操作系统 Operating System

南方科技大学 计算机科学与工程系 11812804 董正

前言 Preface

笔记结构基于《操作系统概念（第九版）》

Based on *Operating System Concepts Ninth Edition*

第三章 进程 Process

3.1 基本概念

- CPU 活动
  - 批处理系统: 作业 (job)
  - 分时系统: 用户程序 (user program) 或任务 (task)

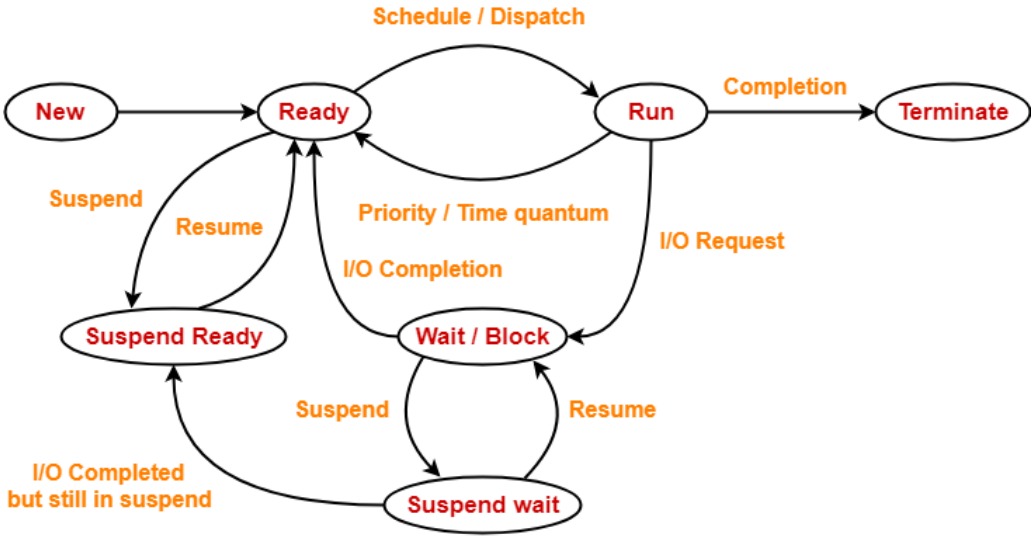
3.1.1 进程的概念

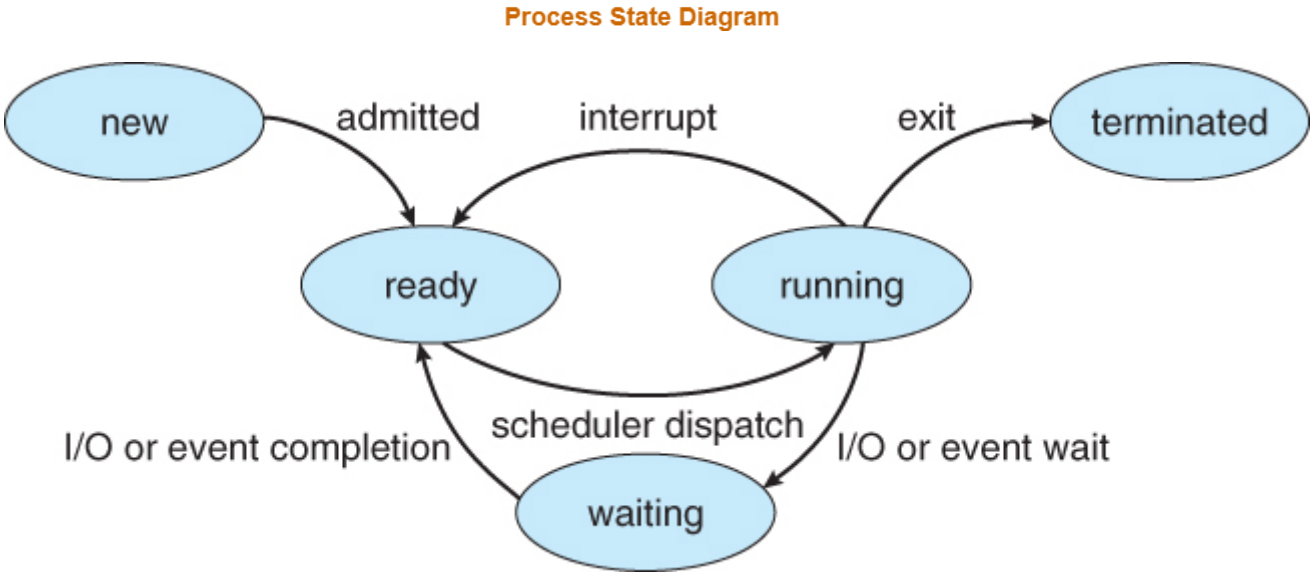
- **进程 (Process)** 是执行的程序

  Process is a program in execution
- 进程还包括:
  - 程序计数器 PC
  - 寄存器 (Register) 的内容
  - 堆 Heap
  - 栈 Stack
  - 数据段 Data Section
  - 文本段 Text Section
  - ...
- 程序 (Program) 和进程
  - 程序是**被动实体 (passive entity)**，如存储在磁盘上包含一系列指令的文件，经常称为可执行文件 (executable file)
  - 进程是**活动实体 (active entity)** 或称**主动实体**，具有一个程序计数器用来表示下个执行命令和一组相关资源
  - 当一个可执行文件被加载到内存时，这个程序就成为进程

3.1.2 进程的状态 Process States

状态	英文	说明
新的	new	进程正在创建
运行	running	指令正在执行
等待	waiting/blocked	进程等待发生某个事件，如 IO 完成或收到信号
就绪	ready	进程等待分配处理器
终止	terminated	进程已经完成执行



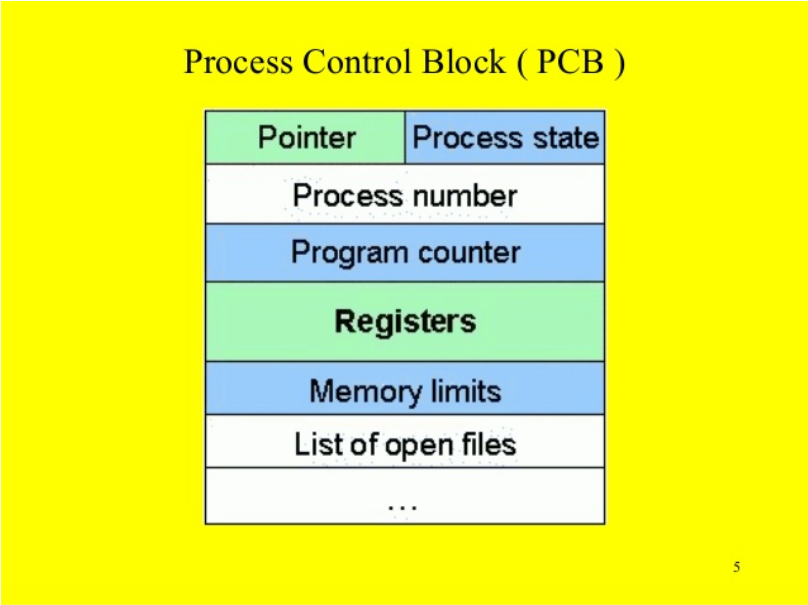


- 等待状态又分为
  - 可中断 (Interruptable)
  - 不可中断 (Un-interruptible)
- 刚 fork 的进程都会变成 ready 状态

3.1.3 进程控制块 Process Control Block

进程控制块 PCB (任务控制块 Task Control Block)

在内存 (Main Memory) 里



PCB 是系统感知进程存在的唯一标志

- 进程状态 Process State
- 程序计数器 PC
- CPU 寄存器 CPU Register
- CPU 调度信息 CPU-scheduling Infomation

  进程优先级，调度队列的指针和其他调度参数
- 内存管理信息 Memory-management Infomation

  基地址，界限寄存器的值，页表或段表等
- 记账信息 Accounting Infomation

  CPU 时间，实际使用时间，时间期限，记账数据，作业或进程数量等
- IO 状态信息 IO Status Infomation

  分配给进程的 IO 设备列表，打开文件列表等

Array of opened files contains:

Array Index	Description
0	Standard Input Stream; **`FILE *stdin;`**
1	Standard Output Stream; **`FILE *stdout;`**
2	Standard Error Stream; **`FILE *stderr;`**
3 or beyond	Storing the files you opened, e.g., **`fopen()`**, **`open()`**, etc.

- That's why a parent process **shares the same terminal output stream** as the child process.

- 几个概念

- PCB = 进程表 = `task_struct` in Linux
- Task list = PCB 组成的双向链表

3.2 进程生命周期

3.2.1 进程标识符 Process Identifier

- 进程标识符 Process Identifier (PID)

  系统的每个进程都有一个唯一的整数 PID

  System call `getpid()`: return the PID of the calling process

- `init` 进程

  PID = 1, 所有用户进程的父进程或根进程

  代码位于 `/sbin /init`

  它的第一个任务是创建进程 `fork()+exec*()`

3.2.2 进程创建 Process Creation

- System call `fork()`: creates a new process by duplicating the calling process.

  fork 出的子进程从 fork 调用的下一行开始执行（因为 PC 也复制了）

Cloned items	Descriptions
Program counter [CPU register]	That's why they both execute from the same line of code after **`fork()`** returns.
Program code [File & Memory]	They are sharing the same piece of code.
Memory	Including local variables, global variables, and dynamically allocated memory.
Opened files [Kernel's internal]	If the parent has opened a file "A", then the child will also have file "A" opened automatically.

- **`fork()`** does not clone the following...
- Note: PCB is in the kernel space.

Distinct items	Parent	Child
Return value of `fork()`	PID of the child process.	0
PID	Unchanged.	Different, not necessarily be "Parent PID + 1"
Parent process	Unchanged.	Parent.
Running time	Cumulated.	Just created, so should be 0.
[Advanced] File locks	Unchanged.	None.

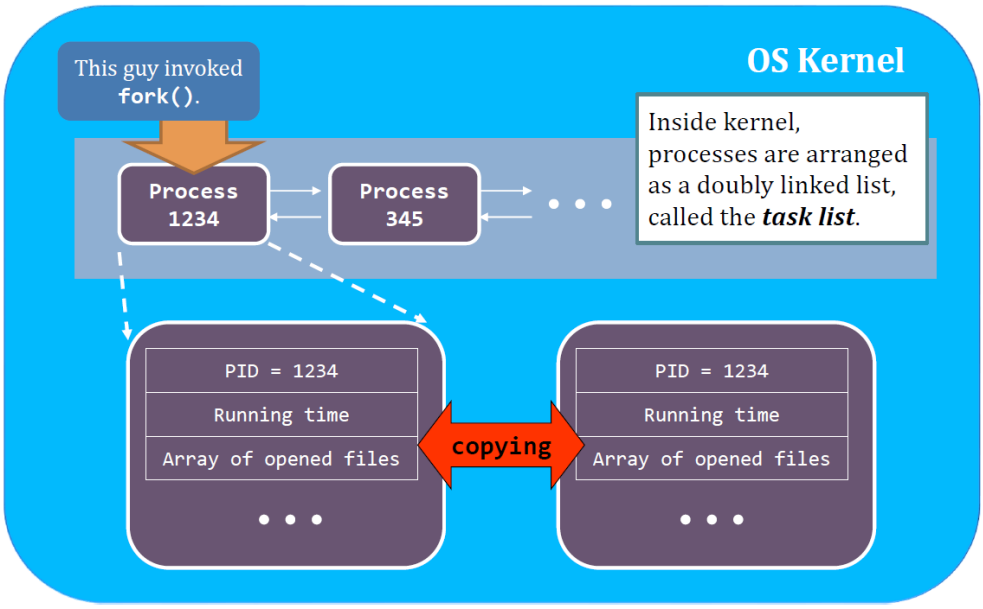
- 在代码中如何区分父进程和子进程

  `pid=fork()`, 则父进程中 `pid` 变量等于子进程的 PID， 子进程中 `pid=0`

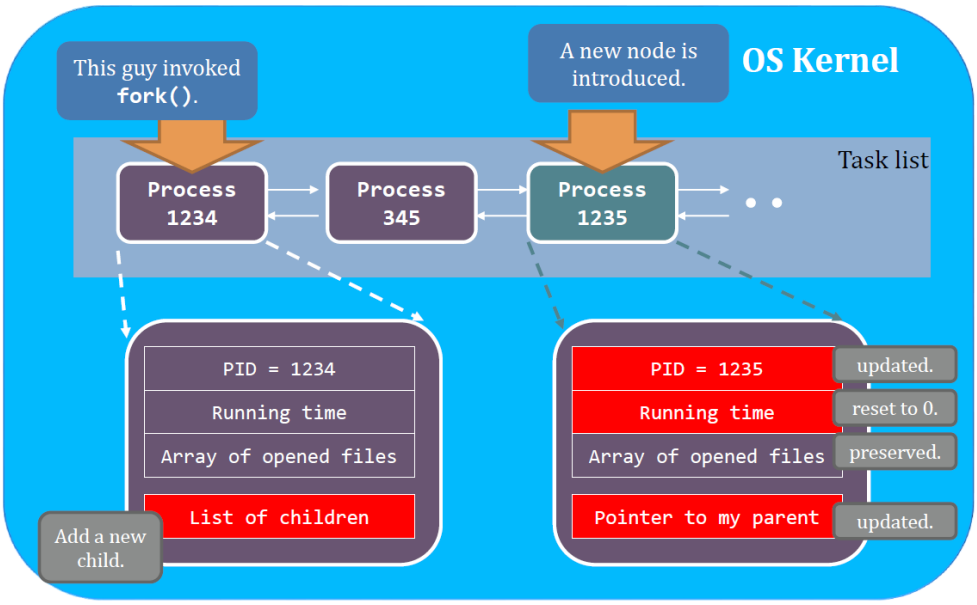
```c
if (!pid) {
    // 只有子进程执行
}
else {
    // 只有父进程执行
}
```

- 父进程和子进程执行顺序不确定
- `fork` 的流程, 内核空间的动作

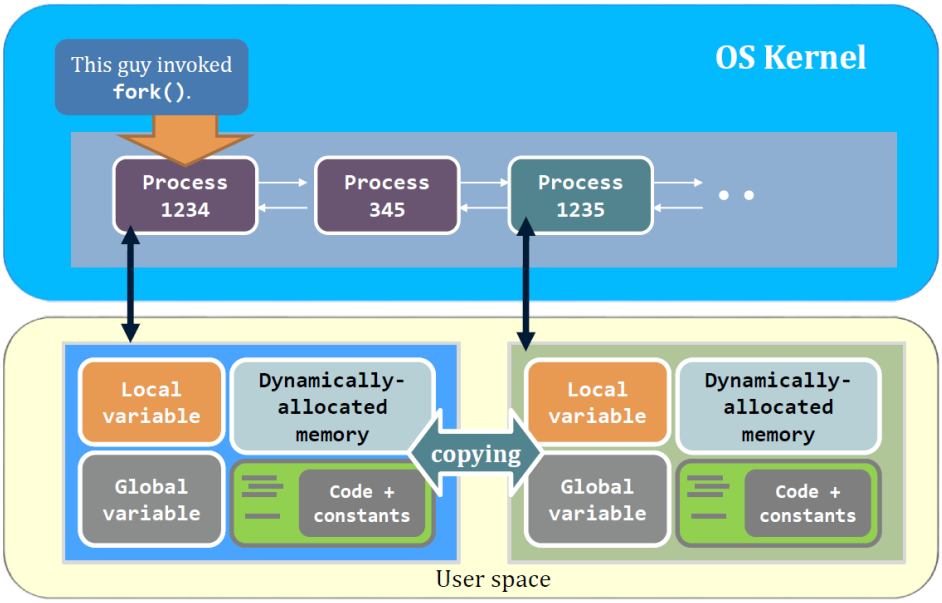
  1. 复制 PCB



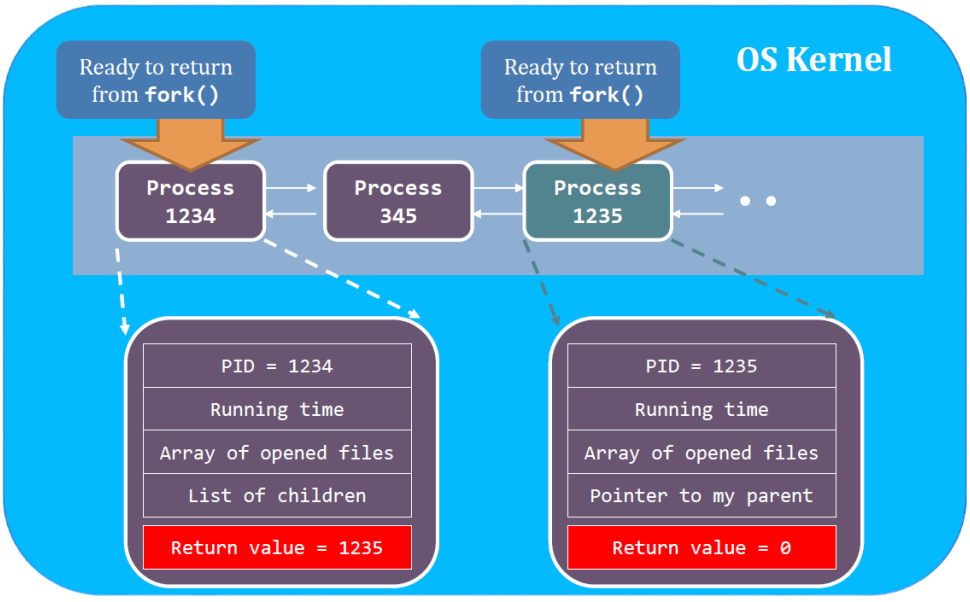
  2. 更新 PCB 和 task list



3. 复制用户空间



4. return



3.2.3 进程执行 Process Execution

- System call `exec*()`
- Example: `ls -l`

```
execl("/bin/ls", "/bin/ls", "-l", NULL);
```

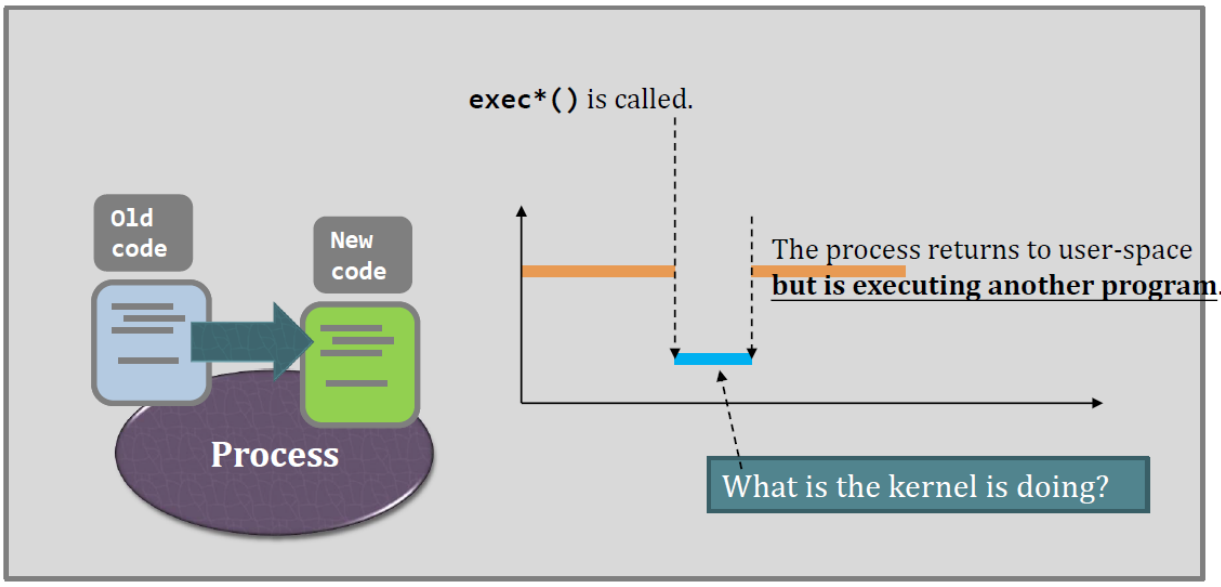
Argument Order	Value in above example	Description
1	`"/bin/ls"`	The file that the programmer wants to execute.
2	`"/bin/ls"`	When the process switches to `"/bin/ls"`, this string is the **program argument[0]**.
3	`"-l"`	When the process switches to `"/bin/ls"`, this string is the **program argument[1]**.
4	`NULL`	This states the end of the program argument list.

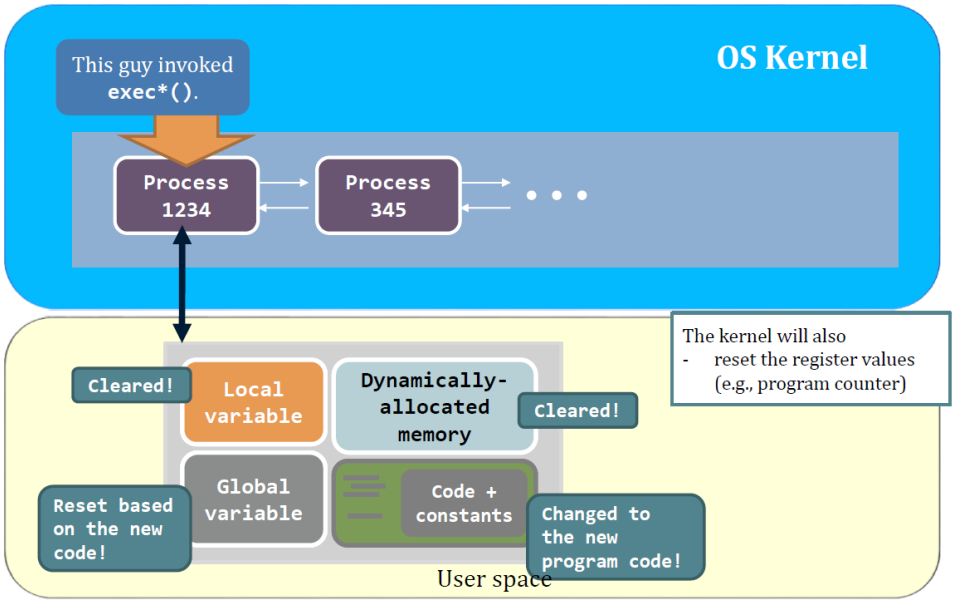
`args[0]` 是程序的名字

- The process is changing the code that is executing and never returns to the original code.

  `exec*()` 之后的代码不会执行了，因为调用之后该进程就去执行 `exec` 指定的程序了

- User space 的信息被覆盖
  - Program Code
  - Memory
    - Local Variables
    - Global Variables
    - Dynamically Allocated Memory
  - Register Value: 如 PC
- Kernel space 的信息保留: PID, 进程关系等
- `exec*()` 的内核执行过程

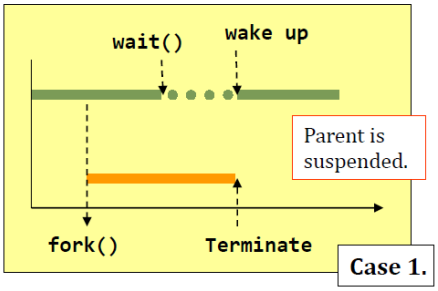


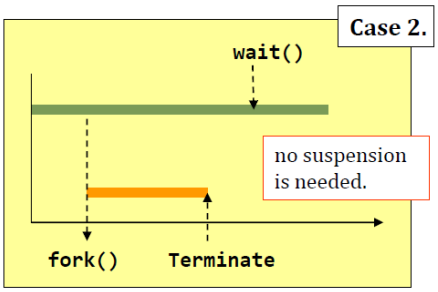


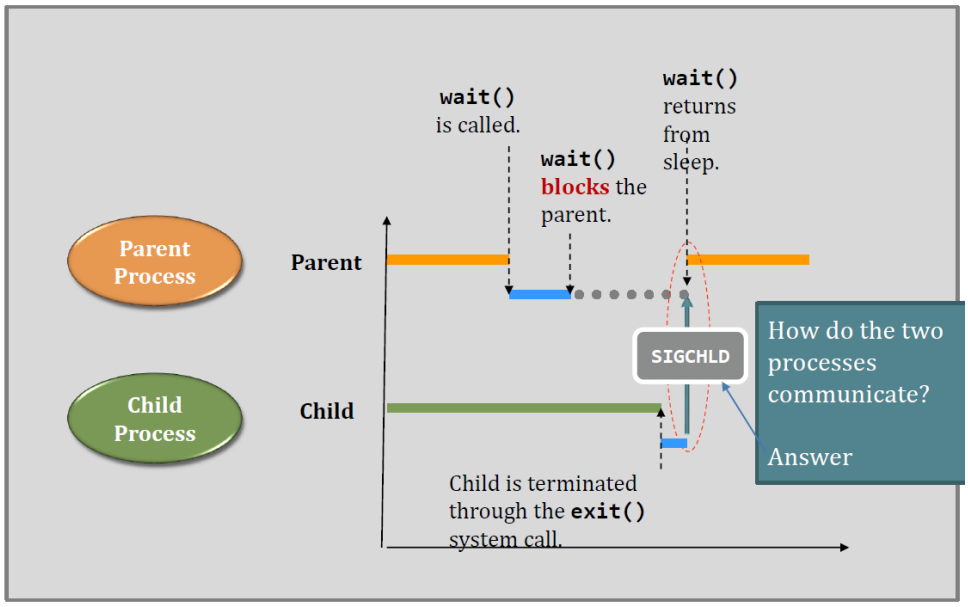
3.2.4 进程等待

1. System call `wait()`
    - Suspend the calling process to waiting state and return (wakes up) when
        - one of its child processes changes from running to terminated
        - received a signal
    - Return immediately (i.e., does nothing) if
        - it has no children
        - a child terminates before the parent calls wait
    - 给子进程收尸 见 3.2.6

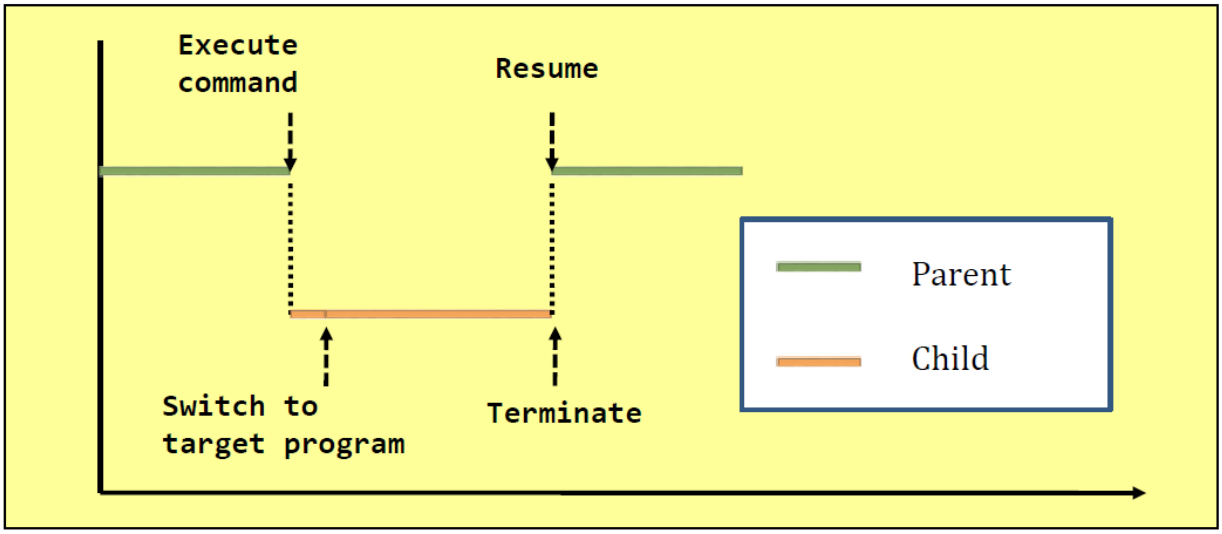
wait()	vs	waitpid()
Wait for any one of the children.		Depending on the parameters, **waitpid()** will wait for a particular child only.
Detect child **termination** only.		Depending on the parameters, **waitpid()** **can detect multiple child's status change**







2. `fork()+exec*()+wait()=system()`



例: shell 里输入命令 -> 执行相应程序 -> 程序终止 -> 返回 shell

- 除了 `init`, 所有的进程都是 `fork()+exec*()` 来的

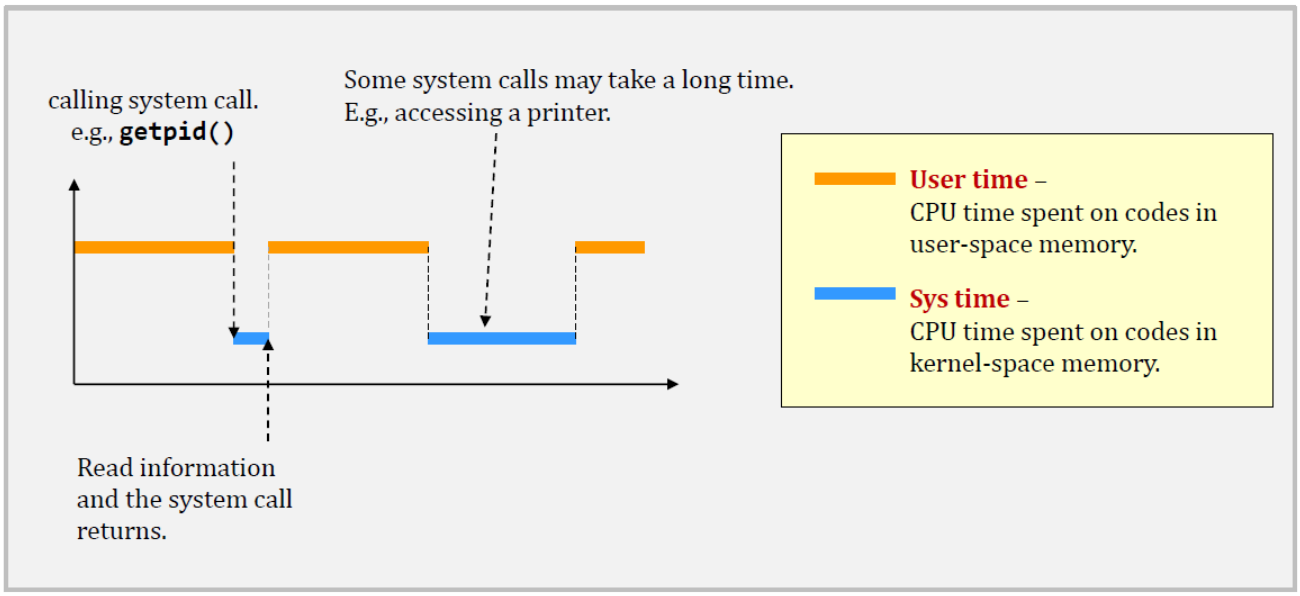
3.2.5 进程时间

- 实际时间 Real Time

  Wall-clock time
- 用户时间 User Time

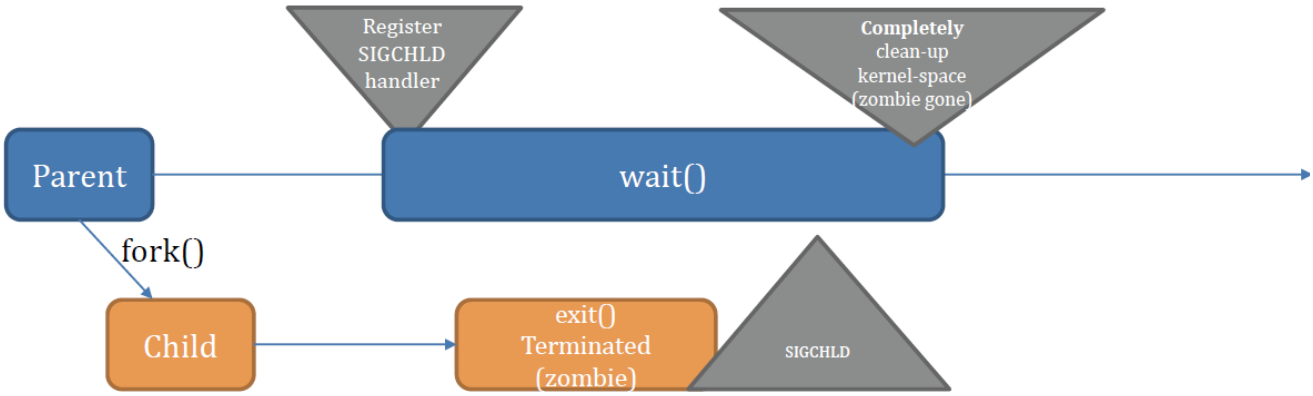
  CPU 在用户空间花费的时间
- 系统时间 System Time

  CPU 在内核空间花费的时间

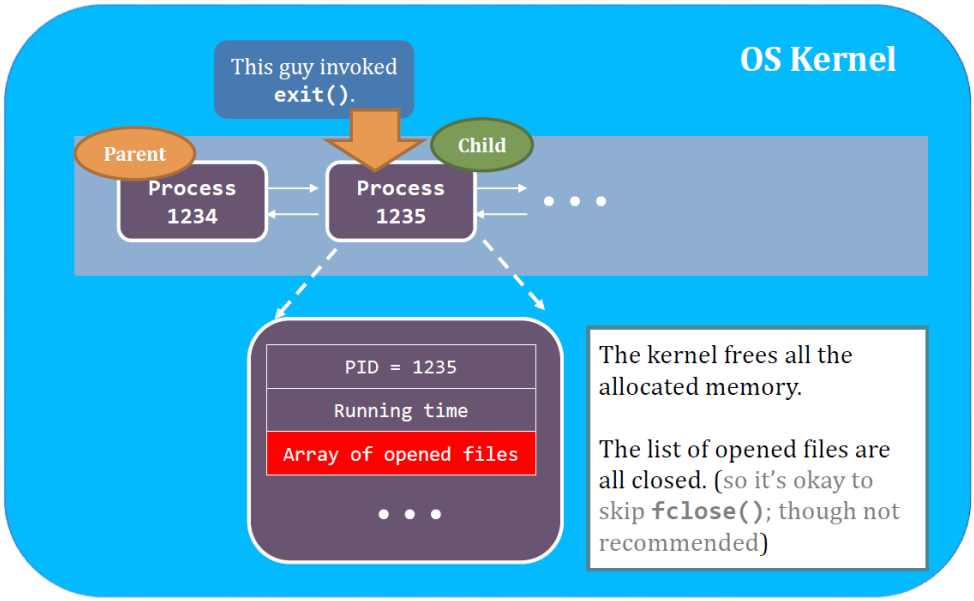


- User Time + Sys Time 决定了程序的性能 (Performance)
  - User Time + Sys Time > Real Time: 单核
  - User Time + Sys Time < Real Time: 多核

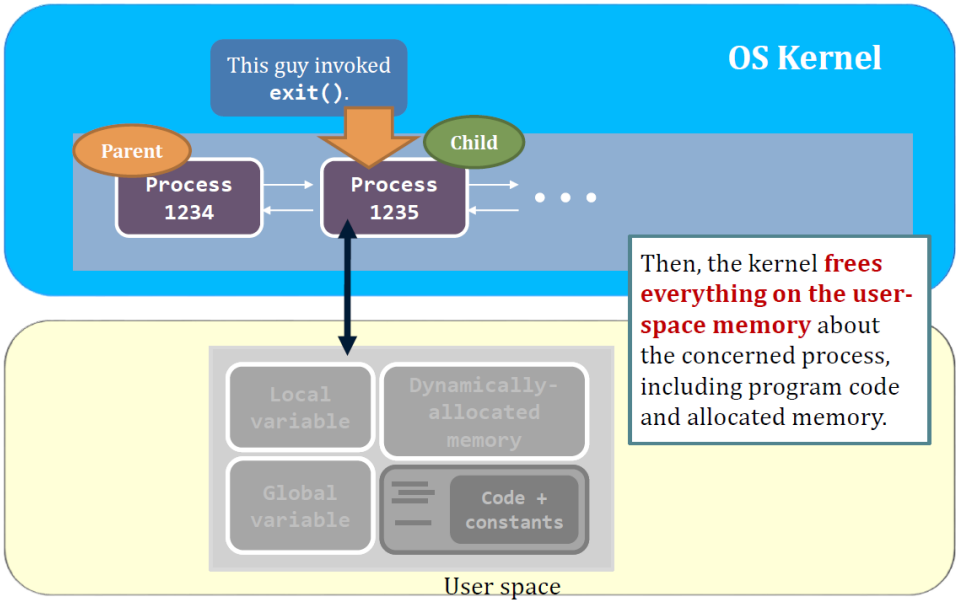
3.2.6 进程终止 Process Termination



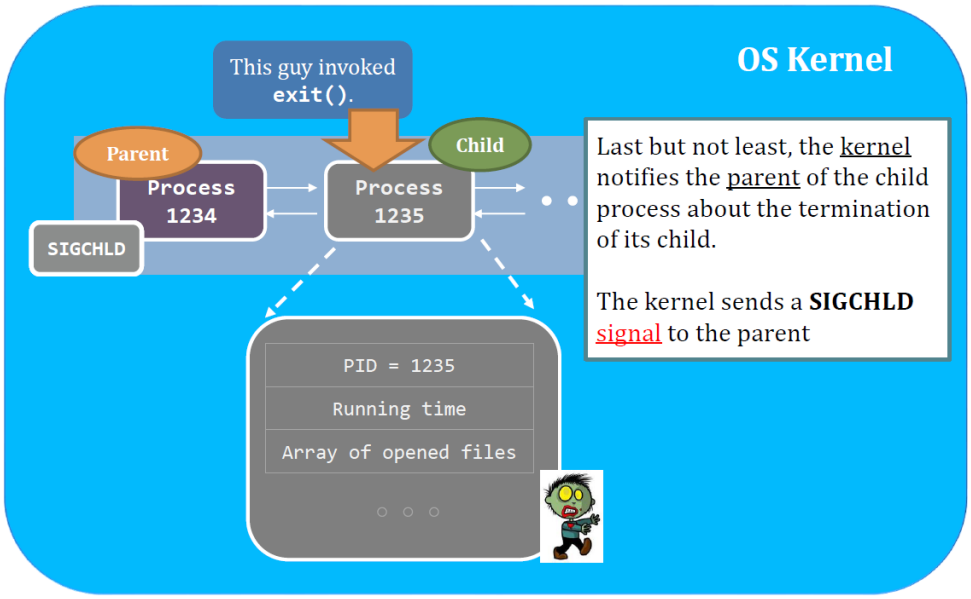
- System call `exit()`: terminate the calling process
- `exit()` 的执行过程
    1. Clean up most of the allocated kernel space memory



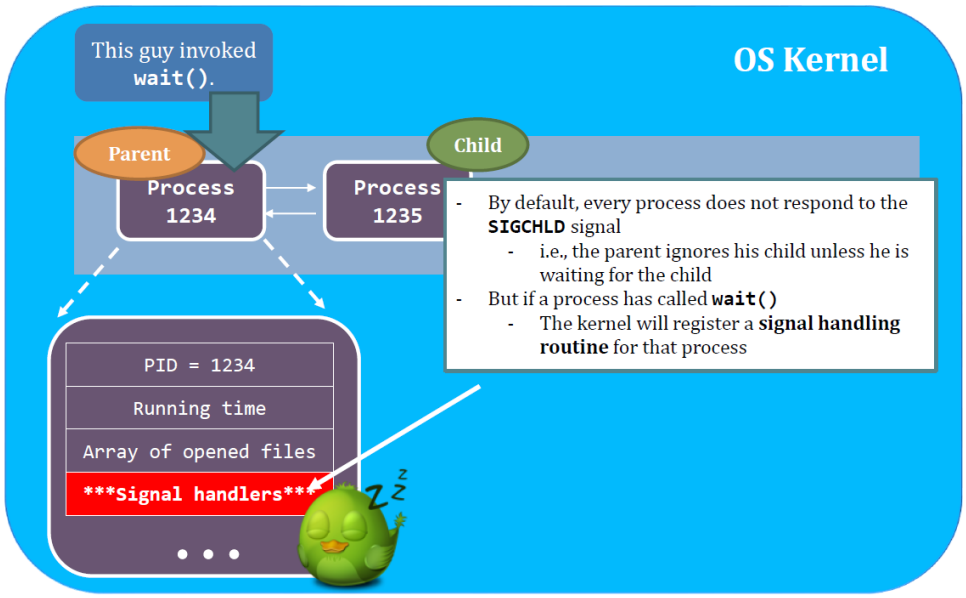
    2. Clean up the exit process's user space memory

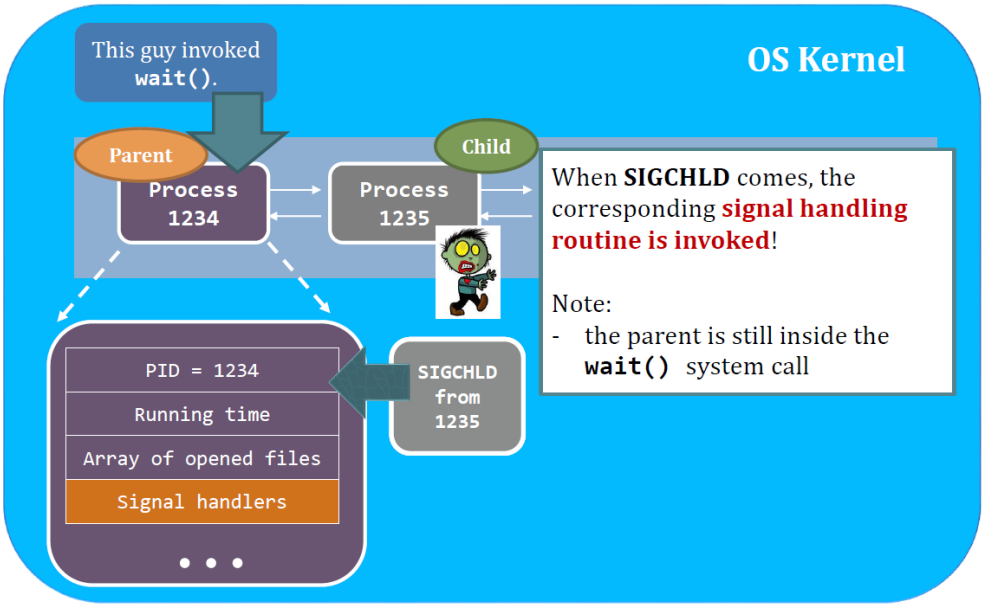


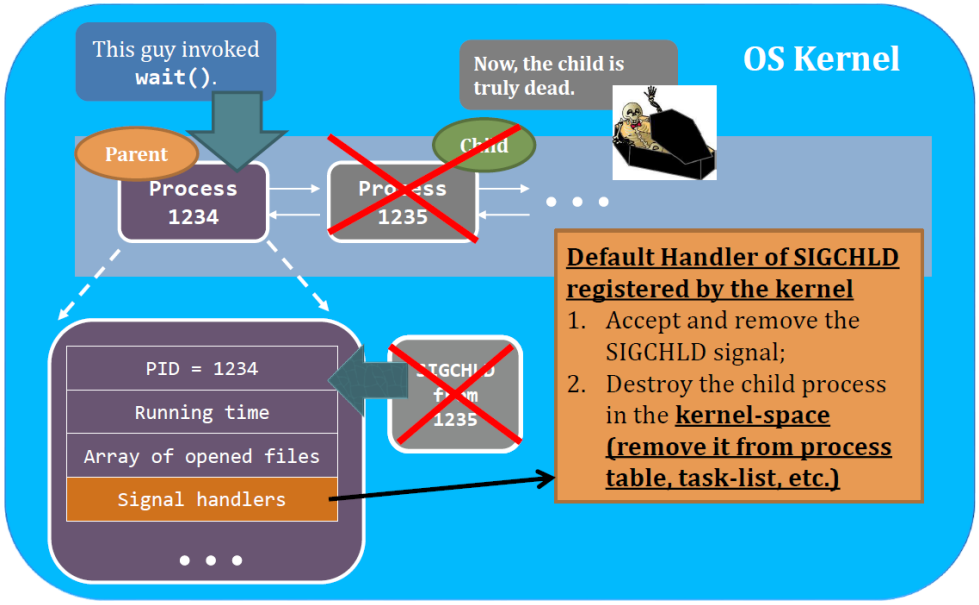
    3. Notify the parent with SIGCHLD.

