

CACHE-BASED COMPRESSION/DECOMPRESSION ✓FIELD OF THE INVENTION

This invention relates to methods and apparatus for data compression and decompression, and more particularly, to such methods and apparatus with synchronously operating cache memories.

BACKGROUND OF THE INVENTION

The goal of data compression is to reduce the number of bits required to represent data or an image. The prior art is replete with many methods for data compression. Those methods which provide the highest levels of data compression generally require the most complex data processing equipment and are often slow in execution. Those methods which offer lower levels of data compression often operate faster and employ less complex hardware. In general, the choice of a data compression method is made based upon a compromise between system complexity and time of execution versus desired level of data compression.

The published prior art contains a number of techniques of data compression methods. "Coding of Two-Tone Images", Hwang, IEEE Transactions on Communications, Vol. COM-25, No. 11, November, 1977, pp. 1406-1424, is a review paper that describes a number of techniques for efficient coding of both alphanumeric data and image data. Both single dimension (run length) and two dimension coding (e.g. per block of pixel data) are considered. Hunter et al. in "International Digital Facsimile Coding Standards", Proceedings of the IEEE, Vol. 68, No. 7, July, 1980, pp. 854-867 describe various algorithms used in facsimile transmission (generally one-dimension coding techniques). They also describe twodimension coding schemes wherein conditions of a subsequent coding line are encoded in dependence upon conditions in a previous reference line.

In a paper entitled "An Extremely Fast Ziv-Lempel Data Compression Algorithm" by Williams, Proceedings of the IEEE Data Compression Conference, April, 1991, pp. 362-371, a fast implementation of the Lempel-Ziv (LZ) compression

algorithm is described that employs the well-known LZ method. That method constructs a dictionary of data strings at both the receiving and transmitting nodes and transmits codes in dependence upon matches found between an input data string and
5 a data string found in the dictionary.

Usubuchi et al. in "Adaptive Predictive Coding For Newspaper Facsimile", Proceedings of the IEEE, Vol. 68, No. 1980, pp. 807-813, describe an adaptive, predictive algorithm which is applied to compression of half-tone image data. A
10 further predictive method of encoding of half-tone picture images is described by Stoffel in "Half-tone Pictorial Encoding", SPIE Applications of Digital Image Processing, Vol. 19, 1977, pp. 56-63. Stoffel's algorithm divides an image into blocks and tries to predict the current block from the
15 previous block. The final coded image consists of prediction errors and block values.

In "Compression of Black-White Images with Arithmetic Coding" by Langdon, Jr. et al., IEEE Transactions on Communications, Vol. COM-29, No. 6, June, 1981, pp.
20 858-867, there is described an arithmetic coding method wherein a pixel by pixel probability is estimated based upon the pixel's context (i.e., surrounding pixels). The arithmetic code of Langdon, Jr. et al. avoids multiplication operations inherent in some earlier arithmetic codes. The Langdon, Jr.
25 et al. compression technique is soon to be an international standard for coding of bi-level image data as indicated by Hampel et al., "Technical Features of the JBIG Standard for Progressive Bi-Level Image Compression", Signal Processing: Image Communication Journal, Vol. 4, No. 2 (1992), pp.
30 103-111.

Bentley et al. in "A Locally Adaptive Data Compression Scheme", Communications of the ACM, April 8, 1986, Vol. 29, No. 4, pp. 320-330 and "Technical Correspondence", Communications of the ACM, September 1987, Vol. 30, No. 9, pp.
35 792, 793 describe methods for compression of textual data using a self-organizing sequential search technique. In specific, frequently accessed words are near the top of a search sequence so they are encountered early in the compression action.

The system described by Bentley et al. is similar to a cache memory in a central processing unit memory system. Specifically, a cache memory with least recently used (LRU) management is employed (the cache taking the form of a list of numbers). The list is ordered in a most recently to least recently used fashion. Every time a value or word is encountered, it is removed from the cache list and placed at the top of the cache list, with the rest of the values being "moved down". If a word is encountered which is not in the cache list, then the least recently used value is removed from the cache to make room for the new value (at the top).

Other publications of interest in the field include: "Data Compression - Techniques and Applications, Hardware and Software Considerations", Gilbert Held and Thomas R. Marshall, John Wiley & Sons, 1989; Conference Paper, "CPU Cache Memories", A.J. Smith, California University, Berkeley, CA, USA, Performance Evaluation Review, Vol: 17 Iss: 1, p. 219, May 1989; Journal Paper, "Caching in on Memory Systems", B. Glass, BYTE, Vol: 14 Iss: 3, p. 281-5, March, 1989; and "Cache and Memory Hierarchy Design: A Performance-Directed Approach", San Mateo, California, Morgan Kaufman, 1990, by Steven A. Przybylski.

The tables shown in Fig. 1, illustrate a sequence of actions in such a cache. Table 20 illustrates a four position cache wherein a most recently used value resides at the "top" of the cache and a least recently used value resides at the "bottom" of the cache, with other values residing in intermediate positions. Cache representations 22, 24 and 26 illustrate successive states of cache 20. Cache state 22 is the initial state of cache 20 and shows it storing four decimal values, with decimal value 10 being the most recently used value. Cache state 24 is the state of cache 20 after a value of 6 is encountered. Since the value 6 was already in cache 20, it is moved to the top of the cache and the values 10 and 5 are pushed down. Cache state 26 occurs after a value of 25 is encountered. Since that value is not in cache 20, the value is inserted at the top of cache 20 and the least recently used value (8) is removed and the other values pushed down.

Encoding and decoding processes employed for data compression and decompression manipulate compressed code words and cache adaptation so as to assure cache state synchronization in both encoder and decoder mechanisms. Cache synchronization assures lossless data handling. As shown in Fig. 1, the state of cache 20 is "adapted" as each new value is encountered. If a value is already in the cache of an encoder, then its cache position is transmitted in the form of a code word that indicates the position of the value. If the value is not in the cache, then a special code word is output by the encoder along with the value, per se. Compression is achieved because in general, the cache position can be transmitted with less bits than the value itself. The decoder interprets received code words and reconstructs the original data stream. The decoder further "adapts" its cache in the same way as the encoder, so as to remain synchronous therewith.

The procedure employed to assure synchronism of caches in both encoder and decoder mechanisms is illustrated in Figs. 2 and 3. Fig. 2 illustrates the encoding procedure and Fig. 3 illustrates the decoding procedure. Referring to Fig. 2, an uncompressed data value (e.g. a byte) is accessed (box 30), and it is then determined whether that uncompressed data value matches a value in the compression cache (decision box 32). If yes, the position code indicating the position of the matching value in the cache is outputted (box 34), and the cache is adapted (box 36) by moving, if necessary, the matching cache value up to the top of the cache and accordingly rearranging remaining values. The procedure then recycles to the next uncompressed data value.

If no cache match is found for a received data value, a "not found" code word is issued (box 38), and the actual data value is also outputted (box 40). The received data value is then inserted in the top of the cache and the remainder of the cache is "adapted" by pushing down the remaining values and eliminating the least recently used value.

Upon decompression (Fig. 3), a compressed code word is accessed (box 44) and it is determined whether it contains

a position code word (decision box 46). If yes, the decoder cache value at that position is outputted (box 48) and the cache is adapted by moving the outputted cache value up to the top of the cache (box 50).

5 If, by contrast, the received data is not a position code word, the data value received is accessed and outputted (box 52) and it is also inserted into the cache, at the top, and the remainder of the cache is adapted.

10 While a cache-based compression procedure such as that described above is efficient, the management of a least recently used cache is often computationally complex. Each time the cache is "adapted", a large number of values need to be rearranged. Such rearrangement, as it occurs many times during a compression sequence, can occupy substantial
15 processing time and renders the compression procedure considerably less efficient.

 It is therefore an object of this invention to render a cache based compression procedure more efficient through the use of improved cache management techniques.

20 It is still another object of this invention to provide an improved compression/decompression procedure that employs cache-based prediction techniques.

 It is yet another object of this invention to provide a cache-based compression/decompression procedure that
25 is particularly adapted to image processing and makes use of data content in the vicinity of a value to be compressed.

SUMMARY OF THE INVENTION

 An apparatus and method for compressing bilevel data includes a first cache memory having a plurality of assigned
30 levels of usage, a first usage level assigned to most recently used data segments and a second level assigned to a plurality of less recently used data segments. A processor determines if a received data segment is found in the cache memory and, if not, it assigns the received data segment to
35 the cache memory's first level in place of a previous data segment stored therein. The previous data segment is assigned to a position in the second level in place of a less recently used data segment. The less recently used data segment that

is displaced is chosen by a pseudo-random method. A not-found indication is then transmitted to a receiving station along with the identity of the received data segment. The receiving station contains identical cache structures and updates its
5 caches in response to received code words and data segments. A further aspect of the invention emplaces data segments into cache memories in accordance with the context in which the data segments are found, the cache memories being assigned in accordance with context derived addresses.

10 The invention embraces a data transmitter employing the data compression technology described herein. It also separately embraces a data receiver employing the data decompression technology described herein. Finally, the invention embraces a combination transmitter and receiver
15 employing the data compression/decompression technology described herein.

DESCRIPTION OF THE DRAWINGS

Fig. 1 illustrates prior art adaptation of a least recently used cache in a data encoding system.

20 Fig. 2 is a flow diagram illustrating a prior art process for encoding of data segments through the use of a least recently used cache store.

Fig. 3 is a flow diagram illustrating a prior art process for decoding of data segments through the use of a
25 least recently used cache store.

Fig. 4 is a block diagram of a system adapted to carry out the invention hereof.

Fig. 5 is a schematic representation of cache states, illustrating a process for encoding data using a
30 layered least recently used cache.

Fig. 6 is a diagram which illustrates a context data segment that is employed as an address of a specific one of a plurality of least recently used cache memories.

Fig. 6a is a diagram of an alternative arrangement
35 of context data segments.

Fig. 7 is a flow diagram illustrating a compression encoding procedure for data segments, the procedure employing context dependent least recently used cache stores.

Fig. 8 illustrates a decompression procedure for decoding data that has been encoded in accordance with the flow diagram of Fig. 7.

Fig. 9 illustrates two cache stores that are used
5 to provide indirect least recently used cache management.

Fig. 10 is a flow diagram illustrating the compression procedure for indirect cache management in accordance with the cache stores shown in Fig. 9.

Fig. 11 is a flow diagram illustrating a
10 decompression procedure for decoding data segments encoded in accordance with the flow diagram of Fig. 10.

Fig. 12 illustrates a look-up table that indicates whether a byte is stored in cache.

Fig. 13 illustrates an encoder cache arrangement
15 having a flag associated with each storage position that indicates a defined or undefined status for the storage position.

DETAILED DESCRIPTION OF THE INVENTION

Hereafter, the invention will be described in the
20 context of a host processor communicating with a printer, with encoding being carried out on bytes (8 bits per byte) of transmitted data. It is to be realized, however, that the procedure and system hereafter to be described apply to a variety of systems that communicate data and employ
25 compression/decompression procedures for transmission bandwidth economies and memory savings.

In Fig. 4, a host processor 60 includes a host central processing unit (CPU) 62, a read only memory (ROM) 64 and a random access memory (RAM) 66. Each of the aforesaid
30 modules is interconnected by a bus 68, with input/output communications being handled by I/O module 70. ROM 64 includes encoding procedures that perform data compression operations under control of host CPU 62, prior to transmitting data via I/O module 70 over a transmission line. RAM 66 contains a
35 plurality of cache memories that are structured in a least recently used (LRU) fashion to enable data compression to be achieved.

Data transmitted from I/O module 70 is received in a printer 72 through I/O module 73. Printer 72 comprises a CPU

74, print engine 76, RAM 78 and ROM 80, which are interconnected by a bus 82. RAM 78 serves as temporary storage for incoming data received through I/O module 73 and also includes cache memories that are identical in structure to the
 5 cache memories contained in RAM 66 in host processor 60. ROM 80 contains decode procedures that enable compressed data received via I/O module 73 to be decompressed and then stored for usage in RAM 78.

As above indicated, according to prior art
 10 arrangements, cache memories contained in RAMs 66 and 78 are structured as LRU caches, wherein data at the top of the cache is the most recently used and data at the bottom of the cache is the least recently used. Management of such a cache can be computationally complex because each time the cache is revised
 15 ("adapted"), a large number of values need to be rearranged.

In Fig. 5, a modified form of a LRU cache is illustrated that reduces computational complexity. Cache 90 is organized as a layered, LRU cache wherein topmost layer 91 is the most recently used data byte and the layers 92-94
 20 comprise a "single" layer that includes the three least recently used data bytes. It is to be understood that cache 90 is shown as only having four entries and two layers for ease of description purposes only. It is possible to have any number of layers and entries per layer.

25 Choosing which value to remove from a particular layer of cache 90 can be accomplished pseudo-randomly while still maintaining synchronism between the encoding and decoding mechanisms. For example, a round robin count approach can be utilized wherein bytes to be removed from the least
 30 recently used layer are selected upon the basis of a sequential count of entries 92-94. In this manner, a larger cache can be maintained with less processing overhead and still be almost effective as an LRU cache.

Cache state representations 95, 96 and 97 show the
 35 state of cache 90 at three different times, i.e., A, B and C. At time A, cache state 95 is as illustrated. At time B, it is assumed that a byte of value 2 is encountered and is placed in most recently used layer 91. Since a byte of value 2 was already in the second level of the cache, the value 78 in the

first level swaps positions with it, thereby causing 78 to be placed in position 94 (as shown in cache state 96). At time C, it is assumed that a byte of value 89 is encountered. Since this value is not in the cache, it is inserted into most recently used layer 91 and a pseudorandomly derived position from the LRU layer (in this case position 93) is overwritten with the value that was displaced by the value 89 (i.e. value 2).

