HITACHI
Inspire the Next

SCSI Device Naming

LinuxCon Japan 2011/06/01

Yokohama Research Laboratory
Linux Technology Center

Nao Nishijima

Overview

HITACHI
Inspire the Next

Introducing preferred name for user operations

This presentation will show the following

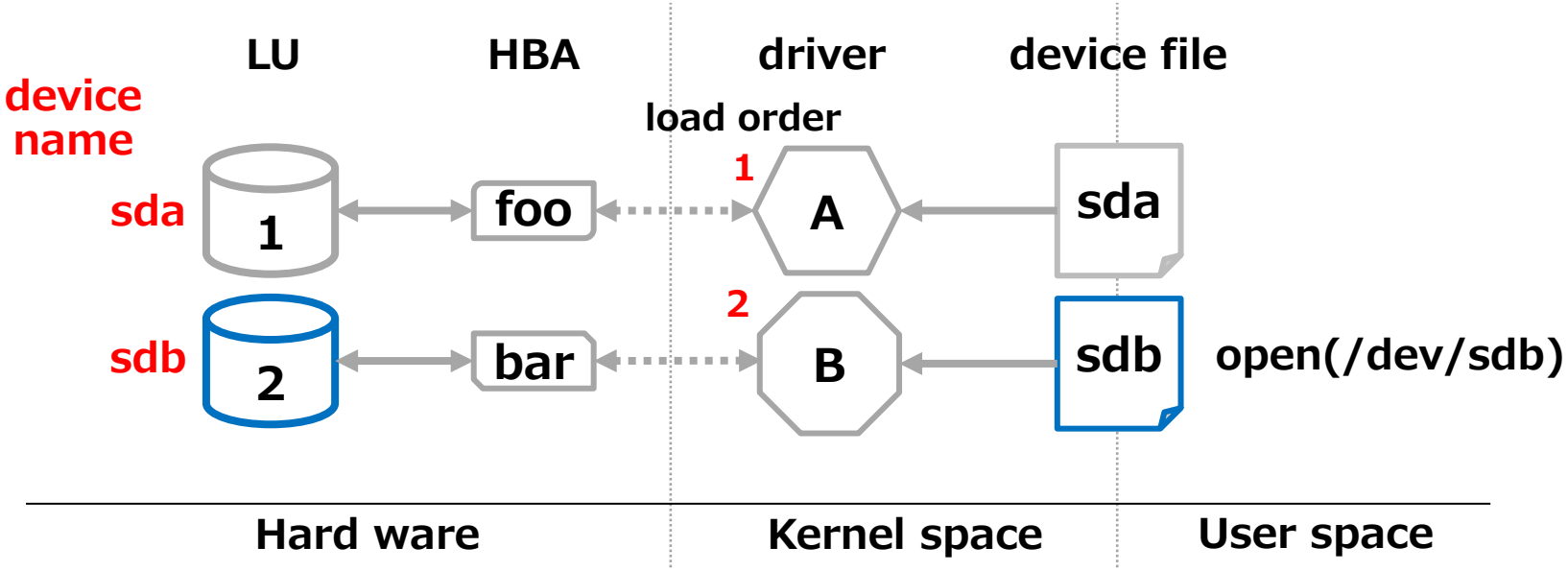
- What is the SCSI device name?
- Issues of udev's persistent device names
- Introducing two suggestions to solve those issues
- Remaining issues and those solutions
- Conclusion

What is the SCSI device name?

HITACHI
Inspire the Next

The SCSI device name identifies each SCSI device

- Kernel assigns a name to a device (e.g. sda)
- The name is used by device file name
- Applications can use these device files to access the device

