ZFS a ZFS on Linux

Milan Beneš
milan.benes@web4ce.cz

files.benesovi.eu/zfsonlinux

ZFS – ZahlendarstellungFormularSystem

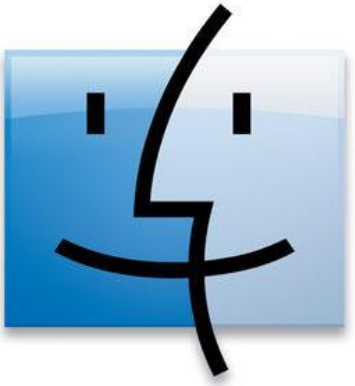
- Přelom 19. a 20. století
- Pokus českého génia Járy da Cimrmana o digitalizaci přebujelé c.k. byrokracie

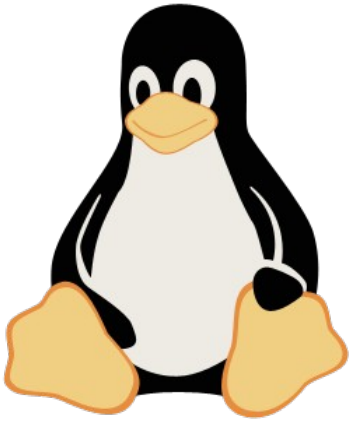
ZFS – Zettabyte File System

- 2001 - Jeff Bonwick, Matt Ahrens – Sun Microsystems
- 1. verze v roce 2005 v OpenSolaris
- Motivace: end-to-end integrita dat
- 128bit FS
- Volume manager
- RAID

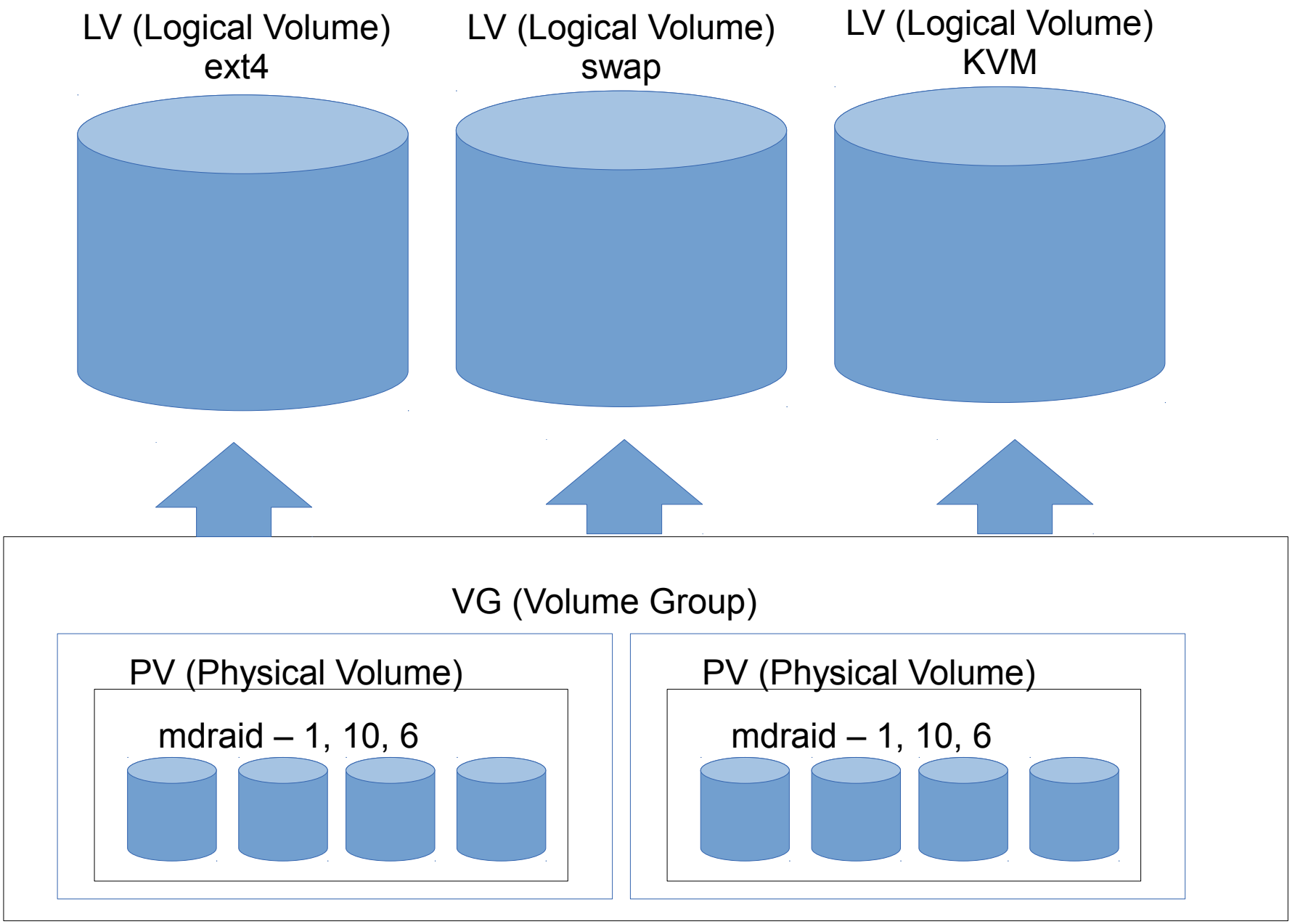
Kde ZFS najdete?

