

Datové struktury I

7. přednáška: Separované řetězce z k -nezávislé hešovací systémy

Jirka Fink

<https://ktiml.mff.cuni.cz/~fink/>

Katedra teoretické informatiky a matematické logiky
Matematicko-fyzikální fakulta
Univerzita Karlova v Praze

Zimní semestr 2024/25

Licence: Creative Commons BY-NC-SA 4.0

Základní pojmy

- Značení: $[n] = \{0, \dots, n - 1\}$
- Máme univerzum U všech prvků
- Chceme uložit podmnožinu $S \subseteq U$ velikosti n
- Uložíme S do pole velikosti m pomocí hešovací funkce $h : U \rightarrow M$, kde $M = [m]$
- Hešovacím systémem $\mathcal{H}$ rozumíme libovolnou množinu hešovacích funkcí
- Dva prvky $x, y \in S$ kolidují, jestliže $h(x) = h(y)$
- Uvažujeme universum $U = [u]$ pro libovolné $u \in \mathbb{N}$, pokud není uvedeno jinak

c-universální systém (ekvivalentní definice)

Systém hešovacích funkcí $\mathcal{H}$ je c -universální, jestliže pro všechna různá $x, y \in U$ ①

- počet hešovacích funkcí $h \in \mathcal{H}$ splňujících $h(x) = h(y)$ je nejvýše $\frac{c|\mathcal{H}|}{m}$
- náhodně zvolená $h \in \mathcal{H}$ splňuje $P[h(x) = h(y)] \leq \frac{c}{m}$ ② ③

(k, c) -nezávislý systém hešovacích funkcí

Nechť $k \in \mathbb{N}$, $K = \{1, \dots, k\}$ a $c \geq 1$.

Systém hešovacích funkcí $\mathcal{H}$ je (k, c) -nezávislý, pokud náhodně zvolená $h \in \mathcal{H}$ splňuje

$$P[h(x_i) = z_i \forall i \in K] \leq \frac{c}{m^k}$$

pro všechna po dvou různá $x_1, \dots, x_k \in U$ a všechna $z_1, \dots, z_k \in M$.

k -nezávislý systém hešovacích funkcí

- Systém $\mathcal{H}$ je k -nezávislý, pokud je (k, c) -nezávislý pro nějaké $c \geq 1$.
- Systém $\mathcal{H}$ je silně k -nezávislý, pokud je $(k, 1)$ -nezávislý.

- 1 Navíc obvykle vyžadujeme, aby hešovací funkci šlo spočítat v čase $\mathcal{O}(1)$ a aby funkci bylo možné popsat $\mathcal{O}(1)$ parametry.
- 2 Náhodný výběr hešovací funkce má vždy rovnoměrné rozdělení na celém systému.
- 3 Úplně náhodný hešovací systém je 1-universální, protože $h(x)$ padne do nějaké přihrádky a $h(y)$ má uniformní distribuci nezávislou na $h(x)$, a proto $P[h(x) = h(y)] = \frac{1}{m}$.

Hešování se separovanými řetězci

Popis

V přihrádce j jsou uloženy všechny prvky $i \in S$ splňující $h(i) = j$.

Implementace

- `std::unordered_map` v C++
- `Dictionary` v C#
- `HashMap` v Java
- `Dictionary` v Python

Otázka

Je možné zajistit složitost operací FIND, INSERT a DELETE $\mathcal{O}(\log n)$ v nejhorším případě? ①

Platí následující tvrzení?

Jestliže $m = \Theta(n)$ a pro hešovací systém platí, že očekávaný počet prvků v libovolné přihrádce je $\mathcal{O}(1)$, pak očekávaná složitost operací FIND, INSERT a DELETE je $\mathcal{O}(1)$. ②

- 1 Pro každou přihrádku vytvoříme vyhledávací strom.
- 2 Pro systém $\mathcal{H} = \{h_a(i) = j; j \in M\}$ a přihrádku $j \in M$ z linearity střední hodnoty platí $E[\{i \in S : h(i) = j\}] = \sum_{i \in S} P[h(i) = j] = \frac{n}{m} = \Theta(1)$, ale všechny prvky jsou ve stejné přihrádce, takže složitost operací je lineární.

