The Annals of Probability
1987, Vol. 15, No. 3, 1180–1202

ON PROBABILISTIC ANALYSIS OF A COALESCED HASHING ALGORITHM

BY B. PITTEL

The Ohio State University

An allocation model [n balls, m ($\geq n$) cells, at most one ball in a cell] related to a hashing algorithm is studied. A ball x goes into the cell $h(x)$, where $h(\cdot)$: $\{1,\ldots,n\} \to \{1,\ldots,m\}$ is random. In case the cell $h(x)$ is already occupied, the ball x is rejected and moved into the leftmost empty cell. This empty cell is found via the sequential search from left to right starting with the cell occupied by the last (before x) rejected ball. Denote $T_2(x)$ the number of the necessary probes. In the end, due to a resulting system of references, the n occupied cells form a disjoint union of ordered chains, and to locate a ball x it suffices to search only the cells of a subchain originating at the cell $h(x)$. Denote $T_1(x)$ the length of this subchain. The main result of the paper is: in probability,

$$\max T_1(x) = \log_b n - 2\log_b \log n + O(1),$$

$$\max T_2(x) = \log_b n - \log_b \log n + O(1),$$

as $n \to \infty$, if n/m is bounded away from 0, $b = (1 - e^{-n/m})^{-1}$.

1. Introduction. Suppose we have n different objects (keys, in computer science terminology) labelled $1, 2, \ldots, n$, which are to be allocated one after another in an m-long array of cells (a table), so that each cell would contain at most one key. (Naturally, $m \geq n$.) Assume also that there is *given* a function $h(\cdot)$: $\{1,\ldots,n\} \to \{1,\ldots,m\}$, so that, for $1 \leq x \leq n$, $h(x)$ is the index of a cell the key x is assigned to, if this cell is still empty. [In applications, the keys may be numbers, records, i.e., elements of a given key space K. Thus, we should begin with a function $H(\cdot)$: $K \to \{1,\ldots,m\}$. But if the keys to be allocated in the table are $k_1,\ldots,k_n$ we may and shall restrict our attention to $h(i) = H(k_i)$, $1 \leq i \leq n$.] An allocation algorithm must include instructions about how to handle collisions, that is, the situations when $h(\cdot)$ has the same value for several keys. In addition, an algorithm must be able to locate a key if it is already stored in the table.

Out of the many existing (so called hashing) algorithms meeting these conditions, we consider in this paper a coalesced hashing algorithm [4], [5], [12], [16], [24] and [25].

This is how it works. Suppose that the keys $1, 2, \ldots, x-1$ are already stored. Suppose also that the $x-1$ occupied cells form a union of disjoint subsets (chains), with an order $\leq$ induced by the labels of the keys occupying them,

Received May 1985; revised February 1986.

AMS 1980 *subject classifications.* 60C05, 60F99, 68P10, 68P20, 68R05, 05C80.

Key words and phrases. Search algorithm, hashing, probabilistic analysis, limiting distributions, largest search time.

having the property: If a key x' is in a cell y, then the cells $h(x')$, y belong to the same chain and $h(x') \leq y$. (Technically, the chains are specified by references associated with the cells; see the example that follows.) Now, if the cell $h(x)$ is not occupied by a key $x' < x$, then the key x moves into this cell, and a new chain of length 1 is formed. Otherwise, the key x is rejected and moves into the leftmost empty cell; the latter joins, as a new last cell, the chain which passes through the cell $h(x)$, thus producing a longer chain. [The empty cell itself is found via the sequential (left-to-right) search starting from the right neighbor of a cell occupied by the key rejected last before the key x.] In the end, all the keys are stored, and the n occupied cells form several disjoint chains. [If needed, we can locate now any key x in the table by searching the cells of a subchain which starts at the cell $h(x)$.]

Efficiency of the algorithm is measured by a sequence of the search times $\{T_1(x), T_2(x)\}_{1 \leq x \leq n}$. Here $T_1(x)$ is the length of a subchain connecting the cell $h(x)$ and a cell which actually contains the key x. $T_2(x)$ is the number of extra cells searched sequentially to accommodate the key x, in case the cell $h(x)$ is occupied by a key $x' < x$.

EXAMPLE. $m = 10$, $n = 8$ and $\{h(x)\colon 1 \leq x \leq 8\} = \{5, 9, 10, 5, 2, 9, 2, 1\}$. The keys are allocated as shown on Figure 1. The chains of cells are $\{10\}$, $\{2, 4\}$, $\{9, 3\}$ and $\{5, 1, 6\}$. Also, $\{T_1(x)\colon 1 \leq x \leq 8\} = \{1, 1, 1, 2, 1, 2, 2, 2\}$ and $\{T_2(x)\colon 1 \leq x \leq 8\} = \{0, 0, 0, 1, 0, 2, 1, 2\}$.

Assume that the hashing function $h(\cdot)$ is chosen at random, which means that $h(1), \ldots, h(n)$ are independent and each $h(x)$ is uniformly distributed on $\{1, \ldots, m\}$. Then the search times become random. Let also $n, m \to \infty$ so that $n = am$ and $0 < a \leq 1$ is fixed. (The parameter a is usually called the load factor.)

Exact and asymptotic formulas for $E(T_i(x))$ and $\sigma^2(T_1(x))$ were obtained in [16], and it was observed there that the *average* search times remain bounded even when $a = 1$, which corresponds to the completely filled table. (For the mean values analysis of other versions of the coalesced hashing, see also [4], [5], [15], [24] and [25].)

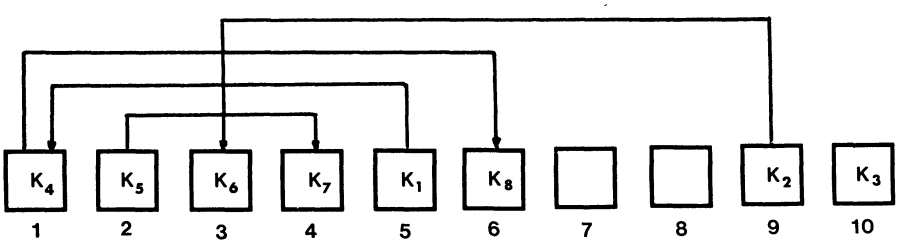

FIG. 1.

Our first result is

THEOREM 1. (a) $\lim P(T_1(n) = 1) = 1 - a$, *and for* $l \geq 0$

$$\lim P(T_1(n) = l + 2)$$

$$= a(1 + H_l) - \frac{a^2}{2} + \sum_{k=1}^{l} (-1)^k \left(\frac{k+1}{k^2}\right)\binom{l}{k}(1 - e^{-ka}), \tag{1.1}$$

where $H_l = 1/1 + \cdots + 1/l$. *Also*

$$P(T_1(n) = l) \sim e^a(1 - e^{-a})^{l-1}/l, \qquad a = n/m, \tag{1.2}$$

if $l \to \infty$ *and* $l^2 = o(n)$.

(b) $T_2(n) \to \delta(1 + \tau)$ *in distribution, where* δ *and* τ *are independent,* $P(\delta = 1) = 1 - P(\delta = 0) = a$ *and* τ *is geometrically distributed with parameter* e^{-a}, *i.e.,* $P(\tau = s) = e^{-a}(1 - e^{-a})^s$, $s \geq 0$.

Consider now $U_i(n) = \max\{T_i(x)\colon 1 \leq x \leq n\}$, $i = 1, 2$; these random variables represent the largest search times. It turns out that $U_i(n)$ grow logarithmically with n. More precisely, we shall prove

THEOREM 2. *In probability,*

$$U_1(n) = \log_b n - 2\log_b \log n + O(1), \tag{1.3}$$

$$U_2(n) = \log_b n - \log_b \log n + O(1), \tag{1.4}$$

where $b = (1 - e^{-a})^{-1}$.

NOTES. (1) It is quite surprising that $U_1(n)$ and $U_2(n)$ behave so similarly, since the corresponding searches are very different. (2) Consider $T(x) = T_1(x) + T_2(x)$ which is the total number of probes needed to accommodate the key x, $1 \leq x \leq n$, and denote $U(n) = \max\{T(x)\colon 1 \leq x \leq n\}$. In view of Theorem 2,

$$\log_b n(1 + o(1)) \leq U(n) \leq 2\log_b n(1 + o(1))$$

in probability, and we conjecture that, out of these two estimates, the upper case is sharp.

The reader interested in a description and probabilistic analysis of other hashing techniques (linear probing, double hashing, etc.) and related search algorithms can look at [1], [3], [6]–[16], [19]–[23] and [26]. (The list is by no means complete!) Undoubtedly, the invention of hashing algorithms and a need to analyze them have stimulated a new interest in the well known combinatorial schemes—random allocations [17] and mappings [18].

