Skip Lists - Some Results on a Recent Data Structure*

Peter Kirschenhofer
Department of Algebra and Discrete Mathematics
Technical University of Vienna †

Abstract

The skip list is a list-based data structure introduced some years ago by Pugh [12]. In [6] an optimized version of the skip list search algorithm has been investigated and an asymptotic result on the variance of the total unsuccessful search cost has been derived. Here we give the precise asymptotics for the difference of the variance of this parameter and the variance of the total successful search cost.

1 Introduction

The *skip list* is a list-based type of data structure that was introduced by Pugh ([12]). In the following we only give a very short description of this structure. We refer to [11], [8], [9], [1], [7] and [6] for detailed descriptions and various analyses on this subject.
A set of n elements is stored in a collection of sorted linear linked lists in the following manner: all elements are stored in increasing order in a linked list called *level* 1 and, recursively, each element which is included in the linked list *level* i is included with independent probability q $(0 < q < 1)$ in the linked list *level* $i+1$. The number of linked lists an element x belongs to is called the *level* of the element. For each element in the skip list, we need a node to store its key and as many pointers as its level indicates. The successor of x at the list level i is given by the i-th pointer of x, also called i-th *forward* pointer of x. A *header* refers to the first element in each of the linked lists and it also holds the *height* of the skip list, which is the maximum level of all elements.
In [7] the asymptotic analysis of the *total (successful) search cost* or *path length*, i.e. the sum of the successful search costs to find all the elements in

*This research was supported by the program Acciones Integradas Hispano-Austríacas (Austrian-Spanish Scientific Exchange Program).
†Wiedner Hauptstr. 8–10, A-1040 Vienna, Austria. e-mail: kirsch@rsmb.tuwien.ac.at

```
x:= header(S); l:= height(S);
alreadyChecked:= NIL;
while l > 0 do
  while (x^.forward [l] <> alreadyChecked) do
    while(x^.forward[l]^.key < search_key) do
            x:= x^.forward[l]
    endwhile
  endwhile;
  alreadyChecked:= x^.forward[l];
  l:= l - 1
endwhile
```

Figure 1: Optimized skip list search algorithm (see [11])

the data structure, was performed for the instance of the simplest form of the search algorithm.

The paper [6] is devoted to the analysis of the total search cost of an *optimized version* of the search algorithm proposed in [11], that reduces the number of expensive key comparisons by guaranteeing that the search key will be compared at most once with the key of any element in the skip list. To this aim, the variable `alreadyChecked` is introduced. At the beginning this variable is set to "NIL". In the following loop we follow forward pointers as long as the elements pointed to are different from `alreadyChecked` and the keys of those elements are smaller than the search key. As soon as this horizontal traversal ends, `alreadyChecked` is set to the element pointed to at this moment and the search continues one level below (see Figure 1).

In [6] an explicit example for a search path in a skip list is depicted together with a comparison of the number of key-to-key comparisons and pointer inspections in both the original and the optimized instance of the search algorithm.

It should be noted that we may also describe a skip list of size n as an n-tuple $(a_1, \ldots, a_n)$, where a_i denotes the level of the i-th element. We adopt the probabilistic model for *random skip lists* of being the outcome of n independent identically distributed random variables. In particular, each $a_i \in \mathbb{N}$ is the outcome of a *geometric* random variable X_i of parameter p, i.e. $\mathrm{Prob}\{X_i = k\} = pq^{k-1}$, where $q = 1 - p$. (Note that in some earlier papers the roles of p and q are interchanged.)

Interestingly enough the search cost parameter, i.e. the number of key-to-key comparisons, can be expressed in terms of *order statistics*: The number $C_{n,i}$ of key comparisons when searching for element i equals the number $l_{n,i}$ of *strict left-to-right maxima* of the sequence $(a_i, \ldots, a_n)$ plus the number r_i of *weak right-to-left maxima* of $(a_1, \ldots, a_{i-1})$.

