Adaptive Online Data Compression

Emmanuel Jeannot

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Adaptive Online Data Compression

Emmanuel Jeannot

N° 4400

Mars 2002

THÈME 1

ISSN 0249-6399 ISRN INRIA/RR--4400--FR+ENG

INRIA
LORRAINE

Adaptive Online Data Compression

Emmanuel Jeannot *

Thème 1 — Réseaux et systèmes
Projet Résédas

Rapport de recherche n° 4400 — Mars 2002 — 11 pages

Abstract: Quickly transmitting huge data in the context of distributed computing on wide area network can be achieved by compressing data before transmission. However, such an approach is not efficient when dealing with high-speed networks. Indeed, the time to compress a large file and to send it is greater than the time to send the uncompressed file. In this paper, we propose an algorithm that allows to overlap communications with compression and to adapt the compression ratio according to the network speed (the slower the network, the more we use efficient and slow compression algorithms). The advantage of such an adaptive algorithm is its generality and that its suitability for a large set of applications.

Key-words: Distributed computing, efficient communications, adaptive network algorithm, online compression algorithm, communication overlap

* LORIA, Université H. Poincaré

Unité de recherche INRIA Lorraine
LORIA, Technopôle de Nancy-Brabois, Campus scientifique,

Compression adaptative et dynamique de données

Résumé : Pour transmettre des données de grandes tailles rapidement, dans le cadre du calcul distribué sur des réseaux à grande échelle, une des solutions possibles consiste à compresser ces données avant de les envoyer. Cependant, une telle approche ne fonctionne pas dans le cadre des réseaux rapides (100 Mbits/s) car le temps de compresser un grand fichier et de l'envoyer est supérieur au temps d'envoi du fichier non compressé. Nous présentons ici un algorithme qui permet de recouvrir la communication par la compression et d'adapter le taux de compression à la vitesse du réseau (plus le réseau est lent plus on utilise des algorithmes efficaces et lents pour compresser). L'avantage d'un protocole adaptatif de ce type est qu'il fonctionne bien pour un très grand nombre de réseaux et un large panel d'applications.

Mots-clés : Calcul distribué, communication efficaces, alagorithme adaptatif pour les réseaux, algorithme de compression dynamique, recouvrement de communication

1 Introduction

Recent developments in the area of grid-computing have focused on the need to efficiently transfer very large amount of data. Indeed, when designing Problem Solving Environments (PSE) such as NetSolve [3], Ninf [8], or Scilab$_{//}$ [2], data (files, internal objects, etc.) are to be sent. Performance of PSE greatly depends on the ability to efficiently transmit these data. However a meta-computing environment is composed of an heterogeneous set of machines interconnected by heterogeneous networks. Available computing resources can be old sequential machine as well as brand new parallel computers. These machines can be interconnected by a slow WAN (Internet), as well as a fast 2.5 Gbit/s backbone (VTHD[1]) or a 100 Mb-it/s LAN, etc.

Since no assumptions can be made on the hardware infrastructure we need a general transmission algorithm that is efficient whatever the speed of the network and the speed of the machines.

When the server is a parallel machine and the data to be transmitted are distributed among the nodes of the server one possibility to efficiently transfer these data is to use multi-socket technics. This solution has the advantage to try to use the maximum bandwidth of the backbone but is limited when a bottleneck exists (such has a slow network connected client) and is not possible when the server is a sequential machine. An other solution is to compress data prior to send them. However, this solution is inefficient in the case of a fast network when the time to compress the data and to send them is greater than the time to send the whole uncompressed data.

In this paper we propose an algorithm called *AdOC* (Adaptive Online Compression), which has the following features. First, it is a general algorithm suited, not only for grid computing, but for any data transfer (ftp, rcp, etc.) whatever the speed of the network. Second, it is a an adaptive algorithm which means the compression ratio depends on the amount of available time to compress (the slower the network the higher the compression ratio). Third, it is a dynamic algorithm with compression/communication overlap. This means that compression and communication are done simultaneously. Therefore, our algorithm is able to decide either to compress the data when useful or to send the data uncompressed when the network is fast.

Dynamic compression has been proposed for improving cache performance as shown in [4, 6]. But, as far as we know, there is no similar algorithm to AdOC proposed in the literature.

