Orosz rulett

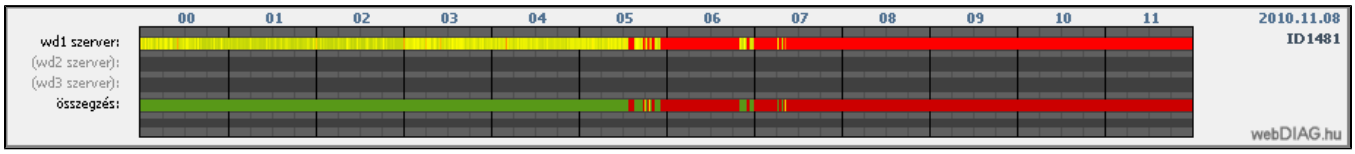
Vannak emberek, akik szeretnek **veszélyesen élni**.

Ezek az emberek akkor érzik magukat jól, ha dl az orrukból is az adrenalin. Elindulnak a szlovák hegyek felé nyári gumival a legnagyobb hózivatarban, esetleg különféle extrém sportokat znek, rendszeresen átfutnak a 2x3 sávos úton a zebrától ötven méterre, vagy bódult állapotban orosz rulettet játszanak hatlövet pisztollyal és öt golyóval... illetve **EXPERIMENTAL** fájlrendszert *használnak produktív szerveren*.

> *WARNING: ZFS is considered to be an experimental feature in FreeBSD.*
> *ZFS filesystem version 6*
> *ZFS storage pool version 6*

Ez még önmagában nem lenne baj, de a FreeBSD operációs rendszerhez 2009 novemberében – igen, alig egy éve – adtak ki stabilnak minsített ZFS alrendszert, de minek frissítsen az ember, amikor **mködik**... jó-jó, néha újra kellett indítani, mielttt az *Adaptive Read Cache* felzabálta az összes memóriát, bár amikor összeért a kernel memóriája és az ARC memória, akkor dobott egy kövér *kernel panic*-ot és újraindult – szóval ha az ember nem volt szemfüles, nem is kellett ilyenkor újraindítani, újraindult magától... 🙂

Hétf 6:05



Egy borús hétf reggel az ember arra az SMS-re ébred, amelyet a www.netdiag.hu küld, ha 10 percig nem éri el az általa figyelt szolgáltatást, jelen esetben ez a szolgáltatás a http://www.javaforum.hu/javaforum volt. A netdiag azt látta, hogy a szerver lassan – 12 másodpercen túl – válaszolt, de amúgy élt a szerencsétlen. Összekapartam a rendelkezésre álló ismereteimet, a percek alatt reagáló konzolt figyeltem, miközben az álmosság és az adrenalin küzdelmének végét vártam a fejemben, ám a szerver csak nem kapta össze magát.

Hétf 7:28

Szk másfél óra agónia után úgy döntöttem, hogy a megoldás legyen a menedzsment konzolról egy jól irányzott **reset**. Ezt követte három perc újabb extra adrenalin adag, amíg a szerver feléledt – egy Sun Fire x2100M2 nem a túl gyors indulásáról híres. Szerver visszajött a menedzsment konzol szerint, aztán a *ping* kérésekre is válaszolni kezdett, de nem válaszolt SSH porton, a menedzsment felületen megnéztem, mi is látszik a szöveges konzolon, és megfagyott bennem a vér: az oprendszer nem találta a dolgait, amelyeket a megszokott helyeken keresett.

Hétf 7:38

Két lehetségem van: vagy a lassú és körülményes menedzsment konzolról valamit tenni, vagy munkába menet befutni a szerverhez és ott helyben megoldani a problémát. Logikus az utóbbi, hiszen dolgozni kell, ezért autóba ülve sebesen megindultam a szerverterem felé.

> *Amikor az ember kocsiba ül Óbudán, de nem gondolja át, hogy a Megyeri-híd kvázi le van zárva, a Margit-híd egy éve teljesen le van zárva, és egyedüli út Budáról Pestre az Árpád-híd – az rossz döntésnek vehet. Fleg, amikor ezt két baleset is súlyosbítja, amelyek miatt az 1-es villamos se jár, de a dugóban ülve már kés bánat, hogy a BKV gyorsabb lenne...*

Hétf 10:20

Két óra alatt a szerverhez érve kiderül, hogy **nagy a baj**, részben el tudott indulni, de a fájlrendszere fura képet mutat:

```
  pool: bpool
 state: FAULTED
status: The pool metadata is corrupted and the pool cannot be opened.
action: Destroy and re-create the pool from a backup source.
   see: http://www.sun.com/msg/ZFS-8000-CS
 scrub: none requested
config:

        NAME        STATE     READ WRITE CKSUM
        bpool       FAULTED      0     0     6  corrupted data
          da0s2     ONLINE       0     0     6

  pool: dpool
 state: FAULTED
status: The pool metadata is corrupted and the pool cannot be opened.
action: Destroy and re-create the pool from a backup source.
   see: http://www.sun.com/msg/ZFS-8000-CS
 scrub: none requested
config:

        NAME        STATE     READ WRITE CKSUM
        dpool       FAULTED      0     0     6  corrupted data
          da0s4     ONLINE       0     0     6
```

ZFS hiba, amitl mindig is féltem... a **FAULTED** pool természetesen nem tartalmaz egy darab **dataset**-et sem, maga a pool se életképes, nem tudja magáról, hogy mekkora és mennyi hely van benne, így nincs más teend, mint rescue módban elindítani a szervert, hogy alapszinten mködjön és távolról is el lehessen érni, s az ember befut a munkahelyére, hiszen dolgozni is kell (sajnos).

Persze néhány kézenfekv javítást azért megpróbáltam helyben, de nem igazán segítettek a helyzeten, így a döntés az lett, hogy ne tegyek rosszabbat, mint ami van, legyen mentés az aktuális állapotról, mielttt komolyabb beavatkozás végzek.

Hétf 12 óra

A szerver hat órája áll, a fájlrendszereinek nagy része nem érhet el, a 320GBájt kapacitásból a 2G alaprendszer megvan, a rescue rendszer megvan, a 8GBájt méret rendszer pool **FAULTED**, a 280G méret adat pool **FAULTED** – vagyis minden, ami adat és bvebb rendszerszolgáltatás, az elérhetetlen vagy elveszett.

Két lehetség van:

- az ember megpróbálja a heti és a napi mentésbl visszaállítani a rendszert, de ekkor elveszik néhol egy heti adat, és egy napi munka a teljes újratelepítés – nyilván a 7.x helyett 8.*friss* került volna a szerverre, majd az adatokat is helyre kellene állítani a mentésbl
- az ember megpróbálkozik a ZFS pool *helyreállításával*, ehhez le kell tölteni a 8GBájt méret kicsi pool teljes tartalmát, majd azon játszadozni, mivel a 7.x FreeBSD-ben nincs normális eszköz erre a célra

Hétf 15 óra

A 8GBájt méret pool kb. 1,5MBájt/s sebességgel letöltve közel két órán át csordogált. Ezek után felgyorsulnak az események, a ZFS-FUSE projekt segítségével a pool – mint fájl – elvileg feltámasztható a józan ész szerint, hiszen komoly trauma nem érte, az adatoknak meg kell lennie valahol a pool mélyén, s a ZFS-FUSE újabb, mint a FreeBSD ZFS alrendszer, hiszen a ZFS pool verzió ez esetben nem **6**, hanem **23** – csak okosabb.

