操作系统原理笔记

陈鸿峥

2019.07*

目录

*Build 20190710

本课程使用的教材为William Stallings《操作系统—精髓与设计原理（第八版）》。其他参考资料包括Stanford CS140、CMU15-460、*Operating System Concepts (10th ed.)*。

关于计算机系统的内容在此不再赘述，详情参见**计算机组成原理**的笔记。

1 操作系统概述

1.1 概述

操作系统核心即怎么虚拟多几个冯诺依曼计算机出来给程序用。操作系统是**控制**应用程序执行的程序，是应用程序和计算机硬件间的**接口**（屏蔽硬件细节）。

这里先解释几个概念

- 并发：两件事情**可以**同时(simultaneously)发生，没有时间限制，$t_1 > t_2$，$t_1 < t_2$，$t_1 = t_2$都可
- 同步：两个事件有确定的时间限制
- 异步：两件事不知道何时发生

例 1. *如果进程A有2条顺序执行的指令，进程B有3条顺序执行的指令，那么在SMP结构两个CPU的机器内并发执行这两个进程，不同的相对时序有几种？*

分析. *实际上就是隔板法解不定方程，设指令分别为ab123，将ab作为隔板，则在ab两侧和ab之间的指令条数为x_1, x_2, x_3，有不定方程*

$$x_1 + x_2 + x_3 = 3, x_i \geq 0$$

由组合数学知识知上面不定方程有$C_{3+3-1}^{3-1} = C_5^2 = 10$个解，即不同相对时序有10种。

1.2 发展历史

- 串行处理/手工操作(1940s)：没有OS，人工调度，准备时间长
- 简单批处理系统(1950s)：
  - 使用监控程序(monitor)，读入用户程序执行
  - 提供内存保护、计时器、特权指令、中断
  - 两种操作模式：用户态、内核态(mode)
  - 单道程序(uniprogramming)批处理：处理器必须等到IO指令结束后才能继续
- 多道程序批处理(1950s末)：多个作业同时进入主存，切换运行，**充分利用处理器**（大块处理时间，少交换上下文的调度）；用户响应时间长，**不提供人机交互能力**；脱机计算环境
- 分时系统(1961)：MIT CTSS(Compatible Time-Sharing System)，满足用户与计算机交互的需要，**减小响应时间**（分时间片小块调度）；多个交互作业，多个用户，把运行时间分成很短的时间片轮流分配
- 实时系统：专用，工业、金融、军事

现代的操作系统通常同时具有分时、实时和多道批处理的功能，因此被称为通用操作系统。而OS也不仅是在PC机上有，网络OS、分布式OS、嵌入式OS层出不穷。

影响现代OS发展的因素：

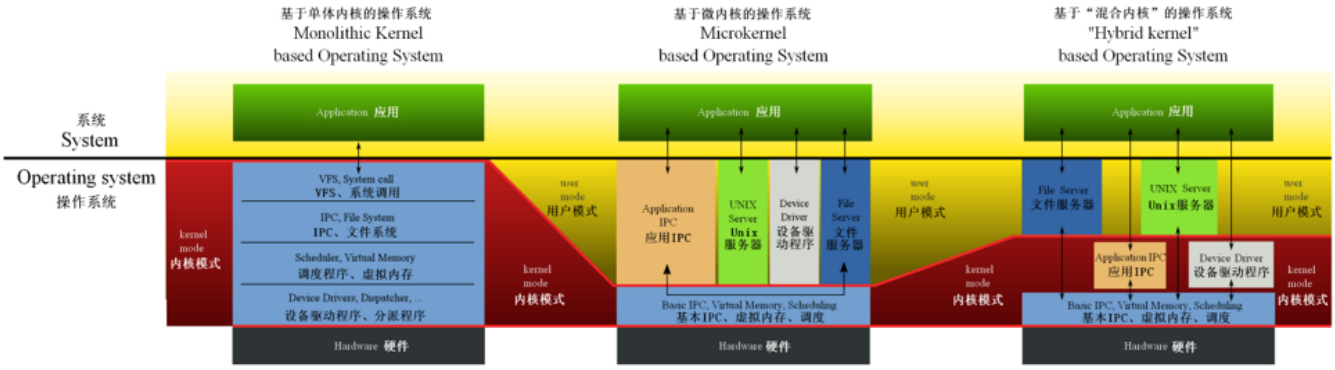
- 新硬件：多核/处理器结构、高速增长的计算机速度、高速网络连接、容量的不断增加各种存储设备
- 新应用：多媒体应用、互联网/Web访问、客户/服务器计算模式
- 新安全威胁：网络使安全问题更突出（病毒、蠕虫、黑客技术）

内核的分类：

- 单体内核(monolithic kernel)/宏内核(macro kernel)
  - 内核实现操作系统所有基本功能（包括调度、文件系统、连网、设备驱动程序、内存管理等等）
  - 一般用一个进程实现，内核代码共享同一个地址空间，每一模块可以调用任意其它模块和使用内核所有核心数据
  - 效率高，但难于修改和扩充
  - 例子：Unix、Linux、Android（基于Linux）、DOS、Windows 9x、Mac OS 8.6以下
- 微内核(microkernel)
  - 内核只实现最基本功能（包括地址空间、IPC[InterProcess Communication，进程间通信]和基本调度）
  - 更多的功能代码组织为多个进程，各自独立使用自己的地址空间，运行于用户态；通常内核服务越少内核越稳定，即微内核要比宏内核稳定

- 一致接口（消息传递[1]）、可扩展性（允许增加新服务）、可移植性（将系统移植到新处理器只需对内核修改），适用于嵌入式与分布式环境，但效率稍低
- 例子：Mach、QNX

- 混合内核
  - 微内核与宏内核的结合（也可算作单体内核中的一类）
  - 具有微内核结构，按宏内核实现
  - 例子：Windows NT(2000/XP/Vista/7/8)、BSD、XNU（Darwin 的核心，源自Mach和BSD，用于Mac OS X 和iOS）



1.3 基本特性总结

- 并发和共享是操作系统两个最基本的特征。
- 操作系统涉及的几个方面进程管理、内存管理、文件管理、设备管理。
- 用户通过命令接口和系统调用使用计算机。
- 提高单机资源利用率的关键技术是多道程序设计
- 多道程序技术的关键硬件基础为中断机制与时钟控制器
- 在操作系统的各个功能组成中，时钟管理、地址映射、中断系统都需要硬件支持，只有进程调度不需要
- 内部异常/内中断包括故障(fault)、陷阱[2](trap)、终止(abort)三类

2 进程

2.1 基本概念

进程(process)是**运行时(running/in execution)**程序的实例。

- 程序并发执行的特征（多道程序）：间断性、无封闭性、不可再现性（破坏冯顺序执行特性）

[1]即使是硬件中断也会被当作消息处理，需要send/receive
[2]让程序从用户态陷入内核态

- 进程的特点：动态性、并发性、独立性、异步性
- 进程的作用
  - 提升CPU利用率：将多个进程重叠（一个进程IO时另一个计算）
  - 降低延迟(latency)：并发执行，不断切换，防止卡住

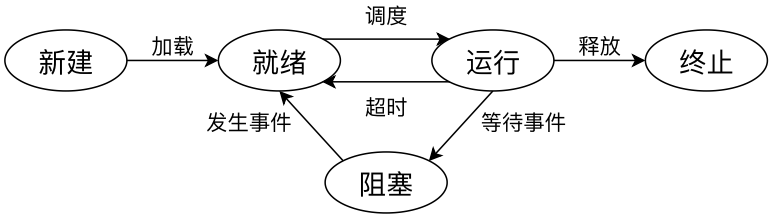
进程控制块(Process Control Block, PCB)，在Unix是`proc`，在Linux是`task_struct`

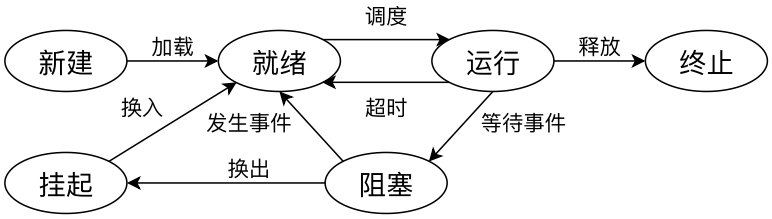
- 进程标识符/ID
- 进程状态：运行、就绪、等待/阻塞、创建、结束、挂起等
- 优先级
- 程序计数器(PC)
- 内存指针：报错指向程序代码、相关数据和共享内存的指针
- 上下文数据(context)：进程被中断时寄存器中的数据
- IO状态信息
- 记账信息(accounting)：占用处理器时间、时钟数总和、时间限制等
- 链表：各状态的进程形成不同的链表：就绪链表、阻塞链表等

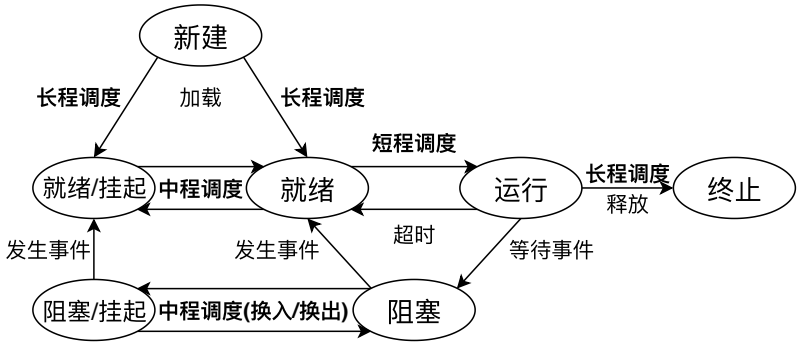
进程映像：进程控制块PCB（用于进程控制于调度）、程序段和相关数据段（用于运行程序）、系统栈（跟踪过程调用及参数传递）、共享地址空间（进程间通信）。

