On the Circuit Complexity of Perfect Hashing

Oded Goldreich and Avi Wigderson

Abstract. We consider the size of circuits that *perfectly hash* an arbitrary subset $S \subset \{0,1\}^n$ of cardinality 2^k into $\{0,1\}^m$. We observe that, in general, the size of such circuits is exponential in $2k - m$, and provide a matching upper bound.

Keywords: Perfect Hashing, Circuit Complexity.

An early version of this work appeared as TR96-041 of *ECCC*. We later found out that, in contrast to our previous impression, the lower bound has been known. In fact, our lower bound argument is analogous to the one presented in [6, pp. 128-129]. The current revision is quite minimal.

Summary

We consider the problem of perfectly hashing an arbitrary subset $S \subset \{0,1\}^n$ of cardinality 2^k into $\{0,1\}^m$, where $k \leq m$. That is, given an arbitrary subset $S \subset \{0,1\}^n$ of cardinality 2^k, we seek a function $h : \{0,1\}^n \rightarrow \{0,1\}^m$ so that $h(x) \neq h(y)$ for every two distinct $x \neq y$ in S. Clearly, such a function always exists, the question is what is its complexity; that is, what is the size of the smallest circuit computing h. Two obvious upper bounds follow.

1. For every $S \subset \{0,1\}^n$, there is a circuit of size $|S| \cdot n$ that perfectly hashes S into $\{0,1\}^{\lceil \log_2 |S| \rceil}$.
   (The circuit is merely a look-up table for S.)
2. For every $S \subset \{0,1\}^n$, there is a circuit of size $\text{poly}(n)$ that perfectly hashes S into $\{0,1\}^{2\lceil \log_2 |S| \rceil}$.
   (The circuit implements a suitable function from a family of Universal$_2$ Hashing [2]. Such a family always contains perfect hashing functions for S [4].)

We show that these upper bounds are the best possible. That is:

Theorem 1 (lower bound): *For every n, k and $m \leq n - 1$, there exists a subset $S \subset \{0,1\}^n$ of cardinality 2^k such that perfectly hashing S into $\{0,1\}^m$ requires a circuit of size $\Omega(2^{2k-m}/n)$.*

Interestingly, this lower bound is tight for all values of $m \in [k, 2k]$ (and not merely for $m \in \{k, 2k\}$). That is:

Theorem 2 (matching upper bound):[1] *For every n, m, k where $k \leq m \leq 2k$, and every subset $S \subset \{0,1\}^n$ of cardinality 2^k, there exists a circuit of size $2^{2k-m} \cdot \text{poly}(n)$ that perfectly hashes S into $\{0,1\}^m$.*

[1] We stress that the circuits guaranteed here cannot, in general, be simply described; that is, this result is inherently nonuniform.

1 Proof of Theorem 1

The proof follows by a simple counting argument, combining an upper bound on the number of circuits of given size with a lower bound on the size of a family of functions that can perfectly hash all subsets of size 2^k. Improved lower bounds for the latter appears in [3, 5, 7]. For sake of completeness, we prove a weaker bound, which is sufficient for our purposes, and present the argument in probabilistic terms.

Suppose, in contrary to Theorem 1, that for every subset $S \subset \{0,1\}^n$ of cardinality $K \stackrel{\rm def}{=} 2^k$ there exists a circuit of size $o(2^{2k-m}/(2k-m))$ that perfectly hashes S into $\{0,1\}^k$. We will show that each circuit can serve as a perfect hashing for too few K-subsets, and hence there are too few circuits to perfectly hash all possible K-subsets. The main observation follows:

Lemma 1.1 (the fraction of sets that are perfectly hashed by any function): *For any $m \leq n-1$, let $C : \{0,1\}^n \to \{0,1\}^m$ be an arbitrary circuit, and let $S \subset \{0,1\}^n$ be a uniformly selected subset of cardinality $K = 2^k$. Then, the probability that C perfectly hashes S into $\{0,1\}^m$ is at most $2^{-\Omega(2^{2k-m})}$.*

Proof: Let $N \stackrel{\rm def}{=} 2^n$ and $M \stackrel{\rm def}{=} 2^m$. Clearly, we may assume that $k \leq m$ (as otherwise the probability is zero). Let $c_1, ..., c_M$ denote the sizes of the preimages of the various m-bit strings under C (i.e., $c_i = |C^{-1}(s_i)|$, where s_i denotes the $i^{\rm th}$ (m-bit long) string by some order). Then, the probability we are interested in is

$$\begin{aligned}
\frac{\sum_{I \subseteq [M]:|I|=K} \prod_{i \in I} \binom{c_i}{1}}{\binom{N}{K}} &\leq \frac{\binom{M}{K} \cdot (N/M)^K}{\binom{N}{K}} \\
&= \prod_{i=0}^{K-1} \frac{1-(i/M)}{1-(i/N)} \\
&= \exp\left\{ -\sum_{i=1}^{K-1} \ln\left(1 + \frac{(i/M)-(i/N)}{1-(i/M)}\right)\right\} \\
&< \exp\left\{ -\sum_{i=1}^{K-1} ((i/M)-(i/N))\right\} \\
&= \exp\left\{ -\frac{K \cdot (K-1)}{2} \cdot \left(\frac{1}{M} - \frac{1}{N}\right)\right\}
\end{aligned}$$

which for $M \leq N/2$ yields $2^{-\Omega(K^2/M)}$. The lemma follows. ■

Deriving Theorem 1. Adding up the contribution of all possible circuits, while applying Lemma 1.1 to each of them, we conclude that if too few circuits are considered then not all K-subsets can be perfectly hashed. Specifically, there are $s^{O(s)}$ possible circuits of size s, and so we need $s^{O(s)} \cdot 2^{-\Omega(2^{2k-m})} \geq 1$. Theorem 1 follows.

2 Proof of Theorem 2

We consider two cases. In the case that $m \leq k + \log_2 n$, the theorem follows by constructing an obvious circuit that maps each string in S to its rank (in S) represented as an m-bit long string. This circuit has size $|S| \cdot n \leq 2^{2k-m} \cdot n^2$ (since $k \leq 2k - m + \log_2 n$), and the theorem follows.

The less obvious case is when $m \geq k + \log_2 n$. Here we use a family of n-wise independent functions mapping $\{0,1\}^n$ onto $\{0,1\}^\ell$, where $\ell \stackrel{\rm def}{=} m - \log_2 n$. Function in such a family can be evaluated by poly(n)-size circuits (cf. [1]). We consider the collisions caused by a uniformly chosen function from this family applied to S. Specifically,

Lemma 2.1 (hashing by n-wise independence functions): *Let H be a family of functions $\{h:\{0,1\}^n \to \{0,1\}^\ell\}$ such that $\mathrm{Prob}_{h\in H}[\wedge_{i=1}^n h(\alpha_i) = \beta_i] = 2^{-n\ell}$, for every n distinct $\alpha_1, ..., \alpha_n \in \{0,1\}^n$ and for every $\beta_1, ..., \beta_n \in \{0,1\}^\ell$. Then, for every $S \subset \{0,1\}^n$ of cardinality $2^k \leq 2^\ell$, there exists $h \in H$ such that*

1. *No value has more than n preimages under h; that is, $|h^{-1}(\beta) \cap S| \leq n$, for every $\beta \in \{0,1\}^\ell$.*
2. *At most $2^{2k-\ell}$ values have more than one preimage under h; that is, $|\{\beta \in \{0,1\}^\ell : |h^{-1}(\beta) \cap S| > 1\}| \leq 2^{2k-\ell}$.*

Proof: Fixing an arbitrary 2^k-subset, S, and uniformly selecting $h \in H$, we consider the probability that the two items (above) hold. Firstly, we consider the probability that h maps n elements of S to the same image. Using the n-wise independence of the family H, the probability of this event is bounded by

$$\binom{2^k}{n} \cdot 2^{-\ell n} \quad < \quad \frac{2^{kn}}{n!} \cdot 2^{-kn} \quad < \quad \frac{1}{2}$$

where the first inequality uses $\ell = m - \log_2 n \geq k$. Thus, the probability that Item (1) does not hold is less than 1/2. Next, we consider the probability that Item (2) does not hold. We start by using the pairwise independence of H to note that the collision probability is $2^{-\ell}$ (i.e., $\mathrm{Prob}_{h\in H}[h(\alpha_1) = h(\alpha_2)] = 2^{-\ell}$, for any $\alpha_1 \neq \alpha_2 \in \{0,1\}^n$). It follows that the expected number of h-images that have more than a single preimage in S is bounded above by the expected number of collisions; that is, by $\binom{2^k}{2} \cdot 2^{-\ell} < \frac{1}{2} \cdot 2^{2k-\ell}$. Applying Markov's Inequality, we conclude that the probability that Item (2) does not hold is less than 1/2. The lemma follows. $\blacksquare$

Deriving Theorem 2. Fixing an arbitrary 2^k-subset, $S \subset \{0,1\}^n$, and using Lemma 2.1, we present a circuit that perfectly hashes S into $\{0,1\}^m$ (where $m \geq k + \log_2 n$). Our construction uses the double hashing paradigm (see, e.g., [4]). Let $h:\{0,1\}^n \to \{0,1\}^{m-\log_2 n}$ be as guaranteed by the lemma (w.r.t the set S). We define a perfect hashing function $f:\{0,1\}^n \to \{0,1\}^m$ for S by letting

$$f(\alpha) \stackrel{\rm def}{=} h(\alpha) \circ \mathrm{rank}_{S\cap h^{-1}(h(\alpha))}(\alpha)$$

where $\mathrm{rank}_R(\alpha)$ is an $\log_2 n$-bit long string representing the rank of α among the elements of R. A circuit computing the function f is constructed as follows. For each β having more than a unique h-preimage in S, we maintain a table ranking these preimages in S. By Item (1) of Lemma 2.1 such a table need only contain n entries, whereas by Item (2) we only need $2^{2k-\ell}$ such tables. (If a string, α, does not appear in any of the tables, then $f(\alpha) = h(\alpha) \circ 0^{\log_2 n}$.) The size of the circuit is $\mathrm{poly}(n) + 2^{2k-\ell} \cdot n^2 = \mathrm{poly}(n) + 2^{2k-m} \cdot n^3$, and so Theorem 2 follows.

References

1. N. Alon, L. Babai, and A. Itai, "A fast and Simple Randomized Algorithm for the Maximal Independent Set Problem", *J. of Algorithms*, Vol. 7, pages 567–583, 1986.
2. L. Carter and M. Wegman, "Universal Classes of Hash Functions", *J. Computer and System Sciences*, Vol. 18, pages 143–154, 1979.
3. M. Fredman and J. Komlós, "On the Size of Separating Systems and Perfect Hash Functions", *SIAM J. Algebraic and Discrete Methods*, Vol. 5, pages 61–68, 1984.
4. M. Fredman, J. Komlós, E. Szemerédi, "Storing a Sparse Table with O(1) Worst Case Access Time", *Journal of the ACM*, Vol. 31, pages 538–544, 1984.
5. J. Korner and K. Marton, "New Bounds for Perfect Hashing via Information Theory", *Europ. J. Combinatorics*, Vol. 9, pages 523–530, 1988.
6. K. Mehlhorn, *Data Structures and Algorithms* (Vol. 1), EATCS Monographs on Theoretical Computer Science, 1984.
7. A. Nilli, "Perfect Hashing and Probability", *Combinatorics, Probability and Computing*, Vol. 3, pages 407–409, 1994.