Understanding the Benefits of SCSI for Linux on System z

Session 15996

John Crossno

EMC Corporation

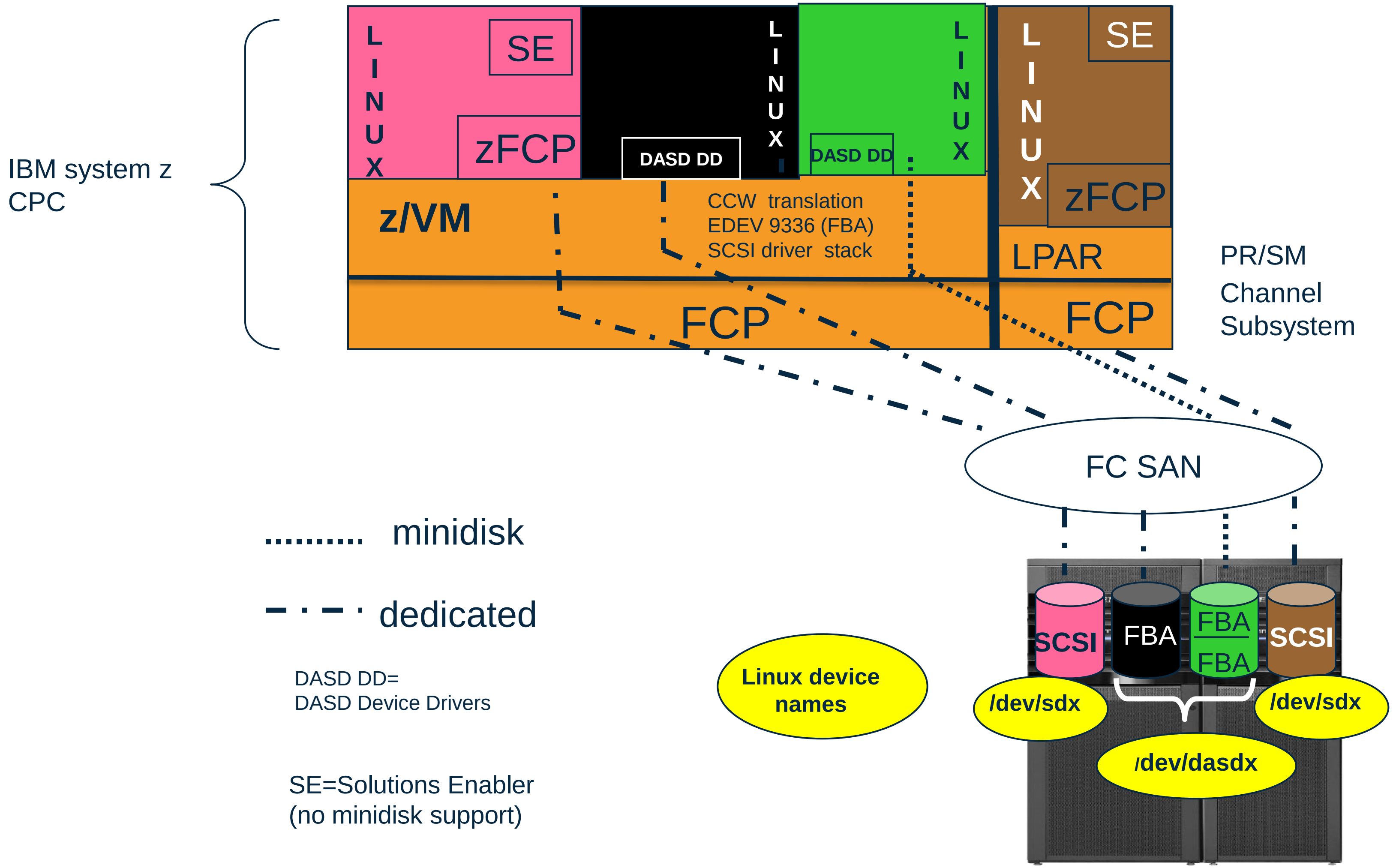
August 6th 2014

Complete your session evaluations online at www.SHARE.org/Pittsburgh-Eval

Objectives

- Examine FBA device attributes
- Look at ease of administration
- Discuss the flexibility of FBA devices
- Explore solutions and innovation with SCSI fiber channel protocol