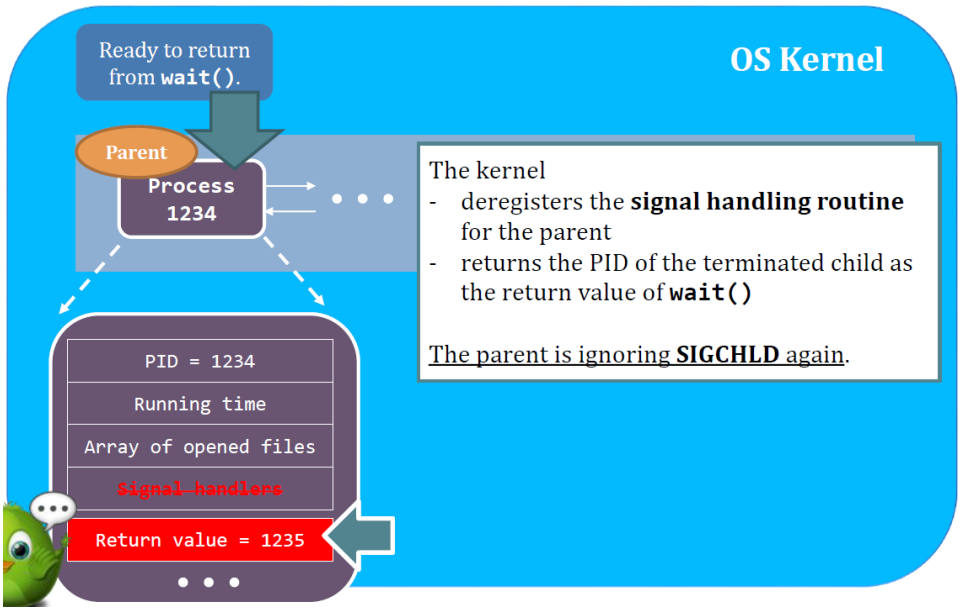


- 僵尸进程 Zombie Process
  - 进程的用户空间和内核空间被释放之后，PID 依然在进程表里，直到父进程调用 `wait()`
  - 当进程已经终止，但是其父进程尚未调用 `wait()`，这样的进程称为僵尸进程
  - 所有进程终止时都会过渡到这个状态
  - 一旦父进程调用 `wait()`，僵尸进程的进程标识符和它在进程表中的条目就会释放









- 子进程先终止，父进程再调用 `wait()` 也可以，SIGCHLD 不会消失，但是这段间隔内僵尸进程就一直存在、占用资源
- Linux 系统中僵尸进程被标为 `<defunct>`

  查看: `ps aux | grep <defunct>`

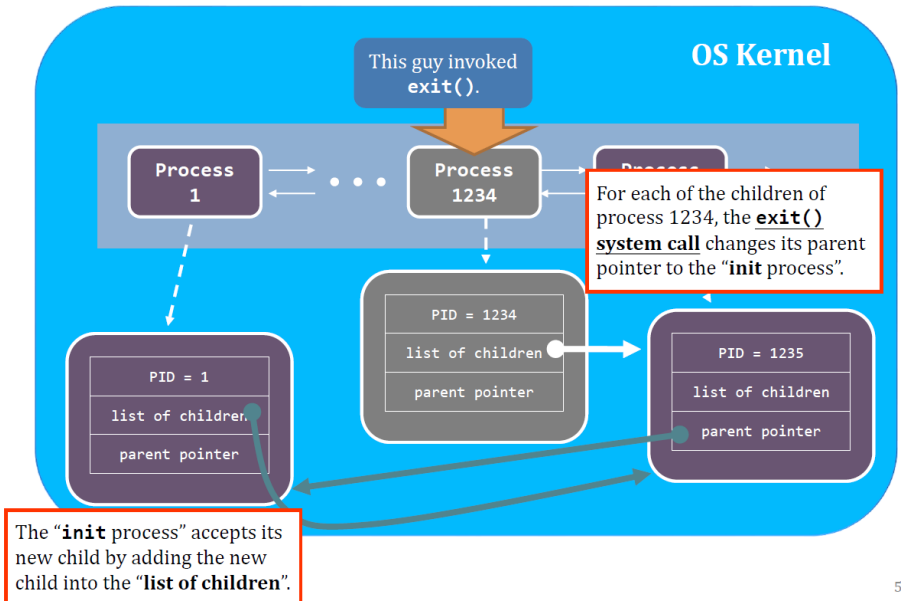
    - `exit()` system call turns a process into a zombie when
        - The process calls `exit()`
        - The process returns from `main()`
        - The process terminates abnormally

          这种情况下 kernel 会帮忙给他调用 `exit()`
- The fork bomb
    - PID 是有限的，Linux 中 PID 最大值为 32768

      `cat / proc /sys/ pid_max`
    - fork bomb (僵尸大军)

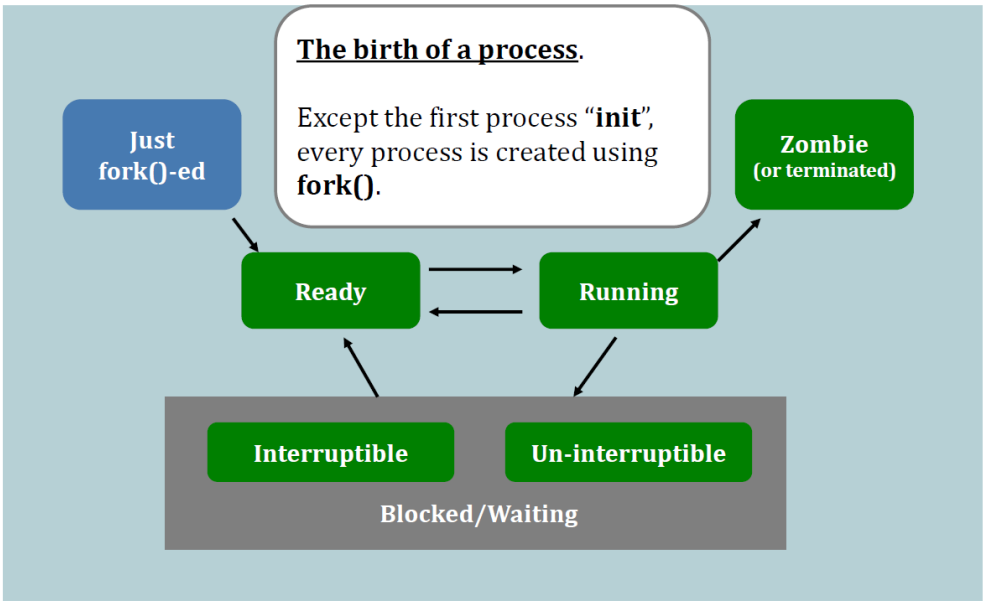
```c
int main() {
    while (fork());
    return 0;
}
```

- 孤儿进程 Orphan Process
    - 父进程没有调用 `wait()` 就终止，子进程变成孤儿进程
    - Linux & UNIX: 将 `init` 进程作为孤儿进程的父进程 (Re-parent)
    - 这个操作在 `exit()` 里完成，见下图
- `init` 进程定期调用 `wait()` 以便收集任何孤儿进程的退出状态，并释放孤儿进程标识符和进程表条目

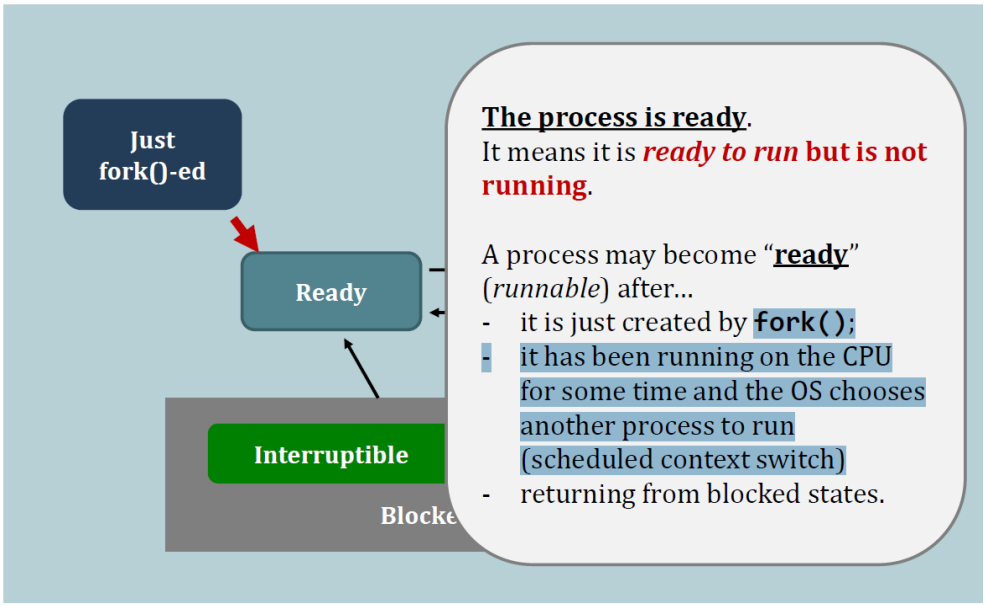


3.2.7 进程生命周期 Process Lifecycle

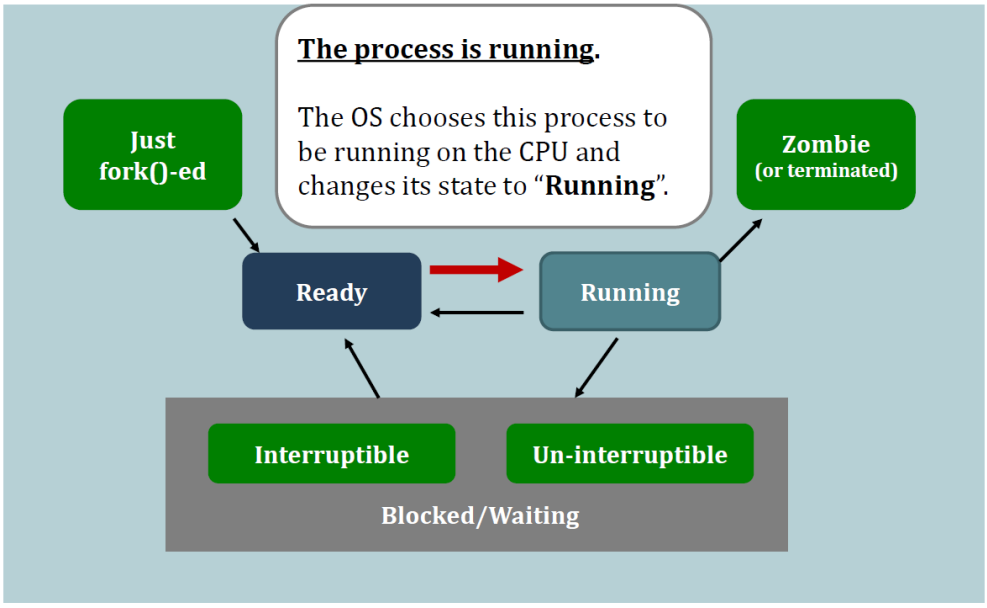
1. forked



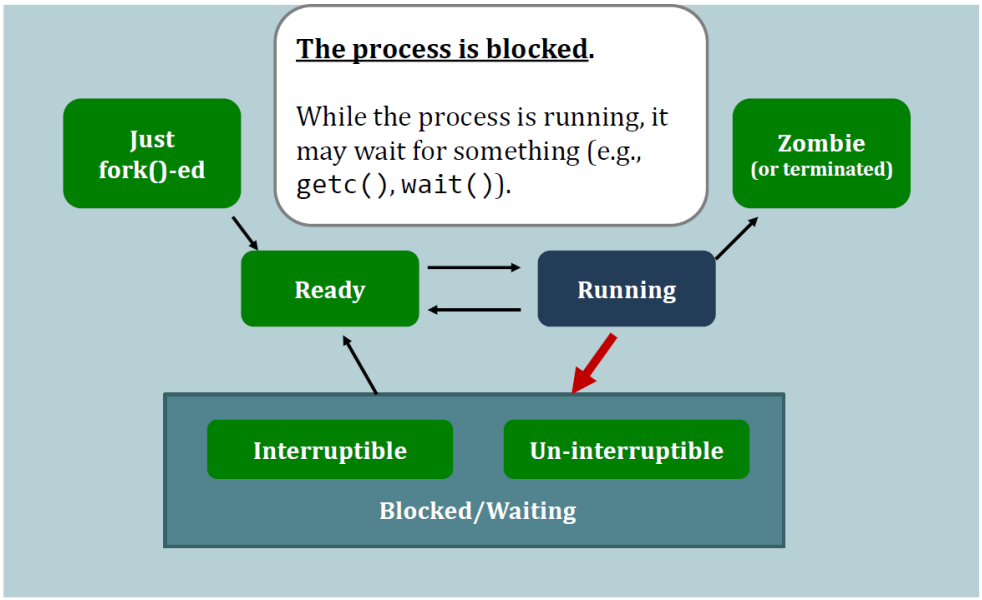
2. Ready



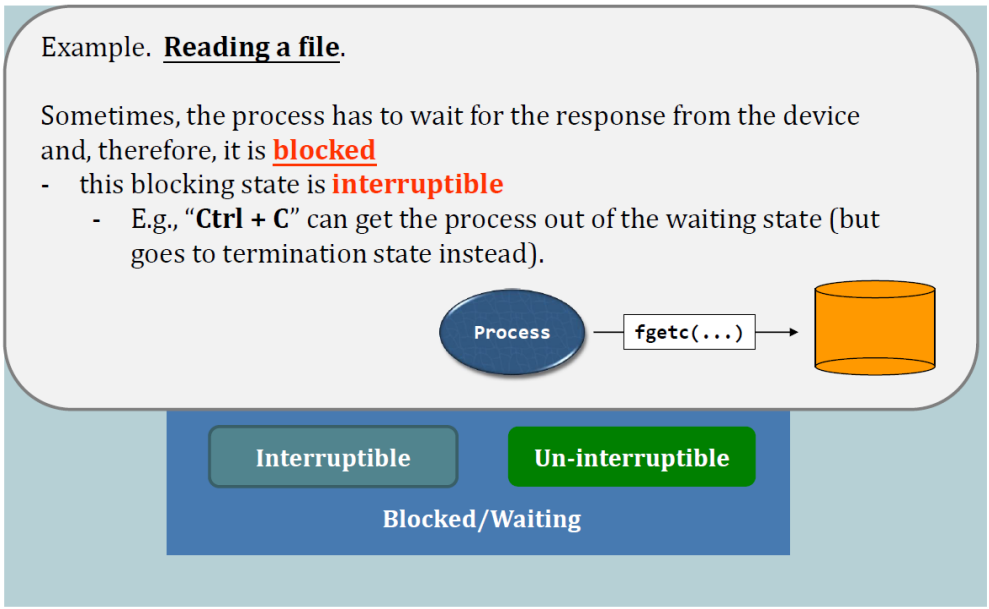
3. Running



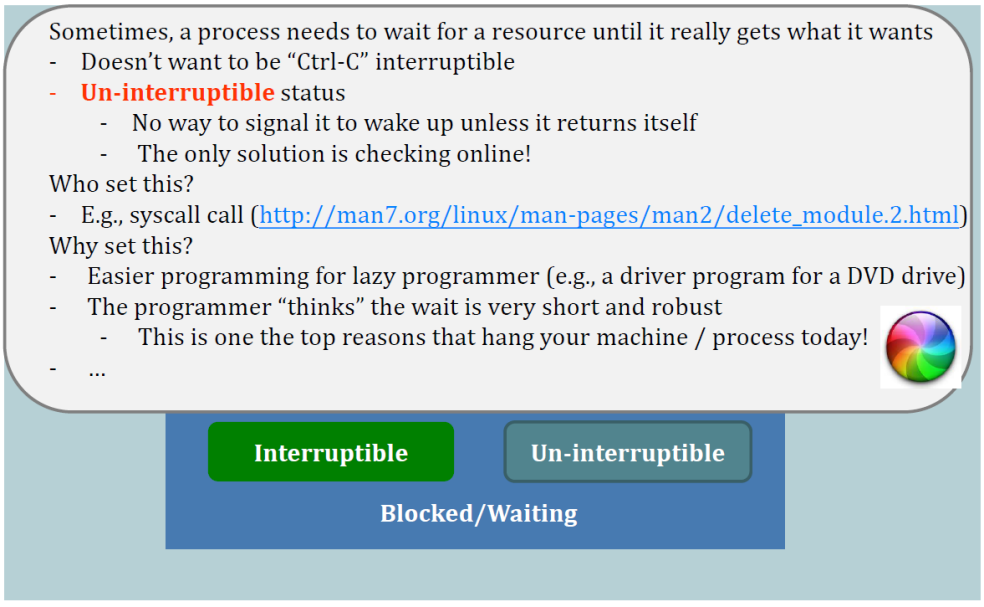
4. Blocked



5. Interruptable waiting

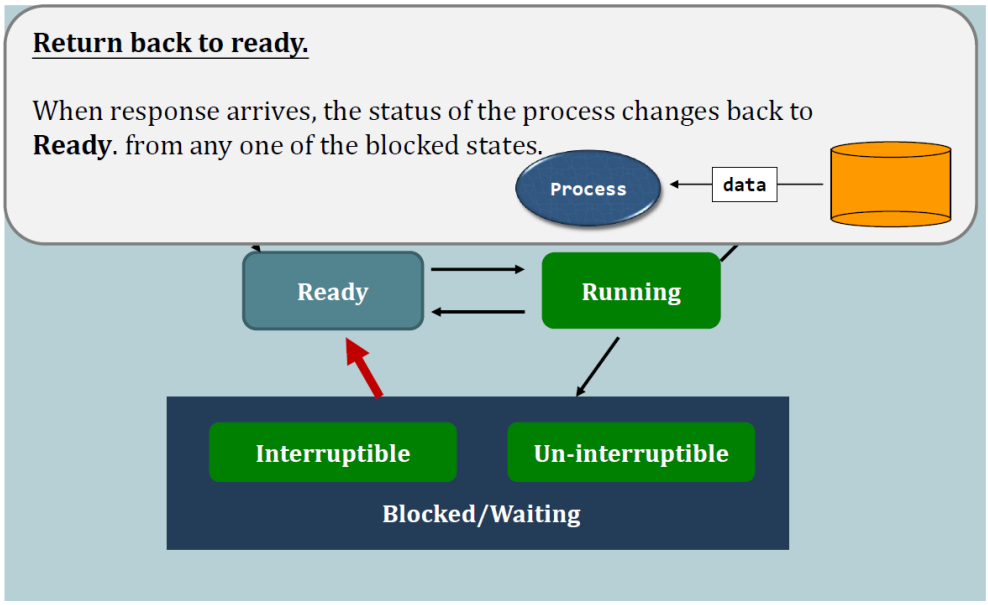


6. Un-interruptable waiting

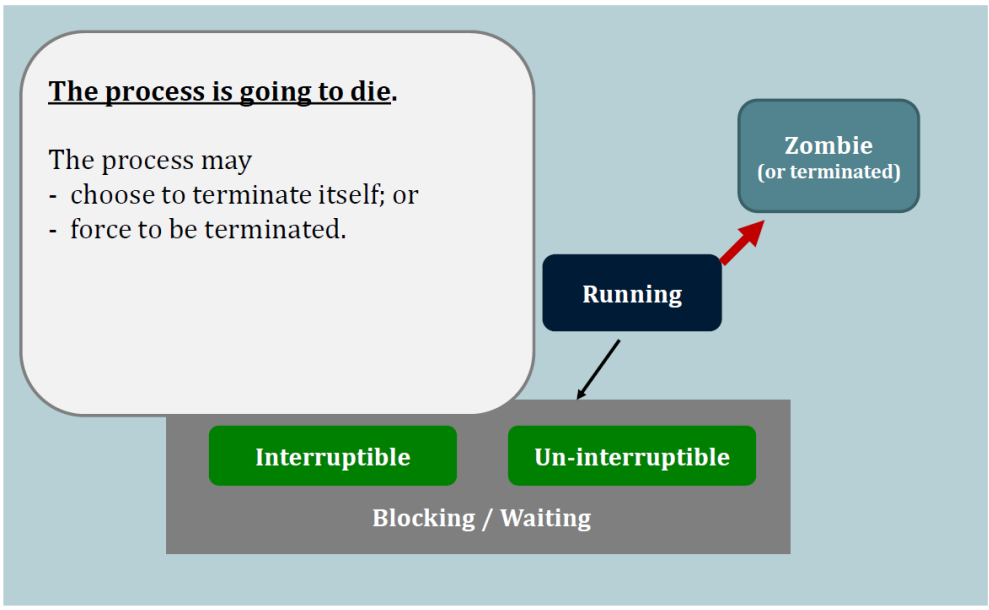


计网的程序里经常碰见，纯贵物，谁设计的抓紧埋了吧

7. Return back to ready



8. Terminated

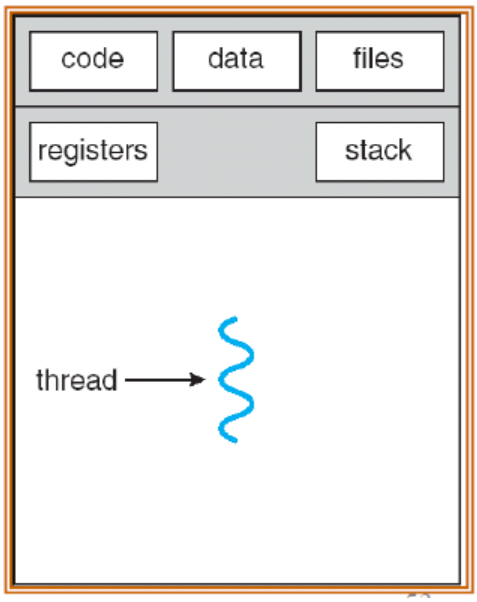


第四章 线程 Thread

4.1 线程的概念

- Heavyweight Process

  A process has a single thread of control



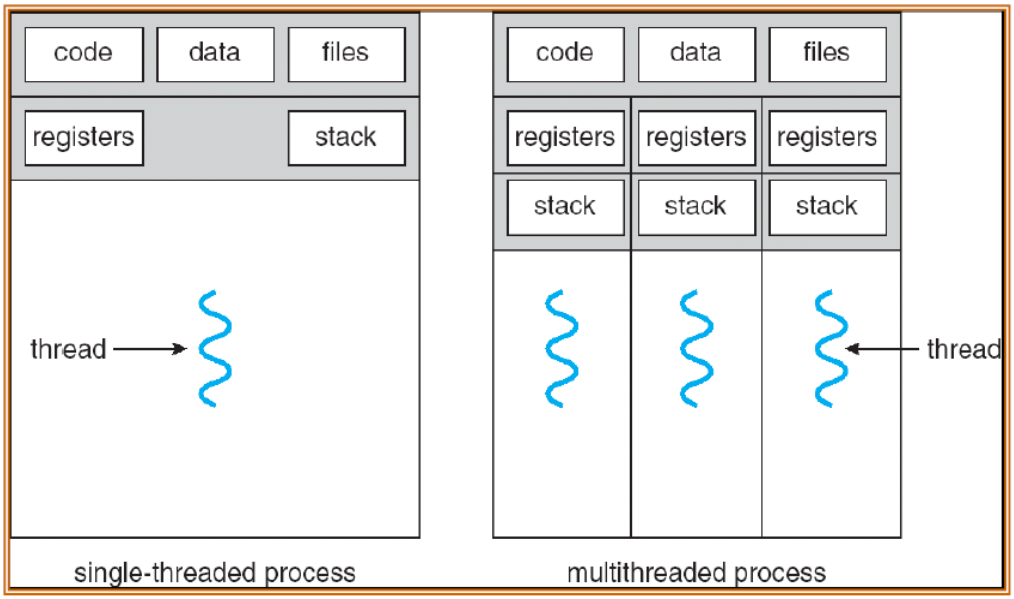
- 线程 Thread

  A sequential execution stream within process

  又称 Lightweight Process

  - Process still contains a single Address Space
  - No protection between threads
- 多线程 Multithreading

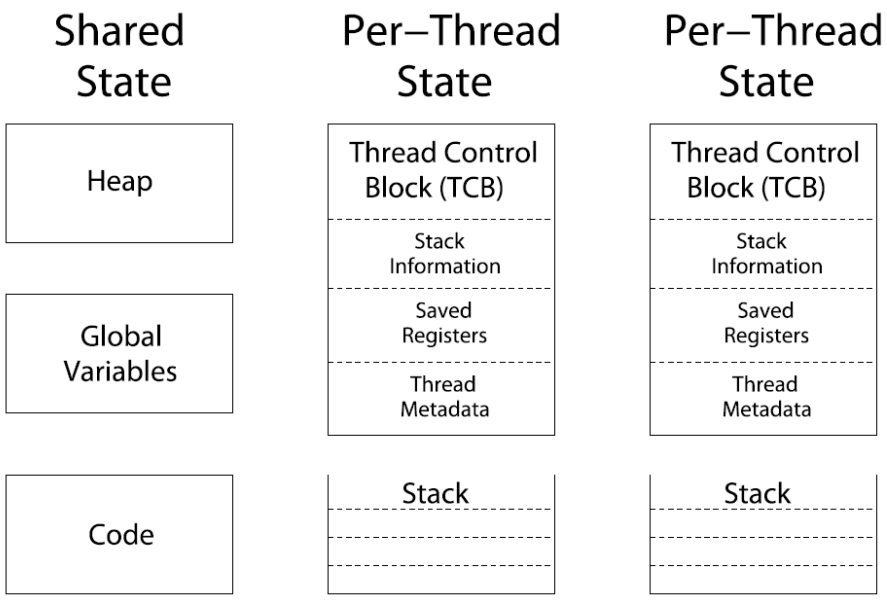
  A single program made up of a number of different concurrent (并发) activities
- 结构图



- Threads encapsulate **concurrency**: "Active" component
- Address spaces encapsulate **protection**: "Passive" part

4.2 线程的组成

- State shared by all threads in process/address space
  - Content of memory (global variables, heap)
  - I/O state (file descriptors, network connections, etc)
- State "private" to each thread
  - Kept in TCB (Thread Control Block)
  - CPU registers (including PC)
  - Execution stack
- 栈
  - Parameters, temporary variables
  - Return PCs are kept while called procedures are executing
  - 回忆计组学的，不多说

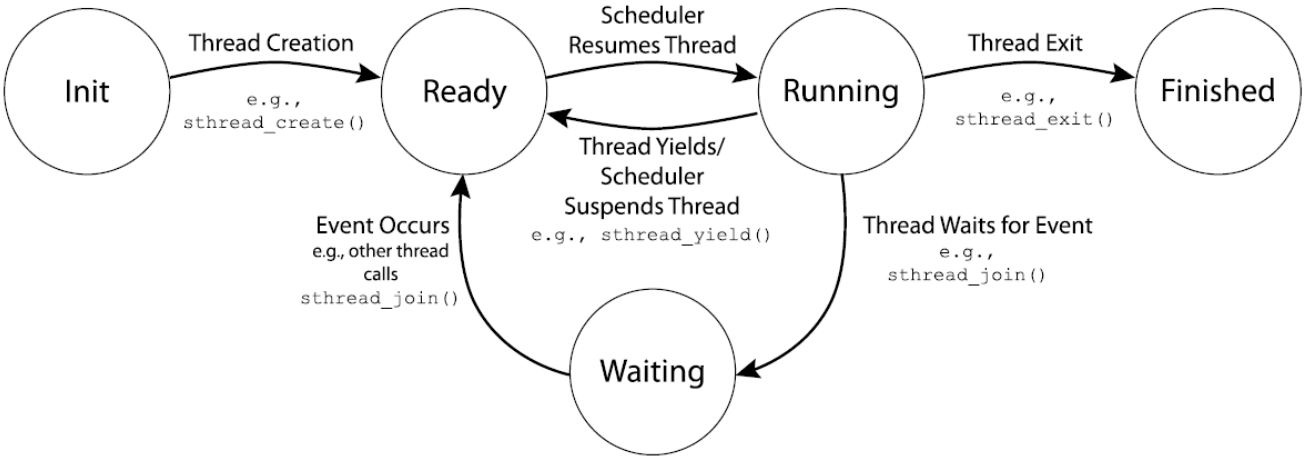


4.3 线程和进程的区别

Process	Thread
Process means any program is in execution.	Thread means segment of a process.
Process takes more time to terminate.	Thread takes less time to terminate.
It takes more time for creation.	It takes less time for creation.
It also takes more time for context switching.	It takes less time for context switching.

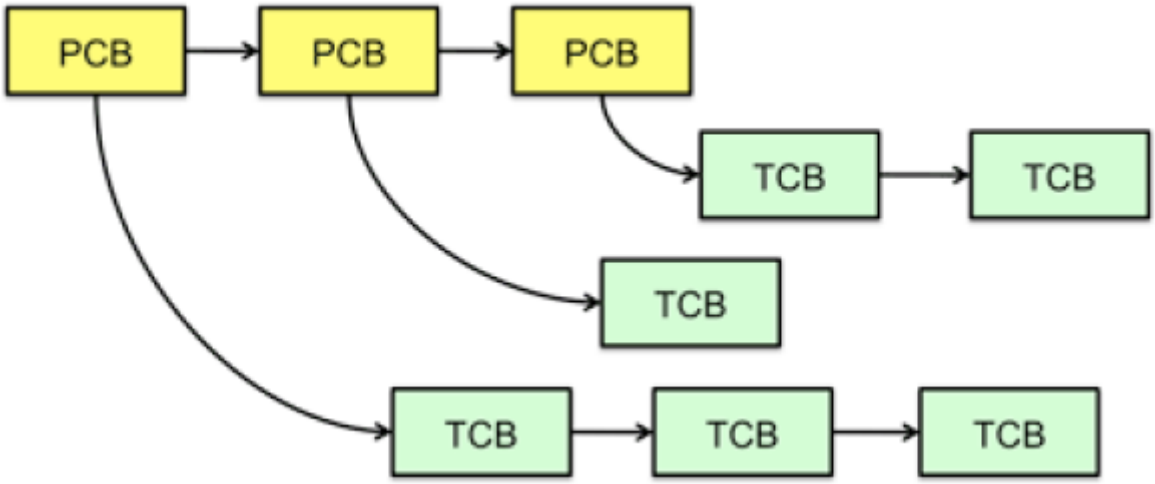
Process	Thread
Process is less efficient in term of communication.	Thread is more efficient in term of communication.
Process consume more resources.	Thread consume less resources.
Process is isolated.	Threads share memory.
Process is called heavy weight process.	Thread is called light weight process.
Process switching uses interface in operating system.	Thread switching does not require to call a operating system and cause an interrupt to the kernel.
If one process is blocked then it will not effect the execution of other process	Second thread in the same task couldnot run, while one server thread is blocked.
Process has its own Process Control Block, Stack and Address Space.	Thread has Parents' PCB, its own Thread Control Block and Stack and common Address space.

4.4 线程的生命周期 Thread Lifecycle



4.5 多线程进程 Multithreaded Process

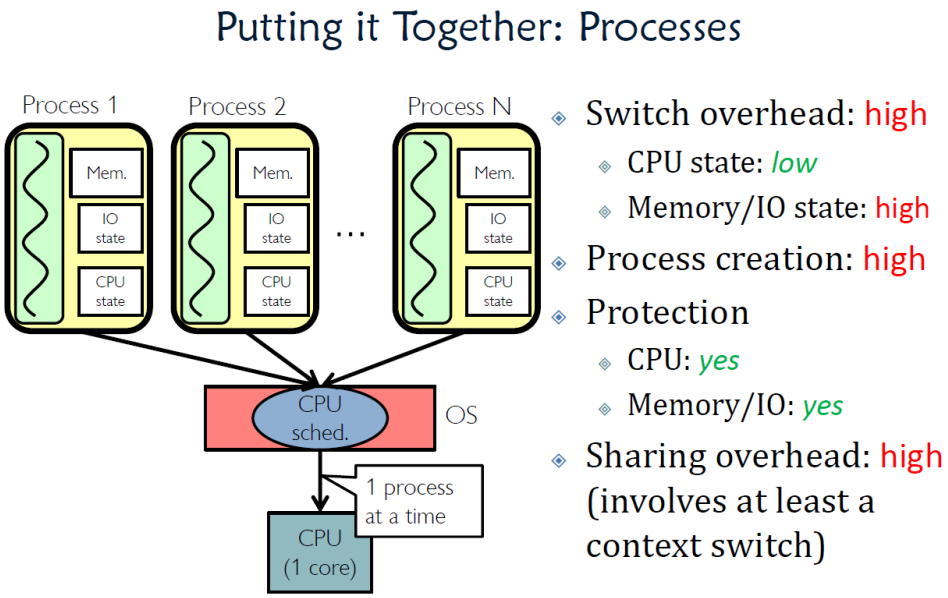
- PCB 指向多个 TCB

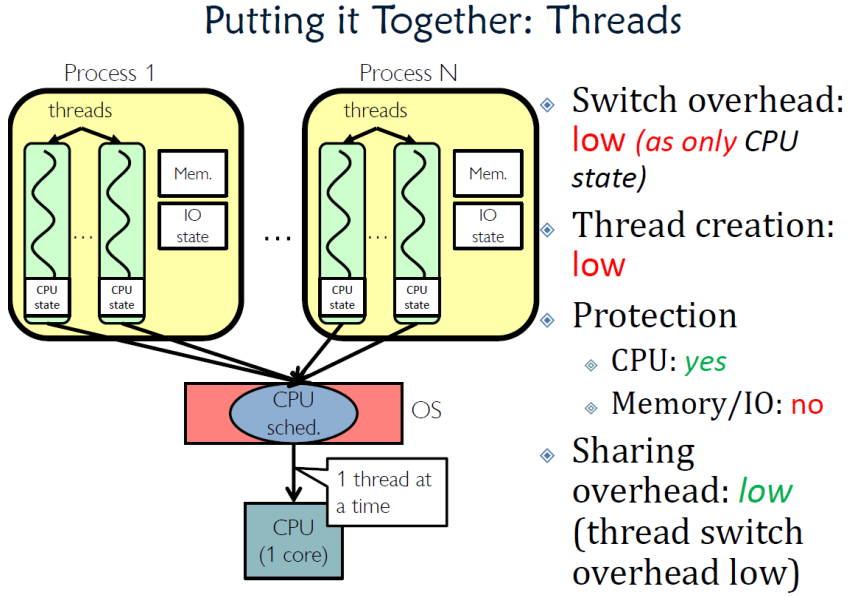


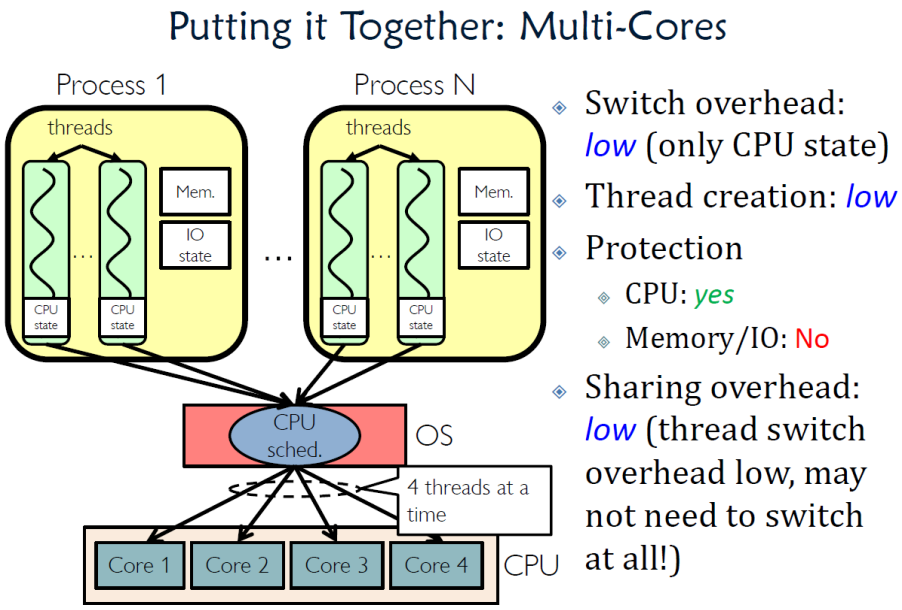
- Switching threads within a block is a simple thread switch
- Switching threads across blocks requires changes to memory and I/O address tables

---

4.6 多线程调度

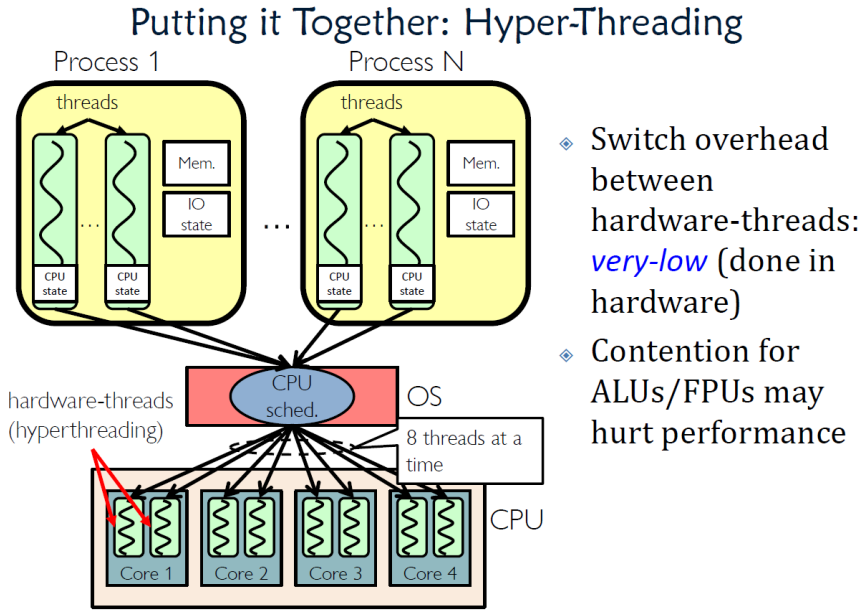






- 超线程 Hyper-Threading

  超线程(hyper-threading)其实就是**同时多线程(simultaneous multi-theading)**，是一项允许一个CPU执行多个控制流的技术。它的原理很简单，就是把一颗CPU当成两颗来用，将一颗具有超线程功能的物理CPU变成两颗逻辑CPU，而逻辑CPU对操作系统来说，跟物理CPU并没有什么区别。因此，操作系统会把工作线程分派给这两颗（逻辑）CPU上去执行，让（多个或单个）应用程序的多个线程，能够同时在同一颗CPU上被执行。注意：两颗逻辑CPU共享单颗物理CPU的所有执行资源。因此，我们可以认为，**超线程技术就是对CPU的虚拟化**



4.7 Multiprocessing, Multithreading and Multiprogramming

- 多进程 Multiprocessing

  Multiple CPUs

  A computer using more than one CPU at a time.

- 多线程 Multithreading

  Multiple threads per Process
- 多道程序设计 Multiprogramming

  Multiple Jobs or Processes

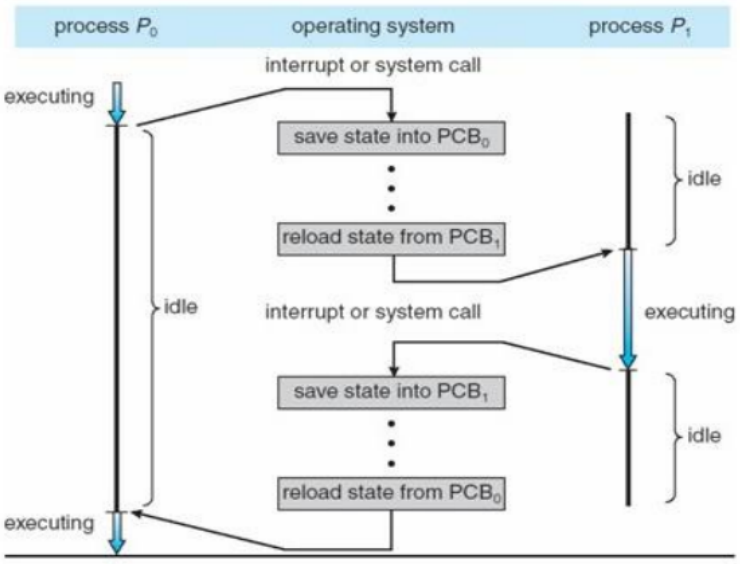
  A computer running more than one program at a time

第五章 进程调度 Process Scheduling

5.1 基本概念

5.1.1 上下文切换 Context Switch

- 切换 CPU 到另一个进程需要**保存**当前进程状态和**恢复**另一个进程的状态，这个任务称为上下文切换
- 当进行上下文切换时，内核会将旧进程的状态保存在其 **PCB** 中，然后加载经调度而要执行的新进程的上下文
- 上下文切换是纯粹的时间开销 (Overhead)，因为 CPU 在此期间没有做任何有用工作
- 上下文切换非常耗时
- 什么时候 Context Switch
    - Get blocked，比如调用 `wait(), sleep()` 等
    - System Call
    - A signal arrives
    - An interrupt arrives
    - 时间片用完
    - 被抢占



5.1.2 调度队列

- 作业队列 Job Queue

  包含所有进程
- 就绪队列 Ready Queue

  等待运行的进程

  PCB 构成的链表

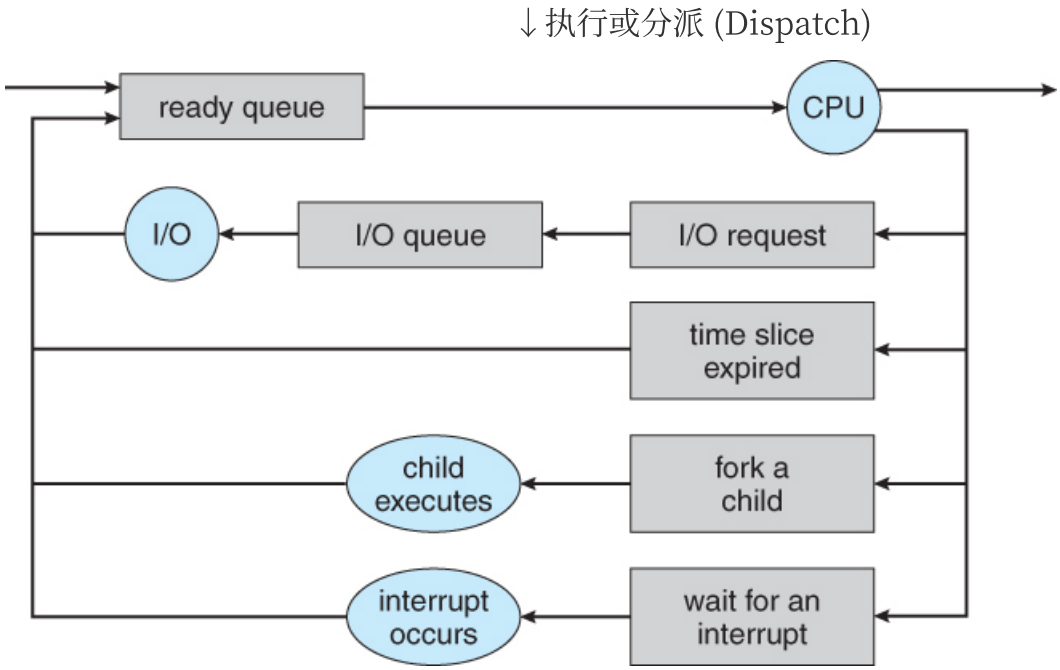
- 设备队列 Device Queue

  等待使用该 IO 设备的进程队列

  每个设备都有

- 队列图 Queueing Diagram

  圆圈代表服务队列的资源， 箭头代表系统内的进程流向



5.1.3 调度程序 Scheduler

- 缓冲池

  通常来说，对于批处理系统，提交的进程多于可执行的，这些进程被保存到大容量存储设备（如磁盘）的缓冲池，以便以后执行

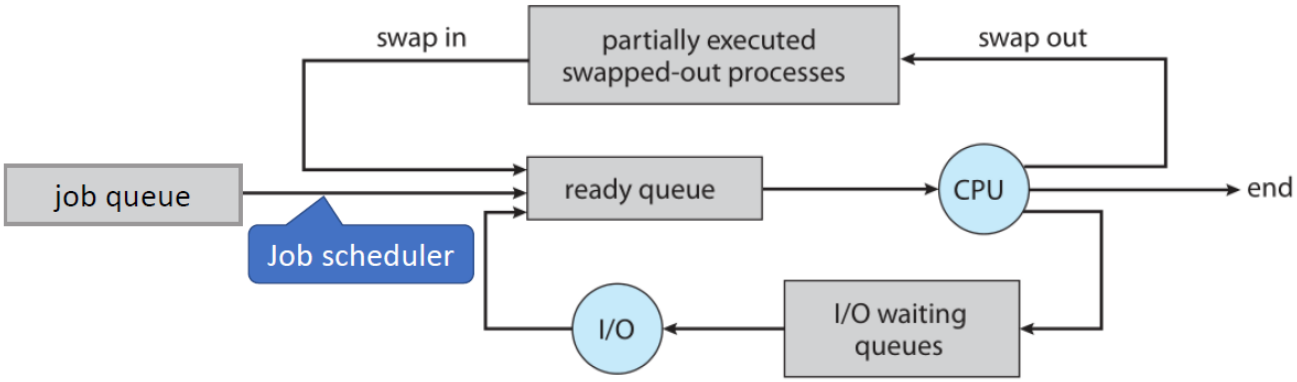
- 调度程序 (调度器)

调度器	别名	作用
长期调度程序 Long-term Scheduler	作业调度程序 Job Scheduler	从缓冲池中选择进程加载到内存
短期调度程序 Short-term Scheduler	CPU 调度程序	从 Ready Queue 中选择进程并分配 CPU
中期调度程序 Medium-term Scheduler		进程交换

- 进程分类

中文	英文	特点
I/O 密集型进程	I/O Bounded Process	执行 I/O 比执行计算耗时
CPU 密集型进程	CPU Bounded Process	很少I/O, 执行计算用时长

  长期调度程序需要选择这两种进程的合理组合才能最大化 CPU 和 IO 设备的利用



5.1.4 Dispatcher

Dispatcher 是一个模块，用来将 CPU 控制交给由 CPU 调度程序选择的进程

- 功能
  - 切换上下文
  - 切换到用户模式
  - 跳转到用户程序的合适位置，以便重新启动程序
- 调度延迟 Dispatch Latency

  Dispatcher 停止一个进程而启动另一个进程所需的时间
- Dispatcher 和 Scheduler 的区别

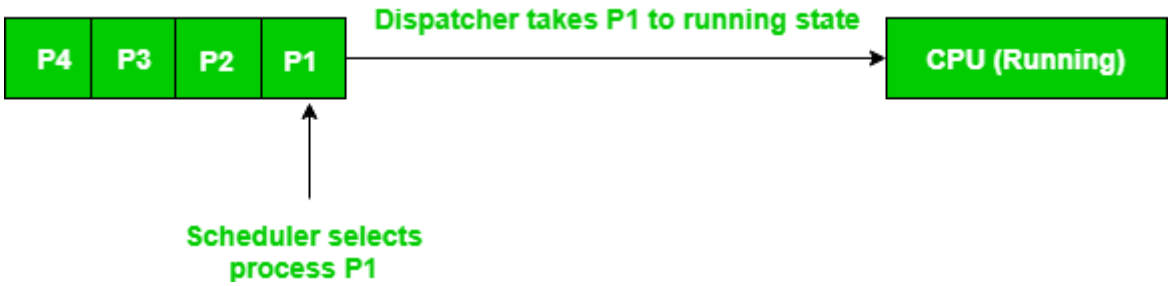
  中文书把 dispatcher 也翻译成调度程序，我真想一拳干碎你的眼镜

  https://www.differencebetween.com/difference-between-scheduler-and-vs-dispatcher

  https://www.geeksforgeeks.org/difference-between-dispatcher-and-scheduler

> The **key difference** between scheduler and dispatcher is that **the scheduler selects a process out of several processes to be executed while the dispatcher allocates the CPU for the selected process by the scheduler.**

Scheduler vs Dispatcher	
A scheduler is special system software that handles process scheduling by selecting the process to execute.	The dispatcher is the module that gives control of the CPU to the process selected by the short-term scheduler.
Types	
There are three types of schedulers known as; • long-term scheduler, • short-term scheduler • medium term scheduler.	There is no categorization for a dispatcher.
Main Tasks	
The **long-term scheduler** selects the process from the job queue and brings it to the ready queue. The **short term scheduler** selects a process in the ready queue. The **medium scheduler** carries out the swap in, swap out of the process.	The dispatcher allocates the CPU to the process selected by the short-term scheduler.



