ALP, January 2002

Digital Trees and Some of their Avatars

by Philippe Flajolet, and many others!

Journal of the ACM
Volume 6, Issue 2 (April 1959)

Radix Exchange—An Internal Sorting Method for Digital Computers*

PAUL HILDEBRANDT AND HAROLD ISBITZ

System Development Corporation, Santa Monica, California

This note describes a new technique—Radix Exchange. The technique is faster than Inserting by the ratio $(\log_2 n)/n$ Its speed compares favorably with internal merging and it has the significant advantage of requiring essentially no working area...

Communications of the ACM
Volume 3 , Issue 9 (August 1960)

Table of Contents

Techniques

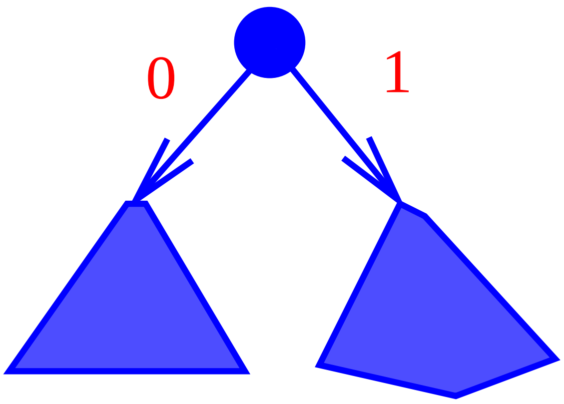
Trie Memory*

EDWARD FREDKIN, *Bolt Beranek and Newman, Inc., Cambridge, Mass.*

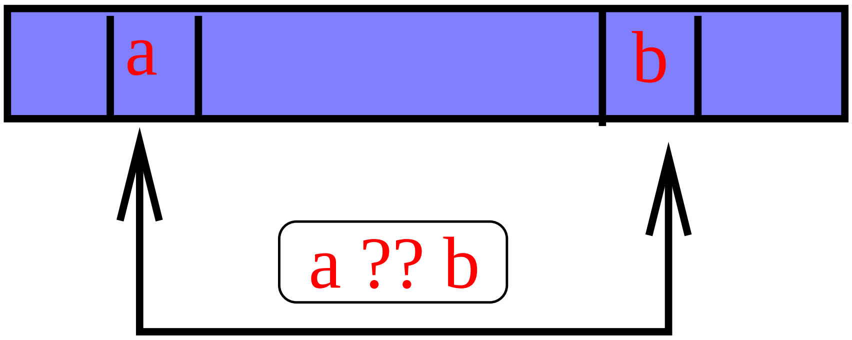
Nexuses and Nexus Chains

In order to permit more general description of trie memory, it is helpful to substitute, for the system of successive addresses used in connection with the illustrations in Fig. 1, a system of directed connections that we may call nexuses. A brief explication of nexuses and nexus chains will facilitate further discussion of trie memory.

