On the Number of Full Levels in Tries

C. Knessl† W. Szpankowski‡

May 31, 2003

*This research is supported by NSF, NSA and NIH.
†Dept. Mathematics, Statistics & CS, University of Illinois at Chicago
‡Department of Computer Science, Purdue University
Outline of the Talk

1. Tries and Their Parameters
2. Some “Multiplicative” Recurrences
   - Fill-up Level
   - Shortest Path
   - Height
3. Exact Solution of the recurrence
4. Main (Asymptotic) Results
   - Symmetric Case
   - Asymmetric Case
5. Sketch of the Proof
   - Saddle Point Method
   - Depoissonization
   - Asymptotic Evaluation of an Integral
Figure 1: A trie and its parameters.

$F_n$ – fill up level;
$R_n$ – shortest path;
$H_n$ – height.

**Memoryless Source**: Binary sequences generated by i.i.d. source with $p$ being the probability of generating a “0”, and $q = 1 - p$. 
Recurrence for the Fillup Level

**Fillup Level** $F_n$: Define $f(n, k) = \Pr[F_n \geq k]$. Then

$$f(n, k + 1) = \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} f(i, k) f(n - i, k)$$

with $f(0, 0) = 0$ and $f(n, 0) = 1$ for $n \geq 1$.

We also have $f(n, k) = 0$ for $n < 2^k$, since we need at least $2^k$ strings to completely fill $k$ levels.

The above recurrence follows from the fact that with probability $\binom{n}{i} p^i q^{n-i}$, $i$ strings end up in the left subtree while $n - i$ move to the right subtree. In order to have the fill-up level of the whole tree equal at least to $k + 1$ we must assure that the fill-up levels in both subtrees are at least equal to $k$. 
**Shortest Path** $R_n$: Define $r(n, k) = \Pr[R_n \geq k]$. Then

$$r(n, k + 1) = \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} r(i, k) r(n - i, k) - \delta_{n,1} \delta_{k,0}$$

with $r(n, 0) = 1$ for $n \geq 0$ where $\delta_{i,j}$ is the Kronecker delta (i.e., $\delta_{i,j} = 1$ for $i = j$ and zero otherwise).

**Height in Tries** $H^T_n$: The distribution $\bar{h}^k_n = \Pr\{H^T_n \leq k\}$ of the height of $b$-tries satisfies

$$\bar{h}^{k+1}_n = \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} \bar{h}^k_i \bar{h}^k_{n-i}, \quad k \geq 0$$

with the initial condition $\bar{h}^0_n = 1$ for $n = 0, 1, 2, \ldots, b$ and $\bar{h}^0_n = 0$ for $n > b$.

**PATRICIA tries** $H^P_n$: the distribution $h^k_n = \Pr\{H^P_n \leq k\}$ satisfies

$$h^{k+1}_n = (p^n + q^n) h^{k+1}_n + \sum_{i=1}^{n-1} \binom{n}{i} p^i q^{n-i} h^k_i h^k_{n-i}, \quad k \geq 0$$

with the initial conditions $h^0_0 = h^0_1 = 1$ and $h^0_n = 0$ for $n \geq 2$. 
Fillup – Some Explicit Solutions

Small Values of $k$:

$$f(n, 1) = 1 - p^n - q^n, \; n \geq 1,$$
$$f(n, 2) = 1 - (1 - p^2)^n - (1 - q^2)^n - 2(1 - pq)^n + 2(p^n + q^n) + (p^2 + q^2)^n + 2^n(pq)^n - (p^n + q^n)^2, \; n \geq 1.$$  

Large Values of $n = 2^k + l$:

$$f(2^k + 1, k) = (2^k + 1)! \frac{1}{2} (\sqrt{pq})^{k2^k},$$
$$f(2^k + 2, k) = (2^k + 2)! (\sqrt{pq})^{k2^k} \left[ \frac{1}{8} + \frac{1}{24} (p^2 + q^2)^k \right].$$
The Poisson transform is $\tilde{F}_k(z) := F_k(z)e^{-z}$.

From the recurrence we find

$$F_{k+1}(z) = F_k(pz)F_k(qz)$$

with $F_0(z) = e^z - 1$. Let $\tilde{F}_k(z) = \exp[H_k(z)]$. Then

$$H_{k+1}(z) = H_k(pz) + H_k(qz)$$

with $H_0(z) = \log(1 - e^{-z})$.

