Theoretical Computer Science 68 (1989) 1-17
North-Holland

ON THE BALANCE PROPERTY OF PATRICIA TRIES: EXTERNAL PATH LENGTH VIEWPOINT*

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

Wojciech SZPANKOWSKI**
Department of Computer Science, Purdue University, West Lafayette, IN 47907, U.S.A.

Communicated by J. Diaz
Received April 1988
Revised July 1988

Abstract. In this paper, we give exact and asymptotic approximations for the variance of the external path length in a symmetric Patricia tree. The problem was open up to now. We prove that for the binary Patricia tree, the variance is asymptotically equal to $0.37\ldots \cdot n + nP(\log_2 n)$ where n is the number of stored records and $P(x)$ is a periodic function with a very small amplitude. This implies that the external path length is asymptotically equal to $n \cdot \log_2 n$ *with probability one* (i.e., almost surely). These results are next used to show that from the practical (average) viewpoint, the Patricia tree does not need to be *restructured* in order to keep it balanced. In general, we ask to what extent simpler and more direct algorithms (for digital search trees) can be expected in practice to match the performance of more complicated, worst-case asymptotically better ones.

1. Introduction

The optimization of the asymptotic worst-case performance is a major aim in the design of most algorithms. In this endeavor lots of insightful, elegant and clever constructions have been made. Along these lines, however, the algorithmic design has often to be targeted at coping efficiently with quite unrealistic, if not pathological, inputs and the possibility is neglected that a simpler algorithm might perform just as well, or even better, in practice. A remedy to this situation is to reconsider the algorithm from the (more natural) average complexity viewpoint. This approach can give a more realistic picture of the overall behavior of an algorithm. In this paper, we apply this strategy to study digital search trees (e.g., Patricia tries) and ask how well on the average these trees are balanced. We will argue that the variance of the external path length in digital search trees is a good measure of the balancing property of the trees.

* A preliminary version of this paper appears in the *Proceedings of ICALP 88*, Tampere 1988.

** The research was supported in part by the National Science Foundation under Grant NCR-8702115.

In 1979, Fagin et al. [2] proposed extendible hashing as a fast access method for dynamic files. In the original version of this method, radix search trees (tries in short) 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. Do we really need to balance the tree? Before we answer this question, let us first consider another, more efficient data structure, namely the Patricia tries for accessing the keys. The Patricia trie was discovered by D.R. Morrison (see [1, 4, 9]), who suggested how to avoid an annoying flaw of regular tries, namely, one-way branching on internal nodes. To recall, a regular trie is a data structure that uses the digital properties of keys. It consists of internal nodes and external nodes. The internal nodes are used to branch a key (e.g., "go left", if the next digit of a key is 0, and "go right" if the next digit is 1), while external nodes contain the minimal prefix information of keys (records). In the Patricia trie, all one-way branches are collapsed on internal nodes, that is, all unary branching nodes are eliminated (for more detailed discussion, see [4, 9]). As with regular tries, the Patricia must be accompanied by an additional procedure in order to balance it, and to achieve good worst-case performance. This restructuring generally changes the entire tree and is rather an expensive operation (compare also binary search trees and AVL trees). Again, the question is whether we really need to balance the Patricia trie. We answer that question from the average complexity viewpoint. Finally, we note that digital search tries find many other applications in computer science and telecommunications such as partial match retrieval of multidimensional data, conflict resolution algorithms for broadcast communications [10], radix exchange sort, polynomial factorization, simulation [4, 9], lexicographical sorting [1, 14].

Two quantities of a digital trie are of special interest, *depth of a leaf* (search time) and the *external path length*. The average depth of a leaf for regular tries and Patricia trie has been studied in [3, 6, 9, 11, 13], the variance in [6, 11, 13] and the higher moments in [11, 13]. The average value of the external path length is closely related to the average depth of a leaf, but *not* the variance. The first attempt to compute the variance was reported in [6], however, it turned out that the variance of the successful search time was estimated, *not* the variance of the external path length. This was rectified by Kirschenhofer et al. in [8], who obtained the correct value for the variance in the symmetric regular tries. In this paper, we propose how to evaluate the appropriate variance for the Patricia trie, which was an open problem up to now. We argue that the variance of the external path length is responsible for a good balance property of the Patricia tries. In addition, we note that the external path length analysis finds direct important applications in such algorithms as modified lexicographical sorting [14], conflict resolution algorithms for broadcast communications [10], etc.

This paper is organized as follows. In the next section, we define our model, establish general methodology to attack the problem and present our main results. In particular, we show that the variance of the external path length for the *binary symmetric* Patricia trie is $0.37\ldots \cdot n + nP(\log_2 n)$ where n is the number of records

and $P(x)$ is a periodic function with small amplitude. This implies that the external path length converges *in probability* and *with probability* 1 to $n \cdot \log_2 n$. Finally, Section 3 contains the proof of our main result.

