数据结构与算法（十一）

张铭 主讲

采用教材：张铭，王腾蛟，赵海燕 编写
高等教育出版社，2008. 6（“十一五”国家级规划教材）

http://www.jpk.pku.edu.cn/pkujpk/course/sjjg

第十一章 索引

- 基本概念
- 11.1 线性索引
- 11.2 静态索引
- 11.3 倒排索引
- 11.4 动态索引
- 11.5 位索引技术
- 11.6 红黑树

输入顺序文件

- **输入顺序文件(entry-sequenced file)**

  按照记录进入系统的顺序存储记录

  - 输入顺序文件相当于一个磁盘中未排序的线性表
  - 因此不支持高效率的检索

主码

- **主码(primary key)** 是数据库中的每条记录的 **唯一** 标识
  - 例如，公司职员信息的记录的主码可以是职员的身份证号码
  - 如果只有主码，不便于各种灵活检索

辅码

- **辅码(secondary key)**

  是数据库中可以出现重复值的码

- 辅码索引把一个辅码值与具有这个辅码值的每一条记录的主码值关联起来
  - 大多数检索都是利用辅码索引来完成的

索引

- **索引(indexing)** 是把一个关键码与它对应的数据记录的位置相关联的过程
  - (关键码，指针)对，即(key, pointer)
  - 指针指向主要数据库文件（即“主文件”）中的完整记录
- 索引文件(index file) 是用于记录这种联系的文件组织结构
- 索引技术是组织大型数据库的一种重要技术
  - 高效率的检索
  - 插入、更新、删除

索引文件

- 一个主文件可以有多个相关索引文件
  - 每个索引文件往往支持一个关键码字段
  - 不需要重新排列重排主文件
- 可以通过该索引文件高效访问记录中该关键码值

稠密索引 vs 稀疏索引

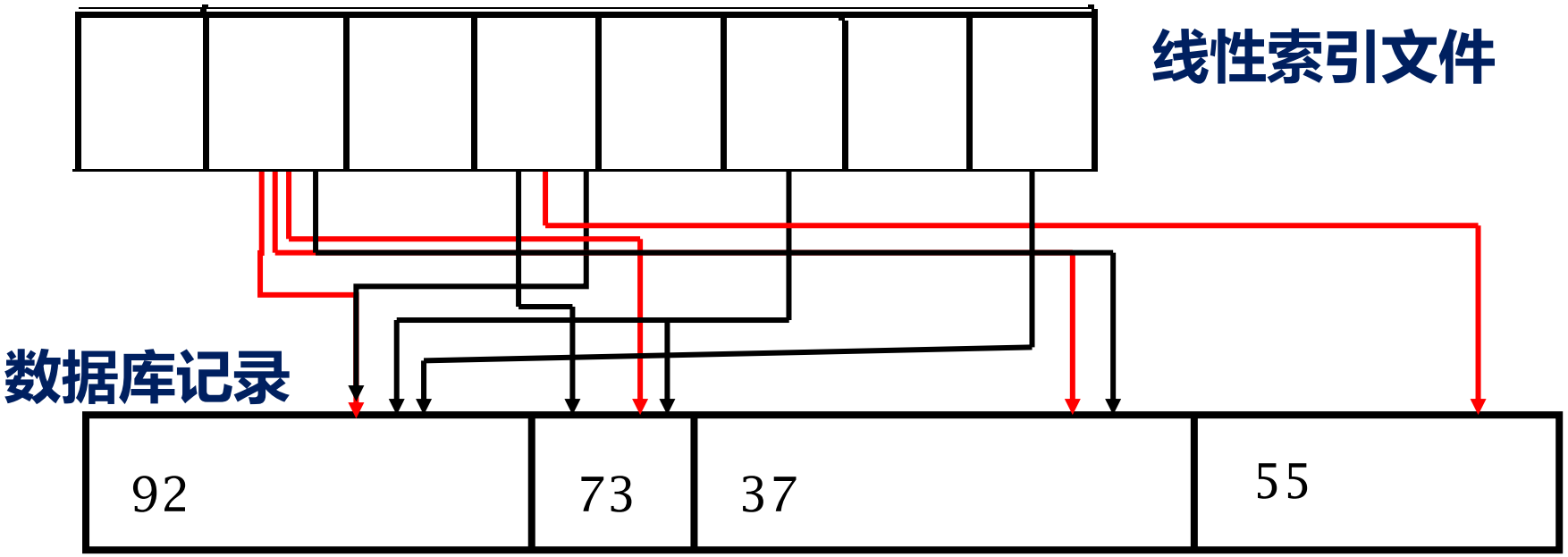
- 稠密索引：对 **每个** 记录建立一个索引项
  - 主文件不按照关键码的顺序排列
- 稀疏索引：对 **一组** 记录建立一个索引项
  - 记录按照关键码的顺序存放
  - 可以把记录分成多个组（块）
  - ➢索引指针指向的这一组记录在磁盘中的起始位置

