Generalized Digital Trees and Their Difference–Differential Equations

Philippe Flajolet*
Algorithms Project, INRIA Rocquencourt, F-78153 Le Chesnay, France

Bruce Richmond†
Department of Combinatorics and Optimization, Univeristy of Waterloo, Waterloo, Ontario, N2L 3G1, Canada

ABSTRACT

Consider a tree partitioning process in which n elements are split into b at the root of a tree (b a design parameter), the rest going recursively into two subtrees with a binomial probability distribution. This extends some familiar tree data structures of computer science like the digital trie and the digital search tree. The exponential generating function for the expected size of the tree satisfies a difference–differential equation of order b,

$$\frac{d^b}{dz^b} f(z) = e^z + 2e^{z/2} f\left(\frac{z}{2}\right).$$

The solution involves going to ordinary (rather than exponential) generating functions, analyzing singularities by means of Mellin transforms and contour integration. The method is of some general interest since a large number of related problems on digital structures can be treated in this way via singularity analysis of *ordinary* generating functions. © 1992 John Wiley & Sons, Inc.

* Supported in part by the Basic Research Action of the E.C. under Contract No. 3075 (Project ALCOM).
† The research of this author was done while visiting INRIA and was also supported in part by NSERC under Grant A-4067.

Random Structures and Algorithms, Vol. 3, No. 3 (1992)

1. INTRODUCTION

We deal with a recursive (tree) partitioning process that depends on some fixed integer parameter b. Given n items, with $n > b$, the process puts b of them aside in the root of a binary tree. The remaining $n - b$ items separate into two subgroups (subtrees), each of them flipping an unbiased coin. The probability that the first subgroup (the left subtree) has size k is thus the binomial probability,

$$\pi_{n,k} = \frac{1}{2^{n-b}} \binom{n-b}{k}. \tag{1}$$

The subgroups again split recursively by the same process. If a group has a cardinality n such that $n \leq b$, then its recursive splitting stops. A realization of this process is clearly attached to a particular binary tree in which internal nodes contain b items, while external nodes contain between 0 and b items. Nodes corresponding to groups of cardinality 0 are called empty nodes.

When $b = 1$, we obtain in this way the classical *digital search tree* structure invented by Coffman and Eve in 1970 [14, p. 489], [21, p. 245]. For $b = 0$, the process defined above determines an infinite tree; however, if we retain only a suitable "finite" part of the tree, we obtain the classical *search "trie"* [14, p. 481], [21, p. 248]. For general b, the corresponding tree structure seems to have been the basis of a folk algorithm in the late 1970s for maintaining paged hashing tables: the idea there is to replace the usual chaining technique by a faster dichotomic access based on bits of hashed records.

Concerning such classical random tree models, we direct the reader to Mahmoud's recent book [18].

Let f_n be the expected number of nonempty nodes of a tree constructed by the basic splitting probabilities (1). (In computer implementations empty nodes need not be represented effectively thanks to the use of "null" pointers). We propose to analyze f_n as a function of n, when the parameter b is kept fixed (we shall suppress the dependence upon b except where it is important). The basic recurrence is

$$\begin{cases} f_0 = 0 & \\ f_n = 1 & \text{if } 1 \leq n \leq b \\ f_n = 1 + \displaystyle\sum_{k=0}^{n-b} \pi_{n,k}(f_k + f_{n-k}) & \text{if } n > b \end{cases} \tag{2}$$

with $\pi_{n,k}$ the Bernoulli probability of Eq. (1).

We introduce the ordinary generating function (OGF) and the exponential generating function (EGF),

$$F(z) = \sum_{n \geq 0} f_n z^n \qquad f(z) = \sum_{n \geq 0} f_n \frac{z^n}{n!}.$$

Lemma 1. *The exponential generating function $f(z)$ satisfies the difference–differential equation*

$$\frac{d^b}{dz^b} f(z) = e^z + 2e^{z/2} f\left(\frac{z}{2}\right), \tag{3}$$

with initial conditions

$$f(z)\Big|_{z+0} = 0, \qquad \frac{d^j}{dz^j} f(z)\Big|_{z=0} = 1 \quad \text{for } j = 1, 2, \ldots, b.$$

The model includes as subcases the usual model of tries ($b = 0$) and the usual model of digital search trees ($b = 1$). When $b = 0$, 1, there is a route by now classical [8, 14] to such equations; it consists of an asymptotic analysis with several stages: (i) explicit solution of the functional equation; (ii) a Taylor expansion providing the coefficients; (iii) analysis of the coefficients via either Mellin transforms or contour integrals of the type used in the calculus of finite differences (the so-called "Rice's method"). We refer to [8] for a partial survey of these techniques.

The itinerary we follow here is different. It starts with the observation that corresponding *ordinary generating functions* satisfy *functional equations* of a simpler form that can be solved by iteration. *Mellin transforms* are then used to determine the behavior of these OGF's near the *dominant singularity*, $z = 1$. This singularity is an accumulation point of simple poles. The behavior of the OGF $F(z)$ is determined by Mellin transforms, through an analysis that requires extended asymptotics in the complex plane; see e.g., [7] for other examples.