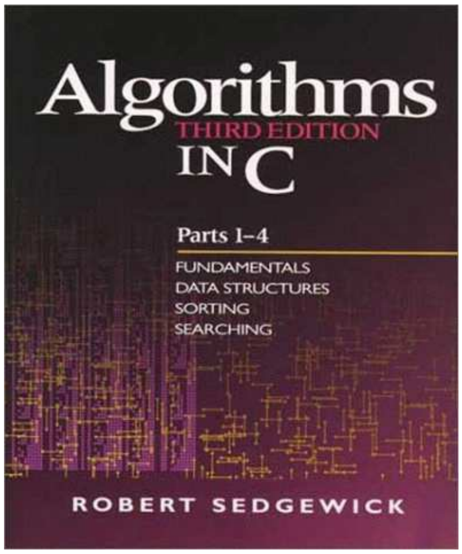
The trie splitting process

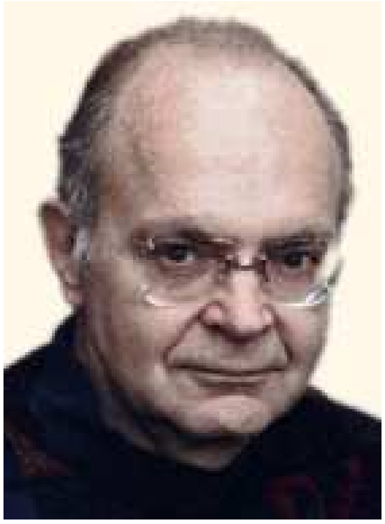


Radix-exchange sort

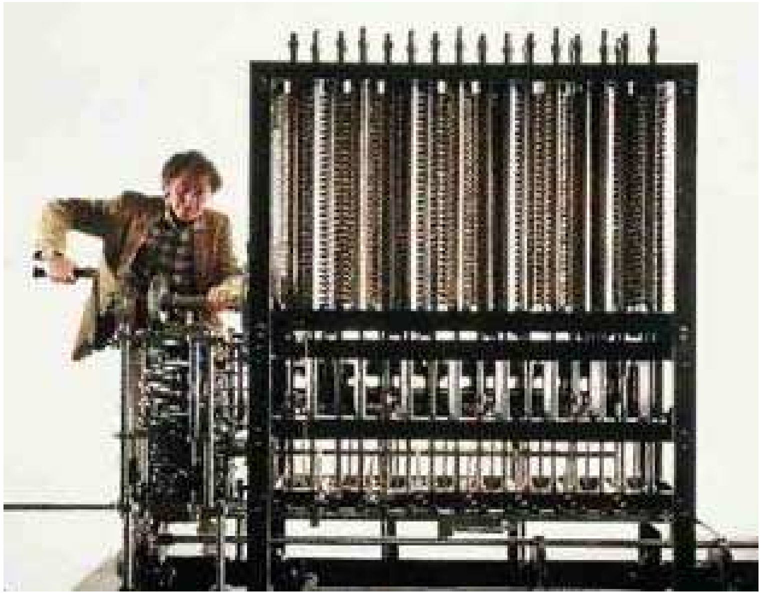


cf Vol 3 & Sedgewick's Algorithms

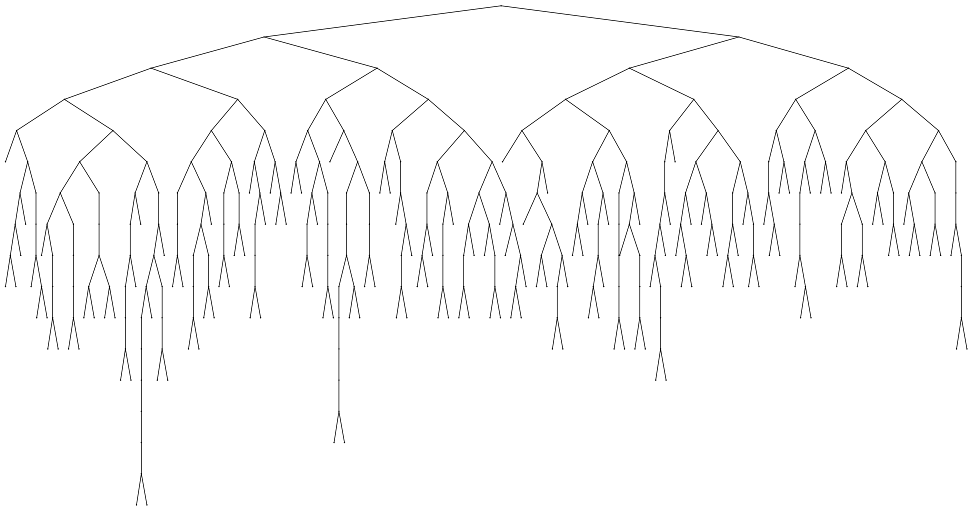


???  ???

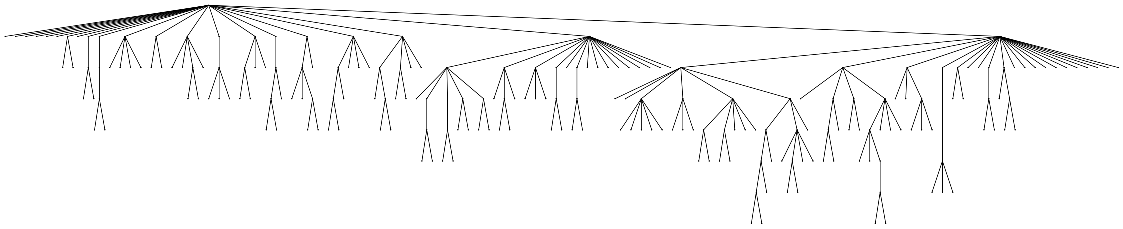
What is the number of turns of the handle?



Binary trie (uniform bits)

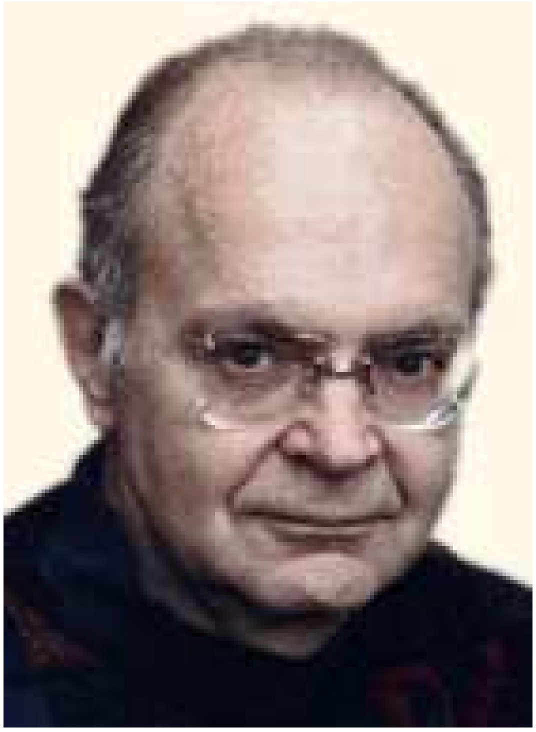


Continued fraction trie by Clément



Weyl trie by Devroye versus ‘beta tree’

𝔆𝔞𝔩𝔦𝔣𝔬𝔯𝔫𝔦𝔞 𝔍𝔫𝔰𝔱𝔦𝔱𝔲𝔱𝔢 𝔬𝔣 𝔗𝔢𝔠𝔥𝔫𝔬𝔩𝔬𝔤𝔶



D.E.K.

