

- 死锁相关概念
- ▼ 死锁产生条件
 - 资源分配图
 - 基本事实
- ▼ 处理死锁的方法
 - 处理死锁的方法1：预防死锁
 - 处理死锁方法2：避免死锁
 - ▼ 死锁避免的算法
 - 银行家算法
 - 处理死锁方法3:死锁的检测与解除

死锁相关概念

- 定义：如果在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，则称一组进程或系统此时发生了死锁。
 - 原因：死锁产生的原因是与资源的类型、资源的数量和相应的使用相关
1. 竞争资源：多个进程竞争资源，而资源又不能同时满足其需求。
 2. 进程推进顺序不当：进程申请资源和释放资源的顺序不当。
- 资源分类

■ 可剥夺资源

- 指某进程获得这类资源后，该资源可以被其他进程或系统剥夺。如CPU，主存储器。

■ 非剥夺资源，又称不可剥夺资源

- 指系统将这类资源分配给进程后，再不能强行收回，只能在进程使用完后主动释放。如打印机、读卡机。

注意：竞争可剥夺资源不会产生死锁！

■ 永久性资源

- 可顺序重复使用的资源，如打印机。

■ 消耗性资源

- 由一个进程产生，被另一个进程使用短暂时间后便无用的资源，又称为临时性资源，如消息。

竞争永久性资源和临时性资源都可能产生死锁。

死锁产生条件

- 4个必要条件



死锁产生的4个必要条件(Coffman 1971)



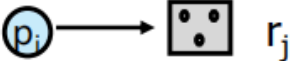
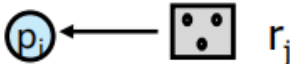
- ① 互斥条件：在一段时间内某资源仅为一个进程所占有。
- ② 请求和保持条件（占有并等待）：又称部分分配条件。当进程因请求资源被阻塞时，已分配资源保持不放。
- ③ 不剥夺条件（非抢占）：进程所获得的资源在未使用完毕之前，不能被其他进程强行夺走。
- ④ 循环等待条件：死锁发生时，存在一个进程资源的循环。

- 注意



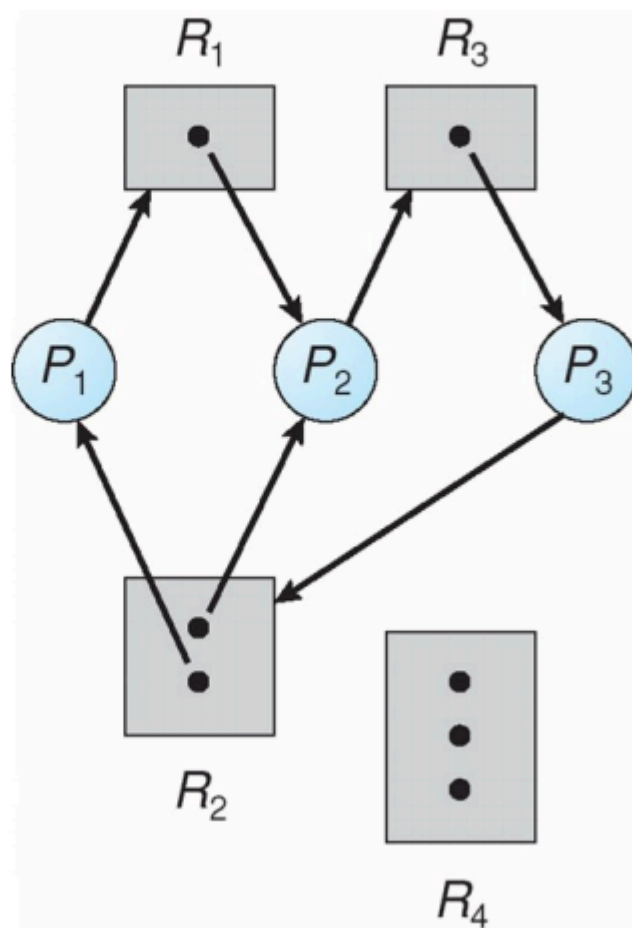
- 死锁是因资源竞争造成的僵局
- 通常死锁至少涉及两个进程
- 死锁与部分进程及资源相关

