Database HM4

120090694 佟臻

▼ **Question 1**

$p_k = \frac{C}{k},\ k = 1, 2, ..., N$

因此，位于文件开头的记录有更高的机会被选中。

▼ a)Let X_N be the (random) number of comparisons to locate a given record present in the file of N records. Derive an expression for the average E(XN) for the Zipf distribution (the expression should not include the constant C).

$\sum_{i=1}^{N} p_k = \sum_{i=1}^{N} \frac{C}{k} = 1$

$C = \frac{1}{\sum_{i=1}^{N} \frac{1}{k}}$

$E(X_N) = \sum p_k k = \frac{N}{\sum_{i=1}^{N} \frac{1}{k}}$

▼ b)Compare the search performance of the Zipf distribution against the uniform distribution ($i.e. p_k = 1/N$) for N = 10. (Note: since N is not large here, the exact formula should be used).

let Y_N be the number to find a record in a uniform distributed file.

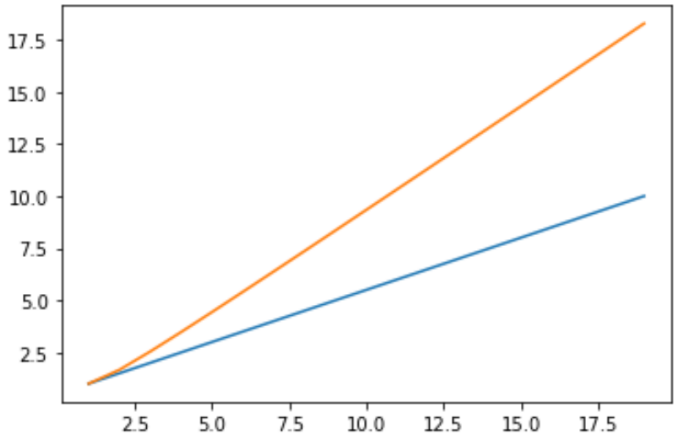
$E(Y_N) = \sum \frac{1}{N} k = \frac{1}{N} \frac{(1+N)N}{2} = \frac{1+N}{2} = 5.5$

$E(X_N) = \frac{10}{\frac{1}{1}+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}+\frac{1}{9}+\frac{1}{10}} = 3.414171521$

Therefore, Zipf distribution has better performance

▼ c)Determine the file size N* where $E(X_{N^*})$is the same for both the Zipf and the uniform distribution.

when $N^*$is 1, the two distributions perform the same



Zipf and Uniform

to find $\frac{N}{\sum 1/k} = \frac{1+N}{2}$

is to find$\sum 1/k = \frac{2N}{1+N}$

becasue $\frac{2N}{1+N} \leq 2$ and $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = 2.083$

only when N = {1,2,3} $\sum 1/k < 2$ holds

when N = 1, $\sum 1/k = \frac{2N}{1+N}$holds

when N = 2, 1.5≠4/3

when N= 3， 11/6 ≠ 3/2

Hence only **when N = 1** holds

▼ d)Consider now the file of N records is arranged in ascending order of the primary key, and assuming the required record is not present in the file. What is

the approximate average number of comparisons required to conclude the search?

By binary search, it will take $log_2 N$ time, which is the worst case

▼ e)Consider the file of N records is now arranged using the heap organization where records are not arranged in any particular order. Determine, under the uniform distribution, the average number of comparisons required to conclude the search for

（i）when the required record is present in the file,

$E(X_N) = \sum \frac{1}{N} k = \frac{1}{N} \frac{(1+N)N}{2} = \frac{1+N}{2}$

and (ii) when the required record is not present in the file.

