Discrete Applied Mathematics 25 (1989) 129–143
North-Holland

ON THE VARIANCE OF THE EXTERNAL PATH LENGTH IN A SYMMETRIC DIGITAL TRIE *

Peter KIRSCHENHOFER and Helmut PRODINGER
Institut für Algebra und Diskrete Mathematik, TU Vienna, A-1040 Vienna, Austria

Wojciech SZPANKOWSKI
Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA

Received 9 January 1989

In this paper we give exact and asymptotic analysis for variance of the external path length in a symmetric digital trie. This problem was open up to now. We prove that for the binary symmetric trie the variance is asymptotically equal to $4.35\ldots \cdot n + nf(\log_2 n)$ where n is the number of stored records and $f(x)$ is a periodic function with a very small amplitude.

1. Introduction

Digital searching is a well-known technique for storing and retrieving information using lexicographical (digital) structure of words. A *radix trie* (in short: trie) is such a digital search tree that edges are labelled by elements from an alphabet (e.g., binary alphabet consisting of 0's and 1's) and leaves (external nodes) contain keys [1, 4, 9]. More precisely, in a binary case a key is a (possible infinite) sequence of 0's and 1's, where 0 means "go left" and 1 means "go right". The keys are stored in external nodes and the access path from the root to a leaf is the minimal prefix information contained in an external node (see Fig. 1 for an example of a trie). There are a number of applications of tries in computer science and telecommunications, e.g., dynamic hashing, radix exchange sort [4, 9], partial match retrieval of multidimensional data, lexicographical sorting [11], tree-type conflict resolution algorithm for broadcast communications [10, 14], etc.

Two quantities of a digital trie are of special interest: the depth of a leaf and the external path length. The average depth of a leaf has been studied in [3, 9], the variance in [6] (binary case) and [14] (general tries) and higher moments of the depth in [14]. The average value of the external path length is closely related to the average depth of a leaf, but *not* the variance. The variance of the external path length was never determined up to now, although the external path length finds important applications in practice, e.g., for modified lexicographical sorting [11] and for conflict

* This research was supported in part by National Science Foundation under grant NCR-8702115.

resolution session in conflict resolution algorithms [10]. Furthermore, it was argued in [14] that the variance of the depth and the external path length provide information on *how well is a trie balanced* in a random environment, that is, under random insertions and deletions of records. This paper deals with the exact and asymptotic approximation for the variance of the external path length.

In Section 2, we state the problem to solve and show that the variance of the external path length is associated with a recurrence equation. This equation is solved exactly in Section 3. Section 2 contains our main result which is formally proved in Section 3. In particular, in Section 3 we derive the exact formula on the variance for an asymmetric trie, that is, when the occurrences of 0's and 1's in a key *are not* the same. The asymptotic approximation for the variance is restricted to symmetric (binary) tries. We prove that the variance for the binary symmetric tries is equal to $4.35\ldots \cdot n + nf(\log_2 n)$, where $f(x)$ is a periodic fluctuating function with a very small amplitude (see the theorem in Section 2). To find the asymptotic approximation we apply either Rice's method or a generalized Mellin transform approach. In fact, these approaches are useful to find an asymptotic approximation for a class of alternative sums. Moreover, the technique used in this paper is novel in the sense that certain properties of the periodic fluctuating function $f(x)$ are exploited to prove our result; in particular, to show that the term at n^2 in the formula on the variance vanishes (for more details see also [8]).

2. Statement of the problem and main results

Let T_n be a family of tries built from n records with keys from random bit streams. A key consists of 0's and 1's (binary case), and we assume that the probability of appearance of 0 and 1 in a stream is equal to p and $q = 1 - p$ respectively. The occurences of these two elements in a bit stream are independent of each other. This defines the so-called *Bernoulli model.*

Let L_n denote the external path length (random variable) in T_n, that is, the sum of the lengths of all paths from the root to all external nodes. We are interested in the average value of L_n, and the variance var L_n. In order to find them, we define the probability generating function $L_n(z)$ of L_n, that is, $L_n(z) = Ez^{L_n}$. Note that in the Bernoulli model the n records are split randomly into left subtree and right subtree of the root. If X denotes the number of keys in the left subtree, then X is Bernoulli distributed with parameter n and p. Then, for $X = k$, $L_n = n + L_k + L_{n-k}$, and finally

$$E\{z^{L_n} \mid X = k\} = z^n E z^{L_k} E z^{L_{n-k}}, \tag{2.1}$$

where L_k, L_{n-k} represent the external path length in the left subtree (with k keys) and right subtrees ($n-k$ keys). Hence, we obtain:

Lemma 1. *The probability generating function $L_n(z)$ satisfies the following recurrence*

$$L_0(z) = L_1(z) = 1,$$
$$L_n(z) = z^n \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} L_k(z) L_{n-k}(z), \quad n \geq 2. \tag{2.2}$$

Let $l_n \overset{\text{def}}{=} EL_n$ and $L_n'' = EL_n(L_n - 1)$, that is, l_n is the average value of the external path length and L_n'' is the second factorial moment of L_n. Note that $l_n = L_n'(1)$ and $L_n'' = L_n''(1)$, where $L_n'(1)$ and $L_n''(1)$ denote the first and the second derivative of $L_n(z)$ at $z = 1$. Simple algebra applied to (2.2) reveals that l_n and L_n'' satisfy the following recurrences

$$l_0 = l_1 = 0,$$
$$l_n = n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (l_k + l_{n-k}), \quad n \geq 2, \tag{2.3}$$

and

$$L_0'' = L_1'' = 0,$$
$$L_n'' = 2n l_n - n(n+1) + 2 \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} l_k l_{n-k} \tag{2.4}$$
$$+ \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} [L_k'' + L_{n-k}''].$$

