SAND2010-8099P

A Scalable Virtualization Environment for HPC

ASC Booth Talk, SC10

November 16, 2010

Kevin Pedretti
Sandia National Laboratories
Albuquerque, NM
ktpedre@sandia.gov

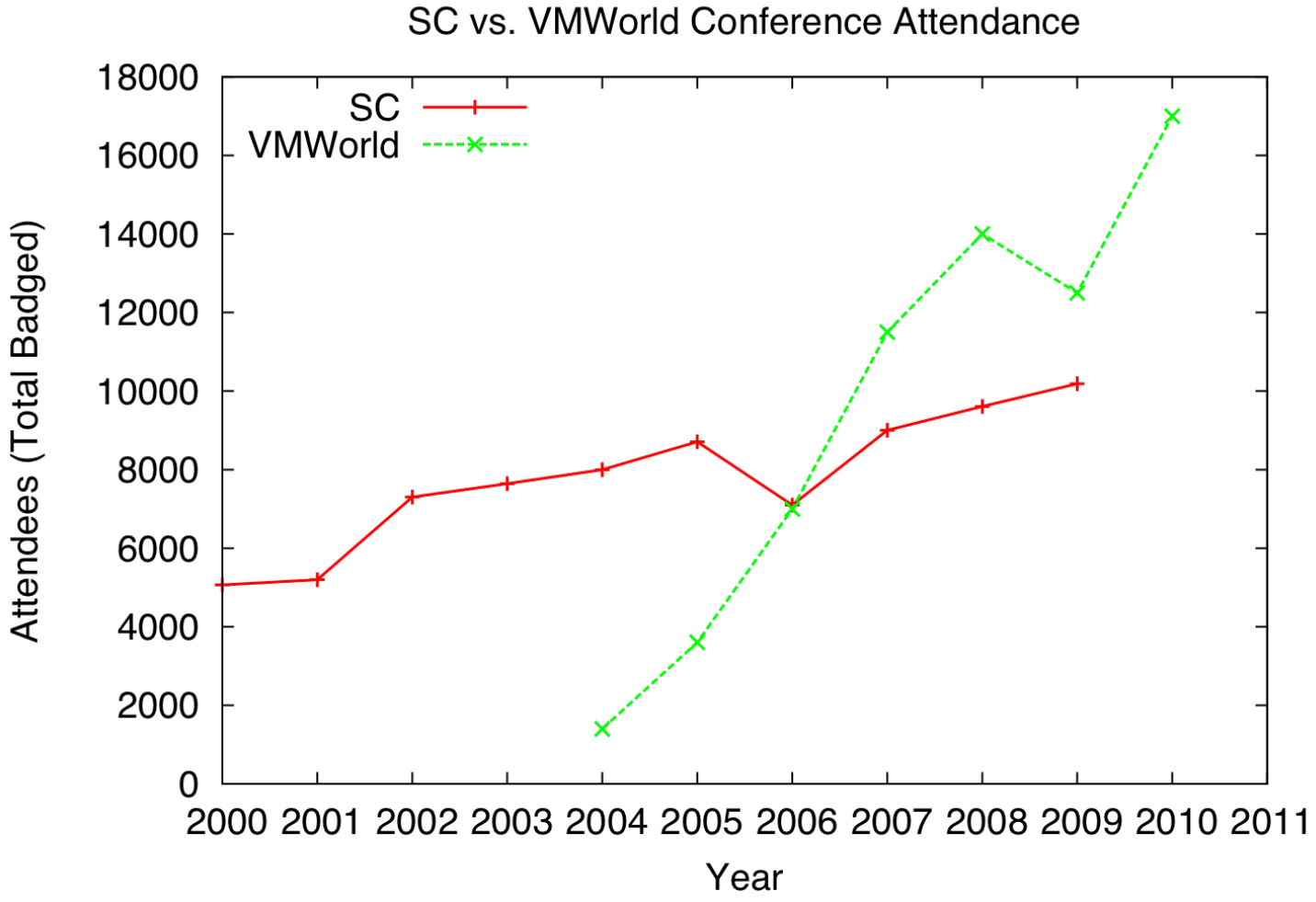
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Background – OS Virtualization

- Treat OS as an application
- Major trend in enterprise data center / IT industry over last several years
- Motivations
  - Server consolidation
  - Dynamic workload balancing
  - Enhanced security isolation
  - On-demand compute capacity, Amazon EC2 “elastic cloud”
- Powerful tool for developers, desktop power users
  - Run Windows on Linux, run Cplant on laptop, etc.

Sandia National Laboratories

Virtualization Seeing Explosive Growth in General Computing Market



Sources: SC web sites, news articles, and blogs

Sandia National Laboratories

HW-accelerated Virtualization Will Be Baked In

- Any commercially viable platform will have a virtualization story; increasingly sophisticated support
  - x86, AMD, Intel, …
  - ARM
  - PowerPC
  - Self-virtualizing devices (NICs, GPUs, …)
- Public clouds beginning to target low/mid HPC
  - Amazon's EC2 Cluster Compute Instances

Can high-end HPC also leverage virtualization? Does it enable new capabilities?

