Lecture 8: 存储与文件结构

Database

Author: Forliage

Email: masterforliage@gmail.com

Date: June 4, 2025

College: 计算机科学与技术学院

Abstract

本讲笔记围绕数据库系统中存储与文件结构的基本原理展开，共分为四个主要部分：物理存储介质概述、磁盘原理与性能、存储访问与缓冲管理，以及文件组织方法。

首先，“物理存储介质概述”介绍了不同类型存储设备的特点与层次，包括缓存（Cache）、主存（Main Memory）、闪存（Flash Memory）、磁盘（Magnetic Disk）、光存储（Optical Storage）与磁带（Tape Storage）等。重点分析了它们在速度、成本、可靠性等方面的差异，阐明了存储层次结构在数据访问效率与持久性保障之间的权衡关系。

其次，“磁盘”部分深入探讨了传统磁盘的结构与性能指标。内容涵盖：磁盘读写头与盘片的扇区组织方式、柱面概念、磁盘控制器的功能（如校验和与坏扇区重映射）、磁头定位与寻道过程；以及磁盘子系统的组成和多个磁盘共享控制器的架构。随后从寻道时间、旋转等待时间、数据传输带宽及平均故障时间（MTTF）等方面，详细解析了磁盘性能的决定因素，并介绍了块（Block）概念与电梯（Elevator）调度算法对磁盘访问优化的原理。

第三部分“存储访问”重点讲解了数据库系统与磁盘之间的数据交互机制，包括缓冲区管理器的角色与工作流程。当应用程序请求某一磁盘块时，缓冲区管理器首先检查该块是否已驻留于内存缓存；如未命中，则需根据替换策略（如LRU、MRU等）腾出空间并将所需块从磁盘读入。章节还讨论了固定块与强制输出策略、引用计数的使用，以及非易失性写缓冲与日志磁盘在加速写操作、保证事务持久性方面的作用。

最后，“文件组织”部分针对数据库文件在磁盘上的存储布局进行了系统说明。首先介绍了定长记录与可变长度记录的存储方法与优缺点；然后阐述了空闲列表技术用于管理已删除记录空间的原理；接着详细说明了槽页（Slotted Page）结构的设计，包括页头的记录条目数、可用空间结束指针，以及记录位置与长度等元数据信息；最后比较了定长表示与指针方法（链式溢出块）在可变长度记录存储中的应用场景与空间效率。

