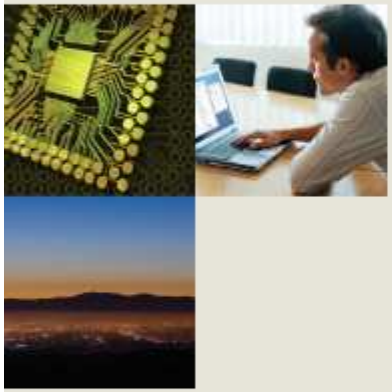


Logical Volume Management
for Linux on System z

Session 9282

Horst Hummel (Horst.Hummel@de.ibm.com)
Linux on System z Development
IBM Lab Boeblingen, Germany

San Jose, August 15th 2008

Agenda

- **Logical volume management overview**
  - RAID levels
  - Striping
  - Mirroring
  - Multipathing
- **Multipathing with zFCP / SCSI**
- **Multipathing with DASD using PAV**
- **Outlook on future development**

Redundant Arrays of Inexpensive / Independent Disks (RAID)

SHARE
Technology · Connections · Results

- **Using multiple disks to share or replicate data to increase**
  - Data integrity
  - Fault-tolerance
  - Throughput
  - Capacity
- **Provides different configurations (RAID Level)**
- **Implemented as Software- or Hardware-RAID**

RAID Level

- **Linear device (JBOD)**
  - Concatenate multiple physical disks to single virtual device
- **RAID-0 (striping)**
  - Data is split evenly across disks (round robin)
  - Fast and efficient (no redundant information stored)
  - No fault-tolerance
- **RAID-1 (mirroring)**
  - exact data copy to 2 or more disks
  - Fast on read slow on write
  - Fault-tolerance (redundant data)
  - Needs additional capacity

RAID Level (cont.)

- **RAID-2**
  - Stripe data at **bit level** across several disks
  - Use 'Hamming code' for error correction
  - Intended for use with no built-in error detection
- **RAID-3**
  - Stripe data at **byte level** across several disks
  - parity stored on dedicated disk (bottleneck)
  - Cannot serve multiple requests simultaneously
  - Parity allows recovery of single disk failure
- **RAID-4 (Striping & Dedicated)**
  - Stripe data at **block level** across several disks
  - Otherwise similar to level 3

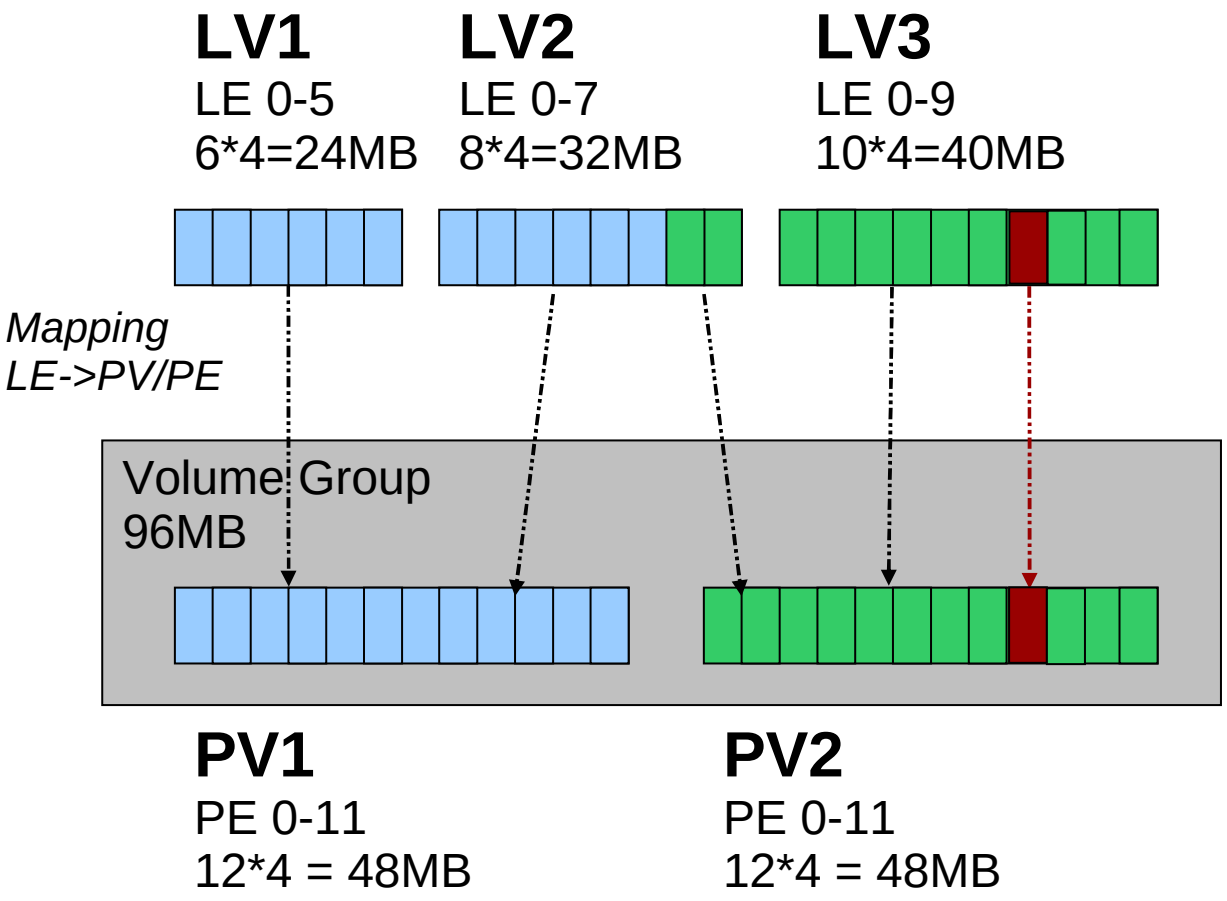
RAID Level (cont.)

- **RAID-5 (Striping & Distributed Parity)**
  - Distribute parity among disks
  - Otherwise similar to level 4
- **RAID-10 (Mirroring & Striping)**
  - Combination of RAID-1 and RAID-0 (mirroring of striped device)
  - Good performance & Fault tolerance

Logical Volume Management (LVM) Terms

- **Physical volume (PV)**
  - Any kind of block device (DASD, SCSI,...)
- **Physical Extend (PE)**
  - Even sized parts of the physical volume (default size 4M)
- **Volume Group (VG)**
  - Pool of physical extends
- **Logical volume (LV)**
  - Virtual block device based on concatenated pooled PEs
- **Logical Extend (LE)**
  - Part of a logical volume
  - Same size as physical extend of the volume group
  - 1:1 mapping LE:(PV:PE)

LVM – Simple Example (linear device)



Logical Volume (LV)

