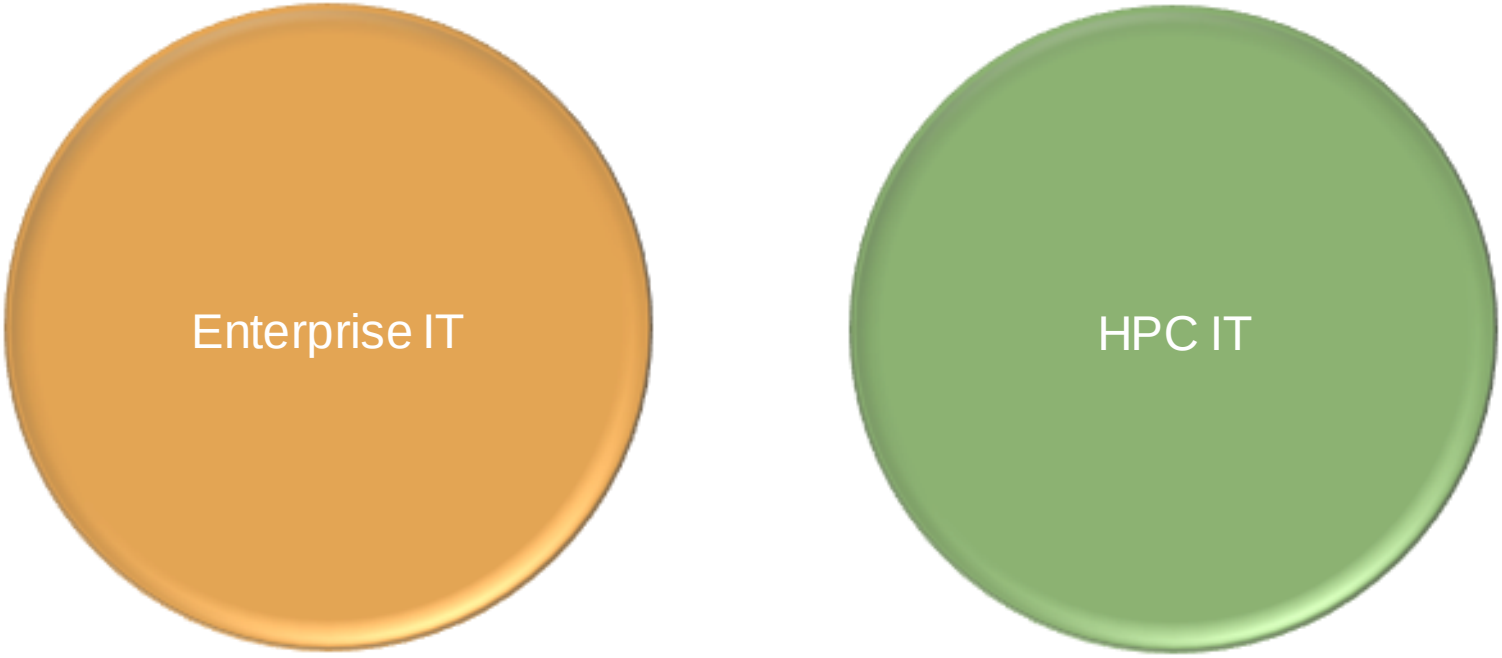
Ultra-low latency in the Cloud: How Low Can We Go?

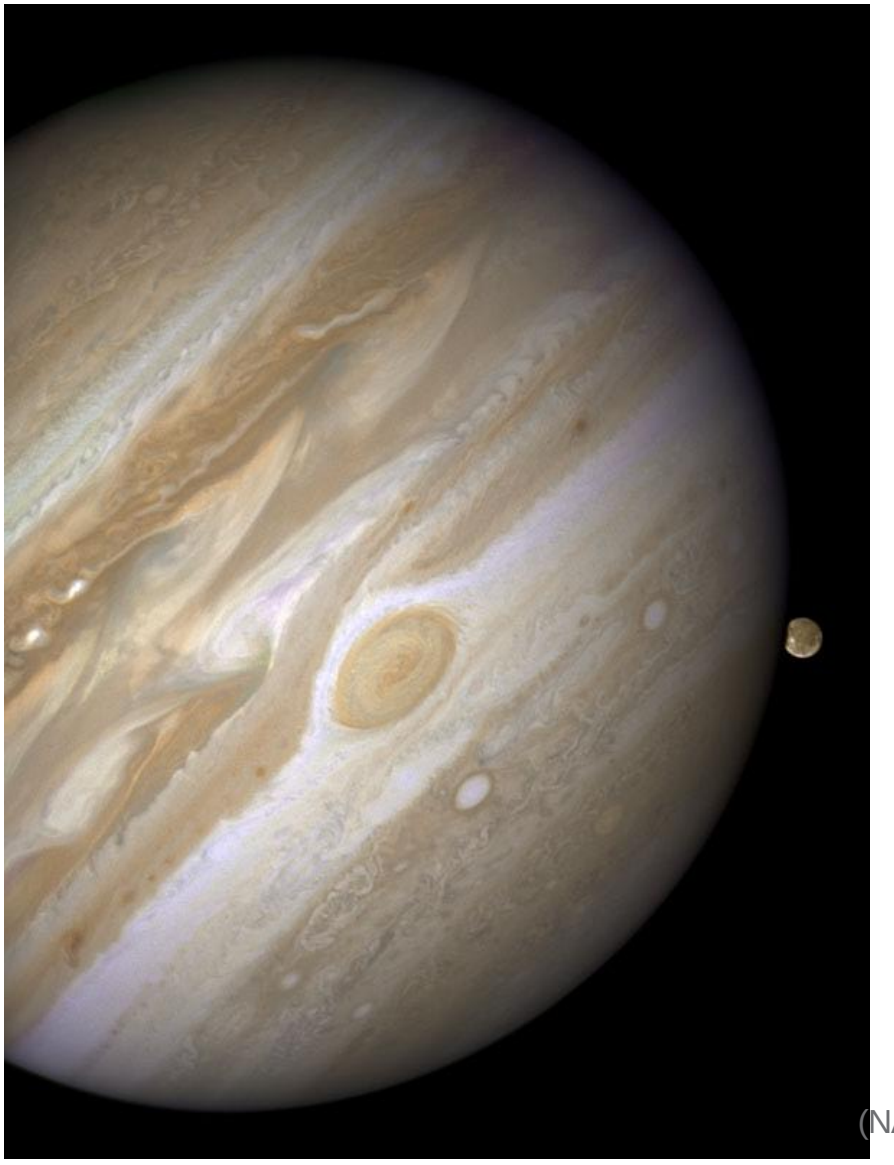
HPC Advisory Council European Conference, June 17, 2012

Josh Simons, Office of the CTO, VMware Inc.

