Additional Feet for the Penguin – SCSI over FCP Multipathing for Linux on System z

Christof Schmitt
IBM Germany Research & Development

2008-08-15
Session 9289

Abstract

Using storage attachments with less than two independent paths is more than grossly negligent. So the solution is a waterproof multipathing setup. But that sounds easier than it is and there are several configuration pitfalls. This presentation will give you a multipathing overview and lights the multipathing configuration for SCSI devices connected over FCP.

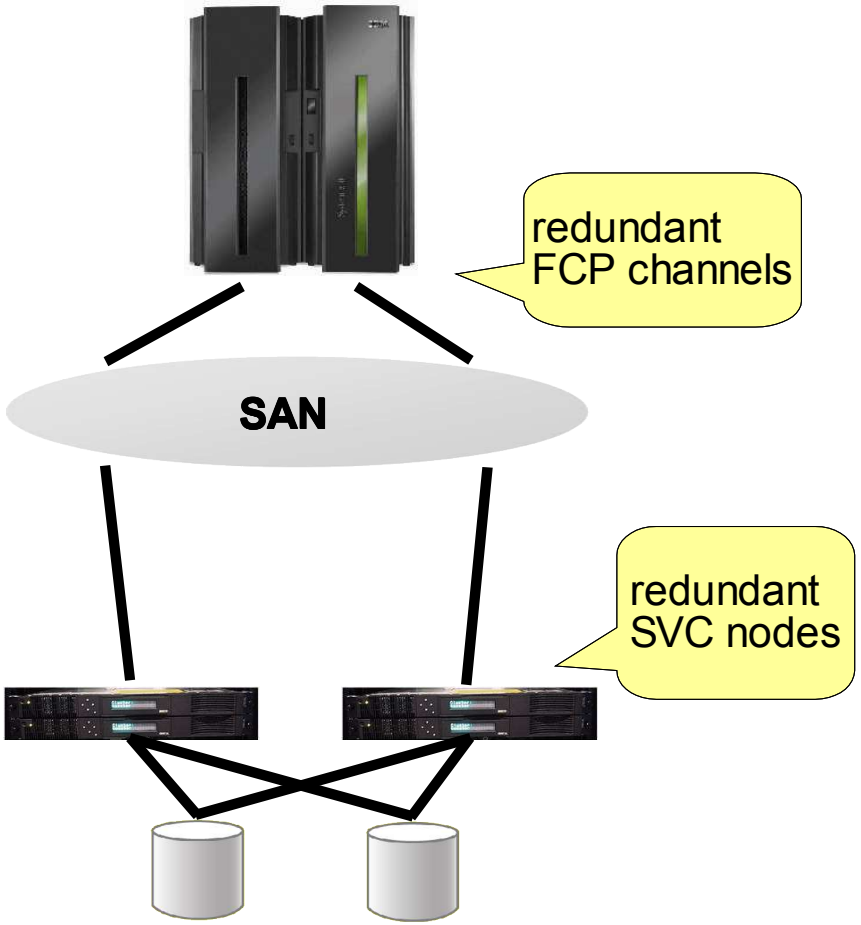
Agenda

- Why multipathing?
- Multipathing for disk storage
- Root filesystem on multipathing device
- Multipathing for IBM tape drives

Why multipathing?

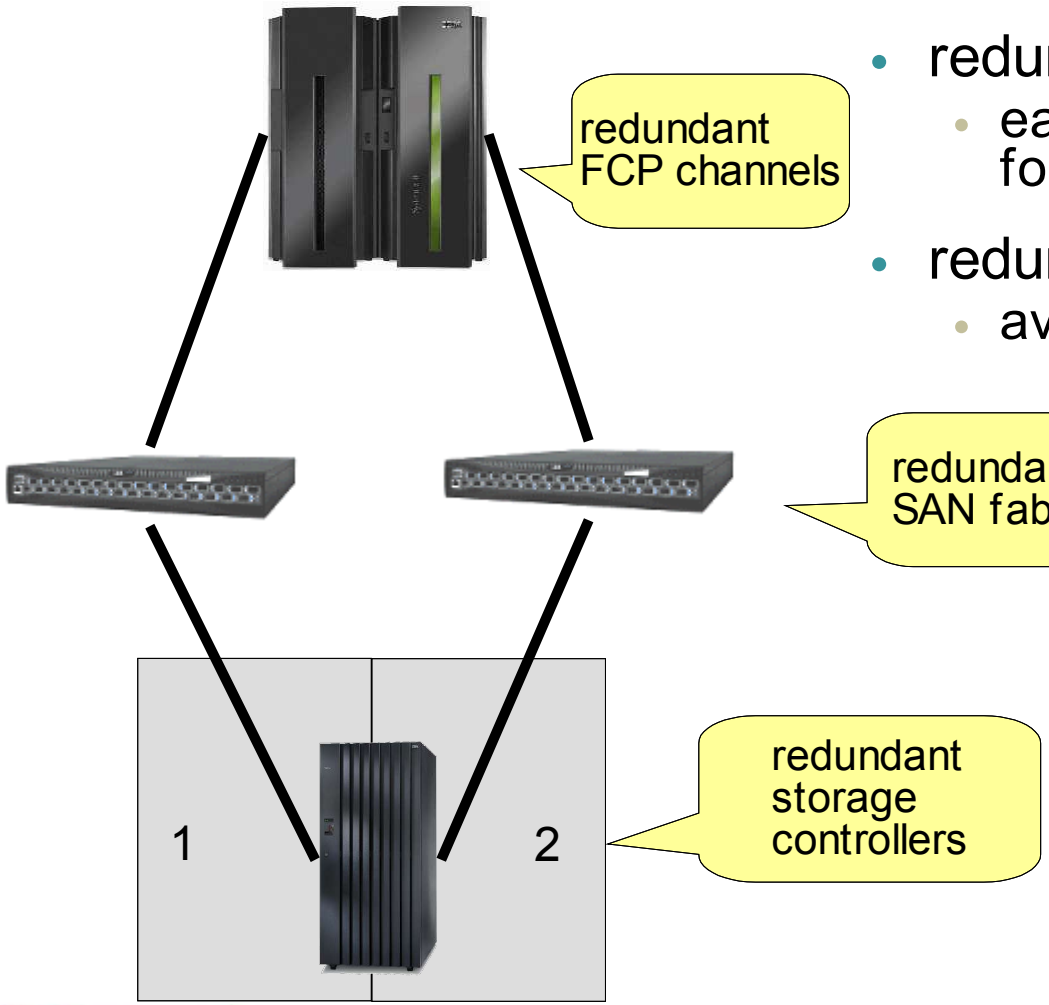
- High availability
  - Access during storage system maintenance
  - Usually required by enterprise disk systems
- Higher performance through load balancing
  - spread I/O load across multiple paths
  - ... across multiple FCP adapters
  - some storage systems use a preferred path
- Failover and Failback
  - hardware maintenance
  - microcode upgrades
  - storage system internal resets

