Making Your Penguins Fly

Introduction to SCSI over FCP

for Linux on System z

Horst Hummel
IBM

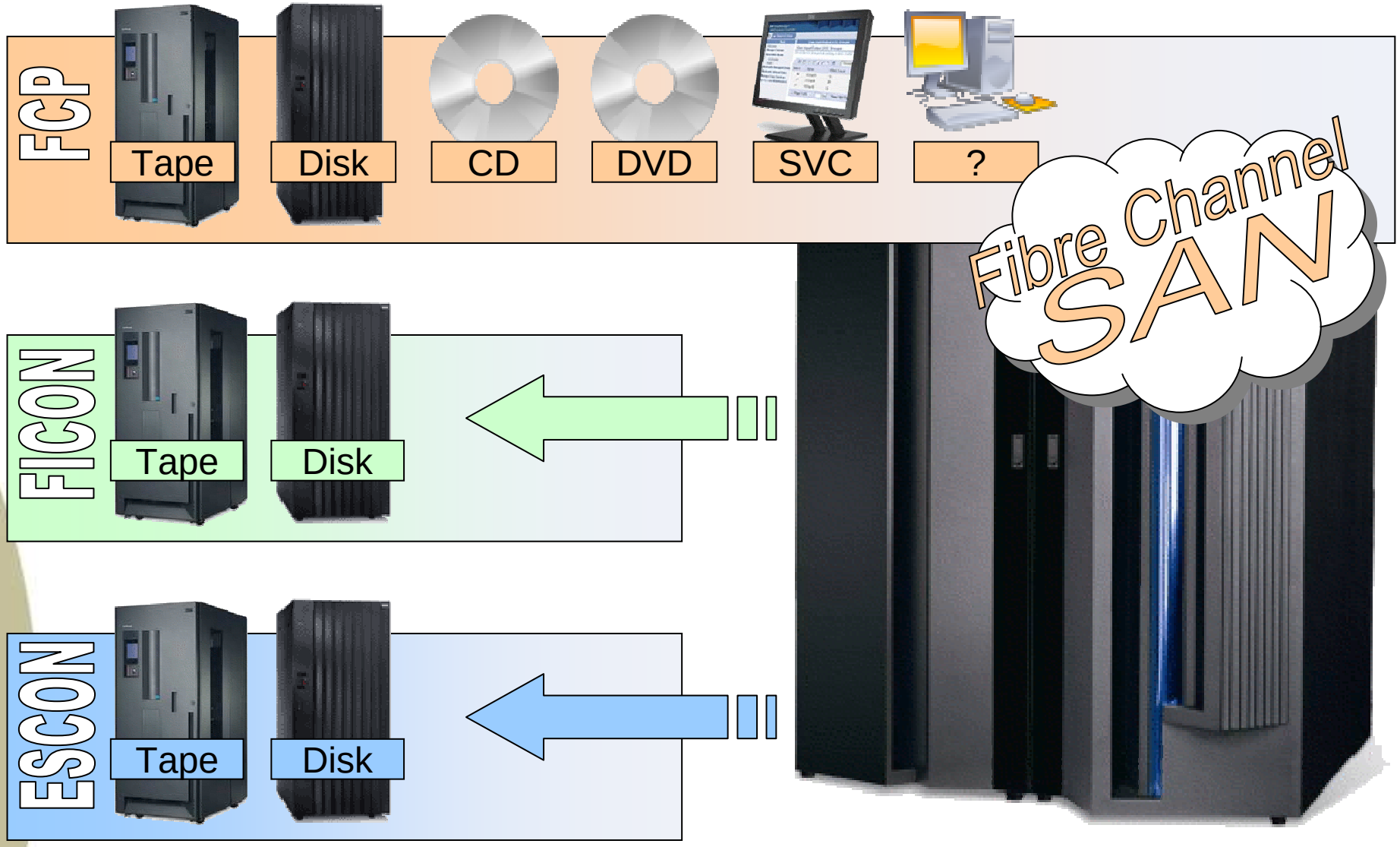
Horst.Hummel@de.ibm.com

February 27th 2008
Session 9259

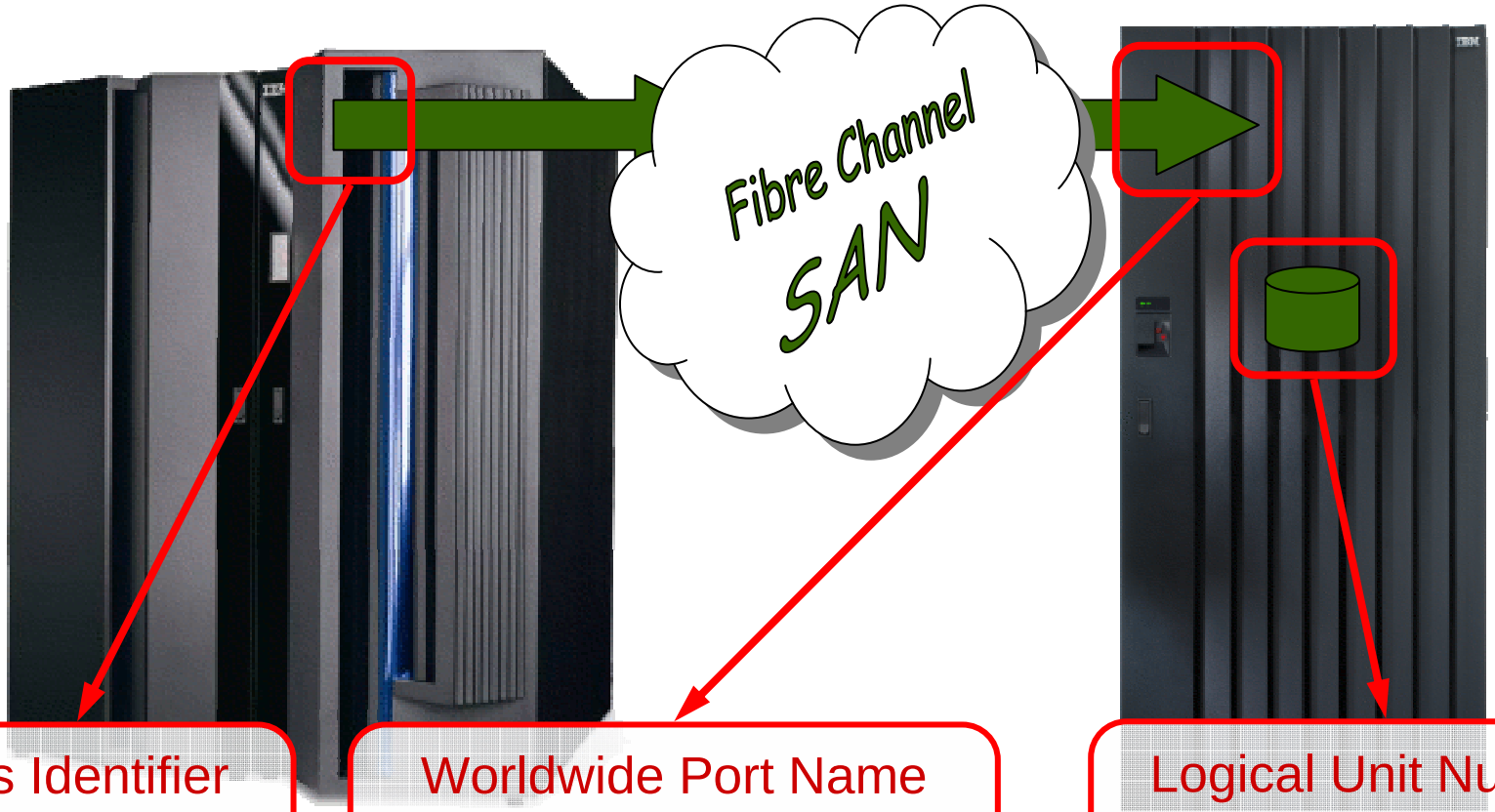
Making Your Penguins Fly – Flight Schedule

- SAN & SAN integration
- Addressing basics
- Requirements
- zfcp device driver
- Why FCP?
- Configuration
- Multipathing
- SCSI tape
- SAN Discovery Tool
- NPIV
- SCSI IPL/Dump

System z in a SAN – Sharing Storage Resources



Navigating in a SAN

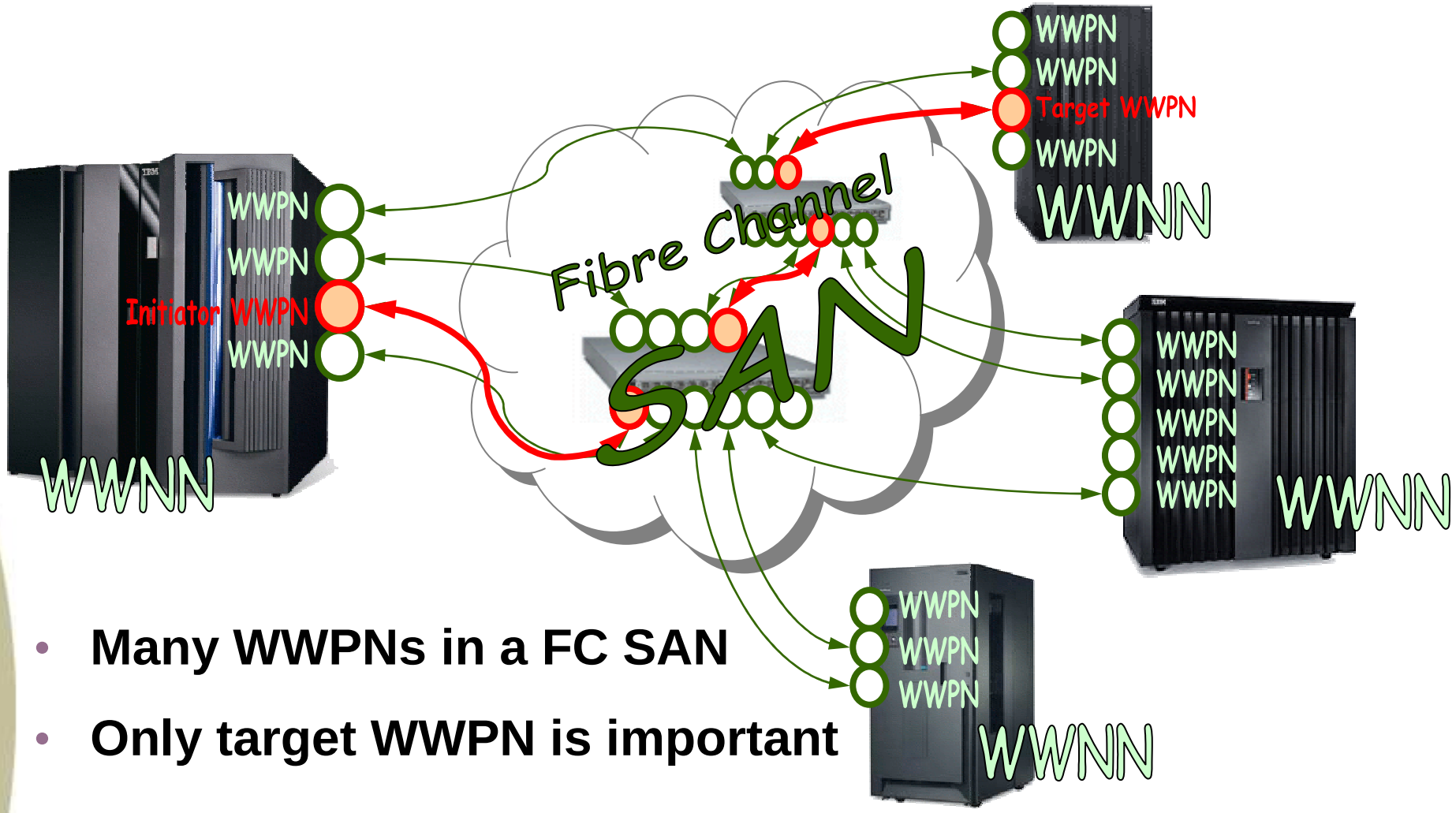


Bus Identifier
(busid)
e.g. 0.0.e962

Worldwide Port Name
(WWPN)
e.g. 0x5005076300ce93a7

Logical Unit Number
(LUN)
e.g. 0x4012403400000000

WWPNs and other Ports



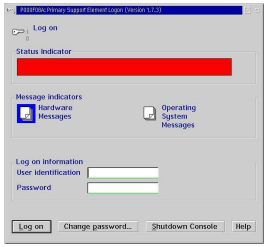
- **Many WWPNs in a FC SAN**
- **Only target WWPN is important**

