SCSI support on Xen

Matsumoto Hitoshi
matsumotohitosh@jp.fujitsu.com
Fujitsu Ltd.
Agenda

- Architecture
- Current status
- Next challenge
Architecture
Conventional application support
Server consolidation
HW fault tolerance
Performance
Conventional application support

- Some applications issue SCSI commands.
  
  ex. DB (enterprise), backup.

  → pvSCSI driver

- Server consolidation
- HW fault tolerance
- Performance
- Data Center management
  Enterprise data are stored in FC/SCSI devices.
  Reliability and availability are required.
- Many SCSI devices in data center.
  - hardware snapshot
  - tape operation

**Data center**

- DB server
- Oracle
- backup server
- SAN
- storage (RAID)
- tape drive
- SCSI command

LAN

load, unload, reset

hardware snapshot

data file

snapshot
example: SCSI command for storage

Minimum backup window

D2T  Disk to Tape

D2D  Disk to disk

D2D  Hardware snapshot

Backup window

on line

on line

minimize
example: SCSI command for tape

- Robot: move cartridge
- Tape: load, unload, rewind

Disk (RAID)

backup
restore

load
unload

out of box

Tape drive

Tape cartridge
pvSCSI driver (SCSI passthrough) consists of SCSI frontend driver and SCSI backend driver.

- Each guest can issue SCSI commands via host.
- Each guest can occupy each FC HBA card.
  - conventional application works on guest.

---

**Diagram Explanation**

- **Host OS** and **guest OS** communicate with **SCSI backend driver**.
- **SCSI command** flows between **SCSI frontend driver** and **guest OS**.
- **FC HBA card** connects to **SAN**.

---

**Key Points**

- pvSCSI driver facilitates direct access to storage for virtual machines.
- Each guest can manage its own storage resources independently.
- Conventional applications can be used on the guest operating systems.
Requirements for VM in data center

- Conventional application support
- **Server consolidation**
  - All resources are consolidated on VM.
    -> pvSCSI driver + NPIV
- HW fault tolerance
- Performance
Many servers in data center
Each server has several storage.
Server consolidation on VM

Many HBA cards are need for data center.
NPIV support

- NPIV: Technology to creates a many vHBA (VP) in a physical HBA.
  - Each guest can have own vHBA.
  - The number of physical HBA can be reduced.
NPIV (N-Port Identifier Virtualization)

- The virtual port can connect to SAN independently as the physical port. The virtual port is allocated to owner guest.
- NPIV is standardized by the SNIA.
Requirements for VM in data center

- Conventional application support
- Server consolidation
- **HW fault tolerance**
  - Containment hardware failure
    - pvSCSI driver + NPIV + driver domain
  - Redundancy hardware failure
    - pvSCSI driver + driver domain + multi path driver
- Performance
Crash on VP to guest 1 does not affect guest 2.
- If host goes down with I/O operation, whole system does not go down.
- Guest cannot access I/O device directly.
Crash on driver domain 1 for guest 1 does not affect guest 2.
multi path driver

- fail over: alternate path retry
- load balance: multi access path

- Linux has a multi path driver as “device mapper”.
- Many vendors prepare their original multi path driver.
Each guest has alternate path to I/O device via driver domain so that each guest can continue to work when a HBA card or a driver domain is crashed.
Requirements for VM in data center

- Conventional application support
- Server consolidation
- HW fault tolerance

**Performance**

- The performance of pvSCSI driver is almost same as VBD.
- Guest issues I/O to device not via host.
  - More performance!!
  - → direct I/O
- PV domain and HVM domain
Each guest can access hardware without host.
direct I/O / NPIV architecture

Managed by Hypervisor/Dom0 but guest domain

Address Translation Table

CPU

chipset for direct I/O

Address Translation

DMA controller

VP

CPU

DMA controller

VP

HBA

for guest A

Access Allowed

Access Denied

for guest B

Memory

Authorized Access

Unauthorized Access

Guarantee that the PCI Express device cannot perform unauthorized access to memory portion

→ The device can be assigned to guest domain
current status
current status

- The basic function of pvSCSI driver code was posted to xen community.
- NPIV works.
- pvSCSI driver on driver domain is under evaluation.
- pvSCSI driver works on HVM domain and PV domain.
- Oracle RMAN works on guest with pvSCSI driver.
**performance (same as VBD)**

- **Dom0 vs VBD vs pvSCSI**
  
  The performance ratio of VBD and pvSCSI to Dom0(100%).
  
  → Performance of pvSCSI is almost same as VBD.

Memory: 2GB

CPU: 2 for each domain

Tool: iogen1.3.6

---

**Dom0 vs VBD vs pvSCSI**

- **read**
  
  Block size: 8k, 128k, 256k

- **write**
  
  Block size: 8k, 128k, 256k
Oracle RMAN works on guest with pvSCSI driver.
Next challenge
Next challenge

- Complexity direct I/O and nondirect I/O
- LUN assignment
- PV domain uses pvSCSI driver.
- HVM domain uses direct I/O.
LUN assignment

- LUN allocation to guest with pvSCSI driver
Special thanks to

Intel Corporation
QLogic Corporation
Emulex Corporation
Brocade Communications Systems Inc
Sun Microsystems Inc
Xenon community engineers!
This work was partly funded by Ministry of Economy, Trade and Industry of Japan as the secure platform project of Association of Super-Advanced Electronics Technologies (ASET).
Appendix
- Direct I/O
- But, All LUNs are assigned to “one” guest domain
→ A HBA must be occupied by an owner guest.
LUN Assignment
(Driver Domain (DD) with pvSCSI)

- Portion of LUNs can be assigned to appropriate guest domain (LUN filtering by pvSCSI)
- But, not direct I/O(via DD)
- Portion of LUNs can be assigned to appropriate guest domain
- But, not yet direct I/O(via DD)
- Portion of LUNs can be assigned to appropriate guest domain
- And direct I/O
Management interface depends on SCSI protocol.

- Virtualization layers
  - server
  - network
  - storage

- SAN
- network virtual storage
- storage box

- Management interface depends on SCSI protocol.
Virtual storage has the online migration function with SCSI protocol.

- Pre copy
- Copying
- Migration complete
Network/storage box virtualization has snapshot copy.
Example 1: Guest Migration

Keep connection to the same v-WWN2.
Example 2: Guest Migration

migrate guest 2

Keep connection to guest 2.