MALA PRIČA O DATOTEČNIM SUSTAVIMA

Dinko Korunić, CARNet
(Grupa za izradu CARNet Debian paketa)

Sadržaj

1. HDD sučelja i karakteristike
2. RAID uvod, performanse
3. Linux RAID i LVM
4. datotečni sustavi: IO scheduleri, datoteke, inode, tipovi
5. Ext3, Ext4, XFS, Btrfs, NFS, OCFS2
6. Microsoft datotečni sustavi
7. diskusija ☺

HDD sučelja – PATA i SCSI

- samo aktualna sučelja
- word-serial:
  - stariji, paralelno sučelje, serijska kom.
  - **IDE/ATA/PATA – Paralel ATA**
    - 40pin te kasnije 80pin, 16/32bit podaci
  - **SCSI – Small Computer System Interface**
    - LVD, HVD varijante signalizacije
    - razvoj: ..., Ultra-160, Ultra-320, Ultra-640
    - 68pin i 80pin, do 12m LVD

HDD sučelja – SATA

- bit-serial:
  - **SATA – Serial ATA**
    - nasljednik PATA standarda, ATA command set
    - karakteristike: half-duplex, hotswap, brža komunikacija, manje vodiča (tanji kablovi, 1m)
    - inačice: 1 (1.5 Gbps), 2 (3 Gbps), 3 (6 Gbps)
    - eksterni – eSATA
    - max read 285MB/s, max write 250 MB/s
    - **NCQ** – interna optimizacija redoslijeda upita (iodepth teoretski 32, realno **31**), manje rotacija, više izvršenih naredbi, odterećenje hosta

HDD sučelja – SAS

- bit-serial:
  - **SAS – Serial Attached SCSI**
    - karakteristike: serijska komunikacija, SCSI command set, **multipath**, serverska namjena, point-to-point linkovi prema uređajima
    - sučelje: full-duplex 3 ili 6 Gbps
    - čest dual-personality (kompatibilno sa SATA)
    - komponente: Initiator, Target, Service Delivery Subsystem, Expanders
    - max 64k uređaja, uređaji imaju WWN
    - protokoli: SSP, STP, SMP

HDD sučelja – FC

- bit-serial:
  - **FC – Fibre Channel**
    - bakar i optika
    - protokol: FCP (transportni, najčešće SCSI makar moguć i ATM, IP)
    - aktualno: 4GFC (800 Mbit), 8GFC (1600 Mbit), 16GFC (3200 Mbit), 10GFCS (2550 Mbit), 20GFC (5100 Mbit)
    - topologije: FC-P2P, FC-AL (petlja/prsten), FC-SW (FC preklopnici)
    - FC **HBA** uređaji za poslužitelje

HDD karakteristike

- disk **sector**:
  - veličina: 512, 1024, novi standard **4096**
  - **partition align** – paziti da se fs blokovi (4K) podudaraju sa fizičkima inače pad performansi zbog 2x read, 2x write
  - fdisk -H 224 -S 56 /dev/sda
  - parted ...
- dimenzije:
  - 0.85”, 1”, 1.8”, 2.5”, 3.5”, 5.25”
  - enterprise: SATA 3.5”, SAS tipično 2.5”

HDD karakteristike

- **shock resistance**:
  - 2D i 3D akceleratori
  - HDAPS, 3D DriveGuard, ...
  - parkiranje glave
- spojenost/lokacija:
  - lokalno spojeno (…) vs. eksterno (FW, USB, eSATA, SAS, FC)
  - NAS (Ethernet), SAN (Ethernet, FC)

HDD karakteristike

- **rotational speed** - pomak do željenog sektora:
  - rotational delay - manji sa većom brzinom
  - ograničavajuće: buka, toplina, vibracije
  - desktop rpm: 4200, 5400, 7200
  - enterprise rpm: 10000, 15000
- (disk-to-buffer) **transfer rate**:
  - sekvencijska brzina: rpm, gustoća zap.
  - 15krpm SAS ~ 120MB/s avg.
  - 10krpm SAS ~ 88MB/s avg.

