Springer - Verlag Lecture Notes in CS #64 1978

ANALYSIS OF A UNIVERSAL CLASS OF HASH FUNCTIONS

George Markowsky
J. Lawrence Carter
Mark N. Wegman

Computer Sciences Department
IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10566

Abstract: In this paper we use linear algebraic methods to analyze the performance of several classes of hash functions, including the class H_2 presented by Carter and Wegman [2]. Suppose H is a suitable class, the hash functions in H map A to B, S is any subset of A whose size is equal to that of B, and x is any element of A. We show that the probability of choosing a function from H which maps x to the same value as more than t other elements of S is no greater than $\min(1/t^2, 11/t^4)$.

Consider a database storage and retrieval system implemented using hashing and a linked list collision resolution strategy. A corollary of the main result is that the probability that the system would perform more than t times more slowly than expected is no greater than $\min(1/t^2, 11/t^4)$. The "performance" being considered can be either the number of memory references required to process any individual request or the number required to process an arbitrary sequence of requests.

Notice that these results do not assume that the requests to the database are random or uniformly distributed. Instead, the averaging is done over the possible choices of the actual hash function from H. Since the system designer can be sure that this choice is made randomly, the probabilities given hold for *any* input.

It is also shown that the bound on poor performance when balanced trees are used in place of linked lists is approximately $\min(1/(4t), 11/(16t))$. The formulas are generalized to any size S.

§ 1 Introduction

Hashing provides a fast method of storing information in and retrieving it from a large database. Only a constant number of memory references are required to process a request, on the average, no matter how large the database is. On the other hand, if the database were structured as a balanced tree, then on the order of log(n) references would be required when there are n items in the database.

A disadvantage of hashing, which has perhaps prevented its more widespread use, is that the good performance cannot be guaranteed - if the data is distributed poorly with respect to

the hash function, a request may require on the order of n memory references. Often, analyses of algorithms which use hashing make the assumption that the data is randomly distributed. But data that arise in real applications are not often truly random, and sometimes a systematic bias in the data makes the algorithm perform more poorly than predicted.

In [2], Carter and Wegman show that the average performance of a storage and retrieval system which uses hashing can be analyzed without considering the distribution of the input. This is done by finding a universal class of hash functions - a class with the property that given any input, the average performance of all the functions when used to hash that input is "good". It follows that for any input, "most" functions perform well on that input. Since one can make sure that the choice of function is random, "most" of the time the function chosen will perform well. Carter and Wegman [2] show that the average time required to process any sequence of requests with their approach is a linear function of the number of requests processed. In addition, they present several classes of hash functions which are universal and can be evaluated fairly easily.

In order to be mathematically certain that the predicted average performance will be achieved, it is necessary to change hash functions periodically. For some applications, such as a compiler, this is feasible. However in a large database system it may not be practical to change the hash function, since this would require moving all the data to new locations. Thus, we want some bound on the probability that a randomly chosen function will perform poorly. This paper gives an improved bound on this probability for the class H_2 over that given in [2].

For some applications of hashing, it is not enough to know that the average performance will be good. There may be some level of performance such that any worse performance would not be tolerated. For instance in an online application, we may want an assurance that no individual transaction will require more than, say, t times the expected time. This paper also gives a bound on the probability that such an undesirable transaction will occur.

Section 2 presents some background and defines the class H_2. Section 3 presents the mathematical analysis of the set of all linear operators between two vector spaces and in section 4 we use these results to derive our results about the classes of hash functions.

§ 2 Background

By an *associative memory* we mean a system which can perform the operation of STORE, RETRIEVE, and DELETE. STORE(KEY, DATA) stores DATA under the identifier KEY and overwrites any data previously associated with KEY. RETRIEVE(KEY) returns the data

associated with KEY, or returns NIL if there is no such data. DELETE(KEY) removes KEY and the associated data from the memory.

One method of implementing an associative memory is to use hashing with a linked list collision resolution strategy. In this implementation, there is a collection of b linked lists. There is also a function f, called a hash function, which maps the set of possible keys into the set {0,1,...,b-1}. Given a STORE, RETRIEVE or DELETE request, the hash function is applied to the given key. The resulting index is used to designate a linked list where the key and its associated data are to be stored. See [1], pages 111-113 for more details. Another implementation would to be to use balanced trees ([1], pages 145-157) in place of the linked lists.

