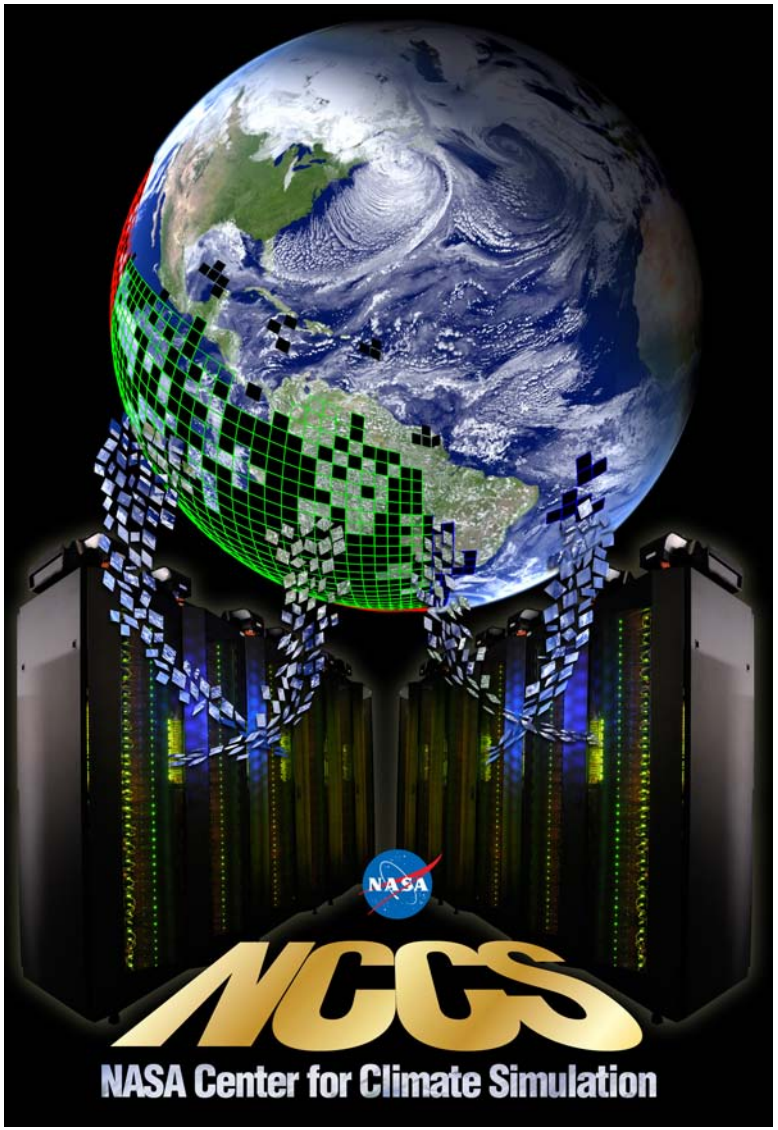


SR-IOV In High Performance Computing

Hoot Thompson & Dan Duffy

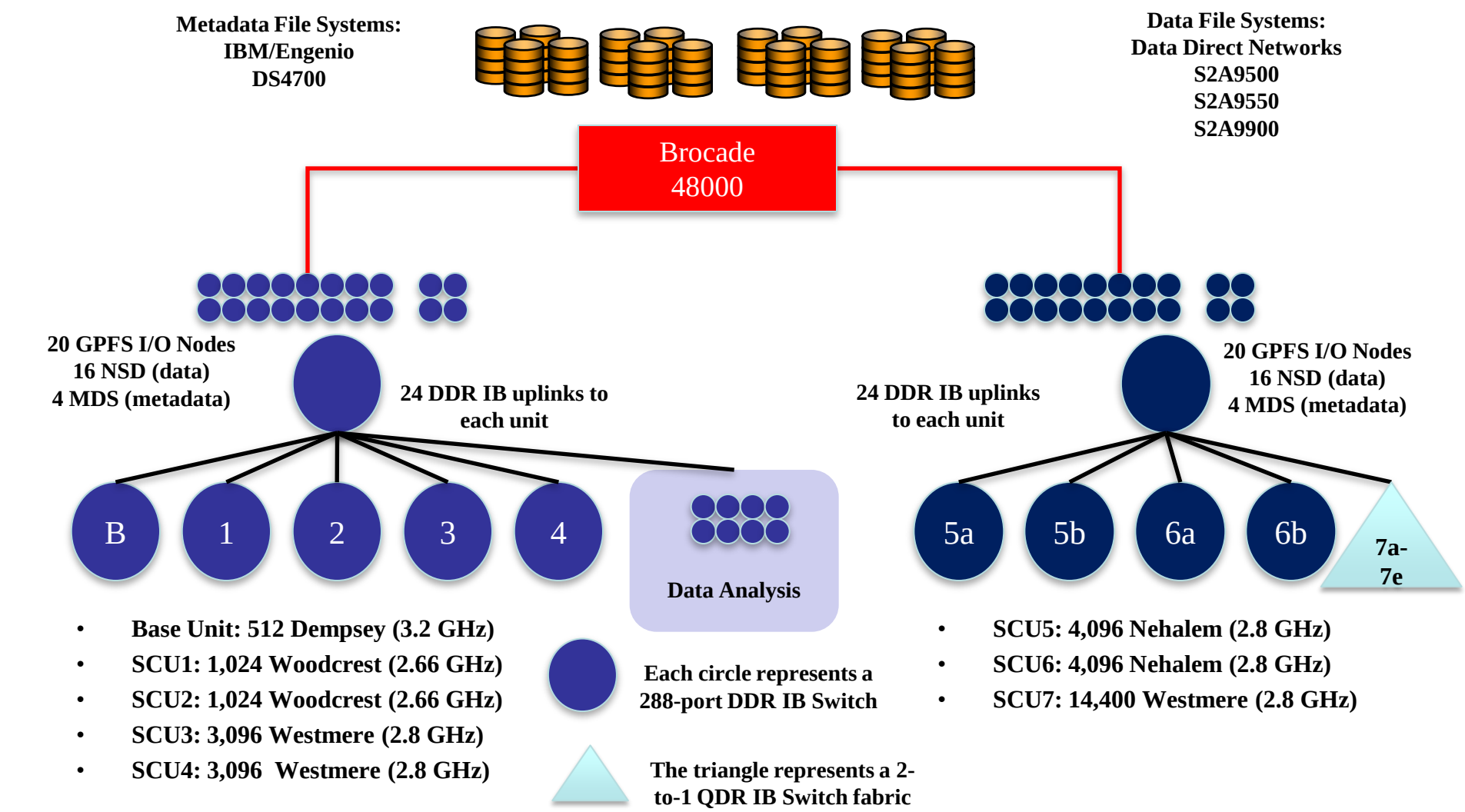
NASA Center for Climate Simulation
NASA Goddard Space Flight Center
Greenbelt, MD 20771
hoot@ptpnow.com
daniel.q.duffy@nasa.gov

www.nccs.nasa.gov

NASA Center for Climate Simulation

- Focus on the research side of climate study (versus NOAA's operational position)
- Simulations span multiple time scales
  - Days for weather prediction
  - Seasons to years for short term climate prediction
  - Centuries for climate change projection
- Examples:
  - High fidelity 3.5 KM global simulations of cloud and hurricane predictions
  - Comprehensive reanalysis of the last thirty years of weather/climate –MERRA
  - Multi-millennium analysis for the Intergovernmental Panel on Climate Change
- Integrated set of supercomputing, visualization and data management technologies
  - Discover computational cluster
    - 30K traditional Intel cores plus 64 GPUs, roughly 400 TFlops
    - DDR/QDR Infiniband (IB) backbone
    - 1 GbE and 10 GbE management infrastructure
    - ~4 PBytes RAID based shared parallel file system (GPFS)
  - Tape archive of over 20 PBytes

Discover IB/GPFS Architecture



- Base Unit: 512 Dempsey (3.2 GHz)
- SCU1: 1,024 Woodcrest (2.66 GHz)
- SCU2: 1,024 Woodcrest (2.66 GHz)
- SCU3: 3,096 Westmere (2.8 GHz)
- SCU4: 3,096 Westmere (2.8 GHz)

Each circle represents a 288-port DDR IB Switch

The triangle represents a 2-to-1 QDR IB Switch fabric

- SCU5: 4,096 Nehalem (2.8 GHz)
- SCU6: 4,096 Nehalem (2.8 GHz)
- SCU7: 14,400 Westmere (2.8 GHz)