资源分配图

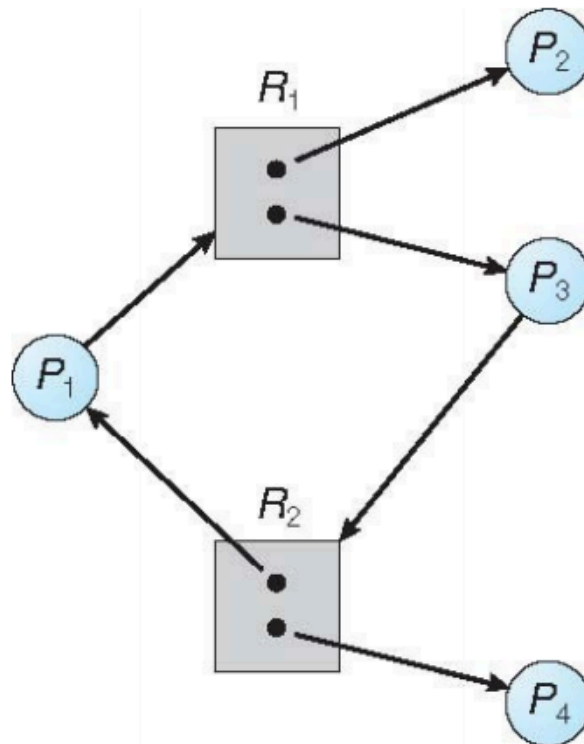
- 系统死锁可利用资源分配图描述。
 - 资源分配图又称“进程——资源”图，由一组结点N和一组边E所构成：
 - N被分成两个互斥的子集：进程结点子集 $P = \{p_1, p_2, \dots, p_n\}$ ，资源结点子集 $R = \{r_1, r_2, \dots, r_m\}$ 。
 - 用圆圈代表一个进程 
 - 用方框代表一类资源，方框中的一个点代表一类资源中的一个资源 
 - E是边集，它连接着P中的一个结点和R中的一个结点
 - $e = \langle p_i, r_j \rangle$ 是资源请求边 
 - $e = \langle r_j, p_i \rangle$ 是资源分配边 

• 一些例子

存在死锁的资源图例



具有环且未死锁的资源分配图



基本事实

- 如果分配图没有环 \Rightarrow **NO deadlock!**
- 如果分配图包含环 \Rightarrow
 - 如果每个资源类型，只包含一个资源实例，则**死锁**
 - 如果每个资源有多个资源实例，则只是存在**死锁的可能**，不一定会死锁

处理死锁的方法

- 用于处理死锁的方法主要有：
 - ① 忽略死锁。这种处理方式又称鸵鸟算法，指像鸵鸟一样对死锁视而不见。被大多数OS采用，因为死锁出现概率低，忽略死锁代价小。
 - ② 预防死锁：设置某些限制条件，通过破坏死锁产生的四个必要条件之一来预防死锁。
 - ③ 避免死锁：在资源的动态分配过程中，用某种方法来防止系统进入不安全状态。
 - ④ 检测死锁及解除：系统定期检测是否出现死锁，若出现则解除死锁。

处理死锁的方法1：预防死锁

- 预防死锁
 - 通过破坏产生死锁的四个必要条件中的一个或几个条件，来防止发生死锁。
 - 考虑破坏必要条件的可能：
 - 互斥条件
 - 请求和保持条件
 - 不可剥夺条件
 - 循环等待条件

(1) 互斥条件

- 破坏条件1：互斥条件
 - 使资源可同时访问，而非互斥使用
 - 如可重入程序、只读数据、时钟等
 - 但互斥对一些资源是固有的属性
 - 如可写文件、互斥锁，此条件往往不能破坏



(2) 请求和保持条件

- 破坏条件2：请求和保持条件
 - 思路：当每个进程申请一个资源时（可能成功或失败），它不能占有其他资源。
 - 方法一：要求进程一次申请它所需的全部资源，若有足够的资源则分配给进程，否则不分配资源，进程等待。这种方法称为**静态资源分配法**。
 - 方法二：允许进程仅在没有资源时才可申请资源。一个进程申请资源并使用，但是在申请更多资源时，**应释放已经分配的所有资源**。
 - 特点：
 - 简单且易于实现；
 - 但资源利用率低，进程延迟运行，可能发生饥饿。

