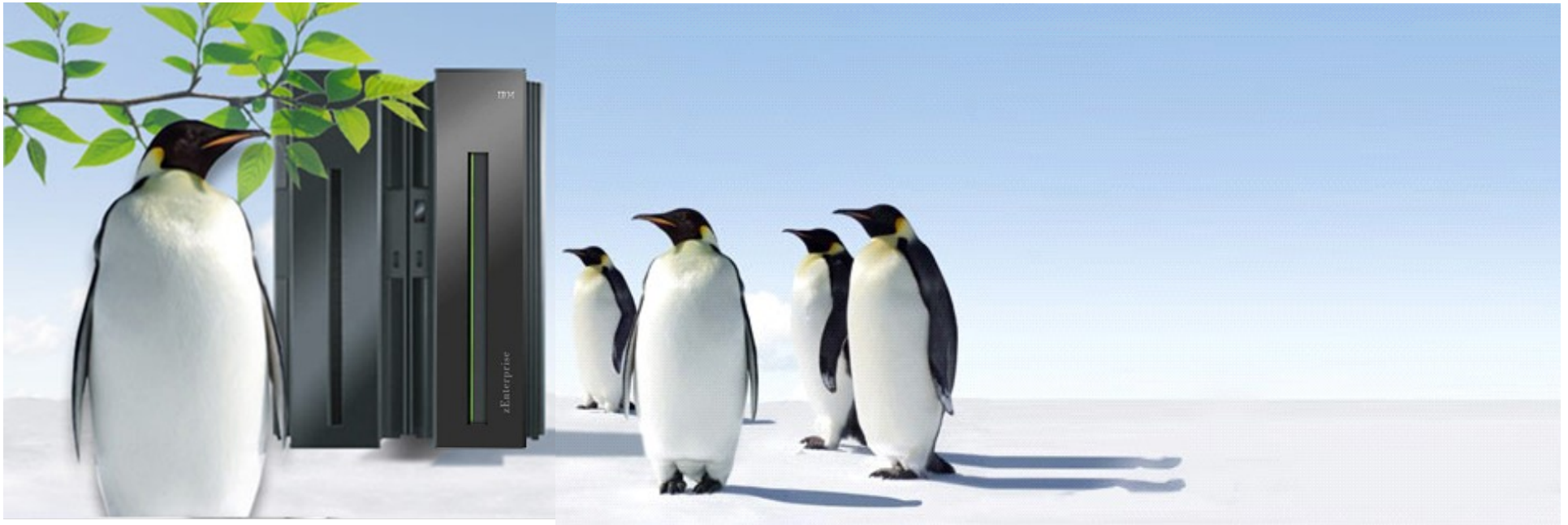
Hyper PAV and Large Volume Support for Linux on System z

Sven Schuetz
Linux on System z Development and Service
sven@de.ibm.com



Trademarks

he following are trademarks of the International Business Machines Corporation in the United States, other countries, or both.

common law marks used by IBM are listed on this page. Failure of a mark to appear does not mean that IBM does not use the mark nor does it mean
product is not actively marketed or is not significant within its relevant market. Those trademarks followed by ® are registered trademarks of IBM in the United Stat
rs are trademarks or common law marks of IBM in the United States.

For a complete list of IBM Trademarks, see www.ibm.com/legal/copytrade.shtml:

*, AS/400®, e business(logo)®, DBE, 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®, System i, System i5, System p, System p5, System x, System z, System z9®, BladeCenter®

The following are trademarks or registered trademarks of other companies.

be logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries
Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
ava-based trademarks are trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.
dows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
ogo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered tra
its subsidiaries in the United States and other countries.
stered trademark of The Open Group in the United States and other countries.
istered trademark of Linus Torvalds in the United States, other countries, or both.
stered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
ure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* All other products may be trademarks or registered trademarks of their respective companies.

rnal 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
on 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
idual user will achieve throughput improvements equivalent to the performance ratios stated here.
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
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 c
ditions. 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
lt 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 with
d 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 co
lity, 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.
ge without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

© 2012 IBM Corporation

Agenda

- PAV/ HyperPAV Support
  - Motivation for PAV
  - PAV and Hyper PAV concepts
  - Device mapper based implementation
  - Configuring PAV volumes with multipath tools
  - DASD device driver based implementation
  - Hyper PAV
  - Tool support
  - Hints and Tips
- Large Volume Support
  - Large Volumes
  - Compatibility

© 2012 IBM Corporation

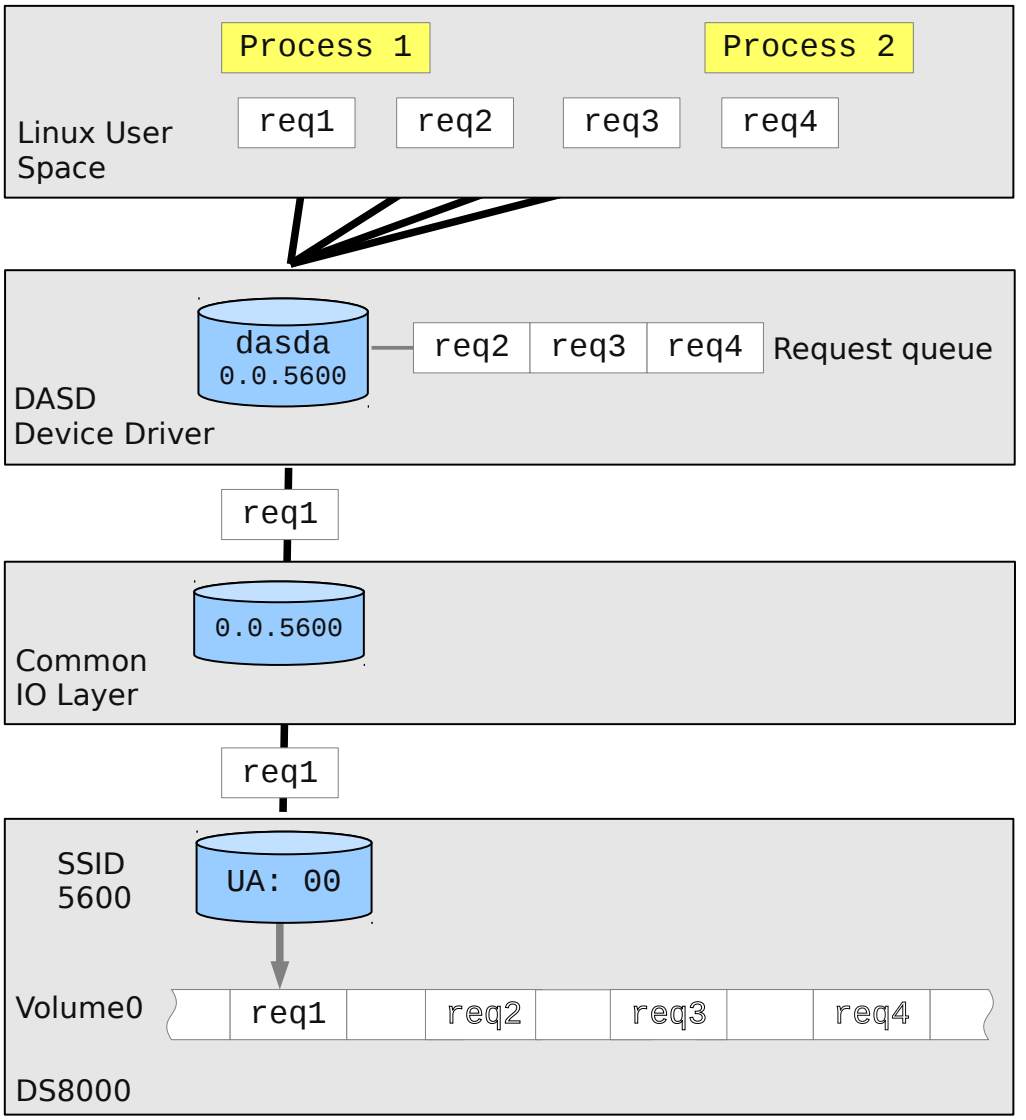
Problem Statement:
Linux on System z customers need to address more data with good performance and high availability!

© 2012 IBM Corporation

Normal DASD I/O – Why do we need PAV?

