

# Parallel Lempel Ziv Coding (Extended Abstract)

Shmuel Tomi Klein<sup>1</sup> and Yair Wiseman<sup>2</sup>

<sup>1</sup> Dept. of Math. & CS, Bar Ilan University  
Ramat-Gan 52900, Israel  
tomi@cs.biu.ac.il

<sup>2</sup> Dept. of Math. & CS, Bar Ilan University  
Jerusalem College of Technology  
wiseman@cs.biu.ac.il

**Abstract.** We explore the possibility of using multiple processors to improve the encoding and decoding tasks of Lempel Ziv schemes. A new layout of the processors is suggested and it is shown how LZSS and LZW can be adapted to take advantage of such parallel architectures. Experimental results show an improvement in compression and time over standard methods.

## 1 Introduction

Compression methods are often partitioned into static and dynamic methods. The *static* methods assume that the file to be compressed has been generated according to a certain model which is fixed in advance and known to both compressor and decompressor. The model could be based on the probability distribution of the different characters or more generally of certain variable length substrings that appear in the file, combined with a procedure to parse the file into a well determined sequence of such elements. The encoded file can then be obtained by applying some statistical encoding function, such as Huffman or arithmetic coding. Information about the model is either assumed to be known (such as the distribution of characters in English text), or may be gathered in a first pass over the file, so that the compression process may only be performed in a second pass.

Many popular compression methods, however, are *adaptive* in nature. The underlying model is not assumed to be known, but discovered during the sequential processing of the file. The encoding and decoding of the  $i$ -th element is based on the distribution of the  $i - 1$  preceding ones, so that compressor and decompressor can work in synchronization without requiring the transmittal of the model itself. Examples of adaptive methods are the Lempel-Ziv (LZ) methods and their variants, but there are also adaptive versions of Huffman and arithmetic coding.

We wish to explore the possibility of using multiple processors to improve the encoding and decoding tasks. In [7] this has been done for static Huffman coding,

focusing in particular on the decoding process. The current work investigates how parallel processing could be made profitable for Lempel Ziv coding.

Previous work on parallelizing compression includes [1,2,3], which deal with LZ compression, [5], relating to Huffman and arithmetic coding, and [4]. A parallel method for the construction of Huffman trees can be found in [8]. Our work concentrates on LZ methods, in particular a variant of LZ77 [13] known as LZSS, and a variant of LZ78 [14] known as LZW. In LZSS [9], the encoded file consists of a sequence of items each of which is either a single character, or a pointer of the form (*off*, *len*) which replaces a string of length *len* that appeared *off* characters earlier in the file. Decoding of such a file is thus a very simple procedure, but for the encoding there is a need to locate longest reoccurring strings, for which sophisticated data structures like hash tables or binary trees have been suggested. In LZW [10], the encoded file consists of a sequence of pointers to a *dictionary*, each pointer replacing a string of the input file that appeared earlier and has been put into the dictionary. Encoder and decoder must therefore construct identical copies of the dictionary.

The basic idea of parallel coding is partitioning the input file of size  $N$  into  $n$  blocks of size  $N/n$  and assigning each block to one of the  $n$  available processors. For static methods the encoding is then straightforward, but for the decoding, it is the compressed file that is partitioned into equi-sized blocks, so there might be a problem of synchronization at the block boundaries. This problem may be overcome by inserting dummy bits to align the block boundaries with codeword boundaries, which causes a negligible overhead if the block size is large enough. Alternatively, in the case of static Huffman codes, one may exploit their tendency to resynchronize quickly after an error, to devise a parallel decoding procedure in which each processor decodes one block, but is allowed to overflow into one or more following blocks until synchronization is reached [7].

For dynamic methods one is faced with the additional problem that the encoding and decoding of elements in the  $i$ -th block may depend on elements of some previous blocks. Even if one assumes a CREW architecture, in which all the processors share some common memory space which can be accessed in parallel, this would still be essentially equivalent to a sequential model. This is so because elements dealt with by processor  $i$  at the beginning of block  $i$  may rely upon elements at the end of block  $i - 1$  which have not been processed yet by processor  $i - 1$ ; thus processor  $i$  can in fact start its work only after processor  $i - 1$  has terminated its own.

