B-trees and Pólya urns

Danièle GARDY

PRiSM (UVSQ)

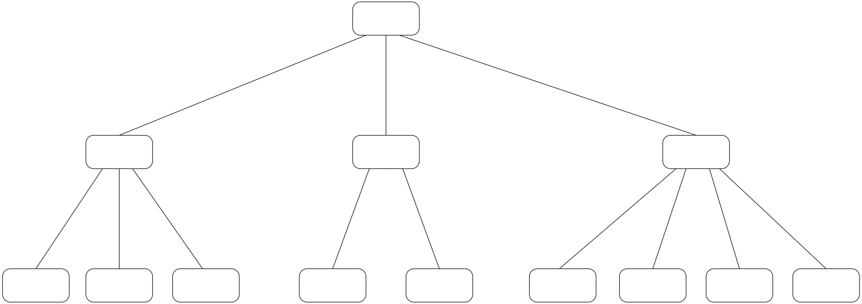
with B. Chauvin and N. Pouyanne (LMV) and D.-H. Ton-That (PRiSM)

AofA, Strobl – June 2015

m integer ≥ 2 : parameter of the B-tree
Database applications : m « large » (several hundreds)

B-tree shape

- planar tree
- root : between 2 and $2m$ children
- other internal nodes : between m and $2m$ children
- nodes without children at same level



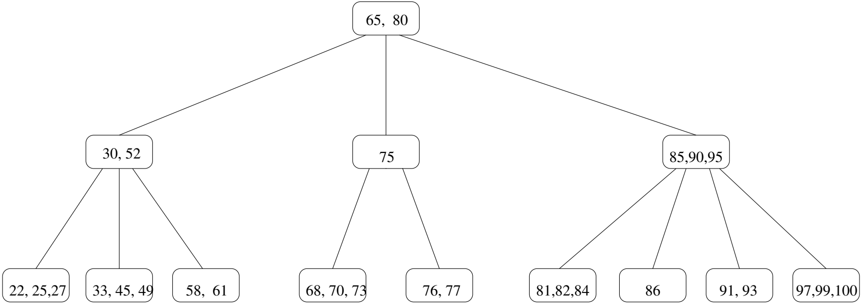
B-tree shape with parameter $m = 2$ and with 13 nodes m integer ≥ 2 : parameter of the B-tree

B-tree

- B-tree shape
- Research tree : nodes contain records (keys) belonging to an ordered set + at each node, the root keys determine the partition of non-root keys into subtrees

Root : between 1 and $2m - 1$ keys
Other nodes : between $m - 1$ and $2m - 1$ keys

All keys distinct : a tree with repeated keys in internal nodes cannot be a B-tree



A B-tree ($m = 2$) : B-tree shape + labelling as a research tree.

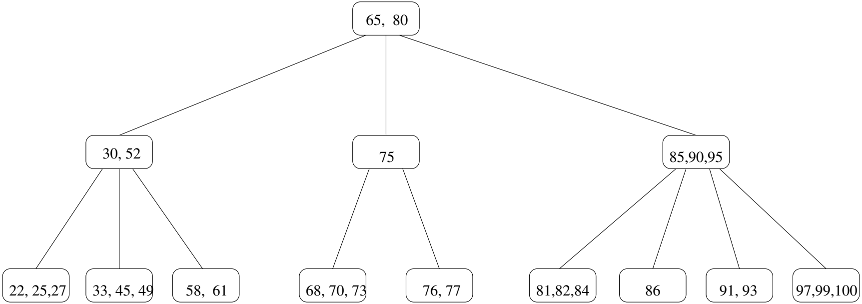
Variations

- Nodes have between m and $2m$ keys (internal nodes : between $m+1$ and $2m+1$ children)
- For such trees and $m = 1$: each node has 1 or 2 keys (internal nodes : 2 or 3 children)

