Simulee: Detecting CUDA Synchronization Bugs via Memory-Access Modeling

Mingyuan Wu
Southern University of Science and Technology
Shenzhen, China
11849319@mail.sustech.edu.cn

Yicheng Ouyang
Southern University of Science and Technology
Shenzhen, China
11610313@mail.sustech.edu.cn

Husheng Zhou
University of Texas at Dallas
Dallas, USA
husheng.zhou@utdallas.edu

Lingming Zhang
University of Texas at Dallas
Dallas, USA
lingming.zhang@utdallas.edu

Cong Liu
University of Texas at Dallas
Dallas, USA
cong@utdallas.edu

Yuqun Zhang*
Southern University of Science and Technology
Shenzhen, China
zhangyq@sustech.edu.cn

1 INTRODUCTION

CUDA [3] is a mainstream parallel computing platform and programming model that allows software developers to leverage general-purpose GPU (GPGPU) computing [4]. CUDA is advanced in simplifying I/O streams to memories and dividing computations into sub-computations since it parallelizes programs in terms of grids and blocks. In addition, CUDA enables more flexible cache management that speeds up the floating point computation of CPUs. CUDA is thus considered rather powerful for accelerating deep-neural-network-related applications where relevant matrix computations can be efficiently loaded.

Unlike traditional CPU multi-thread programs where the synchronization mechanism is built upon managing the shared resource that can be accessed by multiple threads, the typical synchronization mechanism of GPU programs is built upon synchronizing the instruction flows [1]. In particular, since GPU programs use barriers rather than locks for synchronization and enable simplified barrier-based happens-before relations, traditional lockset-based [13, 38] and happens-before-based bug detection approaches [18, 34] for CPUs become obsolete and expensive in detecting parallel-computing-related bugs for GPUs.

Recently, several approaches, e.g., [6, 9, 20, 30, 31, 37, 50, 51], have been proposed to detect synchronization bugs for CUDA kernel functions. In general, they can be categorized into two classes: (1) the compiler-based approaches that leverage compiler instrumentation with/without static analysis to identify race-free locations for detecting data-race bugs [6, 20, 37, 50, 51], and (2) the SMT-solver-based approaches that integrate static analysis and symbolic execution with SMT solver to detect data race and barrier-divergence bugs [9, 30, 31]. Although such approaches can detect CUDA synchronization bugs under their respectively defined environments, they have limited applicability and may rely on certain assumptions. In particular, the compiler-based approaches are limited due to relying on developer-provided test inputs and detecting data races only; while the SMT-solver-based approaches can be heavyweight due to triggering expensive static-analysis overhead and may also involve manually-provided test inputs.

In this paper, to address the aforementioned issues, we develop a systematic lightweight bug detection framework, namely Simulee [7],...
which automatically detects general synchronization bugs for CUDA kernel functions. In particular, Simulee generates a Memory-Access Model that maintains thread-wise memory-access information including thread id, visit order, and action. Accordingly, Simulee utilizes the LLVM bytecode of CUDA kernel functions to initialize the running environmental setups including arguments, dimensions, and global memory if necessary based on the Memory-Access Model. Moreover, Simulee applies Evolutionary Programming [21] to approach error-inducing inputs for exposing the dangerous program execution paths which may possibly induce synchronization bugs. Subsequently, Simulee collects the corresponding memory-access information from such paths. At last, by combining CUDA specifics, such collected memory-access information are analyzed to find whether they can potentially lead to synchronization bugs.

Compared with other CUDA synchronization bug detection approaches, Simulee can detect multiple bug types including data race, redundant barrier function, and barrier divergence fully automatically. Moreover, Simulee benefits from exploring the dangerous execution paths guided by the lightweight evolutionary search, without incurring large overhead (e.g., for constraint solving), such that it is more efficient than existing state-of-the-art approaches [9, 30, 31] that usually rely on heavyweight techniques or manual inputs.

To evaluate the effectiveness and efficiency of Simulee, we first collect 7 popular CUDA-related projects from GitHub (with a total of 17928 commits and 1.36 million LOC by Aug, 2019) as our benchmark suite. Then, as a preliminary study, we manually identified 24 real-world synchronization bugs from a subset of the studied real-world CUDA projects and the GKLEE [30] benchmark suite. We conduct a set of experiments to explore (1) how many of those manually identified bugs can be detected by Simulee for the preliminary analysis; (2) whether Simulee can detect previously-unknown bugs on those real-world CUDA projects; and (3) how Simulee compares with state-of-the-art approaches in CUDA bug detection. The experimental results suggest that Simulee can successfully detect 21 of the 24 manually identified synchronization bugs in our preliminary study. Furthermore, it even detects 24 previously-unknown bugs from all the 7 real-world CUDA projects, 10 of which have already been confirmed by the corresponding developers. The experimental results also demonstrate that Simulee can be much more effective than state-of-the-art approaches, i.e., GKLEE [30], GKLEE-SESA [31], GPUVerify [9], and RaceChecker [6], in detecting synchronization bugs, e.g., Simulee can detect 2X more previously unknown bugs than any of the other studied state-of-the-art approaches. In addition, Simulee can significantly outperform the other studied approaches in terms of the runtime overhead.

In summary, our paper makes the following contributions:

- **Idea.** To the best of our knowledge, we propose the first idea of CUDA synchronization bug detection via evolutionary search guided by memory-access modeling.
- **Implementation.** We have implemented the proposed idea as a lightweight, fully automated, and general-purpose detection framework for CUDA synchronization bugs, namely Simulee, that can automatically detect a wide range of synchronization bugs in CUDA programs which are hard to be captured manually.

2 BACKGROUND

In this section, we give an overview on CUDA, the CUDA parallel computing mechanism, and typical CUDA synchronization bugs.

