CryptOpt: Verified Compilation with Random Program Search for Cryptographic Primitives

JOEL KUEPPER, University of Adelaide, Australia
ANDRES ERBSEN, Massachusetts Institute of Technology, USA
JASON GROSS, Massachusetts Institute of Technology, USA
OWEN CONOLY, Massachusetts Institute of Technology, USA
CHUYUE SUN, Stanford, USA
SAMUEL TIAN, Massachusetts Institute of Technology, USA
DAVID WU, University of Adelaide, Australia
ADAM CHLIPALA, Massachusetts Institute of Technology, USA
CHITCHANOK CHUENGSATIANSUP, University of Adelaide, Australia
DANIEL GENKIN, Georgia Institute of Technology, USA
MARKUS WAGNER, University of Adelaide, Australia
YUVAL YAROM, University of Adelaide, Australia

Most software domains rely on compilers to translate high-level code to multiple different machine languages, with performance not too much worse than what developers would have the patience to write directly in assembly language. However, cryptography has been an exception, where many performance-critical routines have been written directly in assembly (sometimes through metaprogramming layers). Some past work has shown how to do formal verification of that assembly, and other work has shown how to generate C code automatically along with formal proof, but with consequent performance penalties vs. the best-known assembly. We present CryptOpt, the first compilation pipeline that specializes high-level cryptographic functional programs into assembly code significantly faster than what GCC or Clang produce, with mechanized proof (in Coq) whose final theorem statement mentions little beyond the input functional program and the operational semantics of x86-64 assembly. On the optimization side, we apply randomized search through the space of assembly programs, with repeated automatic benchmarking on target CPUs. On the formal-verification side, we connect to the Fiat Cryptography framework (which translates functional programs into C-like IR code) and extend it with a new formally verified program-equivalence checker, incorporating a modest subset of known features of SMT solvers and symbolic-execution engines. The overall prototype is quite practical, e.g. producing new fastest-known implementations for the relatively new Intel i9 12G, of finite-field arithmetic for both Curve25519 (part of the TLS standard) and the Bitcoin elliptic curve secp256k1.

1 INTRODUCTION

Being a foundational pillar of computer security, cryptographic software needs to achieve three often-competing aims. First, being security-critical, the software needs to be correct and protected from implementation attacks. Second, because it is used frequently, it needs to be efficient. Third, for migration to new architectures, the software needs to be portable. Implementations of cryptographic code, therefore, must strike a trade-off between these needs. Implementations that aim for portability tend to use high-level languages, such as Java or C. These allow for easy maintenance and are essentially platform-independent, assuming the existence of suitable development tools like compilers and assemblers.

Authors' addresses: Joel Kuepper, University of Adelaide, Australia; Andres Erbsen, Massachusetts Institute of Technology, USA; Jason Gross, Massachusetts Institute of Technology, USA; Owen Conoly, Massachusetts Institute of Technology, USA; Chuyue Sun, Stanford, USA; Samuel Tian, Massachusetts Institute of Technology, USA; David Wu, University of Adelaide, Australia; Adam Chlipala, Massachusetts Institute of Technology, USA; Chitchanok Chuengsatiansup, University of Adelaide, Australia; Daniel Genkin, Georgia Institute of Technology, USA; Markus Wagner, University of Adelaide, Australia; Yuval Yarom, University of Adelaide, Australia.

At the same time, compiler-based code generation can be a double-edged sword. First, compiler-produced cryptographic code tends to underperform when compared to hand-optimized code [Bernstein et al. 2013, 2014a,b, 2017; Chou 2015, 2016; Chuengsatiansup et al. 2013; Chuengsatiansup and Stehlé 2019; Kannwischer et al. 2019], typically written directly in the platform's assembly language. Beyond slower performance, compilers are typically not designed for maintaining security properties. In particular, compilation bugs could result in incorrect code [Erbsen et al. 2019, Appendix A], while overly aggressive optimizations may even strip side-channel protections [Barthe et al. 2018; D'Silva et al. 2015; Kaufmann et al. 2016].

We note that the difficulties compilers have when operating over cryptographic code are not caused by high code complexity. In fact, cryptographic code tends to be simpler than a typical program code, due to its avoidance of data-dependent control flows and memory-access patterns for reasons of side-channel resistance. Ironically, compiler optimization passes often focus on control flow, as it offers higher impact than fine-tuning straight-line code [Aho et al. 1986].

Instead, the cause is that such code tends to be simpler than a typical program code, and this simplicity deprives the compiler of optimization opportunities. At the same time, we observe that such simplicity may offer opportunities for using strategies not commonly exploited by compilers, such as reordering arithmetic operations within a basic block or exchanging machine instructions with semantically equivalent machine instructions. Thus, our work starts from the question:

How can we exploit the simplicity of cryptographic primitives in order to generate efficient and provably correct implementations of cryptographic functions?

Our Contribution

We present CryptOpt, a new code generator that produces highly efficient code tailored to the architecture it runs on. The task is split between *finding* performant program variants and *checking* that they have preserved program behavior. The former works via random search, and the latter works via a formally verified program-equivalence checker that should be applicable even with other optimization strategies. As a result, the random-search procedures need not be trusted, and, when we compose them with the Fiat Cryptography Coq-verified compiler [Erbsen et al. 2019] and our new equivalence checker, we get end-to-end functional-correctness proofs for fast assembly code – the fastest yet demonstrated for the cryptographic algorithms we study, when compiling automatically from high-level programs as we do.

To *find* performant machine-code variants, instead of relying on human heuristics, CryptOpt represents code generation as a combinatorial optimization problem. That is, CryptOpt considers a solution space that consists of machine-code implementations of the target function and uses techniques from the randomized-search domain to seek the best-performing implementation.

To optimize, CryptOpt first chooses an arbitrary correct implementation of the target function. It then mutates the implementation by either changing the instruction(s) that implements a certain operation or changing the order of operations. If the mutated implementation outperforms the original, the mutated implementation is kept; otherwise it is discarded.

Instead of trying to predict code performance, CryptOpt measures actual execution time. This approach is important because it avoids inaccuracies inherent in hardware models and allows CryptOpt to tailor produced code to the target processor while treating the processor itself as an opaque unit. Thus, once new hardware is released by manufacturers, CryptOpt can simply be run.

To *check* that generated code is correct, CryptOpt integrates with Fiat Cryptography [Erbsen et al. 2019]. That framework, implemented in Coq, already generates low-level IR programs proven to preserve behavior of high-level functional programs, and it has been adopted by all major web browsers and mobile platforms for small but important parts of their TLS implementations, so there is high potential for impact improving performance further without sacrificing formal rigor. CryptOpt begins with Fiat Cryptography IR programs and generates x86-64 assembly code. To avoid needing to model the random-search process, we instead built and proved an equivalence checker, which can compare programs across Fiat Cryptography IR and x86-64 assembly. Its two main pieces implement modest subsets of features known from SMT solvers and symbolic-execution engines. From SMT solvers, we take an E-graph data structure to canonicalize logical expressions via rewrite rules. From symbolic-execution engines, we take maintenance of symbolic states that tie registers and symbolic memory locations to logical expressions known to the solver. Thanks to their combined proof in Coq, none of these details need to be trusted.

We evaluate CryptOpt in two case studies for cryptographic arithmetic:

- In the first, we feed CryptOpt from Fiat Cryptography. Using nine different primes as input to Fiat Cryptography, we obtain a speedup of 1.42 on geometric means compared to GCC 11.3.0 (speedup 1.19 against Clang 14.0.0), across eight different x86-64 platforms (three AMD, five Intel). This x86-64 assembly code has end-to-end Coq proof.
- In the second case study, we create an input function from the C code of `libsecp256k1` [Bitcoin Core 2022], feed it into CryptOpt, and with the generated x86-64 assembly obtain an average speedup of 1.05 against the hand-optimized assembly code in `libsecp256k1` For this second case study, the verified checker lags in functionality and is not used to validate the results of random search.

Summary of Contributions. In summary, we make the following contributions in this paper:

- We present CryptOpt, a code generator that relies on combinatorial optimization instead of compiler heuristics for producing highly efficient code (Section 3). This broad approach is shown to apply to larger routines than past work has tackled successfully, while also being integrated with formal verification for the first time.
- We demonstrate that a relatively modest Fiat Cryptography extension (Section 4) suffices to enable integration with a wide range of backend compiler heuristics. We implemented and verified Coq functional programs with a minimal subset of well-known features from SMT solvers and symbolic-execution engines, leading to a single extractable compiler that checks assembly files for semantic equivalence with high-level functional programs. To our knowledge, this is the first such equivalence checker with mechanized proof from first principles.
- We generate formally verified high-performance cryptographic code optimized for eight CPU architectures, obtaining considerable speedups over GCC and Clang (Section 5).

At the level of detail we have presented so far, there are important similarities with the work of Bosamiya et al. [2020], who also combine random search through assembly programs with formal methods for cryptography. However, their starting point is handwritten assembly code within a relatively shallow metaprogramming framework – the inputs to their tooling literally dictate expected assembly code in performance-critical inner loops. Genetic search then goes on to find superior alternatives for some of those instructions. We show instead that random search can be used as part of a fully automated pipeline that starts from high-level functional programs, allowing us to generate fast code for multiple elliptic curves, not just multiple target architectures from a single algorithm for Bosamiya et al. We also deal with elliptic-curve arithmetic, raising issues of automatic register allocation, spilling, instruction selection (e.g., multiple addition instructions with different consequences for processor pipelines), and reasoning modulo associativity and commutativity, none of which show up in their AES-GCM case study (which naturally leans more on bitwise operations). Finally, our approach has a smaller trusted code base, being built within a proof assistant (Coq) rather than more specialized formal-methods tools (F^* and Z3).

It is also worth noting the close connection to the fields of *peephole optimization* and *superoptimization*. Peephole optimization deals with applying generic rules to replace short instruction sequences with higher-performance alternatives. Superoptimization uses smart enumeration to find the guaranteed shortest implementation of some chunk of functionality, with costs of combinatorial search often implying fairly stringent length limits. For instance, our literature survey did not uncover any tools generating sequences longer than tens of instructions, while we show good performance results using CryptOpt to generate functions of hundreds of instructions. We are also distinguished by handling differences in performance across different kinds of CPU, rather than applying relatively high-level cost models; CryptOpt works by running candidate programs on the CPUs themselves. The tradeoff vs. classic superoptimization is that we do not *guarantee* that we find the shortest or otherwise best implementations. Instead, we apply heuristics that do well empirically. Ours also seems to be the first work connecting this general sort of approach to verified compilation. See section 6 for more detailed coverage of specific past work in these categories.

We intend to release our implementation as open source, and it is attached to this submission as a code supplement.

2 OVERVIEW

Let us give the basic background on the problem we solve and sketch our solution, before devoting the following few sections to more detail.

2.1 Finite-Field Arithmetic for Cryptography

We focus on elliptic-curve cryptography (ECC), which is used widely in Internet standards like TLS. It involves certain geometric aspects that are orthogonal to our tooling, which supports arithmetic modulo large prime numbers, otherwise known as finite-field arithmetic (FFA).

