OPENFABRICS ALLIANCE

HPC, Enterprise, Cloud: Converging Concerns

Josh Simons

Office of the CTO, VMware

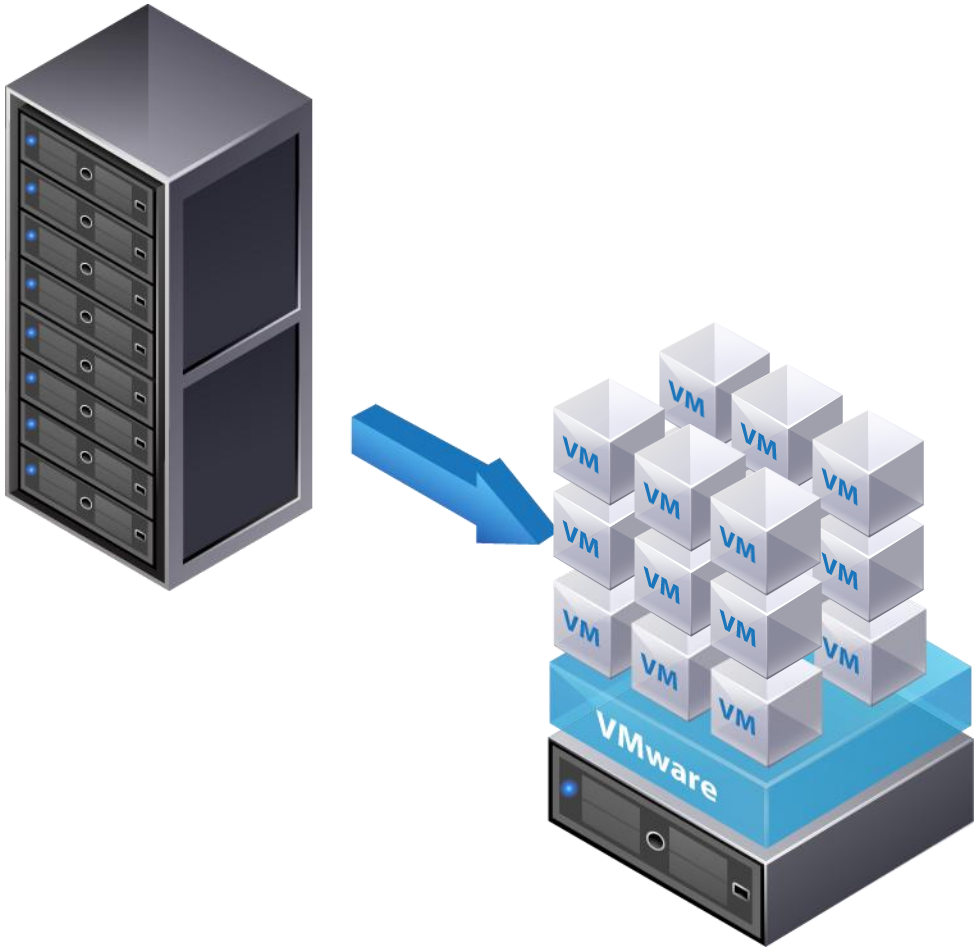
© 2009 VMware Inc. All rights reserved

Virtualization

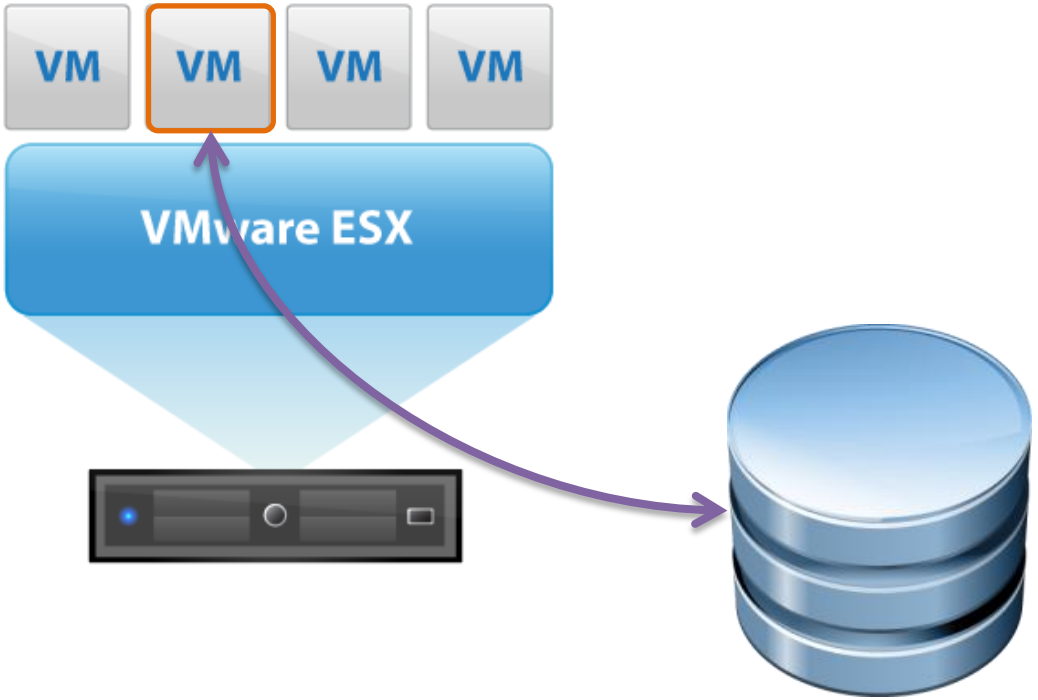
OPENFABRICS ALLIANCE

What Can it Offer?

