HITACHI
Inspire the Next

Better Utilization of Storage Features from KVM Guest via virtio-scsi

Masaki Kimura

<masaki.kimura.kz@hitachi.com>

Information & Telecommunication Systems Company
IT Platform Division Group
Hitachi, Ltd.

Human Dreams.
Make IT Real.

HITACHI
Inspire the Next

Better Utilization of Storage Features from KVM Guest via virtio-scsi

Contents

HITACHI
Inspire the Next

Better Utilization of Storage Features from KVM Guest via virtio-scsi

1. Background

1-1. Background

HITACHI
Inspire the Next

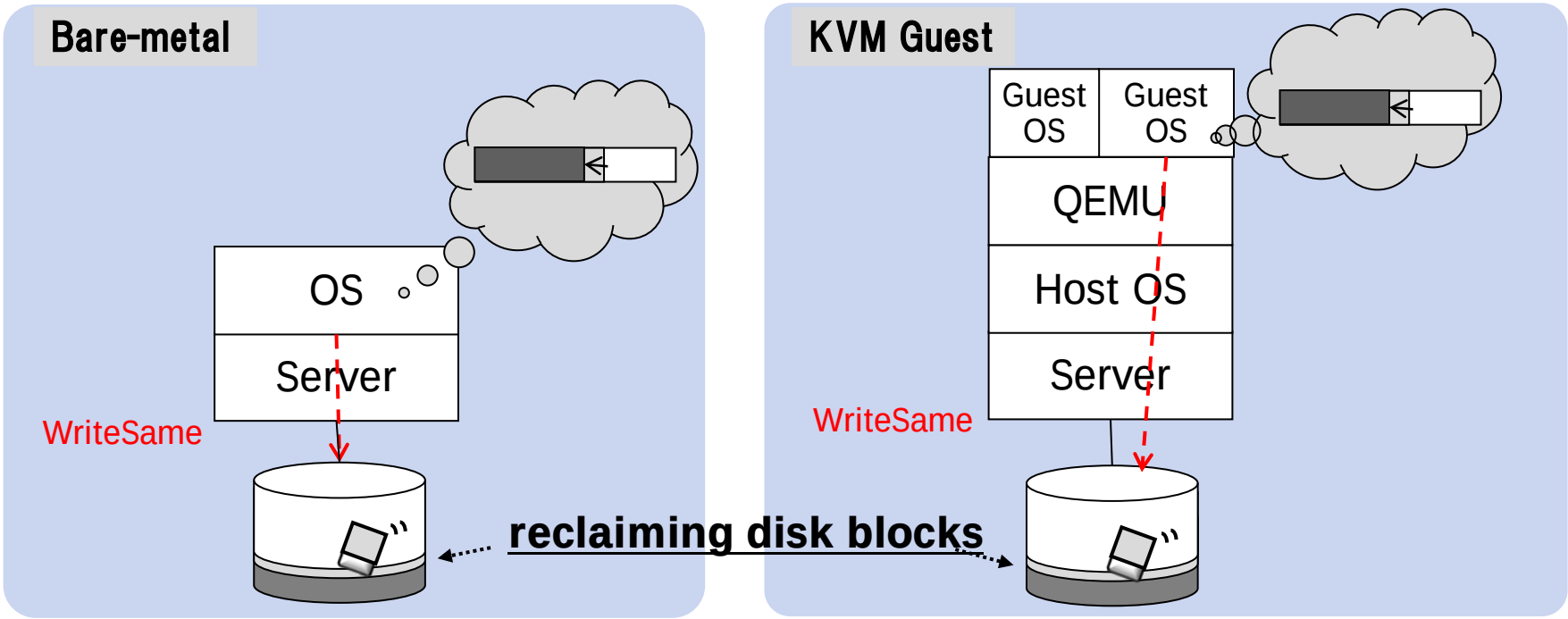
- Enterprise systems expect that virtualized environment has the same level of manageability, availability, and reliability achieved in bare-metal.
- For example:
  1. Thin-provisioned Storage for manageability,
  2. HA cluster for availability,
  3. Backup server for reliability.
- In bare-metal environment, some of these requirements are achieved by using storage features, such as SCSI commands.
- In virtualized environment, the same use cases exists for guests.
  ➔ Issuing SCSI commands from guests are required.

Three use cases, thin-provisioned storage, HA cluster, and backup server will be explained in the next slides.

1-2. Use case #1: Thin-provisioned Storage

HITACHI
Inspire the Next

- Many types of enterprise storage have thin-provision function.
- For achievement of thin-provision, a disk block is allocated on access.
- However, once it is allocated, it can not be reclaimed by storage automatically even when the disk block becomes unused by OS.
- To reclaim unused disk block, OS needs to let storage know unused blocks by issuing WriteSame SCSI command.

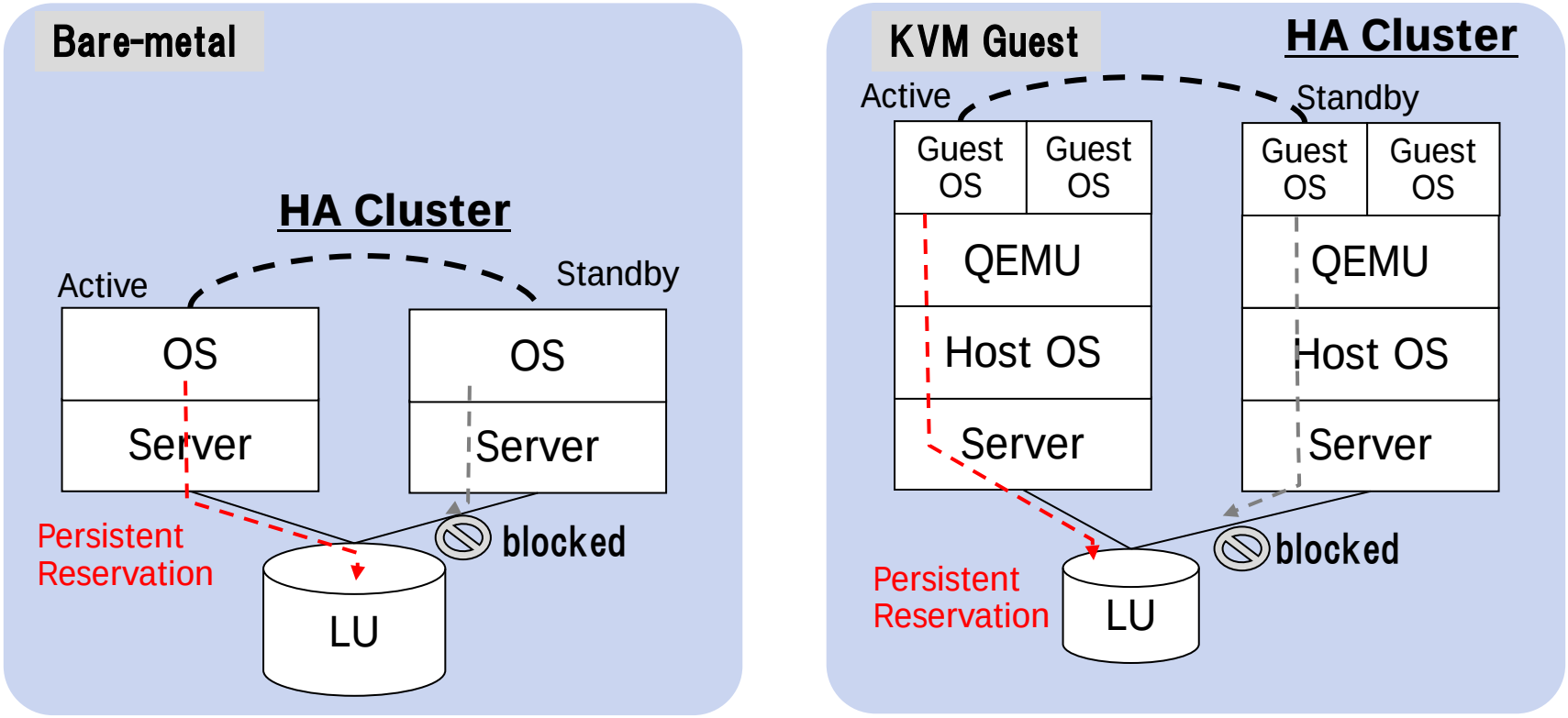


This use case exists for both bare-metal and KVM guests.
➔ **WriteSame is required to be issued to storage from guests.**

1-3. Use Case #2: HA cluster (1/2)

HITACHI
Inspire the Next

- To improve availability, HA cluster is commonly used in bare-metal.
- For HA cluster, Persistent Reservation SCSI command is generally used to guarantee an exclusive access from an active system.

