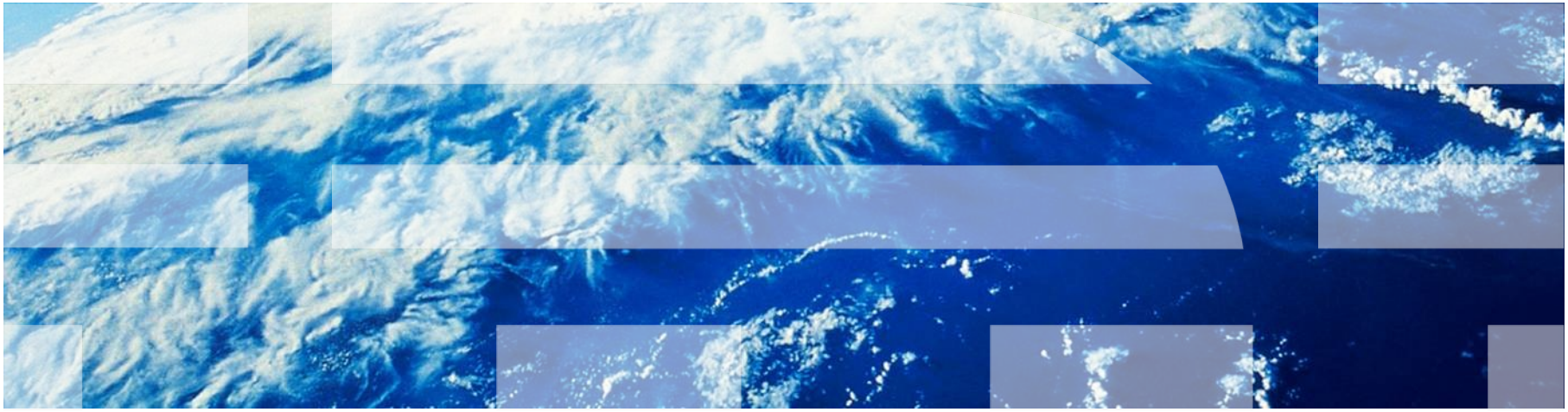
2010-10-20Dr. Holger Smolinski <Holger.Smolinski@de.ibm.com>
IBM Deutschland Research & Development GmbH, LTC Böblingen
4th European IBM/GSE Technical University, Munich, October 2010

Implementing SCSI over FCP with Linux on System z



Trademarks

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

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

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.

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.

Abstract

SCSI over FCP is an open, standard-based alternative and supplement to existing ESCON or FICON connections.

This session will provide an introduction to the storage attachment via the SCSI over FCP protocol.

It includes setup considerations, related features in Linux on System z and troubleshooting basics.

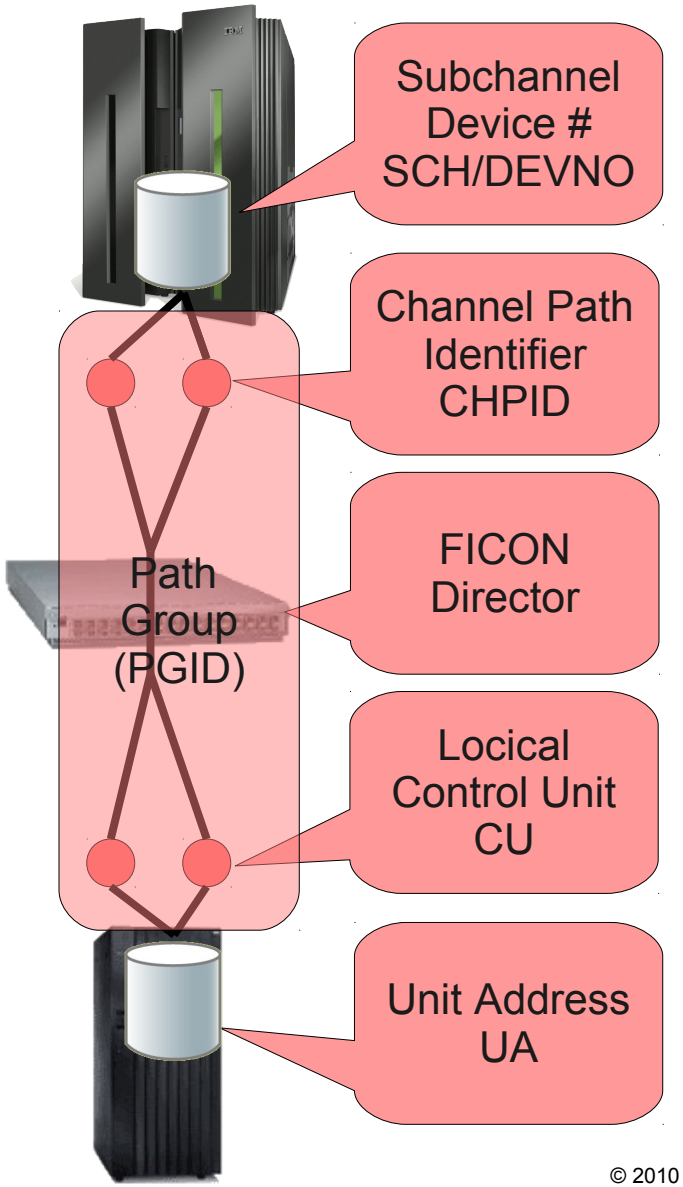
Agenda

- Introduction to SCSI/FCP storage attachment
  - Terminology: FCP vs. FICON channels
  - FCP Channel, Initiator, target & virtual Ports, LUNs
- Setup considerations
  - SAN, FCP devices, NPIV, Storage configuration
- Related features in Linux on System z
  - s390-tools
  - Multipathing
  - SCSI IPL
- Troubleshooting basics
  - Default kernel messages
  - scsi_logging_level
  - zfcp_show, zfcp_ping

Source If Applicable

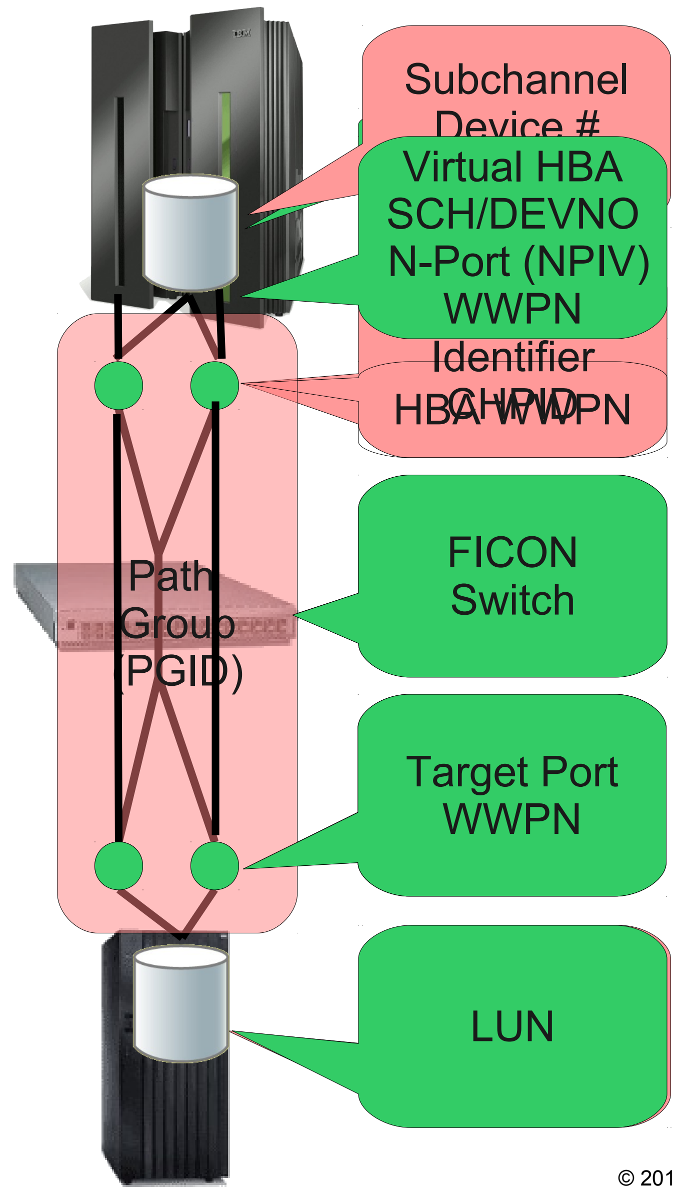
Fibre Channel Storage Area Network (SAN) - FICON terminology

- Each unit (UA) is represented on the host side as a subchannel (SCH) and managed by a device number (DEVNO)
- Each unit is associated to a control unit (CU), which can be reached via a path group (PGID)
- Each path group consists of multiple paths (CHPIDs)
- System z connected via Fibre Channel Connections to FICON Director
- The association is managed by System z IOCDS/IODF