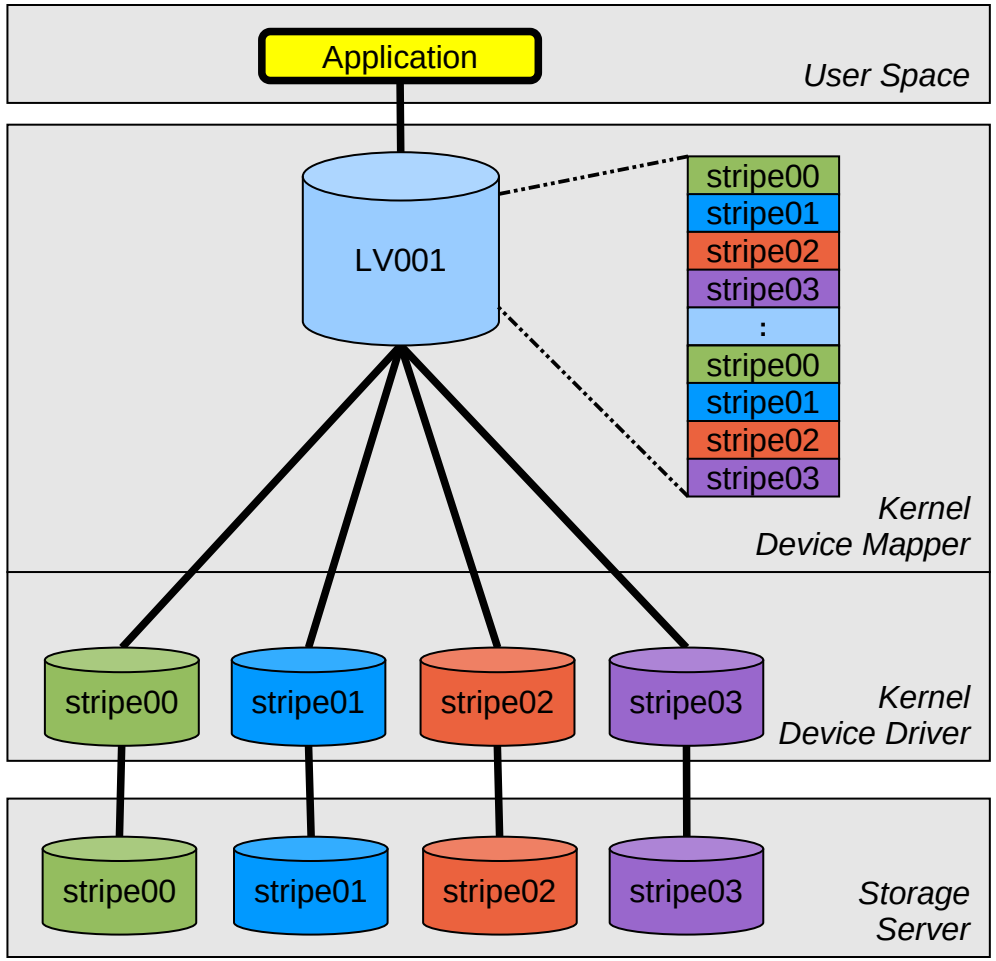
e.g.
LV3:LE7
->PV2:PE9

Physical Volume (PV)

1PE = 1LE = 4MB (default size)

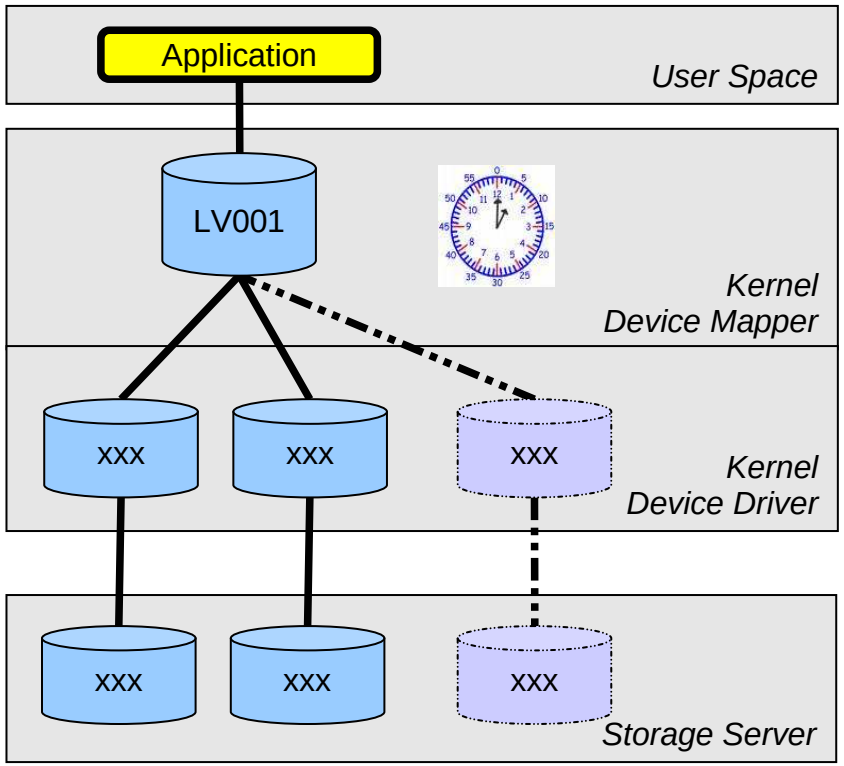
LVM environment for striping

- Performance improvement due to multiple small disks
- No fault-tolerance
- Data evenly split across disks



LVM setup for mirroring

- Same data on each mirror
- Fault-tolerance
  Failing mirror can be recovered non-disruptive
- Needs double (or more) storage capacity
- Enhanced real time capabilities available