Observe that for a fixed element i the two parameters are independent random variables, but that this is no longer true for the *total successful search cost* $\widetilde{C}_n$, which is given by the sum $\sum_{i=1}^n C_{n,i}$. This dependency is the reason that the variance of $\widetilde{C}_n$ cannot be gained by simply adding the variances of the two parameters $\sum_i l_{n,i}$ and $\sum_i r_i$, which where already computed in [7].
The number r_{n+1} of weak right-to-left maxima of $(a_1, \ldots, a_n)$ is never counted in $\widetilde{C}_n = \sum_{i=1}^n C_{n,i}$, which causes some kind of asymmetry of this parameter. Therefore in [6] instead of $\widetilde{C}_n$ the *total unsuccessful search cost* C_n was studied:

$$C_n = \sum_{1 \le i \le n+1} C_{n,i}. \tag{1}$$

This parameter fulfills nice recurrence relations, and its expectation and variance were analyzed asymptotically in the cited paper. The variance of C_n is of order n^2, and it was shown that the difference of the variances fulfills

$$\mathrm{Var}(C_n) - \mathrm{Var}(\widetilde{C}_n) = O(n^{17/9+\varepsilon}), \quad \text{any} \quad \varepsilon > 0, \tag{2}$$

so that $\mathrm{Var}(\widetilde{C}_n)$ is of order n^2, too. It was conjectured, that the difference of the variances is of order $n \log^2 n$. In this paper we will prove that in fact the order of the difference is $n \log n$. We will give a precise asymptotic result, which also allows to describe the asymptotics of the *covariance* of the two random variables C_n and r_{n+1}.
From the technical point of view we will make use of probability generating functions, use a substitution that allows to express the desired coefficients explicitly in terms of alternating sums and apply a technique due to S.O.Rice to get the final asymptotics (compare the very recent article [2] on the latter subject.)

2 Generating Functions and First Order Moments

We start with the following combinatorial decomposition of a skip list of height m (compare [7, 6]): We split up the whole skip list according to the first appearance of an element of height m into a skip list σ of height $< m$, the partitioning element of height m and the remaining skip list τ of height $\le m$. Observing that each successful or unsuccessful search starts with a comparison of the search key with the key of the partitioning element we easily find

$$\begin{aligned} C(\sigma m \tau) &= C(\sigma) + C(\tau) + |S| + 1, \\ \widetilde{C}(\sigma m \tau) &= C(\sigma) + \widetilde{C}(\tau) + |S|. \end{aligned} \tag{3}$$

We denote by $P^*(z,y)$ resp $\widetilde{P}^*(z,y)$ the bivariate generating functions where the coefficient of $z^n y^k$ is the probability that a random skip list of size n has height fulfilling condition $*$ and the total unsuccessful resp. successful search cost is equal to k. Then Eqs. (3) translate to the system of functional equations

$$\begin{aligned} P^{=m}(z,y) &= pq^{m-1}zy^2P^{<m}(zy,y)P^{\le m}(zy,y), \\ \widetilde{P}^{=m}(z,y) &= pq^{m-1}zyP^{<m}(zy,y)\widetilde{P}^{\le m}(zy,y), \quad m\ge 1, \\ P^{=0}(z,y) &= \widetilde{P}^{=0}(z,y) = 1. \end{aligned} \tag{4}$$

It is convenient to use the bgf's $R^*(z,y) := zP^*(z,y), \widetilde{R}^*(z,y) := z\widetilde{P}^*(z,y)$ instead. In order to get the expectations we have to derive w.r.t. y and set $y=1$. We introduce the notations $S^*(z) = R^*_y(z,1)$ resp. $\widetilde{S}^*(z) = \widetilde{R}^*_y(z,1)$ for the generating functions of these partial derivatives as well as the abbreviations

