DIGITAL SEARCH TREES AGAIN REVISITED: THE INTERNAL PATH LENGTH PERSPECTIVE

Peter Kirschenhofer†, Helmut Prodinger†,
and
Wojciech Szpankowski*

†Department of Algebra and Discrete Mathematics
Technical University of Vienna, Austria
and
*Department of Computer
Science, Purdue University
West Lafayette,U.S.A.

Abstract. This paper studies the asymptotics of the variance for the internal path length in a symmetric digital search tree. The problem was open up to now. We prove that for the binary digital search tree the variance is asymptotically equal to $0.26600\ldots \cdot N + N\delta(\log_2 N)$ where N is the number of stored records and $\delta(x)$ is a periodic function of mean zero and a very small amplitude. This result implies that the internal path length becomes asymptotically $N \cdot \log_2 N$ *with probability one* (i.e. almost surely). In our previous work we have argued that the variance of the internal (external) path length is a good indicator how well the digital trees are balanced. We shall show that the digital search tree is the best balanced digital tree in the sense that a random shape of this tree strongly resembles a shape of a complete tree. Therefore, we conclude that a symmetric digital tree is a good candidate for a dictionary structure, and a typical search time is asymptotically equal to the optimal one for these type of structures. Finally, in order to prove our result we had to solve a number of nontrivial problems concerning analytic continuations and some others of numerical nature. In fact, our results and techniques are motivated by the methodology introduced in an influential paper by Flajolet and Sedgewick.

1. INTRODUCTION

Digital trees [2], [7], [14] experience a new wave of interest due to a number of novel applications in computer science and telecommunications. For example, recent developments in the context of large external files and ideas derived from the dynamic hashing (virtual hashing, dynamic hashing, extendible hashing) lead to the

†This research was supported by Fonds zur Förderung der Wissenschaftlichen Forschung Projekt No. P7497-TEC
*This research was primary supported by NSF Cooperative Grant INT-8912631. Additional funding was available from AFOSR grant 90-0107, NSF grant CCR-8900305, and grant R01 LM05118 from the National Library of Medicine

Typeset by $\mathcal{A}_{\mathcal{M}}\mathcal{S}$-TEX

analysis of digital trees [6], [8], [9], [11], [12], [19], [22], [23], [24]. In telecommunications, recent developments in conflict resolution algorithms [16] have also brought a new interest in digital trees. Some other applications are: radix exchange sort, polynomial factorizations, simulation, Huffman's algorithm, etc., [2], [7], [14].

The three primary digital tree search methods are: *digital search trees (DST)*, *radix search tries* (shortly: tries), and *Patricia tries* [2], [7], [14]. In all cases, a digital tree is built over a V-ary alphabet $\mathcal{A} = \{\omega_1, \ldots, \omega_V\}$. Records stored in a tree, say n of them, consist of (possibly infinite) strings (keys) from $\mathcal{A}$. A *digital search tree* [2], [4] is a data structure that leads to much improved worst case performance, by making use of the digital properties of the key. The idea is to build a structure consisting of nodes such that each node has a record containing a key and V links which point to subtrees. The branching policy on a level, say k, is based on the k-th digit (element) of a key. For example, if the k-th element of the key is ω_1, then we go to the leftmost subtree; if it is ω_2, we move to the next of the leftmost subtree, etc. However, if keys are very long, then comparisons of keys at each level of the tree might be quite costly. To avoid this, in the *radix search trie* we do not store keys in the tree nodes (internal nodes), but rather store all the keys in the external nodes of the tree. However, such a radix trie has an annoying flaw: there is "one-way branching" which leads to the creation of extra nodes in this tree. D.R.Morrison discovered a way to avoid this problem in a data structure which he named the *Patricia trie*. In such a tree, all nodes have branching degree greater than or equal to two. This is achieved by collapsing one-way branches on internal nodes, that is, by avoiding unary nodes (cf. [14] and [24]). Note that the number of internal nodes in the digital search tree and the Patricia trie are equal to N and $N - V + 1$, respectively. This does not hold for radix search tries. It can be proved that the average number of internal nodes is larger than N, namely asymptotically N/H, where H is the *entropy* of the alphabet.

In 1979, Fagin *et al.* [5] proposed extendible hashing as a fast access method for dynamic files. In the original version of this method, radix search trees have been used to access digital keys (records). In addition, another procedure was used to balance the tree in order to achieve good worst case performance. This restructuring generally changes the entire tree and is rather an expensive operation (compare binary search trees and AVL trees). So one may ask whether we need such a rebalancing procedure. To answer this question we must analyze a (random) shape of digital trees, and decide whether this shape resembles the shape of a complete tree [2] (the ultimate balanced tree). This problem led us to investigate the *depth* of a node (search time) and the (external or internal) *path length* in digital trees. The average depth of a node for digital trees has been studied in [6], [14], [22], [23], [24], the variance in [9], [22], [23], [24] and limiting distributions in [8], [18], [19]. The average value of the (external or internal) path length is closely related to the average depth of a node, but *not* the variance. The first attempt to compute the variance was reported in [9], however, it turned out that the variance of the depth was estimated, *not* the variance of the path length. This was rectified by Kirschenhofer, Prodinger and Szpankowski in [11], [12] who obtained the correct value for the variance in the symmetric regular tries and Patricia tries, respectively (for asymmetric extensions of these results see [8]). In this paper, we propose to evaluate the appropriate variance for the digital search trees, which was an open problem up to now. It has to be stressed that the variance of the internal path length in a digital search tree is the most difficult to estimate. This was already seen in the paper by Flajolet and Sedgewick [6] who establish an analytical methodology to analyze digital search trees (e.g., the average depth of a node). In our paper in the process of establishing the asymptotics of the internal path length we had to obtain some new analytic continuations of functions, which are mainly based on the famous Euler's product identities. As in [9] and [10], to derive the final results, namely to show the cancellation of the higher order asymptotics, we had to appeal to the theory of modular functions (cf. Section 3). In addition, this problem possesses nontrivial numerical challenge. A very preliminary version of this paper was presented at the IFIP Congress [13].

This paper is organized as follows. In the next section, we define our model, establish the general methodology to attack the problem and present our main results. In particular, we show that the variance of the internal path length for the *binary symmetric* digital search tree is $0.26600\ldots \cdot N + N\delta(\log_2 N)$ where N is the number of records and $\delta(x)$ is a periodic function with a very small amplitude. This implies that the internal path length converges *almost surely* to $N \log_2 N$. Finally, Section 3 contains proofs of our main results.

2. MAIN RESULTS

Let $\mathcal{D}_N$ be the family of digital search trees built from N records with keys from a random stream of bits. A key consists of 0's and 1's with equal probability of appearance. Let L_N denote the random variable "internal path length" of trees in $\mathcal{D}_N$ and $F_N(z)$ the corresponding probability generating functions, i.e., the coefficient $[z^k]F_N(z)$ of z^k in $F_N(z)$ is the probability that a tree in $\mathcal{D}_N$ has internal path length equal to k. Then the following recursion holds which is a direct consequence of the definition:

$$F_{N+1} = z^N \sum_{k=0}^{N} 2^{-N} \binom{N}{k} F_k(z) F_{N-k}(z), \qquad F_0(z) = 1. \tag{2.1}$$

The *expectation* l_N is given by $l_N = F_N'(1)$ and fulfills for $N \geq 0$

$$l_{N+1} = N + 2^{1-N} \sum_{k=0}^{N} \binom{N}{k} l_k, \qquad l_0 = 0 \tag{2.2}$$

This recursion may be solved explicitely by the use of *exponential generating functions*. With $L(z) = \sum_{N\geq 0} l_N \frac{z^N}{N!}$, (2.2) translates into the following functional differential equation

$$L'(z) = ze^z + 2e^{z/2} L(z/2).$$

By the substitution $\hat{L}(z) = e^z L(-z)$ we have the easier equation

$$\hat{L}(z) - \hat{L}'(z) = -z + 2\hat{L}(z/2).$$

With $\hat{L}(z) = \sum_{N\geq 0} \hat{l}_N \frac{z^N}{N!}$ we find for $N \geq 2$

$$\hat{l}_N = Q_{N-2}, \qquad \hat{l}_0 = \hat{l}_1 = 0$$

with the finite product

$$Q_N = \left(1-\frac{1}{2}\right)\left(1-\frac{1}{4}\right)\dots\left(1-\frac{1}{2^N}\right), \tag{2.3}$$

so that finally

$$l_N = \sum_{k=2}^{N} \binom{N}{k} (-1)^k Q_{k-2}. \tag{2.4}$$

The reader should note that an asymptotic evaluation of (2.4) is non elementary due to the fact that terms of almost equal magnitude occur with alternating signs. For this reason sophisticated methods from complex analysis are needed to find the correct order of growth. An essential step is the application of the following lemma from the calculus of finite differences.