Linux on System z
FBA Disk Attachment Options

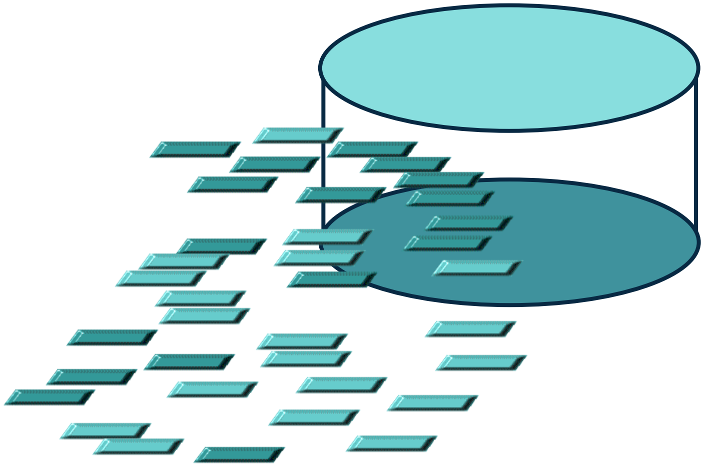


DASD DD=
DASD Device Drivers

SE=Solutions Enabler
(no minidisk support)

Fixed Block Architecture Device Basics

- FBA devices are fixed byte block (512 bytes)
- FBA device size limited by Linux kernel definition
  - Current limitation 2TB maximum
  - Variable device size
- Best use of physical device space

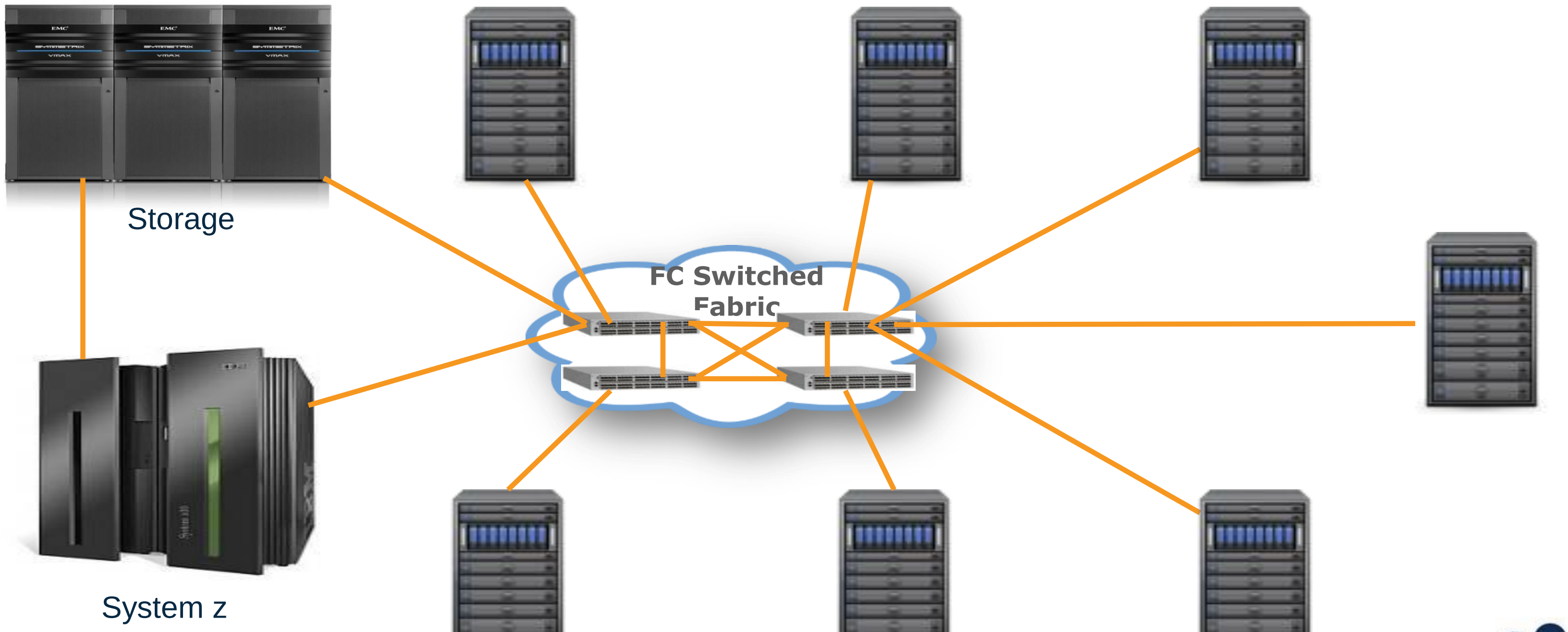


Ease of Administration

- No format is required on a SCSI LUN
- No IOCDS change required
- No additional z/VM changes needed to provision additional SCSI LUNs to a Linux host
    - No directory changes, no additional mdisks
- Utilizes existing SAN infrastructure

