DO WE REALLY NEED TO BALANCE PATRICIA TRIES?

(Extended Abstract)

Peter Kirschenhofer
Helmut Prodinger
Institut für Algebra und
Diskrete Mathematik
TU Wien
A-1040 Wien, AUSTRIA

and

Wojciech Szpankowski*

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

Abstract

In this paper, we give exact and asymptotic approximations for the variance of the external path length in a symmetric Patricia trie. The problem was open up to now. We prove that for the binary Patricia trie, the variance is asymptotically equal to $0.37...n + n\ P(\log_2 n)$ where n is the number of stored records and $P(x)$ is a periodic function with a very small amplitude. This result is next used to show that from the practical (average) viewpoint, the Patricia trie 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 tries) can be expected in practice to match the performance of more complicated, worst-case asymptotically better ones.

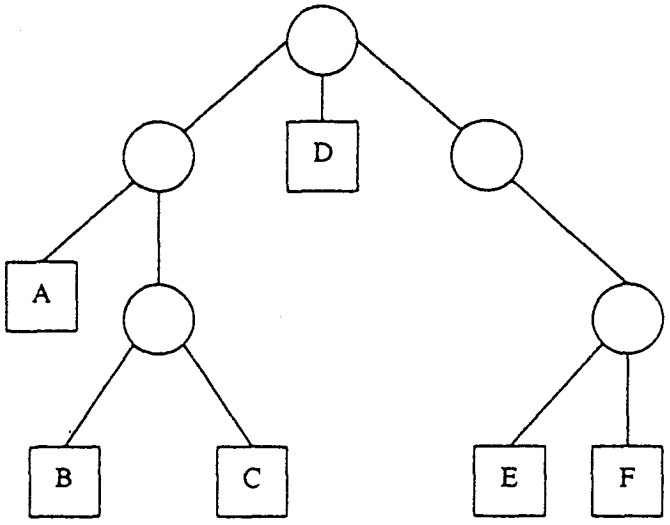
1. INTRODUCTION

Most algorithmic designs are finalized to the optimization of asymptotic worst-case performance. Insightful, elegant and generally useful constructions have been set up in this endeavor. 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 tries (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.

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 a key (record) - see below an example of a regular trie with a ternary alphabet.

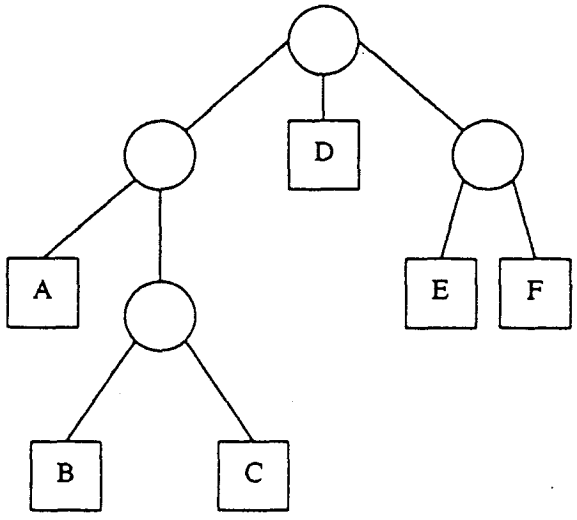
* The research was supported in part by the National Science Foundation under grant NCR–8702115.

A = 100
B = 010
C = 012
D = 100
E = 220
F = 221



In the Patricia trie, all one-way branches are collapsed on internal nodes [9]. For example, in the figure above there is one-way branching on the path from the root to the keys E and F. Collapsing it on an appropriate internal node leads to the Patricia shown below.

A = 000
B = 010
C = 012
D = 100
E = 220
F = 221



As with regular tries, the Patricia must be accompanied with 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], etc.

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, Prodinger and Szpankowski 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 shall 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 directly important applications in such algorithms as modified lexicographical sorting [14], conflict resolution algorithms for broadcast communications [10], etc.

This paper is organized as follow. 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 $0.37\ldots n + n\,P(\log_2 n)$ where n is the number of records and $P(\log_2 n)$ is a periodic function with small amplitude. 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 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 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 of L_n, and the variance var L_n. Let the probability generating function of L_n^P be denoted as $L_n^P(z)$, that is, $L_n^P(z) = Ez^{L_n^P}$. 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 parameters n and p. Then, for $X = k$, the following holds

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

where L_k, L_{n-k} 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 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 $L_n = n + L_k + L_{n-k}$ for all $k = 0, 1, \ldots, n$.