2.2 进程状态模型

多状态进程模型如下，其中挂起进程指进程**映像**在外存中等待被调入内存（又分为六状态和七状态模型）。进程由运行态转入阻塞态的原因是I/O中断、系统调用，这是一个主动的过程。而由阻塞态转入就绪态则是被动的过程。







2.3 进程的操作

进程的创建

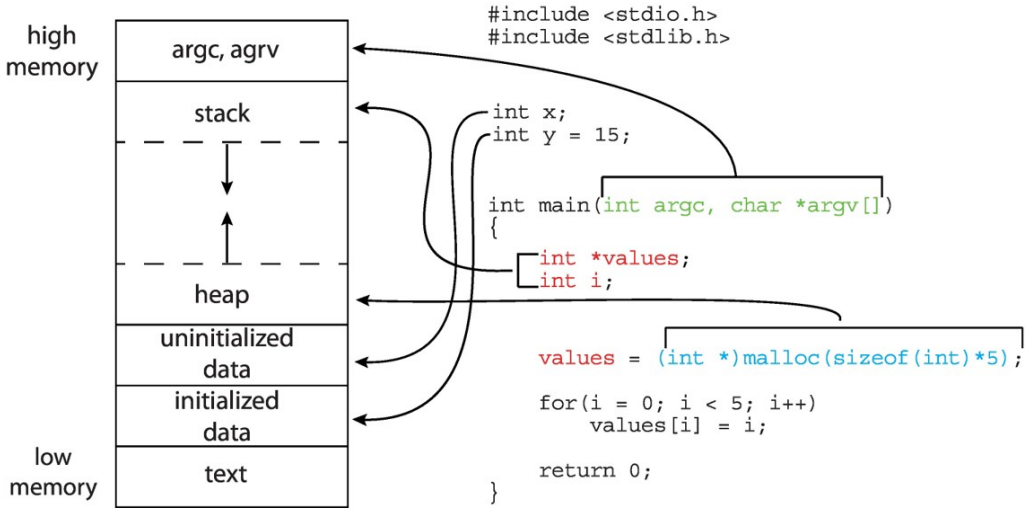
1. 分配唯一的进程ID
2. 分配进程所需的资源：内存空间（页面）、栈
3. 初始化PCB：ID、状态、优先级等
4. 如果可以则加入就绪队列（长程调度）

进程间的通信

- 共享存储：进程到共享空间再到进程
- 消息传递：直接进程到进程
- 管道通信：进程到缓冲区到进程，管道即共享文件，半双工通信

进程控制由原语(primitive)完成，由若干条指令完成，也可被视为原子操作

内存组织：代码段、数据段、堆段、栈段（从小地址往大地址）



可以参考原始的UNIX论文[3]。

- 创建进程：fork、waitpid
- 删除进程：exit、kill
- 执行进程：execve

导致OS获得控制权的事件：

- 时钟中断：时间片结束
- IO中断：IO完成
- 硬件中断/陷阱(trap)/异常
- 系统调用：`int`

可重入性(reentrant)函数：可以被一个以上的任务调用，而不必担心数据的破坏。可重入型函数任何时候都可以被中断，一段时间以后又可以运行，而相应数据不会丢失。可重入型函数或者只使用局部变量，即变量保存在CPU寄存器中或堆栈中。如果使用全局变量，则要对全局变量予以保护。如`strcpy`是可重入的，而`swap`是不可重入的。

可再入式内核：一个进程中断后，内核会保存当前进程的相关信息，然后执行下一个进程，等完成后再将上个进程从中断的位置重新恢复执行。Linux内核是可再入的，这意味着几个进程可能同时在内核模式下执行。（当然单处理器系统，在某一时间只会有一个进程执行，但许多会阻塞在内核模式）这些进程会分时共享CPU、I/O设备等系统资源，给用户的感觉就像是在同时运行。

3 线程

在没有线程概念的系统中，进程是**资源分配**、调度/执行的单位；而在有线程概念的系统中，线程就成了**基本调度单位**/程序执行流最小单元，由线程ID、程序计数器、寄存器集合和堆栈组成。

线程独立拥有寄存器和栈等现场状态，只会与进程共享地址空间、代码和IO文件资源，**不与进程共享上下文寄存器**。

[3] http://www.scs.stanford.edu/19wi-cs140/sched/readings/unix.pdf

线程的优点：

- 创建速度快
- 终止所用时间少
- 切换时间少
- 通信效率高，同一进程无需调用内核，共享存储空间

用户级线程(ULT)：线程管理都由应用程序完成（线程库），内核不知道线程的存在，优点：

- 线程切换不需要模式切换
- 调度算法可以应用程序专用
- ULT不需要内核支持，线程库可以在任何OS上运行

缺点：

- 一个线程阻塞会导致整个进程阻塞（因用户级线程对操作系统不可见，操作系统对整个进程进行调度）
- 不能利用多核和多处理器技术

内核级线程(KLT)：线程管理由内核完成（提供API），调度基于线程进行，优点：

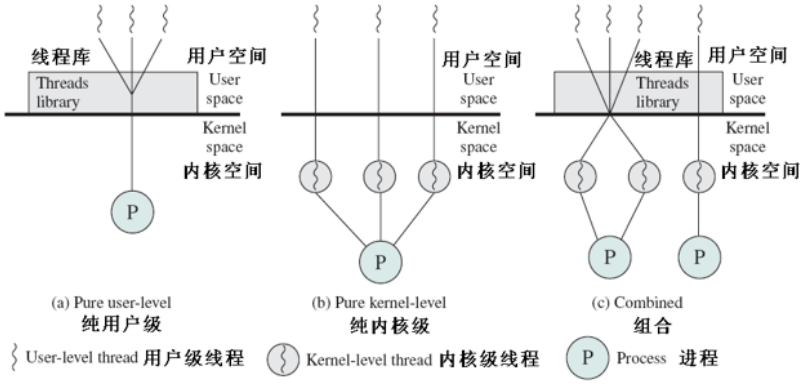
- 线程阻塞不会导致进程阻塞
- 可以利用多核和多处理器技术
- 内核例程本身也可以使用多线程

缺点：

- 线程切换需要进行模式切换

多线程模型

- 多对一：多个用户级线程映射到一个内核级线程，线程管理在用户空间进行，效率高；若内核服务阻塞，则整个进程被阻塞
- 一对一：每个用户级线程映射到一个内核级线程，开销大
- 多对多：折中



User-Level and Kernel-Level Threads
用户级与内核级线程

线程与进程之间的关系

- 1：1，每个进程都有唯一线程，DOS、传统Unix
- M：1，一个进程多个线程，Windows NT、Linux、Mac OS、iOS
- 1：M，一个线程可在多个进程环境中迁移

* Linux并不区分线程和进程，采用Copy On Write (COW)方式

4 并发—互斥与同步

4.1 基本概念

- 原子(atomic)操作：不可分割
- 临界区(critical section)：不允许多个进程同时进入的一段访问**共享资源**的代码
- 死锁(deadlock)：两个及以上进程，因每个进程都在等待其他进程做完某事（如释放资源），而不能继续执行
- 活锁(livelock)：两个及以上进程，为响应其他进程中的变化，而不断改变自己的状态，但是没有做任何有用的工作
- 互斥(mutual exclusion)：当一个进程在**临界区访问共享资源**时，不允许其他进程进入访问
- 竞争条件(race condition, RC)：多个进程/线程读写共享数据，其结果依赖于它们执行的相对速度
- 饥饿(starvation)：可运行的进程长期未被**调度执行**

核心内容

$$\text{并发} \rightarrow \text{共享} \rightarrow \text{RC问题} \rightarrow \text{互斥}$$

共享数据的最终结果取决于进程执行的相对速度（异步性），需要保证进程的结果与相对执行速度无关

同步：有明确的时间先后限制，两个或多个进程之间的操作存在时间上的约束（也包含互斥）

4.2 互斥

互斥的要求

- 在具有相同资源或共享对象的临界区的所有进程中，一次只允许一个进程进入临界区（强制排它）
- 一个在非临界区停止的进程必须不干涉其他进程（充分并发）
- 没有进程在临界区中时，任何需要访问临界区的进程必须能够立即进入（空闲让进）
- 决不允许出现一个需要访问临界区的进程被无限延迟（有限等待）
- 相关进程的执行速度和处理机数目没有任何要求或限制（满足异步）
- 当进程不能进入临界区，应该立即释放处理机，防止进程忙等待（让权等待）

4.2.1 简单的尝试

第一种尝试（单标志法）：两个进程轮流进入临界区

```
while (turn != 0)
    /* do nothing */;
/* critical section */
turn = 1;
```

```
while (turn != 1)
    /* do nothing */;
/* critical section */
turn = 0;
```

可以保证互斥，硬性规定进入的顺序，但是

- 忙等待，白白消耗CPU时间
- 必须轮流进入临界区，不合理，限制推进速度
- 若一个进程失败，则另一个将永远被阻塞

难以支持并发处理

一共四种尝试

- 单标志法
- 双标志先检查：死锁
- 双标志后检查：不能保证互斥
- 双标志延迟礼让：活锁

例 2. 忙等待效率一定比阻塞等待效率低吗

分析. 一般情况下确实如此，因为忙等待一直在消耗*CPU*资源。但特殊情况下，忙等待能立即响应请求完成（条件判断结束），对于性能要求较高的应用有好处；而阻塞等待还需等*OS*调度才能继续执行下面的工作。