Define $M_k(s) = M[H_k(z); s]$ the Mellin transform of $H_k(z)$. Then

$$M_{k+1}(s) = (p^{-s} + q^{-s})M_k(s)$$

and

$$M_0(s) = \int_0^\infty z^{s-1}\log(1 - e^{-z})dz = -\Gamma(s)\zeta(s + 1).$$

Therefore, $M_k(s) = -\Gamma(s)\zeta(s + 1)(p^{-s} + q^{-s})^k$. 

Theorem 1. The distribution $f(n, k)$ has the representation

$$f(n, k) = n! [z^n] \exp [H_k(z) + z],$$

so that by Cauchy’s formula,

$$f(n, k) = \frac{n!}{2\pi i} \oint_{\gamma} \frac{e^z}{z^{n+1}} \exp[H_k(z)] \, dz.$$

Here the Mellin transform of $H_k(z)$ is

$$\mathcal{M}[H_k(z); s] = -\Gamma(s) \zeta(s + 1) (p^{-s} + q^{-s})^k, \quad \Re(s) > 0$$

and by inversion we have

$$H_k(z) = -\frac{1}{2\pi i} \oint_{\gamma} \frac{1}{z} \Gamma(s) \zeta(s + 1) (p^{-s} + q^{-s})^k \, ds.$$

Here $\Gamma(\cdot)$ is the Gamma function and $\zeta(\cdot)$ is the Riemann zeta function.

In the symmetric case $p = q = 1/2$, we have

$$f(n, k) = \frac{n!}{2\pi i} \oint_{\gamma} \frac{e^z}{z^{n+1}} \left[ 1 - \exp \left( -\frac{z}{2^k} \right) \right]^{2^k} \, dz.$$
Main Result - Symmetric Case

**Theorem 2.** Consider $p = q = 1/2$ as $k, n \to \infty$.

(a) $k, n \to \infty$ with $n = 2^k + \ell, \ell = O(1), \ell \geq 0$

\[ f(n, k) \sim \sqrt{2\pi e^{-2^k}} 2^{k(\ell+1/2)} \frac{1}{2^\ell \ell!} \sim \sqrt{2\pi n e^{-n}} n^n \frac{e^{\ell}}{2^\ell \ell!}. \]

(b) $k, n \to \infty$ with $n2^{-k} \in (1, \infty)$

\[ f(n, k) \sim \sqrt{\frac{n}{2^k}} \left( \frac{n}{2^k} \right)^n e^{-n} \zeta^{-n}(e^{\zeta} - 1)^{2^k} \frac{e^{\zeta/2} - e^{-\zeta/2}}{\sqrt{\zeta} \sqrt{e^{\zeta} - 1 - \zeta}} \]

\[ = \sqrt{\frac{e^{\zeta} - 1}{e^{\zeta} - 1 - \zeta}} \left( \frac{e^{\zeta}}{e^{\zeta} - 1} \right)^n e^{-n}(e^{\zeta} - 1)^{2^k}. \]

Here $\zeta_0$ is defined implicitly as the solution to

\[ \frac{\zeta_0 e^{\zeta_0}}{e^{\zeta_0} - 1} = n2^{-k}, \quad 0 < \zeta_0 < \infty \]

which satisfies

\[ \zeta_0 \sim 2(n2^{-k} - 1), \quad n2^{-k} \to 1 \]

\[ \zeta_0 \sim n2^{-k}[1 - \exp(-n2^{-k})], \quad n2^{-k} \to \infty. \]
(c) $n \to \infty$ with $k = O(1)$

$$1 - f(n, k) \sim 2^k (1 - 2^{-k})^n.$$
Probability Accumulation

We wish to compute those values of $k$ where $f$ undergoes the transition from $f \approx 1$ to $f \approx 0$. This occurs in the matching region between cases (b) and (c) above.

**Corollary 1.** For $p = q = 1/2$ we define $L$ to be an integer such that

$$k = \lfloor \log_2 n - \log_2(\log n) \rfloor + L = \log_2 n - \log_2(\log n) - \beta + L$$

where $\beta = \beta(n) = \langle \log_2 n - \log_2(\log n) \rangle$ with $\langle x \rangle = x - \lfloor x \rfloor$ denoting the fractional part of $x$. Then

$$\Pr[F_n \geq k] \approx \exp \left( -\frac{n}{\log n} 2^{L-\beta} n^{-2\beta-L} \right).$$

as $n \to \infty$.

