SHARE
Technology · Connections · Results

Tending the SANity of the Flock

SAN Experiences at Nationwide

Rick Troth <trothr@nationwide.com>

March 2, 2009

SHARE 112 session 9286

Downloadable Proceedings

http://ew.share.org/client_files/callpapers/attach/SHARE_in_Austin/S9286RT204515.pdf

Disclaimer

The content of this presentation is informational only and is not intended to be an endorsement by Nationwide Insurance. Each site is responsible for their own use of the concepts and examples presented.

Or 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.

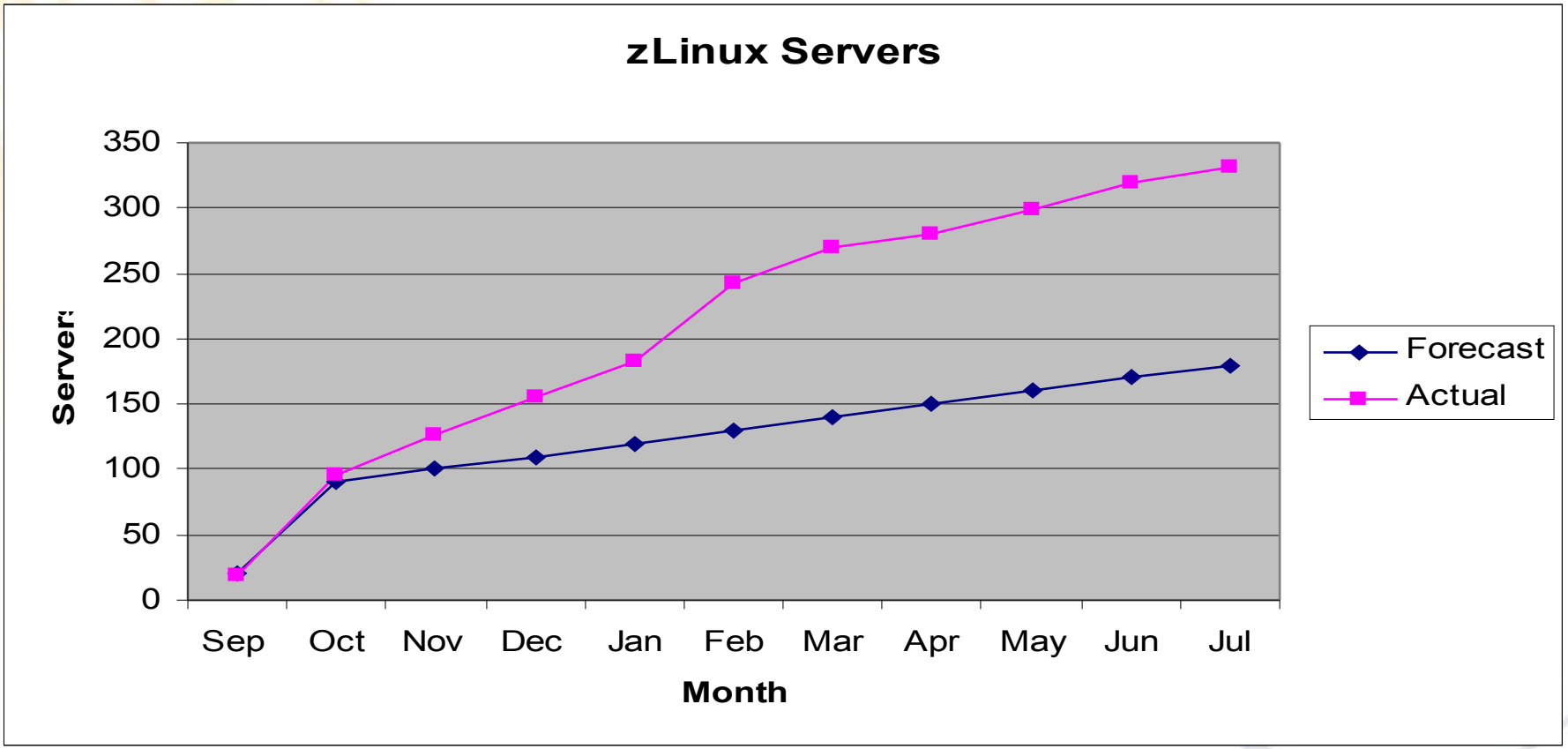
When in doubt, ask! Still in doubt? try it!

