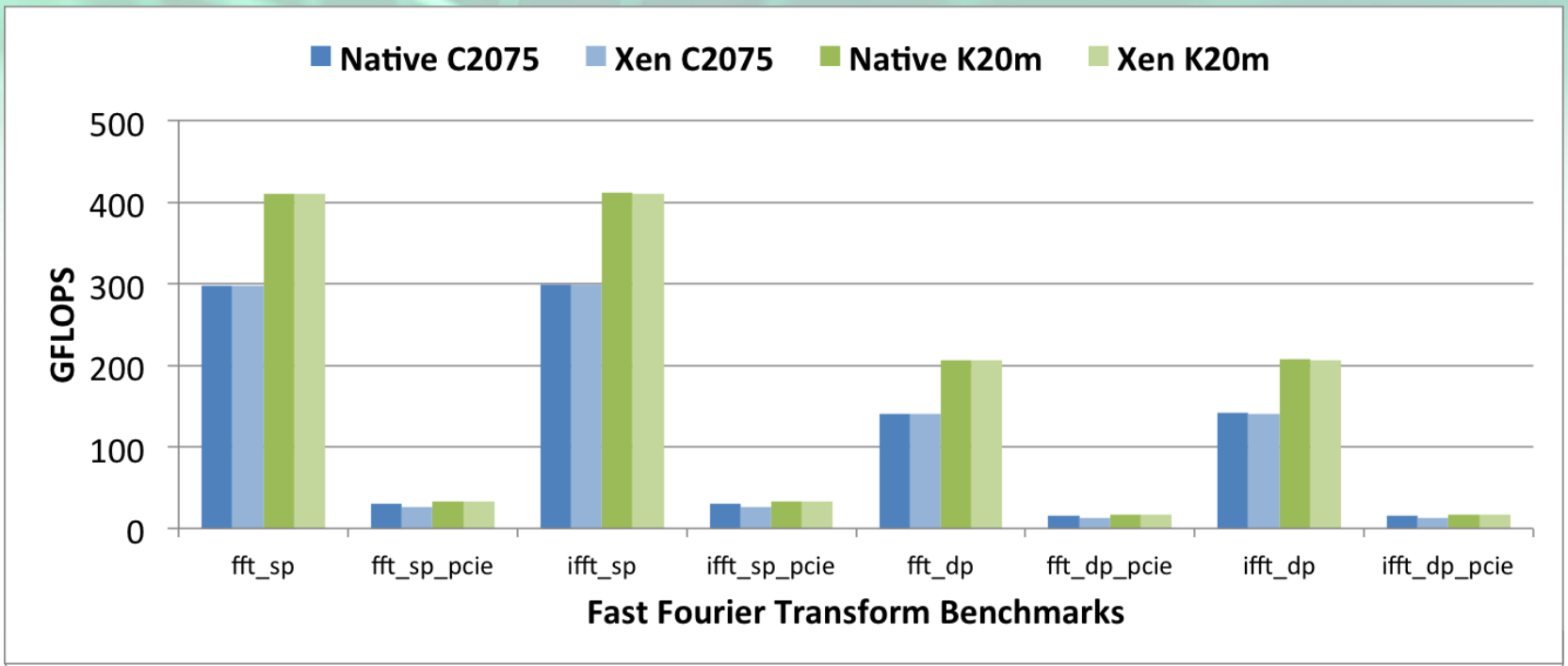


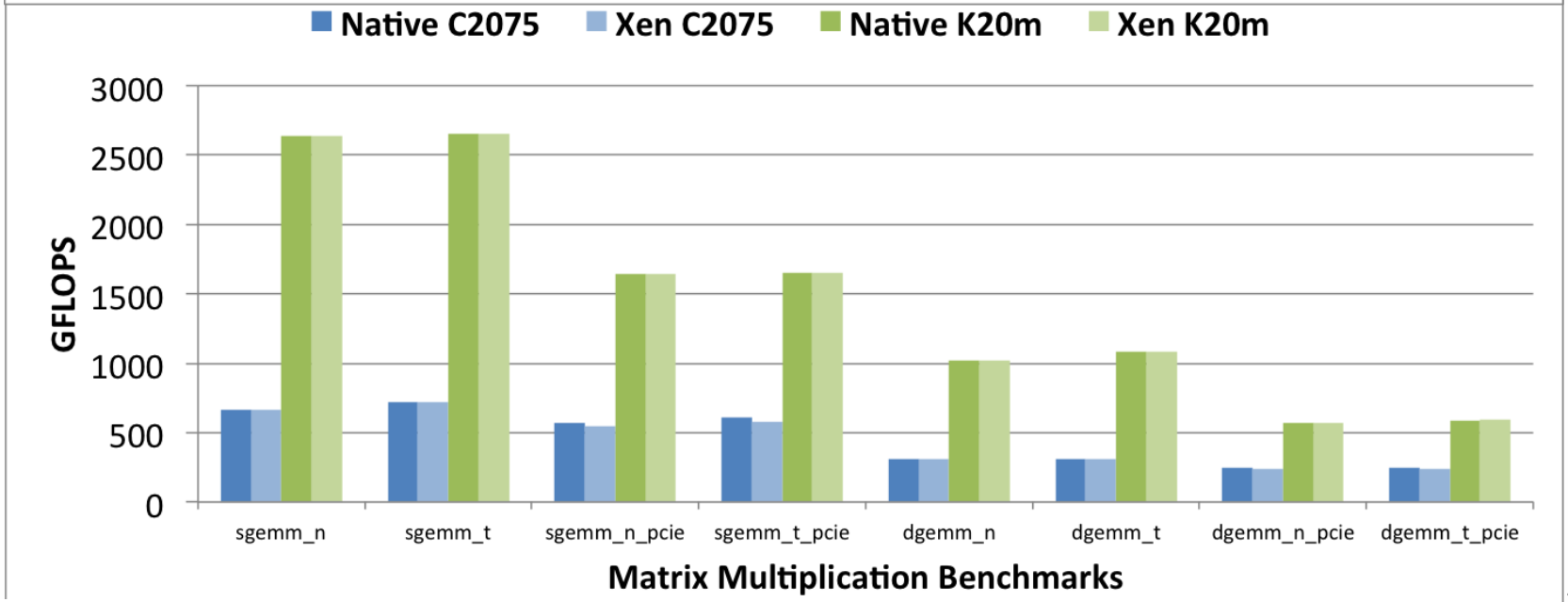
From: Andrew J. Younge, John Paul Walters, Stephen Crago, Geoffrey C. Fox, Evaluating GPU Passthrough in Xen for High Performance Cloud Computing, in High-Performance Grid and Cloud Computing Workshop at the 28th IEEE International Parallel and Distributed Processing Symposium. 2014

Hardware Setup

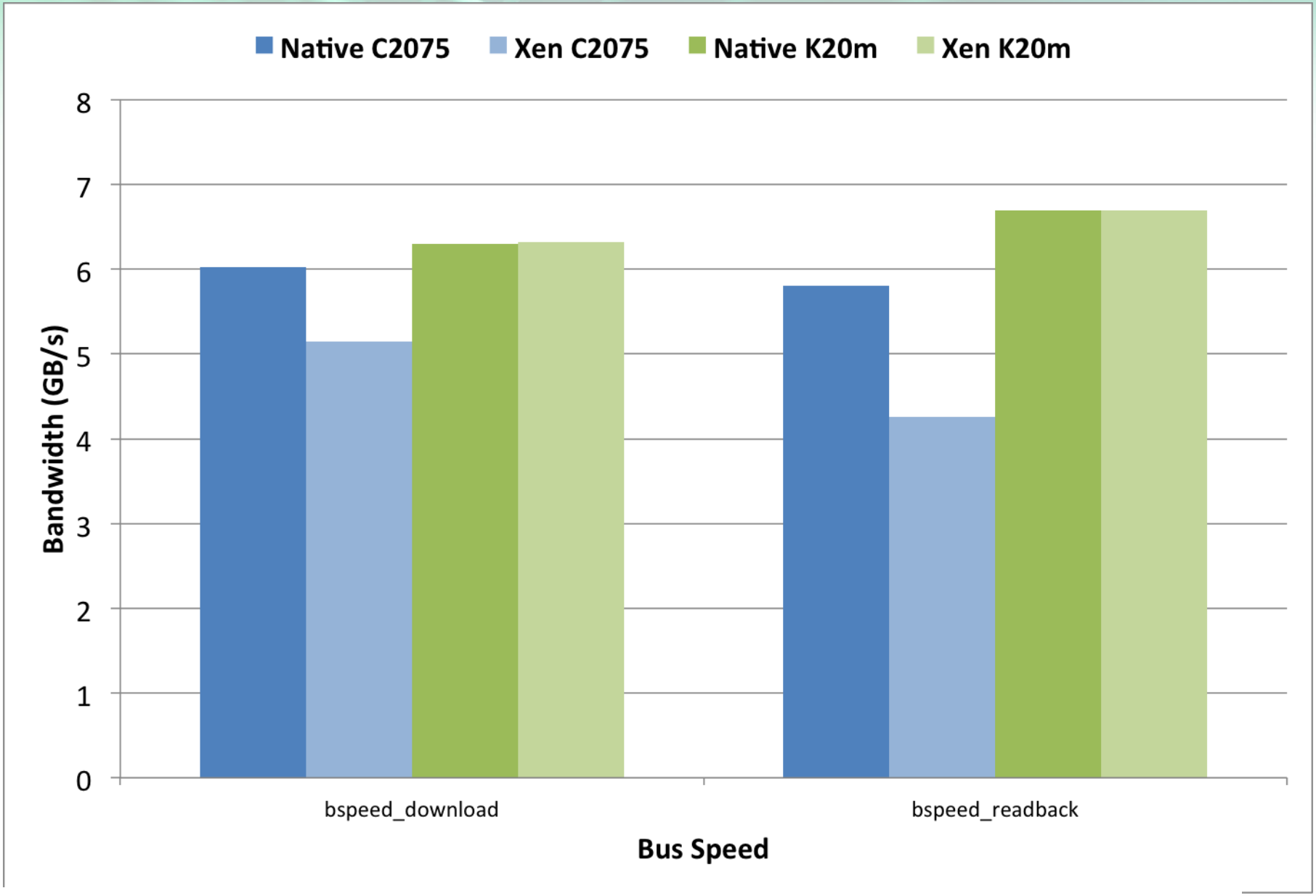
	Westmere + Fermi	Sandy Bridge + Kepler
Name	Delta (IU/FutureGrid)	Bespin (ISI)
CPU (cores)	2xX5660 (12)	2xE5-2670 (16)
Clock Speed	2.6 GHz	2.6 GHz
RAM	192 GB	48 GB
NUMA Nodes	2	2
GPU	2xC2075	1xK20m
PCI-Express	2.0	3.0 (with bug)
Release	~2011	~2013

From:John Paul Walters, Andrew J. Younge, Dong-In Kang, Ke-Thia Yao, Mikyung Kang, Stephen P. Crago, Geoffrey C. Fox, GPU-Passthrough Performance: A Comparison of KVM, Xen, VMWare ESXi, and LXC for CUDA and OpenCL Applications, in Proceedings of the 7th IEEE International Conference on Cloud Computing (CLOUD 2014).





From: Andrew J. Younge, John Paul Walters, Stephen Crago, Geoffrey C. Fox, Evaluating GPU Passthrough in Xen for High Performance Cloud Computing, in High-Performance Grid and Cloud Computing Workshop at the 28th IEEE International Parallel and Distributed Processing Symposium. 2014



From: Andrew J. Younge, John Paul Walters, Stephen Crago, Geoffrey C. Fox, Evaluating GPU Passthrough in Xen for High Performance Cloud Computing, in High-Performance Grid and Cloud Computing Workshop at the 28th IEEE International Parallel and Distributed Processing Symposium. 2014

GPU Discussion

- GPU Passthrough possible in Xen
- Overhead is minimal for GPU computation
  - Sandy-Bridge/Kepler has < 1.2% overall overhead
    Westmere/Fermi has < 1% computational overhead, but worst-case ~15% due to PCI-Express bus
  - PCIE overhead not likely due to VT-d mechanisms
  - NUMA configuration in Westmere CPU architecture
- GPU PCI Passthrough performs better than other front-end remote API solutions

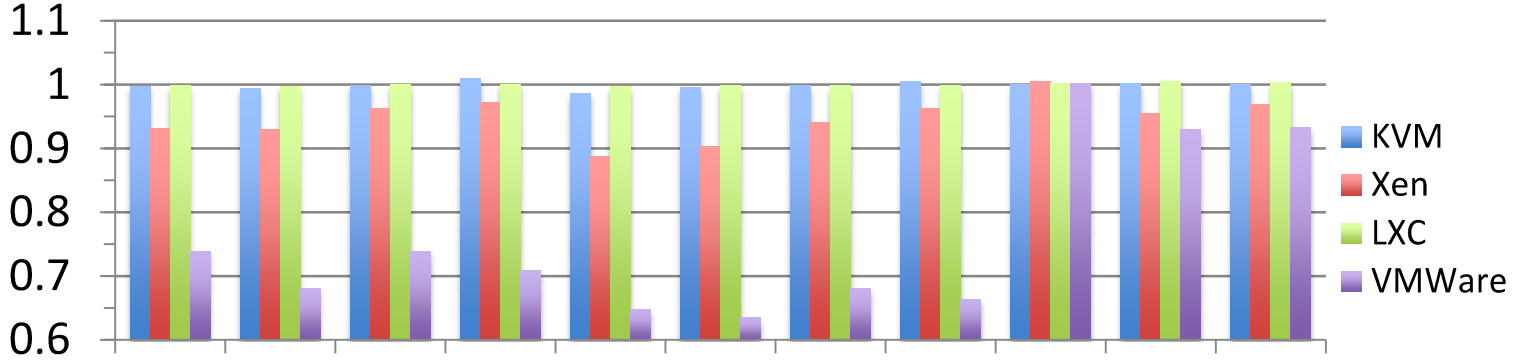
From: Andrew J. Younge, John Paul Walters, Stephen Crago, Geoffrey C. Fox, Evaluating GPU Passthrough in Xen for High Performance Cloud Computing, in High-Performance Grid and Cloud Computing Workshop at the 28th IEEE International Parallel and Distributed Processing Symposium. 2014

GPU Hypervisor Experiment

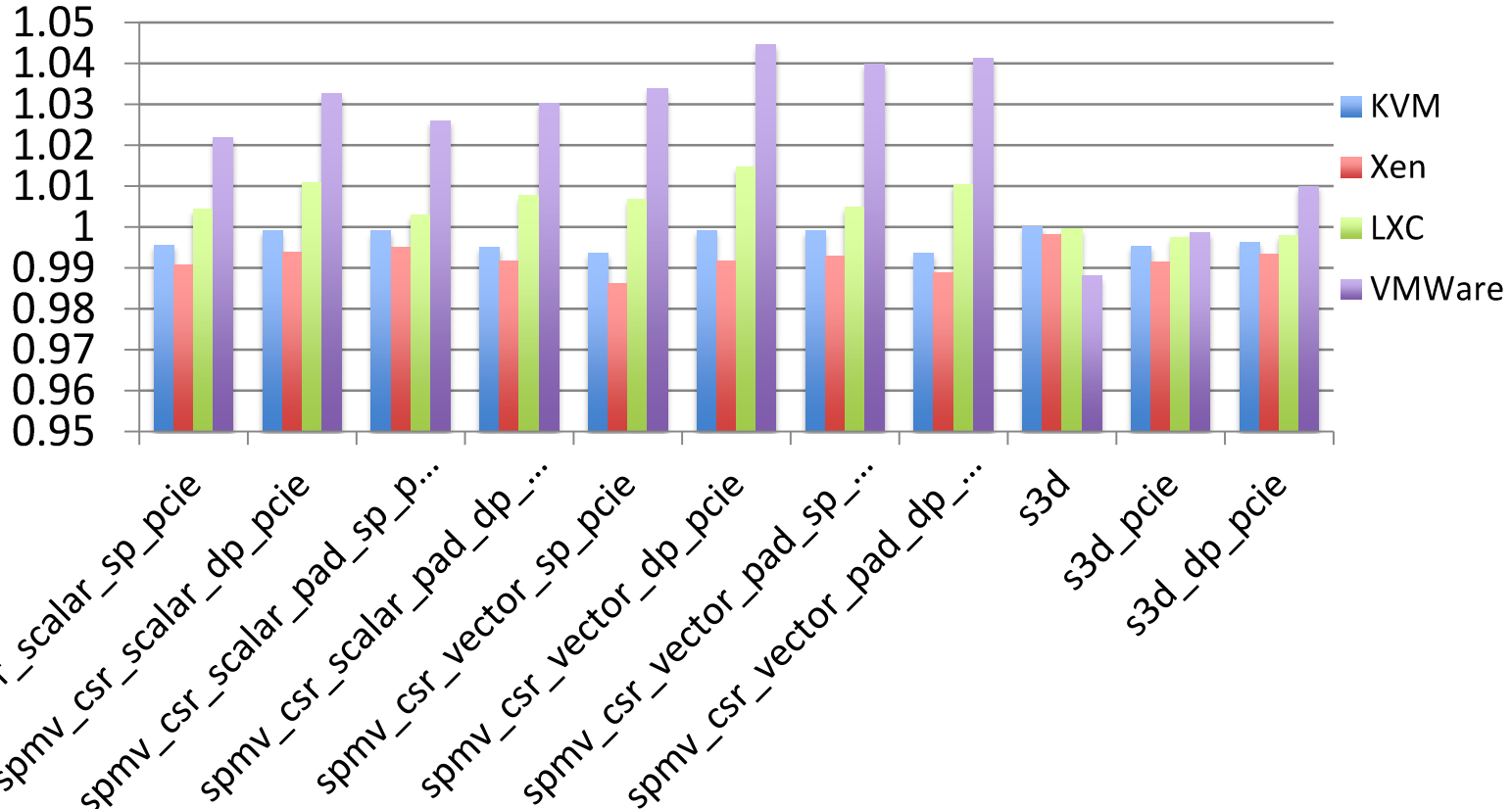
- In 2012, the Xen GPU Passthrough implementation was novel for Nvidia GPUs
- Today GPUs available through most of the major hypervisors
  - KVM, VMWare ESXi, Xen, LXC
- Developed similar methods in KVM
  - Based on kvm/qemu VFIO in new kernel >= 3.9
- What are the performance implications?
  - Near-native performance possible?
- What lessons can we learn?