Sandia National Laboratories

Key Questions

- What are the use cases for high-end HPC?
- What are the virtualization overheads?
  - Compute
  - Virtual Memory
  - I/O
- What can be done to mitigate the overheads?

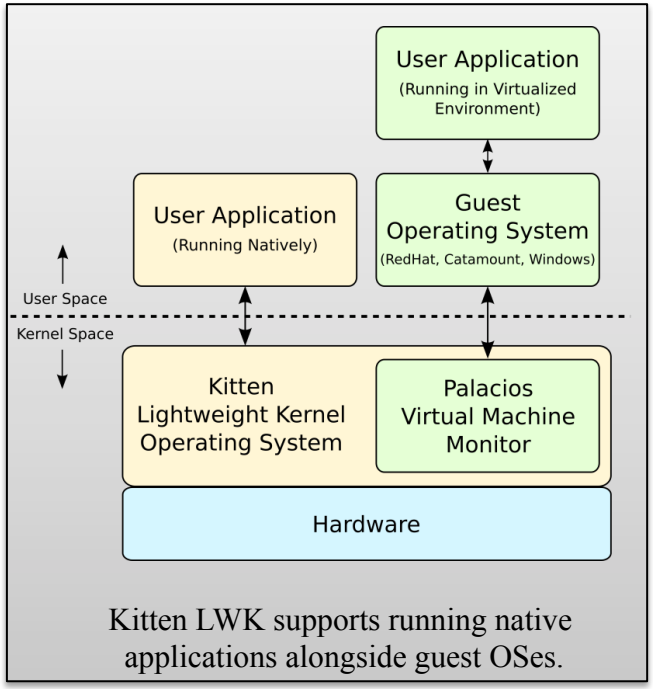
Sandia National Laboratories

Virtualization Use Cases

Sandia National Laboratories

Use Case 1: Augment lightweight kernel with VMM to increase flexibility

- Original motivation
- LWK provides high perf. native environment
- VMM allows full-featured guest OS (e.g., Red Hat Linux) to be loaded on-demand
  - Perl, python, matlab, …
  - COTS databases, simulators, …
  - You name it
- Approach applies to lightweight Linux distributions like CLE as well



Kitten LWK supports running native applications alongside guest OSes.

Sandia National Laboratories

Use Case 2: Tool allowing researchers to test at scale on production machines

- Currently have to request dedicated system time to test prototype system software at scale
  - Long process, difficult to navigate
  - Limited ability to iterate
- Incorporating virtualization into production software stack would allow on-demand loading of custom system software stack(s)
  - Expose effects that only occur at scale
  - VMM can provide enhanced debugging capability compared to native
  - VMM can simulate prototype hardware
  - Issue: performance may be different than native

Sandia National Laboratories

Use Case 3: Enable New Capabilities

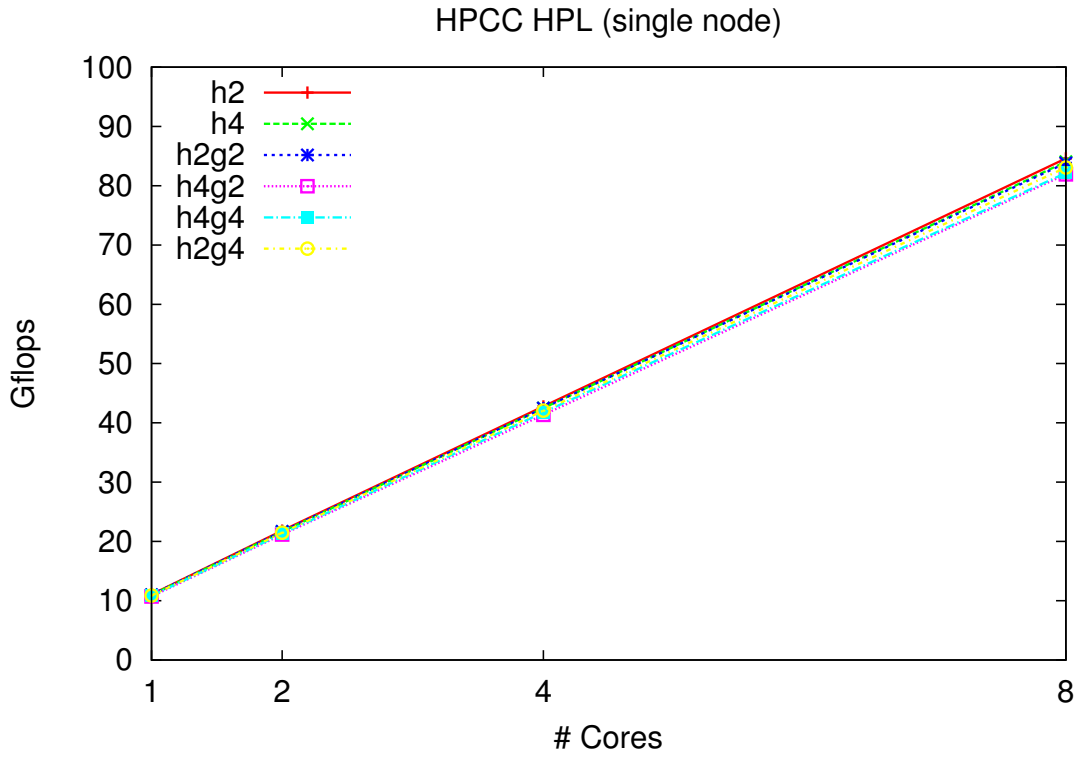
- Perform cybersecurity experiments on capability resources
  - Run commodity OSes + software
  - Multiple virtual nodes per physical node
  - Simulate Internet-scale behavior
- Dynamically replace runtime with one more suitable for the user's workload (e.g., a massive number of small jobs)
- System administrators test new vendor software without taking machine out of production
- Provide backwards capability on future platforms