N.G. de Bruijn

Page 131 of Volume 3 (1973) — The original derivation

♣ Decompose & get a binomial recurrence

♢ Use binomial inversion to get

$$\rightsquigarrow \sum_k \binom{n}{k} \frac{(-1)^k a_k}{1-2^{-k}}$$

♡ Reorganize to derive

$$\rightsquigarrow \sum_\ell \left(1 - e^{-n/2^\ell}\right)$$

♠ Cleverly use Gamma function

$$e^x = \frac{1}{2i\pi} \int_{-1-i\infty}^{1+i\infty} \Gamma(s)\, ds$$

Miraculous factorizations occur, residues fly all around, and ...

The Big Theorem of P. 131$\pm$

Tries and Radix-exchange sort have expected cost $=$ path length, bit comparisons, resp.

$$\sim \boxed{n \log_2 n} + n\left(\frac{\gamma - 1}{\log 2} - \frac{1}{2} + f(n)\right)$$

where $f(n)$ is the rather strange function

$$f(n) = \frac{1}{\log 2} \sum_{k \neq 0} \Gamma\left(-1 - \frac{2ik\pi}{\log 2}\right) \exp\left(2ik\pi \log_2 n\right)$$

Furthermore

$$f(n) < 0.0000001725$$

thus we may safely ignore $f(n)$ for practical purposes.

Size has expectation

$$\sim \boxed{\frac{n}{\log 2}} + n \cdot \widehat{f}(n)$$

with fluctuations!

Criticisms

- ♠ A complicated math exercise. An isolated problem.
- ♠ An expected outcome:

  $O(n \log n)$ by *easy* probabilistic argument

- ♠ A useless answer with 10^{-6} fluctuations!
- ♠ With Moore's law, anyhow, etc.

Early developments

♡ D.E.K. 1973:
— bucket tries;
— multiway tries;
— non-uniform probabilities

♡ Digital Trees (Coffman-Eve) + Konheim-Newman'73
+ Vol 3:
— digital search trees;
— Patricia tries
= "the hardest asymptotic nut"

♡♡ de Bruijn-Knuth-Rice '72 $\leadsto$ Height of Catalan trees
= another Saga...

♢ Analytic approaches here.

♢ Probabilistic methods by Aldous, Shields, Pittel, ... are not covered, unfortunately.

Methodological advances I

♡ The algebra of costs

— Poisson models, generating functions

"Set Representation and Set Intersection" by Luis Trabb Pardo (1978)

Symbolic methods: *cf* Dan Greene's PhD (1983).

F., Régnier, Sotteau (1985)

$$\begin{array}{rcl}
\mathfrak{P}\{\{0,1\}^{\ell+1}\} & \cong & \mathfrak{P}\{\{0,1\}^{\ell}\} \times \mathfrak{P}\{\{0,1\}^{\ell}\} \\
(1+z)^{2^{\ell+1}} & = & (1+z)^{2^{\ell}} \times (1+z)^{2^{\ell}} \\
e^{z} & = & (e^{z/2}) \times (e^{z/2}).
\end{array}$$

$\Longrightarrow$ Difference equations , e.g., *size*:

$$\begin{array}{rcl}
\Phi(z,u) & = & \Phi\left(u\frac{z}{2},u\right)^2 + \mathsf{toll}\,(z,u) \\
\phi(z) & = & 2e^{z/2}\phi\left(\frac{z}{2}\right) + \mathsf{toll}\,(z).
\end{array}$$

Iteration, coefficient extraction, . . .

Methodological advances II

♡ Mellin transforms

$$f(x) \quad \xrightarrow{\mathcal{M}} \quad f^\star(s) := \int_0^\infty f(x) x^{s-1}\, dx.$$