SmartOS





Linux – současné paradigma



Verze ZFS a zpool

ZFS:

- Poslední verze 6
- Poslední FOSS verze 5

ZPOOL:

- Poslední verze 34
- Poslední FOSS verze 28
  - Verze 5000, feature flags

vdev

- V LVM je ekvivalentem PV
- Vdev lze rozšířit pouze vertikálně, výjimkou je přidání disku do mirroru

vdev -pokračování

- Disk (výchozí) – fyzické disky (může být i zvol)
- File – soubor (cesta k existujícímu souboru)
- Mirror – RAID 1 mirror
- RAID-Z1/Z2/Z3 – podobné RAID 5/6/7
- Spare – hot spare disk přiřazený zpoolu
- Cache – zařízení přiřazené k zpoolu sloužící jako L2ARC
- Log – samostatné zařízení sloužící jako ZIL

zpool

- Množina vdevů
- Je možné vertikální rozšíření (skrze vdev) i horizontální (přidáním dalšího vdevu).
- V LVM je ekvivalent volume group
- Zpool stripuje přes všechny vdevy – tj. nutnost redundance na úrovni vdevů
- (Problematika AF disků)

Příklad zpoolu

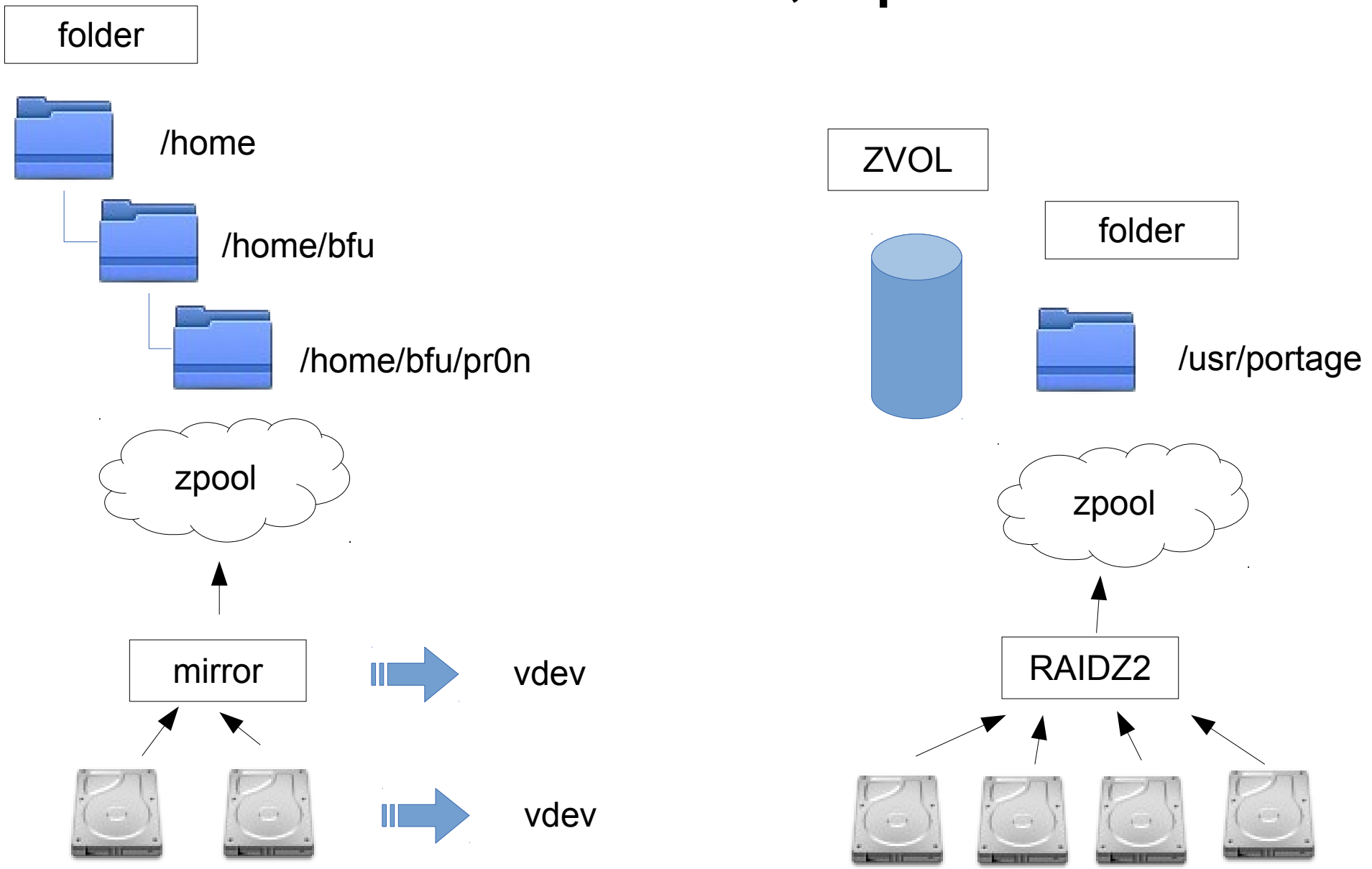
```
                           capacity     operations    bandwidth
pool                       alloc   free   read  write   read  write
-------------------------  -----  -----  -----  -----  -----  -----
data1                       3,52T  1,92T      0      8  2,65K   139K
  raidz2                    3,52T  1,92T      0      8  2,65K   139K
    ata-ST31000340NS_9QJ6WSXV     -      -      0      2     1K  45,8K
    ata-ST31000340NS_9QJ6N831     -      -      0      2    446  45,6K
    ata-ST31000340NS_9QJ59VD1     -      -      0      2    687  45,7K
    ata-ST31000340NS_9QJ7VAAY     -      -      0      2   1007  45,8K
    ata-ST31000340NS_9QJ27VSN     -      -      0      2    307  45,6K
    ata-ST31000340NS_9QJ6Z6YY     -      -      0      2    521  45,7K
logs                            -      -      -      -      -      -
  mirror                    1,57M  3,72G      0      0      0     40
    sda3                       -      -      0      0      0     40
    sdb3                       -      -      0      0      0     40
cache                          -      -      -      -      -      -
  sda4                      41,0G    16M      0      0     22  57,4K
  sdb4                      41,0G    16M      0      0     15  57,3K
-------------------------  -----  -----  -----  -----  -----  -----
```