Hešování se separovanými řetězci: Složitost

Pozorování

Jestliže $\mathcal{H}$ je c -universální, pak očekávaný počet prvků v přihrádce $h(x)$ pro $x \in U$ je nejvýše $\frac{cn}{m}$. $\rightarrow$ pro jeden prvek platí $\frac{cn}{m}$

Důkaz

$$E[|\{y \in S : h(x) = h(y)\}|] = \sum_{y \in S} P[h(x) = h(y)] \leq \frac{cn}{m} \rightarrow |S| = n$$

Důsledek

Jestliže $\mathcal{H}$ je c -universální a $m = \Omega(n)$, pak očekávaná složitost operací FIND, INSERT a DELETE je $\mathcal{O}(1)$. $\rightarrow \frac{cn}{m} = c$

Dynamická velikost tabulky, pokud dopředu neznáme počet prvků

- Tabulku udržujeme ve velikosti $n/4 \leq m \leq n$
- Při překročení mezí tabulku dvakrát zmenšíme/zvětšíme přehešováním všech prvků novou hešovací funkcí
- **Amortizovaná očekávaná** složitost operací INSERT a DELETE je $\mathcal{O}(1)$

Hešování se separovanými řetězci: Nejdelší řetězec

Definice

Posloupnost náhodných jevů $E_n, n \in \mathbb{N}$ se vyskytuje **s velkou pravděpodobností**, pokud existují $c > 1$ a $n_0 \in \mathbb{N}$ takové, že pro každé $n \geq n_0$ platí $P[E_n] \geq 1 - \frac{1}{n^c}$.

Značení

Nechť A_j je počet prvků v j -té přihrádce.

Věta: Délka nejdelšího řetězce

Pokud $m = \Theta(n)$ a systém hešovacích funkcí je úplně náhodný, pak délka nejdelšího řetězce $\max_{j \in M} A_j = \Theta\left(\frac{\log n}{\log \log n}\right)$ s velkou pravděpodobností.

Poznámka

Dokážeme, že $P[\max_j A_j \leq (1 + \epsilon) \frac{\log n}{\log \log n}] > 1 - \frac{1}{n^{\frac{\epsilon}{3}}}$ pro všechna $\epsilon > 0$. ①

> show velkou pravděpodobnost

Důsledek: Očekávaná délka nejdelšího řetězce

Pokud $\alpha = \Theta(1)$ a systém hešovacích funkcí je úplně náhodný, pak očekávaná délka nejdelšího řetězce je $E[\max_{j \in M} A_j] = \Theta\left(\frac{\log n}{\log \log n}\right)$. ②

1 Nebudeme dokazovat, že $\max_{j \in M} A_j = \Omega\left(\frac{\log n}{\log \log n}\right)$ s velkou pravděpodobností.

2 Pro $\epsilon = 3$ dostáváme $P[\max_j A_j \leq 4 \frac{\log n}{\log \log n}] > 1 - \frac{1}{n}$. Tedy $E[\max_{j \in M} A_j] \leq P[\max_j A_j \leq 4 \frac{\log n}{\log \log n}] \cdot 4 \frac{\log n}{\log \log n} + P[\max_j A_j > 4 \frac{\log n}{\log \log n}] \cdot n \leq 4 \frac{\log n}{\log \log n} + 1$.

Důkaz $E[\max_{j \in M} A_j] = \Omega\left(\frac{\log n}{\log \log n}\right)$ vynecháváme.

$$1 - (1 - \frac{1}{n}) = \frac{1}{n}$$

přes všechny prvky: n ✓

Chernoffův odhad

Nechť $X_1, \dots, X_n$ jsou nezávislé náhodné proměnné mající hodnoty $\{0, 1\}$. Označme $X = \sum_{i=1}^n X_i$ a $\mu = E[X]$. Pak pro každé $c > 0$ platí

$$P[X > c\mu] < \frac{e^{(c-1)\mu}}{c^{c\mu}}.$$

Důkaz: $P[\max_j A_j \leq (1 + \epsilon) \frac{\log n}{\log \log n}] > 1 - \frac{1}{n^{\frac{\epsilon}{3}}}$

① I_{ij} je náhodná proměnná indikující, zda i -tý prvek patří do j -té přihrádky

