Steffen Maier – Linux on System z Development
2014-09-24/25

FCP with Linux on System z: SCSI over Fibre Channel Best Practices

© Copyright IBM Corp. 2014

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.

© Copyright IBM Corp. 2014

Agenda

- Introduction and Terminology
- Setup
  - I/O Definition
  - N_Port ID Virtualization (NPIV)
  - Zoning
  - LUN-Masking
  - Multipathing
  - LUN Management with ZFCP
- IPL (booting) over FCP

© Copyright IBM Corp. 2014

Introduction and Terminology

© Copyright IBM Corp. 2014

FCP in a nutshell

- Storage Area Networks (SANs) are specialized networks dedicated to the transport of mass storage data (block/object oriented)
- Today the most common SAN technology used is Fibre Channel (FC) [T11]
- The Fibre Channel (FC) standard was developed by the InterNational Committee for Information Technology Standards (INCITS)
- Over this FC transport, using the Fibre Channel Protocol (FCP) as encapsulation, the SCSI protocol is used to address and transfer raw data between server and storage device [T10]
- Each server and storage is equipped with a least two adapters which provide a redundant physical connection to a redundant SAN
- For System z any supported FCP adapter, such as FICON Express can be used for this purpose.
  - Latest adapter cards are: FICON Express8 and FICON Express8S

Throughout presentation, all royal blue text fragments are clickable hyperlinks!

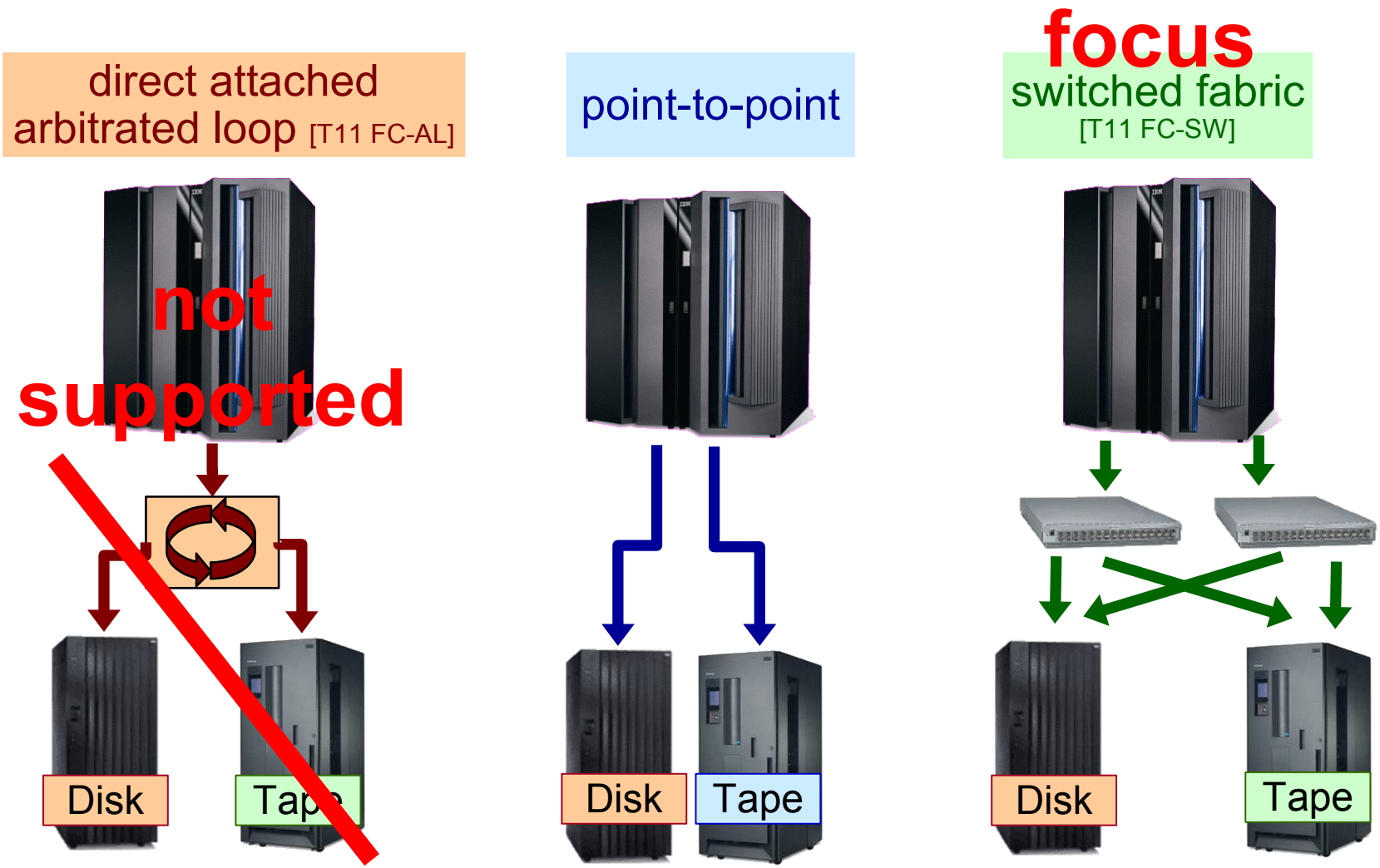
© Copyright IBM Corp. 2014

FCP Compared to Channel I/O

	FCP	Channel I/O
OS	**multipathing** handled in operating systems	multipathing handled in System z firmware
	port and LUN attachment handled in operating systems	port attachment handled in System z I/O configuration
fabric	FCP device represents **virtual adapter** to the Fibre Channel SAN	DASD device represents **disk** volume (ECKD)
	FCP device defined in System z I/O configuration → add new storage without IOCDS change	**disk** defined in System z I/O configuration
	both use **existing FC SAN**: FICON Express cards, switches, cabling, storage subsystems	
	additional **configuration** beyond System z: • Zoning in the SAN fabric switches • LUN masking on the storage server	Switch configuration via System z I/O configuration
disk	no restrictions for SCSI disk size	**disk size** restrictions to Mod 54 / Mod A
	0–15 **partitions** per disk	1–3 partitions per disk
	no low-level formatting	**low-level formatting** → wastes disk space
	no emulation → performance	ECKD **emulation** overhead
	built-in **asynchronous I/O** → performance	async I/O requires Parallel Access Volumes

