Recommendations for Virtualization in HPC

Nathan Regola & JC Ducom*

Center for Research Computing

University of Notre Dame

*now at Scripps Research Institute

Introduction-Why Profile VMs?

- We wanted to know if VMs are useful for HPC (especially related to I/O).
- If they are efficient enough, then perhaps they could be used to extend the HPC Center into the Cloud
  - Support HPC “cloud” servers such as SGE nodes, Condor nodes, and user uploaded VMs.

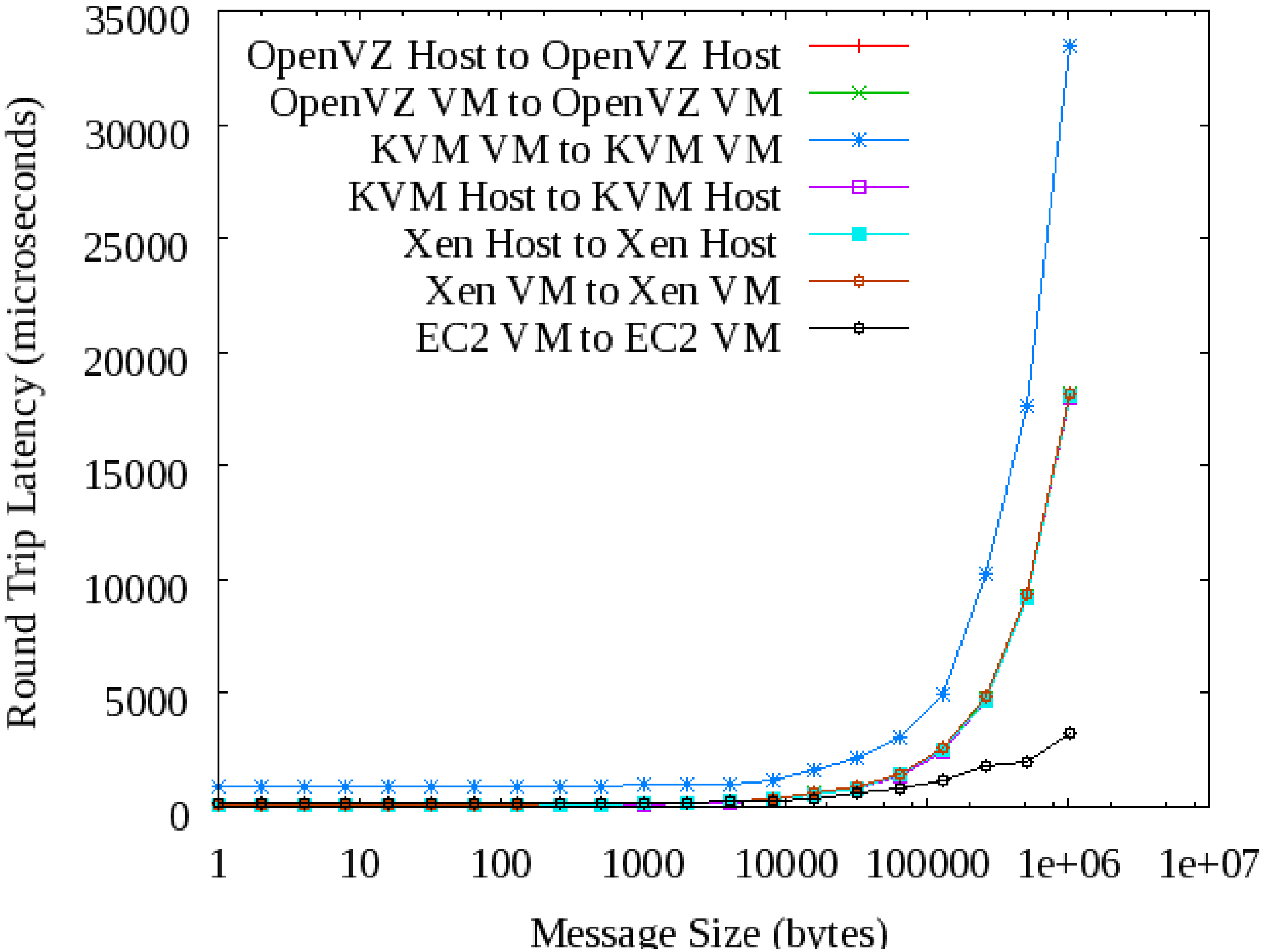
Experiment

- 4 Dell R610 compute nodes with InfiniBand
  - 8 CPU, 12GB RAM (32 cores total)
  - Xen HVM Mode, KVM, or OpenVZ
- 4 Amazon EC2 “Cluster Compute Nodes”
  - 8 CPU, 24GB RAM (32 cores total)
  - 10Gbps Ethernet
  - Xen HVM Mode (not user configurable)