2. Statement of the problem and main results

Let T_n be a family of Patricia tries built from n records with keys from random bit streams. A key consists of 0s and 1s (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 occurrence of these two elements in a bit stream is independent of each other. This defines the so-called *Bernoulli model.*

Let L_n^P 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 and the variance of L_n^P. Let the probability generating function of L_n^P be denoted as $L_n^P(z)$, that is, $L_n^P(z) = \mathrm{E} z^{L_n^P}$. Note that in the Bernoulli model the n records are split randomly into the left subtree and the right subtree of the root. If X denotes the number of keys in the left subtree, then X is Bernoulli distributed with parameters n and p. Then, for $X = k$, the following holds:

$$L_n^P = \begin{cases} n + L_k^P + L_{n-k}^P, & \text{for } k \neq 0, n \\ L_n^P, & \text{for } k = 0, k = n \end{cases} \tag{2.1}$$

where L_k^P, L_{n-k}^P represent the external path length in the left and right subtrees. Note, that if either left or right subtree is degenerate (i.e., $k = 0$ or $k = n$) then in the Patricia an appropriate internal node is "skipped". Using (2.1) we immediately prove, after some elementary algebra:

Lemma 2.1. *The probability generating function $L_n^P(z)$ satisfies the following recurrence*:

$$L_0^P(z) = L_1^P(z) = 1, \tag{2.2a}$$

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

The appropriate recurrence for the generating function, $L_n^T(z)$, of the external path length, L_n^T, in a family of *regular* (radix search) tries is given by (2.2) except that the last term in (2.2b) is dropped (see [8]). This reflects the fact that in regular tries empty subtrees are allowed (one-way branching nodes). In other words, the equivalent recurrence to (2.1) in regular tries is simply $L_n^T = n + L_k^T + L_{n-k}^T$ for all $k = 0, 1, \ldots, n$.

Let now $l_n^P = \mathrm{E}L_n$ and $\bar{L}_n^P = \mathrm{E}L_n^P(L_n^P - 1)$, that is, l_n^P is the average value of the external path length in Patricia trie and $\bar{L}_n^P$ is the second factorial moment of L_n^P. Note that $l_n^P = L_n'(1)$ and $\bar{L}_n^P = L_n''(1)$, where $L_n'(1)$ and $L_n''(1)$ denote the first and the second derivative of $L_n^P(z)$ at $z = 1$. Simple algebra applied to (2.2) reveals that l_n^P and $\bar{L}_n^P$ satisfy the following recurrences:

$$\begin{aligned} &l_0^P = l_1^P = 0, \\ &l_n^P = n(1-p^n-q^n) + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (l_k^P + l_{n-k}^P), \quad n \geq 2 \end{aligned} \tag{2.3}$$

and

$$\begin{aligned} &\bar{L}_0^P = \bar{L}_1^P = 0, \\ &\bar{L}_n^P = 2nl_n^P(1-p^n-q^n) - n(n+1)(1-p^n-q^n) \\ &\qquad + 2\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} l_k^P l_{n-k}^P + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} [\bar{L}_k^P + \bar{L}_{n-k}^P]. \end{aligned} \tag{2.4}$$

Knowing l_n^P and $\bar{L}_n^P$, one immediately obtains the variance of L_n^P, as

$$\operatorname{var} L_n^P = \bar{L}_n^P + l_n^P - (l_n^P)^2. \tag{2.5}$$

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

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

$$\begin{aligned} &u_0^P = u_1^P = 0, \\ &u_n^P = nl_n^P(1-p^n-q^n) + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (u_k^P + u_{n-k}^P), \quad n \geq 2, \end{aligned} \tag{2.7}$$

$$\begin{aligned} &w_0^P = w_1^P = 0, \\ &w_n^P = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} l_k^P l_{n-k}^P + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (w_k^P + w_{n-k}^P), \quad n \geq 2. \end{aligned} \tag{2.8}$$

Then

$$\bar{L}_n^P = 2u_n^P - v_n^P + 2w_n^P. \tag{2.9}$$

We note here that regular tries are analyzed in a similar manner [8]. The average path length, l_n^T, satisfies a recurrence like (2.3), except that the first term, i.e., $n(1-p^n-q^n)$, is replaced simply by n. If one drops the factor $(1-p^n-q^n)$ in (2.4), (2.6) and (2.7), we obtain equivalent quantities for the regular tries, i.e., $\bar{L}_n^T$, v_n^T, u_n^T. The quantity w_n^T for tries satisfies (2.8) with l_k^P, l_{n-k}^P replaced by l_n^T and l_{n-k}^T. This suggests that there is a close relationship between the appropriate parameters

of regular tries and Patricia tries. We explore this fact in the derivation of our main result.

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 [9, 15]) 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 functions of $\hat{a}_n$ and a_n are related by $\hat{A}(-z) = A(z)\,e^{-z}$. Using this, in [11], the following lemma is proved.

Lemma 2.2. (i) *The recurrence* (2.10) *possesses the following solution*:

$$x_n = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \frac{\hat{a}_k + k a_1 - a_0}{1 - p^k - q^k}. \tag{2.12}$$

(ii) *The inverse relative* $\hat{x}_n$ *of* x_n *satisfies*

$$\hat{x}_n = \frac{\hat{a}_n + n a_1 - a_0}{1 - p^n - q^n}, \quad n \geq 2. \tag{2.13}$$

Finally, to find asymptotic approximation for x_n, we apply a general approach proposed either in [3] (Rice's method) or in [12] (Mellin like approach, see also [9]). Namely, we consider an alternating 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 Lemma 2.2.

Lemma 2.3 (Rice's method, see [3, 6]). *Let* C *be a curve surrounding the points* $2, 3, \ldots, n$, *and* $f(z)$ *be an analytical continuation of* $f(k)$ *inside* C. *Then*

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

with

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

Proof. The formula is a direct consequence of Cauchy's residue theorem [5]. For details, see [3]. □

An alternative approach to estimate the asymptotics for the alternating sum S_n is proposed in [12]. It is proved there that

$$S_n = \frac{1}{2\pi i}\int_{-3/2-i\infty}^{-3/2+i\infty} \Gamma(z)f(-z)n^z\,dz$$
$$= \frac{1}{2\pi i}\int_{-3/2-i\infty}^{-3/2+i\infty} B(n+1,z)f(-z)\,dz, \tag{2.15a}$$

where $n^z = \Gamma(n+1)/\Gamma(n+1+z)$, and

$$B(n+1,z) = \frac{n!}{z(z+1)\cdots(z+n)},$$

and $\Gamma(z)$ is the gamma function [1]. Equivalently, if one notices that $n^z = n^{-z}[1+zO(n^{-1})]$, then (2.15a) can be simplified to

$$S_n = \frac{1}{2\pi i}\int_{-3/2-i\infty}^{-3/2+i\infty} I(z)f(-z)n^{-z}\,dz + e_n \tag{2.15b}$$

where

$$e_n = O(n^{-1})\int_{-3/2-i\infty}^{-3/2+i\infty} z\Gamma(z)f(-z)n^{-z}\,dz,$$

that is, $e_n = o(n)$. Formulas (2.15) resemble the Mellin like approach discussed in [9], and first proposed by De Bruijn.