The next time a value in the LRU layer requires replacement, it will be position 94 that will be displaced and then position 92, position 93 etc. As a result, the layering of cache 90 enables less movement of cache values during adaptation. Structurally, the position to be replaced in the LRU layer may be determined by the position of a pointer that is updated, as the round robin count is incremented.

Additional performance gains may be achieved by employing context to select from among a number of independent caches in which to emplace a newly received byte. A printer is generally supplied with a bit map memory that comprises a raster image of pixels, each pixel represented by a binary bit value. Each raster scan row is divided into a series of 8 bit bytes. Surrounding pixels which are spatially positioned so as to have been previously decoded are used as the context for deciding which cache is to be used to store a most recently used byte. Bit values from surrounding pixels are combined to form a number, enabling a context value (or index) to be derived which specifies which cache should be used. This action improves data compression by allowing each cache to better adapt itself specifically for each context.

In the diagram shown in Fig. 6, the eight numbered pixels from a raster scan line (pixels 100) form the context value for the eight shaded pixels on the next raster scan to be coded (pixels 102). Hereafter, pixels 100 will be termed the context byte and pixels 102 will be termed the current byte. The philosophy of a context dependent cache arrangement is based upon the fact that images are often vertically related (e.g. vertical lines in a chart arrangement) and by employing a context byte as a cache address, it is likely that a current byte will encounter, in the most recently used entry

in the addressed cache, an identical data byte - thereby enabling rapid encoding without further cache data movement. While a "vertical" context byte will form the basis of the ensuing description, it is to be understood that other
5 physical relationships are within the scope of the invention.

In the implementation of a context dependent cache structure, all caches may be created and initialized at the start of a compression/decompression procedure. For instance, if an eight bit byte is to be employed, 256 caches may be
10 created at the outset, each cache having a plurality of possible entries. If memory is more important than cache complexity, then not all of the caches need be created at the outset, and caches can be created only when needed.

As shown in Fig. 7, once a new data byte is
15 encountered that is to be transmitted (decision box 104), a specific cache is addressed having an address of the context byte which is vertically aligned on an immediately previous scan row of the pixel image (box 106). (As indicated above, it is likely that the most recently used entry in the
20 addressed cache will be identical to the uncompressed byte value). The byte value is then matched with the values in the addressed cache (decision box 108) and if a match is found, a code word indicating the position of the matching value in the addressed cache is outputted (box 110). The cache is then
25 adapted by moving, if necessary, the matching value up to the most recently used entry in the addressed cache (box 112) and the procedure recycles.

If no match is found in the addressed cache (decision box 108), a "not found" code word is outputted (box
30 114) and the actual byte value is transmitted (box 116). The non-matching byte value is inserted in the addressed cache and the cache is adapted (box 118) so as to move the displaced byte into a lower level of the cache and to displace a byte already in that lower level, etc. The procedure then recycles.

35 In Fig. 8, a decompression procedure is shown that is responsive to data compressed by the procedure shown in Fig. 7. A received compressed code word is accessed (box 120) and a cache is addressed, based upon the context value of the already decoded context byte in the scan row immediately above

(box 122). If the compressed code word is a position code word (decision box 124), the value appearing at the indicated position in the addressed cache is outputted (box 126) and the cache is then adapted, if necessary by rearranging the entries
 5 therein to enable the most recently used byte to be placed at the top of the cache (box 128).

If it is determined that the received data is not a position code word (decision box 124), a received byte value is accessed from memory (box 130) and is outputted. That new
 10 byte value is then inserted into the top of the addressed cache and the cache is adapted (box 132). The procedure recycles until no further received data is available.

In summary, a current row byte will always be directed to a cache whose address is the value of the context
 15 byte immediately above the current byte. The current byte becomes a context byte when the next scan row is accessed. If the context byte value is already an address of a cache, a new cache does not need to be created. If the context byte value is new, a new cache is established with the context byte as
 20 its address. If all caches were established at initialization, no new caches are required, as one cache for each possible context byte value already exists.

In the above discussion, the notion of a context value was introduced. In that discussion the context value
 25 consisted of the value of the data segment from the raster scan line immediately above the current line. That context value was used as the cache address. The context value can also be formed from a combination of previously decoded data segments. The advantage is that more context can sometimes
 30 provide a better means of selecting the best performing cache to code the current data segment. One specific implementation of this concept uses not only the data segment directly above the current data segment but a piece of the data segment immediately above and to the left of the current data segment,
 35 as shown in Fig. 6a. One way to combine these two data segments into a context value is to append two bits from the data segment above and to the left to the data segment above. In Fig. 6a the byte labeled "A" is the current data segment. The bytes labeled "B" and "C" are the data segments which are

used to form the context value. The context value can be generated by concatenating bits 1 and 2 of data segment B to bits 1 through 8 of data segment C, thereby generating a 10 bit data context value.

5 In some systems, it is desirable to reduce the complexity of decoding/decompression procedures at the expense of providing a more complex encoding procedure. Indirect cache management enables a portion of the decoding procedure to be transferred to the encoding procedure. This transfer is made
10 possible due to the fact that in the case of an LRU cache, so long as a newly received data byte is found already stored in the cache, values in the cache do not change but are just rearranged. The only time a set of values in an LRU cache changes is when no match is found between a new data byte and
15 bytes already stored in the cache. In indirect cache management, an encoder of added complexity is employed so that the decoder does not have to perform the value rearrangements necessary to maintain an LRU cache.

Referring to Fig.9, a pair of cache memories 120 and
20 122 are shown that enable an encoding procedure to maintain two lists of numbers i.e., a pointer list in cache 120 and a byte value list in cache 122. The pointer list comprises pointers or indices to entries in the value list. When a match is found between a newly received byte and a byte already
25 stored in the byte value list in cache 122, pointers in the pointer list in cache 120 are rearranged in an LRU manner. The values in the value list do not require rearrangement because no new value has entered the cache and no old value has been removed. Upon finding a match (between a new byte and a byte
30 already stored in cache 122), the encoder emits the corresponding pointer value (i.e., indicates the value entry in cache 122 that corresponds to the received byte).

If a match is not found in cache 122, the value list and pointer list must be modified. The least recently used
35 value in the value list in cache 122 is removed. This is accomplished by placing the current byte value in the value list position in cache 122 pointed to by the least recently used pointer in cache 120.

Then, the pointers in cache 120 are updated to reflect the new order of values in cache 122. The newly entered current value is pointed to by the most recently used position in cache 120. In this case, the encoder emits a special code word signalling
 5 that no match was found, followed by the actual byte value and the pointer to the value list where it is to be inserted (in the decoder).

As is shown in Fig. 9, pointer list entries in cache 120 are arranged in a most recently used (MRU) to least recently used (LRU) sequence. Thus, irrespective of the
 10 position of the byte value in the value list in cache 122, it is known from the position of a pointer in the pointer list what the most recently used byte value is in the value list and all succeeding usage levels for remaining stored byte
 15 values.

In the decoder, only a value list is maintained, not a pointer list. When a code word (e.g.a. pointer value) is decoded, the value from the pointed-to position in the value list is outputted. If a "not found" code word is received, a
 20 pointer and the value are derived from the input stream and the value is placed in the value list in the decoder at the pointed-to position.

Referring to Fig. 10, the procedure for indirect cache management encoding (compression) will be described.
 25 When a new data value is encountered (box 130), it is determined whether that value matches a value already stored in cache 122 containing the value list (box 132). If yes, a pointer value is outputted from the pointer list that identifies where in the value list the matching value resides
 30 (box 134). Next, the pointers in the pointer list are adapted so as to properly indicate the most recently used byte value in the value list.

If no cache value match is found (decision box 132), a "not found" code word is outputted (box 138) accompanied by
 35 the "not found" data value (box 140). In addition, a pointer value is outputted that points to the least recently used value in the value list. The new data value is then inserted in the least recently used value position in the value list and the value previously stored therein is removed. The

procedure then recycles to obtain the next uncompressed data value.

In Fig. 11, the decoding (decompression) procedure is shown and commences by accessing a next compressed code word (box 150). If the code word is found to be a pointer, the "pointed to" value in the corresponding value list is outputted (decision box 152, box 154). If the code word is found not to be a pointer value (decision box 152), the received data value is accessed (box 156) along with the replacement pointer value (box 158). The new data value is then inserted in the pointed-to position in the value list. The value already there is discarded (box 160). The procedure recycles until there are no further compressed code words to process. It should be understood that indirect cache management and layered cache management may be employed in combination in order to gain benefits inherent in both techniques.

The system as described above has been indicated as encoding data in 8 bit bytes. If the data exhibits a different periodicity, i.e. 6 bits, 10 bits, etc., as in some dithered image data, the data can be broken up into units of bit length which match the image data periodicity.

An aspect of the above algorithms is that in order to determine if a value is in a cache, the value must be compared against every stored value in the cache. This comparison procedure may be avoided by maintaining a look-up table which contains a flag as to whether or not a particular value is in the cache (see table 170 in FIG 12). Look-up table 170 must contain every possible value. The procedure still requires that if the value is present in the cache that its position must be determined in the cache.

As stated above, caches in both an encoder and decoder must be in the same initial state to ensure that they are synchronized during compression and decompression actions. This initialization can be eliminated in the decoder at the expense of some added complexity in the encoder. The encoder can associate a flag with every cache entry (such as shown in cache 180 in FIG 13). This flag will indicate whether or not that cache entry has an undefined or defined value. A defined

value is one wherein the cache entry represents an actual data byte value from the image being processed. The encoder can make sure never to match a value with an entry which has an undefined cache value. The decoder need not maintain those
 5 flags as it will never use an un-initialized value, since it will never be instructed to do so by the encoder.

The following example illustrates the operation of the invention with a two layered cache. The first layer has one assigned value and the second layer has three values. Code
 10 words (bit sequences) are assigned as follows:

1	=	first layer prediction correct
000	=	second layer prediction #1 correct
001	=	second layer prediction #2 correct
010	=	second layer prediction #3 correct
15 011	=	correction value is needed

The cache starts out in the following condition:

1st layer	2nd layer - #1	2nd layer - #2	2nd layer #3
20 0	255	255	255

Random replacement in the second level of the cache is performed via a simple round robin approach. A variable "repl_index" is defined and initially set to 1. It determines
 25 which of the 2nd level cache entries will be replaced when a replacement is required. After a replacement is performed, "repl_index" is incremented. If "repl_index" reaches a value of four, it is reset to one.

The data stream to be coded consists of a sequence
 30 of bytes.

The following table traces the coding sequence. The cache state is shown after the current input value is processed.

Input Value	Cache State				Output Description and Action
	1st Layer	2nd Layer #1	3rd Layer #2	4th Layer #3	
100	100	0	0	0	"011" & "01100100" is output; the cache is adapted; repl_num = repl_num + 1 (now=2)
17	17	0	100	0	"011" & "00010001" is output; the cache is adapted; repl_num = repl_num + 1 (now=3)
100	100	0	17	0	"001" is output; the cache is adapted
100	100	0	17	0	"1" is output
9	9	0	17	100	"011" & "00001001" is output; the cache is adapted; repl_num = repl_num + 1 (now=4); repl_num = 1

It should be understood that the foregoing description is only illustrative of the invention. Various alternatives and modifications can be devised by those skilled in the art without departing from the invention. Accordingly, the present invention is intended to embrace all such alternatives, modifications and variances which fall within the scope of the appended claims.

CLAIMS

1. An apparatus for efficient data compression, comprising:

first cache memory means (20,66) having a plurality of levels, a first said level assigned to a most recently used data segment, a second said level assigned to a plurality of less recently used data segments;

means (62,32) for determining if a received data segment is found in said first cache memory means (20,66) and, if ~~it~~ assigning said received data segment to said first level in said first cache memory means in place of a previously received data segment assigned to said first level, and assigning said previously received data segment to a position in said second level in place of a less recently used data segment, said less recently used data segment chosen by a pseudo-random method from among less recently used data segments assigned to said second level;

means (62,38) for transmitting to a receiving means a "not found" indication and said received data segment in uncompressed form; and

means (62,34) responsive to a determination that a received data segment is found in said first cache memory means for causing said transmitting means to transmit to said receiving means a position code indicating where in said first cache memory means said received data segment is assigned, said position code comprising a compressed form of said received data segment.

2. The apparatus as recited in Claim 1, including a receiving means said receiving means including means for decompressing compressed form data comprising:

second cache memory means (20,78) having a plurality of levels, a first said level assigned to a most recently used data segment, a second said level assigned to a plurality of less recently used data segments; and

means (74) for performing said pseudo-random method so as to maintain data segments in conjunction with said second cache memory means so that data assigned thereto is maintained



in synchronism with data segments in said first cache memory means.

3. The apparatus as recited in Claim 1 or 2, wherein said first cache memory means (20,66) comprises more than two assigned levels.

4. The apparatus as recited in any one of the preceding claims, wherein said pseudo-random method sequences a count through data segments assigned to a level having a plurality of less recently used data segments by a round robin count thereof, a data segment to be replaced being determined by a value of said count.

5. The apparatus as recited in any one of the preceding claims, wherein said means for determining (62,32) includes a look-up table (170) including a position for each possible value of a data segment and an associated flag to indicate whether the associated data segment value is present in said first cache memory means.

