

Dask IO for sequentially splitting, merging and resplitting multidimensional arrays

1st Timothée Guédon

Department of Computer Science and Software Engineering
Concordia University
Montreal, Quebec, Canada
t_guedon@encs.concordia.ca

2nd Tristan Glatard

Department of Computer Science and Software Engineering
Concordia University
Montreal, Quebec, Canada
tristan.glatard@concordia.ca

3rd Valérie Hayot-Sasson

Department of Computer Science and Software Engineering
Concordia University
Montreal, Quebec, Canada
valerie.hayot-sasson@concordia.ca

Abstract—In this paper, we define a new algorithm to resplit a multidimensional array stored into a set of files, and we show that it can also be used to split and merge data with the same behavior than the “multiple” strategy. We also give an implementation of this algorithm into a Python library called “Dask IO”, available as an optimization package for the Python big data library Dask.

Index Terms—multidimensional, array, split, merge, resplit, IO, processing, Dask, Python

NOMENCLATURE

- A_i : Area of the i^{th} component
- m : Amount of main memory available for the buffer at initialization
- m' : Amount of main memory available for the buffer during the algorithm execution
- $\text{iterable}[i]$: Access the i^{th} element of an iterable
- $B = (B_i, B_j, B_k)$: Shape of the buffer
- $(B_i B_j B_k)$: Size of the buffer in voxels
- R_i, R_j, R_k : Length of the reconstructed image in the i^{th} dimension
- I_i, I_j, I_k : Length of an input file in the i^{th} dimension
- O_i, O_j, O_k : Length of an output file in the i^{th} dimension
- $C_i(x), C_j(x), C_k(x)$: Overlap size in voxels in the i^{th} dimension, when loading the t^{th} buffer.
- $\Lambda_i, \Lambda_j, \Lambda_k$: Length of an input aggregate in the i^{th} dimension
- $b_i, b_j, b_k = \frac{R_i}{B_i}$: The number of buffers in the i^{th} dimension in the reconstructed image
- f_1, f_2, f_3 : Names of the overlap areas in 2D
- F_1, F_2, F_3, F_4 : Names of the overlap volumes in 3D
- N_s : Number of buffers in a slice
- n_s : Number of buffers read so far from the current slice.
- α : Number of bytes per voxel i.e. the data type of the input and output files. For the sake of simplicity we assume that the data types of all files are the same.

INTRODUCTION

A. Context

With the improvement of acquisition methods in several scientific domains like health sciences, geology and astrophysics, new big data challenges have emerged. It includes processing a big amount of data, as well as processing very big files like ultra high resolution images. An example of a ultra high resolution image in the neuroscience field is Big Brain, a model of the brain providing microscopic data (20 micrometers) for modeling and simulation [1]. Most image processing pipelines process only specific ROI (Region Of Interest) or process data by block. At the end of such pipelines, merging back those blocks into one reconstructed image may be required as well. The problem is that naive algorithms perform poorly due to the amount of seeks occurring on disk [2]. Previous work in [2] introduced two types of sequential algorithms to split and merge multidimensional arrays: the “clustered” and “multiple” strategies. To the best of our knowledge however, no algorithm has been proposed for the resplit task. Moreover, the split and merge tasks are special cases of the resplit task, which gives an opportunity to solve the three problems at once. It is easy to see that the “clustered” and “multiple” strategies are non optimal for the resplit task, that is why we searched for another strategy.

B. Problem definition

Consider a multidimensional array of shape $R = (R_i, R_j, R_k)$, stored in some input files with a given shape $I = (I_i, I_j, I_k)$, all input files having the same shape. Our goal is to optimize the process of sequentially resplitting the input files into output files with a different shape $O = (O_i, O_j, O_k)$, all output files having the same shape, too.

The resplit process has two particular cases:

- It becomes a split process if there is one input file and several output files,
- It becomes a merge process if there are several input files and one output file.

For this I/O process to be fast, one needs to minimize the number of seeks that occur on disk while reading and writing. We consider that a seek occurs either when opening a file or seeking into a file.

Let us consider a basic sequential resplit algorithm: One can repeatedly read the maximum amount of data possible from the input files into a buffer stored in main memory, and then write this buffer down into the output files requiring this data, until all output files have been completely written. This resplit algorithm is described in Algorithm 1.

The algorithm described in Algorithm 1 takes a list of input and output files *inFiles* and *outFiles* as parameters, as well as the amount of memory *m* available in RAM for the buffer and the list of the buffers' coordinates. We call *m'* the amount of memory available in the buffer at a given time during the execution of the algorithm ($m' = m$ at initialization). The list of buffers' coordinates contains the coordinates of each buffer to be loaded in the referential of the reconstructed image. The algorithm successively loads as much data as it can from the input files into the buffer and write it down to the output files that are supposed to contain this data. Although we could use a naive shape for the buffer, we can use the input and output files shapes to elaborate more efficient strategies as we will see in the next sections. The algorithm ends when all buffers have been read. Therefore, the buffers must cover the whole reconstructed image such that when the algorithm ends all the output files have been completely written. Moreover, we must also ensure that all the data read from input files are either stored in RAM or directly written, such that all the output files are completely written at the end of the algorithm.

Given Algorithm 1, the optimization problem that we want to solve can be stated as follows: Given the amount of main memory available *m*, as well as the shapes of the input and output files *I* and *O*, how to select the best buffer shape *B* which will minimize the number of seeks that take place during reading and writing?

We add two restrictions on the buffers: We shall use only non-overlapping buffers, all buffers having the same size, and each buffer has to be written only once.

Algorithm 1 Basic resplit algorithm

Inputs: *inFiles*, *outFiles*, *m*, *buffersList*
for *buffer* in *buffersList* **do**
 read(*inFiles*, *buffer*)
 write(*outFiles*, *buffer*)
end for

C. Consistency with previous works

For the sake of consistency with previous works [2], we call the original array of shape *R* stored in the input files the “reconstructed image” (see Figure 1). Of course, the input files' positions in the reconstructed image have to be stored in some way. Also with a view to be consistent with previous works, we assume the files to be written in column-order. In column ordering (also called “F” ordering) the fastest moving dimension is the last dimension of the array and the slowest moving dimension is the first dimension. For example a 3D array with dimensions *i*, *j* and *k* would be written on disk by writing the complete columns in the *k* dimension first (see Figure 2).

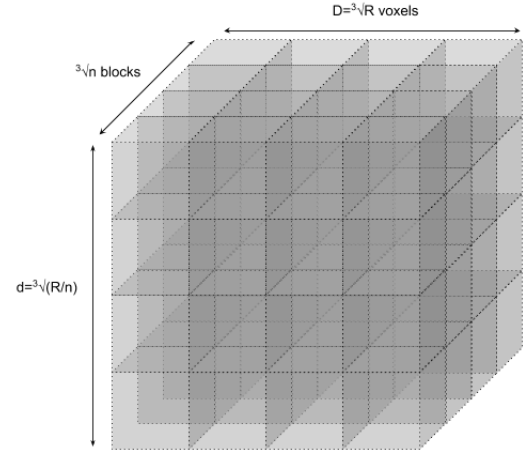


Fig. 1. Illustration of the reconstructed image divided into input files at the initialization of the resplit algorithm.