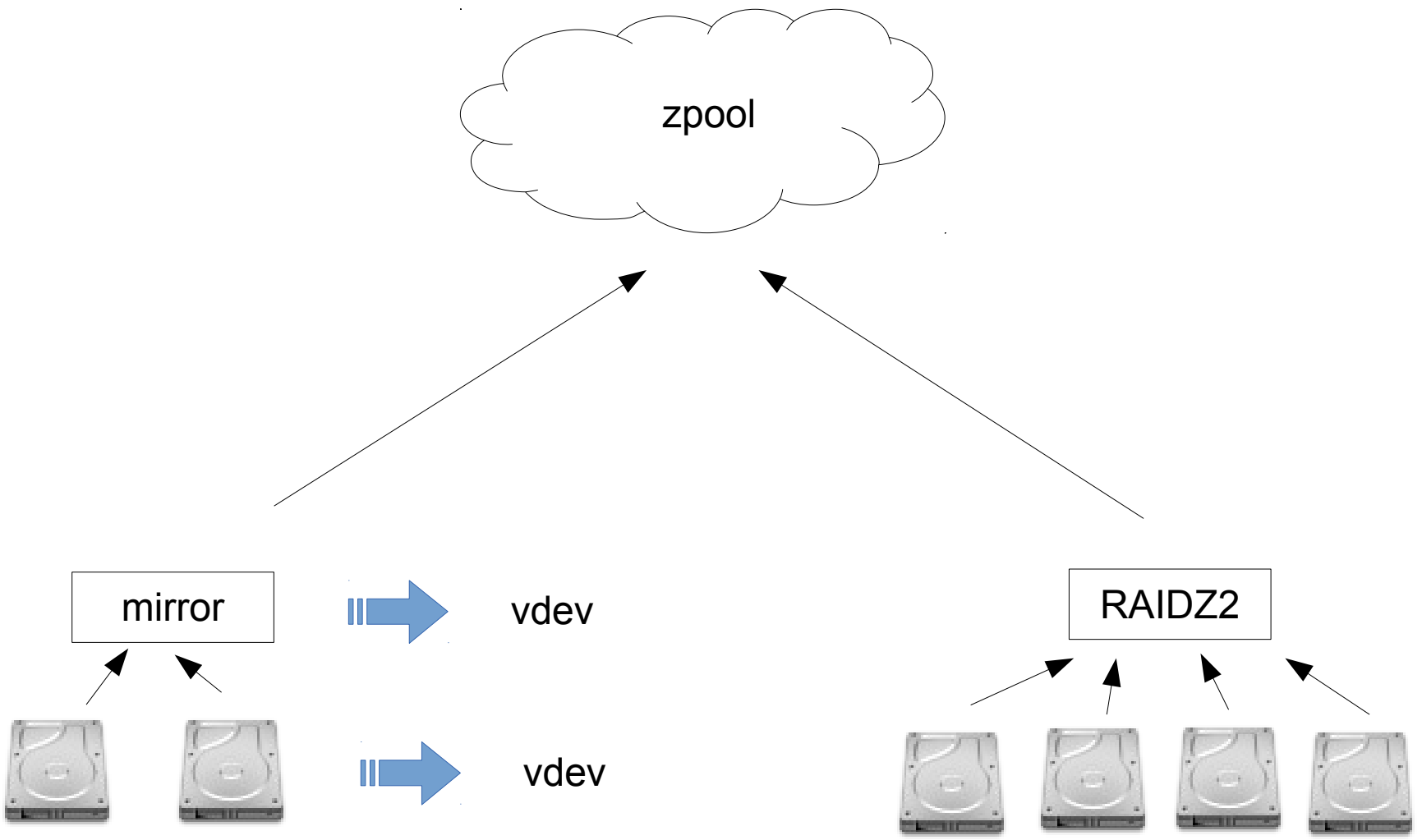
zvols, folders

- Zvol – volume, blokové zařízení
- V LVM je ekvivalentem logical volume

- Folder – samotná instance filesystému
- Z folderů můžeme tvořit hierarchii
- Delegace, qoutas, share, mount, options

