Simplified Tight Bounds for Monotone Minimal Perfect Hashing

Dmitry Kosolobov

Ural Federal University, Ekaterinburg, Russia

Content

- Monotone minimal perfect hash function
- Known upper and lower bounds
- Model and counting argument
- Let's count

Monotone minimal perfect hash function (MMPHF)

Monotone minimal perfect hash function (MMPHF)

Given $a_1 < \cdots < a_n$ from $[1..u] = \{1, 2, ..., u\}$, compute a data structure with queries $f : [1..u] \to [1..n]$:

- $f(x) = k$ if $x = a_k$ for some $k \in [1..n]$
- $f(x)$ is arbitrary if $x \notin \{a_1, \ldots, a_n\}$

Monotone minimal perfect hash function (MMPHF)

Given $a_1 < \cdots < a_n$ from $[1..u] = \{1, 2, ..., u\}$, compute a data structure with queries $f\colon [1..u] \to [1..n]$:

- $f(x) = k$ if $x = a_k$ for some $k \in [1..n]$
- $f(x)$ is arbitrary if $x \notin \{a_1, \ldots, a_n\}$

Example

$$u = 16, \{a_1, ..., a_5\} = \{3, 6, 7, 10, 14\}$$

1 2 <u>3</u> 4 5 <u>6</u> <u>7</u> 8 9 <u>10</u> 11 12 13 <u>14</u> 15 16

MMPHF colors the segment $[1..u]$: the color of x is $f(x)$

color 1, color 2, color 3, color 4, color 5

Monotone minimal perfect hash function (MMPHF)

Monotone minimal perfect hash function (MMPHF)

Restriction throughout the talk: $u \le 2^{2^{\mathrm{poly}(n)}}$

Monotone minimal perfect hash function (MMPHF)

Restriction throughout the talk: $u \leq 2^{2^{\mathrm{poly}(n)}}$
Upper bound: $O(n \log\log\log u)$ bits, $O(\log u)$ query time
[Belazzougui, Boldi, Pagh, Vigna SODA'09]

Monotone minimal perfect hash function (MMPHF)

Restriction throughout the talk: $u \le 2^{2^{\mathrm{poly}(n)}}$
Upper bound: $O(n \log\log\log u)$ bits, $O(\log u)$ query time
[Belazzougui, Boldi, Pagh, Vigna SODA'09]

Can this space be lowered?

Monotone minimal perfect hash function (MMPHF)

Restriction throughout the talk: $u \le 2^{2^{\mathrm{poly}(n)}}$
Upper bound: $O(n \log\log\log u)$ bits, $O(\log u)$ query time
[Belazzougui, Boldi, Pagh, Vigna SODA'09]

Can this space be lowered?
NO...

Lower bound: $\Omega(n)$ bits
[Belazzougui, Boldi, Pagh, Vigna SODA'11]

Monotone minimal perfect hash function (MMPHF)

Restriction throughout the talk: $u \le 2^{2^{\mathrm{poly}(n)}}$
Upper bound: $O(n \log\log\log u)$ bits, $O(\log u)$ query time
[Belazzougui, Boldi, Pagh, Vigna SODA'09]

Can this space be lowered?
NO...

Lower bound: $\Omega(n)$ bits
[Belazzougui, Boldi, Pagh, Vigna SODA'11]
Lower bound: $\Omega(n \log\log\log u)$ bits for all $u \ge n 2^{2^{\sqrt{\log\log n}}}$
[Assadi, Farach-Colton, Kuszmaul SODA'23]

Monotone minimal perfect hash function (MMPHF)

Restriction throughout the talk: $u \leq 2^{2^{\mathrm{poly}(n)}}$
Upper bound: $O(n \log\log\log u)$ bits, $O(\log u)$ query time
[Belazzougui, Boldi, Pagh, Vigna SODA'09]

Can this space be lowered?
NO...