- One subchannel can execute one I/O request at a time.
- Programs running in parallel often access independent areas of one volume.
- Storage server could address independent areas in parallel (cache, striping, etc. on storage server).
- Sending requests in parallel would improve performance.

- Note: The SSID identifies the logical control unit (LCU) on the storage server.



© 2012 IBM Corporation

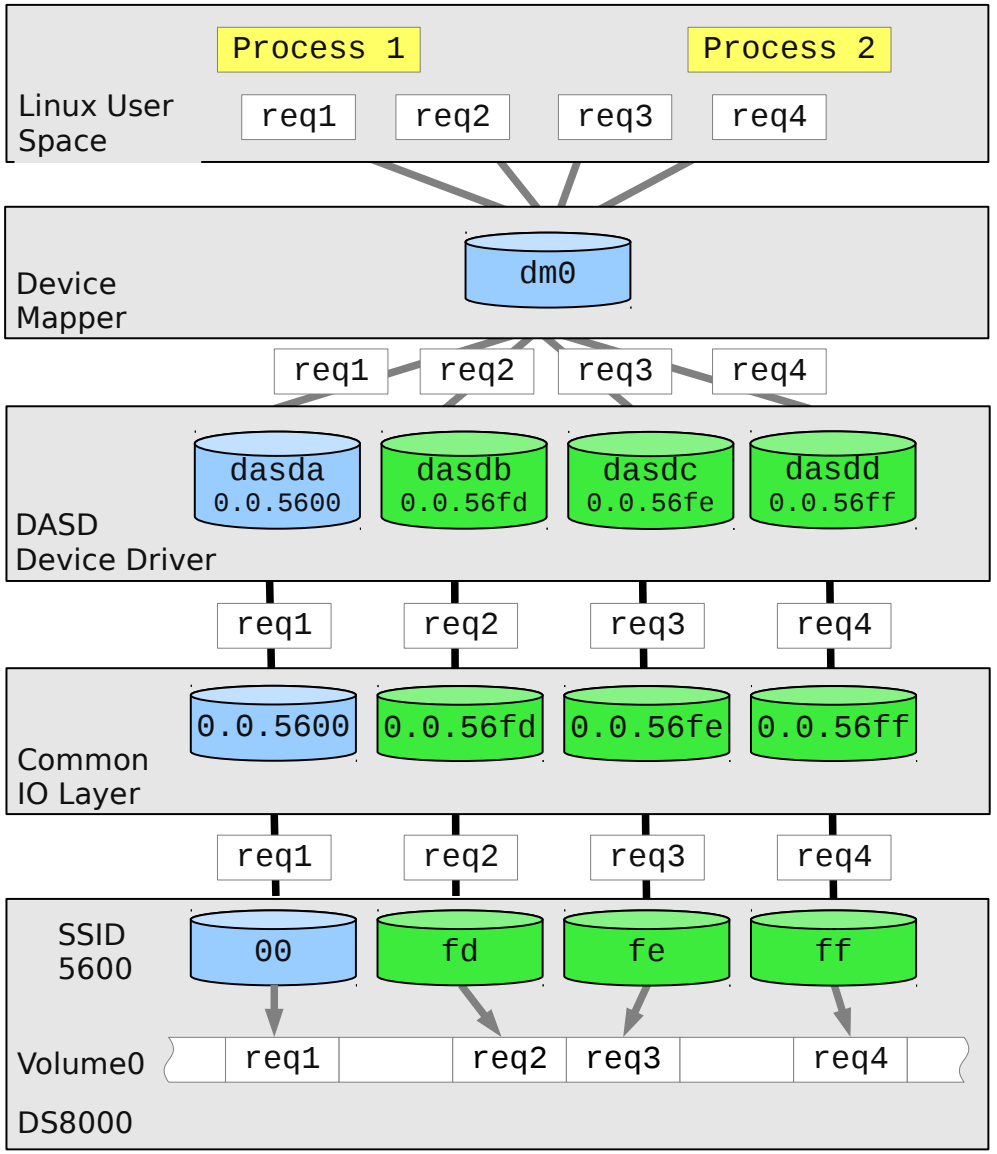
Parallel Access Volumes - Concepts

- Still one I/O per subchannel, but
- Several subchannels per volume.
- One PAV base device per volume.
- Several additional PAV alias devices.
- Base PAV
  - I/O on an alias will be directed to fixed base volume.
- Hyper PAV
  - I/O on alias may be directed to any base volume in the same logical control unit (LCU)
  - Backward compatible: Will work like Base PAV if the Linux version does not support Hyper PAV

© 2012 IBM Corporation

Original device mapper based implementation

- Used on SLES10 (<SP4) and RHEL4 & 5
- Through the PAV feature, storage systems can present the same physical disk space as a base device and one or more alias devices.
- The DASD device driver initially senses the base device
- Each DASD base device and each alias device has a separate device node assigned
- Example: If device 5600 is a base device and devices 56fd, 56fe and 56ff are alias devices, all device are available at the common IO layer and device nodes dasda to dasdd will be created
- This design relies on the fixed base/alias association.
- Finally device-mapper is required to combine the device dasda to dasdd to a single multipath device dm0



© 2012 IBM Corporation

Configuring PAV volumes with multipath tools

- If your Linux instance runs natively in an LPAR, the **nopav** keyword must not have been set for the **dasd=** kernel or module parameter.
- Issue **lsdasd** to ensure that device nodes exist for the PAV base volume and its aliases and that the devices are online.
- Use **chccwdev** to set the DASD online, if needed.
- Ensure that the device is formatted. If it is not already formatted, use **dasdfmt** to format it. Because a base device and its aliases all correspond to the same physical disk space, formatting either the base device or one of its aliases formats the base device and all alias devices.
  - Example: **dasdfmt -f /dev/dasdc**
- Ensure that the device is partitioned. If it is not already partitioned, use **fdasd** to create one or more partitions. The following command creates both a partition ***/dev/dasdc1*** for the base device and also a partition ***/dev/dasdd1*** for the alias.
  - Example: **fdasd -a /dev/dasdc**
- Set the base device and all its aliases offline and back online to assure that the device driver detects the partitions for each device name.
  - Example: **chccwdev -d 0.0.5600,0.0.56ff && chccwdev -e 0.0.5600,0.0.56ff**

© 2012 IBM Corporation

Configuring PAV volumes with multipath tools (cont'd)

- If it is not already loaded, load the dm_multipath module: (e.g: modprobe dm_multipath)
- Make sure that your multipath.conf configuration file does not contain blacklist-entries for dasd devices

```
# cat /etc/multipath.conf
…
blacklist {
        devnode "^(ram|raw|loop|fd|md|dm-|sr|scd|st)[0-9]*"
        devnode "^hd[a-z][[0-9]*]"
        devnode "^cciss!c[0-9]d[0-9]*[p[0-9]*]"
        devnode "^dasd[a-z]+[0-9]*"
}
…
```

- Use the multipath command to detect multiple paths to devices for failover or performance reasons and coalesce them
- Make sure multipathd is started (e.g: **`/etc/init.d/multipathd start`**)
- Enter **`multipath -ll`** to display the resulting multipath configuration.

© 2012 IBM Corporation

Configuring PAV volumes with multipath tools (cont'd)

```
# multipath -ll
IBM.75000000092461.2a00.1b dm-1 IBM,S/390
DASD ECKD [size=2.3G][features=0]
[hwhandler=0
\_ round-robin 0 [prio=4][enabled]
 \_ 0:0:10779:0 dasde 94:16 [active][ready]
 \_ 0:0:10924:0 dasdf 94:20 [active][ready]
 \_ 0:0:10925:0 dasdg 94:24 [active][ready]
 \_ 0:0:10926:0 dasdh 94:28 [active][ready]
IBM.75000000092461.2a00.1a dm-0 IBM,S/390
DASD ECKD [size=2.3G][features=0]
[hwhandler=0]
\_ round-robin 0 [prio=4][enabled]
 \_ 0:0:10778:0 dasdc 94:12 [active][ready]
 \_ 0:0:10927:0 dasdd 94:32 [active][ready]
```

- The DASDs can now be accessed as multipath devices IBM.75000000092461.2a00.1a and IBM.75000000092461.2a00.1b.
- You can find the corresponding device nodes in /dev/mapper.

```
/dev/mapper/IBM.75000000092461.2a00.1a
/dev/mapper/IBM.75000000092461.2a00.1ap1
/dev/mapper/IBM.75000000092461.2a00.1b
/dev/mapper/IBM.75000000092461.2a00.1bp1
/dev/mapper/control
```

- There is a device node for each multipath device and for each partition on these multipath devices.

