Reproducibility Report for ACM SIGMOD 2024 Paper: "ALP: Adaptive Lossless Floating-Point Compression"

Bogdan Nicușor Bindea
bogdan.bindea@cs.utcluj.ro
Technical University of Cluj-Napoca
Cluj-Napoca, Romania

George Katsogiannis-Meimarakis
katso@athenarc.gr
Athena Research Center
Athens, Greece

Luca Zecchini
luca.zecchini@unimore.it
University of Modena and Reggio Emilia
Modena, Italy

Azim Afroozeh
azim@cwi.nl
Centrum Wiskunde & Informatica
Amsterdam, Netherlands

Leonardo Kuffó
lxkr@cwi.nl
Centrum Wiskunde & Informatica
Amsterdam, Netherlands

Peter Boncz
boncz@cwi.nl
Centrum Wiskunde & Informatica
Amsterdam, Netherlands

Abstract

This report describes the reproducibility evaluation carried out for the research paper "ALP: Adaptive Lossless Floating-Point Compression", authored by Azim Afroozeh, Leonardo Kuffó, and Peter Boncz from the Centrum Wiskunde & Informatica (CWI), located in Amsterdam, Netherlands. The paper, presented at SIGMOD 2024, introduces ALP, a novel adaptive lossless compression algorithm for floating-point datasets.

The results of the reproducibility evaluation confirm the core contributions of the original paper, showing that ALP consistently outperforms state-of-the-art techniques in both compression ratio and (de)compression speed.

CCS Concepts

• **Information systems → Data compression**.

Keywords

reproducibility, lossless compression, floating point compression, lightweight compression, vectorized execution, columnar storage, big data formats

ACM Reference Format:
Bogdan Nicușor Bindea, George Katsogiannis-Meimarakis, Luca Zecchini, Azim Afroozeh, Leonardo Kuffó, and Peter Boncz. 2024. Reproducibility Report for ACM SIGMOD 2024 Paper: "ALP: Adaptive Lossless Floating-Point Compression". In *Reproducibility Reports of the 2024 International Conference on Management of Data (SIGMOD ARI Reports '24), June 9–15, 2024, Santiago, Chile.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3687998.3717057

SIGMOD ARI Reports '24, Santiago, Chile.

ACM ISBN 979-8-4007-1179-4/2024/06
https://doi.org/10.1145/3687998.3717057

1 Introduction

This reproducibility report examines the findings presented in the research paper "ALP: Adaptive Lossless Floating-Point Compression" [1], presented at SIGMOD 2024 and authored by Azim Afroozeh, Leonardo Kuffó, and Peter Boncz from the Centrum Wiskunde & Informatica (CWI), located in Amsterdam, Netherlands. The paper introduces ALP, a novel adaptive lossless compression algorithm for floating-point datasets. ALP dynamically adjusts its compression strategies based on data characteristics, offering significant improvements in compression ratio and execution performance compared to existing methods. In particular, the main contributions of the paper include:

- An in-depth analysis of the datasets used to motivate and evaluate state-of-the-art floating-point encodings.
- An innovative approach to lossless compression (ALP) that is adaptable to diverse floating-point datasets, whose C++ implementation is openly available (together with data and useful scripts) on GitHub[1].
- A thorough experimental evaluation showing that ALP outperforms state-of-the-art methods in both compression ratio and (de)compression speed on a wide range of datasets.

Our reproducibility evaluation aims at validating the core claims of the paper. Specifically, we want to:

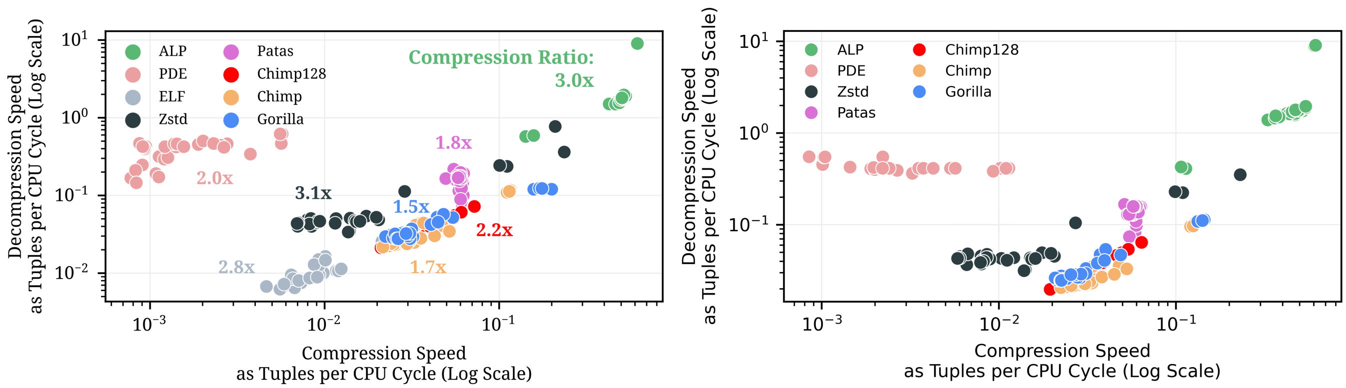
- Confirm the reported compression ratios across datasets.
- Validate the reported performance metrics (i.e., compression and decompression speed).
- Assess the reproducibility of the experiments on alternative hardware configurations.

The authors provided a comprehensive set of artifacts (described in Section 2), including source code, scripts, and datasets, which significantly streamlined the reproducibility process. Despite some challenges were encountered in adapting the experiments to different hardware setups (all details about machine configurations are reported in Section 3), thanks to the help of the authors the

[1] https://github.com/cwida/ALP

Table 1: Hardware platforms used for the experimental evaluation, i.e., Table 3 from the original paper [1].

Architecture	Scalar ISA	Best SIMD ISA	CPU Model	Frequency
Intel Ice Lake	x86_64	AVX-512 (512-bits)	8375C	3.5 GHz
AMD Zen 3	x86_64	AVX2 (256-bits)	EPYC 7R13	3.6 GHz
Apple M1	ARM64	NEON (128-bits)	Apple M1	3.2 GHz
AWS Graviton 2	ARM64	NEON (128-bits)	Neoverse N1	2.5 GHz
AWS Graviton 3	ARM64	NEON (128-bits) SVE (variable)	Modified Neoverse V1	2.6 GHz



(a) Original results. **(b) Reproduced results.**

Figure 1: Original and reproduced (de)compression speed results, i.e., Figure 1 from the original paper [1].

reproducibility process mostly ran smoothly and the core claims of the paper were clearly validated, as described in detail in Section 4. This report provides a critical evaluation of the reproducibility process, offering insights to potentially improve the robustness and portability of similar studies.

2 Submission

The reproducibility submission provided by the authors includes a well-organized set of artifacts that ensure the smooth reproduction of the key results presented in the paper. The submission adheres to reproducibility standards by including the necessary code, scripts, data, and documentation. In particular, the key components provided by the authors are:

- GitHub repository: the authors shared a public repository, available at https://github.com/cwida/ALP, which contains the code, scripts, and data required for reproduction.
- Detailed README file: the repository includes a comprehensive README file that provides instructions for setting up the environment, running the experiments, and recreating the results and figures.
- Scripts for automation: a set of scripts is provided to automate the execution of the experiments and generate results, making the process user-friendly and efficient.
- Master script: A single script is provided to automate the entire process by running and executing all previously mentioned scripts, including benchmarking, generating tables, and creating figures, making it even easier for reviewers to reproduce the results.
- Data sources: the submission includes datasets used in the experiments or links to sources where the data can be obtained. Additionally, data generators are provided for synthetic datasets.
- Self-validation: the repository includes self-check mechanisms to verify the correct setup and functionality of the environment.

The submission meets the ideal criteria for a reproducibility package, reflecting the authors' commitment to transparency and collaboration. Most results, including figures and tables, can be reproduced by running the provided scripts with minimal effort, ensuring accessibility and reliability for reviewers. A few results could not be reproduced solely due to limited access to specific hardware.

3 Hardware Configuration

In the original paper [1], the authors use multiple hardware platforms to carry out their experimental evaluation, whose configurations are reported in Table 1. Even though we did not have access to these machines, we performed the evaluation on machines with heterogeneous features, assessing the reproducibility of the results on different hardware configurations. In particular, each reviewer adopted one of the following machines:

- An HP ZBook Fury G8 notebook with 32 GB of RAM and an Intel Core i7 (11th generation) CPU @ 2.5 GHz.
- A MacBook Air with an Apple M2 chip and 16 GB of RAM.

