RTGPU: Real-Time GPU Scheduling of Hard Deadline Parallel Tasks with Fine-Grain Utilization

An Zou, *Member, IEEE,* Jing Li, *Member, IEEE,*
Christopher D. Gill, *Senior Member, IEEE,* Xuan Zhang, *Member, IEEE,*

Abstract—Many emerging cyber-physical systems, such as autonomous vehicles and robots, rely heavily on artificial intelligence and machine learning algorithms to perform important system operations. Since these highly parallel applications are computationally intensive, they need to be accelerated by graphics processing units (GPUs) to meet stringent timing constraints. However, despite the wide adoption of GPUs, efficiently scheduling multiple GPU applications while providing rigorous real-time guarantees remains a challenge. Each GPU application has multiple CPU execution and memory copy segments, with GPU kernels running on different hardware resources. Because of the complicated interactions between heterogeneous segments of parallel tasks, high schedulability is hard to achieve with conventional approaches. This paper proposes RTGPU, which combines fine-grain GPU partitioning on the system side with a novel scheduling algorithm on the theory side. Through system and theory co-design, RTGPU achieves superior system throughput and real-time schedulability. In this paper, we start by building a model for the CPU and memory copy segments. Leveraging persistent threads, we then implement fine-grained GPU partitioning with improved performance through interleaved execution. To reap the benefits of fine-grained GPU partitioning and schedule multiple parallel GPU applications, we propose a novel real-time scheduling algorithm based on federated scheduling and grid search with uniprocessor fixed-priority scheduling. Our approach provides real-time guarantees to meet hard deadlines, and achieves over 11% improvement in system throughput and up to 57% schedulability improvement compared with previous work. We validate and evaluate RTGPU on NVIDIA GTX1080Ti GPU systems. Our system side techniques can be applied on mainstream NVIDIA GPUs, and the proposed scheduling theory can be used in general heterogeneous computing platforms which have a similar task execution pattern.

Index Terms—GPGPU, Parallel Real-time Scheduling, Persistent Thread, Interleaved Execution, Federated Scheduling, Fixed Priority, Self-suspension Model, Schedulability Analysis.

1 Introduction

Nowadays, artificial intelligence (AI) and machine learning (ML) applications accelerated by graphics processing units (GPUs) are widely adopted in emerging autonomous systems, such as self-driving vehicles and collaborative robotics [1], [2]. For example, Volvo deployed NVIDIA DRIVE PX 2 technology for semi-autonomous driving in 100 XC90 luxury SUVs [3]. These autonomous systems need to execute different AI/ML applications simultaneously in the GPU to perform tasks such as object detection, 3D annotation, movement prediction, and route planning [4], [5]. Moreover, they often need to process images and signals from various sensors and decide the next action in real time. It is thus essential to manage concurrent execution in the GPUs diligently with respect to various timing constraints, since they can have direct and critical impacts on the stability and safety of the whole system.

For general-purpose computing in a non-real-time setting, GPU scheduling aims to minimize the makespan of a single application or to maximize the total throughput of the system [6], [7], [8], [9]. Many state-of-the-art learning frameworks that support GPU acceleration of AI and ML algorithms, such as Caffe [10] and TensorFlow [11], also handle workloads in a sequential manner. This type of sequential execution model is sufficient for large-scale resource-abundant systems (e.g., in data center applications) that aim to maximize the average throughput of a single task. The same model, however, does not translate well in parallel GPU applications with strict timing deadlines. When computing resources are constrained, such as in on-board GPU systems, parallel tasks have to make good use of the limited resource to meet strict deadlines. However, even state-of-art GPU execution patterns pose unique challenges to the real-time scheduling of parallel tasks. First, inside each task, there are multiple serially dependent segments, namely, CPU execution and memory copy segments and GPU kernels. Inside each task, these segments need to access different hardware resources serially. Second, a large GPU kernel in one task may occupy the entire GPU, blocking the GPU kernels in other parallel tasks. This aggravated dependency inside and among tasks may reduce the system's performance or cause extra scheduling pessimism under hard timing constraints.

This paper proposes RTGPU, a general real-time GPU scheduling framework. To overcome such aggravated dependency and pessimism in GPU real-time scheduling, RTGPU provides GPU partitioning and modeling as well as a scheduling algorithm and tight schedulability analysis. First, based on an in-depth understanding of GPU kernel execution and profiling of synthetic workloads, we leverage the persistent threads technique [12] to support Streaming Multiprocessor (SM)-granularity partitioning for concurrent

This work is supported by NSF CNS-1739643, NSF CNS-1948457, and NSFC 62202287.

GPU applications. To fully utilize the GPU resources, we further propose interleaved execution which can achieve 10% to 37% improvement in system utilization compared with SM-granularity resource partitioning without interleaved execution [13]. We then develop a measurement-based task model that introduces the concept of *virtual streaming multiprocessors* (virtual SMs), which allows finer-grained (SM-level) GPU scheduling without any low-level modifications to GPU systems.

Following the flexible task execution pattern, we propose a novel real-time scheduling algorithm leveraging federated and fixed-priority scheduling. The key idea behind federated scheduling is to calculate and statically assign the specific computing resources that each parallel real-time task needs to meet its deadline. Note that preemption between tasks is not required if the correct number of fixed-granularity computing resources can be accurately derived in analysis and enforced during runtime. For the CPU segments and memory copies between CPU and GPU scheduled by the uniprocessor fixed-priority scheduling algorithm, a novel analysis is proposed to calculate the response time upper bounds and lower bounds of the two types of segments alternately. Leveraging the flexibility from GPU partitioning and the scheduling algorithm with tight response time analysis, the proposed RTGPU achieves up to 57% improvement in system schedulability. More generally, our proposed scheduling algorithm and analysis can be applied to other heterogeneous computing systems that have a similar application execution pattern (each task has CPU, memory copy, and heterogeneous core segments), such as AMD GPUs and Google TPUs.

2 Background

2.1 Background on GPU Systems

GPUs are designed to accelerate compute-intensive workloads with high levels of data parallelism. As shown in Fig. 1., a typical GPU program contains three parts — a code segment that runs on the host CPU (the *CPU segment*), the host/device *memory copy segment*, and the device code segment which is also known as the *GPU kernel*. GPU kernels are single instruction multiple threads (SIMT) programs. The programmer writes code for one thread, many threads are grouped into one thread block, and many thread blocks form a GPU kernel. The threads in one block execute the same instruction on different data simultaneously. A GPU consists of multiple streaming multiprocessors (SMs). The SM is the main computing unit, and each thread block is assigned to an SM to execute. Inside each SM are many smaller execution units that handle the physical execution of the threads in a thread block assigned to the SM, such as CUDA cores for normal arithmetic operations, special function units (SFUs) for transcendental arithmetic operations, and load and store units (LD/ST) for transferring data from/to cache or memory.

When GPU-accelerated tasks are executed concurrently, kernels from different tasks are issued to a GPU simultaneously. When kernels are launched, the thread blocks are dispatched to all the SMs on a first-come, first-served basis. An *occupancy factor*, defined as the ratio of active warps (a group of adjacent threads) on an SM to the maximum number of active warps supported by the SM, is used to

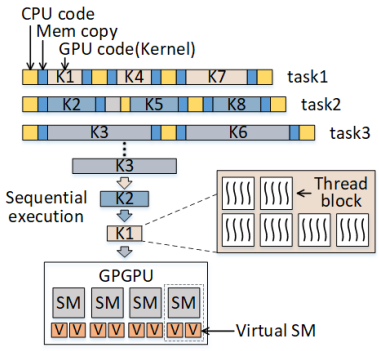


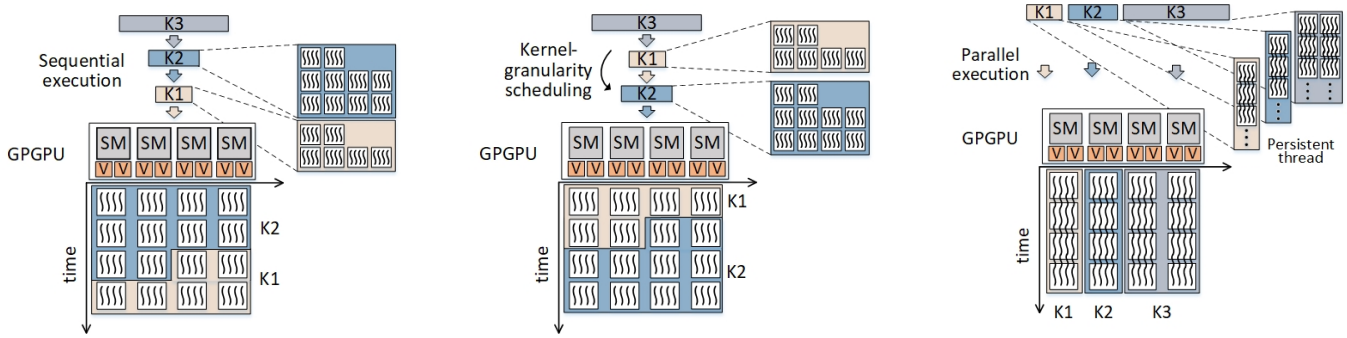
Figure 1: Typical GPU task execution pattern.

describe the capacity of SMs. If the first-launched kernel is large and occupies all the GPU resources (the occupancy factor is 1), the next kernel begins its execution only when the first kernel is about to finish and resources within SMs are freed (occupancy factor below 1). To better manage GPU resources and support multiple kernels concurrently, Multi Process Service (MPS) and Multi-Instance GPU (MIG) have been introduced by NVIDIA. For example, the CUDA contexts belonging to MPS clients funnel their work through the MPS server. It allows client CUDA contexts to bypass hardware limitations associated with time sliced scheduling, and permit CUDA kernels to execute simultaneously [14].

2.2 Persistent Threads

An off-the-shelf GPU supports only kernel-granularity scheduling, as shown in Fig. 2(a). When kernels are launched in the GPU, if the kernel is large enough to fully occupy all the compute resources (SMs and CUDA cores) on the GPU, a GPU is only able to execute one kernel at a time by default even with Multi-Process Service (MPS). The kernel execution orders from different tasks can be changed in kernel-granularity scheduling, as shown in Fig. 2(b).

The persistent threads approach is a new software workload assignment solution proposed to implement finer and more flexible SM-granularity GPU partitioning [12], [15], [16]. Specifically, each persistent threads block links multiple thread blocks of one kernel and is assigned to one SM to execute for the entire hardware execution lifetime of the kernel. For example, in Fig. 2(c), the first thread block in kernel 1 (K1) links the other thread blocks in K1 to form a big linked thread block. When this first thread block is executed by one SM, the other thread blocks in K1, which are linked by the first block, will also be executed in the first SM. Thus, K1 takes one SM to execute. Similarly, in kernel 3 (K3), the first two thread blocks link the other thread blocks and form two big linked thread locks. Thus, the kernel 3 (K3) takes two SMs to execute. When the numbers of linked thread blocks are changed, the resulting number of persistent threads blocks controls how many SMs (i.e., GPU resources) are used by a kernel. In addition, when there are remaining available SMs, CUDA introduces CUDA Streams that support concurrent execution of multiple kernels. By exploiting persistent threads and CUDA Streams, we can explicitly control the number of SMs used by each kernel and execute kernels of different tasks concurrently to achieve SM-granularity scheduling. Persistent threads enabled SM-granularity scheduling fundamentally improves



(a) Default sequential execution (b) Kernel-granularity scheduling (c) SM-granularity scheduling

Figure 2: Comparison of three different GPU application scheduling approaches.

schedulability of parallel GPU applications by exploiting finer-grained parallelism.

