GE-SpMM: General-Purpose Sparse Matrix-Matrix Multiplication on GPUs for Graph Neural Networks

Guyue Huang, Guohao Dai, Yu Wang, Huazhong Yang
Department of Electronic Engineering, BNRist, Tsinghua University, Beijing, China
huang-gy16@mails.tsinghua.edu.cn, {daiguohao, yu-wang, yanghz}@mail.tsinghua.edu.cn

***Abstract*—The acceleration of Graph Neural Networks (GNNs) requires efficient and framework-compatible Sparse-Dense Matrix-Matrix Multiplication (SpMM). From the compatibility perspective, the sophisticated sparse matrix representations in state-of-the-art SpMM designs cause heavy preprocessing overhead for the framework. From the efficiency perspective, optimizations for SpMV (Sparse Matrix-Vector) do not apply well to SpMM, leading to redundant and uncoalesced global memory access.**

We propose GE-SpMM[1], which takes the CSR format consistent with GNN frameworks to enable integration without the format transformation overhead. We use Coalesced Row Caching to ensure coalesced access to both sparse and dense data in the global memory. We use Coarse-grained Warp Merging to reduce redundant data loading among GPU warps. Experiments on a real-world graph dataset demonstrate up to 1.41× speedup over Nvidia cuSPARSE [1] and up to 1.81× over GraphBLAST [2]. We embed GE-SpMM in GNN frameworks and get up to 3.67× speedup on popular GNN models like GCN [3] and GraphSAGE [4].

I. Introduction

Machine learning algorithms on graphs, especially the recently proposed Graph Neural Networks (GNNs), have been successfully applied to tasks such as link prediction and node classification [3]–[8]. Acceleration of GNN systems is crucial to solving real-world problems because of days of execution time (e.g., it takes up to 78 hours to train a GNN model with 7.5 billion edges on 16 GPUs [9]). In GNN algorithms, data are organized in the graph structure composed of vertices (nodes) and edges (links). Each vertex in the graph is associated with a feature vector. The feature vectors are propagated along edges and transformed by small DNN models in every GNN layer. Aggregating the feature vectors of neighbor vertices is a fundamental operation in GNNs, which can be expressed in Equation (1):

$$\overrightarrow{f_u} = reduce\_op(\{\overrightarrow{f_v}\}), \exists edge_{v\to u} \tag{1}$$

In Equation (1), $\overrightarrow{f_u}$ represents the feature vector of a vertex u in the graph, while $\{\overrightarrow{f_v}\}$ denotes the collection of feature vectors of u's in-edge neighbors. The $reduce\_op$ is the aggregation operation to generate a new feature vector of u, which varies in different GNN algorithms [3], [4]. The edges between vertices in real-world graphs are usually sparse. These edges can be represented with a sparse adjacency matrix. Fig. 1 shows that the aggregation operation on the graph can be treated as a customized multiplication between the sparse adjacency matrix (representing the graph topology) and the dense feature matrix (representing feature vectors of vertices). When the $reduce\_op$ in Equation (1) is taking sum, the operation between the adjacency matrix and the feature matrix is a standard Sparse-Dense Matrix-Matrix Multiplication (SpMM). When a customized $reduce\_op$ is adopted (e.g., max pooling), we call it a general SpMM-like operation.

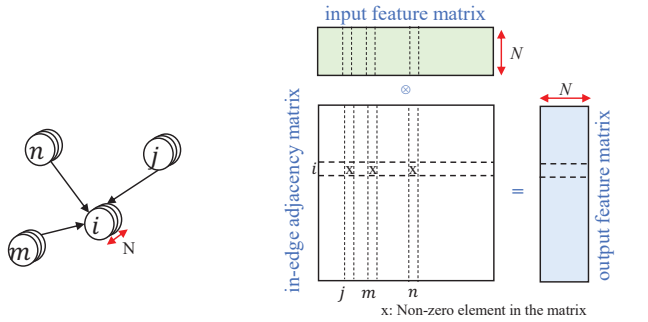

Fig. 1. SpMM operation from the graph- and matrix-perspective. Left: aggregating feature vectors (length=N) from Vertex j, m, n to Vertex i. Right: the SpMM operation between the in-edge adjacency matrix and the feature matrix, involving the j-th,m-th,n-th vectors in the feature matrix. The $\otimes$ operation represents the general-purpose aggregating operation.

Being a primary operation in GNN models, SpMM is a time-consuming step even on parallel hardware like GPUs. We profile the percentage of SpMM operations[2] during a GCN [3] training procedure. In GCN training, the forward and backward of graph convolution layers both involve SpMM. As listed in Table I, SpMM operations take $\sim 30\%$ of the total time in the example code provided by DGL [11] with default model settings. Dense matrix multiplications take $\sim 10\%$, and the rest of the operations all take less than 10%. Thus, accelerating SpMM operations can bring significant benefit to the overall training performance.

However, current SpMM acceleration solutions on GPUs still face challenges from 1) meeting GNN application requirements, and 2) full utilization of global memory bandwidth of the GPU hardware.

From the GNN application perspective, embedding SpMM designs in GNN frameworks has at least two requirements: support for general SpMM-like operators (rather than the standard SpMM), and no (or very low) data format conversion overhead in the whole procedure. In terms of general SpMM-

[1]The project is open-sourced at https://github.com/hgyhungry/ge-spmm

[2]Percentage of CUDA time, reported by PyTorch [10] profiler.

SC20, November 9-19, 2020, Is Everywhere We Are
978-1-7281-9998-6/20/$31.00 ©2020 IEEE

TABLE I
PERCENTAGE OF SPMM IN CUDA [12] TIME DURING GCN TRAINING ON GTX1080TI

Graph	SpMM percentage
Cora	33.1%
Citeseer	29.3%
Pubmed	29.8%

like operators, existing GNN frameworks cannot perform as well as the standard SpMM in cuSPARSE [1] library. For example, DGL [11] relies on Nvidia cuSPARSE to perform standard SpMM, but falls back to its own implementations for SpMM-like operations because they are not provided in cuSPARSE. Table II shows the comparison of the same aggregation step in two models: GraphSAGE-GCN [4], which internally calls SpMM, and GraphSAGE-pool [4], which internally calls SpMM-like. We again use example codes provided by DGL with default parameters. The results show that the current implementation of SpMM-like in DGL cannot compete with the performance of cuSPARSE. On the other hand, although recent studies on SpMM [13]–[15] in high-performance computing fields achieve better performance than cuSPARSE, they cannot be directly adopted by GNN frameworks. These implementations require preprocessing on the standard sparse matrix representation used by GNN frameworks, making them hard to be integrated. The extra time spent on preprocessing cannot be compensated by SpMM performance gain if SpMM is performed only a few times in GNNs, as is the case in GNN direct inference or batched training.

TABLE II
SPMM AND SPMM-LIKE COMPARISON IN DGL [11] ON GTX 1080TI

Graph	SpMM-like perf. loss against SpMM in GraphSAGE [4]
Cora	8.8%
Citeseer	89.2%
Pubmed	139.1%

From the GPU hardware perspective, SpMM exposes column-wise parallelism in the dense output matrix, which does not exist in the widely-studied Sparse Matrix-Vector Product (SpMV). A straightforward generalization by adding parallel threads along the column dimension can result in uncoalesced access patterns to the sparse matrix data, as shown in Fig. 2. Uncoalesced access pattern has been shown inefficient on GPUs [12], [16]. Thereby, SpMM kernel needs to be carefully designed to enable a coalesced access pattern for the data loading. On the other hand, reusing sparse matrix data is crucial for SpMM, while this issue does not rise for SpMV. In real applications, the column dimension of the dense matrix can be up to 512 [17], in which case the amount of memory transactions is substantial. Fig. 3 shows the profiling of SpMM kernel in cuSPARSE. We use as input a synthetic random sparse matrix of 65k rows and 650k non-zeros, to be detailed in Section V-B. We test a range of the dense matrix width (N in Table III) in the dense matrix. The two metrics are 1) the number of global load transactions (in the unit of 32bytes), and 2) the global load throughput (in the unit of GB/s), both reported by Nvidia nvprof. The GPU we use has a maximum global bandwidth of 484GB/s. Fig. 3 shows that the total number of memory transactions grows linearly with

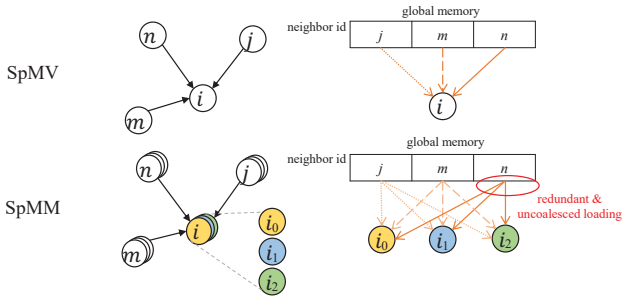

Fig. 2. Differences in data loading in SpMV and its straightforward generalization to SpMM. SpMM exposes more redundant data loading and uncoalesced access pattern.

