Yohandi - 120040025

CSC3170 Assignment 3

1(a) $E(X_N) = \sum_{K=1}^{N} K \cdot \frac{C}{K} = \sum_{K=1}^{N} C = CN$

we know that:

$$\frac{C}{1} + \frac{C}{2} + \dots + \frac{C}{N} = 1$$

$$\Rightarrow C = \frac{1}{\sum_{K=1}^{N} \frac{1}{K}}$$

$$\Rightarrow E(X_N) = \frac{N}{\sum_{K=1}^{N} \frac{1}{K}} = \frac{N}{1 + \frac{1}{2} + \dots + \frac{1}{N}}$$

(b) let Y_N be the (random) number of comparisons to locate a given record present in the file of N records where the probability of choosing any records are the same, we have:

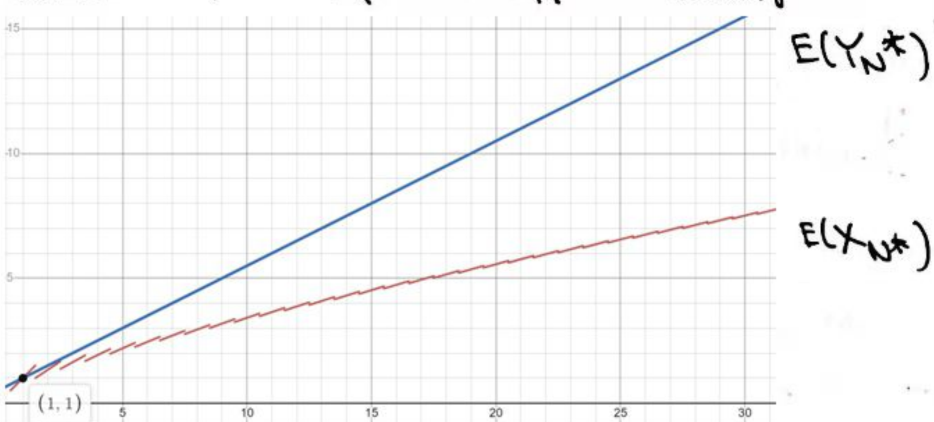
$$E(Y_N) = \sum_{K=1}^{N} K \cdot \frac{1}{N} = \frac{1}{N} \sum_{K=1}^{N} K = \frac{1}{N} \cdot \frac{N(N+1)}{2} = \frac{N+1}{2}$$

when $N=10$,

$$\frac{E(X_N)}{E(Y_N)} = \frac{E(X_{10})}{E(Y_{10})} \approx \frac{3.414}{5.500} \approx \frac{0.621}{1}$$

The Zipf distribution has approximately 62.1% expected comparisons of the uniform distribution's one. This shows about $\frac{1-0.621}{0.621} \approx 61.0\%$ increment in the number comparisons for the Zipf distribution.

(c)



Based on the trend above, we can conclude that the only time where $E(X_{N^*}) = E(Y_{N^*})$ is when $N^* = 1$. For all $N > N^*$, $E(X_N^*) < E(Y_N^*)$; meaning Zipf distribution will always outperform uniform distribution, when $N \neq 1$ and N is a discrete value.

Note that for both (d) and (e), I am using the assumption that it is not possible to do the search by "jumping" over any indexes (Imagine a linked list instead of an array). This to prevent any binary search, ternary search, etc. idea-like

(d) Suppose we have $x_1, x_2, \dots, x_N$ as the sorted version of the N records, then we have:

$$x_1 < x_2 < x_3 < \dots < x_N$$

In here x_i is being taken out with the assumption that the value (suppose it is y) that is taken fulfill: $x_{i-1} < y < x_{i+1}$ for ~~some~~ an i

Then, if we take out x_i for an i, we realize that we need to check for $x_1, x_2, \dots,$ ~~x_i. This implies that~~ x_{i-1}, x_{i+1}. This implies that for an x_i, we require i comparisons.