2.3 Multi-Segment Self-Suspension Model

In the multi-segment self-suspension model, a task τ_i has m_i execution segments and $m_i - 1$ suspension segments between the execution segments. So task τ_i with deadline D_i and period T_i is expressed as a 3-tuple:

$$\tau_i = \left((L_i^0, S_i^0, L_i^1, ..., S_i^{m_i-2}, L_i^{m_i-1}), D_i, T_i\right)$$

where L_i^j and S_i^j are the lengths of the j-th execution and suspension segments, respectively. $[\check{S}_i^j, \hat{S}_i^j]$ gives the lower and upper bounds of the suspension length S_i^j. $\hat{L}_i^j$ is the upper bound on the length of the execution segment L_i^j. The analysis in [17] bounds the worst-case response time of a task under the multi-segment self-suspension model, which is summarized below.

Lemma 2.1. *The following workload function $W_i^h(t)$ bounds on the maximum amount of execution that task τ_i can perform during an interval with a duration t and a starting segment L_i^h:*

$$W_i^h(t) = \sum_{j=h}^{l} \hat{L}_i^{j \bmod m_i} + \min\left(\hat{L}_i^{(l+1) \bmod m_i}, t - \sum_{j=h}^{l} \left(\hat{L}_i^{j \bmod m_i} + S_i(j)\right)\right)$$

where l is the maximum integer satisfying the following condition:

$$\sum_{j=h}^{l} \left(\hat{L}_i^{j \bmod m_i} + S_i(j)\right) \leqslant t$$

and $S_i(j)$ is the minimum interval-arrival time between execution segments L_i^j and L_i^{j+1}, which is defined by:

$$S_i(j) = \begin{cases} \check{S}_i^{j \bmod m_i} & \text{if } j \bmod m_i \neq (m_i - 1) \\ T_i - D_i & \text{else if } j = m_i - 1 \\ T_i - \sum_{j=0}^{m_i-1} \hat{L}_i^j - \sum_{j=0}^{m_i-2} \check{S}_i^j & \text{otherwise} \end{cases}$$

Then the response time of execution segment L_i^j in task τ_k can be bounded by calculating the interference caused by the workload of the set of higher-priority tasks $hp(k)$.

Lemma 2.2. *The worst-case response time $\hat{R}_k^j$ is the smallest value that satisfies the following recurrence:*

$$\hat{R}_k^j = \hat{L}_k^j + \sum_{\tau_i \in hp(k)} \max_{h \in [0, m_i - 1]} W_i^h(\hat{R}_k^j)$$

Hence, the response time of task τ_k can be bounded by either taking the summation of the response times of every execution segments and the total worst-case suspension time, or calculating the total interference caused by the workload of the set of higher-priority tasks $hp(k)$ plus the total worst-case execution and suspension time.

Lemma 2.3. *Hence, the worst-case response time $\hat{R}_k$ of task τ_k is upper bounded by the minimum of $\widehat{R1}_k$ and $\widehat{R2}_k$, where:*

$$\widehat{R1}_k = \sum_{j=0}^{m_k-2} \hat{S}_k^j + \sum_{j=0}^{m_k-1} \hat{R}_k^j \tag{1}$$

and $R2_k$ is the smallest value that satisfies the recurrence:

$$\widehat{R2}_k = \sum_{j=0}^{m_k-2} \hat{S}_k^j + \sum_{j=0}^{m_k-1} \hat{L}_k^j + \sum_{\tau_i \in hp(k)} \max_{h \in [0, m_i - 1]} W_i^h(\widehat{R2}_k) \tag{2}$$

3 CPU and Memory Model

In this work, we target CPU-GPU heterogeneous computing systems. The heterogeneous systems may have different typologies such as different numbers of CPU cores, memory copy engines, and GPUs (and also SMs in one GPU). To propose a general system model, we start from the most fundamental case which only has one CPU core, one memory copy engine, and multiple SMs in one GPU. All applications are written as threads of a single process. This fundamental case is the minimal heterogeneous system with a full heterogeneous computing function. Any heterogeneous system can be regarded as a combination of this fundamental case. This work aims to study the fundamental CPU-GPU heterogeneous computing system real-time scheduling problem as the first step and then extend the study to multiple GPUs in the future.

3.1 CPU Modeling

As represented in Fig. 1, a complete GPU application has multiple segments of CPU code, memory copies between the CPU and GPU, and GPU code (also called GPU kernels). The CPU executes serial instructions, e.g., for communication with IO devices (sensors and actuators) and launches memory copies and GPU kernels. When a CPU executes serial instructions, it naturally behaves as a single-threaded application without parallelism. When the CPU code launches memory copies or GPU kernels, these instructions will be added into multiple FIFO buffers called

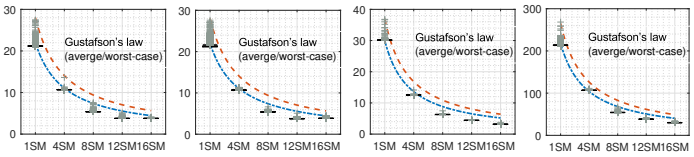
a "CUDA stream". The memory copies and GPU kernels, which are in different CUDA streams, can execute in parallel if there are remaining available resources. The execution order of memory copies and GPU kernels in a single CUDA stream can be controlled by the order in which they are added to it by the CPU code. After the CPU has launched memory copies and GPU kernels into a CUDA stream, it will immediately execute the next instruction, unless extra synchronization is used in the CPU code to wait for the memory copies or GPU kernels to finish. Thus, the CPU segments can be modeled as serial instructions in one thread.

3.2 Memory Modeling

Memory copying between the CPU and GPU execution units includes two stages. In the first stage which is also called global memory copy, data is copied between the CPU memory and the GPU memory through a single peripheral component interconnect express (PCIe) or through a network on chip (NoC). The PCIe and NoC offer packet-based and full-duplex communication between any two endpoints. The number of global memory copies that can happen simutanously are determined by the number of copy engines provided by GPUs. For example, GeForce GTX TITAN Black and Jetson TX2 have 1 copy engine; 1080TI, TITAN X and NVIDIA Xavier [18] have 2 copy engines. In this work, we assume that there is only one copy engine in the minimal heterogeneous system model, which has one CPU core and multiple heterogeneous cores. Also, the memory copy through PCIe/NoC is non-preemptive once it starts. The GPU and other accelerators mainly provide two types of first stage memory movement [19], [20]: direct memory copy (also called traditional memory) and unified memory (introduced in CUDA 6.0). Direct memory copy uses traditional memory, where data must be explicitly copied from CPU to GPU portions of DRAM. Unified memory is developed from zero-copy memory where the CPU and the GPU can access the same memory area by using the same memory addresses between the CPU and GPU. In the following discussion, we focus mainly on direct memory copy, but our approach can also be directly applied to unified memory by setting explicit copy length to zero. The second stage is the memory access from the GPU's execution units to the GPU cache or memory. The GPU adopts a hierarchical memory architecture. Each GPU SM has a local L1 cache, and all SMs share a global L2 cache and DRAM banks. These memory accesses happen simultaneously with the kernel's execution. Therefore, the second stage memory operation is measured and modeled as part of the kernel execution model. Although run-time memory factors, such as the state of the row buffers in the first stage and contention on GPU memory or cache in the second stage, would impact memory copy time, we have to simplify the memory model with static factors given the consideration of real-time scheduling complexity. Therefore, we assume that the memory copy time between CPU memory and GPU memory is a linear function of the copied memory size.

4 Modeling and Management of GPU Fine-Grain Partitioning

Following the persistent thread technique, this section introduces the modeling and management for GPU fine-grain partitioning. The proposed technique in this section takes both throughput and flexibility into account. It develops the system foundation for GPU real-time scheduling with high schedulability.



(a) computation (b) memory (c) branch (d) special

Figure 3: Kernel execution time trends.

4.1 Kernel Execution Model

To understand the relationship between the execution time of a kernel and the number of SMs assigned via persistent threads, we conducted the following experiments. We use five synthetic kernel benchmarks that utilize different GPU resources: a computation kernel, consisting mainly of arithmetic operations; a branch kernel containing large number of conditional branch operations; a memory kernel full of memory and register visits; a special-function kernel with special mathematical functions, such as sine and cosine operations; and a comprehensive kernel including all these arithmetic, branch, memory, and special mathematical operations. Each kernel performs 1000 floating-point operations on a 2^{15}-long vector.

We first run each kernel separately with a fixed workload for 1000 times and record its corresponding execution time with increasing numbers of assigned SMs, as shown in Fig. 3(a). Next, we examine the kernel execution time with increasing kernel sizes and different numbers of assigned SMs. Fig. 3(d) shows that the comprehensive kernel and the other types of kernels have similar trends. From the boxplot, we can see that the kernel execution time follows the formula of Gustafson's law [21] and can be expected of a system whose resources are more flexible:

$$S = N + (1 - N)s \tag{3}$$

Where N is the number of assigned SMs, s is the serial fraction of the workload (which does not benefit from parallelism), and S is the estimated speedup. The S speedup in latency is normalized to the kernel only with computation instruction. From the architecture perspective, the GPU kernels are fully parallel workloads, which can utilize all allocated SMs. The only sequential execution is when the GPU is copying data and launching the kernel. We can also observe that the execution time of a GPU kernel has low variation because it benefits from a single-instruction multiple-threads (SIMT) architecture, in which single-instruction, multiple-data (SIMD) processing is combined with multithreading for better parallelism.

4.2 Interleaved Execution and Virtual SM

Through a close comparison of the GPU kernel execution and the design of GPU architectures, we find that the system throughput can be further improved by exploiting interleaved execution of GPU kernels. On a GPU with M SMs, naive SM-granularity scheduling can first concurrently execute the K_1 and K_2 kernels, each with $M/2$ persistent threads blocks, and then execute the K3 kernel with M

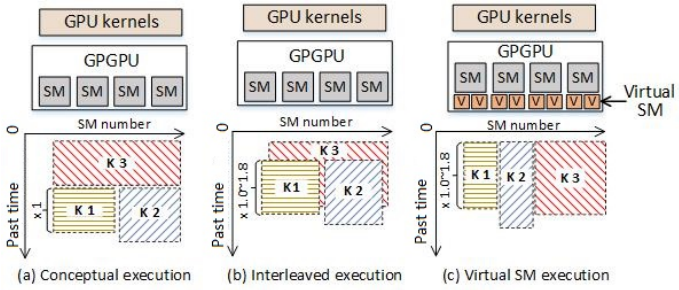


Figure 4: Virtual SM model for interleaved execution

persistent threads blocks, as shown in Fig. 4(a). Each block requires one SM to execute one persistent thread at a time.

On the other hand, an SM actually allows the parallel execution of two or more persistent threads blocks to overlap if the current SM occupancy factor is below 1 which means the number of active warps is less than the maximum [22], [23]. This interleaved execution is similar to the hyper-threading in conventional multithreaded CPU systems that aims to improve computation performance. For example, in an NVIDIA GTX 1080 TI, one SM can hold 2048 software threads, whereas one thread block can have at most 1024 software threads. Thus, two or more thread blocks can be interleaved and executed on one SM. One important consequence of interleaved execution is that the execution time of a kernel increases. Therefore, to improve GPU utilization and efficiency, we can launch all three kernels, as illustrated in Fig. 4(b), where kernel 1 and kernel 2 will simultaneously execute with kernel 3. The execution latency of each kernel is increased by a factor called the interleaved factor, which ranges from 1.0 to 1.8 in the following experiments.