Properties	DISPATCHER	SCHEDULER
Definition	Dispatcher is a module that gives control of CPU to the process selected by short term scheduler	Scheduler is something which selects a process among various processes
Types	There are no different types in dispatcher. It is just a code segment.	There are 3 types of scheduler i.e. Long-term, Short-term, Medium-term
Dependency	Working of dispatcher is dependent on scheduler. Means dispatcher have to wait until scheduler selects a process.	Scheduler works independently. It works immediately when needed
Algorithm	Dispatcher has no specific algorithm for its implementation	Scheduler works on various algorithm such as FCFS, SJF, RR etc.
Time Taken	The time taken by dispatcher is called dispatch latency.	Time taken by scheduler is usually negligible. Hence we neglect it.
Functions	Dispatcher is also responsible for: Context Switching, Switch to user mode, Jumping to proper location when process again restarted	The only work of scheduler is selection of processes.

5.2 调度准则

- CPU 使用率

  应该使 CPU 尽可能忙碌
- 吞吐量

  一个时间单元内进程完成的数量
- 周转时间

从进程提交到完成的时间段称为周转时间 (Turnaround Time)

- 等待时间

  在就绪队列中等待所花时间之和

- 响应时间

  从提交请求到产生第一响应的时间

- Number of Context Switches (from 课件)

  尽可能少做上下文切换

5.3 调度算法 Scheduling Algorithm

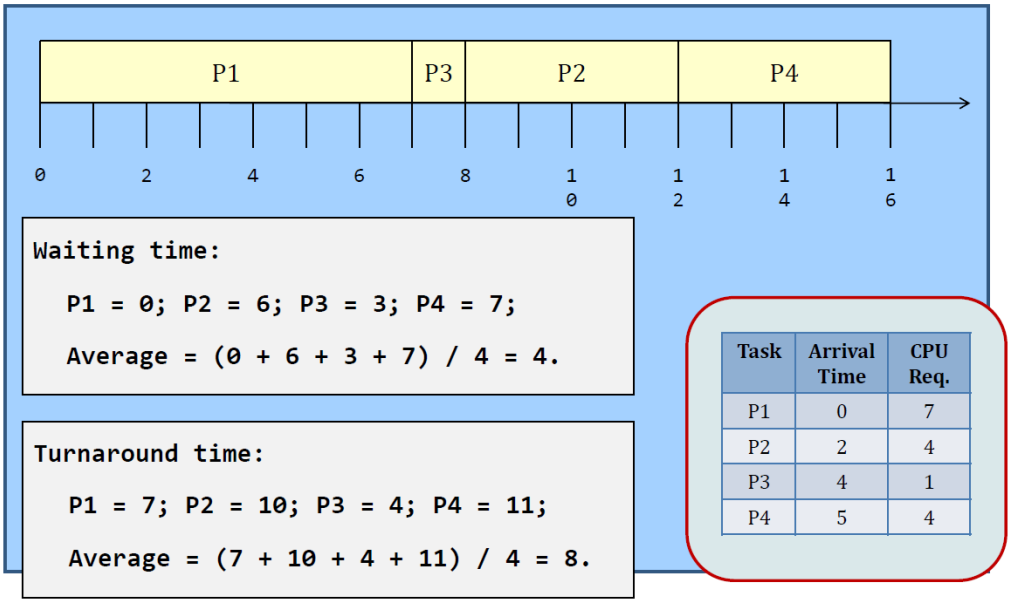
5.3.1 先到先服务调度 First-Come-First-Served (FCFS)

字面意思

5.3.2 最短作业优先调度 Shortest-Job-First (SJF)

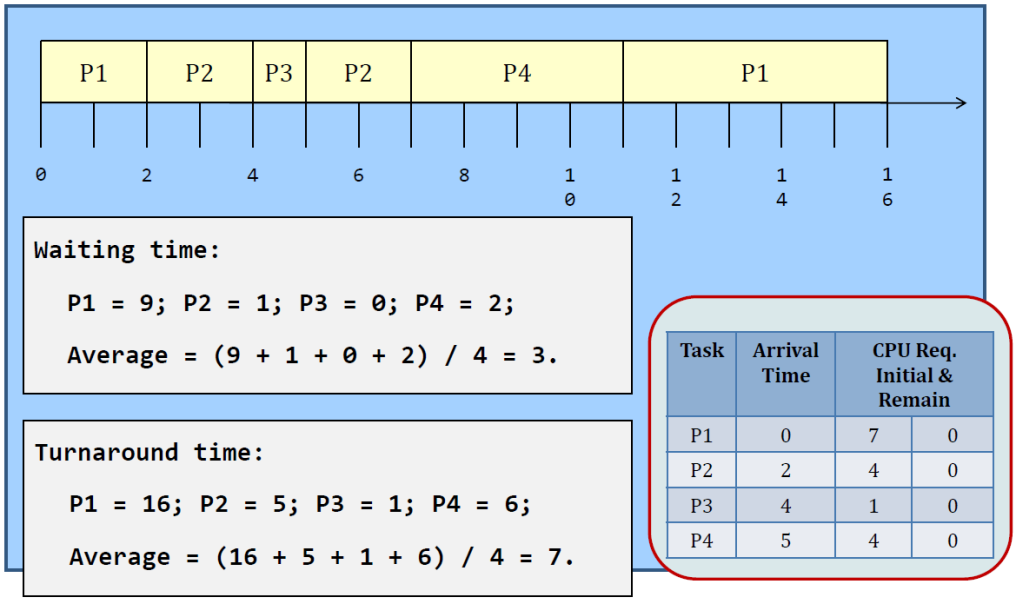
- 选择最短 CPU 执行时间的进程
- 相同，可以使用 FCFS 规则选择
- 又称最短下次 CPU执行 (Shortest-Next-CPU-Burst) 算法
- 可能造成 starvation

5.3.2.1 非抢占 (Non-Preemptive) SJF



5.3.2.2 抢占 SJF

- 最短剩余时间优先 (Shortest-Remaining-Time-First)



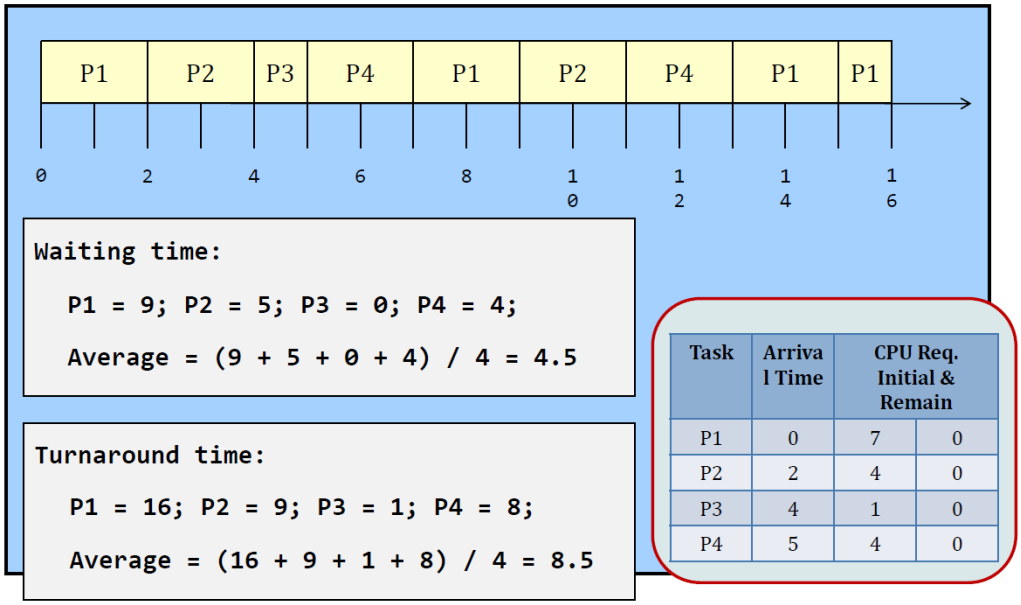
Task	Arrival Time	CPU Req. Initial & Remain	
P1	0	7	0
P2	2	4	0
P3	4	1	0
P4	5	4	0

- 缺点：上下文切换多

5.3.3 轮转调度 Round Robin (RR)

- 每个进程都有一个时间量 (Time Quantum) 或时间片 (Time Slice)

  通常 10~100ms
- 当时间片用完时，该进程就会释放 CPU (相当于抢占)
- 调度程序选择下一个时间片 > 0 的进程
- 如果所有进程都用完了时间片，它们的时间片同时被 recharge 到初始值
- 就绪队列为循环队列，进程被依次执行
  - 刚执行完的进队尾
  - 新来的进队尾
  - 新来的进程不会触发新的 Schedule，就按队列顺序来



Task	Arrival Time	CPU Req. Initial & Remain	
P1	0	7	0
P2	2	4	0
P3	4	1	0
P4	5	4	0

- 缺点：性能较差
- 优点：公平

5.3.4 优先级调度 Priority Scheduling

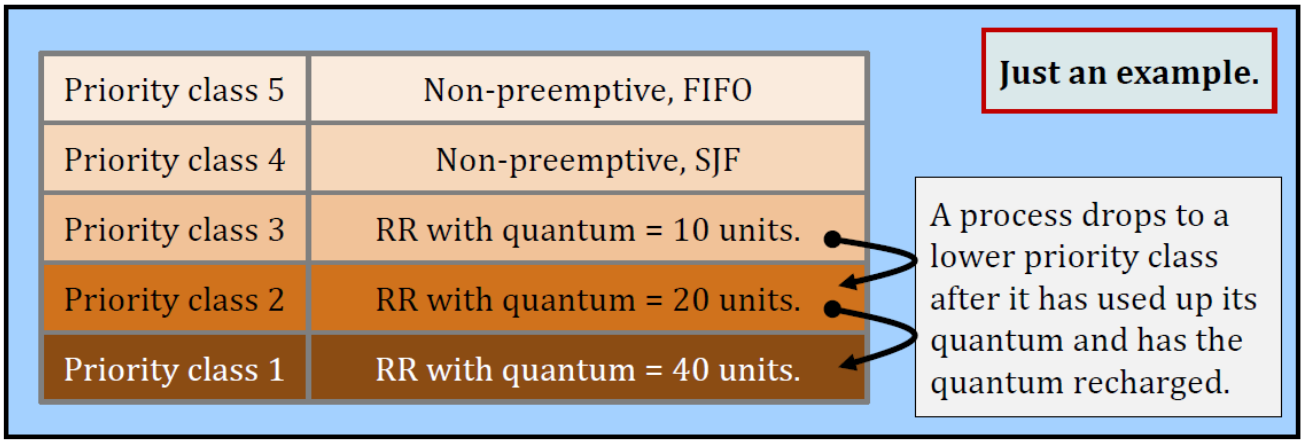
- 每个进程都有一个优先级
- 调度程序根据优先级选择进程
- 优先队列
- 分类

2 Classes	
Static priority	**Dynamic priority**
Every task is given a fixed priority.	Every task is given an initial priority.
The priority is **fixed** throughout the life of the task.	The priority is **changing** throughout the life of the task.

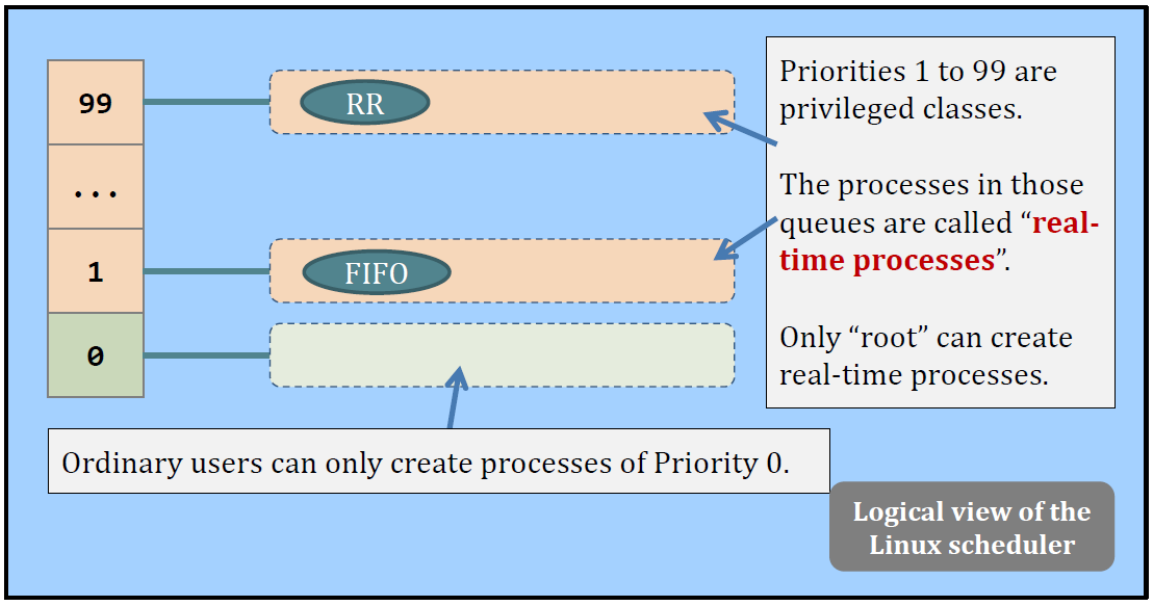
- 新进程到来时，重新 schedule (这里会发生抢占)
- 如果当前进程被抢占，它先出队再入队

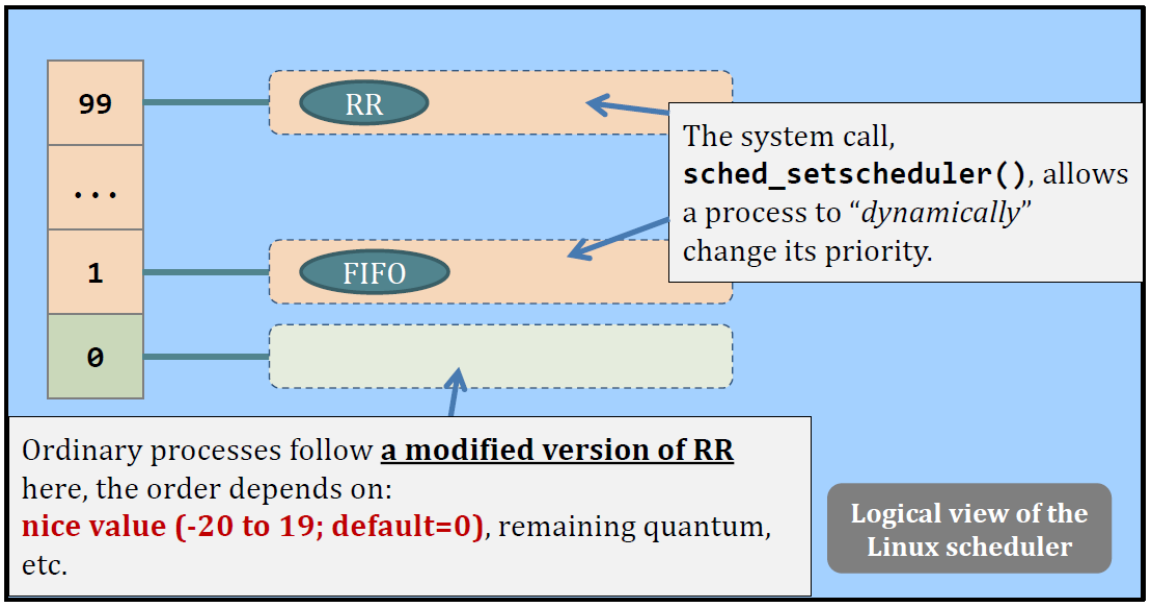
5.3.4.1 Multiple Queue Priority Scheduling

- 依然是 priority scheduler
- 每个优先级有不同的调度方式
- 可以是静态优先级和动态优先级混合



- Linux Scheduler

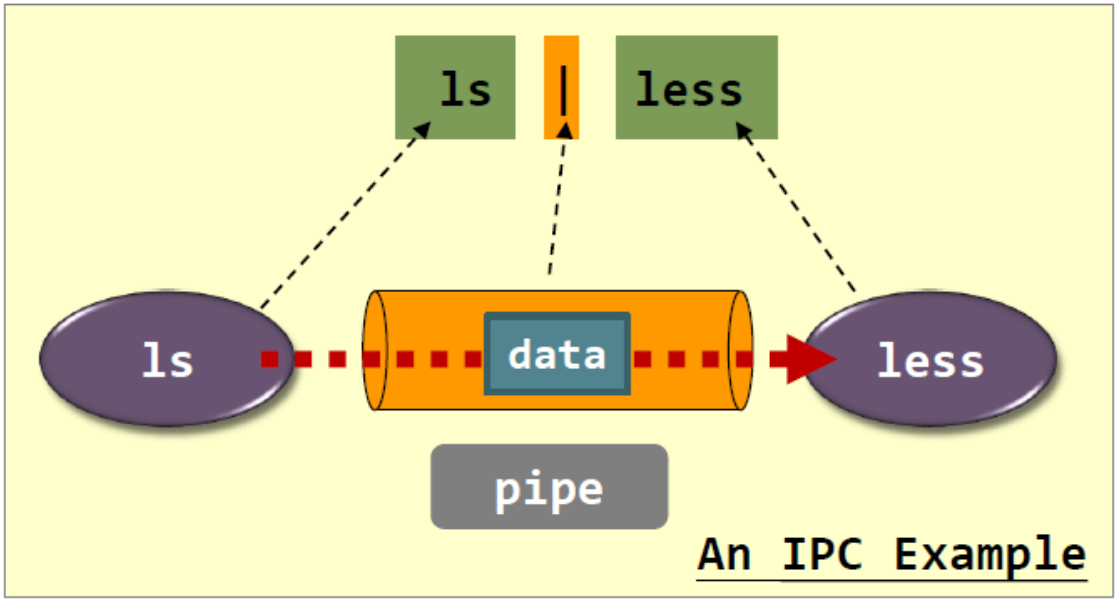

99
...
1
0
RR
FIFO
Priorities 1 to 99 are privileged classes.
The processes in those queues are called "real-time processes".
Only "root" can create real-time processes.
Ordinary users can only create processes of Priority 0.
Logical view of the Linux scheduler


99
...
1
0
RR
FIFO
The system call, sched_setscheduler(), allows a process to "dynamically" change its priority.
Ordinary processes follow a modified version of RR here, the order depends on:
nice value (-20 to 19; default=0), remaining quantum, etc.
Logical view of the Linux scheduler

第六章 同步 Synchronization

6.1 进程间通信 Inter-Process Communication (IPC)

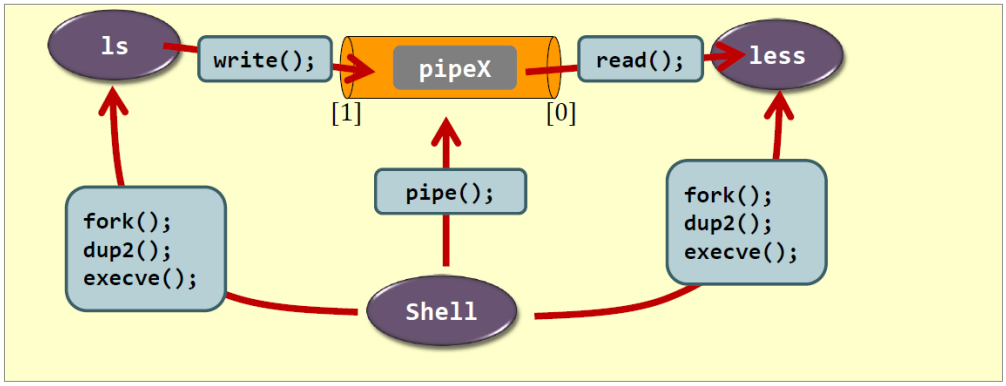
- 匿名管道 Pipe
  - 单向 Unidirectional
  - 匿名管道只能在祖先相同的进程之间建立

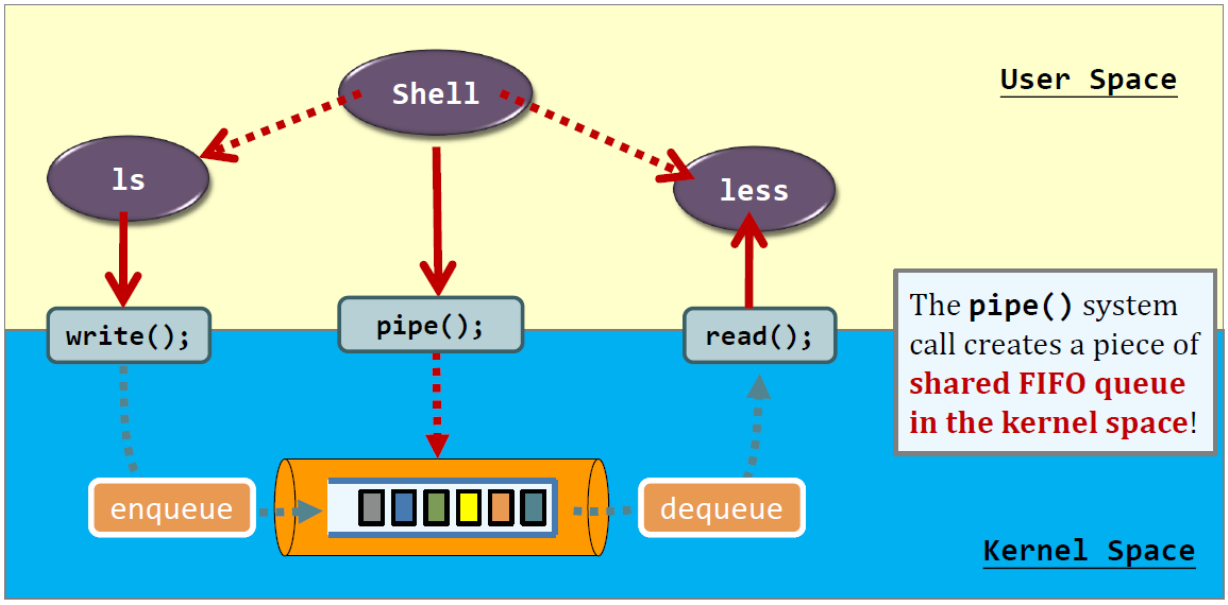


- 信号 Signal
  - More kernel-level
  - Limited (SIGKILL, SIGCHLD, ...)
- 例: `ls | less`

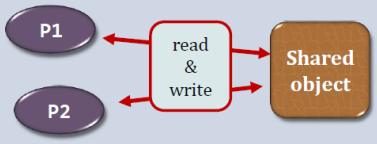
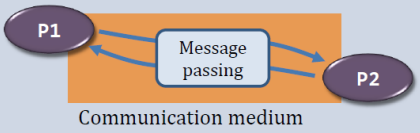
```
fork();
if (pid==0) { // child; "ls"
  //dup2: replace "ls" default stdout
    by the write end of the pipe
  dup2(pipeX[1], STDOUT_FILENO);
  execlp("ls", "ls", NULL);
} else … //parent; "less"
```

In UNIX*, "everything is a file"
- Every resource that can read/write is represented as a file. E.g.,
  - Network, Disk, Keyboard
- A "file" is indexed by a number called *file descriptor*





- IPC Models

Shared Objects	Message Passing
	
• shared files (on disk; slow) • pipes (restricted, but OS takes care of synchronization for you) • shared memory (primitive, general, but synchronization is on you) • shared address space (threading)	• socket programming • message passing interface (MPI) library for computing clusters.
- Usually single-node communication - More efficient - Need to take great care of synchronization because of sharing the same object	- Usually multi-node communication - Less efficient - Less troublesome in synchronization - But need to care of other faults (e.g., what if a network link is broken?)

- User space 里的所有东西都不能 share，所以 pipe 之类的都在 kernel 里

6.2 临界区 Critical Section

6.2.1 竞争条件 Race Condition

- 多个进程并发访问和操作同一数据并且执行结果与特定访问顺序有关，称为竞争条件 (Race Condition)
- Shared Object + Multiple Process + Concurrency

6.2.2 临界区问题 Critical Section Problem

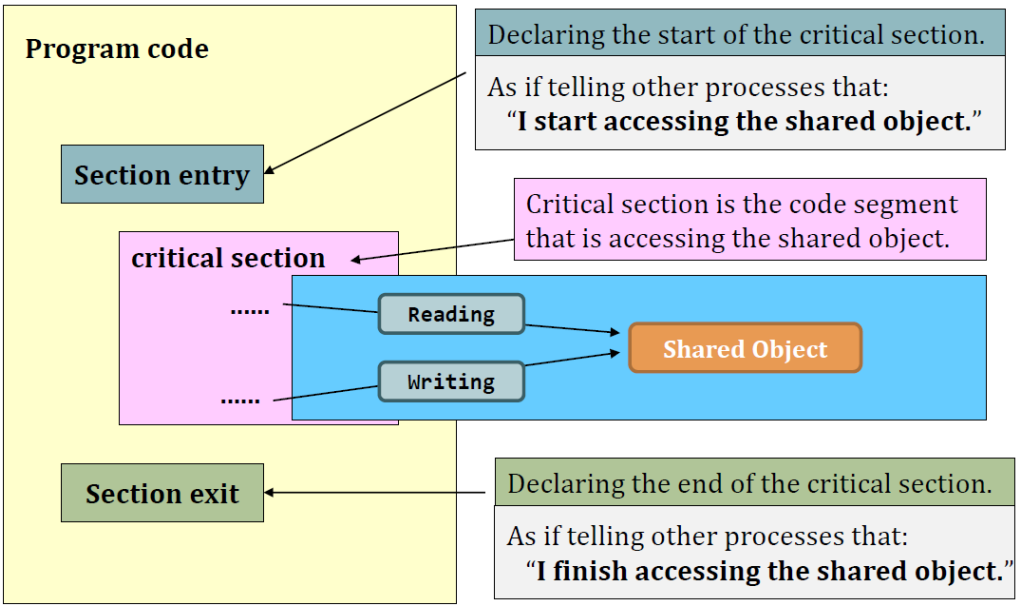
- 临界区 Critical Section

  每个进程有一段代码，进程在执行该段代码时可能修改公共变量、更新一个表、写一个文件等

- 进入区 Entry Section

进入临界区前，请求许可的代码段

- 退出区 Exit Section
- 剩余区 Remainder Section



- 临界区问题 (Critical-Section Problem) 指设计一个协议以便协作进程，使得没有两个进程可以在它们的临界区内同时执行
- 临界区要尽可能紧凑
- 一个临界区里可以访问多个 shared object
- 重点是进入区和退出区的实现

6.2.3 临界区问题的要求

1. 互斥 Mutual Exclusion

   如果一个进程在其临界区内执行，那么其他进程都不能在临界区内执行

2. 进步 Progress

   如果没有进程在临界区内执行，并且有进程需要进入临界区，那么只有那些不在剩余区内的进程可以参加选择，以便确定谁下次进入临界区，而且这种选择不能无限推迟

   别让执行临界区的进程空着，除非大家都不想进临界区

3. 有限等待 Bounded Waiting

   从一个进程做出进入临界区的请求直到这个请求允许为止，其他进程允许进入其临界区的次数有上限

   别让一个进程等一辈子

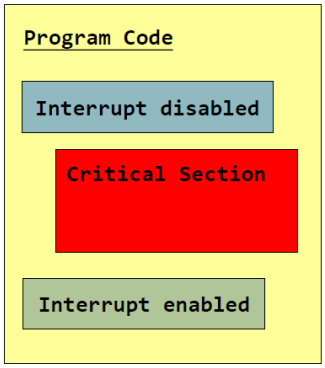
6.3 临界区问题的解决方案 Solutions for Critical Section Problem

- Lock-based
  - Spin-based Lock
    - Basic spinning
    - Peterson's solution
  - Sleep-based Lock
    - POSIX semaphore
    - `pthread_mutex_lock`
- Lock-free

6.3.1 硬件同步 (×) Hardware Synchronization

- 禁止中断

  - **Aim**
    - To **disable context switching** when the process is inside the critical section.
  - **Effect**
    - When a process is in its critical section, no other processes could be able to run.
  - **Correctness?**
    - **Uni-core: Correct but not permissible**
      - at user space: what if one writes a CS that loops infinitely and the other process (e.g., the shell) never gets the context switch back to kill it?
      - At kernel level: yes, correct and permissible
    - **Multi-core: Incorrect**
      - if there is another core modifying the shared object in the memory (unless you disable interrupts on all cores!!!!)



- 单核

  正确，但是不能接受

  如果有个进程在 CS 里写个死循环就全卡这了

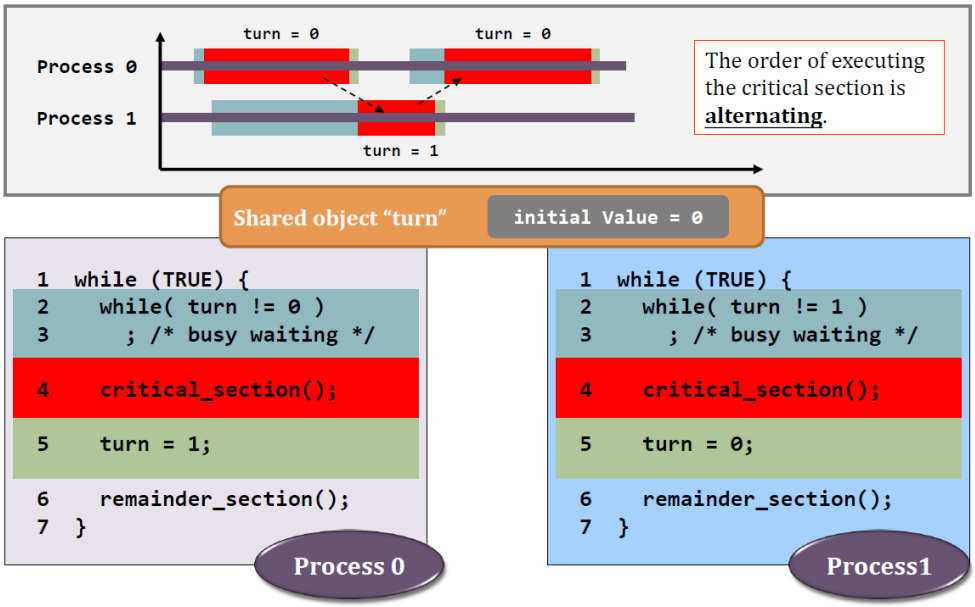
- 多核

  不正确，除非把所有核的中断全都禁止

6.3.2 基本自旋锁 (×) Basic Spin Lock

- 原理

  设置一个公共变量 `turn` 来决定哪个进程可以进 CS

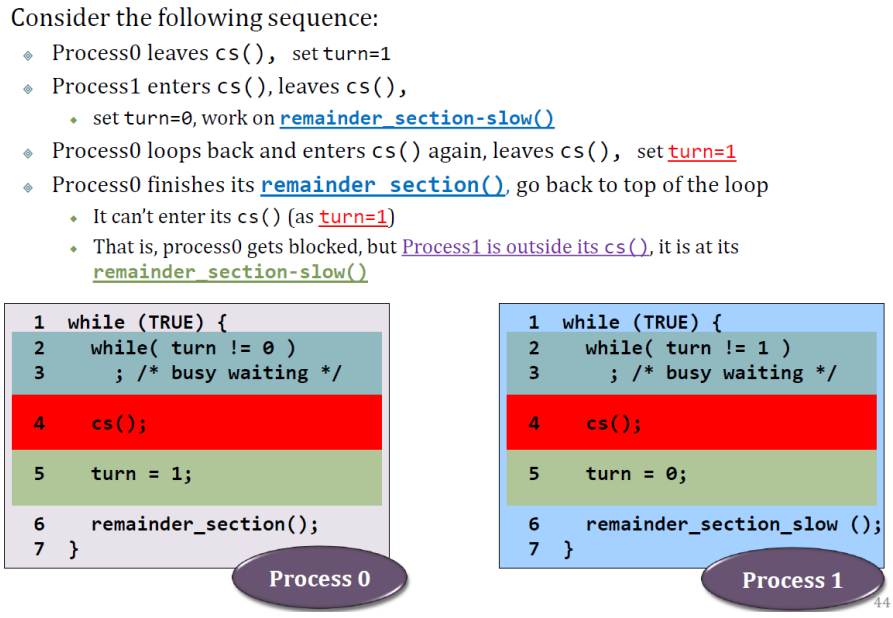


- 太浪费 CPU
- 违反 Progress

  多个进程一定是交替执行

  如果一个进程不打算进 CS 但是另一个进程交出了权限，那就要等很长时间 (**no progress**)

  Example: 这种情况下没人在 CS 里。不能让执行 CS 的进程空着



6.3.3 Peterson's Solution

- 在 `turn` 的基础上新加一个布尔数组 `interested`
  - If I don't show interest

    I let you all go
  - If we both show interest

    Take turns

```c
int turn;
int interested[2] = {false, false};

void lock(int process) {
    int other = 1 - process;
    interested[process] = true;
    turn = other;
    while (turn == other && interested[other]); // busy waiting
}

void unlock(int process) {
    interested[process] = false;
}
```

- 会产生优先级翻转问题 (Priority Invasion)

  优先级 $A < B < C$

  1. A 获得锁, C 来了, C 申请锁
  2. 按理来说 C 应该抢占 A, 但是锁在 A 手里, C 就只能等待
  3. B 不要锁，所以 B 可以被调度上去
  4. 明明 B 优先级低，却比 C 先执行
- 为什么 `turn=other` 不是 `turn=process`

  我们假设是这样

  ```
  turn=自己;
  while (turn==自己 && interested[别人]);
  ```

  如果现在有三个进程 P_1, P_2, P_3

  1. 我们脸比较黑，这三个进程经过调度，都该执行 `turn=自己` 这一行
  2. 那么最终 `turn` 是几，就取决于调度器了
  3. 假设调度器就是按 P_1, P_2, P_3 的顺序调度的，那么最后 `turn=3`
  4. 现在我们检查 `while` 的条件
     - 对于 P_1, `turn=3`, 前半句不成立, 不需要 wait
     - 对于 P_2, `turn=3`, 前半句不成立, 不需要 wait
     - 对于 P_3, 条件成立, 需要 wait
  5. 那么现在 P_1, P_2 都被许可进入 CS，违反了**互斥**原则

  正确是这样:

  ```
  turn=别人;
  while (turn==别人们 && interested[别人们]);
  // while ((turn==x && interested[x]) || (turn==y &&
  interested[y]))
  // while (turn≠自己 && interested[别人们])
  ```

  还是这个例子，我们假设 `turn` 的赋值是 1 给 2，2 给 3，3 给 1

1. 还是都执行到 `turn=别人` 这一行，还是按 123 的顺序调度的
2. 那最终 `turn=1`
3. 检查 `while` 的条件，只有 P_1 可以进 CS

6.3.4 信号量 Semaphore

- 信号量是一个 Structure
  - 一个 `int`，表示剩余多少资源可用
  - 一个等待队列

```
typedef struct {
    int value;
    struct process *list;
} semaphore;
```

- Wait (P 操作)

```
wait(semaphore *s) {
    s→value--;
    if (s→value<0) {
        add this process to s→list;
        block();
    }
}
```

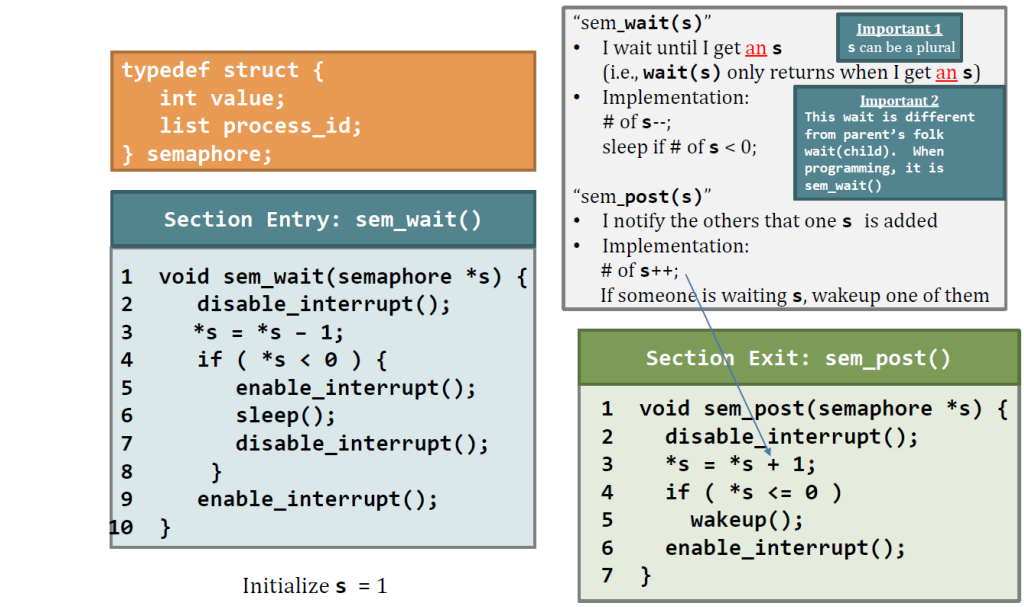
- Post (V 操作)

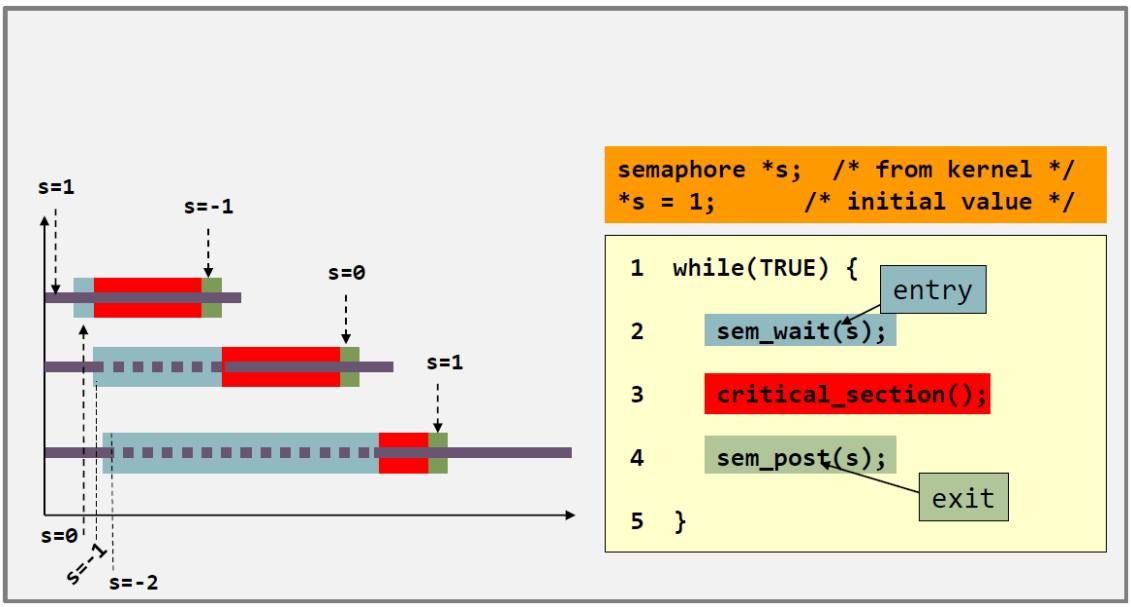
```
post(semaphore *s) {
    s→value++;
    if (s→value⩽0){
        remove a process p from s→list;
        wakeup(p);
    }
}
```

- 分类
  - 二进制信号量 Binary Semaphore

    只能 0 或 1
  - 计数信号量 Counting Semaphore

    可以 > 1



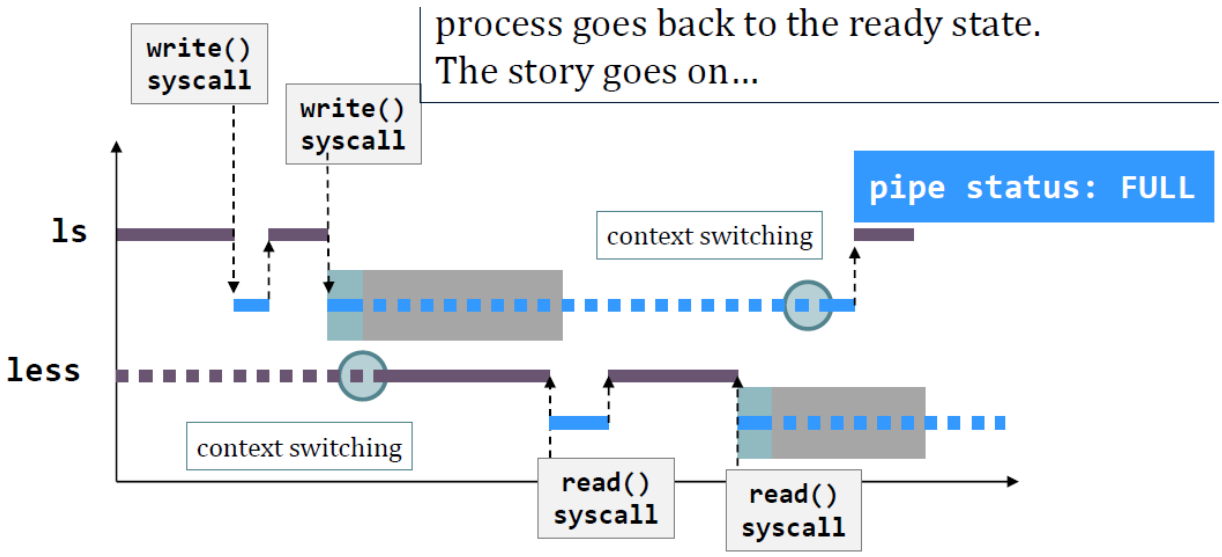


6.4 经典同步问题

6.4.1 有界缓冲问题 Bounded-Buffer Problem

- 又称生产者-消费者问题 (Producer-Consumer Problem)
- 组成
  1. Bounded Buffer
     - Shared object
     - Limited size
     - Queue
  2. Producer Process
     - Produce a unit of data and writes that piece of data to the tail of the buffer at one time
  3. Consumer Process
     - Remove a unit of data from the head of the buffer at one time

- 要求
    1. Producer
        - 当 producer 向 buffer 里放入数据，但是 buffer 已经满的时候，他需要 wait
        - 放入数据后，通知 consumer (wake up)
    2. Consumer
        - 当 consumer 要消费数据，但是 buffer 是空的，他需要 wait
        - 消费数据之后，通知 producer (wake up)
- 例子



- Semaphore 实现

```
semaphore mutex=1;
semaphore avail=N;
semaphore fill=0;

void producer() {
    int item;

    while (true) {
        item=produce_item();

        wait(&avail);
        wait(&mutex);

        insert_item(item);

        post(&mutex);
        post(&fill);
    }
}

void consumer() {
    int item;

    while (true){
```

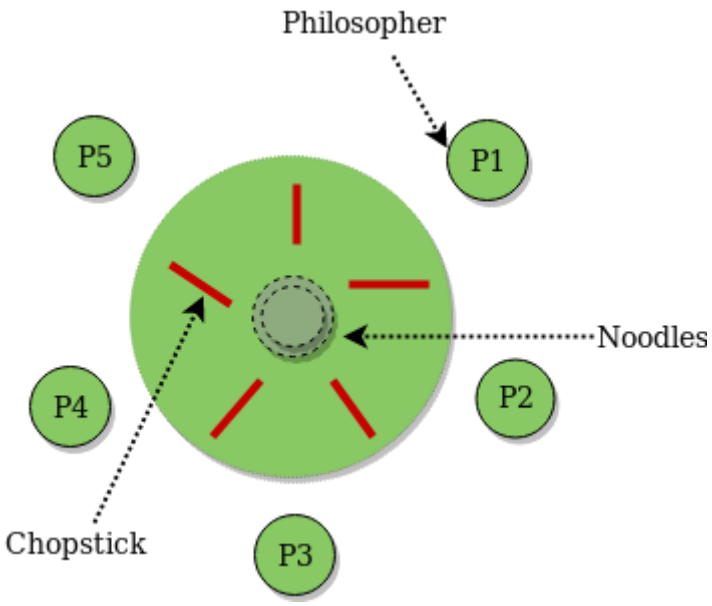
```
25          wait(&fill);
26          wait(&mutex);
27
28          item=remove_item();
29
30          post(&mutex);
31          post(&avail);
32      }
33  }
```

6.4.2 读者-作者问题 Reader-Writer Problem

- 要求
  - 任何数量的 reader 都可以同时 read
  - 同时只能有一个 writer 写
  - 如果有 writer 在写，那么所有 reader 都不能读

6.4.3 哲学家就餐问题 Dining-Philosophers Problem

- 问题描述
  - 有 5 个哲学家，5 根筷子，1 盘面条



  - 每个哲学家有两个可能的动作
    - Think
    - Eat
  - 如果一个哲学家要吃面条，他必须同时获得左右两根筷子
  - 拿起来的筷子不会被别人抢
- 要求

  设计一个 Protocol，保证所有哲学家
  - 不会饿死
  - 不会死锁
- 解决方案设计

- 如果一个哲学家想吃面条，那么他先问左右
- 如果左右都不在吃，那么他拿两根筷子吃
- 如果左右有人在吃，他就饿着等着，直到别人吃完了通知他
- 吃完之后，他放下筷子并且通知左右他吃完了

Shared object

```
#define N 5
#define LEFT  ((i+N-1) % N)
#define RIGHT  ((i+1) % N)

int state[N];
semaphore mutex = 1;
semaphore p[N] = 0;
```

Identifying i's "left" and "right"

The states of the philosophers, including "**EATING**", "**THINKING**", and "**HUNGRY**".