Multipathing for disk storage



- spread load across channels
- keep I/O running during
  - channel recovery
  - configuration changes
- failover and failback
  - during storage maintenance
  - during channel maintenance
  - microcode updates
  - storage internal resets

Mulipathing for disk storage



- redundant controllers in DS8000
  - each controller can be offline for maintenance
- redundant SAN fabrics
  - availability during maintenance

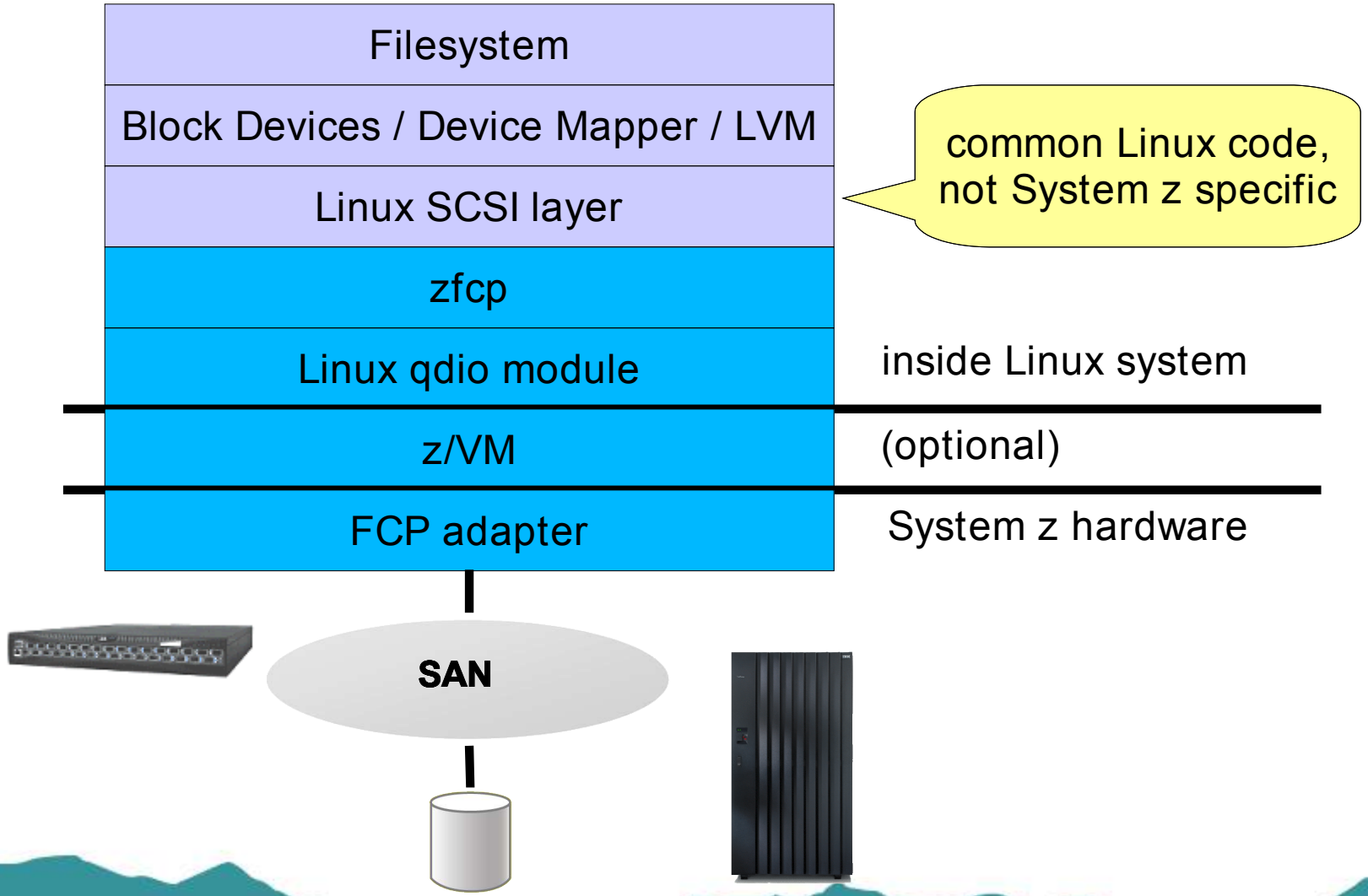
Multipathing for disk storage in Linux

- device-mapper in Linux kernel and multipath-tools
- standard in distributions: RHEL, SLES, ...
- multipathing layer above block devices
- A SCSI device in Linux is now a path!
- Cross-platform: Linux on System z, p, x, ...
- Cross-vendor
- Used throughout storage test and device qualification tests
- supports more than two paths, the following example only uses two