The easiest way to implement parallelization in spite of the above problem is to let each processor work independently of the others. The file is thus partitioned into  $n$  blocks which are encoded and decoded without any transfer of data between the processors. If the block size is large enough, this solution may even be recommendable: most LZ methods put a bound on the size of the history taken into account for the current item, and empirical tests show that the additional compression, obtained by increasing this history beyond some reasonable size, rapidly tends to zero. The cost of parallelization would therefore be a small deterioration in compression performance at the block boundaries, since each

processor has to “learn” the main features of the file on its own, but this loss will often be tolerated as it may allow to cut the processing time by a factor of  $n$ . In [6] the authors suggest letting each processor keep the last characters of the previous block and thereby improve the encoding speed, but each block must then be larger than the size of the history window. On the other hand, putting a lower bound on the size  $N/n$  of each block effectively puts an upper bound on the number of processors  $n$  which can be used for a given file of size  $N$ , so we might not fully take advantage of all the available computing power.

We therefore turn to the question how to use  $n$  processors, even when the size of each block is not very large. In the next section we propose a new parallel coding algorithm, based on a time versus compression efficiency tradeoff which is related to the degree of parallelization. On the one extreme, for full parallelization, each of the  $n$  processors works independently, which may sharply reduce the compression gain if the size of the blocks is small. On the other extreme, all the processors may communicate, forcing delays that make this variant as time consuming as a sequential algorithm. The suggested tradeoff is based on a hierarchical structure of the connections between the processors, each of which depending at most on  $\log n$  others. The task can be performed in parallel by  $n$  processors in  $\log n$  sequential stages. There will be a deterioration in the compression ratio, but the loss will be inferior to that incurred when all  $n$  processors are independent.

In contrast to Huffman coding, for which parallel decoding could be applied regardless of whether the possibility having multiple processors at decoding time was known at the time of encoding, there is a closer connection between encoding and decoding for LZ schemes. We therefore need to deal also with the parallel encoding scheme, and we assume that the same number of processors is available for both tasks.

Note, however, that one cannot assume simultaneously equi-sized blocks for both encoding and decoding. If encoding is done with blocks of fixed size, the resulting compressed blocks are of variable lengths. So one either has to store a vector of indices to the starting point of each processor in the compressed file, which adds an unnecessary storage overhead, or one performs *a priori* the compression on blocks of varying size, such that the resulting compressed blocks are all of roughly the same size. To get blocks of exactly the same size and to achieve byte alignment, one then needs to pad each block with a small number of bits, but in this case the loss of compression due to this padding is generally negligible. Moreover, the second alternative is also the preferred choice for many specific applications. For instance, in an Information Retrieval system built on a large static database, compression is done only once, so the speedup of parallelization may not have any impact, whereas decompression of selected parts is required for each query to be processed, raising the importance of parallel decoding.

## 2 A Tree-Structured Hierarchy of Processors

The suggested form of the hierarchy is that of a full binary tree, similarly to a binary heap. This basic form has already been mentioned in [6], but the way to use it as presented here is new. The input file is partitioned into  $n$  blocks  $B_1, \dots, B_n$ , each of which is assigned to one of the available processors. Denote the  $n$  processors by  $P_1, \dots, P_n$ , and assume, for the ease of description, that  $n + 1$  is a power of 2, that is  $n = 2^k - 1$  for some  $k$ . Processor  $P_1$  is at the root of the tree and deals with the first block. As there is no need to “point into the future”, communication lines between the processors may be unidirectional, permitting a processor with higher index to access processors with lower index, but not vice versa. Restricting this to a tree layout yields a structure in which  $P_{2i}$  and  $P_{2i+1}$  can access the memory of  $P_i$ , for  $1 \leq i \leq (n - 1)/2$ . Figure 1 shows this layout for  $n = 15$ , the arrows indicating the dependencies between the processors. The numbers indicate both the indices of the blocks and of the corresponding processors.

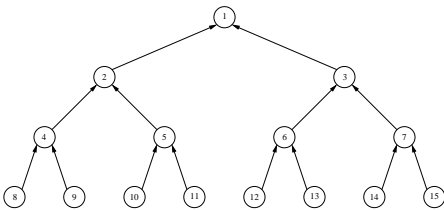


FIGURE 1: Simple tree layout.

