Evaluation des parametres du p-hachage dynamique virtuel

G. Levy

INRIA

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France

Tél. (1) 39 63 55 11

Rapports de Recherche

N° 740

EVALUATION DES PARAMETRES DU p-HACHAGE DYNAMIQUE VIRTUEL

Gérard LEVY

OCTOBRE 1987

EVALUATION DES PARAMETRES DU p-HACHAGE DYNAMIQUE VIRTUEL

Gérard Lévy*

Les méthodes de hachage dynamique (virtuel) (HDV), ont suscité de nombreuses études qui ont permis, dans le cas p=2, d'apprécier leurs performances. La justification mathématique de ces résultats a été donnée, toujours pour p=2, par [1].
L'objet du présent article est de calculer exhaustivement, pour p>2, les paramètres du p-HDV, et d'en étudier le comportement asymptotique. Notre démarche s'inscrit dans le prolongement des travaux de [3], où l'examen du cas général est déjà amorcé.

Mots clés : Hachage dynamique, indexation, performances asymptotique, transformée de Mellin.

* : Université Paris IX-Dauphine et projet SESAME-INRIA

EVALUATION OF PARAMETERS OF DYNAMIC VIRTUAL p-HASHING

Gérard LEVY *

Dynamic (virtual) hashing methods (HDV) has led to many studies that evaluated the performance for p = 2. Mathematical justification of these results for p = 2 was given in [1].
In this article we calculate in exhaustive manner the parameters of p-HDV for p > 2 and we study the asymptotic behavior. Our work extends the one presented in [3], where the study of the general case was initiated.

Key words : Dynamic, virtual hashing, index, asymptotic performance, Mellin transform.

* : University of Paris IX-Dauphine and SESAME Project at INRIA

TABLE DES MATIERES

INTRODUCTION

1.RAPPELS

1.1 Hypothése de "Bernouilli finie"

1.2 Hypothése de "Bernouilli infinie"

1.3 Hypothése de Poisson

2. RECURRENCES VERIFIEES PAR LES PARAMETRES DU P-HACHAGE

2.1Nombre de sommets de l'index

2.2 Coût moyen d'insertion d'une clé

2.3 Coût moyen de suppression d'une clé

2.4 Distribution des feuilles

2.5 Longueur du cheminement externe

2.6 Profondeur de l'index

3. FONCTIONS GENERATRICES DES PARAMETRES

4. ANALYSE ASYMPTOTIQUE DES PERFORMANCES

5. ANNEXE

6. CONCLUSION

7. BIBLIOGRAPHIE

EVALUATION DES PERFORMANCES DU HACHAGE DYNAMIQUE VIRTUEL

INTRODUCTION

Nous allons généraliser au cas des arbres p-homogènes les résultats qui ont été obtenus pour le hachage dynamique dans le cas des arbres binaires.Outre leur intêret mathématique,ces résultats devraient permet tre de juger de l'utilité ou non de se servir d'arbres de degré supé rieur à 2.Les outils et méthodes que nous employerons sont pour la plupart empruntés à [1],[2], [3].Nous avonspréféré opter pour cette approche car,en dépit de l'arsenal mathématique qu'elle nécessite, elle fournit un cadre d'étude unitaire et systématique,situation assez rare en analyse combinatoire.

Aprés quelques rappels mathématiques,nous donnerons les équations de récurrence desdifférents paramétres du hachage.Nous résoudrons ensuite ces équations.Enfin nous étudierons le comportement asymptotique des paramétres.Compte tenu de la necessaire briéveté des rappels,le lecteur se rapportera utilement aux travaux cités en rèférences.

I-RAPPELS

Soient p et b deux entiers positifs,p supérieur ou égal à 2.
Tous les arbres dont il sera question ici sont des arbres p-homogènes aux feuilles desquels sont associées des boîtes de capacité b,à l'in térieur desquelles on peut ranger des clés.Suivant le cas,ces clés sont des mots de longueur fixe s ou de longueur infinie de l'alphabet $L=\{0,1,\ldots,p-1\} = [0..p-1]$.

Soit L^s l'ensemble des mots de longueur s de L,et L^∞ celui des mots de longueur infinie.On désigne par $P^{(s)}$ l'ensemble des parties w de L^s et par $P_n^{(s)}$ celui des parties w de cardinal $|w| = n$,n entier naturel. (Notations similaires pour les mots infinis.)

Pour tout ensemble w de clés de L^s et tout élément i de L,on appelle w/i l'ensemble des mots m de L^{s-1} tels que im soit une clé de w. La correspondance entre w et (w/0,w/1,...,w/p-1) est une bijection entre $p^{(s)}$ et $p^{(s-1)}xp^{(s-1)}x...xp^{(s-1)}$ .De plus on a les relations

$$w=\bigcup_{i\in L}\{i\}\times(w/i) \quad \text{et} \quad |w|=\sum_{i\in L}|w/i| \ .$$

Les méthodes de hache-codage dynamique associent à tout ensemble w de clés un p-arbre de la maniére suivante:

- si w contient au plus b clés il lui correspond un arbre qui n'a qu'une seule feuille dont la boîte associée renferme ces clés.
-sinon on éclate la feuille associée à w en p sous-arbres associés aux différents ensembles $\{i\}x(w/i)$ sur lesquels on réitére la procédure d'éclatement jusqu'à n'avoir plus que des ensembles d'au plus b clés.

Les techniques de codage précédentes font correspondre à chaque w un certain nombre de paramétres tels que la taille de l'arbre,le taux d'occupation des feuilles,le coût moyen d'une insertion ou d'une suppression de clé,etc...Ces paramétres sont interessants par leurs valeurs moyennes ,sous certaines hypothèses statistiques à préciser.Les hypothéses le plus souvent étudiées,parce que les plus plausibles,sont:

BF("Bernouilli-Finie"): les clés sont de longueur fixe s finie,la taille du "fichier" w est figée à n,et tous les w de même cardinal n sont équiprobables;

BI("Bernouilli Infinie"):mêmes hypothéses que ci-dessus,sauf que les clés sont de longueur infinie;

PS("Poisson"):mêmes hypothéses que pour BI,sauf que la taille n du fichier w fluctue en suivant une Loi de Poisson de paramétre ν .

I-1 Hypothèse BF

On définit un paramètre $a^{(s)}$ comme une application de $P^{(s)}$ dans R, c-à-d un moyen d'associer à tout ensemble w de mots de L^s un réel $a^{(s)}(w)$.

Pour tout entier n compris entre 0 et p^s, on pose

$a_n^{(s)}$ = somme des $a^{(s)}(w)$ pour w appartenant à $P_n^{(s)}$.

La fonction génératrice $a^{(s)}$ est définie par

$$a^{(s)}(x) = \sum_{w\in P^{(s)}} a^{(s)}(w)\, x^{|w|} = \sum_{n=0}^{p^s} \Big(\sum_{w\in P_n^{(s)}} a^{(s)}(w) \Big) x^n = \sum_{n=0}^{p^s} a_n^{(s)} x^n \ .$$

I-1-1 Propriétés des fonctions génératrices:

1) si $a^{(s)}(w)=b^{(s)}(w)+c^{(s)}(w)$, alors $a^{(s)}(x)=b^{(s)}(x)+c^{(s)}(x)$;

2) si $a^{(s)}(w)=\alpha b^{(s)}(w)$ alors $a^{(s)}(x)=\lambda b^{(s)}(x)$;

3) si $a^{(s)}(w)=\prod_{i\in L} b_i^{(s-1)}(w/i)$ alors $a^{(s)}(x)=\prod_{i\in L} b_i^{(s-1)}(x)$.