$\mathbf{M}_1$: Asympt. of f $\mapsto$ Singularities of $f^\star$

$\mathbf{M}_2$: Harmonic sums $\mapsto$ factored form

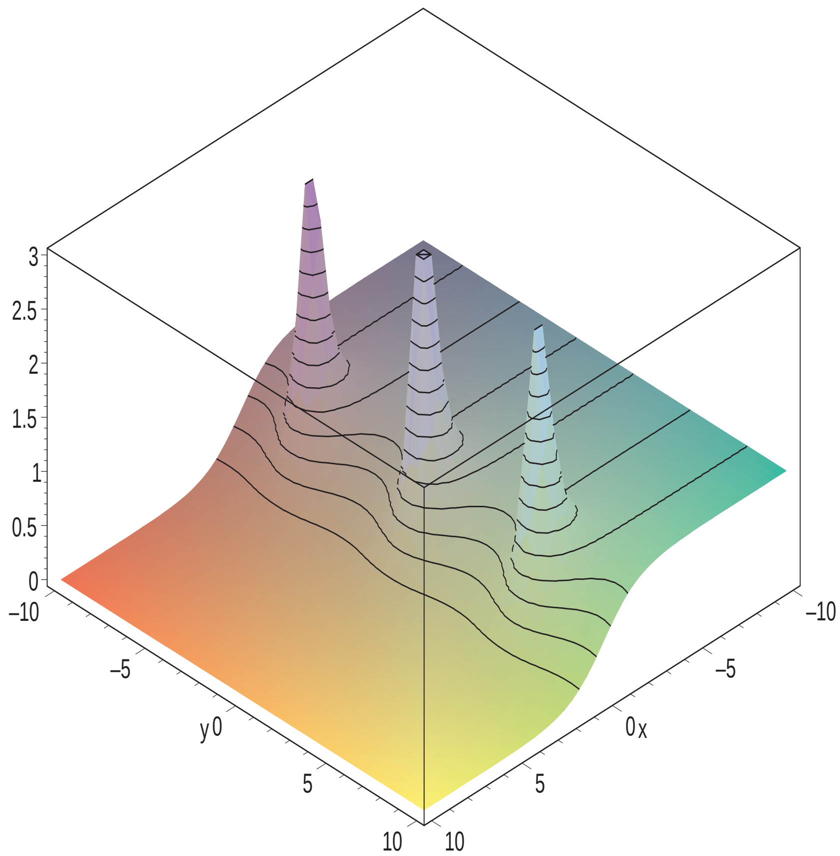
$$\sum \lambda_k f(\mu_k x) \quad \mapsto \left(\sum \lambda_k \mu_k^{-s}\right) \cdot f^\star(s).$$

Guibas+Ramshaw+Sedgewick (unpub)

F+Régnier+Sedgewick (1985) +Dumas+Gourdon (1995)

Difference Eq., Harm. Sum $\mapsto$ factored form $\mapsto$ Asympt.

$$\frac{\Gamma(s)}{1-2^{-s}}$$



Work till 1973 has led to a **systematic approach**

Trie-like algorithms

$\Downarrow$

Algebra of Costs

$\Downarrow$

Asymptotic estimates

corresponding to a major combinatorial process of computer science.

Tries meet data bases

Adaptive hashing schemes

Near 1977-78, several groups discover the virtues of dynamic hashing.

Idea: *Split buckets instead of chaining them.*
Larson; Fagin-Nievergelt-Pippenger-Strong; Litwin
as it turned out $\equiv$ bucket tries of vol. 3.

Expected size of tree is

$$\frac{n}{b \log 2} + fluctu$$

corresponding to

$$69\% \quad \text{filling ratio.}$$

Compare with similar ratio for B–trees (Yao)

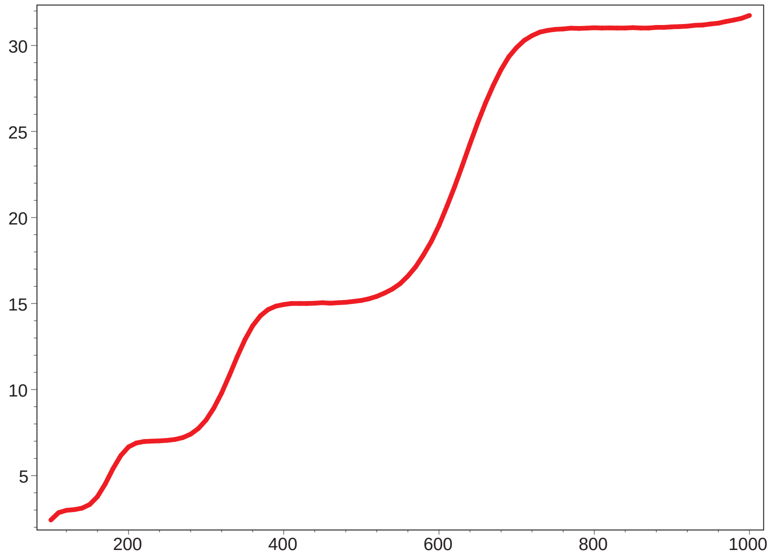
Extendible Hashing transforms the index into a perfect tree $\equiv$ array that can be paginated.