Let now $l_n^P \stackrel{def}{=} EL_n$ and $\bar{L}_n^P = EL_n^P(L_n^P - 1)$, that is, l_n^P is the average value of the external path length in the 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

$$l_0^P = l_1^P = 0$$

$$l_n^P = n(1 - p^n - q^n) + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (l_k^P + l_{n-k}^P) \quad n \geq 2 \tag{2.3}$$

and

$$\bar{L}_0^P = \bar{L}_1^P = 0$$

$$\bar{L}_n^P = 2n\ l_n^P(1 - p^n - q^n) - n(n+1)(1 - p^n - q^n) + 2 \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} l_k^P\ l_{n-k}^P + \tag{2.4}$$

$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} [\bar{L}_k^P + \bar{L}_{n-k}^P]$$

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

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

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

$$v_0^P = v_1^P = 0$$

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

$$u_0^P = u_1^P = 0$$

$$u_n^P = n\ l_n^P(1 - p^n - q^n) + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (u_k^P + u_{n-k}^P) \quad n \geq 2 \tag{2.7}$$

$$w_0^P = w_1^P = 0$$

$$w_n^P = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}\ l_k^P\ l_{n-k}^P + \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} (w_k^P + w_{n-k}^P) \quad n \geq 2 \tag{2.8}$$

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 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), (2.7), we obtain equivalent quantities for the regular tries, that is, $\overline{L_n^T}$, 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 recurrence (2.3)–(2.8), we note that all of these recurrences 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

$$\begin{aligned} x_0 &= x_1 = 0 \\ x_n &= a_n + \sum_{k=0}^{n} \left[\begin{matrix} n \\ k \end{matrix} \right] p^k q^{n-k} (x_k + x_{n-k}) \quad n \geq 2 \end{aligned} \tag{2.10}$$

To solve (2.10), we define a sequence $\hat{a}_n$ (binomial inverse relations [9], [15]) as

$$\hat{a}_n = \sum_{k=0}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix} \right] a_k \Leftrightarrow a_n = \sum_{k=0}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix} \right] \hat{a}_k \tag{2.11}$$

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] it is proved that

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

$$x_n = \sum_{k=2}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix} \right] \frac{\hat{a}_k + k a_1 - a_0}{1 - p^k - q^k} \tag{2.12}$$

(ii) The inverse relations $\hat{x}_n$ 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 approximations for x_n, we apply a general approach proposed either in [3] (Rice's method) or in [12] (Mellin like approach, see also Knuth [9]). Namely, we consider an alternating sum of the form $\sum_{k=2}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix} \right] f(k)$ where $f(k)$ is any sequence. This sum appears in our Lemma 2. Then

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

$$\sum_{k=2}^{n} \left[\begin{matrix} n \\ k \end{matrix} \right] (-1)^k f(k) = \frac{1}{2\pi i} \int_C [n\,;z]\, f(z)dz \tag{2.14}$$

with

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

(ii) [Mellin like approach; see [12]]. Let

$$S_{m,r}(n) = \sum_{k=m}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix} \right] \left\{ \begin{matrix} k \\ r \end{matrix} \right\} f(k)$$

and $f(z)$ be an analytical continuation of $f(k)$ left to the line $(\frac{1}{2} - [m-r]^+ - i\infty,\ \frac{1}{2} - [m-r]^+ + i\infty)$, $a^+ = \max\{0,a\}$. Under certain conditions on the growth of $f(z)$ at infinity (compare [12])

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

where $\int_{(c)}$ stands for $\frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty}$; $\Gamma(z)$ is the gamma function [1], [4] and

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

that is, $e_n = o(n)$.

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 [12].

□

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 3(ii), we find residues *right* to the line $(c - i\infty,\ c + i\infty)$ where $c = \frac{1}{2} - [m-r]^+$. Hence, by the residue theorem and Lemma 3 (for simplicity $r = 0$ is assumed in (2.15))

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

for any $M > 0$ and the sums are taken over all poles, z_k , $k=0,\pm1,\cdots$, 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. 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 method approach.