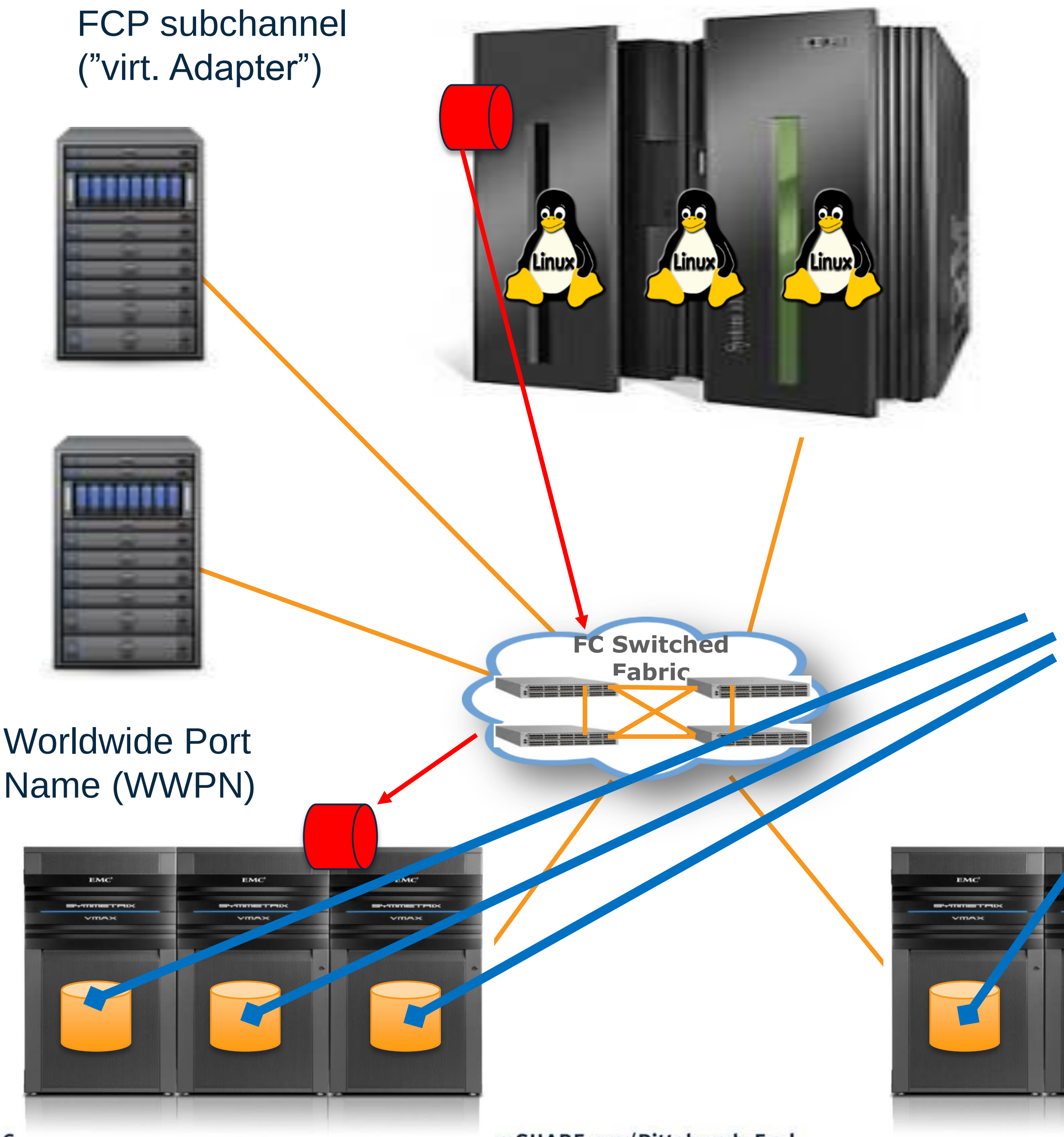
Existing Infrastructure

- Use of existing SAN infrastructure used by open systems
- Use of existing FICON components
  - FICON Express cards
  - FC switches and cabling



Flexibility

- FBA devices can be setup as a SCSI LUN to Linux or defined as a emulated device (edev, 9336) to z/VM
- No matter which setup is used they both communicate to the storage array in SCSI fibre channel protocol
- a SCSI *LUN*, or **logical unit number**, is a number used to identify a **logical unit**, which is a device addressed by the SCSI protocol or protocols which encapsulate SCSI, such as Fibre Channel



Storage devices usually comprise many ***logical units*** - volumes, tape drives, etc.

