

Hachage Parfait

Julien Devevey

2018-2019

Ref : Cormen - Introduction to Algorithms p.277

Définition 1. Soit $\mathcal{H}$ un ensemble fini de fonctions de hachage d'un ensemble E de clés dans $\llbracket 0, m-1 \rrbracket$. Un tel ensemble est dit *universel* si $\forall k, l \in E, k \neq l$ il y a au plus $\frac{|\mathcal{H}|}{m}$ fonctions de $\mathcal{H}$ pour lesquelles (k, l) est une collision.

Théorème 2. Si $E \subset \llbracket 0, p-1 \rrbracket$ avec p premier et $p > m$, on définit $\mathcal{H}_{p,m} = \{h_{a,b} : k \mapsto ((ak + b) \bmod p) \bmod m \mid a \in \mathbb{Z}_p^*, b \in \mathbb{Z}_p\}$. Alors $\mathcal{H}_{p,m}$ est une famille de fonctions de hachage universelle.

Démonstration.

Soient $k \neq l \in \mathbb{Z}_p$. On pose $r = (ak + b) \bmod p$ et $s = (al + b) \bmod p$. Alors $r - s \neq 0$ comme $k \neq l$. Au niveau de la réduction modulo p , il ne peut donc pas encore y avoir de collisions. Or chacun des $p(p-1)$ choix possibles pour (a, b) donne un résultat différent étant donné qu'on peut retrouver a et b avec :

$$\begin{aligned} a &= ((r - s)(k - l)^{-1} \bmod p) \bmod p \\ b &= (r - ak) \bmod p \end{aligned}$$

Et comme on a $p(p-1)$ différents couples possibles pour (r, s) , on a bien bijection entre les couples (r, s) et (a, b) . Donc en choisissant (a, b) uniformément aléatoirement (ce qui revient à choisir une fonction de $\mathcal{H}_{p,m}$ uniformément aléatoirement), (r, s) suit aussi une loi uniforme. Ainsi la probabilité d'avoir une collision est ramenée à la probabilité de l'évènement " $r = s \bmod m$ " pour r et s tirés uniformément aléatoirement dans $\mathbb{Z}_p^2 - \{(i, i), i \in \mathbb{Z}_p\}$. A r fixé, pour les $p-1$ valeurs restantes pour s , il y a au plus $\lceil p/m \rceil - 1 \leq ((p+m-1)/m) - 1 = \frac{p-1}{m}$ valeurs qui créent une collision. La probabilité d'avoir une collision est donc au final au plus $\frac{1}{p} \frac{p-1}{m} \leq \frac{1}{m}$. $\square$

Définition 3. Un hachage est dit *parfait* si dans le pire cas, une recherche s'effectue en un nombre $O(1)$ d'accès mémoire.

Pour obtenir un tel hachage, on utilise un second niveau de hachage.

Théorème 4. Soit un ensemble de taille n fixée qu'on hache à l'aide d'une fonction $h \in \mathcal{H}_{p,m}$ dans $m = n$ sous-tables de hachage de taille $m_j = n_j^2$, où n_j est

le nombre d'éléments hachés dans la j -ème table, ayant chacune une fonction de hachage $h_j \in \mathcal{H}_{p,m_j}$ associée. En choisissant bien (quitte à les changer) h et les $(h_j)_{j \in \llbracket 0, m-1 \rrbracket}$, on a un hachage parfait qui occupe un espace $O(n)$.

Lemme 5. Soient n clés que l'on stocke dans une table de taille $m = n^2$ en utilisant une fonction $h \in \mathcal{H}_{p,m}$ tirée uniformément aléatoirement. Alors la probabilité d'avoir une collision est au plus $\frac{1}{2}$.

Démonstration.

On a $\binom{n}{2}$ couples de clés qui peuvent entrer en collision. De plus, comme h est choisie dans une classe universelle, la probabilité d'avoir une collision pour chaque couple est de $\frac{1}{m}$. On pose alors X la variable aléatoire qui compte le nombre de collisions dans la table. On regarde son espérance :

$$\begin{aligned}\mathbb{E}[X] &= \binom{n}{2} \times \frac{1}{n^2} \\ &= \frac{n^2 - n}{2} \times \frac{1}{n^2} \\ &< \frac{1}{2}\end{aligned}$$

On applique ensuite l'inégalité de Markov pour conclure :

$$\mathbb{P}(X \geq 1) \leq \mathbb{E}[X] < \frac{1}{2}$$

□

Lemme 6. Soient n clés que l'on stocke dans une table de hachage de taille $m = n$ en utilisant une fonction de hachage h tirée uniformément aléatoirement dans une famille de fonctions de hachage universelle. On a alors avec probabilité au moins $\frac{1}{2}$ sur le choix de h que $\left(\sum_{j=0}^{m-1} n_j^2\right) \leq 4n$.

Démonstration.

Comme on a $\binom{n_j}{2} = \frac{n_j^2 - n_j}{2} \forall j \in \llbracket 0, m-1 \rrbracket$, on peut réécrire :

$$\begin{aligned}\mathbb{E}\left(\sum_{j=0}^{m-1} n_j^2\right) &= \mathbb{E}\left(\sum_{j=0}^{m-1} \left(n_j + 2\binom{n_j}{2}\right)\right) \\ &= \mathbb{E}\left(\sum_{j=0}^{m-1} n_j\right) + 2\mathbb{E}\left(\sum_{j=0}^{m-1} \binom{n_j}{2}\right) \\ &= n + 2\mathbb{E}\left(\sum_{j=0}^{m-1} \binom{n_j}{2}\right)\end{aligned}$$

Or, on remarque que la somme de droite correspond au nombre de collisions dans la table de hachage. Mais grâce aux propriétés du hachage universel, on sait que cette quantité est au plus $\binom{n}{2} \frac{1}{m} = \frac{n(n-1)}{2n} = \frac{n-1}{2}$. On a au final que :

$$\mathbb{E} \left(\sum_{j=0}^{m-1} n_j^2 \right) \leq n + 2 \frac{n-1}{2} \leq 2n$$

On peut alors appliquer l'inégalité de Markov :

$$\mathbb{P} \left(\sum_{j=0}^{m-1} n_j^2 \geq 4n \right) \leq \frac{2n}{4n} = \frac{1}{2}$$

□

Démonstration du théorème.

Le lemme 6 montre qu'en essayant plusieurs fonctions de hachage h , il est possible de trouver facilement une fonction de hachage h dont la table de hachage n'utilise qu'un espace en $O(n)$. Pour chaque table secondaire, le lemme 5 montre qu'en essayant plusieurs fonctions choisies aléatoirement, on en trouve telles qu'il n'y ait aucune collision au second niveau : la recherche s'effectue donc dans le pire cas en $O(1)$. On a alors un hachage parfait. □