



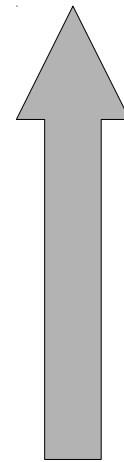
LVM2 – data recovery

Milan Brož
mbroz@redhat.com

LinuxAlt 2009, Brno

Linux IO storage stack

- **[VFS]**
filesystem
- **[volumes]**
MD / LVM / LUKS / MPATH ...
- **[partitions]**
legacy partition table
- **driver / IO scheduler**
block device layer, iSCSI, ...
- **HW**



**recovery
from the bottom up**

(also benchmark and optimize
performance this way)

Common storage problems

- **Storage failures**

- missing disk (cable, power, network – iSCSI)
- bad sectors
- intermittent HW failures

- **Accidental changes**

- **Overwritten metadata**

- **Bugs**

- firmware
- drivers
- volume manager
- filesystem

Planning storage (failures)

- **Failures are inevitable**
- **Losing data - problem?**
 - redundancy (RAID, replication)
 - backups (RAID is NOT a backup)
- **TEST all changes first**
 - most problems can be solved without data loss
 - **data loss is very often caused by operator error when trying to "fix" the problem**

Houston, we have a problem...

- **Don't panic!**
- **Think, try to understand the problem.**
read manual, error messages, logs, ...
- **Don't make changes
before the problem is understood.**
- **Backup.**
- **Test recovery strategy.**
- **Seek advice.**
paid support, mailing list, IRC

HW failures – intermittent / permanent

- **Disks, cables, connectors, power**
- **Firmware bugs**
- **Operator error (again)**
 - wrong cable connection

1) Fix HW

2) recover data

- **Bad sectors – use binary backups**
 - dd, dd_rescue, ...

Driver & disk partition problems

- **Driver / kernel version**
 - what changed – an update?
 - which version works, which not

- **Legacy partitions**
 - fdisk, parted – legacy partition table, GPT
 - partprobe – refresh in-kernel metadata
 - Gpart – guess & recover partition table

- **Check device sizes**
 - blockdev
 - fdisk / parted
 - sysfs

MD / LUKS / multipath – some pointers

- **MD – multiple device (RAID)**

- metadata, configuration
- mdadm, mdadm.conf
- cat /proc/mdstat, sysfs

- **LUKS – disk encryption**

- cryptsetup luksDump, status

- **Multipath**

- multipath -ll, multipath.conf

...