4.2.2 软件方法

Dekker算法：避免无原则礼让，规定各进程进入临界区的顺序；逻辑复杂，正确性难以证明，存在轮流问题，存在忙等待；初始化`flag`都为`false`，`turn`为1

```
void P0()
{
    while(true)
    {
        flag[0] = true; // P0想使用关键区
        while(flag[1]) // 检查P1是不是也想用？
        {
            if(turn == 1) // 如果P1想用，则查看P1是否具有访问权限？
            {
                flag[0] = false; // 如果有，则P0放弃
                while(turn == 1); // 检查turn是否属于P1
                flag[0] = true; // P0想使用
            }
        }
        visit(0); // 访问Critical Partition
        turn = 1; // 访问完成，将权限给P1
        flag[0] = false; // P0结束使用
    }
}

void P1()
{
    while(true)
    {
        flag[1] = true; // P1想使用关键区
        while(flag[0]) // 检查P0是不是也想用？
        {
            if(turn == 0) // 如果P0想用，则查看P0是否具有访问权限？
            {
                flag[1] = false; // 如果有，则P1放弃
                while(turn == 0); // 检查turn是否属于P1
                flag[1] = true; //  P1想使用
            }

        }
          visit(1); // 访问Critical Partition
        turn = 0; // 访问完成，将权限给P0
        flag[1] = false; // P1结束使用
    }
}
```

Peterson算法：flag和turn的含义同Dekker的，但先设turn=别人，且只有flag[别人]和turn=别人同时为真时才循环等待

```
void P0()
{
        while(true)
        {
                flag[0] = true;
                turn = 1;
                while(flag[1] && turn == 1)
                // 退出while循环的条件就是，要么另一个线程
                // 不想要使用关键区，要么此线程拥有访问权限
                {
                        sleep(1);
                        printf("procedure0 is waiting!\n");
                }
                //critical section
                flag[0] = false;
        }
}

void P1()
{
        while(true)
        {
                flag[1] = true;
                turn = 0;
                while(flag[0] && turn == 0)
                {
                        sleep(1);
                        printf("procedure1 is waiting!\n");
                }
                //critical section
                flag[1] = false;
        }
}
```

4.2.3 硬件方法

- 关中断：限制处理器交替执行各进程的能力，不能用于多核
- 专用指令：**比较并交换**，原子指令，一个指令周期内完成，不会被中断
  - TestSet(TS)指令，比较并交换的bool形式

```
int compare_and_swap (int *word, int testval, int newval)
bool testset (int i)
```

– Exchange/swap指令(x86xchg指令)：同上，适用于单核多核，多变量多临界区，但需要忙等待(busy waiting)/自旋等待(spin waiting)，可能饥饿或死锁

```
void exchange (int register, int memory)
```

机器指令方法优点

- 适用于单处理器或共享主存多[核]处理器系统，进程数目任意
- 简单且易于证明
- 可以使用多个变量支持多个临界区

缺点

- 忙等待/自旋等待
- 可能饥饿或死锁

例 3. 利用xchg实现一套互斥机制并给出使用该机制的框架

分析. 由于xchg可以交换两个变量的内容，且为原子操作，故如果临界区未被占用，经过xchg操作后，lock_var被置为*1*；而其他进程再要访问临界区时，lock_var和ax均为*1*，交换后不会发生改变，进而不断进行lock_loop循环。

```
    lock_var db 0 ; not using critical section
lock:
    mov ax, 1
lock_loop:
    xchg [lock_var], ax
    cmp ax, 0
    jnz lock_loop
```

解锁操作则只需将lock_var置*0*即可

```
unlock:
    mov ax, 0
    xchg [lock_var], ax
```

4.3 信号量(semaphore)

解决RC问题一种简单高效的方法

4.3.1 基本操作

记录信号量

- 整数：可用资源数(≥ 0)，需要初始化
- P操作(proberen,semWait)：信号量的值**减1**（申请一个单位的资源），若信号量变为**负数**，则执行P操作进程阻塞，**让权等待**

- V操作(verhogen,semSignal)：信号量的值**加1**（释放一个单位的资源），若信号量**不是正数**（绝对值=现被阻塞的进程数/等待队列的长度），则使一个因P操作被阻塞的进程解除阻塞（唤醒）

需要保证P操作和V操作的**原子性!**

```
struct semaphore {
    int count;
    struct process* L; // 阻塞队列
} s;
void P(semaphore s) { // semWait
    s.count--;
    if (s.count < 0)
        Block(CurruntProcess, s.L);
        // 将当前进程插入该信号量对应的阻塞队列
}
void V(semaphore s) { // semSignal
    s.count++;
    if (s.count <= 0)
        WakeUp(s.L);
}
```

注意P操作是**小于0**，V操作**小于等于0**[4]，且一个进程只会在一个信号量的阻塞队列中。

信号量的优点是简单且表达能力强，用P、V操作可解决多种类型的同步/互斥问题，但不够安全，P、V操作使用不当会产生死锁。

二元信号量省空间，不能代表资源数量，要引入全局变量代表数量。

4.3.2 实现互斥(mutex)

- 对于每一个RC问题，设一个信号量（向系统调用/向内核调用），**初始化为**1
- 所有相关进程在进入临界区**之前**对该信号量进行**P操作**
- 出临界区**之后**进行**V操作**

操作系统实现提供互斥的方法包括信号量和消息传递。

4.3.3 同步

同步：后续动作必须在前驱动作执行完后才能进行

- 对每一个同步关系都要设一个信号量，初值看具体问题（一般为0）
- 在**前驱动作之后**执行V操作（相当于资源产生了）
- 在**后续动作之前**执行P操作

[4]理解为先判断是否小于0（阻塞队列非空），然后再++会比较好

4.3.4 生产者-消费者问题

```
void producer() {
    while (true) {
        produce();
        P(e);

        P(s);
        append();
        V(s);

        V(n); // first
    }
}

void consumer() {
    while (true) {
        P(n); // after

        P(s);
        take();
        V(s);

        V(e);
        consume();
    }
}

// s = initSem(1); // mutex
// n = initSem(0); // # products
// e = initSem(12); // # empty entries in buffer
```

4.3.5 读者写者问题

可以有多个读者，但只有一个写者

读者优先

```
int readercount;
semaphore x = 1, wsem = 1;
void reader() {
    while(true) {
        P(x); // mutex for modifying readercount
        readercount++;
        if (readcount == 1) P(wsem); // mutex, the more readers need not lock
```

```
        V(x);

        READUNIT();

        P(x);
        readercount--;
        if (readcount == 0) V(wsem); // only when no readers, the mutex will be unlocked
        V(x);
    }
}

void writer() {
   while(true) {
       P(wsem);
       WRITEUNIT();
       V(wsem);
   }
}
```

写者优先

```
int readercount, writercount;
semaphore x = 1, y = 1, z = 1, rsem = 1, wsem
    ↪ = 1;
void reader() {
    while(true) {
        P(z); P(rsem);
        P(x);
        readcounter++;
        if (readcounter == 1) P(wsem); // when
            ↪ there is a reader, writer should
            ↪  not write
        V(x);
        V(rsem); V(z);

        READUNIT();

        P(x);
        readcount--;
        if (readcount == 0) V(wsem);
        V(x);
    }
}

void writer() {
    // the same as the original reader
    while(true) {
        P(y);
        writercount++;
        if (writercount == 1) P(rsem);
        V(y);

        P(wsem);
        WRITEUNIT();
        V(wsem);

        P(y);
        writercount--;
        if (writercount == 0) V(rsem);
        V(y);
    }
}
```

读写公平

```
int readercount, writercount;
semaphore x = 1, y = 1, wrsem = 1;
void reader() {
    while(true) {
        P(y);
        P(x);
        readcounter++;
        if (readcounter == 1) P(rwsem);
        V(x);
        V(y);

        READUNIT();

        P(x);
        readcount--;
        if (readcount == 0) V(rwsem);
        V(x);
    }
}

void writer() {
    while(true) {
        P(y);
        P(wrsem);

        WRITEUNIT();

        V(wrsem);
        V(y);
    }
}
```

4.4 管程(monitor)

