SHARE
Technology · Connections · Results

Linux on System z
What's new in the I/O Area

Horst Hummel
IBM

Horst.Hummel@de.ibm.com

February 28th 2007
Session 9280

© Disney

Agenda

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

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

IBM System z9 Enterprise Class

DASD DIAG250 64 bit support
(1Q2006)

- DIAGNOSE 'X'250 (DIAG250) enabled for 64bit kernel
- Requires z/VM 5.2 (or above) DIAG250 interface that supports
  - 64 bit addressing mode as well as
  - 64 bit block numbers on 64 bit guests
- Now DIAG access method available for
  - ECKD, FBA and SCSI emulated FBA (EDEV)
  - Any fixed block sized format
    (not only CMS reserved)

IBM System Storage DS8000

FAILFAST support in DASD device driver (1Q2006)

- Support the request flag 'REQ_FASTFAIL' (return failed requests)
- Implemented the following (DASD specific) way
  - No FAILFAST if Extended Error Reporting (EER) is enabled
  - Process ERP first (including long busy)
  - Return request only if not able to process (no operational device available) and FAILFAST flag is set – otherwise queue the request

FC transport class exploitation (1Q2006)

- Exploit complete fibre channel (FC) transport class functionality
  - ***Provide transport specific attributes for SCSI devices***
  - ***Attributes can be found in sysfs***

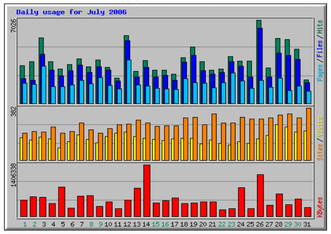
```
/sys/class/fc_host
/sys/class/fc_transport
/sys/class/fc_remote ports
/sys/class/fc_host/hostX                      where X   is the SCSI host number.
/sys/class/fc_host/hostX/statistics  where X   is the SCSI host number
```

  - Also replace existing 'zfcp Adapter/Port Attributes' by 'Transport Class Attributes' and add new attributes

Other I/O features in 1Q2006 code drop

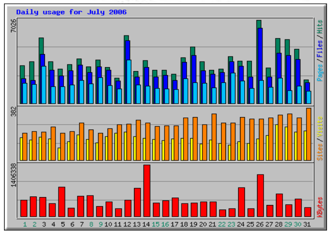
- Multiple Subchannel Set (MSS) support
  Detect and use subchannels and devices in subchannel sets with ID > 0
- Generic Attribute Support in chccwdev
  adding additional '`--attribute`' option for generic attribute modification
- Improved zfcp Traces - Additional lcrash View
  lcrash plugin for zfcp traces
- Improved SAN Notifications
  log 'Unsolicited status notification' to /var/log/messages
- zfcp/SCSI Scripts: `scsi logging level`
  shell script (s390-tools) to create, set, or get the SCSI logging level
- zfcp/SCSI Scripts: `lszfcp`
  shell script (s390-tools) to display information about zfcp adapters, ports and units and their associated class devices

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

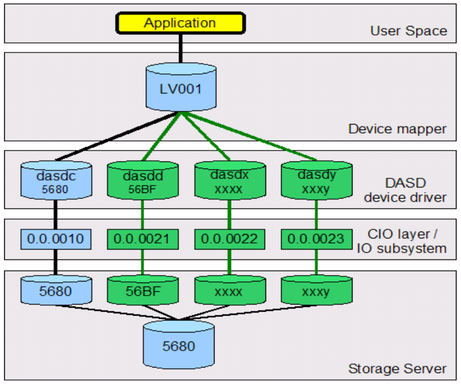
- Attributes added for active measurements
  - **`measurement`**
    Binary, containing the extended channel measurement data
    Consists of eight 32 Bit Channel-Utilization Entries
  - **`measurement_chars`**
    Channel measurement group dependent characteristics
    Consists of five 32 Bit CMG-Dependent Channel-Measurement Characteristics

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

- 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:

