Félelem nélkül élni - RAID és ENBD

- Tomka Gergely
- Szent István Egyetem, Gödöllő
- Jobbára programozó és rendszergazda
- Linux-elkötelezett
- Debian-elkötelezett
- tomka.gergely@ih.szie.hu

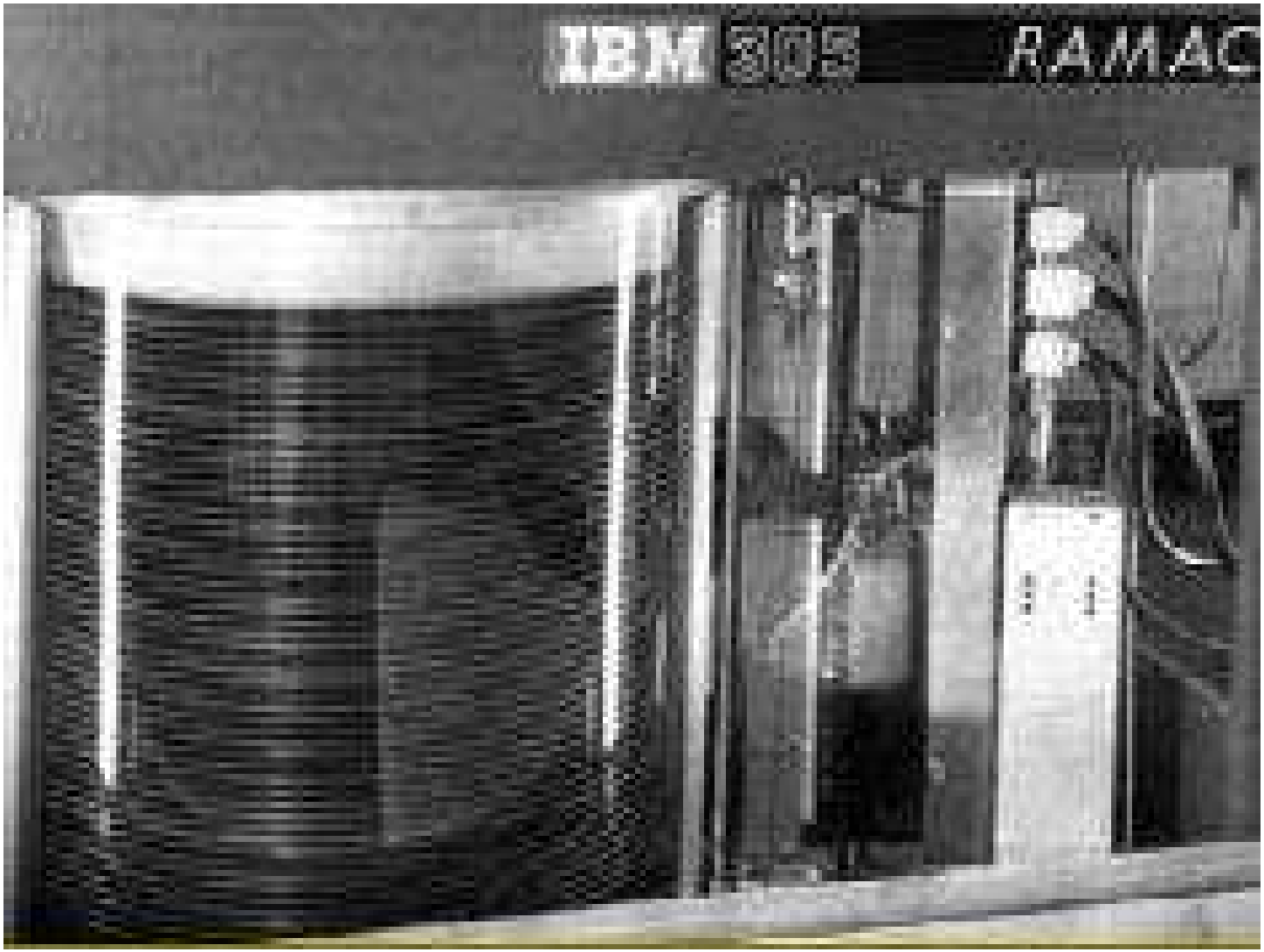
Érinteni tervezett témakörök

- Eszmefuttatás a merevlemezek halandóságáról
- RAID, részletekben
- Eszmefuttatás a RAID buktatóiról
- Eszmefuttatás a fizikai biztonságról
- Bevezető az ENBD használatába
- Egy buta, majdnem valós példa
- Kérdések, melyek zavarba hoznak majd

