



虚存设计关键问题

中国科学院大学计算机学院
2025-12-03





虚存设计

- 设计目标
 - 隔离
 - 隔离不同进程间的地址空间
 - 隔离用户进程和内核间的地址空间
 - 虚拟化物理内存
 - 方便用户编程，内存空间的使用转变为对0-max size的逻辑地址的访问
 - 可以使用外存（例如磁盘）来扩展物理内存的容量



虚存设计

- 虚存设计时要考虑的问题
 - 虚实地址映射
 - 映射机制：基址+长度、分段、分页
 - 映射加速：MMU、TLB、TLB查询流程（命中、缺失流程）
 - 映射开销：**页大小**、页表大小
 - 按需加载和页替换
 - 缺页处理
 - 页替换：替换算法、**颠簸**、pin/unpin
 - **后备存储 (swap)**
 - 虚存使用
 - 清零
 - 写时复制
 - 共享内存



内容提要

- 虚实映射设计问题
 - 页大小
- 页替换设计问题
 - 颠簸与工作集
 - 后备存储
 - pin/unpin
- 页使用设计问题
 - 清零
 - 写时复制 (Copy-on-write)
 - 进程间共享内存
- UNIX和Linux的虚存机制



大小页设置

- 4KB和4MB页时的目录表项 (x86架构)

Page-Directory Entry (4-KByte Page Table)

31	12 11	9 8	7 6	5 4	3 2	1 0	
Page-Table Base Address	Avail	G S	P 0	A	P C W D	P U R T S W	P

Available for system programmer's use
Global page (Ignored)
Page size (0 indicates 4 KBytes)
Reserved (set to 0)
Accessed
Cache disabled
Write-through
User/Supervisor
Read/Write
Present

Page-Directory Entry (4-MByte Page)

31	22 21	13 12 11	9 8 7 6 5 4 3 2 1 0						
Page Base Address	Reserved	P A T	Avail.	G S	D	A	P C W D	P U R T S W	P

Page Table Attribute Index
Available for system programmer's use
Global page
Page size (1 indicates 4 MBytes)
Dirty
Accessed
Cache disabled
Write-through
User/Supervisor
Read/Write
Present



页大小

- 大页
 - 典型大小：2MB、1GB
 - 好处：页表小 & 磁盘I/O高效
 - 不足：
 - 进程申请不同大小内存空间时，会造成页内碎片
 - 大页可能用不满
- 小页
 - 典型大小：4KB
 - 好处：减少内部碎片
 - 不足
 - 页表大，占用更多的内存
 - 虚实地址转换查询路径长
 - 按4KB加载加载磁盘块、执行页替换，相比大页而言磁盘I/O不高效



内容提要

- 虚实映射设计问题
 - 页大小
- 页替换设计问题
 - 颠簸与工作集
 - 后备存储
 - pin/unpin
- 页使用设计问题
 - 清零
 - 写时复制 (Copy-on-write)
 - 进程间共享内存
- UNIX和Linux的虚存机制



颠簸

- 基本概念
 - 频繁发生缺页，运行速度很慢
 - 进程被阻塞，等待内核将页从磁盘读进内存
 - 从磁盘读入一页~10ms，执行一条指令~几ns
- 虚存访问时间
 - 请求平均访问时间 = $h \times$ 内存访问时间 + $(1-h) \times$ 磁盘访问时间
 - h ：应用访存请求中，直接在内存中访问的比例
 - 例：假设内存访问时间=100ns，磁盘访问时间=10ms, $h=90\%$
 - 请求平均访问时间约是 1 ms！



颠簸

- 颠簸原因
 - 进程的工作集 > 可用的物理内存
 - 进程过多，即使单个进程的工作集都小于可用内存
 - 内存没有被很好地回收利用，导致可用内存变少
- 如何避免颠簸
 - 多进程时，考虑哪些工作集放进内存
 - 多进程时，如何分配页



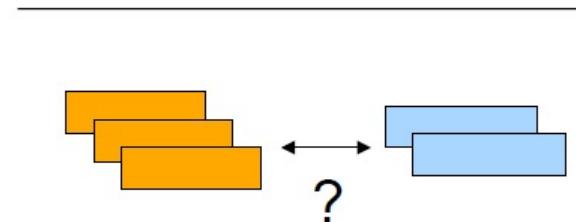
哪些工作集放进内存

- 进程分为两组
 - 活跃组：工作集加载进内存
 - 不活跃组：工作集不加载进内存
- 如何确定哪些进程是不活跃的
 - 等待事件（敲键盘、点击鼠标、...）
 - 等待资源
 - 典型方法：设定一个等待时间阈值，等待时间大于阈值的进程为不活跃进程
- 两种调度器
 - 长期调度器（准入调度器）决定，通常用于批处理系统
 - 哪些进程可以同时运行（CPU vs I/O，工作集之和，...）
 - 哪些是不活跃的进程，把它们的内存页换出到磁盘
 - 哪些是活跃进程，把它们的内存页换入到内存
 - 短期调度器（CPU调度器，执行CPU调度算法）决定
 - 把CPU分配给哪个活跃进程