We propose a virtual SM model to capture this interleaved execution of multiple GPU kernels, as shown in Fig. 4(c). In particular, we double the number of physical SMs to get the number of virtual SMs. Compared with a physical SM, a virtual SM has a reduced computational ability and hence a prolonged execution time, the length of which is related to the type of instructions in the interleaved kernel. To understand the interleaved ratio, we empirically measured the execution time of a synthetic benchmark when it was interleaved with another benchmark. Fig. 5 illustrates the minimum, median, and maximum interleaved execution time, colored from light to dark, normalized over the worst-case execution time of the kernel without interleaving, where the left bar is without interleaving and right bar is with interleaving. We can see that the interleaved execution ratio is at most $1.45\times$, $1.7\times$, $1.7\times$, and $1.8\times$ for special, branch, memory and computation kernels, respectively. The proposed virtual SM model improves throughput by $11\% \sim 38\%$ compared to the naive non-interleaved physical SM model. The number of virtual SMs is determined by how many threads can be physically simultaneously executed on one SM. In this work, we use the NVIDIA GTX 1080 TI GPU as an example. In this GPU one physical SM can hold and execute 2048 threads and one thread block has 1024 threads at most. If one thread block is at its highest capacity with 1024 threads, two thread blocks are executed on one physical SM. Therefore, one physical SM will be mapped to two virtual SMs. If one thread block only has 512 threads, four thread blocks are executed on one physical SM. and one physical SM will be mapped to four virtual SMs.

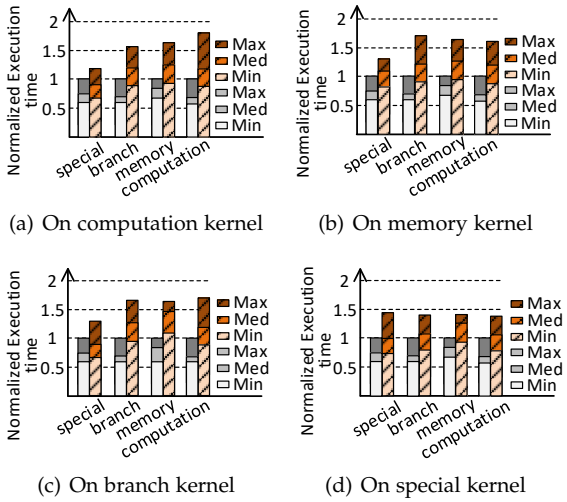


Figure 5: Characterization of the latency extension ratios of interleaved execution

Therefore, by limiting the maximum number of threads in one thread block (in writing the GPU kernels) to $2048/V_{SM}$, we can generate V_{SM} virtual SMs. In this work, we follow the default setting of the NVIDIA GTX 1080 TI GPU where the maximum number of threads in one thread block is 1024.

4.3 Workload Pinning and Self-Interleaving

Using the persistent threads and interleaved execution techniques, multiple tasks can be executed in parallel, and the interleaved execution further improves GPU performance. In real GPU systems, such as NVIDIA GPUs, a hardware scheduler is implemented that allocates the thread blocks to SMs in a greedy-then-oldest manner [24]. Thus, at run time, the thread blocks from a kernel are interleaved and executed with thread blocks from other possible kernels, and the interleaved execution ratio is different when different kernels are interleaved and executed, as shown in Fig. 5. To guarantee a hard deadline, each kernel has to adopt the largest interleaved execution ratio when this kernel is interleaved and executed with other possible kernels. However, using the highest interleaved execution ratio cannot avoid underestimation of the GPU computation ability. Therefore, we introduce workload pinning which pins the persistent threads blocks to specific SMs, and self-interleaving where the kernel interleaves with itself on its pinned SMs.

Workload pinning is implemented by launching Vm (m is the number of physically assigned SMs) persistent threads blocks in each kernel, which is also the number of virtual SMs, so that all virtual SMs will finally have one persistent threads block to execute. If the SM is the targeted pinning SM, the thread block will begin to execute. Persistent threads blocks assigned to undesired SMs (untargeted pinning SMs), will simply return, which takes about 10s μs. When a persistent threads block is assigned to the correct SM, it will not only execute its own workload, but will also execute the workloads from blocks assigned to the undesired SMs. Thus, the kernel is actually executed on the desired SMs, and the undesired SMs execute an empty block within a negligible time. Therefore, the overhead from the GPU fine-grain partitioning is around 10s us.

Algorithm 1: Pseudo Code of Pinned Self-Interleaving Persistent Threads Pseudo Code

```
// Get the ID of current SM with assemble language
static __device__ __inline __ uint32_t __mysmid()
{
uint32_t smid;
asm volatile ("mov.u32 %0, %%smid;" : "=r"(smid));
return smid;  }

// Kernel pinned to desired_SMs with self-interleaved
persistent threads
__global__ void kernel (int *desired_SMs, ...){
int SMs_num = length(desired_SMs);
int Real_SM_id;
Real_SM_id = __mysmid(); // Get the ID of current SM
//Excute on desired SMs, otherwise return
if(Real_SM_id == ∀k, desired_SMs[k]) {
  //Get the global thread index: tid
  int tid = threadIdx.x + k * blockDim.x;
  //off_set links to the next thread block by persistent threads
  int off_set = blockDim.x * SMs_num;
//Divide N threads inside a kernel to V pieces [0 N/V) and
[N/V 2N/V) ... and [(V-1)N/V N) from same kernel
interleaved execute with each other. From the kernel
perspective, the kernel interleaved execute with itself.
  if(blockIdx.x < virtual_SM/V) {
    for(int i = tid; i < N/V; i += off_set) {
      Execute on thread i;}}
  else if(virtual_SM/V⩽ blockIdx.x<2virtual_SM/V){
    for(int i = tid + N/V; i < 2N/V; i += off_set) {
      Execute on thread i;}}
  else if(2virtual_SM/V⩽blockIdx.x<3virtual_SM/V){
    ... }
}
return; }

// Kernel launch
int main () {
int desired_SMs[] = {1, 2, 4}; //The desired SM_id in
this example is 1, 2, 4
dim3 gridsize (number of virtual SM);
dim3 blocksize (Max number of threads per block);
kernel <<< gridsize, blocksize, ..., stream >>>
(desired_SMs, ...);
return 0; }
```

A persistent threads with pinned self-interleaving design and implementation is implemented in the CUDA code, which is detailed described in Algorithm 1. For the usage of the proposed GPU fine-grain partitioning, the researchers will follow the persistent threads style and add the comparison of SM id number at the beginning of the GPU kernel to realize workload pinning and self-interleaving.

5 Practical Full System Scheduling

In this section, we first introduce the RT-GPU scheduling algorithm, and then develop the corresponding timing analysis. One of the key challenges of deriving the end-to-end response times is to simultaneously bound the interference

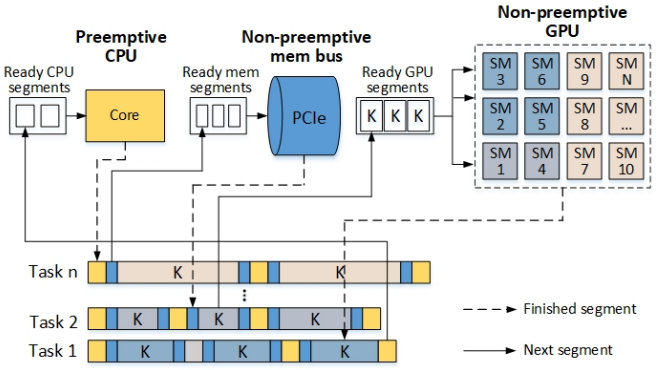


Figure 6: GPU tasks real-time scheduling model.

on CPU, GPU, and memory bus. As the start of memory copy and dispatch of GPU kernel are initialized by the CPU, the scheduling of the full system is managed by the scheduler on the CPU side, i.e. the real-time scheduler in Linux. Therefore, the start of memory copy and GPU kernel will immediately follow its previous CPU segment and memory copy.

5.1 Task Model

Leveraging the platform implementation and the CPU, memory and GPU models discussed in previous sections, the model for the parallel real-time tasks executing on a CPU-GPU platform is shown in Fig. 6. We consider a task set τ comprised of n parallel tasks, where $\tau = \{\tau_1, \tau_2, \cdots, \tau_n\}$. Each task τ_i, where $1 \leqslant i \leqslant n$, has a relative deadline D_i and a period (minimum inter-arrival time) T_i. In this work, we restrict our attention to constrained-deadline tasks, where $D_i \leqslant T_i$, and tasks with fixed task-level priorities, where each task is associated with a unique priority. More precisely, when making scheduling decisions on any resource, such as CPU and bus, the system always selects the segment with the highest priority among all available segments for that resource to execute. Of course, a segment of a task only becomes available if all the previous segments of that task have been completed.

On a CPU-GPU platform, task τ_i consists of m_i CPU segments, $2m_i - 2$ memory copy segments, and $m_i - 1$ GPU segments. As discussed in Section 4.1, a GPU segment G_i^j models the execution of a GPU kernel on interleaved SMs using total work GW_i^j, critical-path overhead GL_i^j, and interleaved execution ratio α_i^j, i.e., $G_i^j = (GW_i^j, GL_i^j, \alpha_i^j)$. Thus, task τ_i can be characterized by the following tuple:

$$\begin{aligned}\tau_i = \Big(\big(&CL_i^0, ML_i^0, G_i^0, ML_i^1, CL_i^1, ML_i^2, G_i^1, ML_i^3,\\ &\cdots, CL_i^j, ML_i^{2j}, G_i^j, ML_i^{2j+1}, \cdots, CL_i^{m_i-2},\\ &ML_i^{2m_i-4}, G_i^{m_i-2}, ML_i^{2m_i-3}, CL_i^{m_i-1}\big), D_i, T_i\Big)\end{aligned} \tag{4}$$

where CL_i^j and ML_i^j are the execution times of the $(j+1)$-th CPU and memory copy segments, respectively. In addition, we use $\check{}$ and $\hat{}$ to denote the lower and upper bound on a random variable. For example, $\widehat{CL}_i^j$ and $\widecheck{CL}_i^j$ are the upper and lower bounds on execution times of the $(j+1)$-th CPU segment of τ_i, respectively.

To derive the end-to-end response time R_i of task τ_i, we will analyze the response times GR_i^j, MR_i^j, and CR_i^j of each individual GPU, memory copy, and CPU segments,

respectively, and calculate their lower and upper bounds in the following subsections.

5.2 Federated Scheduling for GPU Segments

For executing the GPU segments of the n tasks on the shared GPU with VGN virtual SMs (i.e., GN physical SMs), we propose to generalize federated scheduling, a scheduling paradigm for parallel real-time tasks on general-purpose multi-core CPUs, to scheduling parallel GPU segments. The key insight of federated scheduling is to calculate and assign the minimum number of dedicated resources needed for each parallel task to meet its deadline.