② Platí $A_j = \sum_{i \in S} I_{ij}$

③ Mějme $\epsilon > 0$

④ Označme $\mu = E[A_1] = n/m$

⑤ Dále $c = (1 + \epsilon) \frac{\log n}{\mu \log \log n}$

⑥ Platí $P[\max_j A_j > c\mu] = P[\exists j : A_j > c\mu] \leq \sum_j P[A_j > c\mu] = mP[A_1 > c\mu]$

⑦ Aplikujeme Chernoffův odhad na proměnné I_{i1} pro $i \in S$

⑧ Platí $P[\max_j A_j > c\mu] \leq mP[A_1 > c\mu] < m \frac{e^{(c-1)\mu}}{c^{c\mu}} = me^{-\mu} e^{c\mu - c\mu \log c}$

→ úplně náhodný systém,
pro každý systém je to stejné
??

Hešování se separovanými řetězci: Nejdelší řetězec

Důkaz: $P[\max_j A_j \leq (1 + \epsilon) \frac{\log n}{\log \log n}] > 1 - \frac{1}{n^{\frac{\epsilon}{3}}}$

- Označili jsme $c = (1 + \epsilon) \frac{\log n}{\mu \log \log n}$ a odvozujeme

$$P[\max_j A_j > c\mu] < me^{-\mu} e^{c\mu - c\mu \log c}$$

$$= me^{-\mu} e^{(1+\epsilon) \frac{\log n}{\log \log n} - (1+\epsilon) \frac{\log n}{\log \log n} \log \left(\frac{(1+\epsilon) \log n}{\mu \log \log n} \right)}$$

$$= me^{-\mu} n^{\frac{1+\epsilon}{\log \log n} - \frac{1+\epsilon}{\log \log n} \log \left(\frac{(1+\epsilon) \log n}{\mu \log \log n} \right)}$$

$$= me^{-\mu} n^{\frac{1+\epsilon}{\log \log n} - (1+\epsilon) + \frac{1+\epsilon}{\log \log n} \log \left(\frac{\mu}{1+\epsilon} \log \log n \right)}$$

$$= \frac{m}{n^{1+\frac{\epsilon}{2}}} e^{-\mu} n^{-\frac{\epsilon}{2} + \frac{1+\epsilon}{\log \log n} + (1+\epsilon) \frac{\log \left(\frac{\mu}{1+\epsilon} \log \log n \right)}{\log \log n}}$$

