Bouts of inSANity

And How to Recover

Rick Troth <rickt@velocitysoftware.com>

July 30, 2011

VM and Linux Workshop, Columbus, Ohio

Disclaimer

The content of this presentation is informational only and is not intended to be an endorsement by Velocity Software. (I am speaking only for myself.) The reader or attendee is responsible for his/her own use of the concepts and examples presented herein.

In other words: Your mileage may vary. “It Depends.” Results not typical. Actual mileage will probably be less. Use only as directed. Do not fold, spindle, or mutilate. Not to be taken on an empty stomach. Refrigerate after opening.

When in doubt, ask! Don't believe me? Ask the list! Still in doubt? Try it!

A New Iceberg ... Still Hasn't Thawed

- Compare SAN to other storage, MF and PC
- How does SAN connect to zLinux?
- How does SAN connect to z/VM?
- Some interesting issues
- Conclusions

Bouts of inSANity

- “No matter how you slice it,
  SAN in the mainframe world is painful.”
  -- James Vincent

- “Use a spread sheet
  because it gets real confusing real fast.”
  -- Andy Hartman

Bouts of inSANity

- “SAN in the mainframe world is painful.”
- Which pain and how much to endure, that is the question.
- The key is to have a "Good Way" to *manage* SAN. (herding cats comes to mind)

SAN Pro/Con

Advantages of SAN

- common disk hardware interface for all large systems, not just IBM System z
- opportunity to share disk-resident content across platforms
- common skills and work for storage management staff
- potential for more cost effective data storage options (but why?)

SAN Pro/Con

DISAdvantages of SAN

- much larger I/O addressing required (64 bits)
- uniqueness interferes with V12N and DR
- lack of instrumentation
- management, management, management (the return of the abominable wire monster)

Storage Area Network

What SAN is not …

- SAN is not NAS (Net Attached Storage)
- SAN is not a networked filesystem
  - Not Unix NFS protocol
- SAN is not “mapped drives”
  - Not Windows SMB protocol (not CIFS)