Specifically, we allocate VGN_i dedicated virtual SMs to each task τ_i, such that its GPU segment G_i^j can start executing immediately after the completion of the corresponding memory copy ML_i^{2j}. In this way, the mapping and execution of GPU kernels to SMs are explicitly controlled via the persistent thread and workload pinning interfaces, so the effects caused by the black-box internal scheduler of a GPU are minimized. Additionally, tasks do not need to compete for SMs, so there is no blocking time on the non-preemptive SMs. Furthermore, via the self-interleaving technique, we enforce that different GPU kernels do not share any physical SMs. Therefore, the interference between different GPU segments is minimized, and the execution times of GPU segments are more predictable.

In summary, each task τ_i is assigned with VGN_i dedicated virtual SMs where each of its GPU segments self-interleaves and has an interleaved execution ratio α_i^j. In Section 5.5, we will present the algorithm that determines the SM allocation to tasks. Here, for a given allocation, we can easily extend the formula in Section 4.1 to obtain the following lemma for calculating the response time GR_i^j of a GPU segment G_i^j.

Lemma 5.1. *If the GPU segment G_i^j has a total work in range $[\widetilde{GW}_i^j, \widehat{GW}_i^j]$, a critical-path overhead in range $[0, \widehat{GL}_i^j]$ and an interleaved execution ratio in range $[1, \alpha_i^j]$, then when running on VGN_i dedicated virtual SMs, its response time is in $[\widetilde{GR}_i^j, \widehat{GR}_i^j]$ where*

$$\widetilde{GR}_i^j = \frac{\widetilde{GW}_i^j}{VGN_i}, \text{ and } \widehat{GR}_i^j = \frac{\widehat{GW}_i^j \alpha_i^j - \widehat{GL}_i^j}{VGN_i} + \widehat{GL}_i^j.$$

Proof. The lower bound $\widetilde{GR}_i^j$ is the minimum execution time of this GPU segment on VGN_i virtual SMs. In the best case, there is no critical-path overhead and no execution time inflation due to interleaved execution. The minimum total virtual work $\widetilde{GW}_i^j$ is executed in full parallelism on VGN_i virtual SMs, which gives the formula for $\widetilde{GR}_i^j$.

In the worst case, the maximum total virtual work is $\widehat{GW}_i^j \alpha_i^j$, as it demands the most computation and thus longest execution time. Additionally, the maximum critical-path overhead $\widehat{GL}_i^j$ captures the maximum overhead of launching the kernel, which run serially and cannot benefit from parallelism. Since $\widehat{GL}_i^j$ is a constant overhead and is not affected by self-interleaving and multiple virtual SMs, we do not need to apply the interleaved execution ratio α_i^j to $\widehat{GL}_i^j$. After deducting the critical-path overhead using to Gustafson's law in Equation 3, the remaining GPU computation is embarrassingly parallel on VGN_i virtual SMs, which results the formula of $\widehat{GR}_i^j$. □

Note that Lemma 5.1 calculates both the lower and upper bounds on the response time of GPU segment G_i^j, because both bounds are needed when analyzing the total response time of task τ_i. Both the lower and upper bounds can be obtained by profiling the execution time of GPU segments.

During runtime execution of schedulable task sets, the work in Section 4 will generate VGN_i persistent threads blocks for each GPU segment of task τ_i to execute on its own assigned VGN_i virtual SMs. For the less powerful GPU with small numbers of SMs, we need to generate enough virtual SMs to make each task have its virtual SMs. This can be realized by limiting the maximum number of threads in a thread block as described by Section 4.2. For example, in the GTX1080TI GPU, each SM can execute 2048 threads simultaneously. Therefore, if each block has 1024 threads, two blocks are executed on one SM, which means one physical SM generates two virtual SMs. By limiting the maximum number of threads in one thread block to $2048/V_{SM}$, we can generate V_{SM} virtual SMs. Therefore, in most cases, enough virtual SMs could be generated.

5.3 Fixed-Priority Scheduling for memory copy Segments with Self-Suspension and Blocking

From the perspective of executing memory copies over the bus, memory copy segments are "execution segments"; the time intervals where task τ_i spends on waiting for CPU and GPU to complete the corresponding computation are "suspension segments". However, compared with the standard self-suspension model, memory copy over a bus has the following differences. (1) Because memory copy is non-preemptive, a memory copy segment of a high-priority task can be blocked by at most one memory copy segment of any lower-priority task if this lower-priority segment has already occupied the bus. (2) The length of suspension between two consecutive memory copies depends on the response time of the corresponding CPU or GPU segment. (3) The response times of CPU segments are related to the response times of memory copy segments, which will be analyzed in Section 5.4. (4) Moreover, the lower bounds on the end-to-end response times of a task are related to the response times of all types of segments, which requires a holistic fixed-point calculation to be presented in Section 5.5. Please note that above differences are not unique in the CPU-GPU system, they widely present in state-of-the-art heterogeneous systems.

We define the following memory copy workload function $MW_i^h(t)$, which is similar to the workload function defined for standard self-suspension tasks in Section 2.3.

Lemma 5.2. *$MW_i^h(t)$ bounds the maximum amount of memory copy that task τ_i can perform during an interval with a duration t and a starting memory copy segment ML_i^h, where:*

$$MW_i^h(t) = \sum_{j=h}^{l} \widehat{ML}_i^{j \bmod 2m_i-2} + \min\Big(\widehat{ML}_i^{(l+1) \bmod 2m_i-2}, t - \sum_{j=h}^{l} \big(\widehat{ML}_i^{j \bmod 2m_i-2} + MS_i(j)\big)\Big)$$

where l is the maximum integer satisfying the following condition:

$$\sum_{j=h}^{l} \left(\widehat{ML}_i^{j \bmod 2m_i-2} + MS_i(j)\right) \leqslant t$$

and $MS_i(j)$ is defined as follow:

- **If** $j \bmod (2m_i-2) \neq (2m_i-3)$ *and* $j \bmod 2 = 0$, *then* $MS_i(j) = \widecheck{GR}_i^{\left(j \bmod (2m_i-2)\right)/2}$;
- **Else if** $j \bmod (2m_i-2) \neq (2m_i-3)$ *and* $j \bmod 2 = 1$, *then* $MS_i(j) = \widecheck{CL}_i^{\left((j \bmod (2m_i-2))+1\right)/2}$;
- **Else if** $j = 2m_i-3$, *then* $MS_i(j) = T_i - D_i + \widecheck{CL}_i^{m_i-1} + CL_i^0$;
- **Else** $MS_i(j) = T_i - \sum_{j=0}^{2m_i-3} \widehat{ML}_i^j - \sum_{j=1}^{m_i-2} \widecheck{CL}_i^j - \sum_{j=0}^{m_i-2} \widecheck{GR}_i^j$;

Proof. From the perspective of executing memory copies over the bus, the $2m_i-2$ memory copy segments are the execution segments by the definition of self-suspension task in Section 2.3. So the definition of $MW_i^h(t)$ and l directly follows those in Lemma 2.1 by applying $\widehat{ML}$ to $\widehat{L}$ and changing from m_i to $2m_i-2$.

The key difference is in the definition of $MS_i(j)$, which is the minimum "interval-arrival time" between execution segments ML_i^j and ML_i^{j+1}. By the RT-GPU task model, when $j \bmod (2m_i-2) \neq (2m_i-3)$, there is either a GPU or CPU segment after ML_i^j, depending on whether the index is even or odd. So the lower bound on the response time of the corresponding GPU or CPU segment is the minimum interval-arrival time on the bus. For the latter case, the response time of a CPU segment is lower bounded by its minimum execution time. When $j = 2m_i-3$, ML_i^j is the last memory copy segment of the first job of τ_i occurring in the time interval t. In the worst case, all the segments of this job are delayed toward its deadline, so the minimum interval-arrival time between ML_i^j and ML_i^{j+1} is the sum of $T_i - D_i$, the minimum execution time of the last CPU segment $\widecheck{CL}_i^{m_i-1}$, and the minimum execution time of the first CPU segment CL_i^0 of the next job. The last case calculates the minimum interval-arrival time between the last memory copy segment of a job that is not the first job and the first memory copy segment of the next job. Since these two jobs have an inter-arrival time T_i between their first CPU segments, intuitively, $MS_i(j)$ is T_i minus all the segments of the previous job plus the last CPU segment $\widecheck{CL}_i^{m_i-1}$ of the previous job plus the first CPU segment CL_i^0 of the next job, which is the above formula. □

Hence, the response time of memory copy segment ML_k^j can be bounded by calculating the interference caused by the workload of tasks $hp(k)$ with higher-priorities than task τ_k and the blocking term from a low-priority task in $lp(k)$.

Lemma 5.3. *The worst-case response time $\widehat{MR}_k^j$ is the smallest value that satisfies the following recurrence:*

$$\widehat{MR}_k^j = \widehat{ML}_k^j + \sum_{\tau_i \in hp(k)} \max_{h \in [0, 2m_i-3]} MW_i^h(\widehat{MR}_k^j) + \max_{\tau_i \in lp(k)} \max_{h \in [0, 2m_i-3]} \widehat{ML}_i^h \tag{5}$$

Proof. Because the execution of memory copy segments is non-preemptive, the calculation of $\widehat{MR}_k^j$ extends Lemma 2.2 by incorporating the blocking due to a low-priority memory copy segment that is already under execution on the bus. Under non-preemptive fixed-priority scheduling, a segment can only be blocked by at most one lower-priority segment, so this blocking term is upper bounded by the longest lower-priority segment. □

5.4 Fixed-Priority Scheduling for CPU Segments

Now, we will switch the view and focus on analyzing the fixed-priority scheduling of the CPU segments. Looking from the perspective of the uniprocessor, CPU segments become the "execution segments"; the time intervals where task τ_i spends on waiting for memory copy and GPU to complete now become the "suspension segments", since the processor can be used by other tasks during these intervals.

For now, let's assume that the upper bounds $\widehat{MR}_i^j$ and lower bounds $\widecheck{MR}_i^j$ on response times of memory copy segments are already given in Section 5.3. As for GPU segments, the upper bounds $\widehat{GR}_i^j$ and lower bounds $\widecheck{GR}_i^j$ have been obtained in Section 5.2. Similarly, we define the following CPU workload function $CW_i^h(t)$.

Lemma 5.4. *$CW_i^h(t)$ bounds the maximum amount of CPU computation that task τ_i can perform during an interval with a duration t and a starting CPU segment CL_i^h, where:*

$$CW_i^h(t) = \sum_{j=h}^{l} \widehat{CL}_i^{j \bmod m_i} + \min\Big(\widehat{CL}_i^{(l+1) \bmod m_i}, t - \sum_{j=h}^{l} \left(\widehat{CL}_i^{j \bmod m_i} + CS_i(j)\right)\Big)$$

where l is the maximum integer satisfying the following condition:

$$\sum_{j=h}^{l} \left(\widehat{CL}_i^{j \bmod m_i} + CS_i(j)\right) \leqslant t$$

and $CS_i(j)$ is defined as follow:

- **If** $j \bmod m_i \neq (m_i-1)$, *then* $CS_i(j) = \widecheck{ML}_i^{2(j \bmod m_i)} + \widecheck{GR}_i^{j \bmod m_i} + \widecheck{ML}_i^{2(j \bmod m_i)+1}$;
- **Else if** $j = m_i - 1$, *then* $CS_i(j) = T_i - D_i$;
- **Else** $CS_i(j) = T_i - \sum_{j=0}^{m_i-1} \widehat{CL}_i^j - \sum_{j=0}^{2m_i-3} \widecheck{ML}_i^j - \sum_{j=0}^{m_i-2} \widecheck{GR}_i^j$;

Proof. From the perspective of the uniprocessor, the m_i CPU segments are the execution segments by the definition of self-suspension task. So the definition of $CW_i^h(t)$ and l directly follows those in Lemma 2.1 by applying $\widehat{CL}$ to $\widehat{L}$. For the minimum "interval-arrival time" $CS_i(j)$, there are two memory copy and one GPU segments between segments CL_i^j and CL_i^{j+1} by the RT-GPU task model, when $j \bmod m_i \neq (m_i-1)$. So $CS_i(j)$ is the sum of the minimum response times of these segments, where the response time of a memory copy segment is lower bounded by its minimum length. The case of $j = m_i-1$ is the same. The last case considers for a job that is not the first job in interval t. The calculation is similar to the one in Lemma 2.1, except that

both the $2m_i - 2$ memory copy and $m_i - 1$ GPU segments constitute the suspension time. □

Hence, the response time of CPU segment CL_k^j can be bounded by calculating the interference caused by the CPU workload of tasks $hp(k)$ with higher-priorities than task τ_k.

