79-11

Reprinted from JOURNAL OF COMPUTER AND SYSTEM SCIENCES
All Rights Reserved by Academic Press, New York and London

Vol. 21, No. 3, December 1980
Printed in Belgium

Open-Addressing Hashing with Unequal-Probability Keys

GASTON H. GONNET

Department of Computer Science, University of Waterloo,
Waterloo, Ontario N2L 3G1, Canada

Received March 23, 1979; revised January 28, 1980

1. INTRODUCTION

When analyzing hashing algorithms, it is usually assumed that all keys in the table are equally likely to be accessed. In practice, however, this is generally not true. As Knuth [7] points out, the keys most likely to be accessed tend to appear first during the creation of the table; consequently, estimates based on an equal probability of accessing all keys are frequently pessimistic. In this paper we compute the expected number of accessed in a model of open-addressing hashing, assuming a variety of distributions for the probability of accessing individual keys in the table. The results presented here correspond to tables which are created by inserting keys in decreasing order of probability; this is the optimal ordering when we do not know the specific hashing function. The optimal order when we know the hashing function is described in [3].

2. RESULTS

Let n be the number of keys we are working with, and let m be the size of our table ($m \geqslant n$). Also let $p_1, p_2, ..., p_n$ be the probability of accessing the keys $k_1, k_2, ..., k_n$. We require that $p_1 \geqslant p_2 \geqslant \cdots \geqslant p_n > 0$ and furthermore that the table be created by inserting the keys in decreasing order of probability. This is the optimal insertion order when we do not know the particular hashing function. In practice, of course, we seldom know the probability distribution of accessing the keys; normally, however, the keys are inserted into the table as they appear, and the keys with the highest probabilities tend to appear first. Thus, our optimal order may be considered an optimistic approximation to the real situation.

We will be dealing here with the open-addressing hashing scheme. Under this scheme collisions are resolved by computing additional hashing functions until an empty position in the table is found. New probe positions for collision resolution are assumed to be independent of the previous ones; this is what Knuth calls *uniform probing* [7]. Double hashing, and schemes without secondary or higher clustering [4, 5], behaves similarly to uniform probing up to a certain load factor.

0022-0000/80/060354-14$02.00/0

In the open-addressing scheme, the probability of requiring more than i accesses to insert the $(n+1)$st key into the table is

$$\Pr\{\mathit{more\ than\ i\ accesses}\} = \frac{n^{\underline{i}}}{m^{\underline{i}}} = \frac{(m-i)^{\underline{m-n}}}{m^{\underline{m-n}}},$$

where $m^{\underline{k}}$ denotes the descending factorial, $m^{\underline{k}} = m(m-1)\cdots(m-k+1)$. Consequently the expected number of accesses to insert the key *is*

$$\begin{aligned} &E[\mathit{accesses\ to\ insert}\ (n+1)\text{st}\ \mathit{key}] \\ &\quad = \sum_{i\geqslant 0} i \Pr\{\mathit{exactly\ i\ accesses}\} \\ &\quad = \sum_{i\geqslant 0} \Pr\{\mathit{more\ than\ i\ accesses}\} = \sum_{i\geqslant 0} \frac{(m-i)^{\underline{m-n}}}{m^{\underline{m-n}}}. \end{aligned}$$

Using the summation formula for descending factorial this simplifies to

$$E[\mathit{accesses\ to\ insert}\ (n+1)\text{st}\ \mathit{key}] = \frac{(m+1)^{\underline{m-n+1}}}{(m-n+1)\, m^{\underline{m-n}}} = \frac{m+1}{m-n+1}.$$

Thus the expected number of accesses to locate a given key in the table containing n keys is

$$\begin{aligned} E[\mathit{accesses}] &= (m+1)\sum_{i=1}^{n} \frac{p_i}{m+2-i} = \frac{m+1}{m+2}\sum_{i=1}^{n}\left(\frac{p_i}{1-\dfrac{i}{m+2}}\right) \\ &= \frac{m+1}{m+2}\sum_{i=1}^{n}\sum_{k=0}^{\infty}\left(\frac{i}{m+2}\right)^{k} p_i = \frac{m+1}{m+2}\left(1+\sum_{k=1}^{\infty}\frac{\mu'_k}{(m+2)^k}\right) \end{aligned}$$

where $\mu'_k = \sum_{i=1}^{n} i^k p_i$ is the kth moment of the distribution of p_i.