Mellin transforms are computed here via standard techniques of contour integrals. They involve higher order basic hypergeometric functions, and in the particular case of $b = 2$, they lead to q-Bessel functions.

We find that near 1, $F(z)$ satisfies

$$F(z) = \frac{1}{(1-z)^2}\left[q_0 + \hat{S}(\log_2(1-z)^{-1})\right] + O((1-z)^{-3/2}), \tag{4}$$

where $\hat{S}(u)$ is a periodic function of u with mean value 0. We can then apply the technique of *singularity analysis* (for a comparable situation, see Odlyzko's

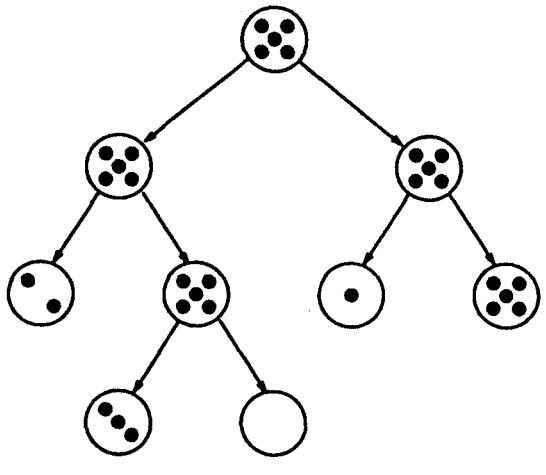

Fig. 1. A generalized tree corresponding to $b = 5$, with $n = 31$ items. The tree has eight nonempty nodes, and one empty node.

counting of 2-3 trees [20], and [6] for general theorems). This enables us to derive an asymptotic form of f_n from the singular expansion (4),

$$f_n = n[q_0 + S(\log n)] + O(n^{1/2}), \tag{5}$$

where S is also a periodic function with mean value 0, see Theorem 1 below.

The path taken here is thus a *two stage analysis* based on (complex) Mellin asymptotics combined with singularity analysis in the style of [6]. A similar two level procedure where (complex) Mellin asymptotics is combined with the saddle point method constitutes the method of Meinardus in the asymptotic theory of integer partitions [2].

Corresponding to a few values of b, here are values of $q_0 \equiv q_0(b)$, as computed by numerical integration.

$$\begin{aligned} q_0(2) &= 0.5747 \\ q_0(3) &= 0.4069 \\ q_0(4) &= 0.3159 \\ q_0(5) &= 0.2585 \\ q_0(10) &= 0.1360\,. \end{aligned} \tag{6}$$

As a byproduct of the analysis, we obtain the estimate $q_0(b) \approx (b \log 2)^{-1}$, for large b. In the context of paging in computer applications, this means that nodes tend to be about 69% full ($\log 2 = 0.69$). Interestingly enough, the same filling ratio is achieved by paged tries and various dynamic hashing strategies [5, 16]. Hoshi and Flajolet [11] provide a review of page occupation in similar tree structures.

2. A STREAMLINED ANALYSIS

In this section, we present the major steps of the analysis in all detail, except for some more technical Mellin transform computations that are relegated to the next section in order not to obscure the proof structure too much.

Our purpose here is to prove the following result:

Theorem 1. *The expected number of nonempty nodes in a random tree built from n elements satisfies*

$$f_n = n[q_0 + S(\log n)] + O(\sqrt{n})$$

where

$$q_0 = \frac{1}{\log 2} \int_0^\infty \frac{(1+t)^{b-1}}{[(1+t)(1+t/2)(1+t/4)\cdots]^b}\, dt\,,$$

and $S(u)$ is a periodic function with mean value 0 and Fourier expansion

$$S(u) = \sum_{k \in \mathbb{Z} \setminus 0} q_k e^{2ik\pi u}$$

with

$$q_k = \frac{1}{(\log 2) \cdot \Gamma(2 + 2ik\pi/\log 2)} \int_0^\infty \frac{(1+t)^{b-1}}{[(1+t)(1+t/2)(1+t/4)\cdots]^b} t^{2ik\pi/\log 2}\, dt\,.$$

A. Basic Equations

We obtain first an explicit form of the OGF, $F(z)$.

Lemma 2. *The ordinary generating function $F(z)$ is given by*

$$F(z) = \frac{1}{1-z} G\left(\frac{z}{1-z}\right), \tag{7}$$

where

$$G(z) = \sum_{j=0}^{\infty} \frac{2^{-bj(j-1)/2} 2^j z^{bj} P(z/2^j)}{[(1+z)(1+z/2)\cdots(1+z/2^j)]^b}, \tag{8}$$

and $P(z) = z(1+z)^{b-1}$.