N, but the bandwidth throughput saturates after N exceeds 32. We can observe that, unlike SpMV, which is typically bounded by low bandwidth utilization [18], SpMM can achieve high utilization, but suffers from too much data movement. Thereby, SpMM design requires a data-reuse mechanism to reduce redundant data transactions.

In this paper, we present GE-SpMM (an acronym for General-purpose SpMM), a CSR-based (Compressed Sparse Row) SpMM kernel on GPUs that tackles all these challenges. We summarize our contributions as follows:

- We present GE-SpMM, an efficient CSR-based SpMM-like kernel on GPUs, to accelerate GNN workloads. GE-SpMM can be integrated into existing GNN frameworks with no data conversion overhead for various GNN algorithms.
- We introduce the Coalesced Row Caching (CRC) method for SpMM, which uses GPU shared memory to cache sparse matrix rows. This method enables coalesced memory access to both sparse and dense matrix, leading to more efficient bandwidth utilization. The average improvement by adopting this method is up to 1.25×.
- We introduce the Coarse-grained Warp Merging (CWM) method for SpMM, which reuses the loaded sparse matrix by merging the workload of different warps. This technique reduces the number of memory transactions and improves instruction-level parallelism. The average speedup by adopting this method is up to 1.51×.
- We conduct extensive experiments on GE-SpMM on real-world network graphs [19] and GNN benchmarks [20]. GE-SpMM achieves up to 1.41× speedup over Nvidia cuSPARSE [1] and up to 1.81× over GraphBLAST [2]. We also embed GE-SpMM in GNN frameworks and get up to 3.67× CUDA time reduction on training popular

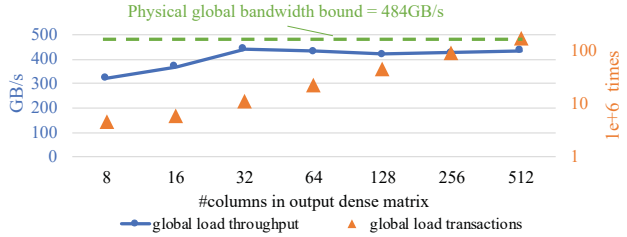

Fig. 3. Profiling of *csrmm2* in cuSPARSE. The loading throughput approaches upper bound when $N \geq 32$ but memory transactions keep growing linearly.

TABLE III
NOTATIONS.

Notation	Description
A	Sparse input matrix with dimension $M \times K$
B	Dense input matrix with dimension $K \times N$
C	Dense output matrix with dimension $M \times N$
M	Number of rows in A. Number of vertices in graph.
K	Number of columns in A, equal to M in graph problems.
N	Number of columns in B. Feature vector length.
nnz	Number of non-zero elements in A. Number of directed edges in graph.

GNN models like GCN [3] and GraphSAGE [4].

The rest of this paper is organized as follows. Background information is introduced in Section II. The designs and optimizations of GE-SpMM will be detailed in Section III, followed by the method to embed GE-SpMM in GNN frameworks. Our experimental setup and results will be presented in Section V. The paper is concluded in Section VI.

II. BACKGROUNDS AND RELATED WORKS

In this section, we introduce background information about both SpMM and GNNs on GPUs. The notations used in this paper are shown in Table III.

A. GPU Preliminaries

We use Nvidia GPUs with CUDA [12] as the programming interface in this paper. GPU is a highly parallel architecture composed of many Streaming Multiprocessors (SMs). An SM executes threads in a SIMT (Single Instruction Multiple Threads) fashion, and a group of 32 threads called a *warp* run simultaneously. The *warp* is transparent in CUDA programming model. Instead, users launch a number of parallel *blocks* and fill each *block* with a certain number of threads. In CUDA programming model, each *block* owns a *shared memory*, which is more efficient to access than the global memory (accessible to all blocks). *Shared memory* can be used for data reusing for different threads in order to reduce data transactions from the global memory.

Organizing threads into warps also have effects on the memory access pattern. GPU always tries to merge the memory request from a warp into as few transactions as possible. From the program perspective, it is recommended in technical materials [12], [16] to make a warp of threads access consecutive, aligned memory in one SIMT instruction. This technique is called *coalesced memory access*.

B. SpMM on GPUs

Because the dense matrix in the SpMM problem can be treated as a vector of vectors, a straight forward SpMM implementation is to perform SpMV for multiple times sequentially, as can be done in GunRock [21]. Obviously, This method does not exploit parallelism along the output column dimension. SpMV design in [18] uses a GPU warp to process a row in the sparse matrix, and previous SpMM design in GraphBLAST [2] inherits this method so that multiple threads can work on one output row in parallel. For the intra-warp data reuse, it uses a warp-level intrinsic (__shfl) to broadcast fetched data to other threads within the same warp. However, GraphBLAST fails to consider reusing the sparse row data among different warps and still has room for improvement. Nvidia cuSPARSE library [1] also provides a high-performance (not open-source) SpMM kernel (*csrmm2*), but general SpMM-like operations in GNN applications are not supported in cuSPARSE.

There are also other researches on high-performance SpMM kernels that perform better than cuSPARSE. Unlike GraphBLAST [2] which takes Compressed Sparse Row (CSR) format as input, these works require preprocessing on the input sparse matrix to form a new sparse representation dedicated for SpMM, such as ELLPACK-R in Fastspmm [22], and sparse formats used by RS-SpMM [13] and ASpT [14], [15]. But they are not practical for GNN frameworks to adopt. These non-standard formats lead to extra memory space and difficulties in software maintenance. Moreover, preprocess time can be up to $5\times$ actual SpMM computation time [13], [14]. Although this cost can be tolerated in iterative algorithms, GNN applications sometimes demand running SpMM only a few times for one matrix, and the overhead cannot be amortized with iterative execution. One example scenario is GNN inference, where trained models are directly used on new graphs to make predictions, such as predicting properties on new protein graphs [4]. Another is sampled batch training [4], [23], where the sampled subgraphs are different for each batch. For these applications, preprocessing cannot be amortized in GNN frameworks since the benefit cannot make up to overhead.

C. Graph Neural Network Frameworks

Many existing systems aim to provide high performance and easy programming abstractions for GNN algorithm developers. Projects like DGL [11] and Pytorch-Geometric (PyG) [24] provide graph APIs on top of deep learning frameworks (e.g., Pytorch [10]). Other systems in the industry [25], [26] and academia [17], [21] also provide programming interfaces and optimizations for GNNs.

As SpMM is a critical operation in many GNN models, DGL and PyG both implement custom SpMM kernel instead of using sparse matrix operators provided by Pytorch.

- DGL internally calls the function, *csrmm2* in cuSPARSE [1], to perform SpMM. However, for SpMM-like operations, cuSPARSE does not provide corresponding functions, so DGL falls back to its own kernels.
- Besides limited support for SpMM-like operations, *csrmm2* produces a column-major output. Since in GNNs both input and output of feature matrix need to be row-major, DGL calls a matrix transpose from cuBLAS [27] to transform the layout.
- PyG uses another abstraction called MessagePassing to represent graph propagation in GNN models. Message-passing first generates messages on all edges explicitly and then reduce them, while SpMM can fuse these two stages into one kernel. The consideration of MessagePassing is to allow more flexible user-defined operation, but

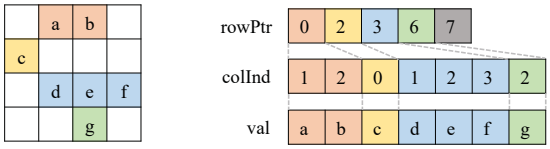

Fig. 4. The sparse matrix (left) and its CSR representation (right).

with generality, it loses the room for improving the performance of specific operations like SpMM.

