Hash Adaptive Bloom Filter: Technical Report

I. APPENDIX

A. Proof on Lemma 4.1 and Lemma 4.2

Lemma 4.1 $\forall u \in V, p \in H_0, 0 \leq p(u) \leq 1$, have the following relation:

$$\prod_{p \in H_0} (1 - p(u)) \geq 1 - \sum_{p \in H_0} p(u). \tag{20}$$

Proof: Let p_i be the distribution of the hash function h_i, then Equation (20) can be expressed as:

$$\prod_{i=0}^{k} (1 - p_i(u)) \geq 1 - \sum_{i=0}^{k} p_i(u). \tag{21}$$

We denote Equation (21) as Ψ. Next we use mathematical induction to prove Ψ, obviously it holds when $k = 0$, we assume that Ψ holds when $k = \alpha - 1$, then we have $\prod_{i=0}^{\alpha-1}(1 - p_i(u)) \geq 1 - \sum_{i=0}^{\alpha-1} p_i(u)$ and we can get:

$$\begin{aligned}
\prod_{i=0}^{\alpha}(1 - p_i(u)) &= (1 - p_\alpha(u)) \prod_{i=0}^{\alpha-1}(1 - p_i(u)) \\
&= \prod_{i=0}^{\alpha-1}(1 - p_i(u)) - p_\alpha(u) \prod_{i=0}^{\alpha-1}(1 - p_i(u)) \\
&\geq 1 - \sum_{i=0}^{\alpha-1} p_i(u) - p_\alpha(u) \prod_{i=0}^{\alpha-1}(1 - p_i(u)) \\
&\geq 1 - \sum_{i=0}^{\alpha} p_i(u).
\end{aligned} \tag{22}$$

Therefore, Ψ holds when $k = \alpha$, this completes the proof. ■

Lemma 4.2 $\forall 0 \leq x \leq 1$, Function $f(x) = \frac{|S| \cdot x}{\frac{1}{(1-x)^{|S|}} - 1}$ is a convex function.

Proof: We rewrite the $f(x)$ as follows:

$$f(x) = \frac{|S| \cdot x(1-x)^{|S|}}{1 - (1-x)^{|S|}} = \frac{|S| \cdot (1-x)^{|S|}}{\sum_{i=0}^{|S|-1} (1-x)^i} \tag{23}$$

Let $\mu = 1 - x$ and $\theta = |S|$, so $f(\mu) = \frac{\theta \mu^\theta}{\sum_{i=0}^{\theta-1} \mu^i}$, and we can derive $f'(\mu)$ as follows:

$$f'(\mu) = \theta \frac{\sum_{i=\theta-1}^{2\theta-2} (2\theta - 1 - i)\mu^i}{(\sum_{i=0}^{\theta-1} \mu^i)^2} > 0 \tag{24}$$

Since $f'(x) = \frac{\delta f(\mu)}{\delta \mu} \frac{\delta \mu}{\delta x} = -f'(\mu) < 0$, then we can derive $f''(\mu)$ as follows:

$$\begin{aligned}
f''(\mu) = & \frac{\theta}{(\sum_{i=0}^{\theta-1} \mu^i)^4} ((\sum_{i=0}^{\theta-1} \mu^i)^2 \sum_{i=\theta-1}^{2\theta-2} i(2\theta - 1 - i)\mu^{i-1} \\
& - 2 \sum_{i=0}^{\theta-1} \mu^i \sum_{i=1}^{\theta-1} i\mu^{i-1} \sum_{i=\theta-1}^{2\theta-2} (2\theta - 1 - i)\mu^i) \\
= & \frac{\theta}{(\sum_{i=0}^{\theta-1} \mu^i)^3} (\sum_{i=0}^{\theta-1} \mu^i \sum_{i=\theta-1}^{2\theta-2} i(2\theta - 1 - i)\mu^{i-1} \\
& - 2 \sum_{i=0}^{\theta-1} i\mu^i \sum_{i=\theta-1}^{2\theta-2} (2\theta - 1 - i)\mu^{i-1})
\end{aligned} \tag{25}$$