Lower bound: $\Omega(n)$ bits
[Belazzougui, Boldi, Pagh, Vigna SODA'11]
Lower bound: $\Omega(n \log\log\log u)$ bits for all $u \geq n2^{2^{\sqrt{\log\log n}}}$
[Assadi, Farach-Colton, Kuszmaul SODA'23]
Lower bound: $\Omega(n \log\log\log \frac{u}{n})$ bits for all $u \geq (1+\epsilon)n$
ours

Monotone minimal perfect hash function (MMPHF)

Restriction throughout the talk: $u \leq 2^{2^{\mathrm{poly}(n)}}$
Upper bound: $O(n \log\log\log u)$ bits, $O(\log u)$ query time
[Belazzougui, Boldi, Pagh, Vigna SODA'09]

Can this space be lowered?
NO...

Lower bound: $\Omega(n)$ bits
[Belazzougui, Boldi, Pagh, Vigna SODA'11]
Lower bound: $\Omega(n \log\log\log u)$ bits for all $u \geq n2^{2^{\sqrt{\log\log n}}}$
[Assadi, Farach-Colton, Kuszmaul SODA'23]
Lower bound: $\Omega(n \log\log\log \frac{u}{n})$ bits for all $u \geq (1+\epsilon)n$
ours

Tight upper bound: $O(n \log\log\log \frac{u}{n})$

Monotone minimal perfect hash function (MMPHF)

Restriction throughout the talk: $u \le 2^{2^{\mathrm{poly}(n)}}$
Upper bound: $O(n \log\log\log u)$ bits, $O(\log u)$ query time
[Belazzougui, Boldi, Pagh, Vigna SODA'09]

Can this space be lowered?
NO...

Lower bound: $\Omega(n)$ bits
[Belazzougui, Boldi, Pagh, Vigna SODA'11]
Lower bound: $\Omega(n \log\log\log u)$ bits for all $u \ge n2^{2^{\sqrt{\log\log n}}}$
[Assadi, Farach-Colton, Kuszmaul SODA'23]
Lower bound: $\Omega(n \log\log\log \frac{u}{n})$ bits for all $u \ge (1+\epsilon)n$
ours

Tight upper bound: $O(n \log\log\log \frac{u}{n})$
For all reasonable $n \le u < (1+\epsilon)n$, known facts give tight bounds

Lower bound of Assadi, Farach-Colton, Kuszmaul

Lower bound of Assadi, Farach-Colton, Kuszmaul

[Assadi, Farach-Colton, and Kuszmaul, SODA'23]:
graph of data structures,
chromatic number,
fractional chromatic number,
non-standard graph products,
duality of linear programming,
intricate probability,
...

Lower bound of Assadi, Farach-Colton, Kuszmaul

[Assadi, Farach-Colton, and Kuszmaul, SODA'23]:
graph of data structures,
chromatic number,
fractional chromatic number,
non-standard graph products,
duality of linear programming,
intricate probability,
...

Lower bound of Assadi, Farach-Colton, Kuszmaul

[Assadi, Farach-Colton, and Kuszmaul, SODA'23]:
graph of data structures,
chromatic number,
fractional chromatic number,
non-standard graph products,
duality of linear programming,
intricate probability,
...

Simpler?

Model of computation

Model of computation

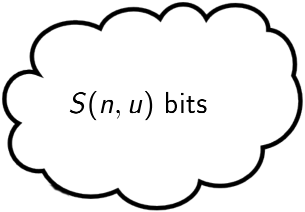
Cell-probe model?

Model of computation

Cell-probe model? Not exactly. Query time is not interesting for us

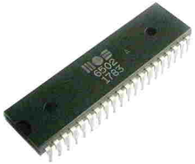
Model of computation

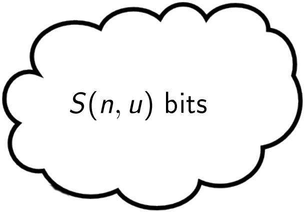
Cell-probe model? Not exactly. Query time is not interesting for us



Model of computation

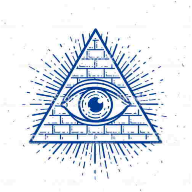
Cell-probe model? Not exactly. Query time is not interesting for us

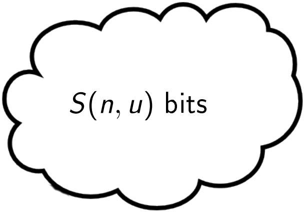




Model of computation

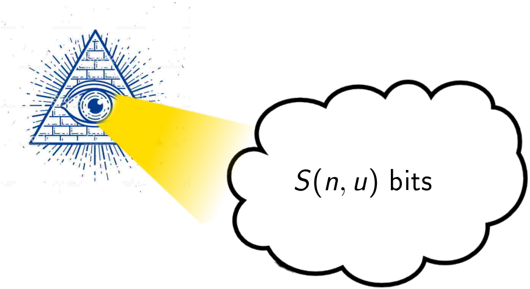
Cell-probe model? Not exactly. Query time is not interesting for us





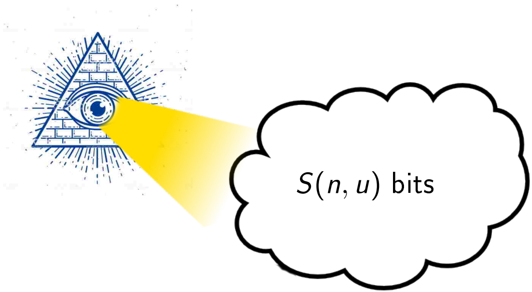
Model of computation

Cell-probe model? Not exactly. Query time is not interesting for us



Model of computation

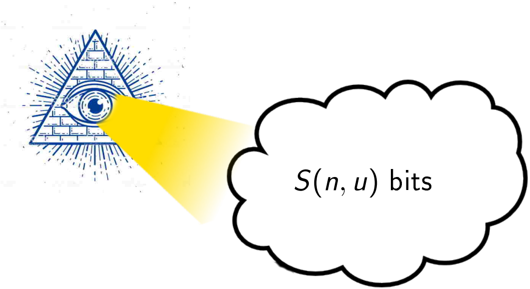
Cell-probe model? Not exactly. Query time is not interesting for us



Queries take $O(1)$ time

Model of computation

Cell-probe model? Not exactly. Query time is not interesting for us



Queries take $O(1)$ time
Modelled as a function $f : [1..u] \times \{0, 1\}^S \to [1..n]$

Counting argument

One data structure corresponds to a coloring of $[1..u]$ in colors $[1..n]$

Counting argument

One data structure corresponds to a coloring of $[1..u]$ in colors $[1..n]$

$S(n, u)$ bits encode at most $2^{S(n,u)}$ colorings

Counting argument

One data structure corresponds to a coloring of $[1..u]$ in colors $[1..n]$

$S(n, u)$ bits encode at most $2^{S(n,u)}$ colorings

One coloring may correctly encode many tuples $\{a_1, \ldots, a_n\} \subset [1..u]$

Counting argument

One data structure corresponds to a coloring of $[1..u]$ in colors $[1..n]$

$S(n,u)$ bits encode at most $2^{S(n,u)}$ colorings

One coloring may correctly encode many tuples $\{a_1, ..., a_n\} \subset [1..u]$

Example

$$\{a_1, ..., a_5\} = \{3, 6, 7, 10, 14\}$$

1 2 <u>3</u> 4 5 <u>6</u> <u>7</u> 8 9 <u>10</u> 11 12 13 <u>14</u> 15 16

color 1, color 2, color 3, color 4, color 5

Counting argument

One data structure corresponds to a coloring of $[1..u]$ in colors $[1..n]$

$S(n, u)$ bits encode at most $2^{S(n,u)}$ colorings

One coloring may correctly encode many tuples $\{a_1, ..., a_n\} \subset [1..u]$

Example

$$\{a_1, ..., a_5\} = \{3, 6, 7, 10, 14\}, \{1, 2, 4, 9, 12\}$$

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

color 1, color 2, color 3, color 4, color 5

Counting argument

One data structure corresponds to a coloring of $[1..u]$ in colors $[1..n]$

$S(n, u)$ bits encode at most $2^{S(n,u)}$ colorings

One coloring may correctly encode many tuples $\{a_1, ..., a_n\} \subset [1..u]$

Example

$$\{a_1, ..., a_5\} = \{3, 6, 7, 10, 14\}, \{1, 2, 4, 9, 12\}, \{1, 6, 11, 15, 16\}$$

$\underline{1}$ 2 3 4 5 $\underline{6}$ 7 8 9 10 $\underline{11}$ 12 13 14 $\underline{15}$ $\underline{16}$

color 1, color 2, color 3, color 4, color 5

Counting argument

A coloring is a map $f\colon [1..u] \to [1..n]$; suppose that $\mathcal{C}$ is a minimal family of colorings such that every tuple $\{a_1, \ldots, a_n\} \subset [1..u]$ is encoded by some $f \in \mathcal{C}$, i.e., $f(a_k) = k$ for all $k \in [1..n]$

Counting argument

A coloring is a map $f\colon [1..u] \to [1..n]$; suppose that $\mathcal{C}$ is a minimal family of colorings such that every tuple $\{a_1, \ldots, a_n\} \subset [1..u]$ is encoded by some $f \in \mathcal{C}$, i.e., $f(a_k) = k$ for all $k \in [1..n]$

Claim

The MMPHF requires $S(n, u) \geq \log |\mathcal{C}|$ bits of space, so we have to prove that $\log |\mathcal{C}| \geq \Omega(n \log\log\log \frac{u}{n})$ for $u \geq (1 + \epsilon)n$

Counting argument

A coloring is a map $f\colon [1..u] \to [1..n]$; suppose that $\mathcal{C}$ is a minimal family of colorings such that every tuple $\{a_1, \ldots, a_n\} \subset [1..u]$ is encoded by some $f \in \mathcal{C}$, i.e., $f(a_k) = k$ for all $k \in [1..n]$

Claim

The MMPHF requires $S(n,u) \geq \log|\mathcal{C}|$ bits of space, so we have to prove that $\log|\mathcal{C}| \geq \Omega(n \log\log\log \frac{u}{n})$ for $u \geq (1+\epsilon)n$

Assadi et al. prove the bound for the case $u = 2^{2^{n^3}}$, when $n \log\log\log \frac{u}{n} = \Theta(n \log n)$:

Counting argument

A coloring is a map $f\colon [1..u] \to [1..n]$; suppose that $\mathcal{C}$ is a minimal family of colorings such that every tuple $\{a_1, \ldots, a_n\} \subset [1..u]$ is encoded by some $f \in \mathcal{C}$, i.e., $f(a_k) = k$ for all $k \in [1..n]$

Claim

The MMPHF requires $S(n, u) \geq \log|\mathcal{C}|$ bits of space, so we have to prove that $\log|\mathcal{C}| \geq \Omega(n \log\log\log \frac{u}{n})$ for $u \geq (1+\epsilon)n$

Assadi et al. prove the bound for the case $u = 2^{2^{n^3}}$, when $n \log\log\log \frac{u}{n} = \Theta(n \log n)$:

- devise a random process generating n-tuples
- prove that any fixed coloring encodes the generated n-tuple with probability $\leq \frac{1}{n^{\Omega(n)}}$
- then there are $\geq n^{\Omega(n)}$ colorings in $\mathcal{C}$

Counting argument

A coloring is a map $f\colon [1..u] \to [1..n]$; suppose that $\mathcal{C}$ is a minimal family of colorings such that every tuple $\{a_1, \dots, a_n\} \subset [1..u]$ is encoded by some $f \in \mathcal{C}$, i.e., $f(a_k) = k$ for all $k \in [1..n]$

Claim

The MMPHF requires $S(n, u) \geq \log |\mathcal{C}|$ bits of space, so we have to prove that $\log |\mathcal{C}| \geq \Omega(n \log\log\log \frac{u}{n})$ for $u \geq (1+\epsilon)n$

Assadi et al. prove the bound for the case $u = 2^{2^{n^3}}$, when $n \log\log\log \frac{u}{n} = \Theta(n \log n)$:

- devise a random process generating n-tuples
- prove that any fixed coloring encodes the generated n-tuple with probability $\leq \frac{1}{n^{\Omega(n)}}$
- then there are $\geq n^{\Omega(n)}$ colorings in $\mathcal{C}$
- then $\log |\mathcal{C}| \geq \log(n^{\Omega(n)}) = \Omega(n \log n)$

Very small and very large u

Very small and very large u

- Suppose $u \geq 2^{2^{n^3}}$

Very small and very large u

- Suppose $u \geq 2^{2^{n^3}}$

  The process generating n-tuples from $[1..2^{2^{n^3}}] \subseteq [1..u]$ gives the probability at most $1/n^{\Omega(n)}$, implying the lower bound $\Omega(n \log n)$, which is $\Omega(n \log\log\log \frac{u}{n})$ when $2^{2^{n^3}} \leq u \leq 2^{2^{\mathrm{poly}(n)}}$

Very small and very large u

- Suppose $u \geq 2^{2^{n^3}}$
  The process generating n-tuples from $[1..2^{2^{n^3}}] \subseteq [1..u]$ gives the probability at most $1/n^{\Omega(n)}$, implying the lower bound $\Omega(n \log n)$, which is $\Omega(n \log\log\log \frac{u}{n})$ when $2^{2^{n^3}} \leq u \leq 2^{2^{\mathrm{poly}(n)}}$
- Suppose $(1+\epsilon)n \leq u < 2^{2^8} n$

Very small and very large u

- Suppose $u \geq 2^{2^{n^3}}$
  The process generating n-tuples from $[1..2^{2^{n^3}}] \subseteq [1..u]$ gives the probability at most $1/n^{\Omega(n)}$, implying the lower bound $\Omega(n \log n)$, which is $\Omega(n \log\log\log \frac{u}{n})$ when $2^{2^{n^3}} \leq u \leq 2^{2^{\mathrm{poly}(n)}}$
- Suppose $(1+\epsilon)n \leq u < 2^{2^8} n$
  Since $n \log\log\log \frac{u}{n} = \Theta(n)$, the lower bound $\Omega(n)$ follows from the known bound $\Omega(n)$

Normal u

Suppose $2^{2^8} n \le u < 2^{2^{n^3}}$

Normal u

Suppose $2^{2^8} n \le u < 2^{2^{n^3}}$
Split $[1..u]$ into $n/\bar{n}$ blocks of size $\bar{u} = u/(n/\bar{n})$, where $\bar{n} = \lfloor (\log\log \frac{u}{n})^{1/3} \rfloor$

1	2	3	4	...	$n/\bar{n}$
$\bar{u}$	$\bar{u}$	$\bar{u}$	$\bar{u}$	...	$\bar{u}$

Normal u

Suppose $2^{2^8} n \le u < 2^{2^{n^3}}$
Split $[1..u]$ into $n/\bar{n}$ blocks of size $\bar{u} = u/(n/\bar{n})$, where
$\bar{n} = \lfloor (\log\log \frac{u}{n})^{1/3} \rfloor$

1	2	3	4	...	$n/\bar{n}$
$\bar{u}$	$\bar{u}$	$\bar{u}$	$\bar{u}$	...	$\bar{u}$

Randomly generate $\bar{n}$-tuple inside each block by our process, which is possible since $\bar{u} \ge 2^{2^{\bar{n}^3}}$: indeed $\bar{u} \ge u/n = 2^{2^{\log\log \frac{u}{n}}} \ge 2^{2^{\bar{n}^3}}$