Lemma 5.5. *The worst-case response time $\widehat{CR}_k^j$ is the smallest value that satisfies the following recurrence:*

$$\widehat{CR}_k^j = \widehat{CL}_k^j + \sum_{\tau_i \in hp(k)} \max_{h \in [0, m_i - 1]} CW_i^h(\widehat{CR}_k^j) \tag{6}$$

Proof. The formula is directly extended from Lemma 2.2. □

5.5 RT-GPU Scheduling Algorithm and Analysis

For a particular virtual SM allocation VGN_i for all tasks τ_i, we can calculate the response times of all GPU, memory copy, and CPU segments using formulas in Section 5.2 to 5.4. Note that a task starts with the CPU segment CL_i^0 and ends with the CPU segment $CL_i^{m_i-1}$. Therefore, we can upper bound the end-to-end response times for all tasks using the following theorem, by looking at the perspective from CPU.

Theorem 5.6. *The worst-case end-to-end response time $\widehat{R}_k$ of task τ_k is upper bounded by the minimum of $\widehat{R1}_k$ and $\widehat{R2}_k$, i.e., $\widehat{R}_k = \min(\widehat{R1}_k, \widehat{R2}_k)$, where:*

$$\widehat{R1}_k = \sum_{j=0}^{m_k-2} \widehat{GR}_k^j + \sum_{j=0}^{2m_k-3} \widehat{MR}_k^j + \sum_{j=0}^{m_k-1} \widehat{CR}_k^j \tag{7}$$

and $R2_k$ is the smallest value that satisfies the recurrence:

$$\begin{aligned} \widehat{R2}_k = &\sum_{j=0}^{m_k-2} \widehat{GR}_k^j + \sum_{j=0}^{2m_k-3} \widehat{MR}_k^j + \sum_{j=0}^{m_k-1} \widehat{CL}_k^j \\ &+ \sum_{\tau_i \in hp(k)} \max_{h \in [0, m_i - 1]} CW_i^h(\widehat{R2}_k) \end{aligned} \tag{8}$$

Proof. The calculations for $\widehat{R1}_k$ and $\widehat{R2}_k$ are extended from Lemma 2.3 by noticing that the time spent on waiting for GPU and memory copy segments to complete are suspension segments from the perspective of CPU execution. □

With the upper bound on the response time of a task, the following corollary follows immediately.

Corollary 5.6.1. *A CPU-GPU task τ_k is schedulable under federated scheduling on virtual SMs and fixed-priority scheduling on CPU and bus, if its worst-case end-to-end response time $\widehat{R}_k$ is no more than its deadline D_k.*

Computational complexity. Note that the calculations for the worst-case response times of individual CPU and memory copy segments, as well as one upper bound on the end-to-end response time, involves fixed-point calculation. Thus, the above schedulability analysis has pseudopolynomial time complexity. Given the system model notation in Section 5.1, the grid search on spatial partitioning of GN SMs has a complexity of $min(O(GN^n), O(n^{GN}))$. The analysis of fixed-priority tasks on the memory copy and on the CPU have a complexity of $O(m_i^2)$ respectively. Therefore, the time complexity of the entire scheduling strategy with response time analysis is

$$min(O(GN^n m_i^4), O(n^{GN} m_i^4)). \tag{9}$$

Algorithm 2: Fixed Priority Self-Suspension with Grid Searched Federated Scheduling

Input: Task set τ, number of physical SMs GN
Output: Schedulability
//Generating enough virtual SMs:
for $Thread\_block\_limit$ = 1024, 512, ... **do**
1 $VGN = \frac{SM\_thread\_limit}{Thread\_block\_limit}$
  if $VGN \geqslant n$ **then**
    Break;
  //Grid search for federated scheduling of GPU segments:
2 **for** VGN_1 = 1, ..., VGN **do**
3 **for** VGN_i = 1, ..., $VGN - \sum_{j=1}^{i-1} VGN_j$ **do**
4 **for** VGN_n = 1, ..., $VGN - \sum_{j=1}^{n-1} VGN_j$ **do**
      //Calculate response times of GPU segments:
5 $\widetilde{GR}_i^j = \frac{\widetilde{GW}_i^j}{2VGN_i}, 1 \leqslant i \leqslant n$;
6 $\widehat{GR}_i^j = \frac{\widehat{GW}_i^j \alpha_i^j - \widehat{GL}_i^j}{2VGN_i} + \widehat{GL}_i^j, 1 \leqslant i \leqslant n$;
7 Calculate worst-case response time $\widehat{MR}_k^j$ for all memory copy segments using Eq.(5);
8 Calculate worst-case response time $\widehat{CR}_k^j$ for all CPU segments using Eq.(6);
9 Calculate worst-case end-to-end response time $\widehat{R}_k$ for all tasks using Theorem 5.6;
10 **if** $\widehat{R}_k \leqslant D_k$ *for all* τ_k **then**
11 Schedulability = 1; break out of all for loops;

Note that the above schedulability analysis assumes a given virtual SM allocation under federated scheduling. Hence, we also need to decide the best virtual SM allocation for task sets, in order to get better schedulability. The following RT-GPU Scheduling Algorithm adopts a brute force approach to deciding virtual SM allocation. Specifically, it enumerates all possible allocations for a given task set on a CPU-GPU platform and uses the schedulability analysis to check whether the task set is schedulable or not. Alternatively, a greedy approach can be applied, if one needs to reduce the running time of the algorithm while a slight loss in schedulability is affordable. Given the number of SMs assigned to the tasks and the CPU and GPU execution time, the schedulability under current resource utilization rate can be calculated following the procedure from subsection 5.2 to subsection 5.5. The full procedure of scheduling GPU tasks can be described as follows: (1) Grid search [25] a federated scheduling for the GPU codes and calculate the GPU segment response time $[\widetilde{GR}_i^j \ \widehat{GR}_i^j]$, details in Section 5.4. (2) The CPU segments and memory copy segments are scheduled by fixed priority scheduling. (3) If all the tasks can meet the deadline, then they are schedulable and otherwise go back to step (1) to grid search for the next federated scheduling. This schedulability test can be summarized with pseudo code in Algorithm 2.

5.6 Roadmap of Extending the Scheduling

Moreover, this end-to-end response time analysis is not limited to CPU-memory-GPU systems. It can also be directly applied to other heterogeneous systems with one type of heterogeneous core, like CPU-memory-FPGA and

CPU-memory-TPU systems. To schedule the systems with multiple GPUs (with the same type of GPU SMs), a new constraint must be added to the GPU SM allocation part (after line 3 in Algorithm 2). In this constraint, the GPU SM allocation can only be valid when all the tasks are not executed on the SMs that belong to different GPUs. Further to schedule the systems with multiple GPUs (with different types of GPU SMs), the above constraint must be added. Also, the lower and upper bounds of GPU segment response time (line 4 and line 5 in Algorithm 2) must be calculated with the corresponding computing power of the different types of GPU SMs. After this updated GPU federated scheduling, the fixed-priority scheduling of CPU and memory copy can be directly applied.

6 Full-System Evaluation

We now present an evaluation of our approach. Section 6.1 describes experiments we conducted to validate our approach. Section 6.2 explains how we implemented persistent threading and workload pinning in those experiments. Section 6.3 discusses our analytical evaluations of schedulability under our approach. Finally, Section 6.4 presents schedulability results on real GPU systems.

6.1 Experiments

We conducted extensive experiments to evaluate the performance of the proposed RTGPU real-time scheduling approach. We choose self-suspension [26], STGM [27]: Spatio-Temporal GPU Management for Real-Time Tasks, and Enhanced MPCP [28]: Analytical enhancements and practical insights for mpcp with self-suspensions as baselines to compare with, as they represent the state-of-the-art in both entire GPU and fine-grained (SM-granularity) GPU real-time scheduling algorithms and schedulability tests. **1. Proposed RTGPU:** the proposed real-time GPU scheduling of hard deadline parallel tasks with fine-grain utilization of persistent threads, interleaved execution, virtual SM, and fixed-priority federated scheduling. **2. Self-Suspension:** real-time GPU scheduling of hard deadline parallel tasks with the persistent threads with self-suspension scheduling, as in [26]. **3. STGM:** real-time GPU scheduling of hard deadline parallel tasks with the persistent threads and busy-waiting scheduling, as in [27]. This work also tested and analyzed the self-suspension scheduling under different scenarios. **4. Enhanced MPCP:** real-time GPU scheduling of hard deadline parallel tasks with hybrid approach of the enhancements and practical insights for MPCP with self-suspension, as in [28]. Please note that there is no previous scheduling algorithm that exactly matches the proposed system model. All the above scheduling algorithms are modified to match the proposed model. Some unique good features may be slightly scarified in the modification. For example, STGM can support multiple CPU cores but in our system model, we only use one CPU core.

To compare the schedulability results for these approaches, we measured the acceptance ratio in each of four simulations with respect to a given goal for taskset utilization. We generated 100 tasksets for each utilization level, with the following task configurations. The acceptance ratio of a level was the number of schedulable tasksets, divided by the number of tasksets for this level, i.e., 100.