IT Today



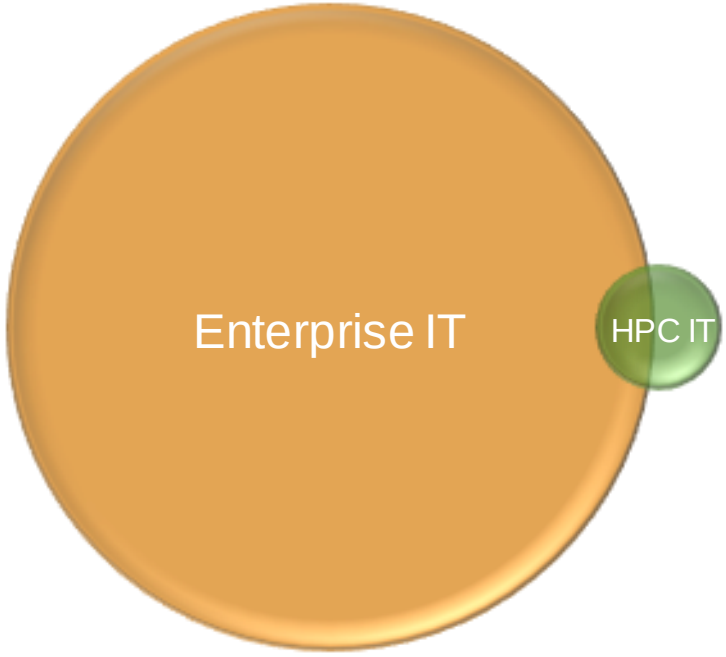
IT Today



(NASA)

Converging IT

Convergence driven by increasingly shared concerns, e.g.:



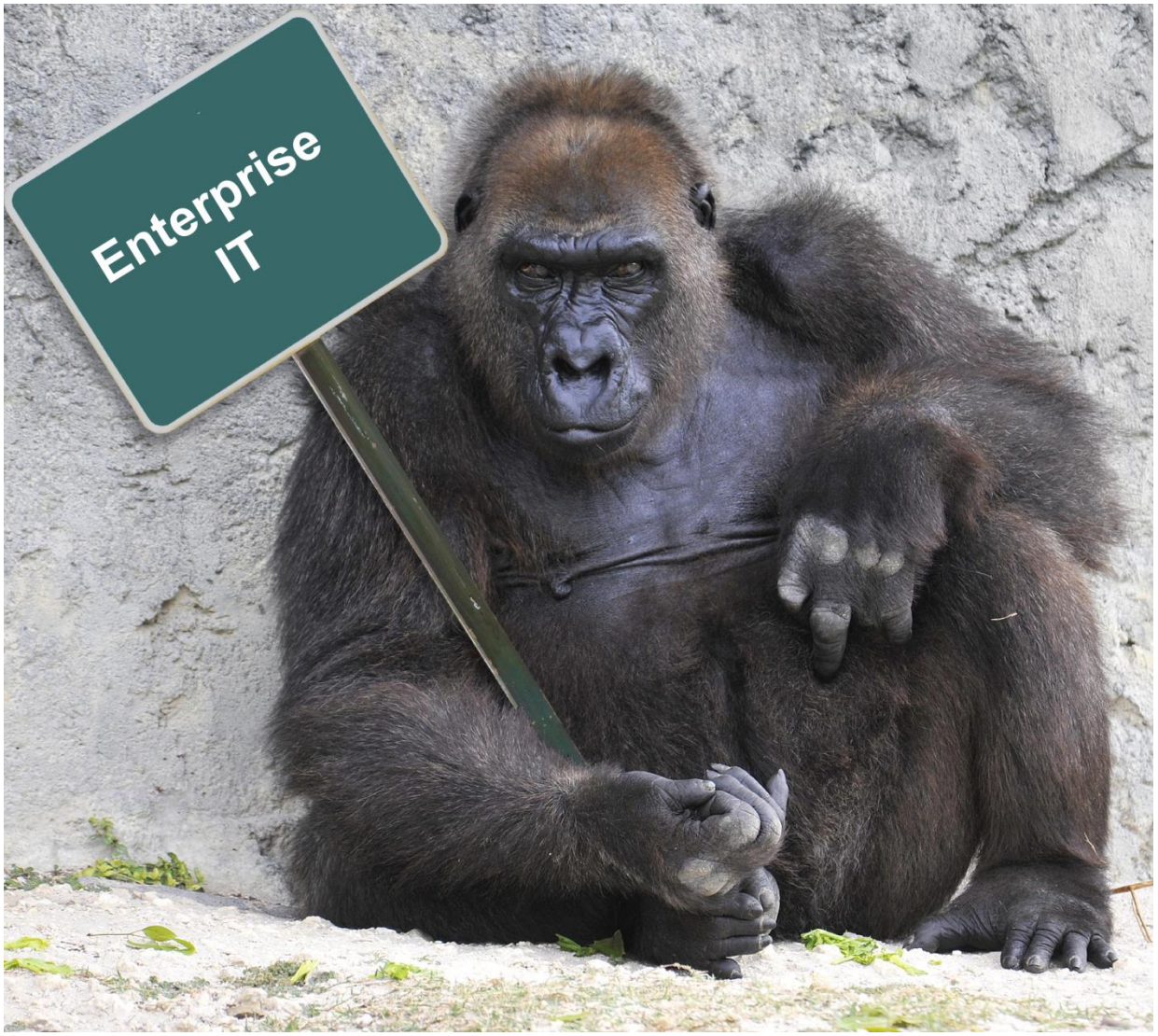
- Scale-out management
- Multi-tenancy and security
- Low latency communication
- High utilization
- Power management
- Dynamic workloads
- Application Parallelism
- Application Resiliency

Cloud Computing

$241B by 2020

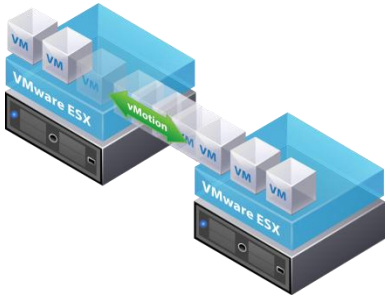
Forrester Research

What Cloud?