Storage Area Network

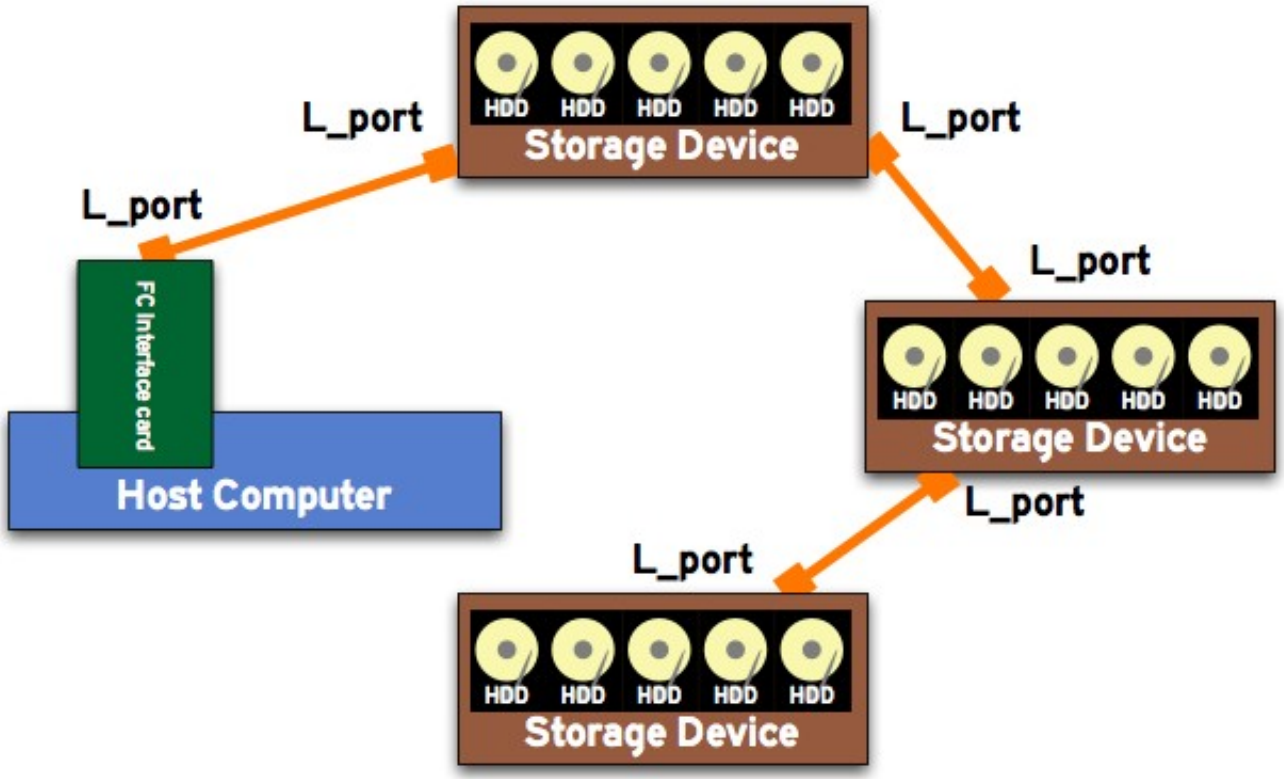
What SAN is …

- External Storage with Long Wires
- Talks like SCSI Disk
- Works like Mainframe Disk (sort of)
- Physically isolated from other networks

SAN and “Normal Disk”

- Topologies
  - Usually Point-to-Point or Multipath Fabric
- Management
  - Pick yer Poison