We note that for a fixed $\beta \in (0, 1)$ and $L = 1$, $f(n, k)$ is exponentially small for $n \to \infty$. When $\beta \in (0, 1)$ and $L = 0$ we have $f(n, k) \sim 1$. This means that if we let $n \to \infty$ along integer subsequences that have $\beta = \beta(n)$ bounded away from 0 and 1, $\Pr[F_n = \lfloor \log_2(n/\log n) \rfloor] \sim 1$, i.e., all the probability mass accumulates at a single point.
Figure 2: The probability $\Pr[F_n \geq \log_2(n/\log(n)) + x]$ and their corresponding lower and upper envelopes for tries with $n = 1000$ (left) and $n = 10000000$ (right).
We can compute the mean as follows

\[
\mathbb{E}[F_n] = \log_2 n - \log_2(\log n) + \Delta(n) + O(n^{-1}),
\]

\[
\Delta(n) = -\beta + \exp \left[ -\frac{n}{\log n} 2^{-\beta} n^{-2\beta} \right] + \exp \left[ -\frac{n}{\log n} 2^{1-\beta} n^{-2\beta-1} \right].
\]

Figure 3: \(\Delta(n)\) as a function of \(n\).
Asymmetric Case

**Theorem 3.** For \( p \in (\frac{1}{2}, 1) \) with \( r = p/q \):

(a) \( k, n \to \infty \) with \( n = 2^k + \ell, \ell = O(1), \ell \geq 0 \)

\[
f(n, k) \sim \sqrt{2\pi} e^{-2^k} 2^{k(\ell+1/2)} (2\sqrt{pq})^{k^2} \frac{1}{2^\ell \ell!} \sim \frac{n!}{2^\ell \ell!} (\sqrt{pq})^{k^2}.
\]

(b) \( n - 2^k = (p-q^{-q})^k e^{b\sqrt{k}}, b = O(1) \)

\[
f(n, k) \sim \frac{n!}{(n - 2^k)^{n-2^k}} e^{n-2^k} \sqrt{2\pi} \left( \frac{\sqrt{pq}}{\sqrt{n-2^k}} \right)^{k^2} [I(b)]^{n-2^k},
\]

\[
I(b) = 1 - \frac{1}{2\sqrt{2\pi}} \int_{b'}^{\infty} e^{-u^2/2} du + O \left( \frac{1}{\sqrt{k}} \right), \quad b' = \frac{b}{\sqrt{p/q} \log(p/q)}.
\]

(c) \( n - 2^k = B^k, B \in (p-q^{-q}, 1/\sqrt{pq}) \)

\[
f(n, k) = \frac{n!}{(n - 2^k)!} (\sqrt{pq})^{k^2} \exp \left( -\Lambda_1(n, k)[1 + O(k^{-1})] \right),
\]

(d) \( n = (pq)^{-k/2} e^{a\sqrt{k}}, a = O(1), a' = 2a / \log r \)

\[
f(n, k) = \exp \left( 2^{k-1} \sqrt{k} (\log r) \frac{1}{\sqrt{2\pi}} \left[ a' \int_{a'}^{\infty} e^{-u^2/2} du - e^{-(a')^2/2} \right] [1 + O(k^{-1})] \right).
\]
(e) \( n = A^k, A \in (1/\sqrt{pq}, 1/q) \)

\[
\begin{align*}
    f(n, k) &= \exp \left( -\Lambda(n, k)[1 + O(k^{-1})] \right), \\
    \Lambda &= (p^{-S_*} + q^{-S_*})^k n^{-S_*} \frac{\Gamma(S_*) \zeta(S_* + 1) (r^{S_* / 2} + r^{-S_* / 2})}{(\log r)^{\sqrt{2\pi k}}}, \\
    S_* &= S_*(A) = -\frac{1}{\log r} \log \left[ \frac{-\log(qA)}{\log(pA)} \right].
\end{align*}
\]

(f) \( nq^k > 0 \)

\[
f(n, k) = \exp \left( -\frac{1}{\sqrt{1 + k(\log^2 r) \bar{S} r^{-\bar{S}}}} \exp \left[ -\bar{S} + \left( \bar{S} + \frac{1}{\log r} \right) \log \left( \frac{\bar{S}}{nq^k} \right) \right] \right) [1 + O(\bar{S}^{-})].
\]

where \( \bar{S} = \bar{S}(n, k) \) is defined implicitly by

\[
\bar{S} = nq^k \exp[k(\log r)r^{-\bar{S}}].
\]

