Design and Analysis of Dynamic Huffman Codes

JEFFREY SCOTT VITTER

Brown University, Providence, Rhode Island

Abstract. A new one-pass algorithm for constructing dynamic Huffman codes is introduced and analyzed. We also analyze the one-pass algorithm due to Faller, Gallager, and Knuth. In each algorithm, both the sender and the receiver maintain equivalent dynamically varying Huffman trees, and the coding is done in real time. We show that the number of bits used by the new algorithm to encode a message containing t letters is $< t$ bits more than that used by the conventional two-pass Huffman scheme, independent of the alphabet size. This is best possible in the worst case, for any one-pass Huffman method. Tight upper and lower bounds are derived. Empirical tests show that the encodings produced by the new algorithm are shorter than those of the other one-pass algorithm and, except for long messages, are shorter than those of the two-pass method. The new algorithm is well suited for on-line encoding/decoding in data networks and for file compression.

Categories and Subject Descriptors: C.2.0 **[Computer-Communication Networks]**: General—*data communications*; E.1 **[Data]**: Data Structures—*trees*; E.4 **[Data]**: Coding and Information Theory—*data compaction and compression; nonsecret encoding schemes*; F.2.2 **[Analysis of Algorithms and Problem Complexity]**: Nonnumerical Algorithms and Problems; G.2.2 **[Discrete Mathematics]**: Graph Theory—*trees*; H.1.1 **[Models and Principles]**: Systems and Information Theory—*value of information*

General Terms: Algorithms, Design, Performance, Theory

Additional Key Words and Phrases: Distributed computing, entropy, Huffman codes

1. *Introduction*

Variable-length source codes, such as those constructed by the well-known two-pass algorithm due to D. A. Huffman [5], are becoming increasingly important for several reasons. Communication costs in distributed systems are beginning to dominate the costs for internal computation and storage. Variable-length codes often use fewer bits per source letter than do fixed-length codes such as ASCII and EBCDIC, which require $\lceil \log n \rceil$ bits per letter, where n is the alphabet size. This can yield tremendous savings in packet-based communication systems. Moreover,

Support was provided in part by National Science Foundation research grant DCR-84-03613, by an NSF Presidential Young Investigator Award with matching funds from an IBM Faculty Development Award and an AT&T research grant, by an IBM research contract, and by a Guggenheim Fellowship.

An extended abstract of this research appears in Vitter, J. S. The design and analysis of dynamic Huffman coding. In *Proceedings of the 26th Annual IEEE Symposium on Foundations of Computer Science* (October). IEEE, New York, 1985. A Pascal implementation of the new one-pass algorithm appears in Vitter, J. S. Dynamic Huffman Coding. *Collected Algorithms of the ACM* (submitted 1986), and is available in computer-readable form through the ACM Algorithms Distribution Service.

Part of this research was also done while the author was at the Mathematical Sciences Research Institute in Berkeley, California; Institut National de Recherche en Informatique et en Automatique in Rocquencourt, France; and École Normale Supérieure in Paris, France.

Author's current address: Department of Computer Science, Brown University, Providence, RI 02912.

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
© 1987 ACM 0004-5411/87/1000-0825 $01.50

the buffering needed to support variable-length coding is becoming an inherent part of many systems.

The binary tree produced by Huffman's algorithm minimizes the weighted external path length $\sum_j w_j l_j$ among all binary trees, where w_j is the weight of the jth leaf, and l_j is its depth in the tree. Let us suppose there are k distinct letters $a_1, a_2, \ldots, a_k$ in a message to be encoded, and let us consider a Huffman tree with k leaves in which w_j, for $1 \leq j \leq k$, is the number of occurrences of a_j in the message. One way to encode the message is to assign a static code to each of the k distinct letters, and to replace each letter in the message by its corresponding code. Huffman's algorithm uses an optimum static code, in which each occurrence of a_j, for $1 \leq j \leq k$, is encoded by the l_j bits specifying the path in the Huffman tree from the root to the jth leaf, where "0" means "to the left" and "1" means "to the right."

One disadvantage of Huffman's method is that it makes two passes over the data: one pass to collect frequency counts of the letters in the message, followed by the construction of a Huffman tree and transmission of the tree to the receiver; and a second pass to encode and transmit the letters themselves, based on the static tree structure. This causes delay when used for network communication, and in file compression applications the extra disk accesses can slow down the algorithm. Faller [3] and Gallager [4] independently proposed a one-pass scheme, later improved substantially by Knuth [6], for constructing dynamic Huffman codes. The binary tree that the sender uses to encode the $(t + 1)$st letter in the message (and that the receiver uses to reconstruct the $(t + 1)$st letter) is a Huffman tree for the first t letters of the message. Both sender and receiver start with the same initial tree and thereafter stay synchronized; they use the same algorithm to modify the tree after each letter is processed. Thus there is never need for the sender to transmit the tree to the receiver, unlike the case of the two-pass method. The processing time required to encode and decode a letter is proportional to the length of the letter's encoding, so the processing can be done in real time.

Of course, one-pass methods are not very interesting if the number of bits transmitted is significantly greater than with Huffman's two-pass method. This paper gives the first analytical study of the efficiency of dynamic Huffman codes. We derive a precise and clean characterization of the difference in length between the encoded message produced by a dynamic Huffman code and the encoding of the same message produced by a static Huffman code. The length (in bits) of the encoding produced by the algorithm of Faller, Gallager, and Knuth (Algorithm FGK) is shown to be at most $\approx 2S + t$, where S is the length of the encoding by a static Huffman code, and t is the number of letters in the original message. More important, the insights we gain from the analysis lead us to develop a new one-pass scheme, which we call Algorithm Λ, that produces encodings of $< S + t$ bits. That is, compared with the two-pass method, Algorithm Λ uses less than one extra bit per letter. We prove this is optimum in the worst case among all one-pass Huffman schemes.

It is impossible to show that a given dynamic code is optimum among all dynamic codes, because one can easily imagine non-Huffman-like codes that are optimized for specific messages. Thus there can be no global optimum. For that reason we restrict our model of one-pass schemes to the important class of one-pass Huffman schemes, in which the next letter of the message is encoded on the basis of a Huffman tree for the previous letters. We also do not consider the worst-case encoding length, among all possible messages of the same length, because for any one-pass scheme and any alphabet size n we can construct a message that is

encoded with an average of $\geq \lfloor \log_2 n \rfloor$ bits per letter. The harder and more important measure, which we address in this paper, is the worst-case *difference in length* between the dynamic and static encodings of the same message.

One intuition why the dynamic code produced by Algorithm Λ is optimum in our model is that the tree it uses to process the $(t+1)$st letter is not only a Huffman tree with respect to the first t letters (that is, $\sum_j w_j l_j$ is minimized), but it also minimizes the external path length $\sum_j l_j$ and the height $\max_j\{l_j\}$ among all Huffman trees. This helps guard against a lengthy encoding for the $(t+1)$st letter. Our implementation is based on an efficient data structure we call a *floating tree.* Algorithm Λ is well suited for practical use and has several applications. Algorithm FGK is already used for file compression in the *compact* command available under the 4.2BSD UNIX[1] operating system [7]. Most Huffman-like algorithms use roughly the same number of bits to encode a message when the message is long; the main distinguishing feature is the coding efficiency for short messages, where overhead is more apparent. Empirical tests show that Algorithm Λ uses fewer bits for short messages than do Huffman's algorithm and Algorithm FGK. Algorithm Λ can thus be used as a general-purpose coding scheme for network communication and as an efficient subroutine in word-based compaction algorithms.

In the next section we review the basic concepts of Huffman's two-pass algorithm and the one-pass Algorithm FGK. In Section 3 we develop the main techniques for our analysis and apply them to Algorithm FGK. In Section 4 we introduce Algorithm Λ and prove that it runs in real time and gives optimal encodings, in terms of our model defined above. In Section 5 we describe several experiments comparing dynamic and static codes. Our conclusions are listed in Section 6.

2. *Huffman's Algorithm and Algorithm FGK*

In this section we discuss Huffman's original algorithm and the one-pass Algorithm FGK. First let us define the notation we use throughout the paper.

Definition 2.1. We define

n = alphabet size;
a_j = jth letter in the alphabet;
t = number of letters in the message processed so far;
$\mathcal{M}_t = a_{i_1}, a_{i_2}, \ldots, a_{i_t}$, the first t letters of the message;
k = number of distinct letters processed so far;
w_j = number of occurrences of a_j processed so far;
l_j = distance from the root of the Huffman tree to a_j's leaf.

