BTRFS

(butter FS , b-tree FS)

Miroslav Brožek
ZČU-FAV

7.12.2011

BTRFS

- B-tree architektura
- Copy –on-write
- Oracle 2007
- GPL licence

https://btrfs.wiki.kernel.org

Stav vývoje

- Stále „alfa“ (nestabilní)
- Fsck chybí (snad v Kernel 3.2)
- Scrubbing

- Potvrzeno a započato
  - Rychlý offline fsck
  - IO priority
  - RAID-5, RAID-6
  - Online RAID konfigurace
  - Alternativní checksum algoritmy
  - Time slider

- Nepotvrzeno
  - Online fsck
  - NFS
  - Hybrid storage
  - Online připojení disku
  - Obnova skupiny bloku

BTRFS

- **Výhody**
  - Max velikost souboru 2^{64} = 16 EiB
  - Efektivní ukládání malých souborů
  - Efekt.indexování adresářů
  - Dynamicky alokované inody
  - Zapisovatelné snapshoty
  - Podsvazky
  - Vestavěný RAID 0,1,10
  - Checksum – několik algoritmů
  - SSD optimalizace
  - Online grow, schrink
  - CRFS
  - Rychlý offline fsck
- **Nevýhody**
  - Nestabilní
  - Mladý fs
  - Možná ztráta dat
  - Fsck zatím neimplementováno
  - Read-only snapshoty ve vývoji
- **Limity**
  - Max velikost souboru 16 EiB
  - Max počet souborů 264
  - Max délka názvu souboru 255 bajtů (délka jména)
  - Max velikost svazku 16 EiB

Instalace, distribuce

1. Stažení zdrojových kódů + btrfs-progs, rozbalení a kompilace, naformátování disku
2. Využít distribuci s implicitní podporou btrfs – snazší ;)

- Distribuce
  - Fedora 16
  - OpenSuse od 11.3
  - Ubuntu od 10.10
- Formát
  - mkfs.btrfs /pokus/mydisks/btrfs1
  - mkfs.btrfs /pokus/mydisks/btrfs2
  - Možnost sloučení
- Návody
  - https://wiki.archlinux.org/index.php/Installing_on_Btrfs_root
  - http://www.linuxbsdos.com/2010/11/13/how-to-install-linux-mint-10-on-a-btrfs-file-system/
  - http://www.linuxbsdos.com/2011/05/30/how-to-install-linux-mint-11-on-a-btrfs-file-system/
  - http://www.linuxbsdos.com/2011/11/16/install-fedora-16-on-an-encrypted-btrfs-file-system/
  - http://www.youtube.com/watch?v=g7BgNBhv7jU

Struktura

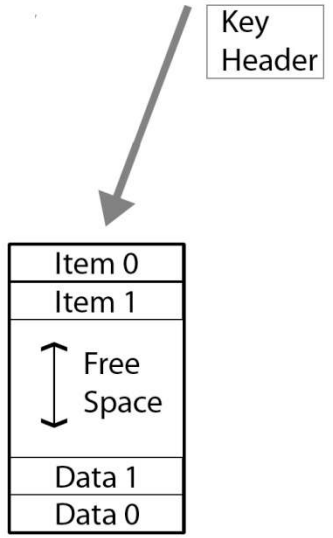
- Obsahuje položky seřazeny dle 136 bit klíče
- Skládá se ze tří typů struktur na disku:
  - Hlaviček bloků, klíčů a položek

```
struct btrfs_header {
    u8 csum[32];
    u8 fsid[16];
    __le64 blocknr;
    __le64 flags;

    u8 chunk_tree_uid[16];
    __le64 generation;
    __le64 owner;
    __le32 nritems;
    u8 level;
}
```

```
struct btrfs_disk_key {
    __le64 objectid;
    u8 type;
    __le64 offset;
}
```

```
struct btrfs_item {
    struct btrfs_disk_key key;
    __le32 offset;
    __le32 size;
}
```



- **Inody**
  - Uloženy v „struct btrfs_inode_item „ s offsetem=0 a hodnotou typu = 1 v klíči
  - Btrs si drží nejvyšší objectId a vždy vrací +1 při vytvoření souboru
  - Btrfs podporuje mnoho stromů. Stromy jsou očíslovány 5,256,257. Objekty uvnitř podsvazku si drží jejich číslo v inodu

