IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 29-33

Comparative Study of Compression Tools like TAR, RAR, ZIP

Prof. Swati D. Ghule, Dr. AnupGirdhar
[1]*Assistant Professor, P.E. S. Modern College of Engineering, Pune – 5*
[2]*Guide TMVEmail:anupgirdhar@gmail.com*

Abstract: *We want to compress files and directories on a computer as there are many reasons. Some of them are conserving less disk space and using minimize bandwidth for network communication.Archiving data generally means taking backup and saving it to a secure location, which is in a compressed format.Archive file is onewhich is composed of one or more files including metadata in the extra fields like filename, permission and so on. Archive files are generally used to combine multiple data files into a single file to achieve f portability and less storage space or simply to compress files for less storage space.Archiving makes file transfer easy than compressed files which require less storage space and are thus faster to move from one system to another.ZIP files are one of the convenient way to wrap up related files together, and can save storage space accordingly*

I. Introduction

The volumes of digital data being produced are growing day by day. According to an International Data Corporation view data were created in huge amount. In the future, this alarmingamount of data is supposed to grow at a 57% annual growth rate, which is faster than the expected growth of storage media capacity. Moreover, therequirement for this is to preserve a larger part of this data. Because of this there is always the requirement for cost-effective digital archives.

Programmers generally know *ar* as it is widely used today to generate static libraries, which is nothing but archives of *compiled* files. As*ar* can be used to generate archives of all kind. Generally, *.deb* package files on Debian systems *are*nothing but *ar* archives! And on MacOS X, *mpkg* packages are gzip-compressed *cpio* archives. Tar is more popular thanar and *cpio* among users. Since the tar command was really good and simpler to use [2].

RAR is a well-known archive file format which is used for data compression, error recovery and file spanning[2].

Structurally, the RAR file consist of variable length blocks of required and optional data. The accuracy of blockcreation evolved over time with the versions. At first, the RAR file consist of marker or introductory block, an archive block which consists of archive header and file header, and closing block which consists of additional comments or other information is required to properly process the file. The order of these blocks may vary, but the first block must be anintroductory block followed by an archive header block. The **archive block** is verycomplicated because it contains the headers of its archives as well as the file headers [5].

The ZIP file format was introduced in late 1980's by Phi Katz for his PKZIP utility. It was updated year byyear, and now includesvarious compression algorithms. It is not always the case that ZIP algorithms are mosteffective or efficient - there are many other competitors which may work better in many situation, that might compresssuperior or faster or both - but its overall performance is good. Therefore, the ZIP file format is being internally used by manyproducts like Java’s JAR files, SAS Enterprise Guide project files, etc., in addition to this as this is a standard file format fordistributing groups of files that a user might extract and use directly [1].

After a decade the introduction of*tar*, *zip*also comes in the MS-DOS world as an *archive format supporting compression*. The most common compression technique used in*zip* is *deflate* which is nothing but the implementation of the LZ77 algorithm. But being developed by PKWARE for commercial purpose, the z*ip*file format has faced to patent hampering for years. So, *gzip* was producedfor the implementation of LZ77 algorithm in free software without violationof any PKWARE patent.*gzip* was used*only* to compress files. Thus to create a *compressed archive file*, first you have to create an *archive*file using the *tar* utility for example. Then, you will *compress* that archive file. Which is a *.tar.gz* file (sometimes abbreviated as *.tgz)*

As computer science developed, some more compression algorithms were also designed to achieve higher compression ratio. For example, the Burrows–Wheeler algorithm used in *bzip2* (abbreviated to *.tar.bz2* archives). Or in recent times*xz* which implementsLZMAalgorithm like the one used in the *7zip* utility.

But the *zip*file format is supported by Windows, so this one is specificallyused in cross-platform environments. You can also find the *zip* file format in manyother places. For example, Zip format was also used by Sun for *JAR* archives which is useful for distributing compiled Java programs. AlsoinOpenDocument files

(*.odf*, *.odp* …) used by LibreOfficeandmany other office suites. All these files formats are nothing but zip archives.

