Analysis of Hashing

COMPSCI 230 — Discrete Math

April 21, 2015
Summary of Hashing Model

• Simple experiment \( B_k = h(\text{key}_k) \in \mathbb{S}_1 = \{0, \ldots, b - 1\} \)
• \( |\mathbb{S}_1| = b \)

• Repeated experiment
  • Either \( B = (B_1, \ldots, B_n) \in \mathbb{S} = \mathbb{S}_1^n \)
  • \( |\mathbb{S}| = b^n \)
  • or \( B = (B_1, B_2, \ldots) \in \mathbb{S} = \mathbb{S}_1^\infty \)
  • \( |\mathbb{S}| = c \)

• Probability model
  • Uniform: \( P(B_k = b_k) = \frac{1}{b} \)
  • Independent: If \( j \neq k \) then
    \[
P(B_j = b_j \ and \ B_k = b_k) = P(B_j = b_j)P(B_k = b_k)
    \]
  • \( P(B) = \frac{1}{b^n} \) for any \( n \)-sequence or \( n \)-subsequence
Probability of Some Collision

- $D(n)$: All $n$ items hash to different buckets
- $P(D(n)) = 0$ for $n > b$
- For $n \leq b$, there are $b$ buckets for first item, $b - 1$ for second
- ... $b - n + 1$ for $n$-th
- $b(b-1) \cdots (b-n+1) = (b)_n = \frac{b!}{(b-n)!}$
- $P(D(n)) = \frac{(b)_n}{b^n}$ for $n \leq b$ and zero otherwise
- $S(n)$: Some collision
- $P(S(n)) = 1 - P(D(n))$ for $n \leq b$ and 1 otherwise
- $b = 10$:

<table>
<thead>
<tr>
<th>$n$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(S(n))$</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.28</td>
<td>0.496</td>
<td><strong>0.70</strong></td>
<td>0.85</td>
<td>0.94</td>
<td>0.98</td>
<td>0.996</td>
<td>0.9996</td>
<td>1</td>
</tr>
</tbody>
</table>
\[ P(\text{collision}) \geq 1/2 \]

- With \( b = 10 \) buckets, it takes \( n = 5 \) items to bring the probability that some collision occurs to at least 1/2
- What about other values of \( b \)? (Table computed numerically)

<table>
<thead>
<tr>
<th>( b )</th>
<th>( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>2</td>
</tr>
<tr>
<td>3-5</td>
<td>3</td>
</tr>
<tr>
<td>6-9</td>
<td>4</td>
</tr>
<tr>
<td>10-16</td>
<td>5</td>
</tr>
<tr>
<td>17-23</td>
<td>6</td>
</tr>
<tr>
<td>24-32</td>
<td>7</td>
</tr>
<tr>
<td>33-42</td>
<td>8</td>
</tr>
<tr>
<td>43-54</td>
<td>9</td>
</tr>
<tr>
<td>55-68</td>
<td>10</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>100</td>
<td>13</td>
</tr>
<tr>
<td>1000</td>
<td>38</td>
</tr>
<tr>
<td>10000</td>
<td>119</td>
</tr>
</tbody>
</table>
Items per Bucket

• Fairly intuitive that with \( n \) items in \( b \) buckets the expected number of items per bucket is \( n/b \)
• Good exercise to prove it. W.l.o.g., consider bucket 0
• Define the sequence of random variables \( X_1, \ldots, X_n \) on the repeated experiment \( B \in \mathcal{S}_1^n \) where

\[
X_i(B) = \begin{cases} 
1 & \text{if the } i\text{-th bucket index is 0} \\
0 & \text{otherwise}
\end{cases}
\]

• \( X_i(B) \) is an event indicator variable
• All buckets equally likely \( \rightarrow p = P(X_i = 1) = 1/b \)
• \( \mathbb{E}[X_i] = p \cdot 1 + (1 - p) \cdot 0 = p = 1/b \)
• Number of items in bucket 0: \( N(B) = \sum_{i=1}^{n} X_i \)
• The sum of an indicator variable is an event count
• From linearity of expectation,

\[
\mathbb{E}[N] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{1}{b} = \frac{n}{b}
\]
Probability that a Given Bucket is Empty

• Probability that item $i$ hashes to bucket 0 is
  \[ p = P(X_i = 1) = 1/b \]
• Probability that it does not is
  \[ q = P(X_i = 0) = 1 - p = (b - 1)/b \]
• Probability that none of the $n$ items hashes to bucket 0 is $q^n$ (because of independence)
• No bucket is special, so probability that any one bucket is empty is

\[ p_e = q^n = \left(\frac{b - 1}{b}\right)^n \]
Expected Number of Empty Buckets