És In, a kis méret pool másodpercek alatt magához tér:

```
authglaptop:/media/SAMSUNG/backup/test # zpool import -f -F -a -d .
Pool bpool returned to its state as of 2010-11-08.
Discarded approximately 14 seconds of transactions.
cannot mount '/tmp': directory is not empty
cannot mount '/usr': directory is not empty
cannot mount '/usr/local': directory is not empty
cannot mount '/usr/src': directory is not empty
cannot mount '/var': directory is not empty
authglaptop:/media/SAMSUNG/backup/test # zpool status
  pool: bpool
 state: ONLINE
status: The pool is formatted using an older on-disk format.  The pool can
        still be used, but some features are unavailable.
action: Upgrade the pool using 'zpool upgrade'.  Once this is done, the
        pool will no longer be accessible on older software versions.
 scrub: none requested
config:

        NAME                                STATE     READ WRITE CKSUM
        bpool                               ONLINE       0     0     0
          /media/SAMSUNG/backup/test/bpool  ONLINE       0     0     0

errors: No known data errors
```

Megnézve kiderül, hogy minden adat megvan, de úgy gondoltam, sikálja át a ZFS-FUSE, hátha vannak hibák:

```
authglaptop:/media/SAMSUNG/backup/test # zpool scrub bpool
authglaptop:/media/SAMSUNG/backup/test # zpool status
  pool: bpool
 state: ONLINE
status: The pool is formatted using an older on-disk format.  The pool can
        still be used, but some features are unavailable.
action: Upgrade the pool using 'zpool upgrade'.  Once this is done, the
        pool will no longer be accessible on older software versions.
 scrub: scrub in progress for 0h0m, 0.00% done, 26h2m to go
config:

        NAME                                STATE     READ WRITE CKSUM
        bpool                               ONLINE       0     0     0
          /media/SAMSUNG/backup/test/bpool  ONLINE       0     0     0

errors: No known data errors
```

Hétf 16 óra

Az átvizsgálás megtörtént, 50 perc alatt talált egy checksum hibát, ami nem vészes:

```
authglaptop:/media/SAMSUNG/backup/test # zpool status
  pool: bpool
 state: ONLINE
status: One or more devices has experienced an unrecoverable error.  An
        attempt was made to correct the error.  Applications are unaffected.
action: Determine if the device needs to be replaced, and clear the errors
        using 'zpool clear' or replace the device with 'zpool replace'.
   see: http://www.sun.com/msg/ZFS-8000-9P
 scrub: scrub completed after 0h51m with 0 errors on Mon Nov  8 16:25:40 2010
config:

        NAME                                STATE     READ WRITE CKSUM
        bpool                               ONLINE       0     0     0
          /media/SAMSUNG/backup/test/bpool  ONLINE       0     0     1  4.50K repaired

errors: No known data errors
```

Megkezddik a tartalom hálózaton való visszatöltése:

```
authglaptop:/media/SAMSUNG/backup/test # dd if=bpool | ssh root@javaforum.hu dd of=/dev/da0s2d
```

Hétf 18 óra

Újabb két óra alatt felmásolódott a megjavított pool, a szerveren is nézzük meg, hogy jó-e:

```
This module (opensolaris) contains code covered by the
Common Development and Distribution License (CDDL)
see http://opensolaris.org/os/licensing/opensolaris_license/
WARNING: ZFS is considered to be an experimental feature in FreeBSD.
ZFS filesystem version 6
ZFS storage pool version 6
freebsd# zpool import -a
cannot import 'bpool': pool may be in use from other system, it was last accessed by authglaptop (hostid:
0x7f0200) on Mon Nov  8 16:26:04 2010
use '-f' to import anyway
freebsd# zpool import -f -a
freebsd# zpool status
  pool: bpool
 state: ONLINE
 scrub: none requested
config:

        NAME        STATE     READ WRITE CKSUM
        bpool       ONLINE       0     0     0
          da0s2d    ONLINE       0     0     0

errors: No known data errors
```

Öröm és bódottág, a módszer megvan, már csak a 8G helyett 280G méretre kell kidolgozni a részleteket: elkezdtem a hordozható 320GBájt kapacitású HDD-rl lementeni a rajta lév dolgokat, ugyanis rá kell majd másolnom a 280GBájt méret második pool tartalmát, mint fájlt. Ehhez olyan fájlrendszer kell, amelyik képes egy fájlként 280G méretet kezelni, a FAT erre nem képes, NTFS-t nem eszik a FreeBSD ebben az állapotában, marad az ext2fs, megformázom gyorsan a HDD-t.

