Linux-Dateisysteme

Lars Thoms

Arbeitsbereich Wissenschaftliches Rechnen
Fachbereich Informatik
Fakultät für Mathematik, Informatik und Naturwissenschaften
Universität Hamburg

2016-04-27

Betreuer: Dr. Julian Kunkel

Gliederung (Agenda)

Definition »Partition«

*Unter einer Partition (lat. partitio = „(Ein)teilung“) versteht man einen **zusammenhängenden Teil** des Speicherplatzes eines geeigneten physischen oder logischen Datenträgers. Eine Partition ist ihrerseits ein **logischer Datenträger** [...].*
Partitionen sind voneinander weitgehend unabhängig und können von Betriebssystemen wie getrennte Laufwerke behandelt werden. [...]

Quelle: Wikipedia [15]

Partitionstabelle

- Erster Datenblock des Datenträgers
- Metadaten über den Datenträger
- Verwaltung der Partitionen
  - Startblock
  - Endblock
  - Flags

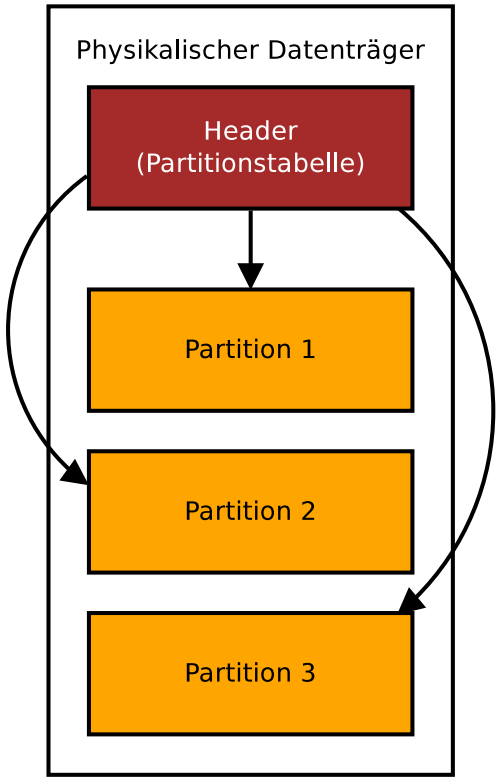

Abbildung: Aufteilung eines Datenträgers

MBR (eine Partitionstabelle)

- Master Boot Record
- 1983 auf IBM-PC XT und MS-DOS/PC DOS 2.0 eingeführt
- Position: CHS 0,0,1 (Cylinder, Head, Sector)
- Größe: 512 Byte
- 2TB mögliche Partitionsgröße

Quelle: Wikipedia [14]

Aufbau des MBR

Adresse	Funktion
0x000 (139B)	Bootloader Programm-code
0x08B (80B)	Fehlermeldung (String)
0x1BE (64B)	Partitionstabelle
0x1FE (2B)	Bootsektor-Signatur

(Bild-)Quelle: An Examination of the Standard MBR [9]

Absolute Sector **0** (Cylinder 0, Head 0, Sector **1**)

```
       0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
0000  FA 33 C0 8E D0 BC 00 7C 8B F4 50 07 50 1F FB FC  .3.....|..P.P...
0010  BF 00 06 B9 00 01 F2 A5 EA 1D 06 00 00 BE BE 07  ................
0020  B3 04 80 3C 80 74 0E 80 3C 00 75 1C 83 C6 10 FE  ...<.t..<.u.....
0030  CB 75 EF CD 18 8B 14 8B 4C 02 8B EE 83 C6 10 FE  .u......L.......
0040  CB 74 1A 80 3C 00 74 F4 BE 8B 06 AC 3C 00 74 0B  .t..<.t.....<.t.
0050  56 BB 07 00 B4 0E CD 10 5E EB F0 EB FE BF 05 00  V.......^.......
0060  BB 00 7C B8 01 02 57 CD 13 5F 73 0C 33 C0 CD 13  ..|...W._s.3...
0070  4F 75 ED BE A3 06 EB D3 BE C2 06 BF FE 7D 81 3D  Ou..........}.=
0080  55 AA 75 C7 8B F5 EA 00 7C 00 00 49 6E 76 61 6C  U.u.....|..Inval
0090  69 64 20 70 61 72 74 69 74 69 6F 6E 20 74 61 62  id partition tab
00A0  6C 65 00 45 72 72 6F 72 20 6C 6F 61 64 69 6E 67  le.Error loading
00B0  20 6F 70 65 72 61 74 69 6E 67 20 73 79 73 74 65   operating syste
00C0  6D 00 4D 69 73 73 69 6E 67 20 6F 70 65 72 61 74  m.Missing operat
00D0  69 6E 67 20 73 79 73 74 65 6D 00 00 00 00 00 00  ing system......
00E0  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
00F0  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
0100  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
0110  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
0120  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
0130  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
0140  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
0150  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
0160  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
0170  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
0180  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
0190  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
01A0  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
01B0  00 00 00 00 00 00 00 00 00 00 00 00 00 00 80 01  ................
01C0  01 00 0B 7F BF FD 3F 00 00 00 C1 40 5E 00 00 00  ......?....@^...
01D0  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
01E0  00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00  ................
01F0  00 00 00 00 00 00 00 00 00 00 00 00 00 00 55 AA  ..............U.
       0  1  2  3  4  5  6  7  8  9  A  B  C  D  E  F
```