Because the FFA dominates scalar multiplication, it is a performance-critical component in the implementations, and many tend to implement it by hand. E.g. targeting different architectures, the ubiquitous cryptographic library OpenSSL [OpenSSL 2022] has many hand-optimized implementations for Curve25519 and NIST P-256, which are both well-known instantiations of ECC. The Bitcoin blockchain uses yet another curve called secp256k1 for their block signatures. Its core library `libsecp256k1` contains hand-optimized code for the field arithmetic as well as a C implementation used as a fallback in architectures for which no optimized version exists.

FFA is not trivial to implement. In particular, a field element is typically represented by multiple limbs using several CPU registers, and thus every field operation requires multiple CPU cycles. However, these implementations tend to be straight-light code, heavy on arithmetic rather than control flow, leading to ineffective optimization by standard compilers. Human experts instead manually apply simultaneous instruction selection, instruction scheduling, and register allocation, which, going well beyond capabilities of off-the-shelf C compilers, should take into account microarchitecture details such as macro-op fusion [Celio et al. 2016; Ronen et al. 2004], cache prediction [Hooker and Eddy 2013; Lepak and Lipasti 2000; Subramaniam and Loh 2006], cache-replacement policies [Vila et al. 2020], and other (potentially undocumented) microarchitectural choices.

2.2 Our Starting Point: Fiat Cryptography

Erbsen et al. [2019] present their Fiat Cryptography framework, which translates descriptions of field arithmetic into code with a Coq proof of functional correctness. The starting point is a library of functional programs giving generic implementations of common implementation strategies (word-by-word Montgomery and Solinas) for field arithmetic. These functional programs are proved correct once and for all. A specific code-generation run selects one of the functional programs and specializes its parameters (e.g. field size) for a particular prime, resulting in an intermediate representation (Fiat IR). Fiat Cryptography then generates provably correct code in various languages (C, Java, Zig) from Fiat IR.

The earlier, verified compilation goes via the Fiat IR, which has translators to various backend languages. Our goal with CryptOpt was to follow the same philosophy of automated compilation alongside formal guarantees, so we chose to build CryptOpt to accept Fiat IR programs as inputs.

We will demonstrate how, compared to the released Fiat Cryptography tooling, we both improved performance and decreased the trusted code base, with theorems bottoming out in the operational semantics of x86-64 assembly instead of Fiat IR, removing compilers like GCC from the picture altogether (specifically not requiring us to trust in their correctness). We note that the generated code from Fiat Cryptography utilizes neither secret-dependent memory accesses nor secret-dependent branching, and thus it follows the constant-time programming paradigm [Almeida et al. 2016].

2.3 Our Solution

Figure 1 diagrams CryptOpt, with connections to preexisting tools. CryptOpt ingests the code of a function in the Fiat Cryptography IR and produces x86-64 assembly code. The CryptOpt optimizer then mutates the current x86-64 assembly code, measuring execution time of both the original and the mutated version. It then discards the slower one and iteratively continues this process until a predefined computational budget is used up (depicted in the middle part of Figure 1).

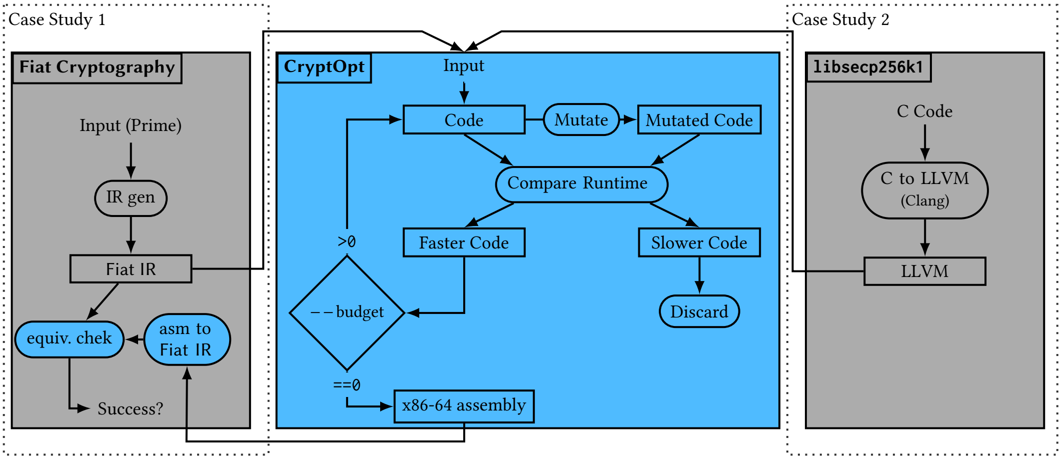


Fig. 1. CryptOpt system architecture

The random optimizer was codesigned with a separate program-equivalence checker, which, given an original IR program and its optimized assembly, is able to verify behavior preservation with no further hints. Thanks to this checker, the optimizer itself need not be trusted. The checker itself is implemented and verified in Coq, and we compose it with the original Fiat Cryptography tooling to extract a command-line tool that takes as input a functional program, the parameters to specialize it to, and a claimed assembly implementation. The classic Fiat Cryptography compiler is run, and its resulting IR is equivalence-checked against the provided assembly. This command-line tool has an end-to-end Coq theorem establishing that the assembly truly implements the original functional program (when a command-line invocation outputs “success”). This usage mode forms “Case Study 1” in Figure 1. As the optimizer is designed to implement the input function correctly, the checker always returns "success" in normal operation, though verification failures may appear during development.

To show CryptOpt's flexibility, we conducted another case study ("Case Study 2" in Figure 1). `libsecp256k1` provides two versions of the field arithmetic: a hand-optimized x86-64 assembly version and a C fallback. We transformed the C fallback to LLVM using Clang. Then, we created a simple LLVM-to-CryptOpt script to transform the necessary operations in LLVM to the input format for CryptOpt. Note that neither the hand-optimized x86-64 assembly nor the C fallback is formally verified, and thus this case study skips the formal-verification aspect, focusing just on further evaluation of the random optimizer.

Now we are ready to fill in the details of CryptOpt (random search in Section 3 and equivalence checking in Section 4) and how we evaluated it (Section 5).

3 RANDOM SEARCH FOR ASSEMBLY PROGRAMS

The first half of our approach is random search for improvements to assembly programs that implement Fiat IR programs. We draw on ideas from the mature field of random search, perhaps surprisingly requiring only some of its simpler and more intuitive elements.

3.1 Input Format

Recall (Section 2.3) that CryptOpt takes Fiat IR as an input. This intermediate language is truly a minimal one (see Figure 2), with the only noteworthy syntactic twist being that integer constants are presented as binary numbers (clearly indicating bitwidth), though we will often abbreviate them in decimal form, when bitwidth is clear from context. Operators may in general take not just multiple operands but also generate multiple results, associated with less common operators like addition with carry or multiplication producing double-wide results via two output words.

Variable	x		
Binary integer	b		
Operand	e	::=	$x \mid b$
Operator	o	::=	! | & | $*$ | + | − | << | = | >> | ~ | or | addcarryx | cmovznz | mulx | static_cast | subborrowx
Expression	E	::=	return $e \mid x, \dots, x \leftarrow o(e, \dots, e); E$

Fig. 2. Fiat IR syntax

Input programs contain no branches or explicit memory accesses, just accesses of local variables. As a result, generated assembly programs will not be too much more complex, avoiding all memory aliasing and restricting pointer expressions to be constant offsets of either function parameters (for data structures passed by reference) or the stack pointer (for spilled variables). The assembly we generate is timing-secure for the same reasons that Fiat IR is timing-secure; we only use primitive program mutations that either preserve relevant behavior (the trace of program memory accesses and control-flow decisions) or add new behaviors in ways independent of secrets (all memory accesses are to constant offsets from either the stack pointer or arrays that are function parameters), so by induction the initially secure programs are still secure after optimization.

Algorithm 1: An Example Function

```
input  : $X, Y, Z$ such that $0 \leq X, Y, Z < 2^{63}$
output : $O = 2^{64} O_1 + O_0 = X \cdot (Y + Z) + Z^2 + Y$
function example($X$, $Y$, $Z$)
begin
    $t_2, \; t_1 \leftarrow \textsc{Mul}_1(Z, \; Z)$
    $c_\varnothing, \; t_0 \leftarrow \textsc{Add}_1(Y, \; Z, \; 0)$
    $t_4, \; t_3 \leftarrow \textsc{Mul}_2(t_0, \; X)$
    $c_0, \; t_5 \leftarrow \textsc{Add}_2(t_3, \; t_1, \; 0)$
    $c_1, \; O_0 \leftarrow \textsc{Add}_3(t_5, \; Y, \; 0)$
    $c_\varnothing, \; t_6 \leftarrow \textsc{Add}_4(t_4, \; t_2, \; c_0)$
    $c_\varnothing, \; O_1 \leftarrow \textsc{Add}_5(t_6, \; 0, \; c_1)$
    return $O_1, \; O_0$
end
```

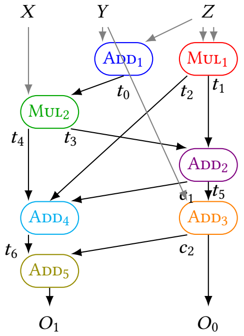


Fig. 3. Data flow of the running example in Algorithm 1

Algorithm 1 shows a Fiat IR program, which we will use throughout this paper as a running example. The program takes three inputs (X, Y, Z) and outputs $X \cdot (Y + Z) + Z^2 + Y$ using two types of operations: Add and Mul. The operation Add adds two 64-bit numbers and one 1-bit carry, then returns the sum as one 64-bit number and one 1-bit carry. The operation Mul multiplies two 64-bit numbers, then returns the 128-bit product as two 64-bit words. For simplicity, we assume that the arguments are in the range $0 \leq X,\ Y,\ Z < 2^{63}$, allowing us to ignore some carries known to be 0. We mark these carries with $c_{\varnothing}$.

In the absence of memory aliasing and control flow, the restrictions on operation evaluation order in the programs we optimize are captured by the data flow. Figure 3 shows the data-flow graph of the code in Algorithm 1.

3.2 Optimization Strategy

A unique feature of CryptOpt is that instead of relying on heuristics for generating the code, CryptOpt explicitly casts the problem as a combinatorial optimization problem. That is, CryptOpt searches the space of assembly-code sequences that implement the input function, looking for the best-performing implementation simply by utilizing the execution time as the objective function. For optimization, we employ the random local search (RLS) strategy [Auger and Doerr 2011; Doerr and Neumann 2019]. RLS is often an efficient and effective baseline for finding and analyzing local optima. Each run of RLS aims to find a local optimum by iteratively moving from one solution to a random neighbor (i.e. mutation) given a defined variation operator. This move happens if that new neighbor performs at least as well.

RLS is often highly sensitive to the initial conditions. To address such erratic behavior, the simple Bet-and-Run heuristic [Fischetti and Monaci 2014; Weise et al. 2019] turns the sensitivity to the initial conditions into an advantage by employing multiple runs. A typical use of Bet-and-Run starts with multiple independent runs of RLS, each optimizing for a predefined number of mutations. After this initialization step, the algorithm selects the best run and lets this run continue optimizing from that step, stopping when a total number of mutations is explored. We adopted Bet-and-Run in CryptOpt, i.e. search begins in parallel on several random assembly realizations of the input Fiat IR program, switching at some point to focus on the most promising one.