Suppose S is the set of keys which have been inserted into the associative memory system. As we will justify in section 4, the time required to perform an operation involving the key x is determined by the number of elements of S which are mapped by f to f(x). Thus, we define $\Delta_{x,S}(f) = |\{y \in S \mid y \neq x,\ f(x)=f(y)\}|$. (Note: In [2], the notation $\delta_f(x,S)$ is used for this quantity.) Notice that $\Delta_{x,S}(f)$ makes sense whenever f is a function, x is a point in the domain of f and S is a subset of the domain of f.

We now define some classes of hash functions. Let $A = (\mathbb{Z}_2)^m$ (that is, the vector space of dimension m over the field of two elements) and $B = (\mathbb{Z}_2)^n$. One class of hash functions is L, the set of all linear transformation from A to B. (Although L is not explicitly mentioned in [2], it follows directly from that paper that L is a universal class of hash functions.) For some choice of j and k, let A' be the set $\{ (a_1,a_2,...,a_j) \mid a_i \in \{0,1,...,k-1\} \}$. [2] gives a definition of the class H_2 of function from A' to B. We give here an equivalent definition of that class. There is a fixed map g from A' to $A = (\mathbb{Z}_2)^{kj}$. Specifically, $g((a_1,a_2,...,a_j))$ is the vector which has 1's in positions $a_1+1, a_1+a_2+2, ..., j+\sum_{i=1}^{j} a_i$. The only property of g we need to use in this paper is that it is one-to-one. Letting L again be the set of linear transformations from A to B, $H_2 = \{T \circ g \mid T \in L\}$.

Since g is one-to-one, for all $x' \in A'$ and $S' \subset A'$, if $f = T \circ g \in H_2$, then $\Delta_{x',S'}(f) = \Delta_{x,S}(T)$ where $x = g(x')$ and $S = g(S')$. Thus the performance of H_2 is the same as the performance of L. We look at some of the structure of L in the next section.

§ 3 Linear Transformations Between Vector Spaces

Throughout this section A is an m-dimensional vector space over $\mathbb{Z}_2$, S is a subset of A and $x \in A\text{-}S$. We will think of $\Delta_{x,S}$ as a random variable over the sample space $L = \{T : A \to B \mid T \text{ is linear}\}$ with the uniform distribution. We use E, Var and σ to denote the expectation, variance and standard deviation of a random variable. A reader unfamiliar with these terms should consult Feller [3]. Whenever a summation is over several elements of a set, such as $\sum_{x,y \in S}$, each subset of distinct elements is used only once. Thus, if $S = \{a,b,c\}$, $\sum_{x,y \in S} \Delta_{x,y} = \Delta_{a,b} + \Delta_{a,c} + \Delta_{b,c}$.

Theorem 1 $E(\Delta_{x,S}) = \frac{|S|}{2^n}$ and $Var(\Delta_{x,S}) = \frac{|S|}{2^n}\left(1 - \frac{1}{2^n}\right) \le E(\Delta_{x,S})$.

Proof: For $y \in A$ and $T \in L$ we let $\Delta_{x,y}(T) = 1$ if $x \neq y$ and $Tx = Ty$

$= 0$ otherwise.

Clearly $\Delta_{x,S} = \sum_{y \in S} \Delta_{x,y}$ and $E(\Delta_{x,S}) = \sum_{y \in S} E(\Delta_{x,y})$. Since T is linear and we are working over $\mathbb{Z}_2$, $Tx = Ty$ iff $T(x+y) = 0$. Since $x \notin S$, $x \neq y$ so $x+y \neq 0$ and we may find a basis $x+y, a_1, \ldots, a_{m-1}$ of A. Since T is uniquely determined by choosing any of the 2^n points of B as a value for each basis vector, $|L| = (2^n)^m$. The number of linear transformation with $T(x+y) = 0$ is $(2^n)^{m-1}$, since we are free to choose any point in B as a value for each of the a_i.

Thus $E(\Delta_{x,y}) = \frac{(2^n)^{m-1}}{(2^n)^m} = \frac{1}{2^n}$ and $E(\Delta_{x,S}) = \frac{|S|}{2^n}$.