11.1 线性索引

- 基本概念
- 线性索引的优点
- 线性索引的问题
- 二级线性索引

线性索引文件

- 按照关键码的顺序进行排序
- 文件中的指针指向存储在磁盘上的文件记录起始位置或者主索引中主码的起始位置



线性索引的问题

- 线性索引太大，存储在磁盘中
  - 在一次检索过程中可能多次访问磁盘，从而影响检索的效率
  - 使用二级线性索引

- 更新线性索引
  - 在数据库中插入或者删除记录时

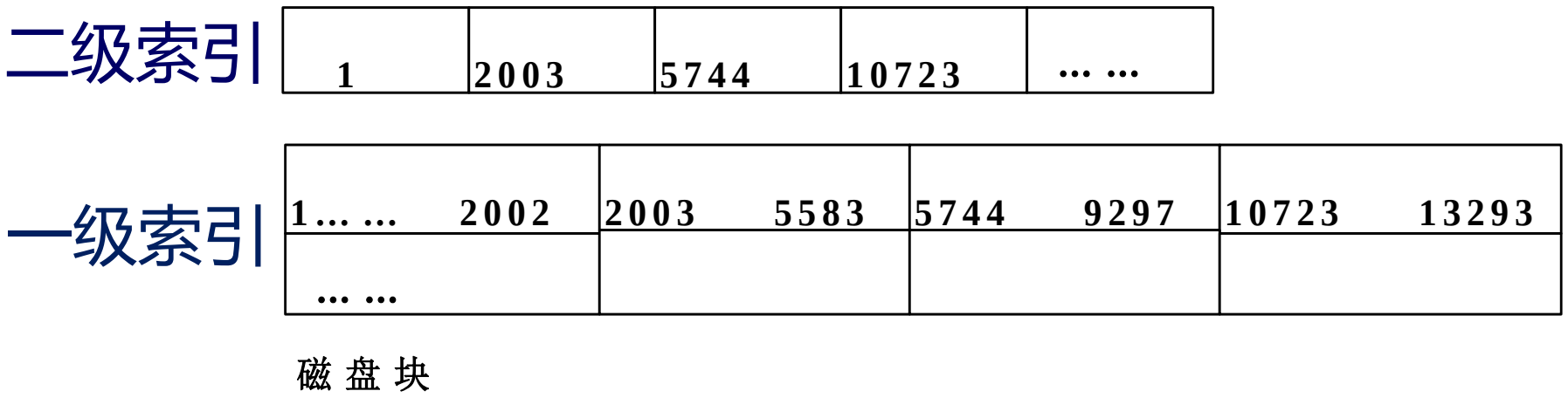
二级线性索引

- 例如，磁盘块的大小是 1024 字节，每对(关键码，指针)索引对需要 8 个字节
  - 1024 / 8 = 128
  - 每磁盘块可以存储 128 条这样的索引对

- 假设数据文件包含 10000 条记录
  - 稠密一级线性索引中包含 10000 条记录
    - 10000/128 = 78.1
    - 那么一级线性索引占用 79 个磁盘块
  - 相应地，二级线性索引文件中有 79 项索引对
  - 这个二级线性索引文件可以在一个磁盘块