Table 2: Reproduced compression ratio results (original values between brackets), i.e., Table 4 from the original paper [1].

Dataset	ALP	Gorilla	Chimp	Chimp128	Patas	Zstd
Air-Pressure	16.4 (16.5)	24.5 (24.7)	23.0 (23.0)	19.2 (19.3)	27.8 (27.9)	9.4 (8.7)
Basel-Temp	30.7 (29.8)	60.5 (61.6)	54.1 (54.1)	31.1 (31.2)	36.4 (36.5)	18.4 (18.3)
Basel-Wind	29.8 (29.8)	62.2 (63.2)	54.7 (54.7)	38.4 (38.4)	48.9 (48.9)	14.7 (14.6)
Bird-Mig	20.1 (20.1)	47.8 (48.7)	42.1 (41.9)	26.4 (26.3)	36.1 (35.9)	21.0 (21.0)
Btc-Price	26.4 (26.4)	54.7 (51.5)	48.1 (48.2)	45.0 (45.1)	56.8 (57.1)	42.1 (49.9)
City-Temp	10.7 (10.7)	58.8 (59.7)	46.3 (46.2)	23.0 (23.0)	24.2 (24.2)	16.8 (16.2)
Dew-Temp	13.4 (13.5)	54.9 (56.2)	51.8 (51.8)	32.6 (32.6)	39.0 (39.0)	25.1 (20.9)
Bio-Temp	10.8 (10.7)	50.7 (51.9)	46.3 (46.3)	18.9 (18.9)	22.9 (22.9)	17.5 (14.5)
PM10-dust	8.6 (8.2)	27.5 (27.7)	24.4 (24.4)	13.7 (13.7)	19.9 (19.9)	7.8 (6.9)
Stocks-DE	11.0 (11.0)	46.2 (46.9)	42.9 (42.9)	13.6 (13.6)	20.8 (20.8)	10.5 (9.4)
Stocks-UK	12.6 (12.7)	34.8 (35.6)	31.3 (31.3)	16.8 (16.8)	21.5 (21.5)	10.3 (10.7)
Stocks-USA	7.9 (7.9)	37.2 (37.7)	35.0 (35.0)	12.2 (12.2)	19.2 (19.2)	8.6 (7.8)
Wind-dir	15.9 (15.9)	58.1 (59.4)	53.9 (53.9)	27.8 (27.8)	28.1 (28.2)	25.5 (24.7)
TS AVG	**16.5 (16.4)**	**47.5 (48.1)**	**42.6 (42.6)**	**24.5 (24.5)**	**30.9 (30.9)**	**17.5 (17.2)**
Arade/4	24.9 (24.9)	58.0 (58.1)	55.6 (55.6)	49.0 (49.0)	59.1 (59.1)	33.9 (33.8)
Blockchain	36.5 (36.2)	62.9 (65.5)	58.3 (58.3)	53.2 (53.2)	62.6 (62.6)	44.0 (38.3)
CMS/1	35.6 (35.7)	37.2 (37.8)	34.8 (34.8)	28.1 (28.2)	36.8 (36.8)	26.6 (24.5)
CMS/25	41.1 (41.1)	64.4 (65.4)	59.5 (59.5)	57.2 (57.2)	70.1 (70.1)	58.3 (56.5)
CMS/9	11.7 (11.7)	16.0 (17.1)	18.7 (18.7)	25.7 (25.7)	26.0 (26.0)	14.7 (14.7)
Food-prices	23.7 (23.7)	38.3 (40.8)	28.0 (28.0)	24.6 (24.7)	28.3 (28.3)	18.3 (16.6)
Gov/10	31.0 (31.0)	57.5 (58.1)	45.8 (45.7)	34.2 (34.2)	35.8 (35.9)	28.1 (27.4)
Gov/26	0.4 (0.4)	2.4 (2.4)	2.3 (2.3)	9.3 (9.3)	16.2 (16.2)	0.2 (0.2)
Gov/30	7.5 (7.5)	10.2 (10.3)	8.9 (8.9)	12.9 (12.9)	19.3 (19.3)	4.5 (4.2)
Gov/31	3.1 (3.1)	5.6 (5.7)	5.0 (5.0)	10.4 (10.4)	17.1 (17.1)	1.6 (1.5)
Gov/40	0.8 (0.8)	2.7 (2.7)	2.6 (2.6)	9.4 (9.4)	16.3 (16.4)	0.5 (0.4)
Medicare/1	39.4 (39.4)	45.5 (45.9)	42.7 (42.7)	32.3 (32.3)	39.9 (39.9)	31.2 (28.7)
Medicare/9	12.3 (12.3)	17.0 (17.9)	19.1 (19.1)	26.0 (26.0)	26.3 (26.3)	15.0 (14.9)
NYC/29	40.4 (40.4)	30.6 (30.8)	29.6 (29.6)	28.7 (28.7)	38.8 (38.8)	27.5 (20.5)
POI-lat	55.5 (55.5)	66.0 (66.0)	57.7 (57.7)	57.5 (57.5)	71.7 (71.7)	59.3 (48.1)
POI-lon	56.6 (56.4)	66.1 (66.1)	63.4 (63.4)	63.2 (63.1)	75.9 (75.9)	61.0 (53.1)
SD-bench	16.2 (16.2)	50.7 (51.1)	45.3 (45.7)	18.8 (19.2)	22.8 (23.0)	11.3 (11.8)
NON-TS AVG	**25.7 (25.7)**	**37.1 (37.7)**	**34.0 (34.0)**	**31.8 (31.8)**	**39.0 (39.0)**	**25.6 (23.3)**
ALL AVG	**21.7 (21.7)**	**41.6 (42.2)**	**37.7 (37.7)**	**28.6 (28.7)**	**35.5 (35.5)**	**22.1 (20.6)**