Hétf 19 óra

A rendszer fájlrendszerek megvannak és mködnek, a normál módon el tud indulni a szerver, már csak a 280G méret adatokkal megrakott pool kell, ehhez hazafelé menet a már említett 320G kapacitású USB HDD-t csatlakoztatom a szerverhez és elkezdem a másolást, amikor nem várt akadályba ütközök:

```
# mount -t ext2fs /dev/da1s1 /mnt
# cd /mnt
cd: not a directory: /mnt
```

Hopszi, ez így nem lesz jó. Fél óra keresgélés után találtam egy fórumot, ahol kiderült, hogy a 7.x FreeBSD csak és kizárólag a 128 alapszámú inode táblázatot képes kezelni, ezért újra kell formázni a hordozható HDD-t, persze ezt a Linux-os laptopon, mivel a FreeBSD-n nincs telepítve ext2fs fájlrendszer kezeléséhez program, telepíteni meg nem tudok:

```
# mkfs.ext2 -I 128 /dev/sdb1
# tune2fs -m 0 /dev/sdb1
```

Ezzel már mködik a dolog, elindul a másolás, ami várhatóan éjfél után fejezdik be.

Hétf kb. 23 óra

A szerver újraindul "mindössze" 240G másolása után, viszont az USB HDD-t elbb ismerte fel, mint a RAID tömböt, így nem tud elindulni, korán reggel legalábbis ez a kép fogad. *Nagyon remélem, hogy nincs hardverhiba.*

Kedd 6 óra

Tanulva a tegnapi dugóból HÉV + villamos kombinációval indultam a szerver felé, hogy az USB HDD-t kihúzzam, újraindítsam a gépet majd visszadugjam a HDD-t.

> *A reggeli dugók 7 óra körül kezddnek, ezért reggel hatkor a kutya se jár az utakon, így mehettem volna autóval is, elbb értem volna oda és kényelmesebben is, de sajnos ebbe nem gondoltam bele: ismét egy rossz döntés.*

Kedd 7 óra

A szerverteremben más ismersként fogadnak, bár a személyi igazolványt továbbra is elkéri az r: "Hátha álarcot húztam..."

A HDD kihúzása, majd újraindítás után a szerver él és virul, most rescue módban elindítva másolja a 280GBájtos pool-t az USB HDD-re, én pedig befutok a munkahelyemre, mert meglep módon kedden is kell dolgozni.

Kedd 11 óra

A másolás befejezdött, közvetlen másolással 3,5 óra alatt megvolt a kb. 280GBájt megmozgatása. Egy laptoppal a kezemben visszamentem a szerverhez, a ZFS-FUSE – a bpool-hoz hasonlóan – egy megtekintéssel megjavította a dpool-t is, USB HDD ment vissza a szerverbe, a másolást elindítottam, és mentem vissza dolgozni.

Kedd 15 óra

A visszamásolás is befejezdött, visszamentem a szerverhez, rescue módban már be tudta tölteni mindkét pool adatait, minden adat meglett. Eddig benntartott leveg kifúj. Észreveszem, hogy esik. Eddig nem tnt fel. 🙂

Kedd 16:03

A normál mködés visszaállt, minden lényeges szolgáltatás fut és elérhet.

Konzekvencia

Legyen mentés. Egy mentés nem mentés, legyen még egy mentés valahol. Gondolkodj, mieltt cselekszel. Ne ítélj túl hamar, lehet, hogy nem annyira rossz a helyzet, mint elsre látszik. Legyen rugalmas munkaidd, ha van maszek munkád. Legyen olyan fnököd, aki szintén maszekol, szintén szervereket üzemeltet és megérti a helyezetet... 🙂

A ZFS fájlrendszerrel továbbra is meg vagyok elégedve, hiszen semmi adat nem veszett el, a hiba a FreeBSD 7.x **EXPERIMENTAL** ZFS alrendszerében volt – két héten belül frissítek 8.x FreeBSD-re, amely megoldja a ZFS probémák jelents részét... remélem. 🙂