In this preliminary report, 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., Lemma 2 and 3), we can easily 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 theorem.

Theorem . For binary symmetric radix tries the following holds

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

$$l_n^T = \sum_{k=2}^{n} (-1)^k \left\{ \begin{matrix} n \\ k \end{matrix} \right\} \frac{k}{1 - 2^{1-k}} \tag{2.17a}$$

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

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

For large n the following also holds

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

where $L = \log 2$ (log means natural logarithm), $\gamma = 0.577...$, $\delta(x)$ and $\delta_1(x)$ are periodic functions with small amplitude and mmean zero.

(ii) [Kirschenhofer, Prodinger and Szpankowski [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.

Proposition. For binary symmetric Patricia tries, the following holds

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

$$l_n^P = \sum_{k=2}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix} \right] \frac{k 2^{1-k}}{1-2^{1-k}} = l_n^T - n + \delta_{n,1} \tag{2.22a}$$

and

$$l_n^P = \frac{n 2^{1-n}}{1-2^{1-n}} = 2^{1-n}\, l_n^T \quad n \geq 2 \tag{2.22b}$$

(ii) The variance, var L_n^P of the external path length is

$$\text{var}\, L_n^P = \text{var}\, L_n^T - n[A_1 + P(\log_2 n)] + O(\log^2 n) = n A_2 + n \cdot P(\log_2 n) + O(\log^2 n), \tag{2.23a}$$

where

$$A_1 = \frac{11}{2L} - 2 - \frac{2}{L}(\nu + \theta) \approx 3.9785 \quad ; \quad A_2 = 1 + \frac{1}{L} - \frac{1}{L^2} - \tau \approx 0.37.. \tag{2.23b}$$

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

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

Numerical evaluation reveals that var $L_n^T = 4.37...\, n + n\, P_1(\log_2 n)$ and var $L_n^P = 0.37...\, n + n\, P(\log_2 n)$.

□

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

Remarks

(i) *Extension to V-ary Patricia tries.* Using our general approach (Lemma 2 and 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, and figures in Section 1). 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 var L_n^P in the symmetric V-ary case is easy to obtain from our analysis (see Section 3).

(ii) *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 i-th 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

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

and this implies (see [11])

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

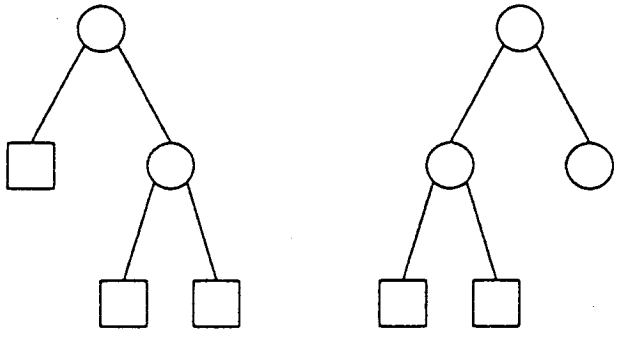
The variance of the depth, var D_n was analyzed in [6], [13]. In particular, it was proved that for binary symmetric tries var D_n = 1.000... (see also [6]). Using our main result and (2.25) we find

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

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

(iii) *How well is the Patricia balanced*? Oh, the Patricia is a very well balanced tree. The random shape of the Patricia is on the average very close to a complete binary tree (the ultimate balance tree). Indeed , note that by remark (ii) any two depths of leaf, say $D_n^{(i)}$ and $D_n^{(j)}$, are *negatively correlated*. This means, that $D_n^{(i)} > ED_n$ and $D_n^{(j)} < ED_n$ tend to occur together and $D_n^{(i)} < ED_n$ and $D_n^{(j)} > ED_n$ 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 the Patricia. Note, that in the regular tries $\text{cov}\{D_n^{(i)}, D_n^{(j)}\} \sim 0.84/n > 0$ 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 likely to be large, too.

The second reason for the well-balanced feature of the Patricia follows from Chebyshev's inequality. It is known that for a random variable X, $Pr\{|X - EX| > \varepsilon\} \leq \frac{\text{var } X}{\varepsilon^2}$, 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 nodes Patricia tree. Two possible shapes may occur as shown below:



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 than 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.