CUDA Overview. CUDA is a parallel computing platform and programming model, which allows developers to use GPU hardware for general-purpose computing, e.g., autonomous driving [48, 52, 53]. CUDA is composed of a runtime library and an extended version of C/C++. In particular, CUDA programs are executed on GPU cores, namely “device”, while they also need to be allocated with resources on CPUs, namely “host”, prior to execution. As a result, developers need to retrieve allocated resources such as global memory after CUDA program execution. To conclude, a complete CUDA program contains 3 runtime stages: (1) host resource preparation, (2) kernel function execution, and (3) host resource retrieval.

CUDA Parallel Computing Mechanism. A CUDA kernel function refers to the part of CUDA programs that runs on the device side. Specifically, thread is the kernel function’s basic execution unit. At the physical level, 32 threads are bundled as a thread warp wherein all the threads execute the same statement at any time except undergoing a branch divergence, while at the logic level, one or more threads are contained in a block, and one or more blocks are contained in a grid. The execution of kernel functions is initialized by setting runtime environments, e.g., dimensions of grids and blocks, passing relevant arguments, such that the computation can be divided into sub-computations and each sub-computation can be dispatched to different threads. Eventually, the results of sub-computations can be merged as the final result of the overall computation through applying algorithms such as reduction [35]. The hierarchy of the parallel computing mechanism of CUDA kernel functions is presented in Figure 1.

To synchronize threads, CUDA applies barriers at which all the threads in one block must wait before any can proceed. In CUDA kernel functions, the barrier function is “_syncthreads()” which synchronizes threads from the same block. When a thread reaches a barrier, it is expected to proceed to next statement if and only if all the threads from the same block have reached the same barrier. Otherwise, the program would be exposed to undefined behaviors.
CUDA Synchronization Bugs. There are three major synchronization bugs in CUDA kernel functions: data race, barrier divergence [14], and redundant barrier function [1]. Specifically, data race indicates that for accessing global or shared memory, CUDA cannot guarantee the visit order of “read&write” actions or “write&write” actions from two or more threads. For example, Figure 2 demonstrates the bug-fixing Revision no. “feb5f15a82” in the file “smo-kernel.cu” of the project “thunder/svm” [42], one of the highly-rated GitHub projects. It can be observed from Figure 2 that the “if” statement writes to the memory of “f_val2reduce”, while inside the device, the function “get_block_min” writes to the same memory. This “write&write” bug is fixed by adding “__syncthreads()” which synchronizes actions among threads.

A barrier function is considered redundant when there is no data race after deleting it from source code. A redundant barrier function compromises the program performance in terms of time and memory usage. For instance, Figure 3 demonstrates bug-fixing Revision no. “31761d27f01” in file “kernel/homography.hpp” from project “arrayfire” [8]. It can be observed that the block is one-dimensional from Line 1, the value of “tid” is assigned only by “threadIdx.x”. That indicates that the “tid’s are identical among different threads from the same block. As a result, “s_median[tid]” and “s_idx[tid]” can only be accessed by one thread, leading to a redundant barrier function in Line 4 because there is no race in “s_median” or “s_idx” after deleting it.

A barrier divergence takes place when some threads in a block complete their tasks and leave the barrier while the others have not reached the barrier yet. Figure 4 demonstrates the bug-fixing Revision no. “9e68ccc5ff” in the file “nearest_neighbour.hpp” from the project “arrayfire” caused by barrier divergence. It can be indicated from Figure 4 that developers make sure all the threads in the same block reach the same barrier in every execution of the kernel function by moving the statement of “__syncthreads()” outside the given branch. Otherwise they will have to handle undefined behaviors.

3 FRAMEWORK OF SIMULEE

In this section, we introduce Simulee, a lightweight, automatic, and device-independent framework to detect real-world CUDA synchronization bugs. Typically, Simulee takes LLVM bytecode translated from CUDA kernel function programs as input. Then, it automatically generates the associated error-inducing test inputs, and yields Memory-Access Model to detect synchronization bugs. Specifically, Simulee is composed of two components—“Automatic Input Generation” and “Bug Detection via Memory-Access Model”. “Automatic Input Generation” is initialized by inputting the LLVM bytecode of CUDA kernel function programs. Next, it slices the memory-access statements (e.g., read and write statements) and inputs them for Evolutionary Programming [21]. Subsequently, Evolutionary Programming helps generate error-inducing environmental setups by iteratively mutating and sorting dimensions/arguments and passes the acceptable ones to “Bug Detection via Memory-Access Model”. At last, “Bug Detection via Memory-Access Model” traces real execution paths by using the error-inducing inputs and collects the memory-access information from the paths to detect whether there are synchronization bugs, as it were “simulating” runtime environment. The details can be found in Figure 5.

3.1 Automatic Input Generation

Generating potentially error-inducing inputs is essentially equivalent to generating the inputs that can lead to the memory-access
conflicts among threads to improve the possibility of CUDA synchronization bug occurrences. However, how to automatically generate such error-inducing inputs remains challenging. Intuitive solutions, e.g., random generation, coverage-oriented generation, can be limited in effectiveness and efficiency, because they are not specially designed for triggering memory-access conflicts. In this section, we introduce how Simulee automatically generates potentially error-inducing inputs for exposing the dangerous program execution paths which could lead to synchronization bugs in an effective and efficient manner.

