



IC Project 3

Diogo Borges - 102954
Henrique Cruz - 103442
Piotr Bartczak - 130327



Overview

- Analysis of the File Structure
- Proposed Solution
- Experimental Results
- Conclusions



Analysis of the File Structure

The file *model.safetensors* stores neural network weights in the **BF16 (Brain Float 16)** format. Each number occupies 2 bytes (16 bits):

- **Most Significant Byte (MSB):** Contains the sign and the exponent.
- **Least Significant Byte (LSB):** Contains the mantissa (precision).

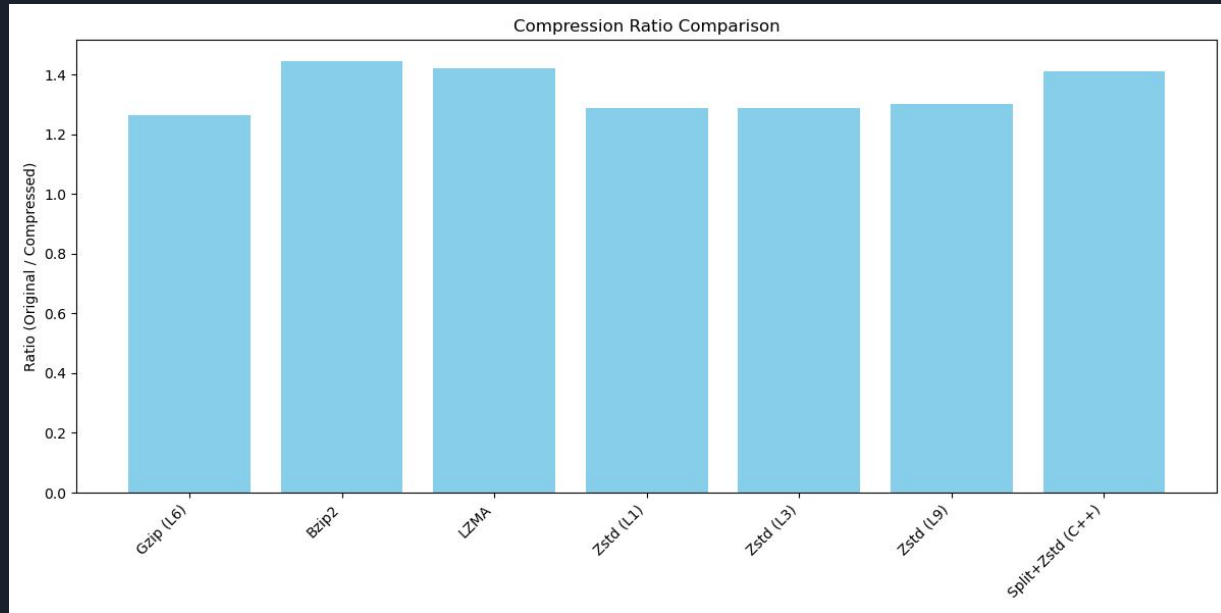


Proposed Solution

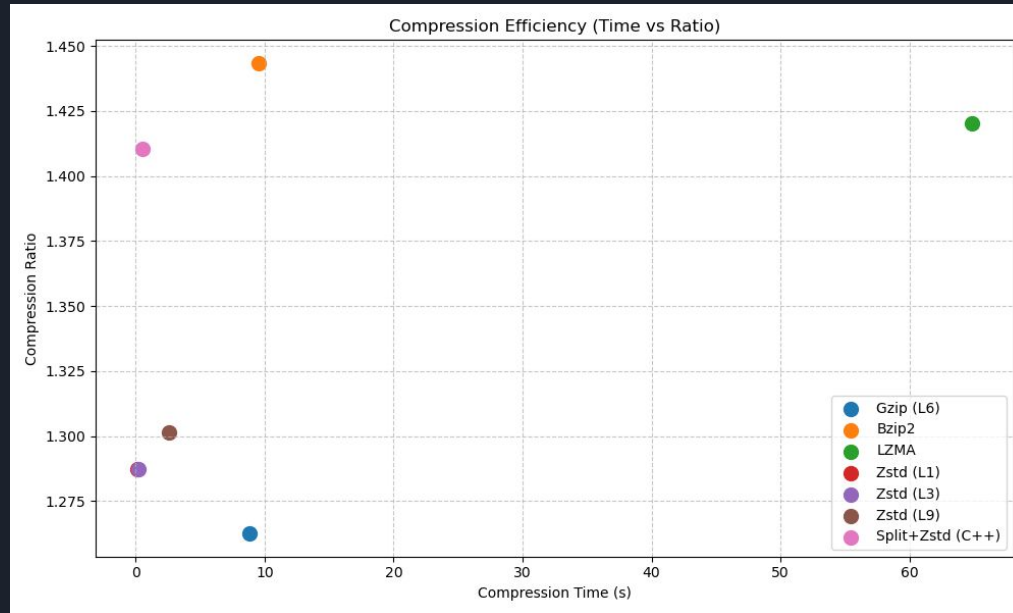
We implemented a **Byte-Plane Splitting** strategy. The process involves:

- **Splitting:** The data stream is separated into two distinct streams: one containing all the even bytes (MSBs) and another containing all the odd bytes (LSBs).
- **Concatenation:** The two streams are concatenated, grouping the highly compressible MSBs together.
- **Compression:** The transformed data is compressed using Zstandard (Zstd), a modern algorithm known for its high speed and good compression ratios.

Experimental Results - Compression Ratio



Experimental Results - Performance (Time vs. Ratio)





Experimental Results - Comparison

Algorithm	Ratio	Comp. Time (s)	Decomp. Time (s)
Gzip (L6)	1.26	8.85	0.64
Bzip2	1.44	9.58	4.92
LZMA	1.42	64.83	2.98
Zstd (L3)	1.29	0.21	0.13
<u>Split+Zstd (C++)</u>	<u>1.41</u>	<u>0.54</u>	<u>0.25</u>



Conclusions

- The **Byte-Plane Splitting** strategy combined with **Zstd** proved to be the optimal solution.
- By exploiting the internal structure of the **BF16** format, we achieved a very good compression ratio with extremely fast compression and decompression times.