System z in a SAN – Hardware Requirements

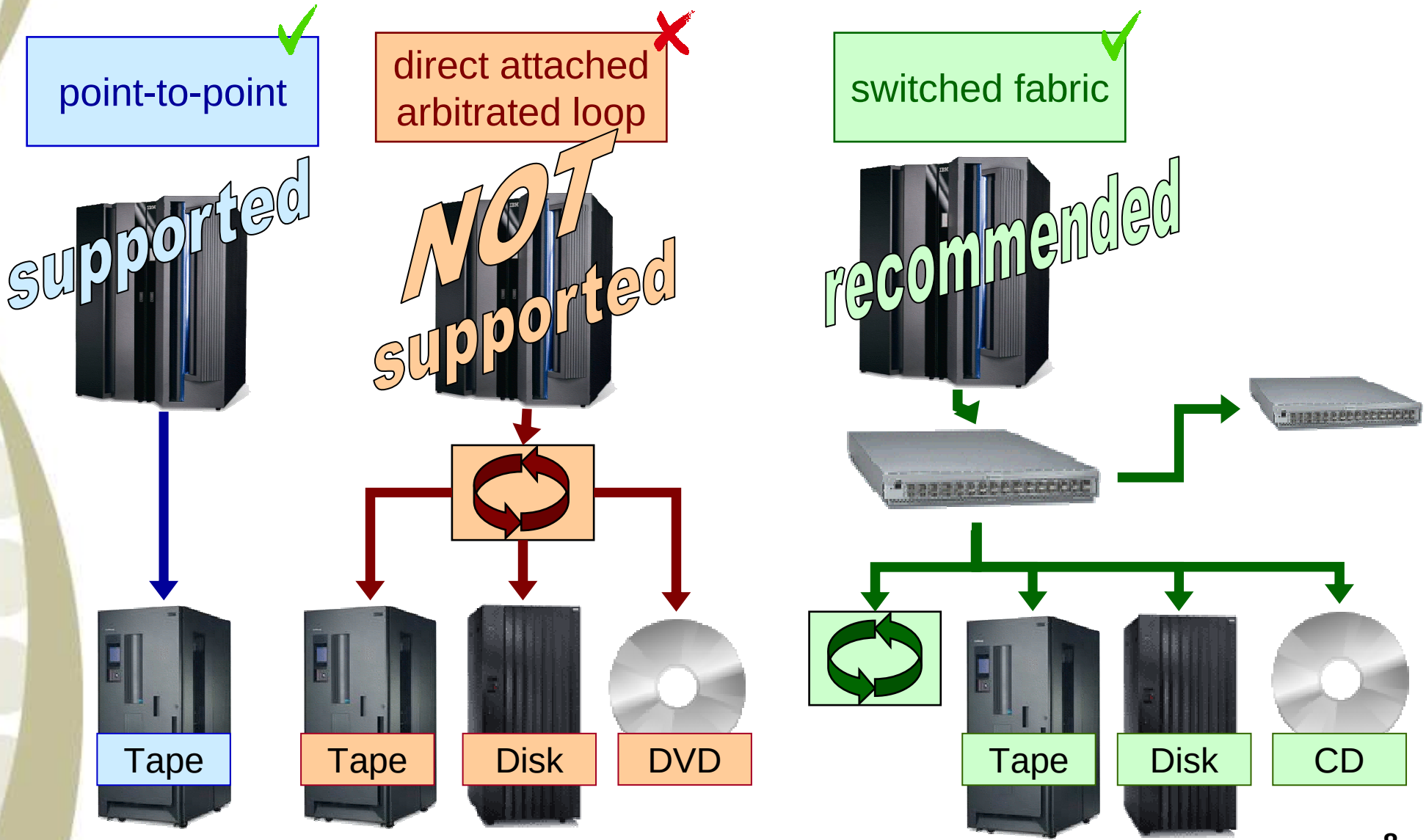
- **IBM zSeries 800, 890, 900 or 990**
  IBM System z9 EC/BC
  (NPIV z9-only)
- **FICON or FICON Express adapter cards**
- **CHPID type FCP**
- **FC fabric switch**
- **FC storage subsystem**
- **Optional: FCP-SCSI bridge + SCSI devices**

IOCDS – FCP Configuration Sample

```
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


CHPID PATH=(CSS(2),58),SHARED,                                        *
      PARTITION=((T29LP32,T29LP33),(=)),PCHID=500,TYPE=FCP
CNTLUNIT CUNUMBR=1781,PATH=((CSS(2),58)),UNIT=FCP
IODEVICE ADDRESS=(1780,064),UNITADD=00,CUNUMBR=(1781),UNIT=FCP
```



System z in a SAN – Topologies



Device Support

- **IBM I/O connectivity website**
  http://www-03.ibm.com/systems/z/connectivity/products/fc.html

- **IBM TotalStorage 3590 Tape Drive**
- **IBM TotalStorage 3592 Tape Drive**
- **IBM TotalStorage 3494 Tape Library**
- **IBM TotalStorage 3584 Tape Library**
- **IBM TotalStorage DS6000**
- **IBM TotalStorage DS8000**
- **Director/Switch Support**
  - **CISCO MDS 9020, 9120, 9140 Fabric Switch (IBM 2061-420, 020, 040)**
  - **CISCO MDS 9216 (IBM 2062-D01, D1A, D1H)**
  - **CISCO MDS 9500 Directors (IBM 2062-D04, D07, E11)**
  - **CNT (INRANGE) FC/9000 Directors (2042-001, -128, -256)**
  - **CNT UltraNet Multi-service Director (2042-N16)**
  - **IBM TotalStorage SAN256N director (2045-N16)**
  - **IBM Total Storage SAN140M (2027-140)**
  - **IBM TotalStorage SAN256M (2027-256)**
  - **McDATA Intrepid 6064 and 6140 Directors (2032-064, 140)**
  - **…**

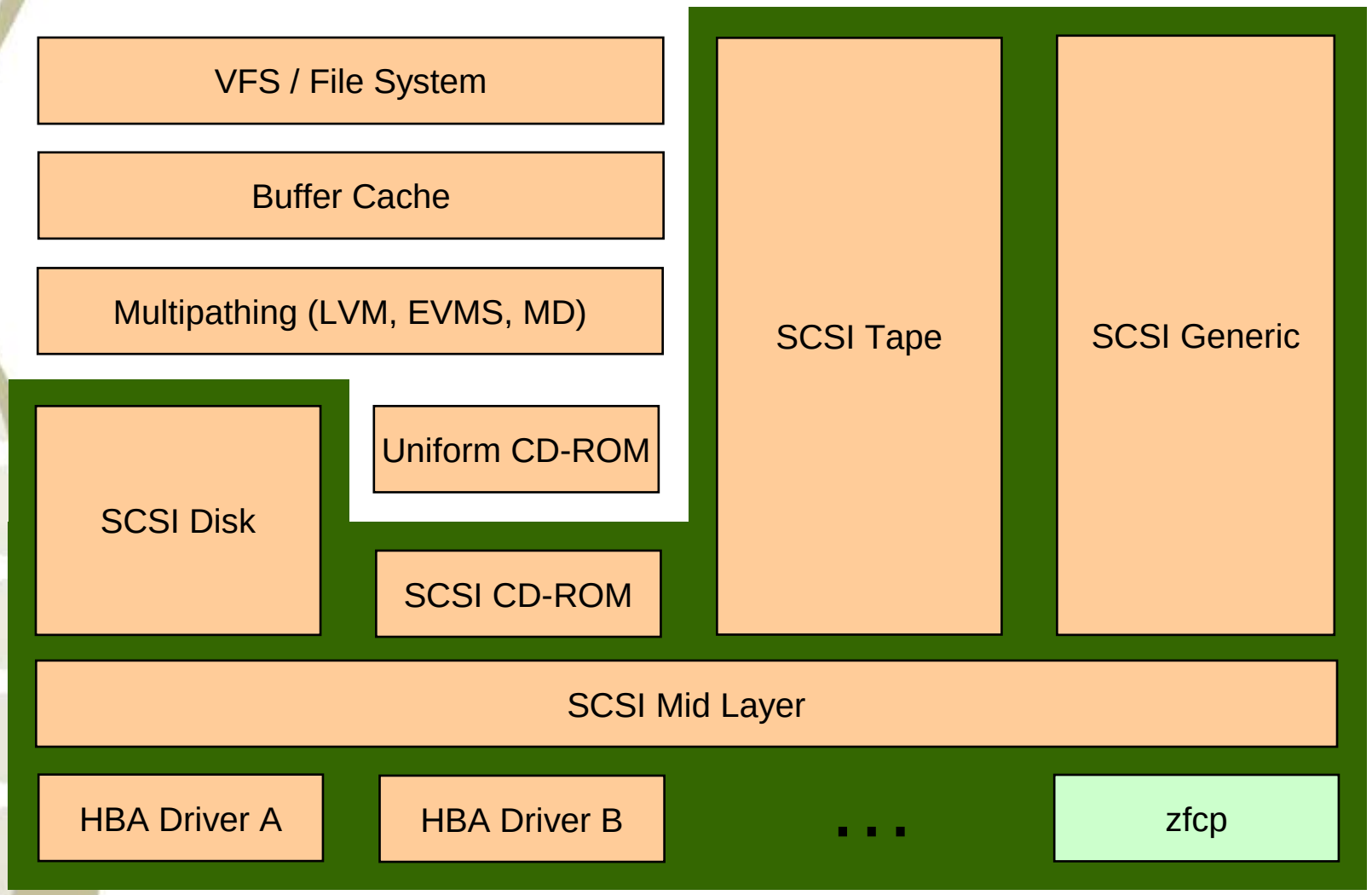
Software Requirements

- **SCSI (IPL) with z/VM**
  - **z/VM Version 4.4 (PTF UM30989) or newer**
  - **z/VM Version 5.3 (current version)**