You can now use LVM2 or an equivalent logical volume manager to configure the multipath device into a volume group, for example, for striping. If you use LVM2 to work with multipath devices, set a filter to ensure only the multipath devices are used and not the underlying base and aliases.

© 2012 IBM Corporation

z/VM PAV setup

- If your Linux system runs as a z/VM guest operating system, you can confirm the mapping of base and alias devices.
- After the hardware configuration with the base and alias device statements has become active, enter the z/VM QUERY PAV command

```
#CP QUERY PAV
00: Device  5600 is a base Parallel Access Volume with the
following aliases: 56FF
00: Device  56FF is an alias Parallel Access Volume device
whose base device is 5600
CP Q PAV  4DE1
00: Device  4DE1 is not a Parallel Access Volume
```

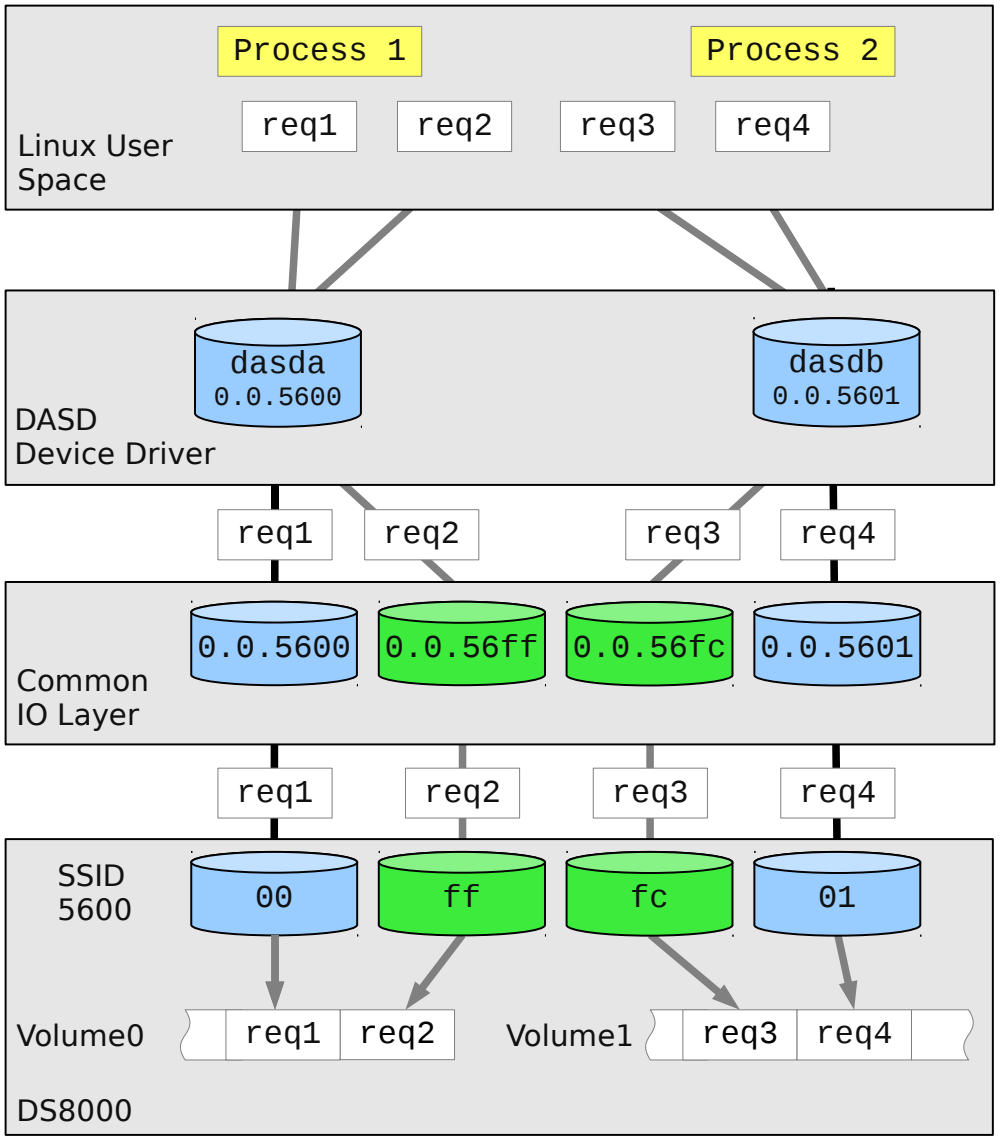
- To make a base device 0x5600 and its alias 0x56FF available to the current z/VM guest, enter the following commands:

```
#CP ATTACH 5600 *
#CP ATTACH 56FF *
```


© 2012 IBM Corporation

New DASD device driver implementation
Base PAV example

- Used on SLES11 and RHEL6
- Through the PAV feature, storage systems can present the same physical disk space as a base device and one or more alias devices.
- The DASD device driver initially senses the base device.
- The DASD device driver creates device nodes for the base devices but not for the aliases.
- The base device is set online.
- The aliases can lead to gaps in the naming scheme for device nodes.
- If multiple user space processes concurrently access a base device, the device driver uses the aliases to issue multiple channel programs.
- Apart from assuring that the corresponding aliases for a base device are online, user space processes need no special handling for accessing a PAV.



© 2012 IBM Corporation

PAV Prerequisites

- Before you can use PAV on your Linux instance, the PAV feature must be enabled on your storage system.
- The PAV feature is available, for example, for the following systems:
  - IBM System Storage DS8000 series systems
  - IBM System Storage DS6000 series systems
  - IBM TotalStorage Enterprise Storage Server (ESS)
- The HyperPAV feature is available, for example, for IBM System Storage DS8000 series systems.
- PAV base and alias volumes require special IOCDS specifications
- You need to know the device numbers of the base devices and their aliases as defined on the storage system.
- If your Linux system runs as a z/VM guest operating system, you need privilege class B authorization.

© 2012 IBM Corporation

IOCDS Configuration

- Configuring base and alias volumes for PAV or HyperPAV on the storage system is beyond the scope of this presentation.
  - See your storage system documentation for details.
  - For information about IOCDS specifications for multiple subchannel sets see the Input/Output Configuration Program User's Guide for your mainframe system.
- The IOCDS examples in this presentation apply to mainframe systems with a single subchannel set.
- Perform the following steps to define the base devices and their aliases to the hardware:
  - Define the base devices to the storage hardware
  - Define the alias devices to the storage hardware

- Example: The following statement defines device number 0x5600 as a base device.

```
IODEVICE ADDRESS=(5600),UNITADD=00,
CUNUMBR=(5600), *
STADET=Y,UNIT=3390B
```