To apply Lemma 2.3 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 formula (2.15), we find residues *right* to the line $(-\frac{3}{2}-i\infty, -\frac{3}{2}+i\infty)$. Hence, by the residue theorem and Lemma 2.3

$$\sum_{k=2}^{n}(-1)^k\binom{n}{k}f(k) = \sum_{k=-\infty}^{\infty}\operatorname{res}\{[n;z_k]f(z_k)\} + O(n^{-M})$$
$$= \sum_{k=-\infty}^{\infty}\operatorname{res}\{\Gamma(z_k)f(-z_k)n^{-z_k}\} + e_n + O(n^{-M}) \tag{2.16}$$

for any $M>0$ and the sums are taken over all singularities, z_k, $k=0,\pm1,\ldots$, of the functions under the integrals (2.14) and (2.15) in the appropriate regions, respectively. By (2.16), the asymptotics of the alternating sum of type (2.12) (Lemma 2.2) is reduced to compute the residues of the functions under the integrals, which is usually an easy task. In [8] we have mainly used a Mellin-like approach to prove our results for the regular (radix) tries. Therefore, in this paper, we exclusively adopt Rice's approach.

In this paper, we concentrate on the analysis of binary *symmetric* Patricia tries, that is, $p=q=0.5$. Note, however, that using our general approach (i.e., Lemmas

2.2 and 2.3), we can produce exact solutions to an asymmetric V-ary Patricia tries. In the following analysis, we shall extensively use the appropriate results obtained by the authors in [8] for the binary symmetric radix search tries. We summarize these results in the next therorem.

Theorem 2.4. *For binary symmetric radix tries the following holds*:

(i) (Knuth [9]). *The exact value of the average of the external path length,* l_n^T, *is*

$$l_n^T = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \frac{k}{1-2^{1-k}} \tag{2.17a}$$

and the inverse, $\hat{l}_n^T$ *of* l_n^T *is given by*

$$\hat{l}_n^T = \frac{n}{1-2^{1-n}}, \quad n \geq 2. \tag{2.17b}$$

For large n the following also holds:

$$l_n^T = n \log_2 n + n\left[\frac{\gamma}{L} + \tfrac{1}{2} + \delta(\log_2 n)\right] - \tfrac{1}{2}L + \delta_1(\log_2 n) \tag{2.18}$$

where $L = \log 2$ (log *means natural logarithm*), $\gamma = 0.577\ldots$ *is the Euler constant,* $\delta(x)$ *and* $\delta_1(x)$ *are periodic functions with small amplitude and mean zero.*

(ii) (Kirschenhofer et al. [8]). *For large n the variance,* var L_n^T *of the external path length is equal to*

$$\text{var } L_n^T = n[A + P_1(\log_2 n)] + O(\log^2 n) \tag{2.19}$$

where

$$A = 1 + \frac{1}{2L} - \frac{1}{L^2} + \frac{2}{L}(\mu + \nu) + \tau, \tag{2.20}$$

$$\mu = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k(2^k - 1)}, \qquad \nu = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{2^k - 1}, \tag{2.21a}$$

$$\tau = \frac{4\pi^2}{\log^3 2} \sum_{k=1}^{\infty} \frac{k}{\sinh(2k\pi^2/\log 2)} \tag{2.21b}$$

and $P_1(x)$ *is a continuous periodic function with period* 1 *and very small amplitude and mean zero (the contribution from* τ *is also very small).*

Using this result, we prove in Section 3 our main result of this paper.

Theorem 2.5. *For binary symmetric Patricia tries, the following holds*:

(i) *The exact solution for the average of the external path length is*

$$l_n^P = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \frac{k2^{1-k}}{1-2^{1-k}} = l_n^T - n + \delta_{n,1} \tag{2.22a}$$

and

$$\hat{l}_n^P = \frac{n2^{1-n}}{1-2^{1-n}} = 2^{1-n}\hat{l}_n^T, \quad n \geq 2. \tag{2.22b}$$

(ii) *The variance* $\operatorname{var} L_n^P$ *of the external path length is*

$$\begin{aligned} \operatorname{var} L_n^P &= \operatorname{var} L_n^T - n[A_1 + P(\log_2 n)] + O(\log^2 n) \\ &= nA + n \cdot P(\log_2 n) + O(\log^2 n) \end{aligned} \tag{2.23a}$$

where

$$A_1 = \frac{11}{2L} - 2 - \frac{2}{L}(\nu + \theta) \approx 3.9785, \qquad A = 1 + \frac{1}{L} - \frac{1}{L^2} - \tau \approx 0.37\ldots \tag{2.23b}$$

with ν *and* τ *defined in* (2.21a,b), *and*

$$\theta = \sum_{j=2}^{\infty} \frac{(-1)^{j-1}2^j}{j(2^j-1)} \left[\frac{j}{2(2^{j-1}-1)} - 1 \right] = 3 - \log 2 - 2\nu - \mu. \tag{2.24}$$

Numerical evaluation reveals that $\operatorname{var} L_n^T = 4.37\ldots \cdot n + nP_1(\log_2 n)$ and $\operatorname{var} L_n^P = 0.37\ldots \cdot n + nP(\log_2 n)$.