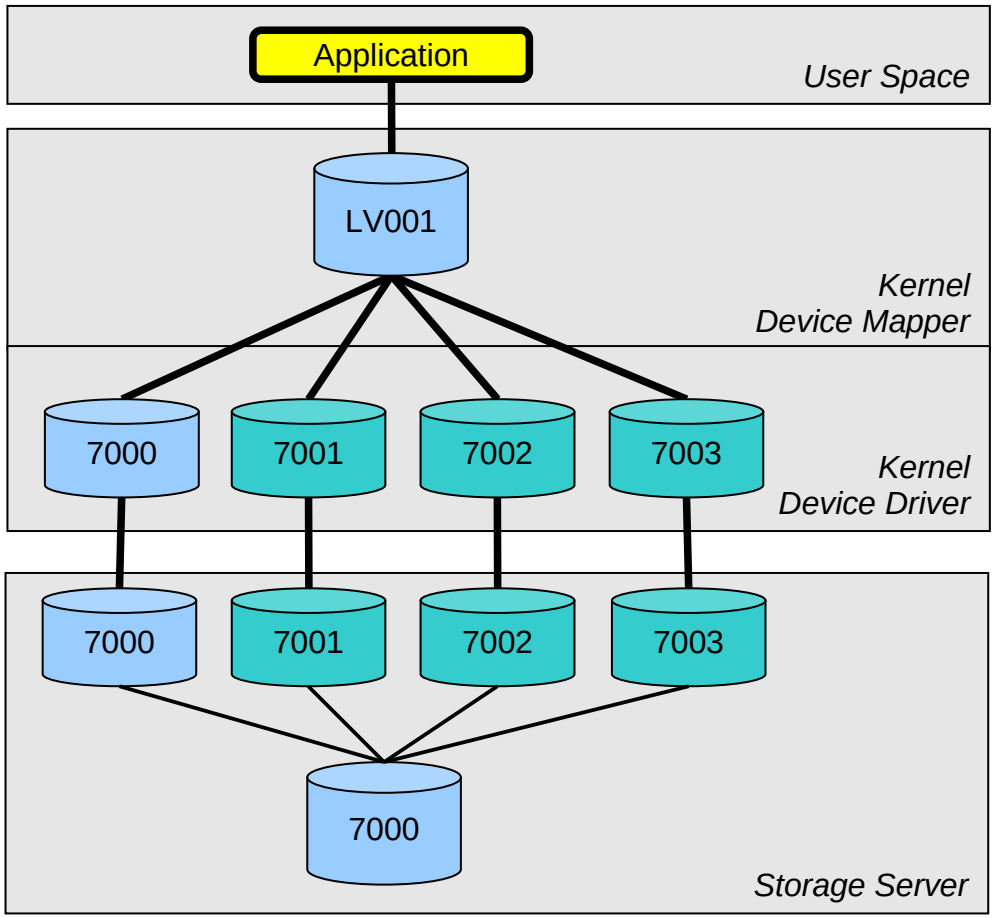


IBM announced a service delivered Data Mirroring Solution for Linux on System z

http://www-03.ibm.com/systems/services/labservices/platforms/labservices_z.html

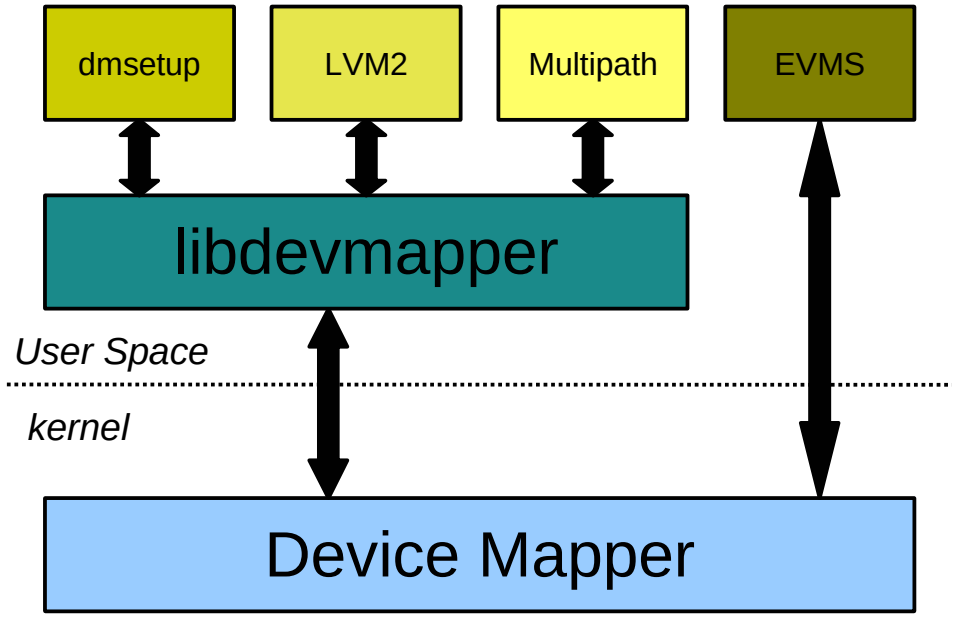
LVM setup for multipathing

- Performance improvement due to path load sharing
- Path fault tolerance (path failover / failback)
- Designed to handle all kind of block devices
- No storage server fault tolerance



Linux LVM Architecture

- **Logical Volume Management applications**
  - dmsetup
    low level logical volume management
  - LVM2
    latest version of Logical Volume Manager
  - Multipath
    multipath configuration tool
  - EVMS
    Enterprise Volume Management System



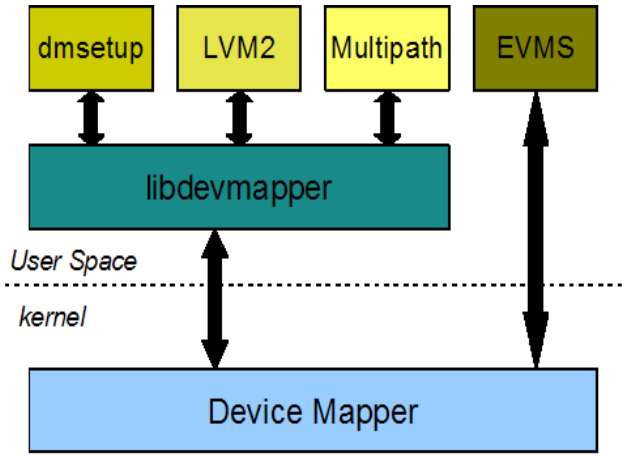
Linux LVM Architecture (cont.)

- **Libdevmapper**

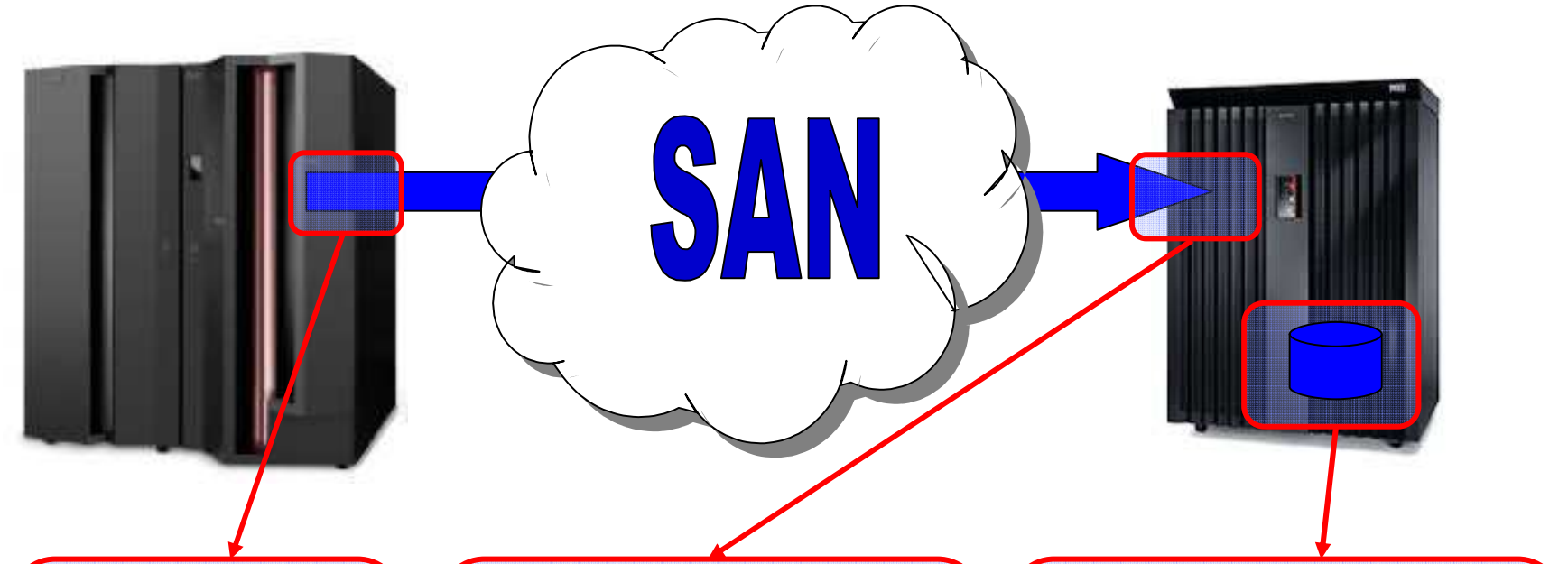
  library for interaction between user and kernel device mapper