- **SLES9 - SUSE Linux Enterprise Server 9 (GA 08/2004)**
  - **Kernel 2.6.5, GCC 3.3.3**
  - **Service Pack 3 (GA 12/2005)**
- **SLES10 - SUSE Linux Enterprise Server 10 (GA 07/2006)**
  - **Kernel 2.6.16, GCC 4.1.0**
  - **Service Pack 1 (GA 06/2007)**

- **RHEL4 - Red Hat Enterprise Linux AS 4 (GA 02/2005)**
  - **Kernel 2.6.9, GCC 3.4.3**
  - **Update 5 (GA 05/2007)**
- **RHEL5 - Red Hat Enterprise Linux AS 5 (GA 03/2007)**
  - **Kernel 2.6.18, GCC 4.1.0**
  - **Available**

Linux SCSI Stack

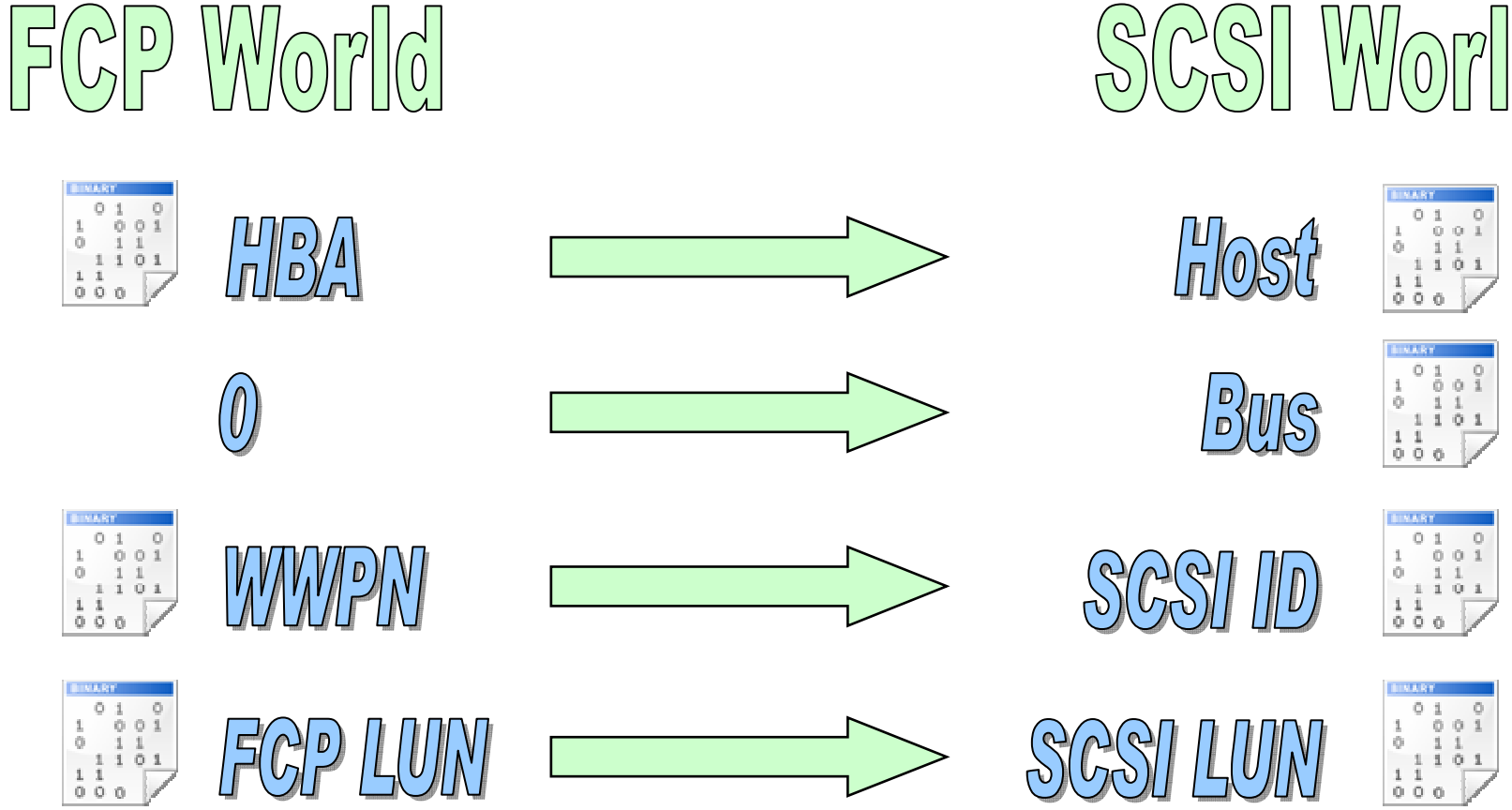


SCSI Stack

zfcp's Task in the Linux SCSI Stack

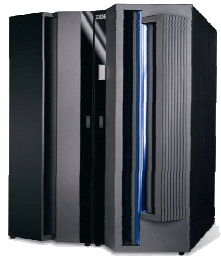
- **zfcp drives the System z FCP host bus adapter.**
  - **maintains connections through the SAN to SCSI devices attached via a System z FCP adapter.**
  - **maps SAN devices to SCSI devices as seen by the Linux SCSI subsystem.**
  - **sends SCSI commands and associated data on behalf of the Linux SCSI subsystem to SCSI devices attached via a System z FCP adapter.**
  - **returns replies and data from SCSI devices to the Linux SCSI subsystem.**

FCP – SCSI Mapping



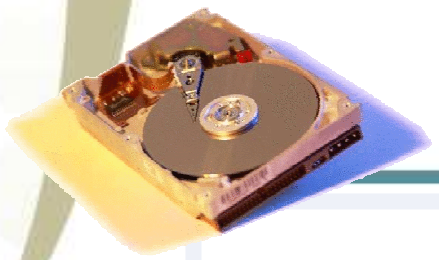
Why FCP?

- **Completely new set of IPL I/O devices**
  - **SCSI over Fiber Channel I/O devices**
  - **Different to any traditional z I/O device**
- **Additional addressing parameters**
- **Performance advantages**
  - **FCP is much faster than FICON**
  - **Reason 1: asynchronous I/O**
  - **Reason 2: no ECKD emulation overhead**
- **No disk size restrictions**
- **Up to 15 partitions (16 minor numbers per device)**
- **SCSI disks do not waste disk space (no low-level formatting)**



Why FCP? – cont.

- System z integration in existing FC SANs
- Use of existing FICON infrastructure
  - FICON/FICON Express adapter cards
  - FC switches / Cabling
  - Storage subsystems
- Dynamic configuration
  - Adding of new storage subsystems possible without IOCDS change
- Does NOT require more CPU than FICON
- SAN access control mechanisms using NPIV (z9 only)
- Get rid of FICON topology constraints, FCP is much more flexible.

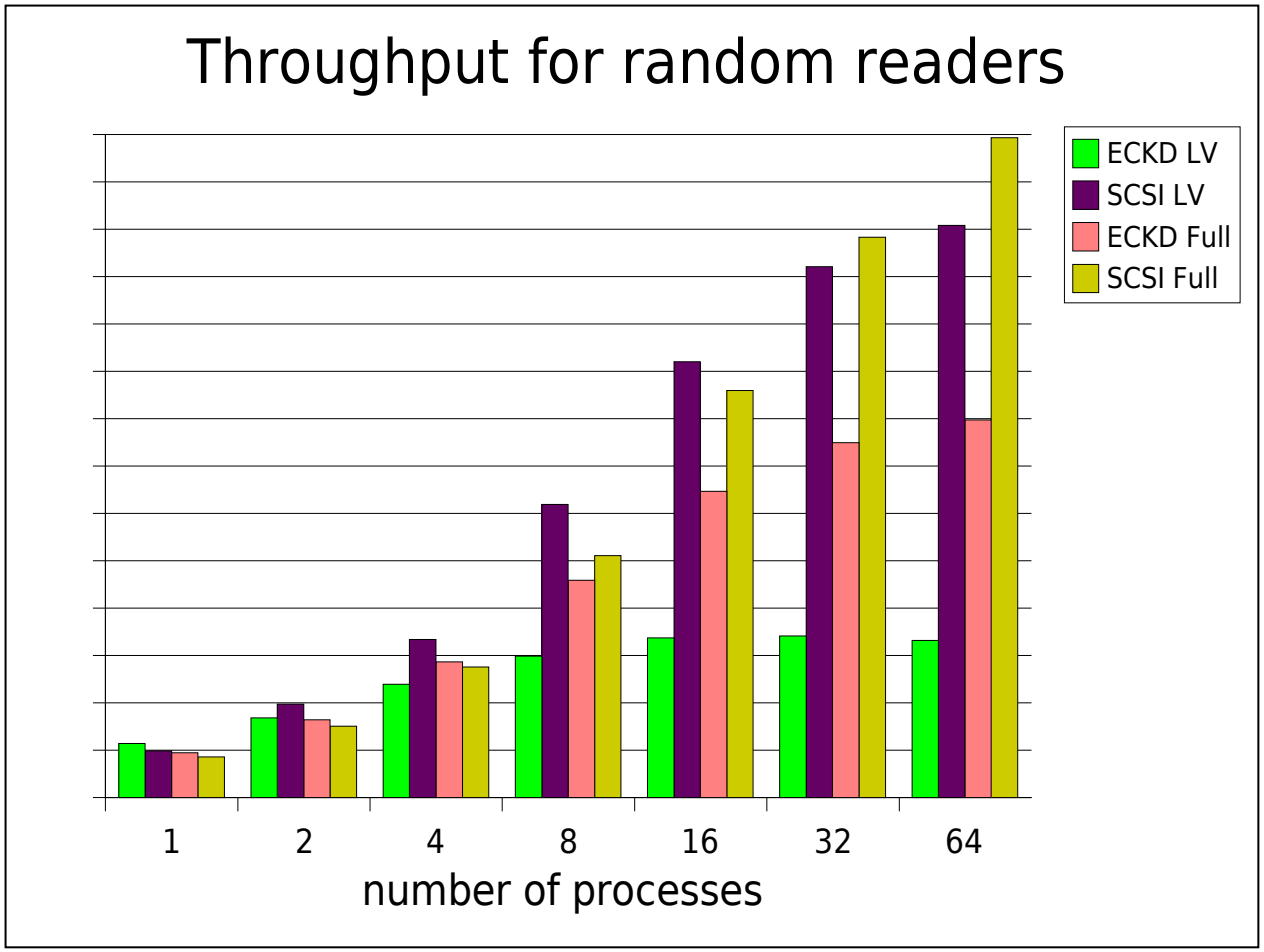


ECKD and SCSI Comparison

SHARE
Technology · Connections · Results

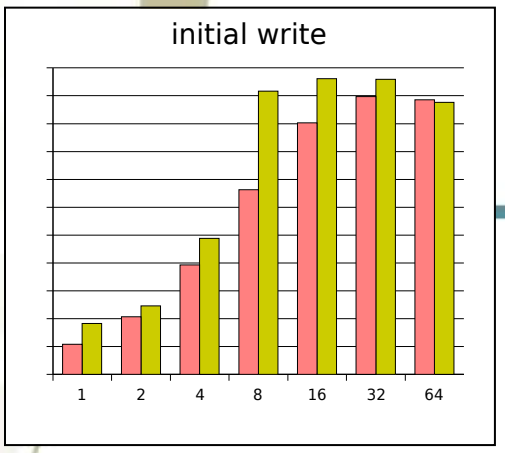
	ECKD DASD	SCSI Disk
Configuration	IOCDS/VM (operator)	IOCDS/VM & SAN & Linux (operator & SAN admin & Linux admin)
Access Method	SSCH/CCW	QDIO
Block Size (Byte)	512, 1K, 2K, 4K	512
Disk Size	3390 Model 3/9	any
Formatting (low level)	dasdfmt	not necessary
Partitioning	fdasd	fdisk
File System	mke2fs (or others)	
Access	Mount	