Sandia National Laboratories

Virtualization Overheads

Sandia National Laboratories

Compute Virtualization Essentially Zero



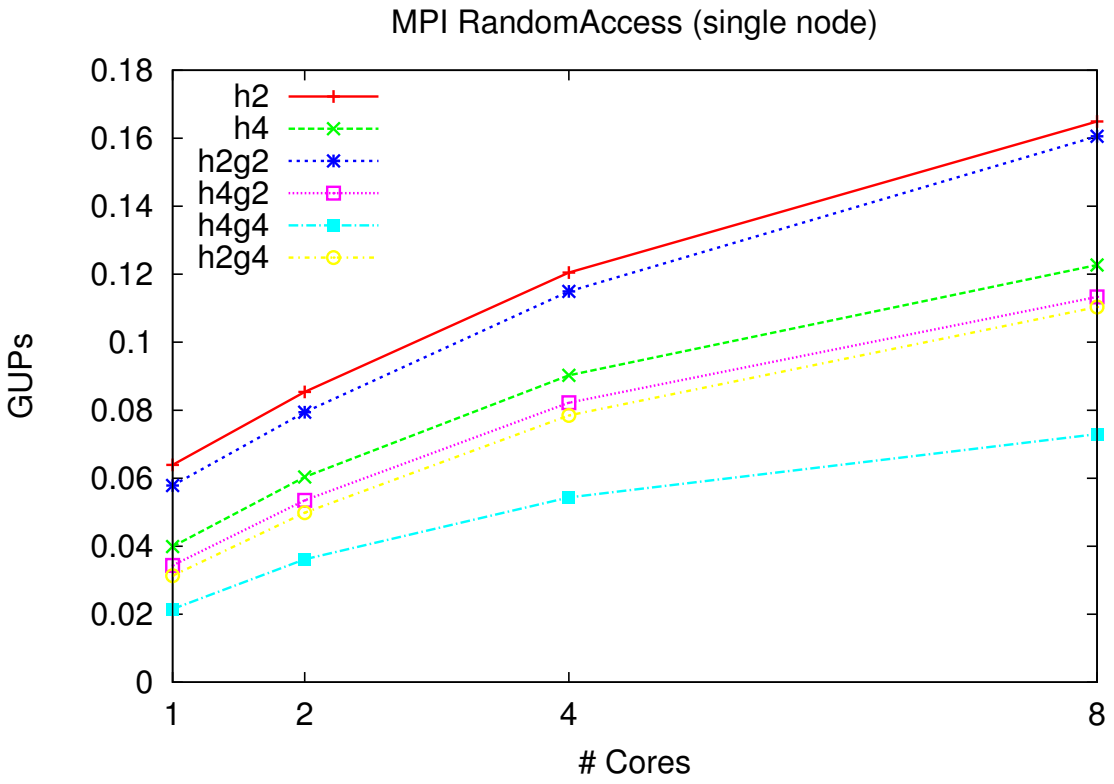
Naming:
h2 = native 2MB paging
n4 = native 4 KB paging
h2g2 = guest memory mapped with 2MB pages, hpcc running in guest using 2 MB pages
h4g2 = guest memory mapped with 4KB pages, hpcc running in guest using 2 MB pages
And so on

Node Configuration:
Intel X5570 2.93 GHz (2 sockets, 8 cores)
24 GB RAM (3x 4GB DDR-1333 per socket)
Hyperthreading disabled
Turboboost disabled

Test Configuration:
Linux 2.6.35, KVM Hypervisor
VCPU to host CPU pinning
Expose NUMA topology to guest
VM uses EPT (aka nested paging)