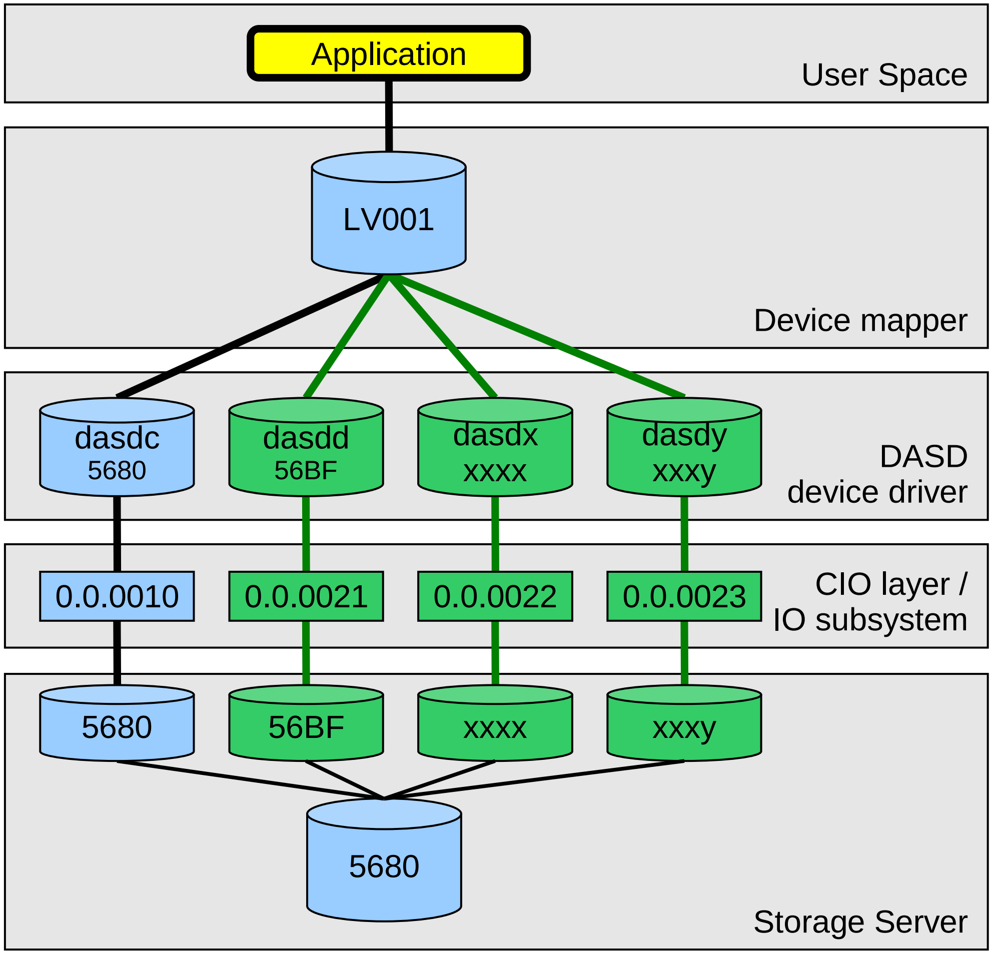
'uid':	**unique-id (vendor.serial.SSID.UA) of the physical (base) device**
'vendor':	**vendor/manufacturer**
'alias':	**0 for base device, 1 for alias device**

- dasdinfo tool to support device-mapper setup
- No DASD internal synchronization done



DASD PAV support for LPAR (cont.)

- 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'

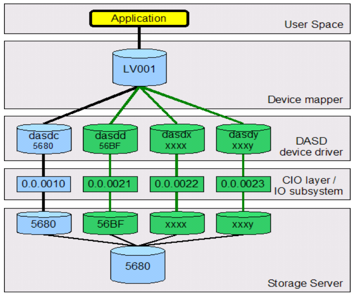


Horst Hummel, IBM STG-LTC Boeblingen

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`'

