CANAL: A Cache Timing Analysis Framework via LLVM Transformation

Chungha Sung  
Los Angeles, CA, USA

Brandon Paulsen  
Los Angeles, CA, USA

Chao Wang  
Los Angeles, CA, USA

ABSTRACT
A unified modeling framework for non-functional properties of a program is essential for research in software analysis and verification, since it reduces burdens on individual researchers to implement new approaches and compare existing approaches. We present CANAL, a framework that models the cache behaviors of a program by transforming its intermediate representation in the LLVM compiler. CANAL inserts auxiliary variables and instructions to allow standard verification tools to handle a new class of cache related properties, e.g., for computing the worst-case execution time and detecting side-channel leaks. We demonstrate the effectiveness of CANAL using three verification tools: KLEE, SMACK and Crab-llvm. We confirm the accuracy of our cache model by comparing with CPU cycle-accurate simulation results of GEMS.

CCS CONCEPTS
• Software and its engineering → Software verification and validation; • Security and privacy → Cryptanalysis and other attacks;

KEYWORDS
cache, execution time, side channel, verification, symbolic execution, abstract interpretation, bounded model checking

ACM Reference Format:

1 INTRODUCTION
Analyzing the cache behaviors of a program is important, e.g., for computing the worst-case execution time of a real-time system [5, 15] and detecting information leaks through side channels [8, 15]. However, existing verification tools are often designed only for checking functional properties, e.g., assertions or pre- and post-conditions. For example, none of the participants of recent software verification competitions [2] can verify non-functional properties such as those related to the execution time. Although specialized tools have been developed to handle such non-functional properties, they are rarely open-source or as well-maintained as mainstream verification tools. As a result, it is difficult for individual researchers to implement new approaches for verifying such properties or evaluate existing approaches.

We fill the gap by developing a lightweight cache modeling framework for standard verification tools, by transforming the LLVM intermediate representation (IR) of a program to add self-modeling capabilities. That is, we insert auxiliary variables and LLVM instructions over these variables to record and update cache statistics related to Load/Store instructions during the program execution. By using the instrumented LLVM bitcode as input, standard (functional) verification tools will have the capability of verifying a new class of (non-functional) properties.

Our modeling framework, named CANAL, takes C/C++ code as input and emits LLVM bitcode as output. Thus, it can be used by any LLVM-based verification tools. For example, symbolic execution tools such as KLEE [4] may take the program instrumented by CANAL to detect side-channel leaks; bounded model checkers such as SMACK [14] may take the program instrumented by CANAL to conduct Must- and May-HIT cache analyses; and static analyzers based on numerical abstract interpretation, such as Crab-llvm [10], may take the program instrumented by CANAL to conduct worst-case execution time (WCET) analysis. CANAL is available on GitHub.

In the remainder of this paper, we shall explain how to combine CANAL with KLEE, SMACK and Crab-llvm to obtain the desired results. We also compare CANAL with the CPU cycle-accurate simulation results of GEMS [9], a standard micro-architectural simulator, to demonstrate the accuracy of our cache model.

2 TOOL OVERVIEW
Figure 1 shows the overall flow of CANAL, which takes the C/C++ code of a program and the cache configuration file of a target computer as input, and returns the instrumented LLVM bitcode as output. After compiling the C/C++ code into LLVM bitcode, it uses a sequence of optimization (opt) passes to insert, before or after each Load/Store instruction, some new instructions that model the change of cache states due to these memory accesses. The inserted instructions can be understood as invocations of two functions: _CSIM_Load(addrInfo) and _CSIM_Store(addrInfo), which updates our model of the cache state whenever Load or Store is executed; addrInfo denotes information of the memory location.

Figure 1: The overall flow of CANAL.
In addition to the automatically inserted calls to __CSIM_Load and __CSIM_Store, the user of CANAL may specify properties using these auxiliary variables: __CSIM_num_hit, __CSIM_num_miss, __CSIM_Load_ret, and __CSIM_Store_ret. They represent the accumulative numbers of hits and misses along a program path, as well as the cache status (hit or miss) associated with each memory access. By feeding the instrumented LLVM bitcode to standard verification tools as input, CANAL allows them to verify a new class of non-functional properties, such as assertions over auxiliary variables that model the cache behaviors of the program.

