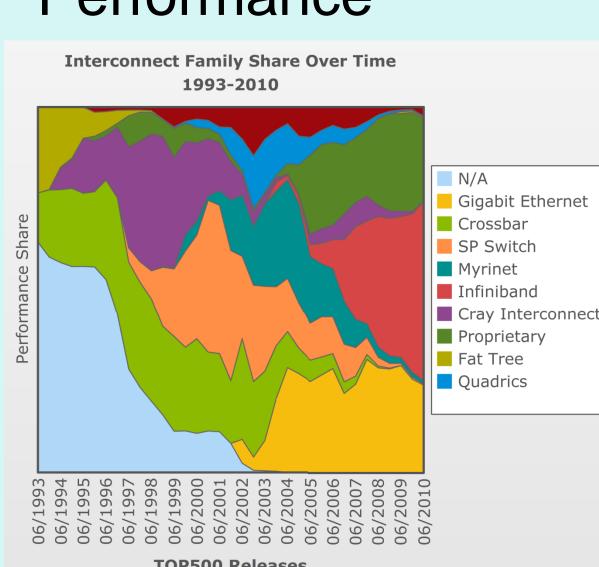
#### Motivation

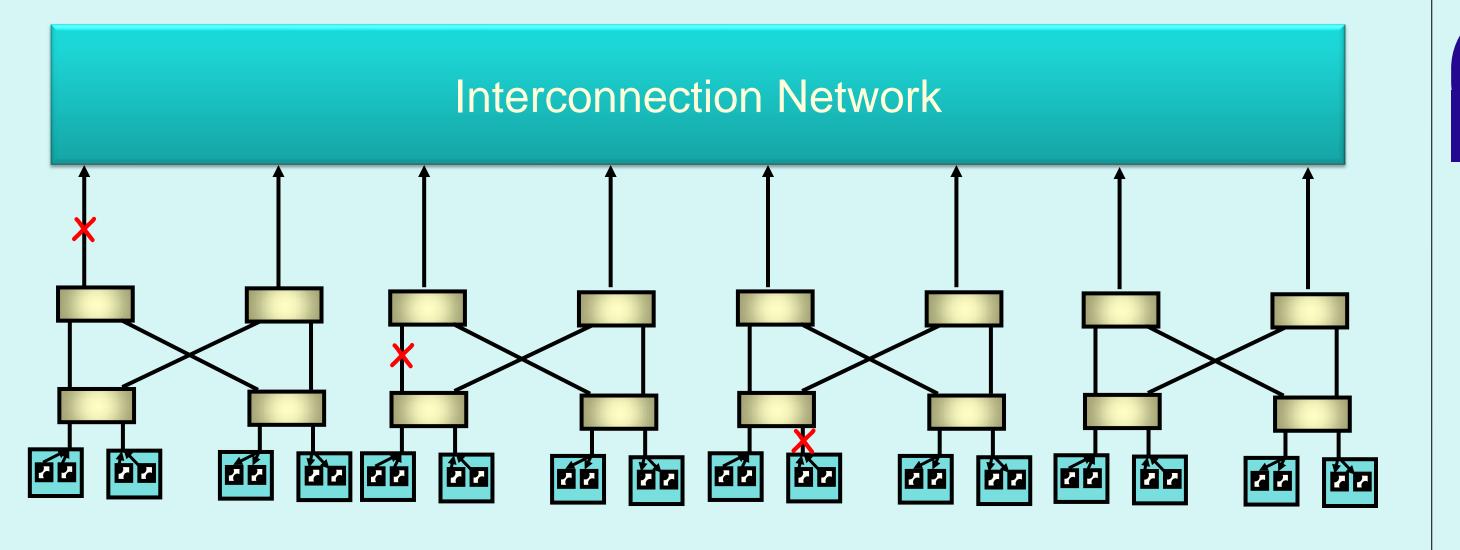
# Systems

# Gigabit Ethernet

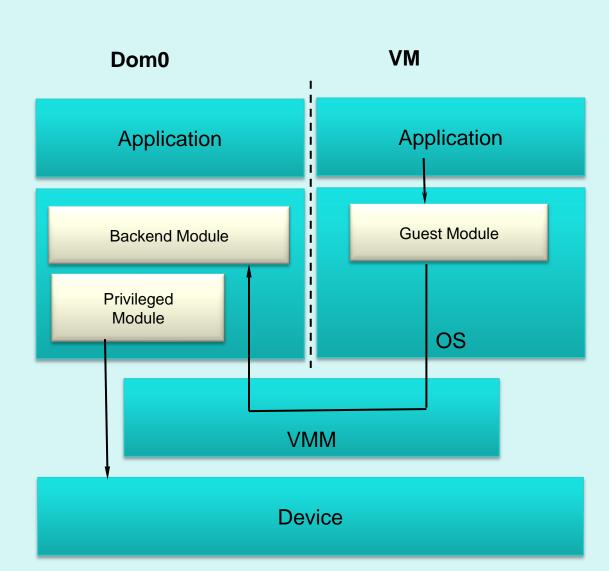
#### Performance



#### **Network Faults**

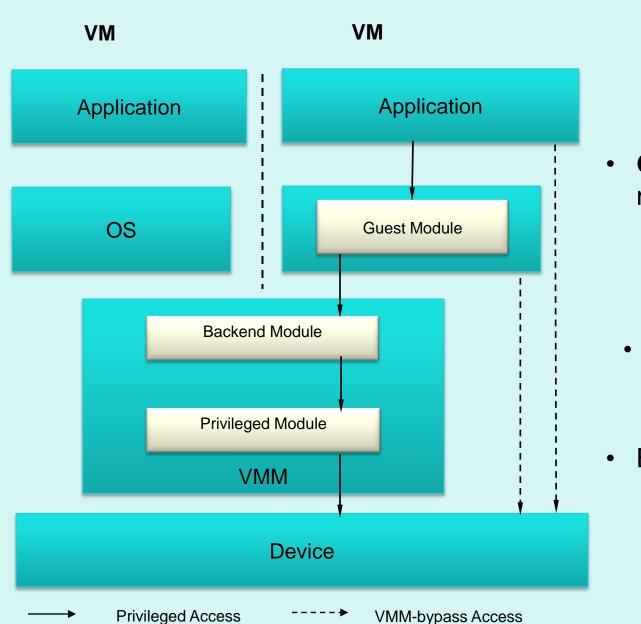


#### I/O Virtualization



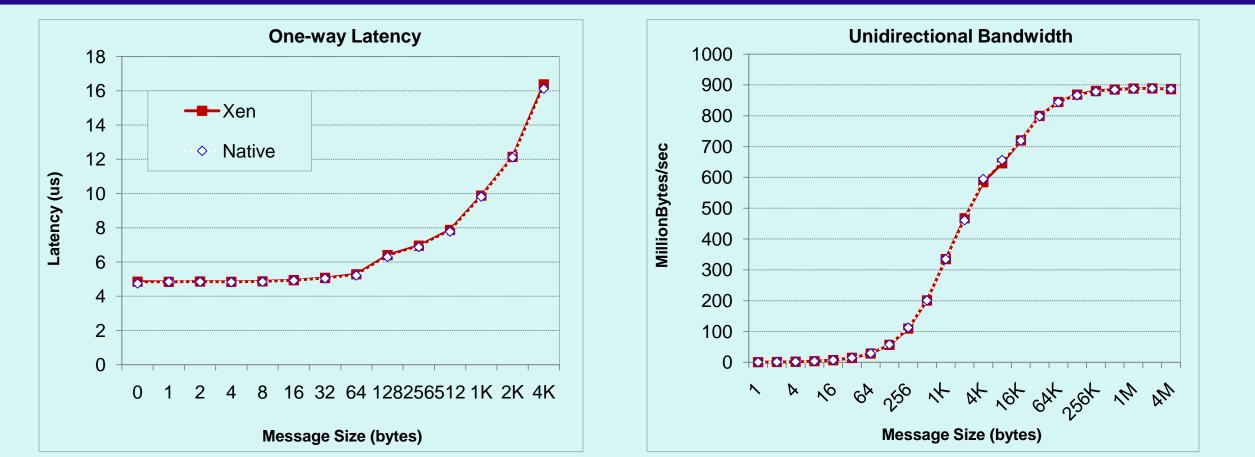
- -I/O in VMM (e.g. VMware ESX)
  - Device drivers hosted in the VMM
  - I/O operations always trap into the VMM - Ensures safe device sharing among VMs
- -I/O in a special VM
- Device drivers are hosted in a special(privileged) VM - I/O ops always involve the VMM and the special VM - E.g.: Xen and VMware Workstation