Remember, this is a shared array.

To guarantee mutual exclusive access to the "**state[N]**" array.

To model a philosopher (instead of a chopstick)

Shared object

```
#define N 5
#define LEFT  ((i+N-1) % N)
#define RIGHT  ((i+1) % N)

int state[N];
semaphore mutex = 1;
semaphore p[N] = 0;
```

Main function

```
1  void philosopher(int i) {
2      think();
3      take_chopsticks(i);
4      eat();
5      put_chopsticks(i);
6  }
```

```
void wait(semaphore *s) {
   disable_interrupt();
   *s = *s - 1;
   if ( *s < 0 ) {
      enable_interrupt();
      sleep();
      disable_interrupt();
    }
   enable_interrupt();
}
```

Section entry

```
1  void take_chopsticks(int i) {
2      wait(&mutex);
3      state[i] = HUNGRY;
4      captain(i);
5      post(&mutex);
6      wait(&p[i]);
7  }
```

Section exit

```
1  void put_chopsticks(int i) {
2      wait(&mutex);
3      state[i] = THINKING;
4      captain(LEFT);
5      captain(RIGHT);
6      post(&mutex);
7  }
```

```
void post(semaphore *s) {
  disable_interrupt();
  *s = *s + 1;
  if ( *s <= 0 )
    wakeup();
  enable_interrupt();
}
```

Extremely important helper function

```
1 void captain(int i) {
2     if(state[i] == HUNGRY && state[LEFT] != EATING && state[RIGHT] != EATING) {
3         state[i] = EATING;
4         post(&p[i]);
5     }
6 }
```

Dining philosopher – Hungry

Tell the captain that you are hungry

If one of your neighbors is eating, the captain just does nothing for you and returns

Then, you wait for your chopsticks (later, the captain will notify you when chopsticks are available)

Section entry

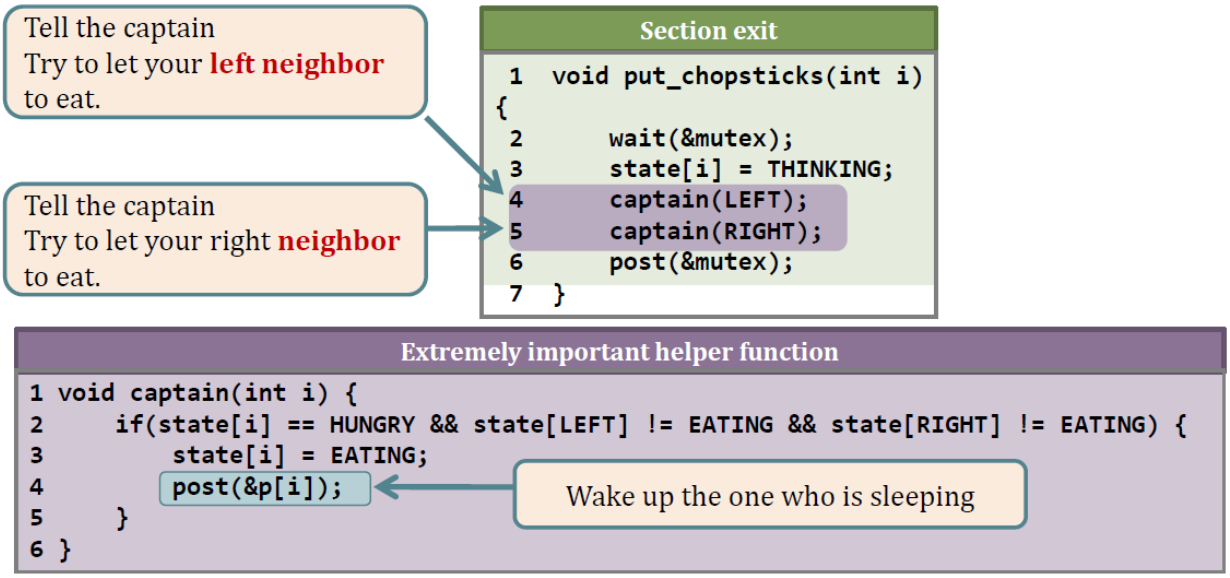
```
1  void take_chopsticks(int i) {
2      wait(&mutex);
3      state[i] = HUNGRY;
4      captain(i);
5      post(&mutex);
6      wait(&p[i]);
7  }
```

Critical Section
The captain is "indivisible"

Extremely important helper function

```
1 void captain(int i) {
2     if(state[i] == HUNGRY && state[LEFT] != EATING && state[RIGHT] != EATING) {
3         state[i] = EATING;
4         post(&p[i]);
5     }
6 }
```

- Finish Eating



第七章 死锁 Deadlock

7.1 死锁的概念

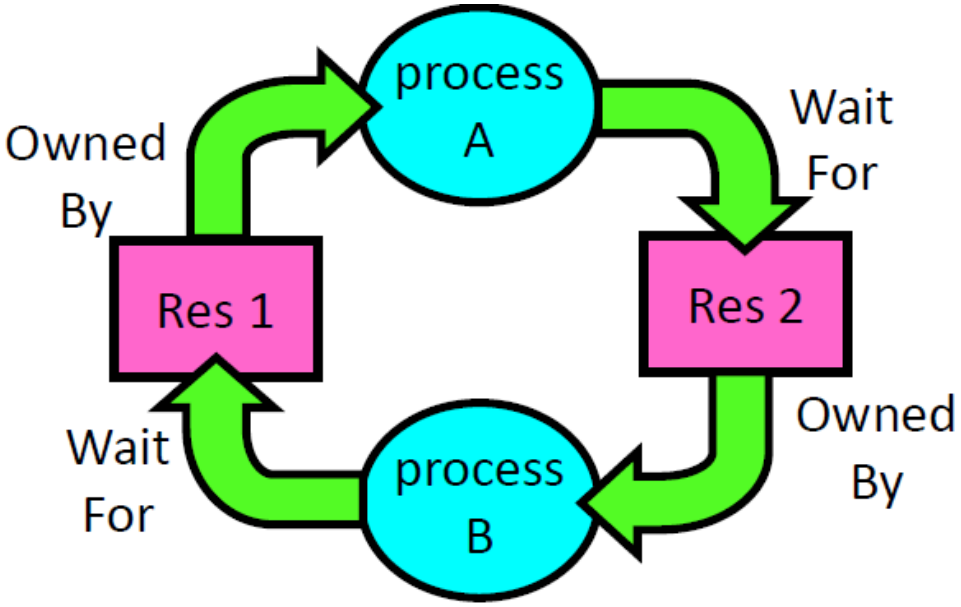
- 在正常操作模式下，进程只能按如下顺序使用资源:
  1. 申请

     进程请求资源。如果进程不能立即被允许，那么它应该等待，直到获取该资源
  2. 使用

     进程对资源进行操作
  3. 释放

     进程释放资源
- 死锁 Deadlock

  Deadlock is a situation where a set of processes are blocked because each process is holding a resource and waiting for another resource acquired by some other process.



- 饥饿 Starvation

  Indefinite Blocking

  A condition in which a process is indefinitely delayed because other processes are always given preference.

  Starvation is the problem that occurs when high priority processes keep executing and low priority processes get blocked for indefinite time.

  Deadlock 一定会造成 starvation

7.2 死锁的特征

7.2.1 死锁的必要条件

1. 互斥 Mutual Exclusion

   Only one thread at a time can use a resource.

2. 占有并等待 Hold and Wait

   一个进程应占有至少一个资源并等待另一个资源，而该资源为其他进程所占有

3. 非抢占 No Preemption

   资源不能被抢占，即资源只能被进程在完成任务后自愿释放

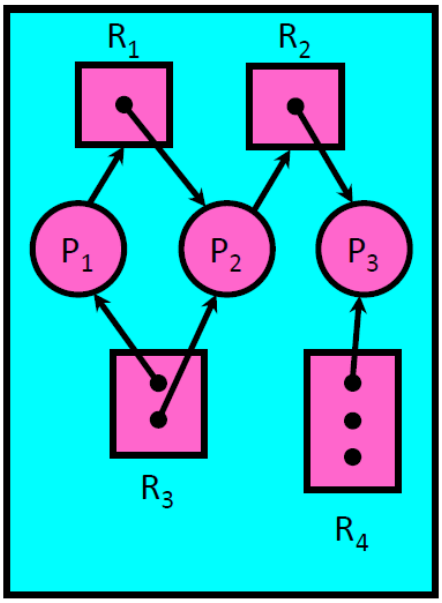
4. 循环等待 Circular Wait

   有一组等待进程 $\{P_0, P_1, P_2, \ldots, P_n\}$

   - P_0 等待的资源被 P_1 占有
   - P_1 等待的资源被 P_2 占有
   - ...
   - P_n 等待的资源被 P_0 占有

注意是必要条件，即使这些条件都满足也不一定死锁，还需要运气比较背

7.2.2 资源分配图 Resource-Allocation Graph

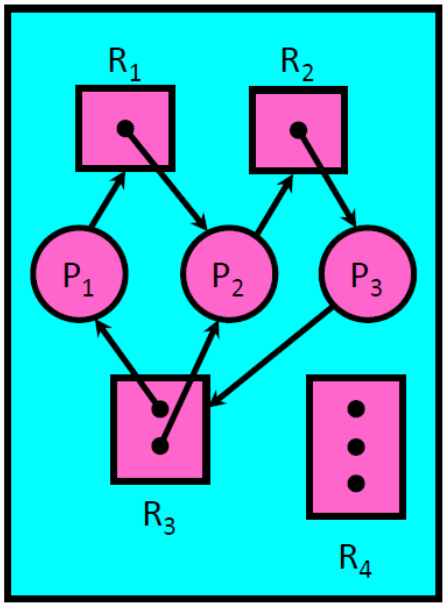


- 圆表示进程
- 矩形表示资源
- 矩形内的点表示资源实例
- 申请边 Request Edge

  进程指向资源的边
- 分配边 Assignment Edge

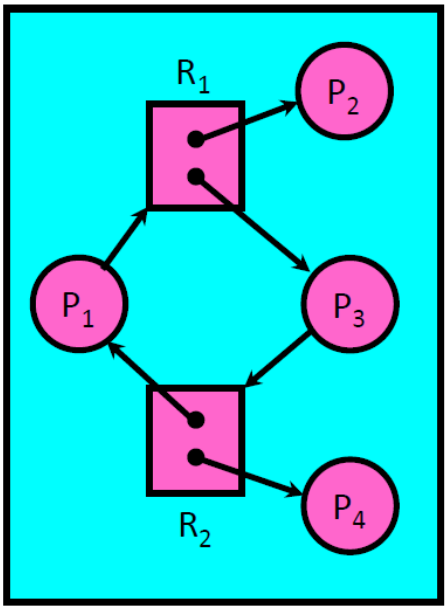
资源指向进程的边

- 如果分配图没有环，那么系统一定没有死锁；如果有环，那么可能存在死锁
- 死锁的例子



- 有环没死锁的例子

  让 P_2, P_4 先执行完



7.3 死锁的处理方法

- 通过协议来预防或避免死锁，确保系统不会进入死锁状态
- 允许系统进入死锁状态，然后检测并恢复
- 忽视，认为死锁不可能在系统内发生

  这种方案被 Linux, Windows 等大多数 OS 采用

  就算出现了死锁，OS 也不管

7.4 死锁检测 Deadlock Detection

7.4.1 死锁检测算法

[xxx] 表示数组

- [FreeResources]: current free resources each type
- [Request_X]: current requests from process X
- [Alloc_X]: current resources held by process X

```
[Avail] = [FreeResources]
Add all nodes to UNFINISHED

do {
    done = true
    Foreach node in UNFINISHED {
        if ([Request_node] ≤ [Avail]) {
            remove node from UNFINISHED
            [Avail] = [Avail] + [Alloc_node]
            done = false
        }
    }
} until(done)
```

7.4.2 死锁恢复

当检测到死锁后:

- 进程终止

  Terminate thread, force it to give up resources
- 资源抢占

  Preempt resources without killing off process
- 回滚

  Roll back actions of deadlocked threads
- Many operating systems use other options

7.5 死锁预防 Deadlock Prevention

核心：打破四个必要条件

1. 互斥
   - 大家都用只读文件
   - 给足够多的资源
2. 持有且等待

    - 每个进程在执行前申请并获得所有资源
    - 进程仅在没有资源时才申请资源
3. 无抢占
    - 如果一个进程持有资源并申请一个不能被立即分配的资源，那么它现在分配的资源都可以被抢占

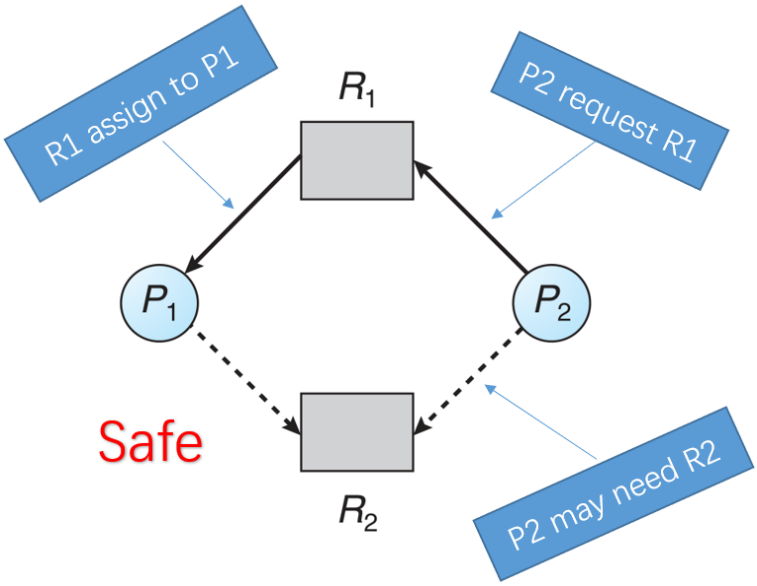
      相当于把它现有的资源都释放了
4. 循环等待
    - 给所有进程一个指定的顺序来申请资源

7.6 死锁避免 Deadlock Avoidance

7.6.1 安全状态

- 如果系统能按一定顺序为每个进程分配资源（不超过其最大需求），可以避免死锁，那么系统的状态就是安全的 (safe)
- 只有存在一个安全序列 (safe sequence)，系统才处于安全状态
- 如果没有这样的序列存在，那么系统状态就是非安全 (unsafe)
- 非安全状态只是可能会导致死锁，不是一定

7.6.2 资源分配图算法



- 需求边 Claim Edge

  进程指向资源，虚线

  进程 P_i 可能在将来申请某个资源 R_j
- 只有在将申请边变成分配边 (反向实线箭头) 并且不会导致资源分配图形成环时，才能允许申请
- 时间复杂度

  $O(n^2)$, n 为进程数量

7.6.3 银行家算法 Banker's Algorithm

n 个进程, m 种资源

- $Available$: 行向量，表示每种资源的可用实例数量
- Max: $n \times m$ 矩阵，每个进程的最大需求
- $Allocation$: $n \times m$ 矩阵，每个进程已经分配的实例数量
- $Need = Max - Allocation$，还缺多少实例才能完事

```
Add all nodes to UNFINISHED

do {
    done = true
    Foreach node in UNFINISHED {
        if ([Max_node] - [Alloc_node] ≤ [Avail]) {
            remove node from UNFINISHED
            [Avail] = [Avail] + [Alloc_node]
            done = false
        }
    }
} until(done)
```

- 例: 可以按 0213 或 0231 的顺序执行完

	Allocation				Max				Available			
	A	B	C	D	A	B	C	D	A	B	C	D
P0	0	0	1	2	0	0	1	2	1	5	2	0
P1	1	0	0	0	1	7	5	0				
P2	1	3	5	4	2	3	5	6				
P3	0	0	1	4	0	6	5	6				

- 现在有个新的 $Request$，比如 P_0 $(1, 3, 1, 0)$
    1. 检查 $Request_0 < Available$
    2. 检查 $Allocation_0 + Request_0 < Max_0$
    3. 假设把资源分配给 P_0
        - 如果分配之后还是 safe（能找到安全序列），那就真的分配给它
        - 如果分配之后 unsafe，拒绝请求

第八章 内存管理策略

8.1 背景

一个内存，多个进程，怎么管理

8.1.1 Aspects of Memory Multiplexing

- Protection

  Prevent access to private memory of other processes

- Controlled Overlap

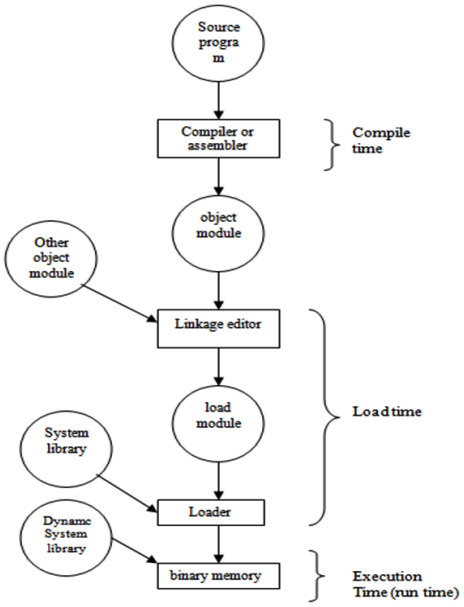
  Sometimes we want to share memory across processes

- Translation

  Ability to translate accesses from one address space (virtual) to a different one (physical)

8.1.2 地址绑定 Address Binding

源程序中的地址通常是用符号表示 (如变量 `count`)。编译器通常将这些符号地址绑定 (bind) 到可重定位的地址 (如“从本模块开始的第 14 字节”)。链接程序或加载程序再将这些可重定位的地址绑定到绝对地址 (如 74014)。每次绑定都是一个从一个地址空间到另一个地址空间的映射。



通常，指令和数据绑定到存储器地址可以在任何一步进行:

- 编译时 Compile Time

  如果在编译时就已经知道进程将在内存中的驻留地址，那么就可以生成绝对代码 (Absolute Code)

  例: MS-DOS 的 .COM 格式程序

- 加载时 Load Time

  如果在编译时并不知道进程将驻留在何处，那么编译器就应生成可重定位代码 (Relocatable Code)。对这种情况，最后绑定会延迟到加载时进行

- 执行时 Runtime time

  如果进程在执行时可以从一个内存段移到另一个内存段，那么绑定应延迟到执行时才进行

  大多数通用 OS 采用

8.1.3 逻辑地址空间与物理地址空间

- 逻辑地址 Logical Address = 虚拟地址 Virtual Address

  CPU 生成的地址

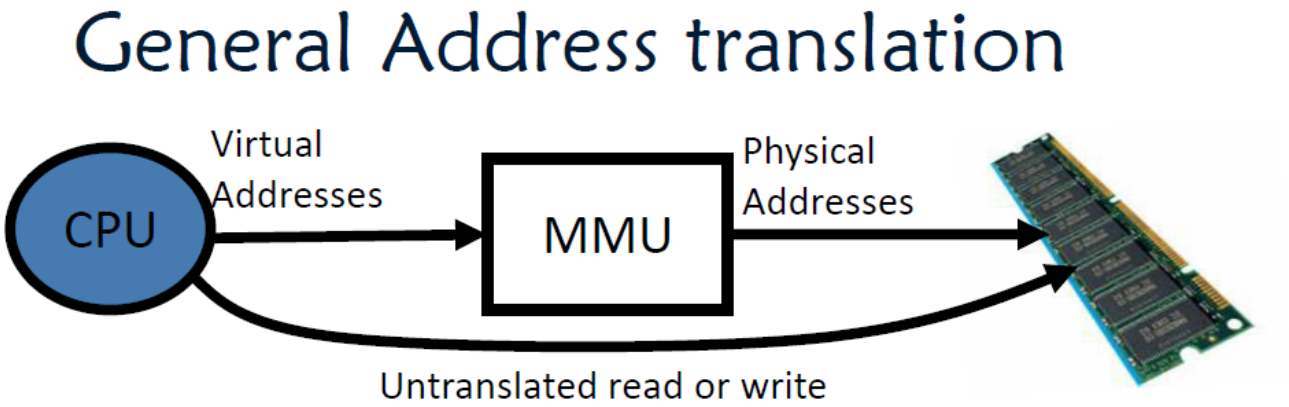
- 物理地址 Physical Address

  真正的内存地址，加载到内存地址寄存器 (Memory-Address Register) 的地址

- 编译时和加载时的地址绑定会生成相同的逻辑地址和物理地址
- 执行时的绑定生成不同的逻辑地址和物理地址
- 内存管理单元 MMU

  从虚拟地址到物理地址的运行时映射是由**内存管理单元 (Memory Management Unit)** 的硬件设备来完成 (包括查页表之类的都是 MMU 干的)

  大多数 on-chip



8.1.4 动态加载 Dynamic Loading

- 一个进程的整个程序和数据如果都必须处于物理内存中，则进程的大小受物理内存大小的限制
- 为了获得更好的内存空间使用率，使用动态加载（Dynamic Loading），即一个程序只有在调用时才被加载

8.1.5 动态链接与共享库

- 动态链接的概念与动态加载相似。只是这里不是将加载延迟到运行时，而是将链接延迟到运行时。这一特点通常用于系统库，如语言子程序库。没有这一点，系统上的所有程序都需要一份语言库的副本，这一需求浪费了磁盘空间和内存空间。
- 存根 Stub

  如果有动态链接，二进制镜像中每个库程序的应用都有一个存根（stub）。存根是一小段代码，用以指出如何定位适当的内存驻留的库程序，或如果该程序不在内存中应如何安装入库。不管怎样，存根会用子程序地址来代替自己，并开始执行子程序。因此，下次再执行该子程序代码时，就可以直接进行，而不会因动态链接产生任何开销。采用这种方案，使用语言库的所有进程只需要一个库代码副本就可以了。
- 举例来说，你在程序里调用了 STL 里的 Map，如果没有动态链接，就相当于你把 STL 里 Map 的源文件复制一份到了你的项目里。在动态链接下，不管多少程序调用，都只会调用那一份代码。
- 动态连接也可用于库更新。一个库可以被新的版本所替代，且使用该库的所有程序会自动使用新的版本。没有动态链接，所有这些程序必须重新链接以便访问。

8.2 交换 Swap

Refer to 进程调度 5.1.3 中期调度程序

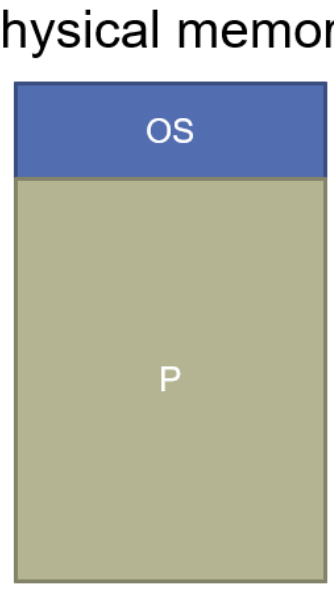
进程需要在内存中以便执行。进程也可以暂时从内存中交换 (swap) 到备份存储 (backing store，一般是磁盘) 上，当需要再次执行时在调回到内存中。

- 换入 Swap In
- 换出 Swap Out

8.3 连续内存分配 Contiguous Memory Allocation

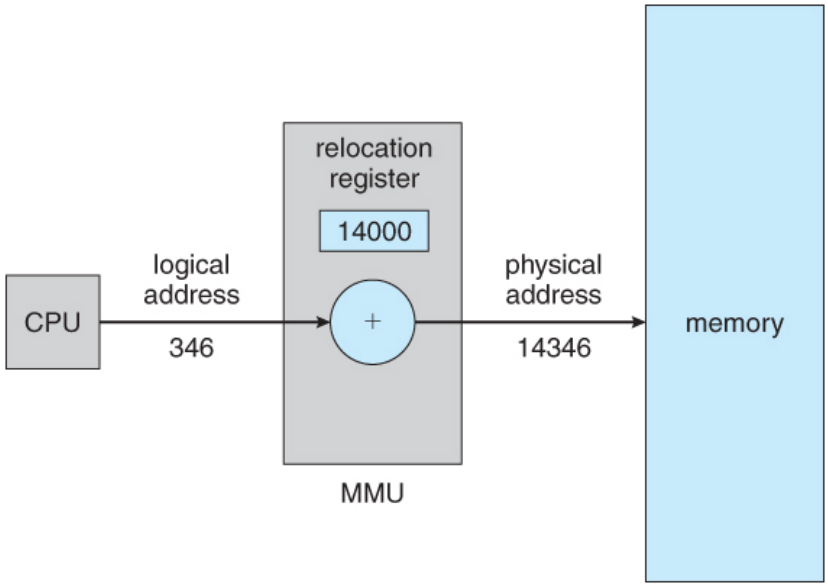
8.3.1 Uniprogramming

- 同时只能有一个程序运行
- Application always runs at same place in physical memory since only one application at a time
- Application can access any physical address

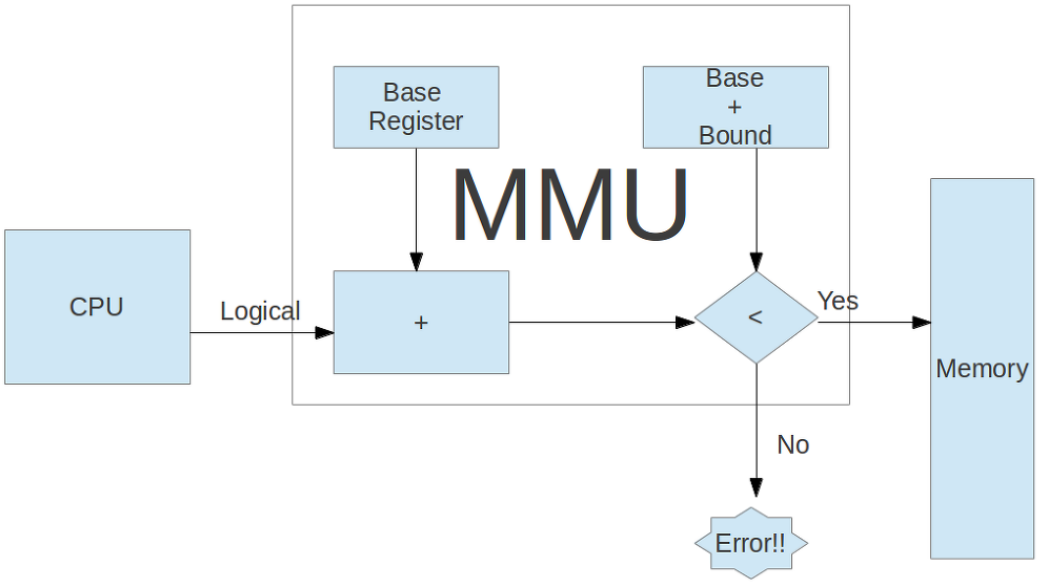


8.3.2 内存保护 Protection

- 重定位寄存器 Relocation Register
- 界限寄存器 Limit Register: 里面是虚拟地址的 bound



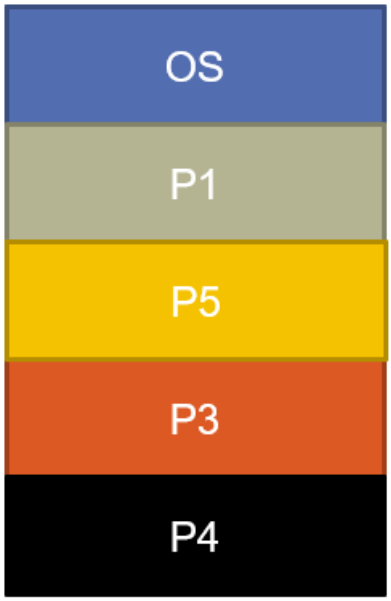
- 保护



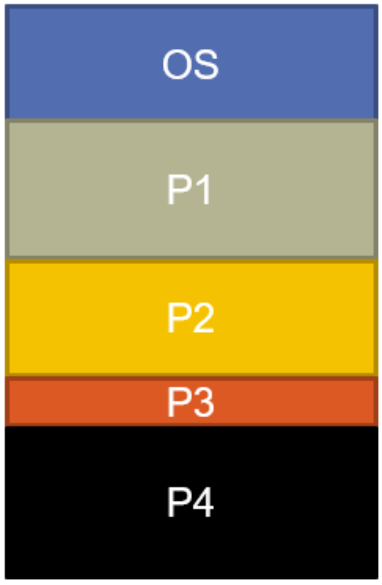
8.3.3 多分区方法 Multiple-Partition Method

- 将内存分为多个分区，每个分区分给一个进程
- 固定分区 Fixed-size Partition

  Each process has same memory size



- 可变分区 Variable Partition



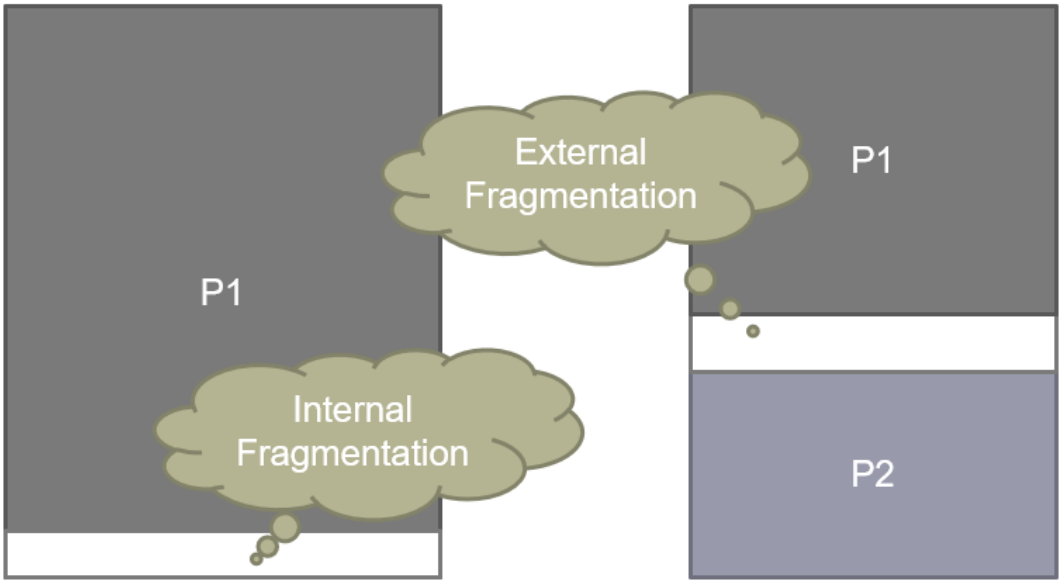
    - 每一块可用的内存称为一个孔 (hole)
    - 可用的内存块为分散在内存里不同大小的孔的集合
- 动态存储分配问题 Dynamic Storage-Alloction Problem
    - 当新进程需要内存时，系统为该进程查找足够大的孔
    - 如果孔太大，那么就分为两块
        - 分配给新进程
        - 合并回孔集合
    - 进程终止时，释放内存，该内存合并回孔的集合
    - 如果新孔与其他孔相邻，则合并成大孔
    - 系统检查是否有等待内存空间的进程，以及新合并的孔能否满足等待进程等
- 从可用孔中选择一个分配的常用方法
    - 首次适应 First-fit

      分配首个足够大的孔
    - 最优适应 Best-fit

      分配最小的足够大的孔
    - 最差适应 Worst-fit

      分配最大的足够大的孔

8.3.3 碎片 Fragmentation



- 内碎片 Internal Fragmentation

  分配给进程的内存比所需的大，多余的那一部分就是内碎片

- 外碎片 External Fragmentation

  两个进程之间的空闲孔，而且这个孔太小，没法分配给别的进程

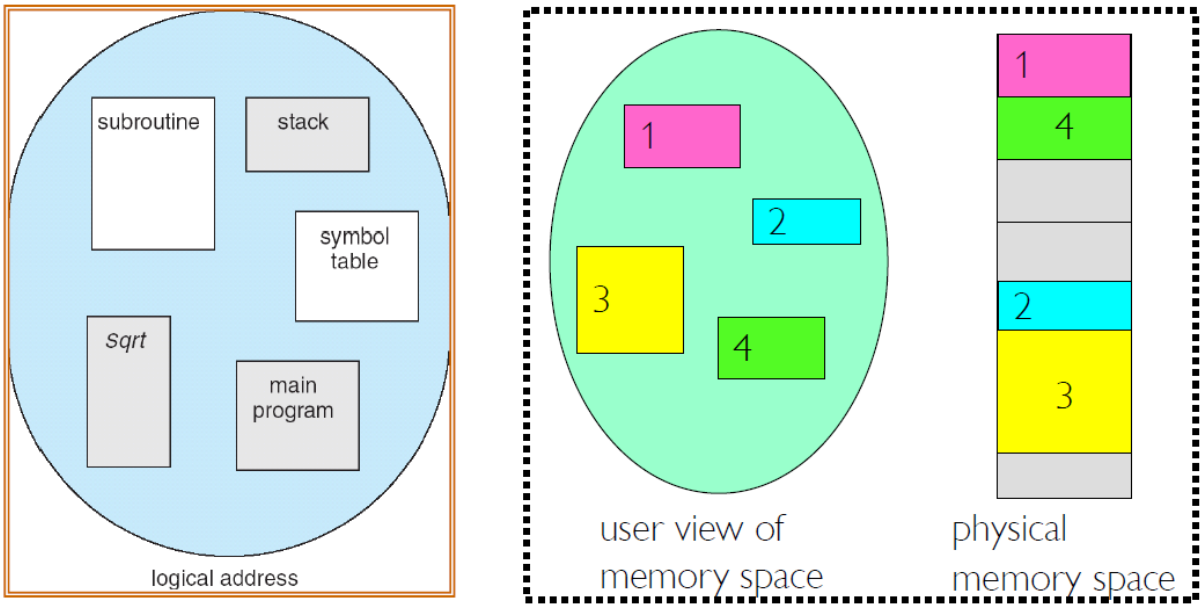
- 紧缩 Compaction

  移动已分配的内存，使得所有外碎片合并成一大块

  只有在运行时绑定才可以使用紧缩，因为要重写基地址寄存器和界限寄存器

  - 编译时：不可能，因为直接就绑定绝对地址
  - 加载时：但是目前进程已经加载到内存里了，这时候也已经是不可变地址了

8.4 分段 Segmentation



- 逻辑地址空间由一组段构成，每个段有名称和长度
- 地址指定了段名称和段内偏移 (Offset)

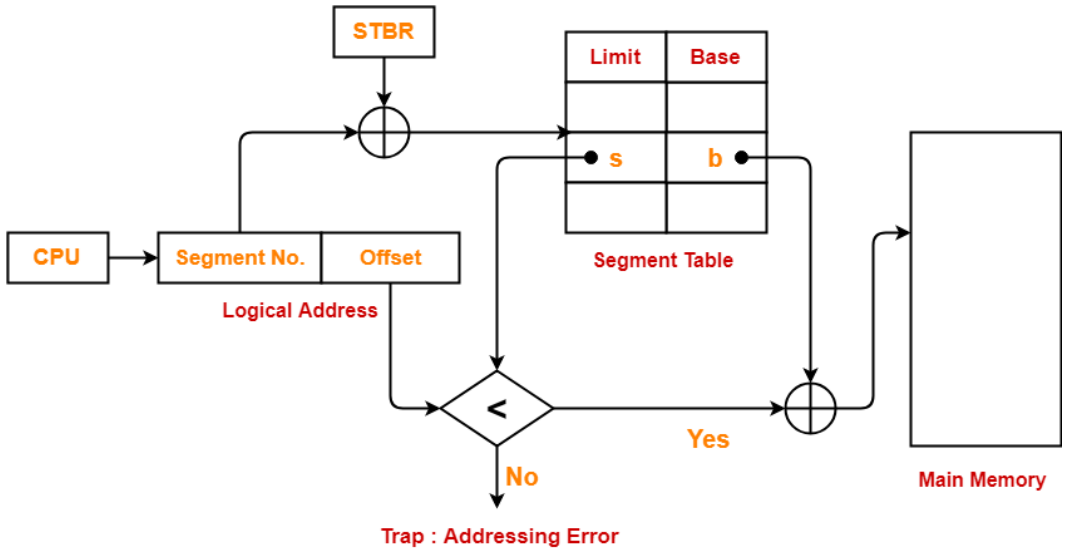
- 逻辑地址由有序对组成 <段号，偏移>
- 段表 Segment Table

  在 CPU 里

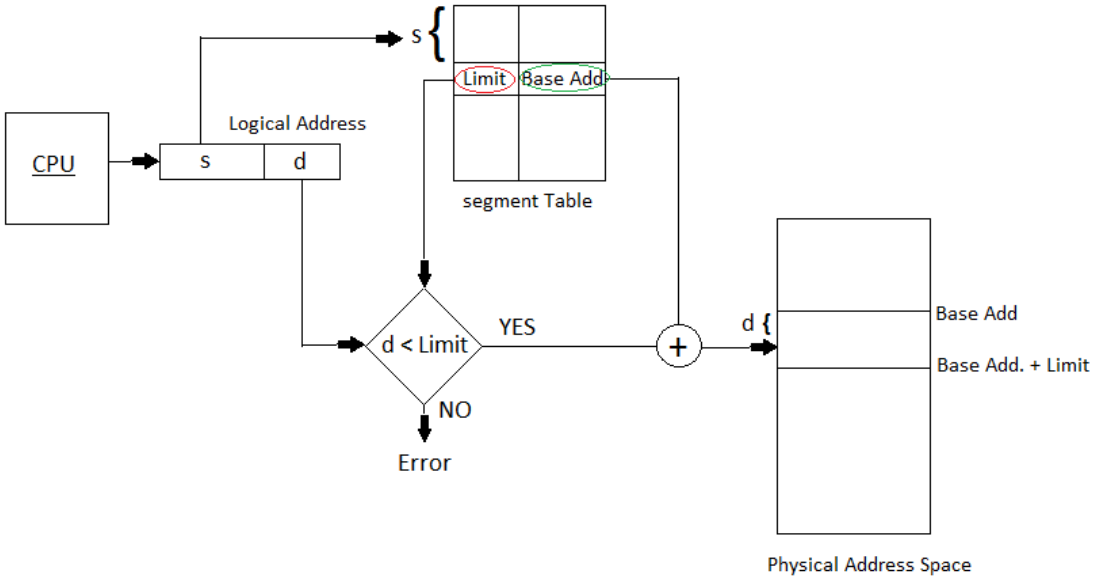
  每个进程都有一个

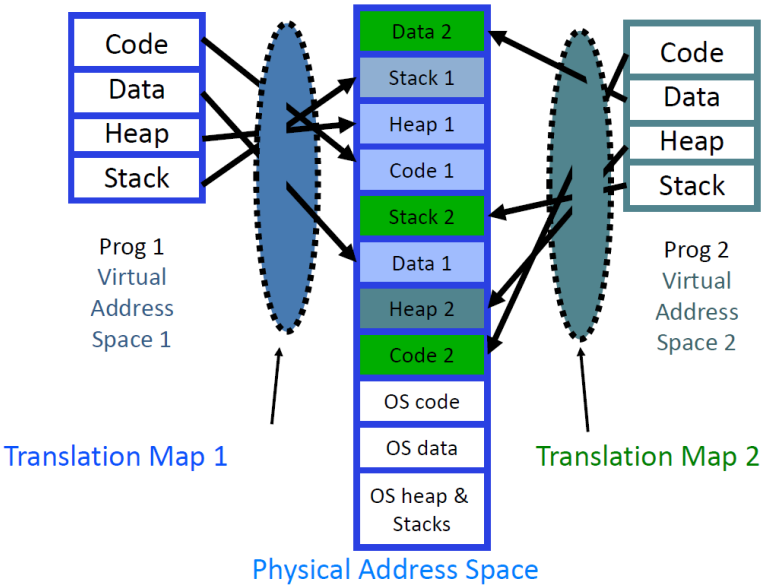
  段表的每个条目包含:

  - 段基地址 Segment Base
  - 段界限 Segment Limit
  - Valid bit

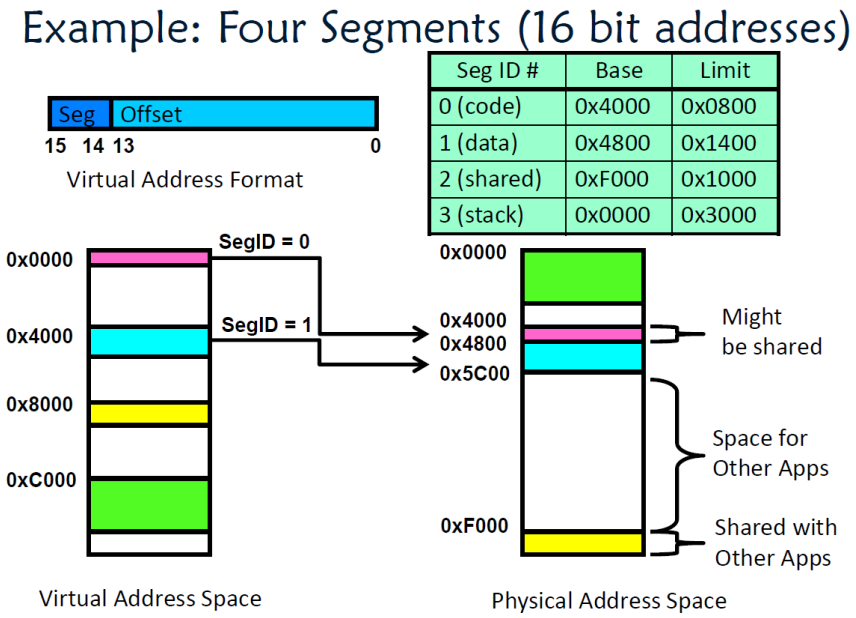


Translating Logical Address into Physical Address





- 图上没有 Check Valid 的内容
- 例 1:



- 例 2:

```
0x240   main:   la $a0, varx
0x244           jal strlen
  …               …
0x360   strlen: li   $v0, 0  ;count
0x364   loop:   lb   $t0, ($a0)
0x368           beq  $r0,$t0, done
  …               …
0x4050  varx    dw   0x314159
```

Seg ID #	Base	Limit
0 (code)	0x4000	0x0800
1 (data)	0x4800	0x1400
2 (shared)	0xF000	0x1000
3 (stack)	0x0000	0x3000

Let us simulate a bit of this code to see what happens (PC=0x240):

1. Fetch 0x240. Virtual segment #? 0; Offset? 0x240
   Physical address? Base=0x4000, so physical addr=0x4240
   Fetch instruction at 0x4240. Get "la $a0, varx"
   Move 0x4050 → $a0, Move PC+4→PC
2. Fetch 0x244. Translated to Physical=0x4244. Get "jal strlen"
   Move 0x0248 → $ra (return address!), Move 0x0360 → PC
3. Fetch 0x360. Translated to Physical=0x4360. Get "li $v0, 0"
   Move 0x0000 → $v0, Move PC+4→PC
4. Fetch 0x364. Translated to Physical=0x4364. Get "lb $t0, ($a0)" Since $a0 is 0x4050, try to load byte from 0x4050, Translate 0x4050 (0100 0000 0101 000). Virtual segment #? 1; Offset? 0x50 Physical address? Base=0x4800, Physical addr = 0x4850, Load Byte from 0x4850→$t0, Move PC+4→PC

不多说，全是计组学过的内容

注意第一条指令 load address 取的是虚拟地址 0x4050, 物理地址只有真正访问内存的时候才翻译过去

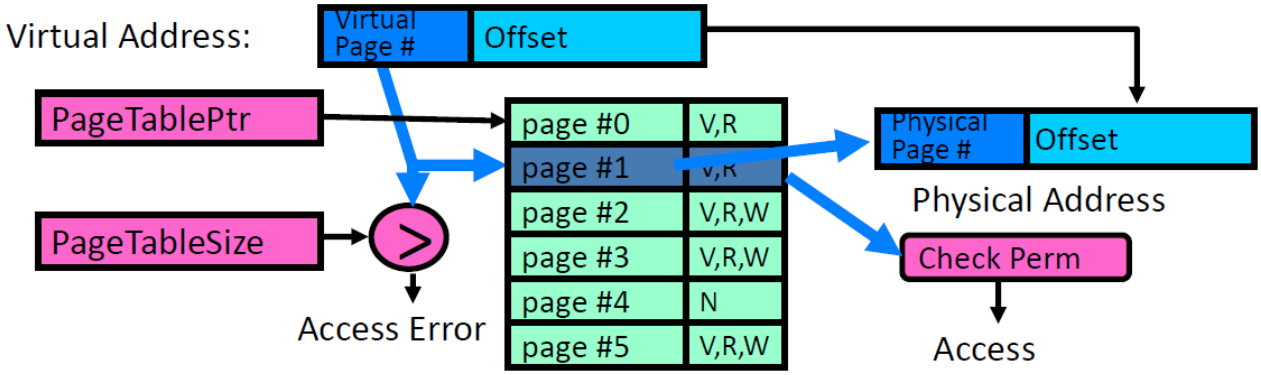
- 分段的特点
  1. Virtual address space has holes
     - Segmentation is efficient for sparse address spaces
     - A correct program should never address gaps
  2. When it is OK to address outside valid range?
     - This is how the stack and heap are allowed to grow
     - For instance, stack takes fault, system automatically increases size of stack
  3. Need protection mode in segment table
     - For example, code segment would be read only
     - Data and stack would be read write (stores allowed)
     - Shared segment could be read only or read write
  4. Fragmentation
- What must be saved/restored on context switch?
  - Segment table that stored in CPU
  - Might store all of processes memory onto disk when switched (8.2 swap)

