SCSI over FCP for Linux on System z – Introduction and New Features

Christof Schmitt <christof.schmitt@de.ibm.com>
IBM Germany Research & Development

2008-08-13
Session 9259

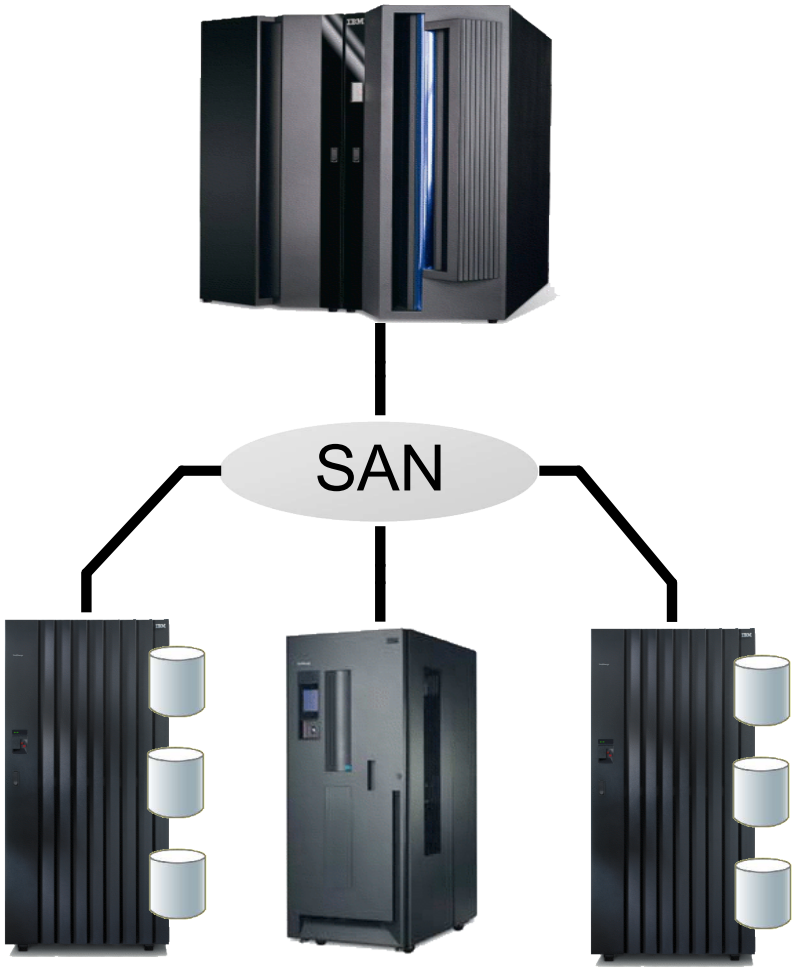
Abstract

The Linux zfcp device driver adds support for Fibre Channel attached SCSI devices to Linux on System z. The Fibre Channel protocol is an open, standard-based alternative and supplement to existing ESCON or FICON connections and becomes more and more important. The intention of this presentation is to give an introduction to the SCSI world on a System z mainframe. This presentation will cover hardware and software requirements, configuration, performance considerations, IPL and dump. Other points will be FCP support in recent Linux distributions and new features to improve the FCP support in Linux on System z.

Agenda

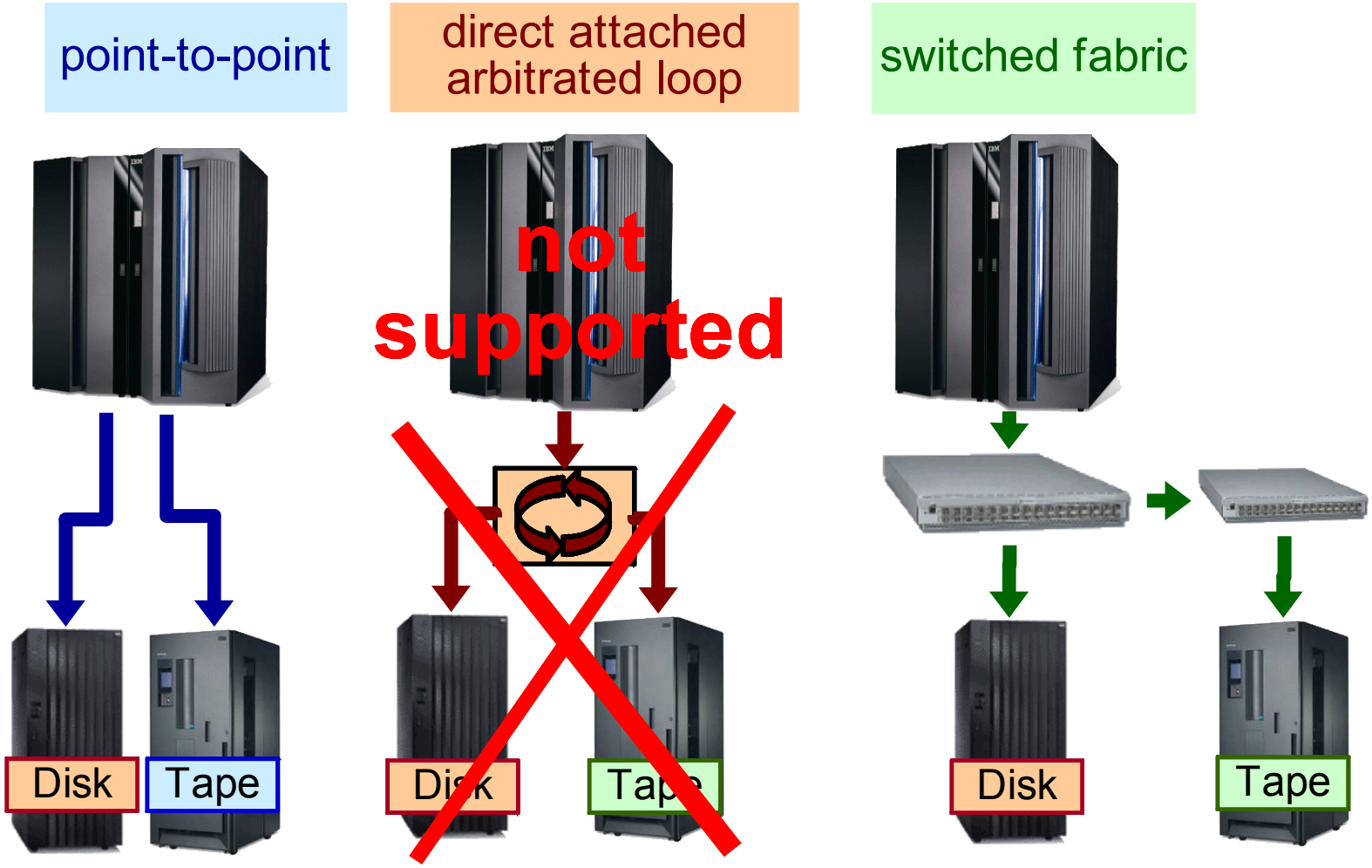
- Introduction
- Channel I/O vs. SCSI
- Hardware requirements
- Software, Configuration
- SCSI IPL
- SCSI dump
- Linux distribution support
- NPIV
- New features