HDD karakteristike

- 15 krpm vs. 10 krpm: s brojem diskova/spindlova u šasiji se smanjuje razlika agregatne propusnosti

- **seek time** - pomak do željene trake:
  - enterprise SAS: 3-5 ms
  - desktop SATA: 5-9 ms (mobile 5-12ms)
  - SSD: ~0.1 ms ☺
- **access time** - vrijeme pristupa podacima:
  - spin-up time + seek time + rotational delay

HDD karakteristike

- **IOPS** – I/O operacije u sekundi:
  - alati: Iometer, IOzone, FIO
  - varijacije: read/write, random/sequential
  - faktori: rpm, seek time, threads, IO queue depth, block size, ...
  - napredna logika: TCQ, NCQ... SAN

HDD	approx IOPS
5400 rpm PATA	50
7200 rpm SATA	75
10000 rpm SAS	125
15000 rpm SAS	175
SSD	6000-10000

RAID uvod

- performanse i/ili pouzdanost
- polje kao jedan disk (**array** vs. **LUN**)
- **HW** (visoka cijena, baterija, dedicirani cache, specijalizirani procesor, background operacije i provjere) vs. **SW** (višak resursa na hostu, jeftino?)
- **hotspare** – tipično jedan po šasiji
- kontroler također ima max IOPS ograničenje!

RAID nivoi

level	min # of disks	fault tolerance	space efficiency	description
RAID 0	2	0	1	stripe
RAID 1	2	n-1	1/n	mirror
RAID 5	3	1	1-1/n	stripe + distrib. parity
RAID 6	4	2	1-2/n	stripe + double distrib. parity
RAID 10	4	n-1	1/n	stripe of mirrors

- **parity**:
  - XOR – uz postojeći paritet je dovoljan bilo koji disk da se izračuna originalna informacija sa preostalog diska

RAID performanse

- RAID IOPS penalitet:
  - omjer operacija izgubljenih zbog RAID nivoa (zaštita, paritet, broj diskova)

level	read penalty	write penalty
RAID 0	1	1
RAID 1 RAID 10	1	2
RAID 5	1	4
RAID 6	1	6

  - traženo je 250 IOPS sa 50% read opterećenjem, 50% write opterećenjem u RAID 6. Potrebno polje mora odraditi 875 IOPS (7x 10krpm HDD ili 5x 15krpm!)

RAID performanse

- savjeti:
  - RAID 0 – kad podaci nisu bitni ☺
  - RAID 1 – intenzivne operacije nad manjim poljima, npr. MySQL, logiranje, e-mail maildir/mailbox (oprez, IOPS!)
  - RAID 5 – veliki statički sadržaj, npr. Apache2 statika, home direktoriji, itd.
  - RAID 6 – previsok penalitet za sve osim high-end storage
  - RAID 10 – izrazito visok IO, npr. Squid cache direktoriji za brze WAN linkove

Linux MD – uvod

- uređaj:
  - /dev/mdN ili /dev/md/N
- standardni nivoi: 1, 4, 5, 6, 10
- pseudo: 0, linear, multipath, faulty
- 128-bitni UUID; tip particije **fd**
- pozadinski check/resync
- alat: **mdadm**
- konfiguracija: **/etc/mdadm.conf**
- md/sync_action: check, repair, idle

Linux MD – tipični zahvati

- stvaranje novog uređaja:
  - mdadm --create /dev/md0 --level=mirror --raid-devices=2 /dev/sda1 /dev/sdb1
  - mdadm --create /dev/md1 --level=5 --raid-devices=3 /dev/sda2 /dev/sdb2 /dev/sdc2
- spajanje postojećeg uređaja:
  - mdadm --assemble /dev/md1 /dev/hda1
  - mdadm --assemble /dev/md3 /dev/hda3
  - mdadm --add /dev/md1 /dev/hdb1
  - mdadm --add /dev/md3 /dev/hdb3

Linux MD – tipični zahvati

- gašenje:
  - mdadm --stop /dev/md0
- stanje svih RAID uređaja:
  - cat /proc/mdstat
- stanje pojedinog uređaja:
  - mdadm --detail /dev/md0
- stvaranje konfiguracije:
  - mdadm -Es >>/etc/mdadm/mdadm.conf
  - mkinitramfs -k all -u

Linux LVM – uvod

- **PV** (physical volume): fizički uređaji sa više **PE** (physical extents)
- **VG** (volume group): nakupina PV
- **LV** (logical volume): dodijeljeni prostor unutar jednog VG
- mogućnosti:
  - dinamičko dodavanje fizičkog prostora (PV) u VG i širenje LV-ova
  - dinamički zahvati nad prostorom
- tip particije **8e**

Linux LVM – alati

- stvaranje:
  - pvcreate, vgcreate, lvcreate
- brisanje:
  - pvremove, vgremove, lvremove
- prikaz:
  - pvdisplay, vgdisplay, lvdisplay
- aktiviranje:
  - vgscan, lvscan
- provjera ispravnosti:
  - pvck

Linux LVM – alati

- preimenovanje:
  - vgrename, lvrename
- proširenje:
  - vgextend, lvextend
- skupljanje:
  - vgreduce, lvreduce
- promjena veličine (alternativa):
  - vgresize, lvresize
- i razni drugi...

