

How to Improve Performance with PAV November 17, 2006

Linux Kernel 2.6 (October 2005 stream)



How to Improve Performance with PAV November 17, 2006

Linux Kernel 2.6 (October 2005 stream)

Note Before using this information and the product it supports, read the information in "Notices" on page 13.				
Second Edition (November 2006)				

This edition applies to Linux kernel 2.6 (October 2005 stream) and to all subsequent releases and modifications until otherwise indicated in new editions.

This edition replaces SC33-8292-00 for the October 2005 stream but does not apply to the April 2004 stream. For the April 2004 stream use SC33-8324-00.

© Copyright International Business Machines Corporation 2004, 2006. All rights reserved. US Government Users Restricted Rights – Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Summary of changes v	Enabling volumes for PAV
About this publication vii Where to get more information vii	guest operating system
Chapter 1. Introduction to PAV 1	multipath tools
Parallel Access Volumes	Chapter 3. Partitioning devices that have been set up for multipathing 11
Chapter 2. Setting up disk volumes for	Notices

Summary of changes

This revision reflects changes to the October 2005 stream until November 2006. This revision does not apply to the April 2004 stream. For the April 2004 stream use SC33-8324-00.

- You can now use PAV for Linux® systems that run in LPAR mode.
- PAV support is now based on multipath tools.

This revision also includes maintenance and editorial changes.

About this publication

This document describes how to set up DASD volumes as parallel access volumes (PAV) using multipath tools.

In this book, System z^{TM} is taken to include System $z9^{\text{TM}}$, $zSeries^{\text{(B)}}$ in 64- and 31-bit mode, as well as $S/390^{\text{(B)}}$ in 31-bit mode.

You can find the latest version of this document on developerWorks[®] at: www.ibm.com/developerworks/linux/linux390/october2005 documentation.html

Where to get more information

This document describes how to set up PAV with the help of multipath tools as included in your distribution. For information on multipathing see christophe.varoqui.free.fr/wiki/wakka.php?wiki=Home.

For information on device mapper, a component of the Linux kernel that supports logical volume management, see sources.redhat.com/dm/.

For more information on the DASD device driver refer to *Device Drivers, Features, and Commands*. You can obtain the latest version of this book on developerWorks at:

ibm.com/developerworks/linux/linux390/october2005 documentation.html

You can use LVM2 to build on the PAV configuration described in this document. For information on LVM2 see sourceware.org/lvm2/.

Chapter 1. Introduction to PAV

The concurrent operations capabilities of the IBM® TotalStorage® enterprise disk storage systems and IBM TotalStorage Enterprise Storage Server® (ESS), support concurrent data transfer operations to or from the same volume from the same system or system image. A volume that can be accessed in this way is called Parallel Access Volume (PAV).

The DASD device driver does not attempt to start more than one I/O operation at a time to a device, but today's storage subsystem design, with large caches and RAID 5 arrays, makes it possible for the storage control unit to do I/O operations in parallel.

When software is using PAV, it can issue multiple channel programs to a volume, allowing simultaneous access to the logical volume by multiple users or processes. Reads can be satisfied simultaneously, as well as writes to different domains. The domain of an I/O consists of the specified extents to which the I/O operation applies. Writes to the same domain still have to be serialized to maintain data integrity.

Prerequisites: Linux on an IBM System z mainframe can use PAV if the volume resides on either of:

- An IBM TotalStorage Enterprise Storage Server (ESS)
- · An IBM TotalStorage enterprise disk storage system

Parallel Access Volumes

Through the PAV feature, storage systems can present the same physical disk space as a base device and one or more alias devices. On the System z mainframe, the base device and the aliases are all configured with a separate device number. Figure 1 illustrates a disk space that can be accessed as a base device or one of two aliases.

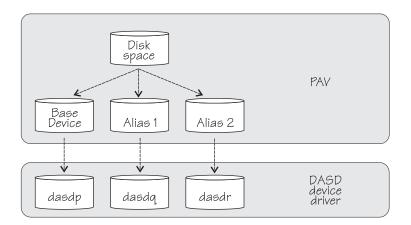


Figure 1. A PAV volume as seen by the DASD device driver

On a Linux system that has access to a base device and its aliases, the DASD device driver senses the base device and each alias as a different, independent DASD and assigns a different device name to each.

