

Systemnahe Programmierung in C (SPiC)

18 Dateisysteme

Jürgen Kleinöder, Daniel Lohmann, Volkmar Sieh

Lehrstuhl für Informatik 4
Verteilte Systeme und Betriebssysteme

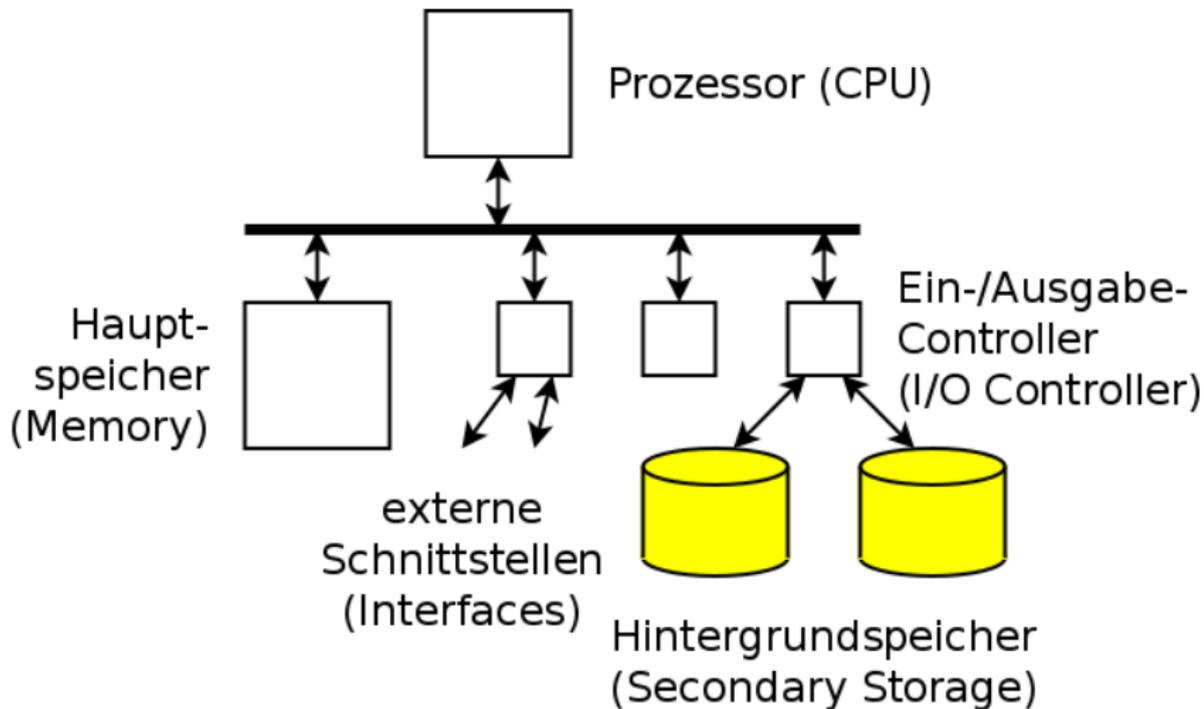
Friedrich-Alexander-Universität
Erlangen-Nürnberg

Sommersemester 2019

http://www4.cs.fau.de/Lehre/SS19/V_SPiC



- Einordnung



- Speichermedien (z.B. Platten, SSD / Flash-Speicher, DVD, CD-ROM) mit Unterschieden; Beispiele
 - Blockgrößen:
 - Festplatten: 512 Bytes/Block
 - CDs: 1024 Bytes/Block
 - Flash: 4096 Bytes/Block
 - Nutzung der Blöcke
 - Flash-Speicher hat nur begrenzte Anzahl von Schreibzyklen pro Block => gleichmäßig beschreiben
 - Festplatten können auf benachbarte Blöcke jeweils schneller zugreifen
 - Größe der Medien (typ.)
 - CD-ROM: ca. 750 MByte
 - DVD: ca. 8,5 GByte
 - Festplatte: ca. 4 TByte
 - SSD: ca. 500 GByte