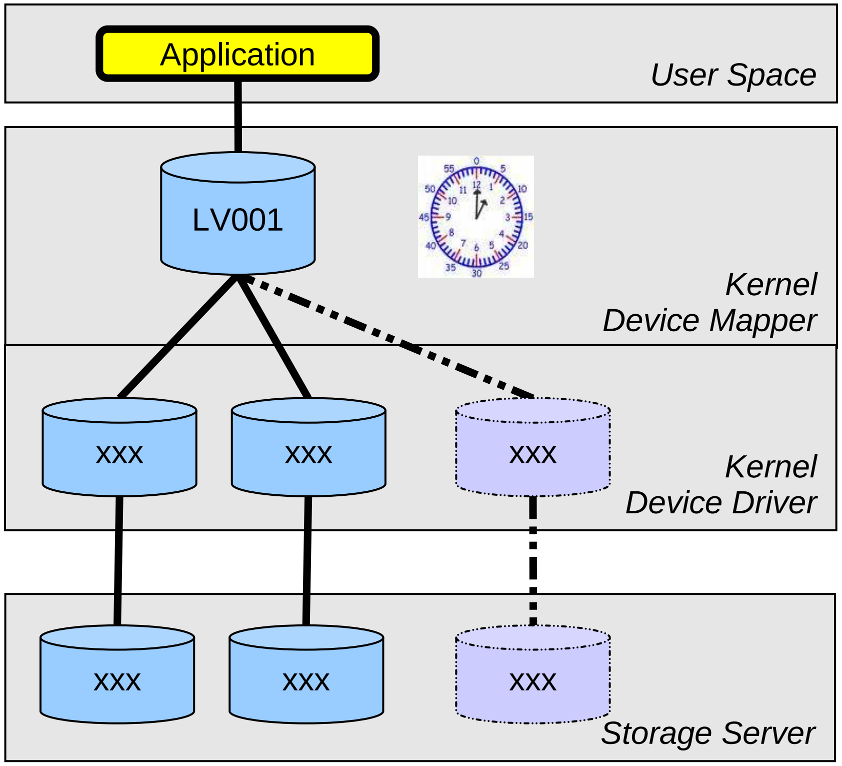
    ```
    # ls -l /dev/mapper/*
    brw-rw---- 1 root disk 253, 0 Oct 19 17:02 /dev/mapper/IBM.75000000092461.2a00.1a
    brw-rw---- 1 root disk 253, 1 Oct 19 17:10 /dev/mapper/IBM.75000000092461.2a00.1ap1
    ```

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
- User defined response time for logical volume
- Higher memory / CPU consumption (memcpy)
- Out of sync handling for mirror path
- No upstream / distro solution yet
  (special customer requirement)



Disk mirroring real time enhancements - Tools

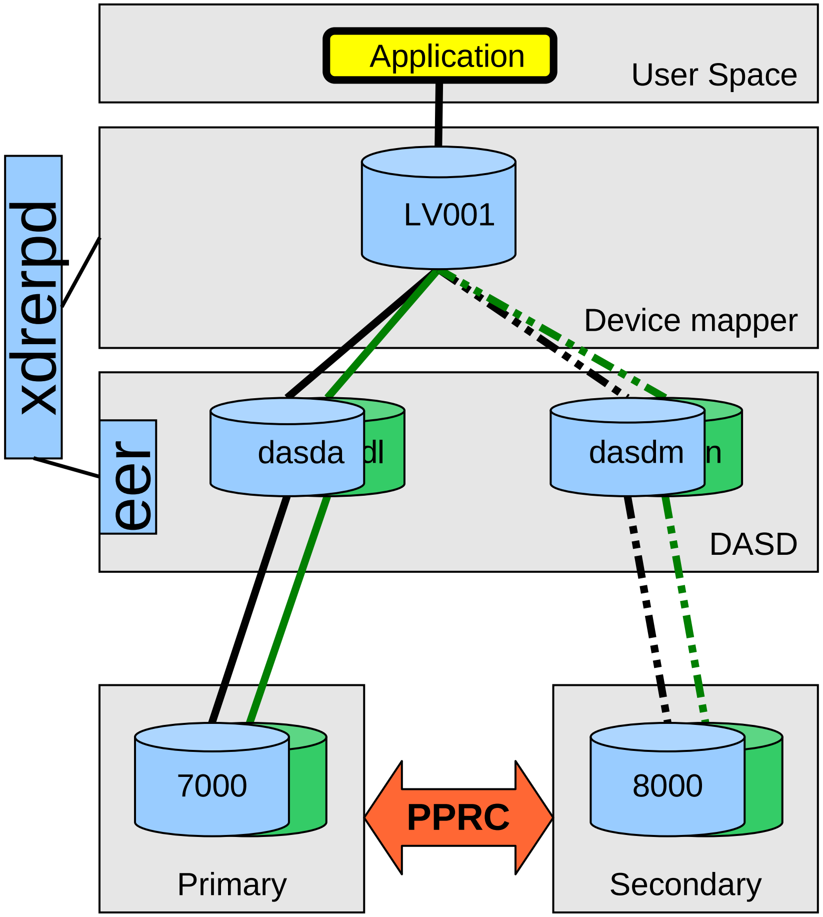
- Adapt user space tools (LVM2) to provide
  - Additional parameter for configuration (e.g. timeout)
  - Tolerance for stalled disks
  - Operation with missing disks
  - Enhanced real time capabilities for disk mirrors
- New perl script (statistics.pl) to extract statistical information like
  - Missed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