Knowing l_n and L_n'' one immediately obtains the variance of L_n, as

$$\operatorname{var} L_n = L_n'' + l_n - (l_n)^2. \tag{2.5}$$

The recurrence (2.4) is a linear one. Hence, let us define three quantities v_n, u_n and w_n as

$$v_0 = v_1 = 0,$$
$$v_n = n(n+1) + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (v_k + v_{n-k}), \quad n \geq 2, \tag{2.6}$$

$$u_0 = u_1 = 0,$$
$$u_n = n l_n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (u_k + u_{n-k}), \quad n \geq 2, \tag{2.7}$$

$$w_0 = w_1 = 0,$$
$$w_n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} l_k l_{n-k} + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} [w_k + w_{n-k}], \quad n \geq 2. \tag{2.8}$$

Then

$$L_n'' = 2u_n - v_n + 2w_n. \tag{2.9}$$

Note that to compute u_n and w_n we need l_n from recurrence (2.3).

In order to find a uniform approach to solve the recurrences (2.3)–(2.8), we note that all of them are of the same type and they differ only by the first term which

we call the *additive term*. Let, in general, the additive term be denoted by a_n, where a_n is any sequence of numbers. Then the pattern for recurrences (2.3)–(2.8) is

$$\begin{aligned} &x_0 = x_1 = 0, \\ &x_n = a_n + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}(x_k + x_{n-k}), \quad n \geq 2. \end{aligned} \tag{2.10}$$

To solve (2.10), we define a sequence $\hat{a}_n$ (binomial inverse relations [12]) as

$$\hat{a}_n = \sum_{k=0}^{n} (-1)^k \binom{n}{k} a_k \;\Leftrightarrow\; a_n = \sum_{k=0}^{n} (-1)^k \binom{n}{k} \hat{a}_k. \tag{2.11}$$

Note that the exponential generating function of $\hat{a}_n$ and a_n satisfies $\hat{A}(-z) = A(z)e^{-z}$. Using this in [14] (see also [9]) it is proved that

Lemma 2. (i) *The recurrence* (2.10) *possesses the following solution*

$$x_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \frac{\hat{a}_k + ka_1 - a_0}{1 - p^k - q^k}. \tag{2.12}$$

(ii) *The inverse relatives* $\hat{x}_n$ *of* x_n *satisfy*

$$\hat{x}_n = \frac{\hat{a}_n + na_1 - a_0}{1 - p^n - q^n}, \quad n \geq 2. \tag{2.13}$$

Finally, to find asymptotic approximations for x_n, we apply a general approach proposed either in [3] (Rice's method) or in [13, 15] (Mellin-like approach, see also Knuth [9]). Namely, we consider an alternative sum of the form $\sum_{k=2}^{n} (-1)^k \binom{n}{k} f(k)$ where $f(k)$ is any sequence. This sum appears in our Lemma 2. Then:

Lemma 3. (i) (Rice's method [3, 6]). *Let* C *be a curve surrounding the points* $2, 3, \ldots, n$ *and* $f(z)$ *be an analytical continuation of* $f(k)$ *in* C. *Then*

$$\sum_{k=2}^{n} \binom{n}{k} (-1)^k f(k) = \frac{-1}{2\pi i} \int_C [n; z] f(z) dz \tag{2.14}$$

with

$$[n; z] = \frac{(-1)^{n-1} n!}{z(z-1)\cdots(z-n)}.$$

(ii) (Mellin-like approach [13, 15]). *Let*

$$S_{m,r}(n) = \sum_{k=m}^{n} (-1)^k \binom{n}{k} \binom{k}{r} f(k),$$

and $f(-z)$ *is an analytical continuation of* $f(k)$ *left to the line* $(\frac{1}{2} - [m-r]^+ - i\infty, \frac{1}{2} - [m-r]^+ + i\infty)$, *where* $a^+ = \max\{0, a\}$. *Then*

$$S_{m,r}(n+r) = \frac{(-1)^r}{r!} \int_{(1/2-[m-r]^+)} \Gamma(z) f(r-z) n^{r-z} \mathrm{d}z + e_n, \tag{2.15}$$

where $\int_{(c)}$ *stands for* $1/2\pi\mathrm{i} \int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty}$; $\Gamma(z)$ *is the gamma function* [1, 5], *and*

$$e_n = \mathrm{O}(n^{-1}) \int_{(1/2-[m-r])} z\Gamma(z) f(r-z) n^{r-z} \mathrm{d}z,$$

that is, $e_n = \mathrm{o}(n)$.

Proof. Both formulas are a consequence of Cauchy's Theorem [5]. The proof of (2.14) is given in [3], while (2.15) is established in [15]. Note, however, that some restrictions on $f(z)$ must be imposed. Roughly speaking, $f(z)$ cannot grow too fast at infinity. The details can be found in [15]. □

To apply Lemma 3(i) for asymptotic analysis, we change C to a larger curve around which the integral is small, and take into account residues at poles in the larger enclosed area. To apply Lemma 3(ii) we find residues *right* to the line $(c-\mathrm{i}\infty, c+\mathrm{i}\infty)$ where $c = \frac{1}{2} - [m-r]^+$. It is proved that (for simplicity $r=0$ is assumed in (2.15))

$$\begin{aligned} \sum_{k=2}^{n} (-1)^k \binom{n}{k} f(k) &= \sum \mathrm{res}\{[n; z] f(z)\} + \mathrm{O}(n^{-M}) \\ &= \sum \mathrm{res}\{\Gamma(z) f(-z) n^{-z}\} + \mathrm{O}(n^{-M}) \end{aligned} \tag{2.16}$$

for any $M > 0$, and the sums are taken over all poles of the functions under the integrals (2.14) and (2.15) in the appropriate regions respectively. By (2.16), the asymptotics of the alternative sum of type (2.12) (Lemma 2) is reduced to compute the residues of the functions under the integrals, which is usually an easy task.