This general result does not provide much insight on how a particular distribution may behave. Thus we will analyze our algorithm using several specific probability distributions, some of which come from experimental observation. The rest of this section will summarize our results. The asymptotic expansion results assume that $m \to \infty$, and that $\alpha = n/m$ $(0 < \alpha < 1)$ is a constant independent of m. The derivations of all these results will be given in Section 3 along with a brief description of each probability distribution used here. These derivations are quite straightforward for the most part, except for the geometric distribution where we have to consider three different cases, and the 80–20% rule.

In Table I we follow Knuth [6] by using H_n to denote the nth harmonic number, i.e.,

$$H_n = \sum_{i=1}^{n} \frac{1}{i} = \ln(n) + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + O(n^{-4}),$$

TABLE I

Type of distribution	p_i	Closed form
Uniform	$\frac{1}{n}$	$\frac{m+1}{n}(H_{m+1}-H_{m-n+1})$
Wedge (general)	$\frac{2(b-i)}{n(2b-n-1)}$	$\frac{2(m+1)(n+(b-m-2)(H_{m+1}-H_{m-n+1}))}{n(2b-n-1)}$
Wedge $(b=n+1)$	$\frac{2(n+1-i)}{n(n+1)}$	
Zipf's law (harmonic)	$\frac{1}{iH_n}$	$\frac{m+1}{m+2}\left(1+\frac{H_{m+1}-H_{m-n+1}}{H_n}\right)$
Generalized harmonic	$\frac{(i+a)^{-1}}{\psi(a+n+1)-\psi(a+1)}$	$\frac{m+1}{m+a+2}\left(1+\frac{H_{m+1}-H_{m-n+1}}{\psi(a+n+1)-\psi(a+1)}\right)$
Lotka's law bi-harmonic	$\frac{1}{i^2H_n^{(2)}}$	$\frac{m+1}{m+2}\left(1+\frac{H_{m+1}-H_{m-n+1}+H_n}{H_n^{(2)}(m+2)}\right)$
Geometric $-m\ln(a)=\beta=\omega(1)$	$\frac{(1-a)\,a^{i-1}}{1-a^n}$	
$-m\ln(a)=\beta=\Theta(1)$	$\frac{(1-a)\,a^{i-1}}{1-a^n}$	
$-m\ln(a)=\beta=o(1)$	$\frac{(1-a)\,a^{i-1}}{1-a^n}$	
80–20% rule	$\frac{i^\theta-(i-1)^\theta}{n^\theta}$	

where γ is Euler's constant, $\gamma=0.57721\ 56649....$ Similarly

$$H_n^{(2)}=\sum_{i=1}^{n} i^{-2}=\frac{\pi^2}{6}-\frac{1}{n}+\frac{1}{2n^2}-\frac{1}{6n^3}+O(n^{-5}).$$

$\mathrm{Ei}(x)$ denotes the exponential integral [1]

$$\mathrm{Ei}(x)=\int_{-x}^{\infty} e^{-t}t^{-1}\,dt.$$

TABLE 1 (*continued*)

$\alpha = n/m$	Full table
$-\alpha^{-1}\ln(1-\alpha) + O(m^{-1})$	$\ln(m) + \gamma - 1 + O(\ln(m)/m)$
$2\alpha^{-2}[\alpha + (1-\alpha)\ln(1-\alpha)] + O(m^{-1})$	$2 + O(\ln(m)/m)$
$1 - \dfrac{\ln(1-\alpha)}{\ln(n)+\gamma} + O(m^{-1})$	$2 - \dfrac{1}{\ln(m)+\gamma} + O(m^{-1})$
$1 + \dfrac{6\ln n}{m\pi^2} + O(m^{-1})$	Same
$1 + \dfrac{\alpha}{(1-\alpha)m} + \dfrac{2\alpha^2}{(1-\alpha)^2 m(m-1)} + O(\beta^{-3})$	Same
$\dfrac{\beta e^{-\beta}}{1-e^{-\alpha\beta}}[\mathrm{Ei}(\beta) - \mathrm{Ei}(\beta(1-\alpha))] + O(m^{-1})$	$\dfrac{\beta e^{-\beta}}{1-e^{-\beta}}[\mathrm{Ei}(\beta) - 1 - \ln(\beta) + \ln m] + O(m^{-1})$
$-\alpha^{-1}\ln(1-\alpha) + \left(1 + \left(\dfrac{1}{\alpha} - \dfrac{1}{2}\right)\ln(1-\alpha)\right)\beta$ $+ O(\beta^2) + O(m^{-1})$	$(\ln m + \gamma - 1)(1 - \beta/2) + \beta + O(\beta^2) + O(m^{-1})$
$C(\alpha) + O(1/m)$	$1 + \theta\ln(m) + C_1 + o(1)$

