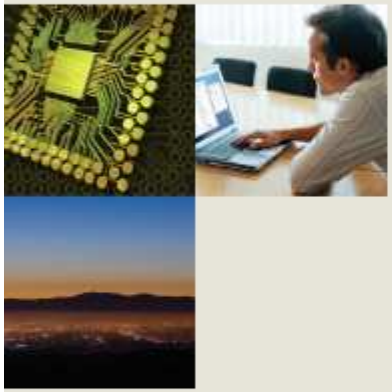


Linux on System z

What's new in the I/O Area

Session 9280

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

San Jose, August 14th 2008

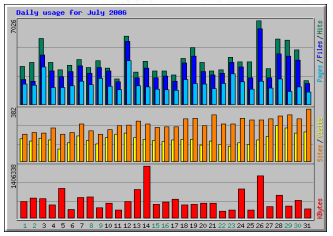
Agenda

- New I/O Features in
  - 4Q2006 code drop
  - 1Q2007 code drop
  - 4Q2007 code drop
  - 2Q2008 code drop
- Distributor support (SLES / RHEL)

  RedHat / SUSE support matrix
- Outlook on future I/O development

IBM System z10

Channel Path Measurement Data *(4Q2006)*

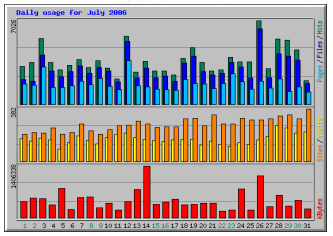


- Collect extended LPAR channel path measurement data from channel subsystem
  - Channel measurement characteristics
    as obtained by the CHSC Store Channel-Measurement Characteristics
  - Channel measurements
    as collected by the channel subsystem and written to the memory area specified by the CHSC Set Extended-Channel Measurements
- Make this data available to user space through sysfs
  - **`/sys/devices/css0/cm_enable`**

    controls enabling/disabling the extended channel path measurement facility

    It can take two values
    - *0: Deactivate facility and remove measurement-related attributes*
    - *1: Activate facility and create measurement-related attributes*

Channel Path Measurement Data (cont.)

