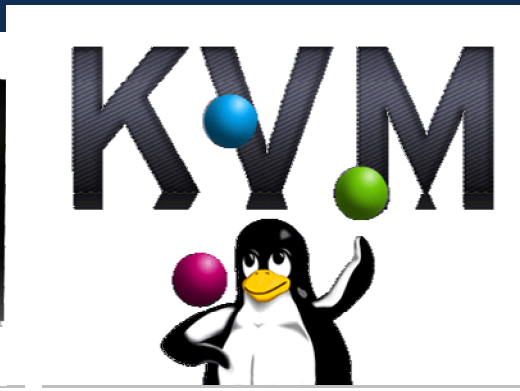
SAND2017-10416C



Enabling Diverse Software Stacks on Supercomputers using High Performance Virtual Clusters

Andrew J. Younge, Kevin Pedretti, Ryan E. Grant,

Brian L. Gaines, Ron Brightwell

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Outline

Sandia National Laboratories

- Motivation
- Background in Virtualization & HPC
- Design of High Performance Virtual Clusters
- Experimental Evaluation
  - HPCC Benchmark Suite
  - HPCG scaling
- Conclusion

Motivation

Sandia National Laboratories

- At the forefront of convergence between High Performance Computing (HPC) and Large-scale Data Analytics (LSDA)
  - Each are based on distributed systems and many-node architectures
  - Each have drastically different system software ecosystems
- Research in LSDA platform services have made advancements towards utilization of HPC resources and clusters
  - However, need crosscutting measure to help bridge the gap
- Introduction of High Performance Virtual Clusters (HPVCs)
  - Can build user-defined software ecosystems on supercomputers
  - Can provide method for in-situ simulation and analytics
  - Can enable LSDA to leverage advanced HW and networks found in HPC

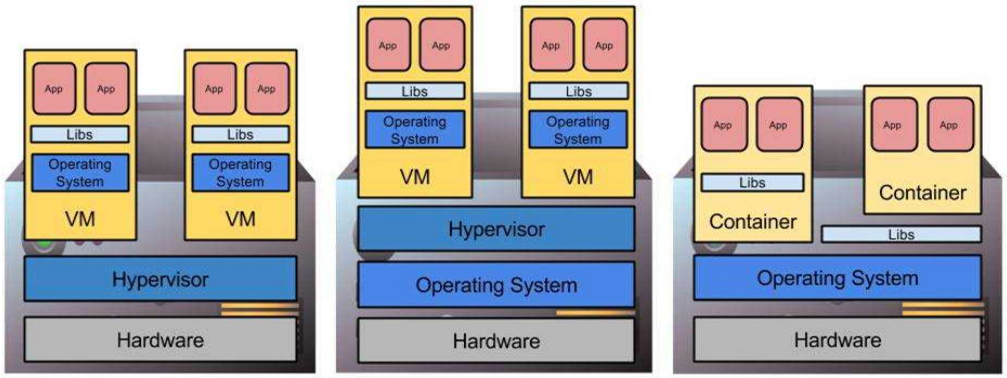
Virtual Clusters

Sandia National Laboratories

- Clusters have been a key aspect of distributed systems.
  - Can build virtual clusters using collections of Virtual Machines (VMs)
- Virtual Clusters provide:
  - Dynamic resource allocation of VMs
  - Clusters built atop VMs
  - OS, software, and apps provisioned by users
  - Leverage both node-local and cluster-wide tools for VM management
    - Cloud-based such as OpenStack IaaS
    - Cluster-based such as batch queueing system
  - User experience is cluster based, not individual VM based
  - Resource consolidation possible through oversubscription
  - Dynamic scaling of VC through add/delete of VMs
  - Back-end virtual storage through persistent storage
- Many Examples: COD, Grid5000, Chameleon, EC2

Hypervisors and Containers

- Type 1 hypervisors insert layer below host OS
- Type 2 hypervisors work as or within the host OS
- Containers do not abstract hardware, instead provide "enhanced chroot" to create illusion of VM
- Where abstraction is inserted can have impact on performance
- All enable custom software stacks on existing hardware



Type 1 Hypervisor Type 2 Hypervisor Containers

KVM

Sandia National Laboratories

- KVM was chosen for providing custom software ecosystems on supercomputers
- Type 2 hypervisors provide good way to augment existing vendor software stacks
  - Simple module within existing Linux kernel
  - Minimal disruption of existing Vendor SW stack
  - Demonstrated by Palacios VMM
- Does not preclude using container solutions like Docker for image management & provisioning
  - Provides additional isolation & security
- Could provide advanced VM features like live migration & cloning in future