Next, we compare $\sum_{i=0}^{\theta-1} i\mu^i$ with $\frac{\theta-1}{2} \sum_{i=0}^{\theta-1} \mu^i$,

$$\begin{aligned}
& \frac{\theta - 1}{2} \sum_{i=0}^{\theta-1} \mu^i - \sum_{i=0}^{\theta-1} i\mu^i \\
= & \sum_{i=0}^{\theta-1} (\frac{\theta - 1}{2} - i)\mu^i \\
= & \sum_{i=0}^{\frac{\theta-1}{2}} (\frac{\theta - 1}{2} - i)(\mu^i - \mu^{\theta-1-i})
\end{aligned} \tag{26}$$

Since $0 \leq \mu \leq 1$, we have $\sum_{i=0}^{\theta-1} i\mu^i < \frac{\theta-1}{2} \sum_{i=0}^{\theta-1} \mu^i$. According to Equation (25), we have:

$$\begin{aligned}
f''(\mu) > & \frac{\theta}{(\sum_{i=0}^{\theta-1} \mu^i)^3} (\sum_{i=0}^{\theta-1} \mu^i \sum_{i=\theta-1}^{2\theta-2} i(2\theta - 1 - i)\mu^{i-1} \\
& - (\theta - 1) \sum_{i=0}^{\theta-1} \mu^i \sum_{i=\theta-1}^{2\theta-2} (2\theta - 1 - i)\mu^{i-1}) \\
= & \frac{\theta}{(\sum_{i=0}^{\theta-1} \mu^i)^2} \sum_{i=\theta-1}^{2\theta-2} (i - (\theta - 1))(2\theta - 1 - i)\mu^{i-1}
\end{aligned} \tag{27}$$

Therefore, $f''(\mu) > 0$ and $f''(x) = \frac{\delta^2 f(\mu)}{\delta^2 \mu} (\frac{\delta \mu}{\delta x})^2 + \frac{\delta f(\mu)}{\delta \mu} (\frac{\delta^2 \mu}{\delta^2 x}) = f''(\mu) > 0$. Since $f'(x) < 0, f''(x) > 0$, $f(x)$ is a convex function.

■

B. Analysis of P'_c

To simplify the analysis, we assume that each bit in Bloom filter is set to 0 with probability p_0 and 1 with probability $1-p_0$. Note that we do not consider the case of *cost exchange*. When all buckets mapped by e_{sk} through all hash functions in H_c are *conflict after adjustment*, we cannot adjust the hash functions of e_{sk}, so we get

$$P'_c = 1 - \prod_{h\in H_c(e_{sk})} (1-(1-p_0^{k-1})^{\chi(h(e_{sk}))}), \tag{28}$$

where $\chi(i)$ represents the number of keys in the i^{th} bucket of Γ. Moreover, according to average value inequality, we have

$$\begin{aligned} 1-P'_c \leq & (\frac{|H|-k-\sum\limits_{h\in H_c}(1-p_0^{k-1})^{\chi(h(e_{sk}))}}{|H|-k})^{|H|-k} \\ \leq & (1-\frac{1}{|H|-k}\sum_{h\in H_c}(1-p_0^{k-1})^{\chi(h(e_{sk}))})^{|H|-k} \\ \leq & (1-\prod_{h\in H_c}(1-p_0^{k-1})^{\frac{\chi(h(e_{sk}))}{|H|-k}})^{|H|-k}. \end{aligned} \tag{29}$$

It is easy to prove that: $\forall 0<\alpha<1, \beta\in\mathbb{N}, (1-\alpha)^\beta < 1-\alpha^\beta$, which is similar to Lemma 4.1, then we have

$$\begin{aligned} 1-P'_c < & 1-(1-p_0^{k-1})^{\sum\limits_{h\in H_c}\chi(h(e_{sk}))} \\ P'_c > & (1-p_0^{k-1})^{\sum\limits_{h\in H_c}\chi(h(e_{sk}))}. \end{aligned} \tag{30}$$