The values for the constants and the function $C(\alpha)$ are given in the next section. Our asymptotic notation follows the conventions used by Knuth [6].

$$f(n) = O(g(n)) \quad \text{if} \quad \text{there exist } k \text{ and } n_0 \text{ such that}$$
$$|f(n)| \leqslant kg(n) \qquad \text{for } n > n_0.$$
$$f(n) = o(g(n)) \quad \text{if} \quad \lim_{n\to\infty} \frac{f(n)}{g(n)} = 0.$$
$$f(n) = \omega(g(n)) \quad \text{if} \quad g(n) = o(f(n)).$$

$$f(n) = \Theta(g(n)) \quad \text{if} \quad \text{there exist } k_1, k_2 \text{ and } n_0 \text{ such that}$$
$$|k_1 g(n)| \leqslant |f(n)| \leqslant |k_2 g(n)| \qquad \text{for } n > n_0.$$

Table II lists some exact numerical results for the different distributions, rounded to five decimal places. For each distribution the first row gives figures for a hash table with $m = 100$, while the second is for a hash table with $m = 1000$. Each case was examined for the five different occupancy factors shown.

3. Derivation of Results

For the uniform distribution, $p_i = 1/n$ and thus

$$E[accesses] = (m+1) \sum_{i=1}^{n} \frac{p_i}{m+2-i}$$
$$= \frac{(m+1)}{n} (H_{m+1} - H_{m-n+1}).$$

In the case of a full table ($n = m$) this becomes

$$E[accesses] = \frac{m+1}{m} (H_{m+1} - 1) = \ln(m) + \gamma - 1 + O(\ln(m)/m),$$

while in a partly filled table with load factor $\alpha = n/m$,

$$E[accesses] = -\alpha^{-1} \ln(1-\alpha) + O(m^{-1}).$$

TABLE II

Occupancy Factors

Distribution	50%	80%	90%	95%	100%
Uniform	1.37050	1.95930	2.44353	2.92079	4.23925
	1.38468	2.00633	2.54609	3.12697	6.49296
Wedge	1.21978	1.47789	1.62904	1.73907	1.91605
$(b = n + 1)$	1.22664	1.49245	1.65115	1.76966	1.98703
Harmonic	1.13951	1.29967	1.41440	1.51982	1.79140
	1.10072	1.21957	1.30887	1.39779	1.86468
Lotka's law	1.02113	1.02895	1.03333	1.03702	1.04592
	1.00354	1.00438	1.00487	1.00531	1.00748
Geometric	1.03201	1.03201	1.03201	1.03201	1.03201
$(a = 3/4)$	1.00302	1.00302	1.00302	1.00302	1.00302
$(a = 9/10)$	1.10684	1.11327	1.11396	1.11428	1.11477
	1.00917	1.00917	1.00917	1.00917	1.00917
$(a = 0.99)$	1.33246	1.77291	2.09726	2.40001	3.19530
	1.11936	1.12792	1.12902	1.12958	1.13130
80–20% rule	1.08120	1.19327	1.27574	1.35200	1.54897
	1.08561	1.20340	1.29445	1.38596	1.86998

The general *wedge* (or truncated wedge) probability distribution is defined by

$$p_i = \frac{2(b-i)}{n(2b-n-1)} \qquad \textit{with } b > n.$$

From this we obtain

$$E[\textit{number of accesses}] = \frac{2(m+1)}{n(2b-n-1)} \sum_{i=1}^{n} \frac{b-i}{m+2-i}$$
$$= \frac{2(m+1)}{n(2b-n-1)} (n + (b-m-2)(H_{m+1} - H_{m-n+1})).$$

If we set $b = n+1$ we obtain the wedge distribution for which the probabilities are proportional to the integers $1, 2, \ldots, n$. The expected value in this case becomes

$$E[\textit{accesses}] = 2\alpha^{-2}[\alpha + (1-\alpha)\ln(1-\alpha)] + O(m^{-1}).$$

