
National Aeronautics and Space Administration
NASA
Virtualized Infiniband:
Enabling HPC in the Cloud
Hoot Thompson
NASA Center for Climate Simulation (NCCS)
john.h.thompson@nasa.gov
www.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

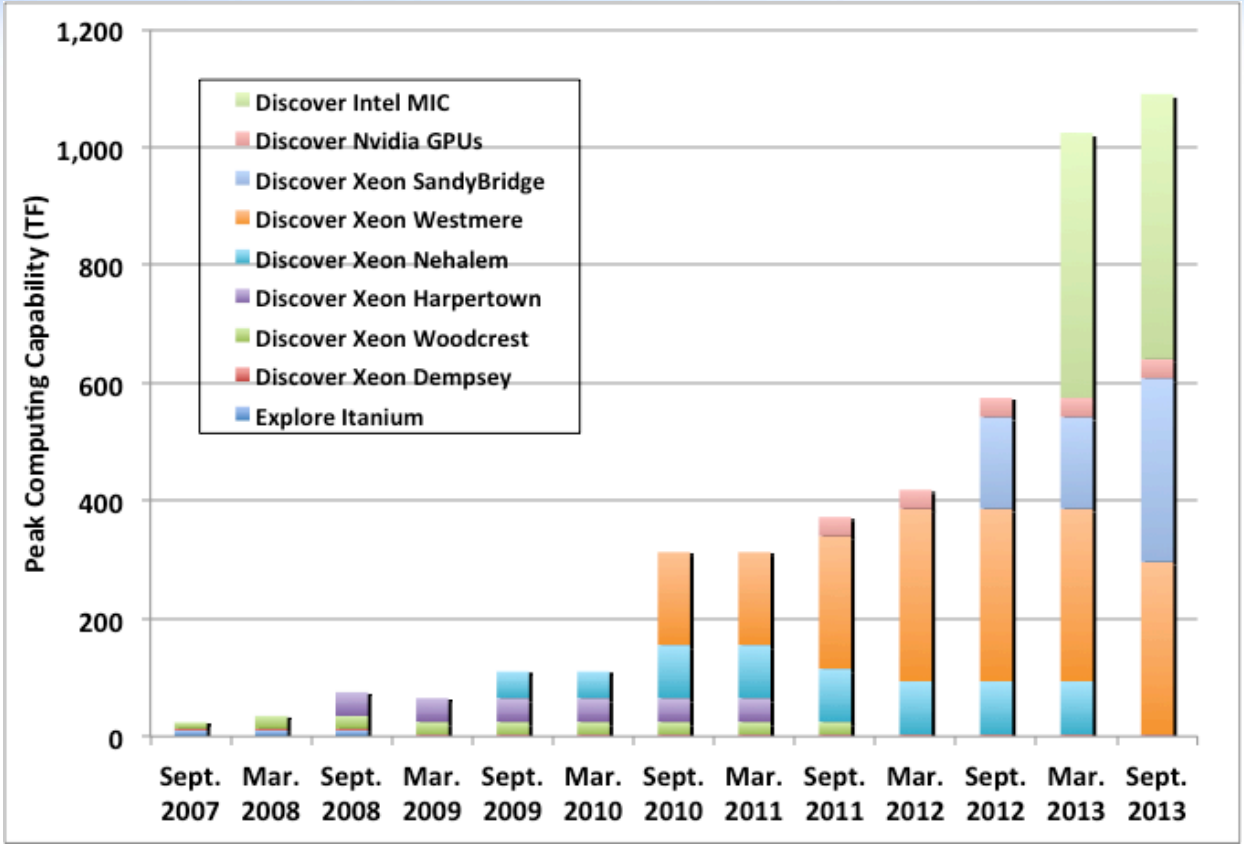
Examples:

- 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
  - » Mix traditional Intel cores, nVidia GPUs and Intel Xeon Phis
  - » DDR/QDR/FDR InfiniBand (IB) backbone
  - » 1 GbE and 10 GbE management infrastructure
  - » ~17 PBytes RAID based shared parallel file system (GPFS)
- Tape archive of over 30 PBytes

Discover Computational Growth



Objective: HPC Science Cloud

Adjunct to Discover hosted science processing

- Special/temporary debug queues
- Customized run-time environments
- Code validation against older system images

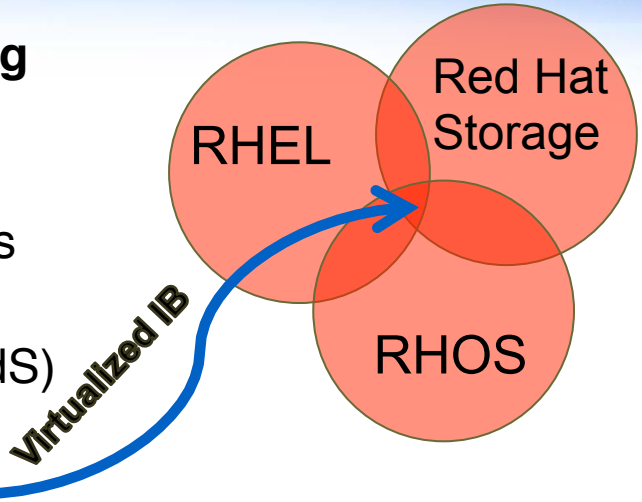
Expanded customer base

- Temporal processing campaigns (e.g. IFloodS)
- Mission support (e.g. SMAP)

Issue is matching HPC levels of performance

- Node-to-node communication critical – high speed, low latency, scalable
- Shared, high performance file system mandatory
- Management and rapid provisioning of resources – cluster formation

Potential obstacle – performance loss in virtualized space



RHEL Virtualized IB Test Bed

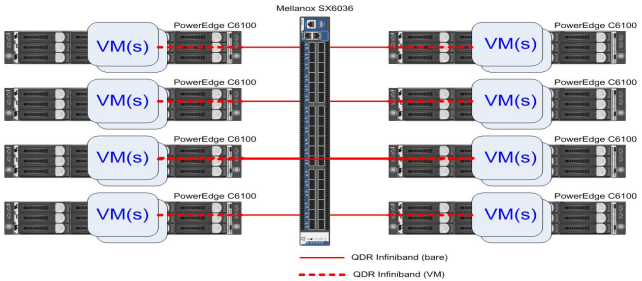
Set-up eight node POC environment – Westmere based

- Ran representative benchmarks
- Contrasted bare Host (KVM hypervisor) with VM (guest)

Benchmark	Description
Stream	Measures sustainable memory bandwidth and the corresponding computation rate.
OSU Micro-benchmarks	Measures performance of OpenSHMEM data movement and atomics operations.
LINPACK	Measures floating point performance by solving a set of linear equations.
NAS Parallel Benchmarks (NPB)	Mimic the computation and data movement in CFD applications.

Investigated multiple techniques for improving performance

- VM tuning – hugepages and NUMA awareness
- Virtualized IB using SR-IOV



Summarized Virtualized IB Results

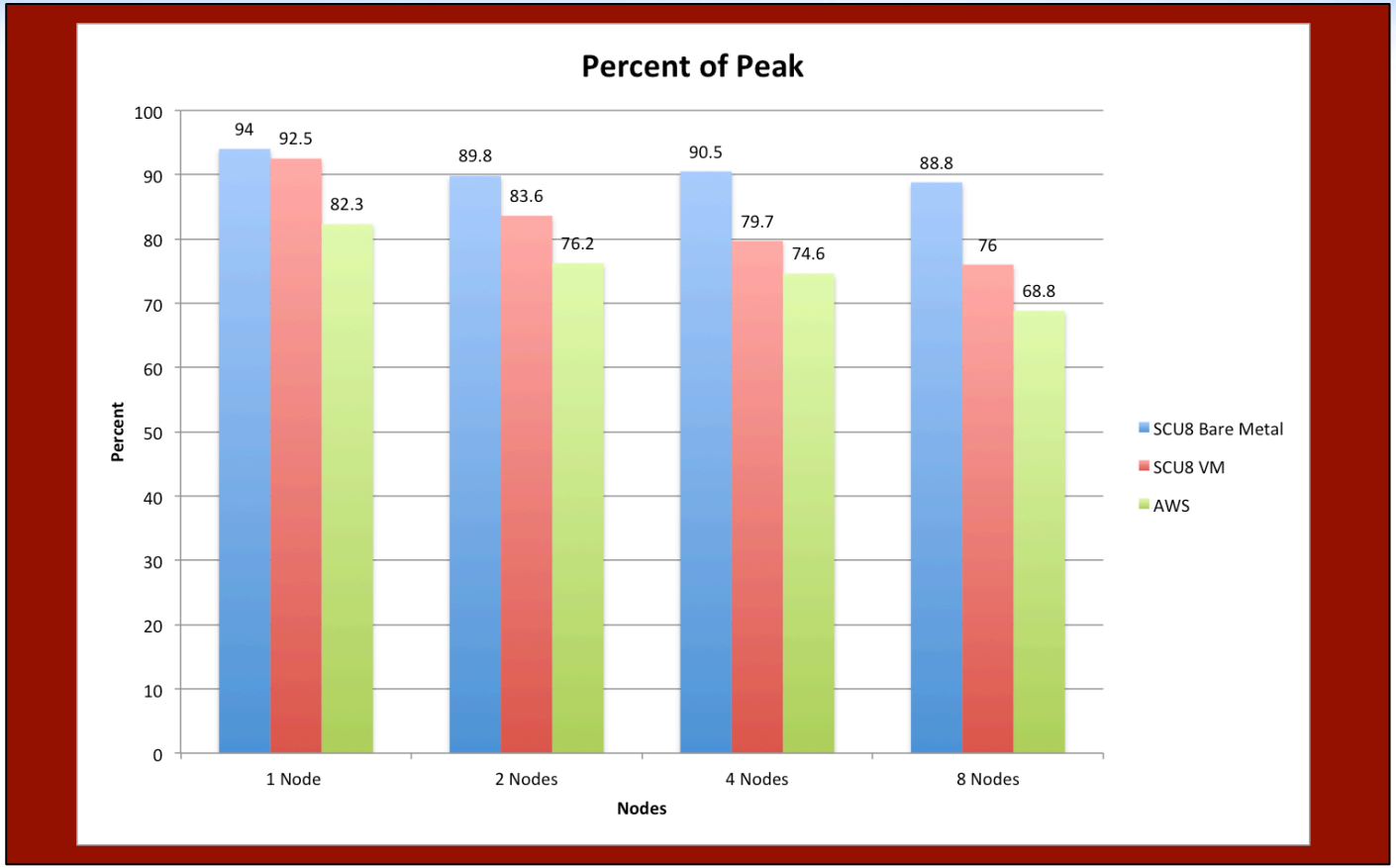
VM memory bandwidth actually exceeded bare-metal