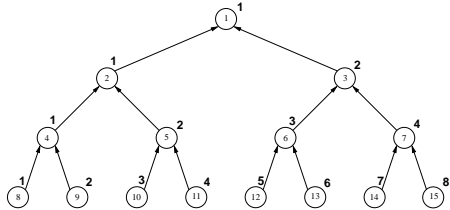


FIGURE 2: Layer-by-layer layout.

The compression procedure for LZSS works as follows:  $P_1$  starts at the beginning of block  $B_1$ , which is stored in its memory. Once this is done,  $P_2$  and  $P_3$  start simultaneously their work on  $B_2$  and  $B_3$  respectively, both searching for reoccurring strings first within the block they have been assigned to, and then extending the search back into block  $B_1$ . In general, after  $P_i$  has finished the processing of block  $B_i$ , processors  $P_{2i}$  and  $P_{2i+1}$  start scanning simultaneously their corresponding blocks. The compression of the file is thus not necessarily done layer by layer, e.g.,  $P_{12}$  and  $P_{13}$  may start compressing blocks  $B_{12}$  and  $B_{13}$ , even if  $P_5$  is not yet done with  $B_5$ .

Note that while the blocks  $B_2$  and  $B_1$  are contiguous, this is not the case for  $B_3$  and  $B_1$ , so that the *(off, len)* pairs do not necessarily point to *close* previous occurrences of a given string. This might affect compression efficiency, as one of the reasons for the good performance of LZ methods is the tendency of many files to repeat certain strings within the close vicinity of their initial occurrences. For processors and blocks with higher indices, the problem is even aggravated. The experimental section below brings empirical estimates of the resulting loss.

The layout suggested in Figure 1 is obviously wasteful, as processors of the higher layers stay idle after having compressed their assigned block. The number

of necessary processors can be reduced by half, or, which is equivalent, the block size for a given number of processors may be doubled, if one allows a processor to deal with multiple blocks. The easiest way to achieve this is displayed in Figure 2, where the numbers in the nodes are the indices of the blocks, and the boldface numbers near the nodes refer to the processors. Processors  $1, \dots, 2^m$  are assigned sequentially, from left to right, to the blocks of layer  $m, m = 0, 1, \dots, k - 1$ . This simple way of enumerating the blocks has, however, two major drawbacks: refer, e.g., to block  $B_9$  which should be compressed by processor  $P_2$ . First, it might be that  $P_1$  finishes the compression of blocks  $B_2$  and  $B_4$ , before  $P_2$  is done with  $B_3$ . This causes an unnecessary delay,  $B_9$  having to wait until  $P_2$  processes both  $B_3$  and  $B_5$ , which could be avoided if another processor would have been assigned to  $B_9$ , for example one of those that has not been used in the upper layers. Moreover, the problem is not only one of wasted time:  $P_2$  stores in its memory information about the blocks it has processed, namely  $B_3$  and  $B_5$ . But the compression of  $B_9$  does not depend on these blocks, but only on  $B_4, B_2$  and  $B_1$ . The problem thus is that the hierarchical structure of the tree is not inherited by the dependencies between the processors.

To correct this deficiency of the assignment scheme, each processor will continue working on one of the offsprings of its current block. For example, one could consistently assign a processor to the left child block of the current block, whereas the right child block is assigned to the next available newly used processor. More formally, let  $S_j^i$  be the index of the processor assigned to block  $j$  of layer  $i$ , where  $i = 0, \dots, k - 1$  and  $j = 1, \dots, 2^i$ , then  $S_1^0 = 1$  and for  $i > 0$ ,

$$S_{2j-1}^i = S_j^{i-1} \quad \text{and} \quad S_{2j}^i = 2^{i-1} + j.$$

The first layers are thus processed, from left to right, by processors with indices: (1), (1,2), (1, 3, 2, 4), (1, 5, 3, 6, 2, 7, 4, 8), etc. Figure 3(a) depicts the new layout of the blocks, the rectangles indicating the sets of blocks processed by the same processor. This structure induces a corresponding tree of processors, depicted in Figure 3(b).

