[Typeset by Helmut Prodinger, November 1997]

NOTES ON "OPEN" ADDRESSING

[MY FIRST ANALYSIS OF AN ALGORITHM, ORIGINALLY DONE DURING SUMMER 1962 IN MADISON.]

Don Knuth 7/22/63

1. Introduction and Definitions. Open addressing is a widely–used technique for keeping "symbol tables." The method was first used in 1954 by Samuel, Amdahl, and Boehme in an assembly program for the IBM 701. An extensive discussion of the method was given by Peterson in 1957 [1], and frequent references have been made to it ever since (e.g. Schay and Spruth [2], Iverson [3]). However, the timing characteristics have apparently never been exactly established, and indeed the author has heard reports of several reputable mathematicians who failed to find the solutions after some trial. Therefore it is the purpose of this note to indicate one way by which the solution can be obtained.

We will use the following abstract model to describe the method: N is a positive integer, and we have an array of N variables $x_1, x_2, \ldots, x_N$. At the beginning, $x_i = 0$, for $1 \le i \le N$.

To "enter the k–th item in the table," we mean that an integer a_k is calculated, $1 \le a_k \le N$, depending only on the item, and the following process is carried out:

(1) Set $j = a_k$.
(2) "The comparison step." If $x_j = 0$, set $x_j = 1$ and stop; we say "the k–th item has fallen into position x_j."
(3) If $j = N$, go to step 5.
(4) Increase j by 1 and return to step 2.
(5) "The overflow step." If this step is entered twice, the table is full, i.e. $x_i = 1$ for $1 \le i \le N$. Otherwise set j to 1 and return to step 2.

Observe the cyclic character of this algorithm.

We are concerned with the statistics of this method, with respect to the number of times the comparison step must be executed. More precisely, we consider all of the N^k possible sequences $a_1, a_2, \ldots, a_k$ to be equally probable, and we ask, **"What is the probability that the comparison step is used precisely m times when the k–th item is placed?"**

2. Non–overflow (self–contained) sequences. Let $\left[{n \atop k}\right]$ denote the number of sequences $a_1, a_2, \ldots, a_k$ $(1 \le a_i \le n)$ in which no overflow step occurs during the entire process of placing k items, if the algorithm is used for $N = n$. (By convention, we set $\left[{n \atop 0}\right] = 1$.)

Lemma 1. *If* $0 \leq k \leq n+1$, *then* $\left[{n \atop k}\right] = (n+1)^k - k(n+1)^{k-1}$.

Proof. This proof is based on the fact that $\left[{n \atop k}\right]$ is precisely the number of sequences $b_1, b_2, \ldots, b_k$ $(1 \leq b_i \leq n+1)$ in which, if the algorithm is carried out for $S = n+1$, then x_{n+1} at the end of the operation. This follows because every sequence of the former type is one of the latter, and conversely the condition implies in particular that $1 \leq b_i \leq n$, and that no overflow step occurs.

But sequences of the latter type are easily enumerated, because the algorithm has circular symmetry; of the $(n+1)^k$ possible sequences $b_1, b_2, \ldots, b_k$, exactly $k/(n+1)$ of these leave $x_{n+1} \neq 0$. This shows that

$$\begin{bmatrix} n \\ k \end{bmatrix} = (n+1)^k \left(1 - \frac{k}{n+1}\right) .$$

3. Sequential pile–up. How many sequences $a_1, \ldots, a_{k-1}$ $(1 \leq a_i \leq N)$ leave

$$X_{N-t-1} = 0,\ X_{N-t} = \cdots = X_{N-1} = 1,\ X_N = 0\ ?$$

Let this number be denoted by $Q(N, k, t)$.

Lemma 2. *If* $0 < k \leq N$, $0 \leq t < N-1$, *then* $Q(N,k,t) = \binom{k-1}{t}\left[{N-t-2 \atop k-t-1}\right]$.