Nebula – NASA's Cloud

- Open-source (OpenStack) cloud computing project and service
- Alternative to costly construction of additional data centers
- Sharing portal for NASA scientists and researchers
  - Large, complex data sets
  - External partners and the public.
- Nebula comprised of two components/containers
  - Nebula west at NASA AMES
  - Nebula east at NASA GSFC

- NCCS team evaluating Nebula as adjunct to Discover hosted science processing
- Key question can clouds match HPC level of capability needed for climate research
- Potential obstacle – clouds primarily exist in virtualized space
  - Overhead or loss due to virtual machine (VM) versus bare metal
  - Node-to-node communication critical – high speed, low latency, RDMA

http://nebula.nasa.gov/

Background And Proposition

- Background
  - Discover's performance tied to it's DDR/QDR IB fabric
  - Nebula, clouds in general, 10 GE based
- Question – can clouds deliver HPC level of performance?
  - Can 10GE compete with high speed, low latency IB?
  - What network performance is lost due to virtualization?
  - What computational performance is lost due to virtualization?
- Proposition – typical NCCS model
  - Build test bed to investigate the virtualization technologies
  - Work with vendors to answer questions and address issues

Methodology and Objectives

- Compare bare metal against virtualized NIC
  - Full software virtualization (SW Virt) – device emulation
  - Virtio – split driver, para-virtualization
  - Single Root IO Virtualization (SR-IOV)
    - Direct assignment
    - Mapped Virtual Function (VF)
- Determine overhead of executing within VM construct
  - VM to VM communication
    - Base Network
    - Message passing environment (mvapich2)
  - Application
    - Single node, multi-core
    - Multi-node, multi-core
- Draw conclusions and comparisons with Discover and Nebula

http://www.intel.com/content/www/us/en/pci-express/pci-sig-sr-iov-primer-sr-iov-technology-paper.html

Benchmarks

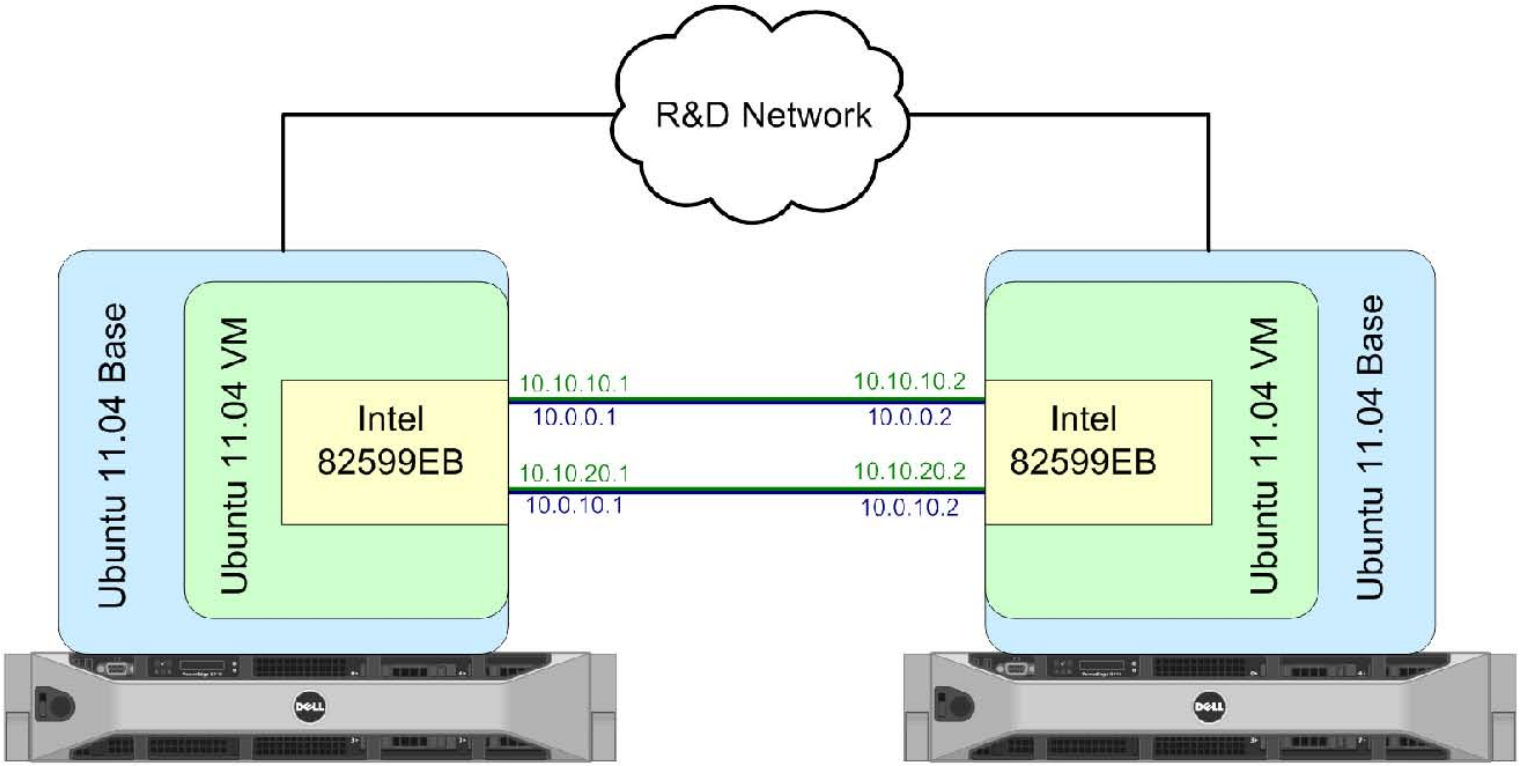
Benchmark	Version	Description	Download
Nuttcp	nuttcp-7.1.5.c gcc compiler	Measure raw network bandwidth, similar to netperf:	http://lcp.nrl.navy.mil/nuttcp
OSU MPI Benchmarks	MVAPICH2 1.7rc1 Intel compiler	Test latencies and bandwidths of most common MPI functions.	http://mvapich.cse.ohio-state.edu/
Linpack	10.2.6 Intel compiler	Intel version of Linpack	http://software.intel.com/en-us/articles/intel-math-kernel-library-linpack-download/
NAS PB	3.3.1 Intel compiler	NASA Parallel Benchmarks; CFD kernel benchmarks	http://www.nas.nasa.gov/Resources/Software/npb.html