Abbildung: Hexdump eines MBR

Primäre/Logische Partition

- MBR ermöglicht nur 4 Partitionen
  - 64B Partitionstabelle, d.h. 16B Metadaten pro Partition
- Logische/Erweiterte Partitionen!
  - Partitionsverwaltung im Datenblock der Logischen Partition

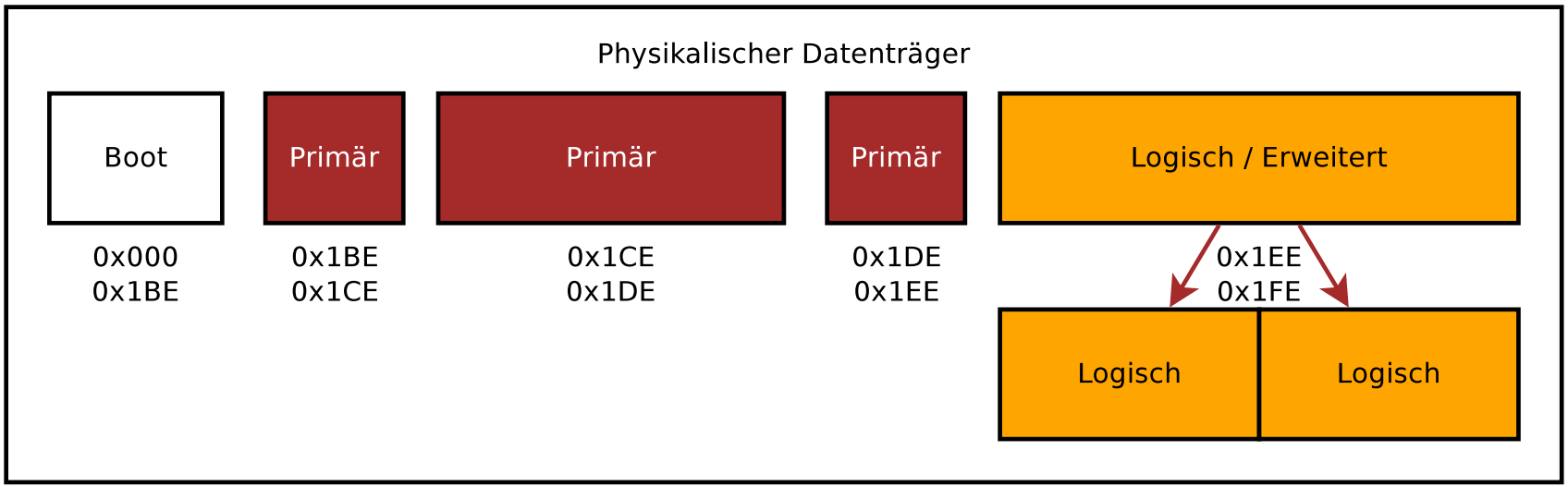

Abbildung: Aufteilung des MBR

GPT (eine bessere Partitionstabelle)