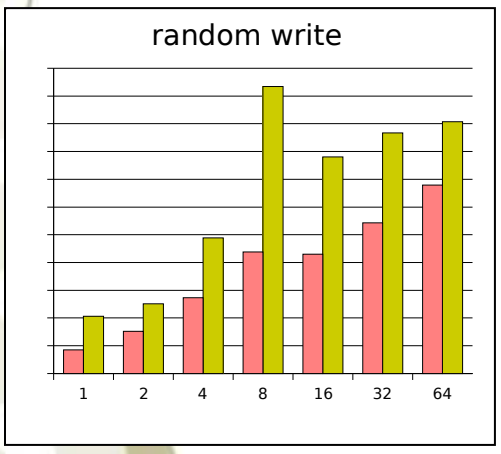
Performance - FCP versus FICON

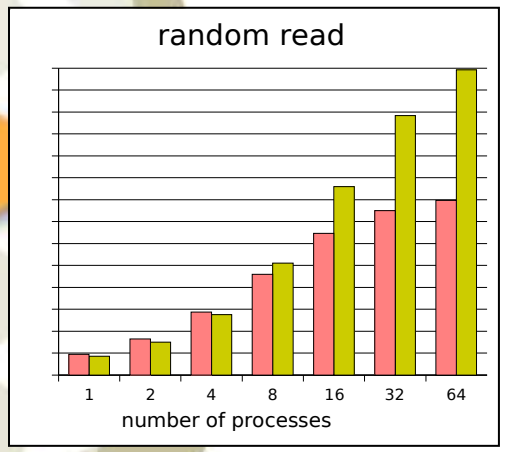


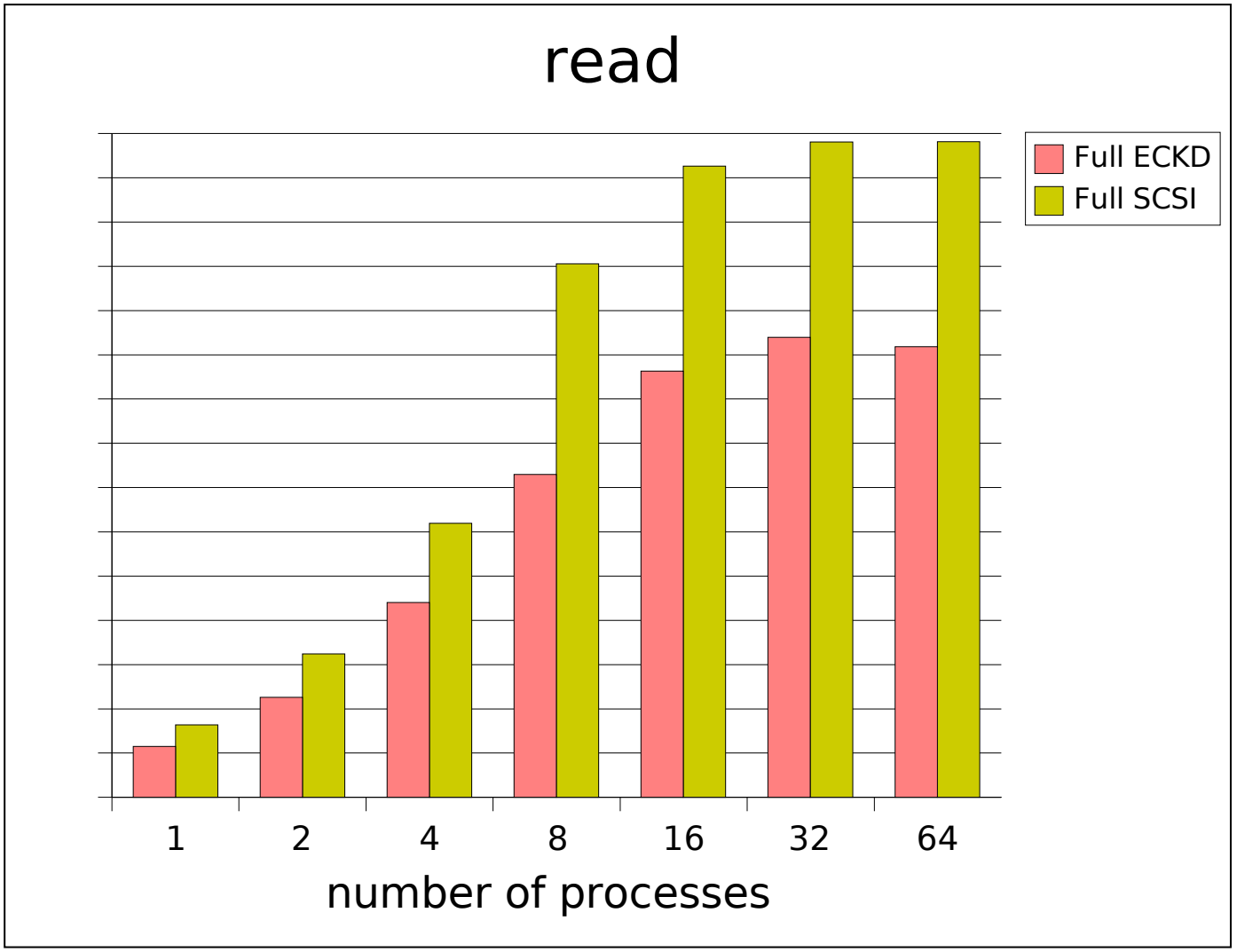
- **DS8300 (2107-92E)**
- **z990 LPAR (2084-B16)**
- **SUSE SLES9 SP2**
  - **8 CPUs**
  - **8 FICON / 8 FCP**
  - **256 MB**
- **Iozone 3.96**

FCP Performance - Throughput





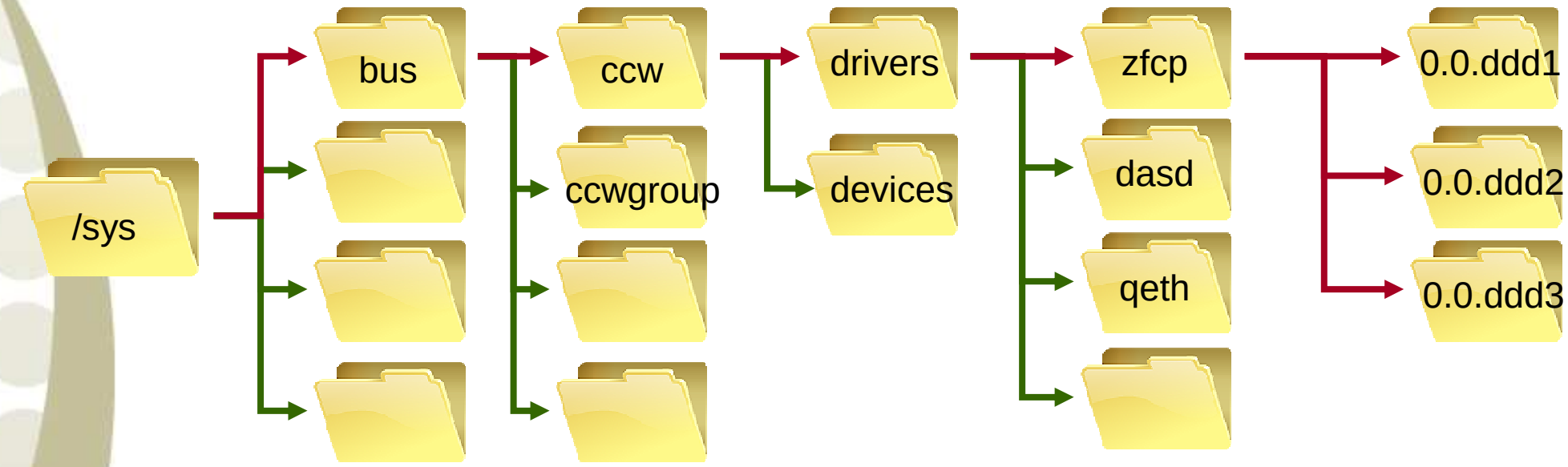




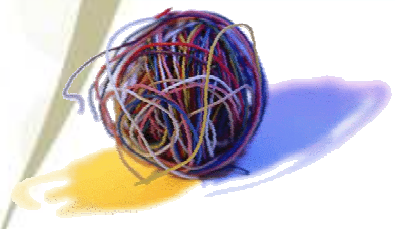
SHARE
Technology · Connections · Results