Related Research

Sandia National Laboratories

- Many efforts have focused on improving performance of VMs or providing user flexibility in HPC
- Few projects focus specifically VMs for HPC
  - Palacios VMM & Hobbes project
  - FutureGrid & Chameleon NSF projects for grid/cloud/hpc
- Adaption of VMs to utilize advanced HW – GPUs, InfiniBand
- Containerization efforts for SW flexibility in HPC
  - Shifter, Singularity, CharlieCloud, etc
  - Provide user-defined software stacks on compute nodes
  - Similar to Docker but without root

Design of a High Performance Virtual Cluster

Sandia National Laboratories

- Commodity clusters different than advanced Supercomputing MPP resources – see top 10 of Top 500 list
- Providing virtualization on a supercomputer like a Cray XC-series system is not the same as a small cluster
  - NOT Ethernet and InfiniBand interconnect
  - Have specialized node OS: Kitten, Catamount, modified Linux, etc.
  - Separated I/O and storage nodes
  - Specially tuned MPI and other HPC system libraries
- If HPVCs can be realized, users can benefit from both extreme scale and performance as well as increased usability

Volta – XC30 Testbed

Sandia National Laboratories

- Volta is part of the Advanced Systems Technology Testbeds project at Sandia through the NNSA's ASC program
- Cray XC30 rack design
  - Dual-socket "Ivy Bridge" E5-2695v2 processors
  - 2x12 cores cores total
  - 64GB DDR RAM
  - Cray Aries Interconnect
- Cray Compute Node Linux (CNL)
  - Cray's tuned Linux OS (ver. 5.2.UP04)
  - 3.0.101 Linux kernel
- Volta a testbed similar to ACES Trinity supercomputer
  - #10 supercomputer on the Top500

Guest Performance

- Running VMs must be performant, so tuning is key
- Ways to significantly reduce overhead in KVM
- Tuning with Libvirt is possible, but has to be a holistic approach
  - CPUs & cores
  - Memory
  - NUMA
  - Network
  - Disk
  - Time

```
<memoryBacking>
 <hugepages>
  <page size="2" unit="M" nodeset="0"/>
  <page size="2" unit="M" nodeset="1"/>
 </hugepages>
 <nosharepages/>
</memoryBacking>
<cpu match='exact'>
 <model>IvyBridge</model>
 <topology sockets='2' cores='12' threads='1'/>
 <vendor>Intel</vendor>
 <numa>
  <cell id='0' cpus='0-11' memory='30' unit='GiB'/>
  <cell id='1' cpus='12-23' memory='30' unit='GiB'/>
 </numa>
</cpu>
<numatune>
 <memory mode='strict' nodeset='0-1'/>
 <memnode cellid="0" mode="strict" nodeset="0"/>
 <memnode cellid="1" mode="strict" nodeset="1"/>
</numatune>
<vcpu>24</vcpu>
<cputune>
 <vcpupin vcpu='0' cpuset='0'/>
 <vcpupin vcpu='1' cpuset='1'/>
  ...
 <vcpupin vcpu='23' cpuset='23'/>
</cputune>
```

Guest CPU & NUMA Configuration

Sandia National Laboratories

- Provide same processor architecture in guest as host
  - Includes IvyBridge AVX and other instruction sets
  - Specify CPU topology – cores and treads
- Assign and pin all available cores to guest VM with <cputune>
- Specify NUMA configuration explicitly
  - Cores utilize local memory & cache
  - Avoid QPI saturation across sockets
- Timing is important!
  - Use kvm-clock for host timing via MSR
  - Ensure constant_tsc for invariant clock

Hugepages

Sandia National Laboratories

- Back entire Guest VM with Hugepages for performance boost
  - Use 2MB pages in guest and host, not 4KB default
- Why?
  - Nested page tables often more efficient than shadow paging
  - Page miss cost decreases by 9 reads for 2MB pages
  - TLB miss count decreases due to increased TLB reach
  - Result is memory intensive codes avoid walking page table
- Use Libvirt's <hugepage> and <memorybacking>
- Cray doesn't support transparent hugepages (THP) on host
  - Have to manually use Cray hugepages
  - HUGETLB_DEFAULT_PAGE_SIZE=2M env variable
  - Pre-allocate hugepages is important
- Enable THP or Libhugetlbfs within guest OS