Introduction

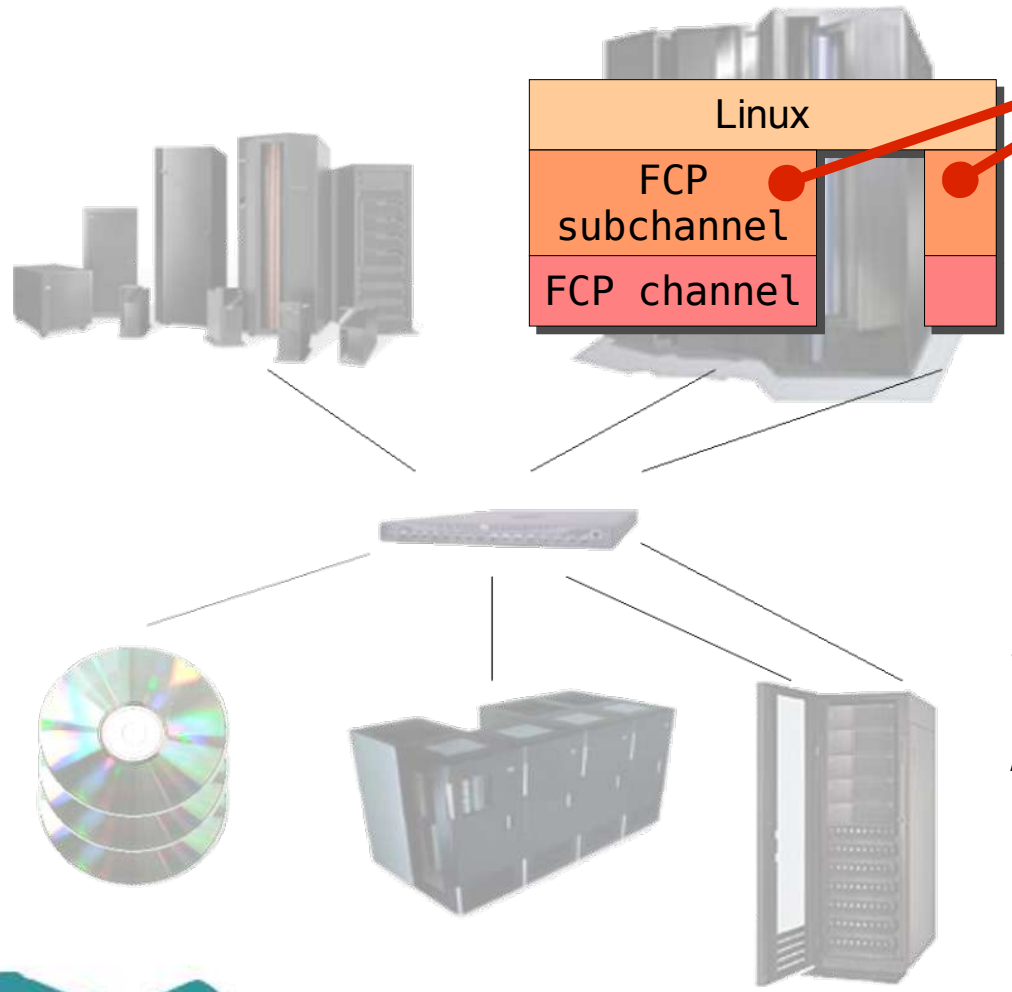


- connect System z to Open Systems storage (disk, tape, ...)
- integrate in existing Storage Area Networks (SANs)
- resource sharing with Open Systems
- I/O uses SCSI over Fibre Channel Protocol (FCP)
- System z FICON Express card
- zfcp as device driver in Linux

SAN topologies and System z



FCP channel and subchannel

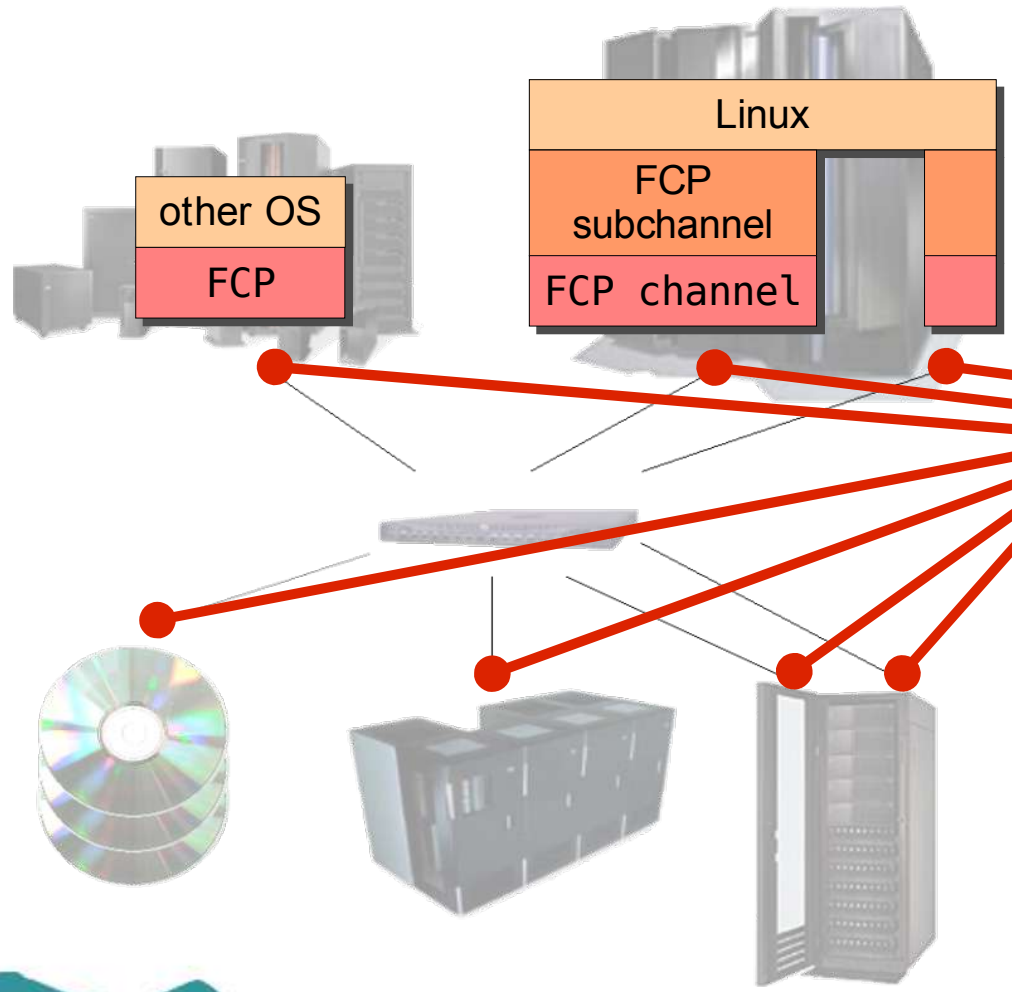


Linux connects through FCP subchannels to FCP attached storage.