2. The proof of $T_1(n)$ and $U_1(n)$ asymptotic characteristics. Denote by $\tilde{h}(x)$ the index of a cell which actually contains the key x. By the definition of

the algorithm, $T_1(n) = 1$ iff $h(n) \neq \tilde{h}(x)$, $x < n$. Since $h(n)$ is independent of $\tilde{h}(x)$, $x < n$, and is uniformly distributed,

$$P(T_1(n) = 1) = 1 - (n-1)/m \to 1 - a. \tag{2.1}$$

How can we evaluate $P(T_1(n) = l + 2)$, $l \geq 0$? Introduce $X(\nu, l)$ the number of the chains induced by the first ν keys, with the length *more* than l. Denote $E(\nu, l) = E(X(\nu, l))$. Clearly, $T_1(n) = l + 2$ iff the cell $h(n)$ belongs to a chain (induced by the first $n - 1$ keys) with length *at least* $l + 1$, and the length of the subchain connecting the cell $h(n)$ and the maximal cell of the chain is *exactly* $l + 1$. Hence, again by independence,

$$\begin{aligned} P(T_1(n) = l + 2) &= E\big(P(T_1(n) = l + 2 | \tilde{h}(x), x < n)\big) \\ &= E(X(n-1, l)/m) = E(n-1, l)/m. \end{aligned} \tag{2.2}$$

In view of (2.2), the relation (1.1) follows directly from

LEMMA 1. *For* $l + 1 \leq \nu \leq m$,

$$\begin{aligned} E(\nu, l) &= \nu(1 + H_l) - m^{-1}\binom{\nu}{2} \\ &\quad + m \sum_{k=1}^{l} (-1)^k \left(\frac{k+1}{k^2}\right)\binom{l}{k}\left[1 - \left(1 - \frac{k}{m}\right)^{\nu}\right], \\ H_l &= \sum_{k=1}^{l} \frac{1}{k}. \end{aligned} \tag{2.3}$$

PROOF. For $A = \{x_1 < \cdots < x_{l+1}\}$, $x_j \leq \nu$, put $\delta(A) = 1$ if the cells $\tilde{h}(x_1), \ldots, \tilde{h}(x_{l+1})$ form an initial segment of a chain; if not, put $\delta(A) = 0$. Observe that $\delta(A) = 1$ iff $h(x_1) \neq \tilde{h}(x)$ for $x < x_1$, $h(x) \neq \tilde{h}(x_1)$ $(= h(x_1))$ for $x_1 < x < x_2$, $h(x_2) = \tilde{h}(x_1), \ldots, h(x_l) \in \{\tilde{h}(x_1), \ldots, \tilde{h}(x_{l-1})\}$, $h(x) \notin \{\tilde{h}(x_1), \ldots, \tilde{h}(x_l)\}$ for $x_l < x < x_{l+1}$, and $h(x_{l+1}) \in \{\tilde{h}(x_1), \ldots, \tilde{h}(x_l)\}$. Obviously, $X(\nu, l) = \Sigma_A \delta(A)$, so

$$E(\nu, l) = \Sigma_A E(\delta(A)) = \Sigma_A P(\delta(A) = 1).$$

Here

$$\begin{aligned} P(\delta(A) = 1) &= [1 - (x_1 - 1)/m](1 - 1/m)^{x_2 - x_1 - 1}(1/m) \cdots \\ &\quad [(l-1)/m](1 - l/m)^{x_{l+1} - x_l - 1}(l/m) \\ &= (1 - s_0/m)(l!/m^l) \prod_{j=1}^{l} (1 - j/m)^{s_j}, \\ &\qquad s_j = x_{j+1} - x_j - 1, \; x_0 = 0. \end{aligned} \tag{2.4}$$

Therefore,

$$E(v,l) = (l!/m^l)\left[\Sigma_s(1 - s_0/m)\prod_{j=1}^{l}(1 - j/m)^{s_j}\right];$$

here the summation is taken over all the tuples $s = (s_0, \dots, s_{l+1})$, s_j being nonnegative integers and $\Sigma_{j=0}^{l+1} s_j = v - (l+1)$. In other words, the sequence $\{E(u + (l+1), l)(l!/m^l)^{-1}\}_{u \geq 0}$ is an $(l+2)$-fold convolution of $\{1 - u/m\}_{u \geq 0}$, $\{(1 - j/m)^u\}_{u \geq 0}$, $1 \leq j \leq l$, and $\{1\}_{u \geq 0}$. Hence

$$E(v,l) = (l!/m^l)\operatorname{coeff}_{z^{v-(l+1)}} F(z,l),$$

$$F(z,l) = \left[(1-z)^{-2} - z(1-z)^{-3}/m\right]\prod_{j=1}^{l}\left[1 - z(1 - j/m)\right]^{-1}, \qquad |z| < 1.$$

So, by the Cauchy formula,

$$E(v,l) = (l!/m^l)(2\pi i)^{-1}\int_C G(z)\,dz, \qquad G(z) = F(z,l)/z^{v-l};$$

here $C = \{z = re^{i\phi}\colon -\pi \leq \phi < \pi\}$, $r < 1$. Outside C, $G(z)$ has the poles $z_j = (1 - j/m)^{-1}$, $0 \leq j \leq l$. Hence, by the residue theorem,

$$E(v,l) = -(l!/m^l)\left[\operatorname{res}_{z=\infty} G(z) + \sum_{j=0}^{l}\operatorname{res}_{z=z_j} G(z)\right].$$

Here $\operatorname{res}_{z=\infty} G(z) = 0$ because $d\max\{|G(z)|\colon |z| = d\} \to 0$ as $d \to \infty$. A direct evaluation of the other residues leads, after simple but tedious manipulations, to the formula (2.3). □

Our next lemma yields immediately [see (2.2)] the formula (1.2), and it will be also used later to prove the relation (1.3).

LEMMA 2. *Suppose that* $l, v, m \to \infty$, $l^2/m = o(1)$ *and* v/m *is bounded away from* 0. *Then*

$$E(v,l) = me^v l^{-1}(1 - e^{-v})^{l+1}(1 + o(1)), \qquad v = v/m. \tag{2.5}$$

PROOF. Introduce two auxiliary functions

$$E_1(t; v, l) = tv + m\sum_{k=1}^{l}(-1)^k k^{-1}\binom{l}{k}\left[1 - \left(1 - t\frac{k}{m}\right)^v\right],$$

$$E_2(t; v, l) = -tvH_l + t^2 m^{-1}\binom{v}{2} - m\sum_{k=1}^{l}(-1)^k k^{-2}\binom{l}{k}\left[1 - \left(1 - t\frac{k}{m}\right)^v\right], \tag{2.6}$$

where $t \in [0,1]$. Clearly,

$$E(v,l) = E_1(1; v, l) - E_2(1; v, l). \tag{2.7}$$

Now, differentiating $E_1(t;\ v, l)$ and $E_2(t;\ v, l)$ with respect to t, we have

$$E_1'(t;\ v, l) = v \sum_{k=0}^{l} (-1)^k \binom{l}{k} \left(1 - t\frac{k}{m}\right)^{v-1}, \tag{2.8}$$

$$\begin{aligned} E_2'(t;\ v, l) &= -vH_l + tm^{-1}v(v-1) - v \sum_{k=1}^{l} (-1)^k k^{-1} \binom{l}{k} \left(1 - t\frac{k}{m}\right)^{v-1} \\ &= tm^{-1}v(v-1) + v \sum_{k=1}^{l} (-1)^k k^{-1} \binom{l}{k} \left[1 - \left(1 - t\frac{k}{m}\right)^{v-1}\right] \qquad (2.9) \\ &= m^{-1}v\left\{t(v-1) + m \sum_{k=1}^{l} (-1)^k k^{-1} \binom{l}{k} \left[1 - \left(1 - t\frac{k}{m}\right)^{v-1}\right]\right\} \\ &= m^{-1}vE_1(t;\ v-1, l) \quad (!). \end{aligned}$$

(We have used here the identity $H_l = \sum_{k=1}^{l} (-1)^{k-1} k^{-1} \binom{l}{k}$, which can be easily proved by induction.)

The sum in (2.8) has a nice probabilistic interpretation. Namely, consider an allocation scheme with w balls and m cells, which differs from the classical scheme in that a ball is accepted by a cell, chosen by the ball, with probability t independently of all other balls and cells. Consider some *fixed* l cells, and denote $P(t;\ w, l)$ the probability that none of these l cells is empty. Since the probability that some k fixed cells are empty equals $(1 - t(k/m))^w$, the inclusion–exclusion principle shows that

$$P(t;\ w, l) = \sum_{k=0}^{l} (-1)^k \binom{l}{k} \left(1 - t\frac{k}{m}\right)^{w}. \tag{2.10}$$

Combining (2.8) and (2.9) with (2.10), and using $E_i(0;\ v, l) = 0$ [see (2.6)], we get

$$E_1(1;\ v, l) = v\int_0^1 P(t;\ v-1, l)\, dt, \tag{2.11}$$

$$E_2(1;\ v, l) = m^{-1}v(v-1)\int_0^1 (1-t)P(t;\ v-2, l)\, dt. \tag{2.12}$$

To proceed, we need a sharp estimate of the function $P(t;\ w, l)$, and it could hardly be done directly via (2.10). Fortunately, using the probabilistic interpretation of $P(t;\ w, l)$, we can show that

$$P(t;\ w, l) = w!\, \mathrm{coeff}_{z^w}\left[e^z(1 - e^{-zt/m})^l\right]. \tag{2.13}$$

Further, an application of the saddle-point method yields that, uniformly over $t \in (0, 1]$,

$$P(t;\ w, l) \sim (1 - e^{-gt})^l, \qquad g = w/m, \tag{2.14}$$

provided that $l^2/m = o(1)$, $g \neq o(1)$. (See Appendix A.)

