Ultra-Low Latency on vSphere with RDMA

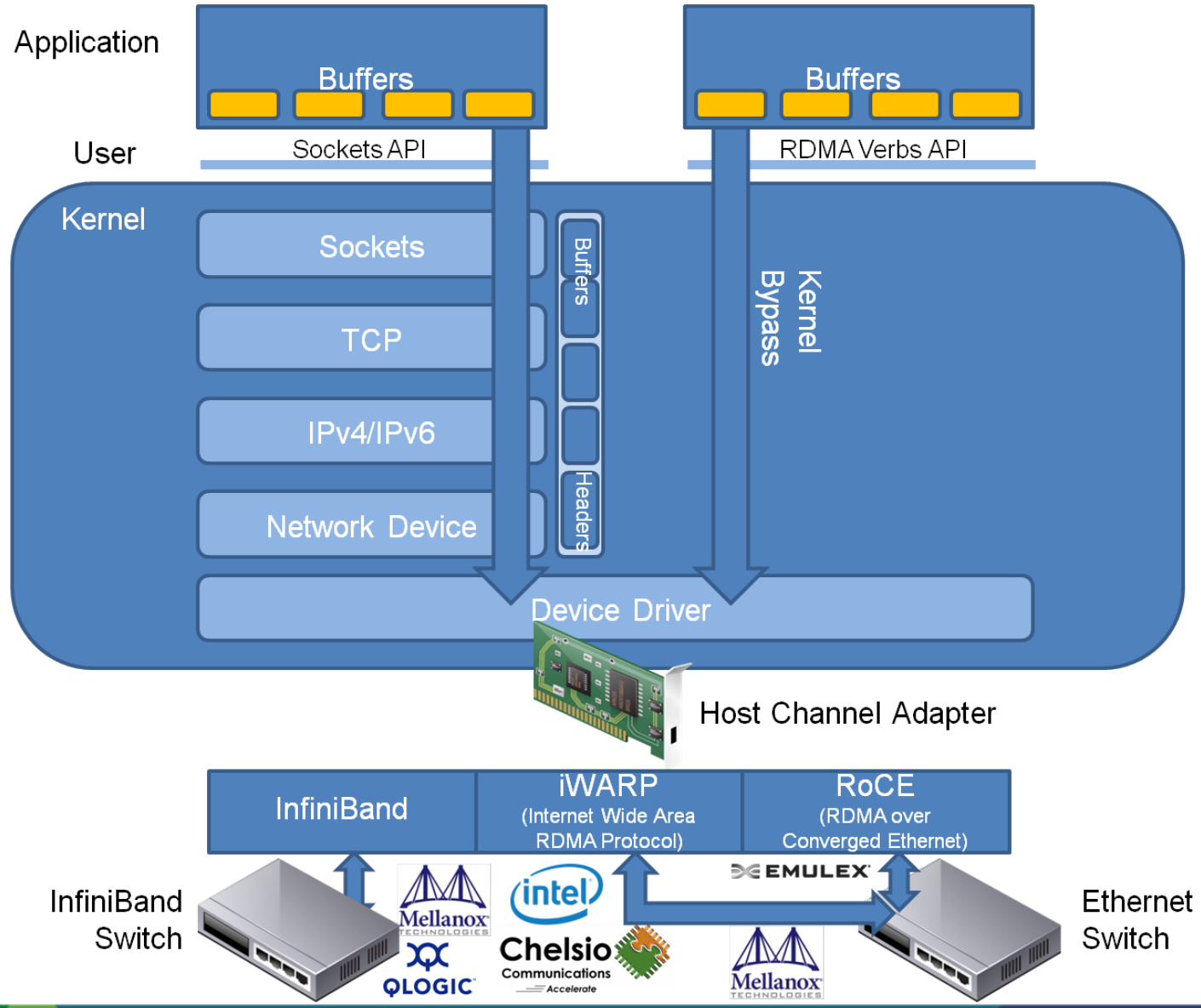
Bhavesh Davda, Office of CTO

VMworld 2012, August 28th, 2012

Agenda

- RDMA primer
- RDMA based applications
- Options for RDMA access in vSphere virtual machines
- RDMA advantages for vSphere hypervisor services

RDMA primer



vmware

Why RDMA: performance

- **Ultra low latency**
  - < 1 microsecond one-way for small messages with Mellanox CX3 FDR HCA in bare-metal (non-virtualized) environment
- **High throughput**
  - > 50 Gbps one-way for large messages with Mellanox CX3 PCIe3 FDR HCA in bare-metal (non-virtualized) environment
- **CPU efficiency**
  - Offloads CPU of running any protocol stack in software, freeing up the CPU to either lower power consumption or run other useful tasks

