

A Framework for Fine-Grained Synchronization of Dependent GPU Kernels

Abhinav Jangda
Microsoft Research
United States

Saeed Maleki
Microsoft Research
United States

Maryam Mehri Dehnavi
University of Toronto
Canada

Madan Musuvathi
Microsoft Research
United States

Olli Saarikivi
Microsoft Research
United States

Abstract—Machine Learning (ML) models execute several parallel computations including Generalized Matrix Multiplication, Convolution, Dropout, etc. These computations are commonly executed on Graphics Processing Units (GPUs), by dividing the computation into independent processing blocks, known as *tiles*. Since the number of tiles are usually higher than the execution units of a GPU, tiles are executed on all execution units in one or more *waves*. However, the number of tiles is not always a multiple of the number of execution units. Thus, tiles executed in the final wave can under-utilize the GPU.

To address this issue, we present **cuSync**, a framework for synchronizing dependent kernels using a user-defined fine-grained synchronization policy to improve the GPU utilization. **cuSync** synchronizes tiles instead of kernels, which allows executing *independent* tiles of *dependent* kernels concurrently. We also present a compiler to generate diverse fine-grained synchronization policies based on dependencies between kernels. Our experiments found that synchronizing CUDA kernels using **cuSync** reduces the inference times of four popular ML models: MegatronLM GPT-3 by up to 15%, LLaMA by up to 14%, ResNet-38 by up to 22%, and VGG-19 by up to 16% over several batch sizes.

Index Terms—CUDA, GPU, Generalized Matrix Multiplication, Convolution, Fine-Grained Synchronization, Machine Learning

I. INTRODUCTION

The trend of larger Machine Learning (ML) models has delivered remarkable results in multiple domains. These results have exploded the demand of ML models in innumerable applications. To serve this demand, the infrastructure for running inference on these large models has also scaled up exponentially. Hence, optimizing for even the last percentage in the inference can lead to huge savings in cost and energy of serving these models.

ML models are typically served using multiple GPUs because these models consist of embarrassingly parallel operations, such as Generalized Matrix Multiplication (GeMM), 2-D Convolution (Conv2D) etc. The traditional approach to execute a computation on a GPU breaks down the computation into multiple independent blocks, known as *tiles*. Each tile is computed by a fixed size block of threads, known as a *thread block*, which runs on an execution unit of the GPU known as a *Streaming Multiprocessor* (SM). Often the number of thread blocks are higher than the number of SMs. Therefore, all thread blocks are executed in one or more *waves*, with initial full waves executing thread blocks that are a multiple of the number of SMs and the final partial wave executing

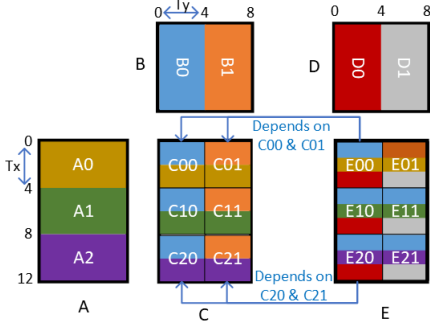
TABLE I: NUMBER OF THREAD BLOCKS (TBs), THREAD BLOCKS PER WAVE, WAVES, AND GPU UTILIZATION OF TWO DEPENDENT GEMMS IN MEGATRONLM GPT-3 [12] ON SEVERAL BATCH SIZES WHEN EXECUTING ON AN NVIDIA TESLA V100 CONTAINING 80 SMS.

Batch	GeMM	TBs	TBs per Wave	Waves	Utilization
256	Producer	[1, 48, 4]	2×80	1.2	60%
	Consumer	[1, 96, 2]	2×80	1.2	60%
512	Producer	[2, 24, 2]	1×80	1.2	60%
	Consumer	[2, 48, 1]	1×80	1.2	60%
1024	Producer	[4, 24, 2]	1×80	2.4	80%
	Consumer	[4, 48, 1]	1×80	2.4	80%

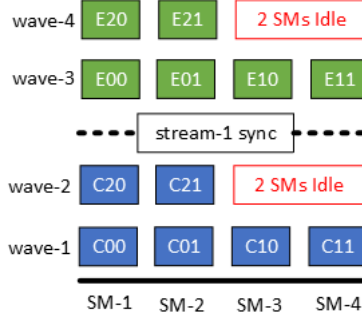
less than the number of SMs thread blocks. When executing a pair of dependent operations, the traditional approach executes these operations on the same *stream*. Executing two or more operations on a stream, ensures that no thread block of a later operation can execute before the thread blocks of all former operations are finished. We call this traditional heavy-weight synchronization approach as *stream synchronization*.

However, this heavy-weight synchronization can lead to the under-utilization of GPU resources in the final wave when thread blocks are not a multiple of SMs. For example, Figure 1a shows that executing 6 tiles of two dependent GeMMs on four SMs require $\lceil \frac{6}{4} \rceil = 2$ waves for each GeMM. With stream synchronization, no thread block of the second GeMM can execute before all thread blocks of the first GeMM are finished. Thus, as Figure 1b shows, the second partial wave of each GeMM utilizes only two out of four SMs. This under-utilization is prevalent in widely used ML models. Table I shows that during the inference of MegatronLM GPT-3 [12], the two dependent GeMMs achieves 60–80% of utilization on an NVIDIA Tesla V100 GPU because the number of thread blocks are not a multiple of the number of SMs.

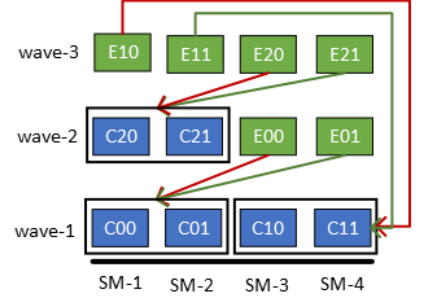
The state-of-the-art technique for executing GeMM computations on GPUs, Stream-K [10], can improve the utilization of the final wave of a workload by partitioning tiles of the final wave among multiple thread blocks. However, Stream-K suffers from three issues. First, partitioning a tile among multiple thread blocks requires each thread block to update the tile elements, leading to extra global memory accesses. Second, Stream-K requires different kernel invocations for



(a) *Tiled GeMM kernels*: A tile $C_{i,j}$ is computed by multiplying sub-matrices A_i and B_j . Similarly, a tile $E_{i,j}$ is computed by multiplying C_i with D_j . Since each tile is computed by one thread block, the tile size of 4×4 gives the grid size of $\{\frac{12}{4}, \frac{8}{4}\} = \{3, 2\}$ for both kernels. Both kernels have the occupancy of 1 thread block per SM.



(b) *Stream Synchronization* synchronizes all thread blocks of both kernels. Thread blocks of both producer ($C_{i,j}$) and consumer ($E_{i,j}$) are executed in two waves. The first wave executes four thread blocks and the second wave executes remaining two. In the second wave of both kernels, SM-3 and SM-4 are not utilized.



(c) *Fine-grained Synchronization* synchronizes only dependent thread blocks (shown as arrows) of both kernels and executes in only three waves. Thread blocks of the consumer-kernel waits using a semaphore until its producer-kernel's thread block has computed the dependent tile. Since in every wave all SMs are utilized, we achieve full utilization.

Fig. 1: Thread block execution with existing stream synchronization and fine-grained synchronization on 4 SMs for two dependent GeMM kernels: $C_{12 \times 8} = A_{12 \times 8} \times B_{8 \times 8}$ and $E_{12 \times 8} = C_{12 \times 8} \times D_{8 \times 8}$.

initial full waves and for the final partial wave. Third, it is not straightforward to extend Stream-K's approach to other tile based computations including Dropout and Softmax.

