EXPECTED LENGTH OF THE LONGEST PROBE SEQUENCE IN HASH CODE SEARCHING

Gaston H. Gonnet

Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada. N2L 3G1
December 1978
CS-RR-78-46

Faculty

of

Mathematics

University of Waterloo
Waterloo, Ontario, Canada

EXPECTED LENGTH OF THE LONGEST PROBE SEQUENCE IN HASH CODE SEARCHING

Gaston H. Gonnet

Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada. N2L 3G1
December 1978
CS-RR-78-46

This research was supported in part by the University of Waterloo, under operating grant 126-7029.

Expected Length of the Longest Probe Sequence in [redacted] Code Searching

GASTON H. GONNET

University of Waterloo, Waterloo, Ontario, Canada.

ABSTRACT. We investigate the expected value of the maximum number of accesses needed to locate any element in a hashing file under various collision resolution schemes. This differs from usual worst case considerations which, for hashing, would be the largest sequence of accesses for the worst possible file. We find asymptotic expressions of these expected values for full and partly full tables. For the open addressing scheme with a clustering-free model we find these values to be $0.6315...\times n$ for a full table and $\approx -\log_\alpha n$ for a partly full table, where n is the number of records, m is the size of the table, and $\alpha = n/m$. For the open addressing scheme which reorders the insertions to minimize the worst case under a random probing model, we find the tight lower bounds $\ln n + 1.077...$ and $\lceil -\alpha^{-1} \ln(1-\alpha) \rceil$ for full and partly full tables respectively. Finally for the separate chaining (or direct chaining) method, we find both expected values to be $\approx \Gamma^{-1}(n)$. These results show that for these schemes, the actual behaviour of the worst case in hash tables is quite good on the average.

KEY WORDS AND PHRASES: table search, hashing, analysis of algorithms, worst case, open addressing, separate chaining, direct chaining, asymptotic analysis, optimal hashing, minimax hashing, expected value, average case.

CR CATEGORIES: 3.74, 4.34, 5.25, 5.30.

1. *Introduction*

It is well known that a hash table with n keys inserted may have a worst case of n accesses to insert (or locate) an element; that is, the last key inserted may require up to n probes. This, however, is the worst case for the worst possible table. For some hashing schemes, this worst case occurs with a ridiculously small probability. This result, besides being discouraging, does not contribute any information about the ordinary behaviour of the length of the longest probe sequence (*llps* for short) in a random file. In other words, the *llps* of a table is a random variable; we know that this random variable has a maximum value of n, but its average value, i.e. the average *llps*, is also of interest.

In this paper we find asymptotic expressions of the average *llps* for full and partly full tables under three hashing schemes: open-addressing with a clustering-free hash function (uniform probing), a reordering scheme for open-addressing which minimizes the *llps*, and separate (or direct) chaining.

Sections 2, 3 and 4 are mathematically oriented and give the derivation of the new results for each of the hashing schemes analyzed. Finally the last section presents a comparative table of results, along with some conclusions.

We use the following standard notation to describe asymptotic performance.

$f(m) = O(g(m))$ when $|f(m)| \leqslant k\,|g(m)|$ for $m \geqslant m_0$,

$f(m) = o(g(m))$ when $\lim_{m\to\infty} f(m)/g(m) = 0$,

$f(m) = \Theta(g(m))$ when $k_1 g(m) \leqslant f(m) \leqslant k_2 g(m)$ for $m \geqslant m_0$ and $k_1 k_2 \neq 0$

and

$f(m) \approx g(m)$ when $\lim_{m\to\infty} f(m)/g(m) = 1$.

Furthermore, we will always use m to denote the size of the table, and n to denote the number of elements inserted in the table.

2. *Average Length of the Longest Probe Sequence for Open-Addressing*

In this section we use a clustering-free model for a hashing function. To insert a key using this scheme, we probe the table following a sequence of positions which is a random permutation of all table locations. This model is usually called uniform probing [14]. The work by Guibas and Szemeredi [9,10] shows that, up to a certain load factor, second or higher order clustering is asymptotically equivalent to uniform probing.

2.1. *Full tables*

It is well known that the average length of the longest probe sequence (*llps*) of a full table organized with uniform probing is $\Theta(n)$ (the last key requires $(n+1)/2$ probes on the average). We are interested in finding the factor which multiplies n for the average *llps*.