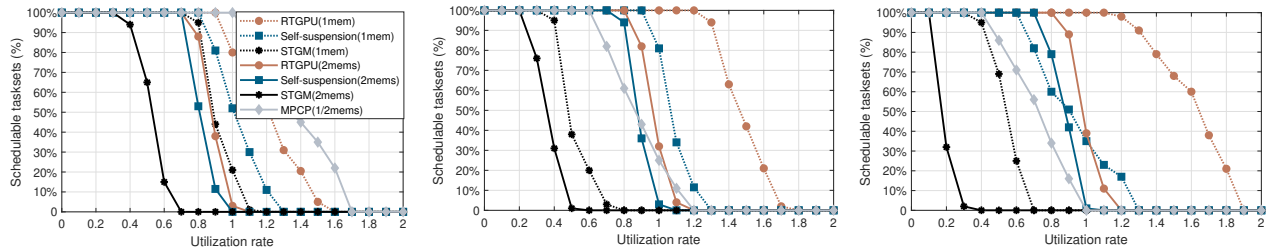
Table 1: Parameters for the taskset generation

Parameters	Value
Number of tasks N in taskset	5
Task type	periodic tasks
Number of subtasks M in each task	5
Number of tasksets in each experiment	100
CPU segment length (ms)	[1 to 20]
Memory segment length (ms)	[1 to 5]
GPU segment length (ms)	[1 to 20]
Task period and deadline	(T_i/D_i)
Number of physical GPU SMs N_{SM}	10
Priority assignment	D monotonic

According to the GPU workload profiling and characterization [29], the memory length upper bound was set to 1/4 of the GPU length upper bound. We first generated a set of utilization rates, U_i, with a uniform distribution for the tasks in the taskset, and then normalized the tasks to the taskset utilization values for the given goal. Next. we generated the CPU, memory, and GPU segment lengths, uniformly distributed within their ranges in Table 1. The deadline D_i of task i was set according to the generated segment lengths and its utilization rate: $D_i = (\sum_{j=0}^{m_i-1} \widehat{CL}_i^j + \sum_{j=0}^{2m_i-3} \widehat{ML}_i^j + \sum_{j=0}^{m_i-2} \widehat{GL}_i^j)/U_i$. In the configuration setting, the CPU, memory, and GPU lengths were normalized with one CPU, one memory interface, and one GPU SM. When the total utilization rate, U, is 1, the one CPU, one memory interface, and one GPU SM are fully utilized. As there are multiple SMs available (and used), the total utilization rate will be larger than 1. The period T_i is equal to the deadline D_i. The task priorities are determined with deadline-monotonic priority assignment. Meanwhile, in each experiment we evaluate two models. The first model has two memory copies: one memory copy from CPU to GPU and one memory copy back from GPU to CPU between a CPU segment and a GPU segment, which is exactly the execution model we introduced in section 4. The second model has one memory copy between a CPU segment and a GPU segment, which combines the memory copy from CPU to GPU and the memory copy from GPU to CPU. These two models can capture not only the CPU-GPU systems but also general heterogeneous computing architectures.

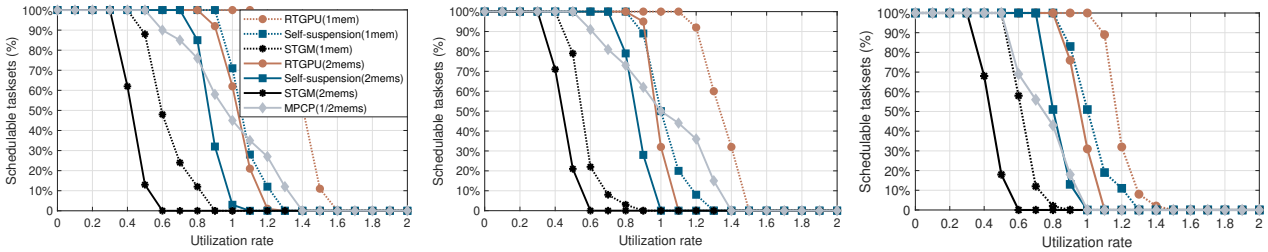
6.2 System Side Implementation

First on the system side, we implement the persistent thread with workload pinning and self-interleaving. As studied in Section 4, the workload pinning and self-interleaving can improve the system throughput under GPU fine-grain partitioning. We further test the workload pinning and self-interleaving with five real benchmarks from the Rodinia and CUDA SDK tasksets representing different but common types of GPU applications. Table 2 shows the interleaved execution ratios measured from these application with different SMs assigned to each kernel. As the kernel is interleaved with itself, each kernel has a relatively stable interleaved execution ratio when different numbers of SM are assigned to the kernel. Since the benchmark is divided into two pieces and these two pieces are executed simultaneously on the given number of SMs, the throughput improvements can be calculated with $2/\alpha$. Therefore,



(a) computation:suspension=2:1 (b) computation:suspension=1:2 (c) computation:suspension=1:8

Figure 7: Schedulability under different computation (CPU) and suspension (memory+GPU) lengths



(a) 3 subtasks (b) 5 subtasks (c) 7 subtasks

Figure 8: Schedulability under different numbers of subtasks

Table 2: Interleaved execution ratios measured from real benchmarks with different SMs assigned to each kernel

Benchmark/Number of SMs	1	2	4	8
Dxtc	1.68	1.66	1.69	1.64
BFS	1.56	1.61	1.59	1.57
Particle Filter	1.42	1.45	1.41	1.46
Vectoradd	1.80	1.77	1.80	1.78
Quasirand	1.22	1.23	1.24	1.23

the benchmarks achieve 11% to 81% throughput improvements on benchmark Vectoradd and Quasirand. Quasirand achieves significant throughput improvement because the original Quasirand uses less than half of the SM resources. Self-interleaving can leverage the remaining resources to achieve throughput improvement. Next, we will evaluate the schedulability of the proposed approach. To have a fair comparison, the following experiments are all based on the workload pinning and self-interleaving.

6.3 Schedulability Analysis

Our analytical evaluation focused on the schedulability of tasksets as the overall utilization increased, with respect to different parameters pertinent to schedulability. The following sub-subsections present the results of four simulations that each varied the different parameters we examined: the ratios of CPU, memory, and GPU segment lengths; the number of subtasks; the number of tasks; and the number of total SMs.

6.3.1 CPU, Memory, and GPU Lengths

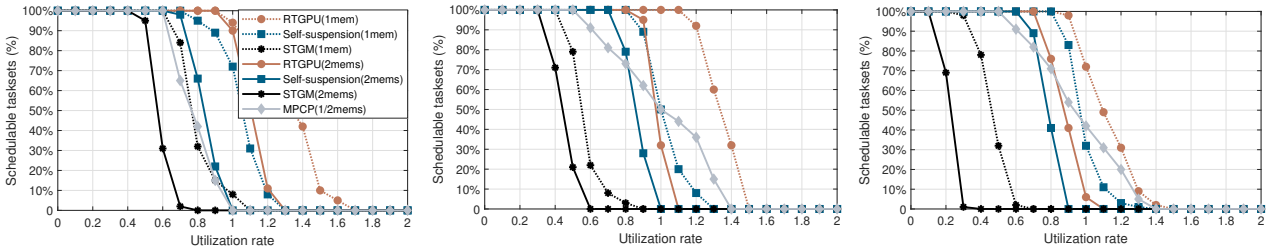
We investigated the impact of CPU, memory, and GPU segment lengths on the acceptance ratio. To study this quantitatively, We tested the acceptance ratio under different length range ratios. The CPU length is shown as Table 1 and we changed the memory, and GPU lengths according to the length ratio. Fig. 7 shows taskset acceptance ratio when the CPU, memory, and GPU length range ratios were set to 2:1, 1:2, and 1:8, which give an exponential scale.

Not surprisingly, the STGM approach is effective only when the memory and GPU segment (suspension segment) lengths are sufficiently short: the STGM approach was developed based on "busy waiting". When tasks are being processed in memory copy and GPU segments, the CPU core is not released and remains busy waiting for the memory copy and GPU segments to finish. Although this is the most straightforward approach, its pessimistic aspect lies in the CPU waiting for the memory copy and GPU segments to finish. Thus, it will be ineffective and hugely pessimistic when the memory copy and GPU segments are large.

Self-suspension scheduling in [17] increases the schedulability performance compared with the straight forward STGM approach. Self-suspension models the memory and GPU segments as being suspended, and the CPU is released during this suspension. The theoretical drawback of this approach is that the suspension does not distinguish between the memory segments and GPU segments. Instead, they are modelled as non-preemptive and will block higher priority tasks. However, in real systems, each task is allocated its own exclusive GPU SMs, and the GPU segments in one task will not interfere the GPU segments in other tasks.

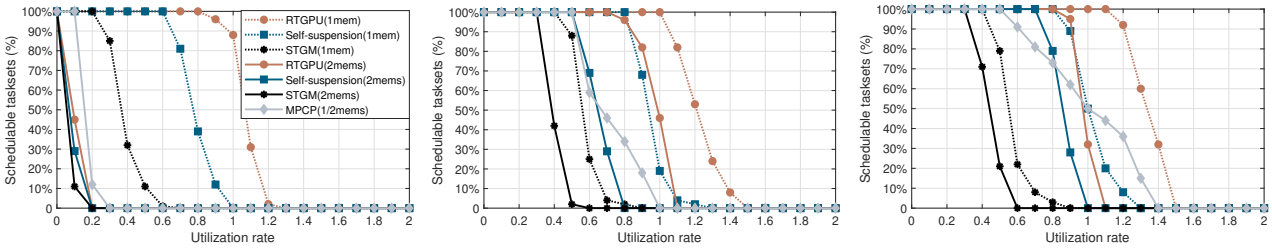
Enhanced MPCP in [28] has the best performance when the CPU segments are large. The schedulability is sensitive to the CPU, memory, and GPU segment lengths ratios. It will decrease obviously when the GPU segments become longer because Enhanced MPCP is designed for the heterogeneous systems with multiple CPU cores which targets the applications with long CPU segments.

The RTGPU schedulability analysis proposed in this paper is effective even when the memory and GPU segment (suspension segment) lengths are long. For example, in the 1:8 length test, RTGPU reaches 1.1 utilization rate with 100% schedulability, which is 57% improvement with previous work. In this approach, we distinguish the CPU, memory, and GPU segments based on their individual properties. For example, if the CPU cores are preemptive, then no blocking will happen. Blocking happens only in non-preemptive



(a) 3 tasks (b) 5 tasks (c) 7 tasks

Figure 9: Schedulability under different numbers of tasks



(a) 5 SMs (b) 8 SMs (c) 10 SMs

Figure 10: Schedulability under different numbers of SMs

memory segments. Meanwhile, because federated scheduling is applied for the GPU segments and each task is allocated its own exclusive GPU SMs, the GPU segments can be executed immediately when they are ready, without waiting for higher priority GPU segments to finish or being blocked by lower GPU segments.

Also, by comparing the models with one memory copy and two memory copies, we notice that the memory copy is the bottleneck in the CPU-GPU systems because of limited resource (bandwidth) and non preemption. Reducing the numbers of memory copies or combining memory copies can increase the system schedulability, especially when the memory copy length is large shown in Fig. 7 (b) and (c).

6.3.2 Number of Subtasks

We then evaluated the impact of the number of subtasks in each task on the acceptance ratio. From the possible values in Table 1, the number of subtasks, M, in each task was set to 3, 5, or 7. The corresponding acceptance ratios are shown in Fig.8. The results show that with more subtasks in a task, schedulability decreases under all approaches but the proposed RTGPU approach still outperforms all other approaches. The Enhanced MPCP approach is the most robust scheduling algorithm as the subtasks increase.

6.3.3 Number of Tasks

In a third simulation, we evaluated the impact of the number of tasks in each taskset on the acceptance ratio. Again, from the possible values in Table 1, the number of tasks, N, in each task was set to 3, 5, or 7. The corresponding acceptance ratios are shown in Fig.9. As with subtasks, schedulability decreases under all the approaches as the number of tasks increases, but the proposed RTGPU approach outperformed the other two.

6.3.4 Number of SMs

Finally, we examined the impact of the number of total SMs on the acceptance ratio. Based on the possible values in Table 1, the number of subtasks M and tasks N in each setting are again set to 5. The corresponding acceptance ratios are shown in Fig.10. All approaches have better schedulability as the number of available SMs increases. From this set of experiments we can see that adding two more SMs will cause the utilization rate to increase for all approaches. Meanwhile, among all the approaches, the proposed RTGPU approach again achieves the best schedulability across different numbers of SMs. The schedulability of the Enhanced MPCP approach also increases significantly with the increased number of SMs. When the number of SMs is large enough, the Enhanced MPCP approach also performs well. As shown in Fig.10 (a), when the computation resources (GPU SMs) are limited, the bottleneck from memory copy is more obvious and serious. The two memories model has a poor scheduability in all approaches and the one memory model has a significant improved performance.