H=height   Index size $\equiv 2^H$

Yao, Régnier, Flajolet, ca 1980.

$$\mathbb{E}(H) \sim \left(1 + \frac{1}{b}\right) \log_2 n; \qquad \mathbb{E}(2^H) \approx 4n^{1+1/b}.$$

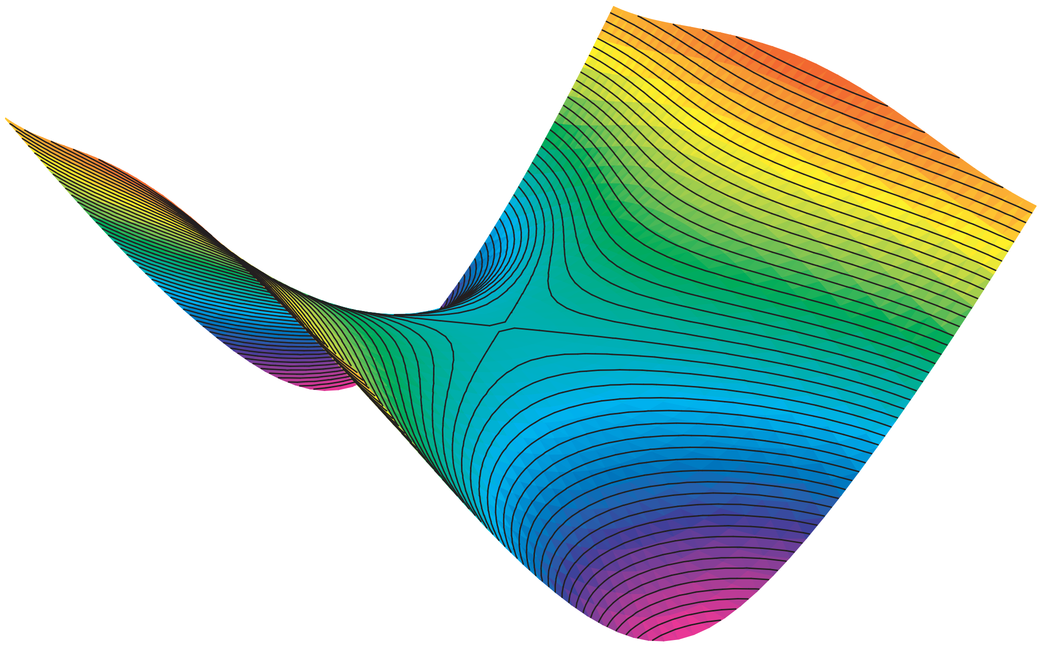
$(b = 50)$



+ Devroye: density models $\int f^2 < \infty$.

Height: One of the very first intrusions of saddle point method in Analysis of Algorithms.

$$[z^n]f(z) = \frac{1}{2i\pi} \oint f(z)\,\frac{dz}{z^{n+1}}.$$



$\leadsto$ Jacquet & Szpankowski's "analytic de-Poissonization" 1995-2000.

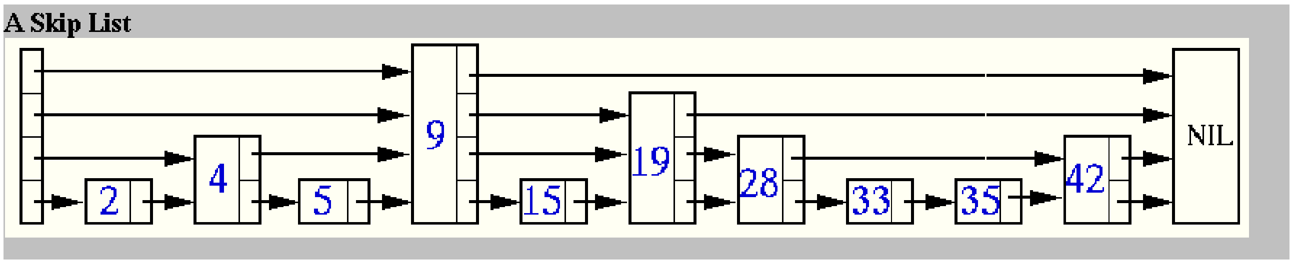
Skip lists

From VSAM's to skip lists

Idea 1 (old): build indexes of indexes of indexes . . .

Idea 2: balance $\leadsto$ B–trees

Idea 2': randomize! $\equiv$ Pugh's skip lists



Much easier to maintain than balanced structures!

Analysis by Papadakis + Munro, Poblete

Kirschenhofer, Martínez, Prodinger entirely based on trie technology.

Probabilistic counting algorithms

Can you estimate to 5% the number of different words in Shakespeare given a pencil and one sheet of paper?

Yes. F.+Martin (1985) for data base query optimization.

Ideas: hash to get uniformity; observe bit patterns.

$\mathtt{0}\ldots = 50\%$ of times; $\mathtt{10}\ldots = 25\%$; $\mathtt{110}\ldots = 12.5\%$

Try 2^K where $\overbrace{11\cdots1}^{K}0\ldots$ is longest initial run of 1's.

The best observable known is trie-like and has accuracy

$$\frac{\boxed{0.78}}{\sqrt{m}} \qquad \text{for } m \text{ words of memory}$$

(+"stochastic averaging"); $\boxed{0.78}$ is a Mellin constant.

