

Napredno korištenje operacijskog sustava Linux

2. Datotečni sustav, RAID, LVM, kvote

Leonard Volarić Horvat
Nositelj: doc.dr.sc. Stjepan Groš

Sveučilište u Zagrebu
Fakultet elektrotehnike i računarstva

09.03.2017

- 1 Datotečni sustav
- 2 RAID
 - Osnovni RAID leveli
 - Ugniježđeni RAID leveli
 - Hardverski RAID
 - Softverski RAID
- 3 LVM
- 4 Kvote
- 5 Napredne mogućnosti

Datotečni sustav

File system

- Datotečni sustav (*file system*) određuje način spremanja i dohvaćanja podataka s medija
 - na tvrdom disku određen za svaku particiju
 - inicijalizacija *formatiranjem*
- Funkcionalnosti *file systema*:
 - normiranje imena datoteka i upravljanje direktorijima
 - metadata na datotekama
 - upravljanje prostorom na mediju:
 - smještanje podataka u sektore
 - grupiranje sektora u blokove
 - briga o fragmentiranim fajlovima

File system

- *File system* sadrži:
 - opisnike fajlova (veličina, lokacija, fragmentacija...)
 - imena fajlova
 - hijerarhiju direktorija (npr. FHS)
 - svoje parametre (npr. veličina bloka)

- Dakle, *file system* je:
 - sučelje između bajtova na disku i njihovog grupiranja u smislene cjeline
 - skup metapodataka koji opisuju pohranjene podatke

Neki datotečni sustavi

- FAT - *File Allocation Table*
 - Masovna podrška
 - FAT12 -> VFAT -> FAT16 -> FAT32 -> exFAT
 - Najveća veličina datoteke 4 GiB (FAT32)
- ext - *extended filesystem*
 - Razvijen za Linux sustave
 - ext2, ext3, ext4
 - ext4 danas najčešće korišten
 - Struktura metapodataka prilagođena Unix file systemu
 - Datoteke predstavljene strukturom *inode*
 - ext3 uvodi *journaling*

Ostali tipovi

- ISO 9660
- Linear Tape FS
- GlusterFS, BGFS
- NTFS - danas najčešći FS na modernim Windows verzijama
- posebni: swap, tmpfs

swap

- *Paging* particija
- Dio virtualne memorije

tmpfs

- Spremanje podataka na RAM
- Obično montiran na `/tmp`

Journaling

- Bilježenje promjena u FS-u
- Drastično se pospješuje robusnost sustava u slučaju kvara:
 - veća vjerojatnost uspješnog vraćanja izgubljenih podataka
 - lakša i puno brža dijagnoza i popravljjanje kvara
- Konfigurabilna granulacija logova

inode

- Struktura koja pohranjuje metapodatke o fajlu
- Dozvole, ID vlasnika, GID, veličina, broj hard linkova, MAC vremena
- Opaska: ime fajla **nije** zapisana u inodeu nego u direktoriju
 - *Everything is a file!*
 - direktorij je poseban fajl koji sadrži imena svih fajlova koje (konceptualno) sadrži
- Ovakav zapis omogućuje efikasno kopiranje i premještanje fajlova
 - kopiranje: dodavanje novog para (ime,inode) u ciljani direktorij
 - premještanje: dodavanje novog para u ciljani i brisanje starog para iz izvornog direktorija
 - **nema potrebe za stvarnim potencijalno dugotrajnim premještanjem sadržaja fajla**
- Ispis inodeova: `ls -li`

ext2

Ispod površine ext2fs:

- FS je podijeljen u blokove, 1-4KiB
- Blokovi su povezani u blok-grupe, veličine 8-512MiB.
- Svaka grupa sadrži:
 - jedan superblok - podaci o FS
 - FS opisnik (sigurnosna redundancija)
 - podatke
- File je predstavljen strukturom inode (*index node*)
- indode ima 15 pointera na podatke
- Ovisno o veličini blokova restrikcije su
 - Max file size 16 GiB - 2 TiB
 - Max FS size 4 TiB - 32 TiB

Upravljanje particijama

Stvaranje ext2 FS:

- stvoriti particiju
- stvoriti FS
- montirati FS

`fdisk`

- Alat za uređivanje particija na disku
- `fdisk` **ne formatira** particije
- Interaktivni način rada