Some aforementioned researches on fast SpMM design require preprocessing [13]–[15]. CuSPARSE [1] has limited support for SpMM-like operation [1]. Other SpMM designs inherited from SpMV fail to consider the column-wise parallelism or sparse matrix data reuse, which may lead to inefficiency when implemented on GPUs. These will further hinder the performance of both SpMM and GNNs.

III. GE-SpMM Design

In Section I, we observe from experiments that SpMM can be memory-bounded, so it is important to load data more efficiently and reduce the total number of memory transactions. In this section, we propose Coalesced Row Caching (CRC) and Coarse-grained Warp Merging (CWM) methods to achieve these goals.

A. Data Organization in GE-SpMM

In order to meet the requirement of compatibility to GNN frameworks with low preprocessing overheads, an universal data format for both SpMM and other sparse matrix operations is required. The Compress Sparse Row (CSR) format is a widely-used format in vendor libraries (e.g., Nvidia cuSPARSE [1]), data science toolkit (e.g., SciPy [28]), and GNN frameworks [11], [21]. As shown in Fig. 4, a sparse matrix is stored by three arrays using the CSR format: $rowPtr$, $colInd$, val. The column indices and values of non-zeros are first packed along the column and then stored in the order of their row indices. $rowPtr$ stores the offset of the first element of each row in $colInd$.

The data structure of CSR format determines the procedure of SpMM. The computation of each output $C[i, j]$, which is a dot-product of sparse row i of A and dense column j of B, begins with accessing $A.rowPtr$ for the offset of row i. Then the program needs to traverse a segment of $A.colInd$ and $A.val$ array to acquire the non-zeros in sparse row i. For each non-zero element, the program needs to use the column index got from $A.colInd$ to locate a specific row in dense matrix B. If the column index gives k, the program then loads $B[k, j]$, multiplies it with the sparse element value from $A.val$, and adds to the final result. When all non-zeros in the sparse row i is consumed, the final result of $C[i, j]$ is returned. Algorithm 1 shows this procedure with pseudo code.

Algorithm 1 A simple parallel CSR-based SpMM

```
Input: A.rowPtr[], A.colInd[], A.val[], B[]
Output: C[]
1:  for i = 0 to M − 1 in parallel do
2:    for j = 0 to N − 1 in parallel do
3:      result = 0
4:      for ptr = A.rowPtr[i] to A.rowPtr[i+1] do
5:        k = A.colInd[ptr]
6:        result += A.val[ptr] * B[k,j]
7:      end for
8:      C[i,j] = result
9:    end for
10: end for
```

Algorithm 2 SpMM with CRC

```
Input: A.rowPtr[], A.colInd[], A.val[], B[]
Output: C[]
1:  i = tb_id
2:  j = tid
3:  lane_id = tid % warp_size
4:  sm_base = tid - lane_id
5:  row_start = A.rowPtr[i]
6:  row_end = A.rowPtr[i+1]
7:  result = 0
8:  for ptr = row_start to row_end-1 step warp_size do
9:    /*load A.colInd and A.val with tile warp_size*/
10:   if ptr+lane_id<row_end then
11:     sm_k[tid]=A.colInd[ptr+lane_id]
12:     sm_v[tid]=A.val[ptr+lane_id]
13:   end if
14:   __syncwarp()
15:   /*consume the loaded elements*/
16:   for kk = 0 to warp_size do
17:     if ptr+kk<row_end then
18:       k = sm_k[sm_base+kk]
19:       result += sm_v[sm_base+kk] * B[k,j]
20:     end if
21:   end for
22: end for
23: C[i,j] = result
```

B. Coalesced Row Caching

When trying to map Algorithm 1 on parallel architectures like GPUs, a simple way is to parallelize the for-loop at Line 1 and 2 since there exists no dependency among each iteration of the loop. The for-loop at line 4 in Algorithm 1 has variable loop bound decided at runtime and involves adding to one same variable, so it cannot be parallelized. As introduced in section II-A, coalesced memory access can improve bandwidth efficiency, and it requires a warp of threads to access consecutive elements in one instruction.

In Algorithm 1, it is easy to ensure coalesced access to the dense matrix B (the instruction at line 9). When we parallelize loop at Line 1 and 2 among threads, we need to ensure that threads within a warp have the same i and contiguous j. However, the current algorithm does not present coalesced memory access to any part of the sparse matrix. Since threads within a warp share the same i, we cannot enable coalesced

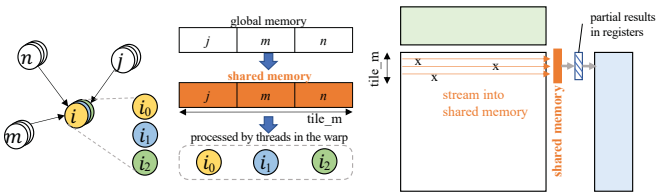

Fig. 5. Overview of CRC which enables coalesced access to sparse rows. Non-zero elements of sparse matrix are first loaded into shared memory in a coalesced way and consumed later. Partial-sums are stored in local variables and written to global memory at the end. Note that no two threads write to the same output, so no atomic operation is needed.

access to $A.rowPtr$. As to $A.colInd$ and $A.val$, in the current algorithm, the sequential execution of the for-loop at line 4 forces threads in the same warp to access the same address (instructions at line 5-6), leading to the broadcast-like pattern in Fig 2. Compared with ideal coalesced access to sparse rows (referring to a segment of $A.colInd$ and $A.val$), the current algorithm is not making full use of data in one memory transaction and issues too many transactions.

Our solution is to partially unroll this sequential for-loop, by a factor of $warp\_size$ (number of threads within a warp, 32 on Nvidia GPUs). During each iteration, a warp of threads first loads a tile of the sparse row i into GPU shared memory. The size of a tile is the same as $warp\_size$, meaning that each thread loads a different element. For now, we assume that the total number of non-zeros in row i is multiples of $warp\_size$, but we will generalize to arbitrary sparse row length later. After loading a tile to shared memory, threads enter an inner for-loop and compute on the loaded data one-by-one, but this time the sparse row data are loaded from shared memory instead of global memory.

A pseudo-code of our method is listed in Algorithm 2. Basically, our method uses a two-phase strategy to load and compute on sparse rows, as shown in Fig. 5. In the first phase, a warp loads a tile of a sparse row into the shared memory to enable coalesced loading of a sparse row. In the second phase, a warp sequentially consumes the previously loaded elements. When the number of non-zeros in the sparse row exceeds a tile, the two-phase procedure continues until all non-zeros are consumed. It is not hard to deal with arbitrary row length. Each thread has a copy of the row length, and always checks if it has exceeded the bound before loading from a sparse row in the first phase (if-condition at line 10). In the second phase, since elements stored in shared memory are consumed sequentially, we check with loop bound in every iteration (condition at Line 17).

The improvement of this algorithm over Algorithm 1 lies in more efficient global memory loading of sparse rows. Ideally, the total number of load requests to these two arrays can be reduced by a factor of $warp\_size$, because a tile of the sparse row is loaded in one coalesced request. In reality, sparse rows are often not perfectly aligned, leading to more than one transaction for a tile. Moreover, many sparse rows have fewer non-zeros than a tile, and the number of load transactions in the Algorithm 1 equals to the number of non-zeros. Thus the reduction in load transactions for these short

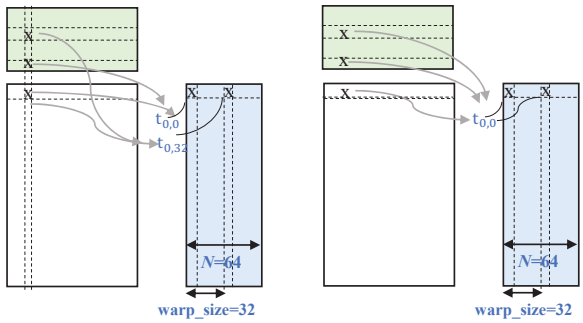

Fig. 6. An example of CWM with $N = 64$ and $CF = 2$. Left: CWM is not applied, and the element in the sparse matrix is loaded twice by two threads. Right: the workloads of the previous two threads are merged, so the element in the sparse matrix is loaded only once.

rows is strictly less than $warp\_size$. Despite all these factors, the improved algorithm still achieves a significant reduction of load transactions and improves the efficiency of global bandwidth, as shown in Section V-B.