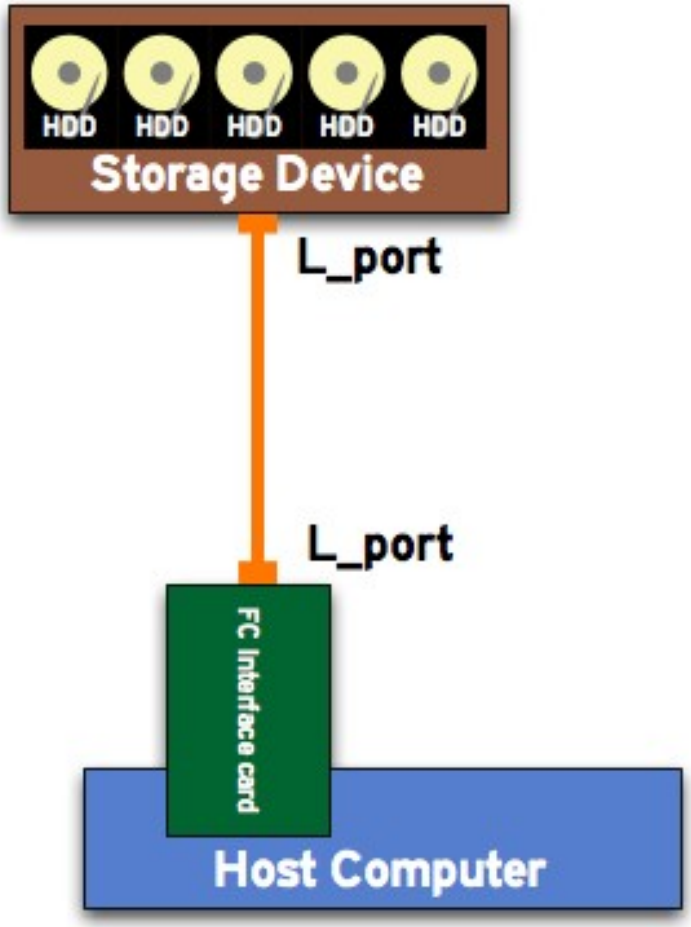
Arbitrated Loop



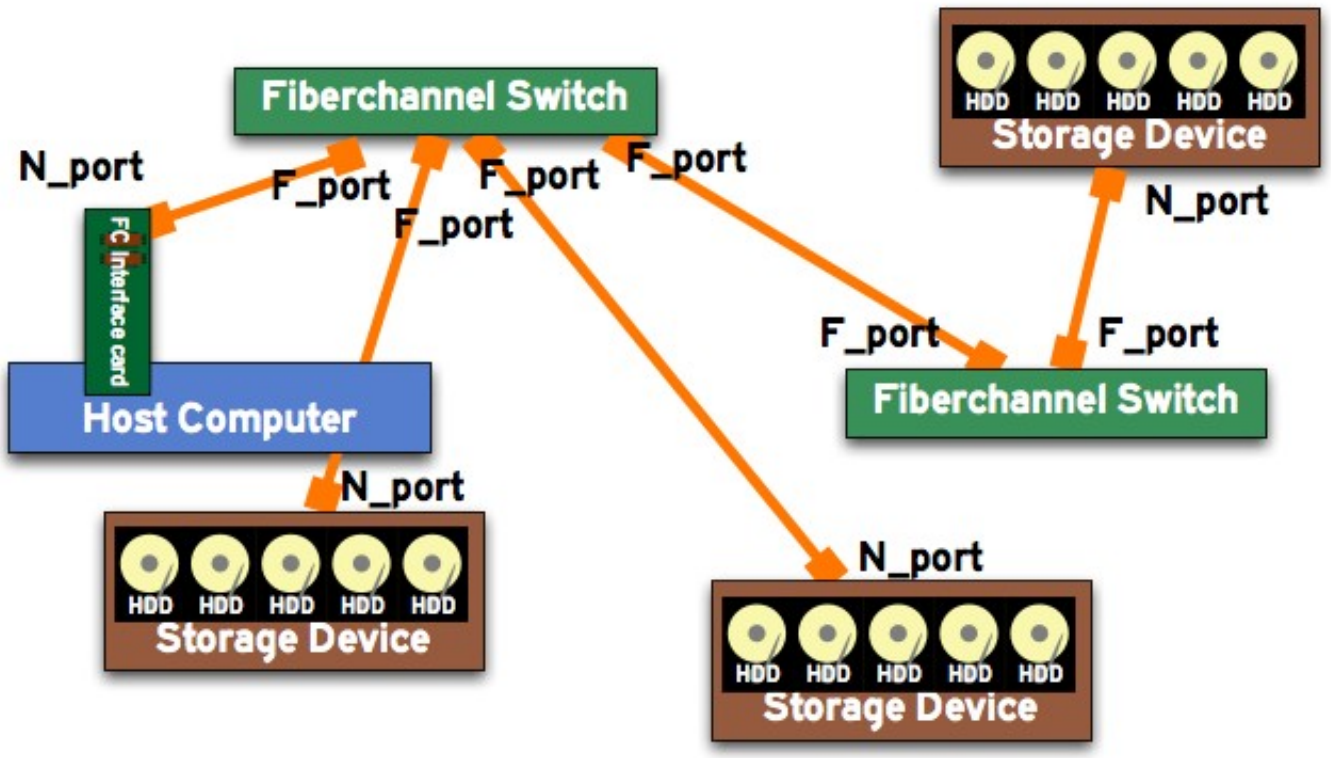
Arbitrated Loop Configuration
Devices are daisy-chained