I-1-2 Valuations particulières :

1) si $a^{(s)}(w)=1$ alors $a_n^{(s)}=|P_n^{(s)}|=C^n_{p^s}=\binom{p^s}{n}$ et $a^{(s)}(x)=(1+x)^{p^s}$;

2) si $a^{(s)}(w)=\delta_{|w|,q}$ (=1 si $|w|=q$, =0 sinon), alors $a_q^{(s)}=\binom{p^s}{q}$, et

$a^{(s)}(x)=\binom{p^s}{q}x^q$;

3) si $a^{(s)}(w)=|w|$, alors $a_n^{(s)}=n\binom{p^s}{n}$ et $a^{(s)}(x)=\sum_{n=0}^{p^s} n\binom{p^s}{n}=p^s x(1+x)^{p^s-1}$

4) si $a^{(s)}(w)=b_i^{(s-1)}(w/i)$ alors $a^{(s)}(w)=b_i^{(s-1)}(w/i) \prod_{j\in L-\{i\}} 1^{(s-1)}(w/j)$, d'où

$$a^{(s)}(x)=b_i^{(s-1)}(x)(1+x)^{(p-1)p^{s-1}} = b_i^{(s-1)}(x)(1+x)^{p^s-p^{s-1}} .$$

5) récurrences: si $a^{(s)}(w)= \sum_{i\in L} a^{(s-1)}(w/i) + b^{(s)}(w)$, alors

$$a^{(s)}(x)=p(1+x)^{(p-1)p^{s-1}}.a^{(s-1)}(x) + b^{(s)}(x).$$

Or si $a^{(s)}(x)=c_s(x).a^{(s-1)}(x)+d_s(x)$, avec $a^{(0)}(x)=c_0(x)$, alors on a

$$a^{(s)}(x)= \sum_{j=0}^{s} d_j(x) \prod_{k=j+1}^{s} c_k(x).$$

Donc $$a^{(s)}(x) = \sum_{j=0}^{s} d_j(x).p^{s-j}.(1+x)^{p^s-p^j} .$$

I-2 Hypothèse BI

Un paramétre $a^{(\infty)}$ est une application de $P^{(\infty)}$ dans R.

Pour tout entier naturel n , on définit

$a_n^{(\infty)}=E\{a^{(\infty)}(w) \mid w\in P_n^{(\infty)}\}$, et la série génératrice associée

$$a^{(\infty)}(x) = \sum_{n=0}^{\infty} a_n^{(\infty)}(x^n/n!)$$

Sous les hypothéses d'équiprobabilité que nous avons prises ici, on a

$$a_n^{(\infty)} = \lim_{s\to\infty} a_n^{(s)}/(p^s_n) = \lim a_n^{(s)}/((p^s)^n/n!) ;$$

Donc $a_n^{(\infty)}/n! = \lim_{s\to\infty} a_n^{(s)}/(p^s)^n$, et $a^{(\infty)}(x)= \lim_{s\to\infty} a^{(s)}(x/p^s)$.

Nous allons appliquer à l'étude de l'hypothése BI le même ordre et les mêmes notations qu'à étude de BF,en prenant soin toutefois de remplacer les indices supérieurs s et s-1 par l'indice ∞ .

I- 2-1 Propriétés des fonctionsgénératrices:

1),2),propriétés de linéarité sont vérifiées;

3) si $a^{(\infty)}(w) = \prod_{i\in L} b_i^{(\infty)}(w/i)$ alors $a^{(\infty)}(x) = \prod_{i\in L} b_i^{(\infty)}(x/p)$

I-2-2Valuations particulières:

1)si $a^{(\infty)}(w)=1$ alors $a_n^{(\infty)}=1$ et $a^{(\infty)}(x) = \sum_{n=0}^{\infty} 1(x^n/n!) = e^x$;

2) si $a^{(\infty)}(w)=\delta_{|w|,q}$ alors $a^{(\infty)}(x) = x^q/q!$;

3) si $a^{(\infty)}(w)= |w|$ alors $a^{(\infty)}(x) = \sum_{n=0}^{\infty} n(x^n/n!) = xe^x$;

4) si $a^{(\infty)}(w)=b_i^{(\infty)}(w/i)$ alors $a^{(\infty)}(x) =b_i^{(\infty)}(x/p).\ e^{\frac{p-1}{p}x}$

5) récurrences:si $a^{(\infty)}(w)= \sum_{i\in L} a^{(\infty)}(w/i) + b^{(\infty)}(w)$, alors

$$a^{(\infty)}(x)= p.e^{\frac{p-1}{p}x}.a^{(\infty)}(x/p) + b^{(\infty)}(x) \quad .$$

Or si $a^{(\infty)}(x) = c(x).a^{(\infty)}(x/p) + d(x)$, on a pour tout k naturel

$$a^{(\infty)}(x)= \sum_{j=0}^{k} d(x/p^j).\prod_{i=0}^{j-1} c(x/p^i) + a^{(\infty)}(x/p^{k+1}).\prod_{i=0}^{k} c(x/p^i) \ .$$

Avec $c(x)=p.\exp(\frac{p-1}{p}x)$, on obtient ,si les produits convergent ,

$$a^{(\infty)}(x) = e^{x}. \sum_{j=0}^{\infty} p^{j}.b^{(\infty)}(x/p^{j}).e^{-(x/p^{j})} \quad .$$

I-3 Hypothèse P

Les clés sont de longueur infinie et la taille du fichier est une variable aléatoire N qui suit une Loi de Poisson de paramétre ν .Donc $Prob(N=n)=e^{-\nu}.(\nu^{n}/n!)$,et la valeur moyenne $a^{(P)}$ du paramétre $a_n^{(\infty)}$ est

$$a^{(P)} = \sum_{n=0}^{\infty} a_n^{(\infty)}.e^{-\nu}.(\nu^{n}/n!) = e^{-\nu} . \sum_{n=0}^{\infty} a_n^{(\infty)}(\nu^{n}/n!) = e^{-\nu}.a^{(\infty)}(\nu) \quad .$$

II-RECURRENCES VERIFIEES PAR LES PARAMETRES DU HACHE-CODAGE.

Pour s fini ou non, nous définissons les fonctions caractéristiques χ à valeurs 0 ou 1 suivant qu'une proposition concernant w de $P^{(s)}$ est fausse ou vraie.Ainsi

$\chi(|w|<=b) = 1$ si $|w|<=b$, et 0 sinon.

On a évidemment

$$\delta_{|w|,j} = \chi(|w|=j) \quad , \quad \chi(|w|<=b) = \sum_{j=0}^{b} \delta_{|w|,j} = 1 - \chi(|w|>b) \quad .$$

L'arbre associé à l'ensemble w de clés s'appelle l'index , et ce sont ses paramétres que l'on étudie.

II-1 Nombre de sommets internes de l'index.

Ce nombre est égal à 0 si w n' pas plus de b clés.Sinon il est égal à la somme des nombres de sommets internes des index des différents w/i plus 1. Donc

$$N^{(s)}(w) = \Big(\sum_{i\in L} N^{(s-1)}(w/i) + 1 \Big) \chi(|w| > b) \text{ ,soit}$$

$$N^{(s)}(w) = \sum_{i\in L} N^{(s-1)}(w/i) + 1 - \chi(|w| <= b) \quad .$$

Le nombre de feuilles de l'index est lié au nombre de sommets internes par l'égalité déjà rencontrée

$N^{(s)}(w)=(F^{(s)}(w)-1)/(p-1)$, d'où $F^{(s)}(w)= 1+ (p-1).N^{(s)}(w)$.

II-2 Coût moyen d'insertion d'une clé dans un fichier w de n clés.

