

VO Operating Systems

20.06.2025

Matti Fischbach
matti.fischbach@web.de

*Erstellt mit den hervorragenden Foliensätzen von Peter Thoman
und den Foliensätzen von Herr Radush,
die die Lehrveranstaltung Betriebssysteme
an der Universität Innsbruck leiten.*

Jegliche Informationen stammen aus den Foliensätzen der Lehrveranstaltung
und demnach der zugrundeliegenden Quelle:
Operating System Concepts (Tenth Edition)
von Abraham Silberschatz, Peter Baer Galvin und Greg Gagne

Der Sourcecode zu dem Script kann auf Github unter dem Link
<https://github.com/jqyDee/vo-os-script>
eingesehen werden!

Contents

1	Betriebssystem	6
1.1	Definition:	6
1.2	Ziele:	6
2	Struktur eines Computersystems	6
2.1	Hardware	6
2.2	Anwendungsprogramme	6
2.3	Benutzer	6
2.4	Userinterface	6
3	Aufbau	6
3.1	Kernel	6
3.2	System services	7
3.3	Middleware	7
3.4	Organisation	7
3.5	Betriebssystem Ablauf	7
4	I/O Operation	8
5	Interrupts und Interrupt based Systems	8
5.1	direct Memory-Access	9
5.2	Dualmode	9
5.3	Timer	10
6	Computer Organisation	10
6.1	Speicherhierarchie	10
6.2	Von Neumann	11
6.3	Gemeinsame Speichersysteme	11
6.4	Clustersysteme	11
6.5	Multiprozess Programmierung und Multitasking	11
7	Ressourcen Verwaltung	12
7.1	Prozessverwaltung	12
7.2	Speicherverwaltung (flüchtig)	12
7.3	Dateisystemverwaltung (nicht flüchtig)	12
7.4	Massenspeicherverwaltung	12
7.5	Cacheverwaltung	12
7.6	Zusammenfassung	13
7.7	Schutz und Sicherheit	13
7.8	Virtuallisierung	13
7.9	Verteilte Systeme	13
8	Systemcalls	13
8.1	POSIX	14
8.2	Systemcall Arten	14
8.3	Microkernel	15
9	Prozesskonzept	15
9.1	Prozess	15
9.2	Speicherstruktur	15
9.3	Zustand	15
9.4	Prozesskontrollblock	16
9.5	Linux	16
10	Prozessoperationen	17

10.1	Prozesserstellung	17
10.2	Prozessterminierung	17
11	Threads	18
11.1	Single vs. Multithreaded Programme	18
11.2	Anwendung (Server)	18
11.3	Concurrency vs. Parallelism	19
11.4	Parallelism	19
11.5	Data vs Task Parallelism	19
11.6	User- und Kernelthreads	19
11.7	Multithreading Modelle	20
11.8	PThread (POSIX Threads)	20
11.9	fork und exec	21
11.10	Signals und Interrupts	21
11.11	Abbruch	21
12	Interprozess Kommunikation (IPC)	21
12.1	Fundamentals Model	22
12.2	Messagepassing	22
12.3	Unnamed Pipe	23
12.4	Named Pipe (FIFO)	24
12.5	Messagequeue	24
12.6	Shared Memory	26
13	Synchronisation	28
13.1	Data Hazards	29
13.2	Critical Sections	29
13.3	Atomics	29
13.4	Atomic Operationen	30
13.5	Shared Memory mit Synchronisation	30
13.6	Mutex	30
13.7	Dekker's Algorithmus	31
13.8	Peterson's Algorithmus	32
13.9	Memorybarrier	33
13.10	Deadlocks	33
13.11	Bankieralgorithmus	34
13.12	Deadlock Detection	35
13.13	Deadlock Behebung	36
13.14	Livelocks	36
13.15	Condition Variables	37
13.16	Semaphores	38
13.17	Ring Buffer	39
13.18	Producer - Consumer Problem	40
13.19	Dining Philosophers Problem	41
13.20	Barriers	42
13.21	Interrupt-based Synchronisation	43
14	Synchronisation (Hardware)	43
14.1	Spinlock / Busy waiting	44
14.2	Semaphoren	45
14.2.1	Implementierung mittels Spinlock:	45
14.2.2	Implementierung ohne Spinlock:	45

14.3	Monitor und Condition Konstrukt (Java)	46
15	Alternative Ansätze der Synchronisation	48
15.1	Transactional Memory	48
16	Input/Output (I/O)	49
16.1	Architektur	49
16.2	Speicher	49
16.3	I/O Bus	50
16.4	Device Communication	50
16.4.1	Beispiel <i>Memory Mapped I/O</i> an einem Arduino Uno ATmega328P:	51
16.4.2	Beispiel <i>DMA-Buffer</i>	51
17	Speicher Hardware und Software	52
17.1	HDD	52
17.2	SSD	53
17.3	NAND und NOR Flash	53
17.4	RAID	53
17.5	RAID Standard Levels	54
17.5.1	RAID 0	54
17.5.2	RAID 1	54
17.5.3	RAID 2	55
17.5.4	RAID 3	55
17.5.5	RAID 4	55
17.5.6	RAID 5	56
17.5.7	RAID 6	56
17.6	RAID Hybrid Levels	57
17.6.1	RAID 01 / RAID 10	57
17.6.2	RAID 50	57
18	Filesystems und Partitioning	57
18.1	Partitiontable	58
18.1.1	MBR	58
18.1.2	GPT	58
18.2	FAT Filesystem	58
18.3	Journaling Filesystem	58
18.4	Multi-Disk Filesystem	59
19	CPU Scheduling / Planung	59
19.1	Contextswitch	60
19.2	Scheduling Typen	60
19.3	Scheduling Kriterien	60
19.4	CPU-Burst Prediction	61
19.5	Scheduling Algorithmus - First come First server (FCFS)	62
19.6	Scheduling Algorithmus - Shortest process first	62
19.7	Scheduling Algorithmus - Process with shortest remaining first	63
19.8	Scheduling Algorithmus - Round-Robin Präemptiv	64
19.9	Scheduling Algorithmus - Prioritäts Planung	65
19.10	Multilevel Queue Scheduling	66
19.11	Multilevel Feedback Queue Scheduling	66
20	Thread Scheduling / Planung	67
21	Multiprocessor Scheduling / Planung	68
21.1	Multicore Prozessoren	68

21.2	Load Balancing	69
21.3	Processor Affinity	69
22	Real-Time CPU Scheduling / Planung	69
22.1	Priority-based Scheduling	70
22.2	Rate-Monotonic Scheduling	70
22.3	Earliest-Deadline-First Scheduling	71
22.4	Proportional Share Scheduling	71
22.5	POSIX Real-Time Scheduling	71
22.6	Linux Scheduling	72
23	Compiling, Linking, Loading and Libraries	73
23.1	Compiling	73
23.2	Ausführung eines Programms	73
23.3	Linker	74
23.4	Metadaten in Ausführbaren Dateien	77
23.4.1	Tool <code>nm</code>	77
23.4.2	Tool <code>objdump</code>	78
23.5	Name Mangling	78
23.6	Linker Typen und Libraries	79
23.6.1	Typ 0 - dynamic Linking	79
23.6.2	Typ 1 - static shared Libraries	79
23.6.3	Typ 2 - dynamic shared Libraries	79
23.7	Code = Data	80
23.8	Alternative Linker	81
24	Virtualisierung	82
24.1	Computer Stack	82
24.2	Simulation	82
24.3	Emulation	83
24.4	Virtualisierungs	83
24.5	Hypervisor	84
24.5.1	Hypervisor - Typ 1	84
24.5.2	Hypervisor - Typ 2	84
24.6	Volle Virtualisierung	84
24.7	Virtualisierung Implementation	85
24.7.1	Virtueller Kernel Modus:	85
24.7.2	Virtuelle CPU (VCPU)	85
24.7.3	Trap and Emulate	86
24.7.4	Binary Translation	86
24.7.5	Shadow Page Tables	87
24.8	Paravirtualisierung und Hardware Support	87
24.9	VM Operationen	87
24.10	Betriebssystem Virtualisierung	87
25	Memory Management	87

1 Betriebssystem

1.1 Definition:

- Software zur Verwaltung der Computerhardware
- Low-Level Grundlage der Anwendungsprogramme
- Interface zwischen User und Hardware

1.2 Ziele:

- Ausführen von User-Programmen
- Benutzerfreundliche/einfache Lösung von Problemen (*higher level programming*)
- effiziente Nutzung der Hardware

2 Struktur eines Computersystems

2.1 Hardware

grundlegende Ressourcen wie:

- CPU,
- GPU,
- RAM,
- usw.

2.2 Anwendungsprogramme

- Verwenden Ressourcen der Hardware um Probleme zu Lösen
- Programme wie Textverarbeitung, usw.

2.3 Benutzer

- Mensch, Maschinen bzw. andere Computer

Das Betriebssystem stellt wie bereits angesprochen die Schnittstelle zwischen Benutzer und Hardware bereit. Dabei stellt es eine angemessene Performance, security und Benutzerfreundlichkeit bereit. Es kann in Ein- bzw. Mehrfachbenutzergeräten und in Embeddedsystems verwendet werden. Hier stellt es die Hardware Ressources und die Kontrolle für die einzelnen Prozesse bereit.

2.4 Userinterface

- CLI (Command Line Interface)
- GUI (Graphical User Interface)

3 Aufbau

3.1 Kernel

- Startet mit als erstes beim Einschalten des Computers
- läuft im Hintergrund bis zum Abschalten des Systems

Es wird zwischen Präemptiven und Kooperativen (nicht präemptiven) Kernels unterschieden:

- **Präemptiv:**

Ein laufender Prozess kann von dem Betriebssystem unterbrochen werden, auch wenn er nicht *freiwillig* den Prozessor freigibt. Dies ist notwendig, wenn eine Prozess mit höherer Priorität bereit ist auf dem Prozessor zu laufen.

- **Kooperativ:**

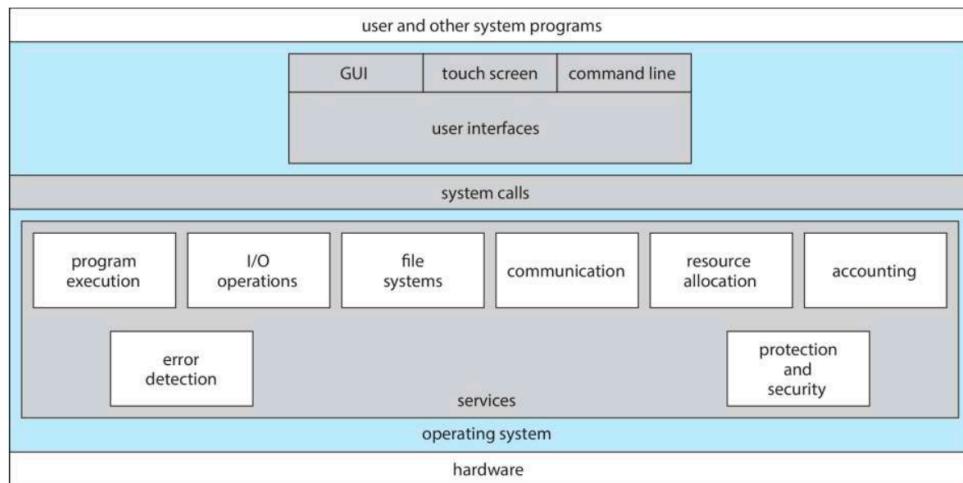
Ein laufender Prozess kann **nicht** von dem Betriebssystem unterbrochen werden. Dieser muss *freiwillig* den Prozessor verlassen. Dies bringt eine langsamere Reaktionszeit mit sich, zugrunde einer geringeren Komplexität. Bei schlechter Programmierung kann ein Prozessor langfristig *verstopfen*.

3.2 System services

- Systemprogramme und Deamons (*im Hintergrund laufende Prozesse, welche meist für den Benutzer nicht direkt sichtbar sind*)
- Treiber
- Systembibliotheken (*Libraries*)
- GUI
- CLI Schnittstelle

3.3 Middleware

- erleichtern die Anwendung (.NET, ...)
- Datenbanken
- Grafik (X-Window System, ...)



3.4 Organistation

Die CPU und der Gerätekontroller ist durch den Systembus mit dem Speicher (RAM) verbunden. Die Treiber der einzelnen Geräte ist die Schnittstelle zwischen der wirklichen Hardware bzw. dem Gerätekontroller und dem restlichen System über den Systembus. Dadurch wird ein gemeinsamer Zugriff auf den Speicher ermöglicht. Durch Interrupts wird eine gleichzeitige Ausführung von der CPU und den Geräten ermöglicht.

3.5 Betriebssystem Ablauf

Bootstrapping:

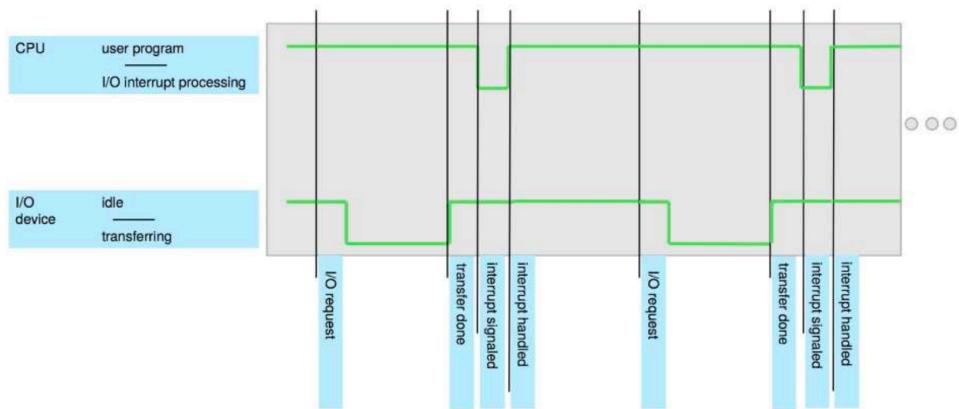
1. Bios: (ROM)
2. Systeminitialisierung

3. Kernel wird geladen
4. Start der Systemdienste

4 I/O Operation

Das Programm stellt eine I/O Request über das Betriebssystem an den Gerätetreiber und lädt die benötigten Register. Der Gerätekontroller prüft die Register und führt die angefragte Aktion aus (*z.b. Zeichen von Tastatur lesen*). Die Daten werden in einem lokalen Buffer gespeichert und der Kontroller informiert über den Gerätetreiber das die Aktion abgeschlossen wurde. Nun findet ein Interrupt ausgelöst vom Systembus statt, was die Kontrolle an das System zurückgibt. Die Daten (und Statusinformationen) können gelesen werden.

Timeline:



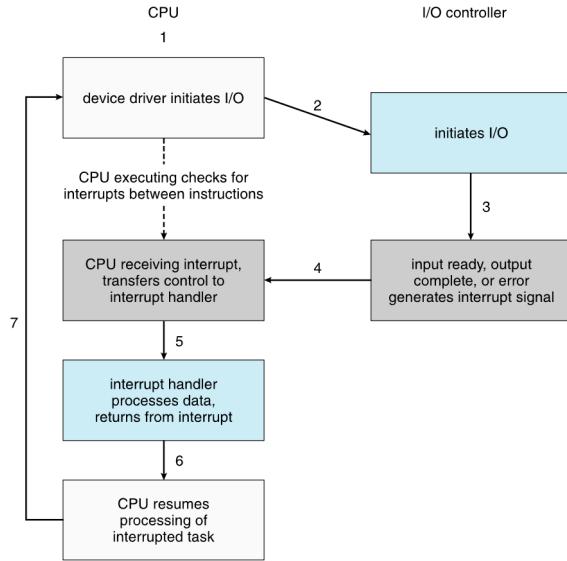
5 Interrupts und Interrupt based Systems

Werden vom Gerätekontroller ausgelöst. Nach jedem CPU Cycle wird überprüft ob ein Interrupt vorliegt. Der Interrupt wird von der CPU abgefangen und an einen Interrupthandler weitergeleitet. Der Interrupthandler speichert die Adresse des Unterbrochenen Befehls und die Register der CPU. Weiter wird der Interrupt verarbeitet. Schlussendlich wird der Zustand vor dem Interrupt wieder hergestellt. Jeder Interrupt hat einen Interruptvektor. Dieser beinhaltet die Interrupt-Zahl, die Adresse des Interrupthandlers und die Interruptchain. Es gibt *Unmaskierte* (*nicht ausschaltbar, z.b. Speicherfehler*) und *Maskierte* (*ausschaltbar*) Interrupts. Interrupts haben unterschiedliche Prioritäten.

Interrupttypen:

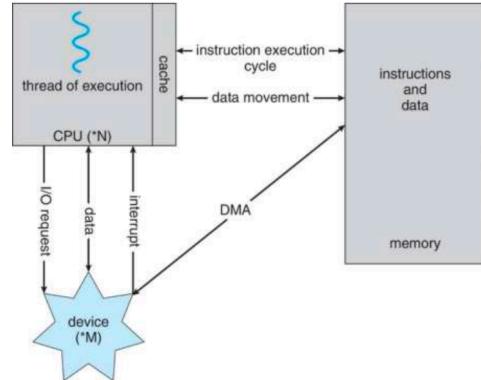
- Hardwareinterrupts (Geräte)
- Softwareinterrupts (Exception, Trap, Error, Syscall, Segfault, ...)

Interrupt-cycle (I/O):



5.1 direct Memory-Access

Zwischen nichtflüchtigem Speicher und dem Gerätekontroller. Wird genutzt um Daten in Blöcken direkt vom Buffer des Kontrollers im Hauptspeicher abzulegen. Der Abschluss wird der CPU durch einen Interrupt kommuniziert. Es findet eine Unterbrechung pro **Block** und **NICHT** pro **Byte** statt. Die CPU ist nicht beteiligt.



5.2 Dualmode

Bietet verbesserte Security. Es wird ein **Modus-Bit** benutzt um zu signalisieren ob der:

- Kernelmode (0),

oder der

- Usermode (1)

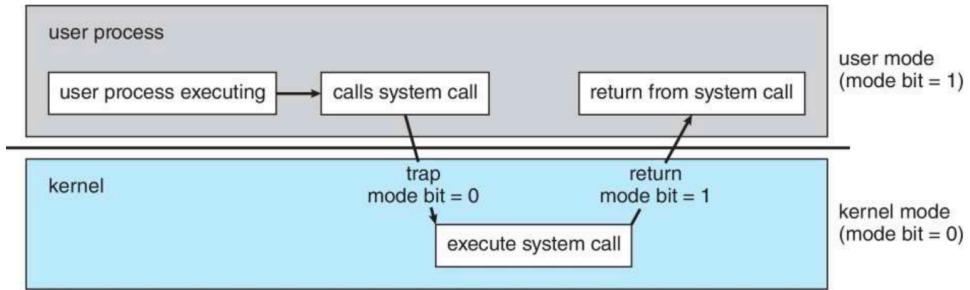
verwendet wird.

Kernelmode: (auch Privilegedmode, Systemmode, Supervisormode)

Startet den Kernel, ist verantwortlich für Systemaufrufe und Interrupts. Lässt Priviligierte Anweisungen zu.

Usermode:

- Usercommands-/code



5.3 Timer

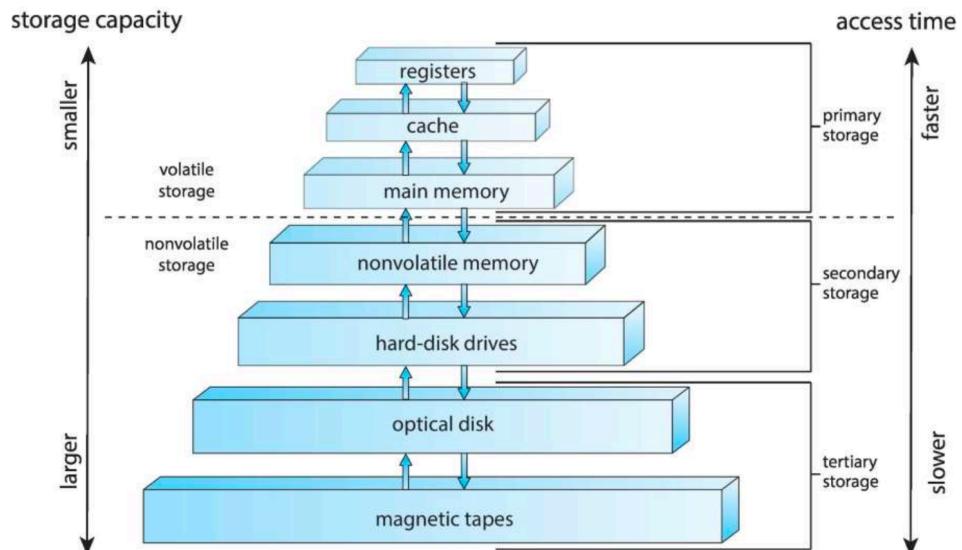
Timer werden genutzt um die Beanspruchung von Ressourcen für bestimmte Prozesse zu begrenzen. Nach einer bestimmten Zeit wird ein Interrupt getriggert um die Kontrolle wieder zu erlangen. Dies kann verhindern, dass zum Beispiel Endlosschleifen die CPU Ressourcen blockieren. Dies kann durch privilegierte Anweisungen unterbunden werden.

Fixed Timer: fixe Zeitspanne (z.B. $\frac{1}{60}$ sec)

Variable Timer: Clockfrequency oder Counter

6 Computer Organisation

6.1 Speicherhierarchie

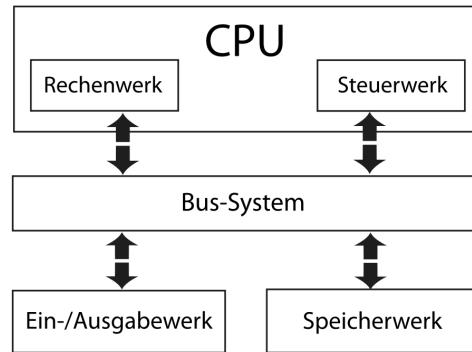


6.2 Von Neumann

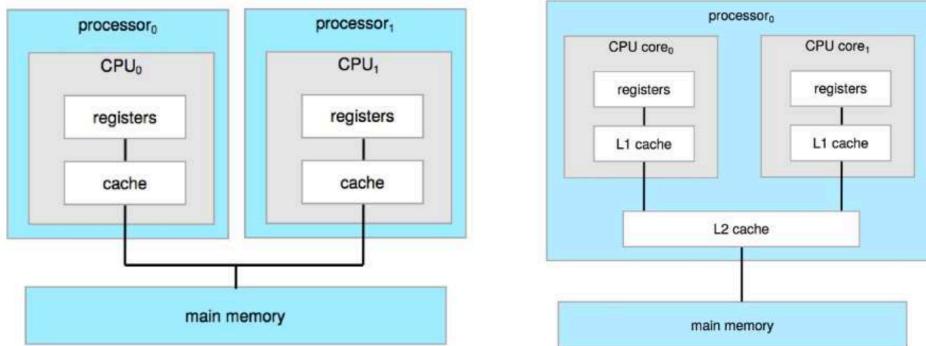
- Bus-System
- Memory

Der Hauptspeicher (RAM) ist das einzige von der CPU direkt zugreifbare Speichermedium. Wird meist als DRAM implementiert und ist flüchtig und wieder beschreibbar.

Der Sekundärspeicher wird in einzelne logische Sektoren unterteilt die vom Speicherkontroller dann die Interaktion mit dem Computer ermöglicht.



6.3 Gemeinsame Speichersysteme



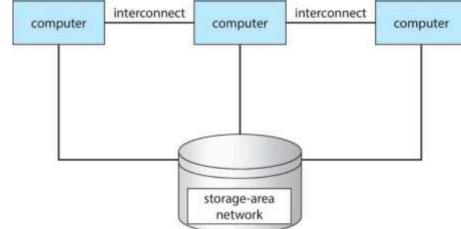
6.4 Clustersysteme

High availability:

- Error tolerant
- (A-)Synchronous Clustering
- Load splitting

High performance:

- Parallel
- speed



6.5 Multiprozess Programmierung und Multitasking

Mehrere Benutzer können System parallel benutzen. Dabei können Programme(/Prozesse) gleichzeitig rechnen. Dies ist besonders nützlich bei Planung, Warteschlangen, Prioritäten, und generellem Lastausgleich, da die Ressourcennutzung maximiert wird. Auch bei jeglicher I/O ist eine Multiprocessing Ansatz sinnvoll, da Anfragen *gleichzeitig* bearbeitet werden.

Multiprocessing und *Multitasking* sind unterschiedlich, da hier auf einem 1 Benutzersystem mehrere Prozesse/Anwendung gleichzeitig laufen.

7 Ressourcen Verwaltung

7.1 Prozessverwaltung

Prozess: Ist ein Programm in Ausführung und eine Arbeitseinheit in dem System. Die Ressourcen (CPU, Memory, I/O) stehen dem Prozess zur Verfügung und werden nach der Beendigung freigegeben.

Betriebssystem: Ist verantwortlich für das erstellen, löschen und unterbrechen der Prozesse, sowie der Zuordnung der Ressourcen. Es ist außerdem verantwortlich für die Synchronisation und der Inter-Prozesskommunikation. Außerdem ist es verantwortlich für die Deadlock Behbung/Behandlung.

Singlethread Prozess: Anweisungen werden sequentiell ausgeführt und in einem Programmcounter wird der jeweilig nächste Schritt/Anweisung gespeichert.

Multithread Prozess: Hier hält jeder Prozess seinen eigenen Programmcounter. Die Programmierung muss parallelisiert werden

Andere: Userprozesse, OS-Prozesse, usw.

7.2 Speicherverwaltung (flüchtig)

Inhalt: Die Gesamtheit bzw. Teile des Programmcodes/Maschinencodes. Nötig um das Programm schlussendlich auszuführen. Außerdem alle weiteren Daten die das Programm benötigt, wie z.b. heap, stack, filehandles usw.

Das Betriebssystem verfolgt die Speicherverwendung und ist zuständig für Speicher Zuordnung und Freigabe. Außerdem ist es zuständig für die Koordination vom Laden und Freigeben von Daten aus dem Sekundärspeicher.

7.3 Dateisystemverwaltung (nicht flüchtig)

Daten werden häufig in Dateisystemen gespeichert. Sie bieten eine einheitliche und logische Ansicht der Informationsspeicherung. Eine Datei ist hier eine logische abstrakte Einheit. Die Daten des Dateisystems werden in der Regel auf dem Sekundär- bzw. Tertiärspeicher gespeichert.

Das Betriebssystem ist zuständig für die Verwaltung und Organisation, der Zugriffskontrolle, dem Erstellen und Löschen, dem Bearbeiten und dem Zuordnen im nichtflüchtigen Speicher.

7.4 Massenspeicherverwaltung

Weiterhin werden Daten oft temporär(USB-Stick) oder langfristig(Backup-Drive) gespeichert. Dies ist entscheidend für die Geschwindigkeit des Computerbetriebs, da mehr Daten = mehr Arbeit für das System.

Das Betriebssystem ist zuständig für das Mounten, Unmounten, Freespace managing, Zuteilung, I/O Planung, Partitionierung und dem Schutz der Datenträger bzw. Daten.

7.5 Cacheverwaltung

Der Cachespeicher ist ein vorübergehender Speicher für das Speichern bei Datenkopien. Der Cache ist meist wesentlich kleiner als der Speicher(RAM) und besteht aus Registern, Befehlscache und Datencache.

Die Hardware ist für die Verwaltung der Cachezeilen, Cachetreffer(Lokalität), Cachefehler, Cachekohärenz(Multiprocessing) und der Cacheersatzrichtlinien zuständig.

7.6 Zusammenfassung

Level	1	2	3	4	5
Name	registers	cache	main memory	solid-state disk	magnetic disk
Typical size	< 1 KB	< 16MB	< 64GB	< 1 TB	< 10 TB
Implementation technology	custom memory with multiple ports CMOS	on-chip or off-chip CMOS SRAM	CMOS SRAM	flash memory	magnetic disk
Access time (ns)	0.25-0.5	0.5-25	80-250	25,000-50,000	5,000,000
Bandwidth (MB/sec)	20,000-100,000	5,000-10,000	1,000-5,000	500	20-150
Managed by	compiler	hardware	operating system	operating system	operating system
Backed by	cache	main memory	disk	disk	disk or tape

Die Übertragung ist in der Reihenfolge:

secondary mem → main memroy(RAM) → cache → hardware registers

7.7 Schutz und Sicherheit

Das Betriebssystem schützt die Daten mit Zugriffskontrolle (*only specific Users and Groups*), Verteidigung gegen Angriffe bzw. unerwünschten Zugriff.

7.8 Virtualisierung

“Betriebssystem in Betriebssystem”

Oft auch als Emulation bekannt, wenn ein anderes (Gast-)Betriebssystem in einem (Host-)Betriebssystem genutzt wird. Nützlich um Programme auf anderen Plattformen zu testen bzw. zu entwickeln.

7.9 Verteilte Systeme

Systeme die durch eine Schnittstelle miteinander verbunden sind. Meistens über TCP/IP bzw. über LAN, WAN, MAN oder PAN. Die Kommunikation findet durch Datenaustausch statt. Hier gibt es sogenannte Netzwerkbetriebssysteme die die Illusion eines einzigen Systems mit vielen einzelnen Computern simuliert.

Beispiele: Traditionell (Server, Client); Peer-To-Peer (Clients that form a system); Cloud Computing; usw.

