Improved LZ77 Compression

Cody (Yingquan) Wu
Tenafe Inc.
Campbell, CA 95008, USA

In essence, LZ77 compression works by finding the longest sequences of data that are repeated in history. In practice, the term "sliding window" is incorporated; all it really means is that at any given point in the data, there is a record of what characters went before. In GZIP, the window width is 15, the minimum match length is 3, and a hash function is used. We propose match-length dependent sliding windows, as illustrated by an example in Table 1. By limiting 2-byte match window to 5-bit, the qualified 2-byte matches become compressible. On the other end, 8-byte match within 21-bit window is clearly compressible.

Table 1: An example of length-dependent sliding windows.

Match L. (l)	2	3	4	5	6	7	≥ 8
Window W. (w)	5	9	12	15	17	19	21

LZ4 achieves extremely fast decompression through purposely enforcing byte granularity in compressed context and branchless form of {length of literals, match length, literal sequence, match distance). In LZ4, match window is set to 16-bit (i.e., 2-byte). We proceed to demonstrate the effectiveness of the proposed improvement in the suite of LZ4. We choose the following two-window-two-hash setup as in Table IV.

Table 2: A two-window-two-hash WLZ.

Match L.	Window W.	Hash B.	Hash Chain W.
3	8	3	
4	8	3	
≥ 5	16	5	16

The Silesia corpus is a widely referred data set of files of total size 212MB that covers the typical data types used nowadays. We run both LZ4 and the proposed LZ4 variant, referred to as WLZ wherein "W" standards for multi-window, on the Silesia corpus on a single-core of Intel i7-8850H (2.60GHz) and achieve the following performance numbers, wherein HC denotes the utilization of hash chain. Note hash-chain search for sliding window of 1-byte is not deployed because of too short window. In WLZ-HC, we do not aim to achieve optimal compression ratio but rather a good trade-off between compression ratio and compression speed. On the other hand, it is interesting to re-derive the existing optimal algorithms that maximize compression ratio under the above new two-window circumstance.

Table 3: Performance comparisons of various algorithms over Silesia corpus

Alg.	Comp. Speed	Decomp. Speed	Comp. Ratio
LZ4	651MB/s	3.54GB/s	2.11
WLZ	265MB/s	3.15GB/s	2.40
LZ4-HC	33.6MB/s	3.64GB/s	2.72
WLZ-HC	62.5MB/s	3.48GB/s	2.71

2021 Data Compression Conference (DCC) | 978-1-6654-0333-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/DCC50243.2021.00066

2375-0359/21/$31.00 ©2021 IEEE
DOI 10.1109/DCC50243.2021.00066