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 allows quiesce / resume and enable / disable of devices



Horst Hummel, IBM STG-LTC Boeblingen

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

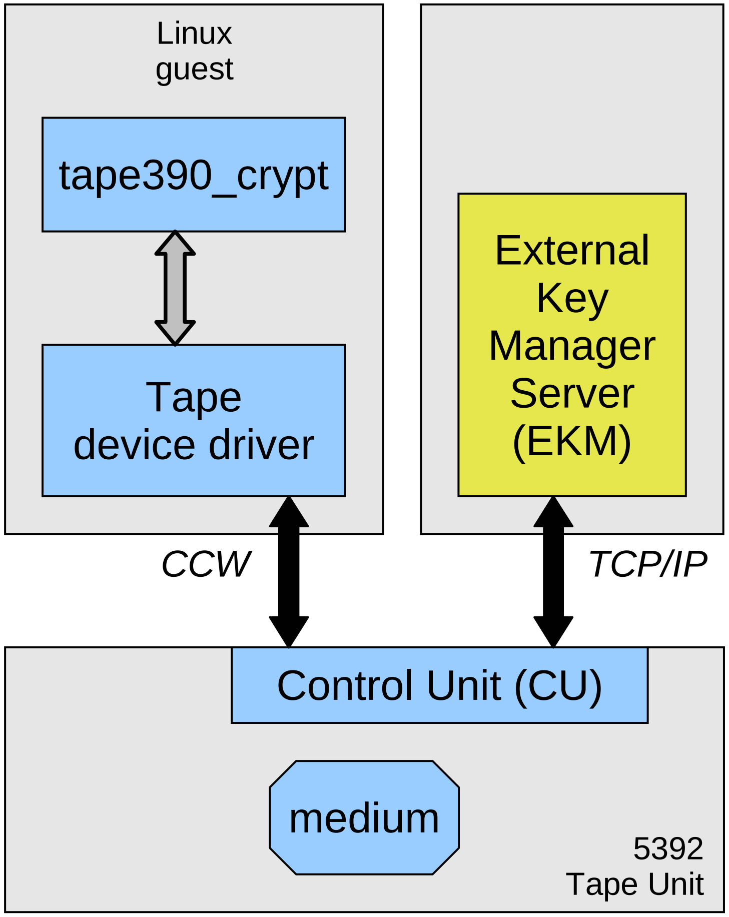
3592 tape encryption support (1Q2007)

- 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)