D. Naive algorithm

As a base case for our problem, let us define a naive algorithm which loads one input file at a time and write it down into the different output files that requires the data. The buffers have the same shape as the input files ($B = I$) and the order is the same order than the storage order i.e. the column order in this study ($[k, j, i]$ order).

E. A particular case

If the input shape is a multiple of the output shape such that one input file covers several output files entirely without falls, then the problem is easily solved: one must read as much input files as one, i.e. the buffer shape *B* is a multiple of the input shape *I*. The algorithm will produce one seek per input file and one seek to write each output file, which is the minimum number of seeks possible for a resplit.

If there is a mismatch between the shapes in any dimension however, one needs a strategy to manage with this overlap while minimizing the number of seeks. We will introduce a strategy to keep falls temporarily into memory in the next section, this strategy's efficiency is completely dependent on

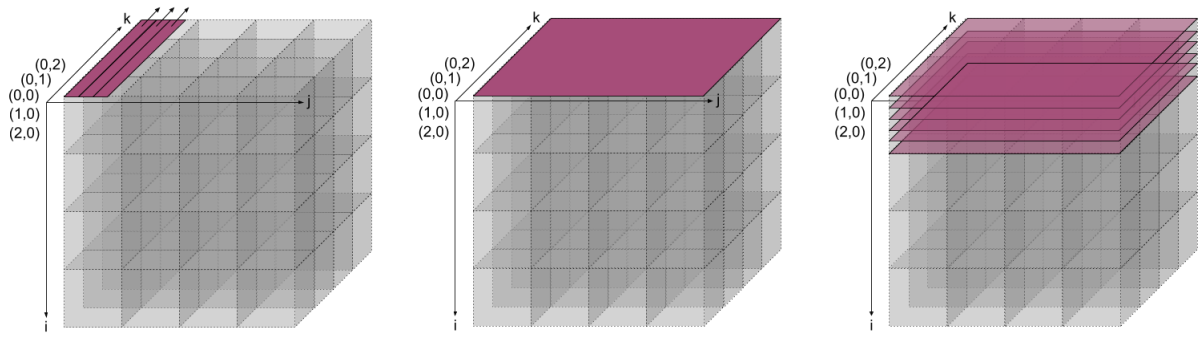


Fig. 2. Illustration of the column-order storage of voxels in a file.

the amount of main memory available.

Also, the resplit process requires multiple buffers to be read. If there is no overlap between input and output files, then the order in which we load buffers is not important. In case of an overlap however, the order may have an important impact on the number of seeks produced.

THE “KEEP” ALGORITHM

The algorithm presented in this paper is called the “keep” algorithm, as it relies on a so-called “keep strategy” that is presented below. The “keep” algorithm is an algorithm to find the best buffer shape for a resplit.

F. The “keep” strategy

At initialization, we assign a shape $B = (1, 1, \dots, 1)$ for the buffer. We will then stretch the buffer in each dimension until all the available memory m has been used while keeping the number of seeks as small as possible.

Let us consider the first dimension f that we increase: Ideally, one wants the buffer to cover both the input file and the output file, I_f and O_f , such that we read and write in one seek each. If there is a mismatch between O_f and I_f (meaning one is not a multiple of the other) we want to read a multiple of O_f or I_f such that we either read partially or write partially but not both at the same time. We must therefore choose between reading a multiple of O_f or I_f . Reading a multiple of I_f in the case of a shape mismatch is equivalent to $B_f = nI_f = mO_f + xO_f$ with $0 < x < 1$, n and m are integers. We call “extra data” the data contained in the overlap areas/volumes between input and output shapes (Figure 3). In this example the extra data is xO_f . An output file that is involved in an overlap is called an “incomplete output file”.

One can try to keep the extra data in memory instead of writing it directly into the output file: this is what we call the “keep strategy” (Figure 4). The idea is to read “more than necessary” from the input files, ideally reading each input file

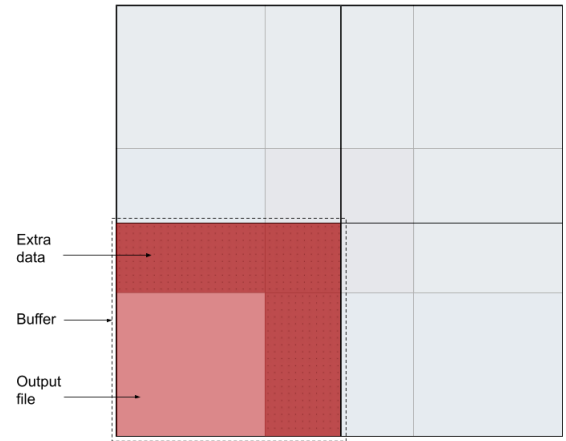


Fig. 3. Illustration of the concept of extra data with a 2D case. In this example the black bordered rectangles represent the input files and the gray bordered rectangles represent the output files. After having read the red buffer we can see that the output file covered by the light red area can be written directly. After having written the data from the light red area however, we are left with the dark red, dotted area which represents some extra data we would like to keep in memory until we read the rest of the incomplete output files in order to read the incomplete output files with one seek.

in one seek, and to keep as much extra data in memory as possible. The goal of this strategy is to read input files in one seek, and write as much output files as possible in one seek as well. If we were to read a multiple of O_f however, we would read an input file partially and there is nothing that one can do about it. It may be that we can only keep some of the overlaps in memory as it could take too much buffers until we can write an incomplete output file in one seek without running out of memory. Extra data that cannot be kept in memory is written directly in the output file(s). By doing so we ensure that when the algorithm ends all the data has successfully been written.