8 Systemcalls

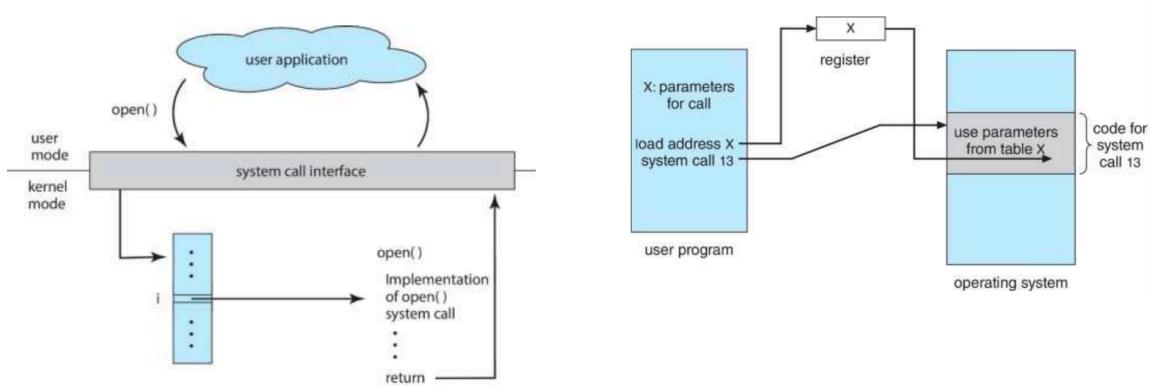
Systemcalls bieten eine Schnittstelle zwischen dem Programmierer und dem Betriebssystem. Sie ermöglichen Low-Level Befehle einfach in Higher-Level Programmiersprachen zu verwenden. Beispiel ist die Win32-API auf Windows oder die POSIX-API für UNIX. Aber auch die Java-API für die Java Virtuelle Maschine.

Für die Verwendung von Systemcalls werden Register und der Stack genutzt um Daten bzw. Parameter an den jeweiligen Aufruf zu übergeben.

Unter einem Modularen Betriebssystem werden die Systemcalls zusätzlich zum Betriebssystem gespeichert. Somit ist die Schnittstelle Modular und kann nach belieben angepasst werden.

8.1 POSIX

POSIX ist eine Standardisierte Programmierschnittstelle zum Betriebssystem UNIX. Sie folgt dem ISO/IEC/IEEE 9945 Standard. Die `libc` Bibliothek folgt dem POSIX Standard. UNIX alleine bietet Basis Definitionen, Konventionen und Konzepte. Sie ist die schlussendliche System-Schnittstelle mit C-Systemaufrufen und Header-Dateien.



8.2 Systemcall Arten

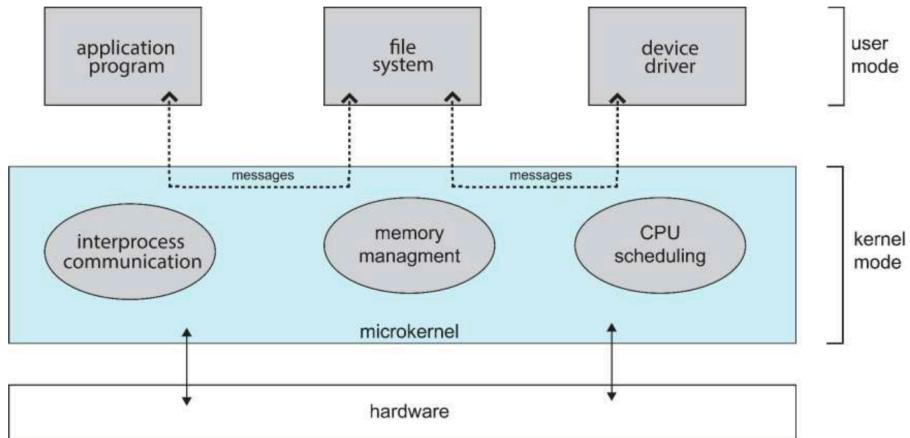
Prozesssteuerung: Zum Prozess Erstellen, Löschen, Laden und Ausführen, Abrufen und Festlegen von Attributen, Warte- bzw. Signalereignisse, Speicher Zuweisung und Freigabe, core dump, Debugger und Locking für gemeinsame Daten.

Dateiverwaltung: Zum Erstellen, Löschen, Öffnen und Schließen von Dateien, Lesen, Schreiben und Bearbeiten von Daten in Dateien und das Abrufen von Dateiattributien.

Devicemanagement: Zum Anfordern und Freigeben von Geräten, Lesen, Schreiben und Bearbeiten von Geräten, Abrufen und Festlegen von Gerät Attributien, sowie das logische An- und Abkoppeln von Geräten

Diese Unterschiedlichen Arten von Systemcalls werden verwendet um bereits vorhandene Informationen zu nutzen bzw. festzulegen. Sie sind wichtig um den Zugriff zu schützen und die Kommunikation zwischen Prozessen zu organisieren.

8.3 Microkernel

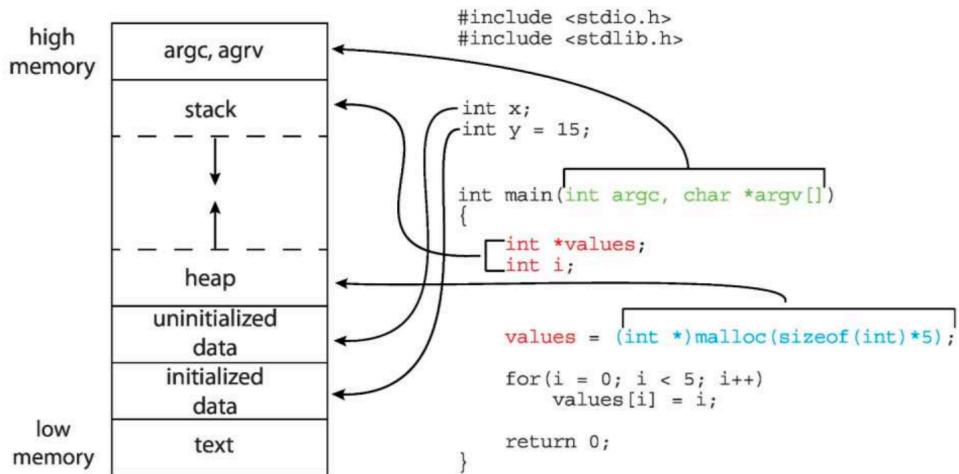


9 Prozesskonzept

9.1 Prozess

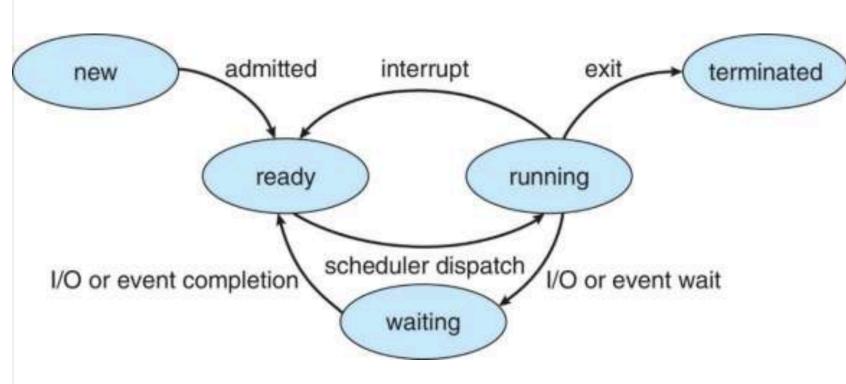
Ein Prozess ist eine Ausführbare Datei auf der Festplatte die Programmcode und/oder Textabschnitte enthält. Die Datei wird geladen und sequenziell ausgeführt. Hier hat der Prozess wie bereits weiter oben angesprochen einen Programcounter, Prozessorregister, einen Stack, globale Daten wie Variablen und einen Heap auf dem Speicher dynamisch alloziert werden kann.

9.2 Speicherstruktur



9.3 Zustand

- **new:** wird gerade erstellt
- **ready:** bereit für Zuweisung (auf Prozessor)
- **running:** Ausführen der Anweisungen
- **waiting:** warten auf Ereignis
- **finished:** Ausführung abgeschlossen



9.4 Prozesskontrollblock

- Zustand
- Programcounter
- CPU-Register
(Data in process dependant registers)
- CPU-Schedulinginformation
(Priority, Schedulingqueue pointer)
- Speicherverwaltungsinformationen
(Registers, Pagetable)
- Abbrechungsinformation
(CPU usage, Time, Timelimits)
- I/O Statusinformation
(Assigned Devices, open files)

process state
process number
program counter
registers
memory limits
list of open files
...

Jeder Thread hat auch einen eigenen Kontrollblock!

9.5 Linux

- **task_struct** Datenstruktur unter `<include/linux/sched.h>`

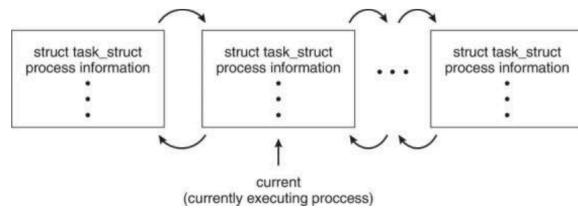
```

pid t_pid;                      /* Prozessbezeichnung */
long state;                      /* Zustand */
unsigned int time_slice;          /* Zustandsinformation */
struct task_struct *parent;       /* Vaterprozess */
struct list_head children;        /* Kinderprozessen */
struct files_struct *files;       /* Offene Dateien */
struct mm_struct *mm;             /* Speicher */

```

- **Prozesstabelle**

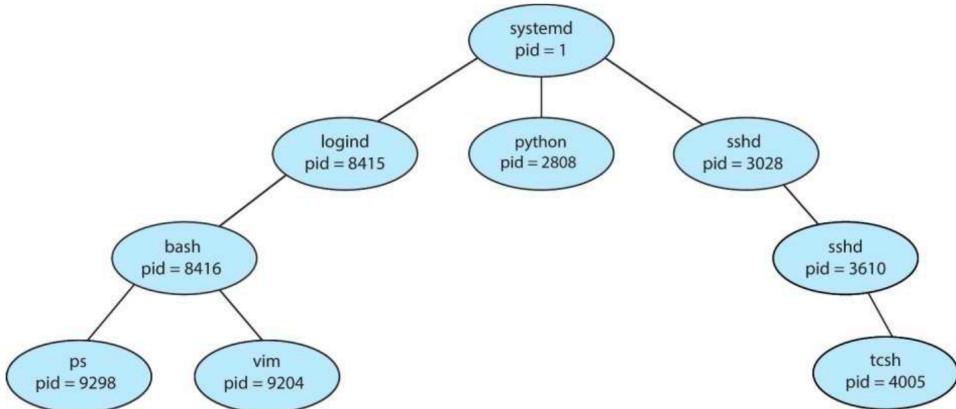
- `current->state = ...`



10 Prozessoperationen

10.1 Prozesserstellung

Bei der Prozesserstellung erstellt ein (jetzt Eltern-)Prozess einen Kindprozess. Dieser hat eine Prozessid (PID bzw. pid_t (POSIX)). Es wird festgelegt ob die Daten des Elternprozesses (teilweise) geteilt werden oder kopiert werden. Außerdem kann festgelegt werden ob auf den Kindprozess schlussendlich gewartet werden soll oder ob dieser von alleine schließt, bzw. ein Daemon ist. Somit wird ein Prozessbaum erstellt. Beispiel in Linux:



Die Erstellung eines Kindprozesses unter UNIX folgt folgendem Schema:

```
#include <sys/types.h>
#include <sys/wait.h>
#include <stdio.h>
#include <unistd.h>

int main(void) {
    pid_t pid = fork();           /* Kindprozess erstellen */
    if (pid < 0) {
        fprintf(stderr, "Fork failed.");
        return 1;
    } else if (pid == 0)          /* Kindprozess */
        execlp("/bin/ls", "ls", NULL);
    } else {                      /* Elternprozess */
        wait(NULL);              /* Kindsbeendigung warten */
        printf("Child completed.");
    }
    return 0;
}
```

fork() erstellt einen Kindprocess als Duplikat des Elternprozesses.

exec() lädt neues Programm in dem Kindprozess. Dieses ersetzt den Speicherplatz des Kindprozesses.

wait() wartet auf die Beendigung des Kindprozesses

10.2 Prozessterminierung

Ein Prozess terminiert wenn entweder die letzte Anweisung ausgeführt wird oder exit(int status) im Kindprozess aufgerufen wird. wait(int *status) wartet auf die Beendigung des Kindes und kann den exit status code des Kindes abrufen. abort() terminiert den Kindprozess vom Elternprozess aus. Wird nicht auf die erstellten Kindprozesse gewartet werden entweder

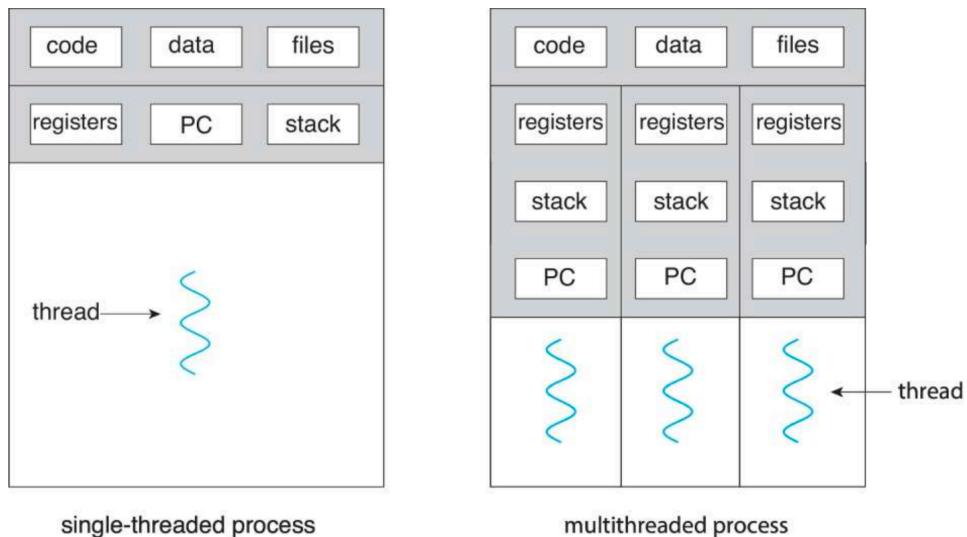
alle nicht terminierten Kinder vom Betriebssystem terminiert (Cascading) oder es verbleiben verwaiste Kindprozesse. Terminiert ein Kind ohne das der Elternprozess auf diesen wartet nennt man das Zombieprozess.

11 Threads

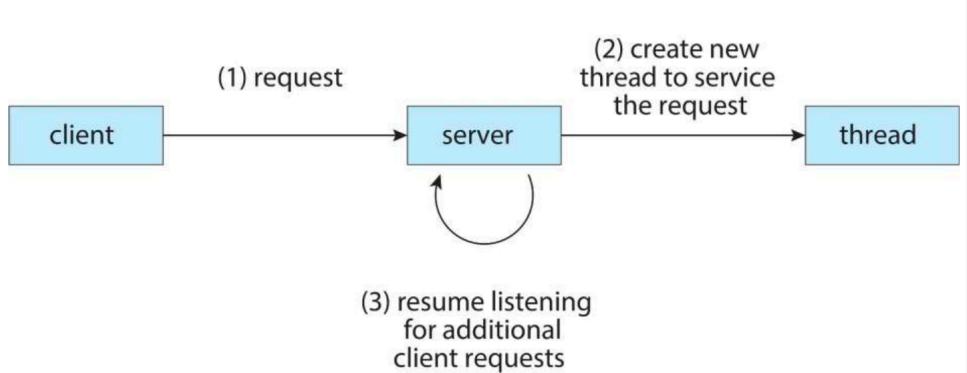
Da viele moderne und Interaktive Anwendungen mehrere Aufgaben parallel ausführen wurde nach einer neuen Methode gesucht um Daten parallel zu verarbeiten. Dafür bieten sich Threads an. Diese sind lightweight und kompatibel mit neuen Multithread-Kernels und Multithread Prozessoren.

Die Vorteile von Threads sind verbesserte Reaktionsfähigkeit auch wenn Teile des Prozesses Blockieren, was besonders wichtig für Grafische Benutzeroberflächen ist. Außerdem lassen sich die Ressourcen eines CPU-Kerns und generell die Ressourcen besser teilen, da kein automatisches Kopieren der Daten des Prozesses stattfindet. Die Threadingerstellung ist zudem wesentlich günstiger und Contextswitches bereiten weniger Overhead, da immernoch auf dem selben Prozessorkern gerechnet wird. Multithread Anwendungen sind zudem besser skalierbar auf neuen Multicore Systemen.

11.1 Single vs. Multithreaded Programme



11.2 Anwendung (Server)



11.3 Concurrency vs. Parallelism

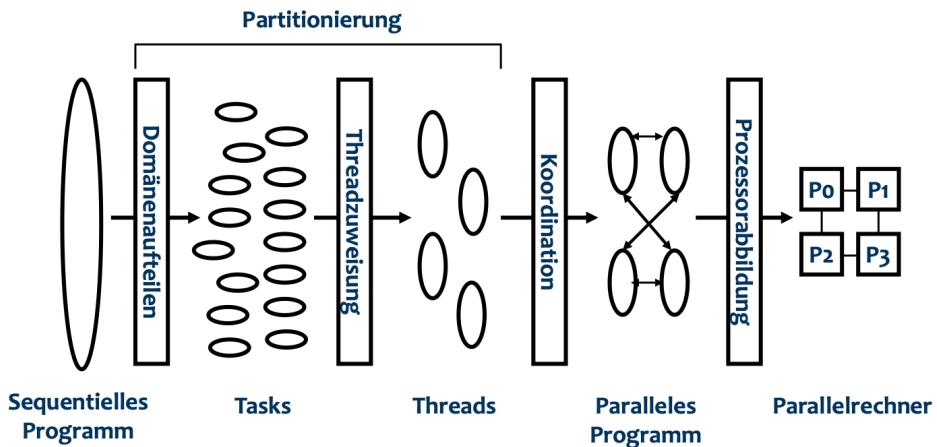
Concurrency:

$$T_1 = \text{Task 1}; T_2 = \text{Task 2}; \\ T_1 \rightarrow T_2 \rightarrow T_1 \rightarrow T_2 \rightarrow T_1 \rightarrow T_2 \rightarrow \dots$$

Parallelism:

$$T_1 = \text{Task 1}; T_2 = \text{Task 2}; \\ T_1 \rightarrow T_1 \rightarrow T_1 \rightarrow \dots \\ T_2 \rightarrow T_2 \rightarrow T_2 \rightarrow \dots$$

11.4 Parallelism



11.5 Data vs Task Parallelism

Bei der Datenparallelisierung findet die selbe Berechnung auf unterschiedlichen Daten statt.
Bei der Taskparallelisierung finden unterschiedliche Berechnungen auf den gleichen bzw. verschiedenen Daten statt.

Bei der Datenparallelisierung wird ein Datensatz in kleinere Teilstücke zerlegt. Diese können mit dem gleichen Code von verschiedenen Prozessen und/oder Threads verarbeitet werden. Dies findet häufig bei Matrix Berechnungen oder Machine Learning statt.

Bei der Taskparallelisierung werden unterschiedliche Aufgabe/Berechnungen auf dem selben oder unterschiedlichen Daten ausgeführt. Dies ist häufig bei komplexen Problemen mit mehreren Arbeitsschritten und ideal wenn das Gesamtproblem in einzelne kleinere Teilprobleme zerlegt werden kann. Betriebssysteme verwenden häufig diesen Ansatz um zum Beispiel gleichzeitig Daten zu verarbeiten und auf der Grafischen Oberfläche anzuzeigen.

11.6 User- und Kernelthreads

Userthreads:

Verwaltung durch Schnittstelle des Betriebssystems:

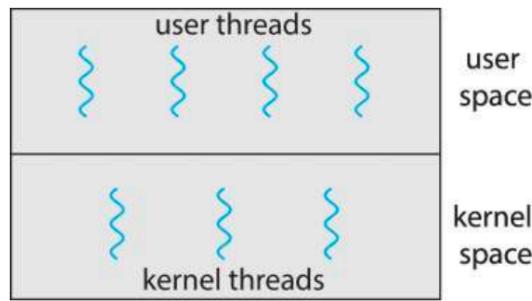
- POSIX: pthread
- Windows: Threads

Kernelthreads:

Verwaltung durch den Kernel

- Linux
- Windows
- MacOS

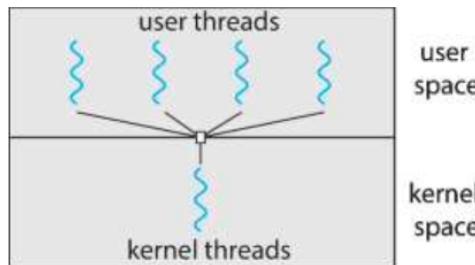
- Java: Threads
- ...



11.7 Multithreading Modelle

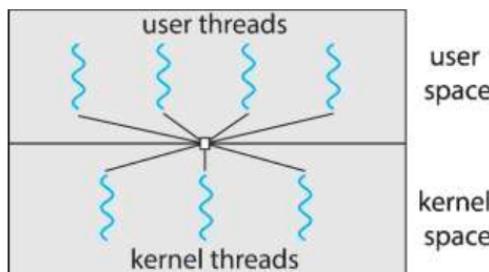
Many-to-One:

Meherere Benutzer werden einem einzelnen Kernel-Thread zugeordnet. Das Blockieren eines Threads blockiert alle anderen. Dadurch ist möglicherweise keine Parallelisierung möglich. Dieser Ansatz wird von wenigen Betriebssystemen genutzt



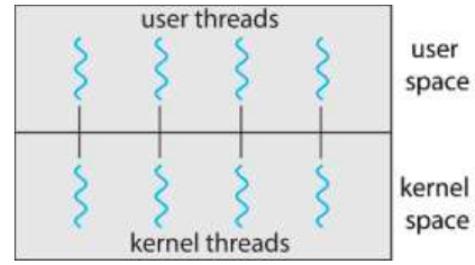
Many-to-Many:

Viele Benutzer-Threads werden vielen Kernel-Threads zugeordnet. Dies benötigt eine ausreichende Anzahl an Kernel-Threads. Dieser Ansatz wird von Windows mit Thread-Fiber-Package verwendet, wird jedoch nicht häufig verwendet.



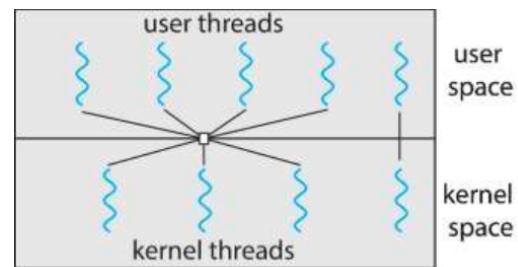
One-to-one:

Jedem Benutzer wird ein Kernel-Thread zugeordnet. Das bietet mehr Parallelität bringt jedoch den Nachteil das die Anzahl der Threads pro Benutzer möglicher begrenzt ist. Linux und Windows benutzen diesen Ansatz.



2-Layers:

Ähnlich wie der Many-to-Many Ansatz. Bindung von Benutzer-Threads an Kernel-Threads möglich, wenn in Benutzung.



11.8 PThread (POSIX Threads)

Wie bereits weiter oben angesprochen ist pthread die POSIX Schnittstelle für Threads auf UNIX Systemen.

```

#include <pthread.h>
#include <stdio.h>
#include <stdlib.h>

/* geteilte Variable */
int sum;

/* Threadfunktion */
void *runner(void *param) {
    int i, n = atoi(param);
    sum = 0;
    for(i = 1; i <= n; i++)
        sum += i;
    pthread_exit(0);
}

int main(int argc, char *argv[]) {
    /* Thread-Bezeichner */
    pthread_t tid;

    /* Thread-Attribute einstellen */
    pthread_attr_t attr;
    pthread_attr_init(&attr);

    /* Thread erstellen */
    pthread_create(&tid, &attr,
                  runner, argv[1]);

    /* Thread-Beendigung warten */
    pthread_join(tid, NULL);
    printf("sum = %d\n", sum);
}

```

11.9 fork und exec

fork dupliziert den aufrufenden Thread bzw. den Prozess der den Thread gestartet hat. exec ersetzt den laufenden Prozess und somit auch alle erstellten Threads.

11.10 Signals und Interrupts

Signale, die erzeugt durch ein bestimmtes Ereignis und an einen Prozess übermittelt, werden vom Kernel Signalhandler bzw. einem Benutzerdefinierten Signalhandler verarbeitet. Die Übermittlung an die Threads des Prozesses findet entweder für alle Threads, bestimmte Threads oder einen bestimmten Thread für alle Signale statt.

11.11 Abbruch

Es wird zwischen Asynchronen und Verzögerten Abbrüchen unterschieden. Ein Asynchroner Abbruch beendet den Thread sofort, wohingegen ein verzögerter Abbruch bei bestimmten Abbruchpunkten beendet.

Für den Abbruch benutzen wir `pthread_setcancelstate(...)` und für den Typen `pthread_setcanceltype(...)`. Mit `pthread_testcancel()` wird überprüft ob der Thread abgebrochen werden soll. `pthread_cancel(thread_id)` bricht den Thread mit der entsprechenden Id ab. Auch Threads die abgebrochen werden müssen mit `pthread_join(tid, NULL)` wieder mit dem main thread bzw. dem Elternprozess verbunden werden. Die Abbruch Typen sind Off/NULL, Deferred und Asynchronous.

12 Interprozess Kommunikation (IPC)

Die Interprozess Kommunikation wird benötigt, wenn ein Problem in mehrere Teilprobleme zerlegt wird (*performance*) und von mehreren Prozessen gleichzeitig bearbeitet wird. Hier ist es nicht immer möglich die einzelnen Aufgaben vollständig auf die einzelnen Prozesse aufzuteilen und es ist nötig Daten zu bestimmten Zeitpunkten zwischen den arbeitenden Prozessen zu teilen um schlussendlich das Gesamtproblem zu lösen.

Da Prozesse keinen Speicher(memory und file handles) Teilen ist meist eine Kommunikation zwischen Prozessen notwendig, die durch das Betriebssystem bereitgestellt wird.

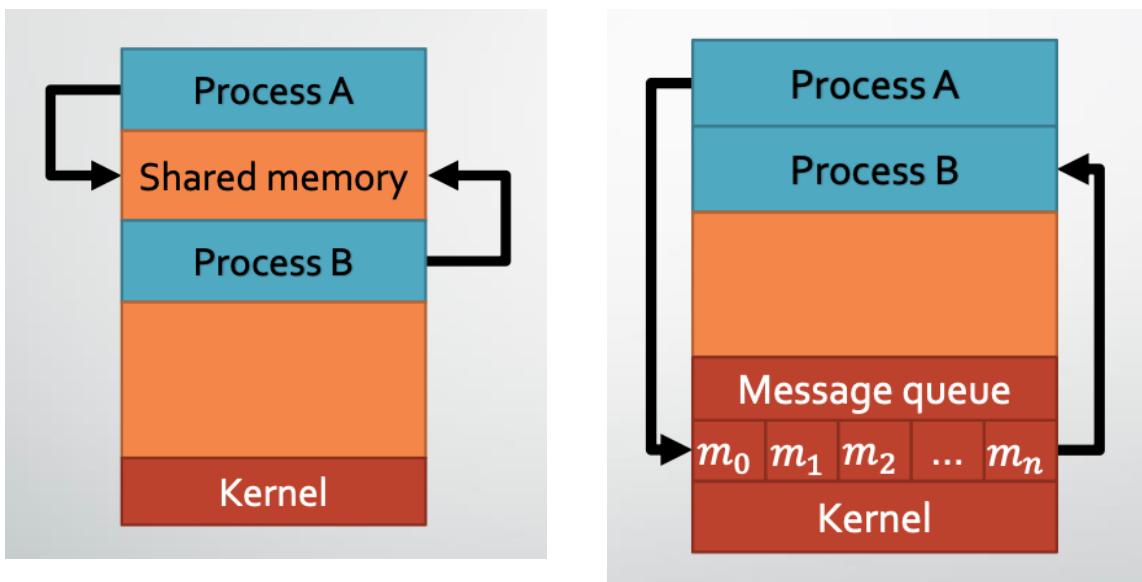
Zu dem bereits angesprochenen Vorteil der schnelleren Berechnung kommt durch das Aufteilen der Probleme auf unterschiedliche Programme/Prozesse eine verbesserte Modularität und somit auch eine bessere Fehlertoleranz bzw. Fehlerlokalität.

Ein *Nachteil* der Aufteilung ist die vermehrt benötigte Synchronisation, auf die in späteren Kapiteln eingegangen wird.

Beispiele für die Interprozesskommunikation sind:

- Dateien
- Pipes
- Shared Memory
- Signals
- Message Queues
- Sockets
- ...

12.1 Fundamentals Model



12.2 Messagepassing

- `send()`
- `receive()`

Erstellen einer IPC und Austausch von Daten über `send()` und `receive()`. Schlussendliches Löschen der IPC. Da die IPC mit mehreren Prozessen geteilt werden soll muss diese entweder im Gemeinsamen Speicher gespeichert werden, oder vor der Erstellung der Kindprozesse angelegt werden. Das Messagepassing kann bi- bzw unidirektional sein.