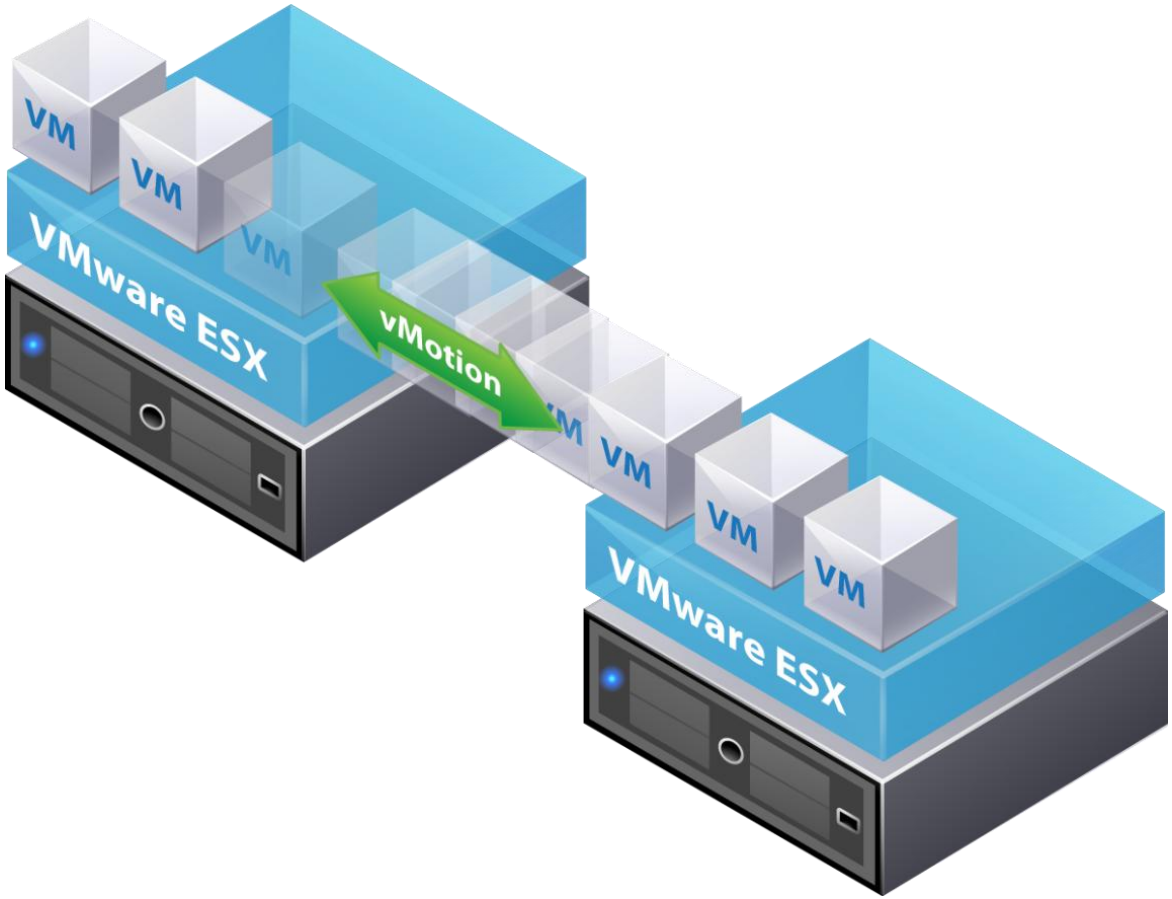
Massive Consolidation



Snapshots



Dynamic Workload Migration

