Using GPUs for compression

Stefan Rua

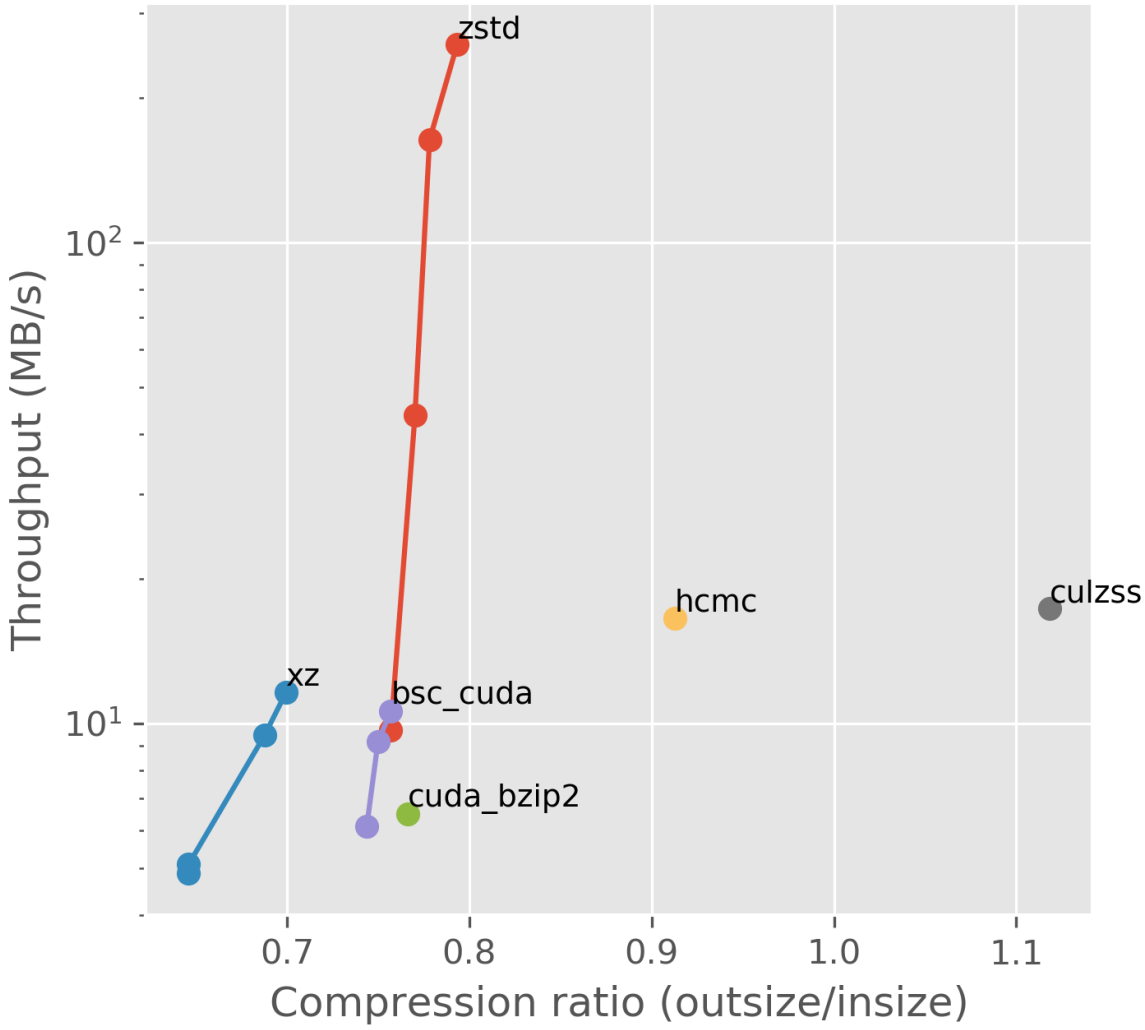
Why?