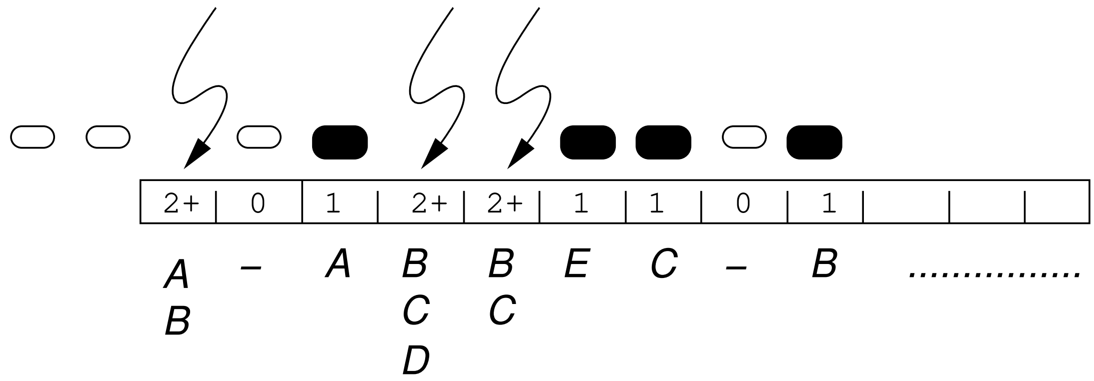
Works in distributed environment:

Yellow pages of New York $\cup$ San Francisco by phone line!

Data mining. Cf Broder.

Tries meet communication protocols

♡ 1970: the shared communication channel

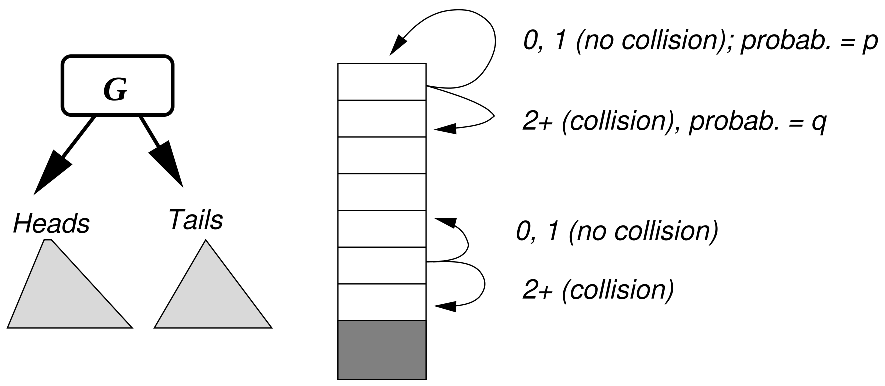


Ethernet: Try; wait $\times 1$, $\times 2$, $\times 4$, etc

⇝ Aldous 1987: *Ethernet is unstable!*

♡ 1977: The **Tree/Stack protocol**

CTM = Capetanakis, Tsybakov, Mikhailov



= A digital trie but with a flow of arrivals!

♠♠ Erroneous analyses missed the wobbles. Variance by Kirschenhofer, Prodinger et al.= Mellin + modular forms.

Tree protocol $\Longrightarrow$ Poisson GenFun solves $(p+q=1)$

$$\psi(z) - \psi(\lambda + pz) - \psi(\lambda + qz) = \mathsf{toll}\,(z).$$

A non-commutative iteration semigroup with a globally invariant measure.

Theorem. Stable till $\lambda_{\max} = 0.36017$ root of:

$$\begin{aligned} -\frac{1}{2} &= \frac{e^{-2x}}{1-2x} \sum_{i \geq 0} 2^i g\left(\frac{x}{2^i}\right), \\ g(y) &:= e^{-2y}\left((e^{-y}(1-y) - 1 + 2y(1+y))\right). \end{aligned}$$

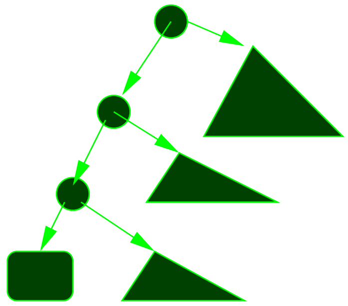
Analyses by Fayolle, F., Hofri, Jacquet, Mathys $\Longrightarrow$
— Ternary tree algorithms gives 10% better throughput
— Protocol is hyperstable at all arrival rates.

The IEEE 802.14 norm. . . a *failed* success story!

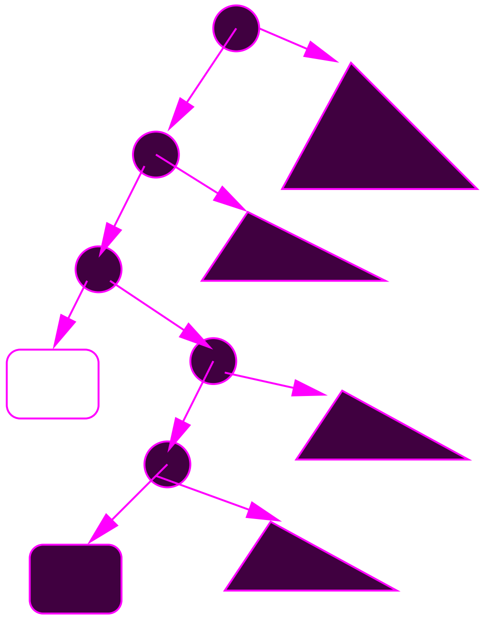
Also analyses by Greenberg+F+Ladner: tree protocol modified to attain 93% of optimal: $\lambda_{\max} = 0.4672$.

