SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Exploring GPU Programming Models for Autonomous Driving: From Coroutine Integration to Persistent Thread Optimization

Jaden Rotter

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Exploring GPU Programming Models for Autonomous Driving: From Coroutine Integration to Persistent Thread Optimization

GPU Coroutines in Autonomes Fahren

Author:	Jaden Rotter
Examiner:	Supervisor
Supervisor:	Jianfeng Gu
Submission Date:	Submission date

Acknowledgments

Abstract

Contents

1 Introduction

1.1 Motivation

Autonomous driving systems place stringent demands on computational performance, predictability and safety. These demands arise from the need to process vast amounts of sensor data and run complex perception and decision making algorithms in real time, while ensuring timely and deterministic responses to a dynamic environment. To meet the computational requirements of such systems, GPUs have become essential due to their performance on machine learning workloads. However, the current GPU programming and execution model is poorly suited to real time constraints. While this thesis does not implement the proposed approach directly within an autonomous driving stack, the challenges of meeting real time requirements in autonomous systems, particularly in GPU workloads, serve as the primary motivation for this research.

1.1.1 Evolution of Autonomous Driving Architectures

Historically, early autonomous driving systems addressed the real time requirements using a distributed architecture. In this approach, major functional modules, such as perception, localization, planning, and control, were mapped to seperate compute units, which together formed a processing pipeline [1]. Each module could thus operate with predictable timing characteristics, avoiding contention with other modules. In this manner, the distributed architecture allowed for fine tuning the timing between modules to achieve low latency responses from the hardware. This modular architecture ensured responsiveness and real time guarantees at the expense of high hardware cost and increased system complexity.

The rise of increasingly powerful GPUs has enabled a shift toward centralized computing in order to simplify the hardware and complexity while reducing costs. In this centralized architecture, all core driving modules share a common compute node consisting of a CPU-GPU system. The compute node integrates a CPU for task scheduling and system control, complemented by a GPU optimized for compute intensive workloads. Moving to a singular compute node allows savings in cost, design, and intermodule latency.

1.1.2 Limitations of Current GPU Execution Models

Despite the benefits afforded by a centralized architecture and the advances in hardware, the system risks GPU oversubscription. As the GPU is simultaneously responsible for all processing tasks, too many simultaneous scheduled tasks can lead to delays in execution time. Typical real time system solutions based on a CPU architecture, ensure system safety under contention by preempting non critical tasks. Tasks requiring high responsiveness can thus be directly executed after preemption of the resident threads and processes. This preemption is supported both natively at the OS and user space levels on the CPU, but not the GPU.

Modern GPUs, are unable to natively support real time programming models and instead rely on a proprietary hardware scheduler that restricts the programmer's scheduling control. The scheduler in particular is optimized for throughput rather than deterministic execution, which is required by real time systems. Critically, tasks with strict timing requirements may suffer delays if long running, lower priority kernels are already resident on the device. The inability to natively preempt GPU kernels or to enforce strict task priorities makes a native utilization of GPUs difficult for safety critical, real time workloads. Particularily in the context of autonomous driving, the repercussions of latencies pose a safety concern.

1.2 Problem Statement

Rather than relying on native kernel launches, this thesis investigates the use of persistent threads to enable low latency execution, and develops the foundational framework needed to integrate coroutines into the system. Persistent threads are specialized kernels that are launched at the start of the application and remain active throughout the lifetime of the system. They function as a user level scheduler, continuously polling for work, executing tasks, and executing scheduling decisions.

Building on this framework, the central research question addressed in this thesis is:

How can a persistent GPU thread model be designed to support the implementation of GPU coroutines for predictable, low latency scheduling on real time systems?

As part of this research, the following aspects will be considered and evaluated:

- **GPU Stream Management:** Organizing and scheduling multiple GPU streams to enable concurrent task execution with memory transfers while maintaining predictable execution orders.

- **GPU Device Memory Management:** Efficient allocation, deallocation, and usage of GPU memory to ensure low latency access and avoid contention between tasks.
- **GPU Task Enqueuing and Dequeuing:** Mechanisms for submitting tasks to the GPU and retrieving results asynchronously.
- **Framework for Coroutines:** Designing a foundation for GPU coroutines that allows cooperative multitasking and the implementation of custom scheduling strategies on top of persistent threads.

1.3 Objectives and Contributions

1.3.1 Original Objective

The original objective of this thesis was to integrate an existing coroutine based GPU scheduling framework into an autonomous driving system. This integration aimed to evaluate the feasability of fine grained GPU scheduling within a complex, real time environment. Furthermore, by measuring scheduling latencies in the system, it would be possible to derive and refine strict timing guarantees.

1.3.2 Challenges and Scope Adjustment

Direct implementation of the coroutine framework proved infeasible within the available time frame. The system was originally designed for graphics rendering, relied on a complex compiler based architecture with little documentation, and required a separate build process. Combined with my limited prior experience in GPU programming and compiler theory, these factors made integration within the timeline challenging.

To simplify the problem, the focus shifted from coroutines to the underlying execution model, persistent threads. No suitable open source implementation was found that could be directly integrated into an autonomous driving system. Instead, a minimal, open source, custom persistent thread scheduler measuring overheads was found from which parts were taken to build a fully functional system. This new scheduler serves as a proof of concept foundation which provides long running GPU kernels to receive and execute tasks efficiently on which coroutines can later be implemented for real time systems. Furthermore it provides the framework on which GPU coroutines can yield and enforce prioritization between tasks, allowing exploration of how real time behaviors can be approximated within the constraints of the CUDA execution model.

1.3.3 Contributions

This thesis designs and implements a number of persistent thread components to enable the efficient execution of GPU code within persistent threads.

Task Management System

An execution management system was developed to allow tasks to be queued and executed independently of resident executing kernels. This system captures the full execution context of functions, enabling persistent threads to retrieve, schedule, and execute tasks without requiring new kernel launches.

Memory Management System

The execution context for each task includes its associated memory allocations. To reduce the overhead of repeated allocations and deallocations, memory is mapped into a running epoch buffer of preallocated memory assigned before the launch of the persistent threads. Tasks then operate within this preassigned memory region, eliminating the latency costs of frequent device memory management operations. This system further provides a logical partition of input and output buffers in order to reduce data interdependencies.

Concurrent Memory and Execution models

To support concurrent memory transfers between the host and device, as well as simultaneous kernel execution, the system leverages CUDA stream manipulation. This reduces interdependencies between data transfers and execution tasks, enabling higher overlap and better utilization of GPU resources.

GPU Task Coordination and Synchronization

The scheduler enables multiple thread blocks to execute distinct tasks concurrently, ensuring that no two blocks perform the same task simultaneously Furthermore, this system allows the host to efficiently schedule and queue tasks to any available thread block, maintaining high utilization of the GPU. By enforcing exclusive task execution, the mechanism prevents race conditions and ensures correctness across all kernels.

1.4 Thesis Outline

2 Related Work

Real time GPU programming models are a result of more autonomous systems and the increased hardware performance of GPUs in those applications. Programmers want to ensure these systems are predictabe. These applications, ranging from robotics to scientific computing, use GPU programming models to reduce kernel overhead while improving resource utilization and predictability. Although GPUs offer high throughput and timely execution important for these domains, they require non native programming models to ensure predictable execution models. Researchers have proposed various solutions to adapt GPU execution models to these real time constraints. The solutions fall into three categories: compiler driven frameworks, runtime scheduling frameworks, and manual implementations.

2.1 Compiler Driven Frameworks

Compiler driven frameworks focus on optimizing GPU workloads through automated code generation. These systems are based on persistent threads, which reduce kernel launch overhead, maximize on chip memory reuse. The compilers generate code using a specific DSL, which abstracts the low level device code. The compiler then parses the DSL into an abstract syntax tree. The abstract syntax tree allows the compiler to transform and optimize the code into an intermediate representation, which is transformed into device code JIT.

One such project, *Mirage* implements these ideas to improve large language model inference, by merging kernels into a singular megakernel. The singular megakernel is essentially a persistent kernel that eliminates extra memory copies between kernels, lowers kernel launch overhead, and retains memory on chip. Similarily, *Halide* provides a framework to automatically make scheduling decisions for users, by decoupling the algorithm from the execution schedule. The execution schedule is then determined through compiler optimizations through autoschedulers that requires minimal manual tuning. Lastly, *Luisa* is structured similar to the other projects, but offers performance benefits specifically for graphics and simulation workloads like ray tracing and rasterization. Luisa has support for acceleration strucutures, ray traversal APIs, and shader abstraction, allowing developers to high level rendering code while retaining low level performance.

Built on top of Luisa's execution model, *LuisaCompute-Coroutines* extends the framework to support coroutines built on persistent threads. Coroutines are expressed simply within the DSL using suspension points. These suspention points allow the device to preserve context and yield, allowing for fine grained scheduling. In particular this approach is efficient and leverages itself for real time systems due to the asynchronous coroutine programming approach.

