



Design and Analysis
of Algorithms I

Data Structures

Universal Hash
Functions: Definition
and Example

Overview of Universal Hashing

Next : details on randomized solution (in 3 parts).

Part 1 : proposed definition of a “good random hash function”.
 (“universal family of hash functions”)

Part 3 : concrete example of simple + practical such functions

Part 4 : justifications of definition : “good functions” lead to “good performance”

Universal Hash Functions

Definition : Let H be a set of hash functions from U to $\{0,1,2,\dots,n-1\}$

H is universal if and only if :

for all x,y in U (with $x \neq y$)

$$Pr_{h \in H} [x, y \text{ collide}] \leq \frac{1}{n} \quad (n = \# \text{ of buckets})$$

ie., $h(x) = h(y)$

When h is chosen uniformly at random from H .

(i.e., collision probability as small as with “gold standard” of perfectly random hashing)

Consider a hash function family H , where each hash function of H maps elements from a universe U to one of n buckets. Suppose H has the following property: for every bucket i and key k , a $1/n$ fraction of the hash functions in H map k to i . Is H universal?

Yes, always.

Yes : Take $H =$ all functions from U to $\{0,1,2,\dots,n-1\}$

No, never.

No : Take $H =$ the set of n different constant functions

Maybe yes, maybe no (depends on the H).

Only if the hash table is implemented using chaining.

Example: Hashing IP Addresses

Let $U =$ IP addresses (of the form (x_1, x_2, x_3, x_4) ,
with each $x_i \in \{0, 1, 2, \dots, 255\}$)

Let $n =$ a prime (e.g., small multiple of # of objects in HT)

Construction : Define one hash function h_a per 4-tuple a
 $= (a_1, a_2, a_3, a_4)$ with each $a_i \in \{0, 1, 2, 3, \dots, n - 1\}$

Define : $h_a : \text{IP addr} \rightarrow \text{buckets}$ by n⁴ such functions

$$h_a(x_1, x_2, x_3, x_4) = \begin{pmatrix} a_1x_1 + a_2x_2 + \\ a_3x_3 + a_4x_4 \end{pmatrix} \text{ mod } n$$

A Universal Hash Function

Define : $H = \{h_a \mid a_1, a_2, a_3, a_4 \in \{0, 1, 2, \dots, n - 1\}\}$

$$h_a(x_1, x_2, x_3, x_4) = \left(\begin{array}{l} a_1x_1 + a_2x_2 + \\ a_3x_3 + a_4x_4 \end{array} \right) \text{ mod } n$$

Theorem: This family is universal

Proof (Part I)

Consider distinct IP addresses $(x_1, x_2, x_3, x_4), (y_1, y_2, y_3, y_4)$.

Assume : $x_4 \neq y_4$

Question : collision probability ?

(i.e., $Prob_{h_a \in H}[h_a(x_1, \dots, x_4) = h_a(y_1, \dots, y_4)]$)

Note : collision \iff

$$a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 = a_1y_1 + a_2y_2 + a_3y_3 + a_4y_4 \pmod{n}$$

$$\iff a_4(x_4 - y_4) = \sum_{i=1}^3 a_i(y_i - x_i) \pmod{n}$$

Next : condition on random choice of a_1, a_2, a_3 . (a_4 still random)

Proof (Part II)

The Story So Far : with a_1, a_2, a_3 fixed arbitrarily, how many choices of a_4 satisfy

$$a_4(x_4 - y_4) = \sum_{i=1}^3 a_i(y_i - x_i) \pmod{n}$$

Still random

$\iff x, y$ collide under h_a

Some fixed number in $\{0, 1, 2, \dots, n-1\}$

Key Claim : left-hand side equally likely to be any of $\{0, 1, 2, \dots, n-1\}$

Reason : $x_4 \neq y_4$ ($x_4 - y_4 \neq 0 \pmod{n}$)

n is prime, a_4 uniform at random

[addendum : make sure n bigger than the maximum value of an a_i]

$\implies \text{Prob}[h_a(x) = h_a(y)] = 1/n$

“Proof” by example : $n = 7$, $x_4 - y_4 = 2$ or $3 \pmod{n}$

Q.E.D.