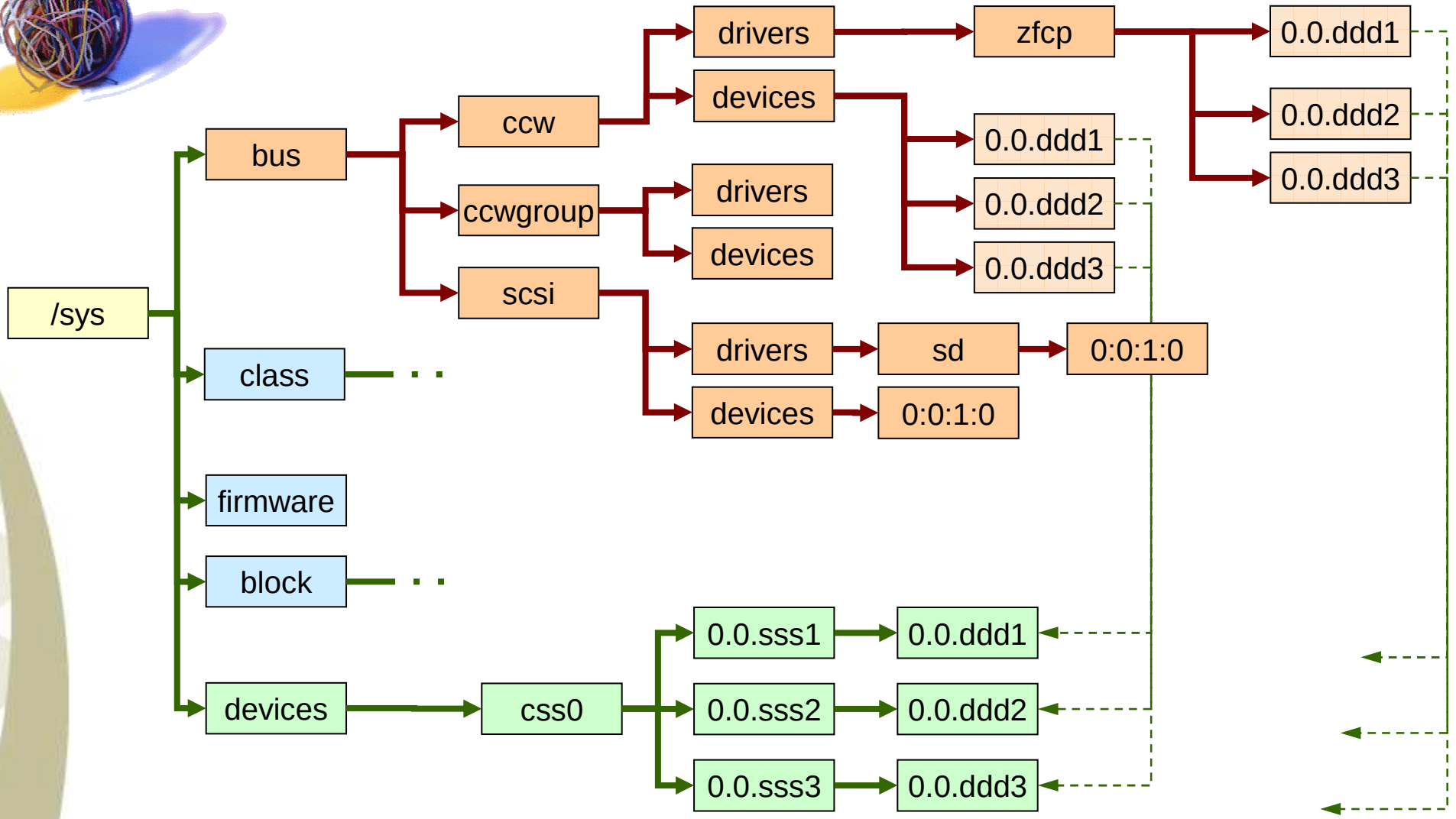
SysFS

- **Configuration file system (Linux kernel 2.6)**
- **Contains all device drivers and device specific information**
- **It is NOT a substitution of the /proc file system**
- **Used to configure device drivers**

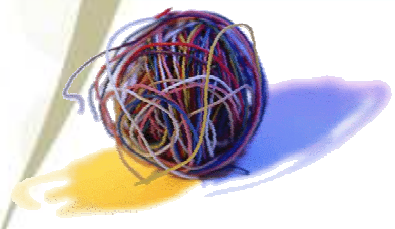


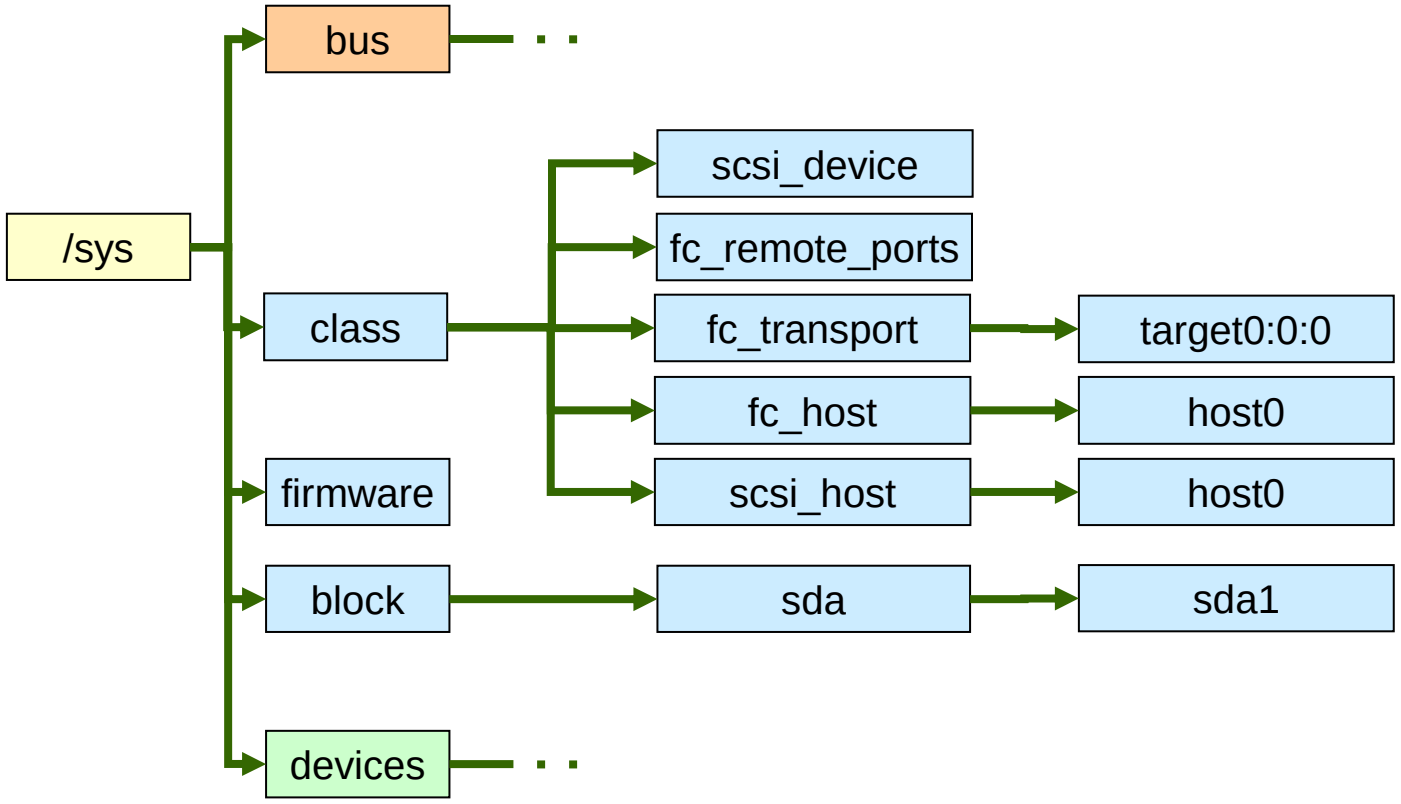
SysFS - Overview



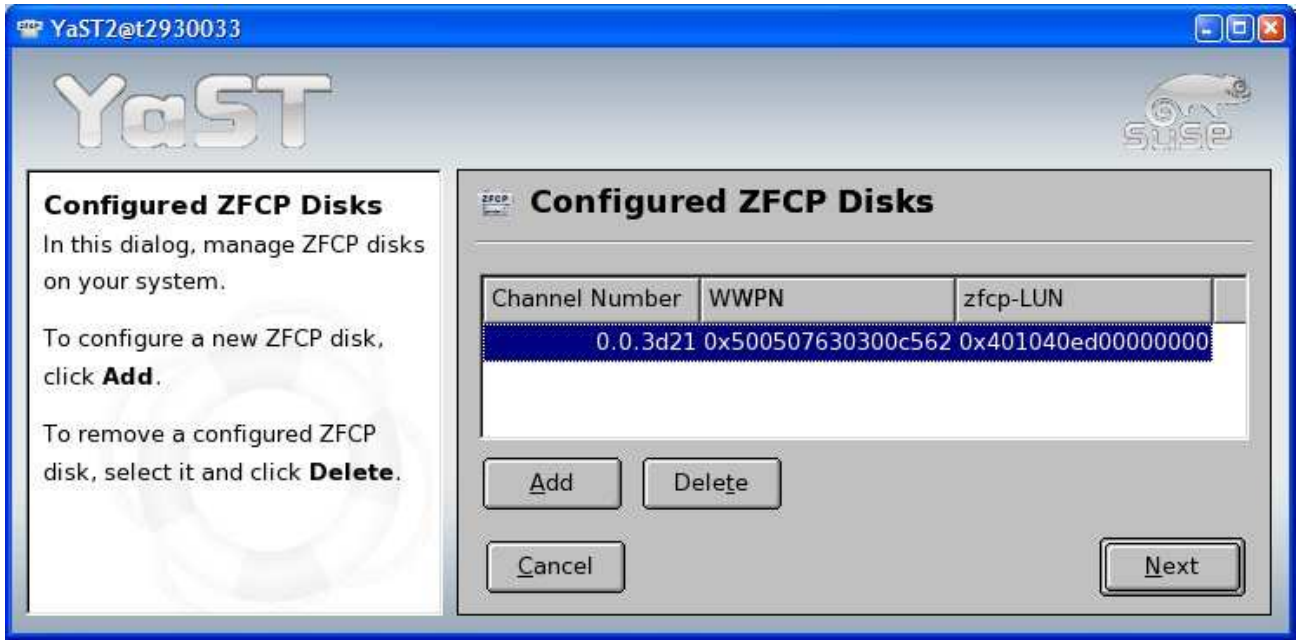


SysFS – Overview (cont)





zfcp Configuration



- **SUSE: yast**
  → hardware
  → zfcp

- **Manual zfcp configuration**

```
# cd /sys/bus/ccw/drivers/zfcp/0.0.5021/
# echo 1 > online    OR     0.0.5021 # chccwdev -e 0.0.5021
# echo 0x500507630303c562 > port_add
# echo 0x4011401600000000 > 0x500507630303c562/unit_add
```

zfcp Configuration – cont.