Using Lemmas 1–3 we prove in Section 3 our main result:

Theorem. *The variance of the external path length in a binary symmetric trie* (*i.e.*, $p=q=0.5$) *consisting of n records* (*external nodes*) *is asymptotically equal to*

$$\mathrm{var}\, L_n = n[A + f(\log_2 n)] + \mathrm{O}(\ln^2 n), \tag{2.17}$$

where

$$A = 1 + \frac{1}{2\ln 2} - \frac{1}{\ln^2 2} + \frac{2}{\ln 2}(\mu + \upsilon) + \tau, \tag{2.18}$$

$$\mu = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k(2^k-1)}, \qquad \upsilon = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{2^k-1}, \tag{2.19}$$

$$\tau = -\frac{4\pi^2}{\ln^3 2} \sum_{k=1}^{\infty} \frac{k}{\sinh(2k\pi^2/\ln 2)} \tag{2.20}$$

and $f(x)$ *is a continuous periodic funtion with period* 1 *and very small amplitude and mean zero* (*the contribution from* τ *is also very small*).

Numerical evaluation reveals that $\operatorname{var} L_n = 4.35\ldots \cdot n + nf(\log_2 n)$.

Before we proceed to the proof of the theorem, we first offer some remarks and extension of the main result.

Remark. (i) *Why the symmetric case?* The reader may be surprised why, having Lemmas 2 and 3, we restrict our asymptotic analysis to the symmetric case. In fact, in the next section we shall see that after applying Lemma 2 an exact formula on L_n'' is available. Nevertheless, for asymptotics, according to Lemma 3, we need an analytical continuation of $\hat{a}_n$ (see (2.12)), where a_n is the additive term in the recurrence (2.10). This is easy to achieve for v_n and u_n (see (2.6) and (2.7)), but very difficult for w_n (see (2.8)). Fortunately, in the symmetric case, such an analytical continuation is available (see equation (3.16)). An easy extension to the asymmetric case is not known up to date. Our guess is that for a symmetric case another approach is necessary (see also Remark (ii)).

(ii) *Extension to V-ary tries.* The methodology provided in this paper can be used to derive exact and asymptotic (symmetric case) analysis for V-ary digital tries. To define this trie, let A be an alphabet containing V elements, i.e., $A = \{\sigma_1, \sigma_2, \ldots, \sigma_V\}$, and let S denote the set of finite numbers, say n, of strings (keys) from A. The probability of occurrence of an element from A, say σ_i, $i = 1, 2, \ldots, V$, in a string is denoted as p_i, where $\sum_{i=1}^{V} p_i = 1$. The branching policy on level k in a V-ary trie is based on the kth element of a key. For example, if the kth element is σ_1, then we go to the leftmost subtree, if it is σ_2, we move to the next leftmost subtree, etc. An example of a 3-ary digital trie is shown in Fig. 1.

Let now L_n be the length of the external path in a V-ary digital trie. Under our Bernoulli model, the V subtrees of the root contain $k_1, k_2, \ldots, k_V$ elements with probability

$$\binom{n}{k_1, \ldots, k_V} p_1^{k_1}, \ldots, p_V^{k_V}$$

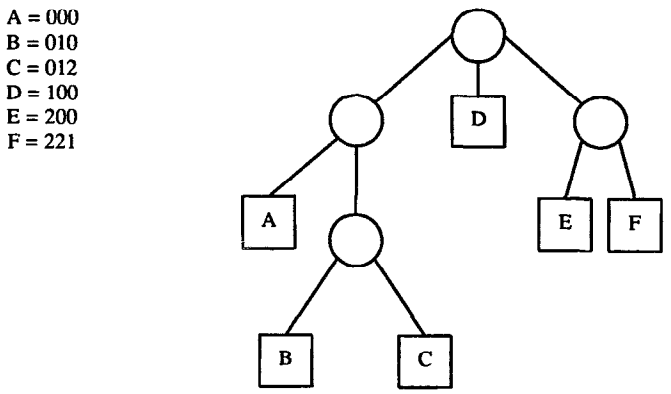

Fig. 1. Example of 3-ary digital trie with $n = 6$.

where

$$\binom{n}{k_1,\ldots,k_V} = \frac{n!}{k_1!k_2!,\ldots,k_V!} \quad \text{and} \quad k_1+k_2+\cdots+k_V = n.$$

Then $L_n = n + L_{k_1} + \cdots + L_{k_V}$ and the probability generating function $L_n(z)$ of L_n satisfies

$$\begin{aligned} L_0(z) &= L_1(z) = 1, \\ L_n(z) &= z^n \sum_{\{k_1+\cdots+k_V=n\}} \binom{n}{k_1\cdots k_V} p_1^{k_1}\cdots p_V^{k_V} L_{k_1}(z)\cdots L_{k_V}(z). \end{aligned} \tag{2.21}$$