- Data from the HLT is compressed
- This is done on CPUs
- GPUs can be very fast
- We have GPUs

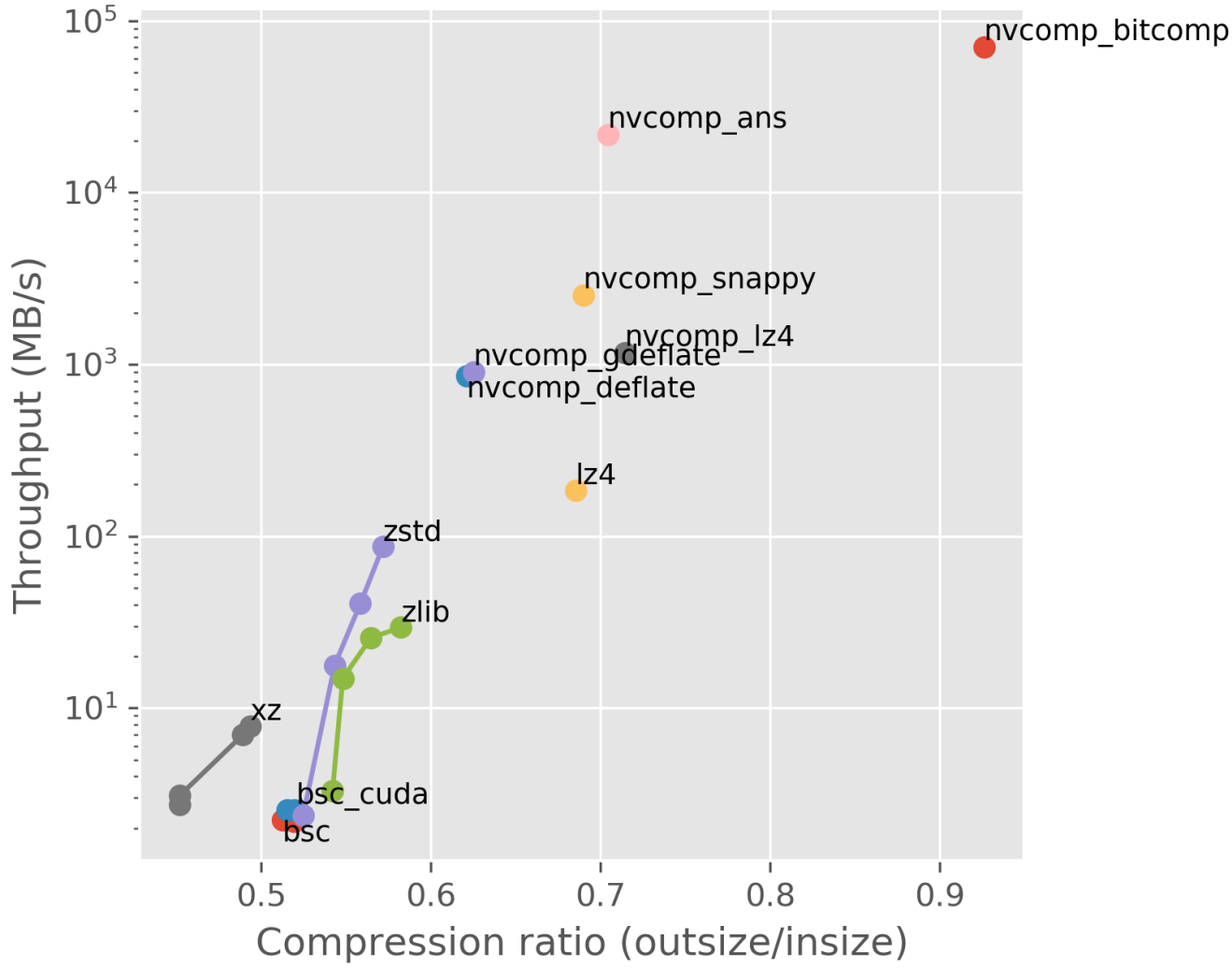
What I've been doing

- Looking for implementations by other people
- Comparing them to traditional CPU compressors

Results (so far)



Results (so far)



Problems

Nvidia

From version 2.3 onwards, the compression / decompression source code will not be released. We'll continue to maintain this Github for documentation and code sample purposes.

Researchers

Hi Stefan

It has been a while that we have worked on that. The best I can find is this code for CULZSS not for bit one, but it might be working or not, i am not sure.

hope that helps.