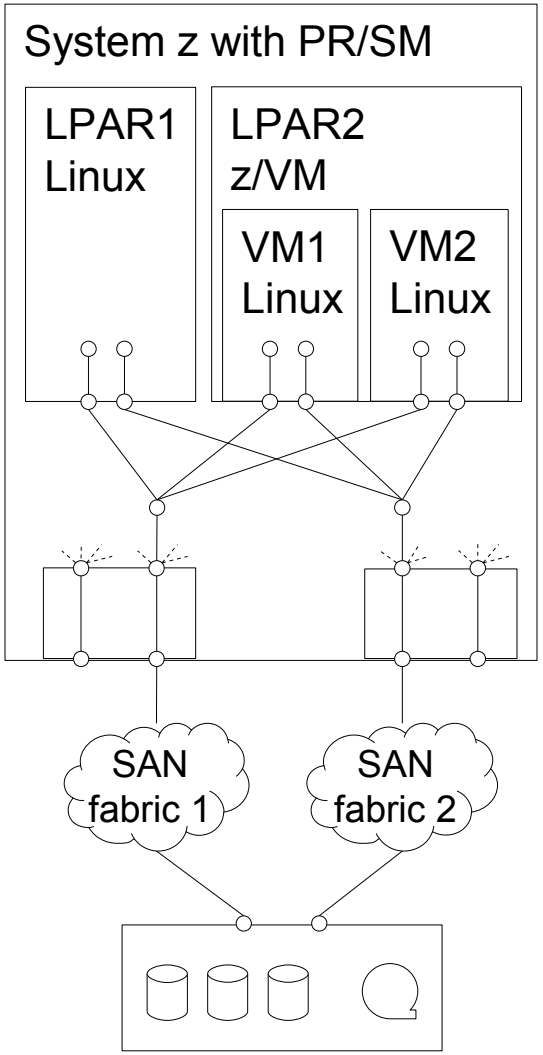
© Copyright IBM Corp. 2014

SAN Topologies and System z

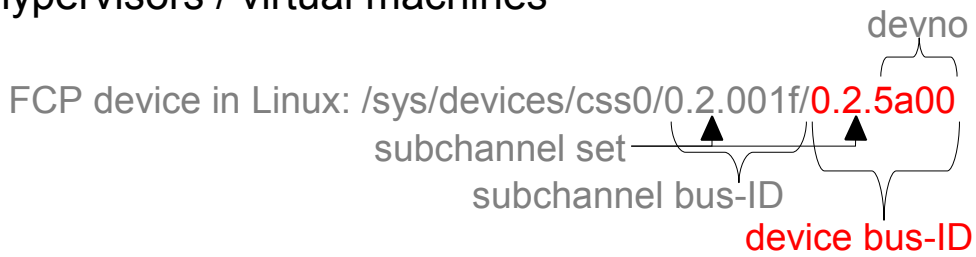


© Copyright IBM Corp. 2014

FCP with System z



hypervisors / virtual machines



FCP devices (direct-attached) / virtual HBAs
  FCP subchannels

CHPIDs (1 per PCHID if spanned), type FCP

I/O Definition

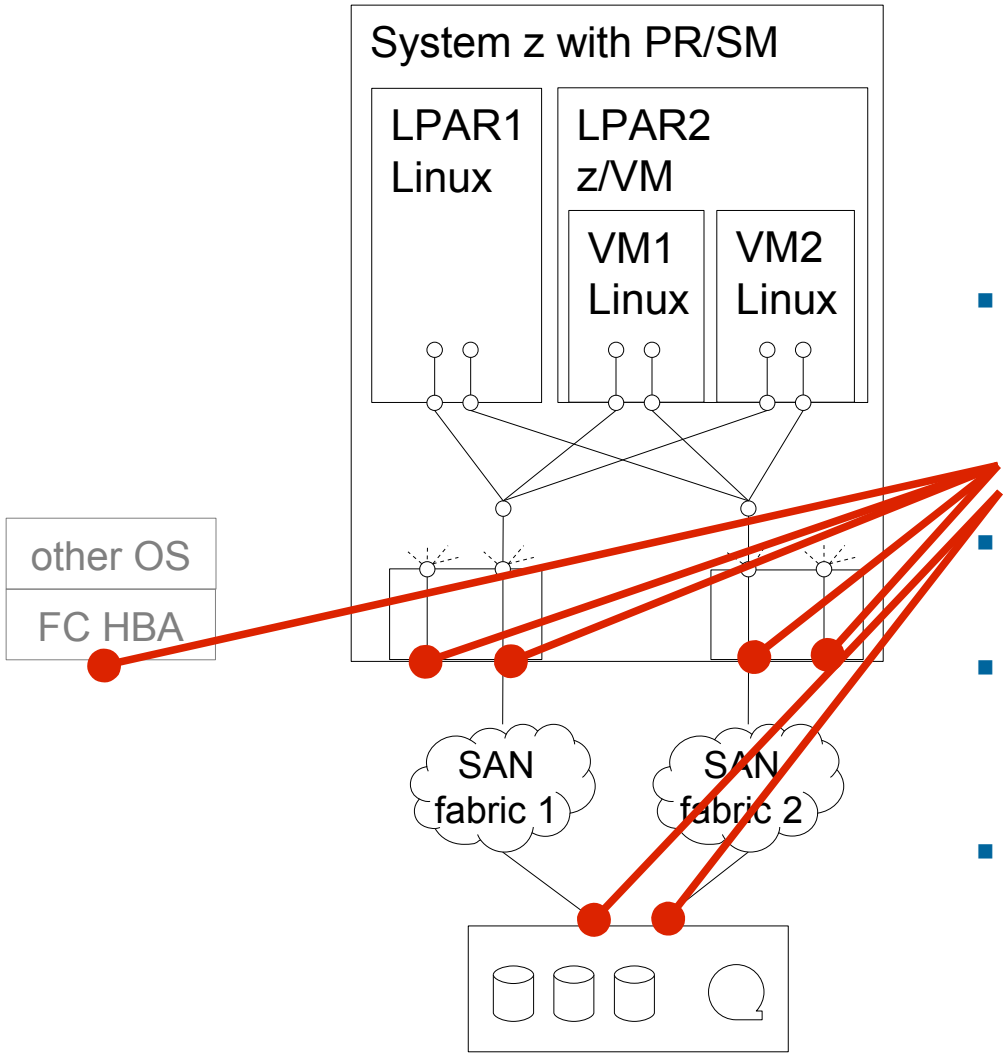
  PCHIDs / FCP Channels / HBAs (2 per card)
FICON Express8S cards
  initiator physical fibre ports (2 per card)

paths over physically redundant FC fabrics

  target physical fibre ports
storage target

© Copyright IBM Corp. 2014

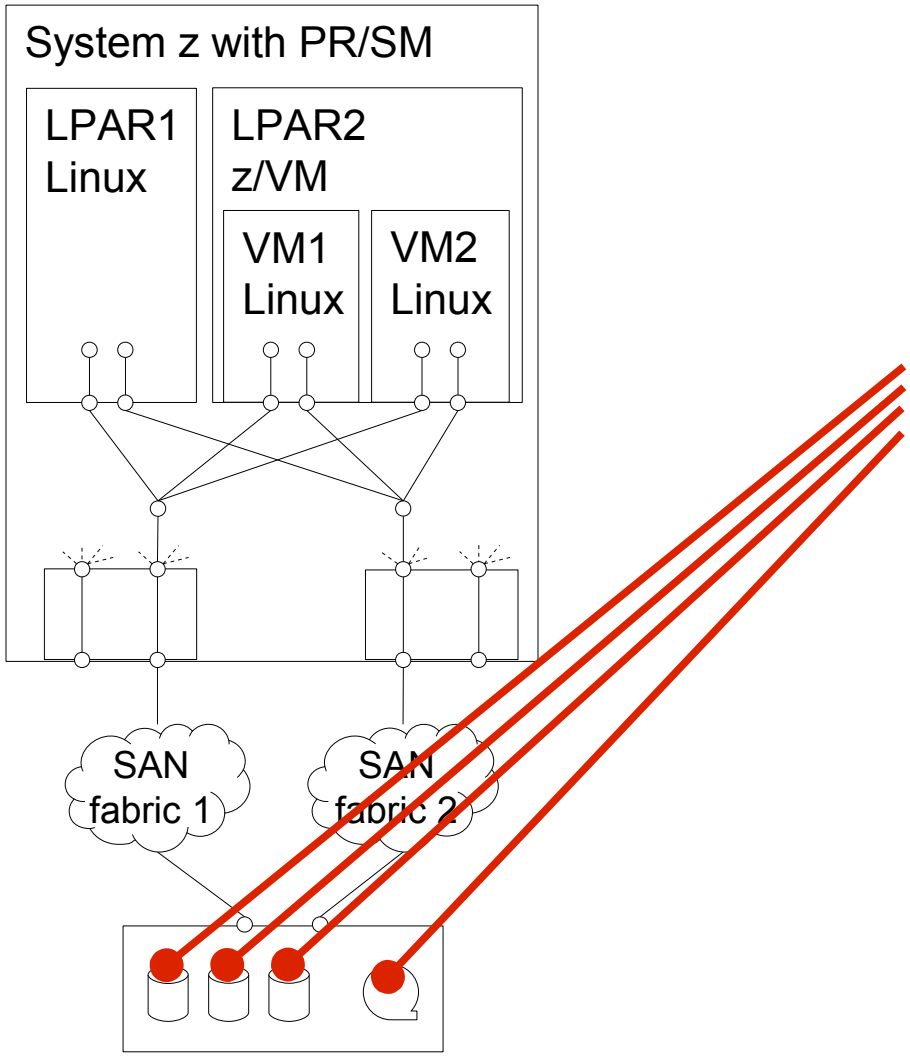
Worldwide Port Names (WWPNs)



- Servers (initiators) and storage devices (targets) attach through Fibre Channel ports (called N_Ports).
- An N_Port is identified by its Worldwide Port Name (WWPN).
- For redundancy, servers or storage should attach through several N_Ports.
- sample WWPNs:
  FCP channel: `0xc05076ffe4803931`
  storage target: `0x5005076303000104`

© Copyright IBM Corp. 2014

Logical Unit Numbers (LUNs)



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

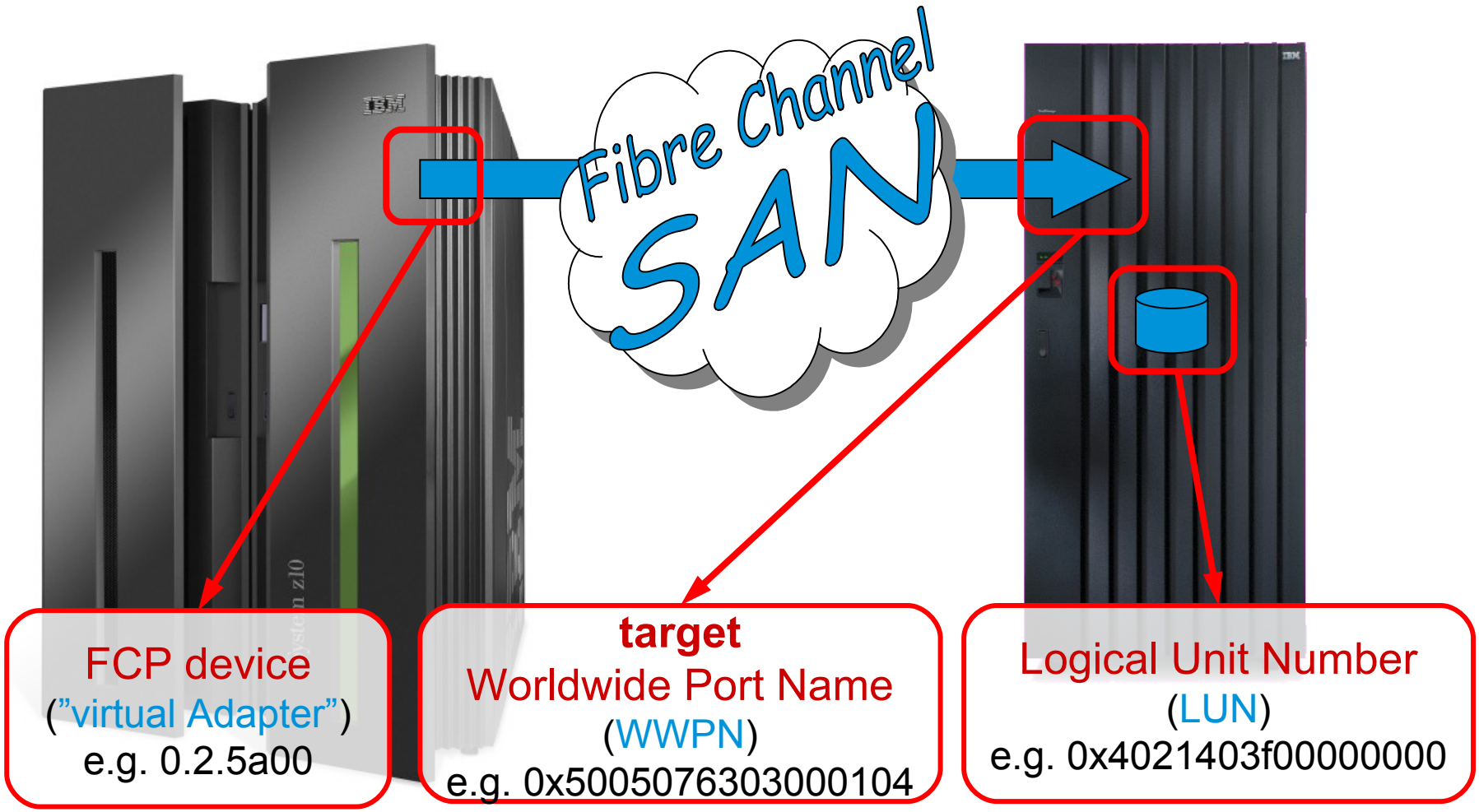
A logical unit behind a target WWPN is identified by its Fibre Channel Protocol Logical Unit Number (FCP LUN).