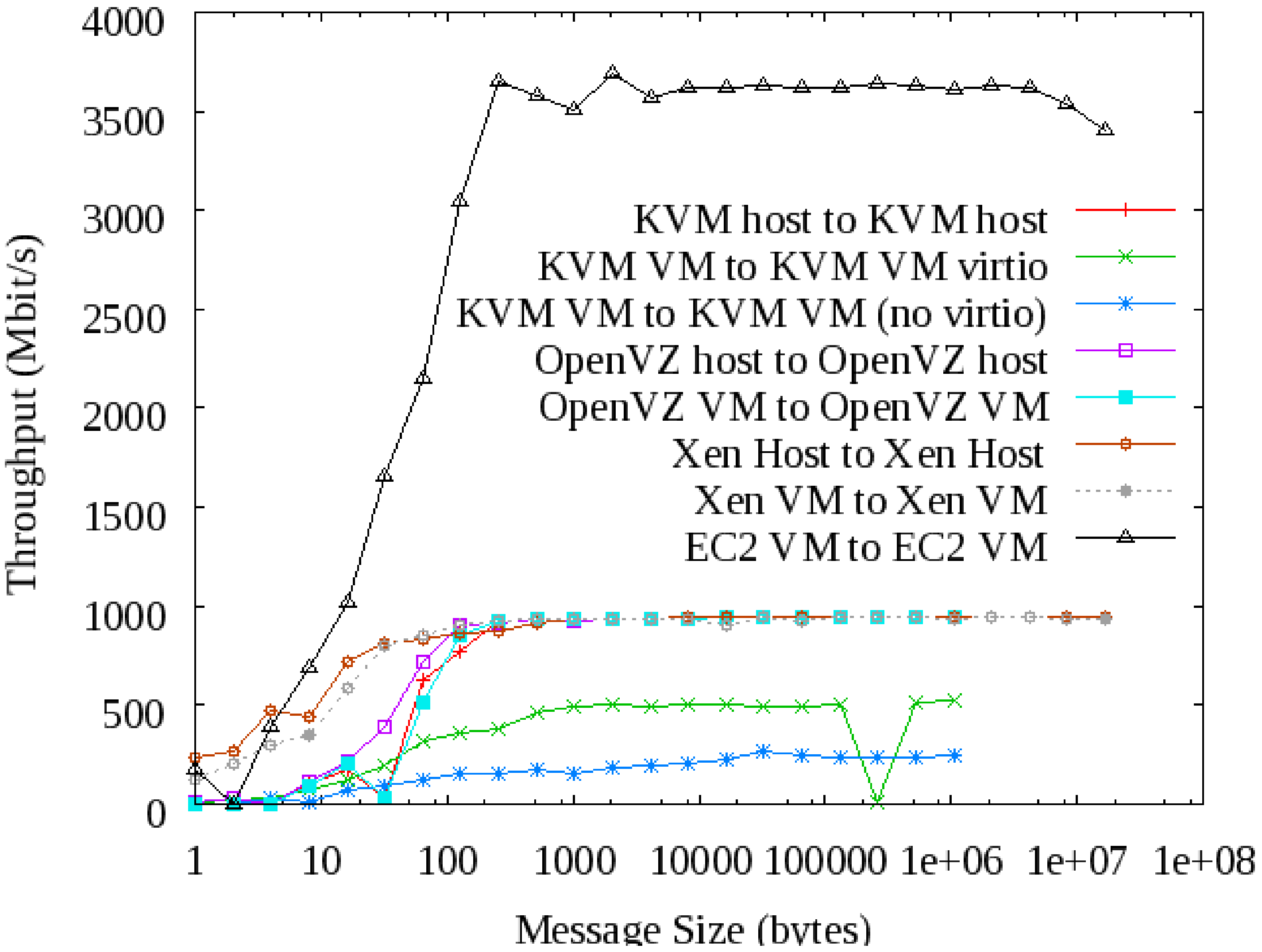
Results

- Operating System virtualization is more efficient (on average) than any paravirtualized or fully virtualized solution for HPC workloads.
- If you must use paravirtualization or full virtualization
  - Currently, KVM isn’t as efficient as Xen

