

APPENDIX

In what follows, we show the detailed steps to derive Formula (10) from Formula (5).

$$\begin{aligned} A_t &= \left(\frac{1}{2} + \left(\frac{1}{p} + \frac{1}{2}\right) \times \left(N_r + \frac{N_r - p}{n_p - 1} + \frac{p^2}{n_h}\right)\right) \times D_t \\ &= \frac{1}{2 \times n_h} \times p^2 + \left(\frac{1}{n_h} - \frac{1}{2 \times (n_p - 1)}\right) \times p + \frac{n_p \times N_r}{n_p - 1} \times \frac{1}{p} + C_2) \times D_t \end{aligned}$$

where $C_2 = \frac{1}{2} \times \left(1 + N_r + \frac{N_r - 2}{n_p - 1}\right)$. To obtain the best access time, we take the derivative of the the above formula:

$$\left(\frac{p}{n_h} + \left(\frac{1}{n_h} - \frac{1}{2 \times (n_p - 1)}\right) - \frac{n_p \times N_r}{n_p - 1} \times \frac{1}{p^2}\right) = 0 \quad (13)$$

multiplying both sides of the equation by p^2 , we obtain,

$$\frac{1}{n_h} \times p^3 + \left(\frac{1}{n_h} - \frac{1}{2 \times (n_p - 1)}\right) \times p^2 - \frac{n_p \times N_r}{n_p - 1} = 0 \quad (14)$$

multiplying both sides of the equation by n_h , we get

$$p^3 + \left(1 - \frac{n_h}{2 \times (n_p - 1)}\right) \times p^2 - \frac{n_h \times n_p \times N_r}{n_p - 1} = 0$$

The above formula can be converted into the following generic form:

$$x^3 + a \times x^2 + b \times x + c = 0$$

where $a = 1 - \frac{n_h}{2 \times (n_p - 1)}$, $b = 0$, and $c = -\frac{n_h \times n_p \times N_r}{n_p - 1}$. Let $x = y - \frac{a}{3}$. Replacing x by its equivalent expression in the above formula, we obtain:

$$y^3 + u \times y + v = 0$$

where $u = -\frac{1}{3} \times a^2$, and $v = \frac{2}{27} \times a^3 + c$. Solving this equation in the real set, we obtain:

$$y = \sqrt[3]{-\frac{v}{2} + \sqrt{\left(\frac{v}{2}\right)^2 + \left(\frac{u}{3}\right)^3}} + \sqrt[3]{-\frac{v}{2} - \sqrt{\left(\frac{v}{2}\right)^2 + \left(\frac{u}{3}\right)^3}}$$

Let us now revisit the values of a , b and c . The approximation of $c = -\frac{n_h \times n_p \times N_r}{n_p - 1} \approx -n_h \times N_r$ is justified because n_p is normally much greater than 1. n_h and n_p are usually of the same magnitude. Therefore, the absolute value of a is negligible compared to the absolute value of

c . For example, consider a scenario where a bucket stores about 20 index entries (n_p), or 100 hashing values and pointers (n_h). The resulting absolute value of coefficient a is approximately 10. This value is much smaller than the absolute value of c , which is usually several orders of magnitude larger. In our experiments, the number of data records (N_r) varies from 1,000 to 100,000. In this case, the corresponding absolute values of c would range from 100,000 to 10,000,000. Therefore, $v \approx c$ and $|u| \ll |v|$ are reasonable. Taking this information into account, we obtain:

$$\begin{aligned}
 p &= x \\
 &= y - \frac{a}{3} \\
 &\approx \sqrt[3]{-v} \\
 &\approx \sqrt[3]{-c} \\
 &\approx \sqrt[3]{n_h \times N_r}
 \end{aligned}$$

To achieve the best access time, the value of p can therefore be represented as:

$$p_v(AT)^* \approx \sqrt[3]{n_h \times N_r}$$

Because of the approximations we used during the computation, $p = p_v(AT)^*$ only yields a nearly optimal solution for access time. However, since the access time varies gradually with p , the resulting access time by $p_v(AT)^*$ should be close to the best access time. Our simulation results appear to support this analysis.