2–3 trees

- Internal nodes may contain just an index, and the actual records are in leaves

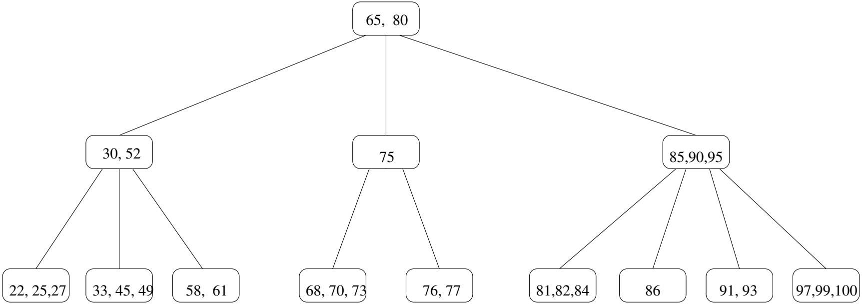
Searching for a key X in a B-tree



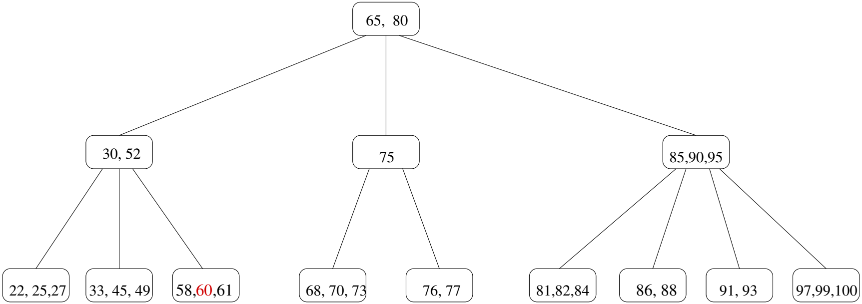
Inserting a key X into a B-tree

- No repeated key
- Insertion of a new key : in a leaf
- Research tree ⇒ a single place in a terminal node to insert X
- B-tree shape ⇒ terminal nodes must be at the same level
- B-tree ⇒ terminal nodes contain between $m-1$ and $2m-1$ keys ; what if the relevant node is already full ?

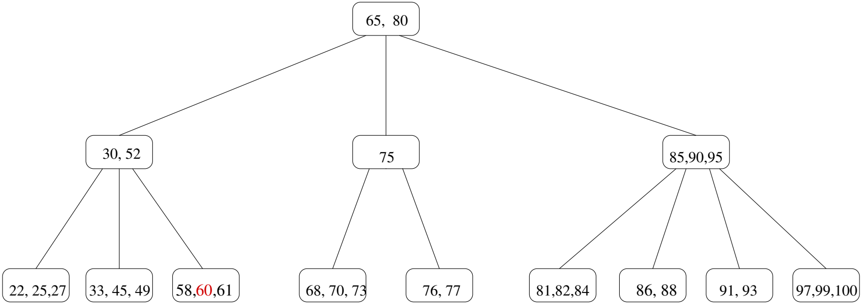
Insertion of 60



Insertion of 60

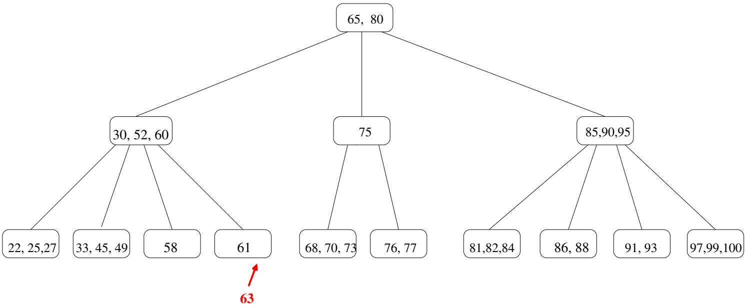


Insertion of 60

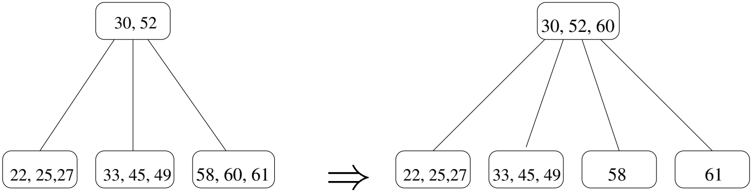


What if we now wish to insert 63 ?

Insertion of 63 ?