2.2 Runtime Scheduling Frameworks

Beyond compiler based transformations, runtime based frameworks provide an alternative approach by enabling real time GPU scheduling. For example, *RT-GPU*, a runtime system, provides deadline aware scheduling of GPU workloads by partitioning GPU resources. Using a reservation based model, RT-GPU enables fine grained control over scheduling to ensure the task deadlines. Additionally, *ROSGM* is a further GPU management framework designed specificially for ROS 2 robotics systems. The ROSGM system interposes a layer to intercept the CUDA API calls to insert metadata to each GPU task. The API interception allows for custom deadlines and priorities that manage how GPU tasks are queued and issued to the device.

2.3 Manual Persistent Kernels

In addition to the other models, manual implementations support fine grained scheduling control specifically tuned to a single application. For example, in scientific computing, simulators like *HOOMD-blue* manually fuse multiple computation steps into a single persistent kernel. Furthermore, Jetson implemented GPU persistent threads to support their real time RedHawk Linux system. Unfortunately, this GPU persistent thread implementation is not open source. These implementations offer low-level control and high efficiency, but demand significant expertise in GPU programming.

2.4 Platform Integration: GPU Scheduling in Apollo

This thesis aims to integrate GPU real time scheduling into the open source autonomous driving platform Apollo. Apollo, developed by Baidu, relies on CyberRT, a real time platform to coordinate CPU task execution. CyberRT manages the different driving modules within the system and coordinates cooperative asynchronous scheduling using coroutines. The coroutine capability; however, does not extend to GPUs, which does not ensure their predictability.

Starting this thesis, I initially explored integrating LuisaCompute-Coroutines into Apollo, to provide the system with GPU coroutines to pair with the existing CyberRT CPU coroutines. Similar to the CPU coroutines, these GPU coroutines were intended to deliver the system predictable execution latencies, with the added benefits GPU persistent thread offer. However, due to several integration barriers, including incompatible build systems, sparse documentation, and time constraints, it became clear that the integrating this system into Apollo would not be feasible within the scope of this thesis.

Consequently, this work's focus shifted to a manual implementation of a GPU persistent thread scheduler using CUDA. This change allowed me the opportunity to study the low level aspects of GPU scheduling from both an architectural and programming perspective. While lacking the automation and abstraction of a compiler driven implementation, this manual approach allows for finer application specific implementations.

3 Background

This background chapter first examines the coroutine based concept from Luisa Compute Coroutines, then real-time systems and finally develops to the architectural and programming constraints of GPUs, particularly in the context of autonomous driving. The goal is to provide the necessary technical background to understand how GPU design influences system behavior and scheduling under real time constraints.

3.1 Luisa Coroutines

The first proposed approach to implementing a GPU scheduler for autonomous driving focuses on integrating the LuisaCompute coroutine platform into Apollo. The goal was to enable GPU coroutines within Apollo, allowing the autonomous driving framework to suspend and resume GPU kernel execution. This capability would allow the scheduler to better enforce bounded response latencies by directly scheduling the highest priority tasks at the appropriate time, rather than waiting for current GPU workloads to complete.

3.1.1 Coroutines

Coroutines are a form of asynchronous programming that enable cooperative multitasking between functions. Unlike thread- or process-level context switches, which involve greater overhead, coroutines maintain only a function-level context, allowing fast and lightweight task switching. This makes them particularly useful for enabling runtime kernel task switching on the GPU. Here, execution of a kernel can be suspended to allow another kernel to run, and then later resumed without blocking other work.

A coroutine suspends execution by capturing its current context, known as the continuation, which contains the execution state needed for later resumption [2]. Once suspended, another task can execute without overwriting or disrupting the saved kernel state. When the interim task completes, the coroutine can continue exactly where it left off by restoring its continuation. This ability to strategically pause and resume execution makes coroutines well suited for real time GPU workloads, where rapid switching between concurrent tasks can help meet hard deadlines without delay.

3.1.2 CPU Coroutines

Unlike threads or processes, which require relatively expensive context switching, coroutines enable simpler and faster context switches between functions. This efficiency arises from the way function calls are handled at the CPU level, particularly on x86 architecture. To switch a function, the GPU merely needs to save its state and then jump to the new function instruction address.

In x86 calling convention, when a function is called, the CPU pushes the return address, the address of the next instruction to execute, onto the stack and jumps to the called function's instruction address. The called function accesses its variables via the stack and registers. Registers are categorized into two groups: volatile, caller saved and non-volatile, callee saved. Volatile registers may be freely modified by the callee without restoration, whereas callee saved registers must be preserved and restored before the function returns.

When the function executes a return instruction, the CPU pops the return address off the stack and continues execution from that point. To yield a function, the minimum context that must be saved and restored includes the callee saved registers, since these are guaranteed to be preserved across function calls. Additional state, such as local variables or other registers, can be manually saved to the stack to be restored later when the coroutine resumes.

Consider the following CPU coroutine code taken from the Apollo project.

```
ctx_swap:
    pushq %rdi
    pushq %r12
    pushq %r13
    pushq %r14
    pushq %r15
    pushq %rbx
    pushq %rbp
    movq %rsp, (%rdi)

    movq (%rsi), %rsp
    popq %rbp
    popq %rbx
    popq %r15
    popq %r14
    popq %r13
    popq %r12
    popq %rdi
    ret
```

Listing 3.1: CPU Coroutine

The CPU coroutine implementation uses the context switching function, ctx_swap, which swaps the continuations by saving the current execution context and loading a new one. As previously stated, standard calling conventions uses registers to pass the function parameters to the function. In this case, this function accepts two parameters, %rdi and %rsp, each representing an addresses used to store and retrieve the coroutine continuations. %rdi is used by the first half of the function to store the current execution state, while %rsi is used in the second half to restore and load the coroutine continuation.

The first section of ctx_swap, up to line 10, is responsible for saving the current continuation. This is achieved by pushing all callee saved registers onto the current stack, preserving the processor state. To resume the continuation later, the memory location of the continuation is stored in %rdi, by copying the current stack pointer into the register. This value allows the function to restore the original context during subsequent switches.

The second section loads the new coroutine context, by reversing the first sections actions to saving the state. The stack pointer is updated with the new continuation, such that all the saved registers can be retrieved and execution can resume. Once within the correct stack frame, all callee saved registers are restored in the reverse order of their saving, effectively loading the processor registers with the new coroutine's state. The callee saved register states are thus preserved across coroutine function calls, with further saved variables easily accessible as local variables on the same stack frame.

In comparison to traditional context switches involving processes or threads, coroutine context switching via ctx_swap operates with significantly lower overhead, resulting in a faster execution. Furthermore, these continuations benefit from the user explicitely defining the suspension points. User defined suspension points typically involve much smaller contexts than preemption points, as they align more closely with convenient locations within the executing program. Consequently, coroutine contexts can often be kept minimal by leveraging only these carefully chosen points that require smaller saved states.

3.1.3 GPU Coroutines

In contrast to CPU coroutines, the GPU coroutines are more complex due to the large number of concurrent threads and concurrent warps that are executing and the specifics of GPU programming, especially in regards to the stack management. GPUs launch kernels that can not be interrupted by the system of programmer throughout the duration of its lifetime. The kernel function will saturate the number of warps and threads that were allocated to it until termination. After each kernel terminates the state afterwards is not preserved for the next incoming kernel. Given these restrictions, the kernel must manually yeild execution using a user space sheduler.

Inlining

Attempting to implement the CPU style coroutines using the ctx_swap function does not work, given that the GPU handles function calls differently than the GPU. CPU function calls store the call stack within the stack, by pushing instruction pointers that can be popped with ret. The GPU does not use these, but instead saves stack pressure by aggressively inlining function calls. Inlining function calls allows the GPU to reduce overhead when beginning a new function as that function code is already readily available and removes the need for a cpu style stack frame.

In particular, GPUs strongly optimize away from call stacks and provide only minimal support for them. Originally, CUDA did not allow recursive functions and only with around CUDA 5.0 did they support them. This support has to be manually implemented using the nvcc flag, lstinline[language=cuda]'-rdc=true'. This recursion comes with limitations of restricted stack size and added performance overhead. Attempting to implement deep call stacks leads to exponentially large instruction Unfortunately for tasks dependent on deep call stacks such as recursion, this leads to exponentially large instruction memory.

Persistent Threads to support Coroutines

Given that call stacks are not effectively or efficiently supported in CUDA, the GPU needs to provide a user level scheduling mechanism to control the execution decisions. Due to the nature of kernels executing until completion, the GPU coroutine scheduler needs to be based on a persistent kernel, which can schedule coroutines. These coroutines themselves need to have suspension points to manually give control back to the scheduler in order to execute the next task. Saving every register value for every thread across every coroutine is to computationally expensive, so the coroutine contexts need to be managed locally and saved in global memory to free up limited SM resources for new tasks.