The rest is simpler. By (2.11) and (2.14),

$$E_1(1;\nu,l) \sim \nu\int_0^1 (1-e^{-g_1 t})^l\,dt, \qquad g_1=(\nu-1)/m. \tag{2.15}$$

Introducing $\psi(t)=\log(1-e^{-g_1 t})$, we observe $\psi'(1)=g_1(e^{g_1}-1)^{-1}$, $\psi''(t)\le 0$, $t\in(0,1]$ and is bounded for $t\in[\frac{1}{2},1]$. By Jensen's inequality,

$$\begin{aligned}\int_0^1 \exp[l\psi(t)]\,dt &\le \int_0^1 \exp\{l[\psi(1)+\psi'(1)(t-1)]\}\,dt\\ &= (1-e^{-g_1})^l\int_0^1 \exp[l\psi'(1)(t-1)]\\ &\sim (1-e^{-g_1})^l(l\psi'(1))^{-1}, \qquad l\to\infty,\\ &= e^{g_1}(lg_1)^{-1}(1-e^{-g_1})^{l+1}.\end{aligned} \tag{2.16}$$

On the other hand, denoting $t_0=1-l^{-2/3}$,

$$\begin{aligned}\int_0^1 (1-e^{-g_1 t})^l\,dt &\ge \int_{t_0}^1 (1-e^{-g_1 t})^l\,dt\\ &= \int_{t_0}^1 \exp\Big\{l[\psi(1)+\psi'(1)(t-1)]+O\big(l(t-1)^2\big)\Big\}\,dt\\ &\sim (1-e^{-g_1})^l(l\psi'(1))^{-1}\int_0^{l^{1/3}\psi'(1)} e^{-u}\,du\\ &\sim e^{g_1}(lg_1)^{-1}(1-e^{-g_1})^{l+1}, \qquad l(t_0-1)^2=l^{-1/3}.\end{aligned} \tag{2.17}$$

By (2.15)–(2.17),

$$E_1(1;\nu,l) \sim me^{v}l^{-1}(1-e^{-v})^{l+1}, \qquad v=\nu/m. \tag{2.18}$$

To estimate $E_2(1;\nu,l)$, observe first that, by the definition, $P(t;w,l)$ increases with w. Hence [see (2.11), (2.12) and (2.14)]

$$\begin{aligned}E_2(1;\nu,l) &\le (\nu/m)\nu\int_0^1 (1-t)P(t;\nu-1,l)\,dt\\ &\sim (\nu/m)\nu\int_0^1 (1-t)(1-e^{-g_1 t})^l\,dt\\ &= (\nu/m)\int_0^1 \Psi(t)\,dt,\end{aligned} \tag{2.19}$$

where

$$\Psi(t)=\nu\int_0^t (1-e^{-g_1 s})^l\,ds. \tag{2.20}$$

Now, the direct differentiation shows that $\Psi''(t)-lc\Psi'(t)\ge 0$, $t\in[0,1]$, $c=(e-1)^{-1}$; therefore, $\Psi'(t)-lc\Psi(t)\ge\Psi'(0)-lc\Psi(0)=0$, and

$$\int_0^1 \Psi(t)\,dt \le (lc)^{-1}\int_0^1 \Psi'(t)\,dt=(lc)^{-1}\Psi(1).$$

Together with (2.15), (2.19) and (2.20), it implies that

$$E_2(1;\nu,l)=O\big(l^{-1}E_1(1;\nu,l)\big)=o\big(E_1(1;\nu,l)\big)$$

(remember, $l \to \infty$). This relation, (2.7) and (2.18) combined complete the proof of Lemma 2. □

Let us turn to $U_1(n) = \max\{T_1(x): 1 \le x \le n\}$. Introduce

$$l_1 = \log_b n - 2\log_b\log_b n + \omega(n), \qquad l_2 = \log_b n - 2\log_b\log_b n - \omega(n), \tag{2.21}$$

where $\omega(n) \to \infty$ and $\omega(n) = o(\log n)$.

PROOF OF (1.3). We need to show that

$$P(U_1(n) > l_1) \to 0, \qquad P(U_1(n) > l_2) \to 1, \qquad n \to \infty.$$

Fix an integer $0 \le l \le n-1$. For a set $A = \{x_1 < \cdots < x_{L+1}\}$, $l \le L \le n-1$, put $\varepsilon(A) = 1$ if the cells $\tilde{h}(x_1), \ldots, \tilde{h}(x_{L+1})$ form an initial segment of a chain, and $T_1(x_1), \ldots, T_1(x_L) \le l$, but $T_1(x_{L+1}) \ge l+1$, that is,

$$h(x_j) \in \begin{cases} \{\tilde{h}(x_1), \ldots, \tilde{h}(x_{j-1})\}, & 2 \le j \le l, \\ \{\tilde{h}(x_{j-(l-1)}), \ldots, \tilde{h}(x_{j-1})\}, & l+1 \le j \le L, \\ \{\tilde{h}(x_1), \ldots, \tilde{h}(x_{L-(l-1)})\}, & j = L+1. \end{cases}$$

Otherwise, put $\varepsilon(A) = 0$.

Introduce the random variables $Y(n,l)$ and $Z(n,l)$ which are the total number of such segments and the total number of segments of length *exactly* $l+1$, respectively. Clearly,

$$Y(n,l) = \sum_{A:\,|A| \ge l+1} \varepsilon(A), \qquad Z(n,l) = \sum_{A:\,|A| = l+1} \varepsilon(A).$$

Also

$$P(U_1(n) > l) = P(Y(n,l) > 0) \le E(Y(n,l)), \tag{2.22}$$

and by Chebyshev's inequality,

$$P(U_1(n) > l) \ge P(Z(n,l) > 0) \ge E^2(Z(n,l))/E(Z^2(n,l)). \tag{2.23}$$

We will use (2.21) with $l = l_1$ and (2.22) with $l = l_2$. To estimate $E(Y(n,l))$, $E(Z(n,l))$ and $E(Z^2(n,l))$, we observe first that, by (2.4), and the definition of $\varepsilon(A)$,

$$\begin{aligned} E(\varepsilon(A)) &= P(\varepsilon(A) = 1) \\ &= (1 - s_0/m)\big[(l-1)!(l-1)^{L-l}(L-l+1)/m^L\big] \\ &\quad \times \prod_{j=1}^{L} (1 - j/m)^{s_j}, \qquad s_j = x_{j+1} - x_j - 1,\ x_0 = 0, \\ &= E(\delta(A))(l-1)^{L-l}(L-l+1)/(L)_{L-l+1}. \end{aligned} \tag{2.24}$$

Let $l \sim \log_b n$, $b = (1 - e^{-\alpha})^{-1}$, $\alpha = n/m$. Denote $\mu = [3\log_b n]$ and write

$$E(Y(n,l)) = \Sigma'_A E(\varepsilon(A)) + \Sigma''_A E(\varepsilon(A)) = \Sigma' + \Sigma'',$$

where $|A| \le \mu$ in Σ' and $|A| > \mu$ in Σ''. Since $\varepsilon(A) \le \delta(A)$ and $E(n, L)$ decreases as a function of L, we have by Lemma 2,

$$\Sigma'' \le \Sigma_A'' E(\delta(A)) = \sum_{L \ge \mu} \sum_{A:\,|A|=L+1} E(\delta(A)) = \sum_{L \ge \mu} E(n, L)$$
$$\le nE(n, \mu) = O(n^2 b^{-\mu}) = O(n^{-1}). \tag{2.25}$$

Further, by (2.24) (and Lemma 2 again),

$$\Sigma' = \sum_{L=l}^{\mu-1} E(n, L)(l-1)^{L-l}(L-l+1)/(L)_{L-l+1}$$
$$\sim me^{a}(l^2 b^{l+1})^{-1} f(n, l),$$
$$f(n, l) = \sum_{L=l}^{\mu-1} (b^{L-l}L/l)^{-1}(l-1)^{L-l}(L-l+1)/(L)_{L-l}.$$

Here

$$1 \le f(n, l) \le \sum_{L=l}^{\infty} (L-l+1)(b^{-1})^{L-l} = e^{2a} \le e^2.$$

Therefore, if $l = l_1$ [see (2.21)], then

$$\Sigma' = O\big(n(\log_b n)^{-2} b^{-l}\big) = O(b^{-\omega(n)}),$$

and [see (2.22) and (2.25)]

$$P(U_1(n) > l) \le \Sigma' + \Sigma'' = O(b^{-\omega(n)}) + O(n^{-1}) = o(1).$$

Let now $l = l_2$ [see (2.21)]. First of all, according to (2.24),

$$E(Z(n, l)) = \sum_{A:\,|A|=l+1} E(\varepsilon(A)) = l^{-1} \sum_{A:\,|A|=l+1} E(\delta(A)) = l^{-1}E(n, l)$$
$$\sim (e^a - 1)m(l^2 b^l)^{-1} = a^{-1}(e^a - 1)b^{\omega(n)} \to \infty. \tag{2.26}$$

Further, since $\varepsilon(A_1)\varepsilon(A_2) = 0$ for $A_1 \ne A_2$ and $A_1 \cap A_1 \ne \varnothing$,

$$E(Z^2(n, l)) = E(Z(n, l)) + \sum_{A_1, A_2} P(\varepsilon(A_1) = 1, \varepsilon(A_2) = 1), \tag{2.27}$$

where the sum is taken over all ordered pairs of disjoint subsets A_1, A_2 with $|A_1| = |A_2| = l + 1$.

