Tiny Data Compression with td512

L. S. Leonard, February 13, 2022

# Introduction

Tiny data compression for as few as 64 bytes is not supported by standard compression programs. Now with td512 you can reasonably compress data from 16 to 512 bytes. td512 is available under the GPL-3.0 License at <https://github.com/lsleonard/tiny-data-compression>. Compared with QuickLZ, a fast compression program that is designed to compress smaller data sets, td512 gets as good or better compression for 512-byte blocks of most data types. QuickLZ steadily declines in compression ratio as the number of bytes decreases to 128, and at 64 bytes, produces compression only for highly compressible files. td512 has good compression at 64 bytes with the td64 interface. td512 combines extended text and string modes for 128 to 512 bytes and uses the td64 interface to compress any remaining bytes in the input. The td512 algorithm emphasizes speed, and based on data in this paper, gets 32% average compression for 512-byte blocks at 300 Mbytes per second on the Squash benchmark test data (see [https://quixdb.github.io/squash-benchmark/#](https://quixdb.github.io/squash-benchmark/)) running on a 2 GHz quad-core processor. For 64-byte blocks on this benchmark data, td512 gets 25% average compression at 260 Mbytes per second.

The Zstandard algorithm produces excellent compression at a good speed for 512-byte blocks using LZ77 and Huffman coding on the Squash benchmark data (see Table 3). But the average compression speed of Zstandard is 100 Mbytes per second, 1/3 that of td512. And like QuickLZ, produces compression only for highly compressible files for 64-byte blocks. Although Huffman coding, with its optimal compression using frequency analysis of values, works well for larger datasets, for tiny datasets the compression modes used in td512 approach or exceed the results of using the Huffman algorithm, and at a higher speed. With the focus on speed of execution, arithmetic coding is not a practical algorithm for applications of tiny data. One area where high-speed compression using td512 might be applied is in a heap manager, an example of where a variety of data types might be encountered.

For compression and speed comparison with td512, this paper presents compression performance using QuickLZ, where its streaming mode supports compression of as few as 200 to 300 characters ([quicklz.com](http://quicklz.com/)). The final section of the paper discusses the fundamentals of td64, the compression interface that the td512 algorithm uses. The appendix shows performance data for 8 to 512 bytes. Also presented in the appendix are compression values from running Zstandard v1.5.1 (<http://facebook.github.io/zstd/>) and the arithmetic encoding program fpaq0 (<http://mattmahoney.net/dc/fpaq0.cpp>) on the Squash benchmark test data using 512-byte blocks (Table 3). These values give a picture of the compression achievable when runtime is less critical to the application.

# Comparison of Compression Performance: td512 and QuickLZ

The td512 algorithm is packaged with a testbed that runs the algorithm iteratively over an input file using a fixed block size until a final block, possibly smaller, concludes the run (see <https://github.com/lsleonard/tiny-data-compression>). In addition to running td512 with a 512-byte block size, the QuickLZ public distribution of version 1.5.1 Beta 7 was modified to iterate over a 512-byte block size. Both programs were run on the Squash benchmark test data on a MacBook with a 2 GHz Quad-Core Intel Core i5 processor. The test beds for td512 and QuickLZ tests read file data into memory and compute loop count based on file size to pick the best (fastest) compression and decompression speed.

Compression performance for td512 v2.1.5 and QuickLZ is shown in Table 1. Across all the benchmark files, average compression percent, assuming a block of data from one file is as likely as from another, for td512 is 32%, and for QuickLZ is 22%. QuickLZ gets better compression for file geo.protodata, and nci and ptt3 are only slightly better than td512. The file geo.protodata has 512-byte blocks with more than 64 unique values, which is the limit that is currently handled by td512 extended string mode. Other files with fewer unique values that primarily use extended string mode get similar results to QuickLZ because string mode is an algorithm that finds repeated strings, just as QuickLZ does. Text files compress well with the td64 text mode algorithm that uses weighted encoding of 23 predefined characters. Also, the files sum and x-ray are much better compressed by td512 than QuickLZ because these files contain frequently repeated values in small blocks. For example, 16 or more 0s in a 64-byte block that are compressed using a td64 compression mode called single value. Blocks that use either the text or single value modes get similar compression results for blocks from 64 to 512 bytes.

On average, td512 compresses about 9% faster and decompresses about 28% slower than QuickLZ on the Squash benchmark files. But with 45% better average compression, this is a good tradeoff for td512.