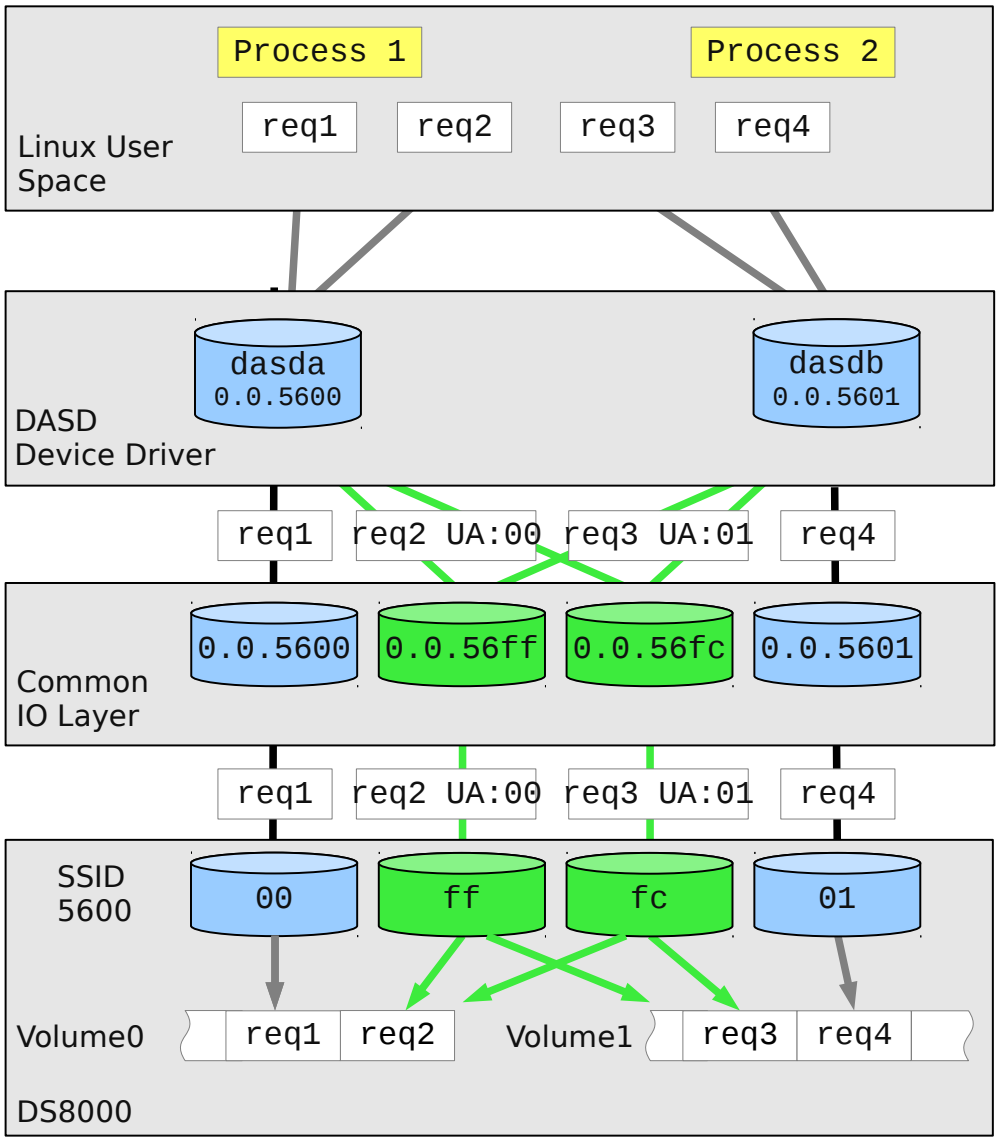
- Example: The following statement defines device 0x56ff as an alias device. The mapping to the associated base device 0x5600 is given by the storage system configuration.

```
IODEVICE ADDRESS=(56FF),UNITADD=FF,
CUNUMBR=(5600), *
STADET=Y,UNIT=3390A
```


© 2012 IBM Corporation

New DASD device driver implementation
Hyper PAV example

- With the old base-PAV support the mapping between base and alias devices was static.
- Now HyperPAV removes the requirement to dedicate alias devices to specific base devices.
- Alias devices are used with all base devices of the same LCU
- Target unit address is encoded into the request itself
- Same user interface as Base PAV
- HyperPAV is activated automatically when the necessary prerequisites are there (DS8000 with HyperPAV LIC, z/VM 5.3)
- SLES11, SLES10 SP4, RHEL6
- If the prerequisites for HyperPAV are not there, base-PAV is used if the PAV feature is enabled on the storage server. Otherwise the DASD driver works without using PAV.



© 2012 IBM Corporation

HyperPAV Setup

- HyperPAV Base and Alias subchannels are defined on control unit's Hardware Management Console and in IOCDS no differently than traditional PAVs
- HyperPAV hardware, priced feature enables floating Alias function associated with the HyperPAV architecture for each LSS (logical control unit)
- Operating system host determines which LCU (logical control unit) is in HyperPAV vs. traditional PAV mode

(1)Setup PAV configuration on Storage Server

(2)System z storage configuration (IOCP)

(3)Basic DASD configuration

(4)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

© 2012 IBM Corporation

z/VM HyperPAV configuration

- VM dedicated DASD support via CP ATTACH command or DEDICATE user directory statement
- VM Minidisk Support:
  - workload balancing for guest's that don't exploit HyperPAV
  - linkable full-pack minidisks for guests that do exploit HyperPAV
  - New `CP DEFINE HYPERPAVALIAS` command creates HyperPAV Alias minidisks for exploiting guests
  - z/VM, z/OS and Linux on System z are current exploiters of HyperPAV
  - Restricted to fullpack minidisks for exploiting guests; architecture change in the works.
- Use the Class B, CP QUERY PAV command to view the current HyperPAV Base and Alias subchannels along with their associated Pools.

© 2012 IBM Corporation

PAV/HyperPAV Toolbox

New sysfs interfaces can allow to identify devices:

```
$ ls -al /sys/bus/ccw/devices/0.0.5600/
total 0
drwxr-xr-x 4 root root    0 Sep  2 16:25 .
drwxr-xr-x 4 root root    0 Sep  2 16:25 ..
-r--r--r-- 1 root root 4096 Sep  2 16:25 alias
-r--r--r-- 1 root root 4096 Sep  2 16:33 availability
drwxr-xr-x 3 root root    0 Sep  2 16:25 block
-rw-r--r-- 1 root root 4096 Sep  2 16:33 cmb_enable
-r--r--r-- 1 root root 4096 Sep  2 16:33 cutype
-r--r--r-- 1 root root 4096 Sep  2 16:33 devtype
-r--r--r-- 1 root root 4096 Sep  2 16:25 discipline
lrwxrwxrwx 1 root root    0 Sep  2 16:25 driver ->
../../../../bus/ccw/drivers/dasd-eckd
-rw-r--r-- 1 root root 4096 Sep  2 16:33 eer_enabled
-rw-r--r-- 1 root root 4096 Sep  2 16:33 erplog
-rw-r--r-- 1 root root 4096 Sep  2 16:33 failfast
-r--r--r-- 1 root root 4096 Sep  2 16:33 modalias
-rw-r--r-- 1 root root 4096 Sep  1 18:35 online
drwxr-xr-x 2 root root    0 Sep  2 16:25 power
-rw-r--r-- 1 root root 4096 Sep  2 16:25 readonly
-r--r--r-- 1 root root 4096 Sep  2 16:33 status
lrwxrwxrwx 1 root root    0 Sep  2 16:33 subsystem ->
../../../../bus/ccw
-rw-r--r-- 1 root root 4096 Sep  2 16:25 uevent
-r--r--r-- 1 root root 4096 Sep  2 16:25 uid
-rw-r--r-- 1 root root 4096 Sep  2 16:33 use_diag
-r--r--r-- 1 root root 4096 Sep  2 16:33 vendor
```

'alias': 0 for base device, 1 for alias device

'uid': unique-id of the base device (vendor.serial.SSID.UA)

© 2012 IBM Corporation

PAV/HyperPAV Toolbox (cont'd)

- Base/Hyper PAV base device:

```
uid = IBM.75000000010671.5600.00  alias = 0
```

- Base PAV alias device:

```
uid = IBM.75000000010671.5600.00  alias = 1
```

- Hyper PAV alias device:

```
uid = IBM.75000000010671.5600.xx  alias = 1
```

- On z/VM multiple minidisks can reside on the same device.
  An additional qualifier allows to distinguish them.
  (needs PTFs for VM APAR VM64273 on z/VM 5.2.0 and higher and SLES 10 SP2 or RHEL 5.3)

  Long uid for VM :

```
uid = IBM.75000000010671.5600.00.0000000000037400000000000000000
```


© 2012 IBM Corporation

PAV/HyperPAV Toolbox (cont'd)

- New option -u /--uid allows to display and sort by uid.
- Groups of base and alias devices are easy to identify:

```
$ lsdasd -u
Bus-ID    Name      UID
============================================
0.0.7500  dasde     IBM.75000000010671.7500.00
0.0.7501  dasdf     IBM.75000000010671.7500.01
0.0.75fb  alias     IBM.75000000010671.7500.xx
0.0.75fc  alias     IBM.75000000010671.7500.xx
0.0.75fe  alias     IBM.75000000010671.7500.xx
0.0.75ff  alias     IBM.75000000010671.7500.xx
0.0.7e9e  dasda     IBM.75000000058251.7e00.9e
0.0.7e9f  dasdb     IBM.75000000058251.7e00.9f
0.0.5600  dasdc     IBM.75000000092461.5600.00
0.0.56fe  alias     IBM.75000000092461.5600.00
0.0.56ff  alias     IBM.75000000092461.5600.00
0.0.5601  dasdd     IBM.75000000092461.5600.01
0.0.56fb  alias     IBM.75000000092461.5600.01
0.0.56fc  alias     IBM.75000000092461.5600.01
```

```
Hyper PAV group

Base PAV group

Base PAV group
```

- For a full description of all features see **'man lsdasd'**.