LuisaCompute-Coroutine

The implementation developed in LuisaCompute-Coroutine enables GPU coroutines by providing a coroutine based API that acts as a JIT compiler for generating GPU kernels at runtime. LuisaCompute offers a DSL embedded in C++, allowing programmers to explicitly define coroutine suspension points using standard C++20 coroutine syntax, such as co_await. These coroutine constructs are not executed immediately but are instead interpreted symbolically into an **AST!**.

At runtime, LuisaCompute builds a symbolic representation of the kernel's control and data flow in an abstract syntax tree (AST), through operator overloading and expression tracking. This symbolic trace is then lowered into an intermediate representation (IR), which encodes the coroutine as a state machine, capturing both the control flow and the coroutine's execution context. The resulting IR is compiled into GPU code, such as PTX for CUDA, using LuisaCompute's JIT backend. Once compiled, these coroutine-based kernels are dispatched and executed on persistent GPU threads, which maintain their state across kernel invocations and facilitate efficient asynchronous execution and task switching on the GPU.

As part of this work, I initially explored the possibility of integrating LuisaCompute coroutines into the Apollo autonomous driving platform. However, due to the lack of documentation and my limited understanding of both Apollo and LuisaCompute in both the implementation of tasks into Apollo and the underlying abstract syntax tree (AST) and intermediate representations (IRs) used in LuisaCompute's JIT compilation system, I struggled with dependency issues and was ultimately unable to complete the integration. Rather than continuing down this path, I decided to simplify the problem and shift focus toward developing a custom implementation of persistent GPU threads, which still reduce the overhead involved with launching GPU kernels.

3.2 Real Time Systems

Real time systems are designed with strict timing constraints to ensure predictable and safe behaviour. [3] These constraints are expressed in terms of the systems ability to meet task deadline, categorized as either soft or hard. Hard deadlines are critical to the safety of the system. Missing these deadlines results in potential system failure or unsafe conditions, such as collision avoidance or brake activation. To ensure the safety of the system, these deadlines need to be prioritized over the less critical soft deadlines. Soft deadlines, for example route planning or navigation updates, are deadlines that should be ensured, but do not lead to system failure if not met. Delays here may lead to suboptimal paths or low display responsiveness, but do not risk safety. Real time systems ensure schedule tasks by preempting soft deadline tasks for high priority hard deadline tasks.

3.3 Integration of GPUs in Autonomous Driving Systems

Early autonomous systems used purely CPU based compute engines for the control and execution of tasks within the system. For example, the Stanley autonomous car, which won the 2005 DARPA Grand Challenge, used 6 Pentium computers running

Linux to run the system. The CPU centric approach allowed systems to meet real time requirements using known strategies, similar to how Apollo manages CPU tasks today. These systems could easily use a real time operating system along with conventional programming mechanisms for synchronization and thread level cooperation techniques. However, as machine learning, in particular deep neural networks, became central to perception and decision making, the computational demands grew beyond the capabilities of CPUs alone. GPUs started to become incorporated to meet these increasing demands, particularily for high throughput workloads like image classification, object detection and sensor data processing.

With the rise of deep learning and CNNs for perception, GPUs began to replace CPUs as the core compute engine. One of the earliest projects using GPUs, was NVIDIA in 2015, which used a GPU to train a CNN that could steer a car end to end from raw camera input. While early self driving systems functioned without GPUs, advances in deep learning made GPUs nearly indispensable. Today, nearly all modern autonomous platforms, rely on GPUs for perception, planning, and sensor processing tasks.

The rapid implementation of GPU into these systems depends heavily on their massive performance speedups in machine learning tasks and in particular deep neural nets. Deep neural nets, structured like neurons in the human brain, learn patterns through layers of weighted nodes. These weighted nodes perform large numbers of matrix multiplications and transformations, operations that are highly parallelizable and therefore suited for GPU architectures. As a result, these GPUs, which are highly parallelized significantly accelerate both the training and inference, critical for these systems.

GPUs support a massively parallel architecture and high memory bandwidth, which perfectly meet the demands of deep learning tasks. Unlike CPUs, which optimize for sequential instruction execution and low latency branching, GPUs are designed to handle large batches of matrix and tensor operations simultaneously. Additionally, modern GPU architectures provide specialized cores, such as tensor cores in NVIDIA GPUs. These cores are explicitly optimized for mixed precision matrix multiplications, a core operation in most machine learning models. Autonomous driving systems can thus use better more complex models and execute inference faster by offloading intensive compute tasks to the GPU.

3.4 GPU versus CPU Architecture

GPUs deliver the vast increase in throughput over CPUs, by simplifying the thread context in order to afford greater parallelism. They were originally developed to accelerate graphics rendering, a task heavy in parallizable computations, which require only a very simple control overhead. The architecture thus adapted to support more and more threads to support larger graphics matrices. These tasks required very little control overhead, which allowed the GPU tasks to be very simple, but slower than pure CPU cores. Typical workloads designed for CPU are based on sequential workloads, such as human input or complex logic. Sequential workloads require complex thread overhead to speed up branches and I/O, using techniques like prefetching, branch prediction, and out of order execution. In order to achieve greater performance, CPUs dedicate a "significant portion of transistors to non computational tasks like branch prediction and caching". GPUs can forgo the control overhead in favor of increasing arithmetic intensity [4].

Consider the following graphic Figure 3.1, which highlights the difference in thread complexity.

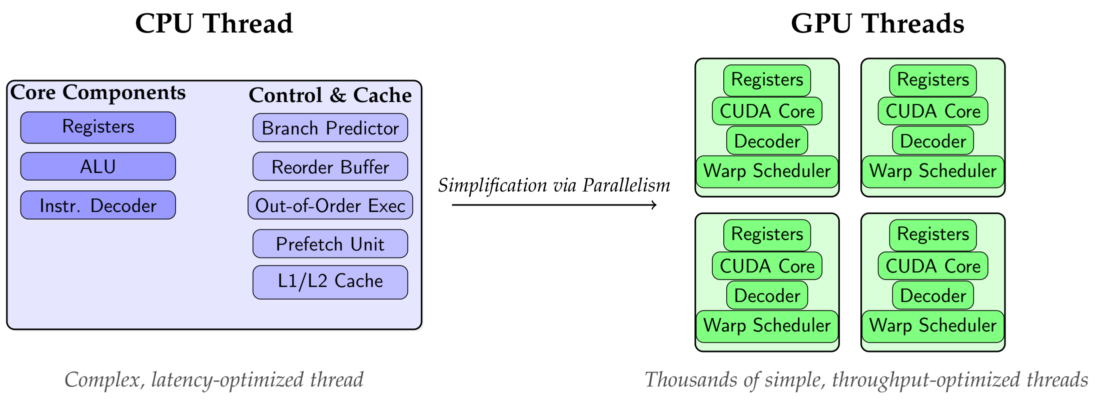

Figure 3.1: CPU vs GPU Thread Architecture

Although core components are named differently, both CPU and GPU threads work similarly with an instruction decoder, registers and an arithmetic unit. The differences arise when trying to maximize a single control flow. The CPU will emply instruction prefetching, out of order execution, speculative execution, and complex branch prediction to maximize performance of a single thread. In contrast, GPU threads execute instructions in order, rely on manual prefetching, and use a simpler, conservative branch predictor, which limits their ability to optimize threads. Instead of optimizing single threaded performance, GPU achieve high throughput by executing thousands of lightweight threads in parallel, amortizing latency across them. For example, threads are grouped into warps that share a common instruction stream, executing the same instruction simultaneous across different to. Furthermore, GPUs allow memory access coalescing to reduce bandwidth overhead. This design favors workloads with high data parallelism, enabling the GPU to hide memory and execution latencies through massive concurrency rather than complex control logic.

The additional complex logic that CPUs use to improve single threaded applications outperforms the single threaded GPU applications, as seen by the following comparison of single threaded matrix multiplication in Figure 3.2 and Figure 3.3.