Table 1. Compression Performance Data: td512 v2.1.7 and QuickLZ with 512-Byte Blocks

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Squash Benchmark  Filename | td512  Compr.  Percent | QuickLZ  Compr.  Percent | td512  Compress  MB/sec | QuickLZ  Compress  MB/sec | td512  Decompress  MB/Sec | QuickLZ  Decompress  MB/Sec |
| alice29.txt | 36.89 | 8.43 | 265889872 | 224983712 | 222677888 | 485827488 |
| asyoulik.txt | 32.37 | 6.93 | 223933808 | 217324640 | 228428832 | 547929792 |
| cp.html | 30.31 | 19.25 | 242394080 | 264548384 | 256281248 | 540131840 |
| dickens | 38.07 | 4.67 | 276825728 | 226428352 | 227611568 | 621070208 |
| fields.c | 40.58 | 33.98 | 437254912 | 441584160 | 459793824 | 494344832 |
| fireworks.jpeg | -0.54 | -1.56 | 1079763200 | 438053408 | 20515500032 | 17554284544 |
| geo.protodata | 9.24 | 13.82 | 206599296 | 263528896 | 472462176 | 679724160 |
| grammar.lsp | 42.68 | 40.15 | 488616160 | 469640800 | 520139776 | 541767424 |
| kennedy.xls | 62.42 | 59.87 | 405730496 | 429776288 | 535488320 | 698529152 |
| lcet10.txt | 36.45 | 8.31 | 257235696 | 225081216 | 225081216 | 551028416 |
| mozilla | 31.53 | 30.86 | 204605296 | 259013712 | 333299584 | 630041408 |
| mr | 44.44 | 30.15 | 229778864 | 257770544 | 630928576 | 653567744 |
| nci | 56.48 | 62.05 | 262119920 | 323940608 | 279656320 | 754891328 |
| ooffice | 16.11 | 12.18 | 150527072 | 216359840 | 222454144 | 557213312 |
| osdb | 1.35 | -1.74 | 182107936 | 265370832 | 640808448 | 10125879296 |
| paper-100k.pdf | 10.73 | 9.41 | 613173632 | 411244960 | 3011764736 | 3531034368 |
| plrabn12.txt | 38.25 | 3.52 | 271318112 | 223913088 | 231775376 | 617682048 |
| ptt5 | 73.2 | 74.94 | 267718304 | 368689664 | 897230784 | 866594560 |
| reymont | 31.44 | 25.51 | 151902496 | 230760192 | 171444880 | 376032224 |
| samba | 40.6 | 33.56 | 209993104 | 264891456 | 253786896 | 598432384 |
| sao | -0.59 | -1.75 | 839345408 | 319103424 | 13209369600 | 11907152896 |
| sum | 39.04 | 27.72 | 185631072 | 240503136 | 308387104 | 456481920 |
| urls.10K | 34.33 | 23.46 | 186675616 | 216159792 | 225896720 | 667889664 |
| webster | 28.35 | 17.60 | 165619520 | 217385648 | 182658384 | 457117056 |
| x-ray | 20.65 | -0.98 | 266025440 | 280883008 | 754741760 | 2247774976 |
| xargs.1 | 27.61 | 18.31 | 460366336 | 418891872 | 394042368 | 489739168 |
| xml | 43.5 | 35.74 | 186220736 | 255693856 | 208360496 | 479010656 |
| Average  Performance | td512  32.05 | QuickLZ  22.01 | td512  322865634 | QuickLZ  295241685 | td512  1689632261 | QuickLZ  2153006402 |

# Fundamentals of td512 and td64 Interfaces

With the td512 interface, you can call the td512 and td512d functions to compress and decompress 1 to 512 bytes. The td512 interface performs compression of 16 to 512 bytes, but accepts 1 to 15 bytes and stores them without compression. Along with its extended text and string modes, td512 acts as a wrapper that uses the td64 interface to compress blocks of 64 bytes until the final block of 64 or fewer bytes is compressed. The number of bytes processed and a pass/fail bit is stored for each block compressed, and the compressed or uncompressed data is output.

With td64, you can call the td5 and td5d functions to compress and decompress 1 to 5 values. This interface is not used by td512 because the number of bytes generated is often more than the number of values to compress. Compression of these miniscule datasets requires bit handling not supported by td512. You can call td64 and td64d functions to compress and decompress 1 to 64 values. For fewer than 6 bytes, td64 calls the td5 interface. The td64 interface returns pass (number of compressed bits) or fail (0) and outputs only compressed values. Decompression requires input of the number of original values and data that successfully compressed.