An internal node was split :



- A terminal node with maximal number of keys disappears
- 2 terminal nodes with minimal number of keys appear
- Parent node could accomodate one more key

Inserting a key X into a B-tree

- Need to keep the tree balanced $\Rightarrow$ intricate algorithm
- Splitting a node may go all the way up to the root $\Rightarrow$ tree grows from the root

Inserting a key X into a B-tree

- Need to keep the tree balanced $\Rightarrow$ intricate algorithm
- Splitting a node may go all the way up to the root $\Rightarrow$ tree grows from the root

- Analysis much more difficult than for other research trees
- Pólya urn approach useful for lower level

Counting issues for B-trees (shapes) with parameter m

Counting issues for B-trees (shapes) with parameter m

- Relation between height h and number of keys n of a tree

$$\log_{2m}(n+1) \leq h \leq \log_m \frac{n+1}{2} + 1.$$

- Number of trees with n keys
- Number of trees with height h

Number of trees with n keys ?

Proposition (Odlyzko 82)

Define $E(z)$ as the g.f. enumerating 2-3 trees w.r.t. number n of leaves = number n-1 of keys in internal nodes

$$E(z) = z + E(z^2 + z^3).$$

Radius of convergence : golden ratio $\frac{1+\sqrt{5}}{2}$
Number of 2-3 trees with n leaves :

$$e_n \sim \frac{\omega(n)}{n} \left(\frac{1+\sqrt{5}}{2}\right)^n (1 + O(1/n)),$$

$\omega(n)$ *periodic : average 0.71208... and period 0.86792...*

Number of trees with n keys ?

Proposition (Odlyzko 82)

Define $E(z)$ as the g.f. enumerating 2-3 trees w.r.t. number n of leaves = number n-1 of keys in internal nodes

$$E(z) = z + E(z^2 + z^3).$$

Radius of convergence : golden ratio $\frac{1+\sqrt{5}}{2}$
Number of 2-3 trees with n leaves :

$$e_n \sim \frac{\omega(n)}{n} \left(\frac{1+\sqrt{5}}{2}\right)^n (1 + O(1/n)),$$

$\omega(n)$ *periodic : average 0.71208... and period 0.86792...*

Similar result for general B-trees ?

Number of trees with height h ?

Proposition (Reingold 79)

The number a_h of 2-3 trees with height h satisfies the recurrence relation

$$a_{h+1} = {a_h}^2 + {a_h}^3$$

with $a_0 = 2$. It is asymptotically equal to

$$a_h = \kappa^{3^h} \left(1 + O\left(\frac{1}{2^{3^h}}\right)\right) \text{ with } \kappa = 2.30992632\ldots$$

First values ($h \geq 0$) : 2, 12, 1872, 6563711232, ...

Known sequence ?

The On-Line Encyclopedia of Integer Sequences

founded in 1964 by N. J. A. Sloane

2,12,1872 | Search | Hints

(Greetings from The On-Line Encyclopedia of Integer Sequences!)

Search: **seq:2,12,1872**

Displaying 1-1 of 1 result found. page 1

Sort: relevance | references | number | modified | created Format: long | short | data

A125295 Number of different non-self-crossing ways of moving a tower of Hanoi from one peg onto another peg. +20 0

1, **2**, **12**, **1872**, 6563711232, 282779810171805015122254036992, 22612323802416302740572466532905158028496454353087246911545156210129751385945830223511552 (list; graph; refs; listen; history; text; internal format)

OFFSET 0,2

COMMENTS In other words, a sequence of moves starting with all disks on the starting peg, ending with all disks on the destination peg and never more than once producing the same distribution of disks among the pegs (assuming 3 pegs).

LINKS Table of n, a(n) for n=0..6.
Wikipedia, Tower of Hanoi

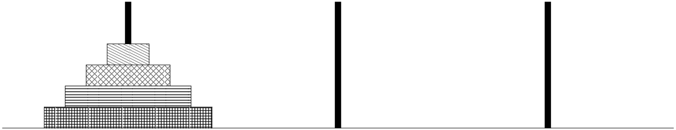
FORMULA a(n+1)=(a(n)^2)(a(n)+1)
log a(n) grows somewhat faster than O(3^n).

