BTRFS on multiple devices

https://serverfault.com/questions/814546/how-to-obtain-read-speeds-of-two-disks-using-mdadm-btrfs-raid1-or-zfs-mirror

Encryption layout

BTRFS on crypted devices

btrfs raid1	
crypt	crypt
sda1	sdb1

- (-) Encryption has to be performed twice (no problem for modern CPUs)
- (+) Bit-flips can be detected immediately and corrected by BTRFS
- (-) not possible to get benefit from hybrid RAID1 (mixed disc HDD & SSD)

BTRFS on crypted devices with LVM

btrfs raid1	
crypt	crypt
LVM	LVM
(DM-Cache on fast SSD)	
sda1	sdb1

- (-) Encryption has to be performed twice (no problem for modern CPUs)
- (+) Bit-flips can be detected immediately and corrected by BTRFS
- (-) not possible to get benefit from hybrid RAID1 (mixed disc HDD & SSD)
- (-) small overhead from LVM but
- (+) DM-cache with SSD disc can be used: https://www.redhat.com/en/blog/improving-read-performance-dm-cache

BTRFS on soft RAID1

btrfs	
crypt	
raid1	
sda1	sdb1

- (+) encryption performed only once (less CPU consumption)
- (-) Bit-flips detected only during RAID1 scan.
- (+) Benefits from hybrid RAID1 (SSH & HHD) using “writemostly” feature). See http://tansi.info/hybrid/

Prepare 2nd disc

Lets assume that we have 2nd disc for other than system stuff, but we want to use it as backup solution for system disc. System should be able to boot if one disc fails.

> By default, metadata will be mirrored across two devices and data will be striped across all of the devices present.
> If only one device is present, metadata will be duplicated on that one device.

Copy partition layouts from sdb to sda drive:

```
sudo sfdisk -d /dev/sdb > parts
sudo sfdisk  /dev/sda < parts
```

Recreate LVM layout if needed.

btrfs 2nd disk

Add 2nd disc to btrfs

```
btrfs device add /dev/disc1/btrfs1 /
btrfs device usage /
nice ionice -c3 btrfs balance start -v -dconvert=raid1 -mconvert=raid1 -
dsoft -msoft /
btrfs device usage /
```

Note: “soft” filter will convert only chunks which are not already converted (usefull when previous balance was incomplete)

Make sure that data on both disc are the same. Especially system data are on both disc in RAID1 mode.

```
# btrfs device usage /
/dev/sda3, ID: 1
   Device size:            27.72GiB
   Device slack:              0.00B
   Data,RAID1:             18.93GiB
   Metadata,RAID1:          2.00GiB
   System,RAID1:           32.00MiB
   Unallocated:             6.75GiB

/dev/sdc5, ID: 2
   Device size:            49.40GiB
   Device slack:              0.00B
   Data,RAID1:             18.93GiB
   Metadata,RAID1:          2.00GiB
   System,RAID1:           32.00MiB
```

```
   Unallocated:            28.44GiB
```

OR

Create blank 2nd disc

NOT POSSIBLE:

~~Create degraded RAID1 with single disc (to add 2nd later)~~

```
mkfs.btrfs -d raid1 -m raid1 -L btrfs_pool2 /dev/mapper/pool2
```

Balance command

```
btrfs balance start -d <filter> -m <filter> /
```

Starts balance action on data (-d) and metadata (-m). For details please look into man btrfs balance
Full command:

```
btrfs balance start -dconvert=<profile> -mconvert=<profile> /
```

Possible profiles are: raid0, raid1, raid10, raid5, raid6, dup, single

- single - is to put single copy of data (not stripping) and duplicated metadata
- dup - duplicate metadata - but it is not guaranteed that metadata is duplicated on separate devices

> For now, RAID-1 means 'one copy of what's important exists on two of the drives
> in the array no matter how many drives there may be in it'.“

It is possible to create raid1 profile in degraded state, to add real 2nd hard disc later. As 2nd disc use small loop device and after creation remove loop device and remount in in degraded mode.

Error: unable to start balance

ERROR in dmesg:

```
[ 2687.757913] BTRFS error (device sda1): unable to start balance with
target data profile 32
```

Update to kernel 4.0.2 and btrfs-tools v4 doesn't help. Probably dup (profile 32) cannot be used for data.

It works:

```
btrfs balance start -dconvert=raid1 -mconvert=raid1 /
```

GRUB

To boot automatically kernel with degraded btrfs, please add “rootflags=degraded” to kernel commandline.

/etc/default/grub

```
GRUB_CMDLINE_LINUX_DEFAULT="quiet rootflags=degraded"
```

Add to grub and fstab. Then update grub:

```
update-grub
grub-install /dev/sda
grub-install /dev/sdb
reboot
```

There is a boot issue (see below) when grub root dir (/boot/grub) is located on BTRFS, and one of disk is missing. During

```
grub-install
```

GRUB sets `prefix` to only one of used disc, and if this device fails, GRUB cannot load rest of GRUB modules and stuck in rescue shell. Maybe in future releases this will be corrected but for now maybe safer is use `/boot` on separate `ext2` (MD Raid1) partition.