## Encoding Used by the td512 Interface

The td512 interface uses the text mode that is part of the td64 interface, an extended string mode, and td64 to compress data in memory. For 1 to 127 values, one block of 64 followed by the remaining values are processed by td64. For 128 to 512 values, the algorithm verifies that of the first 96 values, 94% are standard text characters and that ¾ are predefined characters. After text mode is confirmed, all values are checked for the high bit clear. Only the standard text mode is used. If text mode is not selected, the first 64 values are scanned for data that would be best compressed by td64. If only 1 or 2 or more than 40 unique values, or 35 or more repeats of a single value are found, then td64 is selected. If a single value occurs 18 times and there are more than 14 unique values, then both extended string mode and td64 are called on these 64 values and the mode that compresses best is selected. If extended string mode is called, it runs until 64 unique values are encountered. It then continues processing strings that start with values from these first 64 unique values until up to 64 more values not in the first 64 are encountered. If text mode or string mode fails, or string mode compresses fewer than all values, the remaining values are handed off in blocks of 64, or possibly fewer for the final block, to td64.

## Encoding Used by the td64 Interface

The td64 interface integrates the following encoding modes: fixed bit coding, text mode, single value mode, string mode, extended string mode, and 7-bit mode. Initially, the program searches up to 1/2 of the input values for unique values and count, accumulates a high-bit value, and counts the number of frequently occurring text characters. If more than the accepted limit of unique values is encountered, then if all characters have the high bit clear, 7-bit mode is used; otherwise, the program fails, assuming random data for this block. If the values match to indicate text data, text mode is used if it gets at least 11% compression. Processing continues by looking for any single value that occurs 25% percent of the time in addition to looking for new unique values. If the limit of unique values is exceeded, single value, if active, is used. If fewer than the limit of unique values is encountered, single value mode, if active, is used first if there are 5 or more unique values; otherwise, fixed bit coding is used. When the limit of unique values is exceeded, single value mode is used first, then string mode is used to encode repeated strings of two or more characters if no more than 32 unique values occurred. If string mode fails and all characters have the high bit clear, 7-bit mode is used. If all modes fail, the program fails to compress the input data. Each of the td64 compression modes is described next.

Fixed bit coding is well known as using only the number of bits required to encode the number of unique characters in a dataset. The td64 algorithm predetermines the maximum number of unique values in the data that will allow compression to occur. With these limits known ahead of time (see uniqueLimits25 in td64.c), decisions about compressibility can be made quickly. This mode is programmed to get 25% (16 unique values) to 83% (2 unique values) compression for 64 input values.

Text mode uses the most frequent characters as defined by Morse code plus space, carriage return and comma to identify text data that can be compressed (see Table 2). When 90 percent of the data values are one of these characters, as in most standard text, text mode gets 35% compression for 64 input values. From 3 to 7 bits are output for each character, based on frequency of occurrence. The adaptive text mode implemented in td64 supports replacement of the 8 lowest frequency characters with characters that are likely to occur in non-standard text. At present, XML (or HTML) and C code characters are handled (see XMLTextChars and CTextChars in td64.c).

Table 2: Most Frequently Occurring Text Characters Plus Space, Carriage Return, and Comma

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| # | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
| Bits | 3 | 3 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 7 | 7 | 7 | 7 |
| Char | sp | e | t | a | i | n | o | s | h | r | d | l | u | c | cr | m | g | f | , | y | w | p | b |

Single value mode encodes a value that occurs at least 25% of the time in the data block as a 1 bit and all other values as a 0 bit plus an 8-bit value. This results in 11% compression for 64 input values, but is often much higher as checking for a single value is stopped when the number of occurrences reaches the minimum. When the data block being compressed contains fewer unique values than the unique limit, extended string mode is used to compress the non-single values.

String mode encodes repeated strings of at least two characters in length for data blocks with at most 32 unique values. Single repeated characters are also encoded and save at least one bit. The algorithm uses the first position where unique characters occur as the beginning for repeated strings, which requires at most 5 bits. This algorithm is the most time-intensive of the modes and is used after all but the 7-bit mode. String mode output is used if it gets at least 6% compression, or 12% when 7-bit mode can be used, for 64 input values. String mode encodes the unique values, when all high bits are 0s, by using 7-bit mode.