VM bandwidth/latency between nodes matched bare-metal

Multi-node VM vs. bare-metal very good results – performance/scaling

LINPACK	NPB Class D							
	Kernels					Pseudo Applications		
	IS	EP	CG	MG	FT	BT	SP	LU
88%	94%	98%	94%	96%	100%	90%	88%	91%

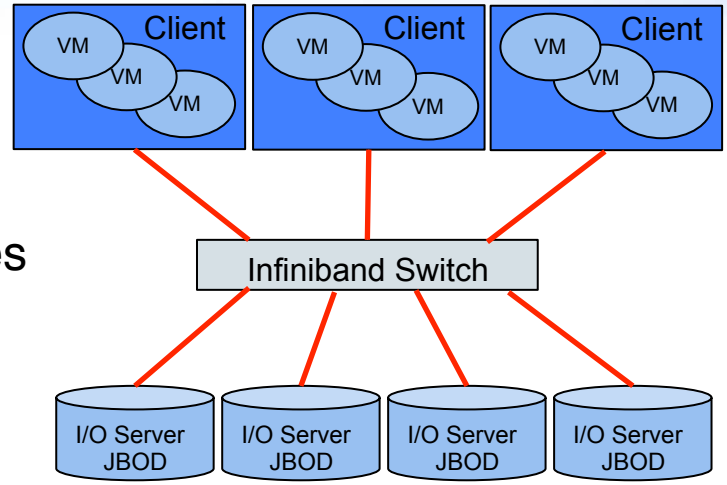
Scale Out Comparisons



GlusterFS Operational Prototype

Recently acquired 960TB raw storage

- Four I/O servers
  - IB frontend
  - SAS backend
- Four 60-bay JBODS populated with 4TB drives
  - One per I/O server
- Various Gluster volume configurations



Bare metal Gluster clients

- Connected to I/O servers using IB/RDMA

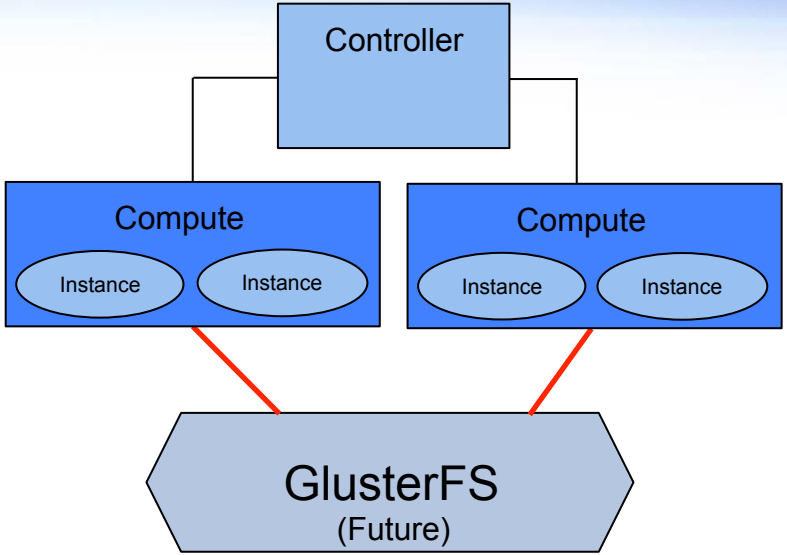
VM Gluster clients

- Connected to VMs servers using virtualized IB/RDMA

Red Hat OpenStack (RHOS) Cloud

Set-up three node evaluation system

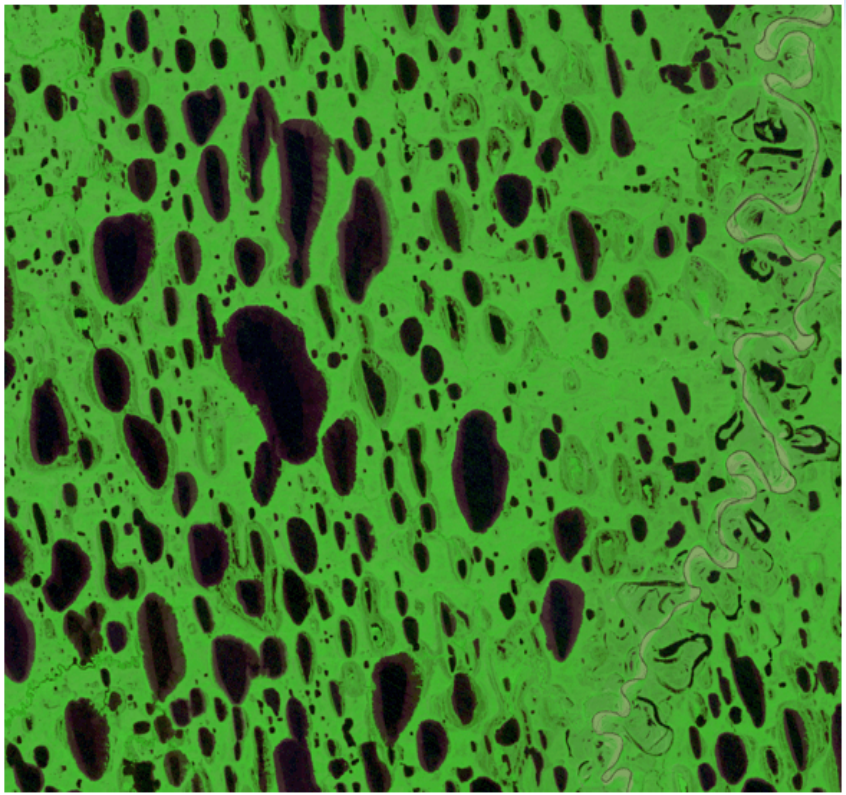
- Havana release
- One controller
- Two compute node
- More compute nodes as available



Objective

- Gain hands on familiarity
- Work with Red Hat/Mellanox
  - Constructs for declaring virtualized IB connections
  - Rapid HPC cluster instantiation
- Define architecture – mix of IB and traditional Ethernet
- Seed the Science Cloud

Decadal Water Products for ABoVE



0 km 5

Landsat image, false color composite, from near Barrow, AK

Small lakes and ponds are a prominent feature of the landscape in the High Northern Latitudes. These ponds will be mapped at 30m spatial resolution at 3 epochs (1991, 2001, 2011) prior to the Arctic Boreal Vulnerability Experiment (ABoVE) field campaign. This will allow researchers to identify areas to study that are either constant or ones that are changing. The effort will take advantage of the time series of Landsat data that is available in this region to provide the max, min, and average condition of each lake/ pond 1ha or larger for each epoch.

Courtesy of Mark Carroll, Sigma Space Corporation

Science Cloud Use Case

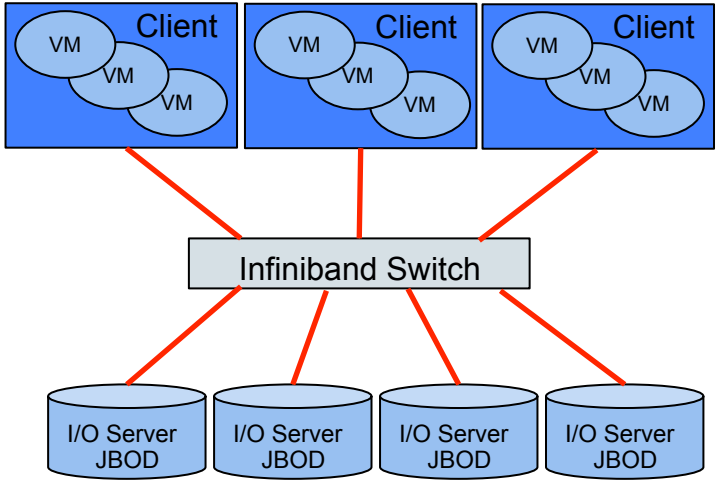
The time series for each epoch will draw from 3 years (1990–1992; 2000–2002; 2010–2012)

To cover the study region this translates to >25,000 scenes to process

Each scene is categorized into land, water, and other (cloud, ice, shadow, undetermined)

The results are then stacked and summed to produce 1 map for each epoch that is the "average" condition for that period

The resultant maps can be used to identify areas of change and areas that are stable



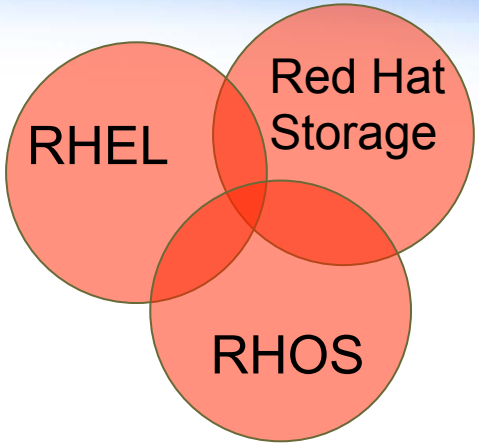
Testing In-Progress / Next Steps

Virtualization overhead elimination (push for 0%)
- Single node VM matches bare metal
- Different VM configurations

Newer Intel CPU performance – virtualized IB
- Eight node Sandy Bridge cluster in test
  - Different I/O structure stirring interest
- Ivy Bridge system in the works
  - 80 node test system – scaling out > 8 nodes

GlufterFS performance tuning
- RDMA write/read rates
- Various volume and file system configurations

Red Hat OpenStack
- Grow test cluster into Science Cloud
- Host science cloud use cases directly – self directed resources



Thanks to …..