Snapshots 1

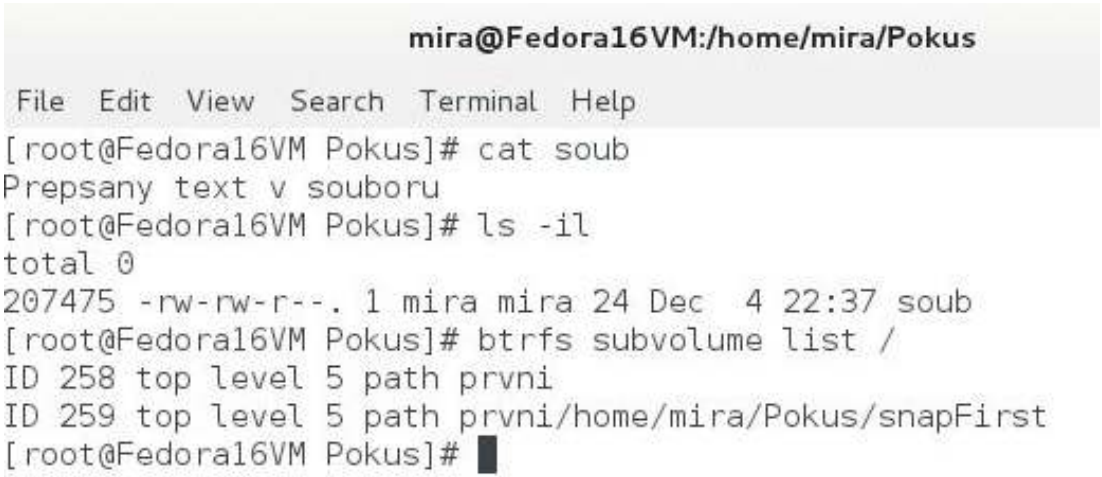
- Pouze pro root
- COW
- Téměř neomezené množství
- Zabírá pouze nezbytné místo

```
[mira@Fedora16VM Pokus]$ echo Text v prvnim souboru > soub
[mira@Fedora16VM Pokus]$ ls
soub
[mira@Fedora16VM Pokus]$ cat soub
Text v prvnim souboru
[mira@Fedora16VM Pokus]$ btrfs subvolume snapshot / prvni
Create a snapshot of '/' in './prvni'
[mira@Fedora16VM Pokus]$ echo Prepsany text v souboru > soub
[mira@Fedora16VM Pokus]$ cat
prvni/ soub
[mira@Fedora16VM Pokus]$ cat
prvni/ soub
[mira@Fedora16VM Pokus]$ cat soub
Prepsany text v souboru
[mira@Fedora16VM Pokus]$ cat prvni/home/mira/Pokus/soub
Text v prvnim souboru
[mira@Fedora16VM Pokus]$ ls -il
total 4
   256 dr-xr-xr-x. 1 root root 122 Nov 30 23:47 prvni
207475 -rw-rw-r--. 1 mira mira  24 Dec  4 22:37 soub
[mira@Fedora16VM Pokus]$
```

Snapshots 2

- btrfs subvolume set-default 258 prvni
- btrfs subvolume snapshot /Pokus snapFirst
- btrfs subvolume set-default 5 /

- Delete:
- sudo btrfs subvolume delete prvni/



Deduplikace

- cp –reflink=auto
- Pokud system „reflink“ neumí, vytvoří se standardní kopie
- Podobnost snapshotům

```
[mira@Fedora16VM Downloads]$ ls -il
total 619520
207780 -rw-------. 1 mira mira 634388480 Dec  5 10:29 Fedora-16-i686-Live-Desktop.iso
[mira@Fedora16VM Downloads]$ df /
Filesystem     1K-blocks    Used Available Use% Mounted on
/dev/sda4       26846208 4877408  19855456  20% /
[mira@Fedora16VM Downloads]$ time cp --reflink=auto Fedora-16-i686-Live-Desktop.iso prepisovKopi
eFedora.iso

real    0m0.012s
user    0m0.000s
sys     0m0.010s
[mira@Fedora16VM Downloads]$ ls -il
total 1239040
207780 -rw-------. 1 mira mira 634388480 Dec  5 10:29 Fedora-16-i686-Live-Desktop.iso
207812 -rw-------. 1 mira mira 634388480 Dec  5 10:43 prepisovKopieFedora.iso
[mira@Fedora16VM Downloads]$
[mira@Fedora16VM Downloads]$ df/
bash: df/: No such file or directory
[mira@Fedora16VM Downloads]$ df /
Filesystem     1K-blocks    Used Available Use% Mounted on
/dev/sda4       26846208 4877416  19855456  20% /
[mira@Fedora16VM Downloads]$
```

Journal

- Nemělo by být třeba , COW ;)
- Pokud dojde k výpadku sítě , ztráta úkolů cca max. posledních 30 sek.
- V souboru btrfs-zero-log
- Obnova:
  - Build a spuštění btrfs-zero-log tool
  - make btrfs-zero-log
  - sudo btrfs-zero-log /dev/sda1