- **Device Mapper**
  - Modular framework for stacking

    target drivers like
    - Linear target
    - Mirror target
    - Multipath target
  - Responsibilities
    - Discover set of associated devices
    - Create mapping table containing configuration information
    - Pass mapping table into kernel
    - Possibly save mapping information



SAN Addressing
Path to FCP device



Device Number
(adapter busid)
e.g. 0.0.190e

Worldwide Port Name
(WWPN)
e.g. 0x500507630300c562

Logical Unit Number
(LUN)
e.g. 0x401040ae00000000

Multipathing with zFCP / SCSI Configuration

- **SCSI disk configuration (first path)**
  with bus ID 0.0.190e *(X)*, WWPN 0x500507630300c562 *(1)* and LUN 0x401040ae00000000 *(A)*.
  - Change to adapter directory
    `cd /sys/bus/ccw/drivers/zfcp/0.0.190e`
  - Set the adapter to online

    0.0.190e # chccwdev -e 0.0.190e

  - Check for messages (in `'/var/log/messages'`)

    ```
    scsi2 : zfcp
    zfcp: The adapter 0.0.190e reported the following characteristics:
    WWNN 0x5005076400c2d09e, WWPN 0x5005076401a07fd4, S_ID 0x00688a13,
    adapter version 0x3, LIC version 0x606, FC link speed 2 Gb/s
    zfcp: Switched fabric fibrechannel network detected at adapter 0.0.190e.
    ```

  - Add target port to FCP adapter

    0.0.190e # echo 0x500507630300c562 > port_add

Multipathing with zFCP / SCSI Configuration (cont.)

- Change to newly created port directory
  0.0.190e # cd 0x500507630300c562/
- Add FCP LUN to that port

```
0.0.190e/0x500507630300c562 # echo 0x401040ae00000000 > unit_add
```

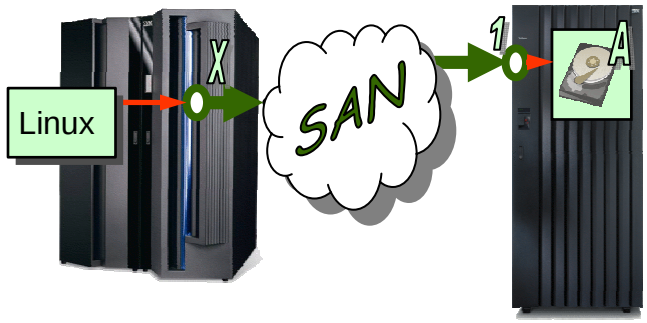
- Find new messages

```
Vendor: IBM Model: 2107900 Rev: .216
Type: Direct-Access ANSI SCSI revision: 05
SCSI device sda: 10485760 512-byte hdwr sectors (5369 MB)
sda: Write Protect is off
SCSI device sda: drive cache: write back
sda: unknown partition table
sd 2:0:0:0: Attached scsi disk sda
sd 2:0:0:0: Attached scsi generic sg0 type 0
```

- SCSI disk is now available

  0.0.190e ***# lsscsi***

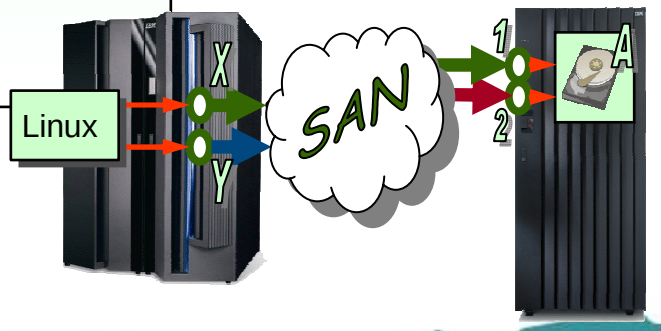
  [2:0:0:0] disk IBM 2107900 .216 /dev/sda



Multipathing with zFCP / SCSI Configuration (cont.)

- **SCSI disk configuration (remaining paths)**
  with additional bus ID 0.0.520e *(Y)*, additional WWPN 0x500507630303c562 *(2)*

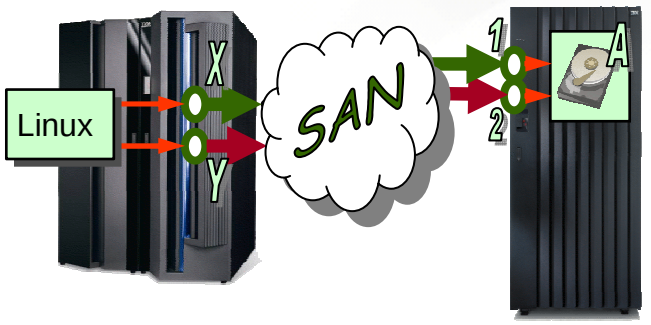
```
# cd /sys/bus/ccw/drivers/zfcp/0.0.190e/
0.0.190e # echo 0x500507630303c562 > port_add
0.0.190e # echo 0x401040ae00000000 >
0x500507630303c562/unit_add
0.0.190e # cd ..

zfcp # cd 0.0.520e/
0.0.520e # echo 0x500507630300c562 > port_add
0.0.520e # echo 0x401040ae00000000 >
0x500507630300c562/unit_add
0.0.520e # echo 0x500507630303c562 > port_add
0.0.520e # echo 0x401040ae00000000 >
0x500507630303c562/unit_add

0.0.520e # lsscsi
[1:0:0:0] disk IBM 2107900 .216 /dev/sdc
[1:0:1:0] disk IBM 2107900 .216 /dev/sdd
[2:0:0:0] disk IBM 2107900 .216 /dev/sda
[2:0:1:0] disk IBM 2107900 .216 /dev/sdb
```



Multipathing with zFCP
Multipath Configuration

- **Start multipathd**
  linux:~ # /etc/init.d/multipathd start
- ***load dm-multipath module, activate mp-tools***
  linux:~ # /etc/init.d/boot.multipath start
- **Check for multipath configuration**

```
linux:~ # multipath -ll
36005076303ffc56200000000000010aeIBM,2107900
[size=5G][features=1 queue_if_no_path][hwhandler=0]
\_ round-robin 0 [prio=4][active]
\_ 2:0:0:0 sda 8:0 [active][ready]
\_ 2:0:1:0 sdb 8:16 [active][ready]
\_ 1:0:0:0 sdc 8:32 [active][ready]
\_ 1:0:1:0 sdd 8:48 [active][ready]
```