Interconnect Utilization

Sandia National Laboratories

- Cray's interconnect is a differentiating factor from clusters
- Exposing the true potential of the underlying hardware to the guest is challenging
  - Aries was never meant for virtualization
  - Proprietary drivers rule out PCI Passthrough
  - Copying vendor libraries to guest not portable
  - SR-IOV not available for Aries
- Need to leverage Ethernet-over-Aries emulated NIC
  - Bridge Aries Ethernet device like normal Eth
  - Match last 3 octets of MAC address
  - Add static ARP entries to all OSes
  - Set MTU size to 65520 for best bandwidth
- Ethernet-over-Aries interconnect creates ~40% overhead
  - Still better than state-of-art 10Gb Ethernet or perhaps even InfiniBand

Experimental Evaluation

Sandia National Laboratories

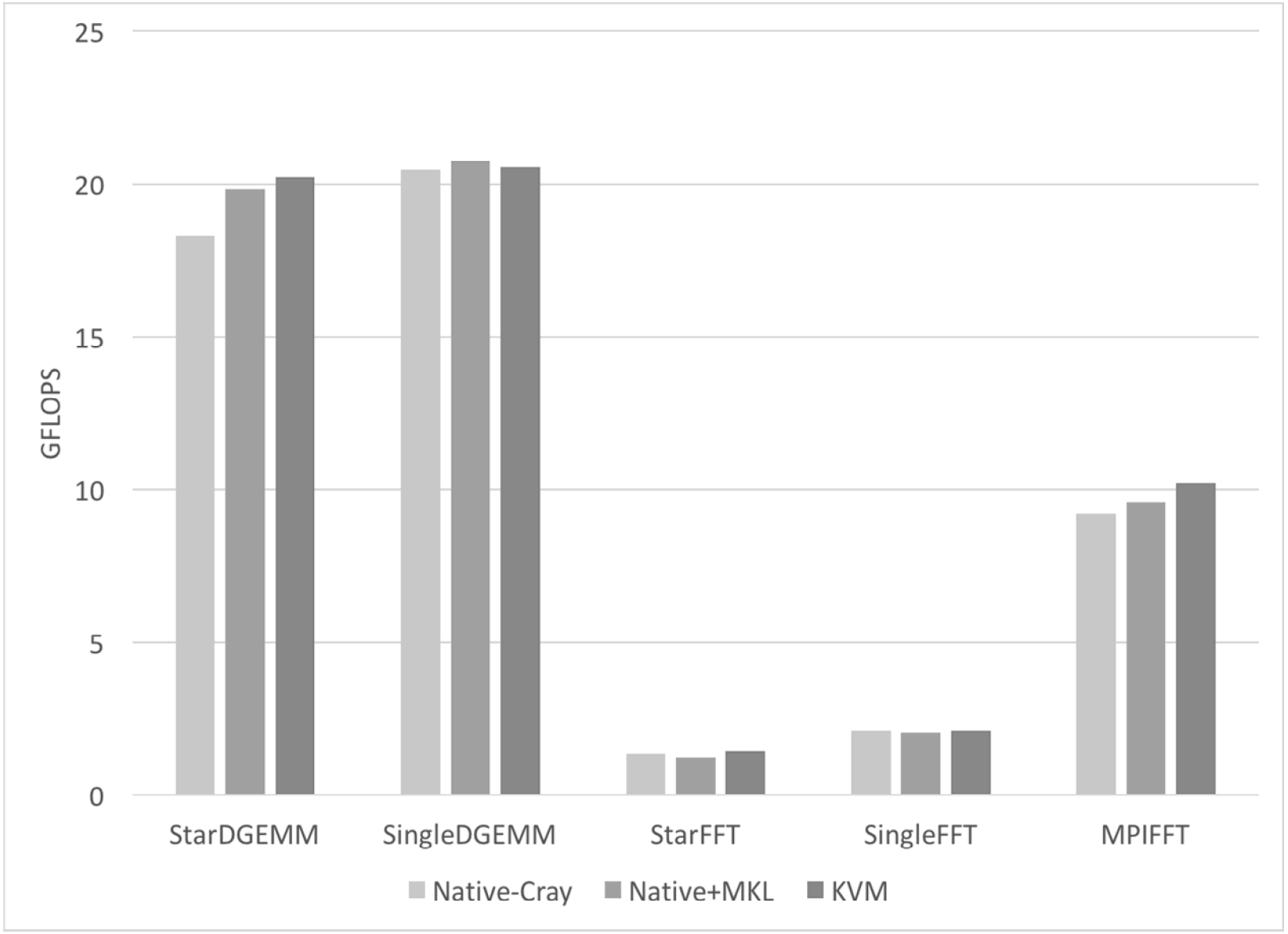
- Ran experiments on 32 nodes of Volta testbed, 24 ppn
- Utilize existing HPC tools for initial evaluation
  - HPCC Benchmark Suite
    - HPL, DGEMM, FFT, STREAM, RA, PingPong, etc
  - High Performance Conjugate Gradient (HPCG)
    - Provides “bookend” to HPL, real-world performance
    - More related to mission apps of interest at Sandia
  - Evaluate both intra-node performance as well as weak scaling
- Evaluation of Apache Spark
  - Significant big data analytics tool for LSDA
  - Not usable on standard Cray XC30 (at time of writing)