Proof. Like several other analysis on digital trees, we start by setting

$$g(z) = e^{-z} f(z), \qquad g(z) = \sum_{n \geq 0} g_n \frac{z^n}{n!} \qquad \text{so that } f_n = \sum_{j=0}^{n} \binom{n}{j} g_j \tag{9}$$

The induced differential equation of $g(z)$ is

$$\sum_{j=0}^{b} \binom{b}{j} \frac{d^j}{dz^j} g(z) = 1 + 2g\left(\frac{z}{2}\right),$$

or equivalently for coefficients ($n > b$)

$$\sum_{j=0}^{b} \binom{b}{j} g_{n+j} = 2^{1-n} g_n\,.$$

It is now natural to solve this recurrence by the use of *ordinary* generating functions. Let $G(z) = \sum_{n \geq 0} g_n z^n$. The relation between f_n and g_n first gives rise to the functional relation $F(z) = (1-z)^{-1} G(z/(1-z))$. The recurrence relation on the g_n leads to the simple functional equation (multiply by z^{n+b} and sum),

$$G(z)(1+z)^b = 2z^b G\left(\frac{z}{2}\right) + P(z), \tag{10}$$

where $P(z)$ is obtained by adjusting the terms of order 0 to b, using the relations

$$f(z) = e^z - 1 + O(z^{b+1}), \quad g(z) = 1 - e^{-z} + O(z^{b+1}), \quad G(z) = \frac{z}{1+z} + O(z^{b+1})\,.$$

Finally, the solution of the functional equation (10) is obtained by iteration:

$$G(z) = \frac{P(z)}{(1+z)^b} + \frac{2z^b}{(1+z)^b} G(z/2)$$
$$= \frac{P(z)}{(1+z)^b} + \frac{2z^b \cdot P(z/2)}{(1+z)^b \cdot (1+z/2)^b} + \frac{2z^b \cdot 2(z/2)^b}{(1+z)^b \cdot (1+z/2)^b} G(z/4),$$

and so on.

A closely related form of $G(z)$ is also especially useful:

$$G(t^{-1}) = \sum_{j=0}^{\infty} \frac{2^j \bar{P}(2^j t)}{[(1+t)(1+2t)\cdots(1+2^j t)]^b},$$

with $\bar{P}(z)$ a reciprocal function of $P(z)$,

$$\bar{P}(z) = z^b P(1/z) = (1+z)^{b-1}.$$

Thus generating functions admit expansions as sums of rational functions. These rational functions can in turn be expanded into partial fractions. From there, expressions could be derived for the coefficients f_n: these expressions appear to be somewhat impractical when b exceeds 2, so that we do not attempt to make them explicit; for $b = 0, 1$, they coincide with the forms obtained for tries and standard digital trees by classical methods, and we thus have a new way of carrying out these analyses.

The singularities of $F(z)$ and $G(z)$ are also apparent from these expressions: $G(z)$ has singularities at $z = -2^j$ with accumulation point at $-\infty$; accordingly, $F(z)$ is singular at $z = (1-2^{-k})^{-1}$, for $k = 1, 2, \ldots$, and at their accumulation point $z = 1$.

B. Mellin Transforms

The idea is to estimate $G(z)$ as z tends to ∞ directly, for z in a neighborhood of $+\infty$. This asymptotic expansion provides via the mapping $z \mapsto z/(1-z)$ the behavior of $F(z)$ near its singularity $z = 1$. The analysis relies on Mellin transforms (see [4] or [23] in the context of analysis of algorithms).

The convenient form is to set $z = 1/t$ and we need to consider $t \to 0$. Define first

$$Q(u) = \prod_{j=0}^{\infty} \left(1 + \frac{u}{2^j}\right).$$

Then, from Lemma 1, we have

$$\frac{G(t^{-1})}{Q(t/2)^b} = \sum_{k=0}^{\infty} \frac{2^k (2^k t)^b P\left(\frac{1}{2^k t}\right)}{(Q(2^k t))^b}.$$

By introducing the infinite product, we have thus reduced the analysis of $G(t^{-1})$ to that of the sum,

$$H(t) = \sum_{k=0}^{\infty} 2^k \frac{\bar{P}(2^k t)}{Q^b(2^k t)},$$

a particular case of a "harmonic sum" which is naturally treated by Mellin transforms.

We recall that the Mellin transform of a function $\psi(t)$ is the function denoted $\psi^*(s)$ such that

$$\psi^*(s) = \int_0^{\infty} \psi(t) t^{s-1}\, dt .$$

Lemma 3. *The Mellin transform of* $H(t) = G(t^{-1})/Q^b(t/2)$ *is defined for* $\mathcal{R}(s) > 1$ *and it satisfies*

$$H^*(s) = \frac{h^*(s)}{1-2^{1-s}} \quad \textit{where} \quad h^*(s) = \int_0^{\infty} \frac{(1+t)^{b-1}}{Q^b(t)} t^{s-1}\, dt .$$

Proof. Let us write $h(t) = \bar{P}(t)/Q(t)$. Since $h(at)$ transforms into $a^{-s}h^*(s)$, by linearity, the Mellin transform of $H(t)$ is found to be

$$H^*(s) = h^*(s) \cdot \sum_{k=0}^{\infty} 2^k 2^{-ks} = \frac{h^*(s)}{1-2^{1-s}} .$$

Sufficient conditions for the validity are: (i) the absolute convergence of the sum, which means $\mathcal{R}(s) > 1$; (ii) the absolute convergence of $h^*(s)$ which necessitates $\mathcal{R}(s) > 0$.