`mkfs` odnosno `mkfs.<type>`

- Kreira filesystem
- **Formatira ciljani uređaj!**

Montiranje

- Postupak dodijeljivanja adrese u strukturi sustava nekom filesystemu
- FS na nekom uređaju (npr. /dev/sdb1) se uvrštava u datotečnu hijerarhiju sustava (FHS)

```
mount <device> <mountpoint>
```

- Bez argumenata - popis montiranih filesystema

```
umount <mountpoint>|<device>
```

- Particija se može identificirati na nekoliko načina:
 - /dev/sda1, /dev/sda2, ...
 - Moguća promjena oznaka
 - *Labela* filesystema
LABEL="Debian", ...
 - **UUID** (Universally **U**nique **I**dentifier)
UUID="de305d54-75b4-431b-adb2-eb6b9e546014"

/etc/fstab

- **File System Table**
- Sadrži opcije za automatsko mountanje filesystema
- Pokretanjem `mount -a` mountaju se filesystemi kako su redom navedeni u `/etc/fstab`

#	<filesystem>	<dir>	<type>	<options>	<dump>	<pass>
	/dev/sda1	/	ext4	defaults,noatime	0	1
	/dev/sda2	none	swap	defaults	0	0
	/dev/sdb1	/home	ext4	defaults,noatime	0	2
	tmpfs	/tmp	tmpfs	nodev,nosuid	0	0

Neke korisne naredbe za rad s diskovima:

- `df` - ispis zauzeća uređaja s FS-om
- `lsblk` - ispis blok-uređaja

FHS struktura

/	- root	
bin		<i>Osnovne korisničke izvršne datoteke</i>
boot		Datoteke bootladera
dev		Device datoteke
etc		Konfiguracija sustava
home		Matični direktoriji korisnika
lib		Biblioteke i kernel moduli
opt		Razni softver
root		Matični direktorij korisnika root
sbin		Sistemske izvršne datoteke
srv		Podaci servisa na računalu
tmp		Privremeni podaci
usr		Dijeljeni dio strukture
var		Često mijenjani i privremeni podaci

RAID

Konfiguracije diskova

Diskovi se u sustavu prikazuju kao logičke jedinice:

`/dev/sda`

`/dev/sdb`

`/dev/hda`

Problem: Pronaći metode za efikasno upravljanje raspoloživim diskovnim prostorom

- **JBOD** - Just a Bunch Of Disks
 - Diskovi se koriste neovisno
- **Spanned**
 - Više se diskova proširuje u jedan logički disk

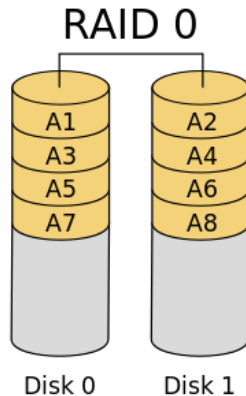
RAID

- **RAID** - **R**edundant **A**rray of **I**ndependent **D**isks
- *RAID polje* - logička jedinica sastavljena od više fizičkih diskova
- Prednosti
 - Povećanje prostora
 - Povećanje performansi
 - Redundancija (zaštita) podataka
- RAID-om se upravlja
 - Sklopovski RAID kontrolerom
 - Softverski md
- *RAID level* - Način rada RAID polja

RAID 0

Striping

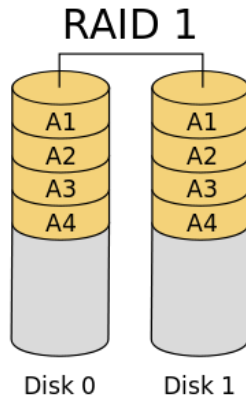
- Podaci se raspodjeljuju na više diskova
- Povećanje prostora
- Povećanje performansi
- Nema zaštite podataka



RAID 1

Mirror

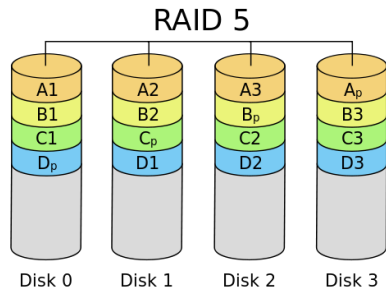
- Podaci se kopiraju na više diskova
- Zaštita podataka
- Nema povećanja prostora ni performansi



RAID 5

Block-striping with distributed parity

- Podaci se raspodjeljuju na više diskova
- Svakom bloku podataka se izračunava paritet i zapisuje na jedan od diskova
- Povećanje prostora
 - Potrebno osigurati dodatni prostor za paritet
- Povećanje performansi
- Zaštita podataka