G. Input aggregates

We call “input aggregate”, Λ , the aggregate of the minimum number of input files that covers one output file entirely in all dimensions (see Figure 6, Figure 5). The input aggregate is the minimal buffer shape we would like to have to be

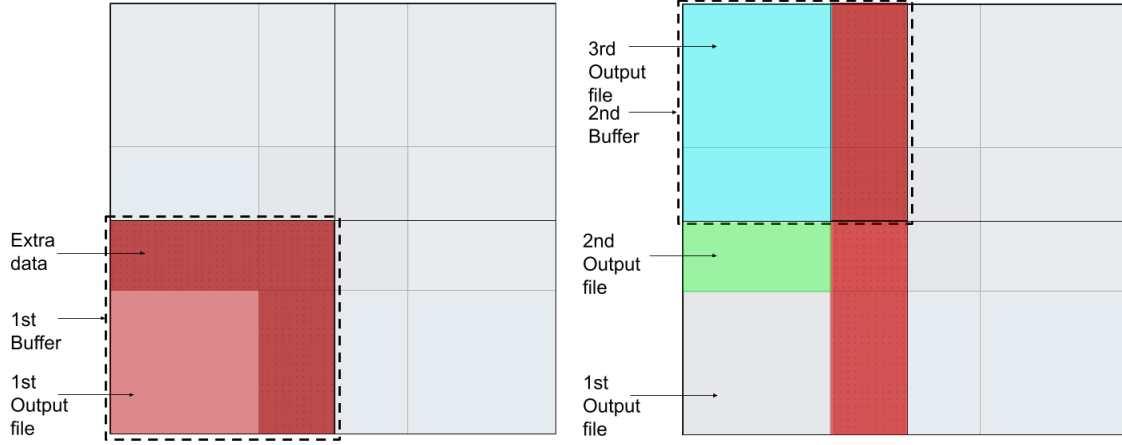


Fig. 4. Illustration of the keep strategy with a 2D case. In this example the black bordered rectangles represent the input files and the gray bordered rectangles represent the output files. As shown on the left figure, the keep strategy consists in reading more than one output file (light red area) into the first buffer. Then the first output file is written and the dark red area is kept in memory. On the left figure the second buffer has been loaded. It allows to free part of the overlap in the k direction (the green area) as the second file data is complete in main memory. The third output file has been read completely and can therefore be written. This lets the overlap in the j direction (dark red area) in main memory for the next buffer.

capable of minimizing the number of seeks as it allows to read the input files in one seek and to use the keep strategy. The mismatch between the input and output shapes and the limited memory available for the buffer may prevent it to be possible.

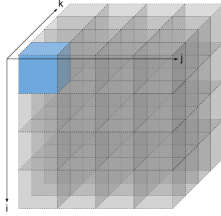


Fig. 5. First buffer in the column-order storage.

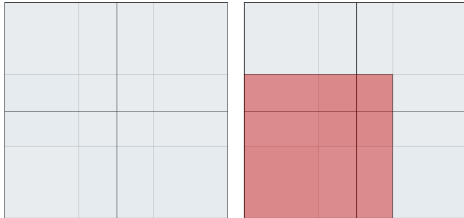


Fig. 6. Illustration of an input aggregate in two dimensions. On the left schema, consider the surface containing all rectangles as being the reconstructed image, with the small rectangles (with black borders) being the input files and the big rectangles (with gray borders) being the output files. On the right side an input aggregate is illustrated by the red area: it is the smallest number of input files such that the surface of at least one output file is completely covered. In this example, four input files are required to cover the first output file.

H. Stretching the buffer in the storage order

The first step of the algorithm is therefore to make the dimensions of the buffer match the dimensions of the input aggregate as much as possible (we want $B = \Lambda$). If more memory is available after that, we may want to increase the buffer's dimension even further (it is discussed in subsection "Stretching beyond the input aggregate shape" below). The question of which dimension should be increased first remains. One should increase the buffer's dimensions in the order of the fastest moving dimensions. For example if one is processing a 3D image stored in files following the column-order, one should increase the buffer in the k dimension first, then the j dimension and finally the i dimension.

Let us first consider the 2D case in which the input and output files overlap in one dimension only. An overlap occurring in the k dimension would incur B_k seeks as opposed to 1 seek only (see Figure 7) in the case of an overlap in the j dimension (see Figure 8). Note that the keep strategy is not useful if we cannot store the extra data in one direction completely (see Figure 8).

Let us define 3 overlapping areas, represented in Figure 9. The f_1 area is the overlap in the k axis. The f_3 area is the upper right overlap which combines the overlaps in both axis. One can see the f_3 area as the continuity of the f_1 area. The f_2 area is the overlap in the j axis only. The f_1 area becomes the F_1 volume in 3D, the F_2 volume corresponds to the f_2 area, etc.

In terms of the number of seeks,

$$\Omega_j > (\Lambda_j - \Omega_j) > 1 \Rightarrow \text{seeksIn}(f_1) > \text{seeksIn}(f_3) > \text{seeksIn}(f_2)$$

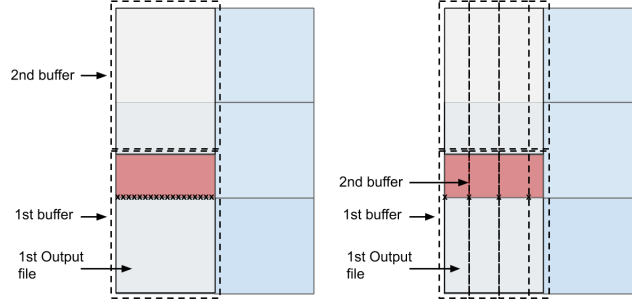


Fig. 7. Left figure: without keep strategy. Right: with keep strategy. Keeping extra data in memory reduces the number of seeks caused by writing but increases the number of buffers needed to write an output file. The crosses represent the number of seeks that happen in both cases.

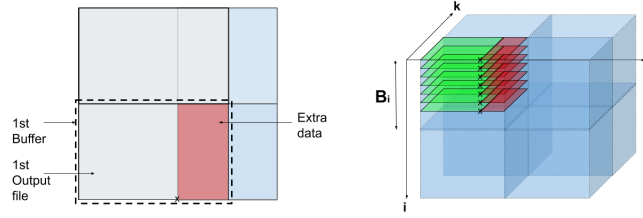


Fig. 8. Illustration of an overlap in the j dimension. The black crosses represents one seek each. In the 2D case on the left we can see that writing down the data into the next output file on the right would take only one seek. In the 3D case however (right side), writing the data down into the next output file would take B_i seeks.

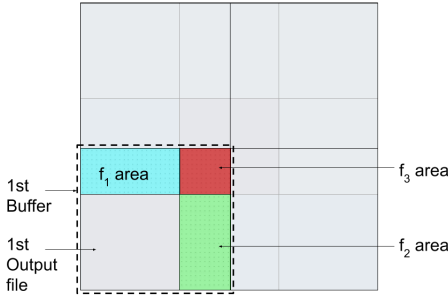


Fig. 9. Illustration of the different overlapping areas in 2D. The blue area is called the f_1 area, the red area is called the f_3 area and we call f_2 area the green one.

in 2D, and this stays true when increasing the number of dimensions. We must therefore use the memory available for stretching the buffer to store f_1 first, then f_3 , and finally f_2 . The same way that it is not useful to save part of the f_2 area as opposed to storing the whole f_2 area, it is not useful to increase B_j if $B_k < \Lambda_k$. Indeed, the number of seeks will be the same than without using the keep strategy. As the f_3 area is the continuity of the f_1 area in the j direction, if the available buffer memory m' allows to store the f_3 area but not the f_3 and f_2 areas together it seems that increasing B_j from Ω_j to Λ_j while storing only the f_3 area could save a lot of seeks.