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

Remark 2.6 (*Extension to V-ary Patricia tries*). Using our general approach (Lemmas 2.2 and 2.3), we are able to present exact solutions to the variance of the external path length in the V-ary asymmetric case (see [8, 9, 13] for definitions). Unfortunately, the asymptotic analysis *cannot* be easily extended to the asymmetric case, since we are not able to find an analytical continuation of the solution of w_k^P (see [8] for more detailed comments). Nevertheless, the asymptotics of $\operatorname{var} L_n^P$ in the symmetric V-ary case is easy to obtain from our analysis.

Remark 2.7 (*The covariance analysis*). The proposition and the results from [6, 13], where the variance of the depth of a leaf in the Patricia was established, provide asymptotics for the covariance between two different depths of leaf in the Patricia. Let D_n be a depth of a (randomly selected) leaf and let $D_n^{(i)}$ be a path from the root to the ith external node. Note that the external path length L_n^P is defined in terms of $D_n^{(i)}$ as $L_n^P = \sum_{i=1}^n D_n^{(i)}$. Then

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

and this implies

$$\operatorname{var} L_n^P = n \operatorname{var} D_n + 2 \sum_{i \neq j} \operatorname{cov}\{D_n^{(i)}, D_n^{(j)}\}. \tag{2.25}$$

The variance of the depth var D_n for symmetric Patricia was analyzed in [6], and for asymmetric Patricia in [13]. In particular, it was proved that for the binary symmetric Patricia var $D_n = 1.000\ldots$. Using our main result and (2.25) we find

$$2 \sum_{i \neq j} \operatorname{cov}\{D_n^{(i)}, D_n^{(j)}\} = -0.63\ldots \cdot n. \tag{2.26}$$

This also implies, in the symmetric case, that $\operatorname{cov}\{D_n^{(i)}, D_n^{(j)}\} \sim -0.63\ldots/n$. Note that the equivalent quantity for regular tries is approximately equal to $+0.84\ldots/n$ [8].

Remark 2.8 (*How well is the Patricia balanced*?). The Patricia is a very well-balanced tree. The random shape of the Patricia is on average very close to a complete binary tree which is the ultimately balanced tree. Therefore, any tree with good balance property should have average depth (external path length) equal to $\log_2 n + O(1)$ ($n \log_2 n + O(n)$), and small variance. Proposition 2.5 implies that the average depth is equal to $\log_2 n + O(1)$, as needed. In addition, we note that by Remark 2.7 any two depths of leaf, say $D_n^{(i)}$ and $D_n^{(j)}$, are *negatively correlated.* This means, that $D_n^{(i)} > \mathrm{E}D_n$ and $D_n^{(j)} < \mathrm{E}D_n$ tend to occur together and $D_n^{(i)} < \mathrm{E}D_n$ and $D_n^{(j)} > \mathrm{E}D_n$ 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 the Patricia. Note, that in the regular tries $\operatorname{cov}\{D_n^{(i)}, D_n^{(j)}\} \sim 0.84/n > 0$ and $D_n^{(i)}$ and $D_n^{(j)}$ in that case are *positively* correlated. This means that if $D_n^{(i)}$ is large, the $D_n^{(j)}$ is also likely to be large.

The second reason for the well-balanced feature of the Patricia follows from Chebyshev's inequality and Proposition 2.5. It is known that for a random variable X, $\Pr\{|X - \mathrm{E}X| > \varepsilon\} \leq \operatorname{var} X/\varepsilon^2$, hence the smaller the variance is, the more balanced X is. In our case $\Pr\{|L_n^P - l_n^P| > \sqrt{n}\varepsilon\} \leq 0.37/\varepsilon^2$. In addition, it seems to us that the external path length is a better measure of the balance property of a tree than the depth of a leaf. To "prove" our claim, consider a three-node Patricia tree. Two possible shapes may occur as shown in Fig. 1. Both possible trees are ultimately well balanced, since they represent different complete binary trees. Note, however, that the variance of the depth of (randomly) chosen leaf is *positive* while the variance of the external path length is equal to *zero.* This heuristic can be extended to more then three-node trees and this suggests that the variance of the external path length can be treated as a measure of how well a tree is balanced.

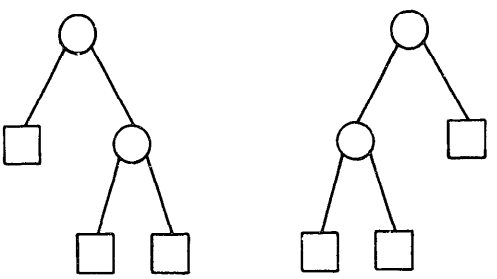

Fig. 1.