MAPLE f:=proc(n) option remember; if n = 0 then 1 else f(n-1)^2*(f(n-1)+1); fi; end;

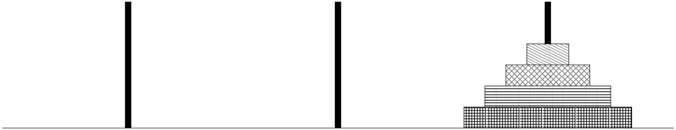
MATHEMATICA t={1, 2}; Do[AppendTo[t, t[[-1]]^3+t[[-1]]^2], {n, 6}]; t (* Vladimir Joseph Stephan Orlovsky, Feb 02 2012 *)

PROG
```
(Scheme)
(define (next n) (* n n (+ n 1)))
(define (list-elements nr-of-elements n0 next)
(let list-elements ((i 0) (n n0))
(show i n)
(let ((i (add1 i)))
(if (< i nr-of-elements) (list-elements i (next n))))))
(define (show i n) (printf "N(~a)=~a~n~n" i n))
(list-elements 6 1 next)
```

Hanoi tower : start from



and move disks (never more than one) ; a disk may never be atop a smaller one ; the end result should be



- The number of different non-self-crossing ways of moving a tower of Hanoi from one peg onto another peg, with $h+1$ disks, is given by the recurrence

$$a_{h+1} = {a_h}^2 + {a_h}^3 \qquad (a_0 = 2)$$

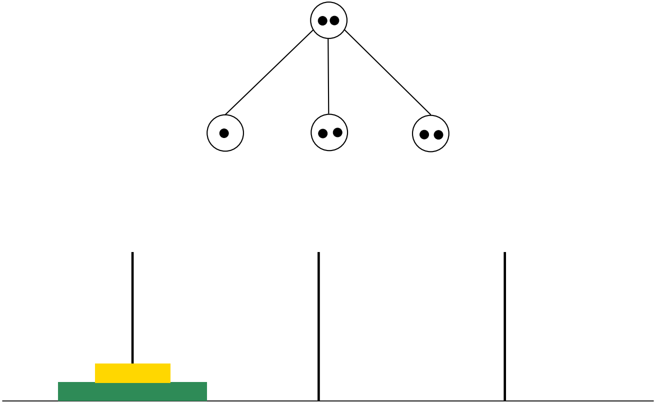
- The number of different non-self-crossing ways of moving a tower of Hanoi from one peg onto another peg, with $h+1$ disks, is given by the recurrence

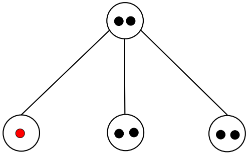
$$a_{h+1} = {a_h}^2 + {a_h}^3 \qquad (a_0 = 2)$$

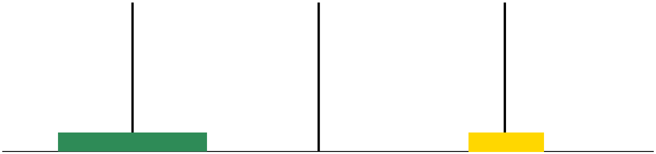
- This is exactly the recurrence for the number of 2-3 trees of height h !

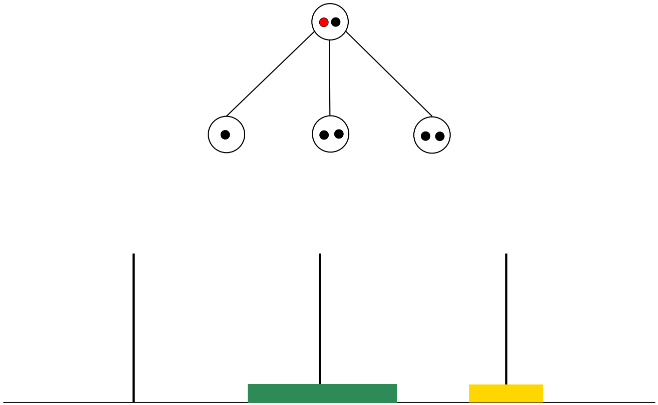
  $\Rightarrow$ bijection between 2-3 trees of height h and sequences of non-self-crossing ways to move $h+1$ disks ?