Lemma 1. *(cf. [14, p.138], [17]). Let $\mathcal{C}$ be a path surrounding the points $j, j+1, \dots, N$ and $f(z)$ be analytic inside $\mathcal{C}$ and continuous on $\mathcal{C}$. Then*

$$\sum_{k\geq j} \binom{N}{k} (-1)^k f(k) = -\frac{1}{2\pi i}\int_{\mathcal{C}} [N;z] f(z)dz \tag{2.5}$$

with $[N;z] = \frac{(-1)^{N-1}N!}{z(z-1)\dots(z-N)}$.

In our application $f(z)$ is a meromorphic function that continues a sequence $f(k)$, e.g., $j=2$ and $f(k) = Q_{k-2}$ in (2.4). Moving the contour of integration, one can obtain the asymptotic expansion of the alternating sum by Cauchy's residue theorem, that is, for any real c (2.5) becomes $\sum_{k\geq j}\binom{N}{k}(-1)^k f(k) = \sum_{z_i\in\mathcal{P}_c} \mathrm{Res}\big([N;z_i]f(z_i)\big) + \mathcal{O}(N^c)$, where the sum is taken over the set of poles $\mathcal{P}_c$ different from $j, j+1, \dots, N$ with real part larger than c.

We note that the function $f(k) = Q_{k-2}$ possesses the analytic continuation $Q_z = Q_\infty / Q(2^{-z})$ where $Q(t) = \prod_{i\geq 1}(1 - t/2^i)$ [6]. Then, applying a refinement of the technique of Flajolet and Sedgewick, we can easily prove the following theorem (cf. Section 3).

Theorem 2. *The expectation l_N of the internal path length of digital search trees built from N records fulfills*

$$\begin{aligned} l_N &= N\log_2 N + N\Big[\frac{\gamma-1}{\log 2} + \frac{1}{2} - \alpha + \delta_1(\log_2 N)\Big] + \log_2 N \\ &\quad + \frac{2\gamma-1}{2\log 2} + \frac{5}{2} - \alpha + \delta_2(\log_2 N) + \mathcal{O}(\log N/N) \end{aligned} \tag{2.6}$$

with $\gamma = 0.57721\ldots$ *(Euler's constant) and* $\alpha = \sum_{n\geq 1} 1/(2^n-1) = 1.60669\ldots$, $\delta_1(x)$ *and* $\delta_2(x)$ *are continuous periodic functions of period* 1, *mean* 0 *and very small amplitude* ($< 10^{-6}$). *For later use we mention the Fourier expansion of* $\delta_1(x)$

$$\delta_1(x) = \frac{1}{\log 2}\sum_{k\neq 0}\Gamma\left(-1-\frac{2k\pi i}{\log 2}\right)e^{2k\pi i x}. \tag{2.7}$$

where $\Gamma(x)$ *is the gamma function* [1].

We mention in passing that the $\mathcal{O}(1)$-term in (2.6) is slightly incorrect in [14].

Now we turn to the *analysis of the variance* which is given by $\mathrm{Var}L_N = s_N + l_N - l_N^2$ with $s_N = F_N''(1)$. From (2.1) we get the recurrence relation (for $N \geq 0$; $s_0 = 0$)

$$\begin{aligned} s_{N+1} &= N2^{2-N}\sum_{k=0}^{N}\binom{N}{k}l_k + N(N-1) \\ &\quad + 2^{1-N}\sum_{k=0}^{N}\binom{N}{k}l_k l_{N-k} + 2^{1-N}\sum_{k=0}^{N}\binom{N}{k}s_k. \end{aligned} \tag{2.8}$$

In order to find an explicit solution to this recurrence, we split it into 3 parts: $s_N = u_N + v_N + w_N$, where

$$u_{N+1} = 2N(l_{N+1}-N) + 2^{1-N}\sum_{k=0}^{N}\binom{N}{k}u_k, \quad N\geq 0, \quad u_0 = 0, \tag{2.9a}$$

$$v_{N+1} = N(N-1) + 2^{1-N}\sum_{k=0}^{N}\binom{N}{k}v_k, \quad N\geq 0, \quad v_0 = 0, \tag{2.9b}$$

$$w_{N+1} = 2^{1-N}\sum_{k=0}^{N}\binom{N}{k}l_k l_{N-k} + 2^{1-N}\sum_{k=0}^{N}\binom{N}{k}w_k, \quad N\geq 0, \quad w_0 = 0. \tag{2.9c}$$

All of the above recurrences, as well as the one for the average internal path length (2.2), fall into the following general recurrence studied in [23]. Let (x_n) be a sequence of numbers satisfying the following

$$x_{n+1} = a_{n+1} + 2^{1-n}\sum_{k=0}^{n}\binom{n}{k}x_k, \qquad n\geq 2, \tag{2.10}$$

where (a_n) is any sequence of numbers. The solution of (2.10) depends on the so called *binomial inverse relations* that are defined as follows

$$\hat{a}_n = \sum_{k=0}^{n}(-1)^k\binom{n}{k}a_k \quad \text{and} \quad a_n = \sum_{k=0}^{n}(-1)^k\binom{n}{k}\hat{a}_k.$$

The second equation justifies the name binomial inverse relations. For more details, see Riordan [20]. A similar treatment as in the case of (2.2) leads to the following explicit solution (for details see [23]).

Lemma 3. *Let $x_0 = x_1 = 0$. Then the recurrence (2.10) possesses the following solution*

$$x_n = x_0 + n(x_1 - x_0) - \sum_{k=2}^{n} (-1)^k \binom{n}{k} \hat{x}_{k-2} \tag{2.11a}$$

where

$$\hat{x}_n = Q_n \sum_{i=1}^{n+1} \big[\hat{a}_i - \hat{a}_{i+1} - a_1\big] / Q_{i-1} \tag{2.11b}$$

and Q_n is defined in (2.3).

Using Lemma 3 we immediately solve our recurrences (2.9a) to (2.9c). In particular, one proves

$$\hat{u}_k = 2Q_{k-2}\left\{4 + \sum_{j=1}^{k-2} \frac{1}{2^j - 1} - \sum_{j=1}^{k-2} \frac{j}{2^j - 1} - \frac{2k}{2^{k-2} - 1}\right\},$$

$$\text{for } k \geq 3, \qquad \hat{u}_0 = \hat{u}_1 = \hat{u}_2 = 0 \tag{2.12a}$$

$$\hat{v}_k = -4Q_{k-2}, \qquad \text{for } k \geq 3, \qquad \hat{v}_0 = \hat{v}_1 = \hat{v}_2 = 0 \tag{2.12b}$$

and

$$\hat{w}_k = Q_{k-2} \sum_{j=4}^{k-1} \frac{2^{1-j}}{Q_{j-1}} \sum_{i=2}^{j-2} \binom{j}{i} Q_{j-2} Q_{j-i-2},$$

$$\text{for } k \geq 5, \qquad \hat{w}_0 = \cdots = \hat{w}_4 = 0 \tag{2.12c}$$

Of course, the "unhatted" solutions u_N, v_N and w_N follow from the binomial relations, as shown in (2.11a). It is also worth to mention that the recurrence for v_N is easy, and after simple algebra one proves

$$v_N = 4\binom{N}{2} - 4l_N, \tag{2.12d}$$

so that the treatment of u_N and w_N remains to be done.