Soit $I^{(s)}(w)$ ce coût. Il est est égal à la moyenne des coûts d'insertion de la nouvelle clé dans les différents w/i, si w contient plus de b clés. Sinon il vaut 1 si w en contient b exactement, puisqu'alors un éclatement est nécessaire. D'où la récurrence

$$I^{(s)}(w) =(1/p).\Big(\sum_{i\in L} I^{(s-1)}(w/i) \Big)+ \delta_{|w|,b} \quad .$$

Une autre maniére d'évaluer ce nombre consiste à dire que pour chaque n $I_n^{(s)}$ est égal au nombre d'éclatements résultant des insertions, donc à la différence $N_{n+1}^{(s)} - N_n^{(s)}$ des nombres de sommets internes des index des fichiers à n+1 et à n clés, puisque chaque éclatement a pour effet d'ajouter un sommet interne nouveau à l'index. D'où

$$I_n^{(s)} = N_{n+1}^{(s)} - N_n^{(s)} \quad .$$

II-3 coût moyen de suppression .

Avec les mêmes notations que çi-dessus, w ayant n clés, la suppression etant l'opération inverse de l'insertion , on peut considérer qu'en moyenne le nombre $S^{(s)}(w)$ ou $S_n^{(s)}$ de suppressions dans un fichier de n clés est égal au nombre d'insertions dans un fichier de n-1 clés. D'où

$$S_n^{(s)} = I_{n-1}^{(s)} \quad .$$

II-4-Distribution des feuilles.

Soit F_q le nombre des feuilles de l'index qui contiennent exactement q feuilles, q entier compris entre 0 et b.

Si w a plus de b clés ce nombre est égal à la somme des nombres F_q correspondants aux différents w/i. Sinon il vaut 1 si w contient q clés, Et 0 sinon. D'où

$$F_q^{(s)}(w) = \Big(\sum_{i\in L} F_q^{(s-1)}(w/i) \Big).\chi(|w|>b) + \delta_{|w|,q}$$

$$= \sum_{i\in L} F_q^{(s-1)}(w/i) + \delta_{|w|,q} - \sum_{i\in L} F_q^{(s-1)}(w/i).\ \chi(|w| <= b) .$$

Or si w a au plus b clés, il en est de même de w/i et $F_q^{(s-1)}(w/i)=0$ ou 1. Plus précisément, ce nombre vaut 1 ssi w/i a q clés et l'ensemble des autres w/j, j≠i, en contient au plus b-q. Donc

$$F_q^{(s-1)}(w/i).\chi(|w|<=b) = \delta_{|w/i|,q}.\chi\Big(\sum_{j\in L-\{i\}} |w/j| <= b-q \Big).$$

Posons E_k = ensemble des w tels que la somme des $|w/j|$, j≠i, est égale à k.

Désignons par R_k l'ensemble des suites (r_j) de p-1 entiers naturels r_j telles que la somme de ces r_j soit égale à k. On a les égalités

$$\{ w \mid \sum_{j\neq i} |w/j| <= b-q \} = \bigcup_{k=0}^{b-q} E_k = \bigcup_{k=0}^{b-q} \bigcup_{(r_j)\in R_k} \{w \mid (w/j)=(r_j)\}.$$

$$\chi\Big(\sum_{j\neq i} |w/j| <= b-q \Big) = \sum_{k=0}^{b-q} \sum_{(r_j)\in R_k} \prod_{j\neq i} \delta_{|w/j|,r_j} .$$

D'où l'expression de la récurrence relative aux F_q:

$$F_q^{(s)}(w) = \sum_{i\in L} F_q^{(s-1)}(w/i) + \delta_{|w|,q} - \sum_{i\in L} \delta_{|w/i|,q} \cdot \sum_{k=0}^{b-q} \sum_{(r_j)\in R_k} \prod_{j\neq i} \delta_{|w/j|,r_j}.$$

II-5 Longueur du cheminement externe.

C'est la somme des longueurs des chemins conduisant de la racine aux feuilles de l'index. Les nombres $L^{(s)}(w)$ satisfont la relation

$$L^{(s)}(w) = \Big(\sum_{i\in L}(L^{(s-1)}(w/i) + |w/i|)\Big).\chi(|w|>b)$$

$$= \sum_{i\in L} L^{(s-1)}(w/i) + |w| - \sum_{j=0}^{b} j\delta_{|w|,j} .$$

II-6 Profondeur de l'index.

sous l'hypothèse BI, soit $P_n^{(\infty)}(k)$ la probabilité qu'un index de taille n soit de profondeur au plus égale à k. Posons

$P_k^{(\infty)}(w)=1$ si la profondeur de w est au plus k, et 0 sinon.

Puisque $P_k^{(\infty)}(n)=E\{ P_k^{(\infty)}(w) \mid |w| =n \}$, on a $P_k^{(\infty)}(n) = P_n^{(\infty)}(k)$.

De plus $P_{k+1}^{(\infty)}(w)= \prod_{i\in L} P_k^{(\infty)}(w/i)$.

De ces relations, on déduit que la probabilité qu'un index de taille n soit égale à k, vaut $P_n^{(\infty)}(k)- P_n^{(\infty)}(k-1)$, et que la profondeur moyenne de l'index est, pour un fichier de taille n,

$$PF(n) = \sum_{k=1}^{\infty} k(P_n^{(\infty)}(k) - P_n^{(\infty)}(k-1))$$

III-FONCTIONS GENERATRICES DES PARAMETRES.

Nous allons nous servir des outils fournis en I pour les appliquer aux récurrences de II,dans l'ordre où elles ont été obtenues,en vue d'obtenir les fonctions génératrices des différents paramétres des index.De ces fonctions nous déduirons l'expression des paramétres,puis ultérieurement leur comportement asymptotique.

III-1 nombre de sommets internes.

Hypothése BF:on a $b^{(s)}(w)=1-\chi(|w|<=b)=1-\sum_{j=0}^{b}\delta_{|w|,j}$.

Donc $b^{(s)}(x)=d_s(x)=(1+x)^{p^s}-\sum_{j=0}^{b}\binom{p^s}{j}.x^j$. ET

$$N^{(s)}(x)=\sum_{k=1}^{s}p^{s-k}(1+x)^{p^s-p^k}.\left(\sum_{j=b+1}^{p^s}\binom{p^k}{j}.x^j\right)$$

$$=\sum_{k=1}^{s}p^{s-k}\left[\sum_{n=b+1}^{p^s}\left(\sum_{j=b+1}^{n}\binom{p^k}{j}\binom{p^s-p^k}{n-j}.x^n\right)\right] .$$

Il en résulte que le coefficient de x^n vaut

$$N_n^{(s)}=\sum_{k=1}^{s}p^{s-k}\left(\sum_{j=b+1}^{n}\binom{p^k}{j}\binom{p^s-p^k}{n-j}\right) \text{ ,si } n>=b+1\text{,et 0 si } n<=b .$$

Hypothése BI:

$$b^{(\infty)}(x)=e^x-\sum_{j=0}^{b}(x^j/j!)=e^x-e_b(x)=\sum_{j=b+1}^{\infty}(x^j/j!).$$

$$N^{(\infty)}(x)=\sum_{k=0}^{\infty}p^k.\exp(1-1/p^k).\left[\exp(x/p^k)-\sum_{j=0}^{b}(x/p^k)^j/j!\ \right].$$

Le coefficient de $x^n/n!$ est donné par

$$N_n^{(\infty)} = \sum_{k=0}^{\infty} p^k \Big[1 - \sum_{j=0}^{b} \binom{n}{j} (1/p^k)^j (1-1/p^k)^{n-j} \Big] .$$

Le taux moyen de remplissage des boîtes attachées aux feuilles de l'index est défini par

$\tau_n = n/(bF_n) = n/b(1+(p-1)N_n)$, et $\tau = \lim_{n\to\infty} \tau_n$.

III-Coût moyen d'insertion.