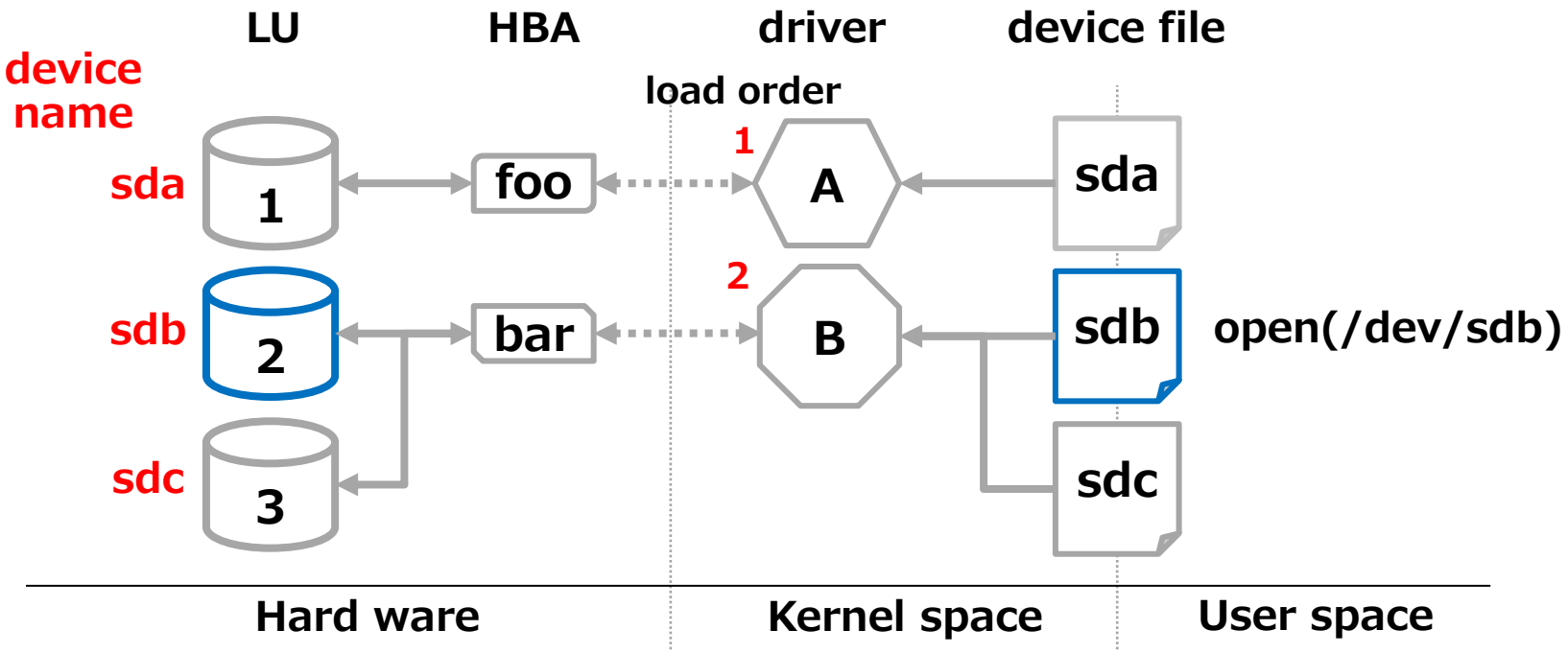


How to assign a device name to a device

HITACHI
Inspire the Next

Device names depend on enumeration of devices

- Device driver loading
- Device scanning (usually from small bus number)