- Benchmarks
  - Microbenchmarks: SHOC OpenCL (70 total benchmarks)
  - LAMMPS: measures hybrid multicore CPU + GPU
  - GPU-LIBSVM: characteristic of big data applications
  - LULESH: hydrodynamics application
- Platforms
  - Westmere with Fermi C2075
  - Sandy Bridge with Kepler K20m
- Control for NUMA effects

Delta - SHOC OpenCL Level 1, Level 2 Outliers

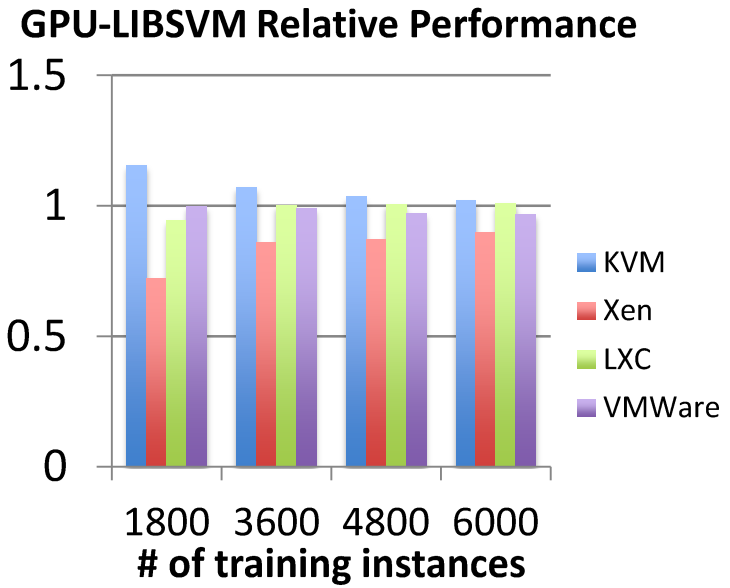


Bespin - SHOC OpenCL Level 1, Level 2 Outliers

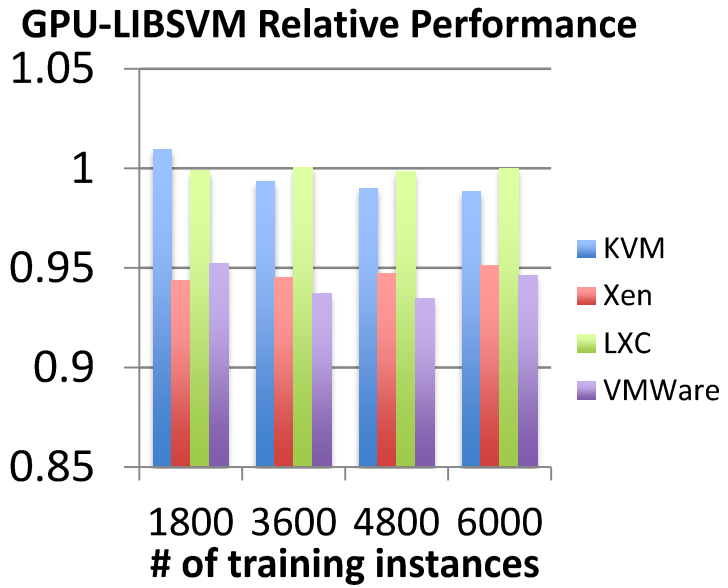


GPU-LIBSVM Results

Delta C2075 Results



Bespin K20m Results

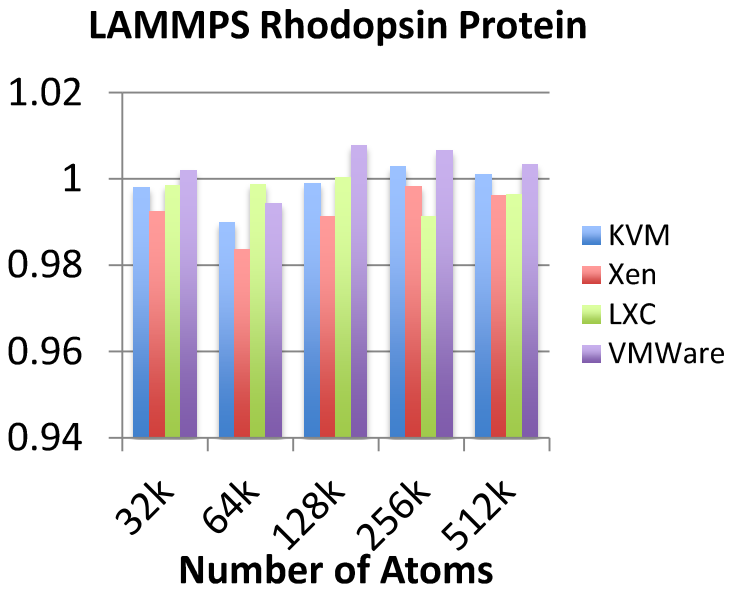


- Unexpected performance improvement for KVM on both systems
  - Most pronounced on Westmere/Fermi platform
- This is likely due to the use of transparent hugepages (THP)
  - Back the entire guest memory with hugepages
  - Improves TLB performance
  - More investigation needed to confirm

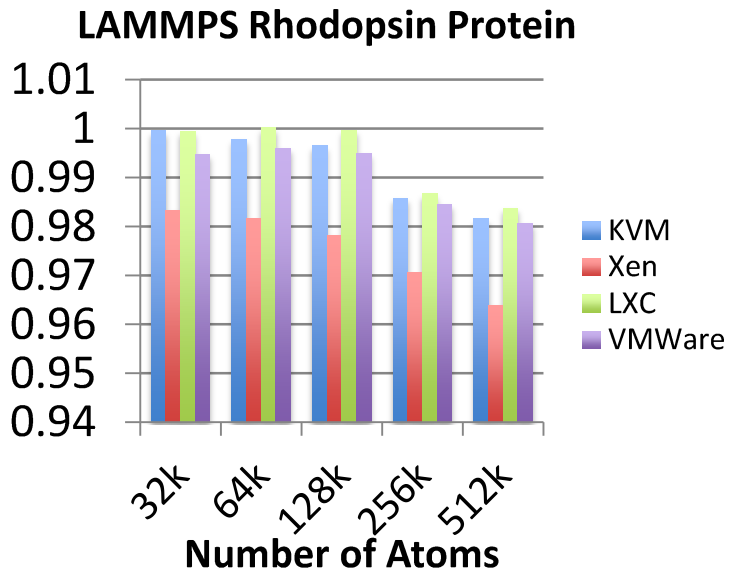
From:John Paul Walters, Andrew J. Younge, Dong-In Kang, Ke-Thia Yao, Mikyung Kang, Stephen P. Crago, Geoffrey C. Fox, GPU-Passthrough Performance: A Comparison of KVM, Xen, VMWare ESXi, and LXC for CUDA and OpenCL Applications, in Proceedings of the 7th IEEE International Conference on Cloud Computing (CLOUD 2014).

LAMMPS Rhodopsin Protein Results

Delta C2075 Results



Bespin K20m Results

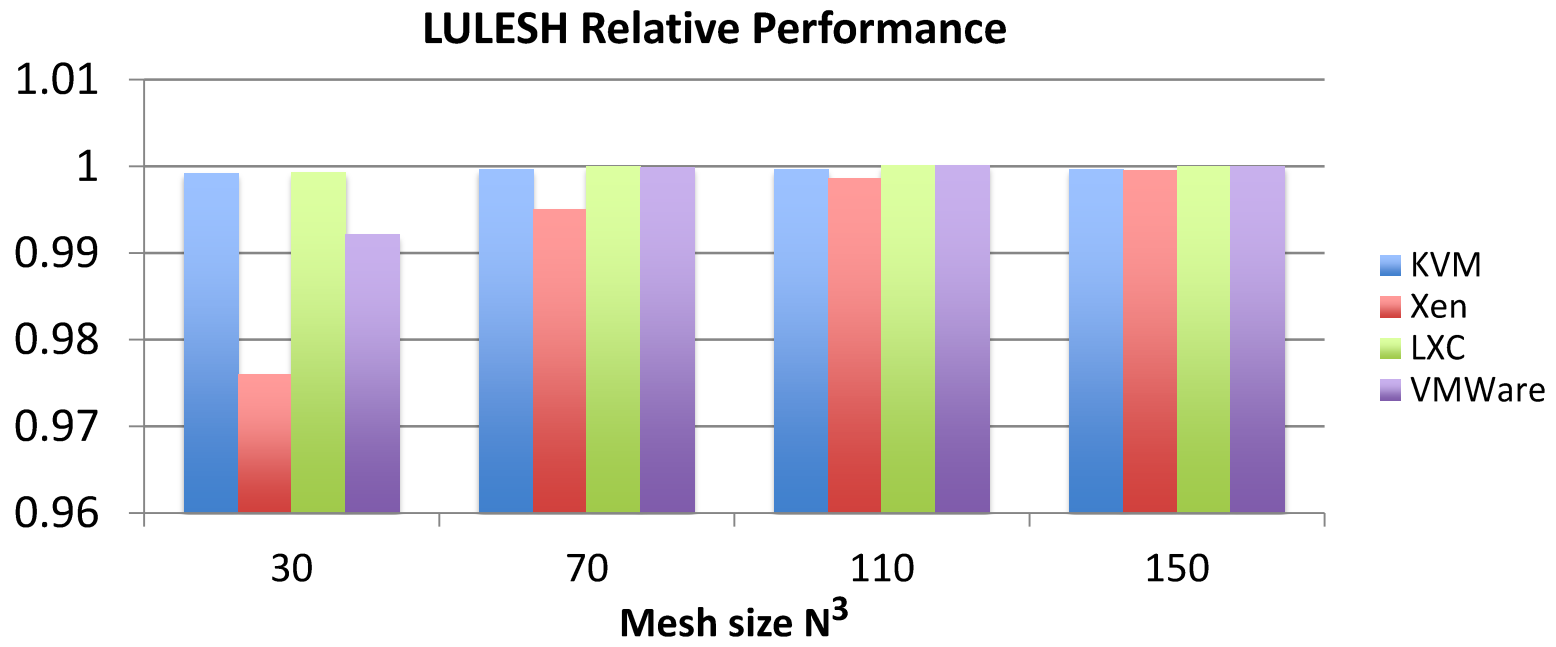


- LAMMPS unique among the benchmarks
  - Exercises multiple CPU cores in addition to GPU
  - Multiple packages available (using GPU)
- Demonstrates high efficiency across both platforms
  - Unexpectedly higher efficiency for Westmere architecture
- Implications for heterogeneous workloads:
  - SMP CPU + GPU efficiency remains high

From:John Paul Walters, Andrew J. Younge, Dong-In Kang, Ke-Thia Yao, Mikyung Kang, Stephen P. Crago, Geoffrey C. Fox, GPU-Passthrough Performance: A Comparison of KVM, Xen, VMWare ESXi, and LXC for CUDA and OpenCL Applications, in Proceedings of the 7th IEEE International Conference on Cloud Computing (CLOUD 2014).

LULESH Hydrodynamics Performance

Bespin K20m Results



LULESH (K20m only)

Highly compute-intensive, little data movement

Expect little virtualization overhead

Initially slight overhead from Xen

Decreases as mesh resolution (N^3) increases

From:John Paul Walters, Andrew J. Younge, Dong-In Kang, Ke-Thia Yao, Mikyung Kang, Stephen P. Crago, Geoffrey C. Fox, GPU-Passthrough Performance: A Comparison of KVM, Xen, VMWare ESXi, and LXC for CUDA and OpenCL Applications, in Proceedings of the 7th IEEE International Conference on Cloud Computing (CLOUD 2014).

Lessons Learned – Hypervisor Performance

- KVM consistently yields near-native performance across architectures
- VMWare's performance inconsistent
  - Near-native on Sandy Bridge, high overhead on Westmere
- Xen performed consistently average across both architectures
- LXC performed closest to native
  - Unsurprising, given LXC's design
  - Trades performance for flexibility
- Given these results we see KVM as holding a slight edge for GPU passthrough

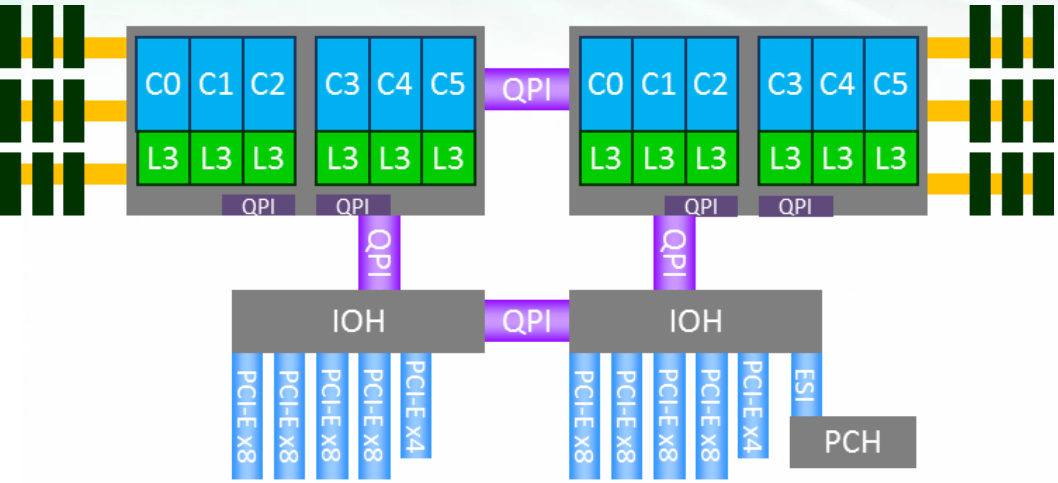
From:John Paul Walters, Andrew J. Younge, Dong-In Kang, Ke-Thia Yao, Mikyung Kang, Stephen P. Crago, Geoffrey C. Fox, GPU-Passthrough Performance: A Comparison of KVM, Xen, VMWare ESXi, and LXC for CUDA and OpenCL Applications, in Proceedings of the 7th IEEE International Conference on Cloud Computing (CLOUD 2014).