$$Q := q^{-1},\ L := \log Q,\ [\![m]\!] := 1 - z(1-q^m)$$

and get

$$\begin{aligned} S^{=m}(z) &= pq^{m-1}\left[\left(\frac{z}{[\![m-1]\!]^2} + S^{<m}(z)\right)\frac{z}{[\![m]\!]} + \left(\frac{z}{[\![m]\!]^2} + S^{\le m}(z)\right)\frac{z}{[\![m-1]\!]}\right], \\ \widetilde{S}^{=m}(z) &= pq^{m-1}\left[\left(\frac{z}{[\![m-1]\!]^2} + S^{<m}(z)\right)\frac{z}{[\![m]\!]}\right. \\ &\quad \left. + \left(\frac{z}{[\![m]\!]^2} - \frac{z}{[\![m]\!]} + \widetilde{S}^{\le m}(z)\right)\frac{z}{[\![m-1]\!]}\right] \end{aligned} \tag{5}$$

Since we know $S^{\le m}(z)$ from [6] we focus our attention on the difference

$$\Delta^*(z) = S^*(z) - \widetilde{S}^*(z). \tag{6}$$

Observing $\Delta^{=m}(z) = \Delta^{\le m}(z) - \Delta^{<m}(z)$, we get from (5) a first order linear recurrence relation for the sequence $(\Delta^{\le m}(z))_{m\ge 0}$, which has the solution

$$\Delta^{\le m}(z) = \frac{p}{q}\frac{z^2}{[\![m]\!]}\sum_{i=1}^{m}\frac{q^i}{[\![i]\!]} \tag{7}$$

For $m\to\infty$, this yields

$$\Delta(z) = \frac{p}{q}\frac{z^2}{1-z}\sum_{i\ge 1}\frac{q^i}{[\![i]\!]}. \tag{8}$$

Although we know already the asymptotics of the coefficients of $\Delta(z)$ from [10], where, amongst other results, the expectation and the variance of the random variable $r_{n+1} = C_n - \widetilde{C}_n$ were derived, we want to sketch here again the method used in [7, 6], since it may be well used for the asymptotic evaluation of the much more complicated expressions for the differences of the second moments, too. We start from the substitution

$$z = \frac{w}{w-1},$$

which in terms of formal residue composition (compare e.g. [3]) reads

$$[z^n]f(z) = (-1)^n [w^n](1-w)^{n-1} f(w/(w-1)).$$

In our instances the above substitution leads to *harmonic sums*

$$G(w) = \sum_i a_i g(b_i w),$$

so that

$$[w^n]G(w) = \sum_i a_i b_i^n \cdot [w^n]g(w),$$

which can be computed explicitly in all our cases. Proceeding in this manner we find

$$\begin{aligned}
\frac{q}{p}[z^n]\Delta(z) &= [z^n]\frac{z^2}{(1-z)}\sum_{i\geq 1}\frac{q^i}{[\![i]\!]} = (-1)^n[w^n]w(1-w)^{n-1}\sum_{i\geq 1}\frac{wq^i}{1-wq^i} \\
&= (-1)^n \sum_{k=1}^{n-1}\binom{n-1}{k}(-1)^{n-1-k}[w^{k-1}]\sum_{i\geq 1}\frac{wq^i}{1-wq^i} \\
&= -\sum_{k=1}^{n-1}\binom{n-1}{k}(-1)^k \sum_{i\geq 1} q^{(k-1)i}[w^{k-1}]\frac{w}{1-w} \\
&= -\sum_{k=1}^{n-1}\binom{n-1}{k}(-1)^k \frac{1}{Q^k-1}. \qquad (9)
\end{aligned}$$

There are several techniques to evaluate the alternating sum. One of them starts from the observation that according to Mittag Leffler's Theorem the meromorphic function $\frac{1}{Q^z-1}$ has a partial fraction decomposition in the complex plane. This decomposition is surprisingly simple and follows from the well-known result (compare e.g. [5] eqn.(7.10)-10)

$$\frac{2\pi}{\mathrm{e}^{2\pi z}-1} = -\pi + \frac{1}{z} + \sum_{n=1}^{\infty}\left(\frac{1}{z-in} + \frac{1}{z+in}\right). \qquad (10)$$