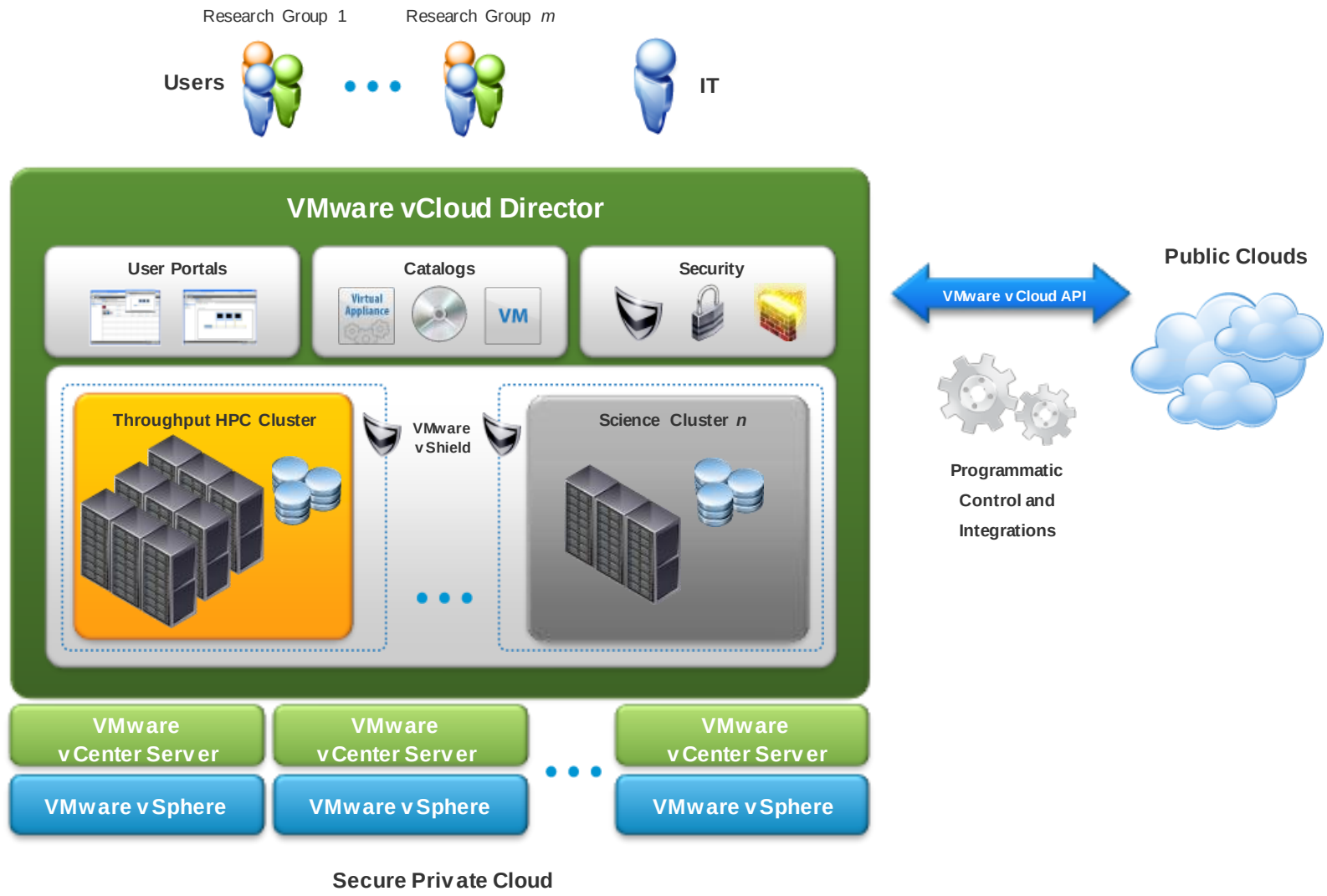
Virtualization for HPC

- **Heterogeneous environments**
  - Run YOUR software stack (OS, libraries, applications), not the site's stack



- **Dynamic resource management**
  - Move running jobs for efficiency, resiliency, power management, preventative maintenance, etc.
- **Workload isolation**
  - For secure multi-tenancy, failure protection, …and even performance
- **Application fault tolerance**
  - Checkpointing (reactive) and Predictive (proactive)
- **For current virtualization users**
  - Unification of IT infrastructure

