Hashing (4)
Cuckoo Hashing
Cuckoo Hashing

A simple scheme for resolving collisions in a hash table
Guaranteed constant time lookup
Expected constant time insertion
Requires stronger assumption for hash functions
We will work with Uniform Hashing Assumption
We will present a simplified version of the scheme, and a simplified analysis
We have a table $T[0..m-1]$ of $m$ slots

Each slot is either null or contains a single data item

Data items are hashed using two hash functions:

$h_1, h_2 : \mathcal{U} \rightarrow \{0, \ldots, m-1\}$

We model each hash function as a truly random function from $\mathcal{U}$ to $\{0, \ldots, m-1\}$

Any data item $a \in \mathcal{U}$ resides in one of two slots: either $h_1(a)$ or $h_2(a)$

Lookup procedure:

- test if $T[h_1(a)] = a$ or $T[h_2(a)] = a$

$\implies$ guaranteed constant time lookup
Procedure to insert a new data item $a$

```plaintext
let $n = \#$ items already in the table
if $T[h_1(a)] = a$ or $T[h_2(a)] = a$ then
    return success  // already in table

pos ← $h_1(a)$
repeat $n$ times
    if $T[pos] = \text{null}$ then
        $T[pos] ← a$
        $T[pos] ← a$
        return success  // found an empty slot

swap $a$ and $T[pos]$

pos ← $h_1(a) + h_2(a) - pos$
// $a$’s alternate position

return failure
// need to rehash
```
The cuckoo graph:

Each slot is a vertex

Each data item \( a \) adds a random edge
\[
e = \{h_1(a), h_2(a)\}
\]
to the graph

- undirected graph
- possibly a multi-graph — repeated edges
Arrows show alternate position of each item

Insert $Z$ at position 0: $0 \rightarrow A \rightarrow 3 \rightarrow B \rightarrow 8$

No cycle $\implies$ no problem!

Insert $Z$ at $h_1(Z) = 7$ ($h_2(Z) = 1$):

$7 \rightarrow W \rightarrow H \rightarrow 7 \rightarrow Z \rightarrow 1 \rightarrow C \rightarrow 2$

Insert $Z$ at $h_1(Z) = 7$ ($h_2(Z) = 4$):

$7 \rightarrow W \rightarrow H \rightarrow 7 \rightarrow Z \rightarrow W \rightarrow H \rightarrow 4 \ldots$

only two slots for three items $\implies$ failure
Lessons learned:

- If a new item is inserted at slot $s$, and there is no cycle in the graph reachable from $s$, then insertions will succeed.
- In particular: if there are no cycles, insertion will succeed.
- Even if there are cycles, insertion may succeed:
  - The exact characterization of failure is a bit more complicated.
Analyzing the probability of insertion failure

We will show that if \( \alpha := n/m \) (the “load factor”) is at most 1/4, then the probability that inserting \( n \) items into a table with \( m \) slots ends in failure is at most 3/4.

**How?** Compute bound on probability \( p \) of a cycle in a multi-graph with \( m \) vertices (slots) and \( n \) random edges (data items) \( e_1, \ldots, e_n \)

**Strategy:** for each \( k = 1, 2, 3, \ldots \), estimate probability \( p_k \) that graph contains a simple cycle of length \( k \)

**Union bound:** \( p \leq \sum_{k \geq 1} p_k \)

**NOTE:** a more careful analysis shows failure probability is much smaller: \( O(1/m) \)
**Typical case:** $p_3 :=$ probability of a 3-cycle:

\[
\begin{array}{c}
\text{s}_1 \\
\text{s}_0 \quad \text{s}_2 \\
\end{array}
\]

$\leq m^3$ ways to pick $s_0, s_1, s_2$, but we count the same cycle 3 times

$\therefore$ # of 3-cycles: $\leq m^3/3$

Probability that

\[
(e_{i_1}, e_{i_2}, e_{i_3}) = (\{s_0, s_1\}, \{s_1, s_2\}, \{s_2, s_0\})
\]

is $(2/m^2)^3$

# of triples $i_1, i_2, i_3$: $\leq n^3$

$\therefore p_3 \leq (m^3)/3 \times (2/m^2)^3 \times n^3 = (2n/m)^3/3$
The general case (exercise):

\[ p_k \leq \frac{(2\alpha)^k}{k}, \quad \text{where } \alpha := \frac{n}{m} \]