8.5 分页 Paging

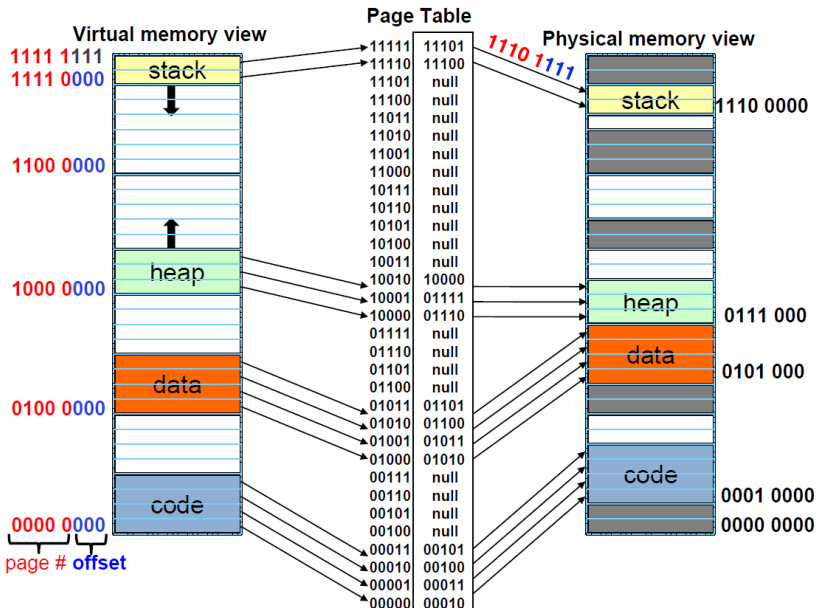
- 将物理内存分为固定大小的块，称为帧或页帧 (frame)
- 将逻辑内存分为同样大小的块，称为页或页面 (page)
- 逻辑地址空间完全独立于物理地址空间，例如一个进程有 64 位逻辑地址空间，而系统的物理内存可以小于 2^{64} 字节

8.5.1 页表 Page Table

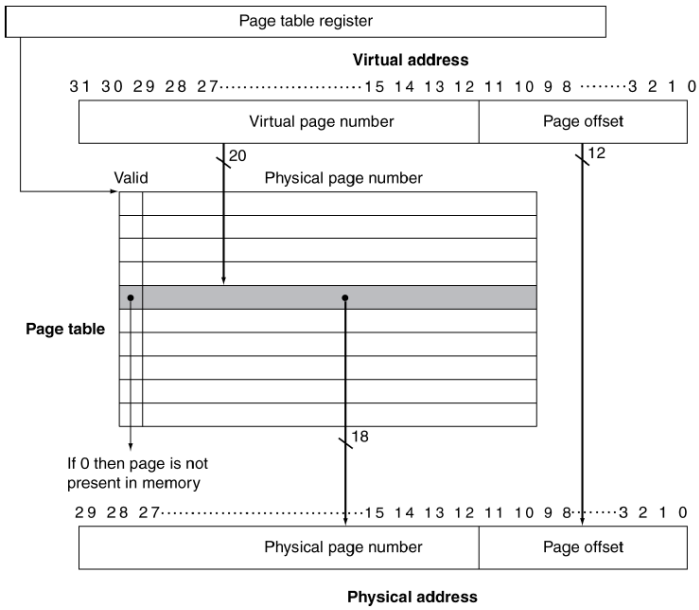
- 每个逻辑地址分为两部分
  - 页码 Page Number: 页表的索引
  - 页偏移 Page Offset
- 与分段对比
  - 分页是先决定一页多大才知道分几页，比如一页 4 KiB, 那么 offset 是 12 位，所以 page number 是 32-12=20 位
  - 分段是先决定分几个段才知道一个段多大，比如分 4 个，那么 segment number 是 2 位，所以 offset 是 32-2=30 位
- 页表条目
  - 物理内存基地址
  - Valid bit, read, write ...
- 在内存里
- 每个进程一个

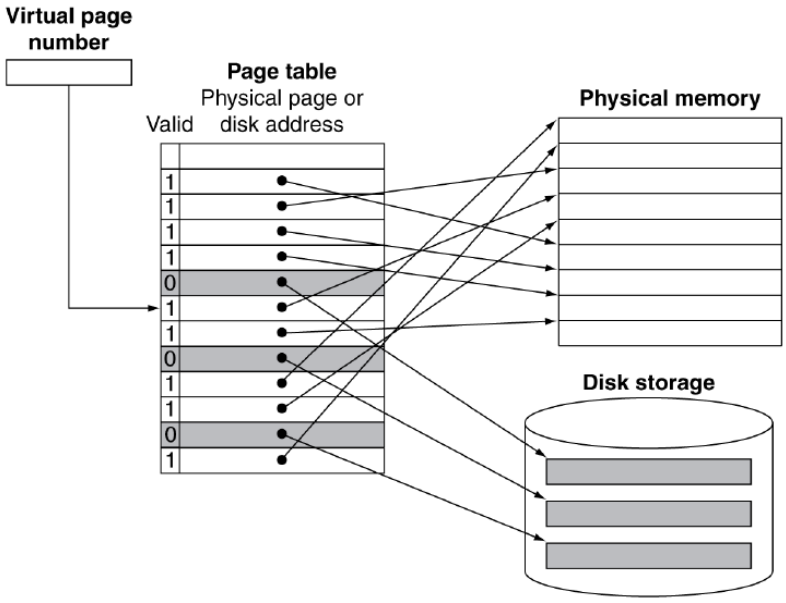


- Page Table (One per process)



- 下图来自计组课件（逻辑地址空间和物理地址空间不必非得一样大）

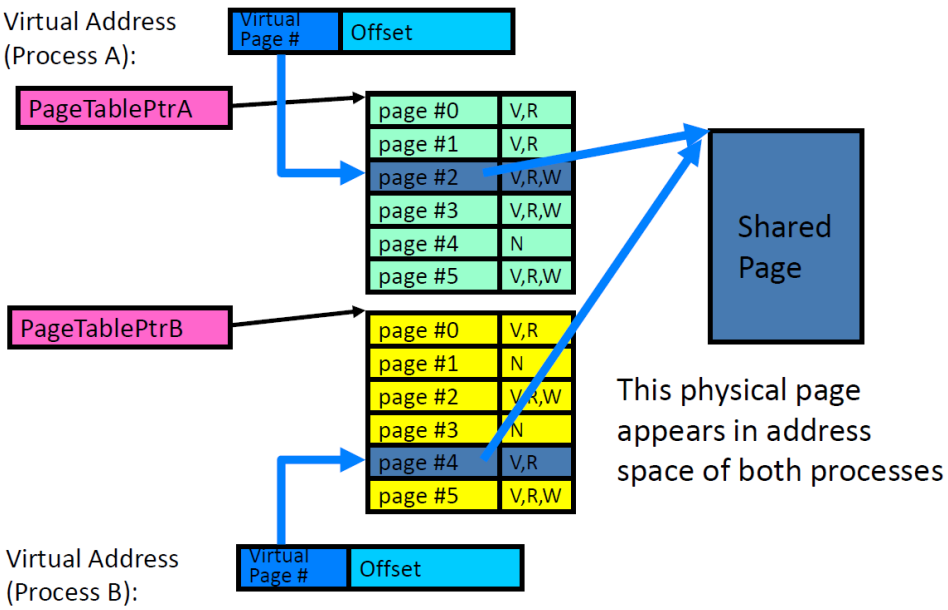




- 分页可以消除外碎片
- 一页的大小需要 trade off
  - 太大，内碎片
  - 太小，页表条目太多，导致页表占用空间太大，页表也是在内存里的
- What needs to be switched on a context switch?
  Page table pointer and limit
- Core Map

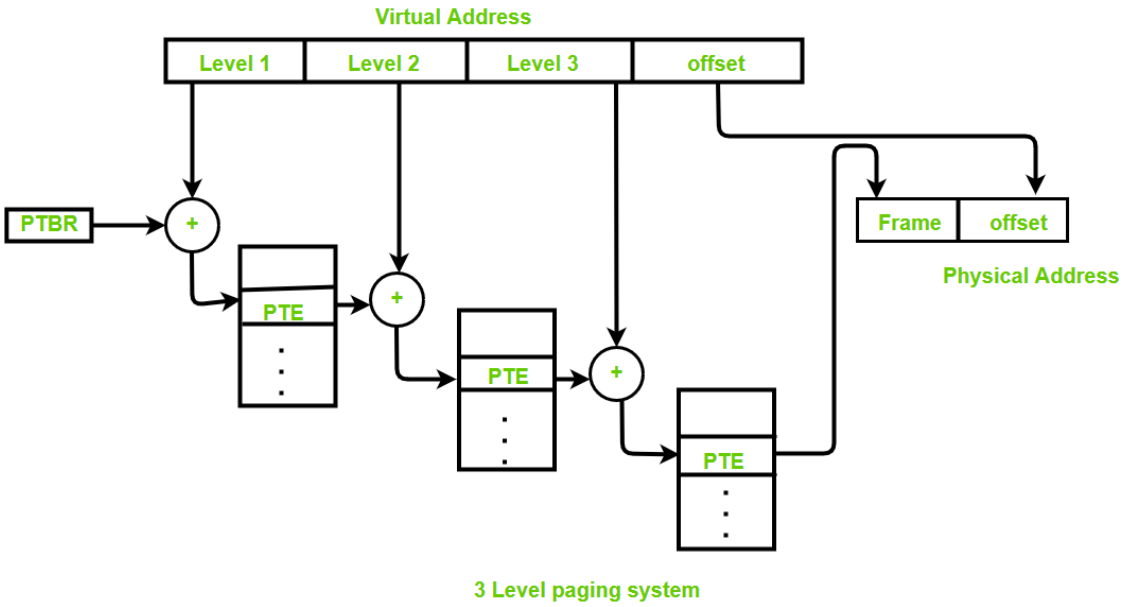
  Do we need a reverse mapping (i.e. physical page -> virtual page)?
  - Yes. Clock algorithm runs through page frames. If sharing, then multiple virtual pages per physical page
  - Can't push page out to disk without invalidating all PTEs

8.5.2 共享页



8.5.3 分层分页 Multilevel Paging

- 向前映射页表 Forward-Mapped Page Table

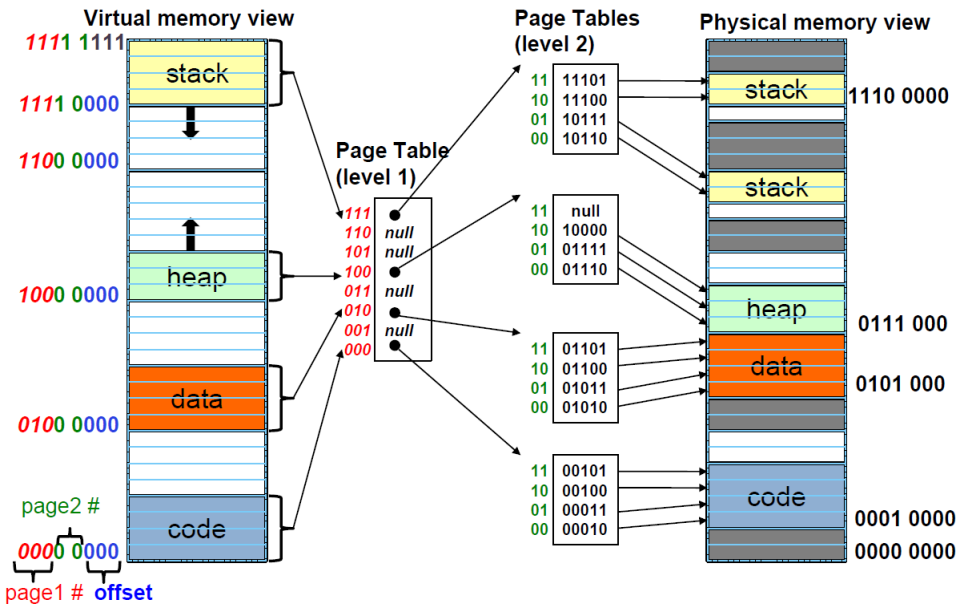


- Page Table Base Register (PTBR): 存的是页表的基地址

1. Reference to PTE in level 1 page table = PTBR value + Level 1 offset present in virtual address.
2. Reference to PTE in level 2 page table = Base address (present in Level 1 PTE) + Level 2 offset (present in VA).
3. Reference to PTE in level 3 page table= Base address (present in Level 2 PTE) + Level 3 offset (present in VA).
4. Actual page frame address = PTE (present in level 3).

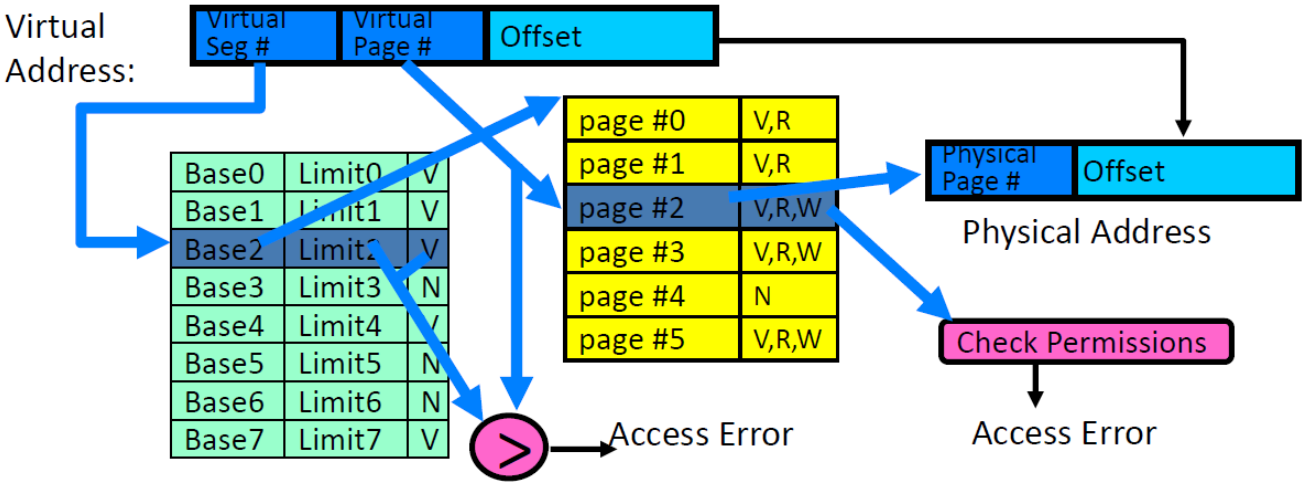
- 注意这里，比如 level 1 offset 10 位，那么二级页表最多可能有 $2^{10} = 1024$ 个

  当然大部分情况都是 < 1024 的，这就是多级页表的作用，解决了之前单个页表里一大堆 null 没用还占地方的问题

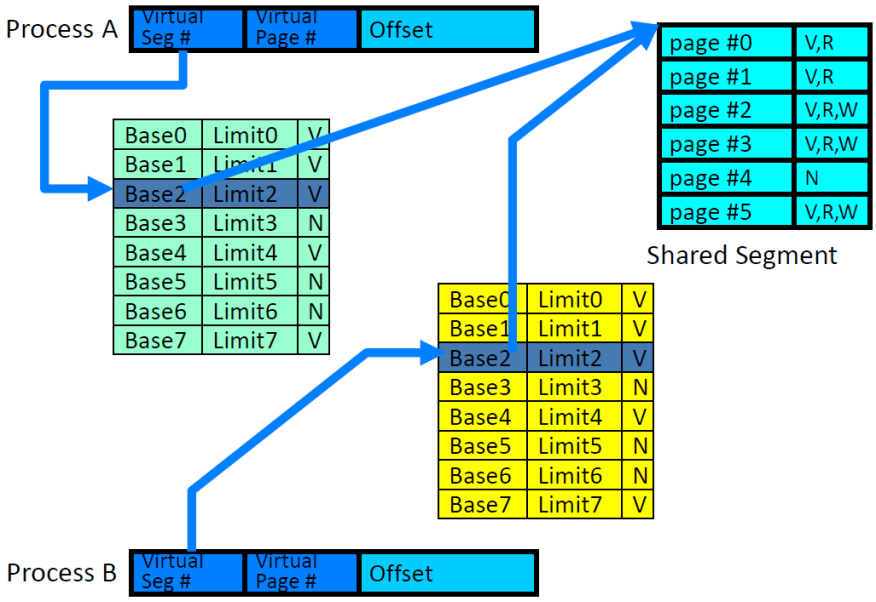


8.5.4 分段+分页

- Tree of tables
  - Lowest level page table: memory still allocated with bitmap
  - Higher levels often segmented



- What must be saved/restored on context switch?
  - Contents of top level segment registers (for this example)
  - Pointer to top level table (page table)
- 共享



第九章 虚拟内存管理

9.1 缓存 Cache

- Cache

  A repository for copies that can be accessed more quickly than the original
- 平均访问时间 Average Access time

  $$Hit\ Rate \times Hit\ Time + Miss\ Rate \times Miss\ Time$$

  注意这里跟计组学的不一样，访问 cache 然后没找到的时间没算进去

  计组: $Hit\ Time + Miss\ Rate \times Miss\ Time$
- 时间局部性 Temporal Locality

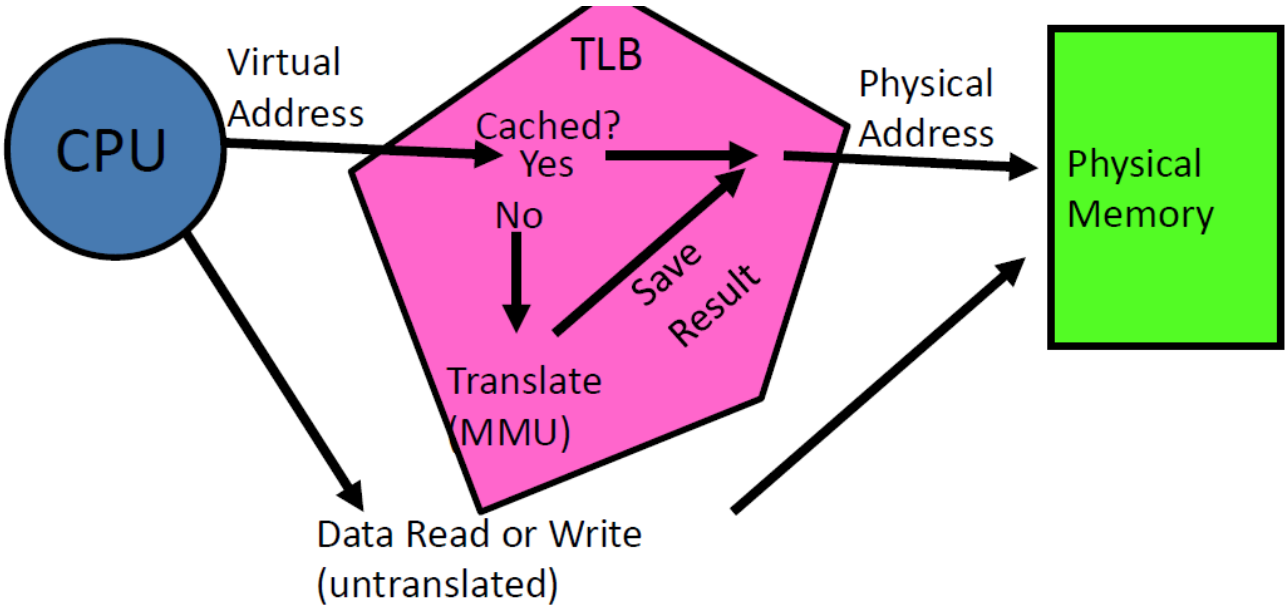
  If you used some data recently, you will likely use it again

  Keep recently accessed data items closer to processor
- 空间局部性 Spatial Locality

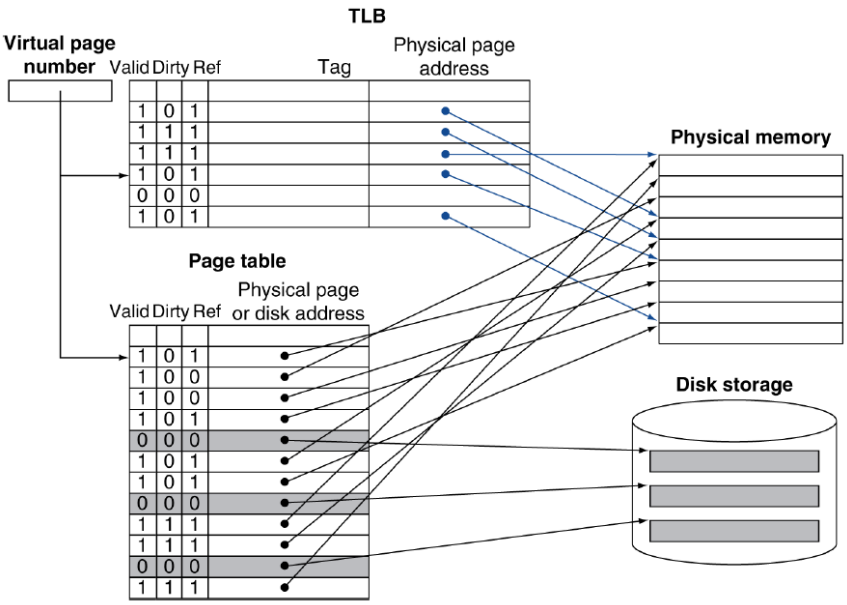
  If you used some data recently, you will likely access its neighbors

  Move contiguous blocks to the upper levels
- 其他内容详见计组课件
- Cache 的应用
  - TLB
  - 虚拟内存
  - 文件系统
  - DNS
  - Web Proxy

9.2 转换表缓冲区 Transition Look-aside Buffer (TLB)

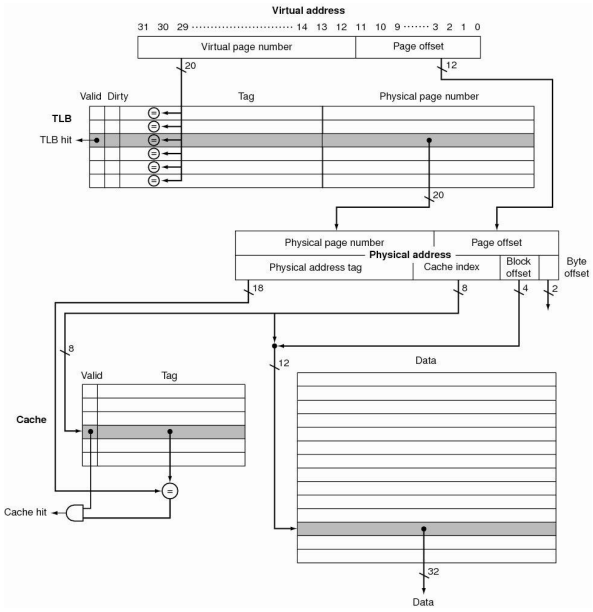


- Page Table Entry (PTE) 的 cache
- 在 CPU 里
- TLB Miss 的处理 (以下来自计组课件 永远滴神)
  - If page is in memory
    - Load the PTE from memory and retry
    - Could be handled in hardware
      - Can get complex for more complicated page table structures
    - Or in software
      - Raise a special exception, with optimized handler
  - If page is not in memory (page fault)
    - OS handles fetching the page and updating the page table
    - Restart the faulting instruction
- TLB 也可以多级 (L1, L2, ...)



- TLB + Cache

- 注意：cache（就是 CPU 里的 L1, L2, ... 那些）用的是物理地址



- Does software loaded TLB need use bit?
  - Hardware sets use bit in TLB; when TLB entry is replaced, software copies use bit back to page table
  - Software manages TLB entries as FIFO list; everything not in TLB is Second Chance list, managed as strict LRU

9.3 请求调页 Demand Paging

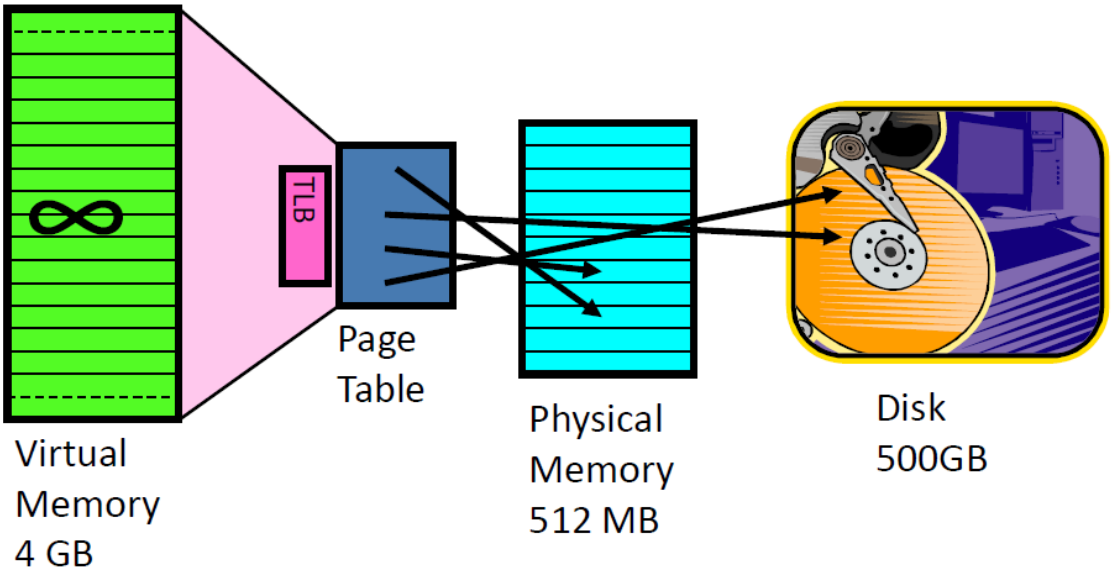
9.3.1 基本概念

- 定义

  Keep all pages of the frames in the secondary memory (外存) until they are required.

  A page is delivered into the memory on demand i.e., only when a reference is made to a location on that page.

- 为什么要请求调页
  - 一个程序运行需要的内存比实际内存大，但是这个程序不是同时需要申请这么多内存
  - 程序所使用的虚拟地址也可能超出物理地址 (比如 64 位系统的虚拟地址空间相当大)
- 内存相当于外存的 cache
  - Block size: 1 page
  - Organization: fully associative
  - How to find a page: first TLB, then page table traversal
  - How to handle write: write-back, need dirty bit

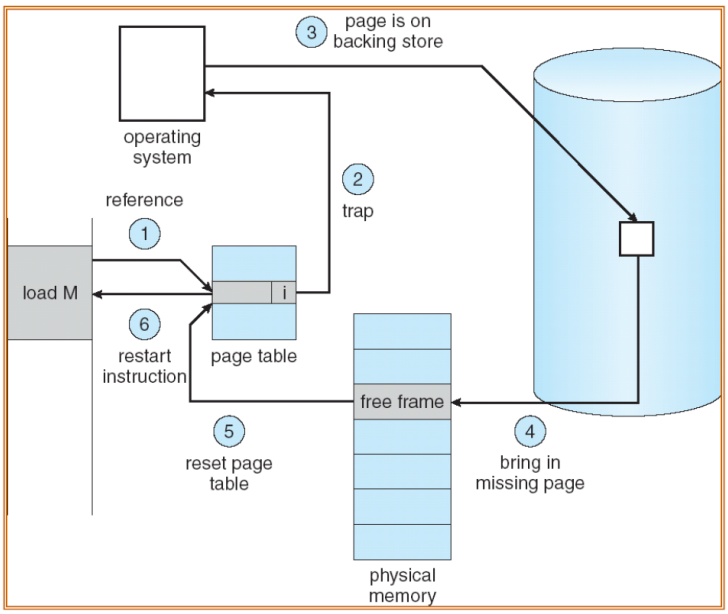


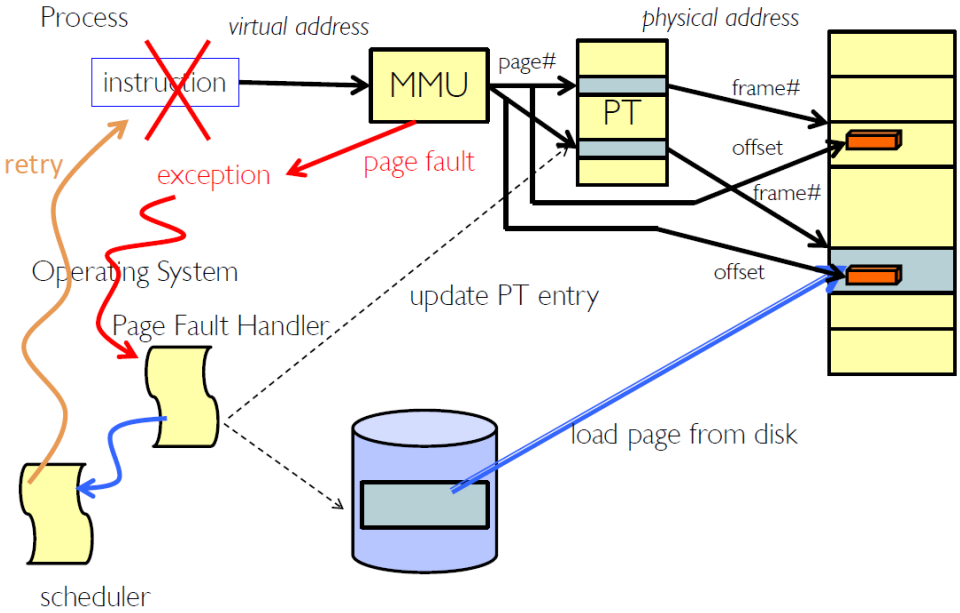
9.3.2 请求调页的性能

- 缺页错误 Page Fault

  对标记为无效 (valid bit) 的页面访问
- 缺页错误的处理
  - 概括
    - 处理缺页错误中断
    - 读入页面
    - 重启进程
  - 具体
    1. 陷入操作系统 (trap)
    2. 保存寄存器和进程状态
    3. 确定中断是否为缺页错误
    4. 检查页面是否合法，并确定页面的磁盘位置
    5. 从磁盘读入页面到空闲帧
       - 在该磁盘队列中等待 (IO 队列)，直到读请求被处理
       - 等待磁盘的寻道或延迟时间
       - 开始传输磁盘页面到空闲帧
    6. 在等待时，将 CPU 分配给其他用户
    7. 收到来自 IO 子系统的中断 (IO 完成)
    8. 保存其他用户的寄存器和进程状态
    9. 确认中断是来自上述磁盘的
    10. 修正页表和其他表，以表示所需页面现在已在内存中
    11. 等待 CPU 再次分配给本进程
    12. 恢复用户寄存器、进程状态和新页表，再重新执行中断的指令
  - 课件

    - Choose an old page to replace
    - If old page modified (`D=1`), write contents back to disk
    - Change its PTE and any cached TLB to be invalid
    - Load new page into memory from disk
    - Update page table entry, invalidate TLB for new entry
    - Continue thread from original faulting location





- OS 如何拿到一个空闲帧
    - Keeps a free list
    - Unix runs a "reaper" if memory gets too full
        - Schedule dirty pages to be written back on disk
        - Zero (clean) pages which have not been accessed in a while
    - As a last resort, evict a dirty page first
- 有效访问时间 Effective Access Time

  p: 缺页错误率 Page Fault Rate

  ma: 内存访问时间

$$EAT = ma + p \times Page\ Fault\ Time$$

注意这里课件和书上不一样，书上是 $(1-p)ma$，一个破公式就不能统一一下？？

9.4 页面置换 Page Replacement

9.4.1 Cache Miss 的分类

- Compulsory Miss (Cold Start Miss)
  - Pages that have never been paged into memory before
  - How might we remove these misses?
    - Prefetching: loading them into memory before needed
    - Need to predict future somehow!
- Capacity Miss
  - Not enough memory. Must somehow increase available memory size
    - Increase amount of DRAM
    - If multiple processes in memory: adjust percentage of memory allocated to each one
- Conflict Miss
  - Happen in direct mapped cache and set-associative cache
  - 刚把他踢走，接着又要访问他
  - Technically, conflict misses don't exist in virtual memory, since it is a "fully associative" cache
- Policy Miss
  - Caused when pages were in memory, but kicked out prematurely because of the replacement policy

9.4.2 FIFO 页面置换

- Throw out oldest page. Be fair let every page live in memory for sameamount of time.
- Bad: may throw out heavily used pages instead of infrequently used
- 理论实现：队列

```
class FIFOCache : public Cache {
    private:
        list<int> lst;

    public:
        FIFOCache(int size) : Cache(size) {}

        bool full() override {
            return lst.size()==this→size;
        }

        bool contains(int x) override {
```

```
            return find(lst.begin(), lst.end(), x)≠lst.end();
        }

        void insert(int x) override {
            if (this→contains(x)) {
                this→hitCount++;
            }
            else {
                if (this→full())
                    lst.pop_front();
                lst.push_back(x);
            }
        }
};
```

- 例: 3 个 frame，4 个 page，访问顺序: `A B C A B D A D B C B`

Ref: / Page:	A	B	C	A	B	D	A	D	B	C	B
1	A					D				C	
2		B					A				
3			C						B		

Page Fault: 7

- Bélády 异常 (Bélády's Anomaly)

  对于 FIFO 策略，Mem size 增大，page fault 有可能反而变多

Ref: / Page:	A	B	C	D	A	B	E	A	B	C	D	E
1	A			D			E					
2		B			A					C		
3			C			B					D	

Ref: / Page:	A	B	C	D	A	B	E	A	B	C	D	E
1	A						E				D	
2		B						A				E
3			C						B			
4				D						C		

9.4.3 最优页面置换 MIN

- 置换最长时间不会使用的页面 (farthest-in-the-future)
- Offline 算法，需要提前知道访问序列
- 理论天花板，实际不可能
- 理论实现: 好多实现方式

```
class MINCache : public Cache {
    private:
        set<int> s;
        priority_queue<pair<int, int>> pq; // <index, pagenumber>

        vector<int> pages;
        vector<int> nxt;

    public:
        MINCache(int size) : Cache(size) {}

        bool full() override {
            return s.size()==this->size;
        }

        bool contains(int x) override {
            return s.find(x)!=s.end();
        }

        void init(int n) {
            int pageNum;
            for (int i=0;i<n;i++) {
                cin>>pageNum;
                pages.push_back(pageNum);
                nxt.push_back(INT_MAX);
            }

            map<int, int> temp;
            for (int i=n-1;i>=0;i--) {
                map<int, int>::iterator p=temp.find(pages[i]);
                if (p!=temp.end())
                    nxt[i]=p->second;
                temp[pages[i]]=i;
            }
        }

        void insert(int i) override {
            if (this->contains(pages[i]))
                this->hitCount++;
            else {
                if (this->full()) {
                    auto t=pq.top();
```

```
                    pq.pop();
                    s.erase(t.second);
                }

                s.insert(pages[i]);
            }

            pq.push(make_pair(nxt[i], pages[i]));
        }

        void start() {
            for (int i=0;i<pages.size();i++)
                this->insert(i);
        }
};
```

- 例：`A B C A B D A D B C B`

Ref: / Page:	A	B	C	A	B	D	A	D	B	C	B
1	A									C	
2		B									
3			C			D					

Page Fault: 5

9.4.4 LRU 页面置换

- 最近最少使用 Least Recently Used
- Replace page that has not been used for the longest time
- 理论实现：双向链表
  - 如果 hit，把他删除然后插回 tail，O(1) 的操作
  - 没 hit，pop head 然后插到 tail
  - 就是查询是否 hit 的时候需要遍历链表，不过可以开一个 set 专门用来查询 (空间换时间)
  - 实际上这个对于硬件来说太复杂了，无法实现

```
class LRUCache : public Cache {
    private:
        list<int> lst;

    public:
        LRUCache(int size) : Cache(size) {}

        bool full() override {
            return lst.size()==this->size;
        }
```

```
11
12          list<int>::iterator findElement(int x) {
13              return find(lst.begin(), lst.end(), x);
14          }
15
16          bool contains(int x) override {
17              return this→findElement(x)≠lst.end();
18          }
19
20          void insert(int x) override {
21              list<int>::iterator it=findElement(x);
22
23              if (it≠lst.end()) {
24                  this→hitCount++;
25                  lst.erase(it);
26                  lst.push_back(x);
27              }
28              else {
29                  if (this→full())
30                      lst.pop_front();
31                  lst.push_back(x);
32              }
33          }
34  };
```

- 例: `A B C D A B C D A B C D`

Ref: / Page:	A	B	C	D	A	B	C	D	A	B	C	D
1	A			D			C			B		
2		B			A			D			C	
3			C			B			A			D

Page Fault: 全是

9.4.5 近似 LRU 页面置换

9.4.5.1 时钟算法 Clock Algorithm

- Arrange physical pages in circle with single clock hand
- Hardware "use" bit per physical page
  - Hardware sets use bit on each reference
  - If use bit is not set, means not referenced in a long time
- On page fault
  - Advance clock hand
  - Check use bit

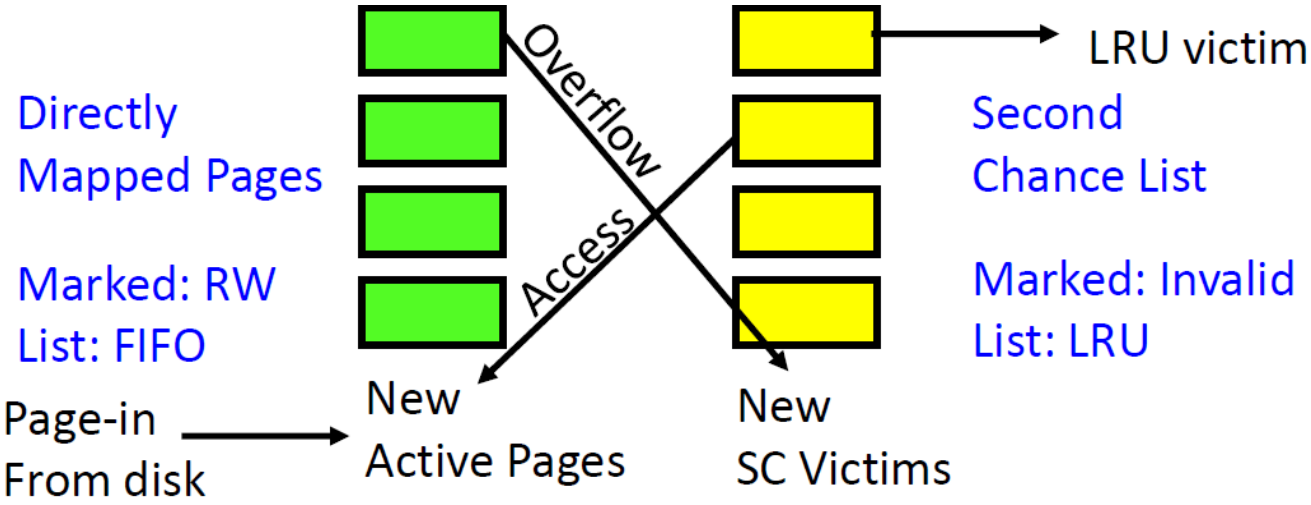
    - 1: used recently; clear (set used bit to 0) and leave alone (go next)
    - 0: selected candidate for replacement
- 最差情况：转了一圈，1 全变 0，又回到最开始 (FIFO)
- hit 的时候时针是不转的
- 稳定时，时针永远指向刚被替换的下一个位置
- 理论实现

```
class ClockCache : public Cache {
    private:
        vector<pair<int, int>> vec;
        int ptr=0;

    public:
        ClockCache(int size) : Cache(size) {}

        bool full() override {
            return vec.size()==this->size;
        }

        int findElement(int x) {
            for (int i=0;i<vec.size();i++)
                if (vec[i].first==x)
                    return i;
            return -1;
        }

        bool contains(int x) override {
            return this->findElement(x)>=0;
        }

        void insert(int x) override {
            int index=this->findElement(x);

            if (index>=0) {
                vec[index].second=1;
                this->hitCount++;
            }
            else {
                if (this->full()) {
                    while (vec[ptr].second!=0) {
                        vec[ptr].second=0;
                        ptr=(ptr+1)%this->size;
                    }
                    vec[ptr]=make_pair(x, 1);
                }
                else
                    vec.push_back(make_pair(x, 1));
                ptr=(ptr+1)%this->size;
            }
        }
```

```
44 };
```

- N^{th} chance algorithm: Used bit 变成一个 counter，从 N 开始倒计
  - N=1K 比较合适
  - N 越小越快
  - Clean pages, use N=1
  - Dirty pages, use N=2 (and write back to disk when N=1)

9.4.5.2 Second Chance List Algorithm



- Split memory in two list
  - Active List
  - Second Chance List
- Access pages in Active list at full speed
- Page Fault
  - Always move overflow page from end of Active list to front of Second chance list (SC) and mark invalid
  - Desired Page On SC List: move to front of Active list, mark RW
  - Not on SC list: page in to front of Active list, mark RW; page out LRU victim at end of SC list

---

9.5 帧分配 Frame Allocation

9.5.1 全局分配与局部分配

- 全局置换 Global Replacement

  process selects replacement frame from set of all frames; one process can take a frame from another

- 局部置换 Local Replacement

  each process selects from only its own set of allocated frames

9.5.2 分配算法

- Equal Allocation (Fixed Scheme)
  - Every process gets same amount of memory
  - Example: 100 frames, 5 processes -> each process gets 20 frames
- Proportional Allocation (Fixed Scheme)

  Allocate according to the size of process

  设 s_i = size of process p_i

  $S = \sum s_i$

  m = total # of frames

  Allocation for p_i is $a_i = \frac{s_i}{S} \cdot m$
- Priority Allocation
  - Proportional scheme using priorities rather than size
    - Same type of computation as previous scheme
  - Possible behavior: If process p_i generates a page fault, select for replacement a frame from a process with lower priority number

---

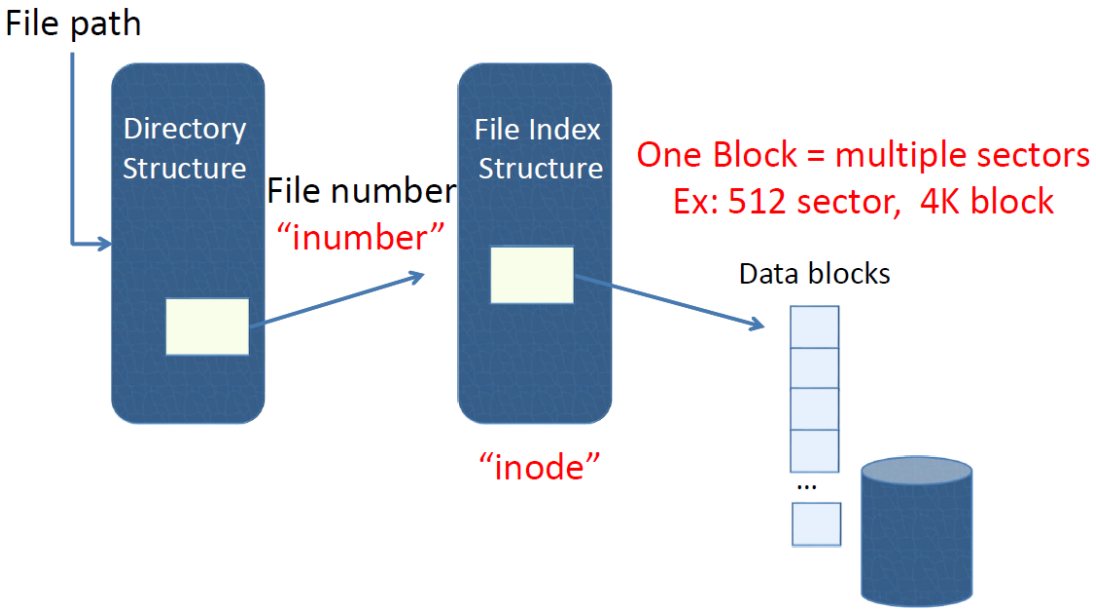
第十 & 十一章 文件系统 File System

10.1 文件系统概念

- 文件系统 File System

  Layer of OS that transforms block interface of disks (or other block devices) into Files, Directories, etc.

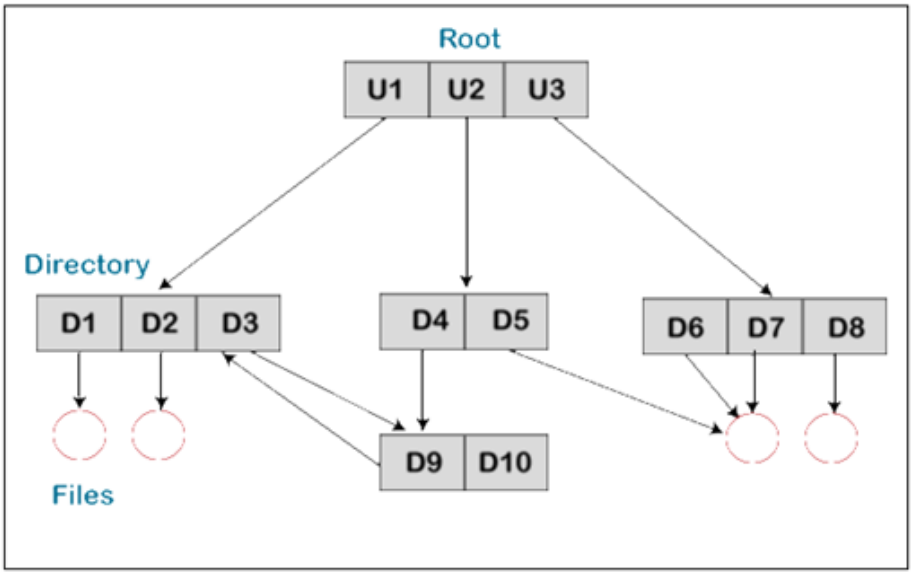
- 文件系统的功能
  - Naming : Interface to find files by name, not by blocks
  - Disk Management: Collecting disk blocks into files
  - Protection: Layers to keep data secure
  - Reliability/Durability: Keeping of files durable despite crashes, media failures, attacks, etc.