A subchannel is identified – in Linux - by its bus identifier which is derived from the subchannel's device number.

sample FCP subchannel (as seen in Linux):
/sys/bus/ccw/drivers/zfcp/0.0.50d4

World Wide Port Names (WWPNs)



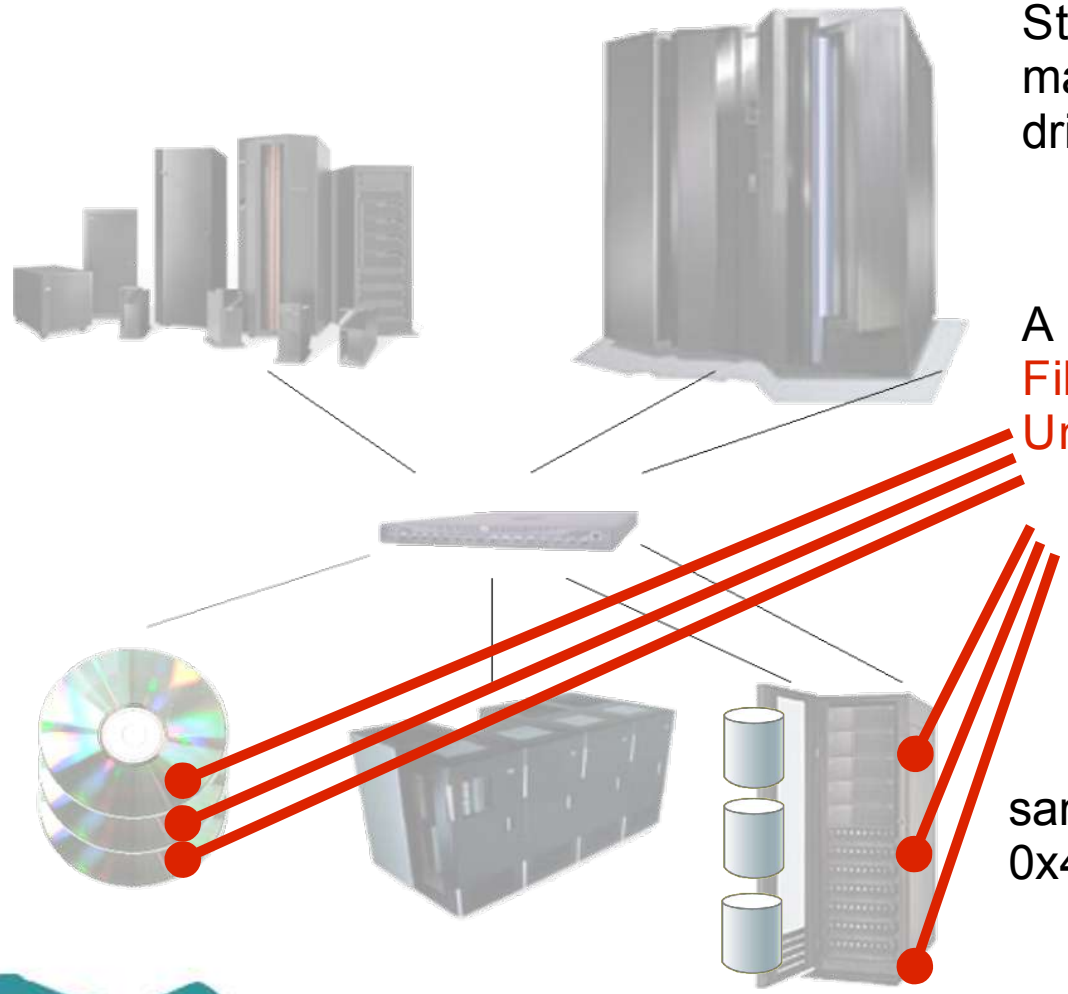
Storage devices and servers attach through Fibre Channel ports (called N_Ports).

An N_Port is identified by its World-Wide Port Name (WWPN).

For redundancy, servers or storage may attach through several N_Ports.

sample WWPN:
0x5005076300c20b8e

Logical Unit Numbers (LUNs)

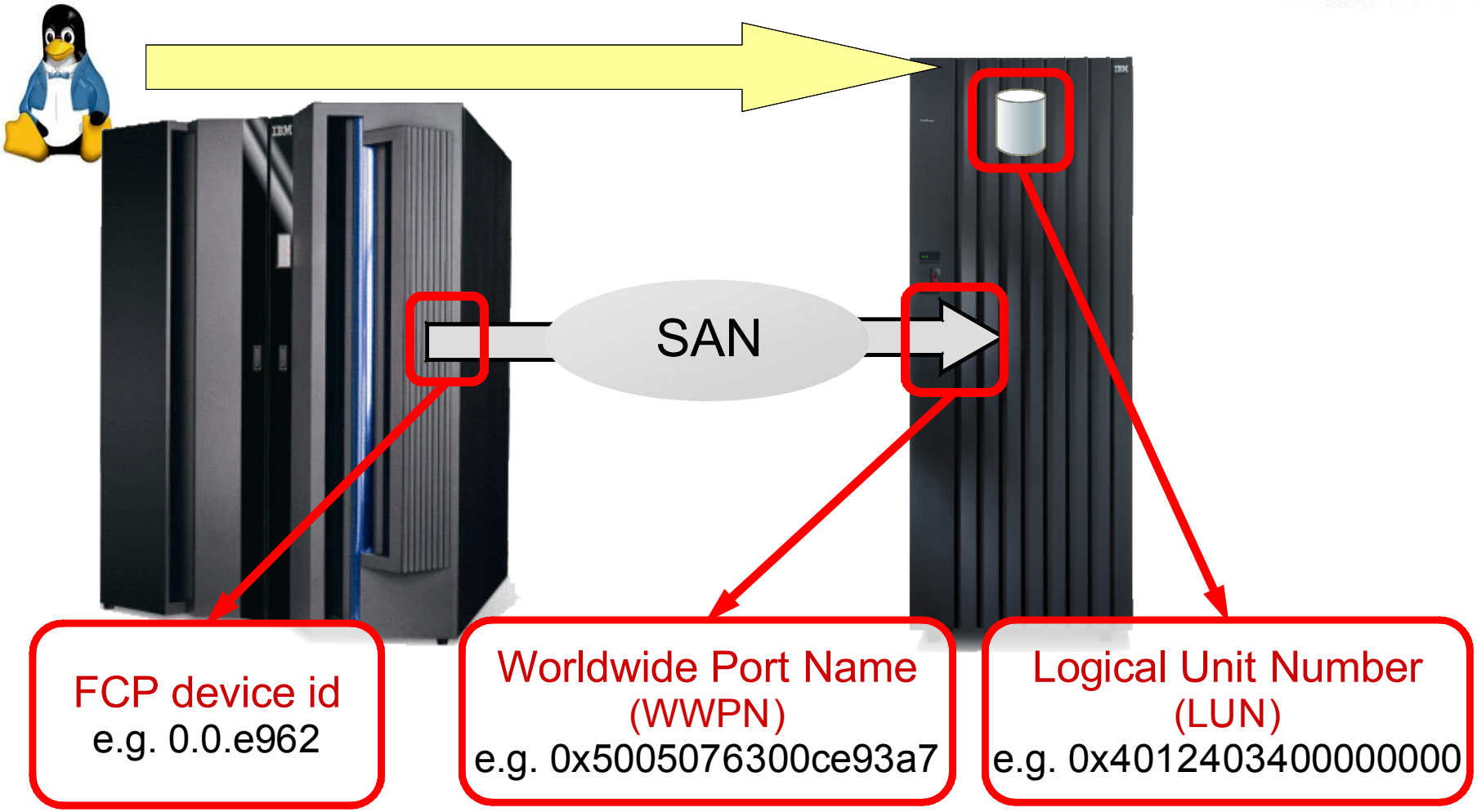


Storage devices usually comprise many logical units (volumes, tape drives, ...).