Red Hat – software and tuning support

Mellanox – hardware loaners and technical support

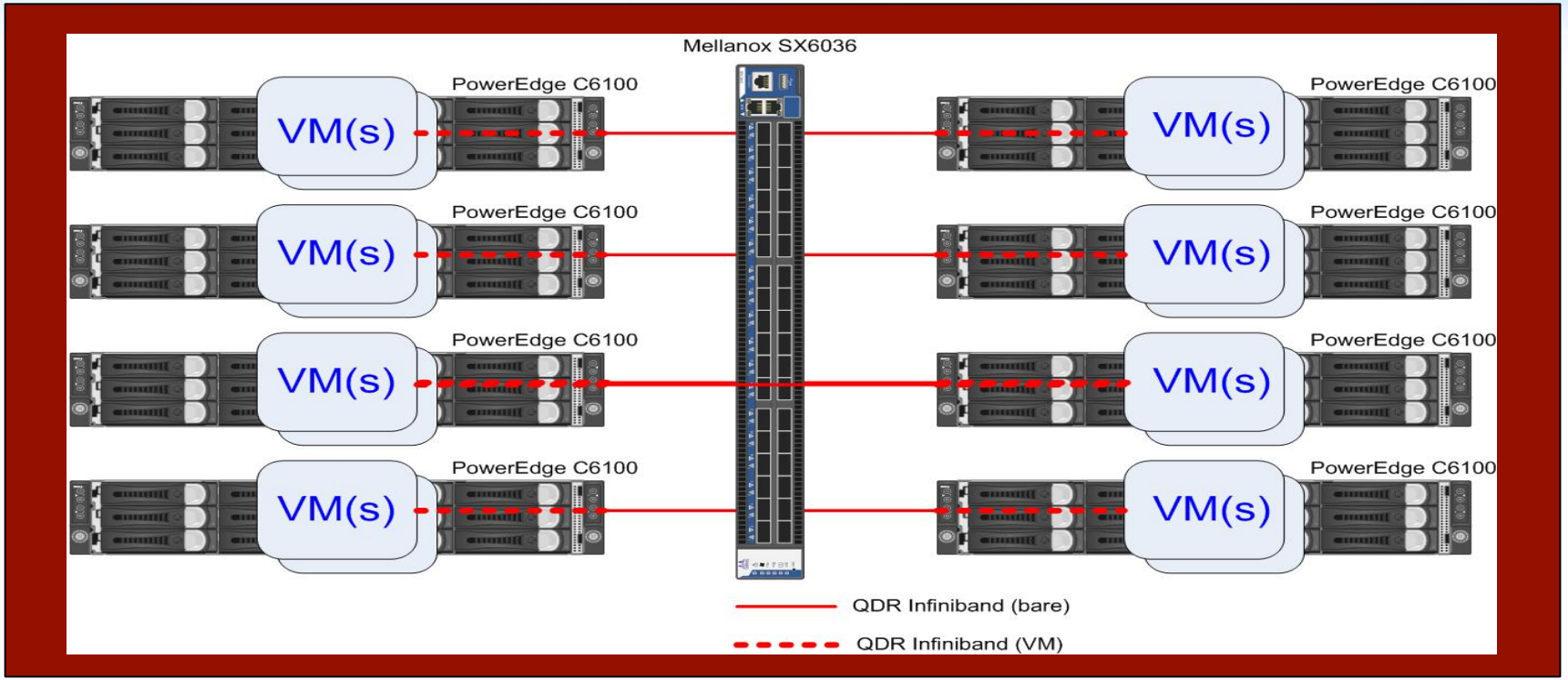
OSU – mvapich2 software support

Questions?

Hoot Thompson
NASA Center for Climate Simulation (NCCS)
john.h.thompson@nasa.gov

Benchmarking Details

Virtualized IB Test Configuration



Configuration Details

Item	Details
Processor Type	Westmere
Processor Number	X5660
Processor Speed	2.8GHz
Sockets per Node	2
Cores per Socket	6
Cores per Node	12
Main Memory	24GB
Interconnect	Mellanox MT26428 QDR IB
Operating System	Red Hat 6.4
Kernel	2.6.32-358.2.1

SR-IOV Basics – Virtual Functions

- BIOS setting
- Kernel iommu enabled
- Special firmware – modified .ini
- Distro Infiniband modules

VM Configuration

- Cloned Westmere features
- 12 cores
- 20 GB memory
- Red Hat 6.4
- 2.6.32-358.2.1 kernel
- Hugepages
- Pinned cpus
- 1 VM per node

Detailed Test Results

Single node

- Memory bandwidth
- LINPACK

Multiple node

- Node-to-node bandwidth
- Node-to-node latency
- Eight node LINPACK
- Eight node NPB

⇒ Spread host file

LINPACK Benchmark Setup

Two different LINPACK versions

- Openmp – single node
- Hybrid – one or more mpi processes each starting 1 to 12 threads

Different block sizes (NB)

- 144, 168, 192 and 216

P Q settings

- P X Q equals number of mpi processes to fill node/cluster
- P always less than Q
- Minimize P – Q (square is best)

NAS Parallel Benchmarks - NPB

Mimic the computation and data movement in CFD applications

Different class levels (C, D) reflect different problem sizes

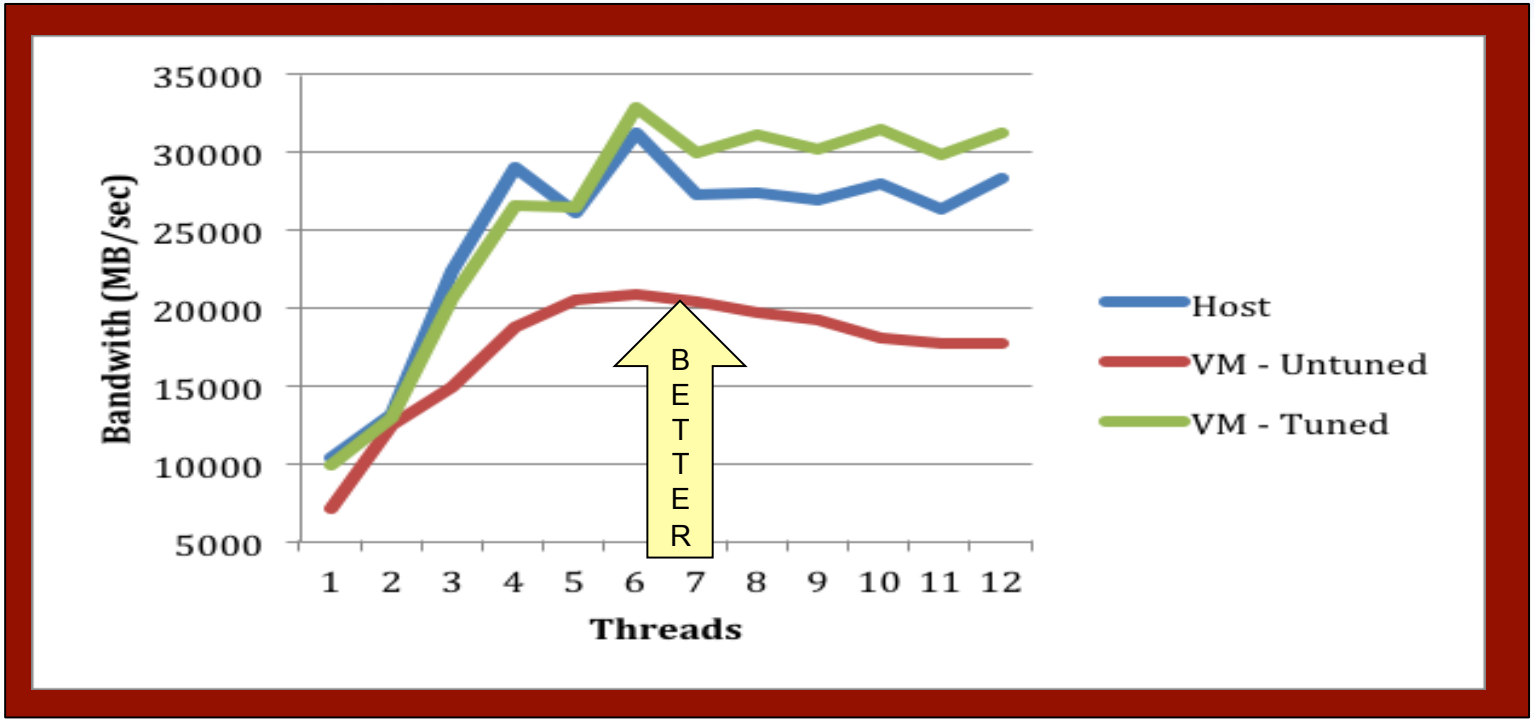
Five Kernels

- IS – Integer Sort, random memory access
- EP – Embarrassingly Parallel
- CG – Conjugate Gradient, irregular memory access and communication
- MG – Multi-Grid on a sequence of meshes, long and short distance
- FT – discrete 3d fast Fourier Transform, all-to-all communication

