On the Analysis of Linear Probing Hashing

Philippe Flajolet
INRIA, France

January 15, 1998

[summary by Hosam M. Mahmoud]

For uniform data, hashing is known to provide fast access schemes [12]. The idea in hashing is to maintain a table, of size m say, and map n keys to the locations of the table. In the absence of complications, later on we can retrieve the key by looking up its hash position in the table. A key x is associated with a hash address $h(x) \in \{1, \ldots, m\}$. In practice, data may collide: the chosen hash function may map two keys to the same location in the table. In this case we must resolve collisions. Standard mechanisms for collision resolution are chaining and linear probing, among other. For hashing a set of n keys to be successful, m must be at least as large as n. The ratio $\alpha = n/m \leq 1$ is called the load factor and plays an important role in the analysis. The special case $\alpha = 1$ corresponds to eventually filling the table at the end of hashing the entire data set; this case will be dubbed the title *full table*. A sparse table (small α) may be viewed as a collection of smaller full tables separated by empty locations. These smaller full tables are also figuratively called islands.

The situation is paralleled to balls-in-urns arguments. In this analogy, the n keys are emulated by n balls, and the m hash locations are emulated by m urns. The random allocation of balls unto urns is the parallel of a uniform hash function. Several results are mentioned to indicate some facts about random allocation of balls in urns and are related to classical theory:

- Collisions occur early (the Birthday Paradox);
- The probability of no collision in a full table is rather small (exponentially so);
- Empty cells disappear late (Coupon Collector's Problem);
- In a sparse table, the maximal share of a bucket is moderately high. For instance if the average share of an urn is $\alpha = n/m = 1/2$, still one of the shares grows on average as fast as $\log n / \log\log n$.

A proof is sketched to argue that when both $m, n \to \infty$, in such a way that $n/m \to \alpha$, the number of urns that receive exactly k (fixed) balls follows a Poisson law with parameter α:

$$P\{\text{an urn receives } k \text{ balls}\} = \frac{\alpha^k}{k!} e^{-\alpha}.$$

Noticeably, even when the number of balls and urns are the same ($\alpha = 1$) the proportion of empty urns ($k = 0$) approaches $e^{-\alpha} \approx 36\%$.

In passing, the analysis of Separate Chaining (when all keys hashed to the same location are linked in a linear chain) is mentioned. The main thrust of the talk, however, focused on Linear Probing Hashing. In this latter collision resolution method, when a key is hashed to an already occupied location, the resolution algorithm looks for the nearest unoccupied position above the hash position (wrapping around to the beginning of the table, if necessary). The distance a key travels till collision is resolved, the *displacement*, is a measure of efficiency for data insertion and retrieval.

Stochastically, the displacement increases as more keys are placed in the table. For example, stochastically the last key has the highest displacement. This last displacement is intuitively small for small α. If α is close to 1, "clotting" occurs and the average displacement is asymptotic to $m/2$.

The problem was first proposed by Knuth in 1962. Over the course of time connections to Abel identities and Ramanujan's function were discovered. "Generating functionology" is a key element in the analysis. A broad array of analytic constructions (a dictionary of formal operators so to speak) together with singularity analysis play a central rôle in the analysis. A starting point is the decomposition of an almost full table ($n = m - 1$) into two full tables:

$$\langle \text{full} \rangle \equiv \langle \text{full} \rangle \star \langle \text{full} \rangle,$$

with the $\star$ indicating an empty slot at any position. In the language of the enumeration generating function $F(z)$, this decomposition corresponds to an integral operator in the dictionary, giving

$$F = \int (zF)' F.$$

The substitution $T = zF$ gives an ordinary differential equation from which it is then demonstrated that $T(z)$ is the *tree function* that solves the equation

$$T = ze^T.$$

By Lagrange's inversion and methods of Eisenstein and Cayley, an explicit formal series is obtained:

$$T(z) = \sum_{n=0}^{\infty} n^{n-1} \frac{z^n}{n!}.$$

Trees also have a decomposition, discussed by Knuth as early as 1963. The number of almost-full tables for n keys is

$$F_n = (n+1)^{n-1}.$$

The talk then shifts focus from counting (the totality of the sample space) to distributional analysis of almost full tables. Conditioned on where the empty slot falls in an almost-full table, one obtains a convolution formula for the probability generating function of full tables:

$$F_n(q) = \sum_{k=0}^{n-1} \binom{n-1}{k} (1 + q + \cdots + q^k) F_k(q) F_{n-1-k}(q).$$

Let $F(z,q)$ be the bivariate generating function of the sequence $F_n(q)$. Via a number of differential operators on $F(z,q)$, moment generating functions are expressed by differential equations involving rational polynomials of the tree function. Let $d_{n,m}$ be the total displacement to place n uniform keys into a hash table of size m. Extraction of coefficients then yields the following result [7, 2].

