Revisiting Huffman Coding: Toward Extreme Performance on Modern GPU Architectures

Jiannan Tian*, Cody Rivera†, Sheng Di‡, Jieyang Chen§, Xin Liang§, Dingwen Tao*, and Franck Cappello†‡¶

*School of Electrical Engineering and Computer Science, Washington State University, WA, USA
†Department of Computer Science, The University of Alabama, AL, USA
‡Mathematics and Computer Science Division, Argonne National Laboratory, IL, USA
§Oak Ridge National Laboratory, TN, USA
¶University of Illinois at Urbana-Champaign, IL, USA

Abstract—Today’s high-performance computing (HPC) applications are producing vast volumes of data, which are challenging to store and transfer efficiently during the execution, such that data compression is becoming a critical technique to mitigate the storage burden and data movement cost. Huffman coding is arguably the most efficient Entropy coding algorithm in information theory, such that it could be found as a fundamental step in many modern compression algorithms such as DEFLATE. On the other hand, today’s HPC applications are more and more relying on the accelerators such as GPU on supercomputers, while Huffman encoding suffers from low throughput on GPUs, resulting in a significant bottleneck in the entire data processing. In this paper, we propose and implement an efficient Huffman encoding approach based on modern GPU architectures, which addresses two key challenges: (1) how to parallelize the entire Huffman encoding algorithm, including codebook construction, and (2) how to fully utilize the high memory-bandwidth feature of modern GPU architectures. The detailed contribution is fourfold. (1) We develop an efficient parallel codebook construction on GPUs that scales effectively with the number of input symbols. (2) We propose a novel reduction based encoding scheme that can efficiently merge the codewords on GPUs. (3) We optimize the overall GPU performance by leveraging the state-of-the-art CUDA APIs such as Cooperative Groups. (4) We evaluate our Huffman encoder thoroughly using six real-world application datasets on two advanced GPUs and compare with our implemented multi-threaded Huffman encoder. Experiments show that our solution can improve the encoding throughput by up to 5.0× and 6.8× on NVIDIA RTX 5000 and V100, respectively, over the state-of-the-art GPU Huffman encoder, and by up to 3.3× over the multi-thread encoder on two 28-core Xeon Platinum 8280 CPUs.

I. INTRODUCTION

With the ever-increasing scale of HPC applications, vast volumes of data are produced during simulation, resulting in a bottleneck for both storage and data movement due to limited capacity and I/O bandwidth. For example, Hardware/Hybrid Accelerated Cosmology Code (HACC) [16] (twice finalist nominations for ACM Gordon Bell Prize) produces 20 petabytes of data to store in one simulation of 3.5 trillion of particles with 300 timesteps, whereas leadership-class supercomputers such as Summit [37] have limited storage capacities (around 50~200 PB) to be shared by hundreds of users. On the other hand, network and interconnect technologies in HPC systems advance much more slowly than computing power, causing intra-/inter-node communication cost and I/O bottlenecks to become a more serious issue in fast stream processing [6]. Compressing the raw simulation data at runtime and decompressing them before post-analysis can significantly reduce communication and I/O overheads and hence improving working efficiency.

Huffman coding is a widely-used variable-length encoding method that has been around for over 60 years [17]. It is arguably the most cost-effective Entropy encoding algorithm according to information theory, though some other coding methods such as arithmetic coding and range coding offer slightly better compression ratios in a few specific cases. As such, Huffman coding algorithm serves as the critical step in many general-purpose lossless compression software or libraries such as GZIP [9], Zstd [47], and Blosc [5]. It is also an integral part of many lossy compressors for image and video, such as JPEG [42]. Moreover, Huffman coding is also extensively used in many error-bounded lossy compressors (such as SZ [10, 40] and MGARD [2]), which have been very effective in compressing big scientific datasets with high data fidelity, as verified by many existing studies [24, 22, 46].

In this paper, we focus on parallelizing the entire Huffman encoding algorithm on GPUs rather than the decoding stage. On the one hand, the Huffman encoding algorithm plays a critical role in more and more HPC applications [18, 7] whose runtime performances are heavily relying on the GPU accelerators of supercomputers. On the other hand, compared with decompression performance, compression performance (or encoding efficiency) is particularly important to HPC applications, since large amounts of data need to be compressed on the fly and poor compression performance may substantially counteract performance improvement resulting from the reduced data size, causing inferior I/O performance [23]. By contrast, decompression generally happens only during the post-analysis, which has nothing to do with runtime performance of the simulation.

However, there are no efficient Huffman encoding algorithms designed for the GPU, leaving a significant gap that needs to be filled to meet modern HPC applications’ requirements. Although many multi-thread Huffman encoding algorithms already exist (e.g., Zstd extended its capability to run on multiple CPU cores [1]), their design is limited to coarse-grained par-
allelism, which is unsuitable for today’s modern GPU architectures (featuring massive single-instruction-multiple-thread (SIMT) mechanisms and high memory-bandwidth features). A few GPU-based Huffman encoders, such as Rahmani et al.’s encoder [32], adopt a rather simple parallel prefix sum algorithm to calculate the location of each encoded symbol, which cannot fully utilize the GPU memory bandwidth due to masses of movements of fragmented and variable-length data cells. Moreover, it is worth noting that traditionally, Huffman coding is used in cases where there are 8 bits per symbol (i.e., 256 symbols in total), which is far less than enough for many emerging HPC use cases. Error-bounded lossy compression, for example, often requires more than 8 bits per codeword (e.g., 16 bits are required if 65536 symbols are used in the codebook), because of potentially large amount of integer numbers produced after the error-bounded quantization step [40]. However, constructing a large Huffman codebook sequentially may incur a significant performance bottleneck to the overall Huffman encoding on GPUs, which was not addressed by any prior work.