在多大范围内分配页

- 多个进程同时运行时，替换哪个进程的页框？

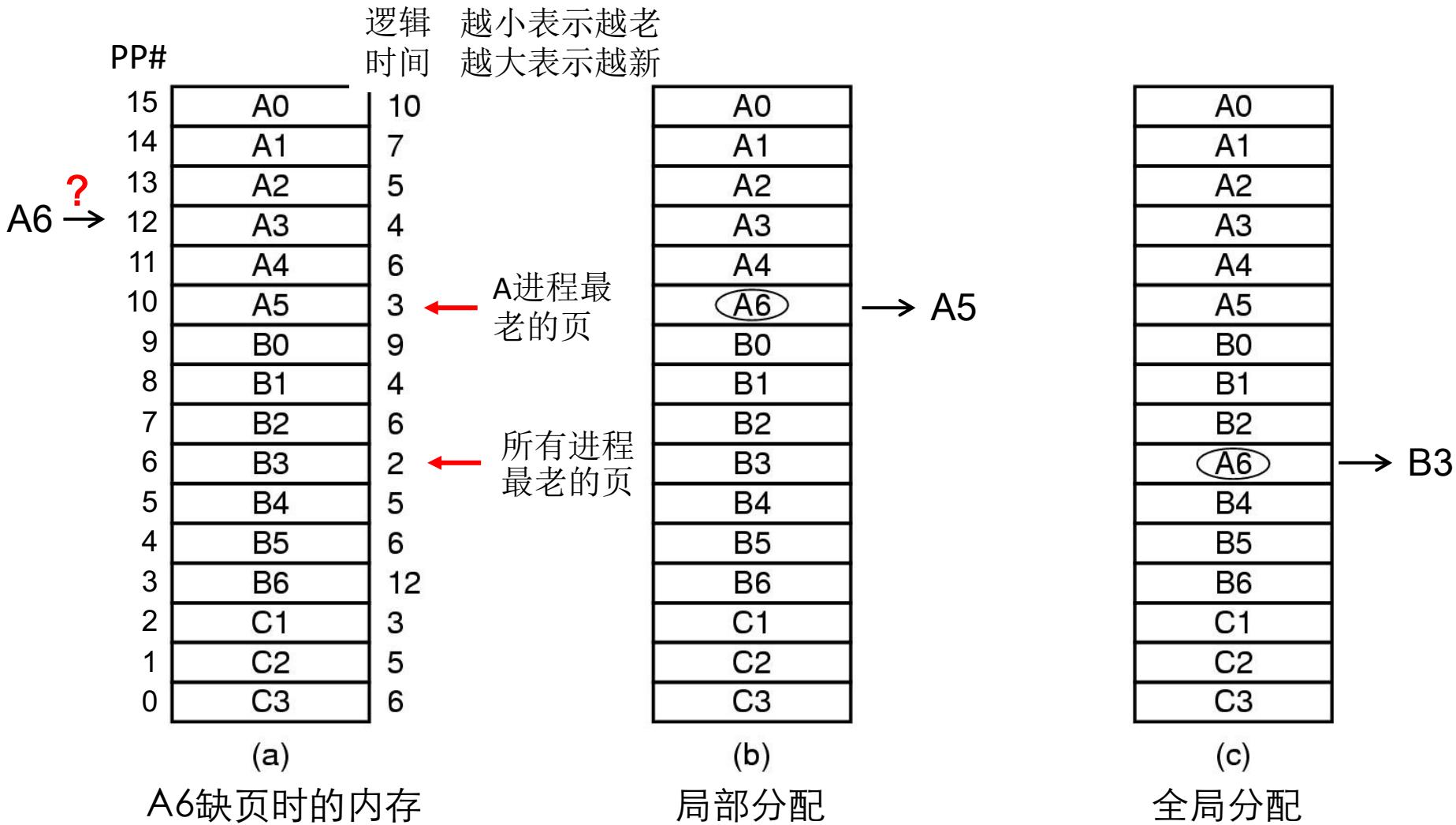


- 全局分配
 - 从所有进程的所有页框中选择
- 局部分配
 - 只从本进程（发生缺页的进程）的页框中选择



页分配

- 例子：局部分配与全局分配





全局分配 vs. 局部分配

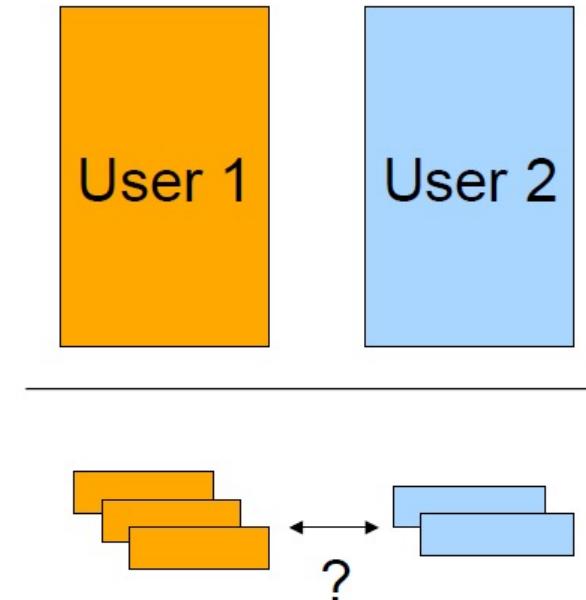
- 全局分配
 - 从所有进程的所有页框中选择
 - 可替换其它进程的页框
 - 每个进程运行期间，其内存大小是动态变化的
 - 好处：全局资源调配
 - 坏处：可能干扰其他进程，或受其它进程的页替换干扰
 - 不能控制各个进程的内存使用量
- 局部分配
 - 只从本进程（发生缺页的进程）自己的页框中选择
 - 一个进程运行期间，固定其内存大小不变
 - 页框池：分配给进程的页框的集合，不同进程的页框池大小可不同
 - 好处：隔离，不影响其它进程
 - 坏处：不灵活
 - 进程使用内存频繁时会出现颠簸
 - 难以充分利用内存：每个进程对内存的需求不一样



平衡分配

- 局部分配 + 池大小动态调整
 - 每个进程有自己的页框池 (pool)
 - 从自己的池中分配页，且从自己的工作集中替换页
 - 用一种机制在运行时动态调整每个池的大小
 - 初始大小 + 动态调整

} 局部
—————> 全局





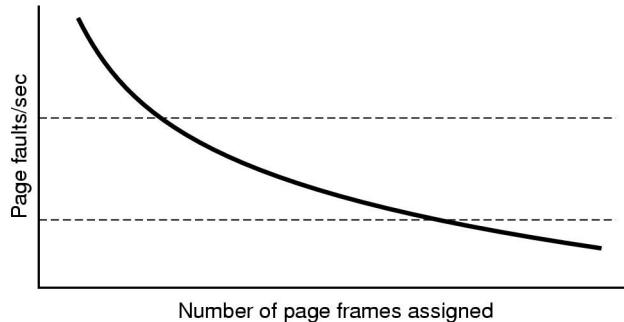
平衡分配

- 初始池大小设置方法
 - 固定大小：所有进程都一样
 - 平均分配
 - 根据进程大小按比例分配
 - 工作集
- 不同的页加载方式
 - 纯粹的按需加载页：加载慢，影响程序执行
 - 预加载：先加载部分页 → 初始池大小



平衡分配

- 动态调整池大小：进程使用内存的大小变化
 - PFF算法 (page fault frequency)
 - 缺页率PFR：进程每秒产生多少次缺页
 - 对于大多数替换策略，PFR随分配给进程的内存增加而减少