In particular, the average value of L_n is equal to $L_n'(1)$ and the variance is related to $L_n'' \overset{\text{def}}{=} L_n''(1)$ as in (2.5). The exact formula for L_n'' follows from the same type of analysis as before. Also, the same type of difficulties arise to obtain asymptotic approximation, hence restriction to the symmetric case (i.e., $p_1 = p_2 = \cdots = p_v = 1/V$) is imposed. Then, copying the analysis from the binary case, one proves that the theorem holds with

$$A = 1 + \frac{1}{V\ln V} - \frac{1}{\ln^2 V} + \frac{V}{\ln V}(\mu+\upsilon) + \tau, \tag{2.22}$$

where

$$\mu = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k(V^k-1)}, \qquad \upsilon = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{V^k-1},$$

τ is very small and can be savely ignored in practice. Finally, let us point out that the variance of the external path length in the *asymmetric* case is qualitatively different from the symmetric case. Although we do not present the analysis here, it is possible to prove, using the results from [14], that in the asymmetric case $\operatorname{var} L_n = \Omega(n \log_2 n)$.

(iii) *The covariance analysis.* The theorem and the results from [6, 14], where the variance of the depth of an external node was established, provide asymptotics for the covariance between two different depths of leaves in a trie. Let D_n be a depth of an external node, and let $D_n^{(i)}$ be a path from the root to the ith external node. Note that the external path length L_n is defined in terms of $D_n^{(i)}$ as $L_n = \sum_{i=1}^n D_n^{(i)}$. Then

$$\operatorname{var} L_n = E\left\{\left[\sum_{i=1}^n D_n^{(i)}\right]^2\right\} - \left\{E\sum_{i=1}^n D_n^{(i)}\right\}^2,$$

and this implies (see [14])

$$\operatorname{var} L_n = n \operatorname{var} D_n + 2\sum_{i\neq j} \operatorname{cov}\{D_n^{(i)}, D_n^{(i)}\}. \tag{2.23}$$

The variance of the depth, $\operatorname{var} D_n$, was analyzed in [6,14]. In particular, it was proved that for binary symmetric tries $\operatorname{var} D_n = 3.507\ldots$. Using our main result and (2.23), we find

$$2\sum_{i\neq j} \operatorname{cov}\{D_n^{(i)}, D_n^{(j)}\} = 0.84\ldots \cdot n. \tag{2.24}$$

This also implies, in the symmetric case, that $\operatorname{cov}\{D_n^{(i)}, D_n^{(j)}\} \sim 0.84/n$.

(iv) The methodology established in this paper can be also applied to estimate the variance of the external path length for other digital trees, that is, Patricia tries and digital search tries [1,4]. In particular, in a forthcoming paper, we shall show that for the Patricia trie $\operatorname{var} L_n \sim 0.35\ldots \cdot n + nf(\log_2 n)$, however, it should be pointed out that the analysis in that case is much more intricate.

3. The analysis

In this section, we present an exact solution for recurrences (2.3)-(2.8), and asymptotic analysis for the binary *symmetric* case ($p=q=0.5$).

3.1. Exact analysis

To solve (2.3) for l_n note that $a_n = n$ and $\hat{a}_n = -\delta_{n,1}$ [9,12], where $\delta_{n,k}$ is the Kronecker delta. Then, immediately from Lemma 2 we find

$$l_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \frac{k}{1-p^k-q^k}, \quad n \geq 2, \tag{3.1}$$

$$\hat{l}_n = \frac{n}{1-p^n-q^n}, \quad n \geq 2. \tag{3.2}$$

To solve (2.6) for v_n, note that $a_n = n(n+1) = 2\binom{n}{2} + 2n$. From [9,12], we know that for $b_n = \binom{n}{r}$ the inverse relation is $\hat{b}_n = (-1)^r \delta_{n,r}$. Hence, $\hat{a}_n = 2\delta_{n,2} - 2\delta_{n,1}$. By Lemma 2 we obtain

$$v_n = \frac{n(n-1)}{1-p^2-q^2} + 2\sum_{k=2}^{n} (-1)^k \binom{n}{k}\binom{k}{1} \frac{1}{1-p^k-q^k}. \tag{3.3}$$

For u_n given by (2.7), we need the inverse relation for $a_n = n l_n$. Using generating functions and the fact $\hat{A}(-z) = A(z)e^{-z}$ one easily proves that $\hat{a}_n = n\hat{l}_n - n\hat{l}_{n-1}$ where $\hat{l}_n$ is given by (3.2). Hence by Lemma 2

$$u_n = 2\sum_{k=2}^{n} (-1)^k \binom{n}{k}\binom{k}{2} \frac{1}{(1-p^k-q^k)^2} + \sum_{k=2}^{n} (-1)^k \binom{n}{k}\binom{k}{1} \frac{1}{1-p^k-q^k}$$
$$-2\sum_{k=3}^{n} (-1)^k \binom{n}{k}\binom{k}{2} \frac{1}{(1-p^k-q^k)(1-p^{k-1}-q^{k-1})}. \tag{3.4}$$

The most difficult is w_n since $a_n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} l_k l_{n-k}$. However, let $a(z)$ and $l(z)$ denote the exponential generating functions for a_n and l_n respectively. Then, $a(z) = l(zp)l(zq)$, and this implies $\hat{a}(-z) = \hat{l}(-zp)\hat{l}(-zq)$. Hence

$$\hat{a}_n = \sum_{k=2}^{n-2} \binom{n}{k} p^k q^{n-k} \hat{l}_k \hat{l}_{n-k}, \quad n \geq 4, \tag{3.5}$$

and $\hat{a}_0=\hat{a}_1=\hat{a}_2=\hat{a}_3=0$. Then the solution for w_n follows from Lemma 2. We return to that problem later, since (3.5) is not very suitable for analytical continuation needed in Lemma 3.

