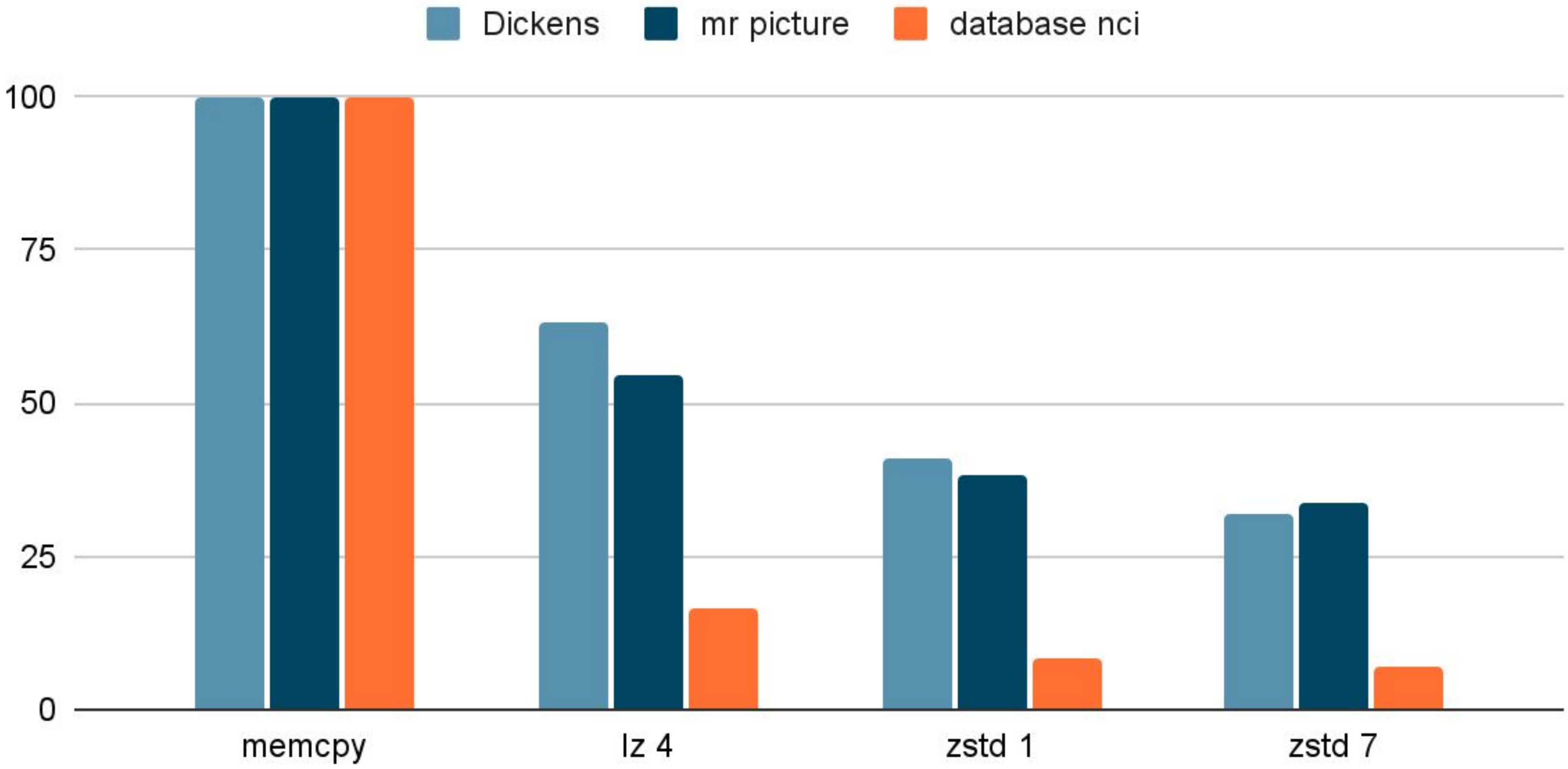
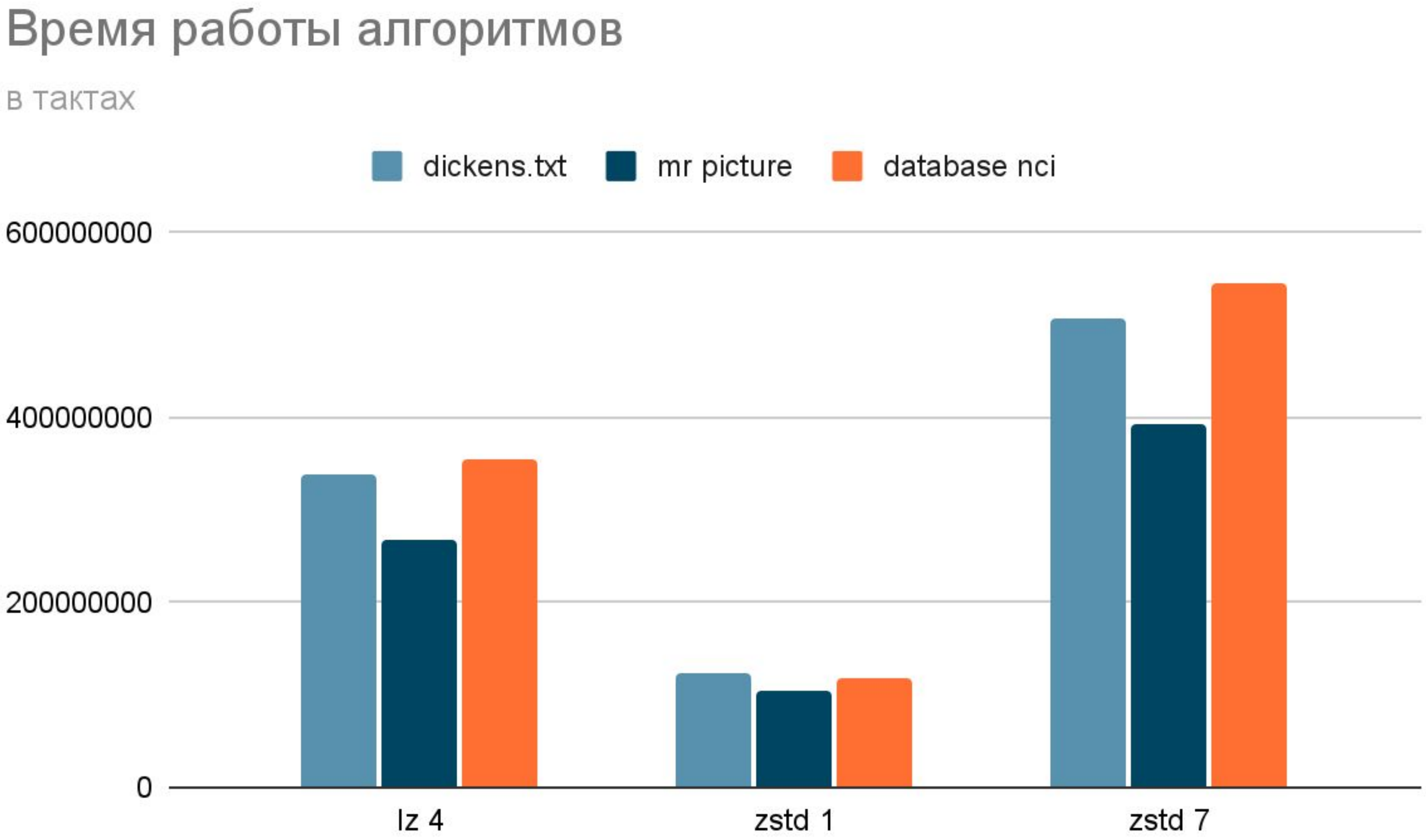
Сравнение алгоритмов сжатия данных zstd и lz4