C. Coarse-grained Warp Merging

With CRC, we can convert the uncoalesced access to the sparse matrix into a more efficient, coalesced way. From the data re-use perspective, the benefit of CRC is to share loaded sparse matrix via GPU shared memory. Although each non-zero element in the sparse matrix is needed to compute an entire row in the output, our CRC method only makes loaded elements shared by threads within the same warp. The consideration behind this is to reduce synchronization overhead. To safely use the shared memory, synchronization needs to be called to avoid read-write races (between line 12 and line 18 in Algorithm 2). If we allow different warps to use the same data in shared memory, we need to insert synchronization in the entire thread block, which brings significant overhead. Since CUDA provides warp-level fine-grained synchronization, which is much less expensive then block-level synchronization, we only make loaded data to be shared within a warp. The drawback is that different warps still perform redundant loading of sparse rows.

Our technique to address redundant data loading is Coarse-grained Warp Merging (CWM). It is similar to thread coarsening in dense matrix multiplication, which can improve bandwidth throughput with instruction-level parallelism (ILP) [29], and reduce the number of memory transactions [30]. The basic idea is to merge the workload of different warps that has redundant data loading. In SpMM, merging workloads means to make each thread produce more than one output. We illustrate CWM by an example in Fig. 6. In Fig. 6, the suffix of t (short for thread) indicates both the indices to this thread and the indices of output it produces. Observe that $thread_{0,0}$ and $thread_{0,32}$ both load data from row 0 of A, but they belong to different warps and cannot share data under Algorithm 2. To dismiss this redundant load, we merge these two threads' workloads, making $thread_{0,0}$ compute both $C[0,0]$ and $C[0,32]$. $thread_{0,0}$ will have two partial sum variables locally, and with every non-zero element in A, it will load two values from matrix B and update two partial

Algorithm 3 SpMM with CRC and CWM(CF=2)

Input: A.rowPtr[], A.colInd[], A.val[], B[]
Output: C[]

```
1:  /* initialization (line 1-6 of Algorithm 2) */
2:  result_1 = 0, result_2=0
3:  for ptr = row_start to row_end-1 step warp_size do
4:    /* load A.colInd, A.val (line 10-14 of Algorithm 2) */
5:    for kk = 0 to warp_size do
6:      k = sm_k[sm_base+kk]
7:      result_1 += sm_v[sm_base+kk] * B[k,j]
8:      result_2 += sm_v[sm_base+kk] * B[k,j+warp_size]
9:    end for
10: end for
11: C[i,j] = result_1
12: C[i,j+warp_size] = result_2
```

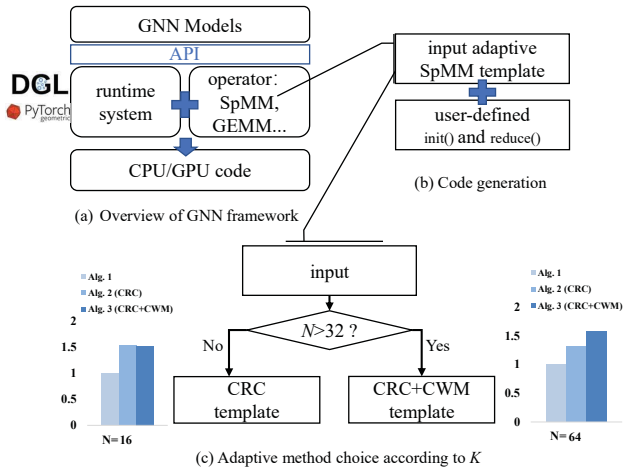

Fig. 7. The overall flow of embedding GE-SpMM in GNN frameworks.

results. In Algorithm 3 we give a pseudo code with CWM adopted.

In Fig. 6, we merge the workloads of two warps and cut the number of threads by half. Intuitively this process can continue, and we can cut down more threads. We call the factor of thread number reduction *coarsening factor* (CF). For example, CF is 2 in Fig. 6, which means each thread is assigned to produce 2 output values. In general, the load transactions of sparse rows can be reduced by CF through this technique. Another benefit of introducing thread coarsening is to improve bandwidth utility via ILP [29]. Line 7-8 in Algorithm 3 is independent memory loading instructions, and GPU architecture can serve these two requests simultaneously, potentially increasing the usage of global bandwidth. Increasing CF can further reduce memory load, but there will be fewer threads on the fly. Large CF also causes each thread to hold more local variables for partial results, and this increment in resource usage may hurt performance. GPUs use massive parallel threads to hide all types of stalls, mostly the latency of memory load. When using CWM, it is significant to balance between data re-use and parallelism. Analytical models for choosing CF could be difficult due to the entangled effects of hardware parameters and sparse matrix properties. We turn to an empirical method and experimented on our dataset of real-world graphs with $N = 512$ to find a general best choice of CF, detailed in Section V-B.

IV. Accelerate GNN Frameworks with GE-SpMM

GE-SpMM is developed to accelerate GNN applications. The CSR format and the support for SpMM-like operations make it easy to be embedded in existing frameworks. This section briefly discusses how we use GE-SpMM to enhance the performance of existing GNN frameworks.

A. GE-SpMM for Different Matrix Sizes

To accelerate real applications, we make a few enhancements to support arbitrary input size (N). CRC and CWM apply well to problems with large N, where the kernel needs to load a large amount of data from the dense matrix and is bottlenecked by bandwidth efficiency. Fig. 7(c) shows the overall benefit of our two techniques when $N = 16$ and $N = 64$, with average performance on the test dataset (detailed in Section V-A) normalized to Algorithm 1. When $N > 32$, we apply both CRC and CWM in the kernel. CWM is not necessary for $N \leq 32$ since $warp\_size$ is 32, and we should directly call Algorithm 2 to dismiss the overhead of unnecessary instructions.

To address the need for SpMM-like operation in GNN models, we modify the basic GE-SpMM to allow user-defined operations. To define an SpMM-like operation, the user needs to provide an initialization function, and a reduce function, both will be inlined at compile time. The parallel execution requires the reduction function to be associative and commutative, but common operations like taking sum or maximum are naturally valid.

B. GNN Acceleration Based on GE-SPMM

Current GNN frameworks like DGL [11] and PyG [24] are often based on other deep learning (DL) frameworks (e.g., PyTorch [10]), but add new APIs for graph operations. Although it is possible to express graph operations with sparse tensor operations provided by DL frameworks, sparse tensor operations' performance is not satisfactory. Hence, DGL implements all graph-related operations in C++/CUDA and exposes to DL frameworks as shared lib. We also follow this method to accelerate GNN application with a high-performance CUDA kernel.

To be more specific, we wrap our kernel inside a custom autograd function, which is an atomic operator with a definition of gradient function in PyTorch. This function represents an aggregation step on the graph and can be used to build GNN layers and modules. Since DGL already implements SpMM-like in CUDA, we substitute their kernel with ours and rebuild the project. PyG is another popular Python library for GNN built on PyTorch. It abstracts GNN as MessagePassing procedure and implements MessagePassing as a versatile module that allows user-defined message and aggregation functions. MessagePassing is a more general interface than SpMM-like, so we cannot use SpMM-like to replace MessagePassing. We

instead implement an SpMM-like operator and replace the MessagePassing function calls in training code with ours. An overview of how we integrate GE-SpMM to GNN frameworks is shown in Fig. 7.

V. Experiment Evaluations

We conduct extensive experiments on GE-SpMM. Results listed in this section evaluate technique effectiveness, kernel performance and application speedup.

A. Experiment Setup

1) Graph Benchmarks: In order to test the proposed GE-SpMM for GNN workloads, we run experiments on three graphs [20] used for node-classification tasks in many GNN models [3], [4], `Cora`, `Citeseer`, and `Pubmed`, whose properties are listed in Table IV. We also test the performance on SNAP dataset collected in SuiteSparse Matrix Collection [31], a sparse matrix benchmark. SNAP group in SuiteSparse contains 66 valid graphs from various domains. The original SNAP dataset maintained in [19] also contains metadata of some graphs which are not collected in SuiteSparse, but we limit our test to graphs in SuiteSparse to save the effort of converting metadata to standard input. We omit two graphs (`FriendSter` and `Twitter`) due to out-of-memory. Some items in SuiteSparse contains more than one matrices. We only run tests on one matrix in each item[3]. This set of 64 sparse matrices has M ranging from 1005 to 4847571, and nnz/row ranging from 1.58 to 32.53. [17] experiments on models with feature size up to 512 and presents the best model accuracy of feature size around 256 with deeply-stacked layers. This should provide an intuition of how large N is in real applications. All average results are geometric means.

TABLE IV
Graphs Used in GNN for Classification [20]

Graph	# Vertices	# Edges	# Classes
`Cora`	2708	5429	7
`Citeseer`	3327	4732	6
`Pubmed`	19717	44338	3

2) SpMM Baselines: We compare our GE-SpMM with the following baselines in our experiments.

- **SpMM kernel by vendor:** *csrmm2*. It is a function in cuSPARSE [1] for SpMM. cuSPARSE has two functions for multiplication of sparse and dense matrices. The *csrmm2* assumes a row-major input dense matrix, while the other one, *csrmm*, assumes a column-major input dense matrix. The *csrmm2* consistently outperforms *csrmm*, and here we show a comparison to *csrmm2*. Note that, the output dense matrix of *csrmm2* is column-major, which is a convention in many Nvidia libraries. As explained in Section II-C, GNN applications require row-major output, so existing solution is forced to perform matrix transpose upon *csrmm2* output. We do not add this to baseline when comparing kernel performance. However, this overhead of cuSPARSE cannot be ignored in real applications.
- **Open-source SpMM kernel:** *rowsplit* in GraphBLAST [2]. It is a most-recent CSR-based SpMM implementation in literature.
- **Graph processing engine on GPUs:** GunRock [21]. Because the SpMM can be executed from a graph perspective by assigning each vertex with a feature vector, we also compare our GE-SpMM with the state-of-the-art graph processing system on GPUs.

3) Environments: We conduct experiments on the following two machines:

- Machine 1. GPU: Nvidia GTX 1080Ti, Compute Capability 6.1 (28 Pascal SMs at 1.481 GHz, 11 GB GDDR5X with 484 GB/s bandwidth). Host CPU: Intel(R) Xeon(R) CPU E5-2643 v4 (24 cores).
- Machine 2. GPU: Nvidia RTX 2080, Compute Capability 7.5 (46 Turing SMs at 1.515 GHz, 8 GB GDDR6 with 448 GB/s bandwidth). Host CPU: Intel(R) Core(TM) i7-9700K (8 cores).

In kernel performance tests, all codes are compiled using NVCC (CUDA compiler provided by Nvidia) in CUDA 10.1 with `-O3` flag. Execution time is measured from the average of 200 kernel runs. Throughput is calculated from theoretical float-operation ($2 * nnz * N$) over measured execution time. For application speedup, all reported items (kernel and model time) refer to CUDA time reported in PyTorch profiler.

B. Benefits of GE-SpMM Design

1) Benefits of Coalesced Row Caching: The aim of CRC is to enable coalesced access to the sparse matrix and improve bandwidth efficiency. To evaluate the effectiveness of CRC, we use Nvidia's nvprof to profile two metrics [12]: gld_transactions (GLT), the number of global load transactions; gld_efficiency (GLT_effi), the ratio of requested global memory load throughput to required global memory load throughput. One issued memory transaction loads a fixed size of data, but the program may only require a part of it. This metric can reflect how efficiently the program uses global bandwidth.

TABLE V
Effects of CRC

Matrix	Method	GLT(×32bytes)	GLT_effi
M=16K	w/o CRC	1.34e+8	68.95%
nnz=160K	w/ CRC	0.55e+8	92.40%
M=65K	w/o CRC	5.36e+8	68.95%
nnz=650K	w/ CRC	2.18e+8	92.40%
M=262K	w/o CRC	21.47e+8	68.95%
nnz=2.6M	w/ CRC	8.73e+8	92.39%

The tests run on three synthetic random graphs[4] with $N = 512$. We only have profiling results on Machine 1 since nvprof in CUDA 10.1 does not support GPU over 7.2 capability [12]. The results in Table V show that using CRC can significantly

[3]The matrix, which has the same filename as this item is considered the default one.

[4]The code to generate a random graph is from the repository of Ligra [32].

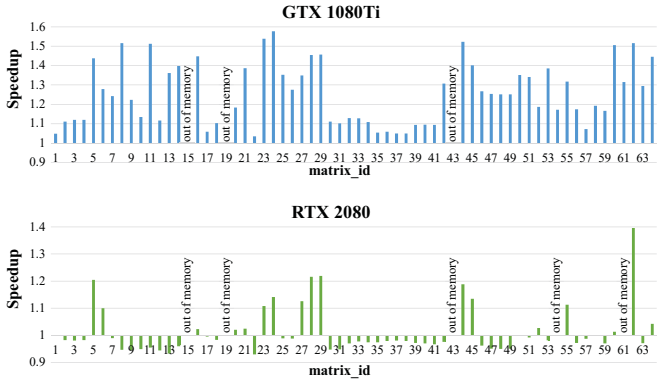

Fig. 8. Relative speedup using Coalesced Row Caching.

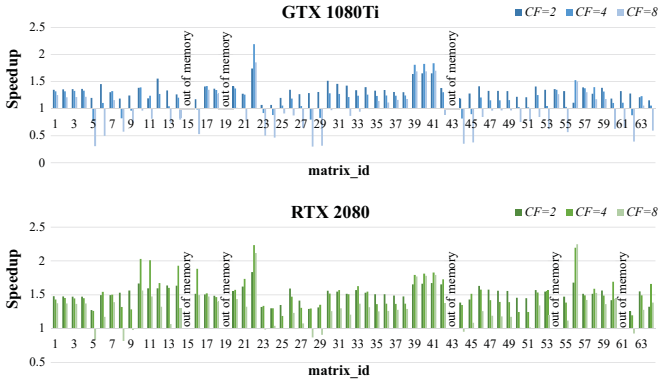

Fig. 9. Relative speedup using Coarse-grained Warp Merging when CF varies.

reduce the total number of load transactions and improve the memory load efficiency due to coalesced memory access.

Fig. 8 shows the relative improvement of applying CRC (Algorithm 2 against Algorithm 1). On GTX 1080Ti, CRC brings an average of 1.246× performance gain. On RTX 2080, simply applying CRC does not bring significant performance gain (average of 1.011×). However, CRC is the foundation of our second technique CWM. One benefit of CWM is a high throughput of loading dense matrix with ILP, but without CRC, the bandwidth is used on a large amount of inefficient access to the sparse matrix. The excessive transactions of the sparse matrix data cause cache throttling of loading the dense matrix. In the next part, we will show that combined with CWM, GE-SpMM still achieve significant improvement over simple SpMM on RTX 2080.

2) Benefits of Coarse-grained Warp Merging: CWM is introduced in order to reuse loaded data and reduce global transactions. However, it reduces the total number of warps and may harm parallelism. We also use nvprof to test the effects brought by CWM. We profile three metrics: gld_transactions (GLT), as previously mentioned it indicates the amount of data loading; gld_throughput, the throughput of the global load; achieved_occupancy (Occ), the ratio of the average active warps to the maximum supported on a multiprocessor, which can be a reflection of achieved parallelism.

TABLE VI
EFFECTS OF CWM

Method	GLT(×32bytes)	gld_throughput(GB/s)	Occ
w/o CWM	2.18e+8	479.54	0.78
CWM (CF=2)	1.93e+8	567.82	0.78
CWM (CF=4)	1.80e+8	479.23	0.75
CWM (CF=8)	1.74e+8	395.22	0.75

The results in Table VI are tested on one of the random graphs ($M = 65K, nnz = 650K, N = 512$). It shows that CWM can reduce global data loading. When the coarsening factor (CF) gets larger, gld_transactions keep decreasing, but the benefit becomes smaller because loading dense matrix takes up most of the transactions. When increasing CF, the occupancy also decreases, indicating parallelism loss. The combination of CRC and CWM brings an average of 1.65× and 1.53× speedup on GTX 1080Ti and RTX 2080 respectively over non-optimized version (Algorithm 1).