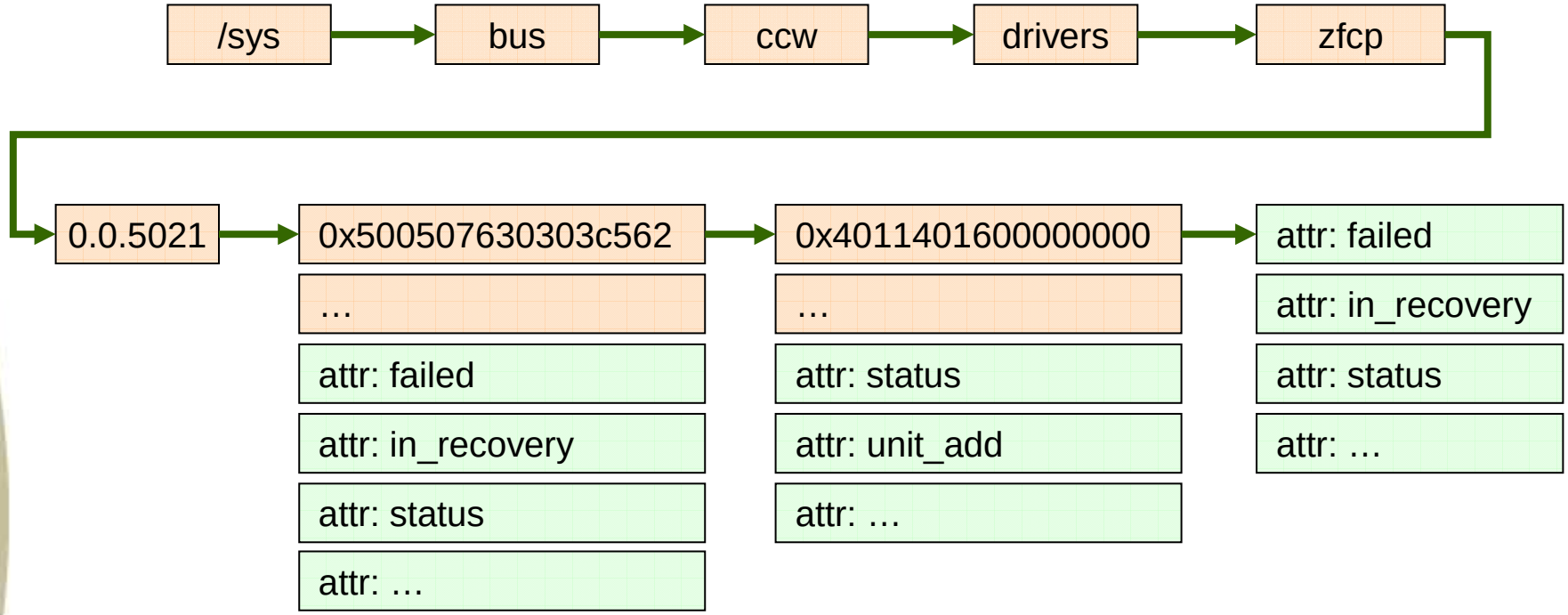
```
/var/log/messages
Jul 10 03:14:12 t2930033 kernel: scsi1 : zfcp
Jul 10 03:14:12 t2930033 kernel: zfcp: The adapter 0.0.5021 reported the following characteristics:
Jul 10 03:14:12 t2930033 kernel: WWNN 0x5005076400cd6aad, WWPN 0x5005076401008fa8, S_ID 0x00651213,
Jul 10 03:14:12 t2930033 kernel: adapter version 0x3, LIC version 0x605, FC link speed 2 Gb/s
Jul 10 03:14:12 t2930033 kernel: zfcp: Switched fabric fibrechannel network detected at adapter 5021
Jul 10 03:14:42 t2930033 kernel:   Vendor: IBM       Model: 2107900          Rev: .203
Jul 10 03:14:42 t2930033 kernel:   Type:   Direct-Access                    ANSI SCSI revision: 05
Jul 10 03:14:42 t2930033 kernel: SCSI device sdb: 104857600 512-byte hdwr sectors (53687 MB)
Jul 10 03:14:42 t2930033 kernel: SCSI device sdb: drive cache: write back
Jul 10 03:14:42 t2930033 kernel:  sdb: sdb1
Jul 10 03:14:42 t2930033 kernel: Attached scsi disk sdb at scsi1, channel 0, id 1, lun 0
Jul 10 03:14:42 t2930033 kernel: Attached scsi generic sg1 at scsi1, channel 0, id 1, lun 0,  type 0
Jul 10 03:14:42 t2930033 /etc/hotplug/block.agent[4105]: new block device /block/sdb
Jul 10 03:14:42 t2930033 /etc/hotplug/block.agent[4122]: new block device /block/sdb/sdb1
```

```
# lsscsi
[0:0:1:0]    disk    IBM      2107900          .203  /dev/sda
[1:0:1:0]    disk    IBM      2107900          .203  /dev/sdb
# mount /dev/sdb1 /mnt
# df
Filesystem           1K-blocks      Used Available Use% Mounted on
/dev/sda1              5156292   1554996   3339368  32% /
tmpfs                   253272         4    253268   1% /dev/shm
/dev/sdb1             51606124   1629436  47355252   4% /mnt
```

zfcp Configuration – cont.

`/sys/bus/ccw/drivers/zfcp/`

- **directory for each subchannel (virtual FCP adapter, e.g. 0.0.5021)**
  - **directory for each configured target port (e.g. 0x500507630303c562)**
    - **directory for each configured FCP LUN (e.g. 0x4011401600000000)**



zfcp Adapter Information

- <directory for each configured target port>
- peer_* - point-to-point attributes
- lic_version - LIC version number
- online - adapter status (online/offline)
- failed - adapter error recovery status
- in_recovery - recovery status
- status - adapter status (debug info)

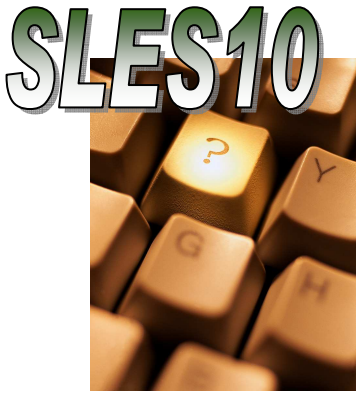
```
# cd /sys/bus/ccw/drivers/zfcp/0.0.170e
# cat lic_version
0x00000708
# cat status
0x5400082e

# lszfcp -H
0.0.170e host0
0.0.180e host1
```

zfcp Port Information

- <directory for each FCP LUN>
- access_* - access control
- failed - port error recovery status
- in_recovery - recovery status
- status - port Status (debug info)

```
# cd /sys/bus/ccw/drivers/zfcp/0.0.170e/0x500507630300c562
# ls
  0x401040ab00000000  access_denied  failed  in_recovery  status  uevent
  unit_add  unit_remove
# cat in_recovery
0
```

```
# lszfcp -P
0.0.170e/0x500507630300c562 rport-0:0-0
0.0.170e/0x500507630303c562 rport-0:0-1
0.0.180e/0x500507630303c562 rport-1:0-0
0.0.180e/0x500507630300c562 rport-1:0-1
```



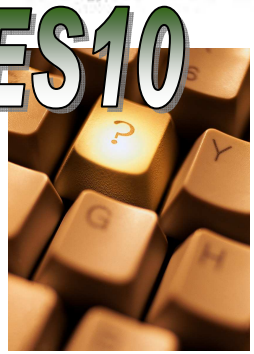
zfcp Unit Information

- **access_*** - access control
- **failed** - unit error recovery status
- **in_recovery** - recovery status
- **status** - unit status (debug info)

```
# # cd /sys/bus/ccw/drivers/zfcp/0.0.170e/0x500507630300c562/0x401040ab00000000
# ls
  access_denied  access_readonly  access_shared  failed  in_recovery  status  uevent
# cat failed
0
# cat in_recovery
0
# cat status
0x54000010
```

```
# lszfcp -D
0.0.170e/0x500507630300c562/0x401040ab00000000 0:0:0:0
0.0.170e/0x500507630303c562/0x401040ab00000000 0:0:1:0
0.0.180e/0x500507630303c562/0x401040ab00000000 1:0:0:0
0.0.180e/0x500507630300c562/0x401040ab00000000 1:0:1:0
```

Fibre Channel transport class




```
# cd /sys/class/fc_transport/
# ls
target0:0:0  target0:0:1  target1:0:0  target1:0:1
```

```
# cd target0:0:0
# ls
device    node_name
port_id   port_name
uevent
# cat node_name
0x5005076303ffc562
# cat port_id
0x652113
# cat port_name
0x500507630300c562
```

```
# cd device/0:0:0:0
# ls
block:sda         ioerr_cnt      scsi_generic:sg0
bus               iorequest_cnt  scsi_level
delete            model          state
device_blocked    queue_depth    timeout
driver            queue_type     type
fcp_lun           rescan         uevent
generic           retries        vendor
hba_id            rev            wwpn
iocounterbits     scsi_device:0:0:0:0
iodone_cnt        scsi_disk:0:0:0:0
```

