Virtualization of Lustre for QA and Benchmarking

DataDirect Networks Japan, Inc.
Senior Solutions Architect
Shuichi Ihara
Infrastructure Challenges

- Massive number of dedicated servers required for testing.
- Challenges with setting up Networking for all the servers.
- Proliferation of hardware and operating system configurations to be tested.
- Infrastructure requires massive amounts of Rack Space and Power.
- CPU cores are increasing – Single threaded applications need to take advantage of it.

Virtualization technology can help us here
KVM (Kernel-based Virtual Machine)

- Strong candidates for Lustre QA infrastructure
- The code is already merged in the Linux mainstream
- Supported by many Linux distributions (SLES11, RHEL5.x, 6, Ubuntu)

- Supports full virtualization
  - No special patched kernel needed on VMs
  - Ability to create an quasi-real server environment
  - Supports CPU affinity and PCI passthrough
  - SR-IOV (future)

- CLI is available and it's scriptable!
  - Easy automation and administration

Benefits
- Less need for physical work in the lab 😊
- Fast implementation of Lustre QA infrastructure
  - For Lustre sanity testing on many types of H/W configuration.
  - For function testing (HA, LNET routing, etc.).
  - For benchmark use.
- To build Lustre RPMs
  - Build systems for different Linux distributions
Lustre on VM example(1)
- HA testing -

- Lustre HA testing
  - VMs: 2 x MDS, 2 x OSS and 1 x Client
  - Attach QDR Infiniband HCA to each VM
• **Testing for 2-hop routing (IB <-> 10GbE <-> 10GbE <-> IB)**
  - 5 VMs: 1 x MDS, 1 x OSS, 2 x Router and 1 x Client
  - Attach a QDR Infiniband HCA to each VM; Router VM also have 10GbE connections
Lustre on VM example(3) - Benchmark use -

- For Lustre Benchmark
  - 2 VMs: 2 x OSS per physical server
  - Attach two QDR Infiniband HCAs to each VM. (One for connecting to Storage, another one for LNET)
Lustre performance on KVM
Benchmark configuration

- **SuperMicro's SuperServer**
  - 2 x Intel Xeon (X5670, 2.93GHz), 48GB Memory
  - 2 x Tylersburg (IOH-36D), 6 x PCIe gen2 slots
- **PCI devices**
  - 4 x Mellanox QDR HCA
  - Dual ports and works as 10GbE/QDR hybrid network card
- **Software**
  - RHEL6 (Host), CentOS5.5(VM)
  - Lustre-1.8.4
Type of benchmark

- **Network performance**
  - RDMA (bandwidth, latency)
  - LNET Selftest
- **Lustre backend performance**
  - obdfilter-survey
- **Lustre performance from the clients**
  - IOR
Network performance
- Physical resource assignment for VMs -

- DDR3
  - X5670
    - IOH2-36D
    - IOH1-36D
    - HCA
    - PCIe gen2 x8
    - QPI

- vm01's resources
- vm02's resources
- vm03's resources
- vm04's resources
Network performance
- RDMA Benchmark results -

- Tested with rdma_bw, rdma_lat
- No performance differences when compared to non-VM (Host).

### RDMA Bandwidth

<table>
<thead>
<tr>
<th></th>
<th>Host</th>
<th>vm01</th>
<th>vm02</th>
<th>vm03</th>
<th>vm04</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth (MB/sec)</td>
<td>3197.23</td>
<td>3239.21</td>
<td>3238.81</td>
<td>3194.67</td>
<td>3194.11</td>
</tr>
</tbody>
</table>

### RDMA Latency

<table>
<thead>
<tr>
<th></th>
<th>Host</th>
<th>vm01</th>
<th>vm02</th>
<th>vm03</th>
<th>vm04</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latency (usec)</td>
<td>1.42</td>
<td>1.36792</td>
<td>1.36861</td>
<td>1.3577</td>
<td>1.35974</td>
</tr>
</tbody>
</table>
Network performance
- RDMA Benchmark (Bandwidth) in detail -

RDMA bandwidth (vm01)

RDMA bandwidth (vm02)

RDMA bandwidth (Host)

RDMA bandwidth (vm03)

RDMA bandwidth (vm04)
Network performance
- RDMA Benchmark (Latency) in detail -

- RDMA Latency (vm01)

- RDMA Latency (vm02)

- RDMA Latency (Host)

- RDMA Latency (vm03)

- RDMA Latency (vm04)
Network performance
- LNET selftest results -

- Tested on 4 servers and 4 clients
- Compared the Lustre clients on VM and non-VM
Lustre backend performance

- **OSS (two VMs)**
  - 6 cores per VM
  - 12GB memory per VM
  - NUMA & NUMIOA aware
  - Two HCA (for SRP and LNET) are assigned with PCI pass-through

- **Storage**
  - SFA10000 (Single Controller)
  - 140 x SATA disk
  - 2 x QDR connections
Lustre backend performance
- I/O path without CPU affinity -