通过以上内容，读者能够掌握数据库系统底层的存储介质特性、磁盘访问原理、缓存管理机制与文件组织策略，为后续学习索引结构、查询执行与性能优化奠定坚实基础。

（此Abstract由ChatGPT-o4-mini-high生成）

Contents

1 物理存储介质概述

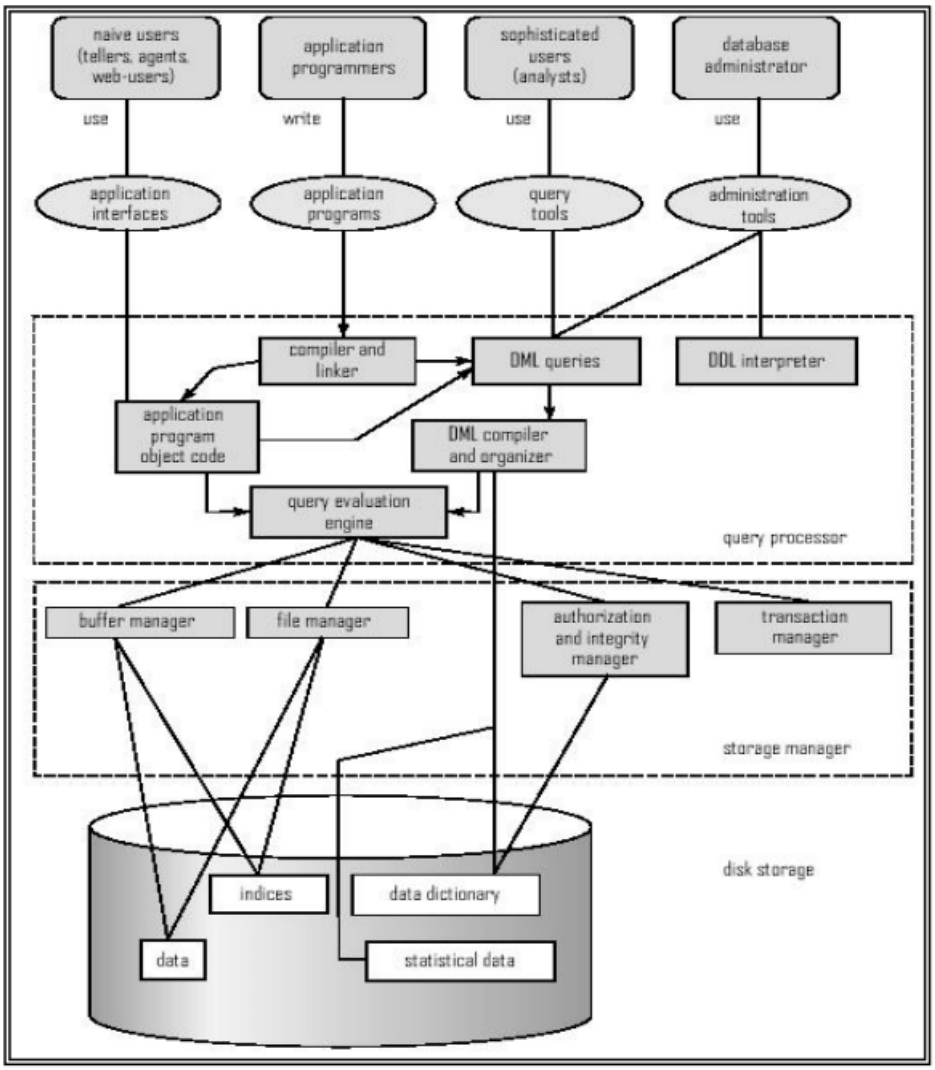

Figure 1. Overall System Structure

1.1 物理存储介质的分类

- 数据访问速度
- 每单位数据的成本
- 可靠性:停电或系统崩溃时的数据丢失/存储设备的物理故障

1.2 存储层次结构

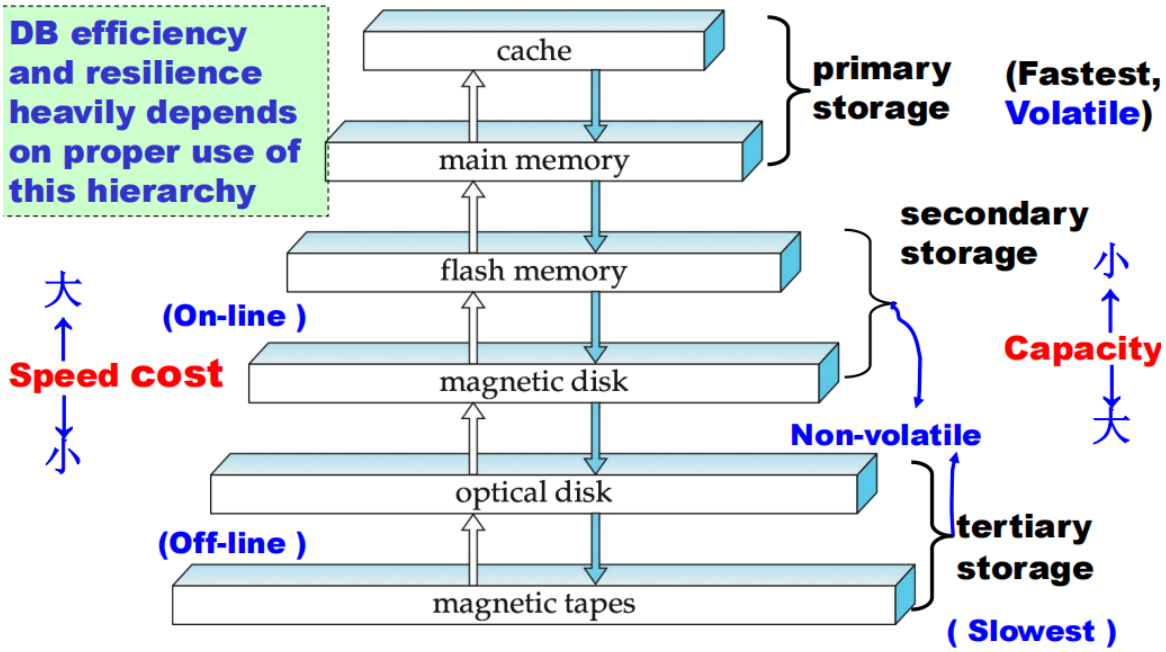

Figure 2. 存储层次结构

- 主存储:最快的存储介质,但易失性(缓存/主存).
- 二级存储(辅助存储器,联机存储器):层次结构中的下一级,非易失性,访问时间适中.也称联机存储,例如闪存/磁盘.
- 三级存储(三级存储器,脱机存储器):层次结构中的最低级别,非易失性,访问时间慢.也成为脱机存储.例如磁带/光存储.

