HPC with Virtual Machines: Experiences with Xen, InfiniBand and MPI

Dhabaleswar K. (DK) Panda
Department of Computer Science and Engineering
The Ohio State University
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda

Presentation Outline

- Introduction
- High performance I/O virtualization with InfiniBand
- Migration Support for InfiniBand
- MPI in VM environment
- Future work

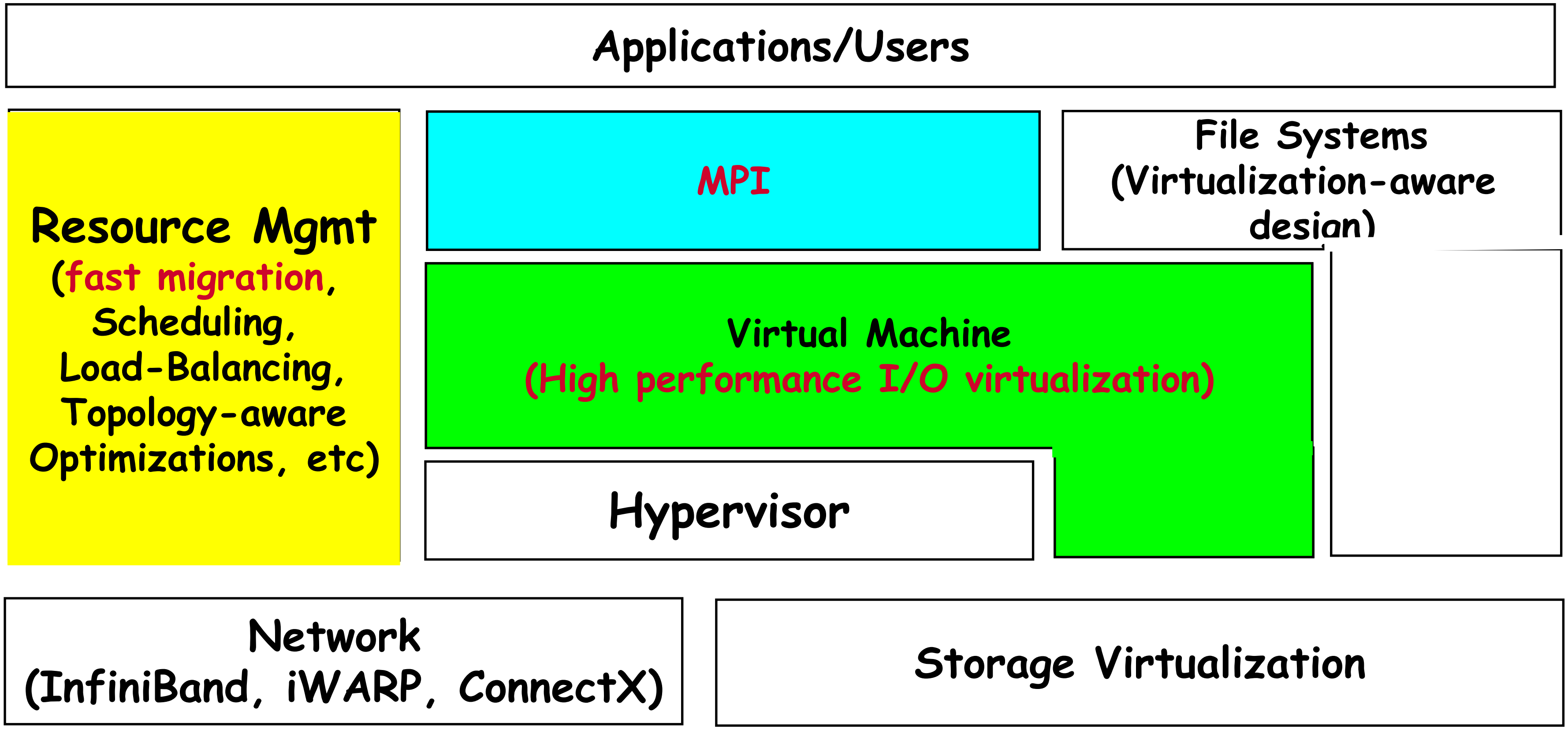
Why Target Virtualization?