Adattárolás régen

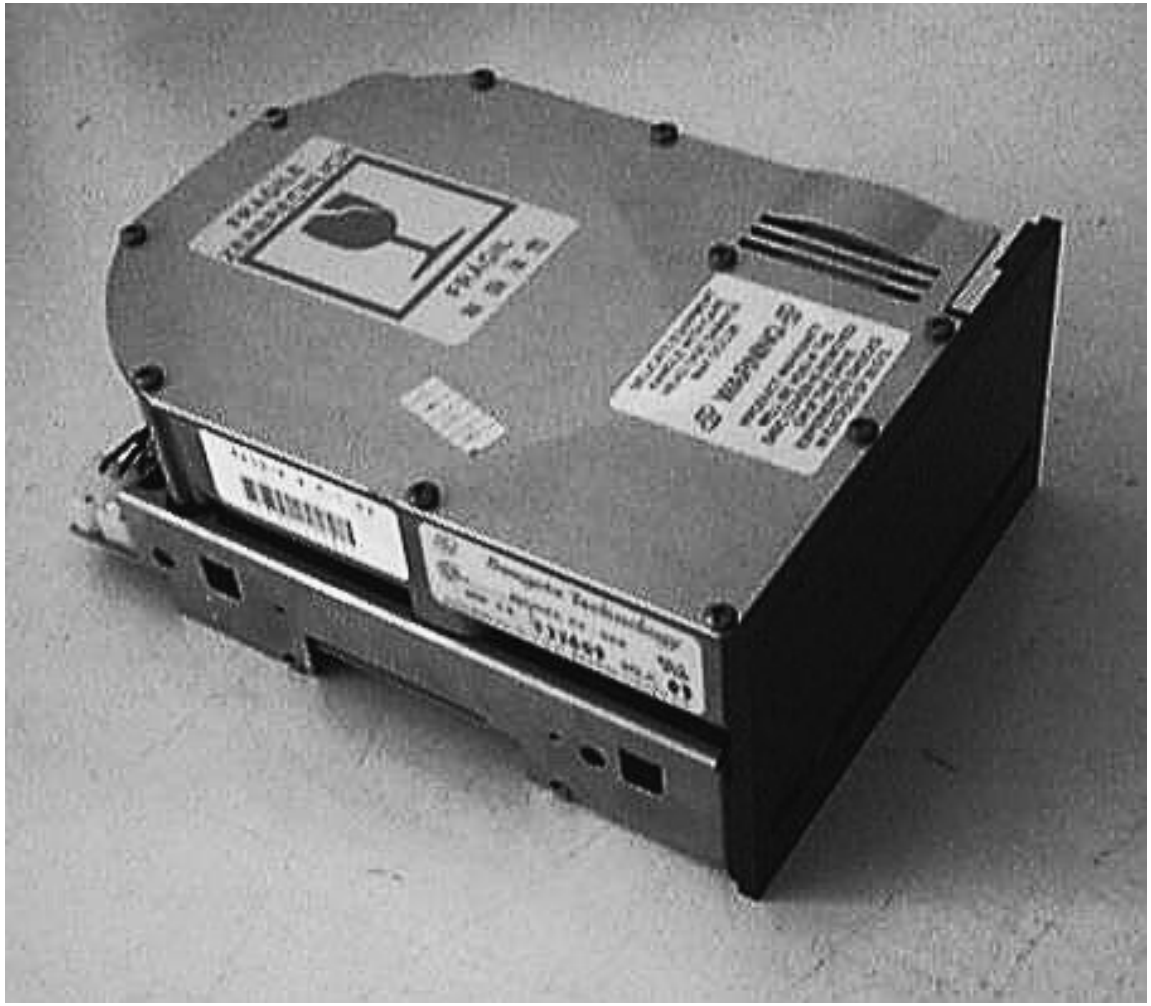
- Szárnyaló igények
- Porban kúszó lehetőségek
- Állandó küzdelem a technikai korlátok ellen
- Mérföldkövek, érdekeségképpen

Az első merevlemez



Mágnesdob memória

Az első PC merevlemez



Szélsőségesen mozgó alkatrészek

- Egyre gyorsabban mozgó alkatrészek
- Hasonló fordulatszámok a Formula 1-ben
- Megállás, elindulás problémája
- MTBF egyre csökken
- MTBF egyre határozottabb
- A modern merevlemezekben nem bízunk
- De legalább olcsóak

A RAID szükségmegoldás

- Mindig keressük az ideális tárolóeszközt
- Gyors, olcsó, megbízható – Szent Grál
- A való világban mindig csak kettő érvényesül
- Hogy melyik kettőt akarjuk, nem egyszerű döntés
- A RAID nem tökéletes megoldás
- Menteni mindig kell

Mit mondok el a RAIDról?

- Eszközök
- “szintek”, szervezési elvek
- Felhasznált eszközök
- Fontosabb RAID szintek
- IDE RAID nehézségek
- Szoftver vagy hardver RAID?

Eszközök

- Debian GNU/Linux – nem vagyok tökéletes
- Mdadm – univerzális raid kezelő
- /proc/mdstat – kernel információk
- bonnie++ - hozzávetőleges teljesítménymérés
- vmstat, top – semmi ördöngösség, figyelünk
- dokumentáció – bizony, bizony!