Volume management - metadata

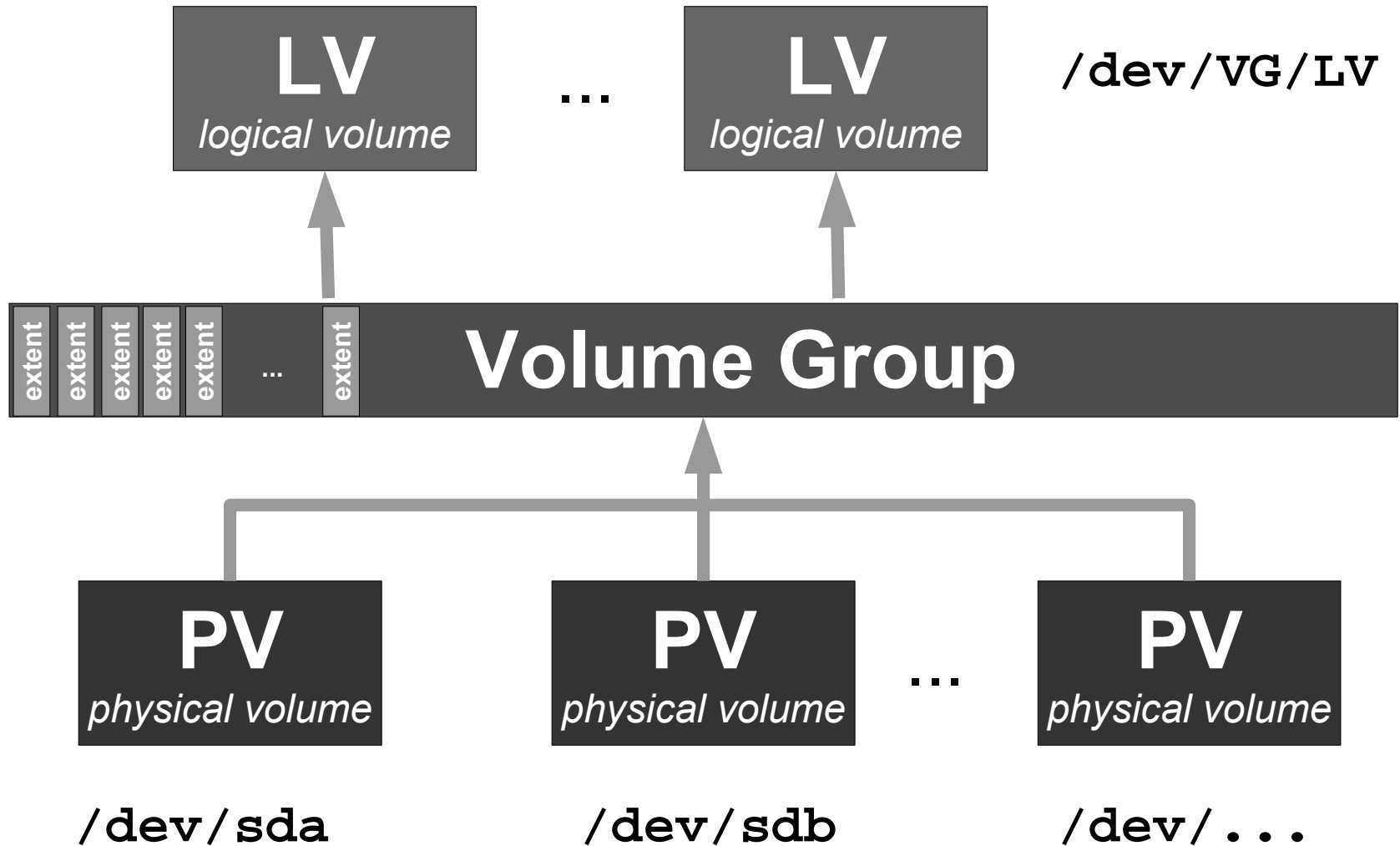
- **Metadata – "how to construct device"**

- **Where are metadata stored?**
 - on-disk (in first or last sectors)
 - in configuration file

- **MD** – on-disk, persistent superblock, handled in kernel
- **LVM2, LUKS** – on disk, handled in userspace
- **multipath** – multipath.conf, userspace daemon

