Hashing
Data Distribution

D.E ZEGOUR

École Supérieure d'Informatique

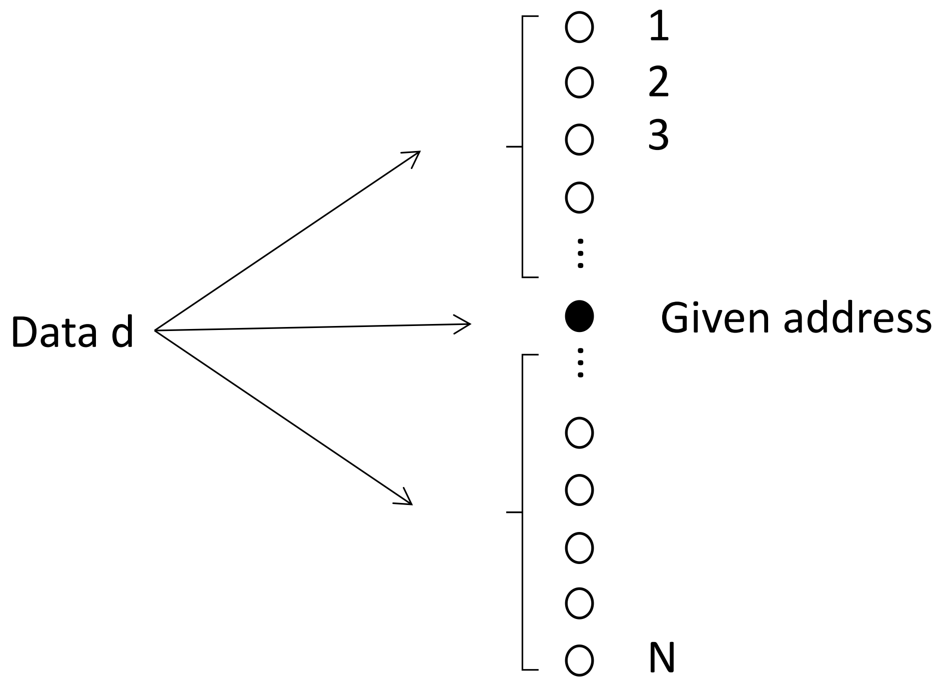
ESI

Hashing / Data Distribution

Poisson Distribution

Suppose

- N possible addresses
- A given address
- Data d to hash

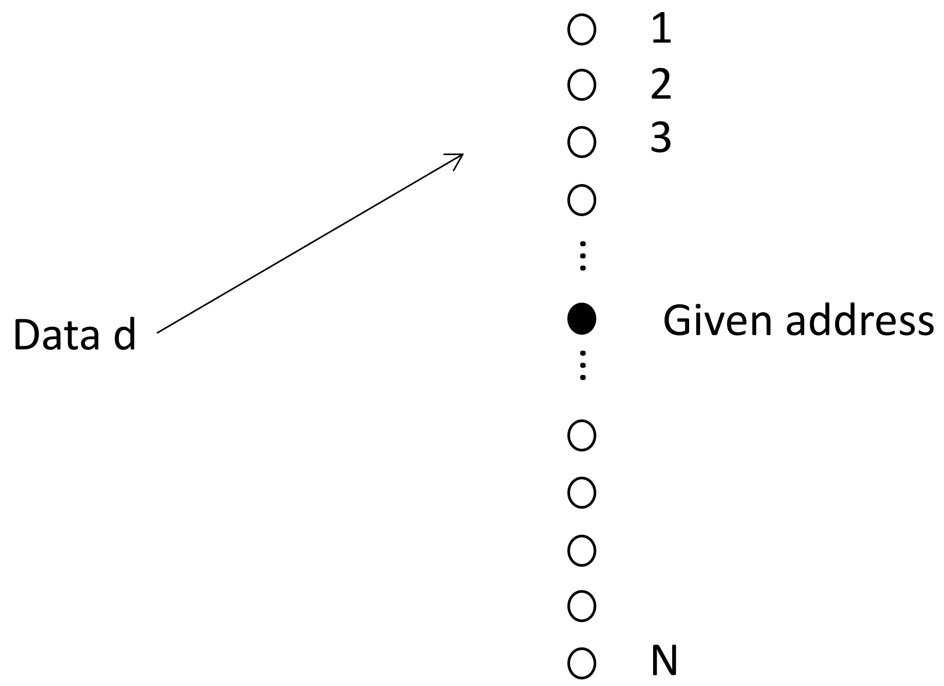


Let’s consider the two events:
A : A given address is not selected
B : A given address is selected

When data is hashed, one of the events (A or B) occurs for a given address.