- Install prerequisites
- ```
  apt-get install mdadm
  ```
- Duplicate boot partition size
- ```
    sudo sfdisk -d /dev/sda > parts
    sudo sfdisk  /dev/sdb < parts
  ```
- copy data from `/boot` somewhere
- Umount `/boot`
- ```
  umount /boot
  ```
- Clear previous RAID headers
- ```
  mdadm --zero-superblock /dev/sdb1
  ```
- Crate soft RAID1
- ```
  mdadm --create /dev/md0 --level=1 --raid-devices=2 /dev/sda1 /dev/sdb1
  ```
- ```
  mkfs.ext2 /dev/md0
  ```

- update `/etc/fstab` to mount `/dev/md0` as `/boot` (ext2). Remember to use `blkid`
- mount `/boot`
- copy backed-up data to `/boot`
- Save mdadm configuration:

```
mdamd --detail --scan >> /etc/mdadm/mdadm.conf
```

- grub-install /dev/sda
- grub-install /dev/sdb
- update-grub
- update-initramfs

Boot issues

No degraded option

```
Boot fail
BTRFS: failed to read the system array on sda1
BTRFS: open_cree failed
```

And initramfs console appear.

```
mkdir /mnt
mount /dev/sda1 /mnt
```

the same error

```
mount /dev/sda1 /mnt -o degraded
```

WORKS!

Grub wrong root=

PROBLEM

grub-pc 2.02~beta2-22+deb8u1 0 incorrectly handle multiple roots. Internal variable ${GRUB_DEVICE} which is used to generate `root=` parameter include newline character, so grub menuconfig entry is broken:

```
    linux    /@/@root/boot/vmlinuz-4.8.0-0.bpo.2-amd64
root=/dev/mapper/disc2-btrfs2
/dev/mapper/disc1-btrfs1 ro rootflags=subvol=@/@root  rootflags=degraded
```

More here:

- https://bugs.launchpad.net/ubuntu/+source/grub2/+bug/1238347
- https://bugs.launchpad.net/ubuntu/+source/grub2/+bug/1582811

SOLUTION

Fixed in Debian Stretch (testing) version **grub-pc_2.02~beta3-3** Now UUID is used multiple roots. UUID of btrfs filesystem placed on 2 disc is the same:

```
    linux    /@/@root/boot/vmlinuz-4.8.0-0.bpo.2-amd64 root=UUID=52944cdd-f8d0-4798-bd1c-80539c45253d ro rootflags=subvol=@/@root  rootflags=degraded
```

```
# blkid
/dev/mapper/disc2-btrfs2: UUID="52944cdd-f8d0-4798-bd1c-80539c45253d" UUID_SUB="54703ddb-5789-4cac-bfd0-691acadfa33c" TYPE="btrfs"
/dev/mapper/disc1-btrfs1: UUID="52944cdd-f8d0-4798-bd1c-80539c45253d" UUID_SUB="9bc49b38-c014-40e3-876e-08f6873293b8" TYPE="btrfs"
```

So, kernel cmdline parameter **root=UUID=…** is now correct.

no symbol table

PROBLEM Grub says:

```
grub error: no symbol table
```

SOLUTION Grub was not reinstalled after update

```
grub-install /dev/sda
grub-install /dev/sdb
update-grub
```

ALERT! /dev/disk/by-uuid/xxxxxxxxx does not exist. Dropping to a shell

PROBLEM Grub loads kernel and initramfs correctly, but: `Begin: Waiting for root file system …` and Grub still boot into initramfs. After some while, initramfs shell is available.

```
ALERT!
Dropping to shell!
```

Unfortunatelly initramfs cannot activate LVM volumes when kernel cmdline `root=UUID=…` is used.

```
lvm lvscan
```

To boot manually:

```
lvm vgchange -ay
mount /dev/disc2/btrfs2 /root -o device=/dev/disc1/btrfs1
exit
```

System should start - but once.

Problem is located in `/usr/share/initramfs-tools/scripts/local-top/lvm2`. Script check if specified root device needs to be activated by LVM. When UUID is used it is executing code:

```
/dev/*/*)
        # Could be /dev/VG/LV; use lvs to check
        if lvm lvs -- "$dev" >/dev/null 2>&1; then
                lvchange_activate "$dev"
        fi
        ;;
```

In result, command

```
lvm lvs -- disc/by-uuid/52944cdd-f8d0-4798-bd1c-80539c45253d<code>
is executed, resulting following output:
<code>"disc/by-uuid/52944cdd-f8d0-4798-bd1c-80539c45253d": Invalid path for
Logical Volume.
```

Debian bugs about similar problem:

- https://bugs.debian.org/cgi-bin/bugreport.cgi?bug=612402
- https://bugs.debian.org/cgi-bin/bugreport.cgi?bug=741342

solution

Install lvm2 package version 2.02.168-1 (Debian Stretch). Previously was 2.02.111-2.2+deb8u1 (Debian Jessie).

```
update-initramfs -k all -u
btrfs scrub /
```

or force LVM activation: Can't find LVM root dropped back to initramfs

Also `GRUB_DISABLE_OS_PROBER=true` can be added.