3.2. Asymptotic approximation

Hereafter we assume $p=q=0.5$, that is, only *binary symmetric* tries are considered. We obtain asymptotic approximations for v_n and w_n, through Lemma 3(ii) and for w_n by Lemma 3(i), however, both methods are equivalent.

Let us start with v_n. Note that $v_n=2l_n+2n^2-2n$. Using the asymptotic expression for l_n from [6, 9, 14] we immediately obtain (see also (3.24))

$$v_n = 2n^2 + \frac{2n\ln n}{L} + n\left[\frac{2\gamma}{L} - 1\right] + 2n\delta(\log_2 n), \tag{3.6}$$

where $\gamma=0.577...$ is the Euler constant, $L \stackrel{\text{def}}{=} \ln 2$, and

$$\delta(x) = \frac{1}{L}\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \omega_k \Gamma(-\omega_k) e^{2k\pi i x}, \tag{3.7}$$

where

$$\omega_k = 1 + \frac{2k\pi i}{L}. \tag{3.8}$$

The ω_k, $k=0, \pm 1, \ldots$, are solutions of the following equation

$$1 - 2^{1-z} = 0, \tag{3.9}$$

where z is a complex number.

The evaluation of u_n is much more intricate. Using (3.4) with $p=q=0.5$ one proves

$$\frac{u_{n+1}}{n+1} = 8n + \sum_{k=2}^{n} (-1)^k \binom{n}{k} \frac{1}{(1-2^{-k})^2}\left[\frac{k}{2(2^{k-1}-1)} - 1\right].$$

Hence by Lemma 3(ii)

$$\frac{u_{n+1}}{n+1} - 8n = \int_{(-3/2)} \frac{\Gamma(z)n^{-z}}{(1-2^z)^2}\left[\frac{-z}{2(2^{-z-1}-1)} - 1\right]dz + O(n^{-1}). \tag{3.10}$$

The evaluation of the integral is standard and appeals to the residue theorem. Note that the function under the integral has two poles: $-\omega_k$ given by (3.8) and $\chi_k=2k\pi i/L$ for $k=0, \pm 1, \pm 2, \ldots$ ($\omega_k=1+\chi_k$). For $k=0$, $-\omega_0=-1$ is a double pole, while $\chi_0=0$ is a triple pole since $z=0$ and $z=-1$ are singular points for $\Gamma(z)$. It is also well known that the main contribution to the asymptotic approximation comes from the real part of the poles, that is, $-\omega_0$ and χ_0. For $-\omega_0=-1$ we use the following Taylor expansions. Let $w=z+1$, then

$$\Gamma(z) = -w^{-1} - (\gamma - 1) + O(w), \tag{3.11a}$$

$$n^{-z} = n(1 - w \ln n) + O(w^2), \tag{3.11b}$$

$$\frac{1}{2^{-z-1}-1} = \frac{-1}{Lw}(1 + \tfrac{1}{2}Lw + \tfrac{1}{12}L^2w^2). \tag{3.11c}$$

For $\chi_0 = 0$ we have [1, 5]

$$\Gamma(z) = z^{-1} + \gamma + \tfrac{1}{2}[\tfrac{1}{6}\pi^2 + \gamma^2]z + O(z^2), \tag{3.12a}$$

$$n^{-z} = 1 - z \ln n + \tfrac{1}{2}z^2 \ln^2 n + O(z^3), \tag{3.12b}$$

$$\frac{1}{2(2^{-z-1}-1)} = -(1 + Lz), \tag{3.12c}$$

$$\frac{1}{(1-2^{-z})^2} = \frac{1}{L^2z^2}(1 + Lz + \tfrac{5}{12}L^2z^2). \tag{3.12d}$$

Multiplying (3.11) and (3.12) and taking the coefficient at z^{-1} and w^{-1} we find the contribution from $-\omega_0$ and χ_0 which yields

$$u_{n,1} = \frac{n}{L}\left[2n \ln n + n(2\gamma - L) + \frac{1}{2L}\ln^2 n + \left(\frac{\gamma+1}{L} - 1\right)\ln n + \left(\frac{\gamma}{L} + \frac{\pi^2}{12L} + \frac{\gamma^2}{2L} + \frac{17}{12}L - 1 - \gamma\right)\right]. \tag{3.13}$$

The contribution from $-\omega_k$ and χ_k, $k \neq 0$ can be found in a similar way. Calculations reveal that

$$u_{n,2} = \frac{n^2}{L}\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} \exp[2k\pi i \log_2 n]\left\{2\left[\omega_k\Gamma(-\omega_k) + \frac{c_1\omega_k}{n} - \frac{\omega_k^2\Gamma(-\omega_k)}{n}\right] + (-\omega_k)[(-\omega_k)\Gamma(-\omega_k)\ln n + \omega_k\Gamma'(-\omega_k) - \Gamma(-\omega_k) - L\Gamma(-\omega_k) - \Gamma(-\omega_k)(1 + 2L\chi_k)]\right\} \tag{3.14}$$

$$c_1 = -\tfrac{1}{2}\chi_k\omega_k\Gamma(-\omega_k). \tag{3.15}$$

Finally, $u_n \sim u_{n,1} + u_{n,2}$ and it turns out that the contribution from $u_{n,2}$ is very small.