A logical unit is identified by its Fibre Channel Protocol Logical Unit Number (FCP LUN).

sample FCP LUN:
0x4010400200000000

Accessing SAN storage



SCSI compared to Channel I/O

- FCP adapter is defined in System z I/O configuration
- Ports and LUNs attachment handled in Operating Systems
- Multipathing handled in Operating System
- No disk size restrictions for SCSI disks
- Additional configuration outside System z necessary
  - Zoning in the SAN fabric
  - LUN masking on the storage server

Hardware requirements

- IBM zSeries 800, 890, 900 or 990
- IBM System z9 or z10
- FICON or FICON Express adapter cards
- Fibre Channel storage system
- Optional:
  - Fibre Channel switch (for Fabric topology)
  - IBM System z9 or z10 (for NPIV support)

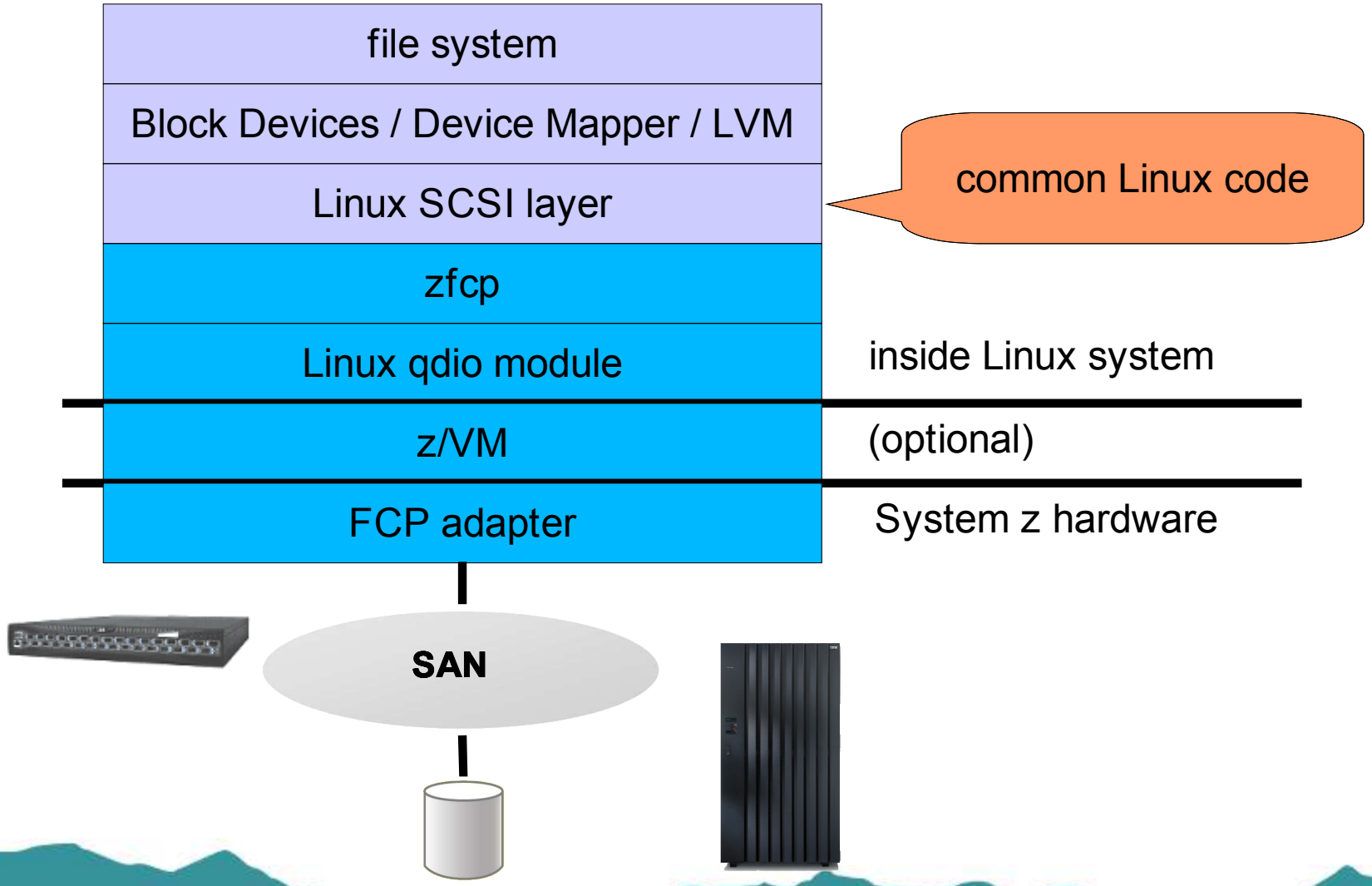
Hardware: Define FCP adapter in IOCDF

```
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
```

Software requirements

- zfcp is part of standard Linux kernel and standard distributions
- supported Linux distributions
  - SLES9
  - SLES10
  - RHEL4
  - RHEL5
- Recommendations
  - start with latest available update / service pack
  - include Linux in maintenance planning
  - check for possibly related z/VM PTFs

I/O stack for SCSI and Linux



Command line setup (SLES9/10, RHEL4/5)

```
# cd /sys/bus/ccw/drivers/zfcp/0.0.3c00/
# echo 1 > online
```

