Title
Harmonic CUDA: Asynchronous Programming on GPUs

Author
Wapman, Jonathan Derek

Publication Date
2022

Peer reviewed|Thesis/dissertation

Harmonic CUDA: Asynchronous Programming on GPUs

By

JONATHAN WAPMAN

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Electrical and Computer Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

John D. Owens, Chair

Jason Lowe-Power

Venkatesh Akella

Committee in Charge

2023

To my family, girlfriend, friends, and everyone who helped me along the way...

CONTENTS

List of Figures

List of Tables

ABSTRACT

Harmonic CUDA: Asynchronous Programming on GPUs

We introduce Harmonic CUDA, a dataflow programming model for GPUs that allows programmers to describe algorithms as a dependency graph of producers and consumers where data flows continuously through the graph for the duration of the kernel. This makes it easier for programmers to exploit asynchrony, warp specialization, and hardware acceleration. Using Harmonic CUDA, we implement two example applications: Matrix Multiplication and GraphSage. The matrix multiplication kernel demonstrates how a key kernel can break down into more granular building blocks, with results that show a geomean average of 80% of cuBLAS performance, and up to 92% when omitting small matrices, as well as an analysis of how to improve performance in the future. GraphSage shows how asynchrony and warp specialization can provide significant performance improvements by reusing the same building blocks as the matrix multiplication kernel. We show performance improvements of 34% by changing to a warp-specialized version compared to a bulk-synchronous implementation. This thesis evaluates the strengths and weaknesses of Harmonic CUDA based on these test cases and suggests future work to improve the programming model.

ACKNOWLEDGMENTS

Graduate school is about the journey as well as the destination. In my journey at UC Davis I've had the pleasure of working with some of the smartest, kindest, and most supportive people in the world.

First, thank you to Professor John Owens. I truly appreciate your support throughout my time at UC Davis. I've learned so much from you, both in and out of the classroom. I'm grateful for your guidance and mentorship. I'm also grateful for your patience and understanding as I've navigated the ups and downs of graduate school, and I'm excited to see what the future holds for you and your research. I hope we'll get more opportunities to work together in the future.

Thank you to my wonderful parents who supported me throughout graduate school. I was able to pursue this path in large part because I knew I had them on my team, encouraging me to pursue my passions both at school and elsewhere. And of course, Emily, Anna, Skye, Toby, and Lily.

I would especially like to thank Sean Treichler and Aamer Jaleel. Their mentorship and advice throughout my graduate school career have been invaluable. I was lucky to work directly for Sean twice as an intern at NVIDIA Research, and Aamer and Sean spent considerable time with me to provide feedback during weekly meetings even when I was no longer at NVIDIA. Also at NVIDIA, I appreciate the guidance from Steve Keckler, who took me on as an intern in the Architecture Research Group at NVIDIA, Michael Garland, who lead the Symphony project I was lucky to be part of and gave me useful feedback on Harmonic CUDA in particular. I also greatly enjoyed working with Jason Clemons, Po-An Tsai, Donghyuk Lee, Duane Merril, Stephen Jones, and many others at NVIDIA.

Thank you very much to Serban Porumbescu, who spent many hours helping me edit my thesis to clarify my ideas and find the right way to express them. Your page cutting skills are legendary.

I would also like to thank all of my labmates, Muhammad Osama, Toluwa Odemuyiwa, Agnieszka Lupinska, Chuck Rozhon, Daniel Loran, Yuxin Chen, Afton Geil, Jason Mak, Vehbi Esref Bayraktar, Collin McCarthy, Ahmed Mahmoud, Zhongyi Lin, Rod Shojaei, Marjorie Suresh, Mythreya Kuricheti, Annie Robinson, Chenfei Yao, Muhammad Awad, Kerry Seitz, and our postdocs Serban Porumbescu and Matthew Drescher. Whether it was talks in the lab, Lazi Cow game nights, help answering difficult technical questions, or chats at our weekly group meetings, I've appreciated getting to know all of you and I hope we have a chance to work together again in the future.

Thank you also to everyone at UC Davis who supported me, including my committee members, Professor Jason Lowe-Power and Professor Venkatesh Akella. I've learned a lot from your classes and research feedback. Also to the incredibly helpful administrative staff, especially Sacksith Ekkaphanh and Michelle Walker.

Thank you to Carl Yang and Kevin Poulet, my managers and mentors at Waymo, who gave me the opportunity to work on a really interesting project and helped me grow as a software engineer. Waymo was an important experience for me, and I'm grateful for the opportunity to work with you both. Also to David Sternberg, Laura Jones-Wilson, Ryan Goldhahn, Professor Diego Yankelevich, and everyone who helped me prepare for graduate school.

Thank you to all my friends who supported me, particularly my former roommate Benjamin Nilsen, my friend Kyle Westphal, and all of my friends from undergrad in Sacramento and elsewhere (you know who you are).

Finally, thank you to Johanna Fröhling, my wonderful girlfriend of almost 3 years. Doing graduate school during a pandemic isn't an easy challenge, and I appreciate your continuous support throughout graduate school. I'll always remember the amazing experience of spending 6 months living in Germany during graduate school, and hearing you talk about "buffets" [29] as if they're the best part of graduate school (you're not wrong [15]) always makes me laugh.

Chapter 1

Introduction

1.1 Background and Motivation

Modern GPUs are more than just a group of thread processors. Over time, GPUs have added specialized hardware units, such as Tensor Cores, of many varieties for machine learning and dense linear algebra, Ray Tracing cores for realistic rendering, Direct Memory Access (DMA), and Tensor Memory Accelerator (TMA) units for asynchronous memory movements between off-chip and on-chip memory, and Transformer Engines for machine learning, with potentially more to come [6, 8, 19, 25]. However, taking advantage of these specialized units typically requires major code rewrites, and orchestrating data movement in a performant way often requires skilled CUDA programming ability. The need to rewrite software for each new GPU architecture is a major barrier to the adoption of new hardware features. NVIDIA may be able to allocate these resources to commonly used and performance-critical libraries that automatically bring new hardware features to end-users, but this does not apply to custom code written by the users themselves. At the same time, GPU programming models have become increasingly asynchronous (Section 2.1) and it is desirable to overlap computation with memory transfers or differing types of computation depending on available accelerator features.

At a higher level of abstraction, NVIDIA and third parties provide many libraries that implement critical GPU kernels such as matrix multiplication [18, 27], block-wide or device-wide collective operations such as prefix sums or reductions [13, 21], or more complex algorithms such as graph algorithms [10]. These libraries provide highly optimized implementations of these kernels, but they lack flexibility. For example, a library such as CUTLASS [18] provides a matrix multiplication kernel that utilizes the entire device and building blocks for implementing custom device-wide kernels but does not support the warp-centric configuration necessary for an application like GraphSage (Chapter 5). This inflexibility makes it difficult to take advantage of the highly optimized kernels provided by these libraries, and instead, forces programmers to write their own kernels from scratch. This is especially difficult for programmers who are not experts in GPU programming, and who do not have the time or resources to optimize their kernels. Even if NVIDIA were to update their libraries to support more flexible building blocks, the user would still have to manage the connections between the building blocks (such as double buffering or pipeline synchronization) and would need to explicitly manage which individual building blocks to use depending on the desired compute location.

Additionally, CUDA libraries typically do not have the flexibility to perform an operation at any level of the GPU's compute hierarchy. A GPU programmer may want to perform an operation with a single thread, warp, or block, a subset of a block, an entire grid, or a subset of a grid. However, CUDA is not well-suited to elegantly expressing operations where the core *algorithm* is the same even though the *location* of the data and the assigned compute group may change.

1.2 Harmonic CUDA

To solve these challenges, we present *Harmonic CUDA*: a programming model for asynchronous producer/consumer computation on modern GPUs that enables programmers to *describe* the dataflow of their code, while relying on highly optimized backends to handle scheduling, synchronization, hardware acceleration, and storage management. The primary goals of Harmonic CUDA are:

- **Computation/Location Abstraction:** Programmers should be able to express the *what* of their computations without worrying about the *where* or *when*. This allows programmers to focus on the algorithmic aspects of their code, and to rapidly experiment with different mappings of computation to hardware.

- **Performance:** Programmers should be able to rely on Harmonic CUDA to use best-available implementations for its backend, which may include highly optimized software libraries, architecture-specific accelerators such as Tensor Cores, or future hardware and software features as they become available. Harmonic CUDA's *Connector* abstraction provides abstractions around optimizations such as double buffering, storage synchronization and management, and backend specialization based on the location of the data buffer. Programmers should also be able to rely on an automatic Node scheduler to give acceptable performance in a wide variety of situations. For more complex situations, programmers should be able to use manual scheduling to achieve the best performance. In some cases, a Dataflow may be able to identify a common pattern of multiple Nodes in the user's program and optimize this into a more performant Dataflow.

- **Composability:** The programmer should be able to construct a Dataflow and treat that Dataflow as an individual Node within other Dataflows. This enables programmers to use third-party Dataflows without needing to understand the implementation of those Dataflows, or to reuse their own Dataflows as a building block within a larger Dataflow.

- **Programmability:** Harmonic CUDA should be an intuitive programming model that makes it *easier* for a programmer to think about their code in terms of data flow, and to write code that is easy to understand and maintain. This includes a simple, easy-to-learn programming model and API.

- **Harmony:** The programmer should be able to use Harmonic CUDA *alongside* traditional CUDA code. This allows the programmer to write new features using Harmonic CUDA without rewriting existing code, or to identify sections of existing code that would be more performant or expressive using Harmonic CUDA. In contrast, other programming models require programmers to rewrite their entire codebases to use the new programming model.

1.3 Main Contributions

This thesis presents the following primary contributions:

- Harmonic CUDA, a dataflow programming model for asynchronous producer/consumer GPU computing (Chapter 3).

- An implementation of a memory copy kernel in Harmonic CUDA that shows the Harmonic CUDA API and backend implementation (Chapter 3), and implementations of matrix multiplication (Chapter 4) and GraphSage (Chapter 5) that demonstrate Harmonic CUDA's performance, the benefits of warp specialization, and the benefits of the reuse of building blocks.

