Linux on System z

How to Improve Performance with PAV May, 2008

Linux Kernel 2.6 - Development stream

SC33-8414-00

Linux on System z

How to Improve Performance with PAV May, 2008

Linux Kernel 2.6 - Development stream

SC33-8414-00

Note

Before using this information and the product it supports, read the information in “Notices” on page 7.

First Edition (May, 2008)

This edition applies to the Linux on System z Development stream and to all subsequent releases and modifications until otherwise indicated in new editions.

SC33-8414 is the Linux on System z Development stream equivalent to SC33-8292, which applies to the Linux on System z October 2005 stream.

Contents

Summary of changes

This edition reflects changes for the May 7th 2008 software drop. This book is the equivalent to SC33-8292, which applies to Linux® on System z™, October 2005 stream.

Changes compared to SC33-8292 are as follows:

- The DASD device driver now handles aliases for block devices for you. A multipath setup is no longer required for PAV.
- The DASD device driver supports PAV and HyperPAV.

This revision also includes maintenance and editorial changes.

About this document

This document describes how to set up DASD volumes as parallel access volumes (PAV) in z/VM® or LPAR and how to use PAV from Linux.

In this book, System z is taken to include System z10™, System z9®, zSeries® in 64- and 31-bit mode, as well as S/390® in 31-bit mode.

You can find the latest version of this document on developerWorks® at ibm.com/developerworks/linux/linux390/development_documentation.html.

Where to get more information

This section points to more information about the DASD device driver and the storage systems that support PAV

For more information about the DASD device driver see *Device Drivers, Features, and Commands*. You can obtain the latest version of this book on developerWorks at ibm.com/developerworks/linux/linux390/development_documentation.html.

For information about PAV, HyperPAV, and IBM System Storage DS8000™ series (DS8000) see the DS0800 information in the IBM® Systems Information Center at http://publib.boulder.ibm.com/infocenter/systems/topic/com.ibm.storage.ssic.help.doc/f2c_ds8000sicparent_3io3pr.html.

For information about PAV and IBM System Storage DS6000™ series (DS6000) see the DS6000 Information Center at http://publib.boulder.ibm.com/infocenter/dsichelp/ds6000ic/index.jsp.

For information about PAV and IBM TotalStorage® Enterprise Storage Server® see the *IBM TotalStorage Enterprise Storage Server Web Interface User's Guide*, SC26-7448.

For information about z/VM, see the documentation for you z/VM version at ibm.com/systems/z/os/zos/bkserv/zvmpdf/.

Chapter 1. Introduction to PAV

Linux on System z can use the parallel access volume (PAV) feature of enterprise storage systems to perform multiple concurrent data transfer operations to or from the same DASD volume.

The storage control units of present-day IBM enterprise storage systems can use large caches and RAID 5 arrays to perform I/O operations in parallel. Multiple users or processes on a Linux instance can concurrently issue channel programs to volumes that are configured as PAV.

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 are serialized to maintain data integrity.

Base and alias devices

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.

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

When the devices are set online, the DASD device driver can distinguish between base devices and aliases. The DASD device driver then creates device nodes for the base devices but not for the aliases. The aliases can lead to gaps in the naming scheme for device nodes. For example, if dasda and dasdd are base devices and dasdb and dasdc the names for alias devices, there will be device nodes /dev/dasda, and /dev/dasdd but no nodes /dev/dasdb and /dev/dasdc. User space processes exclusively access PAV though the device node for the base device.

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.

HyperPAV

With IBM HyperPAV, aliases are not exclusively used for the base device for which they are defined. An alias can be used for any base device within the same logical subsystem on the storage system. When the DASD device driver has to satisfy an I/O request through an alias, it associates an eligible alias with the respective base device. Apart from assuring that the aliases for the logical storage subsystem to which a base device belongs are online, user space processes need no special handling for accessing a volume configured for HyperPAV.

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.

Chapter 2. Making PAV available to Linux

PAV base and alias volumes require special IOCDS specifications. This section provides IOCDS sample specifications and describes additional steps you must perform if your Linux instance runs as a z/VM guest operating system.

Prerequisites:

- 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.

Configuring base and alias volumes for PAV or HyperPAV on the storage system is beyond the scope of this description. See your storage system documentation for details.