Who uses RDMA: bare-metal applications

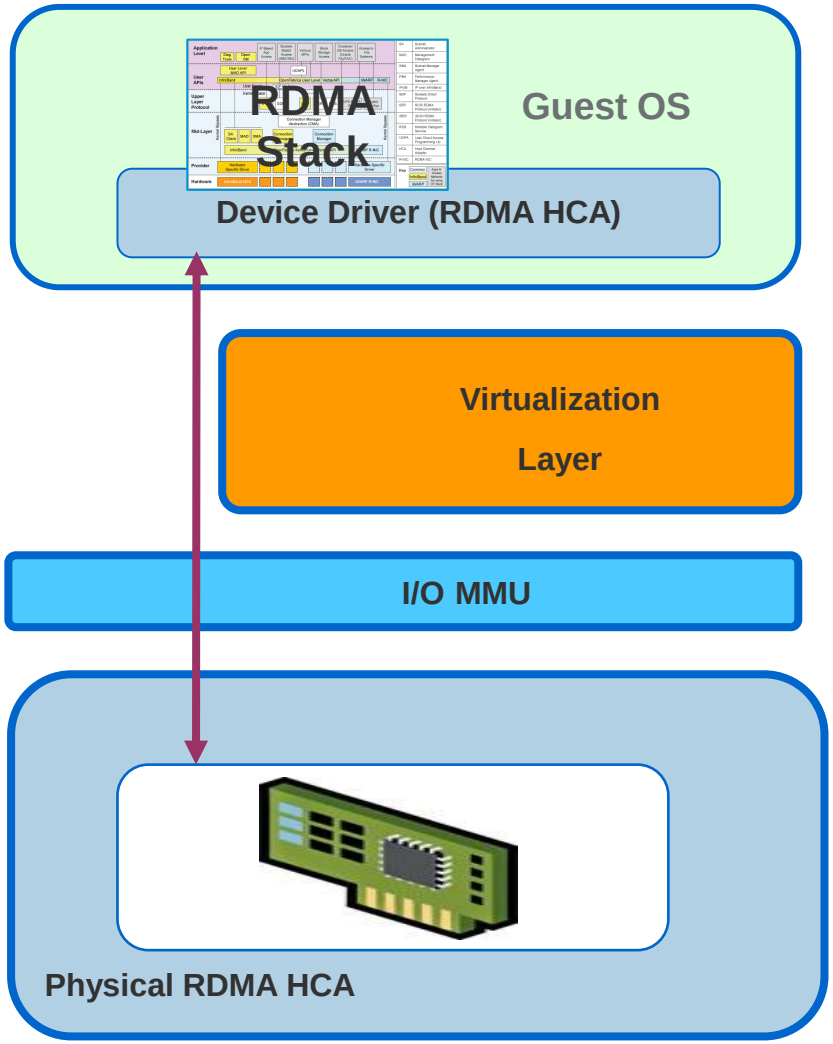
- **High performance computing applications relying on various message passing libraries like MPI**
- **Clustered databases**
  - IBM DB2 pureScale
  - Oracle ExaData/RAC
- **Distributed filesystems**
  - IBM GPFS
  - Open source Lustre
  - Red Hat Storage Server (GlusterFS)
- **Distributed caches**
  - Dell (RNA Networks)
- **Financial trading applications**
- **BigData (Hadoop: Mellanox Unstructured Data Accelerator)**

Options for RDMA access to vSphere virtual machines

1. Full-function DirectPath I/O (passthrough)
2. SR-IOV VF DirectPath I/O (passthrough)
3. Paravirtual RDMA HCA (vRDMA) offered to VM

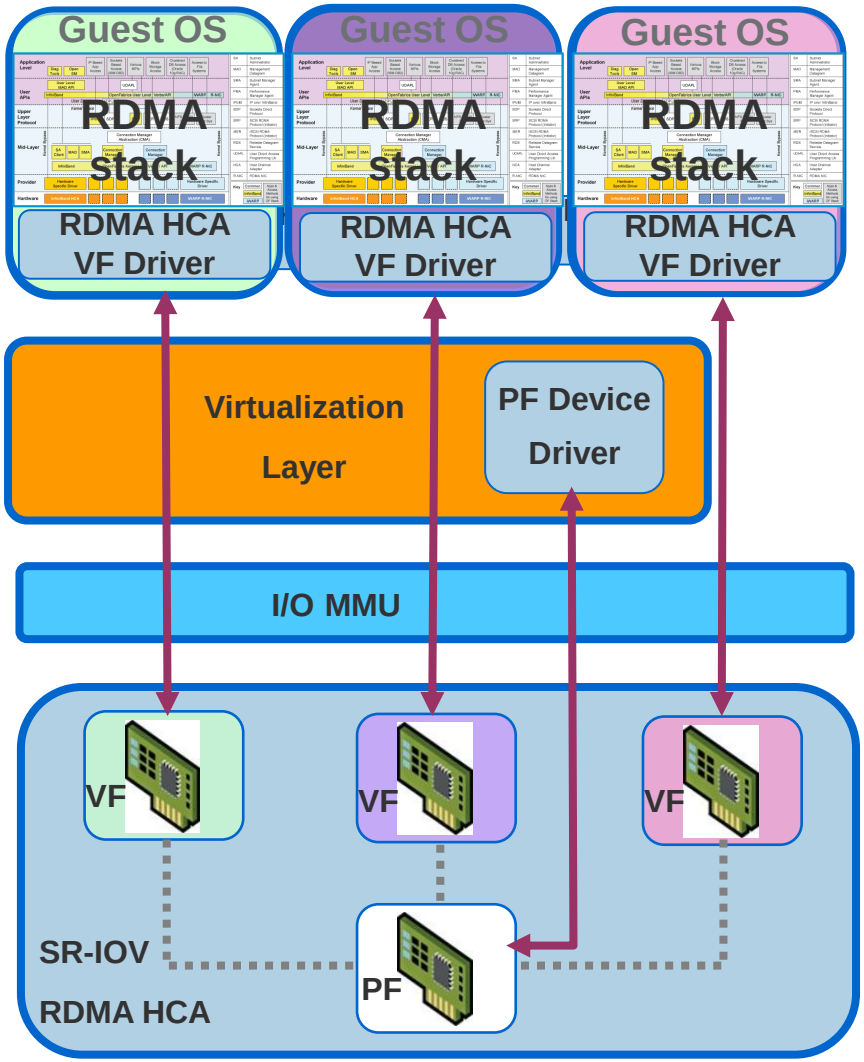
Full-function DirectPath I/O

- **Direct assign physical RDMA HCA (IB/RoCE/iWARP) to VM**
- **Physical HCA cannot be shared between VMs or by the ESXi hypervisor**
- **DirectPath I/O is incompatible with many important vSphere features:**
  - Memory Overcommit, Fault Tolerance, Snapshots, Suspend/Resume, vMotion