LEMMA 3. *For all such pairs* (A_1, A_2),

$$P(\varepsilon(A_1) = 1, \varepsilon(A_2) = 1) \le c(l, m)P(\varepsilon(A_1) = 1)P(\varepsilon(A_2) = 1),$$

where

$$c(l, m) = [1 - 2(l+1)/m]^{-2(l+1)}.$$

PROOF. For $B \subset \{1, \ldots, n\}$, denote $I(B)$ the smallest interval which contains B, and $J(B) = I(B) \setminus B$. Introduce also $f_i(x)$, the total number of the keys

from A_i strictly less than x, $i = 1,2$. Then [see (2.24)]

$$P(\varepsilon(A_i) = 1) = [(l-1)!/m^l](1 - s_{0i}/m) \prod_{x \in J_i} (1 - f_i(x)/m), \tag{2.28}$$
$$J_i = J(A_i),$$

where $s_{0i} = x_{1i} - 1$, x_{1i} being the first key from A_i, $i = 1,2$. It is not difficult to see that

$$P(\varepsilon(A_1) = 1, \varepsilon(A_2) = 1) = [(l-1)!/m^l]^2(1 - s_{01}/m)(1 - s_{02}/m) \times \prod_{x \in J} (1 - f(x)/m), \qquad J = J(A_1 \cup A_2), \tag{2.29}$$

where

$$f(x) = \begin{cases} f_1(x) + f_2(x), & \text{if } \max(f_1(x), f_2(x)) \le l, \\ \min(f_1(x), f_2(x)), & \text{if } \max(f_1(x), f_2(x)) \ge l+1. \end{cases}$$

By (2.28),

$$\begin{aligned} &P(\varepsilon(A_1) = 1)P(\varepsilon(A_2) = 1) \\ &\quad = [(l-1)!/m^l]^2(1 - s_{01}/m)(1 - s_{02}/m)\Pi_1\Pi_2\Pi_3, \\ &\Pi_1 = \prod_{x \in J_1 \cap J_2} (1 - f_1(x)/m)(1 - f_2(x)/m), \\ &\Pi_2 = \prod_{x \in J_1 \setminus J_2} (1 - f_1(x)/m), \\ &\Pi_3 = \prod_{x \in J_2 \setminus J_1} (1 - f_2(x)/m). \end{aligned}$$

But $f_i(x) \le l$ on J_i, $i = 1,2$, so, by the definition of $f(\cdot)$,

$$\begin{aligned} (1 - f_1(x)/m)(1 - f_2(x)/m) &\ge (1 - (f_1(x) + f_2(x))/m) \\ &= (1 - f(x)/m), \qquad x \in J_1 \cap J_2, \\ 1 - f_i(x)/m &\ge 1 - f(x)/m, \qquad x \in J_i,\ i = 1,2. \end{aligned}$$

Hence,

$$\begin{aligned} &P(\varepsilon(A_1) = 1)P(\varepsilon(A_2) = 1) \\ &\quad \ge [(l-1)!/m^l]^2(1 - s_{01}/m)(1 - s_{02}/m) \prod_{x \in J_1 \cup J_2} (1 - f(x)/m). \end{aligned} \tag{2.30}$$

Besides, it can be checked that

$$J_1 \cup J_2 \subset J \cup (I_1 \cap A_2) \cup (I_2 \cap A_1),$$

so

$$|(J_1 \cup J_2) \setminus J| \le |I_1 \cap A_2| + |I_2 \cap A_1| \le |A_2| + |A_1| = 2(l+1). \tag{2.31}$$

Since $f(x) \le 2(l+1)$ as well, the combination of (2.29)–(2.31) leads to

$$P(\varepsilon(A_1) = 1)P(\varepsilon(A_2) = 1) \ge (1 - 2(l+1)/m)^{2(l+1)}P(\varepsilon(A_1) = 1, \varepsilon(A_2) = 1).$$

□

By this lemma and (2.26) and (2.27),

$$E(Z^2(n,l)) \le E(Z(n,l)) + c(l,m)E^2(Z(n,l))$$
$$\le E^2(Z(n,l))\left[1 + O(l^2/m) + O(b^{-\omega(n)})\right].$$

Hence [see (2.23)],

$$P(U_1(n) > l) \ge \left[1 + O(l^2/m) + O(b^{-\omega(n)})\right]^{-1} = 1 + o(1).$$

The proof of (1.3) is now complete. □

NOTE. Introduce $C(n)$, the length of the longest chain ("monster"; cf. [11]). It is obvious that $C(n) \ge T_1(n)$. Using $X(n,l)$ instead of $Y(n,l)$ and $Z(n,l)$, we could have proved in the same way that, in probability,

$$C(n) = \log_b n - \log_b \log n + O(1).$$

3. The proof of $T_2(n)$ and $U_2(n)$ asymptotic behavior. Recall that $T_2(x)$ is the number of extra cells searched sequentially to find an available cell for the key x, in case the cell $h(x)$ is occupied by a previous key, $1 \le x \le n$. [This search starts from the right neighbor of a cell occupied by the last (before x) rejected key, or from the cell 1 if no key $< x$ has been rejected.] In particular, $T_2(x) = 0$ iff $h(x) \notin \{\tilde{h}(x')\}_{x' < x}$. Denote $K(x) = \Sigma_{x' \le x} T_2(x')$ [if $K(x) \ne 0$, then it is the index of a cell filled by a key rejected last among the keys $\le x$]. We shall also need $L(x)$, the length of the maximal block of cells, beginning from the first cell, which are all occupied by the keys $\le x$. ["Maximal" means that no key $\le x$ occupies the cell $(L(x) + 1)$.] Put finally $M(x) = L(x) - K(x)$. (The expected values of $K(x)$, $L(x)$, $M(x)$ were obtained in [16].)

NOTE. Observe that $T_2(x) = \delta(x)[1 + M(x-1)]$, where $P(\delta(x) = 1) = 1 - P(\delta(x) = 0) = (x-1)/m$, and $\delta(x)$ is independent of $M(x-1)$ (in fact of the whole $\{h(x')\colon x' < x\}$).

LEMMA 4. *Let $k \le l \le x \le n$. If $k \ge 1$, then*

$$P(K(x) = k,\ L(x) \ge l) = m^{-x}\binom{m-l}{m-x}(x-1)^x P(x, x-1, x-k)$$
$$= m^{-x}\binom{m-l}{m-x}\sum_{j=0}^{x-k}(-1)^j\binom{x-k}{j}(x-1-j)^x, \tag{3.1}$$

where $P(u,v,w)$ is the probability that in the usual allocation model, with u balls and v cells, some w fixed cells are not empty. Also

$$P(K(x) = 0,\ L(x) \ge l) = m^{-x}\binom{m-l}{m-x}x!. \tag{3.2}$$

PROOF. The relation (3.2) is easy, since $K(x) = 0$ iff no cell is chosen by two or more keys $\le x$. [We say a cell y is chosen by a key x' if $h(x') = y$.]

Let $K(x) = k \ge 1$ and $L(x) = l$. By the definition of the algorithm and $K(x)$, $L(x)$, we have then (a) no key $\le x$ has chosen the cell k, (b) all the cells

$k+1,\dots,l$ are chosen by the keys $\le x$ and (c) exactly $x-l$ cells among those to the right from the cell l are chosen by the keys $\le x$.

Conversely, suppose that the conditions (a), (b), (c) are met. If $K(x) < k$ then [by (a)] the total number of the cells occupied by the keys $\le x$ is at most $x-1$ —impossible. If $K(x) > k$, then [by (b)] $K(x) > l$ and [by (c)] the keys $\le x$ occupy at least $x+1$ cells—impossible again. Hence, $K(x) = k$ and [see (b)] $L(x) \ge l$.

Consequently,

$$P(K(x)=k,\ L(x)\ge l) = P((\text{a}),(\text{b}),(\text{c}) \text{ are satisfied})$$
$$= \binom{m-l}{x-l}\left(\frac{x-1}{m}\right)^x P(x, x-1, x-k)$$
$$= \binom{m-l}{x-l}\left(\frac{x-1}{m}\right)^x \sum_{j=0}^{x-k}(-1)^j\binom{x-k}{j}\left(1-\frac{j}{x-1}\right)^x.$$

□

COROLLARY 1 (straightforward).

$$(3.3)\quad P(K(x)=k) = \begin{cases} m^{-x}\binom{m-k}{m-x}(x-1)^x P(x,x-1,x-k), & k\ge 1,\\ m^{-x}\binom{m}{m-x}x!, & k=0.\end{cases}$$

COROLLARY 2. *If $m-x = O(1)$, then $M(x) \Rightarrow G_1$, where G_1 is geometrically distributed with parameter e^{-1}.*

[In view of the note preceding Lemma 4, the last corollary implies Theorem 1(b) in case $a = 1$.]