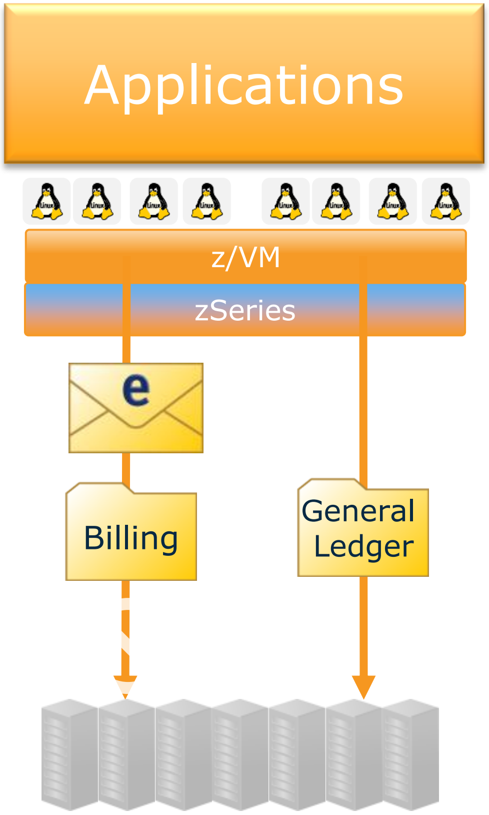
A logical unit is identified by its Fibre Channel Protocol Logical Unit Number (FCP LUN).

FBA as SCSI LUN devices

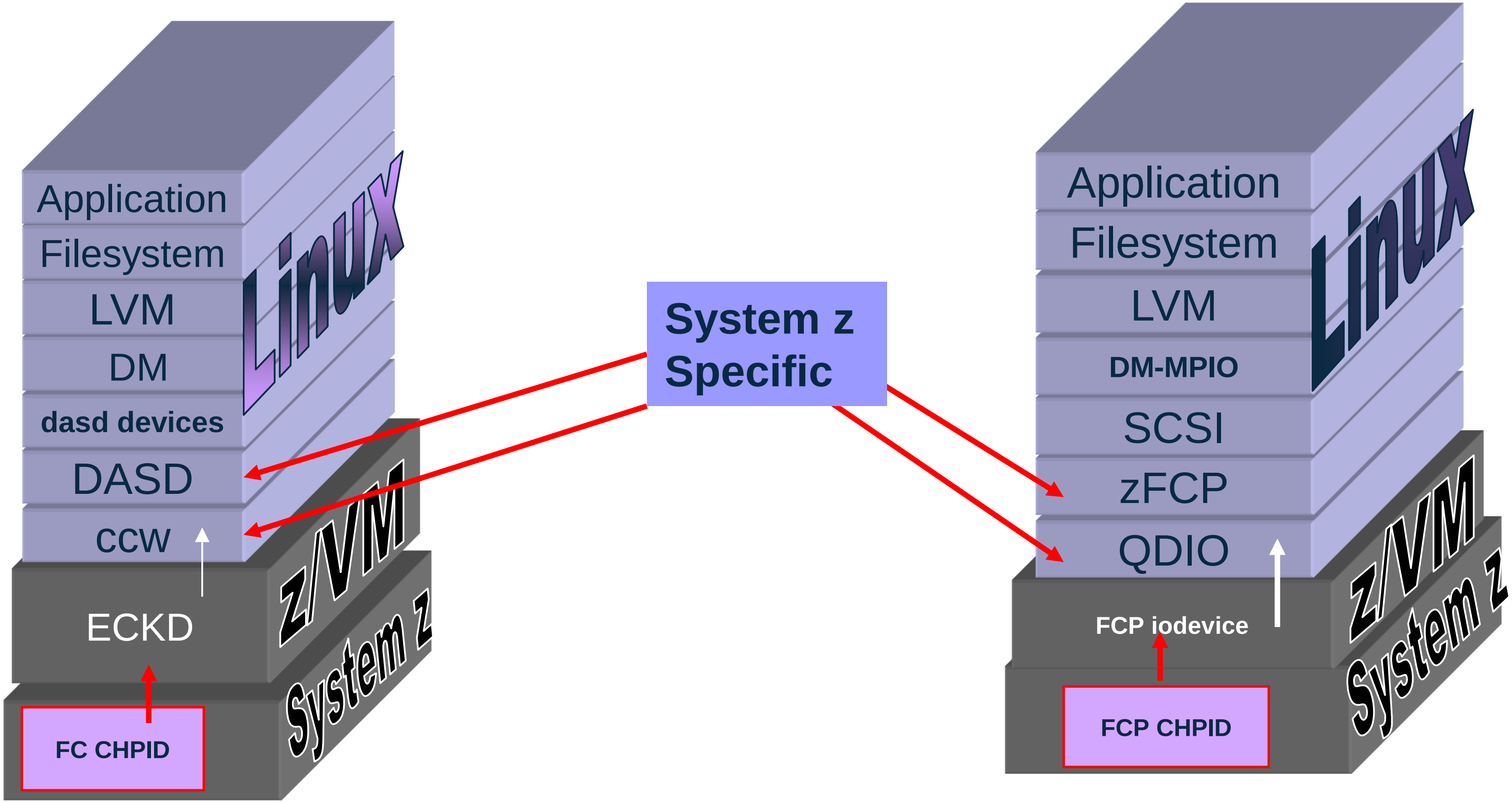
- Provision new FBA devices on storage array
- Dynamic LUN allocation to Linux
- Same protocol as used in open systems environment
- Multipath is handled by Linux on System z
  - Hardware independence
- Many databases utilize SCSI LUN devices
- Ability to exploit open systems features
  - e.g. – DB2 – the *<u>no filesystem caching</u>* option is supported for SCSI LUNs