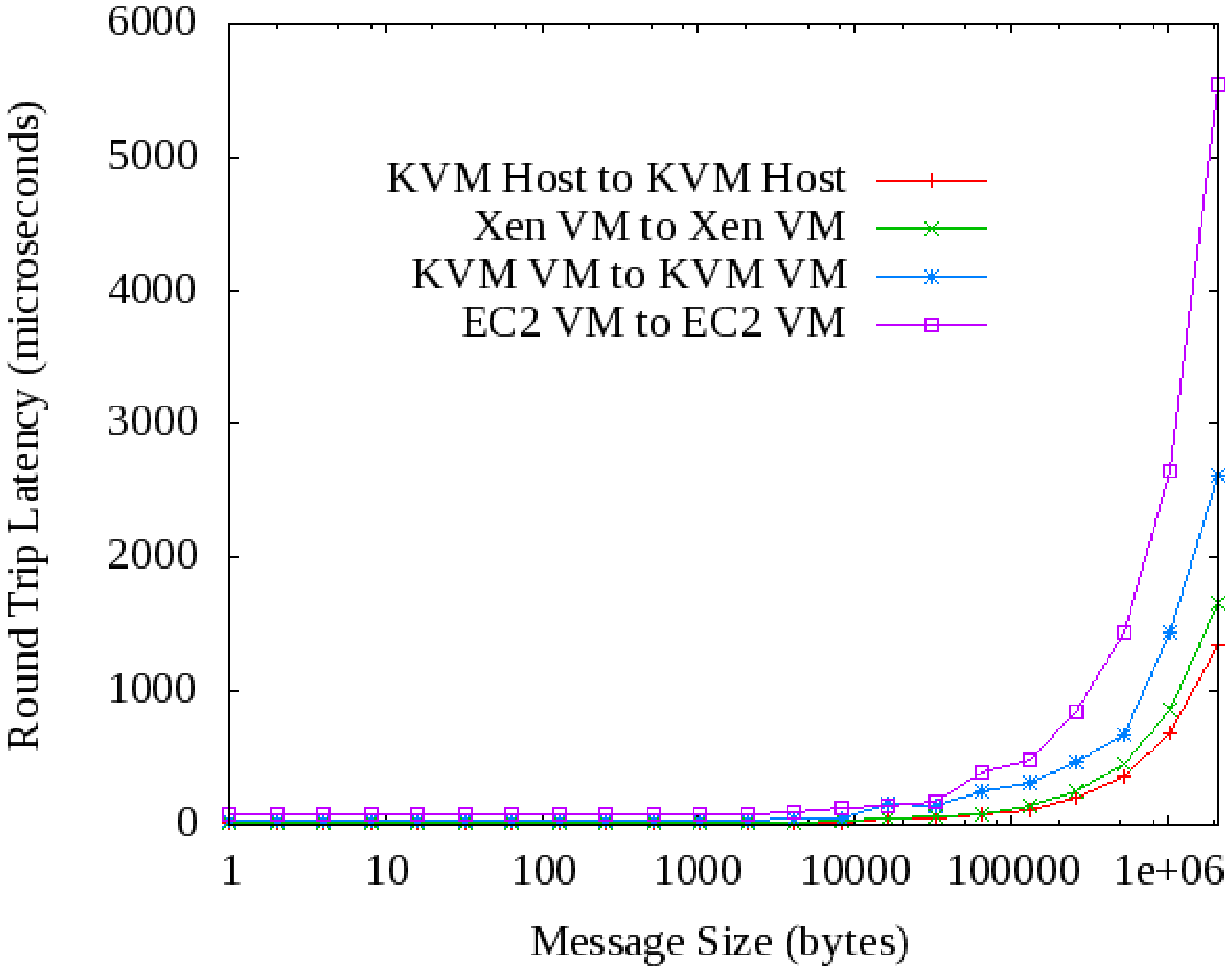
Network Latency—Ethernet



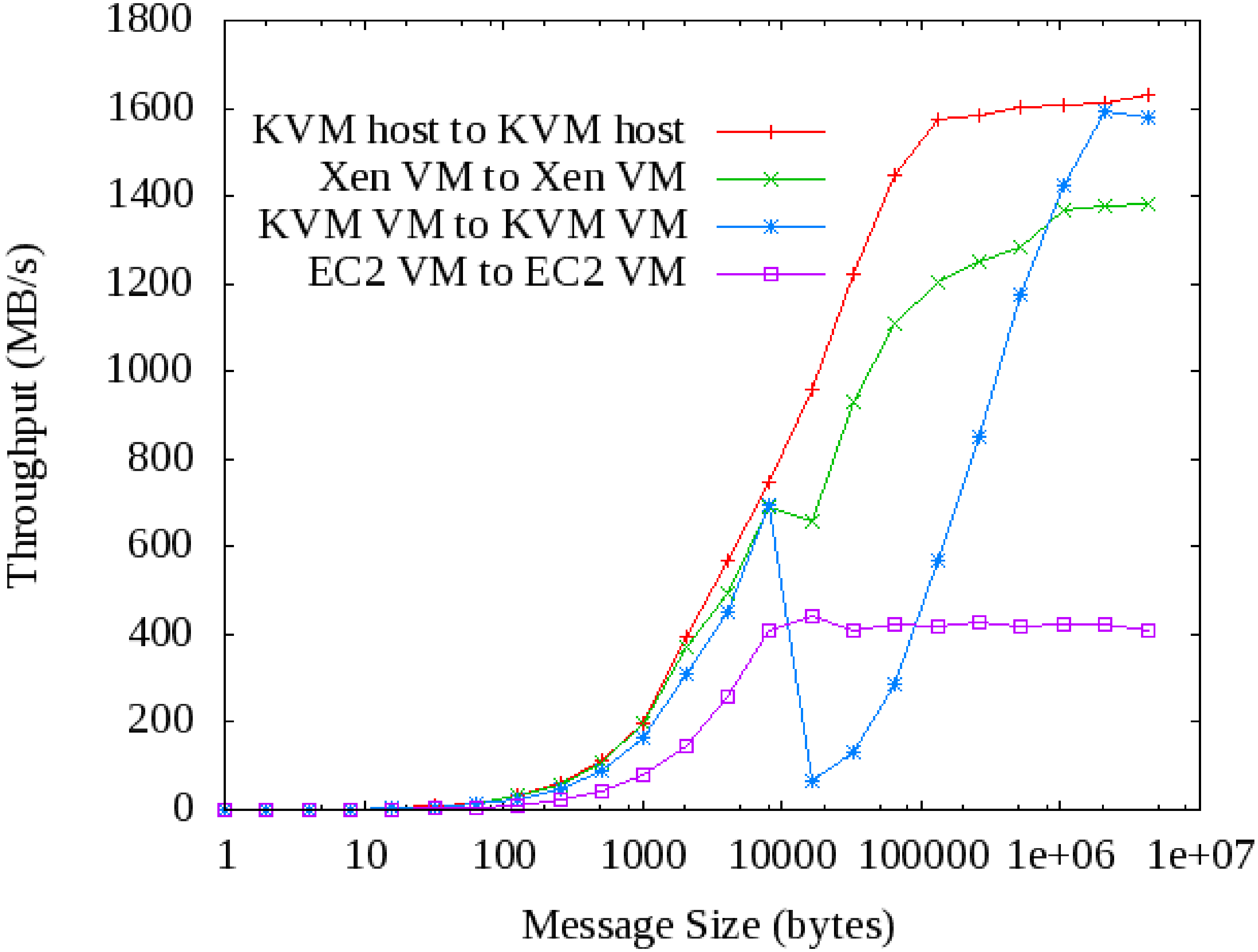
Network Throughput--Ethernet



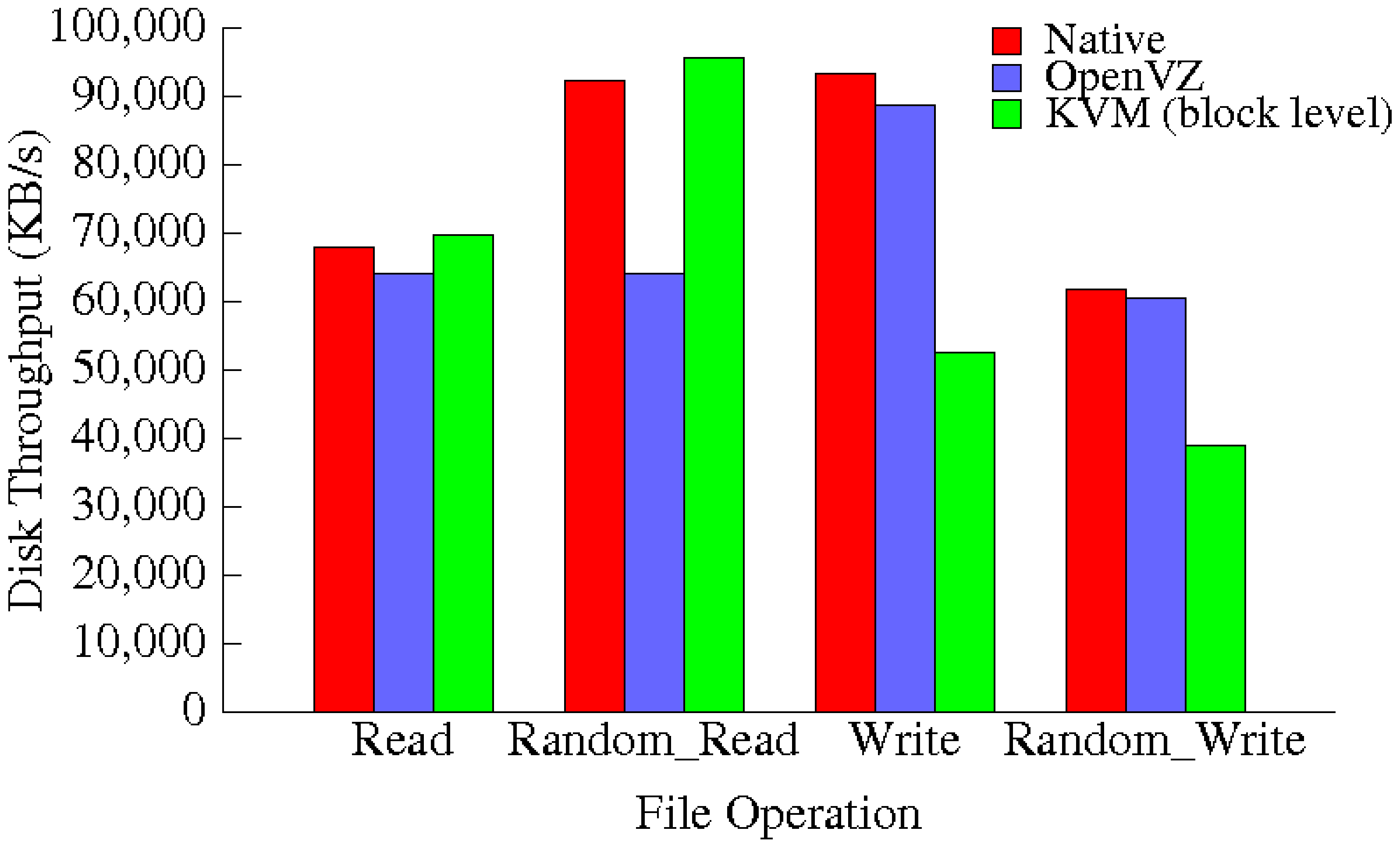
Network Latency—InfiniBand Passthrough