3.3 Code Generation

As for the process whereby those assembly candidates are generated, it can be understood as split between instruction scheduling, instruction selection, and register allocation, in that order. Each phase makes certain arbitrary decisions that may be reversed by a later random mutation. We summarize the phases here before returning to details of generation and mutation.

First, for instruction scheduling (see Section 3.4), dataflow analysis determines dependency structure across instructions, using which, a random topological sort of that graph is chosen to use in the assembly. Then, for instruction selection (see Section 3.5), for each high-level operation, we choose among the compatible x86-64 assembly instructions. Finally, in our setting, register allocation (see Section 3.6) arises mostly in ensuring compatibility with operand restrictions of the instructions that were selected. For example, consider the objective to *multiply two values* (one single instruction can at most read from one memory location), in a context where both operands reside in memory. The relevant dimension of freedom is which value to load into a register explicitly and into which register. After the decisions of the three phases have been recorded, it is easy to read off the chosen assembly program sequentially. Recall that all of these decisions may be revisited later in random mutations.

To present details of the three phases, we focus on the example program from Algorithm 1. Figure 4 shows how operations are initially ordered and potential scheduling intervals; the mutations are shown in Figure 5; until finally we see the emitted code with and without the effects of the mutations in Figure 7. We will reference those illustrations later as needed.

3.4 Instruction Scheduling

Thanks to the simplicity of Fiat IR, every variable is assigned exactly once in the straight-line programs that CryptOpt takes as input. Consequently, any operation can be evaluated whenever all its inputs have been computed. In the example, Mul$_2$ can be evaluated when t_0 and X are computed.

Initial Ordering. To create an initial ordering, CryptOpt simply computes a topological sort of the dataflow graph. Figure 4 shows an example of an initial ordering for our running example. We would like to emphasize that the selection does not rely on any heuristic. While this initial selection does affect the subsequent mutated ordering, our mutation strategy guarantees that any possible ordering can be reached from any initial random starting point via a sequence of mutation steps.

Mutation Step. An instruction-scheduling mutation step of CryptOpt randomly selects one operation in the current ordering. This operation is moved to a randomly chosen location within the interval where it would be valid to move: not before the last assignment responsible for setting a variable that is used here as an operand, nor after the first assignment that reads the variable being set here. We begin with the ordering shown in Figure 4. The vertical colored bars indicate the intervals where the respective operations can be scheduled. Step α of Figure 5 shows the effect of moving Add$_4$ up one position.

Selecting the position to move an operation to is biased towards larger moves, i.e. further away from the initial position, in an effort to minimize spills by minimizing distance between computing a value and using it.

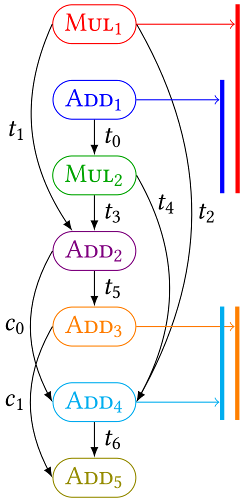


Fig. 4. One ordering. Round rectangles show the operations, which are evaluated top-down. Arrows indicate creation and consumption of intermediate values. Vertical bars show the intervals in which an operation can be scheduled.

3.5 Instruction Selection

An important property of complex instruction sets such as x86-64 is that there can be multiple alternative implementations for each high-level operation, each with a slightly different semantics and impact on the processor pipeline. To match machine instructions to operations, CryptOpt uses templates describing the possible implementations for each operation.

Consider, for example, the Add operation, which can be implemented using multiple instructions, such as `add`, `adcx`, or `adc`. Though semantically equivalent, these choices influence program state differently. For instance, in the case of no input-carry (Add$_{1-3}$), implementing an Add operation using an `adcx` instruction requires clearing the carry flag. The template of the `adcx` instruction accounts for that and issues a clear-carry instruction (`clc`) before the `adcx` instruction, unless CryptOpt determines that the carry is already clear. Lines 7 and 8 of Figure 7 show this choice for Add$_1$.

Initial Template Mapping. When creating the initial code, CryptOpt selects a random template for each operation as an initial mapping. Figure 5 (I) shows an example of a possible mapping of templates to the operations of our example function. The figure indicates the operation Mul$_1$ is implemented using the `mulx`, Add$_1$ uses `adcx`, and so forth.

Mutation Step. To mutate the instruction selection, CryptOpt chooses an operation at random and replaces the template for that instruction with one alternative. Mutation β in Figure 5 shows an example of replacing the template for Add$_1$ to use the `add` instruction instead of the original `adcx`.

Flag Spills. The series of additions $\textsc{Add}_2, \ldots, \textsc{Add}_5$ show not only the influence of different orderings on the resulting code but also how a template is used to handle flag spills. Consider the code in Figure 7a where $\textsc{Add}_2$ results in `CF`=c_0. Later on, $\textsc{Add}_3$ needs to write its own `CF`=c_1. Therefore, the current `CF` needs to be spilled (into another register). In this case, it does so with `setc dl` (line 18). Similarly, $\textsc{Add}_4$ then needs to spill c_1 (line 23) to avoid overwriting it with its $c_\varnothing$.

At this point (line 25), CryptOpt needs to add the values of three operands, i.e. $t_4 + t_2 + c_0$. As the x86-64 assembly language does not have a single three-operand addition instruction, CryptOpt first adds two operands together then adds the sum to the third one. Note that this changes the evaluation order from $t_4+t_2+c_0$ to $(t_2+c_0)+t_4$. As the $\textsc{Add}$ operation is associative and commutative, any evaluation order maintains correctness. The equivalence checker accounts for this change in evaluation order (see Section 4.4).

Strength Reduction. Some of the templates we use support limited forms of strength reduction. For example, we have templates to implement multiplication-by-a-constant, including using a right-shift operation (i.e. $x \times 8 \implies x \gg 3$), a series of multiplications and additions (i.e. $x \times 5 \implies x \times 2 + x$), or a combination thereof. The final template selection is left to the optimizer.

3.6 Register Allocation

The register-allocation step of CryptOpt achieves two aims. It must both decide which values are assigned to registers and which registers to spill to memory when running out of registers. CryptOpt uses random search for the former and a deterministic strategy for the latter.

Register Assignment. Registers need to be assigned in two main cases: when computing a new value and when reading a value from memory, either from the input or following a register spill. CryptOpt keeps track of the live registers, allocating a free register if one is available. If none is available, CryptOpt spills the contents of a register to memory and uses the freed register. To choose the register to spill, CryptOpt scans the future use of all registers and spills the register whose next use is furthest based on the current operation order.

For example, lines 11–12 in Figure 7 implement the $\textsc{Mul}_2$ operation. (For this example, we assume that the architecture only has three general-purpose registers: `r8`, `r9`, and `rdx`.) At this point, registers `r8`, `r9`, `rdx` have been used for t_2, t_1, and t_0, respectively. Hence, we need to spill a register. Observing that t_2 is not required until $\textsc{Add}_4$, whereas t_0 and t_1 are used earlier, CryptOpt spills `r8` (Line 11).

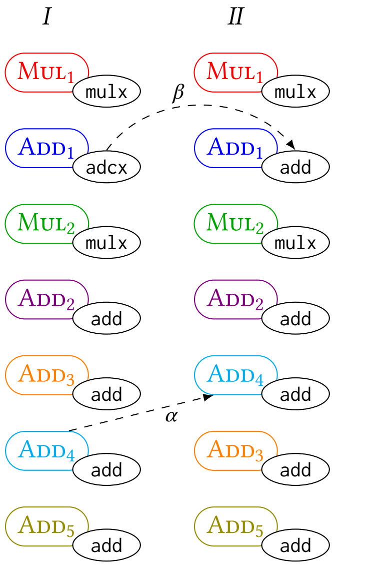


Fig. 5. I: Initial ordering of operations (colored rounded rectangles) with attached templates (black ellipses); II: After two mutations α and β: mutation α in topological ordering, mutation β in instruction-template selection. Dataflow arrows omitted; dashed arrows indicate mutations.

Memory Loads. Most arithmetic operations in the x86-64 architecture support instruction formats that take one argument from memory. When an argument of an operation is in memory, CryptOpt tries to use such an instruction format. When this is not possible, e.g. when the values of two arguments are in memory, CryptOpt resorts to loading a value from memory into a register. In the case of associative operations, such as addition and multiplication, CryptOpt initially randomly chooses the value to load, though mutations may later alter the choice.

```
Inductive REG := rax | rcx | (* ...65 omitted... *) | r15b.
Inductive AccessSize := byte | word | dword | qword.
Record MEM := { mem_bits_access_size : option AccessSize; mem_base_reg : option REG;
                mem_scale_reg : option (Z * REG); mem_base_label : option string;
                mem_offset : option Z }.
Inductive FLAG := CF | PF | AF | ZF | SF | OF.
Inductive OpPrefix := rep | repz | repnz.
Inductive OpCode := adc | adcx | add | adox | and | bzhi | call | clc | cmovb | cmovc | cmovnz | cmp
  | db | dd | dec | dq | dw | imul | inc | je | jmp | lea | mov | movzx | mul | mulx | pop | push
  | rcr | ret | sar | sbb | setc | seto | shl | shlx | shr | shrx | shrd | sub | test | xchg | xor.
Record JUMP_LABEL := { jump_near : bool; label_name : string }.
Inductive ARG := reg (r : REG) | mem (m : MEM) | const (c : Z) | label (l : JUMP_LABEL).
Record NormalInstruction := { prefix : option OpPrefix; op : OpCode; args : list ARG }.
```

Fig. 6. Syntax of Coq embedding of x86-64 assembly

Exploiting Simplicity. Finally, we note that the absence of control flow and avoidance of human heuristics are key enablers for memory spill decisions. The former simplifies dependency analysis, allowing CryptOpt to determine the requirements for downstream operations. The latter allows CryptOpt to examine the entire function rather than focusing on instructions within a peephole window, a commonly used technique by off-the-shelf compilers.