Now $Var(\Delta_{x,S}) = E(\Delta_{x,S}^2) - E^2(\Delta_{x,S}) = \sum_{y \in S} E(\Delta_{x,y}^2) + 2 \sum_{y,z \in S} E(\Delta_{x,y}\,\Delta_{x,z}) - E^2(\Delta_{x,S})$

$= \sum_{y \in S} E(\Delta_{x,y}) + 2 \sum_{y,z \in S} E(\Delta_{x,y}\,\Delta_{x,z}) - E^2(\Delta_{x,S})$ since $\Delta_{x,y} = \Delta_{x,y}^2$.

Thus we have $Var(\Delta_{x,S}) = E(\Delta_{x,S}) - E^2(\Delta_{x,S}) + 2 \sum_{y,z \in S} E(\Delta_{x,y}\,\Delta_{x,z})$,

where $E(\Delta_{x,S}) = \frac{|S|}{2^n}$ and $E^2(\Delta_{x,S}) = \frac{|S|^2}{2^{2n}}$.

For all $y,z \in S$ with $y \neq z$, $x+y$ and $x+z$ are linearly independent. Thus we can choose a basis $x+y, x+z, a'_1, \ldots, a'_{m-2}$ of A. Now $\Delta_{x,y}\,\Delta_{x,z} = 1$ iff $Tx = Ty = Tz$ iff $T(x+y) = T(x+z) = 0$. Arguing as before we see that $E(\Delta_{x,y}\,\Delta_{x,z}) = \frac{(2^n)^{m-2}}{(2^n)^m} = \frac{1}{2^{2n}}$.

Thus $Var(\Delta_{x,S}) = \frac{|S|}{2^n} - \frac{|S|^2}{2^{2n}} + \frac{|S|(|S|-1)}{2^{2n}} = \frac{|S|}{2^n}(1 - \frac{1}{2^n})$. □

With the information in the above theorem, we can use Chebyshev's inequality (see [3]) to derive some basic information about the distribution of $\Delta_{x,S}$.

Theorem 2 Prob($|\Delta_{x,S} - E(\Delta_{x,S})| > t) \leq \frac{Var(\Delta_{x,S})}{t^2}$. □

We will derive more information about this distribution by using the fourth central moment (see [3]). The following theorem is proved in a manner identical to the way Chebyshev's inequality is derived.

Theorem 3 Let X be a random variable with mean μ. Then Prob($|X-\mu| > t) \leq \frac{E((x-\mu)^4)}{t^4}$. □

Theorem 4 Let θ denote $E(\Delta_{x,S})$. Then

$$E((\Delta_{x,S} - \theta)^4) = 24 \sum_{y,z,w,t, \in S} E(\Delta_{x,y}\, \Delta_{x,z}\, \Delta_{x,w}\, \Delta_{x,t}) + (36-24\theta) \sum_{y,z,w, \in S} E(\Delta_{x,y}\, \Delta_{x,z}\, \Delta_{x,w})$$
$$+ (14-24\theta+12\theta^2) \sum_{y,z \in S} E(\Delta_{x,y}) + (\theta-4\theta^2+6\theta^3-3\theta^4)$$

where as before, the sums are over sets of indices no two of which are equal. (Bounds for the sums are derived in the lemmas which follow the proof.)

Proof: Using just the properties of the expectation, we can derive

(Equation *) $E((\Delta_{x,S} - \theta)^4) = E(\Delta^4_{x,S}) - 4\theta E(\Delta^3_{x,S}) + 6\theta^2 E(\Delta^2_{x,S}) - 3\theta^4$.

The argument in Theorem 1 shows that $E(\Delta^2_{x,S}) = \theta + 2 \sum_{y,z \in S} E(\Delta_{x,y}\, \Delta_{x,z})$.

Next, we obtain

$$E(\Delta^3_{x,S}) = \sum_{y \in S} E(\Delta^3_{x,y}) + 3 \sum_{y \in S} \sum_{\substack{z \in S \\ z \neq y}} E(\Delta^2_{x,y}\, \Delta_{x,z}) + 6 \sum_{y,z,w \in S} E(\Delta_{x,y}\, \Delta_{x,z}\, \Delta_{x,w})$$

$$= \theta + 6 \sum_{y,z \in S} E(\Delta_{x,y}\, \Delta_{x,z}) + 6 \sum_{y,z,w \in S} E(\Delta_{x,y}\, \Delta_{x,z}\, \Delta_{x,w}),$$

since $\Delta^2_{x,y}\Delta_{x,z} = \Delta_{x,y}\, \Delta_{x,z} = \Delta_{x,y}\, \Delta^2_{x,z}$ and $\Delta^3_{x,y} = \Delta_{x,y}$.