Кудрявцев Федор
Зорабов Георгий

Размер сжатого файла dickens.txt

в процентах



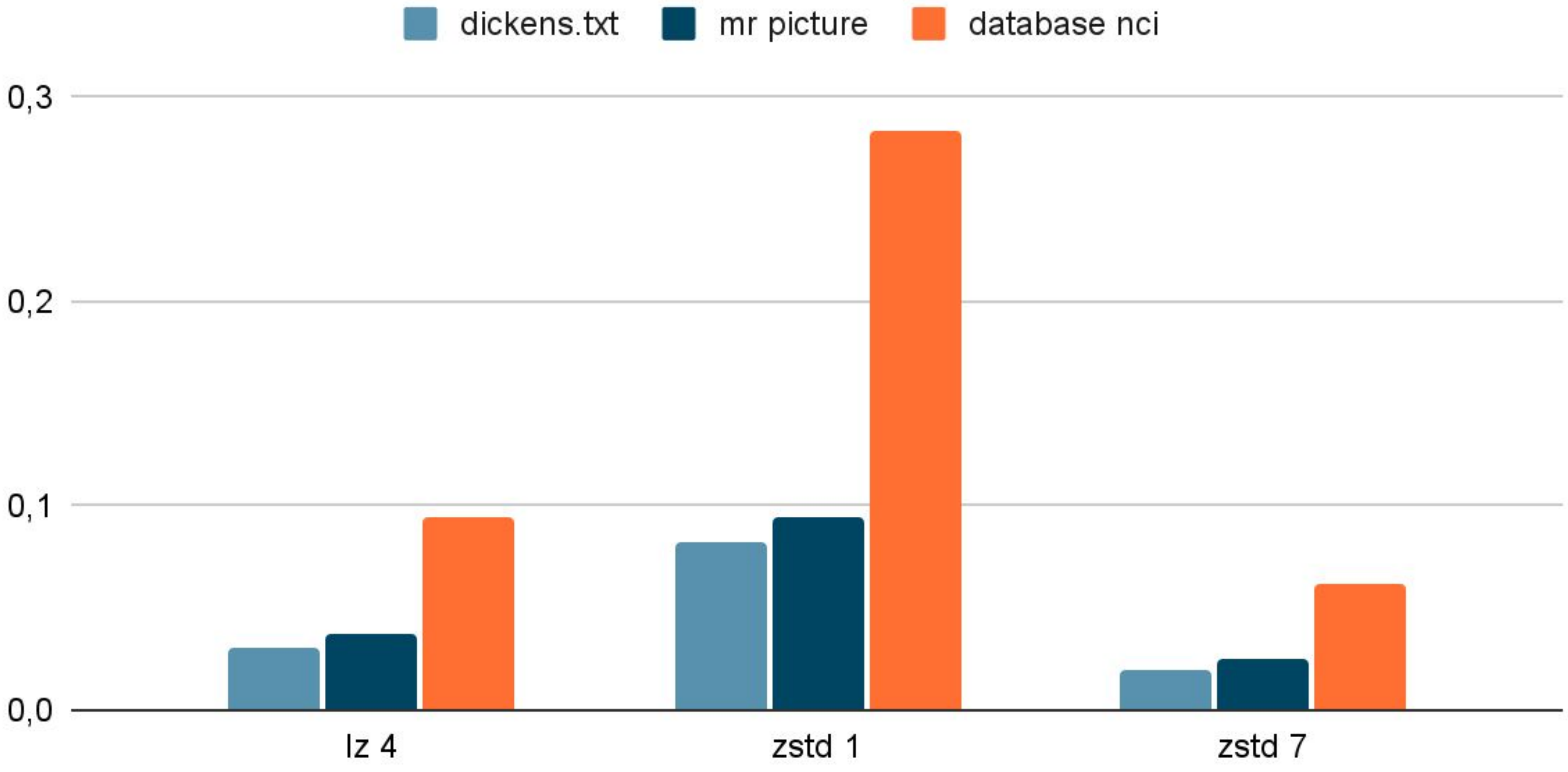


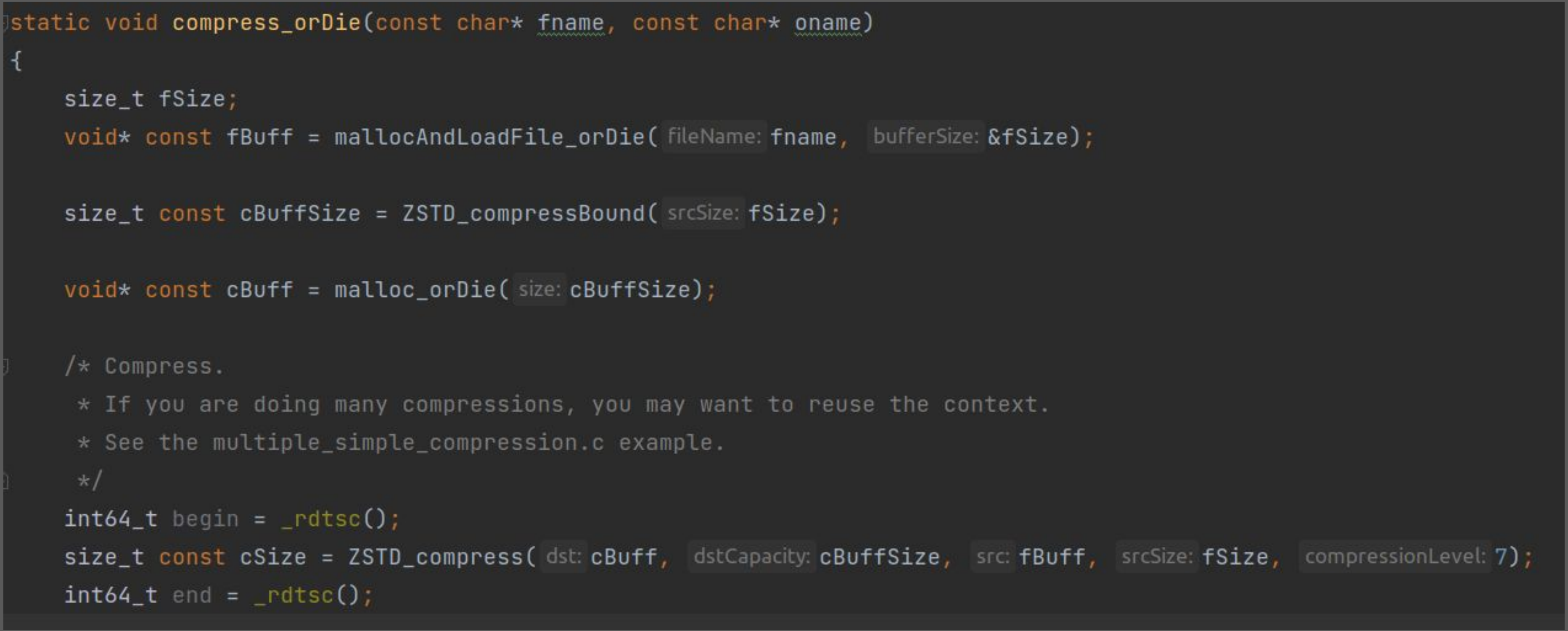
Время работы алгоритмов

в тактах

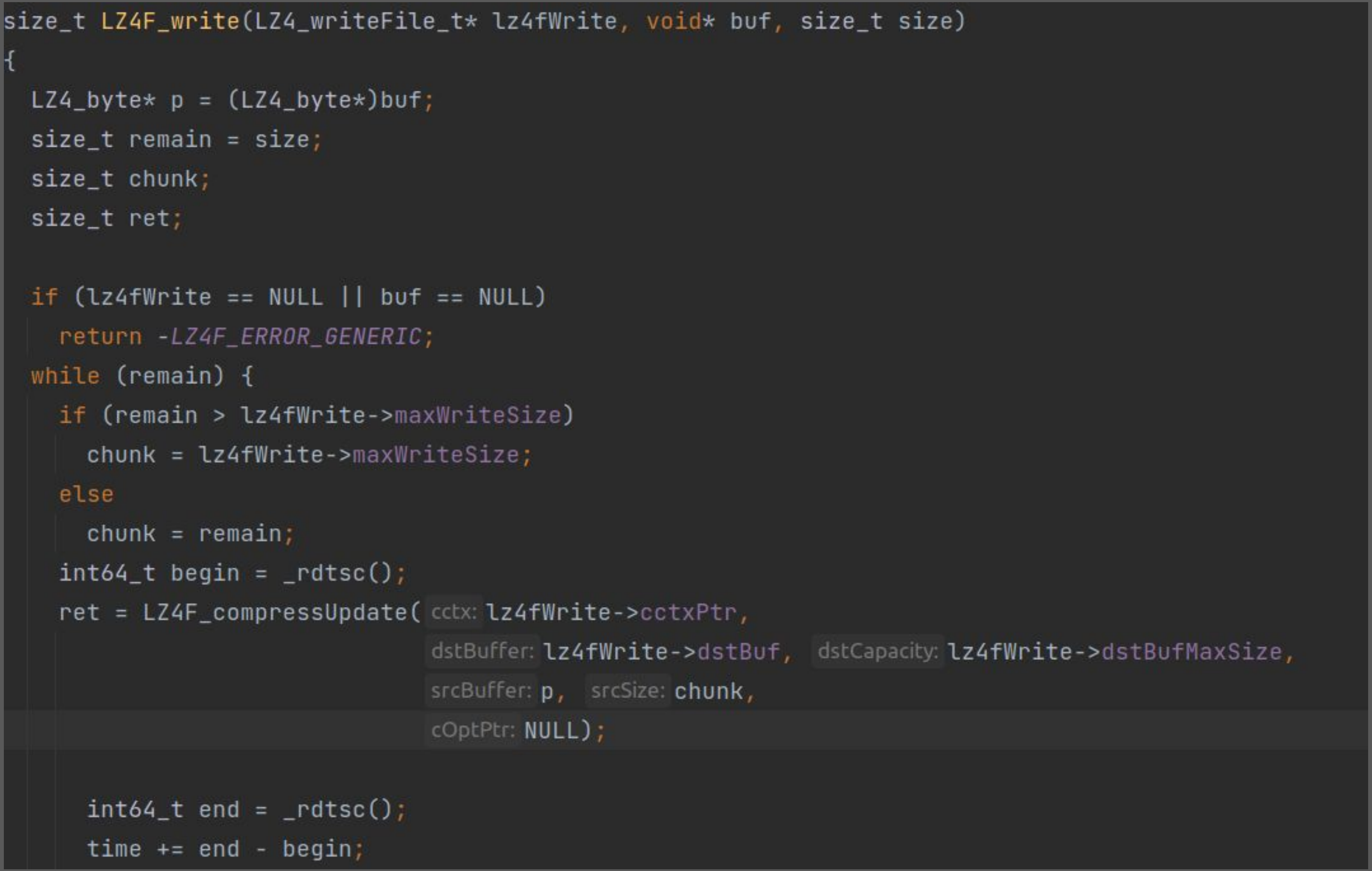
Скорость алгоритмов

байт/такт



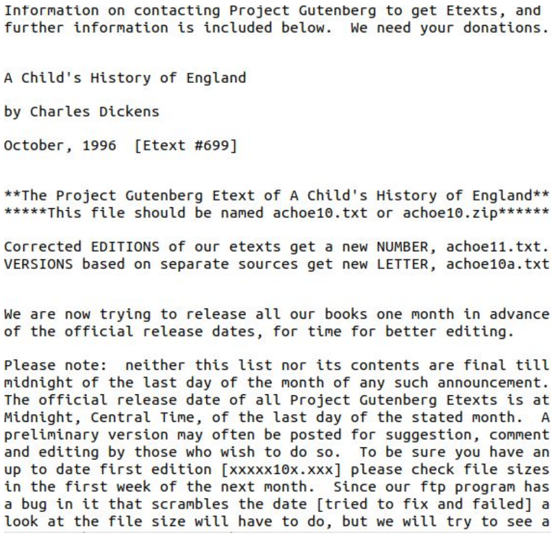


```
static void compress_orDie(const char* fname, const char* oname)
{
    size_t fSize;
    void* const fBuff = mallocAndLoadFile_orDie( fileName: fname,  bufferSize: &fSize);

    size_t const cBuffSize = ZSTD_compressBound( srcSize: fSize);