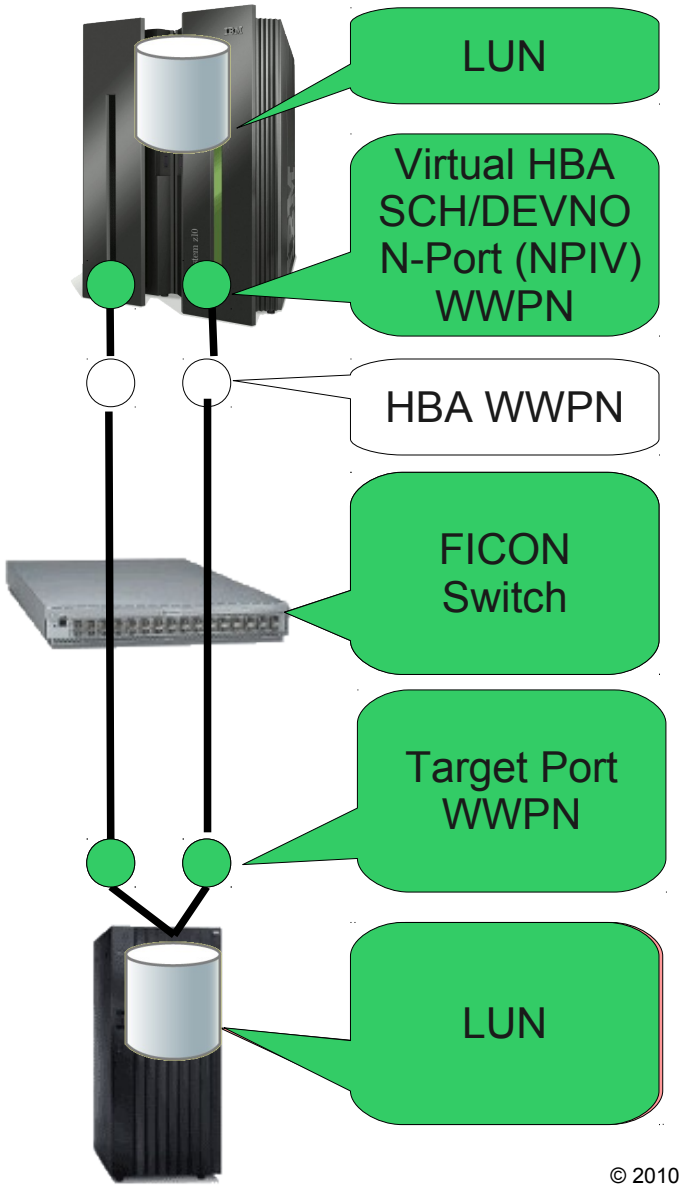
Fibre Channel Storage Area Network (SAN) - FICON vs. FCP terminology

- Each Unit (UA) is represented on the host side as a subchannel (SCH) and managed by a device number (DEVNO)
  Units (LUNs) are managed by the OS. A virtual HBA is presented on the host side as a subchannel (SCH) and DEVNO
- Each unit is associated to a control unit (CU) which can be reached via a path group (PGID)
  Each FICON Adapter (CHPID) hosts multiple HBAs with own N-Ports. Addressing is done through WWPNs.
- Each path group consists of multiple paths (CHPIDs)
  No Path Grouping is performed by the Card-Fw firmware.
- System z connected via Fibre Channel Connections to FICON Director Switch
- The association is managed by System z IOCDS/IODF for Virtual HBAs but not for LUNs



Fibre Channel Storage Area Network (SAN) - FCP terminology

- Units (LUNs) are managed by the OS.
  A virtual HBA is presented on the host side as a subchannel (SCH) and DEVNO
- Each FICON Adapter (CHPID) hosts multiple HBAs with own N-Port Ids.
  Addressing is done through WWPNs.
- No Path Grouping is performed by the Hard- or firmware.
- System z connected via Fibre Channel Connections to FICON Switch
- The association is managed by System z IOCDS/IODF for Virtual HBAs but not for LUNs

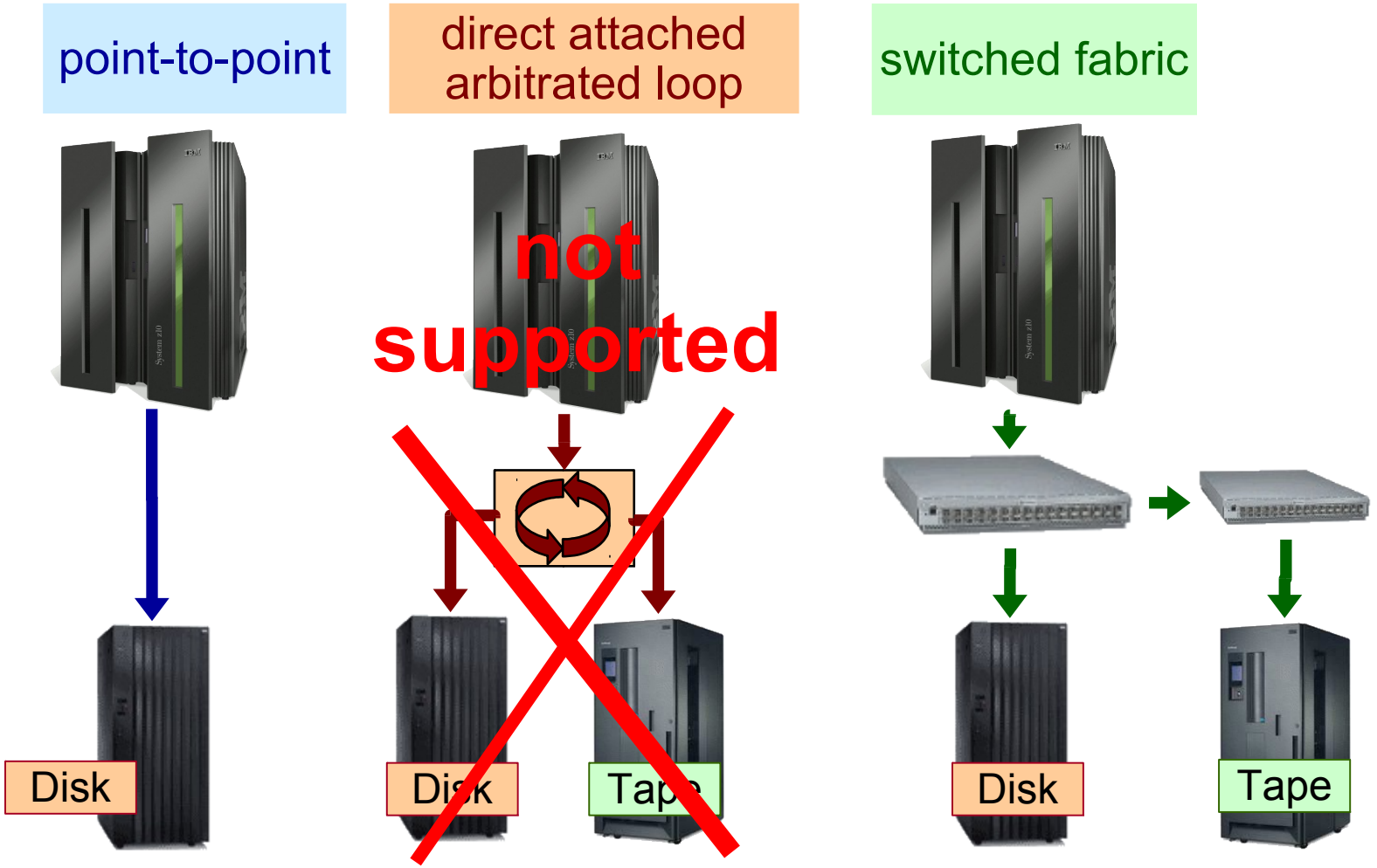


IODF for FCP devices

```
CHPID PATH=(CSS(0,1,2,3),51),SHARED,                    *
     NOTPART=((CSS(1),(TRX1),(=)),(CSS(3),(TRX2,T29CFA),(=)))*
     ,PCHID=1C3,TYPE=FCP
CNTLUNIT CUNUMBR=3D00,                                  *
     PATH=((CSS(0),51),(CSS(1),51),(CSS(2),51),(CSS(3),51)), *
     UNIT=FCP
IODEVICE ADDRESS=(3D00,001),CUNUMBR=(3D00),UNIT=FCP
IODEVICE ADDRESS=(3D01,007),CUNUMBR=(3D00),             *
     PARTITION=((CSS(0),T29LP11,T29LP12,T29LP13,T29LP14,T29LP*
     15),(CSS(1),T29LP26,T29LP27,T29LP29,T29LP30),(CSS(2),T29*
     LP41,T29LP42,T29LP43,T29LP44,T29LP45),(CSS(3),T29LP56,T2*
     9LP57,T29LP58,T29LP59,T29LP60)),UNIT=FCP
IODEVICE ADDRESS=(3D08,056),CUNUMBR=(3D00),             *
     PARTITION=((CSS(0),T29LP15),(CSS(1),T29LP30),(CSS(2),T29*
     LP45),(CSS(3),T29LP60)),UNIT=FCP
```