(g) \( k = O(1) \)

\[
1 - f(n, k) \sim (1 - q^k)^n.
\]
Corollary 2. Define \( L \) as an integer such that

\[
k = \frac{1}{\log(1/q)} \left[ \log n - \log \log \log n + \log(e \log r) \right] - \beta_* + L
\]

where \( \beta_* \) denotes the fractional part

\[
\beta_* = \beta_*(n) = \left\langle \frac{1}{\log(1/q)} \left[ \log n - \log \log \log n + \log(e \log r) \right] \right\rangle.
\]

Then for \( L > L_c = L_c = \left\lfloor \beta_* + \frac{1}{\log q} \right\rfloor \)

\[
\Pr[F_n \geq k] \sim \exp \left( - \frac{1}{\sqrt{\log \log n}} \frac{1}{\sqrt{1 + (\beta_* - L) \log q}} \frac{e \log r}{\log(1/q)[(\beta_* - L) \log q + 1]} \left( \log n \right)^{\frac{(\beta_* - L) \log q}{\log r}} \right)^{\frac{1+(\beta_* - L) \log q}{\log r}}
\]

and \( 1 - \Pr[F_n \geq k] \approx \left( \frac{1}{\log n} \right)^{\frac{qL - \beta_* e^{-1}}{\log r}} \), \( L \leq L_c \)
The integer $L_c$ in Corollary above can only take on the three values $-2$, $-1$, and $0$. We can show that the probability mass concentrates at $L = 0$, i.e.,

$$\Pr[F_n = \log_{1/q}(en \log r / \log \log n)] \sim 1$$

except for some special subsequences of $n$ that lead to mass at two consecutive points.

Finally, Corollary 2 suggests the following formula for the mean $\mathbb{E}[F_n]$

$$\mathbb{E}[F_n] = \log_{1/q} n - \log \log \log n + O(1).$$

Observe that in the symmetric case the second term is $\log \log n$. 
Sketch of Proof: Methods

The proof is based on asymptotic evaluation of some integrals presented in Theorem 1. We use saddle point method and depoissonziation methods.

**Saddle Point Method.** Throughout the proof we need to evaluate integrals of the following form

\[
J(\lambda) = \int_{-\infty}^{\infty} \exp[\lambda f(x) + \lambda^a g(x)] h(x) \, dx,
\]

where \( a \in (0, 1) \) for \( \lambda \to +\infty \). The saddle point \( x_0 \) satisfies

\[
f'(x_0) + \lambda^{a-1} g'(x_0) = 0
\]

and then

\[
J(\lambda) = e^{\lambda f(x_0)} e^{\lambda^a g(x_0)} h(x_0) \sqrt{\frac{2\pi}{\lambda |f''(x_0)|}} \left( 1 + O(\lambda^{a-1}) \right).
\]
Exponential Depoissonization

We consider a Cauchy integral in the form

\[ f(n) = \frac{n!}{2\pi i} \oint \frac{e^z}{z^{n+1}} \tilde{F}(z) dz. \]

where \( \tilde{F}(z) \) is the Poisson transform of \( f(n) \), that is,

\[ \tilde{F}(z) = e^{-z} \sum_{n \geq 0} f(n) \frac{z^n}{n!}. \]

**Theorem 4 (Jacquet and Szpankowski, 1998).** Let \( S_\theta = \{z : \arg(z) \leq \theta, \ |\theta| < \pi/2\} \) be a linear cone around the real axis.

(i) For \( z \to \infty \) the following holds for \( A, B > 0 \):

(I) For \( z \in S_\theta \) \( |\tilde{F}(z)| \leq A \exp(B|z|^\nu) \)

where \( 0 \leq \nu < 1/2 \).

(O) For \( z \notin S_\theta \) \( |\tilde{F}(z)e^z| \leq A_1 \exp(\omega|z|) \)

for \( \omega < 1 \) and \( A_1 > 0 \). Then for \( n \to \infty \)

\[ f(n) = \tilde{F}(n) + O \left(n^{-\nu} \exp(Bn^\nu)\right). \]

(ii) Let conditions (I) and (O) above hold again except that now \( \nu \) in (I) satisfies \( \frac{1}{2} < \nu < 1 \). Then

\[ \log f(n) = \log \tilde{F}(n) + O(n^{2\nu-1}). \]
Evaluation of $H_k(z)$

To establish Theorem 3 we need asymptotic evaluation of

$$H_k(z) = -\frac{1}{2\pi i} \int_{\frac{1}{2}-i\infty}^{\frac{1}{2}+i\infty} z^{-s} \Gamma(s) \zeta(s+1)(p^{-s} + q^{-s})^k ds.$$  

as $z \to \infty$.