For a full table we have $n = m = b - 1$ and

$$E[\textit{accesses}] = \frac{2}{m}[m - (H_{m+1} - 1)]$$
$$= 2 + O(\ln(m)/m).$$

For keys distributed according to Zipf's law, the harmonic distribution [7] we have $p_i = 1/H_n i$. Consequently

$$E[\textit{accesses}] = \frac{m+1}{H_n} \sum_{i=1}^{n} \frac{1}{(m+2-i)\,i} = \frac{m+1}{H_n(m+2)} \sum_{i=1}^{n} \left(\frac{1}{i} + \frac{1}{m+2-i}\right)$$
$$= \frac{m+1}{H_n(m+2)} (H_n + H_{m+1} - H_{m-n+1})$$
$$= \frac{m+1}{m+2} \left(1 + \frac{H_{m+1} - H_{m-n+1}}{H_n}\right).$$

For the full table result we set $n = m$ to obtain

$$E[\textit{accesses}] = 2 - \frac{1}{H_m} + O(m^{-1}).$$

and for the partly filled table with $\alpha = n/m$

$$E[\textit{accesses}] = 1 - \frac{\ln(1-\alpha)}{\ln(n) + \gamma} + O(m^{-1}).$$

We will consider a generalization of the harmonic probability distribution that makes $p_i \propto 1/(i+a)$. After normalization we obtain

$$p_i = \frac{1}{i+a} \frac{1}{\psi(a+n+1)-\psi(a+1)},$$

where a is a real constant and $\psi(x)$ is the logarithmic derivative of the gamma function [1]. Simplifying this in a manner very similar to our work with the Zipf distribution we obtain

$$E[accesses] = \frac{m+1}{m+a+2}\left[1+\frac{H_{m+1}-H_{m-n+1}}{\psi(a+n+1)-\psi(a+1)}\right].$$

Lotka's distribution (bi-harmonic) is given by

$$p_i = \frac{1}{H_n^{(2)} i^2},$$

and thus

$$\begin{aligned} E[accesses] &= \frac{m+1}{H_n^{(2)}} \sum_{i=1}^{n} \frac{1}{(m+2-i)\, i^2}, \\ &= \frac{m+1}{H_n^{(2)}} \sum_{i=1}^{n} \left(\frac{(m+2)^{-2}}{m+2-i} + \frac{(m+2)^{-2}}{i} + \frac{(m+2)^{-1}}{i^2}\right), \\ &= \frac{m+1}{m+2}\left(1+\frac{H_{m+1}-H_{m-n+1}+H_n}{H_n^{(2)}(m+2)}\right). \end{aligned}$$

For any α we conclude that

$$E[accesses] = 1 + \frac{6(\ln n - \ln(1-\alpha)+\gamma)}{m\pi^2} - \frac{1}{m} + O(m^{-2}\ln m).$$

The geometric distribution is shaped by the parameter a, and is given by

$$p_i = \frac{(1-a)\, a^{i-1}}{1-a^n}; \quad 0 < a < 1; \quad \alpha = n/m; \quad \beta = -m\ln(a).$$

The parameter a may be a function of m. To obtain the proper asymptotic expansions we must consider three different cases.

(a) $m\ln(a) = \omega(1)$. Thus for increasing m, β does not remain bounded. Note that this case includes the situation when a does not depend on m. Note also that this condition is equivalent to $m(1-a) = \omega(1)$, since $0 < a < 1$ then $0 < -m\ln(a) = \omega(1)$ or $0 < \ln(a) = \omega(1/m)$ or $1/a = e^{\omega(1/m)} = 1 + \omega(1/m)$. From this we conclude that $m(1-a)/a = \omega(1)$, and either $m(1-a) = \omega(1)$ or $m(1-a) \neq \omega(1)$ and $1/a = \omega(1)$.

It is easy to see that the latter condition is impossible since $m \to \infty$ and consequently $m(1-a) = \omega(1)$. From our formula we obtain

$$\begin{aligned}
&E[accesses]\\
&\quad= \frac{(1-a)(m+1)}{1-a^n} \sum_{i=1}^{n} \frac{a^{i-1}}{m+2-i}\\
&\quad= \frac{(1-a)(m+1)}{1-a^n} \left(\frac{1-a^n}{(1-a)(m+1)} + \sum_{i=1}^{n} \frac{(i-1)\,a^{i-1}}{(m+2-i)(m+1)} \right)\\
&\quad= 1 + \frac{1}{1-e^{-\alpha\beta}} \left(\frac{a - na^n + (n-1)\,a^{n+1}}{(1-a)\,m} + (1-a) \sum_{i=1}^{n} \frac{(i-1)(i-2)\,a^{i-1}}{(m+2-i)\,m} \right)\\
&\quad= 1 + \frac{a}{(1-a)\,m} + \frac{2a^2}{(1-a)^2\,m(m-1)} + O(\beta^{-3}).
\end{aligned}$$