Synchronisation:

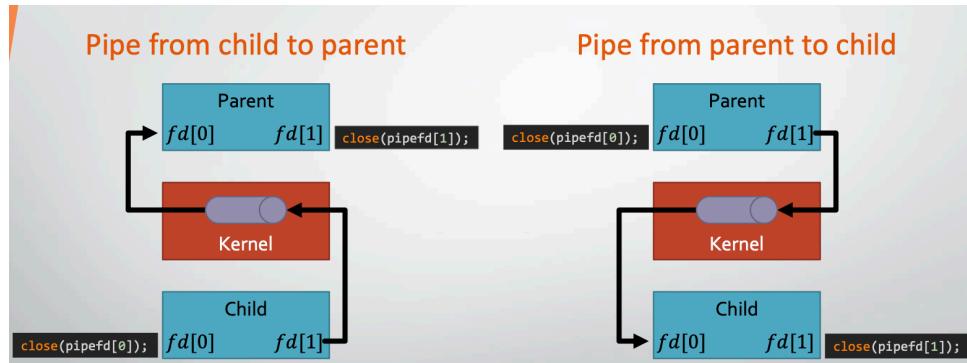
- Synchroner Austausch: blocking sender wartet bis Nachricht empfangen wurde, blocking receiver wartet bis eine Nachricht verfügbar ist.
- Asynchroner Austausch: non blocking sender sendet Nachricht und setzt fort, non blocking receiver empfängt eine Nachricht oder auch keine Nachricht.
- Rendezvous Kommunikation: blocking sender und receiver
- Hybrid: Synchron und Asynchron

Eine einfache IPC kann mit einer Datei implementiert werden. Dies ist jedoch extrem langsam und ineffizient. Die Synchronisation ist zudem nicht trivial und definitiv benötigt.

12.3 Unnamed Pipe

Unnamed Pipes werden von allen UNIX Systemen unterstützt. Sie sind unidirektional, haben also ein Read- und ein Write-end. Die Benutzung ist ähnlich zu den read/write Systemcalls. Daten können nur einmal gelesen werden und können jegliche Form haben. Der Kernel ist zuständig für die nötige Synchronisation.

Um eine Pipe zu erstellen muss der pipe Systemcall vor dem erstellen des Kind/Arbeitsprozesses aufgerufen werden. Passiert das nicht verbindet die Pipe zu dem gerade laufenden Prozess. Nach dem fork Aufruf sind Eltern- und Kindprozess mit der pipe verbunden.



```

void parent(const int pipefd[2]) {
    close(pipefd[0]);                                // Close read-end

    const char* msg = "Hello World!";
    write(pipefd[1], msg, strlen(msg));
    close(pipefd[1]);                                // Close write-end
                                                       // -> reader will see EOF

    wait(NULL);                                     // Wait for child
}

void child(const int pipefd[2]) {
    close(pipefd[1]);                                // Close write-end
    char buf;
    while(read(pipefd[0], &buf, 1) > 0) {
        write(STDOUT_FILENO, &buf, 1);
    }
    write(STDOUT_FILENO, "\n", 1);
    close(pipefd[0]);                                // Close read-end too
}

int main(void) {
    int pipefd[2];                                  // pipefd[0]: read-end
                                                       // pipefd[1]: write-end
    if(pipe(pipefd) != 0) return EXIT_FAILURE;

    const pid_t cpid = fork();
    if(cpid == -1) return EXIT_FAILURE;

    if(cpid == 0) child(pipefd);
    else          parent(pipefd);
}

```

12.4 Named Pipe (FIFO)

Named Pipes oder auch FIFO genannt werden von POSIX unterstützt und funktionieren ähnlich zu normalen Dateien. Sie können geöffnet, geschlossen, gelesen und beschrieben werden. Anders zu wirklichen Dateien werden die Daten jedoch nicht auf dem Sekundärspicher gespeichert. Sie werden vom Kernel verwaltet. Die Daten der FIFO können nur 1 mal gelesen werden. Im Gegensatz zu Unnamed Pipes können Named Pipes von mehreren Prozessen geöffnet werden. Eine Kommunikation ist jedoch erst möglich wenn beide Enden (Read/Write) geöffnet wurden. Wenn nur 1 Prozess schreibt und 1 Prozess liest ist die Synchronisation durch den Kernel sichergestellt.

```
void create_named_pipe_reader(void) {
    const char* name = "named_pipe";

    const mode_t permission = S_IRUSR | S_IWUSR | S_IRGRP | S_IROTH; // 644
    if(mkfifo(name, permission) != 0) return EXIT_FAILURE;

    const int fd = open(name, O_RDONLY);
    if(fd < 0) return EXIT_FAILURE;

    char buf;
    while(read(fd, &buf, 1) > 0) {
        write(STDOUT_FILENO, &buf, 1);
    }
    write(STDOUT_FILENO, "\n", 1);

    close(fd);
    unlink(name);
}

void create_named_pipe_writer(void) {
    const char* name ="named_pipe";
    const int fd = open(name, O_WRONLY);
    if(fd < 0) return EXIT_FAILURE;

    const char* msg = "Hello World";
    write(fd, msg, strlen(msg));

    close(fd);
}
```

12.5 Messagequeue

Messagequeues ermöglichen im Vergleich zu Pipes das Senden und Empfangen von Daten mit einer festgelegten Größe. Diese Größe wird gesendet UND empfangen und wird meist gehard-coded. Diese Pakete werden in der selben Reihenfolge empfangen wie sie gesendet wurden.

POSIX Messagequeues erlauben zu dem senden und empfangen die mitgabe der Priorität eines Pakets. Eine Höhere Priorität bedeutet das frühere Zustellen beim Empfänger, wohingegen die Reihenfolge von Paketen mit der selben Priorität erhalten bleibt.

- mqd_t mq_open(const char *name, int oflag, mode_t mode, struct mq_attr *attr);

Erstellt oder Öffnet eine Messagequeue. Der Name (Konvention: Start mit "/") identifiziert die Queue aber ist nicht im Dateisystem sichtbar. Die oflag ist zuständig für den Zugriffstypen und ob die Queue erstellt werden darf.

mode sind die Berechtigungen der Queue.

attr legt die Eigenschaften der Messagequeue fest.

```

struct mq_attr {
    long mq_flags;           // queue flags, ignored on open
    long mq_maxmsg;          // max number of messages in the queue at any point
    long mq_msgsize;          // max size of each individual message
    long mq_curmsgs;         // number of messages in the queue
};

• int mq_send(mqd_t mqdes, const char* msg_ptr, size_t msg_len, unsigned int msg_prio);

```

Sendet eine Nachricht in die Messagequeue, mit den Daten msg_ptr und der Länge von msq_size. msg_prio legt dir Priorität der Message fest.

```
• ssize_t mq_receive(mqd_t mqdes, char* msg_ptr, size_t msg_len, unsigned int* msg_prio);
```

Empfängt eine Nachricht aus der Messagequeue. Die Daten werden in msg_ptr mit der Länge msq_size gespeichert. msg_prio ist die Priorität der empfangenen Message.

```

struct message {
    char data[32];
    bool quit;
};

typedef struct message message;

bool create_message_queue(const char* name) {
    const int oflag = O_CREAT | O_EXCL;
    const mode_t permissions = S_IRUSR | S_IWUSR; // 600
    const struct mq_attr attr = { .mq_maxmsg = 2, .mq_msgsize = sizeof(message) };
    const mqd_t mq = mq_open(name, oflag, permissions, &attr);
    if(mq == -1) return false;
    mq_close(mq);
    return true;
}

void logging_server(char **msg_queue_name) {
    const mqd_t mq = mq_open(msg_queue_name, O_RDONLY, 0, NULL);
    for(int quit_requests = 0; quit_requests < 2;) {
        usleep(100 * 1000); // Simulate logging being very slow
        message msg = { 0 };
        unsigned int priority = 0;
        if(mq_receive(mq, (char*)&msg, sizeof(msg), &priority) == -1) return;

        if(msg.quit) ++quit_requests;
        else printf("%02u: %s\n", priority, msg.data);
    }
    mq_close(mq);
    mq_unlink(msg_queue_name);
}

void client(char **msg_queue_name, long priority) {
    const mqd_t mq = mq_open(msg_queue_name, O_WRONLY, 0, NULL);
    for(int i = 0; i < 10; ++i) {
        message msg = { .quit = false };
        sprintf(msg.data, "Hello World %d", i);
        if(mq_send(mq, (const char*)&msg, sizeof(msg), priority) != 0) return;
    }
    const message msg = { .quit = true };
    if(mq_send(mq, (const char*)&msg, sizeof(msg), priority) != 0) return;
    mq_close(mq);
}

int main(void) {
    pid_t cpid = fork();
    if(cpid == 0) {
        cpid = fork();
        if(cpid == 0) client(msg_queue_name, 0);
        else client(msg_queue_name, 1);
    } else {
        logging_server(msg_queue_name);
    }
}

```

12.6 Shared Memory

- `int shm_open(const char *name, int oflag, mode_t mode);`

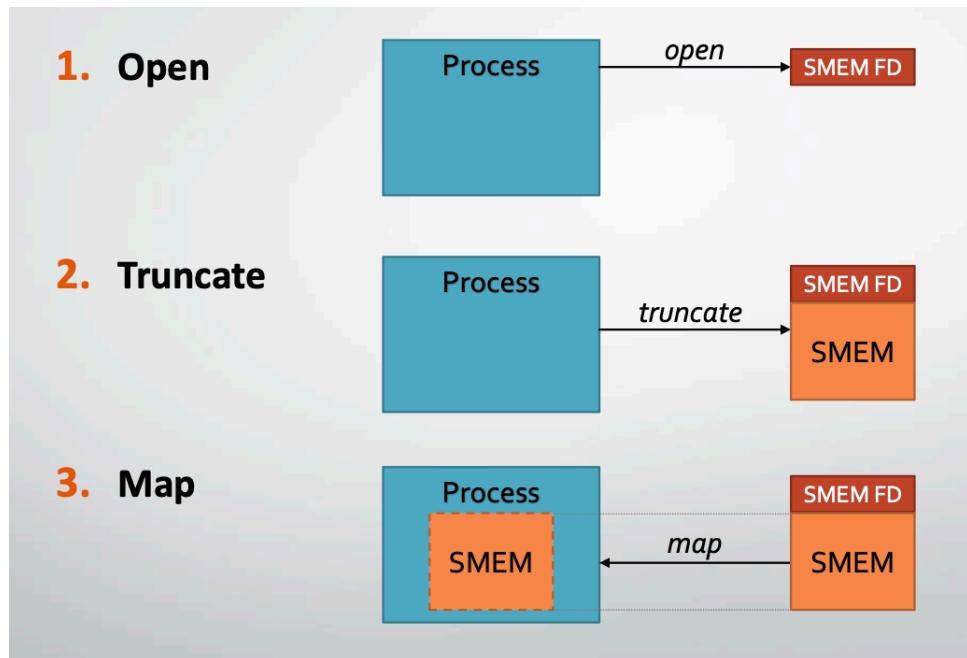
Erstellt oder öffnet ein Shared memory Objekt spezifiziert durch den Namen (Konvention: Start mit "/"). Dieses ist wie bei der Messagequeue nicht im Dateisystem zu finden. Wichtig ist hier das kein Speicher wirklich alloziert wird!

- `int ftruncate(int fd, off_t length);`

Begrenzt (oder erweitert) eine Datei auf die spezifizierte Größe. Hier ist wichtig das ftruncate nur auf allozierte Speicherbereiche angewendet werden darf.

- `void *mmap(void *addr, size_t length, int prot, int flags, int fd, off_t offset);`

Dieser Befehl mapped Dateien oder Geräte in den Prozessspeicher. mmap wird benötigt um das shared_mem Objekt in den wirklichen Speicher zu legen. addr = NULL lässt den Kernel entscheiden wo genau das Memory Objekt alloziert/gemapped werden soll. length beschreibt die Größe des shared_mem Objekts. prot legt die Berechtigungen für den Speicherbereich fest. flags sollte bei shared_mem = MAP_SHARED. fd ist der Filedescriptor für das shared_mem Objekt. offset spezifiziert ob das Speicherobjekt mit einem gewissen offset gemapped werden soll. Bei einem fehlgeschlagenen Mapping wird MAP_FAILED zurückgegeben, sonst ein alloziert Speicherbereich.



```

void writer(void) {
    const char* name = "/shared_memory";
    const int oflag = O_CREAT | O_EXCL | O_RDWR; // create, fail if exists, read+write
    const mode_t permission = S_IRUSR | S_IWUSR; // 600
    const int fd = shm_open(name, oflag, permission);
    if(fd < 0) return EXIT_FAILURE;

    const size_t shared_mem_size = 100;
    if(ftruncate(fd, shared_mem_size) != 0) return EXIT_FAILURE;

    char* shared_mem = mmap(NULL, shared_mem_size, PROT_READ | PROT_WRITE, MAP_SHARED, fd, 0);
    if(shared_mem == MAP_FAILED) return EXIT_FAILURE;

    const char message[] = "Hello World";
    memcpy(shared_mem, message, sizeof(message));

    usleep(10 * 1000 * 1000);

    munmap(shared_mem, shared_mem_size);
    close(fd);
    shm_unlink(name);
}

void reader(void) {

```

```

const char* name = "/shared_memory";
const int oflag = O_RDWR; // open read+write
const int fd = shm_open(name, oflag, 0);
if(fd < 0) return EXIT_FAILURE;

const size_t shared_mem_size = 100;
char* shared_mem = mmap(NULL, shared_mem_size, PROT_READ | PROT_WRITE, MAP_SHARED, fd, 0);

if(shared_mem == MAP_FAILED) return EXIT_FAILURE;

char buffer[shared_mem_size];
memcpy(buffer, shared_mem, shared_mem_size);

munmap(shared_mem, shared_mem_size);
close(fd);

printf("%.*s\n", (int)shared_mem_size, buffer);
}

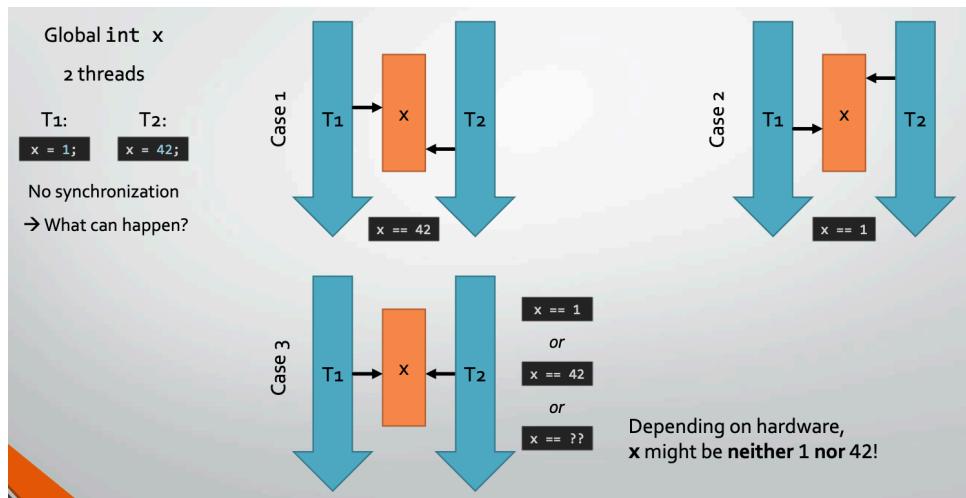
```

Beim Shared Memory stellt sich nun jedoch ein Problem. Wann weiß das Lesende Ende wann Daten verfügbar sind? Was passiert wenn gleichzeitig gelesen **und** geschrieben wird? Was ist wenn das Lesende Ende Daten zurücksenden will?

Hier wird eine Art von Synchronisation benötigt. Der usleep call ist nicht ausreichend.

13 Synchronisation

Wir benötigen Synchronisation, da Prozesse und Threads gleichzeitig Daten verarbeiten. Wenn nun mehrere Prozesse wieder gleichzeitig auf die selben Daten zugreifen kann es zu Fehlern kommen. Diese Fehler werden **race conditions** genannt.



Oder in Code:

```

#define ITERATIONS 100000000
#define THREADS 2
int x = 0;
void* thread(void* param) {
    for(int i = 0; i < ITERATIONS; ++i) {
        ++x;
    }
    return NULL;
}
int main(void) {
    /* spawn THREADS number of threads and wait until they finish */
}

```

```
    printf("x: %d", x);
}
```

13.1 Data Hazards

- **Read-after-Write** (RAW): true dependency

```
1) x = 4;
2) y = x + 7;
3) x = 2;
```

y hängt vom Wert der Variable x ab. 1 und 2 haben eine true dependency. Wird Zeile 1 und 2 getauscht ist der Output ein anderer.

- **Write-after-Read** (WAR): anti dependency

```
1) x = 4;
2) y = x + 7;
3) x = 2;
```

y hängt vom Wert der Variable von x ab. x wird später verändert. Wenn nun Zeile 2 und 3 getauscht werden ist der Output ein anderer.

- **Write-after-Write** (WAW): output dependency

```
1) x = 4;
2) y = x + 7;
3) x = 2;
```

y hängt vom Wert der Variable von x ab. x wird in Zeile 1 und 3 verändert. Wenn nun Zeile 1 und 3 getauscht werden ist der Output ein anderer.

13.2 Critical Sections

In dem oben gezeigten Code Abschnitt ist `++x` die Critical section. Jede Critical Section benötigt eine Synchronisation bei einer parallelen Ausführung.

13.3 Atomics

Atomics sind der Grundstein für die Synchronisation. Sie benötigen Hardware support und können dementsprechend entweder komplett funktionieren oder gar nicht. Mit Atomics können wir die oben angesprochenen Probleme vollständig lösen.

`<stdatomic.h>` stellt primitive Datentypen wie `atomic_bool` oder `atomic_int` zur Verfügung. Benutzerdefinierte Typen können mit `_Atomic` deklariert werden und sind demnach Atomic. Wird `_Atomic` verwendet müssen die Atomic API Funktionen verwendet werden.

Das oben gegebene Beispiel kann durch die Benutzung eines `atomic_int` als Datentyp für x synchronisiert werden, da der parallele Zugriff auf Atomics sicher ist.

13.4 Atomic Operationen

Operation	Explanation
<code>atomic_store</code>	stores a value in an atomic object
<code>atomic_load</code>	reads a value from an atomic object
<code>atomic_exchange</code>	swaps a value with the value of an atomic object
<code>atomic_compare_exchange_*</code>	swaps a value with an atomic object if the old value is what is expected, otherwise reads the old value
<code>atomic_fetch_add</code>	atomic addition
<code>atomic_fetch_sub</code>	atomic subtraction
<code>atomic_fetch_or</code>	atomic bitwise OR
<code>atomic_fetch_xor</code>	atomic bitwise exclusive OR
<code>atomic_fetch_and</code>	atomic bitwise AND

13.5 Shared Memory mit Synchronisation

```
struct shared_data {
    atomic_bool available;
    atomic_bool processed;
    size_t cnt;
    char buf[1024];
};

typedef struct shared_data shared_data;
```

Dieses Konstrukt stehen dem Schreibenden **und** dem Lesenden Ende zur Verfügung.
Der Server:

```
/* shared memory creation and mapping, same as before */
data->processed = true; // Indicate initialization is done
while(!data->available); // Busy wait for data to become available
// Process data
for(size_t i = 0; i < data->cnt; ++i) {
    data->buf[i] = toupper((unsigned char)data->buf[i]);
}
// Signal data has been processed
data->processed = true;
munmap(data, sizeof(shared_data));
close(fd);
shm_unlink(shared_mem_name);
```

Der Client:

```
/* shared memory creation and mapping, same as before */
while(!data->processed); // Busy wait until shared memory has been initialized
data->processed = false; // Reset flag
// Write data
data->cnt = strlen(message);
memcpy(&data->buf, message, data->cnt);
// Signal data has been written
data->available = true;
while(!data->processed); // Busy wait until data has been processed
printf("%.*s\n", (int)data->cnt, data->buf);
munmap(data, sizeof(shared_data));
close(fd);
```

13.6 Mutex

Mutex (von **Mutual Exclusion**), auch genannt Locks können 2 States halten:

- Locked/held/owned/acquired
- Unlocked/free/released.

Ein Mutex kann immer nur von **einem** Thread gehalten/gelocked sein. Jeder weitere Thread der versucht den Mutex zu halten wartet bis dieser erneut freigegeben wird. Die Nutzung eines Mutex bringt jedoch einen großen Performance Verlust mit sich, weshalb die Critical Section klein gehalten werden sollte.

Mutexes werden meist durch eine Binäre Semaphore dargestellt (siehe Semaphores 13.16).

Das obige Problem, zuerst gelöst durch die Verwendung von Atomics kann auch mit der Verwendung eines Mutex gelöst werden:

```
pthread_mutex_t mutex;
int x = 0;
void* thread(void* param) {
    for(int i = 0; i < ITERATIONS; ++i) {
        pthread_mutex_lock(&mutex);
        ++x;
        pthread_mutex_unlock(&mutex);
    }
    return NULL;
}
• int pthread_mutex_init(pthread_mutex_t* mutex, const pthread_mutexattr_t* mutexattr);
```

oder

- `pthread_mutex_t mutex = PTHREAD_MUTEX_INITIALIZER;`

erstellt den Mutex mit den spezifizierten mutexattr bzw. einen default Mutex wenn mutexattr = NULL.

- `int pthread_mutex_lock(pthread_mutex_t* mutex);`

versucht den Mutex zu locken. Ist dieser nicht gelocked, locked der aufrufende Thread den Mutex sofort. Sonst blockt der Thread so lange bis der Mutex unlocked ist.

- `int pthread_mutex_trylock(pthread_mutex_t* mutex);`

versucht den Mutex zu locken. Ist dieser nich gelocked, locked der aufrufende Thread den Mutex sofort. Sonst wird der error code EBUSY zurückgegeben.

Das Locken des Mutex darf im Thread nur einmal passieren.

- `int pthread_mutex_unlock(pthread_mutex_t* mutex);`

unlocked den Mutex. Der Mutex muss zuvor von den Thread gelocked worden sein.

- `int pthread_mutex_destroy(pthread_mutex_t* mutex);`

zerstört den Mutex nach der Benutzung. Der Mutex darf zu diesem Zeitpunkt nicht mehr gehalten werden.

13.7 Dekker's Algorithmus

(1960er)

Hardware support für Mutexes hat es noch nicht immer gegeben und der Dekker Algrotihmus war die erste korrekte Lösung für das Mutex Problem. Der Dekker Algorithmus ist in Software implementiert, ohne jeglichen Hardware support. Dieser arbeitet auf dem Gemeinsamen Speicher. Auf neuen Systemen nicht funktionsfähig und nicht möglich rein in C zu lösen, da der Compiler und die CPU heutzutage in der Lage sind Instruktionen umzustellen wenn der Code nicht synchronisiert ist. Dies verhindert das Benutzen des Dekker Algorithmus.

Dekker Algorithmus in Pesudocode:

```
variables:  
    wants_to_enter: array of 2 booleans  
    turn: integer  
    wants_to_enter[0] ← false  
    wants_to_enter[1] ← false  
    turn ← 0 // or 1  
  
p0:  
    wants_to_enter[0] ← true  
    while wants_to_enter[1] {  
        wants_to_enter[0] ← false  
        while turn != 0 // busy wait  
        wants_to_enter[0] ← true  
    }  
    // critical section  
    turn ← 1  
    wants_to_enter[0] ← false  
    // remainder section  
  
p1:  
    wants_to_enter[1] ← true  
    while wants_to_enter[0] {  
        wants_to_enter[1] ← false  
        while turn != 1 // busy wait  
        wants_to_enter[1] ← true  
    }  
    // critical section  
    turn ← 0  
    wants_to_enter[1] ← false  
    // remainder section
```

13.8 Peterson's Algorithmus

(1981)

Der Peterson's Algorithmus ist wie der Dekker Algorithmus eine Software Lösung für das Mutex Problem. Der Peterson's Algorithmus wird als besser angesehen, da er einfacher zu implementieren ist, weniger Anweisungen und eine klarere Logik besitzt und besser für formale Beweise geeignet ist.

Peterson's Algorithmus in Pseudocode:

```
variables:  
    flag: array of 2 booleans  
    turn: integer  
    flag[0] ← false  
    flag[1] ← false  
  
p0:  
    flag[0] ← true  
    turn ← 1  
    while (flag[1] && turn == 1)  
        // critical section  
        flag[0] ← false  
        // remainder section  
  
p1:  
    flag[1] ← true  
    turn ← 0  
    while (flag[0] && turn == 0)  
        // critical section  
        flag[1] ← false  
        // remainder section
```

Wie auch bereits beim Dekker Algorithmus angesprochen ist das benutzen dieses Algorithmus heutzutage nicht mehr möglich, wegen Compiler und CPU Instruction reordering.

13.9 Memorybarrier

Heutige CPUs optimieren die Ausführung stark. Um schneller zu arbeiten werden häufig Befehl umgeordnet und verzögern Speicheroperationen. In Multi-Core-Systemen können dadurch Speicheränderungen nicht sofort für andere Threads sichtbar sein. Selbst bei korrektem Code, wie die beiden oben angesprochenen Algorithmus, kann es zu race conditions kommen.

Wir Unterscheiden zwischen:

- stark geordnetem Speicher (Änderungen sind sofort für alle Prozessoren sichtbar)
- und
- schwach geordnetem Speicher (Änderungen sind nicht sofort sichtbar. Probleme bei der Parallelität).

Thread 1	Thread 2
-----	-----
while(!flag)	turn = 1;
memory_barrier()	memory_barrier();
print turn;	flag = true;

Hier benutzen wir eine Speicherbarriere auf schwach geordneten Speicher Systemen um sicherzustellen, dass das Speichern in der richtigen Reihenfolge stattfindet. Ohne Barriere könnte die CPU die Instruktionen umordnen und es kommt zu einer race condition.

13.10 Deadlocks

2 Thread halten bzw. geben 2 Mutexes frei. Dies geschieht in entgegengesetzter Reihenfolge. Dies kann nicht funktionieren da nun beide Threads auf die Freigabe des jeweilig anderen Mutex wartet, diese aber erst freigegeben werden, wenn einer der beiden Mutexes freigegeben wird. Dies ist in sich ein Widerspruch und somit unmöglich.

Beispiel in C:

p0:

```
for(int i = 0; i < ITERATIONS; ++i) {  
    pthread_mutex_lock(&mutexes[0]);  
    // hält hier  
    // nächstes locking unmöglich  
    pthread_mutex_lock(&mutexes[1]);  
  
    pthread_mutex_unlock(&mutexes[1]);  
    pthread_mutex_unlock(&mutexes[0]);  
}
```

p1:

```
for(int i = 0; i < ITERATIONS; ++i) {  
    pthread_mutex_lock(&mutexes[1]);  
    // und hier (gleichzeitig)  
    // nächstes locking unmöglich  
    pthread_mutex_lock(&mutexes[0]);  
  
    pthread_mutex_unlock(&mutexes[0]);  
    pthread_mutex_unlock(&mutexes[1]);  
}
```

Eine weitere Art von Deadlock ist das Zirkuläre Warten. Dies geschieht wenn eine begrenzte Anzahl an Threads auf das Lock des jeweilig nächsten wartet und sich dadurch ein Zyklus bildet. Es ist unmöglich für einen der Threads fort zu fahren. Dieses Problem gleicht dem Dining Philosophers Problem.

Deadlock Prävention:

- Hold and Wait: Alle Ressourcen am Anfang anfordern und erst am Schluss wieder freigeben. Insgesamt jedoch eine geringe Ressourcennutzung.
- No preemption: Freigeben aller Ressourcen wenn benötigte Ressourcen nicht angefordert werden können. Ausführung fortfahren wenn alle benötigten Ressourcen angefordert werden können.
- Circular Waiting: Ressourcen in aufsteigender Reihenfolge anfordern.

Hier verweise ich erneut auf die Folien von Herr Radush (*Foliensatz 5, Folien 51 bis 56*), da ich keine Ahnung habe was wirklich passiert!

13.11 Bankieralgorithmus

Wir haben n Threads und m Ressourcentypen:

- einen Ressource-Availability Vector: avail_m
- eine Max-Acquire Matrix: max_{nm}
- eine Ressource-Map Matrix: alloc_{nm}
- eine Ressources-Needed Matrix: need_{nm}

Safety Algorithm:

Input: $n, m, \text{avail}[n][m], \text{alloc}[n][m], \text{need}[n][m]$
Output: True/False (safe or unsafe)

```
finish[1..n] = false  
  
i = 1  
while i <= n && finish[i] == false && need[i] <= avail:  
    finish[i] = true  
    avail = avail + alloc[i]  
  
return finish[1] && ... && finish[n]  
⇒ O(n² · m)
```

Ressourcesneeded

Algorithm:

Input: $n, m, \text{avail}[n][m], \text{alloc}[n][m], \text{need}[n][m], \text{req}[i]$ // resources need by Thread $T[i]$
 Output: True/False (safe or unsafe)

```

if req[i] <= need[i]
    error: needs too many Ressources

if req[i] > avail:
    wait(T[i])
else:
    avail = avail - req[i]
    alloc[i] = alloc[i] + req[i]
    need[i] = need[i] - req[i]
    if !safe(n, m, avail, alloc, need):
        avail = avail + req[i]
        alloc[i] = alloc[i] - req[i]
        need[i] = need[i] + req[i]
    
```

Beispiel:

- Sichere Threadreihenfolge:

$$\{T_1, T_3, T_4, T_2, T_0\}$$

- $\text{Req}_1 = (1, 0, 2) < (3, 3, 2) = \text{Avail}$
 - $\text{Avail} = (3, 3, 2) - (1, 0, 2) = (2, 3, 0)$
 - $\text{Alloc}_1 = (3, 0, 2), \text{Need}_1 = (0, 2, 0)$
- Sichere Threadreihenfolge:

$$\{T_1, T_3, T_4, T_0, T_2\}$$

- $\text{Req}_4 = (3, 3, 0) > (2, 3, 0) = \text{Avail}$
 - Anforderung größer als Available
- $\text{Req}_0 = (0, 3, 0) < (2, 3, 0) = \text{Avail}$
 - $\text{Need}_0 = (7, 1, 3), \text{Alloc}_0 = (0, 4, 0)$
 - Unmöglich da unsicherer Zustand

Avail	Thread	Alloc			Need		
		A	B	C	A	B	C
3 3 2	T_0	0	1	0	7	4	3
	T_1	2	0	0	1	2	2
	T_2	3	0	2	6	0	0
	T_3	2	1	1	0	1	1
	T_4	0	0	2	4	3	1

Avail	Thread	Alloc			Need		
		A	B	C	A	B	C
2 3 0	T_0	0	1	0	7	4	3
	T_1	3	0	2	0	2	0
	T_2	3	0	2	6	0	0
	T_3	2	1	1	0	1	1
	T_4	0	0	2	4	3	1

13.12 Deadlock Detection

Immer nur eine Instanz pro Ressourcentyp. Darstellung durch einen Resourczuteilungsgraph, einem Wartegraph und einem Zykluserkennungsalgorithmus.