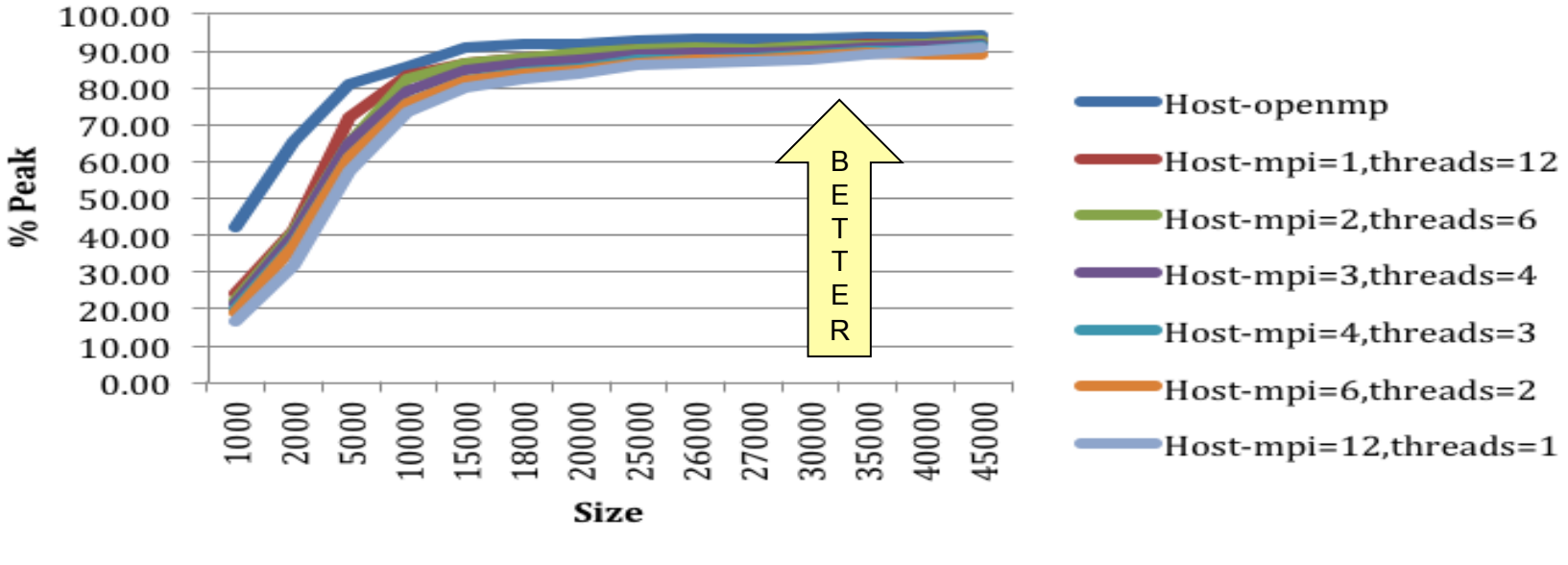
Three pseudo applications

- BT – Block Tri-diagonal solver
- SP – Scalar Penta-diagonal solver
- LU – Lower-Upper Gauss-Seidel solver