(iv) *The path length L_n^P converges to EL_n^P in probability* ! 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 one *in probability* as $n \to \infty$. Indeed, by Chebyshev's inequality one obtains

$$Pr\{|\frac{L_n}{EL_n} - 1| \geq \varepsilon\} \leq \frac{\text{var } L_n}{\varepsilon^2 (EL_n)^2}$$

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

$$Pr\{|\frac{L_n^P}{EL_n^P} - 1| \geq \varepsilon\} \leq \frac{0.37...}{\varepsilon^2 n \log_2^2 n} \to 0.$$

Hence, $L_n^P/EL_n^P \to 1$ in probability as $n \to \infty$.

3. THE ANALYSIS

In this section, we the present sketch of the proof of our Proposition 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 given in [8] (see Theorem), 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

$$l_n^P = \sum_{k=2}^{n} (-1)^k \left\{ \begin{matrix} n \\ k \end{matrix} \right\} \frac{k2^{1-k}}{1 - 2^{1-k}} \tag{3.1a}$$

and

$$\hat{l}_n^P = \frac{k2^{1-k}}{1 - 2^{1-k}} \tag{3.1b}$$

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

$$l_n^P = l_n^T - n + \delta_{n1} \tag{3.2a}$$

$$\hat{l}_n^P = 2^{1-n}\ \hat{l}_n^T, \quad n \geq 2 \tag{3.2b}$$

which proves Proposition (i).

The variance, var L_n^P, of the external path length is given by

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

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

$$\text{var } L_n^P = \bar{L}_n^P + l_n^T - (l_n^T)^2 + \frac{2n^2 \log_2 n}{L} + \frac{2n^2\gamma}{L} - n\,(1 + L^{-1}) + P(\log_2 n)$$

where $L = \log 2$. We shall show that $\bar{L}_n^P = \bar{L}_n^T + g(n)$ for some $g(n)$, and 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 $\bar{L}_n^P$ which is 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 given by recurrences (2.6)–(2.8). Let us first consider v_n^P, that is,

$$\begin{aligned} v_0^P &= v_1^P = 0 \\ v_n^P &= n(n+1)(1 - 2^{1-n}) + 2^{1-n} \sum_{k=0}^{n} \left[\begin{matrix} n \\ k \end{matrix} \right] v_k^P \qquad n \geq 2 \end{aligned} \tag{3.4}$$

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

$$z_n = n(n+1)2^{1-n} + 2^{1-n} \sum_{k=0}^{n} \left[\begin{matrix} n \\ k \end{matrix} \right] z_k \qquad n \geq 2 \tag{3.5b}$$

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 $\hat{a}_n = 4 \left[\begin{matrix} n \\ 2 \end{matrix} \right] 2^{-n} - 4\, n 2^{-n}$ [15], and by Lemma 2

$$z_n = \sum_{k=2}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix} \right] \frac{4 \left[\begin{matrix} k \\ 2 \end{matrix} \right] 2^{-k} - 4k2^{-k} + 2k}{1 - 2^{1-k}} \tag{3.6}$$

We need asymptotics for (3.6), and Lemma 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

$$T_{n,r}(c) = \sum_{k=2}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix} \right] \left[\begin{matrix} k \\ r \end{matrix} \right] \frac{c^k}{1 - 2^{1-k}} \tag{3.7}$$

Then in [11], using our Lemma 3, we have proved after some simple algebra, the following asymptotic approximation for $T_{n,r}(c)$.

Lemma 4. For any r, c and large n, the following holds

$$T_{n,r}(c) = nc \left\{ \log_2 nc + \frac{\gamma}{L} - \frac{\delta_{r,0}}{L} + \frac{1}{2} + (-1)^r P_r(\log_2 nc) \right\} + O(1), \qquad r = 0,1 \tag{3.8a}$$

$$T_{n,r}(c) = (-1)^r\ nc\left[\ \frac{1}{r(r-1)L} + P_r(\log_2 nc)\ \right] + O(1)\ , \quad r \geq 2 \tag{3.8b}$$

where $P_r(x)$ is given by

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

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 4 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

$$u_n^P = n\ l_n^P(1 - 2^{1-n}) + 2^{1-n} \sum_{k=0}^{n} \left[\begin{matrix} n \\ k \end{matrix} \right] u_k^P \quad n \geq 2 \tag{3.12a}$$

$$u_n^T = n\ l_n^T + 2^{1-n} \sum_{k=0}^{n} \left[\begin{matrix} n \\ k \end{matrix} \right] u_k^T \quad n \geq 2 \tag{3.12b}$$