pascal@southbrain.com

Point to Point



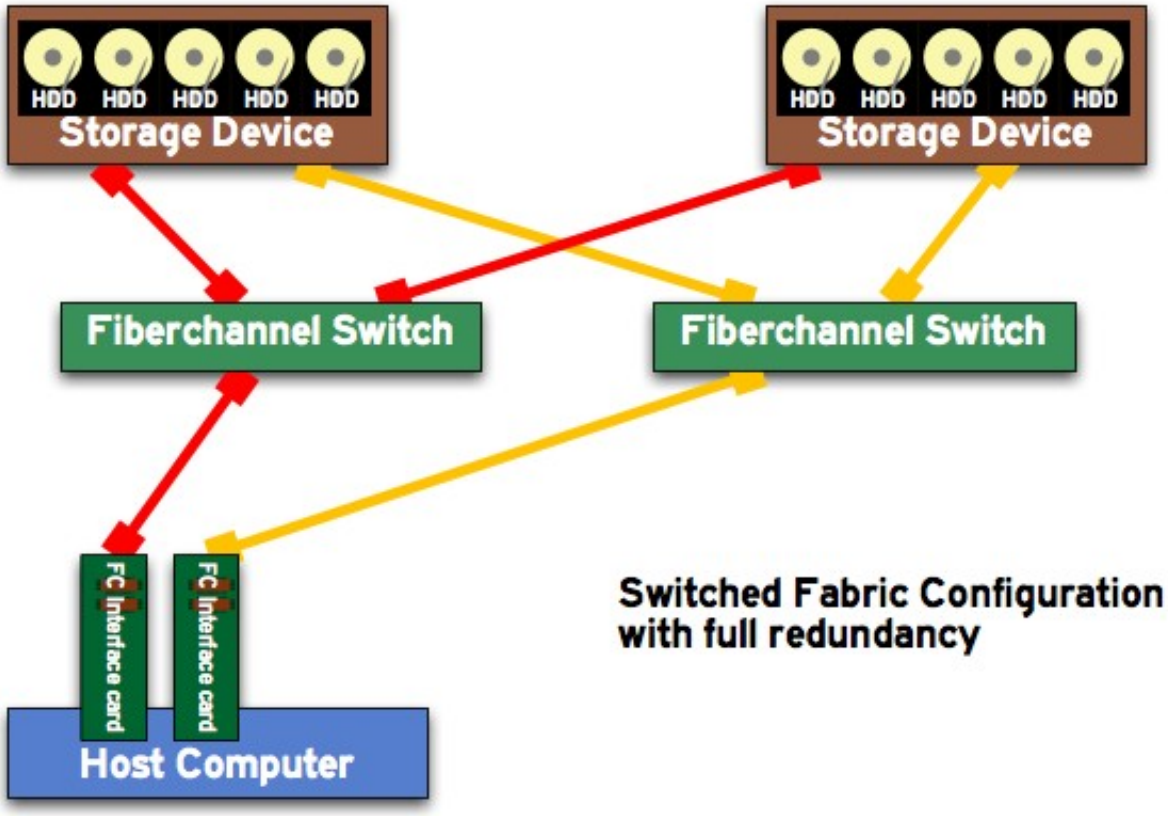
Switched Fabric



Switched Fabric Configuration

pascal@southbrain.com

Switched Fabric with Redundant Paths



Switched Fabric Configuration
with full redundancy

pascal@southbrain.com

SAN with zLinux

- SAN direct to zLinux
  - Dedicated FCP subchannels
  - Multipath support
  - Musical WWPNs
- EDEV to zLinux
  - FBA translation in CP
  - NPIV (for different reasons)

FCP attached to guest

- Requires Multipath Support in Linux
  - Two or more FCP “channels” per guest
- Demands Multipath Management
- Some Loss of Control (over guest storage)
- Coarse Grained Allocations
  - Probably okay if you use LVM
- Should it be N-port Virtualized?

FCP attached to guest – Two HBAs

```
FCP  0100 ON FCP   0304 CHPID D1 SUBCHANNEL = 0018
     0100 DEVTYPE FCP          CHPID D1 FCP
     0100 QDIO ACTIVE          QIOASSIST ACTIVE
…
     WWPN C05076FC7D000D90

FCP  0200 ON FCP   0404 CHPID D5 SUBCHANNEL = 0019
     0200 DEVTYPE FCP          CHPID D5 FCP
     0200 QDIO ACTIVE          QIOASSIST ACTIVE
…
     WWPN C05076FC7D001110
```

FCP attached to guest

```
cd /sys/bus/ccw/drivers/zfcp
echo 1 > $HBA/online
echo $WWPN > $HBA/port_add
echo $LUN > $HBA/$WWPN/unit_add

ls -l $HBA/$WWPN/$LUN/.
```

SAN speak: HBA == FCP adapter

Can You Say “coalesce”?