Input: $n, m, \text{avail}[n][m], \text{alloc}[n][m], \text{req}[n][m]$
 Output: True/False (safe or unsafe)

```

finish[i] = alloc[i] != 0 ? False : True; i = {1..n}

i = 0
while i <= n && finish[i] = False && req[i] <= avail
    finish[i] = True
    avail = avail + alloc[i]

return finish[1] && ... && finish[n]
    
```

Cycledetection Algorithm: $\Rightarrow \mathcal{O}(n^2 \cdot m)$

Beispiel:

- Sicherer Zustand:
 - Kein Deadlock

$\{T_0, T_2, T_3, T_1, T_4\}$

\Rightarrow

Thread	Alloc			Req		
	A	B	C	A	B	C
T_0	0	1	0	0	0	0
T_1	2	0	0	2	0	2
T_2	3	0	2	0	0	0
T_3	2	1	1	1	0	0
T_4	0	0	2	0	0	2

- Unsicherer Zustand:
 - Deadlock

$\{T_0, \dots\}$

\Rightarrow

Thread	Alloc			Req		
	A	B	C	A	B	C
T_0	0	1	0	0	0	0
T_1	2	0	0	2	0	2
T_2	3	0	2	0	0	1
T_3	2	1	1	1	0	0
T_4	0	0	2	0	0	2

13.13 Deadlock Behebung

Um einen Deadlock zu beheben können entweder alle Threads die in dem Deadlock gefangen sind gleichzeitig oder nacheinander terminiert werden. Beim sequenziellen Terminieren wird nach jeder Terminierung überprüft ob das System nun Deadlock frei ist. Die Reihenfolge der Terminierung kann von Threadpriorität, Berechnungszeit und weiteren Faktoren abhängen.

Eine weitere Möglichkeit ist die Präemption von Ressourcen. Hierbei werden Ressourcen die in einem Deadlock gefangen sind Ressourcen entzogen, bzw. einzelne Threads gezielt terminiert.

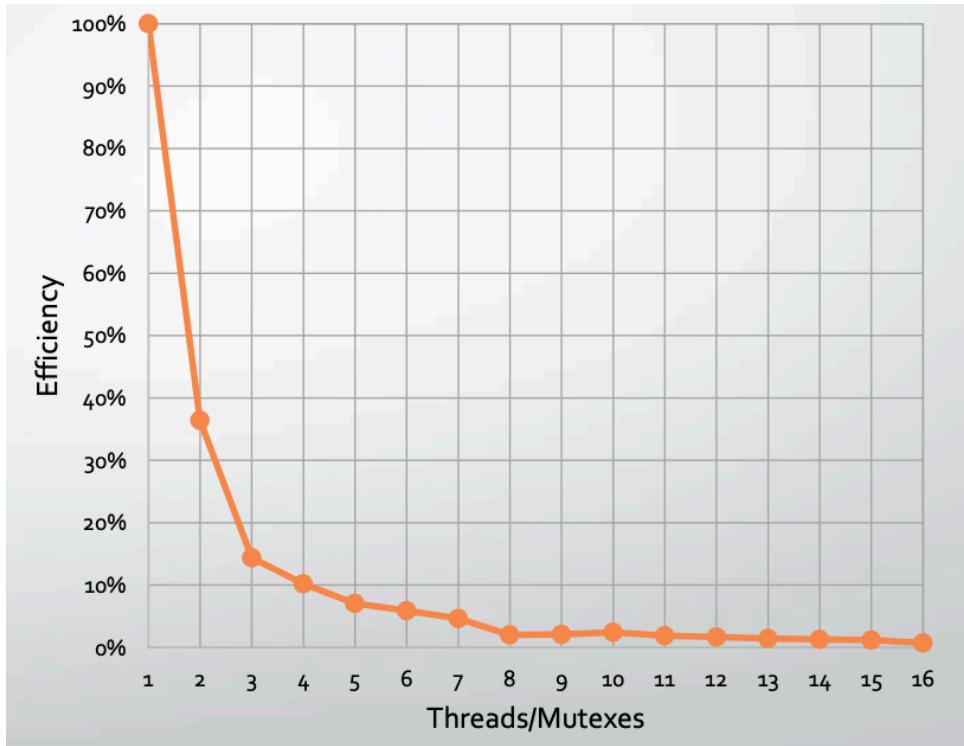
Außerdem können Threads auf einen früheren Zustand zurückgesetzt bzw. neu gestartet werden.

13.14 Livelocks

```
void* thread(void* param) {
    ThreadData* data = param;
    for(int i = 0; i < ITERATIONS; ++i) {
        if(!trylock_all_mutexes(data)) continue;
        ++data->work_done;
        unlock_all_mutexes(data);
    }
    return NULL;
}
```

N Threads versuchen N mutexes zu locken um eine bestimmte Aufgabe zu lösen. Dies ist ähnlich zu einem Deadlock. Die Ausführung des Programms stoppt nicht. Die CPU verbraucht jedoch Ressourcen ohne wirkliche Arbeit zu leisten.

Effizienz:



Effizienz wird als Arbeit definiert die wirklich geschafft wird. Nicht als Versuche Arbeit zu leisten. Wenn hier nun die Anzahl der Threads und somit die Anzahl der Mutexes steigt, nimmt die Effizienz drastisch ab. Durch das zurückziehen finden keinen wirklichen Deadlocks statt.

13.15 Condition Variables

Condition Variables werden genutzt um zu warten/blockieren bis eine bestimmte Kondition wahr wird. Sie können ein Signal erhalten durch das signalisiert wird das eine Bestimmte Kondition erfüllt ist. Die meisten praktischen Anwendung haben aber sogannente *spurious wakeups*. Das bedeutet das die Condition Variable ein Signal bekommt ohne das die Kondition wirklich erfüllt ist. Diese müssen explizit behandelt werden.

- `int pthread_cond_init(pthread_cond_t* cond, pthread_condattr_t* cond_attr);`

initialisiert die Condition Variable, die durch den pointer gegeben wird. cond_attr spezifiziert die Attribute der Condition Variable.

- `int pthread_cond_destroy(pthread_cond_t* cond);`

zerstört die Condition Variable.

- `int pthread_cond_wait(pthread_cond_t* cond, pthread_mutex_t* mutex);`

jede Condition Variable muss mit einem Mutex genutzt werden. Der oben genannte Befehl muss auf einen bereits gehaltenen Mutex folgen. Der Aufruf überprüft ob die Condition Variable ein Signal erhält. Ist dies nicht der Fall bzw. ein *spurious wakeup*-call findet statt, wird der Mutex freigegeben und der Aufruf blockiert die Ausführung. Bekommt die Condition Variable ein Signal werden die folgenden Anweisungen ganz normal ausgeführt. Beim zurückkehren des Aufrufs ist der Mutex entweder immer noch gelocked(Signal erhalten) oder wird geredlocked(Blockieren und dann erst Signal).

Um *spurious wakeups* zu behandeln wird die Kondition nach dem zurückkehren der Funktion erneut überprüft. Dafür wird meistens eine while-Schleife verwendet. Dies bringt den Vorteil, dass bei einer erfüllten Kondition der Aufruf gar nicht erst stattfindet. Ein einfaches Beispiel ist:

```

pthread_mutex_lock(&mutex);
    while(!condition) {
        pthread_cond_wait(&cond, &mutex);
    }
// Critical section
pthread_mutex_unlock(&mutex);

• int pthread_cond_signal(pthread_cond_t* cond);

```

signalisiert genau einem Thread das die Kondition erfüllt ist. Dieser verlässt nun den pthread_cond_wait Aufruf. Warten keine Threads passiert nichts. Warten mehrere Threads verlässt ein unspezifizierter Thread den pthread_cond_wait Aufruf.

- int pthread_cond_broadcast(pthread_cond_t *cond);

signalisiert allen wartenden Threads, und veranlasst diese den pthread_cond_wait Aufruf zu verlassen. Da die Threads durch einen Mutex blocken folgt dies keiner spezifizierten Reihenfolge und die Threads verlassen nacheinander den Aufruf.

Beispiel:

client:

```

pthread_mutex_lock(&data->mtx);           // Aquire mutex
data->cnt = strlen(message);             // Write
memcpy(&data->buf, message, data->cnt); // message
data->available = true;                  // Set condition
pthread_mutex_unlock(&data->mtx);         // Release mutex
pthread_cond_signal(&data->cond);         // Signal condition variable

pthread_mutex_lock(&data->mtx);           // Aquire mutex
while(!data->processed) {                // Wait until condition
    pthread_cond_wait(&data->cond, &data->mtx); // Release mutex and wait
}                                         // Mutex will be re-aquired
printf("%.*s\n", (int)data->cnt, data->buf); // Print
pthread_mutex_unlock(&data->mtx);         // Release mutex

```

server:

```

pthread_mutex_lock(&data->mtx);           // Aquire mutex
while(!data->available) {                // Wait until condition
    pthread_cond_wait(&data->cond, &data->mtx); // Release mutex and wait
}
for(size_t i = 0; i < data->cnt; ++i) {   // Process
    data->buf[i] = toupper((unsigned char)data->buf[i]); // the
}                                         // data
data->processed = true;                  // Set condition
pthread_mutex_unlock(&data->mtx);         // Release mutex
pthread_cond_signal(&data->cond);         // Signal condition variable

```

13.16 Semaphores

Semaphores sind ähnlich zu Mutexes. Semaphoren sind im grössten Sinne ein Zähler. Dieser Zähler startet meistens bei 0 und zeigt meistens an, ob Ressourcen verfügbar sind. Sie ermöglichen 2 Operationen.

- post/release: welches den Zähler um 1 erhöht und die Semaphore semantisch unlocked, und
- wait/acquire: welches den Zähler um 1 verringert und die Semaphore sematisch locked.

Das Verringern ist nur möglich wenn der Zähler > 0 ist. Eine Semaphore die zu jedem Zeitpunkt den Wert 0 bzw. 1 hält wird auch Binäre Semaphore genannt bzw. Mutex genannt.

- int sem_init(sem_t* sem, int pshared, unsigned int value);

initialisiert die Semaphore. Wenn pshared = 0 wird die Semaphore zwischen Threads geteilt. Ist pshared ≠ 0 dann wird die Semaphore zwischen Prozessen geteilt. Die Semaphore muss beim teilen mit Prozessen im Shared memory liegen. Das initialisieren muss nur einmal und nicht pro Thread/Prozess erfolgen. Value gibt den Startwert der Semaphore an.

- `int sem_destroy(sem_t* sem);`

zerstört die Semaphore. Muss nur einmal und nicht pro Thread/Prozess erfolgen.

- `int sem_wait(sem_t* sem);`

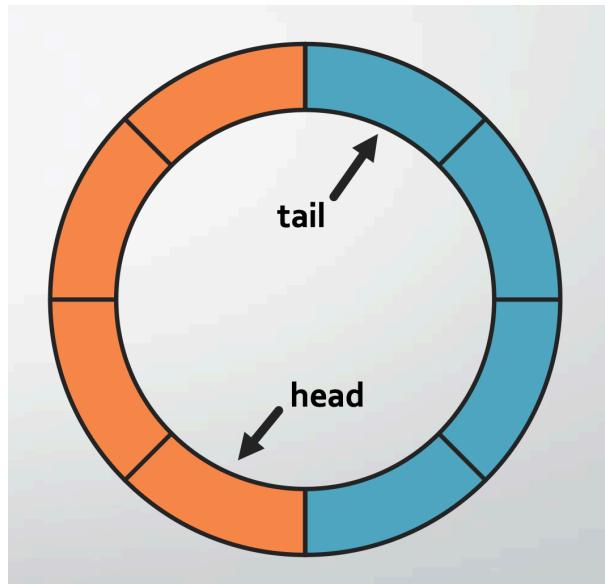
verringert den Wert der Semaphore. Ist der Wert der Semaphore 0, blockiert der Aufruf bis ein anderer Thread die Semaphore erhöht.

- `int sem_post(sem_t* sem);`

erhöht den Wert der Semaphore. War der Wert der Semaphore 0, wird ein blockender `sem_wait` call entblocked.

13.17 Ring Buffer

Ein Ring Buffer wird verwendet um eine Queue zu implementieren. Er hat ein read-end(tail) und ein write-end(head). Wenn head = tail ist der Ring Buffer leer. Wenn $(\text{head} + 1) \bmod N = \text{tail}$ ist der Ring Buffer voll. Die Kapazität des Ring Buffers ist $N - 1$, da sonst "leer" und "voll" nicht klar definiert wäre.



```

struct RingBuffer {
    int head;
    int tail;
    Data data[BUFFER_SIZE];
};

typedef struct RingBuffer RingBuffer;

void ring_buffer_init(RingBuffer* buf) {
    buf->head = 0;
    buf->tail = 0;
}

bool ring_buffer_is_full(RingBuffer* buf) {
    return ((buf->head + 1) % BUFFER_SIZE) == buf->tail;
}

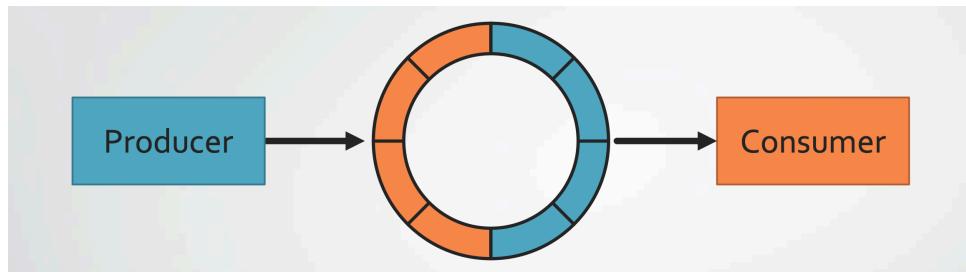
bool ring_buffer_is_empty(RingBuffer* buf) {
    return buf->head == buf->tail;
}

bool ring_buffer_push(RingBuffer* buf, const Data* data) {
    if(ring_buffer_is_full(buf)) return false;
    buf->data[buf->head] = *data;
    buf->head = (buf->head + 1) % BUFFER_SIZE;
    return true;
}

bool ring_buffer_pop(RingBuffer* buf, Data* data) {
    if(ring_buffer_is_empty(buf)) return false;
    *data = buf->data[buf->tail];
    buf->tail = (buf->tail + 1) % BUFFER_SIZE;
    return true;
}
}

```

13.18 Producer - Consumer Problem



Das Producer-Consumer Problem ist ein klassisches Problem in der Informatik. Der Producer und die Consumer sind Threads bzw. Prozesse, die parallel arbeiten. Der Producer erzeugt ununterbrochen Daten und speichert diese in dem Ringbuffer. Die Consumer nehmen sich daten aus dem Ring Buffer und verarbeiten diese. Auf die Frage “Wird hier Synchronisation benötigt” stellen wir uns ein Beispiel eines Buffets vor.

Auf diesem Buffet wird ständig neues Essen gekocht und gegessen. Der Koch fügt neues Essen hinzu wenn Platz auf dem Buffet ist bzw. die Gesamtanzahl der benötigten Essen noch nicht erschöpft ist. Jedoch ist diese Gesamtanzahl viel größer als Platz auf dem Buffet ist. Die Consumer bzw. Gäste essen, solange Essen auf dem Buffet ist bzw. die Gesamtanzahl der Essen konsumiert wurde. Wenn das Buffet leer ist wartet der Consumer auf das nächste Gericht.

Ohne Synchronisation greifen mehrere Gäste gleichzeitig auf das Buffet zu. Dabei können sie sich in die Quere kommen. Am Beispiel: Es befindet sich ein Essen auf dem Buffet. 2 oder mehr Gäste sehen dieses essen und konsumieren dies (virtuell) gleichzeitig. Entweder die Gäste konsumieren essen das eigentlich gar nicht da ist, oder sie *greifen* sich ein Essen das schon nicht mehr vorhanden ist. In beiden Fällen findet eine **race condition** statt bzw. potentiell ein Error statt.

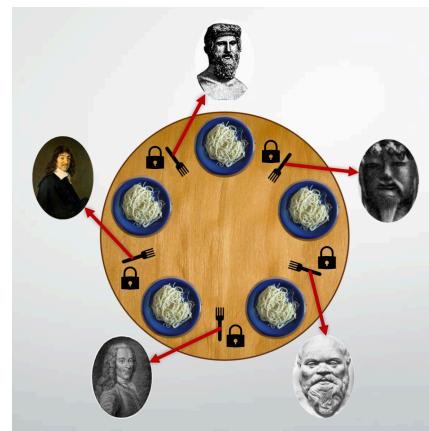
Das Code Beispiel zu diesem Problem kann im *Foliensatz 4, Folien 37-40* nachvollzogen werden. Um das Problem zu beheben verwenden wir 2 Semaphoren, die anzeigen wie viel freier Platz auf dem Buffet ist, und wie viele Essen sich auf dem Buffet befinden. Zudem benötigen wir einen Mutex der den Zugriff auf das Buffet schützt und einen Atomic Boolean der den Consumern anzeigt ob diese Stoppen sollen.

13.19 Dining Philosophers Problem



An einem Tisch befinden sich N Philosophen. Jeder Philosoph hat eine Portion Spaghetti vor sich. Auf dem Tisch befinden sich zudem $N - 1$ Essstäbchen. Ein Philosoph wechselt immer von Denken zu Essen. Um zu essen muss jeder Philosoph die 2 Essstäbchen neben seinen Spaghetti aufheben. Das Aufheben kann durch einen Mutex für jedes Essstäbchen synchronisiert werden.

Dies kann jedoch einfach zu einem Deadlock führen, wenn alle Philosophen gleichzeitig zum Beispiel das Stäbchen rechts seines Essens aufhebt. Nun ist für keinen der Philosophen mehr möglich ein weiteres Stäbchen aufzuheben und seine Spaghetti zu essen. Wir befinden uns in einem Deadlock!



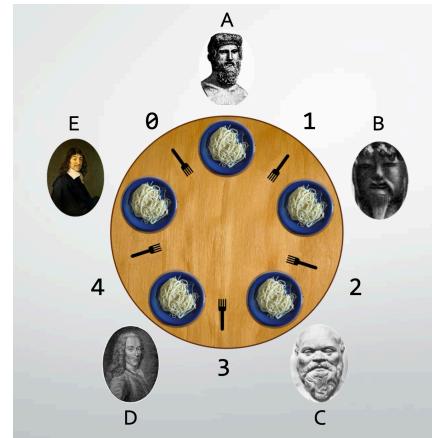
Lösungsansätze:

- **Centralized Locking:**

Es wird eine zentraler Mutex verwendet um die Kontrolle der Esstäbchen zu schützen. Jeder Philosoph kann den Mutex halten und versuchen die beiden Stäbchen links und rechts aufzuheben. Gelingt dies werden die Stäbchen aufgehoben und der Mutex wird freigegeben. Gelingt dies nicht werden die Stäbchen zurückgelegt und es wird zu einem späteren Zeitpunkt erneut versucht. Durch das warten kann es jedoch zu einem Livelock kommen und manche Philosophen essen nie und Verhungern.

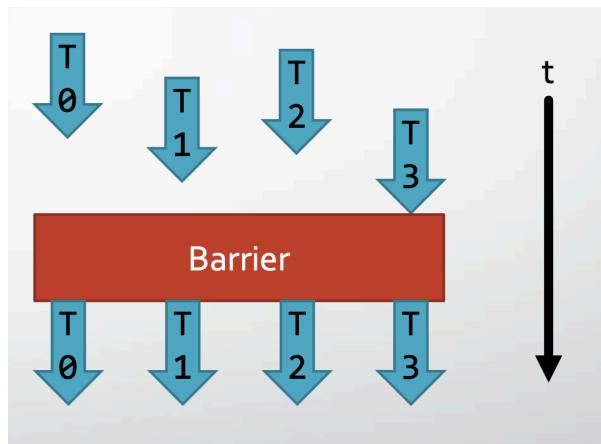
- **Ordering:**

Die Stäbchen bekommen eine Reihenfolge. Jeder Philosoph hebt erst das Stäbchen mit der niedrigeren ID auf. Hier kommt es jedoch dazu, dass manche Philosophen häufiger essen als andere. In dem Beispiel isst Philosoph häufiger als andere.



13.20 Barriers

Barriers synchronisieren alle partizipierenden Threads, indem sie die Ausführungen solange blockieren bis alle Threads an der Barriere angelangt sind.



- `int pthread_barrier_init(pthread_barrier_t* barrier,
 const pthread_barrierattr_t* attr,
 unsigned count);`

initialisiert die Barriere mit den spezifizierten Attributen und dem count als Anzahl der partizipierenden Threads.

- `int pthread_barrier_destroy(pthread_barrier_t* barrier);`

zerstört die Barriere.

- `int pthread_barrier_wait(pthread_barrier_t* barrier);`

das Ankommen an der Barriere wird auch warten genannt. Der Aufruf blockt solange bis alle Threads an der Barriere angekommen sind. Ein unspezifizierte Thread gibt den Wert

PTHREAD_BARRIER_SERIAL_THREAD zurück. Dies kann benutzt werden um bestimmte Logik pro Barrier auszuführen bzw. pro Loop/Step.

Beispiel:

```
void* thread(void* arg) {
    ThreadData* thread_data = arg;
    SharedData* data = thread_data->data;
    for(int t = 0; t < NUM_TIMESTEPS; ++t) {
        compute(thread_data);
        if(pthread_barrier_wait(&data->barrier) == PTHREAD_BARRIER_SERIAL_THREAD) {
            swap_buffers(data);
        }
        pthread_barrier_wait(&data->barrier);
    }
    return NULL;
}
```

13.21 Interrupt-based Synchronisation

Generelles Vorgehen:

1. Eintritt in Critical Section: Deaktivieren der Interrupts
2. Ausführen der Critical Section
3. Austritt aus Critical Section: Reaktivieren der Interrupts

Auf Single-Core-Systemen äußerst effizient und einfach zu Implementieren. Sie garantiert den exklusiven Zugriff in einer Critical-Section. Auf Multi-Core-Systemen ungeeignet, da das De-/Aktivieren der Interrupts **aller** CPU's und das sperren des Systembus extrem teuer ist.

14 Synchronisation (Hardware)

Bei Single-Core-Systemen verwenden wir den Interrupt-based Synchronisation Ansatz.

Auf Multi-Core-Systemen verwenden wir Hardwareanweisungen, wie:

- test_and_set
- compare_and_swap

und atomare Variablen und Funktionen. Somit haben wir eine Synchronisation durch die Hardware.

- test_and_set

```
atomic_bool test_and_set(atomic_bool *lock) {
    atomic_bool old = *lock;
    *lock = true;
    return old
}
```

Beispielnutzung:

```
atomic_bool lock = false;

do {
    // spinlock / busy waiting
    while (test_and_set(&lock));

    // critical section

    lock = false;
    // remainder section

} while (true)
```

- **compare_and_swap**

```
atomic_int compare_and_swap(atomic_int *val,
                           int expected, int new_val) {
    temp = *val;
    if(*val == expected)
        *val = new_val;
    return temp;
}
```

Beispielnutzung:

```
atomic_int lock = 0;

do {
    // spinlock / busy waiting
    while (compare_and_swap(&lock), 0, 1) == 1;

    // critical section

    lock = 0;
    // remainder section

} while (true)
```

- **begrenztes warten mit compare_and_swap**

```
atomic_bool wait[N] = { false };
atomic_int lock = 0;

// Prozess i
do {
    wait[i] = true; // wait[i] = true; Prozess i wartet auf
                    // kritischen Bereich

    // spinlock / busy waiting
    while (wait[i] && compare_and_swap(&lock), 0, 1) == 1;
    wait[i] = false;

    // critical section

    j = (i + 1) % N;
    while((j != i) && !wait[j])
        j = (j + 1) % N;

    if(j == i) lock = 0; // release
    else wait[j] = false; // transfer lock
    // remainder section

} while (true)
```

14.1 Spinlock / Busy waiting

Bei einem Spinlock oder auch busy waiting wartet ein Prozess darauf, dass eine bestimmte Kondition erfüllt wird. Dabei führt er bei jeder Überprüfung einen Rücksprung aus, und prüft erneut. Die verschwendet CPU und ist schlecht in Timesharing-Systemen. Vorteil ist das kein Contextswitch stattfinden muss. Sie sind außerdem gut für kurzes Warten in Multi-Core-Systemen.

Wie bereits oben gesehen kann durch die Verwendung von Spinlocks eine Synchronisation, auf Kosten von CPU Ressourcen stattfinden.

14.2 Semaphoren

Beschreibung siehe Semaphores 13.16.

14.2.1 Implementierung mittels Spinlock:

```
void wait(sem) {
    while (sem <= 0);
    sem--;
}

void signal(sem) {
    sem++;
}
```

Beispielnutzung:

```
semaphore sem = 1
```

```
wait(sem)
// critical section
signal(sem)
```

14.2.2 Implementierung ohne Spinlock:

mittels Semaphore Warteschlange:

```
typedef struct _process {
    // true == busy
    bool locked;
    _process *next;
} Process;

typedef struct {
    int value;
    struct Process *list;
} Semaphore;

// last process in the queue
Process *last = NULL;

• lock

lock(Process *l, Process *p) {           // last process
    p->next = NULL;                      // p
    Prozess pred = fetch_and_store(l, p); // p: new tail: pred = l; l = p;
    if(pred != NULL) {                   // Locked; List not empty
        p->locked = true;              // Locked by pred
        pred->next = p;                // insert p
        sleep();                       // wait on pred
    }
}

• unlock

unlock(Process *l, Process *p) {
    if(p->next == NULL) {             // no successor in the list
        if(compare_and_swap(l, p, NULL)) // if (l == p) l = NULL
            return;
        while(p->next == NULL);       // wait in succ
    }
    p->next->locked = false;         // unlock succ
    wakeup(p->next);
}
```

- process p

```

void wait(Semaphore *sem) {
    sem->value--;
    if(sem->value < 0)
        lock(last, p);
}
void signal(Semaphore *sem) {
    sem->value++;
    if(sem->value <= 0)
        unlock(last, p);
}

```

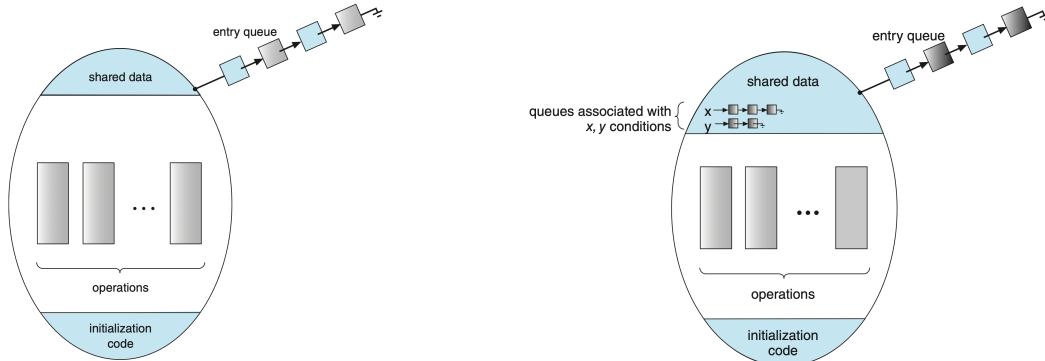
Falls jemand das hier nicht versteht, würde ich gerne auf die Folien von Herr Radush verweisen, mit denen ich genau so viel wie hier steht verstanden habe. (Nicht besonders viel!)

14.3 Monitor und Condition Konstrukt (Java)

Ein Monitor ist ein ADT der Daten und vordefinierte Funktionen bereitstellt. Diese Funktionen sind mutual exclusive. Die Daten auf die Zugriffen werden kann sind nur lokal vorhanden und somit parallel zugreifbar. Der Monitor stellt sicher das zu jedem Zeitpunkt maximal ein Prozess aktiv die Daten des Monitors manipuliert.

Um dies zu ermöglichen wird ein weiteres Konstrukt, die Condition benötigt. Die einzigen 2 Operationen auf den Conditions sind signal() und wait(), wobei wait mit einem Parameter von Typ Condition aufgerufen werden kann. Dies ermöglicht Bedingtes Warten.

Die Conditions werden beim schlafen bzw. warten auf Ausführung in einer Queue gespeichert.