Therefore, the following holds

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

where

$$x_n = n\ l_n^T 2^{1-n} + 2^{1-n} \sum_{k=0}^{n} \left[\begin{matrix} n \\ k \end{matrix} \right] x_k \tag{3.14a}$$

$$y_n = n^2(1 - 2^{1-n}) + 2^{1-n} \sum_{k=0}^{n} \left[\begin{matrix} n \\ k \end{matrix} \right] y_k \tag{3.14b}$$

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

$$a_n = \left[\begin{matrix} n \\ r \end{matrix} \right] c^n \Leftrightarrow \hat{a}_n = \left[\begin{matrix} n \\ r \end{matrix} \right] (-c)^r (1-c)^{n-r} \tag{3.15}$$

Applying Lemma 2 and 4, we immediately obtain

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

The analysis of x_n is more difficult, however, after some algebra one proves

$$x_n = 8 \sum_{k=2}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix} \right] \left[\begin{matrix} k \\ 2 \end{matrix} \right] \frac{2^{1-k}}{1-2^{1-k}} + \sum_{k=3}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix} \right] \frac{2^{1-k}}{1-2^{1-k}} \sum_{j=3}^{k} \left[\begin{matrix} k \\ j \end{matrix} \right] \hat{a}_j^T \tag{3.17}$$

The asymptotics for the first term of (3.17) readily follows from Lemma 4. To obtain the asymptotics for the second term of (3.17) we apply Lemma 2, and finally after some algebra we prove

$$x_n = n \frac{4-\theta}{L} + n[4\delta_3(\log_2 n) + \delta_4(\log_2 n)] + O(1) \tag{3.18}$$

where

$$\theta = \sum_{j=2}^{\infty} \frac{(-1)^{j-1} 2^j}{j(2^j-1)} \left[\frac{j}{2(2^{j-1}-1)} - 1 \right] \tag{3.19}$$

and

$$\delta_4(x) = \frac{1}{L} \sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} \Gamma(-\chi_k) \exp[2\pi i k x] \sum_{j=2}^{\infty} \left[\begin{matrix} \chi_k \\ j \end{matrix} \right] \frac{1}{1-2^{-j}} \left[\frac{j}{2(2^{j-1}-1)} - 1 \right], \tag{3.20}$$

and $\delta_3(\log_2 n)$ is given in (3.9) after some natural modifications. Now, (3.13), (3.16) and (3.18) finally imply

$$u_n^P = u_n^T - 2n^2 - n \log_2 n - n \left[\frac{3+\gamma-\theta}{L} - \frac{5}{2} \right] - n \sigma(\log_2 n) + O(1) \tag{3.21}$$

where $\sigma(x)$ is a linear combination of $\delta_2(x)$, $\delta_3(x)$ and $\delta_3(x)$.

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

$$w_n^P = 2^{-n} \sum_{k=0}^{n} \left[\begin{matrix} n \\ k \end{matrix} \right] l_k^P \, l_{n-k}^P + 2^{1-n} \sum_{k=0}^{n} \left[\begin{matrix} n \\ k \end{matrix} \right] w_k^P \quad n \geq 2 \tag{3.22}$$

We appeal again to our analysis of regular tries. The appropriate recurrence for w_n^T replaces l_n^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)

$$\hat{a}_n^P = 2^{2-n} \cdot 2^{-n} \sum_{k=0}^{n} \left[\begin{matrix} n \\ k \end{matrix} \right] l_k^T \, l_{n-k}^T = 2^{2-n} \hat{a}_n^T \tag{3.23}$$

and this implies

$$w_n^P = 2w_n^T - 2 \sum_{k=3}^{n} (-1)^k \left[\begin{matrix} n \\ k \end{matrix} \right] \hat{a}_k^T \tag{3.24}$$

where d_n^T is given in [8]. We need to estimate the second term in (3.24), which we denote as B_n. Using the results from [8] one immediately proves that

$$B_{n+1} = (n+1)\sum_{k=2}^{n}(-1)^k \begin{Bmatrix} n \\ k \end{Bmatrix} \frac{k}{2^{k-1}-1}\left[1 - 2^{k-2} + \frac{1}{2^{k-1}-1} - \sum_{j=2}^{\infty}\begin{Bmatrix} k-1 \\ j \end{Bmatrix}\frac{1}{2^j-1}\right] \tag{3.25}$$