Theorem 1.

$$E[d_{n,n}] = \frac{n}{2}\big(Q(n) - 1\big), \qquad E[d_{n,n}^2] = \frac{n}{12}\big(5n^2 + 4n - 1 - 8Q(n)\big),$$

where $Q(n)$ is a Ramanujan function.

Asymptotic analysis of the mean and variance gives a series expansion.

Theorem 2.

$$E[d_{n,n}] = \frac{\sqrt{2\pi}}{4} n^{3/2} - \frac{2}{3} n + \frac{\sqrt{2\pi}}{48} n^{1/2} - \frac{2}{135} + O\big(n^{-1}\big),$$

$$\mathrm{Var}[d_{n,n}] = \frac{10 - 3\pi}{24} n^3 + \frac{16 - 3\pi}{144} n^2 + \frac{\sqrt{2\pi}}{135} n^{3/2} - \cdots.$$

Higher moments are "pumped" from the functional equation on $F(z,q)$. Through the rth derivative, one gets a functional equation for the rth moment. The latter functional is solved either exactly or asymptotically. The method has been used before in various combinatorial analyses, such as Quicksort [4], path length in trees [17], Brownian excursions [11], in-situ permutations [9, 6].

The Airy distribution is introduced next. Its distribution function solves the differential equation

$$Y'' - zY = 0,$$

and is known to have the integral representation

$$Ai(z) = \frac{1}{\pi}\int_0^\infty \cos\Big(\frac{1}{3}t^3 + zt\Big)\,dt.$$

The Airy distribution is uniquely characterized by its moments, as its exponential moment generating function converges in a neighborhood of 0. By showing that all moments of the random total displacement converge to the moments of the Airy distribution, one main result of the investigation is obtained.

Theorem 3. *In an almost full table the random total displacement $d_{n,n-1}$ converges in distribution to A, a random variable having the Airy distribution, in the usual sense of convergence of distribution functions: for every real x,*

$$P\Big\{\frac{d_{n,n-1}}{(n/2)^{3/2}} \le x\Big\} \to P\{A \le x\}, \qquad as \quad n \to \infty.$$

Full tables are building blocks of general hash tables. Generally, a table can be decomposed as

$$\langle \text{full} \rangle \equiv \langle \text{full} \rangle \star \cdots \star \langle \text{full} \rangle.$$

Given that there are k islands, the bivariate generating function becomes the convolution $F^k(z,q)$.

The whole analysis package outlined above can then be "recycled" to derive the result for a general load factor, as in [2, 8].

Theorem 4.

$$\begin{aligned} E[d_{m,n}] &= \frac{n}{2}\big(Q_0(m,n-1)-1\big),\\ E[d_{m,n}^2] &= \frac{n}{12}\big[(m-n)^3 + (n+3)(m-n)^2 + (8n+1)(m-n) + 5n^2 + 4n - 1\\ &\qquad - \big((m-n)^3 + 4(m-n)^2 + (6n+3)(m-n) + 8n\big)Q_0(m,n-1)\big]. \end{aligned}$$

where $Q_0(m,n)$ is a Ramanujan function. Asymptotically, these expressions simplify to

$$\begin{aligned} E[d_{m,n}] &= \frac{\alpha}{2(1-\alpha)}n - \frac{\alpha}{(1-\alpha)^3} + O\big(n^{-1}\big),\\ \mathrm{Var}[d_{m,n}] &= \frac{6\alpha - 6\alpha^2 + 4\alpha^3 - \alpha^4}{12(1-\alpha)^4}n - \cdots. \end{aligned}$$

The convolution form of the generating function admits a Gaussian law.

Theorem 5. *The random total displacement is asymptotically normally distributed.*

This result is obtained by a delicate saddle point analysis on the integral

$$\frac{1}{2\pi i}\oint \frac{F^{m-n}(z,q)}{z^{n+1}}\,dz,$$

an expression for the coefficient (a probability generating function) of z^n in $F^{m-n}(z,q)$, the bivariate function for a table of m locations receiving n keys.

This general law for coefficients of functions that are large powers of generating functions has wide applicability and appears later in other contexts, like for example the analysis of Distributive Sort (a flavor of Bucket Sort) with a large number of buckets [13].