Network Throughput--InfiniBand



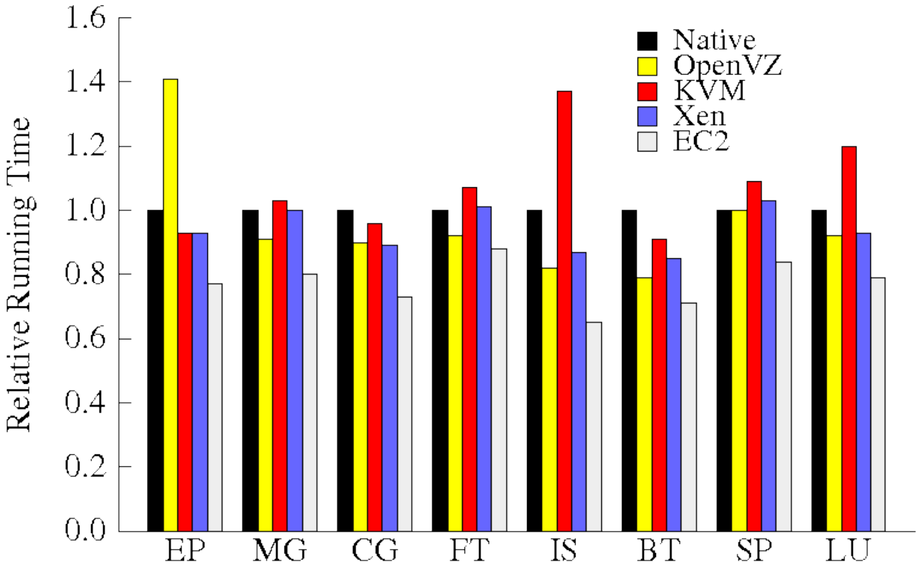
Storage Performance--IOZone

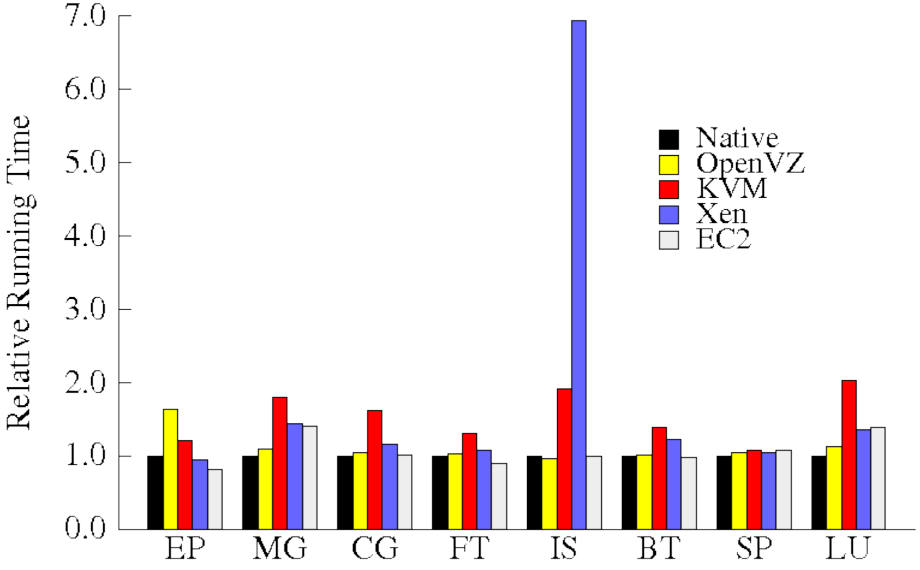


NAS Parallel Benchmarks

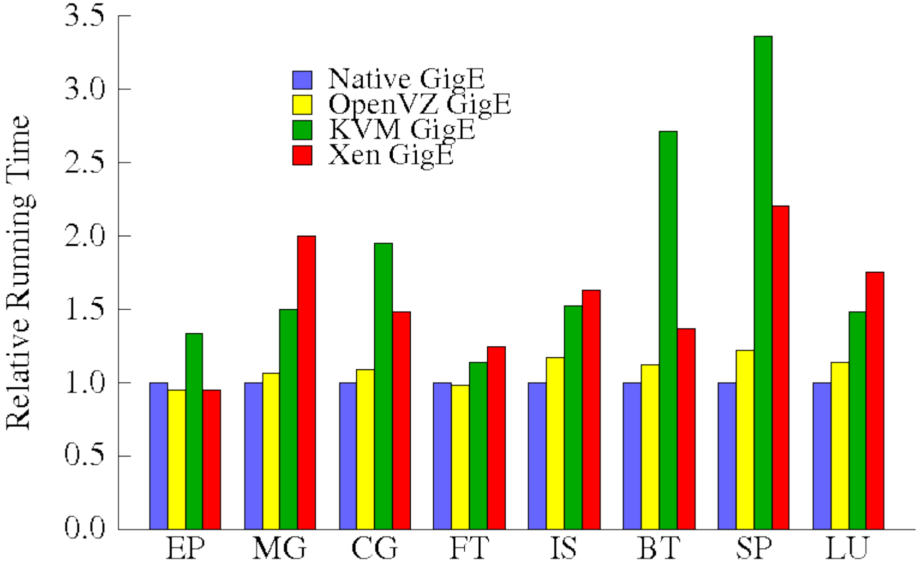
- Suite of five kernels (EP,MG,CG,FT,IS) and three CFD applications (BT,SP,LU)
- NPB benchmarks exhibit large variety of network communications, CPU, memory loads
- Problem size (class): S,W,A,B,C,(D)