The constraints are $1 \leq j, k \leq n$, and $0 \leq w_j \leq t$.

In many applications, the final value of t is much greater than n. For example, a book written in English on a conventional typewriter might correspond to $t \approx 10^6$ and $n = 87$. The ASCII alphabet size is $n = 128$.

Huffman's two-pass algorithm operates by first computing the letter frequencies w_j in the entire message. A leaf node is created for each letter a_j that occurs in the message; the weight of a_j's leaf is its frequency w_j. The meat of the algorithm is the

[1] UNIX is a registered trademark of AT&T Bell Laboratories.

following procedure for processing the leaves and constructing a binary tree of minimum weighted external path length $\sum_j w_j l_j$:

```
Store the k leaves in a list L;
while L contains at least two nodes do
  begin
  Remove from L two nodes x and y of smallest weight;
  Create a new node p, and make p the parent of x and y;
  p's weight := x's weight + y's weight;
  Insert p into L
  end;
```

The node remaining in L at the end of the algorithm is the root of the desired binary tree. We call a tree that can be constructed in this way a "Huffman tree." It is easy to show by contradiction that its weighted external path length is minimum among all possible binary trees for the given leaves. In each iteration of the **while** loop, there may be a choice of which two nodes of minimum weight to remove from L. Different choices may produce structurally different Huffman trees, but all possible Huffman trees will have the same weighted external path length.

In the second pass of Huffman's algorithm, the message is encoded using the Huffman tree constructed in pass 1. The first thing the sender transmits to the receiver is the shape of the Huffman tree and the correspondence between the leaves and the letters of the alphabet. This is followed by the encodings of the individual letters in the message. Each occurrence of a_j is encoded by the sequence of 0's and 1's that specifies the path from the root of the tree to a_j's leaf, using the convention that "0" means "to the left" and "1" means "to the right."

To retrieve the original message, the receiver first reconstructs the Huffman tree on the basis of the shape and leaf information. Then the receiver navigates through the tree by starting at the root and following the path specified by the 0 and 1 bits until a leaf is reached. The letter corresponding to that leaf is output, and the navigation begins again at the root.

Codes like this, which correspond in a natural way to a binary tree, are called *prefix codes*, since the code for one letter cannot be a proper prefix of the code for another letter. The number of bits transmitted is equal to the weighted external path length $\sum_j w_j l_j$ plus the number of bits needed to encode the shape of the tree and the labeling of the leaves. Huffman's algorithm produces a prefix code of minimum length, since $\sum_j w_j l_j$ is minimized.

The two main disadvantages of Huffman's algorithm are its two-pass nature and the overhead required to transmit the shape of the tree. In this paper we explore alternative one-pass methods, in which letters are encoded "on the fly." We do not use a static code based on a single binary tree, since we are not allowed an initial pass to determine the letter frequencies necessary for computing an optimal tree. Instead the coding is based on a dynamically varying Huffman tree. That is, the tree used to process the $(t + 1)$st letter is a Huffman tree with respect to $\mathscr{M}_t$. The sender encodes the $(t + 1)$st letter a_{i_t} in the message by the sequence of 0's and 1's that specifies the path from the root to a_{i_t}'s leaf. The receiver then recovers the original letter by the corresponding traversal of its copy of the tree. Both sender and receiver then modify their copies of the tree before the next letter is processed so that it becomes a Huffman tree for $\mathscr{M}_{t+1}$. A key point is that neither the tree nor its modification needs to be transmitted, because the sender and receiver use the same modification algorithm and thus always have equivalent copies of the tree.

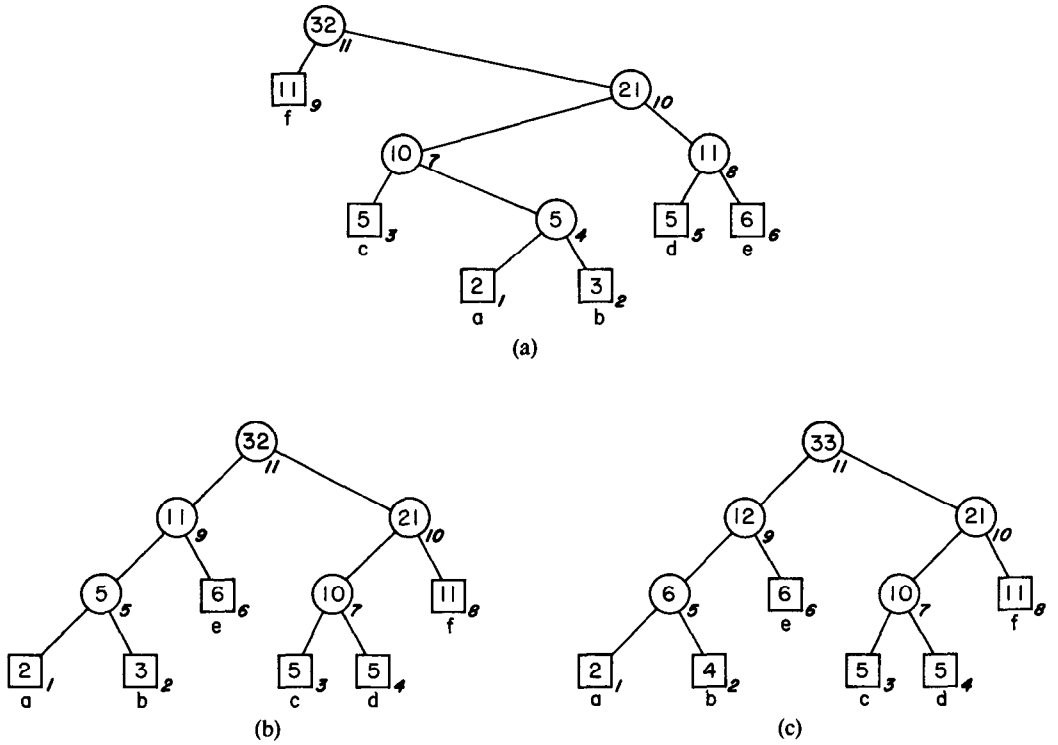


FIG. 1. This example from [6] for $t = 32$ illustrates the basic ideas of the Algorithm FGK. The node numbering for the sibling property is displayed next to each node. The next letter to be processed in the message is "$a_{i_{t+1}}$ = b". (a) The current status of the dynamic Huffman tree, which is a Huffman tree for $\mathcal{M}_t$, the first t letters in the message. The encoding for "b" is "1011", given by the path from the root to the leaf for "b". (b) The tree resulting from the interchange process. It is a Huffman tree for $\mathcal{M}_t$ and has the property that the weights of the traversed nodes can be incremented by 1 without violating the sibling property. (c) The final tree, which is the tree in (b) with the incrementing done, is a Huffman tree for $\mathcal{M}_{t+1}$.

Another key concept behind dynamic Huffman codes is the following elegant so-called characterization of Huffman trees:

Sibling Property: A binary tree with p leaves of nonnegative weight is a Huffman tree if and only if

(1) the p leaves have nonnegative weights $w_1, \ldots, w_p$, and the weight of each internal node is the sum of the weights of its children; and

(2) the nodes can be numbered in nondecreasing order by weight, so that nodes $2j - 1$ and $2j$ are siblings, for $1 \leq j \leq p - 1$, and their common parent node is higher in the numbering.

The node numbering corresponds to the order in which the nodes are combined by Huffman's algorithm: Nodes 1 and 2 are combined first, nodes 3 and 4 are combined second, nodes 5 and 6 are combined next, and so on.