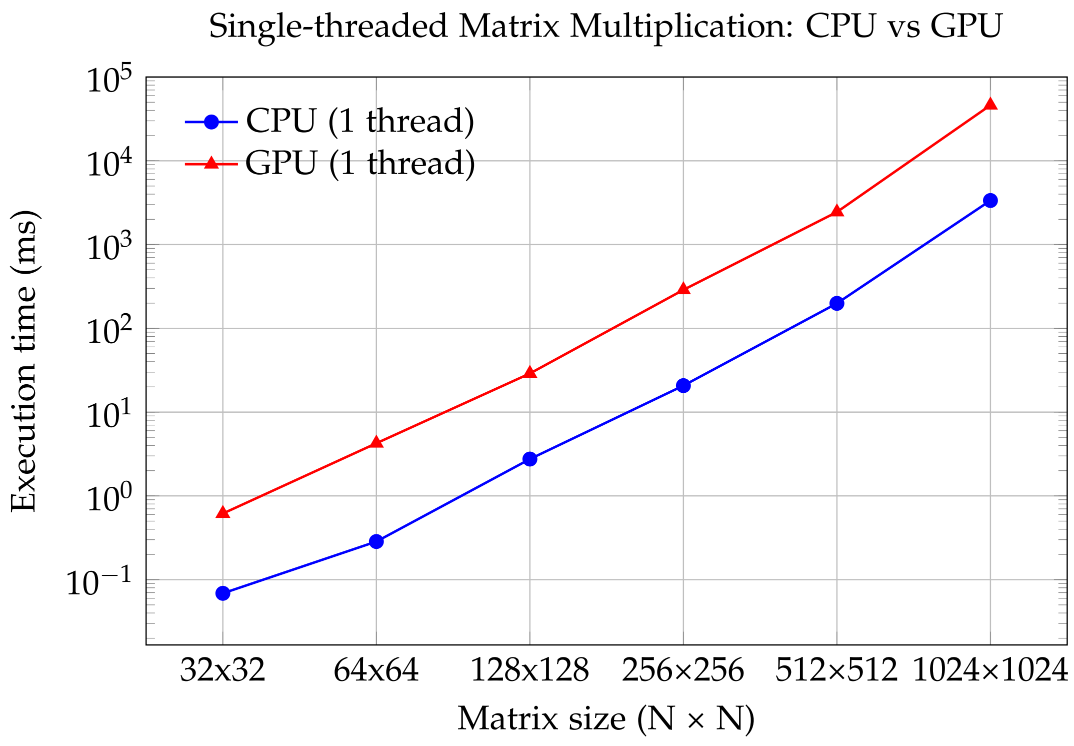

Figure 3.2: Single threaded Matrix Multiplication Execution between CPUs and GPUs averaged over 10 executions

Matrix Size (n × n)	**CPU Time (ms)**	**GPU Time (ms)**	**Speedup (GPU/CPU)**
32 × 32	0.068705	0.615584	11.970438
64 × 64	0.285104	4.249685	15.554119
128 × 128	2.751817	28.923530	11.419879
256 × 256	20.706290	287.933700	13.906730
512 × 512	198.716200	2446.466000	12.323010
1024 × 1024	3356.762000	46097.410000	13.745850

Figure 3.3: Data Matrix from Figure 3.2

As seen in Figure 3.2 and Figure 3.3, GPUs struggle in applications that fail to utilize the architectural parallelism. For each of the matrices tested, the CPU is on aver around 13 times faster than the GPU due to prefetching, a higher clock rate, and branch prediction, despite running the exact same algorithm. These results, show that any complex control and management tasks should be kept on the CPU rather than the GPU, due to the performance difference. Furthermore, the GPU should only be used in place of the CPU, when the application is parallelizable and lacks complex control flow logic. Attempting to programm a GPU scheduler must consider the hardware when determining where scheduling logic is run to not unnecessarily introduce latencies. Understanding both the hardware and programming style for GPUs is crucial in order to achieve scheduling based performance gains.

3.4.1 GPU Architecture

The NVIDIA Tesla V100 GPU, based on the Volta architecture, was selected for this project due to its availability and suitability for high performance computing design. At a high level, the GPU architecture supports 5120 CUDA cores, 640 Tensor Cores and delivers up to 7 TFLOPS of double precision performance. The current system has 16 GB of VRAM and provides 900GB/s of bandwidth between on-chip memory and device memory. The GPU itself connects to the host system over PCIe, with a theoretical maximum of 16GB/s for transfers between CPU memory and VRAM.

Despite these impressive specifications, performance is often constrained by the disparity between compute throughput and memory transfer rates. While the GPU can execute arithmetic operations at extremely high speed, fetching and storing data from VRAM is comparatively slower, which can stall execution if memory access is not optimized. Even slower are the memory transfers between host and VRAM and must be minimized to avoid further bottlenecks.

Understanding this imbalance is fundamental to design of efficient scheduling and memory access patterns, which are essential to fully exploit the device capabilities. The following analysis of the V100's hardware architecture, will provide both the reasoning and implementation fundamentals for designing an efficient GPU scheduling strategy.

Thread Hierarchy and Execution Model

From the programmer's perspective, the GPU appears as an array of independent, highly parallelized processors, called SMs. Each SM receives work in the form of CTAs, blocks of threads executing the same instruction code, which define the organization and grouping of threads for execution. The executing CTA is subdivided into Warps, the smallest execution unit on the GPU, each consisting of 32 GPU threads executing instructions in lockstep. The lockstep execution ensures all threads within a warp execute the same instruction simulataneously. The lockstep execution model simplifies scheduling and dispatching threads to compute units to simplify the GPU design and enable further parallelism.

While lockstep execution enables efficient SIMD style throughput, it also introduces a potential performance hazard. If threads within a warp follow different control flow paths, these threads diverge, and the GPU is forced to execute these different threads sequentially with respect to one another. This serialization stalls part of the warp using masks, which disable the write back of the compute units. In effect, thread divergence, reduces the effective parallelism and, degrades performance.

SM Architecture

The Tesla V100 GPU SM architecture, contains 4 processing partitions, each with its own complete execution pipeline. These partitions share an L1 instruction cache as well as a combined L1 data and shared memory cache, enabling thredas within different warps of the same CTA to efficiently access shared instructions and data. Within the each processing partition there is an L0 instruction cache, a warp scheduler, a dispatch unit, and multiple execution units. An in depth view of each processing partition's architecture is provided in Figure 3.4, taken from the NVIDIA Volta Whitepages.

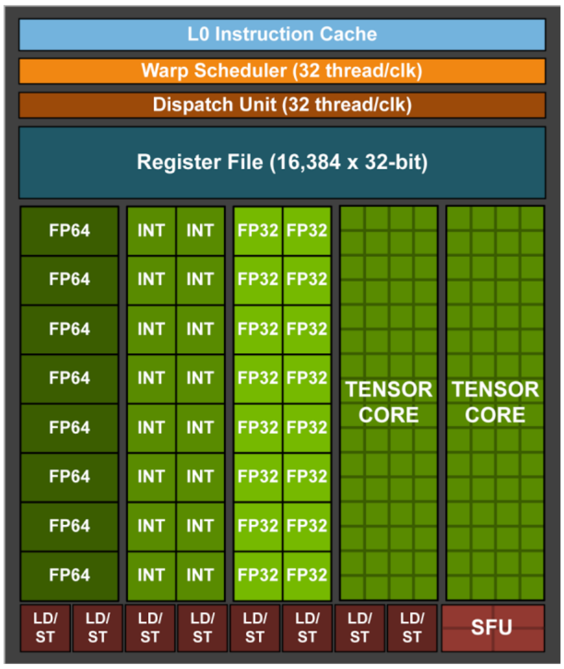
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
FP64
INT
FP32
TENSOR CORE
LD/ ST
SFU

Figure 3.4: SM Processing Partition Architecture taken from the Volta Whitepages

Every clock cycle, the warp scheduler selects a warp of 32 threads to issue to the dispatch unit, which dispatches decoded instructions to the appropiate functional units. If there are not enough execution units of the required type for a given instruction, the instruction is queued. Depending on the current queue and delays, such as global memory accesses or dependencies, the warp scheduler will interweave different instructions from other ready warps, ensuring that the execution units remains busy. This thread interweaving allows GPUs to hide latencies and resource contention through thread oversubscription. Beyond, the warp scheduler and dispatcher, each processing partition contains a number of execution units. In particular, there are Tensor Cores for deep learning, 64 bit-floating point (FP) cores, Load/Store (LD/ST) units, a register file, and SFUs for mathematical functions such as sine and square root.

SM Scaling and Thread Concurrency

Compared to CPU hardware threads, GPUs scale far more aggressively to support further parallelism by supporting a larger number of threads. On the Tesla V100, there are 80 individual SMs, each capable of supporting up to 64 resident warps. With 32 threads per warp, this yields $64 \times 32 = 2048$ threads per SM. Across all 80 SMs, the theoretical maximum concurrency is $80 \times 2048 = 163{,}840$ resident threads.

For comparison, the Intel Tiger Lake i5-1135G7 CPU in my laptop has 4 cores with 2 hardware threads each, for a maximum of 8 concurrent hardware threads. Even high end server CPUs, such as the Intel Xeon Gold 6148, only support 20 hardware threads. While the GPU oversubscribes the number of warps and threads to hide latencies, the total number of dispatch and scheduling units allow for a maximum of $4 * 32 * 80 = 10840$ instructions issued per cycle, when the hardware is fully saturated.

In practice, this theoretical maximum is rarely achieved due to resource bottlenecks. All threads within an SM share the same on-chip L1 data and shared memory cache, as well as the 256 KiB on-chip register file (16,384 × 32 bit registers per processing partition, with four partitions per SM). SM threads all share the same on-chip L1 data and shared memory cache and the 256KiB on-chip register file (16,384 × 32 bit registers per processing partition, with four partitions per SM). Depending on the kernel's resource usage, high register pressure will cause the kernel launch will fail. Furthermore, if multiple thread blocks with high shared memory demands get scheduled to the same SM, they will saturate the L1 data and shared memory cache and force frequent evictions and write backs to global memory. These hardware limits must be carefully considered when mapping tasks to SMs.