Inserting elements into a table of size m we find

$$Pr\{j^{th}\ last\ key\ requires \leqslant k\ probes\} = 1-\frac{(m-j)^{\underline{k}}}{m^{\underline{k}}} \tag{2.1}$$

where $x^{\underline{k}}$ indicates the descending factorial $x^{\underline{k}}=x(x-1)...(x-k+1)$, and the 1^{st} *last* is the last, the 2^{nd} *last* is the next to last, and so on.

In the uniform probing hashing scheme, all probe paths are independent; consequently the *llps* among the last j keys has a probability distribution given by

$$Pr\{llps\ among\ last\ j\ keys \leqslant k\} = \prod_{i=1}^{j}\left[1-\frac{(m-i)^{\underline{k}}}{m^{\underline{k}}}\right]. \tag{2.2}$$

We next compute the expected value of the *llps* among the last j keys based on this probability distribution. If we express the expected value as

$$E[X] = \sum_{i\geqslant 1} i[Pr\{X>i-1\}-Pr\{X>i\}] = \sum_{i\geqslant 0} Pr\{X>i\},$$

we find that the expected value can be given by

$$E[llps\ among\ last\ j\ keys] = \sum_{k=0}^{m}\left\{1-\prod_{i=1}^{j}\left[1-\frac{(m-i)^{\underline{k}}}{m^{\underline{k}}}\right]\right\}. \tag{2.3}$$

For any fixed j, when $m\to\infty$ we may use the transformation

$$\frac{(m-j)^{\underline{k}}}{m^{\underline{k}}} = \frac{(m-j)(m-j-1)...(m-k+1)...(m-j-k+1)}{m(m-1)...(m-j)...(m-k+1)} = \frac{(m-k)^{\underline{j}}}{m^{\underline{j}}}$$

$$= \frac{(m-k)^j-\binom{j}{2}(m-k)^{j-1}+...}{m^j-\binom{j}{2}m^{j-1}+...}$$

$$= \left(\frac{m-k}{m}\right)^j\left[1+\binom{j}{2}\left(\frac{1}{m-k}-\frac{1}{m}\right)+O(m^{-2})\right].$$

Computing uniform asymptotics [4] with respect to k (the maximum value occurs for $k=m/j$) we obtain

$$\frac{(m-j)^{\underline{k}}}{m^{\underline{k}}} = \left(\frac{m-k}{m}\right)^j + O(m^{-1}), \tag{2.4}$$

Using (2.4) in (2.3) gives us

$$E[llps\ among\ last\ j\ keys] = m\left[1-\frac{1}{m}\sum_{k=0}^{m}\prod_{i=1}^{j}\left[1-\left(\frac{m-k}{m}\right)^i+O(m^{-1})\right]\right].$$

We can simplify this using the substitution $y = (m-k)/m$ and by changing the sum to an integral (since $m \to \infty$). This yields

$$E[\textit{llps among last } j \textit{ keys}] = m\{1-\int_0^1 \prod_{i=1}^{j} (1- \quad ^i)\, dy\} + O(1). \tag{2.5}$$

For the latter step we bound the difference between the sum and the integral using the fact that the integrand has total variation 1. For each j, when $m \to \infty$, our expected value has the form

$$E[\textit{llps among } j \textit{ last keys}] = c_j m + O(1).$$

The sequence of constants c_j is bounded above by 1 and is monotonically increasing; thus converges, and has a limiting value which is given by

$$\lim_{j\to\infty} c_j = 1 - \int_0^1 \prod_{i=1}^{\infty} (1-y^i)\, dy. \tag{2.6}$$

To compute this integral we use Euler's identity related to partitions, also a special case of Jacobi's identity [1,12]

$$\prod_{i=1}^{\infty} (1-x^i) = \sum_{i=-\infty}^{\infty} (-1)^i x^{(3i^2+i)/2}.$$

We substitute this into (2.6) and evaluate the integral to obtain

$$E[\textit{llps}] = m\left\{1-\sum_{i=-\infty}^{\infty} \frac{2(-1)^i}{3i^2+i+2}\right\} + O(1), \tag{2.7}$$

Thus we have the following.