管程(monitor)由程序设计语言提供，通过集中管理（封装同步机制与同步策略）以保证安全（类似于OOP中的抽象类），实现同步互斥等功能

主要特点：

- 本地变量只能由管程过程访问（封装）
- 进程通过调用管程过程进入管程（调用）
- 每次只能一个进程在执行相关管程的过程（互斥）

主要缺陷

- 可能增加了两次多余的进程切换
- 对进程调度有特殊要求（不允许插队）

4.5 消息传递

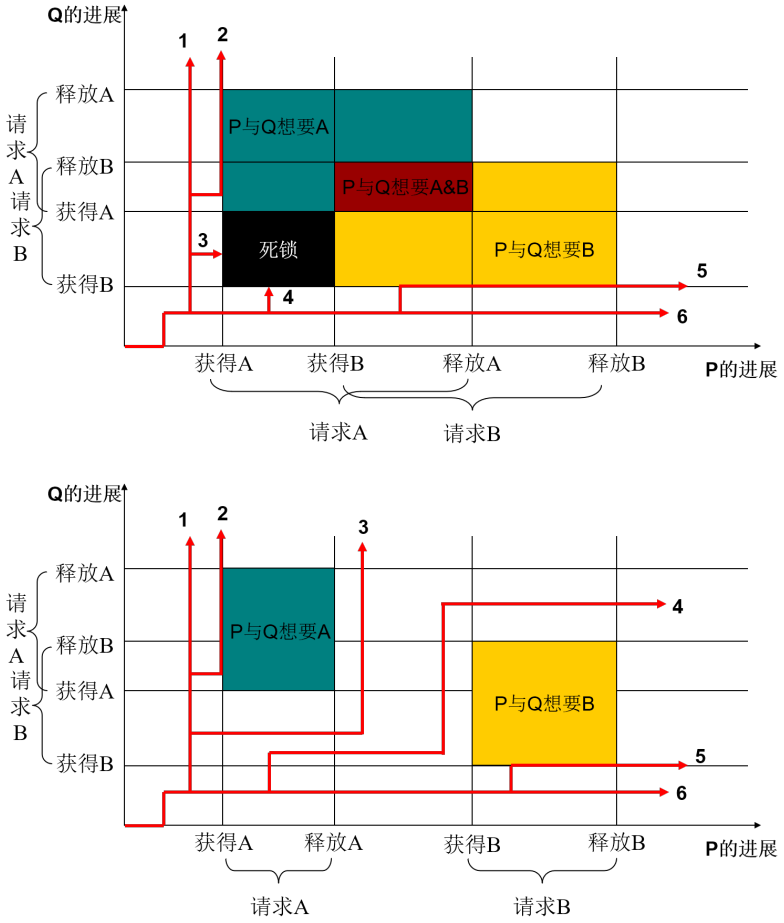
send和receive。

同样可以实现互斥，相当于在进程间传递一个可使用临界区的令牌

5 并发—死锁与饥饿

5.0.1 死锁

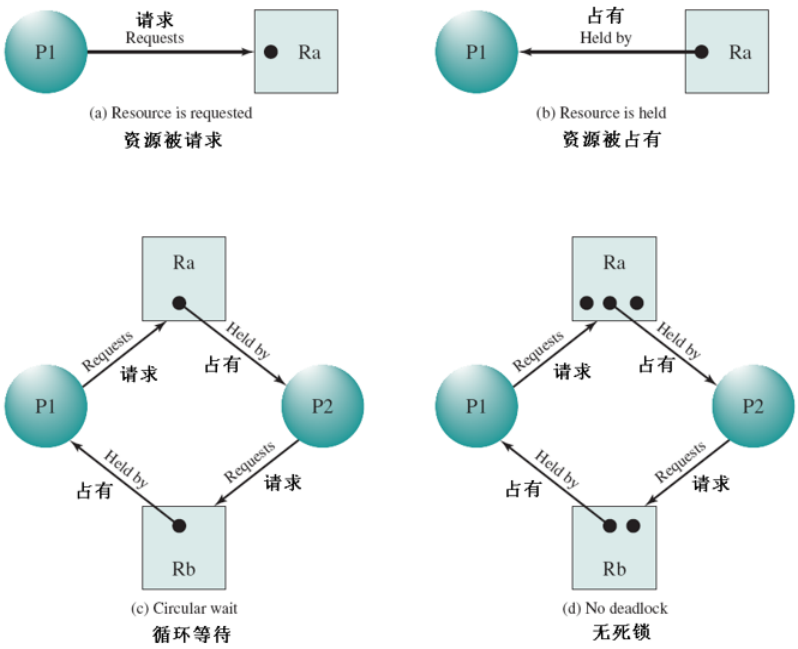
死锁(deadlock)：一个进程集合中的每个进程都在等待只能由该集合中的其他一个进程才能引发的事件（释放占有资源/进行某项操作）



联合进程图，上图死锁，下图无死锁

原则	资源分配策略	不同的方案	主要优点	主要缺点
预防	保守，预提交资源	一次性请求所有资源	• 对执行单一突发行为的进程有效 • 不需抢占	• 低效 • 延时进程的初始化 • 进程必须知道未来的资源请求
		抢占	• 便于用于状态易保存和恢复的资源	• 过多的不必要抢占
		资源排序	• 通过编译时检测可实施 • 由于问题已在系统设计时解决，不需运行时再计算	• 不允许增加资源请求
避免	位于检测和预防中间	操纵以发现至少一条安全路径	• 不需抢占	• OS必须知道未来的资源请求 • 进程可能被长期阻塞
检测	自由	周期性地调用以测试死锁	• 不会延时进程的初始化 • 易于在线处理	• 丢失固有抢占

死锁定理：资源分配图中存在环路是存在死锁的充分必要条件（每个资源类中只包含一个资源实例）。其中指向资源的是申请边，离开资源的是分配边。



Examples of Resource Allocation Graphs

资源分配图的例子

产生死锁必须**同时**满足以下四个条件，只要其中一条不成立，则死锁不会发生

- 互斥：进程对所分配的资源排他性控制，在一段时间内该资源仅为一个进程所占有。
- 不可抢占（不剥夺）：进程所获得的资源在未使用完毕之前，不能被其他进程强行夺走，只能主动释放
- 占有且等待（请求和保持）：进程已经保持了至少一个资源，但又提出了新的资源请求，而该资源已经被其他进程占有。此时请求进程被阻塞，但对自己已获得的资源保持不放。
- 循环等待：存在一种进程资源的循环等待链，链中每一个进程已获得的资源同时被链中下一个进程所请求。

5.1 死锁预防

5.1.1 破坏互斥条件

允许多个进程同时使用资源，但不适用于绝大多数资源，适用条件：

- 资源的固有特性允许多个进程同时使用（如文件允许多个进程同时读）
- 借助特殊技术允许多个进程同时使用（如打印机的SPOOLing技术）

5.1.2 破坏占有且等待条件

禁止已拥有资源的进程再申请其他资源，如要求所有进程在开始时一次性地申请在整个运行过程所需的全部资源；或申请资源时要先释放其占有资源后，再一次性申请所需全部资源

- 优点：简单、易于实现、安全
- 缺点：进程延迟运行，资源严重浪费

5.1.3 破坏不可剥夺条件

一个已经占有了某些资源的进程，当它再提出新的资源请求而不能立即得到满足时，必须释放它已经占有的所有资源，待以后需要时再重新申请；OS可以剥夺一个进程占有的资源，分配给其他进程

适用条件：资源的状态可以很容易地保存和恢复（如CPU）缺点：实现复杂、代价大，反复申请/释放资源、系统开销大、降低系统吞吐量

5.1.4 破坏环路等待条件方法

- 要求每个进程任何时刻只能占有一个资源，如果要申请第二个则必须先释放第一个（不现实）
- 对所有资源按类型进行**线性排队**，进程申请资源必须严格按资源序号递增的顺序（可避免循环等待）

缺点

- 很难找到令每个人都满意的编号次序，类型序号的安排只能考虑一般作业的情况, 限制了用户简单、自主地编程
- 易造成资源的浪费（会不必要地拒绝对资源的访问）
- 可能低效（会使进程的执行速度变慢）

5.2 死锁避免

进程启动拒绝：考虑一个有n个进程和m种不同类型资源的系统。定义以下向量和矩阵：

Resource=$\mathbf{r}=\begin{bmatrix}R_1 & R_2 & \cdots & R_m\end{bmatrix}$	系统中每种资源的总量
Available=$\mathbf{v}=\begin{bmatrix}V_1 & V_2 & \cdots & V_m\end{bmatrix}$	未分配给进程的每种资源的总量
Claim=$C=\begin{bmatrix}C_{11} & \cdots & C_{1m}\\ \vdots & \ddots & \vdots\\ C_{n1} & \cdots & C_{nm}\end{bmatrix}$	C_{ij}为进程i对资源j的请求
Allocation=$A=\begin{bmatrix}A_{11} & \cdots & A_{1m}\\ \vdots & \ddots & \vdots\\ A_{n1} & \cdots & A_{nm}\end{bmatrix}$	A_{ij}为分配给进程i资源j的数目

有下列关系式成立：

1. $\forall j:\ r_j = v_j + \sum_{i=1}^{n} A_{ij}$。所有资源要么可用，要么已经被分配
2. $\forall i,j:\ A_{ij} \le C_{ij} \le R_i$：分配小于等于最大请求，最大请求小于等于资源总量

进而可以定义死锁避免策略：若一个新进程的资源需求会导致死锁，则拒绝启动这个进程，当且仅当

$$\forall j:\ R_j \ge C_{(n+1)j} + \sum_{i=1}^{n} C_{ij}$$

进程分配拒绝：银行家算法(Dijkstra)

$$\forall j:\ C_{ij} - A_{ij} \le v_j$$

- 只要系统处于安全状态（至少有一个资源分配序列不会导致死锁，即所有进程都能运行到结束），必定不会进入死锁状态
- 不安全状态不一定是死锁状态，但不能保证不会进入死锁状态
- 如果一个新进程的资源请求会导致不安全状态，则拒绝启动这个进程

优点

- 比死锁预防限制少
- 无须死锁检测中的资源剥夺和进程重启

缺点

- 必须事先声明每个进程请求最大资源
- 进程必须无关，没有同步要求
- 分配的资源数目是固定的
- 占有资源时进程不能退出

例 4. 在如下条件下考虑银行家算法。

6个进程：P0-P5

4种资源：A（15单位）、B（6单位）、C（9单位）、D（10单位）

时间T0时的情况，左侧为分配矩阵A，右侧为请求矩阵C

	A	B	C	D	A	B	C	D
P0	2	0	2	1	9	5	5	5
P1	0	1	1	1	2	2	3	3
P2	4	1	0	2	7	5	4	4
P3	1	0	0	1	3	3	3	2
P4	1	1	0	0	5	2	2	1
P5	1	0	1	1	4	4	4	4
v	6	3	5	4				
All	15	5	9	10				

分析. 需求矩阵$C - A$

P0	7	5	3	4
P1	2	1	2	2
P2	3	4	4	2
P3	2	3	3	1
P4	4	1	2	1
P5	3	4	3	3

看有没有进程所请求的资源都小于等于可用资源，用完则将当前分配资源还回可用资源中

原来	6	3	5	4
P1	6	4	6	5
P2	10	5	6	7
P3	11	5	6	8
P4	12	6	6	8
P5	13	6	7	9
P0	15	6	9	10

如果某一个进程的请求后，需求矩阵每一行均小于等于可用资源的话，基于死锁避免原则，则该请求应该被拒绝

5.3 死锁检测

检测方法：

- 单个资源实例：检测资源分配图中是否存在环路（依据——死锁定理）
- 多个资源实例：类似银行家算法的安全检查

恢复：

- 剥夺法：连续剥夺资源指导不存在死锁
- 回退法：将每个死锁进程回滚到前面定义的某些检查点(checkpoint)，并重启所有进程（死锁可能重现）
- 杀死进程法

5.3.1 饥饿

饥饿：一组进程中，某个或某些进程**无限等待**该组进程中其他进程所占用的资源
进入饥饿状态的进程可以只有一个，而由于循环等待条件而进入死锁状态的进程必须大于等于两个