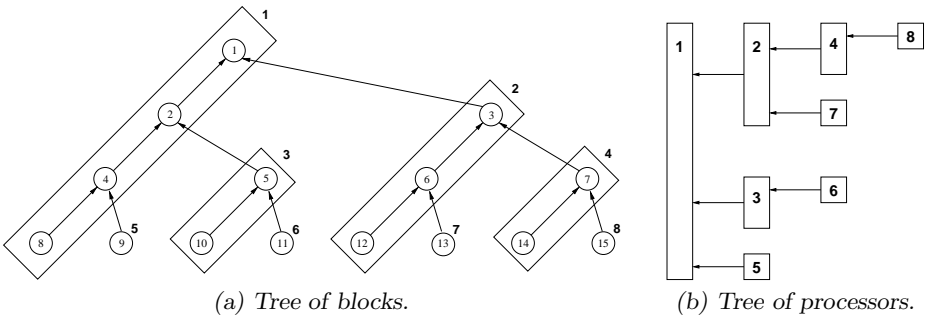


FIGURE 3: New hierarchical structure.

As a results of this method, processor  $P_i$  will start its work with block  $B_{2i-1}$ , and then continue with  $B_{4i-2}$ ,  $B_{8i-4}$ , etc. In each layer, the evenly indexed blocks inherit their processors from their parent block, and each of the oddly indexed blocks starts a new sequence of blocks with processors that have not been used before.

The memory requirements of the processors have also increased by this new scheme, and space for the data of up to  $\log_2 n$  blocks have to be stored. However, most of the processors deal only with a few blocks, and the average number of blocks to be memorized, when amortized over the  $n$  processors is

$$\frac{1}{n} \sum_{i=1}^{\log_2 n} i \frac{n}{2^i} = 2 - \frac{\log_2 n + 2}{n} < 2.$$

For the encoding and decoding procedures, we need a fast way to convert the index of a block into the index of the corresponding processor, i.e., a function  $f$ , such that  $f(i) = j$  if block  $B_i$  is coded by processor  $P_j$ . Define  $r(i)$  as the largest power of 2 that divides the integer  $i$ , that is,  $r(i)$  is the length of the longest suffix consisting only of zeros of the binary representation of  $i$ .

CLAIM: 
$$f(i) = \frac{1}{2} \left( \frac{i}{2^{r(i)}} + 1 \right).$$

PROOF: By induction on  $i$ . For  $i = 1$ , we get  $f(1) = 1$ , which is correct. Assume the claim is true up to  $i - 1$ . If  $i$  is odd,  $r(i) = 0$  and the formula gives  $f(i) = (i + 1)/2$ . As has been mentioned above, any oddly indexed block is the starting point of a new processor and indeed processor  $P_{(i+1)/2}$  starts at block  $B_i$ . If  $i$  is even, block  $B_i$  is coded by the same processor as its parent block  $B_{i/2}$ , for which the inductive assumption applies, and we get

$$f(i) = f(i/2) = \frac{1}{2} \left( \frac{i/2}{2^{r(i/2)}} + 1 \right) = \frac{1}{2} \left( \frac{i}{2 \cdot 2^{r(i)-1}} + 1 \right) = \frac{1}{2} \left( \frac{i}{2^{r(i)}} + 1 \right),$$

so that the formula holds also for  $i$ . ■

## 2.1 Parallel Coding for LZSS

We now turn to the implementation details of the encoding and decoding procedures for LZSS. Since the coding is done by stages, the parallel co-routines will invoke themselves the depending offsprings. For the encoding, the procedure  $PLZSS-encode(i, j)$  given in Figure 4 will process block  $B_i$  with processor  $P_j$ , where  $j = f(i)$ . The whole process is initialized by a call to  $PLZSS-encode(1, 1)$  from the main program.