Inserting the last result in eqn.(10) we arrive at alternating sums that may be computed explicitly using the formula (compare e.g. [4] eqn.(5.41))

$$\sum_{k=0}^{n}\binom{n}{k}(-1)^k\frac{1}{x+k} = \frac{n!}{x(x+1)\dots(x+n)}. \qquad (11)$$

In this manner we find with the abbreviation $\chi_j = 2j\pi i/L$

$$\mathrm{E}(C_n) - \mathrm{E}(\widetilde{C}_n) = [z^{n+1}]\Delta(z) = (Q-1)\left(\frac{H_n}{L} - \frac{1}{2} - \frac{1}{L}\sum_{j\neq 0}\frac{\Gamma(-\chi_j)\Gamma(n+1)}{\Gamma(n+1-\chi_j)}\right). \qquad (12)$$

The third term can be rewritten using Euler's classical formula

$$\frac{1}{\Gamma(z)} = z\mathrm{e}^{\gamma z}\prod_{k\geq 1}\left[(1-\frac{z}{k})\mathrm{e}^{z/k}\right] \tag{13}$$

in the form

$$\frac{1}{L}\sum_{j\neq 0}\frac{\Gamma(-\chi_j)\Gamma(n+1)}{\Gamma(n+1-\chi_j)} = \frac{1}{L}\sum_{j\neq 0}\Gamma(-\chi_j)\mathrm{e}^{\frac{2j\pi i(H_n-\gamma)}{L}}\prod_{k\geq n+1}\left[(1-\frac{2j\pi i}{Lk})\mathrm{e}^{\frac{2j\pi i}{Lk}}\right]. \tag{14}$$

From this last expression we easily see that for $n\to\infty$ the third term converges towards a periodic function in $\log_Q n$, so that we have the asymptotic result (compare [10])

$$\mathrm{E}(C_n) - \mathrm{E}(\widetilde{C}_n) = \mathrm{E}(r_{n+1}) = (Q-1)\left(\log_Q n + \frac{\gamma}{L} - \frac{1}{2} + \delta(\log_Q n)\right), \tag{15}$$

where $\delta(x) = -\frac{1}{L}\sum_{j\neq 0}\Gamma(-\chi_j)\mathrm{e}^{2j\pi i x}$ is a continuous periodic function of period 1 and mean 0 with very small amplitude for reasonable values of Q.
If the alternating sums in consideration involve more complicated terms and we are mainly interested in asymptotic results it is convenient to follow an approach attributed to S.O.Rice. The excellent survey [2] explains this methodology in detail. The basic idea is to express the alternating sum as a complex contour integral:

$$\sum_{k=a}^{n}\binom{n}{k}(-1)^k f(k) = -\frac{1}{2\pi i}\int_{\mathcal{C}} B(n+1,-z)f(z)\,dz, \tag{16}$$

where $B(x,y) = \dfrac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ is the Beta function, $\mathcal{C}$ is a positively oriented closed curve surrounding the points $a, a+1, \ldots, n$ and $f(z)$ is an analytic continuation of the discrete sequence $f(k)$ to the complex plane, with no poles within the region surrounded by $\mathcal{C}$.
If $f(z)$ decreases sufficiently fast towards $\pm i\infty$, the asymptotic evaluation of this expression can be achieved by extending the contour of the integral to the left and collecting the residues at the newly encountered poles. The residue computations can often be performed automatically using a Computer algebra system like MAPLE.

3 Second Order Moments

In this section, we will establish the asymptotic behavior of the difference of the second factorial moments of C_n resp. $\widetilde{C}_n$, thereby getting the desired asymptotics for the difference of the variances resp. the covariance of C_n and r_{n+1}.