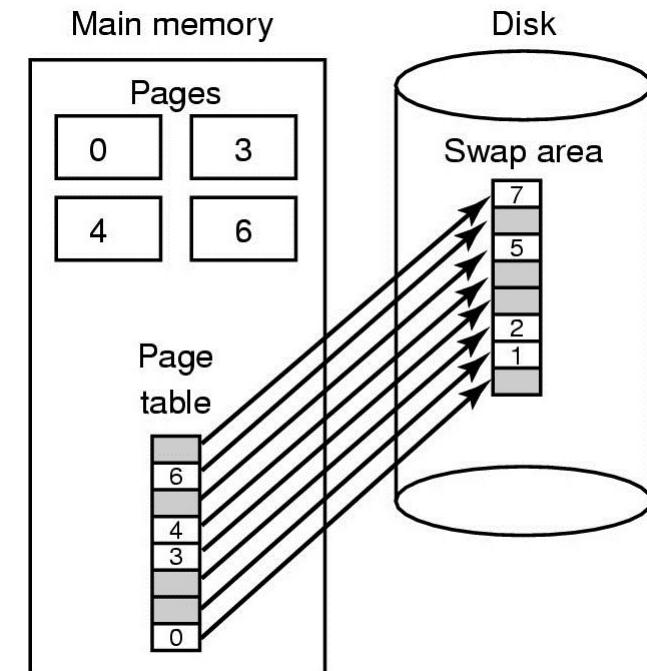
两个阈值：A和B
A为上限，B为下限

- 根据进程的PFR来调整分配给它的内存量 (pool size)
 - 当PFR高于A，就增加其内存 (pool增大)
 - 当PFR低于B，就缩小其内存 (pool减小)
- PFR计算
 - 方法1：只看当前1秒钟内的缺页次数R_i
 - 方法2：求滑动均值
 - 例： $PFR_i = (R_i + PFR_{i-1})/2$



后备存储

- 交换区 (swap area)
 - 在磁盘上
 - 专门用于存储进程换出页
 - 交换分区
 - 用专门的磁盘分区保存换出数据
 - 安装操作系统时预先划分,大小固定,不易调整
 - 连续的磁盘空间,交换性能相对较好
 - 交换文件
 - 用文件保存换出数据
 - 管理灵活,大小可调整
 - 通过文件系统访问,交换性能相对低些

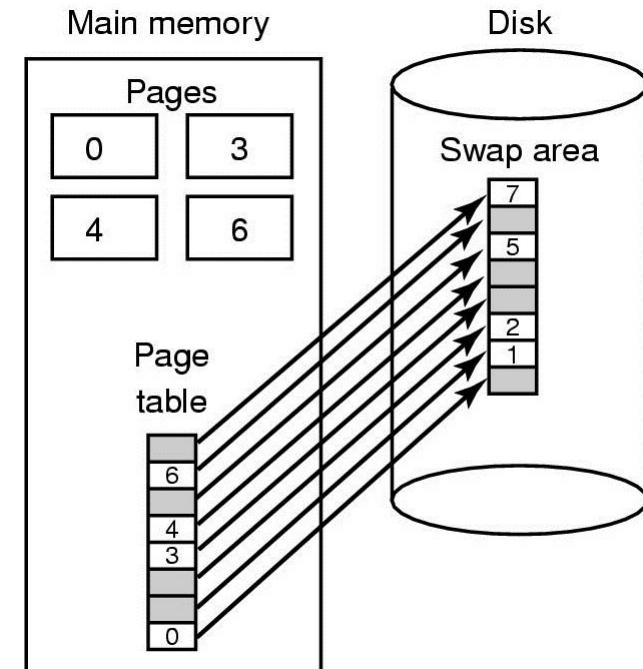




交换空间管理

- 静态分配

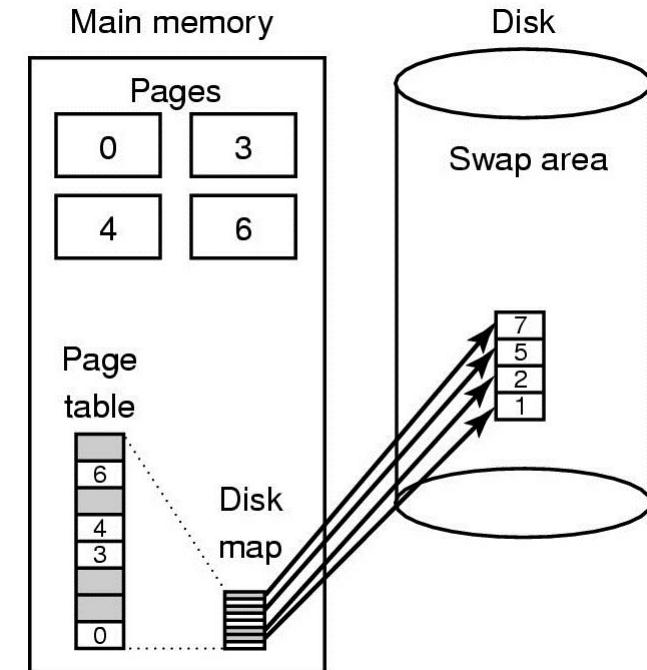
- 创建进程时分配，进程结束时回收
- 大小：进程镜像大小 (process image)
- PCB记录交换空间的磁盘地址
- 绑定：一个虚存页对应一个磁盘块
 - 磁盘块称为shadow page
- 页换出：进程内存页拷贝到交换区
- 页换入：进程镜像加载进内存
- 缺点
 - 难以增长：栈、堆