    void* const cBuff = malloc_orDie( size: cBuffSize);

    /* Compress.
     * If you are doing many compressions, you may want to reuse the context.
     * See the multiple_simple_compression.c example.
     */
    int64_t begin = _rdtsc();
    size_t const cSize = ZSTD_compress( dst: cBuff,  dstCapacity: cBuffSize,  src: fBuff,  srcSize: fSize,  compressionLevel: 7);
    int64_t end = _rdtsc();
```

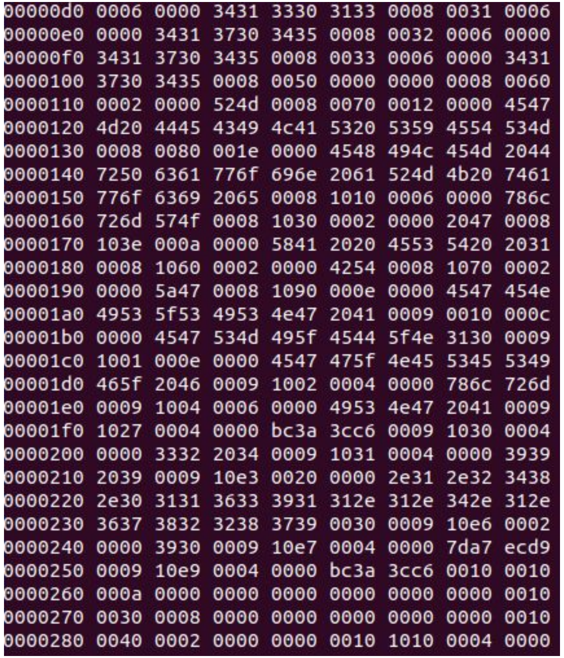


```
size_t LZ4F_write(LZ4_writeFile_t* lz4fWrite, void* buf, size_t size)
{
  LZ4_byte* p = (LZ4_byte*)buf;
  size_t remain = size;
  size_t chunk;
  size_t ret;

  if (lz4fWrite == NULL || buf == NULL)
    return -LZ4F_ERROR_GENERIC;
  while (remain) {
    if (remain > lz4fWrite->maxWriteSize)
      chunk = lz4fWrite->maxWriteSize;
    else
      chunk = remain;
    int64_t begin = _rdtsc();
    ret = LZ4F_compressUpdate( cctx: lz4fWrite->cctxPtr,
                              dstBuffer: lz4fWrite->dstBuf,  dstCapacity: lz4fWrite->dstBufMaxSize,
                              srcBuffer: p,  srcSize: chunk,
                              cOptPtr: NULL);

      int64_t end = _rdtsc();
      time += end - begin;
```