In principle u_N and w_N may be analyzed by making use of Lemma 1 and 3. However, it turns out to be a highly non trivial problem to find an analytical continuation of $\hat{w}_k$. After lengthy and difficult computations the residue calculus leads us to the following main result of this paper, which is proved in the next section.

Theorem 4. *The variance of the internal path length of digital search trees built from N records becomes*

$$\mathrm{Var}\, L_N = N \cdot \big\{C + \delta(\log_2 N)\big\} + \mathcal{O}(\log^2 N / N)$$

where C is a constant that can be expressed as

$$
\begin{aligned}
C = & -\frac{28}{3L} - \frac{39}{4} - 2\beta_1 + \frac{2\alpha}{L} + \frac{\pi^2}{2L^2} + \frac{2}{L^2} \\
& - \frac{2}{L}\sum_{k\geq 3} \frac{(-1)^{k+1}(k-5)}{(k+1)k(k-1)(2^k-1)} \\
& + \frac{2}{L}\sum_{r\geq 1} b_{r+1}\left(\frac{L(1-2^{-r+1})/2-1}{1-2^{-r}} - \sum_{k\geq 2}\frac{(-1)^{k+1}}{k(k-1)(2^{r+k}-1)}\right) \\
& + \frac{2}{L}\hat{w}'(3) - 2[\delta_1\delta_2]_0 - [\delta_1^2]_0
\end{aligned}
\tag{2.13}
$$

with $L = \log 2$, $\alpha = \sum_{n\geq 1} \frac{1}{2^n-1}$, $\beta_1 = \sum_{j\geq 1}\frac{j2^j}{(2^j-1)^2}$, $b_{r+1} = (-1)^r 2^{-\binom{r+1}{2}}$. *The fluctuating function* $\delta(x)$ *is continuous with period 1, mean zero and* $|\delta(x)| \leq 10^{-6}$, *and* $|[\delta_1^2]_0| \leq 10^{-10}$, $|[\delta_1\delta_2]_0| \leq 10^{-10}$. *Finally,* $\hat{w}(z)$ *is a function defined as*

$$
\begin{aligned}
\hat{w}(z+1)/Q_{z-1} = & -2Q_\infty z + \frac{\xi(z+2)}{2^z Q_z} + \frac{\xi(z+3)}{2^{z+1}Q_{z+1}} \\
& + \sum_{j\geq 2}\left(\frac{\xi(z+j+2)}{2^{z+j}Q_{z+j}} - \frac{\xi(j+2)}{2^j Q_j}\right)
\end{aligned}
\tag{2.14}
$$

with $Q_z = Q_\infty/Q(2^{-z})$, *where* $Q(t) = \prod_{i\geq 1}(1-t/2^i)$, $Q_\infty = Q(1)$, *and*

$$
\begin{aligned}
\xi(z+1) = \sum_{r\geq 0}\frac{b_{r+1}}{Q_r}\cdot\frac{Q_\infty}{Q(2^{3-z-r})}\cdot & \left\{2^z - \frac{2}{1-2^{1-z-r}} - \frac{2z}{1-2^{2-z-r}}\right. \\
& \left. + 2\sum_{k\geq 2}\binom{z}{k}\frac{1}{2^{r+k-1}-1}\right\}.
\end{aligned}
\tag{2.15}
$$

Numerical evaluation of the constant C *reveals that* $C = 0.26600\ldots$ *and all five digits after the decimal point are significant. We should point out that in order to achieve the same accuracy in* C *one needs to run the recurrence equations (2.9a)–(2.9c) for* $N \sim 10^6$.

In the following lemma we present an explicit formula for $\hat{w}'(3)$ that is convenient for numerical evaluations.

Lemma 5. *The following identity holds*

$$
\begin{aligned}
\frac{2\hat{w}'(3)}{L} &= -\frac{2Q_\infty}{L} + \sum_{j\ge 2} \frac{1}{2^j Q_j} \sum_{r\ge 0} a_{r+1} Q_{r+j-2} \cdot \\
&\cdot \Bigg\{ -\sum_{n\ge 1} \frac{1}{2^{n+r+j+2}-1} \cdot \Bigg(2^{j+1} - 2j - 4 \\
&\qquad + 2\sum_{k=2}^{j-1} \binom{j+1}{k} \frac{1}{2^{r+k-1}-1} \Bigg) \\
&\quad + \frac{2}{(1-2^{-j-r})^2} + \frac{2j+2}{(1-2^{1-j-r})^2} - \frac{2}{L}\frac{1}{1-2^{1-j-r}} \\
&\quad - 2\sum_{k=2}^{j+1} \binom{j+1}{k} \frac{1}{2^{r+k-1}-1} + \frac{2}{L}\sum_{k=1}^{j+1} \binom{j+1}{k} \frac{1}{2^{r+k}-1} \\
&\quad + \frac{2}{L}\sum_{k=0}^{j+1} \binom{j+1}{k} \sum_{i\ge 1} \frac{(-1)^i}{(i+1)(2^{r+k+i}-1)} \Bigg\} \\
&\quad + \sum_{j\ge 3} \frac{\xi(j+2)}{2^j Q_j} \sum_{k\ge j+1} \frac{1}{2^k - 1}
\end{aligned}
\tag{2.16}
$$

where

$$a_{r+1} = \frac{b_{r+1}}{Q_r} = (-1)^r 2^{-\binom{r+1}{2}}/Q_r, \tag{2.17}$$

and

$$\xi(j+2) = \sum_{k=2}^{j-1} \binom{j+1}{k} Q_{k-2} Q_{j-k-1} \tag{2.18}$$

with Q_k defined in Theorem 4.

Before we proceed to the proof of our results, we first offer some remarks and extensions.

Remark 2.19. *The covariance analysis.* Theorem 4 and our previous result [9] and [23] provide asymptotics for the covariance between two different nodes in a digital search tree (DST). Let D_N be a depth of a (randomly selected) node and let $D_N^{(i)}$ be the length of a path from the root to the i-th node. Note that the internal path length L_N is defined in terms of $D_N^{(i)}$ as $L_N = \sum_{i=1}^N D_N^{(i)}$. Then

$$\mathrm{Var}\, L_N = E\left\{ \left[\sum_{i=1}^N D_N^{(i)} \right]^2 \right\} - \left\{ E \sum_{i=1}^N D_N^{(i)} \right\}^2$$

and this implies

$$\mathrm{Var}\, L_N = N \mathrm{Var}\, D_N + 2 \sum_{i\ne j} \mathrm{Cov}\left\{ D_N^{(i)}, D_N^{(j)} \right\}.$$

The variance of the depth $\mathrm{Var}\, D_N$ for the symmetric DST was analyzed in [9], and for the asymmetric one in [23]. In particular, it was proved that for the binary symmetric Patricia $\mathrm{Var}\, D_N = 2.844\cdots$. Using Theorem 4 and the above we find

$$2\sum_{i\neq j} \mathrm{Cov}\left\{D_N^{(i)}, D_N^{(j)}\right\} = -2.67\ldots \cdot N.$$

This also implies that the average value of

$$\mathrm{Cov}\{D_N^{(i)}, D_N^{(j)}\} \text{ is } \sim -2.67\ldots/N.$$

Note that the equivalent quantity for regular tries is approximately equal to $+0.84\ldots/N$ [11] and for Patricia $= -.63\cdots/N$ [12].