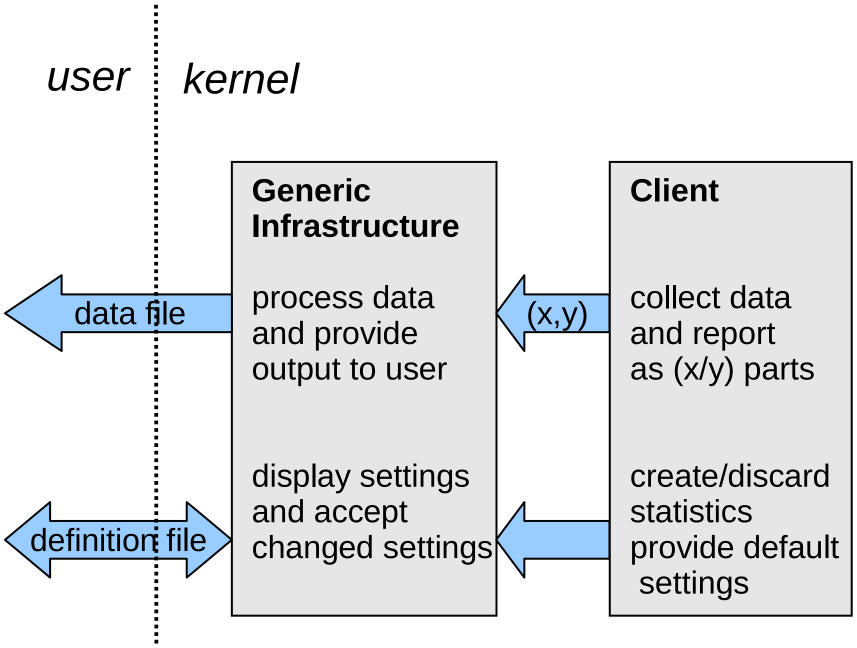
- Generic Infrastructure
  - Data output
    …/statistics/<scsi-lun>/data
  - Definition file
    …/statistics/<scsi-lun>/definition

- 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



user *kernel*

data file

definition file

Generic Infrastructure

process data and provide output to user

display settings and accept changed settings

(x,y)

Client

collect data and report as (x/y) parts

create/discard statistics provide default settings

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)

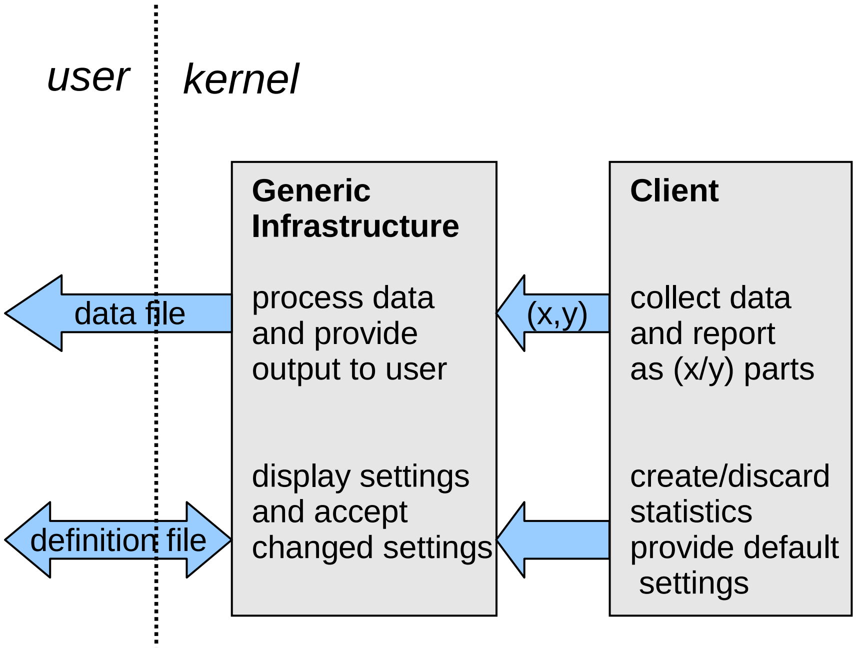
    ```
    # chchp --configure 1 0.40
    ```
  - logical vary on/off

    ```
    # chchp --vary 1 0.40
    ```
- Query configuration state

```
# lschp
CHPID Vary Cfg. Type Cmg Shared
====================================
```

Dynamic CHPID reconfig via SCLP (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