6.4 Schedulability on Real GPU Systems

To test and compare schedulability between the theoretical boundary and the performance on real GPU systems, we empirically evaluated the proposed RTGPU scheduling framework on an NVIDIA 1080TI GPU, under enough and a limited number of SMs. The CPU was an Intel(R) Core(TM) i7-3930K CPU operating at 3.20GHz. We implemented the synthetic benchmarks described in Section 4 in a common real-time scheduling context, since multiple GPU kernel concurrency is supported only within the same CUDA context. To avoid interference from adaptive power setting and guarantee hard deadlines, we manually fixed the SM core and memory frequencies respectively using the nvidia-smi command. We also set the GPUs to persistence mode to keep the NVIDIA driver loaded even when no applications are accessing the cards.

As in the previous schedulability analysis experiments, each task in a taskset was randomly assigned one of the values in Table 1. The deadline was set to the same value as the period. In this work, we use a measurement-based analysis to get the values of the kernel model (including the critical-path overhead). The execution time distributions of different sizes of memory copies through PCIe from CPU to GPU and

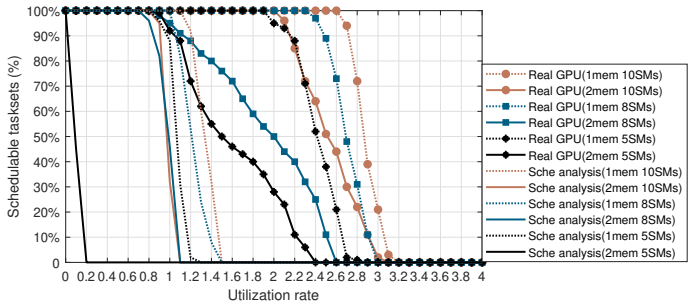


Figure 11: Full system schedulability of 5 parallel tasks under 5, 8, 10 physical SMs

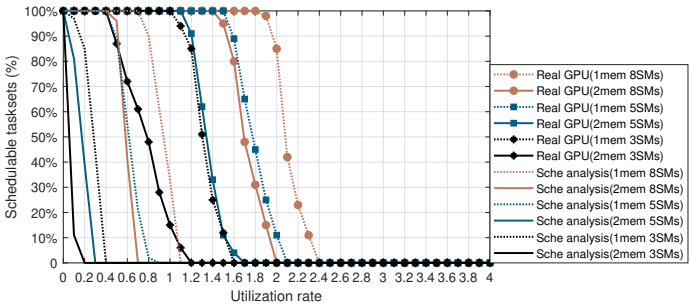


Figure 12: Full system schedulability of 12 parallel tasks under 3, 5, 8 physical SMs

from GPU to CPU and different GPU kernel thread lengths are measured from 10,000 samples. Based on the measured worst case execution time, we calculate the values of the kernel model and build the worst execution time model. Then using the real GPU system, we examined schedulability using different numbers of SMs and compared the results from the schedulability analysis and from the real GPU experiments. Fig. 11 presents the acceptance ratio results of the RTGPU schedulability analysis and experiments on the real GPU system. Both of them have better schedulability as the number of available SMs increases. The gaps between the schedulability analysis and real GPU system arise from the pessimistic aspect of the schedulability analysis and the model mismatches between worst execution time and actual execution time. In the limited computation resource scenarios (5 SMs and 8 SMs), the bottlenecks from memory copy exist in both schedulability test and experiments with real GPU systems. Reducing the numbers of memory copies or combining memory copies are proper methods to deal with the bottlenecks.

Finally, we evaluate the system schedulability when it has a limited number of physical SMs (i.e., the number of physical SMs is smaller than the number of tasks). Fig. 12 presents the system schedulability and corresponding response time analysis under 12 parallel tasks with 3, 5, and 8 physical SMs. To let each task have at least one allocated virtual SM, we generate four virtual SMs from every physical SM. Not surprisingly, the system suffers a lower schedulability when there are more parallel tasks and fewer GPU SMs. If we compare the schedulability from different numbers of parallel tasks or different numbers of physical SMs, the number of physical SMs has a more impact on the schedulability because the number of physical SMs directly determines the full system utilization rate.

7 Related Work

For real-time systems with GPUs, previous work mainly leveraged GPU kernel-granularity scheduling. For example, Kato [30] introduced a priority-based scheduler. Elliott proposed shared resources and containers for integrating GPU and CPU scheduling [31] and GPUSync [32] for managing multi-GPU multicore soft real-time systems with flexibility, predictability, and parallelism. Golyanik [33] described a scheduling approach based on time-division multiplexing; S^3DNN [34] optimized the execution of DNN GPU workloads in a real-time multi-tasking environment through scheduling the GPU kernels. Thermal and energy efficieny in GPU real-time scheduling systems were studied in [35], [35], [36]. However, these approaches focus on predictable GPU control, and hard to support multiple tasks to use the GPU at the same time. Thus, the GPU may be underutilized and a task may wait a long time to access the GPU.

In the past few years, GPU vendors and researchers started to provide a more flexible GPU kernel execution manner such as temporal preemption and spatial partitioning. For example, NVIDIA started the initial preemption since Pascal architecture and in the recent Tegra architecture for embedded systems, the preemption types can be selected by users according to the application priorities. Besides, researchers also developed many frameworks to further support GPU preemption at kernel or even finer granularity. For example, Park [37], Basaran [38], Tanasic [39], and Zhou [40] proposed architecture extensions with hardware and software codesigns to improve the preemption and tested on the GPU simulators. Capodieci [41] further presented a deadline-based real-time scheduling with the support of preemption on a NVIDIA Drive-PX GPU. The scheduler runs as a software partition on top of the NVIDIA hypervisor and leverages pixel-level preemption and thread-level preemption. This preemptive execution pattern implements and tests a preemptive Earliest Deadline First (EDF) scheduler. Extensive experiments demonstrate that preemptive EDF scheduling achieves significant schedulability improvement. The Effisha framework [42] introduced software techniques without any hardware modification to support kernel preemption at the end of any arbitrary thread block. Meanwhile, targeting embedded systems without hardware nor driver stack extensions, Hartmann [43] also developed a fixed point preemption on GPUs called GPUart and evaluated the Gang-Earliest Deadline First and Gang-Fixed Task Priority scheduling strategies on it. According to the experimental results, up to 221x response time improvements are achieved in GPUart.

On spatial partitioning, NVIDIA launched the MPS and MiG process management software to manage kernel parallel execution. AMD released open-source software support for hardware partitioning which has the potential to accelerate and aid the long-term viability of real-time GPU research [44], [45]. Researchers [12], [15], [16] proposed the persistent thread techniques as discussed in Section 2. Following the persistent thread technique, [46] presents an energy-efficient scheduler sBEET by partitioning the computing resources and isolating kernel execution. Experiments on NVIDIA Jetson Xavier AGX demonstrate the sBEET could reduce deadline misses and energy consumption by up to 13% and 21%. Liang [13] introduced a software-hardware

solution for efficient spatial-temporal multitasking for GPU. However, the computation throughput [8], [9] is usually the focus of GPU spatial partitioning. [27], [47] considers the fine-grained real-time GPU scheduling only with the state-of-the-art system side work (persistent threads) and scheduling analysis. However, SM-granularity resource partitioning without an efficient real-time scheduling algorithm is not sufficient to achieve effective SM-Level scheduling with fine granularity and high utilization rates. According to the related works on temporal access and spatial portioning, temporal access based on preemption has a more flexible GPU access but spatial partitioning could achieve a higher schedulability with the additional requirement that the number of virtual SMs should be larger than the number of parallel tasks.

Although flexible task execution can improve system schedulability, rare work provides a complete solution, which can seamlessly link the system improvement with efficient real-time scheduling algorithms. To obtain more universal and effective real-time GPU scheduling, and to piggyback on previous work, we propose real-time GPGPU scheduling: RTGPU. Compared with previous work, RTGPU leverages architecture information to support finer-grain SM-level scheduling and improves the schedulability and increases the throughput of real-time GPU systems.

8 Conclusion

To execute multiple parallel real-time applications on GPU systems, we propose *RTGPU*—a real-time scheduling method including both system work and and a real-time scheduling algorithm with schedulability analysis. RTGPU leverages a precise timing model of the GPU applications with the persistent threads technique and achieves improved fine-grained utilization through interleaved execution. The RTGPU real-time scheduling algorithm is able to provide real-time guarantees of meeting deadlines for GPU tasks with better schedulability compared with previous work. We empirically evaluate our approach using synthetic benchmarks both via schedulability analysis and on real NVIDIA GTX1080Ti GPU systems, the results of which demonstrate significant performance gains compared to existing methods.

References

[1] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[2] Shih-Chieh Lin, Yunqi Zhang, Chang-Hong Hsu, Matt Skach, Md E Haque, Lingjia Tang, and Jason Mars. The architectural implications of autonomous driving: Constraints and acceleration. In *International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 751–766, 2018.

[3] Nvidia accelerates race to autonomous driving at ces. https://blogs.nvidia.com/blog/2016/01/04/drive-px-ces-recap/ note = Accessed: 2019-11-23.

[4] Omid Hosseini Jafari, Dennis Mitzel, and Bastian Leibe. Real-time rgb-d based people detection and tracking for mobile robots and head-worn cameras. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 5636–5643. IEEE, 2014.

[5] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

[6] Christopher J Rossbach, Jon Currey, Mark Silberstein, Baishakhi Ray, and Emmett Witchel. Ptask: operating system abstractions to manage gpus as compute devices. In *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles*, 2011.

[7] Shinpei Kato, Michael McThrow, Carlos Maltzahn, and Scott Brandt. Gdev: First-class {GPU} resource management in the operating system. In *Presented as part of the 2012 {USENIX} Annual Technical Conference ({USENIX}{ATC} 12)*, pages 401–412, 2012.

[8] Onur Kayiran, Nachiappan Chidambaram Nachiappan, Adwait Jog, Rachata Ausavarungnirun, Mahmut T Kandemir, Gabriel H Loh, Onur Mutlu, and Chita R Das. Managing gpu concurrency in heterogeneous architectures. In *Microarchitecture (MICRO), 2014 47th Annual IEEE/ACM International Symposium on*. IEEE, 2014.

[9] Chao-Tung Yang, Chih-Lin Huang, and Cheng-Fang Lin. Hybrid cuda, openmp, and mpi parallel programming on multicore gpu clusters. *Computer Physics Communications*, 182(1):266–269, 2011.

[10] Deep learning framework by bair. http://caffe.berkeleyvision.org/ note = Accessed: 2019-11-23.

[11] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

[12] Kshitij Gupta, Jeff A Stuart, and John D Owens. A study of persistent threads style gpu programming for gpgpu workloads. In *Innovative Parallel Computing-Foundations & Applications of GPU, Manycore, and Heterogeneous Systems (INPAR 2012)*. IEEE, 2012.

[13] Yun Liang, Huynh Phung Huynh, Kyle Rupnow, Rick Siow Mong Goh, and Deming Chen. Efficient gpu spatial-temporal multitasking. *IEEE Transactions on Parallel and Distributed Systems*, 2014.