As a conclusion, we should increase the fastest moving dimension first to allow the reduction of the number of seeks

in the f_1 and f_3 areas in the k direction. Then we should increase B_j as much as possible, while saving only f_1 and then f_3 , even if we cannot cover it entirely. Finally, if we can store the whole f_2 area, we should do it before increasing the i dimension to prevent the 1 seek to be multiplied by B_i .

I. Stretching beyond the input aggregate shape

Using small input and output file shapes, it is probable for the buffer to have been stretched to the input aggregate, which means that a good amount of RAM is still available. One has no reason to stop increasing the shape of the buffer from that point. The first idea which comes to mind is to stretch the buffer such that the buffer length becomes a multiple of the output file length in a given dimension x : $B_x = nO_x$. Again, this would create an overlap in the input files however, which would prevent the keep strategy to be applied and create a maximum number of seeks. Therefore, one can only stretch the buffer by adding one input file length to a given dimension of the buffer shape (see Figure 10). A good idea seems to be to extend the buffer shape in the dimension of the biggest overlap area/volume such that we accumulate extra data from the smallest overlap volumes only. Again, if we can stretch the buffer such that $B_x = R_x$ in the dimension x of the biggest overlap area/volume, we then increase in the dimension of the second biggest overlap size, etc.

J. Special case of non optimality

When the amount of memory available is too small, a special case arises where the algorithm is not optimal. Consider the 2D case where there is an overlap in the k dimension only

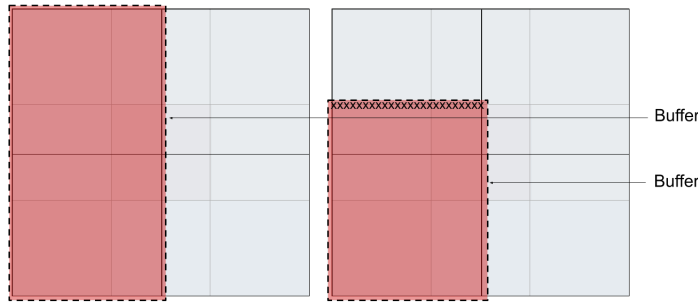


Fig. 10. Left side: Extending the buffer to the next input length. Right side: Extending the buffer to the next output length. The crosses indicates the number of seeks. As we can see, extending the buffer to the next output shape incurs a lot of seeks in the next input file when loading the buffer. As the input file is incomplete, there is no extra data, therefore the keep strategy cannot be used.

Algorithm 2 Pseudocode of the “keep” algorithm

```

1: volumesToKeep,  $B \leftarrow$  bufferExtensionAlgorithm(...)
2: buffers  $\leftarrow$  getBufferCoordinates( $B$ )
3: storage  $\leftarrow$  new Dictionary()
4: for buffer in buffers do
5:   bufferData  $\leftarrow$  read(buffer, inputFiles)
6:   writeOffSurplus(bufferData, outFiles, volumesToKeep,
     storage)
7:   keep(storage, bufferData, outFiles)
8:   for outFile in storage do
9:     if outFile is complete then
10:      write(outFile)
11:      remove outFile from storage
12:   end if
13: end for
14: end for

```

(see Figure 7). Let us say that the buffer size is such that $B_k = \Lambda_k$, but is too small to extend B_j to Λ_j . Keeping the extra data from the overlap in the k dimension in memory allows us to write contiguously in the output files but it implies using more buffers in the j^{th} axis. In the worst case, $B_j = 1$ which incurs Λ_j seeks. As soon as $B_j > 1$ however (keeping $B_k = \Lambda_k$), we divide the number of seeks produced by writing in the incomplete output file per the number of buffers. This observation is true for the other dimensions as well: as soon as $B_x = \Lambda_x$ in a given dimension x , the more we can read in the next dimension $x + 1$, the better. This proves that we should not increase a dimension if $B_x < \Lambda_x$ in the preceeding dimensions.

K. Impact of the buffer order on performance

Using the keep strategy in case of overlaps, one may order the buffer loadings to further reduce the number of seeks. By optimizing the buffer ordering one can reduce the maximum quantity used to store the extra data in memory. For example, if an overlap occurs only in the k axis, loading the next buffers in this direction will enable recycling the extra data kept in memory, resulting in a smallest memory consumption over time. The memory saved thanks to a smart ordering

could enable the storage of more overlaps in memory using the “keep strategy”, further reducing the overall number of seeks.

As we will see, the buffer ordering problem is complex and does not seem easily solvable. Thankfully, the impact of the buffer ordering on performance can be mitigated. Indeed, the impact of the buffer ordering depends on the size of the falls, i.e. the overlaps between the buffer and the incomplete output files’ shapes. One can reduce the falls’ sizes by using smallest chunks: Even if the overlap between the input and output files is big with respect to their size, the area/volume of the falls will be kept small. In particular, we remark that the falls tends to be smaller when the buffer shape is bigger than the output file shape, as the overlaps are smallest and concentrated on the borders (see Figure 11). We can stimulate this property by using small chunks such that we use buffer bigger input aggregates (see Stretching beyond the input aggregate shape), while keeping the overlaps small at the borders.

A MEMORY ANALYSIS OF THE KEEP STRATEGY

The keep algorithm assumes that m is big enough to keep at least a column of length Λ_k . (Note: a modifier car bad assumption) As it has been explained, in 3D the buffer is then stretched in the j^{th} dimension and finally in the i^{th} dimension as long as there is enough main memory available to use the keep strategy. This section covers how we estimate the amount of memory required by the keep strategy for a given buffer shape in order to know how much we can stretch the buffer in each dimension.

Two pieces of information are required to know the worst case memory consumption of the keep algorithm: what is the maximum number of each overlap area/volume we will have to keep in memory during the algorithm and what is the maximum size that those overlap areas/volumes can have. In this analysis we express the overlap size in terms of the number of voxels that constitutes the overlap because the real quantity of memory used depends on the size of a voxel in memory. For convenience one can avoid this extra parameter by setting the number of bytes per voxel to 1.

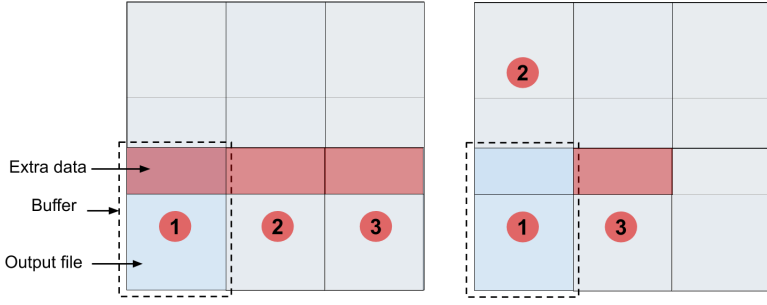


Fig. 11. Comparison of a bad buffer ordering (left side) against a good buffer ordering (right side), given an overlap in the k dimension. The red area represents the amount of extra data kept in memory after loading the third buffer. The right side order allowed to release the extra data from the first buffer after reading the second buffer.