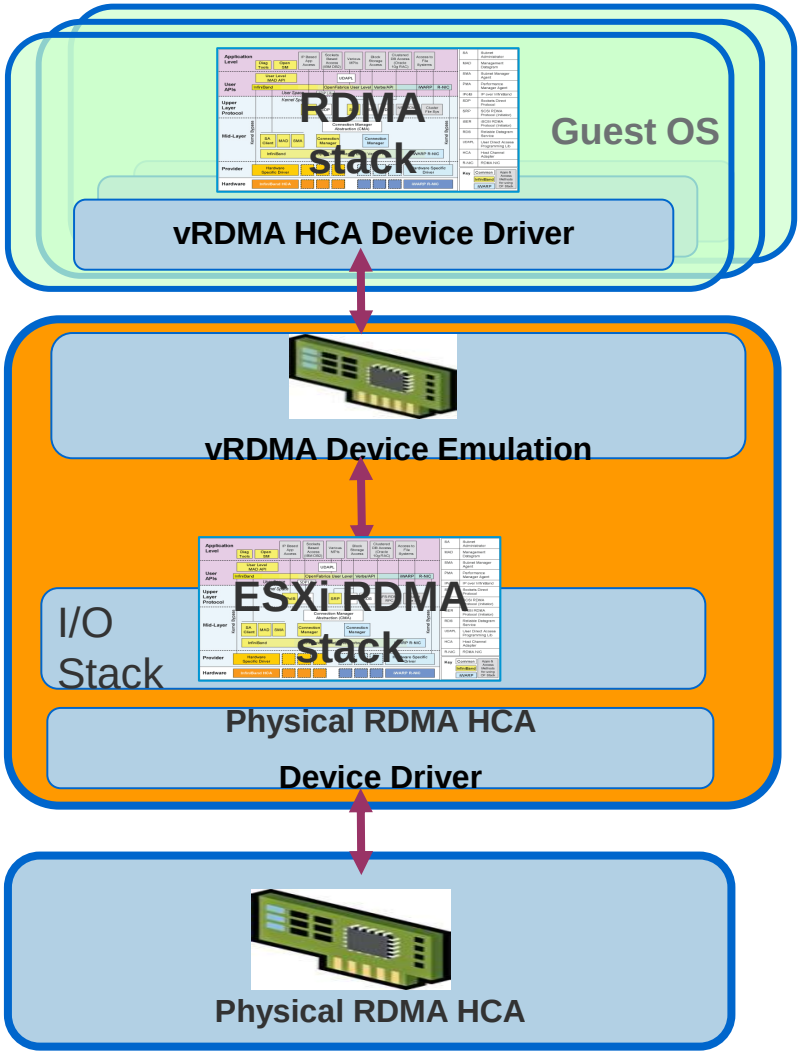
SR-IOV VF DirectPath I/O

- **Single-Root IO Virtualization (SR-IOV): PCI-SIG standard**
- **Physical (IB/RoCE/iWARP) HCA can be shared between VMs or by the ESXi hypervisor**
  - Virtual Functions direct assigned to VMs
  - Physical Function controlled by hypervisor
- **Still DirectPath I/O, which is incompatible with many important vSphere features**