FCP adapter

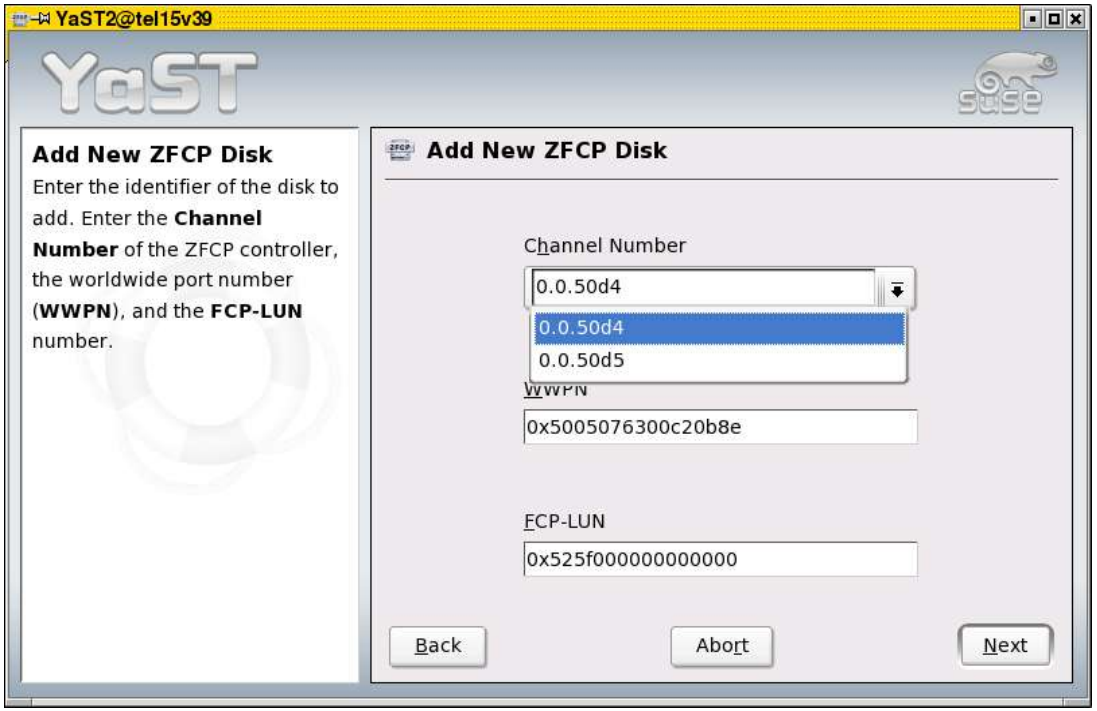
WWPN

```
# echo 0x500507630313c562 > port_add


# echo 0x401040cc00000000 > 0x500507630313c562/unit_add
```

LUN

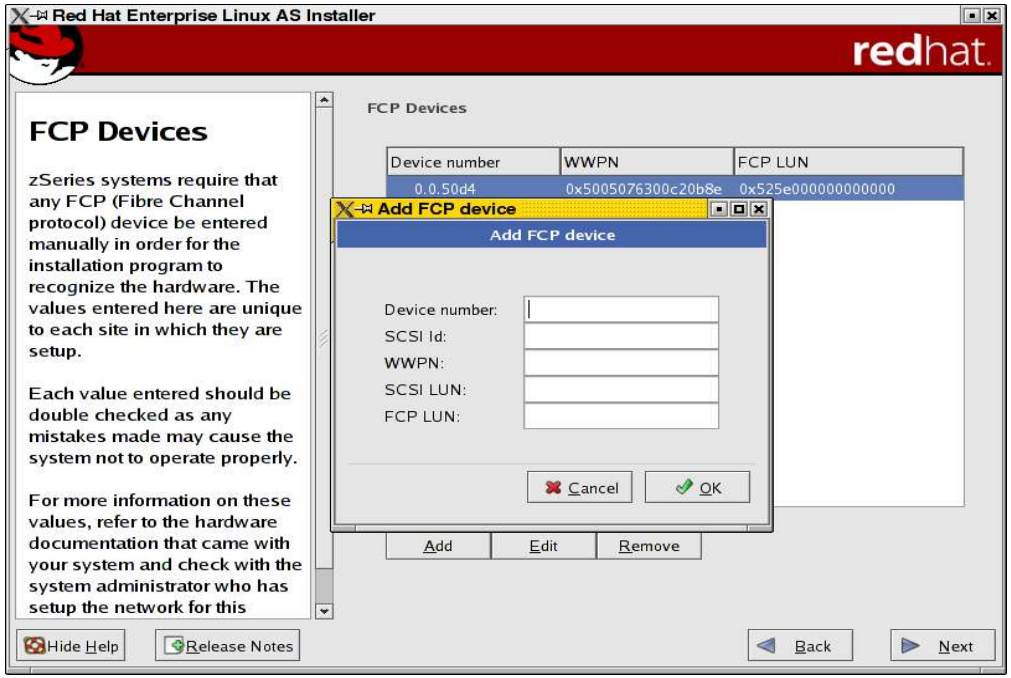
```
# lszfcp -D
0.0.3c00/0x500507630313c562/0x401040cc00000000
0:0:0:1087127568


# lsscsi -t
[0:0:0:1087127568]disk  fc:0x500507630313c562,0x650d13 /dev/
sda
```

Persistent setup: SLES



- through YaST or
- setup in /etc/sysconfig/hardware/hwcfg-zfcp-bus-ccw-0.0.*

Persistent setup: RHEL



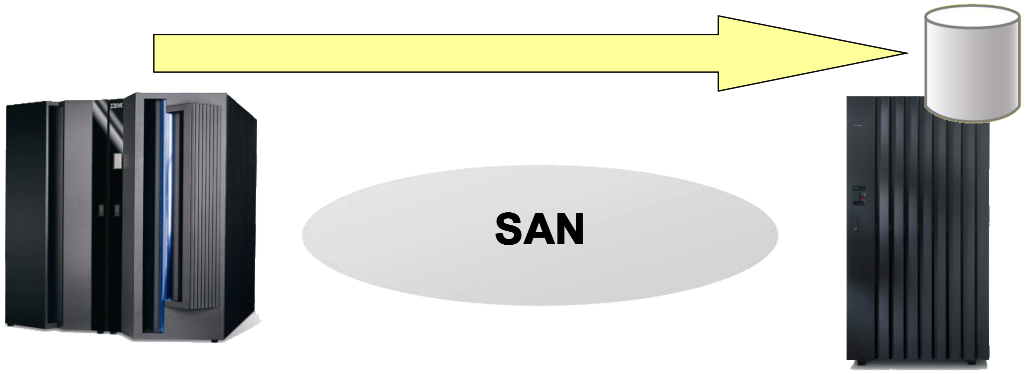
- Installer GUI or
- /etc/zfcp.conf
- SCSI and SCSI LUN are unused fields (use 0, 1, ...)

Multipathing considerations

- storage controllers allow different paths
- path failover required for
  - storage system maintenance
  - SAN fabric maintenance (with dual fabrics)
- implemented inside Linux
- disk storage: multipath-tools
- IBM tape drives: lin_tape driver
- more details: Session 9289, “Additional Feet for the Penguin - SCSI over FCP Multipathing for Linux on System z”

SCSI IPL

- Similar to IPL from DASD
- Requires to address the SCSI disk
  - FCP adapter id
  - Remote port
  - LUN



SCSI IPL example with z/VM

```
00:
00: CP SET LOADDEV PORTNAME 50050763 0313C562 LUN 401040CF 00000000
00:
```

WWPN

LUN

```
00: CP Q LOADDEV
PORTNAME 50050763 0313C562    LUN  401040CF 00000000    BOOTPROG 0
BR_LBA   00000000 00000000
00:
```

FCP adapter

```
00: CP IPL 3C00
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 0.18).
00: MLOPDM003I: Machine loader finished, moving data to final storage location.
Linux version 2.6.16.60-0.9-default (geeko@buildhost) (gcc version 4.1.2 2007011
5 (SUSE Linux)) #1 SMP Mon Mar 17 17:16:31 UTC 2008
We are running under VM (64 bit mode)
Detected 2 CPU's
Boot cpu address  0
Built 1 zonelists
Kernel command line: root=/dev/disk/by-id/scsi-36005076303ffc56200000000000010ce
-part1   TERM=dumb
```

