

Filesystems, LVM, MD

- Typing some stupidities in text files, databases or whatever, where does it fit? why does it fit there, and how do you access *there* ?

Filesystem - Introduction

Short description of what is commonly used

Filesystem - Ext2

- At the beginning, there was ...

ext2 :

- ✓ Based on sysV filesystem
- ✓ Inode table usage
- ✓ Support links, subdirectories, attributes, locks
- ✓ Quota support

Filesystem - Journalizing

- **Fast, faster, fastest ...**

ext3 = ext2 + journal

- ✓ **More data corruption protection**
- ✓ **Fastest recovery**
- ✓ **ext2 compatibility**

Filesystem - Others

- **Because one is not enough**

Reiserfs

- ✓ Created from scratch
- ✓ B(inary)-tree concept

JFS

- ✓ Designed and used by IBM
- ✓ Recently opened to Free Software movement
- ✓ Reiserfs "like"

XFS

- ✓ SGI filesystem (IRIX based)

Filesystem - CDrom

- Because one is everywhere
 - ISO9660 with rockridge extension, allows long-filename and Unix-style symbolic links

Filesystem - Raw

- Because even it go fast, it can go faster

Pseudo raw file-system

- ✓ Linux kernel buffer-cache not used
- ✓ File like interface to read/write

Filesystem - Clustering

- Have the knowledge is good, share it make you good

OCFS – Oracle Cluster Filesystem

- ✓ Open-source
- ✓ Extent based
- ✓ simultaneous access
- ✓ filesystem interface on raw device
- ✓ easily resizable

GFS – Global Filesystem

- ✓ RedHat clustered filesystem
- ✓ Include Logical Volume Manager

Logical Volume Manager

Small introduction to LVM concept

LVM – Physical Volume

- Take some rocks

Physical volume

- ✓ `pvcreate <device_file>`
- ✓ Any "kind" of devices
- ✓ Any sizes

LVM – Physical extents

- Brake them in small equal parts

Physical extents

- ✓ fixed size
- ✓ defined by an unique ID

LVM – Volume Group

- Put all pieces in the same bag

Volume group

- ✓ `vgcreate <vgname> <pv1> <pv2> ...`
- ✓ Define which PV belongs to which VG
- ✓ Like a "big" device

LVM – Logical Volume

- Sort small pieces and link them together

Logical volumes

- ✓ `lvcreate [<options>] "-l <size_in_extent>|-L <size_in_HR_type>" -n <lvname> <vgname> [PVs]`
- ✓ LV is a regroupement of LE linked together in an specified order
- ✓ LV can be seen as an usual "partition"

LVM – Logical extents

- All pieces have a place in the "mosaic"

Logical extents

- ✓ Small piece of a LV
- ✓ Linked to a PE somewhere on a PV
- ✓ Same size as PE size

LVM – LVM tools

- "Mosaic" size is not defined

vgextend

vgreduce

lvextend

lvreduce

pvmove (premove)

- "Mosaic" can be paste on another wall

pvscan

vgscan

LVM – Features (1/2)

- "Mosaic" can't stand on different walls

Striping (LVM2 only)

- ✓ `lvcreate -i ...`
- ✓ Each stripe have the PE size
- ✓ Round-robbing striping is used

- "Mosaic" can evolve in time

Snapshot

- ✓ `lvcreate -s ...`
- ✓ Filesystem independant

LVM – Features (2/2)

- "Mosaic" can be seen from different view angles

Multipathing

- ✓ Not an official system development (IBM one)
- ✓ Available in various major "enterprise" linux release
- ✓ Links to PV have characteristics

MD - Introduction

Some about Software RAID linux driver

MD – Types (1/3)

- **Supported RAID type (1/3)**

Linear

- ✓ 2 devices and more appended
- ✓ No redundancy but FS could recover some chunks
- ✓ Performance enhanced only if read/writes are done on different disks