Ostali RAID leveli

Nisu u (širokoj) upotrebi

RAID 2

- Hammingov kod za zaštitu podataka
- Dedicirani hard diskovi za zaštitne bitove

RAID 3, 4

- Paritetna zaštita
- Dedicirani hard disk za paritetne bitove

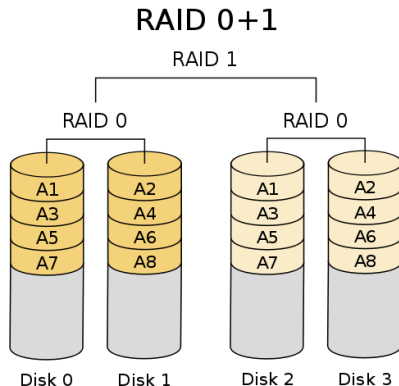
RAID 6

- Distribuirani zaštitni blokovi
- Dvostruki paritetni blokovi

RAID 0+1

Stripe, then mirror

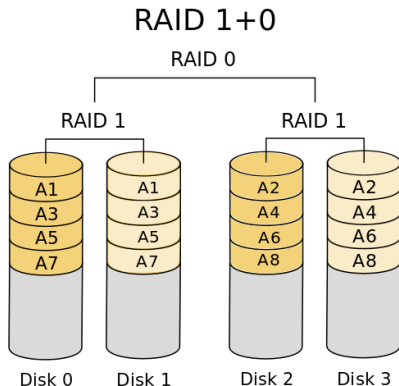
- Podaci se raspodjeljuju unutar jednog polja pa se cijelo polje kopira
- Prednosti RAID 0 na razini jednog polja
- Sigurnost RAID 0 polja



RAID 1+0

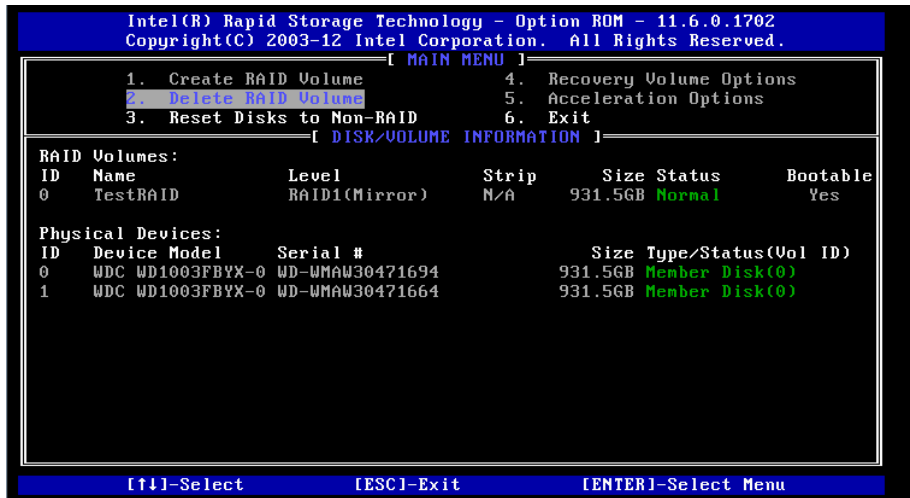
Mirror, then stripe

- Podaci se kopiraju unutar jednog polja pa se cijelo polje raspodjeljuje
- Sigurnost RAID 1 na razini jednog polja



RAID kontroler

RAID ROM



Softverski RAID

md - multiple device

- Linux implementacija softverskog RAID-a
- Podrška

Span, RAID 0, RAID 1, RAID 4, RAID 5, RAID 6, Nested

`mdadm`

`/dev/md*`

Particionirana polja

`/dev/md/md1p1`

`/dev/md/md2p1`

...

`/proc/mdstat` - popis inicijaliziranih polja

`mdadm` ne pamti polja pri ponovnom pokretanju

→ `mdadm --detail --scan >> /etc/mdadm.conf`

RAID boot

Hardverski RAID

OS vidi RAID polja kao i fizičke diskove. Nema izravni pristup fizičkim diskovima.

→ Bootloader radi kao u konfiguraciji bez RAID-a.

Softverski RAID

OS vidi fizičke diskove i iz njih gradi polje i logičke diskove.

→ Bootloader mora imati podršku za takva polja.

Hardware assisted / Fake RAID

Hibridni model. Kontroler ima ograničenu RAID podršku.