C. Mellin Analysis

It is not too hard to use Mellin analysis in order to establish estimates for $G(z)$ when z tends to $+\infty$ along the *real* line, and accordingly, for $F(z)$ when z tends to 1 along the *real* ray [0, 1]. Some of the difficulty of our problems comes from the fact that we need a *continuation into the complex plane* of these asymptotic expansions.

As is usual in a Mellin analysis, we apply the inversion theorem in order to recover $H(t)$ from $H^*(s)$. We have [4]

$$\frac{G(t^{-1})}{Q^b(t/2)} = \frac{1}{2i\pi} \int_{d-i\infty}^{d+i\infty} H^*(s) t^{-s}\, ds , \qquad d > 1 . \tag{11}$$

Apart from the explicit form of $H^*(s)$, we also require in passing some growth estimates for the Mellin transform $h^*(s)$

$$h^*(c+ix) = O(|x|^{b-1} e^{-\pi|x|}) \qquad \text{as } x \to \pm\infty , \quad \text{for } c > 0 \tag{12}$$

whose proof will be deferred till the next section (see Lemma 5).

Lemma 4. *The function $F(z)$ satisfies locally around $z=1$, $|\mathrm{Arg}(z-1)|>\pi/4$,*

$$F(z)=\frac{1}{(1-z)^2}\left[q_0+\hat{S}(\log_2(1-z)^{-1})\right]+O((1-z)^{-3/2}),$$

where $\hat{S}(u)$ is defined by the Fourier expansion

$$\hat{S}(u)=\frac{1}{\log 2}\sum_{k\in\mathbb{Z}\setminus 0} h^*(\xi_k)\,e^{2ik\pi u} \text{ with } \xi_k=1+\frac{2ik\pi}{\log 2}.$$

Proof. We apply the standard technique of Mellin analysis and in the inverse Mellin integral (11) we shift d to the left. If $t=re^{i\phi}$,

$$|t^{-d-ix}|=r^{-d}e^{x\phi}.$$

Thus from the fundamental growth property (12), we can choose any $d>0$ in order to ensure the convergence of (11), provided $|\mathrm{Arg}(t)|<\pi$ and we avoid the line of poles of $(1-2^{1-s})^{-1}$, i.e., $d=1$.

Recall that the standard technique is to pick a contour consisting of two vertical lines with real parts d and d_1, and two horizontal lines with $|\mathscr{I}(s)|\to+\infty$ that pass in between the poles of $H^*(s)$, at $|\mathscr{I}(s)|=2i\pi(k+1/2)/\log 2$. The smallness properties implies that the integrals along horizontal lines are negligible. By Cauchy's residue theorem, we thus get

$$\frac{G(t^{-1})}{Q^b(t/2)}=\sum \mathrm{Res}(H^*(s)t^{-2})_{\mathscr{R}(s)>1/2}+O(t^{-1/2}), \tag{13}$$

provided we take, as we may, $d_1=1/2$. Now the singularities of $H^*(s)$ are at the points

$$\xi_k=1+\frac{2ik\pi}{\log 2},$$

k an integer. These are simple poles with residue $h^*(\xi_k)/\log 2$. Thus, from (13), and using the fact that $Q(t/2)=1+O(t)$, we find our main estimate

$$G(t^{-1})=\frac{1}{t\log 2}\left[h^*(1)+\sum_{k\neq 0} h^*(\xi_k)t^{2ik\pi/\log 2}\right]+O(t^{-1/2}).$$

The residues at the nonreal poles contribute a Fourier series in $\log_2 t$,

$$\hat{S}(\log_2 t)=\frac{1}{\log 2}\sum_{k\in\mathbb{Z}\setminus 0} h^*(\xi_k)e^{2ik\pi\log_2 t}.$$

By the growth conditions on $h^*(s)$, this series has exponentially fast convergence.

From these developments, the asymptotic form of $F(z)$ finally follows via the basic relation $F(z)=(1-z)^{-1}G(z/(1-z))$. The validity condition $|\mathrm{Arg}(t)|<\pi$ for Mellin inversion is amply fulfilled since we have assumed $|\mathrm{Arg}(z-1)|>\pi/4$. We find $q_0=h^*(1)/\log 2$, and periodic fluctuations in the form of the Fourier series $\hat{S}$.

D. Singularity Analysis

We now need to translate term by term an infinite number of periodic fluctuations. The situation is identical to the case of 2–3 trees treated by Odlyzko [20] and also to that of register allocation in [7].

The basic method is the one called singularity analysis and detailed in [6]. Let $[z^n]a(z)$ denote the coefficient of z^n in the Taylor expansion of $a(z)$. We have the following "transfers" from functions to coefficients,

$$\begin{cases} [z^n](1-z)^{-2} & = & n+1 \\ [z^n](1-z)^{-1-\xi_k} & = & \dfrac{n^{\xi_k}}{\Gamma(1+\xi_k)}\left(1+O\left(\dfrac{1}{n}\right)\right) \\ [z^n]O((1-z)^{-3/2}) & = & O(n^{1/2})\,. \end{cases}$$

The last relation is applicable in the context of our problem since the singular expansion holds in an extended area of the complex plane, as guaranteed by Lemma 4.