On a les égalités suivantes

$I^{(s)}(x) = p.\underline{(1/p)}.I^{(s-1)}(x) + \binom{p^s}{b}.x^b$, soit $c_s(x)=(1+x)^{p^s}-p^{s-1}$, et

$d_s(x)=\binom{p^s}{b}.x^b$. D'où $I^{(s)}(x)=(1+x)^{p^s}-p^{s-1} I^{(s-1)}(x)+\binom{p^s}{b}.x^b$;

$$I^{(s)}(x) = x^b . \sum_{j=0}^{s} \binom{p^j}{b} . (1+x)^{p^s-p^j} .$$

Ce qui permet de determiner les $I_n^{(s)}$, en développant l'expression

$$I^{(s)}(x)=x^b . \sum_{j=0}^{s} \binom{p^j}{b} . \sum_{k=0}^{p^s-p^j} \binom{p^s-p^j}{k} x^k = \sum_{k=0}^{p^s-1} \Big(\sum_{j=0}^{\mathrm{Log}_p(p^s-k)} \binom{p^j}{b} \binom{p^s-p^j}{k} \Big) x^{b+k} .$$

D'où $I_n^{(s)} = \sum_{j=0}^{\mathrm{Log}_p(p^s-n+b)} \binom{p^j}{b} \binom{p^s-p^j}{n-b}$.

Sous l'hypothése BI on a

$I^{(\infty)}(x)=I^{(\infty)}(x/p).\exp((1-1/p)x) + x^b/b!$.

Ce qui entraine que

$$I^{(\infty)}(x) = \sum_{k=0}^{\infty} (1/b!).e^{x}.e^{-(x/p^k)} = (x^b/b!).\sum_{k=0}^{\infty} (1/p^k)^b.e^{x(1-1/p^k)} .$$

Le coefficient de $x^n/n!$ permet d'obtenir

$$I_n^{(\infty)} = \binom{n}{b}.\sum_{k=0}^{\infty} (1/p^k)^b(1-1/p^k)^{n-b}$$

Sous l'hypothèse P de Poisson,la valeur moyenne du coût d'insertion est

$$I^{(P)}(\nu) = e^{-\nu}.I^{(\infty)}(\nu) = (\nu^b/b!).\sum_{k=0}^{\infty} (1/p^k)^b.e^{-\nu/p^k} .$$

III-3 Coût moyen de suppression.

Son étude découle immédiatement de celle du coût d'insertion,puisque

$S_n = I_{n-1}$.

III-4- Distribution des feuilles.

A la différence des autres paramétres,les seconds membres des relations de récurrence relatives à la repartition des feuilles ont des expressions assez compliquées. Il nous faudra donc en un premier lieu,calculer leurs fonctions génératrices. Nous allons poser , pour simplifier les notations $g_i((w/j):j\neq i)=\chi(\sum_{j\in L-\{i\}} |w/j| <= b-q)$.

Une expression détaillée de g_i a été dèjà calculée.

Hypothése BF:

On sait que dans $L^{(s)}$,il correspond à $\delta_{|w|,q}$ la fonction génératrice $\binom{p^s}{q}x^q$.Il correspond donc à g_i,compte tenu de son expression détaillée,

$$\sum_{k=0}^{b-q} \sum_{(r_j)\in R_k} \prod_{j\neq i} \binom{p^{s-1}}{r_j} x^{r_j} = \sum_{k=0}^{b-q} \sum_{(r_j)\in R_k} x^{(\Sigma r_j)} \prod_{j\neq i} \binom{p^{s-1}}{r_j}.$$

Or pour toute suite (r_j) de R_k, la somme des r_j est égale à k.Donc

$$g_i(x) = \sum_{k=0}^{b-q} x^k \sum_{(r_j)\in R_k} \prod_{j\neq i} \binom{p^{s-1}}{r_j} = \sum_{k=0}^{b-q} \binom{p^s-p^{s-1}}{k} x^k.$$

Cette égalité s'obtient aisément si on remarque que

$p^s-p^{s-1}=(p-1)p^{s-1}$, et donc que $(1+x)^{p^s-p^{s-1}} = \prod_{j\in L-\{i\}} (1+x)^{p^{s-1}} =$

$$= \prod_{j\neq i} \left(\sum_{r_j=0}^{p^{s-1}} \binom{p^{s-1}}{r_j} x^{r_j} \right) = \sum_{k=0}^{p^s} \sum_{(r_j)\in R_k} x^{(\Sigma r_j)} \prod_{j\neq i} \binom{p^{s-1}}{r_j}$$

Donc $\sum_{(r_j)\in R_k} \prod_{j\neq i} \binom{p^{s-1}}{r_j} = \binom{p^s-p^{s-1}}{k}$.

Si on convient de poser pour toute fonction f qui admet le développement en série

$f(x) = \sum_{n=0}^{\infty} a_n x^n$, $[\![f(x)]\!]_i^j = \sum_{n=i}^{j} a_n x^n$,on constate que

$$g_i(x) = [\![(1+x)^{p^s-p^{s-1}}]\!]_0^{b-q} .$$

Il correspond,à présent,à $\delta_{|w/i|,q} \cdot g_i((w/j))$ la fonction

$\binom{p^{s-1}}{q} x^q . g_i(x)$,et au second membre tout entier de la récurrence,

$$b_q^{(s)}(x) = d_s(x) = \binom{p^s}{q} x^q - p\binom{p^{s-1}}{q} x^q \left(\sum_{k=0}^{b-q} \binom{p^s-p^{s-1}}{k} x^k \right) .$$

Les seconds membres relatifs aux feuillesvides et pleines sont

$$b_0^{(s)}(x) = 1 - p \sum_{k=0}^{b} \binom{p^s-p^{s-1}}{k} x^k \text{ , et } b_b^{(s)}(x) = \left(\binom{p^s}{b} - p\binom{p^{s-1}}{b} \right) x^b .$$

La fonction génératrice $F_q^{(s)}(x)$ du nombre de feuilles qui contiennent q clés est donc

$$F_q^{(s)}(x) = \sum_{j=0}^{s} b_q^{(j)}(x).p^{s-j}.(1+x)^{p^s-p^j}$$

$$= x^q \sum_{j=0}^{s} p^{s-j}\left(\binom{p^j}{q} - p\binom{p^{j-1}}{q}\right)(1+x)^{p^s-p^j}\left(\sum_{k=0}^{b-q}\left(p^s_k - p^{s-1}\right)x^k\right)$$

et le nombre $F_{q,n}^{(s)}$ est le coefficient de x^n de $F_q^{(s)}(x)$.

Hypothése BI.

Ici encore c'est la determination de $b_q^{(\infty)}(x)$ qui pose probléme.

Du fait qu'à $\delta_{|w/j|,r_j}$ il correspond $(x/p)^{r_j}/r_j!$,il correspond à

$g_i((w/j))$ la fonction

$$g_i(x) = \sum_{k=0}^{b-q} \sum_{(r_j)\in R_k} \prod_{j\neq i} (x/p)^{r_j}/r_j!! = \sum_{k=0}^{b-q} \frac{(x^k/k!)}{p^k} \sum_{(r_j)\in R_k} \left(\frac{k!}{\prod_{j\neq i} r_j!}\right)$$

Or la somme des termes portant sur les (r_j) de R_k est égale à $(p-1)^k$.

D'où $g_i(x) = \sum_{k=0}^{b-q} \left(\frac{p-1}{p}x\right)^k/(k!) = e_{b-q}\left(\frac{p-1}{p}x\right) = [\![\exp\left(\frac{p-1}{p}x\right)]\!]_0^{b-q}$.