按可靠性分类的存储设备：

- 易失性存储：断电时内容丢失，例如DDR2、SDR。
- 非易失性存储（非易失性存储器）：即使断电，内容仍会保留,包括二级和三级存储,以及电池备份的主存储器。

按速度分类的存储设备：

- Cache
- Main memory
- Flash memory
- Magnetic disk
- Optical storage
- Tape storage

1.3 物理存储介质

Cache:最快且成本最高的存储形式，易失性，由计算机系统硬件管理。速度≤0.5 ns;大小:~ $KB \sim MB$

主存储器:快速访问(10到100ns);通常太小或太贵,无法存储整个数据库;易失性：如果发电源故障或系统崩溃，主内存中的内容通常会丢失。

闪存(快闪存储器):

- 也称为 EEPROM(电可擦可编程只读存储器)
- 断电时数据仍可保留
- 数据只能在一个位置写入一次，但该位置可以擦除后再次写入:仅支持有限数量（10K - 1M）的写入/擦除周期,必须对整个内存组进行内存擦除操作
- 读取速度 致与主内存相同（< 100 纳秒）
- 但写入速度较慢 (~ 10ns) ，擦除速度更慢
- 每单位存储成本 致与主内存相似
- 广泛应用于数码相机、手机和 U 盘等嵌入式设备中

磁盘:

- 数据存储在旋转的磁盘上，并通过磁方式进行读写
- 数据 期存储的主要介质，通常存储整个数据库
- 访问数据时必须将其从磁盘移动到主内存，存储时再写回:访问速度比主内存慢得多
- 直接访问：与磁带不同，可按任意顺序读取磁盘上的数据
- 可在停电和系统崩溃后保留数据:磁盘故障可能会破坏数据，但这种情况非常罕

光存储:

- 非易失性，使用激光从旋转的磁盘上以光学方式读取数据
- 最流行的形式是CD - ROM（640 MB）和DVD（4.7至17 GB）
- 一次写入、多次读取（WORM）光盘用于存档存储
- 也有多次写入版本（CD - RW、DVD - RW和DVD - RAM）
- 读写速度比磁盘慢。
- 自动光盘机系统，配有大量可移动磁盘、几个驱动 以及用于自动装卸磁盘的机制，可用于存储大量数据

磁带存储:

- 非易失性，主要用于备份（从磁盘故障中恢复）和存档数据
- 顺序访问——比磁盘慢得多

- 容量非常大（有 40 到 300GB 的磁带可供使用）
- 磁带可以从驱动 中取出$\Longrightarrow$存储成本比磁盘便宜得多，但驱动器价格昂贵
- 可用于存储大量数据的磁带自动换带机