From these principles, we derive

$$f_n = n\left[q_0 + \frac{1}{\log 2}\sum_{k\in\mathbb{Z}\setminus 0}\frac{h^*(\xi_k)}{\Gamma(1+\xi_k)}\,n^{2ik\pi/\log 2}\right] + O(n^{1/2})\,,$$

which, apart from notational details, coincides with our main estimate in the statement of Theorem 1.

3. MELLIN INTEGRALS AND q-ANALOGUES

We now establish some of the properties of the Mellin transform $h^*(s)$ that are needed in order to complete the proof of Theorem 1. The computations developed here also provide alternative series forms for the integral representation of q_0 and reveal some connections with basic hypergeometric functions that are especially useable for low values of b (we work out the case $b=2$). Finally, they show that the periodic fluctuations in the form of the function $S(u)$ are from a standard set of functions encountered everywhere in this range of problems.

It will prove convenient to operate here with the simple Mellin integral

$$I^*(s) = \int_0^\infty \frac{1}{Q^b(t)}\,t^{s-1}\,dt\,,$$

on whose properties we now concentrate. Properties of $I^*(s)$ easily carry over to $h^*(s)$ because of the relation

$$h^*(s) = \sum_{j=0}^{b-1}\binom{b-1}{j} I^*(s+j)\,.$$

Lemma 5. *The function $I^*(s)$ admits the representation*

$$I^*(s) = \frac{\pi}{\sin \pi s}\,[A_0(2^s) + (s-1)A_1(2^s) + \cdots + (s-1)(s-2)\cdots(s-b+1)A_{b-1}(2^s)]\,,$$

where the $A_k(x)$'s are entire functions.

We observe right away that the $A_k(2^s)$ are themselves entire functions of s with the complex period $2i\pi/\log 2$, and thus they stay bounded as $s = c + ix$ when $x \to \pm\infty$. Therefore,

$$I^*(c+ix) = O(|x|^{b-1}e^{-\pi|x|}) \qquad \text{as } |x| \to \infty\,, \quad \text{for } c > 0\,. \tag{14}$$

This is sufficient to justify the growth estimates used in the previous section.

Proof. We find it useful to write the Mellin transform as a loop integral, using a Hankel contour. Let

$$J(s) = \int_{\mathscr{H}} \frac{1}{Q^b(t)}\,(-t)^{s-1}\,dt\,,$$

where $\mathscr{H}$ denotes a contour that goes from $+\infty - i0$, circles around 0 clockwise and returns to $+\infty + i0$.

Then by a standard argument [24, p. 244] comparing the determinations in the upper and lower half planes, we find

$$J(s) = 2i \sin(\pi s) I^*(s)\,.$$

By another standard argument (see [24] again) we can evaluate the loop integral by residues,

$$J(s) = 2i\pi \sum_{j=0}^{\infty} \operatorname{Res}\left(\frac{(-t)^{s-1}}{(Q(t))^b}\right)_{t=-2^j}.$$

Each residue provides one term in the power series expansion of the $A_k(x)$.

We now describe a procedure to calculate a residue for any fixed b. The problem is to expand the integrand till order $b-1$ around $t = -2^j$, j an integer.

First, we observe that

$$\frac{(-t)^{s-1}}{Q^b(t)} = 2^{jb}\,\frac{1}{(t+2^j)^b}\left[(-t)^{s-1}\prod_{l\neq j}(1+t2^{-l})^{-b}\right]. \tag{15}$$

Thus the problem reduces to evaluating the $(b-1)$st coefficient in the expansion of the quantity inside square brackets at $t = -2^j$. Let $X(t)$ represent the product

$$X(t) = \prod_{l\neq j}(1+t2^{-l})^{-b}\,.$$

From Eq. (15), we expand separately around $t = -2^j$ the two quantities, $X(t)$ and

$(-t)^{s-1}$. We set $t = -2^j + w$. First we have

$$(-t)^{s-1} = 2^{j(s-1)}\left[1 + \frac{(1-s)}{1!}\frac{w}{2^j} + \frac{(1-s)(2-s)}{2!}\left(\frac{w}{2^j}\right)^2 + \cdots\right]. \tag{16}$$

We next proceed to expand $X(-2^j + w)$ around $w = 0$. We have

$$X(-2^j + w) = \prod_{l\neq j}(1-2^{-l+j})^{-b}\prod_{l\neq j}\left(1 + \frac{w}{2^l - 2^j}\right)^{-b}. \tag{17}$$

Now,

$$\log\left(1 + \frac{w}{2^l - 2^j}\right) = \sum_{\alpha\geq 1}\frac{(-1)^{\alpha-1}}{\alpha}\frac{w^\alpha}{(2^l - 2^j)^\alpha}. \tag{18}$$

Set

$$Q_m = \prod_{l=1}^{m}(1-2^{-l}) \quad \text{and} \quad Q_\infty = \prod_{l=1}^{\infty}(1-2^{-l}).$$

From the major equations (15–18), we get at last (!)