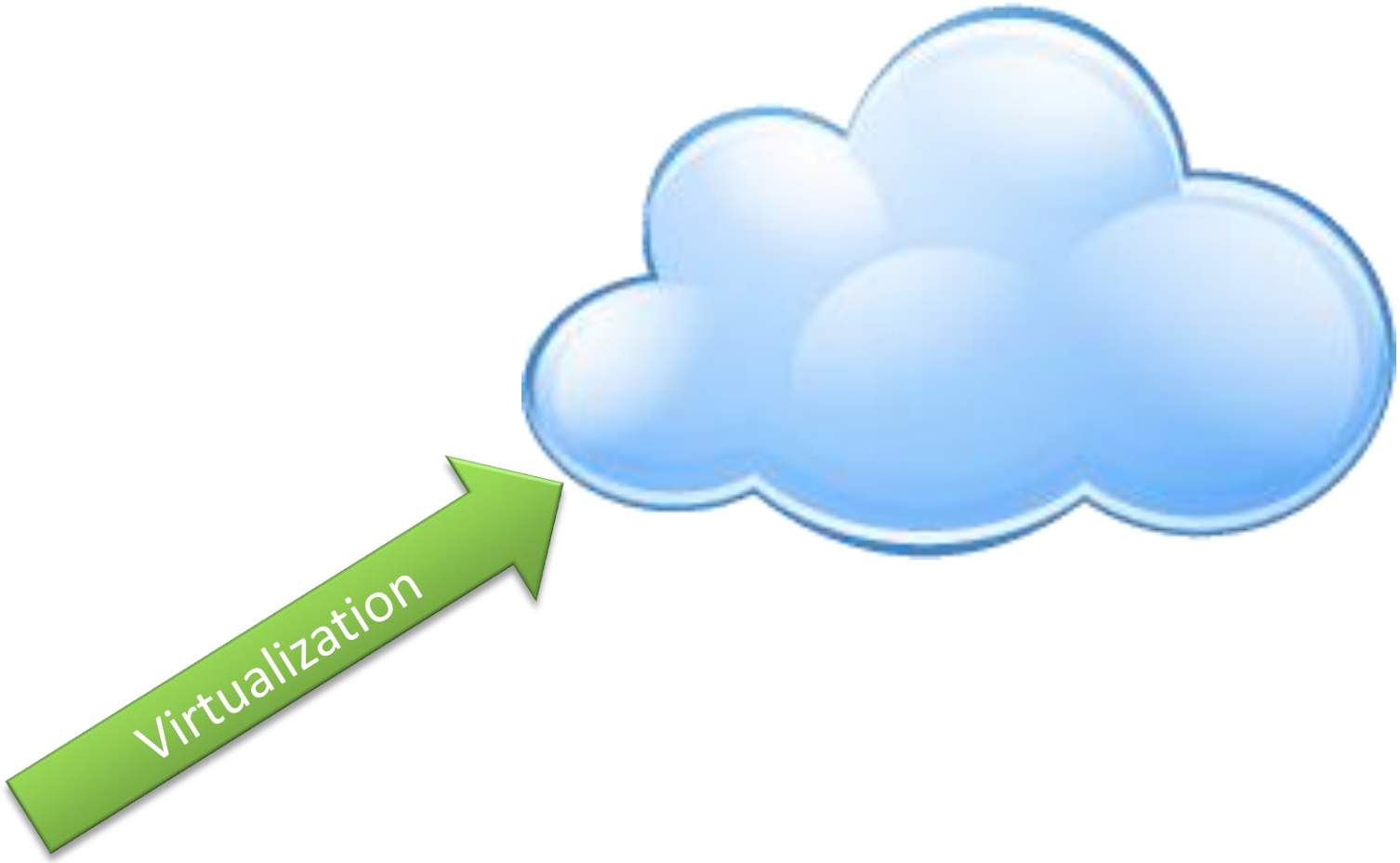


Virtualization

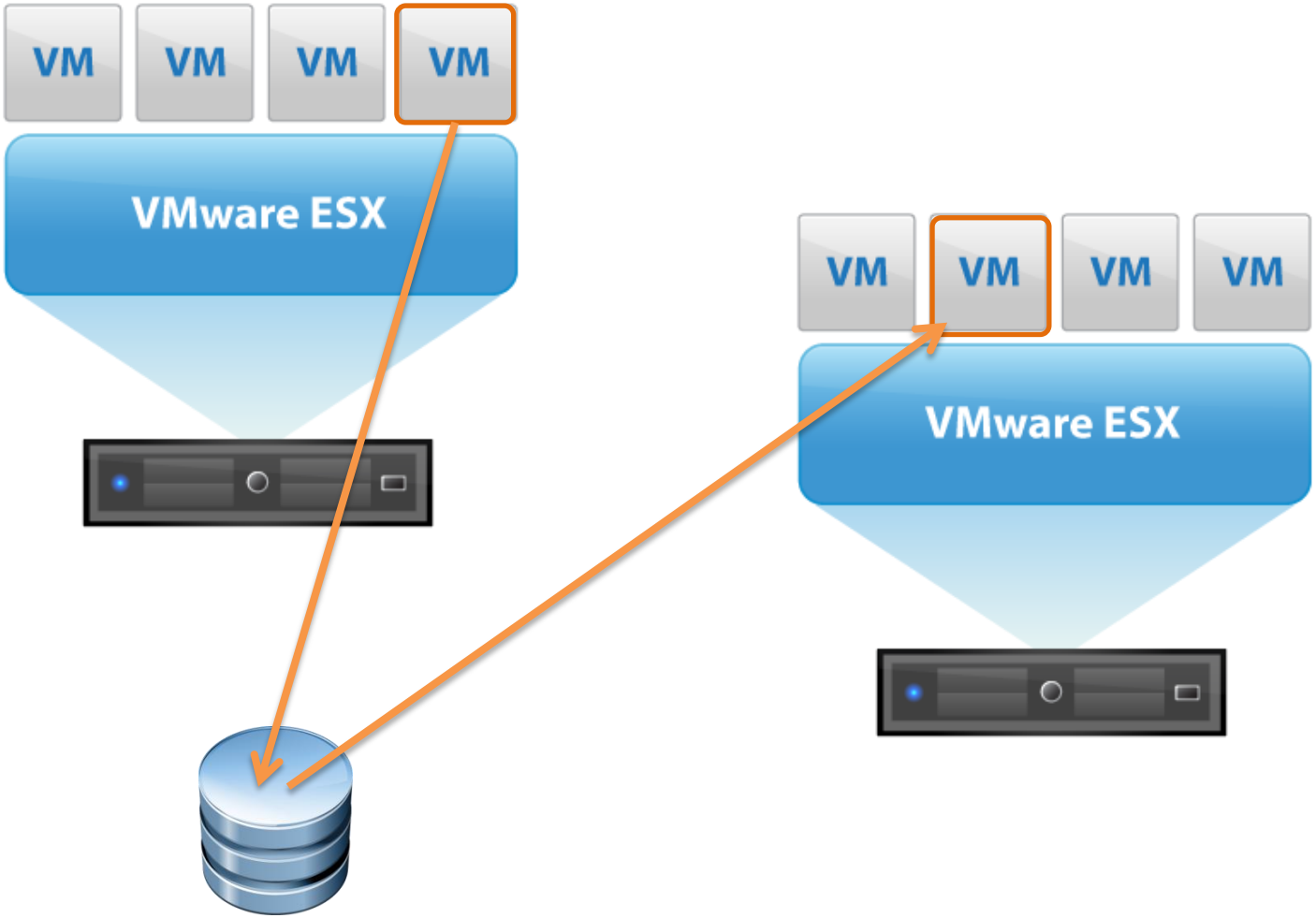
OPENFABRICS ALLIANCE

How Can We Use It?