Average number of comparisons $= \sum_{i=1}^{N} f(x_i) = \sum_{i=1}^{N} i = \frac{N+1}{2} \approx \frac{N}{2}$

define f as a function that denotes the number of comparisons to ~~find x_i in x_i~~ check x_i's existence

(e) (i) Similar to part (d); however, this time we assume that we have:

$$x_{h(1)} < x_{h(2)} < x_{h(3)} < \dots < x_{h(N)}$$

In here, $h(i)$ denotes the index of x_j with rank i in $x_1, x_2, \dots, x_N$.

With the same argument, the average number of comparisons:

$$\underbrace{\sum_{i=1}^{N} f(x_{h(i)}) = \sum_{i=1}^{N} f(x_i)}_{} = \sum_{i=1}^{N} i = \frac{N+1}{2} \approx \frac{N}{2}$$

this is because $h(i) = \{1, 2, \dots, N\}$ and $|h(i)| = N$ and $h(i) \neq h(j)$ when $i \neq j$

(ii) In this case, we don't have the ascending order property. Therefore, for all case, we need N comparisons.

Average number of comparisons $= \sum_{i=1}^{N} N = N$

2. In a B-tree, the average fullness implies $m = n\ln(2)$.
Since $n = 23$, $m = 23\ln(2) \approx 15.94$. We take the fanout as 16, we have:

Level	Number of nodes	Key entries	Children pointers
0	$16^0 = 1$	$1 \times 15 = 15$	$16^1 = 16$
1	$16^1 = 16$	$16 \times 15 = 240$	$\boxed{(i)\ 16^2 = 256}$
2	$16^2 = 256$	$256 \times 15 = 3840$	$16^3 = 4096$
3	$16^3 = 4096$	$4096 \times 15 = 61440$	$\boxed{(ii)\ 16^4 = 65536}$
4	$16^4 = 65536$	$65536 \times 15 = 983040$	$\boxed{(iii)\ 16^5 = 1048576}$

(iv) $15 + 240 + 3840 = 4095$

(v) $15 + 240 + 3840 + 61440 = 65535$

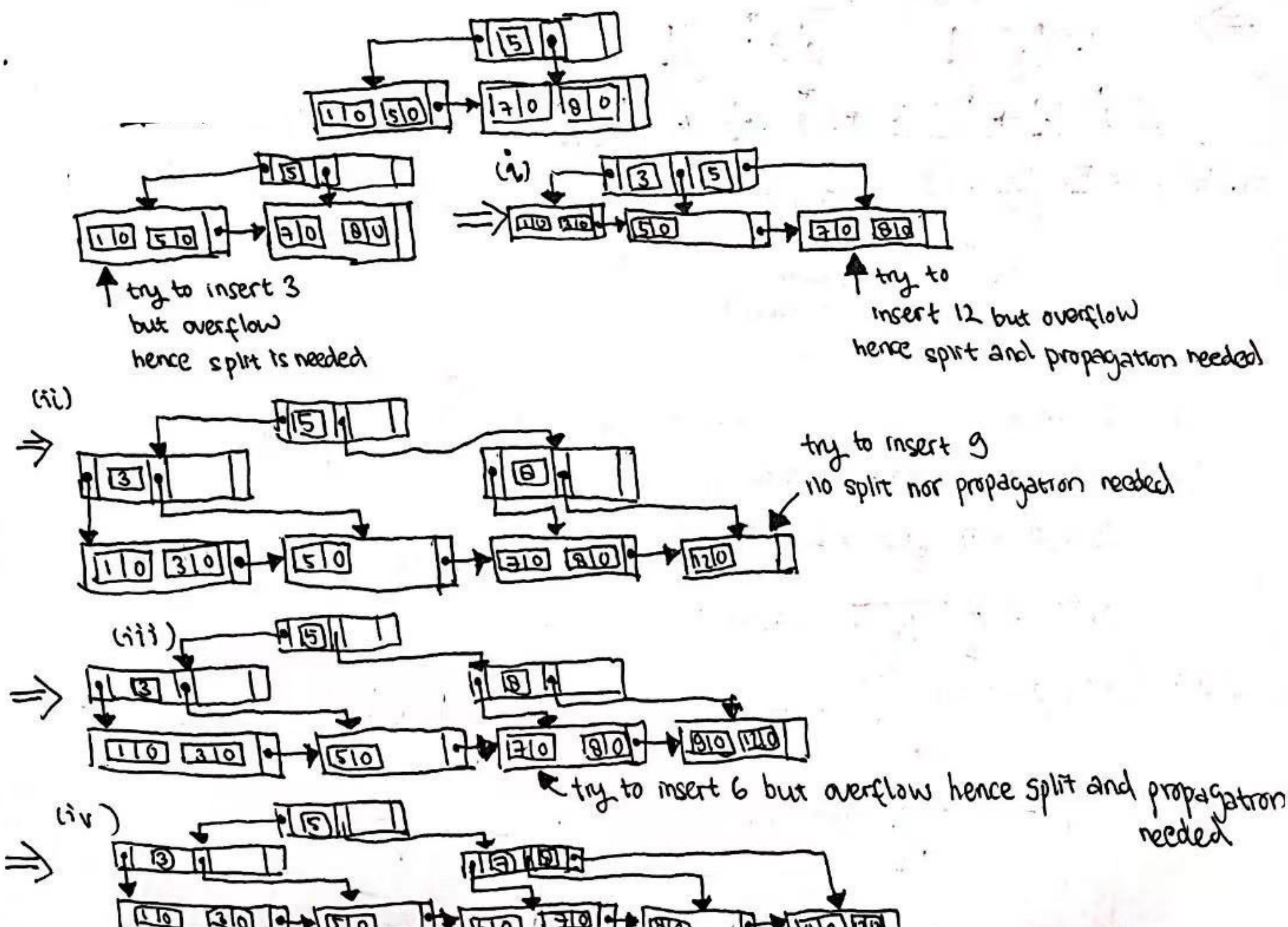
(vi) $15 + 240 + 3840 + 61440 + 983040 = 1048575$