10.2 文件和目录 Files and Directories

10.2.1 目录的组成

- 多级继承结构 Hierarchical Structure
- Each directory entry is a collection of
  - Files
  - Directories: A link to another entries
- Each has a name and attributes
- Links (hard links) make it a DAG, not just a tree



10.2.2 文件

- 文件 File

  操作系统对存储设备的物理属性加以抽象，从而定义逻辑存储单位

  Named permanent storage
- 文件的组成
  - Data

    Blocks on disk somewhere
  - Metadata (Attributes)
    - Owner, size, last opened, ⋯
    - Access rights
      - R, W, X
      - Owner, Group, Other (in Unix systems)
      - Access control list in Windows system

10.3 磁盘管理策略

- Basic entities on a disk
  - File

    User visible group of blocks arranged sequentially in logical space
  - Directory

    User visible index mapping names to files
- Access disk as linear array of sectors
  - Identify sectors as vectors [cylinder, surface, sector], sort in cylinder major order: **not used anymore**

    - Logical Block Addressing (LBA): Every sector has integer address from zero up to max number of sectors
    - Controller translates from address to physical position
- Need way to track free disk blocks
    - Link free blocks together: too slow today
    - Use bitmap to represent free space on disk
- Need way to structure files: File header
    - Track which blocks belong at which offsets within the logical file structure
    - Optimize placement of files' disk blocks to match access and usage patterns

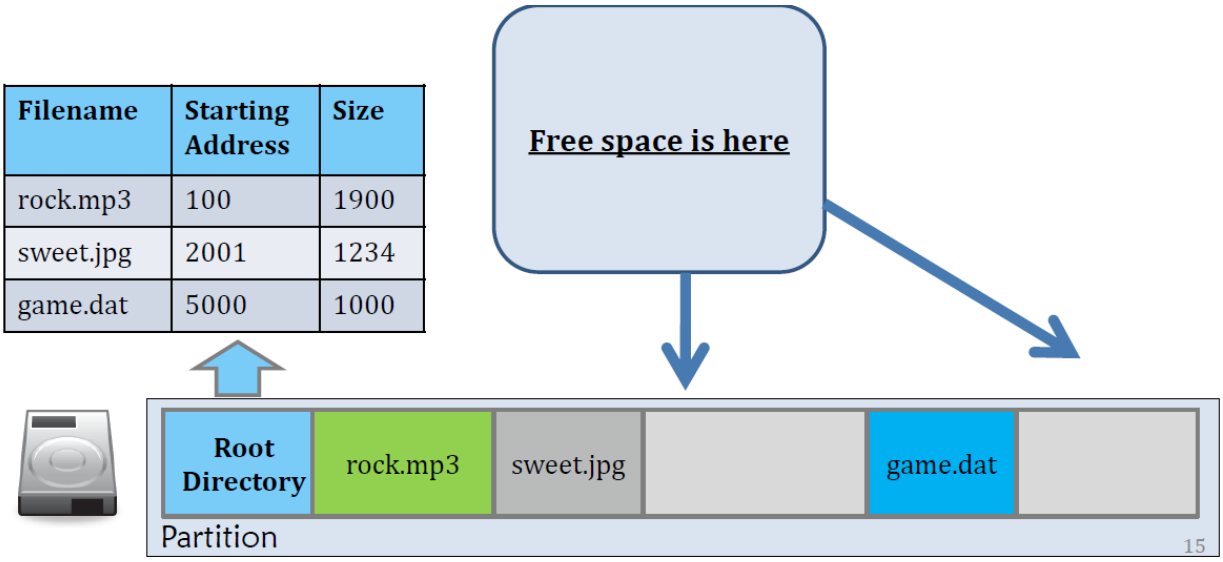
---

10.4 目录分配

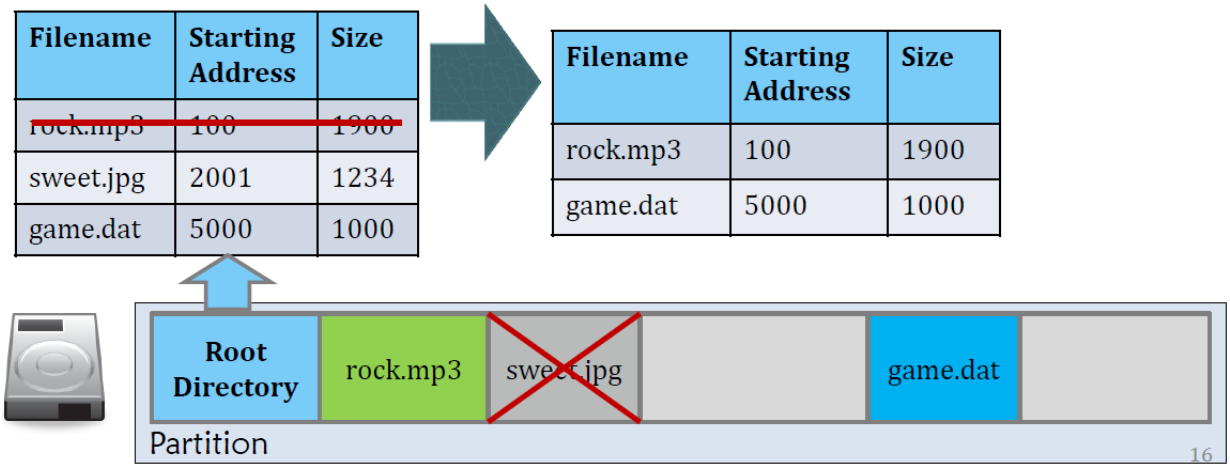
10.4.1 连续分配 Contiguous Allocation

- 连续分配方法要求每个文件在磁盘上占有一组连续的块。磁盘地址为磁盘定义了一个线性排序。有了这个排序，假设只有一个作业正在访问磁盘，在块 b 之后访问块 b+1 时通常不需要移动磁头。当需要磁头移动（从一个柱面的最后扇区到下一个柱面的第一个扇区时），只需要移动一个磁道
- Locate Files

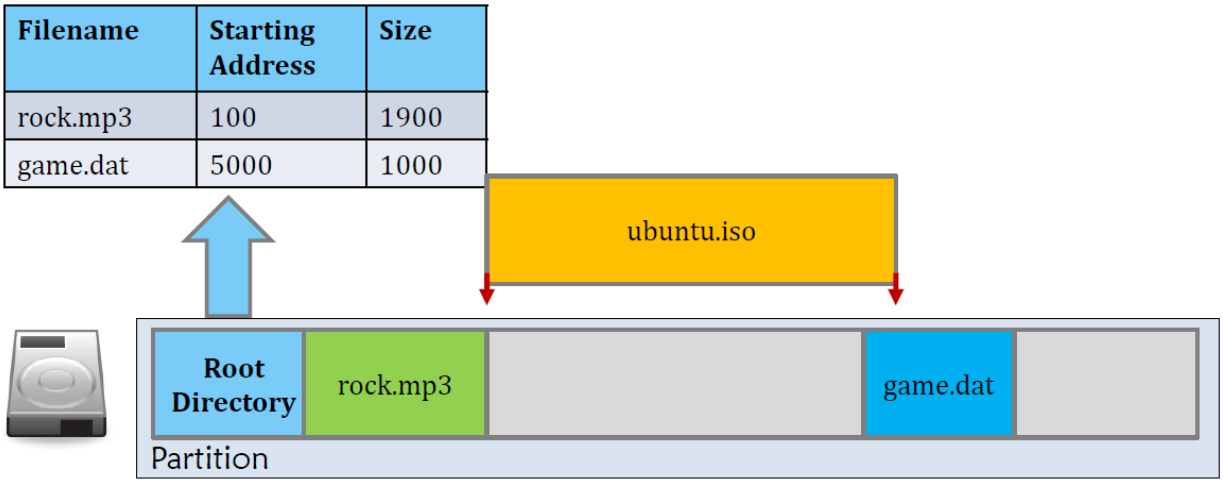
  Start address and size -> easy to random access



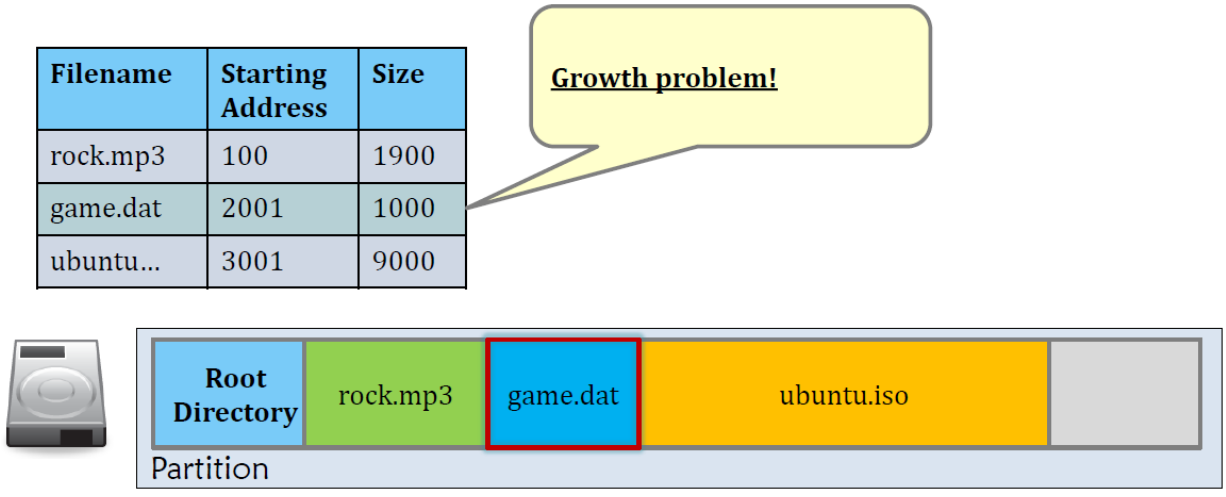
- Delete Files



- 缺点：外碎片 External Fragmentation



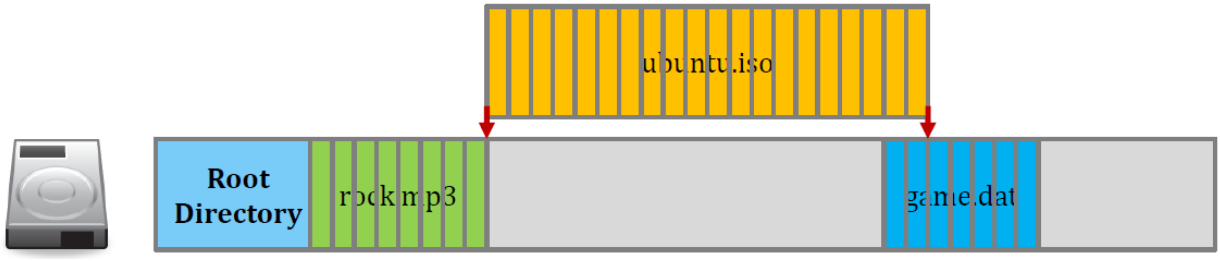
- 缺点：无法应对文件增长 File Growth Problem



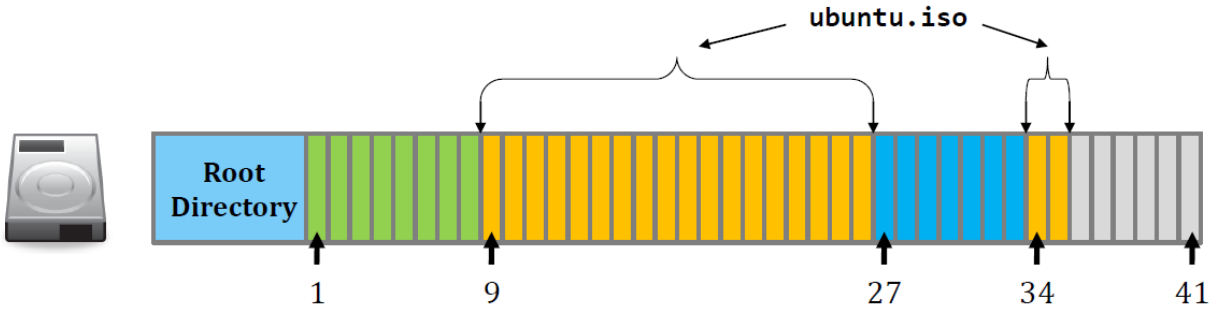
- 连续分配的应用
  - ISO 9660
  - CD-ROM

10.4.2 链接分配 Linked Allocation

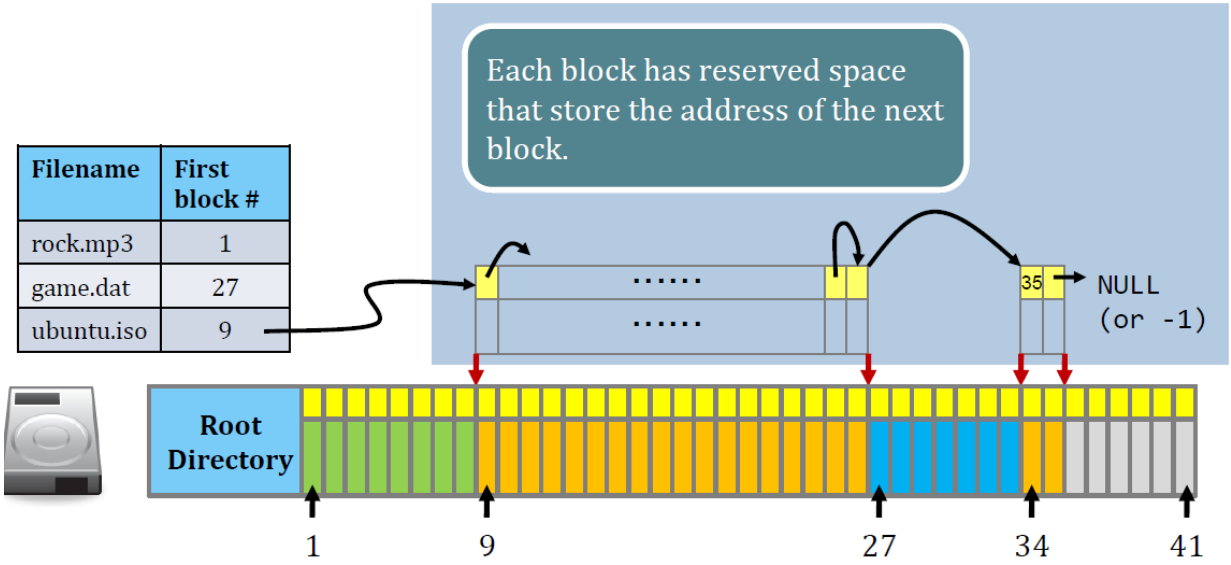
1. Chop the storage device into equal sized blocks



2. Fill the empty space in a block by block manner

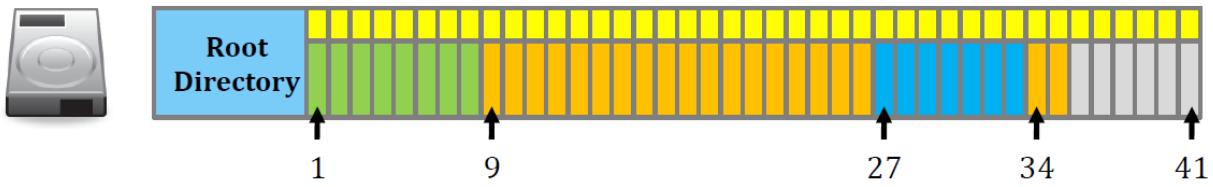


3. Leave 4 bytes from each block as the pointer



4. Keep the file size in the root directory table

Filename	First block #	Size
rock.mp3	1	1900
game.dat	27	1000
ubuntu.iso	9	9000



- 缺点：内碎片 Internal Fragmentation

  The last block of a file may not be fully filled

- 缺点：随机访问性能差

  What if I want to access the 2019-th block of ubuntu.iso?

You have to access blocks 1~2018 of ubuntu.iso until the 2019-th block.

- 优点
  - 没有外碎片
  - 解决了文件增长的问题

10.4.3 索引分配 Index Allocation

- 索引分配通过将所有指针放在一起，即索引块 (Index Block)，实现了高效的随机访问
- 索引块的组织方式
  - 链接方案

    一个索引块通常为一个磁盘块，本身可以读写，通过链表的形式存放大文件
  - 多级索引

    通过第一级索引块指向一组第二级索引块，它又指向文件块
  - 组合方案

    iNode

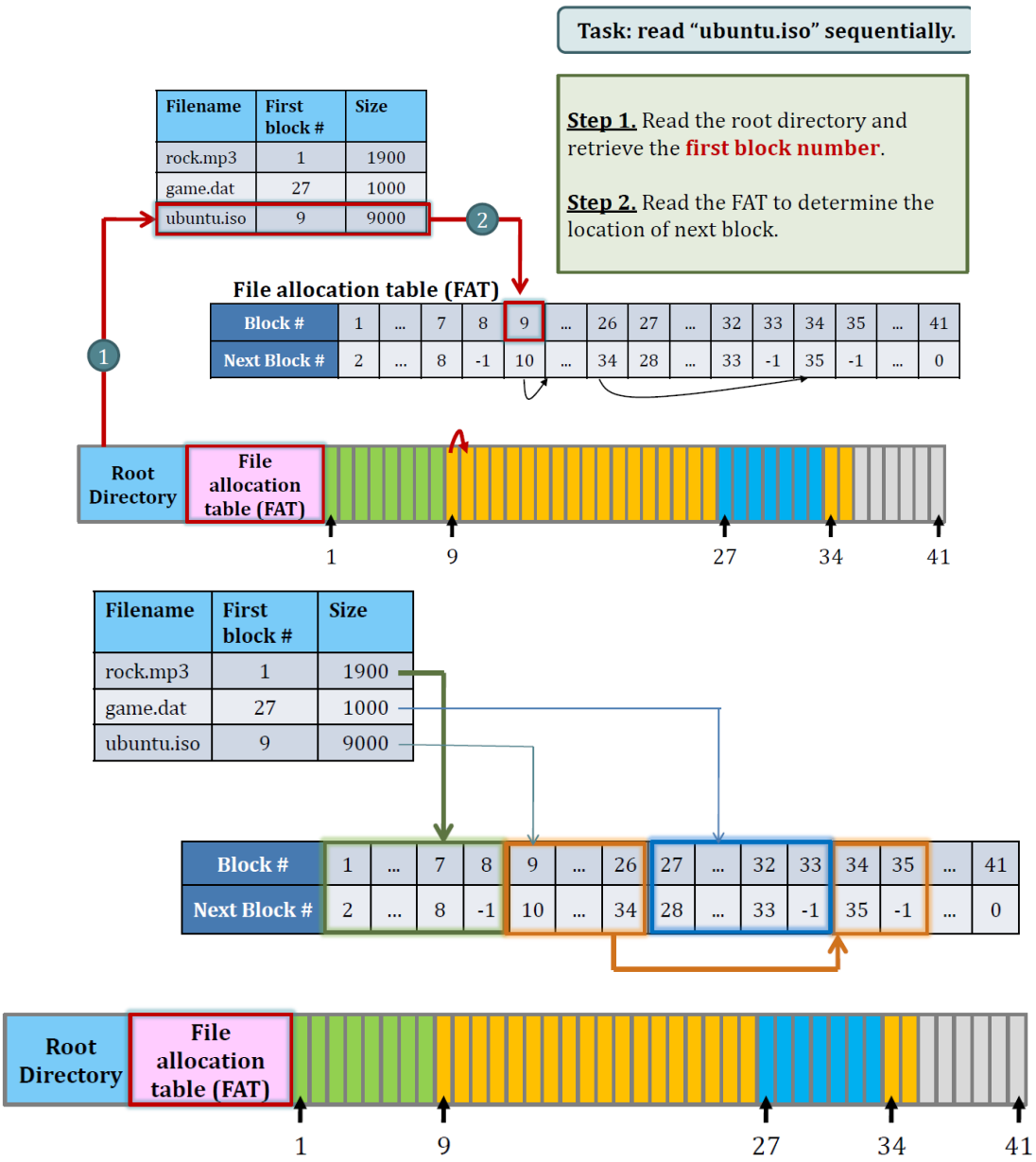
10.5 文件分配表 FAT

10.5.1 FAT 的原理

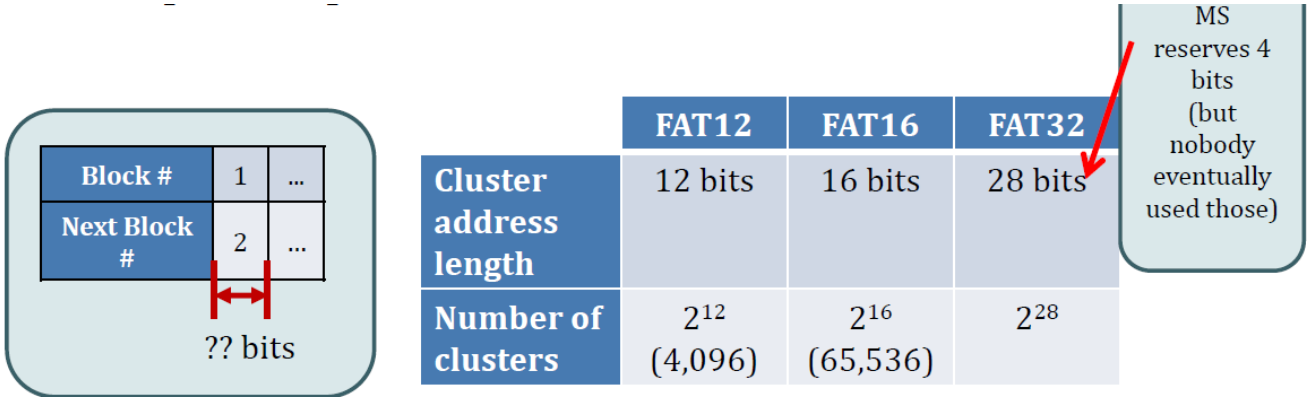
- 文件分配表 File Allocation Table

  Centralize all the block links as FAT

- FAT 相当于一个 next 数组



10.5.2 FAT 文件系统的大小

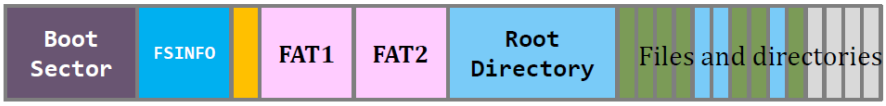


- 在 DOS 里 block 被称为 cluster
- 假设 block size = 32 KB，那么对于 FAT32 来说，它支持的文件总大小为 $32 * 2^{10} * 2^{28} = 2^{43} = 8$ TB
- 然而，微软为了催人用 NTFS，手动把 FAT 文件系统大小的上限设为了 32 GB
  如果你有一个 64 GB 的 U 盘，在 FAT 文件系统下你最多只能往里存 32 GB

- 但是你可以不用微软的工具格式化，用别的手段把它格式化成 FAT 就没有这个限制了

10.5.3 FAT 文件系统结构

	Propose	Size
Reserved sectors: Boot sector	FS-specific parameters	1 sector, 512 bytes
Reserved sectors: FSINFO	Free-space management	1 sector, 512 bytes
Reserved sectors: More reserved sectors	Optional	Variable, can be changed during formatting
FAT (2 pieces)	1 copy as backup	Variable, depends on disk size and cluster size.
Root directory	Start of the directory tree.	At least one cluster, depend on the number of directory entries.



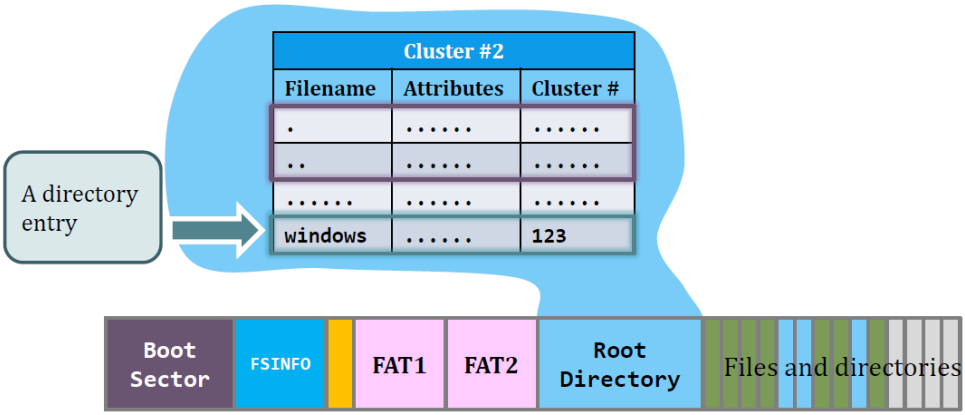
10.5.4 FAT 文件遍历

例: `dir c:\windows`

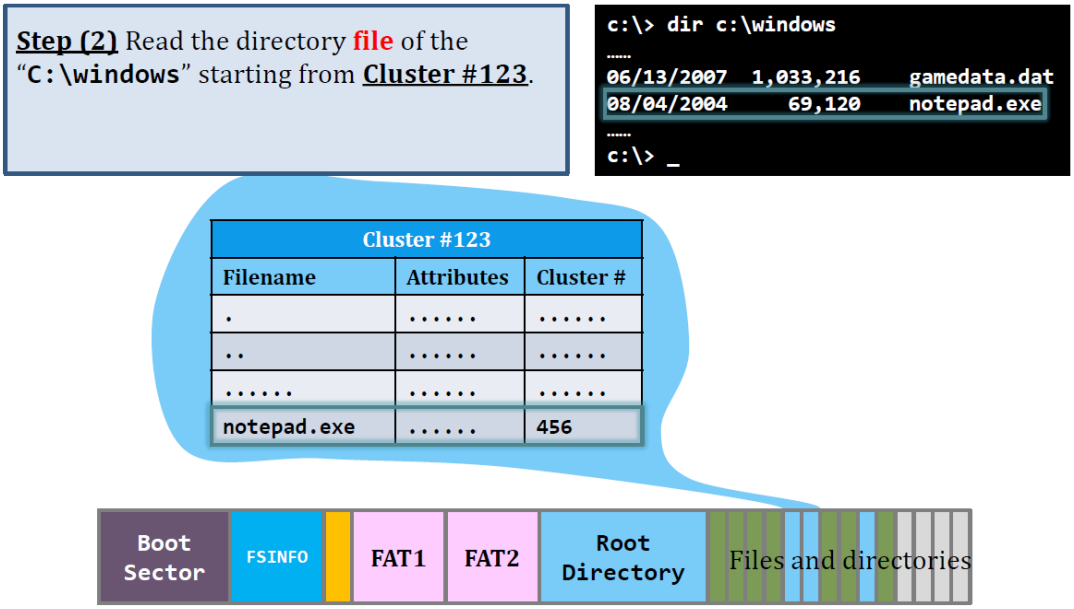
Step (1) Read the directory file of the root directory starting from **Cluster #2**.

"`C:\windows`" starts from Cluster #123.

```
c:\> dir c:\windows
……
06/13/2007  1,033,216    gamedata.dat
08/04/2004     69,120    notepad.exe
……
c:\> _
```



- 为什么是 cluster 2



10.5.5 FAT Directory Entry

- A 32 byte directory entry in a directory file
- A directory entry is describing a file (or a sub directory) under a particular directory

Bytes	Description
0-0	1st character of the filename (0x00 or 0xe5 means unallocated)
1-10	remaining characters of filename + extension.
11-11	File attributes (e.g., read only, hidden)
12-12	Reserved.
13-19	Creation and access time information.
20-21	High 2 bytes of the first cluster address (0 for FAT16 and FAT12).
22-25	Written time information.
26-27	Low 2 bytes of first cluster address.
28-31	File size.

Filename	Attributes	Cluster #
explorer.dat	.....	32

0	e	x	p	l	o	r	e	r	7
8	e	x	e	…	…	…	…	…	15
16	…	…	…	…	00	00	…	…	23
24	…	…	20	00	00	C4	0F	00	31

Note. This is the 8+3 naming convention.

8 characters for name +
3 characters for file extension

- 文件名被规范成了 8+3 的格式
- FAT32 的最大文件大小

  4G - 1 bytes

Bytes	Description
0-0	1st character of the filename (0x00 or 0xe5 means unallocated)
1-10	7+3 characters of filename + extension.
11-11	File attributes (e.g., read only, hidden)
12-12	Reserved.
13-19	Creation and access time information.
20-21	High 2 bytes of the first cluster address (0 for FAT16 and FAT12).
22-25	Written time information.
26-27	Low 2 bytes of first cluster address.
28-31	File size.

Filename	Attributes	Cluster #
explorer.dat	.....	32

0	e	x	p	l	o	r	e	r	7
8	e	x	e	…	…	…	…	…	15
16	…	…	…	…	00	00	…	…	23
24	…	…	20	00	00	C4	0F	00	31

So, what is the largest size of a FAT32 file?

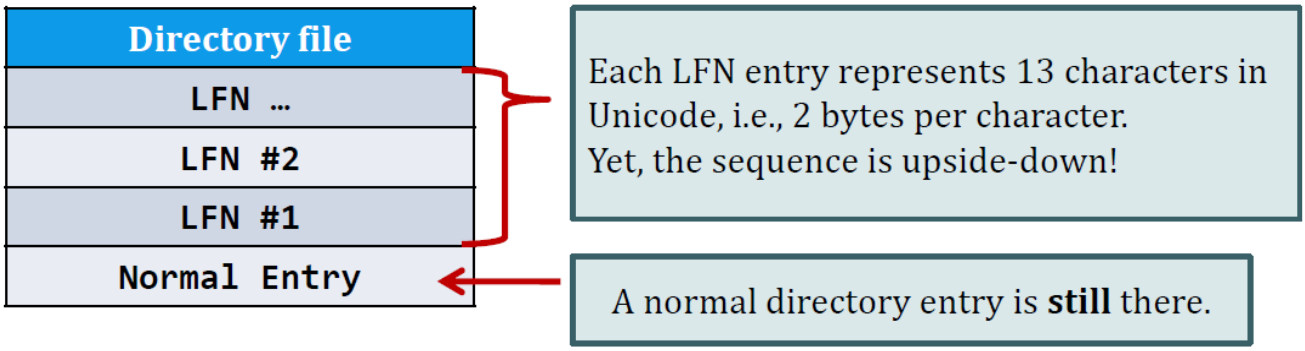
4G - 1 bytes

Bounded by the file size attribute!

Why "– 1"?

- Imagine 3 bits: 000, 001, ..., 110, 111
- Largest number is 111 = 2^3-1
- i.e., we also need to represent "0 bytes"

- 长文件名 Long File Name (LFN) Directory Entry

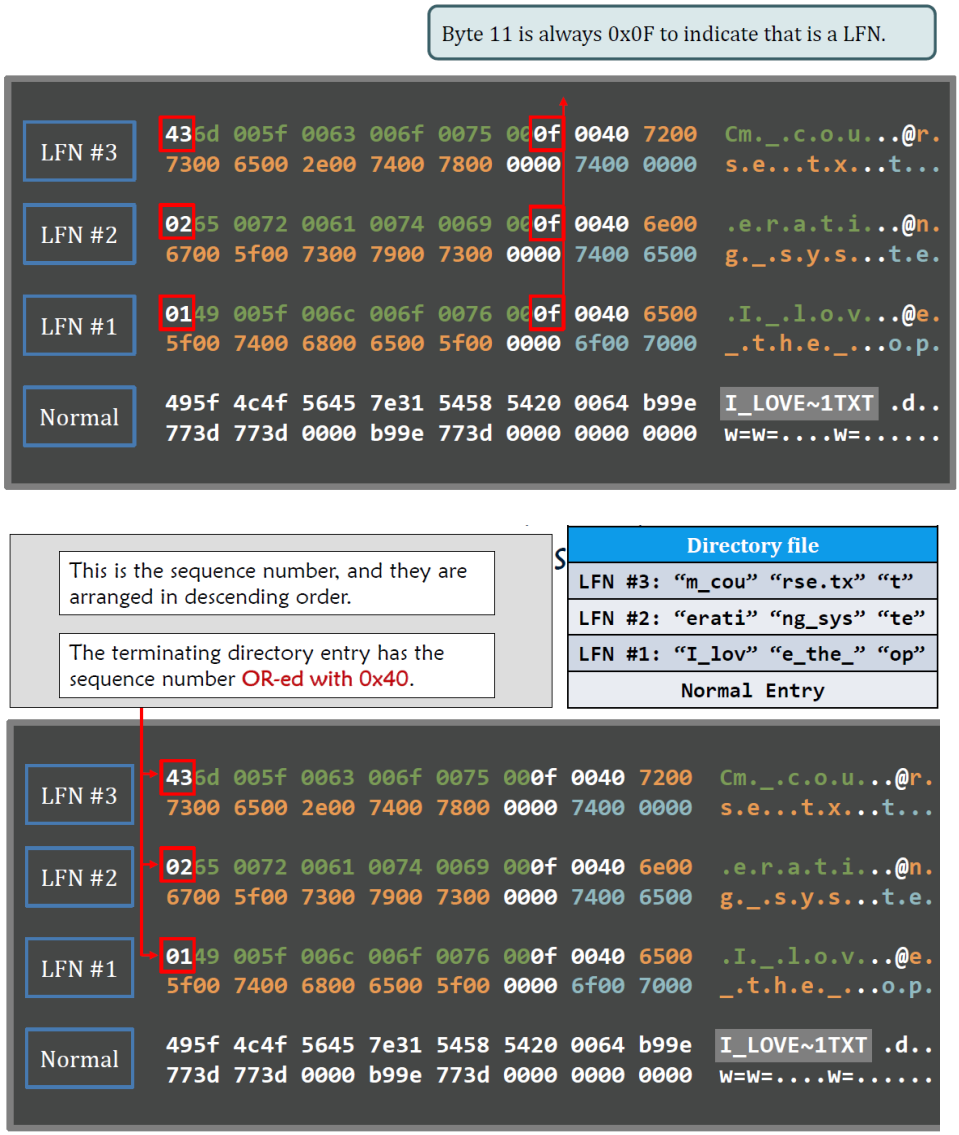


- Normal directory entry vs LFN directory entry

Bytes	Description
0-0	1st character of the filename (0x00 or 0xe5 means unallocated)
1-10	7+3 characters of filename + extension.
11-11	File attributes (e.g., read only, hidden)
12-12	Reserved.
13-19	Creation and access time information.
20-21	High 2 bytes of the first cluster address (0 for FAT16 and FAT12).
22-25	Written time information.
26-27	Low 2 bytes of first cluster address.
28-31	File size.

Bytes	Description
0-0	Sequence Number
1-10	File name characters (5 characters in Unicode)
11-11	File attributes - always 0x0F (to indicate it is a LFN)
12-12	Reserved.
13-13	Checksum
14-25	File name characters (6 characters in Unicode)
26-27	Reserved
28-31	File name characters (2 characters in Unicode)

- Filename:
  “**I_love_the_operating_system_course.txt**”.

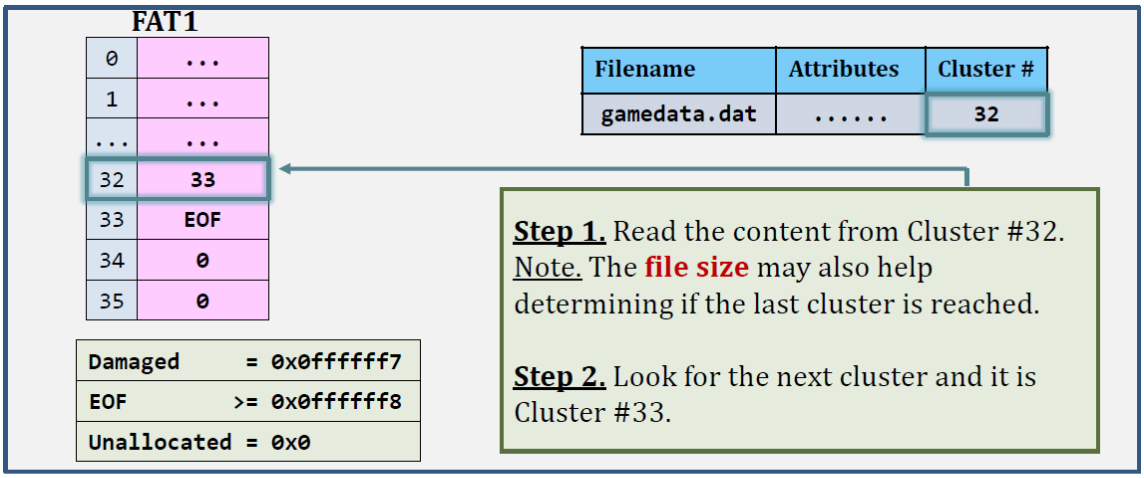


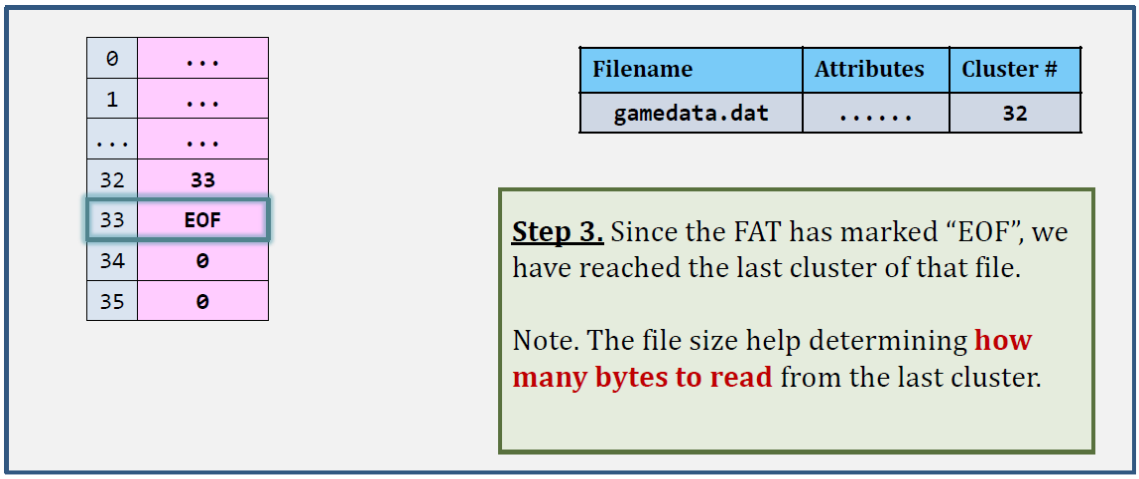
- Directory entry is important
  - It stores the start cluster number.
  - It stores the file size

    Without the file size, how can you know when you should stop reading a cluster?
  - It stores all file attributes

10.5.6 FAT 读文件

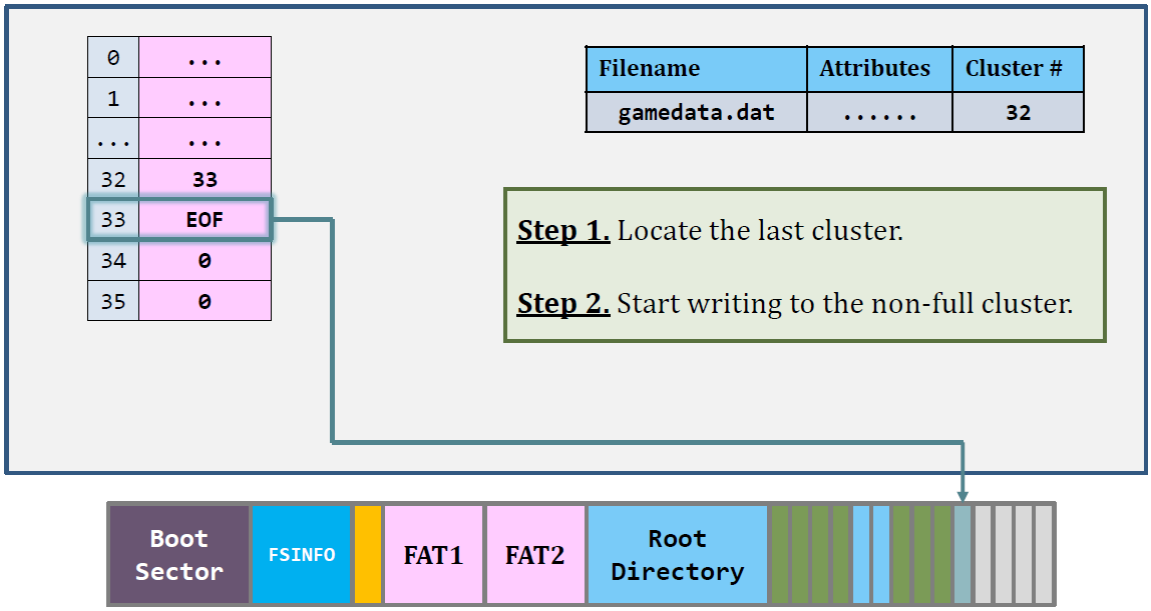
例: 顺序读取 `C:\windows\gamedata.dat`

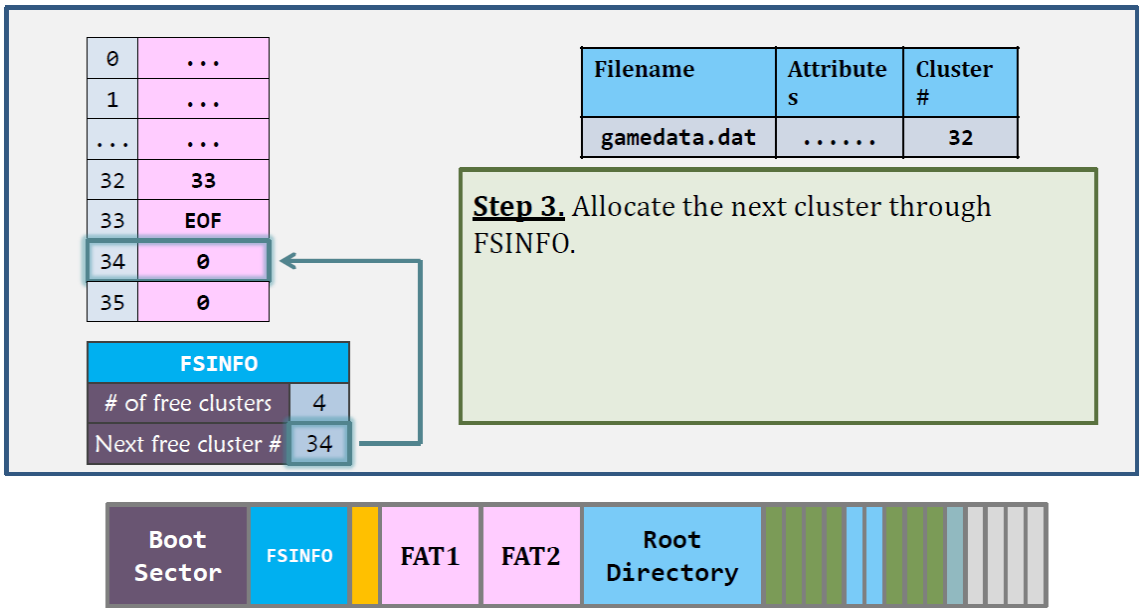


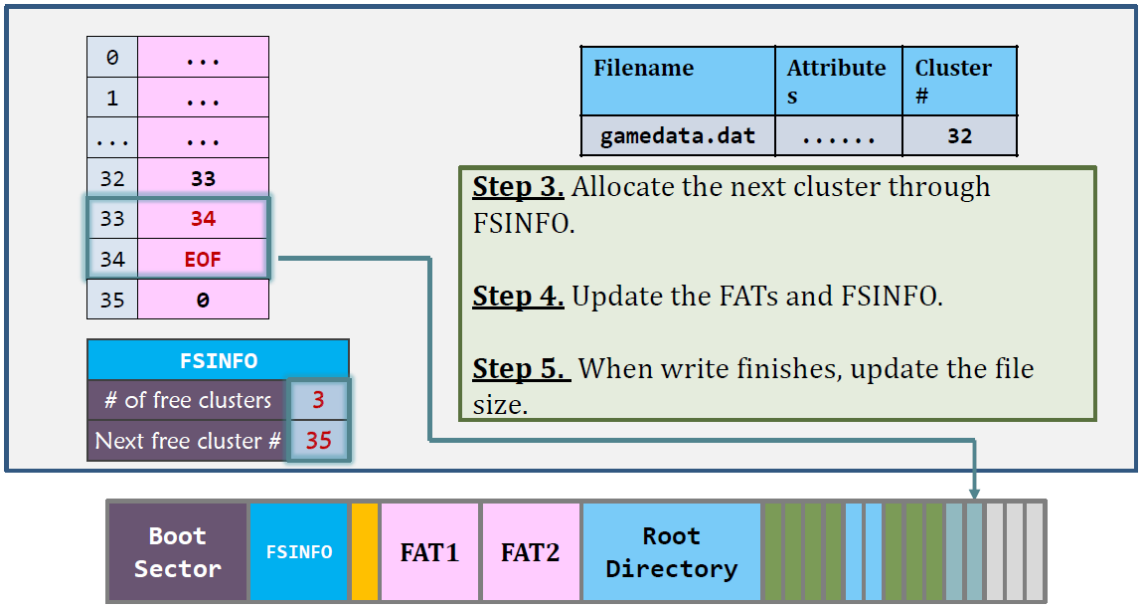


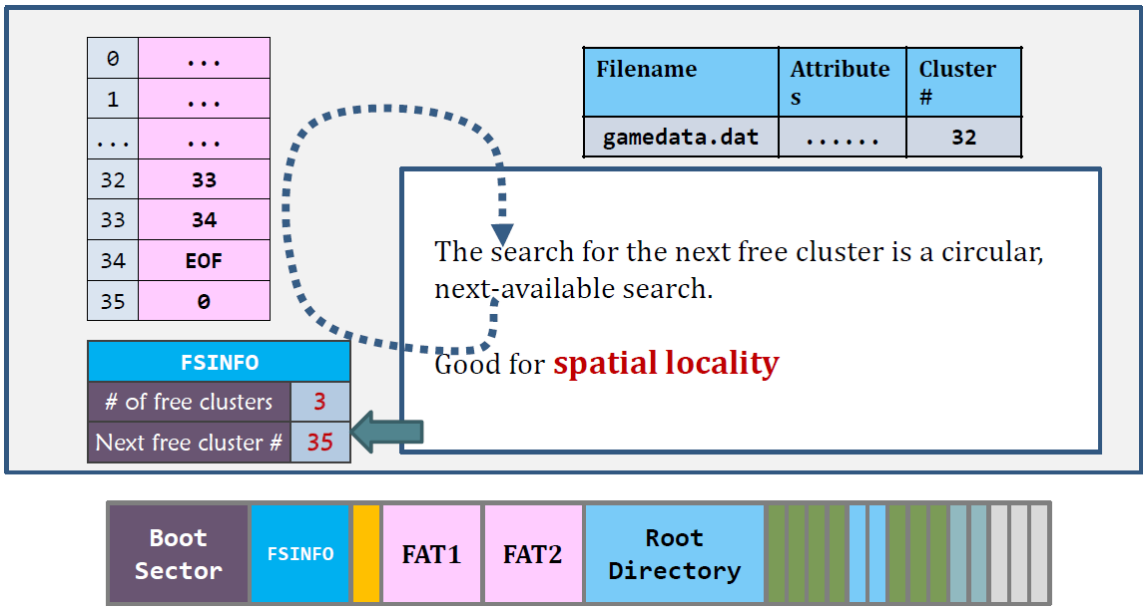
10.5.7 FAT 写文件

例: 向 `C:\windows\gamedata.dat` 写入数据


0 ...
1 ...
... ...
32 33
33 EOF
34 0
35 0
Filename
Attributes
Cluster #
gamedata.dat
.....
32
Step 1. Locate the last cluster.
Step 2. Start writing to the non-full cluster.
Boot Sector
FSINFO
FAT1
FAT2
Root Directory