The generating functions for the second factorial moments are obtained by deriving $P^*(z,y)$ resp. $\widetilde{P}^*(z,y)$ twice w.r.t. y and setting $y=1$. Again we use R resp. $\widetilde{R}$ instead of P resp. $\widetilde{P}$. Denoting the generating functions of the second partial derivatives by $T^*(z) := R^*_{yy}(z,1)$ resp. $\widetilde{T}^*(z) := \widetilde{R}^*_{yy}(z,1)$ we find, starting from eqs.(4), for the difference $\Phi^*(z) = T^*(z) - \widetilde{T}^*(z)$ the relation

$$\begin{aligned} \Phi^{=m}(z) &= \frac{pq^{m-1}z}{[\![m-1]\!]}\left(\Phi^{\le m}(z) + 2z\Delta_z^{\le m}(z) + 2\widetilde{S}^{\le m}(z) + \frac{2z}{[\![m]\!]^2} - \frac{2z}{[\![m]\!]}\right) \\ &+ 2pq^{m-1}\left(\frac{z}{[\![m-1]\!]^2} + S^{\le m-1}\right)\left(\frac{z}{[\![m]\!]} + \Delta^{\le m}(z)\right). \end{aligned} \tag{17}$$

Inserting the functions $\Delta, S, \widetilde{S}$ explicitly, solving the first order linear recurrence for $\Delta^{\le m}(z)$ and performing the limit $m \to \infty$ we get the solution

$$\begin{aligned} &\Phi(z) = \frac{2z^2}{1-z}\frac{p}{q}\{-S_1 + (1+pz)(S_2+S_3)\} \\ &+ \frac{2z^3}{1-z}\frac{p^2}{q^2}\{-qS_4 - S_5 + (1+pz)S_6 - S_7 + (2+q)(S_8+S_{10}) + S_9 + pzS_{10}\} \\ &+ \frac{2z^4}{1-z}\frac{p^3}{q^3}\{-S_{11} + (1+q)S_{12}\} \end{aligned} \tag{18}$$

with the sums

$$\begin{aligned} &\mathcal{S}_1(z) = \sum_{i\ge1}\frac{q^i}{[\![i]\!]},\ \mathcal{S}_2(z) = \sum_{i\ge1}\frac{q^i}{[\![i]\!]^2},\ \mathcal{S}_3(z) = \sum_{i\ge1}\frac{q^i}{[\![i]\!][\![i-1]\!]},\ \mathcal{S}_4(z) = \sum_{i\ge1}\frac{q^{2i}}{[\![i]\!]^3}, \\ &\mathcal{S}_5(z) = \sum_{i\ge1}\frac{q^{2i}}{[\![i]\!][\![i-1]\!]^2},\ \mathcal{S}_6(z) = \sum_{i\ge1}\frac{q^{2i}}{[\![i]\!]^2[\![i-1]\!]},\ \mathcal{S}_7(z) = \sum_{1\le j\le i}\frac{q^{i+j}}{[\![i]\!][\![j]\!]}, \\ &\mathcal{S}_8(z) = \sum_{1\le j\le i}\frac{q^{i+j}}{[\![i]\!]^2[\![j]\!]},\ \mathcal{S}_9(z) = \sum_{1\le j\le i}\frac{q^{i+j}}{[\![i]\!][\![j]\!]^2},\ \mathcal{S}_{10}(z) = \sum_{1\le j<i}\frac{q^{i+j}}{[\![i]\!][\![i-1]\!][\![j]\!]}, \\ &\mathcal{S}_{11}(z) = \sum_{1\le j\le i}\frac{q^{2i+j}}{[\![i]\!][\![i-1]\!]^2[\![j]\!]},\ \mathcal{S}_{12}(z) = \sum_{\substack{1\le j\le i\\ 1\le h<i}}\frac{q^{i+j+h}}{[\![i]\!][\![i-1]\!][\![j]\!][\![h]\!]}. \end{aligned} \tag{19}$$