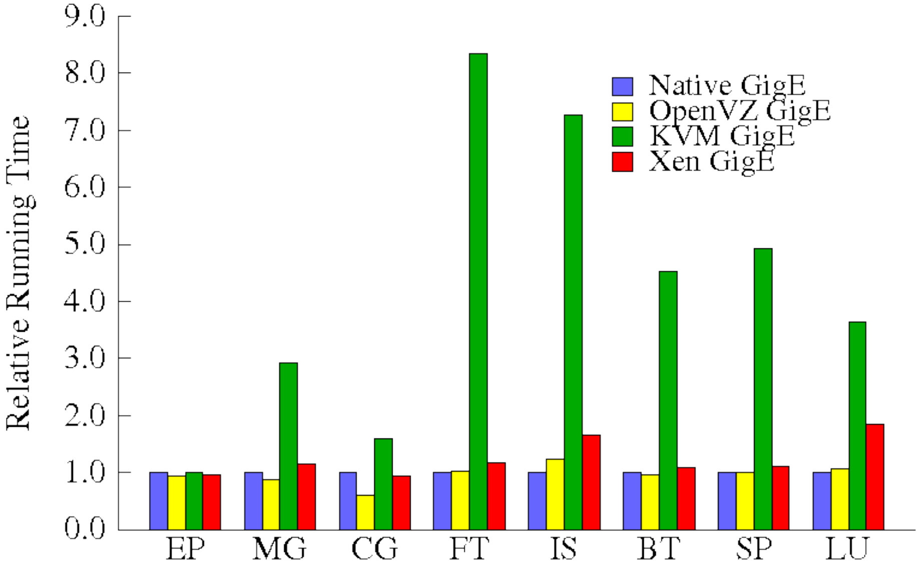
NPB—OpenMP


Relative Running Time
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
■ Native
■ OpenVZ
■ KVM
■ Xen
□ EC2
EP
MG
CG
FT
IS
BT
SP
LU

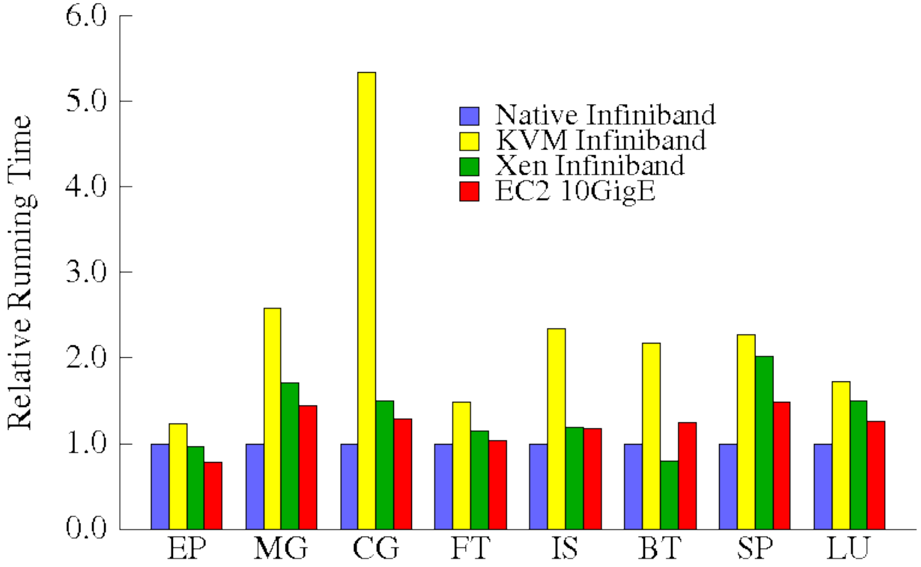

Relative Running Time
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
■ Native
■ OpenVZ
■ KVM
■ Xen
□ EC2
EP
MG
CG
FT
IS
BT
SP
LU