3.1.1 Intuition. An effective and efficient automatic approach to generate potentially error-inducing inputs for triggering CUDA synchronization bugs implies generating as many memory-access conflicts as possible within a short time limit. Given the ith memory address and the kernel function inputs, i.e., grid and block dimensions and arguments, \( f(i) \) is defined as the number of threads that access the ith memory address while \( g(i) \) is a function that returns 1 when the ith memory address is accessed by any thread and returns 0 otherwise. \([\text{start, end}]\) denotes the memory-access range. An intuitive target function \( F(dimensions, arguments) \) can be presented in Equation 1 which denotes the ratio of the total number of the accessed memory addresses to the total number of the memory-access threads:

\[
F(dimensions, arguments) = \frac{\sum_{i=\text{start}}^{\text{end}} g(i)}{\sum_{i=\text{start}}^{\text{end}} f(i)} \tag{1}
\]

It can be derived that the max value of \( F(dimensions, arguments) \) is 1 which denotes that there is no memory-access conflict between any thread pair. On the other hand, the smaller \( F(dimensions, arguments) \) is, the higher chance the memory-access conflict takes place. Therefore, \( F(dimensions, arguments) \) can be used for optimization to obtain error-inducing inputs that trigger CUDA synchronization bugs. Note that since \( F(dimensions, arguments) \) is discrete, we choose Evolutionary Programming [44] as our optimization approach.

3.1.2 Algorithm. The framework of “Automatic Input Generation” is presented in Algorithm 1. First, Simulee randomly initializes arguments and dimensions to create and sort individual solutions for evolving (Lines 3 to 7). In each generation, each solution is mutated to generate two children, which are added to the whole population set (Lines 8 to 14). Next, the population winners survive for the subsequent iterations (Lines 15 to 16). The iterations can be terminated once it finds an acceptable solution. Otherwise, after completing the iterations, it returns the optimal solution.

**Initial Solutions.** The initial dimensions and arguments are randomly generated and passed to fitness functions as initial solutions for future evolution. Note that the dimensions can be extracted from kernel functions. For instance, if a kernel function has “threadIdx.x” and “threadIdx.y”, it means the block is two-dimensional.

**Fitness Function.** Equation 1 is chosen as the primary fitness function for Evolutionary Programming. Specifically, the output of \( F(dimensions, arguments) \) is the fitness score for a solution of dimensions and arguments in Evolutionary Programming. However, it is difficult to derive an optimal solution of dimensions and arguments by only optimizing \( F(dimensions, arguments) \). In particular, since \( F(dimensions, arguments) \) is non-differentiable when the gradient does not exist, it is hard to find an optimal solution given the set of inferior solutions, e.g., all the solutions of \( F(dimensions, arguments) \) are “1”s. To address such issues, we design a secondary fitness function such that they are sorted according to their possibility to be optimal: \( R(\text{start, end}) = \text{end} - \text{start} \). In particular, it indicates that a smaller memory-access range leads to a higher possibility of memory-access conflict. As a result, we define fitness score of the primary fitness function as primary score, and the fitness score of the secondary fitness function as secondary score. During the population evaluation, the primary score is sorted first; if and only if the top-ranked primary score is 1, the secondary score is sorted to decide which solution is more likely to converge to the minimum of \( F(dimensions, arguments) \).

**Mutation.** In Simulee, solutions are generated by mutation, where each solution generates two children in one generation. Specifically, arguments and dimensions are independent from each other during mutation with respective mutation strategies. The mutate strategy for dimensions is trivial: first, Simulee randomly generates an integer vector ranging from -1 to 1 according to the dimension size; next, the child’s dimension is mutated by summing the parent’s dimension and the generated integer vector.

The details of the mutation strategy for arguments is presented in Algorithm 2. Since the memory-access-relevant arguments are numbers, Simulee views them as floating numbers and converts them back to their actual types when executing \( f(i) \). Accordingly, each generation generates two children: one adds a random number generated by standard Normal Distribution \([5]\) \( \mathcal{N}(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \) to the arguments inherited from the parent solution, and the other adds a random number generated by standard Cauchy Distribution \([2]\) \( \mathcal{C}(x) = \frac{1}{\pi(x^2+1)} \) to the arguments inherited from the parent solution. We define the search step length of the arguments as
We next explain why we apply the above two distributions. It can visit_order read four Unit Tuple action of such threads, and thread_id ory address from different threads, <a three-dimensional vector space and is composed of a set of Unit by the kernel functions is defined to be composed of a set of The memory-access model that depicts thread-wise memory-access instances. Based on the Memory-Access Model, Simulee develops a set of criteria to detect synchronization bugs and trigger CUDA synchronization bugs.

3.2 Memory-based Synchronization Bug Detection

With the auto-generated error-inducing inputs, the synchronization bug detection of Simulee is established on building a Memory-Access Model that depicts thread-wise memory-access instances. Based on the Memory-Access Model, Simulee develops a set of criteria to detect synchronization bugs including data race, redundant barrier functions, and barrier divergence.

3.2.1 Memory-Access Model. The Memory-Access Model accessed by the kernel functions is defined to be composed of a set of Memory Units where each Memory Unit corresponds to a memory address and is composed of a set of Unit Tuples. A Unit Tuple is defined as a three-dimensional vector space <visit_order, thread_id, action>, where visit_order represents the visit order to the associated memory address from different threads, thread_id represents the indices of such threads, and action refers to the read or write action from those threads.

An example of Memory Unit is demonstrated in Figure 6 with four Unit Tuples <0, (1 0 0), read>, <0, (2 0 0), read>, <0, (3 0 0), read>, and <1, (3 0 0), read> where threads (1,0), (2,0), and (3,0) read the same memory address in the same visit_order since none of them have reached any barrier function before they read. Assume all the threads reach a barrier function later and thread (3 0 0) reads, the visit_order is then incremented from 0 to 1 for thread (3 0 0) and the other threads afterwards.

3.2.2 Memory Accessing Model Construction. Since Memory-Access Model is only associated with barrier functions and memory-access statements, it is applicable to detect synchronization bugs by obtaining such statements and then extracting/analyzing the memory-access information instead of executing the complete CUDA programs on GPUs, i.e., modeling the execution of CUDA kernel functions. This modeling process is initiated by inputting the auto-generated block and grid dimensions and arguments passed to the kernel functions. Next, it constructs the Memory Unit for each memory address by tracing back the execution path for each thread.

