RDMA in the Cloud: Enabling high-bandwidth, low-latency communication in virtual environments for HPC

Josh Simons
VMware Office of the CTO
Server Virtualization

Traditional Architecture

Virtual Architecture

Virtual Machine (VM)
Secure Private Cloud for HPC

Research Group 1  Research Group \( m \)

Users \( \cdots \) IT

VMware vRealize Automation

User Portals  Blueprints  Security

Research Cluster 1  NSX  Research Cluster \( n \)

VMware vCenter Server  VMware vCenter Server  VMware vCenter Server

VMware vSphere  VMware vSphere  VMware vSphere

Hybrid/Public Clouds

Programmatic Control and Integrations
Task Parallel Performance
Testbed Configuration

- **Hardware**
  - Four two-socket HP DL380 G8 servers (3.3 GHz E5-2667v2 CPUs; 128 GB)
  - Dual-ported Mellanox FDR / 10 Gb RoCE adaptor
  - Mellanox 12-port FDR switch

- **Software**
  - ESXi 5.5u1 hypervisor
  - RHEL 6.5 (native and guest)
  - MLNX OFED 2.2.1
BioPerf Benchmark Suite

Native to Virtual Ratios  
(Higher is Better)

ESXi5.5u1

- CLUSTALW
- GLIMMER
- GRAPPA
- HMMER
- PHYLIP
- PREDATOR
- TCOFFEE
- BLAST
- FASTA
BLAST

Native to Virtual Ratios
(Higher is Better)

ESXi5.5u1

OMP_NUM_THREADS=1
OMP_NUM_THREADS=4
OMP_NUM_THREADS=8
OMP_NUM_THREADS=16
RDMA Performance
Kernel Bypass Model

PHYSICAL

user

application

sockets

tcp/ip

driver

rdma

hardware

VIRTUAL

user

application

sockets

tcp/ip

driver

rdma

guest kernel

vmkernel

hardware
FDR InfiniBand Read Latency

ib_read_lat / passthrough

Half Round trip Latency (µs)

Message Sizes (Bytes)

Native

ESXi 5.5
HPC Challenge Benchmark (HPCC)

Native to Virtual Ratios (Higher is Better)
NAS Parallel Benchmarks (NPB)
Native to Virtual Ratios (Higher is Better)

NAS Parallel Benchmarks (Class C)
NAMD

Native to Virtual Ratios (Higher is Better)
NWCHEM
Native to Virtual Ratios (Higher is Better)
NWChem Profiling
(siosi6)

Number of Messages (Millions)

Message Sizes

Source: NWChem Performance Benchmark and Profiling, HPC Advisory Council
FDR InfiniBand Read Latency: Future

ib_read_lat / passthrough / polling completions
RDMA Storage Performance
Remote Storage Access Path

OS

app
app
app

device driver

HW

PCI device

switch

storage server
Passthrough Mode Limitation

- app
- Guest OS
- Guest OS
- Guest OS
- driver
- PCI device
- hardware
- switch
- storage server
Single-Root I/O Virtualization (SR-IOV)

- Guest OS
  - VF driver
- vmkernel
- hardware
- PCI device
- storage server
- switch
FDR InfiniBand Read Latency

ib_read_lat / passthrough, SR-IOV
IOR Bandwidth Performance

3VM x 4core versus bare-metal Linux 12core

Data provided by Sorin Faibish, EMC
Office of the CTO
Summary

• Virtualized HPC performance for throughput applications generally very close to bare-metal (well under 5% overhead)

• Passthrough RDMA can deliver close to native performance for some MPI benchmarks and applications
  – Will continue to improve as latency overheads are reduced…or eliminated
  – Higher-scale testing

• SR-IOV can enable access to RDMA-connect parallel file systems from virtual environments with good performance
Thank You

Josh Simons
simons@vmware.com