SCSI over Fibre Channel Support for Linux on zSeries

slides: http://www.borntraeger.net/WAVV
Christian Bornträger (cborntra@de.ibm.com)
Linux on zSeries
IBM Lab Boeblingen, Germany

WAVV2005 Colorado Springs
May 20-24, 2005

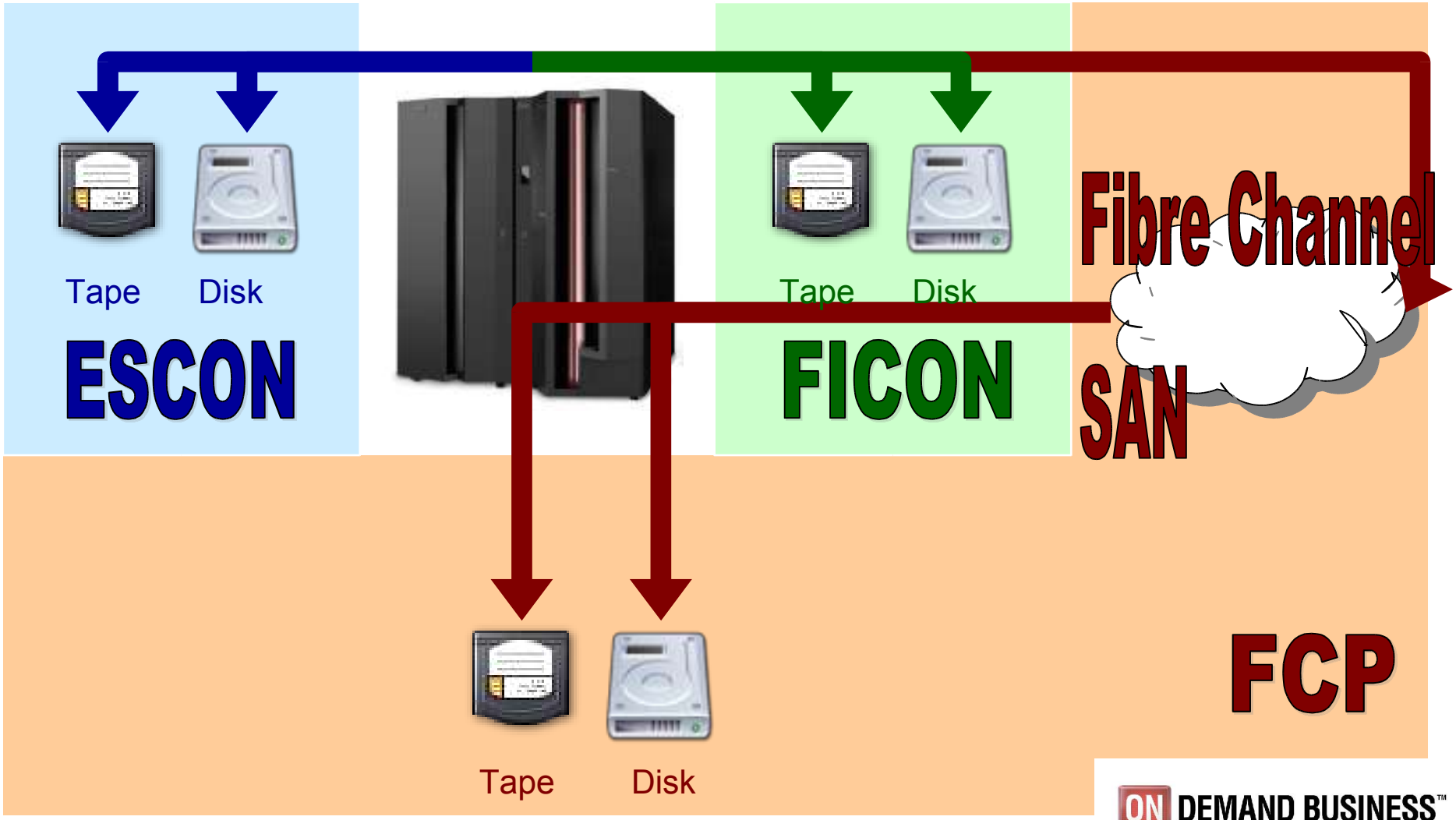
Agenda

- zSeries Hardware
  - Hardware Requirements
  - zSeries in a FC SAN
  - Topologies
- zSeries Software
  - Software Requirements
  - Linux SCSI/FCP Support
  - Multi-Pathing
- Storage Devices
  - Disk, Tape
- SCSI IPL

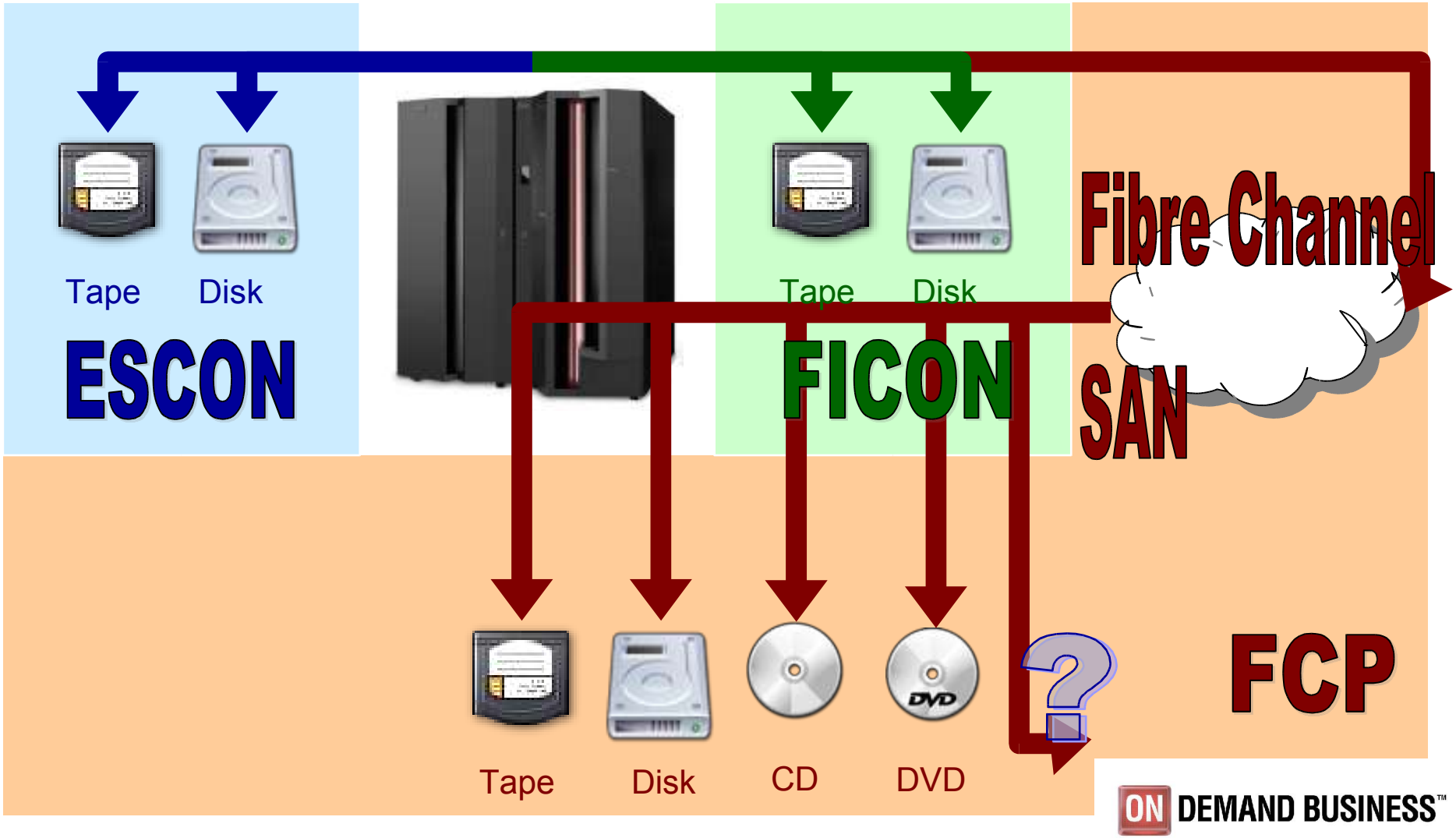
zSeries in a SAN – Sharing Storage Resources

