CS 5968/6968: Data Str & Alg Scalable Comp Spring 2023

Lecture : Hashing (Balls and Bins) — February 1, 2023

Prof. Prashant Pandey *Scribe: Manoj Marneni*

1 Overview

In the previous lectures, we discussed VEB, X-Fast and Y-Fast Trees, which improved the worst time complexity of insert, find, and successor query operations from $O(\log n)$ (using Binary search tree) to $O(\log \log U)$. After that, we discussed Tries, Suffix Tree, and Suffix array for pattern search with better space complexity. We also discussed BWM Transform and FM-Index which are used for pattern search with even better space. Overall we discussed storing and finding, Integers and Strings optimally.

In this lecture, we discussed Hashing. we discussed a little about Chaining based and Open addressing based hash tables. It is discussed that we need to study different hashtables since the performance of hash table may differ with collision rate in input data and the distribution of the data. It is also discussed how the performance of Hash table degrades as we insert data into the hashtable. As we can see in figure 1 (black line), usually the throughput of hashtable reduces as we increase the load factor. Load factor is defined the $\frac{x}{n}$, where x is the number of items in the hash table and n is the capacity of the hashtable. We will see one hashtable whose througput would be constant with increase in load factor of the hashtable as shown in the graph with red line, which is desired. We will discuss in this lecture why the performance degrades with load factor.

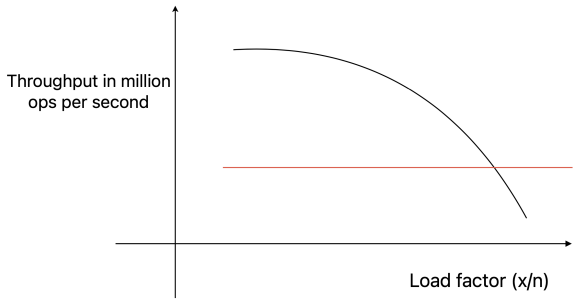


Figure 1: Performance of throughput of hashtable with increase in load factor

2 Balls ans Bins

We introduce "Balls and Bins" game, where we throw b balls equiprobably and independently into n bins (Often b = n). Some applications of this problem are:

1. Hashing

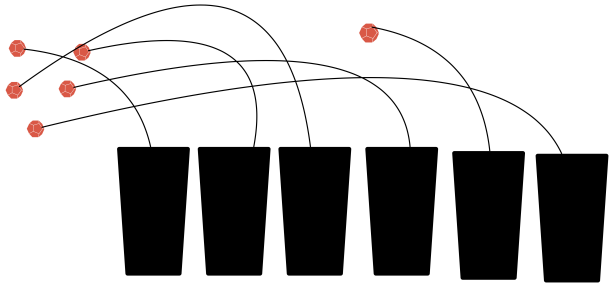


Figure 2: Throwing n balls in b bins

2. Birthday attack
3. Load balancing

2.1 Quesitons:

Some of the questions that we want to answer for this problem are:

1. Expected number of balls in a bin
2. Expected number of balls in the fullest bin
3. expected number of balls that need to be thrown before getting a collision
4. expected number of empty bins
5. expected number of bins with a collision
6. expected number of balls needed to fill all bins

3 Probability Refresher:

Definition 1. *Probability Sample Space (S,P): Let S be the set of outcomes, which is finite or countably infinite.*

$$S = \{S_1, S_2, ...\}$$

Let probability function, $P : S \to [0, 1]$, Where $\sum P(S_i) = 1$

Definition 2. *An Event is a subset of outcomes from the sample space (S,P).*

To solve any probability problem we need to find

1. Find the sample space
2. Define events of interest
3. Determine outcome probability

4. Determine event probability

Definition 3. *A random variable is a function*

$$f : S \to R^+$$

Definition 4. *Expected value:* $E(f) = \sum P(S_i)f(S_i)$

Definition 5. ***Linearity of Expectation:***

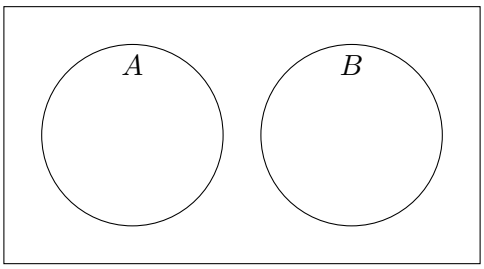
$E(f + g) = E(f) + E(g)$

Definition 6. ***Conditional Probability*** *The notation* $Pr(A|B)$ *denotes the probability of Event A happening, given that event B happens.*

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$

Definition 7. ***Independence*** *Two events A and B are independent if and only if*

$$Pr(A \cap B) = Pr(A) * Pr(B)$$



In the above venn diagram, Events A and B in the sample space are disjoint. They are not independent, since only one of Events A or B can occur at any time.

Example: If a coin is tossed, then event of occurence of head and the event of occurence of tail are not independent since if head occurs then tail cannot occur. They both are disjoint events but not independent events. If we roll 2 dice, The event of the occurence of 1 on dice 1, and the event of the occurence of 2 on dice 2, are independent events.