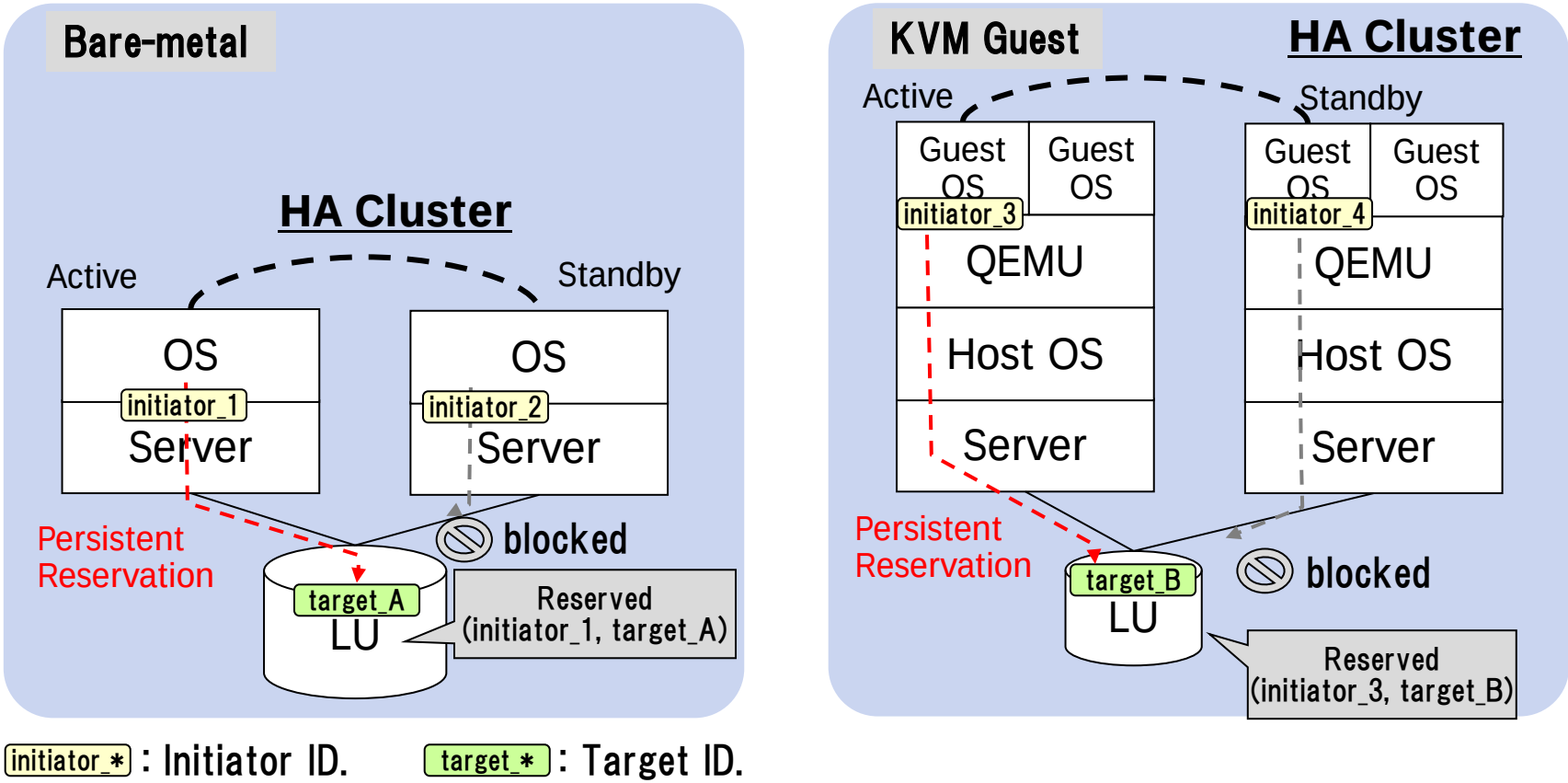


This use case also exists for KVM guests.

➔ **Persistent Reservation is required to be issued to storage from guest.**

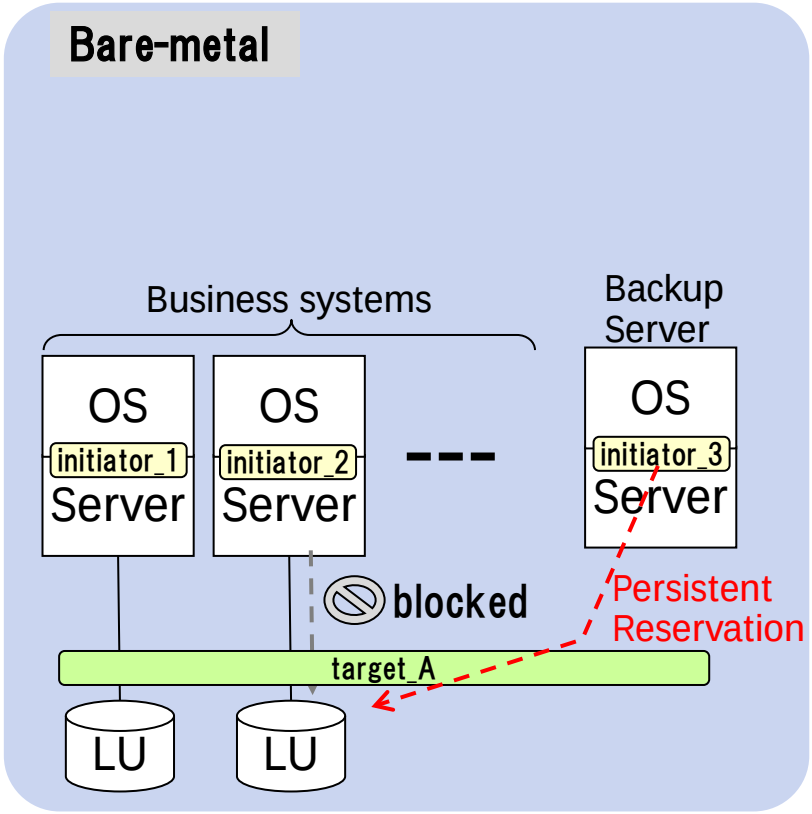
1-4. Use Case #2: HA cluster (2/2)

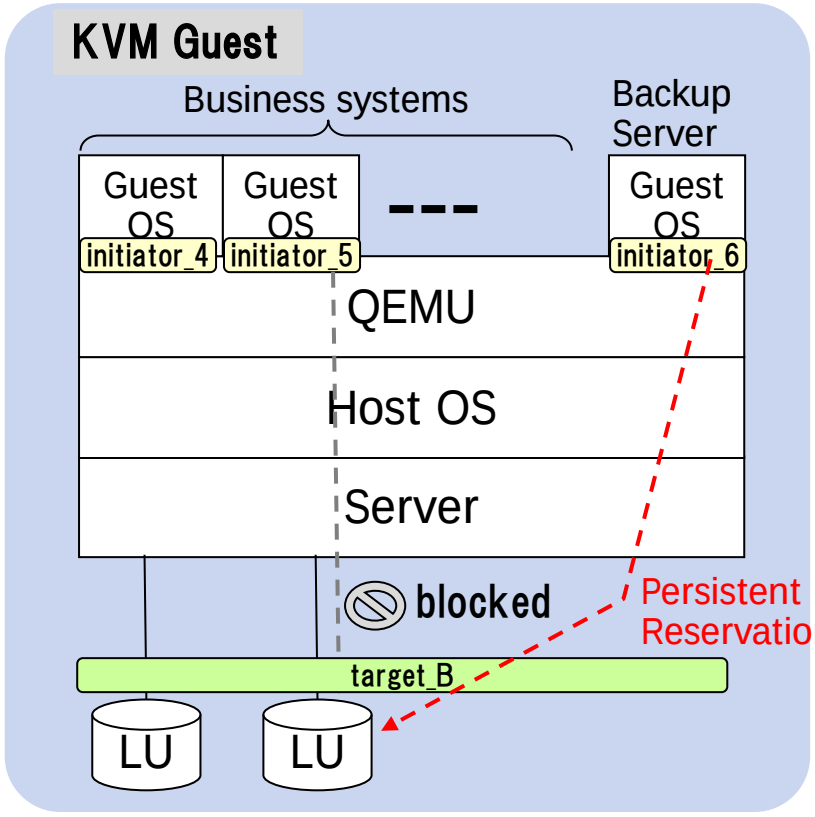
- Persistent Reservation is held by so-called I_T nexus, the combination of initiator ID and target ID.
- I/Os from standby system are blocked, because the I_T nexus is different.
- Therefore, **I_T nexus is required to be unique** for Persistent Reservation to work properly.



1-5. Use Case #3: Backup Server

- Persistent Reservation is also used by some backup-server products to guarantee an exclusive access from a backup server on backup.
- **Persistent Reservation to storage** and **unique I_T nexus** are required by these products.





initiator_* : Initiator ID. target_* : Target ID.

1-6. Summary of Requirements

HITACHI
Inspire the Next

- Requirements from the use cases:
  - Requirement #1: SCSI commands to storage from guests.
    - Thin-provisioned storage requires WriteSame to storage from guests.
    - HA cluster and backup server require Persistent Reservation to storage from guests.
  - Requirement #2: Unique initiator ID across guests.
    - HA cluster and backup server require I_T nexus to be unique.

HITACHI
Inspire the Next

Better Utilization of Storage Features from KVM Guest via virtio-scsi

2. KVM features for guest SCSI commands

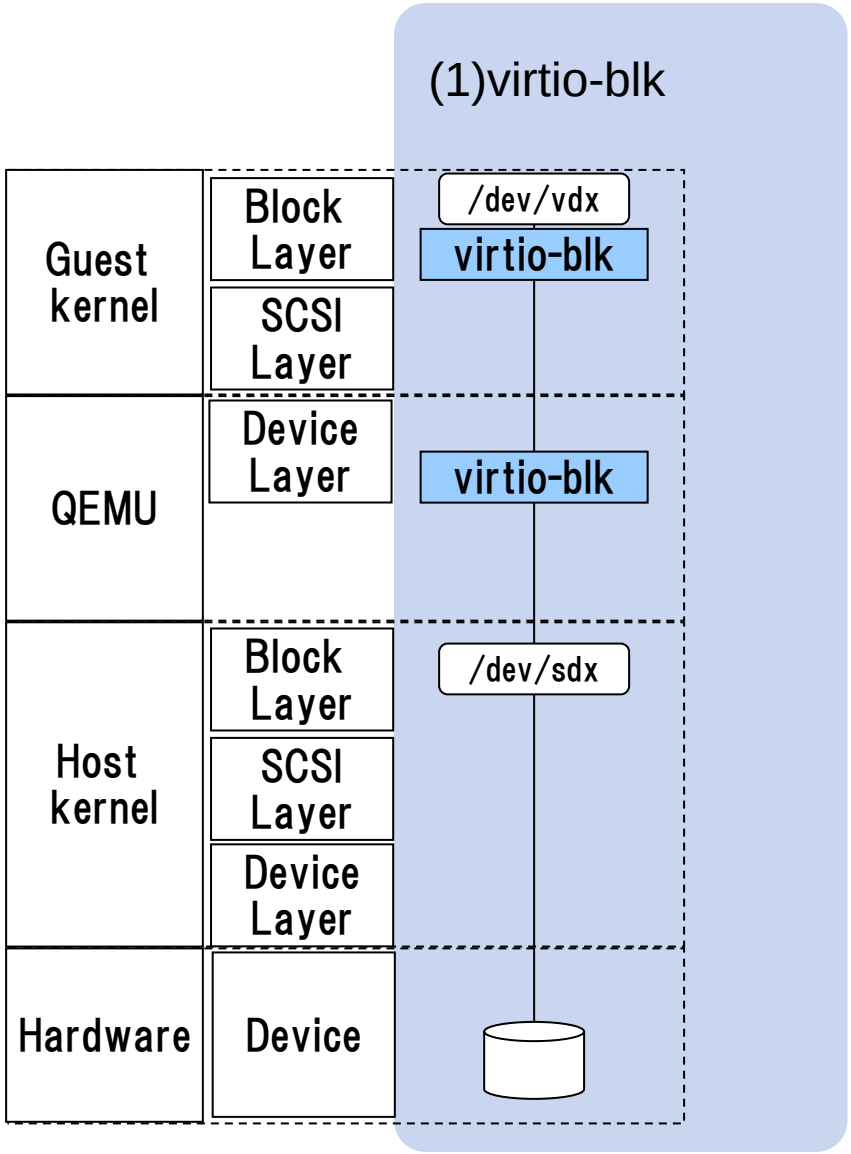
2-1. KVM features for guest disks

HITACHI
Inspire the Next

- This presentation focuses on following three device types and their configurations.