After a DASD is partitioned and the corresponding base device and all aliases are newly initialized, the DASD device driver detects the partitions and assigns additional device names for the partitions on the base device and on each alias.

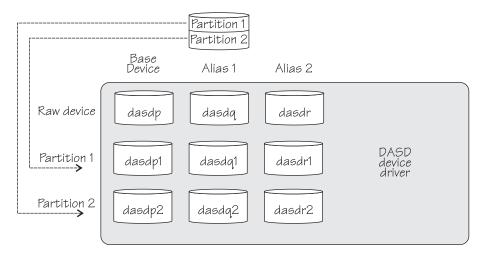


Figure 2. PAV with partitions

As illustrated in Figure 2, the DASD device driver has the following number of device names for the same physical disk space: (number of aliases + 1) × (number of partitions + 1).

Multipathing

To ensure data integrity, Linux must use a multipath tool. To configure your PAV base and alias devices for multipathing, use the multipath tools included in your distribution. The multipath tool detects that a base device and its aliases are, in fact, all representations of the same physical disk space.

As illustrated in Figure 3, the multipath tools summarize the base and alias representations of the raw device and of each partition into multipath devices and assign unique multipath device-names to them. User-space programs must access multipath devices through the multipath device-names. The multipath tools prevent user-space programs from writing to the other device names by occupying the device names with write access.

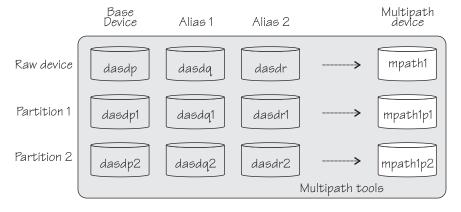


Figure 3. PAV as a multipath device

Depending on your environment, the device names for the multipath devices can be the unique ID that the multipath tools use to identify a multipath device or an alias. Table 1 summarizes the possible formats for raw devices:

Table 1. Formats for device names of multipath devices

Device name based on	Format	Example	
Unique ID using the DASD uid sysfs attribute	IBM. <xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx< td=""><td colspan="2">IBM.75000000092461.2a00.09</td></xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx<>	IBM.75000000092461.2a00.09	
Unique ID using volser	String of up to 6 characters	V0L001	
Generated alias	mpath <n></n>	mpath1	
	where $\langle n \rangle$ is an integer.		
Custom-defined alias	Depends on your specifications.	pav.vol001	

The device names for the first, second, and third partition are the device names of the raw devices with p1, p2, or p3 appended. For example, the three partitions on a multipath device for which the raw device is mpath1 would be mpath1p1, mpath1p2, and mpath1p3.

If the multipath tools in your environment use volsers or custom-defined aliases, be sure that your volsers are unique or that you specify unique aliases.

Unique IDs

The multipath tools use unique IDs to identify multipath devices, regardless of whether these unique IDs are also used as device names or whether alias device names are used. The unique ID for a device can be either of:

- The volser of the DASD
- · The value of the uid sysfs attribute of the DASD

You can use a command of this form to find out which information is used for the unique ID:

```
# multipath -d -v3 /dev/<device-name> |grep uid
```

where *<device-name>* is the device name of any DASD that is eligible for multipathing. In the output, the device name is shown as the value for the unique ID parameter (uid).

Example:

```
# multipath -d -v3 /dev/dasdn |grep uid
...
uid = IBM.75000000092461.2a00.09 (callout)
uid = IBM.75000000092461.2a00.09 (cache)
```

If your multipath tools use volsers for the unique ID, be sure to assign unique volsers to the devices that you want to configure as multipath devices.

Chapter 2. Setting up disk volumes for PAV

Setting up a disk volume for PAV includes these tasks:

- 1. Configuring the volume on the storage system. The volume must be configured as a base device with at least one alias device. There is no separate real disk space associated with alias devices.
 - Refer to your storage system documentation for details. For example, *IBM TotalStorage Enterprise Storage Server Web Interface User's Guide*, SC26-7448.
- 2. Defining the volume to the System z hardware. See "Enabling volumes for PAV."
- 3. Configuring paths to the volume on Linux. See "Configuring PAV base and alias volumes with multipath tools" on page 7.

Important

The procedure described in this section assumes that the volumes to be set up for PAV do not already hold data.