The IOCDS examples in the following sections apply to mainframe systems with a single subchannel set. For information about IOCDS specifications for multiple subchannel sets see the Input/Output Configuration Program User's Guide for your mainframe 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 by using UNIT=3390B in the IOCDS IODEVICE statements.

   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 by using UNIT=3390A in the IOCDS IODEVICE statements.

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

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

3. Optional for z/VM: 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 comand.

   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. Required for z/VM: If your Linux system runs as a z/VM guest operating system, use CP ATTACH commands to make the base devices and their aliases accessible to Linux.

   Example: To make a base device `0x5680` and its alias `0x56BF` available to a z/VM guest operating system with ID `LNX1` enter the following commands:

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

Chapter 3. Using PAV on Linux

You work with PAV or HyperPAV base devices as you would without PAV. To take advantage of PAV, be sure that the corresponding aliases are set online.

Prerequisites:

- If your Linux instance runs natively in an LPAR, the `nopav` keyword must not have been set for the `dasd=` kernel or module parameter.
- 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 start and confirm your PAV environment. In the examples, we assume that device `0x5680` is a base device and `0x56BF` a corresponding alias device.

1. Set your base devices online.

   Example: To set a base device with device number `0x5680` is online, enter the following command in 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.

   Note: If your Linux system runs as a z/VM guest operating, each device has a sysfs attribute `use_diag` that by default is set to `0`. Do not change this attribute to `1` for any of the aliases.
2. Set the alias devices online.
3. Optional: Confirm the mapping of base and alias devices by reading the `uid` attributes from sysfs.

   The uid has several dot-separated sections. The first three sections identify the logical subsystem of the storage system. The forth section identifies a particular volume within the logical subsystem.

   If your Linux instance runs as a z/VM guest operating system, there might be an additional section that identifies individual minidisks on the volume. The presence of this additional section depends on your z/VM version and service level.

   In a basic PAV environment, a base device and all its aliases have matching `uid` attributes.

   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
   ```

   In the example, the `uid` attributes are both the same which confirms that both devices map to the same physical disk space.

   In a HyperPAV environment, alias devices are not dedicated to a particular base device but can be used for any base device in the same logical subsystem on the storage system. Instead of a device identifier, alias devices have `xx` as the

forth section of their `uid` attribute. An alias is eligible for a base device if the first three sections of its `uid` attribute match the first three sections of the `uid` attribute of the base device.

HyperPAV example:

```
# cat /sys/bus/ccw/drivers/dasd-eckd/0.0.5680/uid
IBM.75000000092461.2a00.1a
# cat /sys/bus/ccw/drivers/dasd-eckd/0.0.5681/uid
IBM.75000000092461.2a00.1b
# cat /sys/bus/ccw/drivers/dasd-eckd/0.0.56BF/uid
IBM.75000000092461.2a00.xx
```

In the example, `0.0.56BF` is an alias that is eligible for base devices `0.0.5680`, and `0.0.5681`.

4. Optional: Confirm which devices are base devices and which devices are alias devices. You can find out if a device is a base device or an alias device by reading its `alias` attribute in sysfs.

   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.

You are now ready to work with the base devices as you would without PAV. The DASD device driver automatically uses the aliases as the need arises.

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

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at http://www.ibm.com/legal/copytrade.shtml.

Linux is a registered 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
May, 2008
Linux Kernel 2.6 - Development stream

Publication No. SC33-8414-00

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.
- Send your comments via e-mail to: eservdoc@de.ibm.com

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
SC33-8414-00

IBM®

Cut or Fold
Along Line

Fold and Tape | **Please do not staple** | Fold and Tape

NO POSTAGE
NECESSARY
IF MAILED IN THE
UNITED STATES

BUSINESS REPLY MAIL

FIRST-CLASS MAIL PERMIT NO. 40 ARMONK, NEW YORK

POSTAGE WILL BE PAID BY ADDRESSEE

International Business Machines Corporation
Schoenaicher Strasse 220
71032 Boeblingen
Germany

Fold and Tape | **Please do not staple** | Fold and Tape

SC33-8414-00

Cut or Fold
Along Line

Printed in USA

SC33-8414-00