[14] Multi-process service. https://docs.nvidia.com/deploy/pdf/CUDA_Multi_Process_Service_Overview.pdf.

[15] Chao Yu, Yuebin Bai, Hailong Yang, Kun Cheng, Yuhao Gu, Zhongzhi Luan, and Depei Qian. Smguard: A flexible and fine-grained resource management framework for gpus. *IEEE Transactions on Parallel and Distributed Systems*, 2018.

[16] Bo Wu, Guoyang Chen, Dong Li, Xipeng Shen, and Jeffrey Vetter. Enabling and exploiting flexible task assignment on gpu through sm-centric program transformations. In *Proceedings of the 29th ACM on International Conference on Supercomputing*. ACM, 2015.

[17] Lea Schönberger, Wen-Hung Huang, Georg Von Der Brüggen, Kuan-Hsun Chen, and Jian-Jia Chen. Schedulability analysis and priority assignment for segmented self-suspending tasks. In *IEEE 24th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, pages 157–167. IEEE, 2018.

[18] Ming Yang, Nathan Otterness, Tanya Amert, Joshua Bakita, James H Anderson, and F Donelson Smith. Avoiding pitfalls when using nvidia gpus for real-time tasks in autonomous systems. In *30th Euromicro Conference on Real-Time Systems (ECRTS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[19] Tanya Amert, Nathan Otterness, Ming Yang, James H Anderson, and F Donelson Smith. Gpu scheduling on the nvidia tx2: Hidden details revealed. In *Real-Time Systems Symposium*. IEEE, 2017.

[20] Nathan Otterness, Ming Yang, Sarah Rust, Eunbyung Park, James H Anderson, F Donelson Smith, Alex Berg, and Shige Wang. An evaluation of the nvidia tx1 for supporting real-time computer-vision workloads. In *2017 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 353–364. IEEE, 2017.

[21] John L Gustafson. Reevaluating amdahl's law. *Communications of the ACM*, 31(5):532–533, 1988.

[22] Achieved occupancy. https://docs.nvidia.com/gameworks/content/developertools/desktop/analysis/report/cudaexperiments/kernellevel/achievedoccupancy.htm.

[23] Gwangsun Kim, Jiyun Jeong, John Kim, and Mark Stephenson. Automatically exploiting implicit pipeline parallelism from multiple dependent kernels for gpus. In *Parallel Architecture and Compilation Techniques, 2016 International Conference on*. IEEE.

[24] Ali Bakhoda, George L Yuan, Wilson WL Fung, Henry Wong, and Tor M Aamodt. Analyzing cuda workloads using a detailed gpu simulator. In *International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 2009.

[25] Jian-Jia Chen and Cong Liu. Fixed-relative-deadline scheduling of hard real-time tasks with self-suspensions. In *2014 IEEE Real-Time Systems Symposium*, pages 149–160. IEEE, 2014.

[26] Konstantinos Bletsas, Neil Audsley, Wen-Hung Huang, Jian-Jia Chen, and Geoffrey Nelissen. Errata for three papers (2004-05) on fixed-priority scheduling with self-suspensions. Technical report, CISTER-Research Centre in Realtime and Embedded Computing Systems, 2015.

[27] Sujan Kumar Saha, Yecheng Xiang, and Hyoseung Kim. Stgm: Spatio-temporal gpu management for real-time tasks. In *2019 IEEE*

25th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA), pages 1–6. IEEE, 2019.
[28] Pratyush Patel, Iljoo Baek, Hyoseung Kim, and Ragunathan Rajkumar. Analytical enhancements and practical insights for mpcp with self-suspensions. In *IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 177–189. IEEE, 2018.
[29] Huixiang Chen, Meng Wang, Yang Hu, Mingcong Song, and Tao Li. Gaas workload characterization under numa architecture for virtualized gpu. In *International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 2017.
[30] Shinpei Kato, Karthik Lakshmanan, Raj Rajkumar, and Yutaka Ishikawa. Timegraph: Gpu scheduling for real-time multi-tasking environments. In *Proc. USENIX ATC*, pages 17–30, 2011.
[31] Glenn A Elliott and James H Anderson. Globally scheduled real-time multiprocessor systems with gpus. *Real-Time Systems*, 2012.
[32] Glenn A Elliott, Bryan C Ward, and James H Anderson. Gpusync: A framework for real-time gpu management. In *2013 IEEE 34th Real-Time Systems Symposium*, pages 33–44. IEEE, 2013.
[33] Vladislav Golyanik, Mitra Nasri, and Didier Stricker. Towards scheduling hard real-time image processing tasks on a single gpu. In *International Conference on Image Processing (ICIP)*. IEEE, 2017.
[34] Husheng Zhou, Soroush Bateni, and Cong Liu. Sˆ 3dnn: Supervised streaming and scheduling for gpu-accelerated real-time dnn workloads. In *IEEE Real-Time and Embedded Technology and Applications Symposium*, pages 190–201. IEEE, 2018.
[35] Muhammad Husni Santriaji and Henry Hoffmann. Merlot: Architectural support for energy-efficient real-time processing in gpus. In *2018 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 214–226. IEEE, 2018.
[36] Seyedmehdi Hosseinimotlagh and Hyoseung Kim. Thermal-aware servers for real-time tasks on multi-core gpu-integrated embedded systems. In *2019 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 254–266. IEEE, 2019.
[37] Jason Jong Kyu Park, Yongjun Park, and Scott Mahlke. Chimera: Collaborative preemption for multitasking on a shared gpu. *ACM SIGARCH Computer Architecture News*, 43(1):593–606, 2015.
[38] Can Basaran and Kyoung-Don Kang. Supporting preemptive task executions and memory copies in gpgpus. In *24th Euromicro Conference on Real-Time Systems (ECRTS 2012)*. IEEE, 2012.
[39] Ivan Tanasic, Isaac Gelado, Javier Cabezas, Alex Ramirez, Nacho Navarro, and Mateo Valero. Enabling preemptive multiprogramming on gpus. In *Computer Architecture (ISCA), 2014 ACM/IEEE 41st International Symposium on*, pages 193–204. IEEE, 2014.
[40] Husheng Zhou, Guangmo Tong, and Cong Liu. Gpes: A preemptive execution system for gpgpu computing. In *Real-Time and Embedded Technology and Applications Symposium*. IEEE, 2015.
[41] Nicola Capodieci, Roberto Cavicchioli, Marko Bertogna, and Aingara Paramakuru. Deadline-based scheduling for gpu with preemption support. In *2018 IEEE Real-Time Systems Symposium (RTSS)*, pages 119–130. IEEE, 2018.
[42] Guoyang Chen, Yue Zhao, Xipeng Shen, and Huiyang Zhou. Effisha: A software framework for enabling effficient preemptive scheduling of gpu. In *Proceedings of the 22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, 2017.
[43] Christoph Hartmann and Ulrich Margull. Gpuart-an application-based limited preemptive gpu real-time scheduler for embedded systems. *Journal of Systems Architecture*, 97:304–319, 2019.
[44] Nathan Otterness and James H Anderson. Exploring amd gpu scheduling details by experimenting with "worst practices",". In *Proceedings of the 29th International Conference on Real-Time Networks and Systems*, 2021.
[45] Nathan Otterness and James H Anderson. Amd gpus as an alternative to nvidia for supporting real-time workloads. In *32nd Euromicro Conference on Real-Time Systems (ECRTS 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.
[46] Yidi Wang, Mohsen Karimi, Yecheng Xiang, and Hyoseung Kim. Balancing energy efficiency and real-time performance in gpu scheduling. In *2021 IEEE Real-Time Systems Symposium (RTSS)*, pages 110–122. IEEE, 2021.
[47] Minseok Lee, Seokwoo Song, Joosik Moon, John Kim, Woong Seo, Yeongon Cho, and Soojung Ryu. Improving gpgpu resource utilization through alternative thread block scheduling. In *High Performance Computer Architecture (HPCA), 2014 IEEE 20th International Symposium on*, pages 260–271. IEEE, 2014.

An Zou is an Assistant Professor at the University of Michigan-Shanghai Jiao Tong University Joint Institute in the Shanghai Jiao Tong University. His research focuses on computer architecture and embedded systems. He broadly investigated a set of techniques and solutions via a bottom-up layered approach to improve computing power and performance efficiency. Dr. An Zou received his Ph.D. degree in Electrical Engineering from Washington University in St. Louis in 2021 and his M.S. and B.S. degrees from Harbin Institute of Technology (HIT) in 2015 and 2013. His work has been extensively published and recognized at top-tier conferences and journals including MICRO, DAC, ICCAD, AAAI, TCAD, TACO, and RTAS. He was a recipient of A. Richard Newton Young Student Fellow Award and the Best Paper Nominations at DAC 2017 and MLCAD 2020.

Jing Li is an Assistant Professor in Department of Computer Science at New Jersey Institute of Technology. She received her Ph.D. at Washington University in St. Louis in 2017, where she was advised by Professor Chenyang Lu and Kunal Agrawal. She received B.S. in computer science from Harbin Institute of Technology in 2011. Her research interests include real-time systems, parallel computing, cyber-physical systems, and reinforcement learning for system design and optimization. She has high impact publications in top journals and conferences with 3 outstanding papers.

Christopher D. Gill is a Professor in the Department of Computer Science and Engineering at Washington University in St. Louis. He has published more than 100 technical articles in selective peer-reviewed conferences and journals, and has led or contributed to the development, evaluation, and open-source release of numerous real-time systems research platforms and artifacts, including: the Kokyu real-time scheduling and dispatching framework that was used in several AFRL and DARPA projects and flight demonstrations; the nORB small-footprint real-time object request broker; a number of real-time and fault-tolerant services for The ACE ORB (TAO) and the Component Integrated ACE ORB (CIAO); the Cyber-physical Instrument for Real-time hybrid Structural Testing (CIRST) that established key foundations for real-time hybrid simulation (RTHS), and the CyberMech platform that built on the CIRST project to enable parallel RTHS at millisecond time scales; and the RT-Xen real-time virtualization research platform and the RTDS scheduler that is now part of the Xen open-source software distribution. Professor Gill has served as an Associate Editor for TCPS and Subject Area Editor for the Elsevier Journal of Systems Architecture. He has served in numerous other organizing and technical reviewing roles within the real-time systems research community, including: IEEE TCRTS Chair; IEEE TCRTS Vice-Chair; IEEE RTSS General Chair; ACM SIGBED Vice-Chair; IEEE RTSS Technical Program Committee Chair; IEEE TCRTS Treasurer and IEEE RTSS Finance Chair.

Xuan Zhang is an Associate Professor in the Preston M. Green Department of Electrical and Systems Engineering at Washington University in St. Louis. She works across the fields of VLSI design, computer architecture, and cyber-physical systems and her research interests include hardware/software co-design for efficient machine learning and artificial intelligence, adaptive power and resource management for autonomous systems in analog/mixed-signal and physical domain. Before joining Washington University, Dr. Zhang was a Postdoctoral Fellow in Computer Science at Harvard University. She received her BE degree in Electrical Engineering from Tsinghua University in China, and her MS and Ph.D. degrees in Electrical and Computer Engineering from Cornell University. Dr. Zhang is the recipient of NSF CAREER Award in 2020, AsianHOST Best Paper Award in 2020, DATE Best Paper Award in 2019, and ISLPED Design Contest Award in 2013, and her work has also been nominated for Best Paper Awards at ASP-DAC 2021, DATE 2019 and DAC 2017.