L. Nomenclature

Let us first define the nomenclature we will use (see Figure 12). Given a buffer shape B and the output file shape O , we define the overlap $C_k(x)$ as the overlap length in direction k for the x th buffer:

$$C_k(x) = xB_k \bmod(O_k)$$

$C_k(x)$ is not fixed during the buffer shape extension as it extends progressively. We therefore define $\Omega_k(x)$ as the overlap length in direction k if the buffer had shape $B = \Lambda$:

$$\Omega_k(x) = x\Lambda_k \bmod(O_k)$$

Finally, we define $\Theta_k(x)$ as follows: $\Theta_k(x) = \Lambda_k - \Omega_k(x)$

Thanks to this nomenclature, we can now express the (F_1, F_2, \dots) volumes as follows (see Appendix A):

$$\begin{aligned} F_1 &= \Omega_k \min(B_j, \Theta_j) \min(B_i, \Theta_i) \\ F_2 &= \Theta_k \max(0, \min(B_j - \Theta_j, \Omega_j)) \min(B_i, \Theta_i) \\ F_3 &= \Omega_k \max(0, \min(B_j - \Theta_j, \Omega_j)) \min(B_i, \Theta_i) \\ F_4 &= \Theta_k \Theta_j \max(0, \min(B_i - \Theta_i, \Omega_i)) \\ F_5 &= \Omega_k \Theta_j \max(0, \min(B_i - \Theta_i, \Omega_i)) \\ F_6 &= \Theta_k \Omega_j \max(0, \min(B_i - \Theta_i, \Omega_i)) \\ F_7 &= \Omega_k \Omega_j \max(0, \min(B_i - \Theta_i, \Omega_i)) \end{aligned}$$

M. A “naive” buffer ordering

The maximum number of each overlap to keep in memory depends on the buffer order. Therefore, the buffer order defines the maximum amount of extra data that we will have to keep in memory. As we do not know a good ordering which do not imply a compute-intensive algorithm (see section “Discussion”), let us define a naive ordering. This naive ordering read the buffers in the direction where the overlap is the most important first, then read those buffer columns in the order of the second dimension where the overlap is the biggest, etc. For example if the order is (j, k, i) , we will read the buffers in direction j first. This order reduces the number of big overlaps to keep in memory.

According to this ordering, we can now find the formula of the maximum number of overlap areas/volumes that will be

stored during the keep algorithm execution: If, for example, the ordering was $[k, j, i]$ (see Figure 13), we would keep at a maximum

$$\Sigma = (F_1 + n(F_2 + F_3) + N(F_4 + F_5 + F_6 + F_7) + B_i B_j B_k) \alpha \quad (1)$$

in memory. n is the number of buffers in direction of the biggest overlap and N is the number of buffers in a buffer slice in the direction of the second biggest overlap.

Indeed, if we consider the overlap in the direction of the biggest overlap, direction k on Figure 13 (a), one can see that the extra data from the first buffer (in blue) is being recycled by the second buffer. At any given time, the amount of memory used by the keep strategy is the size of the next buffer plus all the extra data that is pending for usage from the preceeding buffers. If there is no overlap in the j dimension one can see that the amount of memory to keep between the buffers is only F_1 , $\Sigma = (F_1 + B_i B_j B_k) \alpha$, i.e. the overlap in F_1 is always recycled by the next buffer. Note that after loading the last buffer in the k dimension the last F_1 extra data stored is recycled. Therefore one keep only one time F_1 at any given time in the resplit process. With the same reasoning we can see that if the size of F_2 and F_3 is not 0 (Figure 13 (b)) one will have to keep it until a complete buffer column has been processed such that the first buffer of the buffer column at buffer index $j + 1$ will recycle the extra data from the first buffer from the column buffer at buffer index j . This formula is for the 3D case, but it can be extended to the N dimensional case with the same reasoning.

N. The buffer extension algorithm

From the buffer order formula (Equation 1) and the formulas of each F volume we can derive the maximum amount of memory used Σ while extending the buffer and therefore find the optimal value ϕ for each dimension of B at each stretching stage. The computation details are described in Appendix A. The buffer extension algorithm is divided in 3 parts: Algorithm 3, Algorithm 5, Algorithm 5. Algorithm 3 describes the buffer extension in the 2D case (f_1, f_3 and f_2). Algorithm 4 describes the buffer extension in the 3D case until the buffer shape B equals the input aggregate

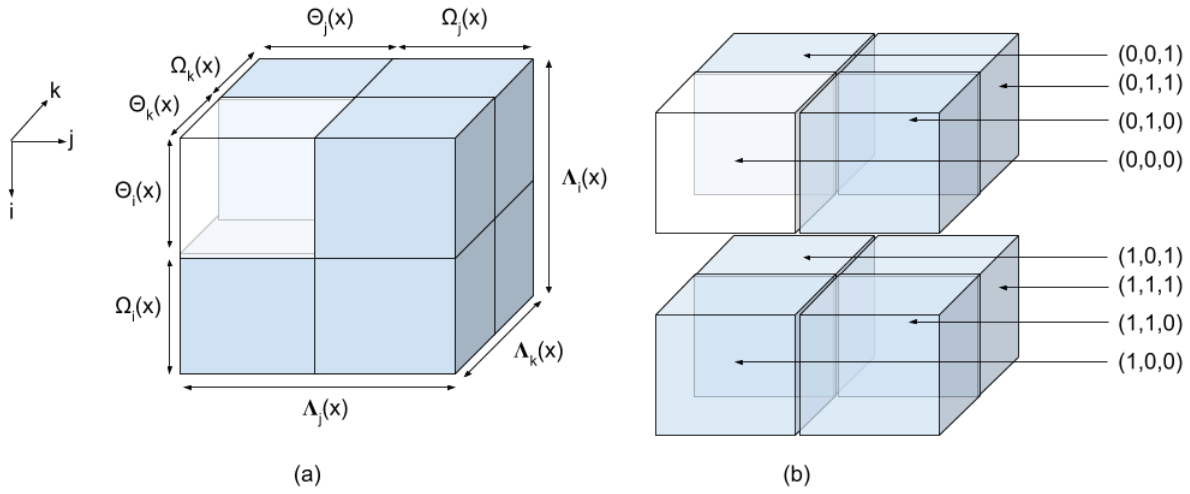


Fig. 12. Figure (a) represents an input aggregate of shape Λ . The white cube represents the data that can be written directly into output files and the other blue blocks represent incomplete files. Figure (b) shows how the different volumes are indexed following the buffer order. Here the buffer order is the storage order. Another way of naming volumes is by using the decimal value or their positions, for example (0,0,1) is block 1, (1,1,0) is block 6.