$E(X_N) = \sum \frac{1}{N} N = \frac{1}{N} \frac{(1+N)N}{2} = N$

▼ **Question 2**

Consider a B-tree where each node can have **a maximum of 23 children**. Assume that each node of the B-tree is of average fullness. Consider that the root node is at Level 0. Determine for each of the following levels, the average number of nodes, the average number of key entries, and the average number of children pointers for that level

$N$is the random number of nodes in the tree

$Average = ln(2) \cdot Max$

▼ (i) level 1

average of nodes = $23 \cdot ln2 = 16$

average of key entities = $15 \cdot 16 = 240$

average of children pointer = $16 \cdot 16 = 240$

▼ (ii)

level 2

average of nodes = 256

average of key entities = 3084

average of children pointer= 4096

level 3

average of nodes = 4096

average of key entities = 61440

average of children pointer= 65536

▼ (iii)

level 4

average of nodes = $16^4 = 65536$

average of key entities = $16^4 \cdot 15 = 983040$

average of children pointer= $16^4 \cdot 16 = 1048576$

Determine the average number of entries that such a tree holds, assuming that:

▼ (iv)

height = 2

average of key entities = $15 + 240 + 3084 = 3339$

▼ (v)

average of key entities = $15 + 240 + 3084 + 61440 = 64779$

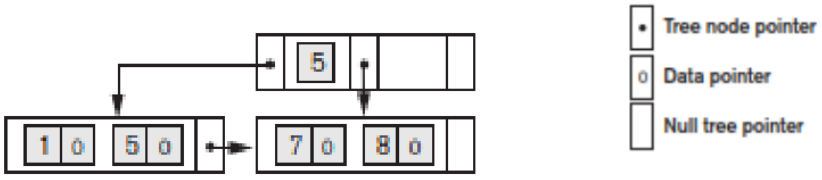
▼ (vi)