Test Suite Configuration

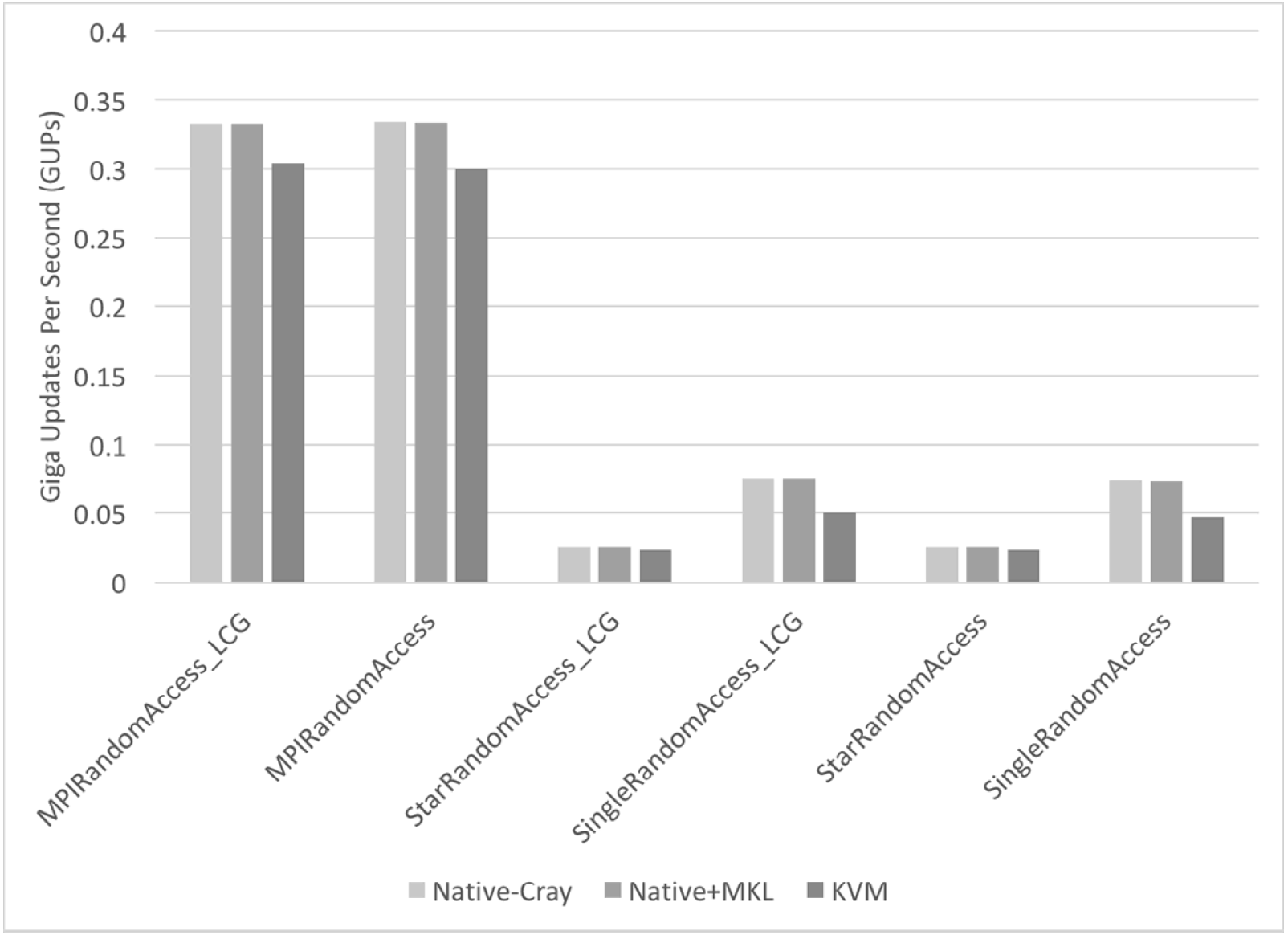
Sandia National Laboratories

- Cray native software configuration – CLE 5.2
  - Cray-Intel programming environment
  - Intel C/C++ compiler 16.0.1
  - Cray MPI library, UGNI protocol, etc
  - LibSci math libs, as well as Intel MKL (explained later)
  - Cray 2MB Hugepages
- HPVC software configuration
  - Standard RHEL 7 guests with stock 3.11 Linux kernel
  - HPC VC
    - MPICH 3.2 MPI library
    - Intel 2017 parallel studio cluster suite
  - Apache Spark VC
    - Java SDK 8 & Maven for Apache Spark v2.1.0