Proof. In order to construct such a sequence, we have a subsequence of t items which fall into the range x_{N-t} through x_{N-1}; there are $\left[{t \atop t}\right]$ such sequences. The remaining terms form a subsequence of $k-t-1$ items which all land in the range x_1 through x_{N-t-2}; there are $\left[{N-t-2 \atop k-t-1}\right]$ such sequences. Finally, there are $\binom{k-1}{t}$ ways to put these two subsequences together. This completes the proof.

Notice that the stated formula for $Q(N, k, t)$ is valid also for the excluded case $t = N-1$, if we adopt the convention that

$$\begin{bmatrix} -1 \\ 0 \end{bmatrix} = 1.$$

4. The probability $P(N, k, m)$**.** Let $P(N, k, m)$ be the probability that m comparison steps are required to place the k–th item, i.e. step 2 of the algorithm is entered m times.

Lemma 3. *The number of sequences* $a_1, \ldots, a_k$ $(1 \leq a_i \leq N)$ *in which the* k*–th item falls in position* x_N *after precisely* m *comparisons, is*

$$\sum_{i=m}^{N} Q(N,k,i-1) = \begin{bmatrix} N-1 \\ k-1 \end{bmatrix} - \sum_{i=1}^{m-1} Q(N,k,i-1), \quad \text{for } 1 \leq k \leq N.$$

Proof. We must have $a_k = N - m + 1$; and after the first $k-1$ steps, we must have $x_i = 1$ for $N - m + 1 \leq i \leq N$, and also $x_N = 0$. Therefore by Lemma 2, the stated formula is obvious, in lieu of the fact that

$$\sum_{i=1}^{N} Q(N,k,i-1) = \begin{bmatrix} N-1 \\ k-1 \end{bmatrix}$$

(the number of sequences which leave $x_N = 0$).

Lemma 4.

$$P(N,k,m) = \frac{1}{N^{k-1}} \sum_{i=m}^{N} Q(N,k,i-1) = 1 - \frac{k-1}{N} - \frac{1}{N^{k-1}} \sum_{i=1}^{m-1} Q(N,k,i-1).$$

Proof. The position x_N in Lemma 3 can be changed to x_i for any other i without affecting the result, by symmetry. There are N^k possible sequences $a_1, \ldots, a_k$ $(1 \le a_i \le N)$, each assumed to be equally probable; hence $P(N,k,m)$ is the appropriate fraction of these sequences, and the result is immediate from Lemma 3.

Now we let $R(N,k,t) = Q(N,k,t-1)/N^{k-3}(N-k)$. By Lemmas 1 and 2, we find that for $1 \le t < N$, $1 \le k < N$,

$$R(N,k,t) = \binom{k-1}{t-1}(N-t)^{k-t-1}(N-k)t^{t-2} \Big/ \Big(N^{k-3}(N-k)\Big)$$

so that

$$R(N,k,t) = \binom{k-1}{t-1} \left(\frac{t}{N}\right)^{t-2} \left(1 - \frac{t}{N}\right)^{n-t-1}.$$

Theorem 1. *If* $1 \le k, m \le N$, *then*

$$P(N,k,m) = 1 - \frac{k-1}{N} - \frac{N-k}{N^2} \sum_{t=1}^{m-1} R(n,k,t).$$

Proof. This formula is immediate from Lemma 4, except in the case $k = N$. But since $P(N,N,m)$ is clearly $1/N$, the above formula holds in all cases.

5. Calculation of the mean. We are interested in the expected number of comparison steps to place the k–th item. This is

$$M(N,k) = \sum_{m=1}^{N} m P(N,k,m).$$

To evaluate $M(N,k)$ in "closed" form, we first observe that $\sum_{m=1}^{N} P(N,k,m) = 1$ which implies (by Theorem 1)

$$1 = N - k + 1 - \frac{N-k}{N^2} \sum_{t=1}^{N} (N-t) R(N,k,t);$$

hence, if $k < N$, $\sum_{t=1}^{N} (1 - \frac{t}{N}) R(N,k,t) = N$. This implies the identity