Therefore, Rice's method (Lemma 3) can be applied to find asymptotics of B_n. After some tedious algebra we get

$$B_n = \frac{n^2}{2L^2}\log^2 n + \frac{1}{L^2}\left[\gamma - \frac{L}{2}\right] n^2 \log n + \frac{n^2}{L^2}\beta_1 - \frac{n}{2L^2}\log^2 n \tag{3.26}$$

$$- \frac{n}{L^2}\left[\gamma + \frac{3}{2} - \frac{L}{2}\right]\log n - \frac{n}{L^2}\left[\frac{\gamma}{2} + \frac{1}{2} - \frac{L}{4} + \beta_1 + \gamma - \frac{L}{2}\right] + O(\log^2 n)$$

where

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

Comparing the above with w_n^T given in [8] we finally obtain

$$w_n^P = w_n^T + (w_n^T - B_n) = w_n^T - \frac{n^2}{L}\log n + \frac{n^2}{L^2}(\beta_2 - \beta_1) \tag{3.27}$$

$$+ \frac{n}{L}\log n + n\left[\frac{1}{4L} + \frac{\nu}{L} + \frac{\gamma}{L} - 2\right] + O(\log_2 n)$$

where

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

Now we are ready to put all results together and prove our proposition. Note that $\overline{L_n^T} = 2u_n^T - v_n^T + 2w_n^T$, so

$$\overline{L_n^P} = \overline{L_n^T} - \frac{2n^2}{L}\log n - n\left[\frac{9}{2L} - 3 - \frac{2\nu}{L} - \frac{2\theta}{L}\right] + O(\log^2 n) \tag{3.28}$$

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), which completes the proof of our proposition.

REFERENCES

1. Aho, A., Hopcroft, J. and Ullman, J., Data Structures and Algorithms, Addison-Wesley (1983).
2. Fagin, R., Nievergelt, J., Pippenger, N. and Strong, H., Extendible hashing: A fast access method for dynamic files, *ACM TODS*, 4, pp. 315–344 (1979)
3. Flajolet, Ph. and Sedgewick, R., Digital search trees revisited, *SIAM J. Comput.*, 15, pp. 748–767 (1986).
4. Gonnet, G., Handbook of algorithms and data structures, Addison-Wesley (1986).
5. Henrici, P., Applied and computational complex analysis, John Wiley & Sons, New York (1977).
6. Kirschenhofer, P. and Prodinger, H., Some further results on digital search trees in: *Automata, Languages and Machines (ICALP'86)* (L. Kott ed.), pp. 177–185, Springer Lecture Notes in Computer Science 226 (1986).
7. Kirschenhofer, P. and Prodinger, H., On some applications of formulae of Ramanujan in the analysis of algorithms, preprint.
8. Kirschenhofer, P., Prodinger, H. and Szpankowski, W., On the variance of the external path length in a symmetric digital trie, *Combinatorics and Complexity Conference*, Abstracts, pp. 53–54, Chicago (1987) (also submitted to a journal).
9. Knuth, D., The art of computer programming. Sorting and searching. Addison-Wesley (1973).
10. Mathys, P. and Flajolet, P., Q-ary collision resolution algorithms in random-access system with free and blocked channel access, *IEEE Trans. Information Theory*, vol. IT-31, 2, pp. 217–243 (1985).
11. Szpankowski, W., Some results on V-ary asymmetric tries, *Journal of Algorithms*, 9 (1988).
12. Szpankowski, W., The evaluation of an alternative sum with applications to the analysis of some data structures, *Information Processing Letters*, (1988).
13. Szpankowski, W., Patricia tries again revisited, Purdue University, CSD-TR 625 (1986) (also submitted to a journal).
14. Paige, R. and Tarjan, R., Three efficient algorithms based on partition refinement, (preprint) (1986).
15. Riordan, J., Combinatorial Identities, John Wiley & Sons (1968).