Started from the basic benchmarks to analyze system performance and build up towards the application layer

System Configurations

Configuration	Bare1	Bare2	VM1	VM2
Processor Type	Intel Nehalem	Intel Nehalem	Intel Nehalem	Intel Nehalem
Processor Number	E5520	E5520	E5520	E5520
Processor Speed	2.27 GHz	2.27 GHz	2.27 GHz	2.27 GHz
Cores per Socket	4	4	4	4
Number of Sockets	2	2	2	2
Cores per Node	8	8	8	8
Theoretical Peak	72.64 GF	72.64 GF	72.64 GF	72.64 GF
Main Memory	48 GB	48 GB	16 GB	16 GB
Operating System	Ubuntu 11.04	Ubuntu 11.04	Ubuntu 11.04	Ubuntu 11.04
Kernel	2.6.38-10.server	2.6.38-10.serve	2.6.38-10.server	2.6.38-10.server
Hypervisor	KVM	KVM	N/A	N/A
Hyperthreading	Off	Off	Off	Off

Test Configuration



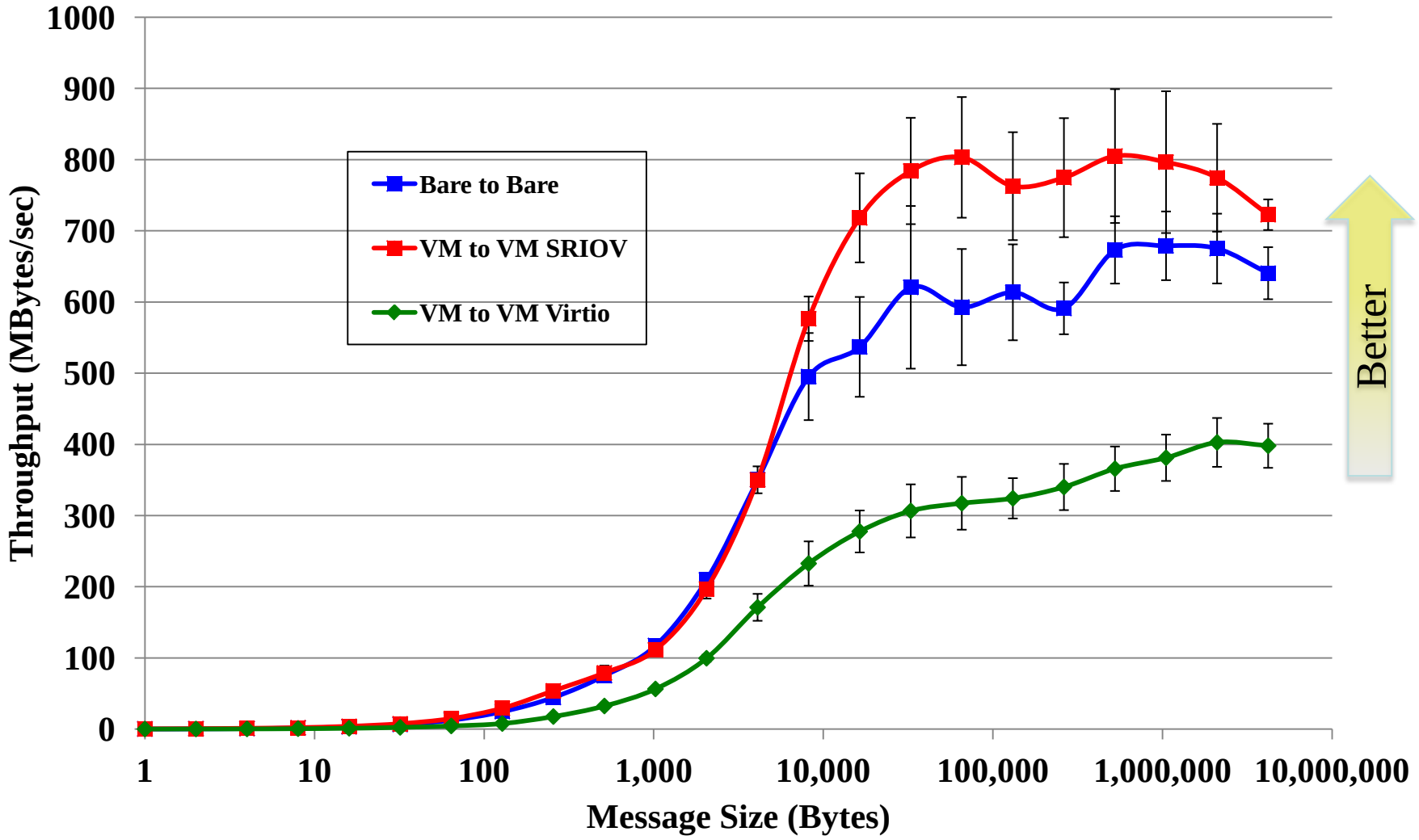
Dell R710
→ E5520 2.27GHz (X2)
→ 48GB

Dell R710
→ E5520 2.27GHz (X2)
→ 48GB

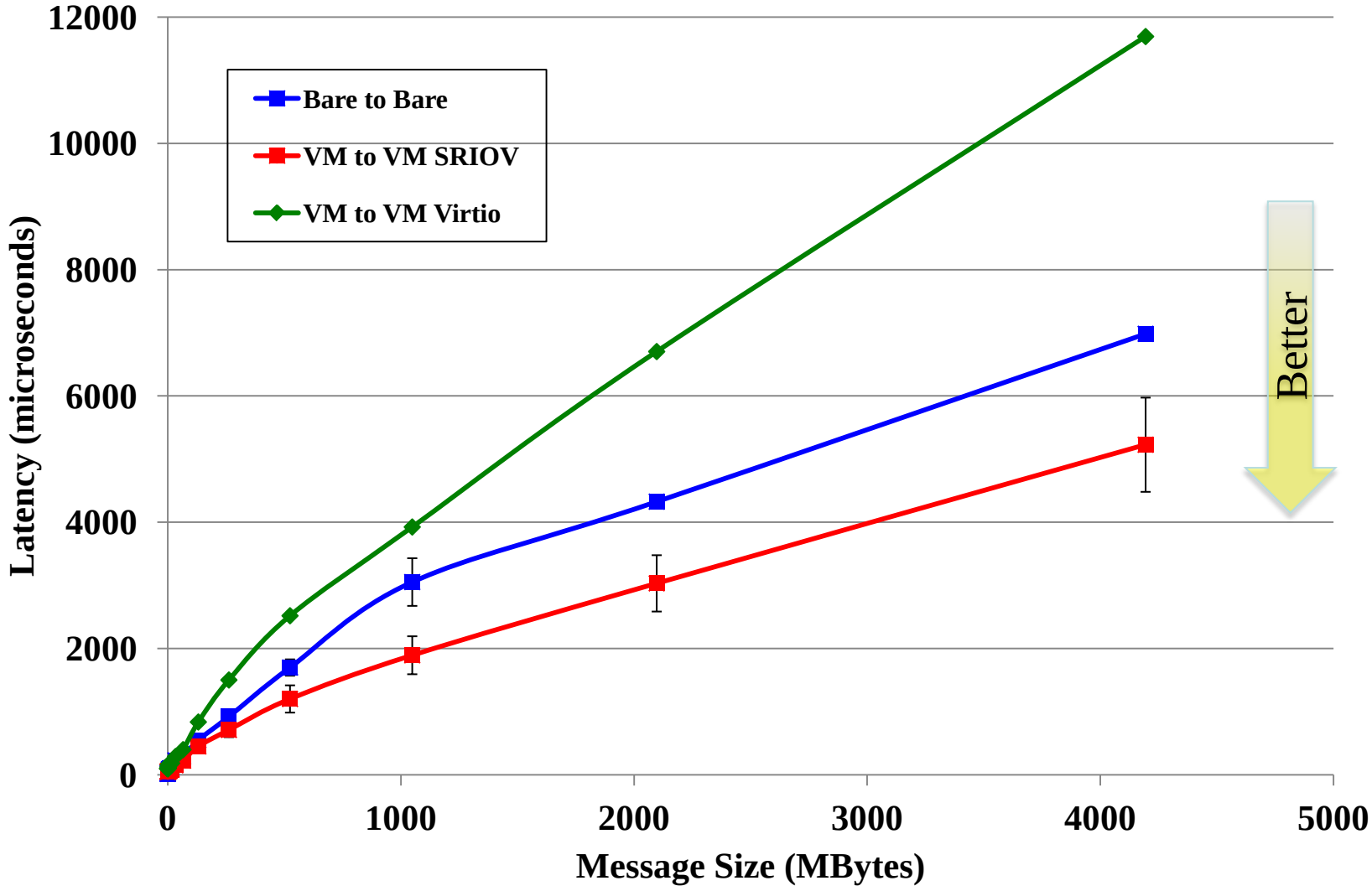
Nuttcp Results

Bare to Bare		VM to VM Sw Virt		VM to VM Virtio		VM to VM SR-IOV	
4418.8401 Mbps	0 retrans	137.3301 Mbps	0 retrans	5864.0557 Mbps	212 retrans	9151.5769 Mbps	0 retrans
8028.6459 Mbps	0 retrans	145.6024 Mbps	0 retrans	5678.0625 Mbps	0 retrans	9408.0323 Mbps	0 retrans
9392.7072 Mbps	0 retrans	145.7500 Mbps	0 retrans	5973.2256 Mbps	0 retrans	8714.4063 Mbps	34 retrans
9415.2675 Mbps	0 retrans	138.5963 Mbps	0 retrans	6309.8478 Mbps	0 retrans	9313.8894 Mbps	7 retrans
9341.4362 Mbps	733 retrans	141.8702 Mbps	0 retrans	6223.4034 Mbps	7 retrans	9251.8453 Mbps	0 retrans
9354.0999 Mbps	208 retrans	146.1092 Mbps	0 retrans	6311.3896 Mbps	0 retrans	9193.1103 Mbps	0 retrans
9414.7318 Mbps	0 retrans	146.3042 Mbps	0 retrans	6316.7924 Mbps	0 retrans	9348.2984 Mbps	0 retrans
9414.8207 Mbps	0 retrans	146.4449 Mbps	0 retrans	5955.8176 Mbps	0 retrans	9101.7356 Mbps	73 retrans
9414.9368 Mbps	0 retrans	146.2758 Mbps	0 retrans	5746.2926 Mbps	0 retrans	8958.5032 Mbps	16 retrans
9415.1618 Mbps	0 retrans	146.1043 Mbps	0 retrans	5692.8146 Mbps	0 retrans	9228.5370 Mbps	0 retrans