• Define the random variables $E_0, \ldots, E_{b-1}$ where

$$E_j(B) = \begin{cases} 
1 & \text{if bucket } j \text{ is empty after } n \text{ insertions} \\
0 & \text{otherwise}
\end{cases}$$

• $p_e = P(E_j = 1) = \left(\frac{b-1}{b}\right)^n$

• $\mathbb{E}[E_j] = p_e \cdot 1 + (1 - p_e) \cdot 0 = p_e$

• Number of empty buckets: $M(B) = \sum_{j=0}^{b-1} E_j$

• From linearity of expectation,

$$\mathbb{E}[M] = \mathbb{E} \left[ \sum_{j=0}^{b-1} E_j \right] = \sum_{j=0}^{b-1} \mathbb{E}[E_j] = \sum_{j=0}^{b-1} \left(\frac{b-1}{b}\right)^n = \frac{(b-1)^n}{b^{n-1}}$$
A Useful Approximation

- Expected number of empty buckets

\[ \mathbb{E}[M] = \frac{(b - 1)^n}{b^{n-1}} \]

- Would have to plot for many combinations of \( b \) and \( n \)
- More useful approach: What if \( b \) is large?

\[
\frac{(b - 1)^n}{b^{n-1}} = b \left( \frac{b - 1}{b} \right)^n = b \left[ \left( \frac{b - 1}{b} \right)^b \right] \frac{n}{b}
\]

- You may recall from calculus that

\[
\lim_{b \to \infty} \left( \frac{b - 1}{b} \right)^b = \frac{1}{e}
\]

so

\[ \mathbb{E}[M] \approx be^{-\frac{n}{b}} \]

<table>
<thead>
<tr>
<th>( b )</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>500</th>
<th>1000</th>
<th>( \frac{1}{e} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \left( \frac{b - 1}{b} \right)^b )</td>
<td>0</td>
<td>0.3277</td>
<td>0.3487</td>
<td>0.3642</td>
<td>0.3660</td>
<td>0.3675</td>
<td>0.3677</td>
<td>0.3679</td>
</tr>
</tbody>
</table>
Approximate Expected **Fraction** of Empty Buckets

\[ \mathbb{E}[M] \approx be^{-\frac{n}{b}} \quad \Rightarrow \quad \frac{\mathbb{E}[M]}{b} = f\left(\frac{n}{b}\right) \approx e^{-\frac{n}{b}} \]

- Get away with a single plot: Fraction of empty buckets versus number of insertions per bucket

![Graph showing the fraction of empty buckets as a function of \( \frac{n}{b} \)]

- Example: \( b = 100, n = 50 \) yields \( \mathbb{E}[M] \approx 60.6 \) empty buckets
Careful with Averages!

- With \( n \) items in \( b \) buckets, there are on average \( n/b \) items per bucket
- If \( n = b \), there is on average one item per bucket
- Should we then expect each bucket to have an item?
- Cannot be, because we expect \( 100e^{-1} \approx 37 \) percent of buckets to be empty (\( n/b = 1 \))
- 1 item per bucket on average does not mean 1 item in every bucket
Expected Number of Collisions

- How many collisions do we expect when inserting \( n \) items into \( b \) buckets?
- We have \( \mathbb{E}[M] = b \left( \left( \frac{b-1}{b} \right)^b \right)^{\frac{n}{b}} \approx be^{-\frac{n}{b}} \) empty buckets on average
- So the remaining \( b - \mathbb{E}[M] \) buckets are occupied
- Only the first item inserted into each of the occupied buckets did not collide
- The remaining \( n - (b - \mathbb{E}[M]) \) ones did
- If \( C(B) \) is the number of collisions with bucket-index \( n \)-sequence \( B \), then

\[
\mathbb{E}[C] = n - (b - \mathbb{E}[M]) = n - b + b \left( \left( \frac{b-1}{b} \right)^b \right)^{\frac{n}{b}} \approx n + b(e^{-\frac{n}{b}} - 1)
\]
Approximate Expected Number of Collisions per Bucket

- Dividing by $b$ and plotting versus $n/b$ yields a universal (approximate) graph again:

\[
\frac{\mathbb{E}[C]}{b} \approx \frac{n}{b} + e^{-\frac{n}{b}} - 1
\]

\[
\approx \frac{n}{b} - 1 \quad \text{when } n \gg b
\]