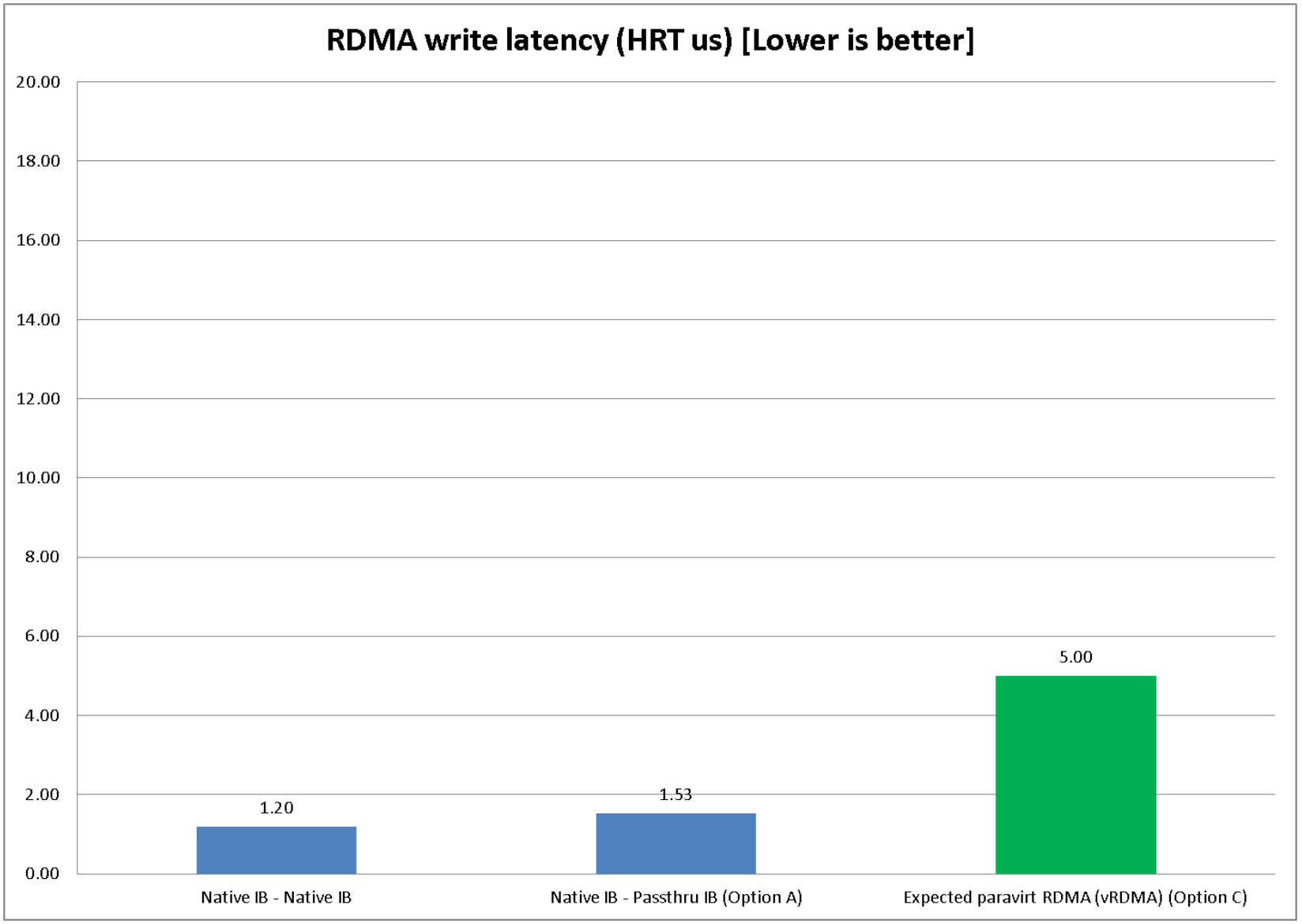


Paravirtual RDMA HCA (vRDMA) offered to VM

- **New paravirtualized driver in Guest OS**
  - Implements “Verbs” interface
- **RDMA emulated in ESXi hypervisor**
  - Translates Verbs from Guest to Verbs to ESXi “RDMA Stack”
  - Guest physical memory regions mapped to ESXi and passed down to physical RDMA HCA
  - Zero-copy DMA directly from/to guest physical memory
  - Completions/interrupts “proxied” by emulation