The most difficult part is the asymptotic approximation for w_n, since we need an analytical continuation for $\hat{a}_n$ given by (3.5), where a_n is the additive term in the recurrence (2.8) on w_n. Fortunately for the symmetric case, it is relatively easy to obtain $\hat{a}_n$, however, we need a further consideration to find it. Note that for $a_n = 2^{-n}\sum_{k=0}^{n}\binom{n}{k} l_k l_{n-k}$ the exponential generating functions $a(z)$ and $l(z)$ are related as $a(z) = [l(\frac{1}{2}z)]^2$, and $\hat{a}(z) = [\hat{l}(\frac{1}{2}z)]^2$. From (2.3) with $p = q = 0.5$ we immediately find

$$\hat{l}(z) = z(e^z - 1) + 2\hat{l}(\tfrac{1}{2}z).$$

Hence

$$[\hat{l}(z)]^2 = z^2(e^{2z} - 2e^z + 1) + 4z(e^z - 1)\hat{l}(\tfrac{1}{2}z) + [\hat{l}(\tfrac{1}{2}z)]^2$$

and equating coefficients of both sides of the above, we finally obtain for $n \geq 3$ (note that $\hat{l}^2(z) = \hat{a}(2z)$)

$$\hat{a}_n = \frac{n(n-1)}{2} \frac{1}{2^{n-2}-1} \left[2^{n-3} - 1 + \sum_{j=1}^{\infty} \binom{n-2}{j} \frac{1}{2^j - 1} - \frac{1}{2^{n-2}-1} \right]. \quad (3.16)$$

Hence, by Lemma 2, w_n has a solution

$$\frac{w_{n+1}}{n+1} = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \frac{2^{k-1}}{2^k - 1} \frac{k}{2^{k-1}-1} \times \left\{ 1 - 2^{k-2} + \frac{1}{2^{k-1}-1} - \sum_{j=1}^{\infty} \binom{n-1}{j} \frac{1}{2^j - 1} \right\}. \quad (3.17)$$

For asymptotic analysis of (3.17), we apply Rice's method to illustrate how it works. Note that the analytical continuation of $f(k)$ in (3.17) is easy, since the last series in (3.17) can be extended as $\sum_{j=1}^{\infty} \binom{z-1}{j} (1/(2^j - 1))$ (it can be proved that the series is convergent for all z). Hence, using Rice's method

$$\frac{w_{n+1}}{n+1} = -\frac{1}{2\pi i} \int_C [n; z] f(z) dz,$$

where

$$f(z) = \frac{z 2^{z-1}}{(2^z - 1)(2^{z-1} - 1)} \left\{ 1 - 2^{z-2} + \frac{1}{2^{z-1}-1} - \sum_{j=1}^{\infty} \binom{z-1}{j} \frac{1}{2^j - 1} \right\}. \quad (3.18)$$

We extend now the circle of the integration such that the poles of $f(z)$ are included, that is, the points ω_k and χ_k, $k = 0, \pm 1, \ldots$. We evaluate separately the residues of the function under the integral for $\omega_0 = 1$, $\chi_0 = 0$ and ω_k, χ_k, $k \neq 0$. We use the Taylor expansions already presented in (3.11) and (3.12). In addition, we have for $w = z - 1$ (see [3, 5, 7])

$$[n; z] \sim \frac{n}{w} \left[1 + w(H_{n-1} - 1) + w^2(1 - H_{n-1} + \tfrac{1}{2}H_{n-1}^2 + \tfrac{1}{2}H_{n-1}^{(2)}) \right], \quad (3.19a)$$

$$\sum_{j=1}^{\infty} \binom{z-1}{j} \frac{1}{2^j - 1} \sim \mu w + O(w^2) \quad \text{for } z \to 1, \quad (3.19b)$$

$$\sum_{j=1}^{\infty} \binom{z-1}{j} \frac{1}{2^j - 1} \sim -\upsilon + O(z) \quad \text{for } z \to 0, \quad (3.19c)$$

where H_n, $H_n^{(2)}$ are harmonic and generalized harmonic numbers [9], μ and υ are defined in (1.3). Then the contributions from $\omega_0 = 1$ and $\chi_0 = 0$ to w_n denoted as $w_{n,1}$, $w_{n,2}$ are respectively

$$w_{n,1} = \frac{n}{L^2} \left\{ \tfrac{1}{2} n \ln^2 n + \gamma n \ln n - \tfrac{3}{2} L n \ln n + n\alpha - \tfrac{1}{2} \ln^2 n + (\tfrac{3}{2}L - \gamma - \tfrac{3}{2}) \ln n - \alpha + \tfrac{9}{4}L - \tfrac{3}{2}\gamma - \tfrac{1}{2} \right\}, \quad (3.20)$$

where $\alpha = \frac{5}{3}L^2 - \frac{3}{2}L\gamma - L\mu + \frac{1}{2}\gamma^2 + \frac{1}{12}\pi^2$, and

$$w_{n,2} = \frac{n}{L}\left\{-\tfrac{5}{4} + v\right\}. \tag{3.21}$$

To find the contribution from ω_k and χ_k, $k \neq 0$, we use the following Taylor expansions for $u = z - \omega_k$:

$$[n; z] \sim n^{\omega_k}\{\Gamma(-\omega_k) + c_1/n + u[-\Gamma'(-\omega_k) + \Gamma(-\omega_k)\ln n + c_1 \ln n/n + c_3/n]\},$$

where

$$c_3 = 0.5\chi_k\omega_k\Gamma'(-\omega_k) - \Gamma(-\omega_k)\chi_k - 0.5\Gamma(-\omega_k),$$