A New Iceberg

- The issue: ECKD constrained
- The solution: put some content on SAN
- The implementation
- The results …
- Lessons Learned
- Changes in z/VM 5.3 and 5.4

Linux has Grown Fast!!

*And I thought we were busy **before** we got Linux!*
Rick Barlow, Aug 1, 2006



Using up our ECKD Space

ECKD is constrained

- Cost (we could buy more, but …)
- Interconnection / Interoperability
- Different Granularity (than other Linux)

So … put “user files” onto SAN

Stretching the Shared Disk Envelope

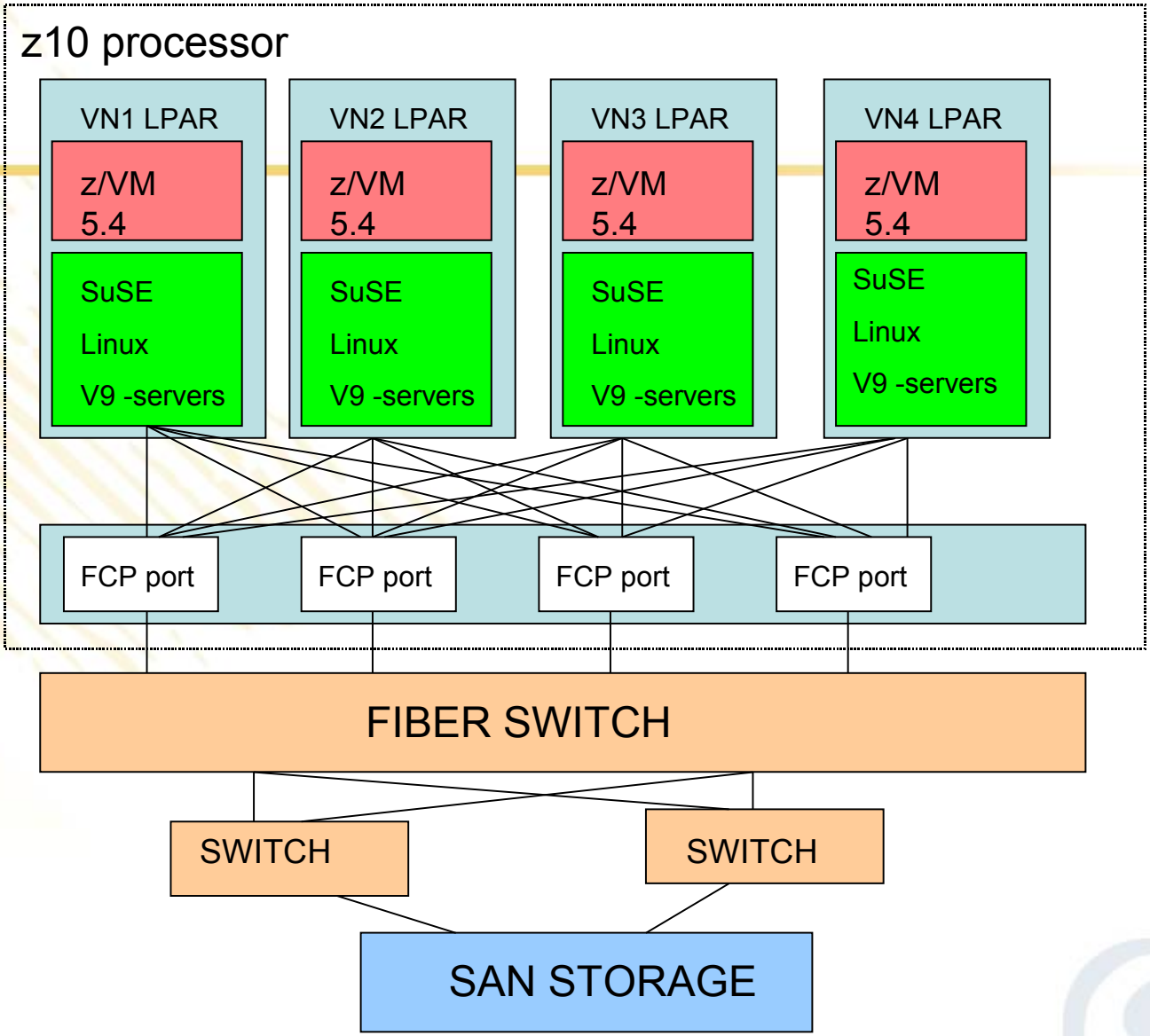
- Can we share SAN volumes? simultaneously? across unlike systems?
- Will discuss shared filesystems more, and especially read-only root, later in the week (session 9216) … no … wait … that was EARLIER, so you missed it! Bummer.