Performance Comparison



vmware

InfiniBand with DirectPath I/O

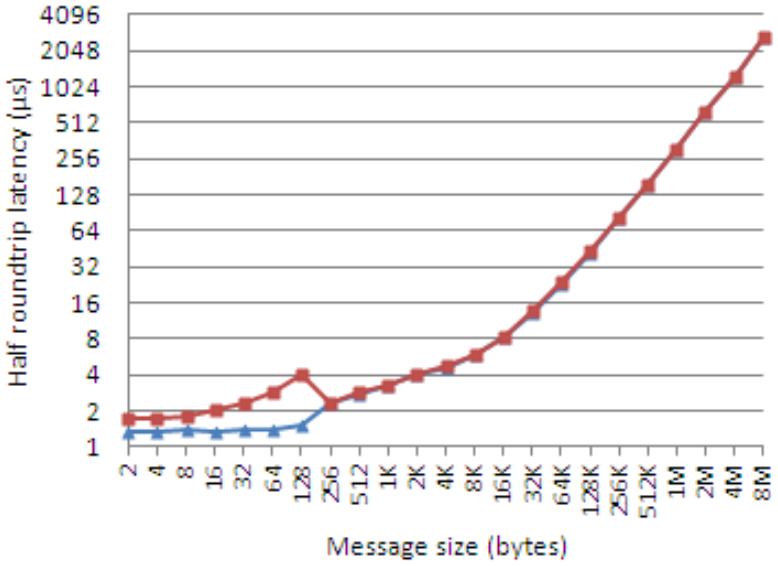


Figure 4: Send latencies using polling completions

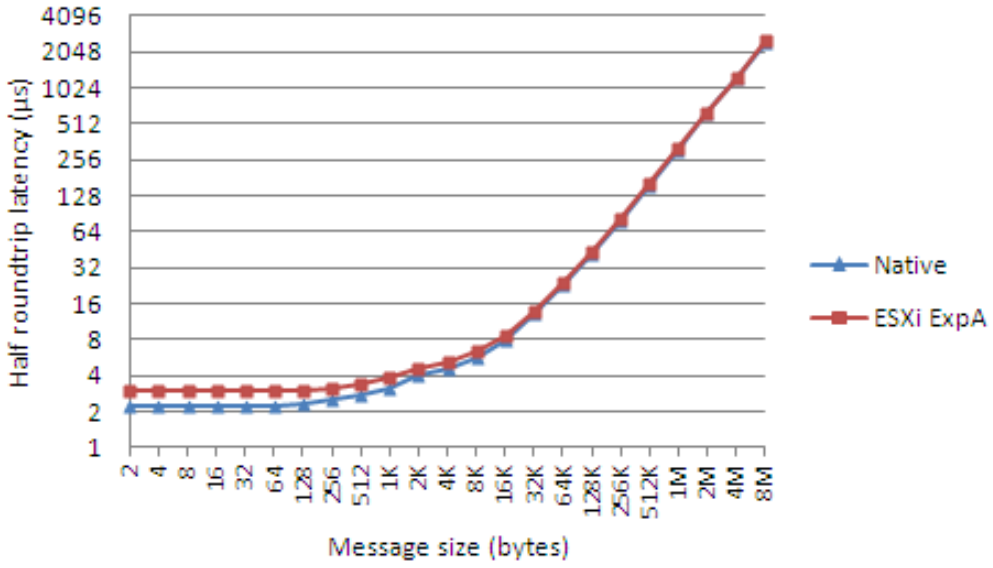


Figure 5: RDMA Read latencies using polling completions

InfiniBand with DirectPath I/O

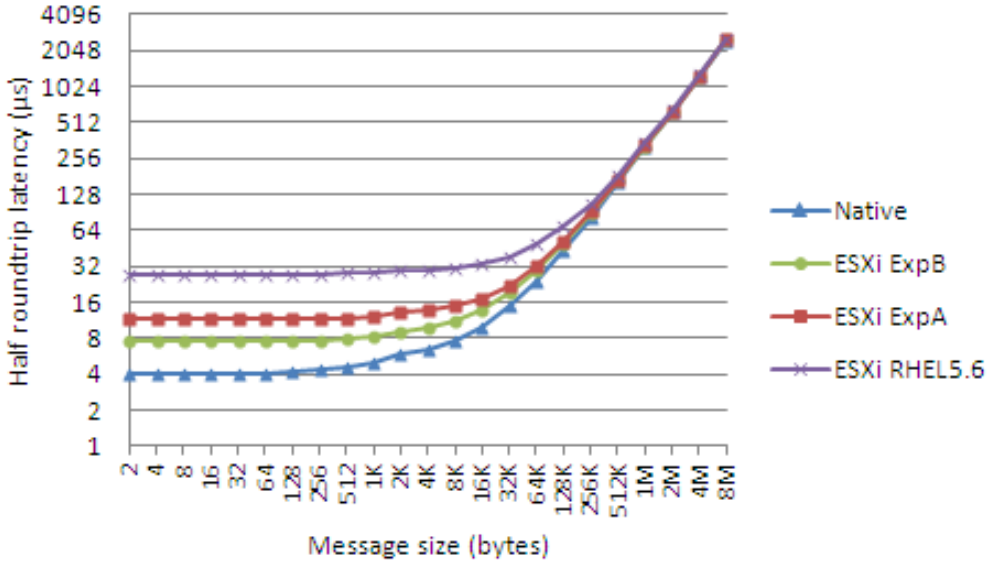


Figure 6: RDMA Read latencies using interrupt completions

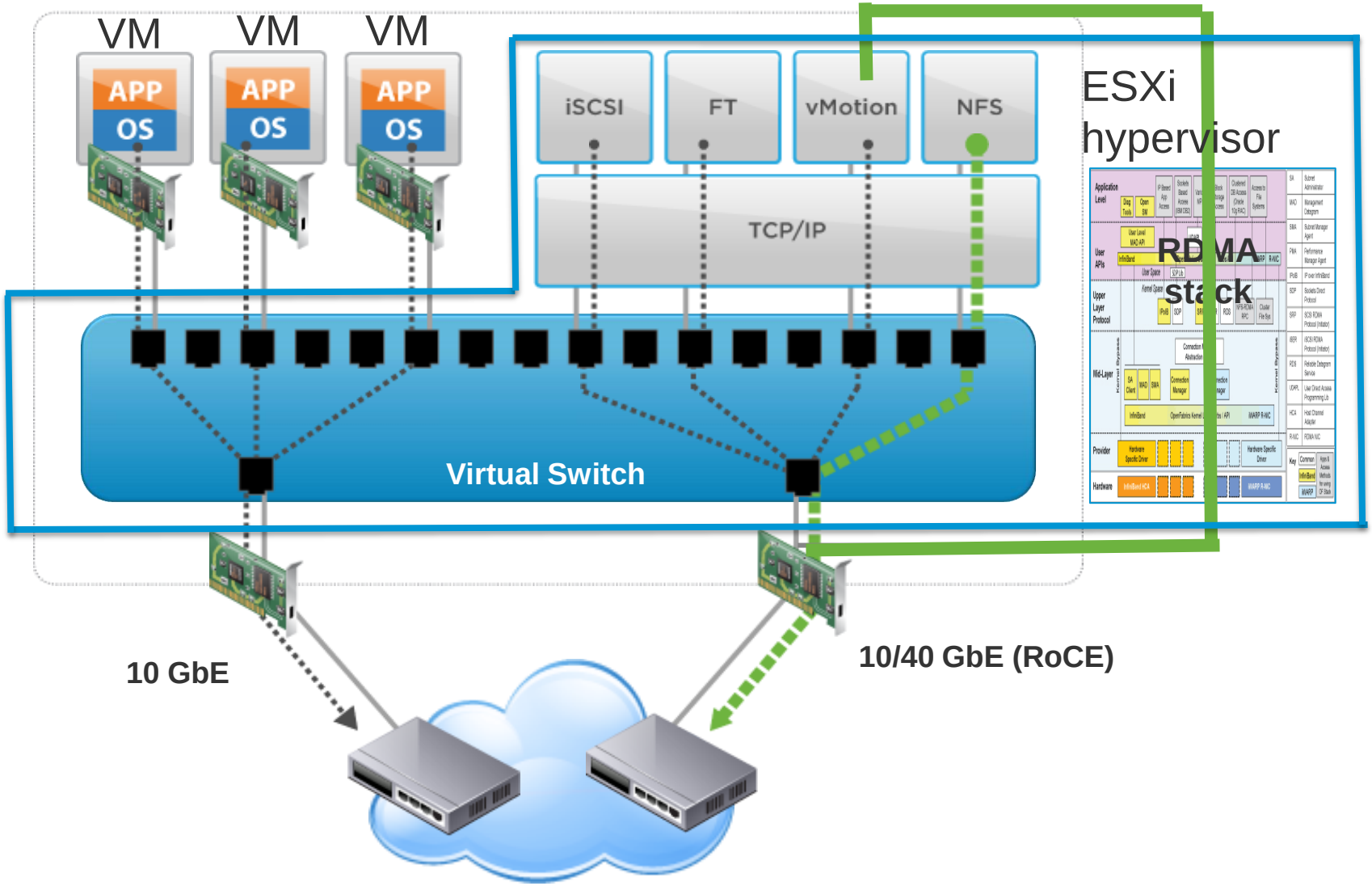
RDMA Performance in Virtual Machines with QDR InfiniBand on vSphere 5

