ESBMC-GPU
A Context-Bounded Model Checking Tool
to Verify CUDA Programs

Felipe R. Monteiro\textsuperscript{1}, Erickson H. da S. Alves\textsuperscript{1}, Isabela S. Silva\textsuperscript{1}, Hussama I. Ismail\textsuperscript{1}, Lucas C. Cordeiro\textsuperscript{1,2}, and Eddie B. de Lima Filho\textsuperscript{1,3}

\textsuperscript{1}Faculty of Technology, Federal University of Amazonas, Brazil
\textsuperscript{2}Department of Computer Science, University of Oxford, United Kingdom
\textsuperscript{3}TPV Technology Limited, Brazil

Abstract
The Compute Unified Device Architecture (CUDA) is a programming model used for exploring the advantages of graphics processing unit (GPU) devices, through parallelization and specialized functions and features. Nonetheless, as in other development platforms, errors may occur, due to traditional software creation processes, which may even compromise the execution of an entire system. In order to address such a problem, ESBMC-GPU was developed, as an extension to the Efficient SMT-Based Context-Bounded Model Checker (ESBMC). In summary, ESBMC processes input code through ESBMC-GPU and an abstract representation of the standard CUDA libraries, with the goal of checking a set of desired properties. Experimental results showed that ESBMC-GPU was able to correctly verify 85\% of the chosen benchmarks and it also overcame other existing GPU verifiers regarding the verification of data-race conditions, array out-of-bounds violations, assertive statements, pointer safety, and the use of specific CUDA features.

Keywords: GPU verification, formal verification, model checking, CUDA

1. Introduction

The Compute Unified Device Architecture (CUDA) is a development framework that makes use of the architecture and processing power of graphics processing units (GPUs) [1]. Indeed, CUDA is also an application programming interface (API), through which a GPU’s parallelization scheme and tools can be accessed, with the goal of executing kernels [1]. Nonetheless, source code is still written by human programmers, which may result in
arithmetic overflow, division by zero, and other violation types. In addition, given that CUDA allows parallelization, problems related to the latter can also occur, due to thread scheduling [2].

In order to address the mentioned issues, an extension to the Efficient SMT-Based Context-Bounded Model Checker (ESBMC) [3] was developed, named as ESBMC-GPU [4, 5, 6], with the goal of verifying CUDA-based programs (available online at http://esbmc.org/gpu/). ESBMC-GPU consists of an extension for parsing CUDA source code (i.e., a front-end to ESBMC) and a CUDA operational model (COM), which is an abstract representation of the standard CUDA libraries (i.e., the native API) that conservatively approximates their semantics.

A distinct feature of ESBMC-GPU, when compared with other approaches [2, 7, 8, 9], is the use of Bounded Model Checking (BMC) [10] allied to Satisfiability Modulo Theories (SMT) [11], with explicit state-space exploration [12, 3]. In summary, concurrency problems are tackled, up to a given loop/recursion unwinding and context bound, while each interleaving itself is symbolically handled; however, even with BMC, state-space exploration may become a very time-consuming task, which is alleviated through state hashing and Monotonic Partial Order Reduction (MPOR) [13]. As a consequence, redundant interleavings are eliminated, without ignoring a program’s behavior.

Finally, existing GPU verifiers often ignore some aspects related to memory leak, data transfer, and overflow, which are normally present in CUDA programs. The proposed approach, in turn, explicitly addresses them, through an accurate checking procedure, which even considers data exchange between main program and kernel. Obviously, it results in higher verification times, but more errors can then be identified and later corrected, in another development cycle.

Existing GPU Verifiers. In addition to ESBMC-GPU, there are other tools able to verify CUDA programs and each one of them uses its own approach and targets specific property violations. For instance, GPUVerify [2] is based on synchronous, delayed visibility semantics, which focuses on detecting data race and barrier divergence, while reducing kernel verification procedures for the analysis of sequential programs. GPU+KLEE (GKLEE) [8], in turn, is a concrete plus symbolic execution tool, which considers both kernels and main functions, while checking deadlocks, memory coalescing, data race, warp divergence, and compilation level issues. In addition, Concurrency Intermediate Verification Language (CIVL) [9], a framework for static analysis and concurrent program verification, uses abstract syntax tree and partial order reduction to detect user-specified assertions, deadlocks, memory leaks, invalid pointer dereference, array out-of-bounds, and division by zero.
In fact, ESBMC-GPU differs from the aforementioned approaches due to its combination of techniques to prune the state-space exploration (i.e., two-thread analysis, state hashing, and MPOR) with COM, which demonstrated effectiveness in the verification of data-race conditions, array out-of-bounds violations, assertive statements, pointer safety, and the use of specific CUDA features (cf. Section 5).