Remark 2.20 *The path length L_N converges almost surely* to E_N! Applying Theorem 4 it is not difficult to prove that L_N/EL_N tends to one *almost surely* (i.e., with probability one) as $N \to \infty$. Indeed, by Chebyshev's inequality one obtains

$$\Pr\{|L_N/EL_N - 1| \geq \epsilon\} \leq \frac{\mathrm{Var}\ L_N}{\epsilon^2(EL_N)^2}.$$

But, by Theorem 4

$$\Pr\{|L_N/EL_N - 1| \geq \epsilon\} \leq \frac{A}{\epsilon^2 N \log_2^2 N} \to 0.$$

This shows that $L_N/EL_N \to 1$ *in probability* as $N \to \infty$ [21]. To prove a stronger result, namely, that $L_N/EL_N \to 1$ *with probability one* (i.e., almost surely) we apply the above and the Borel-Cantelli lemma [21], and then show

$$\sum_{N=1}^{\infty} \Pr\{|L_N/EL_N - 1| \geq \epsilon\} \leq \frac{0.266\cdots}{\epsilon^2} \sum_{N=1}^{\infty} \frac{1}{N \log_2^2 N} < \infty,$$

so, by the Borel-Cantelli lemma $L_N \sim EL_N \sim N \log_2 N$ with probability one.

Remark 2.21. *Comparison of digital trees.* In order to select the best digital tree one needs to compare different characteristics of digital trees, namely regular tries, Patricia tries and Digital Search Trees (DST). The table below contains four important parameters that are often used to predict a random shape of these trees (cf. [6], [9], [11], [12], [14], [15], [22], [23], [24]).

	EL_N	$\mathrm{Var}\, L_N$	$\mathrm{Var}\, D_N$	$\mathrm{Cov}(D_N^i, D_N^j)$
DST	$N(\log_2 N - 1.71)$	$N \cdot 0.26$	2.844	$-\ 2.67/N$
TRIES	$N(\log_2 N + 1.33)$	$N \cdot 4.35$	3.507	$+\ 0.84/N$
PATRICIA	$N(\log_2 N + 0.33)$	$N \cdot 0.37$	1.000	$-\ 0.63/N$

It can be seen from the table that the average external (internal) path length is approximately the same for all three digital trees. However, the variance of the depths and internal (external) path lengths differ significantly. We also notice that the variance of the internal for DST is smaller than the variance of the external path length for Patricia, but the reverse holds true for the variance of the depth. Therefore, in order to answer the question which digital tree is the best (balanced) one needs to decide which parameter (depth or path length) carry more useful information. This is discussed below.

Remark 2.22. *Which digital tree is balanced the best?* A complete binary tree [2] is the ultimately best balanced tree. Therefore, any tree with a good balance property should have the average depth (external path length) equal to $\log_2 N + \mathcal{O}(1)$ $\big($resp. $N\log_2 N + \mathcal{O}(N)\big)$, and a small variance. Such a property is highly desired since then one can expect that a typical search time for a key is approximately equal to the average depth. This was already said in Remark 2.17, nevertheless we might be interested in a relative comparison between different trees that satisfy the above vague criterion. One may ask which digital tree is balanced the best. To solve this problem we must determine which parameter from the table above is the most suited for such an analysis. We first notice that the covariance $\mathrm{Cov}\left\{D_N^{(i)}, D_N^{(j)}\right\}$ for Patricia and digital search tree is *negatively correlated*. This means, that $D_N^{(i)} < ED_N$ and $D_N^{(i)} > ED_N$ also tend to occur together. Thus, for negatively correlated random variables $D_N^{(i)}$ and $D_N^{(j)}$, if one is large, the other is likely to be small. This indicates a good balance property for a tree. Note, that in the regular tries $\mathrm{Cov}\left\{D_N^{(i)}, D_N^{(j)}\right\} \sim 0.84/N > 0$ and $D_N^{(i)}$ and $D_N^{(j)}$ in that case are *positively* correlated. This means that $D_N^{(i)}$ is large, then $D_N^{(j)}$ is likely to be large, too.

So finally the problem under consideration boils down to a choice between the variance of the depth or the variance of the internal (external) path length. In [12] we have argued that for tries (regular or Patricia) the external path length is a better measure of the balance property for a tree. This argument can be made even more convincing for digital search trees. Consider a DST that is completely balanced. Then, since keys are stored in internal nodes the variance of the depth is *positive* no matter how balanced the tree is. But, the variance of the internal path length is *zero* for such an instance. This leads to an obvious conclusion that the variance of the path length should be considered as a criterion for the balance property. Then, one finds out from the table above that digital search trees are the best balanced digital trees.

3. ANALYSIS

As we have already pointed out in Section 2, it is a nontrivial problem to find appropriate analytic continuations for the sequences of values $f(k)$ that occur in alternating sums (2.5). In order to illustrate our approach, we start with the easiest case, namely the evaluation of the expectation l_N. From (2.4) we know

$$l_N = \sum_{k=2}^{N} \binom{N}{k} (-1)^k Q_{k-2}.$$

As in [6] we may rewrite $f(k) = Q_{k-2}$ as

$$Q_{k-2} = \frac{Q_\infty}{Q(2^{2-k})},$$

where

$$Q(x) = \prod_{i \geq 1} \Big(1 - \frac{x}{2^i}\Big) \qquad \text{and} \qquad Q_\infty = Q(1). \tag{3.1}$$

Therefore we have the analytic continuation

$$f(z) = \frac{Q_\infty}{Q(2^{2-z})}. \tag{3.2}$$

The main contribution to l_N is given by $\operatorname{Res}([N;z]f(z); z = 1)$. We have with $u = z - 1 \to 0$

$$[N;z] \sim \frac{N}{u}(1 + u(H_{N-1} - 1))$$

and

$$\frac{Q_\infty}{Q(2^{2-z})} \sim \frac{1}{Lu}\Big(1 + Lu(\frac{1}{2} - \alpha)\Big)$$

(remember $L = \log 2$), since

$$\alpha = \left(\frac{Q_\infty}{Q(x)}\right)'\Bigg|_{x=1}. \tag{3.3}$$

Therefore

$$\operatorname{Res}([N;z]f(z); z = 1) = \frac{N}{L}\Big(H_{N-1} - 1 + L(\frac{1}{2} - \alpha)\Big).$$

Using the well known asymptotics for the *harmonic numbers* H_{N-1} we get the contribution (from $z = 1$)

$$N \log_2 N + N\left(\frac{\gamma}{L} - \frac{1}{L} + \frac{1}{2} - \alpha\right) - \frac{1}{2L} + \mathcal{O}\left(\frac{1}{N}\right). \tag{3.4}$$

Besides $z = 1$ we have with the same real part 1 the simple poles $z_k = 1 + \frac{2k\pi i}{L}$, $k \in \mathbf{Z}$, $k \neq 0$ with

$$\operatorname{Res}\big([N;z]f(z); z = z_k\big) = [N;z_k] \cdot \frac{1}{L},$$

so that we get the contribution

$$\frac{N^{z_k}}{L}\Gamma(-z_k) - \frac{(z_k - 1)z_k N^{z_k-1}\Gamma(-z_k)}{2L} + \mathcal{O}(N^{z_k-2}). \tag{3.5}$$

The reader should take notice of the fact that the first term in (3.5) gives the Fourier coefficients of $\delta_1(x)$ in Theorem 1.

The next relevant pole is $z=0$ and yields a contribution of

$$\log_2 N + \frac{\gamma}{L} + \frac{5}{2} - \alpha. \tag{3.6}$$

The poles $z = z_k - 1$ yield a periodic contribution of order N^0 and so on.

Collecting all contributions gives the expansion (2.6) in Theorem 2.