Accuracy. To demonstrate the accuracy of our cache model, we compare our results with the cache statistics reported by GEM5. Toward this end, note that CANAL may be used as a standalone cache simulator: if we compile the LLVM bitcode instrumented by CANAL to an executable and run it with a concrete input, it will produce the cache statistics associated with that particular execution. Figure 2 shows this usage case, where __CSIM_init_cache() and __CSIM_print_stat() are inserted to the original C program to initialize the cache states and display the result, respectively. The function body of aes_encrypt() will be instrumented by CANAL automatically.

Since the cache statistics reported by GEM5 include not only the main() function but also operating system code executed before and after, we need to create two program versions and then compute their difference. One of the programs consists of the main() function and instructions inserted at the beginning and end of the main() function to flush the cache, while the other program consists of only these cache-flushing instructions with an empty main() function body. By running these two programs and computing the difference, we have obtained the exact numbers of cache hits and misses reported by GEM5.

Table 1 shows the comparison of GEM5 and CANAL for five example programs, including three sorting routines and two cryptographic routines. The results are always identical. The sorting routines exhibit a diverse range of memory-accessing behavior and CANAL allows them to verify a new class of non-functional properties, such as assertions over auxiliary variables that model the cache behaviors of the program.

### Table 1: Accuracy comparison: CANAL versus GEM5

<table>
<thead>
<tr>
<th>Name</th>
<th>Loc</th>
<th>Mem-access</th>
<th>GEM5</th>
<th>CANAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ary Acc</td>
<td>70</td>
<td>522,575</td>
<td>2,881</td>
<td>2,881</td>
</tr>
<tr>
<td>Bit Sort</td>
<td>49</td>
<td>32,192</td>
<td>68</td>
<td>68</td>
</tr>
<tr>
<td>Int Sort</td>
<td>49</td>
<td>2,619,985</td>
<td>15,054</td>
<td>15,054</td>
</tr>
<tr>
<td>AES [1]</td>
<td>788</td>
<td>536</td>
<td>39</td>
<td>39</td>
</tr>
</tbody>
</table>

3 APPLICATION SCENARIOS

We now demonstrate how CANAL may be used by KLEE, SMACK, and Crab-llvm using five example programs taken from the SV-COMP benchmark [2]: copysome, sanfoundry, and standard from the array-programs section, and gcd and sum from the bit-vectors section. We use a 4-way associative cache with LRU and 64-byte cache line while setting the cache size to 1 KB, 16 KB and 32 KB, respectively. We set the timeout to one hour for each program.

3.1 Combined with Symbolic Execution Tools

Symbolic execution is a technique for systematically exploring feasible paths of a program and generating their test inputs. Although it has been used primarily for checking functional properties, with CANAL, it can now be used to detect timing side-channel leaks.

**Timing Side-channel Leaks.** We say that a program P(k) with sensitive input k has timing side-channel leaks if the execution time of P depends on the value of k. That is, 3k1, k2 ∈ dom (k) : τ (P, k1) ≠ τ (P, k2), where k1 and k2 are values in the domain of k and τ is the execution time. Even if the program executes the same number (and type) of instructions, the execution time may still differ if there are different numbers of cache hits/misses. Such side-channel leaks may be detected by CANAL + KLEE.

Figure 3 shows an example, where input1 and input2 are marked as symbolic values and used to run the program prog() twice. After each execution, the numbers of hits and misses are stored in h1, m1, h2, and m2, respectively. Finally, the assertion checks if prog() is leak-free; that is, ∀ input1, input2, the condition (h1=h2 & m1=m2) always holds. KLEE can be used to search for concrete values of input1 and input2 that violate the assertion.

Table 2 shows the results of running KLEE on these programs. In each case, we manually modified the program to mark one or more parameters as the sensitive input. Columns 3-5 show if a leak is detected, together with the total number of tests generated and, among them, the number of tests that manifest the leak. Columns 6-8 show the time taken by KLEE for the different cache sizes.