2. Architecture and Implementation

ESBMC-GPU is built on top of ESBMC, which is an open-source context-bounded model checker based on SMT solvers for ANSI-C/C++ programs [12, 3, 14], and adds four essential models, as described below.

![Figure 1: Overview of ESBMC-GPU's architecture.](image)

**CUDA Operational Model.** An operational model for CUDA libraries that provides support to CUDA functionalities, in conjunction with ESBMC, is shown in Fig. 1. Such an approach, which was previously attempted in the verification of C++ programs [14, 15, 16, 17], consists of an abstract representation that reliably approximates the CUDA library’s semantics; however, COM incorporates pre- and post-conditions into verification processes, which enables ESBMC-GPU to verify specific properties (cf. Section 3). Indeed, COM allows the necessary control for performing code analysis, where both CUDA operation and knowledge for model checking its properties are available. Importantly, COM encloses a representation for the Runtime, Math, and cuRAND APIs, which are important CUDA libraries widely used in real applications [1]. In particular, with respect to the number of methods/functions, COM covers 53%, 31%, and 17% of Runtime, Math, and cuRAND APIs, respectively.

ESBMC was designed to handle multi-threaded software, through the use of an API called POSIX – ISO/IEC 9945 [18]. In order to support the verification of CUDA kernels, COM applies code transformations to kernel calls using ESBMC’s intrinsic functions [6]. In particular, thread/block configurations in a CUDA kernel call are used as parameters in such intrinsic functions, which are responsible for checking for preconditions, configuring block
and threads dimension, and translating GPU threads to POSIX ones. Thus, COM is able to support thread interleaving (to create execution paths) and dynamic creation of threads, in order to check data race and specific C/C++ programming language failures (e.g., array out-of-bounds and pointer safety).

ESBMC models the Pthreads API [12] to support thread synchronization, i.e., mutex locking operations and conditional waiting, and dynamic creation of threads, which makes that representation very similar to the official CUDA scheduler [6], in such a way that our multi-threading model approximates to that of GPU kernels. Therefore, COM is able to use such aspects present in ESBMC, in order to handle variables in different types of memory and also in inter-warp communication.

The ESBMC’s memory model uses static pointer analysis, padding in structures, with the goal of making all fields align to word boundaries, memory access alignment rules enforcement, and byte array allocation, when the type of memory allocation is unclear [19]. Since ESBMC-GPU is built on top of ESBMC, its memory model for the different types of memory is complete.

Two-threads Analysis. Similarly to GPUVerify [2] and PUG [7], ESBMC-GPU also reduces the number of threads (to only two elements), during the verification of CUDA programs, by considering a NVIDIA Fermi GPU architecture [1], in order to improve verification time and avoid the state-space explosion problem. In CUDA programs, whilst threads execute the same parametrized kernel, only two of them are necessary for conflict check. Thus, such an analysis ensures that errors (e.g., data races) detected between two threads, in a given subgroup and due to unsynchronized accesses to shared variables, are enough to justify a property violation [6].

State Hashing. ESBMC-GPU applies state hashing to further eliminate redundant interleavings and also reduce the state space, based on SHA256 hashes [20]. In particular, its symbolic state hashing approach computes a summary for a particular state that has already been explored and then indexes the resulting set, in order to reduce the generation of redundant states. Given any state computed during the symbolic execution of a specific CUDA kernel, ESBMC-GPU simply summarizes it and efficiently determines whether it has been explored before or not, along a different computation path. When this behavior is confirmed, which happens during the ESBMC-GPU’s symbolic-execution procedure, then the current computation path does not need to be further explored in the associated reachability tree (RT). This way, if ESBMC-GPU reaches such a state, i.e., where a context switch can be taken (e.g., before a global variable or synchronization primitive) and all shared/local variables and program counters are similar to another explored node, then ESBMC-GPU just considers that an identical node to be
further explored, since reachability subtrees associated to them are also similar [6, 21].

**Monotonic Partial Order Reduction.** MPOR is used to reduce the number of thread interleavings, by classifying transitions inside a program as dependent or independent. As a consequence, it is possible to determine whether interleaving pairs always lead to the same state and then remove duplicates in a RT, without ignoring any program’s behavior [21].