THEOREM *The expected length of the longest probe sequence using uniform probing in a full hash table of size m is given by (2.7).* This can be evaluated by standard summation techniques [11] to give

$$E[\textit{llps}] = \left\{1-\left(\frac{12}{23}\right)^{1/2} \frac{\sinh(\pi\left(\frac{23}{36}\right)^{1/2})}{\sinh^2(\pi\left(\frac{23}{36}\right)^{1/2})+\frac{1}{4}}\right\} \times m + O(1),$$

$$= 0.631587464\ldots \times m + O(1).$$

Table I shows the exact expected *llps* for some table sizes as computed

using (2.3) (numbers have been rounded to four decimal places).

TABLE I

m	$E[llps]$	$\frac{E[llps]}{m}$
5	3.3696	0.67392
10	6.5226	0.65226
40	25.4676	0.63669
100	63.3624	0.63362
∞	∞	0.631587464...

2.2. *Partly Filled Tables*

When a hash table of size m is partly filled with n entries, the probability of taking k of fewer accesses to insert a new key is

$$Pr\{taking \leqslant k \ accesses\} = 1-\frac{n^{\underline{k}}}{m^{\underline{k}}}. \tag{2.8}$$

The probability that the *llps* is less than or equal to k when inserting the first n keys is

$$Pr\{llps \leqslant k\} = \prod_{i=0}^{n-1}(1-i^{\underline{k}}/m^{\underline{k}}), \tag{2.9}$$

and the expected value of the *llps* is

$$E[llps] = \sum_{k\geqslant 0} 1-\prod_{i=0}^{n-1}\left(1-\frac{i^{\underline{k}}}{m^{\underline{k}}}\right).$$

This summation involving products and descending factorials seems very difficult to solve for fixed $\alpha=n/m$ and $m\to\infty$. Consequently we will try a different approach based on finding the median of the distribution of the *llps* when $m\to\infty$ and then showing that the expected value asymptotically coincides with the median. Taking logarithms in (2.9) we have

$$\ln(Pr\{llps \leqslant k\}) = \sum_{i=0}^{n-1}\ln(1-i^{\underline{k}}/m^{\underline{k}}). \tag{2.10}$$

Let $\alpha=n/m$ be the occupation factor of the table as $n,m\to\infty$. Using the Maclaurin series for $\ln(1-x)$ to expand the right side of (2.10) we find

$$\ln(Pr\{llps\leqslant k\}) = -\left[\sum_{i=0}^{n-1}\frac{i^{\underline{k}}}{m^{\underline{k}}}+\frac{1}{2}\sum_{i=0}^{n-1}\left(\frac{i^{\underline{k}}}{m^{\underline{k}}}\right)^2+\ldots\right]. \tag{2.11}$$

We can simplify this using the summation formula

$$\sum_{i=0}^{n} i^{\underline{k}} = k!\sum_{i=0}^{n}\binom{i}{k} = k!\binom{n+1}{k+1} = \frac{(n+1)^{\underline{k+1}}}{k+1}$$

to obtain

$$-\ln(Pr\{llps\leqslant k\}) = \frac{n^{\underline{k+1}}}{(k+1)m^{\underline{k}}}+\frac{1}{2}\sum_{i-0}^{n-1}\left(\frac{i^{\underline{k}}}{m^{\underline{k}}}\right)^2+\ldots$$

which reduces asymptotically to

$$-\ln(Pr\{llps\leqslant k\}) = \frac{n\alpha^k}{k+1}+O\left(\frac{n\alpha^{2k}}{k}\right).$$

Let j denote the median of the above probability distribution for fixed α. Thus $Pr\{llps \leqslant j\} = \frac{1}{2}$ and

$$\ln 2 = \frac{n\alpha^j}{j+1} + O(\frac{n\alpha^{2j}}{j}). \tag{2.12}$$

With α fixed, $0<\alpha<1$ and $n,m \to \infty$ we conclude that

$$\frac{\alpha^j}{j} = \Theta(m^{-1}).$$

This leads to

$$j = \Theta(\log m) \tag{2.13}$$

and thus (2.12) becomes

$$\log 2 = \frac{m\alpha^{j+1}}{j+1} + O(\frac{\log m}{m}). \tag{2.14}$$

Simple algebraic manipulations of this equation yield

$$\ln 2 = \frac{-m\ln\alpha}{-(j+1)\ln\alpha e^{-(j+1)\ln\alpha}} + O(\frac{\ln m}{m}). \tag{2.15}$$