Each routine starts by copying the text of the current block into the memory of the processor, possibly adding to texts of previous blocks that have been stored there. As in the original LZSS, the longest substring in the history is sought that matches the suffix of the block starting at the current position. The search for this substring can be accelerated by several techniques, and one of the fastest is by use of a hash table [12]. The longest substring is then replaced by

```

PLZSS-encode( $i, j$ )
{
  append text of  $B_i$  to memory of  $P_j$ 
   $cur \leftarrow 1$ 
  while  $cur < |B_i|$ 
  {
     $S \leftarrow$  suffix of  $B_i$  starting at  $cur$ 
     $ind \leftarrow i$ 
    while  $ind > 0$ 
    {
      access memory of  $P_{f(ind)}$  and
      record occurrences in  $B_{ind}$  matching a prefix of  $S$ 
       $ind \leftarrow \lfloor ind/2 \rfloor$ 
    }
    if longest occurrence not long enough
    {
      encode single character    $cur \leftarrow cur + 1$ 
    }
    else
    {
      encode as ( $off, len$ )    $cur \leftarrow cur + len$ 
    }
  }
  perform in parallel {
    if  $2i \leq n$    PLZSS-encode( $2i, j$ )
    if  $2i + 1 \leq n$  PLZSS-encode( $2i + 1, i + 1$ )
  }
}

```

FIGURE 4: Parallel LZSS encoding for block  $B_i$  by processor  $P_j$ .

a pair (*offset, length*), unless *length* is too small (2 or 3 in implementations of [12], such as the patent [11], which is the basis of Microsoft's DoubleSpace), in which case a single character is sent to output and the window is shifted by one.

In our case, the search is not limited to the current block, but extends backwards to the parent blocks in the hierarchy, up to the root. For example, referring to Figure 3, the encoding of block  $B_{13}$  will search also through  $B_6$ ,  $B_3$  and  $B_1$ , and thus access the memory of the processors  $P_7$ ,  $P_2$ ,  $P_2$  and  $P_1$ , respectively. Note that the size of the history window is usually limited by some constant  $W$ . We do not impose any such limit, but in fact, the encoding of any element is based on a history of size at most  $\log_2 n \times$  the block size.

For the decoding, recall that we assume that the encoded blocks are of equal size *Blocksize*. The decoding routine can thus address earlier locations as if the blocks, that are ancestors of the current block in the tree layout, were stored contiguously. Any element of the form (*off, len*) in block  $B_i$  can point back into a block  $B_j$ , with  $j = \lfloor i/2^b \rfloor$  for  $b = 0, 1, \dots, \lfloor \log_2 i \rfloor$ , and the index of this block can be calculated by

$$b \leftarrow \lceil (off - cur + 1) / Blocksize \rceil,$$

where *cur* is the index of the current position in block  $B_i$ . The formal decoding procedure is given in Figure 5.

```

PLZSS-decode( $i, j$ )
{
     $cur \leftarrow 1$ 
    while there are more items to decode
    {
        if next item is a character
        {
            store the character     $cur \leftarrow cur + 1$  }
        else // the item is ( $off, len$ )
        {
            if  $off < cur$  // pointer within block  $B_i$ 
                copy  $len$  characters, starting at position  $cur - off$ 
            else // pointer to earlier block
            {
                 $b \leftarrow \lceil (off - cur + 1) / Blocksize \rceil$ 
                 $t \leftarrow (off - cur) \bmod Blocksize$ 
                copy  $len$  characters, starting at position  $t$ 
                in block  $B_{\lfloor i/2^b \rfloor}$  which is stored in  $P_{f(\lfloor i/2^b \rfloor)}$ 
            }
             $cur \leftarrow cur + len$ 
        }
    }
    perform in parallel {
        if  $2i \leq n$     PLZSS-decode( $2i, j$ )
        if  $2i + 1 \leq n$  PLZSS-decode( $2i + 1, i + 1$ )
    }
}

```

FIGURE 5: Parallel LZSS decoding for block  $B_i$  on processor  $P_j$ .

The input of the decoding routine is supposed to be a file consisting of a sequence of items, each being either a single character or a pointer of the form ( $off, len$ );  $cur$  is the current index in the reconstructed text file.