Beispiel (Serial):

```

monitor Serial {
    condition x = 0;
    bool done = false;
    void f1() {
        Anweisung s1;
        done = true;
        x.signal();
    }

    void f2() {
        if(done == false)
            x.wait();
        Anweisung s2;
    }
}

```

Prozess P_1 :

```

Serial s1;
s1.f1();

```

Prozess P_2 :

```

Serial s2;
s2.f2();

```

Implementierung mit Semaphoren:

```
monitor ParallelSem {
    semaphore mutex = 1; // Monitor lock
    semaphore next = 0; // Lock of Conditions
    int next_count = 0; // Count Conditions

    Function f() {
        // calculation
        if (next_count > 0)
            signal(next);
        else
            signal(mutex);
    }
}

condition X {
    semaphore x_sem = 0;
    int x_count = 0;

    x.wait() {
        x_count++;
        if(next_count > 0)
            signal(next);
        else
            signal(mutex);

        wait(x_sem);
        x_count--;
    }

    x.signal() {
        if(x_count > 0) {
            next_count++;
            signal(x_sem);
            wait(next);
            next_count--;
        }
    }
}
```

Ressourcenzuteilung:

```
monitor ResourceAllocator {
    boolean busy = false;
    condition x;

    void acquire(int time) {
        if(busy)
            x.wait(time);
        busy = true;
    }

    void release() {
        busy = false;
        x.signal();
    }
}

ResourceAllocator R;
Time t;

R.acquire(t); // acquire Ressources
R.release();
```

Dining Philosophers Monitor basiert:

```
monitor DiningPhilosophers {
    enum { THINKING, HUNGRY, EATING } state[N];
    condition self[N];

    int left(int i) {
        return (i + N - 1) % N;
    }

    int right(int i) {
        return (i + 1) % N;
    }

    void test_and_eat(int i) {
        if(state[i] == HUNGRY &&
           state[left(i)] != EATING &&
           state[right(i)] != EATING) {
            state[i] = EATING;
            self[i].signal();
        }
    }
}

initialization() {
    for(int i = 0; i < N; i++)
        state[i] = THINKING;
}

void pickup(int i) {
    state[i] = HUNGRY;
    test_and_eat(i);
    if(state[i] != EATING)
        self[i].wait();
}

void putdown(int i) {
    state[i] = THINKING;
    test_and_eat(left(i));
    test_and_eat(right(i));
}
```

Auch hier habe ich die Folien nicht richtig verstanden und eigentlich nur kopiert!

15 Alternative Ansätze der Synchronisation

Auf neuen Systemen ist wie bereits mehrfach angesprochen ein paralleles Berechnen von Problemen zum Standard geworden. Um dies zu ermöglichen werden meist Mutexes bzw. Semaphoren und andere Konzepte verwendet. Diese sind jedoch nicht sonderlich schnell und schützen nicht vor deadlocks, race conditions und liveness hazards wie zum Beispiel Livelocks. Außerdem verliert man bei zunehmender Thread Anzahl performance durch erschwerte jedoch nötige Synchronisation.

Um diese Probleme zu lösen wurden neue Wege gesucht um das Programmieren zu vereinfachen und trotzdem die Performance der Programme zu erhalten.

Wir gehen nur auf das Konzept des Transactional Memorys ein. Weitere Ansätze sind OpenMP und Funktionale Programmiersprachen (nachzulesen in der zugrundeliegenden Quelle Chapter 7)

15.1 Transactional Memory

Ein Ansatz ist der Transaktionelle Speicher. Ein Traditioneller Ansatz (mit Mutexes) um Daten im Shared memory zu verändern wäre wie folgt:

```
void update() {
    acquire();

    // update shared memory

    release();
}
```

Da dies jedoch potentielle Deadlocks und andere Fehler mit sich bringt kann auf einen anderen Ansatz gesetzt werden.

Wir erstellen ein neues Konstrukt `atomic{S}`, welches sicherstellt, dass Operationen auf S transaktionsell stattfinden.

```

void update {
    atomic {
        // update shared memory
    }
}

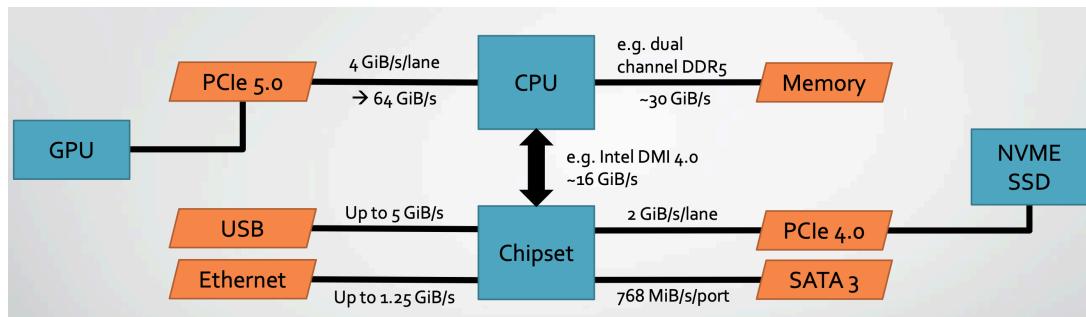
```

Dadurch ist nichtmehr der Programmierer sonder der Mechanismus an sich zuständig für die Synchronisation der Operation.

Dieses Konstrukt wird meist vom Compiler generiert und durch Hardware Transaktionsspeicher im Cache unterstützt.

16 Input/Output (I/O)

16.1 Architektur



Unter externen Geräten(Devices) verstehen wir alles, was nicht CPU bzw. RAM ist (z.B. GPU, Networkcard, SSDs, ...). Diese Geräte sind meistens bestimmten Speicher Adressen zugeordnet.

16.2 Speicher

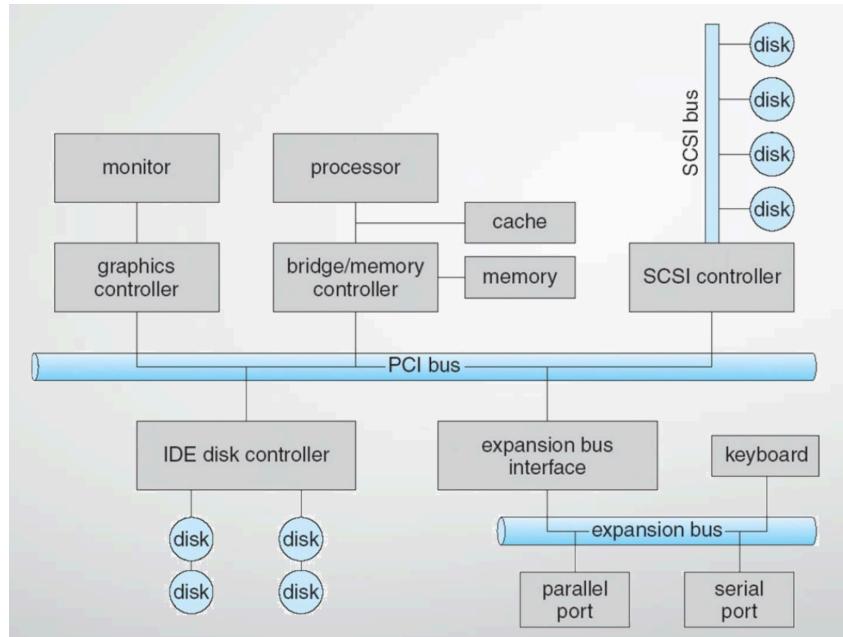
- SRAM (Static Random Access Memory)

Wird meist als Cachespeicher benutzt. Komplex und teuer!

- DRAM (Dynamic Random Access Memory)

Relativ günstig. Dichte Speicher Anordnung. Kapazitoren verlieren Ladung und müssen periodisch neu “beschrieben” werden.

16.3 I/O Bus



Interconnect welches für die Kommunikation mit vielen verschiedenen Geräten gebraucht wird.

16.4 Device Communication

Wir unterscheiden zuerst zwischen Speicher gemapptem I/O (*Memory mapped I/O*), gemappten Gerät Speicher (*Mapped device memory*), Port gemapptem I/O (*Port mapped I/O*) und DMA (*Direct memory access*) unterschieden.

Bei *Memory mapped I/O* wird einem Gerät eine spezifizierte Speicheradresse zugeordnet und read/write kann genau wie auf den restlichen RAM genutzt werden. Die einzelne Semantic ist je nach Gerät spezifisch. Es handelt sich hier jedoch *nicht wirklich um RAM*.

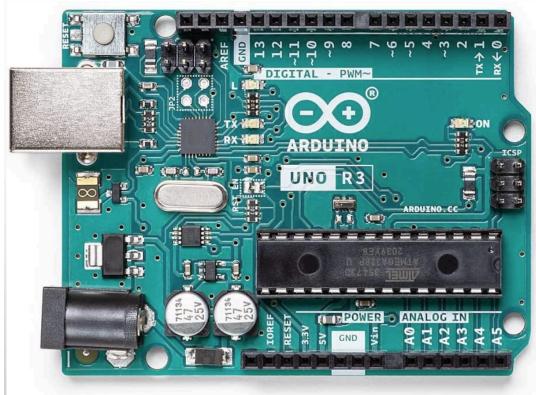
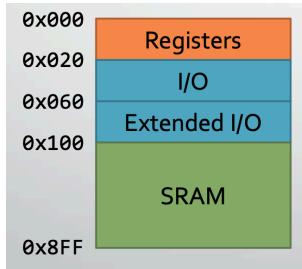
Bei *Mapped device memory* wird eine spezifische Speicheradresse, dem RAM des externen Geräts zugeordnet.

Bei *Port mapped I/O* wird von manchen Architekturen ermöglicht. x86 stellt zu Beispiel die in/out Befehle zur Verfügung mit denen I/O ports/pins gesteuert werden können. Der RAM und die I/O haben einen getrennten Addressraum.

Bei *DMA* schreibt/liest das Gerät automatisch in/aus den/dem RAM. Die CPU ist hier nur der Initiator und kann während dem Transfer andere Aufgaben erledigen.

16.4.1 Beispiel *Memory Mapped I/O* an einem Arduino Uno ATmega328P:

CPU Register, die den Status der I/O pins beinhalten. 8 GPIO (General purpose I/O) pins sind verbunden um einen **GPIO Port** zu bilden.



Der Zugriff erfolgt wie auf normalen Speicher:

```
volatile uint8_t* portb_direction_addr = (volatile uint8_t*)0x24;
volatile uint8_t* portb_addr = (volatile uint8_t*)0x25;

*portb_direction_addr = 0b11111111; // Configure all pins on port B as output

*portb_addr = 0b11111111;           // Set all pins on port B high
*portb_addr = 0b00000000;           // Set all pins on port B low
```

Der ATmega328P hat 3 verschiedene Addressräume.

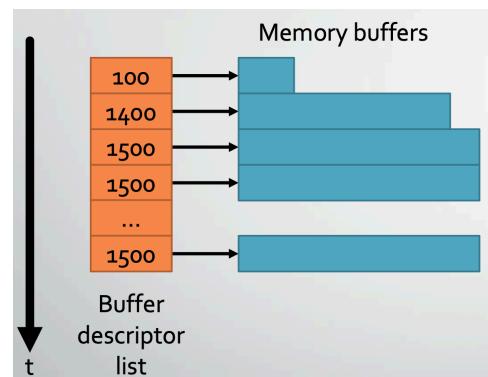
- Regulären RAM: Inklusive dem *Memory mapped I/O*.
- Flash memory: welches das Programm und den Bootloader enthält.
- Non-volatile EEPROM: Dieser kann genutzt werden um permanent Daten zu speichern.

Alle Adressen starten bei 0x0000. Der gewählte Addressraum wird durch eine Assembly Instruktion angegeben.

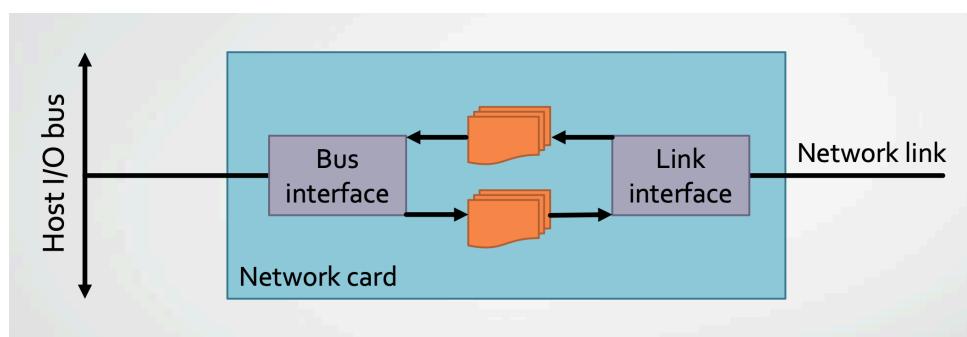
16.4.2 Beispiel DMA-Buffer:

Die CPU wird nur genutzt um den Transfer zu initialisieren und beim Abschluss zu überprüfen. Der Transfer an sich ist unabhängig.

Damit dies möglich ist braucht es gewisse Buffer, damit die initialisierten und in einer Queue wartenden Transfers direkt stattfinden können.



Wir betrachten nun das Beispiel an einem NIC (Network Card Interface).



Das Link-Interface kommuniziert mit der Network hardware. Das Bus-Interface nutzt DMA um Pakete direkt in den RAM zu lesen/schreiben. Für die Operations Queue wird eine FIFO für das Lesen und das Schreiben verwendet. Der Gerätetreiber stellt dem Kernel die benötigte Funktionalität zur Verfügung (reset, ioctl, output, interrupt, read, write, ...). Um das Gerät zu synchronisieren (lesen, schreiben, usw.) kann die einfache Methode des **Polling** oder ein Interrupt-based Ansatz verwendet werden.

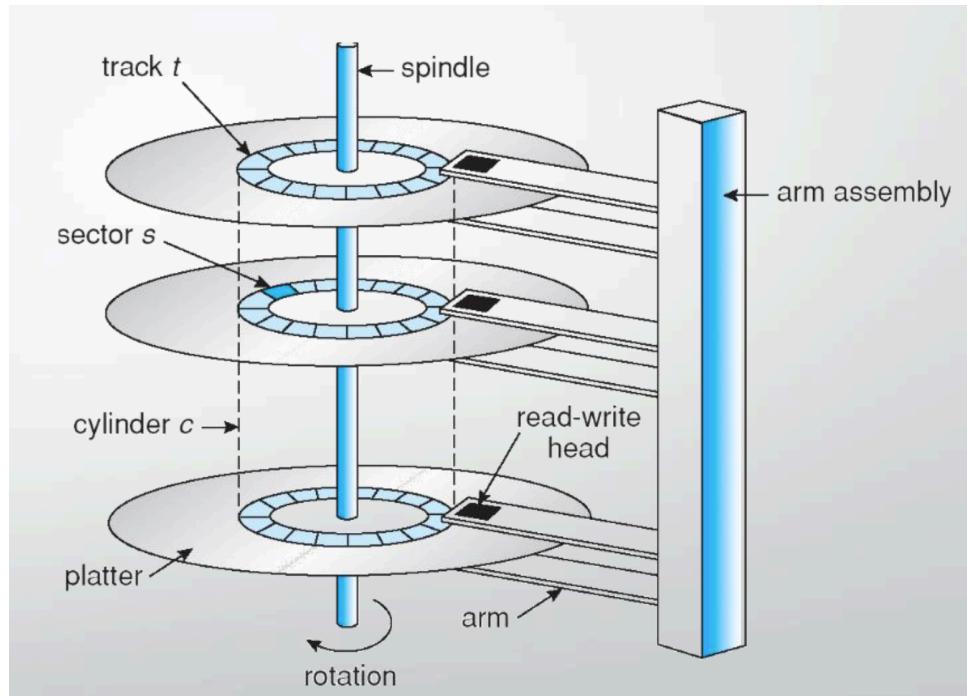
Polling: Überprüft ununterbrochen ob die Operation beendet wurde. Das benötigt viel CPU Resourcen bzw. lässt die CPU busy waiting. Außerdem können keine 2 oder mehr Operation gleichzeitig ausgeführt werden.

Interrupt-based Device Synchronisation: Das Gerät triggert einen CPU Interrupt um etwas zu signalisieren. Der Interrupthandler wird aufgerufen und der Gerätetreiber kann den Grund des Interrupts sofort prüfen. Dadurch können CPU und Gerät unabhängig voneinander Arbeiten. Kommt es zu häufigen Interrupts kann die CPU Geschwindigkeit stark beeinflusst werden, da jedes mal ein Contextswitch getriggert wird. Dieser Ansatz wird von den meisten Betriebssystemen genutzt.

17 Speicher Hardware und Software

17.1 HDD

HDDs (Hard disk drive) sind magnetische Datenträger. Sie nutzen mehrere gestapelte platters, die in konzentrische tracks aufgeteilt werden. Mehrere sectors erzeugen einen track. Die gleichen tracks von allen platters werden cylinder genannt. Die Schreib und Leseköpfe lsens/schreiben auf den cylinder. In einer HDD wird ein großer Teil der Daten für die Error Korrektur reserviert.



Die Geschwindigkeit einer HDD hängt stark von der Ordnung und dem Ort des aufgerufenen Speicher Segments ab. Der sequenzielle Abruf ist wesentlich schneller als der willkürliche Zugriff.

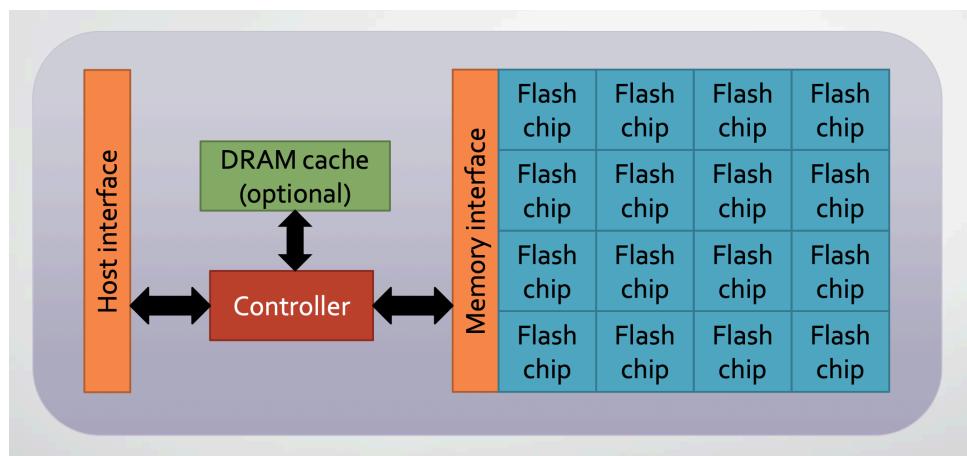
17.2 SSD

SSDs (Solid state drives) sind Flash Datenträger, ohne jegliche Bewegten Teile. Sie haben die HDDs in so gut wie allen Bereichen und Anwendungen ersetzt. Die Daten werden durch elektrische Ladung gespeichert und halten über 10+ Jahre. Speicherzellen können beinahe unendlich oft gelesen werden. Das schreiben ist jedoch auf 10.000 bis 100.000 beschränkt. Eine erhöhte Redundanz und Abnutzungsausgleich kompensiert diesen Nachteil. SSDs sind schnell, zuverlässig und Erschütterungsresistent.

Abnutzungsausgleich: (*Wear leveling*)

Teilt die Schreibvorgänge auf den gesamten Speicher gleichmäßig auf, um die Lebensdauer des Geräts zu verlängern. Dies benötigt eine komplexe Addresslogik, die vom Flash controller erledigt wird. Ist die SSD fast voll kann dies zu Performance Verlust führen.

Architektur:



Der Flash controller ist zuständig den Speicher als Zusammenhängenden Bereich an das System zu übergeben.

17.3 NAND und NOR Flash

NAND Flash ist dicht bedeckt mit einer hohen Kapazität pro Speicherchip. NAND ist anfälliger für Fehler und benötigt Error correction. Die Schreibgeschwindigkeit ist hoch.

NOR Flash hat eine schnelle Lesegeschwindigkeit und kann genutzt werden um Programmcode direkt in-place auszuführen. Es ist zuverlässig und benötigt virtuell keine Error correction. NOR Flash wird häufig für Embedded Systems verwendet.

Typen:

- SLC: (Single) 1 Bit
- MLC: (Multi) 2 Bits
- TLC: (Triple) 3 Bits
- QLC: (Quad) 4 Bits

Das Erhöhen der Bits pro Zelle verringert die Schreibgeschwindigkeit des NAND-Flash. Die Kosten pro Bit sind jedoch geringer.

17.4 RAID

RAID oder auch Redundant Array of Independent Disks ist das kombinieren von mehreren Datenträger zu einem virtuellen Datenträger. Es gibt Unterschiedliche RAID Level, die unterschiedliche Aspekte optimieren. Die Hardware Implementation ist teuer, aber schnell, hat keinen

CPU-Overhead, aber ist nicht portable, da RAID Kontroller von verschiedenen Herstellern verschieden funktionieren. Die Software Implementation ist kostenlos, auf modernen CPUs genau so schnell wie eine Hardware Implementation und portable.

Die Benutzung von RAID ist nicht gleich einem BACKUP!

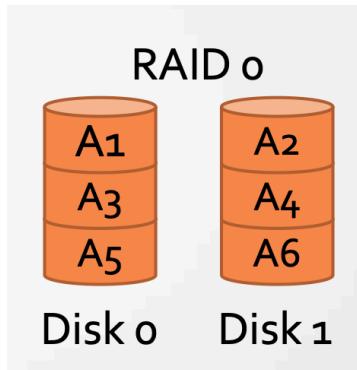
17.5 RAID Standard Levels

$$N = \# \text{disks}$$

$$D = \text{disk size}$$

17.5.1 RAID 0

(*Striping of blocks across disks*)



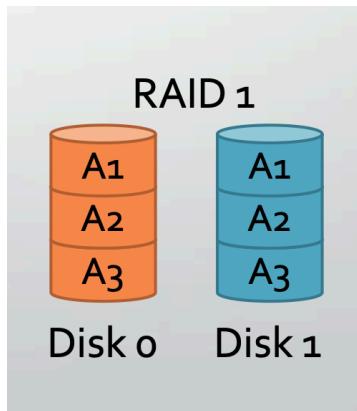
Keine Redundanz. Alle Speicherblöcke werden zu einem großen Block zusammengefasst. Die Lese-/Schreibgeschwindigkeit verbessert sich.

Nutzbare Speicher:

$$N \cdot D$$

17.5.2 RAID 1

(*Mirroring of blocks across disks*)



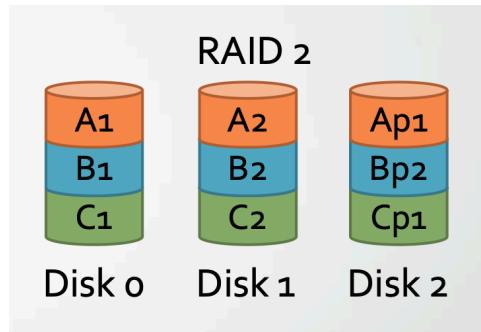
Redundanz bis zum Versagen eines kompletten Datenträgers. Die Lesegeschwindigkeit verbessert sich.

Nutzbare Speicher:

$$D$$

17.5.3 RAID 2

(*Striping of bits with hamming code error correction*)

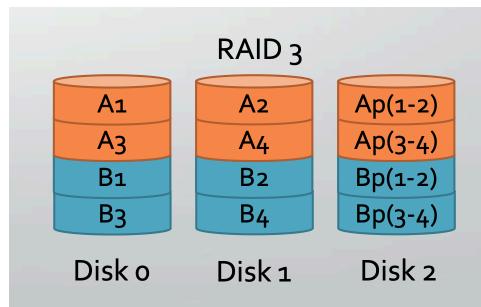


Es werden mindestens 3 Datenträger benötigt. Durch die Error correction ist es möglich, die Daten eines Datenträgers nach dem Versagen wieder herzustellen. Die Lese und Schreib Geschwindigkeit verändert sich nicht. Dieser Ansatz wird in der Praxis so gut wie nie genutzt.
Nutzbarer Speicher:

$$(N - 1) \cdot D$$

17.5.4 RAID 3

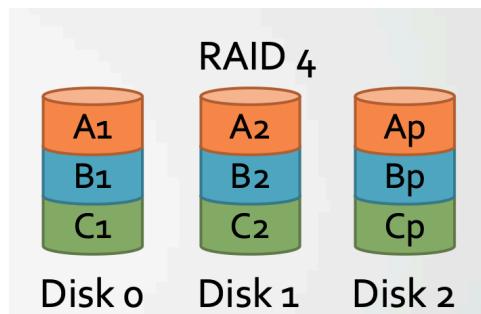
(*Striping of bytes with parity disk*)



Die selben Eigenschaften wie RAID 2, jedoch wesentlich einfachere Berechnung.

17.5.5 RAID 4

(*Striping of blocks with parity disk*)



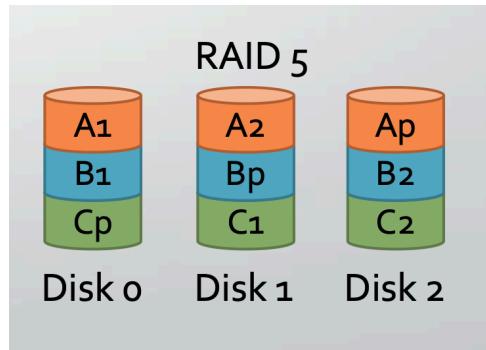
Mindestens 3 Datenträger, komplette Wiederherstellung eines fehlerhaften Datenträgers. Die Schreib Geschwindigkeit ist nicht erhöht, jedoch die Lesegeschwindigkeit. Wird in der Praxis selten genutzt.

Nutzbarer Speicher:

$$(N - 1) \cdot D$$

17.5.6 RAID 5

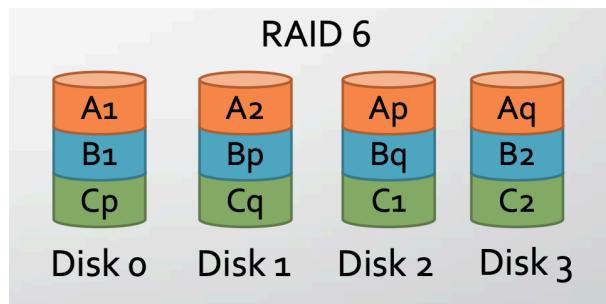
(*Striping of blocks with distributed parity*)



Die selben Eigenschaften wie RAID 4. Durch die Verteilung sind Lese- und Schreibgeschwindigkeit verbessert. Wird häufig genutzt und ersetzt (eigentlich) RAID 2-4.

17.5.7 RAID 6

(*Extension of RAID 5, adding another parity block*)

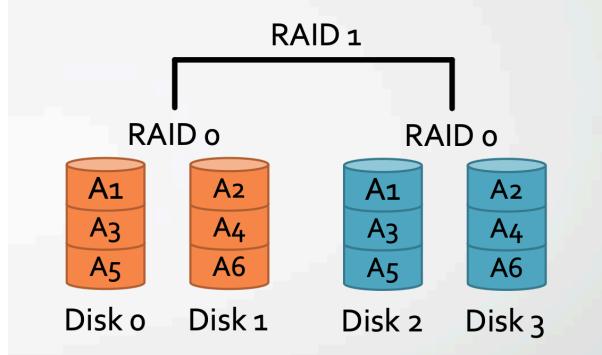


Es werden mindestens 4 Datenträger benötigt. Durch die zweifache Verteilung können 2 fehlerhafte Datenträger komplett wieder hergestellt werden. Wird häufig genutzt um weitere Sicherheit zu gewährleisten. Z.b. nach dem ersetzen eines fehlerhaften Datenträgers.

17.6 RAID Hybrid Levels

17.6.1 RAID 01 / RAID 10

(*Combination of RAID 0 and 1*)



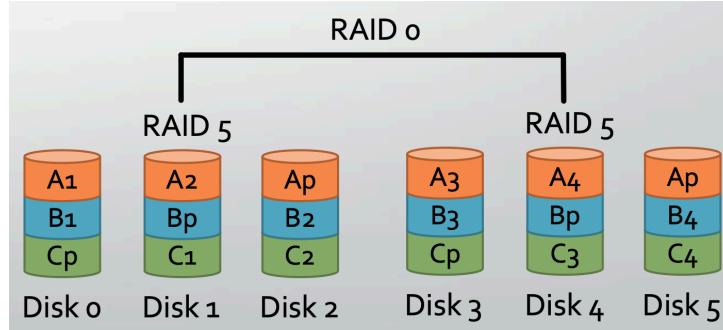
Es werden mindestens 4 Datenträger benötigt. Es handelt sich um eine Kombination von RAID 0 und RAID 1. (*Mirror of stripes*) bzw. (*Stripe of mirrors*).

Nutzbarer Speicher:

$$\left(\frac{N}{2}\right) \cdot D$$

17.6.2 RAID 50

(*Combination of RAID 5 and 0*)



Es werden mindestens 6 Datenträger benötigt. In jedem RAID 5 Block kann ein Datenträger versagen und komplett wieder hergestellt werden. Die Schreibgeschwindigkeit wird erhöht.

Nutzbarer Speicher:

$$M = \#\text{RAID groups}$$

$$M \cdot \left(\frac{N}{M} - 1\right) \cdot D$$

18 Filesystems und Partitioning

Die Datenträger eines Geräts werden dem Betriebssystem als lineares Array an bytes übergeben. Um die einzelnen Daten des Systems zu speichern, wieder zu finden, und viele weitere Operationen auf ihnen auszuführen, wird ein Dateisystem genutzt.

18.1 Partitiontable

Die Partitionstabelle befindet sich am Anfang des Datenträgers bzw. des Speicherarrays. Sie enthält die Position, den Typ, den Name, usw. von jeder Partition.