- GUID Partition Table (*GUID:* Globally Unique Identifier)
- GPT ist Teil des UEFI-Standards
- Nachfolger des MBR
  - Erhöhung der Anzahl der primären Partitionen (128 Stück)
  - Maximale Partitionsgröße erhöht (bei 512 Byte Blockgröße: 8 Zebibyte)

Quelle: Wikipedia [13]

Aufbau des GPT

- Adressierung mit Hilfe des Logical Block Addressing
- MBR aus Kompatibilitätsgründen noch als erster Datenblock enthalten
- LBA1 enthält Metadaten des physikalischen Datenträgers (z.B. UUID)
- Partitionstabelle (LBA2-34) für 128 Partitionen

Bildquelle: Wikimedia-Commons [12]

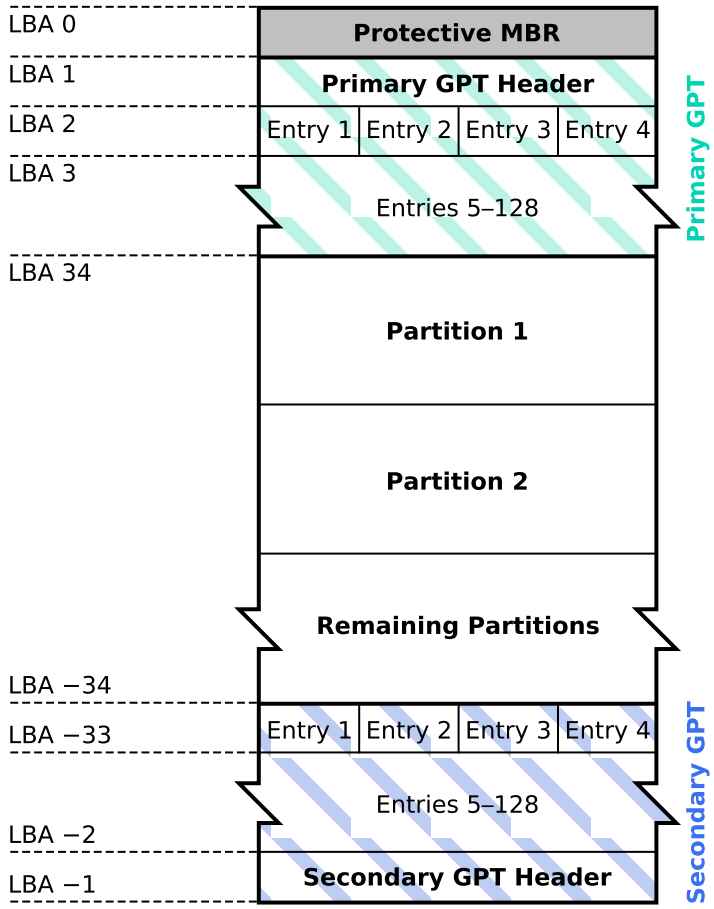

Abbildung: Aufbau des GPT

GPT mit zwei Partitionen erstellen

```
1 $ sudo gdisk /dev/sdc
2
3 // Partitionstabelle erstellen mit o
4 // Partition hinzufuegen mit n
5 // Speichern mit w
```

Mehr Informationen unter: `man gdisk`

RAID

- Redundant Array of Independent Disks
- Verknüpfung von mehreren physikalische Blockgeräten in ein logisches
- Sowohl via Hardware als auch über Software möglich
- Verschiedene Modi, je nach Einsatzzweck
  - RAID 0: Striping
  - RAID 1: Mirroring
  - RAID 5: Striping mit Paritätsbildung