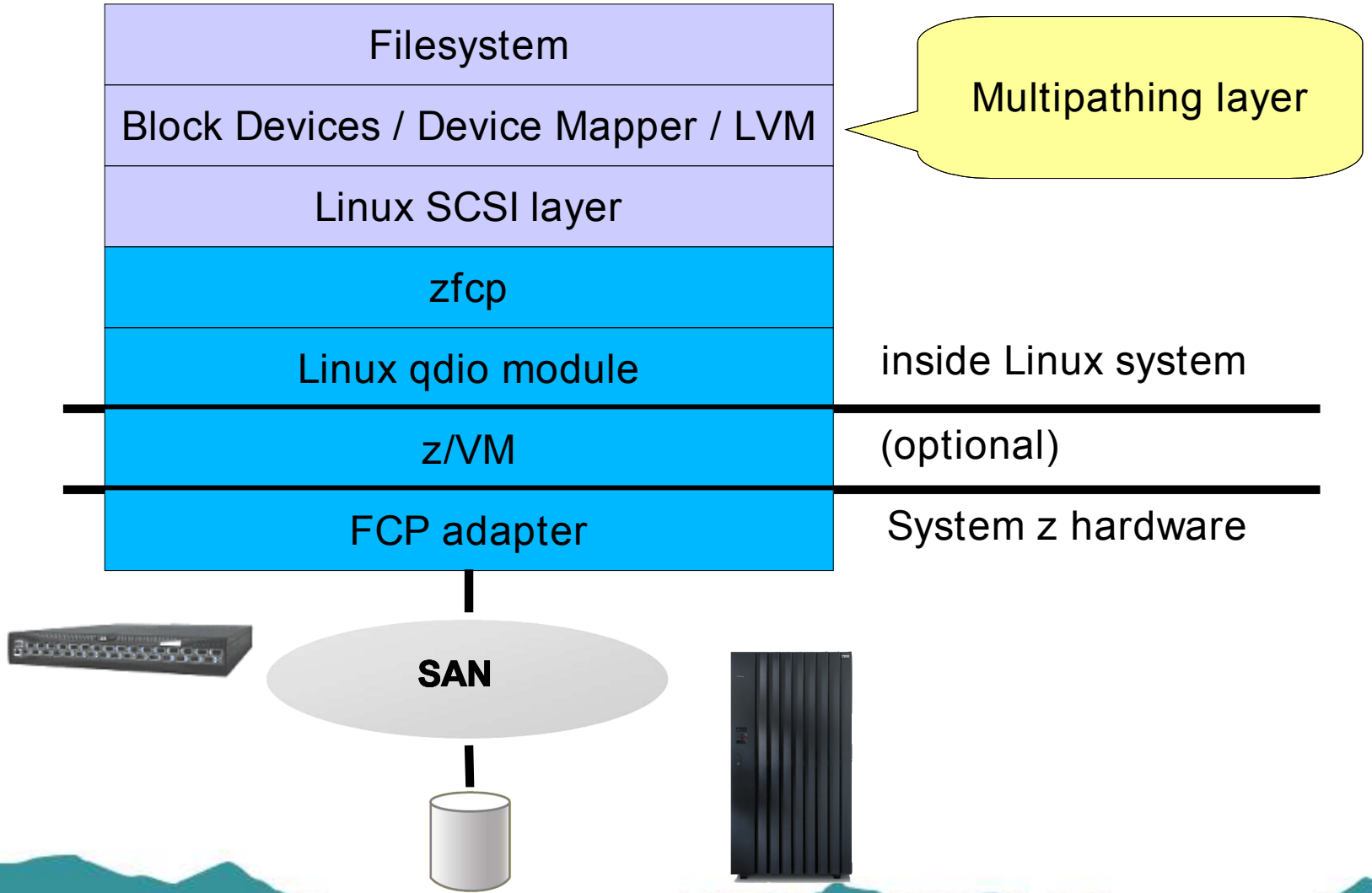
multipathing components

- device-mapper: kernel infrastructure
- multipathd
  - reads configuration
  - establishes setup
  - queries storage
  - checks paths periodically: failback
- command line interface
  - multipath
  - multipathd
- kpartx: helper for partitions on multipath devices
- setup already includes sane default settings