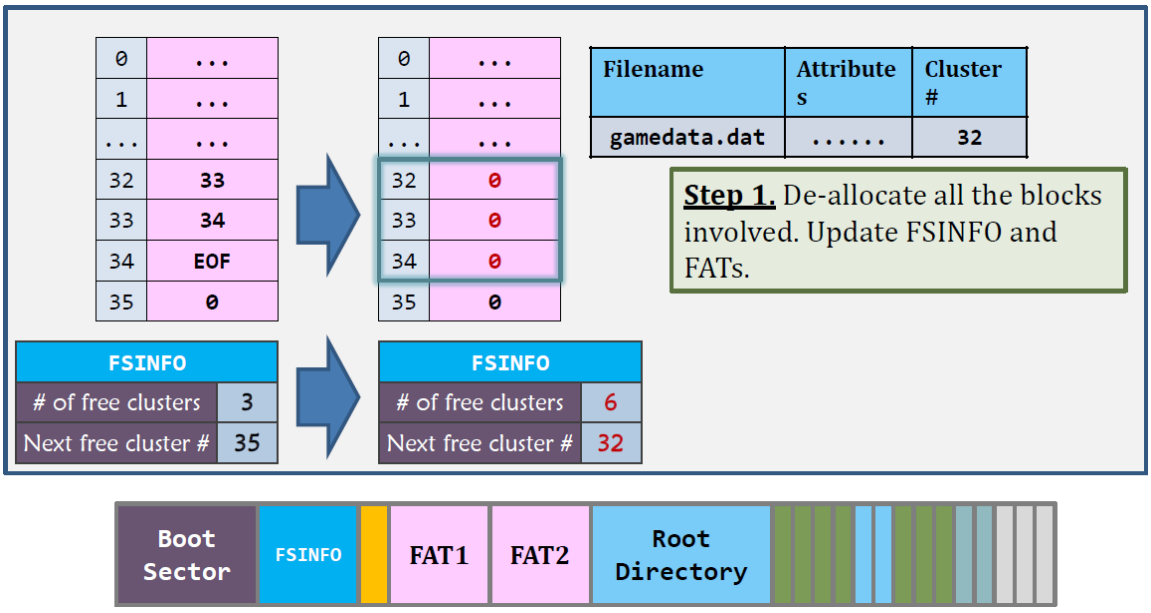

0 ...
1 ...
... ...
32 33
33 EOF
34 0
35 0
FSINFO
# of free clusters 4
Next free cluster # 34
Filename
Attributes
Cluster #
gamedata.dat
.....
32
Step 3. Allocate the next cluster through FSINFO.
Boot Sector
FSINFO
FAT1
FAT2
Root Directory

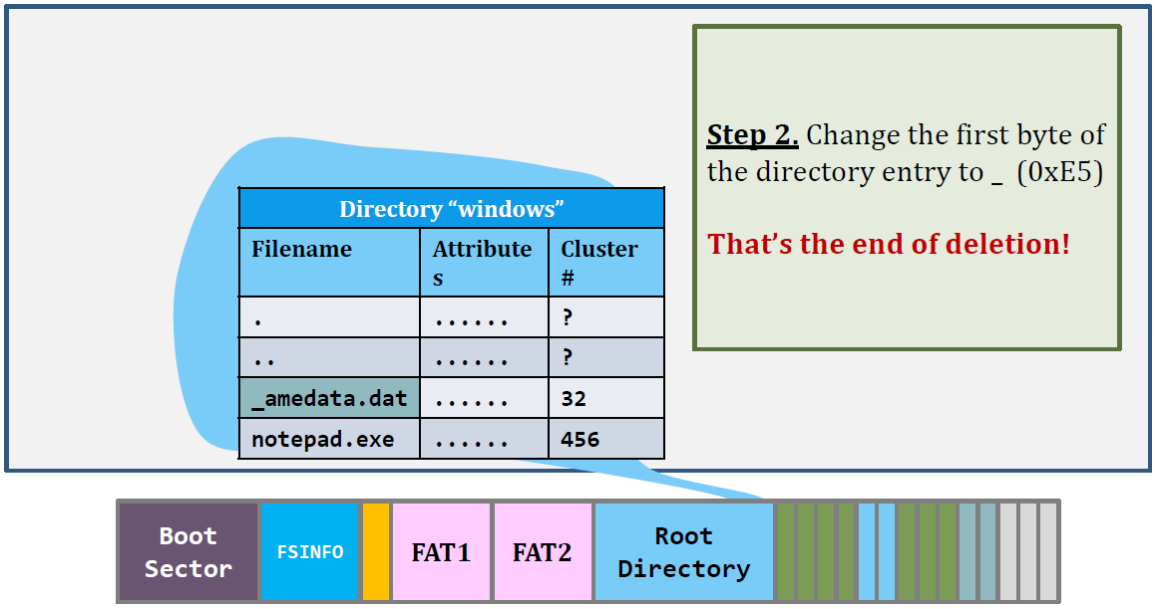

0 ...
1 ...
... ...
32 33
33 34
34 EOF
35 0
FSINFO
# of free clusters 3
Next free cluster # 35
Filename
Attributes
Cluster #
gamedata.dat
.....
32
Step 3. Allocate the next cluster through FSINFO.
Step 4. Update the FATs and FSINFO.
Step 5. When write finishes, update the file size.
Boot Sector
FSINFO
FAT1
FAT2
Root Directory



10.5.8 FAT 删除文件

例: 删除 `C:\windows\gamedata.dat`





- 实际上数据还在硬盘中，直到 de-allocated 的 cluster 被再次使用（覆盖）
- 所以会产生安全问题，如果删除的数据还没被覆盖，就能通过其他手段获取
- Secure Disk Erase

  一种简单的办法是把释放出的 cluster 直接全写成 0
- 数据恢复

  既然还在硬盘里，就可以在没被覆盖之前恢复

  首先要抓紧拔电源，然后拿下硬盘

File size is within one block (cluster)	Because **the first cluster address** in the direct is still readable, the recovery is having a very high successful rate.
File size spans more than 1 block	Because of the next-available search, clusters of a file are likely to be contiguous allocated. This provides a hint in looking for deleted blocks. Can you devise an undelete algorithm for FAT32?

10.5.9 总结

- Space efficient:
  - 4 bytes overhead (FAT entry) per data cluster.
- Delete
  - Lazy delete efficient
  - Insecure

    designed for single user 20+ years ago
- Deployment: (FAT32 and FAT12)

  It is everywhere: CF cards, SD cards, USB drives

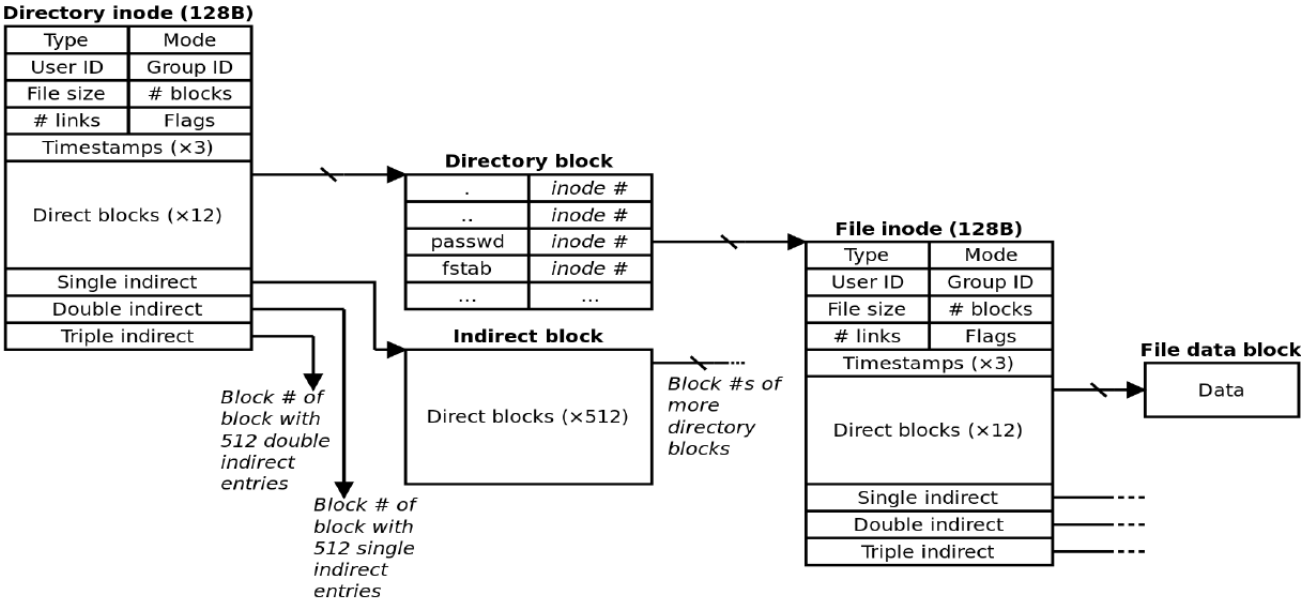
- Search
  - Block addresses of a file may scatter discontinuously
  - To locate the 888-th block of a file?

    Start from the first FAT entry and follow 888 pointers
- The most commonly used file system in the world

10.6 iNode

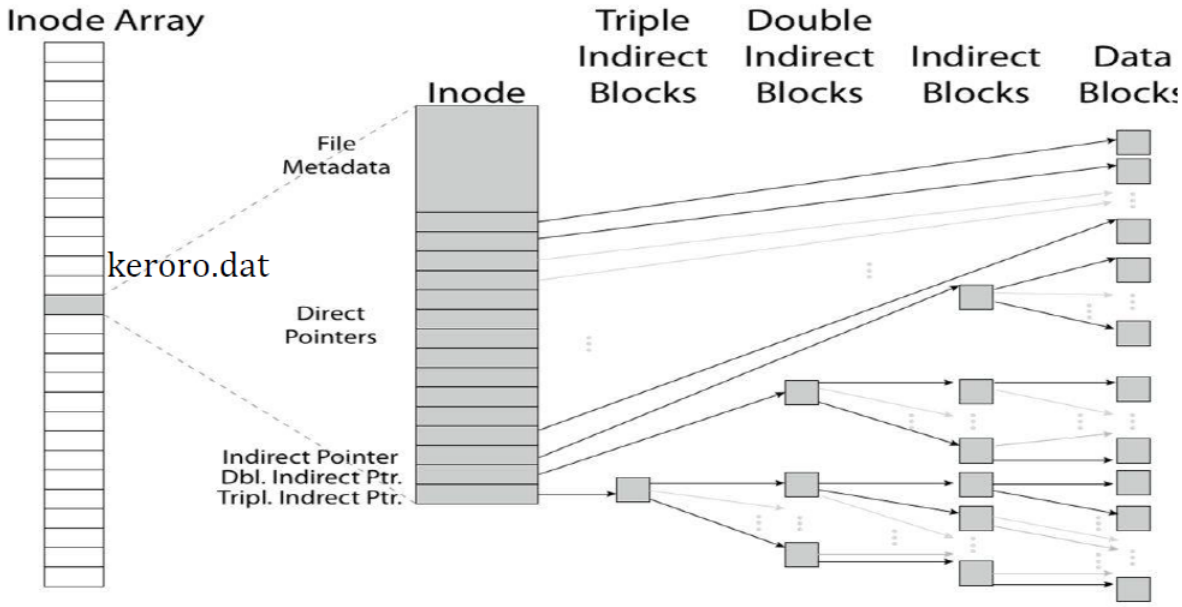
10.6.1 iNode 的原理

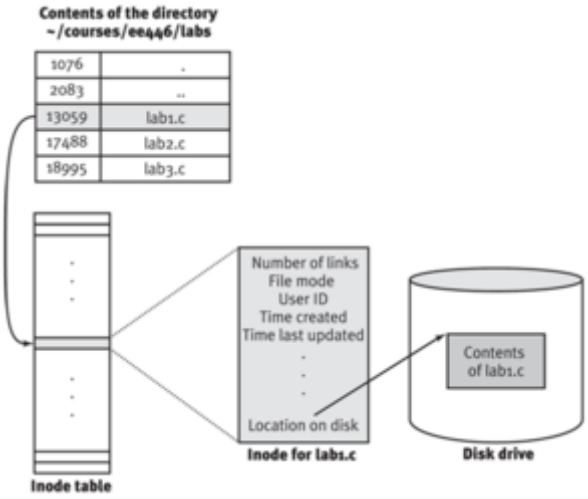
- All pointers of a file are located together
- One directory/file has one iNode

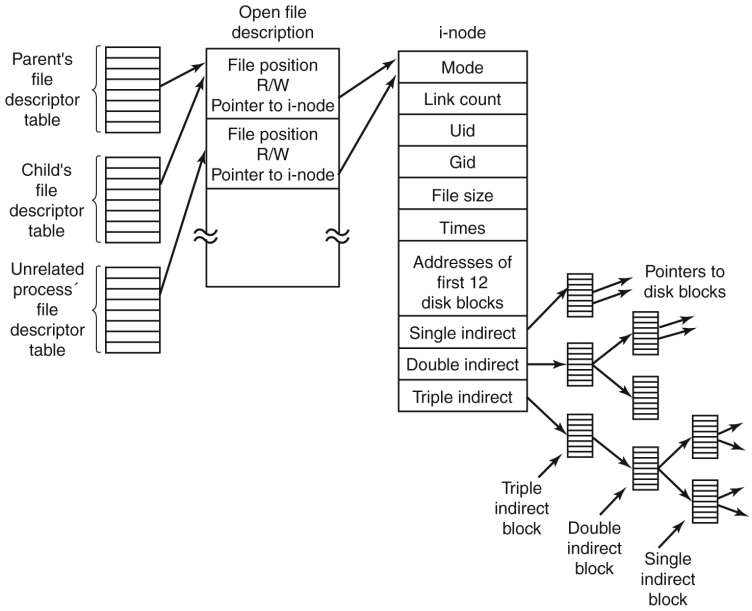


- iNode Table

  An array of iNodes
- Pointers are unbalanced tree based

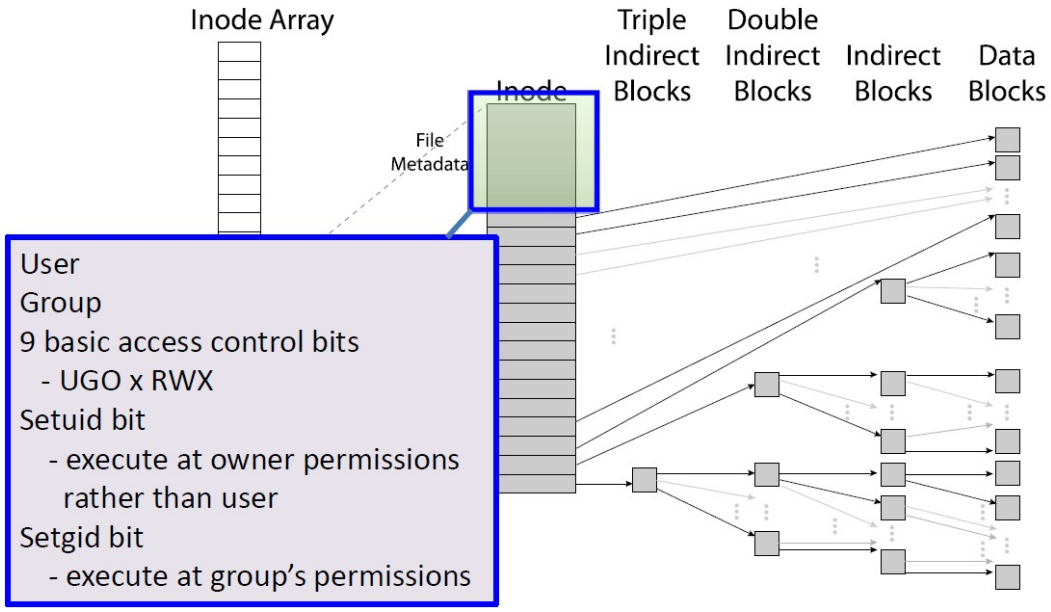




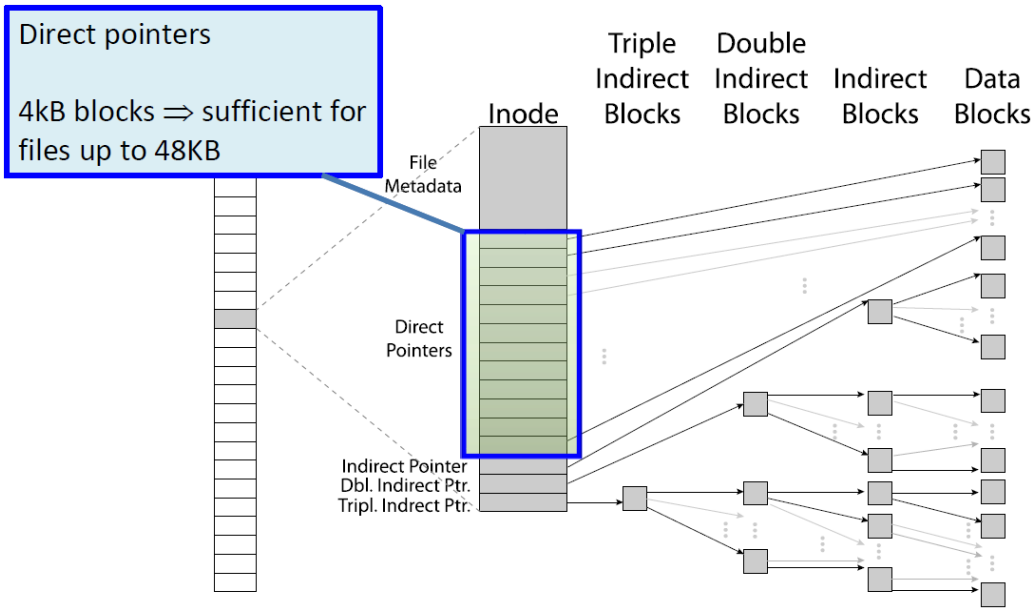


10.6.2 iNode 的结构

- iNode Metadata

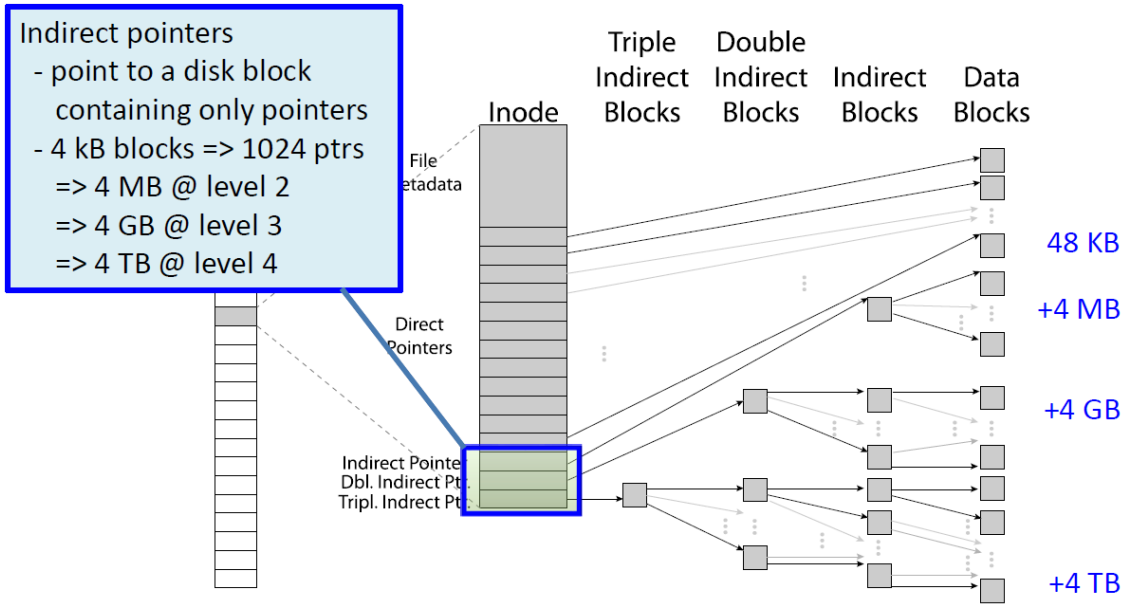


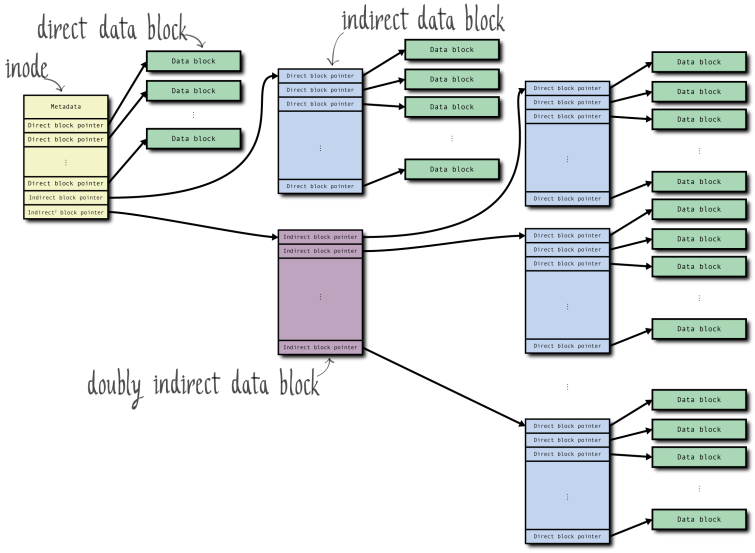
- 12 direct pointers



- Indirect Pointers

  一个 pointer 4 个 bytes (实际上就是个 32 bit 的地址)





10.6.3 iNode 文件大小

Number of direct blocks	12
Number of indirect blocks	1
Number of double indirect blocks	1
Number of triple indirect blocks	1
Block size	2^x bytes
Address length	4 bytes

$$12 \times 2^x + 1 \times 2^x/4 \times 2^x + 1 \times (2^x/4)^2 \times 2^x + 1 \times (2^x/4)^3 \times 2^x$$

File size = number of data blocks * Block size

contains "2^x / 4" addresses

Block size 2^x	Max size
1024 bytes = 2^{10}	approx. 16 GB
4096 bytes = 2^{12}	approx. 4 TB

10.7 可扩展文件系统 Ext

- Extended File System
- The latest default FS for Linux distribution is the Fourth Extended File System, Ext4 for short.
- 基于 iNode

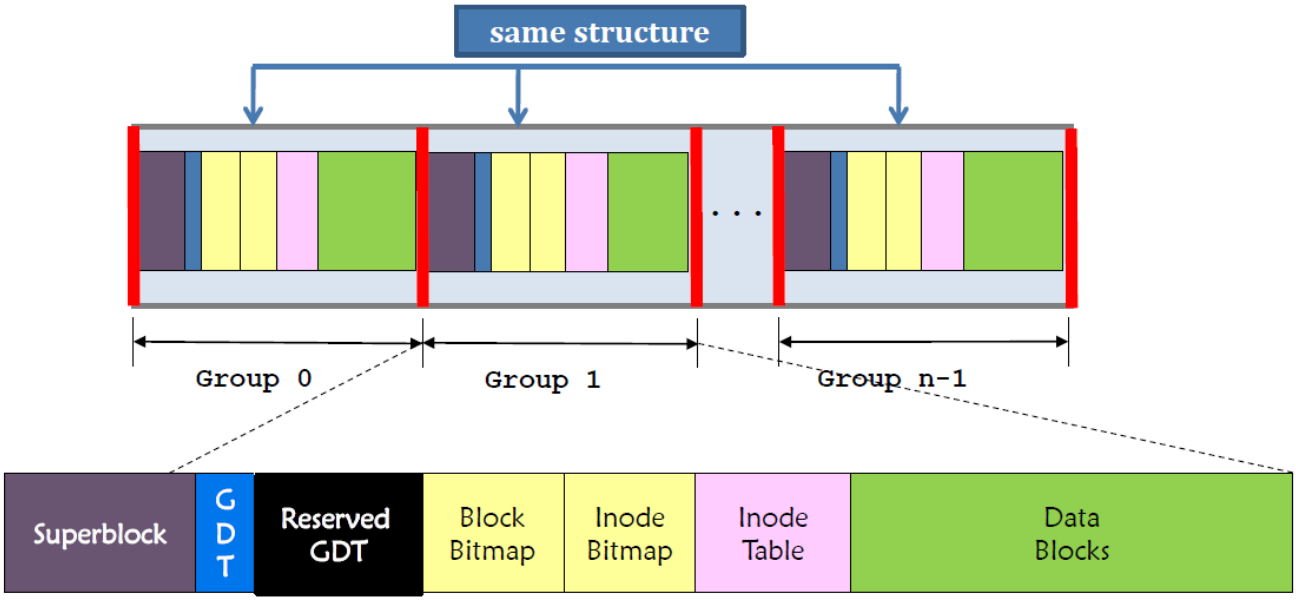
10.7.1 Ext 文件系统的大小

- For Ext2 & Ext3:
  - Block size: 1,024, 2,048, or 4,096 bytes.
  - Block address size: 4 bytes => # of block addresses = 2^{32}

$2^x \times 2^{32} = 2^{32+x}$			
Block size	$2^x = 1024$	$2^x = 2048$	$2^x = 4096$
File System size	4 TB	8 TB	16 TB

10.7.2 Ext 文件系统结构

- The file system is divided into block groups and every block group has the same structure



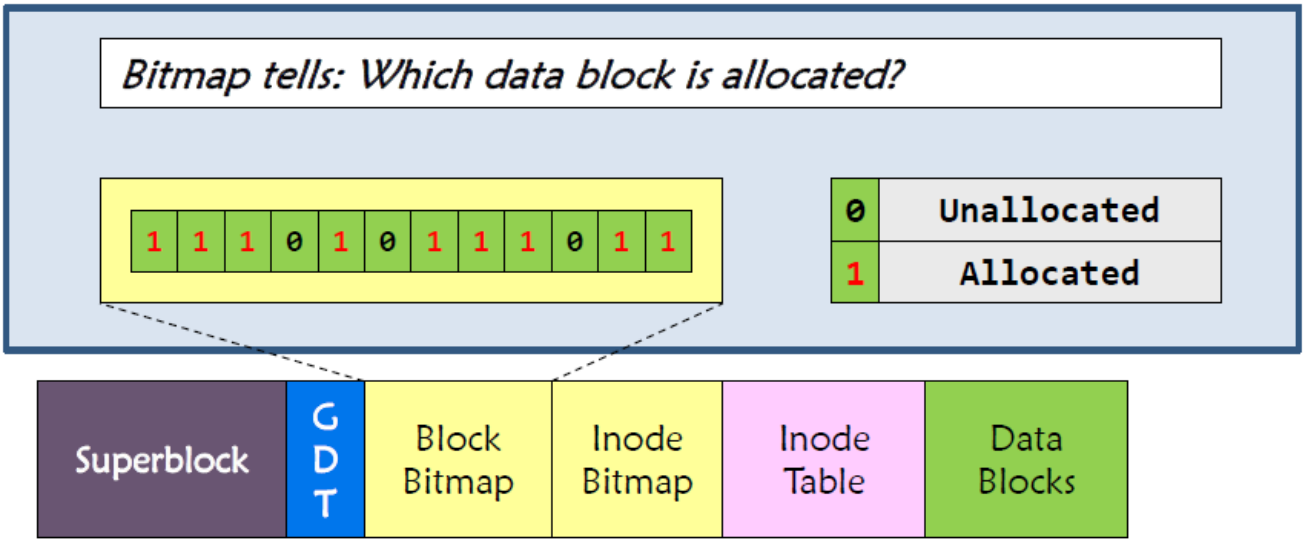
- Block Group 的结构

Superblock	Stores FS specific data. E.g., the total number of blocks, etc.
GDT – Group Descriptor Table	It stores: - The locations of the **block bitmap**, the **iNode bitmap**, and the **iNode table**. - Free block count, free iNode count, etc...
Block Bitmap	A bit string that represents if a block is allocated or not.
iNode Bitmap	A bit string that represents if an inode (index-node) is allocated or not.
iNode Table	An array of inodes ordered by the inode #.
Data Blocks	An array of blocks that stored files.

- Block Bitmap & iNode Bitmap
  - Block bitmap tells which block is allocated
  - iNode Bitmap

    A bit string that represents if an iNode (index node) is allocated or not

    Implies that the number of files in the file system is **fixed**



- Block Group 的优点
  - Performance: spatial locality

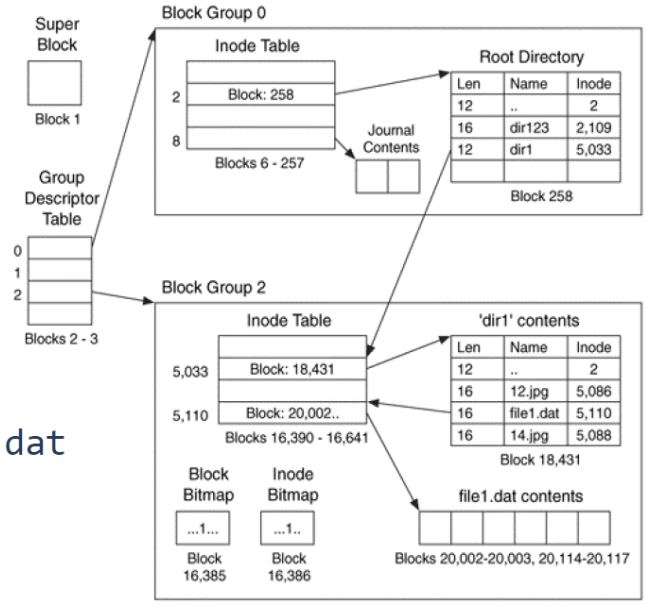
    Group iNodes and data blocks of related files together
  - Reliability

    Superblock and GDT are replicated in each block group

    可以互相校验
- 磁盘管理

  Disk divided into block groups
  - Each group has two block sized bitmaps (free blocks/ inodes)
  - Block sizes settable at format time: 1K, 2K, 4K, 8K···
  - Provides locality

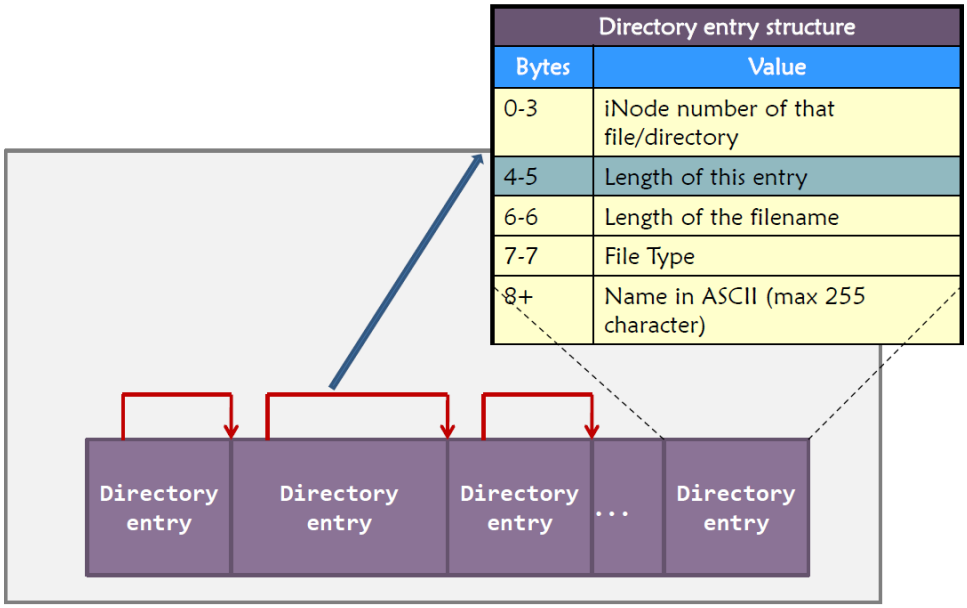
  例: 在 `/dir1/` 目录下创建 `file1.dat`

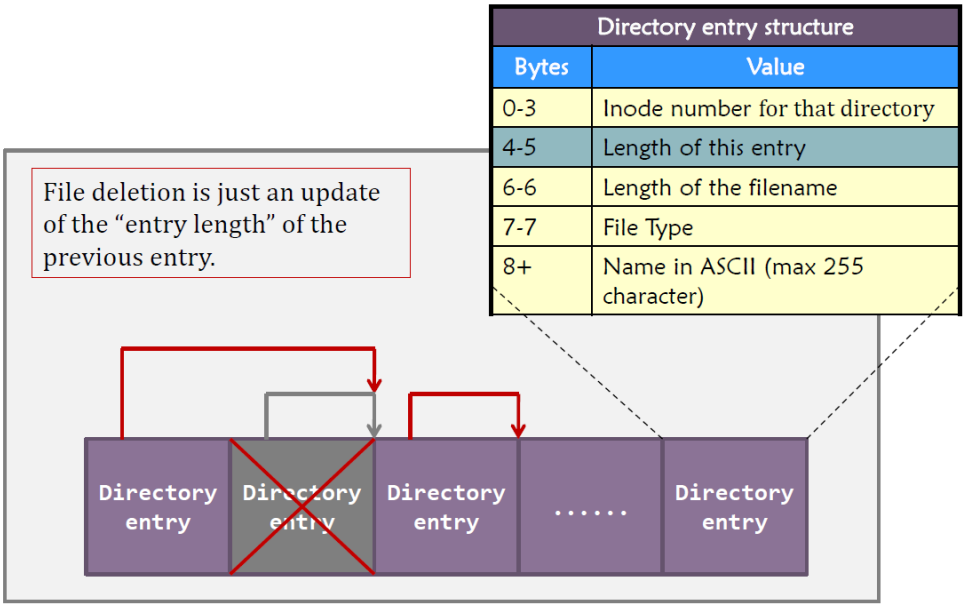


10.7.3 Ext 的 iNode 结构

iNode Structure (128 bytes long)	
Bytes	**Value**
0-1	File type and permission
2-3	User ID
4-7	Lower 32 bits of file sizes in bytes
8-23	Time information
24-25	Group ID
26-27	Link count (will discuss later)
…	…
40-87	12 direct data block pointers
88-91	Single indirect block pointer
92-95	Double indirect block pointer
96-99	Triple Indirect block pointer
…	…
108-111	Upper 32 bits of file sizes in bytes

10.7.4 Ext 删除文件



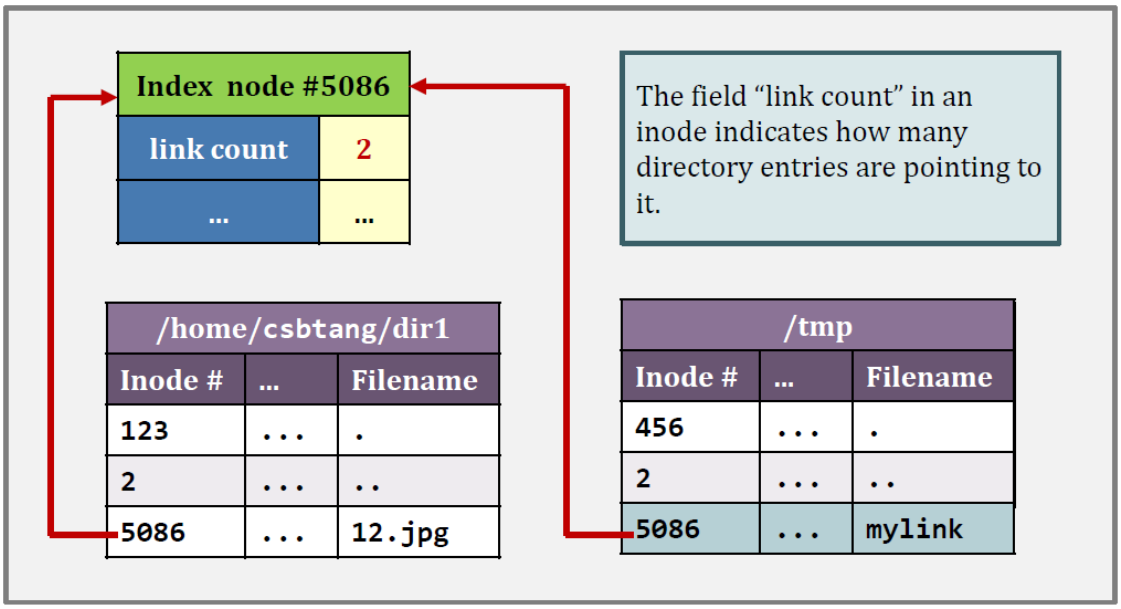


10.7.5 硬链接 Hard Link

- A hard link is a directory entry pointing to the iNode of an existing file

  That file can accessed through two different pathnames

  例: `ln /home/csbtang/dir1/12.jpg /tmp/mylink`

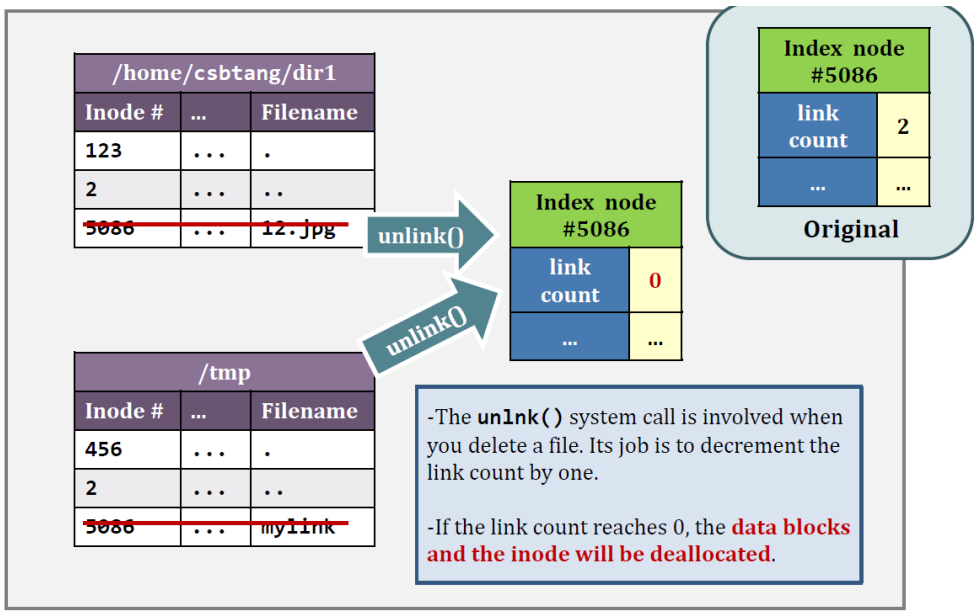


- 例: `/` 下有 20 个目录，`/` 有多少硬链接
  - 20 sub directories: they have link `..`
  - Root directory: `.` and `..` pointing to itself

    根目录比较特殊：`..` 指向自己，因为它没有上级目录
  - 20 + 2 = 22 hard links

```
# ls -F /
bin/    home/           media/  rules.log  tmp/
boot/   initrd.img@     mnt/    sbin/      usr/
cdrom/  initrd.img.old@ opt/    selinux/   var/
dev/    lib/            proc/   srv/       vmlinuz@
etc/    lost+found/     root/   sys/       vmlinuz.old@
# stat /
  File: `/'
  Size: 4096          Blocks: 8          IO Block: 4096   directory
Device: 806h/2054d    Inode: 2           Links: 22
......
$ _
```

- 删除硬链接

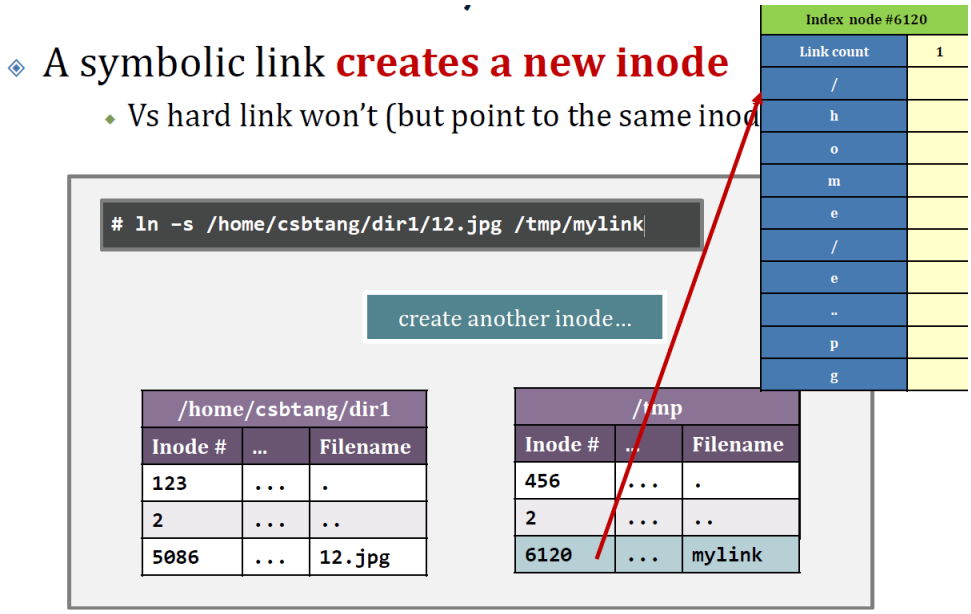


就是把链接的 iNode 的 link count --，当 = 0 的时候就会被 deallocate

10.7.6 符号链接 (软链接) Symbolic (Soft) Link

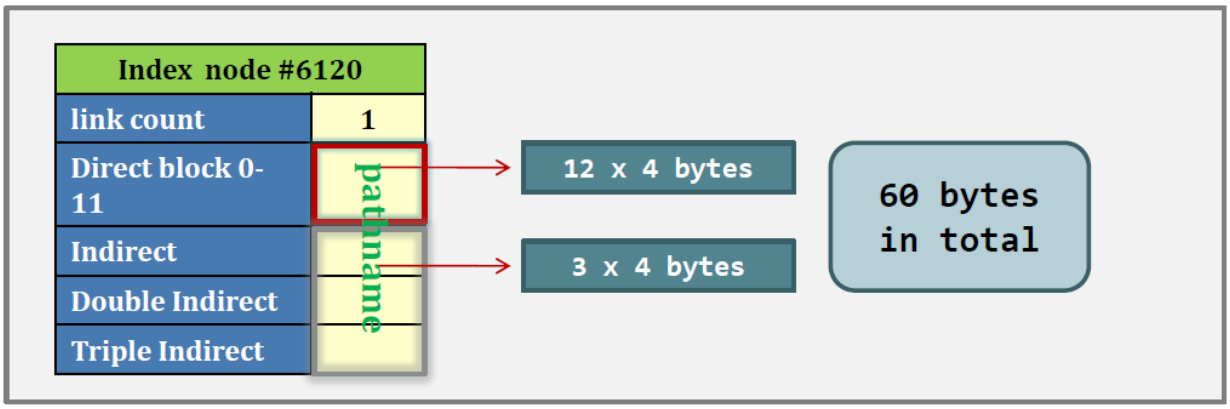
- A symbolic link creates a new iNode

  例: `ln -s /home/csbtang /dir1/12.jpg /tmp/mylink`



- Symbolic link is pointing to a new iNode whose target's **pathname** are stored using the space originally designed for 12 direct block and the 3 indirect block pointers if the pathname is shorter than 60 characters

  Use back a normal iNode + one direct data block to hold the long pathname otherwise