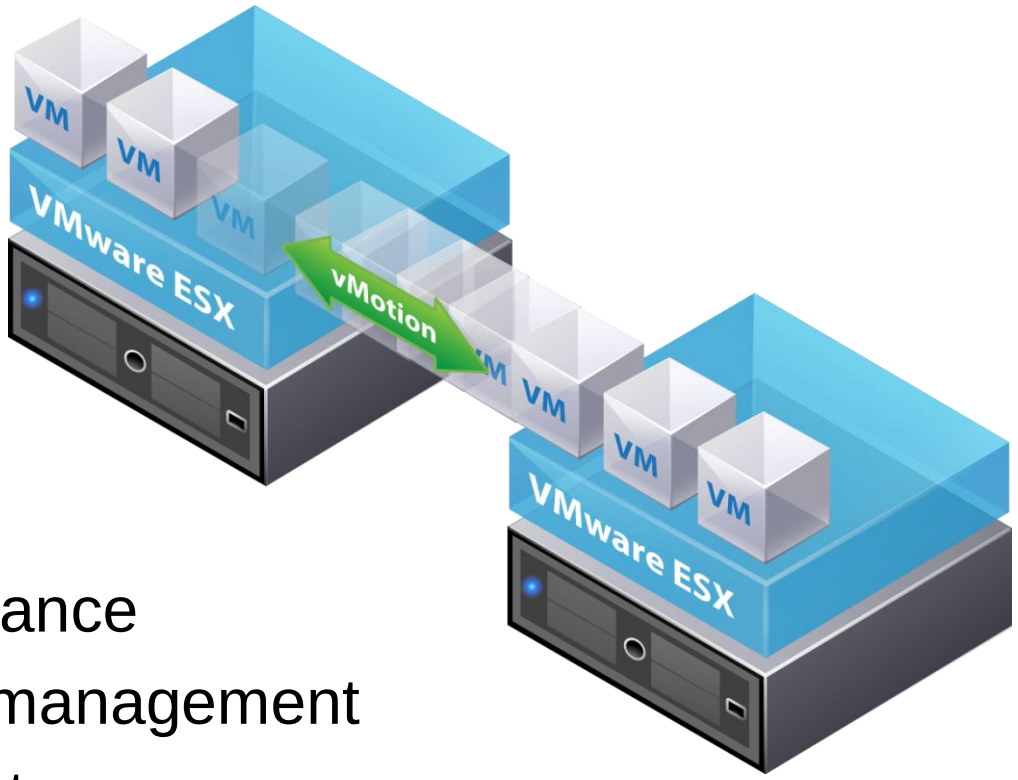
Cloud Computing



Reactive Fault Tolerance: Checkpoint / Restart

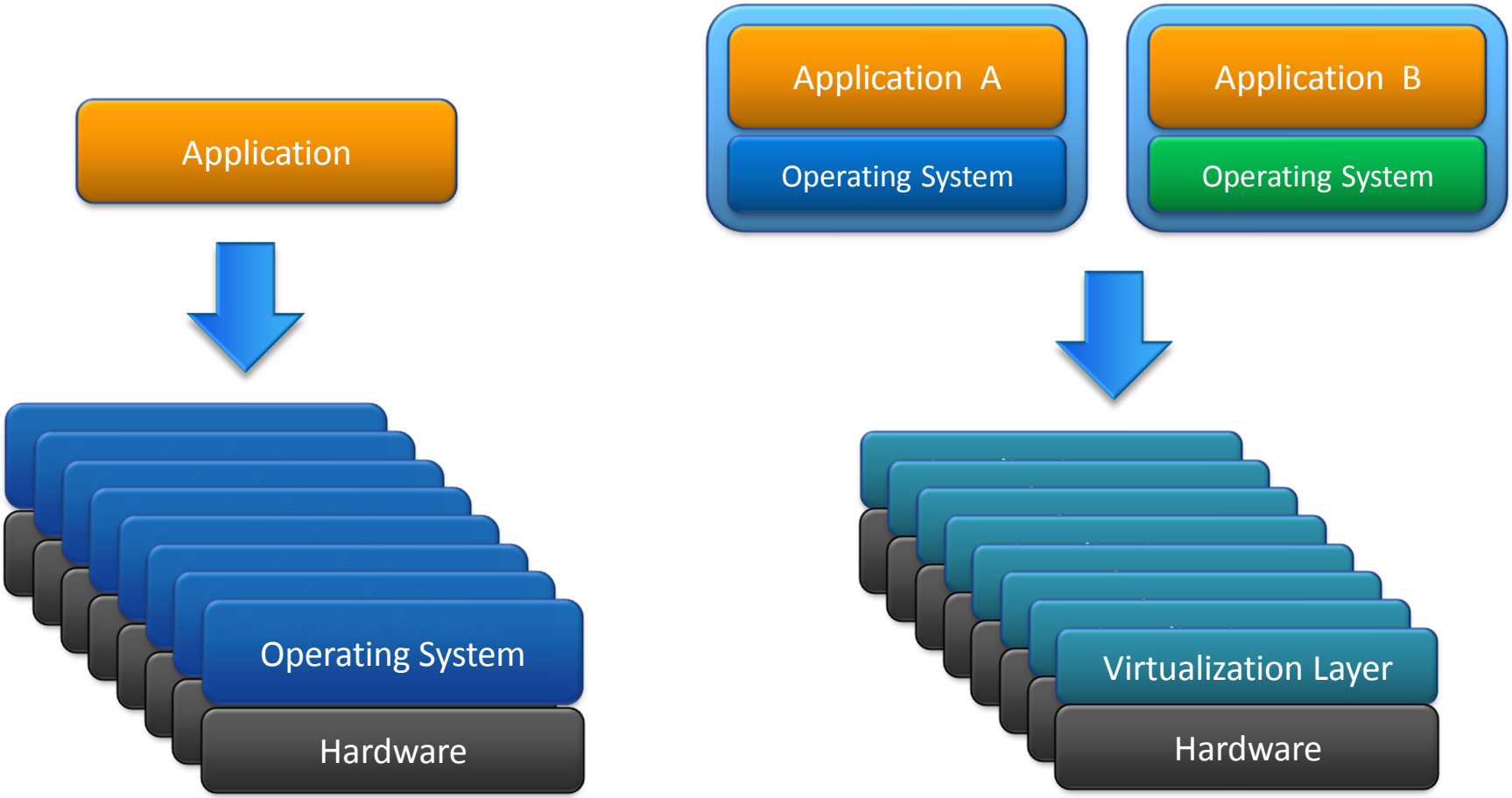


Dynamic Workload Migration

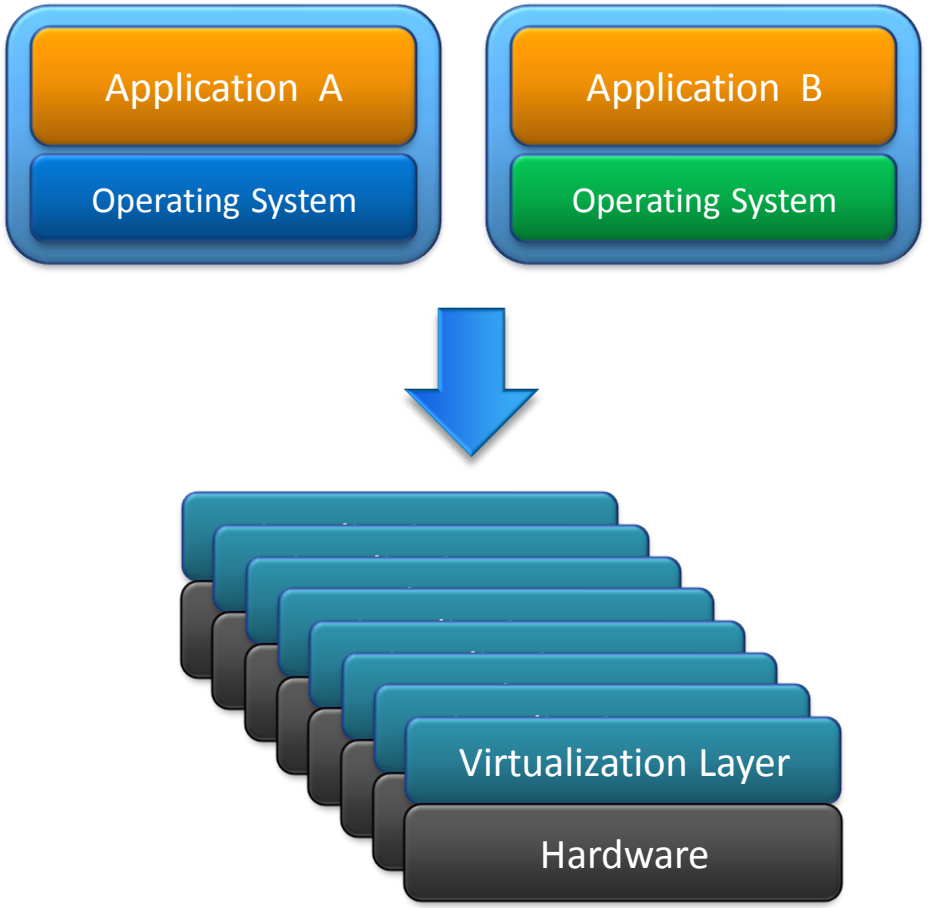


1. Proactive fault tolerance
2. Dynamic resource management
3. Power management

Heterogeneity for End Users and ISVs

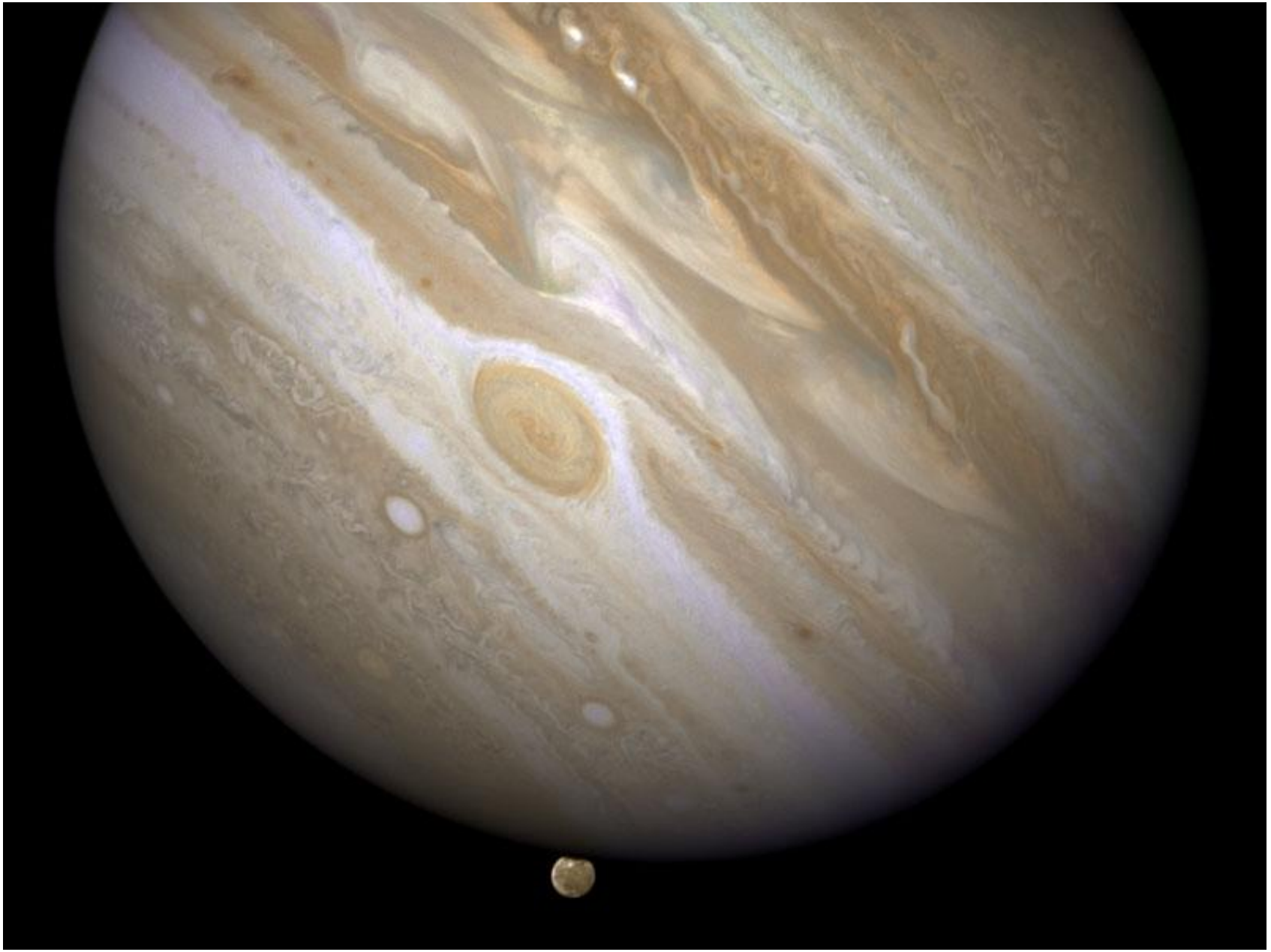


Clean Computing