3.7 Cost-Function Evaluation

We compare different random program variants by running them on the actual processors of interest. Controlling noise in running-time measurement is of utmost importance because with too much noise, random search could be driven into unproductive oscillation. We found the details of such measurement surprisingly difficult to get right. Appendix A gives those details, which rely on running a chosen number of repetitions of the two candidate programs, interleaved in a random order, then returning the median timing observed per program.

```
; t_2, t_1 ← Mul_1(Z, Z)
mov   rdx,   [Z]
mulx  r8,    r9,    [Z]

; c_∅, t_0 ← Add_1(Y, Z, 0)
mov   rdx,   [Y]
clc
adcx  rdx,   [Z]

; t_4, t_3 ← Mul_2(t_0, X)
mov   [rsp], r8 ; spill
mulx  r8,    rdx,   [X]

; c_0, t_5 ← Add_2(t_3, t_1, 0)
add   r9,    rdx

; c_1, O_0 ← Add_3(t_5, Y, 0)
setc  dl          ; dl ← c_0
add   r9,    [Y]
mov   [O_0], r9

; c_∅, t_6 ← Add_4(t_4, t_2, c_0)
setc  r9b         ; r9b ← c_1
movzx rdx,   dl
add   rdx,   [rsp]
add   r8,    rdx

; c_∅, O_1 ← Add_5(t_6, 0, c_1)
movzx r9,    r9b
add   r8,    r9
mov   [O_1], r8
```

(a) Code from I (initial code)

```
; t_2, t_1 ← Mul_1(Z, Z)
mov   rdx,   [Z]
mulx  r8,    r9,    [Z]

; c_∅, t_0 ← Add_1(Y, Z, 0)
mov   rdx,   [Y]

add   rdx,   [Z]

; t_4, t_3 ← Mul_2(t_0, X)
mov   [rsp], r8 ; spill
mulx  r8,    rdx,   [X]

; c_0, t_5 ← Add_2(t_3, t_1, 0)
add   rdx,   r9

; c_∅, t_6 ← Add_4(t_4, t_2, c_0)

adc   r8,    [rsp]



; c_1, O_0 ← Add_3(t_5, Y, 0)

add   rdx,   [Y]
mov   [O_0], rdx

; c_∅, O_1 ← Add_5(t_6, 0, c_1)

adc   r8,    0
mov   [O_1], r8
```

(b) Code from II (code after two mutations)

Fig. 7. Emitted assembly code. Highlighted lines show effects from mutations.

4 CHECKING PROGRAM EQUIVALENCE

Our goal with CryptOpt was to preserve or even strengthen the formal guarantees of Fiat Cryptography. One way to achieve that goal would have been to verify the whole random-search process with Coq, but we wanted to find a simpler strategy that would have the side benefit of also potentially supporting automatic verification of various handwritten assembly solutions. Therefore, we decided to write a program-equivalence checker in Coq and verify it. Industrial-strength translation validation as in Alive [Lopes et al. 2021] is now well-established, but again, proving such a tool from first principles would be a substantial undertaking. We were curious, instead, how far we could get implementing (and proving) our checker from scratch, lifting just those features of more conventional checkers that turned out to be important in our domain.

Figure 6 shows a nearly complete description of the x86-64 assembly syntax accepted by our checker. For better error messages, some control-flow opcodes like `jmp` are included, though they will always be rejected by the checker. This type of syntax trees is given a very standard semantics, in the form of an interpreter as a total Coq function. The simplicity of syntax and semantics is important, since both are referenced by the final theorem for any specific compilation, while the syntax and semantics of Fiat IR drop out of the picture as untrusted.

4.1 Background

Formally proved compilers are the gold standard to address concerns of optimization soundness. For instance, CompCert [Leroy 2009; Leroy et al. 2016] was proved as a correct C compiler using the Coq theorem prover, which we also rely on. However, proving a whole compiler can be very labor-intensive, and thus it is often appealing to prove a *checker* for compiler outputs, known as a translation validator. For instance, CompCert was extended in that way [Tristan and Leroy 2008]. To date, however, the formally proved translation validators have not incorporated reasoning with algebraic properties of arithmetic, as we found we needed in CryptOpt.

In contrast, translation validation with SMT solvers uses rich reasoning to prove equivalence between the high-level input program and the obtained low-level output. The Alive project [Lopes et al. 2021] for LLVM is a good example. Compared to work with formally proved translation validators, Alive and similar tools include much larger trusted bases, for instance including a full SMT solver like Z3 [de Moura and Bjørner 2008]. Some SMT solvers have been extended to produce proofs that can be checked in tools like Coq, as in SMTCoq [Armand et al. 2011].

In our experience, these tools hit performance bottlenecks when working with large bitvectors. Even if we imagine those issues as fixed some day, there are still benefits to creating a customized checker, keeping just the relevant aspects of SMT. A benefit of a slimmed-down custom prover is that it becomes more feasible to prove the prover itself, rather than just a checker for its outputs, which improves performance and reduces surprise from proof-generation bugs.

4.2 Code Verification

The best-performing implementation produced by CryptOpt is assured to be correct through a formally verified equivalence checker. Specifically, we verify the correctness of CryptOpt's output through functional equivalence between programs in the Fiat IR and x86-64 assembly. We developed simple symbolic-execution engines for the relevant subsets of both languages, producing program-state descriptions in a common logical format. The next task is to check that function-output registers and memory locations store provably equivalent values between the two programs. To that end, we developed a simple equivalence theorem prover that borrows from SMT solvers, using a similar (E-graph) data structure.

The public API connecting the checker's two main components is based on the following definition of expressions:

$$\begin{array}{rlcl} \text{Integer constants} & n & & \\ \text{Variables} & x & & \\ \text{Operators} & o & & \\ \text{Expressions} & e & ::= & n \mid x \mid o(e, \ldots, e) \end{array}$$

The E-graph exposes a function `internalize` that takes in an expression and returns a variable now associated with that expression's value. Importantly, the expression will usually mention variables that came out of previous calls, which lets us work with exponentially more compact representations than if we expanded out all variables. The internal E-graph takes advantage of this sharing for efficiency. Also, crucially, every `internalize` invocation proactively infers equalities between previously considered expressions and the new expression and its subterms. Thus, we may check two expressions for equality simply by verifying that `internalize` maps them to the same variable, which becomes a chosen representative of an equivalence class of expressions.

4.3 Symbolic Execution

We built symbolic-execution engines for the two relevant languages, Fiat IR and x86-64 assembly.

The engine for the IR is simpler than for x86-64 assembly. Programs in this IR are just purely functional sequences of variable assignments with expressions that effectively already match the grammar we just gave. Thus, to evaluate such a program symbolically, we just maintain a dictionary associating program variables to logical variables; the latter are effectively handles into the E-graph.

The execution engine for x86-64 assembly is moderately more complicated. Now the symbolic state associates not program variables but *registers and memory addresses* with logical variables. We take advantage of the stylized structure of cryptographic code to simplify the treatment of memory. The only valid pointer expressions are constant offsets from either function parameters (standing for cells within data structures passed to the function) or the stack pointer (standing for spilled temporaries). Thus, it is appropriate to make the symbolic memory a dictionary keyed off of *pairs of logical variables and integers*. The logical variable is the base address of an array in memory, while the integer gives a fixed offset into one of its words.

With this convention fixed, it is fairly straightforward to march through the instructions in a program, updating the register and memory dictionaries with the results of `internalize` calls. The symbolic executor effectively breaks each (possibly complex) x86 instruction down into multiple simpler operations on integers. There are multiple operations because many instructions affect both flags and their explicit destination registers. The explanations of those effects are expressions from the grammar above, using a relatively small vocabulary of bitvector operators.

At the end of symbolic execution of an assembly function, we pull the output values out of calling-convention-designated registers and memory locations. These can then be compared against the explicit return value of a Fiat IR program. Both are expressed as logical variables connected to a common E-graph, so they should be syntactically equal exactly when the E-graph found a proof of equality. Importantly, both symbolic states are initialized with common logical variables.

4.4 The E-graph

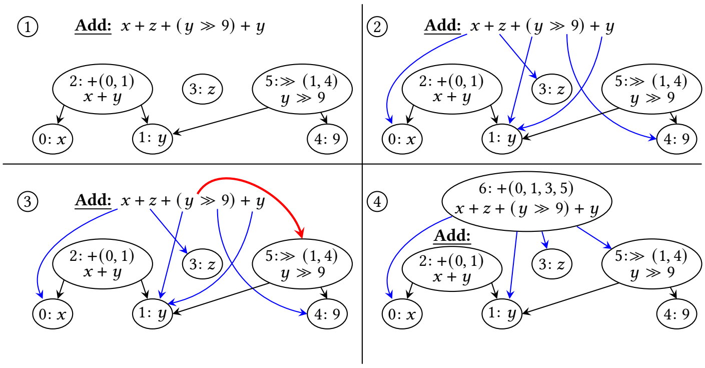


Fig. 8. Example of operations on an E-graph-style structure

Following SMT-solver conventions [Detlefs et al. 2005], our E-graphs include nodes for equivalence classes of symbolic expressions, in addition to the edges representing subterm relationships. Each node is configured to present the most compact representation of its equivalence class. For instance, whenever a node becomes provably equal to a constant, it is labeled with that constant, without outgoing edges. When a node is most succinctly expressed as a sum, it is labeled with a "+" operator and has edges to the other nodes being added.

Figure 8 animates a simple example. It steps through stages of adding a new node to the graph, representing the new symbolic expression: $(x + z + (y \gg 9) + y)$. Step 1 shows the initial state. Nodes 2 and 5 are labeled with operators and IDs of operands. To process the expression we are evaluating, we first look up existing graph nodes for all of its leaf expressions, as Step 2 shows. Then we proceed bottom-up in the expression tree, finding an existing node or building a new one for each subexpression. In this case, we next need to find a node for $y \gg 9$, in Step 3. As we resolved y to node 1 and 9 to node 4, we are able to search the DAG for a node already labeled with operator $\gg$ and argument nodes 1 and 4, finding node 5. In the final step, Step 4, we need to find a node for $x + z + (y \gg 9) + y$. We have node IDs for all four operands of addition, and we *sort* them by ID, taking advantage of associativity and commutativity of addition, to find a canonical description of this node. No existing node has that description, so we add a new one. The main complication absent from this example is normalization with rewrite rules going beyond associativity and commutativity; the approach is parameterized on a set of such rules (which must have proofs in Coq), and they are applied as each node of the input AST is processed.

Algorithm 2 presents the `internalize` algorithm more generally. Without loss of generality, we assume it is called on an expression that is an operator applied to a list of variables, which are used as E-graph node names. To internalize expressions with constants and/or deeper nesting of operands, we can simply traverse their trees bottom-up, internalizing each node to obtain a variable (i.e. E-graph node) to replace it with. Elided from the figure are additional rewrite rules to complement the one we include, which notices that an operation to extract the low byte of a constant is extraneous, when the constant is low enough. The algorithm ends in a linear scan of the E-graph for existing nodes matching the normalized input, which perhaps surprisingly turned out to be more than fast enough for our examples.

Algorithm 2: Internalize expression into the E-graph

```
input  : Op an operator, Args its list of argument variables
output : n, a variable / graph node for the expression's
         equivalence class

function internalize(Op, Args)
begin
    if Op is associative then
        for argument a in Args do
            if a is also labeled with Op then
                Expand a in the list into its own
                  E-graph neighbors
            end
        end
    end
    if Op has identity element e then
        Remove from Args any variable whose node is
          labeled with constant e.
    end
    if Op is LowByte and len(Args) = 1 and Args[0] is labeled
      with a constant below $2^8$ then
        return Args[0]
    end
    /* Many other algebraic rewrite rules */
    if Op is commutative then
        Sort Args by textual variable name.
    end
    for node n in the E-graph do
        if n is labeled with Op and has Args as its edge list
          then
            return n
        end
    end
    n ← new E-graph node;
    n.label ← Op;
    n.edges ← Args;
    return n
end
```

The actual implementation includes a general range analysis to establish upper bounds on variable nodes based on bounds on their inputs. This feature is used to elide truncations whenever appropriate and to gate other rewrite rules: for example, the carry bit resulting from adding small numbers is always 0, and a sum of a couple of carry bits fits in a byte.