- **Device node provided by mp-tools**

```
linux:~ # ls -l /dev/mapper/
total 0
brw------- 1 root root 253, 0 Jan 4 11:47 36005076303ffc56200000000000010ae
lrwxrwxrwx 1 root root 16 Jan 4 11:15 control -> ../device-mapper
linux:~ #
```

Multipathing with zFCP Partitioning

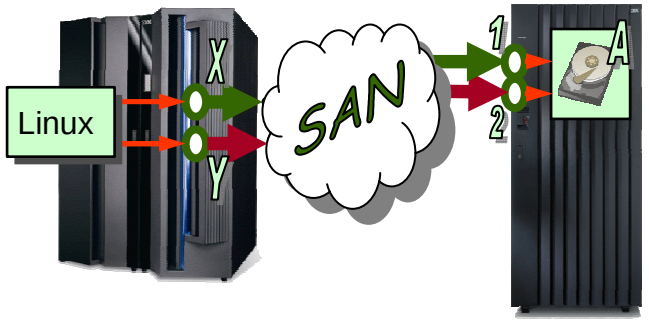
- **Write partition table to disk**

  linux:~ # fdisk /dev/sda

  -> follow instructions to create primary partition

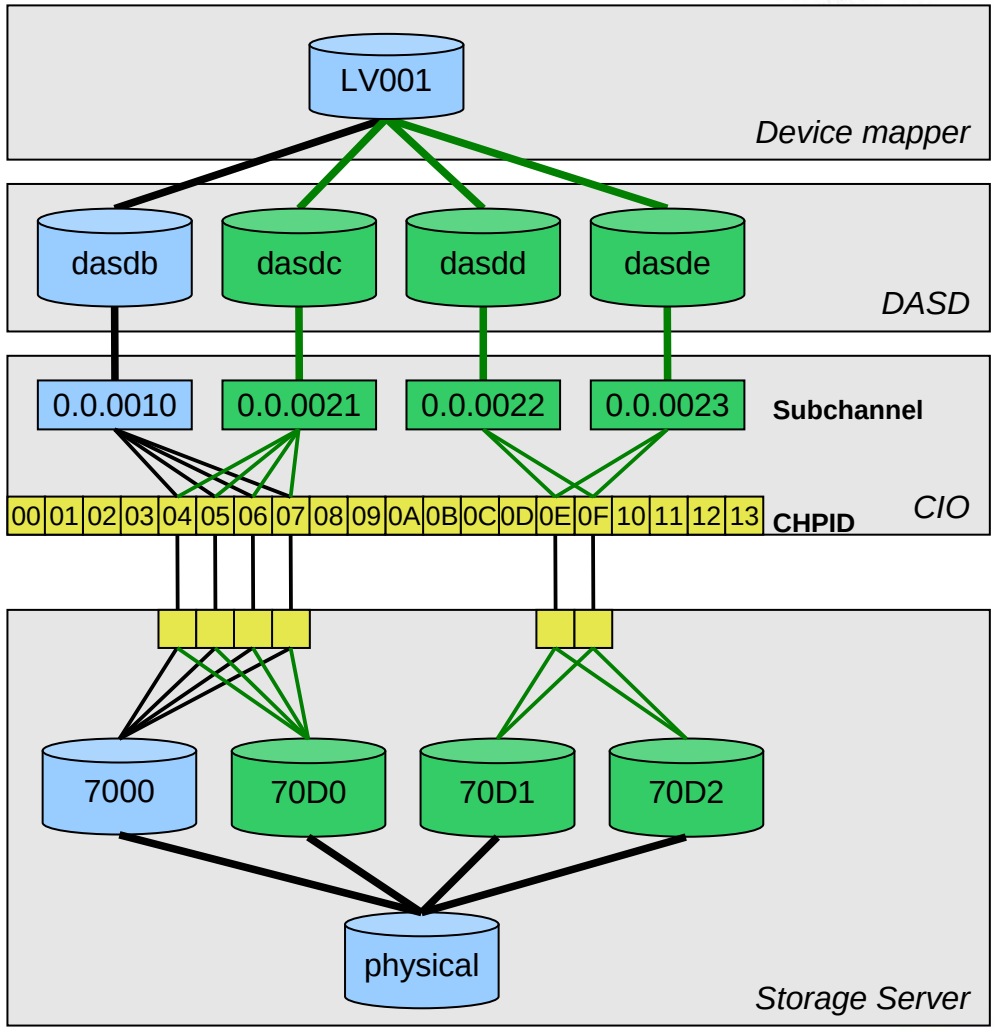
- **Check device nodes**

```
linux:~ # ls -l /dev/mapper/
total 0
brw------- 1 root root 253, 0 Jan 4 12:03 36005076303ffc56200000000000010ae
brw------- 1 root root 253, 1 Jan 4 12:03 36005076303ffc56200000000000010ae-part1
```



Multipathing with DASD using static PAV

- One base path (blue) to physical device
- Additional alias paths (green)
- Increased performance and path-fault tolerance
- Needs additional subchannels



Multipathing with DASD HW configuration

- **PAV configuration on Storage Server**

  please refer to

  IBMTotalStorage Enterprise Storage Server Web Interface User's Guide, SC26-7448

- **zSeries configuration (IOCP)**

```
**************************************************************************
* DEFINE 3390-9 BASE AND ALIASES ADDRESS                                 *
* 16 BASE ADDRESS, 3 ALIASES PER BASE                                    *
**************************************************************************
IODEVICE ADDRESS=(7000,016),CUNUMBR=(5000),STADET=Y,UNIT=3390B
IODEVICE ADDRESS=(70D0,048),CUNUMBR=(5000),STADET=Y,UNIT=3390A
```

Multipathing with DASD
DASD configuration

- DASD parameters / attributes
  - **'nopav'** to disable pav enablement call and device re-probing in DASD / CIO
  - **sysfs attributes** in `'/sys/bus/ccw/device/<busid>/'`
    - **vendor: The vendor of the machine (also known as manufacturer).**
    - **alias: '0' for base device / '1' for alias device**
    - **uid: Containing a string like 'www.xxx.yyy.zzz' where**

www	=vendor (also known as manufacturer)
xxx	= serial (serial of the machine)
yyy	= subsystem id (address of the subsystems)
zzz	= unit address (address of the physical disk

- **DASD device configuration (base device)**
  - **Set base devices online**

```
# chccwdev -e 0.0.7000
```

  - **Check for messages (in `'/var/log/messages'`)**

```
dasd(eckd): 0.0.7000: 3390/0A(CU:3990/01) Cyl:3339 Head:15 Sec:224
dasd_erp(3990):  0.0.7000: EXAMINE 24: No Record Found detected
dasd(eckd): 0.0.7000: volume analysis returned unformatted disk
```

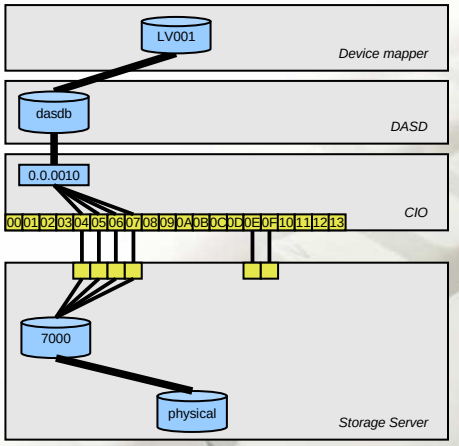
Multipathing with DASD
DASD configuration (cont.)