zSeries in a SAN – Sharing Storage Resources

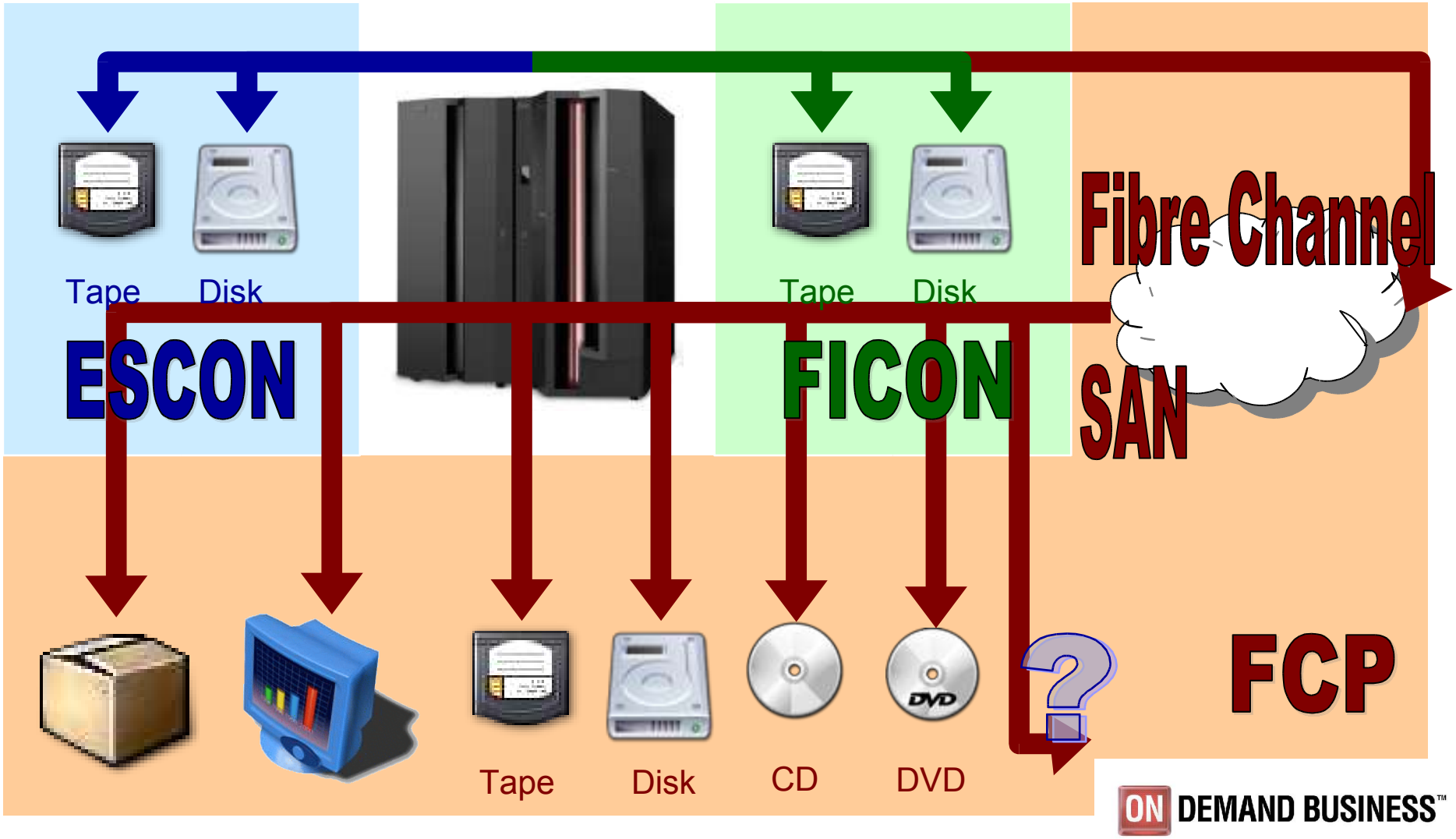
zSeries in a SAN – Sharing Storage Resources



zSeries in a SAN – Sharing Storage Resources



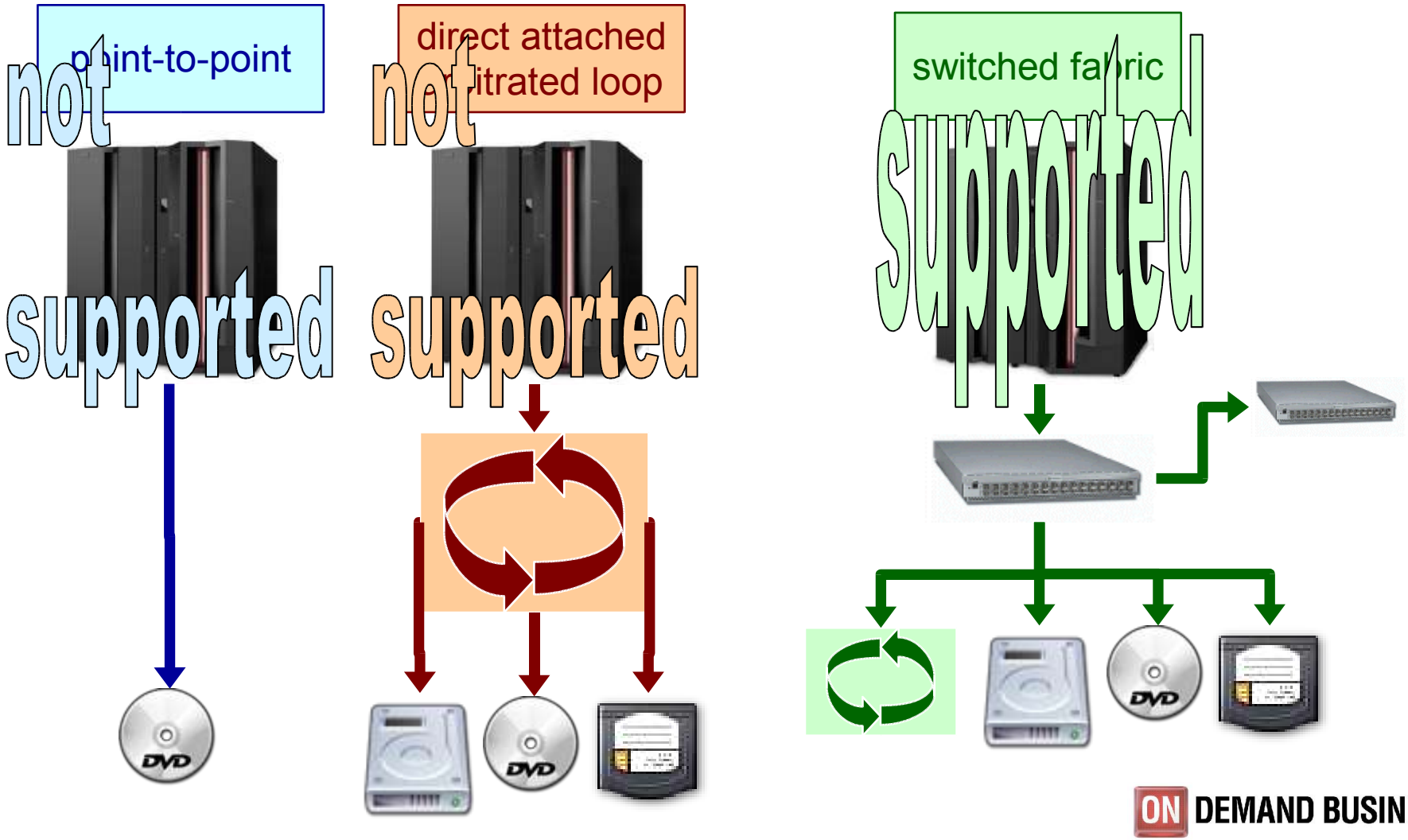
zSeries in a SAN – Sharing Storage Resources



zSeries in a SAN – Hardware Requirements

- IBM zSeries 800, 890, 900 or 990
- FICON or FICON Express adapter card
- Additional CHPID type FCP
- FC fabric switch
- FC attached storage devices
- Optional: FCP-SCSI bridge
  + SCSI devices