Here it is interesting to note which standard features of SMT-based verification tools did *not* need to be implemented, saving us from needing to prove their soundness. We incorporated no SAT-style management of case splits, nor did we need to include specialized cooperating decision procedures for domains like arithmetic on mathematical integers or bitvectors. It sufficed to stick to congruence-closure-style reasoning in the theory of equality with uninterpreted functions, augmented with a modest pool of rewrite rules, as SMT solvers are also often effective at applying. Note also that we omitted a central complexity of E-graph implementations: merging nodes (usually through union-find algorithms) as it is discovered that they are equal. Our domain is simple enough that relevant equalities are always discovered at node-creation time. As for symbolic execution, we avoided the characteristic complexities of control flow (e.g. merging logical states) and memory access (e.g. flexible-enough addressing that pointer aliasing is nontrivial to check).

4.5 Proof

We proved the Coq implementation of equivalence checking correct, in the sense that symbolically executing an assembly (or IR) program produces an expression (E-graph) that would evaluate to the same output as the original program from all starting states. A fairly direct corollary is our main theorem: if two programs are symbolically executed with the same (symbolic) inputs and produce outputs that are represented by the same E-graph nodes, then the programs are equivalent. We combine the Fiat IR symbolic evaluator with the larger Fiat Cryptography pipeline, and we verify it against the existing semantics of the IR, before extracting the pipeline to a command-line program. That program takes as input a choice of cryptographic algorithm, its numeric parameters, and an assembly file, and it checks that the assembly file matches the behavior of the algorithm, by comparing it with the Fiat IR code (itself generated by a verified compiler).

It is important that the top-level theorem of the equivalence checker avoids mentioning any specifics of symbolic execution or E-graphs, as those are relatively complex techniques. Instead, that theorem refers only to formal semantics of Fiat IR and x86-64 assembly. Slightly more precisely, we depend on the following preconditions, often stated using separation logic [Reynolds 2002].

- Calling-convention-designated registers hold input values, input-array base pointers, and output-array base pointers.
- Input-array base pointers point to arrays in memory holding input values.
- Output-array base pointers are valid.
- `rsp` points to the end of the stack, which must be valid.

Then the main theorem concludes the following postconditions.

- Input-array base pointers are still valid.
- Output-array base pointers point to arrays in memory holding the output values.
- The stack base-pointer address is still a valid pointer to an array of the right size.
- Callee-save registers have the same values as before the code execution.
- All other memory is untouched.

A few other concerns must be stated in the full theorem, including that all source-program initial variable values fit in machine words, arrays are laid out contiguously (the symbolic-execution engine allows more flexible specification of memory contexts), and every array has a start address stored directly in a register.

5 PERFORMANCE EVALUATION

In this section, we evaluate the performance of code produced by CryptOpt. Specifically, we answer four main questions:

(1) How does optimization progress over time (Section 5.2)?
(2) How does CryptOpt compare with traditional compilation (Section 5.3)?
(3) Is CryptOpt optimization platform-specific (Section 5.4)?
(4) How does CryptOpt-optimized code perform as part of a cryptographic implementation (Section 5.5)?

We first describe the setup and the procedures we use for the evaluation. We then describe the experiments we carry out and their results.

5.1 Experimental Setup

In this subsection, we describe the hardware platforms and discuss the *generation* of results through optimization, plus the *evaluation* of which one is the best of all the results.

Hardware Platforms. To compare across multiple processor architectures, we evaluate CryptOpt on multiple hardware platforms, summarized in Table 1. We did not observe differences in optimization behavior on machines with SMT enabled or disabled. We use Ubuntu Server 22.10.1 LTS for all machines, with all packages up-to-date. Moreover, because performance counters are not available on the efficiency cores of the i9 12G, we only use the performance cores on the machine.

Table 1. Overview of target machines used in the experiments

Name	CPU	μ-architecture
1900X	AMD Ryzen Theadr. 1900X	Zen 1
5800X	AMD Ryzen 7 5800X	Zen 3
5950X	AMD Ryzen 9 5950X	Zen 3
i7 6G	Intel Core i7-6770HQ	Skylake-H
i7 10G	Intel Core i7-10710U	Comet Lake-U
i9 10G	Intel Core i9-10900K	Comet Lake-S
i7 11G	Intel Core i7-11700KF	Rocket Lake-S
i9 12G	Intel Core i9-12900KF	Alder Lake-S

Generation. Every platform has at least four physical cores. For a fair comparison, we run the same number of optimization processes in parallel on all platforms. We pin the optimization processes to cores to reduce noise due to context switching. We chose to run three optimizations in parallel, keeping one core free for general OS activity. This results in three assembly files for each primitive per platform, of which we report only the best performing. The number of assembly instructions varies from 57 to 1253, depending on the field, the operation, and how well the optimization went. See Table 2 for details.

Bet-and-Run. Each optimization process has a budget of 200 000 mutations. In the bet stage, we explore 20 initial candidate solutions, optimizing each for 1 000 mutations. Hence, overall, we use 20 000 mutations, which are 10% of the total budget, for the bet part. The remaining 180 000 mutations are used for the run stage of the Bet-and-Run strategy. With those parameters, the optimization for the 18 Fiat IR primitives takes between 36 and 70 wall-clock hours depending on the machine.

Code Verification. When combined with Fiat Cryptography, the optimization process verifies that the code it produces is equivalent to the Fiat IR code. We perform verification only on the final output of the optimization. Table 3 shows the time it takes to verify one assembly implementation as an average over the eight platforms.

Table 2. Instruction count; average over all eight platforms

Primitive	Multiply	Square
Curve25519	169.500	120.875
NIST P-224	220.750	219.375
NIST P-256	204.000	199.500
NIST P-384	709.875	698.250
SIKEp434	985.875	964.625
Curve448	588.125	404.625
NIST P-521	542.000	337.500
Poly1305	75.750	61.250
secp256k1	233.375	224.000

Performance Metric. To compare the performance of different implementations we need a stable metric that measures the performance. To reduce system noise we fix the CPU frequency, disable boosting, and set the governor to `performance`. We note that we only apply these settings when evaluating the performance but not during optimization. We observe that the performance loss due to controlling CPU speed outweighs the benefits of the stability it provides. For more details on this aspect of our experimental setup, see Appendix A.

5.2 Optimization Progress

The first evaluation question we answer is how the optimization progresses over time. We include details in Appendix B, but a summary is that optimization progress is roughly logarithmic in the number of mutations, with some interesting differences in measurement stability across platforms.

5.3 CryptOpt vs. Off-the-Shelf Compilers

Table 3. Verification times (average over all eight platforms)

Primitive	Multiply	Square
Curve25519	0.29 *s*	0.18 *s*
NIST P-224	2.16 *s*	1.94 *s*
NIST P-256	1.07 *s*	0.95 *s*
NIST P-384	29.09 *s*	25.81 *s*
SIKEp434	315.23 *s*	284.32 *s*
Curve448	3.02 *s*	1.44 *s*
NIST P-521	3.14 *s*	1.46 *s*
Poly1305	0.07 *s*	0.06 *s*
secp256k1	1.78 *s*	1.65 *s*

To compare CryptOpt with traditional compilers, we use the implementations of finite-field arithmetic as produced by Fiat Cryptography. Specifically, we use Fiat Cryptography to produce implementations of the multiply and the square functions in nine fields. We consider prime fields of the standardized NIST P curves: Curve P-224, Curve P-256, Curve P-384 and Curve P-521 [NIST 2000]. Moreover, we consider the field of the popular ‘Bitcoin’ curve secp256k1 [Certicom Research 2000], the high-speed de-facto standard Curve25519 [Bernstein 2006], and a high-security Curve448 [Hamburg 2015]. In addition to elliptic-curve cryptography, we apply our method to the underlying fields of the post-quantum scheme SIKEp434 [Azarderakhsh et al. 2019] and of the Poly1305 message-authentication scheme [Bernstein 2005]. It is worth noting that Erbsen et al. [2019] reported that their generated C code was roughly the best-performing available for all elliptic curves, up to the usual vagaries of C-compiler optimizers fluctuating in behavior across versions, so it makes sense to use that C code as our performance baseline.

We run CryptOpt on each of the eight platforms summarized in Table 1 and select the best result as described in Section 5.1. Additionally, we compile the equivalent C code, as produced by Fiat Cryptography, with GCC 11.3.0 [GCC 2022] and Clang 14.0.0 [Clang 2022]. We use the highest optimization level the compilers support and enable native support using the compilation switches `-march=native -mtune=native -O3`.

Table 4. Geometric means of CryptOpt vs. off-the-shelf compilers.

	Multiply		Square	
Curve	Clang	GCC	Clang	GCC
Curve25519	1.19	1.14	1.14	1.18
P-224	1.31	1.87	1.24	1.84
P-256	1.27	1.79	1.30	1.85
P-384	1.12	1.66	1.08	1.60
SIKEp434	1.30	1.70	1.29	1.83
Curve448	1.02	0.95	1.00	0.99
P-521	1.20	1.06	1.25	1.11
Poly1305	1.10	1.15	1.09	1.16
secp256k1	1.34	1.73	1.32	1.74

Table 4 shows a summary of the results. For each function the table presents the geometric mean performance gain of CryptOpt over GCC and Clang. The mean is calculated over the different platforms; see Table 6 in Appendix C for full details. The table shows that CryptOpt achieves significant performance gains in the majority of functions. The performance gains are somewhat more modest when the produced code does not require any memory spills, as is the case for operations in the fields of Curve25519 and Poly1305. For large fields, as in P-521 and Curve448, CryptOpt is less successful, achieving modest gains (or on par) for the former and underperforming for the latter. We note that the code produced for these curves is quite large. As an example, the resulting x86-64 assembly files for Curve448-mul are in the range of 525–700 instructions and for P-521-mul 463–630 instructions depending on the platform. For comparison, Curve25519-mul is in the order of 160–180 instructions, and in the range of 70–80 for Poly1305-mul. We suspect that CryptOpt’s simple mutations have less impact on the execution time for those big functions than what would be needed to be measurable and direct the optimizer towards the optimal instruction sequences. We leave more sophisticated genetic-improvement strategies for future work.

5.4 Platform-Specific Optimization

CryptOpt optimizes the execution time of the function on the platform it executes on. Because different platforms have different hardware components, the fastest code on one platform is not necessarily the fastest on another. Surprisingly, sometimes it does pay off to optimize on a different platform than the target, apparently when the host supports more stable performance measurement. See Appendix C for details.

5.5 Scalar Multiplication

So far we have focused on the optimized functions in isolation. In this section, we investigate the use of the field operations within the context of elliptic-curve cryptography. Specifically, we investigate implementations of two popular elliptic curves: Curve25519, and secp256k1. We compare the performance of 15 implementations of these curves, four of which use CryptOpt code for field operations. In Table 5, we summarize the implementations we investigate.

State of the art. For Curve25519, the SUPERCOP benchmark framework [Bernstein and Lange 2022] provides us with many implementations: sandy2x [Chou 2015], amd64-51 and amd64-64 [Chen et al. 2014], as well as donna and donna-c64 [Langley 2022]. OpenSSL [OpenSSL 2022] provides many implementations (two assembly-based and one portable C version) and chooses at run time which one to use, opting by default for OSSL fe-64, a variant of amd64-64. Additionally, Project Everest [HACL 2022] provides special parallelized assembly implementations for the fields. “Parallelized” in this context means, that two independent computations are interleaved.