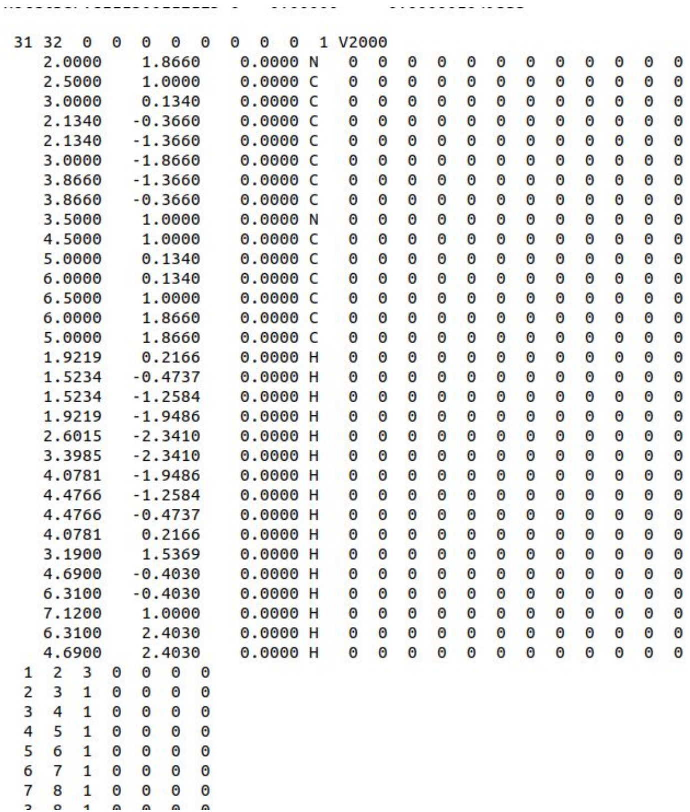
выводы



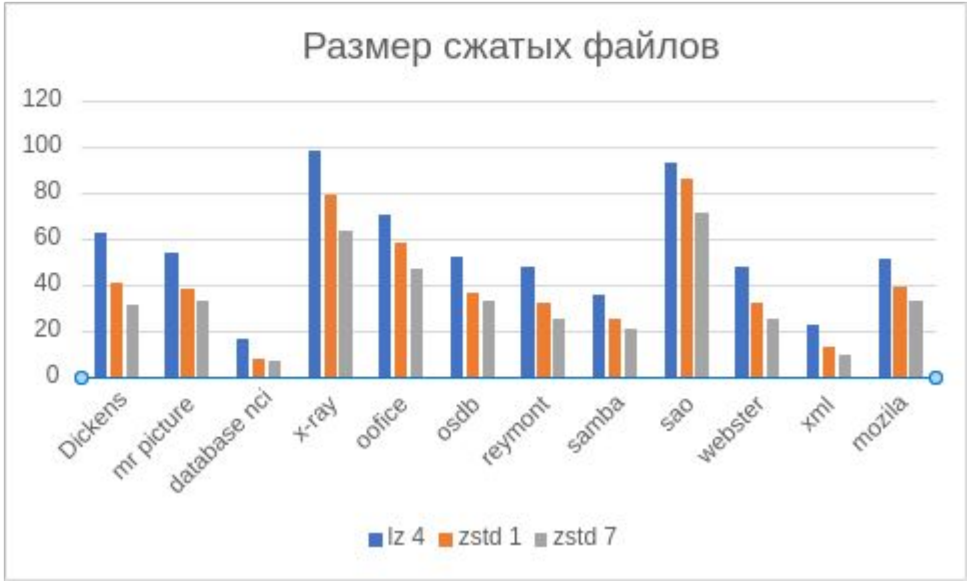
```
Information on contacting Project Gutenberg to get Etexts, and
further information is included below.  We need your donations.


A Child's History of England

by Charles Dickens

October, 1996  [Etext #699]


**The Project Gutenberg Etext of A Child's History of England**
*****This file should be named achoe10.txt or achoe10.zip******

Corrected EDITIONS of our etexts get a new NUMBER, achoe11.txt.
VERSIONS based on separate sources get new LETTER, achoe10a.txt


We are now trying to release all our books one month in advance
of the official release dates, for time for better editing.

Please note:  neither this list nor its contents are final till
midnight of the last day of the month of any such announcement.
The official release date of all Project Gutenberg Etexts is at
Midnight, Central Time, of the last day of the stated month.  A
preliminary version may often be posted for suggestion, comment
and editing by those who wish to do so.  To be sure you have an
up to date first edition [xxxxx10x.xxx] please check file sizes
in the first week of the next month.  Since our ftp program has
a bug in it that scrambles the date [tried to fix and failed] a
look at the file size will have to do, but we will try to see a
```