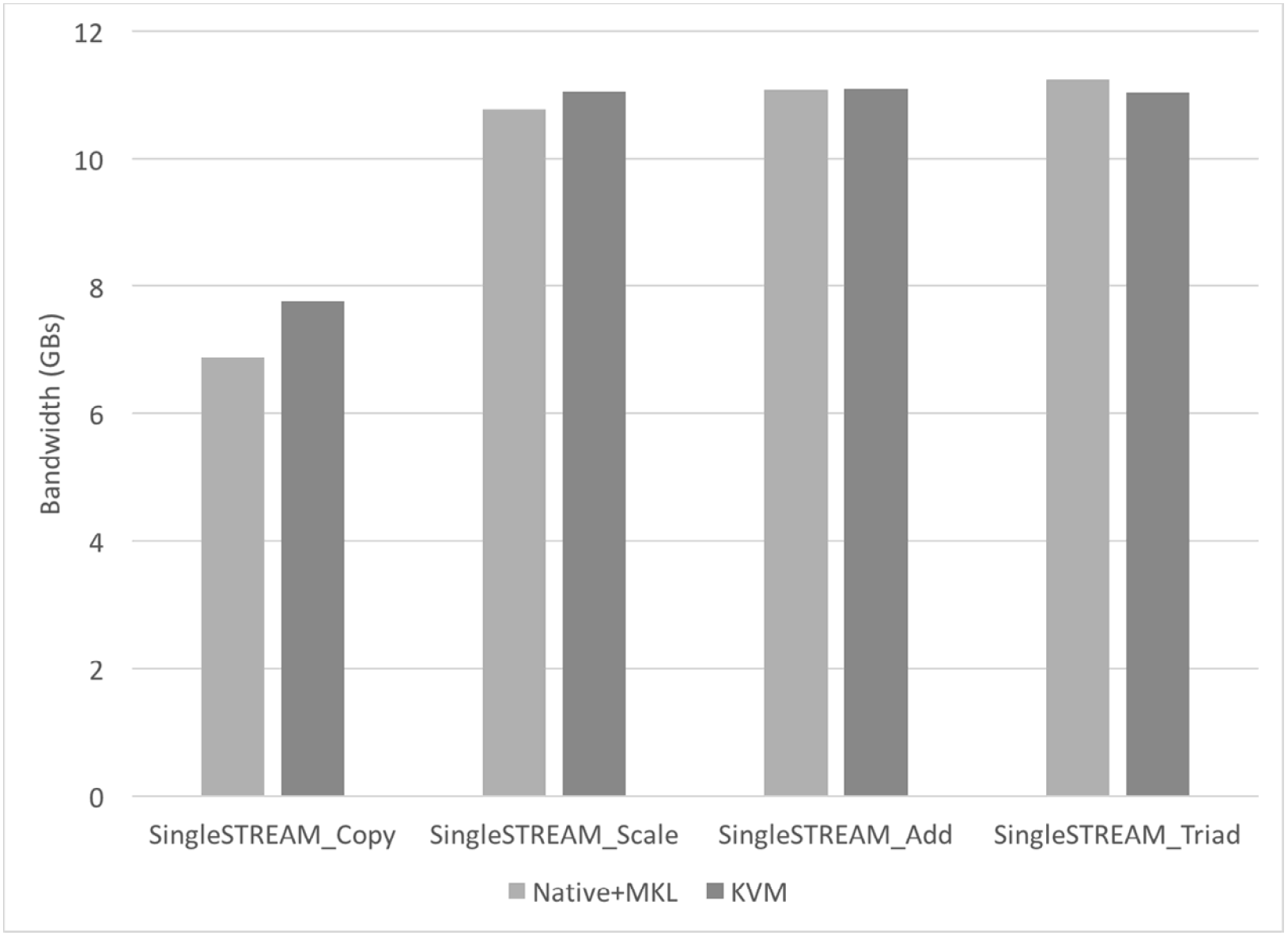
Intra-Node FLOPS



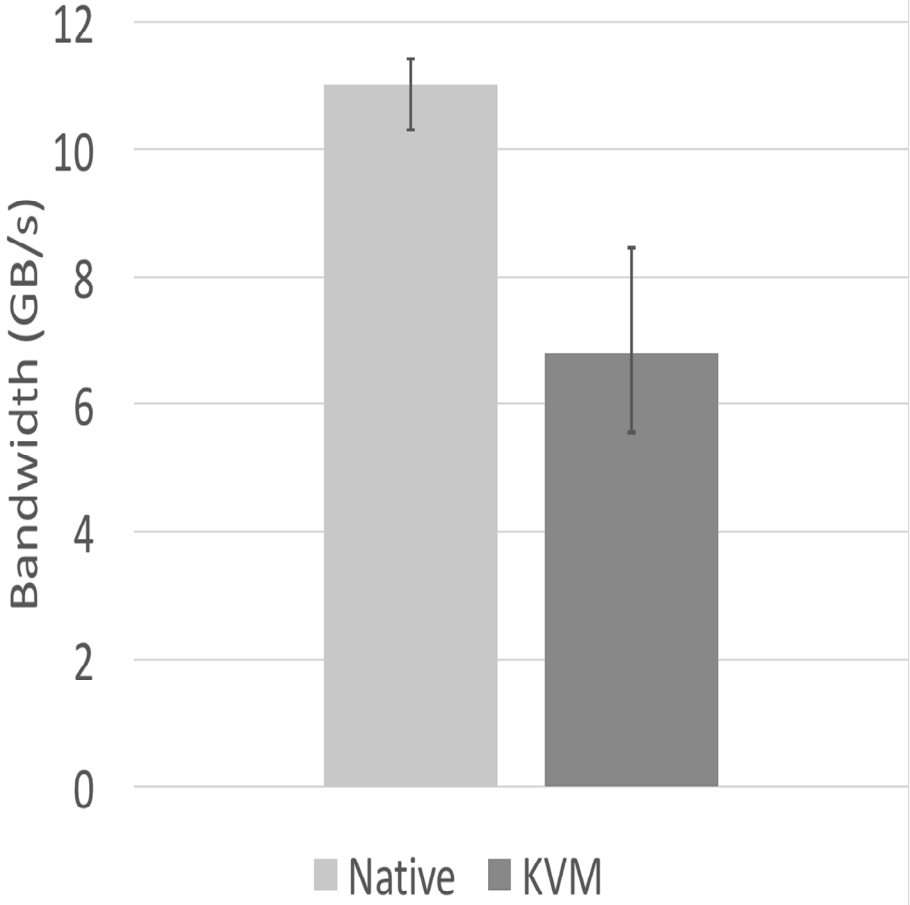
Intra-Node GUPS

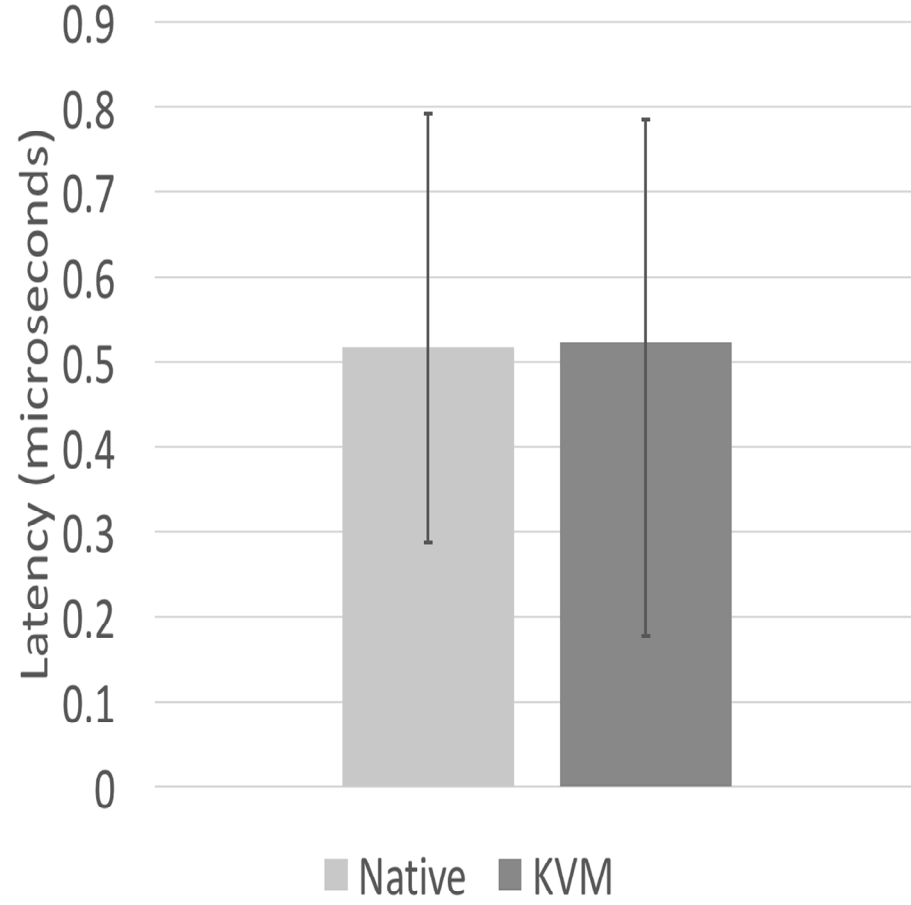


Intra-Node STREAM

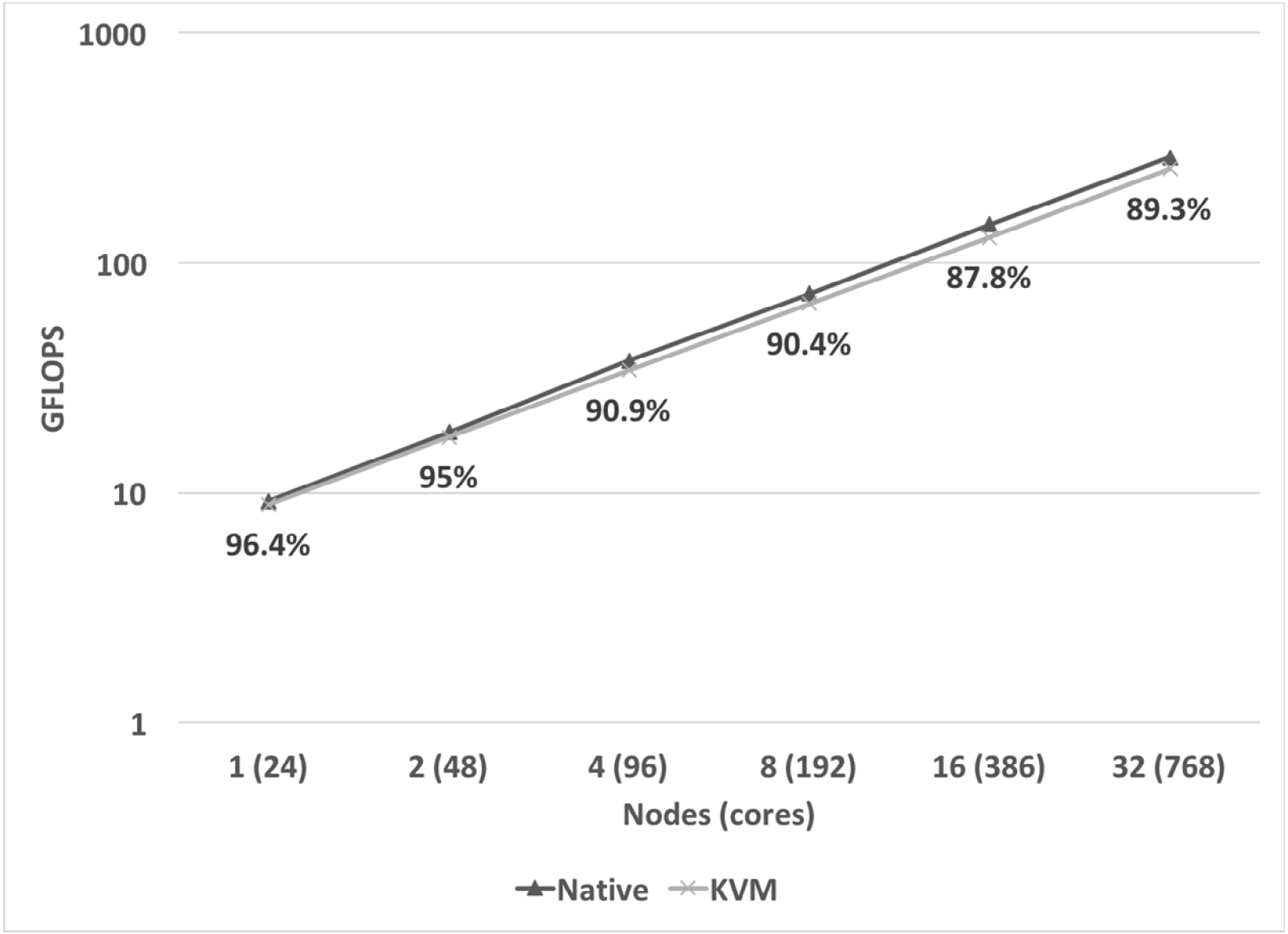


Intra-node MPI Performance





Weak Scaling HPCG



Take-aways

Sandia National Laboratories

- Single-node performance looks to be “near-native”
  - Compiler & library selection more important than virtualization overhead
    - MKL/Libsci selection accounts for FLOPS differences
    - Intra-node PP BW limited by not using XPMEM
  - Hugepages relives TLB pressure and helps memory-intensive codes
- Multi-node scaling limited by interconnect
  - Emulated Ethernet-over-Aries not capable of native performance