II. Need / Importance Of Study

Compression is a way of decreasing the size of a file on disk using different algorithms and some mathematical calculations. Files are organizedin a way that it makes their general structure which can be predict easily, even if their content varies. As the, contentsin files aregenerally repeated. These bothgive the opportunities to apply compression techniques.

A**lossless** compression method generates a file which is small in size than the original that can be used to regenerate the original file.

Lossless compression techniques does not compress files on the basis of approximations, and instead they use certain algorithms to identifythe repeated portions infile. It removes this repetition and replaces them with a placeholder. And continues the process of replacing later occurrences of the pattern with reference. This makes the computer to store the information in less disk space. This process is considered as creating a list of variables that define blocks of data, and then using these variables later on to use in the program. This is actually a two stage process that all lossless compression techniques used: first map highly repeated values with something which is lessrepeated and that can be easily referenced and then change the occurrences of all those values with the reference.

Furthermore, the modern lossless compression techniques are **adaptive**. As they do not analyze the whole input file at the beginning and create the "dictionary" of reference which is to be substitute. Rather, they analyze the file as they start reading and recreate the dictionary based on which data in actual is repeated. The dictionary becomes progressive and more competent as the process continues. It replaces the later occurrences of pattern withreferences with the same placeholder.

III. Statement Of Problem

When you download the file it can be either *.tar*, *.zip* or *.gz* extensions. But don’t you know what isthe difference between all these Tar, Zip and Gz is***?*** Why we use them and which one is more efficient, tar or zip or gz?

The difference between zip and tar and gz:
.tar is uncompressed archive file
.zip is (usually) compressed archive file
.gzis file (archive or not) compressed using gzip technique.

Basically, a tar file format is a suitable way to distribute, store, back up, and handle groups of related files [6]. ZIP file format defines only a limited set of mandatory file attributes to store for each file entry like filename, modification date, permissions. Instead of these basic attributes, an archiver may store some more metadata in the extra field of the ZIP file header. But, these extra fields are depend on implementation, so there is no guarantees though the archiver is efficient to store or retrieve the same set of metadata [4].

IV. Hypothesis

Tar does not compress the data. The meaning is that the size of tar archive file is the same as the sum of the sizes of packed files, plus some overhead metadata. If data compression is required, you can use the other compression tools like gzip or bzip2 with *tar [6]*.

RAR archives generally provide a markably higher compression ratio than ZIP file format.

ZIP files are a suitable way to group related files together, so that the storage space can be saved at the same time.