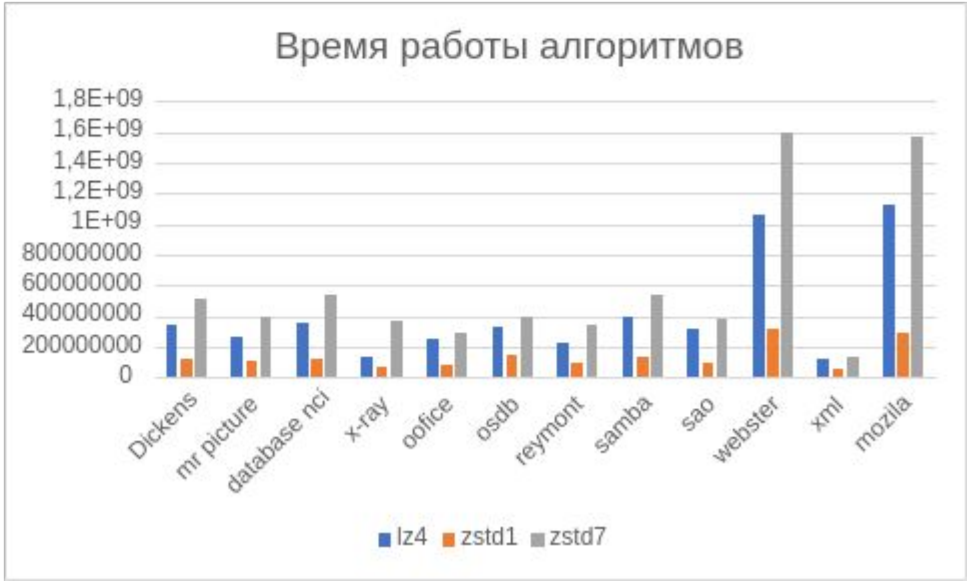


```
00000d0 0006 0000 3431 3330 3133 0008 0031 0006
00000e0 0000 3431 3730 3435 0008 0032 0006 0000
00000f0 3431 3730 3435 0008 0033 0006 0000 3431
0000100 3730 3435 0008 0050 0000 0000 0008 0060
0000110 0002 0000 524d 0008 0070 0012 0000 4547
0000120 4d20 4445 4349 4c41 5320 5359 4554 534d
0000130 0008 0080 001e 0000 4548 494c 454d 2044
0000140 7250 6361 776f 696e 2061 524d 4b20 7461
0000150 776f 6369 2065 0008 1010 0006 0000 786c
0000160 726d 574f 0008 1030 0002 0000 2047 0008
0000170 103e 000a 0000 5841 2020 4553 5420 2031
0000180 0008 1060 0002 0000 4254 0008 1070 0002
0000190 0000 5a47 0008 1090 000e 0000 4547 454e
00001a0 4953 5f53 4953 4e47 2041 0009 0010 000c
00001b0 0000 4547 534d 495f 4544 5f4e 3130 0009
00001c0 1001 000e 0000 4547 475f 4e45 5345 5349
00001d0 465f 2046 0009 1002 0004 0000 786c 726d
00001e0 0009 1004 0006 0000 4953 4e47 2041 0009
00001f0 1027 0004 0000 bc3a 3cc6 0009 1030 0004
0000200 0000 3332 2034 0009 1031 0004 0000 3939
0000210 2039 0009 10e3 0020 0000 2e31 2e32 3438
0000220 2e30 3131 3633 3931 312e 312e 342e 312e
0000230 3637 3832 3238 3739 0030 0009 10e6 0002
0000240 0000 3930 0009 10e7 0004 0000 7da7 ecd9
0000250 0009 10e9 0004 0000 bc3a 3cc6 0010 0010
0000260 000a 0000 0000 0000 0000 0000 0000 0010
0000270 0030 0008 0000 0000 0000 0000 0000 0010
0000280 0040 0002 0000 0000 0010 1010 0004 0000
```