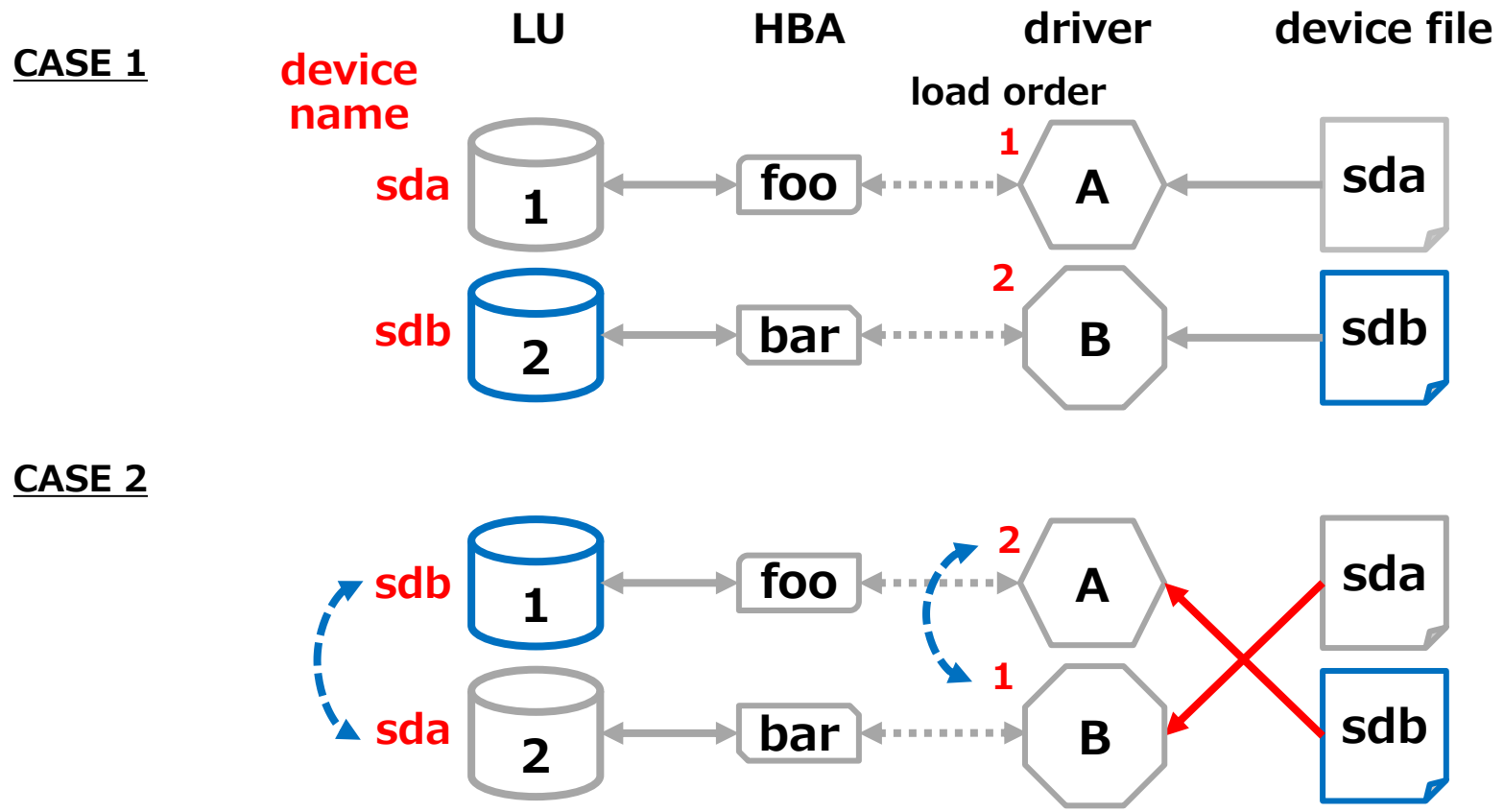


Swapping the device names

HITACHI
Inspire the Next

E.g.) Changing the order of driver loading

- Udev processes load drivers **asynchronously**.



What happens on device name swapping?

HITACHI
Inspire the Next

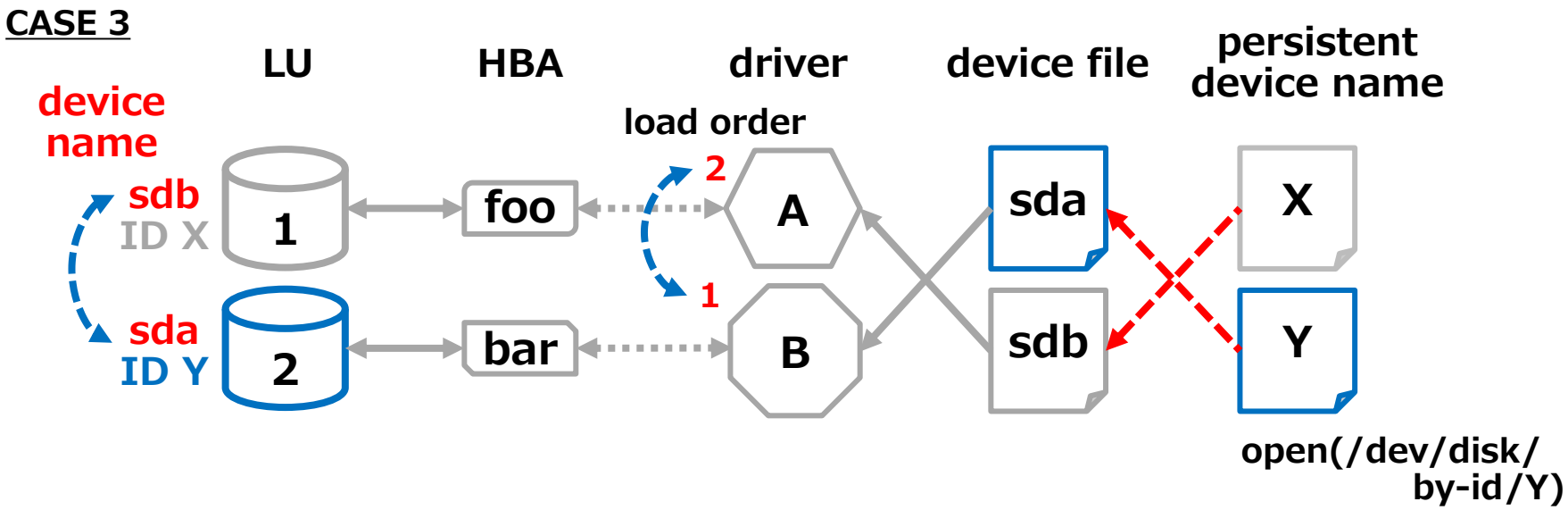
When swapping happens --

- Kernel may fail to mount root device
- You mount device to incorrect place since fstab is broken
- Application may write a data to incorrect device
- Commands(e.g. iostat) may show different device
  - The device name does not always point at the same device

HITACHI
Inspire the Next

We use persistent device name with symbolic links

- Udev makes symbolic links as persistent alias
  - Symbolic links identify a device by device's id (uuid, path, serial)
  - Symbolic links point at device file witch has correct device's id
  - we can access the same device anytime via persistent device name



HITACHI
Inspire the Next

Udev's solution can solve some issues

- Kernel may fail to mount root device
  - Currently, we use persistent device name for identifying devices
- You mount device to incorrect place since fstab is broken
  - Currently, we use persistent device name!
- Application may write a data to incorrect device
  - We can use persistent device name