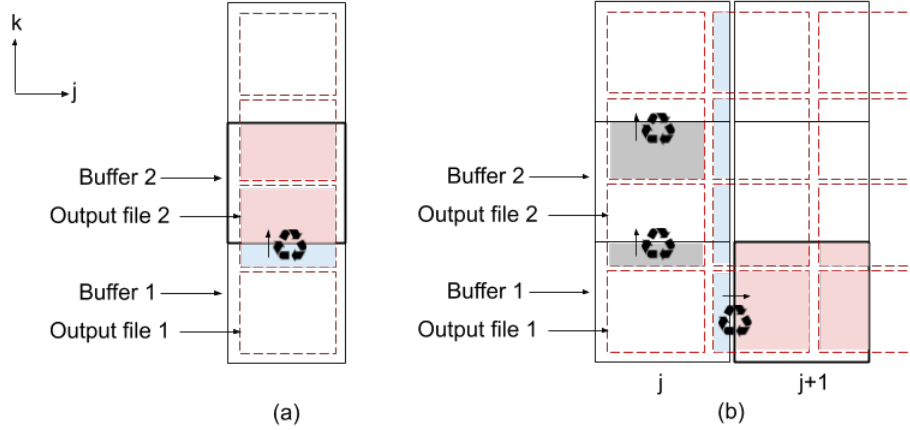


Fig. 13. On both figures the solid black borders blocks are buffers and the dotted red blocks are output files. We consider a buffer of interest with a thick black border. The red areas represent the extra data loaded by the buffer of interest and the blue areas represent the extra data kept in memory from the previous buffer loadings. On figure (a) an overlap in the first buffer order dimension is shown. One can see that each buffer recycles the previous extra data such that one never store more that the size of a buffer plus the size of F_1 in memory during the resplit process. Figure (b) shows that an overlap in the second dimension of the buffer order has to be kept in memory b_k times with b_k the number of buffers in the k^{th} dimension. This can be extended to 3D by adding the overlaps in the third dimension of buffer order N_s times with N_s the number of buffers in a buffer slice.

shape Λ . Finally, Algorithm 5 describes how to extend the buffer beyond the input aggregate shape, by successively adding one input file length to one dimension of the buffer shape (the dimension of biggest overlap). It uses the algorithm *addInputFileLength* described in Algorithm 6 to compute the estimated memory consumption Σ_{test} . While Σ_{test} is less than m one can continue extending the buffer.

IMPLEMENTATION

EXPERIMENTS

RESULTS

DISCUSSION

- O. Usage and support for MINC/Nifti file formats*
- P. Comparison with previous work*
- Q. Solution of the ordering problem*
- R. Extending the algorithm for ROI extraction*
- S. Distributed keep algorithm*
- T. Future works*

CONCLUSION

ACKNOWLEDGMENTS

REFERENCES

- [1] K. Amunts, C. Lepage, L. Borgeat, H. Mohlberg, T. Dickscheid, M.-É. Rousseau, S. Bludau, P.-L. Bazin, L. B. Lewis, A.-M. Oros-Peusquens,

Algorithm 3 Buffer extension algorithm (part 1)

```
 $B = [\Lambda_k, 1, 1]$ 
// storing  $F_1$ 
 $\phi = \lfloor \frac{m}{(\Omega_k + \Lambda_k)\alpha} \rfloor$ 
if  $\phi > \Theta_j$  then
   $B_j \leftarrow \Theta_j$ 
else
   $B_j \leftarrow \phi - 1$ 
  return  $B$  //stop algorithm
end if
// storing  $F_3$ 
 $\phi = \lfloor \frac{m - [\Omega_k \Theta_j](1-n)\alpha}{(n\Omega_k + \Lambda_k)\alpha} \rfloor$ 
if  $\phi > \Lambda_j$  then
   $B_j \leftarrow \Lambda_j$ 
else
   $B_j \leftarrow \phi - 1$ 
  return  $B$  //stop algorithm
end if
if  $B_j > \Theta_j$  then
  storeF3  $\leftarrow$  True
else
  storeF3  $\leftarrow$  False
end if
// take into account buffer order
if  $f_1 > f_2$  then
   $g_1 = f_1, g_2 = f_2, g_3 = f_3$ 
   $\phi = \lfloor \frac{m}{\alpha[\Omega_k \Theta_j + n\Lambda_k \Omega_j + \Lambda_j \Lambda_k]} \rfloor$ 
   $A = F_1 + n(G_3 + F_2)$ 
else
   $g_1 = f_3, g_2 = f_2, g_3 = f_1$ 
   $\phi = \lfloor \frac{m}{\alpha(\Theta_k \Omega_j + n\Omega_k \Lambda_j + \Lambda_j \Lambda_k)} \rfloor$ 
   $A = F_2 + n(G_3 + F_1)$ 
end if
// storing  $F_2$ 
 $\Sigma = [g_1 + n(g_3 + g_2) + B_i B_j B_k]$ 
if  $\Sigma < m$  then
  continue
else if  $\Sigma = m$  then
  storeF2  $\leftarrow$  True
  return  $B$  //stop algorithm
else
  storeF2  $\leftarrow$  False
  return  $B$  //stop algorithm
end if
```

Algorithm 4 Buffer extension algorithm (part 2)

```
// storing  $F_1, F_2, F_3$ 
if  $\phi > \Theta_i$  then
   $B_i \leftarrow \Theta_i$ 
else
   $B_i \leftarrow \phi - 1$ 
  return  $B$  //stop algorithm
end if
// storing  $F_5, F_6, F_7$ 
 $\phi = \lfloor \frac{\frac{m}{\alpha} - A + \Theta_i B}{(B + \Lambda_k \Lambda_j)} \rfloor$ 
if  $\phi > \Delta_i$  then
   $B_i \leftarrow \Delta_i$ 
else
   $B_i \leftarrow \phi - 1$ 
  return  $B$  //stop algorithm
end if
if  $B_i > \Theta_i$  then
  storeF5F6F7  $\leftarrow$  True
else
  storeF5F6F7  $\leftarrow$  False
end if
// storing  $F_4$   $\Sigma = \text{computeSigma3D}(F_1, F_2, \dots, F_7, B, \alpha)$ 
if  $\Sigma > m$  then
  storeF4  $\leftarrow$  True
else
  storeF4  $\leftarrow$  False
end if
```

Algorithm 5 Buffer extension algorithm (part 3)

```
// stretching buffer beyond input aggregate
 $p = [0, 0, 0]$ 
 $ndims = \text{len}(m)$ 
for  $dim$  in  $\text{range}(ndims)$  do
  for  $j$  in  $\text{range}(nbFiles[dim] - \frac{B_{dim}}{I_{dim}})$  do
     $p' = p$ 
     $p'[dim] += 1$ 
     $\Sigma_{test} = \text{addInputFileLength}(B, p')$ 
    if  $\Sigma_{test} < m$  then
       $p[dim] = p'[dim]$ 
    else
      for  $i$  in  $ndims$  do
         $B[i] = B[i] + p[i]I[i]$ 
      end for
      return  $B$  // end algorithm
    end if
  end for
end for
return  $B$  // end algorithm
```

N. J. Shah, T. Lippert, K. Zilles, and A. C. Evans, "Bigbrain: An ultrahigh-resolution 3d human brain model," *Science*, vol. 340, no. 6139, pp. 1472–1475, 2013.