#	Configuration				
	Device Type		Initiator	Target	Backend
1	(1) virtio-blk		-		File
2					Device
3					LUN
4	(2) virtio-scsi		-	(a)qemu	File
5					Device
6					LUN
7			-	(b)lio	block
8					pscsi
9			(c) libiscsi	-	iSCSI storage
10	(3)PCI device assignment	(a)Legacy	-		PCI device
11		(b)VFIO	-		PCI device

2-2. (1) virtio-blk (1/2)



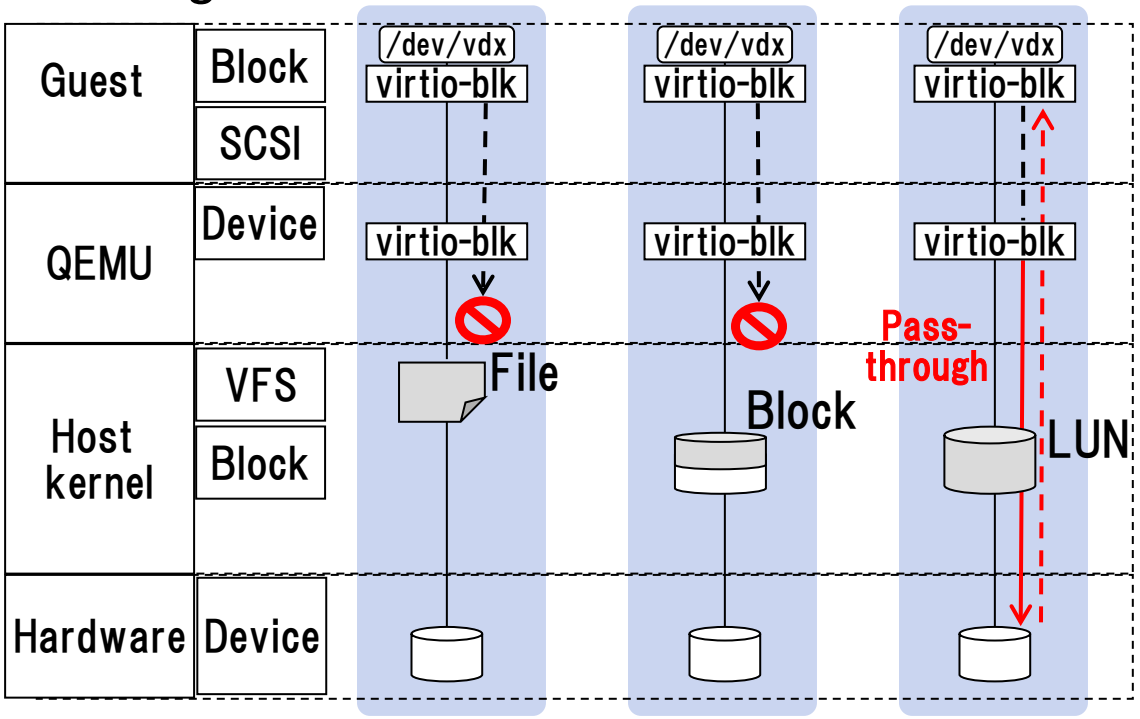
Characteristic

- Para-virtualized disk.
- Shown as /dev/vdX on Linux guests.
- Maximum number of disks is limited by maximum number of PCI devices (32).
- Improving performance with virtio-blk plane and bio-based I/O.

2-3. (1) virtio-blk (2/2)

HITACHI
Inspire the Next

- Configurations and how SCSI commands are handled.



SCSI Command Capability

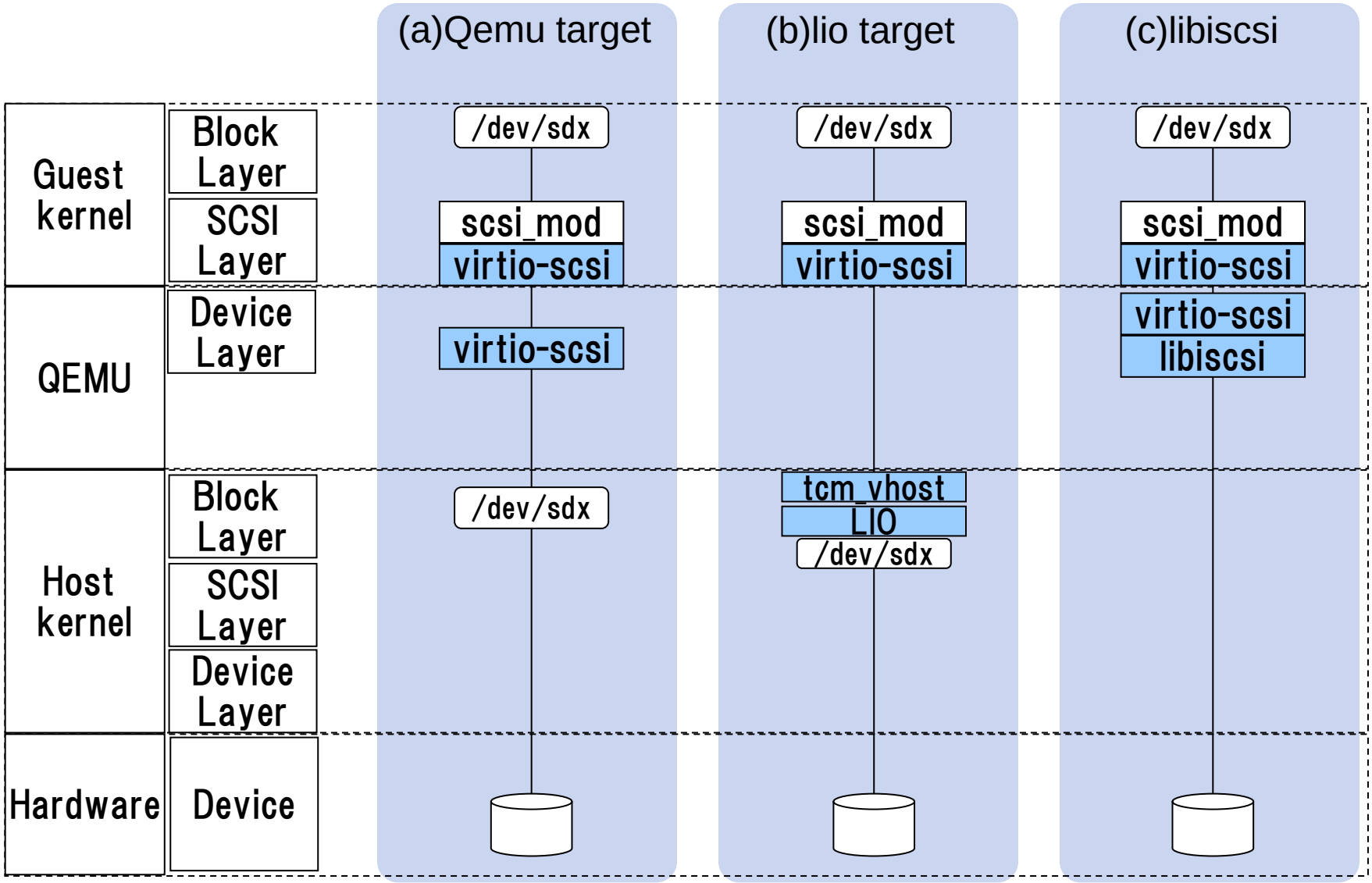
- SCSI command from guest reaches to storage only when attached as LUN.