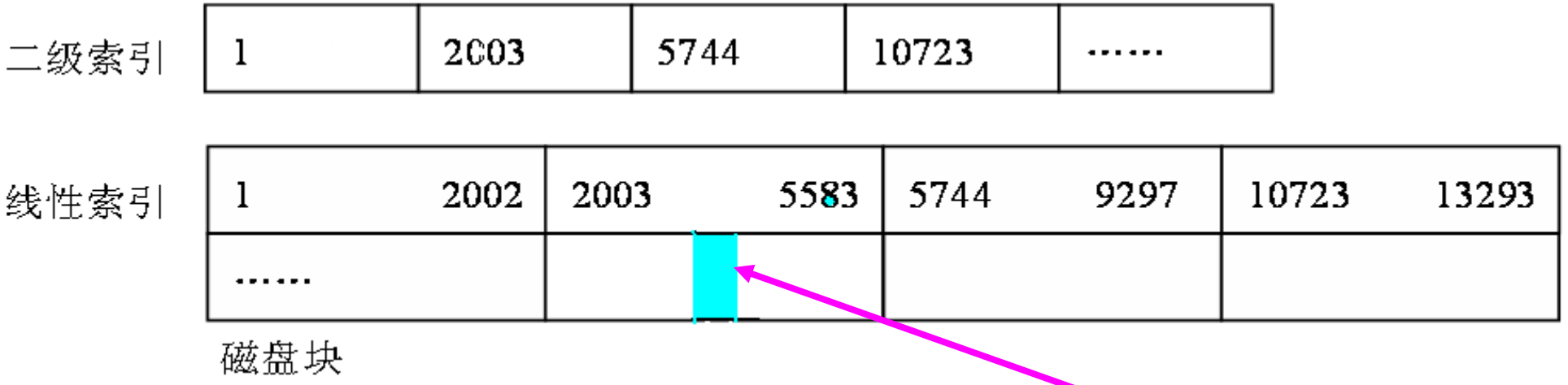
二级线性索引的例子

- 关键码与相应磁盘块中第一条记录的关键码的值相同
- 指针指向相应磁盘块的起始位置



关键字2555的记录指针

例如：检索关键码为2555的记录



1. 二级线性索引文件读入内存
2. 二分法找关键码的值小于等于2555的最大关键码所在一级索引磁盘块地址——关键码为2003的记录
3. 根据记录2003中的地址指针找到其对应的一级线性索引文件的磁盘块，并把该块读入内存
4. 按照二分法对该块进行检索，找到所需要的记录在磁盘上的位置
5. 最后把所需记录读入，完成检索操作

11.2 静态索引

静态索引

- 索引结构在文件创建、初始装入记录时生成
- 一旦生成就固定下来，在系统运行(例如插入和删除记录)过程中索引结构并不改变
- 只有当文件再组织时才允许改变索引结构