Promising things

Facbook has an MIT-licensed GPU-enabled ANS implementation! [1]

- Only a Python API for Torch tensors
- C++ API "coming soon"
- I'm trying to make use of the underlying C++ functions

[1] https://github.com/facebookresearch/dietgpu

That's it for now

Contact info

Stefan Rua

stefan.elias.rua@cern.ch
stefan.rua@iki.fi

Extra

Burrows-Wheeler transform
Sorts the data in a way that repeating patterns are near each other

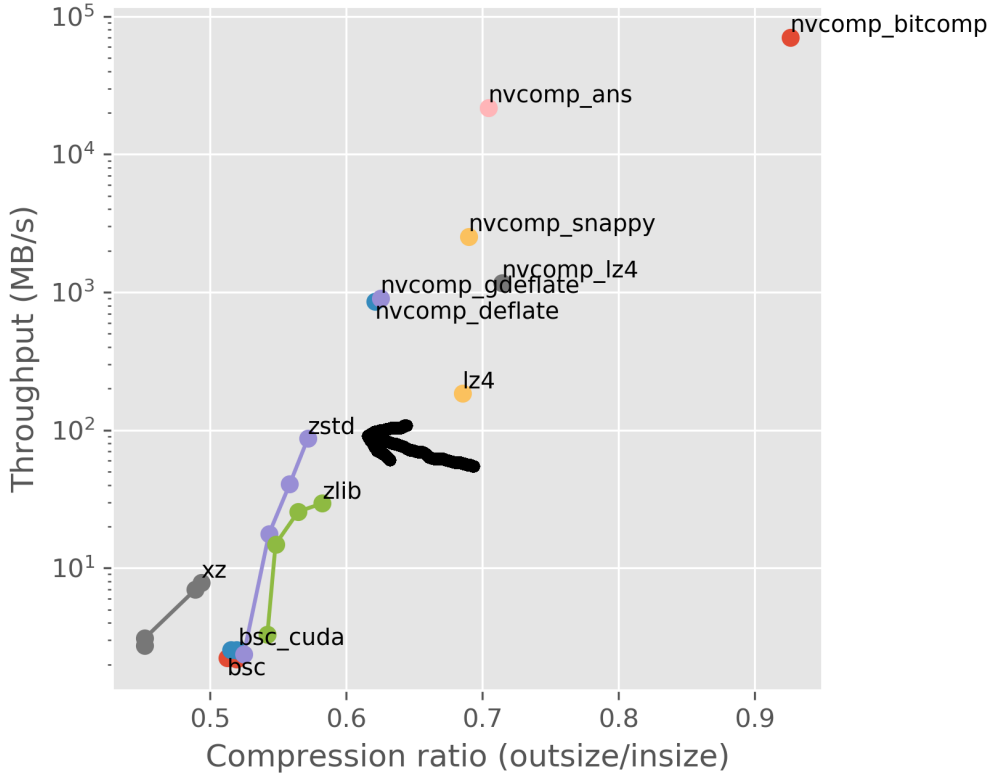
Huffman coding
Gives characters shorter codes based on frequency

Lempel-Ziv compression
Replaces patterns with references to previous occurences

Asymmetric numeral systems
Stores the data in an integer that is constructed based on probability of encountering a certain symbol