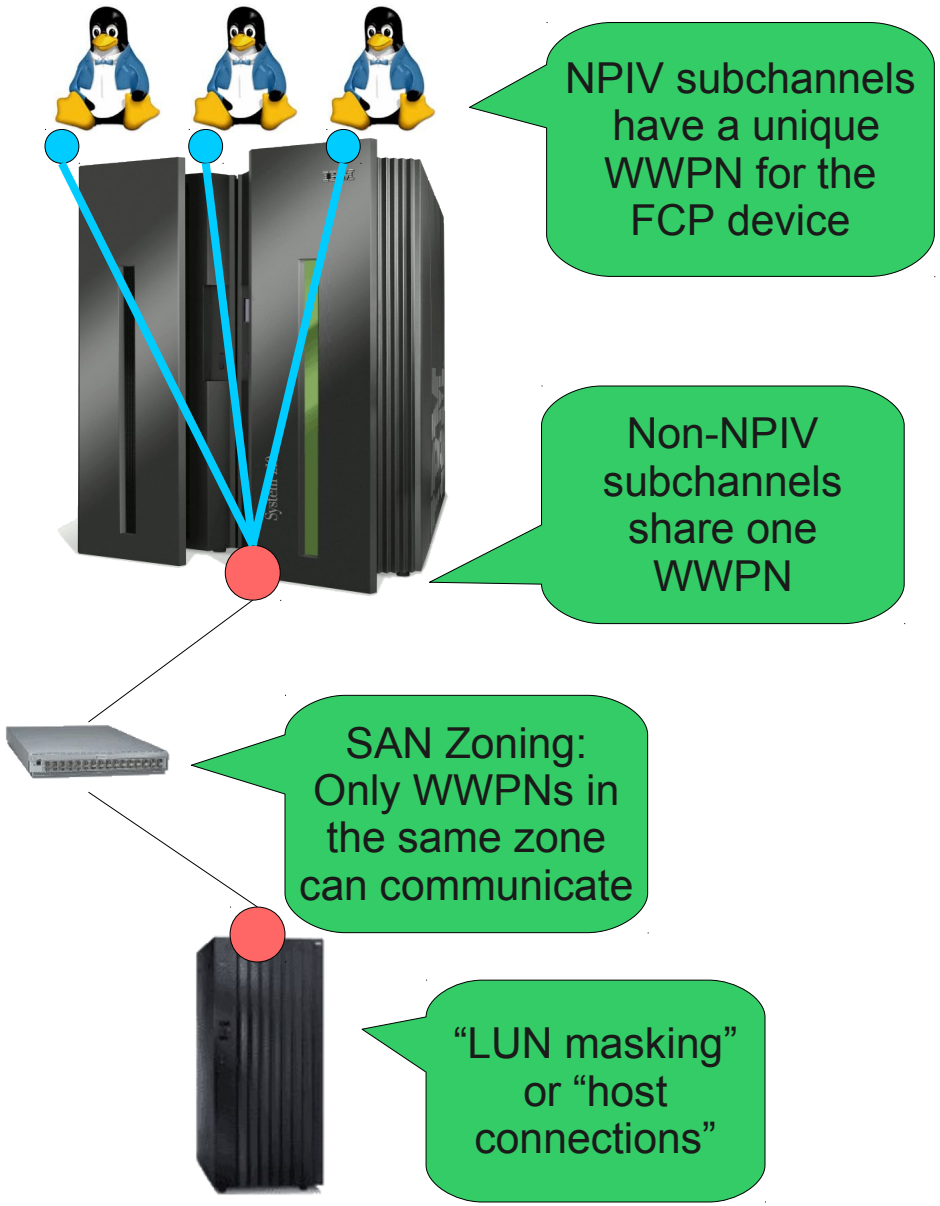
- Defines only FCP adapter
- WWPN and LUN configuration inside operating systems (e.g. Linux on System z)

SAN topologies and System z



- A loop with two nodes looks like a point-to-point connection, but uses the unsupported loop protocol. Check device specifications for details.

SAN Access Control based on N-Port-Id-Virtualization (NPIV)



- NPIV: Each virtual adapter has its own WWPN in the SAN. This is the foundation for restricting access to storage.
- SAN zoning: Only WWPNs in same zone can communicate
- “LUN masking” or “host connections”: Each disk volume is only available for a specific list of WWPNs
- With NPIV: Restricted configuration possible where disk volumes are defined for each operating system
- Without NPIV: First operating system to access a disk volume can use it.

NPIV

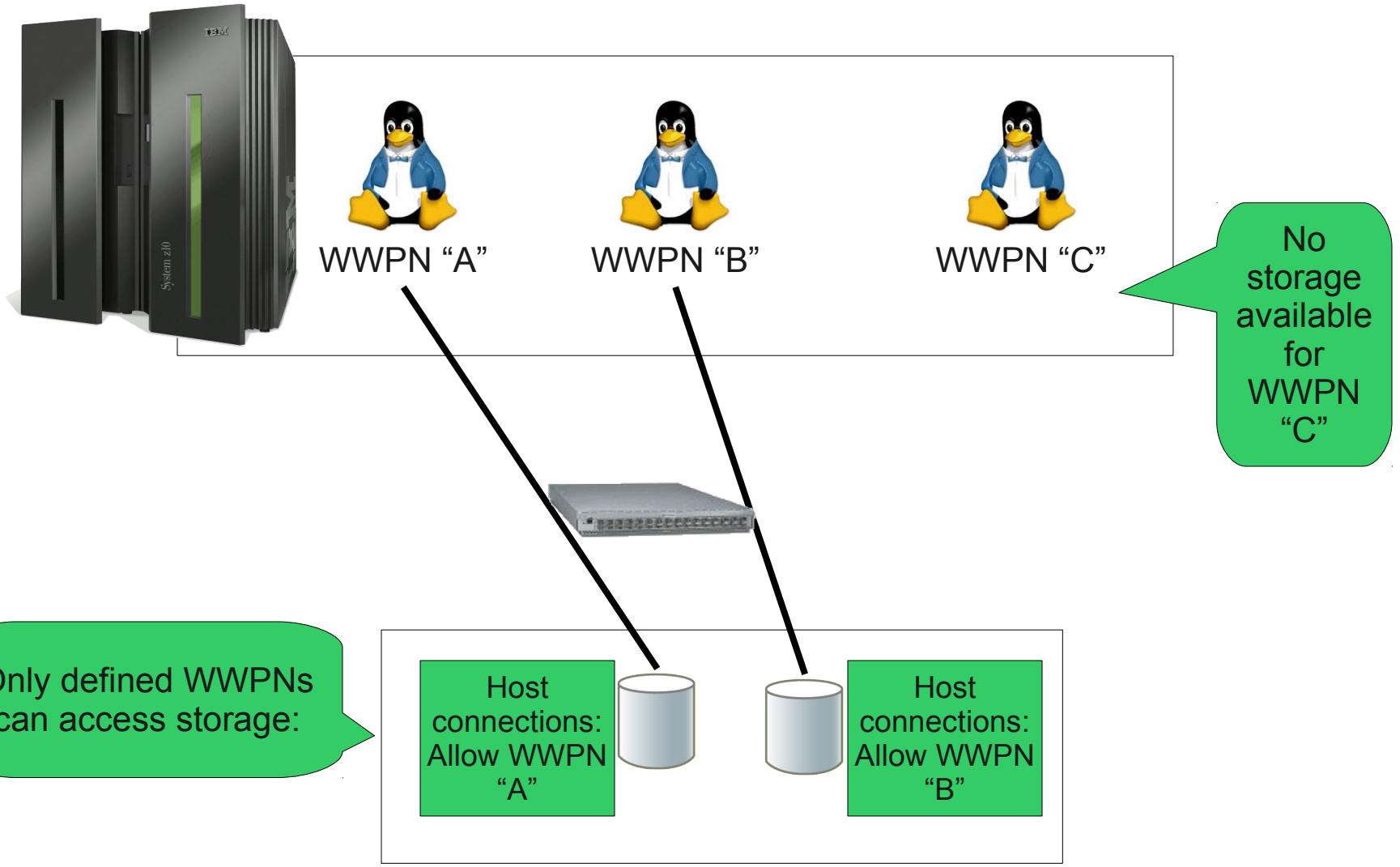
```
# lszfcp -a | grep port_name
    permanent_port_name = "0xc05076ffe5005611"
    port_name           = "0xc05076ffe5005350"
```

- “port_name” is the WWPN used by the FCP subchannel
- “permanent_port_name” is the WWPN assigned to the FCP channel
- Compare to find out if NPIV is in use:
  - If both are the same, the FCP subchannel does NOT use NPIV
  - If they differ, the FCP subchannel uses NPIV
- Worldwide portname prediction tool
  - http://publib.boulder.ibm.com/infocenter/zvm/v6r1/index.jsp?topic=/com.ibm...cpa0/

NPIV requirements

- NPIV is available on System z9, System z10 and z/Enterprise
- z/VM
  - z/VM 5.2 or newer
  - z/VM 5.1 with the PTF for APAR VM63744
- Linux Distribution (LPAR mode or z/VM)
  - SLES9 SP3+, SLES10, SLES11, RHEL5, ...
- NPIV-Capable Switch
  - only required for switch adjacent to System z
  - Mostly firmware upgrades possible (e.g. McData, Brocade)

