第14周小结

1 外排序概述

- 排序数据存放在外存上
- 排序结果存放在外存上
- 排序过程借助内存实现

存在内、外存数据交换（多）
存在关键字比较（多）
存在元素移动（少）

外排序时间=内、外存数据交换+关键字比较

2 磁盘排序

- 通过分割要排序的文件，生成多个初始归并段
- 对初始归并段进行多路归并，产生一个有序文件

❶ 生成多个初始归并段

常规方法：

内存大小为w，每次从外存文件in.dat中读入w个记录，采用某种内排序方法进行排序来产生初始归并段，即产生out_1.dat、…、out_n.dat有序文件

⬇

产生初始归并段个数为$\lceil n/w \rceil$，长度基本相同

置换-选择算法

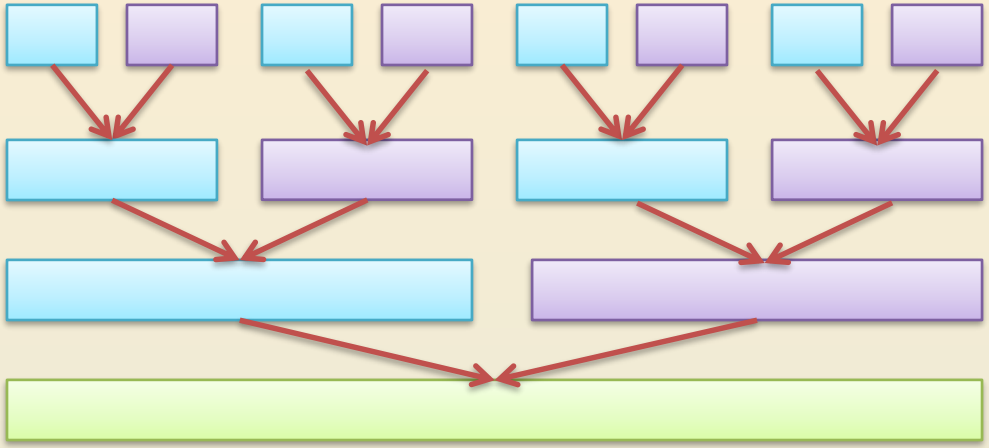
⬇

产生初始归并段个数 $< \lceil n/w \rceil$，长度差异比较大

❷ 多路归并

多路平衡归并：

k=2



记录读写次数 = WPL

k越大WPL越小

k路归并中：大量的操作是从k个记录中找出最小的记录

采用简单比较实现 ⇨ 效率低

采用类似堆的方式即败者树 ⇨ 效率高

归并中关键字比较次数与k无关

尽可能增加k提高外排序效率

按最佳归并树进行归并：

- 根据初始归并段（个数m和每个初始归并段中记录数）和k构造最佳归并树。
- 按照其过程进行归并。

设有11个初始归并段，它们所包含的记录个数为{25，40，16，38，77，64，53，88，9，48，98}。试根据它们做4路归并，要求：

（1）指出采用4路平衡归并时总的归并趟数。

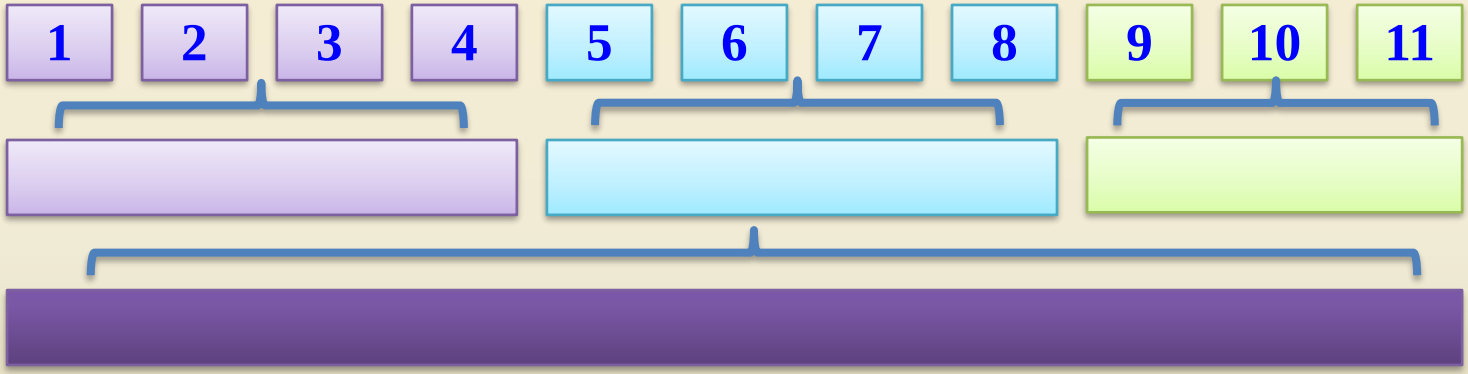
（2）给出采用4路平衡归并时的归并过程。

（3）构造最佳归并树。

（4）根据最佳归并树计算每一趟及总的读记录数。

解：（1）采用4路平衡归并时，$m=11$，$k=4$，总的归并趟数$=\lceil \log_k m \rceil=\lceil \log_4 11 \rceil=2$。