The overall Memory-Access Model construction is demonstrated in Algorithm 3. In particular, the algorithm is launched to initialize the block and grid dimensions as well as the global and shared memory for each thread (Lines 2 to 5). Next, for each block, the shared memory (Line 7) and the thread-wise visit_order for each global and shared memory address (Lines 8 to 9) are initialized. If there are still some unterminated threads, for all of them, their corresponding Memory Units are derived based on the collected parameters, e.g., globalMem and visitOrderGlobal (Lines 10 to 14). The construction of the thread-wise Memory Units for shared memory and global memory are completed if there is no running thread left (Lines 15 to 16).

Algorithm 4 illustrates the details of Memory-Access Model construction for a single thread. Specifically, given a running thread and the parameters passed by Algorithm 3 (Lines 2 to 4), Algorithm 4 is initialized by detecting whether the current statement is the end of file. If so, the thread would be terminated. If there is any thread halting afterwards, we can confirm there is a “barrier divergence” bug because that indicates at least a thread has not reached the barrier function where the other threads of the same block all have completed their tasks and left (Lines 5 to 9).

If the current statement calls barrier function and all the other threads have reached the same barrier function, the visit_order for both global and shared memory would be incremented if they have been visited before (Lines 10 to 13), since it indicates that (1) all the threads in one block have visited the current memory address and (2) the subsequent visits in the same block would be made rigorously later than the previous visits. On the other hand, if the
Algorithm 3 Memory-Access Model construction

Input: gridDim, blockDim, arguments

Output: Memory-Access Model

1: function CONSTRUCT_MEMORY_MODEL
2: BLOCKS ← generateFromDimension(gridDim)
3: THREADS ← generateFromDimension(blockDim)
4: globalMem ← [MemoryUnit() for i in range(globalSize)]
5: sharedMemLst ← list()
6: for blk in BLOCKS do
7:    sharedMem ← [MemoryUnit() for i in range(sharedSize)]
8:    visitOrderGlobal ← [0 for i in range(globalSize)]
9:    visitOrderShared ← [0 for i in range(sharedSize)]
10:   while hasUnterminatedThread() do
11:       for t in THREADS do
12:           env ← Environment(arguments)
13:           PROCESS_THREAD(t, globalMem, sharedMem, env)
14:           visitOrderGlobal, visitOrderShared, env
15:       sharedMemLst.append(sharedMem)
16:   return globalMem, sharedMemLst

The Details of how to detect data race and redundant barrier function based on Memory-Access Model are presented in Algorithm 5. For each Memory Unit, to detect data race, Simulee first groups the Unit Tuples with the same visit_order. For all the Unit Tuples in one group, Simulee checks whether any Unit Tuple has data race with others according to Theorem 3.1 (Lines 4 to 16). To detect redundant barrier function of one Memory Unit, Simulee extracts its visit_order and groups all the Unit Tuples with adjacent visit_order to find out whether any data race can take place (Lines 18 to 22). If there is no data race, Simulee identifies the associated barrier function and increments its recorder by 1 (Lines 23 to 24). At last, it checks whether the total recorder number matches the total number of the changing visit_order caused by that barrier function which can be obtained after constructing the Memory-Access Model. This barrier function is redundant if the two numbers are equivalent (Lines 25 to 28).

Barrier Divergence. As mentioned in Section 3.2.2, barrier divergence can be detected during constructing Memory-Access Model when there is any halting thread after the current execution is terminated, because it indicates that there is at least one thread which has not reached the barrier function while the others have already left.

To conclude, Simulee first applies Evolutionary Programming to generate error-inducing grid and block dimensions and arguments. Next, Simulee inputs such dimensions and arguments to construct Memory-Access Model, Simulee can detect data race in CUDA kernel functions as described in Theorem 3.1.

Theorem 3.1. Given two Unit Tuples \( \psi_i \) and \( \psi_j \) from the identical Memory Unit, a data race between them takes place if the conditions below are met:

- \( \psi_i[\text{visit}_\text{order}] = \psi_j[\text{visit}_\text{order}] \)
- \( \psi_i[\text{thread}_\text{id}] = \psi_j[\text{thread}_\text{id}] \)
- \( \psi_i[\text{action}] = \text{‘write’} \) or \( \psi_j[\text{action}] = \text{‘write’} \)

when the threads of \( \psi_i \) and \( \psi_j \) are (1) from different thread warps or (2) executing the “write” action on the same statements in the same thread warp or (3) underwent branch divergence before the current “write” action.

Redundant Barrier Function. A redundant barrier function indicates that no data race can be detected by removing that barrier function. In CUDA kernel functions, the visit_order is incremented for one Unit Tuple when at least one thread reaches a barrier function. In other words, two Unit Tuples with adjacent visit_order in one Memory Unit indicates the presence of a barrier function, shown in Figure 6. Therefore, to detect whether a barrier function is redundant or not, it is essential to collect all the associated Unit Tuples and analyze whether they together would lead to data race. The barrier function is defined to be redundant if no data race can be detected among such Unit Tuples.

4 EVALUATION

In this section, we conduct an extensive experimental study to evaluate the effectiveness and efficiency of Simulee in detecting synchronization bugs of CUDA kernel functions. In particular, we first perform a preliminary study on 24 manually identified CUDA synchronization bugs to explore the efficacy of Simulee. Next, we explore the capability of Simulee in detecting previously-unknown bugs from all the real-world CUDA projects in our benchmark suite. Furthermore, we also compare Simulee with multiple existing state-of-the-art approaches to explore whether Simulee can outperform them.