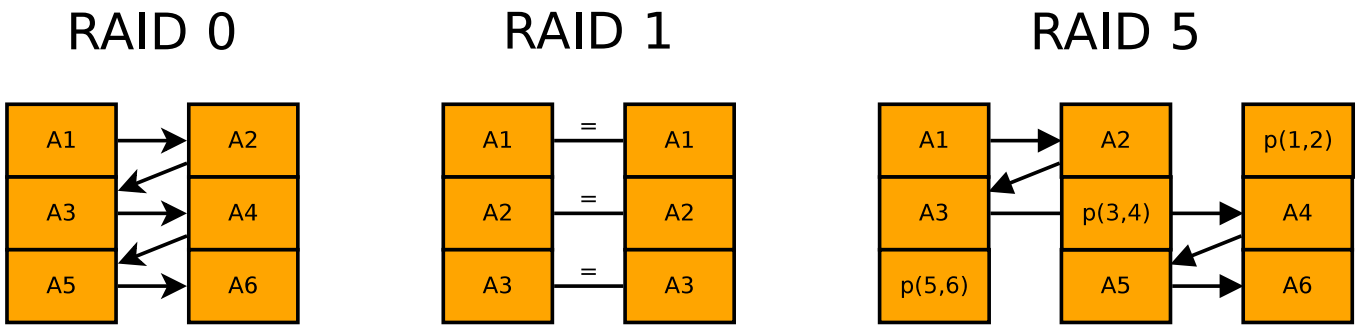

Abbildung: RAID-Modi

Logical Volume Manager

- Fusion von **mehreren** Partitionen
- **Eine** virtuelle Partition mit einem Dateisystem

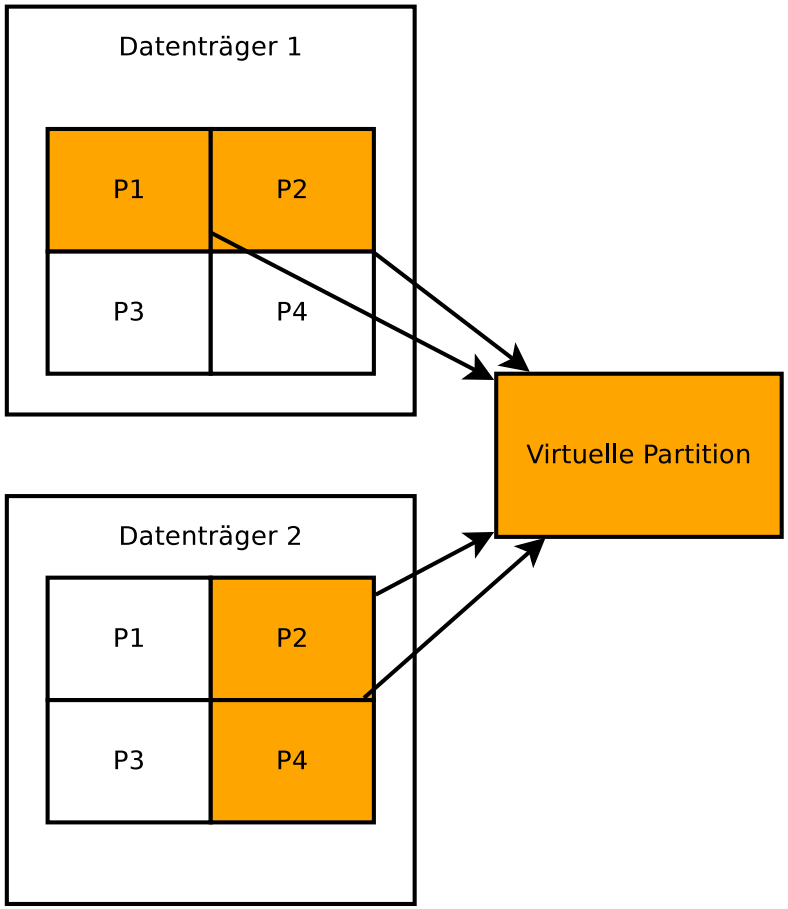

Abbildung: Virtuelle Partition mit LVM

Definition »Dateisystem«

Das Dateisystem (FS) ist Bestandteil des Betriebssystems und bildet die Schnittstelle zwischen diesem und den Laufwerken. Es legt fest, wie der Computer Dateien auf den Datenträgern benennt, speichert, organisiert und verwaltet. Ein Dateisystem besteht aus Dateien, Verzeichnissen und Adressen, über die die Dateien lokalisiert werden.