A titre indicatif , on remarquera qu'une manière de calculer le coeffi cient de $x^k/(k!)$ dans $g_i(x)$ est d'évaluer de deux façons différentes ce coefficient dans

$$\left(e^{\frac{x}{p}}\right)^{p-1} = \prod_{j\neq i} e^{\frac{x}{p}} = \prod_{j\neq i} \sum_{r_j=0}^{\infty} (x/p)^{r_j}/(r_j!) = \sum_{k=0}^{\infty} \left(\frac{p-1}{p}x\right)^k/(k!) \quad .$$

Au second membre tout entier de la récurrence correspond enfin

$$b_q^{(\infty)}(x) = (x^q/q!) - p((x/p)^q/q!).e_{b-q}(\frac{p-1}{p}x) = (x^q/q!)[1-p^{1-q}.e_{b-q}(\frac{p-1}{p}x)].$$

D'où $F_q^{(\infty)}(x) = e^x.\sum_{j=0}^{\infty} p^j.b_q^{(\infty)}(x/p^j).e^{-(x/p^j)}$

$$= e^x.(x^q/q!).\sum_{j=0}^{\infty} p^{(1-q)j}[1-p^{1-q}.e_{b-q}(\frac{p-1}{p}(x/p^j))].e^{-(x/p^j}$$

Le calcul des coefficients de x^n dans les séries génératrices conduit aux résultats suivants:

sous des hypothéses de Bernouilli ,pour des fichiers w de n clés,le nombre de feuilles de l'index qui contiennent q clés est

$$F_{q,n}^{(s)} = \delta_{n,q} + \sum_{k=0}^{s} p^{(s-k)}[\binom{p}{q}\binom{p^s-p^k}{n-q} - p\binom{p^{k-1}}{q}\sum_{i=0}^{b-q}\binom{p^k-p^{k-1}}{i}\binom{p^s-p^k}{n-q-i}] ;$$

$$F_{q,n}^{(\infty)} = \delta_{n,q} + \binom{n}{q}\sum_{k=0}^{\infty} p^{(1-q)k}[(1-p^{-k})^{n-q} - p^{1-q}\sum_{i=0}^{b-q}\binom{n-q}{i}(\frac{p-1}{p}p^{-k})^i(1-p^{-k})^{n-q-i}$$

III-5-<u>Longueur du cheminement externe</u>.

Au second membre $|w| - \sum_{j=0}^{b} j\,\delta_{|w|,j}$ correspond dans le cas BF

$$b^{(s)}(x) = p^s x(1+x)^{p^s-1} - \sum_{j=0}^{b} j\binom{p^s}{j}x^j = \sum_{j=b+1}^{p^s} j\binom{p^s}{j}x^j \; ; \text{ d'où}$$

$$L^{(s)}(x) = \sum_{k=0}^{s} p^{s-k}\, b^{(k)}(x).(1+x)^{p^s-p^k}$$

$$= \sum_{k=0}^{s} p^{s-k}.\Big(\sum_{i=b+1}^{p^s} i\binom{p^s}{i}x^i\Big).\Big(\sum_{j=0}^{p^s-p^k}\binom{p^s-p^k}{j}x^j\Big)$$

Dans le cas BI, il correspond au second membre la fonction

$$b^{(\infty)}(x) = xe^x - \sum_{j=0}^{b} j(x^j/j!) = x(e^x - e_{b-1}(x)). \text{ D'où}$$

$$L^{(\infty)}(x) = e^x \sum_{k=0}^{\infty} p^k . b^{(\infty)}(x/p^k) . e^{-(x/p^k)}$$

$$= xe^x(1-e^{-x}e_{b-1}(x)) + xe^x \sum_{k=1}^{\infty}[1 - e^{-(x/p^k)}(1+\sum_{j=1}^{b-1} \frac{1}{j!} \frac{x^j}{(p^k)^j})] .$$

L'extraction des coefficients de x^n dans ces fonctions génératrices donne les valeurs du paramètre L_n sous les deux hypothèses BF et BI.

III-6- Profondeur de l'index.

Sous l'hypothèse BI, on a vu que $P_{k+1}(w) = \prod_{i \in L} P_k(w/i)$.

De plus on a $P_0(w)=1$ si et seulement si w n'a pas plus de b clés. Donc

$P_k(x) = (P_{k-1}(x/p))^p = \ldots = (P_0(x/p^k))^{p^k}$, avec

$P_0(x) = \sum_{j=0}^{b} x^j/j! = e_b(x)$. Par conséquent

$P_k(x) = (e_b(x/p^k))^{p^k}$.

Il s'ensuit que la probabilité que l'index d'un fichier de n clés soit de profondeur inférieure ou égale à k est :

$$P_n(k) = P_k(n) = n![x^n](e_b(x/p^k))^{p^k}.$$

Dans le cas poissonien, on obtient la probabilité que la profondeur du fichier soit inférieure ou égale à k

$$P_\nu(k) = e^{-\nu}.P_k(\nu) = (f_b(\nu p^{-k}))^{p^k}, \text{ avec } f_b(x) = e^{-x}.e_b(x).$$

Quant à la profondeur moyenne de l'index, elle est donnée par

$$PF(n) = \sum_{k=1}^{\infty} k(P_n(k) - P_n(k-1)) = \sum_{k=0}^{\infty} (1-P_n(k)),$$

ainsi qu'on peut s'en rendre compte par transformation d'Abel.
Sous l'hypothèse de Poisson, on a un résultat similaire en passant aux valeurs moyennes.

IV- ANALYSE ASYMPTOTIQUE DES PERFORMANCES

Nous allons nous interesser au comportement asymptotique de certains des paramétres du hachage dynamique virtuel sous la seule hypothése Poissonienne , lorsqueela taille moyenne ν du fichier w devient infinie. On constate ,comme on le verra un peu plus loin,que l'expression de plusieurs de ces paramétres utilise des séries infinies d'un type particulier qui on déjà été étudiées pour p=2,et dont l'étude se généra lise immédiatement pour p plus grand que 2.

Définition:

Posons $S_j(x)= \sum_k p^{(1-j)k}.\ e^{-xp^{-k}}$,p,j entiers superieurs à 1, x positif.

LemmeAsymptotiquement en x,on a

$$S_j(x)=x^{1-j}.\frac{\Gamma(j-1)}{Ln(p)}\left(1+ \sum_{1\neq 0} e^{2i\pi Log_p x}\ \frac{\Gamma(j-1+2i1\pi/Ln(p))}{\Gamma(j-1)} + O(x^{-m})\right),$$

où m est un réel positif quelconque.

La preuve est identique à celle de [ ] où l'on remplace 2 par p.

La somme qui apparaîtdans les $S_j(x)$ est une fonction périodique de $Log_p x$ de période 1,de moyenne nulle,notée $N_j(Log_p x)$.

IV-Nombre de sommets internes de l'index .

On a vu que le second membre,pour s infini,est $b(x)= \sum_{j=b+1}^{\infty} x^j/j!$.