Enabling volumes for PAV

How you define disk volumes to your hardware depends on whether Linux runs as a z/VM[®] guest operating system or in LPAR mode.

The IOCDS examples in the following sections apply to mainframe systems with a single subchannel set. For information on IOCDS specifications for multiple subchannel sets refer to *System z9 Business Class and Enterprise Class and @server zSeries 890 and 990 Input/Output Configuration Program User's Guide for ICP IOCP*, SB10-7037.

Enabling volumes for Linux running as a z/VM guest operating system

This section applies to Linux running as a z/VM guest operating system. If your Linux instance runs in LPAR mode, skip this section and perform the steps in "Enabling volumes for Linux in LPAR mode" on page 6 instead.

Prerequisites:

- You need to know the device numbers of the base devices and their aliases as defined on the storage system.
- You need privilege class B authorization on z/VM.

Perform the following steps to define the base devices and their aliases to the hardware. In the examples, we assume that device 0x5680 is a base device and 0x56BF an alias device for the same physical disk space on the storage system.

1. Define the base devices to the hardware. In an IOCDS IODEVICE statement, use UNIT=3390B.

Example: The following statement defines device number 0x5680 as a base device.

```
IODEVICE ADDRESS=(5680), UNITADD=00, CUNUMBR=(5680), STADET=Y, UNIT=3390B
```

Define the alias devices to the hardware. In an IOCDS IODEVICE statement, use UNIT=3390A.

Example: The following statement defines device 0x56BF as an alias device. The mapping to the associated base device 0x5680 is in the storage system configuration.

```
IODEVICE ADDRESS=(56BF),UNITADD=18,CUNUMBR=(5680), **
STADET=Y,UNIT=3390A
```

3. After the hardware configuration with the base and alias device statements has become active, use z/VM to check the mapping of base and alias devices.

Example:

```
# CP QUERY PAV
00: Device 5680 is a base Parallel Access Volume with the following aliases: 56BF
00: Device 56BF is an alias Parallel Access Volume device whose base device is 5680
```

4. From z/VM, use CP ATTACH commands to make base devices and their aliases accessible to the Linux guest.

Example: To make a base device 0x5680 and its alias 0x56BF available to a guest with ID "LNX1" issue:

```
# ATTACH 5680 LNX1
# ATTACH 56BF LNX1
```

For more details on the z/VM commands refer to z/VM CP Commands and Utilities Reference, SC24-6081.

Continue with "Configuring PAV base and alias volumes with multipath tools" on page 7.

Enabling volumes for Linux in LPAR mode

This section applies to Linux running in LPAR mode. If your Linux instance runs as a z/VM guest operating system, omit this section and perform the steps in "Enabling volumes for Linux running as a z/VM guest operating system" on page 5 instead.

Prerequisite: You need to know the device numbers of the base devices and their aliases as defined on the storage system.

Perform the following steps to define the base devices and their aliases to the hardware. In the examples, we assume that device 0x5680 is a base device and 0x56BF an alias device for the same physical disk space on the storage system.

1. Define the base devices to the hardware. In an IOCDS IODEVICE statement, use UNIT=3390B.

Example: The following statement defines device number 0x5680 as a base device.

```
IODEVICE ADDRESS=(5680), UNITADD=00, CUNUMBR=(5680), *
STADET=Y, UNIT=3390B
```

2. Define the alias devices to the hardware. In an IOCDS IODEVICE statement, use UNIT=3390A.

Example: The following statement defines device 0x56BF as an alias device. The mapping to the associated base device 0x5680 is in the storage system configuration.

```
IODEVICE ADDRESS=(56BF),UNITADD=18,CUNUMBR=(5680),
STADET=Y,UNIT=3390A
```

3. Ensure that at least one base device is online on Linux.

Example: To ensure that the base device with device number 0x5680 is online, issue the following command from a Linux terminal session:

```
# chccwdev -e 0.0.5680
```

Alias devices might not be represented in sysfs until the base device has been set online.

4. **Optional:** Check that the alias device has the same value for the uid attribute as the base device to confirm that both map to the same physical disk.

Example:

```
# cat /sys/bus/ccw/drivers/dasd-eckd/0.0.5680/uid
IBM.75000000092461.2a00.1a
# cat /sys/bus/ccw/drivers/dasd-eckd/0.0.56BF/uid
IBM.75000000092461.2a00.1a
```

5. **Optional:** Confirm that one is a base device and the other is an alias device.