- **Low level format base device**
  - **get device name using 'lsdasd'**
    ```
    # lsdasd
    ```
  - **Format device**

```
# dasdfmt -b 4096 -y -p /dev/dasdb
cyl   5 of   5 ############################################################| 100%
Finished formatting the device.
Rereading the partition table... ok
```

- **Write partition table**

```
# fdasd -a /dev/dasdb
auto-creating one partition for the whole disk...
writing volume label...
writing VTOC...
rereading partition table...
```



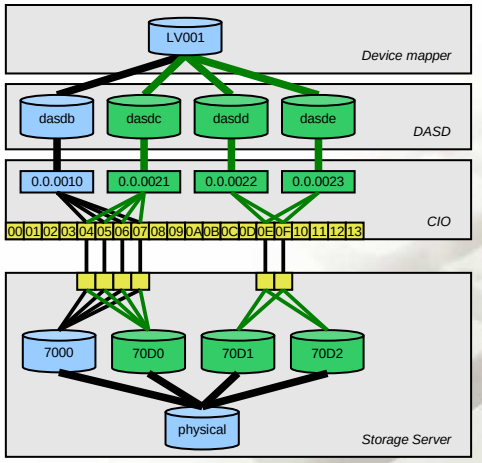
Multipathing with DASD
DASD configuration (cont.)

- **Find new messages**

```
dasd(eckd): 0.0.7000: (4kB blks): 2404080kB at 48kB/trk
compatible disk layout
 dasdb: unknown partition table
 dasdb:VOL1/  0X7000:
 dasdd:VOL1/  0X7000: dasdd1
```

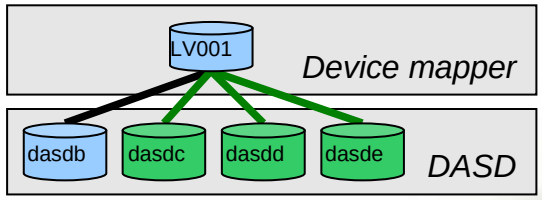
- **DASD device configuration (alias devices)**

```
# chccwdev -e 0.0.70d0-0.0.70d2
```



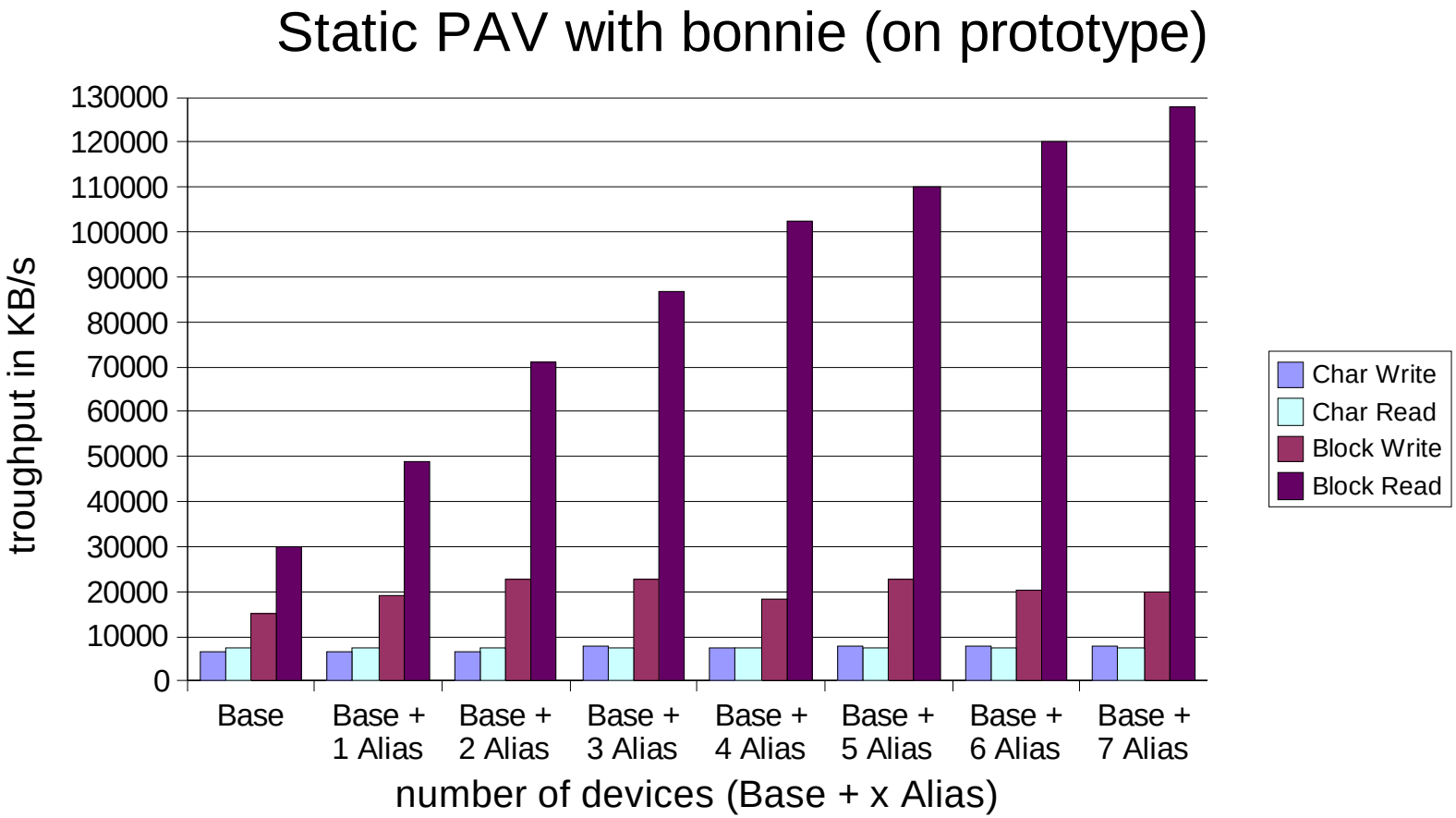
Multipathing with DASD
Multipath configuration

- **Start multipathd**
```
# /etc/init.d/multipathd start
```
- ***load dm-multipath module, activate mp-tools***
```
# /etc/init.d/boot.multipath start
```
- **Check for multipath configuration**
```
# multipath –ll
IBM.75000000092461.2a00.1a IBM,S/390 DASD ECKD
[size=2.3G][features=0][hwhandler=0]
\_ round-robin 0 [prio=4][undef]
\_ 0:0:10778:0 dasdb 94:4  [undef][ready]
\_ 0:0:10927:0 dasdc 94:8  [undef][ready]
\_ 0:0:10778:0 dasdd 94:12 [undef][ready]
\_ 0:0:10927:0 dasde 94:16 [undef][ready]
```

