btrfs

Playing with `btrfs` under `Debian` without any knowledge of this filesystem.

context

- a `Dell R720` server connected to a 60 (3*4*5) disks (3TB each) storage bay (SAS attachment)
- Debian Squeeze installation from scratch

latest kernel installation

dependencies installation

```
aptitude install build-essential kernel-package debconf-utils dpkg-dev
debhelper ncurses-dev fakeroot libncurses-dev
```

kernel download

```
# we are going to compile the kernel as the "maintenance" user
adduser maintenance src
cd /home/maintenance
mkdir src
cd src
wget https://www.kernel.org/pub/linux/kernel/v3.x/linux-3.17.1.tar.xz
tar xvf linux-3.17.1.tar.xz
ln -s linux-3.17.1 linux
chown -R maintenance:maintenance /home/maintenance/src
cd /usr/src
ln -s /home/maintenance/src/linux .
```

kernel configuration

```
su - maintenance
cd ~/src/linux
cp -vi /boot/config-`uname -r` .config
make menuconfig
```

kernel compilation

```
make-kpkg clean
```

```
fakeroot make-kpkg --initrd --append-to-version=-$(date '+%Y%m%d') kernel-
image kernel-headers
```

kernel installation

```
# go back to the "root" identity and...
cd /home/maintenance/src
dpkg -i *.deb
reboot
```

latest btrfs installation

compilation pre-requesites installation

```
aptitude build-dep btrfs-tools
aptitude install uuid-dev libattr1-dev zlib1g-dev libacl1-dev e2fslibs-dev
libblkid-dev liblzo2-dev
aptitude install asciidoc xmlto --without-recommends
```

btrfs-progs installation

```
cd /usr/local/src
git clone git://git.kernel.org/pub/scm/linux/kernel/git/kdave/btrfs-
progs.git
make
./btrfs fi show
```

btrfs kernel module installation

Done at the kernel installation.

playing with btrfs

filesystem creation

Creating a 60 disks RAID6 filesystem.

```
./mkfs.btrfs -f -d raid6 -m raid6 /dev/sdaa /dev/sdab /dev/sdac /dev/sdad
/dev/sdae /dev/sdaf /dev/sdag /dev/sdah /dev/sdai /dev/sdaj /dev/sdak
/dev/sdal /dev/sdam /dev/sdan /dev/sdao /dev/sdap /dev/sdaq /dev/sdar
/dev/sdas /dev/sdat /dev/sdau /dev/sdav /dev/sdaw /dev/sdax /dev/sday
/dev/sdaz /dev/sdba /dev/sdbb /dev/sdbc /dev/sdbd /dev/sdbe /dev/sdbf
/dev/sdbg /dev/sdbh /dev/sdbi /dev/sdbj /dev/sdbk /dev/sdbl /dev/sdbm
/dev/sdbn /dev/sdbo /dev/sdh /dev/sdi /dev/sdj /dev/sdk /dev/sdl /dev/sdm
/dev/sdn /dev/sdo /dev/sdp /dev/sdq /dev/sdr /dev/sds /dev/sdt /dev/sdu
/dev/sdv /dev/sdw /dev/sdx /dev/sdy /dev/sdz
```

Result:

```
root@aigrette:/usr/local/src/btrfs-progs# ./btrfs filesystem show
Label: none  uuid: 648909f8-5393-4cab-b020-6b3f2eb17d33
    Total devices 60 FS bytes used 112.00KiB
    devid    1 size 3.64TiB used 258.88MiB path /dev/sdaa
    devid    2 size 3.64TiB used 238.88MiB path /dev/sdab
...
    devid   59 size 3.64TiB used 239.88MiB path /dev/sdy
    devid   60 size 3.64TiB used 239.88MiB path /dev/sdz

Btrfs v3.17
```

Mounting the new file system:

```
btrfs device scan
mkdir /data
mount /dev/sdaa /data
# we can use any of the device in the FS in the mount command
```

subvolumes creation

Creating 2 subvolumes:

```
./btrfs subvolume create /data/teams
./btrfs subvolume create /data/perso
```

```
root@aigrette:/usr/local/src/btrfs-progs# ./btrfs subvolume list -p /data
ID 288 gen 275 parent 5 top level 5 path teams
ID 290 gen 278 parent 5 top level 5 path perso
```