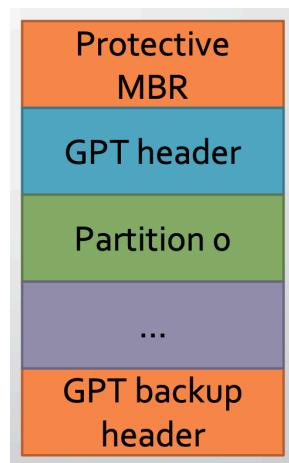
18.1.1 MBR

MBR (Master Boot Record) ist eine simple Partitionstabelle. Sie startet im boot sector (ersten 512bytes) und kann die Partitionsdaten von Datenträgern bis zu 2TB speichern. Bei einer MBR Partitionstabelle ist die Anzahl der Partitionen zudem auf 4 Partitionen beschränkt. MBR wird heutzutage nicht mehr genutzt. Der heutige Standard ist **GPT**.

18.1.2 GPT

GPT (GUID Partition Table) ist eine weitere Art der Partitionstabelle. Sie wird heutzutage hauptsächlich genutzt.

Im Bootsector wird zuerst ein Bereich von 512bytes freigehalten. Dies ist dazu da um Programme daran zu hindern die GPT Tabelle durch schreiben von MBR Daten zu zerstören. Der GPT-Header beinhaltet Informationen zu den einzelnen Partitionen. Mit einer GPT Tabelle sind virtuell unendlich viele Partitionen möglich. Jeder Partition wird durch eine GUID gekennzeichnet. Am Ende des Datenträger findet sich zudem eine Backup Header, der bei Datenkorruption am Anfang des Datenträgers potenziell die Partitionen wieder herstellen kann.



18.2 FAT Filesystem

FAT (File Allocation Table) ist ein simples und dementsprechend naives Dateisystem. Der Datenträger ist in gleich große Cluster unterteilt (meist 16KiB). Eine Datei wird durch eine linked-List von einem oder mehrere dieser Cluster dargestellt. Eine Indexierungstabelle zu Beginn der Partition zeigt auf das Startcluster einer Datei. Folders sind spezielle Dateien die auf die Dateien/cluster in dem Folder zeigen. Das FAT Dateisystem schützt nicht gegen Datenkorruption und hat keinerlei error detection bzw. data recovery.

FAT wird heutzutage noch in manchen Embedded Systems verwendet.

18.3 Journaling Filesystem

Journaling Dateisysteme wie NTFS oder ext4 werden heutzutage häufig verwendet. Sie schützen vor datacorruption. Das schreiben auf das Dateisystem ist jedoch nicht besonders schnell und benötigt mehrere Schritte. Fehler während des Schreibens sind somit gefährlich und können Daten bzw. im schlechtesten Fall Dateisystem Metadaten zerstören.

Wie das Wort Journaling schon vermuten lässt werden Änderungen an Dateien in einem *Journal* gespeichert. Dies passiert als erstes bei einer Schreiboperation. Wenn nun ein Fehler aufgetreten ist können die Daten durch das Journal wieder ersetzt werden.

Fehler im Journal werden durch das inkludieren der checksum in den Journal Einträgen verhindert.

18.4 Multi-Disk Filesystem

Multi-Disk-Filesystems kombinieren Ideen von RAID und dem Journaling Dateisystem. Beispiel für diese Dateisysteme sind ZFS oder btrfs. Das Dateisystem kann hier über mehrere Datenträger verteilt sein. Dies passiert ohne ein RAID Array. Diese Dateisysteme schützen vor bit-rot durch das Speichern der checksum von jeder Datei. Duplikate werden nur einmal auf dem Dateisystem gespeichert und durch Referenzen zugänglich gemacht.

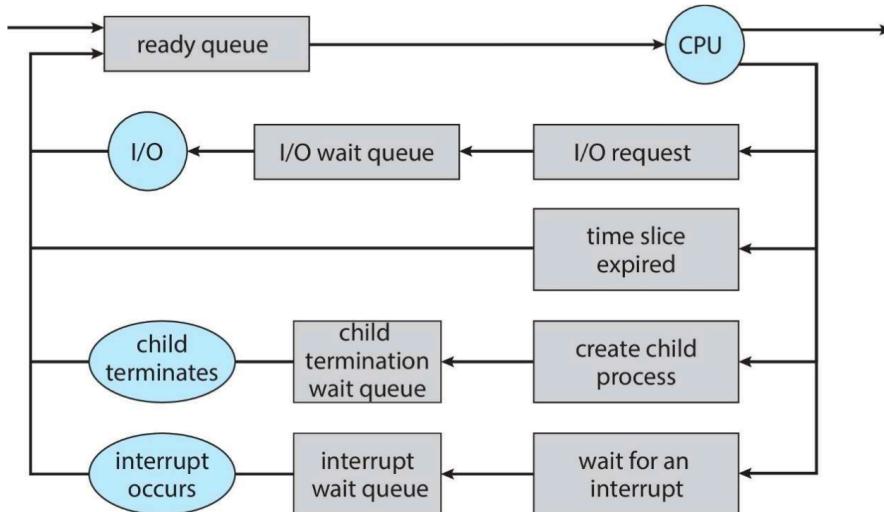
19 CPU Scheduling / Planung

Das Scheduling bzw. die Prozessplangin wird bei der Multi-Core-Programmierung benötigt um eine maximale Prozessorauslastung zu erzielen. Während der Prozessauführung unterscheiden wir zwischen 2 Phasen, dem CPU-Burst und dem I/O Burst.

I/O bound processes sind demnach Programme die eine große Anzahl kurzer CPU-Bursts benötigen. CPU-bound processes haben eine kleine Anzahl an langen CPU-Bursts. Die CPU-Bursts werden über die parallelen Prozesse aufgeteilt.

Ablauf:

Die Prozesse in der Prozessqueue werden von Prozessplaner einem CPU-Kern zugeordnet. Dies ist sinnvoll um die CPU-Auslastung zu maximieren und Prozesse möglichst effizient und schnell einem CPU-Kern zuzuordnen. Dabei muss die Prozessqueue verwaltet werden und die einzelnen CPU-Kerne, sowie I/O und Interrupts auf Bereitschaft überprüft werden. Oft gibt es verschiedenen Warteschlangen zwischen denen gewechselt werden muss.

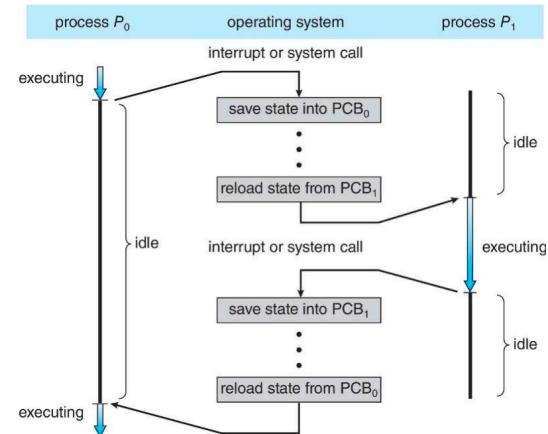


Prozesse die bereit sind auf der CPU ausgeführt zu werden, warten in der Ready-Queue darauf auf der CPU ausgeführt zu werden. Die Queue ist nach verschiedenen Attributen geordnet. Ist "Platz" auf der CPU wird ein neuer Prozess aus der Ready-Queue auf der CPU ausgeführt. Während dem Ausführen des Prozesses kann es dazu kommen dass dieser die CPU wieder verlässt. Dies passiert hier wenn eine I/O Anfrage anfällt, der Timer des Prozesses abgelaufen ist, ein Kindprozess erstellt werden soll oder ein Interrupt die Ausführung unterbricht. Wenn diese Anfragen/usw. verarbeitet wurden wird der Prozess erneut in die Ready-Queue eingefügt. Dies passiert solange bis der Prozess fertig ausgeführt ist und anstatt erneut in die Ready-Queue zu gelangen den Kreislauf verlässt.

Das Dispatchermodule ist zuständig für die Prozessübergabe. Es führt den Context- und Modeswitch (Kernelmode, Usermode) aus und springt an die richtige Stelle im Programm des Prozesses. Als Dispatchlatenz bezeichnen wir die Zeit, die für das Starten und Stoppen bzw. den Contextswitch benötigt wird.

19.1 Contextswitch

Bei dem Kontext handelt es sich um die CPU Register, Zustand und Speicherinformation welche im Prozesskontrollblock gespeichert sind. Bei einem Kontextswitch wird der Kontrollblock eines Prozesses gespeichert und der eines anderen/neuen geladen. Wie bereits oben erklärt finden *ständig* Contextswitches statt, hervorgerufen durch die oben erklärten Anfragen/Interrupts/usw. .



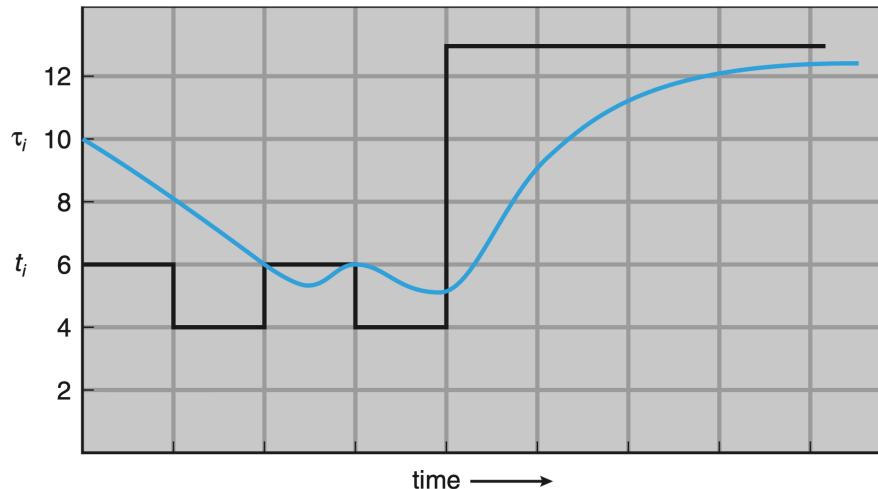
19.2 Scheduling Typen

Wir unterscheiden zwischen Präemptiv und Kooperativ. Wie bereits beim Kernel erklärt kann bei einem Kooperativen Kernel ein Prozess die CPU nur freiwillig verlassen. Bei einem Präemptiven Kernel kann das Betriebssystem entscheiden wann der Prozess die CPU verlässt. Kooperative Betriebssysteme/Kernels sind extrem selten. Präemptive Betriebssysteme/Kernels sind zum Beispiel Linux, Windows und MacOS.

19.3 Scheduling Kriterien

Als Planungskriterien können die Prozessorauslastung, der Durchsatz, die Bearbeitungszeit, die Wartezeit eines Prozesses bzw. die Antwortzeit eines Prozesses herangezogen werden.

19.4 CPU-Burst Prediction



CPU burst (t_i)	6	4	6	4	13	13	13	...	
"guess" (τ_i)	10	8	6	6	5	9	11	12	...

Um die CPU-Burst Zeit vorherzusagen nutzen wir einen exponentielle Glättung.

Die Formel für die Vorhersage ist:

$$\tau_{n+1} = \alpha \cdot t_n + (1 - \alpha) \cdot \tau_n$$

wobei:

- τ_{n+1} : Vorhersage für den nächsten CPU-Burst
- t_n : tatsächliche Dauer des letzten CPU-Burst
- τ_n : Historie der vorigen CPU-Bursts
- α : Glättungsfaktor

Bedeutung des Faktors α :

- $\alpha = 0$: Nur Historie!
- $\alpha = 1$: Nur letzte tatsächliche Zeit!
- $\alpha = 0.5$: Historie und tatsächliche Zeit gleich gewichtet!

Konstante τ_0 : ist der Gesamtsystemdurchschnitt

19.5 Scheduling Algorithmus - First come First server (FCFS)

Beispiele:

P_1 : Burst time : 24

P_2 : Burst time : 3

P_3 : Burst time : 3

- Ankunftsreihenfolge:

P_1, P_2, P_3

- Wartezeiten:

$P_1 : 0; P_2 : 24; P_3 : 27$

$$\emptyset = \frac{0 + 24 + 27}{3} = 17$$

- Ankunftsreihenfolge:

P_2, P_3, P_1

- Wartezeiten:

$P_2 : 0; P_3 : 3; P_1 : 6$

$$\emptyset = \frac{0 + 3 + 6}{3} = 3$$

Hier findet sich ein Konvoi Effekt, spätere Prozesse müssen auf vorherige Prozesse warten. Egal ob diese wesentlich kürzer bzw. länger sind. Dieser Ansatz ist Kooperativ

19.6 Scheduling Algorithmus - Shortest process first

Wenn wir nun alle ankommenden Prozesse aufsteigend nach ihrer geschätzten Burst-Zeit sortieren und in dieser Reihenfolge ausführen, bekommen wir eine minimale durchschnittliche Wartezeit. Auch dieser Ansatz ist Kooperativ.

Beispiel:

P_1 : Burst time : 5

P_2 : Burst time : 8

P_3 : Burst time : 7

P_4 : Burst time : 3

- Reihenfolge:

P_4, P_1, P_3, P_2

- Wartezeiten:

$P_4 : 0, P_1 : 3, P_3 : 9, P_2 : 16$

$$\emptyset = \frac{0+3+9+26}{4} = 7$$

19.7 Scheduling Algorithmus - Process with shortest remaining first

Dies ist die Präemptive version des *Shortest process first* Ansatzes. Die ankommenden Prozesse werden in eine Queue eingesortiert und ausgeführt wenn die Burst-Zeit kürzer als die des geradeigen Prozesses ist.

Beispiel:

P_1 : arrival : 0; Burst time : 8

P_2 : arrival : 1; Burst time : 4

P_3 : arrival : 2; Burst time : 9

P_4 : arrival : 3; Burst time : 5

- Ganttdiagramm (Wartezeiten):

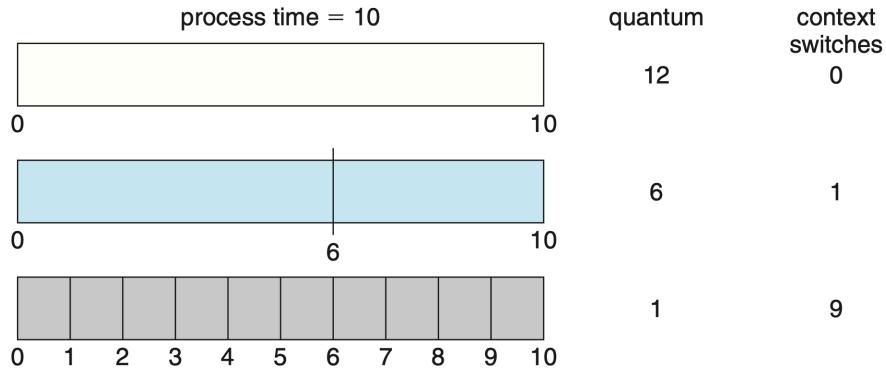


- durchschnittliche Wartezeit:

$$\frac{(10 - 1) + (1 - 1) + (17 - 2) + (5 - 3)}{4} = 6.5$$

19.8 Scheduling Algorithmus - Round-Robin Präemptiv

Der Round-Robin ist der Präemptive Version zu dem *First come First serve* Ansatz. Für diesen Ansatz definieren wir ein **CPU-Zeitquantum** q . Dieses Zeitquantum liegt normalerweise zwischen 10 bis 100 Millisekunden. Im Vergleich zu einem Contextswitch ist dies eine lange Zeit. Wir wählen q so, dass $q > 80\%$ der CPU-Bursts. Die Prozesse werden, wie auch schon bei dem *FCFS* Ansatz, in eine FIFO eingereiht und der Reihe nach abgearbeitet. Dabei darf jeder Prozess eine maximale Zeit von einem Zeitquantum q auf der CPU arbeiten. Der Prozess gibt demnach die CPU frei wenn das Zeitquantum erschöpft bzw. der Prozess abgeschlossen ist.



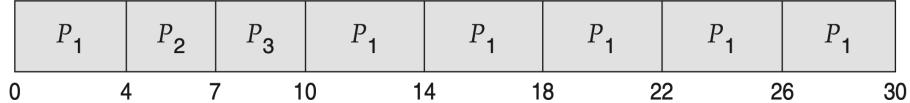
Beispiele:

$$P_1 : \text{Burst time} : 24$$

$$P_2 : \text{Burst time} : 3$$

$$P_3 : \text{Burst time} : 3$$

- $q = 4$:



durchschnittliche Wartezeit:

$$\frac{(10-4) + (4-0) + (7-0)}{3} = 5.66$$

- $n = 3$; Burst time : 10

- $q = 1$

$$\text{Laufzeit: } \frac{28+29+30}{3} = 29$$

$$\text{Wartezeit: } \frac{18+19+20}{3} = 19$$

- $q = 10$

$$\text{Laufzeit: } \frac{10+20+30}{3} = 20$$

$$\text{Wartezeit: } \frac{0+10+20}{3} = 10$$

19.9 Scheduling Algorithmus - Prioritäts Planung

Jeder Prozess hat eine Priorität (z.B. von $[0, 4095]$). Ein kleinerer Wert bedeutet eine höherer Priorität. Der Prozess mit der höchsten Priorität wird der CPU zugewiesen. Das Prinzip funktioniert sowohl Präemptiv als auch Kooperativ. Da die Queue nach der Priorität sortiert wird kann es dazu kommen, dass manche Prozesse mit einer geringen Priorität nie ausgeführt werden. Dies kann behoben werden in dem die Priorität mit dem "Alter" des Prozesses zunimmt. Das Suchen des Prozesses mit der höchsten Priorität hat eine Laufzeit von $\mathcal{O}(n)$.

Beispiele:

- Kooperativ:

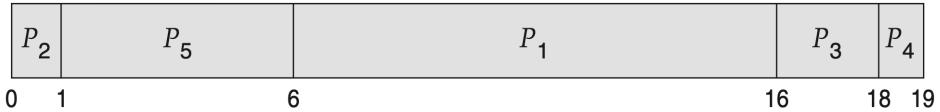
P_1 : Burst time : 10; Priority : 3

P_2 : Burst time : 1; Priority : 1

P_3 : Burst time : 2; Priority : 4

P_4 : Burst time : 1; Priority : 5

P_5 : Burst time : 5; Priority : 2



durchschnittliche Wartezeit:

$$\frac{0 + 1 + 6 + 16 + 18}{5} = 18$$

- Präemptiv:

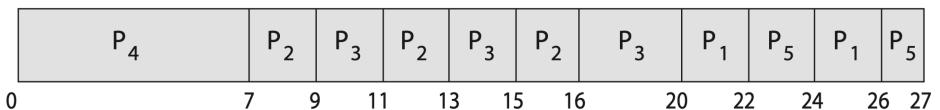
P_1 : Burst time : 4; Priority : 3

P_2 : Burst time : 5; Priority : 2

P_3 : Burst time : 8; Priority : 2

P_4 : Burst time : 7; Priority : 1

P_5 : Burst time : 3; Priority : 3



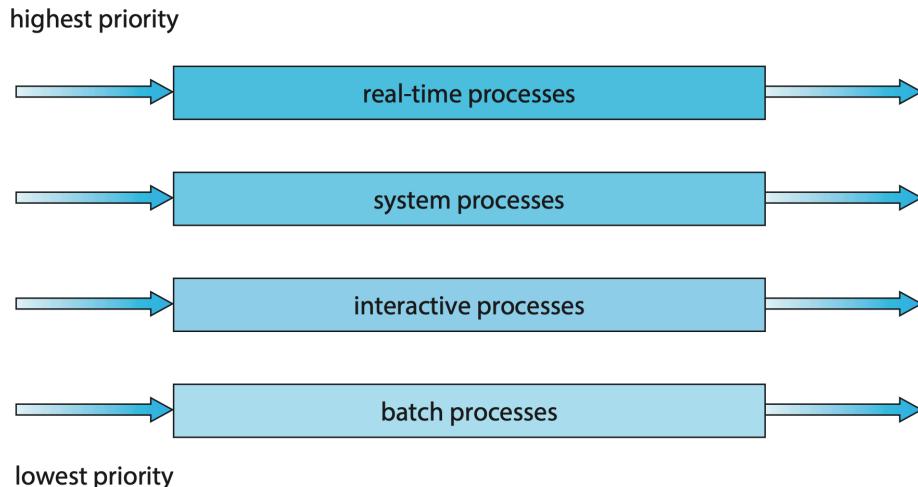
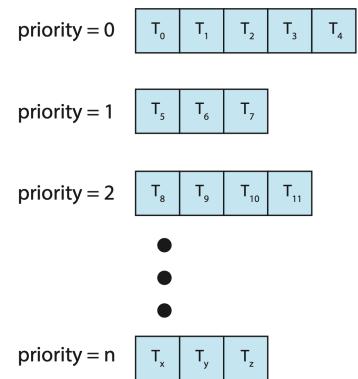
durchschnittliche Wartezeit:

$$\frac{(20 + 24 - 22) + (7 + 11 - 9 + 15 - 13) + (9 + 13 - 11 + 16 - 15) + (22 + 26 - 24)}{5} = 13.8$$

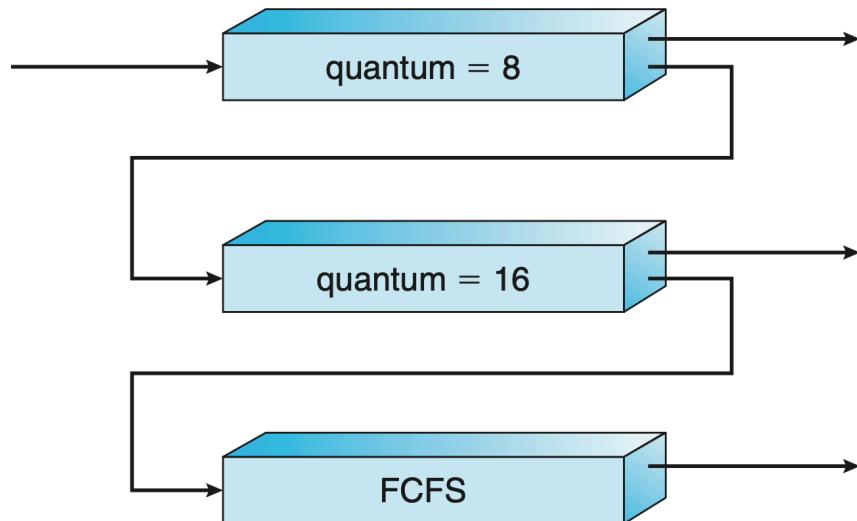
19.10 Multilevel Queue Scheduling

Im Vergleich zum Priority Queue Ansatz wird hier jeder Priorität eine Queue zugeordnet. Das heißt alle Prozesse mit Priorität k kommen in die zugehörige Queue q_k , die alle Prozesse mit Priorität k beinhaltet.

Wir unterscheiden zwischen Vorder- und Hintergrundprozessen. Die Vordergrundprozesse haben eine hohe Priorität und werden mit dem Round-Robin Ansatz abgearbeitet. Die Hintergrund Prozesse haben eine geringere Priorität und werden in FCFS abgearbeitet.



19.11 Multilevel Feedback Queue Scheduling



Ähnelt dem Multilevel Queue Scheduling stark. Hier ist es jedoch möglich, dass Prozesse von einer Prioritäts Queue in eine andere verschoben werden. Meist werden die Prozesse nach CPU-Burst Zeit und nach Zeit in der Queue in die jeweilige höhere bzw. niedrigere Queue verschoben.

Prozesse werden herabgestuft, das Zeitquantum erschöpft wurde, und heraufgestuft wenn der Prozess bereits lange in der Queue auf Ausführung wartet.

20 Thread Scheduling / Planung

Contention Conflict:

Wir unterscheiden zwischen 2 Konflikttarten:

Beim Prozesszugangskonflikt (PTHREAD_SCOPE_PROCESS) konkurrieren Benutzer-Threads innerhalb eines Prozesses um die Zuteilung auf Kernel-Threads. Dabei kommen Many-to-One- oder Many-to-Many-Modelle zum Einsatz, und die Planung erfolgt anhand einer vom Programmierer festgelegten Priorität.

Beim Systemzugangskonflikt (PTHREAD_SCOPE_SYSTEM) wird jedem Benutzer-Thread ein eigener Kernel-Thread zugeordnet (One-to-One-Modell). Die Threads konkurrieren dabei systemweit um Rechenzeit. Dieses Modell ist die einzige Option unter Linux und macOS.

Beispiel:

```
#include <pthread.h>
#include <stdio.h>
#define NUM_THREADS 5

// Each thread will begin control in this function
void *runner(void *param) {
    // do some work ...
    pthread_exit(0);
}

int main(int argc, char *argv[]) {
    int i, scope;
    pthread_t tid[NUM_THREADS];
    pthread_attr_t attr;

    // get the default attributes
    pthread_attr_init(&attr);

    // first inquire on the current scope
    if (pthread_attr_getscope(&attr, &scope) != 0) {
        fprintf(stderr, "Unable to get scheduling scope\n");
    } else {
        if (scope == PTHREAD_SCOPE_PROCESS)
            printf("PTHREAD_SCOPE_PROCESS");
        else if (scope == PTHREAD_SCOPE_SYSTEM)
            printf("PTHREAD_SCOPE_SYSTEM");
        else
            fprintf(stderr, "Illegal scope value.\n");
    }

    // set the scheduling algorithm to PCS or SCS
    pthread_attr_setscope(&attr, PTHREAD_SCOPE_SYSTEM);

    // create the threads
    for (i = 0; i < NUM_THREADS; i++)
        pthread_create(&tid[i], &attr, runner, NULL);

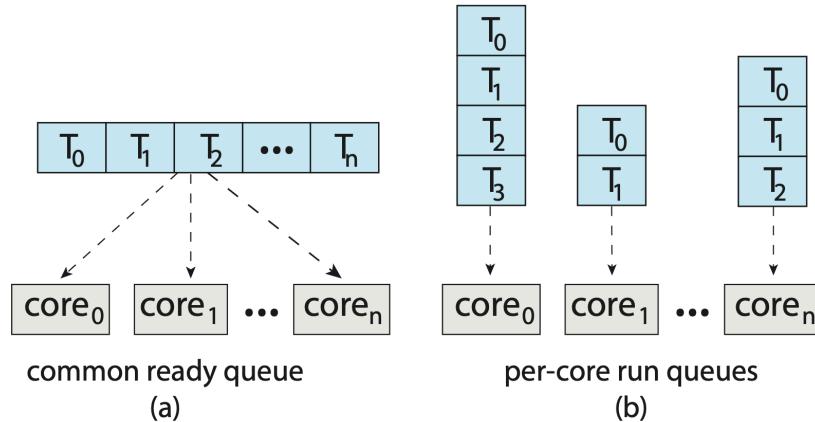
    // now join on each thread
    for (i = 0; i < NUM_THREADS; i++)
        pthread_join(tid[i], NULL);
}
```

21 Multiprocessor Scheduling / Planung

Wir unterscheiden zwischen asymmetrischer und symmetrischer Multiprozessorplanung. Bei der Asymmetrischen Planung ist ein CPU Kern für die Planung zuständig. Bei der Synchronen Planung ist jeder CPU Kern für seine eigene Planung zuständig.

Zudem unterscheiden wir zwischen einer gemeinsamen und einer privaten Threadqueue. Auf die gemeinsame Queue wird von allen Threads zugegriffen. Bei der privaten Queue hat jeder Prozessor eine eigene Warteschlange.

Beide Ansätze haben Vor- und Nachteile. Bei der gemeinsamen Queue kann es zu race conditions kommen, bzw. das Synchronisieren zum Bottleneck werden. Die private Queue hat eine bessere Performance, auf Kosten eines deutlich erschwerteren Lastausgleichs.

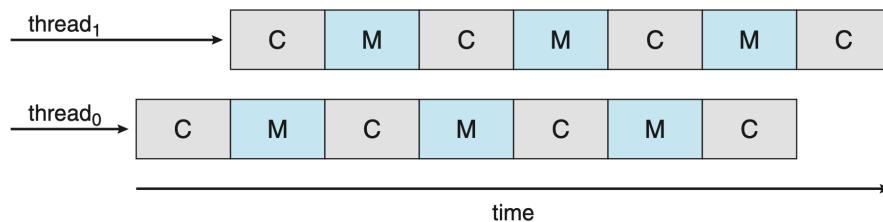


21.1 Multicore Prozessoren

Mehrere Hardware Threads pro CPU Kern. Da bei einem Speicheraufruf die CPU eine lange Wartezeit hat, wird versucht während dieses Speicheraufrufs einen anderen Thread auf der CPU laufen zu lassen.

M : Memory stall cycle

C : compute cycle



Neue Prozessoren ermöglichen außerdem ein sogenanntes Chip-Multithreading bzw. Hyper-threading (Intel). Dabei werden jedem CPU Kern 2 Hardware Threads zugeordnet, was für einen Quad-Core Prozessor ein System mit 8 logischen Prozessoren ergibt. Es ist jedoch anzumerken, dass pro CPU Kern maximal 1 Hardware Thread gleichzeitig ausgeführt werden kann, da der Cache und die Pipelines zwischen den beiden Hardwarethreads geteilt werden.

Wir unterscheiden zudem zwischen Coarsened- und Finegrained Multithreading. Bei Coarsened-grained Multithreading wird ein Thread solange ausgeführt bis eine lange Unterbrechung, wie

eine Memory stall vorliegt. Dann findet ein großer Contextswitch statt, bei dem die gesamte Instruktions Pipeline geflushed werden muss.

Auf Finegrained Systemen findet ein kleiner Contextswitch statt. Dieser ist Hardware unterstützt.

Die wirkliche Planung auf Multi-Core-Systemen wird in 2 Ebenen aufgeteilt. Die Kernelebene und die Betriebssystemebene. Auf der Betriebssystemebene werden die Softwarethreads auf die Hardware-Threads (bzw. logischen Kerne) aufgeteilt. Auf der Kernelebene werden die Hardware-Threads mittels Round-Robin Algorithmus und einer Priorität auf den physischen CPU Kern gescheduled.