http://labs.vmware.com/publications/ib-researchnote-apr2012

Summary: VM/Guest level RDMA

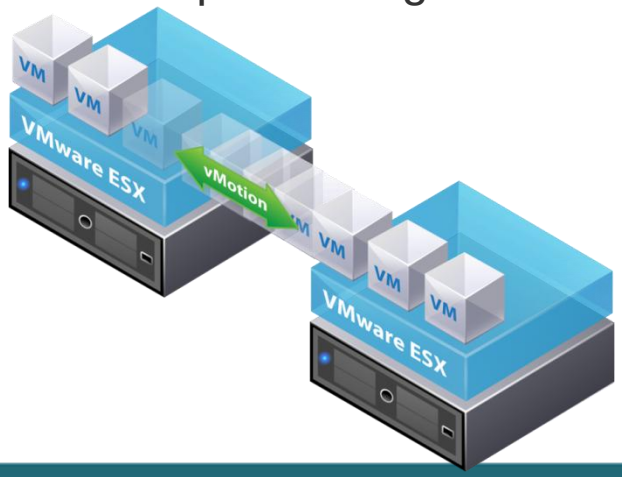
- Full function DirectPath I/O already used by some customers
- SR-IOV DirectPath I/O supported in vSphere 5.1
- vRDMA paravirtual driver/device being prototyped in Office of CTO

RDMA advantages for vSphere hypervisor services



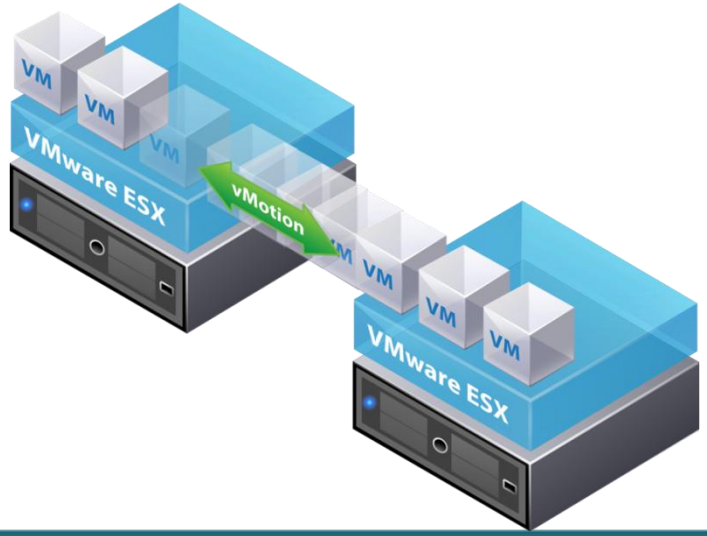
vMotion using RDMA transport

- **vMotion: live migration of virtual machines**
  - Key enabler for Distributed Resource Scheduler (DRS) and Distributed Power Management (DPM)
- **vMotion data transfer mechanism**
  - Currently network (TCP/IP/BSD sockets/mbufs) based
- **Why RDMA? Performance**
  - Zero copy send & receive == Reduced CPU utilization
  - Lower latency
  - Eliminate TCP/IP protocol & processing overheads