- An evaluation of the strengths and weakness of the Harmonic CUDA programming model (Chapter 6.1).

- Analysis of future research directions and applications of Harmonic CUDA (Chapter 6).

Chapter 2

Related Works

2.1 Hardware Asynchrony

In recent years, GPUs have added more asynchronous functionality. NVIDIA's Ampere architecture includes asynchronous Direct Memory Access (DMA) units that use dedicated hardware units to copy sequential regions of memory directly from global memory to shared memory without the need for intermediate copies to thread registers [19]. NVIDIA's Hopper architecture extends this concept to a Tensor Memory Access unit, which performs the same function but for 2-dimensional tiles of a matrix [11, 25]. Additionally, there is a large body of research into domain-specific accelerators [14], which feature asynchronously running hardware units connected with intermediate buffers [29]. Harmonic CUDA addresses the programming challenge of efficiently targeting an increasing number of asynchronous by creating abstractions around the actual implementation of logical operations and the data movements between them.

2.2 GPU Software Asynchrony

CudaDMA proposed an approach to divide a block on the GPU into DMA warps for performing memory transfers, and computation warps for performing any necessary computation [2]. CudaDMA provides robust evidence that warp specialization is a powerful strategy on GPUs, and that it can improve the performance of a wide range of applications. CudaDMA primarily provides the infrastructure for assigning threads to a warp group, and for synchronizing between groups. However, we view CudaDMA's principles as a key building block, but not a complete asynchronous producer/consumer programming model.

Building on the concepts of CudaDMA, Warp Specialization is a well-studied method of programming GPUs that researchers have used to improve the performance of a wide range of applications including graph analytics, physics simulations, and combinatorial optimization [2–4, 17, 22, 31]. To use warp specialization, a programmer defines separate paths in their kernel that a warp can take. The warp then chooses one of these paths depending on some condition. For example, the programmer may assign a warp to a memory movement path if the warp is responsible for moving data from global memory to shared memory, and to a computation path if the warp is responsible for performing a computation on the data in shared memory. All warps within the same block may continue to communicate with each other over shared memory and may take advantage of efficient synchronization mechanisms. Warp specialization can improve performance by increasing the size of a working set, by performing multiple independent computations in parallel, or by reducing memory divergence [3]. While warp specialization can often significantly improve the performance of a kernel, in practice it is not a commonly used programming paradigm in part due to implementation difficulties. Even when programmers use Warp Specialization, they tend to focus on a limited set of configurations, such as dividing a block in half into memory access and computation warps. Harmonic CUDA provides a more general approach to warp specialization that abstracts away the low-level details of warp specialization and allows programmers to use more flexible specialization configurations.

Libcu++ [28] provides abstractions such as pipelines and barriers to manage asynchronous GPU hardware. Harmonic CUDA leverages these primitives as building blocks to create a higher-level programming abstraction.

Several works [1, 16] aim to hide complexities of asynchronous GPU programming (such as synchronization, Direct Memory Accesses, and memory management) using a producer-consumer model, but these focus on kernel-level dataflows between the CPU and one or more GPUs. In contrast, Harmonic CUDA provides a programming model usable *within* a GPU kernel at runtime. Additionally, while Harmonic CUDA may target heterogeneous systems for future work, the programming model itself is fundamentally different, since it treats the CPU or multiple GPUs as just another user-specified compute location.

2.3 CPU Software Asynchrony

LabView and Simulink are node-based graphical dataflow programming models commonly used on the CPU [20, 23]. LabView and Simulink both provide libraries of Nodes that a programmer connects together in a dataflow to perform some computation, with typical examples being signal processing or control loops. Harmonic CUDA differs in that it is a text-based programming model that additionally provides for the concept of "where" to store and/or perform computation.

StreamIt [32] is a text-based DSL for programming streaming applications, but this model focuses primarily on signal processing, rather than on general-purpose parallel computing. It also does not include any concept of "where" to store and/or perform computation beyond offering support for CPU multi-threading.

In the Senders programming model [9], a "sender" object contains a computation that runs asynchronously. The sender returns a single item (such as an int, pointer, struct, etc.) before a dependent sender begins. Although it is an asynchronous programming model, the Senders model is not a true producer/consumer dataflow pipeline where data flows continuously. Harmonic CUDA is compatible with the Senders model. For example, programmers can use Harmonic CUDA to more easily write a CUDA kernel that they can then launch as the computation of a Sender.

There are many examples of parallel programming and dataflow programming languages on both the GPU and the CPU [7]. In particular, Halide [30] similarly separates the logical computation from a schedule, but is not a dataflow programming model.

Chapter 3

Programming Model

3.1 Motivation

Harmonic CUDA is a node-based asynchronous programming model for expressing computation as a dataflow of producers and consumers with the timing and physical location of computations abstracted away. We design the programming model to be productive for programmers, performant, and support software composability. We also design the programming model to take advantage of improvements in future architectures without changing code, and include automatic and manual scheduling frameworks that provide good defaults for the average programmer and extended customizability for the expert programmer.

3.1.0.1 Overview

Figure 3.1 shows an example of a basic memory copy operation implemented as a Harmonic CUDA Dataflow using two instances of data transformation (a producer and a consumer) and an intermediate buffer. We include the full code for this example in Listing 1. The basic unit of Harmonic CUDA is a *Node*, which is an abstraction of a program (Section 3.2). A Node embodies tasks such as "perform a prefix sum on the elements given an input," "perform elementwise additions on the two inputs," or "do a matrix multiplication given two input matrices." A Node may run on a thread, a block, a subset of a block, an entire grid, or a subset of a grid. It may have a software backend or a hardware-accelerated backend, and may coexist on the same hardware with many other Nodes. Using this abstraction, the Node defines *what* the user wants to do while giving the flexibility to easily change *where* the computation runs, and also gives

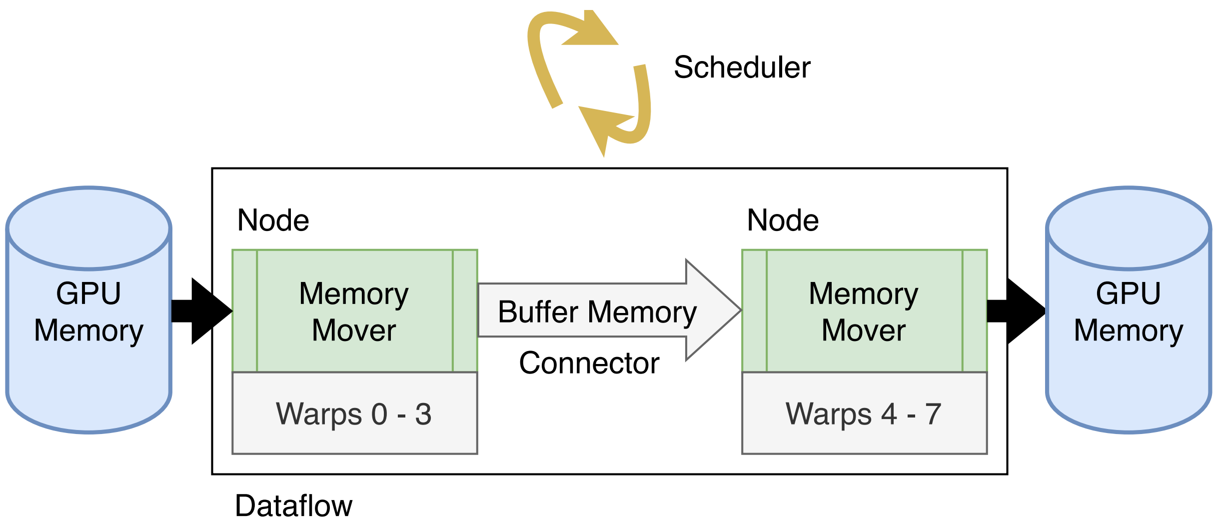

Figure 3.1. A Harmonic CUDA Memory Copy Dataflow. The producer and consumer map to an arbitrary compute location, and the buffers map to an arbitrary storage location. The scheduler takes advantage of hardware asynchrony while respecting resource limitations.

the system the power to decide *when* it happens.

Connectors join Nodes and capture synchronization and storage management (Section 3.3). The user provides Connectors with input and output Nodes as well as the intermediate data storage location, while the Nodes and Connectors choose highly optimized backends based on the functionality of the Nodes, the size of the data chunks transferred between Nodes, the locations of the Nodes, and the locations of the intermediate data storage.

A *Dataflow* collects Nodes and Connectors into a single unit of computation and assists in scheduling of Nodes onto hardware (Section 3.4). Furthermore, Dataflows are composable—that is, they themselves can be used as a Node within a larger Dataflow.

Finally, Harmonic CUDA provides an automatic Node scheduler that is able to efficiently schedule Nodes onto hardware for many common cases, and APIs that the programmer can use to manually schedule Nodes onto hardware for more complex cases (Section 3.6).

To enable Harmonic CUDA to coexist *alongside* traditional CUDA, threads can interact with aforementioned components. For example, Harmonic CUDA could provide Nodes whose function is to read data from a thread-provided location, or to fill data into a location where other GPU threads can consume the data directly.

```cpp
__global__ memcpy_kernel(int *global_data, int size) {
  __shared__ int shared_data[BUFFER_SIZE];

  // Instantiate the nodes.
  auto Global2Shared =
      make_MemMover(BlockGroup, LocationGlobal, LocationShared, global_data,
                    shared_data, size * sizeof(int));

  auto Shared2Global =
      make_MemMover(BlockGroup, LocationShared, LocationGlobal, shared_data,
                    global_data, size * sizeof(int));

  // Abstracts pipelining and storage management.
  __shared__ ConnectorSharedState</*num_pipeline_stages=*/ 2>
    memcpy_connector_shared_state;
  auto memcpy_connector = make_connector(Global2Shared, Shared2Global,
                                         &memcpy_connector_shared_state);

  // Incrementally build the dataflow.
  auto BaseDF = make_dataflow();
  auto [DF_Global2Shared, ID_Global2Shared] = BaseDF.add_node(Global2Shared);
  auto [DF_Shared2Global, ID_Shared2Global] =
      DF_Global2Shared.add_node(Shared2Global);

  auto [DF_Connected, ID_Connected] =
      DF_Shared2Global.add_connection(memcpy_connector);

  // Execute the dataflow using the automatic scheduler.
  DF_Connected.progress();
}
```

Listing 1: Memory Copy Example in Harmonic CUDA with Automatic Scheduling. The user instantiates all Nodes and Connectors, builds the dataflow, and calls the Dataflow's `progress()` method.