In this paper, we present several *fine-grained synchronization* techniques for synchronizing tiles of dependent computations enabling the execution of *independent* tiles of both computations concurrently in the final wave. Figure 1c shows how one of our techniques, tile synchronization, obtains full utilization in our example. We invoke both kernels on separate streams and synchronize only the dependent tiles, thus thread blocks, using a semaphore stored in the GPU memory. Therefore, thread blocks of both kernels are executed in only three waves, leading to full utilization of the GPU. However, as we show in the paper, the granularity of synchronization that provides the best performance depends on computations, data sizes, and GPU architecture. To this end, we propose, *cuSync*, a framework to efficiently synchronize dependent computations based on user-defined synchronization policies. *cuSync* contains mechanisms to: (i) ensure that all thread blocks of the producer are executed before the consumer (Section III-B), (ii) allow processing of producer and consumer tiles in an order that minimizes the wait time of synchronization by consumer tiles (Section III-C), (ii) maintain the dependence between tiles of producer and consumer computations using semaphores and memory fences (Section III-D). Furthermore, we propose a DSL to describe dependencies between GPU kernels and a compiler *cuSyncGen* to generate synchronization policies from the DSL specification for *cuSync* (Section IV). We described dependencies between computations of several ML models in the DSL and generated synchronization policies for diverse GPU computations, such as GeMM, 2-D Convolutions, and Dropout, using *cuSyncGen*. Synchronizing GPU computa-

tions using *cuSync* reduces the inference time of several state-of-the-art open source ML models on 8x NVIDIA Tesla V100 GPUs: MegatronLM GPT-3 145 Billion parameter model [12] by 6–15%, LLaMA 65.2 Billion parameter model [15] by 9–14%, ResNet-38 [6] by 5–22%, and VGG-19 [13] by 6–16% (Section V).

II. BACKGROUND

This section provides a background on NVIDIA GPUs and ML models.

A. NVIDIA Graphics Processing Units and CUDA

A parallel computation executing on NVIDIA GPUs is called a CUDA kernel. A CUDA kernel executes multiple concurrent *threads* organized in a 3-dimensional *grid*, and these threads are grouped into equally sized *thread blocks*. The `dim3` struct in CUDA represents a 3-D grid size and identifier for both threads and thread blocks in *x*, *y*, and *z* dimensions. An NVIDIA GPU contains multiple Streaming Multiprocessors (SMs), each of which executes one or more thread blocks. This number of thread blocks per SM, known as *occupancy*, depends on the register and shared memory usage, thread block size, and number of thread blocks of the CUDA kernel.

Thread block Wave Execution Thread blocks are executed on all SMs in $\lceil \frac{\text{Number_of_TBs_in_Grid}}{\text{occupancy} \times \text{Number_of_SMs}} \rceil$ waves, where the *initial full waves* execute $\text{occupancy} \times \text{Number_of_SMs}$ thread blocks and the *final partial wave* execute the remaining thread blocks. NVIDIA has not documented the mechanism for scheduling thread blocks to SMs followed by CUDA and GPUs.

Stream Synchronization A CUDA *stream* is a sequence of CUDA operations that execute in the order they were issued.

```

1 //X: [B,S,H]; W1: [H,4H/8]; W2: [4H/8,H]
2 //1st GeMM fused with GeLU XW1: [B,S,4*H/8]
3 XW1 = GeLU(X × W1)
4 //2nd GeMM XW2: [B,S,H]
5 XW12 = XW1 × W2

```

(a) Multi-Layer Perceptron (MLP) contains two weight matrices: W_1 of shape $[H, \frac{4H}{8}]$ and W_2 of shape $[\frac{4H}{8}, H]$.

```

1 //X: [B,S,H]; QKV: [H,3H/8]; W2: [H/8,H]
2 //1st GeMM XQKV: [B,S,3H/8]
3 XQKV = X × QKV
4 //XQ: [B,S,H/8]; XK: [B,S,H/8]; XV: [B,S,H/8]
5 XQ = XQKV[:, :, 0:H/8] //1st matrix slice
6 XV = XQKV[:, :, H/8:2*H/8] //2nd matrix slice
7 XK = XQKV[:, :, 2*H/8:] //3rd matrix slice
8 //Cached Attention Mechanism
9 //CachedK: [H/8,S',B]
10 //CachedV: [B, S', H/8]
11 P = XQ × Concat(CachedK, XK.T)
12 R = Softmax(Dropout(P))
13 T = R × Concat(CachedV, XV)
14 CachedV[:, S'+S:] = T
15 CachedK[:, S'+S:] = XK.T
16 //2nd GeMM XW2: [B,S,H]
17 XW12 = R × W2

```

(b) Attention contains two weight matrices: QKV of shape $[\frac{3H}{8}, H]$, and W_2 of shape $[\frac{H}{8}, H]$. Attention caches generated keys and values for each token to avoid recomputation of all previous tokens during inference.

Fig. 2: Architecture of Multi-Layer Perceptron (MLP) and Attention of GPT-3, where H is 12288. Model parallelism on 8 GPUs divides weight matrices of both layers among 8 GPUs. Both takes an input matrix X of shape $[B, S, H]$ and obtain the result XW_{12} of the same shape. B is the number of batched requests, S is the sequence length, H is the hidden dimension, and S' is the sum of processed and generated tokens.

When two dependent kernels are invoked on the same stream, the consumer-kernel is not started before all thread blocks of the producer-kernel have finished their execution. We call this synchronization *stream synchronization*. We can invoke independent CUDA kernels on different streams to execute kernels concurrently. A stream has an associated priority value, such that operations on a higher priority stream are issued before a lower priority stream.

B. Computations in Large ML Models

Contemporary ML models contain embarrassingly parallel computations, such as, Generalized Matrix Multiplication (GeMM), 2-D Convolution (Conv2D), Dropout, and Softmax. We consider four widely used machine learning models: MegatronLM GPT-3 145B [12], LLaMA 65.2B [15], ResNet-38 [6], and VGG-19 [13]. Below we briefly explain computations involved in these models.

1) *Transformers Models*: A *transformer* is a deep learning architecture for Natural Language Tasks and is the basis of two widely used models: MegatronLM GPT-3 [12] and LLaMA [15].

```

1 //X: [B, S, H]; W1: [H, H/3];
2 //V: [H, H/3]; W2: [H/3, H]
3 //1st GeMM XW1: [B, S, H/3]
4 XW1 = GeLU(X × W1)
5 //2nd GeMM XV: [B, S, H/3]
6 XV = X × V
7 //SwiGLU fused with 3rd GeMM XW2: [B, S, H]
8 SwiGLU = Swish(XW1) · XV
9 XW12 = SwiGLU × W2

```

Fig. 3: The LLaMA MLP contains three weight matrices. With model parallelism on 8 GPUs, these matrices are: W_1 of shape $[H, \frac{H}{3}]$, V of shape $[H, \frac{H}{3}]$, and W_2 of shape $[\frac{H}{3}, H]$.

An inference request to a transformer model consists of a prompt and is served in two phases: (i) *prompt processing*, where the prompt is processed, and (ii) *token generation*, where a series of tokens that represents the output response text is generated incrementally. The model can batch B requests into a single inference task. The sequence length S denotes the number of tokens of each request being processed in the prompt processing phase or the number of tokens of each request being generated in the token generation phase. Therefore, during prompt processing $B \geq 1, S > 1$ and during token generation $B \geq 1, S = 1$.

A transformer consists of multiple Multi-Layer Perceptron (MLP) and Attention blocks. The design of MLP and Attention can be different for each model.

GPT-3: In GPT-3, both MLP and Attention takes an input matrix, perform operations with its two weight matrices, and outputs a matrix. With model parallelism these weight matrices are divided among all GPUs [12]. Figure 2 shows computations of GPT-3 with model parallelism of 8 GPUs. Both MLP and Attention first applies a linear transformation on the input, i.e., perform GeMM of the input and the weight matrix. Then, they perform operations, such as, GeLU and the Attention mechanism. Finally, the output of this operation is applied to second linear transformation. Existing MLP implementations fuse the GeLU activation with the first GeMM (line 4 in Figure 2a). State-of-the-art Attention implementations [4] caches the already processed and generated tokens in a KV Cache, such that, after prompt processing number of cached tokens, i.e. S' , is set to S and when generating tokens S' increases incrementally. These implementations also fuses the attention mechanism in a single CUDA kernel (line 11–line 13) in Figure 2b.

LLaMA: LLaMA uses the hidden dimension of size 8192. LLaMA’s MLP contains three GeMMs and SwiGLU [11] activation as shown in Figure 3. State-of-the-art implementations combines first two GeMMs into a single GeMM and fuses the SwiGLU activation with the third GeMM. Moreover, LLaMA uses the same Attention architecture as GPT-3.

2) *Computer Vision Models*: ResNet-38 [6] and VGG-19 [13] are two state-of-the-art computer vision models, where each layer performs several Conv2D operations. Table II shows

TABLE II: INPUT/OUTPUT IMAGE SIZE (P, Q, C), KERNEL SIZE (R, S), CHANNELS (K) FOR EACH CONV2D, NUMBER OF CONV2DS PER LAYER, AND NUMBER OF LAYERS IN RESNET-38 AND VGG-19.