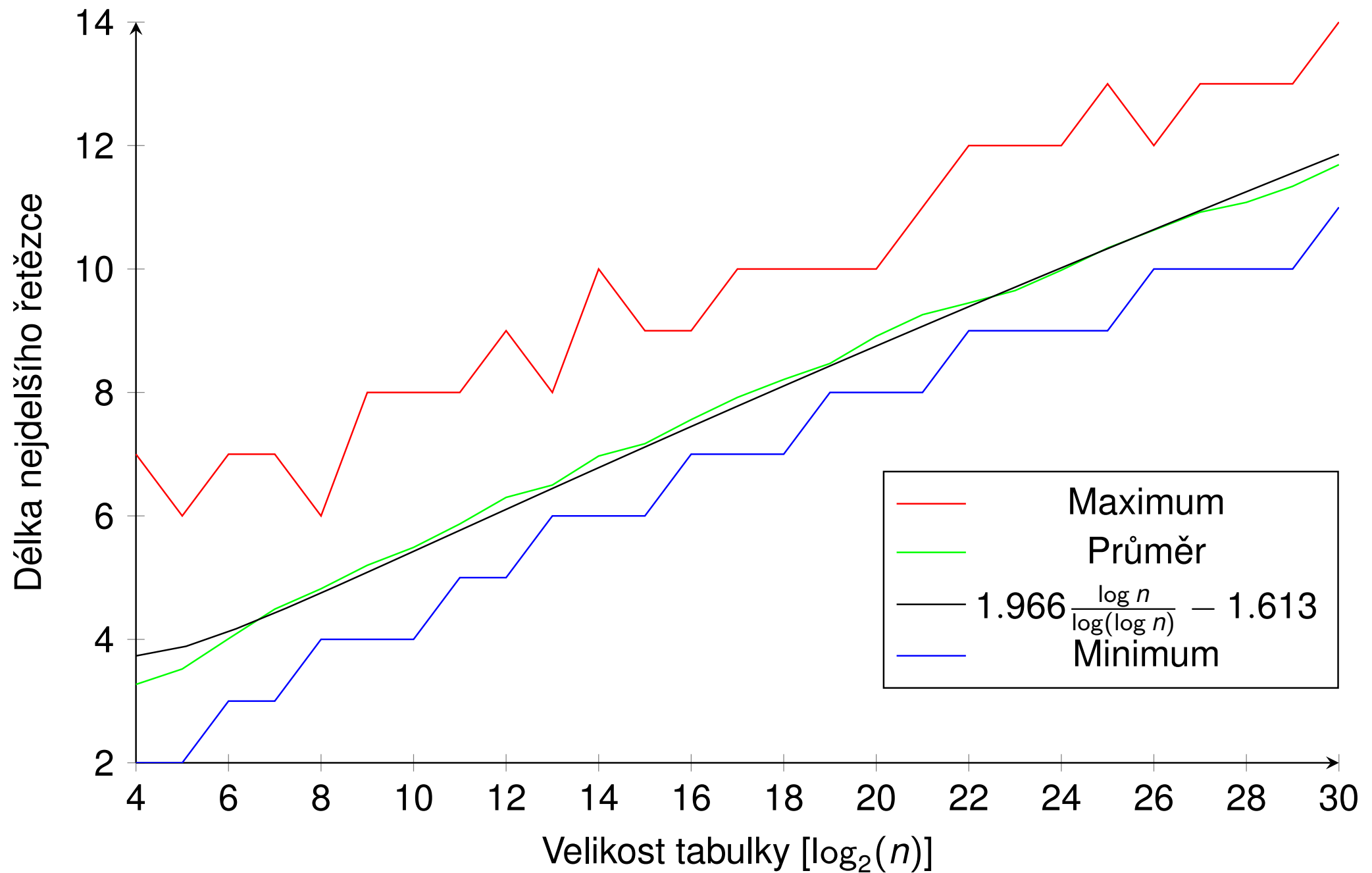
$$< \frac{1}{n^{\frac{\epsilon}{2}}} \frac{me^{-\mu}}{n} n^0 < \frac{1}{n^{\frac{\epsilon}{3}}} \quad \text{... pro dostatečně velká } n$$

sem jsme chtěli dojít.

Protože $-\frac{\epsilon}{2} + \frac{1+\epsilon}{\log \log n} + (1+\epsilon) \frac{\log \left(\frac{\mu}{1+\epsilon} \log \log n \right)}{\log \log n} < 0$ pro dostatečně velká n .

- Tedy $P[\max_j A_j \leq (1 + \epsilon) \frac{\log n}{\log \log n}] > 1 - \frac{1}{n^{\frac{\epsilon}{3}}}$.

Délka nejdelšího řetězce ①



- 1 Hodíme n míčů do n košů. Kolik míčů je v nejplnějším koši v závislosti na n ? Pro každé n provádíme 100 experimentů, ze kterých se díváme na minimum, maximum a průměr. Interpolace funkcí $1.966 \frac{\log n}{\log(\log n)} - 1.613$ má chybu (součet čtverců rozdílů) 0.682 a pro funkci $0.466 \log n + 2.254$ je chyba 0.68, takže na ověření správnosti odhadu $\frac{\log n}{\log(\log n)}$ bychom potřebovali zkoušet větší hodnoty n , ale na to už nemáme dost paměti.

Více-příhrádkové hešování se separovanými řetězci

2-příhrádkové hešování

Prvek x může být uložen v příhrádce $h_1(x)$ nebo $h_2(x)$ a nový prvek vkládáme do příhrádky s menším počtem prvků, kde h_1 a h_2 jsou dvě hešovací funkce.