© 2012 IBM Corporation

PAV/HyperPAV Toolbox (cont'd)

- **`dasdinfo`** provides various device identifiers as used in scripts and configuration files.
- Example 1: export all information (e.g. for use in udev):

```
$ ./dasdinfo -a -e -b dasde
ID_BUS=ccw
ID_TYPE=disk
ID_UID=IBM.7500000092461.4a00.df
ID_XUID=IBM.7500000092461.4a00.df.0000001500000032000000000000000
ID_SERIAL=0X0202
```

- Example 2: print extended uid (e.g. callout in multipath.conf):

```
$ /sbin/dasdinfo -x -b dasda
7500000092461.4a00.df.0000001500000032000000000000000
```

- Example 3: print uid (e.g. callout in multipath.conf):

```
$ /sbin/dasdinfo -u -b dasda
IBM.7500000092461.4a00.df
```

- For a full description of all features see **`'man dasdinfo'`**.

© 2012 IBM Corporation

PAV Performance

- Generally speaking, disk performance depends on the time required to do a single I/O operation on a single disk volume
- This time is generally understood as `response time`, which is composed of `queue time` and `service time`. Queue time is the time a guest's I/O spends waiting to access the real volume.
- Service time is the time required to finish the real I/O operation.
- PAV can help to decrease the access time
- But if a volume is not experiencing queuing, adding aliases does not help to enhance the volume's performance, for the time required to perform an I/O operation is not appreciably changed by the queue time.
- In some cases, increasing the number of aliases for a volume might result in increased service time for that volume.
- Most of the time, the decrease in wait time outweighs the increase in service time, so the response time improves.
- The z/VM Performance Toolkit provides a report on the response time and related performance statistics for the real devices. Use the command DEVICE in the performance toolkit

© 2012 IBM Corporation

General Hints and Tips

- Base PAV and Hyper PAV are priced features (LIC) of your storage server.
- No support for Dynamic PAV!
  - If you use a workload manager to manage PAV alias devices with Dynamic PAV, you must exclude Linux devices.
- On LPAR: Alias devices only become 'visible' after at least one base device of the same LCU (same SSID) has been set online.
  - It may take a few seconds before the alias devices are available.
- Alias devices 'consume' minor numbers and device names.
  - Example: The third device in your configuration will always get the internal name 'dasdc' assigned, even if it is an alias device and not a block device on it's own.

© 2012 IBM Corporation

General Hints and Tips (cont'd)

- Examples in this presentation use a very simple naming scheme:
  - Device number = SSID + Unit Address
- This is just one possible definition.
- Actual device numbers will depend on your I/O configuration (HCD / IOCP).
- It is possible to define alias devices in the second subchannel set.
- Example:
  - Base device: 0.0.5600
  - Alias device: 0.1.5600
- Resources on the web:
  http://www.ibm.com/developerworks/linux/linux390/development_documentation.html
- PDF for download: 'How to Improve Performance with PAV'

© 2012 IBM Corporation

General Hints and Tips (cont'd)

- Basic instructions on using dm multipath:
  - SLES10: Storage administration Guide,
    Chapter: Managing Multipath I/O for Devices
  - RHEL5: DM Multipath, DM Multipath Configuration and Administration
- When working directly with a base device (e.g. dasdfmt, fdasd) make sure to set all alias devices offline.
- DASDs are blacklisted per default.
  Add a blacklist exception to your */etc/multipath.conf*:

```
blacklist_exceptions {
        device {
                vendor IBM
                product S/390.*
        }
}
```


© 2012 IBM Corporation

PAV vs. HyperPAV

- PAV
  - Well defined mapping between base and alias devices.
  - This results in a predictable resource utilization
- HyperPAV
  - Dynamic allocation and utilization of resources via pooled alias devices
  - Advantage: In average this leads to a higher utilization, resulting in I/O transfer rates

If you run current Linux on System z distributions, prefer HyperPAV over PAV

© 2012 IBM Corporation

Problem Statement:
Legacy ECKD size limitation

© 2012 IBM Corporation

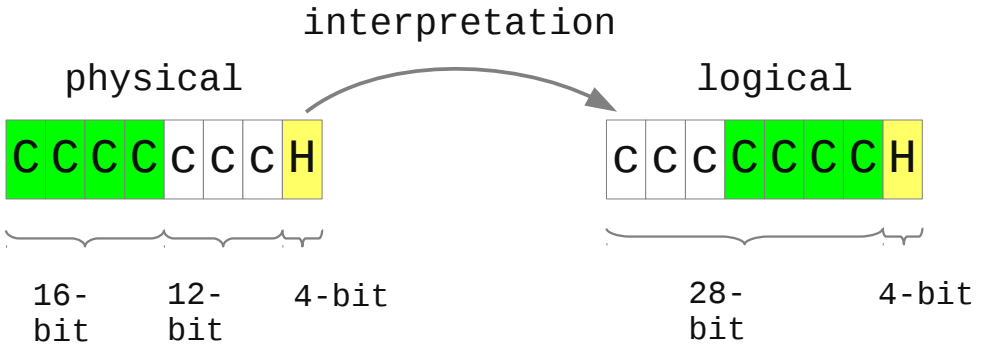
Large Volume Support

- Largest model 9 ECKD DASD:
  - 65520 cylinders (0xFFF0), 15 heads
  - 54GB raw capacity, or
  - 45GB when formatted with 4096 byte blocks
- Size is restricted by the legacy ECKD interface
  - 16-bit cylinder address
  - 16-bit head address, but
    more then 15 heads would break legacy code.
- Legacy ECKD interface is implemented in many software systems (z/OS, z/VM, z/VSE).
- Extension must be compatible to legacy software.

© 2012 IBM Corporation

Large Volume Support (cont'd)

- Number of heads per cylinder cannot be extended without breaking legacy software.
- Idea: Use 'free' head bits to extend the cylinder address:



- 3390 Model A
- Other name: Extended Address Volume (EAV)
- Currently up to 262668 cylinders
  (approximately 223.2 GB raw, 180GB formatted)

© 2012 IBM Corporation

Linux Support

- Upstream Kernel 2.6.30, s390-tools 1.8.2
- Included in distributions:
  - SLES11 SP1
  - SLES10 SP3
  - RHEL6
  - RHEL5.5
- Large Volumes can be used like any other volume
  - Usable as boot / IPL device
  - Usable as dump device
- Can be combined with PAV / Hyper PAV

