Perfect hashing

No collisions in the end

f injective: $x \neq y \Rightarrow f(x) \neq f(y)$

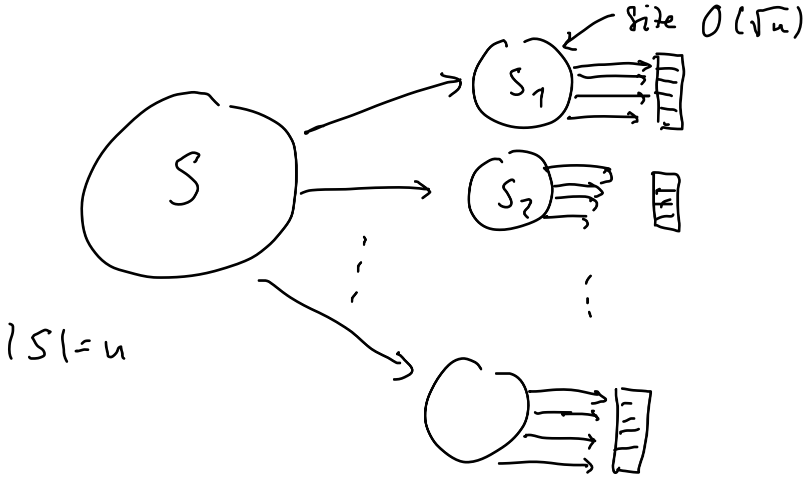
Choose a hash function that is injective (i.e. one-to-one) on the set *S* to be stored. (Assumption: *S* is known in advance.)

Can be removed

Idea

Two-level hashing scheme

1. In the first level, *S* is partitioned into "short lists" (hashing with chaining).
2. In the second level for each list, a separate injective hash function is used.



Construction of injective hash functions

Let $U = [0\ldots N\text{-}1]$, $S \subseteq U$, $|S| = n$, $|T| = m$

For $k \in \{1,\ldots,N\text{-}1\}$, let

$$h_k : U \rightarrow \{0,\ldots,m\text{-}1\}$$
$$x \rightarrow ((kx) \bmod N) \bmod m$$

Let $S \subseteq U$. Is it possible to choose k such that h_k restricted to S is injective?

h_k restricted to S is injective if for all $x,y \in S$, $x \neq y$,

$$h_k(x) \neq h_k(y)$$

A measure for the violation of injectivity

For $\underbrace{0 \le i \le m\text{-}1}_{\text{table pos.}}$ and $\underbrace{1 \le k \le N\text{-}1}_{\text{function parameter for } h_k}$ let

$$\rightarrow \qquad b_{ik} = |\{\, x \in S : h_k(x) = i \,\}|$$

Then: $(x,y) \in S \times S = S^2$

$$\rightarrow \qquad |\{\, (x,y) \in S^2 : \; x \neq y \; \text{ and } h_k(x) = h_k(y) = i \,\}| \; = \; b_{ik}\,(b_{ik} - 1)$$

ordered pairs

Define

$$\rightarrow \qquad B_k = \sum_{i=0}^{m-1} b_{ik}\,(b_{ik} - 1)$$

$$= 2 \cdot \#\text{ collisions caused } h_k$$

B_k measures to which extent h_k restricted to S is not injective.

Injectivity

Choose $k \in \{1, \dots, N-1\}$

Lemma 1: h_k restricted to S is injective $\Leftrightarrow$ $B_k < 2$

Proof:

($\Leftarrow$) $B_k < 2 \Rightarrow B_k \leq 1 \Rightarrow b_{ik}(b_{ik} - 1) \in \{0,1\}$ for all i

$\Rightarrow b_{ik} \in \{0,1\} \Rightarrow h_k$ restricted to S is injective

$b_{ik} \cdot (b_{ik} - 1) = 0 \Rightarrow B_k = 0$

($\Rightarrow$) h_k restricted to S is injective $\Rightarrow b_{ik} \in \{0,1\}$ for all $i \Rightarrow b_{ik}(b_{ik}-1) = 0$