“Host connections” / “LUN masking” with NPIV



zfcp, differences to other Linux platforms

- Linux common code:
  - Query available LUNs from storage server (“REPORT LUNS”)
  - Attach all LUNs that are reported by the storage server
- zfcp differs from other Linux platforms
  - Only attach LUNs configured in Linux
  - Required for FCP channel sharing in non-NPIV mode (e.g. LUN 1 used by one system, LUN 2 by another)
- Future (planned for Linux kernel 2.6.37):
  - For FCP subchannels in NPIV mode
    - allow automatic attachment of LUNs through common code
    - same behaviour as other Linux platforms

Manual LUN configuration with s390-tools

Show available FCP devices:

```
# lscss -t 1732/03,1732/04
Device   Subchan.  DevType CU Type Use  PIM PAM POM  CHPIDs
----------------------------------------------------------------------
0.0.181d 0.0.0010  1732/03 1731/03      80  80  ff   3b000000 00000000
0.0.191d 0.0.0011  1732/03 1731/03      80  80  ff   46000000 00000000
```

Enable FCP devices:

```
# chccwdev -e 0.0.181d,0.0.191d
Setting device 0.0.181d online
Done
Setting device 0.0.191d online
Done
```

Show online FCP devices and corresponding SCSI hosts:

```
# lszfcp
0.0.181d host0
0.0.191d host1
```

Manual zfcp LUN configuration

Attach ports, only required for older distributions (SLES10, RHEL5, ...):

```
# echo 0x500507630313c562 > /sys/bus/ccw/drivers/zfcp/0.0.181d/port_add
# echo 0x500507630300c562 > /sys/bus/ccw/drivers/zfcp/0.0.191d/port_add
```

Show available storage ports:

```
# lszfcp -P
0.0.181d/0x500507630313c562 rport-0:0-0
0.0.191d/0x500507630300c562 rport-1:0-0
```

Attach LUNs:

```
# echo 0x401040c300000000 >
/sys/bus/ccw/drivers/zfcp/0.0.181d/0x500507630313c562/unit_add
# echo 0x401040c300000000 >
/sys/bus/ccw/drivers/zfcp/0.0.191d/0x500507630300c562/unit_add
```

List available LUNs:

```
# lszfcp -D
0.0.181d/0x500507630313c562/0x401040c300000000 0:0:0:1086537744
0.0.191d/0x500507630300c562/0x401040c300000000 1:0:0:1086537744
```

Linux Tools: lsluns

lsscsi: Show SCSI devices and block device nodes:

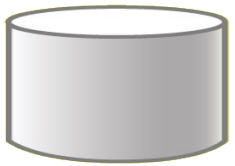
```
# lsscsi
[0:0:0:1086537744]disk    IBM      2107900          .280  /dev/sda
[1:0:0:1086537744]disk    IBM      2107900          .280  /dev/sdb
```

lsluns: Which LUNs are available on the storage system for use?

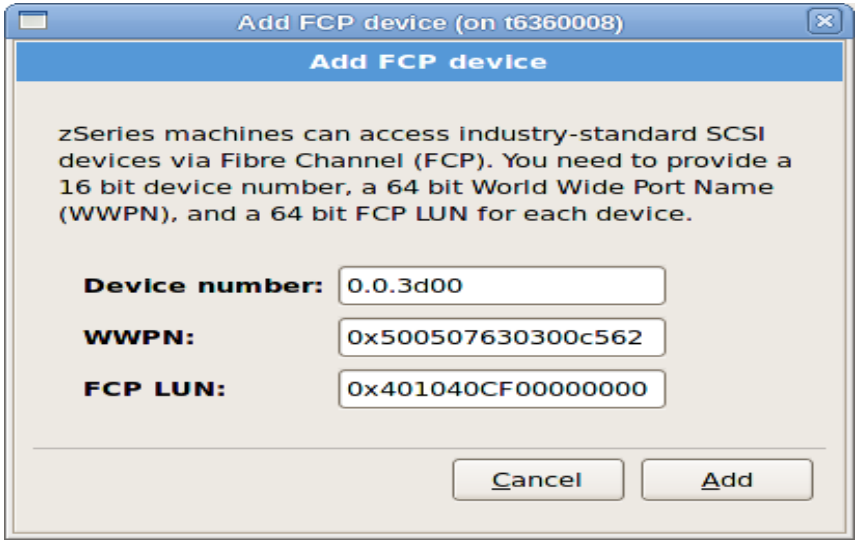
```
# lsluns -c 0.0.181d -p 0x50050763031b0104
Scanning for LUNs on adapter 0.0.181d
    at port 0x50050763031b0104:
        0x4011400000000000
        0x4011400100000000
        0x4011400200000000
        0x4011400300000000
        0x4011400400000000
        0x4011400500000000
        0x4011400600000000
        0x4011400700000000
        0x4011400800000000
        0x4011400900000000
```

zfcp configuration files

- Configuring LUNs manually through sysfs is not persistent, will be lost after IPL
- Use configuration mechanism provided by distribution for persistent LUN configuration
- See documentation from distributions for details
- zfcp configuration file in RHEL
  - /etc/zfcp.conf
- zfcp configuration files in SLES 10
  - /etc/sysconfig/hardware/hwcfg-zfcp-bus-ccw-*
- zfcp configuration files in SLES 11
  - /etc/udev/rules.d/51-zfcp*
-



zfcp LUN configuration in RHEL 5.5

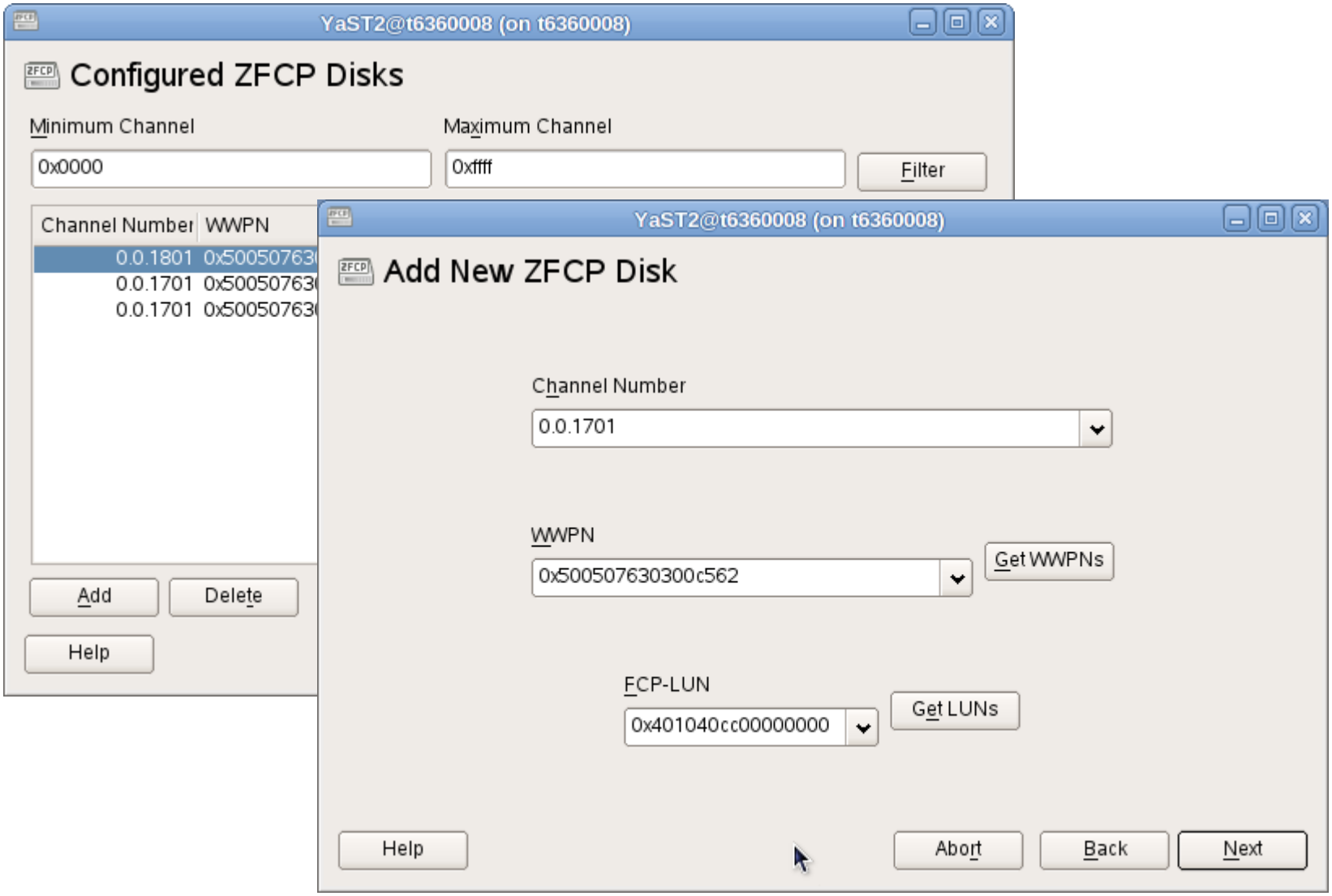


- Dialog to add zfcp LUN configuration during installation (e.g. root filesystem disks)
- Alternatively: Add zfcp LUN configuration to /etc/zfcp.conf config file (e.g. non-root filesystem disks:

```
# cat /etc/zfcp.conf
0.0.3c00    0x500507630313c562 0x401040c300000000
0.0.3d00    0x500507630300c562 0x401040C300000000
```