The remaining of the paper is organized as follow. Section 2 presents some important facts about communication and compression. The AdOC algorithm is proposed section 3. Section 4 presents our experimental results. Section 5 discusses possible optimizations of the algorithm. Concluding remarks are given section 6.

2 Compression and Communication

We present here some facts about compression and communication. We manly show that a none adaptive algorithm cannot be efficient for all kind of network and machines.

In the remaining of this paper we focus on the following problem. Given a file on a machine (the sender), how to transfer it as fast as possible to an other machine (the receiver).

2.1 Facts about Compression

A solution to efficiently transfer a file is to compress this file. However there is an overhead due to the compression time. We evaluate this overhead by measuring the time taken to compress and uncompress various files. We have taken two bench files. The first one is `oilpann.hb`. It is an sparse matrix in the Harwell/Boeing format (ASCII). This file has many redundancy and compresses very well. The second file is an archive of binaries (`bin.tar`), which compresses less than `oilpann.hb`. For sake of simplicity `bin.tar` have been truncated to the size of `oilpann.hb` (52 633 266 bytes).

Table 1 shows the time to compress the bench files using LINUX `gzip` with various compression options on a Pentium III 500MHz running LINUX kernel 2.4.7-10. `gzip` uses the Zempel-Liv compression algorithm [9] which basically checks for redundancy. It uses a "*sliding window*" which size defines the compression level (the larger the window the higher the compression level). Column *level* is the gzip compression option from 1 (fast method / low compression) to 9 (slow method / best compression). Column *c. time* is the LINUX user

[1] `http://www.vthd.org`

time to compress the file in seconds. Column *ratio* is the ration between the compressed size of the file and the uncompressed size. Column *d. time* is the LINUX user time to decompress the file. We remark that

	`oilpann.hb`			`bin.tar`		
level	c. time	ratio	d. time	c. time	ratio	d. time
1	10.2	4.88	9.6	16.9	2.23	12
2	11	5.13	9.3	17.2	2.27	12
3	11.3	5.52	8.9	18.3	2.31	13.4
4	11.9	5.83	9	21.7	2.38	13.7
5	13.6	6.32	8.7	23.6	2.43	13.6
6	16.7	6.64	8.3	28.3	2.44	13.7
7	20.4	6.75	8.7	32.1	2.45	13.14
8	47.6	6.99	7.7	44	2.45	13.1
9	76	7.02	7.8	65.8	2.46	12.3

Table 1: *Compression timings on bench files using gzip*

the compression time and ratio increase with the compression level. Moreover, for a given file, we remark that the decompression time is also roughly constant and is always smaller than the compression time.

2.2 Facts about Communication

One idea to transfer a file is to use FTP (file transfer protocol) or rcp (remote copy). Our experiments show that rcp and FTP have the same performances. Hence, we only show results for FTP. Table 2, shows FTP timings for transferring `oilpann.hb`[2] on various network type and various compression level. 100 Mbit/s experiments were done on the LAN of the LORIA whereas 10Mbit/s experiments were performed between two PCs interconnected by a cross cable. Transatlantic experiments were done between the LORIA (Nancy, France) and the university of Rutgers (New-Jersey, USA). Files were on the sender local disk and sent to the receiver local disk (NFS was not involved). Remark that the compression time are not included in the timing.

Row *FTP* is the time to send the file uncompressed. Row *level=1* (resp. *level=9*) is the time to send the file compressed with gzip using compression level 1 (resp. 9).

A second idea consists in using the `-C` option of the UNIX `scp` command that does a remote copy of a file using encryption and compression. The compression algorithm used by scp is the same used by gzip. The last two rows of table 2 show the timings of transferring the file using scp.

method	100 Mbit/s	10 Mbit/s	Transatlantic	ADSL	56k Modem
FTP	6.34	70.5	234	831	9900
level=1	1.08	14.5	50	171	2017
level=9	0.93	9.5	42.2	121	1452
`scp`	12	70	560	1010	10550
`scp -C`	23	68	220	342	1660

Table 2: *FTP and scp average timings on various network*

We can see that FTP/rcp is a good solution for high speed network. Indeed, we need about one second to transfer the compressed file but from 10 to 76 seconds to compress it and about 8 seconds to decompress it. Therefore sending the uncompress file is faster than compressing sending and then uncompressing the file.