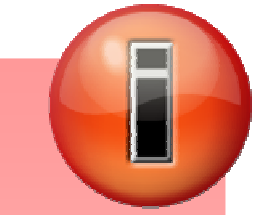


- **Device node provided by mp-tools**
```
# ls -l /dev/mapper/*
brw-rw---- 1 root disk 253, 0 Oct 19 17:02 /dev/mapper/IBM.75000000092461.2a00.1a
brw-rw---- 1 root disk 253, 1 Oct 19 17:10 /dev/mapper/IBM.75000000092461.2a00.1ap1
```

Multipathing with DASD Performance (first glance)

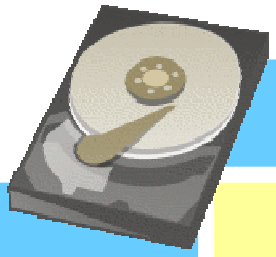


Multipathing with DASD Pitfalls



- Make sure the device is formatted and partitioned prior to mulitpath-setup
- Be careful when formatting / partitioning devices currently in use (see howto)
- Use cio_ignore since base detection does re-probing (performance issue during ipl)
- Use blacklist in multipath-tools to exclude no-PAV DASD devices

Disk usage
ECKD and SCSI Comparison



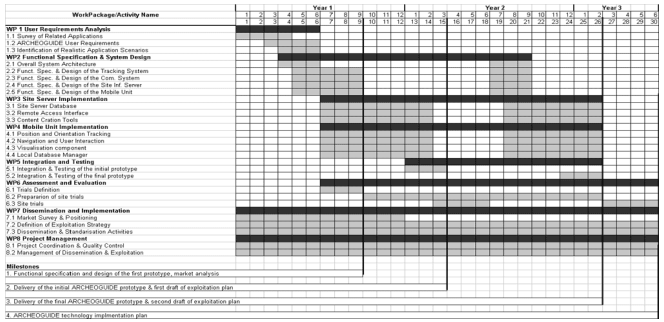
	ECKD DASD	SCSI Disk
Configuration	IOCDS / zVM	IOCDS / zVM
	(operator)	(operator & linux admin)
Access Method	SSCH / CCW	QDIO
Block Size (Byte)	512, 1K, 2K, 4K	512
Disk Size	< ~57GB	?
Formatting (low level)	dasdfmt	not necessary
Partitioning	fdasd	fdisk
File System	mke2fs (or others)	
Access	mount	

Useful Commands

- **`lscss`**
  list channel subsystem devices
- **`lsdasd`**
  list DASD related device information
- **`dasdview`**
  display extended DASD information
- **`lszfcp`**
  list information about zfcp adapters, ports, and units
- **`lsscsi`**
  list all scsi devices
- **`chccwdev -e/-d`**
  enable/disable ccw device
- **`dasdfmt`**
  low level format for DASD (ECKD) devices
- **`fdasd`**
  partitioning tool for DASD
- **`fdisk`**
  partitioning tool for SCSI
- **`multipath -ll`**
  display multipath configuration

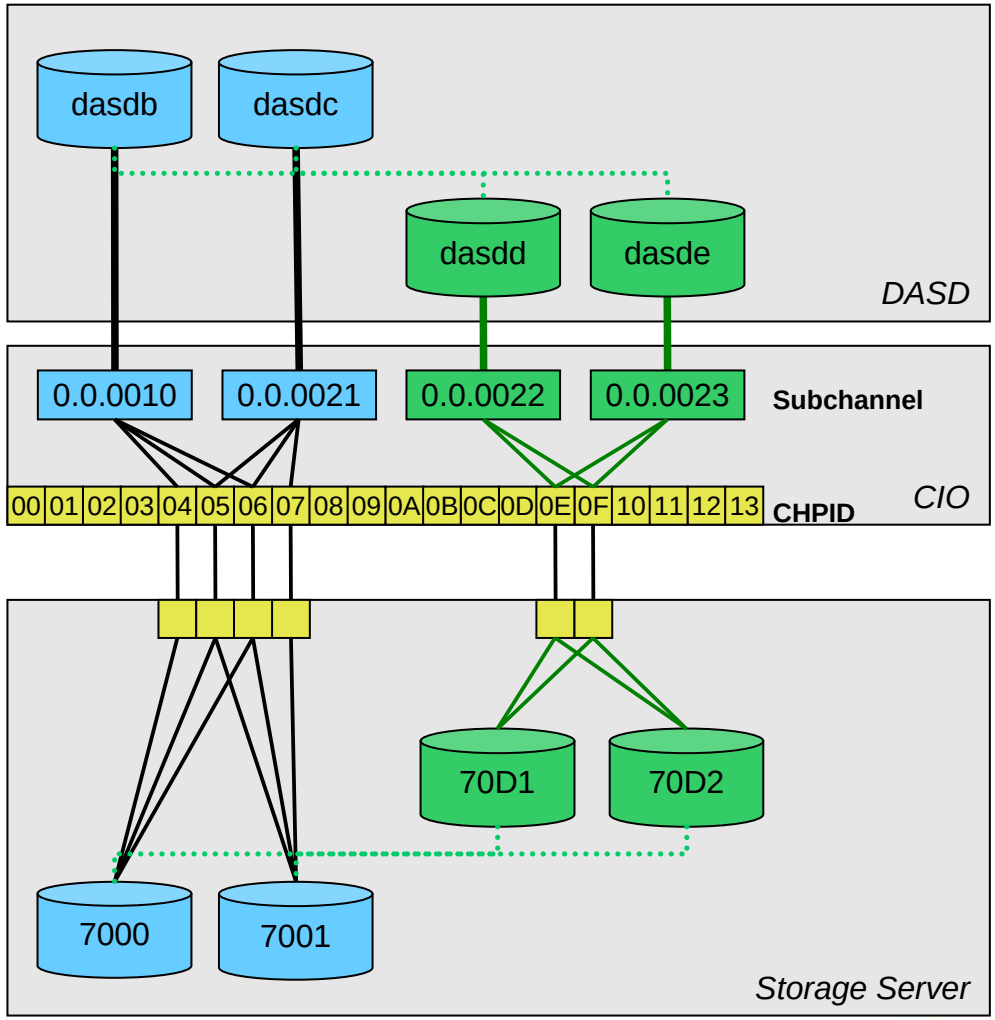
Outlook (subject to change)

- **Enhancements in common block device stack to allow cancellation of a request**
- **Upstream integration of real time mirror solution**
- **PAV / HyperPAV support in DASD device driver**



DASD Next Generation Multipathing using HyperPAV support

- Pool of ALIAS decives can be used for each base device on demand
- Loadbalancing done in DASD device driver
- Configuration autodetection