- Combined 2+ Paths into One PV
  - `http://www.webster.com/dictionary/coalesce`
- “logical volume” in a different sense
  - Physical PV represents an I/O path
  - Logical PV is fed to LVM
- Modify **`/etc/lvm/lvm.conf`** accordingly

Can You Say “coalesce”?

- Modify `/etc/lvm/lvm.conf`:

```
filter = [ "r|^/dev/sd|",
           "r|^/dev/dm|",  …
```

Can You Say “coalesce”?

```
# cat /proc/partitions
  ...
    8      0    35354880 sda
    8     16    35354880 sdb
  253      0    35354880 dm-0
    8     32    35354880 sdc
    8     48    35354880 sdd
  253      1    35354880 dm-1
```

FCP attached to Linux guest

Define paths manually or via YaST, then …

```
/etc/init.d/boot.multipath start
/etc/init.d/multipathd start
pvcreate /dev/mapper/3600604800001901006305330304538 32
vgcreate sanvg1 \
         /dev/mapper/3600604800001901006305330304538 32
lvcreate –L 4G –n sanlv1 sanvg1
```

Management and Inventory

```
:vmid.NZVJT002   :node.VS2
:chpid.51    :realwwpn.5005076401620 8c5
:rdev.0304   :virtwwpn.c05076fc7d800c10
:sanframe.1822   :sandev.0EE0
:targwwpn.50060482d52e4fa3   :lun.0027000000000000
:size.36G   :uuid.360060480000190101822533030454530
```

SAN with z/VM

non-Linux uses of EDEV

- Minidisks
  - All of CMS and CP and LPPs
  - Most guests: VSE, Linux, Solaris, VM
- Paging (and spooling)
  - With a hypervisor short-cut (auto D250)

FCP attached to VM

```
/* make a SAN volume work like an FBA disk */
'CP SET EDEV FF02 TYPE FBA ATTR SCSI' ,
   'FCP_DEV 010A WWPN 50060482D52CC7F2 LUN 0002000000000000' ,
   'FCP_DEV 020A WWPN 50060482D52CC7FD LUN 0002000000000000'
'CP VARY ON FF02'

/* how does it look to CP? */
'CP Q DASD DETAILS FF02'
'CP Q 10A 20A'
...
FF02  CUTYPE = 6310-80, DEVTYPE = 9336-10, VOLSER = SAN002,
      CYLS = 91003, BLKS = 70709760
FCP  010A ATTACHED TO SYSTEM   0000 CHPID 50
FCP  020A ATTACHED TO SYSTEM   0000 CHPID 54
```

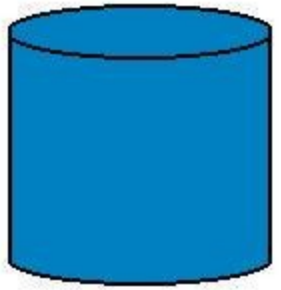
```
cp q dasd san00a
DASD FF0A CP SYSTEM SAN00A   0

cp q edev ff0a details
EDEV FF0A TYPE FBA ATTRIBUTES SCSI
  VENDOR: EMC PRODUCT: SYMMETRIX REVISION: 5771
  BLOCKSIZE: 512 NUMBER OF BLOCKS: 70709760
  PATHS:
    FCP_DEV: 030A WWPN: 50060482D52E4FA3 LUN: 02BB000000000000
      CONNECTION TYPE: SWITCHED
    FCP_DEV: 040A WWPN: 50060482D52E4FAC LUN: 02BB000000000000
      CONNECTION TYPE: SWITCHED
```

EDEV attached to guest

- Attached to CMS to run CPFMTXA (then attach to system for normal use)
- Attached to Linux for full LUN (either dedicated or “DEVNO” mdisk)