Leader Election:

(i) The leftmost branch of a trie



(ii) The leftmost border of a trie

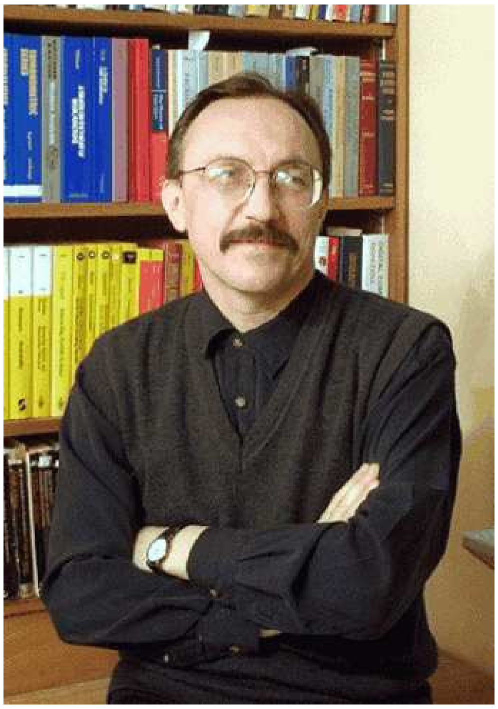


Analyses by Fill, Mahmoud, Szpankowski, Prodinger, F+Sedgewick; includes distributions.

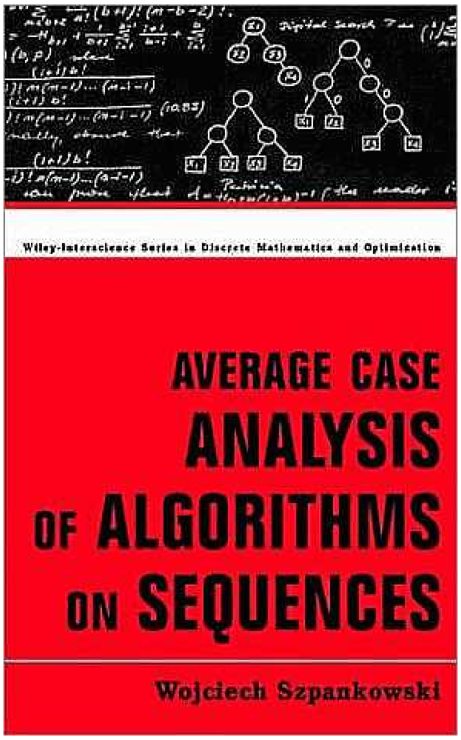
$(2\log n)$ rounds; $\log_2 n$ rounds.

Tries meet texts again!

Szpankowski's *Analysis of Algorithms on Sequences*.



$\leadsto$



Random text: `kwnbpr hwnqqcpq yt nxgfhsd agghos fhskla zmmxnz kasiweyzkcn ejhjsal ehrdjn...`

$\neq$ "Natural" language text: *Cale Pismo przez Boga jest natchnione i pozyteczne do nauki, do wykrywania bledow, do poprawy, do wychowania w sprawiedliwosci, aby czlowiek Bozy byl doskonaly, do wszelkiego dobrego dziela przygotowany.*

• *Lempel & Ziv* invent LZ *compression* (1977+).
Turn out to be related to *digital search trees*.

a|b|r|ac|ad|ab|ra|abr|aca|d|abra|abrac|ada|br|aa|br|acad|abraa|...

• 'Szpankowsquet' do average-case analysis of tries under *Markovian dependencies* (1990).

• Régnier-Jacquet (1987) do *distributional analysis* of tries under Bernoulli models.

• Jacquet-Szpankowski-Louchard (1995+) extend distributional analysis to DST's:

$$\boxed{\frac{\partial}{\partial z}F(z,u) = F(z,pu)F(z,qu) + \text{fudge}}$$

$\leadsto$ Combines everything:
algebra of trie costs, Mellin, analytic dePoissonization...

♡♡ Complete characterizations of *Lempel-Ziv* algorithms, notably: redundancy.

Tries meet thermodynamics

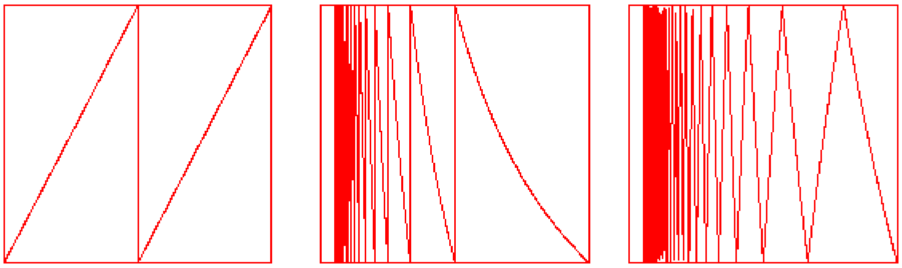
D. Ruelle

B. Vallée

- "thermodynamic formalism" by Ruelle (1970)
- Operators & Euclid's alg. by Babenko, D. Mayer $\approx$1977.
- Related to information theory & tries by Vallée (1998)

Transfer operators

T is a transformation. Iterates?