Name: btrfs-zero-log
Type: executable (application/x-executable)
Size: 154.4 kB (154,364 bytes)

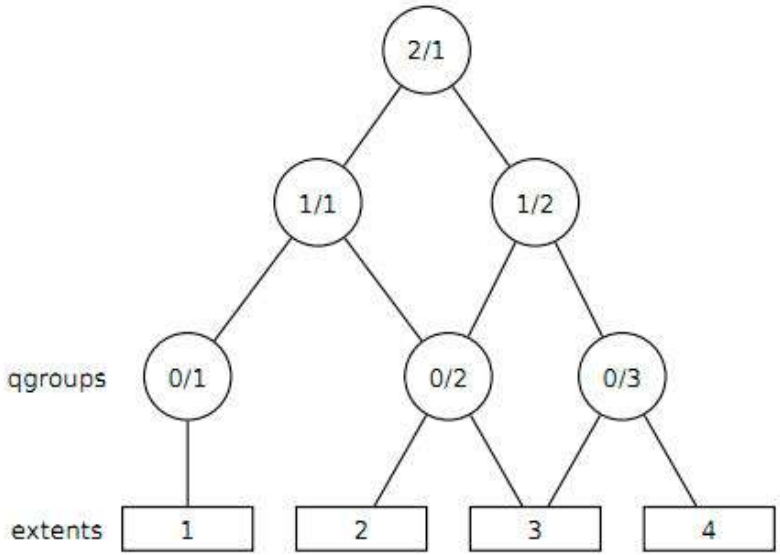
Location: /sbin
Volume: unknown

User quotas

- Zatím neimplementováno

- Flexibilní koncept
- Přidělené podsvazky
- Každý uživatel jeden podsvazek
- Imitace disk.oddílu
- Online zvětšení

- Qgroups <level>/<id>
  - Referované množství místa
  - Exkluzivní množství místa



FSCK, /lost+found

- Zatím neimplementováno, stále ve vývoji

```
[mira@Fedora16VM boot]$ ll
total 43874
-rw-r--r--. 1 root root   124362 Nov  1 21:58 config-3.1.0-7.fc16.i686.PAE
-rw-r--r--. 1 root root   124518 Nov 22 09:54 config-3.1.2-1.fc16.i686.PAE
drwxr-xr-x. 2 root root     1024 Nov 30 23:48 grub
drwxr-xr-x. 3 root root     7168 Nov 30 23:01 grub2
-rw-r--r--. 1 root root 16194739 Nov 30 23:54 initramfs-3.1.0-7.fc16.i686.PAE
.img
-rw-r--r--. 1 root root 16286664 Nov 30 23:33 initramfs-3.1.2-1.fc16.i686.PAE
.img
drwx------. 2 root root    12288 Nov 30 23:46 lost+found
-rw-------. 1 root root  1944686 Nov  1 21:58 System.map-3.1.0-7.fc16.i686.PA
E
-rw-------. 1 root root  1945313 Nov 22 09:54 System.map-3.1.2-1.fc16.i686.PA
E
-rwxr-xr-x. 1 root root  4134144 Nov  1 21:58 vmlinuz-3.1.0-7.fc16.i686.PAE
-rwxr-xr-x. 1 root root  4136576 Nov 22 09:54 vmlinuz-3.1.2-1.fc16.i686.PAE
[mira@Fedora16VM boot]$ df .
Filesystem     1K-blocks  Used Available Use% Mounted on
/dev/sda2         495844 56167    414077  12% /boot
[mira@Fedora16VM boot]$ df -T
Filesystem     Type     1K-blocks    Used Available Use% Mounted on
rootfs         rootfs    26846208 5213764  19711148  21% /
devtmpfs       devtmpfs   1023344       0   1023344   0% /dev
tmpfs          tmpfs      1031604     228   1031376   1% /dev/shm
tmpfs          tmpfs      1031604   40380    991224   4% /run
/dev/sda4      btrfs     26846208 5213764  19711148  21% /
tmpfs          tmpfs      1031604   40380    991224   4% /run
tmpfs          tmpfs      1031604       0   1031604   0% /sys/fs/cgroup
tmpfs          tmpfs      1031604       0   1031604   0% /media
/dev/sda2      ext4        495844   56167    414077  12% /boot
[mira@Fedora16VM boot]$
```

Grow, schrink

- Implementováno a možnost rozšíření/zmenšení za běhu

```
[root@Fedora16VM /]# time btrfs filesystem resize -3g /mnt
Resize '/mnt' of '-3g'

real    0m0.009s
user    0m0.000s
sys     0m0.006s
[root@Fedora16VM /]# time btrfs filesystem resize +4g /mnt
Resize '/mnt' of '+4g'

real    0m1.173s
user    0m0.000s
sys     0m0.026s
[root@Fedora16VM /]# time btrfs filesystem resize 18g /mnt
Resize '/mnt' of '18g'

real    0m0.009s
user    0m0.000s
sys     0m0.005s
[root@Fedora16VM /]#
```