[P, Q, C]	[R, S]	K	Convs/Layer		Layers	
			ResNet	VGG	ResNet	VGG
[56, 56, 64]	[3, 3]	64	2	2	3	1
[28, 28, 128]	[3, 3]	128	2	2	4	1
[14, 14, 256]	[3, 3]	256	2	4	6	1
[7, 7, 512]	[3, 3]	512	2	4	3	1

the details of each convolution layer.

III. FINE-GRAINED SYNCHRONIZATION OF KERNELS

Our fine-grained synchronization of dependent CUDA kernels consists of four novel mechanisms. These mechanisms (i) ensure simultaneous allocation of dependent kernels (Section III-A), (ii) execute thread blocks of producer kernels before consumer kernels (Section III-B), (iii) control the order of tile processing in each kernel to minimize the wait time of synchronization (Section III-C), and (iv) performs fine-grained synchronization of only dependent tiles of producer and consumer kernels (Section III-D). We have implemented these mechanisms in a header-only standalone CUDA library, *cuSync*.

Figure 4a explains these mechanisms using an example of synchronizing the two dependent GeMM kernels of MLP using *cuSync*. The `gemm` function is the standard GeMM GPU kernel (we use NVIDIA CUTLASS [1] for our experiments) with additional code (shown as underlined) to call into *cuSync*. *cuSync* associates each kernel with a `CuStage` object that provide synchronization facilities among kernels. The MLP function creates these stage objects, declares dependencies between them, and invokes kernels.

A. Invoke Dependent Kernels

The first requirement for fine-grained synchronization is to eliminate the stream synchronization between kernels. *cuSync* achieves this by invoking all kernels on different CUDA streams. The example creates producer (`prod`) and consumer (`cons`) stages for both GeMM kernels (lines 18–20 in Figure 4a). Then, the example declares the dependency between the two stages by specifying that the output of the producer is the input of the consumer (line 23). Finally, the example invokes both kernels on different streams associated with respective stages (lines 26 and 30). Section III-D describes how *cuSync* enforces this dependency.

B. Stage Processing Order

The second requirement for fine-grained synchronization is to execute all full waves of the producer kernel before the consumer kernel. However, the CUDA runtime lacks any mechanism to enforce this execution order among kernels belonging to different streams. Hence, there is a possibility that the consumer kernel is scheduled on the GPU before the producer kernel. This can lead to poor performance as thread blocks of the consumer kernel occupy SMs without doing any

useful work. In the worst case, this can lead to deadlocks if no SMs are available for the producer kernel.

cuSync ensures this requirement by enforcing the scheduling order of kernels using its *wait-kernel* mechanism. The *wait* kernel is invoked by the consumer stage on the consumer stream before the consumer kernel (line 28 in Figure 4a). The *wait* kernel contains a single thread, which waits on a global memory semaphore for each consumer kernel using a busy-wait while loop. When the producer kernel calls the `stage.start()` function (line 4), the function sets the semaphore using the first thread of the first thread block, which in-turn exits the *wait-kernel*. After the *wait-kernel* exits, the CUDA runtime can invoke the consumer kernel. Thus, *cuSync* ensures that no thread blocks of the consumer kernel are scheduled before at least one of the thread blocks of the producer kernel.

The *wait-kernel* mechanism assumes that CUDA schedules thread blocks of kernels in the order the kernels are invoked by CUDA. We have found that the latest versions of CUDA 11 and 12 executing on NVIDIA GPUs based on Volta and Ampere architecture follow this schedule.

C. Custom Tile Order

The third requirement for efficient synchronization is minimizing waiting time of consumer kernels. However, the CUDA runtime can schedule thread blocks on SMs in any arbitrary order, which can lead to unpredictable wait times. Ideally, thread blocks of the consumer kernel should be scheduled in the order the producer kernel generate tiles.

cuSync enables execution of both producer and consumer kernel’s thread blocks in a custom scheduling order independent of how the CUDA runtime schedules thread blocks. In our example, each thread block calls `stage.tile()` (line 4) to obtain the tile it needs to compute. The parameter `RowMajor` (lines 18–20) ensures that both kernels produce tiles in a row major order, i.e., first all thread blocks in *x*, then in *y*, and finally in *z*. Figure 4b defines the `RowMajor` order as a function (line 29). A tile order function takes a tile index in the 3-D grid and returns a distinct 1-D index for the tile. Internally, *cuSync* maintains an array that maps a linear tile index to a 3-D index. For each thread block, *cuSync* increments an atomic global counter and returns the 3-D index in the array for the previous counter value. In summary, *cuSync* allows easy experimentation with diverse scheduling orders to obtain the best performance.

D. Synchronizing Dependent Tiles

The final requirement for fine-grained synchronization is to ensure that the dependence between tiles of producer- and consumer-kernels is maintained using a synchronization mechanism. *cuSync* provides two functions, `wait` and `post` to enforce this dependency. For instance, in our example, the `wait` function is called twice (line 6 and 8) before loading the tiles of A and B, and the `post` function is called once (line 12) for the producer kernel after computing the tile. However, the consumer kernel only needs to wait on the output of the producer kernel, i.e., input A of the consumer kernel.

```

1 //CUDA Kernel to compute C = A * B
2 global void gemm(f16* A, f16* B, f16* C,
3                 int K, CuStage stage) {
4     stage.start(); row, col = stage.tile();
5     for (tk = 0; tk < K; tk += TileK) {
6         stage.wait(A, row, tk);
7         LoadTileToShMem(Ash, A, row, tk);
8         stage.wait(B, col, tk);
9         LoadTileToShMem(Bsh, B, col, tk);
10        MultiplyAccumulate(C, Ash, Bsh,
11                           row, col, tk);
12    } stage.post(row, col);
13 }
14 void MLP(int BS, int H, f16* X, f16* W1,
15          f16* XW1, f16* W2, f16* XW12) {
16     dim2 grid1 = {4*H/8, B}/tile1;
17     dim2 grid2 = {H, B}/tile2;
18     CuStage<RowMajor, RowSync>
19         prod(grid1, tile1);
20     CuStage<RowMajor, RowSync>
21         cons(grid2, tile2);
22     // declare prod to cons[XW1] dependency
23     CuSync::dependency(prod, cons, XW1);
24     // invoke the producer gemm
25     gemm<<<grid1, tb1, prod.stream()>>>
26         (X, W1, XW1, H, prod);
27     // invoke waitKernel and then consumer
28     cons.waitKernel();
29     gemm<<<grid2, tb2, cons.stream()>>>
30         (XW1, W2, XW12, 4*H/8, cons);

```

(a) The kernels are invoked on different streams. The wait kernel ensures the order of kernel invocation. The post and wait methods ensure tile dependency. Changes to the GeMM kernel are underlined.

```

1 class CuStage<Policy>
2 void init() {
3     sems = /*Init semaphores using Policy*/
4 void post(dim2 tile, dim2 grid) {
5     __syncthreads();
6     if(threadIdx == {0,0,0})
7         __threadfence_system();
8     sem = &sems[Policy.sem(tile, grid)];
9     atomicAdd(sem, 1);
10 void wait(dim2 tile, dim2 grid) {
11     sem = &sems[Policy.sem(tile, grid)];
12     if(threadIdx == {0,0,0})
13         while(*sem != Policy.value(tile, grid));
14     __syncthreads();
15
16 class TileSync
17 int sem(dim2 tile, dim2 grid) {
18     //Distinct semaphore for each tile
19     return tile.x*grid.y + tile.y;
20 int value(dim2 tile, dim2 grid){return 1;
21
22 class RowSync
23 int sem(dim2 tile, dim2 grid) {
24     //Tiles of same row share semaphore
25     return tile.y;
26 int value(dim2 tile, dim2 grid) {
27     return grid.x;
28
29 int RowMajor(dim2 tile, dim2 grid){
30     return tile.y*grid.x + tile.x;

```

(b) TileSync creates a semaphore for each tile. RowSync trades concurrency for synchronization by creating a single semaphore per row.

Fig. 4: Fine-grained synchronization of two GeMMs of MLP using cuSync’s TileSync and RowSync policies.