Mind different LUN formats [T10 SAM], e.g.:

- DS8000 (pseudo flat space addressing, but with 2nd level): 0x4021403f00000000
- SVC / V7000, XIV, FlashSystem, Tape (peripheral device addressing, single level): 0x01c8000000000000

© Copyright IBM Corp. 2014

SAN Addressing for One (of Multiple) Paths



© Copyright IBM Corp. 2014

Setup

© Copyright IBM Corp. 2014

Setup Overview for FCP with Linux on System z

1) Optionally: Early Preparation (see backup slides at the end)
2) Define FCP devices within the mainframe (I/O Definition File), dedicate in z/VM.
3) Enable NPIV for the FCP devices (Service Element / HMC).
4) Configure zoning for the FCP devices to gain access to desired target ports within a SAN, max. one single initiator (virtual) WWPN per zone.
5) Configure LUN masking for the FCP devices at the target device to gain access to desired LUNs.
6) In Linux, configure multipathing
7) In Linux, configure target WWPNs and LUNs to obtain SCSI devices.

Note: If FCP Channel is directly connected to a target device (point-to-point), steps 3 & 4 do not apply.
After preparation, steps 4 & 5 can be conducted before or in parallel to step 3.

© Copyright IBM Corp. 2014

Define FCP Devices

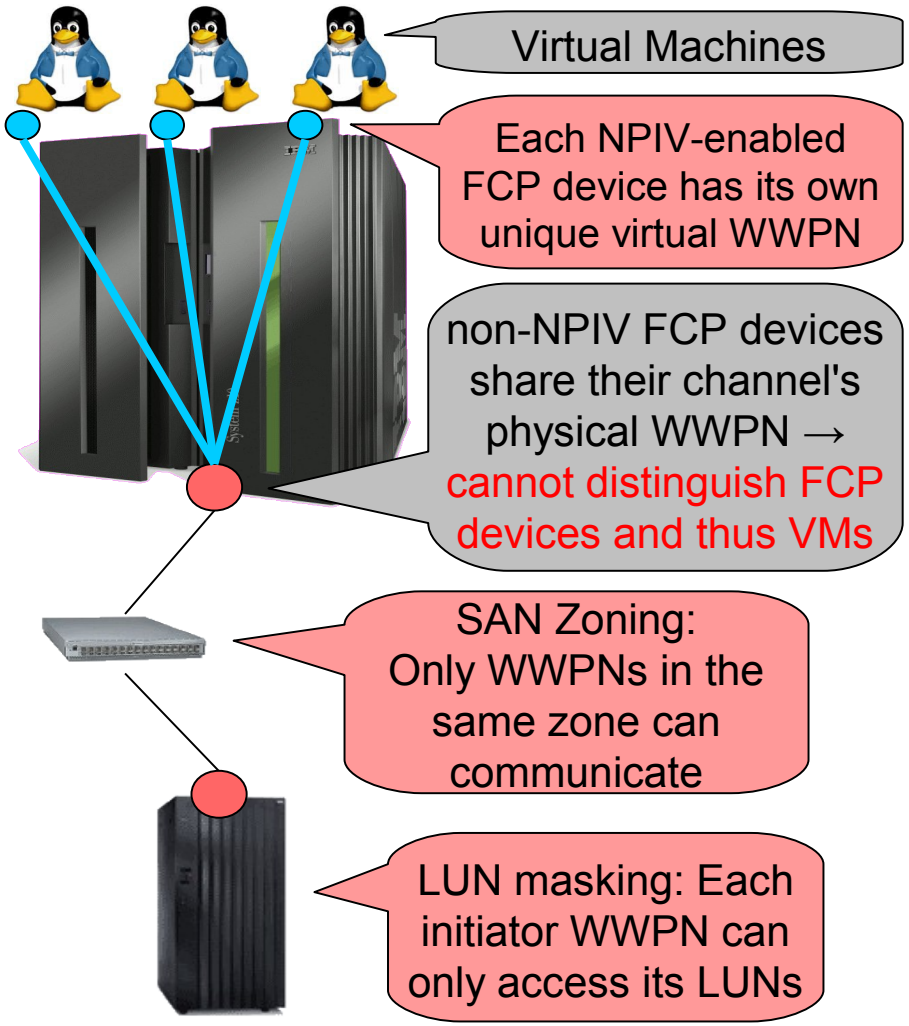
- virtual device config & passthrough for LPAR hypervisor (PR/SM):

```
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,SCHSET=2
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
```

- for z/VM: dedicate 1 FCP device per CHPID per z/VM guest in its user directory

© Copyright IBM Corp. 2014

NPIV: N_Port ID Virtualization

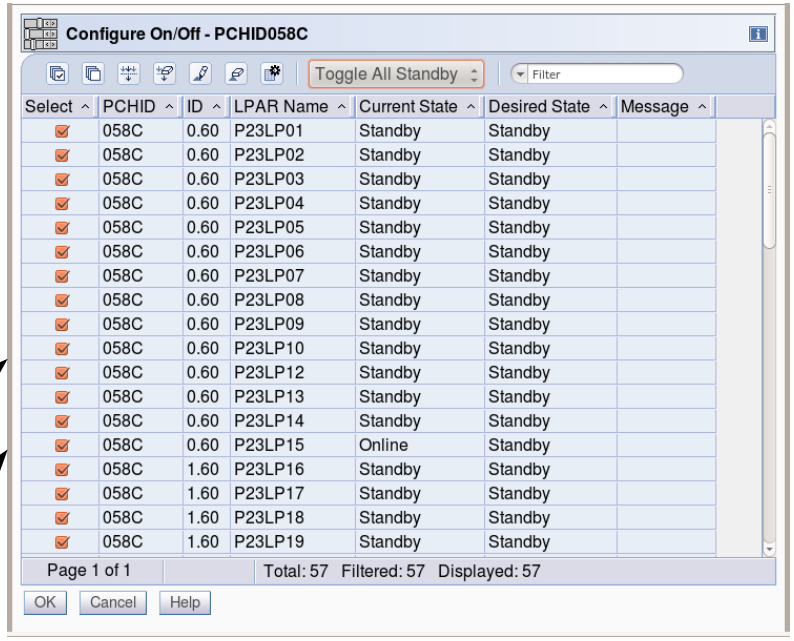


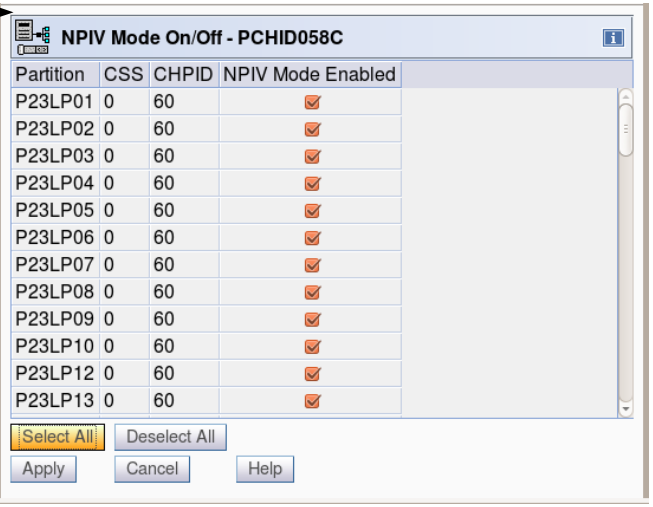
- Each virtual HBA uses FDISC with virtual WWPN to log into fabric and get its own N_Port ID [T11 FC-LS]
- Enable NPIV on the SAN switch before enabling it on the System z server.
- Switches typically limit the number of NPIV-enabled FCP devices per switch.
- Some switches limit the number of NPIV-enabled FCP devices per switch port.
- Each port login from an NPIV-enabled FCP device into a storage target counts as a separate host login, which are limited at storage.

