Making Your Penguins Fly

Introduction to SCSI over FCP for Linux on System z

Christian Borntraeger (cborntra@de.ibm.com)
Linux on zSeries Development
IBM Lab Boeblingen, Germany
Session 9259, Tue Feb 13
Making Your Penguins Fly – Flight Schedule

- SAN & SAN integration
- Addressing basics
- Requirements
- zfcp device driver
- Why FCP?
- Configuration
- Multipathing
- NPIV
- SAN Discovery Tool
- SCSI IPL/Dump
System z in a SAN – Sharing Storage Resources
WWPNs and other Ports

- Many WWPNs in a FC SAN
- Only target WWPN is important
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
zSeries 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
Software Requirements

- SCSI (IPL) with z/VM
  - z/VM Version 4.4 (PTF UM30989) or newer
  - z/VM Version 5.2 (current version)

- SUSE Linux Enterprise Server 8 (SLES8)
  - Service Pack 4

- SUSE Linux Enterprise Server 9 (SLES9)
  - Service Pack 3

- SUSE Linux Enterprise Server 10 (SLES10)
  - Available

- Red Hat Enterprise Linux 3 (RHEL3)
  - Update 8

- Red Hat Enterprise Linux 4 (RHEL4)
  - Update 4

- Red Hat Enterprise Linux 5 (RHEL5)
  - Outlook 2007
**Introduction to SCSI over FCP for Linux on System z**

```
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
```

**IOCDS – FCP Configuration Sample**
Linux SCSI Stack

- VFS / File System
- Buffer Cache
- Multipathing (LVM, EVMS, MD)
- SCSI Disk
- Uniform CD-ROM
- SCSI CD-ROM
- SCSI Tape
- SCSI Generic
- SCSI Mid Layer
- HBA Driver A
- HBA Driver B
- zfcp

© Copyright IBM Corporation 2006,2007
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 zSeries 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 zSeries FCP adapter.
  - returns replies and data from SCSI devices to the Linux SCSI subsystem.
zSeries in a SAN – Topologies

point-to-point

direct attached arbitrated loop

switched fabric

supported

NOT supported

Tape

Tape

Disk

DVD

Tape

Disk

CD
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
  - FCP is much faster than FICON
  - Asynchronous I/O
  - No ECKD emulation overhead
- No disk size restrictions
- Up to 16 partitions
- Get rid of FICON topology constraints, FCP is much more flexible.

- 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
- Requires slightly more CPU than FICON
- SAN access control mechanisms only usable with NPIV (z9 only)
# Disk Usage – ECKD and SCSI Comparison

<table>
<thead>
<tr>
<th>Configuration</th>
<th>ECKD DASD</th>
<th>SCSI Disk</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IOCDS/VM (operator)</td>
<td>IOCDS/VM &amp; Linux (operator &amp; Linux admin)</td>
</tr>
<tr>
<td>Access Method</td>
<td>SSCH/CCW</td>
<td>QDIO</td>
</tr>
<tr>
<td>Block Size (Byte)</td>
<td>512, 1K, 2K, 4K</td>
<td>512</td>
</tr>
<tr>
<td>Disk Size</td>
<td>3390 Model 3/9</td>
<td>any</td>
</tr>
<tr>
<td>Formatting (low level)</td>
<td>dasdfmt</td>
<td>not necessary</td>
</tr>
<tr>
<td>Partitioning</td>
<td>fdasd</td>
<td>fdisk</td>
</tr>
<tr>
<td>File System</td>
<td>mke2fs (or others)</td>
<td></td>
</tr>
<tr>
<td>Access</td>
<td>Mount</td>
<td></td>
</tr>
</tbody>
</table>
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)
- McDATA 3232 (IBM 2031-232)
- McDATA Sphereon 4500 Fabric Switch (IBM 2031-224)
- IBM TotalStorage SAN Switch (2109-F32)
- IBM TotalStorage SAN32B-2 (2005-B32)
Performance - FCP versus FICON

Throughput for random readers

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

![Graphs showing throughput for different operations and process numbers](image-url)
FCP – SCSI Mapping
New file system with 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

Introduction to SCSI over FCP for Linux on System z

SysFS

/bus
/class
/firmware
/block
/devices
/ccw
/ccwgroup
/scsi
/drivers
/devices
/zfcp

0.0.ddd1
0.0.ddd2
0.0.ddd3

0.0.sss1
0.0.sss2
0.0.sss3

0.0.ddd1
0.0.ddd2
0.0.ddd3

0:0:1:0

© Copyright IBM Corporation 2006, 2007
zfcp Configuration

- SUSE: yast
  - hardware
  - zfcp

- Manual zfcp configuration

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

/var/log/messages
Jul 10 03:14:12 t2930033 kernel: scsil : 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 scsil, channel 0, id 1, lun 0
Jul 10 03:14:42 t2930033 kernel: Attached scsi generic sg1 at scsil, 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)
```
Adapter Information

- `<directory for each configured target port>`
- `serial_number` - Adapter serial number
- `lic_version` - LIC version number
- `scsi_host_no` - SCSI host number
- `wwnn` - Worldwide node name
- `wwpn` - Worldwide port name
- `fc_topology` - Fiber Channel topology
- `fc_link_speed` - Link Speed

```bash
# cd /sys/bus/ccw/drivers/zfcp/0.0.3d21/
# cat serial_number
IBM020000000D6AAD
# cat lic_version
0x00000605
# cat scsi_host_no
0x0
```