- **Recover metadata then data**

LVM2 - overview



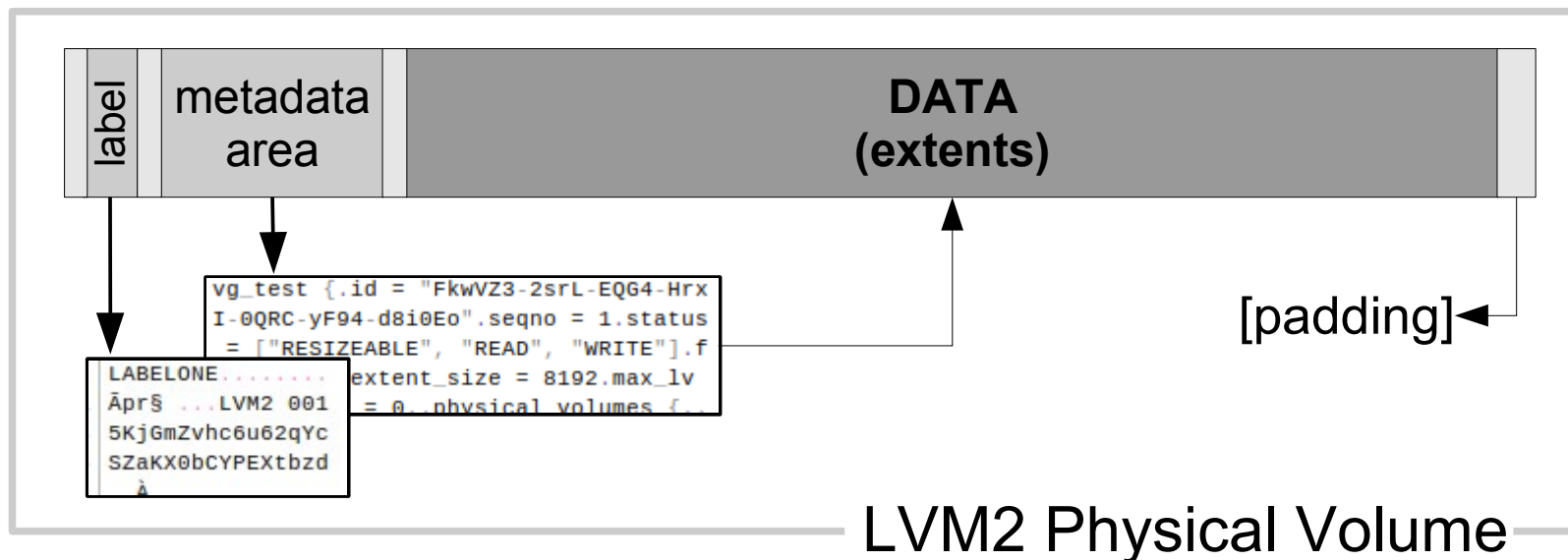
LVM2 – on-disk PV format

▪ Label

- 1 sector: signature, PV UUID, metadata area position

▪ Metadata Area

- 1 sector: **metadata area header** – pointer to metadata
- **circular buffer, text format** (at least 2 versions of metadata)
- **atomic update** – 1) write new version 2) update pointer
- **SEQNO** – sequential number
- checksum, redundancy, autorepair



LVM2 – text metadata example

```
creation_time = ...
description = "Created *after* executing ... "
...
vg_test {
    id = "xxxxxxx-xxxx-xxxx-xxxx-xxxx-xxxx-xxxxxxx"
    seqno = 25
    ...
    physical_volumes {
        pv0 {
            id = "xxxxxxx-xxxx-xxxx-xxxx-xxxx-xx...
            device = "/dev/sdb1" # Hint only
            ...
            pe_start = 384
            pe_count = 50 # 200 Megabytes }
        pv1 { ... }
    }
    logical_volumes {
        lv1 {
            id = "xxxxxxx-xxxx-xxxx-xxxx-xx...
            ...

```

Metadata backup

- **Archive & backup in /etc/lvm**
- **vgcfgbackup, vgcfgrestore**
- **Partial mode**
 - e.g. `vgchange --partial`
- **Test mode**
 - no metadata updates
 - `--test`
- **LVM2 & system info / debug**
 - `lvm dump`
 - `-vvvv` (all commands)

Example: [1/6] missing PV

Rescan all devices on system

```
# vgscan
```

```
Reading all physical volumes. This may take a while...  
Couldn't find device with uuid 'DHmMDP-bqQy-TalG-2GLa-sh6o-fyVW-3XQ3gp'.  
Found volume group "vg_test" using metadata type lvm2
```