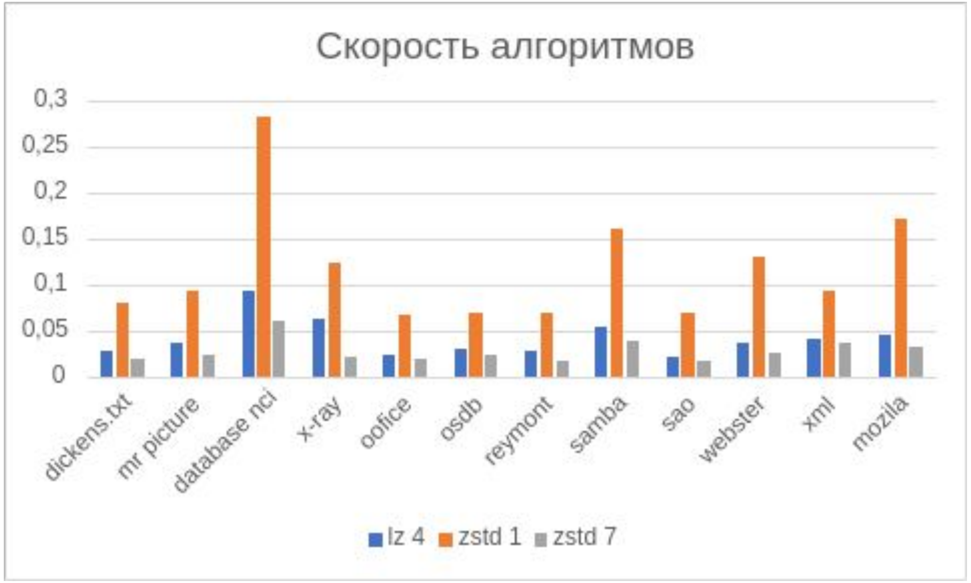


```
 31 32  0  0  0  0  0  0  0  0  1 V2000
    2.0000    1.8660    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    2.5000    1.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.0000    0.1340    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.1340   -0.3660    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.1340   -1.3660    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.0000   -1.8660    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.8660   -1.3660    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.8660   -0.3660    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.5000    1.0000    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    4.5000    1.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.0000    0.1340    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    6.0000    0.1340    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    6.5000    1.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    6.0000    1.8660    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    5.0000    1.8660    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.9219    0.2166    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.5234   -0.4737    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.5234   -1.2584    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.9219   -1.9486    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    2.6015   -2.3410    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    3.3985   -2.3410    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    4.0781   -1.9486    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    4.4766   -1.2584    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    4.4766   -0.4737    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    4.0781    0.2166    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    3.1900    1.5369    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    4.6900   -0.4030    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    6.3100   -0.4030    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    7.1200    1.0000    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    6.3100    2.4030    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
    4.6900    2.4030    0.0000 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  3  0  0  0  0
  2  3  1  0  0  0  0
  3  4  1  0  0  0  0
  4  5  1  0  0  0  0
  5  6  1  0  0  0  0
  6  7  1  0  0  0  0
  7  8  1  0  0  0  0
```

Размер сжатых файлов
120
100
80
60
40
20
0
Dickens
mr picture
database nci
x-ray
oofice
osdb
reymont
samba
sao
webster
xml
mozila
lz 4
zstd 1
zstd 7



Время работы алгоритмов


Скорость алгоритмов
0,3
0,25
0,2
0,15
0,1
0,05
0
dickens.txt
mr picture
database nci
x-ray
oofice
osdb
reymont
samba
sao
webster
xml
mozila
lz 4
zstd 1
zstd 7

выводы

Общие наблюдения про сравнение скорости и качества сжатия для разных алгоритмов сохранились.

Лучше всего сжались текст в pdf и базы данных в текстовом виде из-за наличия большого количества повторяющихся символов. Также неплохо сжался html из-за хорошей структурированности, за счет которой появляются похожие отрезки.

Плохо сжались бинарники.