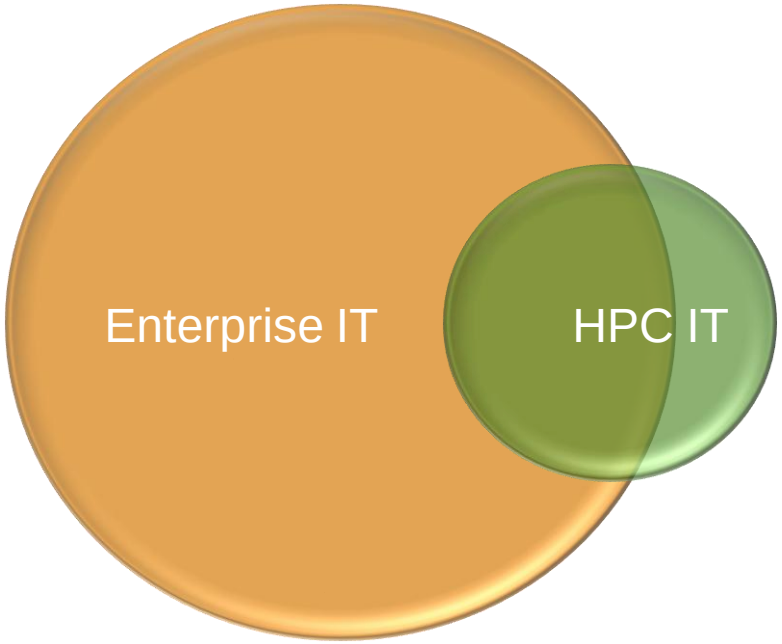
1. Pristine
2. Consistent
3. Isolated
4. Secure

IT Today



(NASA)

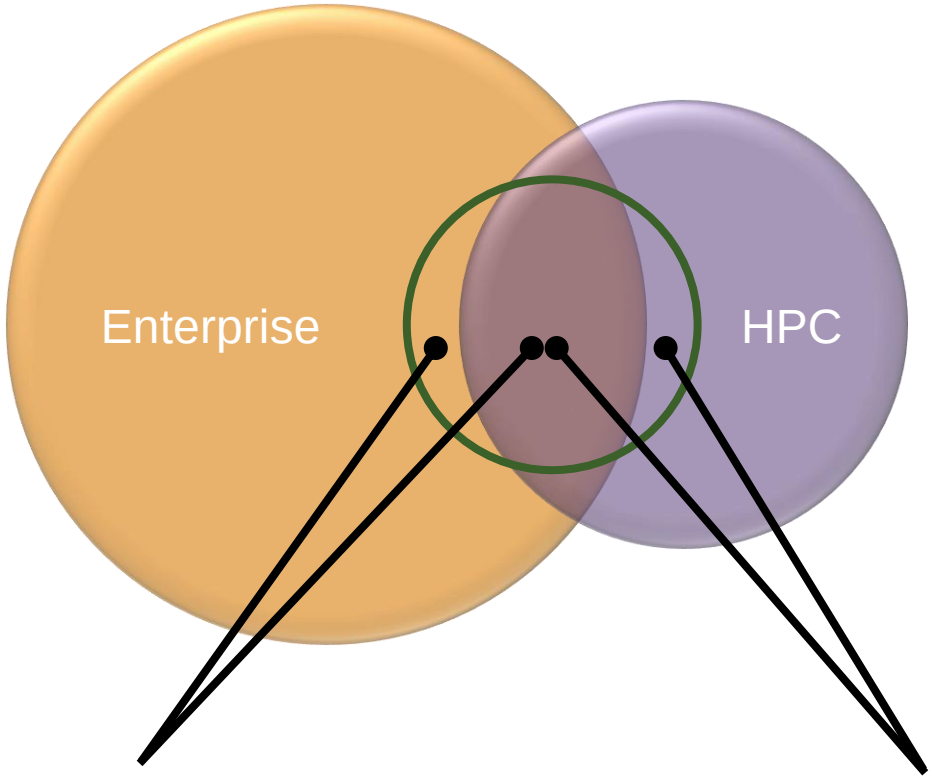
Future IT



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

- Scale-out management
- Power & cooling costs
- Dynamic resource mgmt
- Desire for high utilization
- Parallelization for multicore
- Application resiliency
- Low latency interconnect
- Cloud computing

