Logical Volume Management for Linux on System z

Horst Hummel
IBM
Horst.Hummel@de.ibm.com

February 29th 2008
Session 9282
Agenda

• Logical volume management overview
  • RAID levels
  • Striping
  • Mirroring
  • Multipathing

• Multipathing with zFCP / SCSI

• Multipathing with DASD using PAV

• Outlook on future development
Redundant Arrays of Inexpensive / Independent Disks (RAID)

- Using multiple disks to share or replicate data to increase:
  - Data integrity
  - Fault-tolerance
  - Throughput
  - Capacity

- Provides different configurations (RAID Level)

- Implemented as Software- or Hardware-RAID

Horst Hummel, IBM STG-LTC Boeblingen
RAID Level

• **Linear device (JBOD)**
  - Concatenate multiple to single virtual disk

• **RAID-0 (striping)**
  - Data is split evenly across disks (round robin)
  - Fast and efficient (no redundant information stored)
  - No fault-tolerance

• **RAID-1 (mirroring)**
  - exact data copy to 2 or more disks
  - Fast on read slow on write
  - Fault-tolerance (redundant data)
  - Needs additional capacity
RAID Level (cont.)

• **RAID-2**
  - Stripe data at **bit level** across several disks
  - Use 'Hamming code' for error correction
  - Intended for use with no built-in error detection

• **RAID-3**
  - Stripe data at **byte level** across several disks
  - Parity stored on dedicated disk (bottleneck)
  - Cannot serve multiple requests simultaneously
  - Parity allows recovery of single disk failure

• **RAID-4 (Striping & Dedicated)**
  - Stripe data at **block level** across several drives
  - Otherwise similar to level 3
RAID Level (cont.)

- **RAID-5 (Striping & Distributed Parity)**
  - Distribute parity among disks
  - Otherwise similar to level 4

- **RAID-10 (Mirroring & Striping)**
  - Combination of RAID-1 and RAID-0 (mirroring of striped device)
  - Good performance & Fault tolerance
Logical Volume Management (LVM) Terms

- **Physical volume (PV)**
  - Any kind of block device (DASD, SCSI, ...)

- **Physical Extend (PE)**
  - Even sized parts of the physical volume

- **Volume Group (VG)**
  - Pool of physical extends

- **Logical volume (LV)**
  - Virtual block device based on concatenated pooled PEs

- **Logical Extend (LE)**
  - Part of a logical volume
  - Same size as physical extend of the volume group
  - 1:1 mapping LE:(PV:PE)
LVM – Simple Example (linear device)

Volume Group 96MB

LV1
LE 0-5
6*4 = 24MB

LV2
LE 0-7
8*4 = 32MB

LV3
LE 0-9
10*4 = 40MB

PV1
PE 0-11
12*4 = 48MB

PV2
PE 0-11
12*4 = 48MB

Mapping LE->PV/PE

1PE = 1LE = 4MB (default size)

Logical Volume (LV)
e.g.
LV3: LE7
-> PV2: PE9

Physical Volume (PV)
LVM environment for striping

- Performance improvement due to multiple small disks
- No fault-tolerance
- Data evenly split across disks
LVM setup for mirroring

- Same data on each mirror
- Fault-tolerance
  Failing mirror can be recovered non-disruptive
- Needs double (or more) storage capacity
LVM setup for multipathing

- Performance improvement due to path load sharing
- Path fault tolerance (path failover / failback)
- Designed to handle all kind of block devices
- No storage server fault tolerance
Linux Multipathing Architecture

- **Logical Volume Management applications**
  - **dmsetup**
    - low level logical volume management
  - **LVM2**
    - latest version of Logical Volume Manager
  - **Multipath**
    - multipath configuration tool
  - **EVMS**
    - Enterprise Volume Management System

![Diagram](chart.png)
Linux Multipathing Architecture (cont.)

- **Libdevmapper**
  library for interaction between user and kernel device mapper

- **Device Mapper**
  - Modular framework for stacking target drivers like
    - Linear target
    - Mirror target
    - Multipath target
  - Responsibilities
    - Discover set of associated devices
    - Create mapping table containing configuration information
    - Pass mapping table into kernel
    - Possibly save mapping information
SAN Addressing
Path to FCP device

Device Number
(adapter busid)
e.g. 0.0.190e

Worldwide Port Name
(WWPN)
e.g. 0x500507630300c562

Logical Unit Number
(LUN)
e.g. 0x401040ae00000000
Multipathing with zFCP / SCSI Configuration

- **SCSI disk configuration (first path)**
  with bus ID 0.0.190e, WWPN0x500507630300c562 and LUN 0x401040ae00000000.

- Change to adapter directory
  ```bash
cd /sys/bus/ccw/drivers/zfcp/0.0.190e
  ```

- Set the adapter to online
  ```bash
  0.0.190e # chccwdev -e 0.0.190e
  ```

- Check for messages (in `/var/log/messages`)
  ```bash
  scsi2 : zfcp
  zfcp: The adapter 0.0.190e reported the following characteristics:
  WWNN 0x5005076400c2d09e, WWPN 0x5005076401a07fd4, S_ID 0x00688a13,
  adapter version 0x3, LIC version 0x606, FC link speed 2 Gb/s
  zfcp: Switched fabric fibrechannel network detected at adapter 0.0.190e.
  ```

- Add target port to FCP adapter
  ```bash
  0.0.190e # echo 0x500507630300c562 > port_add
  ```
Multipathing with zFCP / SCSI Configuration (cont.)