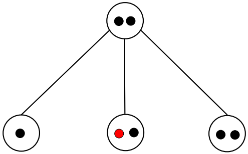
- Leaf with one key $\Leftrightarrow$ move a single disk from initial to final peg in one step
- Leaf with two keys $\Leftrightarrow$ move a single disk from initial to final peg in two steps
- Recursive structure of the tree $\Leftrightarrow$ recursive sequence of disk moves

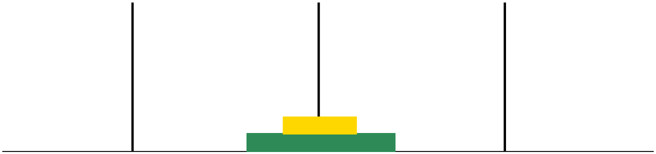


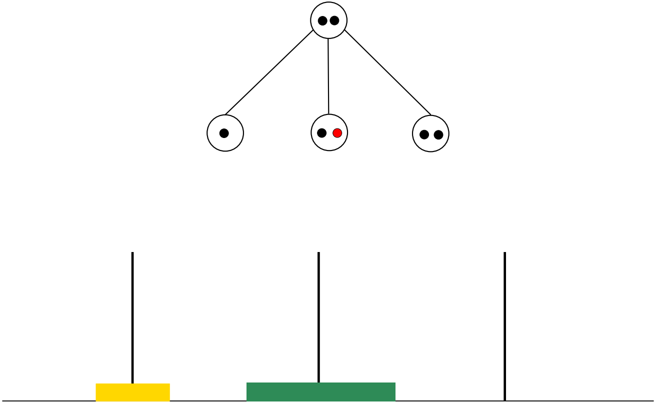


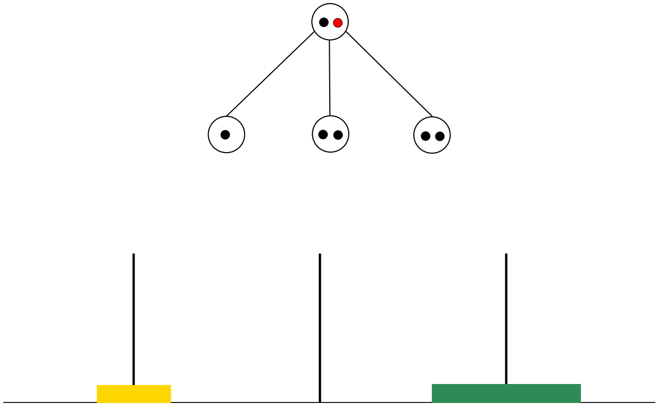


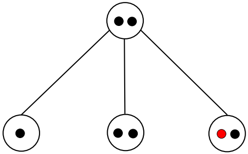


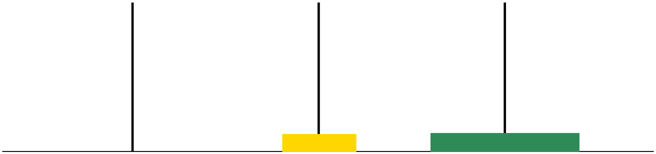


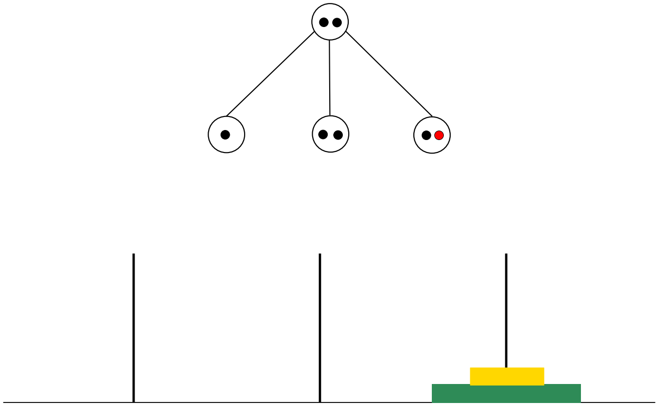












- Quickest way to solve the Hanoi problem $\Leftrightarrow$ "thinnest" 2-3 tree
- Slowest way to solve it without redundant moves $\Leftrightarrow$ "fattest" 2-3 tree
- Number of disk moves $=$ number of keys in the 2-3 tree
- Bottom disk at height 1 $\Leftrightarrow$ root at level 0
  Number of moves of bottom disk $=$ number of keys in the root node
- Number of moves of disk at height $i - 1 =$ Number of keys at level i

Number of trees with height h ?