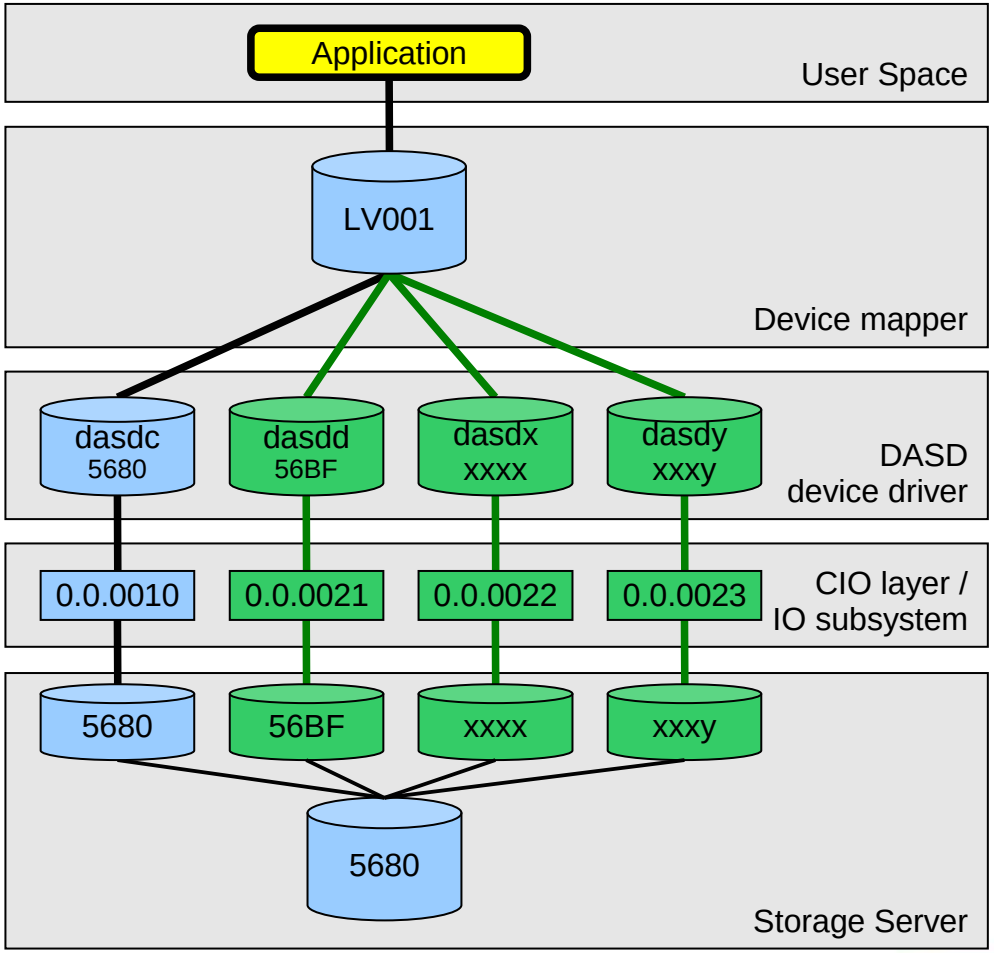


- Attributes for each channel path object
  - **`cmg`**
    Specifies the channel measurement group
  - **`shared`**
    Specifies whether the channel path is shared between LPARs
- Attributes added for active measurements
  - **`measurement`**
    Binary, containing the extended channel measurement data
    Consists of eight 32 Bit Channel-Utilization Entries
  - **`measurement_chars`**
    Channel measurement group dependent characteristics
    Consists of five 32 Bit CMG-Dependent Channel-Measurement Characteristics

DASD PAV support for LPAR (static PAV) *(4Q2006)*

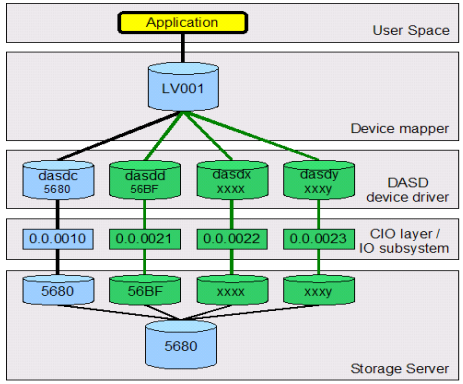
Structure

- One base path
  from application (via device mapper, DASD, CIO,..) to physical device
- Additional optional alias path
  allows simultaneous I/O to logical device using additional subchannel
- Alias paths must be managed by device-mapper doing:
  - Device mapping
  - Workload balancing
  - 'Path-failover'



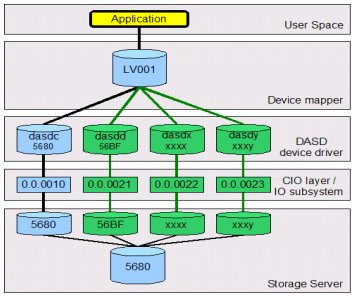
DASD PAV support for LPAR (cont.)

- Support for IBM Parallel Access Volumes (PAV) feature of IBM DASD subsystem
- Simultaneously process multiple I/O operation to single volume
- Significant performance improvement
- Can be deactivated by DASD-parameter 'nopav'
- Introduce new sysfs attributes:
  - **'uid':** unique-id (vendor.serial.SSID.UA) of the physical (base) device
  - **'vendor':** vendor/manufacturer
  - **'alias':** 0 for base device, 1 for alias device
- dasdinfo tool to support device-mapper setup
- No DASD driver internal synchronization done



DASD PAV support for LPAR Configuration

- Storage Server configuration
  Please refer to *storage system documentation*



- IOCDS

```
IODEVICE ADDRESS=(5680),UNITADD=00,CUNUMBR=(5680), *
         STADET=Y,UNIT=3390B
IODEVICE ADDRESS=(56BF),UNITADD=18,CUNUMBR=(5680), *
         STADET=Y,UNIT=3390A
```

- 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 PAV support for LPAR Configuration (cont.)

Device-mapper configuration

- Load dm_multipath module (if not already available)
  ```
  # modprobe dm_multipath
  ```
- Check device availability (optional)
  ```
  # 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
  ```
- Use multipath command to automatically detect paths to device

  ```
  # multipath
  create: 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 dasdc 94:8  [undef][ready]
   \_ 0:0:10927:0 dasdd 94:12 [undef][ready]
  ```

- Access to multipath device

  device nodes for the multipath device are available at '**/dev/mapper**'
  ```
  # 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
  ```

DASD PAV support for LPAR Pitfalls

- Make sure the device is formatted and partitioned prior to multipath-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

