SR-IOV Support for Virtualization on InfiniBand Clusters: Early Experience

Jithin Jose, Mingzhe Li, Xiaoyi Lu, Krishna Kandalla, Mark Arnold and Dhabaleswar K. (DK) Panda

Network-Based Computing Laboratory
Department of Computer Science and Engineering
The Ohio State University, USA
Outline

• Introduction
• Problem Statement
• Challenges in Evaluating SR-IOV
• Performance Evaluation
• Conclusion & Future Work
Introduction

• Cloud computing paradigm has become increasingly popular

• Organizations provide computing, storage, and infrastructure as a service
  – Amazon Cloud, Google Cloud

• Modern Virtual Machine Technology offers attractive features to manage hardware and software components
  – Security guarantees, performance isolation, live migration
HPC on Cloud?

“HPC is all about performance, performance, performance!”

– Marc Snir, Keynote Talk, CCGrid’13

HPC application middlewares (MPI, PGAS) rely extensively on the features of modern interconnects

InfiniBand (IB) – most popular HPC interconnect

– More than 44% of the TOP500 (top500.org) systems use IB
– Offers attractive features such as RDMA, Atomics
– IP-over-IB (IPoIB) for socket applications
– Offers different communication semantics
  • Send-receive and memory semantics
– Offers two communication progress modes
  • Blocking and polling modes

Virtualization techniques have reduced the performance gap between native and virtualized modes, but how far?
State-of-the-art I/O Virtualization Techniques

• Software Based Schemes
  – VMs access physical devices through Virtual Machine Monitors
  – Full Virtualization, Para-virtualization, Software emulation
  – Overheads: context switches, memory copies, extra scheduling!

• Hardware Based Schemes
  – Performance-critical I/O operations carried out in a guest VM by interacting with hardware directly
    • Single Root I/O Virtualization (SR-IOV)
    • Multi Root I/O Virtualization (MR-IOV)
  – Recent studies demonstrate SR-IOV is significantly better than software-based solutions for GigE and 10GigE networks
Single Root I/O Virtualization (SR-IOV)

- SR-IOV specifies native I/O Virtualization capabilities in the PCI Express (PCIe) adapters
- Physical Function (PF) presented as multiple Virtual Functions (VFs)
- Virtual device can be dedicated to a single VM through PCI pass-through
- VM can directly access the corresponding VF

Is the SR-IOV support for InfiniBand networks, ready for “Prime-Time” HPC workloads?
Outline

• Introduction
• Problem Statement
• Challenges in Evaluating SR-IOV
• Performance Evaluation
• Conclusion & Future Work
Problem Statement

• What are the performance characteristics and trade-offs of using the SR-IOV?
• What are the performance characteristics of HPC middlewares when used with SR-IOV over InfiniBand?
• How does different VM deployment policies impact performance when used with SR-IOV?
• Can we offer insights into the performance characteristics of scientific application benchmarks?
Outline

• Introduction
• Problem Statement
• Challenges in Evaluating SR-IOV
• Performance Evaluation
• Conclusion & Future Work
Virtualization on Multi-core Systems

• Nodes are getting fatter
  – Nodes with 32, 64 CPU cores already available!

• Multiple VMs per host requires I/O to be virtualized

• Enables deployment of multiple Virtual Machines (VMs) per host

• VMs can be deployed in many ways
  – VM per CPU core
  – VM per CPU socket
  – VM per host node
Different Communication modes, HPC Middlewares

• InfiniBand Communication Modes
  – Send-Recv and RDMA semantics
  – Blocking and Polling based progress modes

• HPC Middlewares
  – MPI, PGAS models
  – IPoIB for socket-based applications
  – Point-to-point and collective operations
  – Application benchmarks
Challenges in Evaluating SR-IOV

• Need a ‘multi-dimensional’ performance evaluation of SR-IOV with InfiniBand

- HPC Middleware (MPI/PGAS)
  - Performance
    - Point-to-point
    - Collectives
- Performance of IPoIB for socket applications
- Performance of HPC application benchmarks