Remark 2.9 (*The path length L_n^P converges almost surely to* EL_n^P!). Applying our theorem and proposition it is not difficult to prove that L_n^P/EL_n^P (as well as L_n^T/EL_n^T) tends to 1 *almost surely* (i.e., with probability 1) as $n \to \infty$. Indeed, by Chebyshev's inequality one obtains

$$\Pr\left\{\left|\frac{L_n}{EL_n}-1\right| \geq \varepsilon\right\} \leq \frac{\operatorname{var} L_n}{\varepsilon^2 (EL_n)^2}.$$

But, by (2.22a) and (2.23a)

$$\Pr\left\{\left|\frac{L_n^P}{EL_n^P}-1\right| \geq \varepsilon\right\} \leq \frac{A}{\varepsilon^2 n \log_2^2 n} \to 0. \tag{2.27}$$

Therefore, (2.27) implies that $L_n^P/EL_n^P \to 1$ *in probability as* $n \to \infty$ [16]. To prove a stronger result, namely, that $L_n^P/EL_n^P \to 1$ *with probability* 1 (i.e., almost surely), we apply (2.27) and the Borel-Cantelli lemma [16]. Note that (2.27) implies

$$\sum_{n=1}^{\infty} \Pr\left\{\left|\frac{L_n^P}{EL_n^P}-1\right| \geq \varepsilon\right\} \leq \frac{0.37\ldots}{\varepsilon^2} \sum_{n=1}^{\infty} \frac{1}{n \log_2^2 n} < \infty, \tag{2.28}$$

so, by the Borel-Cantelli lemma $L_n^P \sim EL_n^P \sim n \log_2 n$ with probability 1. These results confirm our hypothesis that the Patricia is a very-balanced tree.

3. The analysis

In this section, we prove Theorem 2.5 for symmetric binary Patricia tries (i.e., $p = q = 0.5$). To simplify the derivations, we shall use extensively our previous results from the binary symmetric regular tries (see Theorem 2.4), that is, we represent all quantities for the Patricia in terms of equivalent quantities for the regular tries.

Let us start with the average of the external path length, l_n^P, which is given by (2.3). This equation falls into our general recurrence (2.10) with the additive term $a_n = n(1-2^{1-n})$ (symmetric case). Hence, by (2.12) we need $\hat{a}_n$ which is $\hat{a}_n = \delta_{n1} + n2^{1-n}$, where δ_{n1} is the Kronecker delta (see [15]). Then, by Lemma 2.2

$$l_n^P = \sum_{k=2}^{n} (-1)^k \binom{n}{k} \frac{k2^{1-k}}{1-2^{1-k}} \quad \text{and} \quad \hat{l}_n^P = \frac{k2^{1-k}}{1-2^{1-k}}. \tag{3.1a,b}$$

Comparing (3.1) with (2.17) one immediately sees that

$$l_n^P = l_n^T - n + \delta_{n1}, \qquad \hat{l}_n^P = 2^{1-n}\hat{l}_n^T, \quad n \geq 2 \tag{3.2a,b}$$

which proves part (i) of Theorem 2.5.

The variance, $\operatorname{var} L_n^P$, of the external path length is given by

$$\operatorname{var} L_n^P = \bar{L}_n^P + l_n^P - (l_n^P)^2$$

where $\bar{L}_n^P$ is shown in (2.4). Hence, using (3.2) and (2.18) one proves

$$\begin{aligned}\operatorname{var} L_n^P &= \bar{L}_n^P + l_n^T - (l_n^T)^2 - n + 2nl_n^T - n^2 \\ &= \bar{L}_n^P + l_n^T - (l_n^T)^2 + \frac{2n^2 \log n}{L} + \frac{2n^2\gamma}{L} - n(1+L^{-1}) + P(\log_2 n)\end{aligned} \tag{3.3}$$

where $L=\log 2$. We shall show that $\bar{L}_n^P=\bar{L}_n^T+g(n)$ for some $g(n)$, hence we represent the variance of the Patricia in terms of the variance of the regular tries var $L_n^T=\bar{L}_n^T+l_n^T-(l_n^T)^2$.

We focus now on the computation of $\bar{L}_n^P$ given by (2.4), that is, $L_n^P=2u_n^P-v_n^P+2w_n^P$ (see (2.9)) where the appropriate components, u_n^P, v_n^P and w_n^P are obtained in recurrences (2.6)-(2.8). Let us first consider v_n^P, that is,

$$\begin{aligned} &v_0^P=v_1^P=0,\\ &v_n^P=n(n+1)(1-2^{1-n})+2^{1-n}\sum_{k=0}^{n}\binom{n}{k}v_k^P, \quad n\geq 2. \end{aligned} \tag{3.4}$$

The equivalent quantity, v_n^T, for regular tries satisfies (3.4) with the adaptive term replaced by $a_n=n(n+1)$. We can write

$$v_n^P=v_n^T-z_n, \tag{3.5a}$$

where

$$z_n=n(n+1)2^{1-n}+2^{1-n}\sum_{k=0}^{n}\binom{n}{k}z_k, \quad n\geq 2 \tag{3.5b}$$

and $z_0=z_1=0$. Note that (3.5b) falls into our general recurrence (2.10) with $a_n=n(n+1)2^{1-n}$, hence $\hat{a}_n=4\binom{n}{2}2^{-n}-4n2^{-n}$ [15], and by Lemma 2.2

$$z_n=\sum_{k=2}^{n}(-1)^k\binom{n}{k}\frac{4\binom{k}{2}2^{-k}-4k2^{-k}+2k}{1-2^{1-k}}. \tag{3.6}$$

We need asymptotics for (3.6), and Lemma 2.3 can be applied. Before we deal with (3.6) we first present one more general result from [11]. Let for some real c and integer r

$$T_{n,r}(c)=\sum_{k=2}^{n}(-1)^k\binom{n}{k}\binom{k}{r}\frac{c^k}{1-2^{1-k}}. \tag{3.7}$$

Then in [11], using Lemma 2.3, we have proved after some simple algebra, the following asymptotic approximation for $T_{n,r}(c)$.