Datotečni sustavi – IO sched

- disk scheduler:
  - optimize (reorder, delay, merge), fairness
- tipovi:
  - **CFQ** – completely fair queueing, za multiuser poslužitelje i desktop
  - **deadline** – SCAN, 2x read + 1x write, izbjegava starvation, deadline with expiry
  - **noop** – nema spajanja
  - **anticipatory** – predviđanje + deadline

Datotečni sustavi – IO sched

- podešavanje:
  - kernel boot: elevator=deadline
  - sysfs: echo deadline > /sys/block/sda/queue/scheduler
- deadline:
  - za cluster fs-ove i high IO servere
  - kad je potrebna garancija propusnosti
- CFQ:
  - garancija “jednakosti” (disk-time)
- noop: za Xen virtualke i SSD-ove

Datotečni sustavi – datoteke

- na Unix sustavu je sve datoteka, a ako nije datoteka onda je proces!
- tipovi datoteka:
  - regular file, directory, special file, link, socket, named pipe, block device
- particioniranje
  - data (tip **83**): root (/), /boot, /var, var/log, /home, /usr, /opt
  - swap (tip **82**): u praksi do veličine RAM-a
  - stroga Unix hijerarhija!

Datotečni sustavi – inode

- spremište najvažnijih informacija:
  - veličina datoteke (bytes)
  - koji uređaj sadrži datoteku (device ID)
  - vlasništvo (UID, GID)
  - tip datoteke (regular, directory, ...)
  - dozvole (ACL) i dodatni atributi
  - vrijeme stvaranja (ctime), zadnjeg čitanja (atime) i promjene (mtime)
  - broj linkova (soft, hard) na ovu datoteku
  - pokazivač do stvarne lokacije datoteke

Datotečni sustavi – inode

- važno!
  - ime datoteke nije nužno jedinstveno
  - brisanje čeka dok svi procesi koji mu pristupaju nisu završili
  - inode broj ostaje isti dokle god se datoteka premješta po istom fs-u
  - nije moguće hardlinkati direktorije
  - broj inodeova ograničen pri formatiranju...
- pathname to inode: find -inum
- dinamički rast: JFS, ext4, XFS

Datotečni sustavi – upotreba

- mount
  - povezivanje uređaja sa direktorijem
  - mount point: /media, /mnt, ...

file system	mount point	type	options	dump	pass
UUID=4f7e1a43-97f0-4146-ae21-f850ebcf1e96	/	ext4	defaults,usrjquota=aquota.user,grpjquota=aquota.group,jqfmt=vfsv0,noatime,nodelalloc	0	1

- fs alati:
  - mkfs, fsck, debugfs, tunefs, ...

Datotečni sustavi – mount

- atime update: relatime, noatime
- tipično: defaults,ro,rw
- ograničenja: nosuid, noexec, nodev
- dnevnik: data=journal, data=ordered, data=writeback
- razno: discard, nodelalloc, ...
- quota: quota, usrquota, grpquota, usrjquota=aquota.user, grpjquota=aquota.group, jqfmt=vfsv0

Datotečni sustavi – tipovi

- disk:
  - Ext2, Ext3, Ext4, Btrfs, ISO9660, ZFS, ReiserFS, swapfs, UDF, XFS, JFS, ...
- flash:
  - JFFS, JFFS2, YAFFS, LogFS
- tape
- database
- transactional:
  - podverzija journaling – Ext3, Ext4, XFS, ...

Datotečni sustavi – tipovi

- network:
  - NFS, AFS, SMB/CIFS, SSHFS
- shared disk (cluster):
  - GFS, GFS2, OCFS, OCFS2, GPFS
- specijalni:
  - tmpfs, sysfs, procfs

Datotečni sustavi – Ext3

- karakteristike:
  - in-place nadogradnja sa ext2
  - journal + Htree za velike direktorije
  - **4KiB, 2TiB max filesize, 16TiB fs max**
  - **8KiB, 2TiB max filesize, 32TiB fs max**
  - jedan od najsporijih fs-ova (fsck 4TB ~ 2h)
- alati:
  - e2fsprogs: **e2fsck**, **mke2fs**, itd.
  - jednostavna struktura, mogućnost dobrog oporavka (e2fsck, **Testdisk**, **findsuper**)