$\Rightarrow B_k = 0 \Rightarrow B_k < 2$

Injectivity

Lemma 2: Let N be a prime number, $S \subseteq U = [0\ldots N\text{-}1]$ with $|S| = n$.
Then $k = 1, \ldots, N-1$ $|T| = m$

$$\sum_{k=1}^{N-1} B_k \leq 2\frac{n(n-1)}{m}(N-1)$$

$\Rightarrow$ If $m > n(n\text{-}1)$, then there exists B_k with $B_k < 2$,
i.e. there is an h_k that is injective on S.

$$\underbrace{\sum_{k=1}^{N-1} B_k < 2 \cdot (N-1)}_{\text{lemma}} \Rightarrow \exists\, B_k < 2 \Rightarrow \exists\, k \quad h_k \text{ is injective}$$

Proof of Lemma 2

$$\sum_{k=1}^{N-1} B_k = \sum_{k=1}^{N-1}\sum_{i=0}^{m-1} b_{ik}(b_{ik}-1)$$

$$= \sum_{k=1}^{N-1}\sum_{i=0}^{m-1} |\{(x,y)\in S^2 : x \neq y, h_k(x) = h_k(y) = i\}|$$

$$= \sum_{\substack{(x,y)\in S^2\\ x\neq y}} |\{k : h_k(x) = h_k(y)\}|$$

Let $(x,y) \in S^2$, $x \neq y$, be fixed. How many k exist with $h_k(x) = h_k(y)$?

Proof of Lemma 2

$$h_k(x) = h_k(y)$$

Def h a

$$\Leftrightarrow ((kx) \bmod N) \bmod m = ((ky) \bmod N) \bmod m$$

$$\Leftrightarrow (kx \bmod N - ky \bmod N) \underline{\bmod m} = 0$$

$$\Leftrightarrow k(x-y) \bmod N = cm \qquad c \in \mathbb{Z}$$

$q = k(x\text{-}y) \bmod N$, $\qquad q' = k' \cdot (x-y) \bmod N \qquad$ (without mod m)

-- different $\underline{k, k'}$ yield different q, q'.

$k(x\text{-}y) \bmod N = q \qquad k'(x\text{-}y) \bmod N = q \Rightarrow (k-k') \cdot (x-y) \bmod N = 0$

N is prime, $k, k' \in \{1, \ldots, N-1\}$, $k \neq k'$

$|k-k'| < N$

$(k\text{-}k') \cdot (x\text{-}y) = \underline{c'N} \qquad c' \in \mathbb{Z}$

$x, y \in \{0, \ldots, N-1\} \quad x \neq y$

$|x-y| < N$

neither $|k-k'|$ nor $|x-y|$ is a multiple of N ↯

-- only $\lceil (N\text{-}1)/m \rceil$ many q are mapped into the $\underline{\text{same}}$ (with mod m) residue class $\underline{\bmod m}$

Results

Corollary 1: There are at least (N-1)/2 many k with $\boxed{B_k \le 4n(n-1)/m.}$
Such a k can be determined in expected time O(m+n).

Proof: Suppose that there are less than (N-1)/2 many k with

$$B_k \le 4n(n-1)/m.$$

Then there are at least (N-1)/2 many k with $\boxed{B_k > 4n(n-1)/m}$

$$\Rightarrow \sum_{k=1}^{N-1} B_k > \frac{N-1}{\cancel{2}} \cdot \frac{\overset{2}{\cancel{4}}n(n-1)}{m} = \frac{N-1}{m} 2n(n-1) = 2 \cdot \frac{n(n-1)}{m} \cdot (N-1)$$

lemma ↯

With probability ≥ ½ , a k chosen at random fulfills the condition. The expected number of trials is ≤ 2.