RAID szintek

- Sok van belőlük
- Jelentős részük speciális problémák megoldására
- Fontosabbak: RAID1, RAID5
- Hasznos még a RAID0 és RAID10
- Sok gyári, titkos is van, ne lepődjünk meg

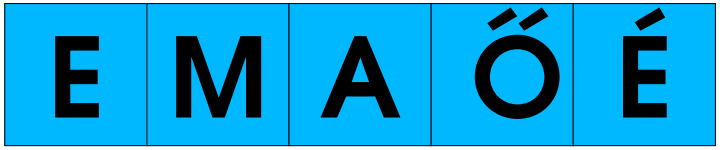
Előkészületek

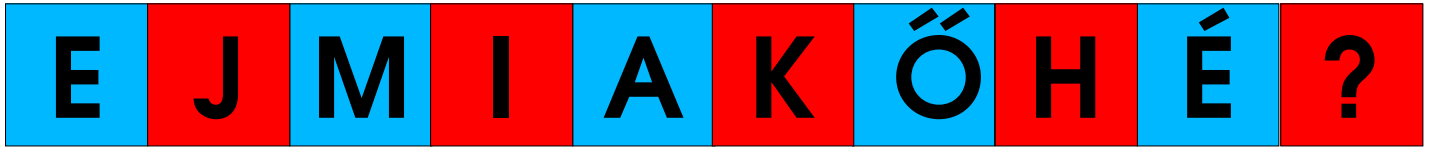
- Linux-specifikus rész
- Kernel RAID támogatás
- FD típusú partíciók
- Persistent Superblock

RAID0

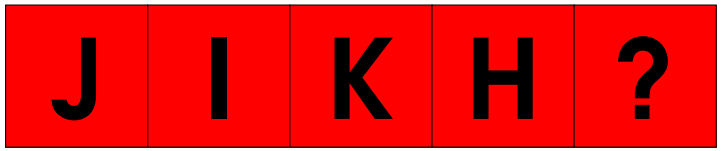
- Olcsó és gyors
- Megbízhatatlan
- Nagymennyiségű, gyorsan ömlő adathoz

RAID0 egészségesen





RAID0 betegen



Parancssorok

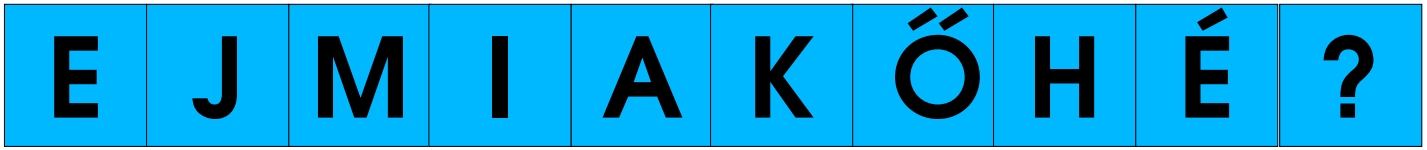
```
mdadm --create /dev/md0 --level=raid0\
--raid-devices=2 /dev/hda5 /dev/hda6
```

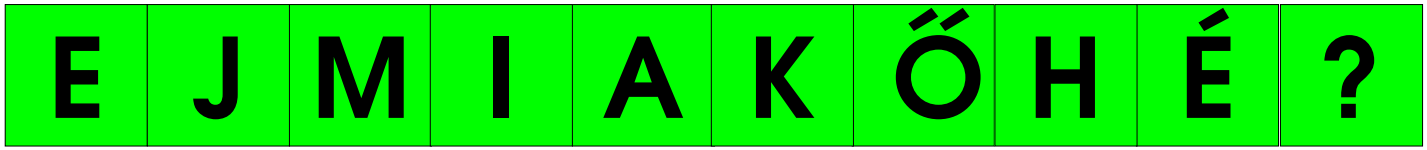
```
#  cat /proc/mdstat
Personalities : [raid0] [raid1] [raid5]
md0 : active raid0 hda6[1] hda5[0]
      9767296 blocks 64k chunks
```

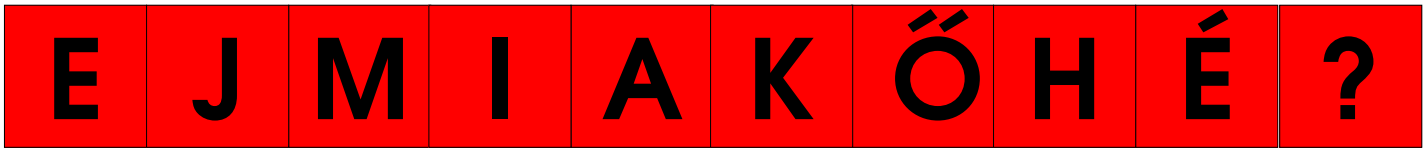
RAID1

- Biztonságos és gyors
- Nem takarékos
- Sokáig szinkronizálhat
- Sok rendszer ismeri, gyakran találkozhatunk vele
- Linux, Windows tud bootolni róla