2-příhrádkové hešování: Délka nejdelšího řetězce (bez důkazu)

Očekávaná délka nejdelšího řetězce je $\mathcal{O}(\log \log n)$.

k -příhrádkové hešování

Prvek x může být uložen v příhrádkách $h_1(x), \dots, h_k(x)$ a nový prvek vkládáme do příhrádky s menším počtem prvků, kde $h_1, \dots, h_k$ jsou hešovací funkce.

k -příhrádkové hešování: Délka nejdelšího řetězce (bez důkazu)

Očekávaná délka nejdelšího řetězce je $\mathcal{O}\left(\frac{\log \log n}{\log k}\right)$.

Lemma

- Necht systém $\mathcal{H}$ je $(2, c)$ -nezávislý z U do $[r]$ a $m \leq r$
- Pak $\mathcal{H} \bmod m = \{x \rightarrow h(x) \bmod m; h \in \mathcal{H}\}$ je $2c$ -universální a $(2, 4c)$ -nezávislý

Lemma

- Necht systém $\mathcal{H}$ je (k, c) -nezávislý z U do $[r]$ a $r \geq 2km$
- Pak $\mathcal{H} \bmod m = \{x \rightarrow h(x) \bmod m; h \in \mathcal{H}\}$ je $(k, 2c)$ -nezávislý

Důkaz

- Zvolme různá $x_1, \dots, x_k \in U$, $z_1, \dots, z_k \in M$ a označme $y_i = h(x_i)$
- $P[h(x_i) \bmod m = z_i \forall i \in K] = \sum P[y_i \equiv_m z_i \forall i \in K]$
- Sumu sčítáme přes nejvýše $\lceil \frac{r}{m} \rceil \leq \frac{r+m-1}{m}$ hodnot $y_1, \dots, y_k$

- Z odhadu $1 + x \leq e^x$ plyne

$$P[h(x_i) \bmod m = z_i \forall i \in K] \leq \frac{c}{r^k} \cdot \left(\frac{r+m-1}{m}\right)^k = \frac{c}{m^k} \cdot \left(\frac{r+m-1}{r}\right)^k \leq \frac{c}{m^k} \cdot \left(1 + \frac{m}{r}\right)^k = \frac{c}{m^k} \cdot e^{km/r} \leq \frac{c}{m^k} \cdot e^{1/2} \leq \frac{2c}{m^k}$$

Lemma

- Necht systém $\mathcal{H}$ je (k, c) -nezávislý z U do $[r]$ a $r \geq 2km$
- Pak $\mathcal{H} \bmod m = \{x \rightarrow h(x) \bmod m; h \in \mathcal{H}\}$ je $(k, 2c)$ -nezávislý

Věta z algebry: Jednoznačnost interpolace polynomem

Pro každé těleso T , $k > 1$ celočíselné, po dvou různá $x_1, \dots, x_k \in T$ a $y_1, \dots, y_k \in T$ existuje právě jeden polynom $p_a(x) = \sum_{i=0}^{k-1} a_i x^i$ stupně $k - 1$ s koeficienty $a_0, \dots, a_{k-1} \in T$ takový, že $p(x_i) = y_i$ pro všechna $i \in K$.

Pozorování: Systém Poly-mod-prime

- Necht p je prvočíslo, $a \in \mathbb{Z}_p^k$, $x \in \mathbb{Z}_p$
- $h_a(x) = \sum_{i=0}^{k-1} a_i x^i$
- Systém $P_k = \{h_a; a \in \mathbb{Z}_p^k\}$ je $(k, 1)$ -nezávislý ①
- Systém $P_k \bmod m$ je $(k, 2)$ -nezávislý pro $p \geq 2km$ ②

- ① Důkaz přímo plyne z jednoznačnosti polynomu
- ② Jestliže p je prvočíslo, tak hešovací funkci lze zapsat jako $h_a(x) = (\sum_{i=0}^{k-1} a_i x^i \bmod p) \bmod m$, kde aritmetické operace jsou nad celými čísly.