PROOF OF COROLLARY 2. We may and shall assume that $m - x = \mu$ is fixed. Denote $R(x) = x - K(x)$. By Lemma 4, for *fixed* $r \ge s \ge 0$,

$$P(R(x)=r,\ M(x)\ge s) = P(K(x)=x-r,\ L(x)\ge x+s-r)$$
$$= \binom{\mu+r-s}{\mu}(1-x^{-1})^x(1+\mu x^{-1})^{-x}$$
$$\times \sum_{j=0}^{r}(-1)^j\binom{r}{j}(1-j(x-1)^{-1})^x$$
$$\to \binom{\mu+r-s}{\mu}(e^{-1})^{\mu+1}(1-e^{-1})^r$$

or

$$P(R(x)=r,\ M(x)=s)$$
$$\to \left[(e^{-1})^\mu(1-e^{-1})^{r-s}\binom{\mu-1+r-s}{\mu-1}\right]\left[e^{-1}(1-e^{-1})^s\right].$$

So

$$(R(x), M(x)) \Rightarrow \left(G_1^{(\mu)} + G_1, G_1\right),$$

where $G_1^{(\mu)}$ is negative binomially distributed with parameters e^{-1}, μ, and $G_1^{(\mu)}$ and G_1 are independent. [Just as easily, we could have proved that $M(x)$ and the lengths of all μ consecutive blocks of occupied cells to the right from the cell $L(x)$ are asymptotically independent and geometrically distributed with parameter e^{-1}.] □

To handle the case $m - x \to \infty$, and later $-U_2(n)$, the following two lemmas are needed.

LEMMA 5. *There exists an absolute constant c_0 such that*

$$P(u, v, w) \le c_0(u/rv)^{u+1/2}\exp(rv - u)(1 - e^{-r})^w, \qquad \forall\, r > 0. \tag{3.4}$$

LEMMA 6. *Introduce*

$$\bar{k} = \bar{k}(x) = m - (m - x)e^{\lambda}, \qquad \lambda = x/m. \tag{3.5}$$

For every $\varepsilon \in (0,1)$, there exists $\mu_0 = \mu_0(\varepsilon)$ such that for $x \ge \varepsilon m$ and $\mu = m - x \ge \mu_0$,

$$P\left(|K(x) - \bar{k}| \ge (m - x)^{3/5}\right) \le \exp\left[-(m - x)^{1/5}/30\right]. \tag{3.6}$$

PROOF OF LEMMA 5. Arguing as in the case of the probability $P(t;\, w, l)$ in Section 2 (see Appendix A), we have

$$\begin{aligned} P(u, v, w) &= (u!/v^u)\operatorname{coeff}_{z^u}\left\{\exp[z(v - w)](e^z - 1)^w\right\} \\ &= (u!/v^u)(2\pi)^{-1}\int_{-\pi}^{\pi} \mathscr{H}(re^{i\phi})\, d\phi, \end{aligned} \tag{3.7}$$

$$\mathscr{H}(z) = \exp[z(v - w)](e^z - 1)^w/z^u, \qquad r > 0.$$

Here

$$\begin{aligned} \left|\mathscr{H}(re^{i\phi})\right| &\le \mathscr{H}(r)\exp[(rv/2)(\cos\phi - 1)] \\ &\le \mathscr{H}(r)\exp\left(-c_1 rv\phi^2\right), \qquad c_1 > 0, \end{aligned} \tag{3.8}$$

since

$$\left|\exp(re^{i\phi})\right| = e^r\exp[r(\cos\phi - 1)]$$

and it can be checked that

$$\left|\exp(re^{i\phi}) - 1\right| \le |e^r - 1|\exp[r(\cos\phi - 1)/2]$$

(see [22, Appendix]).

Hence

$$P(u, v, w) \le c_0(u!/v^u)(rv)^{-1/2}\mathscr{H}(r), \qquad c_2 = 2^{-1}(\pi c_1)^{-1/2},$$

and, using an inequality $u! \le c_3 u^{1/2}(u/e)^u$, we obtain (3.4) with $c_0 = c_2 c_3$. □

PROOF OF LEMMA 6. First of all

$$|K(x) - \bar{k}| = |x - R(x) - \bar{k}| \tag{3.9}$$
$$= |R(x) - (m - x)(e^{\lambda} - 1)|, \qquad \lambda = x/m.$$

Further, by Lemmas 4 and 5 with $r = \lambda$, we have, for $0 \le s \le j < x$,

$$\begin{aligned} &P(R(x) = j,\ M(x) \ge s) \\ &\quad = P(K(x) = x - k,\ L(x) \ge x + s - k) \\ &\quad = m^{-x}(x-1)^x \binom{\mu + j - s}{\mu} P(x, x-1, j) \\ &\quad \le c_0 \binom{\mu + j - s}{\mu} m^{-x}(x-1)^x [m/(x-1)]^{x+1/2} (e^{-\lambda})^{\mu+1} (1 - e^{-\lambda})^j \qquad (3.10) \\ &\quad \le c \binom{\mu + j - s}{\mu} (e^{-\lambda})^{\mu+1} (1 - e^{-\lambda})^j \\ &\quad = cP\big(G_\lambda^{(\mu)} + G_\lambda = j,\ G_\lambda \ge s\big), \qquad \mu = m - x,\ c = 2c_0. \end{aligned}$$

[G_λ and $G_\lambda^{(\mu)}$ are independent, $G_\lambda(G_\lambda^{(\mu)})$ is geometrically (negative binomially) distributed with parameter $e^{-\lambda}$ (with parameters $e^{-\lambda}, \mu$).] Consequently,

$$P(R(x) = j) \le cP\big(G_\lambda^{(\mu')} = j\big), \qquad j < x,\ \mu' = \mu + 1. \tag{3.11}$$

Also [see (3.3)],

$$P(R(x) = x) = P(K(x) = 0) = (m)_x/m^x \le \exp[-x(x-1)/2m]. \tag{3.12}$$

Observe that

$$E\big(G_\lambda^{(\mu')}\big) = \mu' E(G_\lambda) = \mu'(e^{\lambda} - 1),$$

so [see (3.9), (3.11) and (3.12)]

$$\begin{aligned} P\big(|K(x) - \bar{k}| \ge \mu^{3/5}\big) &\le \exp[-x(x-1)/2m] \\ &\quad + cP\Big(\big|G_\lambda^{(\mu')} - E\big(G_\lambda^{(\mu')}\big)\big| \ge \chi(\mu')^{3/5}\Big), \end{aligned}$$

where $\chi \to 1$ as $\mu \to \infty$.

The rest of the proof follows a general idea, owing to Chernoff [2]. Namely, we write

$$P\big(G_\lambda^{(\mu')} \ge (\le) j\big) \le g(z)/z^j, \qquad z \ge (\le) 1,$$

where

$$g(z) = E(z^{G_\lambda^{(\mu')}}) = \big[E(z^{G_\lambda})\big]^{\mu'} = [p/(1 - qz)]^{\mu'}, \qquad p = 1 - q = e^{-\lambda}.$$

Choosing in each case z which minimizes $g(z)/z^j$, we obtain

$$P\big(G_\lambda^{(\mu')} \ge (\le) j\big) \le \exp[\Phi_{mx}(j)], \qquad j \ge (\le) q\mu'/p = E\big(G_\lambda^{(\mu')}\big), \tag{3.13}$$
$$\Phi_{mx}(y) = \mu' \log[p(\mu' + y)/\mu'] + y \log[q(\mu' + y)/y].$$

Now, since $\Phi_{mx}(q\mu'/p) = \Phi'_{mx}(q\mu'/p) = 0$ and

$$\Phi''_{mx}(y) = -\mu'/[y(\mu' + y)],$$

we easily get from (3.13) that ($p = e^{-\lambda}$)

$$P\big(|G_\lambda^{(\mu')} - E\big(G_\lambda^{(\mu')}\big)| \geq \chi(\mu')^{3/5}\big) \leq 2\exp\big[-p^2(m-x)^{6/5}\chi^2/4\mu'\big] \leq \exp\big[-(m-x)^{1/5}/29.6\big], \tag{3.14}$$

provided that μ' is large enough.

Using (3.12) (remember, $x \geq \varepsilon m$) and (3.14), we obtain (3.6). □

This lemma enables us to complete the proof of Theorem 1(b) by proving

LEMMA 7. *If* $\liminf_m x/m \in (0,1]$, $m - x \to \infty$ *and* $s = o((m-x)^{1/5})$, *then*

$$P(M(x) \geq s) \sim (1 - e^{-\lambda})^s, \qquad \lambda = x/m.$$

PROOF. By (3.2),

$$P(M(x) \geq s) = m^{-x}\binom{m-s}{m-x}x! + P_1 + P_2,$$

where

$$P_1 = P\big(M(x) \geq s, |K(x) - \bar{k}| \leq (m-x)^{3/5}\big),$$

$$P_2 = P\big(M(x) \geq s, |K(x) - \bar{k}| > (m-x)^{3/5}\big), \qquad \bar{k} = m - (m-x)e^{\lambda}.$$

It can be checked that, uniformly over k such that $|k - \bar{k}| \leq (m-x)^{3/5}$,

$$\binom{m-k-s}{m-x} = \binom{m-k}{m-x}\big[(x-\bar{k})/(m-\bar{k})\big]^s\big[1 + O\big(s/(m-x)^{2/5}\big)\big] = \binom{m-k}{m-x}(1-e^{-\lambda})^s(1+o(1)). \tag{3.15}$$

By (3.3), (3.15) and Lemma 6, we have then

$$P_1 = (1-e^{-\lambda})^s P\big(|K(x) - \bar{k}| \leq (m-x)^{3/5}\big)(1+o(1)) \sim (1-e^{-\lambda})^s.$$

It remains to observe that

$$P_2 \leq \exp\big[-(m-x)^{1/5}/30\big] = o\big[(1-e^{-\lambda})^s\big],$$

[$s = o((m-x)^{1/5})$] and

$$m^{-x}\binom{m-s}{m-x}x! \leq m^{-x}\binom{m}{m-x}x! = O(\exp(-\varepsilon^2 m)) = o\big[(1-e^{-\lambda})^s\big].$$