Suppose that $\mathcal{M}_t = a_{i_1}, a_{i_2}, \ldots, a_{i_t}$ has already been processed. The next letter $a_{i_{t+1}}$ is encoded and decoded using a Huffman tree for $\mathcal{M}_t$. The main difficulty is how to modify this tree quickly in order to get a Huffman tree for $\mathcal{M}_{t+1}$. Let us consider the example in Figure 1, for the case $t = 32$, $a_{i_{t+1}}$ = "b". It is not good enough to simply increment by 1 the weights of $a_{i_{t+1}}$'s leaf and its ancestors, because

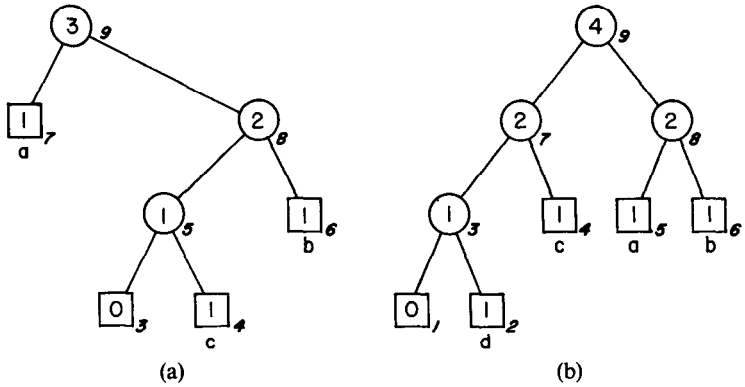


FIG. 2. Algorithm FGK operating on the message "abcd ...". (a) The Huffman tree immediately before the fourth letter "d" is processed. The encoding for "d" is specified by the path to the 0-node, namely, "100". (b) After *Update* is called.

the resulting tree will not be a Huffman tree, as it will violate the sibling property. The nodes will no longer be numbered in nondecreasing order by weight; node 4 will have weight 6, but node 5 will still have weight 5. Such a tree could therefore not be constructed via Huffman's two-pass algorithm.

The solution can most easily be described as a two-phase process (although for implementation purposes both phases can be combined easily into one). In the first phase, we transform the tree into another Huffman tree for $\mathscr{M}_t$, to which the simple incrementing process described above can be applied successfully in phase 2 to get a Huffman tree for $\mathscr{M}_{t+1}$. The first phase begins with the leaf of $a_{i_{t+1}}$ as the current node. We repeatedly interchange the contents of the current node, including the subtree rooted there, with that of the highest numbered node of the same weight, and make the parent of the latter node the new current node. The current node in Figure 1a is initially node 2. No interchange is possible, so its parent (node 4) becomes the new current node. The contents of nodes 4 and 5 are then interchanged, and node 8 becomes the new current node. Finally, the contents of nodes 8 and 9 are interchanged, and node 11 becomes the new current node. The first phase halts when the root is reached. The resulting tree is pictured in Figure 1b. It is easy to verify that it is a Huffman tree for $\mathscr{M}_t$ (i.e., it satisfies the sibling property), since each interchange operates on nodes of the same weight. In the second phase, we turn this tree into the desired Huffman tree for $\mathscr{M}_{t+1}$ by incrementing the weights of $a_{i_{t+1}}$'s leaf and its ancestors by 1. Figure 1c depicts the final tree, in which the incrementing is done.

The reason why the final tree is a Huffman tree for $\mathscr{M}_{t+1}$ can be explained in terms of the sibling property: The numbering of the nodes is the same after the incrementing as before. Condition 1 and the second part of condition 2 of the sibling property are trivially preserved by the incrementing. We can thus restrict our attention to the nodes that are incremented. Before each such node is incremented, it is the largest numbered node of its weight. Hence, its weight can be increased by 1 without becoming larger than that of the next node in the numbering, thus preserving the sibling property.

When $k < n$, we use a single 0-node to represent the $n - k$ unused letters in the alphabet. When the $(t + 1)$st letter in the message is processed, if it does not appear in $\mathscr{M}_t$, the 0-node is split to create a leaf node for it, as illustrated in Figure 2. The

$(t + 1)$st letter is encoded by the path in the tree from the root to the 0-node, followed by some extra bits that specify which of the $n - k$ unused letters it is, using a simple prefix code.

Phases 1 and 2 can be combined in a single traversal from the leaf of $a_{i_{t+1}}$ to the root, as shown below. Each iteration of the **while** loop runs in constant time, with the appropriate data structure, so that the processing time is proportional to the encoding length. A full implementation appears in [6].

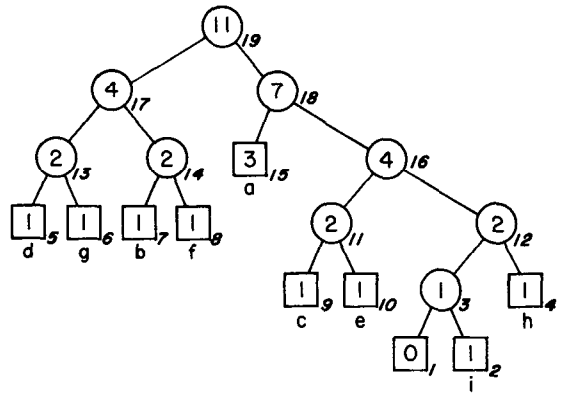
```
procedure Update;
begin
q := leaf node corresponding to a_{i_{t+1}};
if (q is the 0-node) and (k < n - 1) then
   begin
   Replace q by a parent 0-node with two leaf 0-node children, numbered in the order left
      child, right child, parent;
   q := right child just created
   end;
if q is the sibling of a 0-node then
   begin
   Interchange q with the highest numbered leaf of the same weight;
   Increment q's weight by 1;
   q := parent of q
   end;
while q is not the root of the Huffman tree do
   begin  {Main loop}
   Interchange q with the highest numbered node of the same weight;
   {q is now the highest numbered node of its weight}
   Increment q's weight by 1;
   q := parent of q
   end
end;
```

We denote an interchange in which q moves up one level by $\uparrow$ and an interchange between q and another node on the same level by $\rightarrow$. For example, in Figure 1, the interchange of nodes 8 and 9 is of type $\uparrow$, whereas that of nodes 4 and 5 is of type $\rightarrow$. Oddly enough, it is also possible for q to move down a level during an interchange, as illustrated in Figure 3; we denote such an interchange by $\downarrow$.

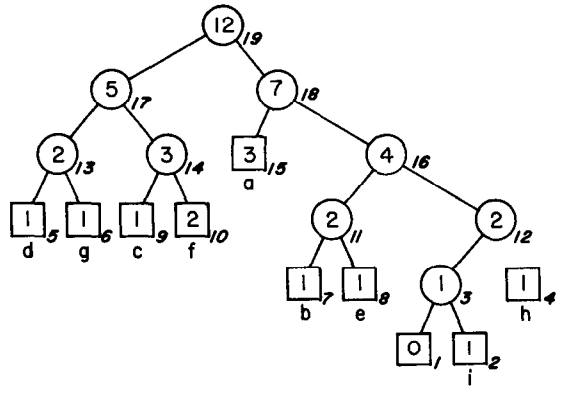
No two nodes with the same weight can be more than one level apart in the tree, except if one is the sibling of the 0-node. This follows by contradiction, since otherwise it will be possible to interchange nodes and get a binary tree having smaller external weighted path length. Figure 4 shows the result of what would happen if the letter "c" (rather than "d") were the next letter processed using the tree in Figure 2a. The first interchange involves nodes two levels apart; the node moving up is the sibling of the 0-node. We shall designate this type of two-level interchange by $\uparrow\uparrow$. There can be at most one $\uparrow\uparrow$ for each call to *Update*.

3. *Analysis of Algorithm FGK*

For purposes of comparing the coding efficiency of one-pass Huffman algorithms with that of the two-pass method, we shall count only the bits corresponding to the paths traversed in the trees during the coding. For the one-pass algorithms, we shall not count the bits used to distinguish which new letter is encoded when a letter is encountered in the message for the first time. And, for the two-pass method, we shall not count the bits required to encode the shape of the tree and the labeling of the leaves. The noncounted quantity for the one-pass algorithms is typically between $k(\log_2 n - 1)$ and $k \log_2 n$ bits using a simple prefix code, and the uncounted



(a)



(b)

FIG. 3. (a) The Huffman tree formed by Algorithm FGK after processing "abcdefghiaa". (b) The Huffman tree that will result if the next processed letter is "f". Note that there is an interchange of type ↓ (between leaf nodes 8 and 10) followed immediately by an interchange of type ↑ (between internal nodes 11 and 14).

quantity for the two-pass method is roughly $2k$ bits more than for the one-pass method. This means that our evaluation of one-pass algorithms will be *conservative* with respect to the two-pass method. When the message is long (that is, $t \gg n$), these uncounted quantities are insignificant compared with the total number of bits transmitted. (For completeness, the empirical results in Section 5 include statistics that take into account these extra quantities.)