(b) $m\ln(a) = \Theta(1)$, or equivalently $m(1-a) = \Theta(1)$, as can be seen using the Taylor expansion of $\ln(a)$. In this case β remains bounded as m increases. We have

$$\begin{aligned}
E[accesses] &= \frac{(m+1)(1-a)}{1-a^n} \sum_{i=1}^{n} \frac{a^{i-1}}{m+2-i},\\
&= \frac{(m+1)(1-a)}{1-a^n}\, a^{m+1} \sum_{j=m+2-n}^{m+1} \frac{(1/a)^j}{j}.
\end{aligned}$$

Using the Euler–Maclaurin summation formula [2, 6] we derive

$$\begin{aligned}
E[accesses] &= \frac{(m+1)(1-a)}{1-a^n}\, a^{m-1} \left\{ \frac{a^{-m-1}}{m+1} \right.\\
&\quad \left. + \left(\mathrm{Ei}(-x\ln(a)) - \frac{a^{-x}}{2x} - \frac{a^{-x}[x\ln(a)+1]}{12x^2} \right)_{x=m-n+2}^{x=m+1} + O(m^{-4})) \right\}\\
&= \frac{\beta e^{-\beta}}{1-e^{-\alpha\beta}} (\mathrm{Ei}(\beta) - \mathrm{Ei}((1-\alpha)\,\beta)) + O(m^{-1}).
\end{aligned}$$

For a full table we use the techniques described in [2] to evaluate the summation and we obtain

$$E[accesses] = \frac{\beta e^{-\beta}}{1-e^{-\beta}} (\mathrm{Ei}(\beta) - 1 - \ln\beta + \ln m) + O(m^{-1}).$$

(c) $m\ln(a) = o(1)$, or equivalently $m(1-a) = o(1)$. In this case we can obtain a good approximation of the sum by using the first two terms of the sum's Taylor expansion around $a = 1$. By doing this or by computing the series expansion for

$\mathrm{Ei}(x) = \gamma + \ln(x) + x + x^2/4 + \cdots$ and $e^x = 1 + x + x^2/2 + \cdots$ in the result for case (b) we obtain

$$E[accesses] = -\alpha^{-1}\ln(1-\alpha) + \left(1 + \left(\frac{1}{\alpha} - \frac{1}{2}\right)\ln(1-\alpha)\right)\beta + O(\beta^2) + O(m^{-1}).$$

For full tables we obtain

$$E[accesses] = (\ln m + \gamma - 1)(1 - \beta + \beta^2/12) + \beta - \beta^2/4 + O(\beta^3) + O(m^{-1}).$$

The final probability distribution which we will examine comes deom the 80–20% rule. This rule indicates that 80% of all accesses will be made on the most active 20% of the keys, and so on "recursively." When we say "recursively," we mean that the rule applies not only to the entire table of keys $k_1, k_2, \ldots, k_n$, but to any subset of the table consisting of the p most active keys $k_1, k_2, \ldots, k_p$, where $p \leqslant n$. The simplest probability distribution which models this rule is given by

$$p_i = \frac{i^\theta - (i-1)^\theta}{n^\theta},$$

where

$$\theta = \frac{\log(0.8)}{\log(0.2)} = 0.13864 \ldots.$$

In our analysis of this distribution, we will use the following terminology: $\zeta(z)$ is Riemann's zeta function. $\Gamma(z)$ is the gamma function, $\psi(z)$ its logarithmic derivative and ${}_2F_1(a, b; c; z)$ the Gauss hyppergeometric series [1]. We will also employ the convention that $0^\theta = 0$ to simplify notation.

The moments of the 80–20% distribution, computed using the Euler–Maclaurin formula and its extensions [2.6], are given by

$$\mu_1' = \sum_{i=1}^{n} ip_i = \frac{\theta n}{\theta+1} + \frac{1}{2} - \frac{\zeta(-\theta)}{n^\theta} - \frac{\theta}{12n} + O(n^{-3}),$$

$$\mu_2' = \sum_{i=1}^{n} i^2 p_i = \frac{\theta n^2}{\theta+2} + \frac{\theta n}{\theta+1} + \frac{2-\theta}{6} - \frac{2\zeta(-\theta-1) + \zeta(-\theta)}{n^\theta} + O(n^{-1}),$$