### Table 2: Results of the side-channel leak detection.

<table>
<thead>
<tr>
<th>Name</th>
<th>Loc</th>
<th>Mem-access</th>
<th>Detection</th>
<th>Leak</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>copysome</td>
<td>79</td>
<td>122,575</td>
<td>No leak</td>
<td>1</td>
<td>1.38</td>
</tr>
<tr>
<td>sanfoundry</td>
<td>95</td>
<td>No leak</td>
<td>1</td>
<td>2.72</td>
<td>16.22</td>
</tr>
<tr>
<td>standard</td>
<td>61</td>
<td>Null</td>
<td>1</td>
<td>3.22</td>
<td>20.68</td>
</tr>
<tr>
<td>gcd</td>
<td>52</td>
<td>Null</td>
<td>1</td>
<td>4.11</td>
<td>40.34</td>
</tr>
<tr>
<td>sum</td>
<td>56</td>
<td>Null</td>
<td>1</td>
<td>6.61</td>
<td>87.87</td>
</tr>
</tbody>
</table>

3.2 Combined with Software Verification Tools

While symbolic execution is geared toward generating tests, verification tools such as SMACK are geared toward generating proofs, e.g., proving that an assertion holds under all test inputs. We show how SMACK can leverage CANAL to prove cache related properties.

**Must-hits and Must-misses.** One type of properties of interest is assertions over auxiliary variables such as __CSIM_Load() and __CSIM_Store(). For example, if a certain Load or Store instruction in the program always leads to a cache hit or miss, regardless of the program path and test input; in such a case, we call it a Must-hit or a Must-miss. CANAL instrument the LLVM bit-code in such a way that calls to __CSIM_Load() and __CSIM_Store() set the values of auxiliary variables __CSIM_Load_ret and __CSIM_Store_ret to reflect the cache status: true means the memory access leads to a hit, whereas false means it leads to a miss.
Figure 4 shows a program where we check if read of buffer[2] is a Must-hit. Thus, we save the value of __CSIM_Load_ret immediately after the read of buffer[2] to the variable named h and add an assertion stating h should always be true. If SMACK can prove the assertion, we know the read of buffer[2] is a Must-hit. Alternatively, we can add assert(h==false) and use SMACK to prove it is a Must-miss.

Since program verification is undecided in general (e.g., equivalent to the Turing-halting problem), SMACK may fail to prove either assertion; in such a case, the result remains inconclusive. In this particular example, however, SMACK is able to find a violation of the Must-hit assertion and generate a counterexample. The counterexample shows a scenario where buffer[0] and buffer[16] resides in two different 64-byte cache lines.

Table 3 shows the results of applying CANAL+SMACK to assertions we manually inserted to check if a Load or Store instruction in the program is a Must-hit/miss. With loop-unrolling bound of SMACK set to 10, and the cache size set to 1 KB, SMACK successfully verified all assertions. However, when the cache size was increased to 16 KB and 32 KB, SMACK started to time-out on some programs. This points out a scalability limitation of SMACK, together with direction for future work: improving the verification algorithms to make SMACK (and similar tools) more scalable for non-functional properties.

### Table 3: Results of the Must-hit analysis

<table>
<thead>
<tr>
<th>Name</th>
<th>LoC</th>
<th>Loop-unroll-bound</th>
<th>Property</th>
<th>Results</th>
<th>1KB</th>
<th>16KB</th>
<th>32KB</th>
</tr>
</thead>
<tbody>
<tr>
<td>canonical</td>
<td>69</td>
<td>10</td>
<td>Must-miss</td>
<td>Verified</td>
<td>225.67</td>
<td>116.71</td>
<td>TO</td>
</tr>
<tr>
<td>standard</td>
<td>75</td>
<td>10</td>
<td>Must-miss</td>
<td>Verified</td>
<td>79.08</td>
<td>1055.55</td>
<td>TO</td>
</tr>
<tr>
<td>sum</td>
<td>42</td>
<td>10</td>
<td>Must-hit</td>
<td>Verified</td>
<td>24.68</td>
<td>139.49</td>
<td>344.71</td>
</tr>
</tbody>
</table>

Figure 5 shows an if-else statement controlled by the value of cond. Assume each cache line contains 64 bytes, the first 16 integers of the array fall into one cache line, whereas the next 16 integers starting with buffer[16] fall into another cache line. However, during static analysis, there is no way of knowing what the value of cond is; therefore, one has to assume both branches may be taken.