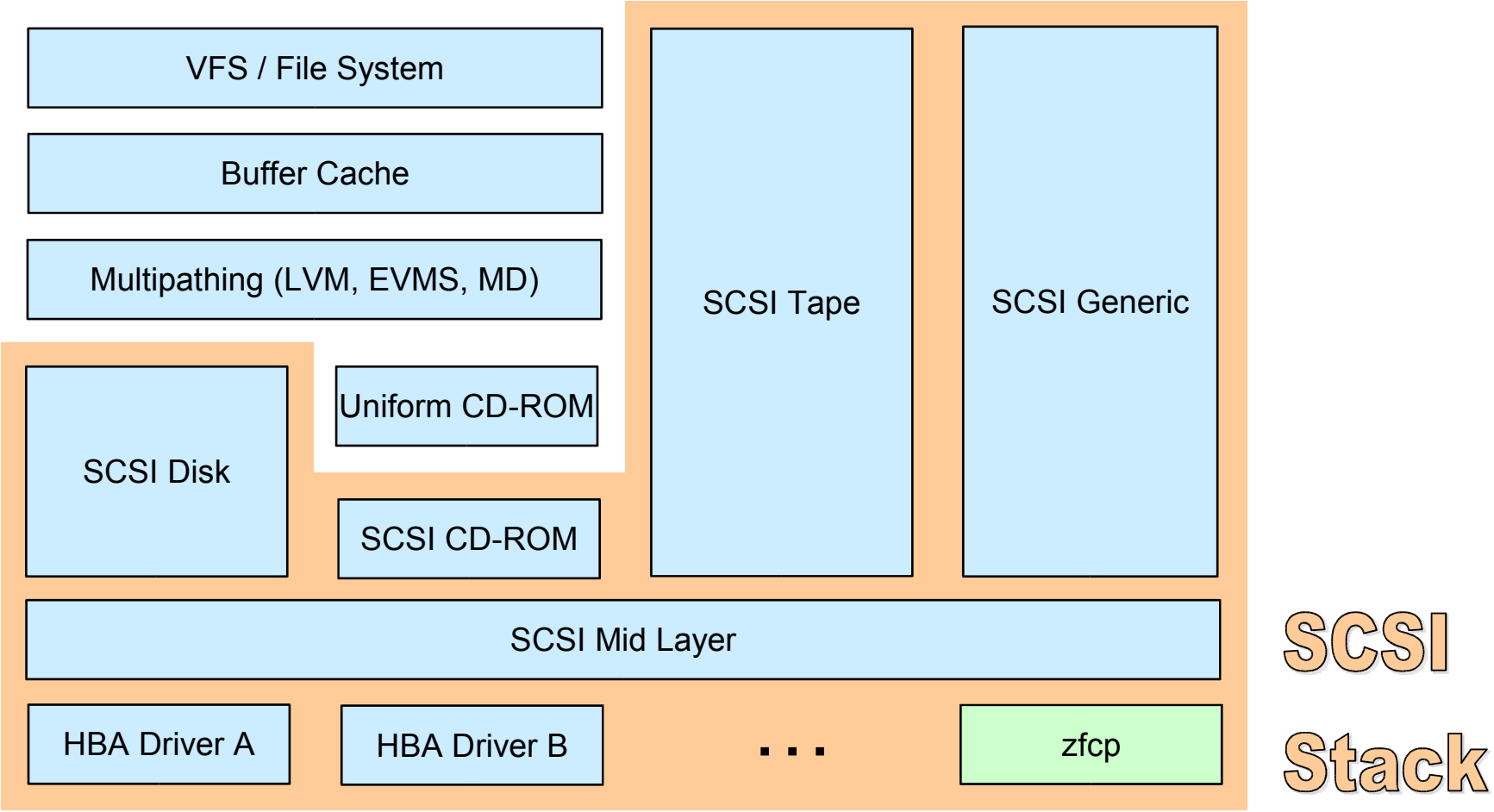
zSeries in a SAN – Topologies



FC and SCSI – Software Requirements

- SUSE Linux Enterprise Server 8 (SLES8)
  - GA November 2002
  - Currently SP3 (submarine)
- SUSE Linux Enterprise Server 9 (SLES9)
  - GA August 2004
    - Currently SP1
- Red Hat Enterprise Linux 3 (RHEL3)
  - GA October 2003
    - Update 4 and RHEL4 are also available
- z/VM 4.3
  - GA May 2002
  - Includes FCP channel guest support for Linux
  - Currently z/VM 5.1

Linux SCSI Stack



zfcp's Task in the Linux SCSI Stack

- zfcp drives the zSeries 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.

SAN Addressing

Device Number

(devno)

e.g. 0x5588

Worldwide Port Name

(WWPN)