Multipathing in Linux

- Multiple paths from OS to storage
- Why?
- Implemented in Linux in multipath-tools package, together with the device-mapper in the Linux kernel, or through 3rd party products
- SCSI device (“LUN”) in Linux represents one path to the disk volume on the storage server
- Multipath devices are block devices in Linux



FICON and FCP IO Path



FICON I/O Path

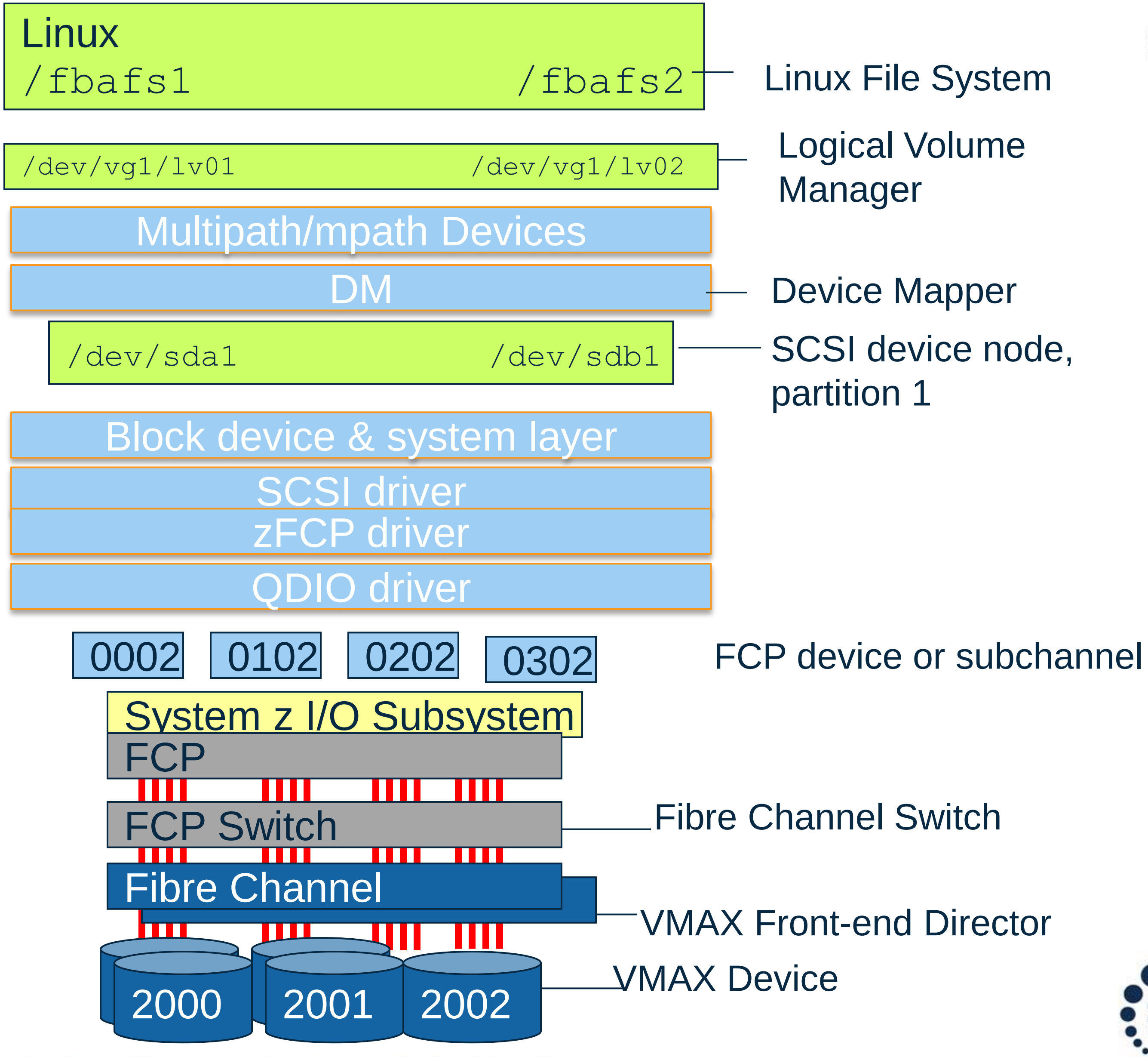
- FICON – no PAV
  - Only one IO can be active on the subchannel and the rest of the IOs need to be queued
- FICON – with HyperPAV
  - Aliases are assigned
  - Each alias is like a subchannel
  - An IO can be active on each subchannel
  - Disk blocksize 4k
  - Serializes IO on each subchannel