$$\operatorname{Res}\left(\frac{(-t)^{s-1}}{(Q(t))^b}\right)_{t=-2^j} = (-1)^{jb}2^{-bj(j+1)/2}\frac{2^{j(s-1)}2^{jb}}{Q_j^b Q_\infty^b} \cdot \sum_{k=0}^{b-1}\frac{(1-s)(2-s)\cdots(k-s)}{k!2^{jk}}Y_{b-1-k}(j), \tag{19}$$

where the $Y_\beta(j)$ are defined by

$$\exp\left(b\sum_{\alpha\geq 1}\frac{(-1)^\alpha w^\alpha}{\alpha}\left(\sum_{l\neq j}\frac{1}{(2^l - 2^j)^\alpha}\right)\right) = \sum_{\beta\geq 0}Y_\beta(j)w^\beta. \tag{20}$$

Finally

$$A_k(x) = \frac{(-1)^k}{k!}\cdot\frac{1}{Q_\infty^b}\sum_{j=0}^{\infty}(-1)^{jb}\frac{2^{-bj(j+1)/2}}{Q_j^b}Y_{b-1-k}(j)(x2^{b-1-k})^j, \tag{21}$$

which concludes the proof of the lemma.

The technique of using loop contours for Mellin transforms gives rise to the famous Hankel representation of the Gamma function [24, p. 244]. Supplemented by a residue calculation as done here, it is one of the ways that are used to derive the functional equation for the Zeta function, following Riemann. It is also the approach used for the computation of Mellin transforms of rational functions [22].

Furthermore, the forms (21) in Lemma 5 relate to the vast domain of q-analogues of special functions—here, basic hypergeometric series with $q = 1/2$; for this, the reader can consult [9], see especially Chapter 4, on integral representations. When $b = 1$, we have $q_0 = 1$ for obvious combinatorial reasons;

analytically, we are brought to q-exponentials (or q-gamma functions [3]) and to a classical identity going back at least to Ramanujan [10, Chap. XI],

$$\int_0^\infty \frac{1}{(1+t)(1+qt)(1+q^2t)\cdots}\, t^{s-1} = \frac{\pi}{\sin \pi s} \prod_{m=1}^{\infty} \frac{1}{1-q^m} \cdot \sum_{j=0}^{\infty} (-1)^j \frac{q^{j(j-1)/2} q^{-j(s-1)}}{(1-q)(1-q^2)\cdots(1-q^j)} = \frac{\pi}{\sin \pi s} \prod_{m=1}^{\infty} \frac{1-q^{m-s}}{1-q^m}.$$

The forms obtained become intricate as b gets large. We cite here:

Theorem 2. *When $b=2$, we have*

$$q_0 \equiv q_0(2) = \frac{1}{\log 2} [2\log 2(2A_0'(4) + 2A_1'(4) - A_0'(2)) + A_1(4) - A_1(2)],$$

where A_0 and A_1 are q-Bessel functions,

$$\begin{cases} A_0(x) = -\dfrac{2}{Q_\infty^2} \displaystyle\sum_{j=0}^{\infty} \frac{2^{-j(j+1)}}{Q_j^2} [\alpha - \zeta_j(1)] x^j \\ A_1(x) = -\dfrac{1}{Q_\infty^2} \displaystyle\sum_{j=0}^{\infty} \frac{2^{-j(j+1)}}{Q_j^2} x^j \end{cases} \tag{22}$$

with

$$\alpha = \frac{1}{1} + \frac{1}{3} + \frac{1}{7} + \frac{1}{15} + \cdots \text{ and } \zeta_j(1) = \frac{1}{1-2^{-1}} + \frac{1}{1-2^{-2}} + \cdots + \frac{1}{1-2^{-j}},$$

the analogues of Euler's constant and of the jth harmonic number, respectively.

From this, we find in a matter of seconds $q_0(2) = 0.57470\ 90927\ 57031\ 98404$.

Proof. At s an integer, the representation provided by Lemma 5 has the indeterminate form $\frac{0}{0}$. We thus apply de L'Hôpital's rule, by which at an integer point $s = 1, 2, \ldots,$ we get

$$I^*(s) = (-1)^s [2^s \log 2(A_0'(2^s) + (s-1)A_1'(2^s)) + A_1(2^s)].$$

The proof then relies on the computation of the A_k's that was detailed in Eqs. (19–21).

The expression of $q_0(2)$ is thus a combination of q-analogues of the Bessel functions J_0 and Y_0. For any given $b \geq 2$, the expressions obtained involve q-analogues of ${}_0F_{b-1}$ hypergeometrics.

From Lemma 5 and the periodicity of the Mellin transforms $h^*(s)$ and $I^*(s)$ that are implied, we also get:

Theorem 3. *The periodic fluctuation $S(u)$ in the coefficient f_n is expressible as a finite linear combination of the standard functions*

$$S^{\langle r\rangle}(u) = \sum_{k\in\mathbb{Z}/0} \Gamma(r-\xi_k)\, e^{2ik\pi u}.$$

Proof. This follows from the form of the Fourier coefficients of $S(u)$, namely

$$\frac{1}{\log 2}\cdot\frac{h^*(\xi_k)}{\Gamma(1+\xi_k)},$$

the decomposition of $I^*(s)$ (with its factor of $\pi/\sin \pi s$ and the periodicities of 2^s) and the complement formula for the Gamma function.