Example:

```
# cat /sys/bus/ccw/drivers/dasd-eckd/0.0.5680/alias
0
# cat /sys/bus/ccw/drivers/dasd-eckd/0.0.56BF/alias
1
```

0 indicates the base device and 1 indicates the alias device.

Configuring PAV base and alias volumes with multipath tools

This section describes how to define a PAV base device and its aliases as a single multipath volume.

Prerequisites:

- You must know the device numbers of the PAV base device and its aliases.
- · You need root authorization on the Linux system

Perform the following steps from a Linux terminal session:

- 1. Ensure that the devices are ready for use.
 - a. Issue **lsdasd** to ensure that device nodes exist for the PAV base volume and its aliases and that the devices are online.

Device nodes are usually created automatically (for example, by udev). If there are no device nodes, create them yourself.

For information on how to create DASD device nodes see the DASD chapter of *Linux on System z Device Drivers, Features, and Commands*. You can find the latest version on the developerWorks Web site at:

www.ibm.com/developerworks/linux/linux390/october2005 documentation.html

Use **chccwdev** to set the DASD online, if needed. For details refer to the **chccwdev** man page.

Example: This **Isdasd** output shows that both the base device, 0x5680, and the alias, 0x56bf, of our example are online and that the standard device names dasdc and dasdd have been assigned to them.

```
# lsdasd
0.0.5601(ECKD) at ( 94: 0) is dasda : active at blocksize: 4096, 1803060 blocks, 7043 MB
0.0.5602(ECKD) at ( 94: 4) is dasdb : active at blocksize: 4096, 1803060 blocks, 7043 MB
0.0.5680(ECKD) at ( 94: 8) is dasdc : active at blocksize: 4096, 1803060 blocks, 7043 MB
0.0.56bf(ECKD) at ( 94: 12) is dasdd : active at blocksize: 4096, 1803060 blocks, 7043 MB
```

b. 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
```

c. Ensure that the device is partitioned. If it is not already partitioned, use fdasd to create one or more partitions. Because a base device and its aliases all correspond to the same physical disk space, partitioning either the base device or one of its aliases creates partitions for the base device and all alias devices.

Example: The following command creates both a partition /dev/dasdc1 for the base device and also a partition /dev/dasdd1 for the alias.

```
# fdasd -a /dev/dasdc
```

d. 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 0.0.5680,0.0.56bf
# chccwdev -e 0.0.5680,0.0.56bf
```

You now have PAV enabled devices for which multiple subchannels are configured. You can display the subchannels for a particular PAV enabled device by issuing a command like this:

```
# lscss | egrep "<devno base device>|<devno alias1>|<devno alias2>| ..."
```

Example: For a base device 0x5680 and alias 0x56BF the command and its output might look like this:

```
# lscss | egrep "5680|56BF"
0.0.5680 0.0.0030 3390/0C 3990/E9 yes FF FF C6C7C8CA CBC90000
0.0.56BF 0.0.0031 3390/0C 3990/E9 yes FF FF C6C7C8CA CBC90000
```

In the example:

- The base device 0x5680 can be accessed through subchannel 0x0030.
- The alias device 0x56BF can be accessed through subchannel 0x0031.
- 2. If it is not already loaded, load the dm_multipath module:

```
# modprobe dm_multipath
```

3. Use the **multipath** command to detect multiple paths to devices for failover or performance reasons and coalesce them:

Example: This example shows the base device dasdc and its alias dasdd of the examples in the previous steps and a device dasde with three aliases: dasdf, dasdg, and dasdh.

IBM.75000000092461.2a00.1a and IBM.75000000092461.2a00.1b are unique IDs obtained from the dasdinfo tool.

- 4. Ensure that the multipathd daemon is running:
 - a. Enter ps -ef | grep multipath to check if the daemon is already running.
 - b. If the command output does not list multipathd as a running process, enter the following command to start it:

```
# /etc/init.d/multipathd start
```

5. Enter multipath -11 to display the resulting multipath configuration.

Example:

Result: 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.