Backend	KVM Command Line	Libvirt XML	SCSI command
Disk (file)	-device virtio-blk-pci,scsi=**off**	<disk type= '**file**' device= '**disk**' > <target dev='vda' bus='virtio'/>	Not Supported
Disk (block)	-device virtio-blk-pci,scsi=**off**	<disk type= '**block**' device= '**disk**' > <target dev='vda' bus='virtio'/>	Not Supported
LUN	-device virtio-blk-pci,scsi=**on**	<disk type= '**block**' device= '**lun**' > <target dev='vda' bus='virtio'/>	Pass-through

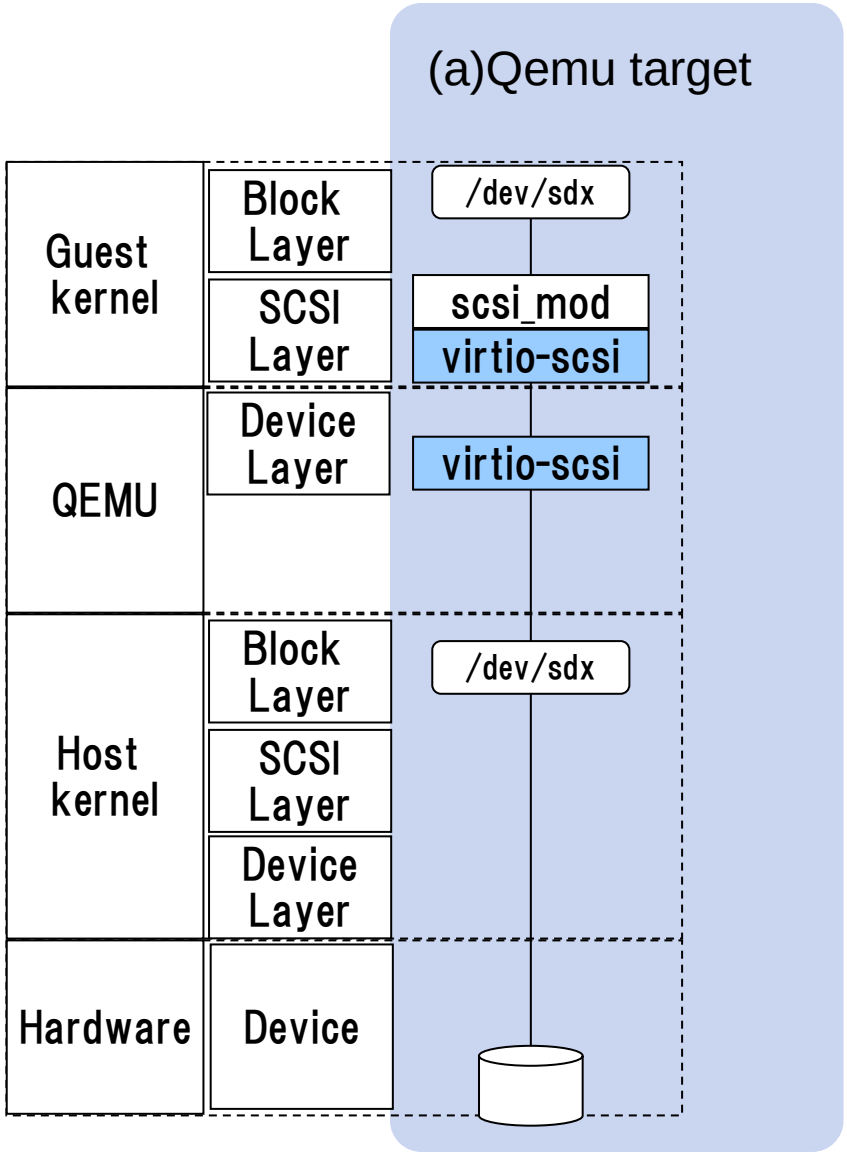
2-4. (2) virtio-scsi

HITACHI
Inspire the Next

- virtio-scsi has three types of configurations.



2-5. (2) virtio-scsi: (a) qemu target (1/2)



Characteristic

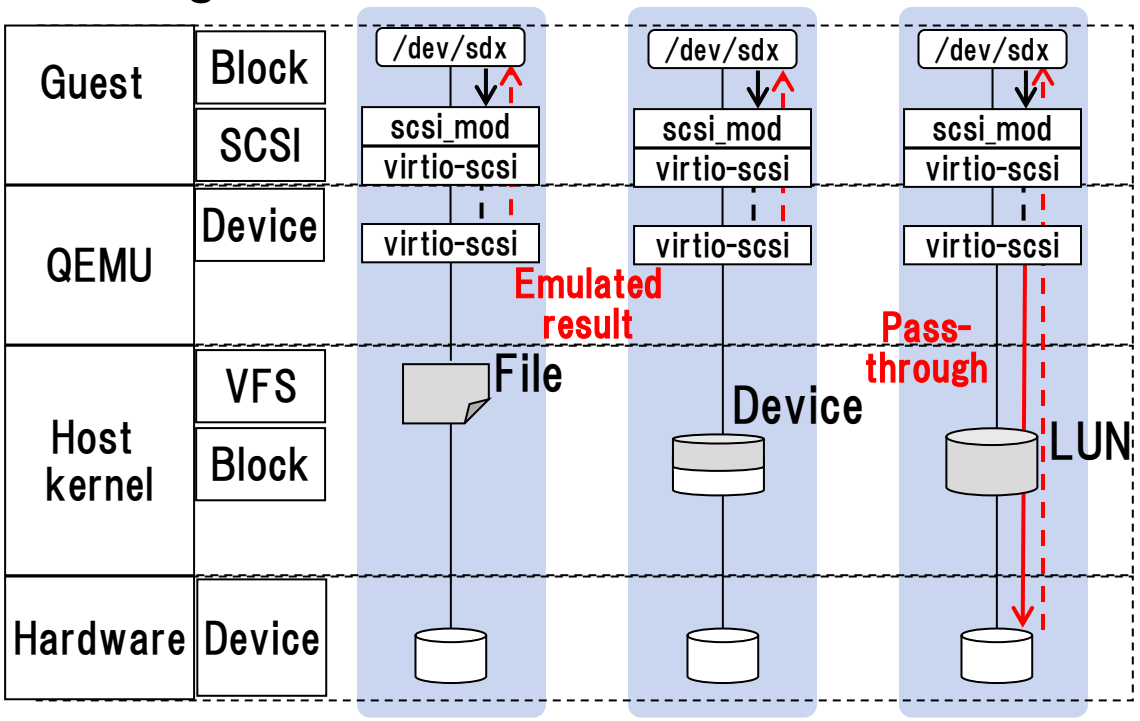
[virtio-scsi]
- Para-virtualized SCSI transport.
- Shown as /dev/sdX on Linux guests.

[Qemu target]
- User space target

2-6. (2) virtio-scsi (a) qemu-target (2/2)

HITACHI
Inspire the Next

- Configurations and how SCSI commands are handled.

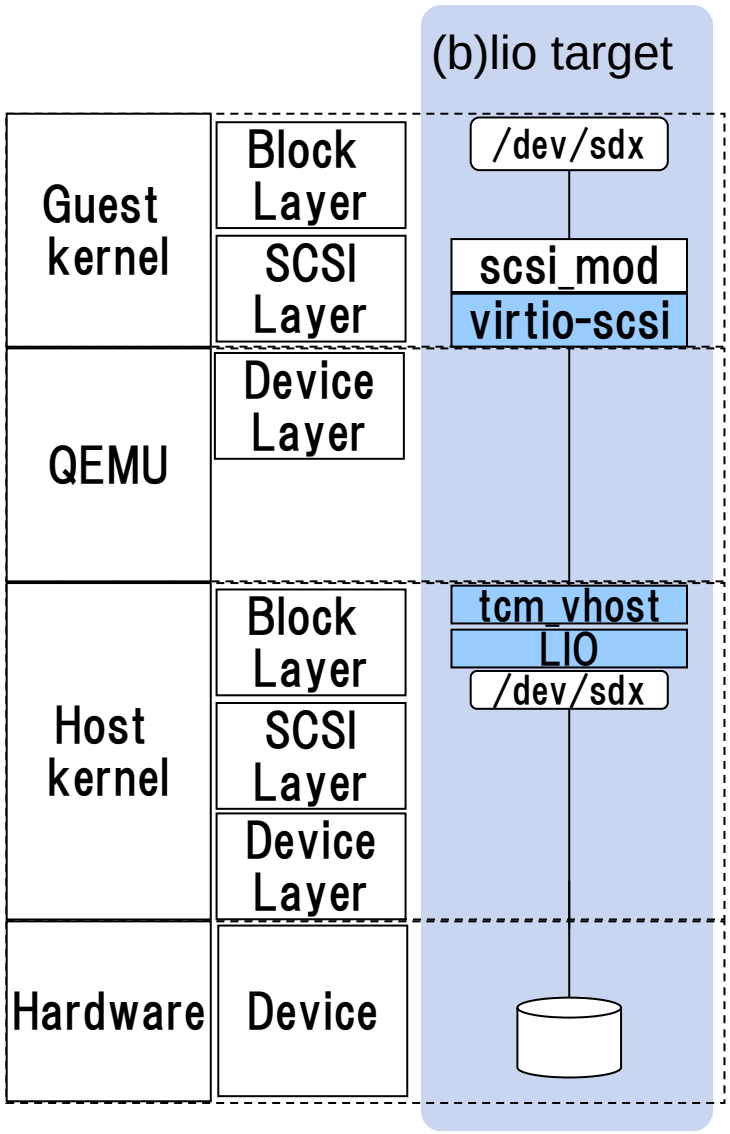


SCSI Command Capability

- SCSI command from guest reaches to storage only when attached as LUN.
- Emulated results return to guest when attached as file or device.