The built-in hardware scheduler normally either handles these constraints or fails the kernel launch. Implementing a custom scheduler on top of the hardware scheduler, requires consideration of these GPU limitations. Forcing new scheduling behavior may conflict with or be constrained by the hardware scheduler's own mechanisms.

Scheduling and Task Mapping

The specific mapping of tasks to SM, TPC, and GPC is determined natively by the proprietary hardware scheduler, the GigaThread Engine. While the exact documentation is not public, this module maps CTAs to the individual SMs based a multitude of factors: hardware resources, parallelism, priorities, and dependencies. Similarly, the global memory and L2 cache utilization are determined by the hardware and transparent to the programmer. After the CTA gets mapped to the specific SM, the device code then executes till completion without interruption.

The CTAs is entirely managed by the SM on which it is currently executing. As shown previously the SM with its own execution pipelines, register files, shared memory and scheduling units. For SMs to communicate with one another, they must use either the global on-chip device HBM2 or through the global L2 cache which is shared and coherent across all SMs. Although these memory accesses allows individual SMs to communicate with each other, accesses require hundreds of cycles, which introduce further latencies when compared to local SM L1 memory caches. Ideally, the SMs execute independently of one another and accumulate answers in global memory, skipping the high memory latency accesses of coordinating synchronous work.

3.5 GPU Programming using the CUDA API

NVIDIA GPUs are intended to be used as computational accelerators for a host system, which manages and schedules tasks using CUDA. In this model, the GPU acts as a standalone processor with its own independent memory and execution pipelines. For tasks to be scheduled on the GPU, the host process must first launch the process using the CUDA API. The CUDA API is an extension of C++, which allows the CPU to interact with the GPU. Using CUDA, the host invokes device functions, called kernels, by specifying the number of threads and the memory parameters. These kernel calls are asynchronous, meaning that once the host issues the command, the GPU executes it independently, while the CPU can continue with other work or synchronize later as needed.

3.5.1 Kernel Launches

The task of launching and running device code begins from a kernel launch, which passes the function, its parameters, pointers, and the grid and block dimensions to the GPU. The launch configuration specifies the number of CTAs and their logical organization into dimensions. Every block is then mapped to a single SM and is constrained by that SM's hardware resources, including the maximum number of threads, registers, resident warps, and available shared memory. If no available SM can meet these requirements, the kernel launch will fail. Conversely, if a CTA does not fully saturate the hardware resources of an SM, additional further blocks may be scheduled concurrently on the same SM.

3.5.2 Memory and Bandwidth Considerations

Both the GPU and CPU maintain distinct memory regions for different hardware architectural and performance reasons. CPU memory is optimized for low latency access to support serial and branch heavy workloads, while GPU memory is optimized for extremely high bandwidth to sustain thousands of parallel threads, accepting higher latency in exchange. These memories are connected by an interconnect such as PCIe or NVLink, whose bandwidth is far lower than that of the GPU's local memory, making constant access to host memory impractical without significant performance loss. By keeping data local, each processor can operate at full efficiency, and the GPU can execute kernels independently without frequent synchronization with the CPU. This separation is thus a deliberate design choice to match each processor's optimization goals, avoid bottlenecks from the limited interconnect speed, and enable independent execution.

As the memory regions are distinct, for the GPU to support the CPU with tasks, the CPU needs to use CUDA to allocate and send data to the GPU memory. When transferring data from the host, the CUDA API does not have access to the CPU's disk memory and requires the memory to be pinned to the CPU's RAM. While the CUDA runtime can perform this pinning automatically, host arrays allocated directly in pinned memory using `cudaMallocHost()` or `cudaHostAlloc()` skip the added step of pinning memory and enable faster transfers. Particularity in applications that are bandwidth bottlenecked, this added transfer latency is significant. If the pinned memory is too large; however, it restricts the memory availability for the programs currently running on the CPU, which may degrade performance by forcing the swapping of memory to disk storage.

In CUDA, understanding how memory is transferred and managed across the host and the device is crucial for optimizing performance. For the programmer, the device memory is partitioned into main memory, the VRAM, with a size of 16 GB on the Tesla V100 architecture, as well as on chip memory. The programmer can decide between three different models of memory management `__constant__`, `__device__`, and `__shared__` memory. Both constant and device memory are allocated to the global memory, with constant memory only being writeable by the CPU and allowing faster access times due to the reduced coherency required. The shared memory is shared among all warps and threads of a given SM in the L1 data and shared memory cache. This memory is far more performant than device memory, but limited in size, due to its location on chip.

3.5.3 Memory Coalescing

For the GPU to coalesce memory operations and enable SIMD-style execution across multiple data points, each thread maintains registers that store its execution context, including its position in the grid. In PTX, these special registers provide unique identifiers such as threadIdx and blockIdx, which indicate a thread's coordinates within its block and a block's coordinates within the grid. These identifiers are essential for structuring parallel computations and optimizing memory access patterns. For example, when threads within a warp access consecutive memory locations, the hardware can coalesce those accesses into a single memory transaction, significantly improving throughput.

3.5.4 Example Kernel Launch

Consider the following example program, which allocates device memory and launches a kernel consisting of one block and 32 threads.

```
__global__ void increment(float *x) {
    x[threadIdx.x] += 1.0f;
}

int main() {
    const int N = 1024;
    float h_x[N];
    for (int i = 0; i < N; ++i)
        h_x[i] = i * 1.0f;

    float *d_x;
    cudaMalloc((void**)&d_x, N * sizeof(float));
    cudaMemcpy(d_x, h_x, N * sizeof(float), cudaMemcpyHostToDevice);

    increment<<<1, 32>>>(d_x);

    cudaMemcpy(h_x, d_x, N * sizeof(float), cudaMemcpyDeviceToHost);
    cudaFree(d_x);
    return 0;
}
```

Listing 3.2: Simple CUDA Kernel

The code block above depicts the launching and execution of a simple GPU kernel that demonstrates the memory allocation scheme used for executing kernel code. The kernel itself is the execution of the GPU device program denoted by `__global__` function, while the `<<<_, _>>>` syntax enables the programmer to specifically partition their execution tasks across waiting threads. The first value in the `<<<1,32>>>` determines the block dimensions, which are either given as an array or a 3 dimensional tensor. Similarly, the second value determines the thread dimensions in the same format as the block dimension. In particular, this code allocates a singular block with an array of 32 threads, which completely saturates a singular warp. These dimensional vectors allow the different threads to maintain lockstep execution, while processing different values of the same array, as seen by using the dimensional properties assigned to the individual threads by the runtime system.

The main function, executed by the CPU or host, initializes the parameters, executes the kernel and then copies the memory back. The host array, `h_x` is allocated to the stack, which exists only in CPU memory, which needs to be passed to the GPU. Passing the array by value, something common in C++ code, seems at first the most simple; however, poses two seperate issues. Firstly, when passing arrays as parameters, they decay to pointers, which CUDA forbids, as the pointer passed to the device does not have any meaning. Secondly, if the array were wrapped in a struct and passed to the function to circumvent the first issue, the array would be allocated to every single thread independently. In the example above, the array would be allocated 32 times, each independent from one another, taking up further memory bandwidth and both on chip and global device memory. In this case, each individual thread, would get the array passed by value, leading to a total 32 threads * 1024 floats * 4 bytes per float or 128KiB. Instead, them memory is allocated in the device memory, transferred once and each thread recieves only the device pointer `d_x`, which can be used to copy the results back to the host.

3.5.5 CUDA Streams

CUDA API calls are queued to the GPU using cuda streams, which enforce the execution order of tasks. `cudaStream_t` defines a command queue for the GPU, which is similar to a Linux file pointer in that it returns an index referring to the specific allocated stream. Each stream allows the queuing of operations such as kernel launches, memory copies, and memory set operations. Commands submitted to the same stream are executed sequentially in the order they were issued, ensuring deterministic behavior within that stream. Multiple streams, however, can run concurrently, enabling overlapping execution of kernels and memory operations to maximize GPU utilization and improve overall performance. By carefully managing streams, developers can optimize task parallelism and resource usage on the GPU.

The Tesla V100 GPU has two seperate hardware copy engines for copying data from the host to the device and back. The copy engines support the transfer in both directions, with one engine specifically being allocated for the unidirectional D2H transfer and the other for H2D. Using only one stream for multiple kernels fails to maximize the device memory bandwidth. For example, consider the launch of two independent kernels, kernel A and kernel B, each on the same CUDA stream. Both A and B, allocate and copy memory onto the device, schedule their kernels and then copy the results back. Regardless of the ordering of API calls, both kernels can not run simultaneously or use both H2D and D2H copy engines simultaneously.