For secp256k1 we use two implementations from the libsecp256k1 library [Bitcoin Core 2022], one is hand-optimized assembly, the other one is portable C.

CryptOpt-Based Implementations. We replace the field operations in four of the implementations with CryptOpt-optimized assembly code. For secp256k1, we use the Fiat+CryptOpt code (CS1), and additionally, we generate a the input for CryptOpt from the C-code of `libsecp256k1` and optimize it (CS2).

Table 5. Performance of Scalar Multiplication (Geometric Mean).

Curve25519		
Implementation	Lang	Cycles
sandy2x [Chou 2015]	asm-v	496k
amd64-64 [Chen et al. 2014]	asm	556k
amd64-51 [Chen et al. 2014]	asm	563k
donna [Langley 2022]	asm-v	988k
donna-c64 [Langley 2022]	C	588k
O’SSL [OpenSSL 2022]	C	544k
O’SSL fe-51 [OpenSSL 2022]	asm	543k
O’SSL fe-51+**CryptOpt**	asm	539k
O’SSL fe-64 [OpenSSL 2022]	asm	463k
O’SSL fe-64+**CryptOpt**	asm	478k
HACL* fe-64 [HACL 2022]	asm	461k
secp256k1		
Implementation	Lang	Cycles
libsecp256k1 [Bitcoin Core 2022]	asm	561k
libsecp256k1 [Bitcoin Core 2022]	C	545k
libsecp256k1+**CryptOpt** (CS1)	asm	545k
libsecp256k1+**CryptOpt** (CS2)	asm	538k

Evaluation. We use the SUPERCOP benchmark framework [Bernstein and Lange 2022] to measure the performance of the evaluated implementations. For each implementation, SUPERCOP tries multiple combinations of compilers and compiler options and reports the execution time of the fastest implementation. We evaluate each implementation on the eight (c.f. Table 1) and report the geometric mean (rounded to the nearest 1000 cycles) in Table 5. (See Table 7 in Appendix D for the full details.)

Results. Comparing CryptOpt with the similarly structured hand-optimized implementations of OpenSSL fe-51 and fe-64, we find that the performance with and without CryptOpt is similar. On average, CryptOpt generates slightly faster implementations for fe-51 and slightly slower implementations for fe-64. Manual optimization of code requires significant expertise and a large time investment, which needs to be repeated for each finite field. In contrast, using CryptOpt is fairly straightforward and only requires moderate computing resources to achieve similar results. CryptOpt also underperforms highly optimized implementations that use a different API (HACL*), which we do not support yet.

For secp256k1, we beat the performance of the hand-tuned assembly, matching the performance of C compiled code with our Case Study 1. That is, we only improve by providing verified formal correctness. In Case Study 2, we achieve higher performance than both the state-of-the-art and our verified implementation.

We leave the tasks of verifying the fastest implementation for secp256k1 and emitting field operations similar to the HACL* API to future work.

CryptOpt on new hardware. When looking at the results on specific machines (Table 7), we see that CryptOpt excels on the i7 11G and i9 12G platforms, providing the overall fastest implementations for fe-64 based field operations. On these platforms, CryptOpt-based implementations of Curve25519 and secp256k1 are the fastest, outperforming hand-optimized implementations, including those that use advanced processor features, such as vector instructions. The 12th generation of Intel processors is a major update of the microarchitecture. We believe that CryptOpt's automated search allows it to exploit the benefits of the new design automatically. In contrast, prior implementations and mainstream compilers need to change to adapt to these new features. We anticipate that in due course, implementations will be adapted to the new design, and hand-tuned implementations will outperform CryptOpt. However, CryptOpt does not require manual effort to adapt to new designs.

6 RELATED WORK

CryptOpt applies genetic improvement (GI), an area within search-based software engineering [Harman and Jones 2001] that automatically searches out improved software versions. Genetic improvement is a relatively young research field; its first survey appeared in 2018 [Petke et al. 2018]. Despite its youth, GI has already had real-world impact: maintainers have accepted GI patches into both open-source [Langdon et al. 2015] and commercial [Haraldsson et al. 2017] projects.

CryptOpt utilizes GI in the code-generation phase to generate fastest code per-microarchitecture. To the best of our knowledge, CryptOpt is the first automatic compiler to offer both this level of microarchitecture-tuned performance (albeit for the limited domain of straightline crypto code) and the highest level of formal assurance (Coq verification of all compiler phases that matter for soundness). However, related work has tackled some of the constituent challenges.

Bosamiya et al. [2020] is most similar to CryptOpt as they also use GI to find fast code per architecture while being provably correct. The primary objective is to parse optimized x86-64 assembly and then use verified transformations to transform it back into a clean form, which is then easier to reason about. From those transformations, they selected the "prefetch insertion" and "instruction-reordering" transformations and conducted a case study on using GI to find fast implementations based on OpenSSL's AES-GCM (which uses AES-NI instructions). They showed that they can generate verified correct per-architecture optimized code. In contrast to CryptOpt, they rely on an initial, cleverly optimized assembly implementation, which they then use as a base to alter the order of instructions, whereas CryptOpt generates the code itself and can thus take advantage of microarchitectural resources, with optimization of register allocation (in particular spills to memory) and instruction selection. For example, using their reordering transformation, they cannot substitute an add-using-overflow adox for an add-with-carry adcx, let alone have those calculate two independent additions in parallel, because they only consider reads and writes to the flag register in general, rather than the granularity of individual flags (i.e. read CF, write OF are independent). CryptOpt also benefits from compatibility with Fiat Cryptography, which makes code generation for finite fields for new primes easy.

Optimization-pass finding. Stephenson et al. [2003] and Peeler et al. [2022] both use GI to select and order existing optimization passes for optimal running time, where the former uses a simulated running time (Trimaran [Chakrapani et al. 2005]) as the objective function and the latter uses the actual running time. CryptOpt, however, is not bound by either applying or not applying those fixed optimization passes from off-the-shelf compilers. Rather, it explores many different variations for any particular code section and is also able to apply optimizations selectively at certain locations and avoid using them at others.

Superoptimization. Massalin [1987] coined the term "superoptimizer" to describe a tool for exhaustive enumeration of all possible programs to implement a given function. The key idea making this feasible is the use of a probabilistic test set, which rejects the majority of incorrect candidates. At the time of writing, it is able to generate programs of 12 instructions after several hours of running (on a 16MHz 68020 computer). Sasnauskas et al. [2017] present Souper, a tool to synthesize new optimizations on the LLVM IR. Working on the IR, by design, they are unable to generate optimizations to exploit target-specific code sequences.

Joshi et al. [2002] present Denali, a superoptimizer for very short programs. They model the architecture of their processor (Alpha EV6) and use solvers to reject conjectures of the form "No program can compute P in at most 8 cycles." Combining this with a binary-search algorithm, they end up with the most efficient program. Schkufza et al. [2013] present STOKE, a superoptimizer which is able to synthesize and optimize programs. It combines correctness indicator and performance into a cost function and then randomly (1) changes opcodes, (2) changes arguments, (3) deletes instructions, and (4) inserts nops. By starting from scratch, they can find algorithmically different solutions, which cannot be found by other superoptimizers. The resulting programs range up to tens of instructions long. Subsequent work extends STOKE to optimize floating-point kernels [Schkufza et al. 2014], optimize loops [Sharma et al. 2013], and more aggressively optimize kernels based on verified runtime preconditions with cSTOKE [Sharma et al. 2015].

However, we believe that STOKE targets a different objective than CryptOpt: finding algorithmically new implementations for very small kernels applicable for a wide range of CPUs. With CryptOpt, we generate code for bigger functions with larger inputs and target-specific microarchitectures and exploit intrinsic effects. Additionally, CryptOpt does not require modeling of the target architecture to estimate instruction costs, and can thus be "just run" on the newest architecture, without the need for complex reverse-engineering and potentially error-prone modeling.

Peephole optimization. Peephole optimizers use a sliding window on instructions (the peephole) and replace sets of instructions with more performant instructions [Aho et al. 1986; Bergmann 2003; Cooper and Torczon 2012]. The replacement is usually done based on a predefined rule set (applying only to short instruction sequences), which itself is based on heuristics for estimating which set of instructions is more performant or shorter. Yet another approach is to find and learn good peephole optimizations automatically: Bansal and Aiken [2006] use machine-learning techniques to characterize small sections of code. Then, based on those characteristics, they replace a code sequence with a semantically equivalent one assumed to be more performant. While they only focus on small sections (on the order of tens of instructions), Pekhimenko and Brown [2010] use machine-learning techniques to characterize entire methods and then apply certain optimization transformations to them. Similarly, CryptOpt considers the entire function as a whole, but rather than characterizing, learning and applying that knowledge to new functions, CryptOpt considers each architecture and function as a black box and eventually finds a fast implementation.

Verified transformations. Instead of proving the correctness of the compiler, translation validation [Pnueli et al. 1998] does not trust the compiler, but verifies that the compiled code preserves the semantics of the source. Bosamiya et al. [2020], as already mentioned, used their tool to transform optimized (manually written) assembly code to easily verifiable code. Similarly, TInA [Recoules et al. 2019] lifts inline assembly to semantically equivalent C code amenable to verification with known tools. Only targeting the code for Curve25519, Schoolderman et al. [2021] used the Why3 proving platform [Filliâtre and Paskevich 2013] to verify an 8-bit AVR implementation.

CryptoLine [Chen et al. 2014; Fu et al. 2019; Polyakov et al. 2018; Tsai et al. 2017] is an automatic verification engine utilizing SMT solvers (BOOLECTOR) and computer-algebra systems (Singular), applying to their own IR. The approach to validating assembly files is similar to ours. CryptoLine has only worked via unverified translators from assembly languages to their IR, and the translator must be trusted, unlike ours, though it likely accepts some correct programs that ours rejects.

Last, Sewell et al. [2013] go further and parse the binary code of the seL4-Linux microkernel and transform it until they could prove equivalence to the already-verified C code.

We would like to emphasize that those works aim to verify *existing* implementations, whereas CryptOpt *generates* them. Targeting a wide range of microarchitectures for performance optimizations manually would quickly become practically infeasible.

Real-world applications of computer-aided cryptography. Provably correct generated code is already deployed in important projects: all major web browsers use finite-field code generated by Fiat Cryptography [Erbsen et al. 2019] (via Google's BoringSSL and other libraries), and Firefox includes routines arising from the more comprehensive efforts of Project Everest [Bhargavan et al. 2017b], including compilation of nonstraightline code to C [Protzenko et al. 2017], verified metaprogramming of assembly [Fromherz et al. 2019], tying it together in the EverCrypt library [Protzenko et al. 2020], and even adding protocol verification [Bhargavan et al. 2017a]. The Everest stack supports many different algorithms for the same functionality (e.g. AES+GCM or ChaCha20+Poly1305 for authenticated encryption) and for each of those many different hand-optimized implementations depending on platform and hardware. We already mentioned the work of Bosamiya et al. [2020] on automatic program search, the only approach to *automatic* assembly generation that we have seen in the Everest ecosystem, and it does not seem to have been applied yet to elliptic curves. CryptOpt also has the usual advantage of proof-assistant work, that, despite our usage of a stack of domain-specific tools, none of them need be trusted.