Proceeding as in Section 2 we gain the following explicit expression for the coefficients of $\Phi(z)$ in terms of alternating sums:

$$\begin{aligned} &[z^n]\Phi(z) = -2(Q-1)^2\sum_{k=2}^{n}\binom{n}{k}\frac{(-1)^k}{Q^{k-1}-1}\Bigg\{-(1+q) + (k-1)(1-\frac{q}{Q-1}) \\ &+2\binom{k-1}{2} + q\sum_{m=1}^{k-2}\frac{m}{Q^m-1} + 2(1+q)\sum_{m=1}^{k-2}\frac{1}{Q^m-1} + (k-1)\sum_{m=1}^{k-2}\frac{1}{Q^m-1}\Bigg\} \\ &+ 2q(Q-1)^2\sum_{k=2}^{n-1}\binom{n-1}{k}\frac{(-1)^k k}{Q^{k-1}-1} + 2(n-1) + \end{aligned}$$

$$- 2(Q-1)^2 \sum_{k=1}^{n-1} \binom{n-1}{k} \frac{(-1)^k}{Q^k-1} \left\{ -\frac{1}{Q-1} + k - 1 + \sum_{m=1}^{k-1} \frac{1}{Q^m-1} \right\}. \qquad (20)$$

Applying Rice's method described in Section 2 allows to obtain the asymptotics of the coefficients now as

$$\begin{aligned} [z^n]\Phi(z) &= 2(Q-1)(Q-q)n\log_Q^2 n \\ &+2(Q-1)^2 n \log_Q n \left(-1 + \frac{2(\gamma-1)(1+q)}{L} + \delta_1(\log_Q n) \right) + O(n), \end{aligned} \qquad (21)$$

where $\delta_1(x) = \frac{1+q}{L} \sum_{j\neq 0}(2+\chi_j)\Gamma(-1-\chi_j)\mathrm{e}^{2j\pi i x}$. From this asymptotic formula we can derive in straightforward manner the desired result:

Theorem 3.1. *The difference of the variances of the total unsuccessful and the total successful search cost of the optimized skip list search algorithm fulfills*

$$\begin{aligned} \mathrm{Var}(C_n) - \mathrm{Var}(\widetilde{C}_n) &= 2\mathrm{Cov}(C_n, C_n - \widetilde{C}_n) + O(\log n) \\ &= -\frac{2(Q-1)(Q-q)}{L} n \log_Q n \; + \; O(n). \end{aligned}$$

References

[1] L. Devroye. A limit theory for random skip lists. *The Annals of Applied Probability*, 2(3):597–609, 1992.

[2] Ph. Flajolet and R. Sedgewick. Mellin transforms and asymptotics: Finite differences and Rice's integrals. *Theoretical Computer Science*, 144:101–124, 1995.

[3] I. Goulden and D. Jackson. *Combinatorial Enumeration.* J. Wiley, 1983.

[4] R.L. Graham, D.E. Knuth and O. Patashnik. *Concrete Mathematics.* Addison Wesley, 1994.

[5] P. Henrici. *Applied and Computational Complex Analysis*, Vol.1. J. Wiley, 1988.

[6] P. Kirschenhofer, C. Martinez and H. Prodinger. Analysis of an optimized search algorithm for skip lists. *Theoretical Computer Science*, 144:199-220, 1995.

[7] P. Kirschenhofer and H. Prodinger. The path length of random skip lists. *Acta Informatica*, 31:775-792, 1994.

[8] T. Papadakis, J. I. Munro, and P. V. Poblete. Average search and update costs in skip lists. *BIT*, 32:316–332, 1992.

[9] Th. Papadakis. *Skip Lists and Probabilistic Analysis of Algorithms.* PhD thesis, University of Waterloo, 1993. Available as Technical Report CS-93-28.

[10] H. Prodinger. Combinatorics of geometrically distributed random variables: Left-to-right maxima. *Discrete Mathematics.* To appear.

[11] W. Pugh. A skip list cookbook. Technical Report CS-TR-2286.1, Institute for Advanced Computer Studies, Department of Computer Science, University of Maryland, College Park, MD, June 1990. Also published as UMIACS-TR-89-72.1.

[12] W. Pugh. Skip lists: a probabilistic alternative to balanced trees. *Comm. ACM*, 33(6):668–676, 1990.