To best utilize the hardware and ensure correct results, a different stream need to be used for each synchronous kernel launch. Each kernel must remain in the same stream as the memory copys related to that stream in order to ensure the arguments used by the kernel are accurate and that the results are not prematurely returned. By placing each kernel in its own stream, the CUDA API, depending on the time and specific situation run multiple kernels, run kernels and memory transfers, or perform both D2H and H2D API calls simultaneously.

4 Design and Implementation Requirements

Persistent GPU threads allow the hardware resources to be partitioned and reduce kernel execution and scheduling overhead and are fundamental for other applications seeking to enhance GPU scheduling. Many other applications such as coroutines or mega kernels are based on persistent threads as a means of reducing kernel launch overhead and allow further control in application specific scheduling decisions, otherwise black boxed by the cuda api. Running persistent threads essentially allows the memory and system configurations to be loaded in ahead of runtime and reduces overhead during execution.

Failing to implement LuisaCompute's coroutines into Apollo, the goal shifted to finding a manual GPU scheduling functionality to implement and use. In restricting the scope from GPU coroutines to simply implementing persistent threads into Apollo, this thesis attempted to find an open source implementation which would allow the fine grained scheduling controll specific to Apollo. Unfortunately, given that most persistent thread implementations are highly specific they are not open source, which led to selecting an implementation that required more work to make it feasible. The only open source persistent thread implementation that was readily available was LightKer, a research project, which measured the hypothetical speedup of using persistent threads over sequential kernel launches. Seeking to implement LightKer into Apollo first required a complete restructuring of the code base to support a real time system.

The LightKer implementation itself intended to measure the performance difference between sequential kernel executions versus a persistent kernel implementation. This implementation constructed simple trivial kernels and tested the overhead difference between calling them explicitely from the host in kernel launches versus implicitely in the persistent kernel. Unfortunately, this application does not support variable tasks or memory transfers at runtime, necessary to pass arguments and results back and forth. As such, the implementation only succeeds in measuring latency differences between scheduling tasks using a device side while loop versus explicit kernel launches. However, the LightKer implementation does provide a simple framework for using a GPU-Host mailbox for scheduling kernel tasks as well as a helpful persistent kernel launch structure.

4.1 Architecture

From a high level, the architecture implemented into this project appears as follows, with the GPU-Host Mailbox, being used from the Lightkernel project as well as the internal structure.

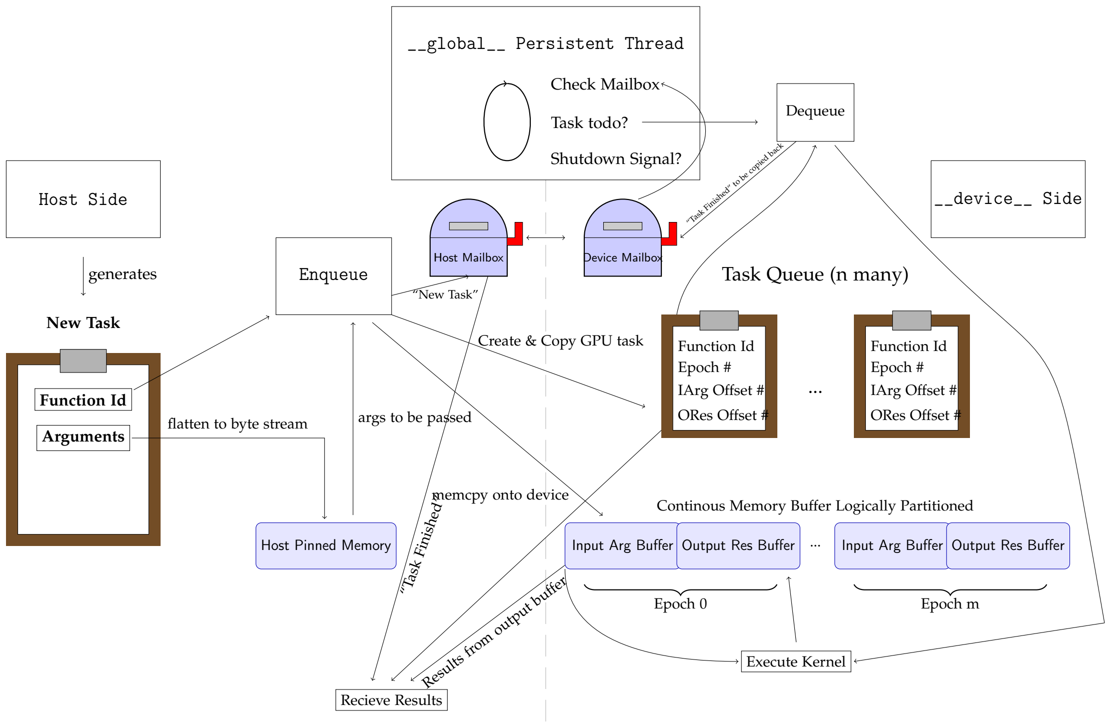

Figure 4.1: Persistent Thread Architecture implemented into LightKer

At a high level, there are three important components: the task queue, memory buffers, and the persistent threads themselves. The task queue is managed, controlled, and allocated by the host and provides the arguments and tasks for the GPU to execute. The implementation of the task queue gives the programmer fine-tuned implementation oppertunities to manually design and schedule workloads. The memory buffers provide an epoch based staging area for input arguments and output results as well as persistent memory for the further extension of coroutines. Lastly, the persistent threads are the fundamental execution units executing the GPU code.

The Figure 4.1 depicts these individual components in a logical program architecture overview. The left side of the graphic shows the host side code and methods, which allow the enqueuing of tasks and launching of the persistent thread kernel. On the right hand side of the graphic, the actual device side code allocation and kernel execution is depicted which allows the processing of memory and write back of results to the memory buffers. The mailboxes are constantly enqueuing and dequeuing new tasks throughout the execution of the kernels.

4.2 Implementation

GPU kernels are essentially device side functions launched by the host process, which execute with parameters and a logical grid of threads to be distributed across the SMs. In order to replicate both the parameterization and execution model without explicitely launching new kernels, the persistent threads must maintain a mechanism for storing function pointers and their associated arguments to assign work to idle compute resources dynamically. Similar to other persistent thread implementations, this project implements a task queue, that captures the full execution context required to execute the GPU code. As shown in the Figure 4.1, the task queue is represented as an array of clipboards each encapsulating a function context. Beneath the task queue, Figure 4.1 also depicts the staging area used for memory management. The memory management is implemented using an epoch based allocation strategy. These memory buffers serve as a real-time memory management structure for GPU kernels, reducing the need for direct runtime memory allocation while enabling in-place memory reuse for tasks.

Based on the results in Figure 3.2, any sequential scheduling mechanism executing directly on the device would incur significantly performance penalties when compared to a CPU based approach. As a result, both the task queue and the associated memory buffers are managed from the host side. This includes deteriming the active epoch and assigning task indices within the corresponding memory or queue structures. The host is responsible for copying all required parameters into device memory and delegates only the dequeuing and execution of tasks to the GPU.

Furthermore, standard GPU kernel launches specify the thread configuration and grid layout of GPU threads executing the code and their physical placement in the architecture. The GPU automatically decides the execution placement of the kernels from the Gigathread Engine during the launch of GPU code from the host. As the persistent threads are already launched at program start, the configuration remains the same throughout the lifetime of the persistent thread. Therefore these persistent threads can not support variable launch configurations at runtime without terminating the kernel and restarting a new kernel with different launch configurations. However, multiple different persistent kernels can be started with various kernel launch configurations, each with different task queues, or through code refactoring the kernels can be adapted to the existing GPU thread block organization.

4.2.1 Task Queue

The task queue functions as a cache like buffer between the host and the executing GPU code, enabling the asynchronous enqueuing of tasks. Rather than relying on the cuda driver to dynamically partition the GPU and assign tasks to threads, task parameters are instead written into a pre-allocated memory region. This memory acts as a staging area and remains allocated throughout the duration of the persistent thread kernel, with periodic cleanup.

Each task entry defines its execution context through an explicit function identifier and its associated parameters. This entry acts exactly like a coroutine continuation, which stores the necessary state to resume execution within the function. The GPU kernel, launched with persistent threads, enters a loop in which each block repeatedly dequeues and executes tasks.

Within this loop, each GPU block retrieves the next task from the queue, processes it, and stores its results. The task queue maintains both input and output buffer offsets for each tasks, allowing blocks to fetch parameters and write results. When a task is selected for execution, its parameters are deserialized from the input buffer, converted into an internal representation, and used to invoke the corresponding device function.

Upon completion of the task, the GPU block writes the results to the specified output buffer offset. This scheme works because the CPU knows the exact size of the input and output arguments, knowledge it already must have for issuing cudaMemcpy operations.

Once a the block finishes executing a task, it signals the host that the task is complete and then continues processing the next available task in the queue.