#### OS-bypass to VMM-bypass



- Guest modules in guest VMs handle setup and management operations (privileged access)
- Guest modules communicate with VMM backend modules to get jobs done
- Original privileged module can be reused Once setup, devices are accessed directly from guest VMs
- (VMM-bypass)
- Either from OS kernel or applications Backend and privileged modules can reside in a special VM

## Near-native Performance with Xen



- Only VMM-bypass operations are used (MVAPICH implementation)

– J. Liu, W. Huang, B. Abali, D. K. Panda. "High Performance VMM-Bypass I/O in Virtual Machines", USENIX '06

- W. Huang, J. Liu, B. Abali, D. K. Panda. "A Case for High Performance Computing with Virtual Machines", ICS '06

- Xen-IB performs similar to native InfiniBand

# Reduced Performance Overhead and Live Migration using Virtualization for HPC

D. K. Panda **The Ohio State University** 

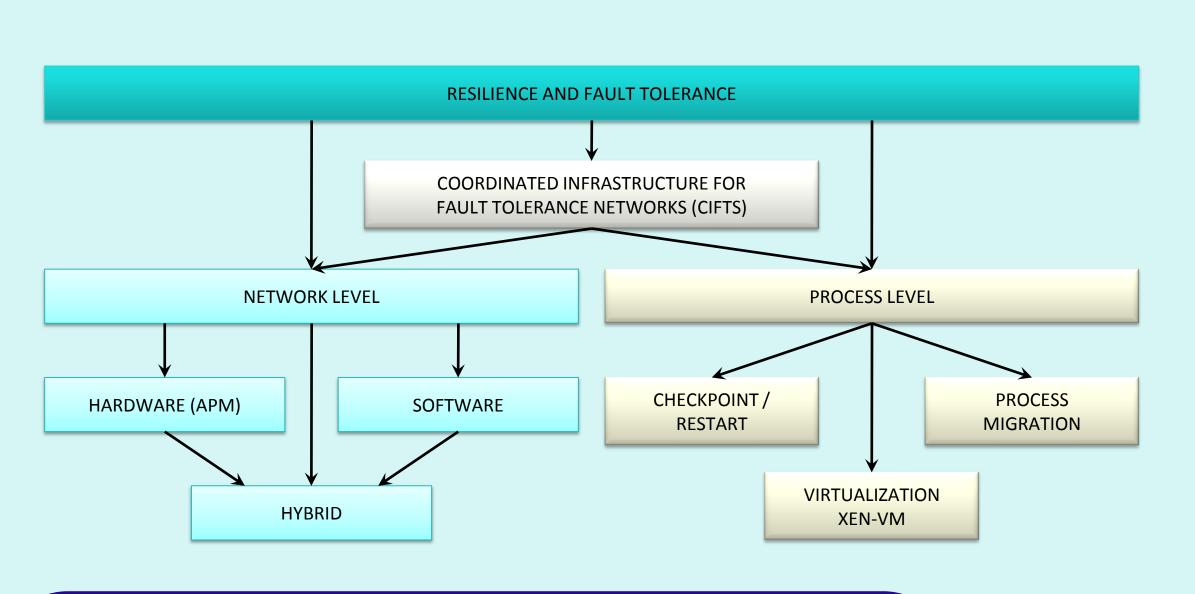
#### Overview

Jevice Manager an

Control Software

**Jnmodified User** 

Software



- High Performance MPI Library for IB and 10GE
- MVAPICH (MPI-1) and MVAPICH2 (MPI-2)
- Used by more than 1275 organizations in 60 countries
- More than 45,000 downloads from OSU site directly
- Empowering many TOP500 clusters
  - 6th, 81,920-core (Pleiades) at NASA
- Available with software stacks of many IB, 10GE and server vendors including Open Fabrics Enterprise Distribution (OFED)