6. In a system for compressing a raster scan representation of bi-level data through a use of multiple cache stores, each cache store for storing a plurality of data segments from said raster scan representation in a hierarchy of usage, a method comprising the steps of:

a. assigning a cache context identifier (100) to each of a plurality of cache stores, each cache context identifier having a value related to a data segment in one scan row of said raster scan representation;

b. accessing (104,106) current data segments contained in a next scan row from said one scan row, each said current data segment positionally aligned with a context data segment in said one scan row;

c. comparing (108) each said current data segment with data segments contained in a cache store having an assigned cache identifier equal in value to a positionally aligned context data segment;

d. transmitting (110) a position identifier of a data segment within a cache store if said current data segment is found in a said cache store, or a "not found" indicator (114) and said current data segment (116) if said current data segment is not found in said cache store; and



e. causing current data segments contained in said next scan row to become context data segments for a succeeding scan row of data segments; and

f. assigning to a new cache store, a new context segment value if no other cache store has that context segment value assigned and repeating steps (b-f) until no further scan rows remain to be processed.

7. The method as recited in Claim 6, further comprising the decompression steps of:

g. establishing a plurality of cache stores, each cache store having an assigned context identifier related in value to a data segment value in a first scan row of said raster scan representation; and

h. decompressing (106,108,110,114,116) a received position identifier by selecting a cache store based upon a context identifier value in a decompressed preceding scan row, and determining from contents of said selected cache store at a position therein indicated by said position identifier, an identity of a data segment assigned to said position.

8. The method as recited in Claim 7, wherein each said cache store (180) comprises a plurality of storage positions, arranged to store data segments in a most recently used to least recently used order, each storage position having an associated flag value indicating whether the position has a defined value data segment or an undefined value data segment stored therein, said comparing step (c) ignoring any storage position having a flag value indicating an undefined value data segment.

9. In a system for compressing a stream of data segments through a use of multiple cache stores, each cache store for storing a plurality of data segments from said stream in a hierarchy of usage, a method comprising the steps of:

a. assigning a cache context identifier (B,C) to each of a plurality of cache stores, each cache context identifier related in value to a context data segment previously encountered in said data stream:

b. accessing (104,106) a current data segment having a predetermined temporal relationship with a context data segment;



c. comparing (108) said current data segment with data segments contained in a cache store having an assigned cache identifier equal in value to the context data segment, but if no cache store exists with a said assigned cache identifier, establishing a cache store with said cache identifier;

d. transmitting (110) a position identifier of a data segment within a cache store if said current data segment is found in a said cache store, or a "not found" indicator (114) and said current data segment if said current data segment is not found in said cache store;

e. causing the current data segment to become a new context data segment for a succeeding current data segment; and

f. assigning to a new cache store, said new context data segment value if no other cache store has that context data segment value assigned and repeating steps (b-f) until no further data segments remain to be processed.

10. The method as recited in Claim 9, further comprising the decompression steps of:

g. establishing a plurality of cache stores, each cache store having an assigned context identifier related in value to a data segment value previously encountered in said data stream; and

h. decompressing (122,124,126,130) a received position identifier by selecting a cache store based upon a context identifier value assigned to said cache store, and determining from contents of said selected cache store at a position therein indicated by said position identifier, an identity of a data segment assigned to said position.

11. The method as recited in Claim 9 or 10, wherein each said cache store (170) comprises a plurality of storage positions, arranged to store data segments in a most recently used to least recently used order, each storage position having an associated flag value indicating whether the position has a defined value data segment or an undefined value data segment stored therein, said comparing step (c) ignoring any storage position having a flag value indicating an undefined value data segment.

12. In a system for compressing a raster scan representation of bi-level data through a use of multiple



cache stores, each cache store for storing a plurality of data segments from said raster scan representation in a hierarchy of usage, a method comprising the steps of:

- a. assigning a cache context identifier to each of a plurality of caches, said plurality equal in number to possible values of said data segments, each cache context identifier equal in value to a possible value of a data segment;

- b. accessing current data segments (104,106) contained in a next scan row from a first scan row, each said current data segment positionally aligned with a context data segment in said first scan row;

- c. comparing (108) each said current data segment with data segments contained in a cache store having an assigned cache identifier equal in value to a positionally aligned context data segment;

- d. transmitting (110) a position identifier of a data segment found in a said cache store, or a "not found" indicator (114) and said current data segment (116) if said current data segment is not found in said cache store; and

- e. causing current data segments contained in said next scan row to become context data segments for a succeeding scan row of data segments; and

- f. repeating steps (b-f) until no further scan rows remain to be processed.

13. The method as recited in Claim 12, further comprising the decompression steps of:

- g. establishing a plurality of cache stores, each cache store having an assigned context identifier equal to a possible value of a data segment; and

- h. decompressing (120,122,124,126,130) a received position identifier by selecting a cache store based upon a context identifier value in a decompressed preceding scan row, and determining from contents of said selected cache store, an identity of a data segment assigned to a position indicated by said position identifiers.

14. The method as recited in Claim 12 or 13, wherein said system includes a look-up table (170) including positions for each possible value of a data segment and an associated flag



to indicate whether the associated data segment value is present in a cache store, said comparing step (c) initially examining said look-up table to determine if a data segment value is contained in said cache store.

15. The method as recited in Claim 12 or 13, wherein each said cache store (180) comprises a plurality of storage positions arranged to store data segments in a most recently used to least recently used order, each storage position having an associated flag value indicating whether the position has a defined value data segment or an undefined value data segment stored therein, said comparing step (c) ignoring any storage position having a flag value indicating an undefined value data segment.

16. In a system for compressing a stream of data segments through a use of multiple cache stores, each cache store for storing a plurality of data segments from said stream in a hierarchy of usage, a method comprising the steps of:

a. assigning a cache context identifier to each of a plurality of caches, said plurality equal in number to possible values of said data segments, each cache context identifier equal in value to a possible value of a data segment;

b. accessing a current data segment (104,106) having a predetermined temporal relationship with a context data segment;

c. comparing (108) said current data segment with data segments contained in a cache store having an assigned cache identifier equal in value to the context data segment;

d. transmitting (110) a position identifier of a data segment found in a said cache store, or a "not found" indicator (114) and said current data segment (116) if said current data segment is not found in said cache store; and

e. causing the current data segment to become a new context data segment for a succeeding current data segment; and

f. repeating steps (b-f) until no further data segments remain to be processed.

17. The method as recited in Claim 16, further comprising the decompression steps of:



g. establishing a plurality of cache stores, each cache store having an assigned context identifier value equal to a possible value of a data segment; and

h. decompressing (120,122,124,126,130) a received position identifier by selecting a cache store based upon a context identifier value and determining from contents of said selected cache store, an identity of a data segment assigned to a position indicated by said position identifiers.

18. The method as recited in Claim 16 or 17, wherein said system includes a look up table (170) including positions for each possible value of a data segment and an associated flag to indicate whether the associated data segment value is present in a cache store, said comparing step (c) initially examining said look-up table to determine if a data segment value is contained in said cache store.

19. The method as recited in Claim 16 or 17, wherein each said cache store (180) comprises a plurality of storage positions arranged to store data segments in a most recently used to least recently used order, each storage position having an associated flag value indicating whether the position has a defined value data segment or an undefined value data segment stored therein, said comparing step (c) ignoring any storage position having a flag value indicating an undefined value data segment.

20. An apparatus for efficient data compression/decompression comprising:

first cache memory means (122) for storing a plurality of data segments in a plurality of memory positions, each position having a position identifier;

second cache memory means (120) having a plurality of levels, a first said level including a position identifier for a most recently used data segment, said position identifier identifying where in said first cache memory means is stored said most recently used data segment and remaining levels including position identifiers that identify storage positions in said first cache memory means wherein are stored for less recently used data segments;

means (62,132) for determining if a received data segment is found in said first cache memory means (122) and if found,



transmitting (134) to receiving means said position identifier from said second cache memory means that indicates the position in said first cache memory means that stores said received data segment.

21. The system as recited in Claim 20, wherein, if said determining means does not find said received data segment in said first cache memory means, a "not found" indication (138), a pointer value (142) indicating a position in said first cache memory means storing a least recently used data segment and said received data segment (140) are transmitted to said receiving means.

22. The apparatus as recited in Claim 21, further comprising:

means (62,144) for storing the "not found" received data segment in a position in said first cache memory means (122) in place of a least recently used data segment and for updating said second cache memory means to indicate said position in said first cache memory means that stores said "not found", received data segment, as being a most recently used data segment, and adapting remaining position identifiers in said second cache memory means accordingly.

23. The apparatus as recited in Claim 22, wherein said receiving means comprises:

receiving cache memory means (122) for storing a plurality of data segments in synchronism with data segments stored in said first cache memory means; and

means (62,152,154,156,158,160) responsive to a received position identifier for accessing a corresponding position in said receiving cache memory means and a data segment stored therein.

24. An apparatus for efficient data compression, comprising:

first cache memory means (20,66) having a plurality of levels for storing recently used data segments in accordance with a predetermined sequence; and

look-up table means (170) containing every possible data value and a flag associated with each data value, said flag indicating whether or not its associated data value is stored in said cache memory, and if it is, determining the position



of the data value in said cache memory, outputting the data value stored in said position, and adapting said cache memory, and if said associated data value is not stored in said cache memory, outputting said data value, inserting said data value in said cache memory, adapting said cache memory, and flagging said data value in said look-up table means.

25. An apparatus for efficient data decompression of compressed form data, comprising:

receiving means (74,44) for receiving transmitted data including a position code or, alternatively a "not found" indication and a data segment in uncompressed form;

means (74,46) for determining if received data is a position code or a "not found" indication and a data segment in uncompressed form;

a cache memory means (20,78) having a plurality of levels, a first said level assigned to a most recently used data segment, a second said level assigned to a plurality of less recently used data segments;

means (74,46) responsive to receipt of a "not found" indication and a data segment in uncompressed form for assigning said received data segment to said first level in said first cache memory means in place of a previously received data segment assigned to said first level, and assigning said previously received data segment to a position in said second level in place of a less recently used data segment, said less recently used data segment chosen by a pseudo-random method from among less recently used data segments assigned to said second level; and

means (74,48) responsive to a determination that a received data is a position code for causing said receiving means to output the data segment in said first cache memory means where said position code indicates, said position code comprising a compressed form of said received data segment.

26. In a system for decompressing a raster scan representation of bi-level data through a use of multiple cache stores, each cache store for storing a plurality of data segments from said raster scan representation in a hierarchy of usage, a method comprising the steps of:



establishing a plurality of cache stores, each cache store having an assigned context identifier related in value to a data segment value in a first scan row of said raster scan representation; and

decompressing (106,108,110,114,116) a received position identifier by selecting a cache store based upon a context identifier value in a decompressed preceding scan row, and determining from contents of said selected cache store at a position therein indicated by said position identifier, an identity of a data segment assigned to said position.

27. In a system for decompressing a stream of data segments through a use of multiple cache stores, each cache store for storing a plurality of data segments from said stream in a hierarchy of usage, a method comprising the steps of:

establishing a plurality of cache stores, each cache store having an assigned context identifier related in value to a data segment value previously encountered in said data stream; and

decompressing (122,124,126,130) a received position identifier by selecting a cache store based upon a context identifier value assigned to said cache store, and determining from contents of said selected cache store at a position therein indicated by said position identifier, an identity of a data segment assigned to said position.