[2] We have taken this file which compresses the best in order to minimize the transfer time

This is no longer true when dealing with slower network. In that case the dominant time is the transfer time and the compression/decompression time becomes negligible.

We also see that using scp is never a good idea, mostly because there is a high overhead due to encryption.

Thus, we can conclude that there is no general solution to efficiently transfer a file for any kind of network.

3 AdOC Algorithm

3.1 Overview

We have designed an algorithm that is able to decide what is the fastest way to transmit the file. Our current implementation is based on the TCP socket. It uses the two following features:

1. *Multi-threading.* The sender is composed of two threads. One thread (called the compression thread) is in charge of reading the file and compressing data when useful. An other thread (called the communication thread) is in charge of sending data. Such an architecture allows compression/communication overlap.
2. *FIFO.* A FIFO data structure is used to store data to be sent. The compression thread writes data to the FIFO. The communication thread reads the FIFO.

The most important value of the AdOC algorithm is the size of the FIFO. If the FIFO size is not growing this means that the sending thread consumes data faster than they are produced by the compression thread. In that case we deduce that we are dealing with a fast network and we have no time to compress data. On the other hand if the FIFO is growing this means that we are dealing with a slow network and hence we have some time to compress data and we can increase the compression level.

3.2 Compression/Communication Overlap

When a process/thread is doing some IO on a device (such as writing data to a file on a disk or to a socket on a network), it may be blocked waiting for the device to become ready. This especially happens when writing data to a socket because memory accesses are faster than network accesses. When this process/thread is blocked the processor is available for some computations. We use this system behavior in order to overlap communication with compression. The sending process is composed with two threads one for compression, one for communication. These two threads communicate using a FIFO data structure.

The compression thread produces data to be sent and the communication thread consumes these data and send them to the network. Such a pipeline schemes allow to hide the communication time with the compression time. Indeed, when the communication thread is suspended while sending data the processor is available for the compression thread to perform some computations.

Since the FIFO is a shared object between two threads we use a mutex and a semaphore to access it.

3.3 Compression Thread

The compression thread has in charge to read the file and to compress data if possible.

We use the *zlib* [7] from Gailly and Adler. It is a compression library that uses the same algorithm than gzip (Zempel-Liv). It provides a C API that is able, given a buffer and a compression level between 1 and 9, to compress this buffer in packets of a given size.

For instance, imagine we want to compress a buffer of 5 000 000 bytes with a packet size of 8192 bytes. If the compressed buffer has a size of 500 000 bytes we actually obtain 61 packets of 8192 bytes and one packet of 288 bytes. The nice thing with this API is that the compression function returns each time a packet is generated. The next call to the function resume the compression and generate the next packet, without any loss in the compression ratio. On the other side we can start decompressing the buffer each time a new packet arrives.

The algorithm of the compression thread is given Figure 1. fifo_size, is a global object that contains the number of non empty packets in the fifo. packet_size is a constant of the program. It is the maximum size of each packet stored in the fifo. When level = 0 (no compression is used), we read a packet from the

file and store it in the fifo. buffer_size is also a constant of the program. It is the size of the buffer to be compressed (level≠0). A buffer is compressed into packets of size at most packet_size with the function compress_next_packet, which returns the compressed packet and its actual size. We add delimiters in the fifo to tell when a buffer is completely compressed or when the file is completely read. At the beginning of each step we save in prev_fifo_size the size of the fifo. At the end of each step we update the level of compression using the difference between the current size of the fifo and the size of the fifo at the beginning.

```
1  level=0;
2  while there is still data in the file
3      prev_fifo_size=fifo_size;
4      if (level=0)
5          packet=readfile(packet_size);
6          add_fifo(packet,packet_size,level);
7          add_fifo(NULL,0,0);
8      else
9          buffer=readfile(buffer_size);
10         while buffer is not fully compressed
11             (packet,size)=compress_next_packet(buffer,packet_size,level);
12             add_fifo(packet,size,level);
13         add_fifo(NULL,0,0);
14     level=update_level(level,fifo_size-prev_fifo_size);
15 add_fifo(NULL,0,0);
```

Figure 1: *Compression Thread Algorithm*

3.4 Updating Level of Compression

As shown Table 1 increasing the level of compression allows to decrease the size of the compressed file but increases the compression time. At the beginning we set the level to 0 (no compression) and we slowly increase this level if we are sure that we have enough time to compress buffers. Updating the level of compression works in a *slow start* conservative fashion : increasing the level is slow and decreasing the level is fast.