(3) 不可剥夺条件

■ 破坏条件3：不可剥夺条件

- 对一个已获得某些资源的进程，若新的资源请求得不到满足，则它已占有的资源都可以被抢占。即这些资源都被**隐式释放**了。
 - 例：进程A已经占有了资源a，并计划申请资源b，此时进程B也处于等待其他资源c的状态，如果：
 - ① 资源b可用，则分配给进程A
 - ② 资源b不可用，则检查资源b是否已经分配给进程B，如果：
 - ① 资源b被进程B所占有，则抢占资源b
 - ② 资源b不被任何一个处于等待资源的进程占有，则资源a也可被其他进程抢占。
- 负面：这种释放有可能造成已有工作的失效，重新申请和释放会带来新的系统开销
- 适用范围：常用于状态易于保存和恢复的资源，如CPU寄存器和内存资源，对于打印机、互斥信号量等不可使用

(4) 循环等待条件

■ 破坏条件4：循环等待条件

■ 层次分配策略

- 资源被分成多个层次
- 当进程得到某一层的一个资源后，它只能再申请较高层次的资源
- 当进程要释放某层的一个资源时，必须先释放占有的较高层次的资源
- 当进程得到某一层的一个资源后，它想申请该层的另一个资源时，必须先释放该层中的已占资源
- 也称为有序资源分配法

■ 层次策略的变种：按序分配策略

■ 把系统的所有资源排一个顺序

- 如系统若共有 n 个进程,共有 m 个资源，用 r_i 表示第 i 个资源，于是这 m 个资源是：

$$r_1, r_2, \dots, r_m$$

■ 规定：

- 进程不得在占用资源 $r_i (1 \leq i \leq m)$ 后再申请 $r_j (j < i)$
- 即，只能申请编号之后的资源，而不许申请编号之前的资源，从而避免资源申请的环路问题。

- 不难证明，按这种策略分配资源时系统不会发生死锁。

处理死锁方法2：避免死锁

■ 死锁的避免

- 允许系统中存在前3个必要条件，通过合适的资源分配算法，确保不会出现第四个必要条件，从而避免死锁。
- 不是对进程随意强加规则，而是在**资源的动态分配**过程中实施
- 用某种方法**防止系统进入不安全状态**，从而避免死锁的发生
- 决策依据：已分配资源情况，当前申请资源情况，以及将来资源的申请与释放情况
- 决策结果：
 - 如果一个进程当前请求的资源会导致死锁，系统就拒绝启动这个进程
 - 如果一个资源分配会导致下一步死锁，系统就拒绝本次分配

• 安全状态

(1) 安全状态

- **思路**：允许进程动态地申请资源，系统在进行资源分配之前，先计算资源分配的安全性。若此次分配不会导致系统进入不安全状态，便将资源分配给进程，否则进程等待。

■ 安全状态

- 是指系统能按某种顺序如 $\langle P_1, P_2, \dots, P_n \rangle$ 来为每个进程分配其所需的资源，直至最大需求，使每个进程都可以顺利完成，则称此时的系统状态为**安全状态**，称序列 $\langle P_1, P_2, \dots, P_n \rangle$ 为安全序列。

• 不安全状态

不安全状态

- 若某一时刻系统中**不存在一个安全序列**，则称此时的系统状态为**不安全状态**。
- 进入不安全状态后，便**可能**进入死锁状态；
 - 不是所有的不安全状态都能导致死锁，因为不安全状态有可能转变为安全状态
- 因此避免死锁的**本质**是使系统不进入不安全状态，而是**始终保持**在安全状态。

死锁避免的算法

- 资源分配图算法

银行家算法

- 对于多类实例资源，最具代表性的死锁避免算法是Dijkstra的银行家算法。
- 背景：
 - 类比于银行业务：顾客类比于进程，顾客想借钱，钱为资源，银行为OS
 - 银行可借出的钱有限，每个顾客都有一定的银行信用额度
 - 顾客可以选择借一部分，但不能保证顾客在借走大量贷款后一定能偿还，除非他能获取全部贷款要求
 - 如果银行存在风险，没有足够的资金提供更多贷款让顾客偿还，则银行家就拒绝贷款给顾客
- 核心思想
 - 检查资源分配后是否会导致系统进入不安全状态
 - 手段：模拟分配资源，然后检查是否满足安全状态