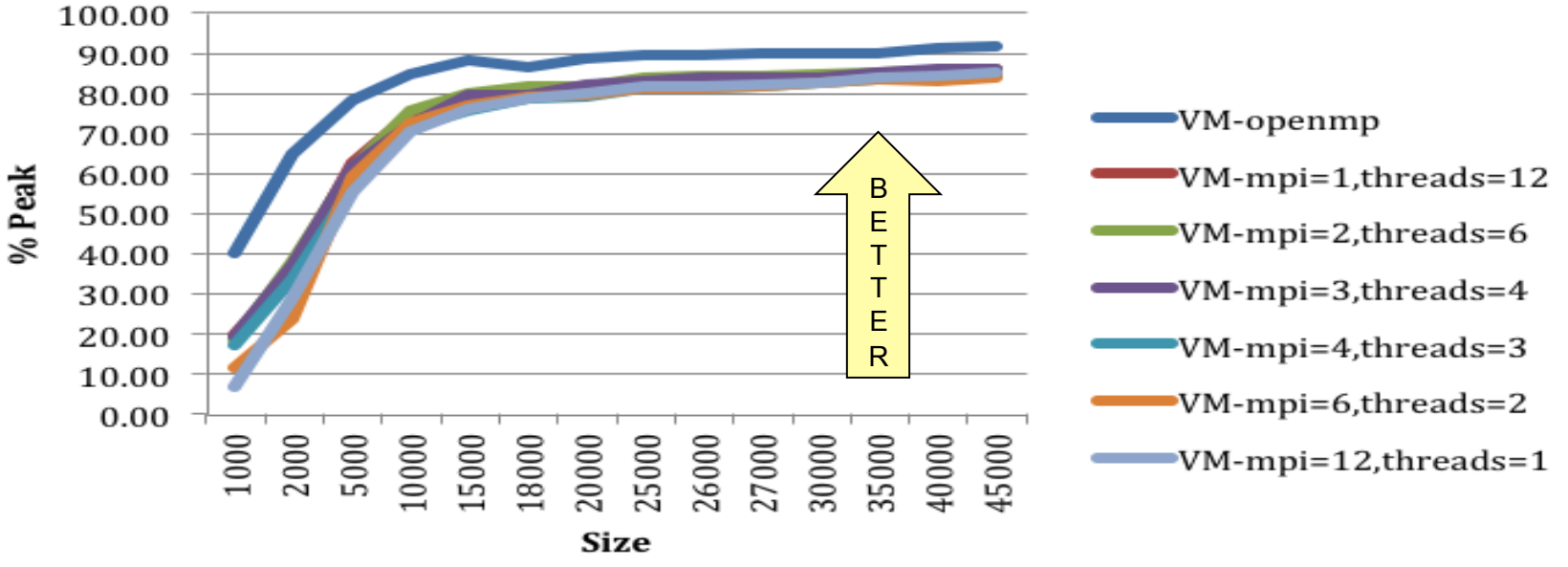
Memory Bandwidth – Single Node



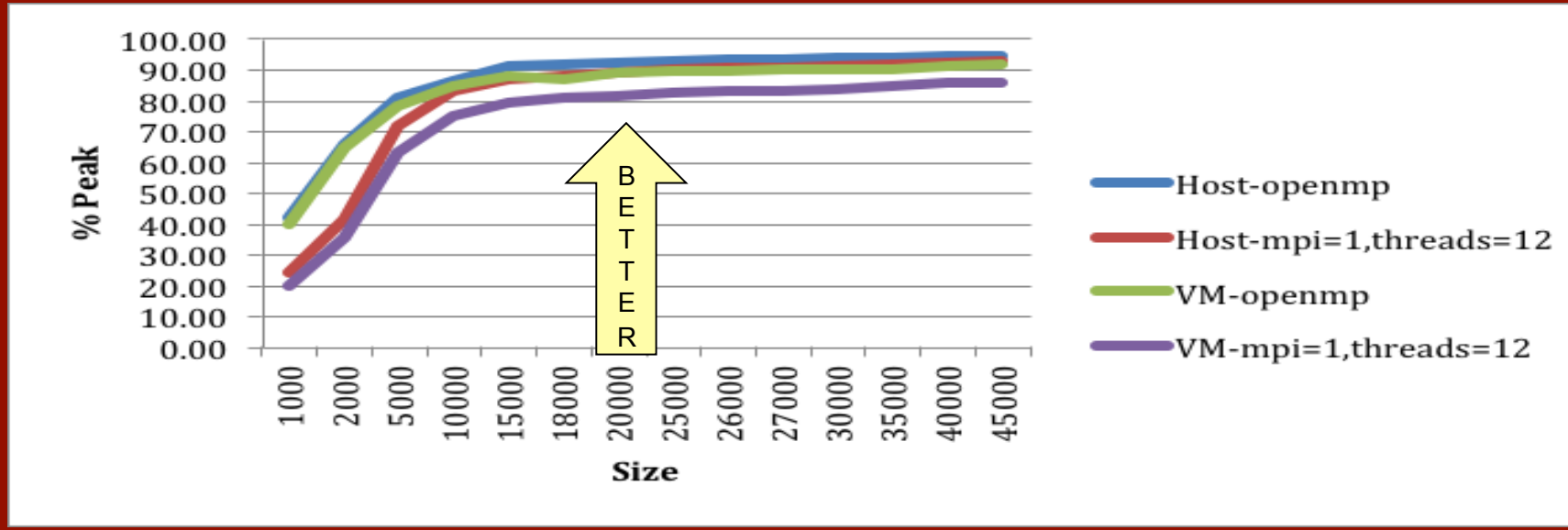
LINPACK – Single Node Host



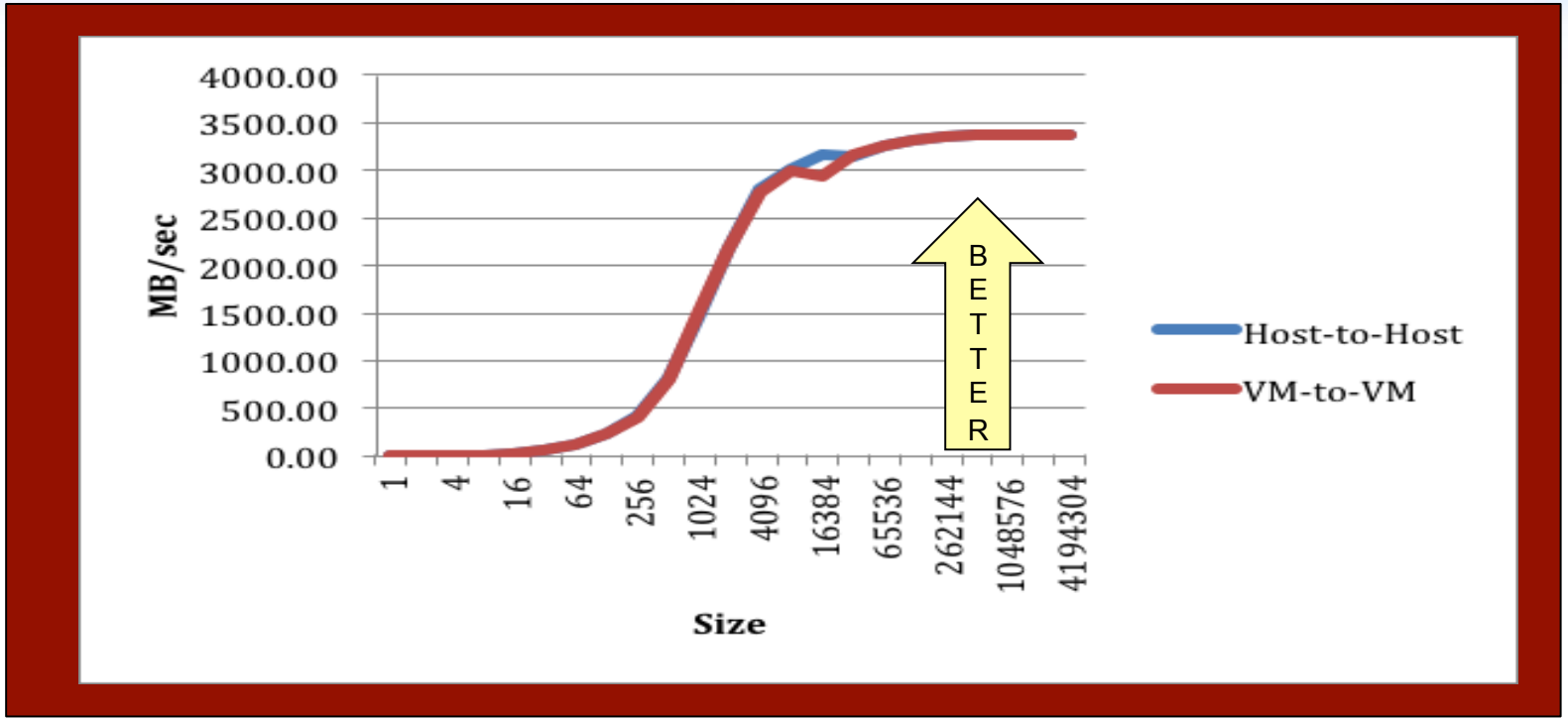
LINPACK – Single Node VM



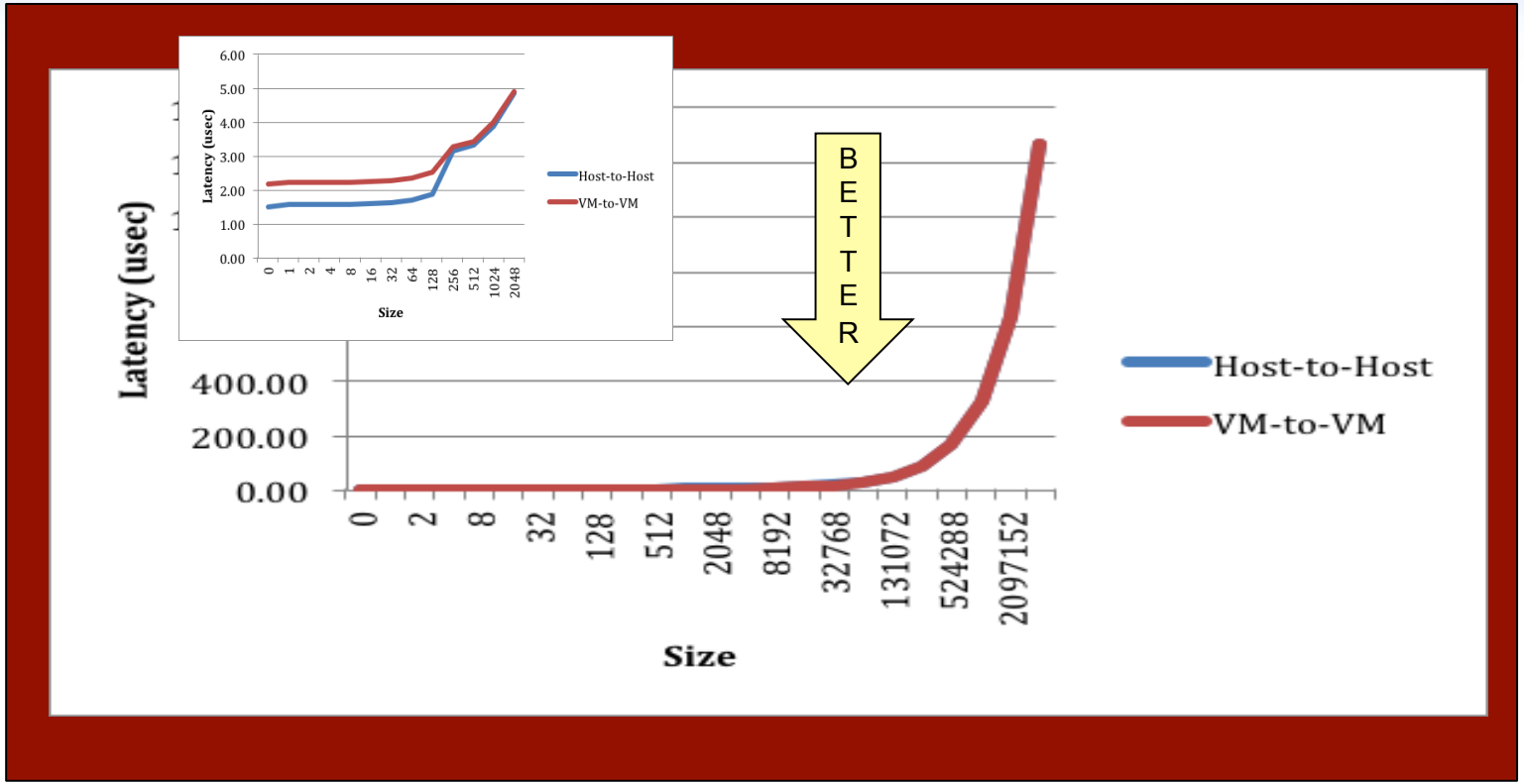
LINPACK – Host versus VM Comparison



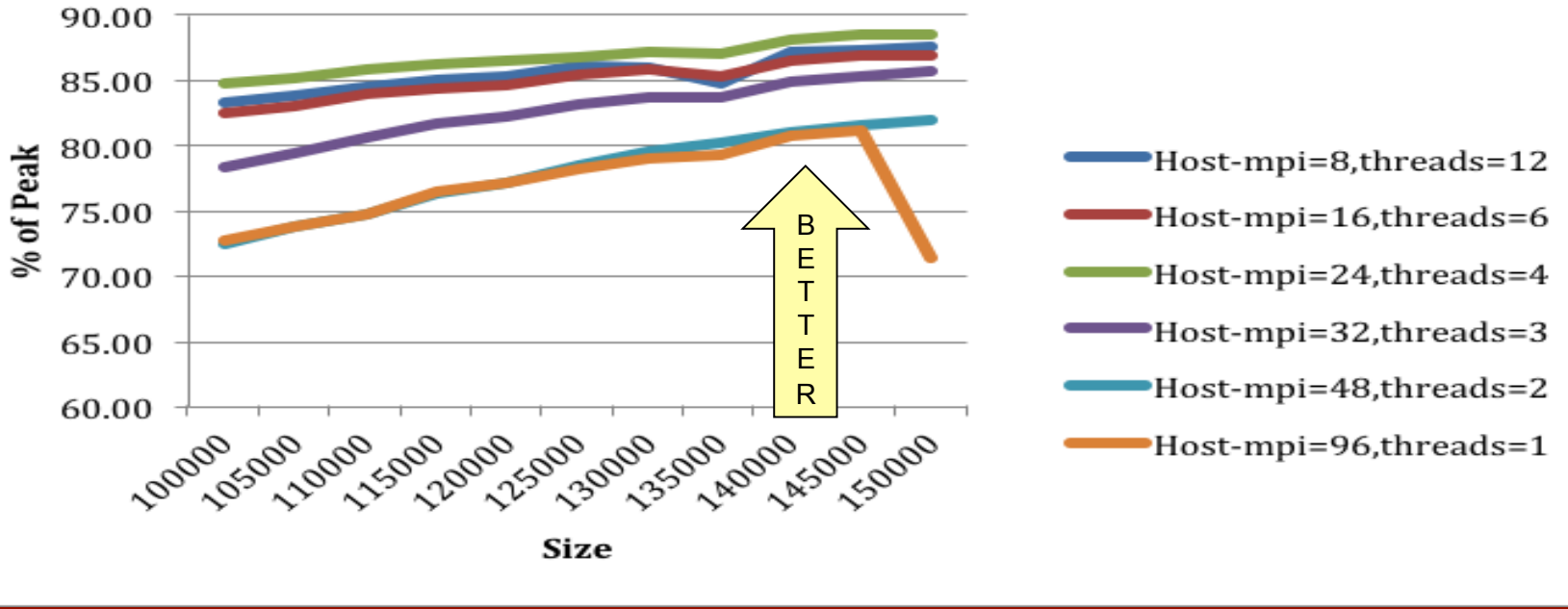
Bandwidth – Node-to-Node



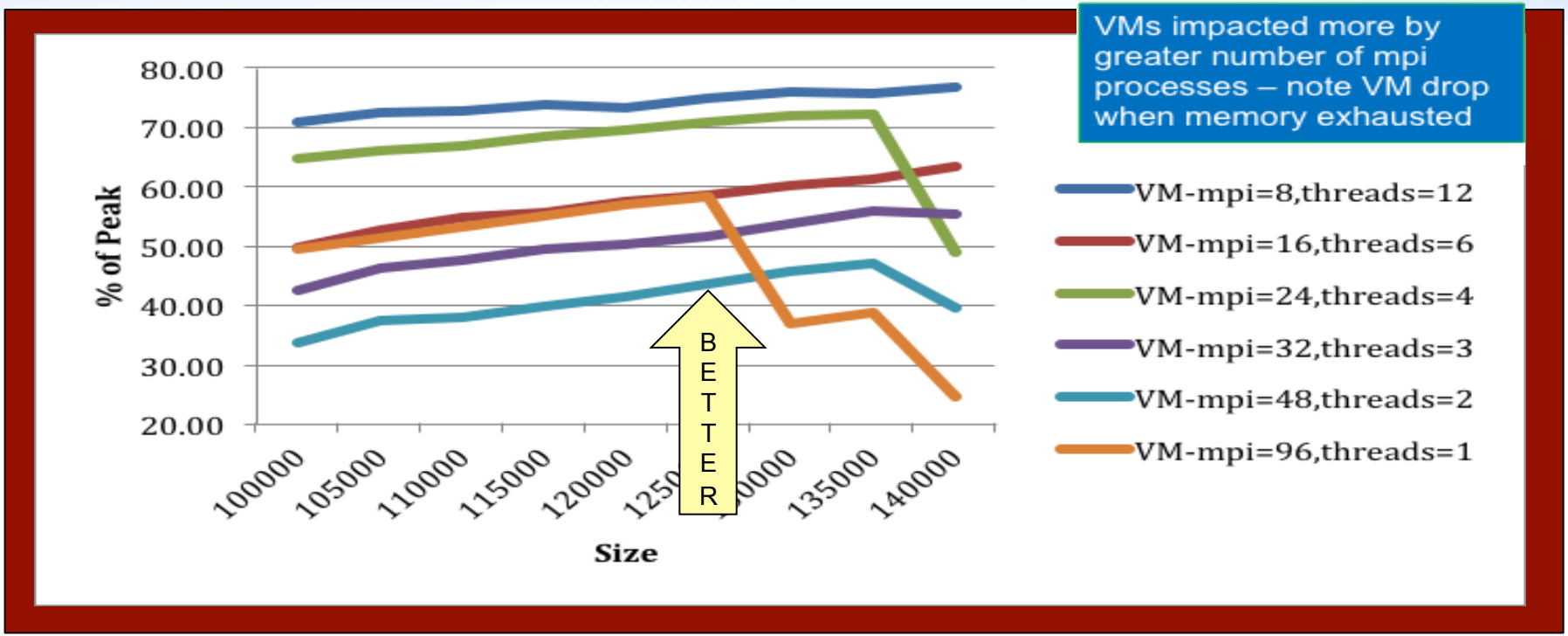