Let's check what is on the missing device:

```
# pvs -o +uuid
```

```
Couldn't find device with uuid 'DHmMDP-bqQy-TalG-2GLa-sh6o-fyVW-3XQ3gp'.  
PV          VG          Fmt  Attr  PSize   PFree  PV UUID  
/dev/sdb    vg_test    lvm2 a-    200.00m  0      5KjGmZ-vhc6-u62q-YcSZ-aKX0-bCYP-EXtbzd  
unknown device vg_test    lvm2 a-    200.00m  0      DHmMDP-bqQy-TalG-2GLa-sh6o-fyVW-3XQ3gp
```

```
# lvs -o +devices
```

```
Couldn't find device with uuid 'DHmMDP-bqQy-TalG-2GLa-sh6o-fyVW-3XQ3gp'.  
LV   VG          Attr  LSize   Devices  
lv1  vg_test    -wi--- 100.00m /dev/sdb(0)  
lv2  vg_test    -wi--- 100.00m unknown device(0)  
lv3  vg_test    -wi--- 200.00m /dev/sdb(25)  
lv3  vg_test    -wi--- 200.00m unknown device(25)
```

Resume: lv1 is OK, lv2 is lost, lv3 – half is lost.

Example: [2/6] missing PV

You can activate only full available LVs now.

```
# vgchange -a y vg_test
```

```
Couldn't find device with uuid 'DHmMDP-bqQy-TalG-2GLa-sh6o-fyVW-3XQ3gp'.  
Refusing activation of partial LV lv2. Use --partial to override.  
Refusing activation of partial LV lv3. Use --partial to override.  
1 logical volume(s) in volume group "vg_test" now active
```

- for `--partial`, missing parts are replaced by device specified in

`/etc/lvm.conf: missing_stripe_filler = "error"`

- default is error (returns IO error on access)

Example: [3/6] missing PV

Let's try to recover at least something from lv3.

Prepare block "zero" device

```
# dmsetup create zero_missing --table "0 10000000 zero"

    missing_stripe_filler = "/dev/mapper/zero_missing"

# vgchange -a y vg_test --partial
...
3 logical volume(s) in volume group "vg_test" now active
```

**Always copy volume to another disk,
"zero" is replacement, not real disk – writes are ignored.**

Example: [4/6] missing PV

Remove all LVs on missing disk.

```
# vgreduce --removemissing vg_test
```

```
Couldn't find device with uuid 'DHmMDP-bqQy-TalG-2GLa-sh6o-fyVW-3XQ3gp'.
WARNING: Partial LV lv2 needs to be repaired or removed.
WARNING: Partial LV lv3 needs to be repaired or removed.
WARNING: There are still partial LVs in VG vg_test.
To remove them unconditionally use: vgreduce --removemissing --force.
Proceeding to remove empty missing PVs.
```

```
# vgreduce --removemissing vg_test --force
```

```
Couldn't find device with uuid 'DhmMDP-bqQy-TalG-2GLa-sh6o-fyVW-3XQ3gp'.
...
Wrote out consistent volume group vg_test
```

```
# pvs
```

PV	VG	Fmt	Attr	PSize	PFree
/dev/sdb	vg_test	lvm2	a-	200.00m	100.00m

```
# lvs -o +devices
```

LV	VG	Attr	LSize	Devices
lv1	vg_test	-wi---	100.00m	/dev/sdb(0)

Done.

Example: [5/6] missing PV

And now ... something completely different.

"operator error" - old device magically reappears!

```
# vgscan
```

```
Reading all physical volumes. This may take a while...
```

```
WARNING: Inconsistent metadata found for VG vg_test - updating to use version 18
```

```
Removing PV /dev/sdc (DHmMDP-bqQy-TalG-2GLa-sh6o-fyVW-3XQ3gp) that no longer belongs  
to VG vg_test
```

```
Found volume group "vg_test" using metadata type lvm2
```

```
# pvs
```

PV	VG	Fmt	Attr	PSize	PFree
/dev/sdb	vg_test	lvm2	a-	200.00m	100.00m
/dev/sdc		lvm2	--	204.00m	204.00m

Fixed automagically – SEQNO 18 – new metadata version.

Example: [6/6] missing PV

Recover the old metadata from backup.

```
# vgcfgrestore -f /etc/lvm/archive/vg_test_01564.vg vg_test
Cannot restore Volume Group vg_test with 1 PVs marked as missing.
Restore failed.
```

Device marked 'missing' – remove flag manually.