• Change to newly created port directory
  `0.0.190e # cd 0x500507630300c562/`

• Add FCP LUN to that port
  `0.0.190e/0x500507630300c562 # echo 0x401040ae00000000 > unit_add`

• Find new messages
  
  Vendor: IBM  
  Model: 2107900  
  Rev: .216  
  Type: Direct-Access ANSI SCSI  
  revision: 05  
  SCSI device sda: 10485760 512-byte hdwr sectors (5369 MB)  
  sda: Write Protect is off  
  SCSI device sda: drive cache: write back  
  sda: unknown partition table  
  sd 2:0:0:0: Attached scsi disk sda  
  sd 2:0:0:0: Attached scsi generic sg0 type 0

• SCSI disk is now available
  `0.0.190e # lsscsi`
  `[2:0:0:0] disk IBM 2107900 .216 /dev/sda`
Multipathing with zFCP / SCSI Configuration (cont.)

• SCSI disk configuration (remaining paths)

```shell
# cd /sys/bus/ccw/drivers/zfcp/0.0.190e/
0.0.190e # echo 0x500507630303c562 > port_add
0.0.190e # echo 0x401040ae00000000 > 0x500507630303c562/unit_add
0.0.190e # cd ..
zfcp # cd 0.0.520e/
0.0.520e # echo 0x500507630300c562 > port_add
0.0.520e # echo 0x401040ae00000000 > 0x500507630300c562/unit_add
0.0.520e # echo 0x500507630303c562 > port_add
0.0.520e # echo 0x401040ae00000000 > 0x500507630303c562/unit_add
0.0.520e # lsscsi
[1:0:0:0] disk IBM 2107900 .216 /dev/sdc
[1:0:1:0] disk IBM 2107900 .216 /dev/sdd
[2:0:0:0] disk IBM 2107900 .216 /dev/sda
[2:0:1:0] disk IBM 2107900 .216 /dev/sdb
0.0.520e #
```
Multipathing with zFCP Multipath Configuration

- **Start multipathd**
  ```
  linux:~ # /etc/init.d/multipathd start
  ```

- **load dm-multipath module, activate mp-tools**
  ```
  linux:~ # /etc/init.d/boot.multipath start
  ```

- **Check for multipath configuration**
  ```
  linux:~ # multipath -ll
  36005076303fffc56200000000000010aeIBM,2107900
  [size=5G][features=1 queue_if_no_path][hwandler=0]
  \_ round-robin 0 [prio=4][active]
  \_ 2:0:0:0 sda 8:0 [active][ready]
  \_ 2:0:1:0 sdb 8:16 [active][ready]
  \_ 1:0:0:0 sdc 8:32 [active][ready]
  \_ 1:0:1:0 sdd 8:48 [active][ready]
  ```

- **Device node provided by mp-tools**
  ```
  linux:~ # ls -l /dev/mapper/
  total 0
  brw------- 1 root root 253, 0 Jan 4 11:47 36005076303fffc56200000000000010ae
  lrwxrwxrwx 1 root root 16 Jan 4 11:15 control -> ../device-mapper
  ```
Multipathing with zFCP Partitioning

• **Write partition table to disk**
  
  `linux:~ # fdisk /dev/sda`

  --> *follow instructions to create primary partition*

• **Check device nodes**
  
  `linux:~ # ls -l /dev/mapper/`

  total 0
  brw------- 1 root root 253, 0 Jan 4 12:03 36005076303ffcc56200000000000010ae
  brw------- 1 root root 253, 1 Jan 4 12:03 36005076303ffcc56200000000000010ae-part1`
Multipathing with DASD using static PAV

- One base path (blue) to physical device
- Additional alias paths (green)
- Increased performance and path-fault tolerance
- Needs additional subchannels
Multipathing with DASD HW configuration

- PAV configuration on Storage Server
  Please refer to
  * IBMTotalStorage Enterprise Storage Server Web Interface User’s Guide, SC26-7448*

- zSeries configuration (IOCP)
  
  ***************************************************
  * DEFINE 3390-9 BASE AND ALIASES ADDRESS*          *
  * 16 BASE ADDRESS, 3 ALIASES PER BASE*              *
  ***************************************************

  IODEVICE ADDRESS=(7000,016),CUNUMBR=(5000),STADET=Y,UNIT=3390B
  IODEVICE ADDRESS=(70D0,048),CUNUMBR=(5000),STADET=Y,UNIT=3390A
Multipathing with DASD

DASD configuration

- 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 device configuration (base device)
  - Set base devices online
    # chccwdev -e 0.0.7000
  - Check for messages (in '/var/log/messages')

    dasd(eckd): 0.0.7000: 3390/0A(CU:3990/01) Cy1:3339 Head:15 Sec:224
dasd_erp(3990): 0.0.7000: EXAMINE 24: No Record Found detected
dasd(eckd): 0.0.7000: volume analysis returned unformatted disk
Multipathing with DASD
DASD configuration (cont.)