SCSI Performance

- There is no emulation overhead
- With SCSI Linux handles IO and errors
  - This is familiar to open systems admin’s
- Multiple IOs can be issued and outstanding
- SCSI uses a customizable field for queuing
  - queue_depth
  - Can be set for each device

```
# lszfcp -l 0x0001000000000000 -a|grep queue_depth
      queue_depth        = "32"
      queue_depth        = "32"
      queue_depth        = "32"
      queue_depth        = "32"
```

default


Linux
/fbafs1
/fbafs2
Linux File System
/dev/vg1/lv01
/dev/vg1/lv02
Logical Volume Manager
Multipath/mpath Devices
DM
Device Mapper
/dev/sda1
/dev/sdb1
SCSI device node, partition 1
Block device & system layer
SCSI driver
zFCP driver
QDIO driver
0002
0102
0202
0302
FCP device or subchannel
System z I/O Subsystem
FCP
FCP Switch
Fibre Channel Switch
Fibre Channel
VMAX Front-end Director
2000
2001
2002
VMAX Device

SCSI Device Driver components

- There are several components that come together to execute SCSI IO
- Using the lsmod command you can see the relationship and other components that are needed in Linux

```
# lsmod|grep zfcp
Module                    Size  Used by
zfcp                    125380  32
scsi_transport_fc        71764  1 zfcp
qdio                     76842  3 qeth_l3,zfcp,qeth
scsi_mod                303205  10
sg,sd_mod,zfcp,scsi_transport_fc,scsi_tgt,scsi_dh_alua
,scsi_dh_hp_sw,scsi_dh_rdac,scsi_dh_emc,scsi_dh
```

FBA as z/VM emulated devices

- Defined in z/VM as 9336 or FB-512 type device
- AKA EDEVs
- Emulation is used at the z/VM and Linux layer
- z/VM communicates to storage array with SCSI fibre channel protocol
- Can be setup as minidisk or direct attached device
- IO handled by Linux and z/VM
- Multipath support handled by z/VM
- Storage can be managed and monitored from z/VM
- Commonly used for Linux OS

Flexibility: Best of Both Worlds

- Mainframe
  - Reliability
  - Availability
  - Serviceability
- Open Systems
  - Open source
  - Worldwide innovation & collaboration
  - Adoption by a community of experts

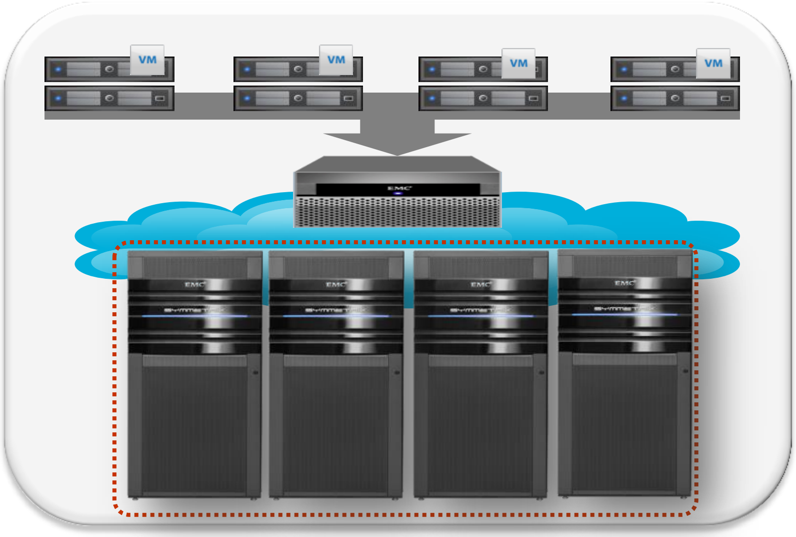
- SCSI continues to evolve…

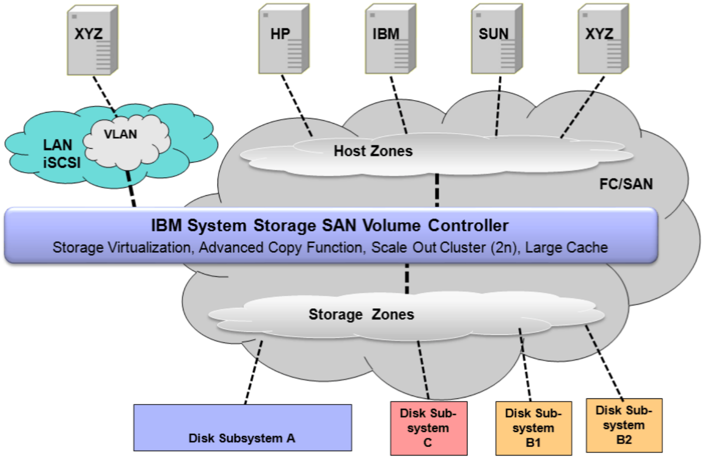
SCSI Innovation

- New host based SCSI commands for thin device cleanup
  - SCSI standard (t10.org) - T10 Technical Committee on SCSI Storage Interfaces
  - SCSI unmap
  - SCSI write same with unmap
  - Support for these SCSI commands are
    - Kernel dependent – Linux vendor and release
    - Storage array dependent