By no means the Airy distribution appears in hash tables as an isolated phenomenon. It seems to be a ubiquitous law in combinatorial analysis. We now know that it appears in full hash tables for Linear Probing with Hashing; inversions in trees; random walks; path length and Dyck or Catalan walks in random trees. These connections to the Airy distribution may be found in [10, 3, 8, 5, 16, 11, 17, 18, 14, 1, 15]. These works connect various areas of combinatorial analysis to each other and eventually to the Airy distribution. Although closed forms for bivariate functions for random variables with the Airy distribution are known, their moments are still hard to find.

Bibliography

[1] Flajolet (Philippe) and Noy (Marc). – *Analytic Combinatorics of Non-crossing Configurations.* – Research Report n° 3196, Institut National de Recherche en Informatique et en Automatique, June 1997.

[2] Flajolet (Philippe), Poblete (Patricio), and Viola (Alfredo). – *On the Analysis of Linear Probing Hashing.* – Research Report n° 3265, Institut National de Recherche en Informatique et en Automatique, September 1997. 22 pages. To appear in *Algorithmica*.

[3] Gessel (Ira) and Wang (Da Lun). – Depth-first search as a combinatorial correspondence. *Journal of Combinatorial Theory. Series A*, vol. 26, n° 3, 1979, pp. 308–313.

[4] Hennequin (Pascal). – Combinatorial analysis of quicksort algorithm. *RAIRO Theoretical Informatics and Applications*, vol. 23, n° 3, 1989, pp. 317–333.

[5] Janson (Svante), Knuth (Donald E.), Łuczak (Tomasz), and Pittel (Boris). – The birth of the giant component. *Random Structures & Algorithms*, vol. 4, n° 3, 1993, pp. 231–358.

[6] Kirschenhofer (P.), Prodinger (H.), and Tichy (R. F.). – A contribution to the analysis of in situ permutation. *Drušstvo Matematičara i Fizičara S. R. Hrvatske. Glasnik Matematički. Serija III*, vol. 22 (42), n° 2, 1987, pp. 269–278.

[7] Knuth (D. E.). – Notes on "open" addressing. – Unpublished memorandum, 1963. Memo dated July 22, 1963. With annotation *"My first analysis of an algorithm, originally done during Summer 1962 in Madison"*.

[8] Knuth (D. E.). – Linear probing and graphs. – Preprint, 1997.

[9] Knuth (Donald E.). – Mathematical analysis of algorithms. In *Information processing 71*. pp. 19–27. – Amsterdam, 1972. Proceedings IFIP, Ljubljana, 1971, Vol. 1: Foundations and systems.

[10] Kreweras (G.). – Une famille de polynômes ayant plusieurs propriétés énumeratives. *Periodica Mathematica Hungarica. Journal of the János Bolyai Mathematical Society*, vol. 11, n° 4, 1980, pp. 309–320.

[11] Louchard (G.). – The Brownian excursion area: a numerical analysis. *Computers & Mathematics with Applications*, vol. 10, n° 6, 1984, pp. 413–417.

[12] Lum (V. Y.), Yuen (P. S. T.), and Dodd (M.). – Key-to-address transform techniques: A fundamental performance study on large existing formatted files. *Communications of the ACM*, vol. 14, n° 4, April 1971, pp. 228–239.

[13] Mahmoud (Hosam), Flajolet (Philippe), Jacquet (Philippe), and Régnier (Mireille). – *Analytic Variations on Bucket Selection and Sorting.* – Research Report n° 3399, Institut National de Recherche en Informatique et en Automatique, 1998. 22 pages, submitted to *Acta Informatica*.

[14] Mallows (C. L.) and Riordan (John). – The inversion enumerator for labeled trees. *Bulletin of the American Mathematical Society*, vol. 74, 1968, pp. 92–94.

[15] Prellberg (T.). – Uniform q-series asymptotics for staircase polygons. *Journal of Physics. A. Mathematical and General*, vol. 28, n° 5, 1995, pp. 1289–1304.

[16] Spencer (Joel). – Enumerating graphs and Brownian motion. *Communications on Pure and Applied Mathematics*, vol. 50, n° 3, 1997, pp. 291–294.

[17] Takács (Lajos). – On the distribution of the number of vertices in layers of random trees. *Journal of Applied Mathematics and Stochastic Analysis*, vol. 4, n° 3, 1991, pp. 175–186.

[18] Wright (E. M.). – The number of connected sparsely edged graphs. *Journal of Graph Theory*, vol. 1, n° 4, 1977, pp. 317–330.