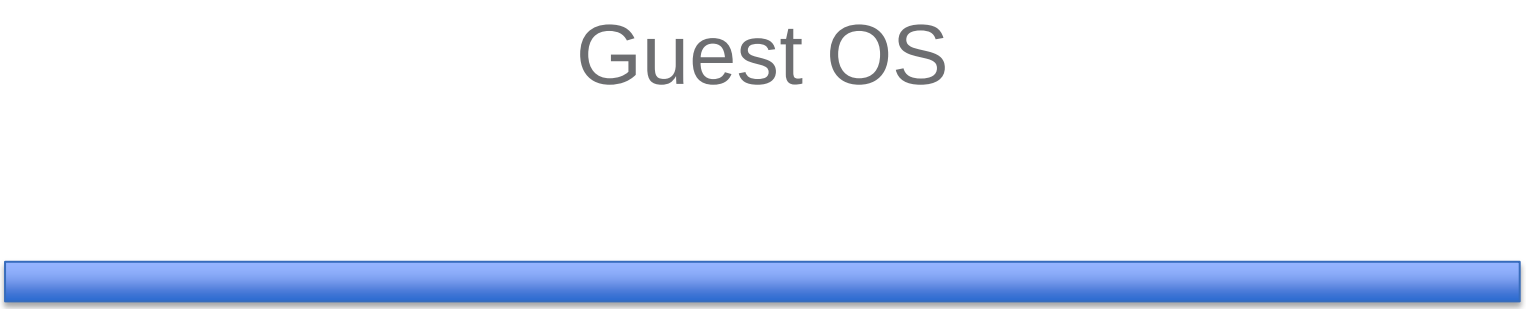
Same Problems, Different Labels



"Big Data Analytics"

"Data Intensive Computing"

RDMA

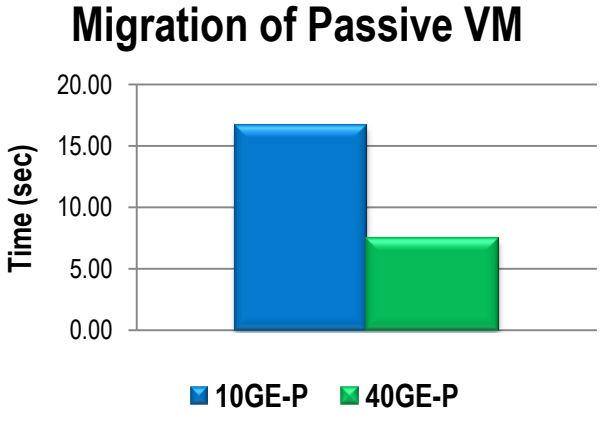


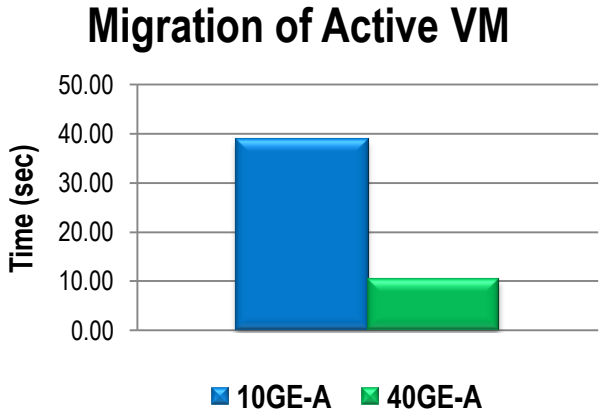
Virtual Infrastructure RDMA

- Distributed services within the platform, e.g.
  - vMotion (live migration)
  - Inter-VM state mirroring
  - Shared storage access
  - Distributed file system
- All would benefit from:
  - Decreased latency
  - Increased bandwidth
  - CPU offload

RDMA for vMotion

- *High Performance Virtual Machine Migration with RDMA over Modern Interconnects*, Huang, Gao, Liu, Panda
  - RDMA reduced total migration time by up to 80%
  - Migration downtime reduced by up to 77%
- Mellanox testing:





Guest OS RDMA

- Bandwidth and latency increasingly important
- Scale-out middleware and applications increasingly important in the Enterprise
  - memcached, redis, Cassandra, mongoDB, …
  - GemFire Data Fabric
- Big Data an important emerging workload
  - Hadoop, Hive, Pig, etc.
- And, of course, HPC -- Traditional and DIC
  - Includes many current VMware customers in EDA, Finance, Digital Content Creation, Life Sciences, etc.

Guest RDMA Approaches

- Fixed passthrough (FPT)
  - VMware DirectPath I/O
  - Allows direct access to HW from guest
  - Kills vMotion
- Mediated passthrough (MPT)
  - Passthrough + state save/restore
  - Could potentially enable RDMA + vMotion
  - Dependencies on external layers undesirable
    - Could do in MPI for HPC, but other cases less clear
- Virtual RDMA device?