→ Bootloader vidi RAID polja kako logičke diskove. Ovisno o hardveru može bootati i bez dodatnih modula.

LVM

LVM

Logical Volume Manager

- Fleksibilnije upravljanje diskovnim prostorom
- Implementacija kroz **device mapper** (dm)
- Moguće dodavanje, uklanjanje i zamjena fizičkih i logičkih diskova za vrijeme rada sustava (čak i bez unmounta)

LVM arhitektura

Physical Volume (PV)

- Particije na fizičkim diskovima
- LVM ih dijeli na manje jedinice - **Physical extent** (PE)

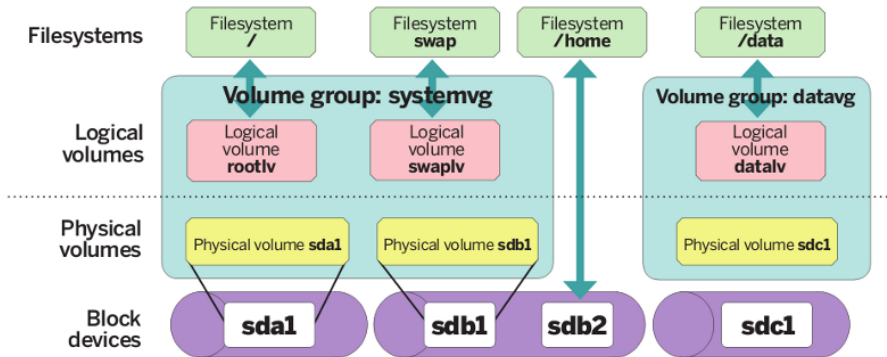
Logical Volume (LV)

- Logički disk (particija)
- LVM ih dijeli da manje jedinice - **Logical extent** (LE)

Volume Group (VG)

- Grupira više PV i LV u jednu skupinu radi mogućnosti upravljanja

LVM arhitektura



- Fizičke particije **sda1** i **sdb1** grupirane su u **systemvg**, a u grupi su stvorene dvije logičke particije: **rootlv** i **swaplv**
- Particija **sd1** je sama u grupi **datavg**
- Particija **sdb2** zaobilazi LVM i montirana je direktno na **/home**

LVM

Primjer

Kreiranje LVM logičke particije korištenjem dviju fizičkih particija

```
# stvaranje fizičkih particija
pvcreate /dev/sda1 /dev/sdb2

# stvaranje grupe moja_grupa i dodavanje navedenih particija
vgcreate moja_grupa /dev/sda1 /dev/sdb2

# Informacije o VG
vgscan
vgdisplay moja_grupa

lvcreate -l 100%FREE -n lvm0 moja_grupa
mkfs.ext3 /dev/lvm-disk/lvm0
```

Zašto LVM?

- Sloj apstrakcije između fizičkih diskova i smještanja podataka u logičke cjeline
- Jednostavnije dodavanje i uklanjanje fizičkih diskova
- Jednostavnija briga o veličinama i zauzeću particija
 - npr. ako je particija premala, bez problema možemo proširiti particiju na drugi disk
- Općenito drastično jednostavnija administracija fizičkih diskova

Kvote

Kvote

- Ograničavaju korištenje diskovnog prostora

`usrquota` Korisničke kvote

`grpquota` Grupne kvote

- Obične kvote
- *Journale*d kvote
 - Vode zapise o promjenama na disku što povećava pouzdanost

`quotacheck`: Your kernel probably supports journaled quota but you are not using it. Consider switching to journaled quota to avoid running `quotacheck` after an unclean shutdown.

Kvote

Podešavanje i naredbe

Datoteka /etc/fstab

```
# Obicne kvote
```

```
/dev/sda2 /home ext4 defaults,usrquota,grpquota 0 1
```

```
# Journaled kvote
```

```
/dev/sda2 /home ext4 defaults,usrjquota=aquota.user,  
                grpjquota=aquota.group,jqfmt=vfsv0 1 1
```

U prvom direktoriju trebaju biti datoteke aquota.user i aquota.group

/home/aquota.user

/home/aquota.group

```
quotacheck -avgum
```

```
quotaon -avgu
```

Kvote

Podešavanje i naredbe

```
# repquota -a
```

```
*** Report for user quotas on device /dev/md0
```

```
Block grace time : 7 days ; Inode grace time : 7 days
```