e.g. 0x5005076300c693cb

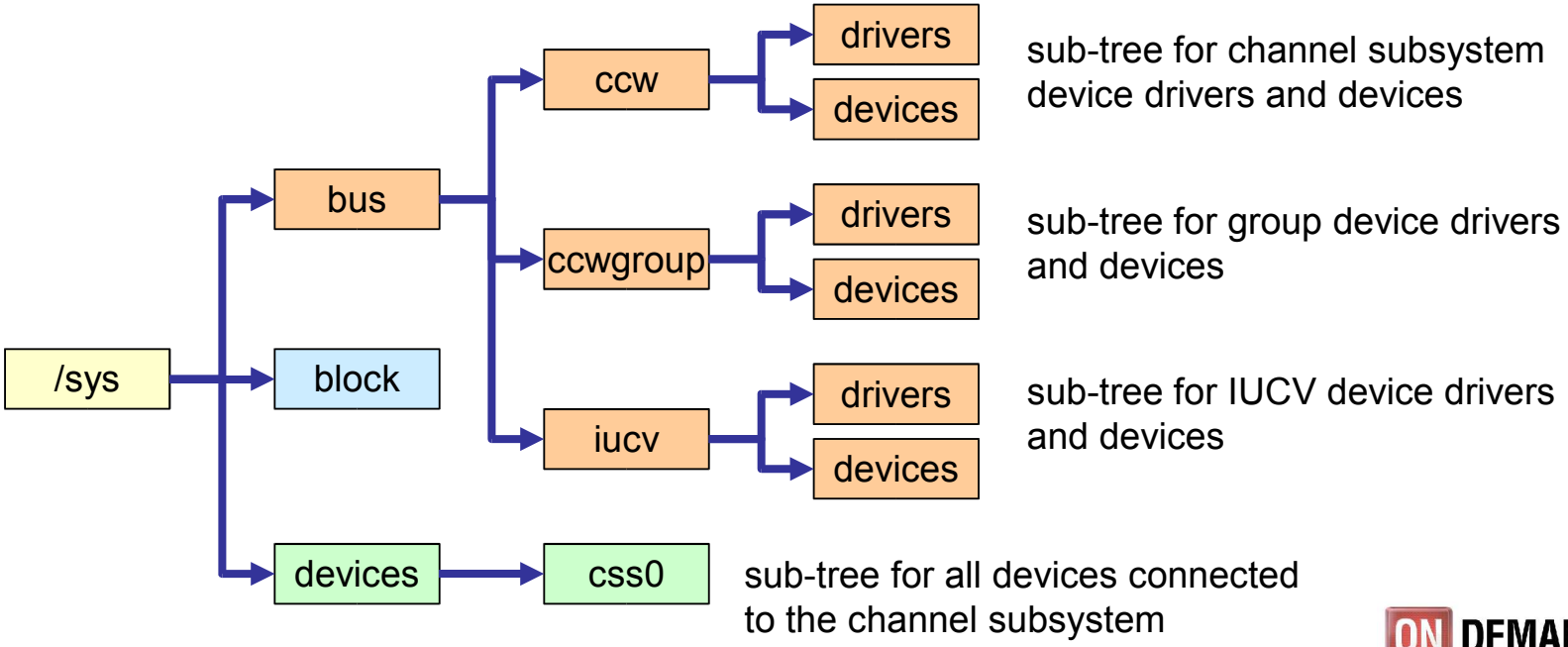
Logical Unit Number

(LUN)

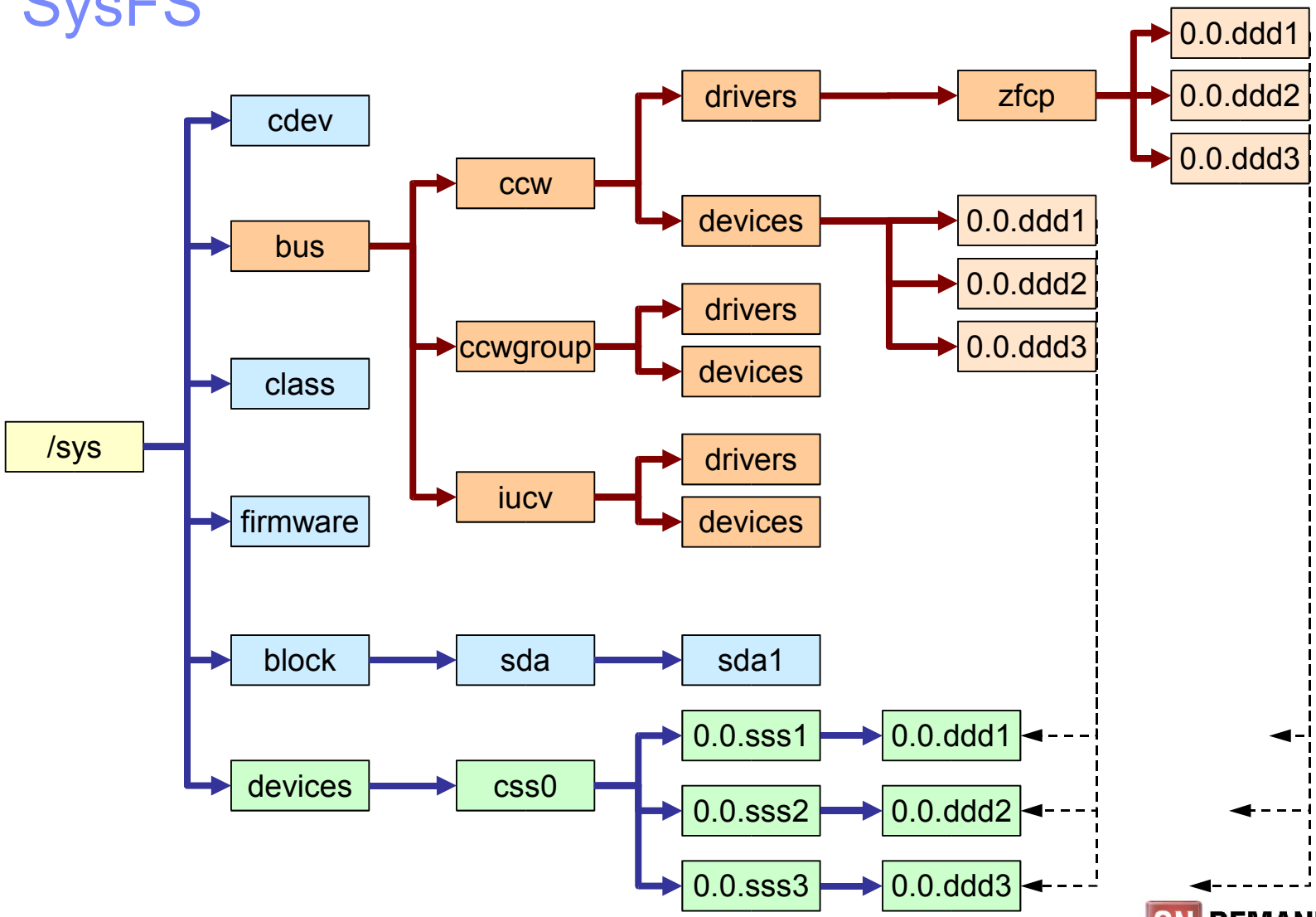
e.g. 0x5125000000000000

SysFS

- 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



Configuration – Set Adapter Online

```
[root: root]# cd /sys/bus/ccw/drivers/zfcp/
[root: zfcp]# ls
0.0.5588 loglevel_cio loglevel_config loglevel_erp
loglevel_fc loglevel_fsf loglevel_other
loglevel_qdio loglevel_scsi version
[root: zfcp]# cd 0.0.5588/
[root: 0.0.5588]# ls
availability card_version cmb_enable cutype detach_state
devtype failed fc_link_speed fc_service_class fc_topology
hardware_version in_recovery lic_version online port_add
port_remove s_id scsi_host_no serial_number status wwnn
wwpn


[root: 0.0.5588]# cat online
0
[root: 0.0.5588]# echo 1 > online
[root: 0.0.5588]# cat online
1
```

Configuration – Add a Port to the Adapter

```
[root: 0.0.5588]# ls
availability card_version cmb_enable
… port_add … status wwpn
[root: 0.0.5588]# echo 0x5005076300c693cb > port_add
[root: 0.0.5588]# ls
0x5005076300c693cb availability card_version
cmb_enable cutype detach_state devtype failed
fc_link_speed fc_service_class fc_topology
hardware_version host0 in_recovery lic_version
nameserver online port_add port_remove s_id
scsi_host_no serial_number status wwnn wwpn
[root: 0.0.5588]# cd 0x5005076300c693cb
[root: 0x5005076300c693cb]# ls
d_id detach_state failed in_recovery scsi_id
status unit_add unit_remove wwnn
```

Configuration – Add a Unit to the Port

```
[root: 0x5005076300c693cb]# ls
d_id detach_state failed in_recovery scsi_id
status unit_add unit_remove wwnn
[root: 0x5005076300c693cb]# echo 0x5125000000000000 > unit_add
[root: 0x5005076300c693cb]# ls
0x5125000000000000  d_id  detach_state  failed
in_recovery  scsi_id  status  unit_add  unit_remove
wwnn
[root: 0x5005076300c693cb]# cd 0x5125000000000000/
[root: 0x5125000000000000]# ls
detach_state  failed  in_recovery  scsi_lun  status


[root: 0x5125000000000000]# lsscsi
[0:0:1:0]  disk   IBM   2105F20    .693  /dev/sda
[root: 0x5125000000000000]#
```