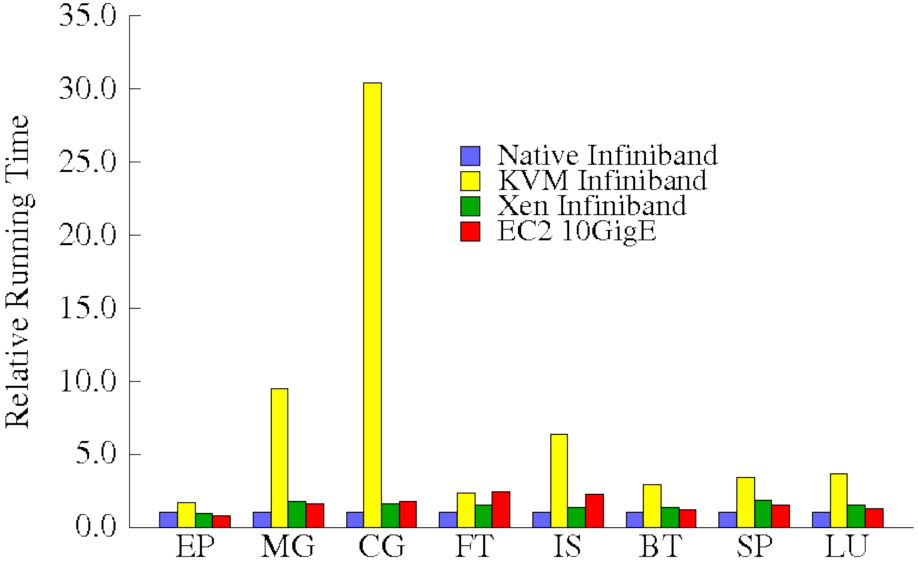
NPB—MPI (GigE)


Relative Running Time
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Native GigE
OpenVZ GigE
KVM GigE
Xen GigE
EP
MG
CG
FT
IS
BT
SP
LU


Relative Running Time
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
Native GigE
OpenVZ GigE
KVM GigE
Xen GigE
EP
MG
CG
FT
IS
BT
SP
LU

NPB—MPI (InfiniBand* passthrough)





Conclusions

- OS virtualization has the lowest overhead on average. Unfortunately no InfiniBand for OpenVZ.

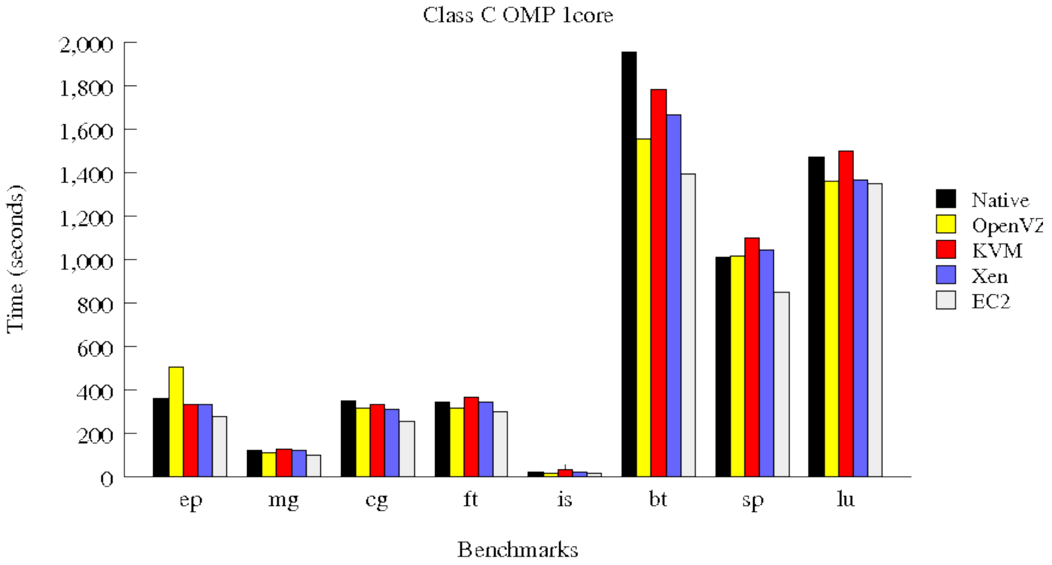
- KVM I/O not mature, under heavy development

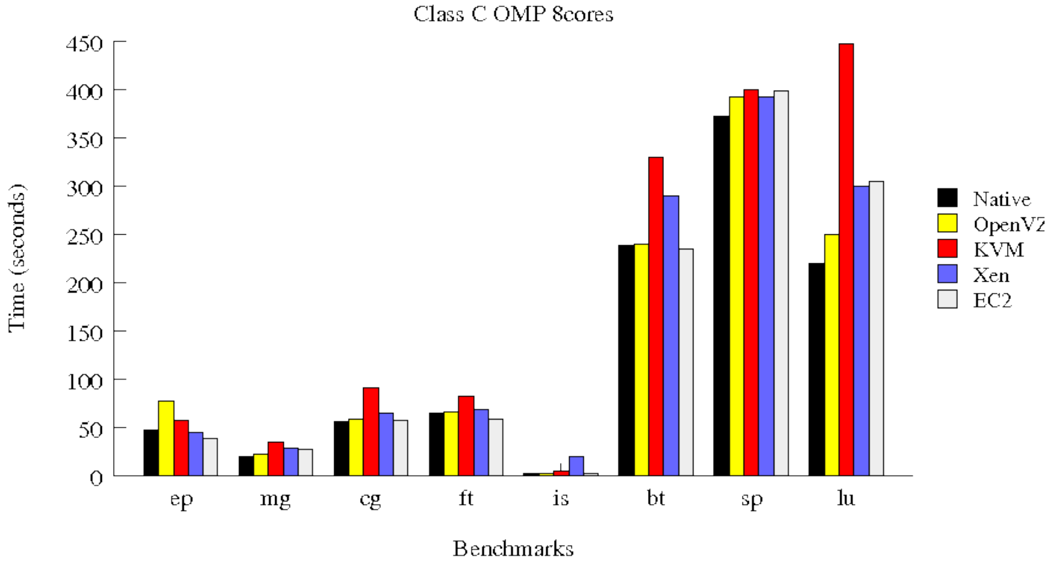
- PCI Passthrough improves scalability but has virtualization overhead

Questions?