Life Sciences / Pharma Customer Example

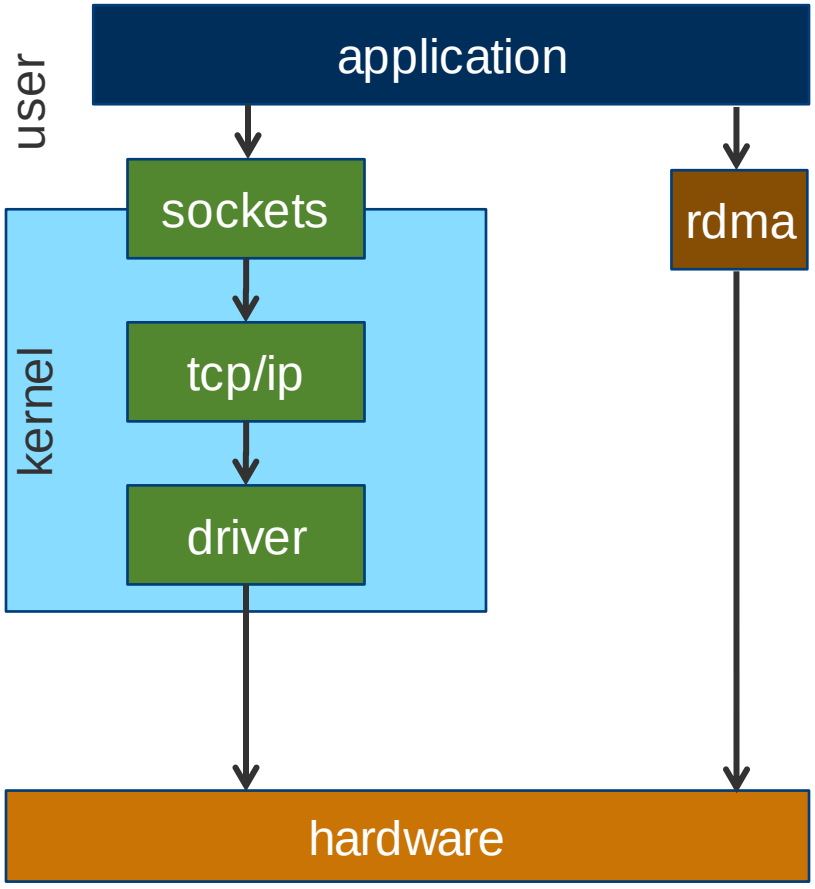


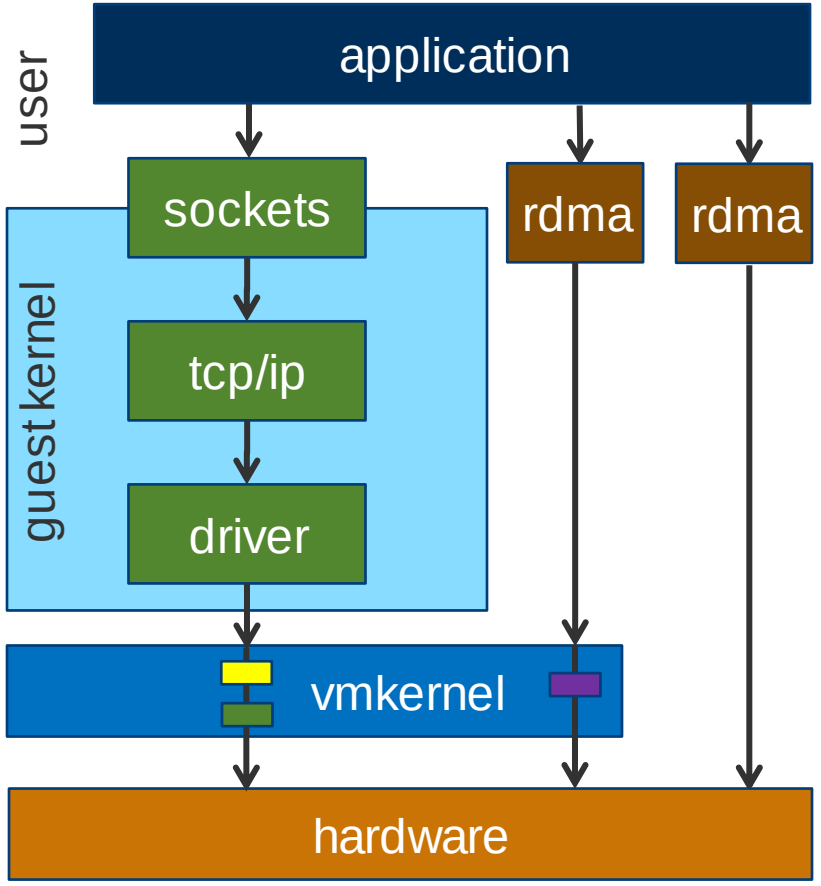
RDMA Performance in Virtual Machines using QDR InfiniBand on VMware vSphere 5

Benchmark Configuration

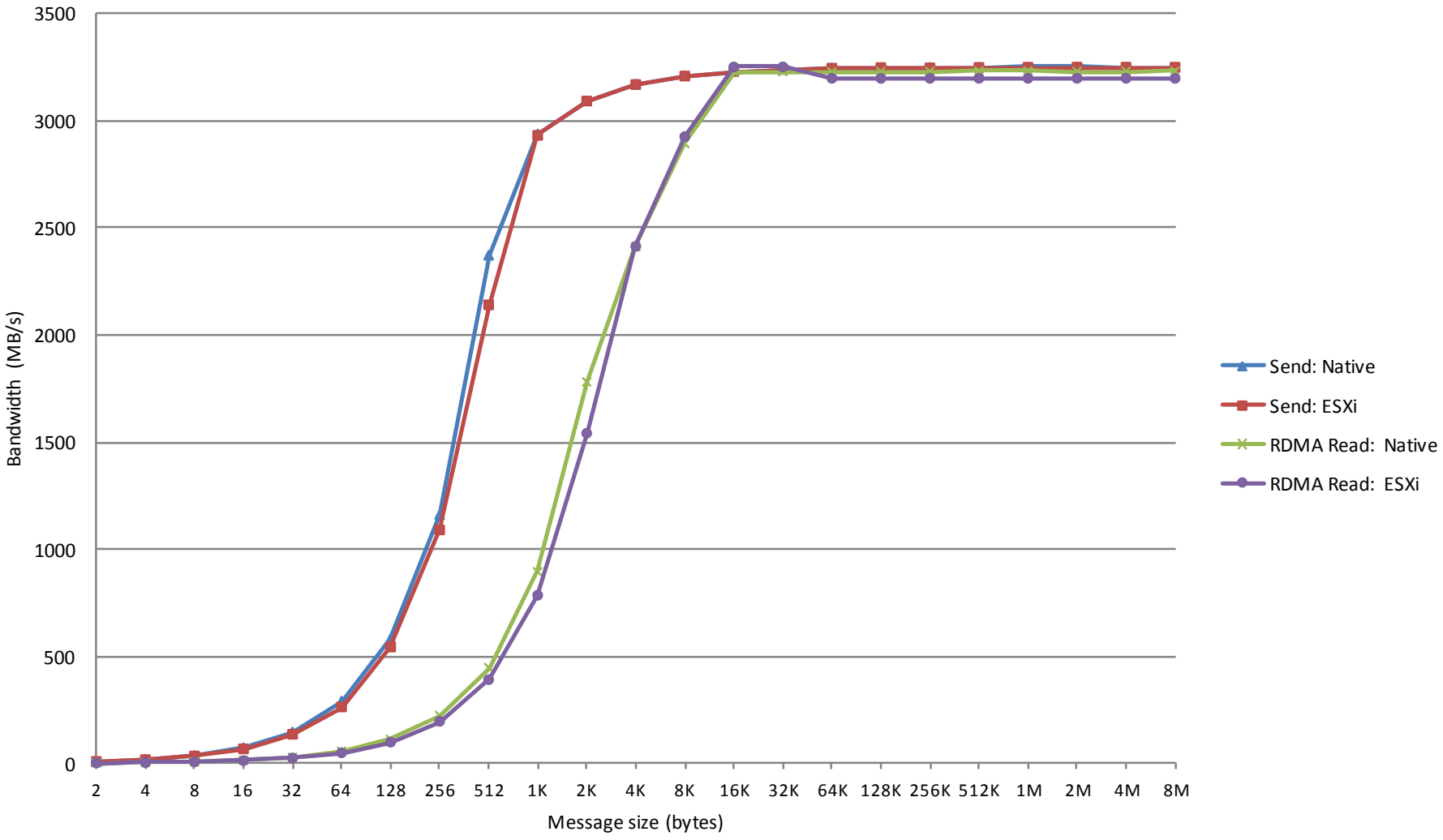
- **Hardware**
  - HP ProLiant ML350 G6, two-socket E5620 (Westmere) 2.4 GHz, 12 GB
  - Mellanox ConnectX-2 MT26428 QDR HCA
  - QSPF cable direct-connect between hosts
- **Software**
  - 64-bit RHEL 6.1 with OFED 1.5.3
  - InfiniBand device configured in ESX as a passthrough device using VMDirect Path I/O
  - All measurements made with OFED performance tests in RC mode
  - Results labeled as “ESXi” used vSphere 5.0
  - Results labeled as “ESXi ExpA” and “ESXi ExpB” included unreleased patches that will be included in a future vSphere release