$$\mu_3' = \sum_{i=1}^{n} i^3 p_i = \frac{\theta n^3}{\theta+3} + \frac{3\theta n^2}{2(\theta+2)} + \frac{\theta(3-\theta)\,n}{4(\theta+1)} + O(1),$$

$$\mu_k' = \sum_{i=1}^{n} i^k p_i = \frac{\theta n^k}{\theta+k} + \frac{\theta k n^{k-1}}{2(\theta+k-1)} + \frac{\theta(k-\theta)\,k n^{k-2}}{12(\theta+k-2)} + O(n^{k-3}) + O(n^{-\theta}),$$

and the variance of the distribution is

$$\sigma^2 = \frac{\theta n^2}{(\theta+1)^2(\theta+2)} + O(n^{1-\theta}).$$

Recall from Section 2 that

$$E[accesses] = \frac{m+1}{m+2}\left(1 + \sum_{k=1}^{\infty} \frac{\mu_k}{(m+2)^k}\right).$$

Consequently with $\alpha = n/m$, $0 < \alpha < 1$, we may substitute the moments of the distribution into the above formula to obtain

$$E[accesses] = C(\alpha) + O(m^{-1}) = 1 + \theta \sum_{k=1}^{\infty} \frac{\alpha^k}{\theta + k} + O(m^{-1}).$$

This last series is inconvenient for the purpose of evaluation when α is close to 1. Thus we use the transformation

$$\begin{aligned}
\sum_{k=1}^{\infty} \frac{\alpha^k}{\theta+k} &= \alpha^{-\theta} \sum_{k=1}^{\infty} \int_0^{\alpha} x^{k-1+\theta}\,dx = \alpha^{-\theta} \int_0^{\alpha} x^{\theta} \sum_{k=1}^{\infty} x^{k-1}\,dx \\
&= \alpha^{-\theta} \int_0^{\alpha} \frac{x^{\theta}}{1-x}\,dx = \alpha^{-\theta} \int_{1-\alpha}^{1} \frac{(1-y)^{\theta}}{y}\,dy \\
&= \alpha^{-\theta} \int_{1-\alpha}^{1} \left(\frac{1}{y} - \frac{\theta y}{y} + \frac{\theta(\theta-1)}{2y} y^2 - \frac{\theta(\theta-1)(\theta-2)}{3!y} y^3 + \cdots\right) dy \\
&= \alpha^{-\theta} \left(\ln y - \theta y + \frac{\theta(\theta-1)}{2!2} y^2 - \frac{\theta(\theta-1)(\theta-2)}{3!3} y^3 + \cdots\right)\Bigg|_{1-\alpha}^{1}.
\end{aligned}$$

Since

$$\begin{aligned}
-\theta &+ \frac{\theta(\theta-1)}{2!2} - \frac{\theta(\theta-1)(\theta-2)}{3!3} + \cdots \\
&= \lim_{b\to 0} \left(\left(\frac{(-\theta)\,b}{1!1!} + \frac{(-\theta)(1-\theta)\,b(b+1)}{2!2!} \right.\right. \\
&\qquad \left.\left. + \frac{(-\theta)(1-\theta)(2-\theta)\,b(b+1)(b+2)}{3!3!} + \cdots\right) b^{-1}\right) \\
&= \lim_{b\to 0} \left(\frac{{}_2F_1(-\theta, b; 1; 1) - 1}{b}\right) = \lim_{b\to 0} \left(\left(\frac{\Gamma(1)\,\Gamma(1+\theta-b)}{\Gamma(1+\theta)\,\Gamma(1-b)} - 1\right) b^{-1}\right) \\
&= -\psi(1+\theta) + \psi(1) = -\psi(1+\theta) - \gamma,
\end{aligned}$$

we have the final expression

$$\begin{aligned}
\sum_{k=1}^{\infty} \frac{\alpha^k}{\theta+k} = \alpha^{-\theta} &\left(-\psi(1+\theta) - \gamma - \ln(1-\alpha) + \theta(1-\alpha) - \frac{\theta(\theta-1)(1-\alpha)^2}{4} \right. \\
&\left. + \frac{\theta(\theta-1)(\theta-2)(1-\alpha)^3}{18} - \cdots\right).
\end{aligned}$$

Note also that

$$\int_0^\alpha \frac{x^\theta}{1-x}\,dx = -\psi(1+\theta) - \gamma - \ln(1-\alpha) + \theta(1-\alpha) + O((1-\alpha)^2).$$