```bash
# cat wwnn
0x5005076400cd6aad
# cat wwpn
0x5005076401c08f98
# cat fc_topology
fabric
# cat fc_link_speed
2 Gb/s
```
Port Information

- <directory for each FCP LUN>
- d_id - Destination ID
- failed - Port error recovery status
- in_recovery - Recovery status
- scsi_id - SCSI ID
- wwnn - Worldwide node name

```bash
# cd /sys/bus/ccw/drivers/zfcp/0.0.3d21/0x500507630300c562/
# ls
  0x401040ed00000000 d_id failed scsi_id unit_add wwnn access_denied
  detach_state in_recovery status unit_remove
# cat in_recovery
  0
# cat scsi_id
  0x1
# cat d_id
  0x632e13
```
Unit Information

- Access Control
- Unit error recovery status
- Recovery status
- Linux SCSI LUN
- Unit status (debug info)

```bash
# cd /sys/bus/ccw/drivers/zfcp/0.0.3d21/0x500507630300c562/0x401040ed00000000/
# ls
  access_denied  access_readonly  access_shared  detach_state  failed
  in_recovery  scsi_lun  status
# cat failed
0
# cat in_recovery
0
# cat scsi_lun
0x0
# cat status
0x54000000
```
FCP Multipathing

- “Failover” on path-failure and “failback”
- Load balancing
- Covers all block devices

- 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

Logical Devices

SCSI Block Devices

Physical Devices
Multipath-Tools Package

# multipath -l
mpath0 (36005076303fffc62000000000000010ed)
[size=5 GB][features="1
  queue_if_no_path"] [hwhandler="0"]
  round-robin 0 [active]
  0:0:1:0    sda  8:0    [active]
  1:0:1:0    sdb  8:16   [active]
  0:0:2:0    sdc  8:32   [active]
  1:0:2:0    sdd  8:48   [active]

mpath1 (IBM.75000000092461.2a00.1a)
[size=2 GB][features="0"] [hwhandler="0"]
  round-robin 0 [active]
  0:0:10778:0 dasdd 94:12 [active]
  0:0:10927:0 dasde 94:16 [active]
  0:0:10928:0 dasdf 94:20 [active]
  0:0:10929:0 dasdg 94:24 [active]

# ls -l /dev/mapper/
total 0
  crw------- 1 root root 10, 63 Jun 27 09:11 control
  brw-rw---- 1 root disk 253, 4 Jun 28 07:51 mpath0
  brw-rw---- 1 root disk 253, 5 Jun 28 07:51 mpath0p1
  brw-rw---- 1 root disk 253, 0 Jun 27 10:05 mpath1
  brw-rw---- 1 root disk 253, 3 Jun 27 10:05 mpath1p1

- Developed by Christophe Varoqui
- RedHat: device-mapper-multipath
- SUSE: multipath-tools
- Development ongoing
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

System z9

Linux A:

shared ID without NPIV: WWPN xx.xx......xx

unique ID with NPIV: WWPN aa.aa......aa

Linux B:

shared ID without NPIV: WWPN xx.xx......xx

unique ID with NPIV: WWPN bb.bb......bb

Unique SAN Identities!

without NPIV:
The SAN sees a shared FCP channel as a single initiator.

with NPIV:
Initiators of I/O and their traffic can be distinguished in the SAN through unique WWPNs or D_IDs respectively.
Introduction to SCSI over FCP for Linux on System z

SCSI IPL & SCSI Dump

- IPL from SCSI disks
- Dump to SCSI disks (LPAR only).

- SCSI disks 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

```
Ready; T=0.01/0.01 22:09:48
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: HCPLD12816I Acquiring the machine loader from the processor controller.
00: HCPLD12817I Load completed from the processor controller.
00: HCPLD12817I Now starting the machine loader.
00: MLOEVL012I: Machine loader up and running (version 0.18).
00: MLOPDMD003I: 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/sdal ro noinitrd
zfcp.device=0.0.3d21,0x500507630300c562,0x401040ee00000000
```
Summary

- FCP/SCSI support for IBM System z.
  - FCP channel based on FICON / FICON Express adapter cards.
  - FCP channel support in z/VM 4.3 and higher for Linux guests.
  - First FCP/SCSI exploitation for System z in SLES8 and RHEL3.
- Integration of your System z into standard based FC SANs.
- New device types.
- Three addressing parameters instead of one
- Performance and other advantages compared to ECKD
- Without NPIV: No LUN sharing or zoning on a single adapter → use separate physical adapters
- With NPIV: SAN access control mechanisms (z9 only)
- Helpful SAN Discovery Tool
Useful Links

- I/O Connectivity on IBM zSeries mainframe servers
  - http://www-03.ibm.com/systems/z/connectivity/
- Getting Started with zSeries Fiber Channel Protocol, IBM Redpaper
- Introducing N_Portal Identifier Virtualization for IBM System z9 (Redpaper)
- How to use FC-attached SCSI devices with Linux on System z
- Linux for IBM System z
- Linux for IBM System z Device Drivers Book and other documentation
- IBM TotalStorage Tape Device Drivers – Installation and User’s Guide
- IBM disk systems
  - http://www-03.ibm.com/servers/storage/disk/
- linuxvm.org
  - http://www.linuxvm.org/
Questions?
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.