- [2] V. Hayot-Sasson, Y. Gao, Y. Yan, and T. Glatard, "Sequential algorithms to split and merge ultra-high resolution 3d images," in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 415–424, Dec 2017.

APPENDIX A

FORMULAS FOR BUFFER EXTENSION

A. Overlap volume memory sizes

This appendix presents the reasoning and computations leading to the formulas used for the keep algorithm to find

the biggest buffer shape possible given m , the amount of memory available.

At step x , for $B(x)$ and the corresponding $\Omega(x)$ and $\Theta(x)$, we define the size of each volume (in number of voxels) as follows (see Figure 12):

$$\begin{aligned} F_1 &= (0, 0, 1) = \Omega_k \min(B_j, \Theta_j) \min(B_i, \Theta_i) \\ F_2 &= (0, 1, 0) = \Theta_k \max(0, \min(B_j - \Theta_j, \Omega_j)) \min(B_i, \Theta_i) \\ F_3 &= (0, 1, 1) = \Omega_k \max(0, \min(B_j - \Theta_j, \Omega_j)) \min(B_i, \Theta_i) \\ F_4 &= (1, 0, 0) = \Theta_k \Theta_j \max(0, \min(B_i - \Theta_i, \Omega_i)) \\ F_5 &= (1, 0, 1) = \Omega_k \Theta_j \max(0, \min(B_i - \Theta_i, \Omega_i)) \\ F_6 &= (1, 1, 1) = \Theta_k \Omega_j \max(0, \min(B_i - \Theta_i, \Omega_i)) \\ F_7 &= (1, 1, 0) = \Omega_k \Omega_j \max(0, \min(B_i - \Theta_i, \Omega_i)) \end{aligned}$$

For the sake of clarity we do not write the (x) suffix (B means $B(x)$). Also, we are only interested in the maximum amount of memory we would have to keep. To that aim, we will not evaluate the above volume formulas for each x but we will replace $\Omega_k(x)$ by its maximum in the formulas instead. By definition of $\Omega_k(x)$ an upper bound is O_k but one can find the exact maximum value by computing $\Omega_k(x)$ for all x , for each dimension k .

B. Maximum amount of overlap to keep in memory

We defined the values of the different F volumes but we do not know which one of F_4, F_1, F_2 is the biggest overlap i.e. we do not know the buffer order ahead of time. In 3D, they are 9 possibilities for the buffer order. If the order of biggest overlap is i, j, k (Figure 14 (b)) as opposed to k, j, i (Figure 14 (a)), the volume indices change and therefore Equation 1 is false (as it defines Σ for the k, j, i buffer order only). We define a more general formula for Σ as follows:

$$\Sigma = (G_1 + n(G_2 + G_3) + N(G_4 + G_5 + G_6 + G_7) + B_i B_j B_k) \alpha$$

At each step in the buffer extension process, we will assign the right F volume to each G (see Figure 14).

C. Keeping $f1$

$$\begin{aligned} g_1 &= f_1 = \Omega_k B_j B_i = \Omega_k B_j \\ \Sigma &= (g_1 + B_i B_j B_k) \alpha \\ \Sigma < m &\Leftrightarrow (\Omega_k B_j + B_j \Lambda_k) \alpha < m \\ &\Leftrightarrow B_j (\Omega_k + \Lambda_k) \alpha < m \\ &\Leftrightarrow B_j < \phi, \phi = \lfloor \frac{m}{(\Omega_k + \Lambda_k) \alpha} \rfloor \end{aligned}$$

D. Keeping $f3$

$$\begin{aligned} g_1 &= f_1 = \Omega_k \Theta_j B_i = \Omega_k \Theta_j \\ g_3 &= f_3 = \Omega_k (B_j - \Theta_j) B_i = \Omega_k (B_j - \Theta_j) \\ \Sigma &= [g_1 + n g_3 + B_i B_j B_k] \alpha \\ \Sigma &= [\Omega_k \Theta_j + n [\Omega_k (B_j - \Theta_j)] + B_j \Delta_k] \alpha \quad \Sigma = [\Omega_k \Theta_j] (1 - n) \alpha + B_j (n \Omega_k + \Lambda_k) \alpha \\ \Sigma < m &\Leftrightarrow B_j < \lfloor \frac{m - [\Omega_k \Theta_j] (1 - n) \alpha}{(n \Omega_k + \Lambda_k) \alpha} \rfloor \end{aligned}$$

E. Keeping $f2$

$$\begin{aligned} f_1 &= \Omega_k \Theta_j B_i = \Omega_k \Theta_j \\ f_2 &= \Omega_k \Omega_j B_i = \Omega_k \Omega_j \\ f_3 &= \Theta_k \Omega_j B_i = \Theta_k \Omega_j \end{aligned}$$

We are asking ourselves if we can keep F_2 . Remember that we only store all F_2 or nothing as it only saves 1 seek in 2D. If $F_1 > F_2$ we compute Σ with $g_1 = f_1, g_2 = f_2$. In the other case, $g_3 = f_3$ and $g_1 = f_2, g_2 = f_3, g_3 = f_3$.

$$\Sigma = [g_1 + n(g_3 + g_2) + B_i B_j B_k]$$

F. Increasing $f1, f2, f3$ to $F1, F2, F3$

$$\begin{aligned} F_1 &= \Omega_k \Theta_j B_i \\ F_2 &= \Theta_k \Omega_j B_i \\ F_3 &= \Omega_k \Omega_j B_i \end{aligned}$$

We are scaling the overlap areas from previous step into volumes. Therefore the buffer order is the same than at the previous step.

If k is the first buffer order ($g_1 = f_1$):

$$\begin{aligned} \Sigma &= [G_1 + n(G_3 + G_2) + B_i B_j B_k] \alpha \\ \Sigma &= [\Omega_k \Theta_j B_i + n(\Omega_k \Omega_j B_i + \Theta_k \Omega_j B_i) + B_i B_j B_k] \\ \Sigma &= B_i \alpha [\Omega_k \Theta_j + n \Lambda_k \Omega_j + \Lambda_j \Lambda_k] \\ \Sigma < m &\Leftrightarrow B_i < \lfloor \frac{m}{\alpha [\Omega_k \Theta_j + n \Lambda_k \Omega_j + \Lambda_j \Lambda_k]} \rfloor \end{aligned}$$

If j is the first buffer order ($g_1 = f_2$):

$$\begin{aligned} \Sigma &= [G_1 + n(G_3 + G_2) + B_i B_j B_k] \alpha \\ \Sigma &= [\Theta_k \Omega_j B_i + n(\Omega_k \Omega_j B_i + \Omega_k \Theta_j B_i) + B_i B_j B_k] \alpha \\ \Sigma &= B_i \alpha (\Theta_k \Omega_j + n \Omega_k \Lambda_j + \Lambda_j \Lambda_k) \\ \Sigma < m &\Leftrightarrow B_i < \lfloor \frac{m}{\alpha (\Theta_k \Omega_j + n \Omega_k \Lambda_j + \Lambda_j \Lambda_k)} \rfloor \end{aligned}$$