Definition 8. ***Mutual Independence*** *Events* $E_1, E_2 \dots E_n$ *are mutually independent events if and only if for every subset of the events, the probability of the intersection, is the product of the probabilities of the individual events.*

$$\implies Pr(E_i \cap E_j) = Pr(E_i) * Pr(E_j)$$

for all distinct i, j*.*

$$\implies Pr(E_1 \cap E_2 \cap \dots E_n) = Pr(E_1) * Pr(E_2) * \dots * Pr(E_n)$$

Example Question:
We flip three fair mutually independent coins.
Event A_1: Coin 1 matches Coin 2
Event A_2: Coin 2 matches Coin 3
Event A_3: Coin 3 matches Coin 1

Are A_1, A_2, A_3 mutually independent?

Answer: No, since if the events A_1, A_2 occur then A_3 will definitely occur, similarly if the events A_2, A_3 occur then A_1 will definitely occur, therefore these 3 events are not mutually independent.
Note: These three events are pairwise independent.

Definition 9. ***Pairwise Independent*** *Events $E_1, E_2 \dots E_n$ are Pairwise independent events if and only if for every 2 events, the probability of the intersection is the product of the probabilities of the individual events.*

$$\implies Pr(E_i \cap E_j) = Pr(E_i) * Pr(E_j)$$

for all distinct i, j.

Theorem 1.

$$Pr(A_1 \cup A_2) = Pr(A_1) + Pr(A_2) - Pr(A_1 \cap A_2)$$

Theorem 2. ***Union bound***

$$Pr(A_1 \cup A_2) = Pr(A_1) + Pr(A_2) - Pr(A_1 \cap A_2)$$

$$\implies Pr(A_1 \cup A_2) \leq Pr(A_1) + Pr(A_2)$$

Lemma 1. *Expected number of balls in a bin is equal to 1.*

Proof. Let X be the random variable of the number of balls in bin 1, if

$$x_i = \begin{cases} 1 & \text{if ball } i \text{ lands in bin 1} \\ 0 & \text{if ball } i \text{ lands in another bin.} \end{cases}$$

$$X = x_1 + x_2 + \dots x_n$$

$$E(x_i) = 1 * \frac{1}{n} + 0 * (1 - \frac{1}{n}) = \frac{1}{n}$$

Using the linearity of expectation,

$$E(X) = E(x_1) + E(x_2) + \dots + E(x_n)$$

$$E(X) = \frac{1}{n} + \frac{1}{n} + \dots + \frac{1}{n} = 1$$

Therefore the expected number of balls in a bin is 1. ■

Definition 10. *Let E_n be an event of problem size n. We say E_n occurs "WITH HIGH PROBABILITY (WHP)" if*

$$Pr(E_n) \geq 1 - \frac{1}{n^c}$$

, for some constant c.

Theorem 3. ***Death Bed Formula (D.B.F)***

$$(\frac{y}{x})^x \leq \binom{y}{x} \leq (\frac{ey}{x})^x$$

Lemma 2. *Expected number of balls in the fullest bin is $O(\frac{\lg n}{\lg \lg n})$ balls with high probability.*

Proof. we start by calculating the probability of bin 1 having l balls.

$$Pr(\text{bin 1 has } l \text{ balls }) = \binom{n}{l}(\frac{1}{n})^l(1 - \frac{1}{n})^{n-l}$$

Now the probability that bin 1 has more than l balls is,

$$Pr(\text{bin 1} \geq l \text{ balls }) \leq \binom{n}{l}(\frac{1}{n})^l$$

$$\implies Pr(\text{bin 1} \geq l \text{ balls }) \leq (\frac{en}{l})^l(\frac{1}{n})^l$$
$$\implies Pr(\text{bin 1} \geq l \text{ balls }) \leq (\frac{e}{l})^l$$

Lets say,

$$l = c * \lg n$$

$$\implies Pr(\text{any bin} \geq c * \lg n \text{ balls }) \leq n * (\frac{e}{c * \lg n})^{c*\lg n}$$

$$\implies Pr(\text{any bin} \geq c * \lg n \text{ balls }) \leq n * (\frac{1}{2})^{c*\lg n}$$

$$\implies Pr(\text{any bin} \geq c * \lg n \text{ balls }) \leq n * n^{-c}$$
$$\implies Pr(\text{any bin} \geq c * \lg n \text{ balls }) \leq n^{1-c}$$

This is not a tighter bound, If we say $l = \frac{c \lg n}{\lg \lg n}$

$$\implies Pr(\text{any bin} \geq \frac{c \lg n}{\lg \lg n} \text{ balls }) \leq n * (\frac{1}{2})^{c*\lg n - O(c*\lg n)}$$

$$\implies Pr(\text{any bin} \geq \frac{c \lg n}{\lg \lg n} \text{ balls }) \leq n^{2-c}$$

Therefore we can say, the fullest bin in the hash table would $\frac{c \lg n}{\lg \lg n}$ with high probability. ■