Latency – Node-to-node Latency



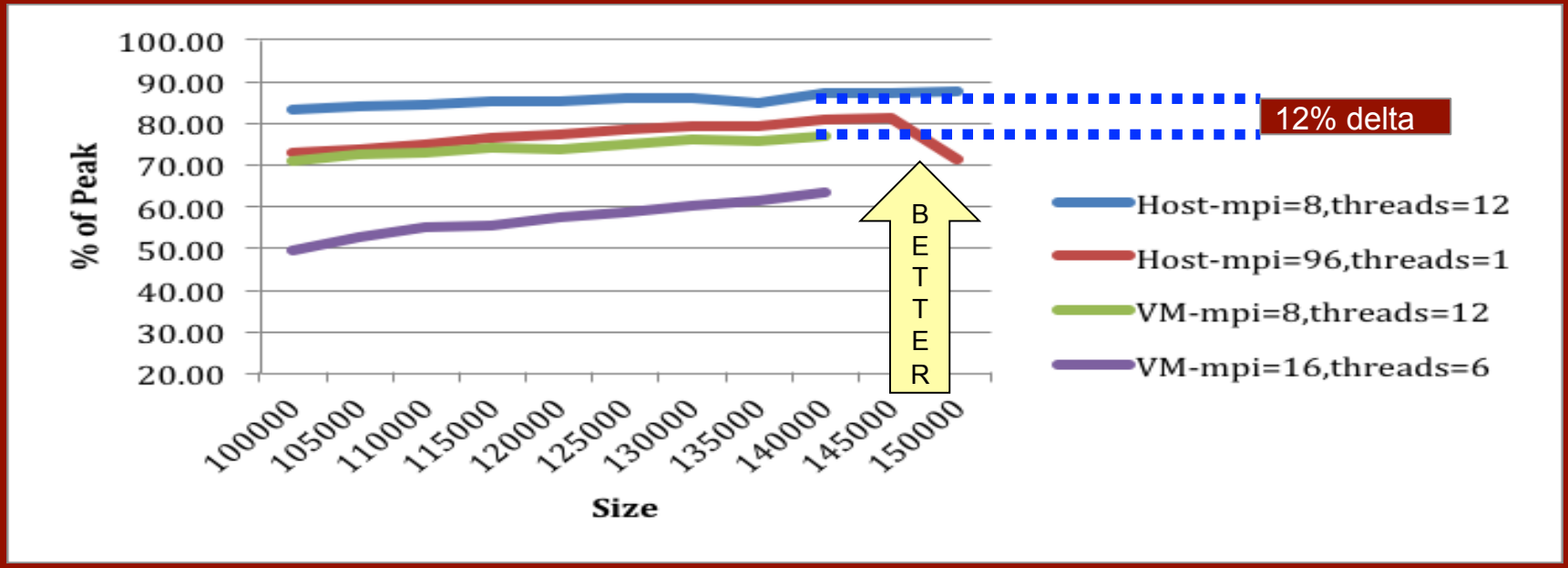
LINPACK – Eight Hosts



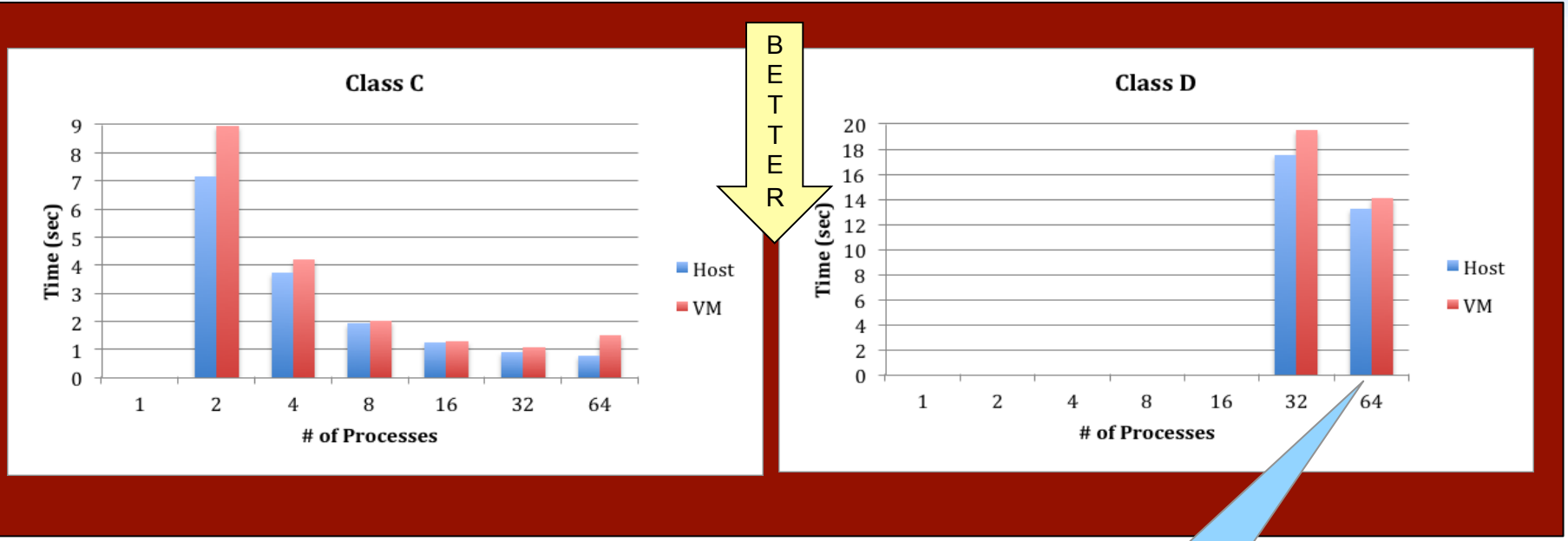
LINPACK – Eight VMs



LINPACK – Eight Hosts Versus Eight VMs



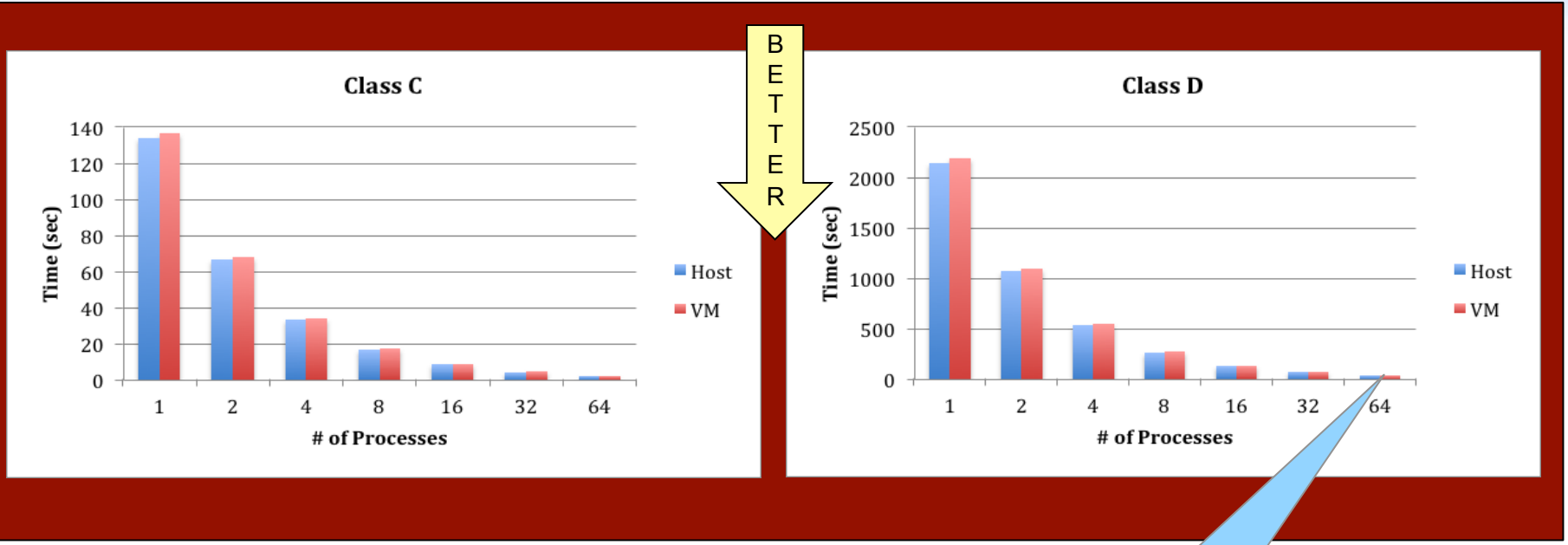
NPB IS – Eight Hosts Versus Eight VMs



Integer Sort, random memory access

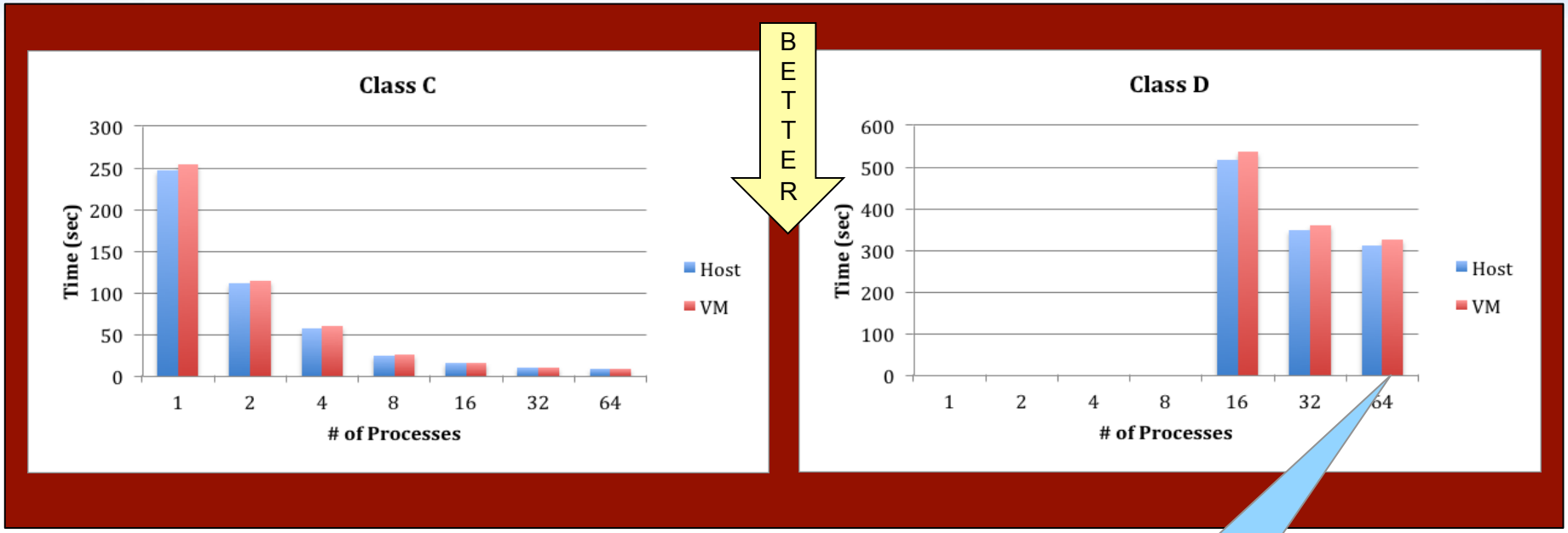
88% Efficiency

NPB EP – Eight Hosts Versus Eight VMs