SCSI IPL for LPARs

Load - H05:H05LP26

CPC:	H05:H05LP26
Image:	H05:H05LP26
Load type	○ Normal ○ Clear ◉ SCSI ○ SCSI dump
☐ Store status	
Load address	* 5900
Load parameter	
Time-out value	60 — 60 to 600 seconds
Worldwide port name	50050763030BC562
Logical unit number	4011400B0000000C
Boot program selector	0
Boot record logical block address	0
Operating system specific load parameters	

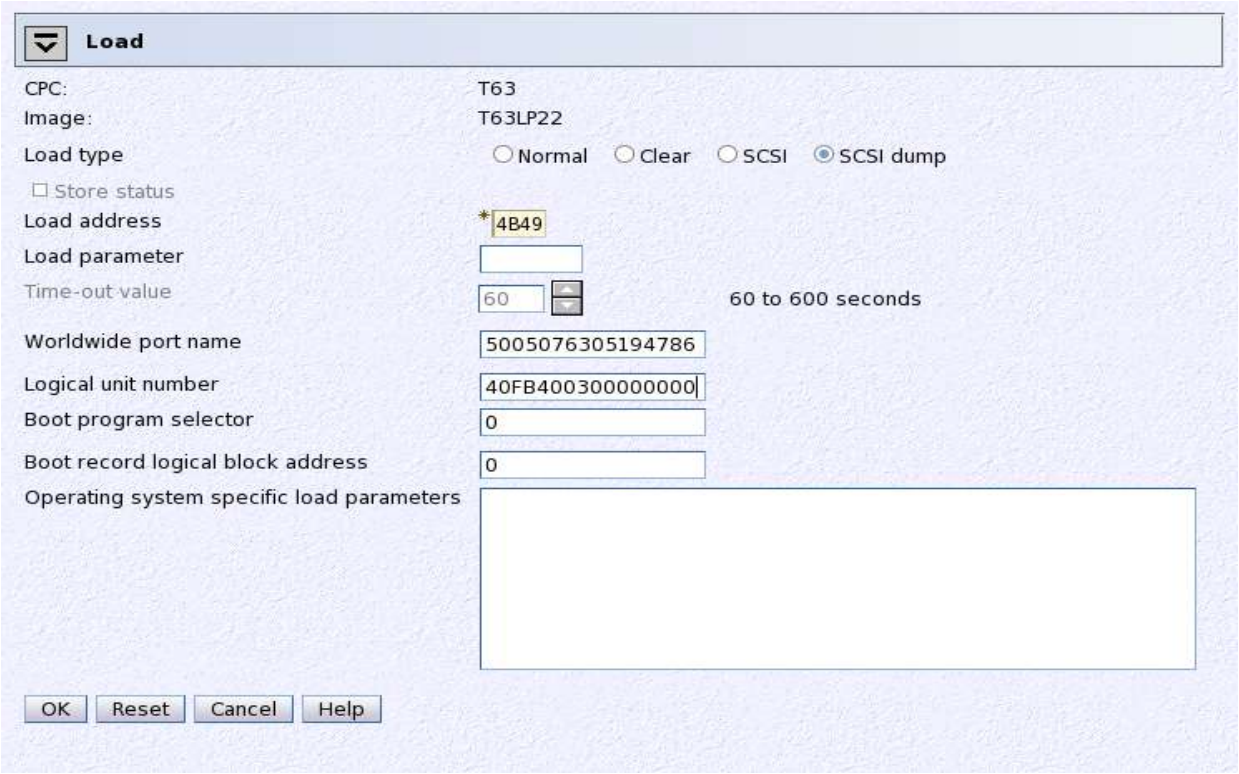
OK | Reset | Cancel | Help

SCSI dump

- Dump memory of one LPAR to disk for problem analysis
- Similar to VMDUMP and dump to DASD
- But: SCSI dump only supported for LPARs, not z/VM
- Preparation summary:
  - large SCSI disk (LPAR memory + 10MB)
  - fdisk /dev/sda
  - mke2fs /dev/sda1
  - mount /dev/sda1 /mnt
  - zipl -D /dev/sda1 -t /mnt
  - umount /mnt

Issue SCSI dump from HMC

- Select CPC image for LPAR to dump
- Goto Load panel
- Issue SCSI dump
  - FCP device
  - WWPN
  - LUN