Definition 3.1. Suppose that a message $\mathscr{M}_t = a_{i_1}, a_{i_2}, \ldots, a_{i_t}$ of size $t \geq 0$ has been processed so far. We define S_t to be the communication cost for a static Huffman encoding of $\mathscr{M}_t$ *using a Huffman tree based only on* $\mathscr{M}_t$; that is,

$$S_t = \sum_j w_j l_j,$$

where the sum is taken over any Huffman tree for $\mathscr{M}_t$. We also define s_t to be the "incremental" cost

$$s_t = S_t - S_{t-1}.$$

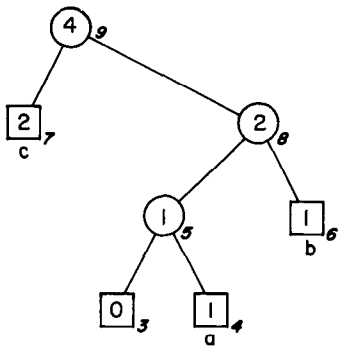


FIG. 4. The Huffman tree that would result from Algorithm FGK if the fourth letter in the example in Figure 2 were "c" rather than "d". An interchange of type ↑↑ occurs when *Update* is called.

We denote by d_t the communication cost for encoding a_{i_t} using a dynamic Huffman code; that is,

$$d_t = l_{i_t}$$

for the dynamic Huffman tree with respect to $\mathcal{M}_{t-1}$. We define D_t to be the total communication cost for all t letters; that is,

$$D_t = D_{t-1} + d_t, \qquad D_0 = 0.$$

Note that s_t does not have an intuitive meaning in terms of the length of the encoding for a_{i_t}, as does d_t. The following theorem bounds D_t by $\approx 2S_t + t$.

THEOREM 3.1. *For each $t \geq 0$, the communication cost of Algorithm FGK can be bounded by*

$$S_t - k + \delta_{k=n} + \delta_{k<n} \min_{w_j>0}\{w_j\} \leq D_t \leq 2S_t + t - 4k + 2\delta_{k=n} + 2\delta_{k<n} \min_{w_j>0}\{w_j\} - m,$$

where δ_R denotes 1 if relation R is true and 0 otherwise, and m is the cardinality of the set $\{j \mid 1 \leq j \leq t;\ a_{i_j} \in \mathcal{M}_{j-1};$ and $\forall x \in \mathcal{M}_{j-1}$ such that $x \neq a_{i_j}$, x appears strictly more often in $\mathcal{M}_{j-1}$ than does $a_{i_j}\}$.

The term m is the number of times during the course of the algorithm that the processed leaf is not the 0-node and has strictly minimum weight among all other leaves of positive weight. An immediate lower bound on m is $m \geq \min_{w_j>0}\{w_j\} - 1$. (For each value $2 \leq w \leq \min_{w_j>0}\{w_j\}$, consider the last leaf to attain weight w.) The minor δ terms arise because our one-pass algorithms use a 0-node when $k < n$, as opposed to the conventional two-pass method; this causes the leaf of minimum positive weight to be one level lower in the tree. The δ terms can be effectively ignored when there is a specially designated "end-of-file" character to denote the end of transmission, because when the algorithm terminates we have $\min_{w_j>0}\{w_j\} = 1$.

The Fibonacci-like tree in Figure 5 is an example of when the first bound is tight. The difference $D_t - S_t$ decreases by 1 each time a letter not previously in the message is processed, except for when k increases from $n - 1$ to n. The following two examples, in which the communication cost per letter D_t/t is bounded by a small constant, yield $D_t/S_t \to c > 1$. The message in the first example consists of any finite number of letters not including "a" and "b", followed by "abbaabbaa ...". In the limit, we have $S_t/t \to \frac{3}{2}$ and $D_t/t \to 2$, which yields $D_t/S_t \to \frac{4}{3} > 1$. The second example is a simple modification for the case of

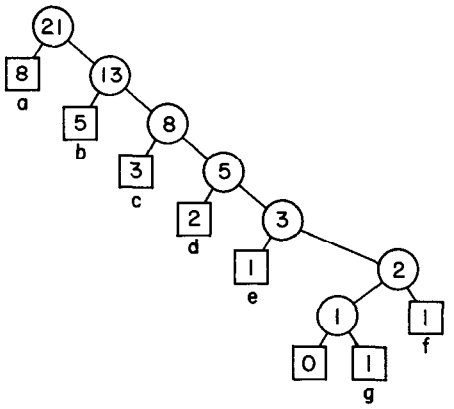


FIG. 5. Illustration of both the lower bound of Theorem 3.1 and the upper bounds of Lemma 3.2. The sequence of letters in the message so far is "abacabdabaceabacabdf" followed by "g" and can be constructed via a simple Fibonacci-like recurrence. For the lower bound, let $t = 21$. The tree can be constructed without any exchanges of types $\uparrow$, $\uparrow\uparrow$, or $\downarrow$; it meets the first bound given in Theorem 3.1. For the upper bound, let $t = 22$. The tree depicts the Huffman tree immediately before the tth letter is processed. If the tth letter is "h", we will have $d_t = 7$ and $h_t = \lceil d_t/2 \rceil - 1 = 3$. If instead the tth letter is "g", we will have $d_t = 7$ and $h_t = \lceil d_t/2 \rceil = 4$. If the tth letter is "f", we will have $d_t = 6$ and $h_t = \lfloor d_t/2 \rfloor = 3$.

alphabet size $n = 3$. The message consists of the same pattern as above, without the optional prefix, yielding $D_t/S_t \to 2$. So far all known examples where $\lim\sup_{t\to\infty} D_t/S_t \neq 1$ satisfy the constraint $D_t = O(t)$. We conjecture that the constraint is necessary:

Conjecture. For each $t \geq 0$, the communication cost of Algorithm FGK satisfies

$$D_t = S_t + O(t).$$

Before we can prove Theorem 3.1, we must develop the following useful notion. We shall denote by h_t the net change of height in the tree of the leaf for a_{i_t} as a result of interchanges during the tth call to *Update*.

Definition 3.2. For each $t \geq 1$, we define h_t by

$$h_t = (\#\ \uparrow\text{'s}) + 2(\#\ \uparrow\uparrow\text{'s}) - (\#\ \downarrow\text{'s}),$$

where we consider the interchanges that occur during the processing of the tth letter in the message.

The proof of Theorem 3.1 is based on the following important correspondence between h_t and $d_t - s_t$:

THEOREM 3.2. *For $t \geq 1$, we have*

$$d_t - s_t = h_t - \delta_{\Delta k=1} + (\delta_{k<n \text{ or } \Delta k=1})\Delta \min_{w_j>0}\{w_j\},$$

where $\Delta f = (f$ after a_{i_t} is processed$) - (f$ before a_{i_t} is processed$)$.

PROOF. The δ terms are due to the presence of the 0-node when $k < n$. Let us consider the case in which there is no 0-node, as in Figure 1. We define $\mathcal{T}_a$ to be the Huffman tree with respect to $\mathcal{M}_{t-1}$, $\mathcal{T}_b$ to be the Huffman tree formed by the interchanges applied to $\mathcal{T}_a$, and $\mathcal{T}_c$ to be the Huffman tree formed from $\mathcal{T}_b$ by incrementing a_{i_t}'s leaf and its ancestors. In the example in Figure 1, we redefine $t = 33$ and $a_{i_t} =$ "b". The trees $\mathcal{T}_a$, $\mathcal{T}_b$, and $\mathcal{T}_c$ correspond to those in Figure 1a, 1b, and 1c, respectively.

Trees $\mathcal{T}_a$ and $\mathcal{T}_b$ represent Huffman trees with respect to $\mathcal{M}_{t-1}$, and $\mathcal{T}_c$ is a Huffman tree with respect to $\mathcal{M}_t$. The communication cost d_t for processing the tth letter a_{i_t} is the depth in $\mathcal{T}_a$ of its leaf node; that is,

$$d_t = l_{i_t}(\mathcal{T}_a). \tag{1}$$

Each interchange of type ↑ moves the leaf for a_{i_t} one level higher in the tree, each interchange of type ↑↑ moves it two levels higher, and each interchange of type ↓ moves it one level lower. We have

$$h_t = l_{i_t}(\mathcal{T}_a) - l_{i_t}(\mathcal{T}_b). \tag{2}$$

The communication costs S_{t-1} and S_t are equal to the weighted external path lengths of $\mathcal{T}_a$ and $\mathcal{T}_c$, respectively. The interchanges that convert $\mathcal{T}_a$ to $\mathcal{T}_b$ maintain the sibling property, so $\mathcal{T}_a$ and $\mathcal{T}_b$ have the same weighted external path length. However, $\mathcal{T}_b$ is special since it can be turned into a Huffman tree for $\mathcal{M}_t$ (namely, tree $\mathcal{T}_c$) simply by incrementing a_{i_t}'s leaf and its ancestors by 1. Thus, we have

$$s_t = S_t - S_{t-1} = l_{i_t}(\mathcal{T}_b). \tag{3}$$

Putting (1), (2), and (3) together yields the result $d_t - s_t = h_t$.