This dependency is specified in line 23. Therefore, the wait before loading a tile of A waits for the corresponding post of the producer kernel, and the wait before loading a tile of B becomes a no-op. Since the producer kernel have no dependency, both waits are no-ops for the producer kernel.

cuSync provides a mechanism for synchronizing producer and consumer tiles based on an arbitrary *synchronization policy* (or policy in short). cuSync uses an array of global memory semaphores for synchronization, where each producer tile is associated with only one semaphore and a semaphore’s value represents the status of its producer tiles. Thus, a policy is a mapping of one or more producer tiles to one semaphore. For example, the finest grained synchronization policy, we call *TileSync*, waits for each producer tile and is defined as a one-to-one map of a producer tile to a semaphore. A policy requires implementation of two methods: (i) *sem*, which returns the semaphore for the given tile, and (ii) *value*, which returns the expected value of semaphore when the tile is ready. We below describe details of three methods of CuStage required for our synchronization mechanism (lines 2–9 in Figure 4b).

init: The *init* method allocates and initializes the array of semaphores in the global memory based on the given policy.

post: The *post* method calls `__syncthreads` and a

memory fence to ensure that all threads of the thread block has computed the tile and all global memory writes are visible to other kernels (line 5–7). Finally, the method obtains the semaphore for the tile using the policy and increments the semaphore (line 9).

wait: The *wait* method obtains the semaphore for the given tile using the policy and then wait on the value of semaphore in a while loop using only the first thread of the thread block (line 13). While the first thread is waiting, all other threads of the thread block are blocked on the `__syncthreads` (line 14). When the semaphore changes to the expected value, all threads of the thread-block proceeds from the `__syncthreads`.

E. Synchronization Policies

cuSync allows implementation of diverse synchronization policies easily. As described earlier, each policy requires implementing *sem* and *value* methods. Below we discuss two general policies that are applicable to all kernels in our workloads.

TileSync is the finest-grained policy that synchronizes on each producer tile (lines 16–20 in Figure 4b). To minimize the wait time of the consumer-kernel, both kernels compute their

tiles in a row major order. The `sem` method returns distinct semaphore for each tile (line 17) and the `value` method returns 1 to signify that the tile is computed (line 20). For example, in Figure 4a to compute a tile E^{xy} , the TileSync policy requires waiting first on C^{x0} and then on C^{x1} .

RowSync synchronizes on each row of the producer kernel requiring less synchronizations than TileSync (lines 22–27 in Figure 4b). For example, for two GeMMs of Figure 4a, TileSync requires 12 synchronizations in total, while RowSync requires 6 synchronizations by sharing the same semaphore for all tiles computing the same row of C . Thus, the `sem` method returns the row of the given tile and the `value` method returns the value when the row is ready, i.e., the number of tiles in a row (line 23–26). To minimize the wait time, both kernels schedule their tiles in a row major order. RowSync can also be used for synchronizing Conv2D kernels. Section V shows that for large GeMMs and Conv2Ds the high number of synchronizations is a bottleneck.

IV. AUTO-TUNING OF POLICIES AND TILE ORDERS

The process of obtaining the best performance involves experimenting with several synchronization policies and tile processing orders. The best policy and tile order depends on computations, data sizes, and the GPU architecture. However, doing this process manually is both tedious and error-prone.

Therefore, `cuSyncGen` is a tool that takes dependencies specified by the user and generates the optimal tile processing order and multiple synchronization policies as CUDA code for `cuSync`. `cuSyncGen` currently requires the user to manually modify the GPU kernels to instantiate `CuStage` with generated policies and tile processing order similar to the MLP example (Figure 4a). The modularity of `cuSync` allows the user to easily plug diverse synchronization policies and tile processing orders.

A. Workflow

The workflow of `cuSyncGen` is as follows:

- 1) The user describes a chain of dependencies between kernel tiles and the grid values for all kernels.
- 2) `cuSyncGen` checks bounds of producer and consumer tiles based on grid values.
- 3) `cuSyncGen` generates a tile processing order as CUDA code that minimizes the wait time.
- 4) `cuSyncGen` generates CUDA code for multiple policies.
- 5) The user modifies the workload to support `cuSync` and plugs the generated CUDA code to `cuSync`.

The rest of the section describes each of these steps.

Describe Dependencies The user describes dependencies between tiles of kernels using a DSL embedded in C++. Figure 5a shows the dependency between both GeMMs of MLP described in the DSL. First, the DSL code must define each kernel’s grid dimensions with their maximum value. The example defines x and y dimensions for both grids (line 1–4). Specifying the exact values for a grid enables generating efficient code and doing bounds checking for correctness.

```
1 Dim x, y;
2 //Max value of all dimensions of both GeMMs
3 Grid g1(x, y,  $\frac{H}{2*TileN}$ ,  $\frac{B*S}{TileM}$ );
4 Grid g2(x, y,  $\frac{B}{TileN}$ ,  $\frac{B*S}{TileM}$ );
5 //Tile is produced by each thread block
6 Tile prod(x, y), cons(x, y);
7 //All col tiles for a row from 0 to  $\frac{H}{2*TileN}$ 
8 ForAll prodCols(prod, x, Range(g1.x));
9 //Tile of 2nd GeMM depends on all
10 //col tiles of 1st GeMM
11 Dep dep({g2, cons}, {g1, prodCols});
```

(a) GPT-3’s MLP

```
1 Dim x, y;
2 //First GeMM Grid
3 Grid g1(x, y,  $\frac{3*H}{8*TileN}$ ,  $\frac{B*S}{TileM}$ );
4 //P, R, and T Grid
5 Grid gP(x, y,  $\frac{B*(S+S')}{TileN}$ ,  $\frac{B*(S+S')}{TileM}$ );
6 Grid gR(x, y,  $\frac{B*(S+S')}{TileN}$ , 1);
7 Grid gT(x, y,  $\frac{B*(S+S')}{TileN}$ ,  $\frac{H}{8*TileM}$ );
8 //Second GeMM Grid
9 Grid g2(x, y,  $\frac{H}{8*TileN}$ ,  $\frac{B}{TileM}$ );
10 //P to 1st GeMM
11 //Strided Tile Dependencies: stride= $\frac{H}{8*TileN}$ 
12 Dep dep1P({gP, Tile(x,y)},
13 {g1, Tile(x,y), Tile(x+ $\frac{H}{8*TileN}$ , y)});
14 Dep depPR({gR, Tile(x,y)},
15 {gP, ForAll(Tile(x,y), y, Range(gP.y))});
16 Dep depTR1({gT, Tile(x,y)},
17 {gR, Tile(x, y)}, {g1, Tile(x+ $\frac{2*H}{8*TileN}$ , y)});
18 //2nd GeMM to T
19 dep23({g3, Tile(x,y)}, {gT, Tile( $\frac{x}{TileM}$ , y)});
```

(b) Attention

```
1 Dim x, y;
2 //First GeMM Grid
3 Grid g1(x, y,  $\frac{C}{TileM}$ ,  $\frac{B*P*Q}{TileN}$ );
4 //Second GeMM Grid
5 Grid g2(x, y,  $\frac{C}{TileM}$ ,  $\frac{B*P*Q}{TileN}$ );
6 //2nd Conv2D to 1st Conv2D
7 Dep dep({g2, Tile(x,y)}, {g1, Tile( $\frac{x}{R*S}$ , y)});
```

(c) Two Conv2Ds

Fig. 5: Dependencies in the `cuSyncGen` DSL. $TileM$ and $TileN$ are tile size of GeMMs in row and column respectively.

Then, the DSL code constructs producer and consumer tiles by specifying an affine function over each dimension of the grid. The example creates a producer and consumer tile for each thread block in the grid and creates a range of column tiles using `ForAll` (line 6–8). Finally, the code specifies the dependence between one consumer tile and one or more producer tiles (line 11).

Generate Tile Processing Order `cuSyncGen` generates a tile processing order for each kernel to minimize the waiting time. To discuss the process, consider a dependency where a consumer tile, $C(x, y)$, depends on N producer tiles, $\{P(x, a_0y + b_0), P(x, a_1y + b_1), \dots, P(x, a_{N-1}y + b_{N-1})\}$.