4.1 Benchmark Construction

To conduct a preliminary study for evaluating Simulee, it is essential to establish a set of CUDA synchronization bugs as the ground truth. To this end, we first consider an existing benchmark, i.e., the GKLEE dataset [22], and select 4 synchronization bugs that can represent all the basic synchronization bug patterns in the dataset. Furthermore, we augment the bug dataset with more real-world CUDA bugs. In this paper, we aim to collect important and influential real-world CUDA benchmark projects for our evaluation by defining a set of policies for selecting open-source CUDA.
projects in GitHub. Specifically, we initialize our CUDA project collection by searching the keyword "CUDA" and collect more than 12,000 projects from GitHub in the first place. Next, we sort these projects in terms of the star number and commit number. We randomly select 7 projects with large star/commit numbers. As a result, we collect "kaldi" [29], "arrayfire" [8], "thundersvm" [42], "cuda-cnn" [54], "cudaSift" [11], "cudpp" [15] and "gunrock" [24] for augmenting our benchmark suite as listed in Table 1.

More specifically, we randomly select projects "arrayfire", "kaldi", and "thundersvm" from our real-world CUDA benchmark suite to retrieve their historical synchronization bugs. Note that we do not consider all the projects from our benchmark suite since the manual bug retrieval process can be quite time-consuming. The synchronization bugs for those selected projects are identified based on their commit messages and "git diff" results. The specific operations are listed as follows. Following prior study on other types of bugs [49], we first filter the commits and only keep the commits with the messages that contain at least one keyword in the set {"fix", "error", "sync"} to retain the commits that have higher chances to contain synchronization bugs. However, the commit messages only with these keywords might not be relevant with CUDA bugs. Therefore, next, among the filtered commit messages, we further filter them according to whether they have at least a keyword in the set {"_global", "_device", "__global", "__device"} or match at least one regular expression in the set {"cuda\w+\s*\[\(\]","\<\<\<\[\(\)\]"} with its parent node’s "git diff" results. To illustrate, "_global" is the modifier of kernel functions and "_device" is the modifier of the device functions that can be called by kernel functions. "cuda\w+\s*\[\(\]" is designed in accordance with the information that the resource is prepared/released in host side before/after executing kernel functions. For instance, "cudaMalloc((void **) &host, sizeof(int) *100)" allocates a global 400-byte memory for kernel functions before execution; "cudaFree(&host)" releases the allocated memory for kernel functions after execution. "\<\<\<\[\(\)\]" is designed in accordance with the scenario that sets up the environment for kernel functions, e.g., "function<<<grid_size, block_size>>>(arguments)". All these regular expressions together deliver the complete life cycle of executing kernel functions such that all the bugs of the whole life cycle can be covered. We further manually review all the remaining commits after the above two rounds of filtering to remove any potential false positive. Due to the tedious and time-consuming manual inspection, all the selected CUDA projects are analyzed within the most recent 1000 commits or all of them if there are less than 1000 commits. As a result, we collected a total of 20 real-world

### Table 1: Subject Statistics

<table>
<thead>
<tr>
<th>Projects</th>
<th>Star Number</th>
<th>Commit Number</th>
<th>LoC</th>
</tr>
</thead>
<tbody>
<tr>
<td>kaldi</td>
<td>6860</td>
<td>8681</td>
<td>364K</td>
</tr>
<tr>
<td>arrayfire</td>
<td>2791</td>
<td>5314</td>
<td>381K</td>
</tr>
<tr>
<td>thundersvm</td>
<td>1014</td>
<td>827</td>
<td>343K</td>
</tr>
<tr>
<td>cuda-cnn</td>
<td>368</td>
<td>135</td>
<td>12K</td>
</tr>
<tr>
<td>cudaSift</td>
<td>123</td>
<td>247</td>
<td>24K</td>
</tr>
<tr>
<td>cudpp</td>
<td>268</td>
<td>302</td>
<td>58K</td>
</tr>
<tr>
<td>gunrock</td>
<td>550</td>
<td>2422</td>
<td>178K</td>
</tr>
</tbody>
</table>
CUDA bugs. By combining with the 4 synchronization bugs from the GKLEE benchmark suite, we obtain a total of 24 CUDA bugs as the ground truth for our preliminary study.

4.2 Environment Setups

We performed our evaluation on a desktop machine, with Intel(R) Xeon(R) CPU E5-4610 and 320 GB memory. The operating system is Ubuntu 16.04. For the Evolutionary Programming settings of “Automatic Input Generation” in Simulee, the population and generation are both set to be 10 by default. Note that the Simulee webpage [7] includes more experimental results under different settings for the Evolutionary Programming component.

We select state-of-the-art CUDA synchronization bug detection approaches, i.e., GPUVerify [9], GKLEE [30], GKLEE-SESA [31], and RaceChecker [6] for the performance comparison with Simulee. More specifically, RaceChecker is the NVIDIA’s official tool and represents state-of-the-art compiler-based approach, while the rest approaches represent state-of-the-art SMT-solver-based approaches. Note that the timeout for all studied techniques are uniformly set to be 5 hours.

4.3 Result Analysis

4.3.1 Preliminary Study on Known Bugs. We first present the experimental results for detecting the 24 manually identified synchronization bugs in our preliminary study in Table 2. In the table, DR, BD, and RB respectively denote data race, barrier divergence, and redundant barrier function. “✓”, “✗”, and “F” respectively denote the successful, failed, and false-positive bug-detection attempts. TO denotes that the associated bug detection attempt incurs timeout, while N/A denotes that the buggy kernel function is out of the scope of the bug-detection capability of the corresponding technique. Note that each row represents one kernel function, which could include multiple bugs (indicated by Column “Bug Num”). For such cases, we present all the reported bugs for each kernel function, e.g., “FFF” denotes that the corresponding technique reports four bugs in total for the kernel function, 3 of which are false positives.