Raid 0

- ✓ 2 devices and more striped
- ✓ Ideally same sized
- ✓ No redundancy and nearly no chance to recover any data
- ✓ Read/write performance enhanced

MD – Types (2/3)

- Supported RAID types (2/3)

Raid 1

- ✓ 2 or more device + 0 or more hot-spare
- ✓ Ideally same sized
- ✓ Write performance worst, read performance slightly improved with a read balancing algorithm dependant of seek time operation

Raid 4

- ✓ Three or more disks striped with a parity device
- ✓ Support one drive failure
- ✓ Mainly not used as parity device become an hotspot bottleneck

MD – Types (3/3)

- **Supported RAID types (3/3)**
 - **RAID 5**
 - Three or more devices with zero or more spare with parity distributed among different devices
 - Support one drive failure
 - Read and write performance increased
 - **RAID 7 (as called as)**
 - Multipathing support

MD - Facts

- **Some facts about RAID**
 - ✓ **Not a data protection**
 - ✓ **Mixed RAID support**
 - ✓ **Filesystem independant**
 - ✓ **Performance should not be a RAID driver**
 - ✓ **No swap on RAID for performance reason**
 - ✓ **Only use one disk per controller**
 - ✓ **Reconstruction is transparent**

MD - Files

- Important files

/etc/raidtab

```
raiddev /dev/md0
raid-level      1
nr-raid-disks  3
nr-spare-disks 1
chunk-size     32
persistent-superblock 1
device         /dev/sda3
raid-disk 0
device         /dev/sdb6
raid-disk 1
device        /dev/sdd5
spare-disk     0
```

/proc/mdstat

MD - Tools

- **Raidtools and mdadm**

- Complementary and or supplementary**

- ✓ **Diagnose monitor and gather**
 - ✓ **Single centralized program**
 - ✓ **Without any needed config file by default**
 - ✓ **Can manager config file if one is needed**

MD - Howto

- **Create a RAID device**

```
mkraid /dev/md0
```

```
mdadm --create --verbose /dev/md0 --  
level=linear --raid-devices=2 /dev/sdb6 /  
dev/sdb5
```

- **Start/Stop RAID device**

```
raidstart /dev/md0 ; raidstop /dev/md0
```

```
mdadm -S /dev/md0 ; mdadm -R /dev/md0
```

MD - Parameters

- **Some parameter**
 - **Persistent-superblock**
 - `/etc/raidtab` on RAID device !!!
 - Superblock written at the beginning of each device
 - Keep `/etc/raidtab` up-to-date for future maintenance
 - **Chunk-size**
 - Size of the smallest striped/mirrored element
 - Mandatory in config file but unused in linear mode
 - Chunk size dependant of what will be hosted and of filesystem parameters

MD - Failure

- Drive failure detection

- ✓ `/var/log/messages`
- ✓ `/proc/mdstat`
- ✓ `mdadm --detail /dev/mdx`
- ✓ `lsraid -a /dev/mdx`

- RAID failure simulation

- Hardware

- Software

- ✓ `raidsetfaulty /dev/md1 /dev/sdc2`
 - ✓ `mdadm --manage --set-faulty /dev/md1 /dev/sdc2`
 - ✓ `raidhotadd /dev/md1 /dev/sdc2`

MD - Proactivity

- **Monitoring**

- **mdadm as a daemon**

- `mdadm --monitor --mail=root@localhost --delay=1800 /dev/md2`
 - **follow mode: test temporary failure**
 - **--program and --alert: reactive**

MD - Bootloaders

- **Boot on RAID**
 - **Lilo**
 - newer release support RAID 1
 - allow boot even if mirror is broken
 - **GRUB**
 - Need to specify the root device in RAID 1

Thanks

Many thanks for your listen. Feel free to ask any questions. I might have an answer

Ideas

- **Google**
- **LVM howto**
- **Software RAID howto**