Extended string mode encodes repeated strings of at least two characters in length, as well as single repeated characters, for data blocks with up to 64 unique values. This mode functions similarly to string mode, but records all two-value occurrences rather than just the first one for each unique value. This means that locations for repeated strings require up to 9 bits for 512-character blocks and for repeated characters up to 6 bits. Extended string mode continues to function once 64 unique values are reached until up to 64 additional values not in the first 64 are encountered. Within the td64 interface, extended string mode is used only by single value mode.

7-bit mode encodes the 7 lower bits of each value, leaving off the high zero bit, in groups of 7 bytes. This mode gets 11% compression for 64 input values and is used last because other modes can get higher compression. 7-bit mode requires at least 16 input values.

## Encoding Used by the td5 Interface

The td5 interface uses three modes to encode 1 to 5 values:

* For 4 or 5 input values, encode 1 or 2 unique values.
* For 1 to 5 input values, encode frequently occurring text characters (see text mode below).
* For 2 or 3 input values, encode 2 unique 4-bit nibbles.

# Summary

This paper has shown that over the Squash benchmark data, td512 compresses 512-byte blocks at 32%, on average, while QuickLZ gets 22%. Although QuickLZ has much better decompression speed, td512 gets 47% better compression. The compression achieved by Zstandard is the best presented in this paper at 36.9%, but for most files the program takes at least twice as long to run as td512. A huge benefit of td512 is that its compression for 64 bytes, a number of values that very few compression programs can support, is 25%. Both QuickLZ and Zstandard produce limited compression for 128-byte blocks, and for 64-byte blocks only 4 benchmark files are compressed (table xxx), while td512 with 16-byte blocks, compresses all but three of the files that are compressible with 512-byte blocks. The arithmetic compressor fpaq0 supports 64-byte blocks and averages 26% compression (table xxx). Over the Squash benchmark data, td512 gets close to the 34% average compression using fpaq0 with 512-byte blocks, showing that td512 closely approaches the compression achieved by this form of arithmetic encoding.

The implementation of td512 uses fixed bit coding, text, single value, string, and 7-bit modes to compress data. This variety of encoding modes reflects the fact that data compression is data dependent. The td64 interface can be easily modified to support additional compression modes for special data types. New modes in the td64 interface will not affect the td512 interface.

## Appendix: Performance Data for td512

Table 3 shows performance data for td512 v2.1.4 run iteratively with 128- to 512-byte blocks, and fpaq0 and Zstandard v1.5.1 run with 512-byte blocks on the Squash benchmark data on a 2 GHz quad-core processor with percent compression and MB per second.

Table 3: Performance Data for td512 v2.1.4 on 128- to 512-byte Blocks and fpaq0 and zstd with 512-byte Blocks Run on the Squash Benchmark Test Data