-

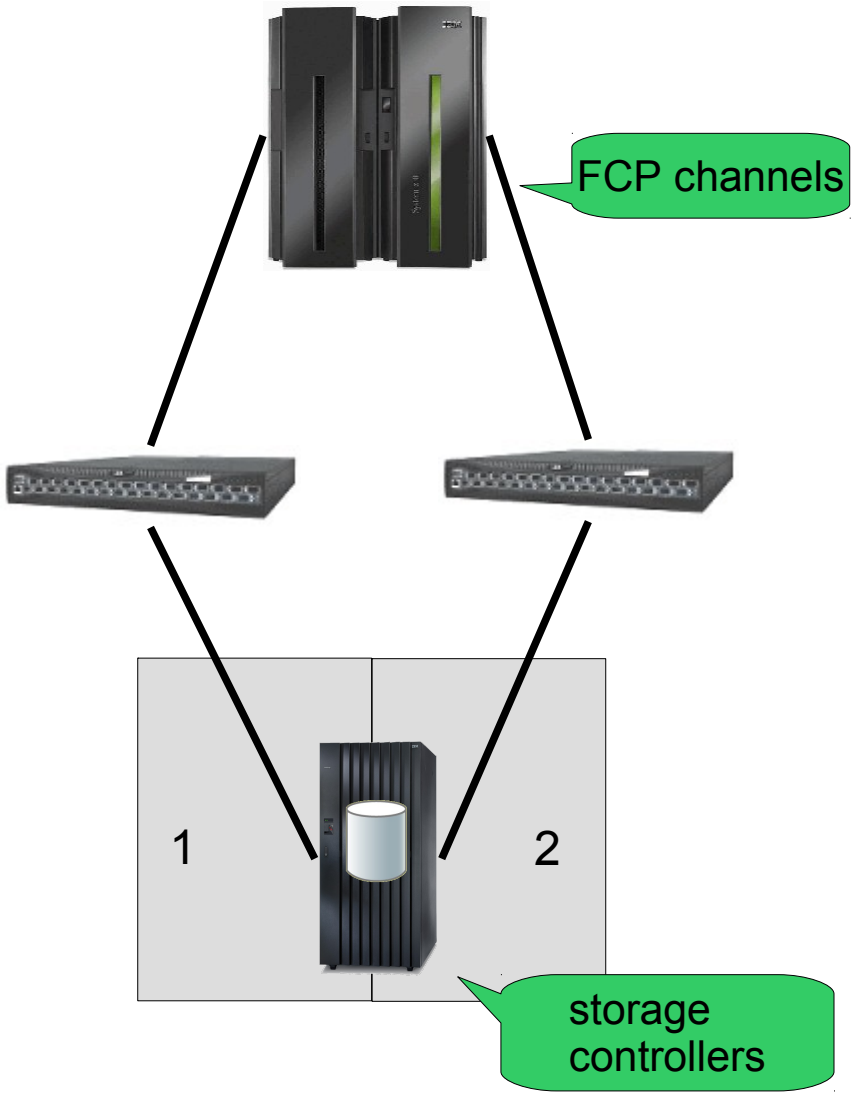
YaST zfcp LUN configuration in SLES11 SP1



Multipathing is mandatory

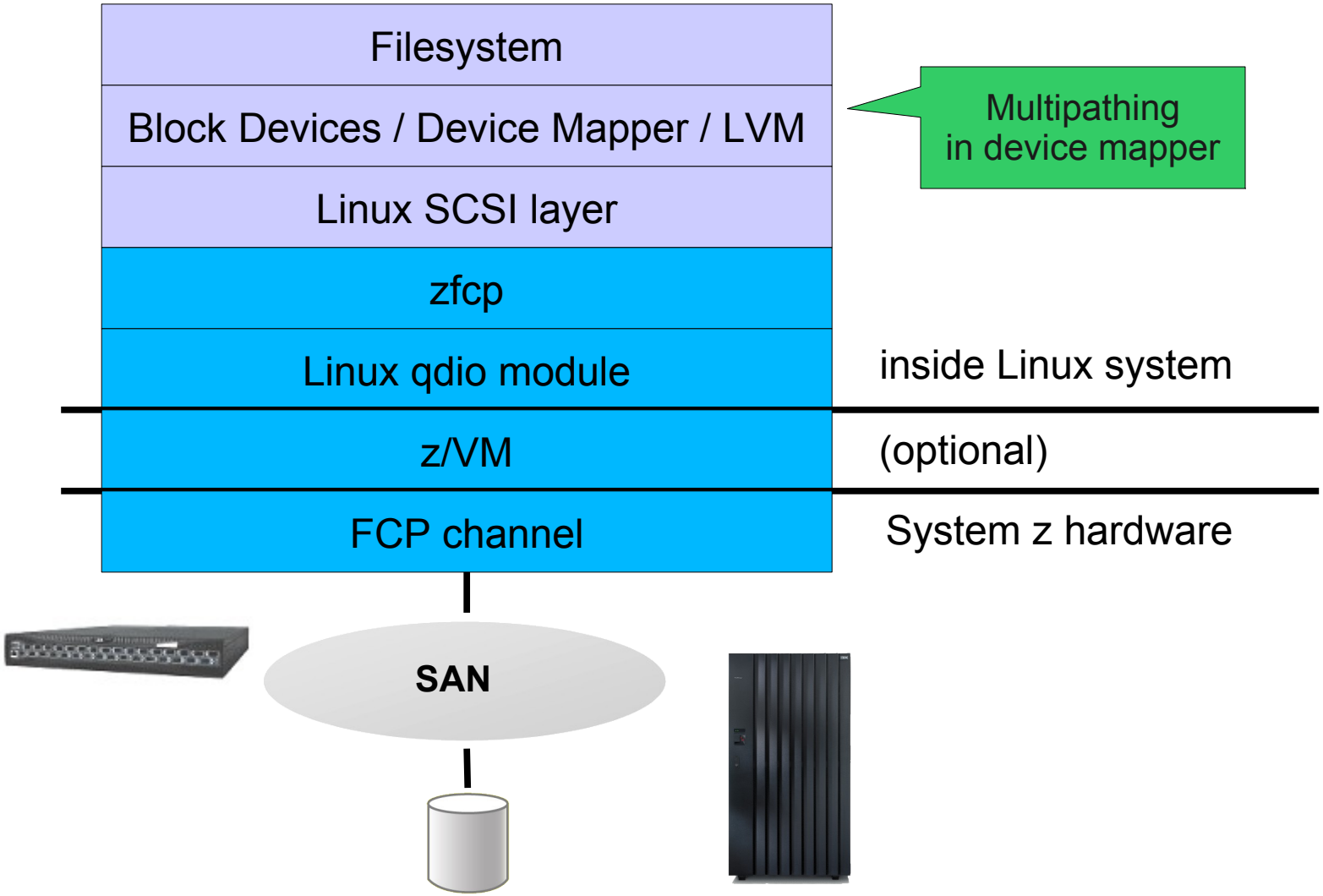
- Use multiple paths from operating system to storage
- Why is multipathing mandatory?
  - Remove single points of failure – single components WILL fail according to standards
  - Performance: I/O requests can be spread across multiple paths,
  - Usually a requirement for SAN connected storage servers
    - e.g. when one storage controller is in maintenance mode I/O continues to run through second controller
- Multipathing in Linux
  - Implemented in Linux in multipath-tools package, together with the device-mapper in the Linux kernel
  - The default configuration is already included, configuration changes only for special requirements
  - SCSI device (“LUN”) in Linux represents one path to the disk volume on the storage server
- Multipath devices are block devices in Linux, must be used e.g.
  - Directly for a filesystem
  - Logical Volume Manager (LVM)
  - Partitioninig is possible

Multipathing overview



- multiple paths to storage volumes
- maintain access to storage volume
  - during storage system update
  - SAN fabric maintenance
  - FCP channel maintenance
- for FCP/SCSI managed by the operating system (Linux)
- WWID for storage volume, e.g. 36005076303ffc5620000000000000010cc

Linux on System z: SCSI stack and multipathing



Multipathing setup

- Config file /etc/multipath.conf
- Ensure devices are not blacklisted and multipathd is running
- Multipath tools include defaults for standard storage systems
- Multipath devices are created automatically when SCSI LUNs are attached

```
# multipath -ll
36005076303ffc56200000000000010cc dm-0 IBM,2107900
size=5.0G features='1 queue_if_no_path' hwhandler='0' wp=rw
`-+- policy='round-robin 0' prio=1 status=active
  |- 0:0:26:1087127568 sda   8:0  active ready running
  `- 1:0:11:1087127568 sdb   8:16 active ready running
```