© Copyright IBM Corp. 2014

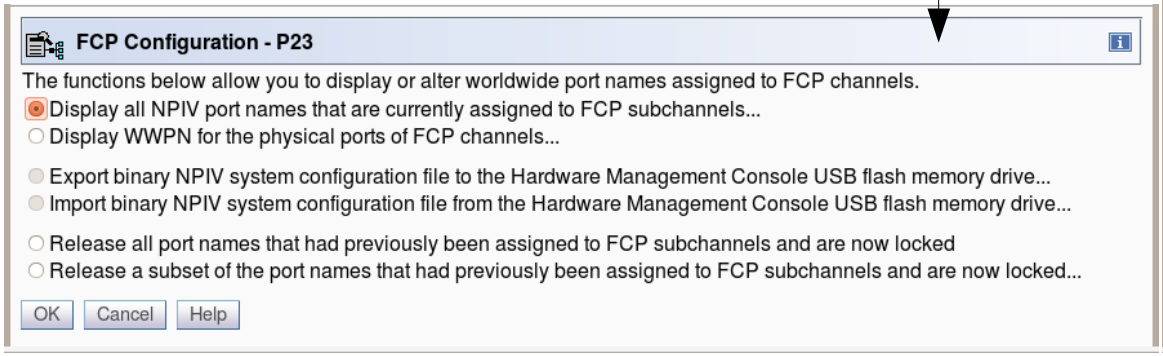
NPIV: Enable for all FCP Devices

- On the service element, for each FCP PCHID for each LPAR:
  1) Configure off its CHPID on LPAR
  2) Enable NPIV mode for LPAR
  3) Configure on its CHPID on LPAR if desired





- Manage FCP Configuration on the SE:



© Copyright IBM Corp. 2014

NPIV: ZFCP Point of View

- Is NPIV enabled for a certain FCP device?:

```
# lszfcp -Ha | grep -e port_type -e ^0
0.2.5a00 host0
    port_type               = "NPIV VPORT"
```

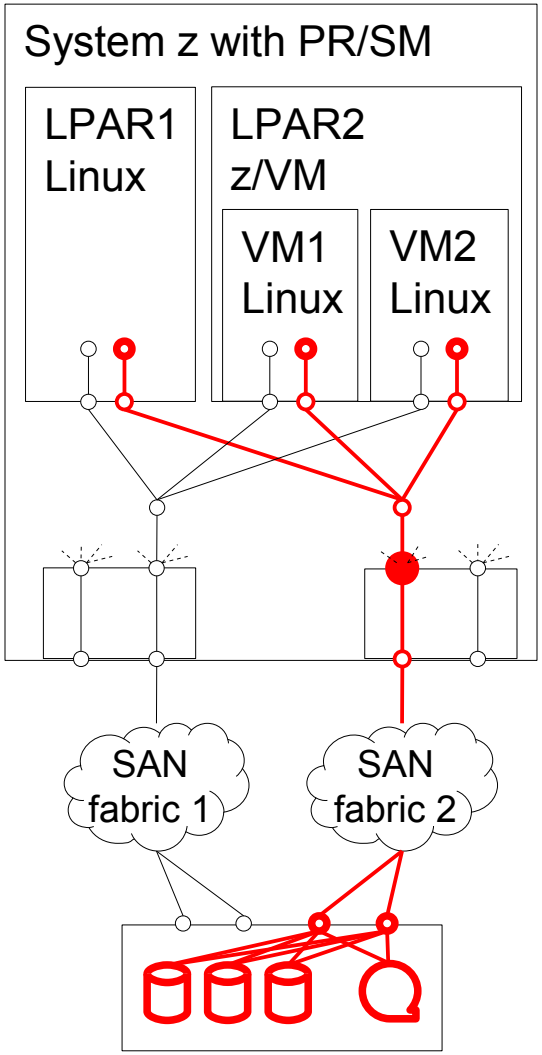
- alternatively for older Linux version (< SLES 11 SP1, < RHEL 6.0, < 2.6.30):

```
# lszfcp -Ha | grep -e port_name -e ^0
0.2.5a00 host0
    permanent_port_name = "0xc05076ffe5005611"
    port_name           = "0xc05076ffe5005350"
```

  - "permanent_port_name" is the WWPN assigned to the FCP channel
  - "port_name" is the WWPN used by the FCP device
  - if both port names differ NPIV is enabled, otherwise not

© Copyright IBM Corp. 2014

System z Hardware for FCP: Limits per Channel



assuming one online FCP device per VM per PCHID
V: # of VMs per PCHID
P: # of target ports per NPIV-enabled FCP device
L: # of LUNs per target port
assuming equal distribution of resources:
$V \le 32 \;\&\&\; V \cdot (P+1) \le 500 \;\&\&\; V \cdot P \cdot L \le 4096$

FCP devices (direct-attached) / virtual HBAs:
- ≤32 online NPIV-enabled FCP devices per PCHID → Linux
- ≤255 defined FCP devices per LPAR per CHPID → IODF
- ≤480 defined FCP devices per CHPID → IODF

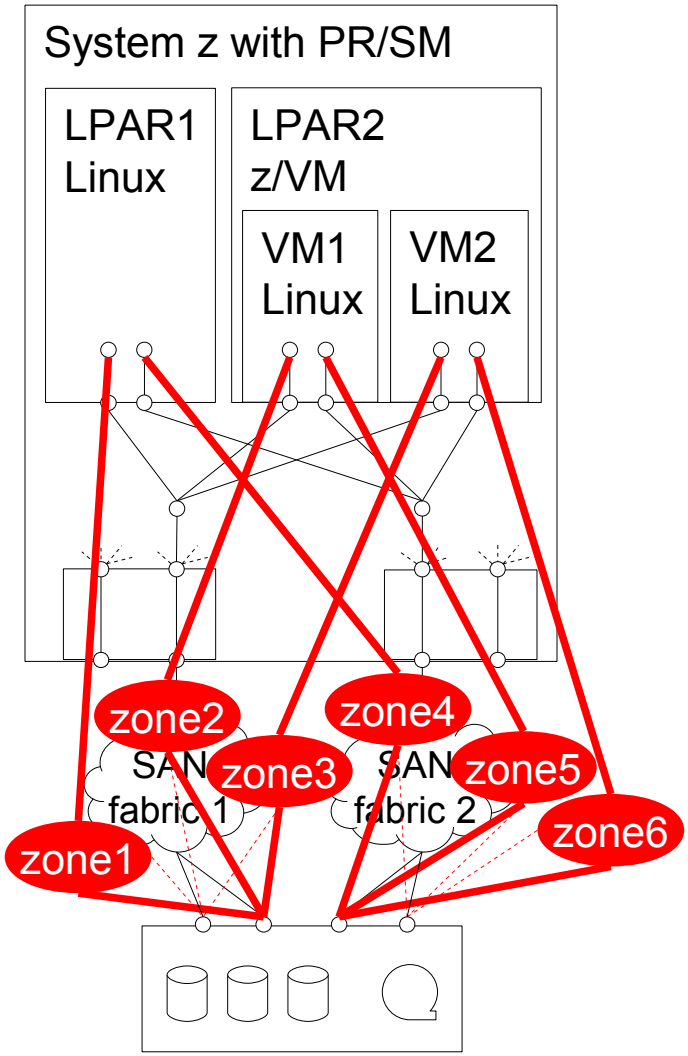
PCHID / FCP Channel / HBA

≤500 open target ports per PCHID → zoning
account for 1 zfcp-internal nameserver port per FCP device!

≤4096 attached LUNs per PCHID → LUN masking & zoning

© Copyright IBM Corp. 2014

Zoning



- **Single Initiator Zoning based on WWPN** (as opposed to based on switch port): Have individual zone for each NPIV WWPN, to avoid storms of change notifications and unnecessary recoveries.
- Each FCP device has its own zone
- Since usually >1 initiator per target port, zones overlap at target ports
- Depending on storage recommendations, a zone can include multiple target ports

© Copyright IBM Corp. 2014

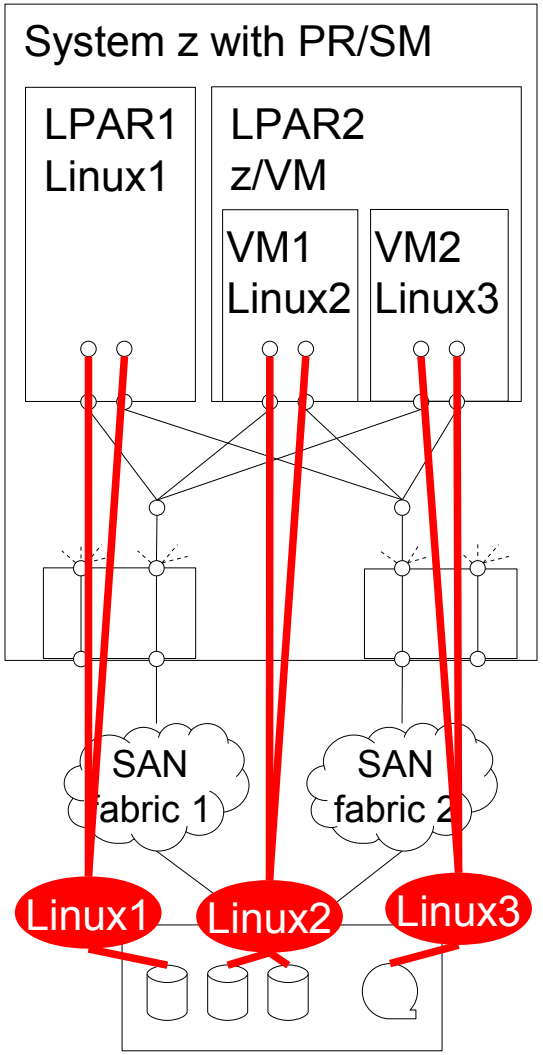
Zoning: No automatic port rescan on events

- The rescan of ports following a fabric change event can cause high fabric traffic, especially when many Linux images share an FCP channel over multiple NPIV-enabled FCP devices in the same zone. This can lead to errors due to timeouts.
- Recommendation: Implement single initiator zones (based on (virtual) WWPNs)
- If single initiator zones are impossible, as a workaround, disable automatic port rescanning by setting kernel parameter in /etc/zipl.conf:
  `zfcp.no_auto_port_rescan=1`
- Ports are still unconditionally scanned when the adapter is set online and when user-triggered writes to the sysfs attribute “port_rescan” occur.
- On fabric changes, manually trigger a port rescan using the following command:
  `# echo 1 > /sys/bus/ccw/drivers/zfcp/0.0.1700/port_rescan`
- Automatic port rescanning is enabled by default.

- IBM is working on improving automatic port scanning to replace this workaround.