- Dateisysteme speichern Daten und Programme persistent in Dateien
 - Benutzer muss sich nicht um die Ansteuerung und Verwaltung verschiedener Speichermedien kümmern
 - einheitliche Sicht auf den Hintergrundspeicher
- Wesentliche Elemente eines Dateisystems:
 - Dateien (Files)
 - Verzeichnisse / Kataloge (Directories)
 - Partitionen (Partitions)

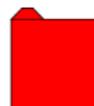


Überblick (4)

- Datei (File)
 - speichert Daten oder Programme
 - enthält Zusatzinformationen
- Verzeichnis / Katalog (Directory)
 - fasst Dateien (u. Verzeichnisse) zusammen
 - erlaubt Benennung der Dateien
 - ermöglicht Aufbau eines hierarchischen Namensraums
- Partition (Partition)
 - eine Menge von Verzeichnissen und deren Dateien
 - sie dienen zum physikalischen oder logischen Trennen von Dateimengen
 - physisch: Festplatte, Diskette
 - logisch: Teilbereich auf Platte oder CD



Datei



Verzeichnis



Partition



- Kleinste Einheit, in der etwas auf den Hintergrundspeicher geschrieben werden kann.
- Unterscheidung:
 - eigentliche Daten (Bild, Text, Programm, ...)
 - Metadaten (Erstellungsdatum, Eigentümer, Zugriffsrechte, ...)

Metadaten / Dateiattribute:

Name: Symbolischer Name, vom Benutzer les- und interpretierbar

- z.B. AUTOEXEC.BAT

Typ: Für Dateisysteme, die verschiedene Dateitypen unterscheiden

- z.B. sequenzielle Datei, zeichenorientierte Datei, satzorientierte Datei

Ort: Wo werden die Daten physisch gespeichert?

- Nummern der Plattenblöcke



Dateiattribute (2)

Größe: Länge der Datei in Größeneinheiten (z.B. Bytes, Blöcke, Sätze)

- steht in engem Zusammenhang mit der Ortsinformation
- wird zum Prüfen der Dateigrenzen z.B. beim Lesen benötigt

Zeitstempel: z.B. Zeit und Datum der Erstellung, letzten Änderung

- für Backup, Entwicklungswerkzeuge, Benutzerüberwachung, ...

Rechte: Zugriffsrechte, z.B. Lese- und Schreibberechtigung

- z.B. nur für den Eigentümer schreibbar, für alle anderen nur lesbar

Eigentümer: Identifikation des Eigentümers

- eventuell eng mit den Rechten verknüpft
- Zuordnung beim Accouting (Abrechnung von Plattenplatz)



■ Erzeugen (Create)

- Nötiger Speicherplatz wird angefordert
- Verzeichniseintrag wird erstellt
- Initiale Attribute werden gespeichert

■ Schreiben (Write)

- Identifikation der Datei
- eventuell Nachfordern von Speicherplatz
- Daten werden auf Platte geschrieben
- eventuell Anpassung der Attribute (z.B. Länge der Datei, Zeitpunkt der letzten Änderung)

■ Lesen (Read)

- Identifikation der Datei
- Daten werden von Platte gelesen
- eventuell Anpassung der Attribute (z.B. Zugriffszeit)



- **Positionieren** des Schreib-/Lesezeigers für die nächste Schreib- bzw. Leseoperation (**Seek**)
 - Identifikation der Datei
 - In vielen Systemen wird dieser Zeiger implizit bei Schreib- und Leseoperationen positioniert
 - Ermöglicht explizites Positionieren
- **Verkürzen (Truncate)**
 - Identifikation der Datei
 - Ab einer bestimmten Position (oder ab Anfang) wird der Inhalt der Datei gelöscht
 - eventuell Freigeben von Speicherplatz
 - Anpassung der Attribute (z.B. Länge der Datei, Zeitpunkt der letzten Änderung)
- **Löschen (Delete)**
 - Identifikation der Datei
 - Entfernen der Datei aus dem Verzeichnis und Freigabe der Plattenblöcke