We achieve minimum wait time when the consumer kernel consumes tiles in the same order as they are produced by the producer kernel. Thus, we schedule all N producer tiles consecutively for each consumer tile using the following code:

```
1 int prodOrder(dim2 tile, dim2 grid) {
2   int linear = bid.y*grid.x + bid.x, y = 0;
3   if (tile.y%a0 <= b0) y = 0;
4   //Similarly for tiles till N-2
5   else if (tile.y%aN-1 <= bN-1) y = N-1;
6   return linear/N+y;}
```

This code obtains the 1-D linear index of a tile, finds the tile index within the group of N tiles, and returns the new linear index. We also set the consumer kernel to follow the row major order of tiles. Our MLP example uses the row major order, i.e., all groups of $\frac{H}{\text{TileN}}$ consecutive producer tiles are scheduled consecutively. It is straightforward to extend this method to a chain of dependent kernels by extending the dependence from the last consumer kernel to the first producer kernel and then generating code for each kernel.

Generating Policies `cuSyncGen` generates multiple synchronization policies for each dependence. For the following discussion, consider a dependence where a consumer tile, $C(x, y)$, depends on N producer tiles, $\{P(x, a_0y + b_0), P(x, a_1y + b_1), \dots, P(x, a_{N-1}y + b_{N-1})\}$. `cuSyncGen` generates two policies for the dependence in each dimension: (i) map each tile to a distinct semaphore, or (ii) map all N tiles to the same semaphore. The code generated for the considered dependence and the value of $M \in \{1, N\}$ is:

```
1 int sem(dim2 tile, dim2 grid) {
2   int y = 0;
3   if (tile.y%a0 <= b0)
4     y = (tile.y-b0)/a0;
5   //Similarly for tiles till M-2
6   else if (tile.y%aM-1 <= bM-1)
7     y = (tile.y-bM-1)/aM-1;
8   else y = tile.y;
9   return y*grid.x + tile.x;}
10 int value(dim2 tile, dim2 grid) {return M;}
```

After considering both cases for the innermost dimension, the phase moves to the outer dimension, and follows the same method. In our MLP example, `cuSyncGen` generates two policies: (i) `TileSync` that maps each tile to a distinct semaphore, and (ii) `RowSync` that maps all column tiles of the same row to the same semaphore.

Running the Generated Code We require the user to modify the workload to support running `cuSync` by adding `wait` calls before every tile load and `post` call after computing a tile. For example, in the case of MLP, we require the user to do the changes of Figure 4a. The modularity of `cuSync` enables plugging multiple policies and tile processing order. So, the user can execute all generated policies and obtain the policy with least execution time.

B. Computation Dependencies in ML Models

We now show how to specify dependencies of Attention and Conv2D cases in `cuSyncGen`.

Attention contains two dependencies between its three kernels

(Figure 5b). In the first dependency, an element of the dot product depends on three elements in the same row with a stride of $\frac{H}{8}$ of the first GeMM output (line 13). In addition to `TileSync` and `RowSync`, for this dependence `cuSyncGen` also generates a policy, we call *StridedSync*, that maps all three producer tiles of the first GeMM to the same semaphore. Thus, *StridedSync* waits until all three tiles of the first GeMM are computed before continuing with the dot product of tiles. Moreover, `cuSyncGen` generates the tile order that schedules these three tiles consecutively. For other dependencies, `cuSyncGen` generates both `TileSync` and `RowSync`, while processing tiles in a RowMajor order.

Conv2D using the implicit GeMM algorithm converts a convolution of B input images of size $[P, Q, C]$ with a kernel matrix of size $[R, S]$ into a GeMM of matrices $[B \times P \times Q, C \times R \times S]$ with $[C \times R \times S, C]$. Figure 5c shows the dependency between two Conv2Ds using the implicit GeMM algorithm. Thus, the dependency describes that each tile of the second implicit GeMM depends on all column tiles of the first implicit GeMM output (line 7). `cuSyncGen` generates two policies for this dependency: (i) `RowSync` to synchronize each row, and (ii) *Conv2D TileSync* policy to synchronize each tile. Moreover, `cuSyncGen` generates a row major order for both Conv2Ds.

C. Optimizations

`cuSyncGen` automatically perform several optimizations to improve the performance of a `cuSync` synchronized workload. These optimizations depend on the architecture details of the GPU, occupancy of CUDA kernels, and grid sizes. The optimizations are as follows:

Avoid Wait Kernel The wait-kernel mechanism ensures that all thread blocks of the producer kernel are scheduled on the GPU before the consumer kernel. However, if both producer and consumer kernels can be executed in less than two waves, we do not need the wait-kernel mechanism.

Avoid Custom Tile Processing Order We can also avoid a custom tile processing order when all tiles of producer and consumer-kernels can be executed in two waves.

Reorder Tile Loads and Synchronization The general workflow of tile based CUDA kernels is to load a tile of all inputs into shared memory or registers and then perform operations on all tiles. We can re-order the waiting of tile of one input with the loading of other tile, to overlap the waiting of one tile with the loading of the other input's tile. For example, in Figure 4a the second GeMM kernel loads a tile of both inputs (A and B) and compute the tile of output matrix (C) (line 6–9). We can reorder the loading of B tile with the waiting on A tile, i.e., swap lines 6–7 with lines 8–9. Since there is no waiting for tile of B, loading a B tile can overlap with waiting of A tile, leading to improved performance. `cuSyncGen` automatically performs the reordering if the user annotate tile loading in kernels with `#pragma tile`.

TABLE III: FRACTION OF LINES OF CODE CHANGED IN GEMM, FUSED SOFTMAX-DROPOUT, AND CONV2D KERNELS TO SUPPORT USING `cuSync`.

Kernel	Implementation	Lines Changed	
		Number	Fraction
GeMM	CUTLASS	25	0.5%
Softmax-Dropout	Ours	5	1%
Conv2D	CUTLASS	22	0.6%

V. EVALUATION

We now evaluate the performance of `cuSync` against state-of-the-art baselines using large open source ML models as our workloads.

A. Experimental Setup

We run our experiments on a machine with a 2.60GHz 12-core Intel Xeon E5-2690 CPU with 448GB RAM and 8 NVIDIA Tesla V100 32GB GPUs connected with NVLINK. We use CUDA 12.2 and report the average time of 20 executions after a warmup of 5 executions.

ML Models We used `cuSync` to synchronize CUDA kernels of four ML models: MegatronLM GPT-3 145 Billion [12], LLaMA 65.2 Billion [15], ResNet-38 [6], and VGG-19 [13]. We used the GeMM and Conv2D CUDA kernels of NVIDIA CUTLASS 3.1 (Figure 2b). We fuse the pointwise computations with GeMM and Conv2D kernels and developed a fused kernel of Softmax and Dropout in the Attention. We evaluate the reduction in inference times of these models using `cuSync` on batch sizes from 1 to the largest supported batch size by each model.

Baselines We consider the following baselines:

StreamSync is the CUDA stream synchronization.

Stream-K [10] partitions the last thread block wave of GeMM and Conv2D among all SMs to improve the GPU utilization.

B. Ease of Programming

Table III shows that the number of lines added and changed to support fine-grained synchronization of GeMM, Conv2D, and Softmax-Dropout kernels using `cuSync` are negligible compared to the lines of code of these kernels. Thus, the `cuSync` approach enables diverse synchronizations of tile based computation kernels through few modifications.

C. Applicability in ML Models Inference

We now discuss the applicability of `cuSync` in improving the performance of ML models from the perspective of kind of computations and the average utilization of GPU. First, ML models majorly consists of tile based GPU kernels, such as GeMM and Conv2Ds. Since `cuSync` supports any tile based kernel, we can use `cuSync` to synchronize kernels of ML models. Second, since the number of waves of each kernel increases with the batch size, the average utilization of all waves also increases. However, each ML model supports a maximum batch size limit during both training and inference phases. For example, the maximum token length supported

by GPT-3 and LLaMA is 2048. We show in our experiments that even for this maximum batch size, GPU kernels suffer from low number of waves leading to low average utilization. In summary, `cuSync` is applicable to diverse ML models because ML models largely contains tile based kernels and the maximum batch size supported by ML models still suffers from under-utilization.

D. Maximum Overhead of Synchronization

The synchronization mechanism has two sources of overhead: global memory accesses and `__syncthreads`. The percentage of total overhead depends on the amount of computations performed by the GPU kernel. A kernel doing large amount of computations on each tile would suffer from less synchronization overhead than a kernel doing less amount of computations. We can obtain an upper bound on the overhead by having two kernels (i) doing minimum computations on each tile, (ii) execute maximum number of thread blocks per wave, and (iii) execute one full wave.