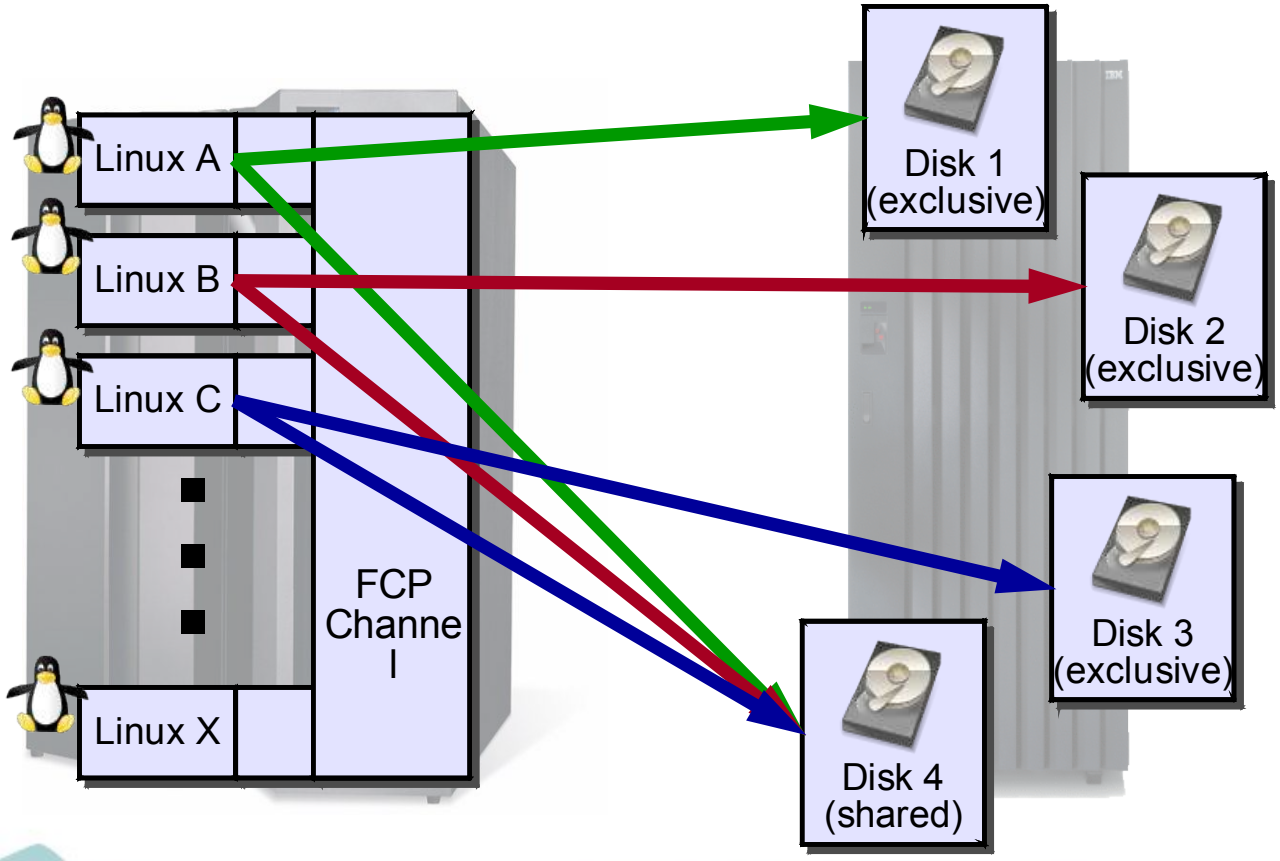


NPIV

- N_Port Identifier Virtualization (NPIV)
- without NPIV: one WWPN for FCP channel
- with NPIV: unique WWPN for each FCP subchannel
- enables proper zoning in SAN fabrics
- enables proper LUN masking in storage devices
- security
- access control