When the Then-branch is taken, buffer[5] will be loaded to the cache, which means the access to buffer[5] is a cache hit. However, when the Else-branch is taken, buffer[5] will not be loaded to the cache, which means the access to buffer[5] is a cache miss. By using numerical abstract interpretation, Crab-llvm can take both cases into consideration and compute value ranges of n_s, n_s_h, and n_s_m. For this example, in particular, the value ranges would be [2,2] for n_s, [0,1] for n_s_h, and [1,2] for n_s_m. Therefore, Crab-llvm can prove the second and the third assertions, while reporting a potential violation of the first assertion.

In addition, an interesting application of the value ranges computed by numerical abstract interpretation is to compute the worst-case execution time (WCET), which depends on the maximum number of cache misses along all program paths.

### Table 4: Results of the numerical abstract interpretation

<table>
<thead>
<tr>
<th>Name</th>
<th>LoC</th>
<th>S.Hit</th>
<th>S.Miss</th>
<th>L.Hit</th>
<th>L.Miss</th>
<th>1KB</th>
<th>16KB</th>
<th>32KB</th>
</tr>
</thead>
<tbody>
<tr>
<td>canonical</td>
<td>69</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>225.67</td>
<td>116.71</td>
<td>TO</td>
</tr>
<tr>
<td>standard</td>
<td>75</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>79.08</td>
<td>1055.55</td>
<td>TO</td>
</tr>
<tr>
<td>sum</td>
<td>42</td>
<td>3</td>
<td></td>
<td>3</td>
<td></td>
<td>24.68</td>
<td>139.49</td>
<td>344.71</td>
</tr>
</tbody>
</table>

Figure 4: CANAL for must-hit analysis.

### 3.3 Combined with Static Analysis Tools

Static analyzers based on numerical abstract interpretation [7], such as Crab-llvm, can generate program invariants. These invariants, computed for each program location, are summaries over all paths and input values. Therefore, they can be used to estimate the worst-case execution time of a program. More specifically, by leveraging CANAL, tools such as Crab-llvm can generate invariants in terms of auxiliary variables such as (5 ≤ __CSIM_num_Load_hit ≤ 18).

Figure 5 shows an if-else statement controlled by the value of cond. Assume each cache line contains 64 bytes, the first 16 integers of the array fall into one cache line, whereas the next 16 integers starting with buffer[16] fall into another cache line. However, during static analysis, there is no way of knowing what the value of cond is; therefore, one has to assume both branches may be taken.

When the Then-branch is taken, buffer[5] will be loaded to the cache, which means the access to buffer[5] is a cache hit. However, when the Else-branch is taken, buffer[5] will not be loaded to the cache, which means the access to buffer[5] is a cache miss. By using numerical abstract interpretation, Crab-llvm can take both cases into consideration and compute value ranges of n_s, n_s_h, and n_s_m. For this example, in particular, the value ranges would be [2,2] for n_s, [0,1] for n_s_h, and [1,2] for n_s_m. Therefore, Crab-llvm can prove the second and the third assertions, while reporting a potential violation of the first assertion.

In addition, an interesting application of the value ranges computed by numerical abstract interpretation is to compute the worst-case execution time (WCET), which depends on the maximum number of cache misses along all program paths.

### 4 CACHE MODELING

We now briefly explain how cache is modeled inside CANAL. It is a lightweight cache model in that the modeling instructions are carefully designed to reduce the overhead of the verification tools. For example, pointers are difficult to handle by verification tools; therefore, we avoid using them in the instrumented code.

#### 4.1 Pre-computing Address-to-Cache Mapping

Inside LLVM, we first obtain the memory address of each program variable by analyzing the symbol table of the pre-compiled code. Then, for the target computer architecture, we generate a memory layout. We try to pre-compute the possible address value for each load or store instruction in the program.