**Computing Process** 

**IVC Kernel Drive** 

Shared Memory Pages

**IVC Kernel Drive** 

VM2

Computing Process 2

**IVC Kernel Driver** 

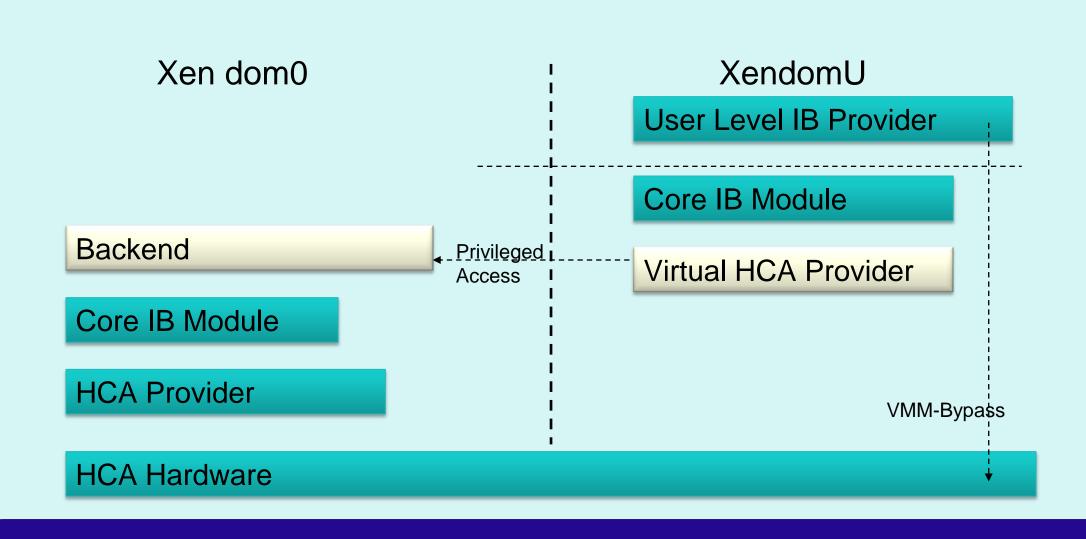
Address Space