Lemma 3.1. *For any r, c and large n, the following holds:*

$$T_{n,r}(c)=\begin{cases} nc\left\{\log_2 nc+\dfrac{\gamma}{L}-\dfrac{\delta_{n,0}}{L}+\dfrac{1}{2}+(-1)^r P_r(\log_2 nc)\right\}+O(1), & r=0,1,\\ (-1)^r nc\left[\dfrac{1}{r(r-1)L}+P_r(\log_2 nc)\right]+O(1), & r\geq 2 \end{cases} \tag{3.8}$$

where $P_r(x)$ is given by

$$P_r(x)=\frac{1}{L}\sum_{\substack{k=-\infty\\ k\neq 0}}^{\infty}\Gamma(r+2\pi ik/L)\exp[-2\pi ik\log_2 x] \tag{3.9}$$

and $\Gamma(z)$ is the gamma function [5]. *The function $P_r(x)$ is periodic with very small amplitude and can be safely ignored in most practical cases.*

Using Lemma 3.1 we immediately obtain

$$z_n = n\left(\frac{1}{L}+2\right)+n\delta_1(\log_2 n)+O(1) \tag{3.10}$$

where $\delta_1(x)$ is a linear combination of $P_2(x)$ and $P_1(x)$. Therefore, we finally find

$$v_n^P = v_n^T - n(L^{-1}+2) - n\delta_1(\log_2 n)+O(1). \tag{3.11}$$

Now we turn to a relationship between u_n^P and u_n^T, where $u_0^T = u_1^T = u_0^P = u_1^P = 0$ and

$$u_n^P = nl_n^P(1-2^{1-n})+2^{1-n}\sum_{k=0}^{n}\binom{n}{k}u_k^P, \quad n\geq 2, \tag{3.12a}$$

$$u_n^T = nl_n^T+2^{1-n}\sum_{k=0}^{n}\binom{n}{k}u_k^T, \quad n\geq 2. \tag{3.12b}$$

Therefore, the following holds:

$$u_n^P = u_n^T - x_n - y_n \tag{3.13}$$

where

$$x_n = nl_n^T 2^{1-n}+2^{1-n}\sum_{k=0}^{n}\binom{n}{k}x_k, \tag{3.14a}$$

$$y_n = n^2(1-2^{1-n})+2^{1-n}\sum_{k=0}^{n}\binom{n}{k}y_k \tag{3.14b}$$

with zero initial conditions. The recurrence (3.14b) on y_n is easy to analyze noting that it falls into (2.10) with $a_n = 2\binom{n}{2}+n-2^{2-n}\binom{n}{2}-n2^{1-n}$ and hence $\hat{a}_n = 2\delta_{n2}-\delta_{n1}-\binom{n}{2}2^{2-n}+n2^{1-n}$. We have used here the result from Knuth [9] which says

$$a_n = \binom{n}{r}c^n \Leftrightarrow \hat{a}_n = \binom{n}{r}(-c)^r(1-c)^{n-r}. \tag{3.15}$$

Applying Lemmas 2.2 and 3.1, we immediately obtain

$$y_n = 2n^2-2n+n\left[\log_2 n+\frac{\gamma}{L}-\frac{1}{2}-\frac{1}{L}+\delta_2(\log_2 n)\right]+O(1). \tag{3.16}$$

The analysis of x_n is more difficult. We need the inverse relation to $a_n^P = nl_n^T 2^{1-n}$. Let $a_n^T = nl_n^T$. We use the following identities proved in [8, 13]:

$$\hat{a}_n^T = n\hat{l}_n^T - n\hat{l}_{n-1}^T, \quad n\geq 3, \qquad \hat{a}_n^P = 2^{1-n}\sum_{k=0}^{n}\binom{n}{k}\hat{a}_k^T. \tag{3.17a,b}$$

For, by (2.17b) we estimate $\hat{a}_0^T = \hat{a}_1^T = 0$, $\hat{a}_2^T = 8$ and [8]

$$\hat{a}_n^T = \frac{n}{1-2^{1-n}}\left[1-\frac{n-1}{2(2^{n-2}-1)}\right], \quad n\geq 3, \tag{3.18}$$

hence, by (3.17b)

$$\hat{a}_n^P = 8.2^{1-n}\binom{n}{2} + 2^{1-n}\sum_{k=3}^{n}\binom{n}{k}\hat{a}_k^T \tag{3.19}$$

and by Lemma 2.2

$$x_n = 8\sum_{k=2}^{n}(-1)^k\binom{n}{k}\binom{k}{2}\frac{2^{1-k}}{1-2^{1-k}} + \sum_{k=3}^{n}(-1)^k\binom{n}{k}\frac{2^{1-k}}{1-2^{1-k}}\sum_{j=3}^{k}\binom{k}{j}\hat{a}_j^T. \tag{3.20}$$

The asymptotics for the first term of (3.20), say $x_{n,1}$, readily follow from Lemma 3.1, and

$$x_{n,1} = \frac{4n}{L} + 4n\delta_3(\log_2 n) + O(1). \tag{3.21}$$

We need asymptotics for the second term of (3.20), say $x_{n,2}$ and we apply Rice's method from Lemma 2.3 (see (2.14)). Note first, that after some simple algebraic manipulations $x_{n,2}$ can be represented as

$$x_{n+1,2} = (n+1)\sum_{k=2}^{n}(-1)^k\binom{n}{k}\frac{1}{2^k-1}\sum_{j=2}^{\infty}\binom{k}{j}\frac{1}{1-2^{-j}}\left[\frac{j}{2(2^{j-1}-1)}-1\right]. \tag{3.22}$$