		Block limits			File limits			
User	used	soft	hard	grace	used	soft	hard	grace
root	-- 52	0	0		10	0	0	
veljko	-- 25585028	40000000	40000000		1123	0	0	
cetko	-- 5162460	40000000	40000000		49	0	0	
marin	-- 6498572	10000000	20000000		183	0	0	
deni	-- 5903852	10000000	20000000		528	0	0	
lovro	-- 3649796	10000000	20000000		19	0	0	
matej	+- 11334792	10000000	20000000	2 days	646	0	0	

Kvote

Podršavanje i naredbe

Soft limit Aktivacija *grace period-a* za vrijeme korisnik još može koristiti prostor

Hard limit Limit nakon kojeg korisnik nema mogućnost pisanja po disku

```
# edquota cetko
```

```
Disk quotas for user cetko (uid 1001):
```

Filesystem	blocks	soft	hard	inodes	soft	hard
/dev/md0	5162460	40000000	40000000	49	0	0

Napredne mogućnosti filesystema

Napredne mogućnosti filesystema

File attributes

File attributes

- Određuju poseban režim rada filesystema kod određenih datoteka / direktorija
- Vrste atributa određene su odabirom filesystema
- Za ext2 i novije ext sustave postoje naredbe za upravljanje atributima `lsattr`, `chattr`

Neki od ext2 atributa:

- a Append only - dopušta samo dodavanje sadržaja fajlu
- A No atime updates
- c Compressed - automatska kompresija fajla
- i Immutable - brani svaku promjenu fajla (**čak i od root!**)
- s Secure deletion - pri brisanju prebriše prostor nulama
- u Undeleteable - omogućuje vraćanje obrisanih podataka

...

Napredne mogućnosti filesystema

Primjer

Kreirajte datoteku koju korisnik root neće moći izbrisati korištenjem naredbe

```
# rm -f file.txt
```

Rješenje je napisano bijelim slovima

Napredne mogućnosti filesystema

Access Control Lists (ACL)

Access Control Lists (ACL)

- Proširenje UNIX dozvola
- Dodjeljivanje različitih dozvola različitim korisnicima i grupama

`setfacl`, `getacl`

- Filesystem mora biti mountan s opcijom `acl`

Napredne mogućnosti filesystema

Extended attributes

Extended attributes

- Parovi ključ:vrijednost koji se mogu po volji pridijeliti datotekama
- Oprez prilikom kopiranja datoteka
 - uobičajene naredbe za kopiranje ne čuvaju extended attribute
- Proučiti man stranice

Klase atributa

- | | |
|------------|-----------|
| • security | • trusted |
| • system | • user |

```
$ setfattr -n user.test -v "podatak" file.txt
$ getfattr -d file.txt
# file: file.txt
user.test="podatak"
```

Extended attributes

Capabilities

- Koncept ograničavanja mogućnosti izvršnih datoteka
- Cilj je izbjeći korištenje *setuid* bita ostavljajući izvršnoj datoteci privilegirani pristup nekim dijelovima sustava

```
$ getcap /bin/ping  
/bin/ping = cap_net_raw+ep
```

- Capabilities se zapisuju kao extended atributi

```
$ getfattr -d -m "^security\\\\" /bin/ping  
# file: bin/ping  
security.capability=0sAQAAAgAgAAAAAAAAAAAAAAAAAAAAA=
```

Loop devices

- Interpretacija običnih datoteka kao uređaja
- Datoteci se dodjeljuje *loop* uređaj u `/dev` folderu kojem se pristupa kao običnom disku
- Datoteka može sadržavati datotečni sustav

Primjer stvaranja loop device-a:

```
# Prazna 100MiB datoteka
```

```
dd if=/dev/zero of=device.img bs=512 count=2048
```

```
losetup /dev/loop0 device.img
```

```
mkfs -t ext3 /dev/loop0
```

```
mount -t ext3 /dev/loop0 /mnt/image
```

Literatura

http://www.ufsexplorer.com/und_fs.php

<http://www.tldp.org/HOWTO/Filesystems-HOWTO-6.html>

<http://www.nongnu.org/ext2-doc/ext2.html>

<https://www.linux.com/news/software/applications/8208-all-about-linux-swap-space> man mdadm

<http://www.ducea.com/2009/03/08/mdadm-cheat-sheet/>

<http://debian-handbook.info/browse/wheezy/advanced-administration.html>

https://www.howtoforge.com/linux_lvm

https://wiki.archlinux.org/index.php/Software_RAID_and_LVM

<http://www.tuxradar.com/content/lvm-made-easy>

https://wiki.archlinux.org/index.php/disk_quota