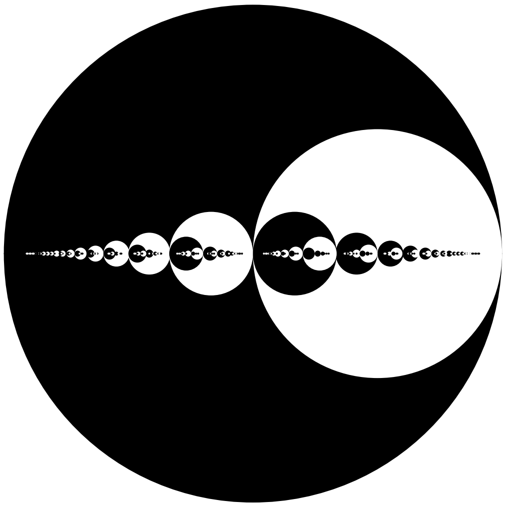


Transfer operator:

$$\mathcal{G}_s[f](x) := \sum_{h \in T^{-1}} (h'(x))^s f \circ h(x).$$

Generalizes probability density transformer ($s = 1$); spectra, functional analysis, etc.

Vallée: serves to generate probabilities of prefixes $\rightsquigarrow$ tries. cf Clément+F+Vallée 2001.



Vallée+Clément+F = Tries under dynamic source models;
$\rightsquigarrow$ Bentley-Sedgewick's Ternary Search Tries

= Entropy for size, depth path-length
= Eigenvalue $\lambda(2)$ for height, etc.

Applies to continued fraction representations & algs:

$\rightsquigarrow$ HAKMEM Algorithm (Gosper, 1972)

$$\frac{36}{113} = \cfrac{1}{3 + \cfrac{1}{7 + \cfrac{1}{\mathbf{5}}}}, \qquad \frac{113}{355} = \cfrac{1}{3 + \cfrac{1}{7 + \cfrac{1}{\mathbf{16}}}}.$$

$\rightsquigarrow$ Sorting with continued fractions, path length:

$$K_0 n \log n + K_1 n + Q(n) + K_2 + o(1),$$

$$K_0 = \frac{6 \log 2}{\pi^2}, \quad K_1 = 18\frac{\gamma \log 2}{\pi^2} + 9\frac{(\log 2)^2}{\pi^2} - 72\frac{\log 2 \zeta'(2)}{\pi^4} - \frac{1}{2}.$$

♡♡♡ Q depends on Riemann hypothesis!!!

Random Trie Encounters

♡ Polynomial factorization (Cantor-Z refinement)

```
> factor(x^13-x^10+x^5-x^2+x^3-1);
              2             6    4         2         2
 (x - 1) (x  - x + 1) (x  - x  + 1) (x  + x + 1)
```

Vol 2., F+Gourdon+Panario

♡ Quadtries and geometry, multiD search

Rivest–Bentley–Samet

♡ Other probabilistic counting algorithms

Morris–Freivalds, Wegner's, etc

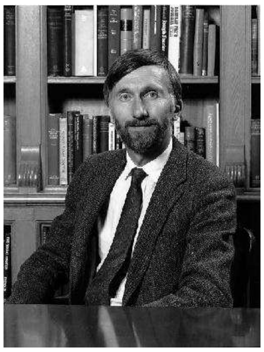
♡ Binary Decision Diagrams (BDD's) by Bryant

= Fully developed tries + common subtree factoring...

♡ Hierarchical data compression by J. Kieffer

♡ Level compressed tries $\Longrightarrow$ fast lookup in routers!

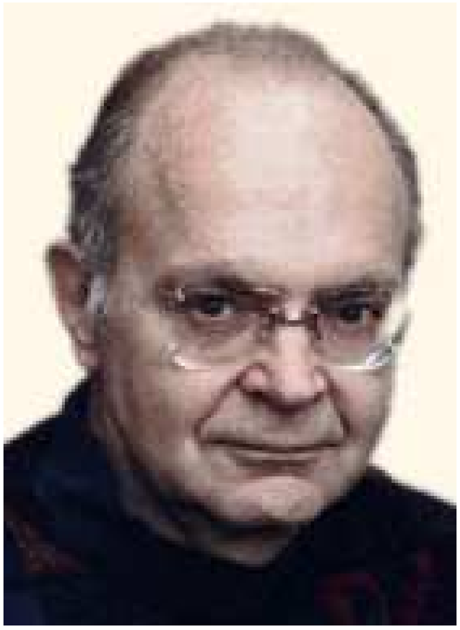
Nilsson et al.

A major combinatorial-algorithmic process of C.S.

$\approx$100 papers

$\rightsquigarrow$ A tribute to Knuth's vision in 1965^{++}

$\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit$  $\heartsuit\heartsuit\heartsuit\heartsuit\heartsuit$