Hashing / Data Distribution

Poisson Distribution



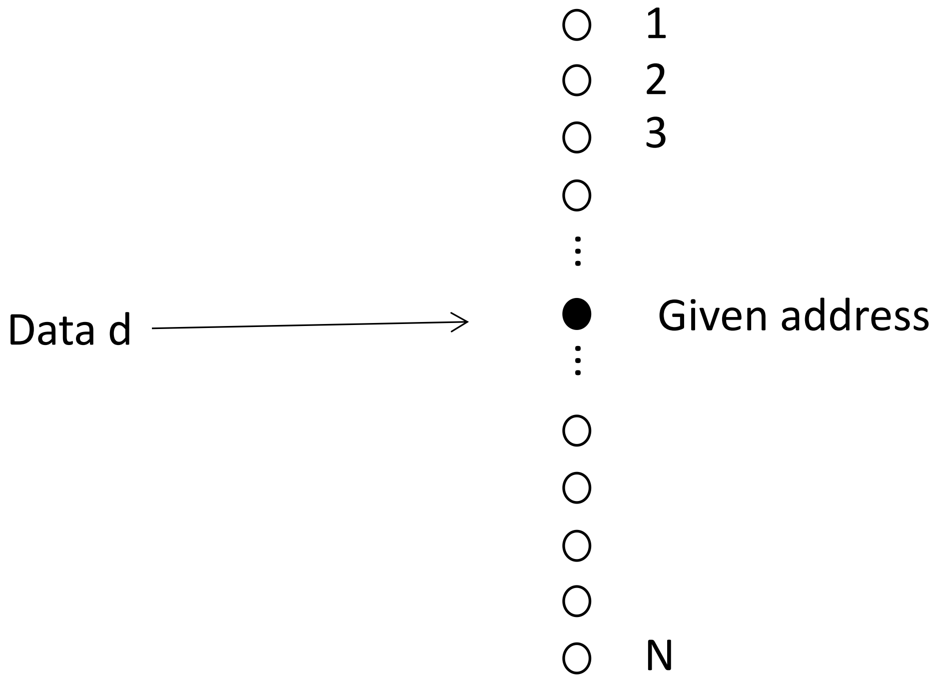
What is the probability that a given address is not chosen?

Event : A

P(A) = a = 1 - 1/N = (N-1)/N= a
If N=10 then a=0.9

Hashing / Data Distribution

Poisson Distribution



What is the probability that a given address is chosen?