Finally, we can derive

$$E(\Delta^4_{x,S}) = \sum_{y \in S} E(\Delta^4_{x,y}) + 4 \sum_{y \in S} \sum_{\substack{z \in S \\ y \neq z}} E(\Delta^3_{x,y}\, \Delta_{x,z}) + 6 \sum_{y,z \in S} E(\Delta^2_{x,y}\, \Delta^2_{x,z})$$

$$+ 12 \sum_{y \in S} \sum_{\substack{z,w \in S \\ y \neq z \\ y \neq w}} E(\Delta^2_{x,y} \Delta_{x,z}\, \Delta_{x,w}) + 24 \sum_{y,z,w,t} E(\Delta_{x,y}\, \Delta_{x,z}\, \Delta_{x,w}\, \Delta_{x,t})$$

$$= \theta + 14 \sum_{y,z \in S} E(\Delta_{x,y}\, \Delta_{x,z}) + 36 \sum_{y,z,w} E(\Delta_{x,y}\, \Delta_{x,z}\, \Delta_{x,w})$$

$$+ 24 \sum_{y,z,w,t} E(\Delta_{x,y}\, \Delta_{x,z}\, \Delta_{x,w}\, \Delta_{x,t}).$$

The theorem follows upon substituting the above values into (Equation *) and simplifying. □

The next lemma follows from the proof of Theorem 1. We retain the convention that whenever k indices appear under a single summation sign without any other indications, we are to sum over all subsets of S consisting of k distinct elements.

Lemma 5 $\sum_{y,z \in S} E(\Delta_{x,y}\, \Delta_{x,z}) = \frac{\binom{|S|}{2}}{2^{2n}}$. □

Lemma 6 $\frac{\binom{|S|}{3}}{2^{3n}} \leq \sum_{y,z,w \in S} E(\Delta_{x,y}\, \Delta_{x,z}\, \Delta_{x,w}) \leq \frac{|S|(|S|-1)}{6 \cdot 2^{2n}} + \frac{|S|(|S|-1)(|S|-3)}{6 \cdot 2^{3n}}$

Proof: Note that $\Delta_{x,y}\, \Delta_{x,z}\, \Delta_{x,w} \neq 0$ iff $\Delta_{x,y} = \Delta_{x,z} = \Delta_{x,w} = 1$ iff the span of Γ ($= \{y-x, z-x, w-x\}$) is in the kernel of T. Any two elements of Γ are linearly independent (e.g., $y-x = z-x$ would contradict $y \neq z$).

If the span of Γ has dimension 2, $E(\Delta_{x,y}\, \Delta_{x,z}\, \Delta_{x,w}) = \frac{1}{2^{2n}}$ while if the span of Γ has dimension 3, $E(\Delta_{x,y}\, \Delta_{x,z}\, \Delta_{x,w}) = \frac{1}{2^{3n}}$. Now the lower bound follows since we have $\binom{|S|}{3}$ triples to consider and $E(\Delta_{x,y}\Delta_{x,z}\Delta_{x,w}) \geq \frac{1}{2^{3n}}$.

To derive the upper bound, it suffices to determine the largest possible number of triples y,z,w such that Γ spans a two dimension space. Equivalently, we can determine the smallest possible number of triples for which Γ will span a three dimensional space. This last number is $\frac{|S|(|S|-1)(|S|-3)}{3!}$. To see this, observe that there are $|S|(|S|-1)$ ways (counting order) of choosing the first two terms the triple. For the third term there are at least $|S|-3$ choices: omit the first two y,z and at most one more element $y+z+x$. Since we want unordered triples we divide by 3!. Thus the largest possible number of triples which span a two dimension space

is $\binom{|S|}{3} - \frac{|S|(|S|-1)(|S|-3)}{3!} = \frac{|S|(|S|-1)}{3!}$ and the upper bound follows. □

Lemma 7 $\frac{\binom{|S|}{4}}{2^{4n}} \leq \sum_{y,z,w,t \in S} E(\Delta_{x,y}\, \Delta_{x,z}\, \Delta_{x,w}\, \Delta_{x,t})$

$$\leq \frac{|S|(|S|-1)(|S|-3)(|S|-7)}{4!\, 2^{4n}} + \frac{5|S|(|S|-1)(|S|-3)}{4!\, 2^{3n}}.$$