4.2.2 Function Pointers

To execute new functions from the persistent threads, the task queue needs to be able to reference the specific function. Generally referencing functions on a CPU requires only the function pointer to execute the code defined at that memory location. When GPU functions are compiled, the device code lives in the GPU address space and is not accessible from the CPU. The CPU only has access to functions denoted by the `__global__` keyword, which allows the execution of GPU kernels, not enqueuing of GPU functions. In order to be able to access and run the functions specified by the CPU, the task queue supports a lookup table to map integers to specific functions. The lookup table allows the host to memcpy in function ids to the task queue when enqueuing new tasks.

4.2.3 Function Parameters

When the CPU assigns tasks to the GPU, it passes either allocated GPU memory pointers or explicit parameters. These explicit parameters then get propogated to all the individual threads executing the kernel code, resulting in greater api memory overhead. When enqueuing new tasks to the task queue, the memory has to be transfered at runtime before the device function calls.

The GPU task in the queue originally had a pointer to the allocated memory and upon recieving compute resources would schedule the task with the memory to the individual persistent thread. Unfortunately, this method is dependent on the specific task and parameters and consumes variable memory requiring further pointers to GPU memory. In order to consolidate the memory pointers, the task queue was simplified to contain only allocated memory pointers in order to automatically load kernel memory.

In this method, enqueing the GPU tasks forces the programmer to streamify the data and automatically load the memory into preallocated memory partitions. The task queue then only consists of the actual memory partition pointers, both start and end. Executing a task then requires the interpretation of the memory and then the loading of it into the device function. Manually extending this method allows the user to manually allocate more memory than is needed and use that memory to yield and run coroutines.

4.2.4 Memory Model

In order to provide the incoming scheduled kernels a staging area for allocated memory, the implementation contains a running epoch memory model. The memory model contains n epochs, with an input and output staging area for each respective epoch. As input are copied into epochs, eventually the corresponding input memory buffer will overflow. The host enqueuing functionality manages the current epoch and the current offset within that epoch, which allows the host to detect when the input buffer will lead to an overflow. Once the buffer is saturated, the host proceeds to allocate further memory to the next epoch. After the device finishes executing the task and copies results to the corresponding epoch's result buffer, the host is notified and that memory is marked as free. After fully allocating the memory for an individual buffer, that buffer remains untouched by the scheduler until the scheduler loops around the entire buffer queue and reaches the same epoch again.

The motivation for this epoch based strategy both asynchronously manages the execution of tasks as well as overhead reduction of continuously freeing and allocating new memory. As these persistent buffers remain allocated throughout the entirety of the kernel, there is no longer any overhead involved in gpu side memory allocation or freeing, as the memory is tied to the lifetime of the kernel and only internaly considered allocated or freed. When the buffer queue loops and returning back, if the specific buffer parameters and memory size has been correctly set, potentially through profiling, the buffer will be free and can be reused. The buffer is only considered free if all tasks within the buffer are considered free, which removes any complicated GPU side memory allocation schemes.

4.2.5 Cuda Streams Optimization

To fully leverage the GPU's memory transfer capabilities, input and output transfers are performed in separate, overlapping CUDA streams. As previously noted, the GPU features distinct engines for H2D and D2H transfers, which can operate concurrently. By assigning input transfers to a dedicated H2D stream and output transfers to a separate D2H stream, this implementation avoids intra-stream dependencies and enables parallel data movement in both directions.

5 Structural Architecture

6 Evaluation

This chapter evaluates the proposed GPU persistent thread model with respect to the objectives of this work:

1. Reducing scheduling latency by minimizing kernel overhead.
2. Establishing a baseline platform for future research in real time GPU scheduling, serving as a foundation for a coroutine based approach.
3. Enhancing understanding of real time GPU scheduling, CUDA architecture, and GPU execution, programming, and scheduling models.

The implementation introduces a GPU persistent thread model designed as both a building block for future scheduling research and as a means to improve the efficiency of executing multiple tasks on the GPU by reducing kernel launch overhead. This persistent GPU model allows for application specific fine tuning with particular focus on hardware resource management. In this model, the persistent kernel executes throughout the lifetime of the application and dynamically processes incoming tasks streamed to the GPU, avoiding repeated kernel launches.

The evaluation focuses on measuring the model's effectiveness in meeting the stated objectives. In particular, it examines scheduling latency, kernel overhead, and suitability as a baseline for real time programming models such as coroutines. The results highlight both the benefits of the approach and its limitations, especially regarding task variability and scaling across the GPU.

6.1 Kernel Overhead Measurement

In order to evaluate the effectiveness of persistent threads in reducing the scheduling latency, the solution was tested on a matrix multiplication benchmark against a baseline implementation. To test the persistent threaded implementation, a simple matrix multiplication task was repeatedly scheduled on the persistent thread implementation and compared with a simple kernel invocation code. The persistent kernel was then tested with a number of varying parameters, including memory overhead, GPU work, and number of persistent kernel blocks.

6.1.1 Matrix Multiplications for Performance Testing

The current execution model requires manual task wrappers for scheduling and executing GPU tasks, making it more complex to provide a varying range of different tasks. The manual task wrappers, designed to reduce the overhead of allocating GPU memory, are required to serialize and deserialize the allocated data during execution. To keep testing manageable, this thesis focuses on matrix multiplications as a representative example to demonstrate the solution's capabilities.

In particular, the selected matrix multiplications were of a predetermined size of 16x16, due to the constraints of the GPU persistent kernel. Normally, the programmer decides at kernel launch the configurations of threads within the kernel; however, this feature is not available for persistent threads. Persistent threads configurations are determined at launch and can not be reallocated throughout the duration of the task. For application based workloads, these values would need to be determined through profiling the workloads and manually implementing the values. In this case, matrix multiplications are significantly faster when the task has at least as many threads as the solution matrix values, which led this implementation to using a 16x16 persistent kernel launch configuration.

6.1.2 Testing Environment

Throughout the matrix multiplication performance testing, as the benchmark and the persistent kernel implementation both vary in application goals, a simulated work environment was designed. For the tests, the persistent kernel task schedules a number of tasks to the GPU and then immediately starts waiting to receive the results. As the tasks are being scheduled, the GPU already begins to execute enqueued tasks. After finishing individual tasks, the host is signaled with the message that the tasks are finished. In comparison, the matrix multiplication benchmark just executes tasks serially in an alloc, memcpy, kernel launch, memcpy loop.

In a real time persistent kernel implementation, the CPU needs to independently run tasks within the system and cannot synchronously wait on results to be streamed back. Rather, the CPU needs the results to be waiting when the it goes to check if the results have returned or not. Unfortunately, this model, when compared to any synchronous execution, will have waiting overhead. This is implemented within the test kernel as a busy waiting scheme; however, this increases the GPU traffic and slightly weakens the GPU implementation as it contests the same streams for data transfers.

6.1.3 Small Kernel and Small Memory Transfers

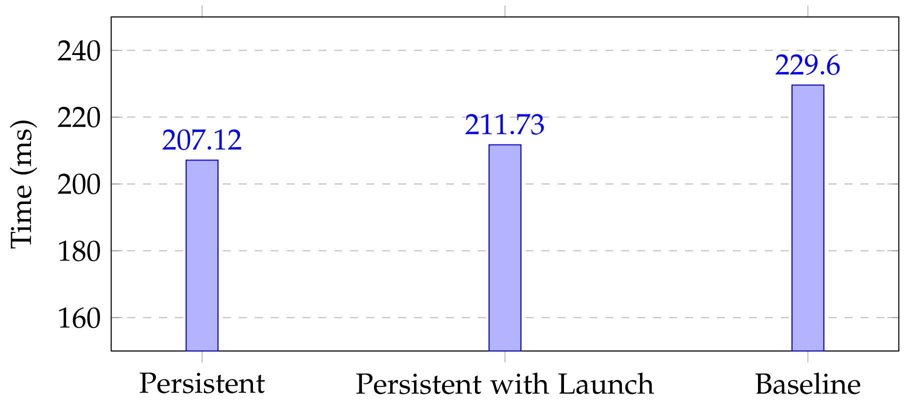

Figure 6.1: Small Kernel and Memory Transfer Tasks

Figure 6.1 shows the first experiment that was run comparing the task scheduling from the GPU persistent kernel in comparison to the baseline model. Here each implementation executes 32 tasks 16x16 matrix multiplication tasks scheduled to the GPU. The matrixes for testing are generated at runtime and compared to the correct CPU implementation using the same functios.

On average over 10,000 runs, the persistent kernel achieved an execution time of 189.165ms, while the baseline model required 229.604ms, demonstrating a clear performance benefit for this small kernel, small memory transfer case. However, the measurements exhibited high variability, with execution times differing by as much as ±50ms between runs, even averaged over 10000 different executions. Despite this noise, the persistent kernel consistently outperformed the baseline in the majority of trials.

6.1.4 Small Kernel and Larger Memory Transfers

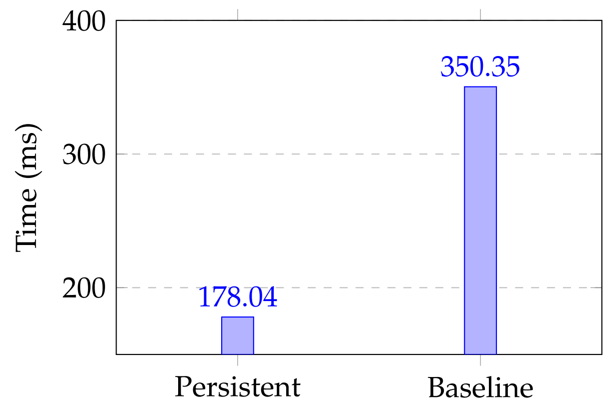

Figure 6.2: Small Kernel and Larger Memory Transfer Tasks

In this secondary evaluation, the GPU executes the same task as before but now transfers an additional 1 MiB of memory to the task and returns it. Interestingly, the persistent thread implementation runs slightly faster than in the previous test, while the baseline control becomes slower. This unexpected result prompted further testing with a larger input size of 10 MiB, which completed in 623.915ms.

The observed increase in speed is likely due to the underlying memory management strategy. To conserve valuable GPU resources, the memory buffers are logically partitioned using offsets. When tasks use very small memory frames, their data ends up placed very close together in memory. During execution, tasks that read from memory locations adjacent to others simultaneously writing to nearby regions can cause contention and reduce performance. By increasing the memory size transferred per task, these memory regions are spaced farther apart, reducing contention and allowing the kernel to run more efficiently.

This separation applies to both input and output buffers. When transferring larger amounts of data (e.g., 1 MiB or more), the serialized GPU memory transfers no longer interfere with kernel execution, leading to the improved performance observed.

6.1.5 Multiple Kernels

6.2 Profiling and Analysis with `nsys`

To better understand the execution characteristics of the implementation, `nsys` profiling was performed for both the baseline (non-persistent) and persistent-threaded approaches. In the baseline version, the profiler output is easy to interpret: each kernel launch is explicitly visible on the timeline, interleaved with host-device memory transfers. This makes it straightforward to determine where computation occurs, identify launch overheads, and correlate them with memory transfer costs.

In contrast, profiling the persistent-threaded implementation presents significant challenges. Since the persistent kernel is launched only once and remains active for the duration of execution, there are no distinct kernel launch entries on the timeline for each individual task. Instead, the majority of the visualized activity in `nsys` consists of frequent host-to-device and device-to-host memory transfers corresponding to the queuing and completion of tasks. The actual execution of the tasks within the persistent kernel is effectively "hidden" from the profiler's high-level view, as it takes place inside the single long-running kernel.

This difference means that the persistent-threaded execution does not lend itself well to the same form of visual, task-by-task inspection available in the baseline case. While memory transfer patterns can still be analyzed, the lack of discrete kernel events makes it difficult to directly measure per-task execution time or to distinguish between overlapping computation and data movement. To obtain deeper insights, low-level instrumentation or custom device-side logging would be required, as standard profiling tools are optimized for discrete kernel launches rather than continuous, event-driven GPU execution.

Consider the evaluation of the current execution.

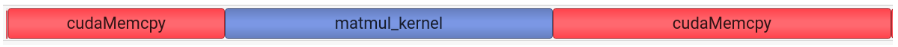

Figure 6.3: Simple `nsys` timelines for the baseline implementations.

6.3 Coroutine Implementation

While the primary objective of this work was the implementation of GPU persistent threads, it became evident that the same architectural foundations could be extended to support coroutine functionality with relatively modest additional effort. The existing design already incorporates a task queue, which the GPU processes in sequence, and a dedicated memory address space for function contexts, making it possible to store and retrieve the continuation state of executing tasks. These features naturally lend themselves to coroutine-style execution, where tasks can yield control and later resume from the same execution point.

6.3.1 Yielding and Changing Tasks

In essence, a coroutine yields its execution by voluntarily releasing its currently allocated compute resources, allowing other tasks to make progress. Within the present implementation, this can be modeled directly using the existing task queue mechanism. The task queue is implemented as a circular buffer, processing tasks in order from the head to the tail. When a task yields, it is effectively repositioned within this queue, relinquishing its slot so that another, potentially higher-priority, task can execute in its place. The most urgent pending task can then take over the vacated slot, ensuring minimal idle time for the GPU.

A more advanced extension would involve replacing the simple circular buffer with a *priority heap*. In such a structure, tasks are automatically sorted and selected for execution based on their priority levels, allowing the scheduler to make more informed decisions about task ordering. While the current system executes tasks to completion once they are selected, this model could still allow tasks to voluntarily halt and later resume execution, improving responsiveness for workloads with mixed task lengths.

6.3.2 Continuation and Resumption

Implementing coroutine behavior also requires explicit mechanisms for saving and restoring task state. When a task yields, the GPU must store sufficient continuation data so that it can later resume execution from precisely the point where it was interrupted. The current memory structure already provides a convenient, per-task memory region where such continuation data can be stored. This includes both the computational state and the control information necessary for resuming execution at the correct instruction address within the device function.

To enable this, each task must be preallocated with enough device memory during scheduling to hold both its input parameters and any continuation data that may be required. This is achieved by copying the task's input stream into the designated memory region and advancing the memory offset to reserve additional space for storing the continuation state. When a task yields, the current execution context—such as register values, loop counters, and program counters—is written into this reserved space. Upon resumption, the scheduler retrieves this data and restores the task's execution context, allowing it to continue seamlessly from its previous stopping point.

By combining these mechanisms with the persistent thread framework, the GPU could effectively execute multiple long-lived, cooperative tasks, enabling finer-grained scheduling and more efficient utilization of GPU resources, especially in irregular or latency-sensitive workloads.

6.4 Limitations and Future Work

Currently, the persistent threads execute using only a single CTA block on a single GPU. This design choice was intentional to establish a minimum viable implementation, but it limits overall GPU utilization and throughput.

Scaling the persistent kernel to launch multiple worker CTAs, ideally one per Streaming Multiprocessor, would improve parallelism and resource usage, potentially unlocking significant performance gains.

Additionally, the current manual streaming of input and output parameters restricts task variability and automation. Developing a generalized parameter serialization and deserialization layer that supports arbitrary kernel arguments would increase flexibility and reduce overhead in task preparation.

Further improvements could also include:

- Implementing task batching, where each worker processes multiple tasks before polling again, to reduce synchronization overhead.
- Optimizing the polling mechanism with adaptive backoff strategies to minimize resource consumption.
- Leveraging asynchronous memory copies with double buffering and pinned memory to better overlap data transfers and computation.
- Extending the evaluation to more complex kernels and real-world workloads beyond matrix multiplication to better characterize the benefits and limitations of persistent threads.

Overall, while the current implementation demonstrates the feasibility of persistent GPU threads, there remains substantial opportunity for optimization and scaling to fully leverage the advantages of this approach.

Abbreviations

GPU Graphics Processing Unit

CPU central processing unit

DSL domain specific language

GPC Graphics Processsing Cluster

TPC Texture Processing Cluster

SM Streaming Multiprocessor

FP floating point

LD/ST Load/Store

SFU special function units

CNN Convolutional Neural Network

CTA Cooperative Thread Array

HBM2 High Bandwidth Memory

VRAM Video Random Access Memory

RAM Random Access Memory

D2H Device to Host

H2D	Host to Device
JIT	just in time

List of Figures

List of Tables

Bibliography

[1] K. Jo, J. Kim, D. Kim, C. Jang, and M. Sunwoo, "Development of autonomous car—part i: Distributed system architecture and development process," *IEEE Transactions on Industrial Electronics*, vol. 61, no. 12, pp. 7131–7140, 2014. DOI: 10.1109/TIE.2014.2321342.

[2] S. Zheng, Z. Zhou, X. Chen, D. Yan, C. Zhang, Y. Geng, Y. Gu, and K. Xu, "Luisarender: A high-performance rendering framework with layered and unified interfaces on stream architectures," *ACM Trans. Graph.*, vol. 41, no. 6, Nov. 2022, ISSN: 0730-0301. DOI: 10.1145/3550454.3555463. [Online]. Available: https://doi.org/10.1145/3550454.3555463.

[3] J. Sun, K. Duan, X. Li, N. Guan, Z. Guo, Q. Deng, and G. Tan, "Real-time scheduling of autonomous driving system with guaranteed timing correctness," in *2023 IEEE 29th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, 2023, pp. 185–197. DOI: 10.1109/RTAS58335.2023.00022.

[4] J. D. Owens, D. Luebke, N. Govindaraju, M. Harris, J. Krüger, A. E. Lefohn, and T. J. Purcell, "A survey of general-purpose computation on graphics hardware," en, *Comput. Graph. Forum*, vol. 26, no. 1, pp. 80–113, Mar. 2007.