If the address is a fixed value, we compute its set and tag fields in the cache, and use these concrete values to simplify the instantiation of __CSIM_Load() and __CSIM_Store(). Otherwise, we resort to the use of if-else statements to dynamically compute the set and tag fields (more difficult to handle by verification tools).

Figure 6 shows a simple case where the address of var is statically known, and thus we can pre-compute its set (242) and tag (1). These concrete values are used to instantiate __CSIM_Store(242, 1).

Figure 6: Pre-computed 'set' and 'tag' values.

## Table of Contents

- [Introduction](#introduction)
- [Background](#background)
- [Related Work](#related-work)
- [Motivations](#motivations)
- [Design](#design)
- [Implementation](#implementation)
- [Evaluation](#evaluation)
- [Conclusions](#conclusions)

**CANAL: A Cache Timing Analysis Framework via LLVM Transformation**

ASE '18, September 3–7, 2018, Montpellier, France
4.2 Simplifying Updates of the Cache Statistics

To simplify the storage and update of cache statistics so verification tools can handle them easily, we use a set of simple variables as opposed to an array indexed by memory addresses. This can drastically reduce the complexity of the cache-modeling instructions inside functions \_\_CSIM\_Load() and \_\_CSIM\_Store().

Figure 8 shows the internals of \_\_CSIM\_Store(), which updates the cache statistics based on the values of set and tag. Instead of using monolithic arrays such as cacheline[set].tag, we use individual variables such as \_\_CSIM\_cacheline00\_\_tag.

The number 00 means the cache line is associated with set 0 and way 0, and the auxiliary variable denotes the tag saved at the line. When a cache miss occurs, for example, we update the value of cacheline00\_\_tag as well as the values of similar auxiliary variables, and evict a victim. In this implementation, LRU policy is used to compute the victim; but other replacement policies may be incorporated into CANAL easily.

Implementations of \_\_CSIM\_Store() and \_\_CSIM\_Load() are specific to each individual program under verification, and therefore they are generated by CANAL automatically.

5 RELATED WORK

CANAL is the first LLVM-based lightweight cache modeling framework designed specifically for software verification tools. Although there are other cache simulators [3, 12] and CPU simulators such as GEM5 [9], they are not designed for this purpose. In particular, they cannot be used in the same way as CANAL to afford existing verification tools the capability of verifying a new class of cache related non-functional properties.

There are also tools designed specifically for WCET analysis based on cache analysis [5, 6, 13] and for detecting cache timing side channels [8, 11, 15]. However, the modeling part of these tools are tied up with the subsequent analysis part, and therefore cannot be used by other verification tools. Furthermore, the analysis part of these tools is rarely open-source, and often not as well-maintained as the mainstream software verification tools, which are updated constantly to keep up with the competition [2].

Although our main contribution in this work is the lightweight cache modeling that facilitates the subsequent analysis and verification, there is still room for improvement in the analysis and verification algorithms. Since cache timing behaviors are non-functional properties, they often have significantly different characteristics from functional properties, and thus may benefit from specialized algorithms to make verification more efficient and scalable.

Our implementation of CANAL has been tested on programs from two domains: real-time software and embedded software. In both cases, the program structure and language constructs are relatively simple. To handle C/C++ programs in other application domains, more sophisticated static analyses may be needed, e.g., to deal with pointer aliasing and complex loops during the pre-computation of address-to-cache mapping and updates of the cache statistics, in order to keep the application of our LLVM based transformation efficient. We also plan to further refine our cache model, e.g., to handle multi-threading as well as multi-level cache.

6 CONCLUSIONS

We have presented CANAL, a framework for modeling cache behaviors of a program based on LLVM transformations, which allows standard software verification tools to check cache timing properties. We demonstrated the accuracy of our cache model in CANAL by comparing with the simulation results of GEMS, as well as the effectiveness of combining CANAL with three existing tools (KLEE, SMACK and Crab-llvm) in verifying cache related properties.

ACKNOWLEDGMENTS

This work was supported in part by NSF under grants CNS-1617203 and CNS-1702824 and ONR under grant N00014-17-1-2896.

REFERENCES