Belyavsky et al. [2020] published work on generating prime-agnostic point arithmetic in C using verified field arithmetic from Fiat Cryptography and claim timing-side-channel-resistant code; however, there is no formal verification of correctness or constant time.

7 CONCLUSION

We presented CryptOpt, a tool that brings a perhaps-surprising confluence of improving performance and increasing formal assurance. It tackles the distinctive simplifications and complexities of straightline cryptographic code. We showed empirically that certain simplifications to established techniques suffice to set new performance records for important routines on some relevant platforms. In *generation* of fast code, we developed a simple set of transformation operators that make genetic search effective. In *checking* of fast code with foundational mechanized proofs, we followed SMT solvers and symbolic-execution engines, while avoiding their most complex aspects, like arithmetic decision procedures or nontrivial pointer-aliasing checks. We hope that these techniques can be generalized to other domains of compilation.

ACKNOWLEDGMENTS

This research was supported by the Air Force Office of Scientific Research (AFOSR) under award number FA9550-20-1-0425; the ARC Discovery Early Career Researcher Award DE200101577; the ARC Discovery Projects DP200102364 and DP210102670; the Blavatnik ICRC at Tel-Aviv University; the National Science Foundation under grants CNS-1954712 and CNS-2130671; the National Science Foundation Expedition on the Science of Deep Specification (award CCF-1521584); the Phoenix HPC service at the University of Adelaide; and gifts from Amazon Web Services, AMD, Facebook, Google, Intel and the Tezos Foundation.

REFERENCES

0xADE1A1DE. 2022. AssemblyLine v1.3.1. https://github.com/0xADE1A1DE/AssemblyLine 26

Alfred V. Aho, Ravi Sethi, and Jeffrey D. Ullman. 1986. *Compilers: Principles, Techniques, and Tools.* Addison-Wesley. 2, 19

Alaa R Alameldeen and David A Wood. 2003. Variability in architectural simulations of multi-threaded workloads. In *HPCA*. 7–18. 25

José Bacelar Almeida, Manuel Barbosa, Gilles Barthe, François Dupressoir, and Michael Emmi. 2016. Verifying Constant-Time Implementations. In *USENIX Security*. 53–70. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/almeida 5

Michaël Armand, Germain Faure, Benjamin Grégoire, Chantal Keller, Laurent Théry, and Benjamin Werner. 2011. A Modular Integration of SAT/SMT Solvers to Coq through Proof Witnesses. In *CPP*. 135–150. 11

Anne Auger and Benjamin Doerr (Eds.). 2011. *Theory of Randomized Search Heuristics: Foundations and Recent Developments*. Series on Theoretical Computer Science, Vol. 1. World Scientific. 7

Reza Azarderakhsh, Matthew Campagna, Craig Costello, Luca De Feo, Basil Hess, Amir Jalali, Brian Koziel, Brian LaMacchia, Patrick Longa, Michael Naehrig, Geovandro Pereira, Joost Renes, Vladimir Soukharev, and David Urbanik. 2019. Supersingular Isogeny Key Encapsulation – Submission to the NIST Post-Quantum Standardization Project, round 2. https://sike.org 16

Sorav Bansal and Alex Aiken. 2006. Automatic generation of peephole superoptimizers. In *ASPLOS*. 394–403. 19

Gilles Barthe, Benjamin Grégoire, and Vincent Laporte. 2018. Secure Compilation of Side-Channel Countermeasures: The Case of Cryptographic "Constant-Time". In *CSF*. 328–343. 2

Dmitry Belyavsky, Billy Bob Brumley, Jesús-Javier Chi-Domínguez, Luis Rivera-Zamarripa, and Igor Ustinov. 2020. Set It and Forget It! Turnkey ECC for Instant Integration. In *ACSAC*. 760–771. 20

Seth D. Bergmann. 2003. Compilers. In *Encyclopedia of Information Systems*. 141–170. 19

Daniel J. Bernstein. 2005. The Poly1305-AES Message-Authentication Code. In *FSE*. 32–49. 16

Daniel J. Bernstein. 2006. Curve25519: New Diffie-Hellman Speed Records. In *PKC*. 207–228. 16

Daniel J. Bernstein, Tung Chou, and Peter Schwabe. 2013. McBits: Fast Constant-Time Code-Based Cryptography. In *CHES*, Vol. 8086. 250–272. 2

Daniel J. Bernstein, Chitchanok Chuengsatiansup, and Tanja Lange. 2014a. Curve41417: Karatsuba Revisited. In *CHES*. 316–334. 2

Daniel J. Bernstein, Chitchanok Chuengsatiansup, Tanja Lange, and Peter Schwabe. 2014b. Kummer Strikes Back: New DH Speed Records. In *ASIACRYPT*. 317–337. 2

Daniel J. Bernstein, Chitchanok Chuengsatiansup, Tanja Lange, and Christine van Vredendaal. 2017. NTRU Prime: Reducing Attack Surface at Low Cost. In *SAC*. 235–260. 2

Daniel J. Bernstein and Tanja Lange. 2022. eBACS: ECRYPT benchmarking of cryptographic systems. http://bench.cr.yp.to/supercop/supercop-20221005.tar.xz 17

Karthikeyan Bhargavan, Bruno Blanchet, and Nadim Kobeissi. 2017a. Verified Models and Reference Implementations for the TLS 1.3 Standard Candidate. In *IEEE SP*. 483–502. 20

Karthikeyan Bhargavan, Barry Bond, Antoine Delignat-Lavaud, Cédric Fournet, Chris Hawblitzel, Catalin Hritcu, Samin Ishtiaq, Markulf Kohlweiss, Rustan Leino, Jay Lorch, Kenji Maillard, Jianyang Pang, Bryan Parno, Jonathan Protzenko, Tahina Ramananandro, Ashay Rane, Aseem Rastogi, Nikhil Swamy, Laure Thompson, Peng Wang, Santiago Zanella-Béguelin, and Jean-Karim Zinzindohoué. 2017b. Everest: Towards a Verified and Drop-in Replacement of HTTPS. In *Proc. SNAPL*. https://project-everest.github.io/assets/snapl2017.pdf 20

Bitcoin Core. 2022. libsecp256k1 - Optimized C Library for ECDSA Signatures and Secret/Public Key Operations on Curve secp256k1. https://github.com/bitcoin-core/secp256k1 3, 17, 30

Mahmoud A. Bokhari, Brad Alexander, and Markus Wagner. 2020. Towards rigorous validation of energy optimisation experiments. In *GECCO*. 1232–1240. 25

Mahmoud A Bokhari, Lujun Weng, Markus Wagner, and Bradley Alexander. 2019. Mind the gap - a distributed framework for enabling energy optimisation on modern smart-phones in the presence of noise, drift, and statistical insignificance. In *IEEE CEC*. 1330–1337. 25

Jay Bosamiya, Sydney Gibson, Yao Li, Bryan Parno, and Chris Hawblitzel. 2020. Verified Transformations and Hoare Logic: Beautiful Proofs for Ugly Assembly Language. In *VSTTE*. 106–123. 3, 18, 20

Christopher Celio, Palmer Dabbelt, David A. Patterson, and Krste Asanovic. 2016. The Renewed Case for the Reduced Instruction Set Computer: Avoiding ISA Bloat with Macro-Op Fusion for RISC-V. arXiv 1607.02318. 4

Certicom Research. 2000. SEC 2: Recommended elliptic curve domain parameters, version 1.0. http://www.secg.org/SEC2-Ver-1.0.pdf. 16

Lakshmi N. Chakrapani, John Gyllenhaal, Wen-mei W. Hwu, Scott A. Mahlke, Krishna V. Palem, and Rodric M. Rabbah. 2005. Trimaran: An Infrastructure for Research in Instruction-Level Parallelism. In *Languages and Compilers for High Performance Computing*. 32–41. 19

Yu-Fang Chen, Chang-Hong Hsu, Hsin-Hung Lin, Peter Schwabe, Ming-Hsien Tsai, Bow-Yaw Wang, Bo-Yin Yang, and Shang-Yi Yang. 2014. Verifying Curve25519 Software. In *CCS*. 299–309. 17, 20, 30

Tung Chou. 2015. Sandy2x: New Curve25519 Speed Records. In *SAC*. 145–160. 2, 17, 30

Tung Chou. 2016. QcBits: Constant-Time Small-Key Code-Based Cryptography. In *CHES*, Vol. 9813. 280–300. 2

Chitchanok Chuengsatiansup, Michael Naehrig, Pance Ribarski, and Peter Schwabe. 2013. PandA: Pairings and Arithmetic. In *Pairing*, Vol. 8365. 229–250. 2

Chitchanok Chuengsatiansup and Damien Stehlé. 2019. Towards Practical GGM-Based PRF from (Module-) Learning-with-Rounding. In *SAC*. 693–713. 2

Clang. 2022. Clang: a C language family frontend for LLVM. https://clang.llvm.org 16

Keith D. Cooper and Linda Torczon. 2012. Chapter 11 - Instruction Selection. In *Engineering a Compiler (Second Edition)*. 597–638. 19

Charlie Curtsinger and Emery D. Berger. 2013. STABILIZER: statistically sound performance evaluation. In *ASPLOS*. 219–228. 25

Leonardo Mendonça de Moura and Nikolaj S. Bjørner. 2008. Z3: An Efficient SMT Solver. In *TACAS*. 337–340. 11

David Detlefs, Greg Nelson, and James B. Saxe. 2005. Simplify: a theorem prover for program checking. *J. ACM* 52, 3 (2005), 365–473. 12

Benjamin Doerr and Frank Neumann. 2019. *Theory of evolutionary computation: Recent developments in discrete optimization*. Springer. 7

Vijay D'Silva, Mathias Payer, and Dawn Xiaodong Song. 2015. The Correctness-Security Gap in Compiler Optimization. In *IEEE SP Workshops*. 73–87. 2

Andres Erbsen, Jade Philipoom, Jason Gross, Robert Sloan, and Adam Chlipala. 2019. Simple High-Level Code for Cryptographic Arithmetic - With Proofs, Without Compromises. In *IEEE SP*. 1202–1219. 2, 4, 16, 20

Jean-Christophe Filliâtre and Andrei Paskevich. 2013. Why3 - Where Programs Meet Provers. In *ESOP*. 125–128. 20

Matteo Fischetti and Michele Monaci. 2014. Exploiting Erraticism in Search. *Operations Research* 62, 1 (2014), 114–122. 7

Aymeric Fromherz, Nick Giannarakis, Chris Hawblitzel, Bryan Parno, Aseem Rastogi, and Nikhil Swamy. 2019. A verified, efficient embedding of a verifiable assembly language. In *POPL*. 63:1–63:30. 20

Yu-Fu Fu, Jiaxiang Liu, Xiaomu Shi, Ming-Hsien Tsai, Bow-Yaw Wang, and Bo-Yin Yang. 2019. Signed Cryptographic Program Verification with Typed CryptoLine. In *CCS*. ACM, 1591–1606. 20

GCC. 2022. GCC, the GNU Compiler Collection. https://gcc.gnu.org 16

HACL. 2022. HACL. https://github.com/hacl-star/hacl-star 17, 30

Mike Hamburg. 2015. Ed448-Goldilocks, a new elliptic curve. *IACR Cryptology ePrint Archive* 2015 (2015), 625. 16