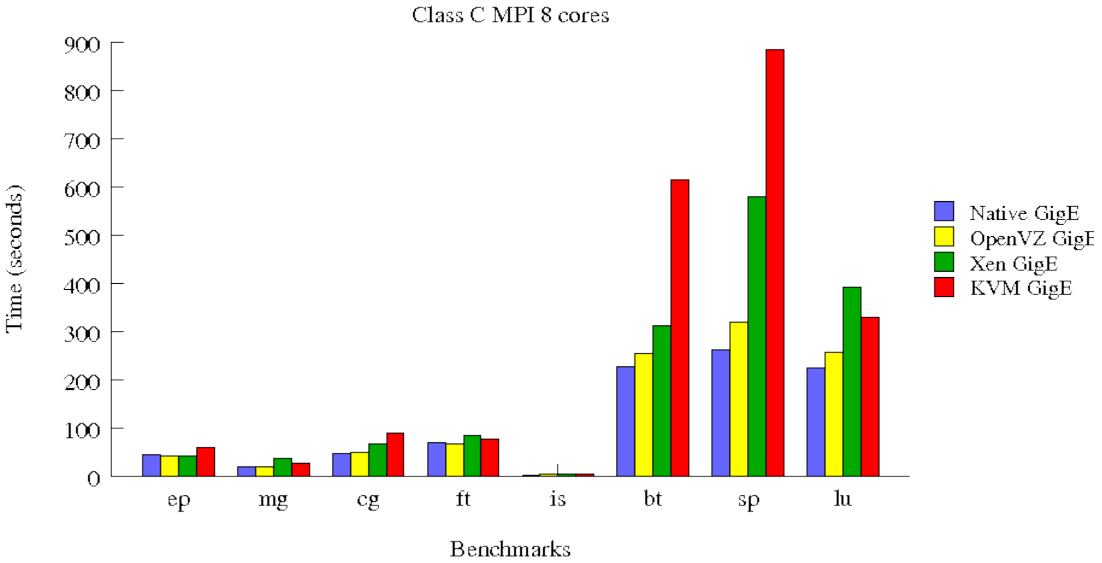
Nathan Regola, nregola@nd.edu

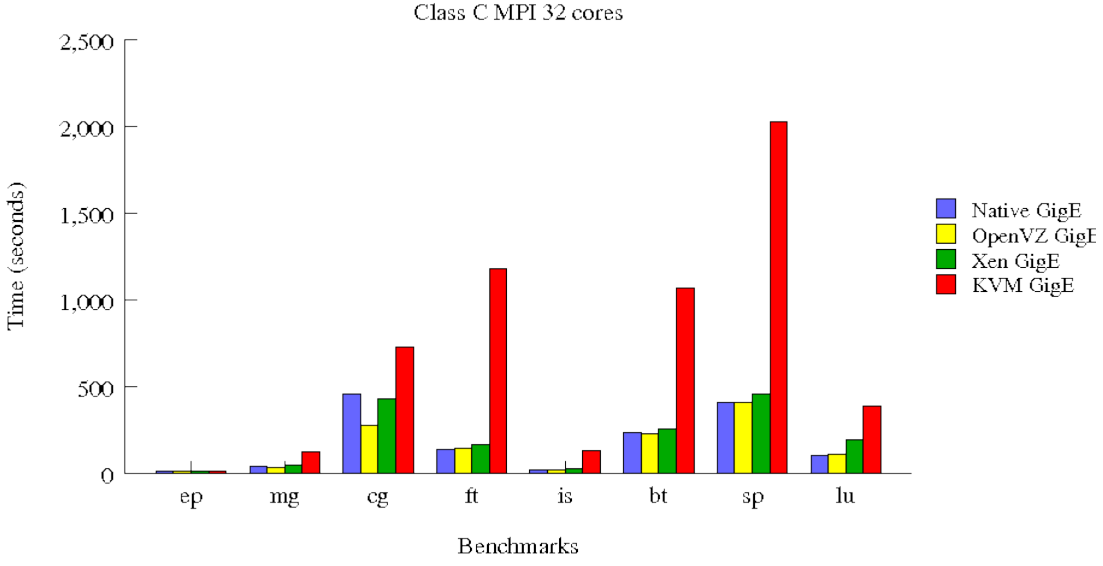
OpenMP—NPB Actual Runtime





MPI-NPB, GigE Actual Runtime


Class C MPI 8 cores


Class C MPI 32 cores