Backend	KVM Command Line	Libvirt XML	SCSI command
Disk (file)	-device scsi-hd	<disk type='file' device='disk' > <target dev='sda' bus='scsi'/>	Emulated
Disk (block)	-device scsi-hd	<disk type='block' device='disk' > <target dev='sda' bus='scsi'/>	Emulated
LUN	-device scsi-block	<disk type='block' device='lun' > <target dev='sda' bus='scsi'/>	Pass-through

2-7. (2) virtio-scsi: (b) lio target



Characteristic

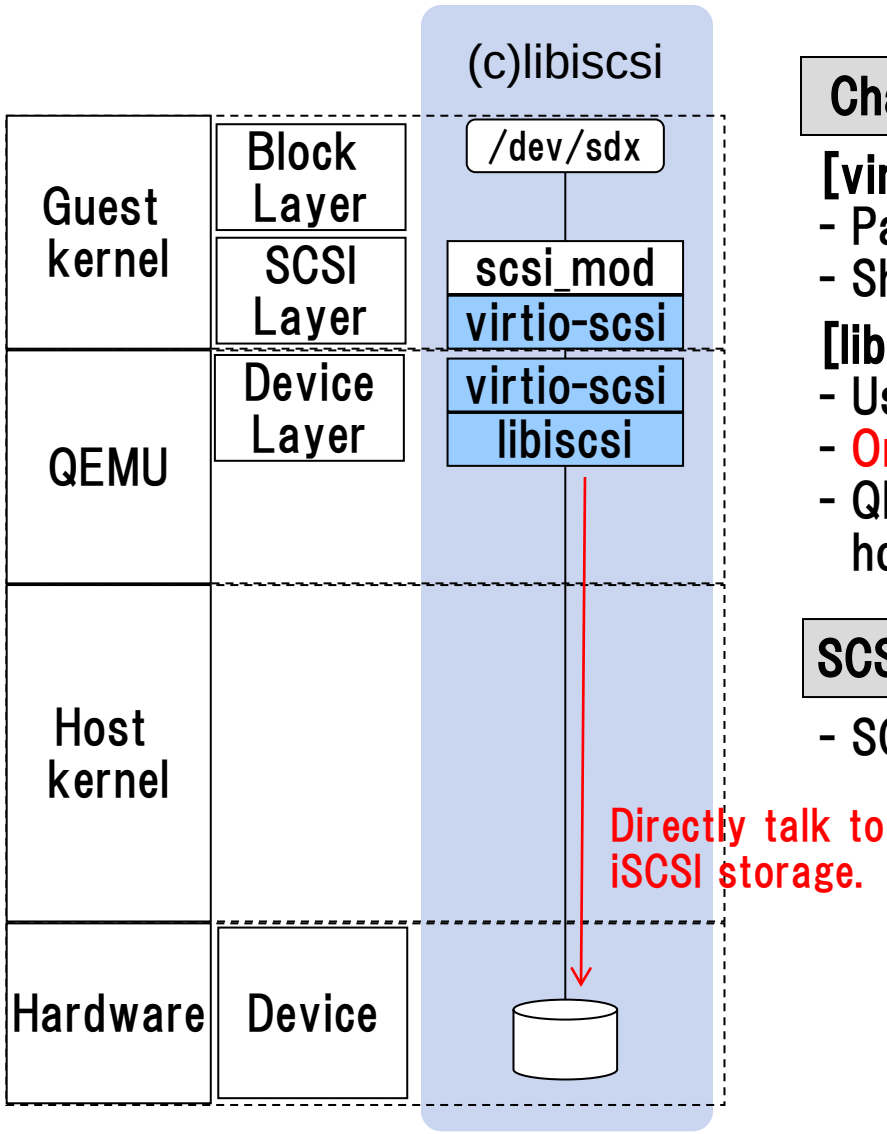
[virtio-scsi]
- Para-virtualized SCSI transport.
- Shown as /dev/sdX on Linux guests.

[lio target]
- Kernel space target.
- Using LIO (linux-iscsi.org) as backend.
- LIO supports following back-stores:
  - block
  - fileio
  - pscsi
  - ramdisk

SCSI Command Capability

- Not yet evaluated.
  (pscsi is pass-through SCSI, therefore it is expected to work well.)

2-8. (2) virtio-scsi: (c) libiscsi



Characteristic

[virtio-scsi]
- Para-virtualized SCSI transport.
- Shown as /dev/sdX on Linux guests.

[libiscsi]
- User space iSCSI initiator.
- Only support iSCSI.
- QEMU directly talk to iSCSI storage, therefore host does not see guest disks.

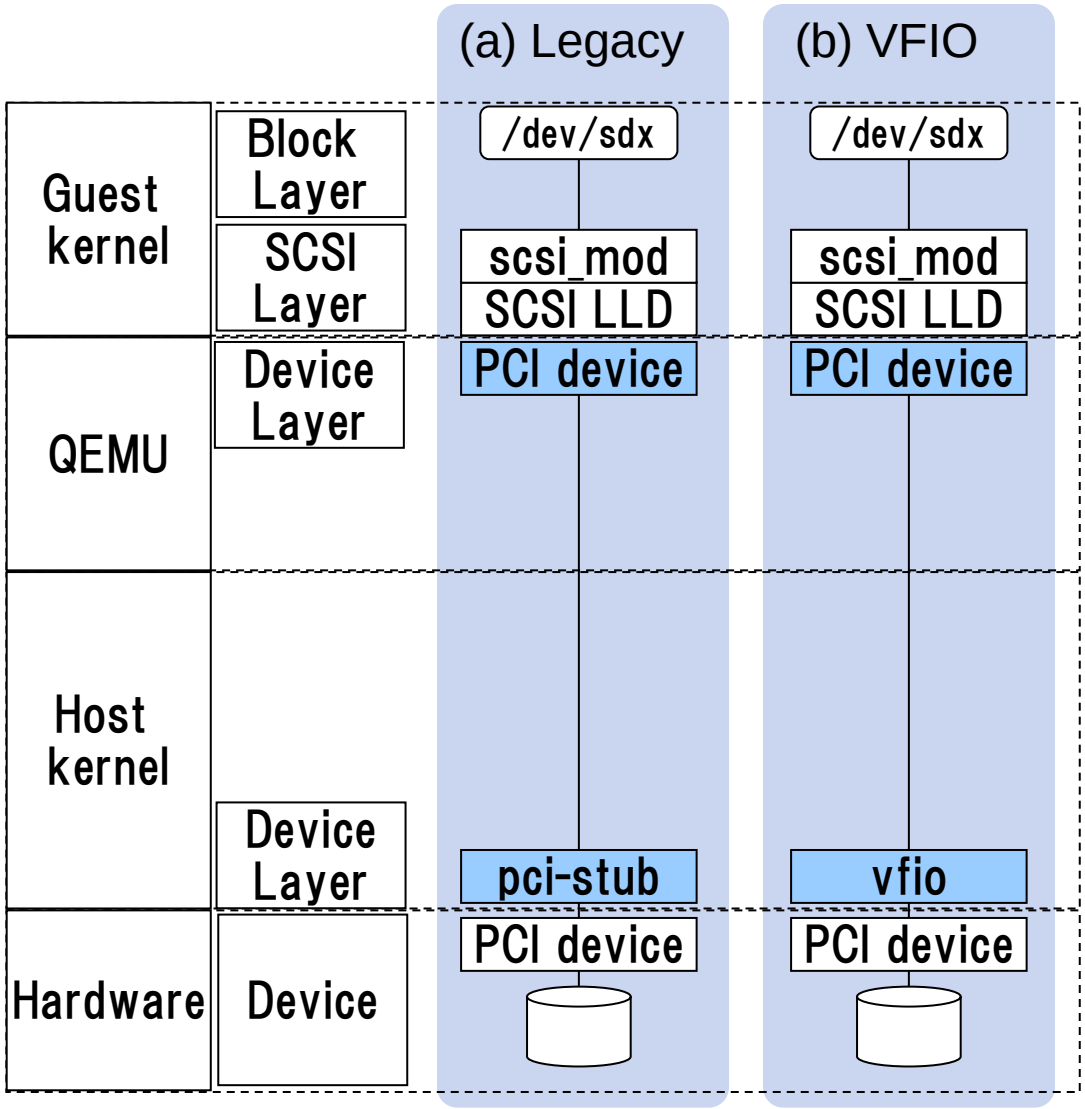
SCSI Command Capability

- SCSI command from guest reaches to storage.

2-9. (3) PCI Device assignment

HITACHI
Inspire the Next

- PCI Device assignment has two types:



Characteristic

- Assign PCI device to guests.
- Host PCI device is dedicated to one guest, therefore the number of guests is limited to the number of PCI devices (or their ports.)

SCSI Command Capability

- SCSI command from guest reaches to storage in both legacy and VFIO configurations.

2-10. Summary of SCSI command capability

HITACHI
Inspire the Next

#	Configuration					Whether guest SCSI commands reach to storage
	Device Type		Initiator	Target	Backend	
1	virtio-blk		-		File	No
2					Device	No
3					LUN	Yes
4	virtio-scsi		-	qemu	File	No
5					Device	No
6					LUN	Yes
7			-	lio	block	No
8					pscsi	???
9			libiscsi	-	iSCSI storage	Yes
10	PCI Device assignment	Legacy	-		PCI device	Yes
11		VFIO	-		PCI device	Yes

In the next chapter, we will see SCSI command capabilities deeper only with configurations marked “Yes” in above table.

HITACHI
Inspire the Next

Better Utilization of Storage Features from KVM Guest via virtio-scsi

3. Current Status of these features

3-1. Evaluation Item and Configuration

HITACHI
Inspire the Next

- Evaluation Items:

#	Evaluation Item	Remarks
(a)	Whether SCSI commands reach to storage.	Requirement #1
(b)	Whether unique initiator ID is assigned.	Requirement #2
(c)	Whether SCSI commands return proper results.	-

- Configurations:

#	Device Type		Initiator	Target	Backend
1	virtio-blk		-		LUN
2	virtio-scsi		-	qemu	LUN
3			libiscsi	-	iSCSI storage
4	PCI Device assignment	Legacy	-		PCI device
5		VFIO	-		PCI device

From next slide, I will share what problems remain in which configurations.