交换空间管理

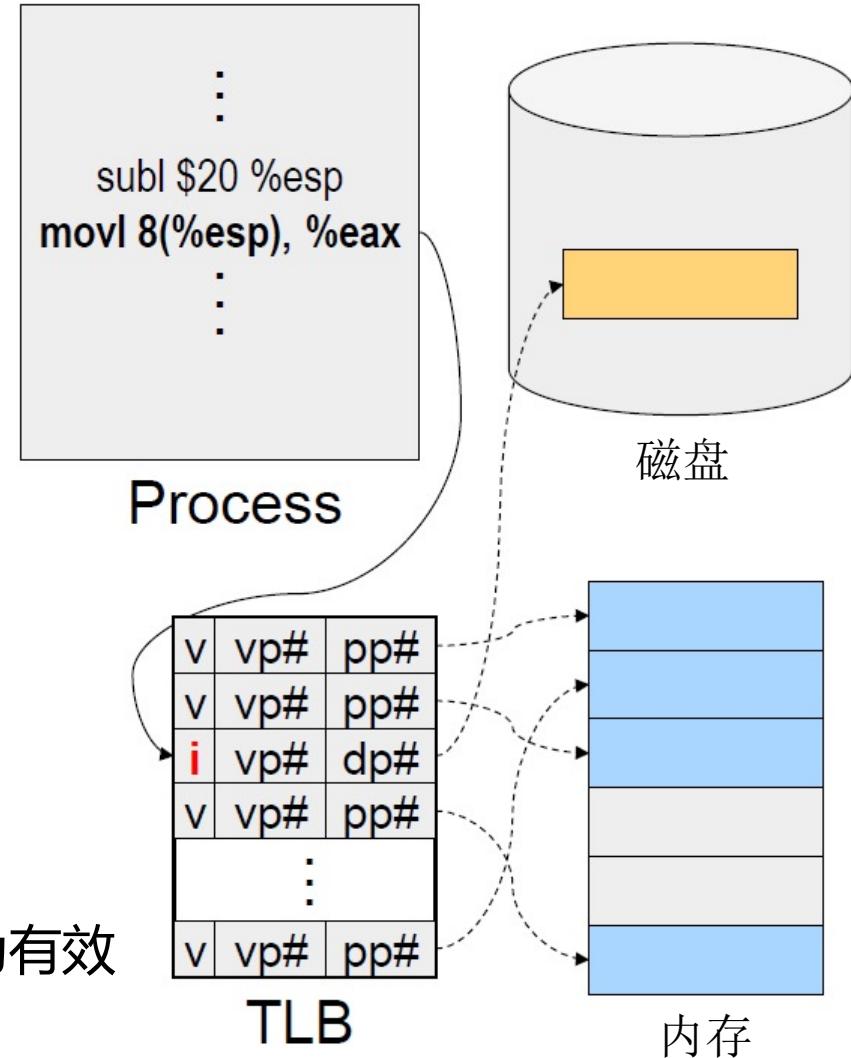
- 动态分配
 - 创建进程时不分配
 - 页换出时分配，页换入时回收
 - 虚页与磁盘页不绑定
 - 多次换出时，分配不同的磁盘页
 - PTE中记录页的磁盘地址
- 一个优化：程序text段
 - 直接用磁盘上的可执行文件作为交换区，减少交换区大小
 - 换出时直接丢弃（只读的），减少不必要的写回





页地址转换

- PTE : 虚页 → 页框和磁盘
 - 如果valid bit=1 , 物理页号pp#
 - 如果valid bit=0 , 磁盘块号dp#
- 换出
 - 将PTE和TLB置为无效
 - 将页拷贝到磁盘
 - 将磁盘块号填入PTE
- 换入
 - 找一个空闲页框 (可能触发页替换)
 - 将磁盘块拷贝到这个页框中
 - 将页框号填入PTE中，并将PTE置为有效





钉住页 (pin/unpin)

- 为什么需要pin ?
 - 有些页会频繁访问，换出后又会被换入，影响系统性能
 - DMA缓冲区对应的内存页不能被换出
- 系统调用接口
 - pin : 把虚页钉在内存，使它们不会被换出
 - unpin : 取消pin，使它们可以被换出
- 如何设计 ?
 - 用一个数据结构来记录所有钉住的页
 - 换页算法在替换页时检查该数据结构
 - 如果页被钉住，则不替换它，重新再选择一页
 - 如果整个内核全部在物理内存中，还需要进行pin和unpin操作么？



内容提要

- 虚实映射设计问题
 - 页大小
- 页替换设计问题
 - 颠簸与工作集
 - 后备存储
 - pin/unpin
- 页使用设计问题
 - 清零
 - 写时复制 (Copy-on-write)
 - 进程间共享内存
- UNIX和Linux的虚存机制



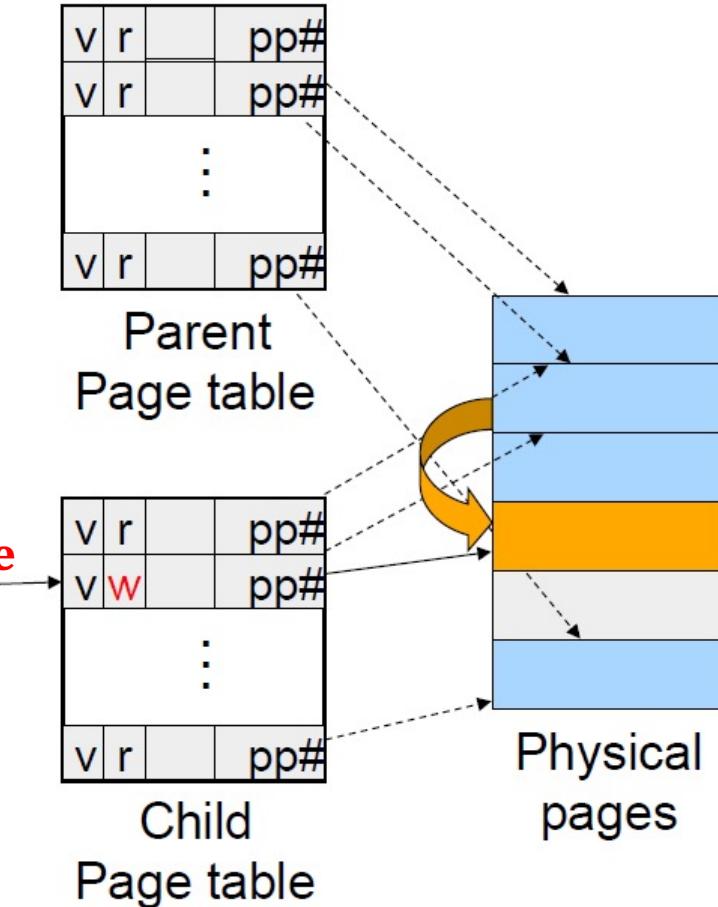
清零页

- 将页清零
 - 把页置成全0
 - 堆和栈的数据都要初始化
- 如何实现
 - 对于堆或栈中的页，当它们第一次发生page fault时，将它们清零
 - 有一个专门的线程来做清零
- 不做页清零可能引起安全问题
 - 敏感数据泄露，例如密码、内核数据结构、内核函数地址等



写时复制 (Copy-On-Write)

