High Performance Migration Framework for MPI Applications on HPC Cloud

Jie Zhang, Xiaoyi Lu and Dhabaleswar K. Panda
{zhanjie, luxi, panda}@cse.ohio-state.edu

Computer Science & Engineering Department,
The Ohio State University
Outline

- Introduction
- Problem Statement
- Design
- Performance Evaluation
- Conclusion & Future Work
Cloud Computing and Virtualization

- Cloud Computing focuses on maximizing the effectiveness of the shared resources
- Virtualization is the key technology for resource sharing in the Cloud
- Widely adopted in industry computing environment
- IDC Forecasts Worldwide Public IT Cloud Services spending will reach $195 billion by 2020

(Courtesy: http://www.idc.com/getdoc.jsp?containerId=prUS41669516)
Drivers of Modern HPC Cluster and Cloud Architecture

- Multi-core/Many-core technologies
- Accelerators (GPUs/Co-processors)
- Large memory nodes
- Remote Direct Memory Access (RDMA)-enabled networking (InfiniBand and RoCE)
- Single Root I/O Virtualization (SR-IOV)
Single Root I/O Virtualization (SR-IOV)

- SR-IOV is providing new opportunities to design HPC cloud with very little low overhead
- Allows a single physical device, or a Physical Function (PF), to present itself as multiple virtual devices, or Virtual Functions (VFs)
- VFs are designed based on the existing non-virtualized PFs, no need for driver change
- Each VF can be dedicated to a single VM through PCI pass-through
Building HPC Cloud with SR-IOV and InfiniBand

- High-Performance Computing (HPC) has adopted advanced interconnects and protocols
  - InfiniBand
  - 10/40/100 Gigabit Ethernet/iWARP
  - RDMA over Converged Enhanced Ethernet (RoCE)

- Very Good Performance
  - Low latency (few micro seconds)
  - High Bandwidth (200 Gb/s with HDR InfiniBand)
  - Low CPU overhead (5-10%)

- OpenFabrics software stack with IB, iWARP and RoCE interfaces are driving HPC systems
Overview of the MVAPICHT2 Project

• High Performance open-source MPI Library for InfiniBand, Omni-Path, Ethernet/iWARP, and RDMA over Converged Ethernet (RoCE)
  – MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.0), Started in 2001, First version available in 2002
  – MVAPICH2-X (MPI + PGAS), Available since 2011
  – Support for GPGPUs (MVAPICH2-GDR) and MIC (MVAPICH2-MIC), Available since 2014
  – **Support for Virtualization (MVAPICH2-Virt support VM, Docker, Singularity, etc.), Available since 2015**
  – Support for Energy-Awareness (MVAPICH2-EA), Available since 2015
  – Support for InfiniBand Network Analysis and Monitoring (OSU INAM) since 2015
  – **Used by more than 2,825 organizations in 85 countries**
  – More than 432,000 (> 0.4 million) downloads from the OSU site directly
  – Empowering many TOP500 clusters (June ‘17 ranking)
    • 1st ranked 10,649,640-core cluster (Sunway TaihuLight) at NSC, Wuxi, China
    • 15th ranked 241,108-core cluster (Pleiades) at NASA
    • 20th ranked 522,080-core cluster (Stampede) at TACC
    • 44th ranked 74,520-core cluster (Tsubame 2.5) at Tokyo Institute of Technology and many others
  – Available with software stacks of many vendors and Linux Distros (RedHat and SuSE)
  – **[http://mvapich.cse.ohio-state.edu](http://mvapich.cse.ohio-state.edu)**
  – Empowering Top500 systems for over a decade
    – System-X from Virginia Tech (3rd in Nov 2003, 2,200 processors, 12.25 TFlops)
    – Sunway TaihuLight (1st in Jun’17, 10M cores, 100 PFlops)
Application Performance on VM with MVAPICH2

- Proposed design delivers up to 43% (IS) improvement for NAS
- Proposed design brings 29%, 33%, 29% and 20% improvement for INVERSE, RAND, SINE and SPEC

• 64 Containers across 16 nodes, pining 4 Cores per Container
• Compared to Container-Def, up to 11% and 73% of execution time reduction for NAS and Graph 500
• Compared to Native, less than 9% and 5% overhead for NAS and Graph 500

Application Performance on Singularity with MVAPICH2

- 512 Processes across 32 nodes
- Less than 16% and 11% overhead for NPB and Graph500, respectively

Application Performance on Nested Virtualization Env. with MVAPICH2

- 256 processes across 64 containers on 16 nodes
- Compared with Default, enhanced-hybrid design reduces up to 16% (28,16) and 10% (LU) of execution time for Graph 500 and NAS, respectively
- Compared with 1Layer case, enhanced-hybrid design also brings up to 12% (28,16) and 6% (LU) benefit.

---

Execute Live Migration with SR-IOV Device

```
[root@sandy1:migration]$ sshh sandy3-vm1 lspci
root@sandy3-vm1's password:
00:00.0 Host bridge: Intel Corporation 440FX - 82441FX PMC [Natoma] (rev 02)
00:01.0 ISA bridge: Intel Corporation 82371SB PIIX3 ISA [Natoma/Triton II]
00:01.1 IDE interface: Intel Corporation 82371SB PIIX3 IDE [Natoma/Triton II]
00:01.2 USB controller: Intel Corporation 82371SB PIIX3 USB [Natoma/Triton II] (rev 01)
00:01.3 Bridge: Intel Corporation 82371AB/EB/MB PIIX4 ACPI (rev 03)
00:02.0 VGA compatible controller: Cirrus Logic GD 5446
00:03.0 Ethernet controller: Red Hat, Inc. Virtio network device
00:04.0 Infiniband controller: Mellanox Technologies MT27700 Family [ConnectX-4 Virtual Function]
00:05.0 Unclassified device [00Ff]: Red Hat, Inc. Virtio memory balloon
[root@sandy1:migration]$ virsh migrate --live --rdma-pin-all --migrateuri rdma://sandy3-ib sandy1-vm1 qemu://sandy3-ib/system
error: Requested operation is not valid: domain has assigned non-USB host devices
[root@sandy1:migration]`
```
Overview of Existing Migration Solutions for SR-IOV

<table>
<thead>
<tr>
<th>Platform</th>
<th>NIC</th>
<th>No Guest OS Modification</th>
<th>Device Driver Independent</th>
<th>Hypervisor Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhai, etc (Linux bonding driver)</td>
<td>Ethernet</td>
<td>N/A</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Kadav, etc (shadow driver)</td>
<td>Ethernet</td>
<td>Intel Pro/1000 gigabit NIC</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Pan, etc (CompSC)</td>
<td>Ethernet</td>
<td>Intel 82576, Intel 82599</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Guay, etc</td>
<td>InfiniBand</td>
<td>Mellanox ConnectX2 QDR HCA</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Han</td>
<td>Ethernet</td>
<td>Huawei smart NIC</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Xu, etc (SRVM)</td>
<td>Ethernet</td>
<td>Intel 82599</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Can we have a hypervisor-independent and device driver-independent solution for InfiniBand based HPC Clouds with SR-IOV?
Problem Statement

• Can we propose a realistic approach to solve this contradiction with some tradeoffs?

• Most HPC applications are based on MPI runtime nowadays, can we propose a VM migration approach for MPI applications over SR-IOV enabled InfiniBand clusters?

• Can such a migration approach for MPI applications avoid modifying hypervisors and InfiniBand drivers?

• Can the proposed design migrate VMs with SR-IOV IB devices in a high-performance and scalable manner?

• Can the proposed design minimize the overhead for running MPI applications inside the VMs during migration?
Outline

- Introduction
- Problem Statement
- Design
- Performance Evaluation
- Conclusion & Future Work
Proposed High Performance SR-IOV enabled VM Migration Framework

- Two challenges need to handle:
  - Detachment/re-attachment of virtualized IB device
  - Maintain IB connection
- Multiple parallel libraries to coordinate VM during migration (detach/reattach SR-IOV/IVShmem, migrate VMs, migration status)
- MPI runtime handles the IB connection suspending and reacting
- Propose Progress Engine (PE) and Migration Thread based (MT) design to optimize VM migration and MPI application performance

Sequence Diagram of VM Migration

Network Based Computing Laboratory
OSU-SC’17
Proposed Design of MPI Runtime

Network Based Computing Laboratory

OSU-SC’17
Outline

• Introduction
• Problem Statement
• Design
• Performance Evaluation
• Conclusion & Future Work
# Experimental Testbed

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Chameleon Cloud</th>
<th>Nowlab</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>Intel Xeon E5-2670 v3 24-core 2.3 GHz</td>
<td>Intel Xeon E5-2670 dual 8-core 2.6 GHz</td>
</tr>
<tr>
<td>RAM</td>
<td>128 GB</td>
<td>32 GB</td>
</tr>
<tr>
<td>Interconnect</td>
<td>Mellanox ConnectX-3 HCA, (FDR 56Gbps), MLNX OFED LINUX-3.0-1.0.1 as driver</td>
<td>Mellanox ConnectX-3 HCA, (FDR 56Gbps), MLNX OFED LINUX-3.2-2.0.0</td>
</tr>
<tr>
<td>OS</td>
<td>CentOS Linux release 7.1.1503 (Core)</td>
<td></td>
</tr>
<tr>
<td>Compiler</td>
<td>GCC 4.8.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MVAPICH2 and OSU micro-benchmarks v5.3</td>
<td></td>
</tr>
</tbody>
</table>
- Migrate a VM from one machine to another while benchmark is running inside
- Proposed MT-based designs perform slightly worse than PE-based designs because of lock/unlock
- No benefit from MT because of NO computation involved
Overlapping Evaluation

- fix the communication time and increase the computation time
- 10% computation, partial migration time could be overlapped with computation in MT-typical
- As computation percentage increases, more chance for overlapping
**Application Performance**

- 8 VMs in total and 1 VM carries out migration during application running
- Compared with NM, MT-worst and PE incur some overhead
- MT-typical allows migration to be completely overlapped with computation
### Available Appliances on Chameleon Cloud*

<table>
<thead>
<tr>
<th>Appliance</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CentOS 7 KVM SR-IOV</td>
<td>Chameleon bare-metal image customized with the KVM hypervisor and a recompiled kernel to enable SR-IOV over InfiniBand. <a href="https://www.chameleoncloud.org/appliances/3/">https://www.chameleoncloud.org/appliances/3/</a></td>
</tr>
<tr>
<td><strong>MPI bare-metal cluster complex appliance (Based on Heat)</strong></td>
<td>This appliance deploys an MPI cluster composed of bare metal instances using the MVAPICH2 library over InfiniBand. <a href="https://www.chameleoncloud.org/appliances/29/">https://www.chameleoncloud.org/appliances/29/</a></td>
</tr>
<tr>
<td><strong>MPI + SR-IOV KVM cluster (Based on Heat)</strong></td>
<td>This appliance deploys an MPI cluster of KVM virtual machines using the MVAPICH2-Virt implementation and configured with SR-IOV for high-performance communication over InfiniBand. <a href="https://www.chameleoncloud.org/appliances/28/">https://www.chameleoncloud.org/appliances/28/</a></td>
</tr>
<tr>
<td>CentOS 7 SR-IOV RDMA-Hadoop</td>
<td>The CentOS 7 SR-IOV RDMA-Hadoop appliance is built from the CentOS 7 appliance and additionally contains RDMA-Hadoop library with SR-IOV. <a href="https://www.chameleoncloud.org/appliances/17/">https://www.chameleoncloud.org/appliances/17/</a></td>
</tr>
</tbody>
</table>

- Through these available appliances, users and researchers can easily deploy HPC clouds to perform experiments and run jobs with
  - High-Performance SR-IOV + InfiniBand
  - High-Performance MVAPICH2 Library over bare-metal InfiniBand clusters
  - High-Performance MVAPICH2 Library with Virtualization Support over SR-IOV enabled KVM clusters
  - High-Performance Hadoop with RDMA-based Enhancements Support

[*] Only include appliances contributed by OSU NowLab
1. Load VM Config
2. Allocate Ports
3. Allocate Floating IPs
4. Generate SSH Keypair
5. Launch VM
6. Attach SR-IOV Device
7. Hot plug IVShmem Device
8. Download/Install MVAPICH2-Virt
9. Populate VMs/IPs
10. Associate Floating IPs

**MPI Complex Appliances based on MVAPICH2 on Chameleon**
Conclusion & Future Work

• Propose a high-performance VM migration framework for MPI applications on SR-IOV enabled InfiniBand clusters
• Hypervisor- and InfiniBand driver-independent
• Design Progress Engine (PE) based and Migration Thread (MT) based MPI runtime design
• Design a high-performance and scalable controller which works seamlessly with our proposed designs in MPI runtime
• Evaluate the proposed framework with MPI level micro-benchmarks and real-world HPC applications
• Could completely hide the overhead of VM migration through computation and migration overlapping
• Future Work – Evaluate our proposed framework at larger scales and different hypervisors, such as Xen; Solution will be available in upcoming release
Thank You!

{zhanjie, luxi, panda}@cse.ohio-state.edu