average of key entities = $15 + 240 + 3084 + 61440 + 983040 = 1047819$

general form for height h, average of key entities =$16^{h+1} - 1$

▼ Question 3

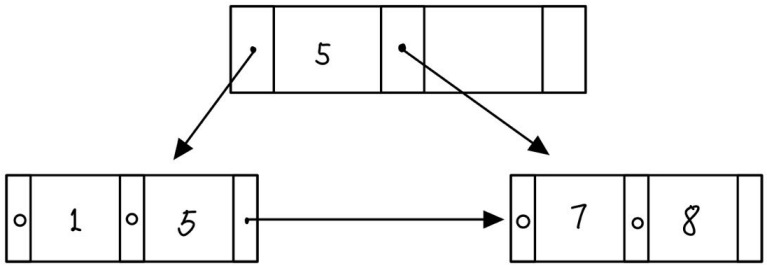
Consider the following B^+-Tree

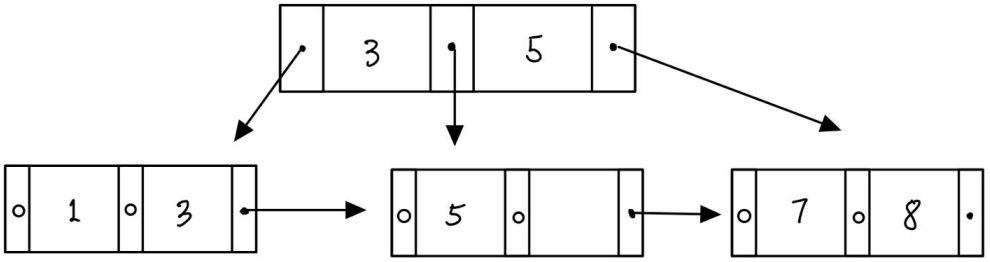


Show each of the tree structures after the insertion of

- (i) Key 3, followed by

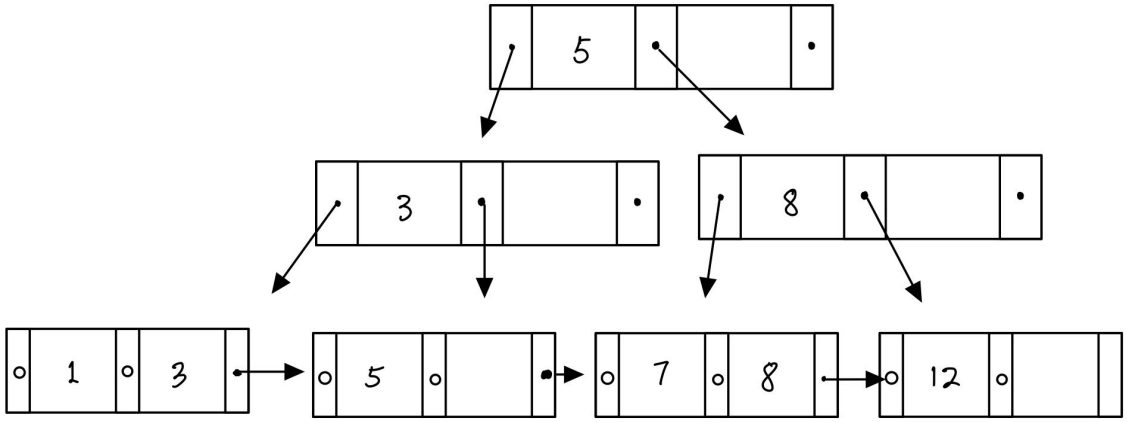
  When Key3 is inserted, the number of Pointers to the leaf node reaches the overflow and the leaf node is split





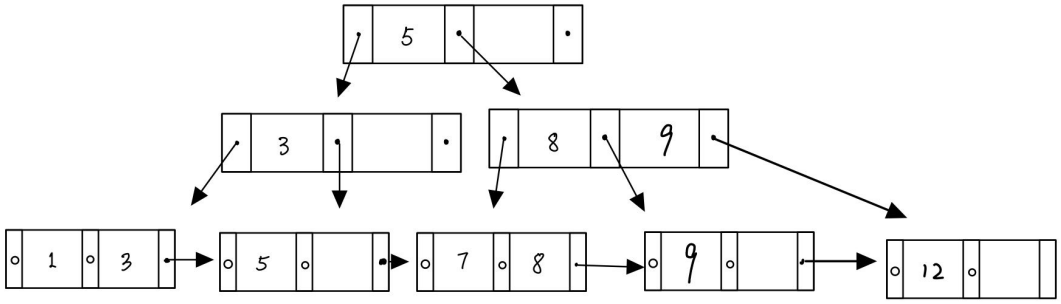
- (ii) Key 12, followed by

  Add a new pointer point to Key12 to the parent node of Key12 , overflow in leaf and non-leaf