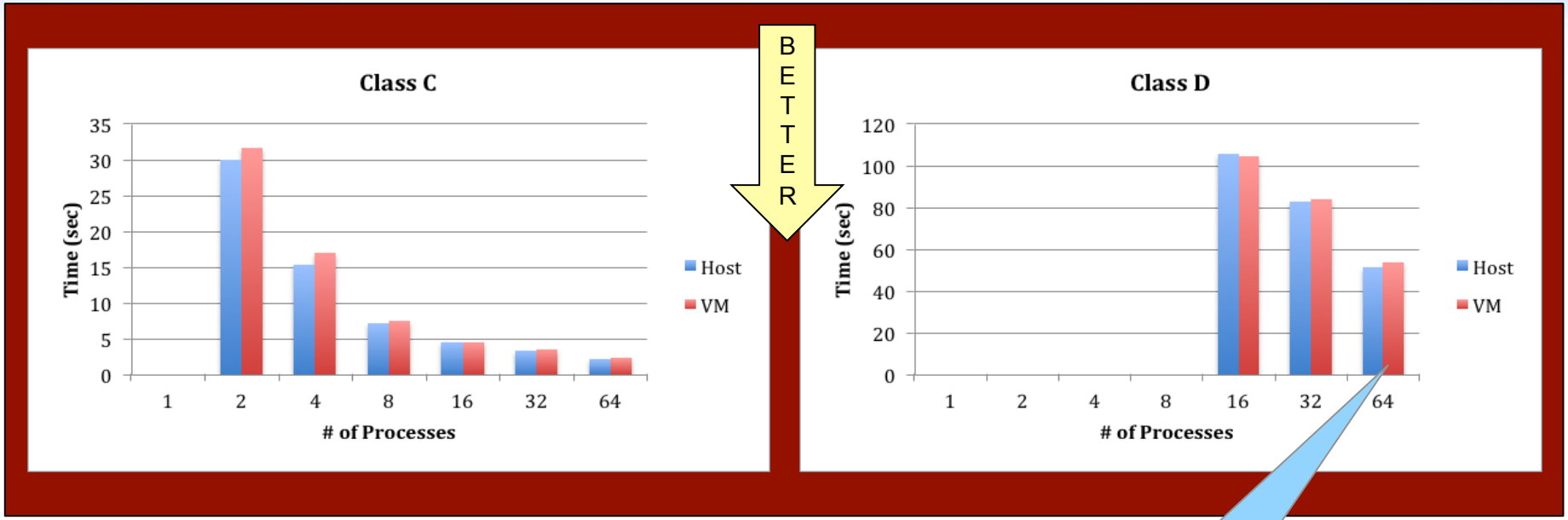
Embarrassingly Parallel

NPB CG – Eight Hosts Versus Eight VMs



Conjugate Gradient, irregular memory access and communication

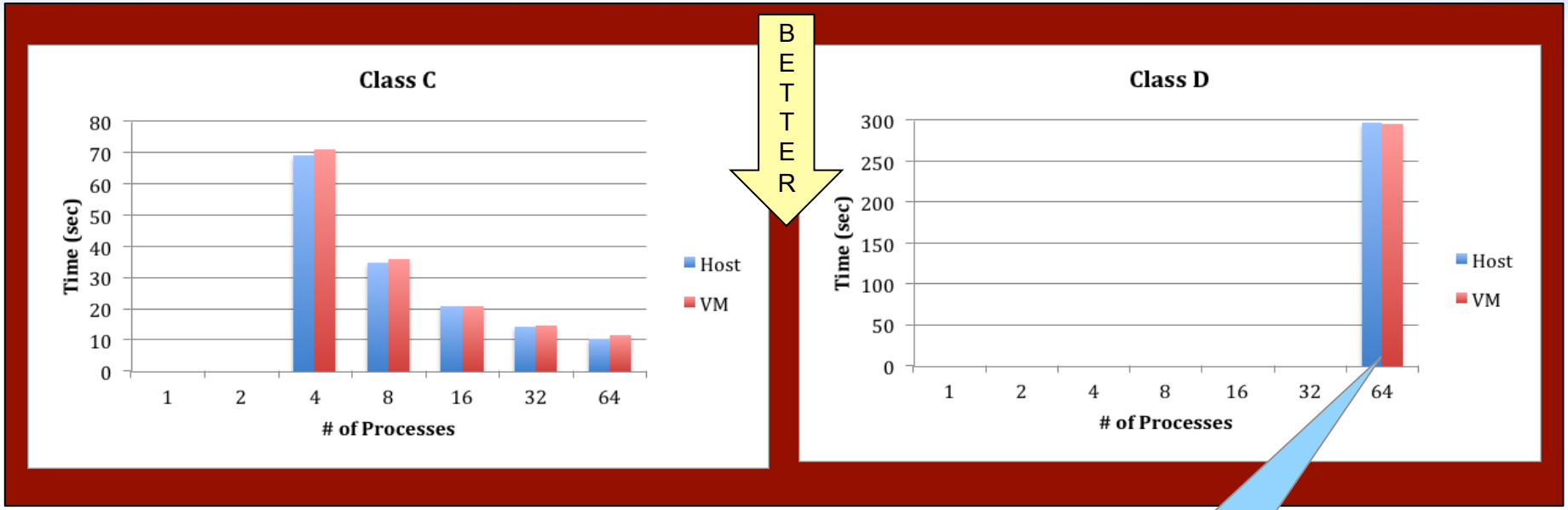
NPB MG – Eight Hosts Versus Eight VMs



Multi-Grid on a sequence of meshes, long and short distance

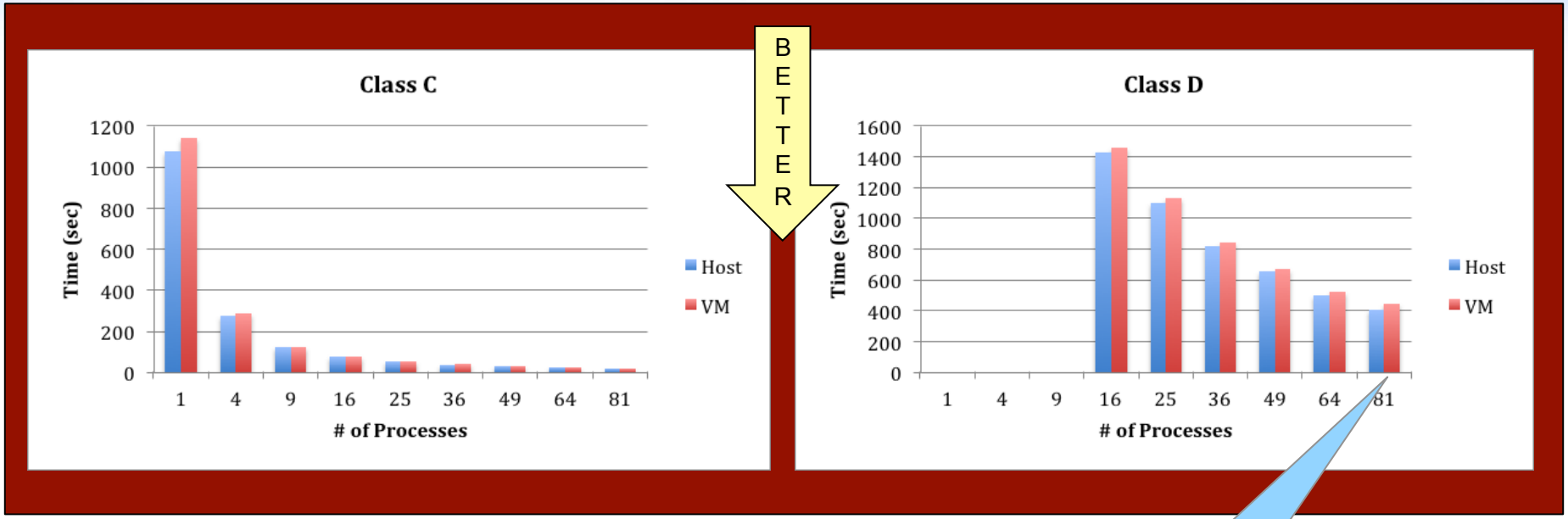
96% Efficiency

NPB FT – Eight Hosts Versus Eight VMs