From the table, we can observe that, in terms of the overall effectiveness, Simulee can detect 21 (87.5%) of the the 24 manually identified synchronization bugs within seconds, e.g., Simulee at most costs 10.73 seconds (when applied on kernel function JacobSVD). We further analyze the 3 cases for which Simulee fails to detect bugs. Simulee fails to detect the data race in kernel function hamming_matcher because this bug can only be exposed under occasional branch coverage, which is out of the scope of the input generation component of Simulee that focuses on the memory-access conflict potentials. Simulee fails to detect bugs for kernel function softmax_reduce and div_rows_vec because they are largely reimplemented to reduce the usage of unnecessary barrier functions while the current version of Simulee is not designed for such challenging bugs requiring large code refactoring.

We next analyze the comparison results between Simulee and state-of-the-art GPUVerify, GKLEE, GKLEE-SESA and RaceChecker from the table in details. Note that GPUVerify requires user-provided dimension settings and the other techniques require the overall user-provided environmental settings. We provide them all with the ideal settings that can trigger the most possible bugs for fair comparison with Simulee. GPUVerify. GPUVerify is designed to detect data-race and barrier-mergence bugs via integrating static analysis with SMT solvers. From the table, we can observe that GPUVerify can successfully detect 17 out of the 24 synchronization bugs.

Meanwhile, GPUVerify also reports 6 false positives. We next manually check all such false positives and observe that GPUVerify tends to report false positives due to two reasons. First, the false positives are triggered by the bottleneck of the static analysis optimization. For instance, in kernel functions deadLock_0 and computeDescriptor, GPUVerify reports false positives due to nonexistent execution paths. Second, its adopted SMT solvers are used to detect data-race bugs caused by thread-wise access conflicts; however, the thread warp mechanism adopted in CUDA kernel functions can resolve some of such bugs, e.g., the inter-instruction “read&write” race. Shown in Figure 7, warpSize is equal to 32 in Revision 5cc9731af4f of function _trace_mat_mat_trans from project kaldi. Suppose there are two threads (0 0 0) and (1 0 0), and the tid of thread (0 0 0) is 0 while the tid of thread (1 0 0) is 1. Moreover, when the loop terminates, shift is set to 1. Meanwhile, for thread (0 0 0), statement ssum[0] += ssum[0 + 1] is executed; for thread (1 0 0), statement ssum[1] += ssum[1 + 1] is executed. In traditional CPU programs, since thread (0 0 0) is reading data from thread ssum[1] while thread (1 0 0) is attempting to write data to ssum[1], it can incur a “read&write” data race. However, in CUDA kernel functions, since thread (0 0 0) and thread (1 0 0) are located in the same warp without branch divergence, they essentially are executing the same instruction at the same time. Since the statement ssum[tid] += ssum[tid + shift] can be compiled to the following two instructions %1 = load ssum[tid+shift]; store %1 ssum[tid], the “read” action is executed strictly prior to the “write” action. As a result, it turns to be a false-positive data race in CUDA kernel functions. We issued this case to its corresponding developers who further verified our finding as follows:

“You are correct that ssum[0] += ssum[0 + 1] and ssum[1] += ssum[1 + 1] are executed at the same time. But since we are now in a warp, all the 32 threads (tid=0…31) are synchronized. So reading data from ssum[1] and ssum[2] always happens before writing data to ssum[0] and ssum[1] for thread tid=0 and tid=1.” — kaldi

GKLEE and GKLEE-SESA. GKLEE and GKLEE-SESA are designed to detect synchronization bugs via integrating concolic execution with SMT solvers. By launching the environmental setups, they collect “read” and “write” statements into different sets and use SMT solvers to detect synchronization bugs accordingly. From our preliminary study, GKLEE and GKLEE-SESA can detect 16 and 4 manually identified bugs, respectively, without any false positive. Interestingly, GKLEE-SESA incurs more timeouts and detects less bugs than GKLEE. The reason is that GKLEE-SESA leverages more static analysis techniques to reduce its dependency on the initial user-provided environmental setups while such techniques are rather time-consuming. Moreover, similar to GPUVerify, because they are both SMT-solver-based approaches, they are also likely to report false positives on data race (further confirmed by our later study
on detecting new bugs) and fail to detect any redundant-barrier-function bugs because their detecting mechanism is not designed for such bugs. We next discuss the failure cases for the better GKEE because GKEE-SEAS timed out on most cases: GKEE fails to detect barrier-divergence bugs in kernel functions select_matches and _add_diag_mat_mat because GKEE models the kernel functions over two parametric threads and tends to ignore important execution paths for detecting barrier-divergence bugs.

**RaceChecker.** RaceChecker is a device-dependent tool designed only for detecting CUDA data-race bugs. Specifically, it can only detect the data-race bugs incurred in shared memory. In particular, RaceChecker can detect all the 10 data-race bugs that are relevant to shared memory, but fails to detect other synchronization bugs including the data-race bugs incurred in global memory from the manually identified bugs for our preliminary study. Note that since RaceChecker does not involve SMT solver, it does not report any false-positive data-race bug as the other SMT-solver-based approaches.

4.3.2 Further Study on Previously Unknown Bugs. After our preliminary study, we further apply Simulee and all the compared techniques to detect previously unknown synchronization bugs for all the 7 projects with the results demonstrated in Table 3, which follows the same format as Table 2. From the table, we can observe that Simulee detects 24 bugs and reports 3 false positives in total, where all the false positives are redundant-barrier-function bugs in kernel function FindMaxCorr. Note that Simulee reports such false positives because it is possible that “Automatic Input Generation” may miss some potential error-inducing inputs that can trigger synchronization bugs such that a barrier function can be reported as a redundant barrier function. Meanwhile, even the most effective existing technique, i.e., GPUVerify can only detect 11 previously unknown bugs with 2 false positives.