Next we focus our attention on the *asymptotics of* u_N. In order to find an appropriate analytic continuation of $\hat{u}_k$ we rewrite the sums appearing in (2.12a) as follows:

$$\sum_{j=1}^{k-2} \frac{1}{2^j-1} = \alpha - \sum_{j\ge 1} \frac{1}{2^{k-2+j}-1},$$

$$\sum_{j=1}^{k-2} \frac{j}{2^j-1} = \sum_{j\ge 1} \frac{j}{2^j-1} - \sum_{j\ge 1} \frac{k-2+j}{2^{k-2+j}-1}.$$

Thus we may continue $\hat{u}_k$ via the function

$$\begin{aligned}\hat{u}(z) = \frac{2Q_\infty}{Q(2^{2-z})}\Bigg[4+\alpha-\sum_{j\ge 1}\frac{1}{2^{z-2+j}-1}\\ -\sum_{j\ge 1}\frac{j}{2^j-1}+\sum_{j\ge 1}\frac{z-2+j}{2^{z-2+j}-1}-\frac{2z}{2^{z-2}-1}\Bigg].\end{aligned} \tag{3.7}$$

Now the main contribution to u_N in Lemma 1 originates from a second order pole of $[N;z]\hat{u}(z)$ in $z=2$. Further contributions that are necessary for the evaluation of the variance come from first order poles in $z = 2 + \frac{2k\pi i}{L}$, $k \neq 0$, a third order pole in $z=1$ as well as second order poles in $z = 1 + \frac{2k\pi i}{L}$, $k\neq 0$. Collecting all the above mentioned contributions we end up with the following expansion of u_N ($\delta_i(x)$ stands in all following formulas for a continuous periodic function of period 1 and mean zero).

Lemma 6.

$$\begin{aligned} u_N \sim\ & 4N^2\log_2 N \\ &+ N^2\left(\frac{4(\gamma-1)}{L} - 6 - 4\alpha + \delta_3(\log_2 N)\right) \\ &- N\log_2^2 N + 2N\log_2 N\cdot\left(\frac{2-\gamma}{L}+8+\alpha+\delta_4(\log_2 N)\right) \\ &+ N\Bigg(-\frac{\gamma^2}{L^2}+\frac{4\gamma}{L^2}+\frac{12\gamma}{L}+\frac{2\alpha\gamma}{L}-\frac{\pi^2}{6L^2}-\frac{4}{L^2}-\frac{10}{L}-\frac{2\alpha}{L} \\ &- \alpha^2+\beta-11\alpha-2\beta_1+\frac{133}{6}+\delta_5(\log_2 N)\Bigg) + \mathcal{O}(\log^2 N)\end{aligned}$$

with $\beta = \sum_{k\ge1} \frac{1}{(2^k-1)^2}$, L, α, β_1 *as in Theorem 4.*

As already mentioned in Section 2,

$$v_N = 4\binom{N}{2} - 4l_N,$$

so that the *asymptotics of* v_N are given by

$$\begin{aligned} v_N &\sim 2N^2 - 4N\log_2 N \\ &\quad + 4N\left(-1 + \alpha - \frac{\gamma}{L} + \frac{1}{L} - \delta_1(\log_2 N)\right) + \mathcal{O}(\log N). \end{aligned} \tag{3.8}$$

The most challenging task is to find an appropriate *analytic continuation of* $\hat{w}(z)$.

From (2.12c) we have

$$\hat{w}_{k+1} = -Q_{k-1}\sum_{j=4}^{k}\frac{\xi(j+1)}{2^{j-1}Q_{j-1}} \tag{3.9}$$

with

$$\xi(j+1) = \sum_{i=2}^{j-2}\binom{j}{i}Q_{i-2}Q_{j-2-i}. \tag{3.10}$$

For the following the reader should note that $\xi(j+1) \sim Q_\infty^2 2^j$. We start by rewriting (3.9) in the following manner:

With $\eta(j+1) = \xi(j+1) - Q_\infty^2 2^j$ we have

$$\begin{aligned} -\frac{\hat{w}_{k+1}}{Q_{k-1}} &= \sum_{j=4}^{k}\frac{\eta(j+1) + Q_\infty^2 2^j}{2^{j-1}Q_{j-1}} \\ &= \sum_{j\ge4}\frac{\eta(j+1)}{2^{j-1}Q_{j-1}} - \sum_{j\ge k+1}\frac{\eta(j+1)}{2^{j-1}Q_{j-1}} \\ &\quad + 2Q_\infty^2\left(\sum_{j\ge4}\left(\frac{1}{Q_{j-1}} - \frac{1}{Q_\infty}\right) - \sum_{j\ge k+1}\left(\frac{1}{Q_{j-1}} - \frac{1}{Q_\infty}\right)\right) \\ &\quad + 2Q_\infty(k-3). \end{aligned} \tag{3.11}$$

Therefore

$$\begin{aligned} \hat{w}_{k+1} = Q_{k-1}\Bigg[&-2Q_\infty(k-3) + \sum_{j\ge0}\frac{\eta(j+k+2)}{2^{j+k}Q_{j+k}} - \sum_{j\ge3}\frac{\eta(j+2)}{2^jQ_j} \\ &+ 2Q_\infty^2\left(\sum_{j\ge0}\left(\frac{1}{Q_{j+k}} - \frac{1}{Q_\infty}\right) - \sum_{j\ge3}\left(\frac{1}{Q_j} - \frac{1}{Q_\infty}\right)\right)\Bigg]. \end{aligned}$$

Since all involved series are now absolutely convergent, we may add them term by term and get

$$\hat{w}_{k+1} = Q_{k-1}\Bigg[-2Q_\infty k + \frac{\xi(k+2)}{2^k Q_k} + \frac{\xi(k+3)}{2^{k+1}Q_{k+1}} + \sum_{j\ge 2}\left(\frac{\xi(k+j+2)}{2^{k+j}Q_{k+j}} - \frac{\xi(j+2)}{2^j Q_j}\right)\Bigg].$$

From this, the representation for $\hat{w}(z+1)$ as in (2.14) is immediate, provided we have an appropriate interpretation for $\xi(z+1)$. This will be our next goal. The following well-known partition identities of Euler are our basic tool:

$$\frac{1}{Q(t)} = \prod_{n\ge 1}\frac{1}{(1-t2^{-n})} = \sum_{n\ge 0}\frac{t^n}{2^n Q_n} \tag{3.12}$$

and

$$Q(t) = \prod_{n\ge 1}\left(1-\frac{t}{2^n}\right) = \sum_{n\ge 0} a_{n+1}t^n \tag{3.13}$$

with

$$a_{n+1} = (-1)^n 2^{-\binom{n+1}{2}}/Q_n.$$

Using (3.2) and (3.12) we have

$$\begin{aligned}\xi(N+1) &= \sum_{k=2}^{N}\binom{N}{k}\frac{Q_\infty}{Q(2^{2-k})}\frac{Q_\infty}{Q(2^{2+k-N})}\\ &= Q_\infty^2\sum_{i,j\ge 0}\frac{2^{i+j}}{Q_iQ_j}\sum_{k=2}^{N-2}\binom{N}{k}(2^{-i})^k(2^{-j})^{N-k},\end{aligned}$$

where the innermost sum is now

$$(2^{-i}+2^{-j})^N - 2^{-iN} - 2^{-jN} - N2^{-i(N-1)-j} - N2^{-i-j(N-1)}.$$

The last expression for $\xi(N+1)$ is symmetric in i and j. However, it turns out that for the purpose of finding an analytic continuation $\sum_{i,j\ge 0}$ should be rewritten as $-\sum_{j=i} + 2\sum_{j\ge i\ge 0}$. Writing $j=i+h$ in the second sum we get

$$\begin{aligned}\xi(N+1) = &-(2^N-2-2N)\sum_{i\ge 0}\frac{Q_\infty^2}{Q_i^2}2^{i(2-N)}\\ &+2\sum_{i,h\ge 0}\frac{Q_\infty^2}{Q_iQ_{i+h}}2^{i(2-N)}2^h\big[(1+2^{-h})^N - 1 - N2^{-h}\big]\\ &-2\sum_{i,h\ge 0}\frac{Q_\infty^2}{Q_iQ_{i+h}}2^{i(2-N)}2^{h(1-N)}\\ &-2N\sum_{i,h\ge 0}\frac{Q_\infty^2}{Q_iQ_{i+h}}2^{i(2-N)}2^{h(2-N)}\end{aligned} \tag{3.14}$$