Example:

```
# 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, 2 Oct 19 17:10 /dev/mapper/IBM.75000000092461.2a00.1ap1
brw-rw---- 1 root disk 253, 1 Oct 19 17:02 /dev/mapper/IBM.75000000092461.2a00.1b
brw-rw---- 1 root disk 253, 3 Oct 19 17:10 /dev/mapper/IBM.75000000092461.2a00.1bp1
crw-rw---- 1 root root 10, 61 Oct 19 16:58 /dev/mapper/control
```

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

Note: 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.

Chapter 3. Partitioning devices that have been set up for multipathing

Do not partition a device while it is configured for multipathing.

Perform the following steps to partition a device that has already been set up for multipathing:

1. Remove the multipath device. Issue a command of this form: multipath -f <multipath-device>

where *<multipath-device>* is the device name that the multipath tool has generated for the multipath device.

Example:

multipath -f IBM.7500000092461.2a00.1a

- Create up to three partitions using fdasd. Because a base device and its aliases all correspond to the same physical disk space, partitioning either the base device or one of its aliases creates partitions for the base device and all alias devices.
- 3. 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 0.0.5680,0.0.56bf
# chccwdev -e 0.0.5680,0.0.56bf
```

4. Restore the multipathing definitions by issuing:

multipath

Result: The following command output is based on the examples of the previous sections and assumes that three partitions have been created on dasdc:

```
# ls -l /dev/mapper/*
brw-rw---- 1 root disk 253, 0 Oct 19 42:09 /dev/mapper/IBM.75000000092461.2a00.1a
brw-rw---- 1 root disk 253, 2 Oct 19 42:17 /dev/mapper/IBM.75000000092461.2a00.1ap1
brw-rw---- 1 root disk 253, 3 Oct 19 42:17 /dev/mapper/IBM.75000000092461.2a00.1ap2
brw-rw---- 1 root disk 253, 4 Oct 19 42:17 /dev/mapper/IBM.75000000092461.2a00.1ap3
brw-rw---- 1 root disk 253, 1 Oct 19 42:09 /dev/mapper/IBM.75000000092461.2a00.1b
brw-rw---- 1 root disk 253, 5 Oct 19 42:17 /dev/mapper/IBM.75000000092461.2a00.1bp1
crw-rw---- 1 root root 10, 61 Oct 19 42:05 /dev/mapper/control
```

Notices

This information was developed for products and services offered in the U.S.A. IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing IBM Corporation North Castle Drive Armonk, NY 10504-1785 U.S.A.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

IBM World Trade Asia Corporation Licensing 2-31 Roppongi 3-chome, Minato-ku Tokyo 106-0032, Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This 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.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

The licensed program described in this information and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement, or any equivalent agreement between us.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information is for planning purposes only. The information herein is subject to change before the products described become available.

Trademarks

The following terms are trademarks of International Business Machines Corporation in the United States, other countries, or both:

developerWorks Enterprise Storage Server **IBM** S/390 System z System z9 TotalStorage z/VM zSeries

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, and service names may be trademarks or service marks of others.

Readers' Comments — We'd Like to Hear from You

Linux on System z How to Improve Performance with PAV November 17, 2006 Linux Kernel 2.6 (October 2005 stream)

Publication No. SC33-8292-01

We appreciate your comments about this publication. Please comment on specific errors or omissions, accuracy, organization, subject matter, or completeness of this book. The comments you send should pertain to only the information in this manual or product and the way in which the information is presented.

For technical questions and information about products and prices, please contact your IBM branch office, your IBM business partner, or your authorized remarketer.

When you send comments to IBM, you grant IBM a nonexclusive right to use or distribute your comments in any way it believes appropriate without incurring any obligation to you. IBM or any other organizations will only use the personal information that you supply to contact you about the issues that you state on this form.

Comments:

Thank you for your support.

Submit your comments using one of these channels:

Send your comments to the address on the reverse side of this form.

If you would like a response from IBM, please fill in the following information:

Name	Address
Company or Organization	
Phone No.	E-mail address

Readers' Comments — We'd Like to Hear from You SC33-8292-01



Cut or Fold Along Line

Fold and Tape Please do not staple

Fold and Tape

PLACE POSTAGE STAMP HERE

IBM Deutschland Entwicklung GmbH Information Development Department 3248 Schoenaicher Strasse 220 71032 Boeblingen Germany

Fold and Tape

Please do not staple

Fold and Tape

SC33-8292-01

