SnappyR: A New High-Speed Lossless Data Compression Algorithm

Rui Chen (ruichen@mathworks.com)

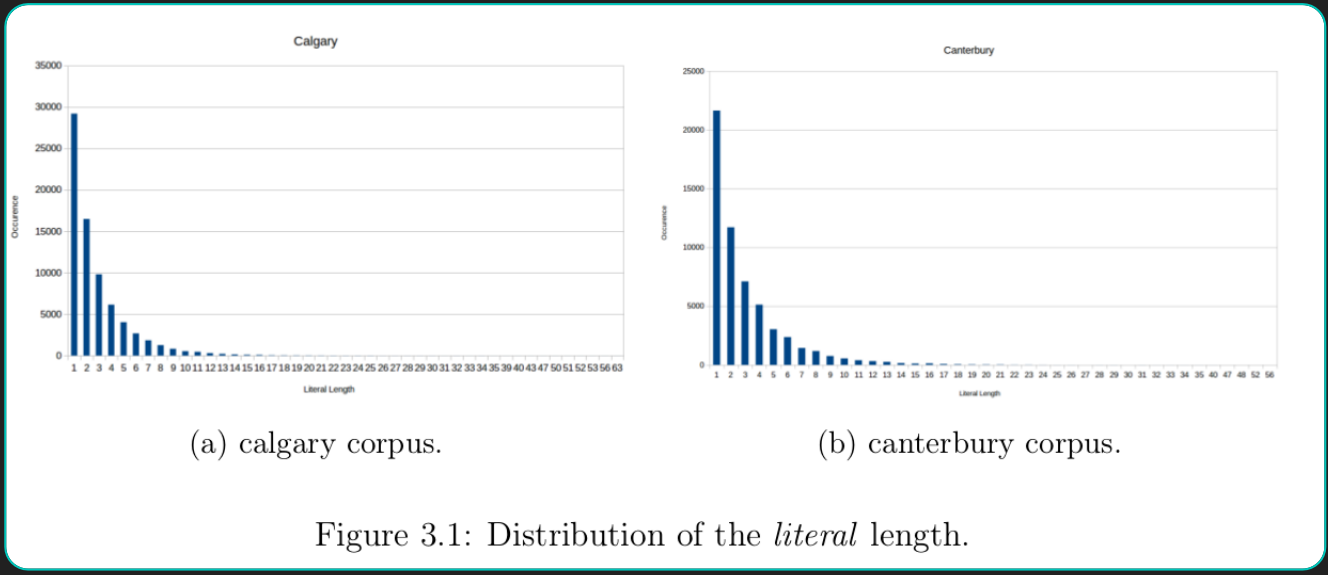
Lihao Xu (lihao@wayne.edu)

Snappy

- Snappy algorithm:
- - developed by Google in 2011
- - open-sourced in C++
- - widely used, e.g. BigTable, MapReduce, Hadoop, etc.

SnappyR Inspiration

- Based on Snappy
- Inspired by two observation during match searching:
- - lots of 0-length literals
- - offset lengths tend to be short



(a) calgary corpus. (b) canterbury corpus.

Figure 3.1: Distribution of the *literal* length.

SnappyR

- Compression/decompression process is the same as Snappy
- New token structure

Token bits	Type	Description
*LLLL_LL***00** *AAAA_AAAA* *AAAA_AAAA* *AAAA_AAAA* *AAAA_AAAA*	Literal	The six *L*-bits indicate the length of the *literal*. A length of 60, 61, 62 or 63 indicates there are additional 1, 2, 3 or 4 byte(s) following to present the length instead of the six *L*-bits. Note the *A*-bits may or may not exist in a *literal* token.
*FFFM_MM***01** *FFFF_FFFF*	Match	The three *M*-bits indicate the length of the *match*. The eleven *F*-bits indicate the length of the *offset*.
*MMMM_MM***10** *FFFF_FFFF* *FFFF_FFFF*	Match	Same as above.
*MMMM_MM***11** *FFFF_FFFF* *FFFF_FFFF* *FFFF_FFFF* *FFFF_FFFF*	Match	Same as above.