Since function $g''(x) = (1-p_0^{k-1})^x$ is a convex function, by the Jensen inequality, we get

$$E(P'_c) > (1-p_0^{k-1})^{E(\sum\limits_{h\in H_c}\chi(h(e_{sk})))}. \tag{31}$$

Let $\psi = \sum\limits_{h\in H_c}\chi(h(e_{sk}))$, and we assume that $\forall h \in H, e_{sk}\in S$, for a certain unit u in V, the probability that u is mapped by e_{sk} through h is only determined by $p(u)$, so we have

$$E(\psi) = E(\sum_{u=1}^{m}\sum_{p\in H_c}\chi(u)p(u)) = E(\sum_{u=1}^{m}\chi(u)\sum_{p\in H_c}p(u)), \tag{32}$$

where $\chi(u) = |O|\sum\limits_{p'\in H_0}p'(u)$, for $\forall p_\alpha\in H_0$, $p_\gamma\in H_c$, p_α and p_γ are independent of each other, we have

$$\begin{aligned} E(\psi) = & \sum_{u=1}^{m}|O|E(\sum_{p\in H_0}p(u))\cdot E(\sum_{p\in H_c}p(u)) \\ < & \sum_{u=1}^{m}\frac{|O|}{4}(\sum_{p\in H}E(p(u)))^2 = \frac{|O|\cdot|H|^2}{4m}. \end{aligned} \tag{33}$$

Since $0<(1-p_0^{k-1})<1$, then

$$E(P'_c) > (1-p_0^{k-1})^{\frac{|O|\cdot|H|^2}{4m}}. \tag{34}$$

C. Analysis of HABF Under Insertion Workloads

In this subsection, we theoretically analyze the performance of HABF under insertion workloads. Let the cost of the i^{th} bit of Bloom filter be $\Theta'(i)$. If α keys have been inserted, the probability that a certain bit remains '0' is $(1-\frac{1}{m})^{k\alpha}$.

Thus, for Bloom filter, the overall cost of false positives in O can be derived as

$$C_{bf} = \sum_{i=1}^{m}(1-(1-\frac{1}{m})^{k\alpha})\Theta'(i). \tag{35}$$

For HABF, in this scenario, when a bit x is set to '1', we compare its cost with the preset threshold τ. If $\Theta'(x)\geq\tau$, we try to customize the hash functions of inserted positive keys to avoid x set to '1'. We denote the probability that a bit i is avoided x set to '1' as $P(i)$, then for HABF, the overall cost of false positives in O can be derived as

$$C_{habf} = \sum_{i=1}^{m}(1-(1-\frac{1}{m}+\frac{P(i)}{m})^{k\alpha})\Theta'(i). \tag{36}$$

As per Equation (36), it is obvious that $C_{habf}\leq C_{bf}$. Since $(1-\frac{1-P(i)}{m})<1$, $\lim\limits_{\alpha\to+\infty}C_{habf} = \sum\limits_{i=1}^{m}\Theta'(i)$, which means C_{habf} will also tend to reach the max overall cost but at a lower speed. Next we analyze P in detail. We denote P_c as the probability that a inserted key can be customized its hash functions successfully and P_c as the probability that the customization results can be inserted into HashExpressor. For simplification, we assume that P_c and P_s are independent of each other, then we have

$$P(i) = Pr(\Theta'(i)>\tau)\cdot P_c\cdot P_s. \tag{37}$$

We set τ to the h^{th} highest value in $\Theta'(i)$ for $i=1,2,...,m$, then $Pr(\Theta'(i)>\tau)=\frac{h}{m}$. Considering P_c, for a inserted key e, and there are $|H_c|$ hash functions for selection, if there exists a hash function which can be used to mapped to a bit with cost below τ, the customization is successful, then

$$P_c = 1-(\frac{h}{m})^{|H_c|}. \tag{38}$$

Considering P_s related to the number of customized keys, which we denote as t, and according to Equation (11) and Lemma 4.1, we have

$$P_s > (1-\frac{kt+k}{\omega})^k > 1-\frac{k^2(t+1)}{\omega}. \tag{39}$$

As per Equation (37), we have $E(t) = k\alpha\cdot\frac{h}{m}\cdot P_c\cdot E(P_s)$, then we can derive

$$\begin{aligned} E(P_s) > & 1-\frac{k^2(E(t)+1)}{\omega} \\ > & 1-\frac{k^2(\frac{k\alpha hP_c}{m}\cdot E(P_s)+1)}{\omega} \\ > & \frac{(\omega+k^2)m}{\omega m+k^3\alpha hP_c}. \end{aligned} \tag{40}$$

As per Equation (37) and (40), then we have

$$E(P(i)) > \frac{(\omega+k^2)}{\frac{\omega m}{P_c h}+k^3\alpha}. \tag{41}$$