ZFS – vdev, zpool

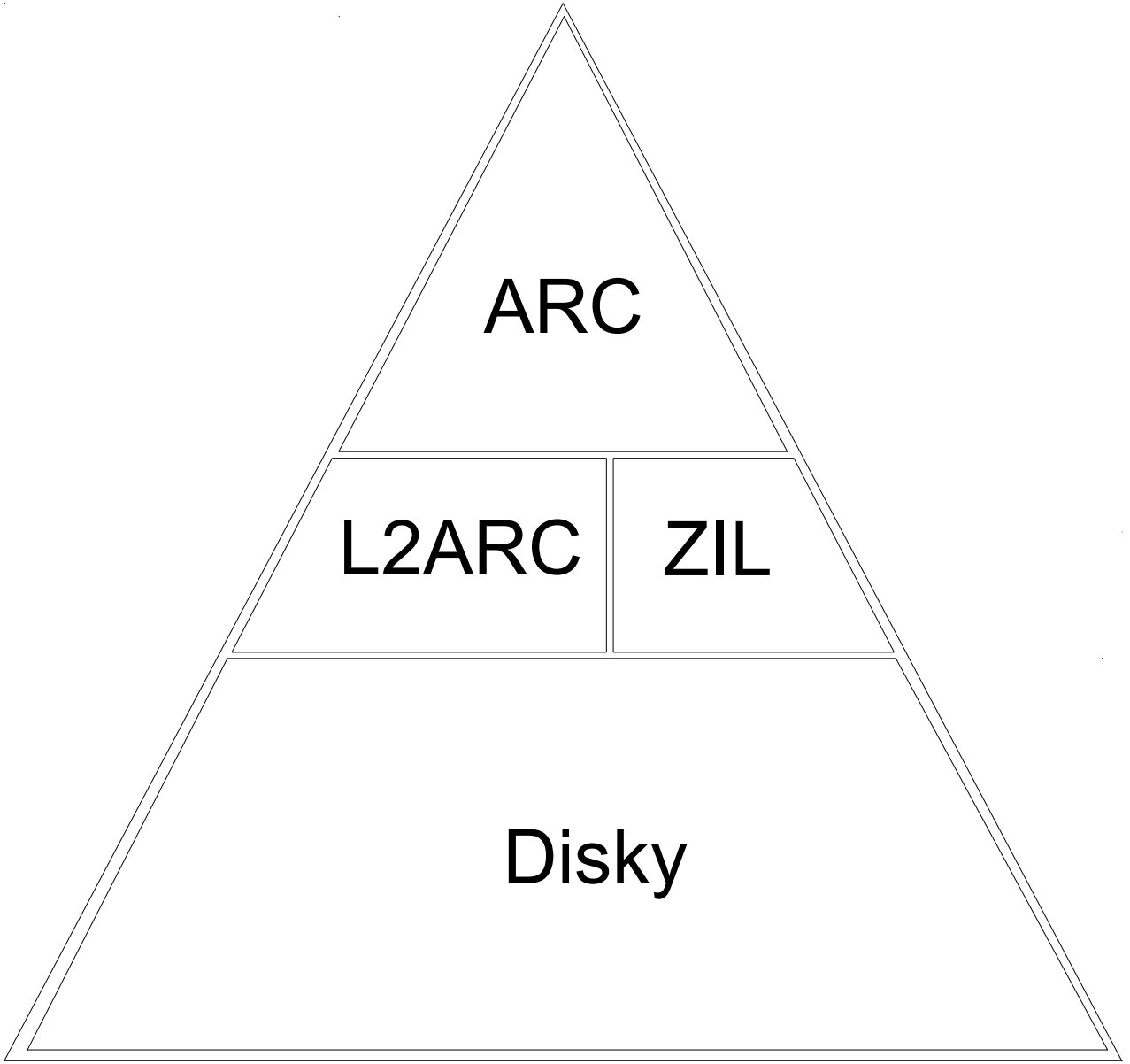



zpool
mirror
vdev
vdev
RAIDZ2

Vlastnosti ZFS

- COW
- Snapshoty, klony
- Send / receive
- Deduplikace
- Komprese
- Hybridní storage