Storage Area Network

- 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?)

Basic Storage Requirements

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


z10 processor
VN1 LPAR
z/VM 5.4
SuSE Linux V9 -servers
VN2 LPAR
z/VM 5.4
SuSE Linux V9 -servers
VN3 LPAR
z/VM 5.4
SuSE Linux V9 -servers
VN4 LPAR
z/VM 5.4
SuSE Linux V9 -servers
FCP port
FCP port
FCP port
FCP port
FIBER SWITCH
SWITCH
SWITCH
SAN STORAGE

Storage Area Network

SHARE
Technology · Connections · Results

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

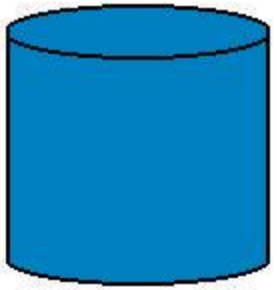
Storage Area Network

Picking out Furniture …

- Point to Point
- Arbitrated Loop
- Switched Fabric ➠

ECKD mainframe disk

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



ECKD

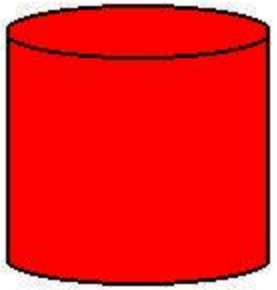
extended count/key/data

(and tracks & records)

ECKD traffic includes non-data

FBA mainframe disk

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



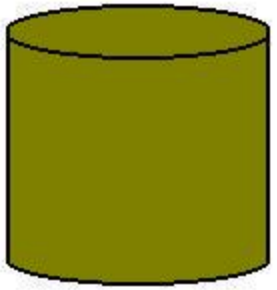
FBA

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



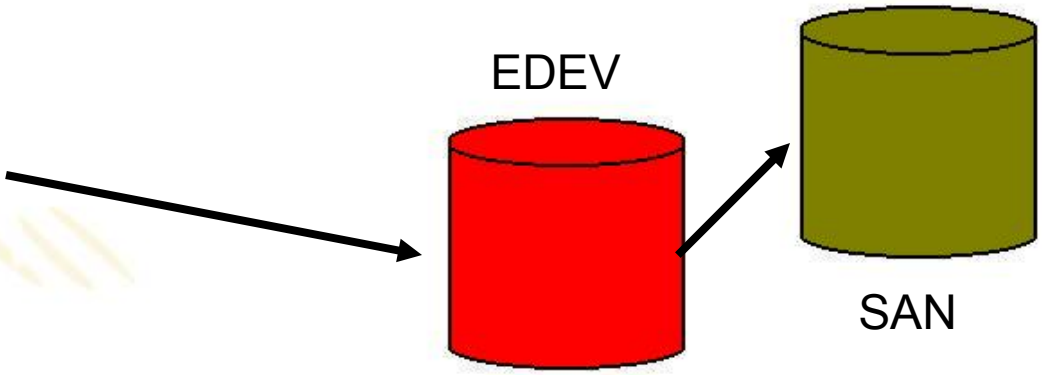
SAN

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.

“The Nucleus Got Bigger”

- 31-bit support dropped,
  CP Nucleus should have gotten smaller
- SAN support (EDEV) added,
  CP Nucleus actually got *larger*

IBM leveraged AIX driver code – cool!

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
```

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

Storage Network – Multipath and LVM

Picking out Appliances …

- EVMS
- MPIO+LVM2 ➟

LVM applies to direct FCP and to EDEV

MPIO only needed for direct FCP

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
```