Try all keys from S
compute the b_{ik} } $O(n)$

Check all entries
in T } $O(m)$

Results

Corollary 2:

a) Let $\boxed{m = 2n(n\text{-}1)+1.}$ Then at least $(N\text{-}1)/2$ of the h_k are injective on S.
Such an h_k can be found in expected time $O(m+n)=O(n^2)$.

$m = 2\cdot n(n-1)+1$

$$\sum_{k=1}^{N-1} B_k \le 2\cdot \frac{n\cdot(n-1)}{\cancel{m}}\cdot(N-1) < \text{[illegible]}\cdot(N-1) \qquad B_k = \begin{cases} \geq 2 \\ 0 \end{cases}$$

b) Let $\boxed{m = n.}$ Then for at least $(N\text{-}1)/2$ of the h_k it holds that $\boxed{B_k \le 4(n\text{-}1).}$
Such an h_k can be found in expected time $O(n)$.

Recall $\quad B_k = \sum_{i=0}^{m-1} b_{ik}(b_{ik}-1)$

Claim $\quad$ If $B_k \le 4(\cancel{m}\; n-1)$ then all $b_{ik} \le 3\cdot\sqrt{n}$

Proof $\quad$ Let i be $b_{ik} > 3\sqrt{n}$

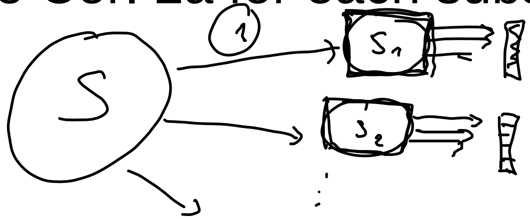
$$B_k \geq b_{ik}(b_{ik}-1) > 3\sqrt{n}\cdot(3\sqrt{n}-1) = 9n - 3\sqrt{n}$$
$$\geq 6n > 4(n-1)$$

Two-level scheme

$|T|$

$S \subseteq U = [0\ldots N\text{-}1]$ $|S| = n$ $m = O(n)$

Idea: Use Corollary 2b and divide S into subsets of size $O(\sqrt{n})$.
Use Cor. 2a for each subset.



1. Choose k with $B_k \leq 4(n\text{-}1) \leq 4n$. $m = n$ Cor 2.b
   $h_k : x \rightarrow ((kx) \bmod N) \bmod n$ Cor 2.a
2. $W_i = \{ x \in S : h_k(x) = i \}$, $b_i = |W_i|$, $m_i = 2b_i(b_i - 1)+1$ for $0 \leq i \leq n\text{-}1$
   Choose k_i such that