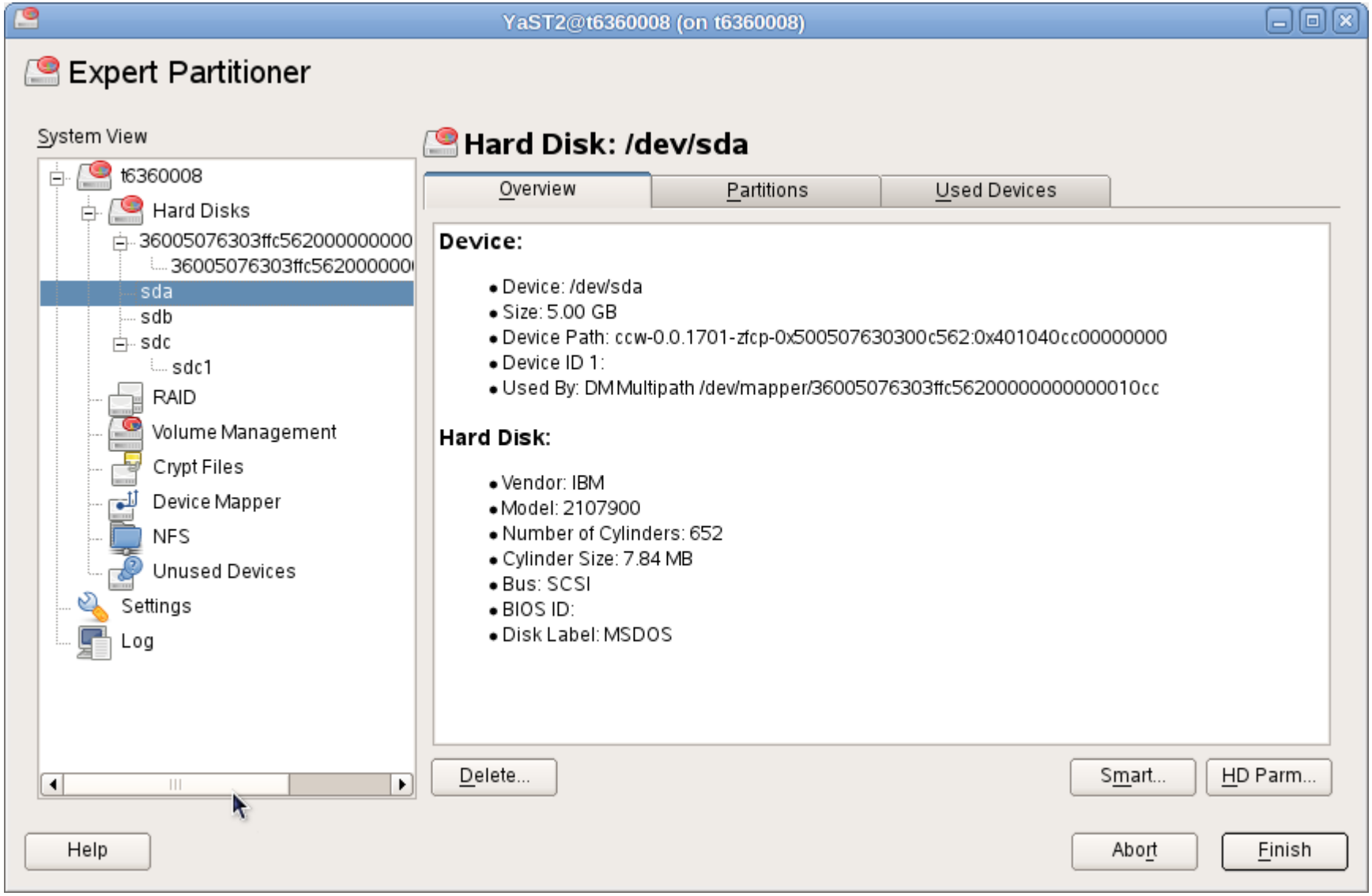
Each, sda and sda represent one path

Resulting multipath block device

```
# mkfs.ext3 /dev/mapper/36005076303ffc56200000000000010cc
…

# mount /dev/mapper/36005076303ffc56200000000000010cc /mnt/
```

SLES11 multipathing setup through YaST



Multipath configuration

```
# cat /etc/multipath.conf
multipaths {
    multipath {
        wwid "36005076303ffc56200000000000010cc"
        rr_min_io  "10"
        path_selector "service-time 0"
    }
}

# multipath -l
36005076303ffc56200000000000010cc dm-0 IBM,2107900
size=5.0G features='1 queue_if_no_path' hwhandler='0' wp=rw
`-+- policy='service-time 0' prio=-1 status=active
  |- 0:0:26:1087127568 sda   8:0  active undef running
  `- 1:0:11:1087127568 sdb   8:16 active undef running
```

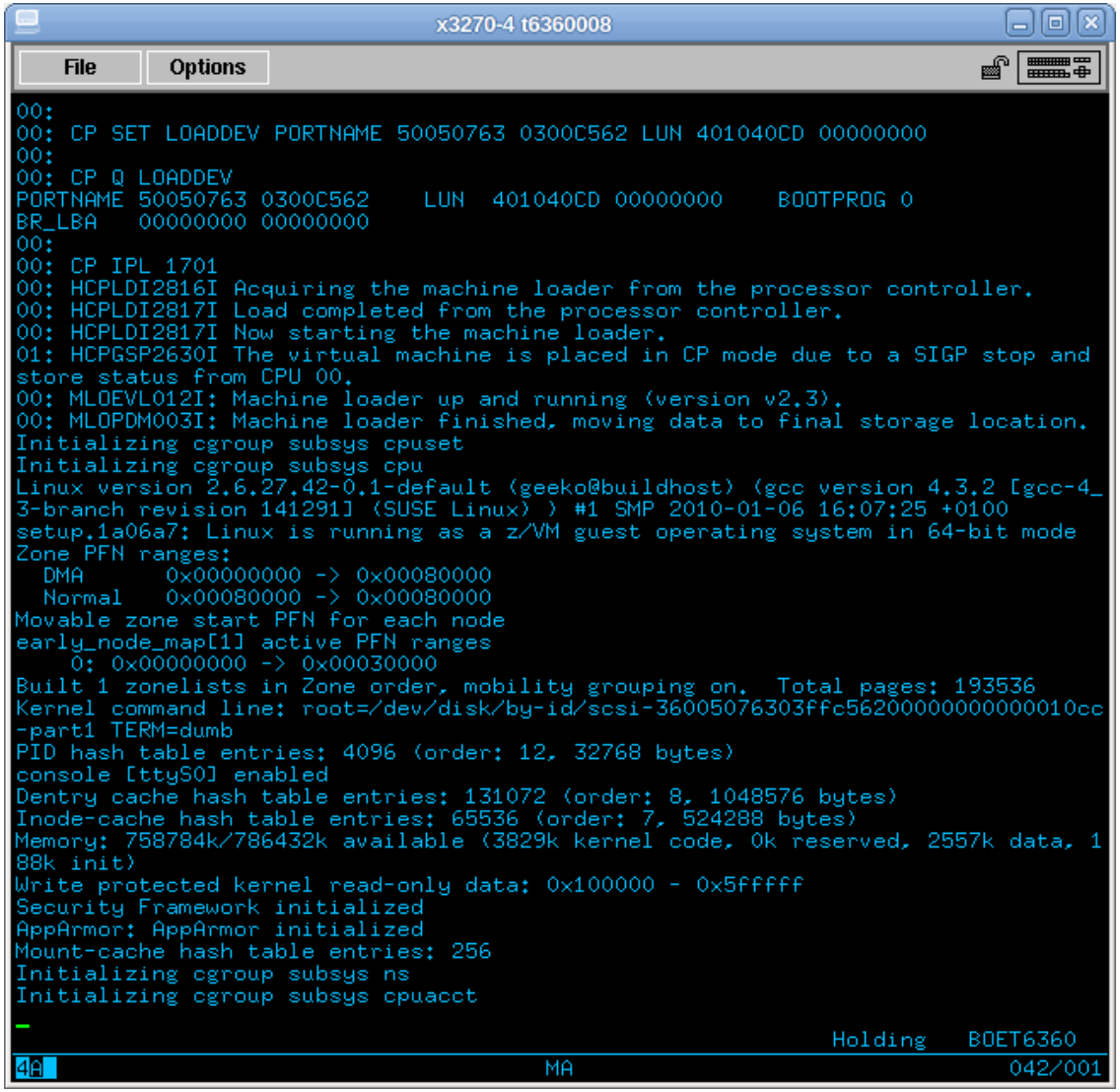
- Default settings are usually good
- Can be overwritten in /etc/multipath.conf
- Example: Change path selector policy from default (“round robin”) to “service time”
- See multipath.conf manpage and documentation from distributions for details

Root filesystem on SCSI multipath volume

- Possibility to put root filesystem on SCSI multipath volume
- Older zipl versions do not support writing on multipath device
- Workaround: Use additional single path volume for /boot, use this volume for IPL
- Newer Linux distributions (e.g. SLES 11 SP1) support root filesystem on multipath device without workaround with s390-tools 1.8.3 or newer
- http://www.ibm.com/developerworks/linux/linux390/s390-tools-1.8.3.html
  - zipl: Add support for device mapper devices.
  - zipl now allows installation of and booting from a boot record on logical devices, i.e. devices managed by device mapper (or similar packages), e.g. multipath devices.
-