When there is a 0-node present in $\mathcal{T}_a$, the communication cost S_{t-1} is $\min_{w_j(\mathcal{T}_a)>0}\{w_j(\mathcal{T}_a)\}$ less than the weighted external path length for $\mathcal{T}_a$, since the presence of the 0-node in $\mathcal{T}_a$ moves a leaf of minimum positive weight one level farther from the root than it would be if there were no 0-node. Similarly, when there is a 0-node in $\mathcal{T}_c$, the communication cost S_t is $\min_{w_j(\mathcal{T}_c)>0}\{w_j(\mathcal{T}_c)\}$ less than the weighted external path length for $\mathcal{T}_c$. If $\mathcal{T}_a$ has a 0-node and a_{i_t} appears in $\mathcal{M}_{t-1}$, then the 0-node will also appear in $\mathcal{T}_c$; this contributes a $\delta_{k<n}\Delta\ \min_{w_j>0}\{w_j\}$ term to $d_t - s_t$. If $\mathcal{T}_a$ has a 0-node and at most $n-2$ leaves of positive weight, and if a_{i_t} does not appear in $\mathcal{M}_{t-1}$, then the 0-node will be split, as outlined in *Update*; this has the effect of moving the leaf of a_{i_t} one level down, thus contributing $-(\delta_{k<n \text{ and } \Delta k=1})$ to $d_t - s_t$. The final special case is when the 0-node appears in $\mathcal{T}_a$ but not in $\mathcal{T}_c$; in this case, a_{i_t} does not appear in $\mathcal{M}_{t-1}$, but the other $n-1$ letters in the alphabet do. This contributes a $(\delta_{k=n \text{ and } \Delta k=1})(\Delta\ \min_{w_j>0}\{w_j\} - 1)$ term to $d_t - s_t$. Putting these three δ terms together, with some algebraic manipulation, gives us the final result. This proves Theorem 3.2. □

Three more lemmas are needed for the proof of Theorem 3.1:

LEMMA 3.1. *During a call to Update in Algorithm FGK, each interchange of type ↓ is followed at some point by an ↑, with no ↓'s in between.*

PROOF. By contradiction. Suppose that during a call to *Update* there are two interchanges of type ↓ with no ↑ in between. In the initial version of the tree before *Update* is called, let a_1 and b_1 be the nodes involved in the first ↓ interchange mentioned above, and let a_2 and b_2 be the nodes in the subsequent ↓ interchange; nodes a_1 and a_2 are one level higher in the tree than b_1 and b_2, respectively, before *Update* is called. Let $k \geq 1$ be the number of levels in the tree separating b_1 and a_2, and let a_1^k be the ancestor of a_1 that is k levels above a_1. Node a_1^k is one level higher than a_2, and we can show that their weights are equal by the following argument: If the weight of a_1^k is $<$ a_2's weight, then we can interchange the two nodes and decrease the weighted external path length. On the other hand, if the weight of a_1^k is $>$ a_2's weight, then it can be shown from the sibling property that at some earlier point there should have been an interchange of type 1 in which one of a_1's ancestors moved down one level. Both cases cause a contradiction. The ↓ interchange between a_2 and b_2 means that b_2 has the same weight as a_2 and is one level lower, which makes b_2 two levels lower than a_1^k but with the same weight. This is impossible, as mentioned at the end of Section 2. □

LEMMA 3.2. *For each $t \geq 1$, we have*

$$0 \leq h_t \leq \begin{cases} \lceil d_t/2 \rceil - 1, & \text{if } a_{i_t}\text{'s node is the 0-node;} \\ \lceil d_t/2 \rceil, & \text{if } a_{i_t}\text{'s node is the 0-node's sibling;} \\ \lfloor d_t/2 \rfloor, & \text{otherwise.} \end{cases}$$

An example achieving each of the three bounds is the Fibonacci-like tree given in Figure 5.

PROOF. Let us consider what can happen when *Update* is called to process the tth letter a_{i_t}. Suppose for the moment that only interchanges of types $\uparrow$ or $\rightarrow$ occur. Each $\uparrow$ interchange, followed by the statement "$q :=$ parent of q", moves q two levels up in the tree. A $\rightarrow$ interchange or no interchange at all, followed by "$q :=$ parent of q", moves q up one level. Interchanges of type $\uparrow$ are not possible when q is a child of the root. Putting this all together, we find that the number of $\uparrow$ interchanges is at most $\lfloor d_t/2 \rfloor$, where d_t is the initial depth in the tree of the leaf for a_{i_t}.

If there are no interchanges of type $\uparrow\uparrow$ or $\downarrow$, the above argument yields $0 \leq h_t \leq \lfloor d_t/2 \rfloor$. If an interchange of type $\downarrow$ occurs, then by Lemma 3.1 there is a subsequent $\uparrow$, so the result still holds. An interchange of type $\uparrow\uparrow$ can occur if the leaf for a_t is the sibling of the 0-node; since at most one $\uparrow\uparrow$ can occur, we have $0 \leq h_t \leq \lceil d_t/2 \rceil$. The final case to consider occurs when the leaf for a_t is the 0-node; no interchange can occur during the first trip through the **while** loop in *Update*, so we have $0 \leq h_t \leq \lceil d_t/2 \rceil - 1$. □

LEMMA 3.3. *Suppose that a_{i_t} occurs in $\mathcal{M}_{t-1}$, but strictly less often than all the other letters that appear in $\mathcal{M}_{t-1}$. Then when the tth letter in the message is processed by Update, the leaf for a_{i_t} is not involved in an interchange.*

PROOF. By the hypothesis, all the leaves other than the 0-node have a strictly larger weight than a_{i_t}'s leaf. The only node that can have the same weight is its parent. This happens when a_{i_t}'s leaf is the sibling of the 0-node, but there is no interchange in this special case. □

PROOF OF THEOREM 3.1. By Lemma 3.2, we have $0 \leq h_t \leq d_t/2 + \frac{1}{2} - \delta_{\Delta k=1}$. Lemma 3.3 says that there are m values of t for which this bound can be lessened by 1. We get the final result by substituting this into the formula in Theorem 3.2 and by summing on t. This completes the proof. □

There are other interesting identities as well, besides the ones given above. For example, a proof similar to the one for Lemma 3.1 gives the following result:

LEMMA 3.4. *In the execution of Update, if an interchange of type $\uparrow$ or $\uparrow\uparrow$ moves node v upward in the tree, interchanging it with node x, there cannot subsequently be more $\uparrow$'s than $\downarrow$'s until q reaches the lowest common ancestor of v and x.*

A slightly weaker bound of the form $D_t = 2S_t + O(t)$ can be proved using the following entropy argument suggested by B. Chazelle (personal communication). The depth of a_{i_j}'s leaf in the dynamic Huffman tree during any of the w_j times a_{i_j} is processed can be bounded as a function of the leaf's relative weight at the time, which in turn can be bounded in terms of a_{i_j}'s final relative weight w_{i_j}/t. For example, during the last $\lfloor w_{i_j}/2 \rfloor$ times a_{i_j} is processed, its relative weight is $\geq w_{i_j}/(2t)$. The factor of 2 in front of the S_t term emerges because the relative weight of a leaf node in a Huffman tree can only specify the depth of the node to within a factor of 2 asymptotically (cf. Lemma 3.2). The characterization we give in

Theorem 3.2 is robust in that it allows us to study precisely how $D_t - S_t$ changes as more letters are processed; this will be crucial for obtaining our main result in the next section that Algorithm Λ uses less than one extra bit per letter compared with the two-pass method.

4. *Optimum Dynamic Huffman Codes*

In this section we describe Algorithm Λ and show that it runs in real time and is optimum in our model of one-pass Huffman algorithms. There were two motivating factors in its design:

(1) The number of ↑'s should be bounded by some small number (in our case, 1) during each call to *Update*.
(2) The dynamic Huffman tree should be constructed to minimize not only $\sum_j w_j l_j$, but also $\sum_j l_j$ and $\max_j\{l_j\}$, which intuitively has the effect of preventing a lengthy encoding of the next letter in the message.

4.1 IMPLICIT NUMBERING. One of the key ideas of Algorithm Λ is the use of a numbering scheme for the nodes that is different from the one used by Algorithm FGK. We use an *implicit numbering*, in which the node numbering corresponds to the visual representation of the tree. That is, the nodes of the tree are numbered in increasing order by level; nodes on one level are numbered lower than the nodes on the next higher level. Nodes on the same level are numbered in increasing order from left to right. We discuss later in this section how to maintain the implicit numbering via a floating tree data structure.