2 磁盘

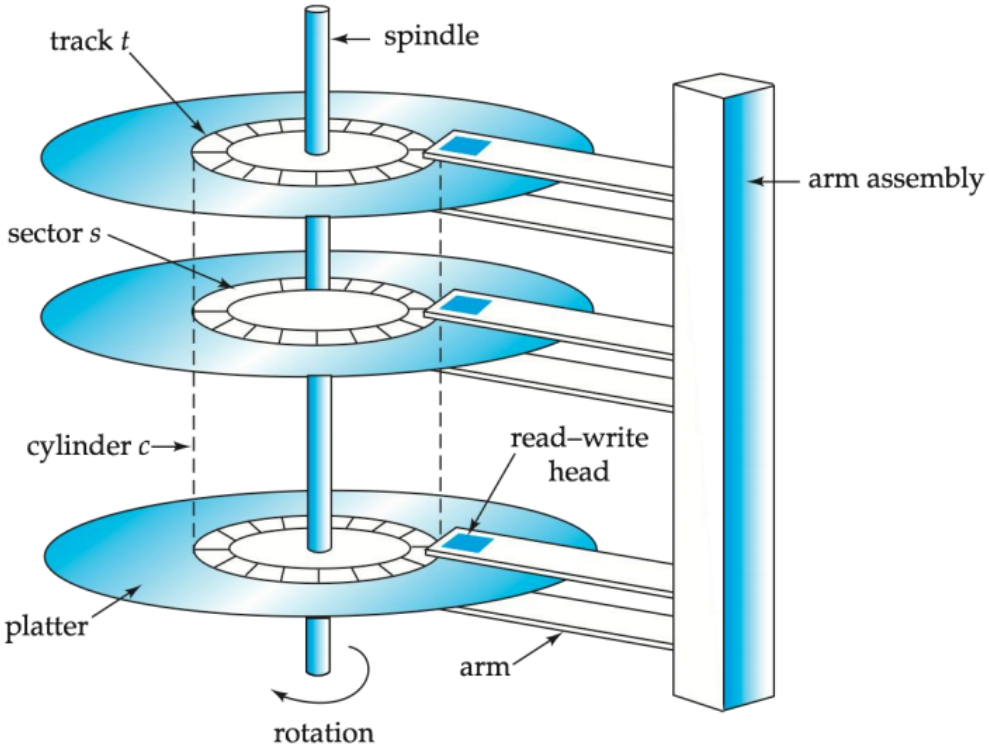

Figure 3. 磁盘结构

2.1 磁盘

读写头:定位在非常靠近盘片表面的位置（几乎接触到它）,读取或写入磁编码信息

盘片表面划分为圆形磁道:典型硬盘每个盘片上有超过 50 K ~ 100 K 条磁道

每个磁道被划分为扇区:

- 扇区是可读写的最小数据单位
- 扇区大小通常为 512 字节
- 每个磁道的典型扇区数：内磁道为 500 到 1000 个，外磁道为 1000 到 2000 个

读写扇区:磁盘臂摆动，使磁头定位到正确的磁道上;盘片持续旋转；当扇区经过磁头下方时，数据被读写

磁头-磁盘组件:单个主轴上有多个磁盘盘片(通常为4到16个),每个盘片有一个磁头,安装在一个公共臂上

柱面i由所有盘片的i^{th}磁道组成.

磁盘控制器:

- 接受读取或写入扇区的高级命令
- 启动诸如将磁臂移动到正确磁道并实际读取或写入数据之类的操作
- 为每个扇区计算并附加校验和,以验证数据是否被正确读回:如果数据损坏,存储的校验和极有可能与重新计算的校验和不匹配.
- 通常在写入扇区后回读该扇区来确保写入成功
- 执行坏扇区的重映射(坏扇区的重映射:将该扇区从逻辑上映射到预留的物理扇区,并且重映射被记录在磁盘或其它非易失性存储器中)

2.2 磁盘子系统

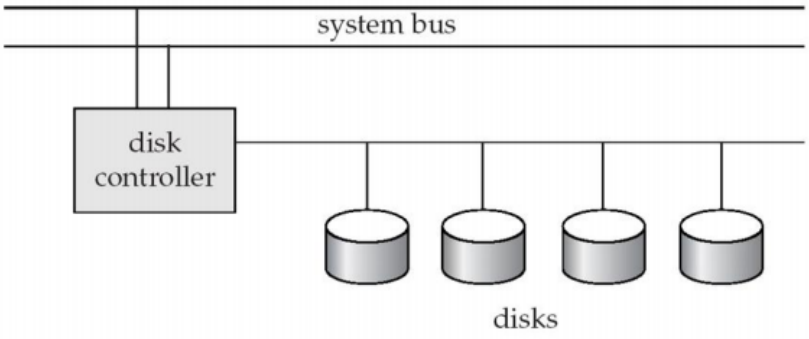

Figure 4. 磁盘子系统

多个磁盘通过控制器连接到计算机系统:控制器功能(校验和/坏扇区重映射)通常由单个磁盘执行;减轻控制器负载.