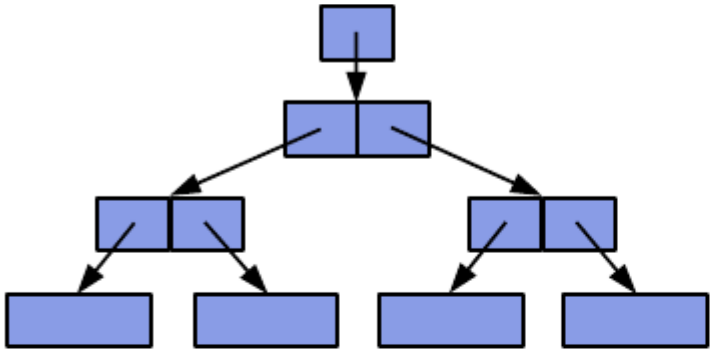
Hybridní storage



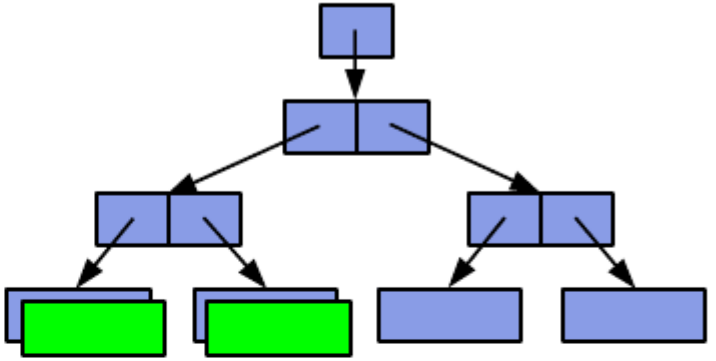
- ARC – adaptive replacement cache, MRU, MFU, ghost MRU, ghost MFU, default 7/8 RAM
- L2ARC – bloky, které běžně vypadnou z ARC, se přesunují do L2ARC
  - Adresování 50GB L2ARC zabírá ~ 1GB RAM
- ZIL – žurnál, používá se pro zajištění synchronicity zápisu, prakticky převádí synchroní zápis na asynchronní

COW, snapshoty

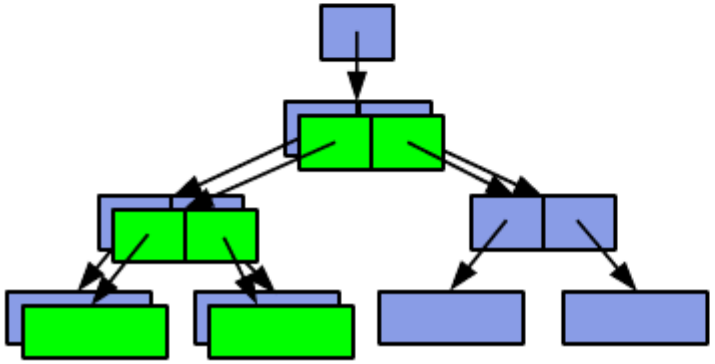
1. Výchozí stav



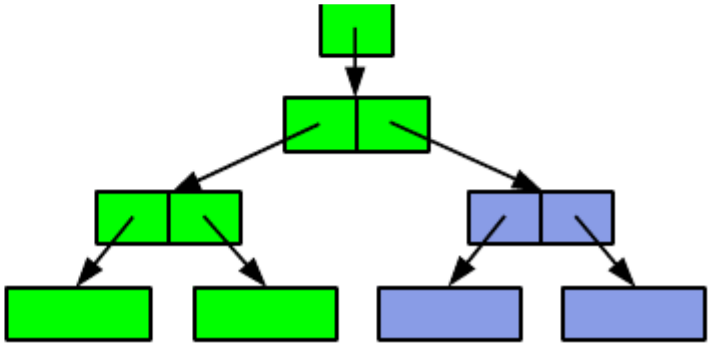
2. COW data



3. COW metadata



4. Přepis überblocku (atomický), free



ZFS na Linuxu

- FUSE implementace
- KQ Infotech
- ZFS on Linux
  - Arch Linux, Debian, Fedora, Funtoo, Gentoo, RHEL / CentOS / SL, Sabayon, SprezzOS, Ubuntu
  - Aktuální verze 0.6.1, 0.6.2

ZFS on Linux

- http://zfsonlinux.org/
- https://github.com/zfsonlinux
- Brian Behlendorf, LLNL
- LUSTRE

ZFS on Linux

- CDDL vs. GPL
- Integrace do kernelu
- / na ZFS
- GRUB2 vs. UEFI
- Nastavení parametrů zfs - /sys/module/zfs/parameters/
- Statistiky /proc/spl/kstat/zfs/

- http://indico.cern.ch/getFile.py/access?contribId=3&sessionId=0&resId=1&materialId=paper&confId=13797
- http://dtrace.org/blogs/brendan/
- http://www.c0t0d0s0.org/
- http://www.solarisinternals.com/wiki/index.php/ZFS_Best_Practices_Guide
- http://www.solarisinternals.com/wiki/index.php/ZFS_Evil_Tuning_Guide
- https://blogs.oracle.com/ahl/entry/double_parity_raid_z