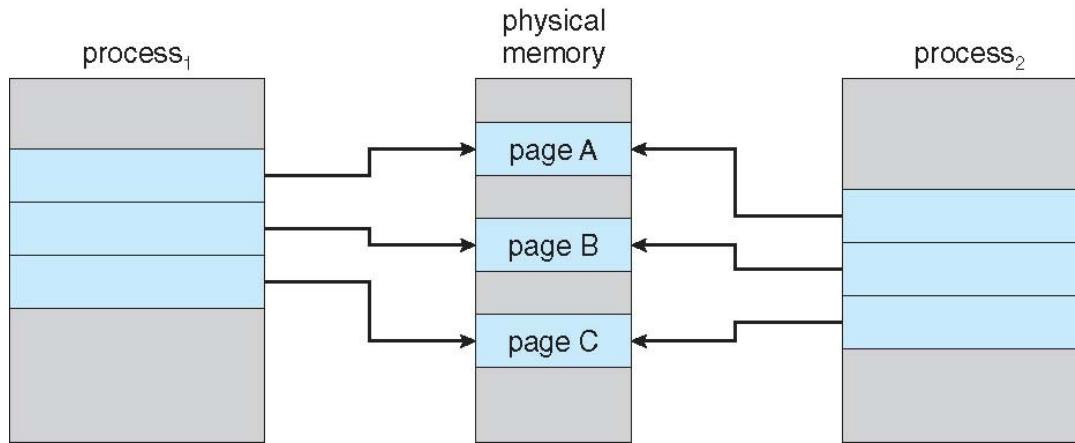
- 该技术用于创建子进程 (fork 系统调用)
- 原理
 - 子进程的地址空间使用其父进程相同的映射
 - 将所有的页置成 read-only
 - 将子进程置成就绪
 - 对于读，没有问题
 - 对于写，产生 page fault
 - 修改PTE，映射到一个新的物理页
 - 将页内容全部拷贝到新物理页
 - 重运行发生缺页的指令



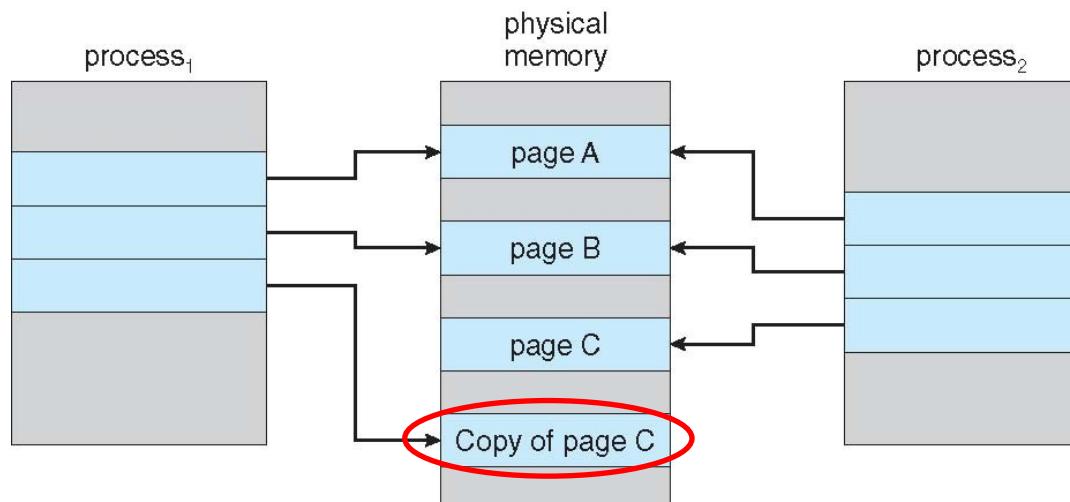


写时复制 (Copy-On-Write)

- 示例



进程P1 fork一个子进程P2

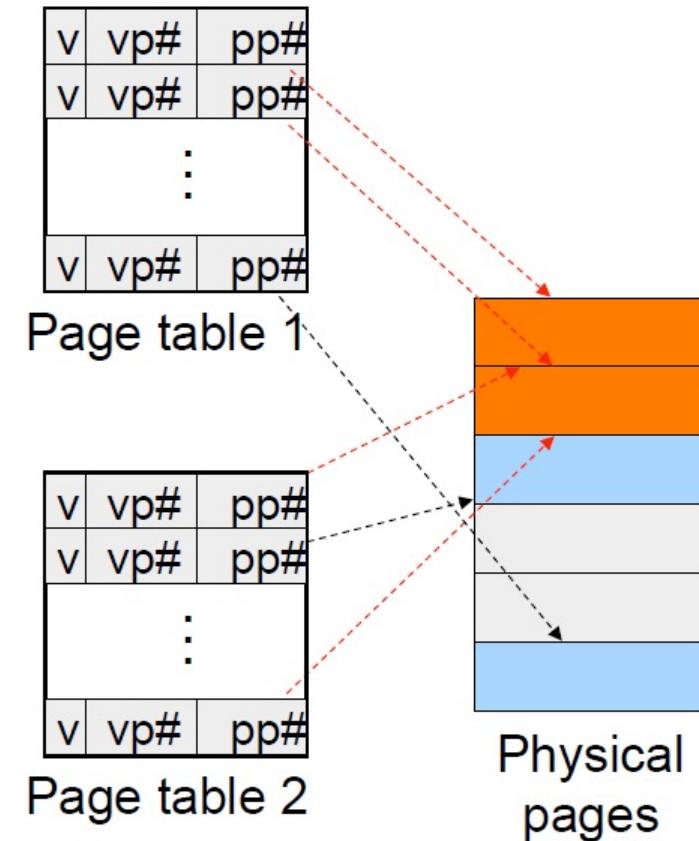


进程P1修改page C



进程间共享内存

- 两个进程的页表共享一些物理页
 - 进程间通信
 - 进程间数据共享
- API
 - 共享内存的系统调用
 - `shm_get`, `shm_ctl`, ...
- 涉及问题
 - 共享页的换入换出
 - 影响多个进程
 - pin和unpin共享页
 - 有共享页的进程的工作集
 - 共享页优先进入工作集
 - 有共享页的进程结束