Thus we can express the expected value as

$$E[accesses] = 1 + \theta\alpha^{-\theta}(-\psi(1+\theta) - \gamma - \ln(1-\alpha) + \theta(1-\alpha) - \cdots) \\ + O(m^{-1}).$$

Direct computation allows us to give a list of some values for the function $C(\alpha)$:

$$C(0.5) = 1.08617\ \ 37741\ \ 28045\ \ 31512\ \ 1541\cdots,$$

$$C(0.8) = 1.20463\ \ 31959\ \ 45617\ \ 83466\ \ 2159\cdots,$$

$$C(0.9) = 1.29674\ \ 51213\ \ 95053\ \ 49342\ \ 7795\cdots,$$

$$C(0.95) = 1.39031\ \ 66674\ \ 21293\ \ 92038\ \ 3669\cdots,$$

$$C(0.99) = 1.61076\ \ 56741\ \ 38615\ \ 66525\ \ 3015\cdots,$$

$$C(0.99940\ 04851\ 43161\ 88579\ 4781\cdots) = 2.0,$$

$$C(0.99999\ 95583\ 01845\ 90154\ 1475\cdots) = 3.0.$$

To consider the case for full tables ($n = m$) we first obtain the inequality

$$\frac{\theta}{n^\theta}\int_{i-1}^{i} \frac{x^{\theta-1}}{m+2-x}\,dx \leqslant \frac{i^\theta - (i-1)^\theta}{n^\theta(m+2-i)} \leqslant \frac{\theta}{n^\theta}\int_{i-1}^{i} \frac{x^{\theta-1}}{m+1-x}\,dx.$$

Since $E[accesses] = (m+1)\sum_{i=1}^{n} [i^\theta - (i-1)^\theta/n^\theta(m+2-i)]$ we have

$$\frac{(m+1)\,\theta}{n^\theta}\int_0^n \frac{x^{\theta-1}}{m+2-x}\,dx \leqslant E[accesses] \leqslant \frac{(m+1)\,\theta}{n^\theta}\int_0^n \frac{x^{\theta-1}}{m+1-x}\,dx.$$

Transforming the left integral with $y = x/(m+2)$ and the right integral with $y = x/(m+1)$ and for $n = m$, we obtain

$$\frac{(m+1)\,\theta}{m^\theta}\int_0^{m/(m+2)} \frac{(m+2)^{\theta-1}}{1-y}\,y^{\theta-1}\,dy \leqslant E[accesses] \\ \leqslant \frac{\theta}{m^\theta}\int_0^{m/(m+1)} \frac{(m+1)^\theta\,y^{\theta-1}}{1-y}\,dy.$$

Finally, we may use part of the previous derivation to get

$$[1 + O(m^{-1})]\,\theta(-\psi(\theta) - \gamma + \ln(m) - \ln(2)) \\ \leqslant E[accesses] \leqslant [1 + O(m^{-1})]\,\theta(-\psi(\theta) - \gamma + \ln(m)).$$

Let us write

$$E[accesses] = 1 + \theta \ln m + C_1 + O(m^{-1} \ln m).$$

The above bounds imply that

$$-1 - \theta(\psi(\theta) + \gamma + \ln(2)) \leqslant C_1 \leqslant -1 - \theta(\psi(\theta) + \gamma).$$

There is an alternative way of computing the number of accesses for a full table. Let $g(i) = (i^\theta - (i-1)^\theta)/(m + 2 - i)$; then

$$\begin{aligned} E[accesses] &= \frac{m+1}{m^\theta} \sum_{i=1}^{m} g(i) \\ &= \frac{m+1}{m^\theta} \left(\int_1^m g(i)\, di + \frac{g(m)}{2} + \frac{g'(m)}{12} - \frac{g'''(m)}{720} + \cdots + C_g \right). \end{aligned}$$

Using straightforward computation and the techniques described in [2] we find

$$C_g = O(m^{-1}),$$

$$\frac{g(m)}{2} = \frac{\theta m^{\theta - 1}}{4} (1 + O(m^{-1})),$$

$$\frac{g'(m)}{12} = \frac{\theta m^{\theta - 1}}{48} (1 + O(m^{-1})),$$

and

$$\frac{g'''(m)}{720} = \frac{\theta m^{\theta - 1}}{1920} (1 + O(m^{-1})).$$

Using the value computed before for $\int_0^\alpha (x^\theta/(1-x))\, dx$ we derive

$$\begin{aligned} \int_1^m g(i)\, di = {} & (-\psi(1+\theta) - \gamma - \ln(2))\, \theta(m+1)^{\theta-1} (\theta \ln(m+1) + 1) \\ & + \frac{1}{(1+\theta)(m+2)} + O(m^{\theta-2}), \end{aligned}$$

and finally

$$\begin{aligned} E[accesses] = {} & \theta(-\psi(1+\theta) - \gamma - \ln(2)) + 1 + \theta \ln(m+1) \\ & + \frac{\theta}{4} + \frac{\theta}{48} - \frac{\theta}{1920} + \cdots. \end{aligned}$$