Kernel and Hypervisor Bypass





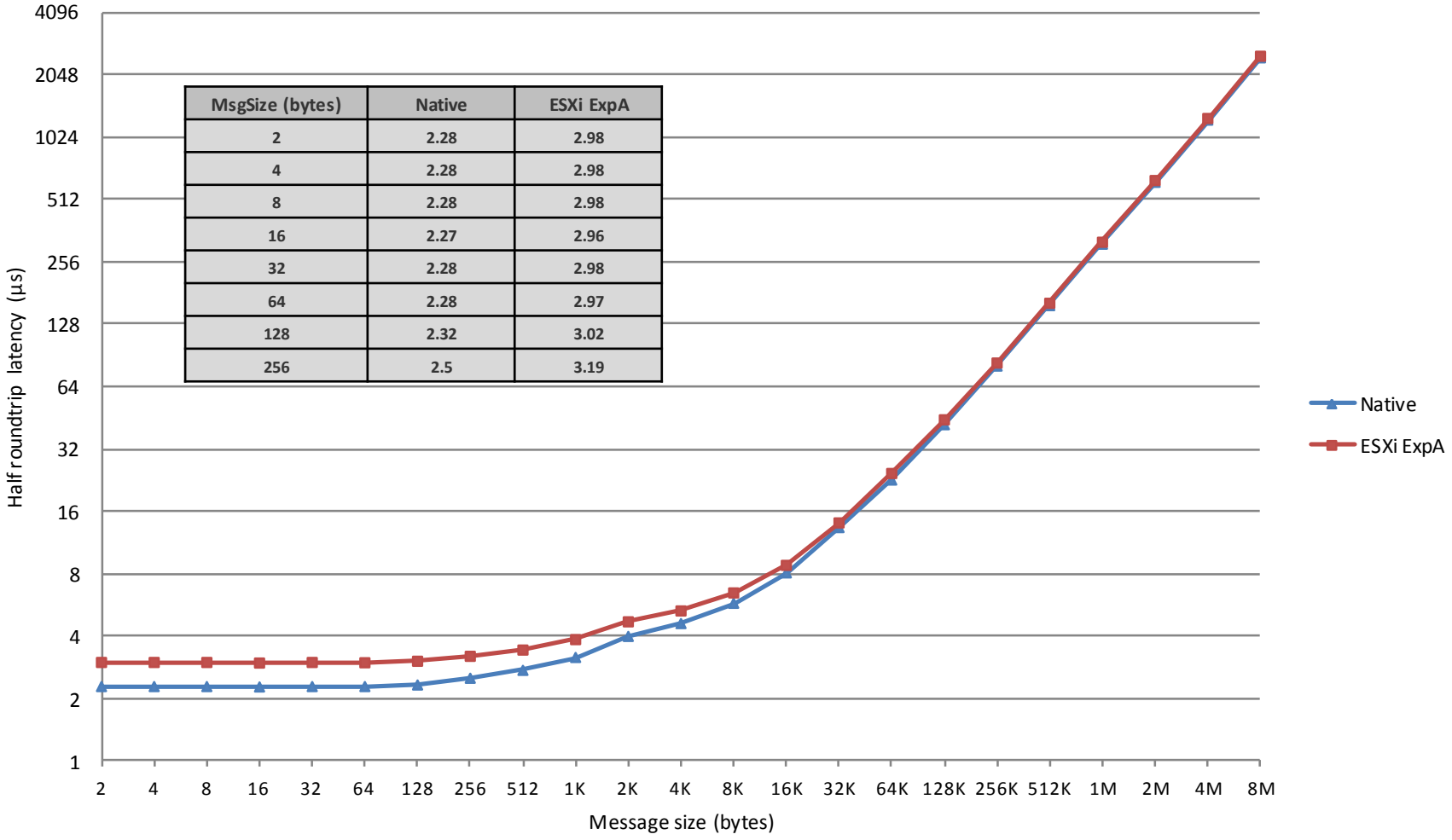
InfiniBand Bandwidth with VM DirectPath I/O