© Copyright IBM Corp. 2014

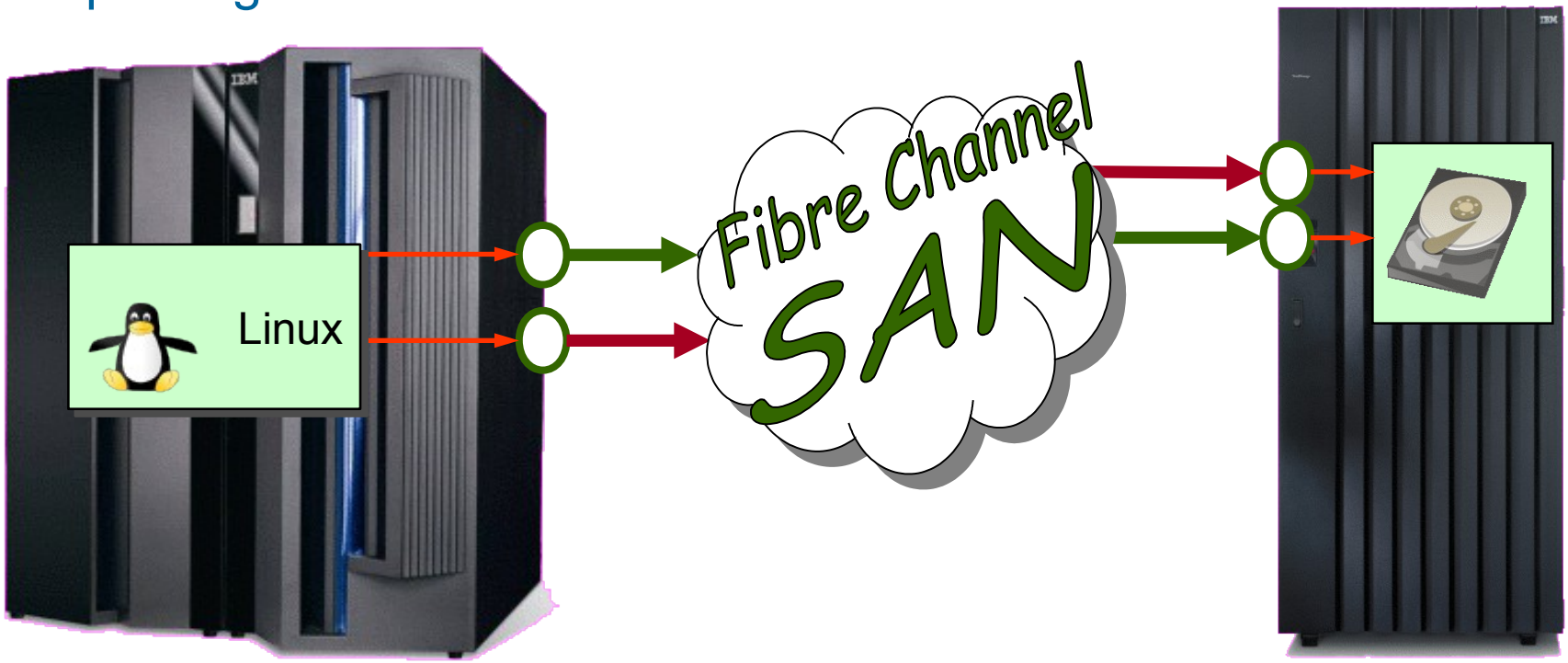
LUN Masking



- In the storage target,
  use virtual initiator WWPNs of NPIV-enabled FCP devices to let each VM only access:
  - Its own exclusive logical units.
  - Logical units shared with other VMs (potentially on other physical machines).
    NOTE: Sharing requires OS support such as clustering file system!
- Depending on storage target type,
  this might require individual volume groups.

© Copyright IBM Corp. 2014

Multipathing



- ≥2 disjoint paths from OS to target device (disk,tape,…); independent FCP cards, independent switches, and independent target ports.
  - Redundancy: Avoid single points of failure
  - Performance: I/O requests can be spread across multiple paths
  - Serviceability: When component of one path is in maintenance mode I/O continues to run through other path(s)
- Linux does multipathing different for disks and tapes …

© Copyright IBM Corp. 2014

Multipathing for Disks – Persistent Configuration

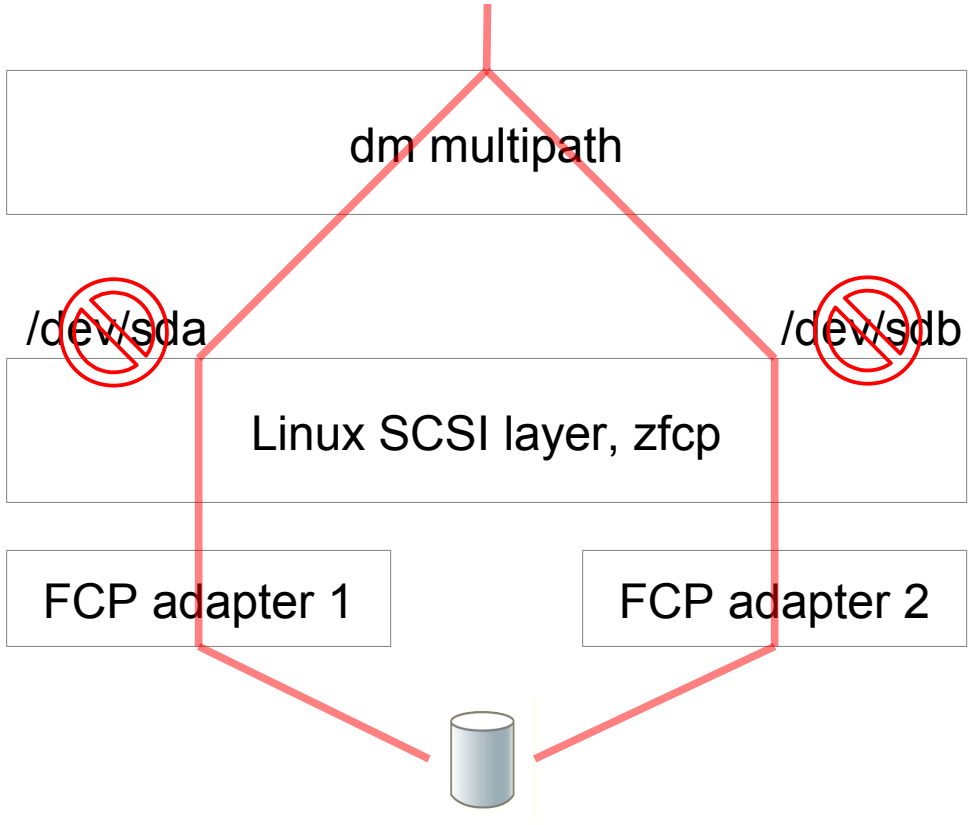
- **Use multipathing on installation** for all disks incl. root-fs and zipl target (/boot):
  SLES, RHEL≥6 out of the box; RHEL5 installer boot parameter in parmfile: mpath.
  Lifting single path to multipath is difficult [≥S10,≥R6] or impractical with LVM [R5].
  - zipl target (/boot):
    use multipathing with sep. mountpoint, or place inside root-fs [S10.4,S11.1,R6],
    if stacking devices on top of multipathing see zipl_helper.device-mapper docs;
    only for RHEL5 use single path SCSI disk device for separate /boot mountpoint
  - root-fs (/):
    always multipathing (optionally stack devices on top, see above if /boot included)
  - any other mountpoint or direct access block device:
    always multipathing (optionally any other virtual block devices such as LVM on top)
- **Post installation [SLES, RHEL]:**
  - **ensure /etc/multipath.conf is suitable (esp. blacklist)**
  - **ensure multipathd is enabled and running (re-activates failed paths)**
  - NOTE: option rr_min_io is called rr_min_io**_rq** in more recent distros

© Copyright IBM Corp. 2014

Multipathing for Disks – device-mapper multipath devices

- device-mapper multipath target in kernel creates one block device per disk:
  /dev/mapper/36005076303ffc5620000000000000010cc

unique WWID



- World-Wide Identifier (not LUN!) from storage server identifies volume / disk / path group
- each SCSI device represents a single path to a target device, do ***not*** use these devices directly!

© Copyright IBM Corp. 2014

Multipathing for Disks – device-mapper multipath devices (cont.)

- Multipath devices are created automatically when SCSI LUNs are attached

WWID for volume

pathgroup

```
# multipath -ll
36005076303ffc56200000000000002006 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
  |- 0:0:2:1074151456 sda   8:0   active ready running
  `- 1:0:5:1074151456 sdb   8:16  active ready running
```

- Multipath devices are virtual block devices, can be used as container for, e.g.
  - Partitions
  - Logical Volume Manager (LVM)
  - Directly for a file system or as raw block device (e.g. for RDBMS)
- Device to work with: e.g. /dev/mapper/36005076303ffc56200000000000002006
  (or user-friendly / alias multipath names such as /dev/mapper/mpatha if enabled)

```
# mkfs.ext4 /dev/mapper/36005076303ffc56200000000000002006
# mount /dev/mapper/36005076303ffc56200000000000002006 /mnt
```


© Copyright IBM Corp. 2014

Multipathing for Disks – LVM on Top

- explicitly ensure that all LVM PVs are **assembled** from multipath devices (/dev/mapper/…) instead of single path scsi devices (/dev/sd…)
  NOTE: pvcreate on multipath devices is necessary but not sufficient!
- otherwise PVs can randomly use only a single path anytime → lack of redundancy
- use a white list of explicitly allowed PV base device names in */etc/lvm/lvm.conf*:

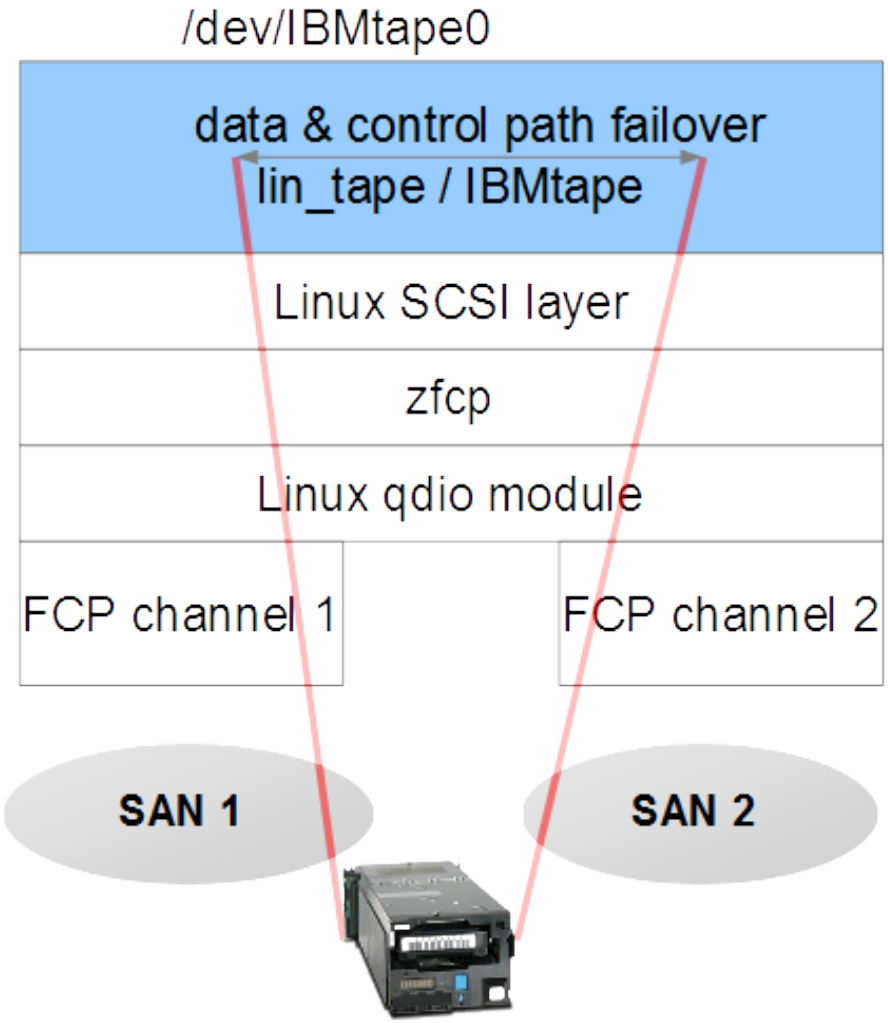
```
filter = [ "a|^/dev/mapper/.*$|", "a|^/dev/dasd.*$|",
           "a|^/dev/scm.*$|", "a|^/dev/dcssblk.*$|", "r|.*|" ]
```

- verify the correct filter for every SCSI disk device node using pvscan, “Skipping (regex)” must be shown:

```
# pvscan -vvv 2>&1 | fgrep '/dev/sd'
...
  /dev/sda: Added to device cache
  /dev/block/8:0: Aliased to /dev/sda in device cache
  /dev/disk/by-path/ccw-0.0.50c0-zfcp-0x1234123412341234:\
    0x0001000000000000: Aliased to /dev/sda in device cache
  ...
  /dev/sda: Skipping (regex)
```


© Copyright IBM Corp. 2014

Multipathing for IBM Tapes



Use Case:

- Backup with Tivoli Storage Manager (TSM) (client & server for Linux on System z)

Setup:

- enable via lin_tape module parameter e.g. in /etc/modprobe.conf.local: options lin_tape alternate_pathing=1
- attach all paths to tape drive

© Copyright IBM Corp. 2014

Multipathing – Error Recovery on FC Transport Layer

- on zfcp detecting broken target port (cable pull, switch maint., target logged out): tell FC transport class which starts fast_io_fail_tmo & dev_loss_tmo for rport
- on fast_io_fail_tmo: zfcp port recovery returns pending IO with result DID_TRANSPORT_FAILFAST
- ~~on dev_loss_tmo: zfcp port recovery returns pending IO with result DID_NO_CONNECT and FC transport deletes SCSI target with its SCSI devices~~ ← issues under IO
  → disable dev_loss_tmo and enable fast_io_fail_tmo (5 seconds):
  - for disks: “infinity” or “2147483647” for dev_loss_tmo in multipath.conf [R,S]
  - double check with “lszfcp -Pa”
- path failover: kernel dm_multipath can re-queue returned IO on another path

© Copyright IBM Corp. 2014

Multipathing – Error Recovery on SCSI Layer

- the following applies if the lower FC transport layer could not detect/recover errors, typically due to dirty fibres or SAN switches suppress RSCNs ← must fix reasons

- on starting IO request: start SCSI command (=block request) timeout
- on SCSI command timeout: start SCSI error handling on SCSI host **as last resort**; multipathd can only see path failure once eh processed path checker IO request;
  - try to abort SCSI command
  - if above failed, escalate and try to reset device (=LUN)
  - if above failed, escalate and try to reset target
  - if above failed, escalate and try to reset host (=FCP device recovery)
  - if above failed, retry scsi_eh a few times; finally give up: set SCSI device offline
- since above handling can take many minutes to complete, recent distros provide “eh_deadline” directly escalating to host reset after deadline

© Copyright IBM Corp. 2014

Multipathing – Handling on Losing Last Path

- if all paths gone at the same time (even for a split second, e.g. during scsi_eh), return IO error (clusters) or queue IO (other):
  disks: /etc/multipath.conf: 'no_path_retry queue' (alias feature queue_if_no_path);
  multipath.conf settings can contradict → double check if queueing is active:

```
# multipathd -k'list maps status'
name                                failback  queueing paths dm-st  write_prot
36005076802870052a000000000000318 immediate on       2     active rw
36005076304ffc3e80000000000002000 -         on       2     active rw
```

  otherwise data corruption can occur if applications don't handle IO errors correctly
- above setting **required** for z/VM SSI live guest relocation (LGR)
  with dedicated FCP devices [z/VM docs1, docs2]

- if IO is stuck due to queueing and paths won't return but you want to flush IO:
  # dmsetup message <*mapname*> 0 fail_if_no_path [SLES,RHEL]

© Copyright IBM Corp. 2014

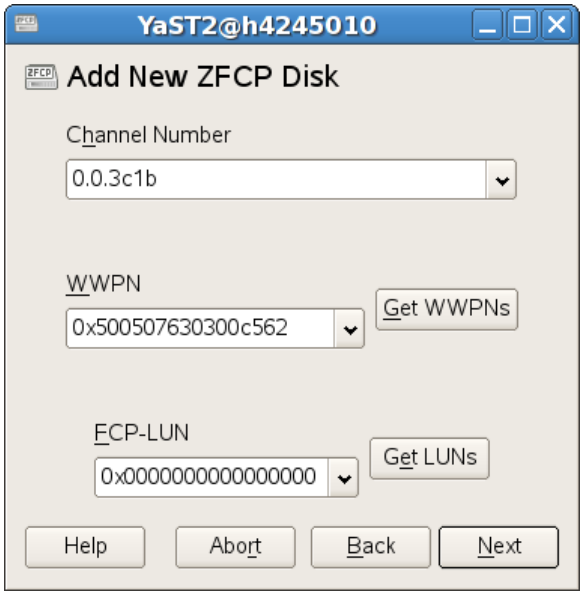
LUN Management with ZFCP: 2 Methods

1) explicit manual LUN whitelist (traditional)
   - user specifies every single path using <FCP device,WWPN,FCP LUN>
   - zfcp only attaches these paths
2) automatic LUN scanning (new and only with NPIV-enabled FCP devices)
   - user specifies to only set FCP device online
   - zfcp attaches all paths visible through fabric zoning and target LUN masking

- to ignore certain LUNs: disable automatic LUN scanning with kernel boot parameter "zfcp.allow_lun_scan=0" in /etc/zipl.conf, and then use explicit manual LUN whitelists for all FCP devices in such Linux instance
- do not mix up automatic LUN scanning (new) with automatic port scanning (no more “port_add”, since RHEL 6.0 and SLES11 SP1)
- do not use zfcp sysfs interface directly, e.g. with own scripting; use tested & supported distribution mechanisms…

© Copyright IBM Corp. 2014

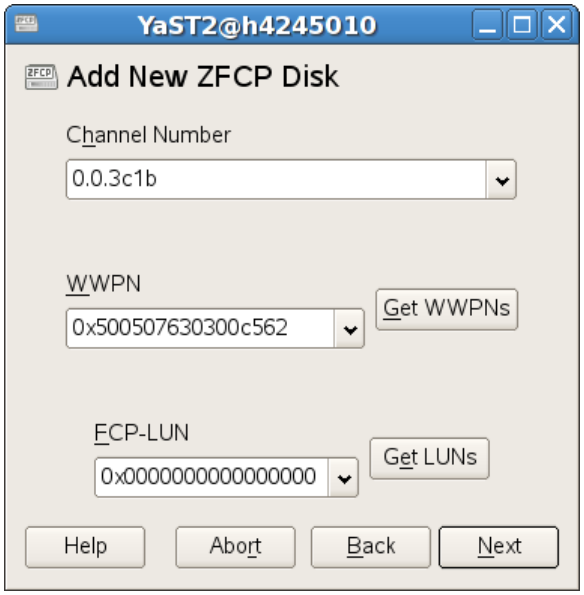
LUN Management with ZFCP: SLES Installation



- interactive
  - GUI / TUI: YaST installer button "Configure ZFCP Disks"
  - GUI and TUI can discover available FCP devices, WWPNs, and LUNs
- unattended
  - AutoYAST: <zfcp> element
- temp. workaround for auto LUN scan: specify just one valid path per FCP device