and c_1 is given in (3.15). Then the contribution $w_{n,3}$ from ω_k, $k \neq 0$ is

$$w_{n,3} = \frac{n^2 \ln n}{L}\delta(\log_2 n) + \frac{n^2}{L^2}\sigma_1(\log_2 n) + \frac{n \ln n}{L^2}\sigma_2(\log_2 n) + \frac{n}{L^2}\sigma_3(\log_2 n), \tag{3.22}$$

where $\delta(x)$ is defined in (3.7) while $\sigma_i(x)$, $i = 1, 2, 3$ are complicated fluctuating functions with very small amplitude (see also (3.27b)). Finally, the contribution from χ_k, $k \neq 0$ is

$$w_{n,4} \sim \frac{1}{L}\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} n^{\omega_k}\Gamma(-\omega_k)\omega_k\chi_k\left[-\frac{5}{4} - \sum_{j=1}^{\infty}\binom{\chi_k - 1}{j}\frac{1}{2^j - 1}\right] \tag{3.23}$$

and $w_n = w_{n,1} + w_{n,2} + w_{n,3} + w_{n,4} + O(\ln^2 n)$.

To complete our analysis, we need an asymptotic approximation for l_n. But from [6, 14] we have

$$l_n \sim \frac{n \ln n}{L} + n\left[\frac{\gamma}{L} + \tfrac{1}{2} + \delta(\log_2 n)\right] - \frac{1}{2L} + \delta_1(\log_2 n), \tag{3.24}$$

where $\delta(x)$ is defined in (3.7) and

$$\delta_1(x) = -\frac{1}{L}\sum_{k\neq 0}\tfrac{1}{2}\omega_k^2\chi_k\Gamma(-\omega_k)e^{2\pi kix}. \tag{3.25}$$

Now, the variance of L_n is given by $\operatorname{var} L_n = 2u_n - v_n + 2w_n + l_n - l_n^2$, and after some tedious algebra, one finds

$$\operatorname{var} L_n = Bn^2 + An + O(\ln^2 n) \tag{3.26}$$

where

$$B = -\frac{11}{12} - \frac{2\mu}{L} + \frac{\pi^2}{6L^2} - \delta^2(\log_2 n) + \left(3 - \frac{2\gamma}{L}\right)\delta(\log_2 n) + \frac{2}{L^2}\sigma_1(\log_2 n), \tag{3.27a}$$

$$\sigma_1(x) = \sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty} e^{2\pi k i x}\Big\{\Gamma(-\omega_k) - \tfrac{1}{2}L\omega_k\Gamma(-\omega_k) - \omega_k\Gamma'(-\omega_k) - L\omega_k\Gamma(-\omega_k)\sum_{n\geq 1}\binom{\chi_k}{n}\frac{1}{2^n-1}\Big\} \tag{3.27b}$$

and A is given in the main theorem (see (2.18)).

To prove the main theorem, we need to show that $B=0$. Let us first consider the Fourier coefficient of $\delta^2(x)$ for $k=0$. We denote it by δ_0. Then from (3.7)

$$\delta_0 = \frac{1}{L^2}\sum_{\substack{l+m=0 \\ l,m\neq 0}} \omega_l\omega_m\Gamma(-\omega_l)\Gamma(-\omega_m)$$

$$= \frac{2}{L^2}\sum_{l=1}^{\infty}\Gamma(\chi_l)\overline{\Gamma(\chi_l)} = \frac{1}{L}\sum_{l=1}^{\infty}\frac{1}{l\sinh(2l\xi)}, \tag{3.28}$$

where $\xi=\pi^2/L$. The last equality follows from [1, 5]

$$|\Gamma(iy)|^2 = \frac{\pi}{y\sinh(\pi y)}.$$

We further can transform (3.28) as

$$\delta_0 = \frac{2}{L}\sum_{l=1}^{\infty}\frac{2}{l}\sum_{n=0}^{\infty}e^{-2l\xi(2n+1)} = \frac{-2}{L}\sum_{n=0}^{\infty}\log(1-e^{2\xi(2n+1)}).$$

This can be rewritten as

$$\delta_0 = \frac{2}{L}\ln\prod_{n=1}^{\infty}\frac{1}{1-e^{-2\xi n}} - \frac{2}{L}\ln\prod_{n-1}^{\infty}\frac{1}{1-e^{-4\xi n}}. \tag{3.29}$$

Using a functional equation for the Dedekind η-function Kirschenhofer, Prodinger and Schoissengeier [6] have proved that δ_0 can be reduced to (see Appendix A)

$$\delta_0 = \frac{\pi^2}{6L^2} - \frac{11}{12} - 2\frac{\mu}{L}. \tag{3.30}$$

Hence the term B in (3.27) can be transformed into

$$B = -\delta_2^2(\log_2 n) + \left[3 - \frac{2\gamma}{L}\right]\delta(\log_2 n) + \frac{2}{L^2}\sigma_1(\log_2 n) = \delta_3(\log_2 n),$$

where $\delta_2^2(x)$ is the function $\delta^2(x)-\delta_0$. Note that now B is expressed in terms of periodic function $\delta_3(x)$, with very small amplitude *and* mean zero. This function is continuous (since the Fourier series associated with the function is absolutely convergent). Assume now $\delta_3(x)$ is not identically zero. Then, $\delta_3(x)$ would take values, say less than $-\varepsilon$, for arguments in an interval, say $[a,b]$. Since $\log_2 n$ is dense modulo 1, the leading factor of the variance would be negative for infinitely many

values of n. This is a contradiction, since $\operatorname{var} L_n \geq 0$ for all n. Hence $\delta_3(x) \equiv 0$ and thus $B=0$. This completes the proof of the main theorem.