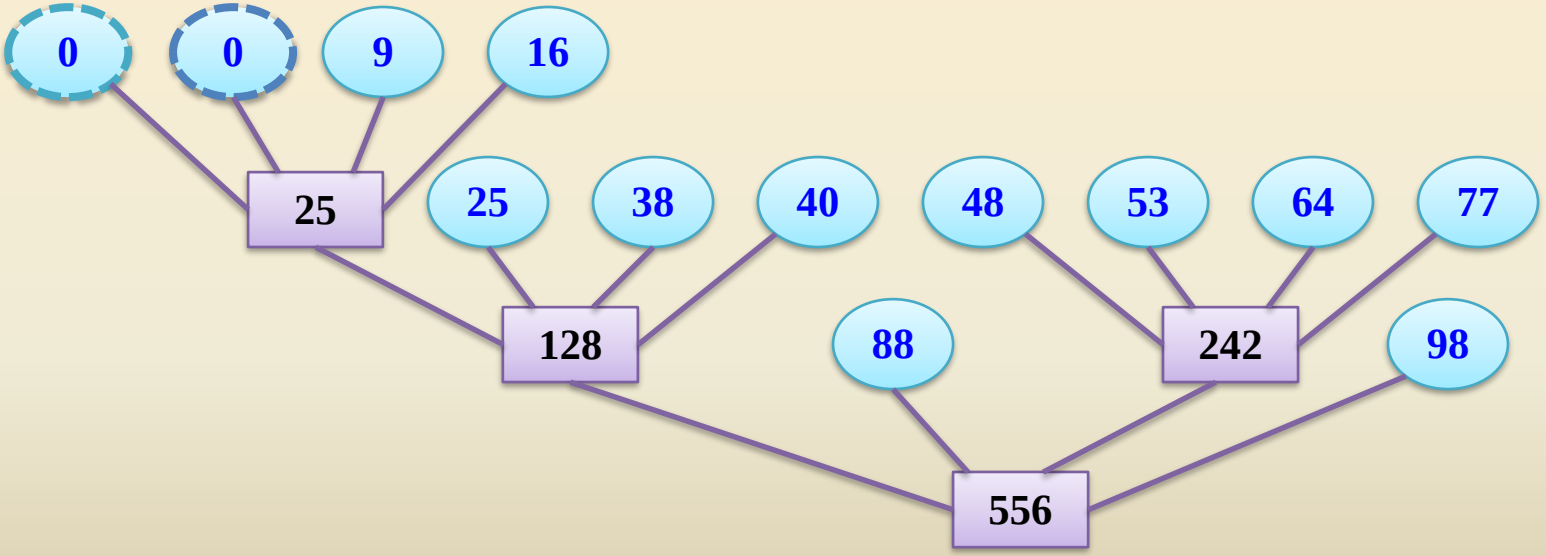
（2）采用4路平衡归并时的归并过程如下（归并段编号1～11）：

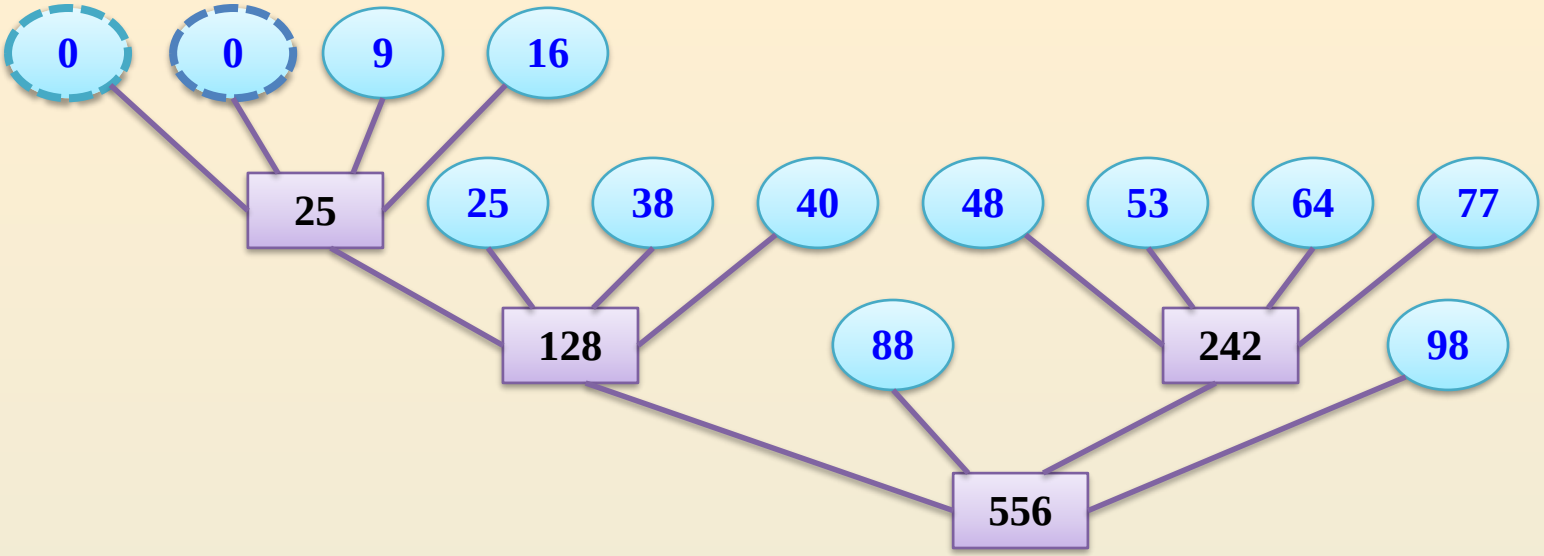


没有考虑归并顺序⇨多路平衡归并适合长度相同的归并段

（3）构造最佳归并树。

m=11，k=4，(m-1) % (k-1)=1≠0，需要附加k-1-(m-1) % (k-1)=2个长度为0的虚归并段，最佳归并树如下。





（4）根据最佳归并树计算每一趟及总的读记录数：

第1趟的读记录数=9+16=25

第2趟的读记录数=25+25+38+40+48+53+64+77=370

第3趟的读记录数=128+88+242+98=556

总的读记录数=25+370+556=951。

WPL最小

归纳：

- 生成多个初始归并段采用常规方法，产生长度相同的归并段 ⇨ 宜采用多路平衡归并（归并中用败者树）
- 生成多个初始归并段采用置换-选择算法，产生长度不相同的归并段 ⇨ 宜采用最佳归并树方案（归并中用败者树）