We design such an experiment where the producer kernel copies values from an input array to an intermediate output array by assigning consecutive threads to contiguous elements, and similarly the consumer kernel copies values from the intermediate array to a final output array. Thus, a thread block of the consumer depends on the same thread block of the producer. We invoke both kernels with the maximum number of thread blocks per wave on Tesla V100, i.e., $Number_of_SMs \times Max_Occupancy = 80 \times 16 = 1280$. We found that synchronization using `cuSync` leads to 2-3% overhead over StreamSync. Hence, `cuSync`'s synchronization mechanism provides low overhead.

E. Large Language Model Inference Results

We now evaluate the reduction in the inference times of GPT-3 and LLaMA with model parallelism on 8 GPUs using `cuSync` for both prompt processing and token generation phase (Figure 2). In prompt processing, we consider the total number of tokens in an inference task, i.e., $B \times S$ from 512 to 2048, and in token generation, we consider batched requests, i.e., B from 1 to 4 with number of already generated tokens, i.e., S' from 512 to 2048. We used `cuSyncGen` to generate the following policies:

RowSync+WRT synchronizes rows and executes thread blocks in the row major order by adding our optimizations of Section IV-C, i.e., avoiding the wait-kernel (W), avoiding custom tile order (T), and reorder tile loads (R).

TileSync synchronizes tiles and executes thread blocks in the row major order.

TileSync+WRT extends TileSync by adding our optimizations of Section IV-C.

Strided+TileSync+WRT, only for Attention, synchronizes the first GeMM with the first GeMM of Cached mechanism using StridedSync, and all other kernels using TileSync (Figure 5b). The policy also add our optimizations of Section IV-C.

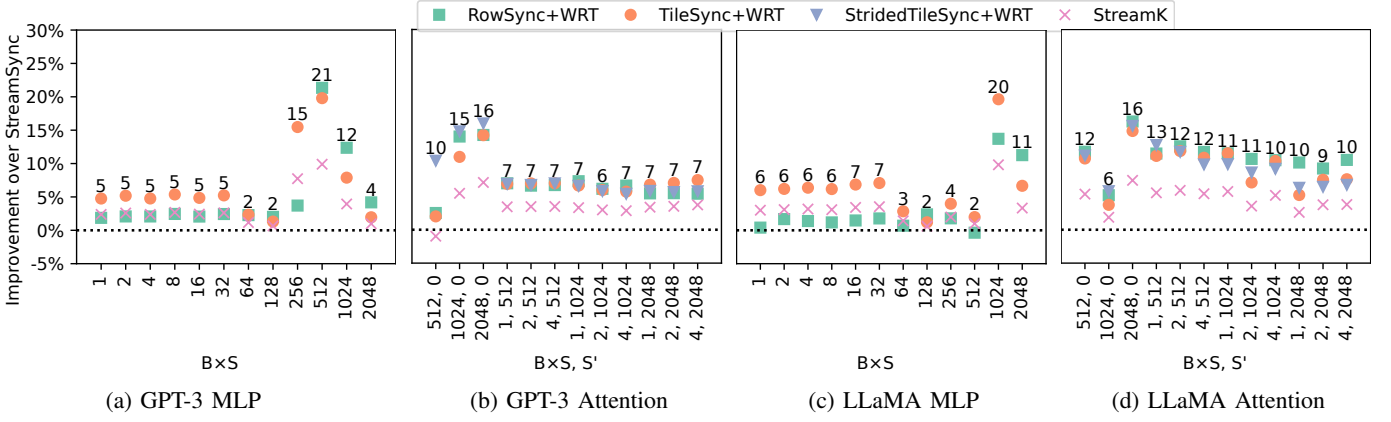


Fig. 6: Improvement of `cuSync`'s policies and StreamK in MLP and Attention over StreamSync for batch sizes 1–2048. During prompt processing $S' = 0, B \times S > 1$ and in token generation $S' > 1, B \geq 1, S = 1$. Numbers shows the maximum speedup out of all policies.

TABLE IV: GRID SIZE, NUMBER OF WAVES, TOTAL WAVES AND EXECUTION TIME IN STREAMSYNC AND `cuSync` FOR BOTH GEMMS OF GPT-3's MLP. THE GRID X AND Y-DIMS ARE OBTAINED BY DIVIDING THE SIZE OF GEMM WITH THE TILE SIZE AND THE Z-DIM IS THE NUMBER OF THREAD BLOCKS USED FOR SPLIT-K.

Batch Size	First GeMM		Second GeMM		StreamSync		cuSync			Decrease in Runtime
	Grid	Waves	Grid	Waves	Waves	Time(μ s)	Waves	Policy	Time(μ s)	
1–64	$1 \times 24 \times 4$	0.6	$1 \times 48 \times 3$	0.9	2	378	1.8	Tile	355	5–6.0%
128	$1 \times 24 \times 3$	0.4	$1 \times 48 \times 3$	0.9	2	530	1.3	Tile	523	2%
256	$1 \times 48 \times 4$	1.2	$1 \times 96 \times 2$	1.2	4	862	2.4	Tile	728	16%
512	$2 \times 24 \times 2$	1.2	$2 \times 48 \times 1$	1.2	6	1500	4.8	Row	1196	21%
1024	$4 \times 24 \times 2$	2.4	$4 \times 48 \times 1$	2.4	5	2111	3.6	Row	1901	10%
2048	$8 \times 24 \times 1$	2.4	$8 \times 48 \times 1$	4.8	8	3730	7.2	Row	3574	4%

1) *MLP Results*: Figure 6a and 6c shows that synchronizing dependent GeMMs of the GPT-3 MLP and LLaMA MLP using `cuSync` decreases the execution time of both MLPs by up to 20% for different sizes. We discuss these results using Table IV that shows the number of waves for all batch sizes for GPT-3 MLP using both StreamSync and `cuSync`.

TileSync+WRT performs best for $B \times S$ of 1 to 256 because there is a single thread block in the x-dimension of grid (Table IV). The improvement at size 256 is higher than small sizes because TileSync+WRT reduces the number of waves by 1 over StreamSync. On small batch sizes, even though the number of waves is not decreased, TileSync+WRT performs 7% faster because the second GeMM can overlap the loading of W_2 tile into the shared memory with the computation of the first GeMM.

RowSync performs best for sizes greater than 512 because synchronizing over a row once reduces memory accesses than synchronizing over multiple tiles and more number of rows provides more opportunities for overlapping. Therefore, increasing the number of rows also increases the speedup of RowSync from 4% at 256 to 20% at 1024. However, the speedup decreases to 4% at 2048 because the fraction of waves reduced by `cuSync` decreases with more thread blocks in the grid.

Effect of Overlapping Kernel Invocations We measured the time of a kernel invocation is $\approx 6\mu$ s, which is significantly lower

than the difference in the execution time of StreamSync and `cuSync`. Table IV shows that the difference in execution times with `cuSync` and StreamSync is significantly higher than the time to invoke a kernel. Hence, the performance improvement of `cuSync` is significantly higher than what would be achieved by only overlapping the invocation of the second GeMM with the first GeMM execution.

2) *Attention Results*: Figure 6b shows that synchronizing all kernels of Attention using `cuSync` provides 6-16% improvement over StreamSync for both GPT-3 and LLaMA.

During prompt processing, i.e. when $S' = 0$, StridedTileSync+WRT works better than both RowSync+WRT and TileSync+WRT because StridedTileSync+WRT performs less number of synchronizations than TileSync and provides larger overlapping opportunities than RowSync. During token generation, i.e. when $S' = 1$ and $S = 1$, all policies works similarly because different synchronization policies provides best performance between different kernels.