3.1.0.2 Workflow

Programmers should use Harmonic CUDA in a three-step process. First, they should conceptualize their algorithm as a dependency graph of producers and consumers without considering the physical mapping of computations or storage locations to the GPU. Next, they must instantiate the dataflow in code and decide where to map the Nodes and Connectors to the GPU. Here, simply supplying the correct flags is enough for the system to take care of the backend. Finally, they should profile their kernel and adjust the mapping based on results, such as separating two Nodes to make them warp-specialized or using shared memory instead of registers.

3.2 Nodes

3.2.1 Specification

A Node is an abstraction of data transformation and movement that specifies *what* computation is being done. *Where* and *when* are instead specified by a programmer-provided compute location parameter and by Harmonic CUDA's scheduler. Nodes utilize NVIDIA's "Cooperative Groups" API [13] to specify which hardware unit(s) to use. The actual implementation of a Node, or its *Backend*, may be highly optimized third-party CUDA backends, hardware accelerators, or custom code written by the end user that builds on top of the Node framework. The compiler can often determine the specific backend of a Node at compile time based on the locations of the input and output data of the Node, the compute location the Node runs on, and the hardware capabilities of the GPU architecture. Other factors, such as runtime flags, may also determine the backend of a Node.

To support the programmability goal, Harmonic CUDA provides Node templates that programmers can customize to their needs, either by building on top of a template or by providing a concise lambda function that implements the Node's behavior. This allows programmers to easily create Nodes that are highly optimized for their specific use case, while still utilizing the high-level programming model.

The following sections describe several "perspectives" of Nodes: what the programmer cares about, what the backend implementer needs to know, and what the scheduler needs.

3.2.2 Perspectives

3.2.2.1 Node User

A Node user must understand at a high level what a Node does and what parameters to change to control its behavior. Given a Node, the programmer is responsible for assigning the Node to a Compute Location based on an algorithmic design decision. The programmer is also responsible for providing a Node with the correct inputs and outputs at all relevant ports. To run the Node, the programmer should provide the desired batch size or total amount of data. This can either be at runtime or compile time. Each Node includes a `progress()` method, which instructs the Node to make forward progress given its available input data and output space. The programmer must understand what action a given Node's `progress()` method takes, and

```
__global__ memcpy_kernel(int *global_data, int size) {
  __shared__ int shared_data[...];
  auto Global2Shared = make_MemMover(/*compute_group=*/GridGroup,
                                     /*input_location=*/LocationGlobal,
                                     /*output_location=*/LocationShared,
                                     /*input_data=*/global_data,
                                     /*output_data=*/shared_data,
                                     /*size=*/size * sizeof(int));
  auto Shared2Global = make_MemMover(/*...*/);
  // ...
}
```

Listing 2: Memory Copy Node Instantiation. The user provides the desired compute location (`GridGroup` for the entire device) alongside the locations of the input and output data and the size of the data.

must also understand what a Node requires to be "ready" or "finished."

In practice, Nodes have many compile-time template parameters that make it overly verbose to instantiate a Node using the form `Node<float, 5, int, ...>` where there may be any number of template parameters that describe types, buffer sizes, compute location, and other information that a Node requires for instantiation. To make it easier to instantiate a Node, each Node provides a helper such as `make_MemMover(...)` that takes in a set of parameters and returns a Node with the correct template parameters. This helper is responsible for determining the correct template parameters to use based on the parameters provided by the user. The `auto` keyword hides the template parameter complexity from the programmer.

Listing 2 shows an example of how to instantiate two "MemMover" Nodes for a basic Memory Copy example. Both Nodes map to a `GridGroup` and internally calculate indices relative to other blocks. Alternatively, the user could assign each Node to a `BlockGroup` and use Harmonic CUDA helpers to calculate global memory pointers and copy sizes per-block. The `LocationGlobal` and `LocationShared` parameters allow Harmonic CUDA to make compile-time optimizations. Variations of `make_MemMover` could provide greater runtime flexibility but higher runtime cost by omitting the need to specify input and output data locations at compile time.

3.2.2.2 Node Implementer

It is the Node implementer's responsibility to provide the functionality of the Node given the input and output locations of the data, the target GPU architecture, the runtime flags of the Node, or the compute location the Node is mapped to. As part of the implementation, the Node must be able to query input and output Connectors to determine how much data is available and where to read from or write to. Internally, a Node must implement `is_finished()` and `is_ready()` methods that are for dynamic scheduling, along with any necessary internal state tracking based on the total amount of data or the batch size. Finally, Nodes must perform all data movement and transformation when calling the `progress()` method, including any internal state updates. It is possible to implement several variants of the `progress()` method, such as supplying pipelines or other synchronization variables as needed depending on the configuration of the overall Dataflow.

Listing 3 shows an example of the MemMover backend from Listing 2, where we demonstrate how for supported hardware such as Ampere GPUs, the Node can use the DMA engine for asynchronous memory copies. This reduces register usage and allows threads to do other work while copies complete.

3.2.2.3 Node Scheduler

From the scheduler's perspective, the functionality of a Node is irrelevant. The scheduler only needs to know if a Node is ready to run, if it is finished, and how much data it can process in a single call to `progress()`. If the scheduler knows the Node's batch size at compile time, it can generate a static schedule for the Node, which has much less overhead than a runtime scheduler. For the dynamic scheduler, `is_finished()` and `is_ready()` methods allow the scheduler to make runtime decisions about which Node to run next.

3.3 Connectors

3.3.1 Specification

Connectors are responsible for managing the intermediate buffer storage between Nodes. A Connector's backing storage may be global memory, a shared memory buffer, thread-local reg-

Table 3.1. Interactions with Node properties from multiple perspectives.

Property	**Node User**	**Node Implementer**	**Scheduler**
Node Functionality	What a Node does at a high level and what parameters to change to control its behavior	Provide highly-optimized functionality given input/output locations, architecture, etc.	Functionality irrelevant as long as the scheduler finds a valid sequencing of all Nodes
Worker and Thread assignment	Selects a group based on algorithmic design decisions	Must run the correct implementation given a supported cooperative group	Schedules Nodes to cooperative groups
Storage Management	Must provide Nodes with their input/output storage locations	Must support querying Connectors for data size and offset	N/A
Batch size / Total size	If the batch size(s) or total amount of data are known at compile time, provide this as template parameters	N/A	If sizes are known at compile time, generate a static Node schedule
Flow Control	User must understand what is required for a Node to be ready or finished	Node must implement `is_finished()` and `is_ready()` for dynamic scheduling. Requires internal state tracking.	Node must have both methods, but the contents are irrelevant.
`progress()`	User must understand what action an individual `progress()` call performs	Nodes must perform all data movement and transformation when using `progress()`, including internal state updates. Several variants of `progress()` interfaces possible, such as supplying pipelines or other synchronization variables	Call `progress()` for each Node and provide Cooperative Groups and synchronization variables as appropriate

```
__device__ progress(int buffer_size, PipelineT &pipe) {
#if __CUDA_ARCH__ >= 800
  // Special asynchronous copy for Ampere GPUs using DMA units.
  cuda::memcpy_async(
      this->get_compute_group(), this->output_base + this->output_offsett,
      this->input_base + this->input_offset, buffer_size * sizeof(int), pipe);
#else
  // For non-Ampere GPUs...
#endif
  internal_state_update(buffer_size);
}

__device__ bool is_finished() { return elements_copied >= max_copies; }
__device__ bool is_ready(/*...*/) {
  /* ... */

  return !(input_elems_avail == 0 || output_space_avail == 0 ||
      elements_copied >= max_copies);
}
```

Listing 3: Architecture-Specialized Node Backend. Since Ampere includes asynchronous DMA units, we target a special case for Ampere while still providing a generic implementation for other architectures.

isters, or other specialized locations such as Tensor Core registers. Connectors must manage synchronization between Nodes, handle optimizations such as double buffering, and assist the scheduler by negotiating the amount of data that a single call to a Node's `progress()` method processes. In many ways, a Connector is similar to a *Buffet* [29], in that it has a fixed capacity, both queue-like and array-like operations, and forms the basis of connections in a dataflow graph.

3.3.2 Perspectives

3.3.2.1 Connector User

The programmer must provide the Connector information about how much storage it has to manage, how many buffers it has, its batch size, and provide a shared state for the Connector. Some parameters may be specified at compile time for better optimization, and others at run time.

3.3.2.2 Connector Implementation

The Connector assists its input and output Nodes with synchronization, pipelining or barrier operations as instructed by the scheduler. It must internally encapsulate any necessary pipelines

Table 3.2. Interactions with Connector properties from multiple perspectives.

Property	Connector User	Connector Implementer	Scheduler
Synchronization	N/A	If instructed by scheduler, Node implementation performs appropriate pipeline/barrier operations	Connector chooses the appropriate synchronization between two Nodes (pipelining, __syncthreads, etc)
Storage Management	Must tell Connector how much storage it has, number of buffers, and provide a shared state for the Connector	Must include pipelines or atomics as needed for synchronization, and track input/output offsets	N/A
Batch Size Negotiation	Must specify Node batch size (ideally as a template parameter)	Nodes must query Connectors to determine if enough data or space is available, as well as input/output offsets. May be compile time or runtime	Must query Connectors to determine the batch size for Node scheduling. This may be compile time or runtime.

```
__shared__ ConnectorSharedState memcpy_connector_shared_state;
auto memcpy_connector = make_connector(Global2Shared, Shared2Global,
                                       &memcpy_connector_shared_state);
```

Listing 4: Connector Instantiation. The programmer provides the input and output Nodes as well as shared state to assist in synchronization.

or synchronization helpers and provide Nodes with a way to query available space/data as well as the read/write offsets.

3.3.2.3 Scheduler

Given two Nodes with user-specified compute locations, the Connector is responsible for choosing the appropriate synchronization between the Nodes. This may be a pipeline, a `__syncthreads()` barrier, or potentially a no-op depending on Node behavior and configuration. The scheduler must query the Connectors to determine the batch size for Node scheduling at either compile time or runtime.

3.4 Dataflows

```
auto BaseDF = make_dataflow();
auto [DF_Global2Shared, ID_Global2Shared] = BaseDF.add_node(Global2Shared);
auto [DF_Shared2Global, ID_Shared2Global] =
    DF_Global2Shared.add_node(Shared2Global);
auto [DF_Connected, ID_Connected] =
    DF_Shared2Global.add_connection(memcpy_connector);
```

Listing 5: Dataflow Instantiaton. The user incrementally builds a Dataflow by adding Nodes and Connectors in a functional programming style.

3.4.1 Specification

Dataflows represent an encapsulation of a diagram of Nodes and Connectors. To build a Dataflow, programmers first instantiate a Node or Connector, and then add it to the Dataflow in a functional programming style, where they incrementally build up a Dataflow by chaining together Nodes and Connectors (Listings 2, 4, and 5). Depending on the configuration of Nodes and Connectors, the Dataflow may be able to do higher-level reasoning to optimize the application's performance, such as fusing Nodes together.

To support Harmonic CUDA's goal of composability, a programmer can pass a Dataflow as a Node to another Dataflow. Note that while our prototype implementation of Harmonic CUDA currently does not support this, the abstraction *does*, and it is a clear next step for future work, with possible challenges being how to support the dynamic scheduling of a Dataflow that is passed as a Node and how to handle cases where the programmer wants to map a complete Dataflow to a variety of compute locations.

3.5 Interaction with CUDA

Harmonic CUDA aims to coexist with traditional CUDA code. To achieve this, threads can fill data into an input buffer of a Dataflow, consume data from an output buffer of a Dataflow, or interact with Node scheduling directly. The Node framework should also enable programmers to pass in CUDA lambda functions as parameters to a Node. For example, an "Elementwise" Node can consume elements one at a time from each input buffer. The programmer can then define the functionality of the Node, such as elementwise addition, summation over the stream, or some other operation.

3.6 Scheduling

3.6.1 Asynchronous Scheduling

Figure 3.1 shows an example of two memory movement Nodes mapped to different halves of the same block (e.g., warps 0–3 of the reader and 4–7 of the writer) following the warp specialization approach in Section 2.2. The Nodes pass data between one another using a shared memory buffer abstracted by the Connector. To optimize, the Connector can use double buffering, allowing the reader and writer to work asynchronously and in parallel. The Connector handles synchronization between the two Nodes using the "pipeline" abstractions from libcu++.

3.6.2 Interleaved Scheduling

The interleaved schedule presented in Figure 3.2 is logically the same Dataflow as before. However, in the interleaved schedule, the *location* to where the Nodes and Connector map are different. Now, both Nodes map to the same Cooperative Group, and the intermediate storage location becomes registers rather than shared or global memory since threads do not need to share data between one another. The functional end result of the algorithm is the same, but not the implementation. In the previous asynchronous schedule, each Node completely owned half of a block and repeatedly performed a single specific action as long as there was available buffer space. Now, the two Nodes share the same compute resources through time slicing in a pattern of read, write, read, write, read, etc. Additionally, while the previous example would have been able to take advantage of Ampere's asynchronous global-to-shared memory transfers, the interleaved schedule does not. This is because the intermediate storage location is a register, rather than shared memory. Rather, it could instead choose to use per-thread vector loads. The backend of the Node is optimized to take advantage of the available hardware resources, and can intelligently pick the correct backend depending on whether the Nodes are on the same or different cooperative group, or depending on what intermediate buffer location they use. This benefits the user because the user can experiment with different Node and Connector locations, without having to manually write the performant implementation of each configuration.

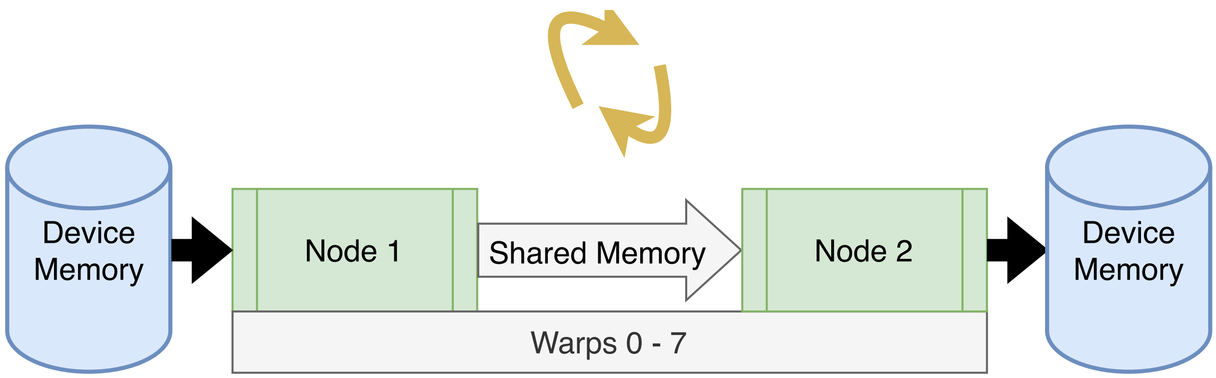

Figure 3.2. Interleaved Node scheduling where both Nodes share the resources of Warps 0–7 through time slicing.

3.6.3 Bulk-Synchronous Scheduling

Algorithms often require barriers between stages, and although Harmonic CUDA is a continuous dataflow producer/consumer programming model, it also supports sequenced operations. For example, a kernel where the programmer wants to perform an in-place sort followed by an in-place scan, they can set the chunk size equal to the data array size, blocking the scan Node until it has all available input data. Harmonic CUDA can also automatically determine if the sort is in-place or not by analyzing input and output storage addresses, in both asynchronous and interleaved schedules.

3.6.4 Manual Scheduling

The Manual Scheduling API enables advanced users to schedule Nodes to run on GPU hardware. Programmers can query a Node's readiness, invoke its `progress()` method, and set termination conditions. Listing 6 shows an example of manual scheduling for the memory copy example.

3.6.5 Automatic Scheduling

From the user's perspective, the only automatic scheduling API is to call the `run()` method on a Dataflow. In many cases the automatic scheduler is able to generate a static schedule at compile time for some portions of the Dataflow. For example, in a GEMM example, the scheduler knows some tile sizes (such as the sizes for the Tensor Cores) at compile time based

```
__global__ memcpy_kernel(int *global_data, int size) {

  // INSTANTIATION ...

  // Prime the pipeline
#pragma unroll
  for (int i = 0; i < PIPELINE_STAGES - 1; i++) {
    DF_Connected.get_node<ID_Global2Shared>().progress(BATCH_SIZE, pipe);
  }

  // Main scheduling loop.
  while (!DF_Connected.get_node<ID_Global2Shared>().is_finished() &&
         !DF_Connected.get_node<ID_Shared2Global>().is_finished()) {
    DF_Connected.get_node<ID_Global2Shared>().progress(BATCH_SIZE, pipe);
    DF_Connected.get_node<ID_Shared2Global>().progress(BATCH_SIZE, pipe);
  }

  // Finish the pipeline.
#pragma unroll
  for (int i = 0; i < PIPELINE_STAGES - 1; i++) {
    DF_Connected.get_node<ID_Shared2Global>().progress(BATCH_SIZE, pipe);
  }
}
```

Listing 6: Manually-scheduled memory copy. The user primes the pipe, schedules Nodes as long as there is data available, and then finally flushes the pipeline.

on the GPU architecture and the data type. In such cases, the automatic scheduler does not need to make each Node query its Connector for available space and can perform other optimizations such as unrolling. For schedules that cannot be determined at compile time, the automatic scheduler queries Nodes to determine if they are ready to run, have enough input data, and have enough output buffer space.

3.7 Summary

When programming with Harmonic CUDA, the programmer has several key design decisions and implementation tasks. First, how can the programmer break down their algorithm into Harmonic CUDA Nodes? In many cases, Harmonic CUDA has off-the-shelf Nodes a programmer can use. In other cases, the programmer may need to implement the desired functionality on top of Node templates in the Harmonic CUDA framework. Second, the programmer must connect all Nodes together into a Dataflow, specify and allocate the intermediate storage locations (future iterations of Harmonic CUDA should make this more automatic if possible), connect all Nodes together into a Dataflow, and specify desired optimizations such as double buffering. Finally, the programmer must determine which hardware units each Node maps to. This will likely call for some trial-and-error experimentation. In many cases the programmer may want to run all Nodes on the same cooperative group in the style of a traditional CUDA program, and in other cases the kernel may benefit from Cooperative Group Specialization.

Chapter 4

Matrix Multiplication

4.1 Motivation

Generalized Matrix Multiplication (GEMM) is typically implemented as a device-wide operation. For example, cuBLAS, NVIDIA's closed-source matrix multiplication library, supports GEMM operations launched from the GPU or CPU that utilize the entire device. CUTLASS, NVIDIA's open-source matrix multiplication library, attempts to support more composability in how the GPU handles tile sizes for a variety of matrix shapes, but this is still primarily at the device level, where all threads, warps, or blocks in a GPU work together to solve a single GEMM problem. This limits the ability of programmers to experiment with GEMM use cases that do not require a full GPU or that combine GEMM building blocks with other operations (e.g., Chapter 5). GEMM also commonly utilizes hardware acceleration units such as DMA, TMA, or Tensor Core units.

As GPUs add more of these units over time, it is important for programmers to have abstractions around these logical operations so they do not need to do major code rewrites. To address this, our Harmonic CUDA GEMM implementation provides generic building blocks for memory movement and compute in a way that abstracts away the location, size, or composition of the building blocks. This allows us to utilize the same core building block no matter where our input and output data sources are. For example, moving a 2D tile of a matrix is logically the same operation whether it is a global-to-shared or a shared-to-register transfer, even though the implementation may change.

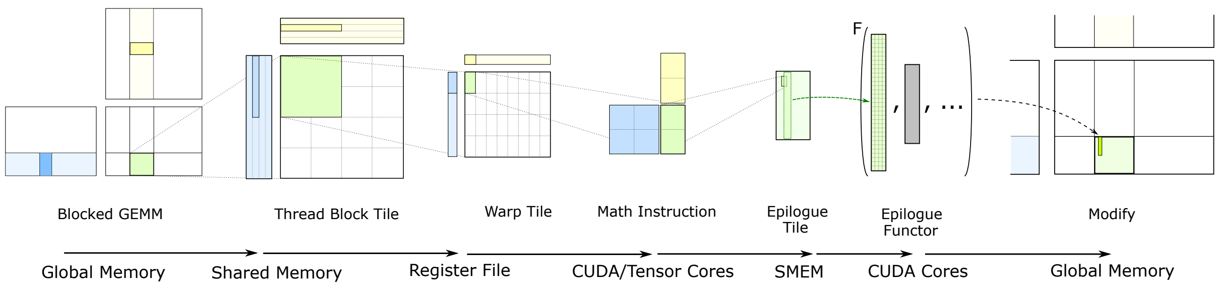

Figure 4.1. Matrix multiplication tiling hierarchy showing the 2D memory movement patterns between global and shared memory, shared memory and registers, and so on. Image used with permission, via CUTLASS [18].

4.2 Implementation

In contrast to Harmonic CUDA, CUTLASS only supports a fixed set of operations at each level of the hierarchy. At the device level, this is the full GEMM operation; at the block level, a fixed-size tile of the matrix; and at the warp level, a smaller fixed-size tile that corresponds to the size of a tensor core, with separate implementations for data movement between each level. Harmonic CUDA allows the programmer to reason about the matrix multiplication building blocks at a higher level of abstraction, while allowing the backend to take care of the performance-critical optimizations that could be provided by CUTLASS as a lower-level backend.

As shown in Figure 4.1, GEMM is a hierarchical algorithm. We first tile the output matrix in global memory and iteratively read in tiles of the *A* and *B* matrices from global to shared memory. We then repeat this pattern, tiling a block-level output matrix for warps, and for threads, and so on. Logically, these are all simply 2D tile movement patterns between different levels of the memory hierarchy. In Harmonic CUDA we can express this as a single "2D Tile Mover" Node regardless of the size of the tile, where it runs on the GPU, or the location of its inputs or outputs. Finally, we include an "MMA" Node that performs the matrix multiplication and accumulation for a single tile of the output matrix. By abstracting away any optimization details common in Matrix Multiplication kernels, we enable programmers to think about GEMM as a small set of building blocks that move memory in 2D tiles between different levels of the memory hierarchy. We show the Harmonic CUDA Dataflow in Figure 4.2.

Although at a high level the only Node the programmer needs to think about is the "2D Tile

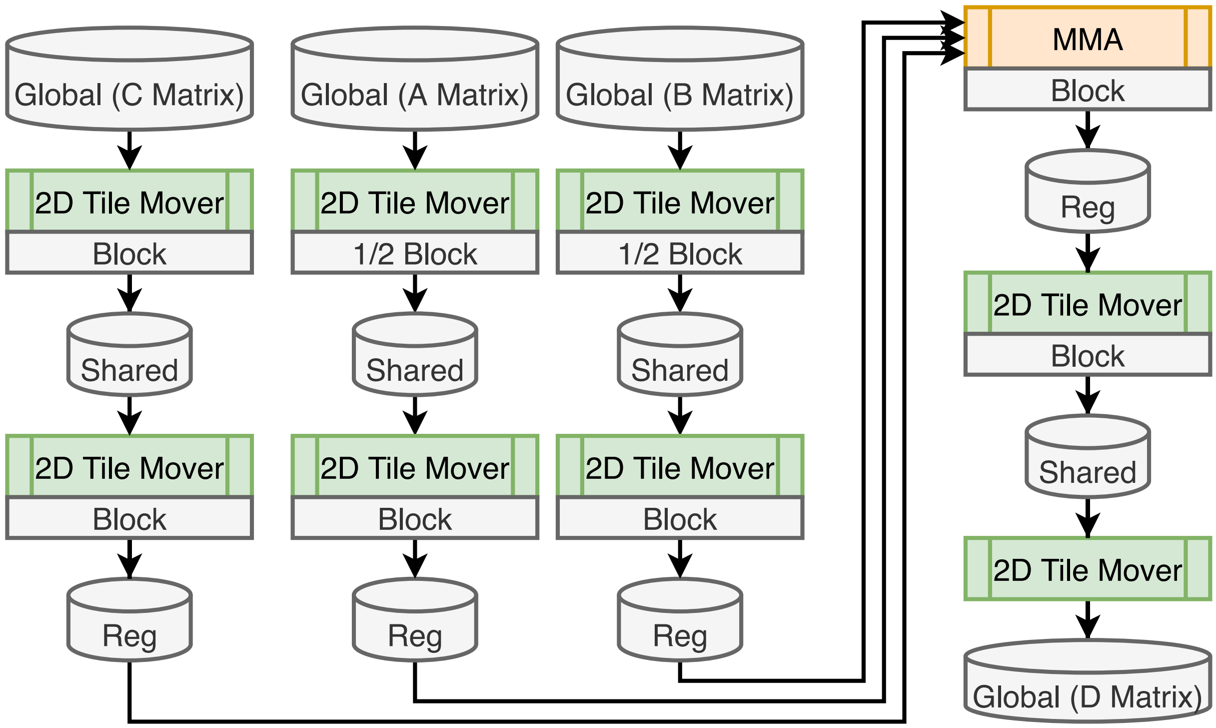

Figure 4.2. Matrix Multiplication using Harmonic CUDA 2D Tile Mover and MMA Nodes.

Mover" Node, internally, the Node specializes its backend based on the input and output locations, which compute group it is assigned to, whether the data is row- or column-major format, the data type of the matrix, and other runtime and compile-time parameters. This specialization should be invisible to the user. Similarly, for the MMA Node we logically map the MMA Node to the entire block, where the inputs are registers. Although internally the MMA Node takes advantage of the GPU's Tensor Cores and complex logic for index mappings, this is again invisible to the user. The MMA Node could be alternatively mapped to a thread, a single warp, or a subset of a block and still *logically* perform the correct implementation.

We adapt NVIDIA's `dmmaTensorCoreGemm` example [26] as the backend for all Nodes, modifying operations to support arbitrary cooperative groups. To match the implementation in the sample code repository, we map the 2D Tile Mover Node to separate halves of each block for the global-to-shared memory copies of A and B matrices, as this is more efficient than mapping each to the entire block due to the required shared memory tile sizes. However, the strength of Harmonic CUDA is that it is easy for the programmer to experiment with configurations of

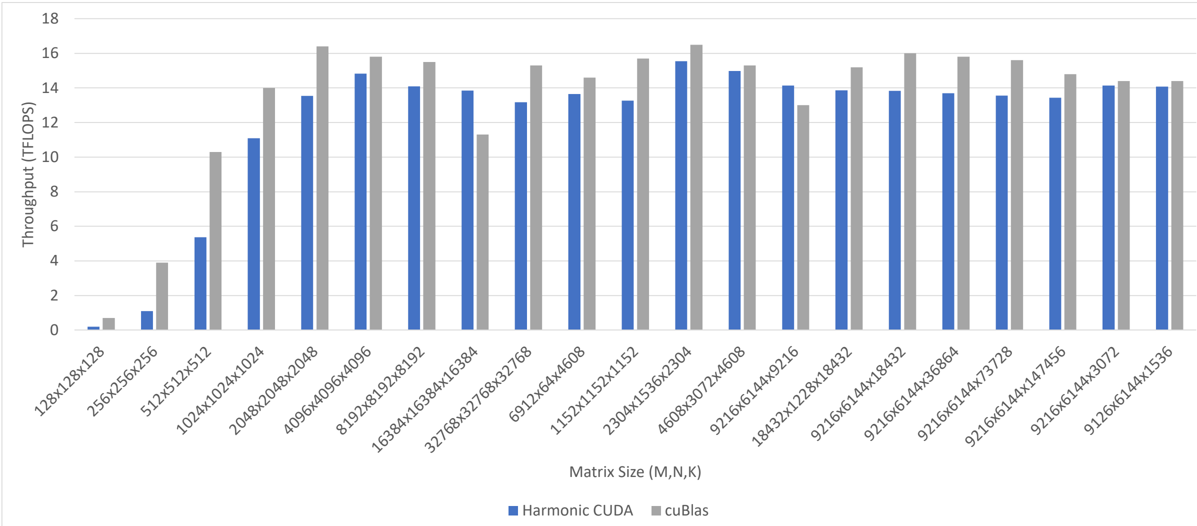

Figure 4.3. GEMM performance results evaluated over a range of matrix shapes including square, perfectly-balanced, mixed aspect ratios, and GraphSage work-equivalent.

Nodes onto compute hardware; for example, mapping all Nodes to the same block would be equally valid (albeit slower).

We manually schedule Nodes to allow optimal kernel performance. We assume a more sophisticated automatic scheduler with minimal overhead could statically determine an appropriate schedule, as the compiler already knows the Tensor Core tile size at compile time. Note that the programmer does not need to specify the tile size, as Nodes and Connectors automatically select the implementation and size based on architecture capabilities, allocated resources, and the data types of the operation.

4.3 Results

We compare our results in Figure 4.3 against cuBLAS, NVIDIA's highly optimized, closed-source GEMM library. We conduct all experiments on an NVIDIA A100 GPU using CUDA Toolkit version 11.6. Matrices evaluated include square matrices (up to 32768 rows/columns), tiles that are an integer multiple of SMs to avoid workload imbalance, mixed aspect ratios, and (M, N, K) = (6912, 64, 4608), approximating the amount of work done in the GraphSage kernel (Chapter 5). Our tests show a geomean average speedup of 0.8X vs. cuBLAS, but excluding the 3 smallest matrices, where cuBLAS has specialized routines for small matrices, our geomean speedup improves to 0.92X.

We believe that the performance discrepancy between Harmonic CUDA and cuBLAS is because although Harmonic CUDA currently takes advantage of the A100's Tensor Cores and DMA units, it lacks many of the low-level optimizations that cuBLAS performs such as pipelined memory movement, assembly-level optimizations, heuristics for tile sizes, and many others. As such, this performance discrepancy is expected. The goal of this experiment is not to *beat* cuBLAS, but rather to show that we are able to express matrix multiplication as a dataflow graph of producers and consumers with a moderately fast backend that allows straightforward implementation and experimentation. We fully expect that in the future, given more development time, an open-source library such as CUTLASS (which achieves performance parity with cuBLAS) could instead become the backend of Harmonic CUDA's GEMM Nodes. Note that the primary barrier to implementing these optimizations in Harmonic CUDA is engineering time, rather than limitations of the programming model itself. Pipelined memory movement fits perfectly into the "Connector" abstraction, assembly-level optimizations may form the backends of Nodes, and heuristics for tile sizes are a fundamental part of the "Node" abstraction, where the Node picks the appropriate backend implementation depending on the architecture, data type, matrix size, and other parameters. We believe that the Harmonic CUDA programming model is well-suited to these optimizations.

In the next chapter, we show how the building blocks used in this kernel can be adapted in new, more flexible ways for another kernel that has similarities to the GEMM kernel, but with application-specific requirements.

Chapter 5

GraphSage

5.1 Motivation

GraphSage is an algorithm for machine learning on graphs that samples the first-hop and second-hop neighbors of the vertices of the graph and then uses features of these neighbors as inputs to train a dense neural network [12]. Computationally, GraphSage is interesting because it includes both a sparse stage (multiple layers of indirection from sampling the first and second hop

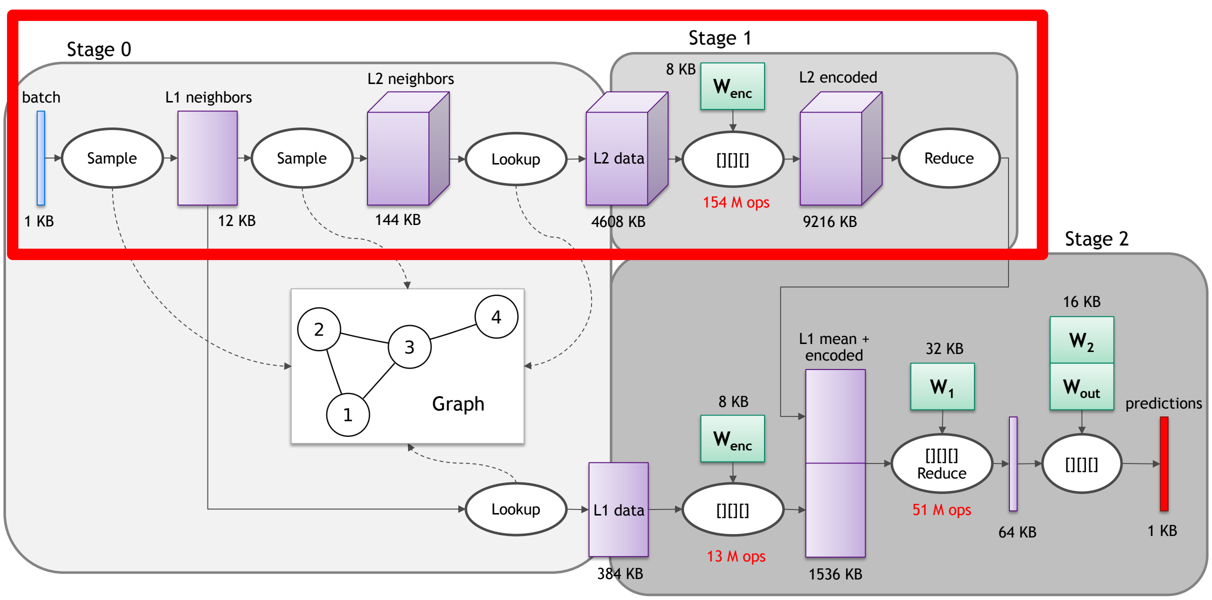

Figure 5.1. GraphSage forward pass with the subset implemented in Harmonic CUDA highlighted in red. Stage 0 represents sparse gathers and Stage 1 represents repeated GEMM operations. Reproduced with permission from Angshuman Parashar.

neighbors of a batch of vertices) and a dense stage (performing many matrix multiplications for training the dense neural network). In this evaluation, we focus on only a portion of the forward pass that isolates the sparse sampling stage and a subset of the dense stage in part as a way to demonstrate the flexibility of the Harmonic CUDA programming model without adding too much complexity to the evaluation and in part due to time constraints. We show the full forward pass and highlight our specific implementation in Figure 5.1.

A naive approach to GraphSage on the GPU is to simply separate the algorithm into two kernels. First, one kernel samples the first-hop and second-hop neighbors and stores their features in global memory. Next, a second kernel performs the dense linear algebra operations. This approach is suboptimal because it requires the first stage to perform an expensive global memory write of all features, followed by an expensive global memory read from the second stage to re-read all features. However, with Harmonic CUDA, it is possible to implement a more optimized approach with minimal additional complexity that combines the two kernels into a single Dataflow and stores the sampled neighbor features in intermediate on-chip buffers, only writing the final output to off-chip memory. Additionally, with Harmonic CUDA, we can reuse many of the building blocks of the previously described GEMM example in new ways to implement shared-memory linear algebra operations and can take advantage of warp specialization to run the sparse and dense stages in parallel.

5.2 Implementation

We construct a Harmonic CUDA implementation of the GraphSage algorithm using building blocks adapted from prior examples in a new context, as shown in Figure 5.2. The strength of Harmonic CUDA is that nowhere in this implementation do we need to explicitly specify the mapping of Nodes to compute resources or the actual implementations of any of the backends of the Nodes. By thinking of GraphSage as a dataflow graph rather than as a sequence of instructions to execute, we can easily experiment with different configurations of the Nodes and can even do more abstract reasoning about *how* to map the algorithm to the GPU. For example, by pipelining the sparse gather stage directly into the dense GEMM stage on-chip, rather than separating these into separate kernels, we can eliminate a significant amount of memory traffic.

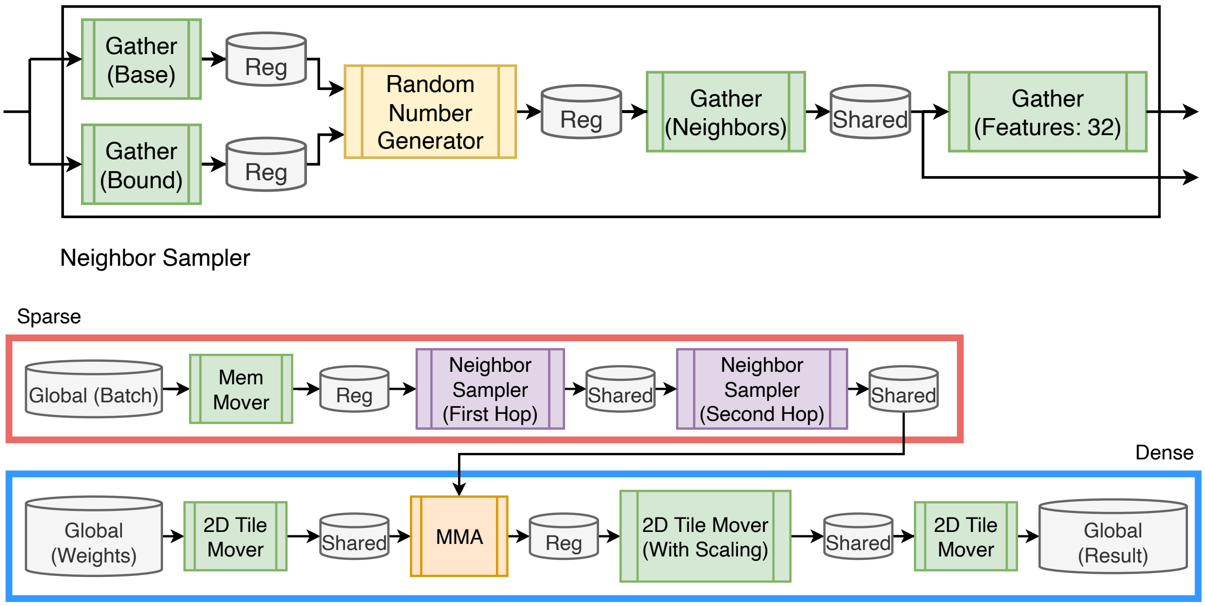

Figure 5.2. Top: GraphSage Neighbor Sampler Dataflow. Bottom: GraphSage Dataflow with Sparse and Dense Stages.

In our experiments in the following sections, we evaluate GraphSage in two configurations: one configuration where the sparse gather stage and the dense GEMM stage share the same compute resources and time slice between them, and another configuration where we assign additional threads dedicated to sparse gathers that send data to the GEMM stage over a shared memory buffer. This is easy to perform since the only changes required are to change the compute locations of the Nodes and to change the locations of the intermediate storage.

5.2.1 Sparse Gather Stage

In the sparse gather stage, the algorithm reads in the batch of vertices and samples 12 first-hop neighbors per batch vertex and 12 second-hop neighbors per first-hop vertex, randomly selected. A vertex with a degree less than 12 will sample duplicate neighbors. Our Harmonic CUDA implementation reuses the “MemMover” Node from Chapter 3 and adds several new Nodes for performing memory indirections (the “Gather” Node) and random sampling (the “RandomNumberGenerator” Node). We also create a modular “Neighbor Sampler” dataflow that we reuse multiple times for the first-hop and second-hop neighbor samples. As shown in Figure 5.2, we can express this algorithm as a dataflow without thinking about how to parallelize it, map it to hardware, or perform low-level optimizations. Instead, the programmer thinks about the algorithm as an abstract sequence of logical transformations done to each vertex in the batch, where data flows sequentially through the dataflow graph over time. The fact that Nodes have parallel internal implementations and produce and consume data in a pipelined manner is abstracted away from the programmer. The end result of this Dataflow is that for each vertex input to a NeighborSampler Dataflow, the Dataflow outputs a 64×32 matrix of feature values as well as a stream of the vertex indices corresponding to each matrix for each second-hop neighbor of the batch vertices.

5.2.2 Dense GEMM Stage

Rather than using a device-wide GEMM where all blocks work on tiles of the same matrix as seen in Chapter 4, our implementation of GraphSage uses a per-block GEMM where each block repeatedly multiplies and accumulates samples of the second-hop vertex features with a small weight matrix. Additionally, where in our GEMM implementation in Chapter 4 we previously mapped the MMA Node to the entire block, in GraphSage, we can instead map the MMA Node to only a portion of the block, leaving the rest of the block free to perform the neighbor indirections to collect the features. We also have different 2D tile movement patterns to consider. Where we previously iterated over tiles of B in global and shared memory, GraphSage instead repeatedly uses the weight matrix as B, with the A matrix changing each iteration. Harmonic CUDA's abstractions allow us to easily reuse the existing building blocks, where the only necessary change is a small modification to the flags that describe where the MMA Node runs, the location of the inputs, and the size of the inputs.

5.3 Results

We compare a warp-specialized, asynchronously-scheduled GraphSage implementation in Harmonic CUDA to a bulk-synchronous, interleaved configuration in Harmonic CUDA, showing significant performance improvements and easier experimentation. To prove these speedups are from more efficient hardware utilization, rather than additional hardware resources allocated during warp specialization, we also compare against cuBLAS using a variant of GraphSage that eliminates sparse lookups to isolate the performance of the dense GEMM stage.

5.3.0.1 GraphSage Performance

Our first Harmonic CUDA implementation is bulk-synchronous and interleaved. In it, we assign all Nodes in the Dataflow to a full block (256 threads) and interleave sparse lookups with dense GEMM operations, yielding a throughput of 8.78 TFLOPS. Harmonic CUDA's flexibility allows us to easily pivot our implementation to a different configuration. In this implementation, we assign sparse lookups to an additional 64 threads, while keeping the GEMM on 256 separate threads. This allows asynchronous scheduling with data passing over a double-buffered shared memory Connector and improves throughput by 34% to 11.75 TFLOPS. The only thing we have to do to make this change is to modify the single parameter that specifies which Cooperative Group each Node runs on.

Now, does this performance improvement stem from the warp specialization approach or from the additional 64 threads per block? Harmonic CUDA's flexibility aids us in performing this design exploration. To begin, we modify GraphSage to eliminate indirection from first- and second-hop neighbor lookups, instead repeatedly using only batch vertex features. This yields a throughput of 13.73 TFLOPS. Each block now only performs a series of small GEMM operations to repeatedly multiply the batch feature matrices with the weight matrix. If this GEMM-like kernel can perform on par with cuBLAS, we can conclude that because cuBLAS is able to efficiently saturate the hardware, adding additional threads to this modified GraphSage experiment would not improve performance.

5.3.0.2 GEMM Performance

We approximate the repeated small GEMM operations of the modified GraphSage kernel as if they were tiles of a larger GEMM. Testing the equivalent matrix[1] in cuBLAS, we achieve 14.6 TFLOPS. Since this is only a 6% improvement over the modified GraphSage kernel, we can conclude that the GraphSage GEMM stage would not benefit from allocating an additional 64 threads (a 25% increase) to the stage.

[1] The equivalent matrix has M as the total batch vertices (108 blocks $\times$ batch size of 64), N as the weights per feature (64), and K as the total number of features (12 first-hop neighbors $\times$ 12 second-hop neighbors per first-hop neighbor $\times$ 32 features per second-hop neighbor) for $(M,N,K) = (6192,64,4608)$.

5.3.0.3 Analysis

This GraphSage example highlights the advantages of considering the algorithm's dataflow and computation mapping separately. With Harmonic CUDA, the programmer can easily experiment with different mappings of Nodes to compute resources. By relocating each Node and the intermediate buffer storage with a simple parameter change, the formerly bulk-synchronous kernel becomes warp-specialized, resulting in improved performance compared to the bulk-synchronous version. Additionally, by viewing GraphSage as a Dataflow of building blocks, we can easily express changes in computation and storage locations, keep intermediate results on-chip, and reuse the building blocks from Chapter 4 in a new context.

Chapter 6

Conclusion

Harmonic CUDA demonstrates a novel approach to asynchrony on GPUs that creates an abstraction between *what* a computation does and *when*/*where* the computation happens, and is designed to be performant, composable, easily programmable, and to coexist with the existing CUDA programming model on NVIDIA GPUs. Rather than thinking about GPU code as a sequence of operations to be executed by a thread, programmers can think about their code in a fundamentally different way. Instead of thinking about how to manage threads and memory, programmers can think about how to connect building blocks together to form a larger computation, where data flows through the compute graph over time. This approach to programming is often more intuitive and easier to write and is particularly useful for programmers who need to keep up with the zoo of hardware accelerators and highly optimized backends, want low-level optimizations for efficient asynchrony, and want to compose pieces from a variety of libraries together in novel ways. It also helps programmers scale their algorithm from a thread all the way to an entire cluster of GPUs just by changing the specification of *where* the algorithm's Dataflow runs, which is a critical contribution both as threads become more and more capable, and as parallel systems gain ever more levels of parallel hierarchy that users want to target. Harmonic CUDA contrasts to other GPU programming models on GPUs that demand to be the only programming model running on a GPU at a given time such as a DSL with a custom CUDA backend compiler or a library that decides that you only need to do certain operations such as GEMM with the entire device. By being harmonious, Harmonic CUDA ensures that a programmer can use it when it makes sense without large code rewrites, and can fall back to traditional CUDA code as needed.We demonstrated Harmonic CUDA's programming model using a memory copy example and further evaluated the programming model on two real-world examples. Matrix Multiplication shows that Harmonic CUDA programmers can implement a key kernel in GPU computing that, while only achieving a geomean performance of 80% of cuBLAS, shows a path forward to a highly-optimized implementation using cuBLAS or CUTLASS as a backend. GraphSage shows that programmers can utilize Harmonic CUDA to implement a real-world application while achieving significant performance improvements over a bulk-synchronous implementation (a 34% improvement for GraphSage), and that implementing this kernel is straightforward and intuitive when reusing the same building blocks as the matrix multiplication kernel.

We identify several areas for future research. First, expanding Harmonic CUDA to new algorithms that focus on irregular parallelism, specializations of subsets of a grid, and algorithms that require more complex Dataflow graphs. We also believe that Harmonic CUDA should expand beyond a single GPU to groups of multiple GPUs on one Node or other architectures such as CPUs and accelerators. This would allow the *where* abstraction to extend not just to layers of the GPU's memory hierarchy, but to CPU cores, domain-specific hardware, and heterogeneous systems. It should also expand to clusters of multiple GPUs or multiple Nodes, where it has the potential to make managing data orchestration in a distributed system easier. In addition, we would like to further improve the model to make it as performant and easy-to-use as possible.

6.1 Programming Model Evaluation

6.1.1 Lessons Learned

Harmonic CUDA enables programmers to think about their algorithms in a fundamentally new way. By thinking about their algorithm as a dataflow of building blocks, rather than as a sequence of instructions for a thread, warp, or block to run, the programmer can reason about how an algorithm transforms data without needing to worry about how specifically to map the algorithm to the GPU. When the programmer *does* want to map the algorithm to the GPU, Harmonic CUDA's ability to specify a Node's functionality independently from its compute location allows a program implementation that can scale from a single thread to a single GPU to an entire cluster of GPUs. Depending on where the Node runs, the backend of the Node will choose the most efficient implementation available. One of Harmonic CUDA's most important ideas is that any building block (such as a GEMM, tiled memory movement, or a reduction) has use cases in many contexts that are not traditionally supported by GPU libraries and that a programmer should be able to reuse the same building block in many applications and at many different granularities.

It is important to note that the question of where to map a Node is an exercise for the programmer. Harmonic CUDA does not claim to be able to automatically map a Node to the best location for the programmer. What it *does* do is make this easy for the programmer to do themselves. We believe that one of the biggest historical barriers to asynchrony on GPUs is the challenge of managing the interactions between different compute units, and we believe that Harmonic CUDA removes this barrier.

However, the Harmonic CUDA experiments in the previous sections do demonstrate several challenges of writing both the Harmonic CUDA backend as well as kernels written with Harmonic CUDA. On the backend, there is a risk that maintaining a robust library of Nodes becomes too complicated, as each Node is expected to target a variety of compute locations, compute capabilities, input and output data locations, runtime flags, compile-time flags, and so on. We expect that although there may be some additional complexity, we will be able to build on a foundation of core features that make up the majority of use cases. For example, NVIDIA's Cooperative Groups API provides some of the building blocks for managing arbitrary groups of threads. To target the variety of hardware accelerators, backend developers can target these slowly over time as new GPU releases include new hardware accelerators.

6.1.2 Next Steps

Harmonic CUDA's scheduler is a key area for future development. The work in this thesis provides a simple implementation of automatic scheduling for the memory copy kernel as a demonstration. However, for the Matrix Multiplication and GraphSage kernels, we chose to use manual Node scheduling to focus on the best-case scenario for the programming model. For future work, the scheduler will need to be able to automatically schedule Nodes to maximize performance. Additional work is required to make the automatic scheduler's performance competitive with manual scheduling and to support nested loops, as seen in the matrix multiplication kernel. In some cases, this will require a static analysis of the dataflow graph to determine the best scheduling order. This analysis will need to take into account the compute location of each Node, the data location of each Node, and the compute capabilities of the hardware. In other cases, the scheduler will need to be able to efficiently query and schedule Nodes at runtime. The scheduler will also need to be able to avoid deadlocks that may occur when a Node has multiple inputs or outputs that consume or produce data at varying rates.

An important task for Harmonic CUDA development going forward is to identify what the core building blocks of GPU computing are. In many cases, this is simply to create wrappers around common GPU primitives such as those found in CUB [21] or more flexible versions of algorithms found in NVIDIA's most-used libraries such as cuBLAS, cuSparse, and others. However, there are likely other cases where the core building blocks are not as obvious, or that they are so common that it would be redundant for a library to implement them (say, an elementwise operation on two vectors). In particular, it is important to develop this suite of Nodes with an eye toward which Nodes may gain future hardware acceleration capabilities.