Let $w(x)$ be the solution of the trascendental equation $w(x)e^{w(x)} = x$. We know that $w(x) = \ln x - \ln(\ln x) + o(1)$ [4, ch 2.4] when $x \to \infty$. If we let $w(y) = -(j+1)\ln\alpha$, equation (2.15) becomes

$$\ln 2 = \frac{-m\ln\alpha}{y} + O(\frac{\ln m}{m})$$

or equivalently,

$$y = -m\log_2\alpha(1+O(\frac{\ln m}{m})).$$

Thus, $w(y) = -(j+1)\ln\alpha$ tells us

$$j = \frac{w(-m\log_2\alpha)}{-\ln\alpha} - 1 + O(\frac{\ln m}{m}). \tag{2.16}$$

To complete our work, we wish to show that the expected value of the *llps* is asymptotically equivalent to its median j, i.e.

$$\lim_{m\to\infty} \frac{E[llps]}{j} = 1$$

This expected value is defined by

$$E[llps] = \sum_{k=0}^{\infty} (1 - Pr\{llps \leqslant k\}). \tag{2.17}$$

To prove the equivalence between the median j and the expected value, we proceed to bound the $E[llps]$. To obtain a lower bound we begin by noting that

$$E[llps] > \sum_{k=0}^{j-\ln j} (1 - Pr\{llps \leqslant k\}) \tag{2.18}$$

$$> (j-\ln j) \times [1 - Pr\{llps \leqslant j-\ln j\}]$$

$$> (j-\ln j) \times [1 - e^{\ln Pr\{llps \leqslant j-\ln j\}}].$$

Substituting $j-\ln j$ for k in (2.12) we have

$$E[llps] > (j-\ln j) \times [1 - e^{-\frac{m\alpha^{j-\ln j+1}}{j-\ln j+1} + O(m^{-1})}].$$

Using (2.14) this becomes

$$E[llps] > (j-\ln j)\times[1-e^{-\frac{\ln 2\,(j+1)\alpha^{-\ln j}}{j-\ln j+1}+O(m^{-1})}]$$

$$> (j-\ln j)\times[1-2^{-\frac{(j+1)\alpha^{-\ln j}}{j-\ln j+1}+O(m^{-1})}].$$

Since

$$\frac{(j+1)\alpha^{-\ln j}}{j-\ln j+1} \geqslant \alpha^{-\ln j} = j^{-\ln\alpha}$$

we derive

$$E[llps] > (j-\ln j)\times[1-2^{-j^{-\ln\alpha}+O(m^{-1})}]. \tag{2.19}$$

But (2.13) tells us that $j\to\infty$ as $n,m\to\infty$ and therefore

$$\lim_{m\to\infty}\frac{E[llps]}{j} \geqslant 1. \tag{2.20}$$

To obtain an upper bound on the $E[llps]$ we derive from (2.7)

$$E[llps] < j+\sum_{k\geqslant j} 1-e^{\ln Pr\{llps\leqslant k\}} \tag{2.21}$$

using the inequality $1-e^{-x}<-x$ we obtain

$$E[llps] < j-\sum_{k\geqslant j}\ln Pr\{llps\leqslant k\}$$

substituting (2.12)

$$E[llps] < j+\sum_{k\geqslant j}\frac{m\alpha^{k+1}}{k+1} < j+\frac{m\alpha^{j+1}}{(1-\alpha)(j+1)}$$

and finally using (2.14)

$$E[llps] < j+\frac{\ln 2}{1-\alpha}+O(\frac{\ln m}{m}).$$

We conclude that

$$\lim_{m\to\infty}\frac{E[llps]}{j} \leqslant 1$$

and this in connection with (2.20) gives us

$$\lim_{m\to\infty}\frac{E[llps]}{j} = 1.$$

Therefore we have the following.

THEOREM *The expected llps under uniform probing for hash tables of size m, partly filled with n entries is given by*

$$E[llps] \approx \frac{w(-m\log_2\alpha)}{-\ln\alpha} - 1 \tag{2.22}$$

$$\approx -\log_\alpha m - \log_\alpha(-\log_\alpha m) + O(1).$$

Table II shows some exact values of $E[llps]$ calculated using formulas (2.9) and (2.17) rounded to four decimal places. The bracketed values in the table give the value of the average *llps* calculated from (2.22) and rounded to two decimal places. Note that for the values shown in the table, α is close to 1; thus the log-log term in (2.22) makes a signigicant contribution.