## 2.2 Parallel Coding for LZW

Encoding and decoding for LZW is similar to that of LZSS, with a few differences. While for LZSS, the “dictionary” of previously encountered strings is in fact the text itself, LZW builds a continuously growing table *Table*, which need not be transmitted, as it is synchronously reconstructed by the decoder. The table is initialized to include the set of single characters composing the text, which is often assumed to be ASCII. If, as above, we denote by  $S$  the suffix of the text in block  $B_i$  starting at the current position, then the next encoded element will be the index of the longest prefix  $R$  of  $S$  for which  $R \in Table$ , and the next element to be adjoined to *Table* will be the shortest prefix  $R'$  of  $S$  for which  $R' \notin Table$ ;  $R$  is a prefix of  $R'$  and  $R'$  extends  $R$  by one additional character.

During the encoding process of  $B_i$ , one therefore needs to access the tables in  $B_i$  itself and in the blocks which are ancestors of  $B_i$  in the tree layout, but the order of access has to be top down rather than bottom up as for LZSS. For each  $i$ , we therefore need a list  $list_i$  of the indices of the blocks accessed on the way from the root to block  $B_i$ , that is,  $list_i[ind]$  is the number whose binary



representation is given by the *ind* leftmost bits of the binary representation of  $i$ . For example,  $list_{13} = [1, 3, 6, 13]$ .

To encode a new element  $P$ , it is first searched for in *Table* of  $B_1$ , and if not found there, then in *Table* of  $B_{list_i[2]}$ , which is stored in the memory of processor  $P_{f(list_i[2])}$ , etc. However, storing only the elements in the tables may lead to errors. To illustrate this, consider the following example, referring again to Figure 3.

Suppose that the longest prefix of the string **abcde** appearing in the *Table* of  $B_1$  is **abc**. Suppose we later encounter **abcd** in the text of block  $B_2$ . The string **abcd** will thus be adjoined to the same *Table*, since both  $B_1$  and  $B_2$  are processed by the same processor  $P_1$ . Assume now that the texts of both blocks  $B_5$  and  $B_3$  start with **abcde**. While for  $B_5$  it is correct to store **abcde** as the first element in its *Table*, the first element to be stored in the *Table* of  $B_3$  should be **abcd**, since the **abcd** in the memory of  $P_1$  was generated by block  $B_2$ , whereas  $B_3$  only depends on  $B_1$ .

```

PLZW-encode( $i, j$ )
{
   $\omega \leftarrow B_i[1]$ 
   $cur \leftarrow 2$ 
  while  $cur \leq |B_i|$ 
  {
     $ind \leftarrow 1$ 
    while  $list_i[ind] \leq i$ 
    {
      while  $cur < |B_i|$  and
              $(\omega B_i[cur], ind) \in Table$  stored in  $P_{f(list_i[ind])}$ 
      {
         $\omega \leftarrow \omega B_i[cur]$ 
         $cur \leftarrow cur + 1$ 
      }
       $last \leftarrow ind$ 
       $ind \leftarrow ind + 1$ 
    }
     $indx \leftarrow index(\omega)$  in Table of  $P_{f(list_i[last])}$ 
    store  $(indx, last)$  in memory of  $P_j$ 
    store  $(\omega B_i[cur], ind)$  in Table in memory of  $P_j$ 
     $\omega \leftarrow B_i[cur]$ 
     $cur \leftarrow cur + 1$ 
  }
}
perform in parallel {
  if  $2i \leq n$     PLZW-encode( $2i, j$ )
  if  $2i + 1 \leq n$  PLZW-encode( $2i + 1, i + 1$ )
}

```

FIGURE 6: Parallel LZW encoding for block  $B_i$  on processor  $P_j$ .

To avoid such errors, we need a kind of a “time stamp”, indicating at what stage an element has been added to a *Table*. If the elements are stored sequentially in these tables, one only needs to record the indices of the last element for each block. But implementations of LZW generally use hashing to maintain the tables, so one cannot rely on deducing information from its physical location, and each element has to be marked individually. The easiest way is to store with each string  $P$  also the index  $i$  of the block which caused the addition of  $P$ . This would require  $\log_2 n$  bits for each entry. One can however take advantage of the fact that the elements stored by different blocks  $B_i$  in the memory of a given processor correspond to different indices  $ind$  in the corresponding lists  $list_i$ . It thus suffices to store with each element the index in  $list_i$  rather than  $i$  itself, so that only  $\log_2 \log_2 n$  bits are needed for each entry. The formal encoding and decoding procedures are given in Figures 6 and 7, respectively.

