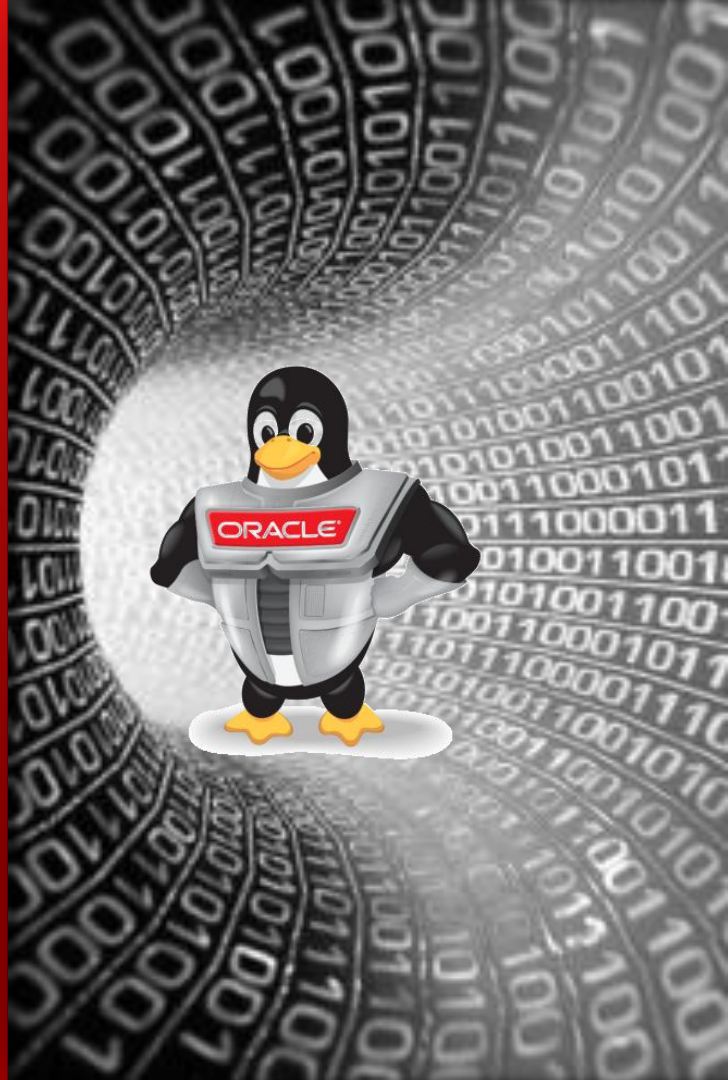

ORACLE®

ORACLE

Snapshots, Checksummen, RAID
Eine Einführung in das Btrfs-Dateisystem

Lenz Grimmer
Senior Product Manager, Oracle Linux



Agenda

- Btrfs History
- Btrfs Architecture
- Btrfs Features
- Btrfs Administration

Btrfs Historie

- Ein modernes Dateisystem für Linux
- Initiiert und koordiniert von Chris Mason (FusionIO)
- Gemeinsam entwickelt durch Beiträge von
  - Fujitsu, FusionIO, Huawei, Intel, Oracle, Red Hat, Strato, SUSE u.a.
- Open Source (GPL)
- In mainline Linux seit 2.6.29 (Jan. 2009)

Btrfs Architektur / Features

- Dateisystem plus Volume Manager (mehrere Platten pro Dateisystem)
- Schreibt Daten und Metadaten via copy-on-write (COW)
- Checksummen (CRC32C) für alle Daten und Metadaten
- Effiziente Snapshots (beschreibbar oder nur-lesend)
- RAID-Unterstützung (RAID0/1/10)
- Transparente Kompression (zlib/LZO)
- Effiziente Speicherung kleiner Dateien
- SSD Optimierungen und Discard/TRIM support

Btrfs Spezialitäten / Besonderheiten

- Migration bestehender ext3/4-Dateisysteme
- Seed Devices
- Schnelle/Inkrementelle Backups
  - btrfs subvolume find-new und btrfs send/receive
- Änderung der RAID-Level für Daten/Metadaten zur Laufzeit
- Online Größenänderung und Defragmentierung
- Snapshots einzelner Dateien (cp --reflink)
- Alternative Kompressionsalgorithmen

Btrfs Skalierbarkeit / Limitierungen

- Max. Dateigröße: 8 EiB
- Max. Anzahl Dateien: 2^64
- Max. Volumengröße: 16 EiB
- Max. Dateinamen-Länge: 255 Byte

- Noch keine (interne) Deduplizierung oder Verschlüsselung
- Behandlung von ENOSPC verbesserungswürdig

Btrfs Architektur

- Btrfs speichert nur eine Art von Metadatenblock
- B-tree Blöcke speichern Schlüssel/Wert-Paare
- Metadatenstrukturen verwenden bestimmte Schlüssel um verwandte Elemente nah beieinander zu speichern
- Logische Addressierung übersetzt Daten- und Metadatenblöcke in physikalische Blöcke auf dem Speichermdedium
- Metadaten für verschiedenen Dateien und Verzeichnisse können im selben B-tree-Block gespeichert werden