ECKD mainframe disk

- z/VM (CP)
- z/VM (CMS)
- Linux
- VSE
- z/OS
- Solaris



ECKD

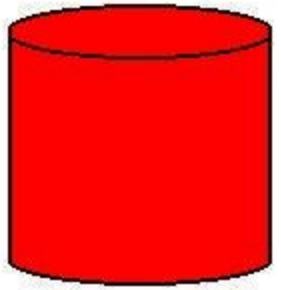
extended count/key/data

(and tracks & records)

ECKD traffic includes non-data

FBA mainframe disk

- **z/VM (CP)**
- **z/VM (CMS)**
- **Linux**
- **VSE**
- z/OS



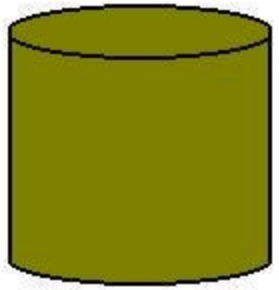
FBA

fixed blocks / just data

z/OS cannot use FBA disks

SAN disk or SCSI disk

- z/VM (CP)
- z/VM (CMS, via EDEV)
- Linux
- VSE
- Solaris, AIX, HP-UX
- Windows



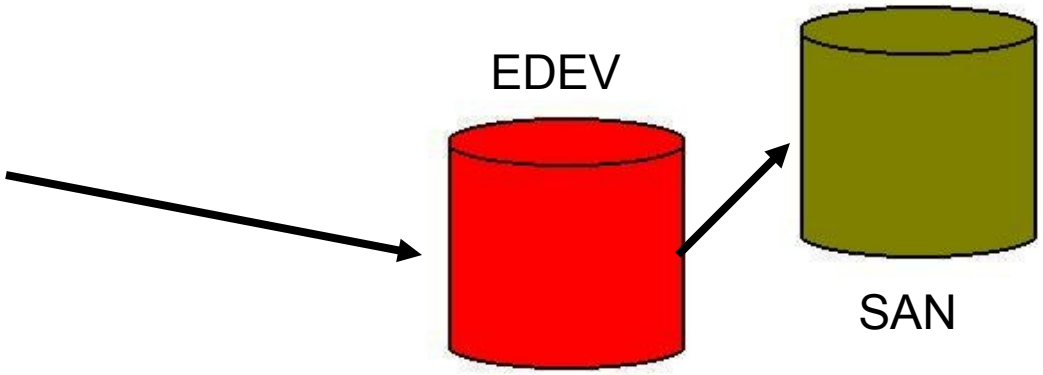
SAN

fixed blocks / just data

SAN disk presented as SCSI

SAN is FBA … sort of

- z/VM (CP)
- z/VM (CMS)
- Linux
- VSE
- Solaris



EDEV makes SAN look like FBA (9336, 3370)

Same format. Same I/O command codes.

Reasons to not run EDEV

- Slower Throughput (protocol translation)
- Increased Overhead (hypervisor CPU)
- Because IBM recommends direct FCP

Reasons to run EDEV

- FBA simpler to configure in Linux
- No need to re-configure when cloning
- Minidisks and CP Dir to manage them
- Minidisk caching
- Can share minidisks (or full volume MD)
- …

Reasons to run EDEV

- …
- No topology issues in the guests
- Easier sharing across LPARs
- EDEV easier to monitor than direct SAN
- Because IBM recommends direct FCP

Interesting and Important Issues

- SAN topology
- deduplication
- interoperability ... does it matter?
- security

Stretching the Shared Disk Envelope

- Can we share SAN volumes? simultaneously? across unlike systems?
- Will discuss shared filesystems more, and especially read-only op sys, later in the workshop … no … wait … that was *earlier*, so you missed it! Bummer.

Basic Storage Requirements

- Replication (more than failover)
- Multipath (failover within I/O space)
- Backup (multiple points of recovery)
- Security / Isolation

Avoid gratuitous partition tables

- Common partitioning: zero, 1, 2, or 3
- Understood by either driver (scsi or dasd)
- Use PC “primary partitions”

But don’t!