$$\sum_{t=1}^{k} \binom{k-1}{t-1} \left(\frac{t}{N}\right)^{t-2} \left(1 - \frac{t}{n}\right)^{k-t} = N, \qquad (*)$$

which is rather difficult to derive directly. Actually the latter formula can be derived independently using Abel's generalized binomial theorem:

$$(x+z)^n = \sum_{r=0}^{n} \binom{n}{r} x(x-ry)^{r-1}(z+ry)^{n-r}$$

which is valid for $x \neq 0$ and arbitrary y, z. Put $x = 1/N$, $Y = -1/N$, $z = 1 - x$ and $n = k-1$, to obtain

$$1 = \sum_{r=0}^{k-1} \binom{k-1}{r} \frac{1}{N} \left(\frac{r+1}{N}\right)^{r+1} \left(1 - \frac{r+1}{N}\right)^{k-r-1}$$

and the required formula $(*)$ follows with $t = r+1$. It is therefore valid for all $k, N > 0$ and in particular for $k = N$.

Now let N be fixed, and let

$$b_k = \sum_{t=1}^{N} \left(1 - \frac{t}{n}\right)^2 R(N,k,t).$$

We have the identity

$$\frac{k-t}{k-1} R(N,k,t) = \left(1 - \frac{t}{N}\right) R(N,k-1,t),$$

which can be rewritten in the form

$$\left(1 - \frac{t}{N}\right) R(N,k,t) = \left(1 - \frac{k}{N}\right) R(N,k,t) + \frac{k-1}{N} \left(1 - \frac{t}{N}\right) R(N,k-1,t).$$

Hence, by $(*)$,

$$b_k = N - k + \frac{k-1}{N} b_{k-1}$$

and using the value $b_1 = N-1$, we find

$$b_{k+1} = N - \frac{k!}{N^k} \left(1 + N + \frac{N^2}{2!} + \cdots + \frac{N^k}{k!}\right), \quad \text{for } k \geq 0.$$

Let us write, for convenience,

$$E_r(N,k) = 1 + \binom{r+1}{r} \frac{k-1}{N} + \binom{r+2}{r} \frac{(k-1)(k-2)}{N^2} + \cdots + \binom{r+k-1}{r} \frac{(k-1)!}{N^{k-1}}.$$

Then

$$b_k = N - 1 - \frac{k-1}{N} E_0(N, k-1).$$

We have

$$
\begin{aligned}
M(N,k) &= \sum_{m=1}^{N} mP(N,k,m) \\
&= \frac{N(N+1)}{2}\left(1-\frac{k-1}{N}\right) - \frac{N-k}{N^2}\sum_{t=1}^{N}\left(\frac{N(N+1)}{2}-\frac{t(t+1)}{2}\right)R(N,k,t) \\
&= \frac{1}{2}\Bigg\{N(N+1)\left(1-\frac{k-1}{N}\right) \\
&\qquad -\left(1-\frac{k}{N}\right)\sum_{t=1}^{N}\left[(2N+1)\left(1-\frac{t}{N}\right)-N\left(1-\frac{t}{N}\right)^2\right]R(N,k,t)\Bigg\}.
\end{aligned}
$$

Putting everything together, we therefore obtain

Theorem 2.

$$
\begin{aligned}
M(N,k) &= \frac{1}{2}\left(1+k-(k-1)\left(1-\frac{k}{N}\right)E_0(N,k-1)\right) \\
&= \frac{1}{2}\Big(1+kE_0(N,k)-(k-1)E_0(N,k-1)\Big) \\
&= \frac{1}{2}\Big(1+E_1(N,k)\Big)
\end{aligned}
$$

As an example, let $N=5$. We find the following values for $P(N,k,m)$:

	m					
$P(5,k,m)$	1	2	3	4	5	$M(5,k)$
1	1	0	0	0	0	1
2	$\frac{4}{5}$	$\frac{1}{5}$	0	0	0	$\frac{6}{5}$
3	$\frac{3}{5}$	$\frac{7}{25}$	$\frac{3}{25}$	0	0	$\frac{38}{25}$
4	$\frac{2}{5}$	$\frac{34}{125}$	$\frac{1}{5}$	$\frac{16}{125}$	0	$\frac{257}{125}$
5	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	3

6. The standard deviation. Let $\sigma(N,k)$ be the standard deviation of the distribution $P(N,k,m)$. Using methods similar to those of the preceding section, and somewhat more pain, the standard deviation can be obtained. We will merely quote the results here.