- Low level format base device (if not already done)
  - get device name using 'lsdaspd'
    # lsdaspd
  - Format device
    # dasdfmt -b 4096 -y -p /dev/dasdb
    cyl  5 of  5
    100%

    Finished formatting the device.
    Rereading the partition table... ok

- Write partition table (if not already done)
  # fdaspd -a /dev/dasdb
  auto-creating one partition for the whole disk...
  writing volume label...
  writing VTOC...
  rereading partition table...
• **Find new messages**
  dasd(eckd): 0.0.7000: (4kB blks): 2404080kB at 48kB/trk compatible
disk layout
  dasdb: unknown partition table
  dasdb:VOL1/ 0X7000:
  dasdd:VOL1/ 0X7000: dasdd1

• **DASD device configuration (alias devices)**
  # chccwdev -e 0.0.70d0-0.0.70d2
  # lsdasd
Multipathing with DASD
Multipath configuration

- **Device-mapper configuration**
  - Load dm_multipath module (if not already available)
    ```bash
    # modprobe dm_multipath
    ```
  - Use multipath command to automatically detect paths to device
    ```bash
    # 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 dasdb 94:4 [undef][ready]
    \_ 0:0:10927:0 dasdc 94:8 [undef][ready]
    \_ 0:0:10778:0 dasdd 94:12 [undef][ready]
    \_ 0:0:10927:0 dasde 94:16 [undef][ready]
    ```
  - Access to multipath device
    ```bash
    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
    ```
Multipathing with DASD Performance (first glance)

Static PAV with bonnie (on prototype)

<table>
<thead>
<tr>
<th>Number of Devices (Base + x Alias)</th>
<th>Throughput in KB/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td></td>
</tr>
<tr>
<td>Base + 1 Alias</td>
<td></td>
</tr>
<tr>
<td>Base + 2 Alias</td>
<td></td>
</tr>
<tr>
<td>Base + 3 Alias</td>
<td></td>
</tr>
<tr>
<td>Base + 4 Alias</td>
<td></td>
</tr>
<tr>
<td>Base + 5 Alias</td>
<td></td>
</tr>
<tr>
<td>Base + 6 Alias</td>
<td></td>
</tr>
<tr>
<td>Base + 7 Alias</td>
<td></td>
</tr>
</tbody>
</table>

Legend:
- Char Write
- Char Read
- Block Write
- Block Read
Multipathing with DASD 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 usage
### ECKD and SCSI Comparison

<table>
<thead>
<tr>
<th>Configuration</th>
<th>ECKD DASD</th>
<th>SCSI Disk</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOCDS / zVM</td>
<td>IOCDS / zVM</td>
<td></td>
</tr>
<tr>
<td>(operator)</td>
<td>(operator &amp; linux admin)</td>
<td></td>
</tr>
<tr>
<td>Access Method</td>
<td>SSCH / CCW</td>
<td>QDIO</td>
</tr>
<tr>
<td>Block Size (Byte)</td>
<td>512, 1K, 2K, 4K</td>
<td>512</td>
</tr>
<tr>
<td>Disk Size</td>
<td>&lt; ~57GB</td>
<td>?</td>
</tr>
<tr>
<td>Formatting (low level)</td>
<td>dasdfmt</td>
<td>not necessary</td>
</tr>
<tr>
<td>Partitioning</td>
<td>fdasd</td>
<td>fdisk</td>
</tr>
<tr>
<td>File System</td>
<td>mke2fs (or others)</td>
<td></td>
</tr>
<tr>
<td>Access</td>
<td></td>
<td>mount</td>
</tr>
</tbody>
</table>
Useful Commands

- **lscss**
  list channel subsystem devices

- **lsdaspd**
  list DASD related device information

- **dasdview**
  display extended DASD information

- **lszfcp**
  list information about zfcp adapters, ports, and units

- **lsscsi**
  list all scsi devices

- **chccwdev -e/-d**
  enable/disable ccw device

- **dasdfmt**
  low level format for DASD (ECKD) devices

- **fdasd**
  partitioning tool for DASD

- **fdisk**
  partitioning tool for SCSI

- **multipath -ll**
  display multipath configuration
Outlook (subject to change)

- Enhancements in common block device stack to allow cancellation of a request
- Upstream integration of real time mirror solution
- PAV / HyperPAV support in DASD device driver
DASD Next Generation Multipath using HyperPAV support

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

Horst Hummel, IBM STG-LTC Boeblingen
DASD Next Generation Multipathing 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
Useful links

- Linux on System z – developerworks page
  http://www-128.ibm.com/developerworks/linux/linux390/

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

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

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

- Device-mapper development
  http://sourceware.org/dm/

- LVM HOWTO
  http://tldp.org/HOWTO/LVM-HOWTO/
Questions