Sandia National Laboratories

Memory Virtualization Has Overhead, Using Large Pages Provides Mitigation



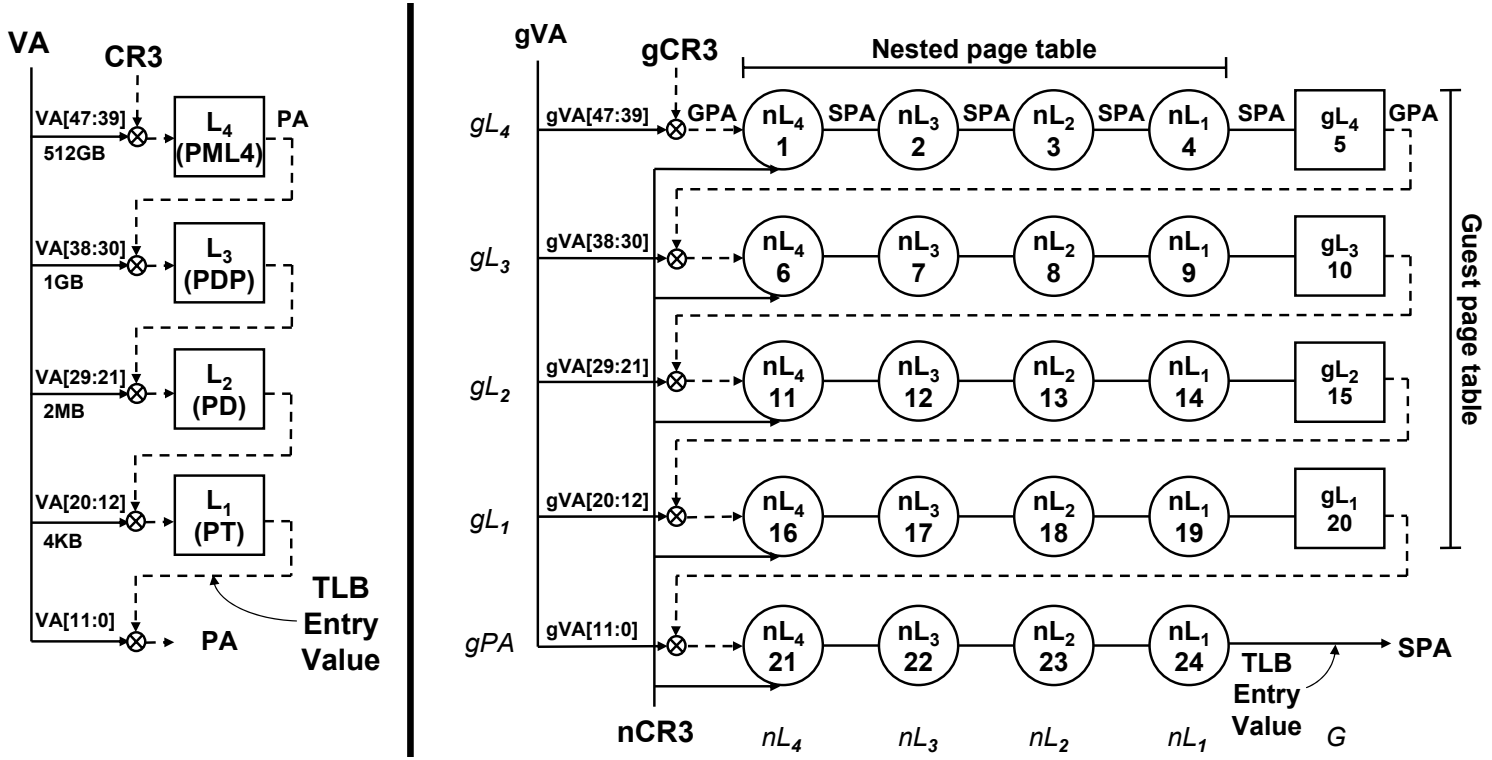
Naming:
h2 = native 2MB paging
n4 = native 4 KB paging
h2g2 = guest memory mapped with 2MB pages, hpcc running in guest using 2 MB pages
h4g2 = guest memory mapped with 4KB pages, hpcc running in guest using 2 MB pages
And so on

Node Configuration:
Intel X5570 2.93 GHz (2 sockets, 8 cores)
24 GB RAM (3x 4GB DDR-1333 per socket)
Hyperthreading disabled
Turboboost disabled

Test Configuration:
Linux 2.6.35, KVM Hypervisor
VCPU to host CPU pinning
Expose NUMA topology to guest
VM uses EPT (aka nested paging)

Sandia National Laboratories

Nested Paging Memory Virtualization



Normal – 4 levels

Nested – up to 24 memory accesses

Figure from: Ravi Bhargava, Ben Serebrin, Francesco Spanini, and Srilatha Manne. Accelerating two-dimensional page walks for virtualized systems. In Proceedings ASPLOS'08, March 2008.

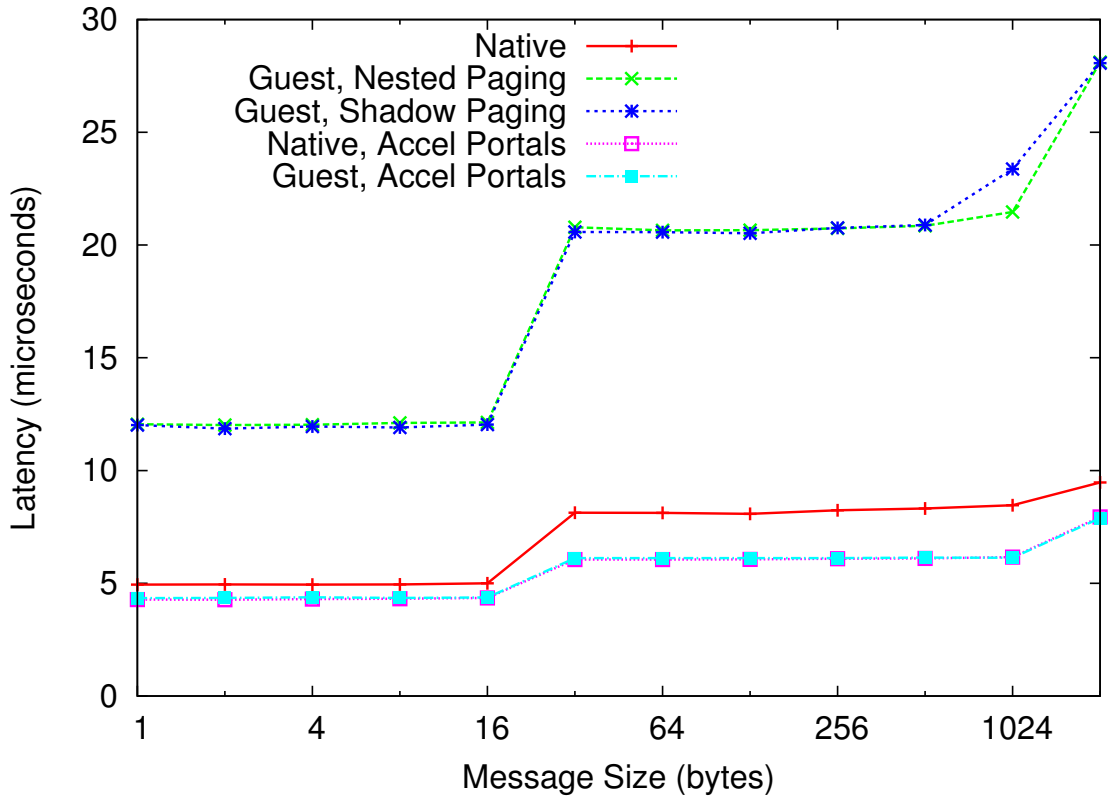
Sandia National Laboratories

Red Storm Virtualization Experiments

- Testing performed on up to 6240 quad-core Red Storm nodes, also on 48-node test system
- Compared native to guest performance
  - Native = Catamount running on bare metal
  - Guest = Kitten+Palacios running on bare metal, Catamount running as guest OS
- Seastar mapped directly through to guest, interrupts managed by Kitten+Palacios, forwarded to guest
  - Also tested “accelerated portals”, no interrupts
- Compared two guest OS memory management strategies: shadow paging and nested paging

Sandia National Laboratories

Red Storm PingPong Latency
(Inter-node, SeaStar Passed Through to Guest)

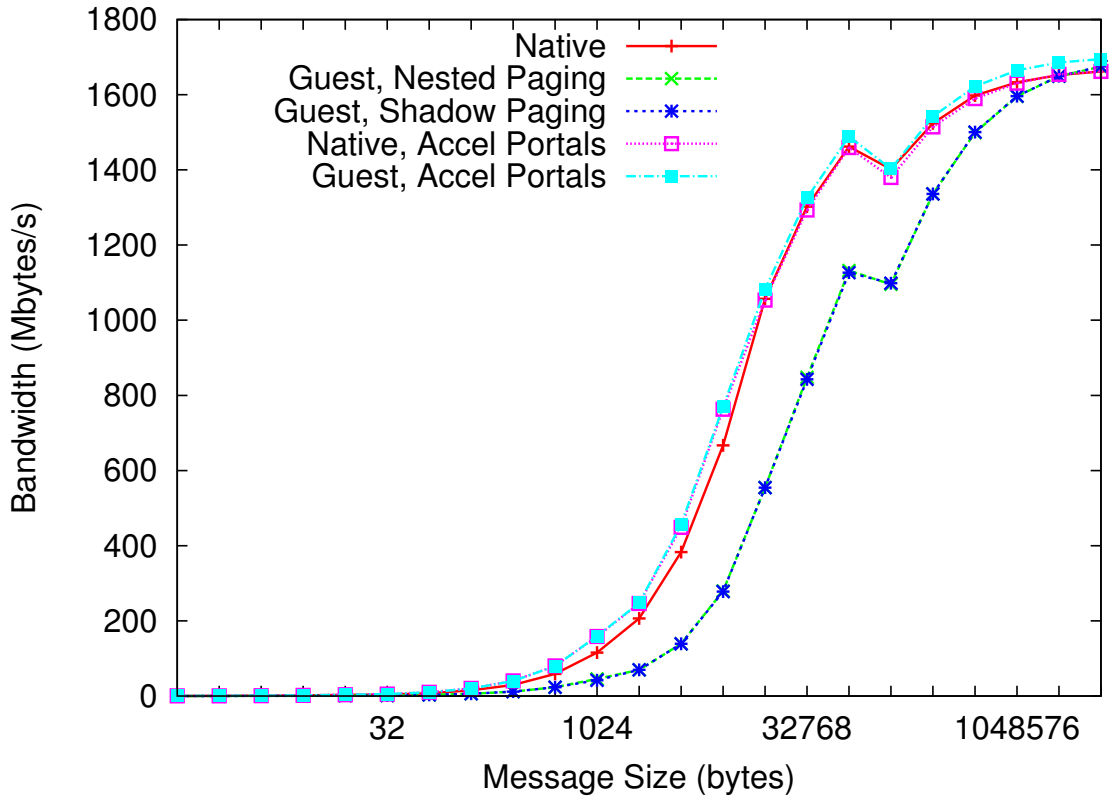


- Interrupt virtualization adds 7 to 14 us overhead for small messages
- Accelerated portals is polling base, so no interrupts.
  - Performance matches native