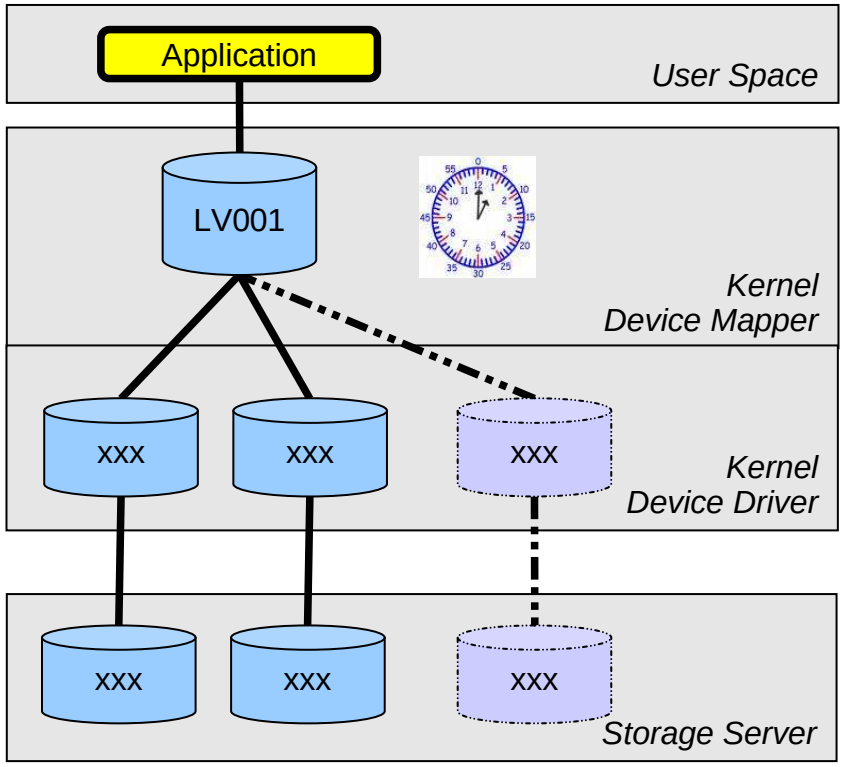
Disk mirroring real time enhancements *(4Q2006)*

Enhanced real time capabilities for disk mirrors

- Mirror fault tolerance / Out of sync handling for mirror path
- User defined response time for logical volume

Issues

- Higher memory / CPU consumption (memcpy)
- No upstream / distro solution yet
  (special customer requirement)



Disk mirroring real time enhancements - Tools

- Adapt user space tools (LVM2) to provide
  - Additional parameter for configuration (e.g. timeout)
  - Tolerance for stalled disks
  - Operation with missing disks
  - Enhanced real time capabilities for disk mirrors
- New perl script (statistics.pl) to extract statistical information like
  - Missed events
  - Recovery duration / distance
  - Degradation duration

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

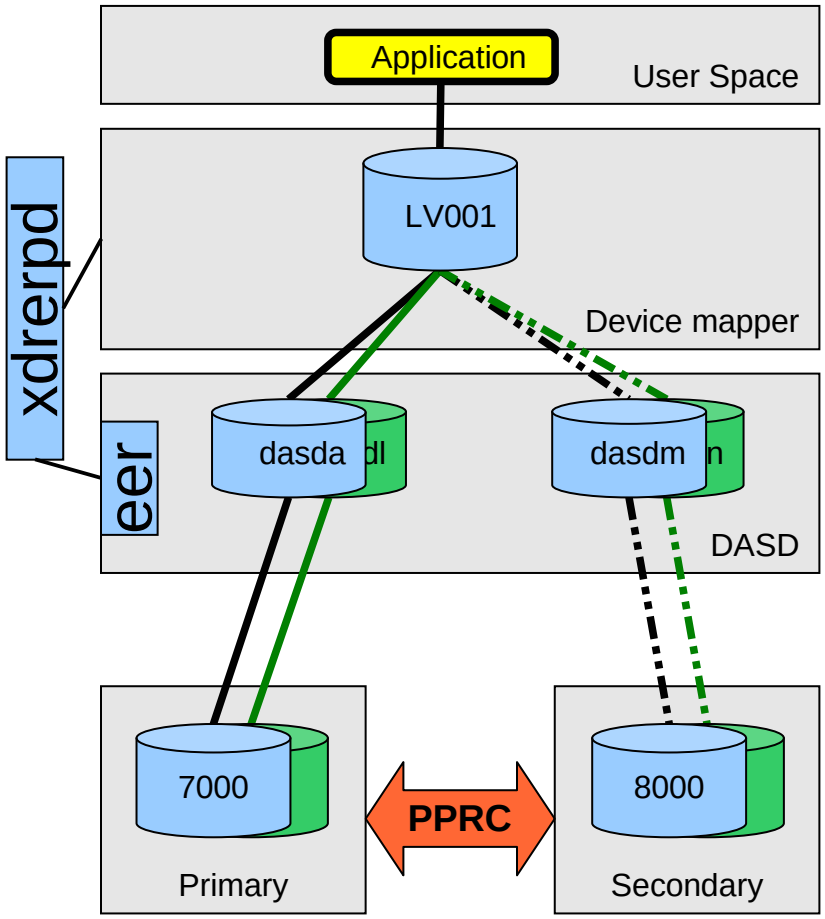
http://www-03.ibm.com/systems/services/labservices/platforms/labservices_z.html