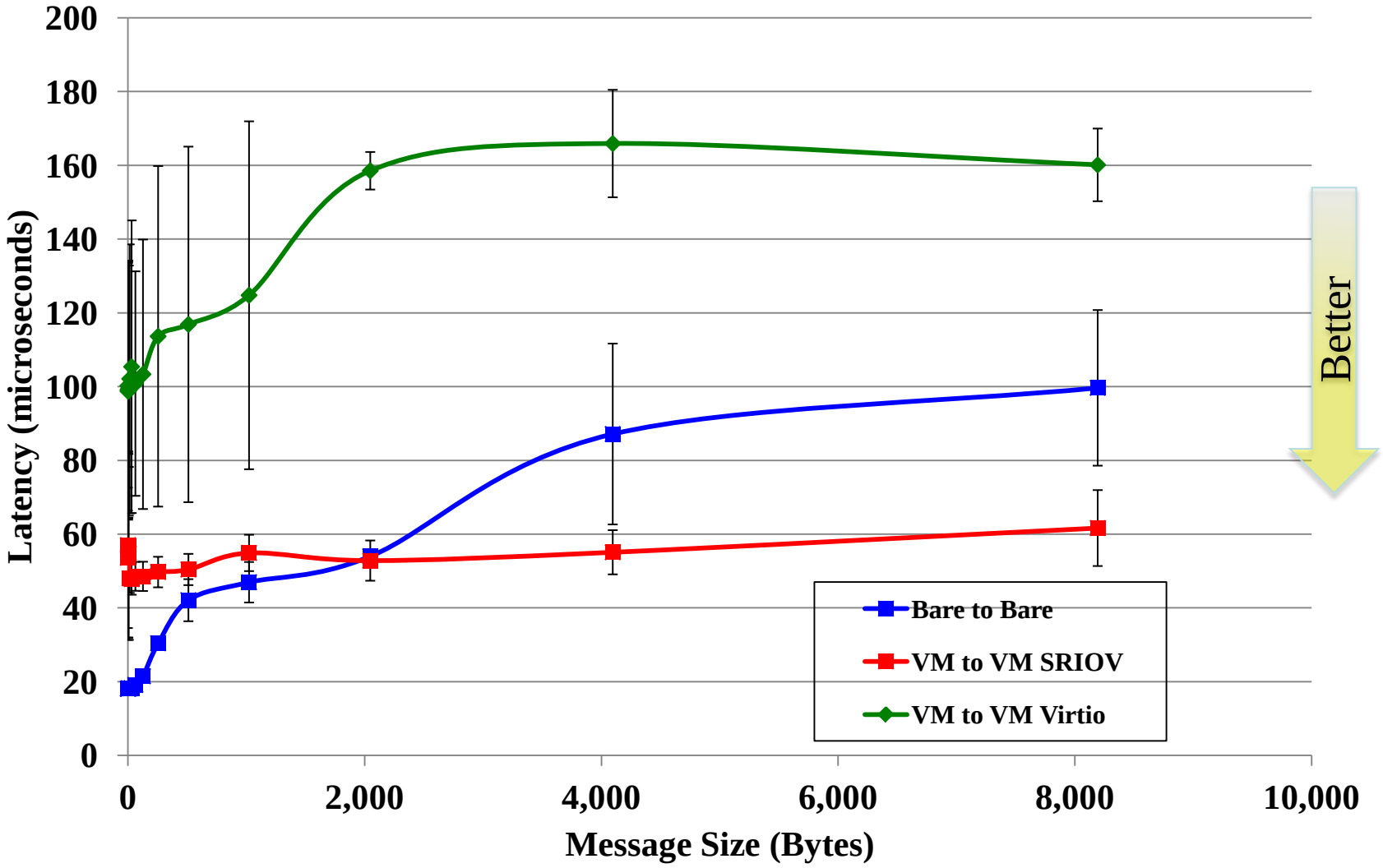
OSU Benchmarks Results – Bandwidth



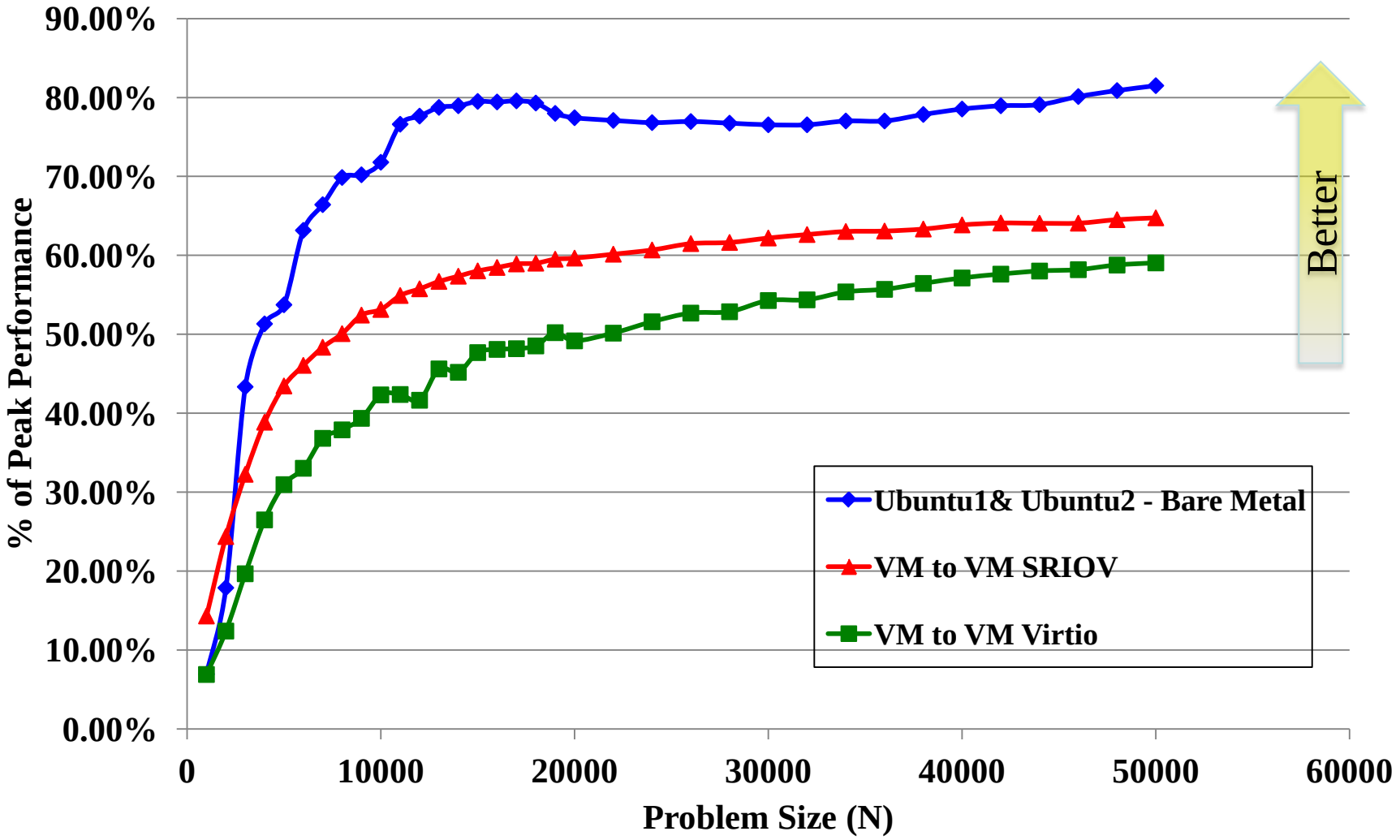
OSU Benchmarks Results – Latency



OSU Benchmarks Results – Latency (Small)



Linpack Benchmarks Results



Going Forward

- Conclusions to-date
  - Clear advantages to SR-IOV technology
  - Cloud based HPC feasible
  - Data requires further analysis to understand Nebula implications
- Issues/concerns
  - TCP Slow start, variability and retran impact on HPC processing
- Additional testing to close the gap
  - More application testing – NAS Parallel and HPCC benchmarks
  - Jumbo frames (9000 MTU)
  - Bare metal-to-bare metal and VM-to-VM IB
  - Different hypervisor – XEN
  - Other VM guest types – RedHat, SUSE
  - Multiple VMs running, bandwidth sharing
  - Add cloud infrastructure to test setup – Openstack, Eucalyptus