Sandia National Laboratories

Red Storm PingPong Bandwidth

(Inter-node, SeaStar Passed Through to Guest)

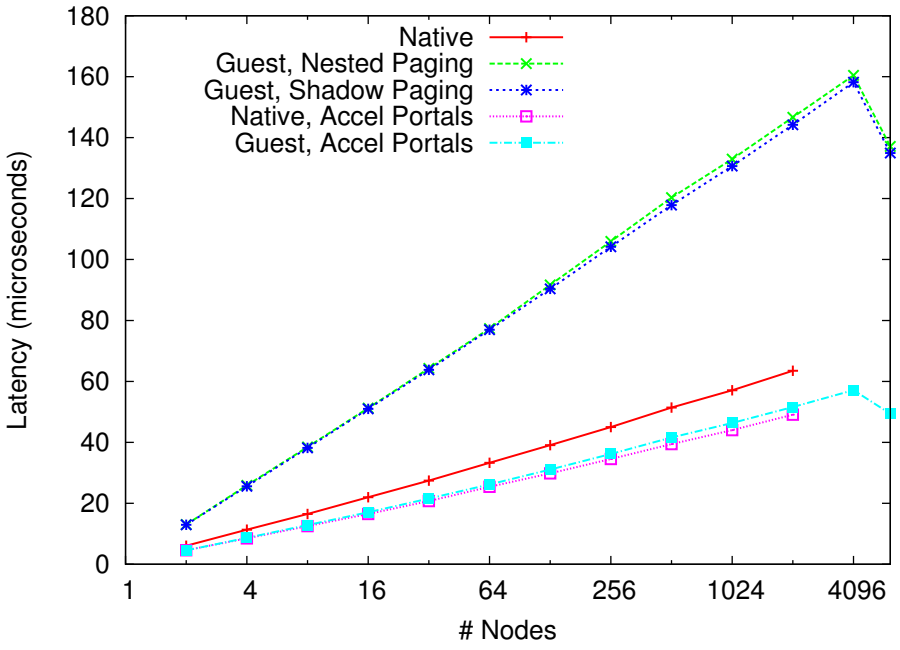


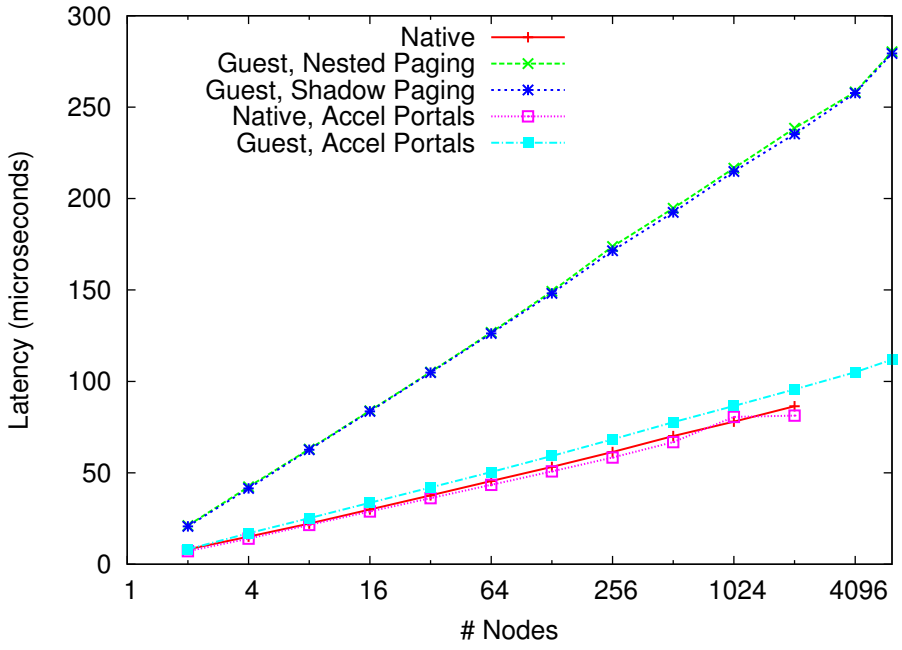
All cases reach same asymptotic bandwidth

Sandia National Laboratories

Red Storm Reduce and AllReduce Latency

(SeaStar Passed Through to Guest)