Virtualized HPC

- Virtualization of high performance workloads historically controversial
  - Nahelem/Westmere results suggest this was *sometimes* legitimate
  - More than 10% overhead possible
- Recent architectures (e.g. Sandy Bridge) and hypervisor advances have nearly erased those overheads
  - Lowest performing hypervisor (Xen) within 95% of native
  - KVM can perform at “near-native”
  - Improved CPU integration with PCI-Express bus

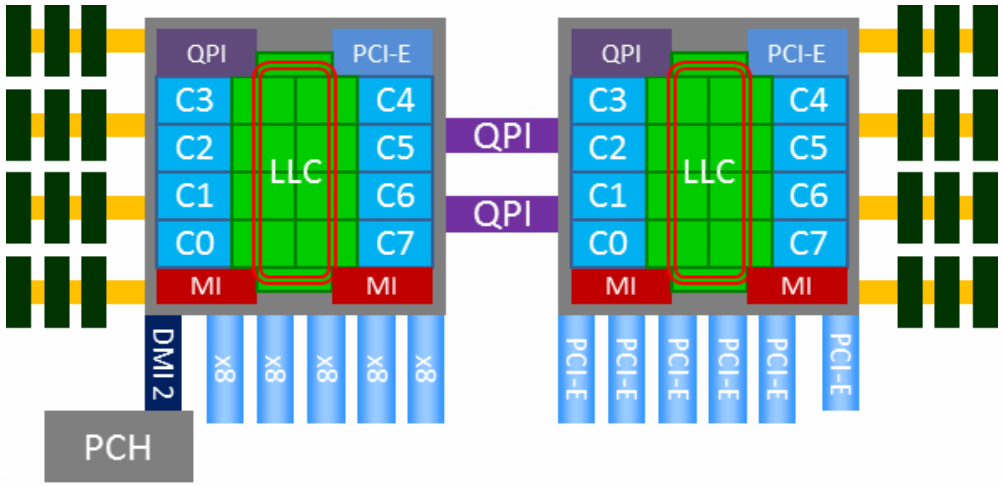
From:John Paul Walters, Andrew J. Younge, Dong-In Kang, Ke-Thia Yao, Mikyung Kang, Stephen P. Crago, Geoffrey C. Fox, GPU-Passthrough Performance: A Comparison of KVM, Xen, VMWare ESXi, and LXC for CUDA and OpenCL Applications, in Proceedings of the 7th IEEE International Conference on Cloud Computing (CLOUD 2014).

CPU Architecture



Westmere/Nehalem

- Single QPI connection between NUMA sockets
- Intel 5500 chipset for I/O Hub (IOH) with own QPI
  - PCI-E from 2 IOHs



Sandy Bridge

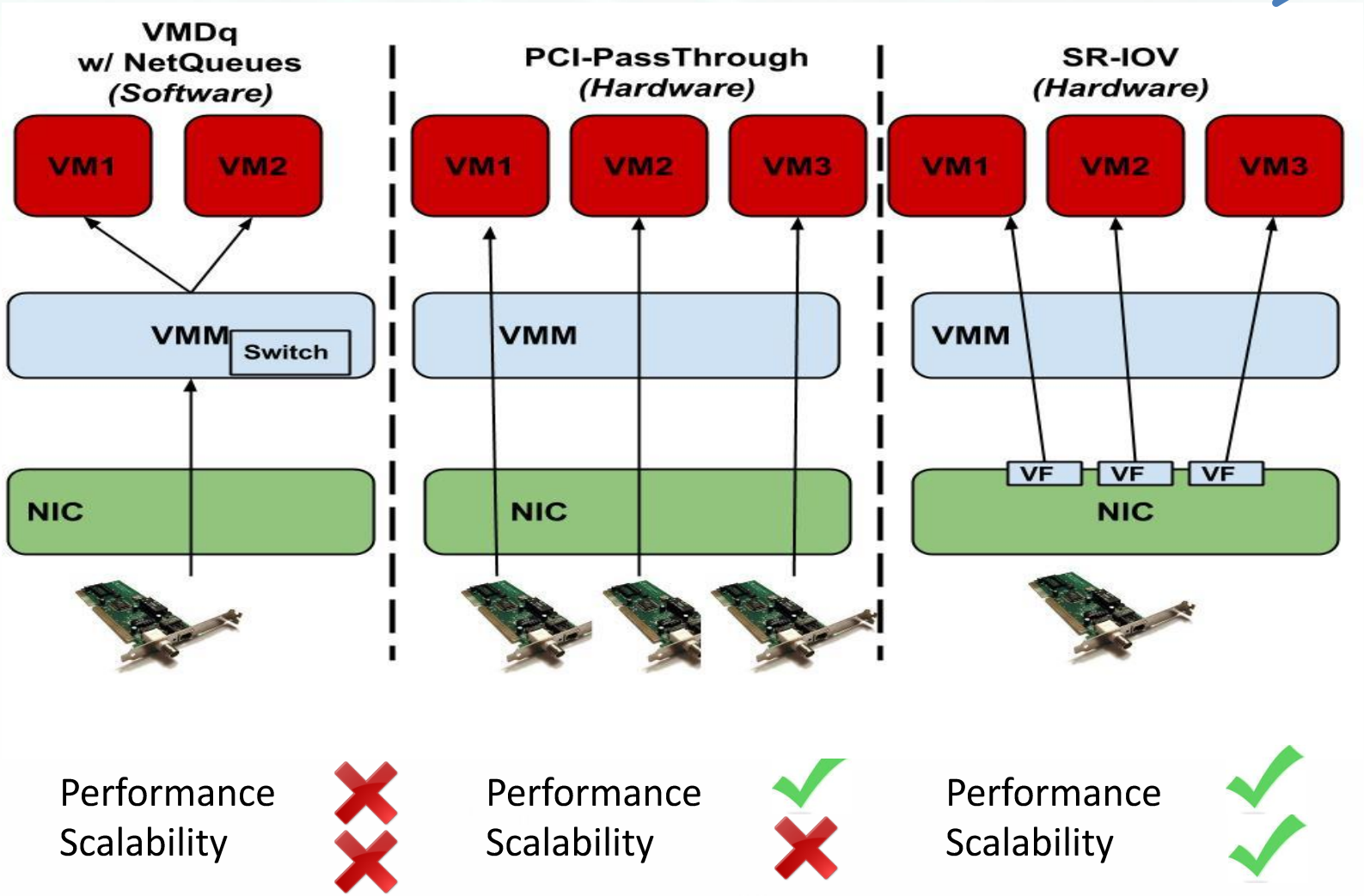
- Dual QPI connection between NUMA sockets
- PCI-E built into processor
  - If VMs pinned, no QPI traversal

I/O Interconnect

- While hypervisor performances improves, I/O support in virtualized environments still suffer
  - Bridged 1GbE or 10GbE often state-of-the-art for IaaS solutions (Amazon EC2, FutureGrid, etc)
  - Latency also suffers with emulated drivers
- Need for high performance, low latency interconnect – InfiniBand

Background - Overview

Overhead Reduction



Future

From: Malek Musleh, Vijay Pai, John Paul Walters, Andrew J. Younge, Stephen P. Crago, Bridging the Virtualization Performance Gap for HPC using SR-IOV for InfiniBand, in Proceedings of the 7th IEEE International Conference on Cloud Computing (CLOUD 2014).

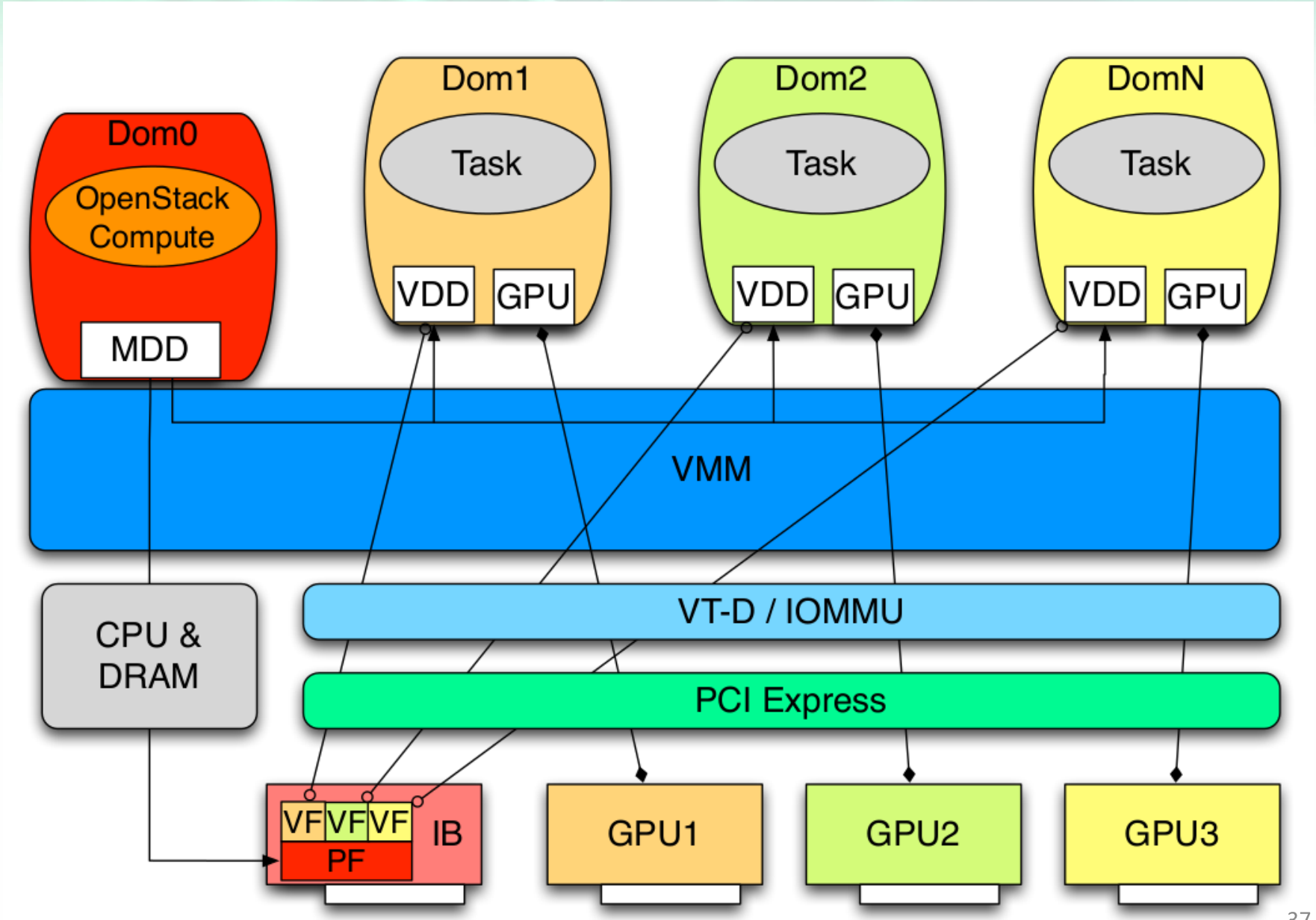
SR-IOV InfiniBand

- SR-IOV enabled InfiniBand drivers now available
  - OFED support with KVM for CX2 & CX3 cards
- Initial evaluation shows promise for IB-enabled VMs
  - *SR-IOV Support for Virtualization on InfiniBand Clusters: Early Experience*, Jose et al – CCGrid 2013
  - *Exploring Infiniband Hardware Virtualization in OpenNebula towards Efficient High-Performance Computing*, Ruivo et al – CCGrid 2014
  - ** *Bridging the Virtualization Performance Gap for HPC Using SR-IOV for InfiniBand*, Musleh et al – IEEE CLOUD 2014 **
  - SDSC Comet

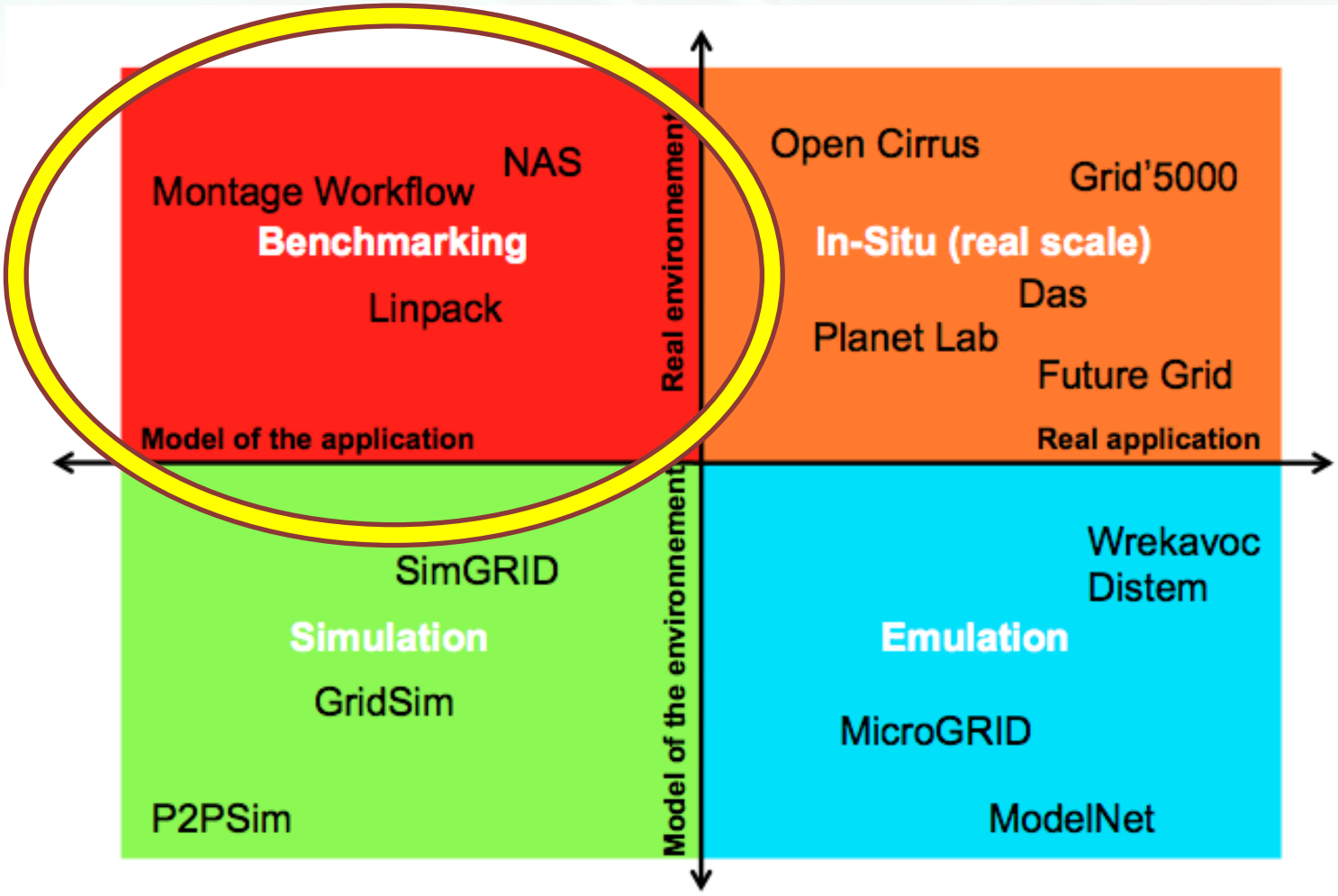
InfiniBand Optimizations

- With InfiniBand SR-IOV, bandwidth is near-native, but high latency overhead remains high
- Observation: Native InfiniBand optimizations may be sub-optimal for SR-IOV
- Possible Solution: Tune parameters for better performance with SR-IOV in VMs
  - Interrupt Moderation & Coalescing
  - IRQ Balancing
  - Shared Receive Queue

From: Malek Musleh, Vijay Pai, John Paul Walters, Andrew J. Younge, Stephen P. Crago, Bridging the Virtualization Performance Gap for HPC using SR-IOV for InfiniBand, in Proceedings of the 7th IEEE International Conference on Cloud Computing (CLOUD 2014).