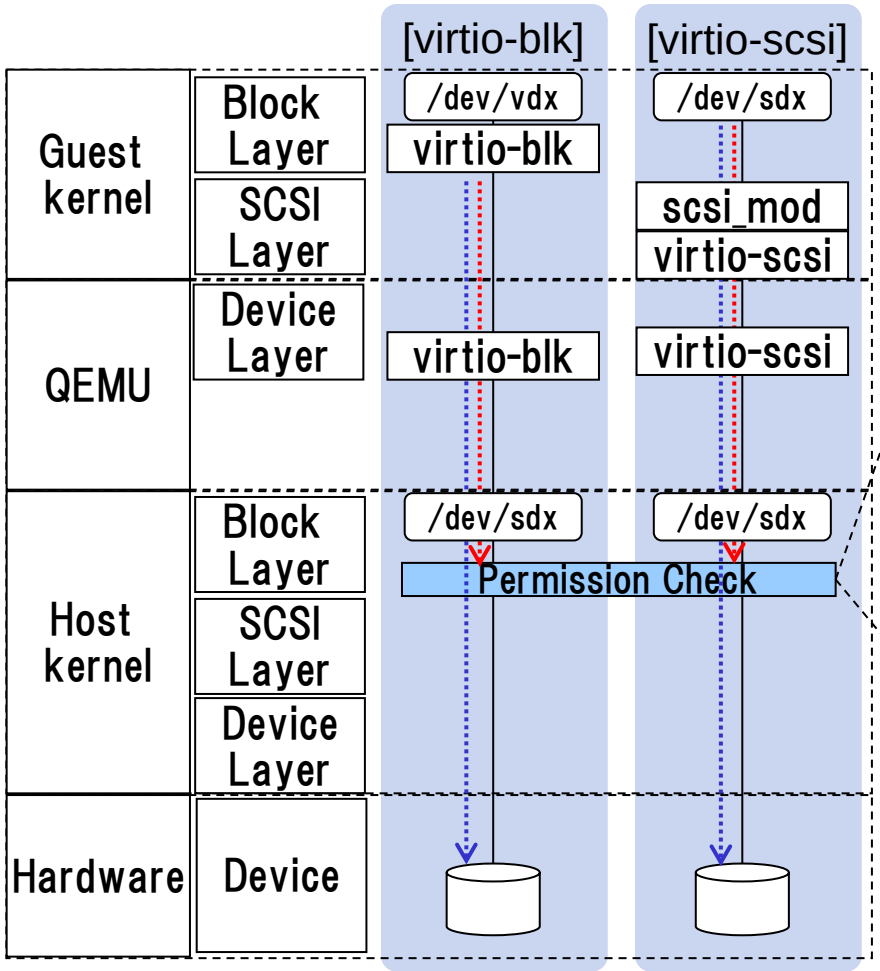
3-2. (a) Whether SCSI commands reach to storage(1/4)

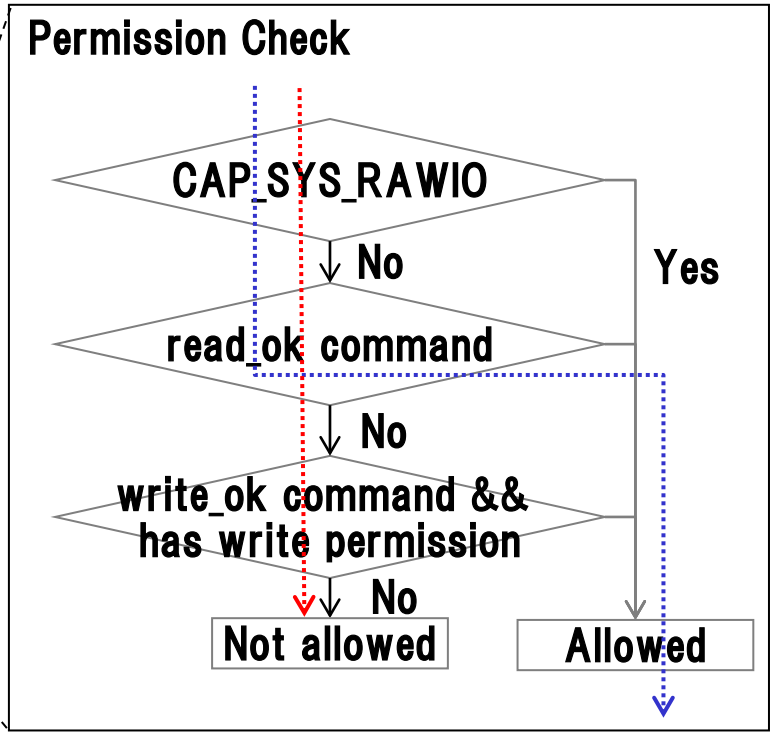
HITACHI
Inspire the Next

Problem

- There is a permission check in host kernel, when guest SCSI commands is issued via virtio-blk or virtio-scsi with qemu target.
- Libvirt-managed KVM guests run as qemu user, who lacks CAP_SYS_RAWIO.

➔ Some SCSI commands, such as **Persistent Reserve and Write Same, are blocked** by this check **unless KVM is running as root user**.





: Inquiry, Report LUN, etc

: Persistent Reserve, WriteSame, etc

3-3. (a) Whether SCSI commands reach to storage(2/4)

HITACHI
Inspire the Next

To solve this issue, following patches have been submitted.

Subject : [PATCH v3 0/2] add per-device sysfs knob to enable unrestricted, unprivileged SG_IO
Date : November 13, 2012
Committer: Paolo Bonzini
URL : https://lkml.org/lkml/2012/11/13/440

Subject : [PATCH 00/13] Corrections and customization of the SG_IO command whitelist (CVE-2012-4542)
Date : January 24, 2013
Committer: Paolo Bonzini
URL : https://lkml.org/lkml/2013/1/24/279

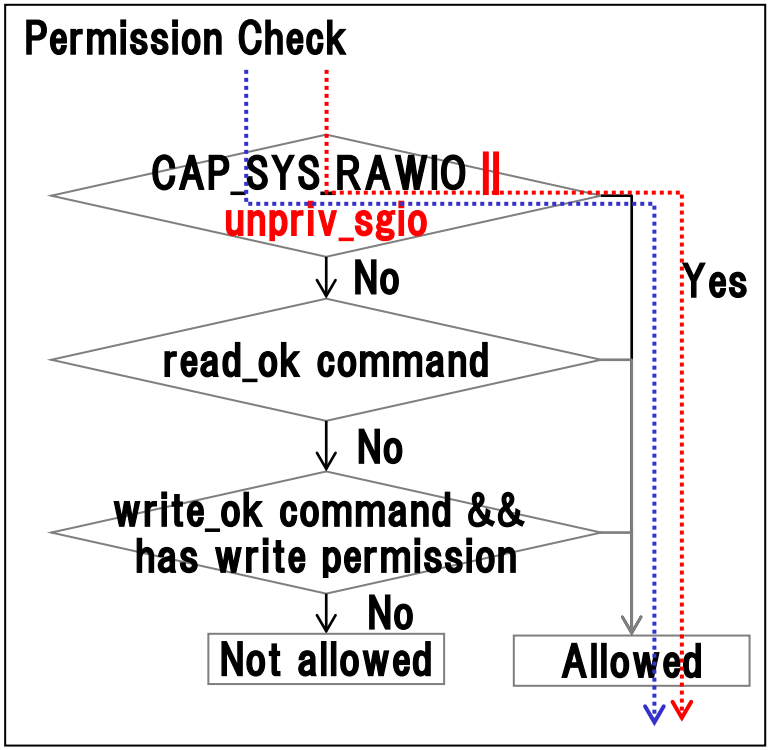
Subject : [PATCH v3 part2] Add per-device sysfs knob to enable unrestricted, unprivileged SG_IO
Date : May 23, 2013
Committer: Paolo Bonzini
URL : https://lkml.org/lkml/2013/5/23/294

However, neither of them has been merged yet.

3-4. (a) Whether SCSI commands reach to storage(3/4)

HITACHI
Inspire the Next

- Concept of these patches is to introduce a flag, unpriv_sgio, to allow non-root users to issue SCSI commands.



: Inquiry, Report LUN, etc

: Persistent Reserve, WriteSame, etc

*** Kernel side interface** (Not merged yet.)

```
# cat /sys/block/sda/queue/unpriv_sgio
0
# echo 1 > /sys/block/sda/queue/unpriv_sgio
```

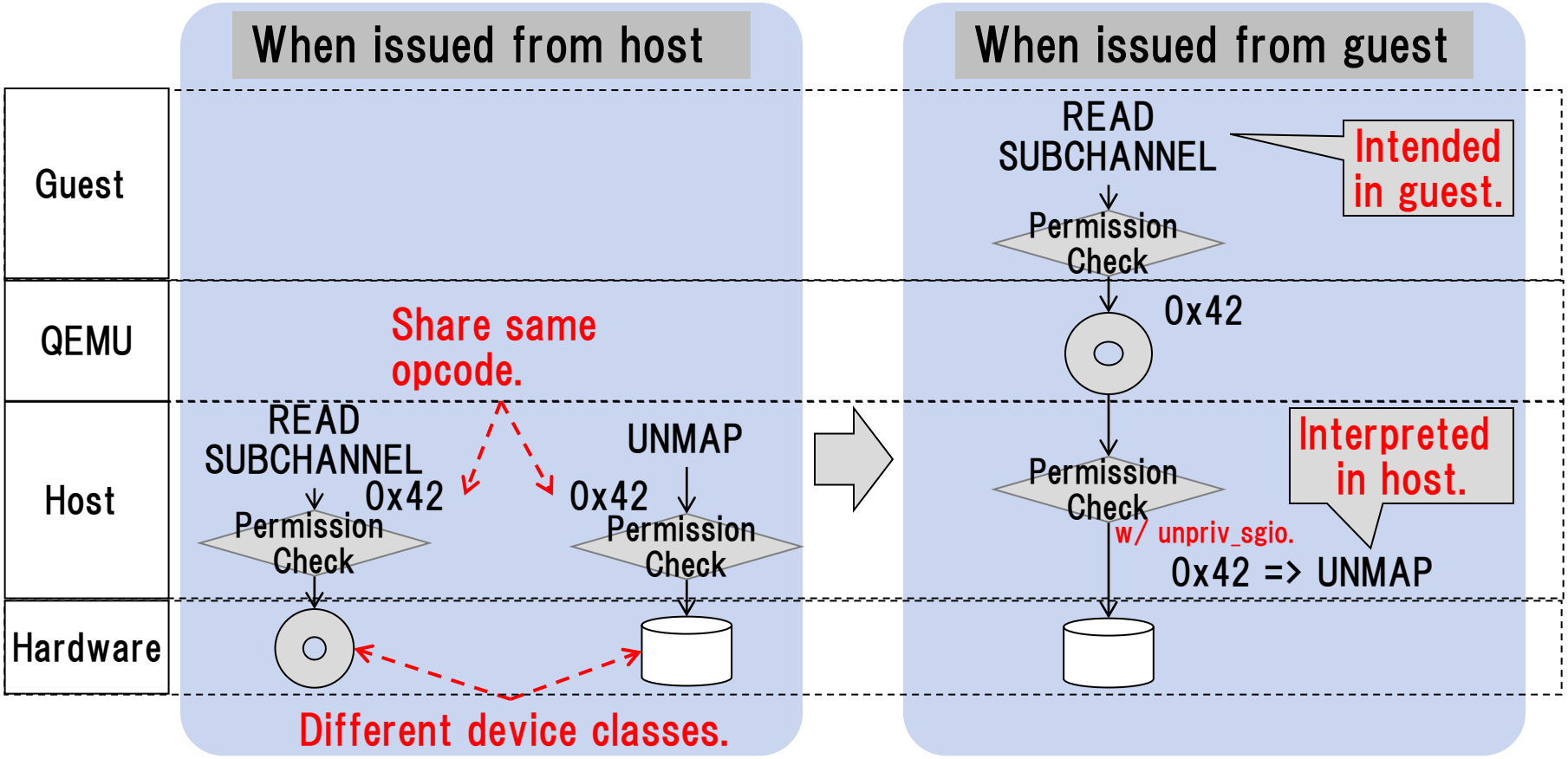
(*) Unpriv_sgio flag can be set per disk.