- Ease of management
  - Virtualized clusters
  - VM migration – deal with system upgrade/failures
- Customized OS
  - Light-weight OS: No wide adoption due to management difficulties
  - VM makes these techniques possible
- System security & productivity
  - Users can do 'anything' in VM, in the worst case crash a VM, not the whole system

Challenges

- Performance overhead
  - CPU and memory
    - HPC applications are highly CPU intensive and spend most of the time in user space
    - Modern VM technologies achieve high performance by executing most instructions natively on host CPUs
  - I/O
    - Bigger problem since the hypervisor lies in the critical path
- Migration of modern OS-bypass network devices
- Management framework to take advantages of VM technology for HPC

Virtual Machine Based HPC: A Roadmap



Applications/Users

Resource Mgmt (fast migration, Scheduling, Load-Balancing, Topology-aware Optimizations, etc)

MPI

File Systems (Virtualization-aware design)

Virtual Machine (High performance I/O virtualization)

Hypervisor

Network (InfiniBand, iWARP, ConnectX)

Storage Virtualization

Our Recent Research Publications

- High Performance I/O virtualization with InfiniBand (VMM-bypass I/O through Xen-IB):
  - 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
- A case deployment of HPC in VM-based environment:
  - 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
- Support for migrating OS-bypass networks (extension to XenIB with Migration support):
  - W. Huang, J. Liu, M. Koop, B. Abali, D. K. Panda. Nomad: Migrating OS-bypass Networks in Virtual Machines, *The Third ACM/USENIX Conference on Virtual Execution Environment (VEE'07),* June, 2007
- High Performance VM migration and MPI Design
  - W. Huang, Q. Gao, J. Liu, D. K. Panda. High Performance Virtual Machine Migration with RDMA over Modern Interconnects, Under Review

External collaborators (J. Liu and B. Abali) from IBM T.J. Watson Research Center

Presentation Outline

- Introduction
- High performance I/O virtualization with InfiniBand
- Migration Support for InfiniBand
- MPI in VM environment
- Future work

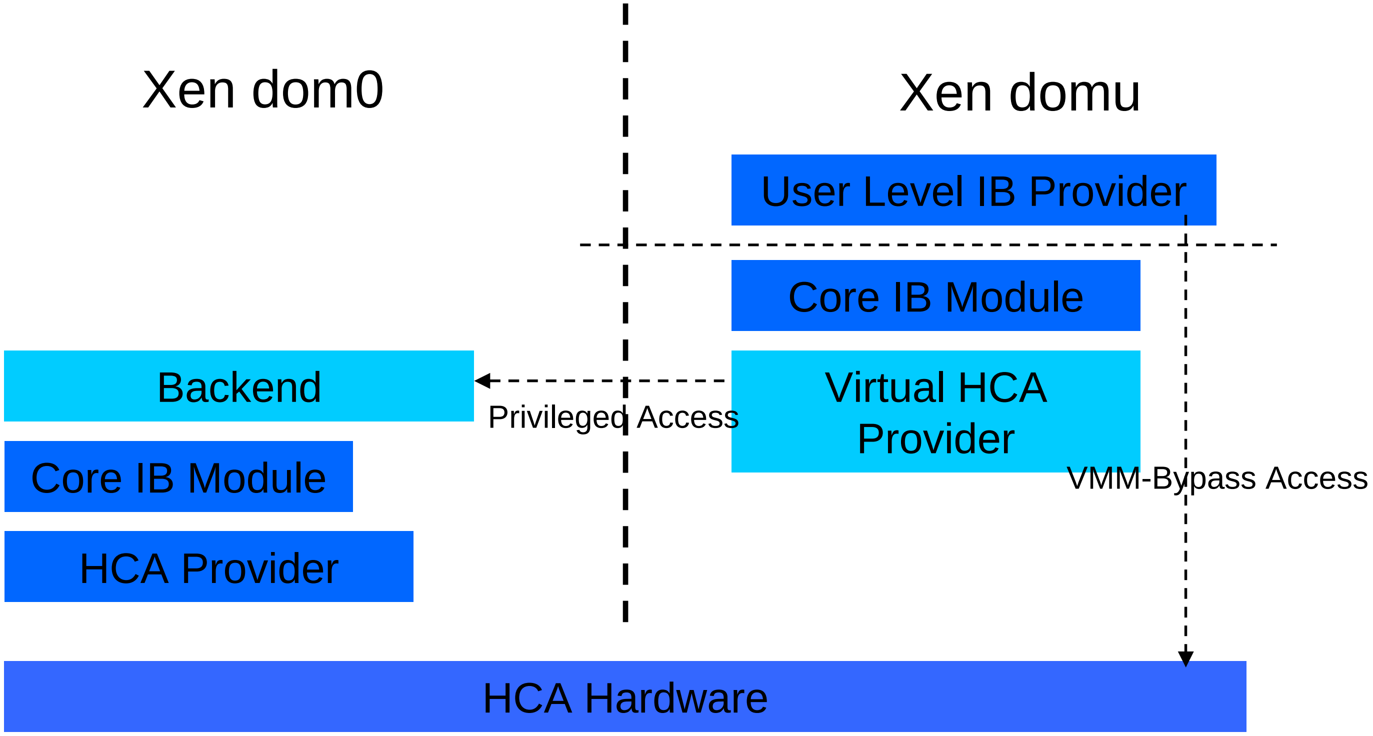
VMM-bypass I/O: Basic Ideas

- VMM-bypass
  - Direct HW access for time-critical I/O operations
  - VMM involved for setup and management
- Extending the concept of OS-bypass in the context of VM environments
  - Requires intelligent I/O adapters
- Para-virtualization
  - Does not emulate the same hardware interface in guest VMs
  - But maintains the same high-level interfaces used by OSes and applications in guest VMs

Xen-IB: InfiniBand Virtualization Driver for Xen

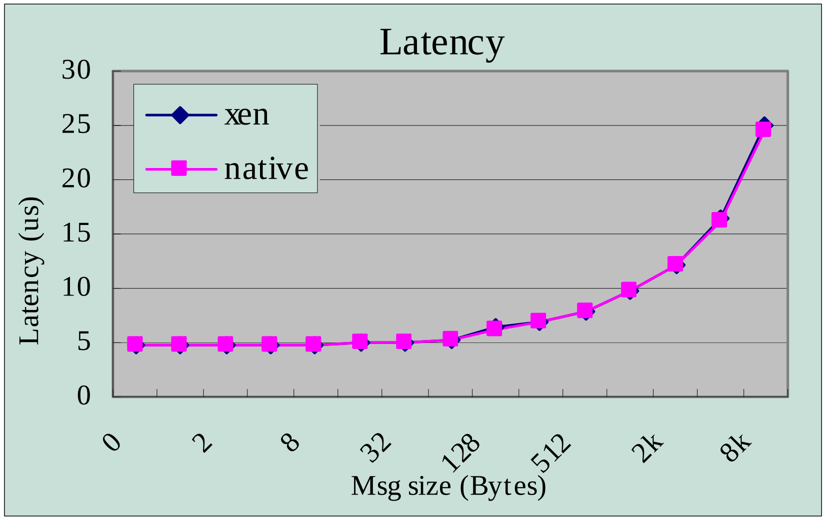
- Follows Xen split driver model
- Presents virtual HCAs to guest domains
  - Para-virtualization
- Two modes of access:
  - Privileged access
    - OS involved
    - Setup, resource management and memory management
  - OS/VMM-bypass access
    - Directly done in user space/guest VM
    - Maintains high performance of InfiniBand hardware

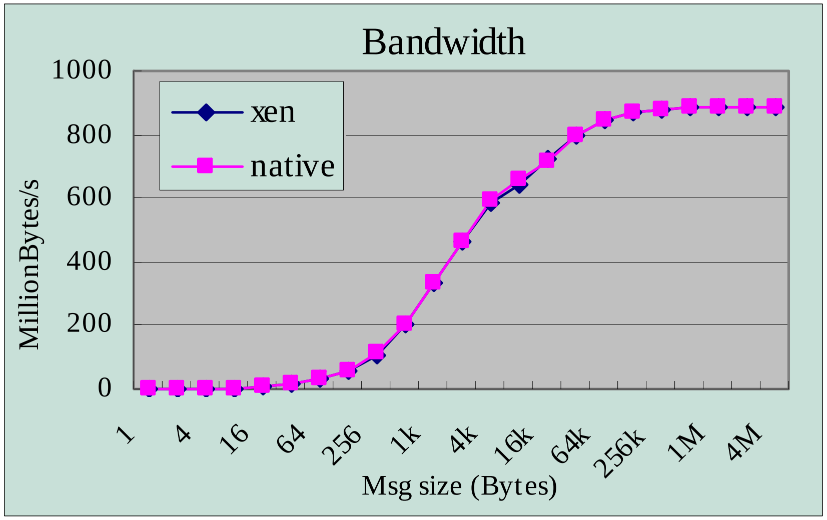
Xen-IB Basic Structure



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

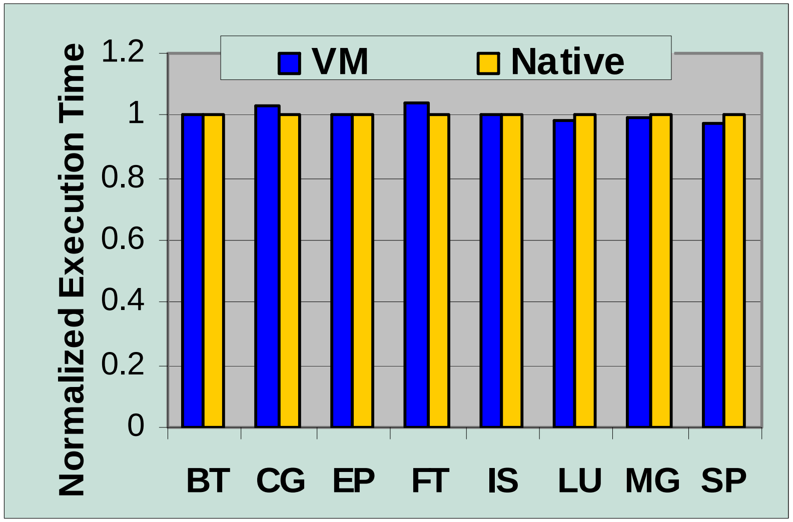
MPI Latency and Bandwidth (MVAPICH)





- Only VMM Bypass operations are used
- Xen-IB performs similar to native InfiniBand
- Numbers taken with MVAPICH-1

HPC Benchmarks (NAS)



	Dom0	VMM	DomU
BT	0.4%	0.2%	99.4%
CG	0.6%	0.3%	99.0%
EP	0.6%	0.3%	99.3%
FT	1.6%	0.5%	97.9%
IS	3.6%	1.9%	94.5%
LU	0.6%	0.3%	99.0%
MG	1.8%	1.0%	97.3%
SP	0.3%	0.1%	99.6%

- NAS Parallel Benchmarks achieve similar performance in VM and native environment (8x2)

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

Presentation Outline

- Introduction
- High performance I/O virtualization with InfiniBand
- Migration Support for InfiniBand
- MPI in VM environment
- Future work

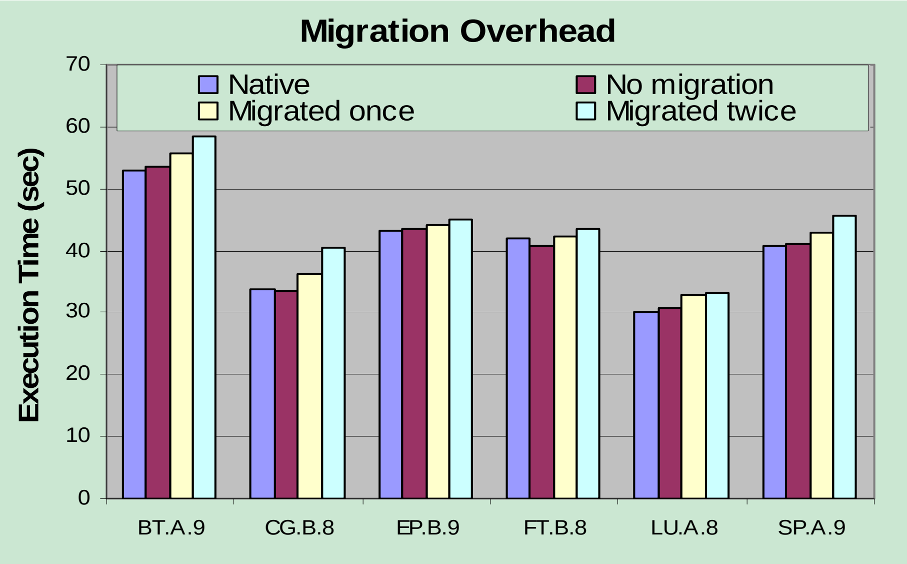
Challenges of Migrating InfiniBand

- Location dependent resources (cannot migrate with VMs):
  - LIDs, QPNs, CQNs
- User level communication:
  - Can be caching handles (memory keys, QPNs, ..) anywhere
  - Hard to suspend communication from kernel
- Hardware managed connection state:
  - Cannot easily achieve reliability during migration

Key Ideas of Nomad: Migration support for InfiniBand in VM environment

- Namespace Virtualization:
  - Virtualize all location dependent resources, such as LIDs, QPNs, CQNs, memory keys, etc.
  - Special handling for memory keys to achieve low overhead in critical path
  - Intercept communication calls at libmthca to achieve application transparency
- Coordination:
  - libmthca coordinates during migration to suspend/resume communication
  - Push QPN, LIDs, memory keys updates to connected peers

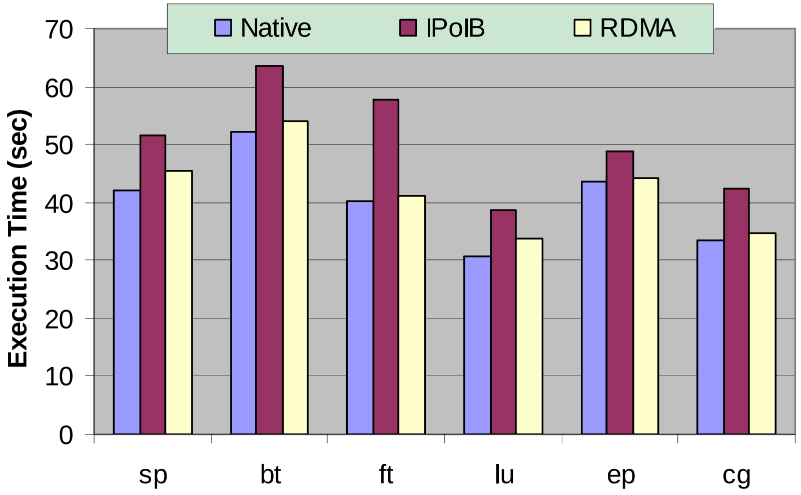
Overhead of Migration

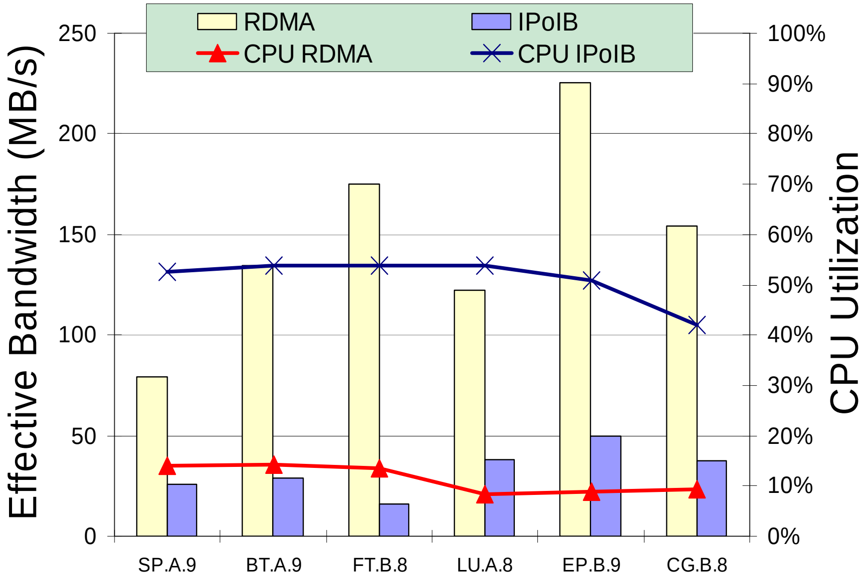


- Each migration costs 0.5 to 3 seconds, depending on the computing and communication patterns
- One process per node (dual processors) to reduce Xen overhead

W. Huang, J. Liu, M. Koop, B. Abali, D. K. Panda. Nomad: Migrating OS-bypass Networks in Virtual Machines, *The Third ACM/USENIX Conference on Virtual Execution Environment (VEE'07)*, June, 2007

Fast Migration over RDMA





- Disable one physical CPU on the nodes
- Migration overhead with IPoIB drastically increases
- RDMA achieves higher migration performance with less CPU utilization

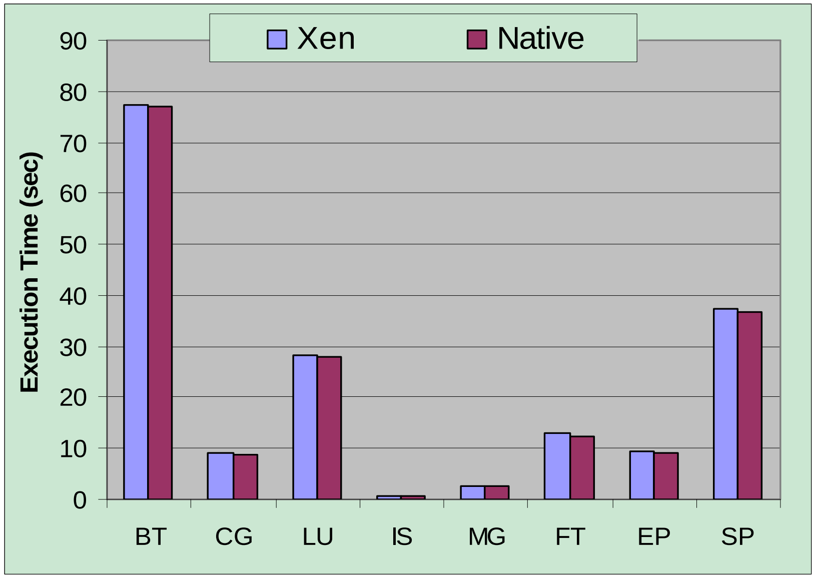
Presentation Outline

- Introduction
- High performance I/O virtualization with InfiniBand
- Migration Support for InfiniBand
- MPI in VM environment
- Future work

MPI in Virtual Machine Environment

- MPI libraries supporting OFA verbs should benefit transparently from VMM-bypass I/O and the migration support
- Extensions: allow efficient inter-VM communication

Evaluation on Larger Cluster



- Numbers taken on 64 nodes (dual processor) using NAS class C
- Overhead is marginal in most cases
- Some gap (FT, SP) is due to the optimized SMP performance of MVAPICH2. We will optimize the Xen case in future

Presentation Outline

- Introduction
- High performance I/O virtualization with InfiniBand
- Migration Support for InfiniBand
- MPI in VM environment
- Future work

Future Work

- System-level support for better virtualization
  - Fully compatible implementation with latest OFA interface (including SDP, MAD service, etc. besides user verbs)
  - Explore migration solutions exploiting hardware features (e.g. Mellanox ConnectX)
    - Achieving inter-operability with unmodified hosts
  - Enhancement to file systems to support effective image management in VM-based environment
  - Scalability studies

Web Pointers

http://nowlab.cse.ohio-state.edu/projects/xen/

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