vmware

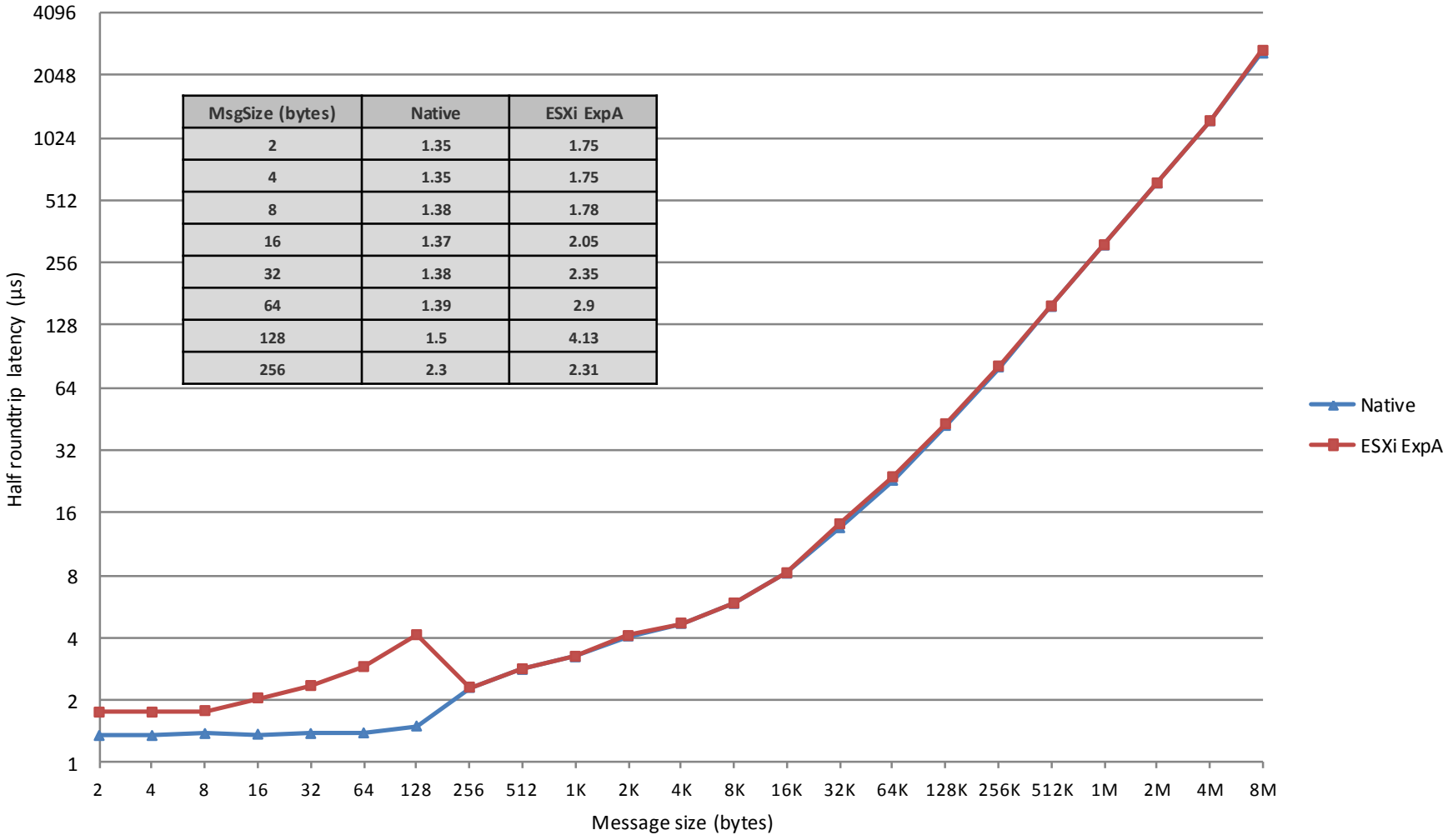
Latency with VM DirectPath I/O (RDMA Read, Polling)

MsgSize (bytes)	Native	ESXi ExpA
2	2.28	2.98
4	2.28	2.98
8	2.28	2.98
16	2.27	2.96
32	2.28	2.98
64	2.28	2.97
128	2.32	3.02
256	2.5	3.19



vmware

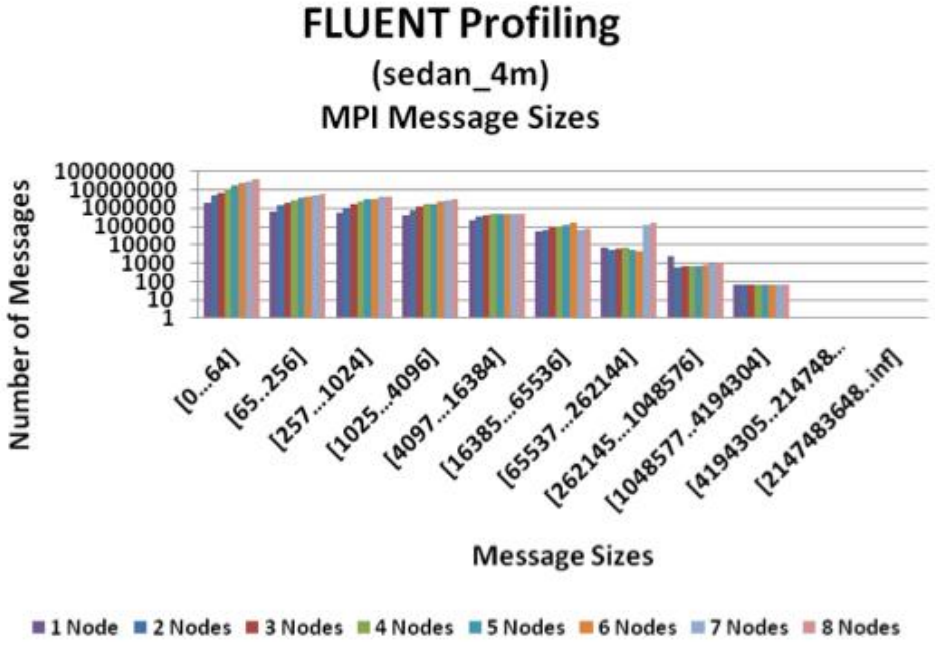
Latency with VM DirectPath I/O (Send/Receive, Polling)

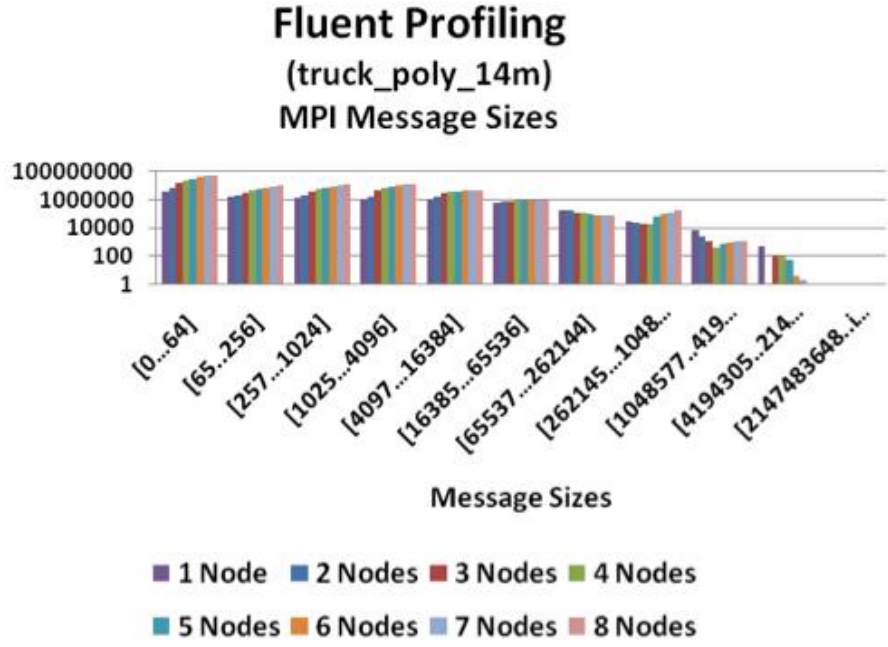


MsgSize (bytes)	Native	ESXi ExpA
2	1.35	1.75
4	1.35	1.75
8	1.38	1.78
16	1.37	2.05
32	1.38	2.35
64	1.39	2.9
128	1.5	4.13
256	2.3	2.31