Proof: The proof is similar to the proof of Lemma 6 so that we shall briefly sketch the differences. For distinct y, z, w and t, $\{y-x, z-x, w-x, t-x\}$ spans either a 3 or 4 dimensional subspace (it cannot be 2-dimensional since there are only 3 distinct nonzero vectors in a 2 dimensional subspace). If it is 3 dimensional, the corresponding expectation is $\frac{1}{2^{3n}}$, etc. Arguing as in Lemma 6, we see that there are at least $\frac{|S|(|S|-1)(|S|-3)(|S|-7)}{4!}$

unordered 4-tuples which span a 4-dimensional subspace of A. □

Theorem 8 Let $\theta = E(\Delta_{x,S})$ as in Theorem 4. Then

$$E((\Delta_{x,S} - \theta)^4) \leq \theta^3 + 9\theta^2 + \theta \text{ if } \theta \leq 3/2$$

$$\leq 5\theta^3 + 3\theta^2 + \theta \text{ if } \theta \geq 3/2.$$

Proof: For $\theta \leq 3/2$, the coefficients of the first two sums appearing on the right-hand side of the equation of Theorem 4 are all nonnegative. Thus we can bound $E((\Delta_{x,S} - \theta)^4)$ by substituting the value derived in Lemma 5 and the upper bounds of Lemmas 6 and 7 into the right hand side of the equation of Theorem 4. Note also that quantities of the form $(|S|-k)/2^n$ are $\leq |S|/2^n = \theta$. Simplification yields $\theta^3+9\theta^2+\theta$.

For $\theta \geq 3/2$, the procedure is almost identical except that the coefficient $36-24\theta$ is ≤ 0. We substitute as before, except that we must now use the lower bound derived in Lemma 6. Simplification yields $5\theta^3+3\theta^2+\theta$. □

Example 9 We now indicate how the preceding results are to be used. Suppose $|S| = 2^n$. Then $\theta = E(\Delta_{x,S}) = 1$. Theorems 1 and 2 imply that $\text{Prob}(|\Delta_{x,S} - 1| > t)$ is bounded above by $1/t^2$, while Theorems 3 and 8 imply that it is bounded above by $11/t^4$. Thus it is bounded by $\min(1/t^2, 11/t^4)$. □

The following lemma will be of use in analyzing sequences of requests to an associative memory. It allows us to bound the variance and fourth central moment of a sum of random variables in terms of the values of these operators on the summands.

Lemma 10 Let $\{X_i\}_{i=1,\ldots,n}$ be a set of random variables defined on the discrete sample space $U = \{x_1,\ldots,x_p\}$. Let $X = \sum_{i=1}^{n} X_i$. Then $E((X-E(X))^m) \leq n^m \max\{E((X_i - E(X_i))^m) \mid i=1,\ldots,n\}$ for m=2 or 4.

Proof: Let $\mathbb{R}^p$ denote the usual Euclidean space. For m = 2 or 4 and $A = (a_1,\ldots,a_p)$ in $\mathbb{R}^p$, we define $\|A\|_m = \left(\sum_{i=1}^{p} a_i^m\right)^{1/m}$. The Minkowski inequality (see Naylor and Sell [4]) implies that for all $A_1,\ldots,A_n \in \mathbb{R}^p$, $\|\sum_{i=1}^{n} A_i\|_m \leq \sum_{i=1}^{n} \|A_i\|_m \leq n \max\{\|A_i\|_m \mid i=1,\ldots,n\}$.

Let $\bar{V} = (v_1,\ldots,v_p)$, $\bar{V}_i = (v_{i1},\ldots,v_{ip})$, $\bar{E} = (e_1,\ldots,e_p)$ and $\bar{E}_i = (e_{i1},\ldots,e_{ip})$ be elements of $\mathbb{R}^p$ where $i=1,\ldots,n$, $v_j = X(x_j)\,(\Pr(x_j))^{1/m}$, $v_{ij} = X_i(x_j)\,(\Pr(x_j))^{1/m}$, $e_j = E(X)\,(\Pr(x_j))^{1/m}$ and $e_{ij} = E(X_i)\,(\Pr(x_j))^{1/m}$ for $i=1,\ldots,n$ and $j=1,\ldots,p$ (recall $x_j \in U$).