□

Next:

LEMMA 8. (a) *If* $\liminf_m x/m \in (0,1]$ *and* $s = o(m)$, *then*

$$P(M(x) \geq s) = O\big[(1-e^{-\lambda})^s\big], \qquad \lambda = x/m. \tag{3.16}$$

(b) *Denote* $\mathscr{M}(x) = \max\{M(x'): x' \le x\}$. *Then, uniformly over* $x \le m$, *and* $s \ge 0$,

$$P(\mathscr{M}(x) \ge s) = O\big[m(1 - e^{-\lambda})^s\big]. \tag{3.17}$$

PROOF. (a) According to (3.11) and (3.12),

$$\begin{aligned} P(M(x) \ge s) &\le \exp[-x(x-1)/2m] + c\sum_{j=s}^{\infty} P\big(G_\lambda^{(\mu)} + G_\lambda = j, G_\lambda \ge s\big) \\ &= \exp[-x(x-1)/2m] + cP(G_\lambda \ge s) = O\big[(1 - e^{-\lambda})^s\big]. \end{aligned}$$

(b) By the definition, $M(x')$ is the length of the block of cells *chosen* by the keys $\le x'$, which begins from the right neighbor of the cell *occupied* by a key rejected last among those keys. Hence, $\mathscr{M}(x) \le B(x)$, where $B(x)$ is the length of the longest block of cells *chosen* by the keys $\le x$. So,

$$P(\mathscr{M}(x) \ge s) \le P(B(x) \ge s) \le mP(x, m, s) \tag{3.18}$$

[see Lemma 4 for the definition of $P(u, v, w)$]. Applying Lemma 5 ($u = x$, $v = m$, $w = s$, $r = \lambda = x/m$), we get

$$P(\mathscr{M}(x) \ge s) \le c_0 m(1 - e^{-\lambda})^s. \qquad \square$$

With the last lemma at our disposal, we can now prove

LEMMA 9 [on an upper bound of $U_2(n)$]. *In probability,*

$$U_2(n) \le \log_b n - \log_b \log_b n + O(1).$$

PROOF. It suffices to estimate from above $\mathscr{M}(n)$, since $1 + M(x) \ge T_2(x)$, $1 \le x \le n$. Denote $n_1 = [n/2]$, $\mathscr{M}(n_1, n) = \max\{M(x): n_1 \le x \le n\}$. Let also $s = \log_b n - \log_b \log_b n + \omega(n)$, where $\omega(n) \to \infty$, $\omega(n) = o(\log n)$ and is otherwise arbitrary. Since

$$\mathscr{M}(n) = \max\{\mathscr{M}(n_1), \mathscr{M}(n_1, n)\},$$

we have

$$\begin{aligned} P(\mathscr{M}(n) \ge s) &\le P(\mathscr{M}(n_1) \ge s) + P(\mathscr{M}(n_1, n) \ge s) \\ &\le P(\mathscr{M}(n_1) \ge s) + \sum_{x=n_1}^{n} P(M(x) \ge s) \\ &= O\left\{n[1 - \exp(-n/2m)]^s + \sum_{x=n_1}^{n} [1 - \exp(-x/m)]^s\right\}. \end{aligned} \tag{3.19}$$

Here $[b = (1 - e^{-a})^{-1},\ a = n/m]$

$$\begin{aligned} n[1 - \exp(-n/2m)]^s &= nb^{-s}[(1 - \exp(-n/2m))/b]^s \\ &= O\big[(\log_b n)\rho^{\log_b n}\big] \qquad \big(\rho = [(1 - \exp(-n/2m))/b]^{1/2}\big) \\ &= O(n^{-c}), \qquad c > 0, \end{aligned}$$

since $\rho < 1$ and is bounded away from 1. Furthermore, estimating the sum in

(3.19) by an integral and arguing as in (2.20)–(2.23), we also have

$$(3.20)\quad \sum_{x=n_1}^{n} [1-\exp(-x/m)]^s \sim n\int_{1/2}^{1} [1-\exp(-at)]^s\,dt \sim c(a)ns^{-1}b^{-s},$$

$c(a) = e^a(ab)^{-1}$. By the choice of s,

$$ns^{-1}b^{-s} = b^{-\omega(n)}.$$

In view of (3.19) and (3.20), and $\omega(n) = o(\log n)$,

$$P(\mathcal{M}(n) \geq s) = O(b^{-\omega(n)}) \to 0, \qquad n \to \infty,$$

and the lemma is proven. □

It remains to establish a similar bound from below. Let us describe our plan of action. Denote $s = \log_b n - \log_b\log_b n - \omega(n)$ $[\omega(n) \to \infty,\ \omega(n) = o(\log n)]$, $n_1 = [n/2]$, $n_2 = n$ if $a = n/m < 1$, and $n_2 = [n - n^{1/2}]$ if $a = n/m = 1$. Define $\Delta(x) = 1$ $[\Delta(x) = 0]$ if $M(x) \geq s$ $[M(x) < s]$, and introduce $Y(n) = \sum_{x=n_1}^{n_2}\Delta(x)$.

We shall (a) estimate $E(Y(n))$ and $E(Y^2(n))$, (b) show—via Chebyshev's inequality—that $Y(n)$ is unbounded in probability and (c) deduce from this that $U_2(n) \geq s$ with probability approaching 1 as $n \to \infty$.

LEMMA 10.

$$(3.21)\qquad E(Y(n)) \sim e^a(ab)^{-1}b^{\omega(n)}, \qquad n \to \infty \ (\textit{and } m \to \infty).$$

PROOF. For m large enough, $m - x \geq n^{1/2}$, $s = o((m-x)^{1/5})$ if $n_1 \leq x \leq n_2$. Since $E(\Delta(x)) = P(M(x) \geq s)$, summing over $n_1 \leq x \leq n_2$ and using Lemma 7, we have [see (3.20)]

$$\begin{aligned} E(Y(n)) &= \Sigma_x P(M(x) \geq s) \sim \Sigma_x(1-e^{-x/m})^s \\ &\sim n\int_{n_1/n}^{n_2/n}[1-\exp(-at)]^s\,dt \sim n\int_{1/2}^{1}[1-\exp(-at)]^s\,dt \sim c(a)ns^{-1}b^s \\ &= e^a(ab)^{-1}b^{\omega(n)}. \end{aligned}$$

□

LEMMA 11.

$$(3.22)\qquad E(Y^2(n)) \sim E^2(Y(n)).$$

(See Appendix B for the proof.)

COROLLARY. $Y(n) \to \infty$ *in probability.*

PROOF. (For the sake of completeness.)

$$\begin{aligned} P(Y(n) \leq E(Y(n))/2) &\leq P(|Y(n) - E(Y(n))| \geq E(Y(n))/2) \\ &\leq 4\sigma^2(Y(n))/E^2(Y(n)) \\ &= 4\left[E(Y^2(n))/E^2(Y(n)) - 1\right] \to 0. \end{aligned}$$

□

Finally,

LEMMA 12. *In probability,* $U_2(n) \geq s$.

PROOF. Introduce

$$t(n) = \begin{cases} \min\{1 \leq x \leq n\colon M(x) \geq s\}, & \text{if } \mathcal{M}(x) \geq s, \\ n, & \text{if } \mathcal{M}(n) < s. \end{cases}$$

[Recall that $\mathcal{M}(n) = \max\{M(x)\colon 1 \leq x \leq n\}$.] Since $Y(n)$ is the number of keys x (between n_1 and n_2) such that $M(x) \geq s$, the previous corollary yields that $n - t(n) \to \infty$ in probability. Also, $t(n)$ is a stopping time adapted to the sequence $\{h(x)\colon 1 \leq x \leq n\}$; in other words, for each j, $\{t(n) = j\}$ belongs to the σ-field generated by $\{h(x)\colon 1 \leq x \leq j\}$. Hence, $\{t(n) = j\}$ is independent of $\{h(x)\colon j + 1 \leq x \leq n\}$.

Consequently,

$$\begin{aligned} P(U_2(n) < s) &= \sum_{j=1}^{n} P(t(n) = j, \forall\, x > j \text{ chooses an empty cell}) \\ &= \sum_{j=1}^{n} P(t(n) = j) \prod_{k=j}^{n-1} (1 - k/m). \end{aligned} \tag{3.23}$$

Fix an integer $d > 0$. Then, it follows from (3.23) that

$$\begin{aligned} P(U_2(n) < s) &\leq P(t(n) > n - d) + \sum_{j=1}^{n-d} P(t(n) = j) \prod_{k=j}^{n-1} (1 - k/m) \\ &\leq P(t(n) > n - d) + \prod_{k=n-d}^{n-1} (1 - k/m) \\ &\leq P(t(n) > n - d) + \exp[-d(n - d)/m]. \end{aligned}$$

Since $P(t(n) > n - d) \to 0$, as $n \to \infty$, we have

$$\limsup_n P(U_2(n) < s) \leq \exp(-da), \qquad a = n/m,$$

and, letting $d \uparrow \infty$,

$$\limsup_n P(U_2(n) < s) = 0,$$

because $a > 0$. The lemma is proven. □

This completes the proof of Theorem 1(b) and (1.4) of Theorem 2.