We conclude this section by investigating the dependency of $q_0 = q_0(b)$ on b. In computer applications, b normally represents a page ("bucket") capacity, measured in the number of records that a disk page can contain. The value of b usually varies from a few tens to a few hundreds.

We prove that for large b, $q_0(b) \approx 1/(b \log 2)$. Therefore, neglecting periodic fluctuations, the number of nodes needed to store a file of size n is about

$$\frac{n}{b \log 2}.$$

Since $1/\log 2 = 1.44$, this represents a loss of about 44%, compared to a perfect packing that would require n/b pages. Alternatively, the global storage occupation behaves as though nodes were 69% full ($\log 2 = 0.69$).

Theorem 4. *As b tends to ∞, we have*

$$q_0(b) \sim \frac{1}{b \log 2}.$$

Proof. Take the integral representation of q_0,

$$q_0 = \frac{1}{\log 2}\int_0^\infty \left(\frac{1+t}{Q(t)}\right)^b \frac{dt}{1+t},$$

and use Laplace's method to evaluate the integral for large b. Everything rests on the local expansion at 0, $(1+t)/Q(t) = \exp(-t + O(t^2))$, from which there follows that the integral is $\sim 1/b$.

4. OTHER ANALYSES

It should be clear from the way that our discussion was organized that any suitable additive parameter of generalized digital trees will also yield to these techniques. We can solve recurrences of the general form

$$f_n = e_n + \sum_{k=0}^{n-b} \pi_{n,k}(f_k + f_{n-k})$$

where e_n is the sum of a polynomial in n and of terms of the type $\sum_{j=0}^{b} \lambda_j \delta_{n,j}$ that may be adjusted to any specific initial conditions. The resulting equation is the one corresponding to Lemma 2, with $P(z)$ being in general a rational function that fully characterizes the problem. This opens a new perspective on the analysis of digital tree structures by means of ordinary generating functions since we can proceed in an almost automatic fashion from a problem specification to its asymptotic estimates. In this way, several analyses of standard tries, b-tries, quadtries, or digital search trees can be cast into a unified framework.

In order to keep our statements simple, let us say that a sequence f_n *fluctuates around* $(q_0 \cdot n)$ if $f_n/n = q_0 + S(\log_2 n) + o(n)$ for some $S(u)$ having period 1 and mean value 0. Our argument shows that if the generating function $E(u) = \sum_{n=0}^{\infty} e_n u^n$ has at most a simple pole at 1, then *the sequence* f_n *fluctuates around* $(q_0 \cdot n)$ *where*

$$q_0 = \frac{1}{\log 2} \int_0^\infty \frac{\bar{P}(t)}{Q^b(t)}\, dt \qquad \text{with} \qquad \bar{P}(t) = t(1+t)^{b-1} E\left(\frac{1}{1+t}\right).$$

For instance, taking the number of internal–external nodes in standard digital trees ($b = 1$) [8], we find immediately from the problem specification that

$$E(u) = u \qquad \text{and} \qquad \bar{P}(t) = t/(1+t).$$

Thus (!), we have:

Theorem 5 (Flajolet–Sedgewick [8]). *The average number of internal–external nodes in a digital search tree fluctuates around* $(q_0^* \cdot n)$, *with*

$$q_0^* = \frac{1}{\log 2} \int_0^\infty \frac{t(1+t)^{-1}}{(1+t)(1+t/2)(1+t/4)\cdots}\, dt\,.$$

This is a new integral representation for the old constant 0.37204 that was found in [8, Thm. 2] and derived there under a sum form.

Similarly, we can count directly the total number of nodes of various types in a randomly grown tree. Considering all nodes *including* empty nodes, we have $E(u) = 1/(1-u)$; considering nodes containing j items for $0 \le j \le b$, we have $E(u) = u^j$.

Theorem 6. (*i*). *The average number of nodes including empty nodes in a random generalized digital tree fluctuates around* $(q_0^{**} \cdot n)$, *where*

$$q_0^{**} = \frac{1}{\log 2} \int_0^\infty \frac{(1+t)^b}{Q^b(t)}\, dt\,.$$

(*ii*). *The average number of nodes containing* j *items,* $0 \le j < b$, *fluctuates around* $(q_{0,j}^{**} \cdot n)$, *where*

$$q_{0,j}^{**} = \frac{1}{\log 2} \int_0^\infty \frac{t^j (1+t)^{b-1-j}}{Q^b(t)}\, dt\,.$$

This theorem provides a characterization of node occupancy in such trees that may be useful for paging strategies (see [11] for a discussion).

Finally, still in the same vein, the trees under consideration are fairly well balanced since their path length, computed by these methods, is found to be $n \log_2 n + O(n)$.