$$h_{k_i} : x \longrightarrow (k_i x \bmod N) \bmod m_i$$

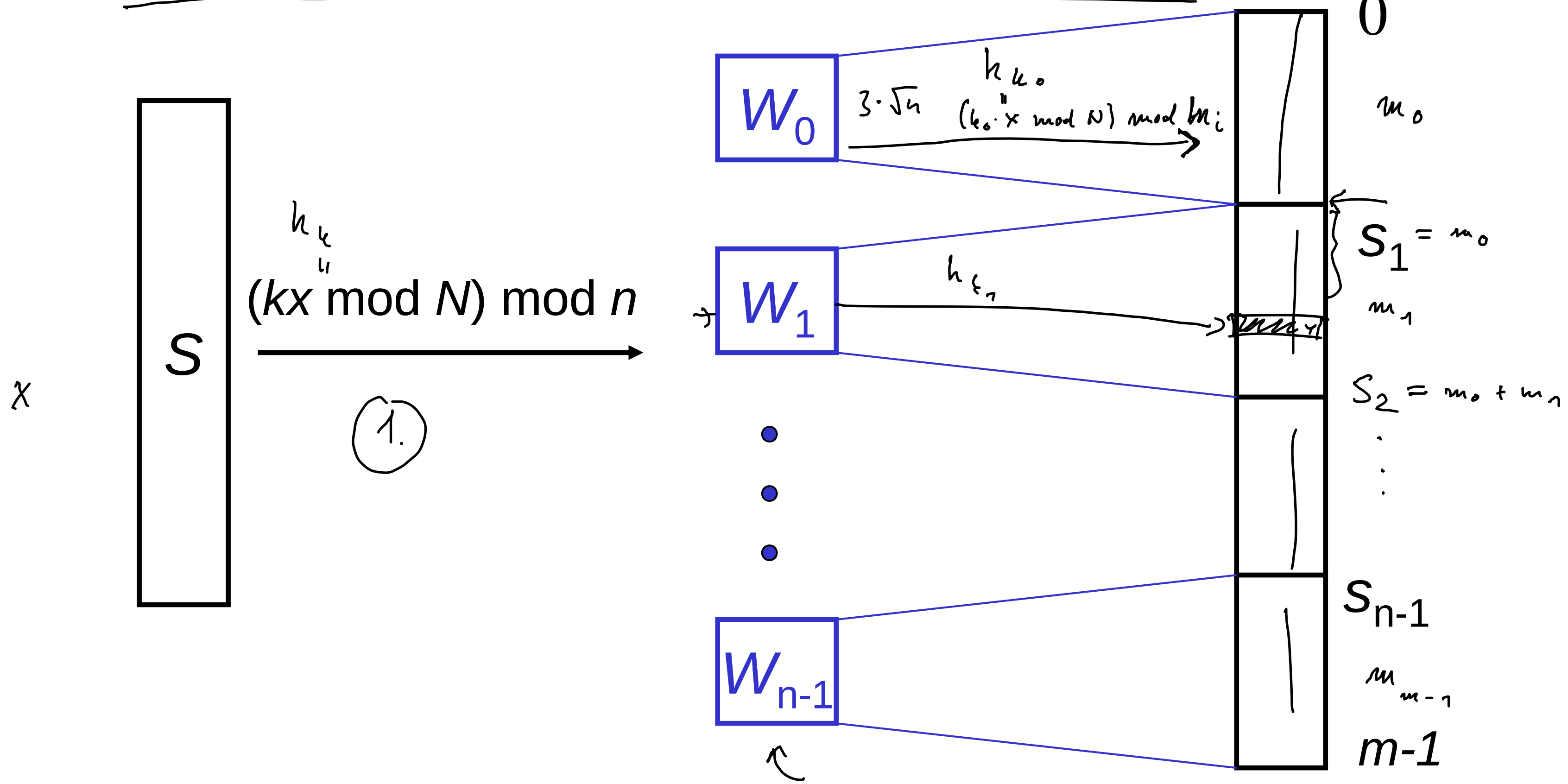
restricted to W_i is injective.

Two-level scheme

3. $s_i = \sum_{j<i} m_j$

Store $x \in S$ in table position $T[s_i + j]$ where

$i = (k\,x \bmod N) \bmod n$ $j = (k_i\,x \bmod N) \bmod m_i$



Space required for hash table and functions

$$m = \sum_{i=0}^{n-1} m_i = \sum_{i=0}^{n-1} (2b_i(b_i - 1) + 1) = n + 2B_k$$

$$\le n + 8(n-1) \le 9n = O(n)$$

$B_k \le 4 \cdot (n-1)$

$O(\sqrt{n})$

Additional space is required for storing k_i, m_i and s_i.

The total space requirement is O(n).

Construction time

- According to Cor. 2b, k can be found in expected time $O(n)$.
- W_i, b_i, m_i, s_i can be computed in time $O(n)$.
- According to Cor. 2a, each k_i can be computed in expected time $O(b_i^2)$.

Total expected time:

$$O\left(n + \sum_{i=0}^{n} b_i^2\right) = O(n + B_k) = O(n)$$

$B_k \leq 4(n-1)$

Main result

Theorem: Let N be a prime number and $S \subseteq U = [0 \ldots N\text{-}1]$ with $|S| = n$.

A perfect hash table of size $O(n)$ and a hash function with access time $O(1)$ can be constructed for S in expected time $O(n)$.