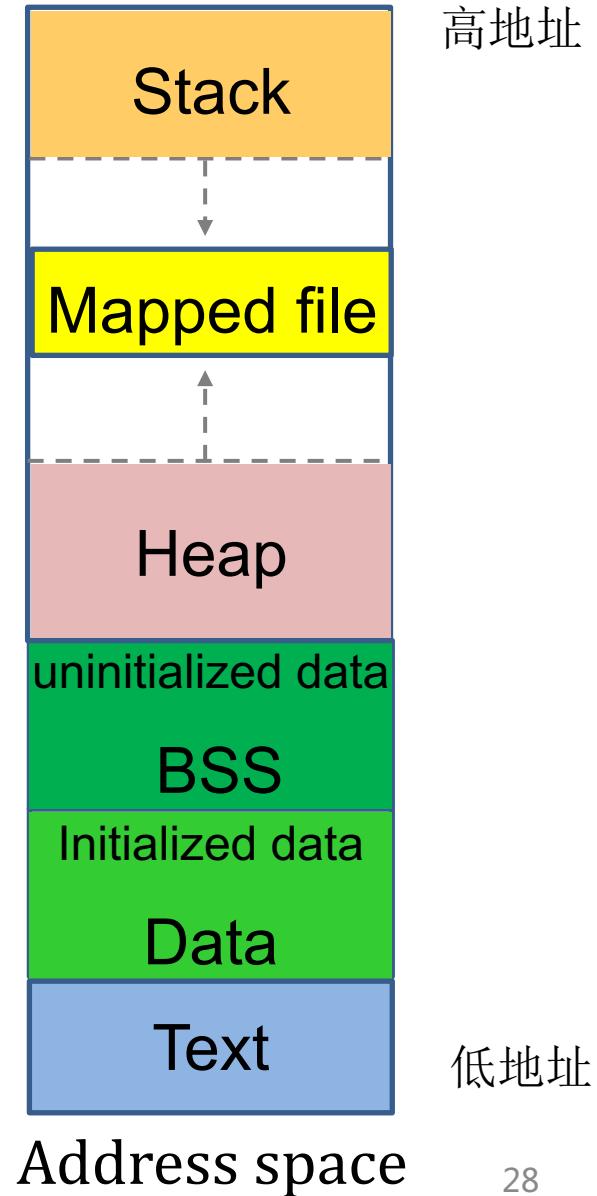
内容提要

- 页替换设计问题
 - 颠簸与工作集
 - 后备存储
 - pin/unpin
- 虚实映射设计问题
 - 页大小
- 页使用设计问题
 - 清零
 - 写时复制 (Copy-on-write)
 - 进程间共享内存
 - 分布式共享内存
- UNIX和Linux的虚存机制



UNIX的地址空间

- Text段：只读 & 大小不变
- 数据段
 - 初始化数据
 - 未初始化数据：BSS (Block Started by Symbol)
 - brk区用于增长或缩小 (Heap)
 - 程序断裂点 (breakpoint) 最初指向数据段的末尾，向上增长用于分配堆内存
- 栈
- 内存映射文件 (memory mapped file)
 - 将一个文件映射进虚存
 - mmap(addr, len, prot, flags, fd, offset)
 - unmap(addr, len)
 - 像访问内存一样访问文件





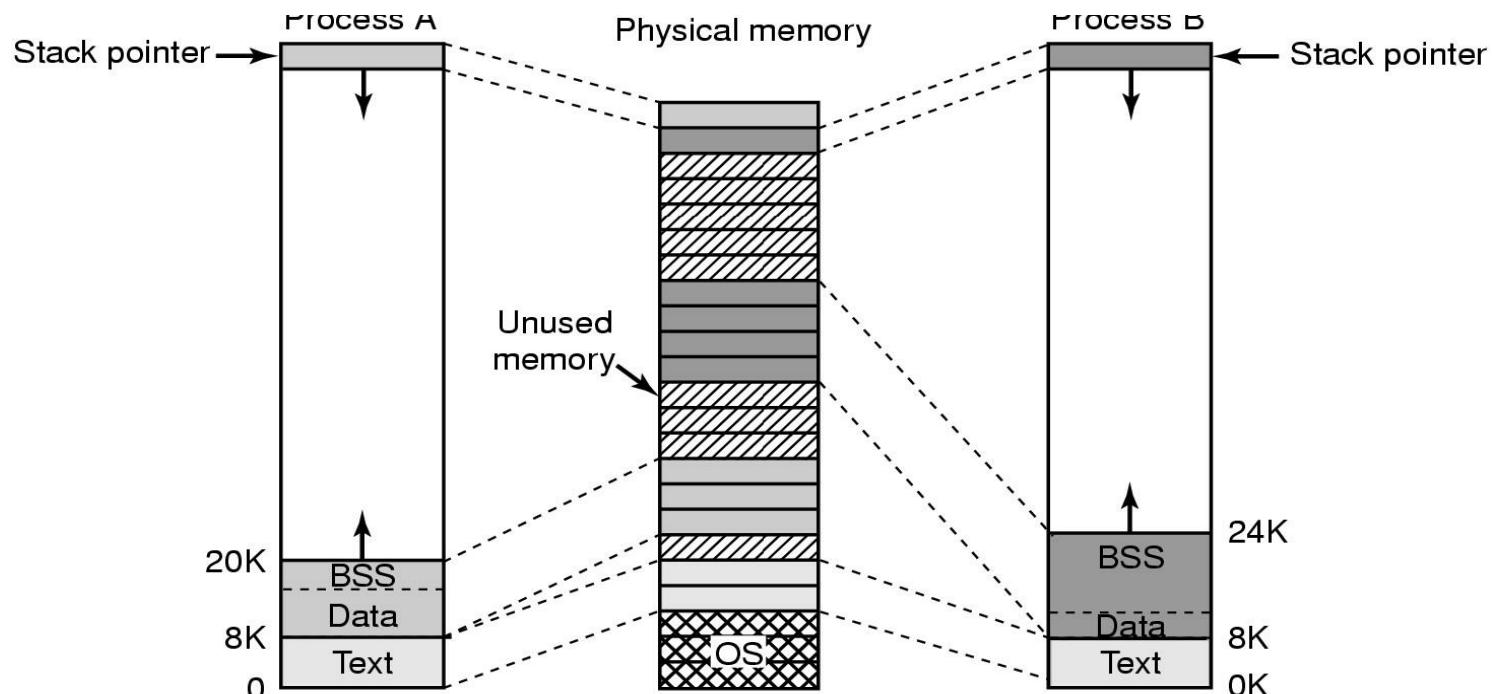
BSD4的虚存

- 物理内存划分
 - Core map (pin) : 所有页框的描述信息
 - 内核 (pin)
 - 其它页框 : 给用户进程
- 页替换
 - 运行换页守护进程 (page daemon) , 直到有足够的空闲页
 - 早期BSD使用原始的Clock替换算法 (有第二次机会的FIFO)
 - 后来的BSD使用双表针的Clock算法
 - 如果page daemon不能得到足够的空闲页 , 则运行swapper
 - 寻找idle时间超过20秒及以上的进程
 - 最大的4个进程



Linux的地址空间

- 32-位机的地址空间
 - 3 GB用户空间
 - 1GB内核空间（用户态不可见）
- 栈段从3GB位置向下增长
 - 初始时保存进程的环境变量和命令行参数
- 数据段：
 - 大小可变：brk vs. malloc
 - BSS：未初始化的全局变量
 - 页加载时初始化为0
 - zero page: 避免分配大量全0的物理页