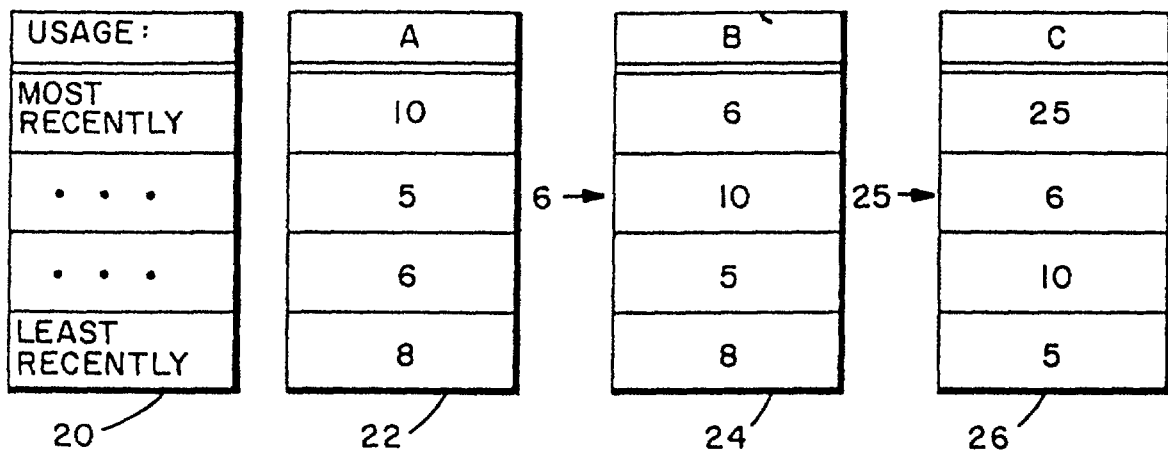


FIG. 1
PRIOR ART

THIS SHEET HAS BEEN REPLACED

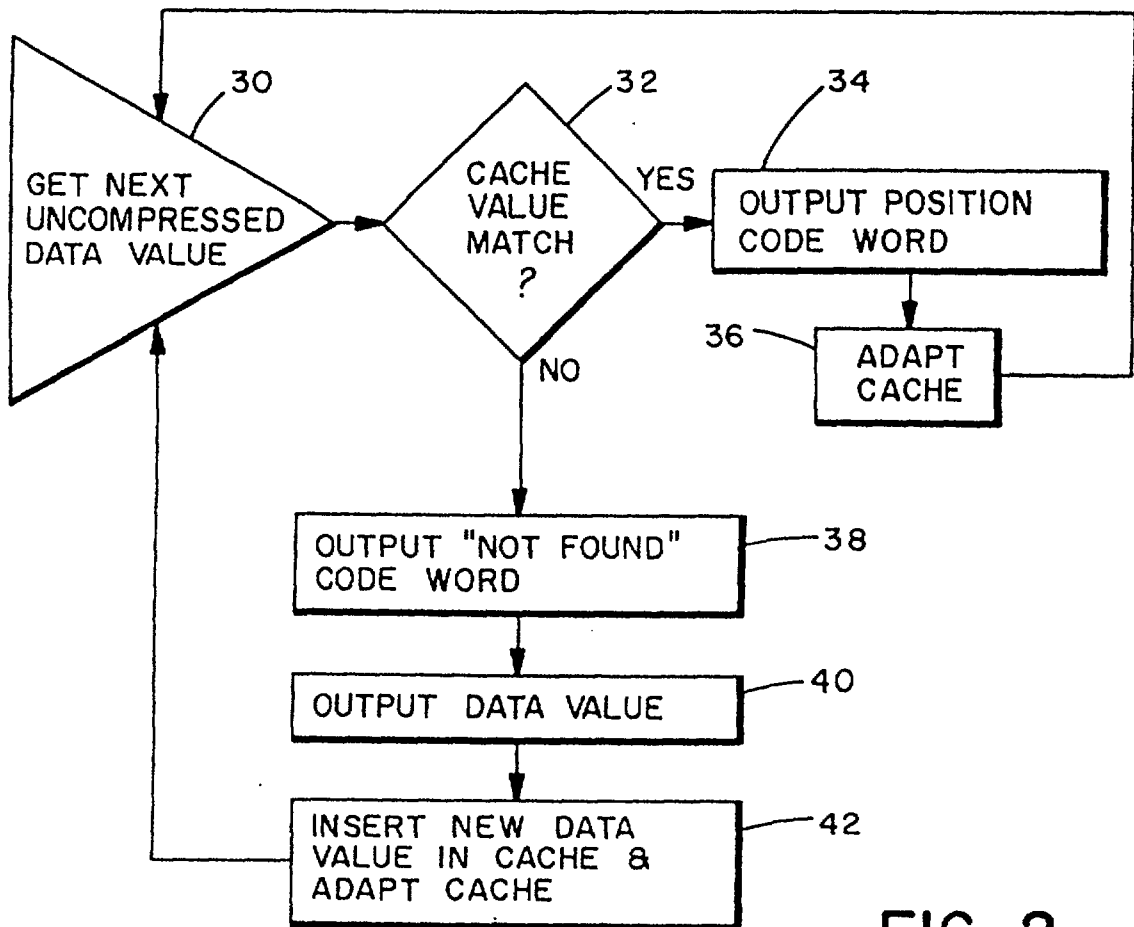


FIG. 2
PRIOR ART



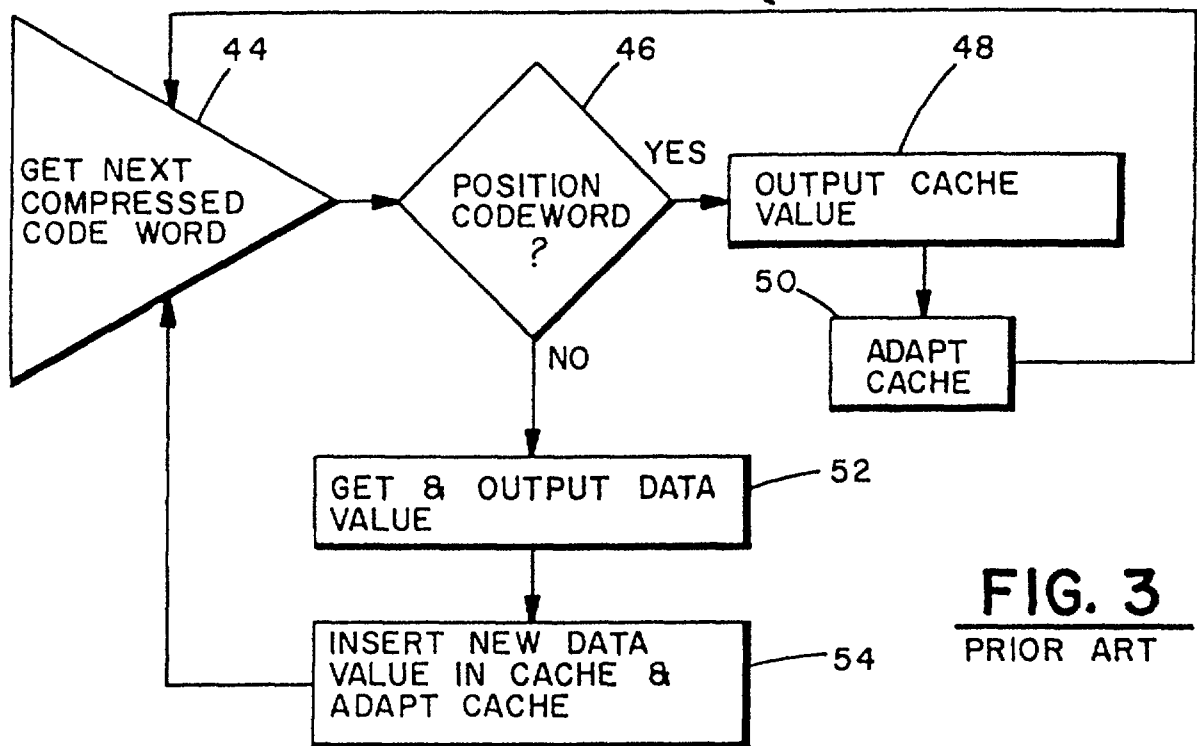


FIG. 3
PRIOR ART

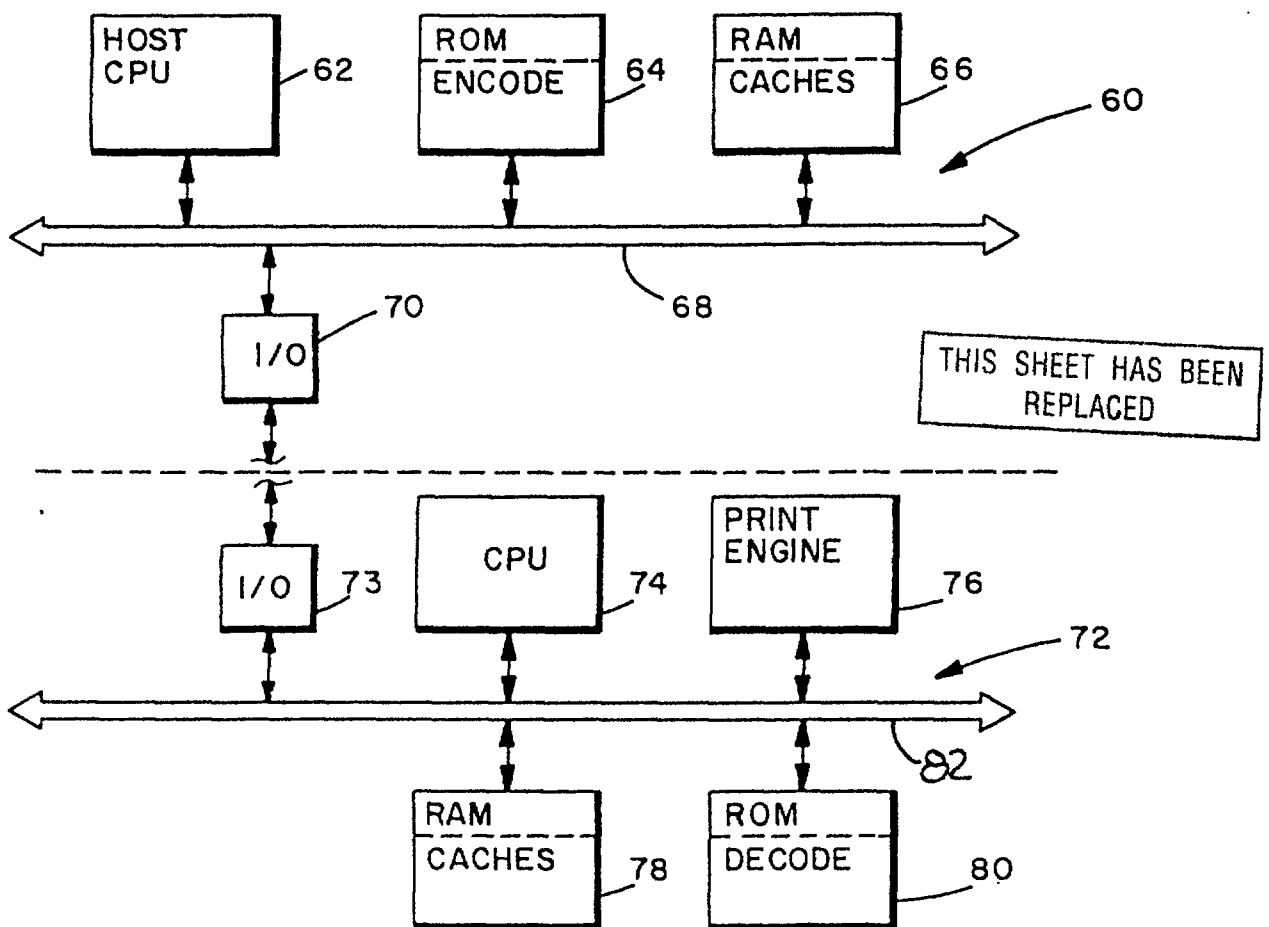


FIG. 4



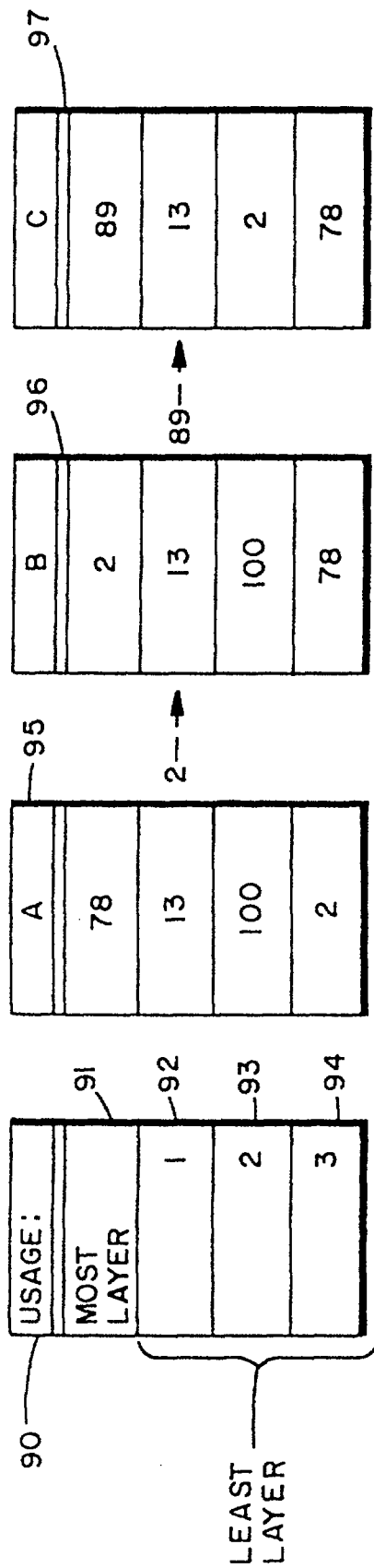


FIG. 5

THIS SHEET HAS BEEN REPLACED