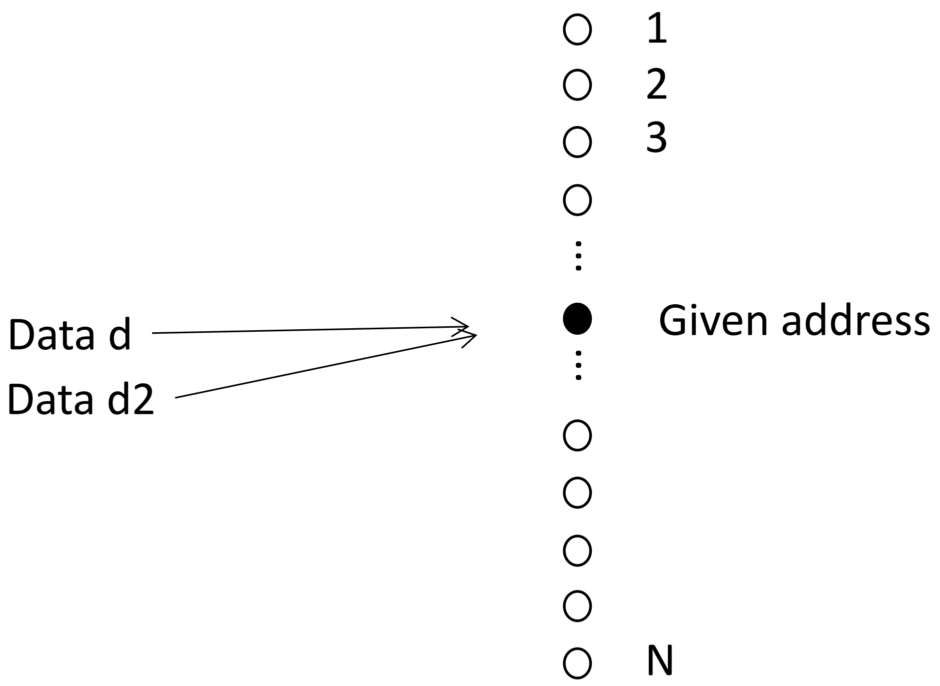
Event : B

P(B) = b = 1/N

If N=10 then a=0.1

Hashing / Data Distribution

Poisson Distribution



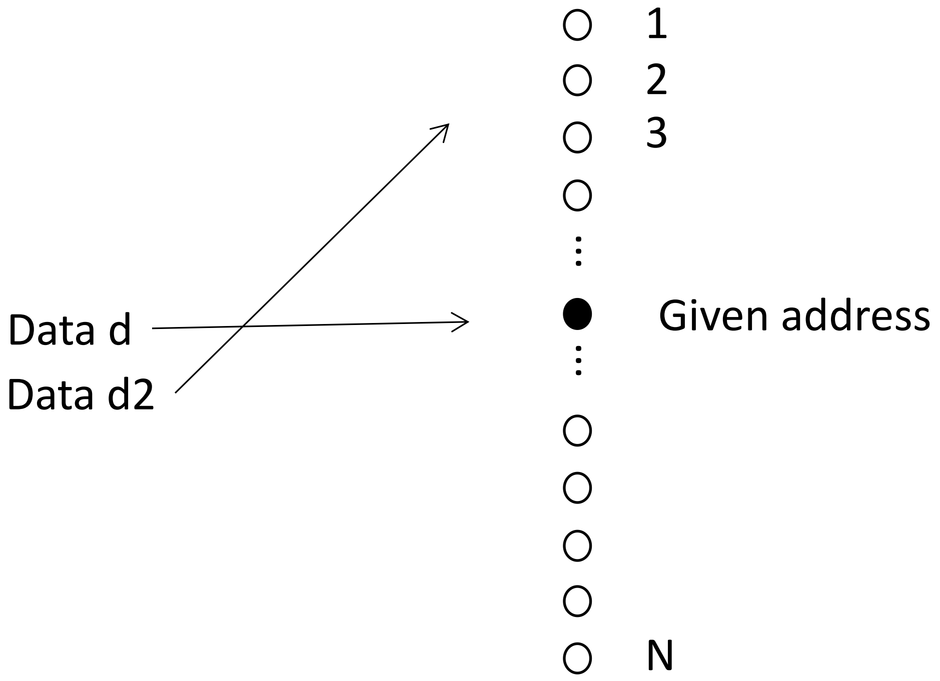
What is the probability that two data hash to the same address?

Event : B B

P(BB) = b.b = 1/N . 1/N (Independent events)

