VIRGINIA TECH®

Implementation and Evaluation of GBDI Memory Compression Algorithm for Different Workloads

Sunil Venkatesh Rao

Introduction

GBDI: Global Base-Delta-Immediate Compression

- GBDI: Going Beyond Base-Delta-Immediate Compression with Global Bases, 2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA) – Chalmers University of Technology.
- Substantially higher bandwidth by selecting global bases.
- Offers a mean compression ratio of 2.3x for SPEC2017 benchmarks [1].
- Requires a data analysis phase which classifies it to the family of statistical compression techniques.

Research Goal

- Most compression techniques cannot be generalized. Some techniques are better for one type of data than others.
- Techniques involve min-maxing every step of the way to obtain good compression ratios.
- Demonstrated for SPEC2017 workloads. What about for others?

Architecture

GBDI Overview

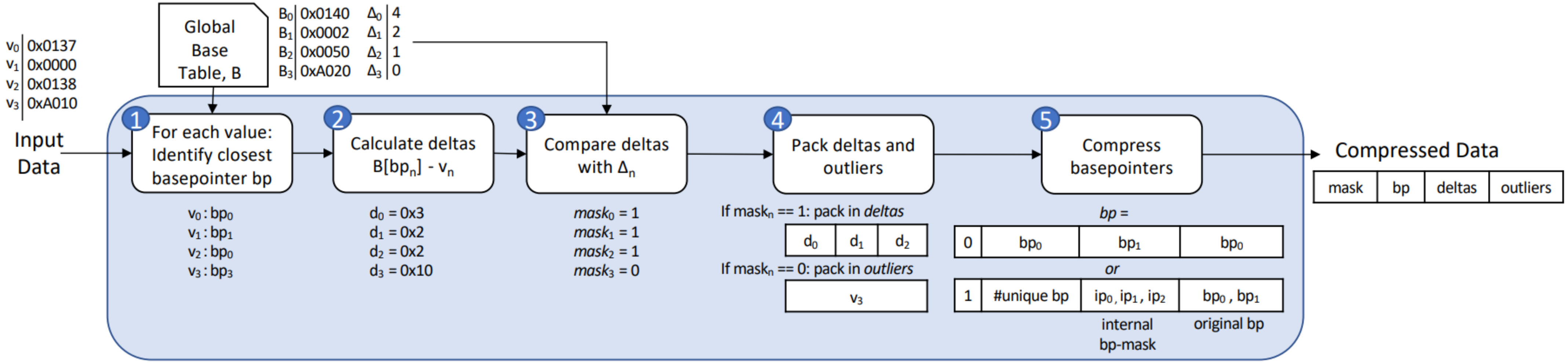


Fig. 1: Overview of the main steps in GBDI compression

Implementation

Step 1: Global Base Table

- 200k values are read for initial data analysis.
- N bins are chosen such that all 200k values are classified into one of the N equally spaced bins.
- Among the N bins, we choose B bins that contain the most values. A value of B = 2048 has been found to be the best.
- Choosing a value of N is critical as it affects the size of deltas. Algorithm is run for different values in steps of 2 bits and the best value chosen.
- $N = 2^{28}$ is found to yield the best compression ratio.

Step 2: Identify Closest Base Pointer and Compute Deltas

- For each of the 16 words in a memory block, identify the closest base pointer from the global base table and find its corresponding delta.
- Delta = Closest Base Value - Data
- Establish a maximum delta for each global base by considering the maximum distance to the closest values.
- An upper bound of 16 – $\log_2(B)$ is chosen for delta to improve compression. This corresponds to 5 signed bits (-16 to +15) for representing delta value.
- Note that the closest base pointer is the index (0 to 2047) of the base table corresponding to the closest base.

Step 3: Compare Deltas

- Compare the delta of each of the 16 words with the max delta corresponding to its closest base pointers.
- A mask is established for each word where the mask bit is 1 if the delta is lesser than the corresponding base’s max delta. Else mask bit is 0.
- Each compressed memory block contains this mask array of 16 bits.

Step 4: Pack Deltas and Outliers

- If mask = 1, the value is considered as an inlier and is compressed using the closest base pointer and delta.
- Since an upper bound of $16 - \log_2(B)$ is chosen for delta, each word can be compressed to *deltasize* + *ptrsize* = $16 - \log_2(B) + \log_2(B) = 16$ bits.
- If mask = 0, the value is left uncompressed and the full 32-bit value is stored.