Figure 2 describes the algorithm for updating the level of compression. It uses the difference D between the size of the fifo at the end of a compression step and the size of the fifo at the beginning of the step. We have three cases :

1. D=0. This means that the number of packets produced by the compression thread is the same that the number of packets consumed by the communication thread.Hence we do not change the level of compression.
2. if D>0 this means that we are dealing with a slow network, hence we increase the level by 1 if the fifo size is greater than 10. We use this threshold in order to avoid the algorithm to switch from the non-compressing state to the compressing state too easily. This is particularly important in the case of a fast network.
3. D<0. The fifo size is decreasing and we divide the compression level by 2.

3.5 Tuning AdOC

We need to decide which value the buffer size and the packet size have to take.

```
input : level : current compression level, D : fifo variation size
output : new compression level.

1  if (D<0)
2      level=level/2;
3  else if ((D>0) and (fifo_size > 10) and (level <9))
4      level++;
```

Figure 2: *Updating the level of compression*

3.5.1 Determining the Buffer Size

Our algorithm split the file into buffers and compresses them independently. Due to the compression algorithm, the obtained compression ratio depends on the size of the given buffer (the larger the buffer, the better the compression ratio). However, since the compression level cannot be change while compressing a buffer, too large buffers have a bad impact on the reactivity of our algorithm. Hence, we need to find a trade-off for this value. Figure 3 shows the evolution of the compressed file size when increasing the buffer size for a level 1 compression of `oilpann.hb`. We see that the size rapidly decreases at the beginning and tends to a limit which is the compressed file size obtained using gzip. With gzip there is no buffer: the file is read and compressed in one step. We obtain similar results for `bin.tar` and various compression level. We have decided to choose a buffer size of 204 800 bytes because in most of the cases, with this size we obtain a difference less than 5% between the compressed file and the optimal compressed file size (the one obtained by gzip).

3.5.2 Determining the Packet Size

When not compressing the file (at the beginning of the algorithm or when dealing with a fast network), the file is splited into packets which are sent uncompressed. If we have too large packets the reactivity of our algorithm at the beginning will be affected. However if we have small packets, we will have a lot of elements added in the fifo which implies more overhead for managing the fifo and sending this packet. We need also here to find a trade-off. In order to have an idea of what is a good packet size we have performed the following computation which is inspired from optimizing the packet size for computation/communication overlap [5].

Let ν the packet size. Let B the size of the buffer to be compressed. Let ρ the compression ratio (hence we obtain a compressed buffer size $B' = B/\rho$). Let C the time to compress one byte (the buffer is compressed in time CB), and C' the time to uncompress one byte (the buffer is uncompressed in time $C'B$). Let γ_c the constant overhead to store a packet in the fifo. Let $t = \frac{B}{\rho\nu}$ the number of packets generated by the compression thread.

Hence the time to generate one packet is $T_a(B,\nu) = C\frac{B}{t} + \gamma_c = C\rho\nu + \gamma_c$, on the over hand the time to uncompress one packet is $T_c(B,\nu) = C'\rho\nu$.

For modeling the network we use a simplistic model that assumes a constant throughput and latency. Let γ_e the constant overhead to extract a packet from the fifo. Let τ the time to transmit one byte and β the network latency. Hence the time to send one packet is $T_b(\nu) = \gamma_e + \beta + \tau\nu$. This is an old and simplistic model but, as shown in the following, it will give us an idea of the order of the packet size.

We have a pipeline scheme, hence, the time to transmit a the whole buffer is :

$$\begin{aligned} T_t(B,\nu) &= T_a(B,\nu) + tT_b(\nu) + T_c(B,\nu) \\ &= C\rho\nu + \gamma_c + \frac{B}{\rho\nu}(\gamma_e + \beta + \nu\tau) + C'\rho\nu \\ &= (C + C')\rho\nu + \frac{B}{\rho\nu}(\gamma_e + \beta) + \gamma_c + \frac{\tau B}{\rho} \end{aligned}$$