APPENDIX A

On the asymptotic formula for $P(t;\ w, l)$. Observe first that by the definition of $P(t;\ w, l)$, rather than (2.10),

$$P(t;\ w, l) = \Sigma_s (w!/m^w) \prod_{j=1}^{m} (s_j!)^{-1} \prod_{k=1}^{l} [1 - (1 - t)^{s_k}], \tag{A.1}$$

where $s_j \geq 0$, $1 \leq j \leq m$ and $\sum_{j=1}^{m} s_j = w$. Indeed, $(w!/m^w)\prod_{j=1}^{m}(s_j!)^{-1}$ is the probability that s_j balls choose the jth cell, $1 \leq j \leq m$, and $\prod_{t=1}^{l}[1-(1-t)^{s_k}]$ is the probability that each of the first l cells accepts at least one of the balls which have chosen it. (Of course, the latter probability equals 0 whenever $s_k = 0$ for some $1 \leq k \leq l$.)

Using (A.1), we can derive a surprisingly simple formula for the exponential generating function of the sequence $\{P(t; w, l)\}_{w \geq 0}$:

$$\begin{aligned}\sum_{w\geq 0} P(t; w, l)z^w/w! &= \sum_{s_1,\ldots,s_m \geq 0} \prod_{j=1}^{m} (z/m)^{s_j}/s_j! \prod_{k=1}^{l}\left[1-(1-t)^{s_k}\right] \\ &= \left\{\sum_{s\geq 0}\left[1-(1-t)^s\right](z/m)^s/s!\right\}^l \left[\sum_{s\geq 0}(z/m)^s/s!\right]^{m-l} \\ &= \{\exp(z/m) - \exp[z(1-t)/m]\}^l \exp[z(m-l)/m] \\ &= e^z[1-\exp(-zt/m)]^l.\end{aligned}$$

So, choosing a contour $C = \{a = re^{i\phi}\colon -\pi \leq \phi < \pi\}$, we have

$$\begin{aligned}P(t; w, l) &= w!(2\pi i)^{-1}\int_C e^z(1-e^{-zt/m})^l z^{-w-1}\,dz \\ &= w!(2\pi)^{-1}\int_{-\pi}^{\pi} \exp\bigl(F(re^{i\phi})\bigr)\,d\phi,\end{aligned} \tag{A.2}$$

where

$$F(z) = z + l\log(1-e^{-zt/m}) - w\log z.$$

To estimate the last integral, we use the saddle-point method. Select r which minimizes $F(u)$, $u \in (0, \infty)$, so that

$$F'(r) = 1 + (e^{rt/m}-1)^{-1}(lt/m) - w/r = 0. \tag{A.3}$$

[Such a point exists since $\lim_{u\to 0^+} F'(u) = \lim_{u\to 0^+}(l-w)/u = -\infty$ $(l < w)$ and $\lim_{u\to\infty} F'(u) = 1$.] By (A.3),

$$r = w - l(rt/m)/(e^{rt/m}-1) = w + O(l)$$

uniformly over $t \in (0, 1]$, since $\eta/(e^\eta - 1) < 1$ for all $\eta > 0$. (All the related estimates below are also uniform.) "Bootstrapping," we get a sharper estimate,

$$r = w - lgt/(e^{gt}-1) + O(l^2/m), \qquad g = w/m.$$

Since

$$F''(u) = (w/u^2)\bigl[1 - (l/w)\eta^2 e^\eta/(e^\eta - 1)^2\bigr], \qquad \eta = ut/m, \tag{A.4}$$

and $F'(r) = 0$, we have

$$\begin{aligned}F(r) &= F(w) - (F''(\tilde{u})/2)(w-r)^2 \qquad (r \leq \tilde{u} \leq w) \\ &= F(w) + O(l^2/w) = w - w\log w + l\log(1-e^{-gt}) + O(l^2/w).\end{aligned} \tag{A.5}$$

In addition, a simple computation shows that

$$\left|r^3 F^{(3)}(re^{i\phi})\right| = 2w(1 + O(l/w)) = O(w)$$

for all small enough $|\phi|$. Then, for $|\phi| \le \phi_0 = w^{-5/12}$,

$$F(re^{i\phi}) = F(r) + \big(r^2F''(r)/2\big)\big(e^{i\phi} - 1\big)^2 + O\big(w|e^{i\phi} - 1|^3\big) \quad [\text{see (A.4)}]$$
$$= F(r) - \big(r^2F''(r)/2\big)\phi^2 + O\big(w\phi_0^3\big)$$

and $O(w\phi_0^3) = O(w^{-1/4})$. So $[\lambda^2 = r^2F''(r) = w(1 + O(l/w))]$

$$\int_{-\phi_0}^{\phi_0} \exp\big(F(re^{i\phi})\big)\,d\phi \sim \exp\big(F(r)\big)\lambda^{-1/2}\int_{-\lambda\phi_0}^{\lambda\phi_0} \exp\big(-u^2/2\big)\,du \tag{A.6}$$
$$\sim (2\pi/w)^{1/2}\exp\big(F(r)\big)$$

because $\lambda\phi_0 \sim w^{1/12}$.

On the other hand, since $|e^z - 1| \le |e^{|z|} - 1|$, we have for $|\phi| \ge |\phi_0|$,

$$\big|\exp\big(F(re^{i\phi})\big)\big| \le \exp\big(F(r) + (1 - tl/m)r(\cos\phi - 1)\big)$$
$$\le \exp\big(F(r) - cw\phi^2\big) \le \exp\big(F(r) - cw^{1/6}\big), \qquad c > 0,$$

whence

$$\left|\int_{|\phi|>\phi_0} \exp\big(F(re^{i\phi})\big)\,d\phi\right| = O\big(\exp\big(F(r) - cw^{1/6}\big)\big). \tag{A.7}$$

Putting together (A.2), (A.5)–(A.7) and applying the Stirling formula to $w!$, we have

$$P(t; w, l) \sim w!(2\pi)^{-1}(2\pi/w)^{1/2}\exp\big(w - w\log w + l\log(1 - e^{-gt})\big)$$
$$\sim (1 - e^{-gt})^l.$$

APPENDIX B

Proof of Lemma 11. By the definition of $Y(n)$,

$$E\big(Y^2(n)\big) = \Sigma_{x_1, x_2} P\big(M(x_i) \ge s,\ i = 1,2\big)$$
$$= \Sigma_x P\big(M(x) \ge s\big) + \Sigma_{x_1 \ne x_2} P\big(M(x_i) \ge s,\ i = 1,2\big)$$
$$= \Sigma_1 + \Sigma_2, \qquad n_1 \le x, x_1, x_2 \le n_2.$$

Here $\Sigma_1 = E(Y(n))$, $\Sigma_2 = \Sigma_{21} + \Sigma_{22}$ and

$$\Sigma_{21} = \Sigma_{x_1 \ne x_2} P\big(M(x_i) \ge s,\ i = 1,2;\ K(x_1) \ne K(x_2)\big),$$
$$\Sigma_{22} = \Sigma_{x_1 \ne x_2} P\big(M(x_i) \ge s,\ i = 1,2;\ K(x_1) = K(x_2)\big).$$

Notice that, for $x_1 < x_2$, the condition $K(x_1) = K(x_2)$ means that no key among the keys $x_1 + 1, \ldots, x_2$ is rejected; so, in particular, $M(x_2) \ge M(x_1)$. Therefore,

$$P\big(M(x_i) \ge s,\ i = 1,2;\ K(x_1) = K(x_2)\big) = P\big(M(x_1) \ge s\big)\prod_{x_1 \le j < x_2}(1 - j/m)$$
$$\le P\big(M(x_1) \ge s\big)\beta^{x_2 - x_1},$$
$$\beta = 1 - [n/2]/m,$$

and

$$\Sigma_{22} \le \Sigma_x P(M(x) \ge s)\left(\Sigma_{j \ge x}\beta^{j-x}\right) = ([n/2]/m)^{-1}E(Y(n)) \tag{B.1}$$
$$= O(E(Y(n))).$$

To estimate Σ_{21}, first we write

$$\Sigma_{21} \le 2\Sigma_{x_1 < x_2} P\Big(M(x_i) \ge s, |K(x_i) - \bar{k}(x_i)| \le (m - x_i)^{3/5}, \quad i = 1,2;\ K(x_1) \ne K(x_2)\Big)$$
$$+ 2\Sigma_{x_1 < x_2} P\left(|K(x_1) - \bar{k}(x_1)| > (m - x_1)^{3/5}\right) \tag{B.2}$$
$$= \Sigma'_{21} + \Sigma''_{21}.$$

Here (see Lemma 6)

$$\Sigma''_{21} \le 2\Sigma_{x_1 < x_2}\exp\left[-(m - x_1)^{1/5}/30\right] = O\left[n^2\exp(-n^{1/10}/30)\right] = o(1) \tag{B.3}$$

and

$$\Sigma'_{21} = 2\Sigma_{x_1 < x_2}\Sigma_{k_1, k_2} P(M(x_i) \ge s,\ K(x_i) = k_i,\ i = 1,2), \tag{B.4}$$

where k_1, k_2 satisfy the conditions

$$|k_i - \bar{k}(x_i)| \le (m - x_i)^{3/5}, \qquad i = 1,2, \tag{B.5}$$
$$k_1 < k_1 + s \le x_1, \qquad k_1 + s < k_2 < k_2 + s \le x_2.$$