We have also reported all the detected bugs to the corresponding developers and show their feedback statistics in Table 4. To date, they have confirmed 10 bugs in total (TT), including 1 data-race bugs (DR), 1 barrier-divergence bug (BD), and 8 redundant-barrier-function bugs (RB). To be specific, the developers of cudpp and CudaSift responded as follows:

"I think you’re right... There are considerably faster ways to do matrix multiply calls..." — cudpp
Table 4: Developer Feedback Statistics

<table>
<thead>
<tr>
<th>Projects</th>
<th>Detected</th>
<th>Confirmed</th>
<th>Under Discussion</th>
<th>Nonresponse</th>
</tr>
</thead>
<tbody>
<tr>
<td>aldi</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>thundercvn</td>
<td>3 3 0 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>CudaSift</td>
<td>4 0 4 0</td>
<td>4 0 4 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>CUDA-CNN</td>
<td>1 0 1 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>cudapp</td>
<td>1 0 3 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>gunrock</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>Total</td>
<td>17 14 7 3</td>
<td>10 13 5</td>
<td>5 0 0</td>
<td>4 3 1</td>
</tr>
</tbody>
</table>

Figure 7: An Example of False Race

```
1  const int32_t cuda_tid = threadIdx.x * blockDim.x
   + threadIdx.y;
2  ... 
3  if (tid < warpSize) {
4      # pragma unroll
5      for ((int shift = warpSize; shift > 0; shift >>= 1) {
6          ssum[tid] += ssum[tid + shift];
7      }
8  } 
9  ...
```

In summary, it can be observed from Section 4.3.1 and Section 4.3.2 that Simulee can correctly detect most of the synchronization bugs while the other approaches are limited in their detection scopes, failing to detect certain bugs that are designed for, or triggering additional false positives. We now summarize the reasons why Simulee can outperform the other approaches: Simulee applies lightweight evolutionary test generation (guided by effective memory-access modeling) and dynamic runtime monitoring in tandem for powerful CUDA synchronization bug detection. On the other hand, the other approaches are usually bounded by heavyweight techniques, such as constraint solvers, which prevent the techniques from exploring all the possible cases.

5 THREATS TO VALIDITY

The threats to external validity mainly lie in the subjects and faults used in our benchmark. Though the projects of our benchmark suite may not represent the overall project distributions, they shall be selected in order to possibly maximize the overall features of the CUDA projects. In this way, our benchmark is derived based on real-world programs from GitHub, i.e., we select seven popular CUDA-related projects with a total of 17928 commits and 1.36 million LOC.

The threats to internal validity mainly lie in the potential bugs in our implementation due to the complicated mechanism of Simulee.

To reduce the threats, three graduate students, closely mentored by three SE/Systems supervisors, have been carefully working for over one year. We manually reviewed all our implementation code and also included corresponding tests for verifying our implementation. In addition, the effectiveness of “Automatic Input Generation” can impact on the performance synchronization bug detection. It is possible that the “Automatic Input Generation” component may miss certain inputs that can trigger synchronization bugs such that Simulee would miss detecting the bug or report a false positive. To reduce this threat, we set a large number of suitable parameters for the evolutionary algorithm adopted in “Automatic Input Generation” to reduce the probability of missing error-inducing inputs.

To threats to construct validity mainly lie in the metrics used in this work. To reduce the threats, we measure the number of both previously known and the identified bugs detected by the studied techniques as well as their false-positive rate and corresponding time cost.

6 RELATED WORK

As our work investigates the automatic bug detection techniques for CUDA programs, the related work includes the following two parts: empirical studies on CUDA programs and techniques of CUDA bug detection. Moreover, since Simulee essentially is a search-based bug-detection technique, we also discuss such relevant work.

Empirical studies for CUDA programs

There are several existing work that study bugs and other features on CUDA programs. For instance, Yang et al. [43] delivered the empirical study on the features of the performance bugs on CUDA programs, Burtscher et al. [10] studied the control-flow irregularity and memory-access irregularity and found that both irregularities are mutually dependent and exist in most of kernels. Che et al. [12] examined the effectiveness of CUDA to express with different sets of performance characteristics. Some researchers are keen on the comparisons between CUDA and OpenCL. For instance, Demidov et al. [17] compared some C++ programs running on top of CUDA and OpenCL and found that they work equally well for problems of large size. Du et al. [19], on the other side, studied the discrepancies in the OpenCL and CUDA compilers’ optimization that affect the associated GPU computing performance. In our previous work, we also conducted empirical studies to explore CUDA program features. For instance, we investigated the features and the distribution of multiple CUDA program bug types based on a collected GitHub dataset [41]. Moreover, we developed an approach that can automatically repair CUDA synchronization bugs via program transformation and validated its performance via an experimental study based on real-world benchmarks [40].

CUDA bug detection

Unlike traditional program bugs which can be deterministically tested [39, 46] and debugged [23, 32, 47], CUDA synchronization bugs, especially data race and barrier divergence can often result in undefined behaviors. To detect such bugs, there are typically two types of approaches—compiler-based approaches and static analysis (SMT solver)-based approaches. In particular, compiler-based approaches, e.g., [37][20][6], usually link the detectors to the applications in the compiling stage and detect the bugs in the runtime process of GPU programs. They are limited by not being “fully automatic” because developers have to manually

5 THREATS TO VALIDITY

The threats to external validity mainly lie in the subjects and faults used in our benchmark. Though the projects of our benchmark suite may not represent the overall project distributions, they shall be selected in order to possibly maximize the overall features of the CUDA projects. In this way, our benchmark is derived based on real-world programs from GitHub, i.e., we select seven popular CUDA-related projects with a total of 17928 commits and 1.36 million LOC.

The threats to internal validity mainly lie in the potential bugs in our implementation due to the complicated mechanism of Simulee.