Block Device View

```
[root: root]# cd /sys/block/
[root: block]# ls
dasda dasdb loop0 loop1 loop2 loop3 loop4 loop5
loop6 loop7 ram0 ram1 ram2 ram3 ram4 ram5 ram6 ram7
ram8 ram9 ram10 ram11 ram12 ram13 ram14 ram15 sda
[root: block]# cd sda
[root: sda]# ls
dev  device  queue  range  sda1 size  stat


[root: sda]# cat dev
8:0
[root: sda]# cat range
16
[root: sda]# cat size
3906304
```

FCP – SCSI Mapping

FCP World		SCSI World
HBA	→	Host
0	→	Bus
WWPN	→	SCSI ID
FCP LUN	→	SCSI LUN

SCSI View

```
[root: root]# cd /sys/bus/scsi/devices/
[root: devices]# ls
0:0:1:0
[root: devices]# cd 0\:0\:1\:0
[root: 0:0:1:0]# ls
block delete detach_state device_blocked
fcp_lun generic hba_id model online
queue_depth rescan rev scsi_level type
vendor wwpn
```

FCP – SCSI Mapping

```
[root: root]# cd /sys/bus/ccw/drivers/zfcp/0.0.5588/
[root: 0.0.5588]# cat scsi_host_no
0x0
[root: 0.0.5588]# cd 0x5005076300c693cb
[root: 0x5005076300c693cb]# cat scsi_id
0x1
[root: 0x5005076300c693cb]# cd 0x5125000000000000
[root: 0x5125000000000000]# cat scsi_lun
0x0
```

```
[root: root]# cd /sys/bus/scsi/devices/0\:0\:1\:0/
[root: 0:0:1:0]# cat hba_id
0.0.5588
[root: 0:0:1:0]# cat wwpn
0x5005076300c693cb
[root: 0:0:1:0]# cat fcp_lun
0x5125000000000000
```

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 - Fibre Channel topology
- fc_link_speed - Link Speed

```
# cd /sys/bus/ccw/drivers/zfcp/0.0.5588/
# cat serial_number
IBM0200000001AB8A
# cat lic_version
0x00000206
# cat scsi_host_no
0x0
```

```
# cat wwnn
0x5005076400c1ab8a
# cat wwpn
0x5005076401602fd8
# cat fc_topology
fabric
# cat fc_link_speed
2 Gb/s
```

Port Information

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

```
# cd /sys/bus/ccw/drivers/zfcp/0.0.5588/0x5005076300c693cb/
# ls
0x5125000000000000 d_id detach_state failed in_recovery
scsi_id status unit_add unit_remove wwnn
# cat in_recovery
0
# cat scsi_id
0x1
# cat d_id
0x632e13
```

Unit Information

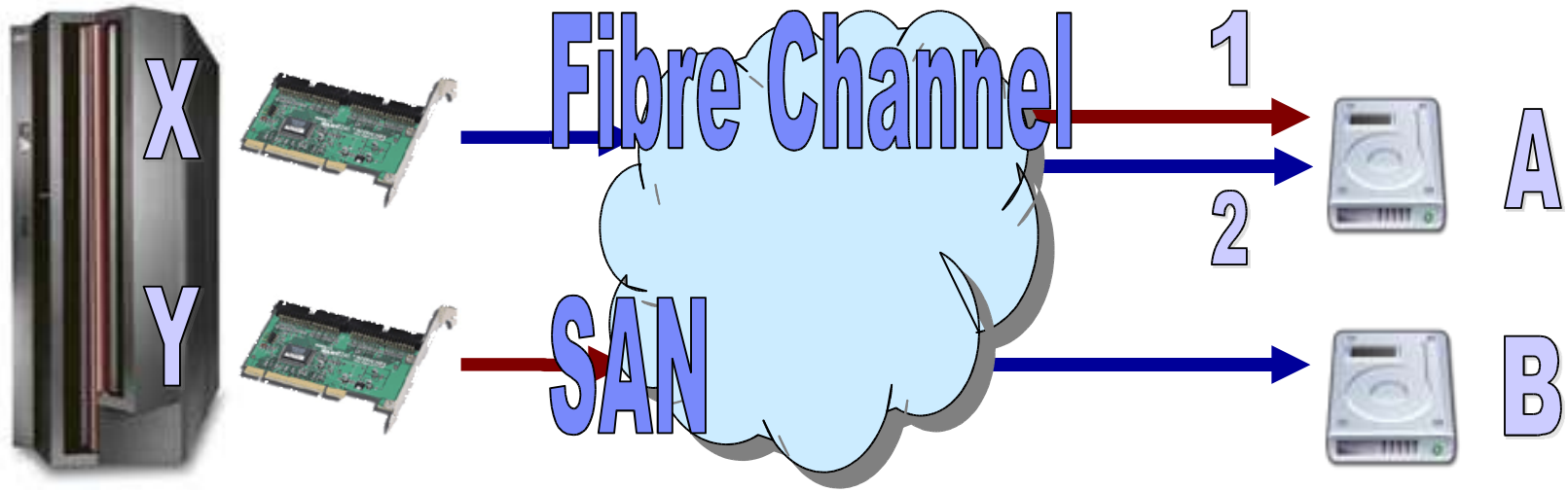
- in_recovery - Recovery status
- scsi_lun - Linux SCSI LUN
- failed - Unit error recovery status
- status - Debug Info

```
# cd /sys/bus/ccw/drivers/zfcp/0.0.5911/0x5005076300cc0b8e/0x5108000000000000/
# ls
detach_state  failed  in_recovery  scsi_lun  status
# cat failed
0
# cat in_recovery
0
# cat scsi_lun
0x0
# cat status
0x60000000
```

FCP Multipathing

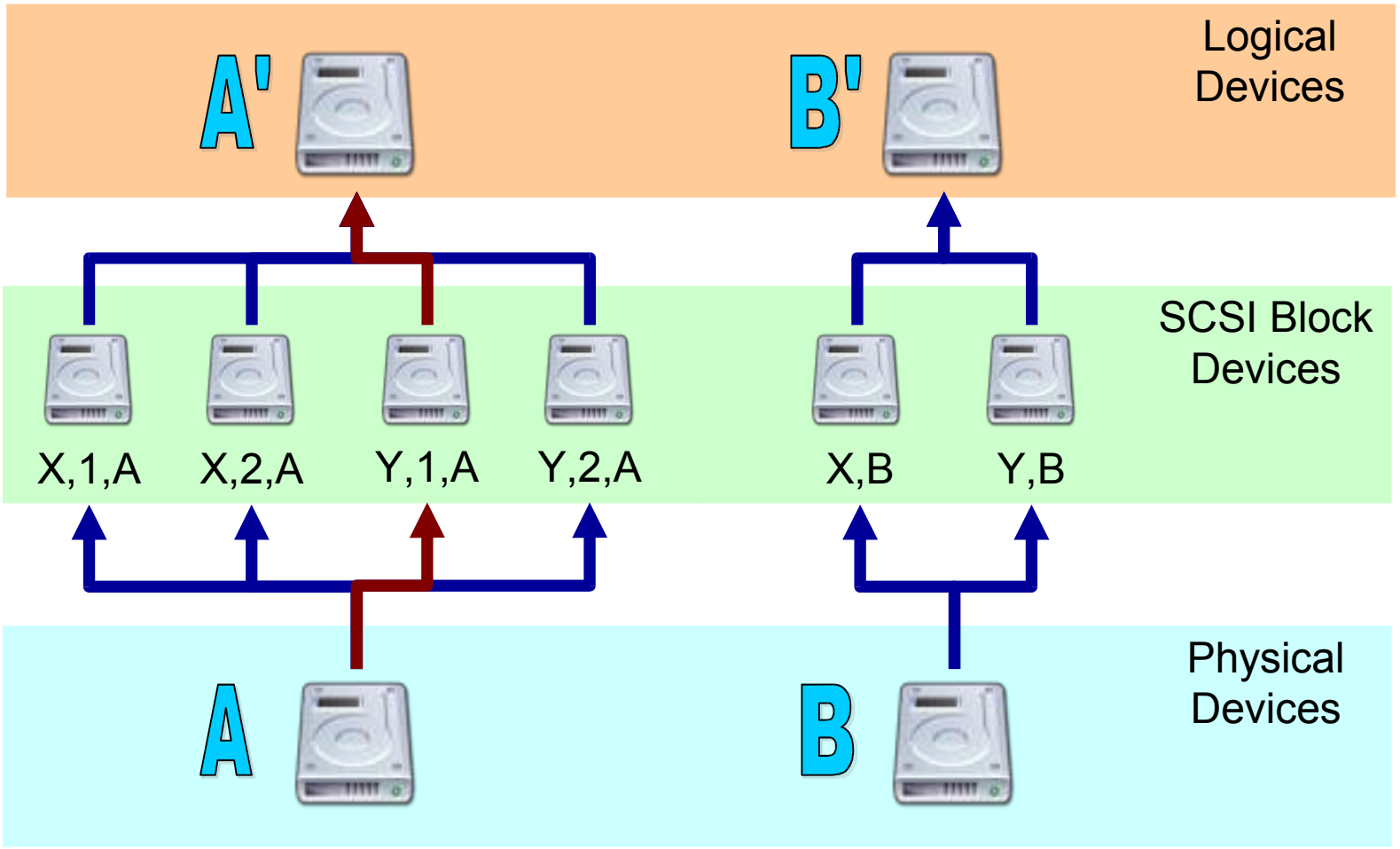
- SLES8
  - LVM – Logical Volume Manager
- SLES9
  - Device Mapper subsystem in 2.6 kernel
  - EVMS – Enterprise Volume Management System
  - LVM2 – Logical Volume Manager
- RHEL3
  - MD
  - mdadm – multiple device administration