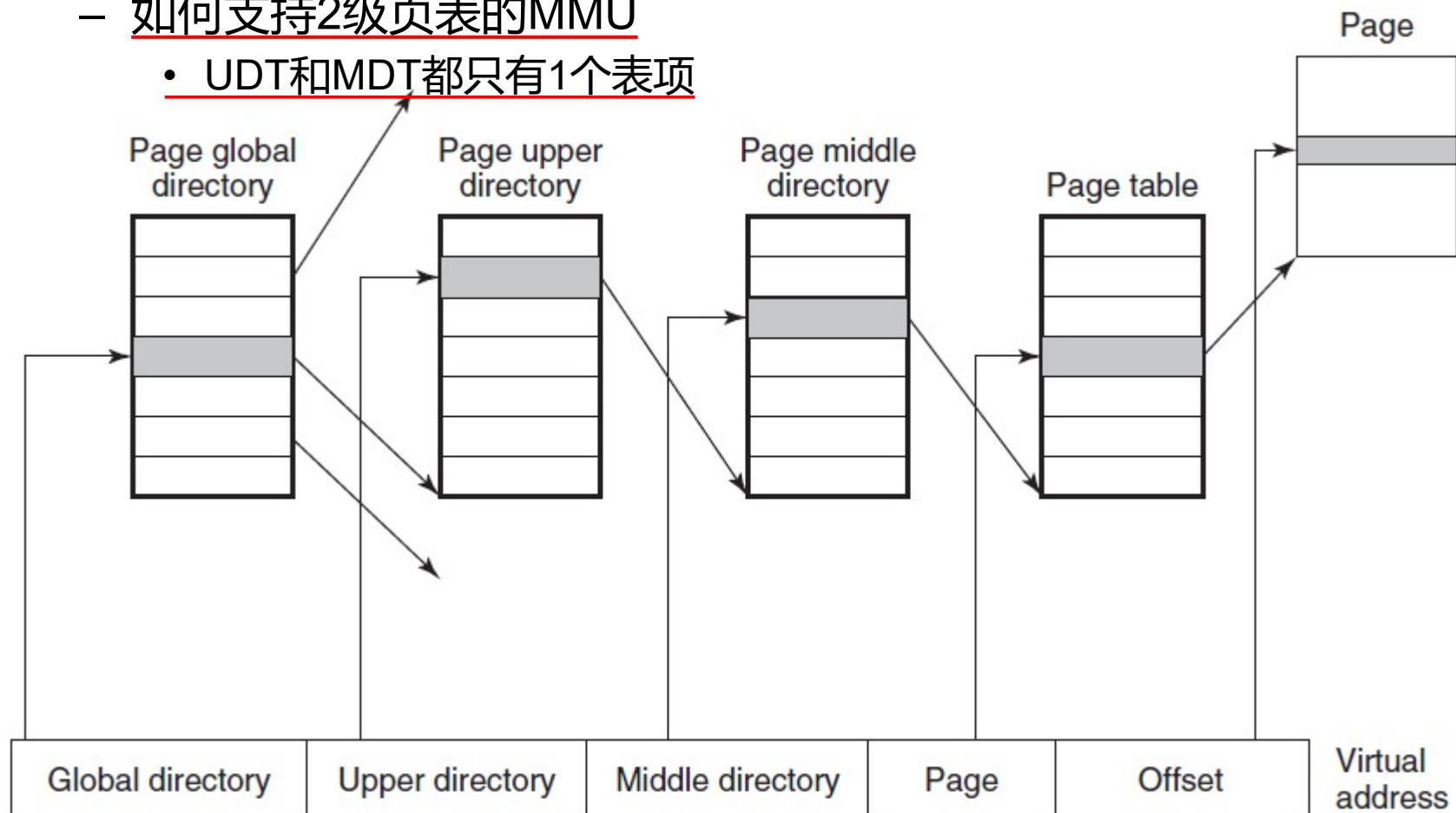
系统调用

System call	Description
s = brk(addr)	Change data segment size
a = mmap(addr, len, prot, flags, fd, offset)	Map a file in
s = unmap(addr, len)	Unmap a file



Linux的地址转换

- 2.6.11及以后的Linux使用4级页表
 - 虚地址划分为5段：GDT、UDT、MDT、PT和页内偏移
 - 如何支持2级页表的MMU
 - UDT和MDT都只有1个表项

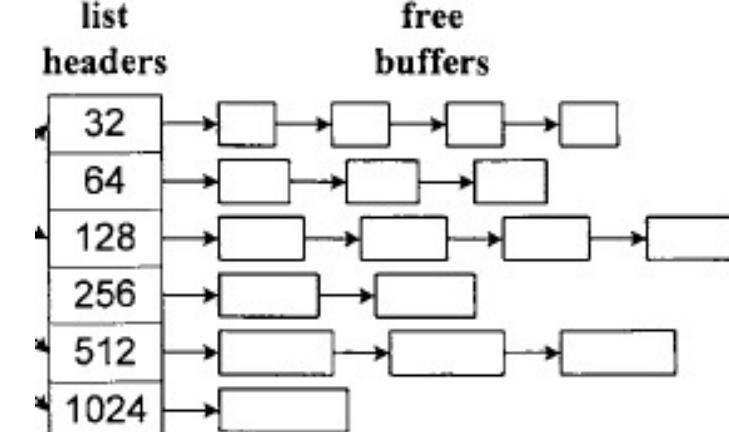




Linux的物理页分配

- 伙伴算法

- 用于分配和释放连续物理空间，减少碎片
- 初始时：一块连续的物理内存空间
- **分配空间**， m 页：找 $\text{size} \geq m$ 且 size 最小的段，假设其 size 为 n
- 如果 $n \geq 2m$ ，则将该段对分为两个 $\text{size} = n/2$ 的段，
- 重复对分，直到 $n/2 < m \leq n$
- **释放空间**， m 页：若释放段的“伙伴”也是空闲的，则合并它们，重
复合并