SCSI IPL with z/VM

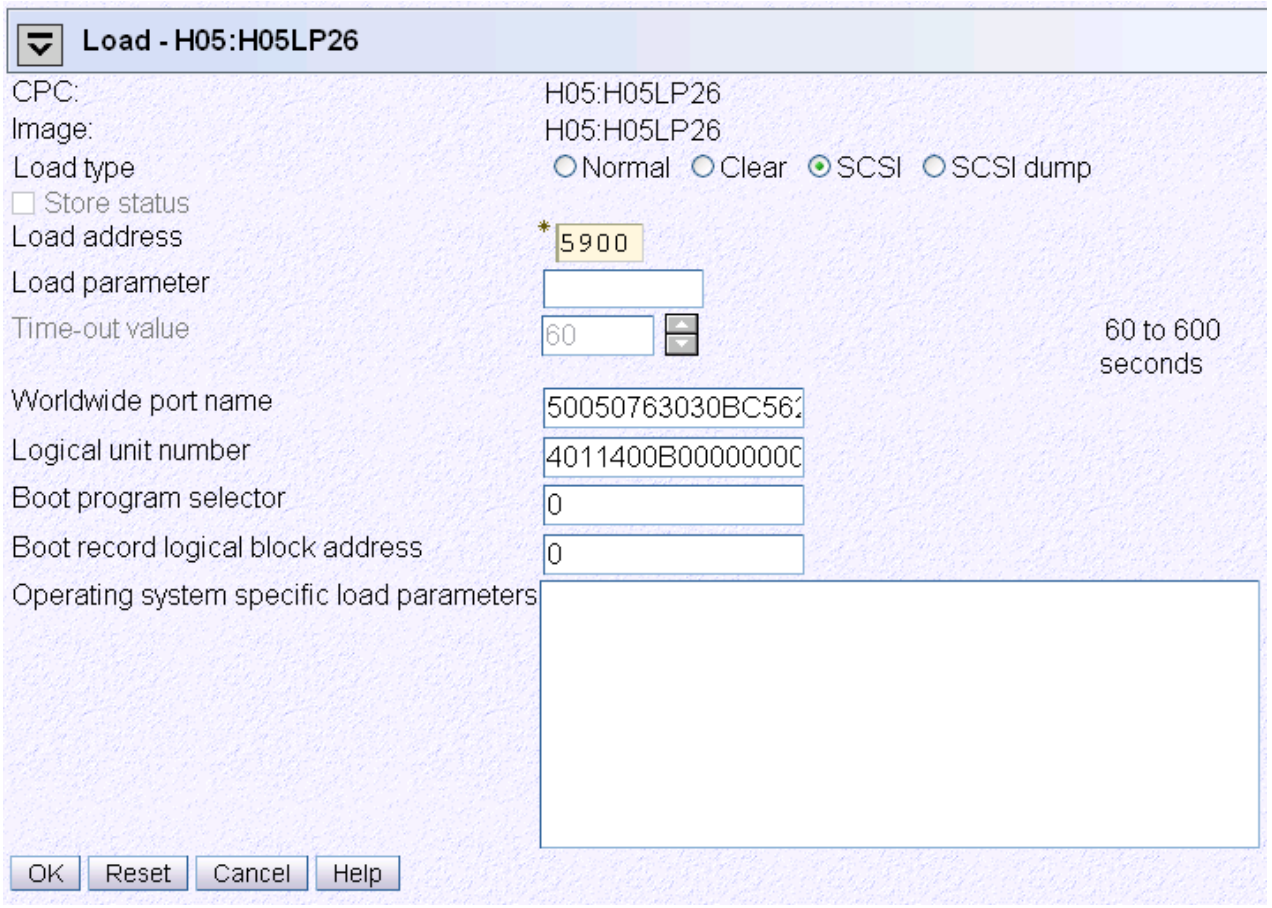
- setup load device
  - WWPN
  - LUN
- IPL from FCP device
- When using dedicated /boot, this LUN becomes IPL device



SCSI IPL with z/VM

```
00:
00: CP SET LOADDEV PORTNAME 50050763 0300C562 LUN 401040CD 00000000
00:
00: CP Q LOADDEV
PORTNAME 50050763 0300C562    LUN  401040CD 00000000    BOOTPROG 0
BR_LBA   00000000 00000000
00:
00: CP IPL 1701
00: HCPLDI2816I Acquiring the machine loader from the processor controller.
00: HCPLDI2817I Load completed from the processor controller.
00: HCPLDI2817I Now starting the machine loader.
01: HCPGSP2630I The virtual machine is placed in CP mode due to a SIGP stop
and
store status from CPU 00.
00: MLOEVL012I: Machine loader up and running (version v2.3).
00: MLOPDM003I: Machine loader finished, moving data to final storage
location.
Initializing cgroup subsys cpuset
Initializing cgroup subsys cpu
Linux version 2.6.27.42-0.1-default (geeko@buildhost) (gcc version 4.3.2 [gcc-
4_
3-branch revision 141291] (SUSE Linux) ) #1 SMP 2010-01-06 16:07:25 +0100
setup.1a06a7: Linux is running as a z/VM guest operating system in 64-bit mode
Zone PFN ranges:
  DMA      0x00000000 -> 0x00080000
  Normal   0x00080000 -> 0x00080000
```

SCSI IPL LPAR from HMC



Troubleshooting

- Check kernel messages that are possibly related to SCSI on Linux on System z:
  - scsi (common SCSI code)
  - sd (SCSI disk)
  - rport (common SCSI code FC remote port messages)
  - qdio (communication between Linux and FCP Channel)
  - zfcp driver kernel messages
    - See “Kernel Messages” book on https://www.ibm.com/developerworks/linux/linux390/
  - “device-mapper: multipath”
- Other syslog messages
  - Multipathd
- zfcp driver traces available in /sys/kernel/debug/s390dbf/
- Collect data with dbginfo.sh when reporting a problem to capture configuration, messages and traces
  -

Troubleshooting: scsi_logging_level

```
# scsi_logging_level -g
Current scsi logging level:
dev.scsi.logging_level = 0
SCSI_LOG_ERROR=0
SCSI_LOG_TIMEOUT=0
SCSI_LOG_SCAN=0
SCSI_LOG_MLQUEUE=0
SCSI_LOG_MLCOMPLETE=0
SCSI_LOG_LLQUEUE=0
SCSI_LOG_LLCOMPLETE=0
SCSI_LOG_HLQUEUE=0
SCSI_LOG_HLCOMPLETE=0
SCSI_LOG_IOCTL=0
```

- More SCSI output in kernel messages
- Higher levels can create lots of messages and slow down system

```
# scsi_logging_level -s -a 1
New scsi logging level:
dev.scsi.logging_level =
153391689
SCSI_LOG_ERROR=1
SCSI_LOG_TIMEOUT=1
SCSI_LOG_SCAN=1
SCSI_LOG_MLQUEUE=1
SCSI_LOG_MLCOMPLETE=1
SCSI_LOG_LLQUEUE=1
SCSI_LOG_LLCOMPLETE=1
SCSI_LOG_HLQUEUE=1
SCSI_LOG_HLCOMPLETE=1
SCSI_LOG_IOCTL=1
```

zfcp_show / zfcp_ping

Query Fibre Channel nameserver about ports available for my system:

```
# zfcp_show -n
Local Port List:
     0x500507630313c562 / 0x656000 [N_Port] proto = SCSI-FCP  FICON
     0x50050764012241e4 / 0x656100 [N_Port] proto = SCSI-FCP
     0x5005076401221b97 / 0x656400 [N_Port] proto = SCSI-FCP
```

Query SAN topology, requires FC management server acccess:

```
# zfcp_show

Interconnect Element Name        0x100000051e4f7c00
Interconnect Element Domain ID   005
Interconnect Element Type        Switch
Interconnect Element Ports       256
    ICE Port 000  Online
        Attached Port [WWPN/ID] 0x50050763030b0562 / 0x650000 [N_Port]
    ICE Port 001  Online
        Attached Port [WWPN/ID] 0x50050764012241e5 / 0x650100 [N_Port]
    ICE Port 002  Online
        Attached Port [WWPN/ID] 0x5005076303008562 / 0x650200 [N_Port]
    ICE Port 003  Offline
...
```

zfcp_ping

Check if remote port responds (requires FC management service access):

```
# zfcp_ping 0x5005076303104562
Sending PNG from BUS_ID=0.0.3c00 speed=8 GBit/s
    echo received from WWPN (0x5005076303104562) tok=0 time=1.905 ms
    echo received from WWPN (0x5005076303104562) tok=1 time=2.447 ms
    echo received from WWPN (0x5005076303104562) tok=2 time=2.394 ms

---------- ping statistics -----------
min/avg/max = 1.905/2.249/2.447 ms
--------------------------------------
```

zfcp_show and zfcp_ping are part of the zfcp-hbaapi 2.1 package:
http://www.ibm.com/developerworks/linux/linux390/zfcp-hbaapi-2.1.html