- 基本数据结构

(1) 可用资源向量Available

- 假定系统中有 n 个进程 P_1, P_2, \dots, P_n ， m 类资源 R_1, R_2, \dots, R_m ，银行家算法
- 可利用资源向量Available是一个含有 m 个元素的数组，其中每一个元素代表一类资源的空闲资源数目。
- 如果 $Available(j) = k$ ，表示系统中现有空闲的 R_j 类资源 k 个。

(2) 最大需求矩阵Max

- 最大需求矩阵Max是一个 $n \times m$ 的矩阵，定义了系统中每个进程对 m 类资源的最大需求数目。
- 如果 $Max(i, j) = k$ ，表示进程 P_i 需要 R_j 类资源的最大数目为 k 。

(3) 分配矩阵Allocation

- 分配矩阵Allocation是一个 $n \times m$ 的矩阵，定义了系统中每一类资源当前已分配给每一个进程的资源数目。
- 如果 $\text{Allocation}(i, j) = k$ ，表示进程 P_i 当前已分到 R_j 类资源的数目为 k 。
- Allocation_i 表示进程 P_i 的分配向量，由矩阵Allocation的第 i 行构成。

(4) 需求矩阵Need

- 需求矩阵Need是一个 $n \times m$ 的矩阵，它定义了系统中每一个进程还需要的各类资源数目。
- 如果 $\text{Need}(i, j) = k$ ，表示进程 P_i 还需要 R_j 类资源 k 个。 Need_i 表示进程 P_i 的需求向量，由矩阵Need的第 i 行构成。
- 三个矩阵间的关系：

$$\text{Need}(i, j) = \text{Max}(i, j) - \text{Allocation}(i, j)$$

• 资源请求算法

- 设 Request_i 是进程 P_i 的请求向量， $\text{Request}_i(j) = k$ 表示进程 P_i 请求分配 R_j 类资源 k 个。
- 当 P_i 发出资源请求后，系统按下述步骤进行检查：
 - ① 如果 $\text{Request}_i \leq \text{Need}_i$ ，则转向步骤2；否则生成出错条件，此时进程请求超出了他的需求。
 - ② 如果 $\text{Request}_i \leq \text{Available}$ ，则转向步骤3；否则进程 P_i 转入等待。
 - ③ 试分配并修改数据结构：
 - ① $\text{Available} = \text{Available} - \text{Request}_i$;
 - ② $\text{Allocation}_i = \text{Allocation}_i + \text{Request}_i$;
 - ③ $\text{Need}_i = \text{Need}_i - \text{Request}_i$;
 - ④ 系统执行安全性算法，检查此次资源分配后得到的新状态是否安全。若安全，才正式分配；否则，试分配作废，让进程 P_i 等待。

- 安全性算法

- ① 设置两个向量

- Work: 表示系统可提供给进程继续运行的各类空闲资源数目, 含有m个元素, 执行安全性算法开始时, 初始化 $Work = Available$ 。
 - Finish: 表示系统是否有足够的资源分配给进程, 使之运行完成, 开始时, $Finish(i) = false$; 当有足够资源分配给进程 P_i 时, 则令 $Finish(i) = true$ 。

- ② 从进程集合中找到一个能满足下述条件的进程i:

- $Finish(i) == false$;
 - $Need_i \leq Work$;
 - 如找到则执行步骤3; 否则执行步骤4。

- ③ 当进程 P_i 获得资源后, 可顺利执行直到完成, 并释放出分配给它的资源, 故应执行:

- $Work = Work + Allocation_i$;
 - $Finish(i) = true$;
 - Goto step 2 ;

- ④ 若所有进程的 $Finish(i)$ 都为true, 则表示系统处于安全状态; 否则, 系统处于不安全状态

一个事实: 这里的安全路径可能有多条!

一个新问题: 是否存在, 有分叉的安全路径情况下, 安全性算法搜索到了一条不安全路径呢?