In practice, tuning for optimal CF requires a balance between data reuse and parallelism, which are related to properties of the input matrix. In Fig. 9, we plot the relative performance when taking different CF. Each bar represents a test on one graph in SNAP dataset. The relative performance means speedup over not using CWM. It can be observed that $CF = 2$ works well for most matrices, while $CF > 4$ shows obvious performance drop. For rare cases (4 and 1 out of 64 on two GPUs), choosing $CF = 2$ causes over 15% performance loss compared to optimal CF. Since our goal is to provide a runtime SpMM kernel, we avoid any parameter tuning and use $CF = 2$ because it provides the best overall effect. The kernel performance reported in the following experiments is all the results of $CF = 2$. Moreover, our previous tests show that on RTX 2080, CRC cannot bring large performance gain, but a combination of CRC and CWM can bring significant improvement (an average of 1.51×) with $CF = 2$.

C. Overall Performance of SpMM Kernel

In this part, we present overall performance compared to kernels in cuSPARSE [1] and GraphBLAST [2]. All matrices in tests are single-precision, as is the case in GNN models. For kernel performance, we test N from 128 up to 512.

1) Graphs for GNNs: We test GE-SpMM on three graphs used in GNN models, and results are shown in Fig. 10.

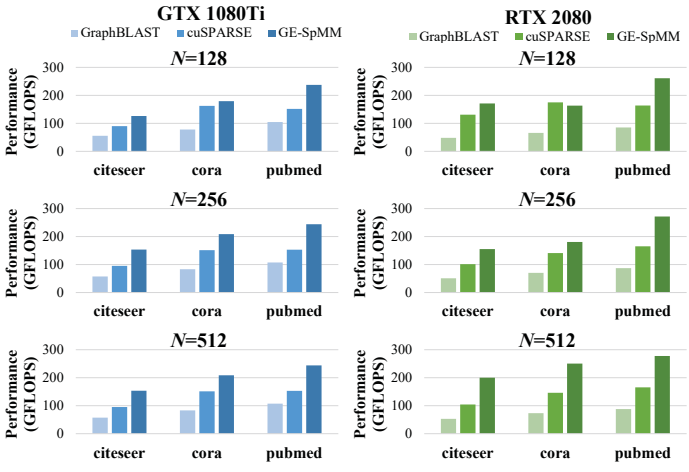

Fig. 10. Performance on GNN graphs.

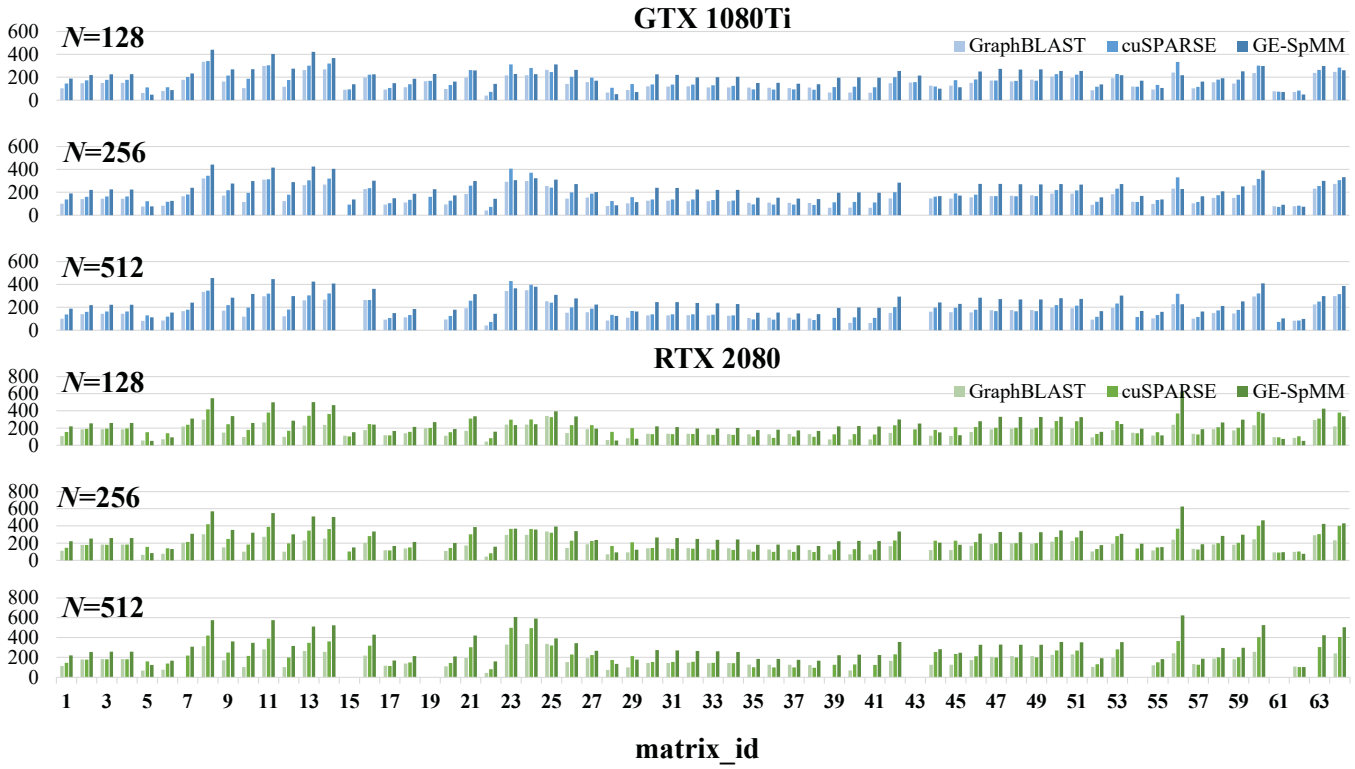

Fig. 11. Overall performance on 64 graphs in SNAP dataset. The omitted bars are due to out-of-memory. The matrix_id corresponds to the alphabetical order of matrix name.

Benefit from our two techniques, GE-SpMM can outperform cuSPARSE by at most 1.62×. Tests on these three graphs show that GE-SpMM can accelerate real GNN models if applied in GNN frameworks.

2) Graphs from SNAP: The performance comparison of GE-SpMM and baselines on SNAP dataset is shown in Fig. 11. To summarize, the average performance gain is listed in Table VII, GE-SpMM achieves up to 1.43× and 1.81× speedup compared with cuSPARSE [1] and GraphBLAST [2]. Although the arbitrary performance is related to graph characteristics, Fig. 11 can demonstrate that for any specific graph, our kernel becomes more competitive over baseline when N gets larger. In other words, our techniques are crucial for applications with large N.

TABLE VII
GE-SPMM AVERAGE IMPROVEMENT ON SNAP DATASET

Machine	Baseline	N=128	N=256	N=512
GTX 1080Ti	cuSPARSE [1]	1.18	1.30	1.37
	GraphBLAST [2]	1.42	1.44	1.61
RTX 2080	cuSPARSE [1]	1.20	1.34	1.43
	GraphBLAST [2]	1.57	1.73	1.81

D. Comparison with Graph Engines on GPUs

It is also possible to build SpMM with graph engines from a graph processing perspective. GunRock [21] provides many APIs and built-in kernels that allow users to write graph algorithms. We use GunRock's API `advance` to write an SpMM program. However, GunRock does not provide any options to parallelize along feature dimension, because

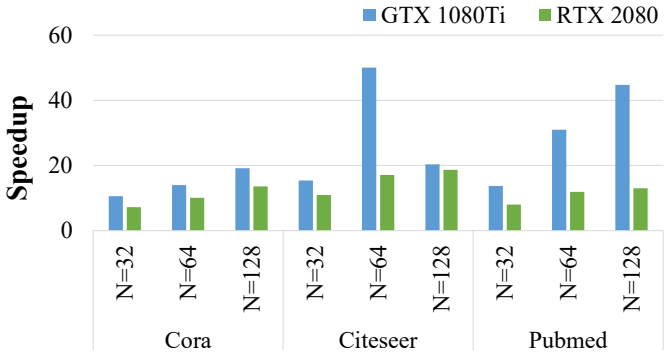

Fig. 12. Speedup of GE-SpMM over GunRock-based SpMM.

in traditional graph algorithms like PageRank, the feature of a vertex is an undividable scalar. GunRock fails to provide feature-dimension parallelism, which significantly harms SpMM performance. Fig. 12 shows the kernel execution time of GE-SpMM compared to the program written with GunRock. GE-SpMM outperforms GunRock-based implementation by 18.27× on average for all test cases, indicating that SpMM and GNN workloads require new primitives in graph processing frameworks rather than SpMV.

