Jak (ne)vyzrát na inode

Aneb co vše může admina po letech překvapit

Úkol: vytvořit adresář s inode 50716024

- HA řešení pro NFS: dva servery + jedno diskové pole
- /mnt/export/nfs4/home - na lokálním disku do nějž se připojuje diskové pole
- NFS export je nastaven jako `/mnt/export/nfs4/` s `fsid=0`
- klienti používají fsid + inode pro identifikaci cesty, kterou chtějí použít

Problém:

- `stor_A# stat --printf=%i /mnt/export/nfs4`
  `50716024`
- `stor_A# stat --printf=%i /mnt/export/nfs4`
  `218129930`

Klient při `cd /home/` dostane jako identifikaci cesty číslo inodu (např. `50716024`), pokud ale v rámci HA přehodíme export na druhý server při dotazu na objekt s inode `50716024` dostane chybu ⇒ většinou nutný restart klienta.

XFS: loterie

1. vytvářej adresáře dokud to jde a testuj zda jsi se netrefil (cca 20 řádků v C)
   - `xfs_info /mnt/export/nfs4/home`
     - agcount=16 (postupně se střídají, adresář dostane “pseudonáhodný” inode)
     - inode s nejvyšším číslem: $16*2^{24}$ = 268.435.456 (velikost jedné skupiny je 2^{24}, viz doc)
   - `df -i /mnt/export/nfs4/home`
     - max. inodes: 62.881.664
2. vytvoř adresář jen ve správné alokační skupině (pro `50716024` je to 4 skupina od 50.331.648 do 67.108.864)
   - adresář zabírá více místa v inode alokační skupině než soubor
   - adresáře: ~50% využití inode alokační skupiny = 100% využití místa alokační skupiny
3. vytvoříme soubory (jsou menší) a ve vhodný okamžik začneme vyrábět adresáře
   - soubory vznikají jinak než adresáře, po smazání souboru (abychom se vyhli bodu 1) se vytvoří soubor se stejným inodem

Dokumentace: http://oss.sgi.com/projects/xfs/papers/xfs_filesystem_structure.pdf

Ext3: kávičku?

Vytvoříme soubor, který naformátujeme a připojíme jej do `/mnt/exports/nfs4`, bude tedy obsahovat jediný adresář, ale se správným číslem inodu:

- spočítat/odhadnout vhodnou velikost disku pro 51M inodů v ext3 (13GB)
- správně naformátovat, tak aby byl co nejmenší

```
mkfs.ext3 -F -m 0 -i 1024 -I 128 -N 51000000 /mnt/export.img
```

A hledáme:

1. tvoříme adresáře dokud nevytvoříme ten správný
   - adresář zabírá příliš mnoho místa, po cca 6M adresářů je plný disk (0% free)
2. vytvoříme soubory, smažeme soubor na `50716024` a vytvoříme adresář na jediném volném inodu
   - v jednom adresáři smí být až 65.535 souborů
3. jako bod 2., ale rozhazujeme soubory do více adresářů
   - Máme ho

Závěr

- nezapomenout použít `find / -i 50716024`
- shell - extrémně pomalé, nutno použít C
- XFS i Ext3 vytváří inody dynamicky
  - XFS - neumí vytvořit na konkrétním inode
  - Ext3 - existuje debugfs, ale nelze vytvořit adresář, s obtížemi jen soubor
- nebude NFS klientům vadit změna typu souborového systému nebo zařízení?
  - další souborové systémy jsme netestovali, nemělo by to přínos, když to neumíme na XFS

Jiná řešení:

- neexportovat lokální kořen - vysvětlete to 500+ NFS klientům bez restartu
- *`stor_A# dd if=/dev/sda1 | ssh stor_B 'dd of=/dev/sda1'`*
  - ze živého na živý systém ;-(
  - chce to hodně štěstí, nemusí pomoci ani `xfs_repair`