The parallel LZW encoding refers to the characters in the input block as belonging to a vector  $B_i[cur]$ , with  $cur$  giving the current index. If  $x$  and  $y$  are strings, then  $xy$  denotes their concatenation. As explained above, since the *Table* corresponding to block  $B_i$  is stored in the memory of a processor which is also accessed by other blocks, each element stored in the *Table* needs an identifier indicating the block from which it has been generated. The elements in the *Table* are therefore of the form  $(string, identifier)$ .

The output of LZW encoding is a sequence of pointers, which are the indices of the encoded elements in the *Table*. In our case, these pointers are of the form  $(index, identifier)$ . There is, however, no deterioration in the compression efficiency, as the additional bits needed for the identifier are saved in the representation of the index, which addresses a smaller range.

For simplicity, we do not go into details of handling the incremental encoding of the indices, and overflow conditions when the *Table* gets full. It can be done as for the serial LZW.

The parallel LZW decode routine assumes that its input is a sequence of elements of the form  $(index, identifier)$ . The empty string is denoted by  $\Lambda$ .

The algorithm in Figure 7 is a simplified version of the decoding, which does not work in case the current element to be decoded was the last one to be added to the *Table*. This is also a problem in the original LZW decoding and can be solved here in the same way. The details have been omitted to keep the emphasis on the parallelization.

### 3 Experimental Results

We now report on some experiments on files in different languages: the Bible (King James Version) in English, the Bible in Hebrew and the *Dictionnaire philosophique* of Voltaire in French. Table 1 first brings the sizes of the files in MB and to what size they can be reduced by LZSS and LZW, expressed in percent of the sizes of the original files. We consider three algorithms: the serial one, using a single processor and yielding the compressed sizes in Table 1, but being slow; a parallel algorithm we refer to as *standard*, where each block is treated

```

PLZW-decode( $i, j$ )
{
   $cur \leftarrow 1$ 
   $old \leftarrow \Lambda$ 
  while  $cur \leq$  number of items in block  $B_i$ 
  {
    ( $indx, ind$ )  $\leftarrow B_i[cur]$ 
    access  $Table$  in  $P_{f(list_i[ind])}$  at index  $indx$ 
    and send string  $str$  found there to output
    if  $old \neq \Lambda$ 
      store ( $old\ first[str], \lceil \log_2(i+1) \rceil$ ) in  $Table$  of  $P_j$ 
     $old \leftarrow str$ 
     $cur \leftarrow cur + 1$ 
  }

  perform in parallel {
    if  $2i \leq n$     PLZW-decode( $2i, j$ )
    if  $2i + 1 \leq n$  PLZW-decode( $2i + 1, i + 1$ )
  }
}

```

FIGURE 7: Parallel LZW decoding for block  $B_i$  on processor  $P_j$ .

independently of the others; and the *new* parallel algorithm presented herein, which exploits the hierarchical layout. The columns headed **Time** in Table 1 compare the new algorithm with the serial one. The time measurements were taken on a Sun 450 with four UltraSPARC-II 248 MHz processors, which allowed a layout with 7 blocks. The values are in seconds and correspond to LZW, which turned out to give better compression performance than LZSS in our case. The improvement is obviously not expected to be 4-fold, due to the overhead of the parallelization, but on the examples the time is generally cut to less than half.

TABLE 1: Size and time measurements on test files.

	Size			Time			
	Full	compressed by		compression		decompression	
		LZSS	LZW	Serial	New	Serial	New
English Bible	3.860	41.6	36.6	5.508	2.296	3.653	1.504
Hebrew Bible	1.471	51.7	44.7	2.134	0.853	1.488	0.566
Voltaire	0.529	49.0	40.6	0.770	0.380	0.456	0.310

For the compression performance, we compare the two parallel versions. Both are equivalent to the serial algorithm if the block size is chosen large enough, as in [6]. The graphs in Figure 8 show the sizes of the compressed files in MB as functions of the block size (in bytes), for both LZSS and LZW. We see that for large enough blocks (about 64K for LZSS and 128K for LZW) the loss relative to a serial algorithm with a single processor is negligible (about 1%) for both the standard and the new methods. However, when the blocks become

shorter, the compression gain in the independent model almost vanishes, whereas with the new processor layout the decrease in compression performance is much slower. For blocks as small as 128 bytes, running a standard parallel compression achieves only about 1–4% compression for LZSS and about 12–15% for LZW, while with the new layout this might be reduced by some additional 30–40%.