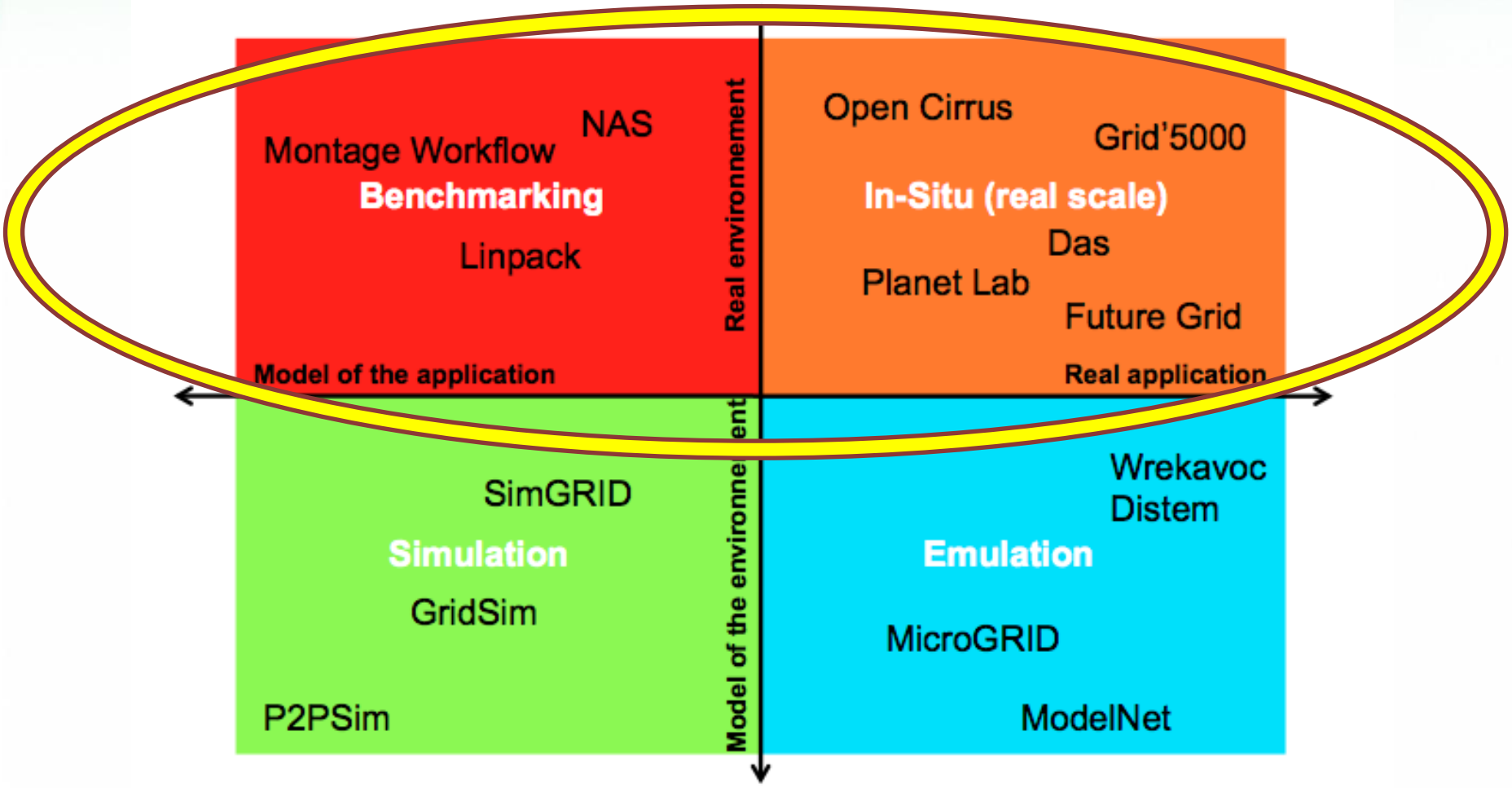

Dom0
OpenStack Compute
MDD
Dom1
Task
VDD
GPU
Dom2
Task
VDD
GPU
DomN
Task
VDD
GPU
VMM
CPU & DRAM
VT-D / IOMMU
PCI Express
VF
VF
VF
PF
IB
GPU1
GPU2
GPU3

Experimental Computer Science



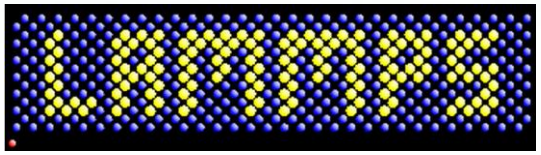
From "Supporting Experimental Computer Science"

Experimental Computer Science



From "Supporting Experimental Computer Science"

Real-world Applications – Molecular Dynamics Simulation

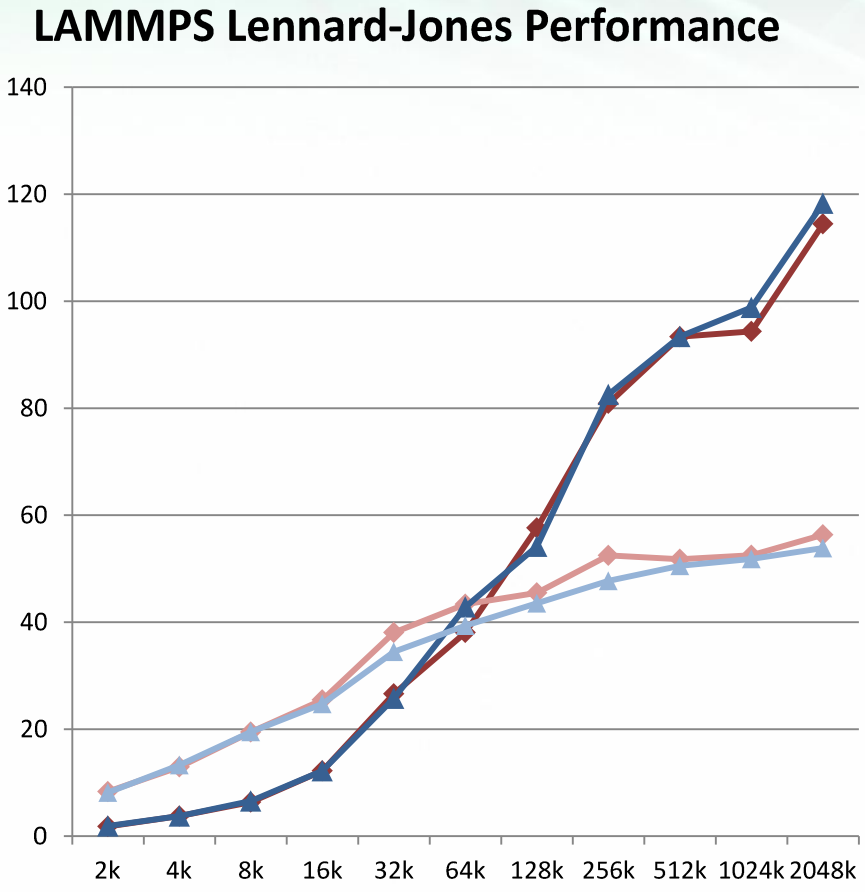


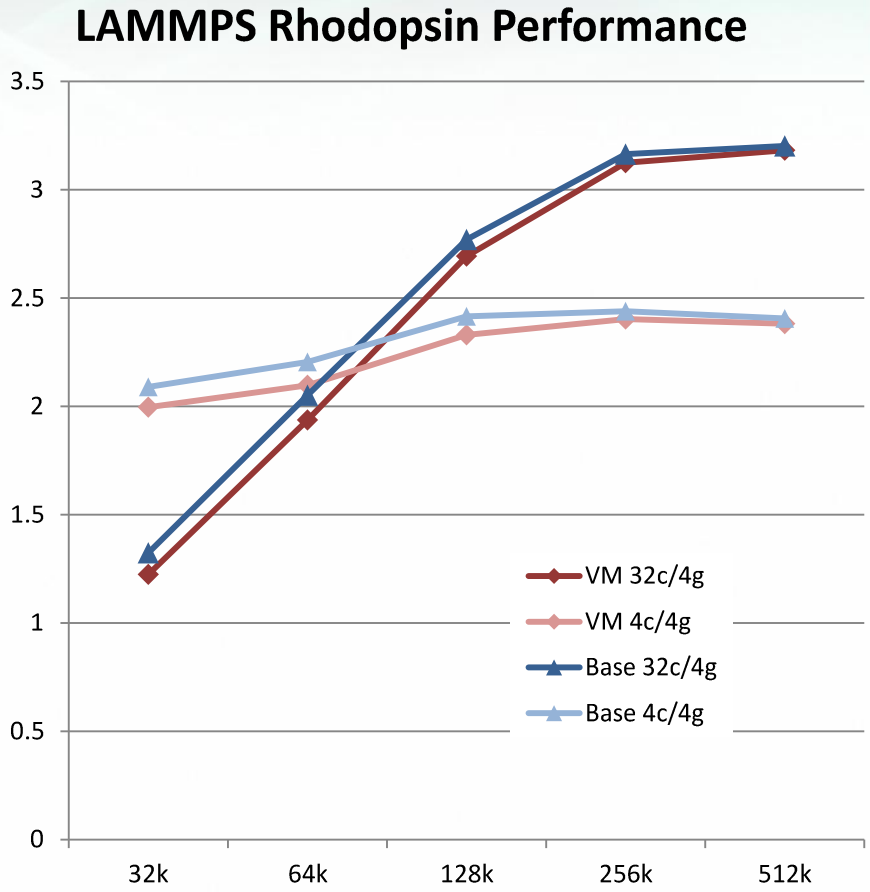
- LAMMPS - "Large-scale Atomic/Molecular Massively Parallel Simulator“
- Very common MD simulator
- From Sandia National Laboratories
- Uses MPI and has the GPU package for hybrid CPU and GPU computation

- HOOMD-blue is a general-purpose particle simulation toolkit
- From University of Michigan
- It scales from a single CPU core to thousands of GPUs with MPI
- HOOMD also has support for GPUDirect, introduced in CUDA 5

Future Grid

LAMMPS LJ & RHODO

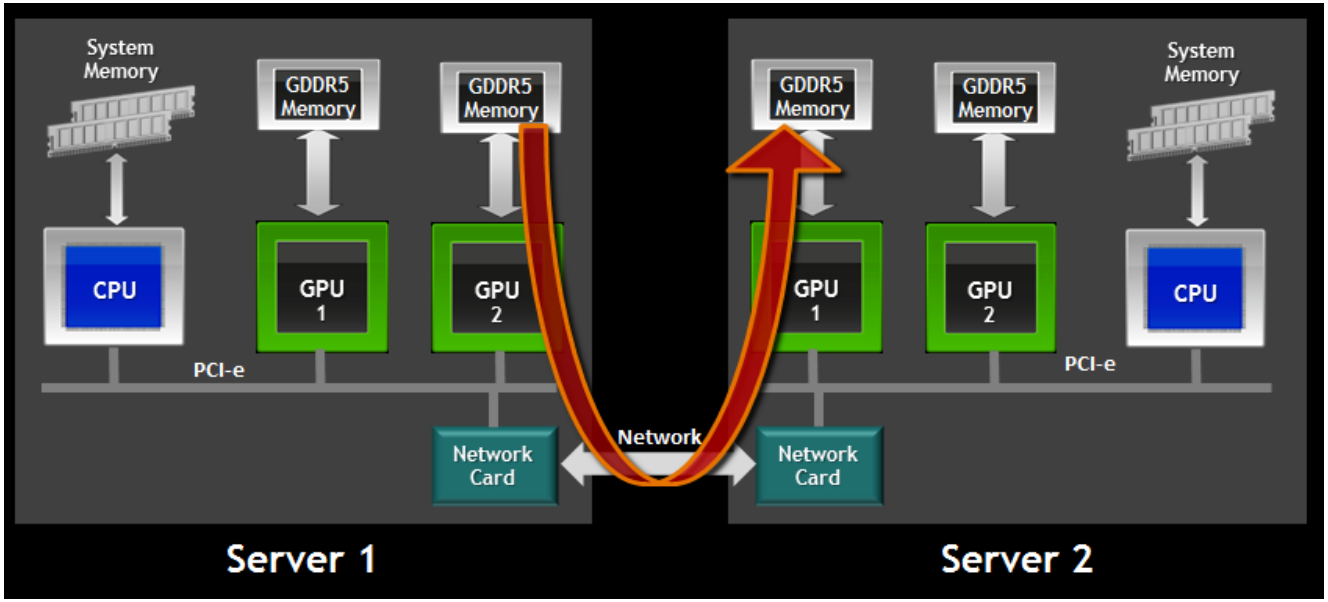




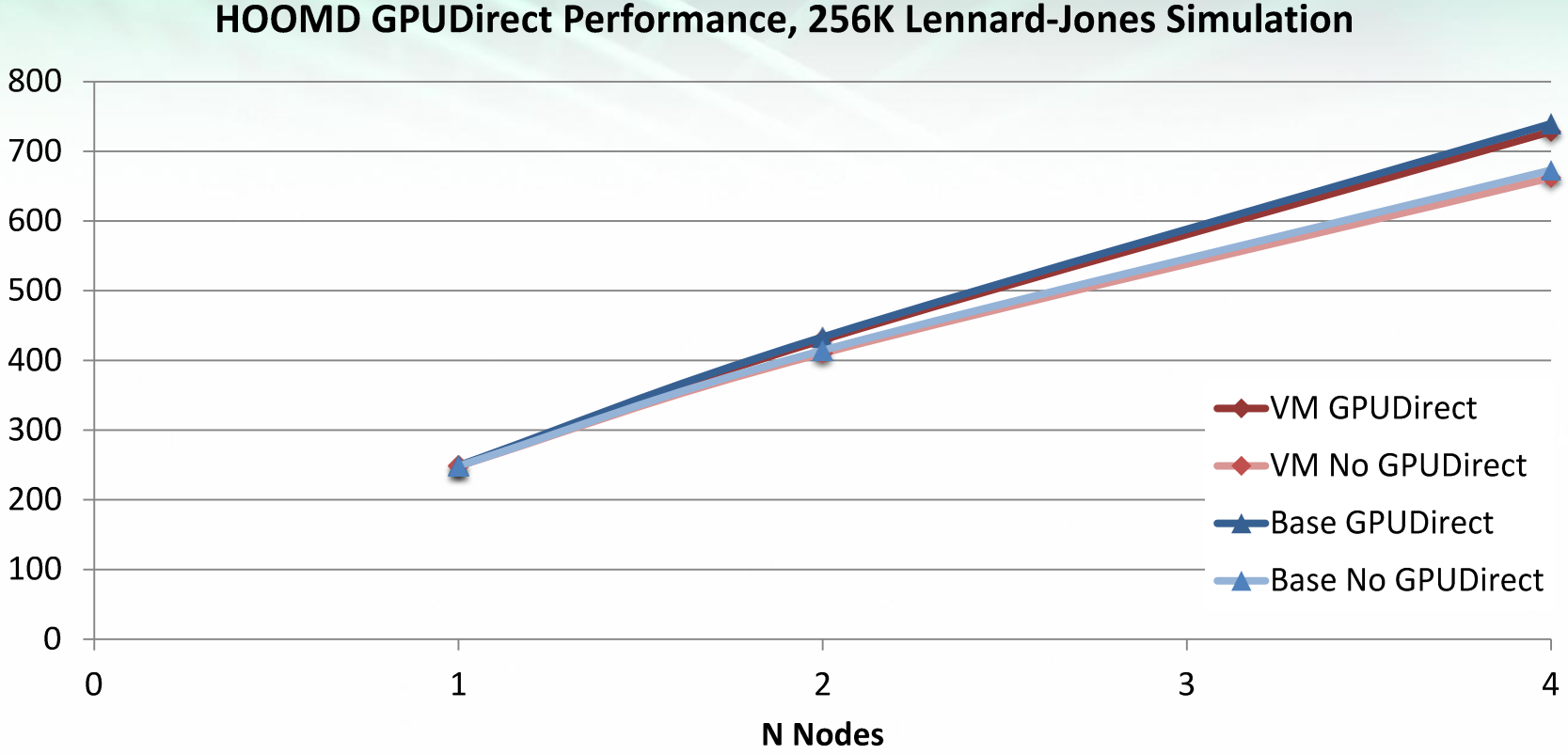
- VMs running LAMMPs achieve near-native performance at 32 cores & 4GPUs
  - 96.7% efficiency for LJ
  - 99.3% efficiency for Rhodo