The following facts should now be clear:

(i) $\bar{V} - \bar{E} = \sum_{i=1}^{n} (\bar{V}_i - \bar{E}_i)$;

(ii) $\| \bar{V} - \bar{E} \|_m = (E((X-E(X))^m))^{1/m}$;

(iii) $\| \bar{V}_i - \bar{E}_i \|_m = (E((X_i - E(X_i))^m))^{1/m}$.

It follows that $E((X-E(X))^m)^{1/m} \leq n \max\{E((X_i - E(X_i))^m)^{1/m} \mid i=1,\ldots,n\}^{1/m}$ from which the theorem follows immediately. □

§ 4 Analysis of the Class H_2

We now relate the results of the previous section to the classes H_2 and L of hash functions defined in section 2. Suppose one implements an associative memory using the linked list collision resolution strategy with hash function f. If S is the set of keys which have been inserted into the associative memory and r is a request involving the key x, then a reasonable definition of the cost (performance) of r is $C_L(f,r,S) = 1 + \Delta_{x,S}(f)$. If keys are stored on some secondary storage device such as a disk, this cost function gives the maximum number of secondary storage references required to process r. This maximum is required if either x is the last element of the list being searched, or r is an "insert" request and $x \notin S$. The subscript "L" in $C_L(f,r,S)$ reminds us that we are considering the linked list strategy.

If R is a sequence of requests and S is the set of keys which are inserted by requests in R, then we can define the cost of processing R to be $C_L(f,R) = \sum_{r \in R} C_L(f,r,S)$. Again, this cost is an upper bound for the number of secondary storage references necessary to process R, assuming the associative memory had nothing stored in it at the beginning.

If the associative memory is implemented using balanced trees in place of the linked lists, then appropriate cost functions would be $C_T(f,r,S) = 2 + \log_2(1 + \Delta_{x,S}(f))$ and $C_T(f,R) = \sum_{r \in R} C_T(f,r,S)$

The remaining theorems are direct applications of the work of the previous section. They are stated in terms of the class H_2, but also apply to the class L.

Theorem 11 Suppose S is any set of keys and r is any request. Let θ be the load factor $\frac{|S|}{2^n}$. If f is chosen at random from H_2, then

$$\Pr[\,|\,C_L(f,r,S)-(1+\theta)\,|\,> t] \leq \min\left(\frac{\theta}{t^2}, \frac{m(\theta)}{t^4}\right)$$

where $m(\theta) = \theta^3+9\theta^2+\theta$ if $\theta \le \frac{3}{2}$

$= 5\theta^3+3\theta^2+\theta$ if $\theta \ge \frac{3}{2}$.

Proof Suppose r is a request involving the key x. Since $C_L(f,r,S) = 1 + \Delta_{x,S}(f)$, theorem 1 tells us that the expectation of $C_L(f,r,s)$ is $1+\theta$ and the variance is θ. Chebyshev's inequality (Theorem 2) then tells us $\Pr[\ |C_L(f,r,S) - (1+\theta)\ | > t\ ] \le \frac{\theta}{t^2}$.

Since $C_L(f,r,S) - (1+\theta) = \Delta_{x,S}(f) - \theta$, Theorem 8 tells us that $E((C_L(f,r,S) - (1+\theta))^4) \le m(\theta)$. Theorem 3 then implies that $\Pr[\ |C_L(f,r,S) - (1+\theta)| > t] \le \frac{m(\theta)}{t^4}$. □

Theorem 12 Let $R = \langle r_1, r_2,\ldots,r_k\rangle$ be any sequence of requests, and let $\theta = \frac{|S|}{2^n}$ where S is the set of keys inserted by requests in R. If f is chosen at random from H_2, then $\Pr[\ |C_L(f,R) - k(1+\theta)| > kt] \le \min\left(\frac{\theta}{t^2}, \frac{m(\theta)}{t^4}\right)$, where m is defined as in Theorem 11.