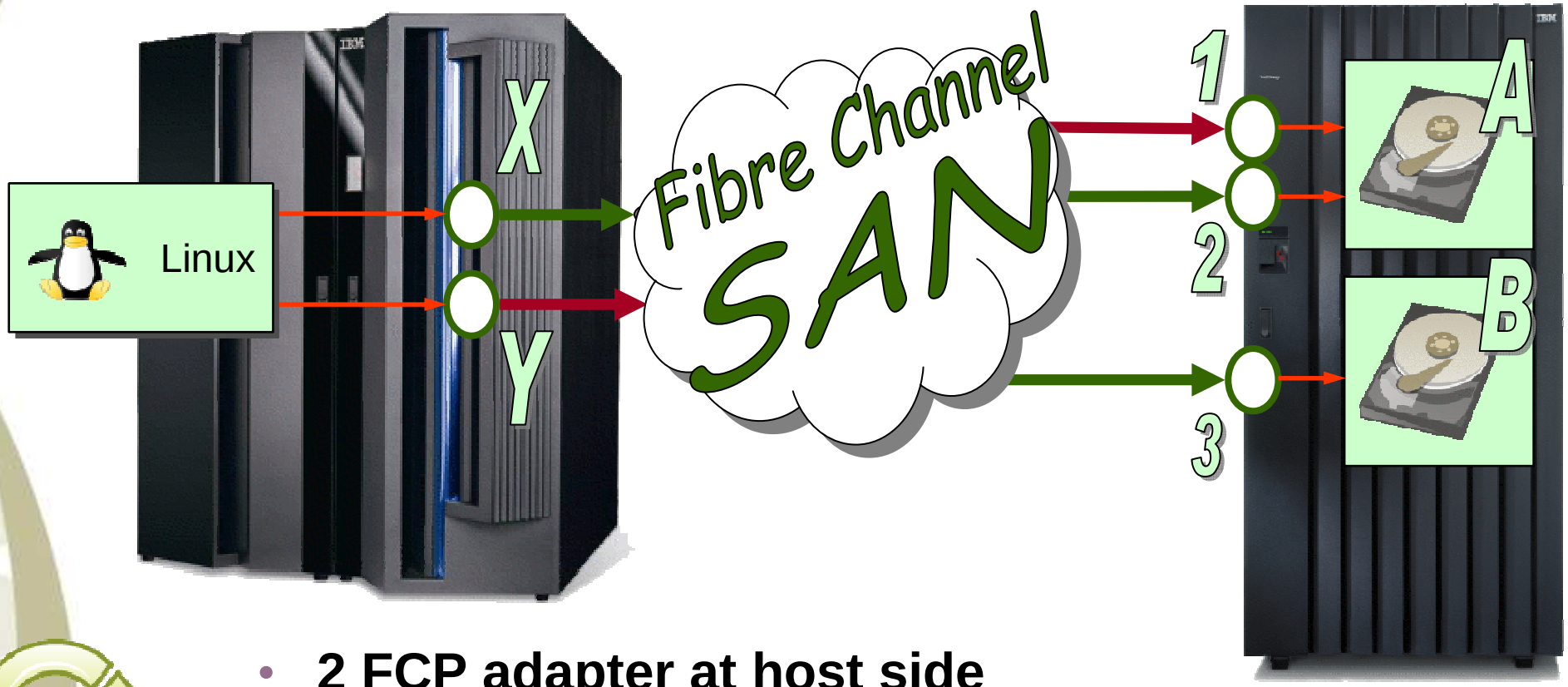
FCP Multipathing

- **“Failover” on path-failure and “failback”**
- **Load balancing**

MP-Tools
LVM
EVMS
MD
LVM2

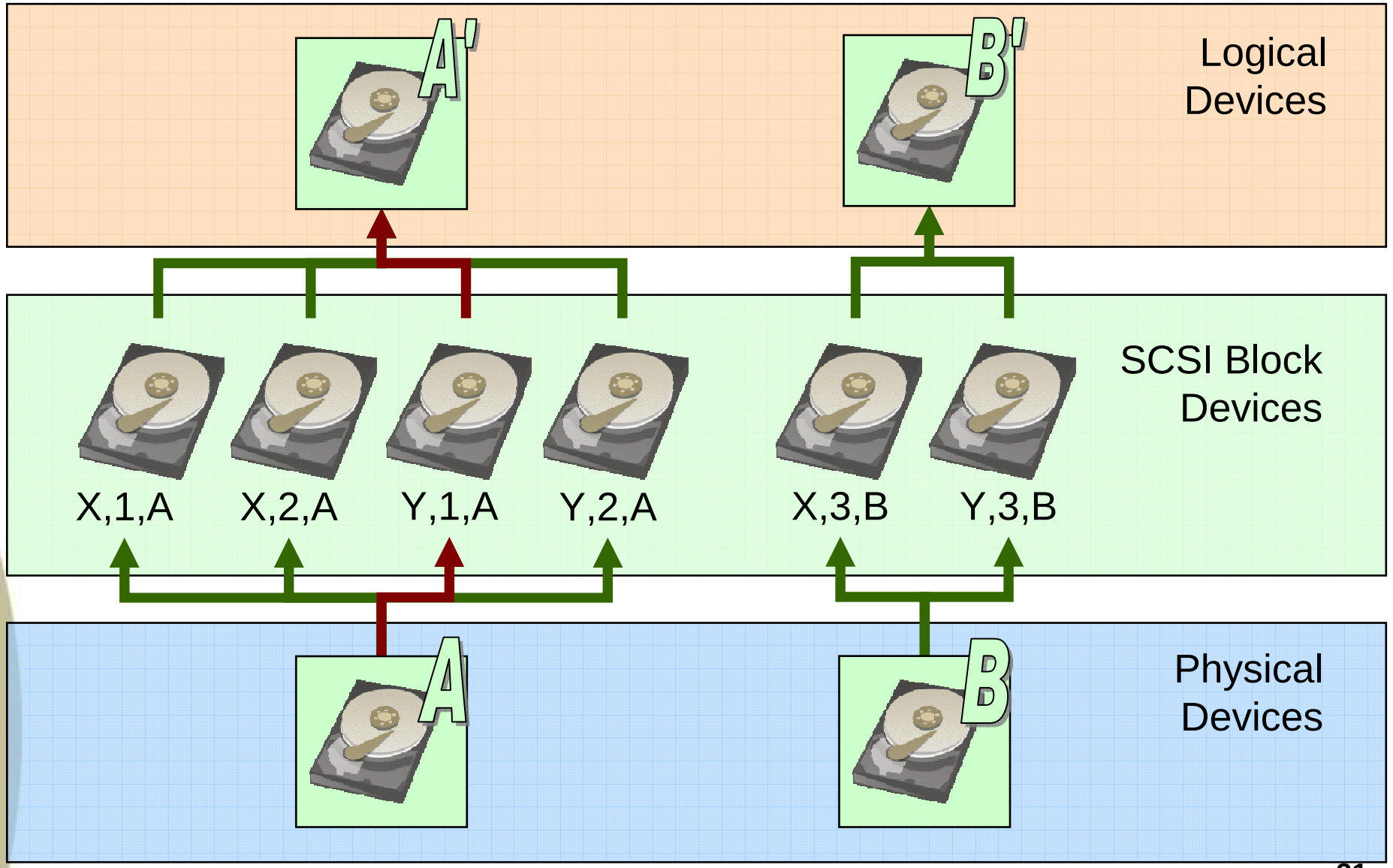
- **LVM – Logical Volume Manager**
- **Device Mapper subsystem in 2.6 kernel**
  - **EVMS – Enterprise Volume Management System**
  - **LVM2 – Logical Volume Manager 2**
  - **MP-Tools – Multipath-Tools**
- **MD – Multiple devices**

FCP Multipathing



- 2 FCP adapter at host side
- 3 FCP adapter at storage side
- 4 paths to disk A and 2 paths to disk B

FCP Multipathing – Devices



Multipath-Tools Package

```
# multipath -ll
36005076303ffc56200000000000010abdm-0 IBM,2107900
[size=5.0G][features=1
   queue_if_no_path][hwhandler=0]
\_ round-robin 0 [prio=4][active]
 \_ 0:0:0:0     sda   8:0   [active][ready]
 \_ 1:0:0:0     sdb   8:16  [active][ready]
 \_ 0:0:1:0     sdc   8:32  [active][ready]
 \_ 1:0:1:0     sdd   8:48  [active][ready]
```

- Developed by Christophe Varoqui
- Link: http://christophe.varoqui.free.fr/wiki/wakka.php?wiki=Home
- RedHat: device-mapper-multipath
- SUSE: multipath-tools
- Development ongoing

```
# cat /etc/multipath.conf
…
blacklist {
        devnode "^(ram|raw|loop|fd|md|dm-|sr|scd|st)[0-9]*"
        devnode "^hd[a-z][[0-9]*]"
        devnode "^cciss!c[0-9]d[0-9]*[p[0-9]*]"
        devnode "^dasd[a-z]+[0-9]*"
}
…
```


```
# ls -l /dev/mapper/
total 0
brw------- 1 root root 253,  0 Aug 13 13:15 36005076303ffc56200000000000010ab
brw------- 1 root root 253,  1 Aug 13 13:15 36005076303ffc56200000000000010ab-part1
lrwxrwxrwx 1 root root      16 Aug 13 13:14 control -> ../device-mapper
```

SCSI Tape

- **Several possibilities to access SCSI tapes**
  - **st - SCSI tape device driver**
  - **sg - SCSI generic device driver**
  - **IBMtape device driver (object code only)**
  - **lin_tape device driver (open source version of IBMtape)**

- **IBMtape/lin_tape**
  - **Designed to use the features provided by the IBM tape drives**
  - **Basic tape operations (e.g. backup and restore)**
  - **Medium changer operations (e.g. mount and demount the cartridges)**
  - **Advanced functions for full tape management systems**
  - **Multipathing (2 paths)**
  - **Powerfull IBMtapeutil package**

lin_tape Device Driver

```
# ls -l /dev/IBM*
crw-rw-rw-  1 root root 254, 512 Aug 13 15:51 /dev/IBMchanger0
crw-r--r--  1 root root 254, 767 Aug  9 16:45 /dev/IBMtape
crw-rw-rw-  1 root root 254,   0 Aug 10 09:30 /dev/IBMtape0
crw-rw-rw-  1 root root 254, 256 Aug 10 09:30 /dev/IBMtape0n
```

```
# cat /proc/scsi/scsi
Host: scsi2 Channel: 00 Id: 01 Lun: 00
  Vendor: IBM Model: 03590H11 Rev: F29C
  Type: Sequential-Access ANSI SCSI revision: 03
Host: scsi2 Channel: 00 Id: 01 Lun: 01
  Vendor: IBM Model: 03590H11 Rev: F29C
  Type: Medium Changer    ANSI SCSI revision: 03
```

```
# cat /proc/scsi/IBMtape
lin_tape version: 1.6.0
lin_tape major number: 254
Attached Tape Devices:
Nr model    SN           HBA  …
0  03590H11 0000000CB722 zfcp …
```

```
# ls -l /sys/class/lin_tape/
drwxr-xr-x  2 root root 0 Aug 13 15:51 IBMchanger0
drwxr-xr-x  2 root root 0 Aug 13 15:56 IBMtape0
drwxr-xr-x  2 root root 0 Aug 10 09:30 IBMtape0n
```

```
# ls -l /etc/lin_taped.conf
-rw-r--r--  1 root root 2386 May 29 20:10 /etc/lin_taped.conf
```

```
# ps -ef | grep lin_taped
root  1192     1 0 Aug09 ?     00:00:00 /usr/bin/lin_taped start
```

IBMtapeutil Package

```
# IBMtapeutil -f /dev/IBMtape0 tur
Issuing test unit ready...
Unit ready.

# IBMtapeutil -f /dev/IBMtape0 status
Generic ANSI SCSI-2 tape drive
File number=-1, block number=0, partition=0.
Tape block size 0 bytes. Density code 0x2c
Soft error count since last status=0
General status bits on (41000000):
 BOT ONLINE

# IBMtapeutil -f /dev/IBMchanger0 elementinfo
Getting element information...

  Number of Robots ..... 1
  First Robot Address ..... 0
  Number of Slots ..... 10
  First Slot Address ..... 32
  Number of Import/Exports ..... 1
  First Import/Export Address ... 31
  Number of Drives ..... 1
  First Drive Address ..... 16
```

- General Subcommands:
  - **tur, inquiry, print, reserve, release, reqsense, qryversion, logpage, modepage, qrypath, enablepath, disablepath**
- Tape Subcommands:
  - **bsf, bsr, eof, fsf, fsr, weof, fsfm, bsfm, asf, compress, tell, seek, nocompress, rewind, sync, load, erase, display, unload, retension, read, qrypos, seod, write, setpos, status, rtest, offline, parms, wtest, rewoffl, list, rwtest, prevent, lock, setblk, allow, unlock, density, qryinquiry, qrysense, append, mtdevice, encryption**
- Medium Changer Subcommands:
  - **allow, prevent, audit, inventory, mount, position, elementinfo, unmount, move, devids, exchange**
- Service Aid Subcommands:
  - **dump, forcedump, ucode, resetdrive**

SAN Discovery Tool

- **Identification of SAN resources**
  - List of host adapters, ports, units
- **Helpful to uncover configuration problems**
  - E.g. zoning or LUN masking problems
- **Does not configure zfcp automatically**