Quelle: ITWissen [4]

Aufbau eines Dateisystems

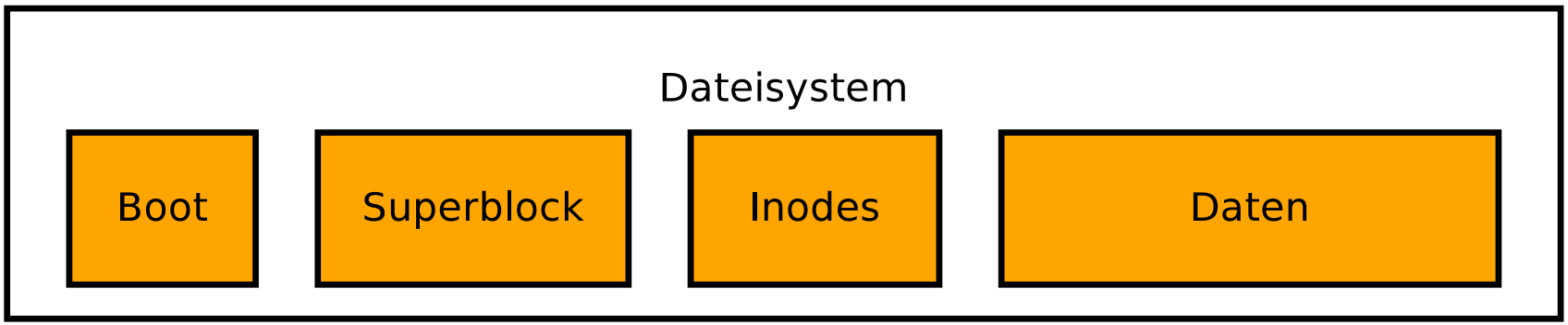

Abbildung: Grundlegende Struktur eines Dateisystems

Boot Reservierter/Ungenutzter Bereich für Bootloader etc.

Superblock Metadaten des Dateisystems (Größe des DS, Größe der Inode-Liste, Anzahl freier Blöcke, ...)

Inodes Liste mit Inodes (Einträge für eine Datei, Ordner, etc.)

Daten Bereich für die Nutzerdaten

Mehr über Superblocks: Wikipedia [16]

POSIX

- Portable Operating System Interface
- Verfasst von der IEEE Computer Society
- Beschreibt die Mindestanforderungen für Schnittstellen, u.A.:
  - Hierarchisches Dateisystem
  - Rechteverwaltung (`chmod` und `chown`)
  - Drei Zeitstempel (atime, ctime, und mtime)

Quelle: IEEE [3]

»Durability«

> *Durability is when you write something or change the filesystem and it's still there after the system crashes or loses power unexpectedly. Durability is what you need at a high level to say 'your email has been received' or 'your file has been saved'.*

Quelle: Chris Siebenmann, University of Toronto [11]

Linux PageCache

- Beschleunigung des Lese-Zugriffs
- Geladene Dateien bleiben im **ungenutzten** Arbeitsspeicher für spätere Zugriffe
- Auslesen mit Hilfe von `free -mh` (Spalte **buff/cache**)

```
1                total   used   free   shared   buff/cache   available
2  Mem:          7,6G   2,2G   524M     370M          4,9G          4,9G
3  Swap:           0B     0B     0B
```

Quelle: Linux Page Cache [2]

VFS

- Virtual File System Switch
- Abstraktionsebene zwischen Benutzer und Betriebssystem
- Erledigt u.A. das Laden von benötigten Kernelmodulen
- Verwaltet eine nahtlose, lokal sichtbare Verzeichnisstruktur

Quelle: The Virtual File System [8]