However, udev's solution can not solve this issue

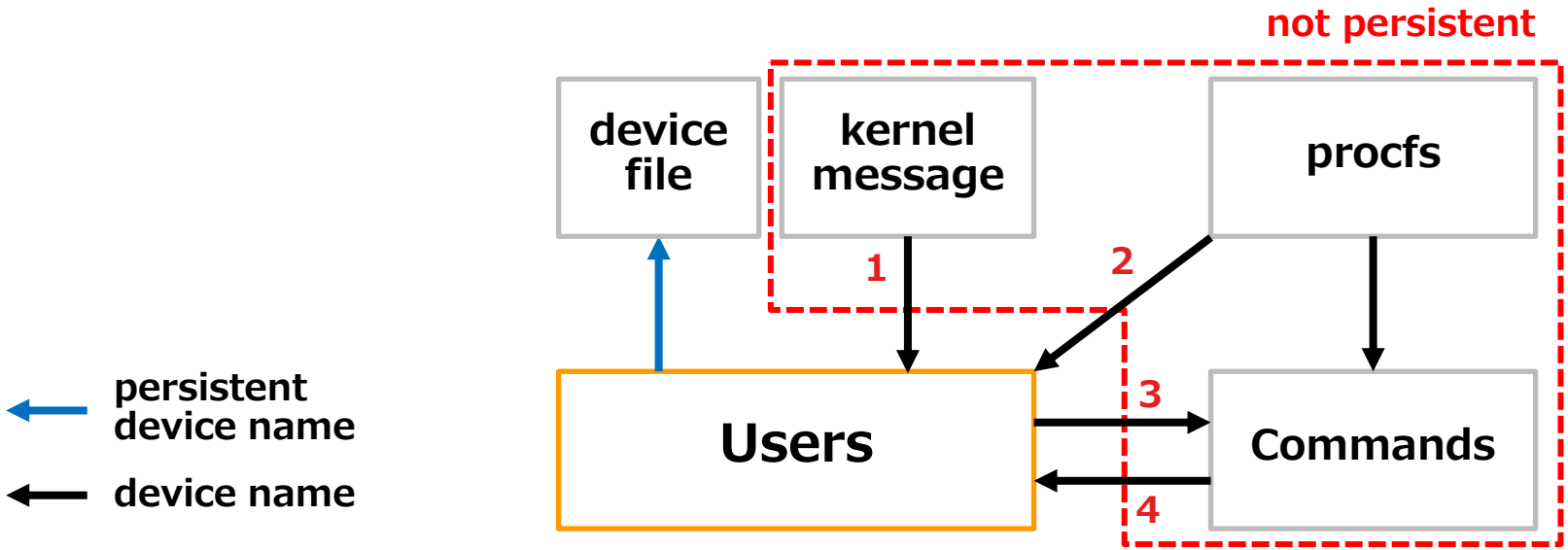
- Commands(e.g. iostat) may show different device

Udev's solution is not enough to solve issues

HITACHI
Inspire the Next

We still see device names in those component

1. The kernel messages
2. The procfs
3. Command arguments
4. Command messages



The kernel messages

HITACHI
Inspire the Next

The persistent device name doesn’t appear in the kernel messages

- The kernel notifies errors via kernel messages
- It is hard for users to know which disk has the errors!
  - Users recognize physical disks by persistent device name
  - Persistent device name source can be changed for each boot

```
# dmesg
…
EXT4-fs (sda3): re-mounted. Opts: (null)
EXT3-fs: barriers not enabled
kjournald starting.  Commit interval 5 seconds
EXT3-fs (sda6): using internal journal
EXT3-fs (sda6): mounted filesystem with ordered data mode
EXT3-fs: barriers not enabled
kjournald starting.  Commit interval 5 seconds
EXT3-fs (sda1): using internal journal
```

The procfs

HITACHI
Inspire the Next

The persistent device name doesn't appear in the procfs

- /proc/{partitions, diskstats} shows device names
- There is no persistent device names

```
# cat /proc/partitions
major minor  #blocks  name

   8        0  488386584 sda
   8        1     194560 sda1
…
   8       16    1969152 sdb
```

```
# cat /proc/diskstats
   1       0 ram0 0 0 0 0 0 0 0 0 0 0 0
…
   7       0 loop0 0 0 0 0 0 0 0 0 0 0 0
…
   8       0 sda 48191 17318 1608679 152013 17665 48659 538708 506492 0 58462
   8       1 sda1 601 240 4732 1098 7 0 20 154 0 1249 1251
…
   8      16 sdb 324 12 2688 397 0 0 0 0 0 396 396
…
```

Command arguments

HITACHI
Inspire the Next

The persistent device name is unavailable in arguments

- Some commands do not support persistent device name in arguments
- E.g.) iostat

```
# iostat /dev/sda
…
Device:            tps   Blk_read/s   Blk_wrtn/s   Blk_read   Blk_wrtn
sda               0.51        11.86         1.75    1039041     153418
…

# iostat /dev/disk/by-id/scsi-SATA_WDC_WD5000AAKS-_WD-WCASY6088049
…
Device:            tps   Blk_read/s   Blk_wrtn/s   Blk_read   Blk_wrtn
```