Flexibility

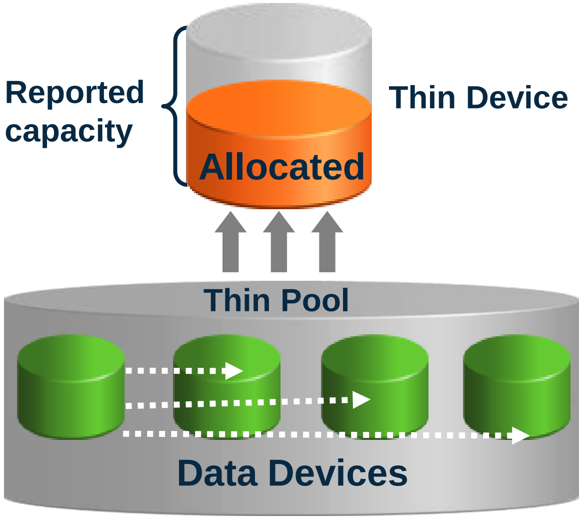
- Ability to exploit open systems solutions
  - Storage virtualization appliances
    - EMC VPLEX, IBM SVC
  - Virtual provisioning or Thin provisioning





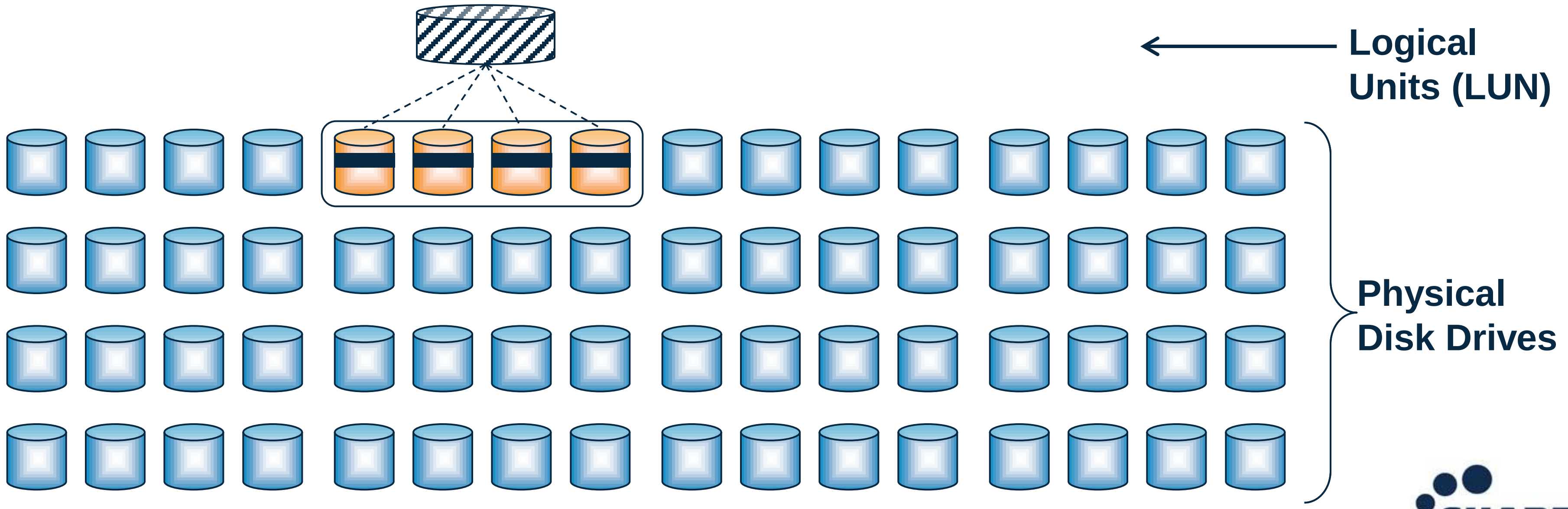
Private Cloud Storage Optimization

- Virtual Provisioning (VP) simplifies Storage Management for FBA
  - Removes data placement requirements from administrators
  - Introduces *thin devices*
  - *Allows for over subscription of storage*