Edit the backup file:

```
...
pv1 {
    id = "DHmMDP-bqQy-TalG-2GLa-sh6o-fyVW-3XQ3gp"
    device = "unknown device"

    flags = ["MISSING"]
...

```

```
# vgcfgrestore -f vg_test_edited.vg vg_test
Restored volume group vg_test
```

Done.

Example: [1/3] overwritten PV header

Original system configuration (live example)

```
# pvs
PV          VG          Fmt  Attr  PSize  PFree
/dev/sda2   system_vg   lvm2  a-    7.89G   0
/dev/sdb    system_vg   lvm2  a-    8.00G   0

# lvs
LV          VG          Attr    LSize
root        system_vg   -wi-ao  13.86G
swap        system_vg   -wi-ao   2.04G
```

Example: [2/3] overwritten PV header

```
Found volume group "system_vg" using metadata type lvm2
Activating logical volumes
  Couldn't find device with uuid 'x8NH50-2CTA-LCHV-BSUF-ebJP-9531-j8d47f'.
  Couldn't find device with uuid 'x8NH50-2CTA-LCHV-BSUF-ebJP-9531-j8d47f'.
  Refusing activation of partial LV root. Use --partial to override.
  Couldn't find device with uuid 'x8NH50-2CTA-LCHV-BSUF-ebJP-9531-j8d47f'.
  1 logical volume(s) in volume group "system_vg" now active
Trying to resume from /dev/system_vg/swap
No suspend signature on swap, not resuming.
Creating root device.
Mounting root filesystem.
mount: could not find filesystem '/dev/root'
Setting up other filesystems.
Setting up new root fs
setuproot: moving /dev failed: No such file or directory
no fstab.sys, mounting internal defaults
setuproot: error mounting /proc: No such file or directory
setuproot: error mounting /sys: No such file or directory
Switching to new root and running init.
unmounting old /dev
unmounting old /proc
unmounting old /sys
switchroot: mount failed: No such file or directory
Kernel panic - not syncing: Attempted to kill init!
```

Example: [3/3] overwritten PV header

Rescue boot from DVD (log from live example)

1) Check the system state, volumes, partitions...

```
# lvm pvs
# lvm lvs -o name,devices
# lvm pvs -a
# fdisk -l /dev/sdb
# lvm pvs /dev/sdb1
# lvm pvs /dev/sdb
```

Resume: there is new partition table on /dev/sdb and new swap partition on sdb1 but correct lvm2 label on sdb is still present (but maybe metadata corrupted).

Task: remove partition table and recreate lvm metadata.

2) recover LVM metadata

```
# fdisk /dev/sdb
# lvm vgcfgbackup -f /vg system_vg
# (remove MISSING flag from /vg file)
# lvm pvcreate -u <missing PV UUID> --restorefile /vg system_vg
# lvm vgcfgrestore -f /vg system_vg

# lvm vgchange -a y system_vg
# fsck -f -C /dev/system_vg/root
...
```

Some common problems

- **Duplicated volume UUID**
- **Checksum error**
- **pvmove & suspended devices**
- **udev**
- **initrd missing drivers**

Storage performance

IO scheduler

CFQ / deadline / noop

per device / default (elevator= ...)

CFQ for desktop vs avoid for NFS, iSCSI, KVM

SSD – no seek time (noop, deadline)

Partitions

offset alignment (for SSD, fdisk -u to use sector units)

Beware of magic 63 sector offset (legacy DOS)

MD & LVM

MD chunk size aligned to LVM data offset

Readahead

LVM2 storage performance

Most values automatically detected now

automatic alignment for MD device

automatic readahead by default

detects IO topology layer (kernel > 2.6.31)

Large VG (many PVs)

only several metadata areas needed

`pvcreate --metadatascopies 0`

Read more in Red Hat Summit presentation

by Mike Snitzer

https://www.redhat.com/f/pdf/summit/msnitzer_1120_optimize_storage.pdf