The appropriate analytical continuation of the function in (3.22) is

$$f(z) = \frac{1}{2^z-1}\sum_{j=2}^{\infty}\binom{z}{j}\frac{1}{1-2^{-j}}\left[\frac{j}{2(2^{j-1}-1)}-1\right] \tag{3.23}$$

since the series in (3.23) is convergent. To apply Rice's method and (2.16), we need residues of $f(z)$ and $[n;z]$ (see (2.14)) at the poles of $f(z)$ (roots of $2^z-1=0$), that is,

$$\chi_k = \omega_k - 1 = \frac{2\pi i k}{L}. \tag{3.24}$$

The main contribution to the asymptotics comes from $\chi_0 = 0$. Using the following Taylor expansions:

$$[n;z] = -z^{-1} + O(1), \qquad (2^z-1)f(z) = z\theta + O(z^2) \tag{3.25a,b}$$

where

$$\theta = \sum_{j=2}^{\infty}\frac{(-1)^{j-1}2^j}{j(2^j-1)}\left[\frac{j}{2(2^{j-1}-1)}-1\right], \tag{3.26}$$

one immediately proves

$$x_{n,2} = -\frac{n\theta}{L} + n\delta_4(\log_2 n) + O(1) \tag{3.27a}$$

where

$$\delta_4(x)=\frac{1}{L}\sum_{\substack{k=-\infty\\ k\neq 0}}^{\infty}\Gamma(-\chi_k)\exp[2\pi ikx]\sum_{j=2}^{\infty}\binom{\chi_k}{j}\frac{1}{1-2^{-j}}\left[\frac{j}{2(2^{j-1}-1)}-1\right]. \quad (3.27b)$$

So, finally by (3.21), (3.22) and (3.27), we prove

$$x_n=n\frac{4-\theta}{L}+n[4\delta_3(\log_2 n)+\delta_4(\log_2 n)]+O(1)$$

and by (3.13), (3.16) and the above

$$u_n^P=u_n^T-2n^2-n\log_2 n-n\left(\frac{3+\gamma-\theta}{L}-\frac{5}{2}\right)-n\sigma(\log_2 n)+O(1)$$

where $\sigma(x)$ is a linear combination of $\delta_2(x)$, $\delta_3(x)$ and $\delta_4(x)$. Finally, the formula on θ can be simplified a little. Noting that

$$\frac{2^j}{(2^j-1)(2^{j-1}-1)}=-\frac{2}{2^j-1}+\frac{2}{2^{j-1}-1},$$

one obtains

$$\begin{aligned}\theta&=\sum_{j=2}^{\infty}\frac{(-1)^j}{2^j-1}+\sum_{j=1}^{\infty}\frac{(-1)^j}{2^j-1}-\sum_{j=2}^{\infty}\frac{(-1)^{j-1}}{j}-\sum_{j=2}^{\infty}\frac{(-1)^{j-1}}{j(2^j-1)}\\&=1-2\nu-(\log 2-1)-(\mu-1)=3-\log 2-2\nu-\mu\end{aligned} \quad (3.28)$$

where ν and μ are defined in (2.21a).

The most intricate analysis is required for w_n^T which is given by the following recurrence

$$w_n^P=2^{-n}\sum_{k=0}^{n}\binom{n}{k}l_k^P l_{n-k}^P+2^{1-n}\sum_{k=0}^{n}\binom{n}{k}w_k^P, \quad n\geq 2. \quad (3.29)$$

We appeal again to our analysis of regular tries. The appropriate recurrence for w_n^T replaces l_k^P and l_{n-k}^P with l_k^T and l_{n-k}^T. The inverse relation to the additive term a_n^P in (3.29) can be computed as (we use here (2.22b))

$$\hat{a}_n^P=2^{2-n}\cdot 2^{-n}\sum_{k=0}^{n}\binom{n}{k}\hat{l}_k^T\hat{l}_{n-k}^T=2^{2-n}\hat{a}_n^T. \quad (3.30a)$$

In [8] we have proved that for regular tries

$$\hat{a}_n^T=\frac{n(n-1)}{2}\frac{1}{2^{n-2}-1}\left[2^{n-3}-1+\sum_{j=1}^{\infty}\binom{n-2}{j}\frac{1}{2^j-1}-\frac{1}{2^{n-2}-1}\right], \quad n\geq 3 \quad (3.30b)$$

hence, after some algebra

$$w_n^P=2w_n^T-2\sum_{k=3}^{n}(-1)^k\binom{n}{k}\hat{a}_k^T. \quad (3.31)$$

We need to estimate the second term in (3.31), which we denote as B_n. After some algebra, we prove

$$B_{n+1} = (n+1)\sum_{k=2}^{n}(-1)^k\binom{n}{k}\frac{k}{2^{k-1}-1} \times\left[1-2^{k-2}+\frac{1}{2^{k-1}-1}-\sum_{j=2}^{\infty}\binom{k-1}{j}\frac{1}{2^j-1}\right]. \tag{3.32}$$

Therefore, Rice's method (Lemma 2.3) can be applied with the analytical continuation function $f(z)$ as below

$$f(z)=\frac{z}{2^{z-1}-1}\left[1-2^{z-2}+\frac{1}{2^{z-1}-1}-\sum_{j=2}^{\infty}\binom{z-1}{j}\frac{1}{2^j-1}\right]. \tag{3.33}$$