Donc $N^{(\infty)}(x)= e^x \sum_{k=0}^{\infty} p^k.b(x/p^k).e^{-x/p^k} = e^x \sum_{k=0}^{\infty} p^k\left(\sum_{j=b+1}^{\infty} x^j/(p^k)^j\right).e^{-x/p^k}$

$$=e^x \sum_{j=b+1}^{\infty} \frac{x^j}{j!}\left(\sum_{k=0}^{\infty} p^{(1-j)k}.\ e^{-x/p^k}\right) = e^x \sum_{j=b+1}^{\infty} \frac{x^j}{j!}\ S_j(x) .$$

Or $N^{(P)}(\nu) = e^{-\nu}.\ N^{(\infty)}(\nu)$. Donc

$$N^{(P)}(\nu) = \sum_{j=b+1}^{\infty} (\nu^j/j!).S_j(\nu) \ .$$

Compte tenu du lemme , on a ,lorsque le paramétre devient infini:

$$N^{(P)}(\nu) = \sum_{j=b+1}^{\infty} (1/j!).\nu^j.\nu^{1-j}.\frac{\Gamma(j-1)}{Ln(p)}.(1+N_j(Log_p\nu)+ O(\nu^{-m})) \ .$$

$$=(\nu/Ln(p)).\sum_{j=b+1}^{\infty} \frac{1}{j(j-1)} (1 + N_j(Log_p\nu) + O(\nu^{-m})) \ .$$

Or la série de terme $1/j(j-1)$, j allant de b+1 à l'infini, est convergente et a pour somme $1/b$. Par conséquent on a le

Théorème

Le nombre moyen de sommets internes de l'index d'un fichier dont la taille est une variable aléatoire de Poisson de paramétre ν,est asymptotiquement en ν:

$$N^{(P)}(\nu) = \frac{\nu}{bLn(p)} (1 + N_S(Log_p\nu) + O(\nu^{-m})) \ ,$$

où m est un réel positif quelconque , et N_S une fonction périodique de période 1,de moyennne nulle de $Log_p\nu$, et dont les coefficients de Fourier sont

$$c_l = \sum_{j=b+1}^{\infty} \frac{b}{j(j-1)} .\frac{\Gamma(j-1+2il\pi/Ln(p))}{\Gamma(j-1)} \quad , \text{ pour } l \neq 0 \ .$$

Puisque le nombre total de feuilles de l'index $F^{(P)}$ est lié au nombre de sommets internes par $F = 1 + (p-1)N$, on en déduit facilement une expression de F .De cette derniére découle le taux de remplissage asymptotique

Corollaire:

Sous les mêes hypothéses que précédemment, le taux de remplissage asymptotique des boftesattachées aux feuilles de l'index est

$$\tau = \frac{Ln(p)}{p-1} \quad .$$

Remarque :pour p=2, on retrouve la valeur connue $\tau = Ln(2)$. De plus on constate que le taux de remplissage moyen est une fonction décroissante de p.

Statistiques sur les feuilles.

Nous allons effectuer la même démarche pour calculer lecomportement asymptotique du nombre $F_q^{(P)}(\nu)$. On a

$$F_q^{(P)}(x) = \sum_k p^k . b_q(x/p^k) . e^{-x/p^k}$$

$$= \sum_k p^k (1/q!) . (x/p^k)^q . [\, 1 - p^{1-q} \sum_{j=0}^{b-q} (1/j!) (\frac{p-1}{p})^j . (x/p^k)^j \,] . e^{-x/p^k}$$

$$= (x^q/q!)[\, S_q(x) - p^{1-q} \sum_{j=0}^{b-q} (1/j!) (\frac{p-1}{p})^j x^j S_{j+q}(x) \,] .$$

Pour q au moins égal à 2, on peut utiliser le lemme.on obtient

Théorème:

asymptotiquement en ν, le nombre moyen de feuilles de l'index qui contiennent q clés est

$$F_q^{(P)}(\nu) = \frac{\nu}{q!\,Ln(p)} \left[\Gamma(q-1) - p^{1-q} \sum_{j=0}^{b-q} (1/j!) \left(\frac{p-1}{p}\right)^j \Gamma(j+q-1) \right] \left[1 + F_S(\mathrm{Log}_p \nu) + O(\nu^{-m}) \right]$$

où $F_S(\mathrm{Log}_p \nu)$ est une fonction périodique de période 1, de moyenne nulle, dont les coefficients de Fourier se calculent comme ceux de $N^{(P)}(\nu)$, et où m est un réel positif quelconque. (Résultat vrai pour $q>=2$)

Corollaire

Asymptotiquement, le nombre moyen de feuilles qui contiennent q clés est

$$F_q^{(P)}(\nu) \sim \frac{\nu}{q(q-1)Ln(p)} \left[1 - p^{1-q} \sum_{j=0}^{b-q} \binom{j+q-2}{j} \cdot \left(\frac{p-1}{p}\right)^j \right] \text{, pour } q>=2.$$

Preuve : elle se fonde sur le fait que $\Gamma(n) = (n-1)!$.

En particulier si q=b, le nombre moyen de feuilles pleines est donné par

$$F_b^{(P)}(\nu) \sim \frac{\nu}{b(b-1)Ln(p)} \left(1 - p^{1-b}\right) .$$

Remarque: ici encore on constate que le nombre de feuilles pleines est fonction décroissante de p. Ce qui était, bien sûr, prévisible.

Cas des feuilles vides (q=0)

Ainsi qu'on l'a vu, les résultats précédants ne sont valables que si q est supérieur à 1. Celà vient du fait que les séries $S_j(x)$ divergent pour q=0 ou 1. Or l'étude du nombre de feuilles vides est un indicateur intéressant dans le problème de hachage, puisqu'il est égal au nombre de cases inoccupées. Pour déterminer ce nombre nous allons suivre la

démarche de [1] qui consiste à déterminer le nombre de sommets internes de l'index qui ont au moins un fils qui contient q clés exactement. Ces nombres $M_q^{(\infty)}(w)$ sont liés, si le cardinal de w est supérieur à b, par

$$M_q(w) = \sum_{i\in L} M_q(w/i) + \sum_{i\in L} \delta_{|w/i|,q} \cdot \chi\Big(\sum_{j\neq i} |w/j| > b-q\Big) .$$

De plus $M_q(w) = \delta_{|w|,q}$, si w est de cardinal inférieur à b+1.

Il en résulte, dans le cas BI, en omettant l'indice supérieur (∞),

$$M_q(x) = p.e^{((p-1)/p)x}.M_q(x/p) + p.(1/q!)(x/p)^q.\sum_{j=b-q+1}^{\infty} (1/j!)\Big(\frac{p-1}{p}x\Big)^j$$

$$M_q(x) = e^x \cdot \sum_{j=0}^{\infty} p^j b_q(x/p^j).e^{-xp^{-j}} = pe^x \sum_{m=b+1}^{\infty} (x/p)^m \frac{(p-1)^{m-q}}{q!(m-q)!}\Big(\sum_{k=0}^{\infty} p^{(1-m)j} e^{-xp^{-j}}\Big)$$

Soit, en tenant compte du cas où w est de cardinal non supérieur à b,

$$M_q(x) = (x^q/q!) + pe^x.\sum_{m=b+1}^{\infty} (x/p)^m.\frac{(p-1)^{m-q}}{q!(m-q)!}.S_m(x).$$

Sous l'hypothèse P, on a $M_q^{(P)}(\nu) = e^{-\nu}.M_q(\nu)$; ce qui donne pour q=0,

$$M_0^{(P)}(\nu) = e^{-\nu} + p.\sum_{m=b+1}^{\infty} (\nu/p)^m \frac{(p-1)^m}{m!}.S_m(\nu).$$

$$M_0^{(P)}(\nu) = e^{-\nu} + p.\sum_{m=b+1}^{\infty} (\nu/p)^m \frac{(p-1)^m}{m!}.\nu^{1-m}\frac{\Gamma(m-1)}{Ln(p)}\{1+N_m(Log_p\nu)+O(\nu^{-m})\}$$

$$M_0^{(P)}(\nu) = ((p\nu)/Ln(p)).U_b.[1 + W(Log_p\nu) + O(\nu^{-m})] \text{ , où } U_b = \sum_{m=b+1}^{\infty} \frac{1}{m(m-1)}u^m ,$$

$u=\frac{p-1}{p}$, et W est une fonction périodique de $Log_p\nu$, de moyenne nulle et de période 1.

De plus $U_b = \frac{1}{b(b+1)} u^{b+1}\left[1 + \sum_{m=1}^{\infty} \frac{b(b+1)}{(b+k)(b+k+1)} u^k\right]$.