IVC Kernel Driver

Xen Virtual Machine Monitor

http://mvapich.cse.ohio-state.edu/

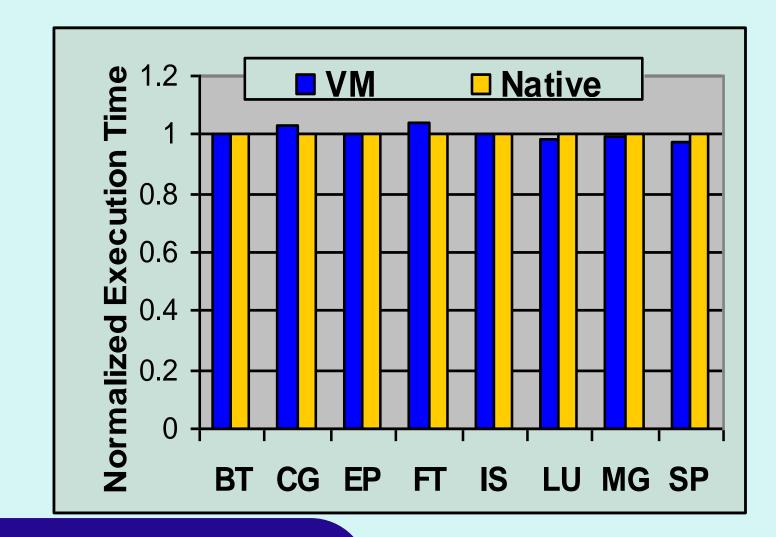
#### Virtualization-Xen-IB



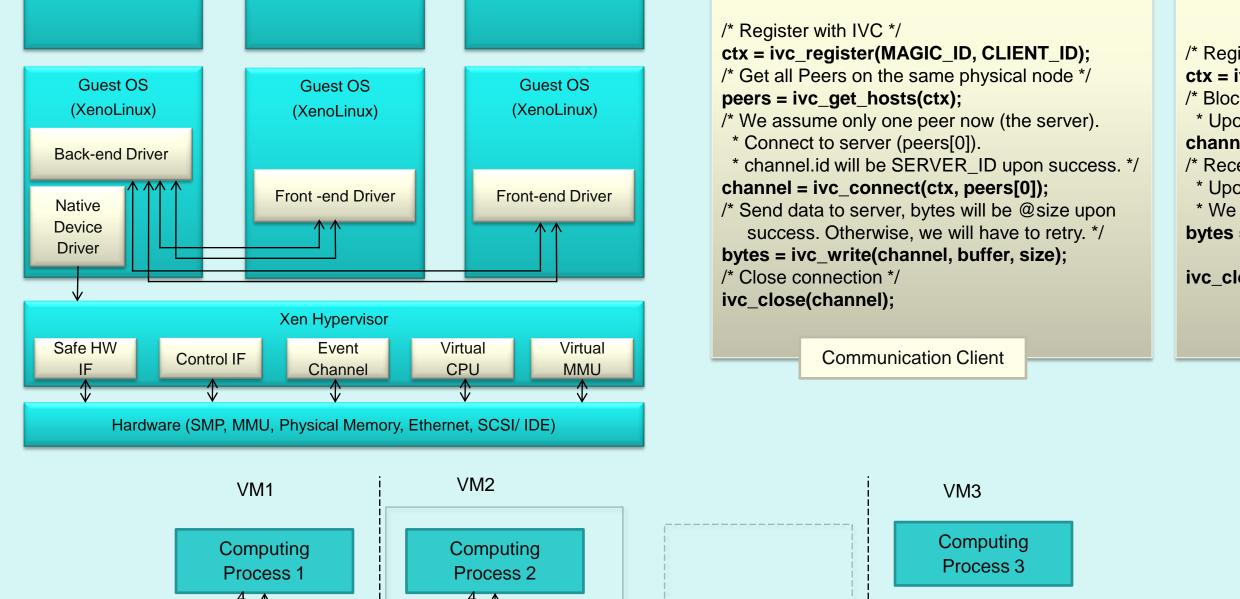
(Guest Domain)

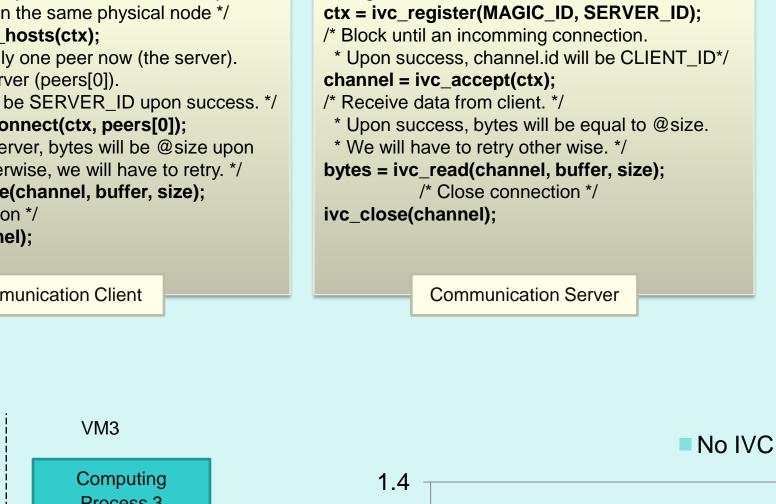
Inmodified User

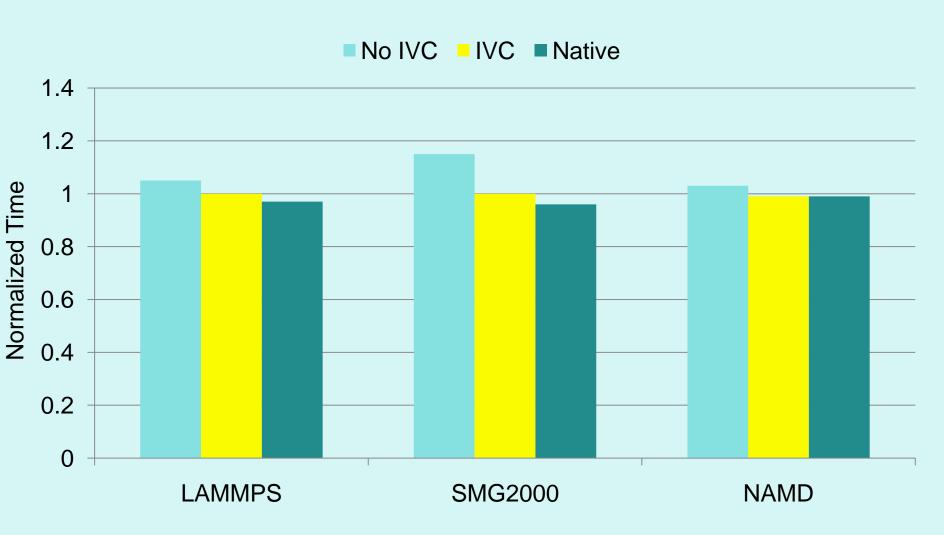
Software



#### VM aware Communication Libraries for HPC



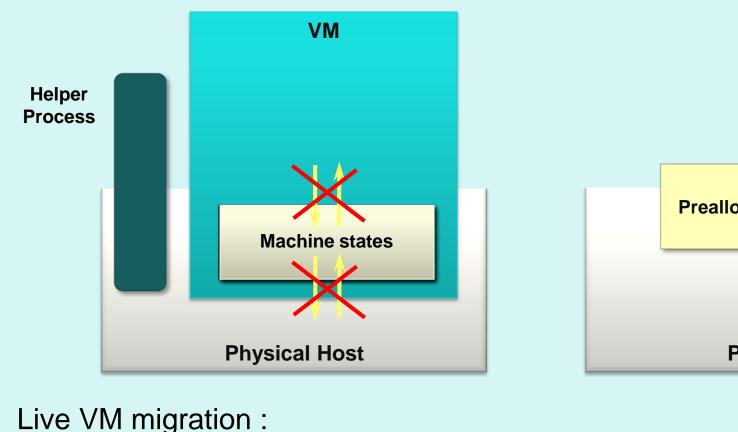




#### VM Migration through RDMA

IVC Library and

Kernel Driver



Step 4: Restart VM on the new host

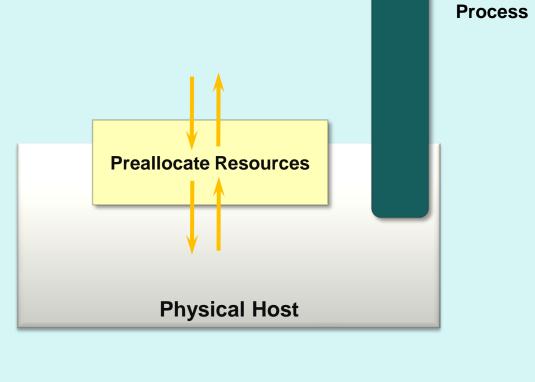
Step 1: Pre-allocate resources on the target host

Step 2: Pre-copy machine states for multiple iterations

Step 3: Suspend VM and copy the latest updates to machine states

Xen Virtal Machine Monitor

Physical Node 1



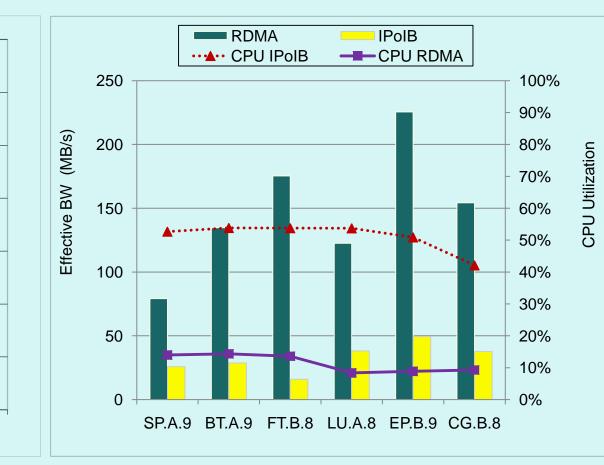
IVC Library and

Kernel Driver

Xen Virtal Machine Monitor

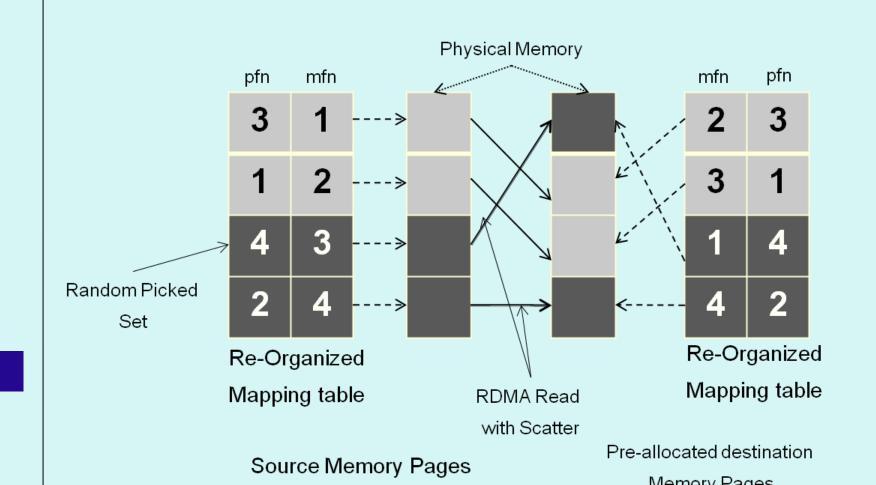
Physical Node 2



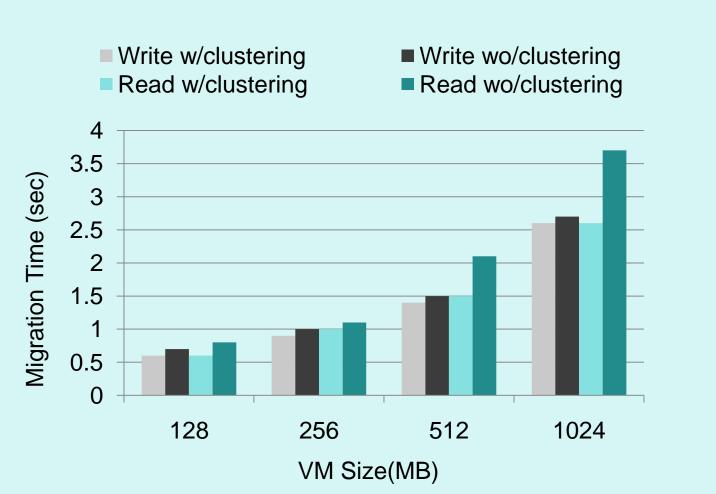


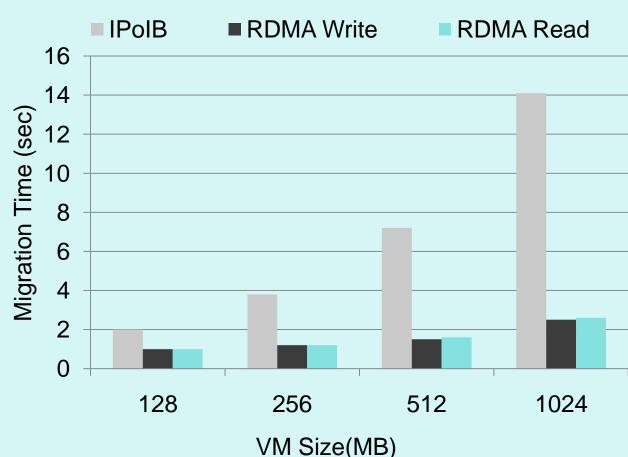
W. Huang, Q. Gao, J. Liu, D. K. Panda. "High Performance Virtual Machine Migration with RDMA over Modern Interconnects", Cluster '07 (Selected as a Best Paper)