软链接的 iNode 把本来存 pointer 的位置用来存链接到的文件的路径，如果路径大于 60 bytes 就再用一个 direct block 存

- 硬链接和软链接比较
  - Hard link
    - Sets another directory entry to contain the file number for the file
    - Creates another name (path) for the file
    - Each is "first class"
  - Soft link or Symbolic Link
    - Directory entry contains the path and name of the file
    - Map one name to another name

Property/Action		Symbolic link	Hard link
When the link is deleted		Target remains unchanged	Reference counter is decremented; when it reaches 0, the target is deleted
When target is moved		Symbolic link becomes invalid	Hard link remains valid
Relative path		Allowed	N/A
Crossing filesystem boundaries		Supported	Not supported (target must be on same filesystem)
Windows	For files	Windows Vista and later[20] (administrator rights required)	Yes
	For folders		No
Unix	For files	Yes	Yes
	For directories	Yes	Partial[21]

10.8 NTFS

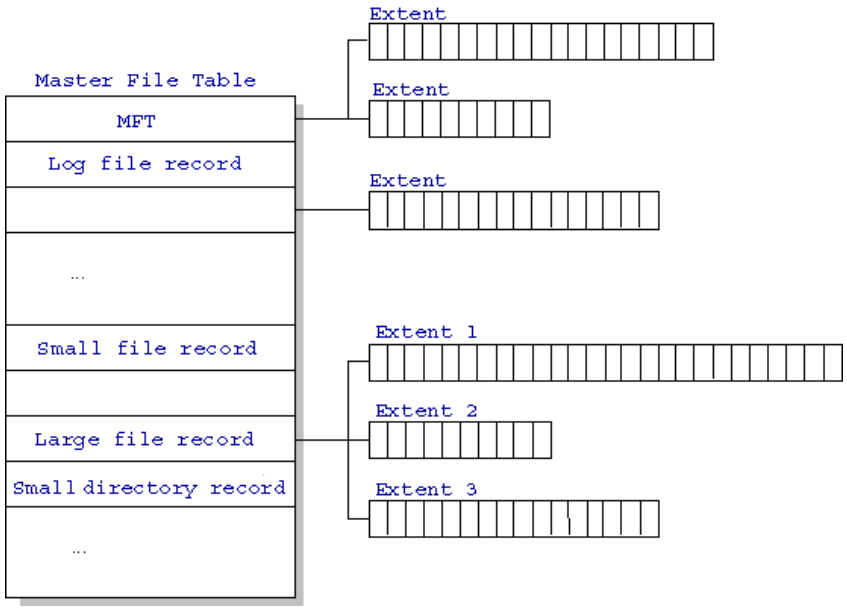
- New Technology File System (NTFS)

  Default on Microsoft Windows systems
- Variable length extents
- Everything (almost) is a sequence of `<attribute: value>` pairs
- Mix direct and indirect freely
- Directories organized in B-tree structure by default

10.8.1 NTFS 文件系统结构

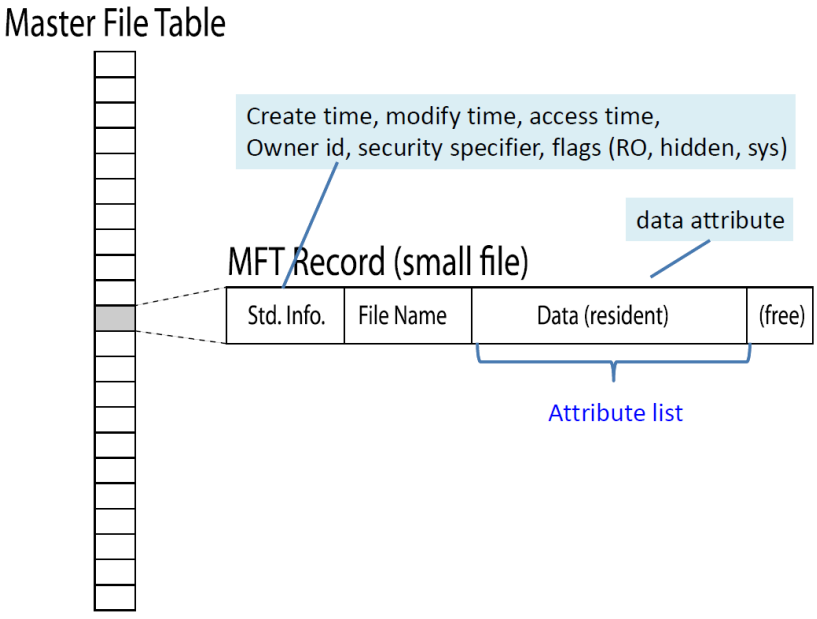
- Master File Table
  - Database with flexible 1KB entries for metadata/data
  - Variable sized attribute records (data or metadata)
  - Extend with variable depth tree (non resident)
- Extents

  Variable length contiguous regions
  - Block pointers cover runs of blocks
  - Similar approach in Linux (ext4)
  - File create can provide hint as to size of file
- Journaling for reliability 日志记录



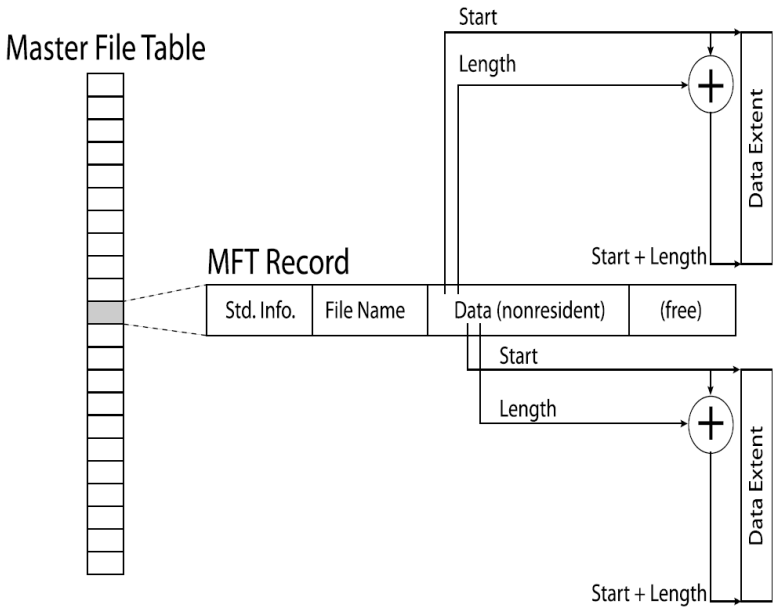
10.8.2 NTFS 文件存储

- 小文件

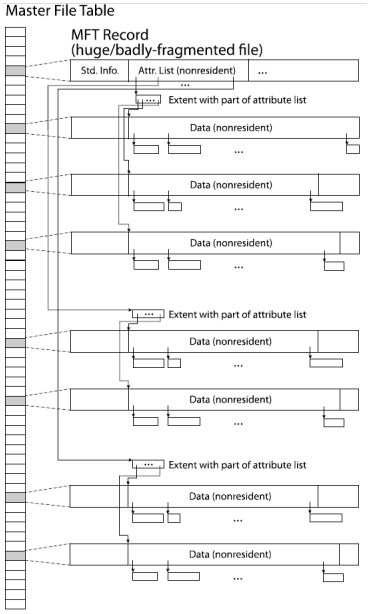


直接把数据存在 MFT Entry 里，不需要 data block

- 中文件



- 大文件



10.9 内存映射文件 Memory Mapped File

- 定义

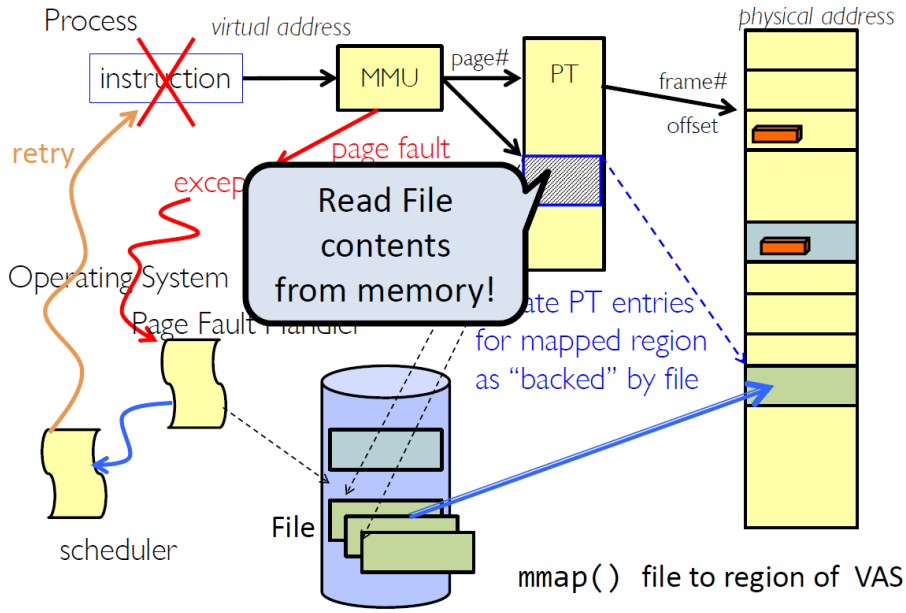
  A memory-mapped file contains the contents of a file in virtual memory. This mapping between a file and memory space enables an application, including multiple processes, to modify the file by reading and writing directly to the memory

- Traditional I/O involves explicit transfers between buffers in process address space to/from regions of a file

  This involves multiple copies into caches in memory, plus system calls

- Map the file directly into an empty region of our address space
  - Implicitly "page it in" when we read it
  - Write it and "eventually" page it out

- Executable files are treated this way when we `exec` the process



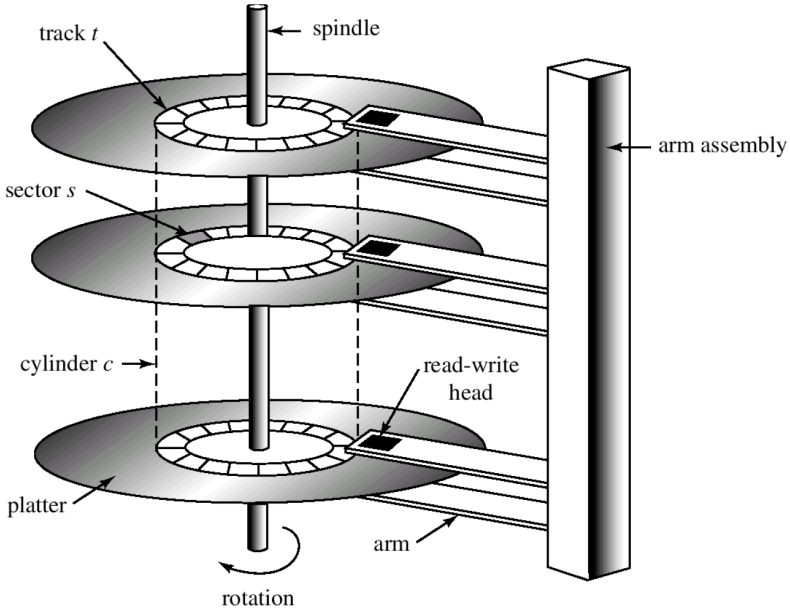
10.10 文件系统总结

- File System
  - Transforms blocks into Files and Directories
  - Optimize for size, access and usage patterns
  - Maximize sequential access, allow efficient random access
- File defined by header, called iNode
- Naming: translating from user visible names to actual sys resources
  - Directories used for naming for local file systems
  - Linked or tree structure stored in files
- Multilevel Indexed Scheme
  - iNode contains file info, direct pointers to blocks, indirect blocks, doubly indirect, etc..
  - NTFS: variable extents not fixed blocks, tiny files data is in header
- File Allocation Table (FAT) Scheme
  - Linked list approach
  - Very widely used: Cameras, USB drives, SD cards
  - Simple to implement, but poor performance and no security
- 4.2 BSD Multilevel index files
  - iNode contains pointers to actual blocks, indirect blocks, double indirect blocks, etc.
  - Optimizations for sequential access: start new files in open ranges of free blocks, rotational optimization
- File layout driven by freespace management
  - Integrate freespace , iNode table, file blocks and dirs into block group
- Deep interactions between memory management, file system, sharing
  - `mmap()`: map file or anonymous segment to memory

第十二章 大容量存储结构

12.1 大容量存储结构概述

12.1.1 磁盘 Magnetic Disk (Hard Disk)

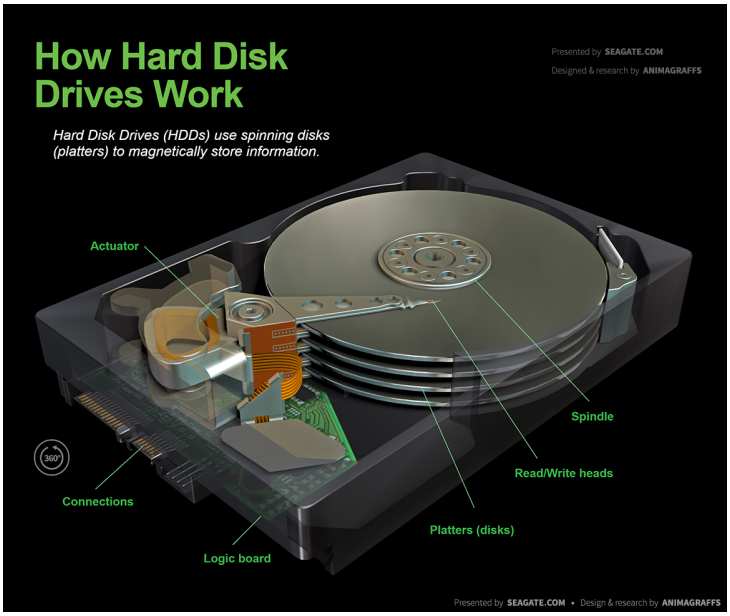


- 盘片 Platter
  - 磁道 Track

    每个 track 都有一个编号，一般 0 号指最外侧的 track
    - 扇区 Sector

      Sector 是磁盘存数据的最小 unit

      通常一个 sector 的 size 比一个 block 要小
- 柱面 Cylinder
- 磁臂 Disk Arm



12.1.2 磁盘性能

- 每分钟转数 Rotation Per Minute (RPM)
- 传输速率 Transfer Rate

  驱动器和计算机之间的数据流的速率
- 定位时间 (Positioning Time) 或随机访问时间 (Random Access Time)
  - 寻道时间 Seek Time

    移动磁臂到柱面所需时间
  - 旋转延迟 Rotational Latency

    旋转磁臂到所要扇区所需的时间
- Data R/W Time

  Positioning Time + Transfer Time (transfer a block of bits (sector) under r/w head)

  = Seek Time + Rotational Latency + Transfer Time
- Disk Latency

  Queueing Time + Controller time + Seek Time + Rotation Lantency + Transfer Time
- 例:
  - Assumptions
    - Ignoring queuing and controller times for now
    - Avg seek time of 5 ms
    - 7200 RPM: Time for rotation: 60000 (ms/min) / 7200 (rev/min) ~= 8 ms (转一圈 8 ms)
    - 平均 Rotational Latency 为 0.5 * 8 = 4 ms
    - Transfer rate of 4 MB/s, sector size of 1 Kbyte $1024bytes \div 4 \times 10^6 (bytes/sec) = 256 \times 10^{-6} sec \ = 0.26ms$
  - Read sector from random place on disk

    Seek Time + Rotational Latency + Transfer Time = 5 + 4 + 0.26 = 9.26 ms

    Approx 10ms to fetch/put data: 100 KByte /sec
  - Read sector from random place in same cylinder

    Rotational Latency + Transfer Time = 4 + 0.26 = 4.26 ms

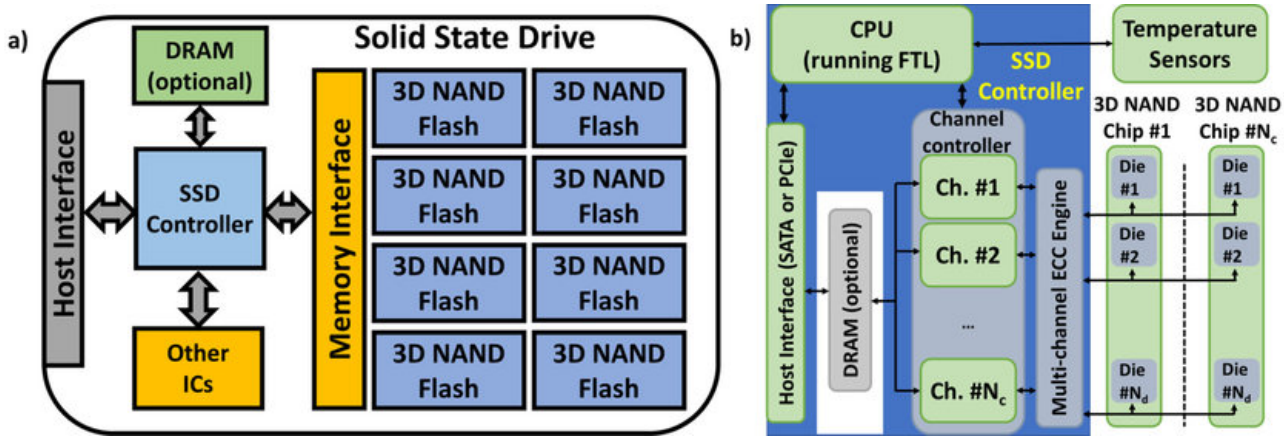
    Approx 5ms to fetch/put data: 200 KByte /sec
  - Read next sector on same track

    Transfer Time = 0.26 ms

    4 MByte /sec
- Typical Numbers for Magnetic Disk

Parameter	Info / Range
Space/Density	Space: 8TB (Seagate), 10TB (Hitachi) in 3½ inch form factor! Areal Density: ≥ 1Terabit/square inch! (SMR, Helium, …)
Average seek time	Typically 5-10 milliseconds. Depending on reference locality, actual cost may be 25-33% of this number.
Average rotational latency	Most laptop/desktop disks rotate at 3600-7200 RPM (16-8 ms/rotation). Server disks up to 15,000 RPM. Average latency is halfway around disk so 8-4 milliseconds
Controller time	Depends on controller hardware
Transfer time	Typically 50 to 100 MB/s. Depends on: • Transfer size (usually a sector): 512B – 1KB per sector • Rotation speed: 3600 RPM to 15000 RPM • Recording density: bits per inch on a track • Diameter: ranges from 1 in to 5.25 in
Cost	Used to drop by a factor of two every 1.5 years (or even faster); now slowing down

12.1.2 固态磁盘 Solid State Disk (SSD)



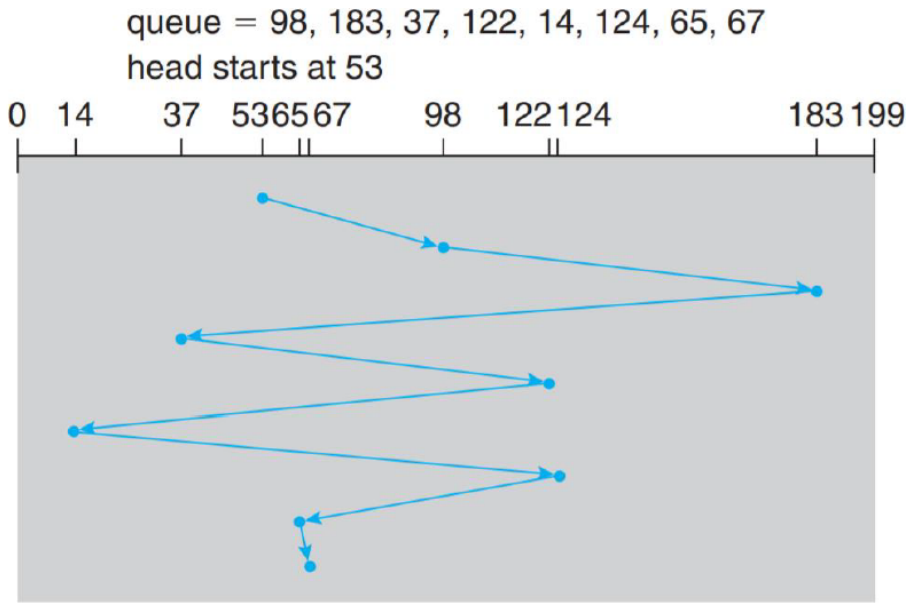
12.2 磁盘调度 Disk Scheduling

- Given a sequence of access pages in the HDD
  - 98, 183, 37, 122, 14, 124, 65, 67
  - Head point (now): 53
  - Pages: 0 ~ 199

  How to minimize seek time?

12.2.1 FCFS 调度

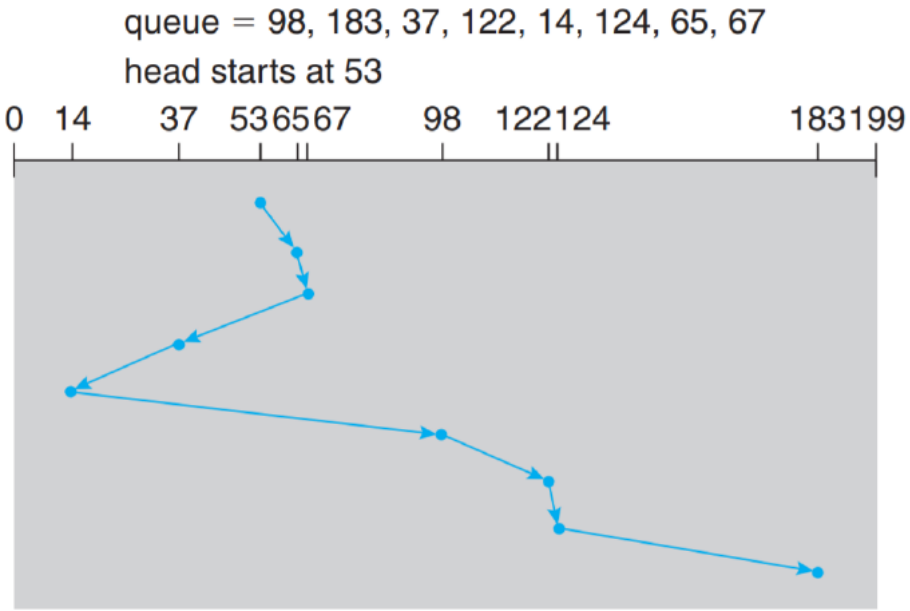
First come, first service



Total head movement distance = 640

12.2.2 SSTF 调度

最短寻道时间优先 Shortest-Seek-Time-First

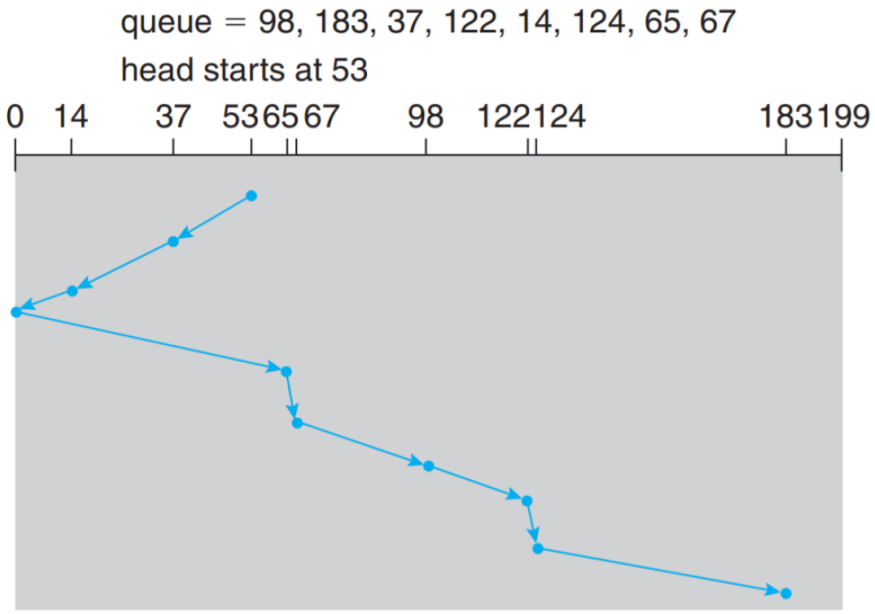


Total distance = 236

- 可能会出现饥饿 (Starvation)

12.2.3 SCAN 调度

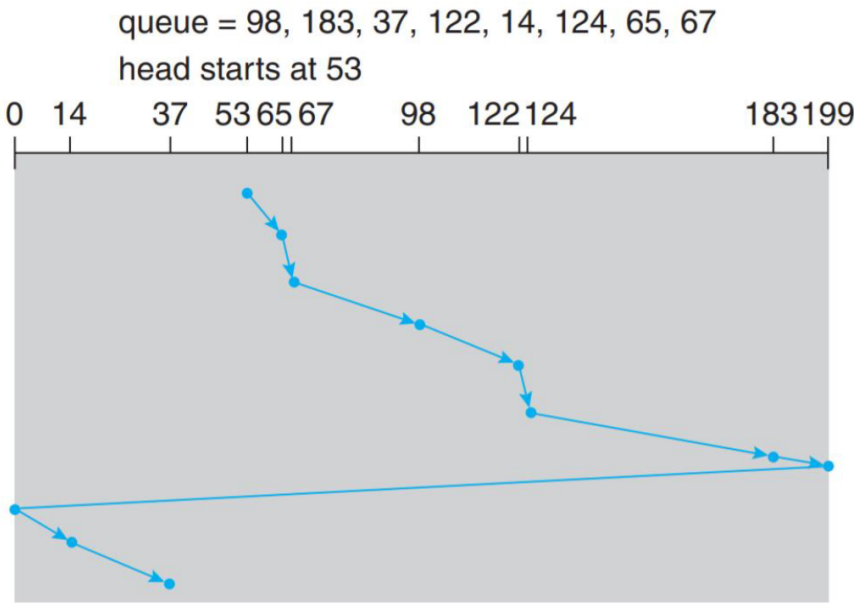
- 扫描算法 or 电梯算法 (elevator algorithm)
- 磁臂从磁盘的一头开始，向另一头移动，在移过每个柱面时处理请求；当到达另一端时，磁头移动方向反转并继续处理



Total distance = 236

12.2.4 C-SCAN 调度

- 循环扫描 (Circular SCAN)
- 磁头到另一端时，马上回到开头，不处理返回时的请求



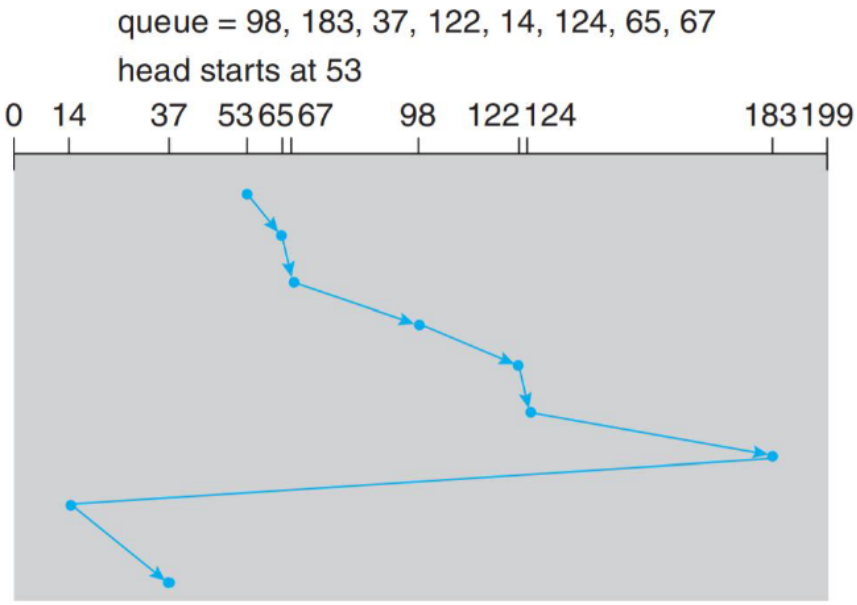
Total distance = 382

但是它的 avg wait time 比较小

12.2.5 LOOK 与 C-LOOK 调度

- 聪明点的 SCAN，不走到头了，走到最远的请求

C-LOOK:



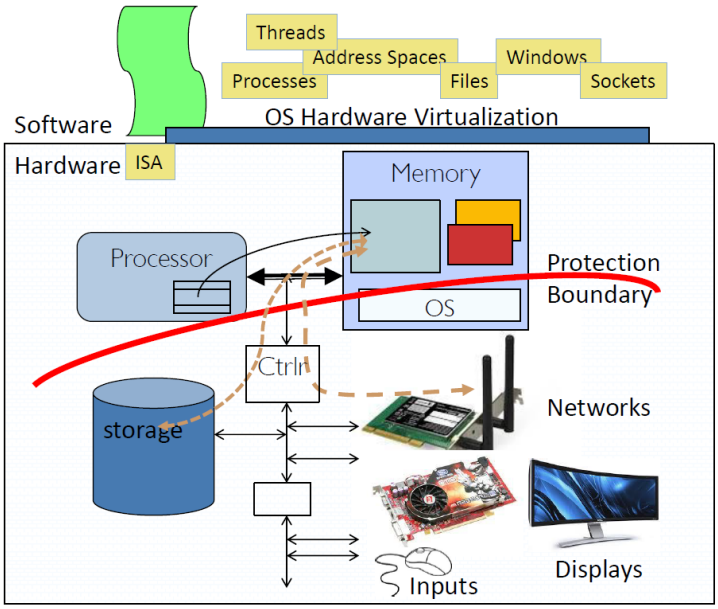
Total distance = 350

12.2.6 调度算法的选择

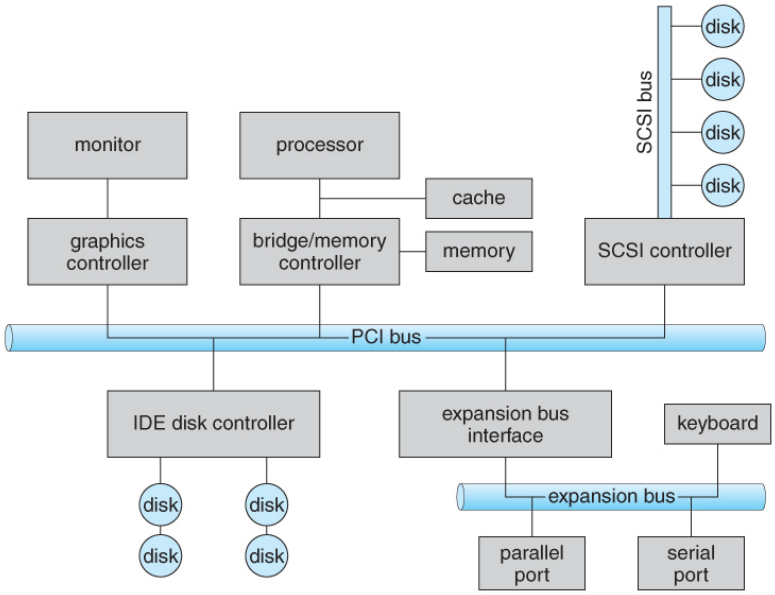
- SSTF is common and has a natural appeal
- SCAN and C-SCAN perform better for systems that place a heavy load on the disk (less starvation)
- Either SSTF or LOOK is a reasonable choice for the default algorithm
- Performance depends on the number and types of requests
- The disk scheduling algorithm should be written as a separate module of the operating system, allowing it to be replaced with a different algorithm if necessary

---

第十三章 I/O 系统



13.1 I/O 硬件



- 端口 Port

  设备与计算机的通信连接点

  - 数据输入寄存器 Data-in Register

    被主机读出以获取数据

  - 数据输出寄存器 Data-out Register

    被主机写入以发送数据

  - 状态寄存器 Status Register

    包含一些主机可以读取的位，表示一些状态，如当前命令是否已完成

  - 控制寄存器 Control Register

可由主机写入，以便启动命令或更改设备模式

- 总线 Bus

  一组线路和通过线路传输信息的严格定义的一个协议

  - PCI 总线
  - 扩展总线 Expansion Bus
  - SCSI 总线
    - 小型计算机连接接口 Small Computer System Interface (SCSI)
- 控制器 Controller

  可以操作端口、总线或者设备的一组电子器件

  - 磁盘控制器 Disk Controller
    - 串行高级技术连接 Serial Advanced Technology Attachment (SATA)
  - SCSI 控制器

    主机适配器 Host Adapter 或单独的电路板

以下为课件内容，我实在不知道该放在哪一节了

Operational Parameters for I/O

- Data granularity: Byte vs. Block
  - Some devices provide single byte at a time (e.g., keyboard)
  - Others provide whole blocks (e.g., disks, networks, etc.)
- Access pattern: Sequential vs. Random
  - Some devices must be accessed sequentially (e.g., tape)
  - Others can be accessed “randomly” (e.g., disk, cd, etc.)
    - Fixed overhead to start transfers
  - Some devices require continual monitoring
  - Others generate interrupts when they need service
- Transfer Mechanism: Programmed IO and DMA

13.2 CPU 访问 I/O 设备

- CPU 与 Contorller 交互

  控制器有一个或多个寄存器，用于数据和控制信号

  处理器通过读写这些寄存器的位模式来与控制器通信

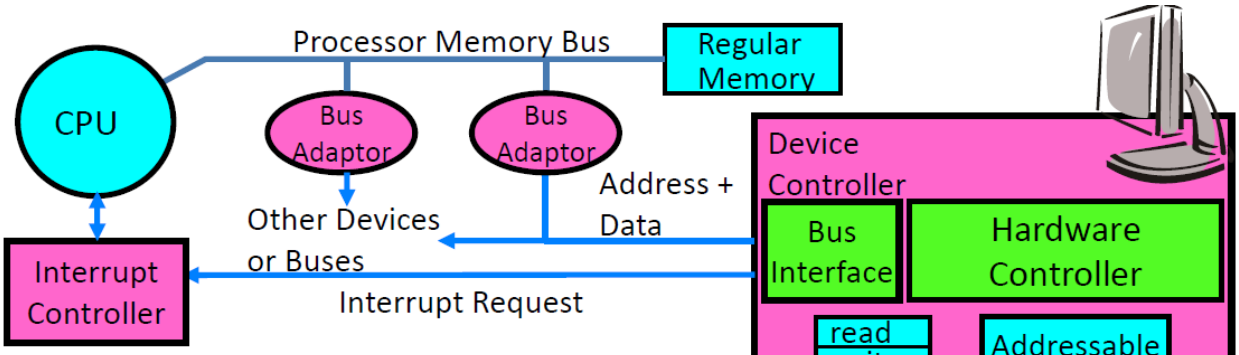
  - I/O Instuction

    通过特殊 IO 指令针对 IO 端口地址传输一个字节或字

    IO 指令触发总线线路，选择适当设备，并将位移入或移出设备寄存器
  - 内存映射 (Memory Mapped) I/O

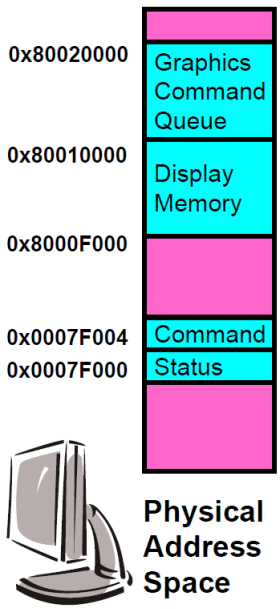
    设备控制寄存器被映射到处理器的地址空间

处理器执行 IO 请求是通过标准数据传输指令读写映射到内存的设备控制器



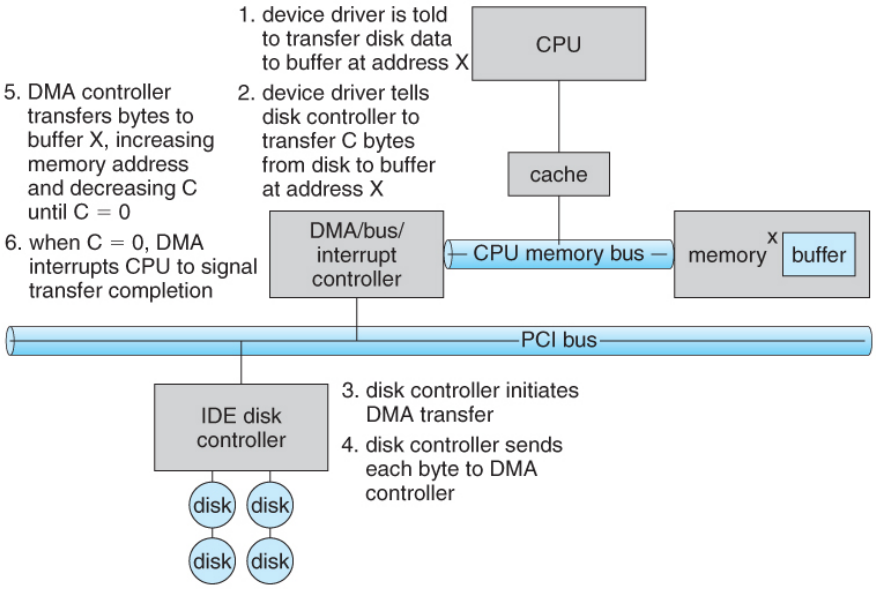
- CPU interacts with a *Controller*
  - Contains a set of *registers* that can be read and written
  - May contain memory for request queues or bit-mapped images

- 例: Memory Mapped Display Controller
  - Hardware maps control registers and display memory into physical address space
    - Addresses set by HW jumpers or at boot time
  - Simply writing to display memory (also called the "frame buffer") changes image on screen
    - Addr: 0x8000F000 — 0x8000FFFF
  - Writing graphics description to cmd queue
    - Say enter a set of triangles describing some scene
    - Addr: 0x80010000 — 0x8001FFFF
  - Writing to the command register may cause onboard graphics hardware to do something
    - Say render the above scene
    - Addr: 0x0007F004
  - Can protect with address translation



13.3 控制器与 I/O 设备的数据传输

- 程序控制 (Programmed) I/O
  - Each byte transferred via processor in/out or load/store
  - Pro: Simple hardware, easy to program
  - Con: Consumes processor cycles proportional to data size
- 直接内存访问 Direct Memory Access (DMA)
  - Give controller access to memory bus
  - Ask it to transfer data blocks to/from memory directly



13.4 I/O 设备与 CPU 通信

13.4.1 轮询 Polling

- OS periodically checks a device specific status register
  - I/O device puts completion information in status register
- Pro: low overhead
- Con: may waste many cycles on polling if infrequent or unpredictable I/O operations

主机与控制器之间的握手协调:

1. 主机重复读取忙位 (busy bit, 在 status register 里)，直到该位清零
2. 主机设置命令寄存器的写位，并写出一个字节到数据输入寄存器
3. 主机设置命令就绪位
4. 当控制器注意到命令就绪位已设置，则设置忙位
5. 控制器读取命令寄存器，并看到写命令。它从数据输出寄存器中读取一个字节，并向设备执行 I/O 操作
6. 控制器清除命令就绪位，清除状态寄存器的故障位表示设备 I/O 成功，清除忙位表示完成

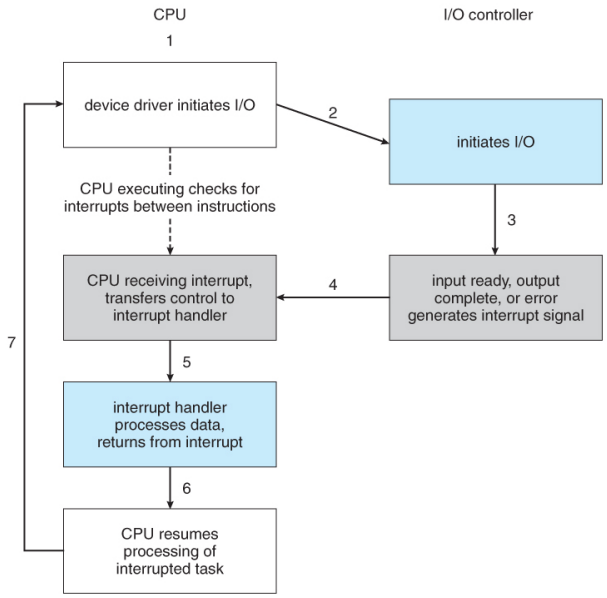
在步骤 1 中，主机一直处于忙等待 (busy waiting) 或轮询 (polling)。在该循环中，一直读取状态寄存器，直到忙位被清除

13.4.2 I/O 中断 I/O Interrupt

- Device generates an interrupt whenever it needs service
- Pro: handles unpredictable events well
- Con: interrupts relatively high overhead

CPU 硬件有一条中断请求线 (Interrupt-Request Line, IRL)。CPU 在执行完每条指令后，都会检测 IRL。当 CPU 检测到控制器已在 IRL 上发出了一个信号时，CPU执行状态保存并且跳到内存固定位置的中断处理程序 (Interrupt Handler Routine)。中断处理程序确定中断原因，执行必要处理，执行状态恢复，并且执行返回中断指令以便 CPU 回到中断前的执行状态。

总结：设备控制器通过 IRL 发送信号从而引起 (raise) 中断，CPU捕获 (catch) 中断并且分派 (dispatch) 到中断处理程序，中断处理程序通过处理设备来清除 (clear) 中断。



- 中断请求线 IRL: 两条
  - 非屏蔽中断 (Nonmaskable Interrupt)

    保留用于诸如不可恢复的内存错误等事件
  - 可屏蔽中断 (Maskable Interrupt)

    在执行不得中断的关键指令序列之前，它可以由 CPU 关闭。可屏蔽中断可由设备控制器用来请求服务
- 中断向量 Interrupt Vector

  是一个地址，根据这个地址+偏移量来选择特定的中断处理程序
- 中断优先级 Interrupt Priority Level
- Top-half/bottom-half interrupt architecture

  https://stackoverflow.com/questions/45095735/top-halves-and-bottom-halves-concept-clarification

https://www.tutorialsdaddy.com/uncategorized/top-half-vs-bottom-half/

- 上半部 Top Half Handler (硬中断)

  快速处理中断，它在中断禁止模式下运行，主要处理跟硬件紧密相关的或时间敏感的工作

- 下半部 Bottom Half Handler (软中断)

  延迟处理上半部未完成的工作，通常以内核线程的方式运行

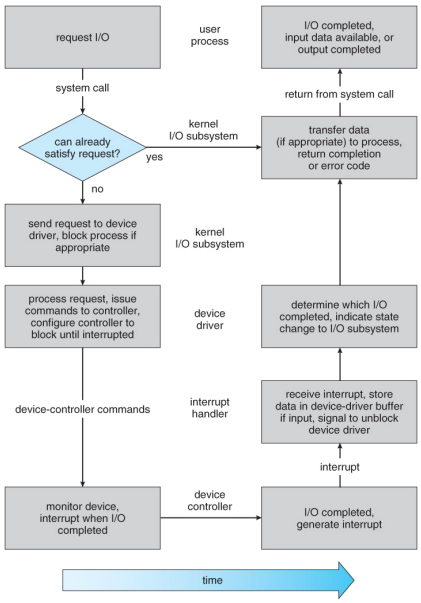
- 设备驱动 Device Driver

  http://www.differencebetween.net/technology/difference-between-device-driver-and-device-controller

  Device driver is a specialized software program running as part of the operating system that interacts with a device attached to a computer. It is just a code inside the OS that allows to be empowered with the specific commands needed to operate the associated device.

  - Supports a standard, internal interface
  - Same kernel I/O system can interact easily with different device drivers
  - Special device specific configuration supported with the `ioctl` system call

13.5 I/O 请求生命周期



13.6 I/O 性能

- Response Time or Latency

  Time to perform an operation(s)

  Response Time = Queue + I/O device service time

- Bandwidth or Throughput

  Rate at which operations are performed (op/s)

  - Files: MB/s, Networks: Mb/s, Arithmetic: GFLOP/s

    - Effective BW per op = transfer size / response time

      EffBW(n) = n / (S + n/B) = B / (1 + SB/n)

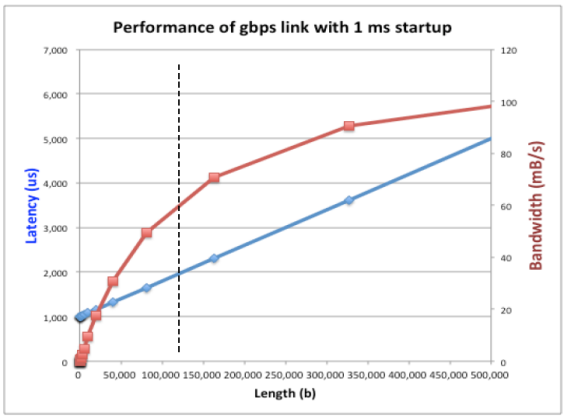
- Start up or "Overhead"

  Time to initiate an operation

    - Syscall overhead
    - Operating system processing
    - Controller Overhead
    - Device Startup
        - Mechanical latency for a disk
        - Media Access + Speed of light + Routing for network
    - Queuing
- Most I/O operations are roughly linear in n bytes

  $Latency(n) = Overhead + n/TransferCapacity$
- 例: Fast Network

  Consider a 1 Gb/s link (Transfer capacity B = 125 MB/s)
    - With a startup cost S = 1 ms



    - Latency(n) = S + n/B
    - Bandwidth = n/(S + n/B) = B*n/(B*S + n) = B/(B*S/n + 1)
- 影响最大带宽的因素
    - Bus speed
        - PCI X: 1064 MB/s = 133 MHz x 64 bit (per lane)
        - ULTRA WIDE SCSI: 40 MB/s
        - Serial Attached SCSI & Serial ATA & IEEE 1394 (firewire): 1.6 Gb/s full duplex (200 MB/s)
        - USB 3.0: 5 Gb/s
        - Thunderbolt 3: 40 Gb/s
    - Device transfer bandwidth
        - Rotational speed of disk
        - Write / Read rate of NAND flash
        - Signaling rate of network link

END