- Partitioned requires double layer admin
- Non-partitioned gives simpler LVM admin
- Non-partitioned makes sharing easier

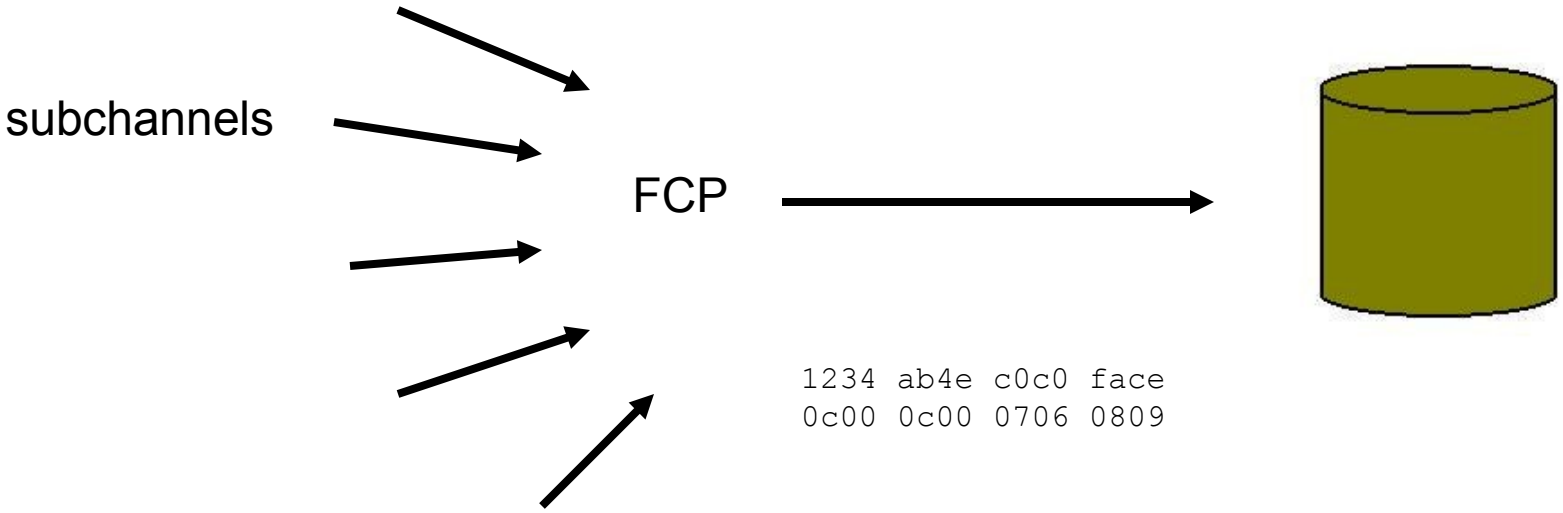
If you must partition …

disk type	*driver*	*format with*	*partition with*
ECKD	**dasd**	**dasdfmt**	**fdasd**
FBA	**dasd**		**fdisk**
SAN	**zfcp+scsi**		**fdisk**
EDEV	**dasd**		**fdisk**