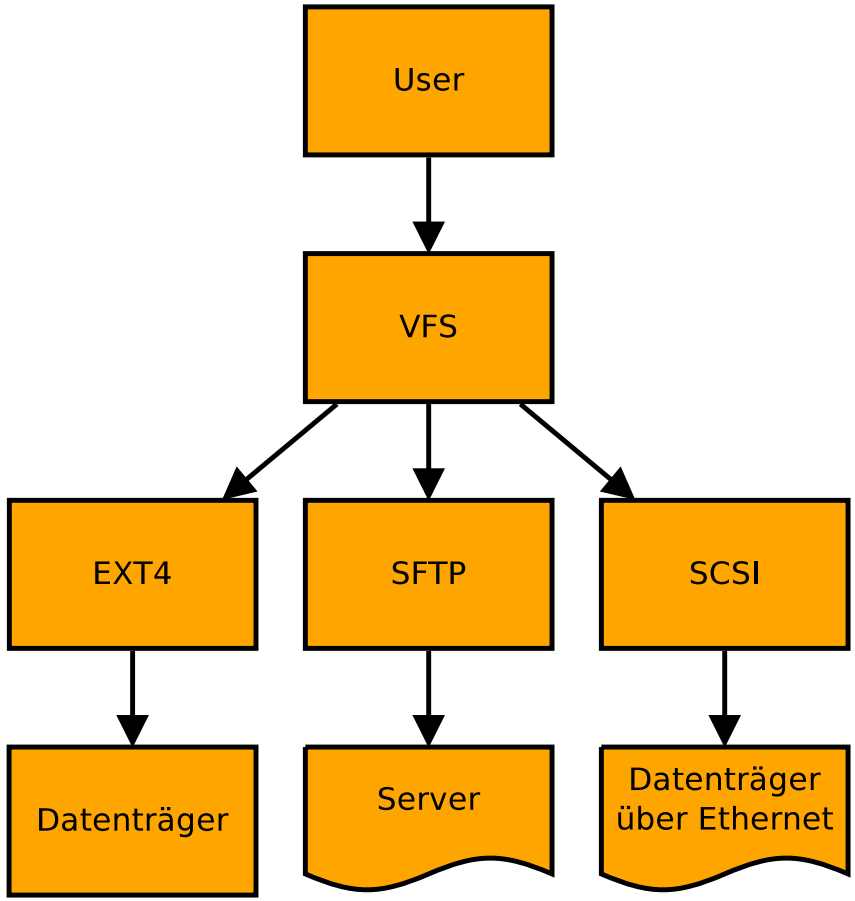

Abbildung: Schema von VFS

Übersicht über Dateisysteme in Linux

»Normale« Dateisysteme	EXT, EXT2, EXT3, EXT4, XFS, BTRFS, ZFS, ReiserFS, Reiser4, JFS, NEXT3, Tux3, ...
Pseudo-Dateisysteme	ramfs, tmpfs, procfs, sysfs, swapfs, ...
Verteile Dateisysteme	OrangeFS, Lustre, BeeGFS, HDFS, Ceph, ...

Beispiel: EXT4

- Wurde 2008 vom Linux-Kernel Entwicklerteam vorgestellt
- Nachfolger von EXT3 und quasi neues Standard-FS bei vielen Linux-Distributionen
- Neue Features:
  - 48 bit Blockadresse
    (2^{48} Blockadressen $\cdot$ 4KiB Blockgröße = 1EiB Partitionsgröße)
  - Extents (Adressierung von großen zusammenhängenden Blöcken)

Quelle: heise Open Source [7]

Erzeugen von EXT4 auf Partition 1

```
1 $ sudo mkfs.ext4 /dev/sdc1
2
3 // Blocksize: 4k (261888 Bloecke)
4 // Inodes:    65536
5 // Journal:   4096 Bloecke
6
7 $ sudo mount /dev/sdc1 /mnt
```

Mehr Informationen unter: `man mke2fs` und `man mount`

Inodes

- Die Inodes bilden das Inhaltsverzeichnis des Dateisystems
- Jede Datei / jedes Verzeichnis hat einen eigenen Inode
- Enthält folgende Metadaten:
  - Typ und Zugriffsrechte
  - Anzahl der Hardlinks
  - Benutzernummer (UID)
  - Gruppennummer (GID)
  - Größe der Datei in Bytes
  - Datum der letzten Veränderung (mtime)
  - Datum der letzten Statusänderung (ctime)
  - Datum des letzten Zugriffs (atime)
  - Adresse von Datenblock 0-9
  - Adresse des ersten Indirektionsblocks
  - Adresse des Zweifach-Indirektionsblocks
  - Adresse des Dreifach Indirektionsblocks

Quelle: Linux-Praxis [5]

Inode auslesen

```
1  // Datei anlegen
2  $ touch Seminar.txt
3
4  // Inode auslesen
5  $ stat Seminar.txt
6    File: 'Seminar.txt'
7    Size: 0     Blocks: 0     IO Block: 4096
8  Device: 821h/2081d  Inode: 12          Links: 1
9  Access: (0644/-rw-r--r--)  Uid: (1000)   Gid: (1000)
10 Context: unconfined_u:object_r:unlabeled_t:s0
11 Access: 2016-04-24 12:00:16.129690659 +0200
12 Modify: 2016-04-24 11:58:58.902414338 +0200
13 Change: 2016-04-24 11:58:58.902414338 +0200
14  Birth: -
```