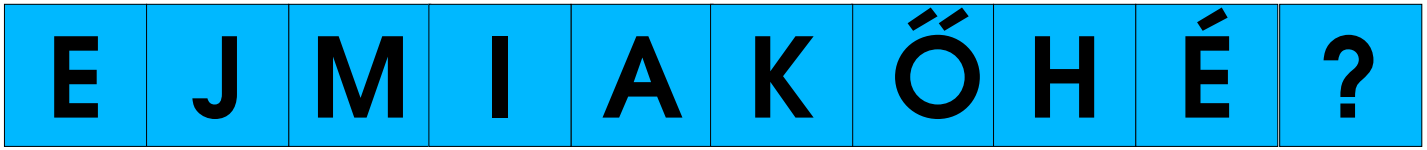
RAID1 egészségesen

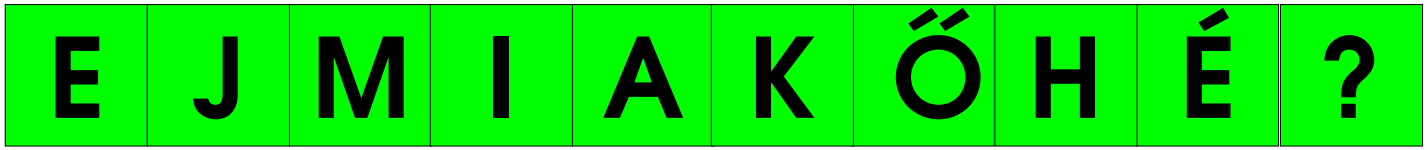


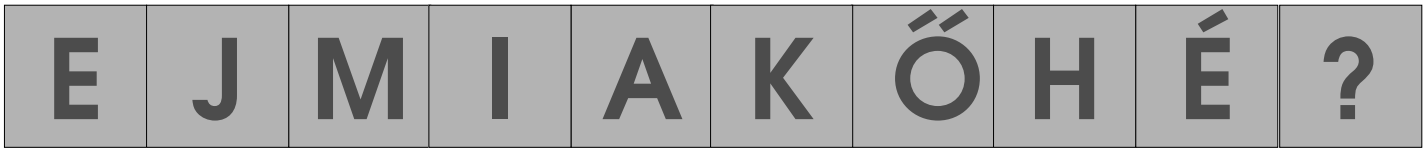




RAID1 betegen







Parancssorok

```
mdadm --create /dev/md1 --level=raid1\
  --raid-devices=2 /dev/hda5 /dev/hda6


md1 : active raid1 hda6[1] hda5[0]
      4883648 blocks [2/2] [UU]
      [=====>.............]  resync =  23.3%
   (1141056/4883648)\
finish=4.1min speed=15159K/sec
```

A tömb elrontása

```
mdadm /dev/md1 --fail /dev/hda6
mdadm: set /dev/hda6 faulty in /dev/md1
```

A tömb elrontása

logfejléc: raid1: Disk failure on hda6, disabling device.

logfejléc: Operation continuing on 1 devices

logfejléc: RAID1 conf printout:

logfejléc: --- wd:1 rd:2

logfejléc: disk 0, wo:0, o:1, dev:hda5

logfejléc: disk 1, wo:1, o:0, dev:hda6

logfejléc: RAID1 conf printout:

logfejléc: --- wd:1 rd:2

logfejléc: disk 0, wo:0, o:1, dev:hda5

A tömb megjavítása

```
md1 : active raid1 hda6[2](F) hda5[0]
      4883648 blocks [2/1] [U_]

mdadm /dev/md1 --add /dev/hda7
mdadm: hot added /dev/hda7
```

A tömb megjavítása

```
md1 : active raid1 hda7[2] hda6[3](F) hda5[0]
      4883648 blocks [2/1] [U_]
      [=========>.....]  recovery =  36.3%
    (1774464/4883648)\
 finish=3.5min speed=14515K/sec
```

A tömb megjavítása

```
md1 : active raid1 hda7[1] hda6[2](F) hda5[0]
      4883648 blocks [2/2] [UU]

mdadm /dev/md1 --remove /dev/hda6
mdadm /dev/md1 --add /dev/hda6

md1 : active raid1 hda6[2] hda7[1] hda5[0]
      4883648 blocks [2/2] [UU]
```