E. Comparison with preprocess-based approaches

Some previous researches propose specialized sparse formats for SpMM problem that exploit regular memory access [22] or data reuse [13], [14]. As far as we know, ASpT [14] is the best SpMM implementation publicly available. ASpT [14] exploits reusing dense matrix data with tiling. It requires a special sparse format composed of CSR and

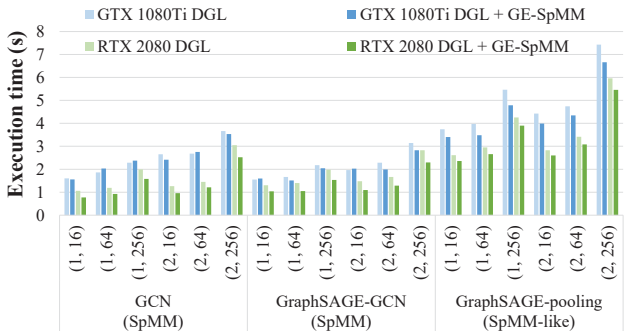

Fig. 13. Accelerating GNNs using GE-SpMM in DGL [11], (x, y) represents the number of layers (x) and the length of features (y) in a GNN model, which is the input parameter for GNNs.

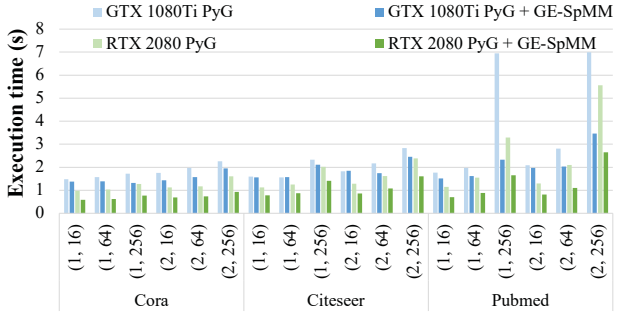

Fig. 14. Accelerating GNNs using GE-SpMM in PyG [24], (x, y) represents the number of layers (x) and the length of features (y) in a GNN model, which is the input parameter for GNNs.

additive arrays to mark the positions of locally-dense blocks explicitly. Our optimizations in this paper to reuse sparse matrix data is orthogonal to their techniques.

We test their source code on our machines on the SNAP dataset and results are listed in VIII. Our GE-SpMM achieves average of 0.93X, 0.97X, 1.00X for N=128, 256, 512 on GTX1080Ti GPU, and average 0.85X, 0.93X, 0.98X on RTX2080, against ASpT (without preprocess time).

The preprocess overhead varies significantly on different sparse matrices, from 0.01× to 64.53× of actual SpMM execution time, and average overhead is 0.47× execution time on GTX1080Ti and 0.34× on RTX2080. With preprocess time added (one preprocess + one run), our kernel is average 1.43×∼2.06× against ASpT.

TABLE VIII
GE-SPMM AVERAGE SPEED AGAINST ASPT [14]

Machine	Baseline	N=128	N=256	N=512
GTX 1080Ti	ASpT [1]	0.93	0.97	1.00
	ASpT w/ preproc [2]	1.88	1.97	2.06
RTX 2080	ASpT [1]	0.85	0.93	0.98
	ASpT w/ preproc [2]	1.43	1.57	1.69

F. End-to-End Performance for GNNs

Our kernel is developed to accelerate GNN applications, and the CSR-based property makes it easy to be embedded in existing frameworks. As introduced in Section IV, we embed GE-SpMM in DGL [11] and PyG [24]. We test application speedup on GCN, GraphSAGE, and other popular GNN models that involve SpMM or SpMM-like operations.

1) GNNs based on SpMM Operators: **Comparison with DGL.** SpMM is a very common operator in GNN, and GE-SpMM can benefit many GNN models. We test the benefit of integrating GE-SpMM in DGL with two models: GCN [3] and GraphSAGE-gcn [4]. For both models, we use example codes provided by DGL, and run tests on different model settings (number of layers and length of feature vectors in each layer). We use PyTorch profiler to monitor the entire training process and record total CUDA time from the report. The results are shown in Fig. 13. GE-SpMM brings speedup in most of these applications. However, in 4 tests on GTX 1080Ti, GE-SpMM does not bring acceleration over original DGL. This is because the feature-length of the last layer in GNN models usually equals the number of classes in the classification problem. Therefore, the output layer involves SpMM with a small N, when GE-SpMM is not very competitive.

Comparison with PyG. PyG [24] provides example code for GCN. We repeat running GCN on different graphs with different model settings on PyG and report the total reduction of CUDA time in Fig. 14. Improvements on PyG is more obvious than on DGL. This is because the general MessagePassing API in PyG explicitly generates 'message' on all edges and then performs reduction, while DGL fuses the two phases into one SpMM kernel. To summarize, integrating GE-SpMM brings up to 3.67× and 2.10× CUDA time reduction on two GPUs.

2) GNNs based on SpMM-like Operators: GE-SpMM is intended to accelerate a class of SpMM-like operators that is not yet supported by cuSPARSE. One good example is the pooling operation in GraphSAGE-pool [4], where each vertex aggregates the features of its neighbors by taking maximum. cuSPARSE does not provide this operation. Future GNN models may also use customized reduction functions for pooling, and relying on cuSPARSE is not flexible to these user-defined operations. It is not hard to generalize GE-SpMM to support SpMM-like operations since these operations follow similar memory access patterns, and the code is almost the same. We implement a GE-SpMM kernel for the aggregation step in GraphSAGE-pool. Our kernel brings up to 1.14× acceleration on the CUDA time of GraphSAGE-pool training on Pubmed graph. Note that for SpMM-like operation only, GE-SpMM can achieve 2.39× to 6.15× speedup over DGL's SpMM-like kernel. This shows the value of this work in supporting flexible, user-defined SpMM-like operation in new GNN models, which are not covered by the vendor library, and achieve even better performance.

TABLE IX
SUMMARY OF GRAPHSAGE-POOL CUDA TIME REDUCTION ON DGL.

	GTX 1080Ti		RTX 2080	
(#layer, #feature)	SpMM-like	Total	SpMM-like	total
(1,16)	2.89	1.10	3.24	1.11
(1,64)	3.92	1.14	3.44	1.11
(1,256)	4.04	1.14	3.46	1.09
(2,16)	2.39	1.11	3.03	1.09
(2,64)	3.09	1.09	3.37	1.11
(2,256)	6.15	1.12	3.51	1.09

VI. Conclusion

In this paper, we propose an efficient CSR-based SpMM design, GE-SpMM, for Graph Neural Network applications on GPUs. GE-SpMM considers GNN applications' requirements, including compatibility in the sparse matrix format and the need for SpMM-like operations. GE-SpMM introduces two techniques: Coalesced Row Caching and Coarse-grained Warp Merging, to improve the efficiency of global data access, leading to 1.25× and 1.51× speedup, respectively. By adopting these optimizations, GE-SpMM achieves up to 1.41× speedup over Nvidia cuSPARSE [1] and up to 1.81× over GraphBLAST [2]. We also embed GE-SpMM in GNN frameworks (e.g., DGL [11], PyG [24]) and achieve significant training time reduction.

Acknowledgement

This work was supported by National Key Research and Development Program of China (No. 2018YFB0105000, 2017YFA0207600), National Natural Science Foundation of China (No. U19B2019, 61832007, 61622403, 61621091), China Postdoctoral Science Foundation (No. 2019M660641), Beijing National Research Center for Information Science and Technology (BNRist), and Beijing Innovation Center for Future Chips.

References

[1] M. Naumov, L. Chien, P. Vandermersch, and U. Kapasi, "Cusparse library," in *GPU Technology Conference (GTC)*, 2010.

[2] C. Yang, A. Buluç, and J. D. Owens, "Design principles for sparse matrix multiplication on the gpu," in *European Conference on Parallel Processing*. Springer, 2018, pp. 672–687.

[3] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[4] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in neural information processing systems*, 2017, pp. 1024–1034.

[5] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European Semantic Web Conference*. Springer, 2018, pp. 593–607.