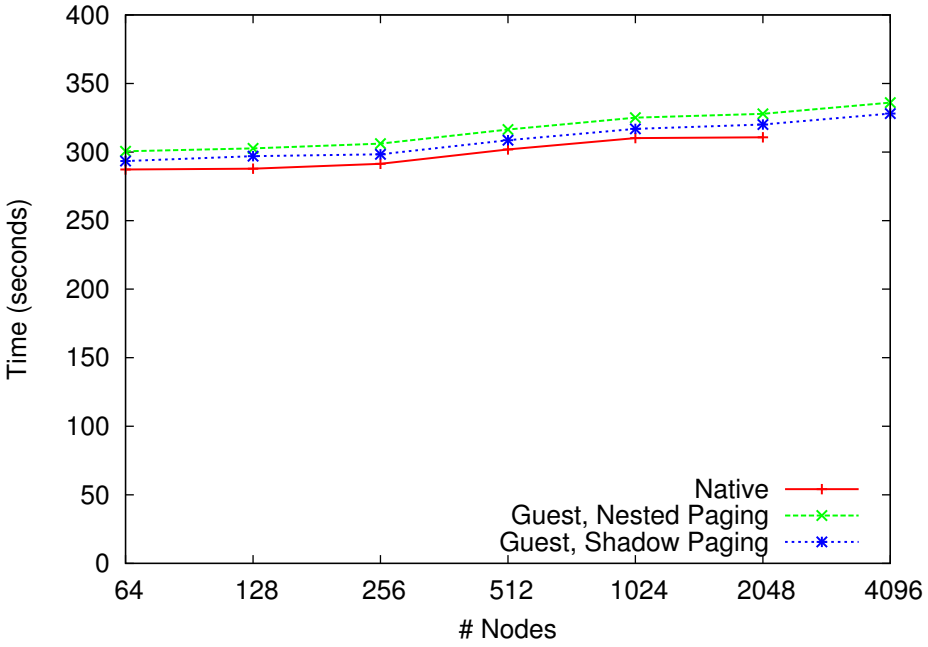


Accelerated Portals Matches Native;
Generic Portals suffers from Interrupt Virtualization Overhead

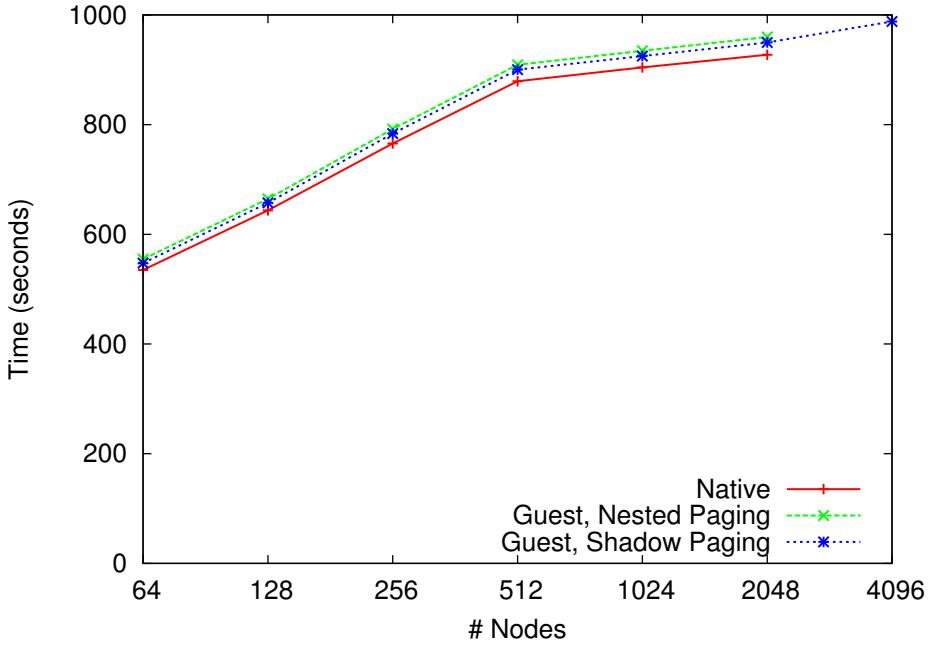
Sandia National Laboratories

Application Results from Red Storm Virtualization Experiments

CTH Hydrocode (SNL App)



Sage Hydrocode (LANL App)



Measured < 5% virtualization overhead for both applications

Sandia National Laboratories

Current Project: DOE/ASCR X-Stack

- **Objective: Enable X-Stack research and HW/SW co-design for exascale systems by leveraging the virtualization capabilities in modern processors**
- **Desired Capabilities**
  - **Enable X-Stack researchers to run new OS stacks at scale on production ASCR systems**
  - **Test potential architectural innovations at scale as extensions to the virtual machine**
  - **Measure system performance across multiple hardware/software boundaries**
- **Example Research**
  - **Scalable virtualization, VM management tools on modern HPC systems**
  - **Integration with cycle-accurate simulation/large-scale emulation techniques**
  - **Explore novel techniques in the VMM, both proposed and potentially in collaboration with other X-Stack or Critical Tech. researchers**
- **Consortium of researchers from Univ. New Mexico, Northwestern University, Oak Ridge, and Sandia**

Sandia National Laboratories

Conclusion

- Applying virtualization technology to HPC
  - Compelling use cases, enable new capabilities
  - Manageable overheads even at scale

- Next steps:
  - Test more applications, better characterize overheads for different workload classes
  - Push vendors to incorporate virtualization support in production software stacks
  - Leverage virtualization in exascale research