- A Google Cloud Platform virtual machine with an Intel Ice Lake processor, namely a Xeon CPU @ 2.60 GHz.

4 Reproducibility Evaluation

The experimental evaluation carried out in the original paper [1] demonstrates how ALP outperforms state-of-the-art techniques in both compression ratio and (de)compression speed. In particular, Table 4 shows the results about compression ratio, measured in bits per value, while (de)compression speed is reported in Table 5 and graphically in Figure 1. Thus, to mark the results as reproduced, we focused the reproducibility evaluation on such experiments. In Section 4.1, we describe the challenges met during the evaluation process, while we report the achieved results in Section 4.2.

4.1 Process

The reproducibility process mostly ran smoothly for all reviewers. Using hardware platforms different from the ones adopted by the authors, some challenges were encountered in adapting the experiments to the reviewers' setups. Nevertheless, the authors were very collaborative and helpful, promptly answering messages when a punctual clarification was needed and making corrections and additions to their code (e.g., to fix minor bugs causing data not being read correctly or include results reported in Table 7 of the original paper) to facilitate the reproducibility process.

4.2 Results

Here we report the results achieved through our reproducibility evaluation, focused on the reproduction of the experiments assessing the performance of ALP in terms of *(i)* compression ratio (i.e., Table 4 from the original paper) and *(ii)* (de)compression speed (i.e., Figure 1 from the original paper).

The reproduced results about compression ratio achieved by ALP and the baselines made available in the repository, consistent across all reviewers' hardware configurations, are shown in Table 2, where we also report between brackets the corresponding value appearing in Table 4 from the original paper. The green cell background highlights the top performer on each dataset among ALP, Gorilla, Chimp, Chimp128, and Patas (i.e., not accounting for Zstd, coherently with the original table). Overall, the obtained measures are extremely close – often identical – to the ones reported in the original table. Further, net of PDE and Elf (not included in the repository), the best-performing solution on each dataset is always the same as the original table. Thus, the results about compression ratio are clearly reproduced, confirming that ALP consistently outperforms state-of-the-art methods.

A comparison between the original and reproduced results about compression and decompression speed achieved by ALP and the baselines is shown instead in Figure 1. While results reported in Figure 1b were obtained through the MacBook Air described in Section 3, conclusions were consistent across all reviewers' configurations. Results are successfully reproduced, displaying a similar trend to the paper and showing that ALP achieves the best compression and decompression speed overall.

5 Summary

The core contributions and claims presented in the original paper [1] were successfully reproduced through a smooth evaluation process, assessing how ALP outperforms state-of-the-art techniques in both compression ratio and (de)compression speed.

References

[1] Azim Afroozeh, Leonardo Kuffó, and Peter Boncz. 2023. ALP: Adaptive Lossless Floating-Point Compression. *Proceedings of the ACM on Management of Data (PACMMOD)* 1, 4, Article 230 (2023), 26 pages. doi:10.1145/3626717