Proposition

Asymptotic number b_h of B-trees with parameter m, height h

$$b_h = \kappa_m^{(\mu+1)^h} \left(1 + O\left(\frac{1}{(m+1)^{(\mu+1)^h}}\right)\right),$$

with $\mu = 2m$ or $2m - 1$ and

$$\kappa_m = v_0 \prod_{\ell \geq 0} \left(1 + \frac{1}{c_\ell} + \dots + \frac{1}{c_\ell^m}\right)^{\frac{1}{(\mu+1)^{h+1}}}.$$

where $c_0 = m + 1$ and

$$c_{h+1} = c_h^{\mu+1} \left(1 + \frac{1}{c_h} + \dots + \frac{1}{c_h^m}\right).$$

Back to insertion in a B-tree

Can we analyze the evolution of a B-tree (as done for binary search trees) ?

- Balancing condition $\Rightarrow$ an insertion can have far-reaching consequences : modify the ancestor nodes on a path up to the root plus the sister nodes
- We can analyze what happens at the lower level

Fringe : Terminal nodes, according to the number of keys in each of them

A terminal node has *type* k when it contains exactly $m + k - 2$ keys $(1 \leq k \leq m + 1)$

There are $m + k - 1$ distinct ways to insert a key in such a node

B-tree with n keys

- $X_n^{(k)}$: number of terminal nodes of type k
- $G_n^{(k)}$: number of ways to insert a key in nodes of type k

$$X_n = \begin{pmatrix} X_n^{(1)} \\ \vdots \\ X_n^{(m)} \end{pmatrix}; \qquad G_n = \begin{pmatrix} G_n^{(1)} \\ \vdots \\ G_n^{(m)} \end{pmatrix}$$

$$G_n = DX_n \qquad \text{with} \qquad D = \begin{pmatrix} m & & & \\ & m+1 & & \\ & & \ddots & \\ & & & 2m-1 \end{pmatrix}$$

$G_n^{(k)}$ number of insertion possibilities of type k in a tree with n keys

$G_n = \begin{pmatrix} G_n^{(1)} \\ \vdots \\ G_n^{(m)} \end{pmatrix}$ is a Pólya urn with m colors, balance $S = 1$,

and replacement matrix

$$R_m = \begin{pmatrix} -m & m+1 & & & \\ & -(m+1) & m+2 & & \\ & & \ddots & & \\ & & & -(2m-2) & 2m-1 \\ 2m & & & & -(2m-1) \end{pmatrix}.$$

The eigenvalues satisfy the equation

$$(\lambda + m) \dots (\lambda + 2m - 1) = \frac{(2m)!}{m!}$$

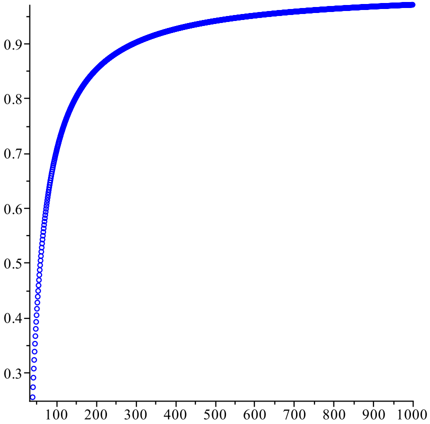
Equation for eigenvalues λ_j

$$(\lambda + m)\dots(\lambda + 2m - 1) = \frac{(2m)!}{m!}$$

- $\lambda_1 = 1$
- $\lambda_2, \overline{\lambda_2}$ conjugate with maximal real part < 1 ; $\sigma_2 := \Re(\lambda_2)$

m	σ_2
57	0.4775726941
58	0.4866133472
59	0.4953467200
60	0.5037882018
61	0.5119521623
62	0.5198520971

Variation of σ_2 according to m



Theorem ($\sim$ Janson) G_n vector for insertion possibilities

- Gaussian if $m \leq 59$: $\dfrac{G_n - nv_1}{\sqrt{n}}$ converges in distribution towards G
- non Gaussian if $m \geq 60$

$$G_n = nv_1 + 2\Re\left(n^{\lambda_2} W v_2\right) + o\left(n^{\sigma_2}\right)$$

with

- v_1, v_2 are deterministic vectors
- W is the limit of a complex-valued martingale
- $o(\;)$ is for a.s. and in all $L^p, p \geq 1$.