Step 5: Compress Base Pointers

- Combinations of several optimization methods are used to compress the size of base pointer array.
- *#unique_bp* is determined for every block. If *#unique_bp* ≤ *4* then the number is stored first followed by an internal base pointer mask for each of the inliers. Finally, the unique base pointers are also stored.
- If *#unique_bp* > *4*, each of the base pointers are simply concatenated.
- Irrespectively, each of the base pointers are also Huffman-encoded [2] if it is found to be beneficial. This greatly reduces the bits used from a constant 11 bits per inlier to variable bits and improves compression ratio.

Other Optimizations

- The scenario when all blocks are equal is identified. In this case, an additional step is added to the compression algorithm to store only the unique value and format encoding bits to indicate the scenario.
- Exploiting Intra-Block Bases is a technique suggested in the paper which combines BDI with GBDI on a per block basis to select the best compression method for that block. This, however, shows negligible improvement according to the paper and has not been implemented.

GBDI Compression Formats

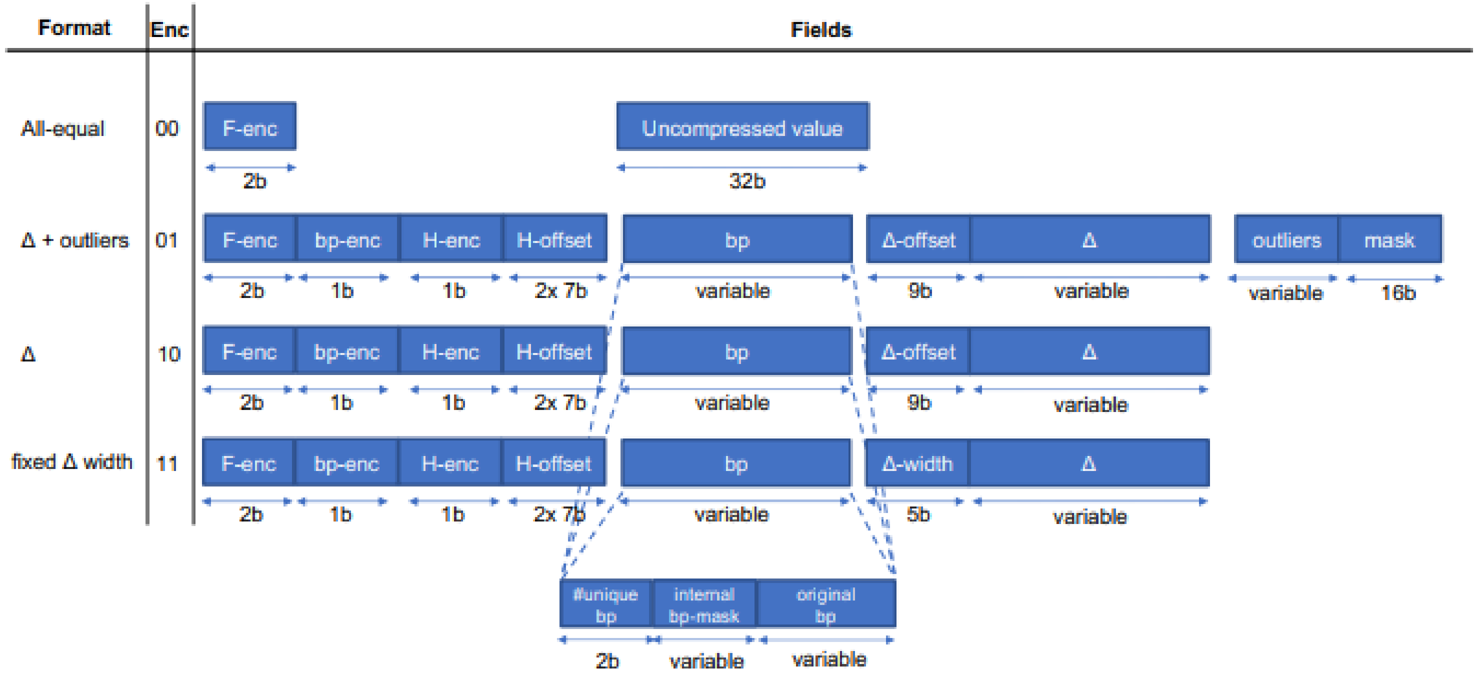
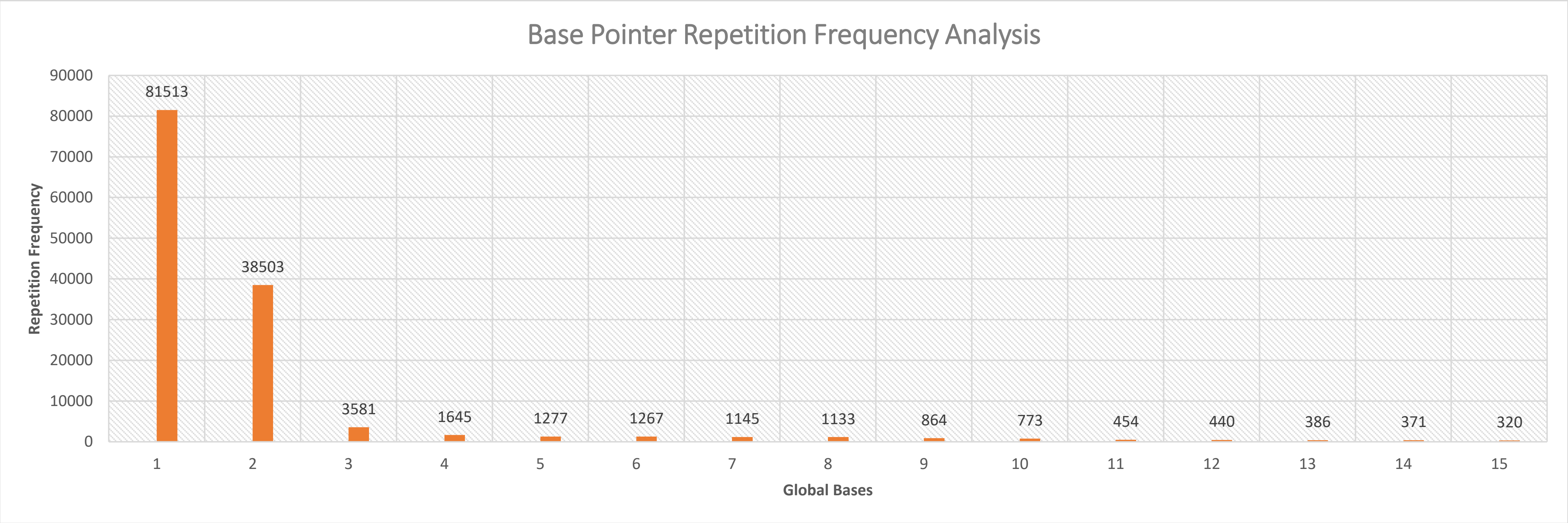


Fig. 2: GBDI compression formats

Experiment and Results

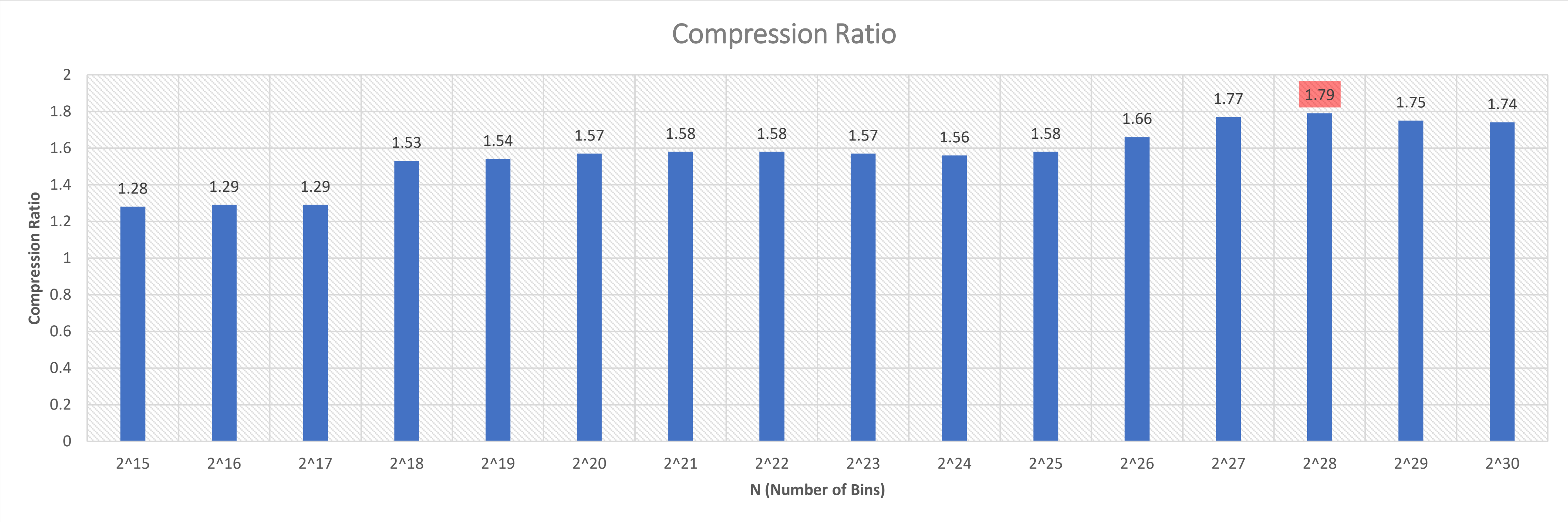
Histogram Binning



Huffman Encoding

Base Pointer	Huffman Code	Bits Used
0	1	1
1	01	2
2	00101	5
3	000111	6
4	000010	6
5	000001	6
6	0011110	7
7	0011101	7
8	0010001	7
9	0001100	7

Effect of Number of Bins on Compression Ratio

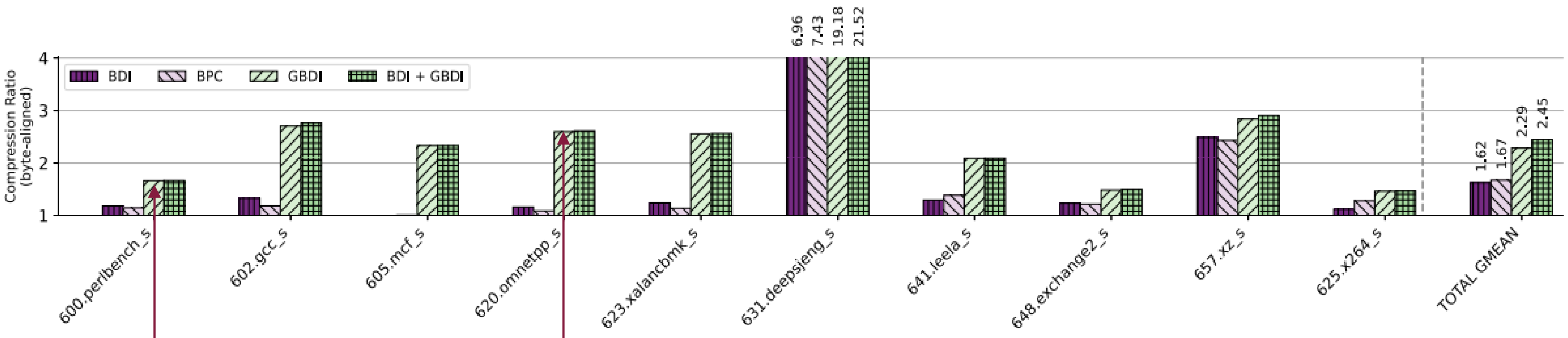


VIRGINIA TECH

Inlier vs Outlier Count

No. of Bits	Inlier Count	Outlier Count
15	13978822	32949818 (125MB)
21	22088104	24840536
22	22278696	24649944
23	21754691	25173949
24	21221966	25706674
25	21973207	24955433
26	23874782	23053858
27	26299537	20629103
28	26567826	20360814 (77MB)
29	26054602	20874038

Expected vs Obtained CR



(b) SPEC2017 Integer + total geometric mean (both integer and floating-point applications).

Expected = ~1.8
Obtained = 1.79

Expected = ~2.6
Obtained = 2.34

Compression Ratios for Other Workloads

Workload	Compression Ratio
parsec_fluidanimate5	2.77 (416MB -> 150MB)
parsec_freqmine5	1.18 (529MB -> 450MB)

VIRGINIA TECH.

Future Work

Future Scope

- Currently, only the size of the compressed blocks are calculated, and the compressed blocks are not actually packed into bit arrays and stored.
- Depending on the value of N chosen, run time can range from 5 to 10 minutes. May be implemented on hardware to significantly make things run faster and in real time.
- Why stop at BDI + GBDI? Try to incorporate multiple algorithms and choose in real-time the algorithm that yields best compression ratio for that workload.

References

References

[1] A. Angerd, A. Arelakis, V. Spiliopoulos, E. Sintorn and P. Stenström, "GBDI: Going Beyond Base-Delta-Immediate Compression with Global Bases," 2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA), Seoul, Korea, Republic of, 2022, pp. 1115-1127, doi: 10.1109/HPCA53966.2022.00085.

[2] D. A. Huffman, “A method for the construction of minimum-redundancy codes,” Proceedings of the IRE, vol. 40, no. 9, pp. 1098–1101, 1952.

Thank You