GPU Direct

- GPUDirect facilitates multi-GPU computation
  - v1 avoids dual CPU buffers (2010)
  - v2 P2P communication between intra-GPUs (2011)
  - v3 avoid CPU entirely with RDMA via InfiniBand (2013)



HOOMD-Blue



- GPUDirect has small but noticable improvement (~9%) in performance for MPI+CUDA applications.
- Both HOOMD simulations, with and without GPUDirect, perform very near-native.
  - GPUDirect 98.5% efficiency
  - non-GPUDirect 98.4% efficiency

Future Grid

Next Steps

- Deploy on Delta
  - Investigate OpenStack Icehouse status
  - Upgrade network to FDR?
- Scale up LAMMPS and HOOMD experiments
  - Evaluate GPUDirect utility
- Test VirtualCalifornia Earthquake Simulation
  - Compare with GigE and native IB results

OpenStack Integration

- Integrated into OpenStack “Havana” fork
  - Xen support for full virtualization with libvirt
  - Custom Libvirt driver for PCI-Passthrough
  - Use instance_type_extra_specs to specify PCI devs


```
root@test-nvidia-xqcow2-vm-58 ~]# lspci
...
00:04.0 3D controller: NVIDIA Corporation Device 1028 (rev a1)
00:05.0 Network controller: Mellanox Technologies MT27500 Family [ConnectX-3]
```

Future Grid

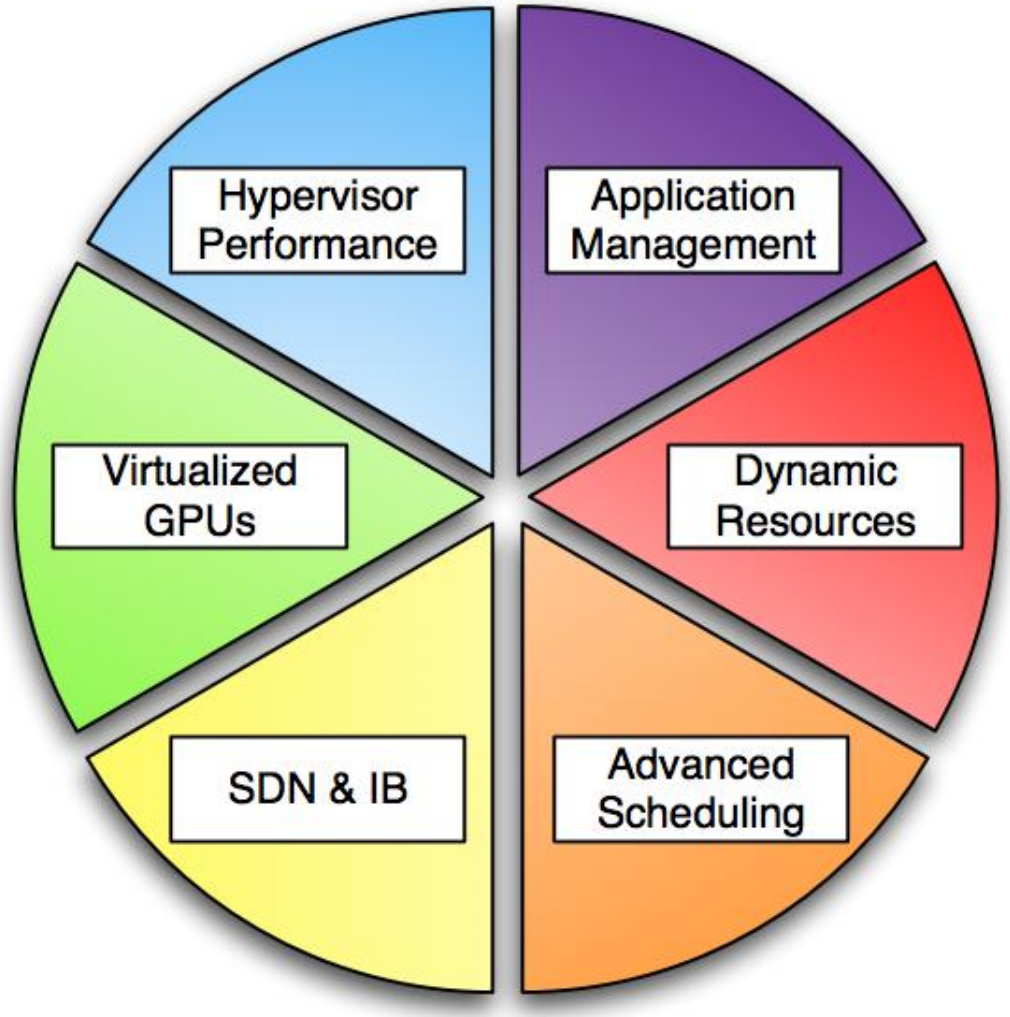
Experimental Deployment: Delta

- 16x 4U nodes in 2 Racks
  - 2x Intel Xeon X5660
  - 192GB Ram
  - Nvidia Tesla C2075 Fermi
  - QDR InfiniBand - CX-2
- Management Node
  - OpenStack Keystone, Glance, API, Cinder, Nova-network
- Compute Nodes
  - Nova-compute, Xen, libvirt

VirtualCalifornia

- Models California's earthquake fault system
- Need for dynamic simulations on Cloud infrastructure
- Stress interaction calculations computationally expensive
- Uses large matrix to avoid infrequent calculations
  - Increased memory requirement as element resolution decreases.
  - Communication quickly becomes limiting factor in parallel computation
  - Ethernet fails to scale past 32 processors
- TODO: Evaluate system architecture using VirtualCalifornia simulations

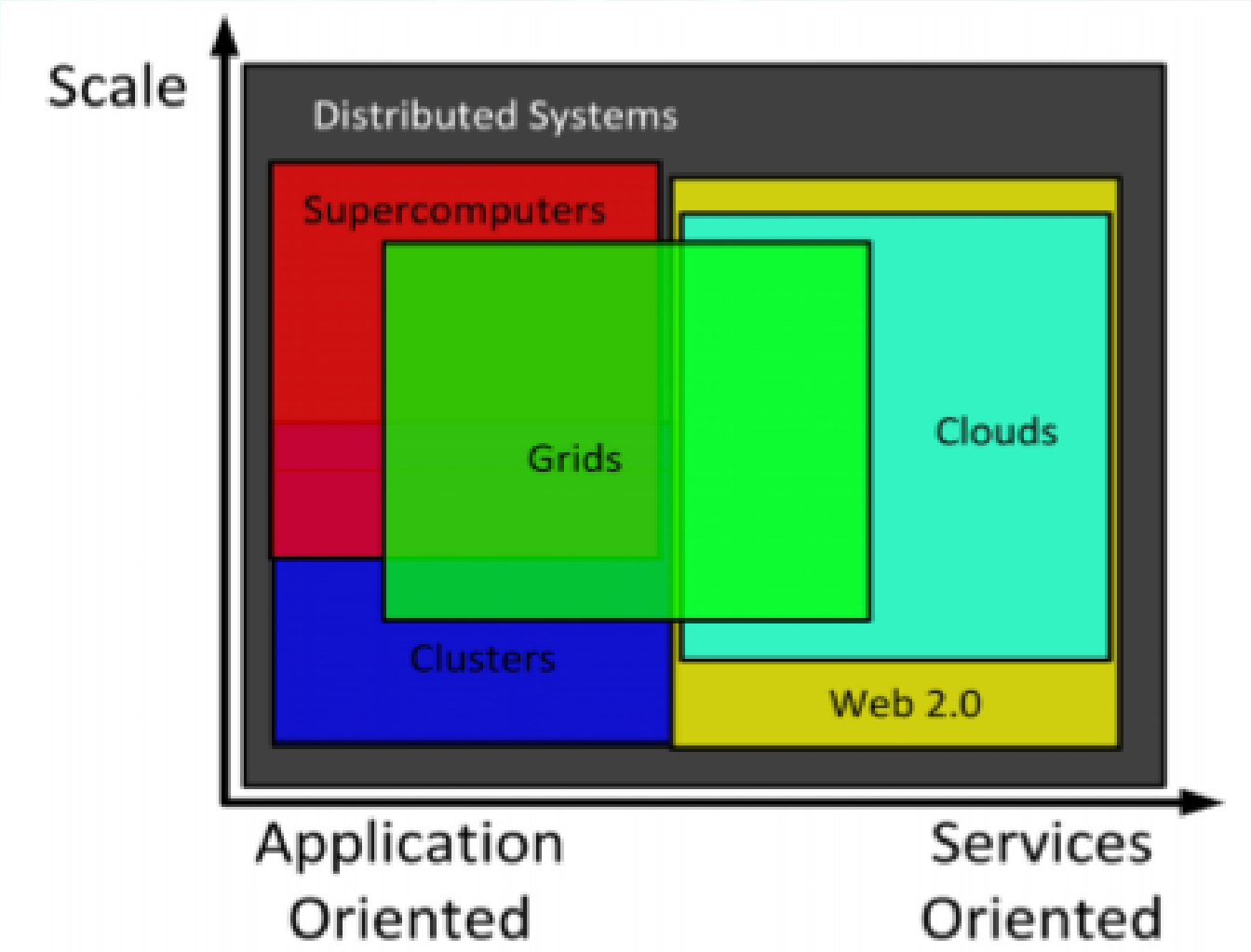
High Performance Cloud Computing Environment



Conclusion

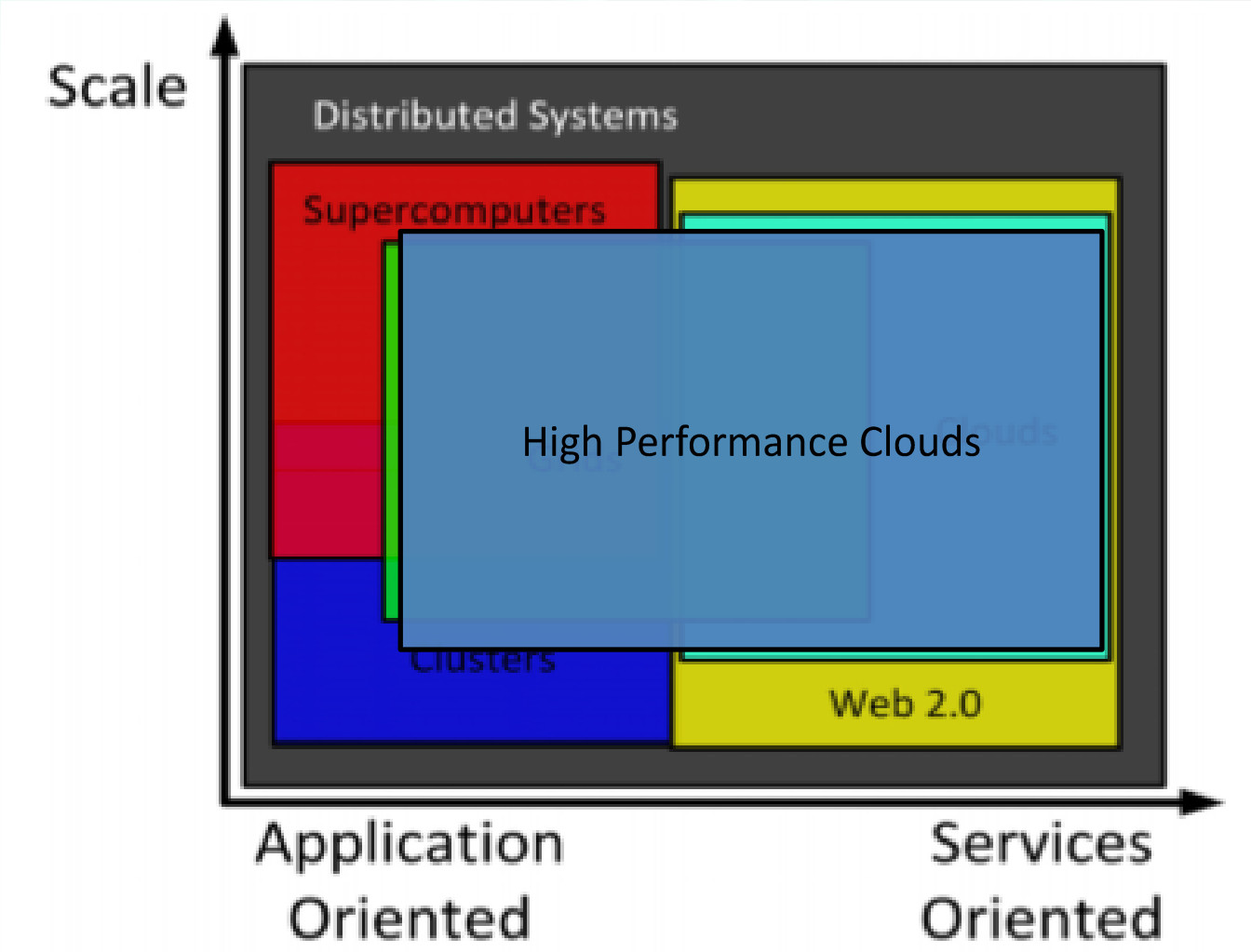
- Today's hypervisors can provide near-native performance for many HPC workloads
  - Careful configuration necessary for best performance
  - NUMA effects still not well understood
- GPUs in VMs now a reality
  - Promising performance via PCI Passthrough
  - Some overhead, but best with new architectures
- InfiniBand SR-IOV = leap in interconnect for IaaS
  - Interrupt tuning may help reduce latency overhead
- Integrate work into OpenStack IaaS Cloud
- Support large scale scientific applications in HPC Cloud
  - Molecular Dynamics simulations
  - NASA Earthquake simulation

Cloud Computing



From: *Cloud Computing and Grid Computing 360-Degree Compared*, Foster et al.

Cloud Computing



From: *Cloud Computing and Grid Computing 360-Degree Compared*, Foster et al.

QUESTIONS?

Andrew J. Younge

Ph.D. Candidate
Indiana University
ajyounge@indiana.edu
http://ajyounge.com

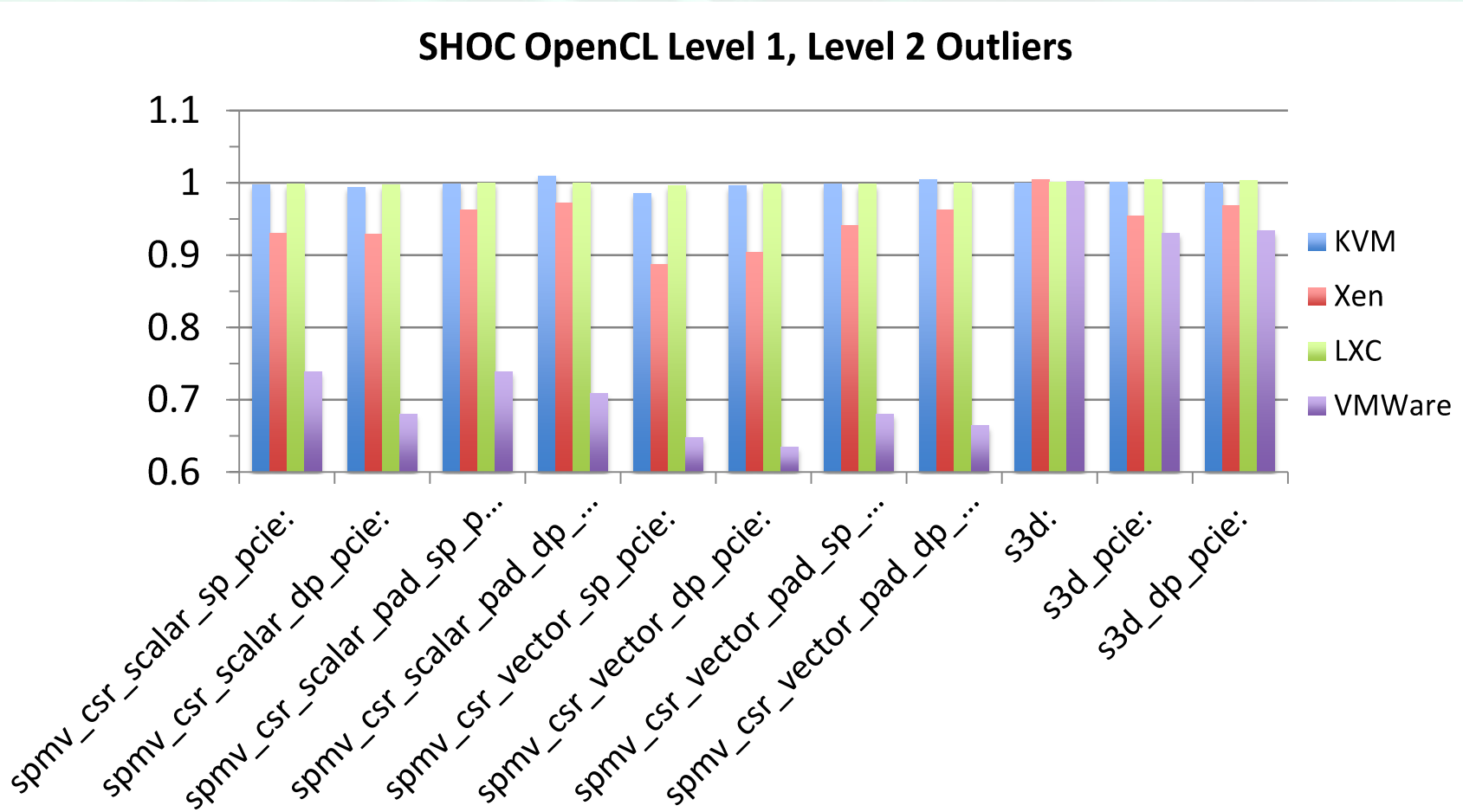
Moving Forward - Horizontally(post PhD)

- Advanced IaaS scheduling
  - Take advantage of NUMA awareness
  - Proximity schedule based on network locality
- Extend PCI Passthrough accellerator model
  - Intel Phi (mic), FPGAs, etc
  - SR-IOV possible?
- Continue work on InfiniBand
  - Auto-tuning of interrupt parameters
  - Evaluate 40GbE vs FDR
  - SDN network integration
- Experiment with alternative “lightweight” architectures (ARM)
- Scale applications to support mid-tier science
  - Utilize test-beds and experimental systems
  - To Petascale and beyond!