To reduce the threats, three graduate students, closely mentored by three SE/Systems supervisors, have been carefully working for over one year. We manually reviewed all our implementation code and also included corresponding tests for verifying our implementation. In addition, the effectiveness of “Automatic Input Generation” can impact on the performance synchronization bug detection. It is possible that the “Automatic Input Generation” component may miss certain inputs that can trigger synchronization bugs such that Simulee would miss detecting the bug or report a false positive. To reduce this threat, we set a large number of suitable parameters for the evolutionary algorithm adopted in “Automatic Input Generation” to reduce the probability of missing error-inducing inputs.

To threats to construct validity mainly lie in the metrics used in this work. To reduce the threats, we measure the number of both previously known and the identified bugs detected by the studied techniques as well as their false-positive rate and corresponding time cost.

6 RELATED WORK

As our work investigates the automatic bug detection techniques for CUDA programs, the related work includes the following two parts: empirical studies on CUDA programs and techniques of CUDA bug detection. Moreover, since Simulee essentially is a search-based bug-detection technique, we also discuss such relevant work.

Empirical studies for CUDA programs

There are several existing work that study bugs and other features on CUDA programs. For instance, Yang et al. [43] delivered the empirical study on the features of the performance bugs on CUDA programs, Burtscher et al. [10] studied the control-flow irregularity and memory-access irregularity and found that both irregularities are mutually dependent and exist in most of kernels. Che et al. [12] examined the effectiveness of CUDA to express with different sets of performance characteristics. Some researchers are keen on the comparisons between CUDA and OpenCL. For instance, Demidov et al. [17] compared some C++ programs running on top of CUDA and OpenCL and found that they work equally well for problems of large size. Du et al. [19], on the other side, studied the discrepancies in the OpenCL and CUDA compilers’ optimization that affect the associated GPU computing performance. In our previous work, we also conducted empirical studies to explore CUDA program features. For instance, we investigated the features and the distribution of multiple CUDA program bug types based on a collected GitHub dataset [41]. Moreover, we developed an approach that can automatically repair CUDA synchronization bugs via program transformation and validated its performance via an experimental study based on real-world benchmarks [40].

CUDA bug detection

Unlike traditional program bugs which can be deterministically tested [39, 46] and debugged [23, 32, 47], CUDA synchronization bugs, especially data race and barrier divergence can often result in undefined behaviors. To detect such bugs, there are typically two types of approaches—compiler-based approaches and static analysis (SMT solver)-based approaches. In particular, compiler-based approaches, e.g., [37][20][6], usually link the detectors to the applications in the compiling stage and detect the bugs in the runtime process of GPU programs. They are limited by not being “fully automatic” because developers have to manually

5 THREATS TO VALIDITY

The threats to external validity mainly lie in the subjects and faults used in our benchmark. Though the projects of our benchmark suite may not represent the overall project distributions, they shall be selected in order to possibly maximize the overall features of the CUDA projects. In this way, our benchmark is derived based on real-world programs from GitHub, i.e., we select seven popular CUDA-related projects with a total of 17928 commits and 1.36 million LOC.

The threats to internal validity mainly lie in the potential bugs in our implementation due to the complicated mechanism of Simulee.
provide test cases. Moreover, given inferior inputs, the synchronization bugs might not be triggered and detected because such bugs could only occur under limited conditions. They can also be expensive since such runtime detection demands compiling process and GPU computing environment. To the best of our knowledge, these approaches fail to detect barrier divergence and redundant barrier function because of their detection mechanisms. On the other hand, various automatic synchronization bug detection approaches, e.g., [30][9], depend on static analysis and SMT solver [16] which could lead to poor runtime performance when handling complicated GPU programs. Besides, such approaches tend to report false positives or false negatives because it lacks runtime information. Although developers do not have to provide whole test inputs for them, they still need to provide heuristic settings in order to avoid path explosions, e.g., dimension settings for GPUTrack, main functions and initial environments of kernel functions for GKLEE.

Compared with these approaches, Simulee can automate the detection process to achieve superior detection performance by enabling the “Automatic Input Generation” component and the “Memory-based Synchronization Bug Detection” component.

**Search-based Software Engineering.** The optimization approaches such as Evolutionary Programming are widely used to solve software engineering problem by modeling them into optimization problems [26]. Yu et al. [45] proposed a metric, PNet constraint to detect CPU-based synchronization bugs. Harman et al. [27] applied evolving pareto front approximation to refactor software systems. Hierons et al. [28] used Many-Objective Evolutionary Optimisation to optimize the process of software product selection. McMinn et al. [33] evolved coverage criteria to improve the performance of bug detection. Oumi et al. [36] modeled the process of refactoring into a multiple-objective search-based problem for generating refator patches.

7 CONCLUSIONS

In this paper we develop a fully automated approach, namely Simulee, that can successfully detect CUDA synchronization bugs efficiently based on accurate memory-access modeling. More specifically, Simulee consists of two different components: the “Automatic Input Generation” component that applies Evolutionary Computation for automatically generating bug-inducing test inputs, and the “Bug Detection via Memory-Access Model” component that builds an accurate memory model for deriving the underlying CUDA synchronization bugs. To evaluate the efficacy of Simulee, we construct a benchmark from real-world CUDA-related projects. Our evaluation results suggest that Simulee can detect most of the manually identified synchronization bugs out of the studied projects, and successfully detect 24 previously unknown bugs which have never been reported/detected before. In addition, Simulee can achieve better effectiveness and efficiency than multiple state-of-the-art approaches.

8 ACKNOWLEDGEMENT

This work is partially supported by the National Natural Science Foundation of China (Grant No. 61902169), Shenzhen Peacock Plan (Grant No. KQTD2016112514355531), and Science and Technology Innovation Committee Foundation of Shenzhen (Grant No. JCYJ20170817110848086). This work is also partially supported by National Science Foundation under Grant Nos. CCF-1763906 and CCF-1942340, as well as Ant Financial Services Group.

REFERENCES

generating_computing_on_graphics_processing_units.