V. Research Methodology

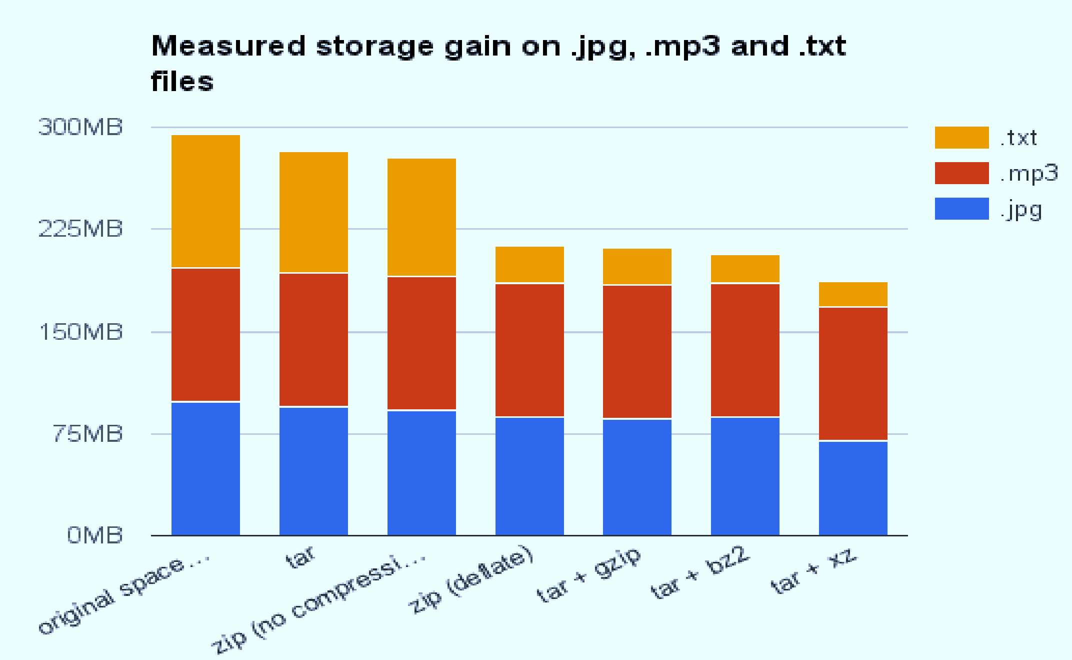
Tar vs. Zip vs.Gz Efficiency Test

Space efficiency — as we can observe, more potentially efficient is a compression technique, more CPU it requires.

Here are the result obtained (all size as reported by *du -sh*):

File type	.jpg	.mp3	.mp4	.odt	.png	.txt
Number of files	2163	45	279	2990	2072	4397
Space on disk	98M	99M	99M	98M	98M	98M
tar	94M	99M	98M	93M	92M	89M
zip (no compression)	92M	99M	98M	91M	91M	86M

zip (deflate)	87M	98M	93M	85M	77M	28M
tar + gzip	86M	98M	93M	82M	77M	27M
tar + bz2	87M	98M	93M	42M	71M	22M
tar + xz	70M	98M	22M	348K	51M	19M



VI. Result And Discussion

Tar supports a many compression programs such as gzip, bzip2, lzip, lzma, lzop, xz and compress. When creating compressed tar archives normally we append the compressor's suffix to the archive file name.

gzip is probably the most widely used storing tool used in Unix and Linux systems. It uses the *Lempel-Ziv coding* (LZ77) for data compression.

The gzip tool uses a compression technique known as "DEFLATE". This algorithm is also used in some other popular technologies like the PNG image file format, the HTTP web protocol, and the SSH secure shell protocol.

One of the main advantages of it is itsspeed. It can compress as well as decompress data at much higher speed than other competing technologies, especially when comparing these utility's most compact compression formats. It is also very effective in terms of memory usage during compression and decompression and requiresless memory when optimizing for best compression.

While gzipuses the "DEFLATE" algorithm,and bzip2 is an implementation of "Burrows-Wheeler algorithm".

The important balance for users is greater compression at the cost of longer compression time. The bzip2 can create more compact files than gzip, but take much long time to achieve the results because of more complex algorithm.

The decompression time requirement is less as compare to compression time, so it can be an advantage to distribute files using the bzip2file format since you need only to suffer the time penalty during compression and can be able to distribute smaller files that can be decompressed in comparatively less amount of time. The time required for decompression is still much greater than gzip, but does not have as big impact as the compression operation.

Another thing which can be noted is that the memory requirements are larger than gzip.

The xz compression utilities influence a compression algorithm known as LZMA2. This algorithm has more compression ratio as compare to the above two examples, to make it a good format when you required to store data on limited disk space. It gives comparatively smaller files.

This again comes to cost, in most of the areas that bzip2have. While the compressed files that xzcreates are smaller than the other tools, it takes *considerably*longer timefor compression.

The xz compression utility also has more memory requirements, sometimes equal to an order of magnitude over the other utilities. If you are working on a system with sufficient memory, this may not be the problem, but this is a concern to think over it.

While the compression time may be quite more than is preferable, but the decompression time is relatively good. While it can'treache togzip in terms of decompression speed, it is usually much faster at decompression than bzip2.

The main difference is that bzip2 uses *Burrows-Wheeler block sorting text compression algorithm* in combination with *Huffman coding* instead of the LZ77 algorithm which is used in gzip. The compression technique of bzip2 gives more efficient compression than gzip's. However, computing the bzip2 compression technique usually is more complex and takes more time (i.e., uses more CPU cycles) than gzipcompression. [6]

RAR uses optional AES encryption, which is a type of block cipher and uses an algorithm that encrypts data in each block. There are various types of the AES standards and the implementations used by RAR which changes with various versions. RAR5 (current version) uses AES-256, rather than AES-128 used in RAR4 [5].