discrete 3d fast Fourier Transform, all-to-all communication

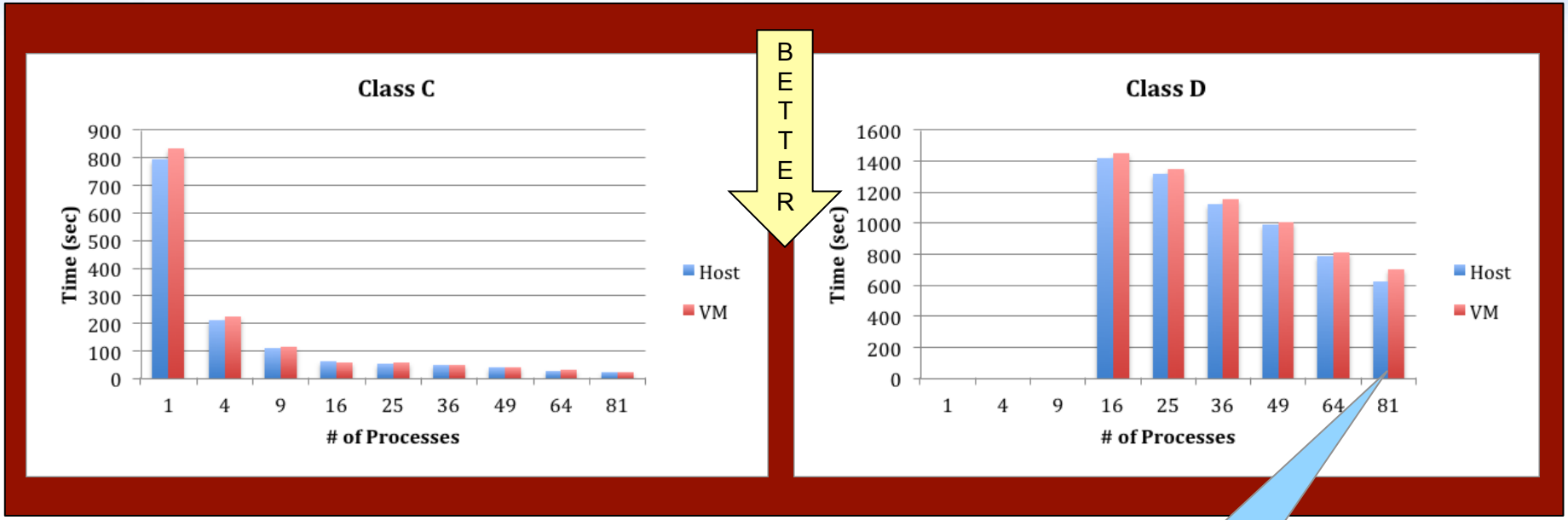
NPB BT – Eight Hosts Versus Eight VMs



Block Tri-diagonal solver

90% Efficiency

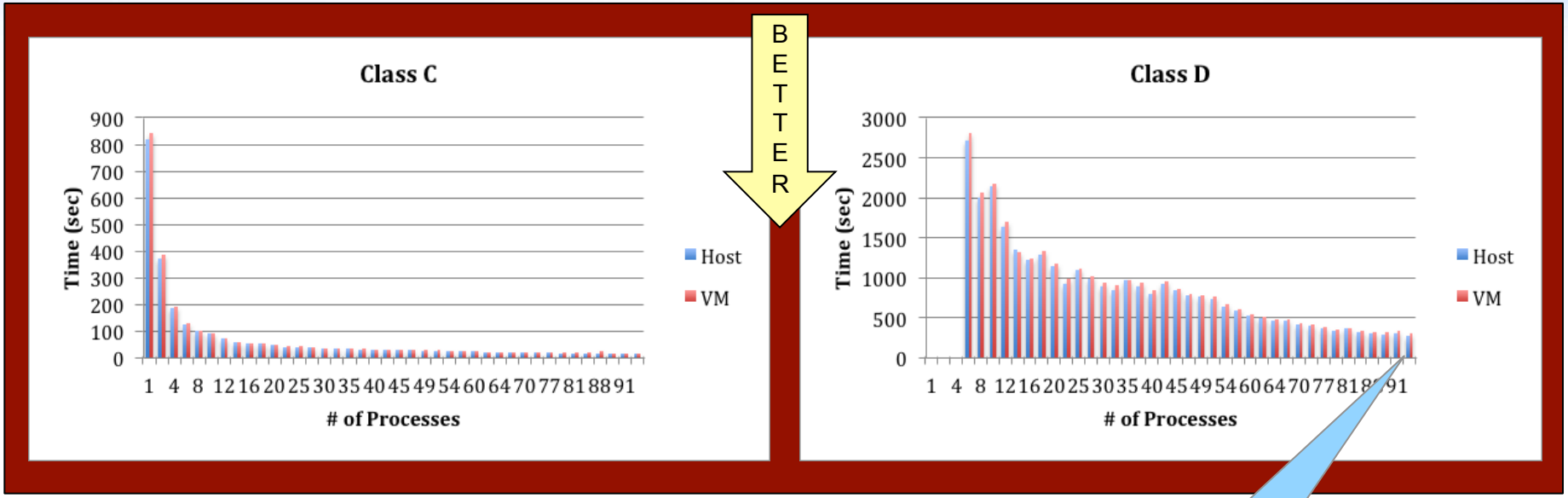
NPB SP – Eight Hosts Versus Eight VMs



Scalar Penta-diagonal solver

88% Efficiency

NPB LU – Eight Hosts Versus Eight VMs



Lower-Upper Gauss-Seidel solver

91% Efficiency