Multiply-shift

- Předpokládáme, že $|U| = 2^w$ a $m = 2^l$
- $h_a(x) = (ax \bmod 2^w) \gg (w - l)$ — *→ vezmeme jen MSb část čísla*
- $\mathcal{H} = \{h_a; a \text{ je liché } w\text{-bitové číslo}\}$

Implementace v C

```
uint64_t hash(uint64_t x, uint64_t l, uint64_t a)
{ return (a*x) >> (64-l); }
```

Vlastnosti systému multiply-shift

- 2-universální
- Velmi rychlý na reálných počítačích ✓
- V praxi často používaný ✓
- Celý výpočet musí být proveden v neznaménkových celočíselných typech, protože ze součinu ax potřebujeme získat posledních w bitů

Tabulkové hešování

- $\oplus$ značí bitový XOR
 - Předpokládáme, že $u = 2^w$ a $m = 2^l$ a w je násobek $d \geq 2$
 - Bitový zápis čísla $x \in U$ rozdělíme na d částí $x^1, \dots, x^d$ po $\frac{w}{d}$ bitech
 - Pro každé $i = 1, \dots, d$ vybereme náhodnou hešovací funkci $T_i : [2^{w/d}] \rightarrow M$
 - Hešovací funkce je $h(x) = T_1(x^1) \oplus \dots \oplus T_d(x^d)$
 - K vygenerování h potřebujeme $d \cdot 2^{w/d}$ náhodných čísel z rozsahu $M = [2^l]$
- projeďte čísla
číslo do M*

Ilustrativní příklad

- Uvažujme $w = 12$ a $d = 3$
- Nejprve vygeneruje náhodné funkce $T_1, T_2, T_3 : [2^4] \rightarrow M$
- Číslo $x = 101100111001$ rozdělíme na $x^1 = 1011$, $x^2 = 0011$ a $x^3 = 1001$
- Výpočet hešovací funkce je
$$h(x) = T_1(x^1) \oplus T_2(x^2) \oplus T_3(x^3) = T_1(1011) \oplus T_2(0011) \oplus T_3(1001)$$

Univerzalita

Tabulkové hešování je silně 3-nezávislé, ale není 4-nezávislé.

$\leq \frac{1}{n^3}$

Tabulkové hešování

- Předpokládáme, že $u = 2^w$ a $m = 2^l$ a w je násobek d
- Bitový zápis čísla $x \in U$ rozdělíme na d částí $x^1, \dots, x^d$ po $\frac{w}{d}$ bitech
- Pro každé $i = 1, \dots, d$ vybereme náhodnou hešovací funkci $T_i : [2^{w/d}] \rightarrow M$
- Hešovací funkce je $h(x) = T_1(x^1) \oplus \dots \oplus T_d(x^d)$

Univerzalita

Tabulkové hešování je 3-nezávislé, ale není 4-nezávislé.

Důkaz 2-nezávislosti (3-nezávislost je ponechána na cvičení)

- Mějme dva prvky x_1 a x_2 lišící se v i -tých částech
- Nechť $h_i(x) = T_1(x^1) \oplus \dots \oplus T_{i-1}(x^{i-1}) \oplus T_{i+1}(x^{i+1}) \oplus \dots \oplus T_d(x^d)$
- $P[h(x_1) = z_1] = P[h_i(x_1) \oplus T_i(x_1^i) = z_1] = P[T_i(x_1^i) = z_1 \oplus h_i(x_1)] = \frac{1}{m}$ ①
- Náhodné jevy $h(x_1) = z_1$ a $h(x_2) = z_2$ jsou nezávislé
 - Náhodné proměnné $T_i(x_1^i)$ a $T_i(x_2^i)$ jsou nezávislé
 - Náhodné jevy $T_i(x_1^i) = z_1 \oplus h_i(x_1)$ a $T_i(x_2^i) = z_2 \oplus h_i(x_2)$ jsou nezávislé
- $P[h(x_1) = z_1 \text{ a } h(x_2) = z_2] = P[h(x_1) = z_1]P[h(x_2) = z_2] = \frac{1}{m^2}$

NEZÁVISLOST

- 1 $T_i(x_1^j)$ nabývá všech hodnot z M se stejnou pravděpodobností $\frac{1}{m}$ a náhodné proměnné $T_i(x_1^j)$ a $z_1 \oplus h_i(x_1)$ jsou nezávislé.