Hence by computing the Euler–Maclaurin formula with three and four terms we can bound C_1 by

$$\theta\left(-\frac{1}{1920}+\frac{1}{48}+\frac{1}{4}-\ln 2-\gamma-\psi(1+\theta)\right)$$
$$\leqslant C_1 \leqslant \theta\left(\frac{1}{48}+\frac{1}{4}-\ln 2-\gamma-\psi(1+\theta)\right).$$

To seven-digit accuracy these bounds are

$$-0.0873946\cdots \leqslant C_1 \leqslant -0.0873224\cdots.$$

The distribution given by

$$p_i=\frac{i^{\theta-1}}{H_n^{(1-\theta)}}$$

is asymptotically equivalent to the 80–20% rule. For this distribution, using the same derivations as before, we find that for full tables

$$E[accesses]=1+\theta[\ln m-\psi(1+\theta)-1]+O(m^{-1}\ln m).$$

From this result we derive the strong conjecture that

$$C_1=-\theta[\psi(1+\theta)+1]=-0.08738\ 76749\ 82611\ 29115\ 5901\cdots,$$
$$(\psi(1+\theta)=-0.36971\ 05008\ 49560\ 89275\ 0609\cdots).$$

The formulas we have derived are valid for any value of θ; thus they could be used to calculate results for a 75–25% rule, for example. In particular when $\theta=1$ we have the uniform distribution.

4. Direct Chaining

For the purpose of comparison, we show the effect of unequal probability keys for direct chaining hashing (or separate overflow chaining). In this method we hash into a sequential list that contains all the elements that share the same hashing address. If we insert all the elements in decreasing probability order, we obtain

$$E[accesses]=\sum_{i=1}^{n} p_i \sum_{j=0}^{i-1}(j+1)\binom{i-1}{j} m^{-j}(1-1/m)^{i-1-j}$$
$$=\sum_{i=1}^{n} p_i\left(\frac{i-1}{m}+1\right)=1+\frac{\mu_1'-1}{m}.$$

5. Conclusion

Our analysis has shown that if a hashing table is constructed in the optimal order, the average number of accesses remains very low for most of the probability distributions we have studied, even when the table is full. For many distributions the expected number of accesses is always less than two. As has been noted, in practical applications the probability distribution involved is usually not known. However, elements with higher accessing probability are more likely to appear first and be inserted first. The real average values can be expected to lie somewhere between between our optimistic results and the values which arise from the uniform distribution.

ACKNOWLEDGMENTS

The author wishes to acknowledge the contribution of Tatuo Nakanishi in the computations of the 80–20% rule, and the anonymous referee who read the manuscript very carefully and made precise remarks.

REFERENCES

1. M. ABRAMOWITZ AND I. A. STEGUN, "Handbook of Mathematical Functions," Dover, New York, 1964.
2. G. H. GONNET, Notes on the derivation of asymptotic expressions from summations. *Inform. Processing Lett.* **7**, No. 4 (1978), pp. 165–169.
3. G. H. GONNET AND J. I. MUNRO, Efficient ordering of hash tables, *SIAM J. Computing*. **8**, No. 3 (1979), pp. 463–478.
4. L. J. GUIBAS, The analysis of hashing techniques that exhibit *l*-ary clustering, *J. Assoc. Comput. Mach.* **25**, No. 4 (1978), pp. 544–555.
5. L. J. GUIBAS AND E. ZZEMEREDI, The analysis of double hashing, *J. Comput. System Sci.* **16**, No. 2 (1978), pp. 226–274.
6. D. E. KNUTH, "The Art of Computer Programming," Vo. I, "Fundamental Algorithms," 2nd ed., Addison–Wesley, Reading, Mass. 1973.
7. D. E. KNUTH, "The Art of Computer Programming," Vol. III, "Sorting and Searching," Addison–Wesley, Reading, Mass., 1973.

Printed by the St. Catherine Press Ltd., Tempelhof 37, Bruges, Belgium