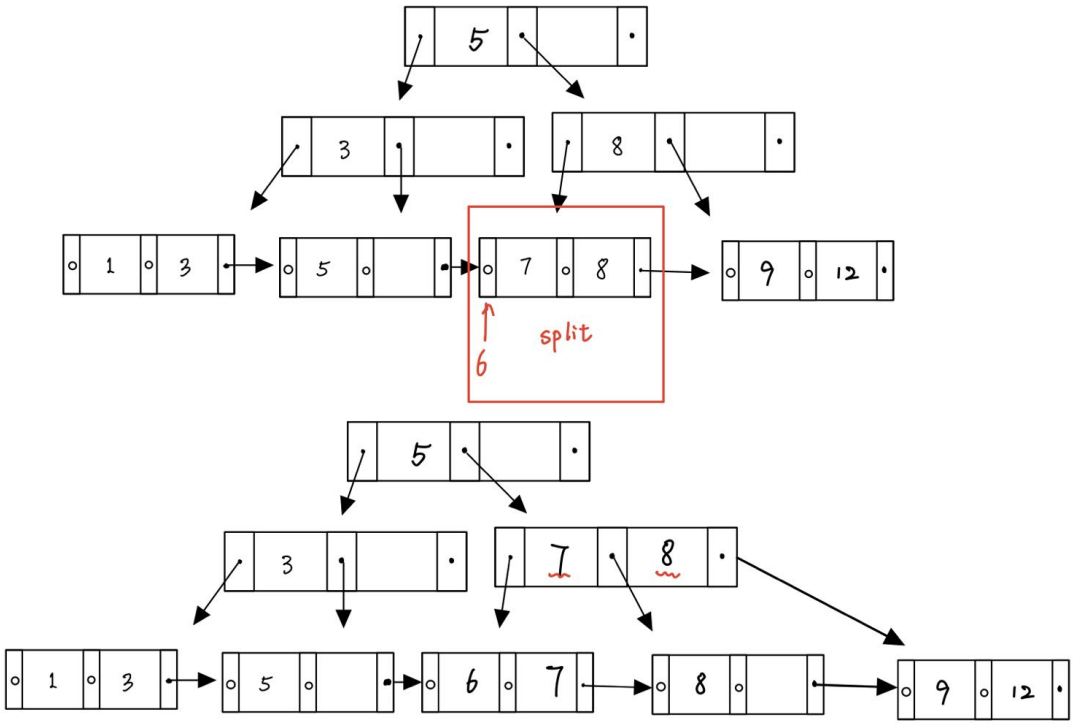


▼ (iii) Key 9, followed by

insert 9 after 8, overflow in leaf

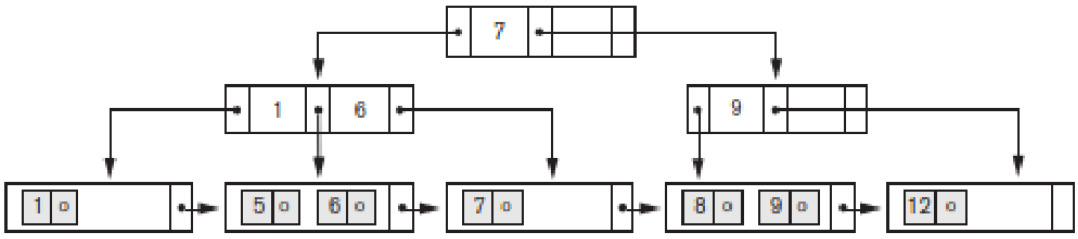


▼ (iv) Key 6



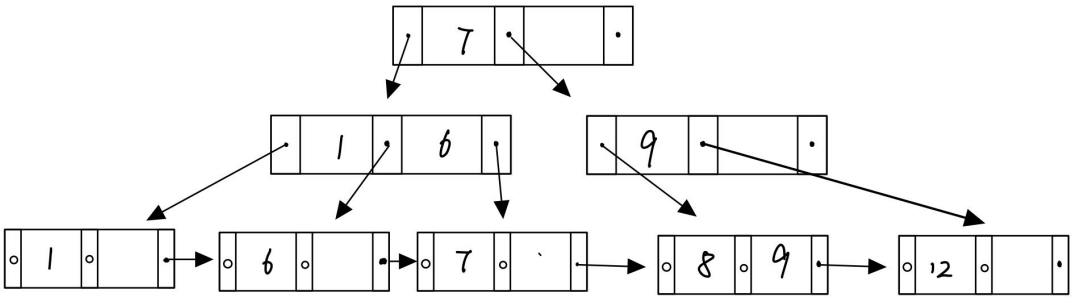
▼ **Question 4**

Consider the following B+-Tree



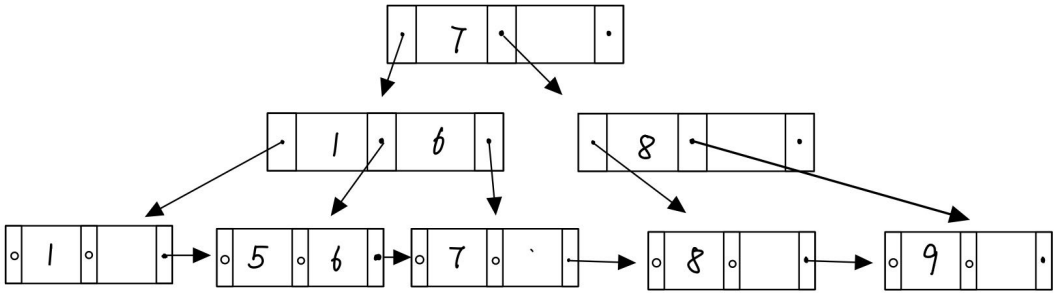
Show each of the tree structures after the deletion of

▼ (i) Key 5, followed by no underflow



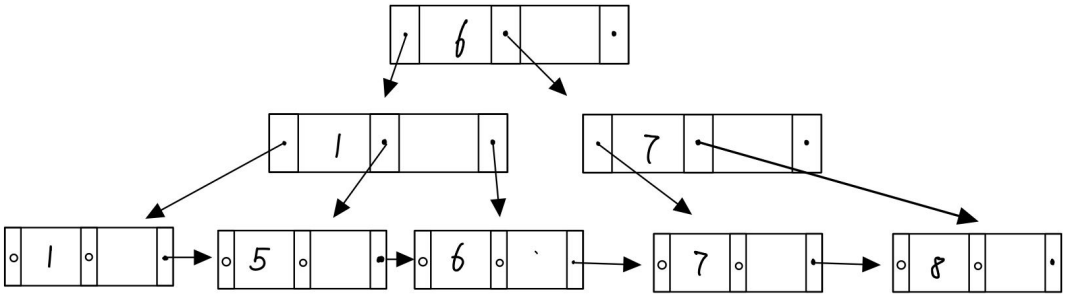
▼ (ii) Key 12, followed by

delete 12 and underflow happen, split leaf node 8,9 and change their father node to 8



▼ (iii) Key 9

underflow happen and move 8 to right leaf and move 7 to right leaf , change the father node to 7 and change root node to 6



▼ **Question 5**

Consider a generalization of the B-tree with minimum fullness factor $f = ¾$.

▼ (i)the mean storage utilization $E(\rho)$

Denote Kas all the search key values inthe tree.
Demote n as the largest capacity of a node
Denote Nas a random number of tree nodes

$$N \in \left[\frac{k}{n}, \frac{k}{nf}\right] \cdot f = \frac{3}{4}$$
$$E(\rho) = E\left(\frac{k}{Nn}\right) = \frac{K}{n} E\left(\frac{1}{N}\right)$$
$$= \frac{K}{n} \times \frac{nf}{K(1-f)} \int_{\frac{k}{n}}^{\frac{k}{nf}} \left(\frac{1}{t}\right) dt = \frac{f}{1-f} \ln \frac{1}{f}$$
$$= 0.8630$$

▼ (ii)the standard deviation of the storage utilization $SD(\rho)$

$$\sigma_f^2 = f - \left(\frac{f}{f'}\right)^2 \left[\ln\left(\frac{1}{f}\right)\right]^2$$

$$\sigma = SD(\rho) \approx 0.07$$

▼ (iii)the probability that the storage utilization lies between 80% and 90%;

$i.e. Prob[0.8 \leq \rho \leq 0.9]$.

$$P(0.8 \leq \rho \leq 0.9)$$
$$= P(\rho \leq 0.9) - P(\rho \leq 0.8)$$
$$= \frac{1}{1-f}(1 - \frac{f}{0 \cdot 9}) - \frac{1}{1-f}\left(1 - \frac{f}{0.8}\right)$$
$$= 0.4167$$

▼ (iv)the median of the storage utilization and compare it with the mean

$$\frac{1}{1-f}\left(1-\frac{f}{x}\right)=\frac{1}{2}$$

$$x=\frac{6}{7}=0.8571$$

It is less than the mean

▼ (v)a general expression of the median for a tree with arbitrary minimum fullness factor f.

$$x=\frac{f}{1-\frac{1}{2}(1-f)}$$

▼ Question 6

Consider a PARTS file with Part# as hash key, which includes records with the following Part# values

▼ (i)

hash value:

h(2305) = 1,h(1168) = 0, h(2580) = **4**,,h(4871) = **7**,h(5659) = 3,h(1821) = 5,

h(1047) = 2, h(7115) = 3, h(1620) = **4**,h(2428) = **4**, h(3943) = **7**, h(4750) = 6,

h(6975) = **7**, h(4981) = 5, h(9208)= 0

average number of block accesses = 17/15 = 1.13

▼ (ii)

binary transform and Hash value:

2305:100100000001, h(2305) = 1:001

1168:10010010000, h(1168) = 0:000

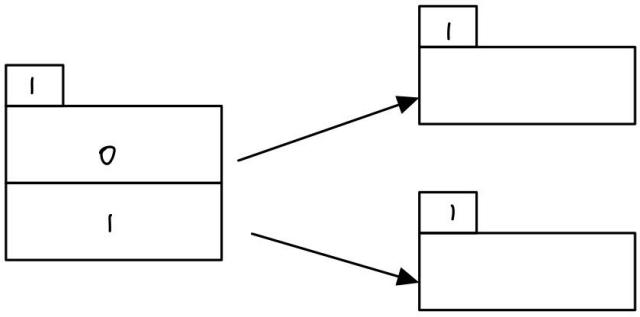
2580:101000010100, h(2580) = 4:100

4871: 1001100000111,h(4871) = 7:111

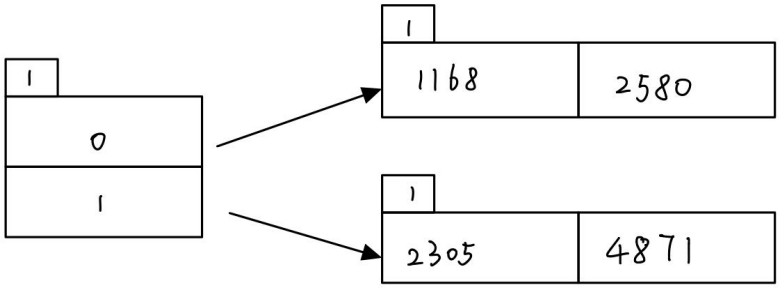
5659: 1011000011011, h(5659) = 3:011

1821: 11100011101, h(1821) = 5:101

Initially, the global-depth and local-depth is always 1. Thus, the hashing frame looks like this:

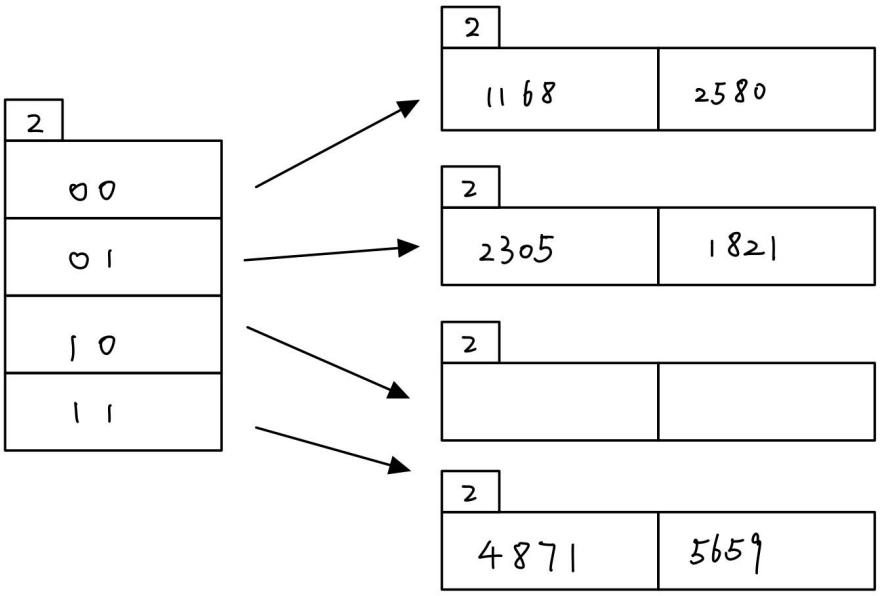


After inserting the first 4 number, we get



Next, inserting 5659 and overflow occurs.

Since Local Depth = Global Depth, the bucket splits and directory expansion takes place.


2
00
01
10
11
2
1168
2580
2
2305
1821
2
2
4871
5659