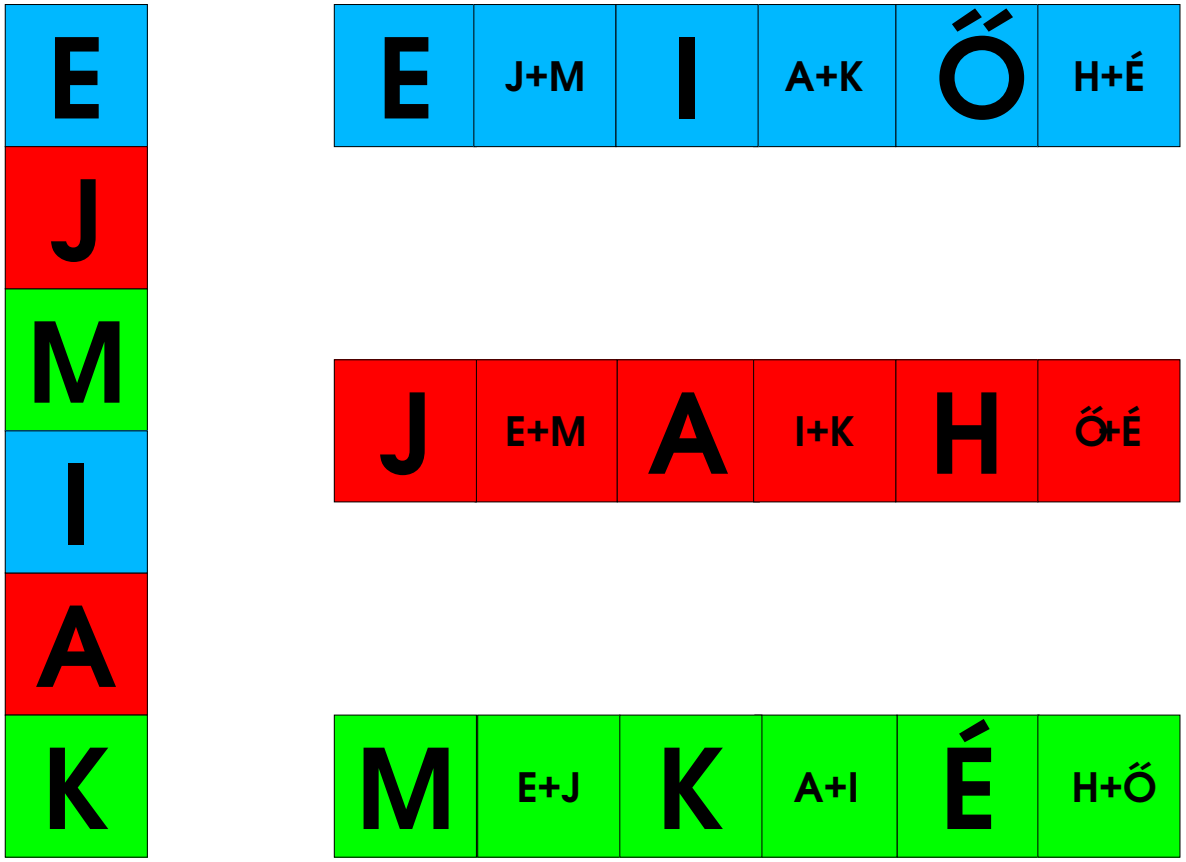
RAID10

- Megbízható és gyors
- RAID1 tömbök RAID0-ba szervezve
- Vagy fordítva
- 50% overhead, drága
- Akár a diszkek fele is tönkremehet, baj nélkül
- Gazdag ember RAID5-je

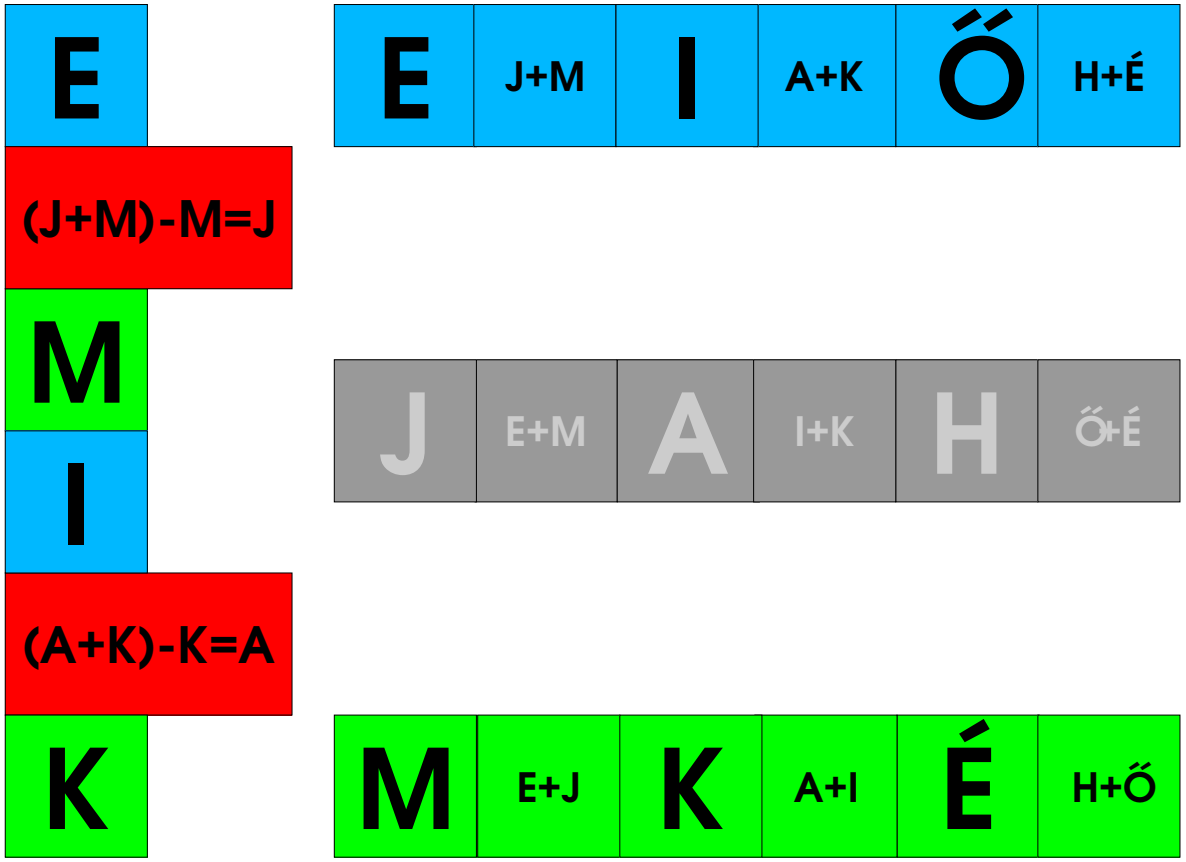
RAID5

- Szegény ember RAID10-ja
- Legalább 3 diszk
- Egy egységnyi overhead
- Elméletben lassú, az ellenőrzőösszeg miatt
- Gyakorlatban lehet gyors
- Egy diszket veszíthet csak
- Tartalékot ha lehet, mindig hagyjunk

RAID5 egészségesen



RAID5 betegen



Parancssorok

```
mdadm --create /dev/md2 --level=raid5 --raid-devices=3 \
--spare-devices=1 /dev/hda[5-8]

md2: active raid5 hda7[3] hda8[4] hda6[1] hda5[0]
      9767296 blocks level 5, 64k chunk, algorithm 2 [3/2] [UU_]
      [=====>.............]  recovery =  30.6%
(1497088/4883648) finish=8.0min speed=7044K/sec
```

RAID5 érdekességek

- Féllábúan indul, így 10%-al gyorsabb
- 5% CPU terhelés, nem nagy gond a paritás

```
md2: active raid5 hda7[2] hda8[3] hda6[1] hda5[0]
      9767296 blocks level 5, 64k chunk, algorithm 2 [3/3] [UUU]
```

Tönkretétel

```
mdadm /dev/md2 --fail /dev/hda5
```

```
md2:active raid5 hda7[2] hda8[3] hda6[1] hda5[4](F)
      9767296 blocks level 5, 64k chunk, algorithm 2 [3/2] [_UU]
      [= = > .....]  recovery =  12.0% (591744/4883648)
    finish= 10.1min speed= 7048K/sec
```

Tönkretétel

logfejléc: raid5: Disk failure on hda5, disabling device.

Operation continuing on 2 devices

logfejléc: RAID5 conf printout:

logfejléc: --- rd:3 wd:2 fd:1

logfejléc: disk 0, o:0, dev:hda5

logfejléc: disk 1, o:1, dev:hda6

logfejléc: disk 2, o:1, dev:hda7

Tönkretétel

logfejléc: RAID5 conf printout:

logfejléc: --- rd:3 wd:2 fd:1

logfejléc: disk 1, o:1, dev:hda6

logfejléc: disk 2, o:1, dev:hda7

logfejléc: RAID5 conf printout:

logfejléc: --- rd:3 wd:2 fd:1

logfejléc: disk 0, o:1, dev:hda8

logfejléc: disk 1, o:1, dev:hda6

logfejléc: disk 2, o:1, dev:hda7

logfejléc: md: syncing RAID array md2

Megsemmisítés

```
mdadm /dev/md2 --fail /dev/hda7


Personalities : [raid0] [raid1] [raid5]
md2 : active raid5 hda7[3](F) hda8[4] hda6[1] hda5[5](F)
      9767296 blocks level 5, 64k chunk, algorithm 2 [3/1] [_U_]
        resync= DELAYED
```

RAID5 katasztrófa tanulságai

- A lehető leghamarabb cseréljünk diszket
- Kell tartalék diszk
- Szinkronizálás alatt élőáldozat, ima
- Innen is van kiút, de ne építsünk rá
- A rendszeres mentést semmi sem helyettesíti
- Ne felejtsük el leállítani, mert próbálkozik
- mdadm monitor módját használjuk

RAID IDE eszközökből

- Olcsó, és egyre olcsóbb, a SATA gyors is
- Lassan otthoni gépbe is követelmény
- IDE csatornát ne osszunk meg
- Lassú eszközökkel ne keverjük
- Alaplapi IDE meghajtók beállítása