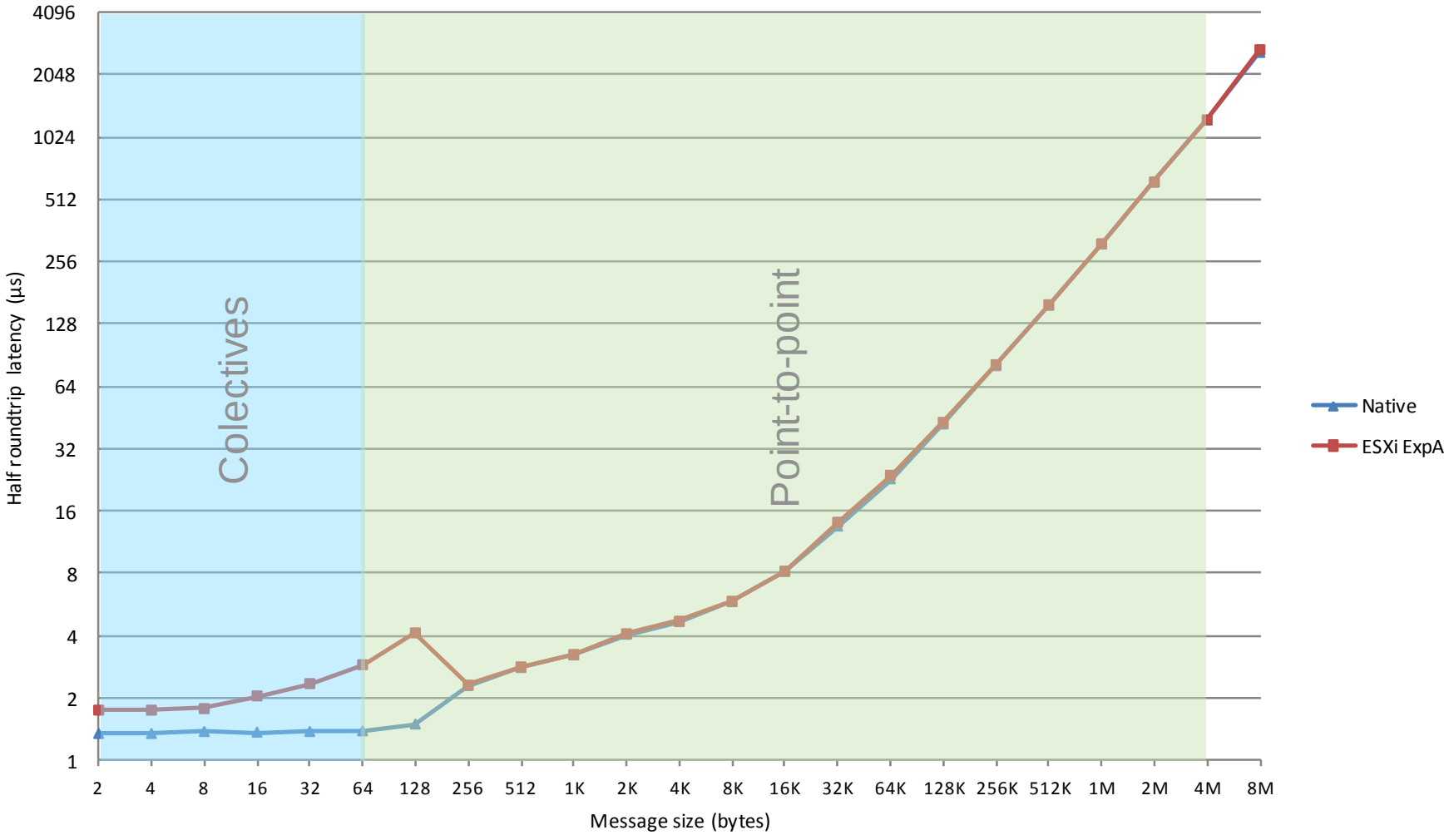
FLUENT Profiling – MPI Message Sizes

- **MPI message sizes are concentrated in range of small message sizes**
  - Majority are in the range of 0B and 64B
  - Small messages are typical used for synchronization, implies FLUENT is latency sensitive
- **Larger message sizes also appeared but at a smaller percentage**
  - Larger messages (65B to 4MB) responsible for data transfers between the MPI ranks
  - Implies that FLUENT also requires high network throughput





Latency with VM DirectPath I/O (Send/Receive, Polling)