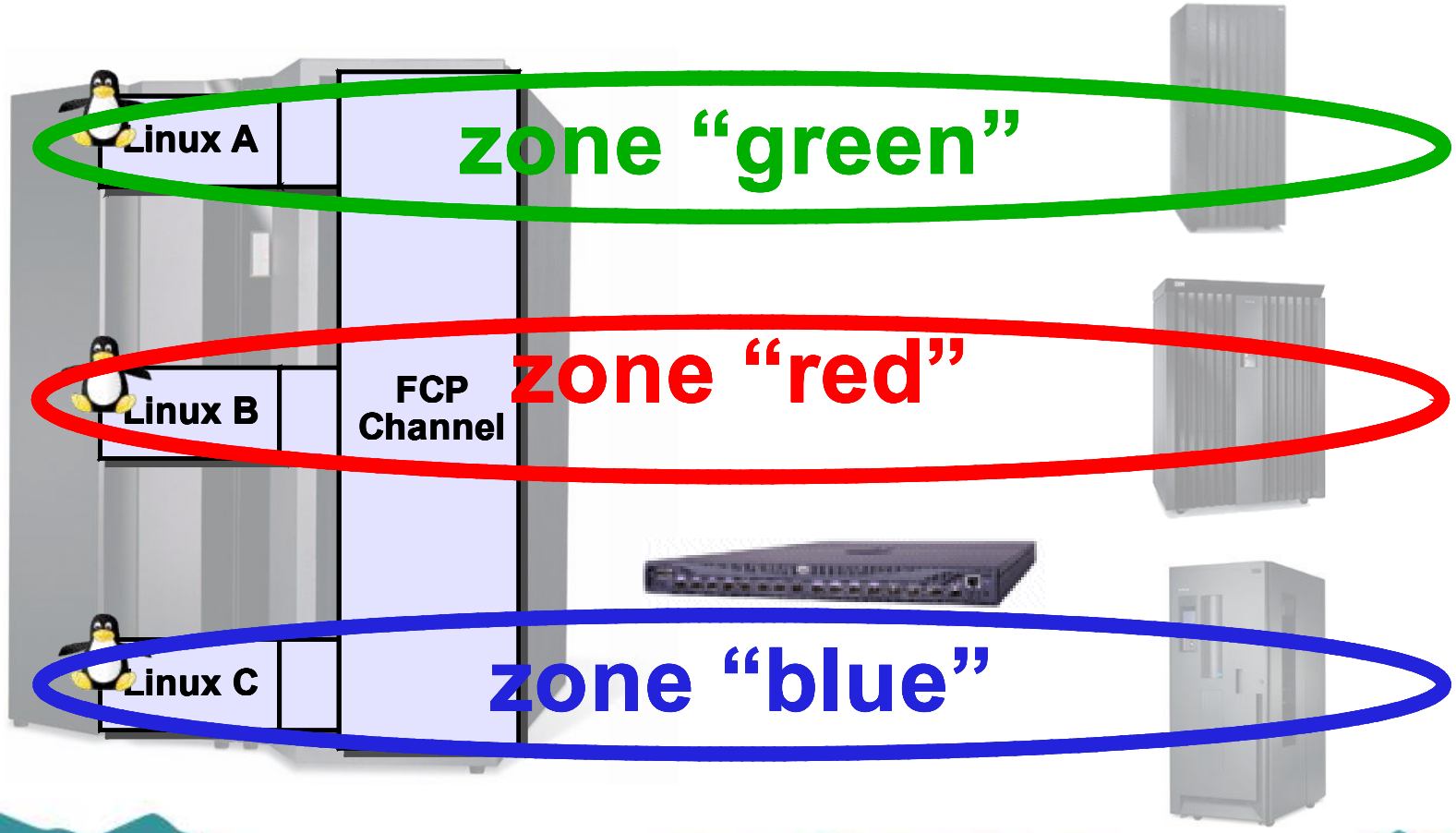
LUN masking with NPIV

Storage server can identify Linux guests via WWPNs



SAN zoning with NPIV

- Different Linux guests in different zones



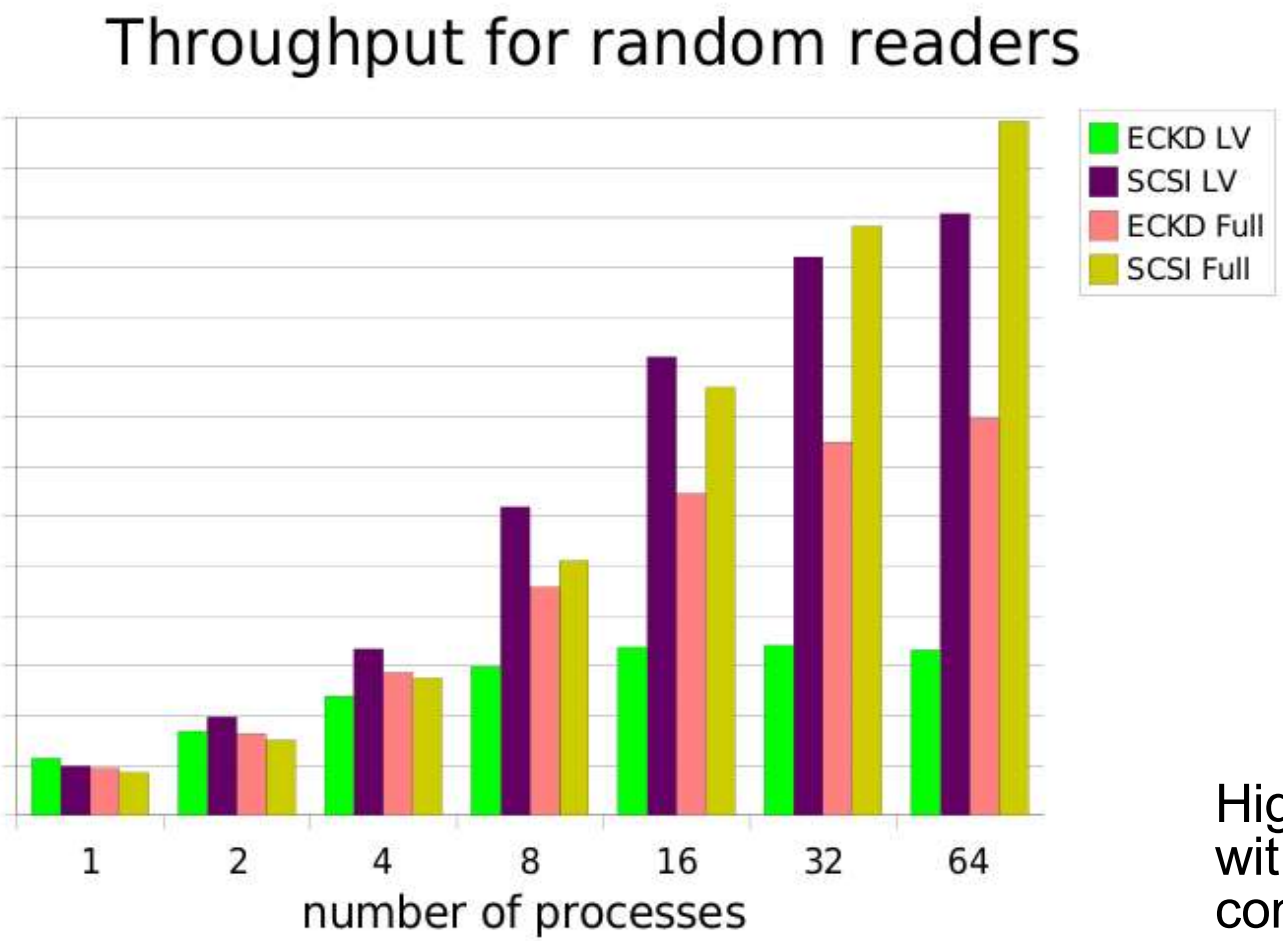
NPIV requirements

+

+

- NPIV is available on System z servers.
  - FICON Express 2 adapter running with MCL003 on EC J99658
- z/VM
  - z/VM 5.2 or 5.3
  - z/VM 5.1 with the PTF for APAR VM63744
- Linux Distribution
  - Currently SLES9 SP3/4, SLES10, RHEL5 (LPAR mode or z/VM)
- NPIV-Capable Switch
  - only required for switch adjacent to System z

Performance considerations



- DS8300
- z990 LPAR
- SLES9 SP2
  - 8 CPUs
  - 8 FICON
  - 8 FCP
  - 256 MB
- Iozone 3.96

Higher performance with SCSI disks compared to ECKD

New features overview

- Blktrace for I/O and latency tracing
- Channel statistics from sysfs (2.6.26)
- Channel and fabric latencies in upstream code (2.6.27)
- Automatic port discovery (2.6.27)
- Message cleanup (2.6.27)

blktrace

- “old” zfcp specific statistics only available as “add-on” patch
- new kernel infrastructure and tools: blktrace
- common I/O tracing infrastructure in Linux
- request sizes / latencies
- functional replacement for most of the zfcp specific statistics

```
==================== All Devices ====================
       ALL           MIN            AVG         MAX            N
--------------- ------------- ------------- ------------- -----------
Q2Q               0.000000072   0.000086313   5.453721801   1257686
Q2I               0.000000359   0.000000516   0.023150311   1257687
I2D               0.000000933   0.003573727   0.487170508   1267275
D2C               0.000363719   0.034028080   0.708048174   1257687
Q2C               0.000395336   0.037609824   0.708064315   1257687
```

(Queued, Issued, Dispatched, Complete)
D2C == Dispatched to Complete

New statistics from sysfs

Subchannel

```
# cat /sys/class/scsi_host/host0/megabytes
16 1
```

read, written

```
# cat /sys/class/scsi_host/host0/requests
3963 37 2
```

input, output, control requests

```
# cat /sys/class/scsi_host/host0/seconds_active
2871
```

more channel data in /sys/class/fc_host/host0/statistics/

FCP Channel utilization

```
# cat /sys/class/scsi_host/host0/utilization
2 10 0
```

channel processor, channel bus, adapter

Channel and fabric latencies

- previously part of zfcp statistics add-on
- available through sysfs (2.6.27)

```
# cat /sys/block/sda/device/write_latency
273 4562 67446 32 616 9212 37

# cat /sys/block/sda/device/read_latency
129 77471 18507807 21 67 110541 3963

# cat /sys/block/sda/device/cmd_latency
92 98 190 23 28 52 2
```

request count

fabric: min, max, sum (micro seconds)

channel: min, max, sum (micro seconds)

Automatic port discovery

- Discover and attach ports automatically
  - when setting adapter online,
  - on changes in SAN.

```
# cd /sys/bus/ccw/drivers/zfcp/0.0.181d/
# echo 1 > online
# ls -d 0x*
0x500308c141699001  0x5005076300cbb130  0x5005076303048335
0x500507630310c562  0x500507630e0202aa  0x500308c141699004
0x5005076300cc0b8e  0x5005076303098335  0x500507630313c562
```

- Manual trigger available:

```
# echo 1 > port_rescan
```

- Does not change handling of LUNs
- LUNs have to be attached manually:

```
# echo 0x401040C300000000 > 0x500507630310c562/unit_add
```

Message cleanup

- removed debug and trace messages
- removed information also available in sysfs or lszfpc
- standard format (zfcp: 0.0.XXXX: ...)

old

```
zfcp: The adapter 0.0.181d reported the following characteristics:
WWNN 0x5005076400c2d09e, WWPN 0x5005076401e071b2, S_ID 0x00689313,
adapter version 0x3, LIC version 0x170b, FC link speed 2 Gb/s
zfcp: Switched fabric fibrechannel network detected at adapter 0.0.181d.
zfcp: adapter 0.0.181d: no path
zfcp: adapter 0.0.181d: operational again
```

new

```
zfcp: 0.0.181d: Switched fabric fibrechannel network detected.
zfcp: 0.0.181d: no path
zfcp: 0.0.181d: operational again
```

References

- Supported devices
  http://www.ibm.com/systems/z/hardware/connectivity/products/fc.html
- Storage device interoperability
  http://www.ibm.com/systems/storage/disk/ds6000/pdf/interop.pdf
  http://www.ibm.com/systems/storage/disk/ds8000/interop.pdf
  http://www.ibm.com/systems/storage/software/virtualization/svc/interop.html
  http://www.ibm.com/systems/support/storage/config/ssic/index.jsp
- Linux on System z Documentation
  http://www.ibm.com/developerworks/linux/linux390/
  - Device Drivers, Features, and Commands
  - How to use FC-attached SCSI devices with Linux on System z
  - Using the Dump Tools
- Tuning hints & tips
  http://www.ibm.com/developerworks/linux/linux390/perf/
- Device driver for IBM tape drives
  http://ftp.software.ibm.com/storage/devdrvr/Doc/

SHARE
Technology · Connections · Results

Trademarks

SHARE
Technology · Connections · Results

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.