- Ein Verzeichnis gruppiert Dateien und evtl. weitere Verzeichnisse
- Gruppierungsalternativen
 - Verknüpfung mit Benennung
 - Verzeichnis enthält Namen und Verweise auf Dateien und andere Verzeichnisse (z.B. UNIX, Windows)
 - Gruppierung über Bedingung
 - Verzeichnis enthält Namen und Verweise auf Dateien, die einer bestimmten Bedingung gehorchen:
 - z.B. gleiche Gruppennummer in CP/M
 - z.B. eigenschaftsorientierte und dynamische Gruppierung in BeOS-BFS
- Verzeichnis ermöglicht das Auffinden von Dateien
 - Vermittlung zwischen externer und interner Bezeichnung (Dateiname – Plattenblöcke)



- **Lesen** der Einträge (**Read, Read Directory**)
 - Daten des Verzeichnisses werden gelesen und meist eintragsweise zurückgegeben
- **Erzeugen** und **Löschen** der Einträge erfolgt implizit beim Anlegen bzw. Löschen der Dateien
- **Erzeugen** von Verzeichnissen (**Create, Create Directory**)
- **Löschen** von Verzeichnissen (**Delete, Delete Directory**)

Attribute von Verzeichnissen

- Die meisten Attribute von Dateien treffen auch auf Verzeichnissen zu
 - Name, Ortsinformation, Größe, Zeitstempel, Rechte, Eigentümer, ...



■ Datei

- einfache, unstrukturierte Folge von Bytes
- beliebiger Inhalt; für das Betriebssystem ist der Inhalt transparent
- dynamisch erweiterbar

■ Dateiattribute

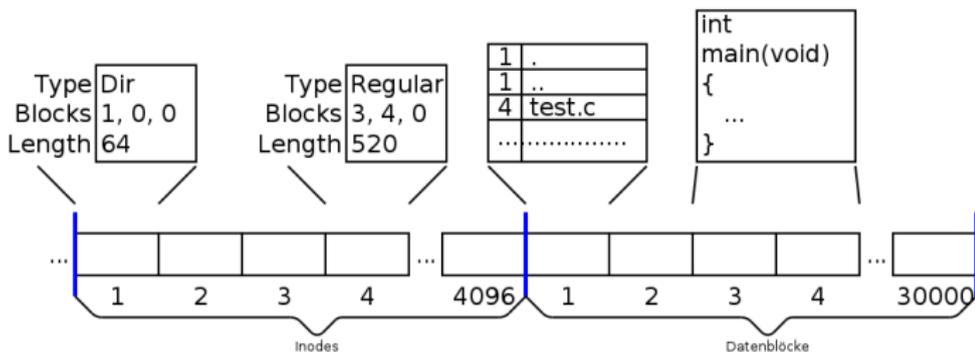
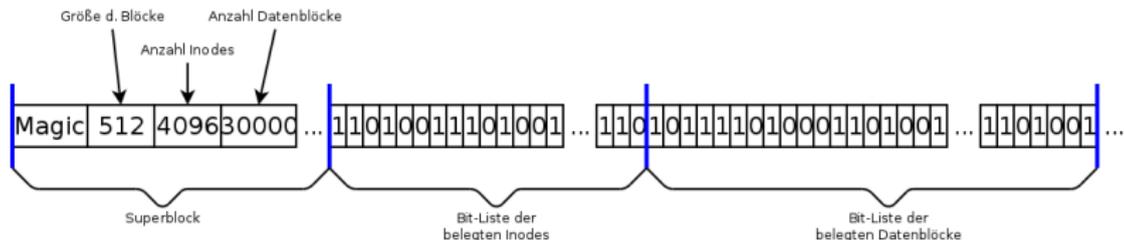
- das Betriebssystem verwaltet zu jeder Datei eine Reihe von Attributen (Rechte, Größe, Zugriffszeiten, Datenblöcke, ...)
- die Attribute werden in einer speziellen Verwaltungsstruktur, dem *Dateikopf*, gespeichert
 - Linux/UNIX: *Inode*
 - Windows NTFS: *Master File Table*-Eintrag