Proof $E(C_L(f,R)) = \sum_{i=1}^{k} E(C_L(f, r_i, S)) = k(1+\theta)$. Lemma 10 tells us the variance and 4^{th} moment of $C_L(f,R)$ are bounded by, respectively, $k^2\theta$ and $k^4m(\theta)$. Thus, as in theorem 11,

$\Pr[\ |C_L(f,R) - k(1+\theta)| > kt] \le \min\left(\frac{k^2\theta}{(kt)^2}, \frac{k^4m(\theta)}{(kt)^4}\right) = \min\left(\frac{\theta}{t^2}, \frac{m(\theta)}{t^4}\right)$. □

Note: The bound on the probability given in Theorem 12 cannot be improved without an improvement in Theorem 11. To see this, pick $k >> |S|$ and let R consist of insertions of the elements of S, followed by $k - |S|$ identical request r. Then $C_L(f,R) \approx k\ C_L(f,r,S)$ so

$\Pr[\ |C_L(f,R) - k(1+\theta)| > kt] \approx \Pr[\ |C_L(f,r,S) - (1+\theta)| > t]$.

Theorem 13 Suppose S is any set of keys and r is any request. Let θ and $m(\theta)$ be as in Theorem 11. If f is chosen at random from H_2, then

$$\Pr[\ C_T(f,r,S) > t] \le \min\left(\frac{\theta}{(e(t))^2}, \frac{m(\theta)}{(e(t))^4}\right) \quad \text{where } e(t) = 2^{t-2} - (\theta+1).$$

Proof: Recall that $C_T(f,r,S) = 2 + \log_2(1 + \Delta_{x,S}(f)) = 2 + \log_2(C_L(f,x,S))$.

If $C_T(f,r,S) > t$, then $\log_2(C_L(f,x,S)) > t-2$, so $C_L(f,x,S) - (1+\theta) > 2^{t-2} - (\theta+1) = e(t)$. Thus $\Pr[C_T(f,r,S)>t] \le \Pr[|C_L(f,x,S) - (1+\theta)| > e(t)]$. Applying Theorem 11 to the right side completes the proof. □

The following lemma will be used in the proof of Theorem 15. The proof is omitted, but the lemma follows directly from the fact that the logarithm function has a nonpositive second derivative over the interval from a_1 to a_n.

Lemma 14 Let $a_1 \le a_2 \le \ldots \le a_n$ be positive numbers and $\lambda_1,\ldots,\lambda_n$ nonnegative numbers which sum to 1. Then $\sum_{i=1}^{n} \lambda_i a_i \in [a_1,a_n]$ and $\sum_{i=1}^{n} \lambda_i \log_2 a_i \le \log_2\left(\sum_{i=1}^{n} \lambda_i a_i\right)$.

Theorem 15 Let $R = \langle r_1,r_2,\ldots,r_k \rangle$ be any sequence of requests, and let $\theta = \frac{|S|}{2^n}$ where S is the set of keys inserted by requests in R. Let $m(\theta)$ and $e(t)$ be as in Theorem 13. Then if f is chosen at random from H_2,

$$\Pr[\, C_T(f,R) > kt] \le \min\left(\frac{\theta}{(e(t))^2}, \frac{m(\theta)}{(e(t))^4} \right)$$

Proof Suppose $C_T(f,R) > kt$. Then $\sum_{i=1}^{k}\left(2 + \log\left(C_L(f,x_i,S)\right)\right) > kt$

so $\sum_{i=1}^{k} \frac{1}{k} \log\left(C_L(f,x_i,S)\right) > t-2$. By Lemma 14 this implies $\log\left(\sum_{i=1}^{k} \frac{1}{k} C_L(f,x_i,S)\right) > t-2$

so $\sum_{i=1}^{k} \frac{1}{k} C_L(f,x_i,S) > 2^{t-2}$ or $\left(\sum_{i=1}^{k} C_L(f,x_i,S)\right) - k(\theta+1) > k(2^{t-2} - (\theta+1)) = k\, e(t)$.

Thus, $\Pr[C_T(f,R) > t] \le \Pr[\, |C_L(f,R) - k(\theta+1)| > k\, e(t)\,]$.

Applying Theorem 12 completes the proof. □

References

[1] Aho, A. V., Hopcroft, J. E., and Ullman, J. D., The Design and Analysis of Computer Algorithms, Addison-Wesley, Reading Mass. (1974).

[2] Carter, J. L. and Wegman M. N., "Universal Classes of Hash Functions," Proceedings of Ninth Annual SIGACT Conference May, 1977.

[3] Feller, W., An Introduction to Probability Theory and Its Applications, John Wiley, New York (1957).

[4] Naylor, A. W. and G. R. Sell, Linear Operator Theory, Holt, Rinehart & Winston, New York (1971).