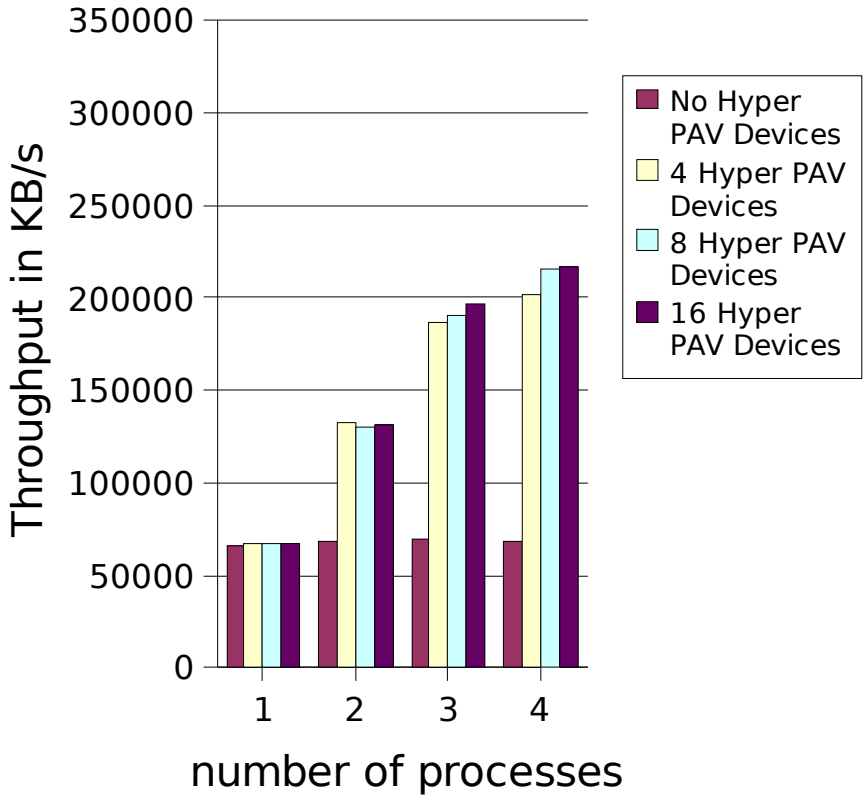
DASD Next Generation Multipathing Configuration

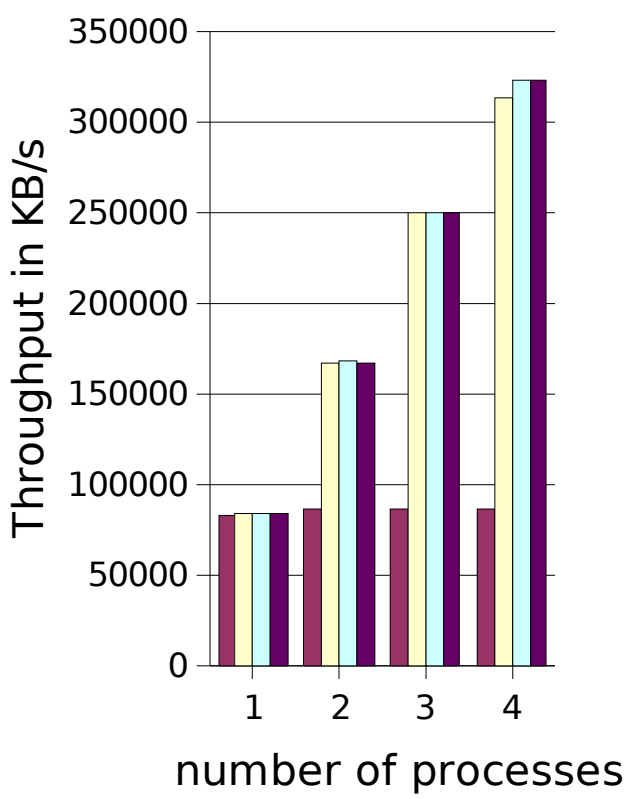
- **PAV configuration on Storage Server**
- **zSeries configuration (IOCP)**
- **Basic DASD configuration**
- **That's it – nothing else to do**
  - no multipath configuration needed
  - no formatting / partitioning related pitfalls

HyperPAV simplifies systems management and improves performance using an on demand I/O model

DASD Next Generation Multipathing

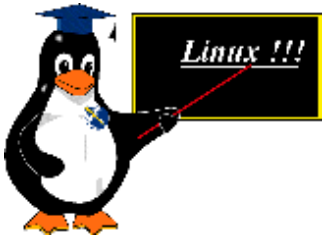
Performance Single Disk Test – Sequential DIO - 700MB file size - 256MB Memory



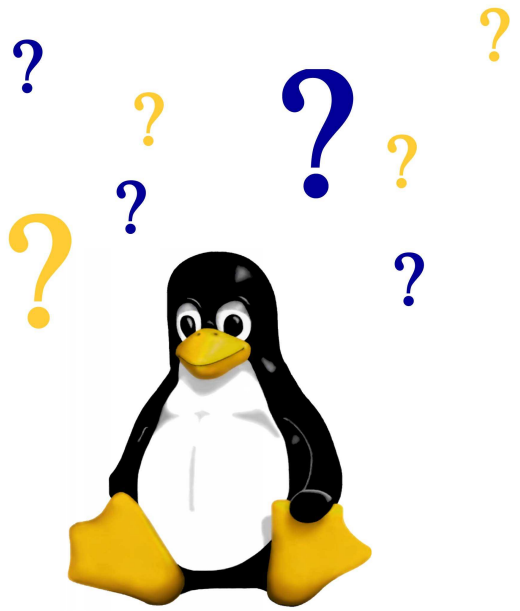


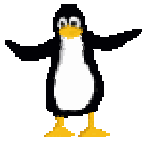
Useful links

- Linux on System z – developerworks page
  http://www-128.ibm.com/developerworks/linux/linux390/
- Device Drivers, Features and Commands (SC33-8411-00)
  http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux390/docu/l26ddd00.pdf
- How to Improve Performance with PAV (SC33-8414-00)
  http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux390/docu/l26dhp00.pdf
- How to use FC-attached SCSI devices with Linux on System z (SC33-8413-00)
  http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux390/docu/l26dts00.pdf
- *Device-mapper development*
  http://sourceware.org/dm/
- *LVM HOWTO*
  http://tldp.org/HOWTO/LVM-HOWTO/

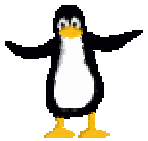


Questions





Thank You



Trademarks

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries. For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml: AS/400, DBE, e-business logo, 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

The following are trademarks or registered trademarks of other companies

Lotus, Notes, and Domino are trademarks or registered trademarks of Lotus Development Corporation
Java and all Java-related trademarks and logos are trademarks of Sun Microsystems, Inc., in the United States and other countries
LINUX is a registered trademark of Linux Torvalds
UNIX is a registered trademark of The Open Group in the United States and other countries.
Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.
SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.
Intel is a registered trademark of Intel 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 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.

References in this document to IBM products or services do not imply that IBM intends to make them available in every country.

Any proposed use of claims in this presentation outside of the United States must be reviewed by local IBM country counsel prior to such use.

The information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.