Theorem 3.

$$
\sum_{t=1}^{N}\left(1-\frac{t}{N}\right)^3 R(N,k,t) = N+(k-1)!\sum_{i=0}^{k}\left(\binom{i+3}{2}-2k-3\right)\frac{1}{(k-i)!N^i}
$$

$$
\sum_{m=1}^{N} m^2P(N,k,m) = \frac{1}{6}+E_3(N,k)-\frac{2}{3}E_2(N,k)+\frac{1}{2}E_1(N,k)
$$

The reader may at least verify this theorem in the case $N=5$. The standard deviation may now be obtained from the latter formula.

7. Searching the table. Once the table has been constructed, let k be the number of items in it. A similar algorithm can be used to search the table, and the number of comparisons needed to find a given item is the same as the number of comparisons which were needed to enter it in the first place.

If all items in the table are referred to with equal probability, the average number of comparisons needed when searching the table is

$$T(N,k) = \frac{1}{k}\sum_{t=1}^{k} M(N,t) = \frac{1}{2}\Big(1 + E_0(N,k)\Big),$$

by Theorem 2.

8. Approximations. Here we give simple upper bounds for $M(N,k)$ and $T(N,k)$. Since all quantities depend on $E_r(N,K)$ for various r, the natural approximation is derived by writing

$$E_r(N,k) \approx 1 + \binom{r+1}{r}\frac{k-1}{N} + \binom{r+2}{r}\left(\frac{k-1}{N}\right)^2 + \cdots = \left(\frac{1}{1-\rho}\right)^{r+1}.$$

Here $\rho = (k-1)/N$ is the approximate density of the items in the table. We then have the approximation

$$\begin{aligned} M(N,k) &\approx \frac{1}{2}\left(1 + \frac{1}{(1-\rho)^2}\right) \\ T(N,k) &\approx \frac{1}{2}\left(1 + \frac{1}{1-\rho}\right) \\ \sigma^2(N,k) &\approx \frac{3}{4}\frac{1}{(1-\rho)^4} - \frac{2}{3}\frac{1}{(1-\rho)^3} - \frac{1}{12} \end{aligned}$$

The estimates for $M(N,k)$ and $T(N,k)$ are upper bounds; we see that the estimates are dependent upon N and k only with respect to their ratio. The estimates are best for small values of ρ. As ρ approaches 1, these estimates become less and less accurate. The worst case is when $k = N$, i.e. $\rho = 1 - 1/N$. Then the approximations give

$$(1+N^2)/2, \qquad (9N^4 - 3N^3 - 1)/12$$

respectively for $M(N,N)$, $\sigma^2(N,N)$; the true values are

$$(1+N)/2, \qquad (N^2-1)/12$$

respectively. As N becomes large, however, the approximation is good for fixed ρ; in fact, it is not hard to show that

$$E_r(N,\rho N) \uparrow \frac{1}{(1-\rho)^{r+1}} \quad \text{as } N \to \infty.$$

9. Relations to the paper of Schay and Spruth. In the article by Schay and Spruth [2], a method is described which the authors claim is superior to open addressing: the elements are shifted as they are stored, so that the elements whose starting address is x_j are stored before (in the cyclic sense) those whose starting address is x_{j+1}. However, the method described there, although more complicated than the open addressing method, actually **has exactly the same performance characteristics**. For it was shown by Peterson [1] that $T(N, k)$ is independent of the order in which the elements enter the table; and there is an order for which the table takes the same form as stipulated by Schay and Spruth.