I/O Stack for SCSI on Linux on System z

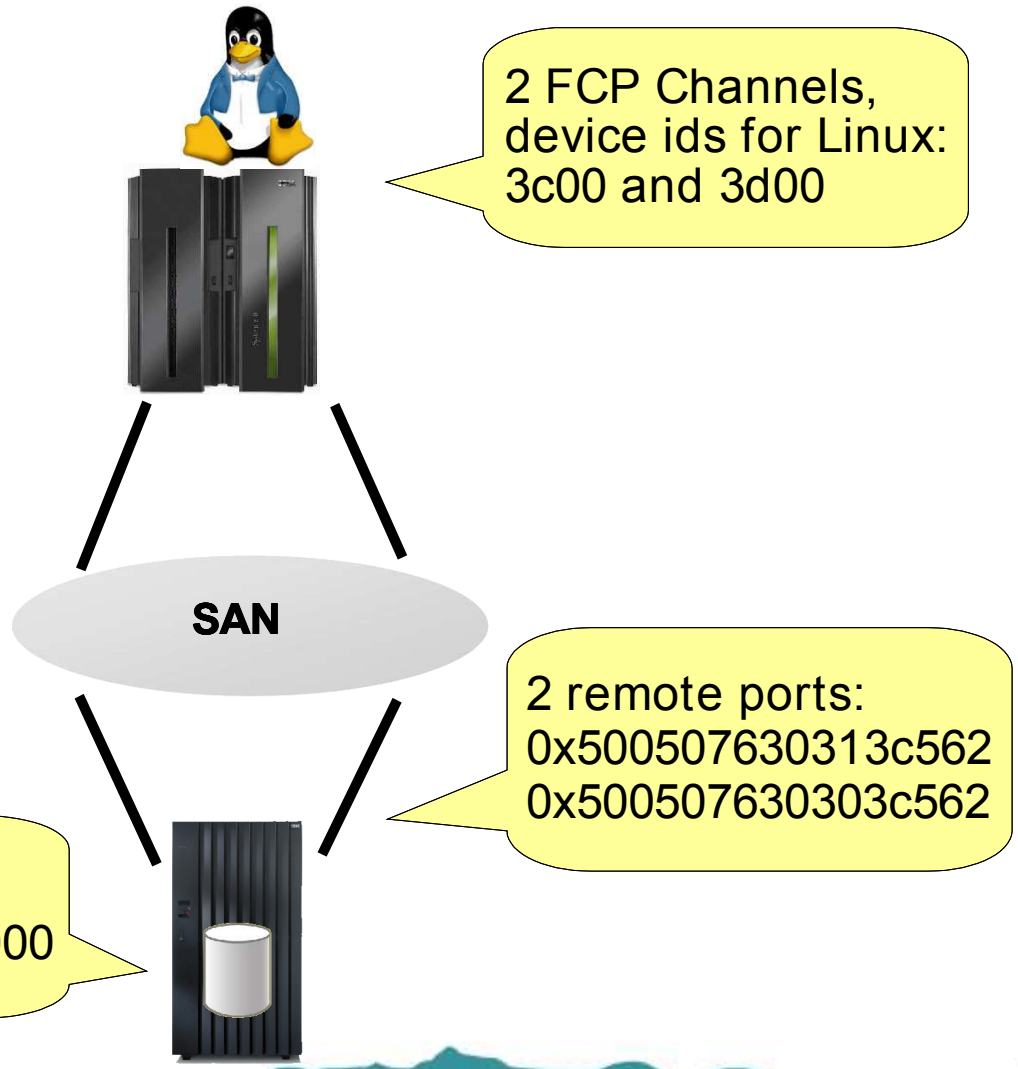


I/O Stack for SCSI on Linux on System z



multipathing example

- one disk
- two paths



multipathing example

- attach ports and units for all paths
- path 1: 0.0.3c00 -> 0x500507630313c562 -> LUN
- path 2: 0.0.3d00 -> 0x500507630303c562 -> LUN
- recommendation: use zfcp config file from distribution
  - RHEL: /etc/zfcp.conf
  - SLES: /etc/sysconfig/hardware/hwcfg-zfcp-bus-ccw-0.0.*

multipathing example

- recommendation: use zfcp config file from distribution
- manual steps would be:

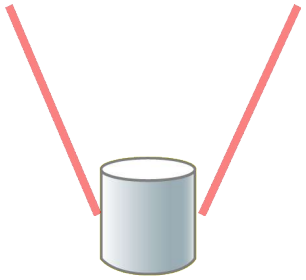
```
# cd /sys/bus/ccw/drivers/zfcp/
# echo 1 > 0.0.3c00/online
# echo 1 > 0.0.3d00/online
# echo 0x500507630313c562 > 0.0.3c00/port_add
# echo 0x500507630303C562 > 0.0.3d00/port_add
# echo 0x401040CC00000000 > 0.0.3c00/0x500507630313c562/unit_add
# echo 0x401040CC00000000 > 0.0.3d00/0x500507630303c562/unit_add
```

multipathing example

- zfcp and SCSI report each path as device

```
# lszfcp -D
0.0.3d00/0x50050763030bc562/0x401040cc00000000 0:0:0:1087127568
0.0.3c00/0x500507630313c562/0x401040cc00000000 1:0:0:1087127568

# lsscsi
[0:0:0:1087127568]disk    IBM      2107900      2.27  /dev/sda
[1:0:0:1087127568]disk    IBM      2107900      2.27  /dev/sdb
```



multipathing example

- 2 SCSI devices for the same disk volume:

```
# scsi_id -g -s /block/sda
36005076303ffc56200000000000010cc
# scsi_id -g -s /block/sdb
36005076303ffc56200000000000010cc
```

Worldwide Identifier (WWID)
Id for storage System + Id for disk volume

- queried from storage system
- multipathd uses this mechanism for mapping paths to disks

multipathing example

- Manually start multipathing (not recommended):

```
# modprobe dm-multipath

# multipathd

# multipath -ll
36005076303ffc56200000000000010cc dm-0 IBM,2107900
[size=5.0G][features=0][hwhandler=0]
\_ round-robin 0 [prio=2][active]
 \_ 1:0:0:1087127568 sdb   8:16  [active][ready]
 \_ 0:0:0:1087127568 sda   8:0   [active][ready]
```

multipathing setup for SLES10

- add all paths to system
  - YaST or edit /etc/sysconfig/hardware/hwcfg-zfcp-*
  - hwup zfcp-bus-ccw-0.0.3c00
- enable device scanning and multipathd
  - chkconfig multipathd on
  - chkconfig boot.multipath on
- reboot or manually start multipath scripts
  - /etc/init.d/boot.multipath start
  - /etc/init.d/multipath start

multipathing setup for RHEL5

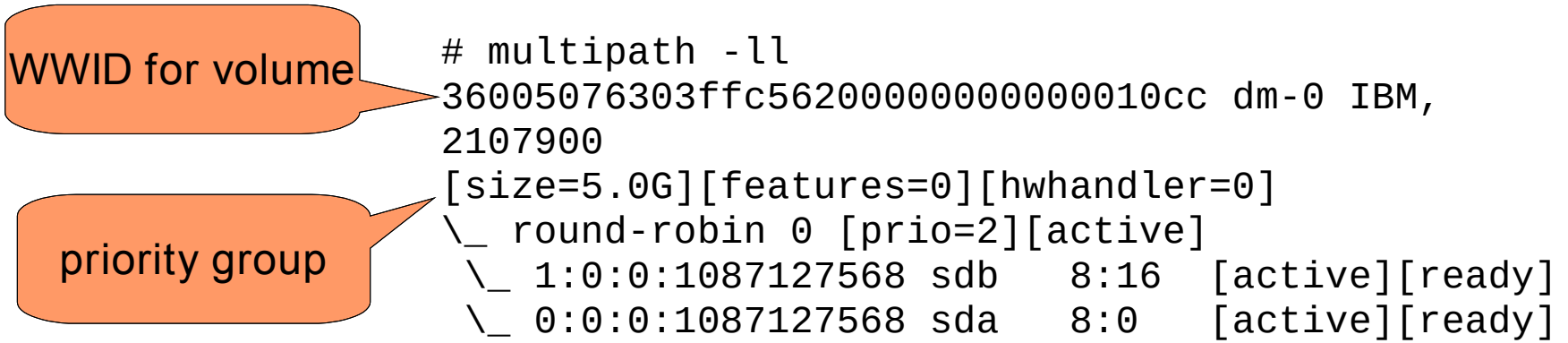
- attach all paths to system
  - /etc/zfcp.conf
  - /sbin/zfcpconf.sh or reboot
- Adjust provided /etc/multipath.conf

```
#blacklist {
#        devnode "*"
#}

#defaults {
#       user_friendly_names yes
#}
```