Journal

- Ein Journal protokolliert alle Änderungen am Dateisystem
- Unterschied zwischen Metadaten- und Full-Journal (protokolliert auch Daten)
- Jeder abgeschlossene Operationenblock wird mit einem »Commit« versehen
- Beim Booten wird das Journal überprüft

Quelle: Journaling Dateisysteme [10]

Rechteverwaltung

- Jede Datei gehört immer einem Benutzer und einer Gruppe (änderbar via `chown`)
- Dateirechte werden mit 4 Oktets beschrieben (Sticky-Bit, Besitzer, Gruppe, Andere) (i.d.R. werden die letzten drei benutzt) (änderbar via `chmod`)

	Besitzer			Gruppe			Andere		
Read	R	W	X	R	W	X	R	W	X
Write	2^2	2^1	2^0	2^2	2^1	2^0	2^2	2^1	2^0
eXecute									

Abbildung: Rechteverwaltung via chmod

Rechteverwaltung

```
1 // Besitzer/Gruppe des Ordners aendern
2 $ chown user:group ./
3
4 // Dateirechte auslesen
5 $ ls -l
6
7 // Nur der Besitzer darf lesen/schreiben
8 $ chmod 600 Seminar.txt
```

Linking

- Anstatt Dateien zu kopieren, kann man sie auch verlinken
- Hardlink: Gleicher Inode wird genutzt (geht nur partitionsintern)
- Symbolic-Link: Pfad zur Datei/zum Ordner wird abgespeichert

```
1 // Hardlink
2 $ ln Seminar.txt Proseminar.txt
3
4 // Link-Counter hat sich erhoeht
5 $ ls -l
6
7 // Symbolic-Link
8 $ ln -s Seminar.txt Oberseminar.txt
```

Beispiel: BTRFS

- Wird seit 2007 von Oracle Corporation entwickelt (steht unter GPL)
- Soll auf lange Sicht EXT4 ablösen
- Bietet viele interessante Features:
  - Bis zu 2^{64} Dateien sind möglich
  - Copy-On-Write
  - Kompression
  - Subvolumes/Snapshots
  - Integriertes RAID

Erzeugen von BTRFS auf Partition 2

```
$ sudo mkfs.btrfs /dev/sdc1 -f

Node size:          16384
Sector size:        4096
Filesystem size:    1.88GiB
Block group profiles:
  Data:             single            8.00MiB
  Metadata:         DUP             104.12MiB
  System:           DUP              12.00MiB
Number of devices:  1
Devices:
   ID        SIZE  PATH
    1     1.88GiB  /dev/sdc1
```

Copy-on-Write

- Bei Änderungen wird der komplette Inhalt an einen anderen Speicherplatz kopiert
- Erst nach dem Kopiervorgang wird ein »Commit« ausgeführt
- Das Risiko für inkonsistente Daten wird minimiert

Kompression

- BTRFS benutzt standardmäßig den LZO-Algorithmus
- Jede Datei wird ohne zusätzlichen Benutzeraufwand komprimiert
- Dem Benutzer steht mehr Speicher zur Verfügung

Subvolumes

> *A btrfs subvolume is not a block device (and cannot be treated as one) instead, a btrfs subvolume can be thought of as a POSIX file namespace. This namespace can be accessed via the top-level subvolume of the filesystem, or it can be mounted in its own right.*

Quelle: Btrfs-Wiki [1]

- Versionierung wird dadurch sehr einfach gemacht (Snapshot)
- Keine konkrete Speicherverteilung im Gegensatz zu einer Partition

Subvolumes

```
1 $ sudo btrfs subvolume create ./Subvolume1
2 Create subvolume './Subvolume1'
3
4 $ sudo btrfs subvolume list ./
5 ID 256 gen 7 top level 5 path Subvolume1
6
7 $ ls -l
8 total 0
9 drwxr-xr-x. 1 root root 0 24. Apr 09:27 Subvolume1/
```

Mehr Informationen unter: `man btrfs`

Quota

- Jede Subvolume kann mit einer Quota begrenzt werden
- Praktisch für Benutzerverzeichnisse o.Ä.

```
1  // Quota-Feature aktivieren
2  $ sudo btrfs quota enable
3
4  // Quota anzeigen
5  $ btrfs qgroup show -r ./
6  qgroupid         rfer         excl     max_rfer
7  --------         ----         ----     --------
8  0/5          16.00KiB     16.00KiB         none
9  0/256        16.00KiB     16.00KiB         none
10
11 // Quota setzen
12 $ sudo btrfs qgroup limit 10M 0/256 ./
```