F. Computer Vision Model Inference Results

We now evaluate the decrease in inference times of Resnet-38 and VGG-19 by synchronizing all Conv2D kernels of each layer of both models using `cuSync` (Table II). We used `cuSyncGen` to generate the following policies:

RowSync+WRT synchronizes rows and execute thread blocks in a row major order with our optimizations in Section IV-C, i.e., apply avoid wait-kernel (W), avoiding custom tile ordering

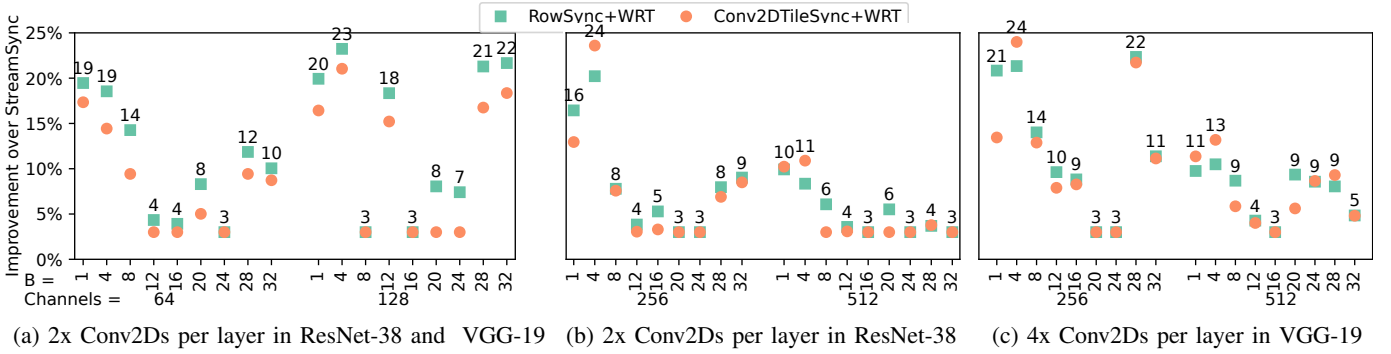


Fig. 7: Performance improvement of `cuSync` policies for all Conv2D kernels of each layer over `StreamSync` in ResNet-38 and VGG-19 for different batch sizes.

(T), and reordering tile loads optimizations (R).

Conv2DTileSync synchronizes tiles of Conv2Ds and execute thread blocks in a row major order.

Conv2DTileSync+WRT extends `Conv2DTileSync` with our optimizations in Section IV-C.

Figure 7b shows that synchronizing all Conv2D kernels of each layer of ResNet-38 and VGG-19 using `cuSync` provides up to 24% improvement over `StreamSync` for different channels and batch sizes. For each channel, the improvement follows an oscillating behavior with increasing batch size, i.e., increases to a local maximum then decreases to a local minimum and finally increases to another local maximum. For example, for 128 channels, the improvement increases from 20% at batch size 1 to 24% at batch 4 and then decreases to 3% at batch size 8, while increasing again to 18% at batch size 12 and then again decreases to 3% at batch size 16. This oscillating behavior is due to the fact that increasing batch size increases invoked number of thread blocks leading to the oscillating behavior of fraction of waves reduced by `cuSync`.

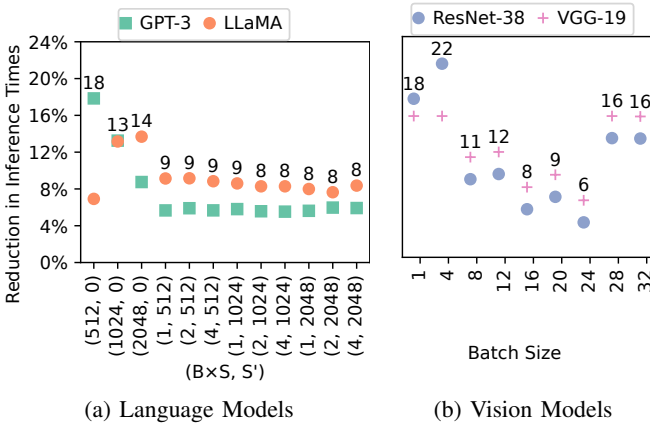


Fig. 8: Reduction in end-to-end inference times of using `cuSync`.

G. End-to-End Inference of ML Models

We integrated `cuSync` synchronized CUDA kernels in all four ML models and then evaluate the improvement in end-to-

end inference times of these models. Figure 8 shows that using `cuSync` synchronized kernels decreases the inference times of GPT-3 by 6–15%, LLaMA by 9–13%, of ResNet-38 by 5–22%, and VGG-19 by 6–16%. Hence, `cuSync` significantly reduces the inference times of popular ML models.

H. Comparison with Stream-K

We also evaluated the performance of `cuSync` against `Stream-K` for GeMMs kernels. The best policy of `cuSync` performs up to 15% better than `Stream-K` in GPT-3 and LLaMA (Figure 6). The speedup of `cuSync` over `Stream-K` is because `Stream-K` divides the GeMM workload into two kernel calls. The first kernel computes GeMM using the traditional tiled approach for full waves while the second kernel partitions workload of the final wave among all SMs. This design requires multiple memory accesses while `cuSync` performs a single atomic add to post the status of a computed tile and a read to wait on the status of a producer tile. Moreover, it is not straightforward to apply the idea of `Stream-K` to all tile-based kernels. Currently, `Stream-K` only supports GeMM computations in NVIDIA CUTLASS. This is why we cannot apply `Stream-K` to Conv2D, while `cuSync` is valid for any tile based kernels.

I. Impact of Optimizations

We now discuss the performance improvements provided by the optimizations on top of `TileSync` for ResNet-38 and GPT-3’s MLP. Table Va shows that applying all optimizations decreases execution times for kernels with low thread blocks.

VI. RELATED WORK

Several works have focussed on efficient software-based synchronization between threads of the same CUDA kernel for irregular applications [9], [16], [17]. Li et. al. [9] developed an approach for inter-thread synchronizations by reassembling the micro-instructions of shared memory atomic operations in an efficient manner. Kai et. al. [16] presented a hierarchical synchronization approach for irregular applications by synchronizing thread blocks using global memory and threads of a thread block using shared memory. Xu et. al. [17] present a

TABLE V: EXECUTION TIMES OF TILESYNC WITH OPTIMIZATIONS IN GPT-3’S MLP AND CONV2DTILESYNC IN RESNET FOR SMALLER GRID SIZES. +W AVOIDS THE WAIT-KERNEL. +WR ALSO REORDERS THE TILE LOADING. +WRT ALSO AVOIDS CUSTOM TILE ORDER.

(a) EXECUTION TIMES IN μ s OF TILESYNC OF GEMM KERNELS IN GPT-3’S MLP WITH AND WITHOUT OPTIMIZATIONS FOR DIFFERENT BATCH SIZES.

B	TileSync			
	Vanilla	+R	+WR	+WRT
1-64	378	365	360	355

(b) EXECUTION TIMES μ s OF CONV2DTILESYNC OF RESNET-38’S CONV2D WITH AND WITHOUT OPTIMIZATIONS FOR ALL CHANNELS AND SMALL BATCH SIZES.

C	B	Conv2DTileSync			
		Vanilla	+R	+WR	+WRT
64	1	50	45	41	37
128	1	60	56	50	45
256	1	65	61	56	51
512	1	100	94	89	85
	4	135	128	120	115

lock design that uses lock stealing to avoid deadlocks. COCONET [8] performs synchronization between computation and communication kernel to overlap the communication transfers with the computation. cuSync targets synchronization between threads of multiple CUDA kernels and provide abstraction to easily design several synchronization policies, both of these are missing from above mentioned works. Moreover, some works have focussed on hardware-supported synchronization primitives for inter-kernel threads. GLocks [2] is the first hardware supported implementation for highly-contented locks using message passing. HQL [18] is a hardware-accelerated fine-grained lock scheme for GPUs, which adds support for queuing locks in L1 and L2 caches and uses a customized communication protocol for faster lock transfer and reduced lock retries. ElTantawy et. al. [5] propose a hardware warp scheduling policy that reduces lock retries by de-prioritizing warps whose threads are waiting in their spin lock. They also propose a hardware mechanism for accurately detecting busy-wait synchronization on GPUs. Dalmia et. al. [3] designed multi-level barrier and priority mechanisms for semaphores for GPU based synchronization primitives. cuSync is a software solution for synchronizing threads of multiple CUDA kernels and these hardware-supported mechanisms are complementary to cuSync.

Lingqi et. al. [19] studied the performance and pitfalls of several CUDA synchronization methods for reduction operations. Sinclair et. al. [14] presented a benchmark suite to measure the performance of synchronization primitives for different coherence protocols and consistency models.

Stream-K [10] is a GeMM implementation that improves the utilization of SMs of a GPU by dividing the workload among all SMs. However, Stream-K is not straightforward to apply to computations other than GeMMs. In contrast, cuSync fits thread blocks of multiple kernels in each wave and is applicable to any tile based computations.