HyperSwap Support in DASD and CIO *(4Q2006)*

- 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

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



Other I/O features in 4Q2006 code drop

- Deprecate DASD FBA driver
  Document that native FBA access is no longer recommended – use DIAG instead
- 3592 CU recognition
  Enable access to 3592 tape device in 3590 mode
- Upstream 3590 Tape Device Driver
  Release driver under GPL license
- Improved handling of FCP adapter failures
  Introduce unique request ID (do not reuse ID)

Improved handling of dynamic subchannel mapping (1Q2007)

- Enable CIO to handle detached devices re-appearing on different subchannel
  - Move ccw device in common driver core
    Provide '**`device_move`**' that moves device to different parent
  - Make use of 'device_move' in CIO if
    - Disconnected device appears on another subchannel
    - Another ccw device appears on already disconnected subchannel (disconnected device is moved to pseudo subchannel)
    - A disconnected device under the pseudo subchannel appears again
  - Device view in sysfs may change
    `/sys/devices/css0/<sch>/<ccw-device>` *connected device*
    `/sys/devices/css0/defunct/<ccw-device>` *for pseudo subchannel*
  - User space needs to handle **`KOBJ_MOVE`** uevents

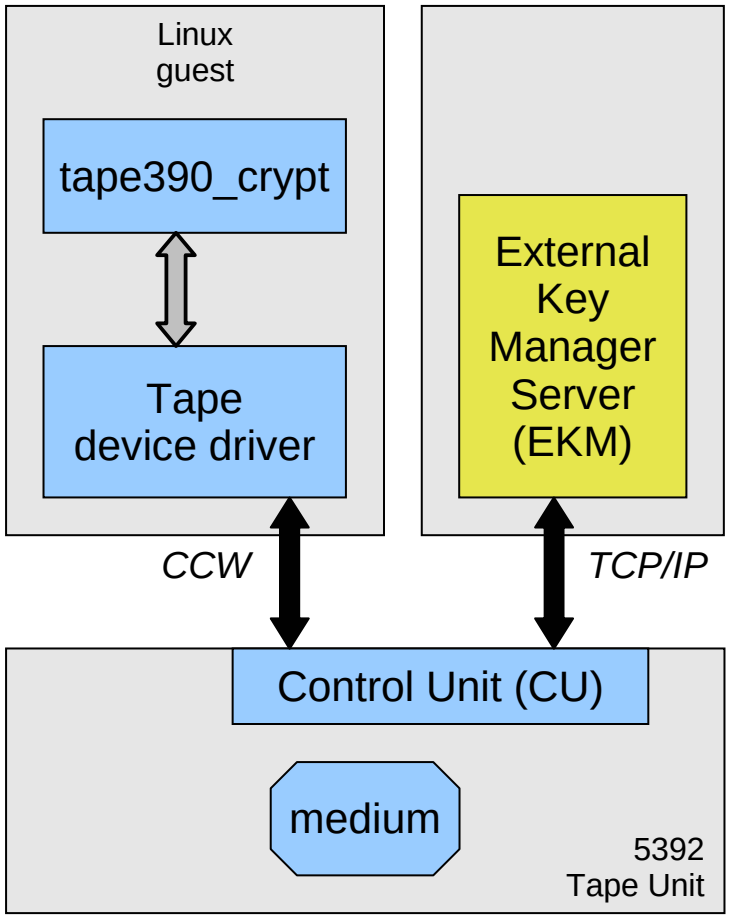
3592 tape encryption support (1Q2007)

- Encryption support for channel attached 3592 tape devices
- Data encrypted on medium using Data Key
  - Data key also stored on medium (max 2) in External Encrypted Data Keys (EEDKs) field
- Key Encrypting Key (KEK)
  - Addressed by operating system (hash or label)
- New tool 'tape390_crypt' controlling encryption feature
- Encryption support can be activated / deactivated

3592 tape unit (TS1120)

3592 tape encryption support Overview

- External key Manager Server (EKM)
  - Store encryption keys (KEK)
  - Communicate with tape control unit ('out of band' control unit based encryption)
  - Create External Encrypted Data Key (EEDK) based on Key Encrypting Key (KEK)
  - Running on any machine with Java and TCP/IP support