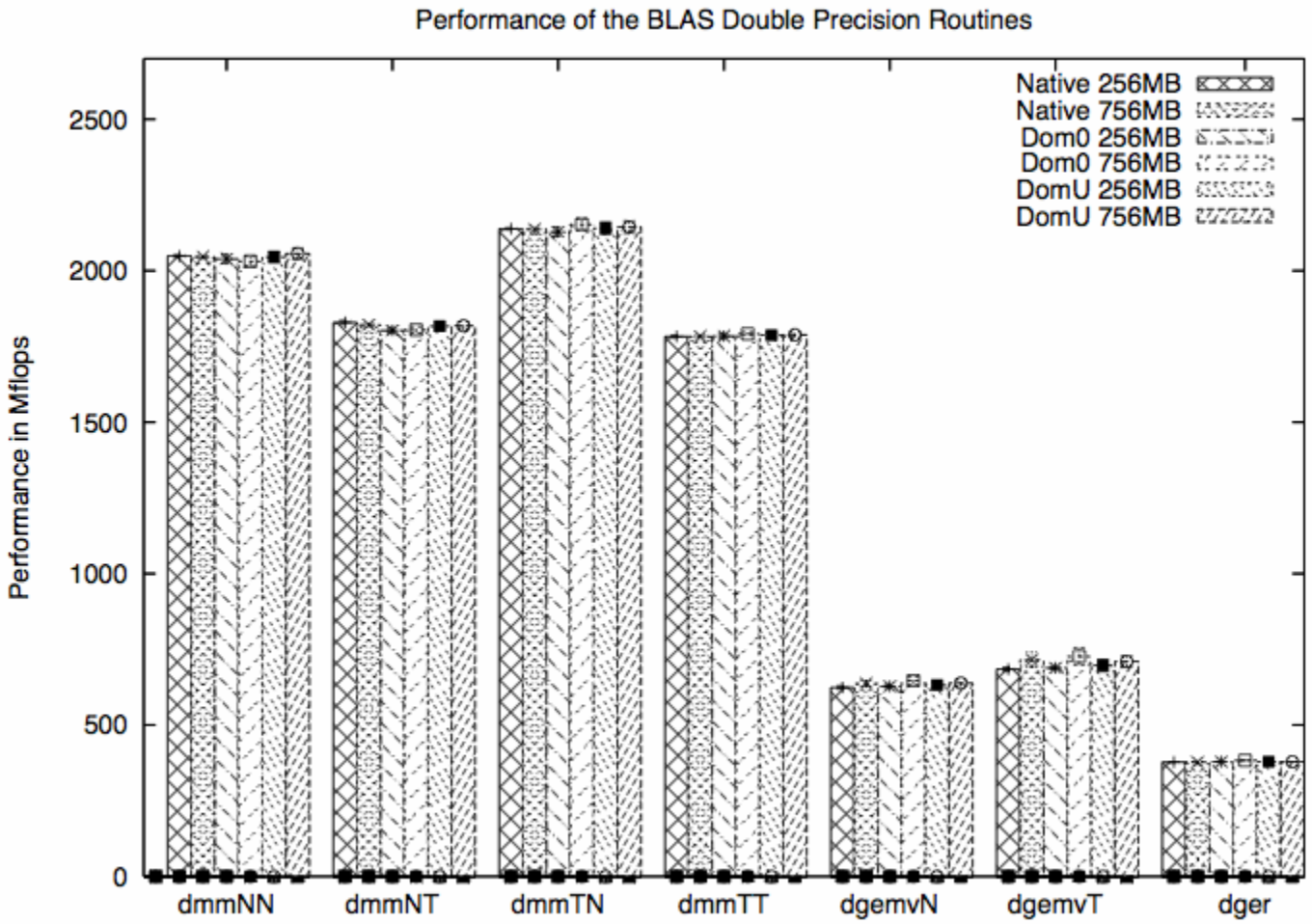
Guest RDMA Requirements

- vMotion
  - Including vMotion to hosts lacking RDMA hardware
    - Enterprise requirement, not an HPC cluster requirement
    - Benefit: Supports incremental RDMA adoption (e.g. in cloud)
- Multi-VM hardware access
- Interoperability with physical endpoints (e.g. filesystem access)
- Low latency (or lowish?)
- API
  - Verbs okay for HPC, but something nicer/easier for Enterprise (ultra-fast sockets?) highly desirable

Final Thoughts

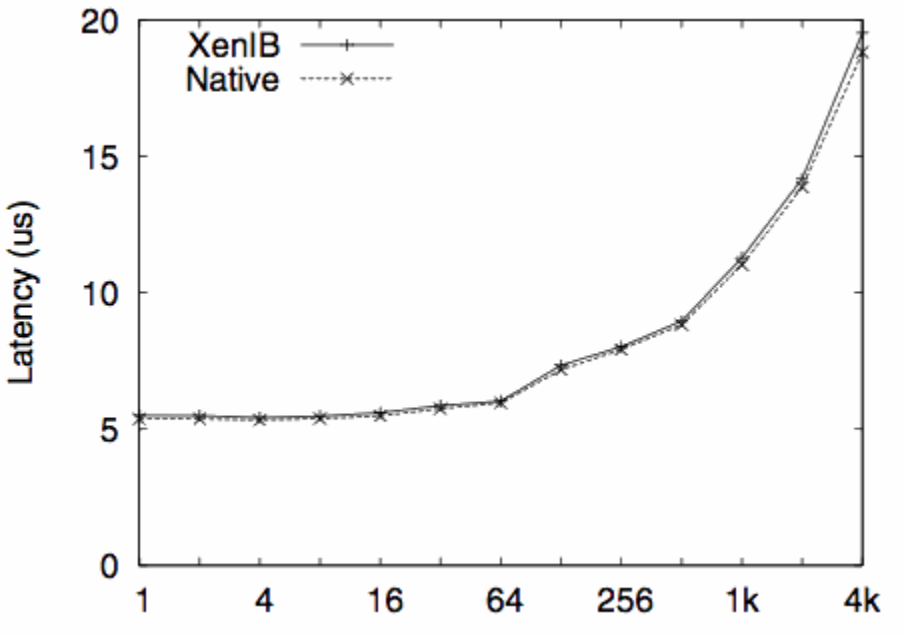
- Moore's Law, multicore, and Dodo Birds
- To Lloyd's comments…
  - Difficult to address markets different than those represented by current organization members, especially torchbearers
  - Very similar to Innovator's Dilemma
- Fit and Finish
- Compelling Enterprise OFED demos would be helpful for non-HPC business case
- Impatient for RDMA ubiquity. Go OFA!

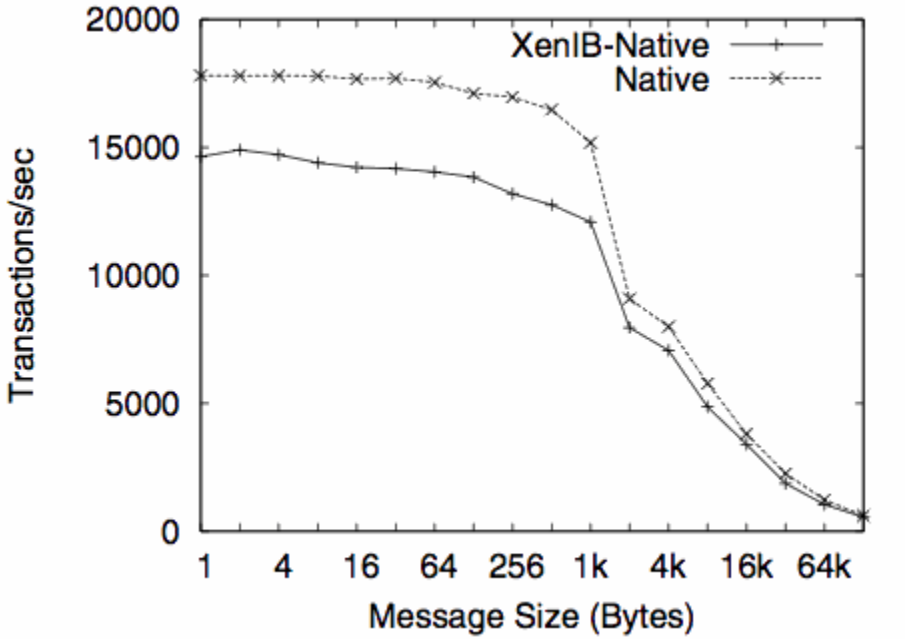
Compute Performance



The Impact of Paravirtualized Memory Hierarchy on Linear Algebra Kernels and Software, Youseff, *et al*, HPDC '08