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Filename | 128 | 128 Comp | 128 Decomp | 256 | 256 Comp | 256 Decomp | 512 | 512 Comp | 512 Decomp | fpaq0 | zstd |
| alice29.txt | 34.64 | 223331856 | 167130768 | 36.04 | 237268336 | 168987776 | 36.60 | 255182880 | 172046384 | 39.19 | 36.09 |
| asyoulik.txt | 30.26 | 207250000 | 168932528 | 31.55 | 238891232 | 171243504 | 32.05 | 245931248 | 179596848 | 35.94 | 34 |
| cp.html | 13.16 | 136683328 | 201663936 | 23.54 | 180904416 | 234314288 | 30.67 | 210282048 | 234314288 | 30.83 | 38.49 |
| dickens | 36.26 | 250872448 | 173683568 | 37.54 | 265843664 | 171957648 | 37.97 | 275545984 | 173234864 | 39.61 | 35.43 |
| fields.c | 28.61 | 202727264 | 318571424 | 35.79 | 285897440 | 384482752 | 40.91 | 359677440 | 446000000 | 36.10 | 46.22 |
| fireworks.jpeg | -1.51 | 559513600 | 13676999680 | -0.74 | 715656960 | 17584713728 | -0.54 | 860790208 | 24618600448 | -3.69 | -1.72 |
| geo.protodata | 7.38 | 177526944 | 484032640 | 10.80 | 195046064 | 401993216 | 9.64 | 205525120 | 434388256 | 9.84 | 21.75 |
| grammar.lsp | 32.30 | 265785712 | 413444416 | 40.07 | 286230784 | 413444416 | 43.11 | 338272736 | 465124992 | 39.51 | 53.24 |
| kennedy.xls | 56.16 | 271556960 | 461148224 | 63.03 | 327526720 | 517979872 | 62.74 | 378304192 | 520072736 | 59.72 | 68.27 |
| lcet10.txt | 35.79 | 224607360 | 174114240 | 36.43 | 245684512 | 176417520 | 36.49 | 258013296 | 175402384 | 39.13 | 36.84 |
| mozilla | 26.35 | 153839744 | 284960320 | 31.54 | 172823216 | 279589280 | 32.75 | 192302272 | 307287232 | 32.27 | 38.26 |
| mr | 45.61 | 154145056 | 557357248 | 45.79 | 194733776 | 555463168 | 44.45 | 221302528 | 593697984 | 56.16 | 49.38 |
| nci | 50.76 | 142713808 | 236925888 | 55.78 | 162516320 | 256303376 | 57.06 | 180862576 | 266430384 | 67.40 | 74.88 |
| ooffice | 13.74 | 124813696 | 199772432 | 16.87 | 135693152 | 192659392 | 16.61 | 146956624 | 211444592 | 21.63 | 21.21 |
| osdb | 0.03 | 157986240 | 798549824 | 1.11 | 174628752 | 541251648 | 1.52 | 204698192 | 679536704 | 13.39 | 5.2 |
| paper-100k.pdf | 7.79 | 270184704 | 1651612928 | 9.31 | 372363648 | 1896296320 | 10.39 | 455111104 | 2226086912 | 7.38 | 11.3 |
| plrabn12.txt | 36.45 | 244475376 | 167603824 | 37.79 | 262165936 | 176635264 | 38.27 | 265195936 | 169252192 | 39.08 | 34.06 |
| ptt5 | 72.20 | 200162256 | 794452032 | 72.92 | 235312240 | 909957440 | 73.22 | 264680768 | 926382656 | 84.02 | 80.69 |
| reymont | 26.01 | 113022752 | 151739024 | 28.36 | 125076944 | 157956000 | 31.65 | 139180144 | 168096432 | 37.05 | 38.7 |
| samba | 31.43 | 142977968 | 207606128 | 37.64 | 169080032 | 223703472 | 40.71 | 190148640 | 233161760 | 35.02 | 46.72 |
| sao | -1.56 | 422140064 | 8297418752 | -0.78 | 584315840 | 10058174464 | -0.58 | 679466368 | 11313485824 | 3.33 | -1.76 |
| sum | 35.58 | 132318344 | 277101440 | 38.71 | 156721312 | 287518816 | 39.16 | 171479824 | 294153856 | 43.49 | 43.42 |
| urls.10K | 22.84 | 133324536 | 192828064 | 30.99 | 157843312 | 211281072 | 34.88 | 173269248 | 206071904 | 32.54 | 43.54 |
| webster | 20.94 | 122053312 | 158053520 | 25.04 | 136448224 | 164450784 | 28.59 | 144456416 | 172701424 | 34.35 | 38.55 |
| x-ray | 20.03 | 174048352 | 660141760 | 19.87 | 213059792 | 646296512 | 21.02 | 263346896 | 801649792 | 26.12 | 15.86 |
| xargs.1 | 29.38 | 264187504 | 301928576 | 28.77 | 325153856 | 301928576 | 27.80 | 325153856 | 352249984 | 35.53 | 39.82 |
| xml | 33.52 | 132138832 | 177767136 | 40.03 | 153701584 | 193019184 | 43.98 | 178927488 | 212781344 | 35.16 | 48.47 |
| Block Size  Performance | 128  27.42 | 128 Comp.  207569927 | 128 Decomp.  1161316308 | 256  30.88 | 256 Comp.  248540299 | 256 Decomp  1380667388 | 512  32.26 | 512 Comp.  280891260 | 512 Decomp.  1724194525 | fpaq0  34.44 | zstd  36.92 |

Table 4 shows performance data for td512 v2.1.1 run iteratively with 32-, 64- and 96-byte blocks on the Squash benchmark test data on a 2 GHz quad-core processor. Performance is percent compression and MB per second. Processing of 96 values involves compression of the first 64 bytes followed by compression of the final 32 bytes. The compression achieved is approximately half the sum of the compression of 32- and 64-byte blocks.