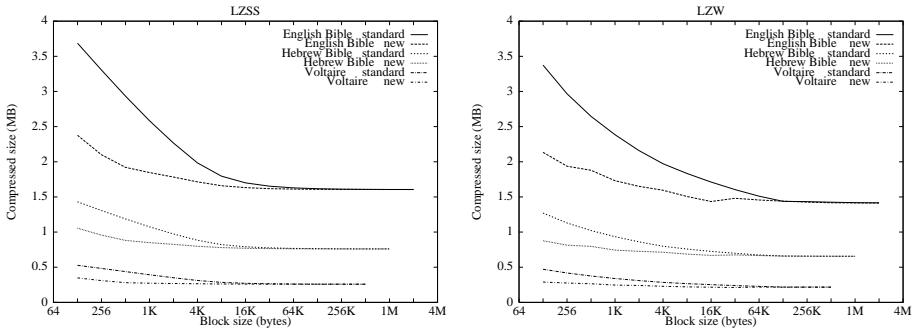


FIGURE 8: Size of compressed file as function of block size.

We conclude that the simple hierarchical layout might allow us to considerably reduce the size of the blocks that are processed in parallel without paying too high a price in compression performance. As a consequence, if a large number of processors is available, it enables a better utilization of their full combined computing power.

## References

1. DE AGOSTINO S., STORER J.A., Near Optimal Compression with Respect to a Static Dictionary on a Practical Massively Parallel Architecture, *IEEE Computer Society Press* (1995) 172–181.
2. DE AGOSTINO S., STORER J.A., Parallel Algorithms for Optimal Compression using Dictionaries with the Prefix Property, *IEEE Computer Society Press* (1992) 52–61.
3. GONZALEZ SMITH M.E., STORER J.A., Parallel Algorithms for Data Compression, *Journal of the ACM* **32**(2) (1985) 344–373.
4. HIRSCHBERG D.S., STAUFFER L.M., Parsing Algorithms for Dictionary Compression on the PRAM, *IEEE Computer Society Press* (1994) 136–145.
5. HOWARD P.G., VITTER J.S., Parallel lossless image compression using Huffman and arithmetic coding, *Proc. Data Compression Conference DCC-92*, Snowbird, Utah (1992) 299–308.
6. IWATA K., MORII M., UYEMATSU T., OKAMOTO E., A simple parallel algorithm for the Ziv-Lempel encoding, *IEICE Trans. Fundamentals* **E81-A** (1998) 709–712.
7. KLEIN S.T., WISEMAN Y., Parallel Huffman decoding, *Proc. Data Compression Conference DCC-2000*, Snowbird, Utah (2000) 383–392.

8. LAWRENCE L.L., PRZYTYCKA T.M., Constructing Huffman Trees in Parallel, *SIAM Journal of Computing* **24**(6) (1995) 1163–1169.
9. STORER J.A., SZYMANSKI, T.G., Data compression via textual substitution, *J. ACM* **29** (1982) 928–951.
10. WELCH T.A., A technique for high-performance data compression, *IEEE Computer* **17** (June 1984) 8–19.
11. WHITING D.L., GEORGE G.A., IVEY G.E., Data Compression Apparatus and Method, U.S. Patent 5,126,739 (1992).
12. WILLIAMS R.N., An extremely fast Ziv-Lempel data compression algorithm, *Proc. Data Compression Conference DCC-91*, Snowbird, Utah (1991) 362–371.
13. ZIV J., LEMPEL A., A universal algorithm for sequential data compression, *IEEE Trans. on Inf. Th.* **IT-23** (1977) 337–343.
14. ZIV J., LEMPEL A., Compression of individual sequences via variable-rate coding, *IEEE Trans. on Inf. Th.* **IT-24** (1978) 530–536.