- chkconfig --add multipathd
- /etc/init.d/multipathd start

Checking multipathing status

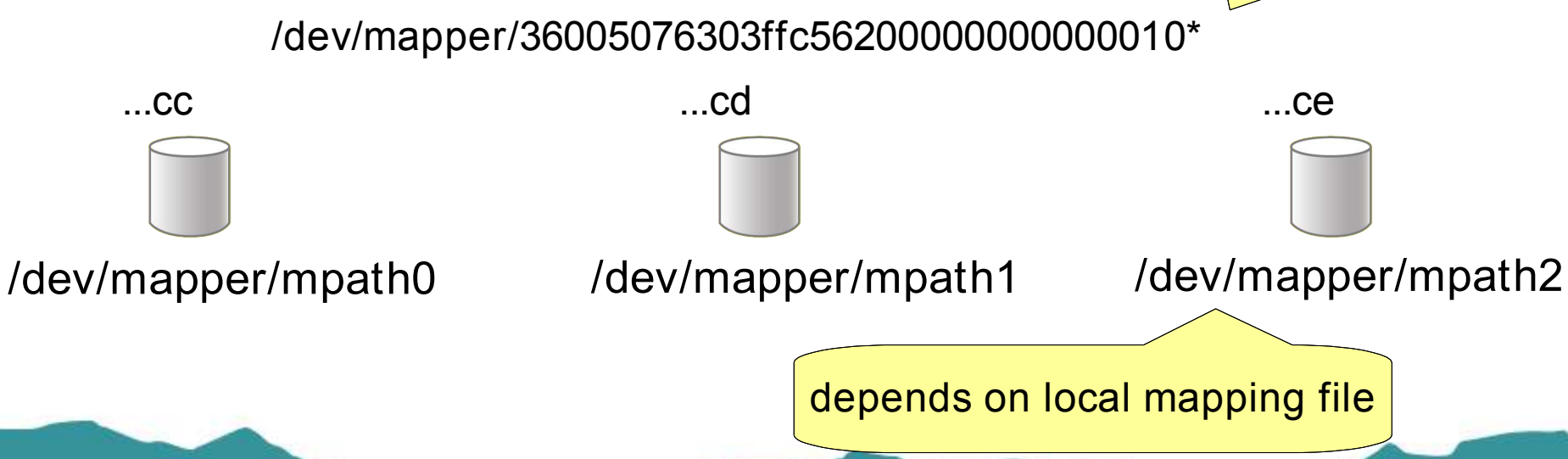


- Paths are combined automatically
- Each path is in one priority group
- multipathing device file /dev/mapper/36005076303ffc56200000000000010cc
- Default settings are good, but can also be changed

Multipathing names and aliases

- user_friendly_names and aliases
  - /dev/mapper/mpath0 instead of /dev/mapper/36005076303ffc56200000000000010cc
- But: WWID is unique, alias maybe not
  - mapping depends on file /var/lib/multipath/bindings
- Recommendation: Use WWIDs

WWIDs from storage system



depends on local mapping file

Multipathing with preferred path

- active / passive controller (DS6000)
- standard storage devices in default hardware table
- 2 pathgroups
  - active/enabled
  - automatically queried from storage (ALUA)

2 priority groups

```
# multipath -ll
3600507630efffca2000000000001229 dm-0 IBM,1750500
[size=3.0G][features=1 queue_if_no_path]
[hwhandler=0]
\_ round-robin 0 [prio=50][active]
 \_ 0:0:0:1076445202 sdaw 67:0   [active][ready]
\_ round-robin 0 [prio=10][enabled]
 \_ 1:0:0:1076445202 sdcb 68:240 [active][ready]
```

hardware table

- combination of
  - default settings
  - redefined settings in /etc/multipath.conf

```
# multipath -t
...
devices {
...
        device {
                vendor "IBM"
                product "1750500"
                path_grouping_policy "group_by_prio"
                path_checker "tur"
                features "1 queue_if_no_path"
                prio "alua"
                failback "immediate"
        }
...
```