© Copyright IBM Corp. 2014

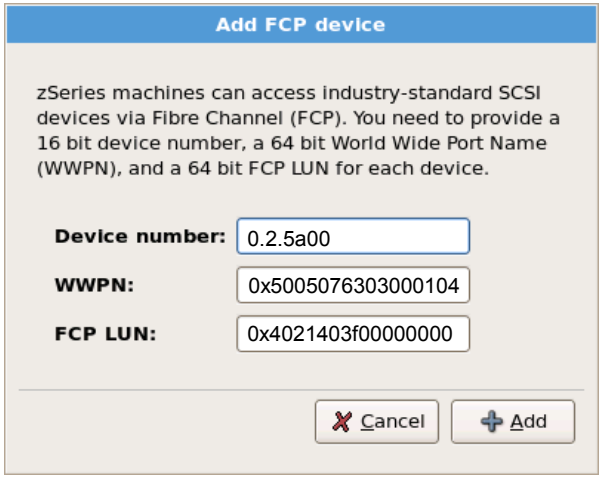
LUN Management with ZFCP: SLES Post-Installation



- GUI: yast2 zfcp
- TUI: yast zfcp
- command line:
  - optionally discover WWPNs or LUNs manually:
    zfcp_san_disc
  - enable/disable FCP device:
    zfcp_host_configure 0.2.5a00 1/0
  - attach/detach FCP LUN to/from enabled FCP device:
    zfcp_disk_configure 0.2.5a00 0x5005076303000104 0x4021403f00000000 1/0
- GUI and TUI can discover available
  FCP devices, WWPNs, and LUNs
- if changes affect root-fs dependencies, process changes: mkinitrd && zipl
- temp. workaround for auto LUN scan: only use zfcp_host_configure, nothing else

© Copyright IBM Corp. 2014

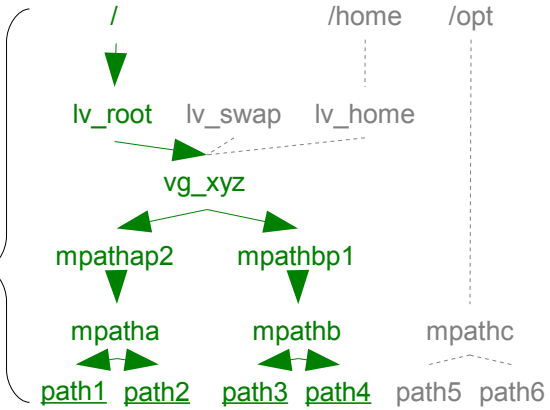
LUN Management with ZFCP: RHEL Installation



- interactive
  - GUI of anaconda installer
- unattended
  - kickstart: “zfcp” option
- both interactive and unattended
  - FCP_*n*='*device_bus_ID WWPN FCP_LUN*' in generic.prm or in a CMS conf file
  - RHEL7 also in generic.prm: rd.zfcp=*device_bus_ID*,*WWPN*,*FCP_LUN*
  - can also be used for e.g. install from SCSI LUN
- RHEL5 installer boot parameter in generic.prm parmfile: “mpath”
- temp. workaround for auto LUN scan: specify just one valid path per FCP device

© Copyright IBM Corp. 2014

LUN Management with ZFCP: RHEL Post-Installation



- GUI only available during installation.
- SCSI disk paths (indirectly) required to mount root-fs, e.g. each path of all multipath PVs of a VG with root-LV
  - RHEL5: /etc/zfcp.conf (see below)
  - RHEL6: /etc/zipl.conf:
    … rd_ZFCP=0.2.5a00,0x5005076303000104,0x4021403f00000000 rd_ZFCP=…
  - RHEL7: /etc/zipl.conf:
    … rd.zfcp=0.2.5a00,0x5005076303000104,0x4021403f00000000 rd.zfcp=…
  - process changes: mkinitrd && zipl
- any other SCSI devices such as data volumes or tapes
  - RHEL5/6/7: /etc/zfcp.conf:
    …
    0.2.5a00 0x5005076303000104 0x4021403f00000000
  - activate additions to /etc/zfcp.conf: zfcp_cio_free [RHEL≥6] && zfcpconf.sh
- optionally discover LUNs manually: lsluns
- temp. workaround for auto LUN scan: specify just one valid path per FCP device

© Copyright IBM Corp. 2014

LUN Management with ZFCP: Automatic LUN Scanning for NPIV-enabled FCP Devices

11.2
7.0

- With this feature, NPIV-enabled FCP devices attach LUNs automatically.
- Needs zoning and LUN masking per Linux image to only access desired LUNs.
- Automatic LUN scanning is disabled by default in SLES11.
  To enable automatic LUN scanning
  set the kernel boot parameter "zfcp.allow_lun_scan=1" in /etc/zipl.conf
- to manually trigger a LUN discovery:
  # rescan-scsi-bus.sh
- then check with lszfcp -D

```
# lszfcp -D
0.0.1700/0x500507630503c1ae/0x4022400000000000 0:0:12:1073758242
0.0.1700/0x500507630503c1ae/0x4022401000000000 0:0:12:1073883778
0.0.1700/0x500507630503c1ae/0x4022402000000000 0:0:12:1073889314
```

- there are no sysfs directories in the zfcp branch for automatically attached LUNs!
  /sys/bus/ccw/drivers/zfcp/<*FCP device bus-ID*>/0x<*WWPN*>/~~0x<*FCP LUN*>~~

© Copyright IBM Corp. 2014

IPL (booting) over FCP

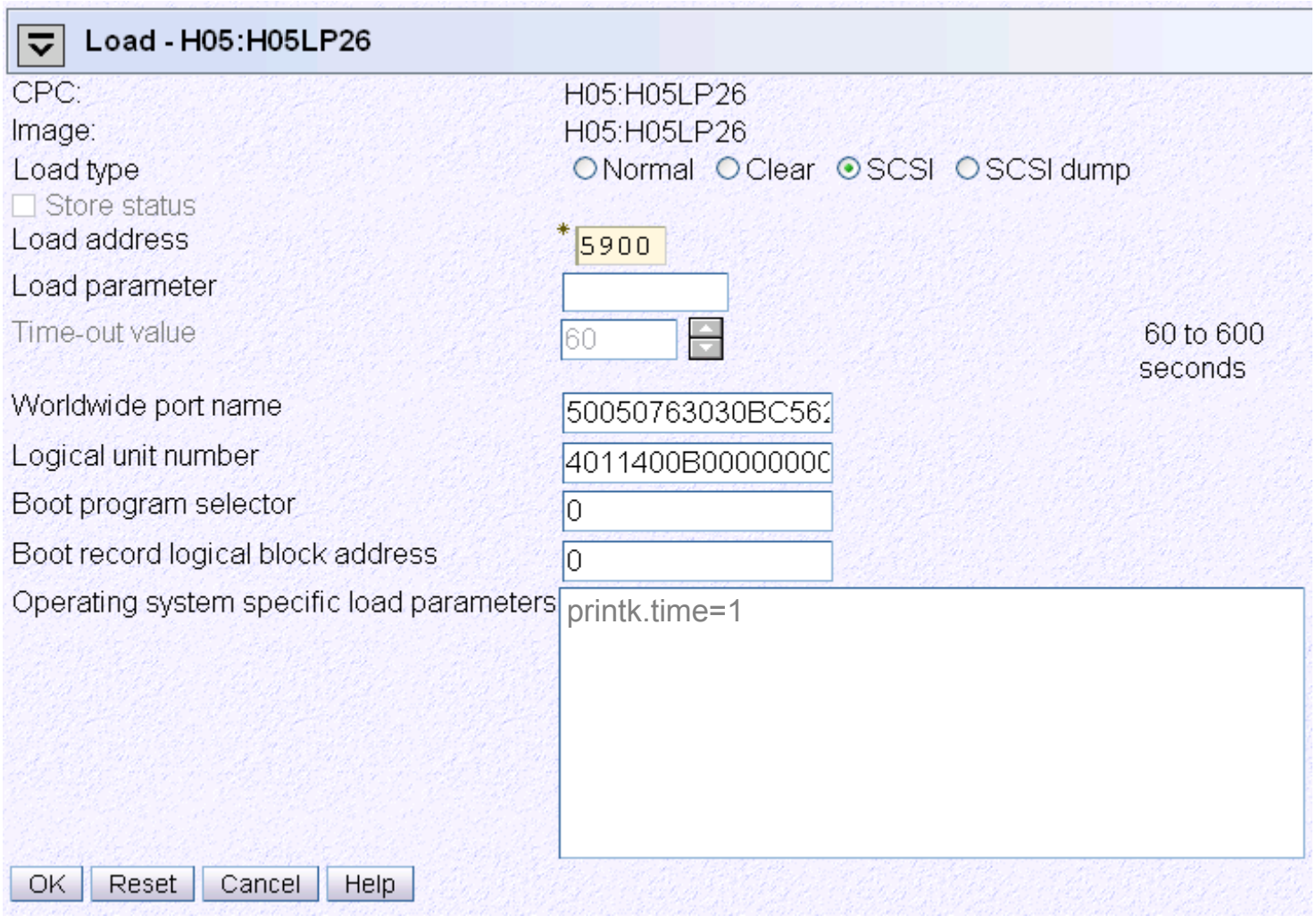
© Copyright IBM Corp. 2014

SCSI IPL

- SCSI IPL expands the set of IPL'able devices
  - SCSI disk to boot Linux (“zipl target”, /boot mountpoint or inside root-fs)
  - SCSI disk for standalone zfcpdump (hypervisor-assisted system dumper)
- New set of IPL parameters
  - Requires to address the SCSI disk
    - FCP device number
    - target WWPN
    - LUN
  - Select zipl boot menu entry with “bootprog”, no interactive menu as with DASD
  - Pass arbitrary kernel boot parameters with “OS load parm”/“scpdata” [S11SP1,R6]
- LPAR and z/VM guests supported
- SCSI (IPL) with z/VM Version 4.4 (with PTF UM30989) or newer