Normal u

Suppose $2^{2^8} n \le u < 2^{2^{n^3}}$
Split $[1..u]$ into $n/\bar{n}$ blocks of size $\bar{u} = u/(n/\bar{n})$, where $\bar{n} = \lfloor (\log\log \frac{u}{n})^{1/3} \rfloor$

1	2	3	4	...	$n/\bar{n}$
$\bar{u}$	$\bar{u}$	$\bar{u}$	$\bar{u}$	...	$\bar{u}$

Randomly generate $\bar{n}$-tuple inside each block by our process, which is possible since $\bar{u} \ge 2^{2^{\bar{n}^3}}$: indeed $\bar{u} \ge u/n = 2^{2^{\log\log \frac{u}{n}}} \ge 2^{2^{\bar{n}^3}}$
Fix any coloring of $[1..u]$: the probability that the random n-tuple is encoded by this coloring is $\le (1/\bar{n}^{\Omega(\bar{n})})^{n/\bar{n}} = 1/\bar{n}^{\Omega(n)}$

Normal u

Suppose $2^{2^8} n \le u < 2^{2^{n^3}}$
Split $[1..u]$ into $n/\bar{n}$ blocks of size $\bar{u} = u/(n/\bar{n})$, where
$\bar{n} = \lfloor (\log\log \frac{u}{n})^{1/3} \rfloor$

1	2	3	4	...	$n/\bar{n}$
$\bar{u}$	$\bar{u}$	$\bar{u}$	$\bar{u}$	...	$\bar{u}$

Randomly generate $\bar{n}$-tuple inside each block by our process, which is possible since $\bar{u} \ge 2^{2^{\bar{n}^3}}$: indeed $\bar{u} \ge u/n = 2^{2^{\log\log \frac{u}{n}}} \ge 2^{2^{\bar{n}^3}}$
Fix any coloring of $[1..u]$: the probability that the random n-tuple is encoded by this coloring is $\le (1/\bar{n}^{\Omega(\bar{n})})^{n/\bar{n}} = 1/\bar{n}^{\Omega(n)}$
Then there are $\ge \bar{n}^{\Omega(n)}$ colorings in $\mathcal{C}$

Normal u

Suppose $2^{2^8} n \le u < 2^{2^{n^3}}$
Split $[1..u]$ into $n/\bar{n}$ blocks of size $\bar{u} = u/(n/\bar{n})$, where
$\bar{n} = \lfloor (\log\log \frac{u}{n})^{1/3} \rfloor$

1	2	3	4	...	$n/\bar{n}$
$\bar{u}$	$\bar{u}$	$\bar{u}$	$\bar{u}$	...	$\bar{u}$

Randomly generate $\bar{n}$-tuple inside each block by our process, which is possible since $\bar{u} \ge 2^{2^{\bar{n}^3}}$: indeed $\bar{u} \ge u/n = 2^{2^{\log\log \frac{u}{n}}} \ge 2^{2^{\bar{n}^3}}$
Fix any coloring of $[1..u]$: the probability that the random n-tuple is encoded by this coloring is $\le (1/\bar{n}^{\Omega(\bar{n})})^{n/\bar{n}} = 1/\bar{n}^{\Omega(n)}$
Then there are $\ge \bar{n}^{\Omega(n)}$ colorings in $\mathcal{C}$
Then $\log|\mathcal{C}| \ge \log(\bar{n}^{\Omega(n)}) = \Omega(n \log \bar{n}) = \Omega(n \log\log\log \frac{u}{n})$

Omitted: upper and lower bounds for $n \leq u < (1+\epsilon)n$, upper bound $O(n \log\log\log \frac{u}{n})$ for $u \geq (1+\epsilon)n$, randomized MMPHF, the random process of Assadi et al. for $u = 2^{2^{n^3}}$

Omitted: upper and lower bounds for $n \leq u < (1+\epsilon)n$, upper bound $O(n \log\log\log \frac{u}{n})$ for $u \geq (1+\epsilon)n$, randomized MMPHF, the random process of Assadi et al. for $u = 2^{2^{n^3}}$

Thank you!