In expression (3.14) N can be replaced by z, yielding a meromorphic function, since all series converge uniformly. However, we are able to simplify $\xi(z+1)$ in the following way. Consider for example the last term in (3.14):

$$2z \sum_{i,h \geq 0} \frac{Q_\infty^2}{Q_i Q_{i+h}} 2^{(i+h)(2-z)} = 2z \sum_{i,h \geq 0} \frac{Q_\infty Q(2^{-i-h})}{Q_i} 2^{(i+h)(2-z)}$$

which is by Euler's identity (3.13)

$$= 2z \sum_{r \geq 0} a_{r+1} \sum_{i \geq 0} \frac{Q_\infty}{Q_i} (2^{-r+2-z})^i \sum_{h \geq 0} (2^{-r+2-z})^h$$

and by Euler's identity (3.12)

$$= 2z \sum_{r \geq 0} a_{r+1} \frac{Q_\infty}{Q(2^{3-z-r})} \cdot \frac{1}{1 - 2^{2-z-r}}.$$

Rewriting the other terms from (3.14) in a similar way, especially using

$$(1 + 2^{-h})^z - 1 - z 2^{-h} = \sum_{k \geq 2} \binom{z}{k} 2^{-hk}$$

for the second term, we finally get (2.15).

Our next task is to investigate the poles of $[N;z]\hat{w}(z)$ different from $z = 5, 6, \ldots, N$. ∎

From (3.11) we see that $\hat{w}(4) = \hat{w}(3) = 0$ (observe that $\xi(4) = 0$), so that the first poles occur with real part 2. In order to determine the residues of $[N;z]\hat{w}(z)$ in $z = 2$ resp. $z = 1$ we need the local behaviour of $\hat{w}(z)$. Because of (2.14) this behaviour will depend on the behaviour of $\sigma(z) := \xi(z)/2^{z-2} Q_{z-2}$ near $z = 2, 3, \ldots$. From (2.15) we see that $\xi(z)$ has second order poles for $z = 2$ and $z = 3$ and is analytic for $z = 4, 5, \cdots$. Since $[N;z]Q_{z-2}$ has already a second order pole for $z = 1$, it will be necessary to expand $\sigma(z)$ near $z = 2, 3, \ldots$ up to the linear terms. In particular, the reader should note that *all* the derivatives $\sigma'(z)$ for $z = 4, 5, \ldots$ will occur in $\mathrm{Res}\big([N;z]\hat{w}(z); z = 1\big)$. This is the main reason that the constant C in the final result is rather a complicated one.

We start with the expansion of $\sigma(z)$ about $z = 3$. Let $u = z - 3$; then we find from (2.15) after laborious computations:

$$\begin{aligned}
\sigma(3+u) &\sim -\frac{4}{L^2u^2} + \frac{1}{u}\left(\frac{6}{L} - \frac{2}{L^2}\right) \\
&+ u^0\left(\frac{16}{3} + \frac{5}{L} + \frac{2\beta_2}{L} + 2\sum_{r\ge 2}\frac{b_{r+1}}{1-2^{-r}}\left(1 - \frac{2}{1-2^{-r}}\right)\right) \\
&+ u^1\Bigg(-\frac{223}{18}L - \frac{23}{6} - 3C_1 + \frac{C_2}{L} - 2C_3 \\
&+ 2\sum_{r\ge 2}\frac{b_{r+1}}{1-2^{-r}}\Bigg\{1 - 3L + L\sum_{i=2}^{r-1}\frac{1}{2^i-1} + \frac{2^{-1-r}L}{(1-2^{-1-r})^2} \\
&+ \frac{2L-1}{1-2^{-r}} - \frac{2L}{1-2^{-r}}\sum_{i=2}^{r-1}\frac{1}{2^i-1} + \frac{2^{1-r}L}{(1-2^{-r})^2} + D_{r,1}\Bigg\}\Bigg)
\end{aligned} \tag{3.15}$$

where

$$\beta_2 = 2\sum_{k\ge 2}\frac{(-1)^k}{(k+1)k(k-1)(2^k-1)},$$

$$C_1 = \sum_{h\ge 0} 2^h\left[(1+2^{-h})^2\log(1+2^{-h}) - 2^{-h}\right],$$

$$C_2 = \sum_{h\ge 0} 2^h(1+2^{-h})^2\log^2(1+2^{-h}),$$

$$C_3 = \sum_{h\ge 0}\left[(1+2^{-h})^2\log(1+2^{-h}) - 2^{-h}\right],$$

$$D_{r,1} = \sum_{h\ge 0} 2^{h(1-r)}\left[(1+2^{-h})^2\log(1+2^{-h}) - 2^{-h}\right].$$

The constant in the u^0-term can be simplified according to the remarkable identity

$$\sum_{r\ge 1}\frac{b_{r+1}}{1-2^{-r}}\left(1 - \frac{2}{1-2^{-r}}\right) = \sum_{j\ge 1}\frac{2^j}{(2^j-1)^2} = \alpha + \beta. \tag{3.16}$$

For the proof of (3.16) we observe that the left-hand side equals

$$\sum_{r\ge 1} a_{r+1}Q_{r-1} - 2\sum_{r\ge 1} a_{r+1}\frac{Q_{r-1}}{1-2^{-r}}.$$

Using (3.12) and (3.13) this expression becomes

$$\sum_{i\ge 0} Q(2^{-i})\left(Q(2^{-i}) - 1\right) - 2\sum_{i,k\ge 0} Q(2^{-i})\left(Q(2^{-i-k}) - 1\right)$$

$$= -\Bigl(\sum_{i\ge 0} \left(Q(2^{-i})-1\right)\Bigr)^2 - \sum_{j\ge 0}(2j+1)\left(Q(2^{-j})-1\right)$$

$$= -E_1^2 - 2E_2 - E_1,$$

$$\text{where}\quad E_1 = \sum_{r\ge 1} \frac{a_{r+1}}{1-2^{-r}} \quad\text{and}\quad E_2 = \sum_{r\ge 1} a_{r+1}\frac{2^{-r}}{(1-2^{-r})^2}.$$

Now we observe

$$E_1 = \lim_{t\to 2}\left[\frac{Q_\infty}{Q(t)} - \frac{1}{1-t/2}\right] = -\left(\frac{Q_\infty}{Q(t)}\right)'\bigg|_{t=1} = -\alpha$$

and

$$E_2 = 2\lim_{t\to 2}\left[\left(\frac{Q_\infty}{Q(t)}\right)' - \frac{1/2}{(1-t/2)^2}\right]$$
$$= -\frac{1}{2}\left(\frac{Q_\infty}{Q(t)}\right)''\bigg|_{t=1} = -\frac{\alpha^2+\beta}{2}$$

and get the right-hand side of (3.16).