21.2 Load Balancing

Es ist sinnvoll die Arbeit des Systems gleichmäßig auf alle CPU Kerne aufzuteilen. Hier unterscheiden wir zwischen Push- und Pullmigration. Bei der Pushmigration wird die Auslastung der Kerne in regelmäßigen Abständen geprüft und Aufgaben von überlasteten Kernen auf andere Kerne mit weniger Auslastung verschoben. Bei der Pullmigration werden wartenden Aufgaben von überlasteten Kernen, von Kernen mit wenig oder keiner Auslastung, abgezogen.

21.3 Processor Affinity

Threadaffinität **zu** einem Prozessor bedeutet, dass ein Thread bevorzugt auf demselben Prozessor ausgeführt wird, um den Cachezugriff zu optimieren. Bei einem Lastausgleich kann diese Affinität verloren gehen.

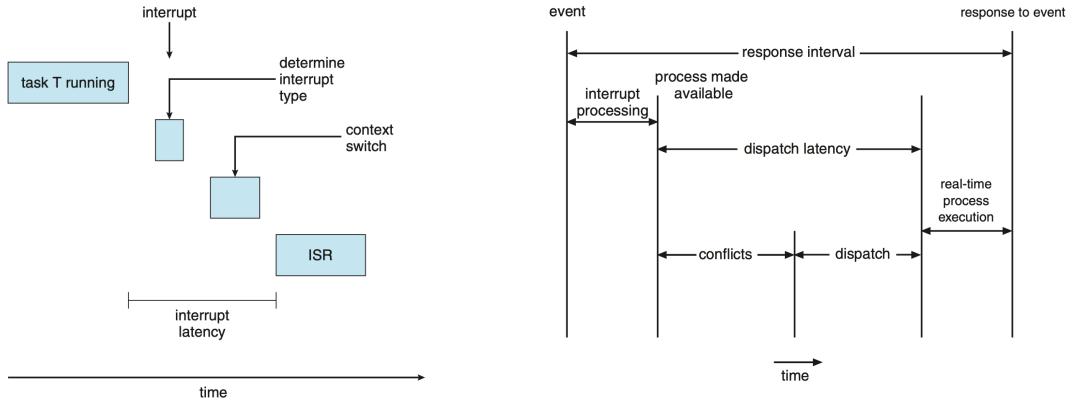
Soft Affinity, versucht Threads auf demselben Prozessor laufen zu lassen, jedoch ohne Garantie. Hard Affinity, erlaubt das angeben einer festen Menge an Prozessoren, auf denen ein Thread ausgeführt werden darf.

22 Real-Time CPU Scheduling / Planung

Echtzeitsysteme sind Computersysteme, die Aufgaben innerhalb einer vorgegebenen Zeit erledigen müssen. Klassische UNIX-Systeme arbeiten mit Timesharing und reagieren auf Unterbrechungen, garantieren aber keine festen Zeiten. Ereignis-basierte Echtzeitsysteme verwenden Prioritäten und können laufende Aufgaben zugunsten wichtigerer unterbrechen. Weiche Echtzeitsysteme versuchen, Termine für kritische Aufgaben einzuhalten, geben aber keine absolute Garantie. Harte Echtzeitsysteme müssen Aufgaben strikt fristgerecht erledigen, ein Verpassen der Deadline ist nicht akzeptabel.

Interrupt-Latency: Unterbrechungsverzögerung ist die Zeitspanne zwischen dem Eintreffen einer Unterbrechung und dem Start der zugehörigen Unterbrechungsroutine. Das bedeutet, vom Moment, in dem eine Unterbrechung erkannt wird, bis zu dem Punkt, an dem die entsprechende Bearbeitung beginnt, vergeht eine gewisse Zeit.

Dispatch-Latency: Dispatch-Verzögerung bezeichnet die Zeit, die benötigt wird, um einen laufenden Prozess zu stoppen und einen neuen Prozess zu laden. In dieser sogenannten Konfliktphase werden alle Prozesse im Kernel-Modus unterbrochen und Ressourcen werden für Prozesse mit höherer Priorität freigegeben. Erst danach kann das System mit der Bearbeitung des neuen Prozesses beginnen.

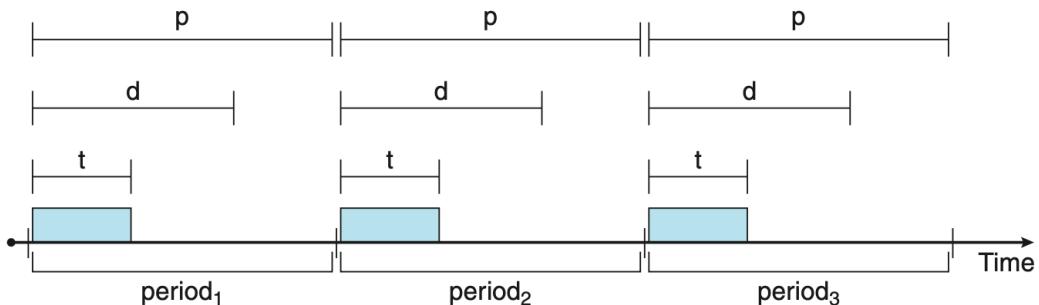


22.1 Priority-based Scheduling

Bei der prioritätsbasierten Echtzeitplanung werden Aufgaben (Prozesse) nach ihrer Priorität eingeplant und können von höher priorisierten Aufgaben unterbrochen werden (präemptiv). Diese Methode eignet sich oft für weiche Echtzeitsysteme.

Die Prozesse wiederholen sich regelmäßig (periodisch) und haben jeweils eine Periode p , eine Frist d , und eine Bearbeitungszeit t . Die Bearbeitung eines Prozesses muss im Zeitraum von 0 bis zur Frist d , die wiederum kleiner oder gleich der Periode p ist, abgeschlossen sein. Die periodische Aufgaberate gibt an, wie oft eine Aufgabe pro Zeiteinheit kommt, also $\frac{1}{p}$.

Die Auslastung der CPU durch einen Prozess P_i wird mit $U_i = \frac{t_i}{p_i}$ berechnet, also als Verhältnis von Bearbeitungszeit zu Periode.

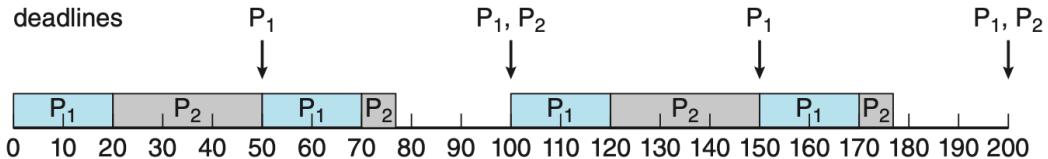


22.2 Rate-Monotonic Scheduling

Bei der ratenmonotonischen Planung wird die Priorität eines Prozesses durch den Kehrwert seiner Periode bestimmt: Prozesse mit kürzeren Perioden (die häufiger laufen müssen) erhalten eine höhere Priorität, während Prozesse mit längeren Perioden eine niedrigere Priorität bekommen.

Beispiel:

- Prozess P_1 hat eine Periode von 50 (läuft alle 50 Zeiteinheiten) und eine Bearbeitungszeit von 20. Seine Auslastung ist $U_1 = 0.4$.
- Prozess P_2 hat eine Periode von 100 und braucht 35 Zeiteinheiten zur Bearbeitung. Seine Auslastung ist $U_2 = 0.35$. Die Gesamtauslastung $U = U_1 + U_2 = 0,75$ ist kleiner als 1, das heißt, beide Prozesse können rechtzeitig fertiggestellt werden.

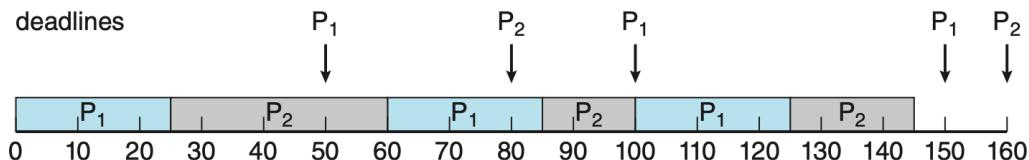


22.3 Earliest-Deadline-First Scheduling

Bei einer Planung nach der frühestmöglichen Deadline werden die Prioritäten der Prozesse dynamisch und immer wieder neu nach ihren aktuellen Deadlines vergeben: Der Prozess mit der nächsten Deadline erhält die höchste Priorität.

Im gezeigten Beispiel hat Prozess P_1 eine Periode von 50 und eine Bearbeitungszeit von 25, während Prozess P_2 eine Periode von 80 und eine Bearbeitungszeit von 35 hat. Die Gesamttauslastung der CPU berechnet sich zu $U = 0.9375$. Dieser Wert liegt zwischen der Schranke für zwei Prozesse nach der ratenmonotonischen Planung 0.83 und 1. Das bedeutet, das System ist für dieses Beispiel planbar.

Im schlimmsten Fall gilt für die maximale CPU-Ausnutzung eine obere Grenze $U \leq N \cdot (\sqrt{2} - 1)$, wobei N die Anzahl der Prozesse ist. Bei einem Prozess sind 100% Auslastung möglich, bei zwei etwa 83%, und bei sehr vielen Prozessen nähert sich die Grenze etwa 69%.



22.4 Proportional Share Scheduling

Bei der proportionalen Anteilplanung wird die gesamte verfügbare Prozessorzeit in Anteile aufgeteilt und den verschiedenen Prozessen zugewiesen. Jeder Prozess erhält entsprechend seinem Bedarf einen Anteil der gesamten Prozessorzeit.

Im Beispiel stehen insgesamt 100 Anteile zur Verfügung. Die Prozesse A, B und C erhalten 50, 15 und 20 Anteile, was zusammen 85 Anteile ergibt. Es bleiben noch 15 Anteile übrig. Da Prozess D aber 30 Anteile benötigt, kann er nicht zugelassen werden, weil die verfügbare Gesamtzeit überschritten würde.

Die Tabelle zeigt die jeweiligen Anteile und die CPU-Nutzung pro Prozess – Prozess D wird abgelehnt, weil nicht genug Ressourcen übrig sind.

22.5 POSIX Real-Time Scheduling

Der POSIX.1b-Standard definiert zwei Klassen für Echtzeit-Threads. Bei der Planungsrichtlinie `SCHED_FIFO` werden Threads nach dem Prinzip *FCFS* und mit einer FIFO-Warteschlange geplant, wobei Threads mit gleicher Priorität keine feste Zeitaufteilung erhalten. Bei `SCHED_RR` (Round Robin) ist die Funktionsweise ähnlich, allerdings gibt es eine Zeitaufteilung für Threads mit gleicher Priorität – sie bekommen also eine festgelegte Zeitscheibe zugewiesen. Programmatisch lassen sich die Planungsrichtlinien mit den POSIX-

Funktionen pthread_attr_getsched_policy und pthread_attr_setsched_policy abrufen und setzen.

```
#include <pthread.h>
#include <stdio.h>
#define NUM_THREADS 5

// Each thread will begin control in this function
void *runner(void *param) {
    // do some work ...
    pthread_exit(0);
}

int main(int argc, char *argv[]) {
    int i, policy;
    pthread_t tid[NUM_THREADS];
    pthread_attr_t attr;

    // get the default attributes
    pthread_attr_init(&attr);

    // get the current scheduling policy
    if (pthread_attr_getschedpolicy(&attr, &policy) != 0) {
        fprintf(stderr, "Unable to get policy.\n");
    } else {
        if (policy == SCHED_OTHER)
            printf("SCHED_OTHER\n");
        else if (policy == SCHED_RR)
            printf("SCHED_RR\n");
        else if (policy == SCHED_FIFO)
            printf("SCHED_FIFO\n");
    }

    // set the scheduling policy - FIFO, RR, or OTHER
    if (pthread_attr_setschedpolicy(&attr, SCHED_FIFO) != 0)
        fprintf(stderr, "Unable to set policy.\n");

    // create the threads
    for (i = 0; i < NUM_THREADS; i++)
        pthread_create(&tid[i], &attr, runner, NULL);

    // now join on each thread
    for (i = 0; i < NUM_THREADS; i++)
        pthread_join(tid[i], NULL);
}
```

22.6 Linux Scheduling

Im Linux-System gibt es zwei Planertypen: den völlig fairen Planer (CFS) und den Echtzeit-Planer. Beim völlig fairen Planer wird die Prozesspriorität durch den Nice-Wert bestimmt, der zwischen -20 (höchste Priorität) und 19 (niedrigste Priorität) liegt. Standardmäßig ist der Wert 0 . Die Verteilung der CPU-Anteile hängt davon ab. Statt eines festen Zeitquants sorgt eine gezielte Verzögerung dafür, dass jeder Prozess mindestens einmal ausgeführt wird. Außerdem wird die Ausführungsreihenfolge nach der virtuellen Prozesslaufzeit geregelt: Die Laufzeit wird je nach Priorität angepasst, und der Prozess mit der geringsten virtuellen Laufzeit wird zuerst ausgeführt.

Beim Echtzeit-Planer gibt es Echtzeitprozesse mit sehr hoher Priorität (Werte zwischen 0 und 99 , mit SCHED_FIFO und SCHED_RR), während normale Prozesse niedrigere Priorität haben (Werte zwischen 100 und 139 , was auf die Nice-Werte $[-20, 19]$ abgebildet wird)

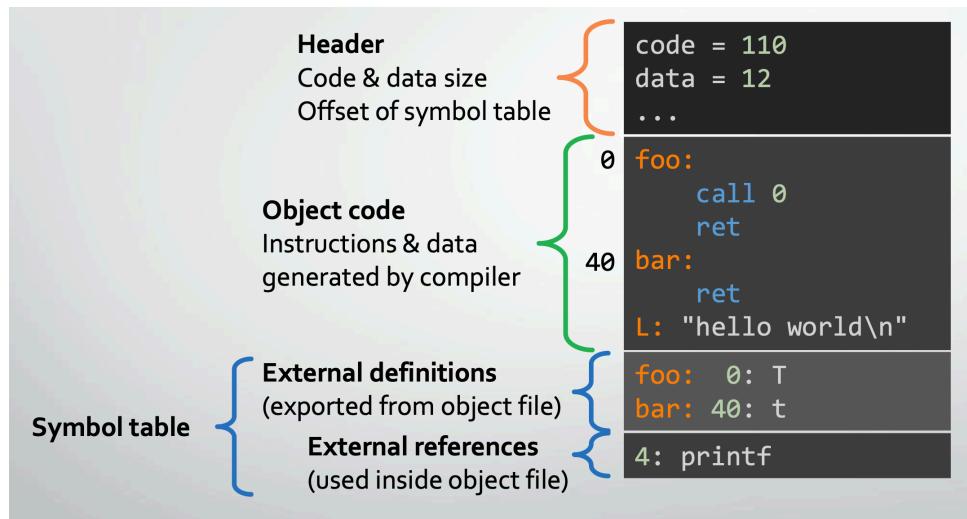
23 Compiling, Linking, Loading and Libraries

23.1 Compiling

Ein Linker erstellt mit einem oder mehreren Objectfiles eine Ausführbare Programmdatei. Dabei können auch sogenannte Libraries eingebunden werden. Diese können statisch(static) und dynamisch(dynamic/shared) sein.

Bei einer Statischen Library wird der Code direkt in den Programmcode eingebunden wohingegen dynamische Libraries erst beim Ausführen des Programms geladen werden.

Eine Kompiliertes Programm besteht aus einen Header, dem Objectcode und aus dem Symbol table der aus externen definitionen und externen Referenzen besteht.



23.2 Ausführung eines Programms

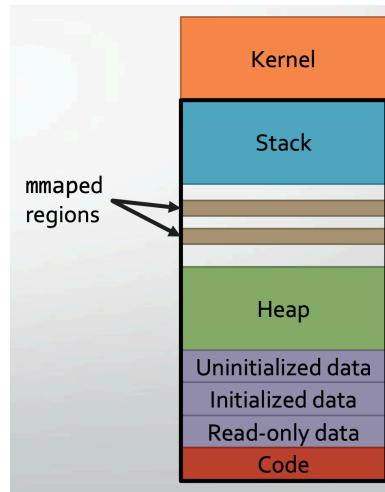
Auf UNIX-Systemen wird das Programm mit einem loader gelesen. Dieser liest alle Daten und die Code-Segmente in einen Cache Buffer. Der Code wird als read-only Bereich und die initialisierten Daten als read/write-able in den Addressraum geladen.

Während dem Laden finden viele Optimierungen statt. Daten die mit 0 initialisiert werden müssen zum Beispiel nicht gelesen werden und Teile des Programms werden erst bei der Benutzung geladen. Außerdem kann der Code des Programms geteilt werden, wenn mehrere Instanzen des gleichen Programms bzw. Teile und Funktionen gleichzeitig ausgeführt werden.

Der Prozessspeicher unter UNIX sieht wie folgt aus:

Der Heap wird am Anfang der Laufzeit des Programms mit malloc alloziert. Dadurch sind Compiler, sowie Linker nicht involviert. Die Variablen des Programms werden vom Programm selbst verwaltet und befinden sich in dem Heap.

Der Stack wird auch am Anfang der Laufzeit des Programms alloziert. Das Layout des Heaps wird durch den Compiler festgelegt. Variablen auf dem Stack werden als pointer relativ zu dem Stackpointer dargestellt. Der Linker ist nicht involviert und der Stack wird wie bereits angesprochen vom Compiler organisiert und gemanaged.



Globale Daten und Code werden durch den Compiler alloziert, wobei das Layout durch den Linker erstellt wird. Der Compiler erstellt symbolische Referenzen und der Linker übersetzt diese Referenzen.

Beispiel: Ein Programm das nur aus dem Aufruf `printf("hello world\n");` besteht, und mit dem Befehl `cc -m32 -fno-builtin -S hello.c`. Die flag `-S` bewirkt, dass da Programm ohne den Assembler kompiliert wird.

Der erstellte Code bzw. `hello.s` sieht dann wie folgt aus:

```
.LC0:  
    .string "hello world\n"  
main:  
...  
    subl $12, %esp  
    pushl $.LC0  
    call printf  
...  
    ret
```

Wie zu sehen ist enthält der code eine symbolische Referenz auf `printf`.

23.3 Linker

Der Linker – unter UNIX aufrufbar mit dem Befehl `ld` – wird normalerweise automatisch vom Compiler gestartet. Dies kann beispielsweise im Befehl `gcc -v hello.c` nachvollzogen werden. GCC ruft dabei `collect2` auf, einen internen Wrapper für `ld`.

Generell führt der Linker 3 Operationen aus:

1. Sammle alle Teile des Programms zusammen (objectfiles)
2. unify/vereinige alle gleichen Segmente
3. Passe die Adressen des Codes und den Daten je nach Programm an

Das Ergebnis ist eine ausführbare Datei bzw. ein ausführbares Programm.

C- und C++-Compiler können immer nur eine einzelne Translation Unit (bzw. eine Quelldatei) „sehen“. Deshalb ist es nicht möglich, direkt aus einem einzelnen Source-File ein komplettes, ausführbares Programm zu generieren. Dafür ist der Linker notwendig. Neben dem bloßen Zusammenfügen der einzelnen kompilierten (nun „object files“ genannten) Source Files führt der Linker zusätzliche Optimierungen am Code aus. Dabei wird die Reihenfolge der Codesequenzen meist nicht verändert, jedoch kann die Reihenfolge der Funktionsaufrufe optimiert werden. Dies führt zu einer besseren Cache-Ausnutzung und dazu, dass unbenötigte Segmente entfernt werden.

Einfacher Linker:

1. Durchlauf:

- Gleiche Code-Segmente werden zusammengefasst und so angeordnet, dass sie sich nicht überschneiden.
- Die Symbol-Tabellen der einzelnen Object-Files werden eingelesen und zu einer globalen Symbol-Tabelle ohne Duplikate zusammengefasst.
- Die virtuellen Adressen aller Segmente (als Offset zur Basisadresse) werden berechnet.

2. Durchlauf:

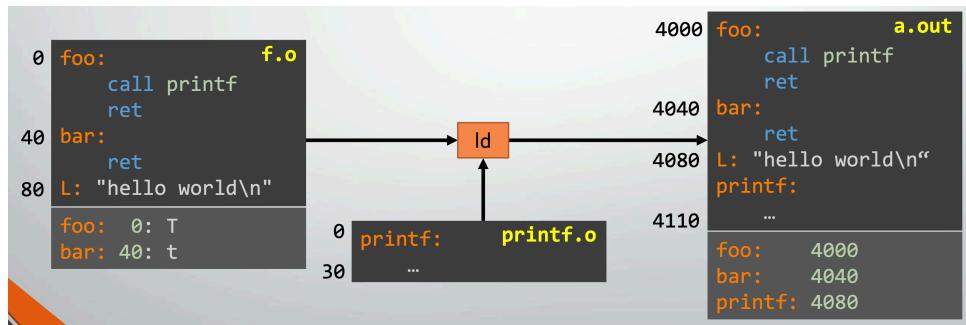
- Alle Referenzen auf Symbole werden anhand der globalen Symbol-Tabelle ersetzt.
- Der endgültige Code wird zurückgegeben.

Die Symboletabellen speichern Informationen über das Programm – z. B. Namen, Größe, alte und neue Position der Segmente. Außerdem werden die einzelnen Symbole mit Namen, Typ und Offset gespeichert.

Die vom Assembler erstellten Object-Files enthalten Code mit einem Offset, wobei die Start- bzw. Basisadresse (0) ans Ende der .s-Datei geschrieben wird.



Der Linker liest diese .s-Dateien erneut ein, berechnet die tatsächlichen Segmentgrößen und speichert alle verwendeten Symbole nebst ihrem Code in einer abschließenden, globalen Symboletabellen mit endgültigen virtuellen Adressen.



Beispiel:

- Source Files:

main.c

```
extern float sin();
extern int printf();
extern int scanf();
float val = 0.0f;

int main() {
    static float x = 0.0f;
    printf("enter number: ");
    scanf("%f", &x);
    val = sin(x);
    printf("sine is %f\n", val);
}
```

libc

```
int scanf(char *fmt, ...) { /* ... */ }
int printf(char *fmt, ...) { /* ... */ }
```

math.c

```
float sin(float x) {
    float tmp1, tmp2;
    static float res = 0.0f;
    static float lastx = 0.0f;
    if(x != lastx) {
        lastx = x;
        // compute sin(x)
    }
    return res;
}
```

- Object Files:

main.o	math.o																																	
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center; background-color: #e0e0e0;">symbols</th></tr> </thead> <tbody> <tr> <td>def: val @ 0:D</td></tr> <tr> <td>def: main @ 0:T</td></tr> <tr> <td>def: x @ 4:d</td></tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center; background-color: #e0e0e0;">relocation</th></tr> </thead> <tbody> <tr> <td>ref: printf @ 0:T, 12:T</td></tr> <tr> <td>ref: scanf @ 4:T</td></tr> <tr> <td>ref: x @ 4:T, 8:T</td></tr> <tr> <td>ref: sin @ ?:T</td></tr> <tr> <td>ref: val @ ?:T, ?:T</td></tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center; background-color: #e0e0e0;">data</th></tr> </thead> <tbody> <tr> <td>0 val:</td></tr> <tr> <td>4 x:</td></tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center; background-color: #e0e0e0;">text</th></tr> </thead> <tbody> <tr> <td>0 call printf</td></tr> <tr> <td>4 call scanf(&x)</td></tr> <tr> <td>8 val = call sin(x)</td></tr> <tr> <td>12 call printf(val)</td></tr> </tbody> </table>	symbols	def: val @ 0:D	def: main @ 0:T	def: x @ 4:d	relocation	ref: printf @ 0:T, 12:T	ref: scanf @ 4:T	ref: x @ 4:T, 8:T	ref: sin @ ?:T	ref: val @ ?:T, ?:T	data	0 val:	4 x:	text	0 call printf	4 call scanf(&x)	8 val = call sin(x)	12 call printf(val)	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center; background-color: #e0e0e0;">symbols</th></tr> </thead> <tbody> <tr> <td>def: sin @ 0:T</td></tr> <tr> <td>def: res @ 0:d</td></tr> <tr> <td>def: lastx @ 4:d</td></tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center; background-color: #e0e0e0;">relocation</th></tr> </thead> <tbody> <tr> <td>ref: lastx @ 0:T, 4:T</td></tr> <tr> <td>ref: res @ 24:T</td></tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center; background-color: #e0e0e0;">data</th></tr> </thead> <tbody> <tr> <td>0 res:</td></tr> <tr> <td>4 lastx:</td></tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center; background-color: #e0e0e0;">text</th></tr> </thead> <tbody> <tr> <td>0 if(x != lastx)</td></tr> <tr> <td>4 lastx = x;</td></tr> <tr> <td>... ... compute sin(x) ...</td></tr> <tr> <td>24 return res;</td></tr> </tbody> </table>	symbols	def: sin @ 0:T	def: res @ 0:d	def: lastx @ 4:d	relocation	ref: lastx @ 0:T, 4:T	ref: res @ 24:T	data	0 res:	4 lastx:	text	0 if(x != lastx)	4 lastx = x; compute sin(x) ...	24 return res;
symbols																																		
def: val @ 0:D																																		
def: main @ 0:T																																		
def: x @ 4:d																																		
relocation																																		
ref: printf @ 0:T, 12:T																																		
ref: scanf @ 4:T																																		
ref: x @ 4:T, 8:T																																		
ref: sin @ ?:T																																		
ref: val @ ?:T, ?:T																																		
data																																		
0 val:																																		
4 x:																																		
text																																		
0 call printf																																		
4 call scanf(&x)																																		
8 val = call sin(x)																																		
12 call printf(val)																																		
symbols																																		
def: sin @ 0:T																																		
def: res @ 0:d																																		
def: lastx @ 4:d																																		
relocation																																		
ref: lastx @ 0:T, 4:T																																		
ref: res @ 24:T																																		
data																																		
0 res:																																		
4 lastx:																																		
text																																		
0 if(x != lastx)																																		
4 lastx = x;																																		
... ... compute sin(x) ...																																		
24 return res;																																		

- Linker Pass 1: Reorganization:

symbols	symbols																													
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center; background-color: #e0e0e0;">data</th></tr> </thead> <tbody> <tr> <td>0 val:</td></tr> <tr> <td>4 x:</td></tr> <tr> <td>8 res:</td></tr> <tr> <td>12 lastx:</td></tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center; background-color: #e0e0e0;">text</th></tr> </thead> <tbody> <tr> <td>16 main:</td></tr> <tr> <td>... ...</td></tr> <tr> <td>26 call printf</td></tr> <tr> <td>30 sin:</td></tr> <tr> <td>... ...</td></tr> <tr> <td>50 return res;</td></tr> <tr> <td>... ...</td></tr> <tr> <td>64 printf:</td></tr> <tr> <td>... ...</td></tr> <tr> <td>80 scanf:</td></tr> <tr> <td>... ...</td></tr> </tbody> </table>	data	0 val:	4 x:	8 res:	12 lastx:	text	16 main:	26 call printf	30 sin:	50 return res;	64 printf:	80 scanf:	<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center; background-color: #e0e0e0;">data</th></tr> </thead> <tbody> <tr> <td>starts @ 0</td></tr> </tbody> </table> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center; background-color: #e0e0e0;">text</th></tr> </thead> <tbody> <tr> <td>starts @ 16</td></tr> <tr> <td>def: val @ 0</td></tr> <tr> <td>def: x @ 4</td></tr> <tr> <td>def: res @ 8</td></tr> <tr> <td>def: main @ 16</td></tr> <tr> <td>... ...</td></tr> <tr> <td>ref: printf @ 26</td></tr> <tr> <td>ref: res @ 50</td></tr> <tr> <td>... ...</td></tr> </tbody> </table>	data	starts @ 0	text	starts @ 16	def: val @ 0	def: x @ 4	def: res @ 8	def: main @ 16	ref: printf @ 26	ref: res @ 50
data																														
0 val:																														
4 x:																														
8 res:																														
12 lastx:																														
text																														
16 main:																														
... ...																														
26 call printf																														
30 sin:																														
... ...																														
50 return res;																														
... ...																														
64 printf:																														
... ...																														
80 scanf:																														
... ...																														
data																														
starts @ 0																														
text																														
starts @ 16																														
def: val @ 0																														
def: x @ 4																														
def: res @ 8																														
def: main @ 16																														
... ...																														
ref: printf @ 26																														
ref: res @ 50																														
... ...																														

- Linker Pass 2: Relocation:

Virtual base address 4000

symbols	
	data
0 val:	4000
4 x:	4004
8 res:	4008
12 lastx:	4012
	text
16 main:	4016
...	...
26 call ?? // printf(val)	4026
30 sin:	4030
...	...
50 return load ?? // res	4050
...	...
64 printf:	4064
...	...
80 scanf:	4080
...	

symbols	
	data
data starts @ 4000	
text starts @ 4016	
def: val @ 4000	
def: x @ 4004	
def: res @ 4008	
def: main @ 4016	
...	

Die Referenzen (rechts) werden meistens nicht mit gespeichert. Diese können jedoch für das Debuggen genutzt werden.

- Final Output:

Die Referenzen (hier nicht zu sehen) werden mit der globalen Symboltabelle überarbeitet.

symbols	
	data
val:	4000
x:	4004
res:	4008
lastx:	4012
	text
main:	4016
...	...
call 4064(4000)	4026
sin:	4030
...	...
return load 4008	4050
...	...
printf:	4064
...	...
scanf:	4080
...	