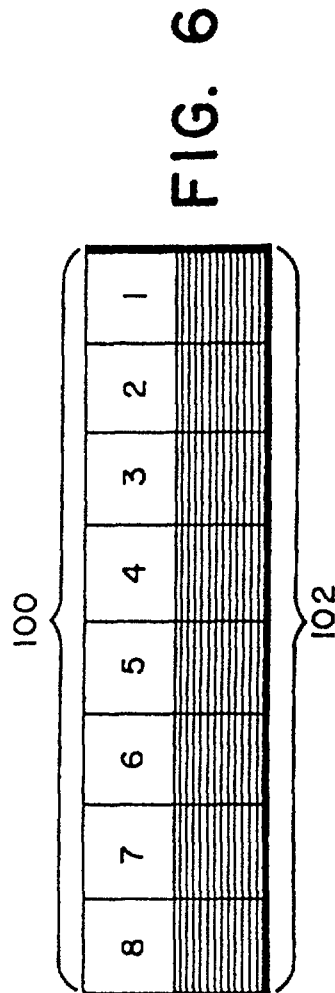


FIG. 6

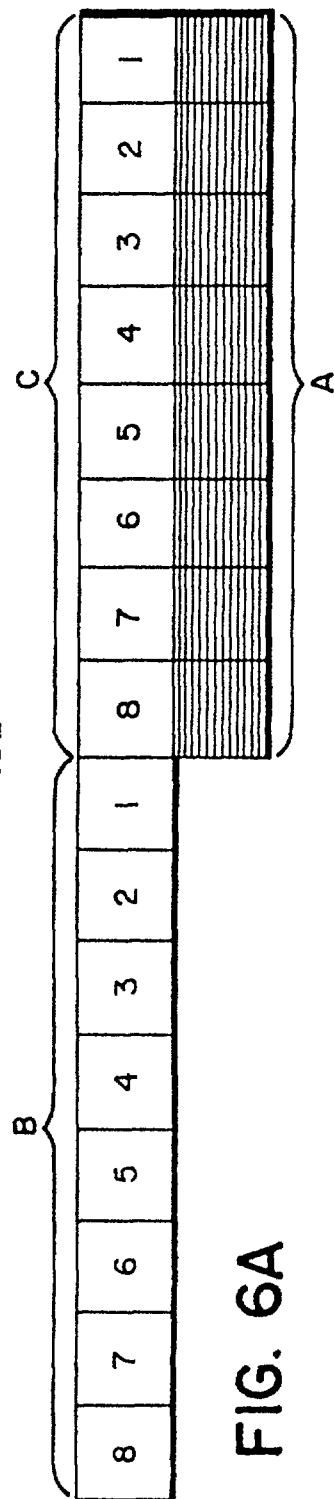


FIG. 6A

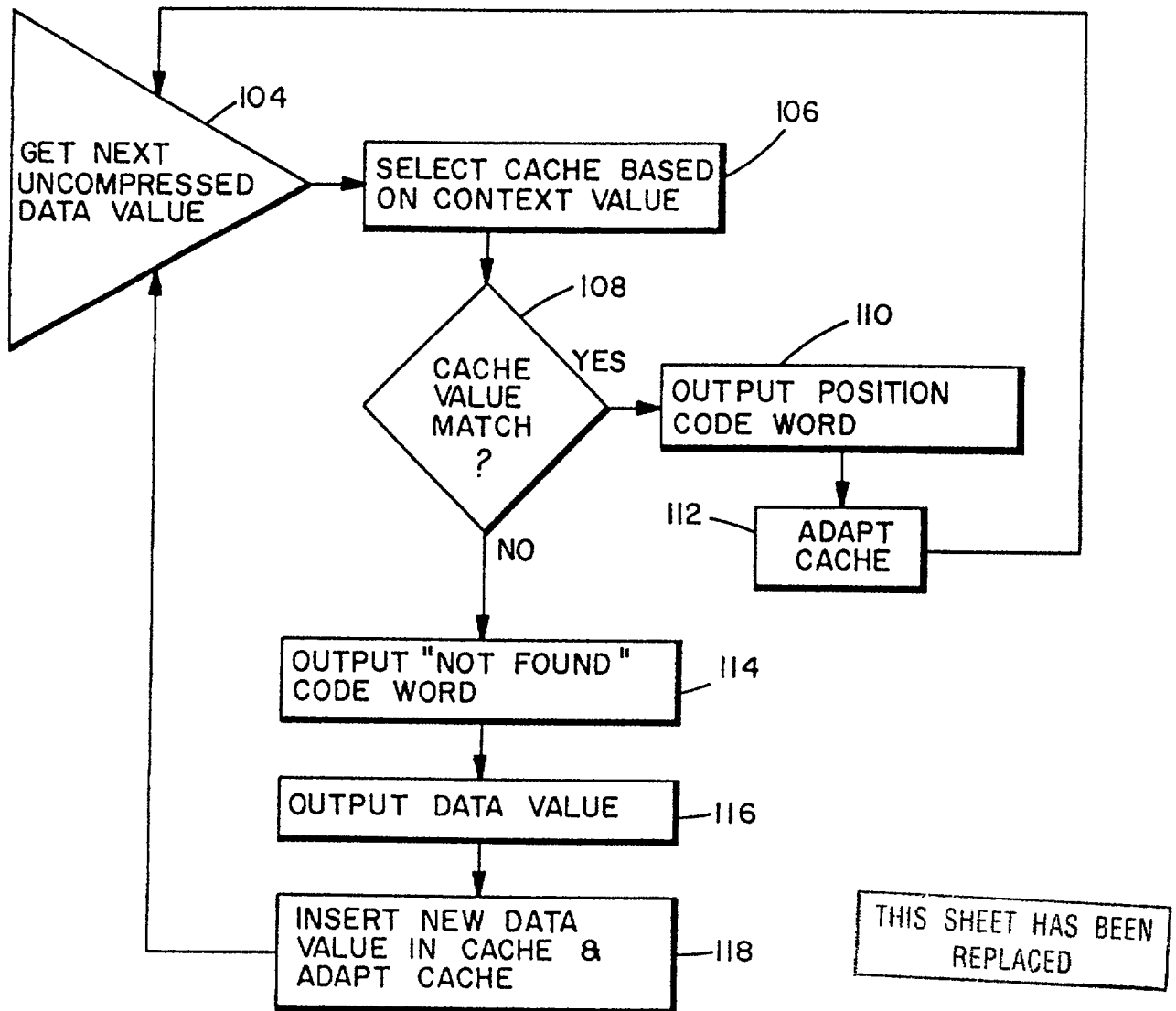


FIG. 7



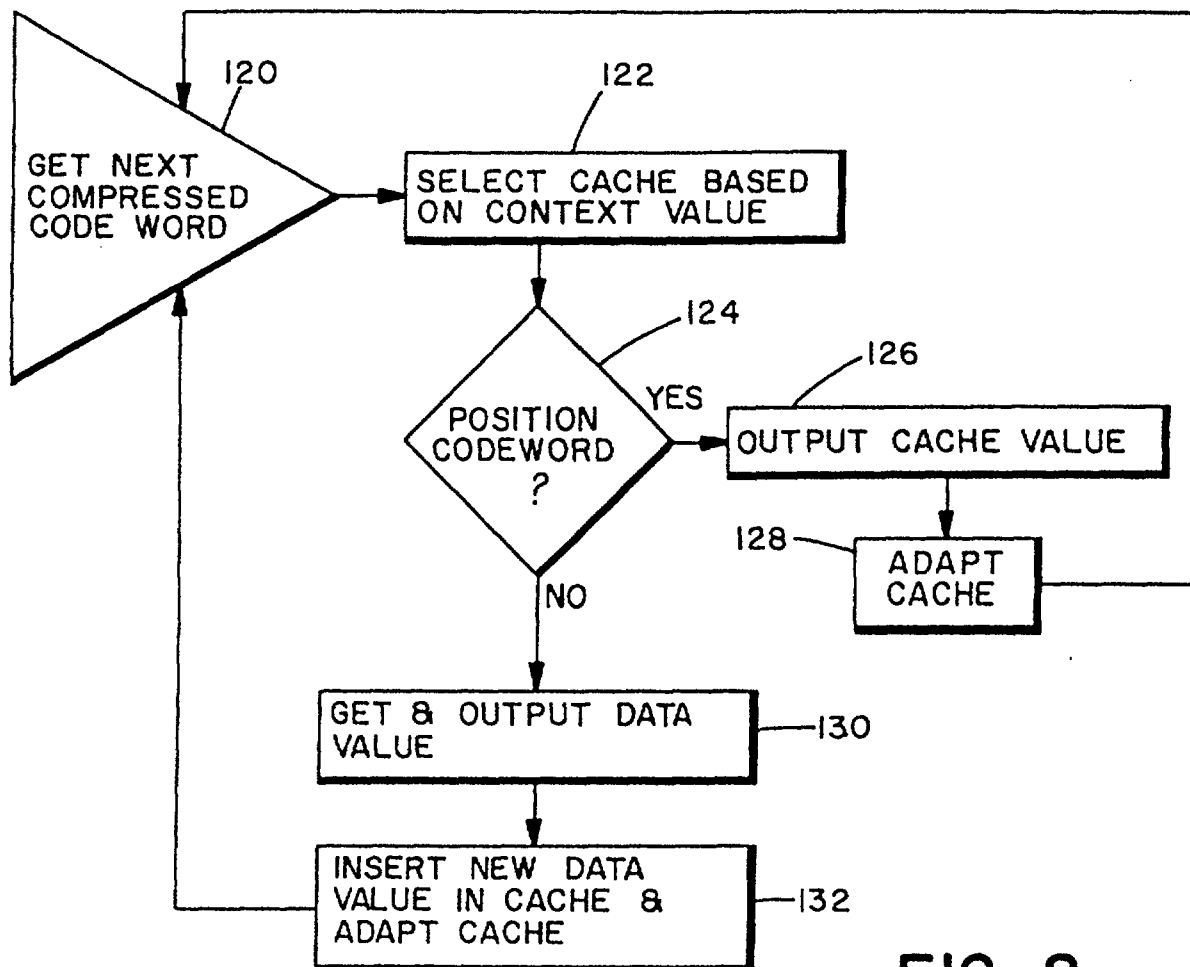


FIG. 8

THIS SHEET HAS BEEN REPLACED

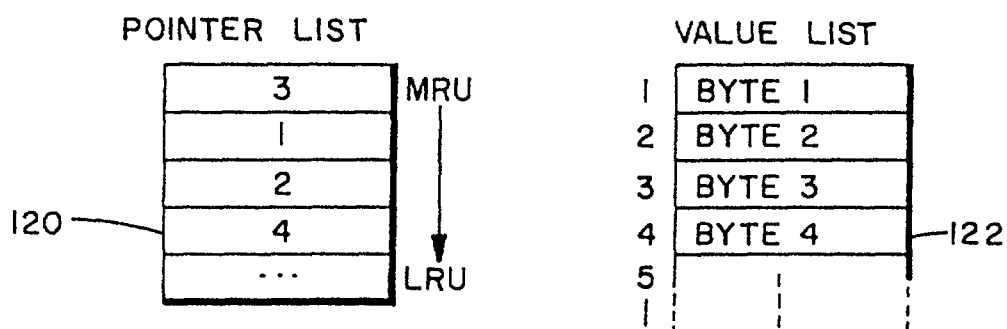
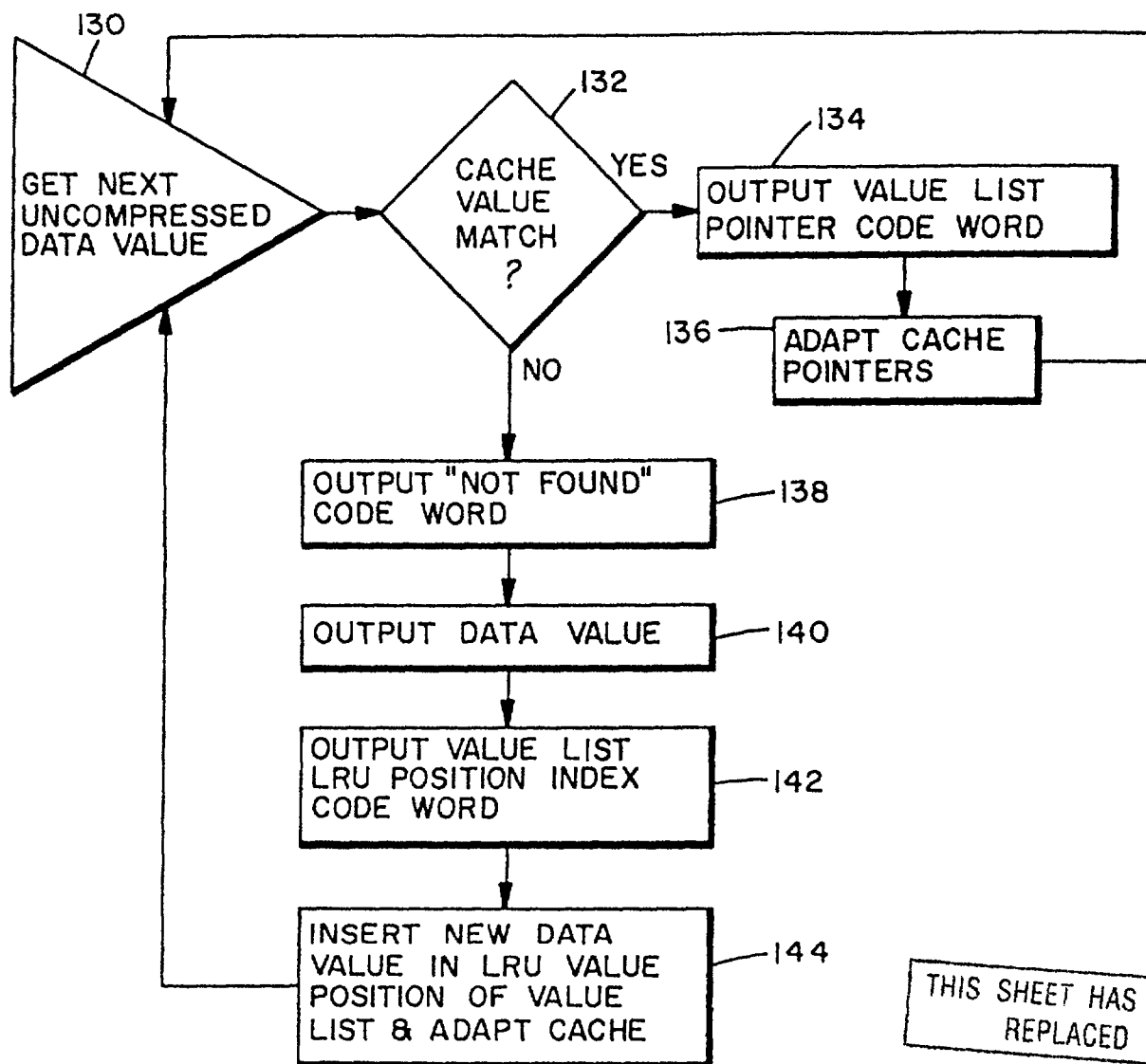