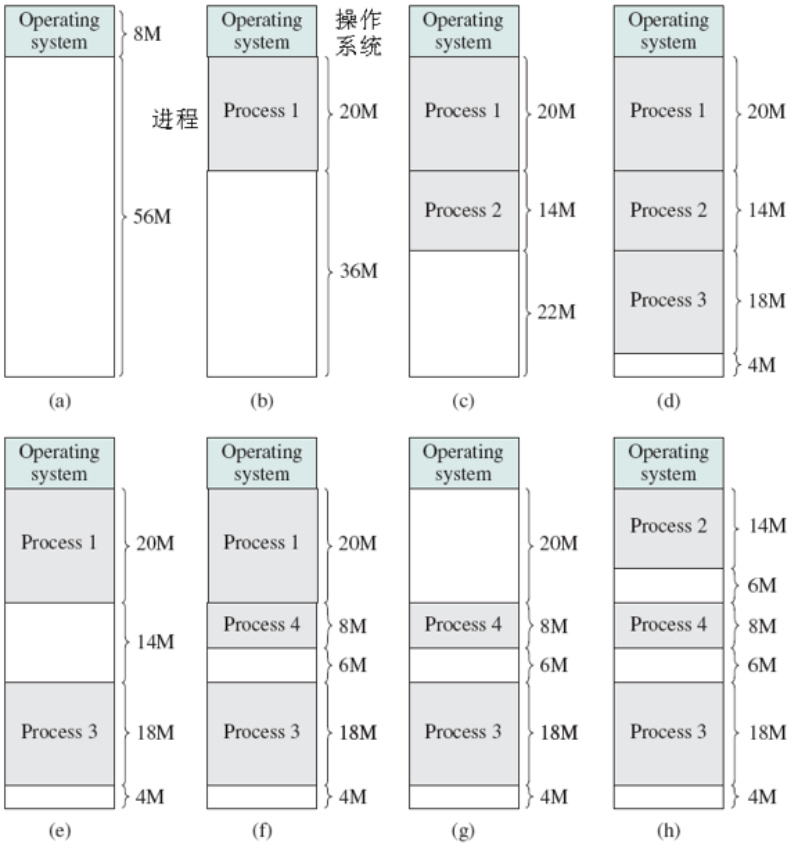
6 内存管理

6.1 分区存储管理

固定分区

- 等长分区：大进程则只能部分载入，小进程将产生内碎片
- 不等长分区：一定程度上缓解等长分区的问题

动态分区：会存在外碎片；若采用压缩方式移动进程使其紧靠，则非常耗时，需要进行动态重定位



The Effect of Dynamic Partitioning
动态分区的效果

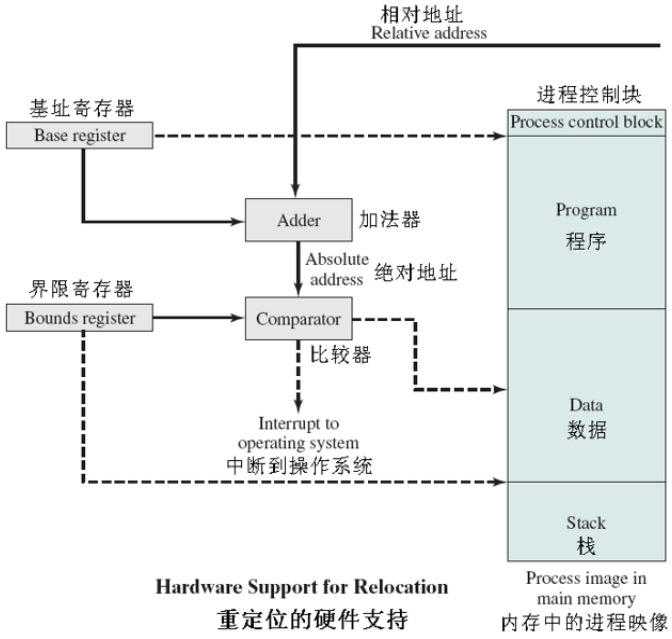
动态分区放置算法：

- 首次适配(first fit)：**最简单也性能最好**，从前端开始扫描内存，直到找到一个足够大的空闲区，通常性能比较好
- 下次适配/邻近适配(next fit)：从上次分配结束的地方开始扫描内存，直到找到一个足够大的空闲区
- 最佳适配(best fit)算法：**性能最差**，扫描整个内存，找出一个足够大的最小的空闲区，会产生很多外部碎片

分区存储管理中，存储保护硬件由基址和限长两个寄存器配合地址转换。

伙伴系统(buddy system)：固定分区和动态分区的折中方案

- 可用内存块大小为$2^K, L \leq K \leq U$
- 初始空间大小为2^U
- 若请求空间大小$s < 2^{U-1}$，则对分现有块



重定位的硬件机制

6.2 页式存储管理

分页(paging)

- 将主存划分为许多等长的帧/页框(frame)
- 将进程划分为若干页(page)
- 进程加载时，所有页面被载入可用帧，同时建立页表

设页大小为L，逻辑地址A，物理地址E，则

$$\text{页号}P = A/L \qquad \text{页内偏移量}W = A\%L$$

例 5. *16*位编址，若页面大小为*1K(1024)*，则需（低）*10*位表示页内偏移，剩下（高）*6*位表示页号，则

- 相对地址为1502的逻辑地址$= 1024 + 478 = (1, 478)$
- 逻辑地址为$(1, 478)$的相对地址$= 1 * 1024 + 478 = 1502$

类似固定分区，不同在于：

- 分页中的页帧非常小（从而内碎片也小）
- 分页中一个进程可占用多个页帧（从而不需要覆盖）
- 分页中不要求一个进程占用的多个页帧连续（充分利用空闲“分区”）

存在问题：

- 不易实现共享和保护（不反映程序的逻辑组织）
- 不便于动态链接（线性地址空间）
- 不易处理数据结构的动态增长（线性地址空间）

6.3 段式存储管理

将程序及数据划分成若干段(segment)（不要求等长，但不能超过最大长度）

- 分页是出于系统管理的需要，分段是出于用户应用的需要：一条指令或一个操作数可能会跨越两个页的分界处，而不会跨越两个段的分界处
- 页大小是系统固定的，而段大小则通常不固定
- 逻辑地址表示
  - 分页是一维的，各个模块在链接时必须组织成同一个地址空间
  - 分段是二维的，各个模块在链接时可以每个段组织成一个地址空间
- 通常段比页大，因而段表比页表短，可以缩短查找时间，提高访问速度
- 分段对程序员可见，从而可用来对程序和数据进行模块化组织
- 分段方便实现模块化共享和保护，如程序可执行、数据可读写（段表表项要有保护位）
- 都存在外碎片，但分段中可通过减少段长来减轻外碎片浪费程度
- 分段中一个进程可占用多个“分区”，不要求一个进程占用的多个“分区”连续（但一般要求一个段所占用的多个“分区”连续）
- 分段克服了分页存在的问题（数据结构的动态增长、动态链接、保护和共享）
- 分段存在外碎片，分页只有小的内碎片，分页内存利用率比分段高

总的来说，分段反映了程序的逻辑组织、易实现保护和共享、便于动态链接和数据结构的动态增长（线性地址空间），但是分段会产生外部碎片，段的长度不一，不利于虚拟存储；分页采用较小的等长分块、内部碎片小，而且可以不连续存储、无外部碎片，易于部分加载和交换、支持虚拟存储，但是分页不支持保护、共享、动态链接和增长；所以分段用于**内存保护**、分页用于**虚拟存储**。

段表只能有一个，而页表可以有多个。

段页式系统中，逻辑地址被分为段号S、页号P和页内偏移量W。

逻辑地址偏移量只需小于段的长度即可。

6.4 虚拟存储

6.4.1 简介

传统的存储方式都是**一次性**加载，并且**驻留**在内存中。而虚拟存储器则是基于程序的局部性原理，在程序装入时，将程序一部分装入内存，其余部分留在外存。 动态地址转换、不连续分配和部分加载是虚拟存储的主要特征。

- 采用部分加载，内存中可同时容纳更多的进程。每个进程都只加载一部分，更多进程中应该也会有更多的就绪进程，从而提高CPU利用率
- 采用部分加载，进程可以比内存大，实现了虚拟存储
  - 用户程序可以使用的独立于物理内存的逻辑地址单元组成存储空间(虚拟存储)

- 逻辑地址空间可以比物理地址空间大，例如，设物理内存64KB，1KB/页，则物理地址需要16位，而逻辑地址可以是28位!
- 虚拟存储由内存和外存结合实现

- 程序重定位

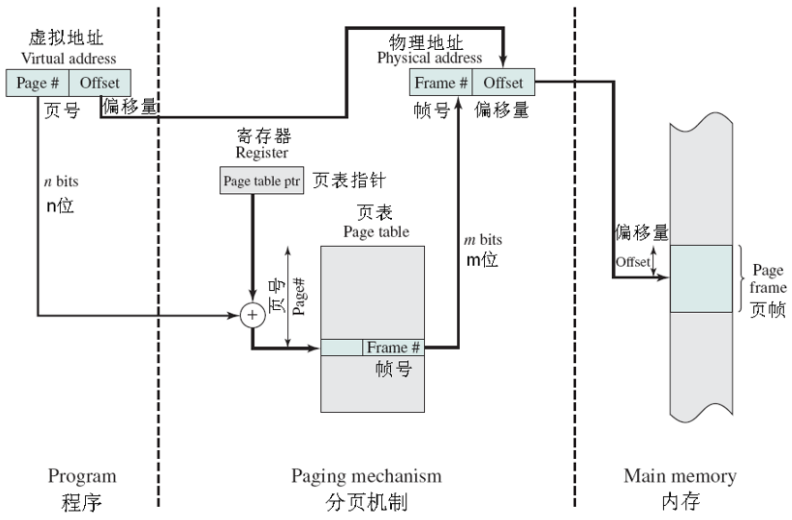
抖动(thrashing)问题：交换操作太过频繁

6.4.2 页表

- 页表项（Page Table Entry，简称为PTE）的一般内容：
  - Present：在/不在内存
  - Modified：有没有被修改
  - Protection：保护码，1位或多位(rwe：读/写/执行)
  - Referenced：有没有被访问
  - Cache：是否禁止缓存
- 页表长度不定，取决于进程大小
- 不适合用寄存器存储页表，而是存放在内存
- 页表起始地址保存在一个CPU专用寄存器里(`cr3`)