23.4 Metadaten in Ausführbaren Dateien

23.4.1 Tool nm

```
int uninitialized;
int initialized = 1;
const int constant = 2;

int main() {
    return 0;
}
```

const Variablen haben den Typen R, was bedeutet das diese im read-only Speicher liegen. Uninitialisierte Daten haben Typ B und liegen in dem BBS Segment.

```
user@PC $ nm -n a.out
...
08049000 T _init
08049040 T _start
08049160 T main
0804a008 R constant
0804c010 W data_start
0804c018 D initialized
0804c020 B uninitialized
```

Dieser ist nicht Teil des Programms und kann in den Speicherpages die mit 0 initialisiert werden gefunden werden.

23.4.2 Tool objdump

```

user@PC $ objdump -h a.out

a.out:      file format elf32-i386

Sections:
Idx Name      Size    VMA     LMA     File off  Align
...           ...
11 .init     00000020 08049000 08049000 00001000 2**2
              CONTENTS, ALLOC, LOAD, READONLY, CODE
13 .text     000001a5 08049040 08049040 00001040 2**4
              CONTENTS, ALLOC, LOAD, READONLY, CODE
15 .rodata   0000000c 0804a000 0804a000 00002000 2**2
              CONTENTS, ALLOC, LOAD, READONLY, DATA
23 .data     0000000c 0804c010 0804c010 00003010 2**2
              CONTENTS, ALLOC, LOAD, DATA
24 .bss      00000008 0804c01c 0804c01c 0000301c 2**2
              ALLOC
...

```

Truncated for readability

Load Memory Address and File Offset have same alignment
→ Allows for easy mapping

No contents contained in the file for .bss segment

23.5 Name Mangling

Funktionen in C++ (bzw. Objektorientierten Sprachen) können den selben Namen mit unterschiedlichen Funktionsparametern haben. Dies wird auch Overloading genannt.

```

// C++
int foo(int a) {
    return 0;
}
int foo(int a, int b) {
    return 0;
}

```

```

user@PC $ nm mangling.o
0000000000000000 T _Z3fooi
000000000000000e T _Z3fooii
user@PC $ nm mangling.o | c++filt
0000000000000000 T foo(int)
000000000000000e T foo(int, int)

```

Demangling tool

Der Compiler benutzt name mangling um Funktionen zu unterscheiden: Er erstellt für jede Funktion aufgrund der name mangling Regeln einen eindeutigen Namen. Dies passiert nicht nur für Funktionen sondern auch für namespaces, ihre member functions und operators. Die erstellten mangled names sind nicht cross compiler kompatibel.

Initialisierung:

```

// C++
int a_foo_exists;
struct foo_t {
    foo_t() { a_foo_exists = 1; }
};
foo_t foo;

```

```

user@PC $ cc -S -o- initialization.cpp | c++filt
.text
...
foo_t::foo_t():
    ...
__static_initialization_and_destruction_0(int, int):
    ...
    call    foo_t::foo_t()

```

Die Initialisierung findet vor dem Aufruf von main statt. Hierbei wird ein plattformspezifischer Mechanismus verwendet. Der Compiler erzeugt in jedem Object File, das Initialisierungscode enthält, entsprechenden Code. Der Linker generiert dann eine Funktion __main, die sämtliche Initialisierungscodes aufruft, bevor das eigentliche main-Programm startet.

Zusatz Informationen:

```
// C++
struct foo_t {
    ~foo_t() { /* ... */ }
    void except() { throw 0; }
};

void fn() {
    foo_t foo;
    foo.except();
    /* ... */
}
```

Wird eine Exception geworfen werden alle Stack Variablen zerstört. Alle Variablen die durch die Exception betroffen sind müssen gefunden und zerstört werden, bis die Exception gehandelt wird. Diese Informationen müssen in speziellen Programm Sektionen gespeichert werden.

Debug Symbole sind weitere Daten die gespeichert werden. Sie müssen alle Variablen-, Funktions-, usw. Namen und die Position im Code speichern.

23.6 Linker Typen und Libraries

23.6.1 Typ 0 - dynamic Linking

Wird dynamisch gelinked ist nicht vorher spezifiziert, wann dies passiert. Potentiell kann erst zur Laufzeit des Programms mit einer Library gelinked werden. Z.b. wird dieser Ansatz oft für Plugins verwendet. Dies entspricht dem Prinzip des lazy-loadings. Code wird erst geladen wenn er wirklich benötigt wird.



23.6.2 Typ 1 - static shared Libraries

So gut wie alle Programme werden mit der C-Standardlibrary gelinked. Diese benötigt Platz im Programm. Jedoch wird System weit nur eine Implementation der Standardlibrary benötigt. LibC befindet sich nicht in jedem Programm sondern sie wird statisch mit dem Programm gelinked.

Das Programm bekommt ein “Shared library segment” an der selben virtuellen Adresse jedes Programms. Jeder Library wird eine einzigartige Adresse in diesem Segment zugeordnet und der Linker linkt das Programm mit der wirklichen Library. Kein Code der Library wird in das Programm übernommen.

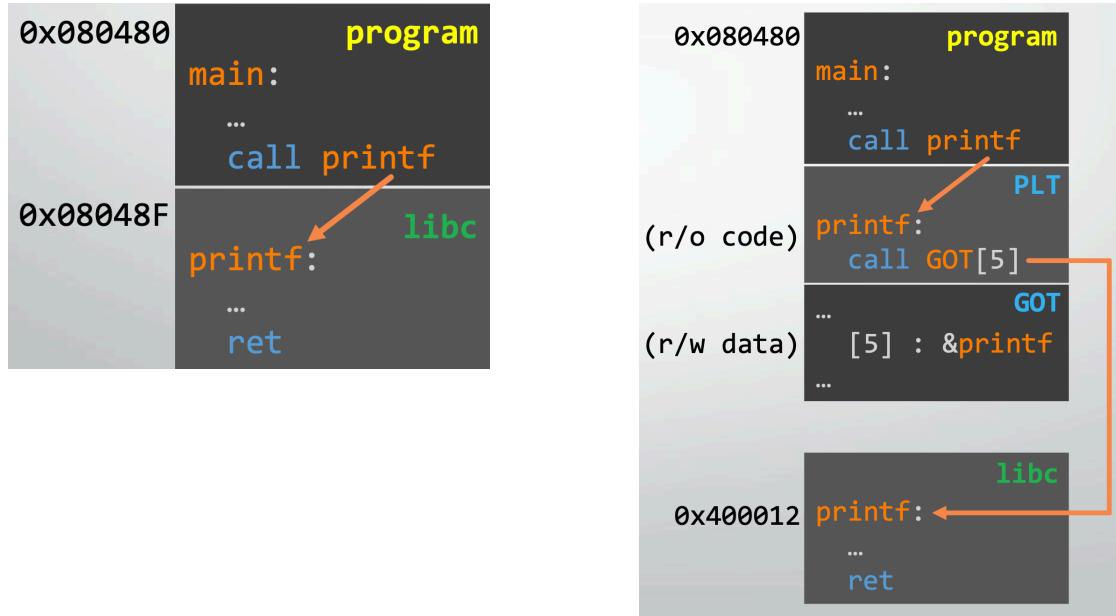
Somit können mehrere Programme den **selben** Code ausführen.

23.6.3 Typ 2 - dynamic shared Libraries

Typ 1 Libraries benötigen das systemweite festlegen bzw. preallocate im Adressspeicher für bestimmte Libraries. Dies ist aufwändig und nicht besonders elegant.

Um dies eleganter (und auch sicherer) zu lösen erlauben wir jeder Library an egal welcher Speicher Adresse zu liegen. Der Linker weiß nun jedoch nicht ob die Symbole wirklich vorhanden sind. Die Lösung dafür ist die Benutzung von sogenannten Stub-objects. Diese enthalten Symbole ohne Implementation.

Damit Funktionen aufgerufen werden können die an virtuell jeder Speicheradresse liegen können verwenden wir sogenannten position independent code (PIC). Dies benötigt einen weiteren Schritt:



23.7 Code = Data

Es gibt keinen wirklichen Unterschied zwischen Code und Daten. Code ist eine Sequenz an Instruktionen, die auf der CPU ausgeführt werden können und keine illegalen Instruktionen bzw. nicht ausführbare Daten enthalten. Code kann demnach auch während der Laufzeit geschrieben und gelesen werden.

Der Grund ist, dass dynamische Code Generierung einen extremen boost an Geschwindigkeit mit sich bringt. Interpretations Overhead in dynamischen Programmiersprachen wird eliminiert und die Geschwindigkeit verbessert sich um einen Faktor [10, 100].

Generell werden für Optimierungen Informationen benötigt. Diese Information sind zahlreicher während der Ausführung eines Programms als bei dem Kompilieren des Programms.

Nachteil der dynamischen Code Generierung ist das sich die Laufzeit des Programms auf die Programmlaufzeit + die Generierungslaufzeit erhöht.

Beispiel:

Bei der dynamischen Codegenerierung wird zunächst die binäre Kodierung von Assembler-Befehlen ermittelt. Anschließend wird dieser Binärkode in einen Speicherbereich (Buffer) geschrieben. Um Sicherheitslücken wie Buffer Overflow Exploits zu vermeiden, hat der Stack normalerweise keine Ausführungsrechte. Deshalb müssen explizit Ausführungsrechte für den Stack-Speicher gesetzt werden, damit der Code dort ausgeführt werden kann.

```
const int64_t a = 15;
const int64_t b = 27;
int64_t result = 0;

uint8_t code[] = {
    0x48, 0x89, 0xf8, // movq %rdi, %rax
    0x48, 0x01, 0xf0, // addq %rsi, %rax
    0xc3                // ret
};

enable_stack_code_execution(code);

result = ((int64_t(*)(int64_t, int64_t))code)(a, b);
```

Unter Unix x86-64 werden die ersten beiden Funktionsparameter in den Registern rdi und rsi übergeben. Schließlich wird der Code in einen Funktionszeiger des passenden Typs umgewandelt (gecastet), sodass er wie eine normale Funktion aufgerufen und ausgeführt werden kann.

Linking und Sicherheit:

Ein Buffer Overflow kann dazu führen, dass ein Angreifer ausführbaren Code in den Speicher legt und durch Überschreiben einer Rücksprungadresse ausführt. Um das zu verhindern, sorgen Linker für Sicherheitsmechanismen: Erstens darf ein Speichersegment niemals gleichzeitig schreibbar und ausführbar sein (W^X-Regel). Zweitens erschwert die zufällige Anordnung von Speicherbereichen ASLR(Address Space Layout Randomization) gezielte Angriffe, da sich die Speicherorte beim Programmstart jedes Mal ändern. Dadurch werden Angriffe schwieriger, aber nicht unmöglich.

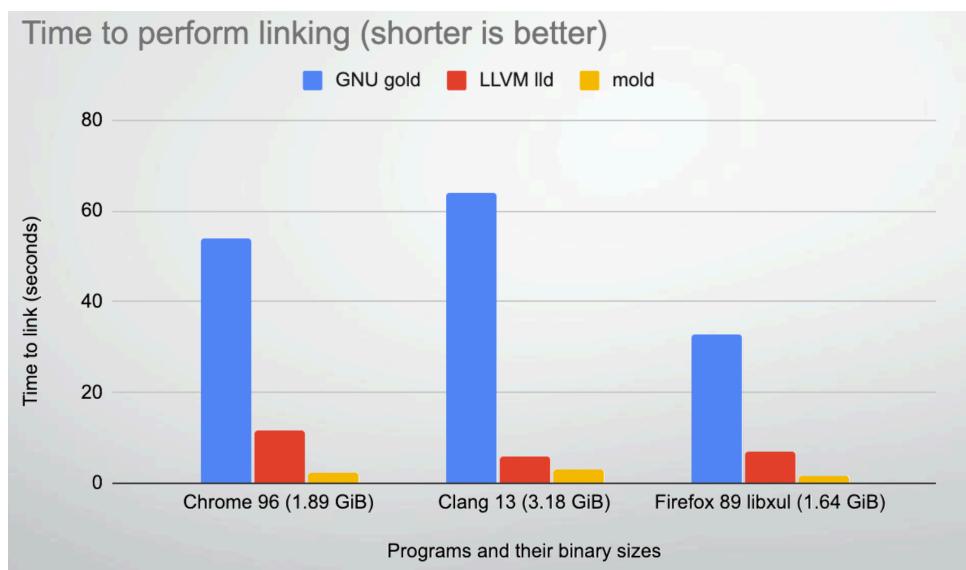
23.8 Alternative Linker

GNU gold ist ein von Google entwickelter Linker der Alternativ zu GNU ld für besonders große C++ Projekte nützlich ist. Auf den meisten Systemen steht er in dem package binutils enthalten. LLVM lld ist der default linker von LLVM (clang). Dieser ist signifikant schneller als zum Beispiel GNU gold.

mold ist ein moderner Linker. Er wird als Ersatz für GNUs ld/gold und LLVMs lld verwendet und ist signifikant schneller als lld. mold macht sich hier neue Multi-Core CPUs zunutze.

Um einen alternativen linker zu benutzen kann die flag `-f_use-ls=<linker>` mit dem Komplileraufruf kombiniert werden. In Makefiles wird der Linker mit `LD = <linker>` festgelegt.

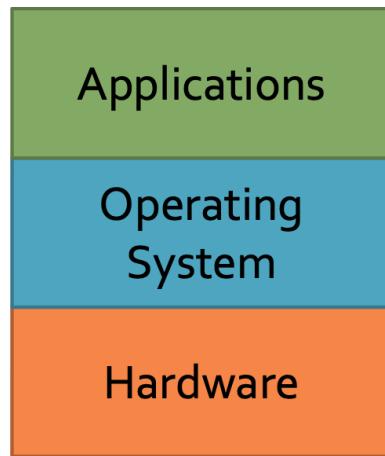
Performance Vergleich:



24 Virtualisierung

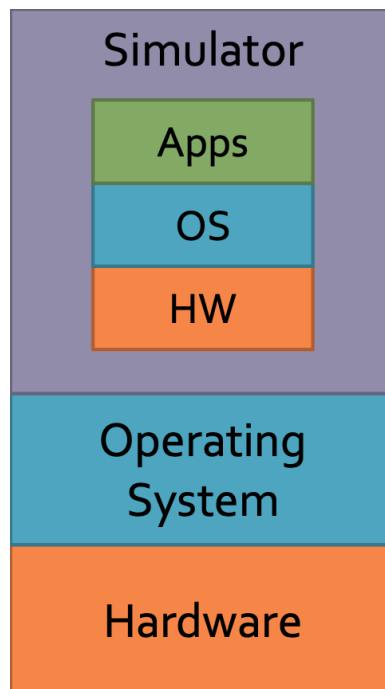
24.1 Computer Stack

Das Betriebssystem läuft direkt auf der Hardware des Systems. Es abstrahiert die Hardware und stellt eine Schnittstelle bereit, über die Programme auf dieser Hardware ausgeführt werden können. Programme können mithilfe des Betriebssystems direkt mit der Hardware kommunizieren. Damit Programme auf einer bestimmten Hardware lauffähig sind, müssen sie für den zugrunde liegenden Instruktionssatz kompiliert werden.



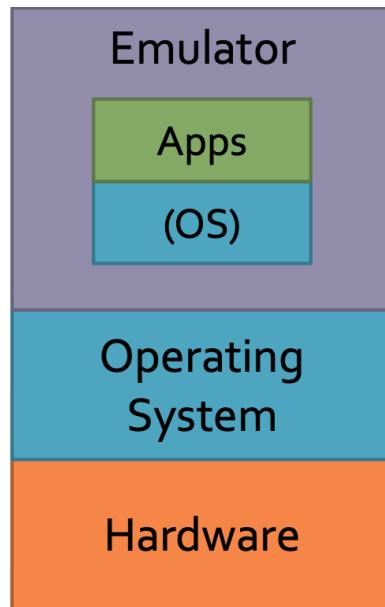
24.2 Simulation

Wenn Programme nicht auf der vorhandenen Hardware ausgeführt werden können, kann eine Simulation eingesetzt werden, die sämtliche Aspekte der benötigten Hardware nachbildet. Dadurch lassen sich Programme trotz inkompatibler Hardware ausführen, wobei sich sowohl das Betriebssystem als auch das Programm exakt so verhalten wie auf einem realen System. Ein Simulator ist dabei ein Programm, das auf einer anderen Hardware als dem ursprünglich vorgesehenen System läuft. Die Implementierung einer solchen Simulation ist jedoch äußerst komplex, ineffizient und meist sehr langsam.



24.3 Emulation

Ein weiterer Ansatz, um Programme auszuführen, die nicht für die vorhandene Hardware geschrieben wurden, ist die Emulation. Dabei werden sowohl das Betriebssystem als auch das Programm emuliert. Das bedeutet, dass sich Programm und Betriebssystem genauso verhalten, als würden sie auf der ursprünglichen Hardware laufen. Ein Emulator ist ein Programm, das die Hardware eines anderen Systems nachbildet. Im Vergleich zur Simulation sind Emulationen oft einfacher zu implementieren und in der Regel auch etwas schneller.

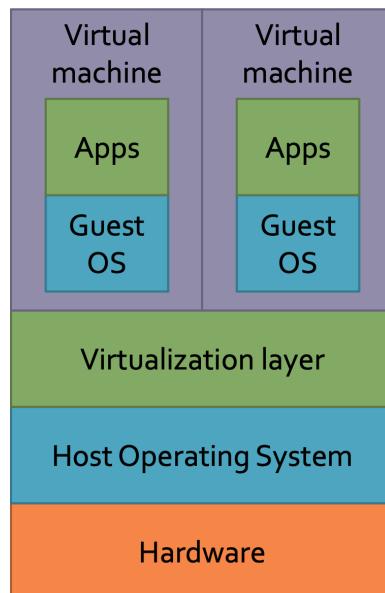


Ein häufig verwendetem Emulator ist QEMU. Er ist kostenlos und Open Source. Qemu hat unterschiedliche Modi der Emulation:

- **User-mode emulation:** Ermöglicht das Ausführen von ARM Programmen auf x86 System.
- **System emulation:** Ermöglicht das Ausführen eines Gastbetriebssystems (potenziell auf einer anderen Hardware). Betriebssysteme wie Linux, Windows, usw. können auf demselben System bzw. für eine andere Architektur emuliert werden
- **Hosting:** Ermöglicht das Integrieren anderer Virtualisierungs Technologien, wie KVM oder Xen.

24.4 Virtualisierungen

Die Virtualisierung ermöglicht es, Gastbetriebssysteme in einer virtuellen Maschine auszuführen. Der Code der einzelnen Virtuellen Maschinen wird auf der wirklichen Hardware des Systems ausgeführt. Dies benötigt die selbe Architektur für Gast- und Hostbetriebssystem. Ein gewisses Maß an hardware support wird benötigt. Die Virtuelle Maschine ist komplett isoliert zum Host bzw. anderen Virtuellen Maschinen.



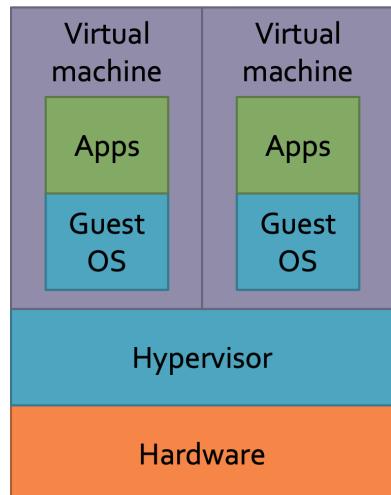
24.5 Hypervisor

Ein Hypervisor (oder Virtual Machine Monitor, VMM) ist für das Erstellen, Verwalten und Ausführen von Virtuellen Maschinen zuständig. Der Hypervisor verwaltet die Ressourcen von allen virtuellen Maschinen, wie zum Beispiel die Anzahl der CPU Kerne, die Größe des jeweils nutzbaren RAMs, usw..

24.5.1 Hypervisor - Typ 1

Ein Typ-1-Hypervisor, auch als „native Hypervisor“ oder „Bare-Metal-Hypervisor“ bezeichnet, läuft direkt auf der Hardware des Systems und benötigt kein Host-Betriebssystem. Er hat die vollständige Kontrolle über die Hardware und verwaltet alle Ressourcen eigenständig. Eine zusätzliche Abstraktionsschicht ist nicht erforderlich, da der Hypervisor die Hardware direkt anspricht und selbst abstrahiert.

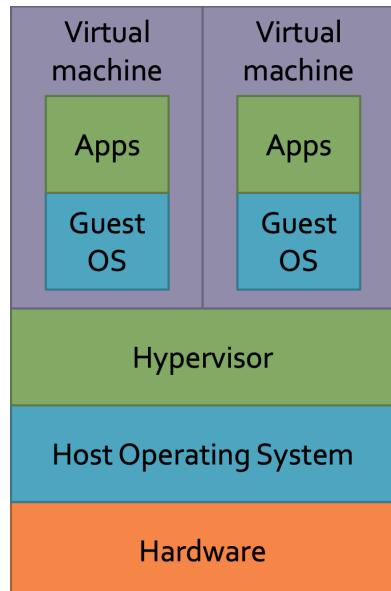
Beispiele hierfür sind Microsoft Hyper-V oder Citrix XenServer.



24.5.2 Hypervisor - Typ 2

Ein Typ-2-Hypervisor, auch als „Hosted Hypervisor“ bezeichnet, läuft auf einem bestehenden Betriebssystem. Der Hypervisor wird dabei wie ein normaler Prozess auf dem Host-Betriebssystem ausgeführt. Das Host-Betriebssystem bemerkt nicht, dass zusätzlich weitere Gastbetriebssysteme auf dem System laufen. Potenziell können mehrere Hypervisoren gleichzeitig auf dem System betrieben werden. Dieser Ansatz ist in der Regel langsamer, da die Kommunikation zwischen Host-Betriebssystem und Hypervisor zusätzliche Ressourcen benötigt.

Beispiele hierfür sind VMware Workstation oder Oracle VirtualBox.

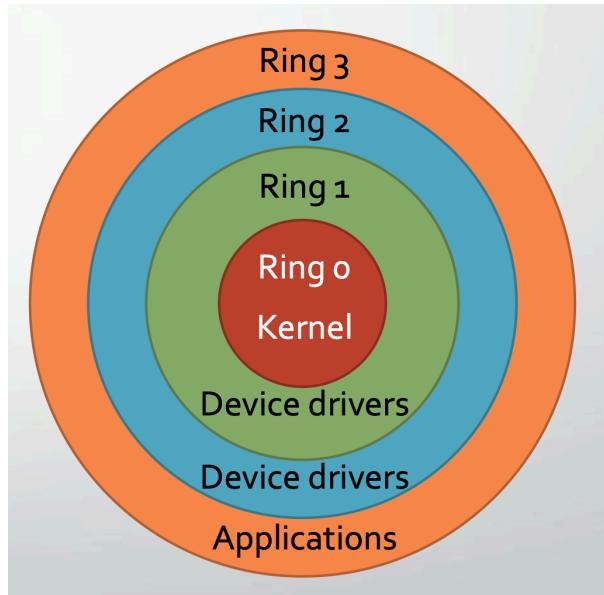


24.6 Volle Virtualisierung

Bei einer vollständigen Virtualisierung kann das Gastbetriebssystem unverändert auf der Hostmaschine ausgeführt werden. Das Gastbetriebssystem erkennt dabei nicht, dass es in einer virtuellen Maschine läuft. In der Regel wird hierfür ein Typ-1-Hypervisor verwendet. Die virtuelle Maschine ist vollständig isoliert und kann nur Daten bzw. Attribute innerhalb der eigenen Umgebung verändern. Um dies zu ermöglichen, wird Hardware-Unterstützung benötigt, wie zum Beispiel Intel VT-x oder AMD-V für die x86-64-Architektur.

24.7 Virtualisierung Implementation

Ursprünglich hatte x86 keine Schutzmechanismen (“real mode”), sodass alle Prozesse auf alles zugreifen konnten. Mit dem Protected Mode wurde es möglich, den Zugriff von Anwendungen einzuschränken. Der Prozessor startet im Real Mode und wechselt dann in den Protected Mode. Das Betriebssystem nutzt Ring 0 für privilegierte Instruktionen (Kernel Mode), während normale Anwendungen in Ring 3 (User Mode) laufen.



Ring 0 erlaubt den Zugriff auf alle CPU-Befehle und die Hardware, während Ring 3 keinen direkten Zugriff auf die Hardware erlaubt und privilegierte Befehle nicht ausgeführt werden können. Fehler in Ring 3 werden vom Betriebssystem in Ring 0 abgefangen und verarbeitet. Die Ringe 1 und 2 waren ursprünglich für Ein-/Ausgabe-Operationen gedacht, werden aber von modernen Betriebssystemen wie Windows oder Linux kaum verwendet. Operationen in Ring 1 oder 2 verursachen zusätzlichen Aufwand, da sie oft einen Wechsel zu Ring 0 erfordern.

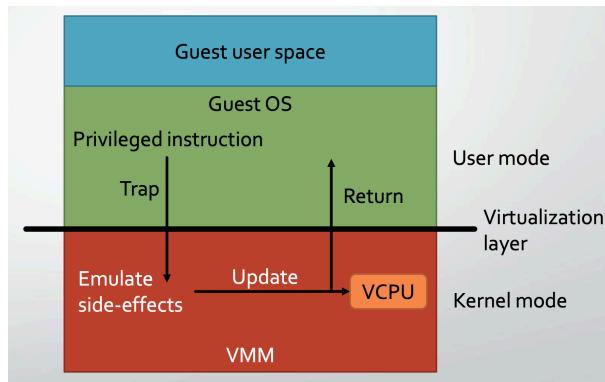
24.7.1 Virtueller Kernel Modus:

Benutzercode in einer virtuellen Maschine kann im Ring 3 ausgeführt werden und hat damit keinen Zugriff auf kritische Hardware. Der Kernel des Gastsystems weiß jedoch nicht, dass er virtualisiert ist und möchte daher in Ring 0 laufen. Das Host-Betriebssystem darf dem Gast-Kernel aber keinen Zugriff auf Ring 0 erlauben, da sonst die virtuelle Maschine verlassen werden könnte. Daher muss das Gastbetriebssystem sicher innerhalb der virtuellen Maschine eingeschlossen werden.

24.7.2 Virtuelle CPU (VCPU)

Der Hypervisor speichert den CPU-Zustand jeder virtuellen Maschine. Wenn eine virtuelle Maschine ausgeführt werden soll, kann der Hypervisor auf den gespeicherten VCPU-Zustand zugreifen und diesen wiederherstellen. Das funktioniert gut für normalen, nicht privilegierten Code. Es stellt sich jedoch die Frage, wie damit Code vom Gast-Betriebssystem-Kernel umgegangen wird.

24.7.3 Trap and Emulate

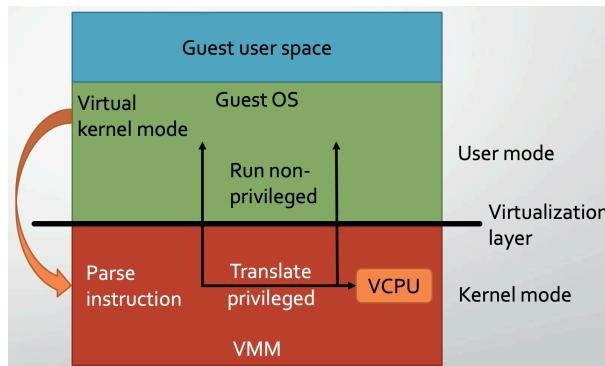


Der Virtual Machine Monitor (VMM) stellt jeder virtuellen Maschine zwei Modi zur Verfügung: den virtuellen User- und den virtuellen Kernel-Modus. Beide laufen im echten User-Modus auf der Hardware. Das Gastbetriebssystem darf nicht in den echten Kernel-Modus wechseln, sondern nur in den vom VMM bereitgestellten virtuellen Kernel-Modus.

Wenn das Gastbetriebssystem versucht, eine privilegierte Instruktion im User-Mode auszuführen, entsteht ein Fehler. Der Virtual Machine Monitor (VMM) erhält dadurch die Kontrolle, analysiert den Fehler und emuliert die Wirkung der privilegierten Instruktion. Anschließend gibt der VMM die Kontrolle wieder an das Gastbetriebssystem zurück.

Benutzercode in der virtuellen Maschine verursacht keinen Performance-Verlust. Kernelcode dagegen muss vom Hypervisor abgefangen und emuliert werden, was die Leistung verringert. Dieser Overhead tritt bei jeder weiteren virtuellen Maschine erneut auf.

24.7.4 Binary Translation



Bei sehr alten x86-CPUs entsteht bei privilegierten Befehlen kein Fehler für Trap-and-Emulate. Deshalb analysiert der Virtual Machine Monitor beim Eintritt in den virtuellen Kernel-Modus den Instruktionsstrom. Normale Befehle werden direkt ausgeführt, während privilegierte Befehle in nicht privilegierte übersetzt werden. Die Nebenwirkungen werden dabei im VCPU-Zustand gespeichert.

Das Lesen, Parsen, Übersetzen und Ersetzen von Gast-Kernel-Code ist sehr langsam. Dieser Performance-Verlust kann jedoch verringert werden, indem die Übersetzung nur einmal durchgeführt und das Ergebnis im VMM zwischengespeichert wird. Beim nächsten Mal kann dann das zwischengespeicherte Ergebnis verwendet werden.

24.7.5 Shadow Page Tables

Der Virtual Machine Monitor (VMM) muss auch den Zugriff auf den virtuellen Speicher verwalten. Das Gastbetriebssystem darf nicht direkt auf die echten Seitentabellen zugreifen, da es sonst aus der virtuellen Maschine ausbrechen könnte. Deshalb verwaltet der VMM sogenannte Shadow Page Tables und fängt Zugriffe darauf ab, um sie in Software zu emulieren.

24.8 Paravirtualisierung und Hardware Support

24.9 VM Operationen

24.10 Betriebssystem Virtualisierung

25 Memory Management