Hashing / Data Distribution

Poisson Distribution



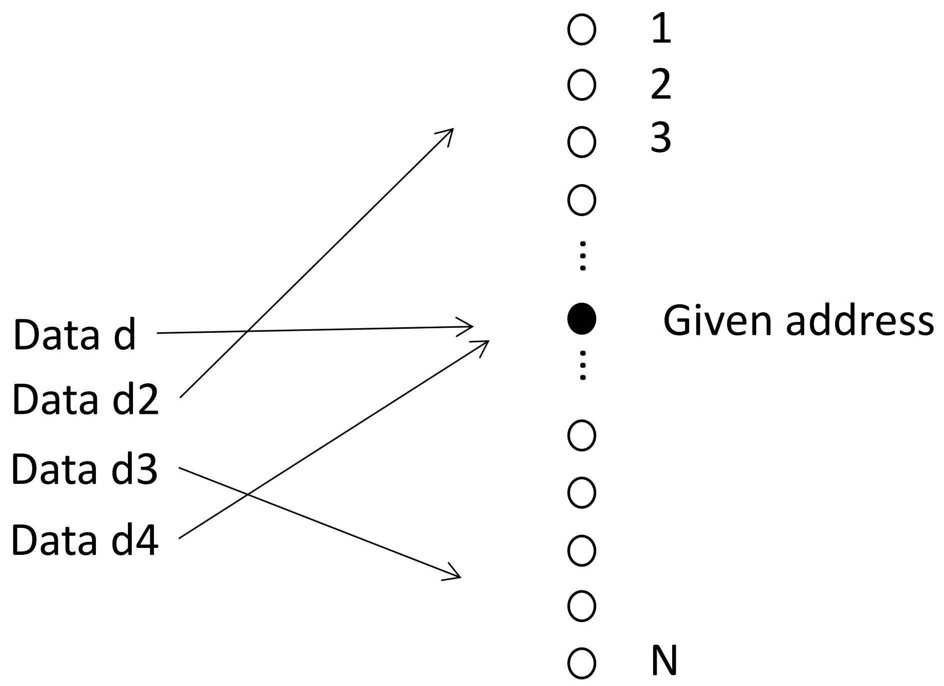
What is the probability that the first data is hashed to the given address and the second data is hashed to an address different from the first one?

Event : B A

P(BA) = b.a = 1/N . (N-1/N)

Hashing / Data Distribution

Poisson Distribution



What is the probability associated with the event BAAB?

P(BAAB) = b.a.a.b = a^2b^2 = $(1/N)^2$ $(N-1/N)^2$

Hashing / Data Distribution

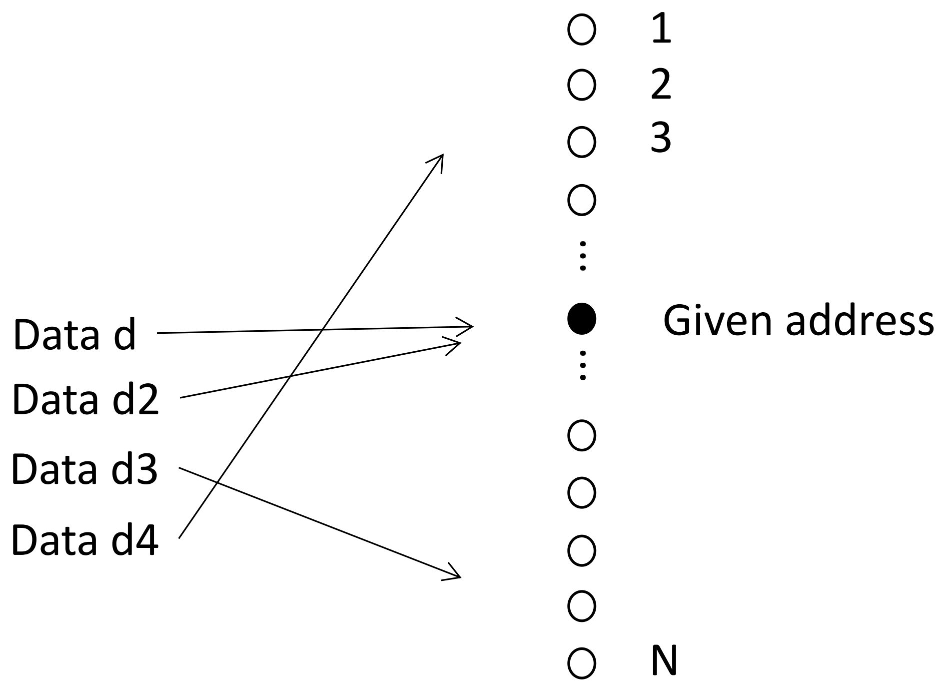
Poisson Distribution

Probability that two out of four data hash to the same address?