FCP Multipathing



- Failover on path-failure
- Failback if recovered path is detected (retries)
- Load balancing (use of multiple paths for concurrent I/Os according to assigned priorities)
- Designed to cover all block devices

FCP Multipathing – Devices

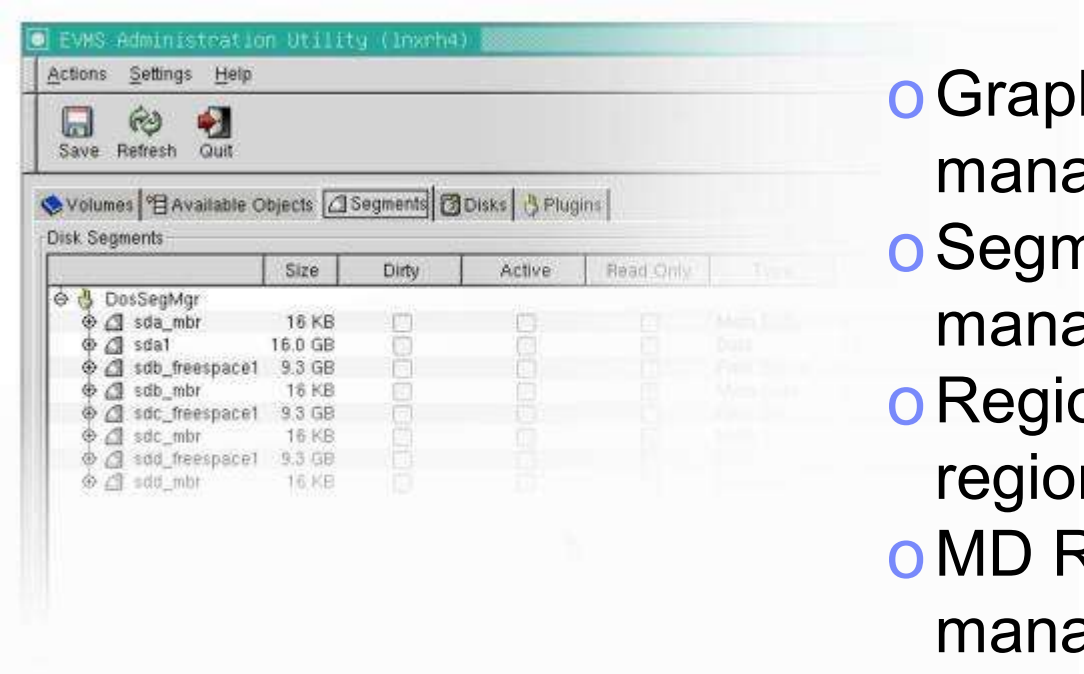


FCP Multipathing – LVM

- Notations
  - Physical volumes
  - Logical volumes
  - Volume groups
- /etc/zfcp.conf
- Only one path enabled by default
- /proc/lvm/

- Standard LVM commands
  - pvcreate
  - vgcreate
  - vgdisplay
  - lvcreate
- Multipath LVM commands
  - pvpath
  - pvpathsave
  - pvpathrestore

FCP Multipathing – EVMS



- Graphical EVMS management tool
- Segment, segment manager
- Region and MD multipath region manager
- MD Raid 0 Region manager

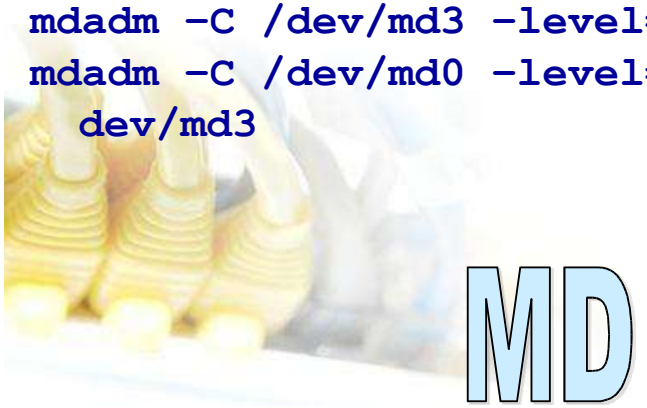
ON DEMAND BUSINESS™

FCP Multipathing – MD

- No load balancing
- Primary – secondary path or actual path – spare path
- Attention: md subsystem is quite verbose

- FCP mapping in modules.conf on ramdisk (single line!)
- Create device nodes (mknod /dev/sda b 8 0)
- Configure mdadm (/etc/mdadm.conf)
- /etc/rc.d/rc.sysinit – enabling on Linux startup