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`**

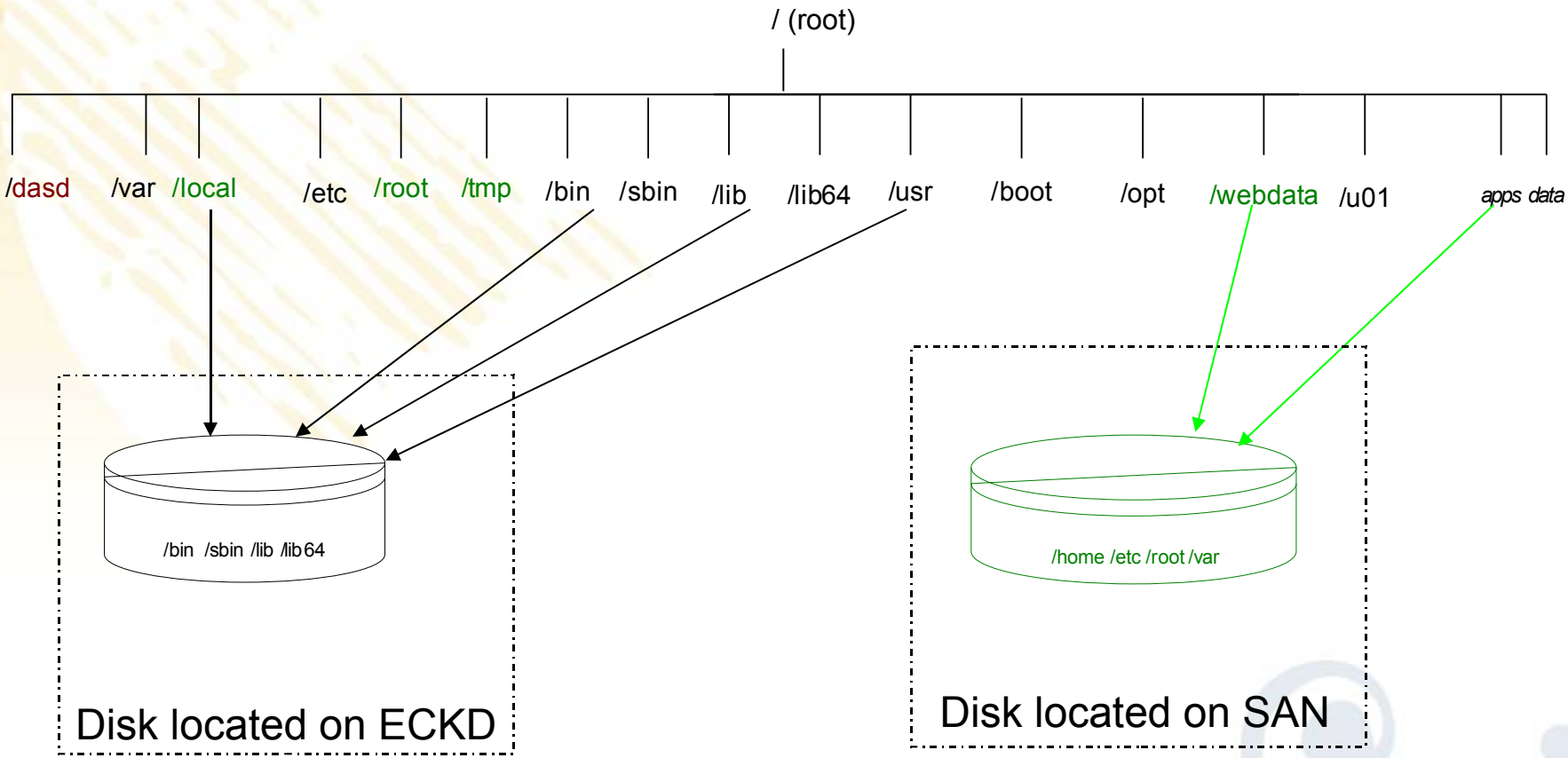
Multipath Management

```
: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 is Seamless

- Operating systems stay on ECKD for now
- No Change of application file access
- Open-Ended storage capacity
- Secured at the hardware level

Mixed Media Methodology



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

About EDEV …

- CP emulates FBA on SAN
- Device type 9336 (like 3370 or 3310)
- Attach EDEVs just like real DASD

EDEV attached to VM

- Minidisks
- Paging (and spooling)

EDEV attached to guest

- Attached to CMS to run CPFMTXA
- Attached to Linux for “full LUN”

Reasons to not run EDEV

- Slower Throughput (protocol translation)
- Increased Overhead (hypervisor CPU)
- Multipath Failover (VM lags Linux)
- Dynamic Multipath Management

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 driver currency issues (in Linux)
- Easier sharing across LPARs
- EDEV easier to monitor than direct SAN

Stretching the Brand X Envelope

- It's all about interoperability …

SAN does for disk

what z/VM does for systems … sort of

Thank You!!

Richard Troth

Senior VM Systems Programmer

Nationwide Services Co., LLC

One Nationwide Plaza, MB-02-201
Columbus, OH 43215-2220
Voice: 1-614-249-7642
Cell: 1-614-849-8255
trothr@nationwide.com

```
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
```

```
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`**