Since the more complex method of [2] is therefore exactly the same as the one we have described, with respect to its performance, it is interesting to compare the timing estimates derived in that paper with the ones derived here. The formula

$$T(N, k) \approx 1 + \frac{\rho}{2(1-\rho)}$$

was derived in [2], and this coincides with

$$\frac{1}{2}\left(1 + \frac{1}{1-\rho}\right),$$

the approximation we derived above, except the value $\rho = k/N$, not $(k-1)/N$, was used in [2]. In that paper, the approximation occurred when the binomial probability was replaced by a Poisson distribution. The derivation is completely different from the one described here, and the authors were not concerned with the quantities $M(N, k)$ or $P(N, k, m)$.

10. Selected values of $M(N, k)$, $\sigma(N, k)$, $T(N, k)$. Precise calculations of $E_r(N, k)$ may be rapidly performed by computer, and hence the exact values of the mean, standard deviation, and average search time. The tables on the following page, for example, were computed by the B–5000 in approximately five seconds, which is the time needed to print that amount of information. The table indicates to what extent the quantities depend on the ratio k/N; all values are exact, except of course for rounding.

References

1. W. W. Peterson, *Addressing for Random-Access Storage*, IBM J. **1** (1957), 130–146.
2. G. Schay, Jr., and W. G. Spruth, *Analysis of a File Addressing Method*, Comm. ACM **5** (1962), 459–461.
3. K. E. Iverson, *A Programming Language*, Wiley, 1962, pp. 153–154.

Table 1. $M(N,k)$

N	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
10	1.000	1.100	1.230	1.4?2	1.634	1.954	2.404	3.055	4.022	5.500
100	1.102	1.256	1.476	1.8?5	2.327	3.218	4.899	8.532	18.029	50.500
1000	1.116	1.279	1.516	1.8?0	2.480	3.576	5.897	12.213	40.357	500.50
10000	1.117	1.281	1.520	1.8?8	2.498	3.620	6.039	12.914	49.116	5000.5
100000	1.117	1.281	1.520	1.8?9	2.500	3.624	6.054	12.991	50.356	5000.1
1000000	1.117	1.281	1.520	1.8?9	2.500	3.625	6.055	13.000	50.486	500001

Table 2. $\sigma(N,k)$

N	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
10	0.000	0.300	0.487	0.6?0	0.930	1.220	1.573	1.994	2.463	2.873
100	0.344	0.610	0.945	1.420	2.146	3.338	5.444	9.455	17.458	28.866
1000	0.377	0.661	1.036	1.595	2.514	4.200	7.787	17.390	55.931	288.68
10000	0.381	0.667	1.046	1.615	2.561	4.321	8.187	19.369	78.421	2886.8
100000	0.381	0.667	1.047	1.618	2.565	4.334	8.230	19.603	82.209	28868.
1000000	0.381	0.668	1.047	1.618	2.566	4.335	8.235	19.627	82.618	288657

Table 3. $T(N,k)$

N	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
10	1.000	1.050	1.110	1.1?3	1.273	1.387	1.532	1.722	1.978	2.330
100	1.049	1.116	1.200	1.312	1.463	1.682	2.020	2.598	3.747	6.605
1000	1.055	1.124	1.213	1.331	1.496	1.742	2.149	2.941	5.101	20.151
10000	1.055	1.125	1.214	1.333	1.500	1.749	2.165	2.994	5.451	63.000
100000	1.056	1.125	1.214	1.333	1.500	1.750	2.166	2.999	5.495	198.50
1000000	1.056	1.125	1.214	1.333	1.500	1.750	2.167	3.000	5.500	626.99

Observe that the entries in the lower row of each table agree well with the approximations we have obtained. Observe also that the standard deviation tends to be large compared with the mean, especially for $k \geq N/2$. Such instability is familiar to anyone who has watched the open addressing method in operation, as the table begins to get full.

This table gives strong reason to believe that $T(N,N) = \sqrt{\frac{\pi N}{8}} + O(1)$, but I have not been able to prove this. Proved 5/20/65