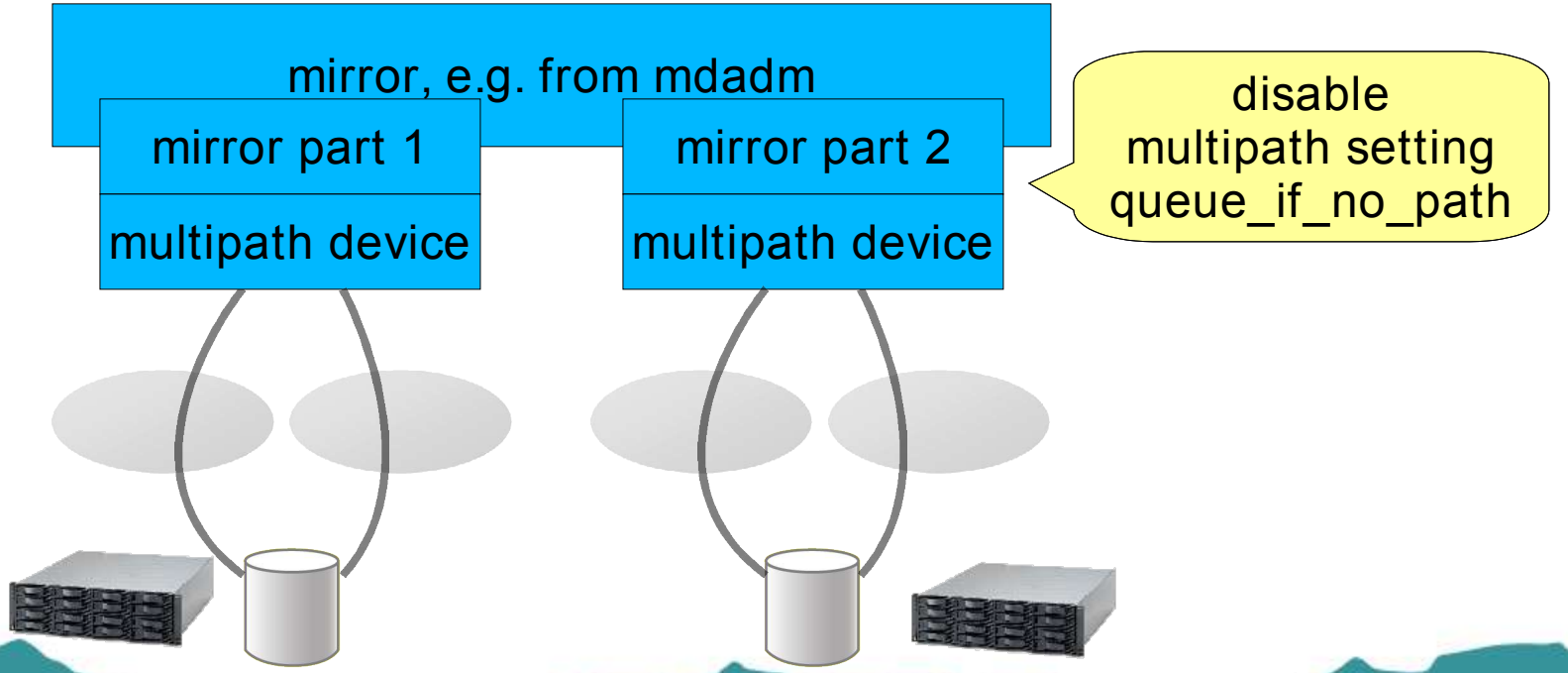
blacklist

- add and change in /etc/multipath.conf

```
# multipath -t
...
blacklist {
        devnode ^(ram|raw|loop|fd|md|dm-|sr|scd|st)[0-9]*
        devnode ^hd[a-z]
        devnode ^dcssblk[0-9]*
        device {
                vendor DGC
                product LUNZ
        }
        device {
                vendor IBM
                product S/390.*
        }
}
blacklist_exceptions {
}
...
```

queue_if_no_path

- set as default
- queues I/O in memory in case all paths are down
- hides path failures from next layer
- disable for software mirror



LVM2 and md on multipathing

- consider queue_if_no_path setting
- mirror and LVM on multipath devices, not SCSI device files!
- setup in /etc/lvm/lvm.conf
  - filter = ["r|/dev/sd*”]

multipathd: more status information

```
# multipathd -k
multipathd> show paths
hcil             dev dev_t pri dm_st   chk_st  next_check
1:0:0:1087127568 sdb 8:16  1   [active][ready] XXXXXXX... 15/20
0:0:0:1087127568 sda 8:0   1   [active][ready] XXXXXXX... 15/20


multipathd> show multipaths status
name                                failback queueing paths dm-st
36005076303ffc56200000000000010cc -        off      2     active


multipathd> show multipaths stats
name                                path_faults switch_grp map_loads
total_q_time q_timeouts
36005076303ffc56200000000000010cc 0           0          1         0
      0


multipathd> help
...
```

Root file system on multipathing device

- example assumed root filesystem on dasd
- root file system on SCSI possible with SCSI IPL
- reliability requirements demand multipathing
- the same applies to swap partition
- Issues:
  - no support for multipathing in distro installers
  - start multipathing before mounting root filesystem
  - zipl does not write IPL record on multipath device file

zipl for multipath device

- add boot entry for single path
  - update procedure
    - boot to single path
    - update
    - zipl
    - boot to multipath
  - reliability during maintenance?
- or use additional disk volume for /boot
  - / on multipath
  - /boot on single path disk volume
  - write zipl IPL record on /boot disk volume
  - IPL /boot disk volume
  - uses additional disk volume
- recommendation: use additional disk volume for reliability

Installing root filesystem on multipath

- Install on single path.
- additional small disk for /boot
- Change to multipath setup after first boot
  - setup second path
  - use multipath device for root filesystem
  - recreate initrd with multipathing
  - zipl for changed initrd