Therefore,

\[ p \leq \sum_{k \geq 1} p_k \leq \sum_{k=1}^{\infty} \frac{(2\alpha)^k}{k} = \ln \left( \frac{1}{1 - 2\alpha} \right) \]

Wolfram Alpha says: \( x \leq 1/2 \implies \ln(1/(1 - x)) \leq 3/4 \)

Implication:

\( \alpha \leq 1/4 \implies \text{failure probability} \leq 3/4 \)
Building a cuckoo hash table

Suppose we attempt to insert \( n \) distinct items \( a_1, \ldots, a_n \) items into an empty hash table, and stop when an insertion fails.

For \( r = 1 \ldots n \), let \( X_r \) be the number of swaps performed when we attempt to insert \( a_r \).

Note: \( X_r = 0 \) if the insertion procedure fails on one of \( a_1, \ldots, a_{r-1} \).

Assume that \( \alpha := n/m \leq 1/4 \).

**Claim:** \( \mathbb{E}[X_r] \leq 3/2 \).

It follows that

- Expected cost of attempting to insert \( n \) items: \( O(n) \)
- Probability that such an attempt succeeds: \( \geq 1/4 \)
- Expected number of attempts until success: \( \leq 4 \)
- Expected cost of building a table: \( O(n) \)
**Claim:** $E[X_r] \leq 3/2$

**Proof:**

Suppose the $h_1(a_r) = s_0$ and consider the cuckoo graph corresponding to items $a_1, \ldots, a_{r-1}$

We have

$$E[X_r] = \sum_{k=1}^{n} \Pr[X_r \geq k]$$

If $X_r \geq k$, then in the cuckoo graph: either

(i) there is a *simple path* of length $k$ starting at $s_0$:

$$s_0 \rightarrow s_1 \rightarrow \cdots \rightarrow s_k,$$

or

(ii) there is a *simple loop* starting at $s_0$:

$$s_0 \rightarrow \cdots \rightarrow s_{\ell-1} \rightarrow s_j \ (j < \ell)$$
Let’s estimate the probability \( q_k \) that there is a simple path of length \( k \) starting at \( s_0 \):

\[
    s_0 \rightarrow s_1 \rightarrow \cdots \rightarrow s_k
\]

# of choices for \( s_1, \ldots, s_k \): \( \leq m^k \)

Probability that

\[
    (e_{i_1}, \ldots, e_{i_k}) = (\{s_0, s_1\}, \ldots, \{s_{k-1}, s_k\})
\]

is \( (2/m^2)^k \)

# of tuples \( i_1, \ldots, i_k \): \( \leq n^k \)

Therefore,

\[
    q_k \leq m^k \times (2/m^2)^k \times n^k = (2n/m)^k = (2\alpha)^k
\]

\[
    \leq 2^{-k} \quad \text{(since } \alpha \leq 1/4)\]
Let’s estimate the probability $\tilde{q}_\ell$ that there is a simple loop of length $\ell$ starting at $s_0$

$$s_0 \rightarrow \cdots \rightarrow s_{\ell-1} \rightarrow s_j \quad (j < \ell)$$

Homework:

$$\tilde{q}_\ell \leq \frac{\ell(2\alpha)^\ell}{m}$$

Let $\tilde{q} :=$ probability of any simple loop starting at $s_0$:

$$\tilde{q} \leq \sum_{\ell \geq 1} \tilde{q}_\ell \leq \frac{1}{m} \sum_{\ell=1}^{\infty} \ell(2\alpha)^\ell = \frac{1}{m} \cdot \frac{2\alpha}{(1 - 2\alpha)^2} \leq \frac{2}{m} \quad (\text{since } \alpha \leq 1/4)$$
Putting it all together:

\[ E[X_r] = \sum_{k=1}^{n} \Pr[X_r \geq k] \leq \sum_{k=1}^{n} (2^{-k} + 2/m) \leq 1 + \frac{2n}{m} = \frac{3}{2} \]