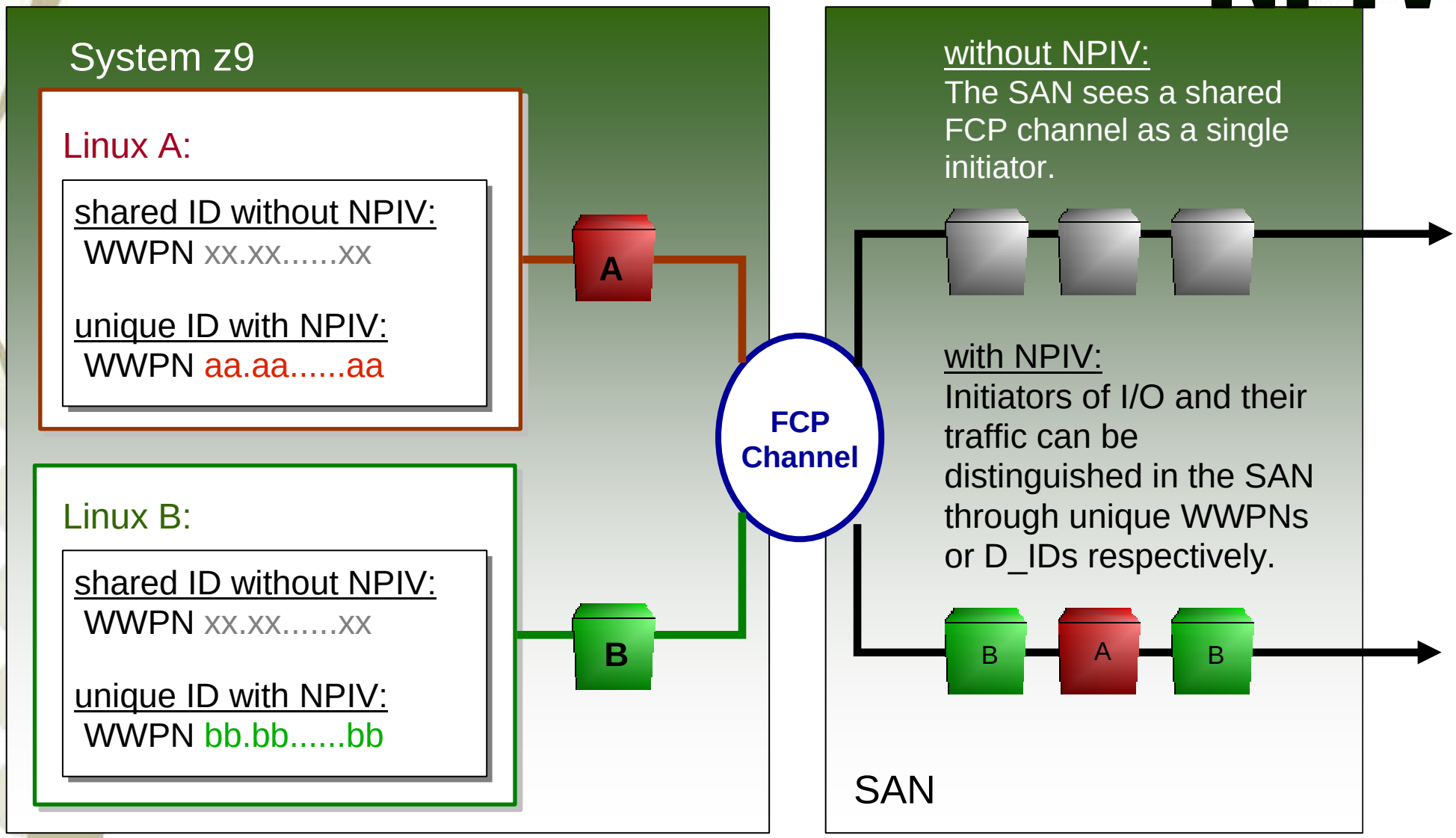
```
# san_disc -c PORT_LIST -a 1
 #  Port WWN           Node WWN           DID      Type
  1 0x500507640140863c 0x5005076400cd6aad 0x650613 N_Port
  2 0x50050764010087ef 0x5005076400cd6aad 0x650713 N_Port
    ...
 97 0x500507640140863c 0x5005076400cd6abd 0x650613 N_Port
```

Port list

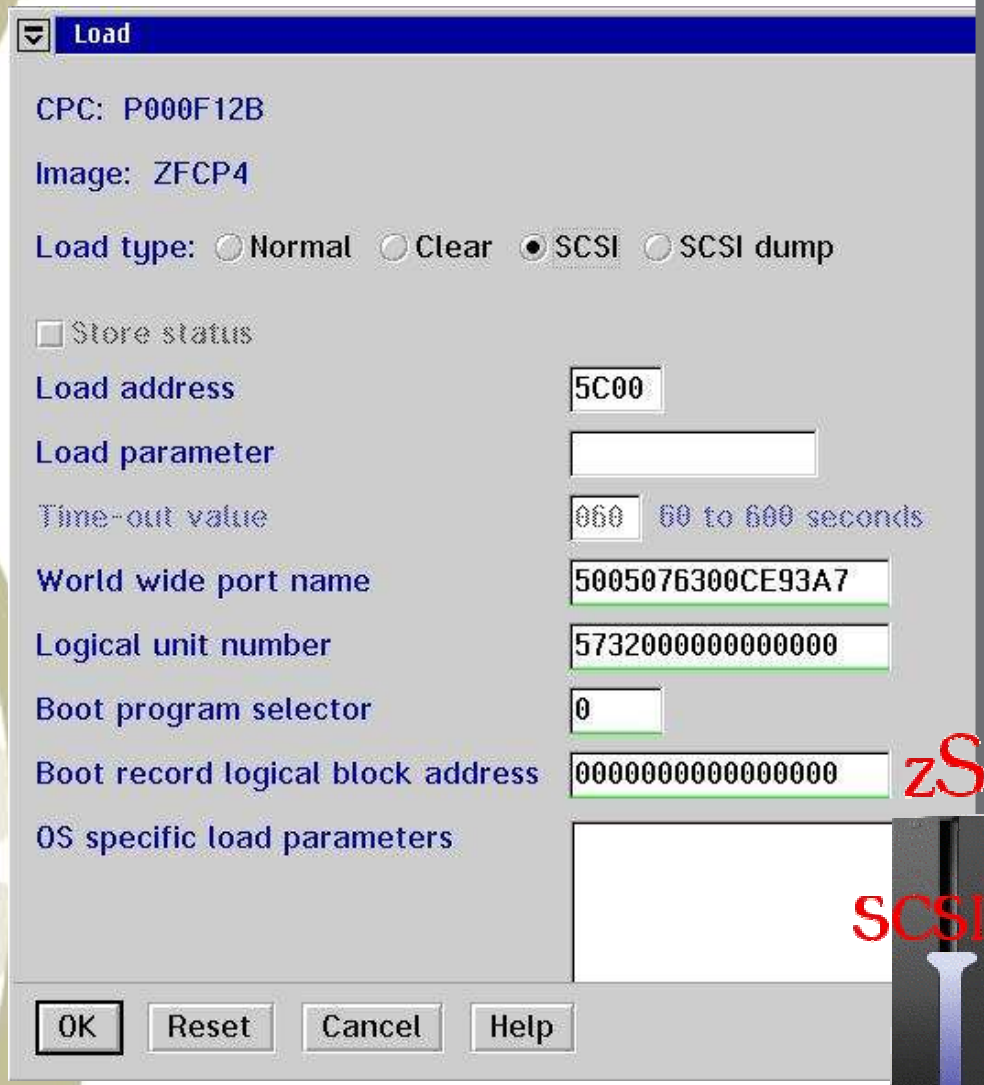
```
# san_disc -c REPORT_LUNS -a 1 -p 0x500507640140863c
Number of LUNs: 97
#    LUN
   1 0x4010400000000000
   2 0x4010400100000000
     ...
  97 0x4010406000000000
```

LUN list

NPIV – N-Port ID Virtualization



SCSI IPL & SCSI Dump



- **IPL from SCSI disks**
- **Dump to SCSI disks (LPAR only).**
- **SCSI IPL expand the set of IPL'able devices**
- **SCSI disks as Linux root file system possible**
- **New set of IPL parameters.**
- **LPAR and z/VM guests supported.**

Requirements

- **z800, z890, z900, z990, z9**
- **Requires enablement by FC9904**
- **FCP Channels**
- **FC attached SCSI Disks**

SCSI IPL – example z/VM

```
set loaddev port 50050763 0300C562 lun 401040EE 00000000
Ready; T=0.01/0.01 22:11:01

query loaddev
PORTNAME 50050763 0300C562    LUN  401040EE 00000000    BOOTPROG 0
BR_LBA   00000000 00000000
Ready; T=0.01/0.01 22:11:06

i 5021
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.
00: MLOEVL012I: Machine loader up and running (version 0.18).
00: MLOPDM003I: Machine loader finished, moving data to final storage location.
Linux version 2.6.16-18.x.20060403-s390xdefault (wirbser@t2944002) (gcc version
4.1.0) #1 SMP PREEMPT Mon Apr 3 09:56:54 CEST 2006
We are running under VM (64 bit mode)
Detected 4 CPU's
Boot cpu address  0
Built 1 zonelists
Kernel command line: dasd=e960-e962 root=/dev/sda1 ro noinitrd
  zfcp.device=0.0.3d21,0x500507630300c562,0x401040ee00000000
```

Summary

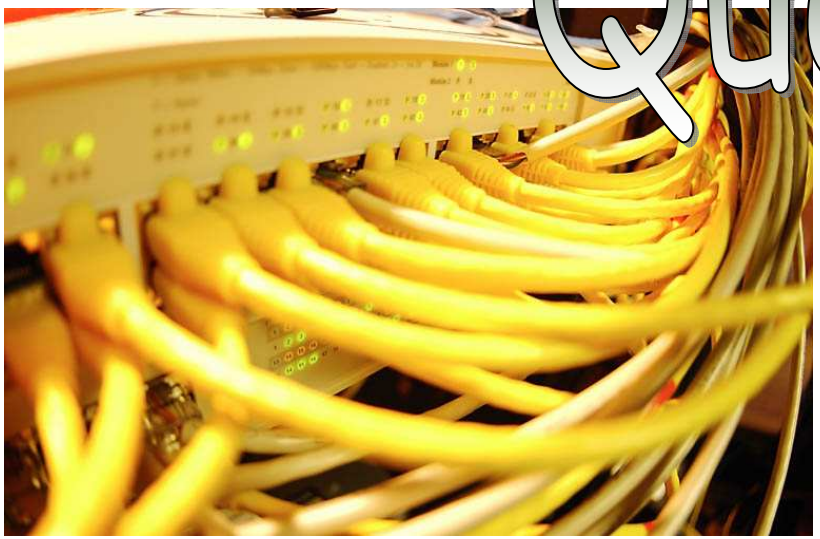
- **FCP/SCSI support for IBM System z**
- **Integration of your System z into standard based FC SANs**
- **FCP channel based on FICON / FICON Express adapter cards**
- **Three addressing parameters instead of one**
- **Many information hidden in sysfs**
- **Performance and other advantages compared to ECKD**
- **If possible use multipathing**
- **If possible use NPIV**
- **Helpful SAN discovery tool**

- **Give it a try, you probably have all you need.**

Useful Links

- I/O Connectivity on System z
  - http://www-03.ibm.com/systems/z/connectivity/
- Getting Started with zSeries Fiber Channel Protocol, IBM Redpaper
  - http://www.redbooks.ibm.com/redpapers/pdfs/redp0205.pdf
- z/VM
  - Version 5.3: http://www.vm.ibm.com/zvm530/
- Introducing N_Port Identifier Virtualization for IBM System z9 (Redpaper)
  - http://www.redbooks.ibm.com/redpapers/pdfs/redp4125.pdf
- How to use FC-attached SCSI devices with Linux on System z
  - http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux390/docu/l26cts00.pdf
- Linux for IBM System z
  - http://www-128.ibm.com/developerworks/linux/linux390/
- Linux for IBM System z Device Drivers Book and other documentation
  - http://www-128.ibm.com/developerworks/linux/linux390/april2004_documentation.html (SLES9)
  - http://www-128.ibm.com/developerworks/linux/linux390/october2005_documentation.html (SLES10)
- IBM TotalStorage Tape Device Drivers – Installation and User‘s Guide
  - ftp://ftp.software.ibm.com/storage/devdrvr/Doc/
- IBM disk systems
  - http://www-03.ibm.com/servers/storage/disk/
- linuxvm.org
  - http://www.linuxvm.org/

SHARE
Technology · Connections · Results



Questions?

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries:

Enterprise Storage Server, IBM*, IBM logo*, System z9*, IBM eServer, z/VM, zSeries

*Registered trademarks of IBM Corporation

Linux is a registered trademark of Linus Torvalds.

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

Making Your Penguins Fly
Introduction to SCSI over FCP for Linux on System z
Horst Hummel