Distributional Aspects of Hashing

Mark E. Lehr Ph.D.

May 2021

1 The Distribution of Collisions

A continuation of determining the efficacy related to hashing, this derivation sets to determine the maximum expected number of collisions that would be experienced in distributed information.

Let's set the stage with an interesting problem. Suppose a data center needs a procedure to spread traffic across available servers. This requires some kind of a traffic management system. The ideal system would look at each servers work load to spread the incoming traffic equally amongst all servers. Of course, this requires communication with each server relating their current load, etc..... We need to compare this to a baseline arrangement with a management system that requires no networking communication.

Let's begin the process by describing an array of servers X. We would like to establish a queue that randomly assigns each work segment to one of the servers without regard to any existing workload on that server. Given that each server can handle k jobs, let us see how to calculate the probability of exceeding that number k. So, we start with defining the probability:

$$P(X < k) = P(\bigcup_{i=1}^{N}(X_i < k))$$

However to solve for the upper bound it will be easier to use the compliment:

$$P(X \geq k) = P(\bigcup_{i=1}^{N}(X_i \geq k))$$

The approach will be to calculate a maximum upper bound so we can use the Probability Union Upper Bound approximation:

$$P(\bigcup_{i=1}^{N}(X_i \geq k)) \leq \sum_{i=1}^{N} P(X_i \geq k)$$

Next, let's choose a value like a half and determine an absolute maximum of the individual component:

$$P(\bigcup_{i=1}^{N}(X_i \geq k)) = \frac{1}{2} \leq \sum_{i=1}^{N} P(X_i \geq k)$$

$$\frac{1}{2} \leq N * (P(X_i \geq k))$$

$$\frac{1}{2} \leq N * (\frac{1}{2N})$$

Now we have a bound on an individual element in the Array X of N

$$\frac{1}{2N} \geq P(X_i \geq k)$$

Using a biased coin analogy and the binomial distribution given N hashes

$$P(X_i \geq k) = \sum_{j=k}^{N} \binom{N}{j} P^j (1-P)^{N-j}$$

where the probability of any position being hashed is uniform $P = \frac{1}{N}$, therefore:

$$P(X_i \geq k) = \sum_{j=k}^{N} \binom{N}{j} (\frac{1}{N})^j (1-\frac{1}{N})^{N-j}$$

Now we need to simplify and bound this term by letting

$$e^{-\frac{1}{N}} \approx (1-\frac{1}{N})$$

$$(\frac{N}{j})^j \leq \binom{N}{j} \leq \frac{N^j}{j!} \leq (\frac{Ne}{j})^j$$

Apply the approximation and the upper bound of the binomial coefficient:

$$P(X_i \geq k) \leq \sum_{j=k}^{N} \frac{N^j}{j!} (\frac{1}{N})^j e^{-\frac{N-j}{N}}$$

$$\leq e^{-1} \sum_{j=k}^{N} \frac{e^{\frac{j}{N}}}{j!}$$

which we can factor out the first term

$$\leq \frac{e^{-1} e^{\frac{k}{N}}}{k!}(1 + \frac{e^{1/N}}{(k+1)} + \frac{e^{2/N}}{(k+1)(k+2)} + \frac{e^{3/N}}{(k+1)(k+2)(k+3)} + \ldots)$$

Realizing $1/2 > 1/(k+1)$ and $1/4 > 1/(k+1)/(k+2)$ and $1/8 > 1/(k+1)/(k+2)/(k+3), etc....all < 2$

$$\leq \frac{2}{k! e^{e^{\frac{-k}{N}}}}$$

$$\leq \frac{1}{k!}$$

So, Let $2N$ be the inverse of $k!$ implying a maximum value of k

$$\frac{1}{2N} \leq \frac{1}{k!}$$

max collisons

$$k \leq (!^{-1})(2N)$$

or using Stirling's approximation

$$k! \approx (\frac{k}{e})^k \sqrt{2\pi k} \leq 2N$$

and in terms of the distribution we can use the approximation from above for $N \gg k \wedge j$

$$P(X_i = j) \approx \frac{e^{-1}}{j!}$$

and the expected value of the number of hits to find a value in the hashed array that may or may not be there is

$$E[Collisions] \approx 1 + e^{-1} \approx 1.36$$

but if it is known to be there then

$$E[Collisions] \approx \frac{1}{1 - e^{-1}} \approx 1.58$$

which leads to the following conclusion that

$$1.36 \approx 1 + e^{-1} \leq E[Collisions] \leq \frac{1}{1 - e^{-1}} \approx 1.58$$

$$E[Collisions] = O(1)$$