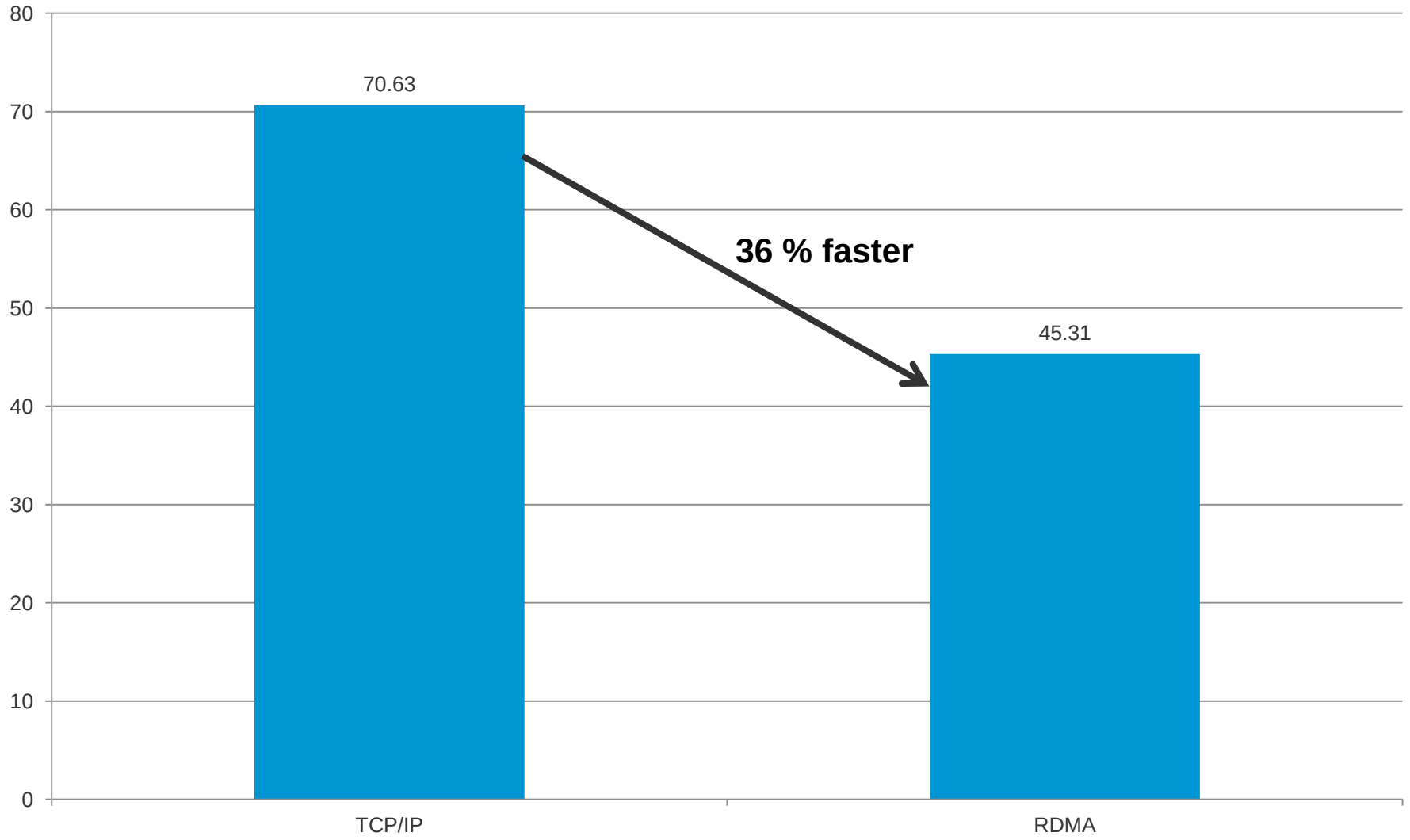


Test setup

- **Two HP ProLiant ML 350 G6 machines, 2x Intel Xeon (E5520, E5620), HT enabled, 60 GB RAM**
- **Mellanox 40GbE RoCE cards**
  - ConnectX-2 VPI PCIe 2.0 x8, 5.0 GT/s
- **SPECjbb2005 50GB workload**
  - 56 GB, 4 vCPU Linux VM
  - Run-time config switch for TCP/IP or RDMA transport
  - Single stream helper for RDMA transport vs. two for TCP/IP transport

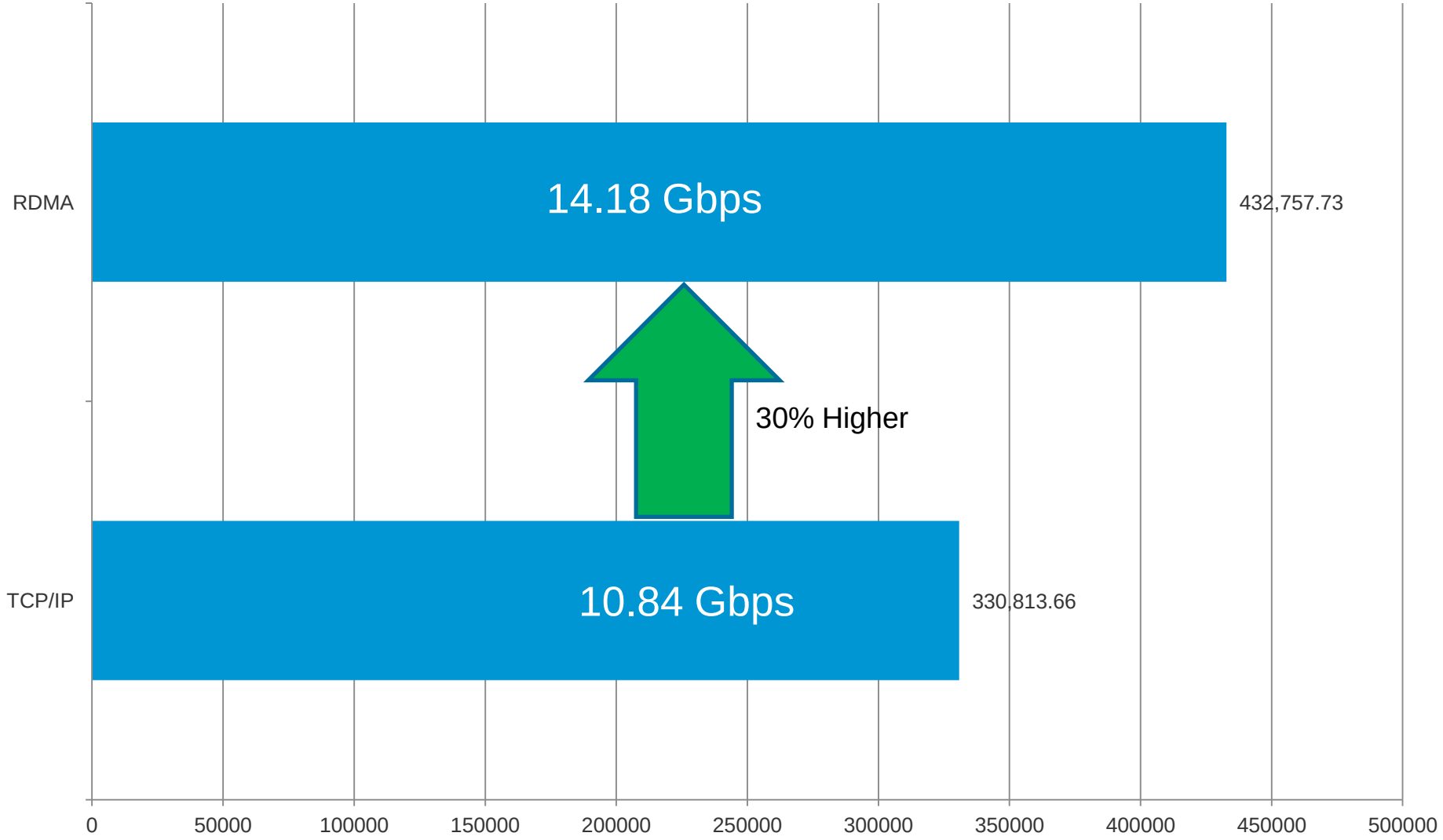


Total vMotion Time (seconds)