虚拟地址转换为物理地址流程：

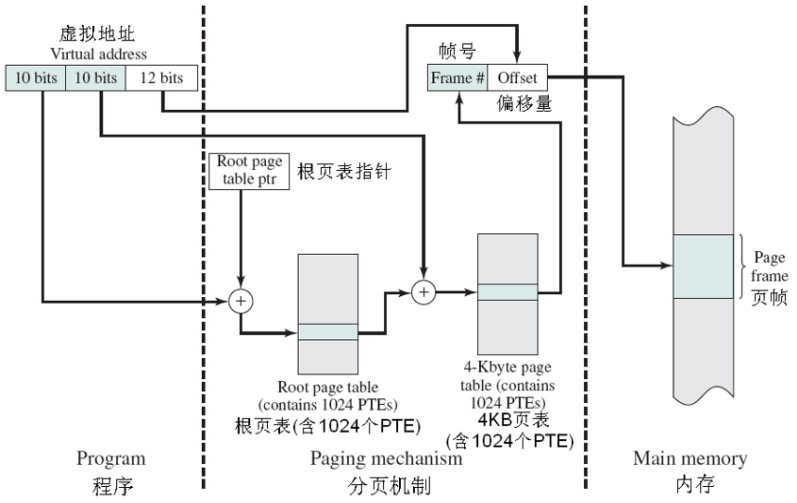
- 将虚拟地址分割为虚页号和偏移量两个部分
- 通过虚页号在页表中寻找对应的表项
- 从中获取页框号后，乘页尺寸，加上偏移量，得到物理地址



Address Translation in a Paging System

分页系统中的地址转换

由于分页后页表项太多了，故要采用多级页表，通常32位CPU用两级页表，64位CPU用三级页表



Address Translation in a Two-Level Paging System

二级分页系统中的地址转换

例 6. *在Intel x86系列的32位CPU 中，分页硬件用二级页表结构。页大小为4KB，一级页表/根页表/页目录和二级页表/用户页表的每个表项占4B。回答下列问题：*

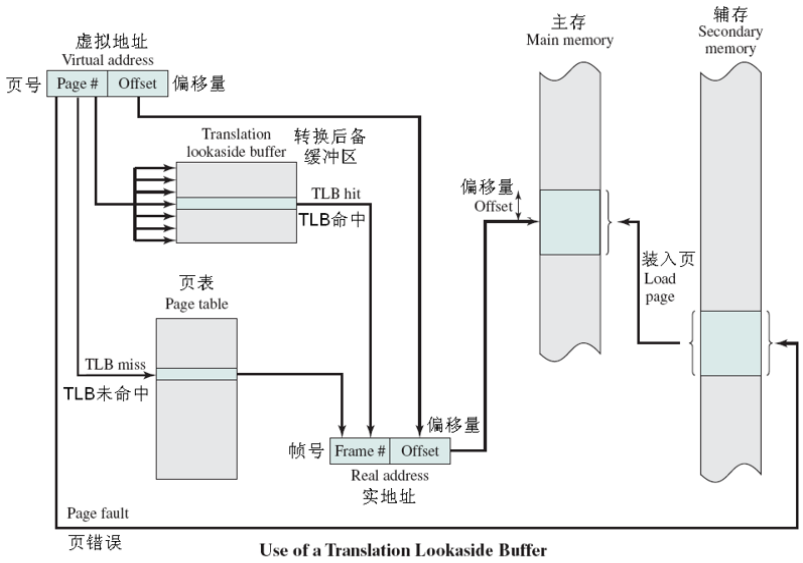
1. *32位的线性地址中，根页表的索引、用户页表的索引和页内偏移量各占哪些位。*
2. *如果有一个十六进制的线性地址为 01E5F1A4，那么对应的页目录索引值、页表索引值和页内偏移量分别是多少？*
3. *如果进程实际地址空间使用了20MB，那么该进程的根页表和用户页表中有用表项占用多少内存？*

分析. *用户地址空间4GB，二级页表4MB，一级页表4KB*

1. *二级页表：$2^{32}/2^{12} = 2^{20}$页表项，每个表项4B，故是4MB*
   一级页表是二级页表的页表，同样以4KB划分一页，故有$4MB/4KB = 2^{10}$个页表项
   因此32位线性地址[5]，高10位为根页表索引，中间10位位用户页表索引，低12位为页内偏移量
2. *按照上面的划分可得一级0x007，二级0x25F，页内偏移0x1A4*
3. *该进程需要用$20 * 2^{20}B/2^{12}B = 20 * 2^8$个用户页，在二级页表中占$20 * 2^8 * 4B/2^{12}B = 5$个页大小，故总共的表项占用空间为$(256 * 20 + 5) * 4B = 20500B$*

快表/联想存储器(TLB)相当于页表的cache

[5]逻辑地址cs:eip，线性地址即虚拟32位地址，物理地址为真实32位地址



Use of a Translation Lookaside Buffer

转换后备缓冲区的用法

常见页面大小介于1KB-8KB

- 分配策略：确定驻留集大小（每个进程读入主存多少页），分配给一个进程的存储量越小，则任何时候驻留在主存中的进程数就越多，提高CPU利用效率，但单一程序的局部性差
  - 固定分配局部置换：为每个进程分配一定数目的物理块，在整个运行时间都不改变。若进程在运行时发生缺页，则只能从该进程在内存的页面中选出一页换出，再调入需要的页。难以确定每个进程应分配的物理块数目：太少会频繁出现缺页中断，太多会使CPU及其他资源利用率下降。
  - 可变分配全局置换：为每个进程分配一定数目的物理块，操作系统自身也保持一个空闲的物理块队列。缺页时则从空闲物理块中取出一个分配给该进程。可以动态增加进程物理块，但是盲目增加物理块数目会导致系统多道程序并发能力下降。
  - 可变分配局部置换：为每个进程分配一定数目的物理块，进程缺页时，只允许进程从内存的页面中选一页换出；若该进程频繁缺页，则操作系统再为该进程分配若干物理块，直到该进程缺页率趋于适当程度；反之缺页率低的就减少分配的物理块。但开销较大。
- 调页策略：按需调页、预先调页
- 替换策略：
  - Opt(Belady)：置换下次访问距当前时间最长的页，最长时间内不被访问，如页1之后都不再用，则将页1替换出去
  - LRU：最近最少使用
  - FIFO：先进先出，可能出现Belady异常（物理块数增大而页故障数不降反增）
  - Clock：时钟/最近未用(NRU)，比较实用的算法。需要附加位，**首次/被访问**时置为1；指针始终指向下一访问的表项；不管命中哪一个表项，指针都不会移动；**当需要置换时**，顺序循环扫描页表，**将1置0**，并选择第一个原来就是0的页表项放置

例 7. 时钟调度的例子如下

1	*4*	*5*	*2*	*1*	*4*	*3*	*5*	*4*	*3*	*1*	*2*	*1*	*5*
*1**	*1**	*1**	*→1**	*→1**	*→1**	*3**	*3**	*3**	*3**	*→3**	*3*	*3*	*3*
→	*4**	*4**	*4**	*4**	*4**	*→4*	*→4*	*→4**	*→4**	*4*	*2**	*2**	*2**
	→	*5**	*5**	*5**	*5**	*5*	*5**	*5**	*5**	*5*	*→5*	*→5*	*→5**
		→	*2**	*2**	*2**	*2*	*2*	*2*	*2*	*1**	*1**	*1**	*1**
F	*F*	*F*	*F*			*F*				*F*	*F*		

Linux的内存管理

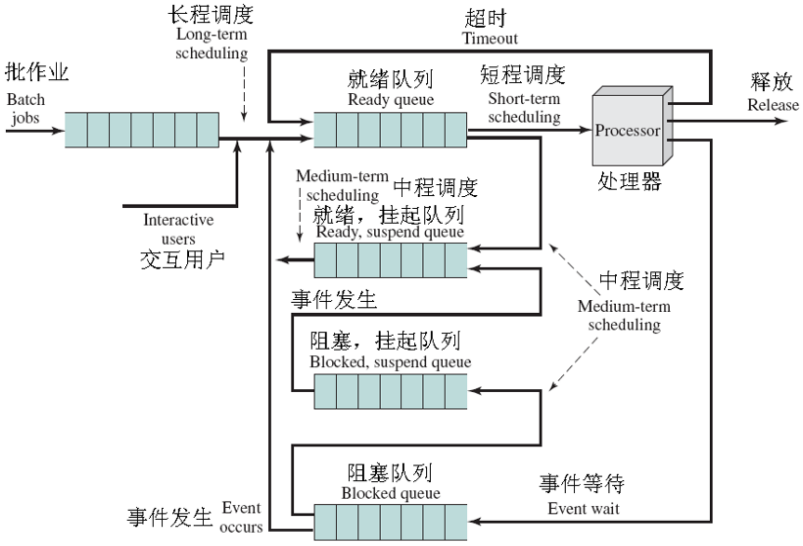
- 虚拟存储采用三级页表
- 页面分配采用伙伴系统
- 页面替换采用时钟算法

7 调度

7.1 单处理器调度

三级调度层次

- 长程调度(Long-term scheduling)/任务调度
  - 决定哪些新建进程可进入系统**准备执行**(ready)
  - 控制多道程序系统的并发程度
  - 进程越多则各进程对CPU的使用百分比越小
- 中程调度(Medium-term scheduling)
  - 决定交换哪些主存-辅存（内存-外存）进程
  - 基于多道程序设计的管理需要
- 短程调度(Short-term scheduling)/CPU调度
  - 决定下一个使用CPU的进程（dispatcher，分派程序）