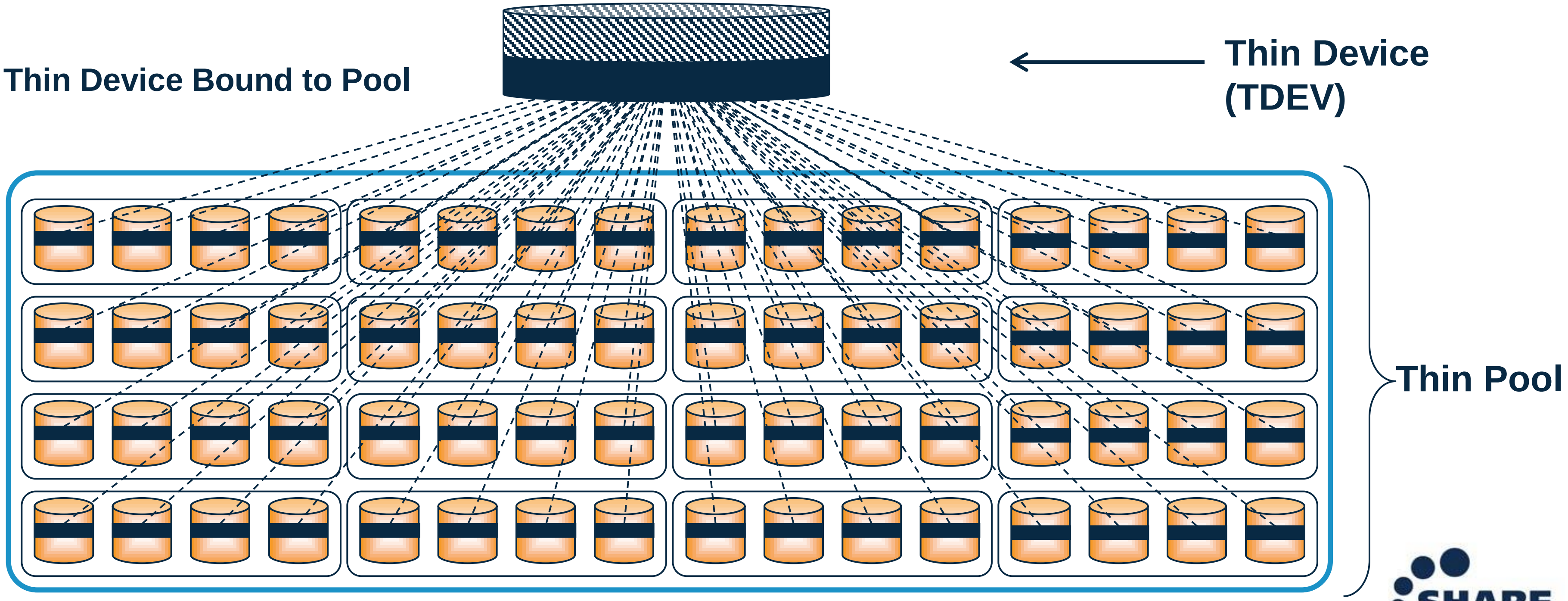
Data Layout – RAID group Allocation

- Capacity for a single logical volume is allocated from a group of physical disks
  - Example: RAID 5 with striped data + parity
- Workload is spread across a few physical disks



Data Layout – Pool-based Allocation Virtual Provisioning

- Storage capacity is structured in pools
- Thin devices are disk devices that are provisioned to hosts



Storage Requirement: Performance

- Storage Layout

Go Wide Before Deep!

- Goal is to spread workload across all available system resources
  - Optimize resource utilization
  - Maximize performance
  - Use what is needed

Thin Provisioning Cleanup for Linux on System z

- SCSI commands
  - Unmap -sent to thin device to unmap (or deallocate) one or more logical blocks
  - Write Same (with unmap flag) - writes at least one block and unmap(s) other logical blocks
- fstrim – executable, batch command used on filesystems
- Discard
  - option on mkfs and mount command for ext4 and xfs filesystems
  - controls if filesystem supports the SCSI unmap command so it can free specific blocks on thin devices at file deletion

Benefits – Why FCP & SCSI

- Performance advantages
  - SCSI continues to evolve in performance
  - Reason 1: asynchronous I/O
  - Reason 2: no emulation overhead
- User definable FBA disk up to 2TB (today)
- Up to 15 partitions (16 minor numbers per device)
- FBA as SCSI LUNs maximize disk space
  - no low-level formatting
- System z integration in existing FC SANs
- Use of existing FICON infrastructure
  - FICON Express adapter cards
  - FC switches / Cabling
  - Storage subsystems
- Dynamic configuration
  - Adding of new LUNs is possible without IOCDS change

Summary

- FBA has best use of physical device space
- SCSI LUNs
  - Can be provisioned rapidly, enabling cloud deployment
  - Is favored for performance
  - Solution innovations

Questions?

EMC²

Johnathan Crossno

VMAX Principal Product Manager
z/VM and Linux on System z

+1 508.249.2246
johnathan.crossno@emc.com

EMC Corporation 176 South Street, Hopkinton, Massachusetts 01748-9103 **www.emc.com**