Speaking of Security …

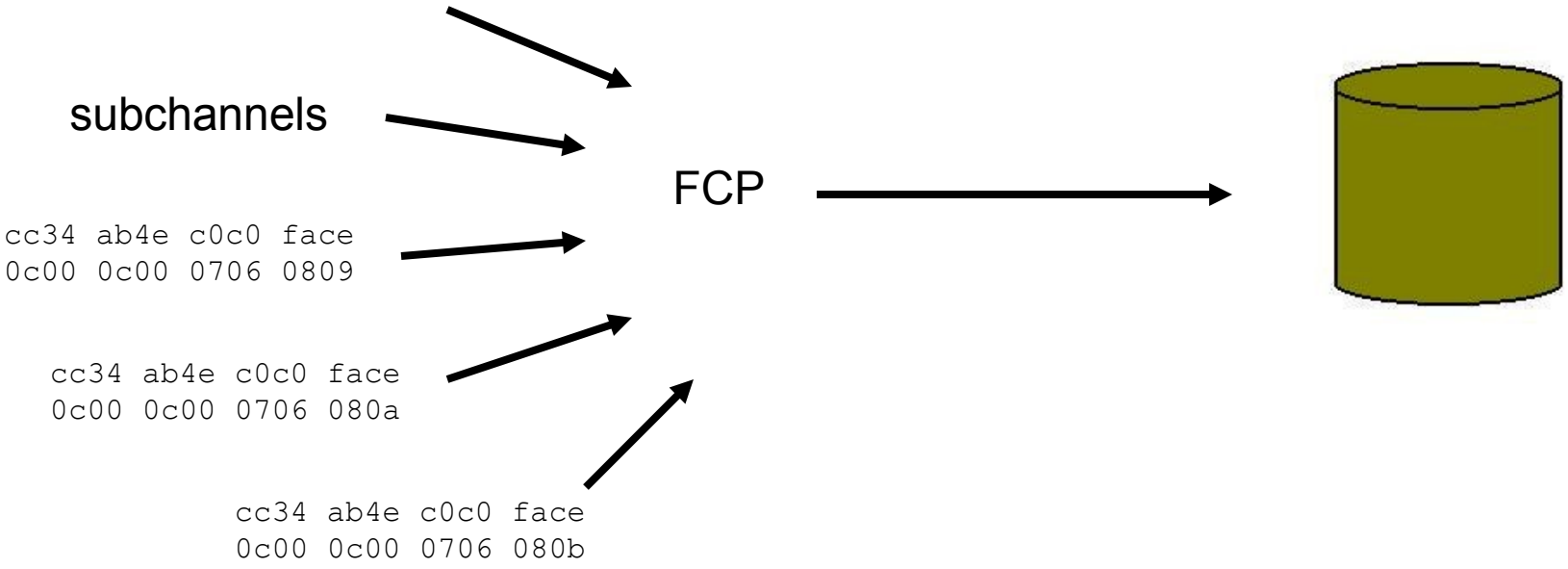
- LUNs are zoned and masked
- NPIV enabled for the fabric
- Without NPIV
  - One (real) WWPN for CHPID
- With NPIV
  - Unique (virtual) WWPN per subchannel

NPIV -- the HBA Multiplier



Without NPIV, all traffic from the single FCP looks like the same host.

NPIV -- the HBA Multiplier



With NPIV, each subchannel
(on each LPAR) looks like a different host.

NPIV for shared EDEVs

```
cp q userid
TROTHR   AT VS1
cp q 30a 40a
FCP  030A ATTACHED TO SYSTEM   0000 CHPID 51
     WWPN C05076FC7D800B28
FCP  040A ATTACHED TO SYSTEM   0000 CHPID 55
     WWPN C05076FC7D801028

cp q userid
TROTHR   AT VS2
cp q 30a 40a
FCP  030A ATTACHED TO SYSTEM   0000 CHPID 51
     WWPN C05076FC7D800C28
FCP  040A ATTACHED TO SYSTEM   0000 CHPID 55
     WWPN C05076FC7D801128
```

Whole disk == “partition zero”

When can you use it?

- **`dasdfmt -l cdl`** <<< NOT okay
- **`dasdfmt -l ldl`**
- CMS **`format`**
- SAN
- FBA

If you must partition …

disk type	*driver*	*format with*	*partition with*
ECKD	**dasd**	**dasdfmt**	**fdasd**
FBA	**dasd**		**fdisk**
SAN	**zfcp+scsi**		**fdisk**
EDEV	**dasd**		**fdisk**

inSANe Summary

- Brand X storage (more than speed)
- “System Any” interconnect
- VM should manage the storage
- Know your layers (MPIO+LVM2)
- Understand NPIV effects
- Avoid gratuitous partitioning

Thank You!!

Rick Troth

Senior Developer

Velocity Software, Inc

<rickt@velocitysoftware.com>

http://www.velocitysoftware.com/