Deduplikation

- Deduplikation auf Blockebene
- Es gibt zwei Varianten:
  - Während des Schreibvorganges – benötigt sehr viel RAM (Lookup-Table mit Block-Hashes)
  - »Cronjob« – ein Tool vergleicht übergebene Files auf identische Blöcke

Zusammenfassung

- Dateisysteme sind mächtiger als einige meinen ;-)
- Moderne Dateisysteme übernehmen viele Aufgaben von externer Software (RAID, Versionierung, Deduplikation, Kompression, ...)
- Es gibt immer eine Abwägung zwischen Lese-/Schreibgeschwindigkeit und Datensicherheit (Full-Journal, Meta-Journal, kein Journal, CoW, Prüfsummenbildung, ...)
- Wichtig: Dateisystem immer für den spezifischen Einsatzzweck auswählen!

Vielen Dank :-)

Noch Fragen?

[1] Btrfs-Wiki. *SysadminGuide*. 2015. URL: https://btrfs.wiki.kernel.org/index.php/SysadminGuide#Subvolumes (besucht am 24. 04. 2016).

[2] Werner Fischer. *Linux Page Cache*. 2013. URL: https://www.thomas-krenn.com/de/wiki/Linux_Page_Cache (besucht am 21. 04. 2016).

[3] IEEE. *IEEE Std 1003.1*. 2004. URL: http://pubs.opengroup.org/onlinepubs/009695399/utilities/xcu_chap01.html#tag_01_07_01_03 (besucht am 26. 04. 2016).

[4] ITWissen. *Dateisystem*. 2016. URL: http://www.itwissen.info/definition/lexikon/Dateisystem-file-system.html (besucht am 21. 04. 2016).

[5] Linux-Praxis. *Das I-Node System*. 2015. URL: http://www.linux-praxis.de/lpic1/lpi101/inode.html (besucht am 24. 04. 2016).

[6] Samara Lynn. *RAID Levels Explained*. 2014. URL: http://www.pcmag.com/article2/0,2817,2370235,00.asp (besucht am 20. 04. 2016).

[7] heise Open Source. *Extents*. 2009. URL: http://www.heise.de/open/artikel/Extents-221268.html (besucht am 24. 04. 2016).

[8] David A. Rusling. *The Virtual File System*. 1997. URL: http://www.science.unitn.it/~fiorella/guidelinux/tlk/node102.html (besucht am 21. 04. 2016).

[9] Daniel B. Sedory. *An Examination of the Standard MBR*. 2012. URL: http://thestarman.pcministry.com/asm/mbr/STDMBR.htm (besucht am 20. 04. 2016).

[10] SelfLinux. *Journaling Dateisysteme*. URL: http://www.selflinux.org/selflinux/html/dateisysteme_journaling02.html (besucht am 24. 04. 2016).

[11] Chris Siebenmann. *Consistency and durability in the context of filesystems*. 2015. URL: https://utcc.utoronto.ca/~cks/space/blog/tech/FSConsistencyAndDurability (besucht am 26. 04. 2016).

[12] Wikimedia-Commons. *GUID Partition Table Scheme*. 2007. URL: https://en.wikipedia.org/wiki/File:GUID_Partition_Table_Scheme.svg (besucht am 20. 04. 2016).

[13] Wikipedia. *GUID Partition Table*. 2016. URL: https://en.wikipedia.org/wiki/GUID_Partition_Table (besucht am 20. 04. 2016).

[14] Wikipedia. *Master Boot Record*. 2016. URL: https://de.wikipedia.org/wiki/Master_Boot_Record (besucht am 20. 04. 2016).

[15] Wikipedia. *Partition (Datenträger)*. 2015. URL: https://de.wikipedia.org/wiki/Partition_(Datentr%C3%A4ger) (besucht am 20. 04. 2016).

[16] Wikipedia. *Superblock*. 2015. URL: https://de.wikipedia.org/wiki/Superblock (besucht am 26. 04. 2016).