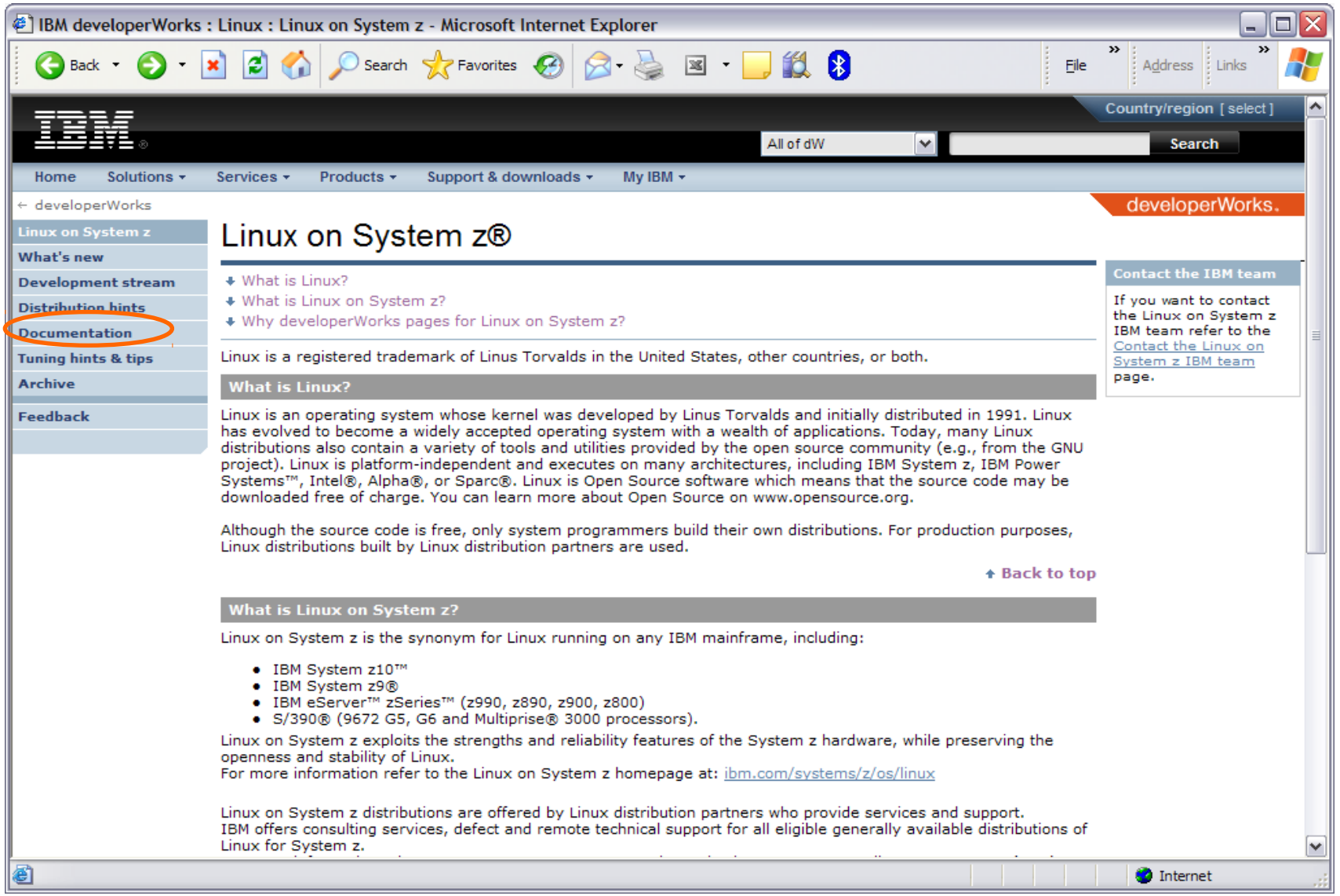
Summary

- Use standard FCP/SCSI storage with Linux on System z
- FCP subchannels defined in System z IODF; ports and LUNs managed in Linux
- Use NPIV to enable storage management and access control
- Use multipathing to avoid single points of failure (check storage requirements)
- Pure SCSI setup possible with root on multipath SCSI device and SCSI IPL
- Troubleshooting through error messages and utilities in s390-tools

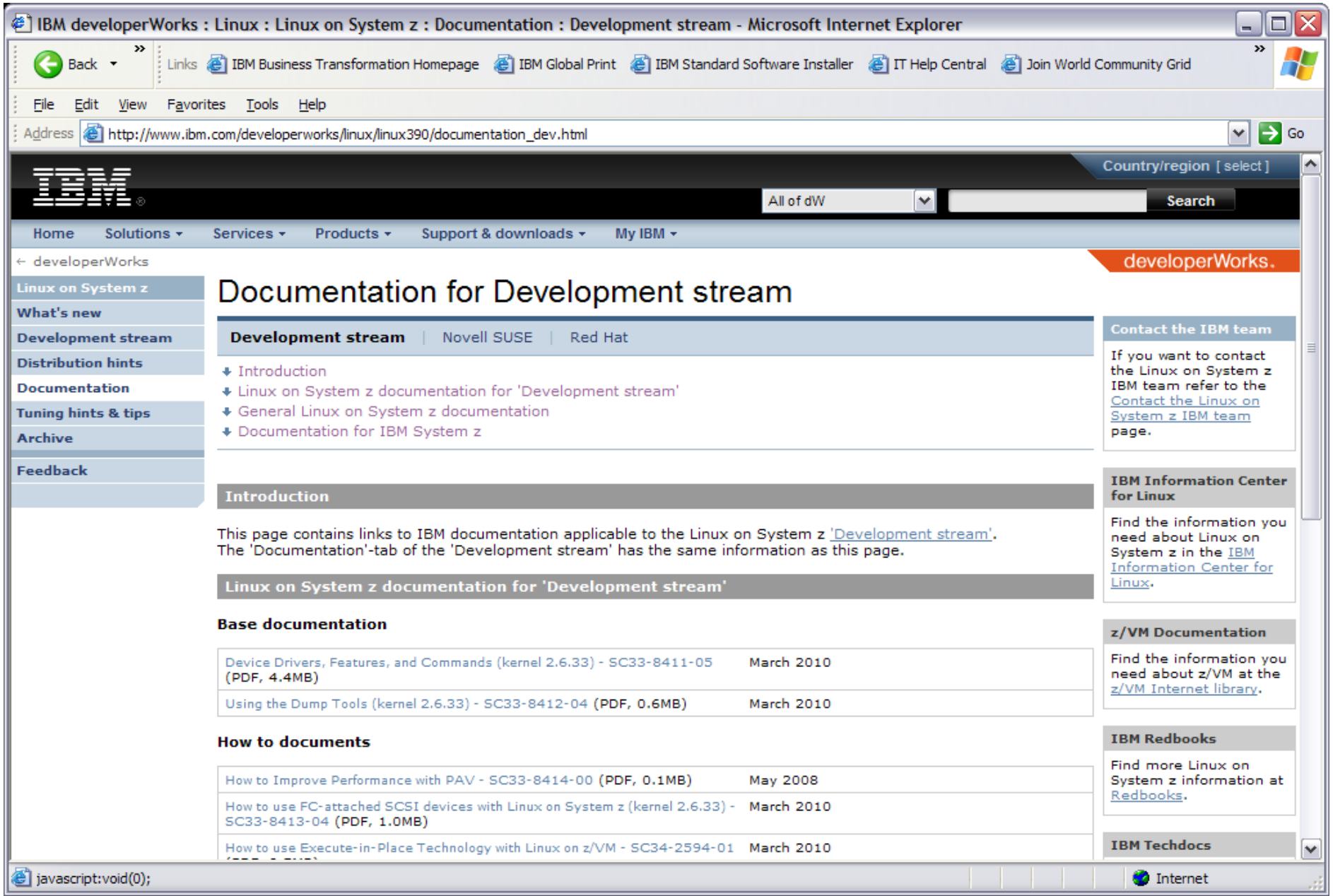
Resources

- Device Drivers, Features, and Commands
  - Chapter 5. SCSI-over-Fibre Channel device driver
  - http://www.ibm.com/developerworks/linux/linux390/documentation_dev.html
- How to use FC-attached SCSI devices with Linux on System z
  - http://www.ibm.com/developerworks/linux/linux390/documentation_dev.html
- lsscsi utility for linux
  - http://sg.danny.cz/scsi/lsscsi.html
- SLES 11 SP1: Storage Administration Guide
  - http://www.novell.com/documentation/sles11 /stor_admin/?page=/documentation/sles11/stor_admin/data/bookinfo.html
- Red Hat Enterprise Linux 6: Storage Administration Guide
  - http://docs.redhat.com/docs/en-US/Red_Hat_Enterprise_Linux/6/html/Storage_Administration
- System Storage Interoperation Center (SSIC)
  - http://www.ibm.com/systems/support/storage/config/ssic/
- System z FCP channels
  - http://www.ibm.com/systems/z/hardware/connectivity/fcp.html

developerWorks – entry page for documentation



Development stream – Novell SUSE – Red Hat documentation



More information

ibm.com/systems/z/linux

www.vm.ibm.com

Thank You!

Questions?

Holger.Smolinski@de.ibm.com