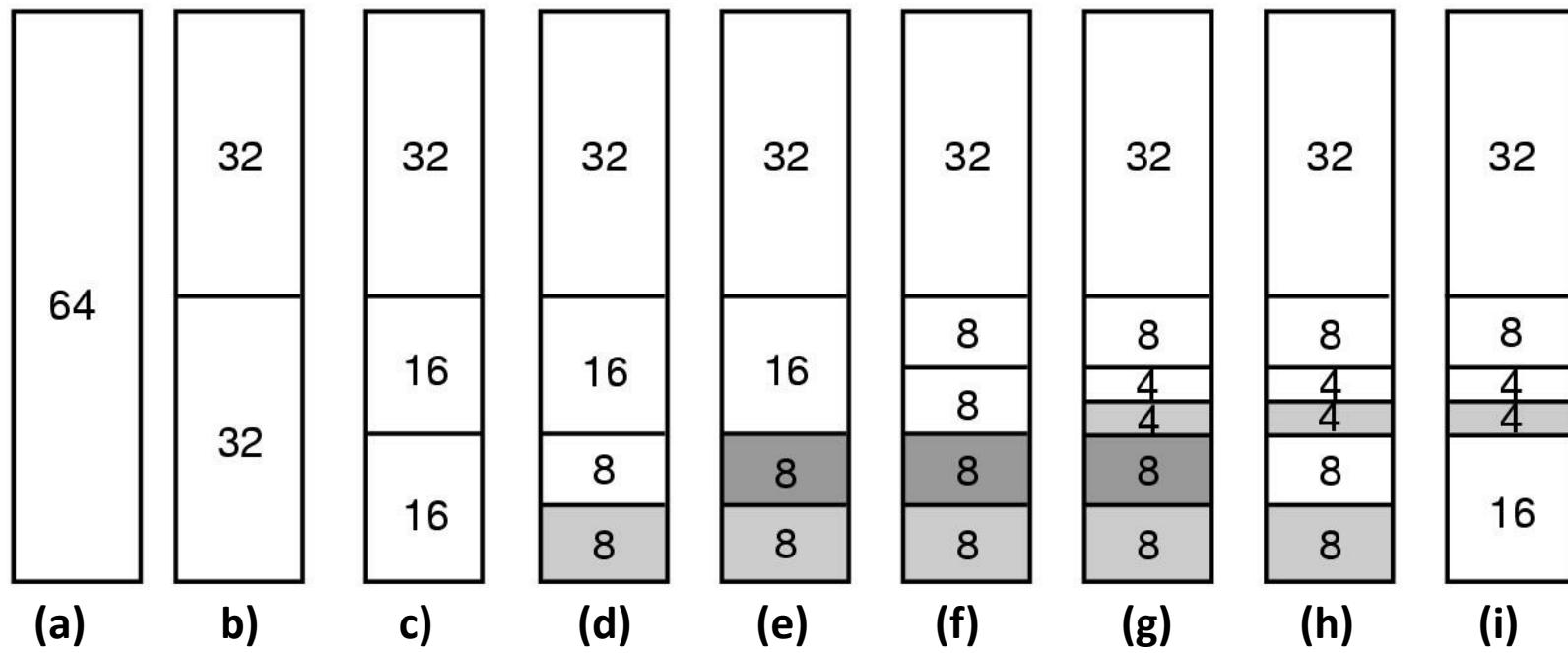
- 数据结构

- 空闲链表数组：每条链上的页大小为
2的幂



Linux的物理页分配

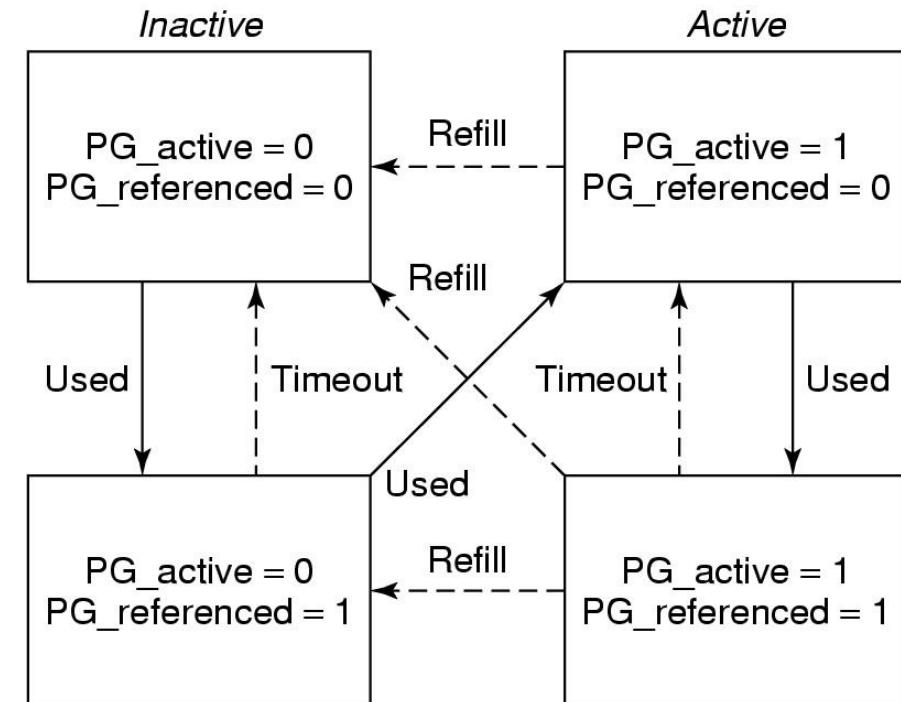
- 大块空间分配采用伙伴算法
 - 例子：请求1，分配8页，(d)完成，3次对分
 - 请求2，分配8页，(e)完成
 - 请求3，分配4页，(g)完成，1次对分
 - 请求4，释放8页，(h)完成
 - 请求5，释放8页，(i)完成，1次合并





Linux的页替换

- 方法
 - 保持一定数量的空闲页
 - 文件缓存 (page cache) 使用 Clock算法
 - 未使用的共享页使用Clock算法
 - 用户进程的内存使用改进的 Clock算法
- 改进的Clock算法
 - 两条链
 - Active list : 所有进程的工作集
 - Inactive list : 回收的候选页
 - Refill是将页从active list移动到 inactive list，在内存资源紧张时用于回收内存页面





总结：内存管理

- 虚存：进程地址空间是独立的、连续的
 - 隔离 & 保护 & 方便程序员
- 地址转换：虚地址→物理地址，保护&错误隔离
 - 每个内存访问都要进行，由硬件MMU完成
 - TLB：专门硬件加速地址转换(查页表)
- 地址映射
 - 分段、分页
 - 页表结构：用于地址转换
 - 页表项 (PTE)：PP# & 控制位 V、R、M、R/W、C
 - 大地址空间：分段+分页、多级页表、反向页表
- 访存操作在正常情况下全部是硬件完成
 - 只有在缺页、无权限等异常情况下，才需要OS参与



总结：内存管理

- 按需加载页
 - 一个进程并不是每时每刻都需要所有的页
 - 频繁使用的页放内存，其它的页放磁盘
 - 要访问的页不在内存，透明地将它加载进内存 → OS缺页处理
 - 加载页时内存可能没有空闲页框 → 进行页替换
- 缺页处理
 - 分配空闲页框 或 替换一个页框
 - 写回页 & 加载页 (磁盘I/O)
 - 刷新TLB、修改页表项、加载TLB
- 页替换算法
 - FIFO & Second chance
 - Clock
 - LRU
 - 工作集