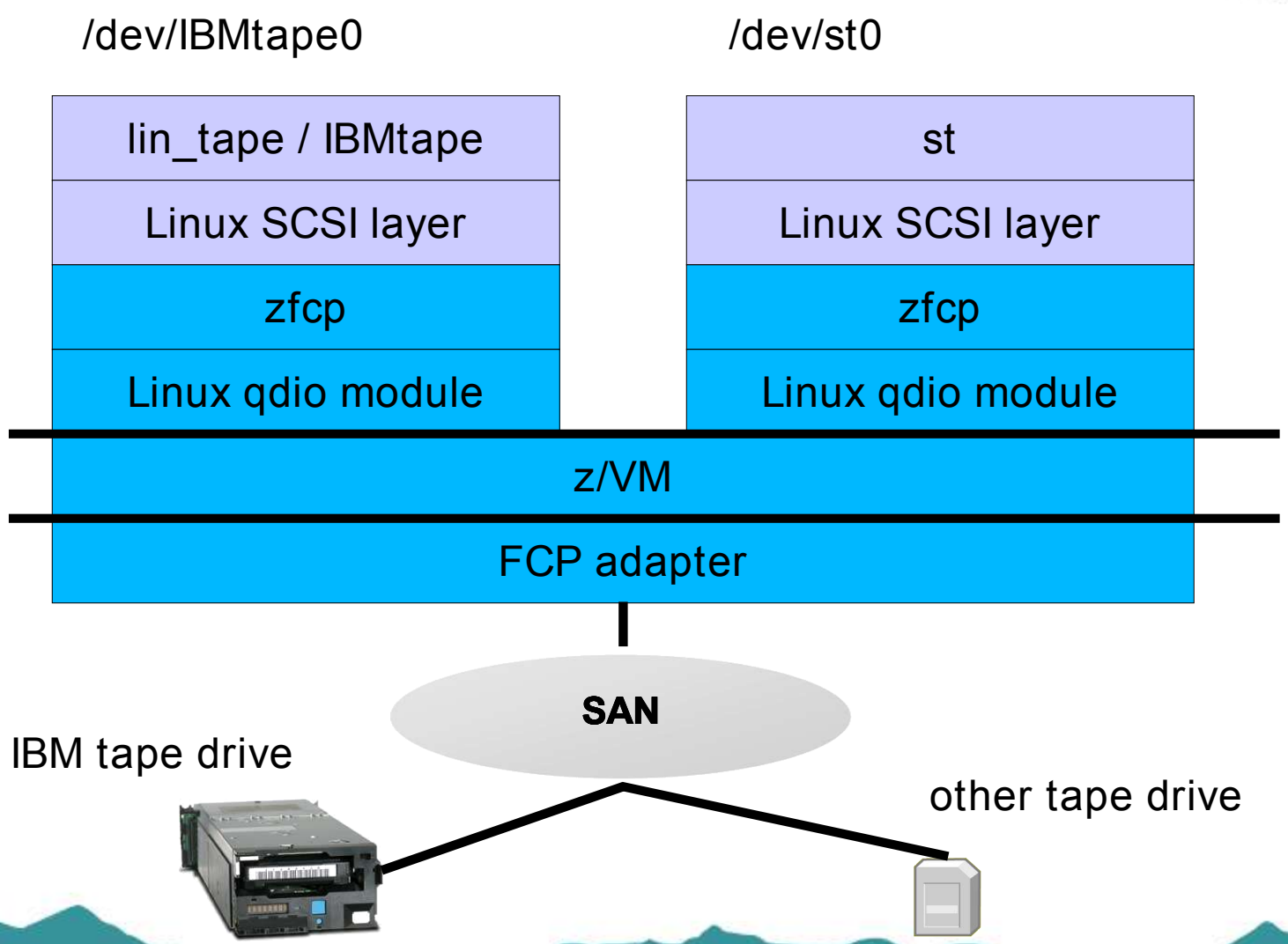
root filesystem on multipath device

- final setup after reboot

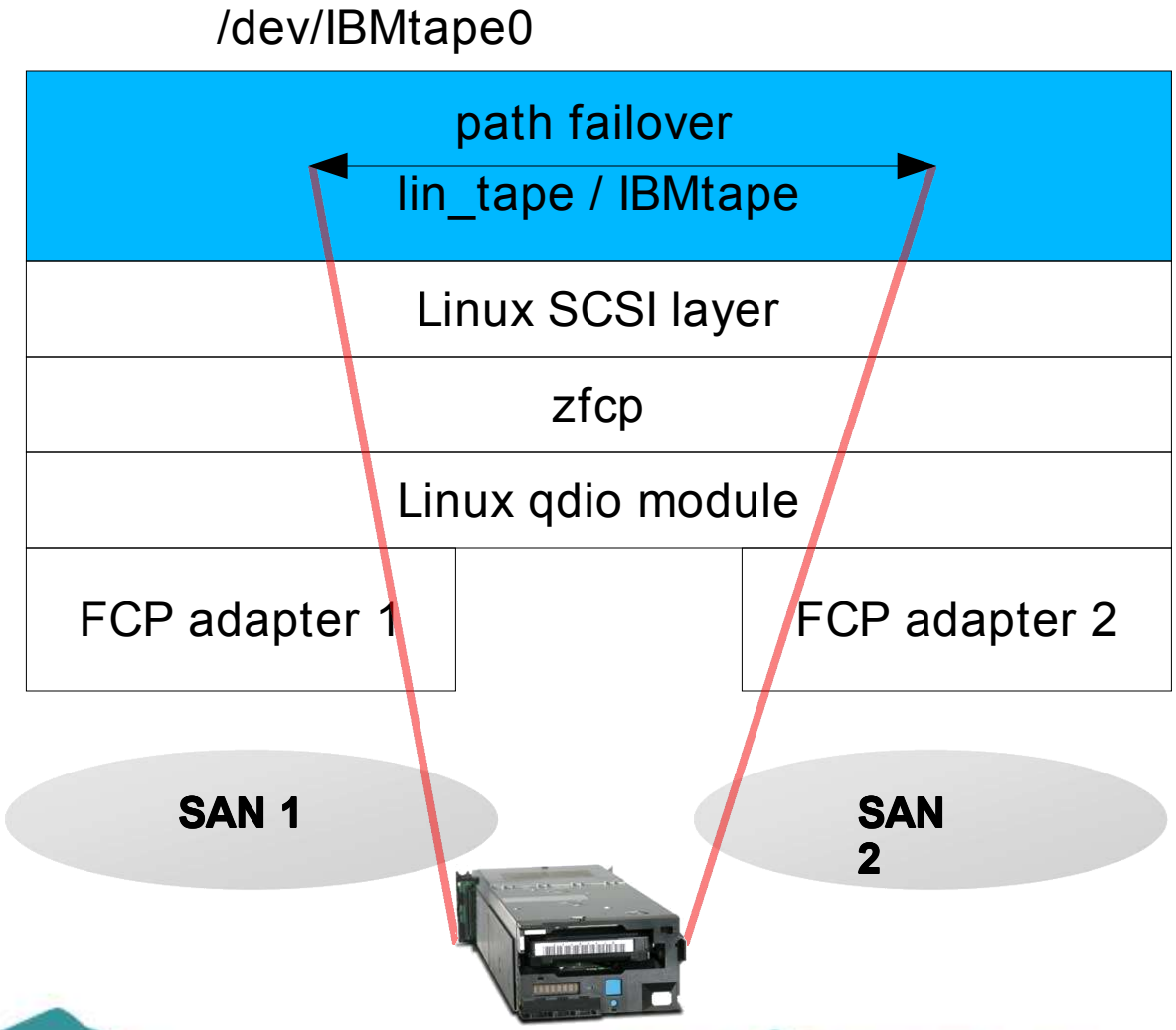
```
# multipath -ll
36005076303ffc56200000000000010cc dm-0 IBM,2107900
[size=5.0G][features=1 queue_if_no_path][hwhandler=0]
\_ round-robin 0 [prio=2][active]
 \_ 1:0:0:1087127568 sdc 8:32  [active][ready]
 \_ 0:0:0:1087127568 sda 8:0   [active][ready]

# mount
/dev/mapper/36005076303ffc56200000000000010cc-part1 on /
type ext3 (rw,acl,user_xattr)
/dev/sdb1 on /boot type ext3 (rw,acl,user_xattr)
```