To address the significant gap, we present an efficient Huffman encoder on GPUs, which is compatible with many emerging HPC scenarios. The basic idea is leveraging a battery of techniques to optimize performance on modern GPU architectures, based on an in-depth analysis of Huffman encoding stage. Specifically, we optimize the parallel Huffman codebook construction for GPUs and significantly reduce the overhead of constructing the codebook that involves a large number of symbols. Moreover, we propose a novel reduction-based encoding scheme, which can significantly improve the memory bandwidth utilization by iteratively merging codeword groups. To the best of our knowledge, our proposed and implemented Huffman encoder is the first work that achieves hundreds of GB/s encoding performance on V100 GPU. The detailed contributions are listed as follows.

- We carefully explore how parallelization techniques can be applied to the entire Huffman encoding algorithm including histogramming, codebook construction, and encoding, and optimize each stage using state-of-the-art CUDA APIs such as Cooperative Groups.
- We develop an efficient parallel codebook construction on GPUs, especially for scenarios requiring a codebook with a large number of symbols, opening new possibilities for non-traditional use cases of Huffman coding.
- We propose a novel reduction-based encoding scheme that iteratively merges the encoded symbols, significantly improving GPU memory bandwidth utilization.
- We evaluate our Huffman encoder on six real-world datasets using two state-of-the-art GPUs and compare it with other state-of-the-art Huffman encoders on both CPUs and GPUs. Experiments show that our solution can improve the encoding throughput by up to $6.8 \times$ on V100 and $3.3 \times$ on CPUs.

In §II, we present the background for Huffman coding and parallel algorithms for its codebook construction. In §III, we discuss the limitation of current Huffman encoding on GPUs. In §IV, we present our proposed parallel Huffman codebook construction and reduction-based encoding scheme on GPUs. In §V, we show the experimental evaluation results. In §VI and §VII, we discuss the related work and conclude our work.

II. BACKGROUND

A. Huffman Coding and Its Emerging Applications

Huffman coding is a fundamental data compression algorithm proposed by David Huffman in 1952 [17]. In essence, it assigns codes to characters such that the length of the code depends on the relative frequency of the corresponding character (a.k.a., input symbol). Huffman codes are variable-length and prefix-free. Here prefix-free means no code is a prefix of any other. Any prefix-free binary code can be visualized as a binary tree (called the Huffman tree) with the encoded characters stored at the leaves.

In recent years, data reduction attracts more and more attention in the HPC field, and Huffman coding becomes an integral part of many data reduction techniques such as error-bounded lossy compression [10, 25, 2]. For multiple HPC use cases, Huffman coding usually needs to be customized with a large number of symbols, instead of using the classic Huffman coding with only 256 symbols/characters in the codebook. For example, SZ requires a customized Huffman coding with 65536 quantization bins in default, such that a large majority of integer codes generated by its quantization could be covered by the encoding scheme. Such a customized Huffman coding is particularly critical when the data is difficult to be predicted accurately, which is very common in scientific datasets.

Another important scenario is n-gram compression [21]. For example, some languages have morphology in the structure of words or morphemes, and it is important to utilize this syllable-based morphology for developing an efficient text compression approach for these languages. Nguyen et al. proposed a method [28] to partition words into its syllables and then to produce their bit representations for compression. The number of bits in syllables (symbols) depends on the number of entries in the dictionary file. As another example, segmenting, encoding, and decoding DNA sequences based on n-gram statistical language model is a critical research topic in bioinformatics to handle the vast volumes of DNA sequencing data. Specifically, in this work [21], researchers find the length of most DNA words/symbols (e.g., 12~15 bits) and build an n-gram biology language model by analyzing the genomes of multiple model species. Then, they design an approach to segment the DNA sequences and encode them accordingly.

In all the above cases, Huffman coding may require generating a codebook with a large number of symbols which is usually far smaller than that of the input codewords. However, since such a large codebook is generated serially, codebook construction can become a significant bottleneck, especially on small to medium-sized datasets. Thus it is vital to develop an approach to build a Huffman codebook efficiently.

B. PRAM Model

PRAM is a classic model to describe parallel algorithms where multiple processors are attached to a single memory en-
C. Parallel Huffman Codebook Construction

The serial Huffman codebook construction algorithm with the complexity of $O(n \log n)$ constructs a naïve binary tree—a data structure not well-suited for the GPU memory. Specifically, the naïve Huffman tree has an inefficient GPU memory access pattern, which would incur a significant performance overhead on codebook construction. This is confirmed by our experiment: constructing a Huffman codebook with 8,192 input symbols takes 144 ms on NVIDIA V100 GPU, which degrades the throughput of compressing 1 GB data to less than 10 GB/s.

Obviously, it is very important to develop an efficient parallel codebook construction algorithm on GPU to match the high speed of other stages in Huffman encoding. To this end, we review the literature carefully for existing parallel Huffman tree and codebook construction algorithms. Larmore and Przytycka [20] proposed a parallel Huffman tree construction algorithm using $n$ processors with $O(\sqrt{n} \log n)$ time per processor, the first algorithm whose processor count scales linearly with the number of input symbols. Here both the number of processors and the number of input symbols are $n$. However, the proposed algorithm is known for its inefficiency of performing $O(n^2)$ work. Millidu et al. [27] later proposed another CREW PRAM algorithm for the same problem, with $n$ processors, $O(H \cdot \log \log \frac{H}{n})$ time per processor, and $O(n)$ work, where $H$ is the length of the longest codeword. Since Huffman codes can be up to $O(n)$ in length, the proposed algorithm has a worst case performance of $O(n)$ per processor, but this is rarely encountered in practice, especially in HPC scenarios whose floating-point data tends to be mostly smooth and predictable.

The previously discussed algorithms all output Huffman trees, where the trees still need to be traversed serially to generate a codebook. To address this issue, Ostadzadeh et al. proposed a CREW PRAM algorithm that directly generates the codebook [31]. To do this, it generates the length of each symbol, with $n$ processors and $O(H \cdot \log \log \frac{H}{n})$ time per processor, and then converts the generated symbol lengths into codes, with $n$ processors and $O(H)$ time per processor. We propose our Huffman codebook construction method based on this algorithm due to its direct output of Huffman codes as well as its outstanding performance on most Huffman work.

III. Problem Statement

In this section, we first discuss the scalability constraints posed by modern GPU architectures, especially the CUDA architecture. We then analyze the performance bottleneck of the state-of-the-art Huffman encoding method. Finally, we formulate our research problem under certain constraints.

A. Scalability Constraints from GPU Hardware

First, we analyze the scalability constraints from the GPU hardware perspective. The thread is the basic programmable unit that allows the programmer to use the massive amount of CUDA cores. CUDA threads are grouped at different levels, including warp, block, and grid levels.

1) Rigid SIMD-ness Against Randomness: The warp is a basic-level scheduling unit in CUDA associated with SIMD (single-instruction multiple-data). Specifically, the threads in a warp achieve convergence when executing exactly the same instruction; otherwise, warp divergence happens. In the current CUDA architecture, the number of threads in a warp is 32, hence, it works as 32-way SIMT when converging. However, when diverging happens, it may cause discrepancy from the PRAM model, because diverged threads add extra overhead to the execution. Thus, we relax the use of the PRAM model. Nevertheless, the GPU's massive parallelism allows our implementation to exhibit the theoretical complexity of parallel Huffman coding under PRAM.

2) Block-Shared Memory Lifecycle Binding: Unlike the warp, the thread block (or simply block) is a less hardware-coupled description of thread organization, as it is explicitly seen in the kernel configuration when launching one. Threads in the same block can access the shared memory, a small pool of fast programmable cache. On one hand, shared memory is bound to active threads, which are completely scheduled by the GPU hardware; however, on the other hand, a grid of threads may exceed the hardware supported number of active threads at a time. As a result, the data stored in the shared memory used by the previous batch of active threads may be invalid when the current or following batch of active threads are executing.

Therefore, we must make use of both coarse- and fine-grained parallelization in our design due to the scalability constraints from the CUDA architecture. For coarse-grained parallelization, we divide the data into multiple independent chunks, not only because it is easy to map chunks to thread blocks and utilize local shared memory, but also because it will facilitate the reverse process, decoding. In addition, coarse-grained chunking can improve performance significantly with only a minimum overhead in compression ratio. For fine-grained parallelization, the state-of-the-art work [32] only addresses the encoding stage rather than Huffman codebook construction. In the next section, we will further analyze the performance issue of the existing fine-grained encoding approach.

B. Fully Enabling GPU’s High Memory Bandwidth

Compared to compute-bound algorithms such as matrix-matrix multiplication, Huffman encoding tends to be more memory-bound [15]. In practice, Huffman encoding that has
canonizing the codebook and Huffman codebook generation/construction

calculating the frequencies of all input symbols, namely, encoding to develop an adaptive

to get blockwise code len

to get codeword lengths

to get codewords

to canonize

to get numl array

to get reverse codebook

to reduce-merge

to shuffle-merge

to coalescing copy

Table I: Parallelism implemented for Huffman coding’s subprocedures (kernels). “sequential” denotes that only 1 thread is used due to data dependency. “coarse-grained” denotes that data is explicitly chunked. “fine-grained” denotes that there is a data-thread mapping with little or no warp divergence.

IV. DESIGN METHODOLOGY

In this section, we propose our novel GPU Huffman encoding design for the CUDA architecture. We propose several optimizations for different stages. Specifically, we modularize the Huffman encoding into the following four stages: 1) calculating the frequencies of all input symbols, namely, histogramming; 2) Huffman codebook generation/construction based on the frequencies; 3) canonizing the codebook and generating the reverse codebook for decoding; and 4) encoding according to the codebook, and concatenate Huffman codes into a bitstream. We first propose an efficient, fine-grained parallel codebook construction on GPUs, especially for scenarios requiring a large number of symbols (stage 2). We then propose a novel reduction-based encoding scheme that iteratively merges the encoded symbols, which significantly improves memory bandwidth utilization (stage 4).

A. Histogramming

The first stage of Huffman encoding is to build a histogram representing the frequency of each integer-represented symbol from the input data. The GPU histogramming algorithm in use is derived from that proposed by Gómez-Luna et al. [13]. This algorithm minimizes conflicts in updating the histogram bin locations by replicating the histogram for each thread block and storing the histogram in shared memory. Where possible, conflict is further reduced by replicating the histogram such that each block has access to multiple copies. All threads inside a block read a specified partition of the input and use atomic operations to update a specific replicated histogram. As each block finishes its portion of the predicted data, the replicated
histograms are combined via a parallel reduction into a single global histogram, which is used to construct the final codebook.

B. Two-Phase Canonical Codebook Construction

In the second stage, we implement an efficient parallel Huffman codebook construction algorithm on the GPU and modify it to produce canonical codes for fast decoding.

1) Codebook Construction: Now that we have a single global histogram, the next step is to efficiently construct a base codebook. We implement the parallel codebook construction algorithm proposed by Ostadzadeh et al., as described earlier, on the GPU [31]. This algorithm provides a parallel alternative to the original Huffman codebook construction algorithm in \( O(n \log n) \) and directly generates codewords. To the extent of our knowledge, this algorithm has not been implemented on the GPU elsewhere. The algorithm is split into two phases, GenerateCL, which calculates the codeword length for each input symbol, and GenerateCW, which generates the actual codeword for each input symbol. Both phases utilize fine-grain parallelism, with one thread mapped to one input symbol or intermediate value. Additionally, both phases are implemented as single CUDA kernels with Cooperative Groups [8], which we use to synchronize an entire CUDA grid. We describe both phases in Algorithm 1, with our modifications colored blue, and emphasize our GPU implementation in the following discussion. We refer readers to [31] for more details of the original algorithm.

GenerateCL takes \( F \), a sorted \( n \)-symbol histogram, and outputs \( CL \), a size \( n \) array of codeword lengths for each symbol. This phase of the algorithm runs in \( O(H \cdot \log \log n) \) time on PRAM, where \( H \) is the longest codeword. Its parallelism can be derived from the fact that for a given set of Huffman sub-trees, all sub-trees whose total frequencies are less than the sum of the two smallest sub-tree frequencies can be combined in parallel, a result which Ostadzadeh et al. prove [31].

Before GenerateCL is launched, the histogram is sorted in ascending order using Thrust [30]. This operation is low-cost, as \( n \) is relatively small compared to the input data size. Once launched, lines 1–4 of Algorithm 1 initialize the array \( iNodes \) with each input symbol’s leaf node, and initialize the array (and queue) \( iNodes \) as empty. Each array element describes a Huffman node, storing its total frequency, its leader, or topmost parent, and auxiliary information. To increase memory access efficiency, each of these arrays are stored in structure-of-arrays format, rather than the intuitive array-of-structures format. The advantage of this is that accesses to single fields of consecutive elements are coalesced.

Next, lines 5–26 construct the Huffman tree while there are still leaf and internal nodes to process. Lines 6–16 create a new node \( t \) from the smallest two leaf or internal nodes, and selects leaf nodes whose frequencies are less than \( t \). ParMerge (line 17) merges these selected leaf nodes and internal nodes, which are sorted by ascending frequency, together. This is done using a \( O(\log \log n) \) parallel merge under the PRAM model [31].

To implement this merge, we customize the parallel GPU Merge Path algorithm proposed by Green et al. [14] for our intermediate structure-of-arrays representations in GenerateCL. In practice, this algorithm does not attain the proposed theoretical time complexity, as its time complexity is \( O(n/p + \log n) \), with \( p \) being the number of partitions (i.e., the number of thread blocks). However, we use a number of thread blocks proportional to the number of streaming multiprocessors (SMs), making \( n/p \) in practice \( O(\log n) \). This component of parallel codebook construction employs coarse-grain parallelism, as once the merge partitions are determined, each partition is merged serially. To remove the overhead of calling a separate kernel with dynamic parallelism, we incorporate ParMerge into the same kernel as the rest of GenerateCL, keeping unnested threads idle until the merging phase. Once these nodes are merged, they are melded together into new nodes, with the appropriate CL values and leaf nodes being updated, on lines 18–25, and the iteration is repeated as appropriate.

GenerateCW takes CL as input and outputs CW (i.e., the actual codewords). It takes \( O(H) \) time per thread in the PRAM model, where \( H \) is the longest codeword, and our GPU implementation of this algorithm is shown in Algorithm 2.
is consistent with this theoretical complexity (see Table III). This phase of the algorithm utilizes fine-grained parallelism, as codewords are generated by individual threads. It assigns numerically increasing codewords to all symbols with a given codeword length CCL in parallel (lines 31–39). If there are more codewords to generate that are longer than CCL, first, CCL and other local variables are updated (lines 42–44). Also, the first codeword for the new codeword length is generated by incrementing the existing codeword and left-shifting the codeword by the difference between the old and new codeword lengths (line 43). Finally, lines 31–39 are repeated. Once all codewords are generated, \( C\) is reversed, and the codewords are sorted by the actual input symbol they represent to generate the forward codebook.

It is worth noting that throughout codebook construction, we use the state-of-the-art CUDA Cooperative Groups instead of existing block synchronization for global synchronization [8]. This is because CUDA blocks are limited to 1024 threads, and we employ fine-grained parallelism for codebook sizes greater than 1024. An alternative technique to achieve the same global synchronization is to separate the parallel regions into different kernels. We chose Cooperative Groups over this technique to avoid the overhead of kernel launches and cudaDeviceSynchronize. Our profiling for the NVIDIA V100 GPU reveals that a CUDA kernel launch takes about 60 microseconds (60 \(\mu s\)), and each of our parallel regions performs very little work. Also, since many of these parallel regions are performed in a loop, we effectively avoid unnecessary CPU-GPU transfers.

2) Canonizing Codebook: A canonical Huffman codebook \([33]\) holds the same bitwidth of each codeword as the original Huffman codebook (i.e., base codebook). Its bijective mapping between input symbol and Huffman codeword is more memory-efficient than Huffman-tree traverse for encoding/decoding. The time complexity of serially building a canonical codebook from the base codebook is \(O(n)\), where \(n\) is the number of symbols, and is sufficiently small compared with the data size. By using a canonical codebook, we can (1) decode without the Huffman tree, (2) efficiently cache the reverse codebook for high decoding throughput, and (3) maintain exactly the same compression ratio as the base Huffman codebook.

We start our implementation which contains a partially-parallelized canonization CUDA kernel, utilizing Cooperative Groups. It performs (1) linear scanning of the base codebook (sequentially \(O(n)\)), which is parallelized at fine granularity with atomic operations; (2) loose radix-sorting of the codewords by bitwidth (sequentially \(O(n)\)), which cannot be parallelized because of the intrinsic RAW dependency; and (3) building the reverse codebook (sequentially \(O(n)\)), which is enabled with fine-grained parallelism. The canonization process is relatively efficient—it only costs about 200 us to canonize a 1024-codeword codebook on V100.

Nevertheless, our choice of codebook construction algorithm provides a sufficient base for further optimization. Since the output of GenerateCL and input of GenerateCW, CL, is sorted by codeword length, the intrinsic RAW dependency of \(2\) is removed. In fact, the existing codewords generated by GenerateCW are almost canonical, except for the fact that given two codewords \(c_1\) and \(c_2\) where \(c_2\) is longer than \(c_1\) and \(l\) is \(c_1\)'s length, the most significant \(l\) bits of \(c_2\) are numerically greater than \(c_1\). The opposite should be the case, as efficient decoding relies on this fact. To alleviate this problem, the codewords for each level are generated in decreasing order per level (line 38 of Algorithm 1), and the bits in each codeword are then inverted before they are stored (line 47). Moreover, additional metadata to facilitate decoding also needs to be generated in \(O(1)\) extra time with little storage overhead. The metadata that we generate consists of two \(H\)-element arrays, where \(H\) is the longest codeword's length. The First array contains the first codeword for a given length, and the Entry array contains a prefix sum of the number of codewords shorter than that length. Right after all the codewords for a given length are generated, the First and Entry arrays are updated appropriately at that index (lines 40–41). Both arrays are used for efficient treeless canonical decoding.

The theoretical time complexity of our codebook construction, including our modifications to generate canonized codes, is \(O(H \cdot \log \log \frac{H}{p})\); however, due to the particular implementation of merge, which is the most expensive operation, the practical complexity is increased to \(O(H \cdot \frac{H}{p} + \log \frac{H}{p})\), where \(H\) is the longest codeword length, \(n\) is the number of symbols, and \(p\) is the number of blocks launched. Nevertheless, due to \(H\) in practical being small and \(p\) being sufficiently large, we observe the complexity of \(\widetilde{O}(\log n)\), and our experiments are consistent with this (see Table III).

C. Encoding

We propose an iterative merge comprised of reduce-merge and shuffle-merge that is the key to improving memory bandwidth utilization for encoding. In each iteration, every two codewords are merged into one, with their lengths summed up. Formally, given two code-length tuples \((a, \ell)_{2k}\) and \((a, \ell)_{2k+1}\), we define

\[
\text{Merge}\left((a, \ell)_{2k}, (a, \ell)_{2k+1}\right) = (a_{2k} \oplus a_{2k+1}, \ell_{2k} + \ell_{2k+1}),
\]

where \(\oplus\) represents for concatenating bits of \(a_{2k+1}\) right after bits of \(a_{2k}\). Note that the merge is not commutative for the encoded symbols and must follow the original order.

We further split this merge into reduce-merge and shuffle-merge phases. Note that the first merge includes a codebook lookup to get the codewords. After that, the merge is performed iteratively on two codewords each time.

a) Reduce-Merge: The practical average bitwidth of Huffman codewords can be fairly low (e.g., most are 1 or 2 bits in many HPC datasets), while data movement is in terms of single-/multi-byte words (i.e. a multiple of 8 bits, such as uint\{8,16,32\}_t). This can significantly hurt the performance due to an extravagant use of threads on the GPU if the data-thread mapping is too fine, e.g., 1-to-1. In each iteration of reduction, active threads for data movement halve every iteration before bits saturates the representing words, leading to a waste of parallelism. Hence, we map multiple
two segments | t0 t1 t2 t3 t4 t5 t6 t7 t8 t9 t10 t11 t12 t13 t14 t15
---|---
two parts of the right | t0 t1 t2 t3 t4 t5 t6 t7 t8 t9 t10 t11 t12
move part 1/2 | t8 t9 t10 t11 t12
move part 2/2 | t8 t9 t10 t11 t12

Fig. 2: Two-step batch move of grouped and typed data. By batch-moving the right grouped data, warp divergence is decreased.

codewords to one thread to merge until the merged bitwidth exceeds that a representing word can handle. More precisely, the condition of stopping reduce-merge is that the average bitwidth of the merged codeword exceeds half of the bitwidth of the data type. For example, merging codewords with an average bitwidth of 2.3 bits for 3 times is expected to result in an 8× length, averagely 18.4 bits, and end before going beyond 32 bits. The number of iterations (denoted by reduction factor r) can be determined by the entropy in bits of input data (from histogram). After reduce-merge is done, the number of merged codewords is shrunk by 2^r×.

b) **Shuffle-Merge**: After reduce-merge completes, threads are grouped to move the corresponding merged codewords, with one thread assigned to one unit of typed data. Following the same scheme as reduce-merge, a right group of typed data is moved to its corresponding left group. To provide more detail, a left group (A, ℓW)2k, with data segment (the representation data type or word is marked with W) and its length ℓW, has a starting index i2k (already known) and an ending index i2k• = (i2k + ℓ2k/ℓW) easy to calculate. Moreover, the ending bit’s location can be calculated as ℓ2k• = ℓ2k + ℓW, and the number of residual bits is ℓ2k• = ℓW − ℓ2k•. Thus, the right group (A, ℓ)2k+1 needs [ℓ2k+1/ℓH] threads for data movements. For each thread, the ℓ2k• bits are first moved to fill the residual bits, and then the ℓ2k• bits are moved right after the ℓ2k• bits (in the next typed data cell), as shown in Figure 2. Note that this process is free of data contention. The iterative process will be performed for s times (denoted by shuffle factor) until a dense bitstream is formed. We also note that shuffle-merge can be finished within a continuous memory space of 2^s typed data cells.

c) **Interface**: Our encoding kernel is interfaced as

```
ReduceShuffleMerge(M, r) (in, out, metadata)
```

We expose two independent parameters to describe the problem size and the two merge phases—magnitude M, reduction factor r, and shuffle factor is derived from s = M − r. In total, there are M iterations in the entire encoding stage. After that, we generate a densely encoded bitstream for this chunk. We use N = 2^M to denote the problem size for each chunk, and n = 2^n to denote the reduced size before shuffle-merge.

We use the average codeword bitwidth to determine the reduction factor r. Specifically, given a word W of length ℓW, a “proper” r is tentatively determined according to [log β]+1 = ℓW, such that the length of the r-time-merged codeword ℓ(r) is expected to be ℓW/2 ≤ ℓ(r) ≤ ℓW, toward maximized memory bandwidth utilization. Note that there are certain codewords that exceed 32 bits after r-time merge, and these (denoted by “breaking” in Table II) are filtered out and handled otherwise.

We give a quantitative example to illustrate the importance of determining a proper reduction factor and how magnitude may affect the performance due to longer shuffle-merge. We compare a magnitude of {12, 11, 10} and reduction factor of {4, 3, 2} on 8×-Quant from baryon-density field, whose average bitwidth is 1.02717 with 1024 symbols. The performance is shown in Table II. We find that the combination of M=10 and r=3 results in the highest performance, which can be empirically generalized. Although reducing magnitude would result in more metadata, as we prioritize the performance in this work, we choose the combination of M = 10 and r = 3 for the following evaluation.

**TABLE II**: Performance (in GB/s) of our Huffman encoding with different chunk magnitudes (mag.) and reduction factors on Longhorn and Frontera.

<table>
<thead>
<tr>
<th>mag</th>
<th>r=4</th>
<th>r=3</th>
<th>r=2</th>
<th>r=1</th>
<th>breaking</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>227.60</td>
<td>274.40</td>
<td>291.04</td>
<td>199.41</td>
<td>314.63</td>
</tr>
<tr>
<td>8</td>
<td>110.94</td>
<td>124.42</td>
<td>133.84</td>
<td>94.27</td>
<td>105.85</td>
</tr>
<tr>
<td>4</td>
<td>42.70</td>
<td>55.53</td>
<td>79.45</td>
<td>0.000434K</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>119.41</td>
<td>124.42</td>
<td>133.84</td>
<td>94.27</td>
<td>105.85</td>
</tr>
<tr>
<td>1</td>
<td>42.70</td>
<td>55.53</td>
<td>79.45</td>
<td>0.000434K</td>
<td></td>
</tr>
</tbody>
</table>

```
d) **Complexity**: For reduce-merge, we map multiple codewords for reduction such that we can effectively move more bits against the given holding codeword W with length ℓW. We maintain a block of 2^s threads for data movement. The time complexity is 2^r−1 i for the i-th iteration, and the time complexity for reduction in parallel is ∑̇ 2^r−1. Note that the operations are homogeneous (i.e. there is only one if-branch) without being affected from warp divergence.

For shuffle-merge, the magnitude of reduced data chunk remains s, and each typed data cell is assigned with a thread. At the i-th iteration, the chunk is split into (s + 1 − i) groups. Compared with reduction, shuffle creates warp divergence at a factor of 2, given the groups are to merge with their corresponding other groups. With s parallel shuffle-merges, the total time complexity is O(s). Note that bank conflicts may affect performance, as the read and written locations inevitably
overlap due to the variable length \(^1\). This proof-of-concept work is intended to show the effectiveness of the method, the effect of bank conflict is to be further investigated.

V. PERFORMANCE EVALUATION

In this section, we present our experimental setup (including platforms, baselines, and datasets) and our evaluation results.

A. Experiment Setup

1) Evaluation Platforms: We conduct our experimental evaluation using the Frontera supercomputer \([11]\) and its subsystem Longhorn \([26]\). We perform our experiments on an NVIDIA Tesla V100 GPU from Longhorn and an NVIDIA Quadro RTX 5000 from Frontera, and compare with CPU implementations on two 28-core Intel Xeon Platinum 8280 CPUs from Frontera.\(^2\) We use NVIDIA CUDA 10.1 and its default profiler to measure the time. In this section, we use \textbf{TU} to denote Turing RTX 5000 and \textbf{V} to denote Volta V100.

2) ComparisonBaselines: CUHD \([43]\) and cuSZ \([41]\) are two state-of-the-art Huffman encoders for GPUs, but both of them are coarse-grained and embarrassingly parallelized. We note that CUHD’s source code \([44]\) only focuses on GPU Huffman decoding and implements a serial CPU Huffman encoder, so we compare our GPU Huffman encoder only with cuSZ \([41]\). We also compare our GPU encoder with the serial encoder implemented in SZ \([39]\) and with our implemented multi-thread encoder\(^3\) on single and multiple CPU cores, respectively.

3) Test Datasets:

a) Single-Byte Based Datasets: Generic Huffman coding takes one byte per symbol, hence, at most 256 symbols in total. Without reinterpreting a byte stream into multi-byte (un)signed integers or floating-point numbers, the generic encoding simply treats all input data as uint8_t. Our evaluation includes these popular datasets: \(1\) enwik8 and enwik9 from \textit{Large Text Compression Benchmark} \([19]\), the first 10\(^8\) and 10\(^9\) bytes of XML-based English Wikipedia dump; \(2\) nc from \textit{Silesia Corpus} \([35]\), a file for chemical database of structures; \(3\) mr from \textit{Silesia Corpus} \([35]\), a sample file of medical magnetic resonance image; and \(4\) Flan\_1565 from \textit{SuiteSparse Matrix Collection} \([36]\), a sparse matrix in Rutherford Boeing format.

b) Multi-Byte Based Datasets: We also evaluate two datasets with multiple bytes as a symbol: \(5\) Nyx-Quant is the quantization codes generated by SZ (a famous error-bounded lossy compression for HPC data) based on Nyx’s (cosmological simulation) baryon density from \textit{Scientific Data Reduction Benchmarks} \([34]\); and \(6\) gbcbti.seq is a sample DNA sequence data from \textit{GenBank} \([4]\), where every \(k\) nucleotides (\(k\)-mer) forms a symbol. We test \(k = \{3, 4, 5\}\) in our evaluation.

\(^1\)Due to the page limit, we refer readers to the discussion of time complexity at https://github.com/szcompressor/huffre/blob/main/doc/benchmark.md.

\(^2\)NVIDIA Tesla V100 GPU from Longhorn and an NVIDIA Quadro RTX 5000 from Frontera.

\(^3\)Note that SZ’s current OpenMP version \([38]\) only divides data into multiple blocks and applies its compression to each block independently.

B. Experimental Results

1) Parallel Codebook Construction: Call back to §IV-B1, we observe a practical complexity of \(O(H \cdot \log(n/H))\) or approximately \(O(\log(n))\), where \(H\) and \(n\) are the height of the built tree and the problem size, respectively. Although we exhibit speedups in codebook construction for \(n = 256\) ranging from 2.0–2.9 in Table V, greater benefits come from using more input symbols since the \(O(n \cdot \log(n))\) serial construction algorithm scales slower. To demonstrate this speedup, we use the gbbcti.seq gene dataset with \(k\)-mer analytics, where \(k = 3, 4, 5\). We also evaluate our codebook construction on the quantization codes generated by cuSZ from the Nyx dataset. Note that data other than the 4 bases of DNA are stored in gbbcti.seq, and as a result, the number of input symbols needed is greater than \(4^k\).

<table>
<thead>
<tr>
<th>#sym.</th>
<th>serial</th>
<th>gen. codebook</th>
<th>canonize</th>
<th>total time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nyx-Quant</td>
<td>1024</td>
<td>0.045</td>
<td>3.051</td>
<td>3.089</td>
</tr>
<tr>
<td>3-mer</td>
<td>2048</td>
<td>0.208</td>
<td>8.381</td>
<td>9.760</td>
</tr>
<tr>
<td>4-mer</td>
<td>4096</td>
<td>0.695</td>
<td>20.148</td>
<td>24.684</td>
</tr>
<tr>
<td>5-mer</td>
<td>8192</td>
<td>1.806</td>
<td>61.748</td>
<td>59.092</td>
</tr>
</tbody>
</table>

Table III compares GPU codebook construction between cuSZ’s serial implementation and our parallel implementation on several datasets with different numbers of input symbols. Ours exhibits more dramatic speedups over cuSZ’s when using more input symbols, consistent with our theoretical analysis and performing up to 45.5 \(\times\) faster when creating a codebook for 8192 symbols. Note that ours is no faster than the CPU serial construction when the number of symbols is below 8192. This is because caching, high frequency, and superior branch prediction in combination result in low latency of CPU threads. However, to avoid long histograms and other CPU-GPU interaction, it is desirable to purely perform codebook construction on the GPU.

<table>
<thead>
<tr>
<th>#sym.</th>
<th>serial</th>
<th>gen. CL</th>
<th>gen. CW</th>
<th>total time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nyx-Quant</td>
<td>1024</td>
<td>0.045</td>
<td>0.315</td>
<td>0.208</td>
</tr>
<tr>
<td>3-mer</td>
<td>2048</td>
<td>0.208</td>
<td>0.494</td>
<td>0.570</td>
</tr>
<tr>
<td>4-mer</td>
<td>4096</td>
<td>0.695</td>
<td>0.633</td>
<td>0.682</td>
</tr>
<tr>
<td>5-mer</td>
<td>8192</td>
<td>1.806</td>
<td>1.330</td>
<td>1.145</td>
</tr>
</tbody>
</table>

Table IV: Performance comparison of Huffman codebook construction time (in milliseconds) on RTX 5000 and V100 with different numbers of symbols.

<table>
<thead>
<tr>
<th>#sym.</th>
<th>serial</th>
<th>1 core</th>
<th>2 cores</th>
<th>4 cores</th>
<th>6 cores</th>
<th>8 cores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nyx-Quant</td>
<td>1024</td>
<td>0.045</td>
<td>0.219</td>
<td>0.419</td>
<td>0.622</td>
<td>0.700</td>
</tr>
<tr>
<td>3-mer</td>
<td>2048</td>
<td>0.208</td>
<td>0.361</td>
<td>0.691</td>
<td>1.101</td>
<td>1.122</td>
</tr>
<tr>
<td>4-mer</td>
<td>4096</td>
<td>0.626</td>
<td>1.006</td>
<td>1.309</td>
<td>1.456</td>
<td>1.702</td>
</tr>
<tr>
<td>5-mer</td>
<td>8192</td>
<td>1.187</td>
<td>1.167</td>
<td>1.513</td>
<td>1.657</td>
<td>1.836</td>
</tr>
<tr>
<td>Synthetic</td>
<td>16384</td>
<td>3.671</td>
<td>1.683</td>
<td>1.796</td>
<td>1.785</td>
<td>2.055</td>
</tr>
<tr>
<td>Synthetic</td>
<td>32768</td>
<td>7.783</td>
<td>2.974</td>
<td>2.858</td>
<td>2.826</td>
<td>2.873</td>
</tr>
<tr>
<td>Synthetic</td>
<td>65536</td>
<td>15.641</td>
<td>5.221</td>
<td>4.850</td>
<td>4.411</td>
<td>4.952</td>
</tr>
</tbody>
</table>

Table IV: Performance (in milliseconds) of multi-thread codebook construction with different numbers of input symbols. The length of the bar under the number reflects the execution time.

Moreover, since SZ \([39]\) currently does not support building the Huffman tree in parallel, we also implement a multi-thread
codebook construction using OpenMP. We evaluate its performance and compare it with SZ’s serial codebook construction on our tested datasets with 1024 ~ 8192 symbols and synthetic normally-distributed histograms4 with 16384~65536 symbols, as shown in Table IV. We note that in many cases, even with only one thread, its performance is better than the serial construction, as it uses internal cache-friendly arrays rather than the binary trees and priority queues used by the serial construction. We also note that on our test datasets (with codebooks in the order of 10^3), the multi-thread construction does not improve the performance because the OpenMP introduces more overhead than it reduces from multiple threads.

We find that our multi-thread CPU codebook construction needs least 32768 symbols to be able to overcome the OpenMP overhead and obtain a speedup using multiple threads. Unlike CPU, GPU parallel construction can always yield a speedup over serial construction in our tested cases. This is due to the relatively high latency and low performance of a single GPU thread. On the other hand, parallelization and synchronization are relatively low-latency on the GPU. Furthermore, since the GPU can launch vastly more threads than the CPU, it takes advantage of fine-grained parallelism in our parallel codebook construction algorithm.

2) Encoding: Without loss of generality, we evaluate our Huffman encoder on multiple datasets with various types, as shown in Table V. Most of our tested datasets have a relatively large average bitwidth (e.g. 4 or more bits vs. uncompressed 8 bits), leading to a relatively low compression ratio. According to the aforementioned r decision-making mechanism, only nci and Nyx-Quant can use r=3 (potentially r=4 for Nyx-Quant), making their throughputs over 100 GB/s. While compared with Nyx-Quant, nci has a relatively small data size, so it is difficult to undersaturate the memory bandwidth. Moreover, Nyx-Quant has much lower writing effort due to 2.66× higher compression ratio than nci, so its encoding throughput is high, at 314.6 GB/s.

As mentioned in §IV-C, the rigid fixed-size typed data makes it possible to meet conflict. For example, when merging Flan_1565, there is less than 1.4e-6% of the data that breaks the fixed size, while there is 3e-3% of the data for Nyx-Quant.

![Table V: Breakdown comparison of Huffman performance on tested datasets. Gathering time is excluded.](image)

TABLE VI: Performance of multi-thread Huffman encoder on Nyx-Quant.

<table>
<thead>
<tr>
<th>cores</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>256</th>
</tr>
</thead>
<tbody>
<tr>
<td>hist. (GB/s)</td>
<td>2.24</td>
<td>4.42</td>
<td>8.83</td>
<td>17.41</td>
<td>34.97</td>
<td>63.59</td>
<td>81.47</td>
<td>63.14</td>
<td>145.20</td>
</tr>
<tr>
<td>par. efficiency</td>
<td>1.00</td>
<td>0.99</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.94</td>
<td>0.94</td>
<td>0.78</td>
<td>0.44</td>
</tr>
<tr>
<td>codebook (ms)</td>
<td>0.22</td>
<td>0.45</td>
<td>0.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hist-enc (GB/s)</td>
<td>0.79</td>
<td>1.57</td>
<td>3.12</td>
<td>6.23</td>
<td>12.38</td>
<td>23.73</td>
<td>29.22</td>
<td>20.61</td>
<td>45.35</td>
</tr>
<tr>
<td>par. efficiency</td>
<td>1.00</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
<td>0.97</td>
<td>0.94</td>
<td>0.37</td>
<td></td>
</tr>
</tbody>
</table>

Our solution is to backtrace the breaking points in batch, which starts at the last iteration of reduce-merge and refers to the 2^3 points that go beyond 32-bit limit together. The reduction is about 300 µs, including one-time read from global memory. The total number of breaking points (in percentile) are shown in the “breaking” column in Table V, which is negligible to affect the compression ratio. After we filter out the breaking data, we can use the state-of-the-art cuSPARSE API or perform a “device-wide” using ready tools such as NVIDIA::cub [29] to perform a dense-to-sparse conversion to save them. In addition, optional gathering coarse-grained chunks into an even denser format can be done, using the merged length; due to the compression ratio (CR), there is only as much as 1/CR additional data movement, which is a short-time memcpy to the Reduce/Shuffle-merge time. Compared with cuSZ’s implementation, our encoder can improve the performance by 3.1× ~ 5.0× and 3.8× ~ 6.8× on RTX 5000 and V100, respectively, on the tested data.

Finally, we evaluate the performance of our implemented multi-thread encoding and our overall encoder on Nyx-Quant5 with different numbers of CPU cores, as shown in Table VI. For encoding, the multi-thread version achieves a peak performance of 56 GB/s, and maintains high parallel efficiency up to 32 cores (with 56 available cores). However, this is still about 5.6× lower than the performance of our fine-grained Huffman encoding on the V100 (i.e., 314.6 GB/s). For the overall encoder (including histogramming, codebook construction, and encoding), compared with the multi-thread encoder on the CPUs, our GPU version on the V100 improves the histogram-encoding performance by 3.3× (i.e., 29.22 GB/s v.s. 96.01 GB/s). This is because Huffman encoding tends to be memory-bound, and the GPU memory such as HBM2 in the V100 has a much higher bandwidth than state-of-the-art CPU memory such as DRAM4.

4The symbol numbers in the tested real datasets are no more than 8192, so we use synthetic data for more than 8192 symbols. Also, note that 8192 is limited by the current optimal GPU histogramming.

5We note that multi-thread histogramming/encoding have relatively stable throughput, thus we only evaluate on Nyx-Quant for demonstration purpose.
VI. RELATED WORK

We note that several approaches [3, 43] have been proposed to accelerate Huffman decoding on GPUs, yet few studies considered optimizing Huffman encoding by fully utilizing GPU computing power. Despite the limited work on GPU-based Huffman encoding, we still search for some related works and discuss them as follows.

In general, parallel Huffman coding obtains each codeword from a lookup table (generated by a Huffman tree) and concatenates codewords together with other codewords. However, a severe performance issue arises when different threads write codewords of varying lengths, which results in warp divergence on GPU [45]. The most deviation between methods occurs in concatenating codewords. Fuentes-Alventosa et al. [12] proposed a CUDA implementation of Huffman coding with a given table of variable-length codes, reaching 20× the serial CPU encoding performance. Rahmani et al. [32] also proposed a CUDA implementation of Huffman coding based on serially constructing the Huffman codeword tree and parallel generating the bytestream, which can achieve up to 22× over the serial CPU performance by disregarding constraint on the maximum codeword length or data entropy. Lal et al. [18] proposed a Huffman-coding-based memory compression technique for GPUs (called E²MC) based on a probability estimation of symbols. Recently, Tian et al. [41] developed a coarse-grained parallel Huffman encoder for error-bounded lossy compressor on GPUs; however, this implementation does not address the non-coalesced memory issue, reaching only 30 GB/s on V100.

VII. CONCLUSION AND FUTURE WORK

In this work, we propose and implement an efficient Huffman encoder for NVIDIA GPU architectures. Specifically, we develop an efficient parallel codebook construction and a novel reduction based encoding scheme for GPUs. We also implement a multi-thread Huffman encoder for a fair comparison. We evaluate our encoder using six real-world datasets on NVIDIA RTX 5000 and V100 GPUs. Compared with the state-of-the-art Huffman encoder, our solution can improve the parallel encoding performance up to 5.0× on RTX 5000, 6.8× on V100, and 3.3× on CPUs. We plan to further optimize the performance in the future work. For example, we seek to (1) tune the performance for low-compression-ratio data, (2) explore more efficient gathering methods, and (3) explore how intrinsic data feature affects the compression ratio and the throughput.

ACKNOWLEDGEMENTS

This research was supported by the Exascale Computing Project (ECP), Project Number: 17-SC-20-SC, a collaborative effort of two DOE organizations—the Office of Science and the National Nuclear Security Administration, responsible for the planning and preparation of a capable exascale ecosystem, including software, applications, hardware, advanced system engineering and early testbed platforms, to support the nation’s exascale computing imperative. The material was supported by the U.S. Department of Energy, Office of Science, under contract DE-AC02-06CH11357. This work was also supported by the National Science Foundation under Grants CCF-1619253, OAC-2003709, OAC-2034169, and OAC-2042084. We acknowledge the Texas Advanced Computing Center for providing HPC resources that have contributed to the research results reported within this paper.

REFERENCES