### 3. Functionalities

COM models CUDA libraries and provides a multi-threading model much similar to the CUDA scheduler (cf. Section 2). Moreover, it is able to simulate CUDA’s program structure and memory, being susceptible of handling CUDA programs. Thus, through the integration of COM into ESBMC (i.e., ESBMC-GPU), one is able to analyze CUDA programs and verify the following properties:

- **Data race.** ESBMC-GPU checks data race conditions, in order to detect if multiple threads perform unsynchronized access to the same memory locations;

- **Pointer safety.** ESBMC-GPU also ensures that (i) a pointer offset does not exceed object bounds and (ii) a pointer is neither NULL nor invalid;

- **Array bounds.** ESBMC-GPU performs array-bound checking, in order to ensure that any variable, used as an array index, is within known bounds;

- **Arithmetic under- and overflow.** ESBMC-GPU checks whether a sum or product exceeds the memory limits that a variable can handle, which can cause an error capable of spreading through the entire execution path;

- **Division by zero.** ESBMC-GPU analyzes whether denominators, in arithmetic expressions, lead to divisions by zero;

- **User-specified assertions.** ESBMC-GPU considers all assertions specified by users, which is essential to a thorough verification process, as some specific possible violations must be explicitly pointed out.

In order to check the aforementioned properties, ESBMC-GPU explicitly explores the possible interleavings (up to the given context bound) and calls
the single-threaded BMC procedure on each one, whenever it reaches a RT
leaf node. Then, the mentioned procedure will stop if it finds a bug or when
all possible RT interleavings have been systematically explored [6]. Further-
more, ESBMC-GPU has the following additional command-line options:

    --no-assertions: to ignore assertions;
    --no-bounds-check: to skip array bounds check;
    --no-div-by-zero-check: to skip division by zero check;
    --no-pointer-check: to skip pointer check;
    --memory-leak-check: to enable memory leak check;
    --overflow-check: to enable arithmetic under- and overflow check;
    --deadlock-check: to enable global/local deadlock check with mutex;
    --data-races-check: to enable data races check;
    --lock-order-check: to enable for lock acquisition ordering check;
    --atomicity-check: to enable atomicity check at visible assignments;
    --force-malloc-success: to consider that there is always enough
    memory available in the device;

Thus, ESBMC-GPU is able to check CUDA programs for: deadlock, asser-
tion, lock acquisition error, division by zero, pointer safety, arithmetic over-
flow, and/or out-of-bounds array violation. The precision and performance
of ESBMC-GPU will be further discussed in Section 5.

4. Illustrative Example

In this part, ESBMC-GPU usage is demonstrated, by using the CUDA
program shown in Fig. 2. First of all, users must replace the default ker-
nel call (line 16) by an intrinsic function of ESBMC-GPU (line 17). Then,
the resulting CUDA program can be passed to the command-line version of
ESBMC-GPU, as follows:

    esbmc-gpu <file>.cu --unwind <k> --context-switch <c>
    --state-hashing -I <path-to-CUDA-OM>,

6
```c
#include "...>
#define BLOCKS 1
#define THREADS 2

__global__ void kernel(int *A) {
    A[threadIdx.x + 1] = threadIdx.x;
}

int main() {
    int *a;
    int *dev_a;
    int size = THREADS * sizeof(int);
    a = (int*) malloc(size);
    cudaMemcpy(dev_a, a, size, cudaMemcpyHostToDevice);
    // kernel<<<BLOCKS, THREADS>>>(dev_a);
    cudaMemcpy(dev_a, a, size, cudaMemcpyDeviceToHost);
    free(a);
    return 0;
}
```

Figure 2: Illustrative CUDA code example.

where <file>.cu is the CUDA program, <k> is the maximum loop unrolling, <c> is a context-switch bound, --state-hashing reduces redundant interleavings, and <path-to-CUDA-OM> is the location of the COM library.

In the mentioned example, ESBMC-GPU detects an array out-of-bounds violation. Indeed, this CUDA-based program retrieves a memory region that has not been previously allocated, i.e., when threadIdx.x = 1, the program tries to access a[2]. Importantly, the cudaMemcpy() function’s operational model has a precondition that checks if the memory size to be allocated is greater than zero. In addition, an assertion checks if the result matches to the expected postcondition (line 22). The verification of this program through ESBMC-GPU produces 54 successful and 3 failed interleavings. For instance, one possible failed interleaving is represented by the threads executions t₀ : a[1] = 0; t₁ : a[2] = 1, where a[2] = 1 represents an incorrect access to the array index a. It is worth noticing that CIVL, ESBMC-GPU, and GKLEE are also able to detect this array out-of-bounds violation, but GPUVerify fails, as it reports a true incorrect result (i.e., a missed bug).

5. Experimental Evaluation

In order to evaluate ESBMC-GPU’s precision and performance, a benchmark suite was created, which comprises 20 CUDA kernels from NVIDIA GPU Computing SDK v2.0 [22], 20 CUDA kernels from Microsoft C++ AMP Sample Projects [23], and 114 CUDA-based programs that explore a wide range of CUDA functionalities. In summary, the chosen suite contains 47.4% bug-free and 52.6% buggy benchmarks, which were organized into 5
sets, in order to simplify our discussion, according to the properties it tackles: array bounds (5), assertive statements (7), data-race conditions (17), pointer safety (7), and other specific CUDA functionalities (118). The latter includes 
\_device\_ function calls, general CUDA functions (e.g., cudaMemcpy), general libraries in CUDA (e.g., curand.h), type modifiers (e.g., unsigned), type definitions, and intrinsic CUDA variables (e.g., uint4).

The present experiments answer two research questions: (i) How accurate is ESBMC-GPU when verifying the chosen benchmarks? (ii) How does ESBMC-GPU’s performance compare to other existing verifiers? In order to answer both research questions, all benchmarks were verified with 4 GPU verifiers (ESBMC-GPU v2.0, GKLEE v2012, GPUVerify v1811, and CIVL v1.7.1), on an otherwise idle Intel Core i7-4790 CPU 3.60 GHz, with 16 GB of RAM, running Ubuntu 14.04 OS. Importantly, all presented execution times are actually CPU times, i.e., only the elapsed time periods spent in the allocated CPUs, which was measured with the times system call (POSIX system). An overview of the experimental results is shown in Fig. 3, where True represents bug-free benchmarks, False represents buggy benchmarks, Not supported represents benchmarks that could not be verified, Correct represents the percentage of benchmarks correctly verified, and Incorrect represents the percentage of benchmarks incorrectly verified (i.e., a verification tool reports an unexpected result).

As one may notice, the present experimental results show that ESBMC-GPU reached a successful verification rate of approximately 85%, while GKLEE, GPUVerify, and CIVL reported 72%, 50%, and 35%, respectively. More precisely, ESBMC-GPU correctly detected all data-race conditions present in the benchmarks, which is due to the COM’s multi-threading model that under-approximates the GPU kernels. It also outperformed GKLEE, GPU-
Verify, and CIVL, in the verification of array out-of-bounds violations (100%),
assertive statements (86%), and pointer safety (72%), which is related to
ESBMC’s capacity to handle arrays and pointers [3]. Furthermore, ESBMC-
GPU presented the highest coverage rate for specific CUDA functionalities
(82%) that is once again due to COM, which incorporates specific pre- and
post-conditions into its verification processes.

Limitations. ESBMC-GPU was unable to correctly verify 24 benchmarks,
which are related to constant memory access (2%), CUDA’s specific libraries
(e.g., curand.h) (4.5%), and the use of pointers to functions, structures, and
char type variables, when passed as kernel call arguments (4.5%). In addition,
it only reported 3% of incorrect true, which are due to NULL pointer
accesses, and 1% of incorrect false results, due to partial coverage of the cu-
daMalloc function for copies over float variables. The remaining verifiers
(i.e., GKLEE, GPUVerify, and CIVL) were unable to detect mostly data-
race conditions, assertive statements, and array out-of-bounds violations.
In addition, they lack support of CUDA specific features, e.g., GPUVerify
does not support the use of the memset function nor function point-
ers and CIVL does not support several CUDA features, such as atomic
functions, cudaThreadSynchronize, threadIdx, curand functions, dim3,
math_functions, uint4, __constant__ variables, among others.

Performance. MPOR resulted in a performance improvement of approx-
imately 80%, by decreasing the verification time from 16 to 3 hours, while
the two-threads analysis further reduced that to 789.6 sec. Although such
techniques have considerably improved the ESBMC-GPU’s performance, it
still takes longer than the other evaluated tools: GPUVerify (98.36 sec), GK-
LEE (105.18 sec), and CIVL (708.52 sec). On the one hand, this is due to
thread interleavings, which combine symbolic model checking with explicit
state-space exploration [6]. On the other hand, ESBMC-GPU still presents
the highest accuracy, with less than 6 seconds per benchmark.

Availability of Data and Tools. The performed experiments are based on
a set of publicly available benchmarks. All benchmarks, tools, and results,
associated with the current evaluation, are available at www.esbmc.org/gpu/.

6. Conclusions and Future Work

ESBMC-GPU marks the first application of an SMT-based context-BMC
tool that recognizes CUDA directives [6]. Besides, it also applies MPOR,
two-thread analysis, and state hashing, in order to further simplify verifi-
cation models and provides fewer incorrect results, compared with GKLEE,
GPUVerify, and CIVL. Indeed, it presents improved ability to detect array out-of-bounds and data race violations.

Future work aims to extend ESBMC-GPU, in order to fully support the verification of CUDA (parallel) streams and events [1]. In addition, more models of libraries will be integrated into COM, with the goal of increasing the coverage of CUDA’s API such as CUDA Driver API, NPP, and cu-SOLVER. Finally, we also aim to implement further techniques (e.g., invariant inference via abstract interpretation [24]), in order to prune the state-space exploration, by taking into account GPU symmetry.

Acknowledgements

This paper is based on research sponsored by the Institute of Development and Technology (INdT) and by the National Council for Scientific and Technological Development (CNPq) under agreement number 475647/2013 – 0.

References


Required Metadata

Current executable software version

Ancillary data table required for sub version of the executable software.

<table>
<thead>
<tr>
<th>Nr.</th>
<th>(executable) Software metadata description</th>
<th>Please fill in this column</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>Current software version</td>
<td>2.0</td>
</tr>
<tr>
<td>S2</td>
<td>Permanent link to executables of this version</td>
<td><a href="http://esbmc.org/gpu/">http://esbmc.org/gpu/</a></td>
</tr>
<tr>
<td>S3</td>
<td>Legal Software License</td>
<td>Apache v2.0</td>
</tr>
<tr>
<td>S4</td>
<td>Computing Operating System</td>
<td>Ubuntu Linux OS</td>
</tr>
<tr>
<td>S5</td>
<td>Installation requirements &amp; dependencies</td>
<td>GNU Libtool; Automake; Flex &amp; Bison; Boost C++ Libraries; Multi-precision arithmetic library developers tools (libgmp3-dev package); SSL development libraries (libssl-dev package); CLang 3.8; LLDB 3.8; GNU C++ compiler (multilib files); libc6 and libc6-dev packages</td>
</tr>
<tr>
<td>S6</td>
<td>Link to user manual</td>
<td><a href="http://esbmc.org/gpu/">http://esbmc.org/gpu/</a></td>
</tr>
<tr>
<td>S7</td>
<td>Support email for questions</td>
<td><a href="mailto:lucas.cordeiro@cs.ox.ac.uk">lucas.cordeiro@cs.ox.ac.uk</a></td>
</tr>
</tbody>
</table>

Table 1: Software metadata (optional)
**Current code version**

Ancillary data table required for subversion of the codebase.

<table>
<thead>
<tr>
<th>Nr.</th>
<th>Code metadata description</th>
<th>Please fill in this column</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Current code version</td>
<td>v2.0</td>
</tr>
<tr>
<td>C2</td>
<td>Permanent link to code/repository used for this code version</td>
<td><a href="https://github.com/ssvlab/esbmc-gpu">https://github.com/ssvlab/esbmc-gpu</a></td>
</tr>
<tr>
<td>C3</td>
<td>Legal Code License</td>
<td>GNU Public License</td>
</tr>
<tr>
<td>C4</td>
<td>Code versioning system used</td>
<td>git</td>
</tr>
<tr>
<td>C5</td>
<td>Software code languages, tools, and services used</td>
<td>C++</td>
</tr>
<tr>
<td>C6</td>
<td>Compilation requirements, operating environments &amp; dependencies</td>
<td>GNU Libtool; Automake; Flex &amp; Bison; Boost C++ Libraries; Multi-precision arithmetic library developers tools (libgmp3-dev package); SSL development libraries (libssl-dev package); CLang 3.8; LLDB 3.8; GNU C++ compiler (multilib files); libc6 and libc6-dev packages</td>
</tr>
<tr>
<td>C7</td>
<td>Link to developer documentation</td>
<td><a href="http://esbmc.org/gpu">http://esbmc.org/gpu</a></td>
</tr>
<tr>
<td>C8</td>
<td>Support email for questions</td>
<td><a href="mailto:lucas.cordeiro@cs.ox.ac.uk">lucas.cordeiro@cs.ox.ac.uk</a></td>
</tr>
</tbody>
</table>

Table 2: Code metadata (mandatory)