Table 4: Performance Data for td512 v2.1.1 on 32-, 64- and 96-byte Blocks Run on the Squash Benchmark Test Data

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Filename | 32 | 32 Comp. | 32 Decomp. | 64 | 64 Comp. | 64 Decomp. | 96 | 96 Comp. | 96 Decomp. |
| alice29.txt | 30.54 | 176847680 | 148815072 | 34.31 | 174014864 | 149253200 | 33.09 | 165674288 | 152393776 |
| asyoulik.txt | 26.32 | 163419056 | 160898464 | 30.32 | 168251344 | 154541968 | 28.99 | 170080160 | 156669600 |
| cp.html | 16.73 | 174489360 | 198411296 | 21.30 | 195261904 | 179583952 | 19.87 | 173260560 | 161861856 |
| dickens | 31.47 | 172566128 | 148225728 | 35.28 | 162204544 | 142189760 | 34.01 | 158723760 | 145575184 |
| fields.c | 19.70 | 359677440 | 293421056 | 26.66 | 384482752 | 271951232 | 24.53 | 327941184 | 247777776 |
| fireworks.jpeg | -3.09 | 1043160960 | 5595136512 | -1.52 | 932522688 | 10257750016 | -2.03 | 961664064 | 10257750016 |
| geo.protodata | 0.43 | 230715952 | 1151339776 | 3.39 | 238128512 | 1185880064 | 2.35 | 221246272 | 1058821440 |
| grammar.lsp | 19.65 | 372100000 | 265785712 | 23.92 | 372100000 | 248066672 | 22.95 | 338272736 | 232562496 |
| kennedy.xls | 27.46 | 316357600 | 703857856 | 47.37 | 222406912 | 388143232 | 40.76 | 236234000 | 422199264 |
| lcet10.txt | 31.43 | 172356224 | 153619152 | 34.98 | 178558160 | 154621008 | 33.83 | 168477696 | 152412144 |
| mozilla | 11.85 | 224500576 | 830409408 | 19.39 | 212513712 | 524042912 | 16.86 | 184447360 | 581859200 |
| mr | 42.47 | 343599296 | 1160042368 | 47.12 | 220924944 | 647439232 | 45.32 | 256299520 | 698854976 |
| nci | 35.14 | 228088688 | 541490240 | 47.27 | 127943520 | 245902864 | 43.26 | 143883312 | 288882752 |
| ooffice | 2.08 | 172349616 | 923198080 | 7.27 | 158133712 | 474852736 | 5.52 | 160355312 | 543912256 |
| osdb | 2.43 | 161275472 | 925631808 | 0.57 | 218271776 | 3972305664 | 1.19 | 177624272 | 1884822272 |
| paper-100k.pdf | 4.02 | 512000000 | 1575384576 | 5.50 | 553513536 | 1706666752 | 5.11 | 487619040 | 1706666752 |
| plrabn12.txt | 31.56 | 167138736 | 150159248 | 35.35 | 161103648 | 145489424 | 34.08 | 157009120 | 152874688 |
| ptt5 | 70.52 | 301714272 | 1204732416 | 72.64 | 260515728 | 1032627776 | 71.91 | 282452384 | 981292480 |
| reymont | 7.77 | 88742512 | 370400320 | 20.56 | 96796928 | 162491168 | 16.30 | 84056744 | 169190752 |
| samba | 20.62 | 150385936 | 237819744 | 26.36 | 152185616 | 193494768 | 24.41 | 143949584 | 178381008 |
| sao | -3.12 | 913918656 | 5737297408 | -1.56 | 734002432 | 7960421376 | -2.08 | 692905024 | 8250220544 |
| sum | 22.67 | 147076928 | 626885248 | 29.50 | 138050528 | 316033056 | 26.95 | 126622520 | 367692320 |
| urls.10K | 11.52 | 121510376 | 190473968 | 17.36 | 130233168 | 174909568 | 15.42 | 107139784 | 174215136 |
| webster | 19.00 | 130613144 | 175077504 | 24.17 | 127334872 | 149651504 | 22.46 | 128427888 | 153494464 |
| x-ray | 19.95 | 320945312 | 906045120 | 21.91 | 312979744 | 987788800 | 21.26 | 340180640 | 778453120 |
| xargs.1 | 24.15 | 352249984 | 248647056 | 27.70 | 384272736 | 248647056 | 26.35 | 352249984 | 248647056 |
| xml | 13.41 | 115304368 | 242306448 | 26.27 | 130102952 | 173813296 | 22.00 | 115185104 | 169224048 |
| Block Size  Performance | 32  19.87 | 32 Comp.  282707566 | 32 Decomp.  920944873 | 64  25.31 | 64 Comp.  264696712 | 64 Decomp.  1194391076 | 96  23.50 | 96 Comp.  254147493 | 96 Decomp.  1122841014 |