VII. CONCLUSION

State-of-the-art ML models consist of thousands of individual computations that are executed on one or more GPUs. However, these models under-utilize the GPUs because individually each of these computations cannot completely utilize a GPU and these models largely consists of dependent computations. In this paper, we presented cuSync, a framework for fine-grained synchronization of tiles of dependent computations. By synchronizing only, the dependent tiles, our framework allows concurrent execution of independent tiles, thus improving the utilization of GPU. Our experiments show that synchronizing computations of existing machine learning models using cuSync can reduce inference times of these models significantly.

APPENDIX

The artifact [7] contains cuSync CUDA implementation and scripts to reproduce all of our results. The artifact provides both a Dockerfile, which contains all prerequisites installed, and source code. Latest source code is available at <https://github.com/microsoft/cusync>. The artifact reproduces Figure 6, 7, and 8 in Section V.

System We executed our experiments on a NVIDIA DGX-2 system containing 8 NVIDIA Tesla V100 GPUs connected using NVLINK. Our experiments will run on any system with a GPU, however, the end-to-end inference results in Figure 8 might not be reproducible on another system.

Extract Artifact Download the artifact from [7] and extract the zip file.

```
unzip cusync-cgo-24.zip
cd cusync-cgo-24
```

A. Docker Container

To run artifact inside a Docker container follow these steps:

Install docker Install docker engine by following steps on <https://docs.docker.com/engine/install/ubuntu/>.

Install NVIDIA Container Toolkit Install NVIDIA Container Toolkit by following steps on <https://docs.nvidia.com/datacenter/cloud-native/container-toolkit/latest/install-guide.html>.

Create Container Create docker container using the Dockerfile, start the container, and cd to the directory:

```
docker build -t cusync-cgo-24 .
docker run -it --gpus all cusync-cgo-24
cd /cusync
```

Check PyTorch and CUDA Install: Check if torch supports CUDA if `torch.cuda.is_available()` returns True:

```
python
>>> import torch
>>> torch.cuda.is_available()
True
```

B. Running Natively

We can also run code natively, which requires installing all dependencies. These steps can be ignored if using docker in above steps.

Linux Installation We recommend using Ubuntu 22.04 as the Linux OS. We have not tested our artifact with any other OS but we believe Ubuntu 20.04 and 23.04 should also work.

Install Dependencies Execute following commands to install dependencies.

```
sudo apt update
sudo apt install gcc linux-headers-$(uname -r) \
make g++ git python3 wget \
unzip python3-pip build-essential cmake
```

Install CUDA We need to install CUDA before proceeding further. In our experiments we used CUDA 12.2 on Ubuntu

22.04. CUDA 12.2 toolkit can be downloaded from <https://developer.nvidia.com/cuda-12-1-0-download-archive>. After installing CUDA, set `nvcc` and CUDA paths.

```
export PATH="/usr/local/cuda/bin:$PATH"
export LD_LIBRARY_PATH=
"/usr/local/cuda/lib64:$LD_LIBRARY_PATH"
```

Check CUDA Installation To check CUDA installation, run `nvidia-smi` and it should print all GPUs in the system. Otherwise there is a problem with the CUDA installation.

Install Pytorch: Install PyTorch using pip.

```
sudo pip3 install torch torchvision torchaudio
```

Check Pytorch CUDA Install: Check if torch supports CUDA if `torch.cuda.is_available()` returns True:

```
python
>>> import torch
>>> torch.cuda.is_available()
True
```

Obtain source code The source code can be downloaded from [7]. Latest source code is available from cuSync repository and CGO AE branch:

```
git clone --recurse-submodules \
https://github.com/microsoft/cusync
cd cusync
git checkout cgo-24-ae
```

C. Functionality and Reusability

The `README.md` contains instructions of how code can be compiled to other NVIDIA GPU architectures, an example and test cases. The functionality can be checked by executing these test cases. To run tests execute:

```
make tests -j
```

If all tests passes then we are ready for reproducing results.

D. Reproduce Results

We will now reproduce our main results of Figure 6, 7b, and 8. All commands should be executed in the `cusync` directory.

Large Language Model Inference Results [Time 60 mins] Following commands will run all experiments to gather the results

```
cd src/ml-bench/volta_transformer
python3 eval_llm.py mlp gpt3
python3 eval_llm.py attention gpt3
python3 eval_llm.py mlp llama
python3 eval_llm.py attention llama
python3 allreduce_times.py
```

Computer Vision Inference Results [Time 60 mins] Following commands will run all experiments to gather results for Figure 7b.

```
cd src/ml-bench/volta_conv2d
python3 eval_conv.py resnet
python3 eval_conv.py vgg
```

Generate Plots [Time 5 mins] Generate all Figures by running below commands:

```
cd src/ml-bench/plots
make -j
```

The current directory will have figures as PDF and they can be checked against figures in the paper.

REFERENCES

- [1] NVIDIA cuTLASS: CUDA Templates for Linear Algebra Subroutines. <https://github.com/NVIDIA/cutlass>, Accessed: 2023-07-30.
- [2] Jose L. Abellán, Juan Fernández, and Manuel E. Acacio. GLocks: Efficient Support for Highly-Contended Locks in Many-Core CMPs. In *Proceedings of the 2011 IEEE International Parallel & Distributed Processing Symposium*, IPDPS '11, pages 893–905, USA, 2011. IEEE Computer Society.
- [3] Preyesh Dalmia, Rohan Mahapatra, Jeremy Intan, Dan Negrut, and Matthew D. Sinclair. Improving the Scalability of GPU Synchronization Primitives. *IEEE Transactions on Parallel and Distributed Systems*, 34(1):275–290, 2023.
- [4] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In *Advances in Neural Information Processing Systems*, 2022.
- [5] Ahmed ElTantawy and Tor M. Aamodt. Warp Scheduling for Fine-Grained Synchronization. In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 375–388, 2018.
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385, 2015.
- [7] Abhinav Jangda. (Artifact) A Framework for Fine-Grained Synchronization of Dependent GPU Kernels. 12 2023.
- [8] Abhinav Jangda, Jun Huang, Guodong Liu, Amir Hossein Nodehi Sabet, Saeed Maleki, Youshan Miao, Madanlal Musuvathi, Todd Mytkowicz, and Olli Saarikivi. Breaking the Computation and Communication Abstraction Barrier in Distributed Machine Learning Workloads. In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '22, pages 402–416, New York, NY, USA, 2022. Association for Computing Machinery.
- [9] Ang Li, Gert-Jan van den Braak, Henk Corporaal, and Akash Kumar. Fine-Grained Synchronizations and Dataflow Programming on GPUs. In *Proceedings of the 29th ACM on International Conference on Supercomputing*, ICS '15, pages 109–118, New York, NY, USA, 2015. Association for Computing Machinery.
- [10] Muhammad Osama, Duane Merrill, Cris Cecka, Michael Garland, and John D. Owens. Stream-K: Work-Centric Parallel Decomposition for Dense Matrix-Matrix Multiplication on the GPU. In *Proceedings of the 28th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming*, PPOPP '23, page 429–431, New York, NY, USA, 2023. Association for Computing Machinery.

- [11] Noam Shazeer. GLU Variants Improve Transformer, 2020.
- [12] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism, 2020.
- [13] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, 2015.
- [14] Matthew D. Sinclair, Johnathan Alsop, and Sarita V. Adve. HeteroSync: A benchmark suite for fine-grained synchronization on tightly coupled GPUs. In *2017 IEEE International Symposium on Workload Characterization (IISWC)*, pages 239–249, 2017.
- [15] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, 2023.
- [16] Kai Wang, Don Fussell, and Calvin Lin. Fast Fine-Grained Global Synchronization on GPUs. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS '19*, pages 793–806, New York, NY, USA, 2019. Association for Computing Machinery.
- [17] Yunlong Xu, Lan Gao, Rui Wang, Zhongzhi Luan, Weiguo Wu, and Depei Qian. Lock-Based Synchronization for GPU Architectures. In *Proceedings of the ACM International Conference on Computing Frontiers, CF '16*, page 205–213, New York, NY, USA, 2016. Association for Computing Machinery.
- [18] Ayse Yilmazer and David Kaeli. HQL: A Scalable Synchronization Mechanism for GPUs. In *Proceedings of the 2013 IEEE 27th International Symposium on Parallel and Distributed Processing, IPDPS '13*, pages 475–486, USA, 2013. IEEE Computer Society.
- [19] Lingqi Zhang, Mohamed Wahib, Haoyu Zhang, and Satoshi Matsuoka. A Study of Single and Multi-device Synchronization Methods in Nvidia GPUs. In *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 483–493, 2020.