The poles of $f(z)$ are at

$$\omega_k = 1+\chi_k = 1+\frac{2\pi i k}{L}, \quad k=0,\pm 1,\ldots .$$

As before, the main contribution comes from $\omega_0 = 1$. We use the following Taylor's expansions with $u = z-1$ [6]:

$$[n;z]=\frac{n}{u}(1+\lambda_1 u+\lambda_2 u^2)+O(u^2),$$

$$\frac{1}{(2^{z-1}-1)^2}=\frac{1}{L^2u^2}-\frac{1}{Lu}+\frac{5}{12}+O(u),$$

$$\frac{1}{2^{z-1}-1}=\frac{1}{Lu}-\frac{1}{2}+O(u),$$

$$\sum_{j=1}^{\infty}\binom{z-1}{j}\frac{1}{2^j-1}=\mu\cdot u+O(u^2)$$

where

$$\lambda_1 = H_{n-1}-1, \qquad \lambda_2 = 1-H_{n-1}+\tfrac{1}{2}H_{n-1}^2+\tfrac{1}{2}H_{n-1}^{(2)}$$

and μ is defined in (2.21a), while H_n, $H_n^{(2)}$ are harmonic numbers of the first and second order [9]. Multiplying $[n;z]$ and $f(z)$, and identifying the coefficient at u^{-1} (residue value), one proves, after tedious algebra,

$$B_n=\frac{n^2}{2L^2}\log^2 n+\frac{1}{L^2}\left(\gamma-\frac{L}{2}\right)n^2\log n+\frac{n^2}{L^2}\beta_1-\frac{n}{2L^2}\log^2 n -\frac{n}{L^2}\left(\gamma+\frac{3}{2}-\frac{L}{2}\right)\log n-\frac{n}{L^2}\left(\frac{\gamma}{2}+\frac{1}{2}-\frac{L}{4}+\beta_1+\gamma-\frac{L}{2}\right) +O(\log^2 n) \tag{3.34}$$

where

$$\beta_1=\tfrac{1}{2}\gamma^2+\tfrac{1}{12}\pi^2-\tfrac{1}{2}L\gamma-\mu L-\tfrac{1}{3}L^2.$$

From [8] we know that the appropriate asymptotics for w_n^T is

$$w_n^T = \frac{n^2}{2L^2}\log^2 n + \frac{n^2}{L^2}\left(\gamma - \frac{3L}{2}\right)\log n + \frac{n^2}{L^2}\beta_2 - \frac{n}{2L^2}\log^2 n$$
$$-\frac{n}{L^2}\left(\gamma - \frac{3L}{2} + \frac{3}{2}\right)\log n + \frac{n}{L^2}\left(L - \frac{3\gamma}{2} - \frac{1}{2} - \beta_2 + L\nu\right)$$
$$+ O(\log^2 n) \tag{3.35}$$

where

$$\beta_2 = \tfrac{5}{3}L^2 - \tfrac{3}{2}L\gamma - L\mu + \tfrac{1}{2}\gamma^2 + \tfrac{1}{12}\pi^2.$$

Hence, by (3.31) and the above, we finally obtain

$$w_n^P = w_n^T + (w_n^T - B_n) = w_n^T - \frac{\kappa^2}{L}\log n + \frac{n^2}{L^2}(\beta_2 - \beta_1)$$
$$+ \frac{n}{L}\log n + n\left(\frac{1}{4L} + \frac{\nu}{L} + \frac{\gamma}{L} - 2\right) + O(\log_2 n). \tag{3.36}$$

Now we are ready to put all the results together and prove our theorem. Note that $\bar{L}_n^T = 2u_n^T - v_n^T + 2w_n^T$, so

$$\bar{L}_n^P = \bar{L}_n^T - \frac{2n^2}{L}\log n - n\left(\frac{9}{2L} - 3 - \frac{2\nu}{L} - \frac{2\theta}{L}\right) + O(\log^2 n) \tag{3.37}$$

and by (3.3)

$$\operatorname{var} L_n^P = \operatorname{var} L_n^T - n[A_1 + P(\log n)] + O(\log^2 n)$$

with A_1 given by (2.23b). Finally using (3.28) we obtain the constant A in (2.23b), which completes the proof of our theorem.

4. Conclusions

In this paper we investigated asymptotics of the external path length in the Patricia tree. In particular, we concentrated on the variance of the external path length, and proved that the variance is asymptotically equal to $0.37\ldots \cdot n + P(\log n)$. This result was used to prove that the external path length L_n is almost surely (with probability 1) equal to $EL_n^P \sim n\log_2 n$, hence we concluded that the Patricia is a very well-balanced tree, and in most practical situations it does not need to be additionally rebalanced.

Finally, the reader may wonder why we have used the results from regular tries to prove the appropriate result for the Patricia. Is it not simpler to focus only on Patricia, and, since we have our general lemmas (Lemmas 2.1, 2.2, 2.3), to derive directly the variance for the Patricia? It is, of course, possible. However, we had to

cope with the following problem. When deriving the results directly for the Patricia, we would obtain

$$\text{var } L_n^P = Bn^2 + An + O(\log^2 n)$$

where A is the coefficient obtained in Proposition 2.5, while B is a fluctuating function. We have used in [8], the Dedekind η-function to prove that $B \equiv 0$ (see also [7]). To avoid this problem in the above derivation, we have chosen another, simpler approach in this paper.

References

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

[2] R. Fagin, J. Nievergelt, N. Pippenger and H. Strong, Extendible hashing: A fast access method for dynamic files, *ACM TODS* **4** (1979) 315-344.

[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* (*ICALP'86*), Lecture Notes in Computer Science **226** (Springer, Berlin, 1986) 177-185.

[7] P. Kirschenhofer and H. Prodinger, On some applications of formulae of Ramanujan in the analysis of algorithms, Preprint (1987).

[8] P. Kirschenhofer, H. Prodinger and W. Szpankowski, On the variance of the external path length in a symmetric digital trie, *Discrete Appl. Math.*, Special Issue on "Combinatorics and Complexity", to be published.

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

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

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

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

[13] W. Szpankowski, Patricia tries again revisited, Purdue University, Tech. Rept. CSD-TR 625 (1986).

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

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

[16] A. Renyi, *Probability Theory* (North-Holland, Amsterdam, 1970).