Concerning *.jpg*, *.mp3* and *.mp4* you know all these are the compressed data files. Also, you may know that they use *destructive compression*. That means we can'treproduce*exactly* the same image after a JPEG compression technique. And it's true. But do you know after the destructive compression phase, when the data are compressed for the second time using the non-destructive Huffman variable word-length algorithm used to remove data redundancy.

Because of this, it was assumed that compressing JPEG images or MP3/MP4 files will not gain much. Please keep in mind thatgenerallya file contains both the highly compressed data and some uncompressed metadata; still we can gain something there.

VII. Findings

.tar file is just a plain*archive*file in which data are not in compressed form. In other words, if you create a tar of 100 files of 50kB, you will get an archive whose size will be around 5000kB. The only gain that can be expected using tar alone would be it avoids the space wasted by the file system.As most of them allocate space at some granularity (for example, on some system, a one byte long file uses 4kB of disk space, 1000 of them will require 4MB but the corresponding tar archive it just requires 1MB)[6].

RAR format has becomemore popularduring the years as compared to its competitor archive formats like 7Z, zip, etcBecause it has better data compression rate than ZIP and uses a lossless compression technique [2].

RAR has many advantages over ZIP files like "more convenient multipart (multivolume) archives, tight compression including special solid, multimedia and text modes, strong AES-128 encryption, recovery records which helps to repair an archive even in case of physical data damage, Unicode is used for processing non-English file names and many more"[5].

Similar in purpose to ZIP files, RAR files are also data containers in which one or more files are stored in compressed form [4].

VIII. Recommendation And Suggestions

Now a days we can freely use any archive file format both on Linux & Windows.But the *zip*file format is having built in support on Windows, this is specificallyused in cross-platform environments. Also you can see the *zip* file format in various places. For example, it wasalso used by Sun for *JAR* archives to distribute compiled Java programs. Or for Open Document files (*.odf*, *.odp* …) used by LibreOfficeandsome other office suites. All these files formats are nothing but zip archives. If you're curious, try to *unzip* one of them to see what's inside:

Still in favor of *tar* archive type because the *zip* file format does not support all the Unix file system metadata reliably. For some concrete reasons of that statement, you must keep in mind that the ZIP file format only defines a limited set of mandatory file attributes to store for each file entry: filename, modification date, permissions. Beyond these basic attributes, an archiver may store some additional metadata in the extra fields of the ZIP file header. But, these extra fields are implementation-specific, there is no guarantees even for efficientarchivers to store or retrieve the same set of metadata.

IX. Conclusion

It is important that we must be conscious of the performance drawbacks and compatibility issues that may be included with each solution. How much importance you give to these issues depends fully on the machines you are working on and what type of clients you arein support?On most modern machines it might happen that you should not have to pay too much attention to these details, but they can cause disputes if you blindly implement a compression technique when interacting with older machines.

Compressing files saves storage space and makes data transfer faster, but it may take a lot of time. Data compression is CPU intensive and compressing a data set of several terabytes may require tremendous computing capability.

X. Future Scope For Further Research

As per the *.jpg*, *.mp3* and *.mp4*file formats you know all these are *already* compressed data files. Also, you may know that they use *destructive compression*. That means you can’t createexactly same original image after a JPEG compression technique. And its true. But do you known that after the destructive compression phase, the data are compressed a second time using the non-destructive Huffman variable word-length algorithm to remove data redundancy.

References

[1]. “Creating ZIP Files with ODS”, Jack Hamilton, Division of Research, Kaiser Permanente, Oakland, California, Paper 131-2013, SAS Global Forum 2013,

[2]. “The Weakness Of Winrar Encrypted Archives To Compression Side-channel Attacks”,Open Access Theses, Fall 2014, Kristine Arthur-Durett, Purdue University.

[3]. “Creating ZIP Files with ODS”, Jack Hamilton, Division of Research, Kaiser Permanente, Oakland, California, Paper 131-2013.

[4]. http://www.pkware.com/documents/casestudies/APPNOTE.TXT>

[5]. http://www.win-rar.com/rarproducts.html>

[6]. https://itsfoss.com/category/linux/ Tar Vs Zip VsGz : Difference And Efficiency.