If this kernel patch is merged, KVM guest running as qemu user will be able to configure to issue any SCSI commands to storages.

3-5. (a) Whether SCSI commands reach to storage(4/4)

HITACHI
Inspire the Next

Why these patches have not been merged yet?
➔ Still under discussion on how to avoid opcodes-overlap problem.



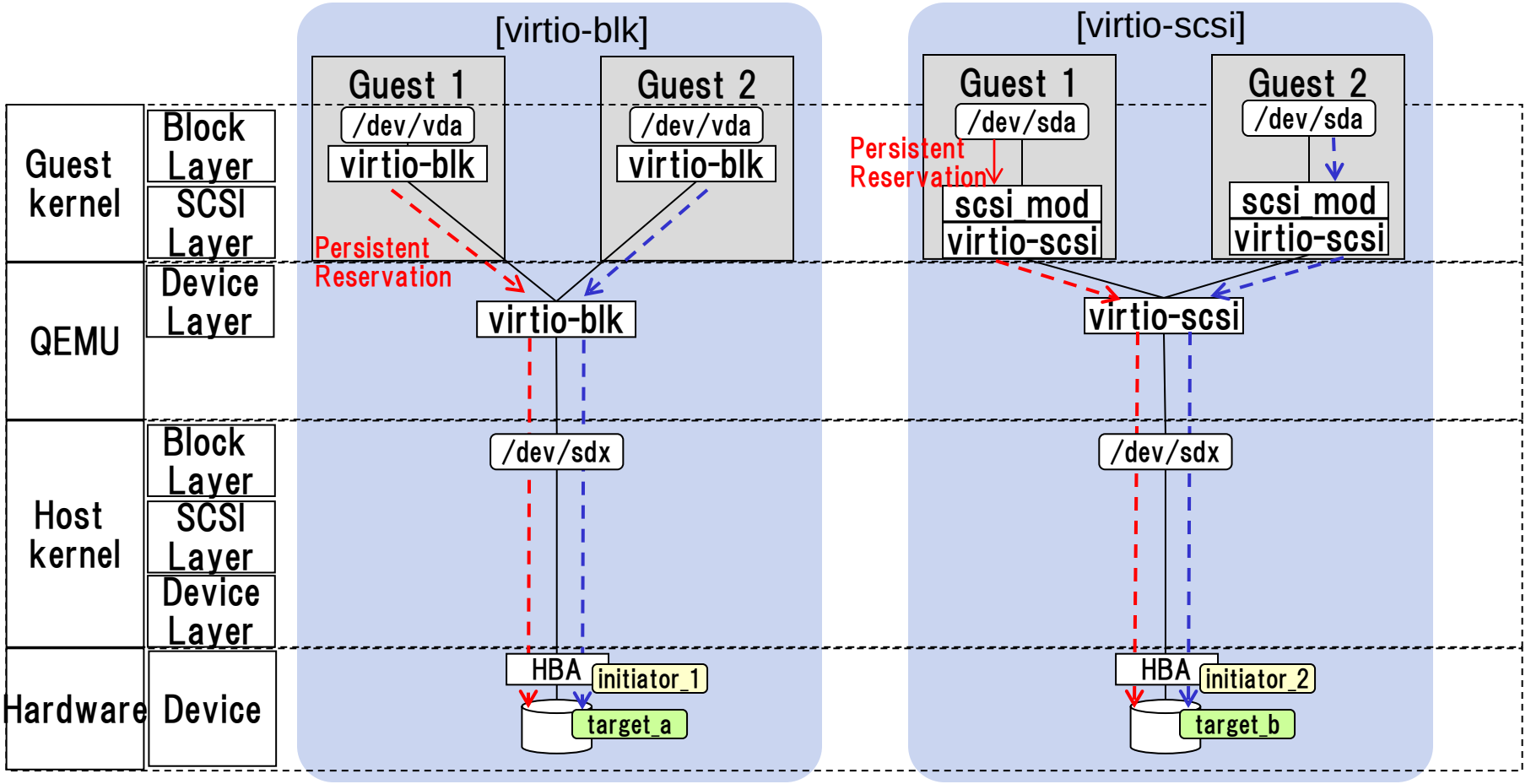
[Problem] : Destructive commands might pass-throughed to the host from guests.
[Implementation]: Split permission check by device class (*1)
or Introduce per-device filter w/o unpriv_sgio (*2)

(*1) https://lkml.org/lkml/2013/1/24/299 (*2) https://lkml.org/lkml/2013/5/27/230

3-6. (b) Whether unique initiator ID is assigned(1/3)

HITACHI
Inspire the Next

Problem When **both guest1 and guest2** are on the same host and use the same HBA, they **share the same initiator ID**. (virtio-blk or vitio-scsi w/ qemu target)



initiator_* : Initiator ID. target_* : Target ID.

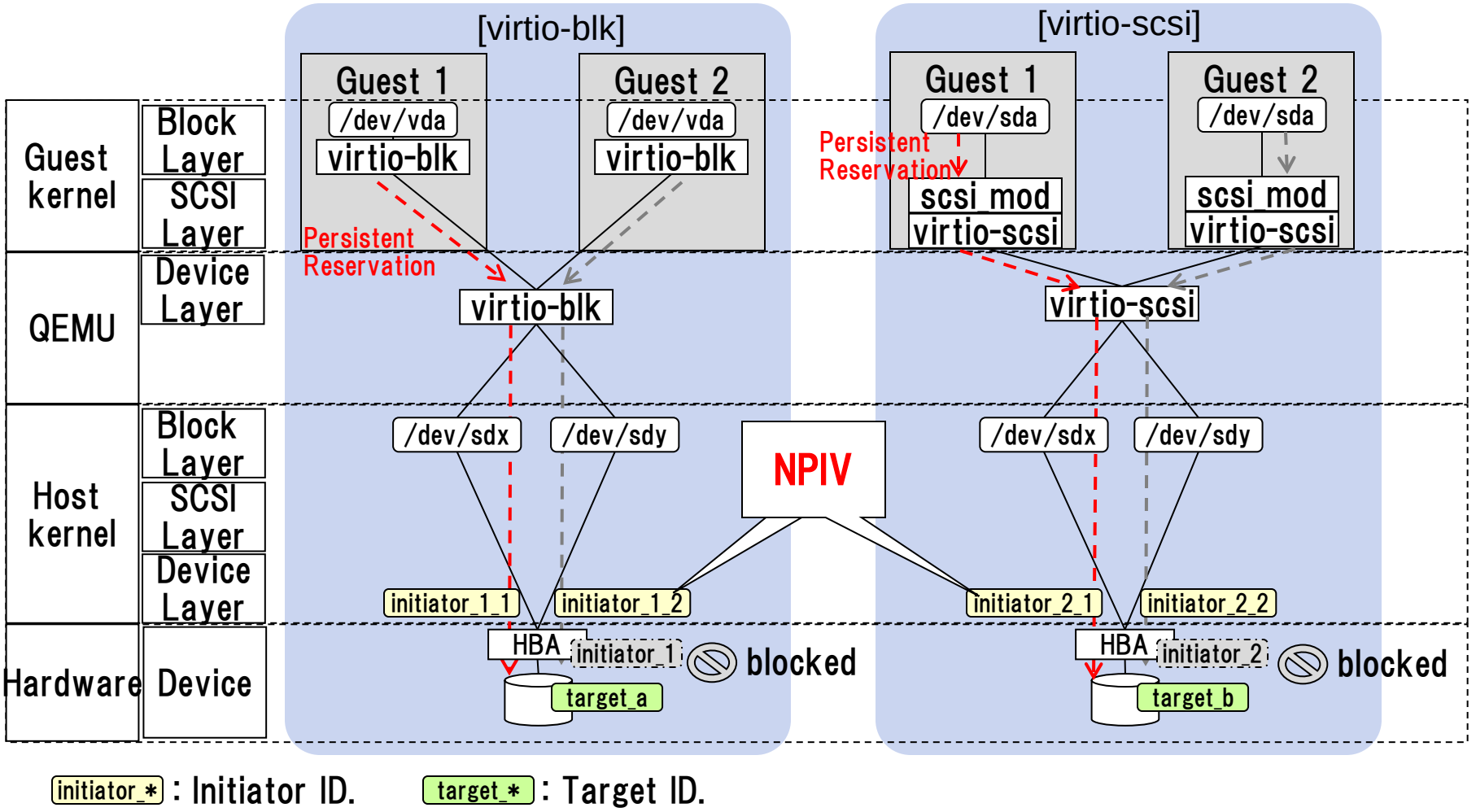
➔ In such a condition, exclusive access is not guaranteed.

3-7. (b) Whether unique initiator ID is assigned(2/3)

HITACHI
Inspire the Next

Solution

With NPIV (N-Port ID Virtualization),
virtio-blk and vitio-scsi w/ qemu target can assign unique initiator ID.