# VM Migration with Page-Clustering

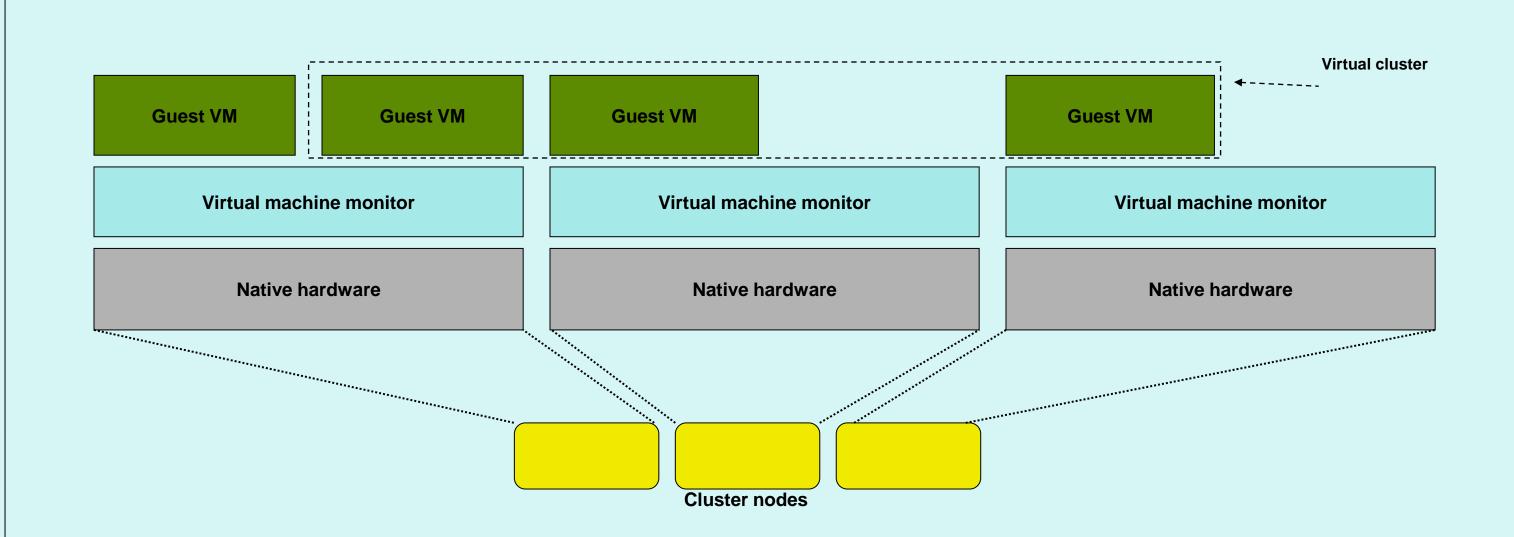


- Improves network utilization
- Reduce non-contiguous transfer
- Xen randomly picks pages to transfer
- Novelty: Page-clustering
- Transfer several pages at once by 'clustering' pages together





#### Issues with In-band Management on VM



- Root access on a single VM would allow a user to control the subnet manager and subnet agents on all IB components
  - Virtualization does not help, since each "virtual network adapter" is connected to the entire network
- Subnet agents do not require any authentication from the subnet manager
- Other networks such as Ethernet only support out-of-band management
  - More secure since who gets access to the management network can be restricted
  - User-defined authentication possible on out-of-band interfaces

### Publications

- W. Huang, M. Koop and D.K. Panda, Efficient One-Copy MPI Shared Memory Communication in Virtual Machines, IEEE Cluster 2008, September 2008
- W. Huang, M. Koop, Q. Gao and D. K. Panda, Virtual Machine Aware Communication Libraries for High Performance Computing, Supercomputing (SC), 2007
- W. Huang, Q. Gao, J. Liu and D. K. Panda. High Performance Virtual Machine Migration with RDMA over Modern Interconnects. IEEE Conference on Cluster Computing (Cluster'07), September 2007 (Best Paper Award)
- W. Huang, J. Liu, M. Koop B. Abali and D. K. Panda, Nomad: Migrating OS-Bypass Networks in Virtual Machines, The 3<sup>rd</sup> ACM/USENIX Conference on Virtual Execution Environments (VEE), 2007
- J. Liu, W. Huang, B. Abali, D. K. Panda. High Performance VMM-Bypass I/O in Virtual Machines, USENIX Annual Technical Conference (USENIX'06), May, 2006
- W. Huang, J. Liu, B. Abali, D. K. Panda. A Case for High Performance Computing with Virtual Machines, ACM International Conference on Supercomputing (ICS '06), June, 2006

#### Acknowledgements & Collaborations









The IB support package for XEN Hypervisor can be downloaded from: http://www.mellanox.com/content/pages.php?pg=xen\_ib\_drivers http://www.openfabrics.org/