■ Namensraum

- flacher Namensraum: Inodes sind einfach durchnummeriert
- hierarchischer Namensraum: Verzeichnisstruktur bildet Datei- und Pfadnamen in einem Dateibaum auf Inode-Nummern ab



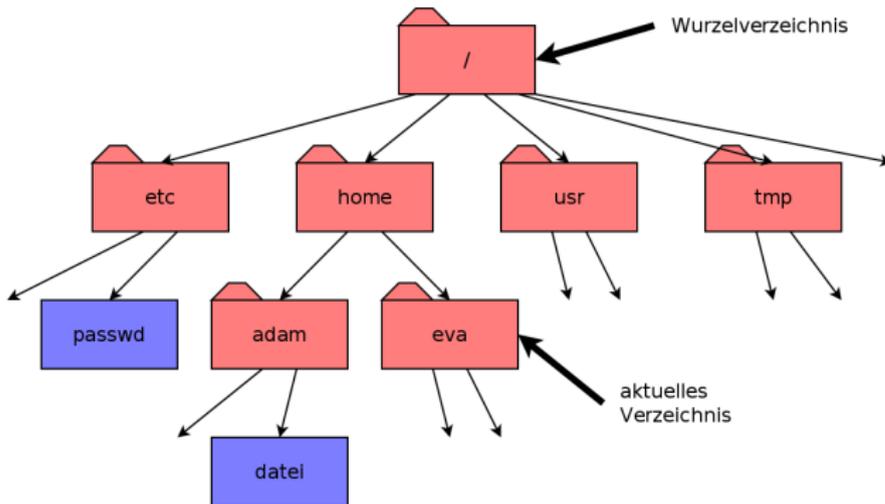
Struktur auf Medium (vereinfacht)



mkfs legt leere Struktur an; fsck überprüft Struktur



■ Baumstruktur



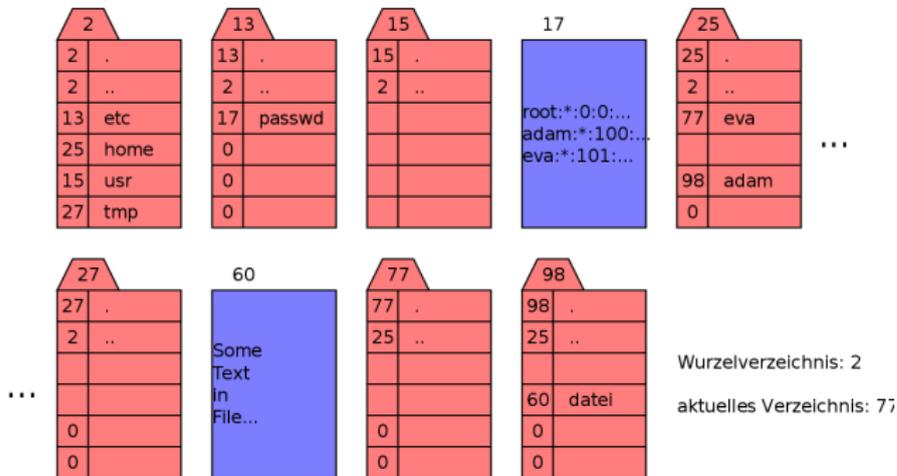
■ Pfade

- z.B. `/home/adam/datei`, `/tmp`, `../adam/datei`
- `/` ist Trennsymbol (*Slash*)
- beginnender `/` ist Wurzelverzeichnis; sonst Beginn implizit mit dem aktuellen Verzeichnis



Pfadnamen (2)

■ eigentliche „Baumstruktur“



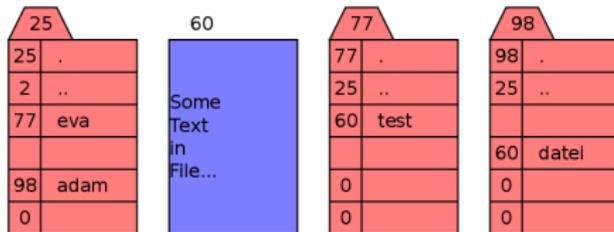
■ Beispiel Pfadauflösung „../adam/datei“:

- 77 + „../adam/datei“ \leadsto 25 + „adam/datei“
- 25 + „adam/datei“ \leadsto 98 + „datei“
- 98 + „datei“ \leadsto 60

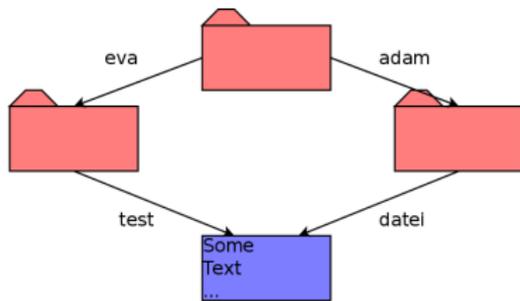


Pfadnamen (3)

- Es können mehrere Verweise (**Hard Links**) auf eine Datei existieren:



aktuelles Verzeichnis: 25



- Beispiel Pfadauflösung „adam/datei“:

- 25 + „adam/datei“ \leadsto 98 + „datei“
- 98 + „datei“ \leadsto 60

- Beispiel Pfadauflösung „eva/test“:

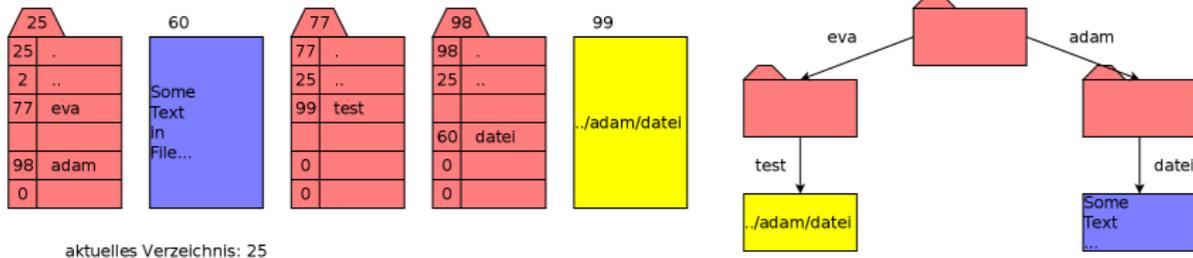
- 25 + „eva/test“ \leadsto 77 + „test“
- 77 + „test“ \leadsto 60

- Datei wird gelöscht, wenn keine Verweise auf sie mehr existieren



Pfadnamen (4)

- Es können mehrere symbolische Verweise (**Symbolic Links**) auf eine Datei oder ein Verzeichnis existieren:



- Beispiel Pfadauflösung „eva/test“:
 - 25 + „eva/test“ \leadsto 77 + „test“
 - 77 + „test“ \leadsto 99 \leadsto 77 + „../adam/datei“
 - 77 + „../adam/datei“ \leadsto 25 + „adam/datei“
 - 25 + „adam/datei“ \leadsto 98 + „datei“
 - 98 + „datei“ \leadsto 60

- Symbolischer Name kann auch bestehen, wenn Datei oder Verzeichnis noch nicht bzw. nicht mehr existiert.



- Eigentümer
 - Jeder Eigentümer wird durch eindeutige Nummer (UID) repräsentiert
 - Ein Benutzer kann einer oder mehreren Benutzergruppen angehören, die jeweils durch eine eindeutige Nummer (GID) repräsentiert werden
 - Eine Datei oder ein Verzeichnis ist genau einem Benutzer und einer Gruppe zugeordnet
- Rechte auf Dateien
 - Lesen, Schreiben, Ausführen (nur vom Eigentümer änderbar)
 - Einzeln für den Eigentümer, für Angehörige der Gruppe und für alle anderen einstellbar
- Rechte auf Verzeichnissen
 - Lesen, Schreiben (Anlegen und Löschen von Dateien/Verzeichnissen), Durchgangsrecht
 - Schreibrecht ist einschränkbar auf eigene Dateien