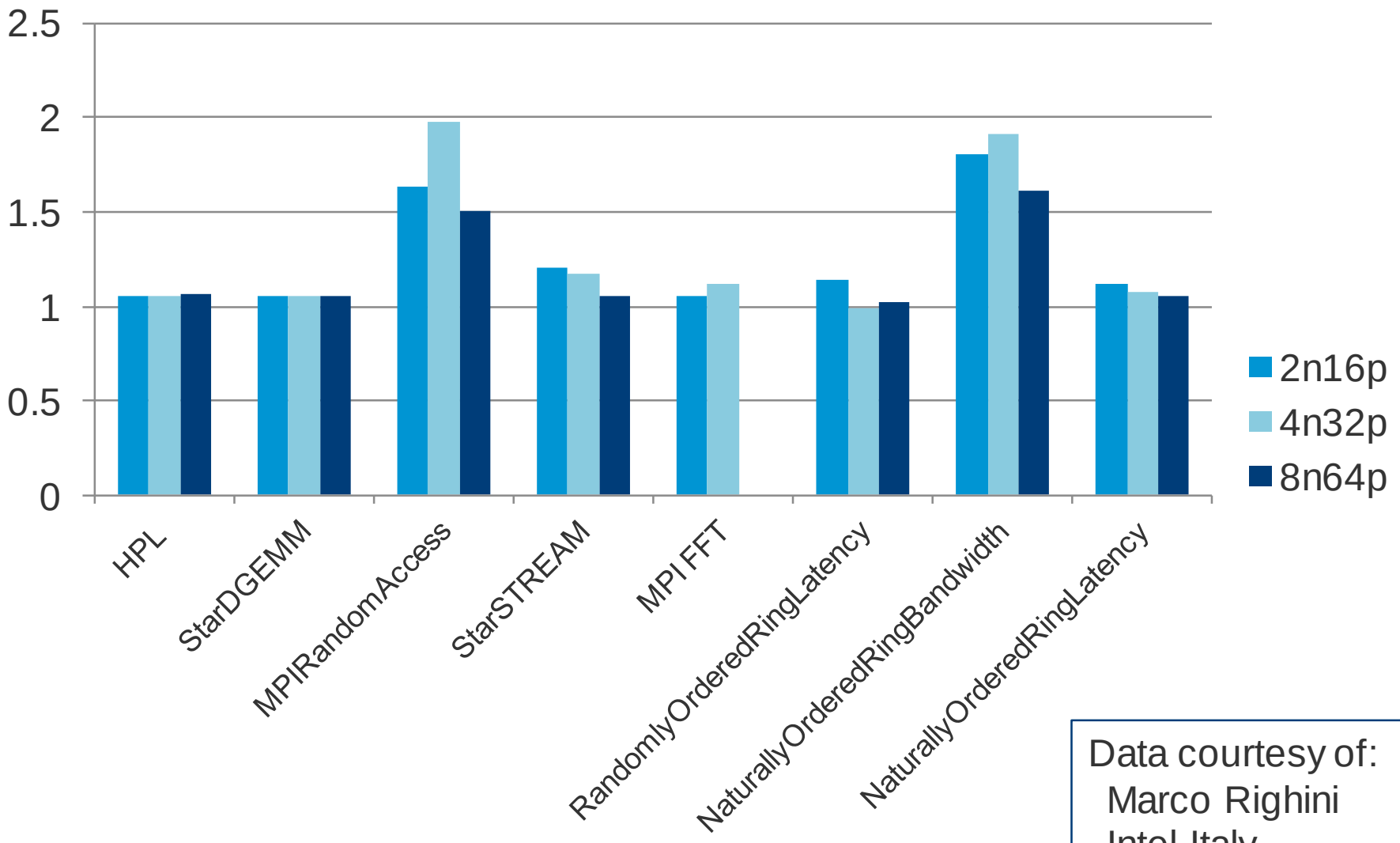


vmware

Intel Experiment: HPCC

- **Dated, but still useful**
- **Hardware**
  - 8 two-socket 2.93GHz X5570 nodes, 24 GB
  - Dual-port Mellanox DDR InfiniBand adaptor
  - Mellanox 36-port switch
- **Software**
  - vSphere 4.0 (current version is 5.0)
  - Platform Open Cluster Stack (OCS) 5 (native and guest)
  - Intel compilers 11.1
  - HPCC 1.3.1

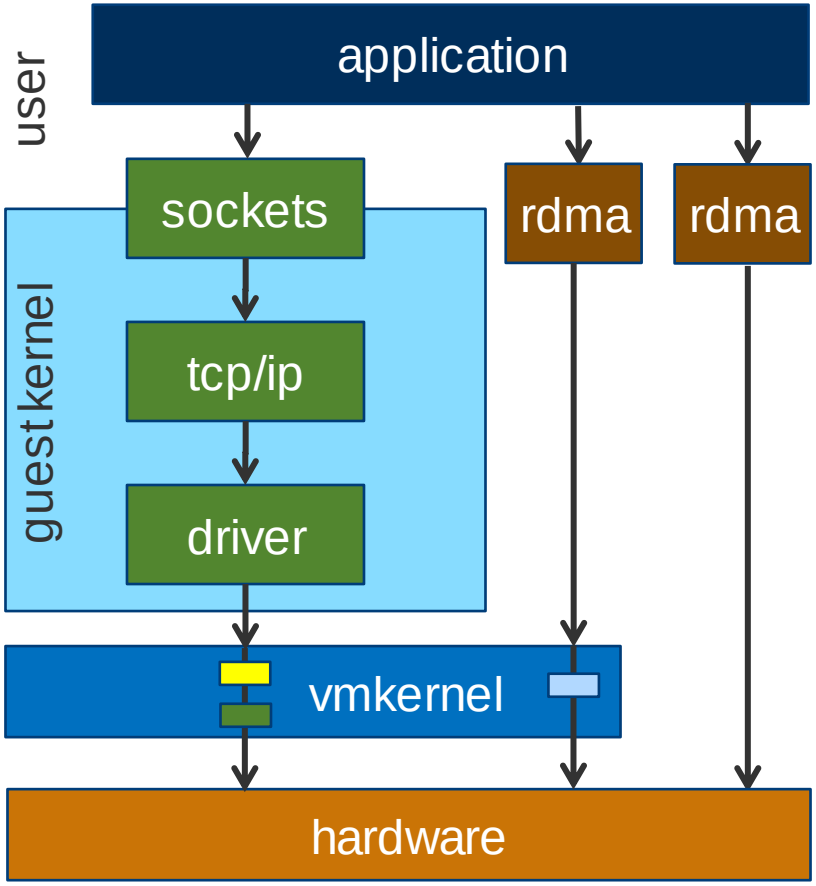
HPCC Virtual to Native Run-time Ratios (Lower is Better)



Data courtesy of:
Marco Righini
Intel Italy

Next Steps

- Evaluate FLUENT and HPCC on vSphere 5.x, QDR InfiniBand
- Continue to drive latency improvements into the vSphere platform
- Build a vRDMA prototype this summer



vRDMA

References

- **RDMA Performance in Virtual Machines with QDR InfiniBand on vSphere 5**
  - http://labs.vmware.com/publications/ib-researchnote-apr2012
- **Best Practices for Performance Tuning of Latency-Sensitive Workloads in vSphere VMs**
  - http://www.vmware.com/files/pdf/techpaper/VMW-Tuning-Latency-Sensitive-Workloads.pdf
- **VMware HPC Blog**
  - http://cto.vmware.com/tag/hpc/