Now, according to an argument in the proof of Lemma 4, $M(x_i) \ge s$, $K(x_i) = k_i$, $i = 1,2$, iff (a) no key $\le x_i$ has chosen the cell k_i, $i = 1,2$, and (b) all the cells $k_i + 1, \ldots, l_i = k_i + s$ and exactly $x_i - l_i$ cells to the right from the cell l_i are chosen by the keys $\le x_i$, $i = 1,2$. To evaluate the probability $P(M(x_i) \ge s, K(x_i) = k_i, i = 1,2)$, observe that each allocation of the keys $\le x_2$ satisfying the conditions (a) and (b) is achieved as follows. Consider the sets of cells $A_1 = \{1, \ldots, k_1 - 1\}$, $|A_1| = k_1 - 1$, $A_2 = \{k_1 + 1, \ldots, l_1\}$, $|A_2| = l_1 - k_1$, $A_3 = \{l_1 + 1, \ldots, k_2 - 1\}$, $|A_3| = k_2 - 1 - l_1$, and $A_4 = \{k_2 + 1, \ldots, l_2\}$, $|A_4| = l_2 - k_2$. First, we choose a set A_5, $|A_5| = x_2 - l_2$, among the cells $\ge l_2 + 1$ (the number of ways is $\binom{m - l_2}{x_2 - l_2}$). Second, we choose the set A_6, $|A_6| = x_1 - l_1$, out of $A_3 \cup A_4 \cup A_5$. If $|A_6 \cap (A_4 \cup A_5)| = j$, then the number of ways is

$$\binom{|A_4 \cup A_5|}{j}\binom{|A_3|}{x_1 - l_1 - j} = \binom{x_2 - k_2}{j}\binom{k_2 - 1 - l_1}{x_1 - l_1 - j}.$$

Third, we allocate the keys $\le x_1$ on the set $A_7 = A_1 \cup A_2 \cup A_6$, $|A_7| = x_1 - 1$, so that the cells of $A_8 = A_2 \cup A_6$, $|A_8| = x_1 - k_1$, are nonempty. Finally, fourth, we allocate the keys $x_1 + 1, \ldots, x_2$ on the set $A_9 = \bigcup_{i=1}^5 A_i \cup \{k_1\}$, $|A_9| = x_2 - 1$, so that the cells of $A_{10} = (A_4 \cup A_5) \setminus A_6$, $|A_{10}| = x_2 - k_2 - j$, are nonempty.

Hence,

$$
\begin{aligned}
&P(M(x_i) \geq s,\ K(x_i) = k_i,\ i = 1,2) \\
&\quad = m^{-x_2}\binom{m - l_2}{x_2 - l_2}\Bigg[\sum_j \binom{x_2 - k_2}{j}\binom{k_2 - 1 - l_1}{x_1 - l_1 - j} \\
&\qquad\qquad \times N(x_2 - x_1, x_2 - 1, x_2 - k_2 - j)\Bigg] \\
&\qquad \times N(x_1, x_1 - 1, x_1 - k_1),
\end{aligned} \tag{B.6}
$$

where $N(u, v, w)$ is the *number* of ways to allocate u balls among v cells, so that some w fixed cells are nonempty.

The rest is simple. Since $|k_2 - \bar{k}(x_2)| \leq (m - x_2)^{3/5}$, we can use (3.15) to obtain from (B.6) that, for k_1, k_2 satisfying the conditions (B.4),

$$
\begin{aligned}
&P(M(x_i) \geq s,\ K(x_i) = k_i,\ i = 1,2) \\
&\quad \sim (1 - e^{-\lambda_2})^s m^{-x_2}\binom{m - k_2}{x_2 - k_2}\Bigg[\sum_j \binom{x_2 - k_2}{j}\binom{k_2 - 1 - l_1}{x_1 - l_1 - j} \\
&\qquad\qquad \times N(x_2 - x_1, x_2 - 1, x_2 - k_2 - j)\Bigg] \\
&\qquad \times N(x_1, x_1 - 1, x_1 - k_1) \\
&\quad = (1 - e^{-\lambda_2})^s P(M(x_1) \geq s;\ K(x_i) = k_i,\ i = 1,2), \\
&\qquad\qquad \lambda_2 = x_2/m.
\end{aligned} \tag{B.7}
$$

Besides, by Lemma 7,

$$P(M(x) \geq x) \sim (1 - e^{-\lambda})^s, \qquad \lambda = x/m,$$

uniformly over $n_1 \leq x \leq n_2$. Combining (B.4) and (B.7) we obtain then

$$
\begin{aligned}
\Sigma_{21} &\leq (1 + o(1))2\Sigma_{x_1 < x_2}(1 - e^{-\lambda_2})^s \Sigma_{k_1, k_2} P(M(x_1) \geq s; \\
&\qquad\qquad K(x_i) = k_i,\ i = 1,2) \\
&\leq (1 + o(1))2\Sigma\Sigma_{x_1 < x_2}(1 - e^{-\lambda_2})^s P(M(x_1) \geq s) \\
&\leq (1 + o(1))\left[\Sigma_x (1 - e^{-\lambda})^s\right]^2 \sim E^2(Y(n)).
\end{aligned} \tag{B.8}
$$

Since $E(Y(n)) \to \infty$, the relations (B.1)–(B.3) and (B.8) imply that

$$E(Y^2(n)) \leq [1 + o(1)]E^2(Y(n)).$$

It remains to observe that $E^2(Y(n)) \leq E(Y^2(n))$. □

Acknowledgment. I am grateful to the referee for many helpful suggestions regarding the presentation of the paper.

REFERENCES

[1] AJTAI, M., FREDMAN, M. and KOMLÓS, J. (1978). There is no fast single hashing algorithm. *Inform. Process. Lett.* **7** 270–273.

[2] BILLINGSLEY, P. (1979). *Probability and Measure*. Wiley, New York.

[3] BLAKE, I. F. and KONHEIM, A. G. (1977). Big buckets are (are not) better! *J. Assoc. Comput. Mach.* **24** 591–606.

[4] CHEN, W.-C. and VITTER, J. S. (1983). Analysis of early-insertion standard coalesced hashing. *SIAM J. Comput.* **12** 667–676.

[5] CHEN, W.-C. and VITTER, J. S. (1984). Analysis of new variants of coalesced hashing. *ACM Trans. Database Systems* **9** 616–645.

[6] DEVROYE, L. (1982). A note on the average depth of tries. *Computing* **28** 367–371.

[7] DEVROYE, L. (1984). A probabilistic analysis of the height of tries and of complexity of triesort. *Acta Inform.* **21** 229–237.

[8] DEVROYE, L. (1985). The expected length of the longest probe sequence for bucket searching when the distribution is not uniform. *J. Algorithms* **6** 1–19.

[9] FAGIN, R., NIEVERGELT, J., PIPPENGER, N. and STRONG, H. R. (1979). Extendible hashing—a fast access method for dynamic files. *ACM Trans. Database Systems* **4** 315–344.

[10] FLAJOLET, PH. and STEYAERT, J. M. (1982). A branching process arising in dynamic hashing, trie searching and polynomial factorization. *Proc. Ninth ICALP Colloquium. Lecture Notes in Computer Sci.* **140** 239–251. Springer, Berlin.

[11] GONNET, G. H. (1981). Expected length of the longest probe sequence in hash code searching. *J. Assoc. Comput. Mach.* **28** 289–304.

[12] GREENE, D. H. and KNUTH, D. E. (1982). *Mathematics for the Analysis of Algorithms*, 2nd ed. Birkhäuser, Boston.

[13] GUIBAS, L. J. (1978). The analysis of hashing techniques that exhibit k-ary clustering. *J. Assoc. Comput. Mach.* **25** 544–555.

[14] GUIBAS, L. J. and SZEMEREDI, E. (1978). The analysis of double hashing. *J. Comput. Sci.* **16** 226–274.

[15] KNOTT, G. D. (1984). Direct chaining with coalescing lists. *J. Algorithms* **5** 7–21.

[16] KNUTH, D. E. (1973). *The Art of Computer Programming* **3**. Addison-Wesley, Reading, Mass. (Section 6.4).

[17] KOLCHIN, V. F. (1982). *Random Mappings*. Nauka, Moscow. (In Russian.)

[18] KOLCHIN, V. F., SEVAST'YANOV, B. A. and CHISTYAKOV, V. P. (1978). *Random Allocations*. Wiley, New York.

[19] KONHEIM, A. G. and WEISS, B. (1966). An occupancy discipline and applications. *SIAM J. Appl. Math.* **14** 1266–1274.

[20] MENDELSON, H. (1982). Analysis of extendible hashing. *IEEE Trans. Software Engrg.* **8** 611–619.

[21] PITTEL, B. (1985). Asymptotical growth of a class of random trees. *Ann. Probab.* **13** 414–427.

[22] PITTEL, B. (1986). Paths in a random digital tree: limiting distributions. *Adv. in Appl. Probab.* **18** 139–155.

[23] PITTEL, B. (1987). Linear probing: The probable largest search time grows logarithmically with the number of records. *J. Algorithms* **8** 1–14.

[24] VITTER, J. S. (1981). A shared-memory scheme for coalesced hashing. *Inform. Process. Lett.* **13** 77–79.

[25] VITTER, J. S. (1983). Analysis of the search performance of coalesced hashing. *J. Assoc. Comput. Mach.* **30** 231–258.

[26] YAO, A. C. (1980). A note on the anlaysis of extendible hashing. *Inform. Process. Lett.* **11** 84–86.

DEPARTMENT OF MATHEMATICS
THE OHIO STATE UNIVERSITY
COLUMBUS, OHIO 43210