Btrfs Architektur

- B-Baum als fundamentale Datenstruktur
- B-Bäume speichern beliebige Schlüssel/Wert-Paare
- Metadatenstrukturen verwenden bestimmte Schlüssel um verwandte Elemente nah beieinander zu speichern
- Speicherplatz wird in Blockgruppen verwaltet
- Blockgruppen werden in Chunks unterteilt
- 1GB für Daten
- 256 MB für Metadaten

Btrfs Architektur Speicherzuweisung

- Chunk Tree verwaltet Verteilung auf Disks
- Daten und Metadaten können verschiedene RAID Level haben

Btrfs Fehlertoleranz / Datenintegrität

- Checksummen für Daten und Metadaten
  - CRC-32C (andere möglich)
- Btrfs Scrub scannt Daten- und Metadatenblocks
  - Automatische Korrektur (wenn intakte Kopie existiert)
- RAID 0/1/10
  - Verschiedene RAID-Level für Daten/Metadaten
  - Chunk-basiert, schnelle Wiederherstellung
  - (RAID 5/6 in Entwicklung)
- `mount -o recover`, `btrfschk` und `btrfs-restore`

Btrfs Scrubbing

- Btrfs CRCs allow us to verify data stored on disk
- CRC errors can be corrected by reading a good copy of the block from another drive
- Scrubbing code scans the allocated data and metadata blocks
- Any CRC errors are fixed during the scan if a second copy exists
- Will be extended to track and offline bad devices

Btrfs Administration

- Kommandozeile, GUIs in Arbeit (z.B. `btrfs-gui`, YaST)
- `mkfs.btrfs`, `btrfs-convert`
- `btrfs <subcommand>`
  - `btrfs device`
  - `btrfs subvolume`
  - `btrfs filesystem`
- Mount-Optionen
- `btrfsck`, `btrfs-restore`

Btrfs Adding Storage

- The “classic” way (LVM/ext4)

```
pvcreate /dev/sdc1
vgextend VolGroup /dev/sdc1
lvextend -L +10G VolGroup/myvolume
resize2fs /dev/mapper/VolGroup-myvolume
```

- Using Btrfs

```
btrfs device add /dev/sdc1 /space
```

Btrfs Seed Devices

- A readonly device can be used as a filesystem seed
- Read/write devices can be added to store modifications
- Changes to the writable devices are persistent across reboots
- The readonly device can be removed at any time
- Multiple read/write filesystems can be built from the same seed

Btrfs SSD Support (Discard/Trim)

- Trim and discard notify storage when we are done with a block
- Btrfs supports both real-time trim and batched trim
- Real-time trims blocks as they are freed
- Batched trims all free space via an ioctl
- Newer kernels will have less penalty for online discard

Btrfs Storage Hints

- Discard and Trim allow the device to ignore blocks the FS isn't using
- Devices may be tiered internally
- Frequently modified or deleted blocks stay on faster cells
- Long lived blocks moved to less expensive storage
- New APIs and standards will allow the FS to give hints to the device
- Large arrays and high end flash can use the hints to improve performance
- Low end flash can use hints to increase cell lifetime
- Btrfs block group layout separates shorter lived metadata from data

Btrfs yum-Integration

- Requires package `yum-plugin-fs-snapshot`
- Creates a snapshot of the root fs prior to installing/updating RPMs (e.g. `/yum_20110926132957`)
- Roll back by rebooting into this snapshot subvolume using mount option `subvol=<snapshot name>` or the `subvolid=<id>`

Btrfs Conversion of existing ext3/4 file systems

- Only few pieces of Btrfs metadata live in fixed locations
- First 1MB of device copied to alternate location
- COW allows keeping an unmodified copy of the original FS as a snapshot for undoing the conversion
- Only metadata blocks are copied, file data blocks are preserved
- Example:

```
fsck.ext3 -f /dev/sdb1
btrfs-convert /dev/sdb1
mount -t btrfs /dev/sdb1 /btrfs
```

Btrfs When Bad Things Happen to Good Data

- Barrier bugs in Btrfs lead to most of the corruptions seen with kernels before v3.2
- Filesystem repair tool in btrfs-progs git
  - Repairs extent allocation tree corruptions in place
  - More repair modes in progress
- Filesystem recovery tool from Josef Bacik
  - Risk free – copies data out of the corrupt FS
- Tree root history log to recover from many hardware errors
  - Jumps back to older versions of the tree roots

Links, Ressourcen

- Oracle Linux Administrator's Solutions Guide for Release 6: http://docs.oracle.com/cd/E37670_01/E37355/html/ol_btrfs.html
- Oracle Linux Hands-on Lab: Storage Management with Btrfs https://wikis.oracle.com/display/oraclelinux/Hands-on+lab+-+Storage+Management+with+Btrfs
- Btrfs Wiki: https://btrfs.wiki.kernel.org/
- Mailing List: http://vger.kernel.org/vger-lists.html#linux-btrfs
- #btrfs channel IRC (freenode.net)

Hardware and Software

ORACLE

Engineered to Work Together


ORACLE®