Command messages

HITACHI
Inspire the Next

The persistent device name doesn't appear in command messages

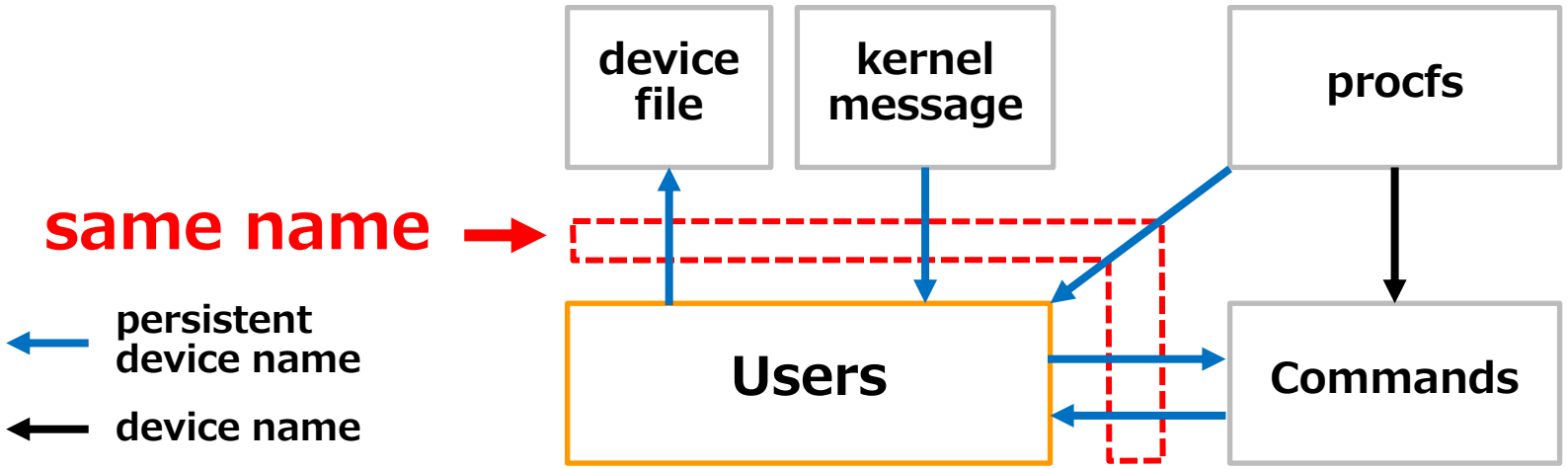
- The device name of command messages does not always point at the same device every boot-up time

```
# iostat
…
Device:            tps   Blk_read/s   Blk_wrtn/s   Blk_read   Blk_wrtn
sda               1.35        11.05         0.02       5577          8
sdb               0.66         5.37         0.00       2712          0
sdc               5.07       198.79         7.45     100370       3760  ←
…                                                                sdc points at LU 1
# reboot
…
# iostat
…
Device:            tps   Blk_read/s   Blk_wrtn/s   Blk_read   Blk_wrtn
sda               0.11         0.89         0.00       5577          8
sdb               0.55        17.15         2.20     107906      13856
sdc               0.05         0.43         0.00       2712          0  ←
…                                                                sdc points at LU 2
```

Our goal

We want to use same name for user operations

- Users can identify the device of
  - The kernel messages
  - The procfs
  - Command messages
- Users can use persistent device name in command arguments



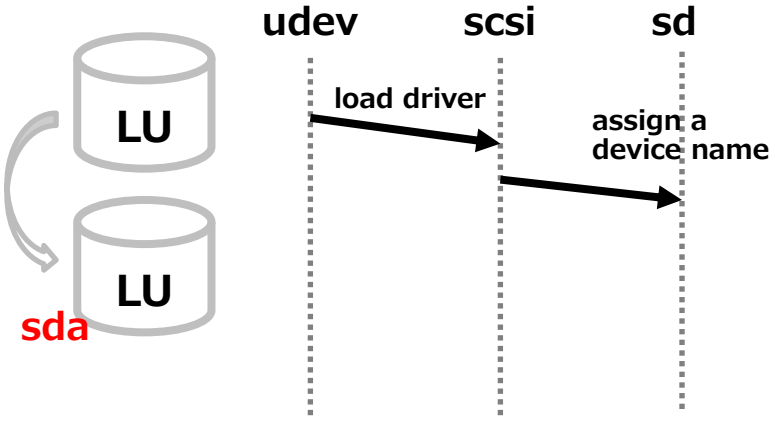
First proposal; the intermediate device

HITACHI
Inspire the Next

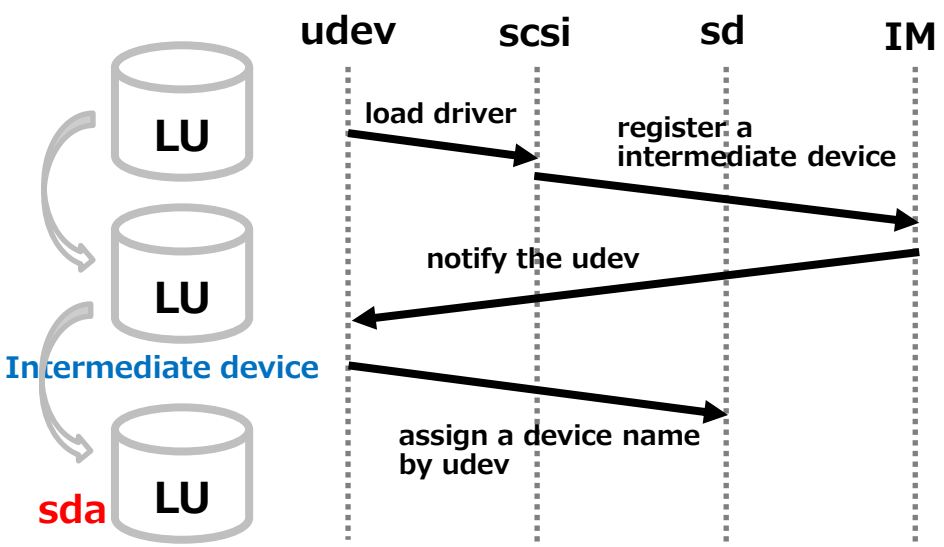
Udev determines **device names**

- New state of device
- Not the order of driver loading
- Small amount of changes

Current device naming flow



Intermediate device naming flow



First proposal Nacked by Maintainers

HITACHI
Inspire the Next

First proposal was rejected

- The kernel cannot always obtain a device ‘s ID
- Not flexible
- Upstream had decided to use udev

“We have been discussing this problem several times in the past.

… We have agreed on using udev to provide persistent device names.”

-- ***Hannes Reinecke***

SCSI driver maintainer

Second proposal; preferred name

HITACHI
Inspire the Next

Add new attribute "**preferred name**"

- Users can assign preferred name to the disk
- The kernel messages show this new attribute
- Not change naming mechanism
- I am going to send the preferred name patch to LKML as soon as possible
- This idea was suggested by James Bottomley
  - **Thank you!**

How to use preferred name

HITACHI
Inspire the Next

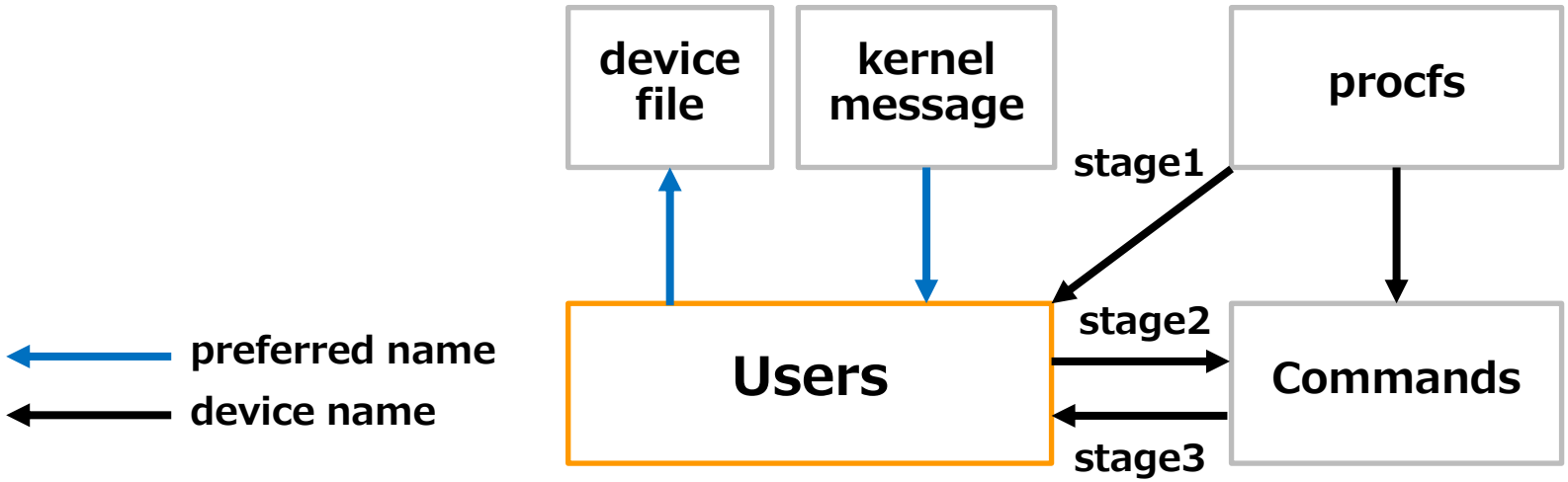
E.g.) The user assigns “foo” to sda

```
# dmesg
…
EXT4-fs (sda1): warning: maximal mount count reached, running e2fsck is
recommended
EXT4-fs (sda1): mounted filesystem with ordered data mode. Opts: (null)
…
# echo foo > /sys/block/sda/preferred_name
# cat /sys/block/sda/preferred_name
foo

# ls -l /dev/disk/by-preferred/foo
lrwxrwxrwx 1 root root 10 May 16 14:25 /dev/disk/by-preferred/foo -> ../../sda

# umount /dev/disk/by-preferred/foo1
# mount /dev/disk/by-preferred/foo1 /mnt
# dmesg
…
EXT4-fs (foo1): warning: maximal mount count reached, running e2fsck is
recommended
EXT4-fs (foo1): mounted filesystem with ordered data mode. Opts: (null)
```

HITACHI
Inspire the Next

Preferred name can solve just one issue

- We still have three issues
- Separate three stage
  - stage1: The procfs issue
  - stage2: Command arguments issue
  - stage3: Command messages issue



HITACHI
Inspire the Next

The procfs should output preferred names

- Output form (discussion point)
  - Replace
  - Add column
  - New partitions and so on.
- Users can see preferred names in the procfs

E.g. Replace

```
# cat /proc/partitions
major minor  #blocks  name

   8        0  488386584 root
   8        1     194560 root1
…
```

E.g. Add column

```
# cat /proc/partitions
major minor  #blocks  name preferred

   8        0  488386584 sad root
   8        1     194560 sda1 root1
…
```

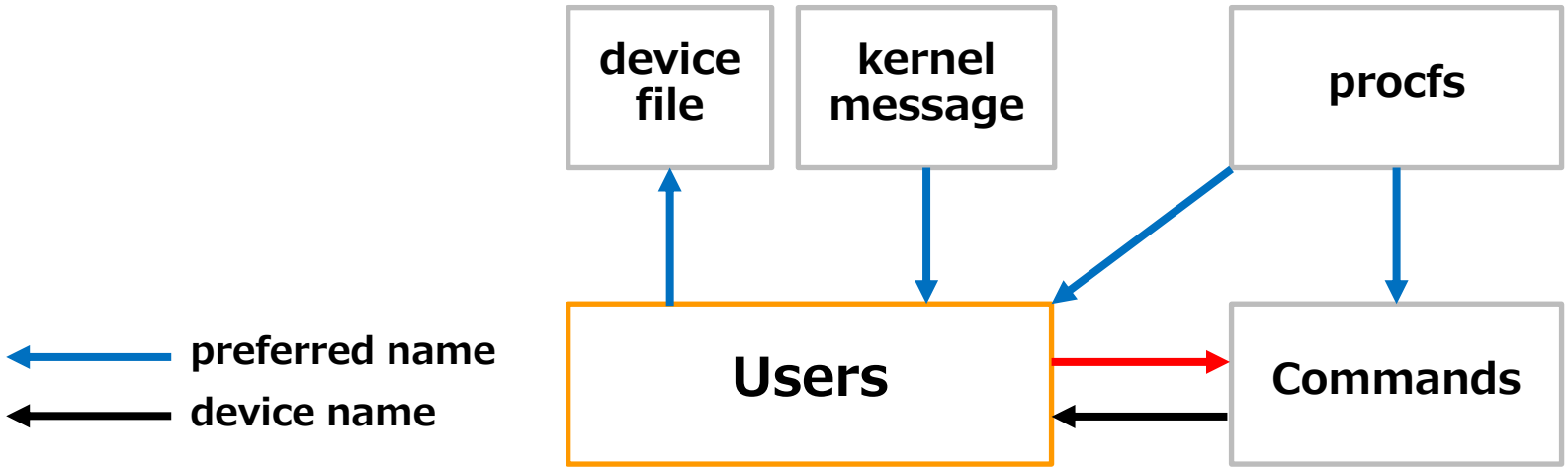
E.g. New partitions

```
# cat /proc/preferred_partitions
major minor  #blocks  name

   8        0  488386584 root
   8        1     194560 root1
…
```

HITACHI
Inspire the Next

Commands should accept preferred names

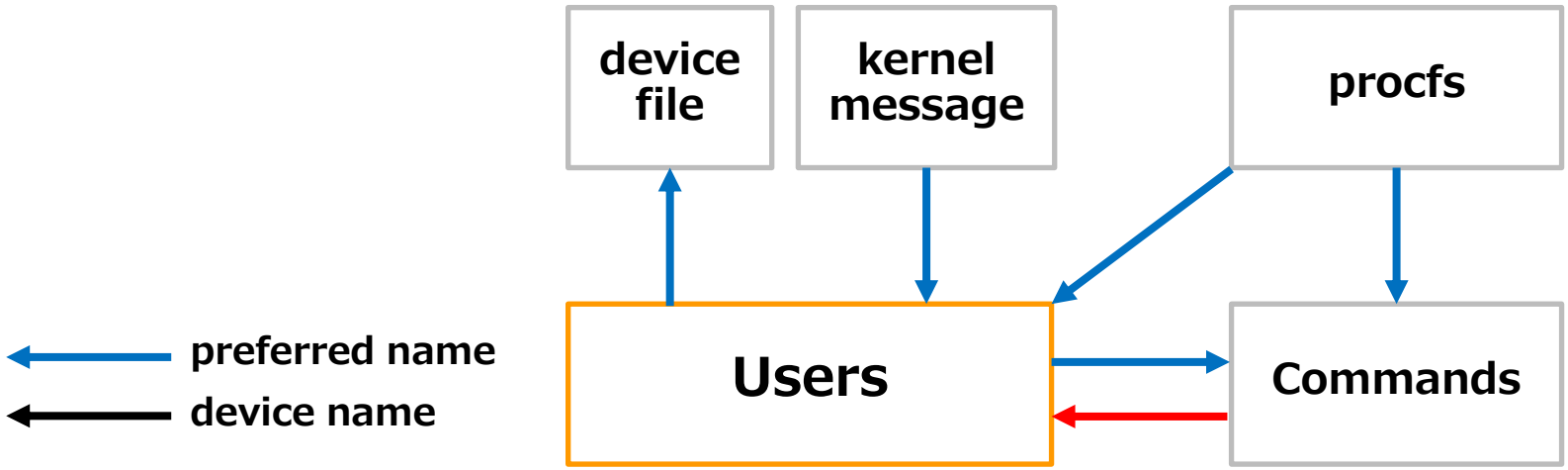
- To use preferred names in arguments, commands must be fix to use the readlink()



HITACHI
Inspire the Next

Commands should show preferred names

- To use preferred names of the procfs
- To read preferred name from /sys/block/***device name***/preferred_name file.