Queueing Diagram for Scheduling
调度的队列图

进程调度方式

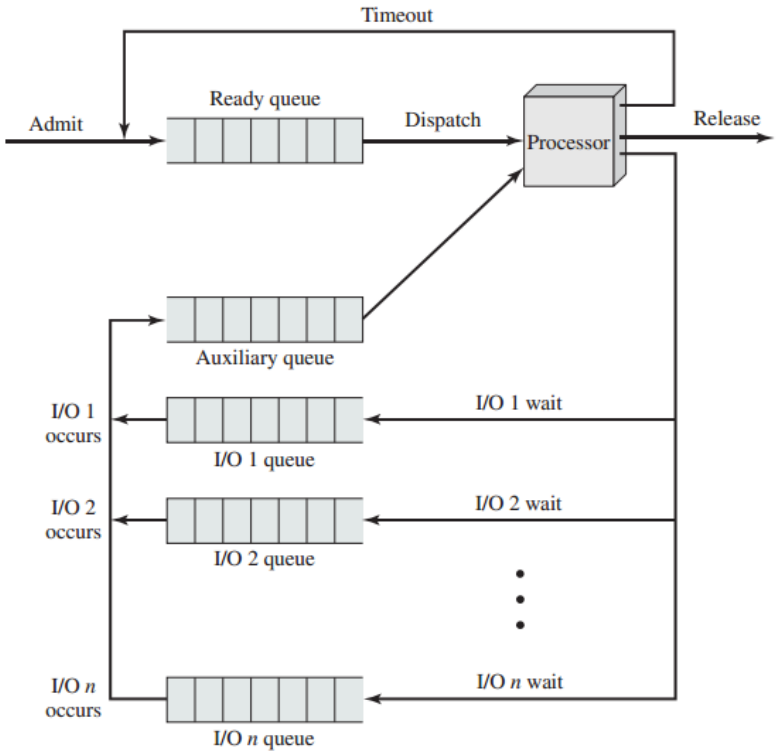
- 剥夺式：立即分配
- 非剥夺式：当前进程执行完再分配给新进程

评价指标：

- 周转时间/驻留时间T_r：作业**提交到完成**所经历的时间，等待时间+服务时间
- 归一化周转时间：周转时间与服务时间的比值，表示一个进程的相对延迟情况
- 吞吐量：单位时间内**完成的进程数量**

进程短程调度算法

- 先来先服务(First Come First Served，FCFS)：公平，更有利于CPU密集型和长作业，不利于短作业
- 时间片轮转(time slicing Round Robin，RR)：最公平，兼顾长短作业，也是剥夺式调度（时间片用完时）；平均等待时间较长，上下文切换浪费时间，对IO密集型进程最不利
- 虚拟时间片轮转(VRR)：进程因IO而阻塞会进入到专门的IO队列中，解除了IO阻塞的进程会被转移到一个FCFS的辅助队列中。进行调度决策时，辅助队列中的进程优先于就绪队列中的进程。提升了公平性。



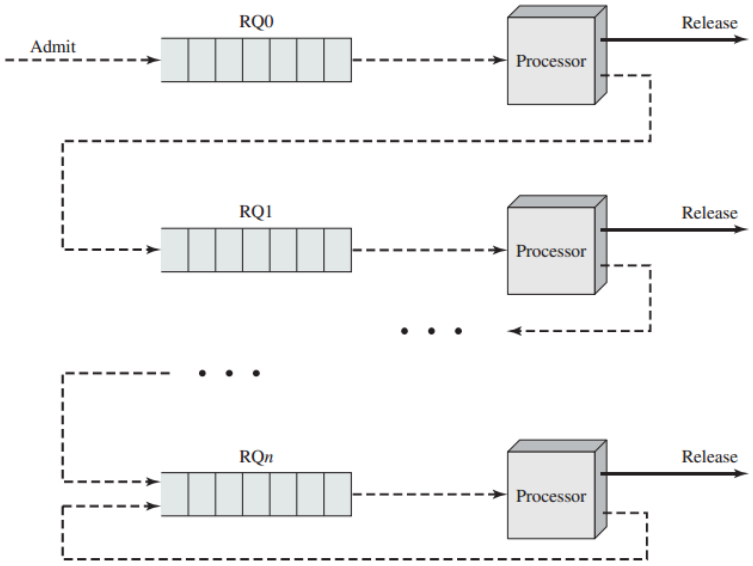
- 最短进程优先(Shortest Process Next，SPN)：非抢占式策略，但需要用EWMA预测，吞吐量最大，长作业会饥饿

$$S_{n+1} = aT_n + (1-a)S_n$$

- 最短剩余优先(Shortest Remaining Time，SRT)：SPN结合**抢占策略**；但必须记录过去的服务时间，增加开销
- 最高响应比优先(Highest Response Ratio Next，HRRN)：兼顾长短作业，计算响应比开销大，**非剥夺型**；注意T_{wait}是截至当前等待处理器的时间，因此长进程由于得不到服务，R_P会一直增加，最终在竞争中胜过短进程。

$$R_P = \frac{T_{wait} + T_{serve}}{T_{serve}}$$

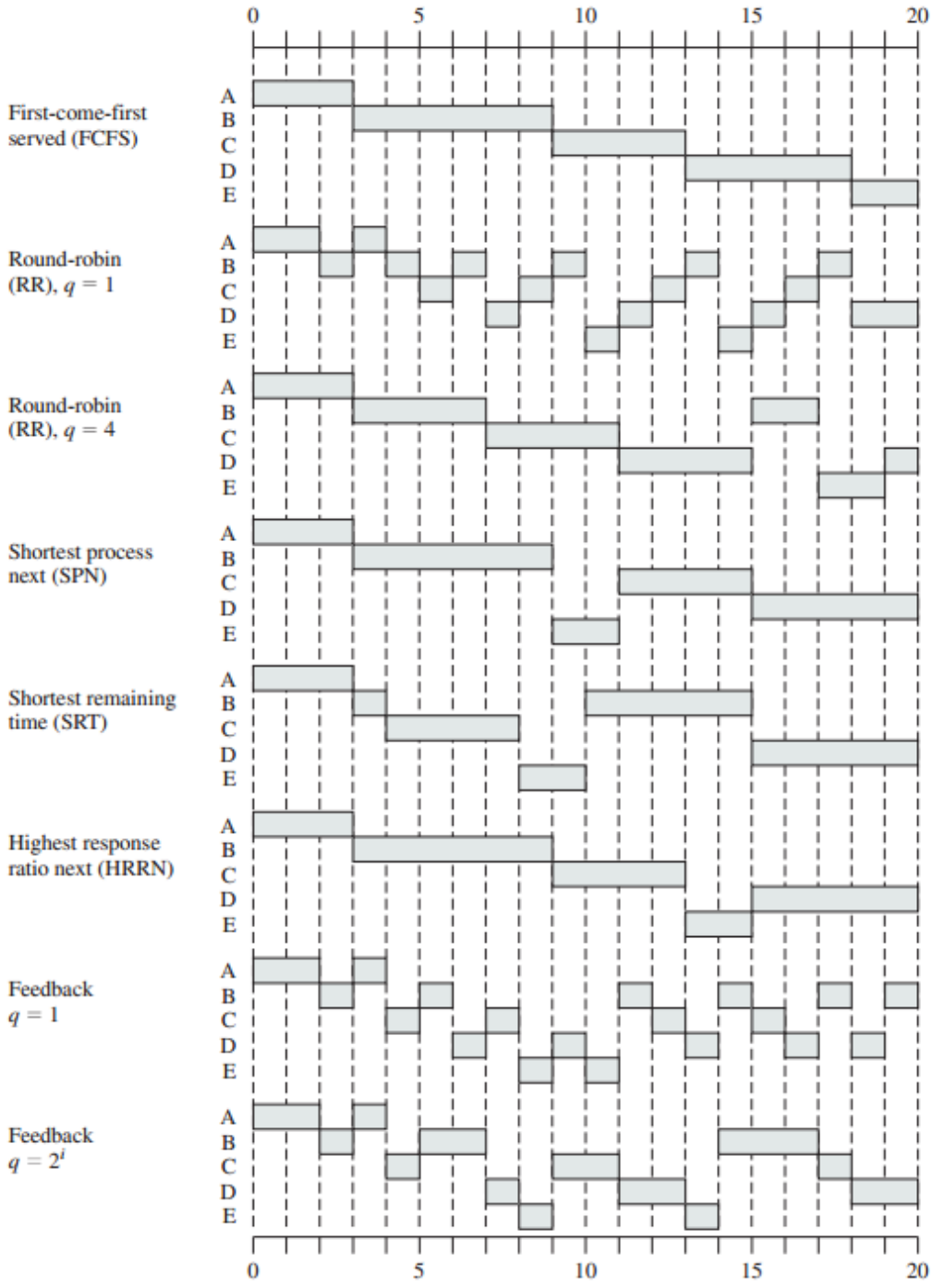
- 最高优先级优先(Highest Priority First，HPF)
- 多级队列反馈(Multilevel Feedback，MF/FB)：结合FCFS、RR和HPF，设置多条优先级不同的队列，前面的队列用FCFS，最后一条队列用RR；每次时间片到都会降级到下一个优先级中，短进程很快执行完毕，而长进程则会多次降级；$q = 1$类似于时间片为1的轮转法，$q = 2^i$则是对于第i个队列，可以执行$q = 2^i$个周期才被抢占



例 8. 考虑下面的进程

进程	到达时间	服务时间
A	*0*	*3*
B	*2*	*6*
C	*4*	*4*
D	*6*	*5*
E	*8*	*2*

分析. 有下面的时序图


First-come-first served (FCFS)
Round-robin (RR), $q = 1$
Round-robin (RR), $q = 4$
Shortest process next (SPN)
Shortest remaining time (SRT)
Highest response ratio next (HRRN)
Feedback $q = 1$
Feedback $q = 2^i$
A
B
C
D
E
0
5
10
15
20