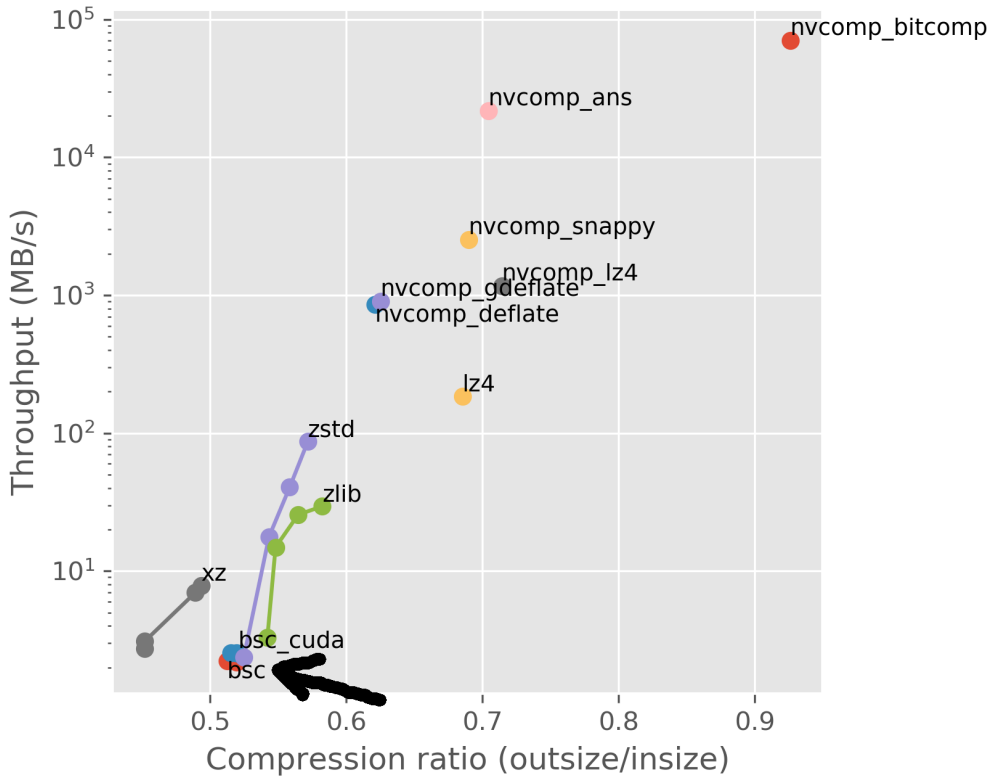
Extra

zstd: LZ77 + ANS



Extra

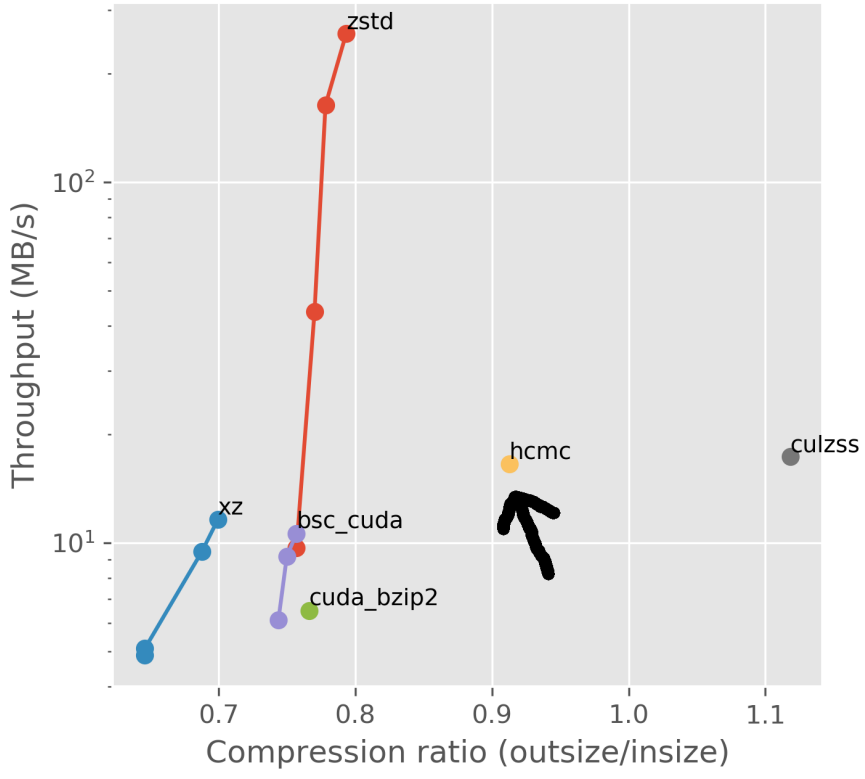
bsc: Burrows-Wheeler transform / higher order sort transform + LZ