- Attribute (Zugriffsrechte, Eigentümer, usw.) einer Datei, eines Verzeichnisses werden in **Inodes** gespeichert (vereinfacht):

```
int st_mode;      /* Typ und Zugriffsrechte */
int st_nlink;     /* Anzahl der Hard Links */
int st_uid;       /* Eigentuerer */
int st_gid;       /* Gruppe */
long st_size;     /* Laenge der Datei in Bytes */
int st_block[...]; /* Liste der (indirekten) Bloecke */
time_t st_atime;  /* Letzter Lesezeitpunkt */
time_t st_mtime;  /* Letzter Modifikationszeitpunkt */
time_t st_ctime;  /* Letzte Aenderung an Attributen */
```

- Jede Inode hat eine Nummer und einen Speicherort (Platte/Partition):

```
int st_ino;       /* Inode-Nummer */
int st_dev;       /* Platte/Partition-Nummer */
```



Programmierschnittstelle für Inodes

- stat, lstat liefern Dateiattribute aus Inodes

- Funktionsschnittstelle:

```
#include <sys/types.h>
#include <sys/stat.h>
int stat(const char *path, struct stat *buf);
int lstat(const char *path, struct stat *buf);
```

- Argumente:

- path: Pfadname
- buf: Zeiger auf Puffer, in den Inode-Informationen eingetragen werden

- Rückgabewert

- 0, wenn OK
- -1, wenn Fehler (errno-Variable enthält Fehlernummer)

- Beispiel:

```
struct stat buf;
stat("/etc/passwd", &buf); /* Fehlerbehandlung...! */
printf("Inode-Nummer: %d\n", buf.st_ino);
```



■ Verzeichnisse, Links verwalten

- Erzeugen (eines leeren Verzeichnisses)

```
int mkdir(const char *path, mode_t mode);
```

- Löschen (eines leeren Verzeichnisses)

```
int rmdir(const char *path);
```

- Hard Link anlegen

```
int link(const char *existing, const char *new);
```

- Symbolischen Link anlegen

```
int symlink(const char *existing, const char *new);
```

- Link löschen (und damit ggf. auch Datei)

```
int unlink(const char *path);
```

- Symbolischen Link auslesen

```
int readlink(const char *path, char *buf, int size);
```



Programmierschnittstelle für Verzeichnisse (2)

- Verzeichnisse lesen (Schnittstelle des Linux-Kerns)
 - `open(2)`, `getdents(2)`, `close(2)`
 - Linux-spezifisch und damit nicht portabel
- Verzeichnisse lesen (Schnittstelle der C-Bibliothek)

- Verzeichnis öffnen

```
DIR *opendir(const char *path);
```

- einen Eintrag lesen

```
struct dirent *readdir(DIR *dirp);
```

- Verzeichnis schließen

```
int closedir(DIR *dirp);
```

- Struktur `struct dirent` (vereinfacht)

```
struct dirent {  
    int d_ino;                /* Inode-Nummer */  
    char d_name[NAME_MAX + 1]; /* Name */  
};
```



Verzeichnisse (3): opendir/closedir

- Funktionsschnittstelle:

```
#include <sys/types.h>
#include <dirent.h>

DIR *opendir(const char *path);
int closedir(DIR *dirp);
```

- Argument von opendir:

- path: Verzeichnisname

- Rückgabewert von opendir:

- Zeiger auf Datenstruktur vom Typ DIR, wenn OK
- NULL, wenn Fehler (errno-Variable enthält Fehlernummer)



Verzeichnisse (4): readdir

- Funktionsschnittstelle:

```
#include <sys/types.h>
#include <dirent.h>

struct dirent *readdir(DIR *dirp);
```

- Argument:

- `dirp`: Zeiger auf `DIR`-Datenstruktur

- Rückgabewert:

- Zeiger auf Datenstruktur vom Typ `struct dirent`, wenn OK
- `NULL`, wenn Verzeichnis zu Ende gelesen wurde (`errno`-Variable nicht verändert)
- `NULL`, wenn Fehler aufgetreten ist (`errno`-Variable enthält Fehlercode)

- Hinweis: Der Speicher für `struct dirent` wird u.U. beim nächsten `readdir`-Aufruf überschrieben!