TABLE II

α	m 20	100	1000	10000	10^6
80%	5.4733	10.8611	19.4495		
	(5.61)	(10.38)	(18.33)	(26.99)	(45.36)
90%	7.6583	17.8671	35.7468		
	(9.03)	(18.16)	(34.23)	(52.18)	(90.72)
95%	9.5475	26.7479	62.2090		
	(13.04)	(29.38)	(60.51)	(96.43)	(174.72)
99%	(N/A)	48.2629	187.6690		
		(69.81)	(196.74)	(364.67)	(751.13)

3. *Lower Bounds on the Average Length of the Longest Probe Sequence*

The minimax hash coding problem can be stated as follows: given a set of keys, find the insertion order which minimizes the maximum number of accesses to locate any single element. This minimum maximum number of accesses will be called the *minimax* value. A lower bound on the minimax is also a lower bound on the average length of the longest probe sequence (*llps*) of any other open-addressing scheme.

3.1. *Full Tables*

We first consider the case of a full table, i.e., $\alpha=1$. We will require that for any key the hashing function of this model produces an independent random sequence of probes from a discrete rectangular distribution in $(1,m)$. This model is called random probing, and differs slightly from the usual hashing schemes in that we allow probe positions to be repeated.

A necessary, though not sufficient, condition to generate a full hashing table requiring at most k accesses is that for each location i from 1 to m, there must be some key for which i is among the key's first k probe positions. For each table, the smallest k satisfying this condition is a lower bound on the corresponding minimax value. Consequently

$$E[minimax] \geqslant E[smallest\ k].$$

Given k, the probability of all the table positions $(1,m)$ appearing among $k \times m$ probe position is an occupancy distribution, also known as Arfwedson's distribution [2,16,13,5] denoted by $A_m(k \times m)$. The distribution is given by

$$A_m(k \times m) = \sum_{i=0}^{m} (-1)^i \binom{m}{i} (1-i/m)^{km} \tag{3.1}$$

$$= m!\, m^{-km} \left\{ {km \atop m} \right\}$$

where the brace brackets denote Stirling numbers of the second kind [1]. Feller [5, Ch IV.2] gives the approximation

$$A_m(k \times m) = e^{-me^{-k}}.$$

The expected value of the smallest k, i.e. the average lower bound for the distribution above, is

$$E[smallest\ k] = \sum_{k=1}^{\infty} k\,[A_m(k \times m) - A_m((k-1) \times m)], \tag{3.2}$$

$$= \sum_{k=0}^{\infty} (1-A_m(k \times m)) \approx \sum_{k=0}^{\infty} (1-e^{-me^{-k}}).$$

Let

$$Q(m) = \sum_{k=0}^{\infty} (1-e^{-me^{-k}}).$$

By inspection, we find the functional equation

$$Q(m) = Q(me) - 1 + e^{-m}. \tag{3.3}$$

The general solution of this functional equation [8, 14 Ch.5.2.2] is

$$Q(m) = \ln m + \gamma + \frac{1}{2} + P(\ln m) + O(e^{-m}), \tag{3.4}$$

where γ=0.57721... is Euler's constant, and $P(x)$ is a periodic function with period 1 and magnitude

$$\left|P(x)\right| \leqslant 0.0001035.$$

Changing the constant so that it includes the periodic factor we have

THEOREM *The average llps of a full table under random probing hashing is bounded below by*

$$E[llps] \geqslant ln\ m + 1.077 + o(1).$$

Recently Rivest [15] proved that it is possible to construct a table having a minimax value that is $O(ln\ m)$. Consequently both upper and lower bounds are optimal within a constant factor. Simulation results reported by Gonnet and Munro [7] indicate that the lower bound is tight.

Table III shows the exact lower bounds computed from (3.2) (rounded to four decimal places); bracketed number indicate the asymptotic value computed with (3.5).

TABLE III

m	lower bound	
5	2.6956	(2.6867)
10	3.3761	(3.3798)
20	4.0737	(4.0729)
50	4.9893	(4.9892)
1000	(7.9850)	