Another tree model (based on order statistics rather than random bits) is discussed by Mahmoud and Pittel in [19]; their model also has internal nodes that contain b elements and external nodes that can contain from 0 to b elements. They show that, under certain conditions, the model leads to an asymptotically Gaussian distribution for the size of the tree. A similar limit law holds for the number of nodes in random tries as shown by Jacquet and Régnier [12]. In view of these facts, we expect the number of nodes in a generalized digital tree of size n to be asymptotically normally distributed for large n.

Distributions of other parameters on digital search trees, corresponding to $b = 1$, are discussed by Louchard in [17], who obtained the expected "profile" (i.e., the proportion of nodes at each level in the tree) of standard digital trees. Louchard's result (see also [15]) can be extended to generalized digital trees via our approach, and the profile is again non-Gaussian. Precise height estimates might yield to the probabilistic techniques of Aldous and Shields [1]. It could also be of some interest to investigate the variance of node levels, after Kirschenhofer and Prodinger revealed surprising connections between some of the corresponding asymptotic analysis and certain identities that belong to the elementary theory of modular forms [13]. Again, the recent book by Mahmoud [18] provides a clear perspective on such questions.

ACKNOWLEDGMENT

We are indebted to Prof. Hosam M. Mahmoud for rekindling our interest in this problem.

REFERENCES

[1] D. Aldous and P. Shields, A diffusion limit for a class of randomly growing binary trees, *Probability Theory Related Fields*, **79**(4) 509–542, 1988.

[2] G. E. Andrews, *The Theory of Partitions*, *Encyclopedia of Mathematics and its Applications*, Vol. 2, Addison-Wesley, Reading, MA, 1976.

[3] R. Askey, Ramanujan's extensions of the Gamma and Beta functions, *Am. Math. Monthly*, **87**, 346–359 (1980).

[4] G. Doetsch, *Handbuch der Laplace Transformation*, Vol. 1–3. Birkhäuser Verlag, Basel, Switzerland, 1955.

[5] R. Fagin, J. Nievergelt, N. Pippenger, and R. Strong, Extendible hashing: A fast access method for dynamic files, *A.C.M. Trans. Database Syst.*, **4**, 315–344, (1979).

[6] P. Flajolet and A. M. Odlyzko, Singularity analysis of generating functions, *SIAM J. Discrete Math.*, **3**(2), 216–240 (1990).

[7] P. Flajolet and H. Prodinger, Register allocation for unary-binary trees, *SIAM J. Comput.* **15**(3), 629–640 (1986).

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

[9] G. Gasper and M. Rahman, *Basic Hypergeometric Series, Encyclopedia of Mathematics and its Applications*, Vol. 35, Cambridge University Press, Cambridge, England, 1990.

[10] G. H. Hardy, *Ramanujan: Twelve Lectures on Subjects Suggested by his Life and Work*, third ed., Chelsea Publishing Company, New York, 1978. Reprinted and corrected from the first ed., Cambridge, 1940.

[11] M. Hoshi and P. Flajolet, Page usage in quadtree indexes. Report 1434, Institut National de Recherche en Informatique et an Automatique, May 1991, 19 pages. Accepted for publication, *BIT*.

[12] P. Jacquet and M. Régnier, Trie partitioning process: Limiting distributions, *CAAP'86*, P. Franchi-Zanetacchi, Ed., of *Lecture Notes in Computer Science*, Vol. 214, 1986, pp. 196–210. Proceedings of the 11th Colloquium on Trees in Algebra and Programming, Nice France, March 1986.

[13] P. Kirschenhofer and H. Prodinger, On some applications of formulae of Ramanujan in the analysis of algorithms, *Mathematika*, **38**, 14–33 (1991).

[14] D. E. Knuth, *The Art of Computer Programming, Vol. 3: Sorting and Searching*, Addison-Wesley, Reading, MA, 1973.

[15] A. G. Konheim and D. J. Newman, A note on growing binary trees, *Discrete Math.*, **4**, 57–63, (1973).

[16] P. A. Larson, Dynamic hashing, *BIT*, **18**, 184–201 (1978).

[17] G. Louchard, Exact and asymptotic distributions in digital and binary search trees, *RAIRO Theoretical Inform. Applications*, **21**(4), 479–495 (1987).

[18] H. Mahmoud, *Evolution of Random Search Trees*, Wiley, New York, 1992.

[19] H. M. Mahmoud and B. Pittel, Analysis of the space of search trees under the random insertion algorithm, *J. Algorithms*, **10**, 52–75, (1989).

[20] A. M. Odlyzko, Periodic oscillations of coefficients of power series that satisfy functional equations. *Adv. Math.*, **44**, 180–205, (1982).

[21] R. Sedgewick, *Algorithms*, second ed., Addison-Wesley, Reading, MA, 1988.

[22] I. N. Sneddon, *The Use of Integral Transforms*, McGraw-Hill, New York, 1972.

[23] J. S. Vitter and P. Flajolet, Analysis of algorithms and data structures, in *Handbook of Theoretical Computer Science*, J. van Leeuwen, Ed., *Vol. A: Algorithms and Complexity*. North Holland, Amsterdam, The Netherlands, 1990, Chap. 9, pp. 431–524.

[24] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, fourth ed., Cambridge University Press, Cambridge, England, 1927. Reprinted 1973.

Received May 14, 1991