Harmonic CUDA exposes a number of implementation-side tasks that future GPU architectures and programming models should implement to improve performance and programmability. Harmonic CUDA shows the value of separating *what* computation is being done from *where* the computation is being done. NVIDIA's *Cooperative Groups* API [13] does this in minor ways, for example by providing an arbitrary Cooperative Group as input to a prefix sum or reduction function, but this is not a widespread practice across GPU libraries despite the large benefits in the flexibility offered by this abstraction. Additionally, the Cooperative Groups API does not support groups that are a subset of a grid (i.e. spanning multiple blocks). This feature would be useful for kernels where data in multiple stages of a pipeline can be processed by different groups of blocks, with a global memory buffer between them. Finally, we observed that using Cooperative Grid synchronization methods was significantly slower than a `__syncthreads()` function call. Future GPUs should provide a hardware synchronization primitive for Cooperative Groups that is as fast as `__syncthreads()`.

6.1.3 Future Work

One promising future research direction for Harmonic CUDA is to map the programming model to additional architectures such as CPUs, AMD GPUs, and accelerators. This would enable programmers to use the model in a wider variety of domains, and would also allow for a more thorough comparison of the model to other programming models. Since the programming model abstracts the type of computation away from the backend, a programmer could use Harmonic CUDA as an efficient way to specify the dataflow of an application, but then map the Nodes and Connectors of the Dataflow to accelerator-specific hardware.

For long-term future work, Harmonic CUDA should expand to even more kernels. The Breadth-First Search kernel is a good candidate for this, as it has irregular parallelism and load balancing, which are both features that Harmonic CUDA is well-suited to handle. Prior BFS works have shown the value of asynchronous programming models where parallel workers can dynamically generate new work within a kernel and add it to a work queue [5]. The ConvNet kernels are another good example, as they rely on CTA specialization for a performant implementation. Additionally, Harmonic CUDA should be able to scale to multi-GPU or multi-Node systems. The programming model's ability to hide backend details from the programmer makes it well-suited to this task since all a programmer would need to do is change the location a Node maps to. NVSHMEM [24] provides a powerful multi-GPU data sharing API that the programming model could use as a backend.

REFERENCES

[1] Farhoosh Alghabi, Ulrich Schipper, and Andreas Kolb. A scalable software framework for stateful stream data processing on multiple GPUs and applications. In *GPU Computing and Applications*, pages 99–118. Springer Singapore, September 2014. doi: 10.1007/978-981-287-134-3_7.

[2] Michael Bauer, Henry Cook, and Brucek Khailany. CudaDMA: Optimizing GPU memory bandwidth via warp specialization. In *Proceedings of the 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '11, pages 1–11, November 2011. doi: 10.1145/2063384.2063400.

[3] Michael Bauer, Sean Treichler, and Alex Aiken. Singe: Leveraging warp specialization for high performance on GPUs. In *Proceedings of the 19th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, PPoPP '14, pages 119–130, February 2014. doi: 10.1145/2555243.2555258.

[4] Zhi bin Huang, Guang-Tao Fu, Tian-Hao Fa, Dan-Yang Dong, Peng Bai, and Chen Xiao. High performance ant colony system based on GPU warp specialization with a static-dynamic balanced candidate set strategy. *Future Generation Computer Systems*, 125:136–150, December 2021. doi: 10.1016/j.future.2021.06.041.

[5] Yuxin Chen, Benjamin Brock, Serban Porumbescu, Aydın Buluç, Katherine Yelick, and John D. Owens. Atos: A task-parallel GPU scheduler for graph analytics. In *Proceedings of the International Conference on Parallel Processing*, ICPP 2022, August/September 2022. doi: 10.1145/3545008.3545056.

[6] Jack Choquette, Oliver Giroux, and Denis Foley. Volta: Performance and programmability. *IEEE Micro*, 38(2):42–52, April 2018. doi: 10.1109/MM.2018.022071134.

[7] Federico Ciccozzi, Lorenzo Addazi, Sara Abbaspour Asadollah, Björn Lisper, Abu Naser Masud, and Saad Mubeen. A comprehensive exploration of languages for parallel computing. *ACM Comput. Surv.*, 55(2), January 2022. doi: 10.1145/3485008.

[8] William J. Dally, Stephen W. Keckler, and David B. Kirk. Evolution of the graphics processing unit (GPU). *IEEE Micro*, 41(6):42–51, 2021. doi: 10.1109/MM.2021.3113475.

[9] Michał Dominiak, Georgy Evtushenko, Lewis Baker, Lucian Radu Teodorescu, Lee Howes, Kirk Shoop, Michael Garland, Eric Niebler, and Bryce Adelstein Lelbach. std::execution. C++ Standards Committee Papers, April 2022. URL `https://www.open-std.org/jtc1/sc22/wg21/docs/papers/2022/p2300r5.html`.

[10] Alex Fender, Brad Rees, and Joe Eaton. RAPIDS cuGraph. In *Massive Graph Analytics*, pages 483–493. Chapman and Hall/CRC, May 2022. doi: 10.1201/9781003033707-22.

[11] Tae Jun Ham, Lisa Wu, Narayanan Sundaram, Nadathur Satish, and Margaret Martonosi. Graphicionado: A high-performance and energy-efficient accelerator for graph analytics. MICRO-49, October 2016. doi: 10.1109/micro.2016.7783759.

[12] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 1025–1035, December 2017. URL `https://proceedings.neurips.cc/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf`.

[13] Mark Harris and Kyrylo Perelygin. Cooperative groups: Flexible CUDA thread programming, October 2017. URL `https://developer.nvidia.com/blog/cooperative-groups/`.

[14] Kartik Hegde, Hadi Asghari Moghaddam, Michael Pellauer, Neal Clayton Crago, Aamer Jaleel, Edgar Solomonik, Joel S. Emer, and Christopher W. Fletcher. ExTensor: An accelerator for sparse tensor algebra. In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO-52, pages 319–333, October 2019. doi: 10.1145/3352460.3358275.

[15] Nadhira Hill. A grad student's guide to free food, February 2019. URL `https://eidolon.pub/a-grad-students-guide-to-free-food-8b7e7a6229f4`.

[16] Dominique Houzet, Sylvain Huet, and Anis Rahman. SysCellC: a data-flow programming model on multi-gpu. *Procedia Computer Science*, 1(1):1035–1044, May 2010. doi: https://doi.org/10.1016/j.procs.2010.04.115. ICCS 2010.

[17] Arpith Chacko Jacob, Alexandre E Eichenberger, Hyojin Sung, Samuel F Antao, Gheorghe-Teodor Bercea, Carlo Bertolli, Alexey Bataev, Tian Jin, Tong Chen, Zehra Sura, Georgios Rokos, and Kevin O'Brien. Efficient fork-join on GPUs through warp specialization. In *2017 IEEE 24th International Conference on High Performance Computing*, HiPC 2017, pages 358–367, December 2017. doi: 10.1109/HiPC.2017.00048.

[18] Andrew Kerr, Duane Merrill, Julien Demouth, and John Tran. CUTLASS: Fast linear algebra in CUDA C++, December 2017. URL `https://devblogs.nvidia.com/cutlass-linear-algebra-cuda/`.

[19] Ronny Krashinsky, Olivier Giroux, Stephen Jones, Nick Stam, and Sridhar Ramaswamy. NVIDIA Ampere architecture in-depth, May 2020. `https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/`.

[20] MathWorks Corporation. Simulink, October 2022. URL `https://www.mathworks.com/help/simulink/index.html`.

[21] Duane Merrill. CUB: Flexible library of cooperative threadblock primitives and other utilities for CUDA kernel programming. `https://github.com/NVIDIA/cub`, 2013–2022.

[22] Nicholas Moore, Miriam Leeser, and Laurie Smith King. Kernel specialization provides adaptable GPU code for particle image velocimetry. *IEEE Transactions on Parallel and Distributed Systems*, 26(4):1049–1058, April 2015. doi: 10.1109/TPDS.2014.2317721.

[23] National Instruments Corporation. LabVIEW documentation, July 2022. URL `https://www.ni.com/docs/en-US/bundle/labview/page/lvhelp/labview_help.html`.

[24] NVIDIA Corporation. NVSHMEM, 2016. URL `https://developer.nvidia.com/nvshmem`.

[25] NVIDIA Corporation. NVIDIA H100 tensor core GPU architecture, 2020. `https://resources.nvidia.com/en-us-tensor-core`.

[26] NVIDIA Corporation. CUDA samples. `https://github.com/NVIDIA/cuda-samples`, 2022.

[27] NVIDIA Corporation. CUDA cuBLAS library (v11.6), 2022. `http://developer.nvidia.com/cublas`.

[28] NVIDIA Corporation. libcu++: The C++ standard library for your entire system, November 2022. URL `https://nvidia.github.io/libcudacxx/`. Version 1.8.1.

[29] Michael Pellauer, Yakun Sophia Shao, Jason Clemons, Neal Crago, Kartik Hegde, Rangharajan Venkatesan, Stephen W. Keckler, Christopher W. Fletcher, and Joel Emer. Buffets: An efficient and composable storage idiom for explicit decoupled data orchestration. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '19, pages 137–151, April 2019. doi: 10.1145/3297858.3304025.

[30] Jonathan Ragan-Kelley, Connelly Barnes, Andrew Adams, Sylvain Paris, Frédo Durand, and Saman Amarasinghe. Halide: A language and compiler for optimizing parallelism, locality, and recomputation in image processing pipelines. In *Proceedings of the 34th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI '13, pages 519–530. ACM Press, June 2013. ISBN 9781450320146. doi: 10.1145/2491956.2462176.

[31] Rahul Sharma, Michael Bauer, and Alex Aiken. Verification of producer-consumer synchronization in GPU programs. In *Proceedings of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI '15, pages 88–98, June 2015. doi: 10.1145/2737924.2737962.

[32] William Thies, Michal Karczmarek, and Saman P. Amarasinghe. StreamIt: A language for streaming applications. In R. Nigel Horspool, editor, *Proceedings of the 11th International Conference on Compiler Construction*, Lecture Notes in Computer Science, pages 179–196. Springer-Verlag, April 2002. ISBN 3-540-43369-4. doi: 10.1007/3-540-45937-5_14.