3.2. *Partially Filled Tables.*

When a table is not full, we define the random variable T_k to be the number of different table positions which are accesible in k or fewer probes based on the n keys. If $T_k < n$ we cannot construct the table with at most k accesses for any given key; construction of such a table requires $T_k \geqslant n$ (and even this condition is not sufficient). Using the same reasoning as in the previous section, we conclude

$$Pr\{not\ succeeding\ with\ k\ accesses\} \geqslant Pr\{T_k < n\}. \tag{3.6}$$

The distribution of T_k is also an occupancy distribution [16,13]; its expected value is given by

$$E[T_k] = m[1-(1-1/m)^{nk}] = m(1-e^{-k\alpha})+O(1), \tag{3.7}$$

and its variance is

$$\begin{aligned} var(T_k) &= m(1-1/m)^{nk}+m(m-1)(1-2/m)^{nk}-m^2(1-1/m)^{2nk} \\ &= m[e^{-k\alpha}-(1+\alpha^2)e^{-2k\alpha})+O(1). \end{aligned} \tag{3.8}$$

For fixed α and k, when $n,m\to\infty$ we can apply Chebyshev's inequality [5]

$$Pr\{|X-\bar{x}| \geqslant t\} \leqslant \frac{\sigma^2}{t^2} \qquad (t>0)$$

to obtain

$$\begin{aligned} Pr\{T_k \geqslant n\} &\leqslant Pr\{|T_k-E[T_k]| \geqslant (n-E[T_k])\} \\ &\leqslant \frac{var(T_k)}{(n-E[T_k])^2} \end{aligned}$$

Using (3.7) and (3.8) we find

$$Pr\{T_k \geqslant n\} \leqslant \frac{m(e^{-k\alpha}-(1+\alpha^2)e^{-2k\alpha})+O(1)}{(n-m(1-e^{-k\alpha})+O(1))^2} = O(m^{-1})$$

if and only if $E[T_k]<n$. Finally then

$$\lim_{m\to\infty} Pr\{T_k \geqslant n\} = 0 \quad iff\ E[T_k]<n.$$

Similarly, we have

$$Pr\{T_k \leqslant n\} = 1+O(m^{-1}) \quad iff\ E[T_k]>n.$$

(Note that $E[T_k]$ cannot equal n in realistic cases, since (3.6) shows $E[T_k]$ cannot be an integer for $nk>1$.) The expected value of the lower bound on the minimax is the smallest integer k for which $E[T_k]>n$. According to (3.6), this means

$$m(1-e^{-k\alpha}) > n$$

and a little algebra yields the equivalent condition

$$k \geqslant -\alpha^{-1}\ln(1-\alpha).$$

Thus we have the following.

THEOREM *The expected value of the lower bound on the minimax for partly filled hash tables with occupation factor* α *is*

$$k = \lceil -\alpha^{-1}\ln(1-\alpha)\rceil.$$

Note that the distribution is single valued when $n,m\to\infty$.

Table IV shows the ranges of the occupation factor α which correspond to an expected lower bound k.

TABLE IV

Range of α	Lower Bound
$0<\alpha<0.7968$	2
$0.7968<\alpha<0.94048$	3
$0.94048<\alpha<0.98017$	4
$0.98017<\alpha<0.993023$	5

4. *Separate Chaining*

Separate chaining hashing (also called direct chaining or separate overflow chaining) forms a linked list with all keys that hash to each location. The length of each linked list is the number of keys that hash to a particular location. This length follows a simple binomial distribution: the probability of a key hashing to a particular location is $1/m$ and so the probability of k of the n keys hashing to the location is

$$Pr\{list\ of\ length\ k\} = \binom{n}{k} m^{-n}(m-1)^{n-k} \sim \frac{e^{-\alpha}\alpha^k}{k!} \tag{4.1}$$

where $\alpha = n/m$ is the load factor. The final expression in (4.1) is the Poisson approximation to the binomial distribution [5], which for fixed α is very accurate. We will therefore consider our model of separate chaining as one which produces linked lists with independent Poisson-distributed lengths.

Let $e_i(\alpha)$ be the cumulative distribution of (4.1) i.e.

$$e_i(\alpha) = \sum_{k=0}^{i} \frac{e^{-\alpha}\alpha^k}{k!} \tag{4.2}$$

and define $d_i(\alpha) = 1 - e_i(\alpha)$. Since we are assuming each linked list is independent of the rest, we have that

$$Pr\{llps \leqslant i\} = e_i(\alpha)^m \tag{4.3}$$

hence

$$E[llps] = \sum_{i=0}^{\infty} (1 - e_i(\alpha)^m). \tag{4.4}$$

This expression for the expected value does not seem to have a simple solution. Therefore we will use the same kind of approach we used in section 2.2. If j is the median of the distribution of the *llps*, we have from (4.3) that