Find all the possible events:
BBAA,   BAAB, BABA, AABB, ABBA, ABAB

Hashing / Data Distribution

Poisson Distribution

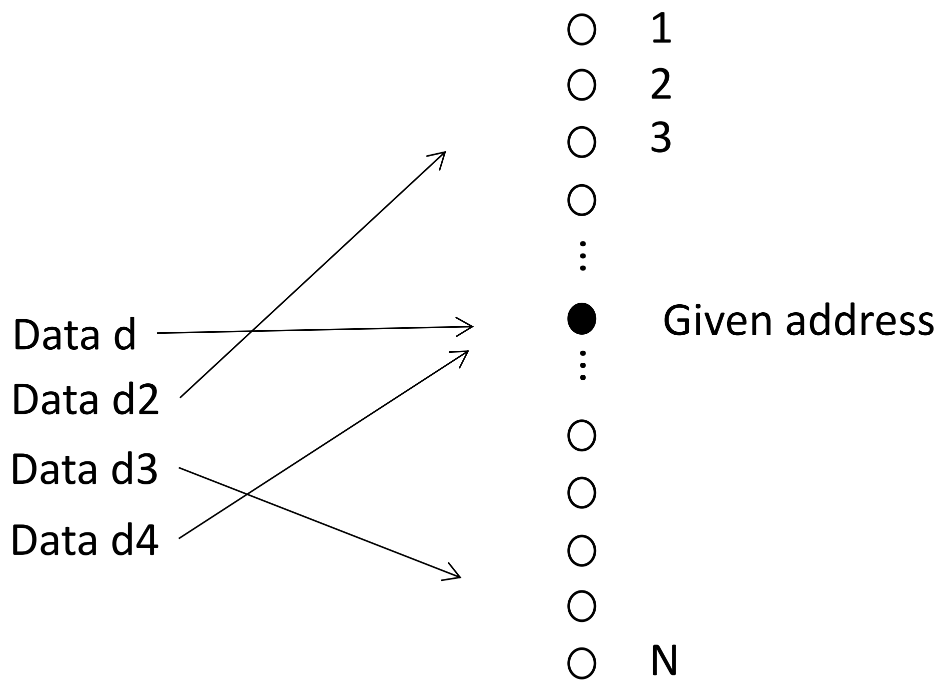


Probability that two out of four data hash to the same address?

Find all the possible events:
BBAA, BAAB, BABA, AABB, ABBA, ABAB

Hashing / Data Distribution

Poisson Distribution



Probability that two out of four data hash to the same address?

Find all the possible events:
BBAA, BAAB, BABA, AABB, ABBA, ABAB

Hashing / Data Distribution

Poisson Distribution

Probability that two out of four data hash to the same address?

Find all the possible events:
AABB, BAAB, BABA, BBAA, ABBA, ABAB

P= P(AABB) + P(BAAB) + ... = 6 a2b2
P= C_4^2 a^2 b^2.

C_4^2 represents the number of ways two A's (and two B's) can be placed in four slots.

Hashing / Data Distribution

Poisson Distribution

Generalisation
Probability that x out of r data hash to the same address?

x times the event B and (r-x) times the event A

All the possible events : C_r^x

$$P(x) = C_r^x \cdot a^{r-x} \quad b^x$$

with $C_r^x = r! / (x! (r-x)!)$

This also means: Probability that a given address is chosen x times and not chosen r-x times.

Hashing / Data Distribution

Poisson Distribution

Calculation:
If N possible addresses
$P(x) = C_r^x \; a^{r-x} \; b^{x} = P(x) = C_r^x \; (1-1/N)^{r-x} \; (1/N)^{x}$

P(x=0) means the probability that a given address is never chosen.
$P(0) = C_r^0 \, (1-1/N)^r \, (1/N)^0$

P(x=1) means the probability that a given address is chosen only once.$P(1) = C_r^1 \, (1-1/N)^{r-1} \, (1/N)^1$

Drawback of the formula: difficult to calculate for large N and r.
The POISSON function is a good approximation.

$P(x) \; =/ \; f(x) = (\; (r/N)^{x} \, . \; e^{-(r/N)} \;) \; / \; x!$

Hashing / Data Distribution

Synthesis

If N is the number of possible addresses, r is the number of inserted data, and x is the number of data with the same address (x times event B and r-x times event A in C_r^x possible ways),

P(x) gives the probability that x data out of r inserted ones hash to the same address.