Tape drives and FCP



Multipathing for IBM tape drives



Multipathing for IBM tape drives

- multipath-tools only cover disk storage
- lin_tape device driver provides multipathing for IBM tape drives
  - (IBMtape is the previous name for the same device driver)
- supported together with tape hardware
- does not cover data mirroring or drive failover, handled by
  - backup application
  - media management application
- lin_tape setup in /etc/modprobe.conf.local
    - options lin_tape alternate_pathing=1

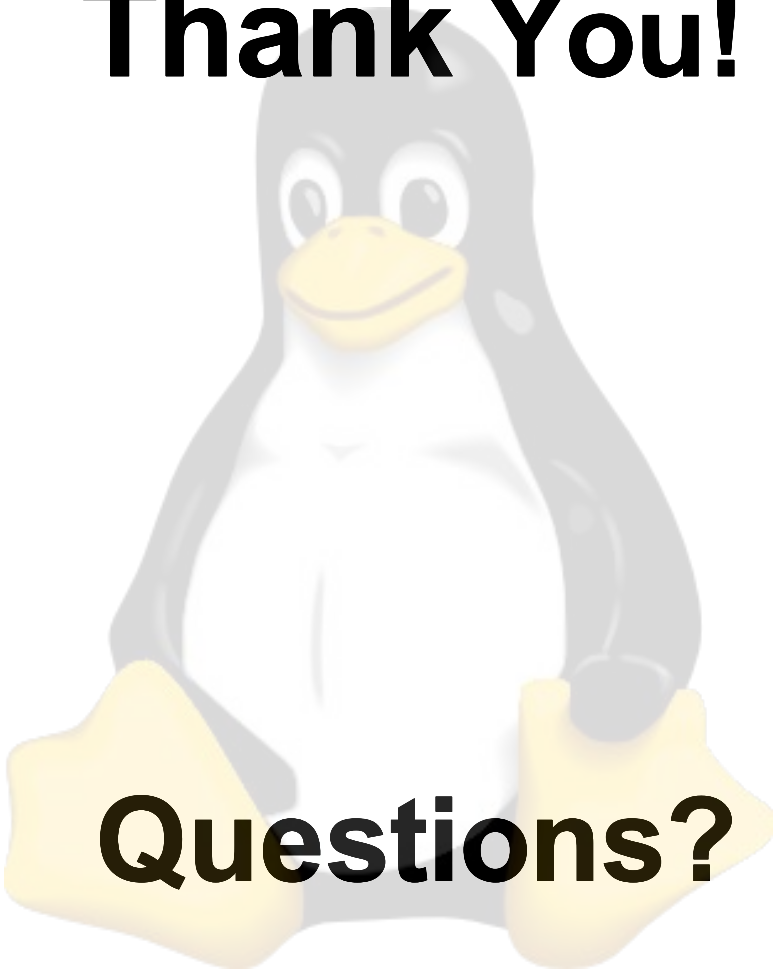
Summary

- multipathing is required for reliability
- multipath-tools are the standard solution for disks
- go with default settings, only do minimal changes
- SCSI device files are paths in multipathing
- basic setup is simple
- root filesystem on multipath device requires more effort
- multipathing for IBM tape drives available

Resources

- Device-mapper Resource Page (link to Multipath bug tracking)
  http://sources.redhat.com/dm/
- Device-mapper and LVM2 Wiki
  http://sources.redhat.com/lvm2/wiki/MultipathUsageGuide
- multipath tools FAQ
  http://git.kernel.org/?p=linux/storage/multipath-tools/.git;a=blob;f=FAQ
- How to setup / use multipathing on SLES
  http://support.novell.com/techcenter/sdb/en/2005/04/sles_multipathing.html
- Enabling root-on-multipath for SLES9 on zSeries
  http://linuxvm.org/Info/HOWTOs/root-on-multipath.html
- Redhat: Using Device-Mapper Multipath
  http://www.redhat.com/docs/manuals/enterprise/RHEL-5-manual/en-US/RHEL510/DM_Multipath/
- IBMtape/lin_tape driver and documentation
  ftp://ftp.software.ibm.com/storage/devdrvr/
- multipath-tools
  http://christophe.varoqui.free.fr/

Thank You!



Questions?

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States, other countries, or both.

Not all common law marks used by IBM are listed on this page. Failure of a mark to appear does not mean that IBM does not use the mark nor does it mean that the product is not actively marketed or is not significant within its relevant market. Those trademarks followed by ® are registered trademarks of IBM in the United States; all others are trademarks or common law marks of IBM in the United States.

For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml:

*, AS/400®, e business(logo)®, DBE, ESCO, eServer, FICON, IBM®, IBM (logo)®, iSeries®, MVS, OS/390®, pSeries®, RS/6000®, S/30, VM/ESA®, VSE/ESA, WebSphere®, xSeries®, z/OS®, zSeries®, z/VM®, System i, System i5, System p, System p5, System x, System z, System z9®, BladeCenter®

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
Java and all Java-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:
Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.
IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.
All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.
This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.
All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.
Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.
Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.