G. Keeping $F5, F6, F7$

$$\begin{aligned} F_1 &= \Omega_k \Theta_j B_i \\ F_2 &= \Theta_k \Omega_j B_i \\ F_3 &= \Omega_k \Omega_j B_i \\ F_5 &= \Omega_k \Theta_j (B_i - \Theta_i) \\ F_6 &= \Theta_k \Omega_j (B_i - \Theta_i) \\ F_7 &= \Omega_k \Omega_j (B_i - \Theta_i) \end{aligned}$$

$$\Sigma = [G_1 + n(G_3 + G_2) + N(G_5 + G_6 + G_7) + B_i B_j B_k] \alpha$$

We define A and B as follows:

$$\begin{aligned} A &= G_1 + n(G_3 + G_2) \\ B &= (\Omega_k \Theta_j + \Theta_k \Omega_j + \Omega_k \Omega_j) N \\ \Sigma &= [A - \Theta_i B + B_i B + \Lambda_k \Lambda_j B_i] \alpha \\ \Sigma < m &\Leftrightarrow B_i < \lfloor \frac{m - A + \Theta_i B}{(B + \Lambda_k \Lambda_j)} \rfloor \end{aligned}$$

H. Keeping $F4$

Our goal is to link the good G to its corresponding F volume size (see Figure 14). To that aim we know that:

1) Given β old the base of storage order (i, j, k) we have the

$$F \text{ indices. } \beta \text{ old} = \left(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right)$$

2) Any block index (a, b, c) is in fact $(a * \beta[0] + b * \beta[1] + c * \beta[2])$ with β the base used for indexing blocks.

3) Saying that the buffer order is i, j, k is equivalent to indexing the blocks using the base (k, j, i) .

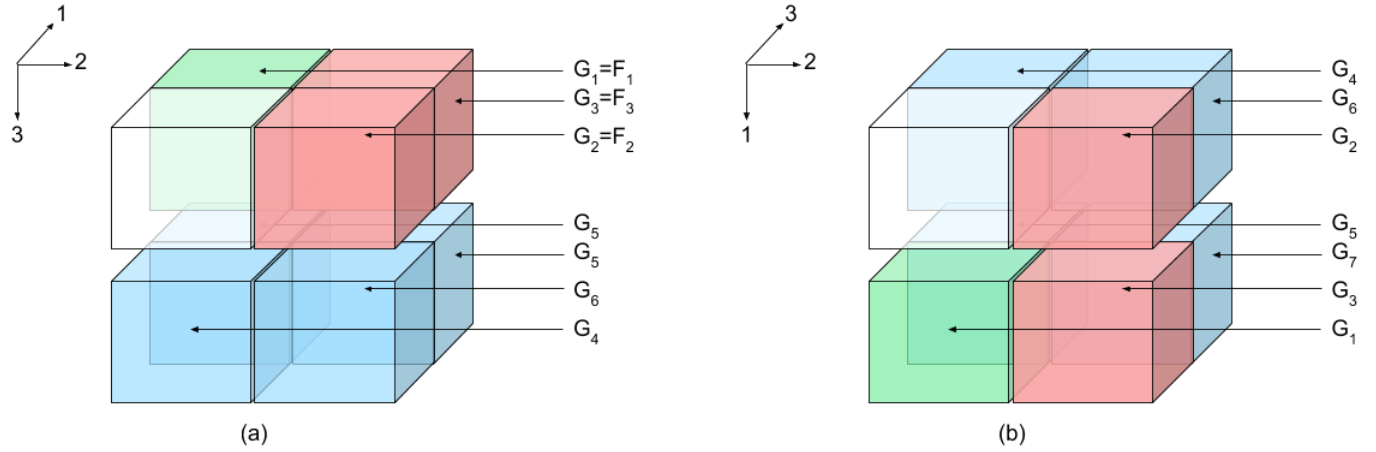


Fig. 14. Figure (a). Figure (b).

Therefore, let us say that we have a certain buffer order. Any index of a block G in this base can be written $(a_{new} * \beta_{new}[0] + b_{new} * \beta_{new}[1] + c_{new} * \beta[2]_{new})$. To find the corresponding volume in the storage order (because we know the F values) we can replace β_{new} by β_{old} . The only information we need is β_{new} . This can be found by ordering F_1, F_2, F_4 by their volume size. If F_1 is the biggest

volume, $\beta_{new}[0] = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ for example. At this point we just

have to fill in a dictionary $G(i)$ associating index i to the corresponding volume size. We then replace G by $G(i)$ in $\Sigma = [G_1 + n(G_3 + G_2) + N(G_4 + G_5 + G_6 + G_7) + B_i B_j B_k] \alpha$. As for F_2 we either store all F_4 or we do not store it. If $\Sigma < m$ then we can store F_4 and move on to the final step of the algorithm.

Algorithm 6 Pseudocode for addInputFileLength

Inputs: B, p, α
 $B[i] = B[i] + p[0]I[i]$
 $B[j] = B[j] + p[1]I[j]$
 $B[k] = B[k] + p[2]I[k]$
 $F_1 = O_k(B_j - O_j)(B_i - O_i)$
 $F_2 = (B_k - O_k)O_j(B_i - O_i)$
 $F_3 = O_k O_j (B_i - O_i)$
 $F_4 = (B_k - O_k)(B_j - O_j)O_i$
 $F_5 = O_k(B_j - O_j)O_i$
 $F_6 = (B_k - O_k)O_j O_i$
 $F_7 = O_k O_j O_i$
return $computeSigma3D(F_1, F_2, \dots, F_7, B, \alpha)$

Algorithm 7 Pseudocode for computeSigma3D

Inputs: $F_1, F_2, \dots, F_7, B, \alpha$
 $G = dict()$
 $G_{old} = \{1 : F_1, 2 : F_2, \dots, 7 : F_7\}$
 $bufferOrder = sortByValue((0, 1, 2), \{0 : F_1, 1 : F_2, 2 : F_4\})$
 $\beta_{new_to_old} = replace(bufferOrder, \{0 : \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, 1 : \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, 2 : \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\})$
for num_{new} **in** $range(8)$ **do**
 $a_{new} = decToBin(num_{new})$
 $a_{old} = a_{new}[0] * \beta_{new_to_old}[0] + a_{new}[1] * \beta_{new_to_old}[1] + a_{new}[2] * \beta_{new_to_old}[2]$
 $num_{old} = binToDec(a_{old})$
 $G[num_{new}] = G_{old}[num_{old}]$
end for
 $\Sigma = [G_1 + n(G_3 + G_2) + N(G_4 + G_5 + G_6 + G_7) + B_i B_j B_k] \alpha$
return Σ