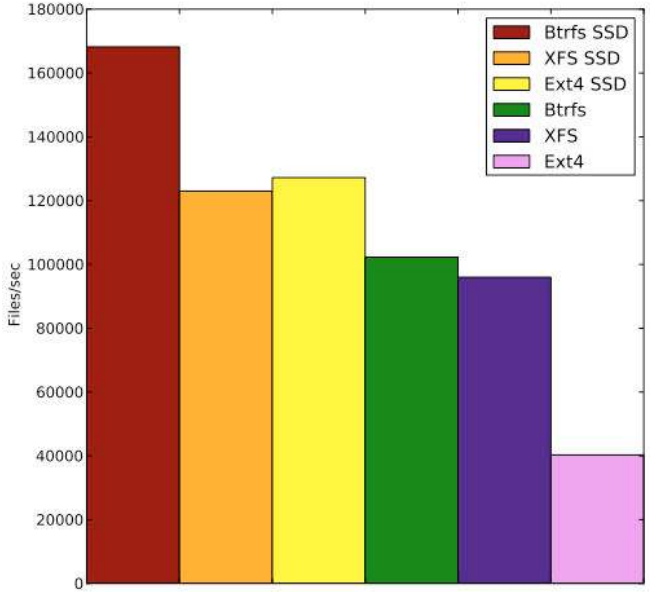
BTRFS a SSD?

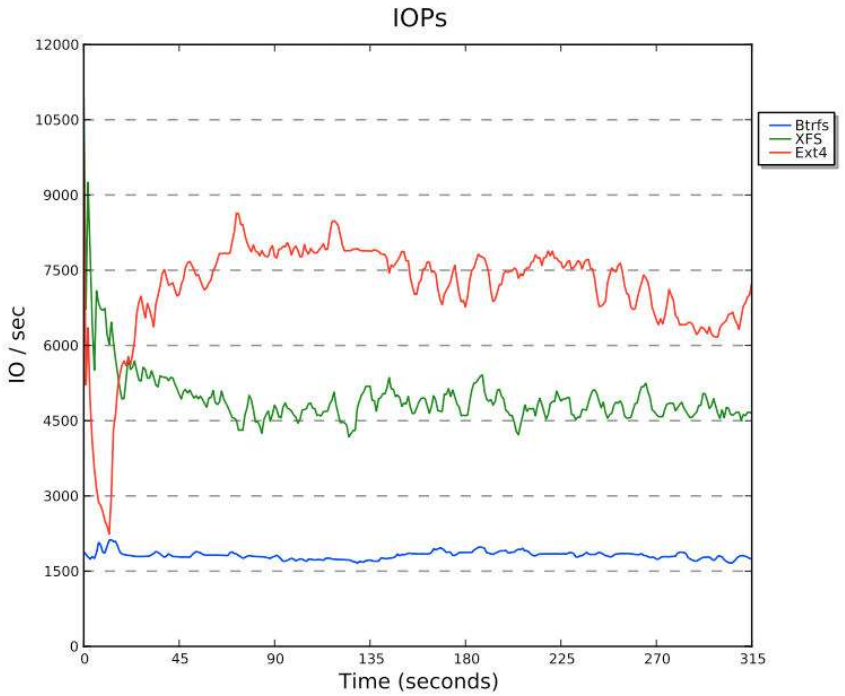
- Původně primárně pro rotační HDD
- Na serveru Phoronix test v 2009/05 , kde btrfs s SSD pomalejší
- Nyní mnoho optimalizací ve vývoji
- Předpokládá se zrychlení souborového systému

- Plán optimalizací:
  - Vysoké IOPs SSD
  - Uživatelské SSD
  - Optimalizace pro jediný disk
  - Hybridní disky
  - Ukládání vytížených kritických rozsahů na SSD
  - Metadata
  - fsync log
  - Přesun na pomalejší a větší disky po „vychladnutí“ dat

BTRFS a SSD?

- Benchmark – vytváření souborů:
  - Ext4 hledá ve velkém prostoru disku
  - XFS prochází místa na disku rovnoměrně a přímočaře
  - XFS a Ext4 vykazují velkou log aktivitu
  - Btrfs sekvenčně zapisuje a někdy náhodně čte





Závěr

- Budoucnost souborových systému pro Linux
- Unikátnost
- Velký vývoj a optimalizace
- Bugy

- Rozšíření do distribucí
- User friendly

Děkuji za pozornost