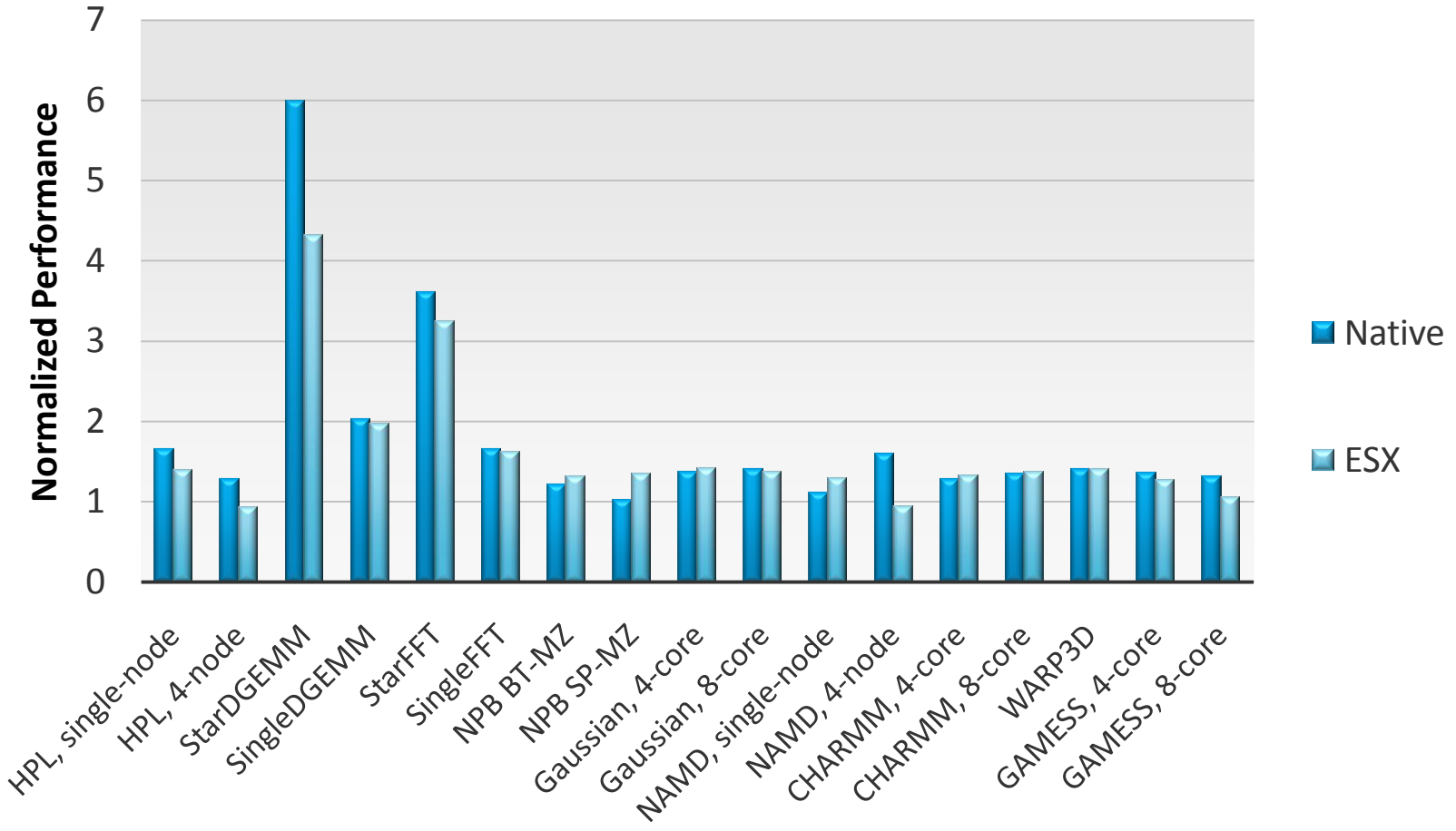
Interconnect Performance





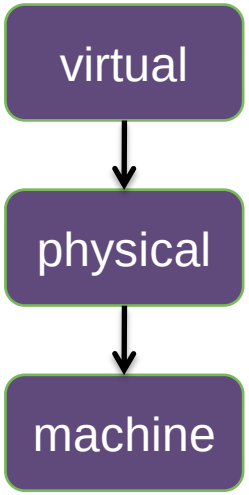
High Performance VMM-Bypass I/O in Virtual Machines, Liu *et al*, USENIX '06

Application Performance



Vblock HPC prototype, Purdue University

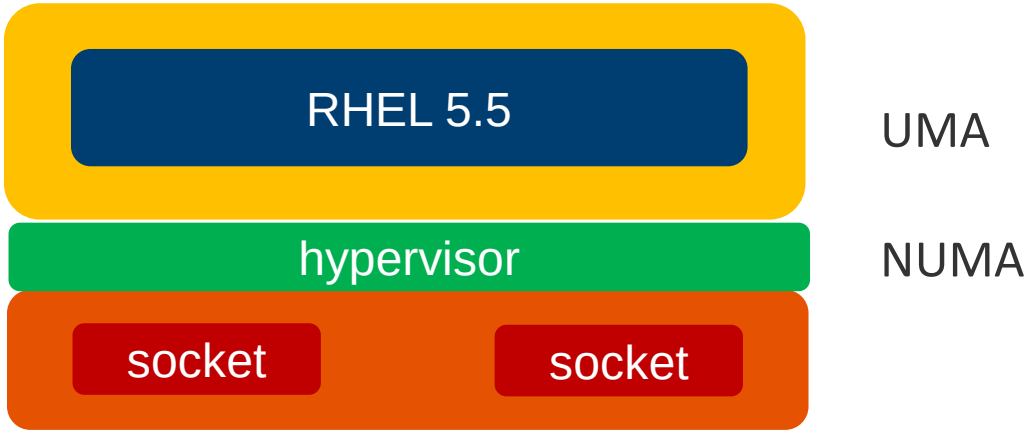
Case Study: Memory Virtualization



HPL	Native	Virtual	
		EPT on	EPT off
Small pages	37.04	36.04 (97.3%)	36.22 (97.8%)
Large pages	37.74	38.24 (100.1%)	38.42 (100.2%)

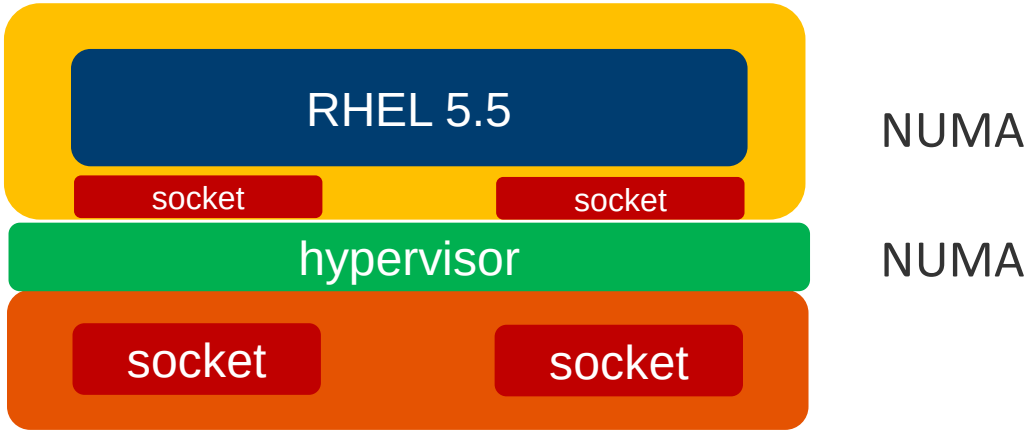
StarRA	Native	Virtual	
		EPT on	EPT off
Small pages	0.01842	0.01561 (84.8%)	0.01811 (98.3%)
Large pages	0.03956	0.03805 (96.2%)	0.03900 (98.6%)

Case Study: NUMA



- STREAM benchmark (memory bandwidth)
- Initially 26% slower virtualized relative to bare metal

Case Study: NUMA



- STREAM benchmark (memory bandwidth)
- Initially 26% slower virtualized relative to bare metal
- With an internal version of ESXi that exposes NUMA to guest
  - Performance disparity eliminated