[6] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *arXiv preprint arXiv:1810.00826*, 2018.

[7] A. Fout, J. Byrd, B. Shariat, and A. Ben-Hur, "Protein interface prediction using graph convolutional networks," in *Advances in neural information processing systems (NeurIPS)*, 2017, pp. 6530–6539.

[8] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, p. early access, 2020.

[9] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," in *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2018, pp. 974–983.

[10] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8024–8035.

[11] M. Wang, L. Yu, D. Zheng, Q. Gan, Y. Gai, Z. Ye, M. Li, J. Zhou, Q. Huang, C. Ma, Z. Huang, Q. Guo, H. Zhang, H. Lin, J. Zhao, J. Li, A. J. Smola, and Z. Zhang, "Deep graph library: Towards efficient and scalable deep learning on graphs," *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. [Online]. Available: https://arxiv.org/abs/1909.01315

[12] "Nvidia cuda toolkit," =https://docs.nvidia.com/cuda/index.html.

[13] C. Hong, A. Sukumaran-Rajam, B. Bandyopadhyay, J. Kim, S. E. Kurt, I. Nisa, S. Sabhlok, Ü. V. Çatalyürek, S. Parthasarathy, and P. Sadayappan, "Efficient sparse-matrix multi-vector product on gpus," in *International Symposium on High-Performance Parallel and Distributed Computing (HPDC)*, 2018, pp. 66–79.

[14] C. Hong, A. Sukumaran-Rajam, I. Nisa, K. Singh, and P. Sadayappan, "Adaptive sparse tiling for sparse matrix multiplication," in *Symposium on Principles and Practice of Parallel Programming (PPoPP)*, 2019, pp. 300–314.

[15] P. Jiang, C. Hong, and G. Agrawal, "A novel data transformation and execution strategy for accelerating sparse matrix multiplication on gpus," in *Proceedings of the 25th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, 2020, pp. 376–388.

[16] D. B. Kirk and W. H. Wen-Mei, *Programming massively parallel processors: a hands-on approach*. Morgan kaufmann, 2016.

[17] Z. Jia, S. Lin, M. Gao, M. Zaharia, and A. Aiken, "Improving the accuracy, scalability, and performance of graph neural networks with roc."

[18] N. Bell and M. Garland, "Implementing sparse matrix-vector multiplication on throughput-oriented processors," in *International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2009, pp. 1–11.

[19] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," http://snap.stanford.edu/data, Jun. 2014.

[20] Z. Yang, W. W. Cohen, and R. Salakhutdinov, "Revisiting semi-supervised learning with graph embeddings," in *International Conference on Machine Learning (ICML)*, 2016, pp. 40–48.

[21] Y. Wang, A. Davidson, Y. Pan, Y. Wu, A. Riffel, and J. D. Owens, "Gunrock: A high-performance graph processing library on the gpu," in *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*, ser. PPoPP '16, 2016.

[22] G. Ortega, F. Vázquez, I. García, and E. M. Garzón, "Fastspmm: An efficient library for sparse matrix matrix product on gpus," *The Computer Journal*, vol. 57, no. 7, pp. 968–979, 2014.

[23] J. Chen, T. Ma, and C. Xiao, "Fastgcn: fast learning with graph convolutional networks via importance sampling," *arXiv preprint arXiv:1801.10247*, 2018.

[24] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[25] R. Zhu, K. Zhao, H. Yang, W. Lin, C. Zhou, B. Ai, Y. Li, and J. Zhou, "Aligraph: a comprehensive graph neural network platform," *Proceedings of the VLDB Endowment (VLDB)*, vol. 12, no. 12, pp. 2094–2105, 2019.

[26] L. Ma, Z. Yang, Y. Miao, J. Xue, M. Wu, L. Zhou, and Y. Dai, "Neugraph: parallel deep neural network computation on large graphs," in *USENIX Annual Technical Conference (ATC)*, 2019, pp. 443–458.

[27] C. Nvidia, "Cublas library programming guide," *NVIDIA Corporation. edit*, vol. 1, 2007.

[28] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–, [Online; accessed]. [Online]. Available: http://www.scipy.org/

[29] V. Volkov and J. W. Demmel, "Benchmarking gpus to tune dense linear algebra," in *International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*. IEEE, 2008, pp. 1–11.

[30] Y. Yang, P. Xiang, J. Kong, M. Mantor, and H. Zhou, "A unified optimizing compiler framework for different gpgpu architectures," *ACM Transactions on Architecture and Code Optimization (TACO)*, vol. 9, no. 2, pp. 1–33, 2012.

[31] T. A. Davis and Y. Hu, "The university of florida sparse matrix collection," *ACM Trans. Math. Softw.*, vol. 38, no. 1, Dec. 2011. [Online]. Available: https://doi.org/10.1145/2049662.2049663

[32] J. Shun and G. E. Blelloch, "Ligra: a lightweight graph processing framework for shared memory," in *ACM Symposium on Principles and Practice of Parallel Programming (PPoPP)*, 2013, pp. 135–146.

Appendix: Artifact Description/Artifact Evaluation

SUMMARY OF THE EXPERIMENTS REPORTED

The artifact contains the CUDA code for Sparse Matrix-Matrix Multiplication(SpMM) described in our paper. In addition it contains the codes, shell scripts and instructions needed to carry out other experiments in the evaluation section of the paper, including

(1) We ran our SpMM code on Group SNAP of SuiteSparse Matrix Collections (https://sparse.tamu.edu/SNAP) on NVIDIA GTX 1080Ti GPU and RTX 2080 GPU, and record execution time of average of 200 kernel runs, and compare with the baselines described in the paper. The program is compiled with CUDA v10.1 with -O3 flag.
(2) We modify the source code of DGL v0.4 library, rebuild the library and run the GCN and GraphSAGE sample codes provided by DGL v0.4 on the same two GPUs with PyTorch v1.4.0, and record CUDA time reported by PyTorch autograd.profiler and compare to results from DGL library before our modification.
(3) We create a PyTorch extension operator with our SpMM and modify GCN code provided by pytorch_geometric library to use this extension and run on the same two GPUs and record CUDA time reported by PyTorch autograd.profiler and compare to results with unmodified code.
(4) We write SpMM using GunRock v1.0 as described in the paper and ran on the same two GPUs on 3 citation graphs and record kernel time of 200 runs and compare with execution time of our SpMM.

The source code and associated scripts are available at https://github.com/hgyhungry/ge-spmm. One can follow the instructions in the repo to verify the results in the paper. To verify results in figure 10-14 and table 7 in the paper, one needs to use same hardware and software version, but this artifact is functional under more settings.

ARTIFACT AVAILABILITY

Software Artifact Availability: All author-created software artifacts are maintained in a public repository under an OSI-approved license.

Hardware Artifact Availability: There are no author-created hardware artifacts.

Data Artifact Availability: All author-created data artifacts are maintained in a public repository under an OSI-approved license.

Proprietary Artifacts: No author-created artifacts are proprietary.

Author-Created or Modified Artifacts:

```
Persistent ID: https://github.com/hgyhungry/ge-spmm
Artifact name: GE-SpMM CUDA code
```

BASELINE EXPERIMENTAL SETUP, AND MODIFICATIONS MADE FOR THE PAPER

Relevant hardware details: NVIDIA GTX1080Ti GPU, Host Machine Intel(R) Xeon(R) CPU E5-2643 v4(24 cores); NVIDIA RTX2080 GPU, Host Machine Intel(R) Core(TM) i7-9700K (8 cores)

Operating systems and versions: Linux (the results presented are based on experiments on Ubuntu 16.04)

Compilers and versions: NVCC v10.1

Libraries and versions: NVIDIA CUDA v10.1, NVIDIA cuSPARSE 10.1, DGL v0.4, PyTorch v1.4.0, GunRock v1.0

Input datasets and versions: Group SNAP in SuiteSparse Matrix collection (https://sparse.tamu.edu/SNAP). Two not valid (com-Orkut, soc-Pokec), two omitted due to out-of-memory (com-FriendSter, twitter). The matrix_id in experiment results are alphabetical order of the matrix names. The matrix name and their matrix_id are also listed in a file in the open-source repository.

ARTIFACT EVALUATION

Accuracy and precision of timings: The GPU kernel execution time is average of 200 kernel runs and only one program is running on GPU.