- IB Communication Modes
  - InfiniBand Communication Semantics
    - Send-receive and Memory
  - InfiniBand Progress Modes
    - Polling and blocking

- VM Subscription
  - VM per core
  - VM per CPU socket
  - VM per node
- VM Scalability
Outline

- Introduction
- Problem Statement
- Challenges in Evaluating SR-IOV
- Performance Evaluation
- Conclusion & Future Work
Experiment Setup

• Experimental testbed
  – Four compute nodes with Intel Sandy Bridge-EP platform
  – Intel Xeon E5-2670 2.6GHz eight-core processors
  – 32 GB of main memory per node
  – Mellanox ConnectX-3 FDR cards (56 Gbps), connected to a Mellanox FDR switch SX6036
  – Mellanox OpenFabrics Enterprise Edition (MLNX OFED) SRIOV-ALPHA-3.3.0-2.0.0008
  – KVM as the Virtual Machine Monitor (VMM)
MVAPICH2/MVAPICH2-X Software

• High Performance open-source MPI Library for InfiniBand, 10Gig/iWARP and RDMA over Converged Enhanced Ethernet (RoCE)
  – MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and initial MPI-3.0), Available since 2002
  – MVAPICH2-X (MPI + PGAS), Available since 2012
  – Used by more than 2,000 organizations (HPC Centers, Industry and Universities) in 70 countries
  – More than 168,000 downloads from OSU site directly
  – Empowering many TOP500 clusters
    • 7th ranked 204,900-core cluster (Stampede) at TACC
    • 14th ranked 125,980-core cluster (Pleiades) at NASA
    • 17th ranked 73,278-core cluster (Tsubame 2.0) at Tokyo Institute of Technology
      and many others
  – Available with software stacks of many IB, HSE and server vendors including Linux Distros (RedHat and SuSE)
    – http://mvapich.cse.ohio-state.edu
• Partner in the U.S. NSF-TACC Stampede (9 PFlop) System
Performance Evaluation

- InfiniBand Communication Semantics
  - Send-recv and Memory
- InfiniBand Progress Modes
  - Polling and blocking

- HPC Middleware (MPI/PGAS) Performance
  - Point-to-point
  - Collectives
- Performance of IPoIB
- Performance of HPC application benchmarks

- VM Subscription
  - VM per core
  - VM per CPU socket
  - VM per node
- VM Scalability
InfiniBand Communication Semantics

- Significant performance difference for small messages
  - 0.87us (native) and 1.53 us (SR-IOV) for two byte message size
- Performance gap because of lack of inline message support
- Large message performance is comparable
- Performance similar for send-recv and memory semantics
Performance of MPI Latency, Bandwidth

- Performance evaluations using OSU MPI benchmarks
- Used MVAPICH2-1.9a2 as the MPI Library
- Comparable performance for Native and SR-IOV
  - 1.02us (native) and 1.39us (SR-IOV) for one byte message size
- MVAPICH2 uses ‘RDMA-FastPath’ optimization for small messages
  - Similar characteristics as that of RDMA write
Performance of PGAS Get operation

- Performance evaluation with OSU Unified Parallel C (UPC) Get benchmark
- Used MVAPICH2-X-1.9a2 as the UPC Stack
- Significant performance gap between Native and SR-IOV modes
  - 1.81us (native) and 2.16us (SR-IOV) for one byte message size
- ‘upc_memget’ implemented directly over RDMA Get operation
Performance Evaluation

- InfiniBand Communication Semantics
  - Send-receive and Memory
- InfiniBand Progress Modes
  - Polling and blocking

- HPC Middleware (MPI/PGAS) Performance
  - Point-to-point
  - Collectives
- Performance of IPoIB
- Performance of HPC benchmarks

- VM Subscription
  - VM per core
  - VM per CPU socket
  - VM per node
- VM Scalability
Performance Polling vs. Blocking Modes
(verbs-level)

- Polling Mode
  - 0.83us (native) and 1.53us (SR-IOV) for one byte message size

- Blocking Mode
  - 6.19us (native) and 28.43us (SR-IOV) for one byte message size