Tabulkové hešování není 4-nezávislé

- 1 Zvolíme prvky x_1, x_2, x_3 a x_4 takové, že
 - části x_1 splňují $x_1^1 = 0, x_1^2 = 0, x_1^i = 0$ pro $i \geq 3$
 - části x_2 splňují $x_2^1 = 1, x_2^2 = 0, x_2^i = 0$ pro $i \geq 3$
 - části x_3 splňují $x_3^1 = 0, x_3^2 = 1, x_3^i = 0$ pro $i \geq 3$
 - části x_4 splňují $x_4^1 = 1, x_4^2 = 1, x_4^i = 0$ pro $i \geq 3$
- 2 Platí $h(x_1) \oplus h(x_2) \oplus h(x_3) = h(x_4) \rightarrow$ leví strana dá 110..., pravá je 110...
- 3 Zvolme libovolná z_1, z_2, z_3 a nechť $z_4 = z_1 \oplus z_2 \oplus z_3$
- 4 Jestliže $h(x_1) = z_1, h(x_2) = z_2$ a $h(x_3) = z_3$, pak platí $h(x_4) = z_4$
- 5 $P[h(x_1) = z_1 \text{ a } h(x_2) = z_2 \text{ a } h(x_3) = z_3 \text{ a } h(x_4) = z_4] = \frac{1}{m^3} > \frac{c}{m^4}$

takže je určeno předchozími body

tedy 5) = $P[h(x_1) = z_1 \text{ a } h(x_2) = z_2 \text{ a } h(x_3) = z_3] = \frac{1}{m^3}$

Scalar-mod-prime

- Chceme hešovat d -tici $x_1, \dots, x_d \in \mathbb{Z}_p$, kde p je prvočíslo
- $\left\{ x_1, \dots, x_d \rightarrow \sum_{i=1}^d a_i x_i \bmod p; a \in \mathbb{Z}_p^d \right\}$ je 1-universální $\frac{1}{m}$
- $\left\{ x_1, \dots, x_d \rightarrow b + \sum_{i=1}^d a_i x_i \bmod p; a \in \mathbb{Z}_p^d, b \in \mathbb{Z}_p \right\}$ je (2,1)-nezávislý $\frac{1}{m^2}$
- $\left\{ x_1, \dots, x_d \rightarrow \left(b + \sum_{i=1}^d a_i x_i \bmod p \right) \bmod m; a \in \mathbb{Z}_p^d, b \in \mathbb{Z}_p \right\}$ je (2,4)-nezávislý $\frac{1}{m^2}$

Důkaz 1-universálnosti ①

- Mějme různé $x, y \in \mathbb{Z}_p^d$ a BÚNO předpokládejme, že $x_1 \neq y_1$
- $P[a \cdot x \equiv_p a \cdot y] = P[a \cdot (x - y) \equiv_p 0] = P \left[a_1 \equiv_p \frac{\sum_{i=2}^d a_i (y_i - x_i)}{x_1 - y_1} \right] = 1/p$ ②

- ① Pro 2-nezávislost stačí podobně nahlédnout, že a_1, b jsou jednoznačně určeny.
- ② Náhodná proměnná a_1 musí nabývat jednu konkrétní hodnotu, což nastane s pravděpodobností $1/p$.

Poly-mod-prime pro různě dlouhé řetězce I

- Chceme hešovat řetězec $x_1, \dots, x_d \in \mathbb{Z}_p$, kde p je prvočíslo
- $\left\{ x_1, \dots, x_d \rightarrow \sum_{i=0}^{d-1} x_{i+1} a^i \bmod p; a \in [p] \right\}$ je d -universální $\frac{d}{m}$
- Dva různé polynomy stupně nejvýše $d - 1$ mají nejvýše d společných bodů, takže existuje nejvýše d kolidujících hodnot a .

Poly-mod-prime pro různě dlouhé řetězce II

- Chceme hešovat řetězec $x_1, \dots, x_d \in U$ do M , kde $p \geq m$ je prvočíslo
- $h_{a,b,c}(x_1, \dots, x_d) = \left(b + c \sum_{i=0}^{d-1} x_{i+1} a^i \bmod p \right) \bmod m$
- $\mathcal{H} = \{h_{a,b,c}; a, b, c \in [p]\}$
- $P[h_{a,b,c}(x_1, \dots, x_d) = h_{a,b,c}(x'_1, \dots, x'_{d'})] \leq \frac{2}{m}$ pro různé řetězce délek $d, d' \leq \frac{p}{m}$.