© 2012 IBM Corporation

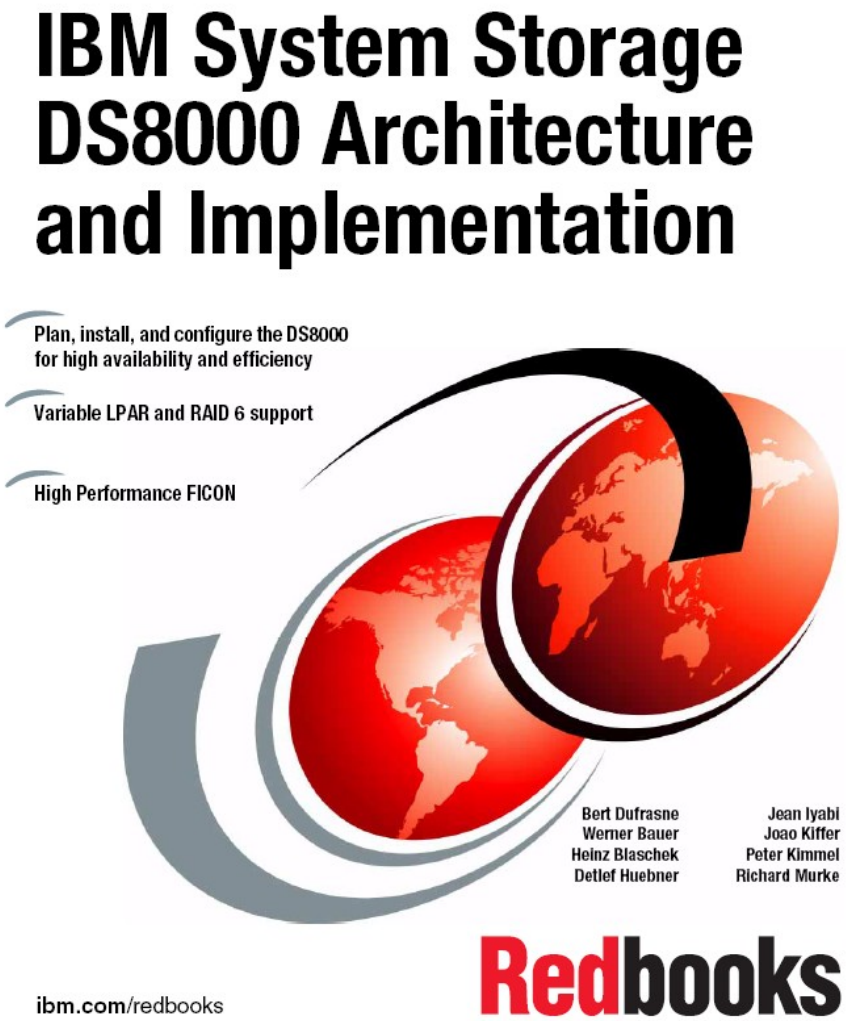
Compatibility

- Kernel and tools without Large Volume support will recognize such a volume with 65535 (0xFFFE) cylinders.
- When formatted without Large Volume support, part of the DASD will stay unformatted and unusable.
  - When moved to a system with large volume support, the formatted part will stay usable.
- Caution:
  When partitions were created on a system with Large Volume support, these partitions are not recognized correctly on a system without support.
  - CDL partitions are not recognized at all
  - LDL partition will be recognized with a wrong size
- Recommendation:
  Use Large Volumes only with kernel and tools that support them.

© 2012 IBM Corporation

More Information

http://www.vm.ibm.com/storman/pav/pav3.html



© 2012 IBM Corporation

More Information (cont'd)

ibm.com/systems/z/linux

www.vm.ibm.com

© 2012 IBM Corporation

Questions?

© 2012 IBM Corporation

Appendix

© 2012 IBM Corporation

Impact of DS8000 Storage Pool Striping

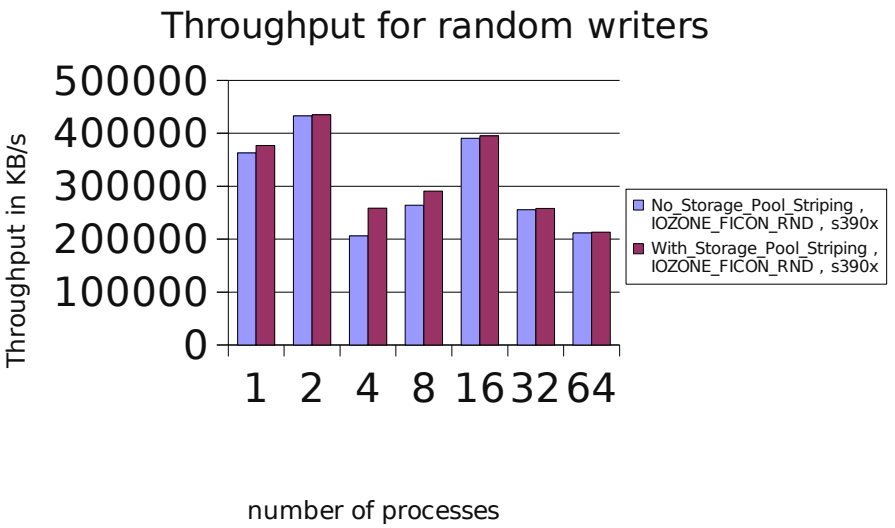
Starting with DS8000 License Machine Code 5.30xx.xx, it is possible to stripe the extents of a DS8000 Logical Volume across multiple RAID arrays.

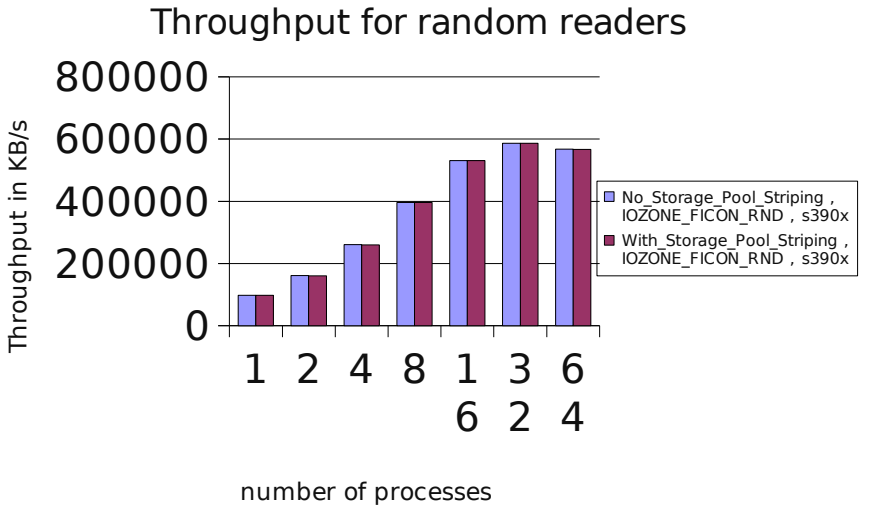
DS8000 Storage Pool Striping will improve throughput for some workloads.

It is performed on a 1 GB granularity, so it will generally benefit random workloads more than sequential ones.

If you are already using a host-based striping method (for example, LVM Striping or DB2 database container striping), there is no need to use Storage Pool Striping.

However, it is possible. You should then combine the wide stripes on DS8000 with small granularity stripes on the host. The recommended size for these is usually between 8 and 64 MB. If large stripes on both DS8000 and attached host interfere with each other, I/O performance may be affected.