$$Pr\{llps < j\} = \frac{1}{2} = e_j(\alpha)^m = (1 - d_j(\alpha))^m \tag{4.5}$$

Taking logarithms gives us

$$-\ln 2 = m \ln(1 - d_j(\alpha))$$

and expanding with the Maclaurin series for $\ln(1-x)$ yields

$$-\ln 2 = -m[d_j(\alpha) + d_j(\alpha)^2/2 + \ldots].$$

Clearly then

$$d_j(\alpha) = \Theta(m^{-1}),$$

and

$$\frac{\ln 2}{m} = d_j(\alpha) + O(m^{-2}).$$

Since (4.2) tells us

$$d_j(\alpha) = 1 - \sum_{k=0}^{j} \frac{e^{-\alpha}\alpha^k}{k!} = \sum_{k=j+1}^{\infty} \frac{e^{-\alpha}\alpha^k}{k!},$$

equation (4.6) becomes

$$\frac{\ln 2}{m} = \frac{e^{-\alpha}\alpha^{j+1}}{(j+1)!}\left(1 + \frac{\alpha}{j+2} + \frac{\alpha^2}{(j+2)(j+3)} + \ldots\right) + O(m^{-2}),$$

which we can write as

$$\ln 2 = m\frac{e^{-\alpha}\alpha^{j+1}}{(j+1)!} + O(1/j) + O(1/m). \tag{4.7}$$

Thus for $\alpha=1$ we have

$$(j+1)! \approx \frac{m}{e \ln 2}$$

or

$$j \approx \Gamma^{-1}\left[\frac{m}{e \ln 2}\right] - 2 = \Gamma^{-1}(m) + O(1) \tag{4.8}$$

where Γ^{-1} is the inverse of the usual gamma function which satisfies $\Gamma(k+1) = k\,\Gamma(k)$. For $\alpha \neq 1$ we have

$$\frac{\Gamma^{-1}(m)}{j} = 1 + O(1/\ln j),$$

but as the O term indicates, this asymptotic form is approached much more slowly in files of a reasonable size than in the case $\alpha=1$.

To summarize, we have the following.

THEOREM *In a separate chaining model whose linked lists have lengths distributed independently according to a Poisson distribution, the median llps is given by (4.8) for a full table, and by (4.9) for a partly full table.*

Following the same approach we used in Section 2.2, we can prove that this median asymptotically coincides with the expected value; that is,

$$\lim_{m\to\infty} \frac{E[llps]}{j} = 1.$$

Consequently

$$E[llps] \approx \Gamma^{-1}(m) \tag{4.10}$$
$$\approx \frac{\ln m}{\ln \ln m}.$$

Table V shows some values of the expected *llps* calculated using (4.4); the bracketed numbers give values for the median j calculated using (4.8). All numbers are rounded off to four decimal places.

TABLE V

m	$\alpha=\frac{1}{2}$	$\alpha=1$		$\alpha=2$
24	2.2360	3.3347	(3.5650)	5.1688
120	3.0363	4.3359	(4.6209)	6.4423
720	3.8444	5.3432	(5.6568)	7.7055
5040	4.6541	6.3493	(6.6823)	8.9519

5. *Conclusions*

Table VI summarizes the new results concerning the average length of longest probe sequence, together with known results for the average number of accesses. For completeness we present various hashing schemes.

TABLE VI

Algorithm	Average Number of accesses		Average Length of the Longest Probe Sequence	
	Full Table	$\alpha = n/m$	Full Table	$\alpha = n/m$
Open Addressing (uniform probing) [14 6.4.D]	$\ln m - 1 + \gamma + o(1)$	$-\frac{\ln(1-\alpha)}{\alpha}$	$0.63158... \times m + O(1)$	$-\log_\alpha m + O(\log_\alpha(-\log_\alpha m))$
Optimal reordering to minimize average [7]	$\geqslant 1.668...$ [6] ~ 1.83 (*e*)	$\geqslant 2 - \frac{1-e^{-\alpha}}{\alpha}$	$\Theta(\ln m)$ (*e*)	?
Optimal reordering to minimize worst case [7,15]	~ 1.83 (*e*)	~ 1.83 (*e*)	$\geqslant \ln m + O(1)$ $\Theta(\ln m)$[15]	$\lceil -\frac{\ln(1-\alpha)}{\alpha} \rceil$
Separate chaining (direct chaining) [14 6.4]	1.5	$1 + \alpha/2$	$\approx \Gamma^{-1}(m)$	$\approx \Gamma^{-1}(m)$

The results marked by (*e*) are experimental results found by simulation.