The node numbering used by Algorithm FGK does not always correspond to the implicit numbering. For example, the numbering of the nodes in Figures 1, 2, and 4 does agree with the implicit numbering, whereas the numbering in Figure 3 is quite different. The odd situation in which an interchange of type ↓ occurs, such as in Figure 3, can no longer happen when the implicit numbering is used. The following lemma lists some useful side effects of implicit numbering.

LEMMA 4.1. *With the implicit numbering, interchanges of type ↓ cannot occur. Also, if the node that moves up in an interchange of type ↑ is an internal node, then the node that moves down must be a leaf.*

PROOF. The first result is obvious from the definition of implicit numbering. Suppose that an interchange of type ↑ occurs between two internal nodes a and b, where a is the node that moves up one level. In the initial tree, since a and b are on different levels, it follows from the sibling property that both a and b must have two children each of exactly half their weight. During the previous execution of the **while** loop in *Update*, q is set to a's right child, which is the highest numbered node of its weight. But this contradicts the fact that b's children have the same weight and are numbered higher in the implicit numbering. □

4.2 INVARIANT. The key to minimizing $D_t - S_t$ is to make ↑'s impossible, except for the first iteration of the **while** loop in *Update*. We can do that by maintaining the following invariant:

(*) For each weight w, all leaves of weight w precede (in the implicit numbering) all internal nodes of weight w.

Any Huffman tree satisfying (*) also minimizes $\sum_j l_j$ and $\max_j\{l_j\}$; this can be proved using the results of [8]. We shall see later that (*) can be maintained by floating trees in real time (that is, in $O(d_t)$ time for the tth processed letter).

LEMMA 4.2. *If the invariant (*) is maintained, then interchanges of type ↑↑ are impossible, and the only possible interchanges of type ↑ must involve the moving up of a leaf.*

PROOF. We shall prove both assertions by contradiction. We remarked at the end of Section 2 that no two nodes of the same weight can be two or more levels apart in the tree, if we ignore the sibling of the 0-node. The effect of the invariant (*) is to allow consideration of the 0-node's sibling. Let us denote the sibling by p and its weight by w. Suppose that there is another node p' of weight w two levels higher in the tree. By the invariant, node p' must be an internal node, since it follows p's parent (which also has weight w) in the implicit numbering. Each child of p' has weight $< w$, but follows p in the implicit numbering, thus contradicting the sibling property. For the second assertion, suppose there is an interchange of type ↑ in which an internal node moves up one level. By Lemma 4.1, the node that moves down must be a leaf. But this violates the invariant, since the leaf initially follows the internal node in the implicit numbering. □

The main result of the paper is the following theorem. It shows for Algorithm Λ that $D_t - S_t < t$.

THEOREM 4.1. *For Algorithm Λ, we have*

$$S_t - k + \delta_{k=n} + \delta_{k<n} \min_{w_j>0}\{w_j\} \le D_t \le S_t + t - 2k + \delta_{k=n} + \delta_{k<n} \min_{w_j>0}\{w_j\} - m.$$

The δ terms and the term $m \ge \min_{w_j>0}\{w_j\} - 1$ have the same interpretation as in Theorem 3.1.

PROOF. By Lemma 4.2, we have $0 \le h_t \le \delta_{\Delta k=0}$. In addition, there are m values of t where the upper bound on h_t can be decreased from 1 to 0. The theorem follows by plugging this into the bound in Theorem 3.2 (which holds not only for Algorithm FGK, but also for Algorithm Λ) and by summing on t. □

Remark. It is important to note that *the version of Update given in Section 2 is never executed by Algorithm* Λ. An entirely different *Update* procedure, which is given later in this section, is called to maintain invariant (*). But, for purposes of analysis, a hypothetical execution of the former version of *Update* does provide, via Theorem 3.2, a precise characterization of $d_t - s_t$.

The lower bound for D_t in Theorem 4.1 is the same as in Theorem 3.1, and the same example shows that it is tight. We can show that the upper bound is tight by generalizing the $D_t/S_t \to \frac{4}{3}$ and $D_t/S_t \to 2$ examples in Section 3. For simplicity, let us assume that $n = 2^j + 1$, for some $j \ge 1$. We construct the message in a "balanced" "balanced" fashion, so that the weights of $n - 1$ letters are within 1 of one another, and the other letter has zero weight. The message begins with $n - 1$ letters in the alphabet, once each. After $\mathcal{M}_{n-1}$ is processed, the leaves of the Huffman tree will be on the same level, except for two leaves on the next lower level, one of which is the 0-node. At each step, the next letter in the message is defined inductively to be the current sibling of the 0-node, so as to force d_t to be always one more than the "average" depth of the leaves. We have $S_t = jt$ and $D_t = S_t + t - 2n + 3$, which matches the upper bound in Theorem 4.1, since $k = n - 1$ and $m = \min_{w_j>0}\{w_j\} - 1 = \lfloor t/(n-1) \rfloor - 1$. Another example consists of appending the nth letter of the alphabet to the above message. In this case we get $S_t = jt + \lfloor (t-1)/(n-1) \rfloor + 1$ and $D_t = S_t + t - 2n + 2 - \lfloor (t-1)/(n-1) \rfloor$, which again matches the upper bound, since $k = n$ and $m = \min_{w_j>0}\{w_j\} - 1 = \lfloor (t-1)/(n-1) \rfloor - 1$.

It is important to note that this construction works for any dynamic Huffman code. The corresponding value of D_t will be at least as large as the value of D_t for Algorithm Λ. This proves that Algorithm Λ is optimum in terms of the model we defined in Section 1:

THEOREM 4.2. *Algorithm Λ minimizes the worst-case difference $D_t - S_t$, over all messages of length t, among all one-pass Huffman algorithms.*

4.3 OUTLINE OF ALGORITHM Λ. In order to maintain the invariant (*), we must keep separate blocks for internal and leaf nodes.

Definition 4.1. *Blocks* are equivalence classes of nodes defined by $v \equiv x$ iff nodes v and x have the same weight and are either both internal nodes or both leaves. The *leader* of a block is the highest numbered (in the implicit numbering) node in a block.

The blocks are linked together by increasing order of weight; a leaf block always precedes an internal block of the same weight. The main operation of the algorithm needed to maintain invariant (*) is the *SlideAndIncrement* operation, illustrated in Figure 6. The version of *Update* we use for Algorithm Λ is outlined below:

```
procedure Update;
begin
leafToIncrement := 0;
q := leaf node corresponding to a_{i_{t+1}};
if (q is the 0-node) and (k < n − 1) then
  begin {Special Case #1}
  Replace q by an internal 0-node with two leaf 0-node children, such that the right child
      corresponds to a_{i_{t+1}};
  q := internal 0-node just created;
  leafToIncrement := the right child of q
  end
else begin
  Interchange q in the tree with the leader of its block;
  if q is the sibling of the 0-node then
    begin {Special Case #2}
    leafToIncrement := q;
    q := parent of q
    end
  end;
while q is not the root of the Huffman tree do
  {Main loop; q must be the leader of its block}
  SlideAndIncrement(q);
if leafToIncrement ≠ 0 then {Handle the two special cases}
  SlideAndIncrement(leafToIncrement)
end;

procedure SlideAndIncrement(p);
begin
wt := weight of p;
b := block following p's block in the linked list;
if ((p is a leaf) and (b is the block of internal nodes of weight wt))
  or ((p is an internal node) and
      (b is the block of leaves of weight wt + 1)) then
    begin
    Slide p in the tree ahead of the nodes in b;
    p's weight := wt + 1;
    if p is a leaf then p := new parent of p
    else p := former parent of p
    end
end;
```

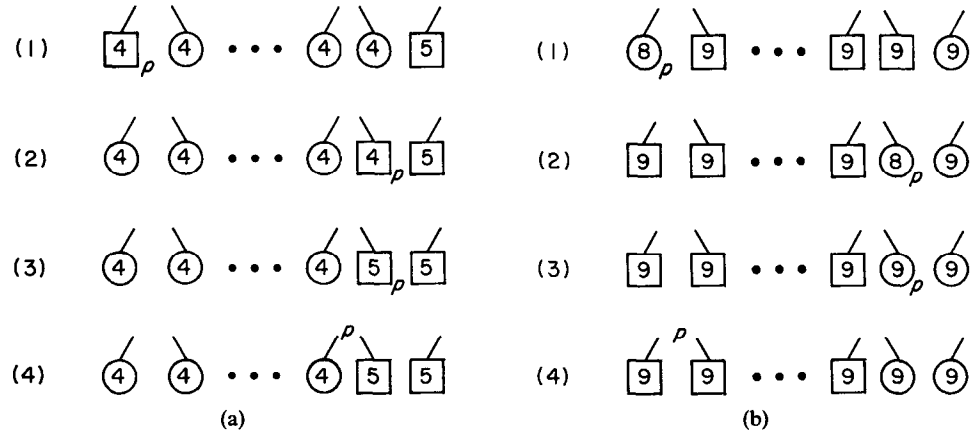


FIG. 6. Algorithm Λ's *SlideAndIncrement* operation. All the nodes in a given block shift to the left one spot to make room for node p, which slides over the block to the right. (a) Node p is a leaf of weight 4. The internal nodes of weight 4 shift to the left. (b) Node p is an internal node of weight 8. The leaves of weight 9 shift to the left.