We note that the integrand has a double pole at $s = 0$ and simple poles at $s = -1, -2, -3, \ldots$ and a saddle-point at

$$s = S_* \equiv \frac{-1}{\log(p/q)} \log \left( \frac{-\log z - k \log q}{\log z + k \log p} \right).$$

1. The saddle lies on the real axis for $p^{-k} < z < q^{-k}$ (i.e., $k/\log z \in (-1/\log q, -1/\log p)$).
2. It coalesces with the pole at $s = 0$ when $z = z_0 \equiv (pq)^{-k/2}$ and with the pole $s = -1$ when $z = z_1 \equiv (p^{-p}q^{-q})^k$.
3. We also note that $S_* \to +\infty$ as $z \uparrow q^{-k}$ and $S_* \to -\infty$ as $z \downarrow p^{-k}$. 
Case A: \((pq)^{-k/2} < z < q^{-k}\)

For \((pq)^{-k/2} < z < q^{-k}\) we have \(S_* \in (0, \infty)\) and then the asymptotics of \(H_k(z)\) follow by shifting the contour leading to

\[
H_k(z) \sim H_{\text{saddle}}(k, z)
\]

\[
\equiv -\frac{1}{\sqrt{2\pi \Delta_0}} \Gamma(S_*) \zeta(S_* + 1) z^{-S_*} (p^{-S_*} + q^{-S_*})^k.
\]
Case B: \((p^{-}q^{-})^{k} < z < (\sqrt{pq})^{-k}\)

For \((p^{-}q^{-})^{k} < z < (\sqrt{pq})^{-k}\), the saddle satisfies \(-1 < S_{*} < 0\) and hence the contribution of the pole at \(s = 0\) must be considered. Then

\[
H_{k}(z) \sim 2^{k} \log[z(\sqrt{pq})^{k}] + H^{\text{saddle}}(k, z).
\]

Case C: \(z < (p^{-}q^{-})^{k}\)

For \(z < (p^{-}q^{-})^{k}\) the pole at \(s = -1\) also contributes and we obtain

\[
H_{k}(z) \sim 2^{k} \log[z(\sqrt{pq})^{k}] - \frac{z}{2} + H^{\text{saddle}}(k, z).
\]
We consider transition ranges where the saddle lies close to one of the poles. For \( \frac{(\log z)}{k} \approx -p \log p - q \log q \) we introduce \( w \) via the scaling
\[
z = \left( p^{-q} q^{-q} \right)^k e^{w \sqrt{k}},
\]
expand the integrand about \( s = -1 \), and set \( s + 1 = T/\sqrt{k} \) with \( T = O(1) \). We have
\[
-s \log z + k \log(p^{-s} + q^{-s}) = \log z - w \sqrt{k}(s + 1) + \frac{1}{2}qpk(\log r)^2(s + 1)^2 + O(k(s)
\]
and \( \Gamma(s) \zeta(s + 1) \sim -\frac{1}{s+1} \zeta(0) = \frac{1}{2(s+1)} \). Upon shifting the line of integration to the range \( \Re(s) \in (-1, 0) \),
\[
H_k(z) \sim 2^k \log[(\sqrt{pq})^k] - \frac{z}{2} \frac{1}{2\pi i} \int_{B_{r+}} e^{-wT} e^{q(\log r)^2T^2/2} dT
\]
\[
= 2^k \log[z(\sqrt{pq})^k] - \frac{z}{2} \frac{1}{\sqrt{2\pi}} \int_{w'} \frac{1}{\sqrt{-u/2}} e^{-u^2/2} du [1 + O(k^{-1})],
\]
where \( w' = \frac{w}{\sqrt{pq} \log r} \).
Case E: Saddle Close to the Double Pole $s = 0$

The saddle is close to the double pole at $s = 0$. For $s \to 0$ we have

$$\Gamma(s)\zeta(s + 1) = s^{-2} + O(1).$$

We assume that now $(\log z)/k \approx -\frac{1}{2} \log(pq)$ and introduce the scaling

$$z = (pq)^{-k/2} e^{\alpha\sqrt{k}} \quad \alpha = O(1).$$

we have

$$H_k(z) = 2^k \sqrt{k} \frac{\log r}{2} \frac{1}{\sqrt{2\pi}} \int_{\alpha'}^{\infty} (\alpha' - v)e^{-v^2/2} dv [1 + O(k^{-1})],$$

where $\alpha' = \frac{2\alpha}{\log r}$. 