RAID hardverek

- IDE/SATA eszköz ritkán jobb, mint a szoftveres raid
- SCSI RAID eszköz lehet jó, de általában drága
- A rossz hardver teljes adatvesztést is okozhat
- A döntés nehéz, fontoljuk meg
- Egyre nagyobb diszkek, egyre gyorsabb csatolók
- Egyre távolabb a határ, ahol már komoly storage eszköz kellhet

Minimális RAID tömb

Közepes méretű RAID tömb

Pár adat Gödöllőről

Gép	Típus	Vezérlő	Diszk	Írás	Olvasás	Átlag MB/sec
sziszi	x345	qla2340/r5-8	cx700	85	127	106
sziszi	x345	qla2340/r5-4	cx700	68	120	94
sziszi	x345	fusion/swr5	10k/320	75	110	93
rserver	x345	qla2340/r5-8	cx700	67	116	92
hera	x335	fusion/swr1	15k/320	72	75	74
noc	x330	aic7892/swr1	10k/160	44	53	49
rserver	x345	fusion/swr1	10k/160	26	40	33
gaia	x235	ibm5i/r5	10k/160	32	26	29

Fizikai biztonság

- Elmélet és gyakorlat különbsége
- Diszkek halnak megy egyszerre
- Szerverszoba kiéghet, beázhat
- Jó, ha nem egy helyen vannak az adataink

Megoldási lehetőségek

- Hosszú SCSI kábel (FC)
- Storage eszközök, HA clusterek gyártótól
- Igen drága, igen jó – 9/11 kihagyás nélkül
- Szabad forrású megoldások
- Kicsit hoszabb átállás
- Kicsit több munka
- Nevetségesen olcsó

Szabad forrású praktikák

- Csak képzelet szabhat határt
- Pár ötlet: Rsync rendszeresen
- Alkalmazások replikálása
- ENBD – általános célú, erre is jó
- DRBD – kész megoldás
- heartbeat, vagy saját scriptek

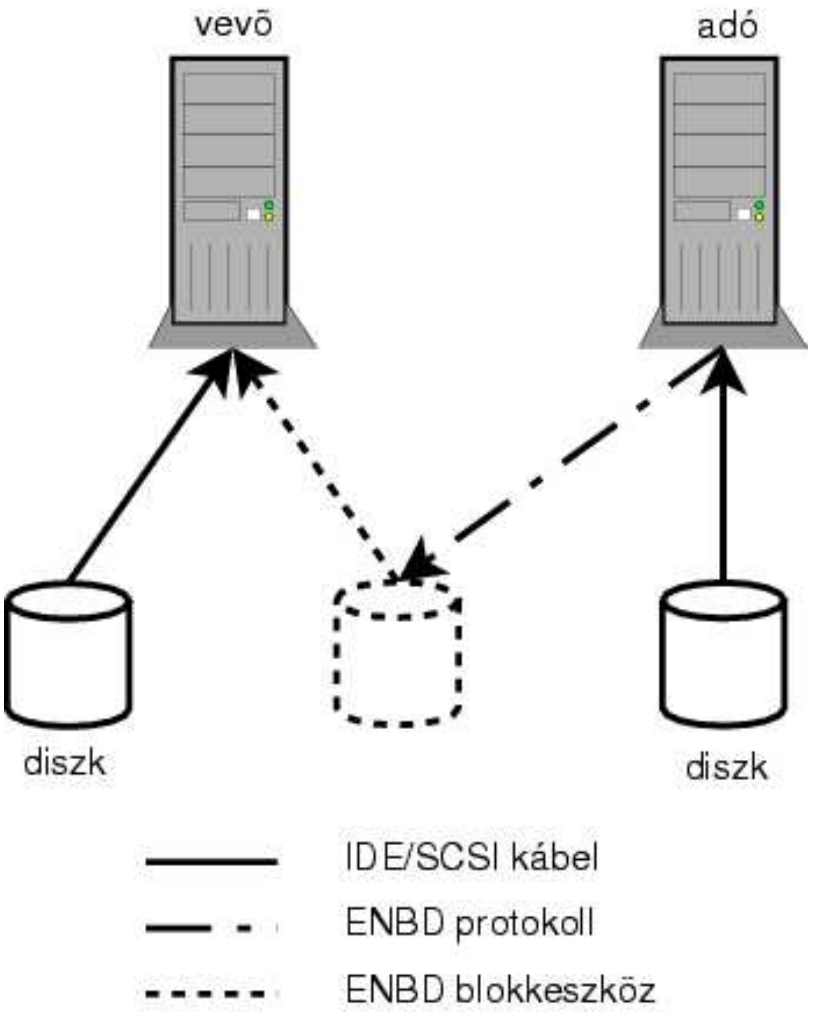
HA cluster megvalósítása

- Tartalék gép beszerzése, beállítása
- Kapcsolat a két gép között
- HA megvalósítása
- Mindegyik részfeladat önmagában több kötet
- Ezért itt csak a középsővel foglalkozom
- És ezzel is inkább csak gondolatébresztőként

Mit mondok el az ENBD-ről?