3592 tape encryption support tape390_crypt

- Enable / Disable encryption

```
# tape390_crypt -e on /dev/ntibm0
```

- Specify encryption key (KEK)

```
# tape390_crypt -k my_first_key:label -k my_second_key:hash /dev/ntibm0
--->> ATTENTION! <<---
All data on tape /dev/ntibm0 will be lost.
Type "yes" to continue: yes
SUCCESS: key information set.
```

- Query encryption status

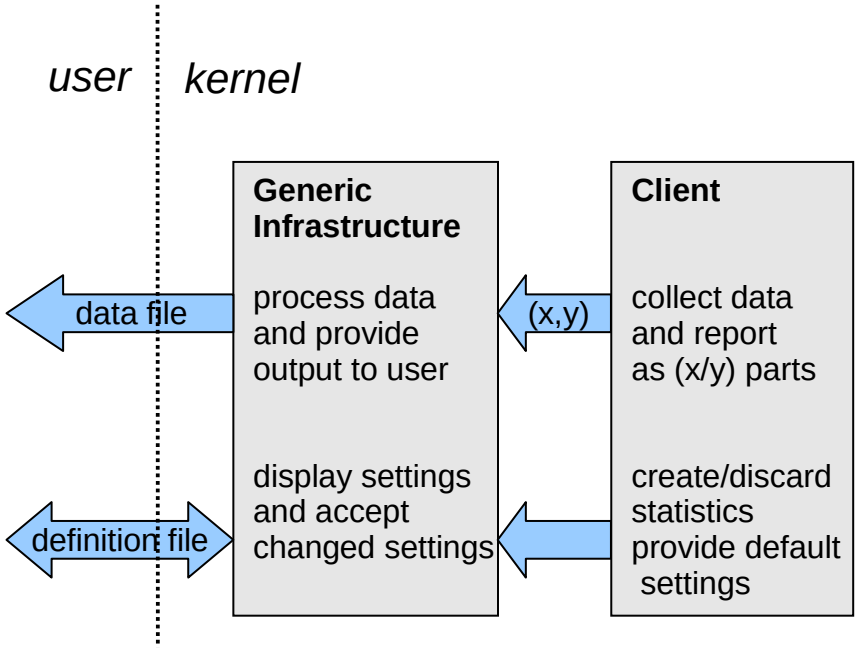
```
# tape390_crypt -q /dev/ntibm0
ENCRYPTION: ON
MEDIUM: ENCRYPTED
KEY1:
 value: my_first_key
 type: label
 ontape: label
KEY2:
 value: my_second_key
 type: label
 ontape: hash
```

FCP measurement data I/O Statistics (1Q2007)

- Generic Infrastructure
  - Data output
    `…/statistics/<scsi-lun>/data`
  - Definition file
    `…/statistics/<scsi-lun>/definition`
- Client

  SCSI collected data including
  - Request latency (read, write, nodata)
  - Request size (read, write, nodata)
  - Result
  - Utilization (queue_used_depth)
- NOT accepted upstream

  Needs rework



Other I/O features in 1Q2007 code drop

- DASD runtime switch for logging

  Activate and de-activate ERP-related logging for a running system using **'dasd='** parameter or sysfs attribute **'erplog'**

- No XML in System Dumper

  Get rid of no longer supported XML formated data in system dumper (zfcpdump in s390-tools), use binary block instead

Dynamic CHPID reconfig via SCLP (4Q2007)