The expansion of $\sigma(z)$ about $z=2$ reads with $u=z-2$:

$$\begin{aligned}
\sigma(2+u) \sim{}& \frac{4}{L^2u^2} + \frac{1}{u}\left(\frac{2}{L}+\frac{2}{L^2}\right)\\
&+ u^0\left(-\frac{16}{3} - \frac{6}{L} - \frac{2}{L}(C_4+C_6) + 2E_3\right)\\
&+ u^1\Biggl(\frac{L}{6} - \frac{55}{6} - 3C_4 - \frac{C_5}{L} + C_6 - \frac{C_7}{L}\\
&\quad + 2\sum_{r\ge 2}\frac{b_{r+1}}{(1-2^{1-r})(1-2^{-r})}\Biggl\{D_{r,2} + 1 - 2L\\
&\qquad + L\sum_{i=1}^{r-2}\frac{1}{2^i-1} + \frac{L-1}{1-2^{1-r}}\\
&\qquad + L\Biggl(\frac{1}{(1-2^{-r})^2} - \frac{1}{1-2^{-r}}\sum_{i=1}^{r-2}\frac{1}{2^i-1}\\
&\qquad + \frac{2^{1-r}}{(1-2^{1-r})^2} - \frac{1}{1-2^{1-r}}\sum_{i=1}^{r-2}\frac{1}{2^i-1}\Biggr)\Biggr\}\Biggr)
\end{aligned}\tag{3.17}$$

where

$$C_4 = \sum_{h\ge 0} 2^h\left[(1+2^{-h})\log(1+2^{-h}) - 2^{-h}\right],$$

$$C_5 = \sum_{h\ge 0} 2^h(1+2^{-h})\log^2(1+2^{-h}),$$

$$C_6 = \sum_{h\geq 0} \left[(1+2^{-h}) \log(1+2^{-h}) - 2^{-h} \right],$$

$$C_7 = \sum_{h\geq 0} (1+2^{-h}) \log^2(1+2^{-h}),$$

$$D_{r,2} = \sum_{h\geq 0} 2^{h(1-r)} \left[(1+2^{-h}) \log(1+2^{-h}) - 2^{-h} \right],$$

$$E_3 = \sum_{r\geq 2} \frac{b_{r+1}}{(1-2^{1-r})(1-2^{-r})} \left(1 - \frac{1}{1-2^{-r}} - \frac{1}{1-2^{1-r}} \right).$$

For later simplifications we note that

$$-\frac{2}{L}\left(C_4 + C_6\right) = -8 + \frac{3}{L} - \frac{2\beta_2}{L} \tag{3.18}$$

and

$$E_3 = 2 - \alpha - \beta, \tag{3.19}$$

where (3.18) follows from the expansion of the logarithm and (3.19) by partial fraction decomposition and rearrangements of the sums.

We finally note that

$$C_5 + C_7 = C_2. \tag{3.20}$$

Next we discuss $\sigma(z)$ for z close to $j = 4, 5, \cdots$.

$$\begin{aligned} \sigma(j+u) &\sim \sigma(j) \\ &+ \frac{u}{2^{j-2}Q_{j-2}} \left(\xi'(j) - L\xi(j) + L\xi(j) \sum_{k\geq 1} \frac{1}{2^{k+j-2}-1} \right). \end{aligned} \tag{3.21}$$

From (3.21)

$$\begin{aligned} &\sum_{j\geq 4} \Big(\sigma(j+u) - \sigma(j) \Big) \\ &\quad \sim u \Bigg(\sum_{j\geq 4} \frac{\xi'(j) - L\xi(j)}{2^{j-2}Q_{j-2}} \\ &\qquad L \sum_{j\geq 4} \frac{\xi(j)}{2^{j-2}Q_{j-2}} \sum_{k\geq 1} \frac{1}{2^{k+j-2}-1} \Bigg). \end{aligned} \tag{3.22a}$$

From (2.14) we find that the last expression equals

$$u\big(2\hat{w}'(3) + 2Q_\infty\big) \tag{3.22b}$$

where $\hat{w}'(3)$ may be computed from (2.15) to get the constant from Lemma 5.

Regarding (3.15), (3.17) and (3.22a) we find that $[N;z]\hat{w}(z)$ has third order poles in $z = 2$ and $z = 1$, and second order poles in $z_k = 2 + \frac{2k\pi i}{L}$, $k \in \mathbf{Z}$, $z \neq 0$, as well as in $z_k - 1$.

Our local expansions allow (after some lengthy but straightforward computations) to find the following asymptotic behaviour of w_N:

Lemma 7.

$$
\begin{aligned}
w_N = N^2\log_2^2 N + N^2\log_2 N\cdot\Big(-3-\frac{2}{L}+\frac{2\gamma}{L}-2\alpha+\delta_6(\log_2 N)\Big)\\
+N^2\Big(\frac{1}{3}+\alpha^2+3\alpha+\frac{2\alpha}{L}-\frac{3\gamma}{L}-\frac{2\alpha\gamma}{L}-\frac{\beta_2}{L}\\
+\frac{2}{L}+\frac{2}{L^2}+\frac{\gamma^2}{L^2}+\frac{\pi^2}{6L^2}-\frac{2\gamma}{L^2}+\delta_7(\log_2 N)\Big)\\
+3N\log_2^2 N + N\log_2 N\cdot\Big(-\frac{7}{L}-3-6\alpha-\frac{10\gamma}{L}+\delta_8(\log_2 N)\Big)\\
+N\Big(-22-\frac{41}{6L}+\frac{\beta_2}{L}-\frac{3\gamma}{L}-\frac{7\gamma}{L^2}\\
+2\alpha-\beta+\frac{7\alpha}{L}+3\alpha^2-\frac{6\alpha\gamma}{L}+\frac{3\gamma^2}{L^2}\\
+\frac{\pi^2}{2L^2}+\frac{6}{L^2}-\frac{2}{L}\sum_{k\ge3}\frac{(-1)^{k+1}(k-5)}{(k+1)k(k-1)(2^k-1)}\\
+\frac{2}{L}\sum_{r\ge1}b_{r+1}\left(\frac{L(1-2^{-r+1})/2-1}{1-2^{-r}}-\sum_{k\ge2}\frac{(-1)^{k+1}}{k(k-1)(2^{r+k}-1)}\right)\\
+\frac{2}{L}\hat{w}'(3)+\delta_9(\log_2 N)\Big)+\mathcal{O}\left(\frac{\log^2 N}{N}\right)
\end{aligned}
$$

with L,α,β,b_{r+1} as in Theorem 4 resp. Lemma 6 and β_2 from (3.15).

It remains to combine the previous results to get an asymptotic expansion for

$$\operatorname{Var} L_N = u_N + v_N + w_N + l_N - l_N^2. \tag{3.23}$$

We start with an important observation concerning leading terms formed by periodic fluctuations of mean zero.

Let us assume that, at any stage, we are able to prove

$$\operatorname{Var} L_N = \delta_{10}(\log_2 N)\cdot N^\mu \log^\nu N + R_N \tag{3.24}$$

where $\delta_{10}(x)$ is continuous and periodic with period 1 and mean zero and $R_N = o(N^\mu\log^\nu N)$. We claim that $\delta_{10}(x)$ must vanish identically under these conditions:

Let us assume $\delta_{10}(x)\not\equiv 0$. Then, since $\delta_{10}(x)$ is continuous with mean 0, there esists an $\epsilon>0$ and an interval, say $[a,b]\subseteq[0,1]$, such that $\delta_{10}(x)<-\epsilon$ for $x\in[a,b]$. Since $\log_2 N$ is dense modulo 1, $\operatorname{Var} L_N$ would be negative for an infinity of values, an obvious contradiction.

In other words: From (3.24) we may deduce that

$$\operatorname{Var} L_N \sim R_N, \qquad N\to\infty, \tag{3.25}$$

so that, in order to prove that $\operatorname{Var} L_N = \mathcal{O}(N)$ we need not collect explicitely the fluctuating contributions of mean zero.