2.3 磁盘的性能指标

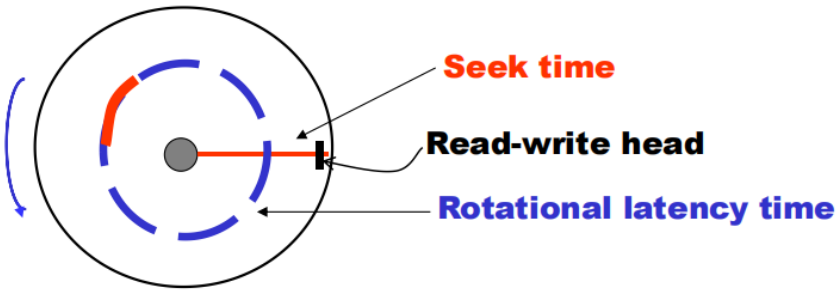

Figure 5. 磁盘的性能指标

访问时间:从发出读写请求到开始数据传输所花费位的时间(=寻道时间+旋转等待时间)

- 寻道时间:将磁臂重新定位到磁道所需的时间,典型磁盘上为4到10毫秒.
- 旋转等待时间:待访问扇区旋转到磁头下方所需的时间.平均等待时间是最坏情况下等待时间的1/2,典型磁盘(5400到15000转/分钟)上为4到11毫秒.

数据传输速率:从磁盘检索数据或向磁盘存储数据的速率.

- 25至100MB每秒最大速率,内圈磁道速率较低
- 多个磁盘可能共享一个控制器,因此控制器能够处理的速率也很重要.

平均故障时间(MTTF, Mean Time To Failure):磁盘预期连续运行且不发生任何故障的平均时长.

- 通常为3到5年
- 新的磁盘的故障概率相当低,对应新磁盘的平均故障时间为1,200,000小时意味着,假设有1000块较新的磁盘,平均每1200小时会有一块磁盘发生故障
- 平均故障时间随磁盘使用年限增加而降低

2.4 磁盘块访问优化

块:来自单个磁道的连续扇区序列

- 数据以块为单位在磁盘和主存之间传输
- 大小范围从512字节到几千字节
  - 较小的块:从磁盘进行更多次传输
  - 较大的块:由于块部分填充而浪费更多的空间
  - 如今,典型的块大小为4到16千字节

磁盘臂调度算法对挂起的磁道进行排序,以使磁盘臂的移动最小化: **电梯算法**:沿一个方向(从外磁道到内磁道或反之)移动磁盘臂,处理该方向上的下一个请求,直到该方向上没有更多请求,然后反转方向并重复此过程.

文件组织:通过组织块以对应数据的访问方式来优化块访问.

- 例如,将相关信息存储在相同或相邻的柱面上.
- 随着时间推荐,文件可能会碎片化:
  - 例如,如果向文件中插入数据或从文件中删除数据
  - 或者磁盘上的空闲块分散,新创建的文件的块也分散在磁盘上

  - 顺序访问碎片化文化会导致磁盘臂移动增加
- 一些系统提供了对文件系统进行碎片整理的实用程序,以加快文件访问速度

(但这些实用程序运行时,系统通常无法正常使用.)

非易失性写缓冲区通过立即将块写入非易失性随机存取存储器(RAM)缓冲区来加速磁盘写入.

- 非易失性随机存取存储器:由电池供电的随机存取存储器或闪存即使断电,数据也是安全的,并且在恢复供电时会被写入磁盘.
- 然后,只要磁盘没有其它请求或者某个请求已经挂起一段时间,控制器就会将数据写入磁盘.
- 要求在继续操作之前安全存储数据的数据库操作可以在不等待数据写入磁盘的情况下继续进行.
- 然后可以对写入操作进行重新排序,以最小化磁盘臂的移动.

日志磁盘:专门用于写入块更新顺序日志的磁盘.

使用方式与非易失性随机存取存储器完全相同:

- 由于无需寻道,写入日志磁盘的速度非常快
- 无需特殊硬件(非易失性随机存取存储器)

文件系统通常会对磁盘写入操作进行重新排序以提高性能

- 日志文件系统按安全顺序将数据写入非易失性随机存取存储器(NV-RAM)或日志磁盘
- 无日志记录的重新排序:存在文件系统数据损坏的风险

3 存储访问

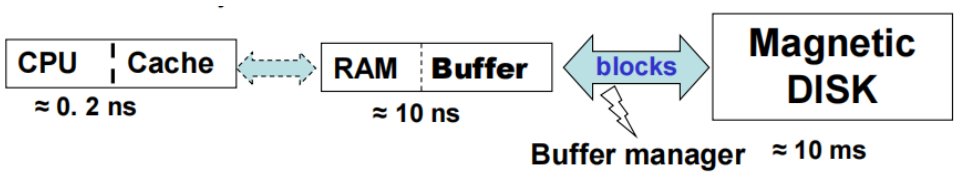

Figure 6. Storage Access

3.1 存储访问

数据库文件在逻辑上被划分为称为块的固定长度存储单元.块是数据库系统中存储分配和数据传输的单位.