FIG. 9

This Sheet Has Been Replaced





THIS SHEET HAS BEEN REPLACED

FIG. 10



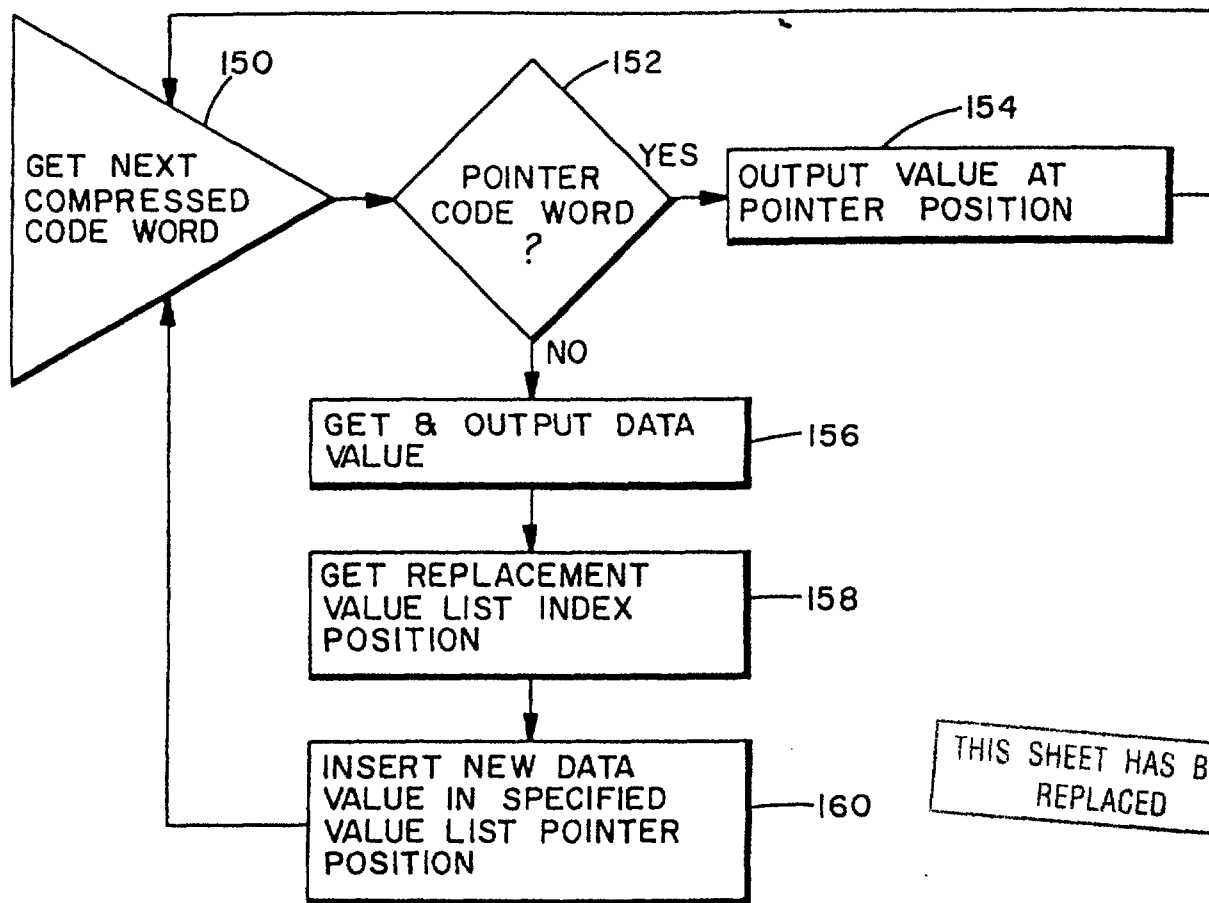


FIG. 11

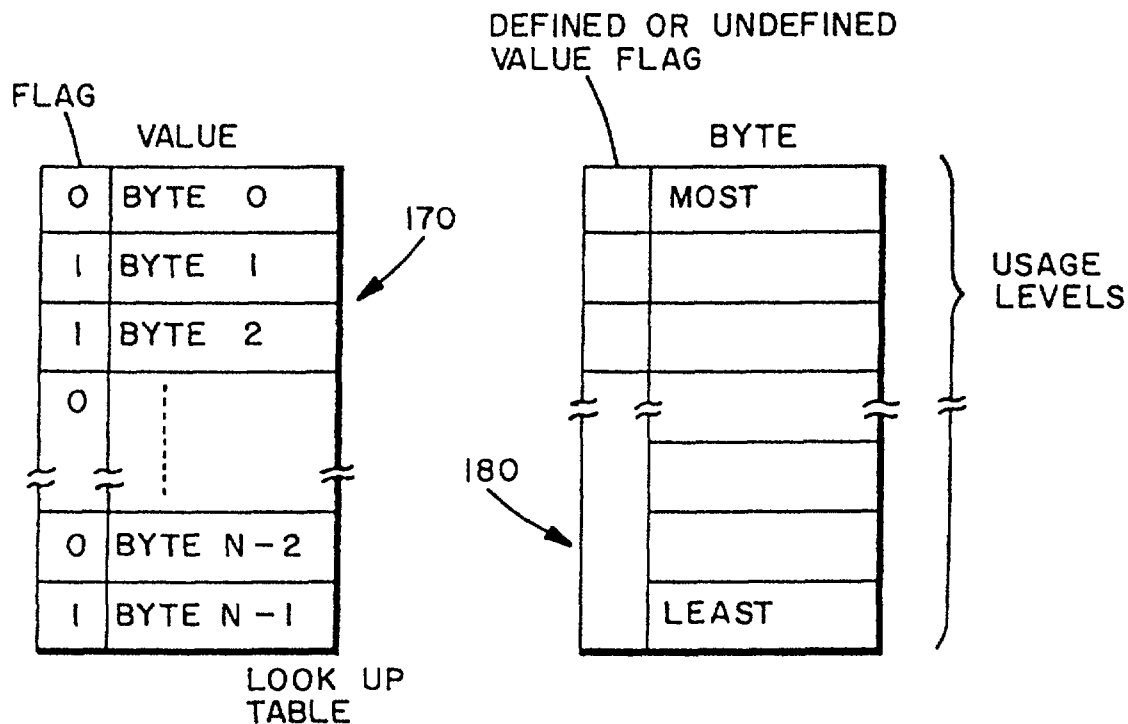


FIG. 12

FIG. 13

This Sheet Has Been Replaced



ABSTRACT

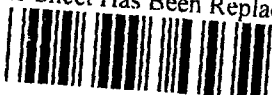
An apparatus (60) for compressing bilevel data includes a first cache memory (20,66) having a plurality of assigned levels of usage, a first usage level assigned to most recently used data segments and a second level assigned to a plurality of less recently used data segments. A processor (62) determines if a received data segment is found in the cache memory (20,66) and, if not, it assigns the received data segment to the cache memory's first level in place of a previous data segment stored therein. The previous data segment is assigned to a position in the second level in place of a less recently used data segment. The less recently used data segment that is displaced is chosen by a pseudo-random method. A "not-found" indication (38,62) is then transmitted to a receiving station along with the identity of the received data segment. The receiving station (72) contains identical cache structures and updates its caches in response to received code words and data segments. A further version of the invention emplaces data segments into cache memories in accordance with the context in which the data segments are found, the cache memories being assigned in accordance with context derived addresses.

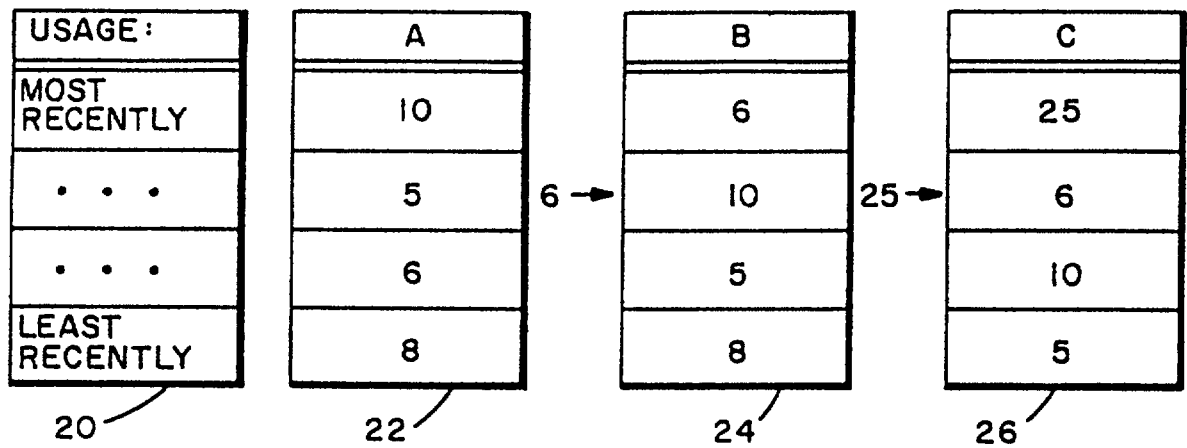
E:\SPEC\DRFT\HPSPEC.EPO

(Fig. 4)

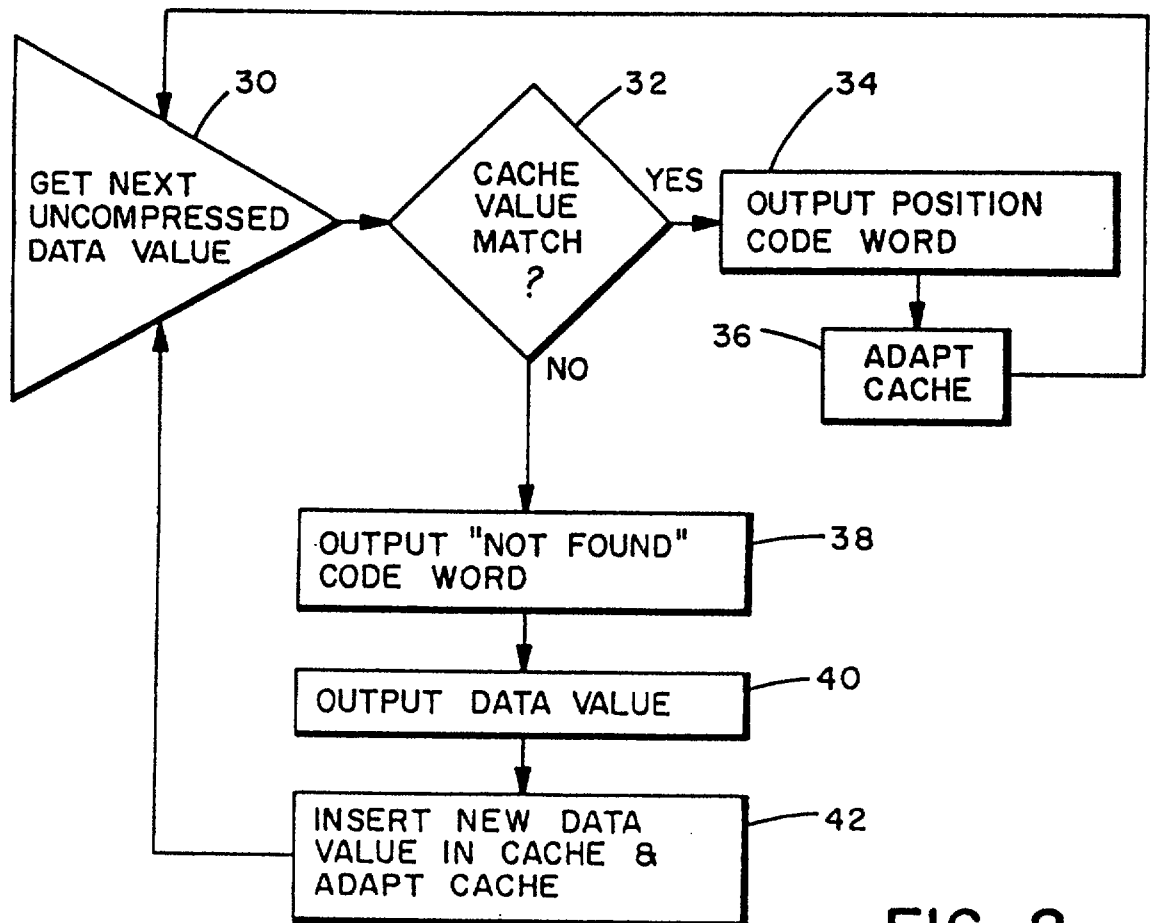
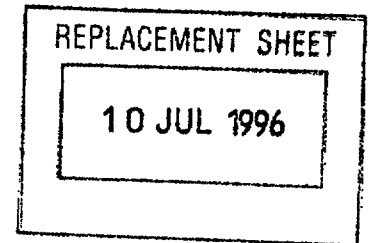
THIS SHEET HAS BEEN
REPLACED