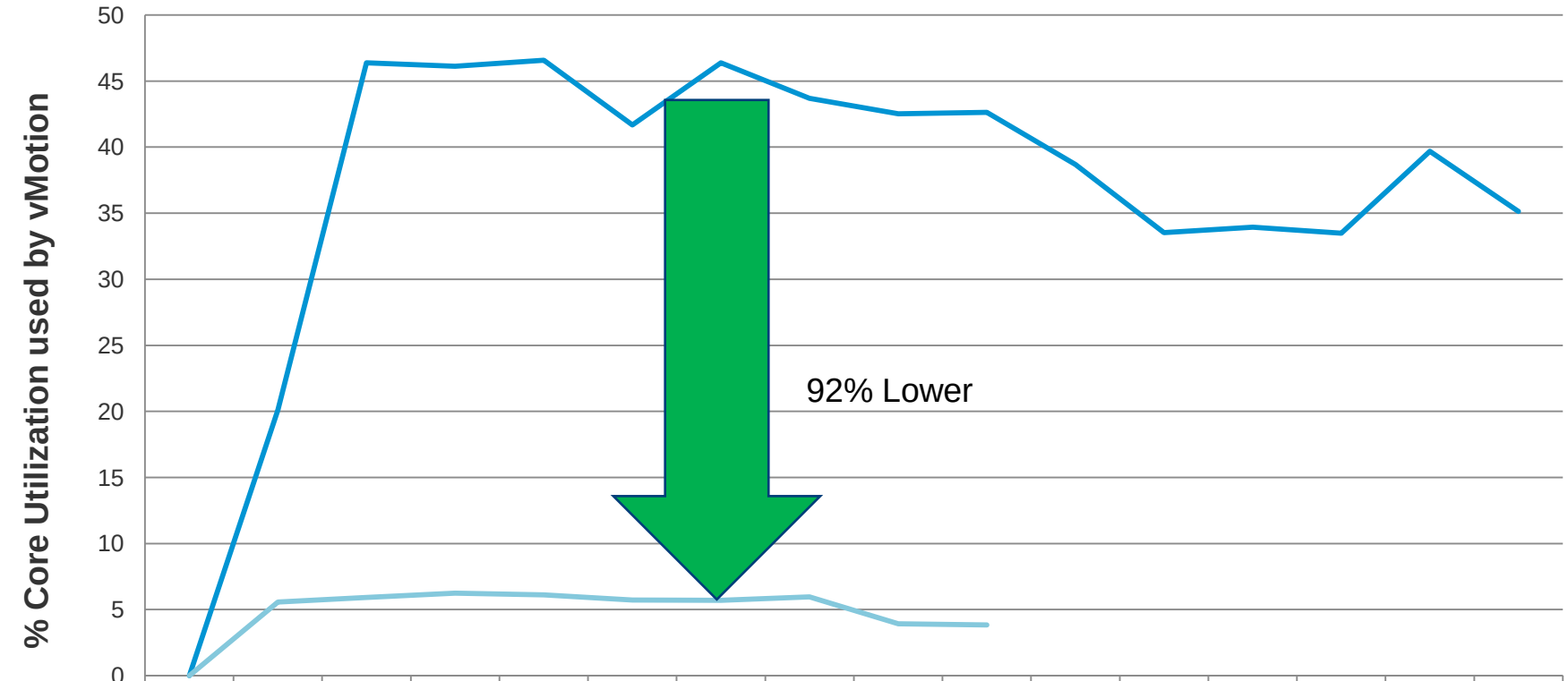
vmware

Precopy bandwidth (Pages/sec)



vmware

CPU Utilization



	0:00	0:05	0:11	0:16	0:21	0:26	0:31	0:36	0:41	0:46	0:51	0:56	1:01	1:06	1:11	1:16
TCP/IP	0	20.11	46.38	46.13	46.58	41.68	46.38	43.69	42.53	42.64	38.69	33.53	33.93	33.49	39.68	35.13
RDMA	0	5.58	5.92	6.24	6.11	5.73	5.71	5.96	3.92	3.83						

- **Destination CPU utilization 92% lower**
- **Source CPU utilization 84% lower**

Overall summary

- **Hypervisor-level RDMA**
  - Proven benefits for hypervisor services (vMotion, etc.)
  - Currently under discussion internally
- **Passthrough InfiniBand delivers credible performance**
- **Paravirtual vRDMA most attractive option for VM-level RDMA**
  - Maintains key virtualization capabilities
  - Delivers better latencies than other approaches
  - Prototyping underway

Acknowledgements

- **VMware**
  - Adit Ranadive (Intern), Anupam Chanda, Gabe Tarasuk-Levin, Josh Simons, Margaret Petrus, Scott Goldman
- **Mellanox**
  - Ali Ayoub, Dror Goldenberg, Gilad Shainer, Liran Liss, Michael Kagan, Motti Beck
- **System Fabric Works:**
  - Bob Pearson, Paul Grun