/etc/mkinitcpio.conf

```
BINARIES="/bin/btrfs"
```

```
update-init -u -k all
```

Disc fail (removed)

Make sure, all system, metadata and data is balanced to RAID1 mode. Without this it cannot be possible to mount degraded btrfs in RW mode. During RW operation in degraded mode, single system, metadata and data structures will be created on BTRFS so it will need rebalance after missing disc will be connected again.

error: disk not found

```
error: disk 'lvmid/8vc3Xl-....' not found.
Entering rescue mode...
grub rescue>
```

GRUB 2 is unable to find the grub folder or its contents are missing/corrupted. The grub folder contains the GRUB 2 menu, modules and stored environmental data.

grub.cfg sets prefix variable to missing LVM (lvimd/<VG UUID/<LV UUID>), and then call to load module which fails:

```
prefix='(lvmid/8vc3Xl-a40o-ILx2-AoRf-9Z37-4h51-jq0OR5/Hr8CPf-dVqs-uIm1-pkfH-fJdI-14io-rwtx3l)/@/@root/boot/grub'
insmod gettext
```

Rescue shell is very limited. Try to start `normal` shell.

```
set prefix='(lvm/disc1-btrfs1)/@/@root/boot/grub'
insmod normal
normal
```

Keeping /boot/grub in multidevice BTRFS is stupid idea. Another problem comes when BTRFS is not clean, and cannot be used by GRUB to load own modules.

Replace bad disc

```
btrfs device add /dev/disc3/btrfs3 /
btrfs device delete missing /
```

Note: Word 'missing' is special device name for this command

```
  ERROR: error removing the device 'missing' - No space left on device
```

```
btrfs filesystem show
btrfs replace start 6 /dev/disc3/btrfs3 /  # change 6 to your setup
```

Don't forget to restore correct data state:

```
btrfs scrub start /
btrfs balance start -v -dconvert=raid1 -mconvert=raid1 /
```

Fine tuning

- Run dropbear on alternative SSH port

```
apt-get install dropbear
vi /etc/default/dropbear
```

- Add your ssh public key to `/etc/initramfs-tools/root/.ssh/authorized_keys`
- `apt-get install busybox-static`
- Change MAC address of eth0 (for initramfs remote access)

/etc/udev/rules.d/75-mac-spoof.rule

```
ACTION=="add", SUBSYSTEM=="net", ATTR{address}=="08:00:27:f9:3d:3e",
RUN+="/sbin/ip link set dev %k address 08:00:27:f9:12:34"
```

Another option is to add `local-top` script like described here for IPv6
https://www.danrl.com/2015/10/21/debian-jessi-ssh-fde-unlock.html

Don't forget to propgate changes:

```
update-grub
update-initramfs -k all -u
```

TODO

- backup GPT header
- backup LVM header
- backup LUKS header

?????????? ?????????? ?????????? ?????????? ??????????

Grub installed to /dev/sda /dev/sdb Test - remove sda,

In grub menu - additional option to boot debian 8 from sdb1 shows. Booting from sdb1 stops an initramfs, because option to mount degraded is missing.

Booting from original options, stuck on

```
"A start job is running for dev-disk-by\x2duuid-fb8d9... 55s / 1min 30s"
```

After timeout 1:30, system boots into emergency mode.

```
"Give root password for maintenance"
"(or type Control+D to continue)
```

```
journalctl -a
```

```
Jul 16 22:22:17 debian kernel: work still pending
Jul 16 22:23:43 debian systemd[1]: Job dev-disk-by\x2duuid-
fb8d97df\x2d625b\x2d440d\x2db63b\x2dcb529586df24.device/start timed out.
```

```
Jul 16 22:23:43 debian systemd[1]: Timed out waiting for device dev-disk-
by\x2duuid-fb8d97df\x2d625b\x2d440d\x2db63b\x2dcb529586df24.device.
Jul 16 22:23:43 debian systemd[1]: Dependency failed for /home.
Jul 16 22:23:43 debian systemd[1]: Dependency failed for Local File Systems.
Jul 16 22:23:43 debian systemd[1]: Dependency failed for File System Check
on /dev/disk/by-uuid/fb8d97df-625b-440d-b63b-cb529586df24.
Jul 16 22:23:43 debian systemd-journal[177]: Runtime journal is using 4.0M
(max allowed 9.8M, trying to leave 14.8M free of 94.4M available → current
limit 9.
```

/lib/systemd/system/local-fs.target

This is because /home directory is missing (it was located on /dev/sda2) but whole /dev/sda removed. This bad behaviour, because I want system to be bootable even without additional disc.