Generic Infrastructure

process data and provide output to user

display settings and accept changed settings

Client

collect data and report as (x/y) parts

create/discard statistics provide default settings

Feature-Matrix
for 1Q2006 code drop

Feature 1Q2006	RHEL		SLES	
	4	5	9	10
DASD DIAG 64bit support	U5	GA	SP3U1	GA
FAILFAST support in DASD device driver	U6	GA		GA
FC transport class exploitation		GA		GA
Multiple Subchannel Set (MSS) support		GA		GA
Generic Attriute Support in chccwdev		GA		GA
Improved zfcp Traces – Additional lcrash view		GA		GA
Improved SAN Notifications		GA		GA
zfcp/SCSI scripts: scsi logging level		GA		GA
zfcp/SCSI scripts: lszfcp		GA		GA

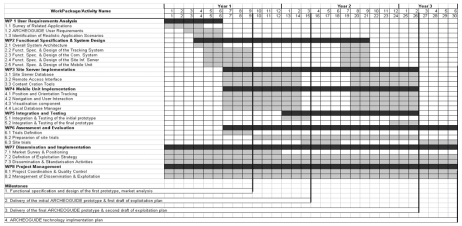
Feature-Matrix
for 4Q2006 code drop

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	--	--	--	--
FCP measurement data – Adapter statistics	--	--	SP4	--

Outlook
(subject to change)



- Multipath IPL / IPL trough IFCC
- SIM Handling for DASD ECKD devices
- DASD Hyper PAV enablement
  - support for Hyper PAV feature
  - automatic configuration detection
- zfcp performance statistics rework
  - use blktrace for common statistics
  - seperate part for z-specific adapter statistics
- performance improvements
- enhanced configuration support

Useful links

- Linux on System z – developerworks page

  http://www-128.ibm.com/developerworks/linux/linux390/

- Device Drivers, Features and Commands (SC33-8289-03)

  http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux390/docu/l26cdd03.pdf

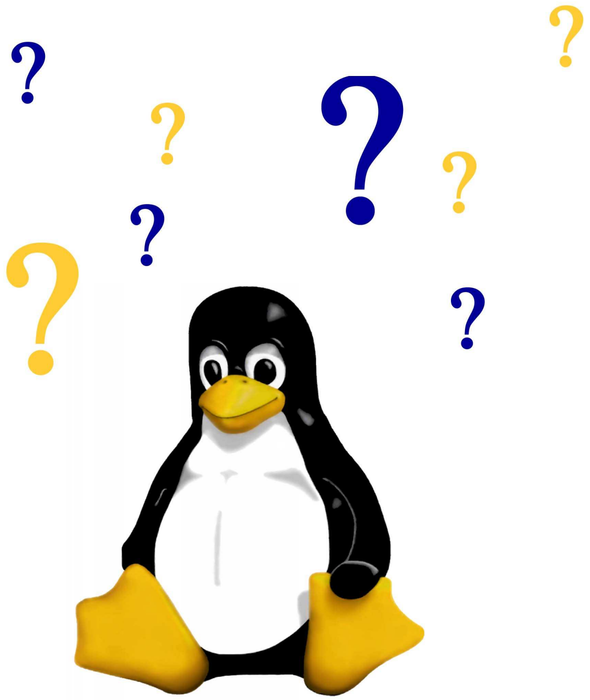
- How to Improve Performance with PAV (SC33-8292-01)

  http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux390/docu/l26chp01.pdf

- How to use FC-attached SCSI devices with Linux on System z (SC33-8287-00)

  http://download.boulder.ibm.com/ibmdl/pub/software/dw/linux390/docu/l26cts02.pdf

Questions



Trademarks

Trademarks

The following are trademarks of the International Business Machi nes 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 trademarksof 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 SE T 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 onmeasurements 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 individu al user will achieve throughput improvements equivalent to the p erformance 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 ar e presented as illustrations of the manner in which some custome rs have used IBM products and the results they may have achieved. Actual environmental costs and performance characteris tics will vary depending on individual customer configurations a nd conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this docum ent in other countries, and the information may be subject to change without notice. Consult your local IBM business contact f or information on the product or services available in your area.

All statements regarding IBM's future direction and intent are s ubject 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 pro ducts 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 produ cts.

Prices subject to change without notice. Contact your IBM repr esentative or Business Partner for the most current pricing in y our geography.

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

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

The information could include technical inaccuracies or typograp hical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of th e publication. IBM may make improvements and/or changes in the pr oduct(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 an y 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.