https://github.com/IlyaGrebnov/libbsc

Extra

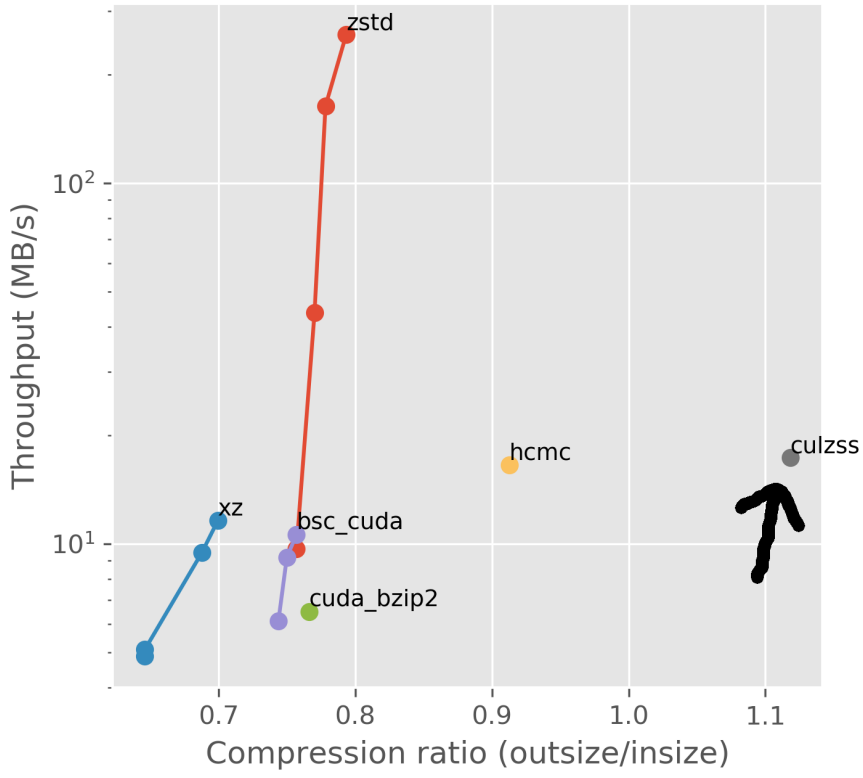
HuffmanCoding_MPI_CUDA: Huffman coding



https://github.com/smadhiv/HuffmanCoding_MPI_CUDA

Extra

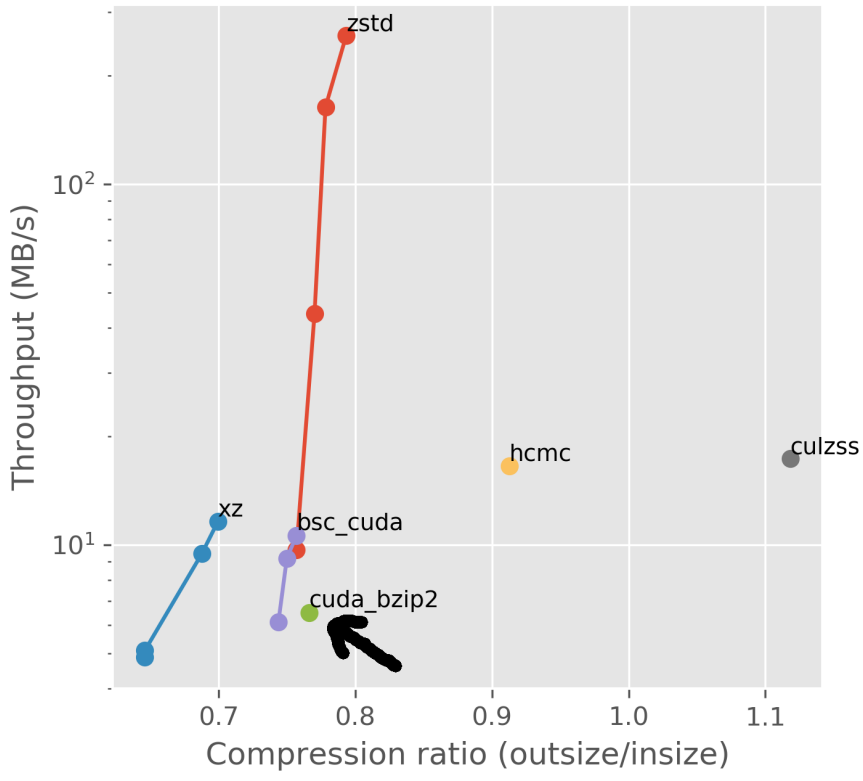
CULZSS: LZSS



https://github.com/adnanozsoy/CUDA_Compression

Extra

cuda_bzip2: BWT + Huffman



https://github.com/aditya12agd5/cuda_bzip2

Extra

ans: ANS
snappy: Modified LZ77
deflate: LZ77 + Huffman