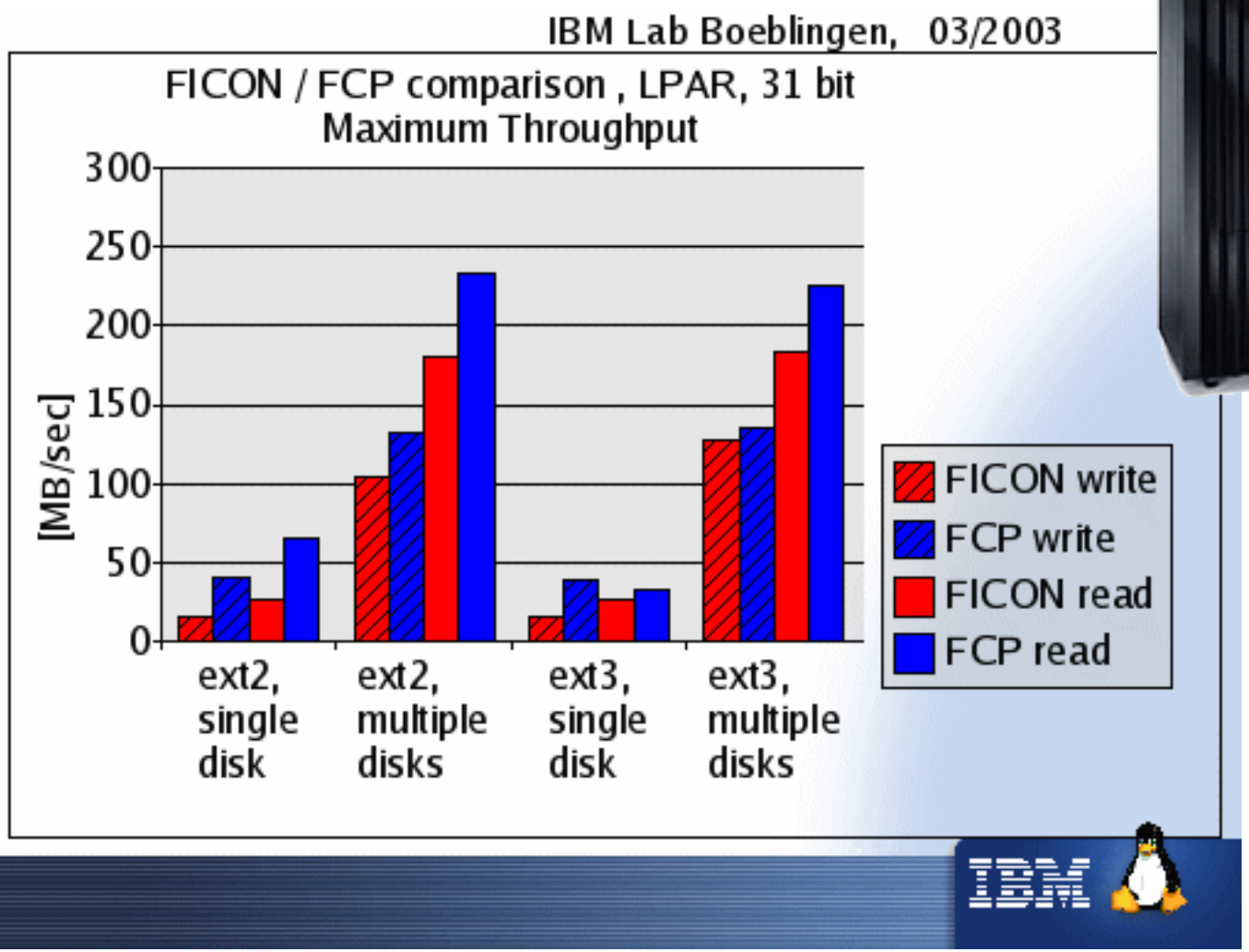
```
mdadm -C /dev/md1 -level=multipath -raid-device=2 /dev/sda1 /dev/sdd1
mdadm -C /dev/md2 -level=multipath -raid-device=2 /dev/sdb1 /dev/sde1
mdadm -C /dev/md3 -level=multipath -raid-device=2 /dev/sdc1 /dev/sdf1
mdadm -C /dev/md0 -level=raid0 -raid-devices=3 /dev/md1 /dev/md2 /
  dev/md3
```



Disk Usage – ECKD and SCSI Comparison

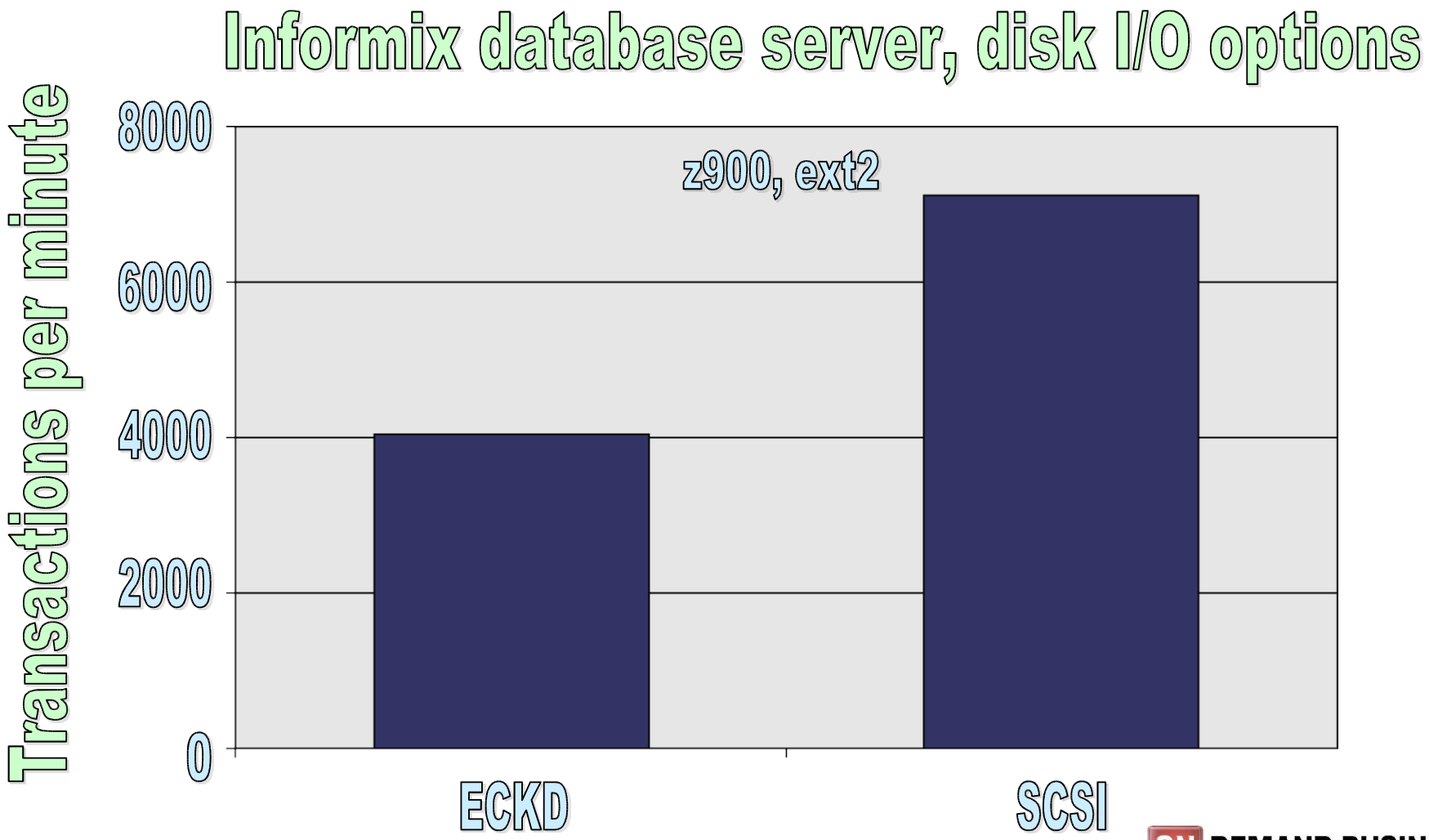
	ECKD DASD	SCSI Disk
Configuration	IOCDS/VM (Operator)	IOCDS/VM & Linux (Operator & Linux Admin)
Access Method	SSCH	QDIO
Block Size (Byte)	512, 1K, 2K, 4K	512
Disk Size	3390 Model 3/9/27 now variable too	any
Formatting (low level)	dasdfmt	not necessary
Partitioning	fdasd	fdisk
File System	mke2fs (or others)	
Access	mount	

FICON and FCP Performance



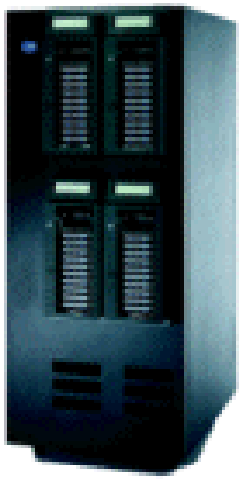
ON DEMAND BUSINESS™

OLTP Workload Informix – I/O Options



FCP/SCSI Tape Support

- Tape Devices:
  - IBM TotalStorage Enterprise Tape System 3590.
  - IBM TotalStorage Enterprise Tape Drive 3592.
  - IBM TotalStorage Enterprise Tape Library 3494.
  - IBM TotalStorage UltraScalable Tape Library 3582, 3583 and 3584 w/ Ultrium 2 Fibre Channel Tape Drives.