© Copyright IBM Corp. 2014

SCSI IPL example LPAR



© Copyright IBM Corp. 2014

SCSI IPL example z/VM

WWPN

LUN

in hexadecimal format with a blank between the first 8 from the final 8 digits

```
set loaddev port 50050763 03000104 lun 40214000 00000000
set loaddev bootprog 3 scpdata 'printk.time=1'

query loaddev
PORTNAME 50050763 03000104    LUN  40214000 00000000    BOOTPROG 3
BR_LBA   00000000 00000000
SCPDATA
     0----+----1----+----2----+----3----+----4----+----
0000 PRINTK.TIME=1

i 1900
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 v2.4.4).
00: MLOPDM003I: Machine loader finished, moving data to final storage location.
…
Linux version 3.0.101-0.29-default (geeko@buildhost) (gcc version 4.3.4 [gcc-4_3-branch
  revision 152973] (SUSE Linux) ) #1 SMP Tue May 13 08:40:57 UTC 2014 (9ec28a0)
setup.1a06a7: Linux is running as a z/VM guest operating system in 64-bit mode
setup.dae2e8: Reserving 128MB of memory at 896MB for crashkernel (System RAM: 1024MB)
…
Kernel command line: root=/dev/mapper/36005076303ff0104000000000002100
                     TERM=dumb crashkernel=256M-:128M BOOT_IMAGE=0 printk.time=1
```

device number of FCP device with access to SCSI boot disk (zipl target, typically /boot).

© Copyright IBM Corp. 2014

Summary of FCP

- available for zSeries and System z
- based on existing Fibre Channel infrastructure
- integrates System z into standard SANs
- connects to switched fabric or point-to-point
- runs on all available z/VM and RHEL/SLES versions
- multipathing for SCSI disks & tapes is a must
- gives you new storage device choices
- buys you flexibility at the cost of complexity
- tooling available, receiving better integration

© Copyright IBM Corp. 2014

More Information

- I/O Connectivity on IBM System z mainframe servers http://www.ibm.com/systems/z/connectivity/
- IBM System Storage Interoperation Center http://www.ibm.com/systems/support/storage/ssic/
- Linux on System z documentation by IBM http://www.ibm.com/developerworks/linux/linux390/distribution_hints.html
  - How to use FC-attached SCSI devices with Linux on System z
  - Device Drivers, Features, and Commands
  - Using the Dump Tools
  - Kernel Messages
- IBM Redbooks
  - Fibre Channel Protocol for Linux and z/VM on IBM System z http://www.redbooks.ibm.com/abstracts/sg247266.html
- SUSE Linux Enterprise Server 11:
  - Release Notes https://www.suse.com/documentation/sles11/
  - Deployment Guide https://www.suse.com/documentation/sles11/book_sle_deployment/data/sec_i_yast2_s390_part.html
  - Administration Guide https://www.suse.com/documentation/sles11/book_sle_admin/data/sec_zseries_rescue.html
  - Storage Administration Guide https://www.suse.com/documentation/sles11/stor_admin/data/bookinfo.html
  - AutoYAST for unattended installation https://www.suse.com/documentation/sles11/book_autoyast/data/createprofile_partitioning.html
- Red Hat Enterprise Linux 6:
  - Release Notes https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/6.5_Release_Notes/
  - Technical Notes https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6/html-single/6.5_Technical_Notes/
  - Installation Guide
    https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/Installation_Guide/Storage_Devices-s390.html#idp22053792
    https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/Installation_Guide/s1-kickstart2-options.html#idp40003296
    https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/Installation_Guide/ap-s390info-Adding_FCP-Attached_LUNs.html
  - DM Multipath https://access.redhat.com/site/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/DM_Multipath/
  - Storage Administration Guide
    https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/6/html/Storage_Administration_Guide/

© Copyright IBM Corp. 2014

Questions?

Steffen Maier

Linux on System z
Development

Schönaicher Strasse 220
71032 Böblingen, Germany

Phone +49 (0)7031-16-2354
maier@de.ibm.com

© Copyright IBM Corp. 2014

Backup

© Copyright IBM Corp. 2014

Setup

© Copyright IBM Corp. 2014

Early Preparation

- Installation of a new machine using the WorldWide PortName Prediction Tool [http://www.ibm.com/servers/resourcelink/]
  - Input: System z I/O definition
  - Output: all virtual NPIV WWPNs for all FCP devices
    - can be used for early SAN zoning and storage target LUN masking even before activation of System z machine
- MES upgrade of a machine migrating existing FCP workload without changing zoning or LUN masking
  - export WWPNs on old machine and import on new machine
  - Always transparent to Linux (it does not care about initiator WWPNs, only about target WWPNs and they only change with the storage)

© Copyright IBM Corp. 2014

Troubleshooting

© Copyright IBM Corp. 2014

Troubleshooting: scsi_logging_level

- More SCSI output in kernel messages
- Default is: 0
- Higher levels can create lots of messages and slow down system due to synchronous output of kernel messages on the console → undesired errors!
  → low level and/or filter console kernel messages with /proc/sys/kernel/printk
- Find issues with LUN discovery and SCSI error handling (recovery) such as dirty fibres but only negligible impact on regular I/O
- Can be added to kernel boot parameters in /etc/zipl.conf:
  "scsi_mod.scsi_logging_level=4605"

```
# scsi_logging_level -s \
  –mlcomplete 1 -T 7 -E 5 \
  -S 7 -I 0 -a 0
New scsi logging level:
dev.scsi.logging_level = 4605
SCSI_LOG_ERROR=5
SCSI_LOG_TIMEOUT=7
SCSI_LOG_SCAN=7
SCSI_LOG_MLQUEUE=0
SCSI_LOG_MLCOMPLETE=1
SCSI_LOG_LLQUEUE=0
SCSI_LOG_LLCOMPLETE=0
SCSI_LOG_HLQUEUE=0
SCSI_LOG_HLCOMPLETE=0
SCSI_LOG_IOCTL=0
```


© Copyright IBM Corp. 2014

Troubleshooting

- Check kernel messages that are possibly related to FCP with Linux on System z:
  - "device-mapper: multipath"
  - sd (SCSI disk)
  - lin_tape* (IBM tape)
  - scsi (common SCSI code)
  - rport (common SCSI code FC remote port messages)
  - zfcp
    - See "Kernel Messages" book on http://www.ibm.com/developerworks/linux/linux390/distribution_hints.html (for RHEL, chose book from development stream with matching kernel version, there are no message IDs so you have to find by matching a message substring)
  - qdio (communication between zfcp and FCP Channel)
- Other syslog messages
  - multipathd (path management daemon for disks)
  - lin_taped (path management daemon for IBM tapes)
- zfcp driver traces available in /sys/kernel/debug/s390dbf/
- Collect data with dbginfo.sh (s390-tools) when reporting a problem to capture configuration, messages and traces

© Copyright IBM Corp. 2014

What's New

© Copyright IBM Corp. 2014

FCP Hardware Data Router Support

- FCP hardware data router reduces path length and improves throughput depending on workload
- To enable the hardware data router feature in zfcp
  set the kernel boot parameter "zfcp.datarouter=1" in /etc/zipl.conf
- check whether the zfcp module parameter datarouter was enabled or disabled:
  `# cat /sys/module/zfcp/parameters/datarouter`
  Y
- under z/VM: show if datarouter is active per FCP device: **`#CP Q V FCP`**
- Note: The hardware data routing feature becomes active only for FCP devices that are based on adapter hardware with hardware data routing support.
- Hardware data router requirements:
  - at least zEnterprise 196 GA2 and zEnterprise 114 with FICON Express8S
  - z/VM: support available beginning with z/VM 6.3
  - RHEL 6.4 & 7[enabled by default], SLES11 SP3

© Copyright IBM Corp. 2014

End-to-end (E2E) data integrity (T10 DIF)

- End-to-end data integrity checking is used to confirm that a data block originates from the expected source and has not been modified during the transfer between the storage system and the FCP device
- To turn end-to-end data integrity checking on
  set the kernel boot parameter "zfcp.dif=1" in /etc/zipl.conf
- check whether the FCP device supports end-to-end data integrity checking, use the lszfcp command and limit the query to a specific FCP device

```
# lszfcp -b 0.0.1700 -Ha |grep prot_capabilities
1
```

  - 0 means: FCP device does not support end-to-end data integrity.
  - 1 means: FCP device supports DIF type 1.
- E2E data integrity checking requirements:
  - at least zEnterprise 196 GA2 and zEnterprise 114 for FICON Express8 & 8S
  - z/VM: support since 5.4 & 6.1 (both with PTFs for APAR VM64925), and later
  - T10 DIF support for SCSI disk only (e.g. DS8000 with release 6.3.1)
  - RHEL 6.4 & 7, SLES11 SP2

© Copyright IBM Corp. 2014

End-to-end (E2E) data integrity extension (DIX)

- Data integrity extension (DIX) builds on DIF to extend integrity checking, e.g. to the operating system, middleware, or an application.
- SCSI devices for which DIX is enabled must be accessed as raw block device with direct I/O (unbuffered I/O bypassing the page cache) or through a file system that fully supports stable page writes, e.g. XFS. Expect error messages on invalid checksums with other access methods.
- Find out about end-to-end data integrity support of an FCP device:

```
# lszfcp -b 0.0.1700 -Ha |grep prot_capabilities
17
```

  - 0 means: FCP device does not support end-to-end data integrity.
  - 1 means: FCP device supports DIF type 1.
  - 16 means: FCP device supports DIX type 1.
  - 17 means: FCP device supports DIF type 1 with DIX type 1.

© Copyright IBM Corp. 2014