Table 5shows performance data for td512 v2.1.1 run iteratively with 8- and 16-byte blocks on the Squash benchmark test data on a 2 GHz quad-core processor. Performance is percent compression and MB per second. To get 8-byte blocks requires changing the value of MIN\_VALUES\_TO\_COMPRESS in td512.h to 8. The default value of 16 causes any number of values from 1 to 15 to be immediately copied to the output without compression.

Table 5: Performance Data for td512 v2.1.1 on 8- and 16-byte Blocks Run on the Squash Benchmark Test Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Filename | 8 | 8 Comp. | 8 Decomp. | 16 | 16 Comp. | 16 Decomp. |
| alice29.txt | 7.50 | 165854960 | 151785440 | 22.90 | 166399344 | 151634112 |
| asyoulik.txt | 5.39 | 181156288 | 159463696 | 20.00 | 177811072 | 161521280 |
| cp.html | -6.28 | 203330576 | 243594064 | 8.56 | 227805552 | 225715600 |
| dickens | 8.42 | 171824304 | 147447344 | 23.83 | 168206064 | 142665424 |
| fields.c | -1.42 | 265476192 | 309722208 | 13.55 | 359677440 | 285897440 |
| fireworks.jpeg | -12.46 | 494349376 | 1578115456 | -6.21 | 826127488 | 3419249920 |
| geo.protodata | -10.11 | 358271904 | 750556992 | -2.92 | 472462176 | 1058821440 |
| grammar.lsp | -2.04 | 286230784 | 310083328 | 12.74 | 310083328 | 265785712 |
| kennedy.xls | -5.27 | 366196288 | 827103552 | 12.98 | 395143520 | 858119936 |
| lcet10.txt | 9.16 | 173689040 | 157357680 | 24.24 | 169279648 | 153453440 |
| mozilla | -3.78 | 365455968 | 762640768 | 5.82 | 393685696 | 876462720 |
| mr | 17.08 | 324774080 | 698952960 | 34.14 | 380396192 | 1043273344 |
| nci | 5.42 | 270332864 | 342906944 | 28.58 | 310496032 | 540879232 |
| ooffice | -8.84 | 341314400 | 1133417856 | -2.04 | 430886112 | 1549670528 |
| osdb | -11.18 | 264980416 | 819707776 | -1.88 | 256705024 | 967266112 |
| paper-100k.pdf | -5.81 | 416260160 | 1044897984 | 1.48 | 640000000 | 1347368448 |
| plrabn12.txt | 8.07 | 167837344 | 154442624 | 23.94 | 163453520 | 127611488 |
| ptt5 | 48.42 | 378477888 | 734214592 | 63.79 | 349364192 | 1051672128 |
| reymont | -12.72 | 257387056 | 1040540480 | 0.23 | 203638224 | 922880128 |
| samba | 1.36 | 187365264 | 254504336 | 14.63 | 194799664 | 240193872 |
| sao | -12.52 | 507235360 | 1464743168 | -6.25 | 778690432 | 2966030336 |
| sum | 0.31 | 303492064 | 570746240 | 14.67 | 267412592 | 616774208 |
| urls.10K | -8.60 | 166529168 | 205288592 | 5.44 | 157171936 | 194000272 |
| webster | -3.67 | 181384544 | 206613760 | 12.17 | 171712192 | 187384816 |
| x-ray | -2.94 | 311541472 | 618873856 | 12.61 | 370215808 | 855033792 |
| xargs.1 | 2.22 | 281800000 | 264187504 | 16.73 | 325153856 | 222473696 |
| xml | -8.58 | 241835040 | 359588288 | 5.98 | 207012896 | 310916704 |
| Block Size  Performance | 8  -0.10 | 8 Comp.  282754919 | 8 Decomp.  567092500 | 16  13.32 | 16 Comp.  328658889 | 16 Decomp.  768250227 |