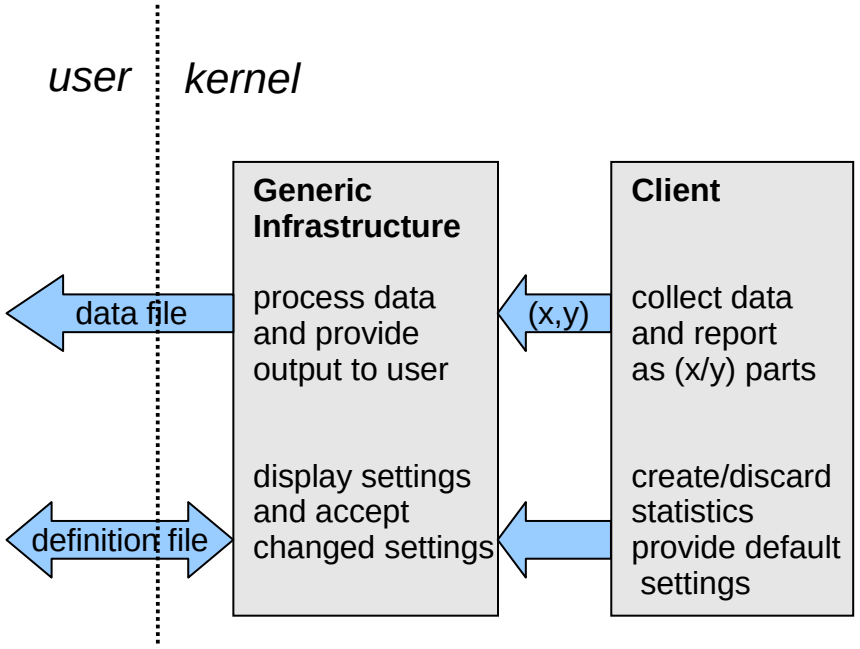
- Change (chchp) configuration state of an I/O subchannel
  - Available state
    0: the channel-path is in standby state
    1: the channel-path is in configured state
    2: the channel-path is reserved
    3: the channel-path is not recognized
  - Configure device (offline/online)
    `# chchp --configure 1 0.40`
  - Logical vary on/off
    `# chchp --vary 1 0.40`

- Query configuration state

```
# lschp
CHPID Vary Cfg. Type Cmg Shared
==================================
0.40  1    1    1b   2   1
```

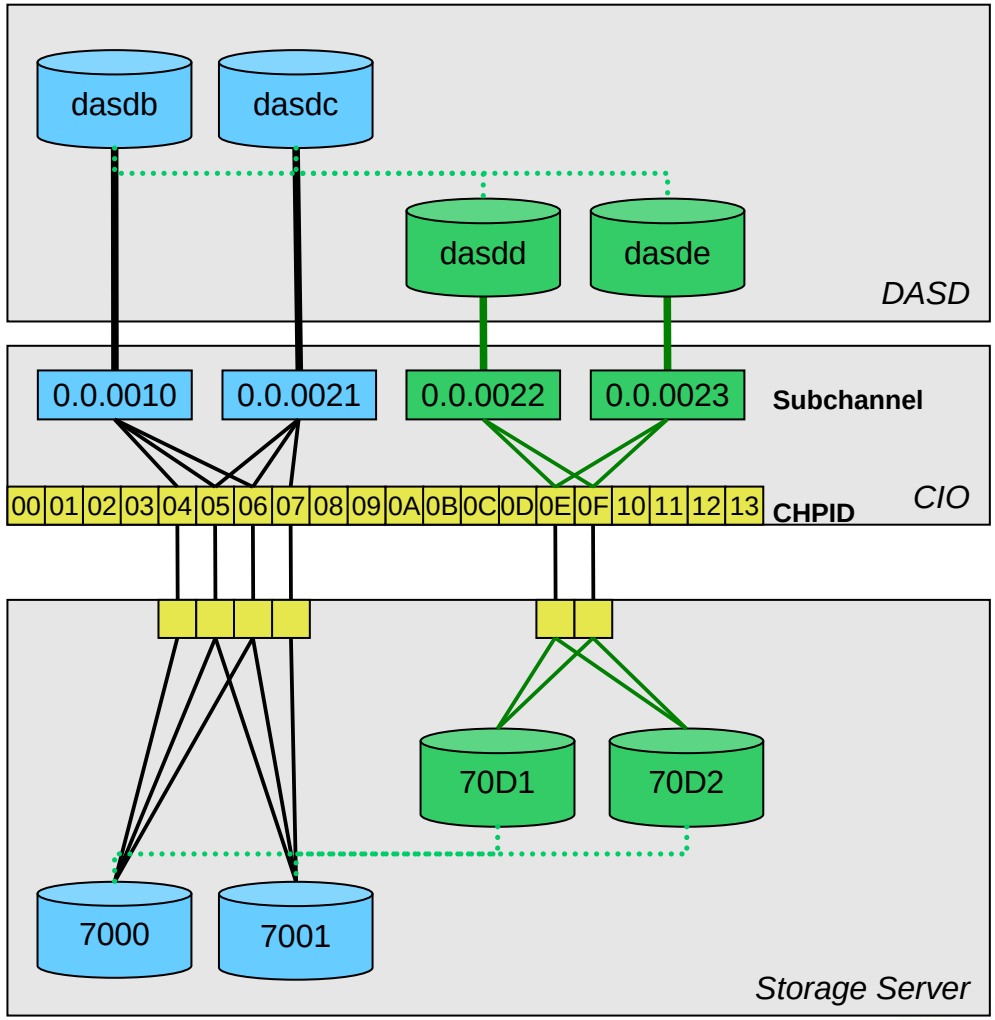
FCP measurement data
Adapter statistics (4Q2007)