Oups ! I have forgotten to set a label to my btrfs filesystem. Fixing this:

```
btrfs filesystem label /dev/sdaa btrfs_vol
```

Testing:

```
btrfs filesystem show btrfs_vol
```

Now we can define the btrfs filesystem on the `/etc/fstab` file:

```
...
LABEL=btrfs_vol      /data             btrfs   defaults,noauto
0 0
...
```

We can now mount it with:

```
mount /data
```

Creating a new file:

```
cd /data
touch toto
```

adding a new root subvolume

We want to define a new root subvolume containing the `perso` and `teams` subvolumes.

```
cd  /data
btrfs subvolume create root
btrfs subvolume set-default 307 /data # 307 is the id of root
mv teams root/
mv perso root/
```

Now when we remount the btrfs filesystem:

```
umount /data
mount /data
# mount -o remount /data
# did not work...
```

we see the `teams` and `perso` directories in `/data` but not `root`.

Anyway to access the root of the btrfs filesystem we can use the `subvolid=0` mount option.

```
LABEL=btrfs_vol      /data             btrfs   defaults
0 0
# another mount point to access the root filesystem
LABEL=btrfs_vol       /btrfs     btrfs   defaults,noauto,subvolid=0   0 0
```

```
mount /btrfs

root@aigrette:/data# ls /btrfs/
root
root@aigrette:/data# ls /btrfs/root/
BaS  perso  teams
```

We will use this feature for snapshots.

snapshots

Creating a new subvolume to play with snapshots.

```
cd /data
btrfs subvolume create BaS
```

```
root@aigrette:/data# btrfs subvolume list /data/
ID 288 gen 482 top level 307 path teams
ID 290 gen 480 top level 307 path perso
ID 307 gen 483 top level 5 path root
ID 352 gen 483 top level 307 path BaS
```

Creating a snapshot of BaS:

```
btrfs subvolume snapshot /data/BaS /btrfs/BaS-snap1
```

Adding content and snapshoting:

```
root@aigrette:/btrfs# echo 'a' > /data/BaS/foo
root@aigrette:/btrfs# btrfs subvolume snapshot /data/BaS /btrfs/BaS-snap2
Create a snapshot of '/data/BaS' in '/btrfs/BaS-snap2'
root@aigrette:/btrfs# echo 'b' >> /data/BaS/foo
root@aigrette:/btrfs# btrfs subvolume snapshot /data/BaS /btrfs/BaS-snap3
Create a snapshot of '/data/BaS' in '/btrfs/BaS-snap3'
```

Checking:

```
root@aigrette:/btrfs# cat /btrfs/BaS-snap2/foo
a
root@aigrette:/btrfs# cat /btrfs/BaS-snap3/foo
a
b
```

It works like a charm.

file transfert

I have started a transfert of a 3T directory from a remote server:

```
# remote server
cd /data-backup/teams
tar cf - mydir | mbuffer -s 128k -m 1G -r 500M | nc -q 1 aigrette 7000
# btrfs server
```

```
cd /data/teams
nc -q 1 -l -p 7000 | mbuffer -s 128k -m 1G | tar xv
```

NFS export of the btrfs volumes

```
aptitude install nfs-kernel-server
```

My `/etc/exports` file:

```
# for tests purposes
/data/teams 140.77.82.0/24(rw,sync,no_subtree_check)
140.77.250.0/24(rw,sync,no_subtree_check)
/data/perso 140.77.82.0/24(rw,sync,no_subtree_check)
140.77.250.0/24(rw,sync,no_subtree_check)
```

kernel upgrade and first issue

I have installed the last 3.17.2 linux kernel and rebooted.

```
...
[  630.696055] BTRFS: failed to read the system array on sdbo
```

Could only mount the filesystem in degraded mode:

```
mount -o degraded /data
```

Tried to remove the faulty disk:

```
btrfs device delete /dev/sdbo /data
# btrfs device delete missing /data
# may have been a better idea ?
```

then:

```
umount /data
mount /data
```

leads to a `segmentation fault`.