This Sheet Has Been Replaced

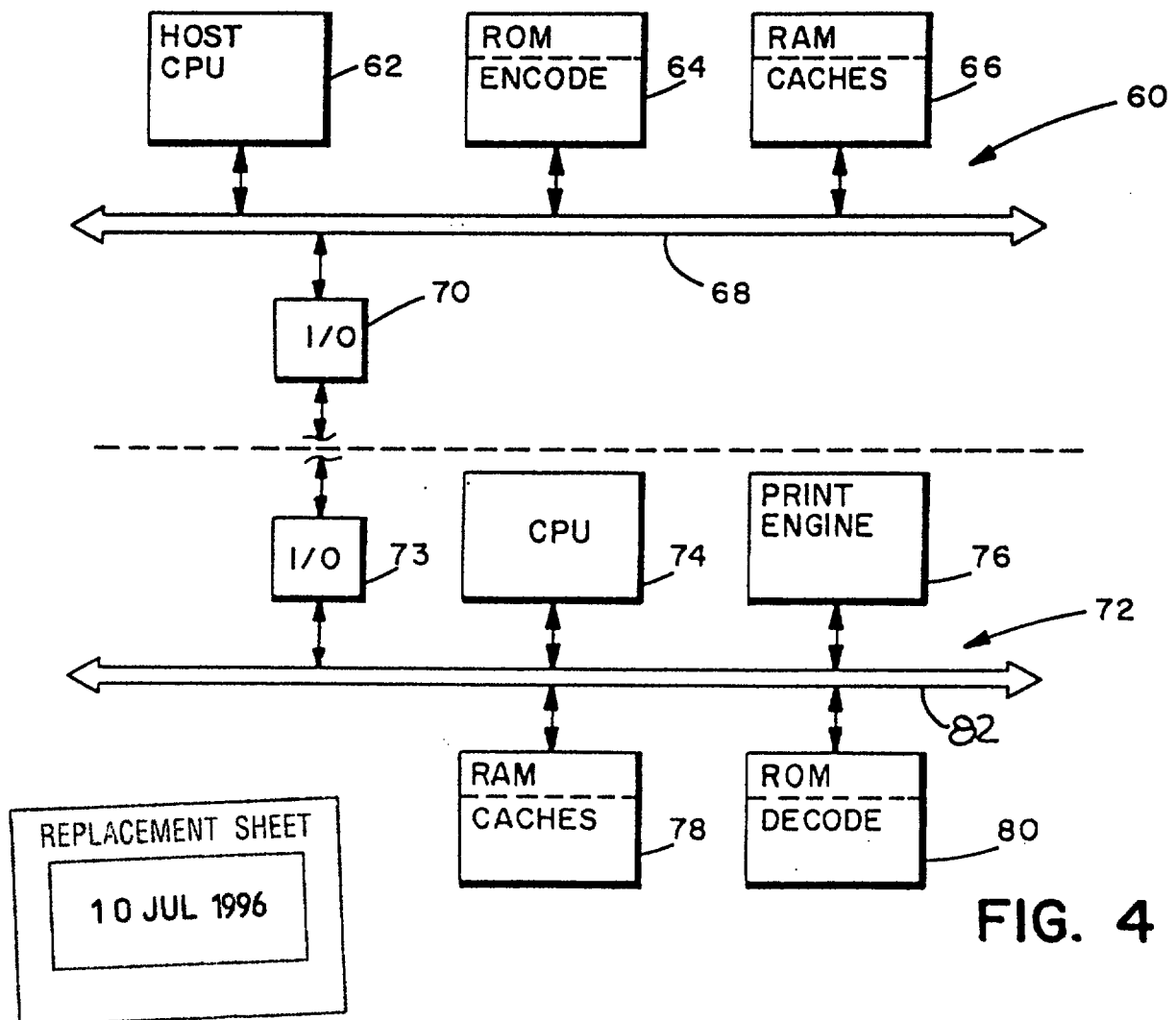
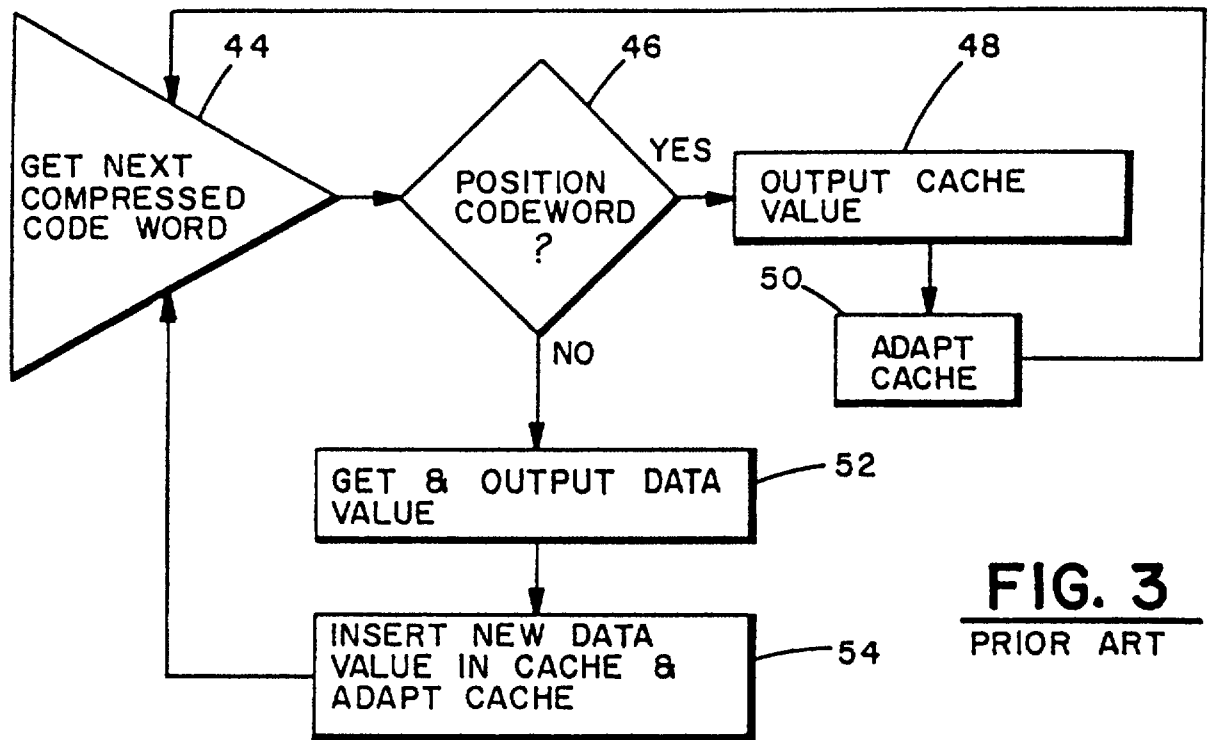


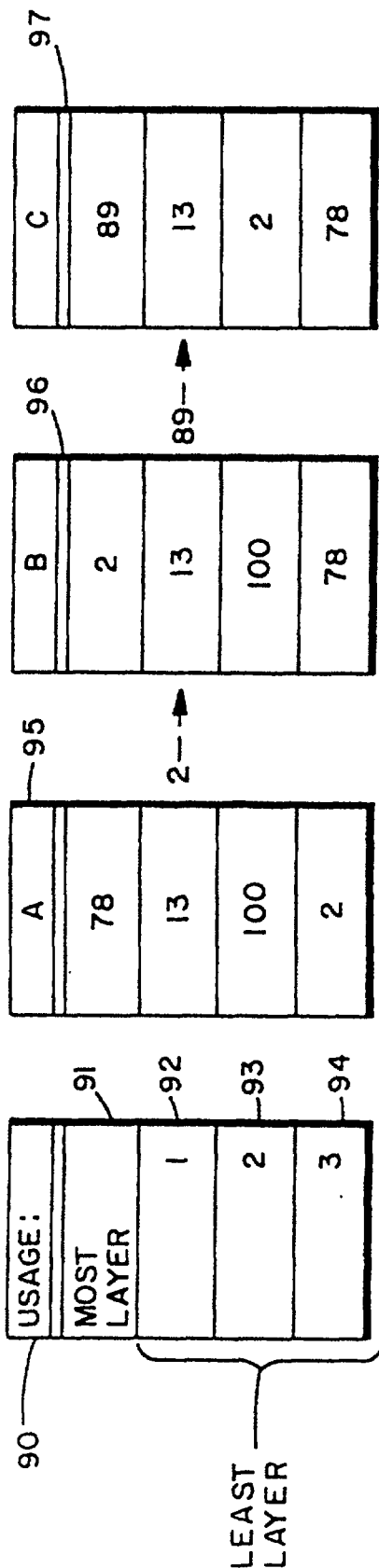
**FIG. 1**

PRIOR ART

**FIG. 2**

PRIOR ART





4/7

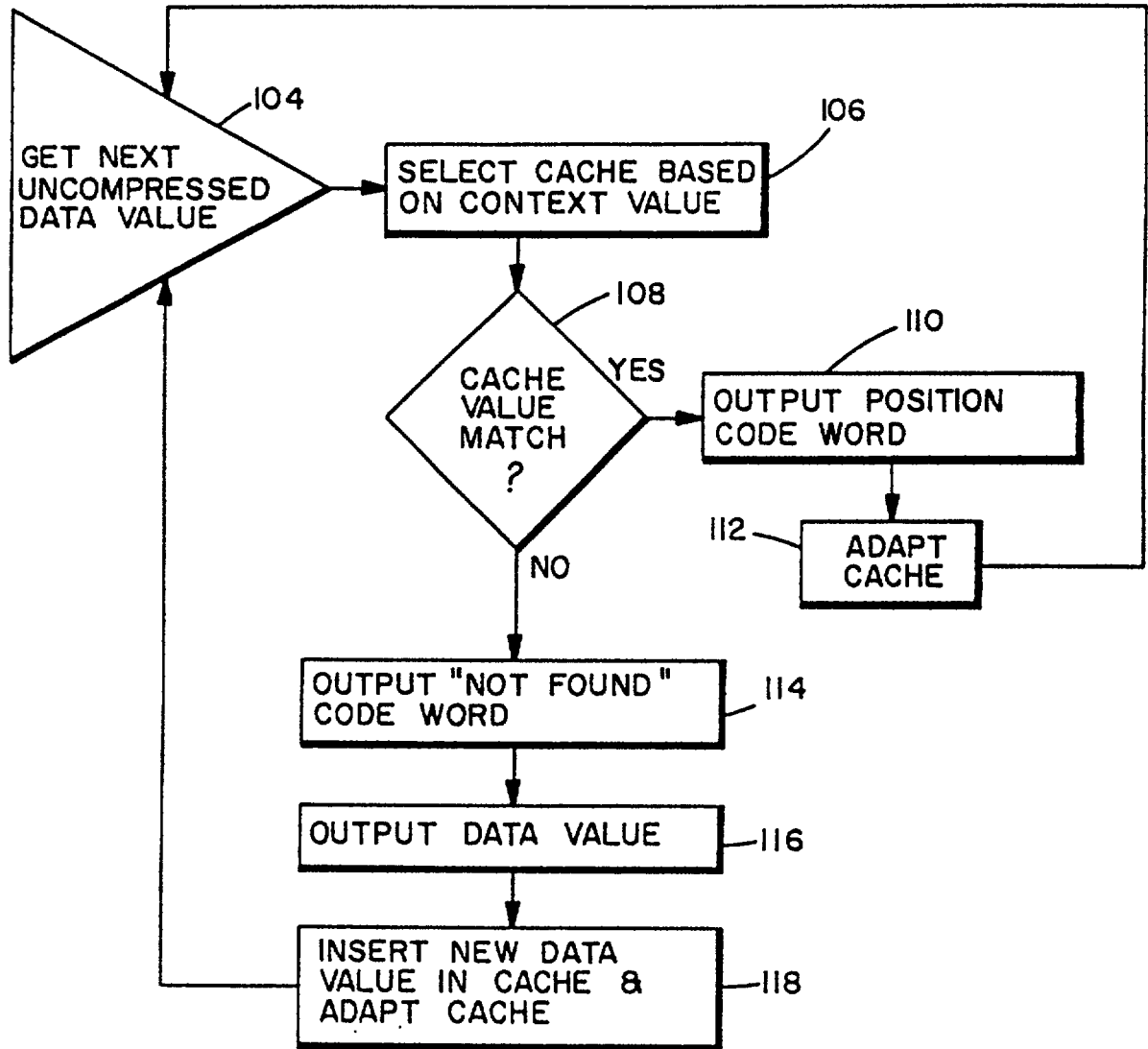
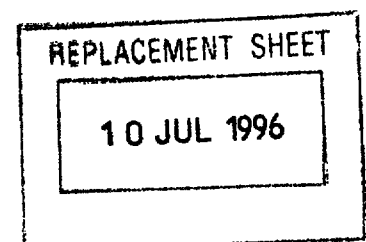