Or u est positif et strictement inférieur à 1, et le coefficient de u^k est également inférieur à 1.

Par conséquent

$$\frac{1}{b(b+1)} u^{b+1} < U_b < \frac{1}{b(b+1)} u^{b+1} (1/(1-u)).$$

En résumé, on a le résultat suivant:

Thèoréme: asymptotiquement en ν, sous l'hypothèse P, le nombre moyen de feuilles de l'index qui sont inoccupées, $F_0^{(P)}(\nu)$, est tel que

$F_0^{(P)}(\nu) \sim \frac{p\nu}{Ln(p)} . U_b$, avec une précision de $O(\nu^{-m})$, m réel positif, et

$V_b < U_b < p.V_b$, en posant $V_b = \frac{1}{b(b+1)} \left(\frac{p-1}{p}\right)^{b+1}$.

Remarque: du fait que $p/Ln(p)$ et que $\left(\frac{p-1}{p}\right)^{b+1}$ sont fonctions croissantes de p, on constate que le nombre de feuilles inoccupées de l'index est lui aussi fonction croissante de p; ce qui était prévisible.

IV-3-Coût moyen d'insertion ou de suppression .

Nous avons établi que $I^{(P)}(\nu) = (\nu^b/b!) \sum_{k=0}^{\infty} (1/p^k)^b . e^{-\nu/p^k}$.

Nous pouvons réécrire $I^{(P)}(\nu) = (\nu^b/b!) \sum_{k=0}^{\infty} p^{(1-(b+1))k} . e^{-\nu/p^k}$.

Soit $I^{(P)}(\nu) = (\nu^b/b!).S_{b+1}(\nu)$.

En utilisant le lemme ,on obtient le

Théoréme :

Asymptotiquement en ν ,le coût moyan d'insertion ou de suppression d'une clé dans un fichier est ,sous l'hypothése P,

$$I^{(P)}(\nu) = \frac{\nu}{b.Ln(p)}[\ 1 + N_{b+1}(Log_p \nu) + O(\nu^{-m})\ ] .$$

(Avec les mêmes notations que précédemment) .

On voit donc que ce coût est une fonction périodique de $Log_p \nu$,de période de 1,et de valeur moyenne $\nu/(b.Ln(p))$.

IV-4-Longueur du cheminement externe .

Désignons par $L_b^{(P)}(\nu)$ ce nombre pour des arbres dont les boîtes associées aux feuilles sont de capacité b. Nous avons montré que

$L_b(x) = e^x \sum_k p^k(x/p^k).(e^{x/p^k} - e_{b-1}(x/p^k)).e^{-x/p^k}$. Soit

$L_b(x) = e^x.x.\sum_k (1- e^{-x/p^k}) - \sum_k \sum_{j=1}^{b-1} \frac{1}{j!} x^j.p^{-jk}).e^{-x/p^k})xe^x$. D'où

$$L_b(x) = L_1(x) - x.e^x. \sum_{j=1}^{b-1} (x^j/j!). \sum_k p^{(0-j)k}.e^{-x/p^k}$$

$$= L_1(x) - x.e^x. \sum_{j=1}^{b-1} (x^j/j!).S_{j+1}(x) \quad , \text{ avec}$$

$$L_1(x) = x.e^x. \sum_k (1 - e^{-x/p^k}) = x.e^x.T(x), \text{avec } T(x) = \sum_k (1 - e^{-x/p^k}) .$$

Il résulte de ces égalités que

$$L_1^{(P)}(\nu) = \nu.T(\nu), \text{ et que}$$

$$L_b^{(P)}(\nu) = L_1^{(P)}(\nu) - \nu. \sum_{j=1}^{b-1} (\nu^j/j!).S_{j+1}(\nu) \quad , \text{ si b est supérieur à 1.}$$

Du fait que l'on sait calculer les $S_j(\nu)$, la détermination des $L_b^{(P)}(\nu)$ se raméne à celle de $T(\nu)$.

Lemme 2:

Asymptotiquement en x, on a

$$T(x) = (\mathrm{Log}_p x + \frac{1}{2} + \frac{\gamma}{\mathrm{Ln}(p)})[\ 1 + U(\mathrm{Log}_p x) + O(x^{-m})\ ] ,$$

où U est une fonction périodique de $\mathrm{Log}_p x$,de période 1, de moyenne nulle , et m un réel positif quelconque.

Preuve:elle est identique à celle de [1], en remplaçant 2 par p.

L'utilisation des deux lemmes conduit au

Théorème

Asymptotiquement en ν,sous l'hypothése P,la longueur du cheminement externe de l'index est une fonction périodique de $\mathrm{Log}_p \nu$,de période 1, de valeur moyenne

$$L_b^{(P)}(\nu) \sim \nu[\ \mathrm{Log}_p \nu + 1/2 + (\gamma/\mathrm{Ln}(p))(1 + \delta_{p>1} . \sum_{j=1}^{b-1} (1/j!)\ )] .$$

ANNEXE

A titre d'exemple, étudions la série $T(x) = \sum_{k=0}^{\infty} (1 - e^{-x/p^k})$.

1) Sa convergence est assurée par le fait qu'asymptotiquement en k

$1 - e^{-x/p^k} \sim x/p^k$, terme général d'une série convergente .

2) Nous allons effectuer une transformation de Mellin sur T(x) :

posons $\alpha_k = 1$, $\beta_k = p^{-k}$, $f(x) = 1 - e^{-x}$.

On a $T(x) = \sum_{k=0}^{\infty} \alpha_k . f(\beta_k x)$.

Pour tout nombre complexe s, la transformée de Mellin $T^*(s)$ de T(x) est

$T^*(s) = \int_0^{\infty} T(x) . x^{s-1} \, dx$. Elle vérifie

$T^*(s) = f^*(s) . \omega(s)$, $\omega(s) = \sum_k \alpha_k . \beta_k^{-s} = \sum_k (p^{-k})^{-s} = \sum_k (p^s)^k$, et

$f^*(s) = - \Gamma(s)$.

ω est convergente pour $Re(s)<0$, et vaut alors $\omega(s) = 1/(1 - p^s)$.

$f^*(s)$ est définie, holomorphe pour $-1 < Re(s) < 0$, et y vaut $-\Gamma(s)$.

On a donc $T^*(s) = - \Gamma(s)/(1 - p^s)$, pour $-1 < Re(s) < 0$.

Cette fonction peut être prolongée analytiquement pour $Re(s) > 0$.

3) pour tout réel c, compris entre -1 et 0, on a par inversion :

$$T(x) = (1/2i\pi).\int_{c-i\infty}^{c+i\infty} T^{*}(s).x^{-s}ds \ .$$

Prenons un contour C pour évaluer cette intégrale: soit m un réel positif, et $t_n = (2\pi+1)n/Ln(p)$, n entier positif tendant vers l'infini. Le contour est délimité par les droites d'abscisses c et m, et les droites d'ordonnées t_n et $-t_n$. Il est parcouru dans le sens direct. On désigne par I_c, I_m, I_{sup} et I_{inf} les intégrales de $T^{*}(s).x^{-s}$ sur ces différentes côtés de ce rectangle parcourus dans le sens direct, et par B l'intérieur de ce rectangle. La fonction à intégrer est analytique à l'intérieur du contour, sauf en ses pôles que nous étudierons plus loin. On peut donc lui appliquer le théorème de Cauchy :

$$I_c + I_m + I_{sup} + I_{inf} = \sum_{s_1 \in S} Res(T^{*}(s).x^{-s}) \ ,$$

où S est l'ensemble des pôles s_1 de $T^{*}(s).x^{-s}$ qui se trouvent dans B.