Rebooting and trying another mount option:

```
mount -t btrfs -o recovery,nospace_cache,clear_cache  /dev/sdaa /data
```

Not better the command freezes.

restarting with a new FS and kernel

The RAID5/6 is currently experimental, I have decided to restart with a `RAID10` filesystem and a fresh `3.18-RC3` kernel.

I have then run a `btrfs filesystem balance /data` and start to retrieve data from another server at the same time.

```
dstat
----total-cpu-usage---- -dsk/total- -net/total- ---paging-- ---system--
usr sys idl wai hiq siq| read  writ| recv  send|  in   out | int   csw
  0   1  99   0   0   0|  25M  138M|   0     0 |   0    13B|1957    27k
  0   8  91   1   0   0| 641M  585M| 574M 1263k|   0     0 |  25k  249k
  0   8  91   0   0   0| 924M    0 | 661M 1541k|   0     0 |  25k  296k
  0   8  91   0   0   0| 843M   41M| 530M  851k|   0     0 |  22k  220k
  0   8  89   2   0   1|  30M 1992M| 819M 2743k|   0     0 |  28k  330k
  0  10  77  13   0   1|  37M 1783M| 762M 2419k|   0     0 |  22k  399k
  0   9  79  13   0   0|  28M 2179M| 620M 2277k|   0     0 |  25k  336k
  0   6  85   9   0   1|  32M 2051M| 572M 2885k|   0     0 |  25k  295k
  0   6  91   2   0   0|  16M 1946M| 619M 2409k|   0     0 |  19k  297k
  0   9  87   3   0   0|  36M 1564M| 546M 2871k|   0     0 |  27k  369k
  0  10  77  12   0   0|  35M 2291M| 556M 1453k|   0     0 |  23k  331k
  0   5  84  10   0   0|  52M 1988M| 532M 1324k|   0     0 |  18k  229k
  0   6  88   5   0   1|  13M 2036M| 582M 2930k|   0     0 |  19k  279k
  0   6  85   8   0   0|  25M 1976M| 630M 3623k|   0     0 |  23k  247k
  0   2  95   2   0   0|  35M 2010M| 612M 3279k|   0     0 |  21k  144k
  0   7  91   2   0   0| 281M  352M| 549M 1407k|   0     0 |  15k  277k
  0   9  89   1   0   0| 330M   31M| 594M 1674k|   0     0 |  20k  285k
  0   6  92   2   0   0|   0   598M| 575M 2160k|   0     0 |  24k  311k
```

```
top
load average: 6,08, 4,30, 2,57
# light for a 40 cores machine
```

The copy has finished succesfully but just after a kernel error appeared (balance process ?).

```
[101046.700011] ------------[ cut here ]------------
[101046.700027] WARNING: CPU: 11 PID: 44077 at fs/btrfs/super.c:260
__btrfs_abort_transaction+0x46/0xf8 [btrfs]()
[101046.700029] BTRFS: Transaction aborted (error -27)
[101046.700030] Modules linked in: btrfs xor raid6_pq dm_mod ses enclosure
nfsd auth_rpcgss oid_registry nfs_acl nfs lockd grace fscache sunrpc 8021q
garp stp llc loop joydev hid_generic usbhid hid snd_pcm snd_timer
x86_pkg_temp_thermal coretemp kvm_intel kvm ghash_clmulni_intel aesni_intel
aes_x86_64 ablk_helper cryptd lrw gf128mul glue_helper snd soundcore
iTCO_wdt iTCO_vendor_support lpc_ich mfd_core sb_edac edac_core dcdbas
microcode shpchp pcspkr evdev tpm_tis tpm ehci_pci ehci_hcd usbcore
usb_common ipmi_si ipmi_msghandler acpi_pad wmi acpi_power_meter button
processor thermal_sys ext4 crc16 jbd2 mbcache sg sd_mod crc32c_intel ixgbe
dca mdio mpt2sas raid_class tg3 megaraid_sas scsi_transport_sas ptp pps_core
```

```
scsi_mod libphy
[101046.700070] CPU: 11 PID: 44077 Comm: btrfs Not tainted 3.18.0-
rc3-20141103 #1
[101046.700071] Hardware name: Dell Inc. PowerEdge R720/08RW36, BIOS 2.2.3
05/20/2014
[101046.700072]  0000000000000000 0000000000000009 ffffffff813a2441
ffff881fb6c0fa28
[101046.700074]  ffffffff81038267 ffff88164891c800 ffffffffa04db5ce
ffff88191292ec80
[101046.700076]  00000000ffffffe5 ffff880ffcf1d000 ffff881fb8fda8e0
ffffffffa055ee20
[101046.700078] Call Trace:
[101046.700085]  [<ffffffff813a2441>] ? dump_stack+0x41/0x51
[101046.700089]  [<ffffffff81038267>] ? warn_slowpath_common+0x78/0x90
[101046.700094]  [<ffffffffa04db5ce>] ? __btrfs_abort_transaction+0x46/0xf8
[btrfs]
[101046.700096]  [<ffffffff81038317>] ? warn_slowpath_fmt+0x45/0x4a
[101046.700101]  [<ffffffffa04db5ce>] ? __btrfs_abort_transaction+0x46/0xf8
[btrfs]
[101046.700110]  [<ffffffffa04f065e>] ?
btrfs_create_pending_block_groups+0x121/0x156 [btrfs]
[101046.700119]  [<ffffffffa04fe777>] ? __btrfs_end_transaction+0x7b/0x2d6
[btrfs]
[101046.700127]  [<ffffffffa04ef87e>] ? btrfs_set_block_group_ro+0x112/0x11d
[btrfs]
[101046.700139]  [<ffffffffa053e083>] ?
btrfs_relocate_block_group+0x6b/0x267 [btrfs]
[101046.700149]  [<ffffffffa051dd33>] ?
btrfs_relocate_chunk.isra.68+0x30/0x9f [btrfs]
[101046.700158]  [<ffffffffa051f095>] ? btrfs_balance+0x9a5/0xb92 [btrfs]
[101046.700168]  [<ffffffffa0526200>] ? btrfs_ioctl_balance+0x21a/0x297
[btrfs]
[101046.700177]  [<ffffffffa0529793>] ? btrfs_ioctl+0x116d/0x211e [btrfs]
[101046.700182]  [<ffffffff8111fbf9>] ? path_openat+0x233/0x4c5
[101046.700188]  [<ffffffff8102d207>] ? __do_page_fault+0x339/0x3df
[101046.700191]  [<ffffffff810f0811>] ? vma_link+0x6b/0x8a
[101046.700194]  [<ffffffff811223ec>] ? do_vfs_ioctl+0x3ed/0x436
[101046.700196]  [<ffffffff8112247e>] ? SyS_ioctl+0x49/0x77
[101046.700199]  [<ffffffff813a7ee2>] ? page_fault+0x22/0x30
[101046.700201]  [<ffffffff813a6512>] ? system_call_fastpath+0x12/0x17
[101046.700202] ---[ end trace 655013971a074e54 ]---
[101046.700204] BTRFS: error (device sdz) in
btrfs_create_pending_block_groups:9214: errno=-27 unknown
[101046.700230] BTRFS info (device sdz): forced readonly
```

commands

- `btrfs device scan` # scan for btrfs filesystems

- `btrfs filesystem show` # gives you a list of all the btrfs filesystems
- `btrfs subvolume create <path>` # create a subvolume
- `btrfs subvolume delete <path>` # delete a subvolume (or a snapshot)
- `btrfs subvolume list -p <path>` # list subvolumes
- `btrfs filesystem df <path>` # df command (basic df command displays wrong informations)
- `btrfs subvolume get-default <path>` # displays the ID of the default subvolume that is mounted for the specified subvolume
- `btrfs subvolume set-default 258 <path>` # set the default subvolume for the specified subvolume
- `btrfs filesystem label <device> <label>` # set la filesystem label
- `btrfs subvolume snapshot <path-snapname>` # snapshot creation
- `btrfs filesystem balance <path>` # balance the chunks across the device.

references

- https://btrfs.wiki.kernel.org
- https://btrfs.wiki.kernel.org/index.php/Btrfs%28command%29
- https://btrfs.wiki.kernel.org/index.php/Using_Btrfs_with_Multiple_Devices
- https://docs.oracle.com/cd/E37670_01/E37355/html/ol_btrfs.html
- https://www.kernel.org/
- http://www.isalo.org/wiki.debian-fr/index.php?title=Compiler_et_patcher_son_noyau
- https://lwn.net/Articles/577961/