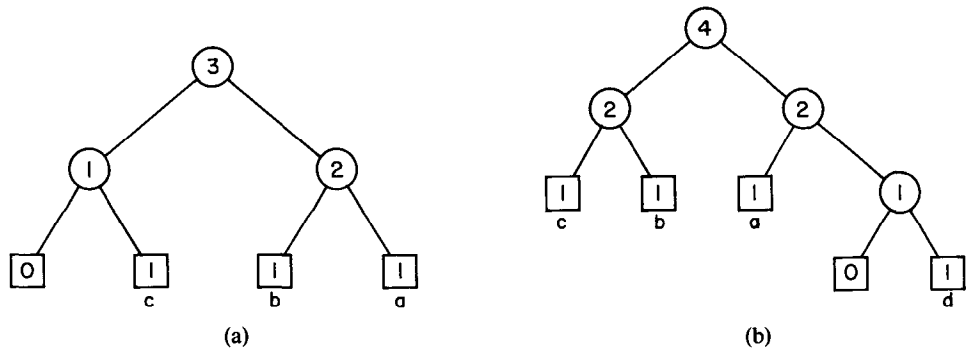


FIG. 7. Algorithm Λ operating on the message "abcd . . .". (a) The Huffman tree immediately before the fourth letter "d" is processed. (b) After *Update* is called.

Examples of Algorithm Λ in operation are given in Figures 7–9; they depict the same examples used to illustrate Algorithm FGK in Figures 2, 4, and 5. As with Algorithm FGK, the processing can be done in $O(d_{t+1})$ time, if the appropriate data structure is used.

4.4 DATA STRUCTURE. In this section we summarize the main features of our data structure for Algorithm Λ. The details and implementation appears in [9]. The main operations that the data structure must support are as follows:

—It must represent a binary Huffman tree with nonnegative weights that maintains invariant (*).
—It must store a contiguous list of internal tree nodes in nondecreasing order by weight; internal nodes of the same weight are ordered with respect to the implicit numbering. A similar list is stored for the leaves.
—It must find the leader of a node's block, for any given node, on the basis of the implicit numbering.
—It must interchange the contents of two leaves of the same weight.

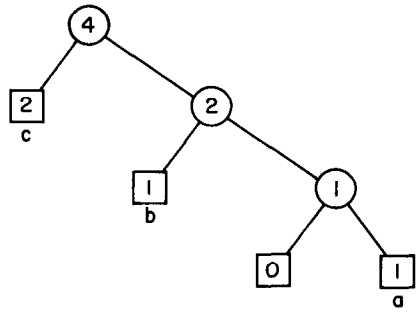


FIG. 8. The Huffman tree that would result from Algorithm Λ if the fourth letter in the example in Figure 7 were "c" rather than "d".

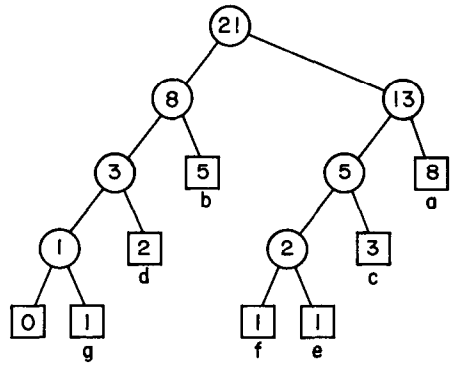


FIG. 9. The Huffman tree constructed by Algorithm Λ for the same message used in Figure 5. Note how much shorter this tree is compared with the one in Figure 5.

—It must increment the weight of the leader of a block by 1, which can cause the node's implicit numbering to "slide" past the numberings of the nodes in the next block, causing their numberings to each decrease by 1.

—It must represent the correspondence between the k letters of the alphabet that have appeared in the message and the positive-weight leaves in the tree.

—It must represent the n–k letters in the alphabet that have not yet appeared in the message by a single leaf 0-node in the Huffman tree.

The data structure makes use of an *explicit numbering*, which corresponds to the physical storage locations used to store information about the nodes. This is not to be confused with the implicit numbering defined in the last section. Leaf nodes are explicitly numbered n, $n - 1$, $n - 2$, ... in contiguous locations, and internal nodes are explicitly numbered $2n - 1$, $2n - 2$, $2n - 3$, ... contiguously; node q is a leaf iff $q \leq n$.

There is a close relationship between the explicit and implicit numberings, as specified in the second operation listed above: For two internal nodes p and q, we have $p < q$ in the explicit numbering iff $p < q$ in the implicit numbering; the same holds for two leaves p and q.

The tree data structure is called a *floating tree* because the parent and child pointers for the nodes are not maintained explicitly. Instead, each block has a *parent* pointer and a *right_child* pointer that point to the parent and right child of the leader of the block. Because of the contiguous storage of leaves and of internal nodes, the locations of the parents and children of the other nodes in the block can be computed in constant time via an offset calculation from the block's *parent* and *right_child* pointer. This allows a node to slide over an entire block without having to update more than a constant number of pointers. Each execution of *SlideAndIncrement* thus takes constant time, so the encoding and decoding in Algorithm Λ can be done in real time.

The total amount of storage needed for the data structure is roughly $16n \log n + 15n + 2n \log t$ bits, which is about $4n \log n$ bits more than used by the implementation of Algorithm FGK in [6]. The storage can be reduced slightly by extra programming. If storage is dynamically allocated, as opposed to preallocated via arrays, it will typically be much less. The running time is comparable to that of Algorithm FGK.

One nice feature of a floating tree, due to the use of implicit numbering, is that the parent of nodes $2j - 1$ and $2j$ is less than the parent of nodes $2j + 1$ and $2j + 2$ in both the implicit and explicit numberings. Such an invariant is not maintained by the data structure in [6]; see Figure 3a, for example.

5. *Empirical Results*

We shall use S_t, D_t^{Λ}, and D_t^{FGK} to denote the communication costs of Huffman's algorithm, Algorithm Λ, and Algorithm FGK. As pointed out at the beginning of Section 3, our evaluation of one-pass algorithms with respect to Huffman's two-pass method is conservative, since we are granting the two-pass method a handicap of $\approx 2k$ bits by not including in S_t the cost of representing the shape of the Huffman tree. The costs S_t, D_t^{Λ}, and D_t^{FGK} also do not count the bits required to encode the correspondence between the leaves of the tree and the letters of the alphabet that occur at least once in the message, but this can be expected to be about the same for the one-pass and two-pass schemes, roughly $k(\log_2 n - 1)$ to $k \log_2 n$ bits using a simple prefix code.

In this section we report on several experiments comparing the three algorithms in terms of coding efficiency. The tables below list not only the costs S_t, D_t^{Λ}, and D_t^{FGK}, but also the corresponding average number of bits used per letter of the message (denoted b/l for each of the three methods), which takes into account the bits needed to describe the tree and the labeling of the leaves. In terms of bits per letter b/l, *Algorithm* Λ *actually outperformed the two-pass method in all the experiments for which* $t \leq 10^4$. Algorithm FGK used slightly more bits per letter, but also performed well.

Algorithm Λ has the advantage of using fewer bits per letter for small messages, where the differences in coding efficiency are relatively more significant. It can be shown using convergence theorems from statistics that, in the limit as $t \to \infty$, the communication cost of the one-pass Huffman algorithms is asymptotically equal to that of the two-pass method for messages whose letters are generated independently according to some fixed probability distribution (discrete memoryless source). Even though the messages used in the longer of our experiments were not generated in such a manner, they are "sufficiently random" that it is not surprising that the statistics for the methods are very close for large t.