- IBMtape and IBMtapeutil packages required
  - /lib/modules/*(Your system's kernel name)*/kernel/drivers/scsi/IBMtape.o
  - /usr/bin/IBMtapeconfig
  - /usr/bin/IBMtaped
  - /usr/bin/IBMtapeutil

FCP/SCSI Tape Support

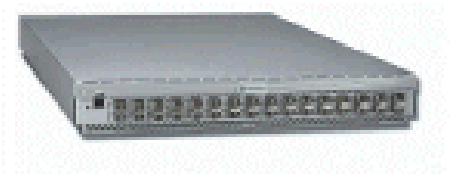
- IBMtape special files (created by IBMtapeconfig):
  - /dev/IBMtape0
  - /dev/IBMtape0n
  - /dev/IBMchanger0

- Tape utility program (IBMtapeutil):

```
# Mount cartridge from slot 3
IBMtapeutil -f /dev/IBMchanger0 mount 3

# Backup myfile.tar to tape
IBMtapeutil -f /dev/IBMtape0 write -s myfile.tar
```

Device Support - Summary



- Devices (via switch)
  - IBM TotalStorage Enterprise Tape System 3590
  - IBM TotalStorage Enterprise Tape Drive 3592
  - IBM TotalStorage Enterprise Tape Library 3494
  - IBM TotalStorage Enterprise Storage Server Models 750, 800, F20, F10
  - IBM TotalStorage UltraScalable Tape Library 3582, 3583 and 3584 w/ Ultrium 2 Fibre Channel Tape Drives
- Director/Switch Support
  - CISCO MDS 9000 Family (IBM 2062)
  - CNT (INRANGE) FC/9000 64-port, 128-port and 256-port models (IBM 2042)
  - McDATA Intrepid 6064 (IBM 2032) and 6140 (IBM 2032)
  - McDATA 3232 (IBM 2031-232)
  - McDATA Sphereon 4500 Fabric Switch (IBM 2031-224)
  - IBM total Storage SAN Switch 2109-M12, 2109-F16 and S16/S08
  - IBM 2108-G07 SAN Data Gateway (parallel SCSI connectivity to non-IBM storage)
  - McDATA ES-1000 Loop Switch (IBM 2031-L00) FCP-to-FC-AL Bridge
  - McDATA ED-5000 (IBM 2032-001)

SCSI IPL & SCSI Dump

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

- Expand the world of open I/O attachments on zSeries from pure data access to allow IPL and Dump support.
- Enhances the setup to allow Linux on zSeries to run completely on SCSI disks - incl. IPL, Data access and Dump support.

- New set of IPL parameters.
- LPAR and z/VM guests supported.

SCSI IPL & SCSI Dump – Cont.

- Disk preparation with Linux „zipl“ tool
- Up to 31 boot configurations possible
- Requirements
  - Requires enablement by FC9904
  - Requires FCP channels
  - IBM zSeries server 800, 890, 900 or 990
  - z/VM 4.4 (PTF UM30989)

ON DEMAND BUSINESS™

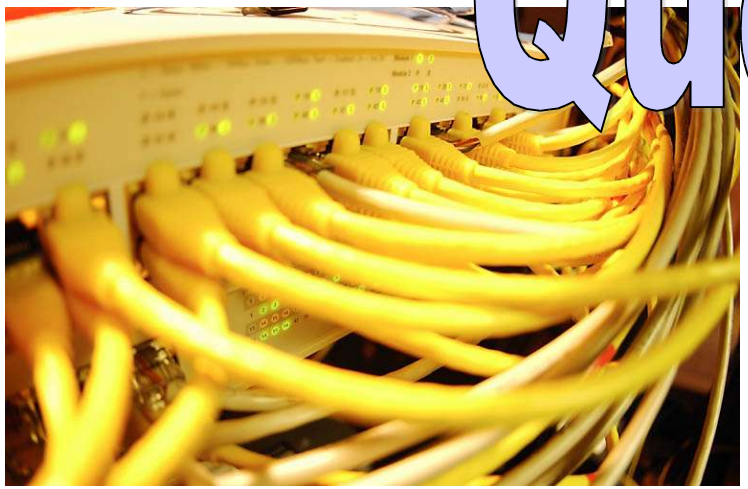
FCP/SCSI on Linux For zSeries - Summary

- FCP/SCSI support for IBM zSeries.
  - New FCP channel based on FICON / FICON Express cards.
  - FCP channel support in z/VM 4.3 and higher for Linux guests.
  - First FCP/SCSI exploitation for zSeries in SLES8 and RHEL3.
- Integration of your zSeries into standard based FC SANs.
- New device types.
- Reduced emulation overhead in OS and ESS compared to ECKD due to native use of fixed block I/O.
- Larger disks in comparison to ESCON/FICON.
- Current restrictions:
  - Only switched fabric supported.
  - No LUN sharing or zoning on a single adapter → use separate physical adapters.

Useful Links

- I/O Connectivity on IBM zSeries mainframe servers
  - http://www-1.ibm.com/servers/eserver/zseries/connectivity/#fcp
- Getting Started with zSeries Fibre Channel Protocol, IBM Redpaper
  - http://www.redbooks.ibm.com/redpapers/pdfs/redp0205.pdf
- z/VM Version 4 Release 4
  - Version 4.4: http://www.vm.ibm.com/zvm440/
  - Version 5.1: http://www.vm.ibm.com/zvm510/
- SUSE Linux Enterprise Server 8
  - http://www.suse.de/de/business/products/server/sles/index.html
- Linux for zSeries and S/390
  - Kernel 2.4: http://awlinux1.alphaworks.ibm.com/developerworks/linux390/june2003_recommended.shtml
  - Kernel 2.6: http://awlinux1.alphaworks.ibm.com/developerworks/linux390/april2004_recommended.shtml
- Linux Device Drivers and Installation Commands
  - Kernel 2.4: http://awlinux1.alphaworks.ibm.com/developerworks/linux390/docu/lx24jun03dd02.pdf
  - Kernel 2.6: http://awlinux1.alphaworks.ibm.com/developerworks/linux390/docu/lx26apr04dd00.pdf
- IBM TotalStorage Tape Device Drivers – Installation and User‘s Guide
  - ftp://ftp.software.ibm.com/storage/devdrvr/Doc/
- ESS Fibre Channel Attachment White Paper
  - http://www.storage.ibm.com/disk/ess/support/essfcwp.pdf

SCSI over Fibre Channel Support for Linux on zSeries



Questions ?

Trademarks

- **The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.**
  - AIX, e-business logo, on-demand logo, IBM, IBM logo, OS/390, PR/SM, z900, z990, z800, z890, zSeries, S/390, z/OS, z/VM, FICON, ESCON
- **The following are trademarks or registered trademarks of other companies.**
  - LINUX is a registered trademark of Linus Torvalds
  - Penguin (Tux) complements of Larry Ewing
  - Tivoli is a trademark of Tivoli Systems Inc.
  - Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries
  - UNIX is a registered trademark of The Open Group in the United States and other countries.
  - SMB, Microsoft, Windows are registered trademarks of Microsoft Corporation.
- * 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 Germany. 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.

ON DEMAND BUSINESS™