According to the pattern, we notice that at level h, the ~~number~~ average total number of entries is $16^{h+1} - 1$. To prove it with general m, we ~~have~~ have:

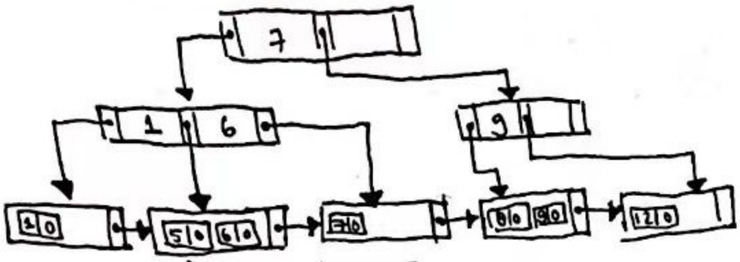
$$\sum_{K=1}^{h} (m^k (m-1)) = (m-1)\sum_{k=1}^{h} m^k = (m-1)\frac{m^{h+1}-1}{m-1} = m^{h+1} - 1$$

as the average total number of entries ~~with~~ at height h.

3.



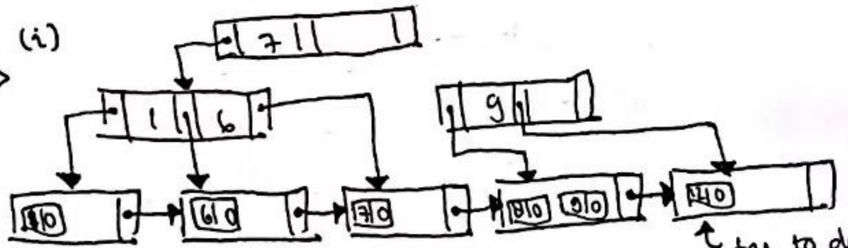
4.



try to delete 5

(i)