P(x) Probability that a given address is chosen x times and not chosen r-x times.

Hashing / Data Distribution

Synthesis

If N is the number of possible addresses, r is the number of inserted data, and x is the number of data with the same address (x times event B and r-x times event A in C_r^x possible ways),

P(x) is also the proportion of addresses with x data assigned to them by hashing.

N.P(x) is the number of addresses that have x data assigned.

This allows us to predict the number of collisions (overflow data).

Formula : N (P(2) + 2P(3) + ...iP(i+1) + ...)

Hashing / Data Distribution

Collision prevention

Let's consider N=10,000 possible addresses and r=10,000 inserted data.

What is the number of addresses that have no data assigned?
10,000 P(0) = 3,679

What is the number of addresses that have only one data assigned?
10,000 P(1) = 3,679

What is the number of addresses that have only two data assigned?
10,000 P(2) = 1,839
→ 1,838 will be in overflow.

What is the number of addresses that have only three data assigned?
10,000 P(3) = 613
→ 613 * 2 will be in overflow.

Not ideal distribution : we have thousands of addresses (3,679) with no data assigned.

More than 1839 + 1226 (= 2 * 613) data will be in overflow.

Hashing / Data Distribution

Collision reduction

Let's demonstrate, using examples, how,

on the one hand,
increasing the number of possible addresses

and, on the other hand,
using boxes,

can reduce collisions.

Hashing / Data Distribution

Increase of the address space

We define the density d as follows: d = r / N (r: number of stored data; N: number of possible addresses).

Let's examine the behavior of hash functions (collisions) for different values of d.

$$P(x) = \left((r/N)^{x} \; e^{-(r/N)} \right) / x! = \left(d^{x} \,.\, e^{-d} \right) / x!$$

P(x) depends on the ratio r/N, that is, on d.

Also, we observe the same behavior for 500 data distributed among 1000 addresses as for 500,000 data distributed among 1 million addresses (d = 50% for both cases).

Hashing / Data Distribution

Increase of the address space

Let's take d = 0.5 (N = 1000 and r = 500 data) with P(0) = 0.607; P(1) = 0.303; P(2) = 0.0758; P(3) = 0.0126; P(4) = 0.0016; etc.

How many addresses will have 0 data assigned?
1,000 * P(0) = 607

How many addresses will have 1 data assigned?
1,000 * P(1) = 303

How many addresses will have 2 data assigned?
1,000 * P(2) = 76

How many addresses will have at least two data assigned?
→ 1,000 * (P(2) + P(3) + P(4) + ...) = 90

Hashing / Data Distribution

Increase of the address space

What is the number of data in overflow?
→ 1,000 * (P(2) + 2 * P(3) + 3 * P(4) + 4 * P(5) ......) = 107

What is the percentage of data in overflow?
→ 107/500 = 21.4%

Conclusion:
If the density is 50%, we can expect 78.6% of data stored in their primary address and 21.4% stored elsewhere.

Hashing / Data Distribution

Use of the boxes (b > 1)

We accept b data per possible address.

In this case: $d = r / (b.N)$

where:
b: number of data per slot
N: number of addresses
r: number of inserted data.

	b=1	b=2
Number of data (r)	750	750
Number of addresses (N)	1000	500
Density (d)	0.75	0.75
Ratio (r/N)	0.75	1.5

Hashing / Data Distribution

Use of the boxes (b > 1)

P(x)	b=1 (r/N = 0.75)	b=2 (r/N = 1.5)
P(0)	0.423	0.223
P(1)	0.354	0.335
P(2)	0.113 (Collisions)	0.251
P(3)	0.033 (Collisions)	0.126 (Collisions)
P(4)	0.006 (Collisions)	0.047 (Collisions)

Number of data in overflow in each case?

b=1:
1 000.[P(2) + 2.P(3) + 3.P(4) + ...]> 197

b=2
5 00.[P(3) + 2.P(4) + 3 P(5) + ...] > 110

The larger the box, the better the performance.