VM02

DDR3

Lustre Threads

X5670

QPI

DDR3

QPI

PCIe gen2 x8

PCIe gen2 x8

LNET

SRP for Storage

VM01

DDR3

Lustre Threads

X5670

QPI

Not expected I/O Path

Shortest I/O Path

oss02's resources

oss01's resources

SRP for Storage

PCIe gen2 x8

PCIe gen2 x8
Lustre backend performance
- Obdfilter-survey results -

Write

vm01: 2.6GB/sec  vm02: 2.6GB/sec  vm01+vm02: 5GB/sec

Read

vm01: 2.7GB/sec  vm02: 2.7GB/sec  vm01+vm02: 5.4GB/sec
Lustre performance from the clients - IOR -

• Run IOR from 7 x Lustre clients

Write: 5GB/sec  Read: 5.4GB/sec

• This is almost the same results which we are seeing on the physical Lustre servers with SFA10000 (single controller).

• Could see double performance by using 4 VMs and dual SFA10000 controllers.
Summary

• **A Virtualized Infrastructure based on KVM works well**
  • Only a couple of minutes needed for all server setup!
  • Various types of Lustre testing are possible.
  • Achieves almost equal performance numbers when compared to physical servers without VMs.
  • SR-IOV will provide a basis for much more flexible configurations!
  • Will investigate SR-IOV, FC and testing on more servers in future!
  • Will continue to invest Lustre on KVM
Introducing the DDN SFA10000E
Multi-Platform Architecture

Block Storage Array
- SFA10000
  - Block Storage Target

Clustered Filer
- SFA10000E
  - DDN File Storage
  - EXAScaler
  - GridScaler

Open Appliance
- SFA10000E
  - Customer Applications
- SFA10000E
  - Embedded Storage Server
- SFA10000
  - Block Storage Target

Product Evolution

Flexible Deployment Options: 3 System Modalities
SFA10000E Appliance

- SFA10000E initially available with DataDirect Networks’ s parallel clustered file system solutions
- Integrate multiple appliances to scale to over 200GB/s and 10’ s of Petabytes

- Reduce complexity, infrastructure and administration
- Reduce cost as well as lower operational cost
- Increase performance for latency sensitive applications
  - Shared Memory
  - Eliminate SCSI Overhead
SFA10000 Embedded ExaScaler

- **8 x 10GigE or QDR IB ports** (4 x Dual-port Cards)

- **Dedicated I/O Bridge**

- **Application Memory**

- **Multi-core CPU Application Processor (AP)**

- **Dedicated PCI-e I/O**

- **Virtual Disk Block Driver**

- **High Speed Bus**

- **Block Services (Virtual Disks)**

- **Multi-Core CPU RAID Processor (RP)**

- **Dedicated I/O Bridge**

- **Cache Memory**

- **Back-End SAS HBAs**
HPC Storage on the SFA10000E Appliance

Storage Fusion Architecture not only reduces complexity, it streamlines IO by reducing latency and protocol conversions.
Sample Customers

Efficient Storage Lustre Users Worldwide – Nearing our 10-Yr Lustre Deployment Milestone

TB/s of Lustre Performance Powering HPC Worldwide

Thanks To DDN Customers For Your Partnership in HPC!
Thank You
Backup Slides
How to setup Lustre on KVM with IB - Quick example -

It's basic KVM setup, no any special operations!

1. Create first VM and Install Operating system (and Lustre)
   
   ```
   # virt-install --name=oss01 --vcpus=6 --ram=8192 --os-type=linux --hvm --connect=qemu:///system --network bridge:br0 --location /var/www/html/os_images/centos5.5 --file /vmimage/oss01.img -s 10 --accelerate --nographics --mac=52:54:00:aa:aa:00 --extra-args='console=tty0 console=ttyS0,115200n8 ks=http://192.168.122.1/centos5.ks'
   ```

2. Create clone VM image for Second VM
   
   ```
   # virt-clone --original oss01 --name oss02 --mac=52:54:00:aa:aa:02 --file /vmimage/oss02.img
   ```

3. Find PCI device ID of HCA
   
   ```
   # lspci | grep InfiniBand
   02:00.0 InfiniBand: Mellanox Technologies MT26428 [ConnectX VPI PCIe 2.0 5GT/s - IB QDR / 10GigE](rev b0)
   ```

4. Detach HCAs from Host
   
   ```
   # virsh nodedev-list | grep pci_0000_02
   pci_0000_02_00_0
   ```
5. Create an definition file of HCA
   # cat mellanox_hca_bus02_00_0.xml
   <hostdev mode='subsystem' type='pci'>
   <source><address bus='0x02' slot='0x00' function='0x00'/></source>
   </hostdev>

6. Attach HCA to VM
   # virsh attach-device vm01 mellanox_hca_bus02.xml

7. CPU assailment and affinity setting
   # virsh vcpupin vm01 0 0
   # virsh vcpupin vm01 1 1
   ...

8. Now, VMs with Infiniband is ready. Move forward formatting the Lustre.

   This all procedures are scriptable and don't many typing 😊