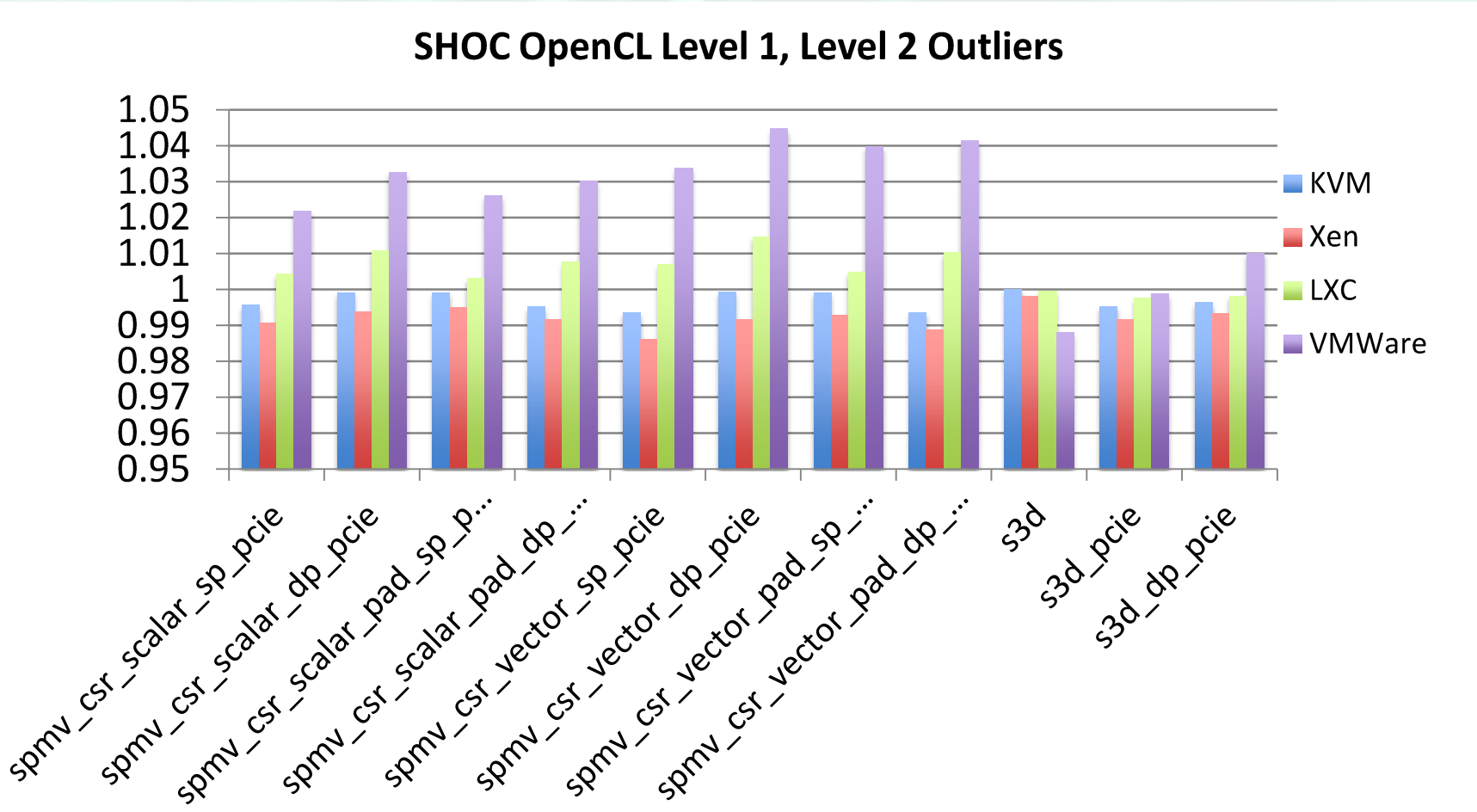
Moving Forward - Vertically (post PhD)

- Cloud Infrastructure now provides new hardware
- Nee for Platform services (PaaS) to leverage new advances
  - Forget about TCP/IP?!
  - Enable InfiniBand usage with MapReduce paradigms
  - Transparent RDMA for “ABDS” solutions
- Evaluate data intensive scientific applications
- Evaluate existing problems with new platforms

C2075 Results – SHOC Outliers

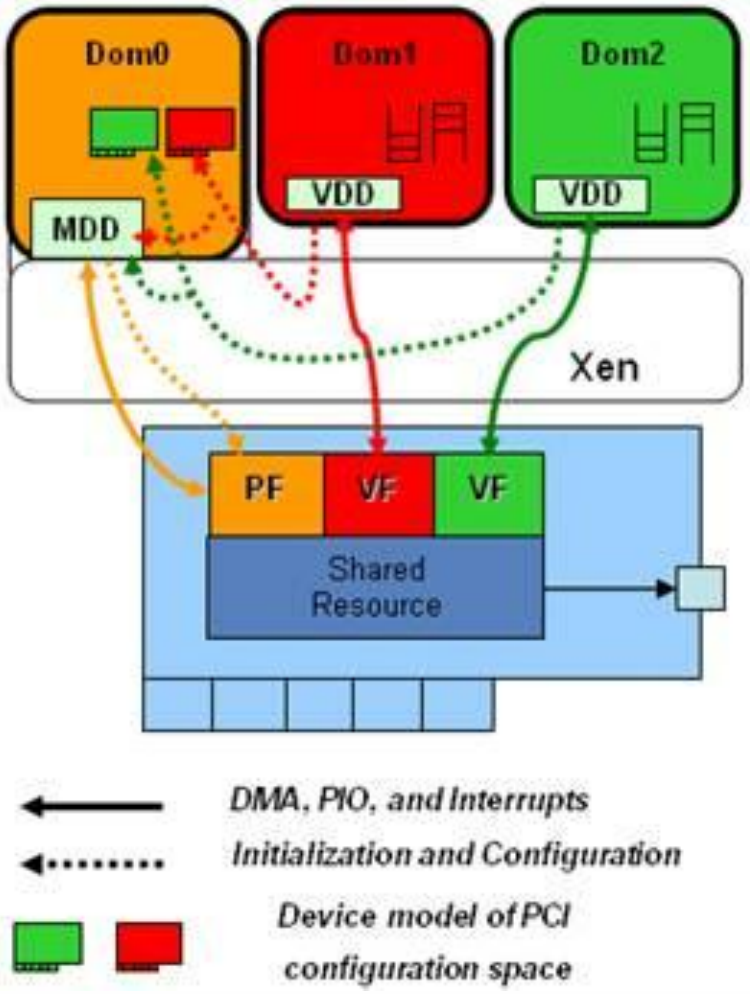


K20 Results – SHOC Outliers



SR-IOV VM Support

- Can use SR-IOV for 10GbE and InfiniBand
  - Reduce host CPU utilization
  - Maximize Bandwidth
  - “Near native” performance
- Requires extensive device driver support
  - Mellanox now supports KVM SR-IOV for CX2 and CX3 cards



From “SR-IOV Networking in Xen: Architecture, Design and Implementation”

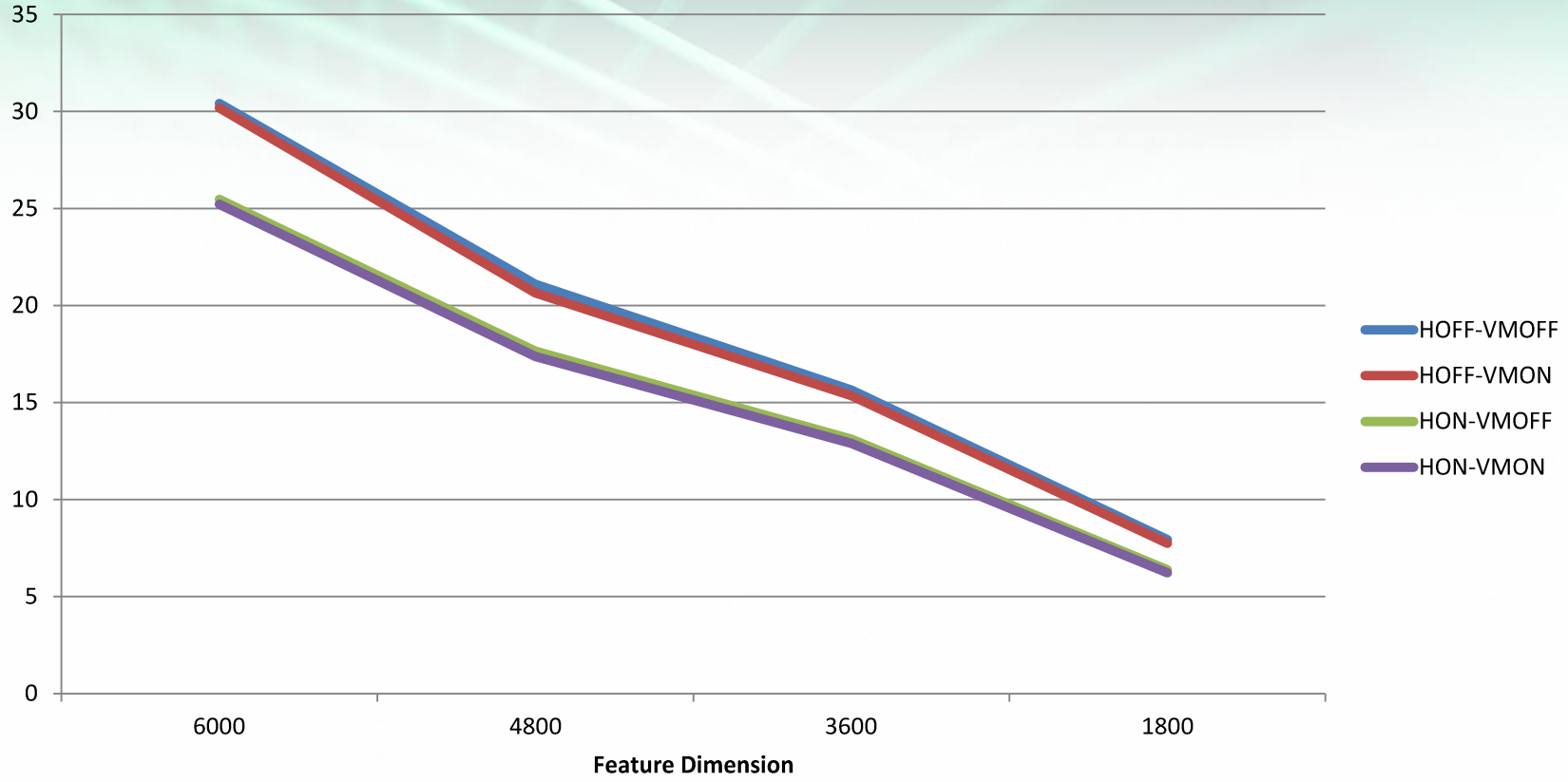
GPU comparison

- In 2012, the Xen GPU Passthrough implementation was first of its kind for Nvidia Tesla GPUs
- Recently, more hypervisors added support
- Developed similar methods in KVM (new)
  - Userspace driver interface
  - Based on kvm/qemu VFIO in new kernel >= 3.9
- Can now make apples-to-apples comparison

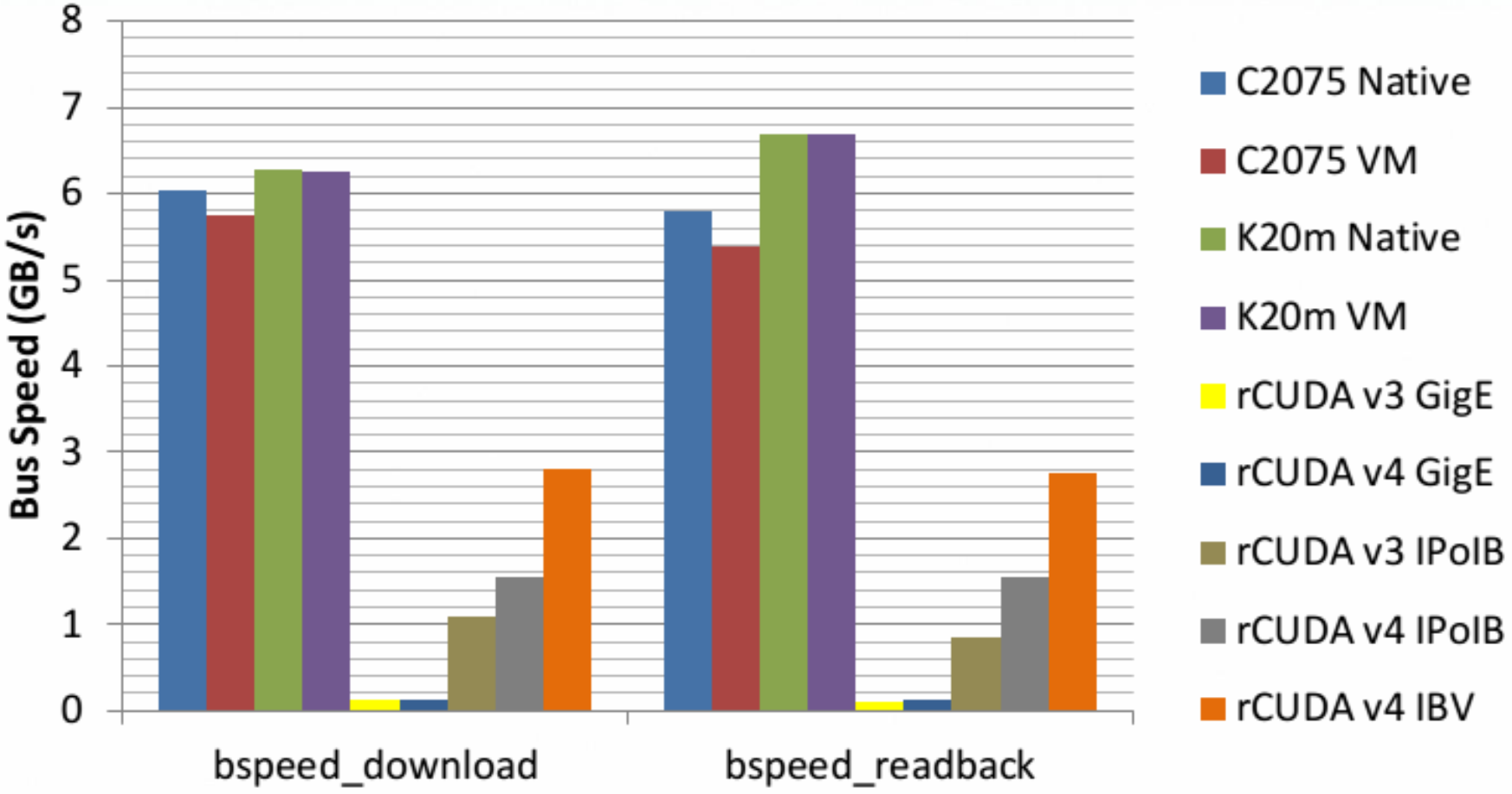
THP EPT and TLB

- THP – Transparent Huge Pages
  - Allocate memory blocks in 2MB and 1GB sizes
  - Easily allocation in userspace
- EPT – second level address translation
  - Intel technique to avoid multiple address lookups
  - Treats guest addresses as host-virtual addresses (in hardware)
- TLB – cache for virtual memory pages
  - Want to minimize TLB misses whenever possible
  - Each TLB miss requires “hypercall”