Observing these comments we easily find that all terms of order $N^2 \log^2 N$, $N^2 \log N$, $N \log^2 N$ and $N \log N$ in $\operatorname{Var} L_N$ cancel. The coefficient of N^2 is of a more delicate nature. The reader should note carefully that the coefficient of N^2 in l_N^2 will contain the square δ_1^2 of the periodic fluctuation δ_1 from Theorem 2 and that the mean $[\delta_1^2]_0$ of δ_1^2 will *not* be zero. Therefore we have to extract this term to end up with a fluctuation of mean zero and get for the coefficient of N^2 in $\operatorname{Var} L_N$ the expression

$$\frac{1}{L^2} + \frac{\pi^2}{6L^2} - \frac{1}{L} - \frac{\beta_2}{L} - \frac{47}{12} - [\delta_1^2]_0 + \delta_{11}(\log_2 N). \tag{3.26}$$

The following Lemma is crucial now:

Lemma 8.

$$\begin{aligned}[\delta_1^2]_0 &= \frac{1}{L^2} \sum_{k \neq 0} \left| \Gamma\left(-1 - \frac{2k\pi i}{\log 2}\right) \right|^2 \\ &= \frac{1}{L^2} + \frac{\pi^2}{6L^2} - \frac{1}{L} - \frac{\beta_2}{L} - \frac{47}{12}.\end{aligned}$$

Sketch of Proof[1]*.* The proof heavily relies on the following two series transformation results due to Ramanujan. The first one is

$$\begin{aligned}&\alpha^{-N}\left(\frac{1}{2}\zeta(2N+1) + \sum_{k\geq 1} \frac{k^{-2N-1}}{e^{2\alpha k}-1}\right) \\ &\quad = (-\beta)^{-N}\left(\frac{1}{2}\zeta(2N+1) + \sum_{k\geq 1} \frac{k^{-2N-1}}{e^{2\beta k}-1}\right) \\ &\quad - 2^{2N} \sum_{k=0}^{N+1} (-1)^k \frac{B_{2k}}{(2k)!} \frac{B_{2N+2-2k}}{(2N+2-2k)!} \alpha^{N+1-k} \beta^k\end{aligned} \tag{3.27}$$

Here and in the next identity, α and β have to be positive numbers with $\alpha\beta = \pi^2$, $\zeta(s)$ is the Riemann ζ–function; N has to be a positive integer and B_n indicates the n-th Bernoulli number defined by

$$\frac{z}{e^z - 1} = \sum_{n \geq 0} B_n \frac{z^n}{n!}.$$

The second identity used in the proof is

$$\begin{aligned}&\sum_{k\geq 1} \frac{1}{k(e^{2\alpha k}-1)} - \frac{1}{4}\log\alpha + \frac{\alpha}{12} \\ &\qquad = \sum_{k\geq 1} \frac{1}{k(e^{2\beta k}-1)} - \frac{1}{4}\log\beta + \frac{\beta}{12}.\end{aligned} \tag{3.28}$$

[1] A full proof of Lemma 8 is long and difficult and included in [10].

In fact, (3.28) is equivalent to a transformation result on Dedekind's η-function (compare [3])

$$\eta(\tau) = e^{\pi i\tau/12}\prod_{n\geq 1}\left(1-e^{2\pi i n\tau}\right), \qquad \Im(\tau) > 0, \tag{3.29}$$

namely

$$\eta(-\frac{1}{\tau}) = (-i\tau)^{1/2}\cdot\eta(\tau), \qquad \Im(\tau) > 0. \tag{3.30}$$

(This is a special instance of Dedekind's famous result on the behaviour of η under a transformation of the modular group.)

The consequences of Lemma 8 are twofold. On the one hand, we find from (3.26) that the N^2-term in $\operatorname{Var} L_N$ cancels, so that $\operatorname{Var} L_N = \mathcal{O}(N)$. On the other hand we may use the identity to express β_2 by the other terms occurring in Lemma 8, including $[\delta_1^2]_0$ which yields the final form of the constant C in Theorem 2. $[\delta_1\delta_2]_0$ is the mean of $\delta_1(x)\delta_2(x)$, originating from l_N^2, which has to be extracted to end up with a fluctuation $\delta(x)$ of mean zero.

We would like to point out some final remarks concerning this analysis:

i) The occurrence of the finite products Q_k gives rise to use results from the *theory of partitions*, especially Euler's product identities (3.12) and (3.13).

ii) A periodic fluctuation $\delta_1(x)$ which has mean zero and very small amplitude may be safely neglected for practical purposes as long as we are only interested in the *expectation*. In order to establish the correct order of the *variance* it is of vital importance to study the behaviour of $\delta_1(x)$, especially the mean of $\delta_1^2(x)$.

iii) The predicted value $0.26600\ldots\cdot N$ matches perfectly with the values obtained by *computer simulations*.

iv) As we have mentioned in the Introduction, with this paper we have finally achieved very good information on the average case behaviour of *Tries, Patricia Tries* and *Digital Search Trees* in the symmetric case (i.e. we start from $0,1$-sequences which contain 0 and 1 with equal probability). Nevertheless the methods of this paper do not seem to be confined to this instance: indeed we hope that sharp results for the asymmetric case may also be established with a similar technical apparatus.

Acknowledgement: The authors would like to thank R.F.Tichy for some helpful remarks.

References

1. M.Abramowitz, I.A.Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1970.
2. A.Aho, J.Hopcroft, J.Ullman, *Data Structures and Algorithms*, Addison Wesley, Reading, MA, 1983.
3. T.M.Apostol, *Modular Functions and Dirichlet Series in Number Theory*, Springer, New York, 1976.
4. E.G.Coffman Jr., J.Eve, *File Structures Using Hashing Functions*, Comm. ACM **13** (1970), 427–436.
5. R.Fagin, J.Nievergelt, N.Pippenger, H.Strong, *Extendible Hashing: A fast Access Method for Dynamic Files*, ACM TODS **4** (1979), 315–344.
6. P.Flajolet, R.Sedgewick, *Digital Search Trees Revisited*, SIAM J. Comput. **15** (1986), 748–767.
7. G. Gonnet, *Handbook of Algorithms and Data Structures*, Addison-Wesley, Reading MA, 1983.
8. P. Jacquet, M.Regnier, *Normal Limiting Distributions for the Size and the External Path Length of Tries*, INRIA Technical Report **827** (1988).
9. P.Kirschenhofer, H.Prodinger, *Further Results on Digital Search Trees*, Theor. Comput. Sci. **58** (1988), 143–154.
10. P.Kirschenhofer, H.Prodinger, *On some Applications of Formulae of Ramanujan in the Analysis of Algorithms*, preprint (1988).
11. P.Kirschenhofer, H.Prodinger, W.Szpankowski, *On the Variance of the External Path Length in a Symmetric Digital Trie*, Discrete Appl. Math. **25** (1989), 129–143.
12. P.Kirschenhofer, H.Prodinger, W.Szpankowski, *On the Balance Property of Patricia Tries: External Path Length Viewpoint*, Theor. Comput. Sci. **68** (1989), 1–17.
13. P.Kirschenhofer, H.Prodinger, W.Szpankowski, *Digital Search Trees - Further Results on a Fundamental Data Structure*, Proceedings of IFIP 89,G.X.Ritter, ed. (1989), 443–447.
14. D.E.Knuth, *The Art of Computer Programming Vol. 3*, Addison-Wesley, Reading MA, 1973.
15. A.G.Konheim, D.J.Newman, *A Note on Growing Binary Trees*, Discrete Mathematics **4** (1973), 57–63.
16. P.Mathys, P.Flajolet, *Q-ary Collision Resolution Algorithms in Random-Access Systems with Free and Blocked Channel Access*, IEEE IT **31** (1985), 217–243.
17. N.E.Nörlund, *Vorlesungen über Differenzenrechnung*, Chelsea, New York, 1954.
18. B.Pittel, *Paths in a Random Digital Tree: Limiting Distributions*, Adv. Appl. Prob. **18** (1986), 139–155.
19. M.Régnier, P.Jacquet, *New Results on the Size of Tries*, IEEE Trans.Inf.Theory **35** (1989), 203–205.
20. J.Riordan, *Combinatorial Identities*, John Wiley & Sons, New York, 1968.
21. A.N.Shiryayev, *Probability*, Springer, New York, 1984.
22. W.Szpankowski, *Some Results on v-ary Asymmetric Tries*, J.Algorithms **9** (1988), 224-244.
23. W.Szpankowski, *A Characterization of Digital Search Trees from the Successful Search Viewpoint*, Theor. Comput. Sci. (in press) (1990).
24. W.Szpankowski, *Patricia Tries again Revisited*, Journal of the ACM (in press) (1990).

TU Vienna, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria
and
Purdue University, West Lafayette, IN 47907, USA