SnappyR Evaluation

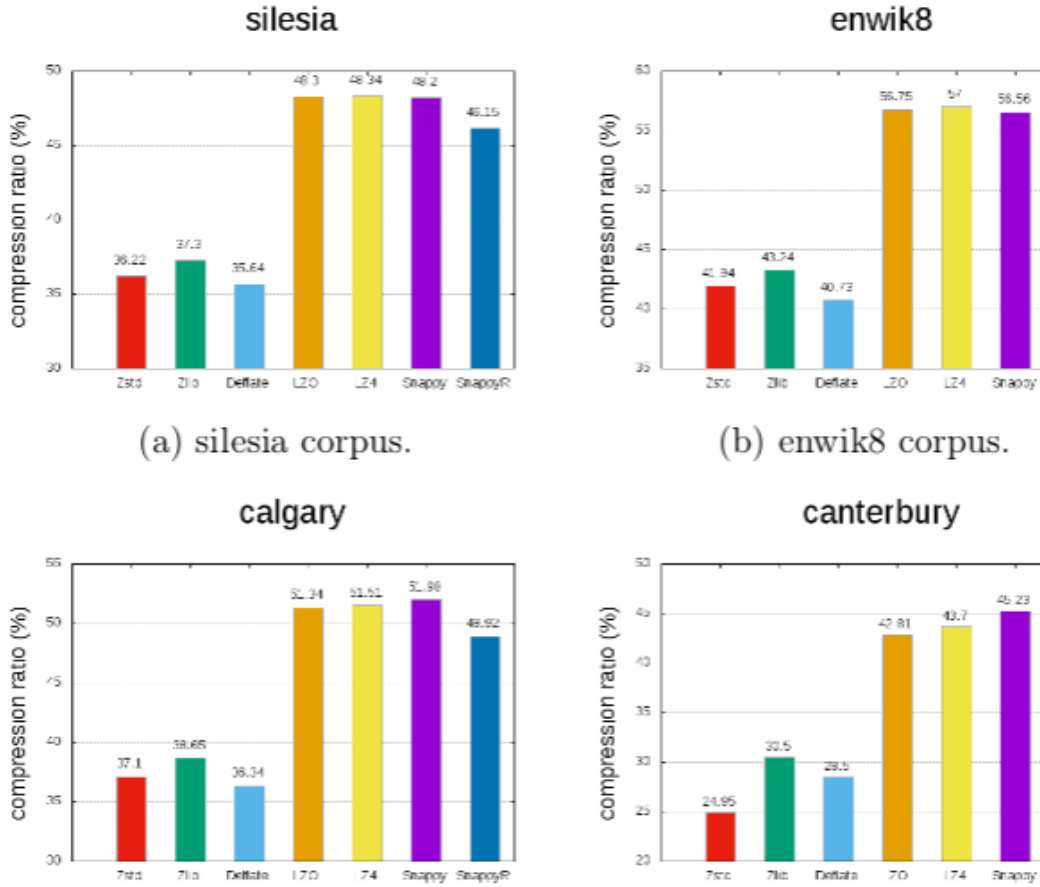
- Test platforms:

Platform	CPU Model	L1 Cache	L2 Cache	L3 Cache	Memory	OS
Y700	Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz	8 × 64KB	8 × 256KB	6MB	16GB	Ubuntu 16.04
Grid	Intel(R) Xeon(R) CPU E5-2698 v3 @ 2.30GHz	2 × 64KB	2 × 256KB	4MB	8GB	CentOS 7

- Comparison algorithms:
  - Zlib 1.2.11, Deflate 1.3, LZO1x 2.10, Snappy 1.1.4, LZ4 1.9.2
- Corpus sets:
  - Silesia, Calgary, Canterbury, enwik8

SnappyR Evaluation – Compression Ratio

- Compression ratio: 5%-10% better (LZO, Snappy, LZ4)



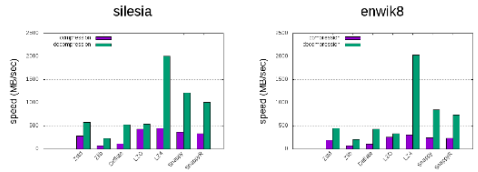
(a) silesia corpus. (b) enwik8 corpus.

(c) calgary corpus. (d) canterbury corpus.

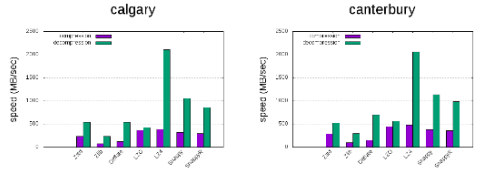
Figure 3.5: Compression ratio with *blocksize* = *64KB*.

SnappyR Evaluation – Speed

- Compression speed: similar (LZO, LZ4), 5%-10% faster (Snappy)
- Decompression speed: similar (Snappy)

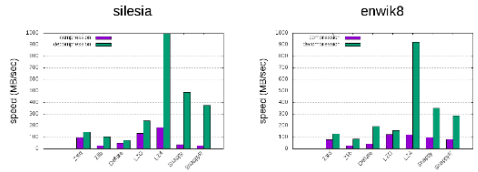


(a) silesia corpus. (b) enwik8 corpus.

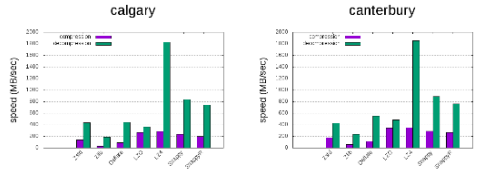


(c) calgary corpus. (d) canterbury corpus.

Figure 3.12: Performance on Y700 with *blocksize=64KB*.



(a) silesia corpus. (b) enwik8 corpus.



(c) calgary corpus. (d) canterbury corpus.

Figure 3.13: Performance on Grid with *blocksize=64KB*.

SnappyR Evaluation

- SnappyR can become another viable replacement or alternative to Snappy, LZ4 or LZO for computing and storage systems and applications, where high-speed lossless data compression is needed.

Thank you!

Rui Chen (ruichen@mathworks.com)

Lihao Xu (lihao@wayne.edu)