Summary of stages

HITACHI
Inspire the Next

There are three stages for solving all issues

- Persistent device names has four issues
- Preferred name can solve the kernel messages issue
- We need to solve remaining issues step-by-step

	P1: The kernel messages	P2: The procfs	P3: Command arguments	P4: Command messages
persistent device name	**NG**	**NG**	**partially**	**partially**
preferred name	**OK**	**NG**	**partially**	**partially**
+ stage 1	**OK**	**OK**	**partially**	**partially**
+ stage 2	**OK**	**OK**	**OK**	**partially**
+ stage 3	**OK**	**OK**	**OK**	**OK**

Conclusion

HITACHI
Inspire the Next

Summary

- The device names may change
- Persistent device names has four issues
- Preferred name can solve one of them
- I'd like to solve remaining issues step-by-step
  - The procfs issue
  - Commands arguments issue
  - Commands messages issue

Future plans

- I would like to discuss the preferred name on LKML
- I would like to talk to command developer about the preferred name

Trademarks

HITACHI
Inspire the Next

- Linux is a trademark of Linus Torvalds in the United States, other countries, or both.
- Other company, product, or service names may be trademarks or service marks of others.

HITACHI
Inspire the Next

END

SCSI Device Naming

LinuxCon Japan 2011/06/01

Yokohama Research Laboratory
Linux Technology Center

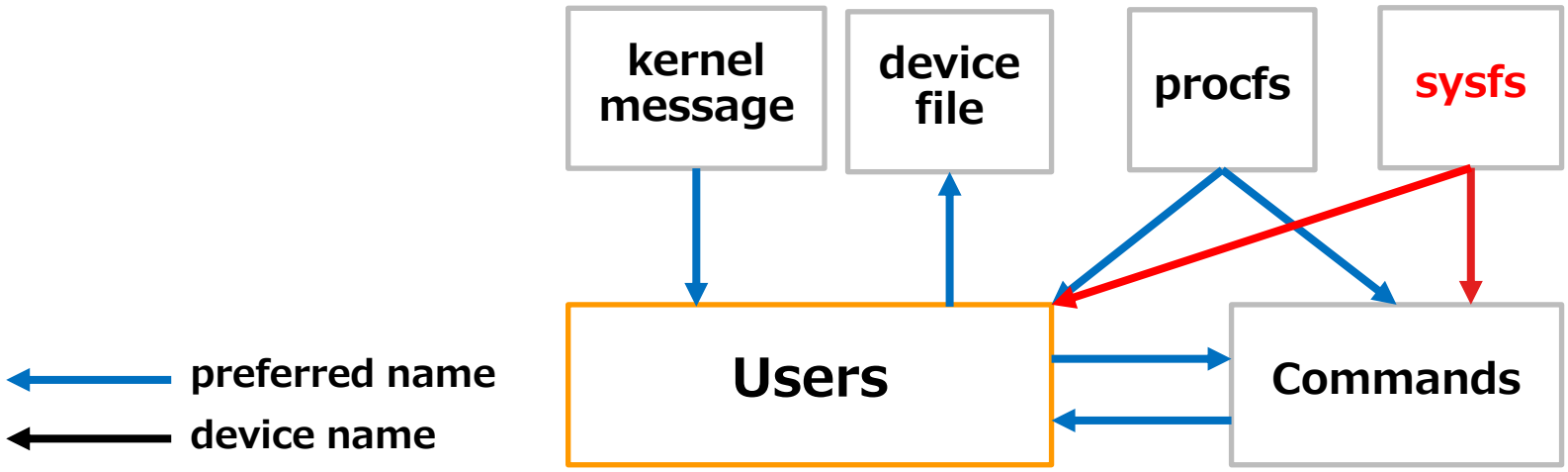
Nao Nishijima


HITACHI
Inspire the Next

HITACHI
Inspire the Next

Do we change all device names to preferred names?

- Sysfs uses device name as symlink
  - /sys/block/sdX -> /sys/devices/pci0000:00/…/sdX
  - /sys/class/block/sdX -> /sys/devices/pci0000:00/…/sdX
- Some commands refers to sysfs to get a device name
- We need to rename symlink name



HITACHI
Inspire the Next

Udev rule

- Udev rule
  - To give the preferred name automatically, we need udev rule.
  - To identify the device, we used by-{id, uuid, path}.
- To do
  - Introduce new udev key (PREFERRED)

udev rules

```
# using by-id
SUBSYSTEM==“block”, ACTION==“add”, ENV{ID_SERIAL}==“WDC_WD5000AAKS-75A7B2_WD-
WCASY6088049”, SYMLINK+=“/disk/by-preferred/pda”, PROGRAM=“write_preferred_name
pda %p”
```

(In the future)

```
# using by-path
SUBSYSTEM==“block”, ACTION==“add”, ENV{ID_PATH}==“pci-0000:00:1f.2-scsi-0:0:0:0”,
SYMLINK+=“/disk/by-preferred/pdb”, PREFERRED=“pdb”
```