FIG. 7



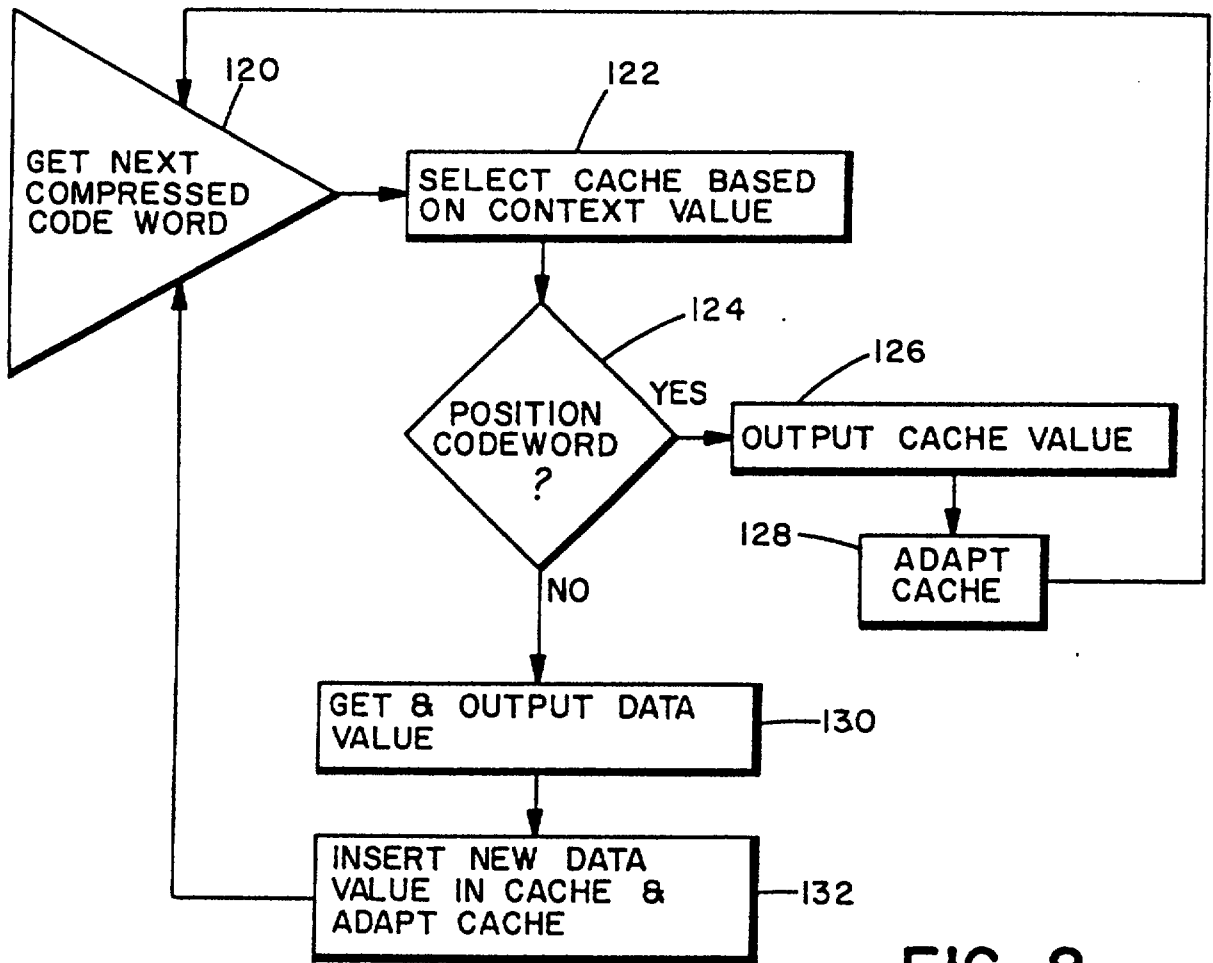


FIG. 8

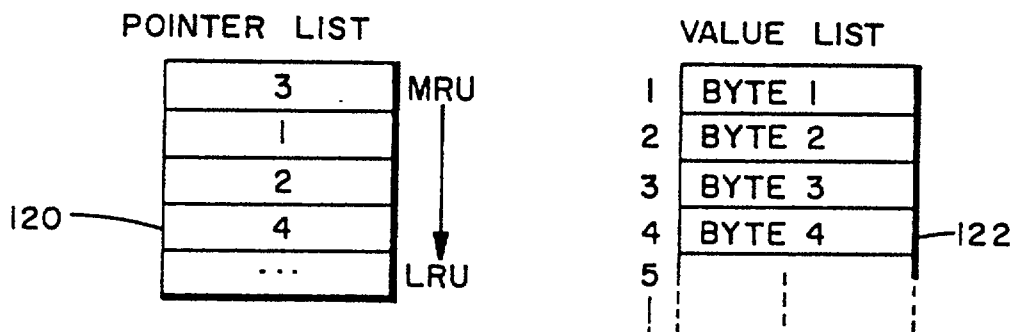
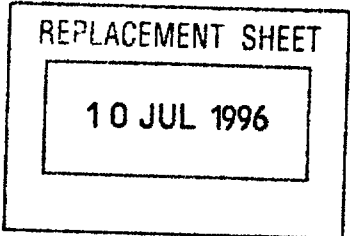


FIG. 9



6/7

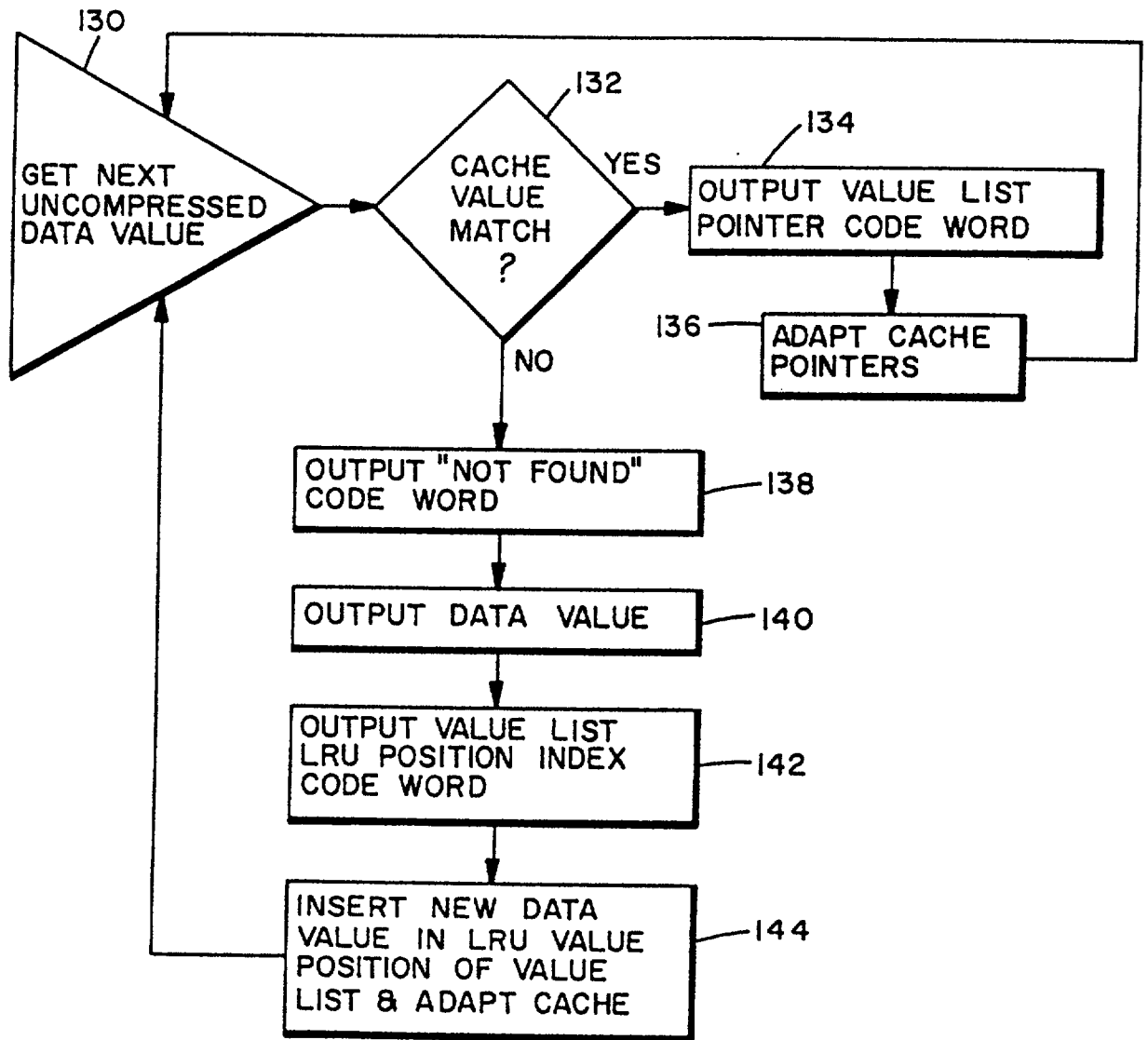


FIG. 10

REPLACEMENT SHEET

10 JUL 1996

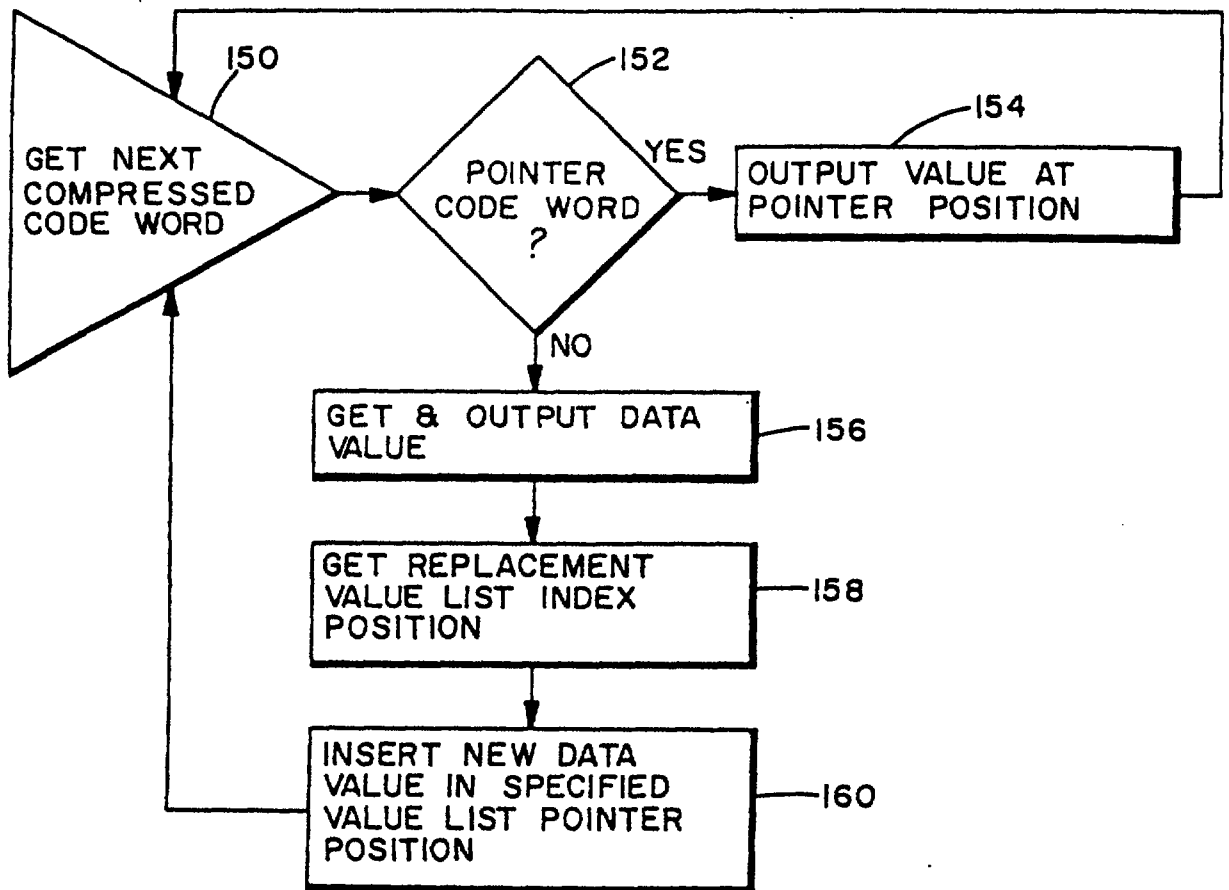
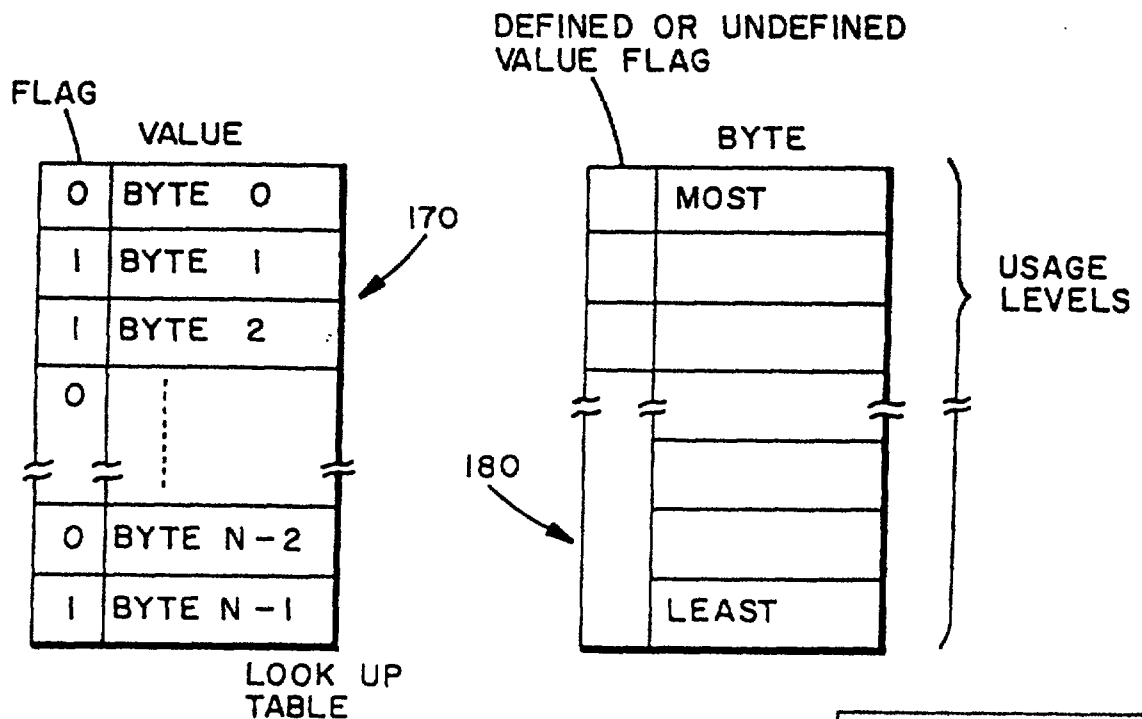


FIG. II



REPLACEMENT SHEET

10 JUL 1996

REPLACEMENT SHEET

10 JUL 1996

ABSTRACT

CACHE-BASED COMPRESSION/DECOMPRESSION

An apparatus (60) for compressing bilevel data includes a first cache memory (20,66) having a plurality of assigned levels of usage, a first usage level assigned to most recently used data segments and a second level assigned to a plurality of less recently used data segments. A processor (62) determines if a received data segment is found in the cache memory (20,66) and, if not, it assigns the received data segment to the cache memory's first level in place of a previous data segment stored therein. The previous data segment is assigned to a position in the second level in place of a less recently used data segment. The less recently used data segment that is displaced is chosen by a pseudo-random method. A "not-found" indication (38,62) is then transmitted to a receiving station along with the identity of the received data segment. The receiving station (72) contains identical cache structures and updates its caches in response to received code words and data segments. A further version of the invention emplaces data segments into cache memories in accordance with the context in which the data segments are found, the cache memories being assigned in accordance with context derived addresses.

(Figure 4)