- Hozzávalók
- Telepítés/konfigurálás
- RAID1 tömb lokális és távoli blokkeszközzel
- Katasztrófa és helyreállítás
- Apróságok

Mi az ENBD?


vevõ
adó
diszk
diszk
IDE/SCSI kábel
ENBD protokoll
ENBD blokkeszköz

ENBD telepítése

- Kernel, segédprogramok fordítása
- inetd.conf, services fájlok kitöltése
- Vevőn a /dev-ben a megfelelő nodeok létrehozása
- enbd.conf kitöltése
- Adón a megfelelő erőforrás biztosítása

inetd.conf, services fájlok

```
enbd-cstatd 5051/tcp   #  NBD statd (client side)
enbd-sstatd 5052/tcp   #  NBD statd (server side)

enbd-sstatd stream tcp nowait root \
/usr/sbin/enbd-sstatd enbd-sstatd

enbd-cstatd stream tcp nowait root \
/usr/sbin/enbd-cstatd enbd-cstatd
```

Adó beállítása

- Az erőforrás lehet fájl vagy blokkeszköz
- Készüljünk föl időnként jelentős IO-ra, ha más feladatai is vannak a gépnek
- enbd.conf:

```
server elso 1234 eroforras  -b 512
```

- /var/state/nbd könyvtár
- /etc/init.d/enbd start

Adó ellenőrzése

- portmap – adat és manager portok (1234-1237 és 5051-5052) nyitva legyenek
- ps axf – megfelelő számú processz
- syslogban bőséges locsogás
- Az adó dolga egyszerű, baj ritkán van vele

Vevő beállítása

- /dev/nd[a-p], major 43
- enbd.conf:

client elso /dev/nda ado_ip 1234 -n 4 -b 512

- /var/state/nbd könyvtár
- /etc/init.d/enbd start

Vevő ellenőrzése

- /proc/nbdstatus – threadokra elosztott IO-statisztika, állapotjelzés, egérmozinak is jó
- ps axf – ha valamelyik processz D állapotú, akkor bukta
- Eleinte gyakoroljunk, mert sokszor kell hidegindítani...

Baj esetén mit tegyünk

- -e opció fontossága – jelentse-e a hibát fölfelé, vagy várjon
- Az inetdbe telepített kicsi programok újra tudják éleszteni a kapcsolatot, ha valamelyik fél hibájából megszakadt
- Általában nem kell, de ha minden kötél szakad:
- kill -SIGUSR1 <pid>
- echo "0" > /proc/nbdinfo

RAID1 tömb ENBD-vel

```
mdadm –create /dev/md3 –level= = raid1 –raid-devices= 2
   /dev/hda5 /dev/nda


md3 : active raid1 nda[1] hda5[0]

      4883648 blocks [2/2] [UU]

      [= = = = = = = = = = = = > .........]  resync =  56.2%
   (2747712/4883648)finish= 6.6min speed= 5324K/sec
```

Teendők, ha az adó elgyengül

- Gatyába kell rázni az adót
- A sikeres újraindulást követően a segítőprogramok újraépítik a kapcsolatot, a kernel újraszinkronizálja a tömböket
- Ehhez kell a -e opció
- fr1 driverrel sokkal kevesebbet szinkronizál

Teendők, ha a vevő elgyengül

- Eldönteni, mi gyorsabb – gatyába rázni a vevőt, vagy elindítani az adót, mint tartalékszervert.
- Ha automatika van, nincs választási lehetőség
- Ha a vevőt rázzuk gatyába, akkor lásd az adónál felsorolt tudnivalókat
- Célszerűbb elindítani az adót, hátha nem egyszerű baja van a vevőnek.

Az adó, mint tartalékszerver

```
#  file raid1.bin

raid1.bin: Linux rev 1.0 ext3 filesystem data (needs journal
    recovery)

fifi:~ #  losetup /dev/loop0 raid1.bin

fifi:~ #  fsck.ext3 /dev/loop0

e2fsck 1.35 (28-Feb-2004)

/dev/loop0: recovering journal

/dev/loop0: clean, 72/611648 files, 27975/1220912 blocks
```

Eltitkolt apróságok

- Az igazi HA-nak ez csak karikatúrája volt
- Erre a feladatra a DRDB való
- Az ENBD sokkal többmindenre képes: multipath, fr1 és fr5 hálózatra optimalizált raid driverek (sajnos csak 2.4-es kernelekre), titkosított adatfolyamok, readonly, asszimmetrikusan írható kapcsolatok, stb.
- Az ENBD-vel létrehozott RAID1 tömböt sync opcióval mountoljuk, jobb a békesség

További információk

- http://www.tldp.org/HOWTO/Software-RAID-HOWTO.html
- http://www.it.uc3m.es/~ptb/nbd/
- http://www.drbd.org/
- http://gergely.tomka.hu/raidenbd.pdf (félkész)
- Tomka.Gergely@ih.szie.hu

Kérdések, jótanácsok?

Csak őszintén, a linux-flame levelezőlistán úgyis megkapom a magamét