- Example: $b = 100$, $n = 100$ yields $E[M] \approx 36.8$ empty buckets and the same number of collisions
- $b = 100$, $n = 300$ yields about 5 empty buckets and 205 collisions
- For many more insertions than buckets, $\mathbb{E}[C]$ is close to linear (slope 1, as many collisions as there are insertions)
New-Occupancy Time

• How many insertions are needed on average until every bucket has at least one item?
• Answer may be useful when designing heuristics to increase $b$ as $n$ grows
• We cannot bound the number of insertions
• Need to refer to $\mathbb{S}_1^\infty$ as the underlying sample space
• Let $T_i$ be the number of insertions needed to go from $i - 1$ to $i$ occupied buckets
• “$i$-th new-occupancy time” for $i = 1, \ldots, b$
• Example 1: $T_1 = 1$, because the first insertion leads from 0 to 1 occupied bucket
• $\mathbb{E}[T_1] = 1$
Expected New-Occupancy Time

• Let $T_i$ be the number of insertions needed to go from $i - 1$ to $i$ occupied buckets
• Example 1: $T_1 = 1$ and $\mathbb{E}[T_1] = 1$
• Example 2: $T_2$: If exactly one bucket is occupied, subsequent items fall into the same bucket ("failure") with probability $q_2 = 1/b$, and in a different bucket ("success") with probability $p_2 = 1 - q_2 = (b - 1)/b$
• $T_2 = k$ if there are $k - 1$ failures followed by one success: Bernoulli trial
• So $P(T_2 = k) = q_2^{k-1}p_2$, and $\mathbb{E}[T_2] = 1/p_2$ (from previous theory)
• In general, if $i - 1$ buckets are occupied
• Probability of "failure" $q_i = (i - 1)/b$ (fall into occupied bucket)
• Probability of "success" $p_i = 1 - q_i = \frac{b - i + 1}{b}$
• Expected $i$-th new-occupancy time

$$\mathbb{E}[T_i] = \frac{1}{p_i} = \frac{b}{b - i + 1}$$
Expected Table Occupancy Time

- To occupy all buckets we need to occupy one, then a second, then a third,...
- Takes time (in number of insertions) \( T = \sum_{i=1}^{b} T_i \)
- Expected table occupancy time:

\[
\mathbb{E}[T] = \mathbb{E} \left[ \sum_{i=1}^{b} T_i \right] = \sum_{i=1}^{b} \mathbb{E}[T_i]
\]

\[
= \sum_{i=1}^{b} \frac{b}{b-i+1}
\]

\[
= \sum_{\ell=b-1+1}^{b} \frac{1}{\ell} = \sum_{\ell=1}^{b} \frac{1}{\ell}
\]

\[
= b \sum_{\ell=1}^{b} \frac{1}{\ell}
\]
Harmonic Numbers $H(\ell) = \sum_{\ell=1}^{b} \frac{1}{\ell}$

$$\int_{1}^{b} \frac{1}{x} \, dx < \sum_{\ell=1}^{b-1} \frac{1}{\ell} = \sum_{\ell=1}^{b} \frac{1}{\ell} - \frac{1}{b}$$

$$\sum_{\ell=2}^{b} \frac{1}{\ell} = \sum_{\ell=1}^{b} \frac{1}{\ell} - 1 = \int_{1}^{b} \frac{1}{x} \, dx$$

$$\ln(b) + \frac{1}{b} < \sum_{\ell=1}^{b} \frac{1}{\ell} < \ln(b) + 1 \quad \text{for} \quad b \geq 2$$

$$\sum_{\ell=1}^{b} \frac{1}{\ell} \approx \ln(b) + \frac{1}{2} \quad \text{for large } b$$
Approximate Expected Table Occupancy Time

- On average, all $b$ buckets of a hash table are occupied after
  \[ \mathbb{E}[T] \approx b \left( \ln(b) + \frac{1}{2} \right) \] insertions

- It takes many more insertions than buckets to fill all the buckets
Summary

• On average, with \( b \) buckets (\( b \) large) and \( n \) insertions
• \( n/b \) items per bucket
• \( \approx b e^{-\frac{n}{b}} \) empty buckets (decays exponentially with \( n \))
• \( n \ll b \) insertions are enough to bring the probability of some collision above 1/2
• \( \approx \frac{n}{b} - 1 \) collisions
• All buckets occupied after \( \approx b \left( \ln(b) + \frac{1}{2} \right) \) insertions
• Analysis also shows small-\( b \) statistics
• Stats are useful for designing hash tables given performance and cost specs
• Assumes a good hash function that yields quasi-random bucket indices
• Good exercise on probabilities, too!