Appendix A

Proof of (3.30). Let us define a function

$$g(x) = \ln \prod_{n=1}^{\infty} \frac{1}{1-e^{-2\pi n x}}. \tag{A.1}$$

Then, by (3.24)

$$\delta_0 = \frac{2}{L} g(\xi/\pi) - \frac{2}{L} g(2\xi/\pi), \tag{A.2}$$

where $\xi = \pi^2/L$. To estimate (A.2), we introduce a new function

$$f(x) = \ln \prod_{n=1}^{\infty} (1+e^{-nx}), \tag{A.3}$$

which can be rewritten as

$$f(x) = \sum_{n=1}^{\infty} \ln(1+e^{-nx}) = \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} (-1)^{k-1} \frac{e^{-nxk}}{k} = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k(e^{kx}-1)}.$$

Note that μ as defined in (2.14) is equal to

$$\mu = f(\ln 2). \tag{A.4}$$

Since $\prod_{n=1}^{\infty} (1+q^n) = \prod_{n=0}^{\infty} 1/(1-q^{2n+1})$ the function $f(x)$ can be represented in terms of $g(x)$ as

$$f(x) = \ln \prod_{n=0}^{\infty} \frac{1}{1-e^{-(2n+1)x}} = g(x/2\pi) - g(x/\pi). \tag{A.5}$$

To estimate the RHS of (A.5), we apply a functional equation for the Dedekind η-function. The η-function is defined as [2]

$$\eta(x) = e^{\pi i x/12} \prod_{n=1}^{\infty} (1-e^{2\pi i n x}), \quad \operatorname{Im} x > 0, \tag{A.6}$$

and it satisfies the following functional equation [2]

$$\ln \eta(i/x) - \ln \eta(ix) = \tfrac{1}{2} \ln x. \tag{A.7}$$

But in [2, p. 48], it is also shown that

$$\ln \eta(ix) = \tfrac{1}{12} \pi x - g(x),$$

where $g(x)$ is defined in (A.1). Therefore, the above and (A.7) imply

$$g(1/x) - g(x) = \tfrac{1}{12} \pi (x - 1/x) - \tfrac{1}{2} \ln x, \quad x > 0. \tag{A.8}$$

Using now (A.5), (A.8) and the following

$$f(x) = g(x/2\pi) - g(x/\pi) - g(2\pi/x) + g(\pi/x) - f(2\pi^2/x)$$

one proves immediately that

$$f(x) = \tfrac{1}{12}\frac{\pi^2}{x} - \tfrac{1}{2}\ln 2 + \tfrac{1}{24}x - f(2\pi^2/x), \tag{A.9}$$

and by (A.2)

$$\delta_0 = \frac{2}{L} f(2\xi). \tag{A.10}$$

But $f(\ln 2) = \mu$, so by (A.9)

$$\mu = \tfrac{1}{12}\frac{\pi^2}{\ln 2} - \tfrac{1}{2}\ln 2 + \tfrac{1}{24}\ln 2 - f(2\xi),$$

and (3.30) follows from (A.10) and the above. □

References

[1] A. Aho, J. Hopcroft and J. Ullman, Data Structures and Algorithms (Addison-Wesley, Reading, MA, 1983).

[2] T. Apostol, Modular Functions and Dirichlet Series in Number Theory (Springer, New York, 1976).

[3] Ph. Flajolet and R. Sedgewick, Digital search trees revisited, SIAM J. Comput. 15 (1986) 748–767.

[4] G. Gonnet, Handbook of Algorithms and Data Structures (Addison-Wesley, Reading, MA, 1986).

[5] P. Henrici, Applied and Computational Complex Analysis (Wiley, New York, 1977).

[6] P. Kirschenhofer and H. Prodinger, Some further results on digital search trees, in: L. Kott, ed., Automata, Languages and Machines, Lectures Notes in Computer Science 226 (Springer, Berlin, 1986) 177–185.

[7] P. Kirschenhofer, H. Prodinger and J. Schoissengeier, Zur Auswertung gewisser Reihen mit Hilfe modularer Functionen, in: E. Hlawka, ed., Zahlentheoretische Analysis 2, Lecture Notes in Mathematics 1262 (Springer, Berlin, 1987) 108–110.

[8] P. Kirschenhofer and H. Prodinger, On some applications of formulae of Ramanujan in the analysis of algorithms, Preprint (1988).

[9] D. Knuth, The Art of Computer Programming 3: Sorting and Searching (Addison-Wesley, Reading, MA, 1973).

[10] P. Mathys and P. Flajolet, Q-ary collision resolution algorithms in random-access system with free and blocked channel access, IEEE Trans. Inform. Theory 31(2) (1985) 217–243.

[11] R. Paige and R. Tarjan, Three efficient algorithms based on partition refinement, Preprint (1986).

[12] J. Riordan, Combinatorial Identities (Wiley, New York, 1968).

[13] W. Szpankowski, Two problems on the average complexity of digital trees, Proceedings Performances-87, Brussel (1987) 189–208.

[14] W. Szpankowski, Some results on V-ary asymmetric tries, J. Algorithms 9 (1988) 224–244.

[15] W. Szpankowski, The evaluation of an alternative sum with applications to the analysis of some data structures, Inform. Process. Lett. 28 (1988) 13–19.