缓冲区:用于存储磁盘块副本的内主存部分.

数据库系统力求最小化磁盘和内存之间的块传输次数.

- 为减少磁盘访问次数:在主内存中尽可能多地保留块—缓冲区.
- 但缓冲区大小有限,如何处理

缓冲区管理区:负责在主内存中分配缓冲区空间的子系统.

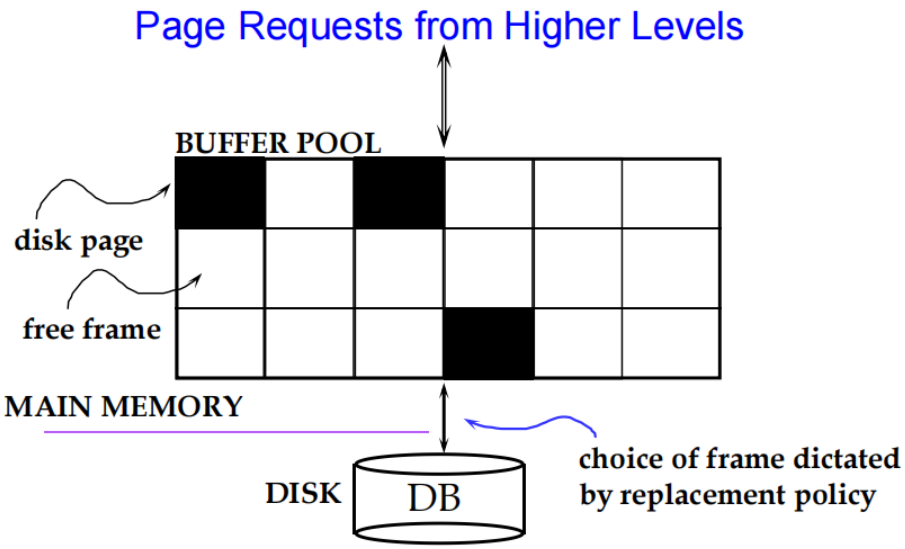

Figure 7. Structure

数据必须存于RAM中,DBMS才能对齐进行操作

会维护<frame#,pageid>对的表格.

页:数据的单位;块:磁盘空间的单位:帧:缓冲池的单位.

3.2 缓冲区管理器

应用程序在需要从磁盘获取一个块时会调用Buffer Manager.

- 如果该块已在缓冲区中,请求程序将获得该块在主存中的地址
- 如果不在缓冲区中:
  - 缓冲区管理器在缓冲区中为该块分配空间,替换(丢弃)一些旧页面;如果没有空闲空间,则为新块腾出空间.(在buffer中为新页分配空间).
  - 只有当被丢弃的块自最近一次写入磁盘或从磁盘读取以来被修改过时,才会将其写回磁盘(将被覆盖的旧块若已被修改过,则写回磁盘).
  - 一旦在缓冲区中分配了空间,缓冲区管理器就会将数据块从磁盘读取到缓冲区,并将该数据块在主存中的地址传递给请求者(从磁盘读入新块放buffer).

缓冲区替换策略:LRU(最近最少使用),MRU(最近最多使用).

固定块:不允许写回磁盘的内存块.(如当前块正在被使用时)

立即丢弃策略:一旦处理完某个块的最后一个元组,就立即释放该块占用的空间.(用后立即丢弃)

强制输出块:块的请求者必须解除对其的固定,并指明页面是否已被修改——为此使用dirty bit.

pool中的页面可能会被多次请求(被多个事务使用):使用一个引用计数,当且仅当引用计数=0时,页面才是替换候选对象.

3.3 缓冲区替换策略

当Buffer的空闲区不够,不能容下新读入的Block时,需要将Buffer中原有Block覆盖(替换).主要策略为:

(1) **LRU策略**:替换最近最少使用的块.

