wiki:

<https://en.wikipedia.org/wiki/Zip_(file_format)>

<https://tools.ietf.org/html/rfc1951>

<https://users.cs.jmu.edu/buchhofp/forensics/formats/pkzip.html>

zip文件格式：



最后面是EOCD，用来记录有多少file entry和central directory开始的位置等。可以通过EOCD用来查找到central directory的位置。

EOCD前面的是central directory。每一个文件，在这里都有一个记录，里面有filename，文件数据开始的offset。可以通过central directory查找到文件关联数据的位置。

最开始的是文件定义，分为local file header和具体的data。local file header用来记录文件名字和文件压缩数据开始的地方。

central directory和local file header里面有重复的地方，如文件名和文件长度。一般已central directory里面的数据为准。但文件压缩数据开始的地方，还是要从local file header里面获取。

文件数据的存储方式一般有2种，**STORE**和**DEFLATE**。

STORE表述文件数据直接存储在文件里面，没有结果压缩。

DEFLATE表示文件数据经过压缩后再存储到文件里面，压缩算法一般是LZ77。

**EOCD field data:**

|  |  |  |
| --- | --- | --- |
| **Offset** | **Bytes** | **Description**[[25]](https://en.wikipedia.org/wiki/Zip_(file_format)#cite_note-appnote-25) |
| 0 | 4 | End of central directory signature = 0x06054b50 |
| 4 | 2 | Number of this disk |
| 6 | 2 | Disk where central directory starts |
| 8 | 2 | Number of central directory records on this disk |
| 10 | 2 | Total number of central directory records |
| 12 | 4 | Size of central directory (bytes) |
| 16 | 4 | Offset of start of central directory, relative to start of archive |
| 20 | 2 | Comment length (*n*) |
| 22 | *n* | Comment |

**Central Directory file header**

|  |  |  |
| --- | --- | --- |
| **Offset** | **Bytes** | **Description**[[25]](https://en.wikipedia.org/wiki/Zip_(file_format)#cite_note-appnote-25) |
| 0 | 4 | Central directory file header signature = 0x02014b50 |
| 4 | 2 | Version made by  upper byte:  0 - MS-DOS and OS/2 (FAT / VFAT / FAT32 file systems)  1 - Amiga  2 - OpenVMS  3 - UNIX  4 - VM/CMS  5 - Atari ST  6 - OS/2 H.P.F.S.  7 - Macintosh  8 - Z-System  9 - CP/M  10 - Windows NTFS  11 - MVS (OS/390 - Z/OS)  12 - VSE  13 - Acorn Risc  14 - VFAT  15 - alternate MVS  16 - BeOS  17 - Tandem  18 - OS/400  19 - OS/X (Darwin)  20 - 255: unused  lower byte:  zip specification version |
| 6 | 2 | Version needed to extract (minimum) |
| 8 | 2 | General purpose bit flag |
| 10 | 2 | Compression method  00: no compression  01: shrunk  02: reduced with compression factor 1  03: reduced with compression factor 2  04: reduced with compression factor 3  05: reduced with compression factor 4  06: imploded  07: reserved  08: deflated  09: enhanced deflated  10: PKWare DCL imploded  11: reserved  12: compressed using BZIP2  13: reserved  14: LZMA  15-17: reserved  18: compressed using IBM TERSE  19: IBM LZ77 z  98: PPMd version I, Rev 1 |
| 12 | 2 | File last modification time  stored in standard MS-DOS format:  Bits 00-04: seconds divided by 2  Bits 05-10: minute  Bits 11-15: hour |
| 14 | 2 | File last modification date  stored in standard MS-DOS format:  Bits 00-04: day  Bits 05-08: month  Bits 09-15: years from 1980 |
| 16 | 4 | CRC-32 |
| 20 | 4 | Compressed size |
| 24 | 4 | Uncompressed size |
| 28 | 2 | File name length (*n*) |
| 30 | 2 | Extra field length (*m*) |
| 32 | 2 | File comment length (*k*) |
| 34 | 2 | Disk number where file starts |
| 36 | 2 | Internal file attributes |
| 38 | 4 | External file attributes |
| 42 | 4 | Relative offset of local file header. This is the number of bytes between the start of the first disk on which the file occurs, and the start of the local file header. This allows software reading the central directory to locate the position of the file inside the .ZIP file. |
| 46 | *n* | File name |
| 46+*n* | *m* | Extra field |
| 46+*n*+*m* | *k* | File comment |

* External file attributes

查看zip30里面unix/ unix.c函数filetime()

表示文件的参数信息，如果version made by是unix，则这个字段的值为：

((ulg)s.st\_mode << 16) | !(s.st\_mode & S\_IWUSR)

即高16位为文件信息，可以根据st\_mode判断文件是目录/链接文件/socket等。

在glibc source code bits/stat.h里面找到st\_mode为\_\_mode\_t类型，

\_\_mode\_t为\_\_MODE\_T\_TYPE，实际上是\_\_U32\_TYPE，为一个32位的数据。

#define \_\_S\_IFMT 0170000 /\* These bits determine file type. \*/

/\* File types. \*/

#define \_\_S\_IFDIR 0040000 /\* Directory. \*/

#define \_\_S\_IFCHR 0020000 /\* Character device. \*/

#define \_\_S\_IFBLK 0060000 /\* Block device. \*/

#define \_\_S\_IFREG 0100000 /\* Regular file. \*/

#define \_\_S\_IFIFO 0010000 /\* FIFO. \*/

stat.h里面有提供api来判断st\_mode，如：

S\_ISDIR（st\_mode）==》 判断文件是否为目录

S\_ISREG(mode) ==》 判断是否为普通文件

S\_ISLNK(mode) ==》 判断是否为链接文件

**local file header**

|  |  |  |
| --- | --- | --- |
| **Offset** | **Bytes** | **Description**[[25]](https://en.wikipedia.org/wiki/Zip_(file_format)#cite_note-appnote-25) |
| 0 | 4 | Local file header signature = 0x04034b50 (read as a little-endian number) |
| 4 | 2 | Version needed to extract (minimum) |
| 6 | 2 | General purpose bit flag  Bit 00: encrypted file  Bit 01: compression option  Bit 02: compression option  Bit 03: data descriptor  Bit 04: enhanced deflation  Bit 05: compressed patched data  Bit 06: strong encryption  Bit 07-10: unused  Bit 11: language encoding  Bit 12: reserved  Bit 13: mask header values  Bit 14-15: reserved |
| 8 | 2 | Compression method  00: no compression  01: shrunk  02: reduced with compression factor 1  03: reduced with compression factor 2  04: reduced with compression factor 3  05: reduced with compression factor 4  06: imploded  07: reserved  08: deflated  09: enhanced deflated  10: PKWare DCL imploded  11: reserved  12: compressed using BZIP2  13: reserved  14: LZMA  15-17: reserved  18: compressed using IBM TERSE  19: IBM LZ77 z  98: PPMd version I, Rev 1 |
| 10 | 2 | File last modification time  stored in standard MS-DOS format:  Bits 00-04: seconds divided by 2  Bits 05-10: minute  Bits 11-15: hour |
| 12 | 2 | File last modification date  stored in standard MS-DOS format:  Bits 00-04: day  Bits 05-08: month  Bits 09-15: years from 1980 |
| 14 | 4 | CRC-32 |
| 18 | 4 | Compressed size |
| 22 | 4 | Uncompressed size |
| 26 | 2 | File name length (*n*) |
| 28 | 2 | Extra field length (*m*) |
| 30 | *n* | File name |
| 30+*n* | *m* | Extra field |