© 2012 IBM Corporation

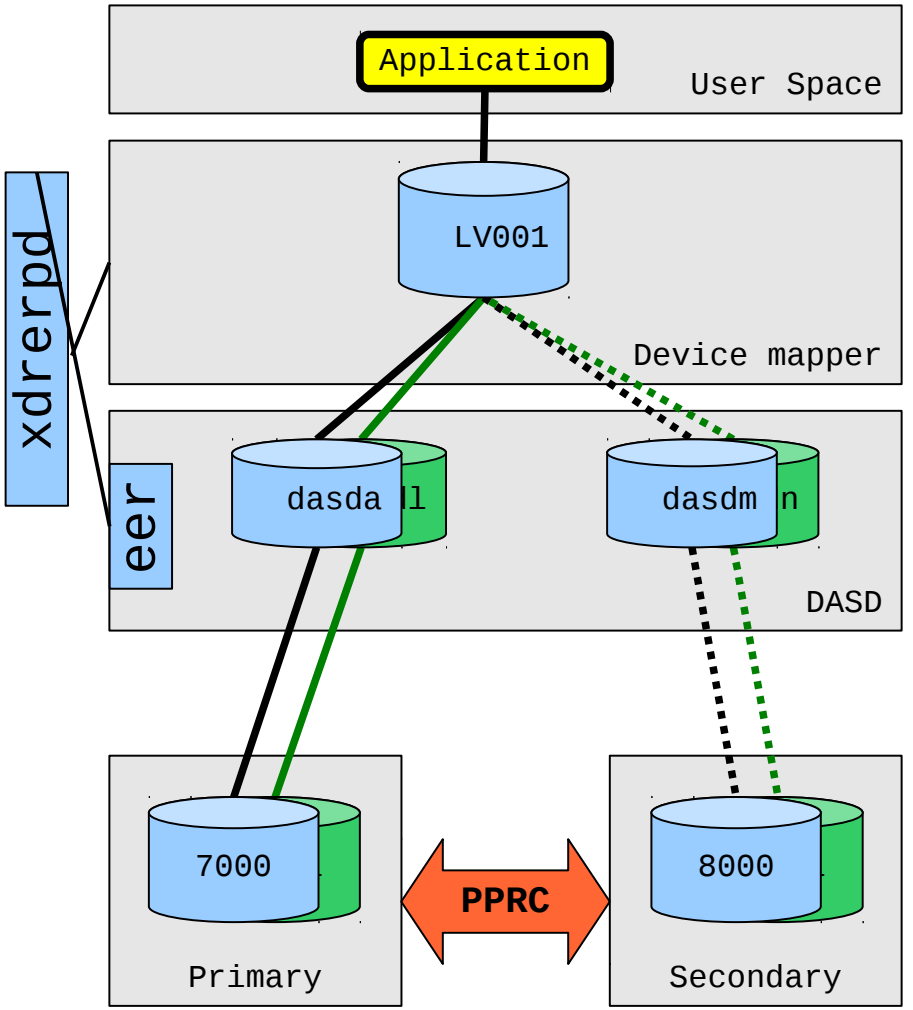
HyperSwap Support in DASD and CIO

- Base support needed to join GDPS/PPRC environment with linux running on LPAR
  - Continuous availability solution
  - Protect against local area disasters
- Switchable through sysfs attribute `'eer_enabled'`
  /sys/bus/ccw/device/<busid>/eer_enabled
- Configurable buffer size for reporting device
  DASD module parameter 'eer_pages' determines number of pages user for internal error record buffering

© 2012 IBM Corporation

HyperSwap support in DASD and CIO - Structure

- System managed by GDPS running on z/OS
- DASD (CIO) supports detection, internal handling and reporting of I/O errors (eer)
- Device swap performed by device-mapper
- DASD driver supports quiesce / resume and enable / disable of devices



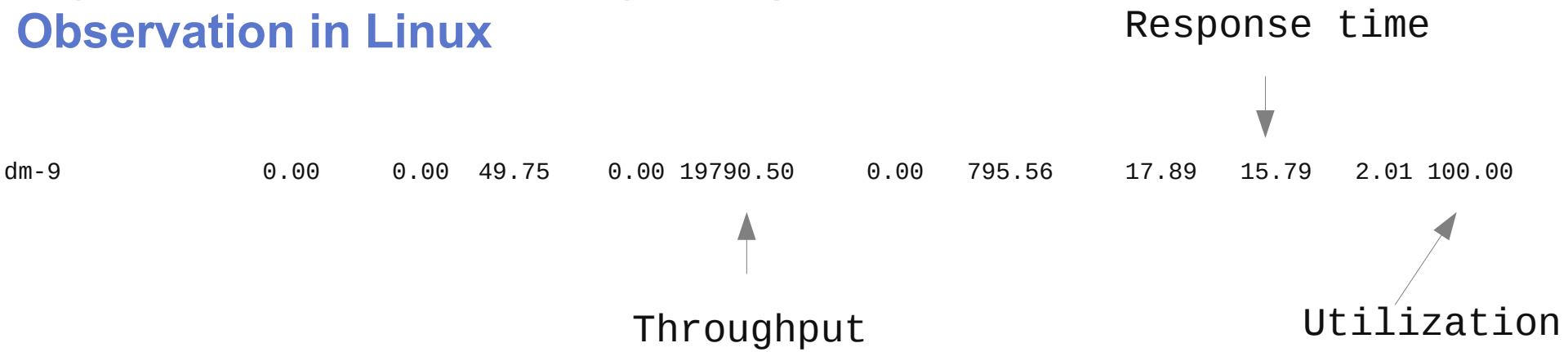
© 2012 IBM Corporation

High Disk response times

- **Configuration:**
  - z10, HDS Storage Server (Hyper PAV enabled)
  - z/VM, Linux with Oracle Database
  - VM controlled Minidisks attached to Linux, LVM on top
- **Problem description:**
  - I/O throughput not matching expectations
  - Oracle Database shows poor performance because of that
  - One LVM volume showing significant stress
- **Tools used for problem determination:**
  - dbginfo.sh
  - sadc/sar
  - z/VM Monitor data

© 2011 IBM Corporation

High Disk response times (cont'd)
Observation in Linux

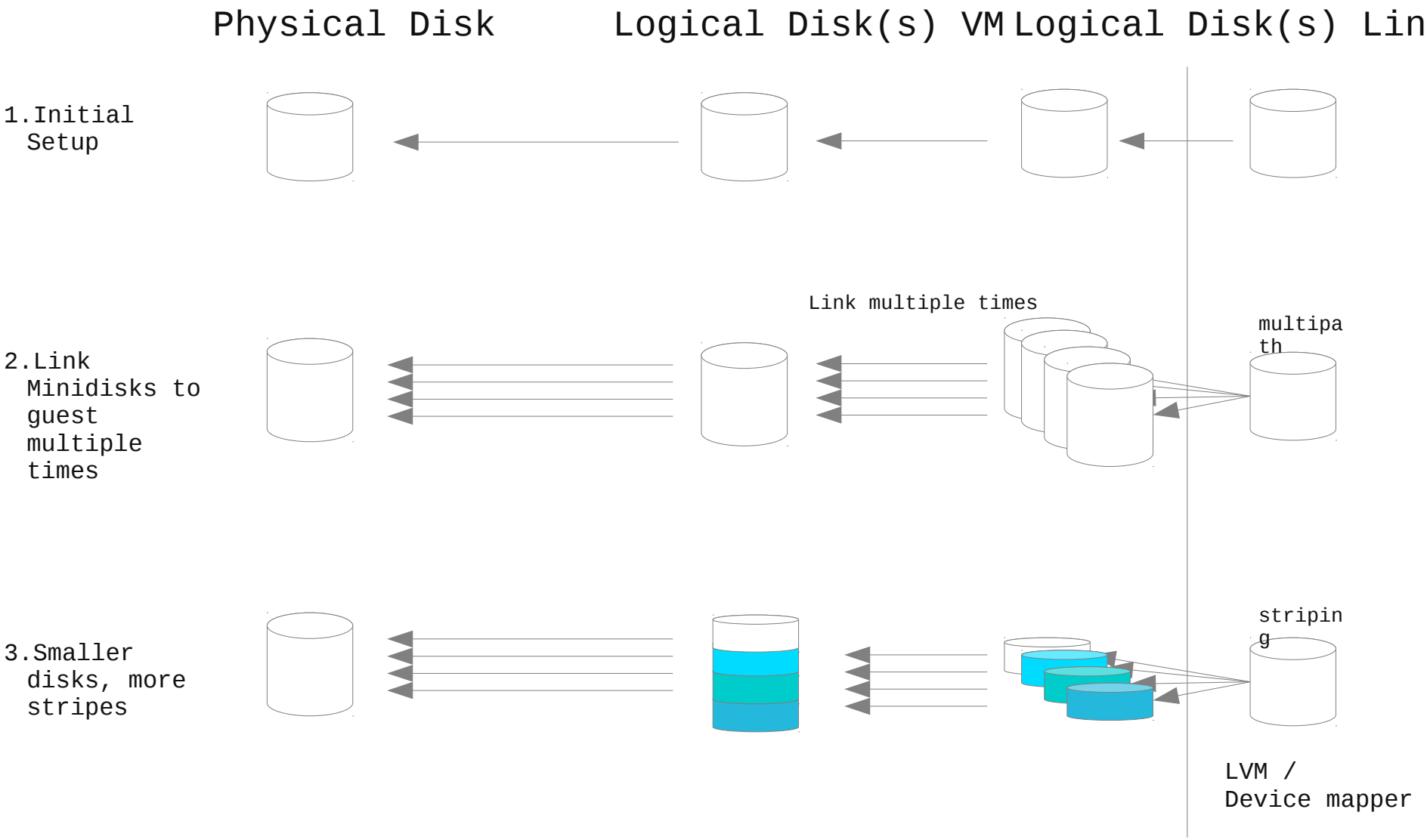


Conclusion

- **PAV not being utilized**
- **No Hyper PAV support in SLES10 SP2**
- **Static PAV not possible with current setup (VM controlled minidisks)**
- **Need to look for other ways for more parallel I/O**
  - Link same minidisk multiple times to a guest
  - Use smaller minidisks and increase striping in Linux