- Enancement to the FCP measurement data item (1Q2007)
- Using generic FCP measurement infrastructure
- Collecting adatper statistis
  - FCP subchannel (virtual HBA)
    - Number of input, output and control requests
    - Number of bytes sent and received;
    - Seconds since activation.
  - FCP channel (physical HBA)
    - Processor, bus and adapter utilization
- NOT accepted upstream



DASD HyperPAV enablement (2Q2008)

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



DASD HyperPAV enablement 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

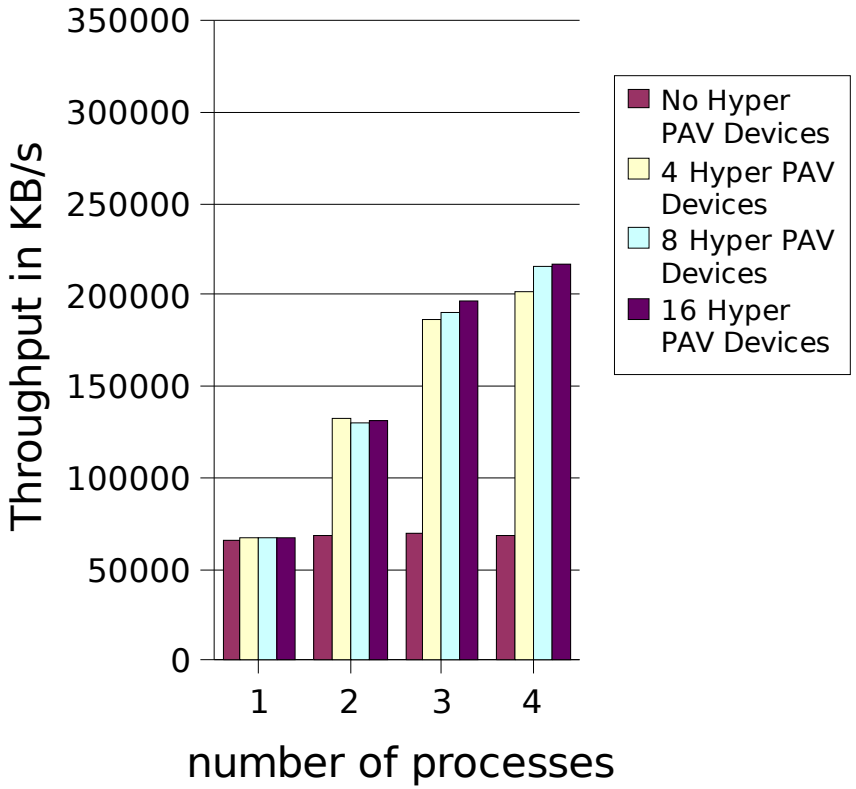
HyperPAV simplifies systems management and improves performance using an on demand I/O model

DASD HyperPAV - Performance

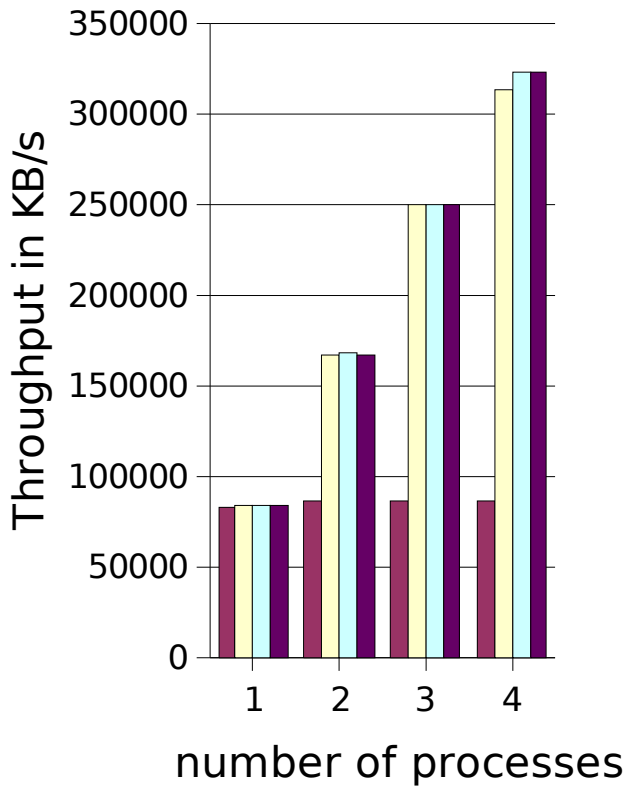
Single Disk Test – Sequential DIO - 700MB file size - 256MB Memory