Datotečni sustavi – Ext3

- nivoi dnevnika:
  - **journal** – sve u dnevnik pa tek na disk
  - **ordered** – metadata u dnevnik pa na disk, postoji garancija spremanja na disk
  - **writerback** – samo metadata u dnevnik
- mane:
  - sporost!, nema online defragmentacija (Shake, defrag), “nemoguće” vraćanje obrisanih datoteka, nema transparentne kompresije, nema snapshota, nema journal checksumminga, 32000 subdirs

Datotečni sustavi – Ext4

- dio osnovne Linux jezgre od kraja 2008 – stabilna verzija od 2.6.28
- prednosti:
  - fs do 1EiB, file max do 16TiB
  - **extents** - nakupine blokova umjesto indirektnog mapiranja blok-datoteka
  - **kompatibilnost**; **alokacija više blokova** odjednom; **odgođena alokacija**
  - vrlo **brzi fsck** (2-20x ubrzanje!)
  - **journal checksumming**; online defrag, SSD podrška, ...

Datotečni sustavi – Ext4

- alati: e2fsprogs
- **offline konverzija** (i ne potpuna...):
  - tune2fs -O extents,uninit_bg,dir_index /dev/disk
  - fsck -yfDC0 /dev/disk
- također potrebno:
  - Grub2, quota (Ext4-compatible)
  - sto noviji kernel (npr. 2.6.32 ili noviji)
- Debian Lenny – standardno nije Ext4 kompatibilan ☹

Datotečni sustavi – XFS

- originalno sa IRIX sustava
- prednosti:
  - 64-bit, fs do 16EiB, file max do 8EiB
  - garantirana propusnost (streaming)
  - journaling, allocation groups, striped allocation, extent based allocation, variable block sizes, delayed allocation, sparse files, direct IO, DMAPI, extended ACL, snapshots(*), online defragmentation, online resize, atomic disk quota, dump/restore, …

Datotečni sustavi – XFS

- alati:
  - xfs_check – provjera konzistencije, nema gornje granice potrošnje memorije!
  - xfs_repair – popravlja greške
  - xfs_freeze – suspendira pisanje radi stvaranja snapshota (LVM...)
  - xfs_admin – promjene parametara pojedinog fs-a
  - xfs_growfs – širenje pojedinog fs-a tijekom rada (online)

Datotečni sustavi – Btrfs

- CoW fs, Oracle proizvod, fs nove generacije... ali još nije dovršen ☹
- enterprise namjena! (ZFS alternativa)
  - pooling, snapshots, checksums, multi-device spanning
- implementirano:
  - online grow/shrink, online block device add/remove, online defrag, online balancing, transparent compression, subvolumes, snapshots, file cloning

Datotečni sustavi – Btrfs

- implementirano:
  - object-level striping/mirroring (RAID 0, RAID1, RAID10), data & metadata checksums, in-place conversion (ext3/ext4), SSD podrška, ...
- upotreba:
  - mkfs.btrfs
  - btrfsctl –a ili btrfs device scan
  - btrfsck
- multi-device – mount sa svih uređaja

Datotečni sustavi – NFS

- brojne verzije protokola: v1 - v4
- brojni problemi:
  - sigurnost i udaljene ovlasti (root squash)
  - performanse TCP vs. UDP
  - locking
  - asinhroni rad
  - centraliziranost rješenja
- implementacije:
  - kernel (v3, v4) vs. userspace (UNFSv3)
  - userspace – rješenje za OpenVZ

Datotečni sustavi – OCFS2

- alternativa GFS/GFS2, integralni dio Oracle RDBMS, dio std. Linux jezgre
- karakteristike:
  - shared-disk, multi-node, cache-coherent, parallel IO, journalling, arch/endian neutral
  - distribuirano zaključavanje (flock(), integrirani DLM), interni heartbeating/quorum/fencing
- moguće koristiti:
  - za virtualizaciju, baze podataka, spremišta statičkih datoteka, ...

Datotečni sustavi – OCFS2

- kod deriviran iz ext3, jednostavan i efikasan
- konfiguracija:
  - /etc/ocfs2/cluster.conf
- alati:
  - mkfs.ocfs2
  - tunefs.ocfs2
  - fsck.ocfs2
  - o2cb_ctl
  - o2image

Datotečni sustavi – MS svijet

- NTFS
  - FUSE NTFS-3G driver, puna rw podrška
  - **ntfsprogs**
  - najčešće uzrokuje povišeno opterećenje
- FAT12, FAT16, FAT32
  - msdos (8.3), **vfat** (LFN), umsdos (LFN + Unix semantika)
- FAT64/exFAT
  - beta FUSE driver