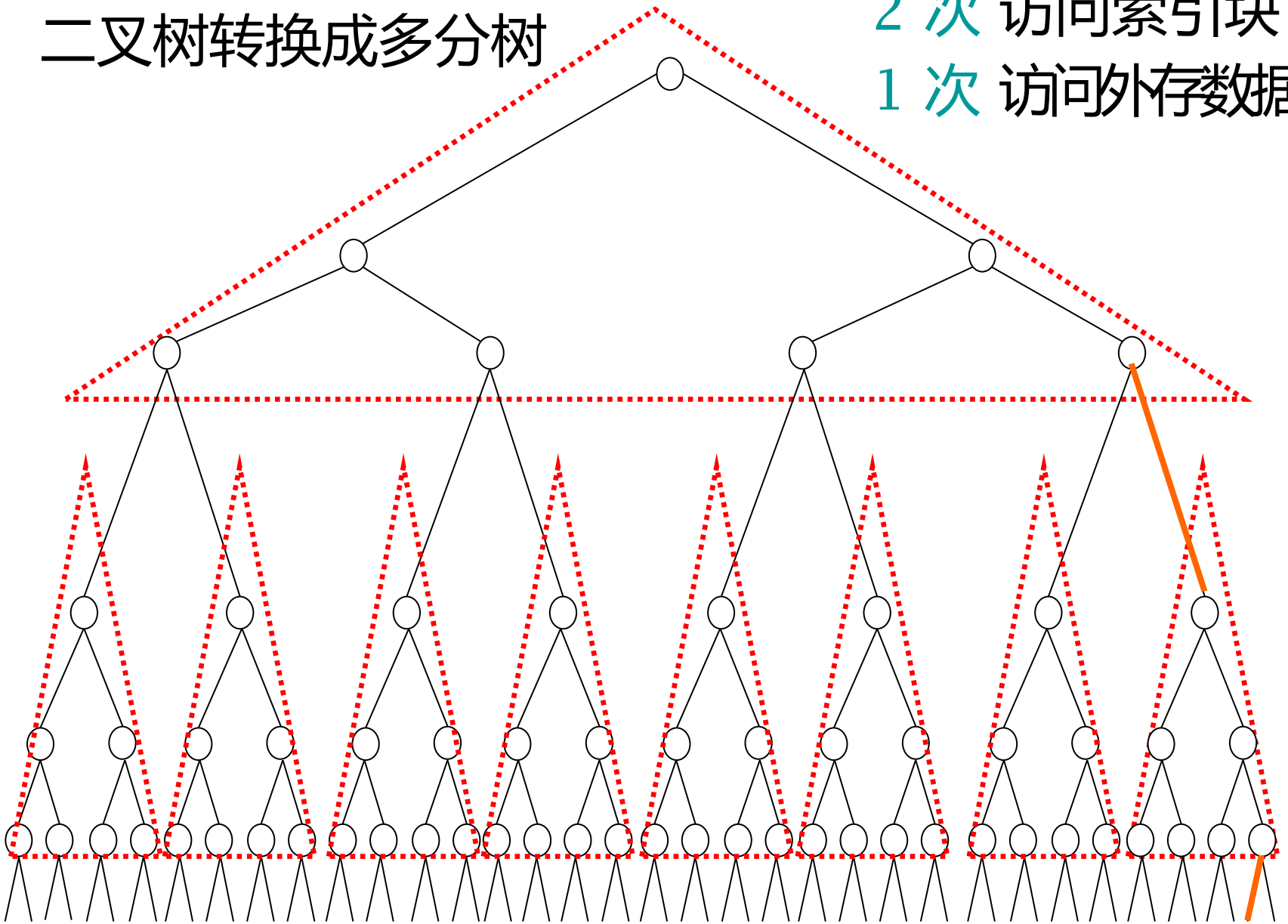
多分树

- 组织索引一般不用二叉树而采用多分树
- 大大减少访问外存的次数

二叉树转换成多分树

2 次 访问索引块

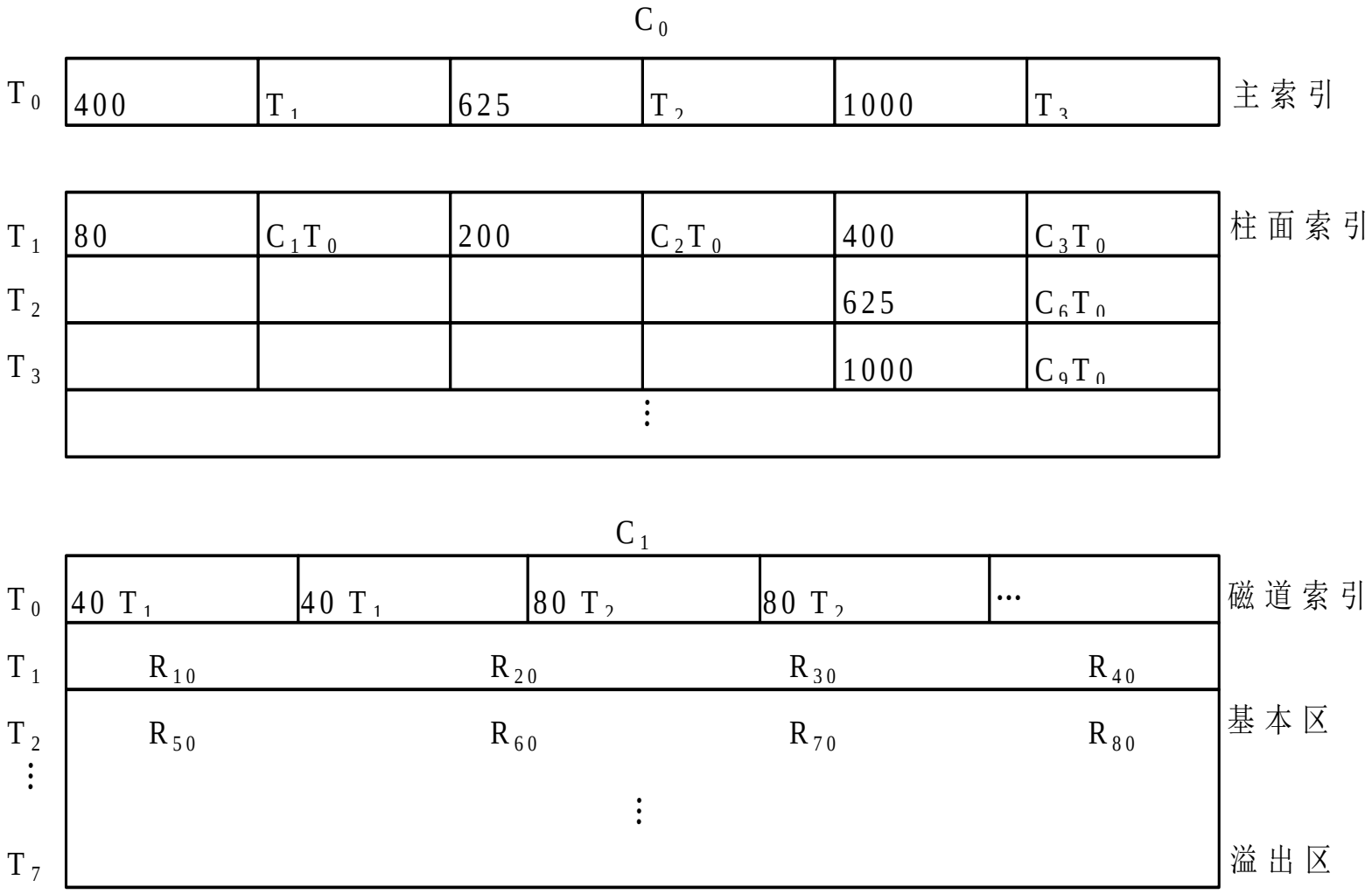
1 次 访问外存数据块



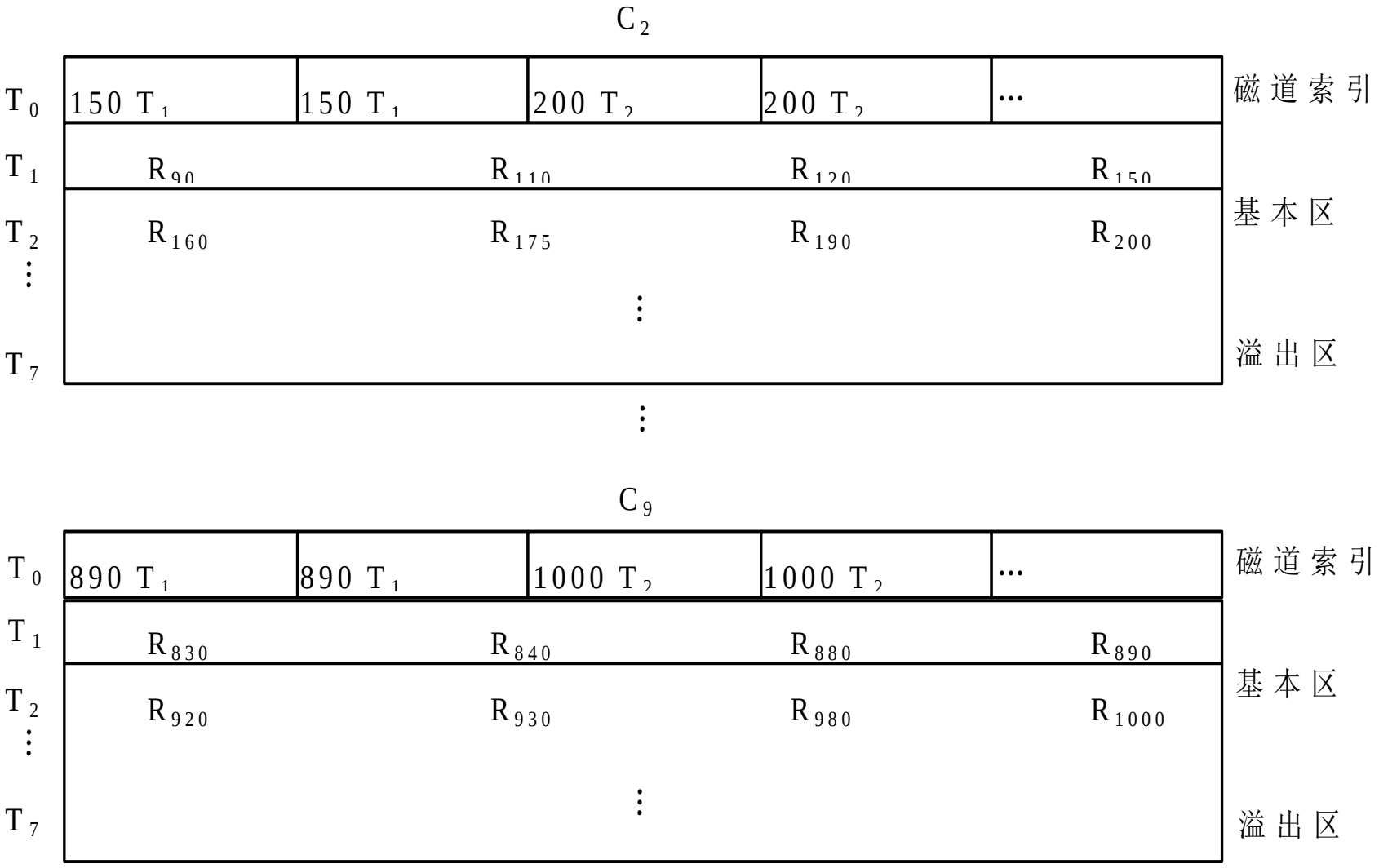
ISAM

- 基于多分树的 ISAM (Index Sequential Access Method)
  - 为磁盘存取而设计
  - 结构采用多级索引
    - 主索引
    - 柱面索引
    - 磁道索引