Sandia National Laboratories

Backup Slides

Sandia National Laboratories

Shadow vs. Nested Paging
No Clear Winner

Shadow Paging
O(N) memory accesses per TLB miss

Palacios managed page tables used by the CPU

Page Faults

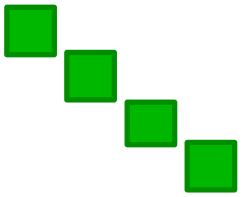
Page tables the guest OS thinks it is using

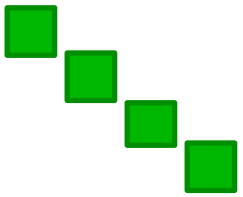
Nested Paging
O(N^2) memory accesses per TLB miss

Palacios managed guest phys to host phys page tables

CPU MMU

Guest OS managed guest virt to guest phys page tables



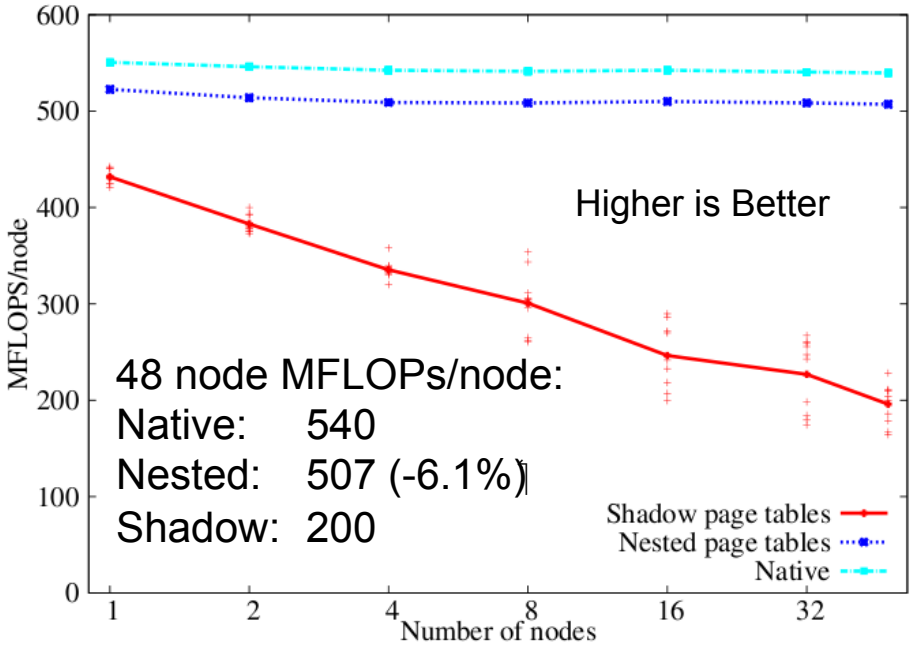


Sandia National Laboratories

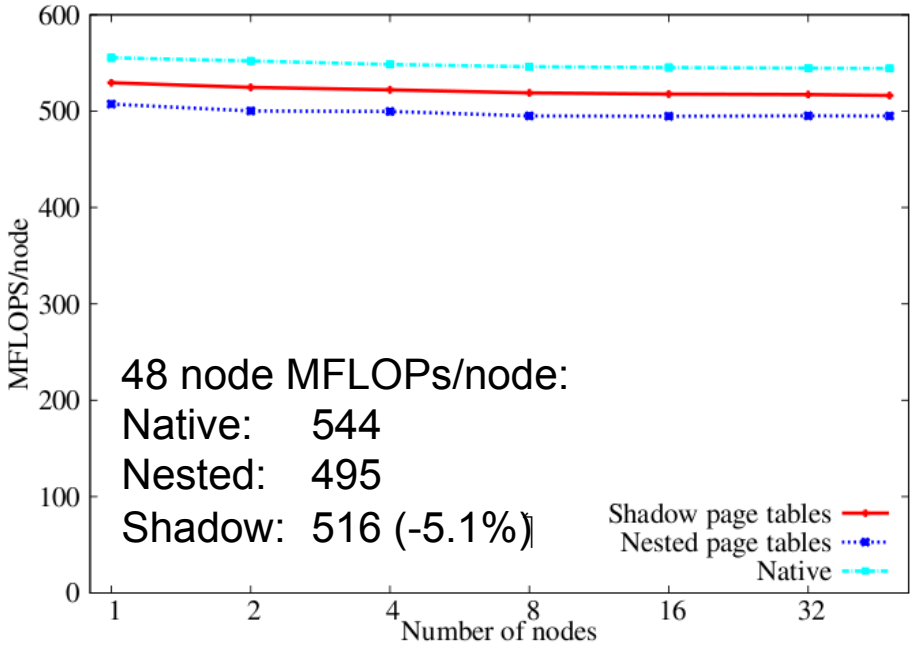
Memory Management Depends on Guest

HPCCG CG “Mini-application”

Compute Node Linux



Catamount



- Poor performance of shadow paging on CNL due to context switching. Could be partially avoided by adding page table caching to Palacios.
- Catamount is essentially doing no context switching, benefiting shadow paging (2n vs. n^2 page table depth issue)

Sandia National Laboratories