Verzeichnisse (5): Beispiel

- Ausgabe der Dateinamen im aktuellen Verzeichnis:

```
#include <sys/types.h>
#include <dirent.h>

DIR *dirp;
struct dirent *de;
int ret;

dirp = opendir(".");           // akt. Verz. oeffnen
if (dirp == NULL) ...       // Fehler

while (1) {
    errno = 0;
    de = readdir(dirp);      // Eintrag lesen
    if (de == NULL && errno != 0) ... // Fehler
    if (de == NULL) break;   // Ende erreicht

    printf("%s\n", de->d_name);
}

ret = closedir(dirp);        // Verz. schliessen
if (ret < 0) ...            // Fehler
```



- Funktionsschnittstelle:

```
#include <sys/types.h>
#include <fcntl.h>

int open(const char *path, int flags, ...);

int close(int fd);

ssize_t read(int fd, void *buf, size_t count);
ssize_t write(int fd, const void *buf, size_t count);
```



■ Kopierprogramm

```
#include <fcntl.h>

int ret;

int src_fd = open("src", O_RDONLY);
if (src_fd < 0) ... // Fehler
int dst_fd = open("dst", O_CREAT | O_TRUNC | O_WRONLY, 0777);
if (dst_fd < 0) ... // Fehler

while (1) {
    char buf[1024];
    len = read(src_fd, buf, sizeof(buf));
    if (len < 0) ... // Fehler
    if (len == 0) break;
    ret = write(dst_fd, buf, len);
    if (ret < 0) ... // Fehler
}

ret = close(dst_fd);
if (ret < 0) ... // Fehler
ret = close(src_fd);
if (ret < 0) ... // Fehler
```



- **write-Aufruf** muss
 - den File-Deskriptor überprüfen (Datei geöffnet, Datei beschreibbar?)
 - die Pufferadresse/-länge überprüfen
 - den/die zu beschreibenden Blöcke des Mediums ermitteln
 - den/die Blöcke vom Medium lesen (wenn nicht ganzer Block geschrieben wird)
 - die entsprechenden Bytes im gelesenen Block überschreiben
 - den/die Blöcke auf das Medium zurückübertragen
 - die Attribute anpassen (Datum letzte Modifikation, Länge der Datei)
 - und ist ein Betriebssystem-Aufruf
- => **write** ist eine zeitlich teure Operation (**read** entsprechend)!
- => Besser: viele Bytes (am Besten: Vielfache der Blockgröße) am Stück lesen/schreiben
- => **fopen-, fclose-, fread-, fwrite-, getchar-, putchar-, fscanf-, fprintf-, ...** -Funktionen aus der C-Bibliothek benutzen!



- Periphere Geräte (Platte, Drucker, CD, Terminal, Scanner, ...) werden als Spezialdateien repräsentiert (`/dev/sda`, `/dev/lp0`, `/dev/cdrom0`, `/dev/tty`, ...)
- in Inode steht
 - Typ:
 - Block-orientiertes Gerät (Platte, CD, DVD, SSD, ...)
 - Zeichen-orientiertes Gerät (Drucker, Terminal, Scanner, ...)
 - statt Blocknummern:
 - Major-Number: Typ des Gerätes (Platte, Drucker, ...)
 - Minor-Number: Nummer des Gerätes (3. Drucker, 5. Terminal, ...)
- Öffnen der Geräte schafft eine (evtl. exklusive) Verbindung zum Gerät, die durch Treiber hergestellt wird
- Geräte können dann mit `read-`, `write-` und `ioctl-`Operationen angesprochen werden



■ Ausgabe auf Drucker

```
#include <linux/lp.h>
int fd, ret;

/* Verbindung zum Drucker 0 herstellen. */
fd = open("/dev/lp0", O_WRONLY);
if (fd < 0) ...

/* Druckerstatus abfragen. */
ret = ioctl(fd, LPGETSTATUS, &state);
if (ret < 0) ...
if (state & LP_POUTPA) {
    fprintf(stderr, "Out of paper!\n"); exit(1);
}

/* Auf Drucker schreiben. */
ret = write(fd, "Hallo, Drucker!\n\f", 17);
if (ret < 0) ...

/* Verbindung abbauen. */
ret = close(fd);
if (ret < 0) ...
```