- 在采用基于 B^+ 树的 VSAM (Virtual Storage Access Method) 技术之前，IBM 公司曾经广泛地采用 ISAM 技术

11.2 静态索引


C0
T0
400
T1
625
T2
1000
T3
主索引
T1
80
C1T0
200
C2T0
400
C3T0
柱面索引
T2
625
C6T0
T3
1000
C9T0
C1
T0
40 T1
40 T1
80 T2
80 T2
…
磁道索引
T1
R10
R20
R30
R40
T2
R50
R60
R70
R80
基本区
T7
溢出区

11.2 静态索引


C2
T0
150 T1
150 T1
200 T2
200 T2
…
磁道索引
T1
R90
R110
R120
R150
T2
R160
R175
R190
R200
基本区
T7
溢出区
C9
T0
890 T1
890 T1
1000 T2
1000 T2
…
磁道索引
T1
R830
R840
R880
R890
T2
R920
R930
R980
R1000
基本区
T7
溢出区

思考

- 在什么情况下需要组织二级线性索引？
- 多分树的阶（子结点的个数）应该怎么确定？

数据结构与算法

谢谢聆听

国家精品课“数据结构与算法”
http://www.jpk.pku.edu.cn/pkujpk/course/sjjg/

张铭，王腾蛟，赵海燕
高等教育出版社，2008. 6。“十一五”国家级规划教材