The first row deals with results for open-addressing hashing [14 Ch.6.4]. Schemes of this type resolve collisions by computing new probe positions until all the table has been searched.

The next row shows results obtained from the optimal reordering during insertion which minimizes the average number of accesses. These results are interesting since they provide lower bounds for all possible reordering schemes which use open-addressing.

Similarly, the optimal reordering of insertions to minimize the length of the longest probe sequence (*llps*) provides lower bounds on the average *llps* for any open-addressing scheme. The results for full tables show that we may do even better than binary search for the *llps*, since we have a length $\ln n$ rather than $\log_2 n$, while the $O(1)$ for the average number of accesses is preserved. For partly filled tables we find an integer (depending only on α) coming from a very familiar formula, i.e. the average number of accesses for open-addressing.

The last row shows the results for the separate (or direct) chaining technique, described in Section 4. As we showed there, the average *llps* for any α depends on the inverse of the gamma function. This function grows very slowly ($O(\ln m / \ln \ln m)$) and hence displays a desirable property of separate chaining hashing.

Except for full tables in simple open addressing, these expected *llps* are very slow growing functions (logarithm and inverse factorial). For some critical applications we may use minimax reordering schemes, and thereby obtain an $O(1)$ behaviour in general for the average, and a worst case which is better than binary search. If we can afford the use of pointers in the table, these results show an advantageous property of separate chaining hashing.

All the results presented are asymptotic with respect to the size of the table. However the calculated values we give in Tables I-V show that these asymptotic results are close approximations when we consider more reasonably sized tables.

ACKNOWLEDGMENTS. The author wishes to acknowledge the anonymous referees for many helpful suggestions that improved this manuscript substantially.

REFERENCES

[1] Abramovitz, M., and Stegun, I.A., *Handbook of Mathematical Functions*. Dover Publications, New York, 1964.

[2] Arfwedson, G., A Probability Distribution Connected with Stirling's Second Class Numbers. Skandinavisk Aktuarietidskrift, 34-3 (1951), pp. 121-132.

[3] Brent, R.P., Reducing the Retrieval Time of Scatter Storage Techniques, CACM 16, 2 (Feb. 1973), pp. 105-109.

[4] De Bruijn, N.G., *Asymptotic Methods in Analysis*, North Holland, Amsterdam, 1970.

[5] Feller, W., *An Introduction to Probability Theory and its Applications*, John Wiley, New York, 1957, Vol I. 3rd Edition.

[6] Gonnet, G.H. Average Lower Bounds for Open-Addressing Hash Coding, Proceedings of the Conference on Theoretical Computer Science, University of Waterloo, Waterloo, Ontario, Canada, (Aug. 1977), pp. 159-162.

[7] Gonnet, G.H., and Munro, J.I., Efficient Ordering of Hash Tables, to appear in SIAM Journal on Computing.

[8] Gonnet, G.H., Notes on the Derivation of Asymptotic Expressions from Summations, Information Processing Letters, Vol. 7-4 (June 1978), pp. 165-169.

[9] Guibas, L.J., The Analysis of Hashing Techniques that Exhibit k-ary Clustering, J.ACM, Vol. 25-4, (Oct. 1978), pp. 544-555.

[10] Guibas, L.J., and E. Szemeredi, The Analysis of Double Hashing, Journal of Computer and System Sciences, Vol. 16-2, (April 1978), pp. 226-274.

[11] Hansen, E.R. *A Table of Series and Products*, Prentice Hall, Englewood Cliffs, NJ, 1975.

[12] Hardy, G.H., and Wright, E.M. *An Introduction to the Theory of Numbers*, Oxford at the Clarendon Press, 1959.

[13] Johnson, N.L., and Kotz, S., *Distributions in Statistics*, Houghton Mifflin, Boston, 1969, Vol. I (Discrete Distributions).

[14] Knuth, D.E., *The Art of Computer Programming*, Addison-Wesley, Don Mills, (1973), Vol. III (Sorting and Searching).

[15] Rivest, R.L., Optimal Arrangement of Keys in a Hash Table, JACM, Vol. 25-2, (April 1978), pp. 200-209.

[16] Stevens, W.L., Significance Grouping, Annals of Eugenics, Vol. 8 (1937), pp.57-69.