Process	A	B	C	D	E	
Arrival Time	0	2	4	6	8	
Service Time (T_s)	3	6	4	5	2	Mean
FCFS						
Finish Time	3	9	13	18	20	
Turnaround Time (T_r)	3	7	9	12	12	8.60
T_r/T_s	1.00	1.17	2.25	2.40	6.00	2.56
RR $q = 1$						
Finish Time	4	18	17	20	15	
Turnaround Time (T_r)	4	16	13	14	7	10.80
T_r/T_s	1.33	2.67	3.25	2.80	3.50	2.71
RR $q = 4$						
Finish Time	3	17	11	20	19	
Turnaround Time (T_r)	3	15	7	14	11	10.00
T_r/T_s	1.00	2.5	1.75	2.80	5.50	2.71
SPN						
Finish Time	3	9	15	20	11	
Turnaround Time (T_r)	3	7	11	14	3	7.60
T_r/T_s	1.00	1.17	2.75	2.80	1.50	1.84
SRT						
Finish Time	3	15	8	20	10	
Turnaround Time (T_r)	3	13	4	14	2	7.20
T_r/T_s	1.00	2.17	1.00	2.80	1.00	1.59
HRRN						
Finish Time	3	9	13	20	15	
Turnaround Time (T_r)	3	7	9	14	7	8.00
T_r/T_s	1.00	1.17	2.25	2.80	3.5	2.14
FB $q = 1$						
Finish Time	4	20	16	19	11	
Turnaround Time (T_r)	4	18	12	13	3	10.00
T_r/T_s	1.33	3.00	3.00	2.60	1.5	2.29
FB $q = 2^i$						
Finish Time	4	17	18	20	14	
Turnaround Time (T_r)	4	15	14	14	6	10.60
T_r/T_s	1.33	2.50	3.50	2.80	3.00	2.63

7.2 多处理器调度

多处理器线程调度方案

- 负载共享(load sharing)
- 组调度(gang scheduling)
- 专用处理器分配
- 动态调度

7.3 调度算法总结

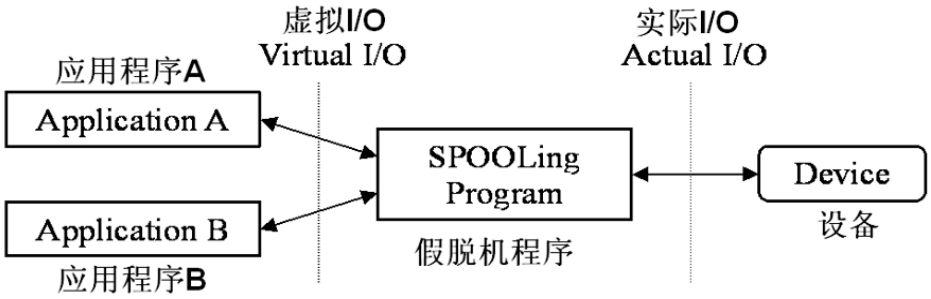
- 动态分区放置算法：First-fit、Best-fit、Next-fit
- 页的替换策略：Opt、LRU、FIFO、Clock
- 单处理器进程调度：FCFS、SPN、SRT、HRRN、RR、HPF、MF/FB
- 磁盘调度算法：FCFS、SSTF、SCAN、C-SCAN

8 IO管理与磁盘调度

控制设备和内存或CPU之间的数据传送方式：

- 程序控制(programmed IO)：需要不断读取IO控制器的状态寄存器，才能决定是否读写数据
- 中断驱动方式(interrupt-driven IO)：IO设备主动打断CPU运行并请求服务
- 直接内存访问(Direct Memory Access, DMA)：直接在IO设备和内存之间开辟直接的数据交换通路

假脱机技术(Simultaneous Peripheral Operations On Line, SPOOL)/虚拟设备技术：专门利用一道程序来完成对设备的IO操作，而无需使用外围IO处理机



优点是高速虚拟IO操作，实现对独享设备的共享

IO缓冲：CPU快IO慢，提高外设利用率，尽可能使外设处于忙的状态（多道程序并发）；增加缓冲区有利于提高命中率，操作系统采用软件方法实现缓冲技术

- 单方向缓冲：单缓冲、双缓冲、环形缓冲
- 双方向缓冲：缓冲池(buffer pool)
- 循环缓冲

磁盘调度算法

- 先来先服务FCFS：公平简单，平均寻道距离大
- 最短寻找时间优先(shortest seek time first, SSTF)：与当前磁头邻近的磁道，性能比FCFS好，可能出现饥饿；移臂距离最小
- 电梯算法/扫描算法SCAN：比较实用的调度算法：在最短寻找时间优先算法的基础上规定了磁头的方向，不利于磁头一端的访问请求
- 循环扫描算法C-SCAN：到头会回到另一侧重新开始，移臂方向改变最少

Windows支持两类RAID配置

- 硬件RAID：若干独立的物理磁盘，通过磁盘控制器或磁盘存储柜，组合成逻辑磁盘。冗余信息的创建和重新生成由控制器负责处理
- 软件RAID：不连续的磁盘空间通过容错软件磁盘驱动程序(Fault-Tolerant software DISK driver, FTDISK)组合成逻辑分区。Windows的软件RAID机制实现了RAID1（镜像）和RAID5（交错块分布式奇偶校验）

9 文件管理

9.1 文件系统概述

文件是以计算机硬盘为载体存储在计算机上的信息集合。在系统运行时，计算机以进程为基本单位进行资源调度和分配；而在用户进行的输入输出中，则以**文件**为基本单位。

文件的存取方式

- 顺序存取
- 随机/直接存取：非顺序或按关键字，现代OS常用

文件目录：将所有文件控制块(File Control Block, FCB)组织在一起，一个FCB称之为一个目录项，提供文件名到文件内容/数据的映射，包含

- 基本信息：文件名、文件类型、文件结构
- 地址信息：存放位置、文件长度
- 访问控制信息：主人、权限
- 使用信息：创建时间、最后一次访问时间和用户

每创建一个文件，系统都会分配一个FCB放在文件目录中，成为目录项。

UNIX采用文件名与文件描述信息分开的方法，文件描述信息单独形成一个称为索引结点的数据结构，称为i结点。

FAT12文件系统中，一个FCB大小为32B，软盘扇区大小为0.5KB，则每个扇区可以存16个目录项。一般地，一个FCB大小为64B，盘块大小1KB，则在每个盘块中可以存放16个FCB（FCB必须连续存放）。而在UNIX系统中一个目录项仅占16B，其中14B为文件名，2B为i结点指针，在1KB的盘块中可以存放64个目录项。

- 一级目录：整个目录组织是一个线性结构，系统中所有文件都建立在一张目录表中
  优缺点：
  - 按名存取，结构简单、易实现
  - 文件多时目录检索时间长，从而平均检索时间长
  - 有命名冲突：如多个文件有相同的文件名或一个文件有多个不同的文件名
- 二级目录：在根目录/第一级目录/主文件目录MFD下，每个用户对应一个第二级目录/用户目录UFD，在用户目录下是该用户的文件，而不再有下级目录

- 多级目录：上下级关系，但需要按路径名逐级访问中间结点，增加磁盘访问次数，进而影响查询速度
  - 当前目录/工作目录.
  - 父目录..
  - 子目录(subdirectory)
  - 根目录(root directory)/

Unix文件系统

- Unix磁盘文件系统结构
- 引导块（块0）
- 超级块（块1）
- i-索引结点表

现代操作系统所采用的文件系统

- DOS文件系统：FAT12、FAT16、FAT32
- Windows文件系统：NTFS
- Linux文件系统：ext2/3

文件共享的方式

- 基于索引结点的共享方式（硬链接）：多个指针指向一个索引结点，当还有一个指针指向索引结点时，索引结点就不能删除
- 利用符号链实现文件共享（软链接）：把到达共享文件的路径记录起来，当要访问文件时，根据路径寻找文件

9.2 文件系统实现

文件系统层次结构通常分为（自顶向下）：用户接口、文件目录系统、存储控制模块、逻辑文件系统与文件信息缓冲区、物理文件系统。

目录实现：

- 线性列表：存储文件名和数据块指针。
- 哈希表：根据文件名得到一个值，并返回指向线性列表中的指针

文件的分配方式：

- 连续分配：磁盘上连续的块，支持顺序访问和直接访问；文件长度不宜动态增加
- 链接分配：目录项给出首尾盘块号(FAT)；离散分配，消除了外部碎片，显著提高磁盘空间利用率；只能依照顺序链查找，且需要多次访存，随机存储效率最差
- 索引分配：直接把每个文件的所有盘块号都集中在一起构成索引块（多开一个单独的扇区）；多层索引使第一层引导块指向第二层引导块，第二层引导块再指向文件块；索引表增加存储空间开销；现代主流操作系统常用

例 9. 某操作系统的文件物理组织方式采用三级索引分配，在FCB中，有*10*个直接数据块指针、*1*个一级间接块指针、*1*个二级间接块指针和*1*个三级间接块指针，每个索引指针占*4B*，磁盘块大小为 *4KB*。回答下列问题：

1. 该文件系统中最大的单个文件有多大？
2. 对一个*20MB*大小的文件，描述其存储组织中有效指针的使用情况。

分析. 每个磁盘块可以存$4KB/4B = 1024 = 1K$个索引指针。

1. *10* 个直接块$10*4KB = 40KB$、*1*个一级间接块$1K*4KB = 4MB$、*1*个二级间接块$1K*1K*4KB = 4GB$、*1*个三级间接块$1K*1K*1K*4KB = 4TB$，最大文件为$4TB4GB4MB40KB$。
2. $20MB = 20480KB$，*10* 个直接块全被使用（$10*4KB = 40KB$）、*1*个一级间接块也被使用（$1K*4KB = 4MB$）、*1*个二级间接块中的第*0-2*个一级间接索引块全部被使用（$3*4MB = 12MB$）以及第*3*个一级间接索引块中的第*0-1013*个索引被使用（$1014*4KB = 4056KB$）、三级间接块没有被使用，总共$40KB + 4MB + 12MB + 4056KB = 16MB + 4096KB = 16MB + 4MB = 20MB$。

注意间接索引块编号从*0*开始！