In the first experiment, the alphabet consisted of the 95 printable ASCII characters, along with the end-of-line character, for a total of $n = 96$ letters. The message contained 960 letters: The 96 distinct characters repeated as a group 10 times. This is the type of example where all the methods can be expected to perform poorly. The static code does the worst. The results are summarized below at intervals of $t = 100$, 500, and 961:

t	k	S_t	b/l	D_t^{Λ}	b/l	D_t^{FGK}	b/l
100	96	664	13.1	569	10.2	659	11.2
500	96	3320	7.9	3225	7.4	3335	7.6
960	96	6400	7.1	6305	6.8	6415	6.9

The next example was a variation in which all the methods did very well. The message consisted of 10 repetitions of the first character of the alphabet, followed by 10 repetitions of the second character, and so on, for a total message of $t = 960$ letters.

t	k	S_t	b/l	D_t^{Λ}	b/l	D_t^{FGK}	b/l
100	10	340	4.2	340	4.0	345	4.1
500	50	2860	6.5	2820	6.2	2863	6.3
960	96	6400	7.1	6305	6.8	6393	6.9

The third experiment was performed on the Pascal source code used to obtain the statistics for S_t and D_t^{Λ} reported in this section. Again, alphabet size $n = 96$ was used.

t	k	S_t	b/l	D_t^{Λ}	b/l	D_t^{FGK}	b/l
100	34	434	7.1	420	6.3	444	6.5
500	52	2429	5.7	2445	5.5	2489	5.6
1000	58	4864	5.3	4900	5.2	4953	5.3
10000	74	47710	4.8	47852	4.8	47938	4.8
12280	76	58457	4.8	58614	4.8	58708	4.8

The fourth experiment was run on the executable code compiled from the Pascal program mentioned above. The "letters" consisted of 8-bit characters in extended ASCII, so alphabet size $n = 256$ was used.

t	k	S_t	b/l	D_t^{Λ}	b/l	D_t^{FGK}	b/l
100	9	124	2.1	117	1.9	122	2.0
500	9	524	1.2	517	1.2	522	1.2
1000	9	1024	1.1	1017	1.1	1022	1.1
10000	249	52407	5.5	52608	5.4	52868	5.5
34817	256	205688	6.0	206230	6.0	206585	6.0

The message for the final experiment consisted of the device-independent code for a technical book [10], written in TEX, with alphabet size $n = 256$.

t	k	S_t	b/l	D_t^{Λ}	b/l	D_t^{FGK}	b/l
100	40	372	7.7	345	6.7	378	7.0
500	123	2566	7.5	2514	6.9	2625	7.1
1000	177	5904	7.6	5875	7.2	6029	7.3
10000	248	67505	7.0	67769	6.9	67997	7.0
100000	256	691897	6.9	692591	6.9	692858	6.9
588868	256	4170298	7.1	4171314	7.1	4171616	7.1

6. *Conclusions and Open Problems*

The proposed Algorithm Λ performs real-time encoding and decoding of messages in a single pass, using less than one extra bit per letter to encode the message, as compared with Huffman's two-pass algorithm. This is optimum in the worst case, among all one-pass Huffman methods. The experiments reported in Section 5 indicate that the number of bits used by Algorithm Λ is roughly equal (and often better!) than that of the two-pass method. It has much potential for use in file compression and network communication and for hardware implementation.

Algorithm FGK performed almost as well in the experiments. We conjecture that Algorithm FGK uses $O(1)$ extra bits per letter over the two-pass method in the worst case. Note that this does not contradict the examples that appear after Theorem 3.1, for which $D_t/S_t > 1$, since in each case the communication cost per letter is bounded, that is, $D_t = O(t)$. Figure 5 shows that $d_t \neq s_t + O(1)$, so a proof of the conjecture would require an amortized approach.

The one-pass Huffman algorithms we discuss in this paper can be generalized to d-way trees, for $d \geq 2$, for the case in which base-d digits are transmitted instead of bits. Algorithm Λ can also be modified to support the use of a "window" of size $b > 0$, as in [6]. Whenever the next letter in the message is processed, its weight in the tree is increased by 1, and the weight of the letter processed b letters ago is decreased by 1. This technique would work well for the second experiment reported in the previous section.

Huffman coding does not have to be done letter by letter. An alternative well suited for file compression in some domains is to break up the message into maximal-length alphanumeric words and nonalphanumeric words. Each such word is treated as a single "letter" of the alphabet. One Huffman tree can be used for the alphanumeric words, and another for the nonalphanumeric words. The final sizes of the Huffman trees are proportional to the number of distinct words used. In many computer programs written in a high-level language, for example, the vocabulary consists of some variable names and a few frequently used keywords, such as **"while"**, **"iff"**, and **"end"**, so the alphabet size is reasonable. The alphabet size must be bounded beforehand in order for one-pass Huffman algorithms to work efficiently.

Algorithm Λ can also be used to enhance other compression schemes, such as the one-pass method described and analyzed in [1], which is typically used in a word-based setting. A self-organizing cache of size c is used to store representatives of the last c distinct words encountered in the message. When the next word in the message is processed, let l, where $1 \leq l \leq c$, denote its current position in the cache; if the word is not in the cache, we define $l = c + 1$. The word is encoded by an encoding of l, using a suitable prefix code. If $l = c + 1$, this is followed by the encoding of the individual letters in the word, using a separate prefix code. The word's representative is then moved to the front of the cache, bumping other representatives down by one if necessary, and the next word in the message is processed. Similar algorithms are also considered in [2]. The algorithm can be made to run in real time by use of balanced tree techniques, and it uses no more than $S_t + t + 2t \log(1 + S_t/t)$ bits to encode a message containing t words, not counting the extra bits required when the representative is not in the cache. (It is interesting to compare this bound with the corresponding bound $S_t + t - 1$ for Algorithm Λ, which follows from Theorem 4.1.) For any given word that appears more than once in the message, its representative can potentially be absent from the cache each time it is processed, and whenever it is absent, extra bits are required. The method achieves its best coding efficiency when the two prefix codes (used to encode l and the letters in the words for which $l = c + 1$) are dynamic Huffman codes constructed by Algorithm Λ.

ACKNOWLEDGMENTS. The author would like to thank Marc Brown, Bernard Chazelle, and Bob Sedgewick for interesting discussions. Marc's animated Macintosh implementation of Algorithm FGK helped greatly in the testing of Algorithm Λ and in the preparation of the figures. The entropy argument

mentioned at the end of Section 3 is due to Bernard. Bob suggested the $D_t/S_t \to \frac{4}{3}$ example in Section 3. Thanks also go to the referees for their very helpful comments.

REFERENCES

1. BENTLEY, J. L., SLEATOR, D. D., TARJAN, R. E., AND WEI, V. K. A locally adaptive data compression scheme. *Commun. ACM 29*, 4 (Apr. 1986), 320–330.
2. ELIAS, P. Interval and recency-rank source coding: Two online adaptive variable-length schemes. *IEEE Trans. Inf. Theory.* To be published.
3. FALLER, N. An adaptive system for data compression. In *Record of the 7th Asilomar Conference on Circuits, Systems, and Computers.* 1973, pp. 593–597.
4. GALLAGER, R. G. Variations on a theme by Huffman. *IEEE Trans. Inf. Theory IT-24*, 6 (Nov. 1978), 668–674.
5. HUFFMAN, D. A. A method for the construction of minimum redundancy codes. In *Proc. IRE 40* (1951), 1098–1101.
6. KNUTH, D. E. Dynamic Huffman coding. *J. Algorithms 6* (1985), 163–180.
7. MCMASTER, C. L. Documentation of the *compact* command. In *UNIX User's Manual*, 4.2 Berkeley Software Distribution, Virtual VAX-11 Version, Univ. of California, Berkeley, Berkeley, Calif., Mar. 1984.
8. SCHWARTZ, E. S. An Optimum Encoding with Minimum Longest Code and Total Number of Digits. *Inf. Control 7*, 1 (Mar. 1964), 37–44.
9. VITTER, J. S. Dynamic Huffman Coding. *ACM Trans. Math. Softw.* Submitted 1986.
10. VITTER, J. S., AND CHEN, W. C. *Design and Analysis of Coalesced Hashing.* Oxford University Press, New York, 1987.

RECEIVED JUNE 1985; REVISED JANUARY 1987; ACCEPTED APRIL 1987