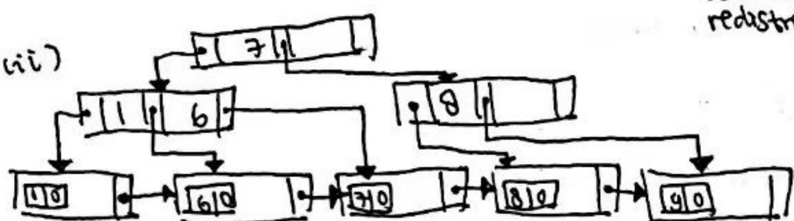
$\Rightarrow$



try to delete 12,
becomes empty (underflow)
redistribution needed

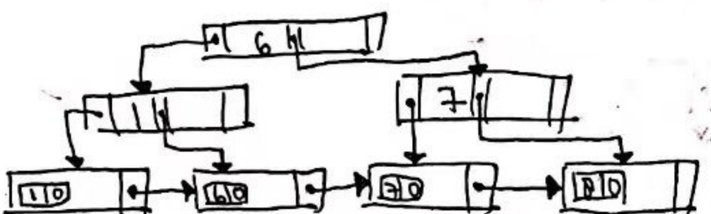
(ii)

$\Rightarrow$



try to delete 9,
becomes empty (underflow)
remerge & redistribution needed

$\Rightarrow$



5 (i) $E(\rho) = E\left(\frac{K}{Nn}\right) = \frac{K}{n}E\left(\frac{1}{N}\right) = \frac{K}{n}\cdot\frac{nf}{Kf'}\int_{\frac{K}{n}}^{\frac{K}{nf}}\left(\frac{1}{t}\right)dt = \frac{f}{f'}\ln\left(\frac{1}{f}\right)$

(here, it is assumed that the distribution uses uniform distribution from $\frac{K}{n}$ to $\frac{K}{nf}$)

since f denotes the fullness factor, we have $E(\rho) = \frac{\left(\frac{3}{4}\right)}{\left(\frac{1}{4}\right)}\ln\left(\frac{1}{\left(\frac{3}{4}\right)}\right) \approx 0.863$

(ii) By the formula derived from slides,

$$Var(\rho) = \sigma_f^2(\rho) = f - \left(\frac{f}{f'}\right)^2\left[\ln\left(\frac{1}{f}\right)\right]^2 \approx \frac{3}{4} - 0.863^2 \approx 0.00523$$

$$\Rightarrow \sigma_f(\rho) = \sqrt{Var(\rho)} \approx 0.0723$$

(iii) $P(0.8 \le \rho \le 0.9) = \int_{0.8}^{0.9} g(x)\,dx = \int_{0.8}^{0.9}\frac{f}{f'}\cdot\frac{1}{x^2}dx = \frac{\left(\frac{3}{4}\right)}{\left(\frac{1}{4}\right)}\left[-\frac{1}{x}\right]\Big|_{x=0.8}^{0.9} = \cancel{0.13} = 0.41\overline{6}$

(iv) $P(f \le \rho \le m) = \int_{f}^{m} g(x)\,dx = \int_{3/4}^{m}\frac{f}{f'}\cdot\frac{1}{x^2}dx = \frac{\left(\frac{3}{4}\right)}{\left(\frac{1}{4}\right)}\left[-\frac{1}{x}\right]\Big|_{x=3/4}^{m} = \frac{1}{2}$

$$\Rightarrow \frac{4}{3} - \frac{1}{m} = \frac{1}{6} \Rightarrow m = \frac{6}{7}$$

(v) $P(f \le p \le m) = \int_f^m \frac{f}{f'} \cdot \frac{1}{x^2}\, dx = \frac{f}{f'}\left[-\frac{1}{x}\right]_{x=f}^{m} = \frac{f}{f'}\left(\frac{1}{f} - \frac{1}{m}\right) = \frac{1}{2}$

$$\Rightarrow m = \frac{2f}{1+f}$$

6(i)

K	h(K) = K mod 8	block access
2305	1	1
1168	4	1
2580	0	1
4871	7	1
5659	3	1
1821	5	1
1074	2	1
7115	3	1
1620	4	1
2428	4	2
3943	7	1
4750	6	1
6975	7	2
4981	5	1
9208	0	1

$E(\text{block access}) = \overline{\text{block access}} = \frac{1 \times 13 + 2 \times 2}{15} = 1.1\overline{3}$

(ii)

K	h(K) = K mod 128	$h(K)_{(2)}$
2305	1	0000001
1168	16	0001000
2580	20	0001010
4871	7	0000111
5659	27	~~0011111~~ 000 0011011
1821	29	0011101