LibSVM - KVM Transparent Huge Pages (Lower is Better)

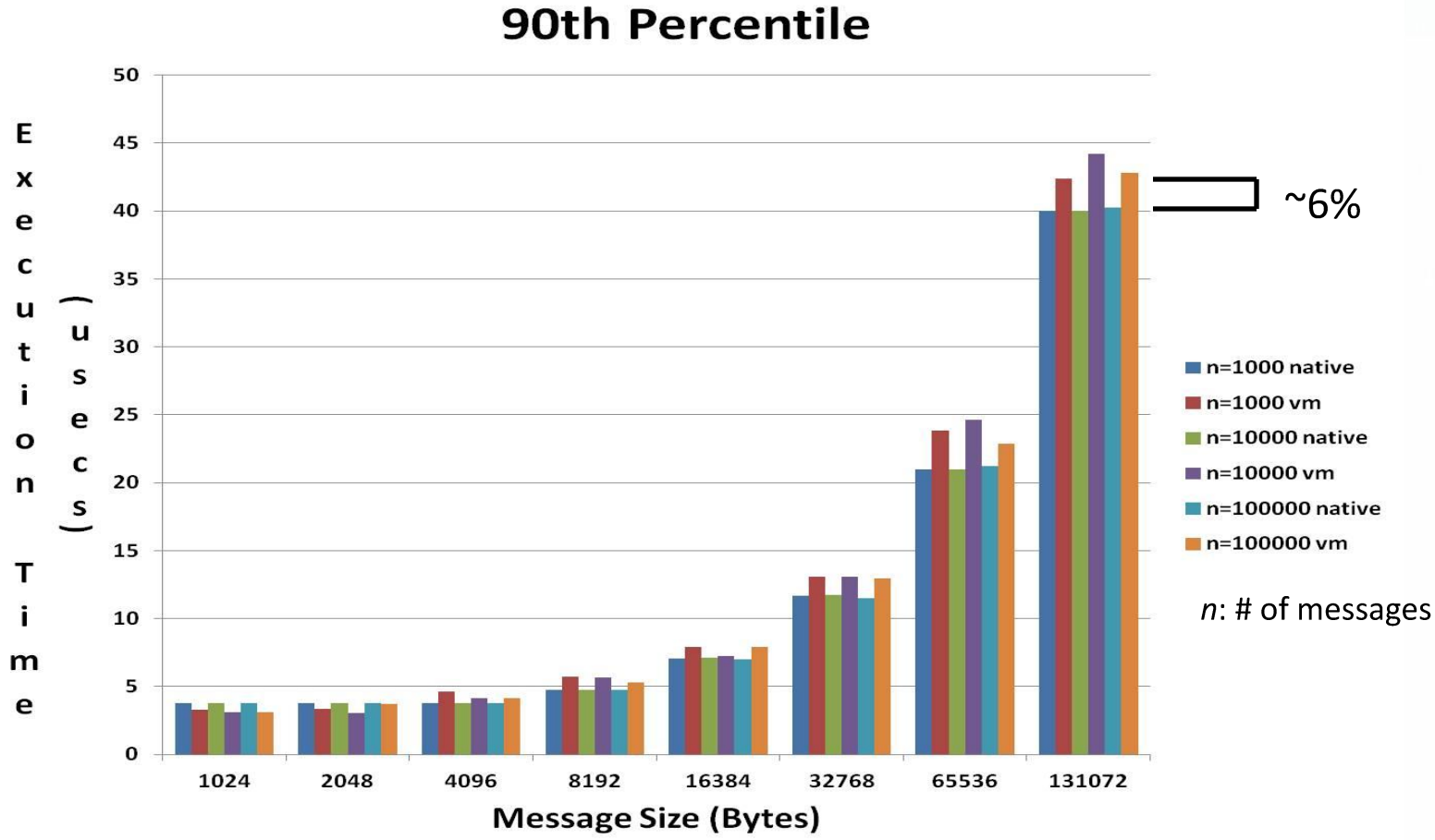


GPU Bus Speed



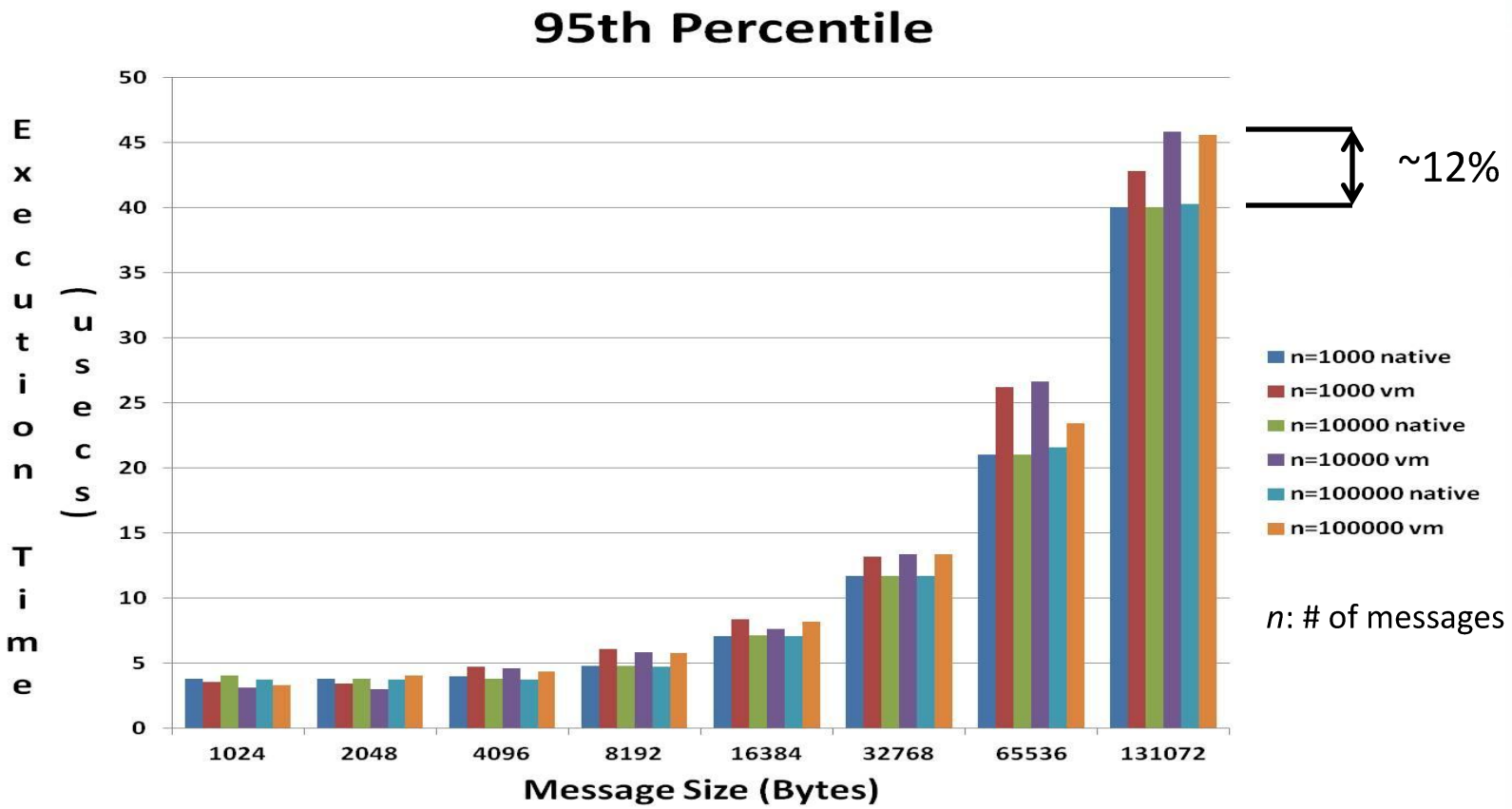
Future Grid

Results: Latency Distribution ib_write-lat *(Network-Level)*



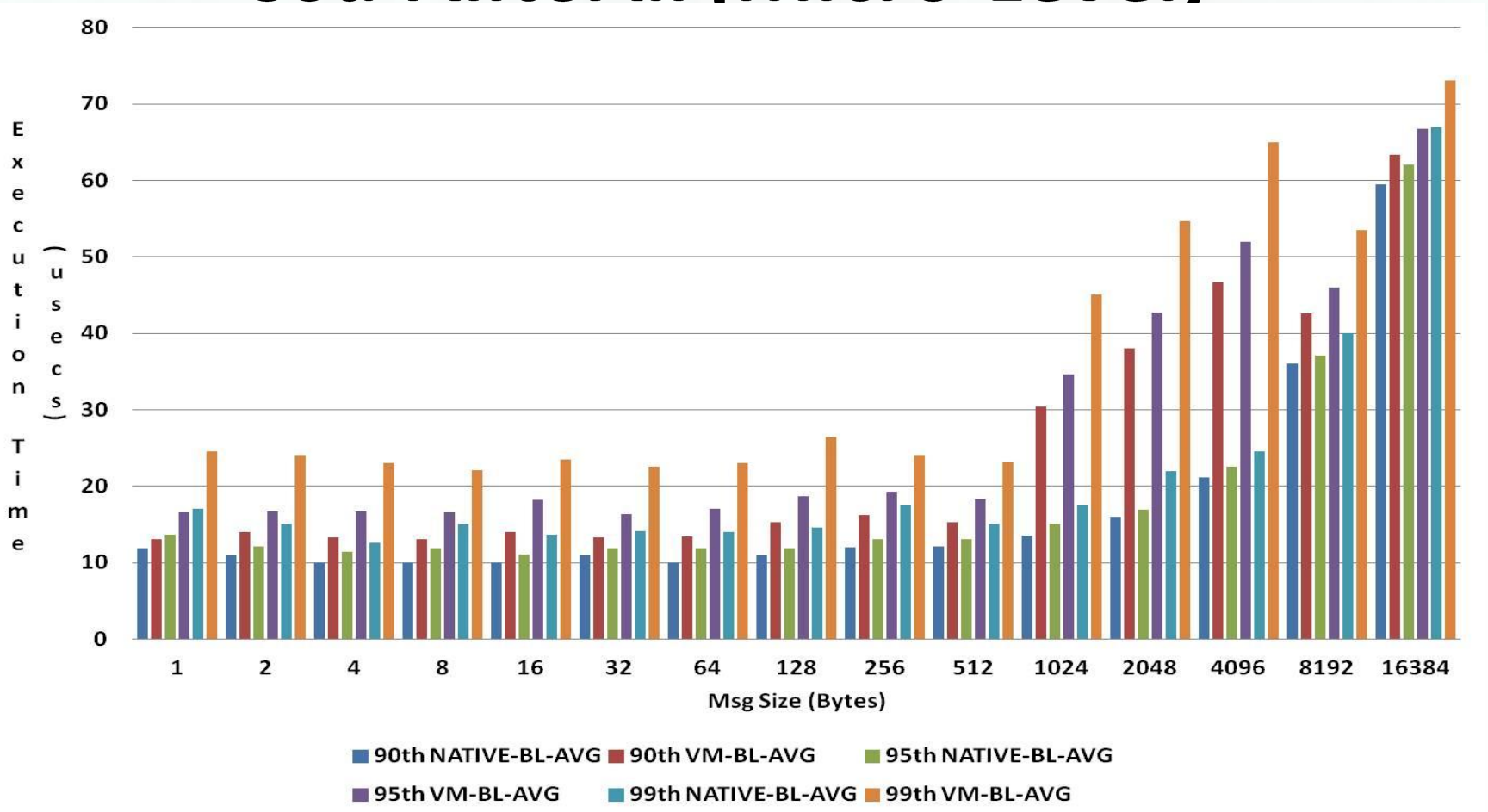
 VM perf. Competitive with native

Results: Latency Distribution ib_write-lat *(Network-Level)*



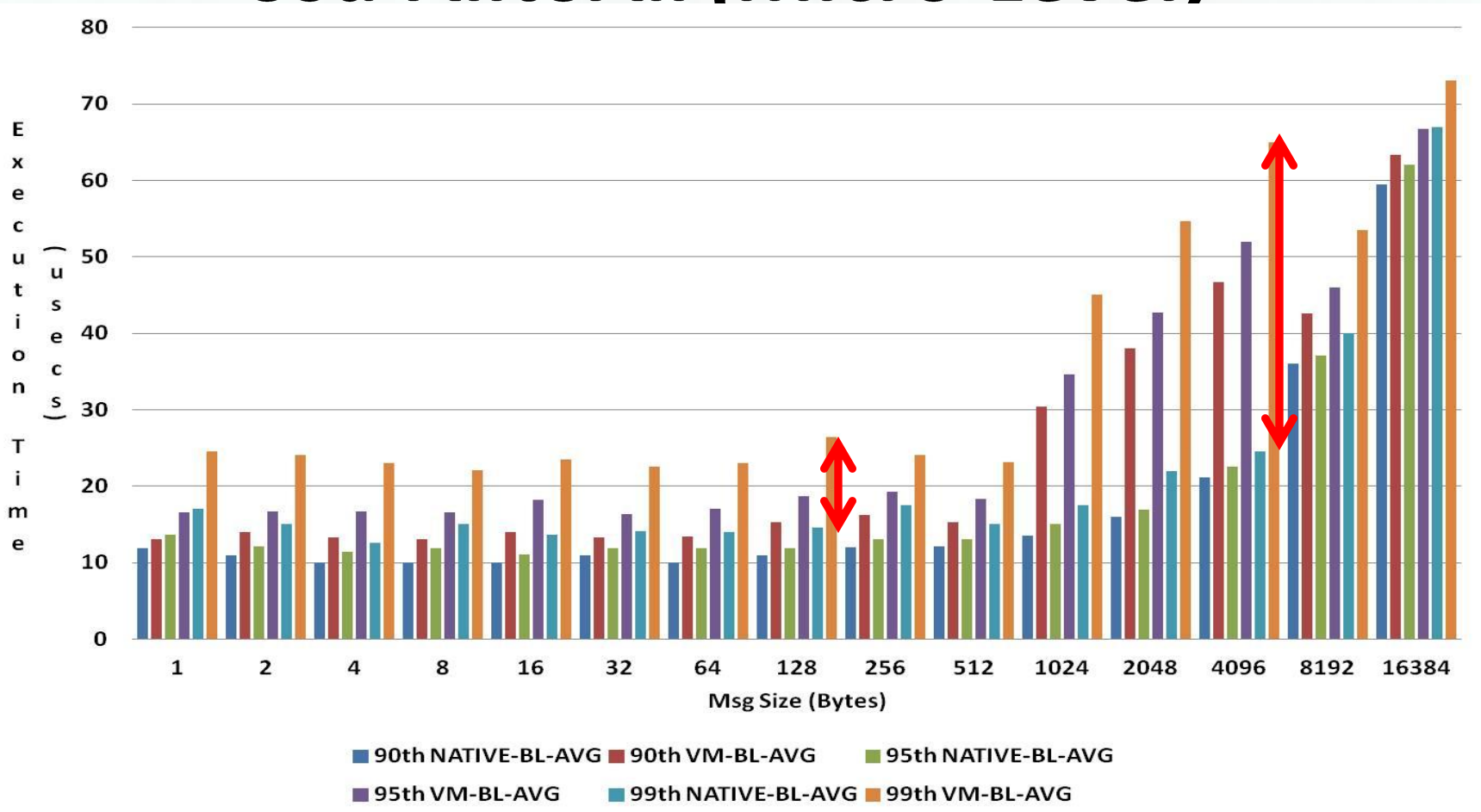
 Majority of overhead in tail-latency

Results: Latency Distribution osu-AlltoAll *(Micro-Level)*



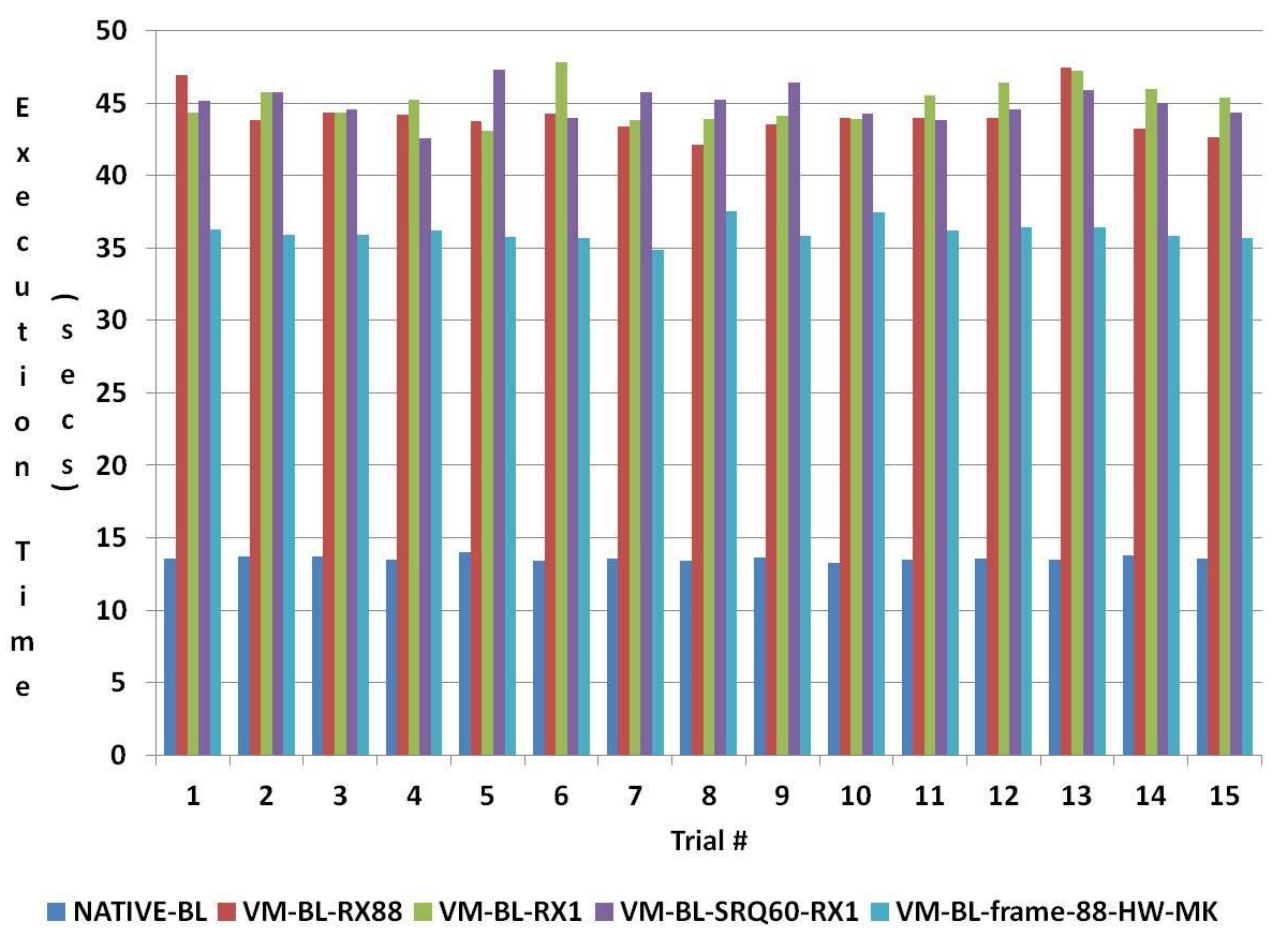
- Majority of overhead in tail-latency