© 2011 IBM Corporation

High Disk response times (cont'd)
Initial and proposed setup



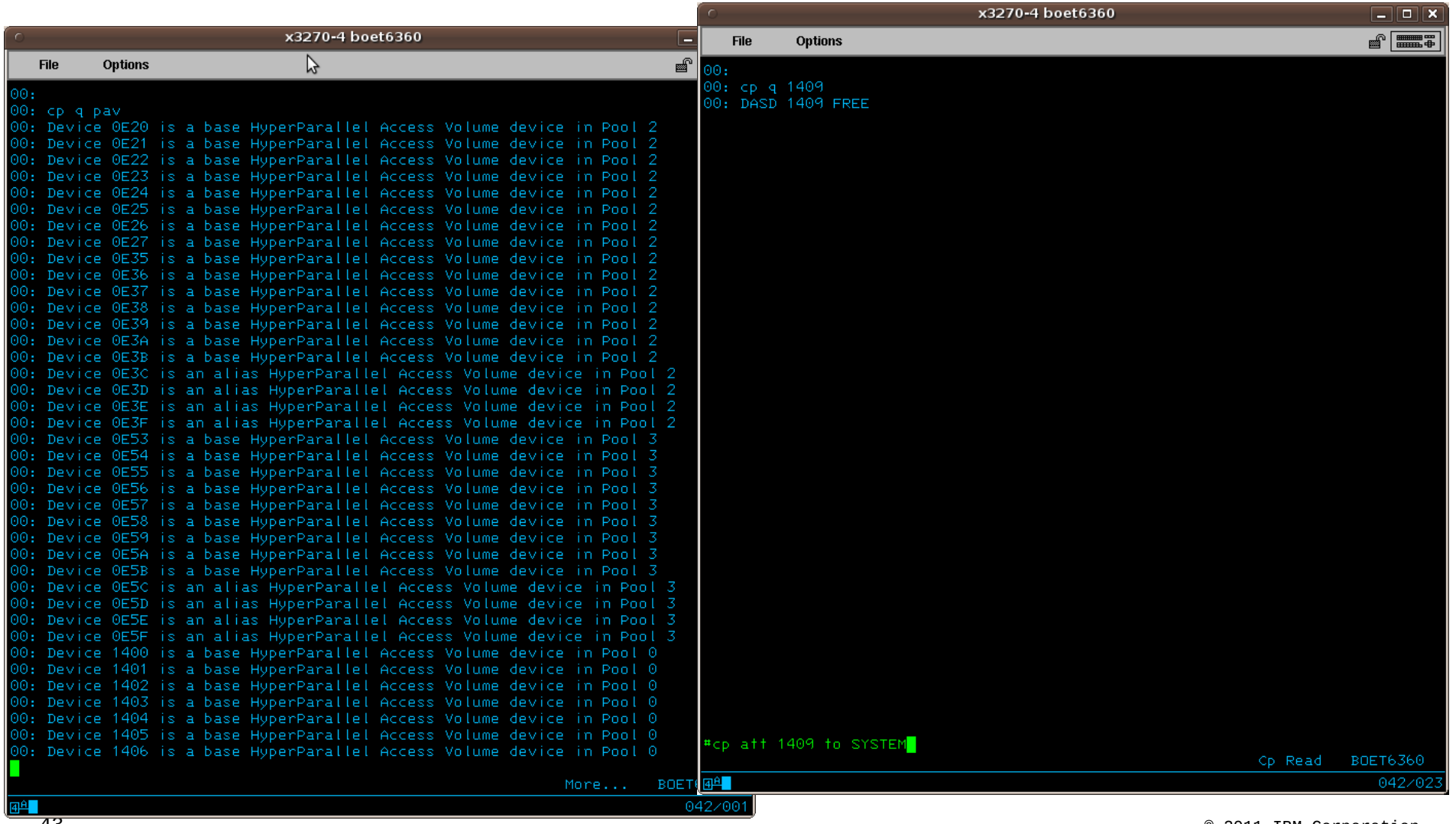
© 2011 IBM Corporation

High Disk response times (cont'd)
New Observation in Linux

- Response times stay equal
- Throughput equal
- No PAV being used!!

© 2011 IBM Corporation

High Disk response times (cont'd)
Solution: check PAV setup in VM



© 2011 IBM Corporation

High Disk response times (cont'd)
Finally:

dasdaen	133.33	0.00	278.11	0.00	12298.51	0.00	88.44	0.79	2.83	1.48	41.29
dasdcbt	161.19	0.00	248.26	0.00	12260.70	0.00	98.77	0.91	3.47	1.88	46.77
dasdfwc	149.75	0.00	266.17	0.00	12374.13	0.00	92.98	1.88	7.07	2.54	67.66
dasdael	162.19	0.00	250.25	0.00	12483.58	0.00	99.77	1.90	7.57	2.86	71.64
dasddyz	134.83	0.00	277.61	0.00	12431.84	0.00	89.56	0.75	2.71	1.68	46.77
dasdaem	151.24	0.00	266.17	0.00	12595.02	0.00	94.64	2.01	7.61	2.82	75.12
dasdcbr	169.65	0.00	242.79	0.00	12386.07	0.00	102.03	1.72	7.05	2.83	68.66
dasdfwd	162.69	0.00	249.25	0.00	12348.26	0.00	99.08	1.92	7.70	2.83	70.65
dasddyy	157.21	0.00	259.70	0.00	12409.95	0.00	95.57	2.58	9.96	3.05	79.10
dasddyx	174.63	0.00	237.81	0.00	12374.13	0.00	104.07	1.76	7.38	2.93	69.65
dasdcbs	144.78	0.00	272.14	0.00	12264.68	0.00	90.14	2.53	9.31	2.89	78.61
dasda	0.00	0.00	0.00	1.00	0.00	3.98	8.00	0.01	10.00	5.00	0.50
dasdq	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
dasdss	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
dasdadx	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.			
dasdawh	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
dasdamk	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
dasdaek	160.70	0.00	255.22	0.00	12382.09	0.00	97.03	2.27	8.95	2.88	73.63
dasdcbq	148.76	0.00	265.67	0.00	12372.14	0.00	93.14	2.14	8.01	2.85	75.62
dasddyw	162.19	0.00	254.23	0.00	12384.08	0.00	97.42	2.12	8.40	2.90	73.63
dasdfwe	146.27	0.00	271.64	0.00	12419.90	0.00	91.44	2.63	9.71	2.80	76.12
dasdfwf	162.19	0.00	249.75	0.00	12455.72	0.00	99.75	0.71	2.83	1.79	44.78
dasdb	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
dm-0	0.00	0.00	1646.77	0.00	49494.53	0.00	60.11	5.08	3.04	0.36	59.70
dm-1	0.00	0.00	1665.17	0.00	49482.59	0.00	59.43	15.00	9.04	0.56	93.53
dm-2	0.00	0.00	1660.70	0.00	49432.84	0.00	59.53	13.46	8.11	0.55	90.55
dm-3	0.00	0.00	1647.26	0.00	49490.55	0.00	60.09	12.05	7.32	0.53	87.56
dm-4	0.00	0.00	1646.77	0.00	49494.53	0.00	60.11	5.08	3.04	0.36	59.70
dm-5	0.00	0.00	1665.17	0.00	49482.59	0.00	59.43	15.00	9.04	0.56	93.53
dm-6	0.00	0.00	1660.70	0.00	49432.84	0.00	59.53	13.46	8.11	0.55	90.55
dm-7	0.00	0.00	1647.26	0.00	49490.55	0.00	60.09	12.06	7.32	0.53	87.56
dm-8	0.00	0.00	0.00	1.99	0.00	7.96	8.00	0.00	0.00	0.00	0.00
dm-9	**0.00**	**0.00**	**497.51**	**0.00**	**197900.50**	**0.00**	**795.56**	**7.89**	**15.79**	**2.01**	**100.00**

© 2011 IBM Corporation