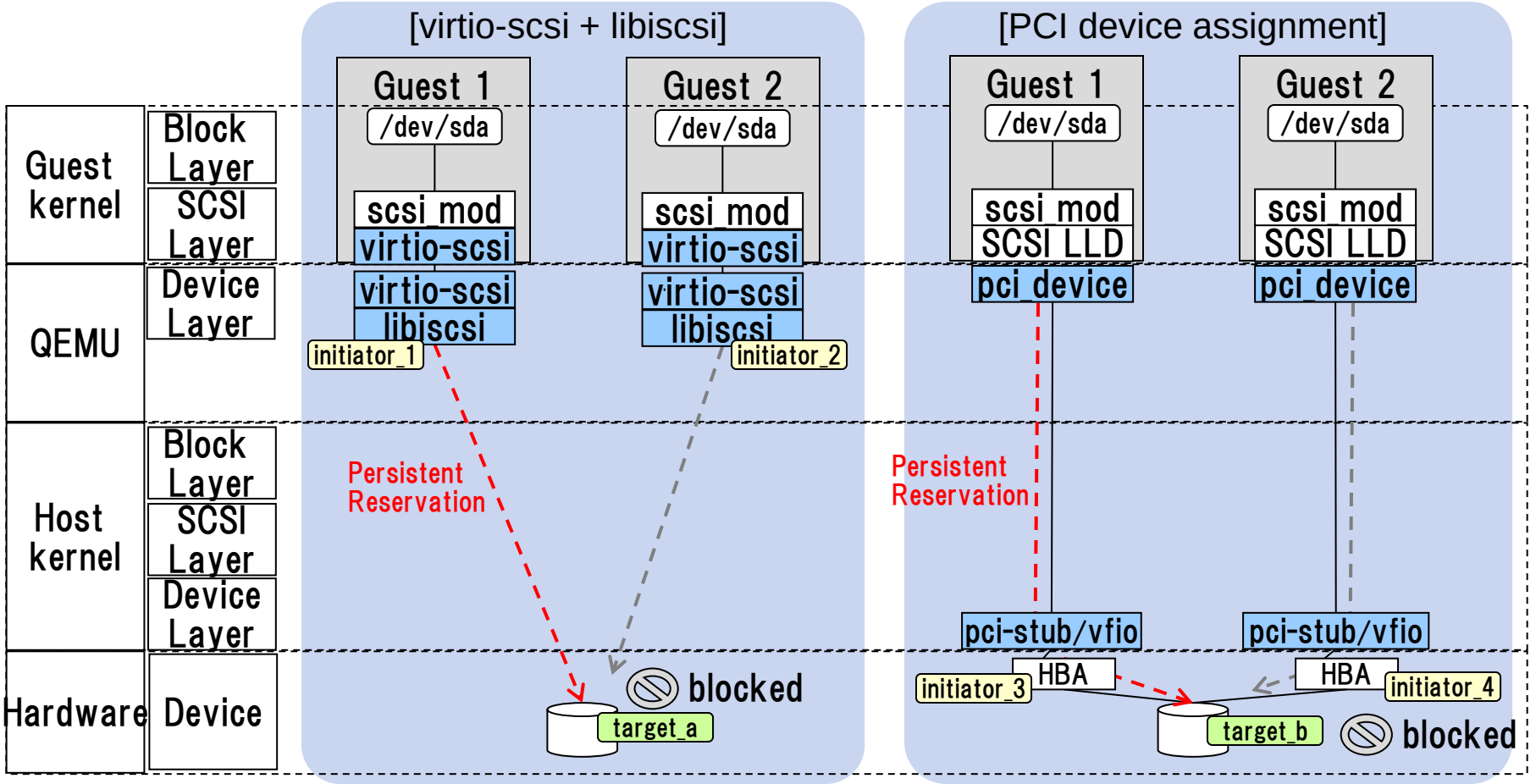
initiator_* : Initiator ID. target_* : Target ID.

3-8. (b) Whether unique initiator ID is assigned(3/3)

HITACHI
Inspire the Next

FYI

With libiscsi or PCI device assignment, exclusive access is guaranteed, because initiator IDs are unique with these configurations.



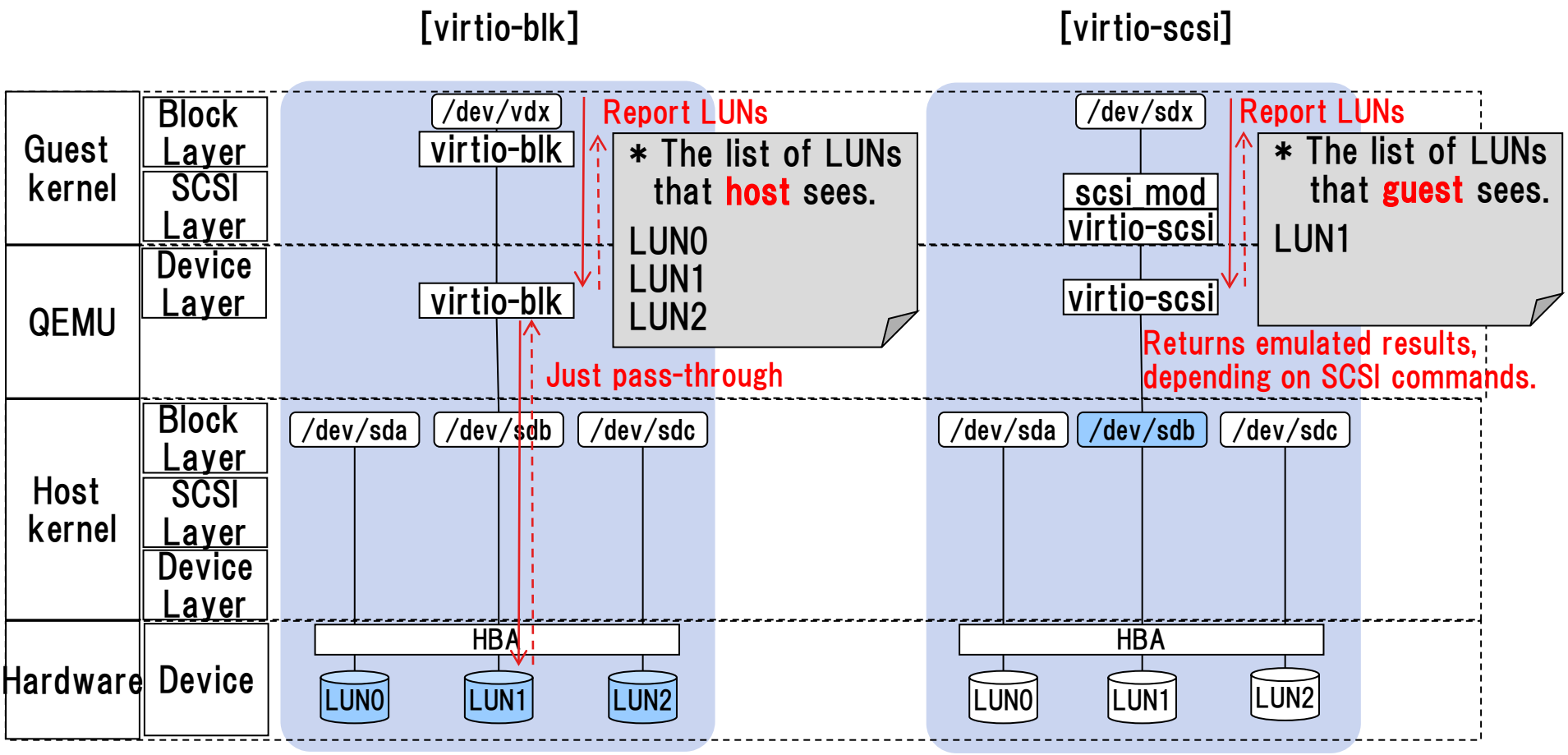
initiator_* : Initiator ID. target_* : Target ID.

3-9. (c) Whether SCSI commands return proper results

HITACHI
Inspire the Next

Problem

With virtio-blk, Report LUNs returns the list of LUNs including LUNs which are not assigned to the guest.



➔ Virtio-blk needs emulation functions to return proper results for particular SCSI commands, such as Report LUNs.

HITACHI
Inspire the Next

Better Utilization of Storage Features from KVM Guest via virtio-scsi

4. Summary

4. Summary

HITACHI
Inspire the Next

- Enterprise system requires SCSI commands in virtualized environments.
- KVM has some configurations which can issue SCSI commands to storage from guests, however each configuration has some restrictions.

#	Configuration		SCSI command				Unique WWN	Restriction
			Persistent Reservation	Write Same	Inquiry	Report LUNs		
1	virtio-blk	-	No	No	Yes	No	Yes	Requires NPIV for unique WWN.
2	virtio-scsi	qemu	No	No	Yes	Yes	Yes	Requires NPIV for unique WWN.
3		libiscsi	Yes	Yes	Yes	Yes	Yes	iSCSI only.
4	PCI Device assignment	Legacy	Yes	Yes	Yes	Yes	Yes	Max number of guests is limited by the number of HBA ports.
5		VFIO	Yes	Yes	Yes	Yes	Yes	

: Already available.
: Patch exists, but not merged yet.
: No patch, but could be fixed.

HITACHI
Inspire the Next

Better Utilization of Storage Features from KVM Guest via virtio-scsi

5. Future Work

5. Future work

HITACHI
Inspire the Next

1. To allow qemu user to issue Persistent Reservation and WriteSame, proper permission check in host kernel is needed.

2. To make Report LUNs return proper results with virtio-blk, an emulation function for Report LUNs is needed.

3. To make SCSI command capability of virtio-scsi w/ lio target clear, evaluation is needed for virtio-scsi w/ lio target.

Questions?

HITACHI
Inspire the Next

END

Better Utilization of Storage Features from KVM Guest via virtio-scsi

LinuxCon North America

Masaki Kimura

<masaki.kimura.kz@hitachi.com>

Information & Telecommunication Systems Company
IT Platform Division Group

Hitachi Ltd.

Trademarks

HITACHI
Inspire the Next

- Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries.

Human Dreams.
Make IT Real.

We will launch innovations that make people's dreams come true through IT, through control technology, and through social infrastructure systems.


HITACHI
Inspire the Next