可以证明: 只要存在一个序列不是安全序列, 那么任意路径都不是安全序列。

只要有一个序列是安全序列, 那么在算法进行过程中出现的任何分叉点所构成的其它序列就都是安全序列。

- 例子

第二张图的alloc就是第一张图的allocation, 就是已经分配给该进程的资源数

- 假定系统中有5个进程 P0、P1、P2、P3、P4和三种类型的资源A、B、C，数量分别为12、5、9，在T0时刻的资源分配情况如下所示。

资源情况 进程	Max			Allocation			Need			Available		
	A	B	C	A	B	C	A	B	C	A	B	C
P0	8	5	3	1	1	0	7	4	3	3	3	2
P1	3	2	3	2	0	1	1	2	2			
P2	9	0	3	3	0	3	6	0	0			
P3	2	2	2	2	1	1	0	1	1			
P4	5	3	3	1	0	2	4	3	1			

资源情况 进程	Work			Need			Alloc			Work+Alloc			Finish
	A	B	C	A	B	C	A	B	C	A	B	C	
P1	3	3	2	1	2	2	2	0	1	5	3	3	true
P3	5	3	3	0	1	1	2	1	1	7	4	4	true
P4	7	4	4	4	3	1	1	0	2	8	4	6	true
P2	8	4	6	6	0	0	3	0	3	11	4	9	true
P0	11	4	9	7	4	3	1	1	0	12	5	9	true

- 从上述分析得知，T₀时刻存在着一个安全序列< P1、P3、P4、P2、P0 >，故系统是安全的，T₀是安全的

处理死锁方法3:死锁的检测与解除

■ 基本思想

- 对资源的分配不施加限制，也不采取死锁避免措施，系统定时的运行“死锁检测”程序
- 判断系统内是否已经出现死锁，如果系统出现死锁，则采取某种措施解除死锁。

■ 特点：

- 死锁检测和解除可使系统获得较高的利用率
- 需要确定何时运行检测算法，执行频率如何

• 依据资源分配图来判定死锁

资源分配图与死锁状态的关系：

- 如果资源分配图中无环路，则此时系统没有发生死锁
- 如果资源分配图中有环路，且每个资源类中仅有一个资源，则系统中发生死锁，此时，环路是系统死锁的充分必要条件，环路中的进程就是死锁进程
- 如果资源分配图中有环路，且涉及的资源类中包含多个资源，则环路的存在只是产生死锁的必要条件，而非充分条件，系统未必会发生死锁。
- 如何求解死锁的充分条件？**死锁定理**

• 资源分配图的化简

自己是这样做的，看每个进程节点是否能够删除与其相关联的边

- ① 在资源分配图中，找出一个**既不阻塞又非孤立的进程结点** p_i
 - 非孤立：指该结点是一个有边与之相连的
 - 非阻塞：该结点没有因为资源请求与分配而导致等待
 - 即资源申请数量不大于系统已有空闲资源数量的进程
 - 亦即该进程节点的出边+被申请资源节点的出边 \leq 被申请资源的数量
- ② 当进程 p_i 获得了它所需要的全部资源，则能运行完成，然后释放所有资源
 - 即相当于消去 p_i 的所有请求边和分配边，使之成为孤立结点。
 - 进程 p_i 释放资源后，可以唤醒因等待这些资源而阻塞的进程，从而可能使原来阻塞的进程变为非阻塞进程。
- ③ 在进行一系列化简后，若能消去图中所有的边，使所有进程都成为孤立结点，则称该图是**可完全简化的**；若不能使该图完全化简，则称该图是**不可完全简化的**。

• 死锁解除

- 一旦检测出系统中出现了死锁，就应将陷入死锁的进程从死锁状态中解脱出来，常用的死锁解除方法有：
 - 系统重启法：结束进程执行，重新启动系统
 - 进程终止：终止进程
 - 进程撤销法：撤消全部死锁进程，使系统恢复到正常状态
 - 逐步撤销法：按照某种顺序逐个撤消死锁进程，直到有足够的资源供其他未被撤消的进程使用，消除死锁状态为止。
 - 当逐步撤销时，如何选取终止哪些进程？计算代价
 - 资源抢占：剥夺陷于死锁进程占用资源，但不撤销进程，直至死锁解除。
 - 选择牺牲进程：计算代价
 - 回滚：将牺牲进程回滚到安全状态，完全回滚 or 回滚到打破死锁
 - 饥饿：如何保证资源不会总从同一个进程中被抢占？