  - Still able to achieve ~89% performance @ 768 cores
  - Likely better than Ethernet solutions deployed today

LSDA – Apache Spark

Sandia National Laboratories

- Built Apache Spark virtual cluster on Volta Cray XC
  - 32 VMs configured as slave, 1 VM master & 1 NFS server
- Used TeraSort benchmark with 10 and 100 GB sizes
  - All-to-all shuffle between the Map and Reduce phases
  - Stresses data movement significance
- Terasort: 3m9s for 10 GB size, 8 hrs for 100GB size
    - Stressing limits of NFS server – no node-local disk on Cray
  - Leads to future work w/ Burst Buffers
  - Expect Lustre performance to be much better

Spark-Perf Results

Sandia National Laboratories

- Spark_perf benchmark suite - tests common operations in MapReduce platforms
- Runtime doubles for 10x problem size for smaller scales
- Aggregate-by-key benchmarks see roughly a 4-5x increase
- Aggregate-by-key with at 10x scale over-saturates cluster
  - Aggr 4b records with 10m unique values for 400t unique integers

Scale	Throughput	Aggr-by-key	Aggr-by-key-int	Aggr-by-key-naive	Sort-by-key	Sort-by-key-int	Count	Count-filter
0.001	2.6585	0.106	0.1085	0.199	0.114	0.1125	0.034	0.0575
0.01	2.6285	0.219	0.1905	0.4135	0.3065	0.3765	0.0395	0.0935
0.1	2.683	0.474	0.437	0.9605	0.839	0.7075	0.056	0.1495
1	2.6975	2.24	1.886	5.19	2.976	1.797	0.162	0.2665
10	2.642	15.429	47.629	32.9335	5.378	3.9455	1.1095	1.1935

Conclusion

Sandia National Laboratories

- Created first High Performance Virtual Cluster on a Cray supercomputer
- Demonstrated both HPC and LSDA workloads on virtualized infrastructure
  - HPC benchmarks utilized to evaluate & tune performance
  - Apache Spark workloads run on Cray
- Single-node approaches near-native performance
- Scale limited by Ethernet-over-Aries network emulation
  - SR-IOV or new NIC design would help in the future
  - Still on-par or better(!) than commodity Ethernet
- Vendor-specific libraries would also help integration
- Success on XC30 testbed demonstrates feasibility to support VMs on Trinity supercomputer

Thanks!

Questions?

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.