- Higher overhead in blocking mode
  - Lack of optimizations related to serving interrupts
Performance Polling vs. Blocking Modes (MPI-level)

- Performance Evaluations using MVAPICH2
- MVAPICH2 employs a hybrid scheme in blocking configuration
  - Polls for a specific number of times, then switches to blocking mode
- Polling Mode: 1.02us (native) and 1.39us (SR-IOV) for one byte message size
- Blocking Mode: 1.46us (native) and 1.89us (SR-IOV) for one byte message size
- Similar performance for MPI bandwidth
Performance Evaluation

- HPC Middleware (MPI/PGAS) Performance
  - Point-to-point
  - Collectives
- Performance of IPoIB
- Performance of HPC benchmarks

IB Communication Modes
- InfiniBand Communication Semantics
  - Send-recv and Memory
- InfiniBand Progress Modes
  - Polling and blocking

HPC Middlewares

Virtual Machine Configurations
- VM Subscription
  - VM per core
  - VM per CPU socket
  - VM per node
- VM Scalability
Performance of IP-over-IB (IPoIB)

- Performance evaluations using ‘Netperf’ benchmark
- Significant performance difference for IPoIB
  - 25.65us (native) and 53.74us (SR-IOV) for one byte message size
- TCP Stack overheads are significant in virtualized mode!
Performance Evaluation

- HPC Middleware (MPI/PGAS) Performance
  - Point-to-point
  - Collectives
- Performance of IPoIB
- Performance of HPC benchmarks

IB Communication Modes
- InfiniBand Communication Semantics
  - Send-recv and Memory
- InfiniBand Progress Modes
  - Polling and blocking

HPC Middlewares

Virtual Machine Configurations
- VM Subscription
  - VM per core
  - VM per CPU socket
  - VM per node
- VM Scalability
• VMs can be deployed as
  – VM per host CPU core, VM per host CPU socket, VM per host

• Evaluations with OSU collective benchmarks

• Number of processes was kept as constant (28)

• VM per node performs better for both collectives

• Performance difference compared to native mode
  – Lack of shared memory communication in virtualized mode
• Similar trends for message rate evaluation
• Native mode offers higher message rate
  – 2.5 Million messages/sec
• Best message rate for VM-per-node configurations
  – 2.1 Million messages/sec
Performance Evaluation

- HPC Middleware (MPI/PGAS) Performance
  - Point-to-point
  - Collectives
- Performance of IPoIB
- Performance of HPC benchmarks

- InfiniBand Communication Semantics
  - Send-recv and Memory
- InfiniBand Progress Modes
  - Polling and blocking

- VM Subscription
  - VM per core
  - VM per CPU socket
  - VM per node
- VM Scalability
Virtual Machine Scalability

- Evaluations with MPI Graph500 benchmark
  - Communication intensive, irregular benchmark
- Varied the number of VMs per node, and compared with number of processes per node, while keeping the problem size constant
- Execution time reduces with increase in number of VMs initially
- Performance decreases after 8 VMs per node
- Indicates performance limitations with fully subscribed mode!
Outline

• Introduction
• Problem Statement
• Challenges in Evaluating SR-IOV
• Performance Evaluation
• Conclusion & Future Work
Conclusion & Future Work

- Presented our initial evaluation results of SR-IOV over InfiniBand
- Explored different dimensions for performance evaluation
  - InfiniBand communication semantics, progress modes, VM configurations, VM scalability, HPC middlewares
- Evaluation Highlights
  - Higher latency for small messages
  - Comparable point-to-point performance for medium and large messages
  - Overheads with ‘blocking’ mode for communication progress
  - Performance limitations for collective operations, message rate evaluations, and for fully-subscribed VM modes
- Plan to evaluate real-world HPC applications with SR-IOV
- Plan to explore designs for improving middleware (MPI/PGAS) performance in virtualized environment
Thank You!

{jose, limin, luxi, kandalla, arnoldm, panda}
@cse.ohio-state.edu

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

MVAPICH Web Page
http://mvapich.cse.ohio-state.edu/