SHARE
Technology · Connections · Results

Throughput for initial writers



Throughput for readers



Other I/O features in 2Q2008 code drop

- SIM Handling for ECKD DASD devices

  Enable System Information Messages (SIM) for DASD devices

- Multipath IPL / IPL trough IFCC

  Continue IPL on alternate path in case of an Interface Control Check (IFCC)

- FCP performance statistics
  - ‚blktrace‘

    Join common block trace (blktrace) approach
  - Architecture specific

    zFCP specific statistics

Feature-Matrix
for **4Q2006** code drop

Feature 4Q2006	RHEL		SLES	
	4	5	9	10
Channel Path Measurement Data	--	GA	--	SP1
DASD PAV support for LPAR	U5	GA	--	SP1
Disk mirroring real time enhancements	--	--	--	--
HyperSwap support in DASD and CIO	--	GA	--	GA
Deprecate DASD FBA driver	n/a	n/a	n/a	n/a
3295 CU recognition	U6	GA	SP4	SP1
Upstream 3590 Tape device driver	U5	GA	SP3U1	GA
Improved handling of FCP adapter failures	--	GA	--	GA

Feature-Matrix
for **1Q2007 / 4Q2007** code drop

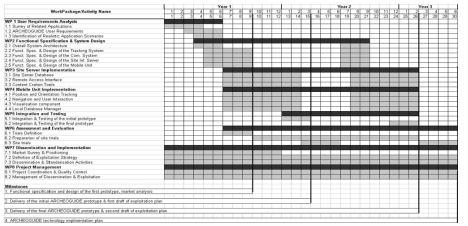
Feature 1Q2007	RHEL		SLES	
	4	5	9	10
Improved handling of dynamic subchannel mapping	--	--	--	--
3592 tape encryption support	U6	U1	SP4	SP1
FCP measurement data – I/O statistics	--	--	SP3U1	SP1
DASD runtime switch for logging	U6	U1	SP3U1	GA
No XML in System Dumper	--	--	SP4	SP1
Feature 4Q2007				
Dynamic CHPID reconfiguration via SCLP	--	U2	--	SP2
FCP measurement data – Adapter statistics	--	--	SP4	SP2

Feature-Matrix
for **2Q2008** code drop

Feature 2Q2008	RHEL		SLES	
	4	5	9	10
DASD HyperPAV enablement	--	--	--	--
SIM Handling for ECKD DASD devices	--	--	--	--
Multi Path IPL / IPL trough IFCC	--	--	--	--
FCP perfomance statistics – blktrace	--	--	--	--
FCP perfomance statistics – achritecture specific	--	--	--	--

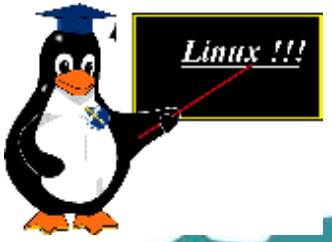
Outlook
(subject to change)

- Support for new Storage features
  - Support for Extended Address Volumes
  - ...
- Enhanced configuration support
  - Automatic discovery (Port/LUN)
  - Configuration simplification
  - Enhanced functionality
- Performance improvements

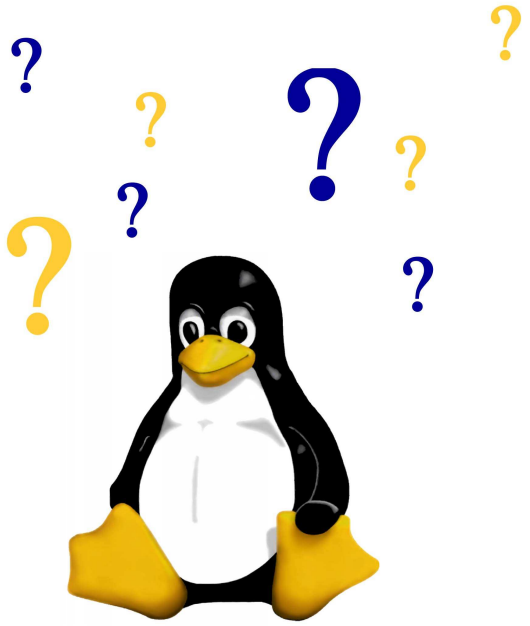


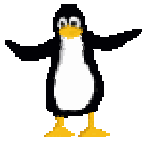
Useful links

- 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

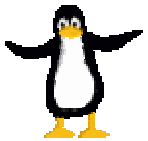


Questions





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.