- jede Festplatte kann als Ganzes ein Dateisystem enthalten
 - Festplatte entspricht dann einer Partition
- jede Festplatte kann aber auch unterteilt werden in mehrere Partitionen
 - erster Block der Platte enthält Partitionstabelle
 - Partitionstabelle enthält Informationen
 - wieviele Partitionen existieren
 - wie groß die jeweiligen Partitionen sind
 - wo sie beginnen
- jede Partition
 - wird durch eine Spezialdatei repräsentiert; z.B.
 - /dev/sda, /dev/sdb (ganze Platte)
 - /dev/sda1, /dev/sda2, /dev/sdb1 (Teile der jeweiligen Platte)
 - enthält eigenes Dateisystem

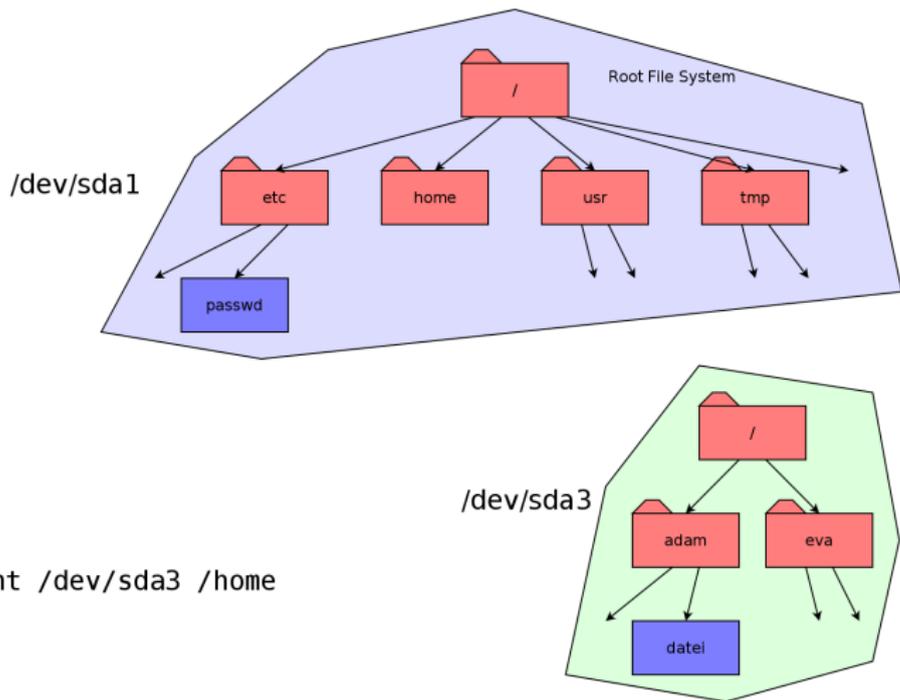


- Bäume der Partitionen können zu einem homogenen Dateibaum zusammengesetzt werden (Grenzen für Anwender nicht sichtbar!)
 - „Montieren“ von Dateibäumen (*mounting*)
- Ein ausgezeichnetes Dateisystem ist das *Root File System*, dessen Wurzelverzeichnis gleichzeitig das Wurzelverzeichnis des Gesamtsystems ist
 - Andere Dateisysteme können mit dem `mount`-Befehl in das bestehende System hineinmontiert werden bzw. mit dem `umount`-Befehl wieder entfernt werden.
 - Über das *Network File System* (NFS) können auch Verzeichnisse anderer Rechner in einen lokalen Dateibaum hineinmontiert werden => Grenzen zwischen Dateisystemen verschiedener Rechner werden unsichtbar



Montieren des Dateibaumes

■ Beispiel



Montieren des Dateibaumes (2)

- Nach Ausführung des Montierbefehls