On constate que $T(x) = -\lim I_c$, pour n tendant vers l'infini.
Nous allons nous employer à calculer la somme des résidus et à montrer le rôle négligeable des autres intégrales.

4) Pôles et résidus de $T^{*}(s).x^{-s}$ dans B:
x^{-s} n'ayant pas de pôle dans B, les seuls pôles à prendre en considération sont ceux de $1/(1-p^{s})$ et ceux de $\Gamma(s)$.

Les pôles de la premiére de ces fonctions vérifient

$p^{s} = 1$, soit $e^{s.Ln(p)} = e^{2il\pi}$, l entier relatif. Donc ce sont les

$s_l = 2il\pi/Ln(p)$, $l \in Z$.

A l'intérieur de B, $\Gamma(s)$ n'a qu'un seul pôle , s = 0.

La fonction $T^*(s).x^{-s}$ a donc un pôle double à l'origine et une infinité de pôles simples $s_l = 2il\pi/Ln(p)$, $l \in Z^*$.

Résidu à l'origine:

au voisinage de s=0, on a $\Gamma(s) = \Gamma(s+1)/s \sim (1 - \gamma.s)/s$, où γ est la constante d'Euler ,([4], p.257 et suivantes).

$$1 - p^s = 1 - e^{s.Ln(p)} \sim -sLn(p)(1 + \frac{s}{2}Ln(p)) ;$$

$$x^{-s} = e^{-s.Ln(x)} \sim 1 - s.Ln(x) .$$

Donc $T^*(s).x^{-s} \sim (1/s^2Ln(p)).(1 - (\gamma + \frac{1}{2}Ln(p) + Ln(x))s)$.

Le résidu en s=0 est le coéfficient de 1/s, soit

$$\rho_0 = -(L_p(x) + 1/2 + \gamma/Ln(p)) .$$

Résidu en chacun des autres pôles s_l :

$$\rho_l = -\Gamma(s_l).x^{-s_l}/(1 - p^s)'_{s=s_l} = \frac{\Gamma(s_l)}{Ln(p)} .x^{-s_l} .$$

5) Etude de I_{sup} et I_{inf} :

Sur la droite d'équation $s = i.t_n$, on a ,

$$1 - p^s = 1 - p^{s_n} .e^{in/Ln(p)} = 1 - e^{in/Ln(p)} = 1 - \cos(n/Ln(p)) + i\sin(Ln(p))$$

Donc $|1 - p^s| = 2.\sin(n/2.Ln(p))$, qui ne peut nul pour des raisons évidentes .Il existe donc un réel poitif d tel que $d <= |1-p^s| <= 2$.

De plus,on sait que ,quand v tend vers l'infini,on a

$\Gamma(u+iv) = O(|v|^{u-1/2}.e^{-(\pi/2)|v|})$.

Donc $\Gamma(it_n) = O(|t_n|^{1/2}.e^{-(\pi/2)|t_n|})$.

Enfin $x^{-s} = x^{-t_n i} = e^{-it_n Ln(x)}$ et $|x^{-s}| = 1$.

Par conséquent $T^x(s).x^{-s}$ est majorée en module par une fonction de t_n qui tend vers 0 quand n tend vers l'infini, et l'on a

$\lim I_{sup} = \lim I_{inf} = 0$, quand n tend vers l'infini.

6) Etude de I_m :

c'est l'intégrale de $T^x(s).x^{-s}$ sur le côté $s = m + iy$, y allant de $-t_n$ à $+t_n$ .On a,ici encore,

$|\Gamma(m+iy)| = O(|y|^{m-1/2}.e^{-(\pi/2)|y|})$;

$p^s = p^m.e^{iyLn(p)}$, $1 - p^s = 1-p^m.\cos(yLn(p)) - i\, p^m \sin(yLn(p))$, donc

$p^m - 1 <= |1 - p^s| <= p^m + 1$;

$x^{-s} = x^{-m}.e^{-iyLn(x)}$, soit $|x^{-s}| = x^{-m}$, pour x positif .

Il s'ensuit que

$|I_m| <= (x^{-m}/(p^m-1)).|\int\Gamma(m+iy)|$.

Or cette dernière intégrale converge vers une limite finie ,quand n tend vers l'infini, compte tenu de la majoration obtenue pour $|\Gamma(m+iy)|$:

Lorsque x tend vers plus l'infini, on voit donc que le module de I_m est un $O(x^{-m})$, quelque soit le réel positif m.

On a donc, en résumé,par passage à la limite, quand n tend vers l'infini,

$T(x) = -\rho_0 + \sum_{l\in Z^*} \rho_l + O(x^{-m})$, m réel positif quelconque, soit encore

$$\sum_{k=0}^{\infty}(1 - e^{-x/p^k}) = L_p(x) + 1/2 + \gamma/Ln(p) + \sum_{l\in Z^*}(1/Ln(p))\Gamma(2il\pi/Ln(p))e^{-2il\pi Log_p x} + O(x^{-m})$$; ou enfin

$$\sum_{k=0}^{\infty}(1- e^{-x/p^k}) = Log_p x + 1/2 + \gamma/Ln(p) + V(Log_p x) + O(x^{-m}),$$

où V est une fonction périodique de $Log_p x$, de période 1 et de moyenne nulle , et m un réel positif quelconque .

CONCLUSION

Ainsi que nous l'avions indiqué, nous avons généralisé au cas des arbres p-homogènes les résultats connus pour le hachage dynamique dans le cas des arbres binaires.

deux constatations s'imposent:

d'une part,méthodologiquement,les outils crées et utilisés pour p=2 ont pu être généralisés sans trop de difficulté à p entier quelconque; d'autre part,au plan des résultats ,le fait de prendre p supérieur à 2 n'améliore pas les performances du hachage.En effet,comme on l'a vu, les valeurs de certains des paramétres se dégradent quand p augmente. Néanmoins,sauf à se résigner à coder en binaire les clés écrites dans un alphabet à p éléments,on peut considérer que les procédés de hachage au moyen d'arbres p-homogènes sont trés utiles.

Pour terminer,nous voudrions suggérer une orientation de recherche,qui, à notre avis,devrait apporter une amélioration des performances des techniques de hachage dynamique : alors que les méthodes actuelles consistent à séparer,de façon aveugle,les clés qui sont rangées dans une boîte attachée à une feuille de niveau k dans l'index,suivant les valeurs de leurs k-iémes chiffres,on pourrait penser à les séparer suivant les valeurs d'un chiffre-pas nécessairement le k-iéme-qui partagerait "le mieux possible!! ,en un sens à préciser,cet ensemble de clés.

BIBLIOGRAPHIE

[1] M. Régnier, "Evaluation des performances du hachage dynamique", Thèse, Université Paris-Sud (1983).

[2] P. Flajolet, M. Régnier, R. Sedgewick : "Some uses of the Mellin integral transform in the analysis of algorithms". Proc. Advanced Worshop on "Combinatorics on Words", Springer NATO ASI Ser. F12, pp 241-254.

[3] P. Flajolet, M. Régnier, D. Sotteau : "Algebraic methods for trie statistics in analysis and design of algorithms for combinatorial problems". Annals of Discrete Math., Vol 25-1985 pp 145-188.

[4] Abramovitz and Stegun : "Handbook of mathematical functions".

[5] R, Fagin., J, Nievergelt, N. Pippenger., H.R. Strong. "Extendible Hashing - a fast access method for dynamifiles". ACM-TODS, 4, 3, (Sep. 1979), 315-344.

[6] P.A. Larson., "Dynamic Hashing" BIT 18, (1978), 184-201.

[7] W. Litwin., "Virtual Hashing : a dynamically changing hashing". VLDB 78. ACM, (Sept. 1978), 517-523.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique