The Summer Undergraduate Research Fellowship (SURF) Symposium
2 August 2018
Purdue University, West Lafayette, Indiana, USA

Expected Length of the Longest Chain in Linear Hashing

Pongthip Srivarangkul, Hemanta K. Maji
Department of Computer Science, Purdue University

ABSTRACT

Hash table with chaining is a data structure that chains objects with identical hash values together with an entry or a memory address. It works by calculating a hash value from an input then placing the input in the hash table entry. When we place two inputs in the same entry, they chain together in a linear linked list. We are interested in the expected length of the longest chain in linear hashing and methods to reduce the length because the worst-case look-up time is directly proportional to it.

The linear hash function used to calculate hash value is defined by $ax+b \bmod p \bmod m$, for any $x \in \{0,1, \ldots, p-1\}$ and a, b chosen uniformly at random from the set $\{0,1, \ldots, p-1\}$, where p is a prime and $p \geq m$. This class of hash functions is a 2-wise independent hash function family. For any 2-wise independent hash functions, the expected length of the longest chain is $O(n^{1/2})$. Additionally, Alon et al. (JACM 1999) proved that when using a similar class of 2-wise independent hash function, the expected length of the longest chain has a tight lower bound of $\Omega(n^{1/2})$. Recently, in 2016, Knudsen (FOCS 2016) showed that the upper bound of the expected length of the longest chain of the linear hashing function is surprisingly $n^{1/3+o(1)}$. This bound is strictly better than $O(n^{1/2})$, which, due to Alon et al.'s result, is already known to be tight for 2-wise independent hash functions. Consequently, there are exclusive properties of the linear hashing function, in addition to being 2-wise independent, that results in this phenomenon. Even though Knudsen's upper bound on the expected length of the longest chain is remarkable, it is still unknown whether it is tight. In other words, does there exist a set of n inputs such that, when hashed using the linear hash function, the expected length of the longest chain is roughly $n^{1/3}$. If Knudsen's bound is not tight, then there is an additional motivation to study further and tighten the upper bound.

Another focus of our research is to reduce the expected length of the longest chain by using the load balancing power of "two choices." The idea is, instead of choosing one bin (hash table entry) for a ball (input), to choose two or more bins and put the ball in the bin with the least load at that moment. Mitzenmacher et al. proved that the power of two choices exponentially improves the expected max-load (from $\Theta(\log n/\log \log n)$) to $\Theta(\log \log n)$) for the hash table that uses two truly random hash functions. We shall conduct an empirical study by simulation with SageMath (System for Algebra and Geometry Experimentation) to verify whether similar improvements are observed for the linear hash function as well. We anticipate that the length of the longest chain of our linear hash table can be significantly improved when used with two linear hash functions.

KEYWORDS

hashing, hashing with chaining, linear hashing, two choices paradigm, power of two choices