- Overhead decreases with larger msg size

Results: Latency Distribution osu-AlltoAll *(Micro-Level)*

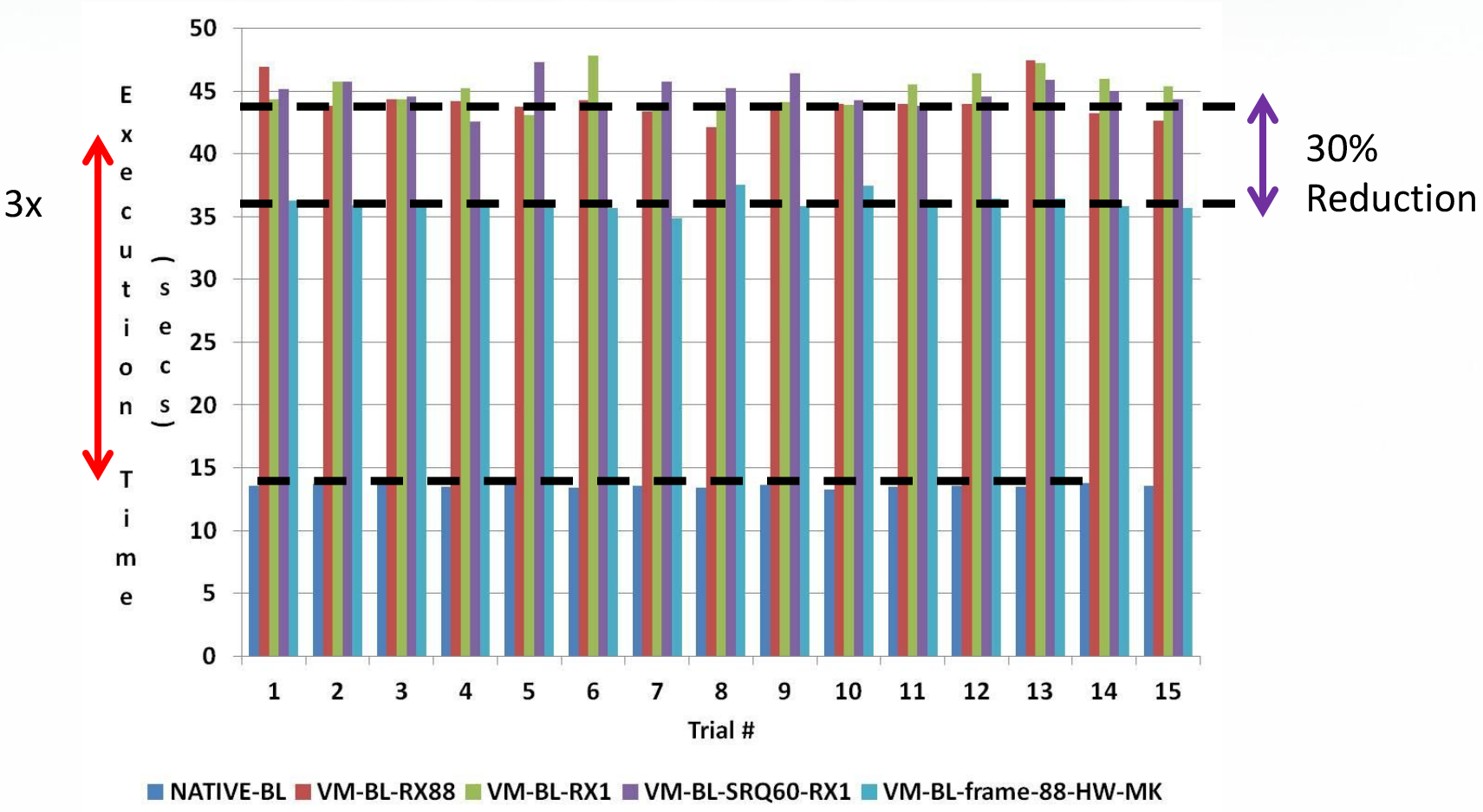


- Majority of overhead in tail-latency
- Overhead decreases with larger msg size

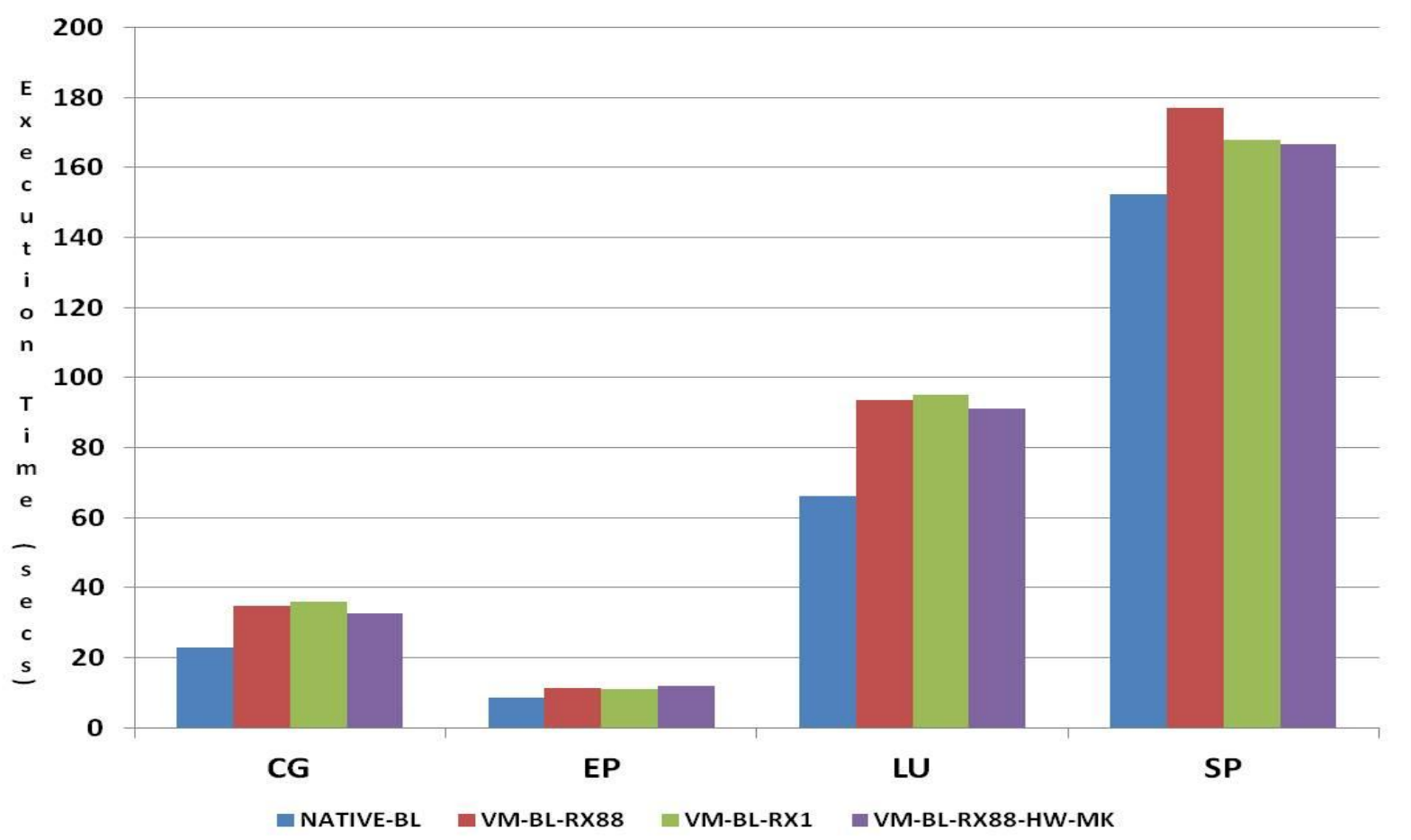
Results: Network Tuning NPB-CG *(Macro-Level)*

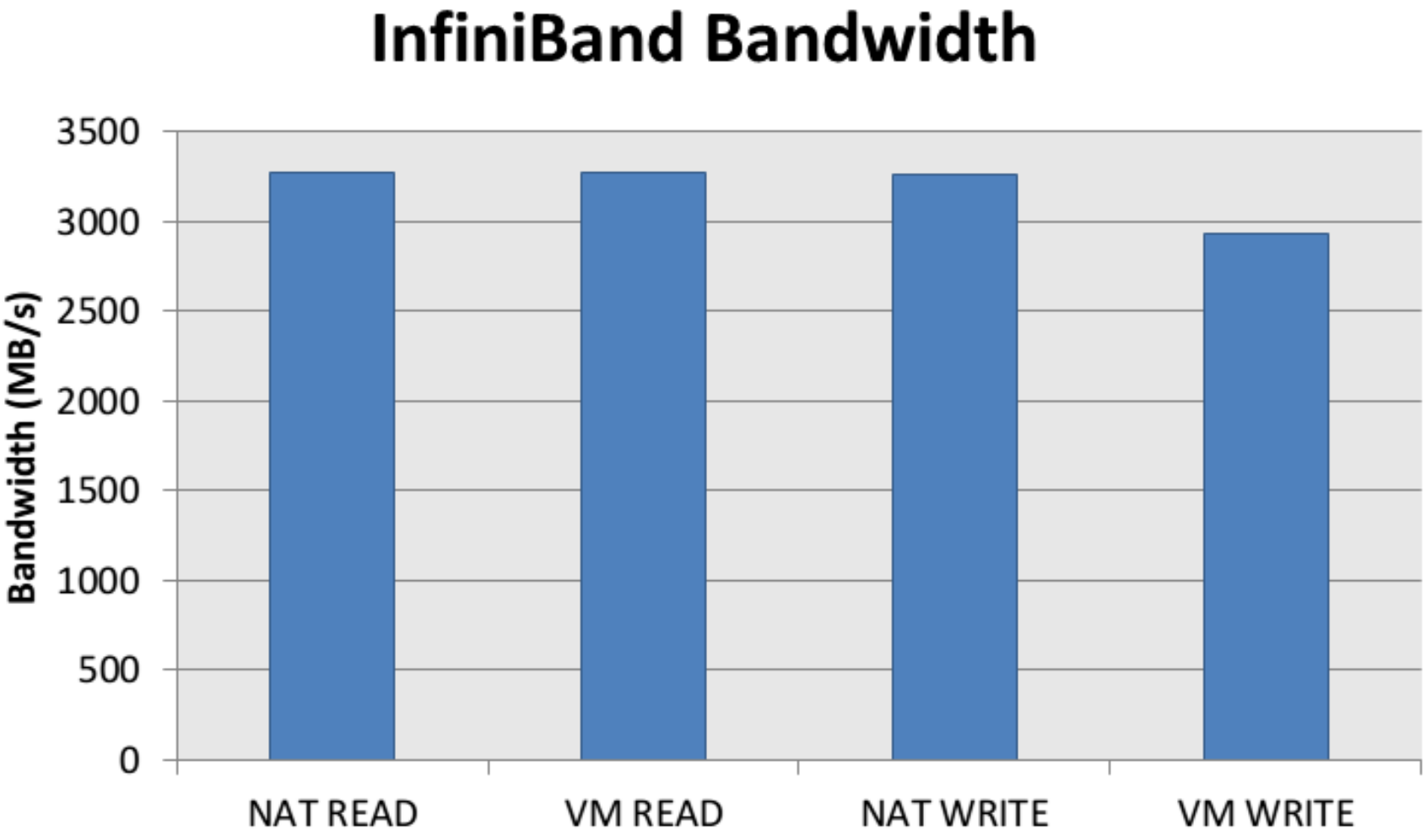


Results: Network Tuning NPB-CG *(Macro-Level)*



Results: Network Tuning
NPB *(Macro-Level)*




InfiniBand Bandwidth
Bandwidth (MB/s)
3500
3000
2500
2000
1500
1000
500
0
NAT READ
VM READ
NAT WRITE
VM WRITE