- LRU背后的理念:将过去的块引用模式用作对未来引用的预测器.
- 查询具有明确的访问模式,数据库系统可以利用用户查询中的信息来预测未来的引用.
- 但对于某些涉及重复扫描数据的访问模式,LRU可能是一种糟糕的策略

(2) **MRU策略**:系统必须固定当前正在处理的块.处理完该块的最后一个元组后,该块被解除固定,并成为最近最常用的块.

- 缓冲区管理器可以使用关于请求引用特定关系的概率的统计信息:例如,数据字典经常被访问.启发式方法:将数据字典块保留在主内存缓冲区中.
- 缓冲区管理器还支持为恢复目的强制输出快.
- 采用查询优化器提供的替换策略提示的混合策略更为可取.

4 文件组织

4.1 文件组织

数据库以文件集合的形式存储.

每个文件是一系列记录.

一条记录是一系列字段.

两种记录:

- 定长记录.
- 可变长度记录.

4.2 定长记录

优点:方法简单:

- 从字节$n*(i-1)$开始存储记录i,其中n是每条记录的大小.
- 记录访问简单,但记录可能跨块.
- 修改:不允许记录跨越块边界

删除记录i:可选方法:

- 方法1:将记录$i+1,...,n$移动到$i,...,n-1$
- 方法2:将记录n移动到i
- 方法3:不移动记录,但将所有空闲记录链接到空闲列表上

4.3 空闲列表

将文件头中第一个已删除记录的地址存储(还有其它信息).

使用第一条记录来存储第二条已删除记录的地址,以此类推.

可以将这些存储的地址视为指针,因为它们"指向"记录的位置.

3个优势:更多空间高效表示:复用普通空间免费记录的属性用于存储指针.(无存储在使用中的指针记录)

4.4 可变长度记录

可变长度记录在数据库系统中有多种产生方式:

- 在文件中存储多种记录类型
- 允许一个或多个字段(如字符串(varchar))具有可变长度的记录类型
- 允许重复字段的记录类型(在一些旧的数据模型中使用)

属性按顺序存储.

由固定大小(偏移量/长度)表示的可变长度属性,实际数据存储在所有固定长度属性之后

由空值位图表示对空值.

4.5 槽页结构

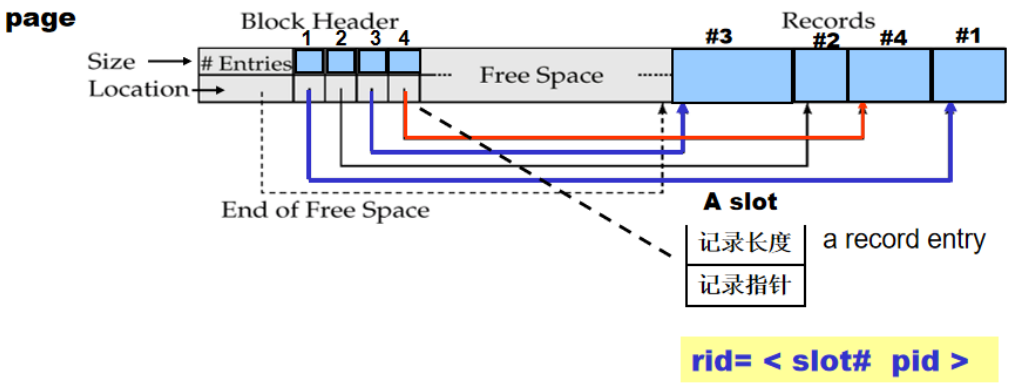

Figure 8. Slotted Page Structure

rid = <slot# pid>

槽式页面头部包含:

- 记录条目数量
- 块中可用空间结束
- 每条记录的位置和大小

记录可以在页面内移动,以保持它们连续且彼此之间没有空白空间;必须更新头部中的条目.(页内无碎块;删除时页内移动记录)

指针(索引)不应直接指向记录—相反,它们应指向头部中记录的条目—间接指针

4.6 定长表示

定长表示:预留空间,指针

预留空间:可以使用已知最大长度的定长记录;较短记录中的未使用空间用空字符或记录结束符号填充.

4.7 指针方法

可变长度记录由一系列固定长度记录表示,这些记录通过指针链接在一起.

即使不知道最大记录长度也可以使用.

指针结构的缺点:除了链中的第一条记录外,所有记录都会浪费(用于分支名称的)空间.

解决方案是允许文件中有两种块:

- 锚块:包含链的第一条记录
- 溢出块:包含除链的第一条记录外的其它记录.