Saemundur O. Haraldsson, John R. Woodward, Alexander E. I. Brownlee, and Kristin Siggeirsdottir. 2017. Fixing bugs in your sleep: how genetic improvement became an overnight success. In *GECCO (Companion)*. 1513–1520. 18

Mark Harman and Bryan F. Jones. 2001. Software engineering using metaheuristic innovative algorithms: workshop report. *Inf. Softw. Technol.* 43, 14 (2001), 905–907. 18

Rodney E. Hooker and Collin Eddy. 2013. Store-to-load forwarding based on load/store address computation source information comparisons. US Patent 8533438. 4

Rajeev Joshi, Greg Nelson, and Keith H. Randall. 2002. Denali: A Goal-directed Superoptimizer. In *PLDI*. ACM, 304–314. 19

Tomas Kalibera, Lubomir Bulej, and Petr Tuma. 2005a. Automated detection of performance regressions: The Mono experience. In *13th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*. 183–190. 25

Tomas Kalibera, Lubomir Bulej, and Petr Tuma. 2005b. Benchmark precision and random initial state. In *2005 International Symposium on Performance Evaluation of Computer and Telecommunications Systems (SPECTS)*. 853–862. 25

Tomas Kalibera and Richard Jones. 2013. Rigorous benchmarking in reasonable time. *ACM SIGPLAN Notices* 48, 11 (2013), 63–74. 25

Matthias J. Kannwischer, Joost Rijneveld, and Peter Schwabe. 2019. Faster Multiplication in $\mathbb{Z}_{2^m}[x]$ on Cortex-M4 to Speed up NIST PQC Candidates. In *ACNS*. 281–301. 2

Thierry Kaufmann, Hervé Pelletier, Serge Vaudenay, and Karine Villegas. 2016. When Constant-Time Source Yields Variable-Time Binary: Exploiting Curve25519-donna Built with MSVC 2015. In *CANS*. 573–582. 2

William B. Langdon, Brian Yee Hong Lam, Justyna Petke, and Mark Harman. 2015. Improving CUDA DNA Analysis Software with Genetic Programming. In *GECCO*. 1063–1070. 18

Adam Langley. 2022. Curve25519-donna. https://github.com/agl/curve25519-donna 17, 30

Kevin M. Lepak and Mikko H. Lipasti. 2000. On the value locality of store instructions. In *ISCA*. 182–191. 4

Xavier Leroy. 2009. A Formally Verified Compiler Back-end. *J. Autom. Reason.* 43, 4 (2009), 363–446. 11

Xavier Leroy, Sandrine Blazy, Daniel Kästner, Bernhard Schommer, Markus Pister, and Christian Ferdinand. 2016. CompCert — A Formally Verified Optimizing Compiler. In *ERTS*. 11

Nuno P. Lopes, Juneyoung Lee, Chung-Kil Hur, Zhengyang Liu, and John Regehr. 2021. Alive2: bounded translation validation for LLVM. In *PLDI*. 65–79. 10, 11

Henry Massalin. 1987. Superoptimizer - A Look at the Smallest Program. In *ASPLOS*. ACM Press, 122–126. 19

Todd Mytkowicz, Amer Diwan, Matthias Hauswirth, and Peter F Sweeney. 2009. Producing wrong data without doing anything obviously wrong! *ACM SIGARCH Computer Architecture News* 37, 1 (2009), 265–276. 25, 27

NIST. 2000. FIPS PUB 186-2: Digital signature standard. 16

OpenSSL. 2022. OpenSSL. https://www.openssl.org/ 4, 17, 30

Hannah Peeler, Shuyue Stella Li, Andrew N. Sloss, Kenneth N. Reid, Yuan Yuan, and Wolfgang Banzhaf. 2022. Optimizing LLVM Pass Sequences with Shackleton: A Linear Genetic Programming Framework. arXiv 2201.13305. 19

Gennady Pekhimenko and Angela Demke Brown. 2010. Efficient Program Compilation Through Machine Learning Techniques. In *Software Automatic Tuning, From Concepts to State-of-the-Art Results*. Springer, 335–351. 19

Justyna Petke, Saemundur O. Haraldsson, Mark Harman, William B. Langdon, David Robert White, and John R. Woodward. 2018. Genetic Improvement of Software: A Comprehensive Survey. *IEEE Trans. Evol. Comput.* 22, 3 (2018), 415–432. 18

Amir Pnueli, Michael Siegel, and Eli Singerman. 1998. Translation Validation. In *TACAS*. 151–166. 20

Andy Polyakov, Ming-Hsien Tsai, Bow-Yaw Wang, and Bo-Yin Yang. 2018. Verifying Arithmetic Assembly Programs in Cryptographic Primitives (Invited Talk). In *CONCUR (LIPIcs, Vol. 118)*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 4:1–4:16. 20

Jonathan Protzenko, Bryan Parno, Aymeric Fromherz, Chris Hawblitzel, Marina Polubelova, Karthikeyan Bhargavan, Benjamin Beurdouche, Joonwon Choi, Antoine Delignat-Lavaud, Cédric Fournet, Natalia Kulatova, Tahina Ramananandro, Aseem Rastogi, Nikhil Swamy, Christoph M. Wintersteiger, and Santiago Zanella Béguelin. 2020. EverCrypt: A Fast, Verified, Cross-Platform Cryptographic Provider. In *IEEE SP*. 983–1002. 20

Jonathan Protzenko, Jean Karim Zinzindohoué, Aseem Rastogi, Tahina Ramananandro, Peng Wang, Santiago Zanella Béguelin, Antoine Delignat-Lavaud, Catalin Hritcu, Karthikeyan Bhargavan, Cédric Fournet, and Nikhil Swamy. 2017. Verified low-level programming embedded in F*. *Proceedings of the ACM on Programming Languages* 1 (2017), 1 – 29. 20

Kishore Kumar Pusukuri, Rajiv Gupta, and Laxmi N Bhuyan. 2012. Thread tranquilizer: Dynamically reducing performance variation. *ACM TACO* 8, 4 (2012), 46. 25

Frédéric Recoules, Sébastien Bardin, Richard Bonichon, Laurent Mounier, and Marie-Laure Potet. 2019. Get Rid of Inline Assembly through Verification-Oriented Lifting. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. 577–589. https://doi.org/10.1109/ASE.2019.00060 20

John C. Reynolds. 2002. Separation Logic: A Logic for Shared Mutable Data Structures. In *LICS*. 14

Ronny Ronen, Alexander Peleg, and Nathaniel Hoffman. 2004. System and method for fusing instructions. US Patent 6675376B2. 4

Raimondas Sasnauskas, Yang Chen, Peter Collingbourne, Jeroen Ketema, Jubi Taneja, and John Regehr. 2017. Souper: A Synthesizing Superoptimizer. *CoRR* abs/1711.04422 (2017). 19

Eric Schkufza, Rahul Sharma, and Alex Aiken. 2013. Stochastic superoptimization. In *ASPLOS*. ACM, 305–316. 19

Eric Schkufza, Rahul Sharma, and Alex Aiken. 2014. Stochastic optimization of floating-point programs with tunable precision. In *PLDI*. ACM, 53–64. 19

Marc Schoolderman, Jonathan Moerman, Sjaak Smetsers, and Marko C. J. D. van Eekelen. 2021. Efficient Verification of Optimized Code - Correct High-Speed X25519. In *NFM*. 304–321. 20

Thomas Arthur Leck Sewell, Magnus O. Myreen, and Gerwin Klein. 2013. Translation validation for a verified OS kernel. In *PLDI*. 471–482. 20

Rahul Sharma, Eric Schkufza, Berkeley R. Churchill, and Alex Aiken. 2013. Data-driven equivalence checking. In *OOPSLA*. ACM, 391–406. 19

Rahul Sharma, Eric Schkufza, Berkeley R. Churchill, and Alex Aiken. 2015. Conditionally correct superoptimization. In *OOPSLA*. ACM, 147–162. 19

Mark Stephenson, Una-May O'Reilly, Martin C. Martin, and Saman P. Amarasinghe. 2003. Genetic Programming Applied to Compiler Heuristic Optimization. In *EuroGP*. 238–253. 19

Samantika Subramaniam and Gabriel H. Loh. 2006. Fire-and-Forget: Load/Store Scheduling with No Store Queue at All. In *MICRO*. 273–284. 4

Jean-Baptiste Tristan and Xavier Leroy. 2008. Formal verification of translation validators: a case study on instruction scheduling optimizations. In *POPL*. 17–27. 11

Ming-Hsien Tsai, Bow-Yaw Wang, and Bo-Yin Yang. 2017. Certified Verification of Algebraic Properties on Low-Level Mathematical Constructs in Cryptographic Programs. In *CCS*. ACM, 1973–1987. 20

Pepe Vila, Pierre Ganty, Marco Guarnieri, and Boris Köpf. 2020. CacheQuery: learning replacement policies from hardware caches. In *PLDI*. 519–532. 4

Thomas Weise, Zijun Wu, and Markus Wagner. 2019. An Improved Generic Bet-and-Run Strategy with Performance Prediction for Stochastic Local Search. In *AAAI*. 2395–2402. 7

A ON RELIABLE PERFORMANCE MEASUREMENT

A.1 Background

The execution time of a program can be affected by a myriad of factors, such as, the state of cache memory [Alameldeen and Wood 2003]; context switches [Pusukuri et al. 2012]; memory layout and OS environment variables [Curtsinger and Berger 2013; Mytkowicz et al. 2009]. Even when a platform is rebooted between runs there can be significant variation in benchmark run times [Kalibera et al. 2005a]. Furthermore, measurements of the benchmark operation can be subject to inherent and external random changes to the benchmark operation, as well as the system state prior to execution [Kalibera et al. 2005b; Kalibera and Jones 2013]. In addition, sensor readings themselves can drift over time [Bokhari et al. 2019]. Consequently, all these factors together result in measurement noise, which can affect both the sensitivity and specificity of performance measurements. One recent work, which surveyed and compared validation approaches in the context of energy consumption optimization [Bokhari et al. 2020], found that a number of measurement approaches failed to mitigate the effects of measurement noise. Only the proposed approach R3-validation was not misled: it exercised the tests in a rotated-round-robin fashion, while accommodating the necessary regular restarting and recharging of the test platform.

A.2 Cost-Function Evaluation

Recall that CryptOpt generates solutions, mutates them, and measures their respective performance. Measuring this performance on physical machines is inherently noisy. To lower the effect of noise and enable more stable and fair comparisons, we base our measurement on R3-validation, which exercises tests in a rotated-round-robin fashion. This approach was used in previous work [Bokhari et al. 2020] in the context of energy consumption in order to mitigate the effects of measurement noise.

In our work, we adapt R3-validation in two ways. First, we forgo the restart of the computer as we do not observe any measurement drift over time. Second, we perform a random scheduling of program variants for performance measurement instead of strictly following a given order of measurements.

Measuring Performance. Algorithm 3 presents our performance-measurement routine. We now explain the rationale behind this from the inside out (inside the loop of the algorithm to the outside of the measure function).

Cycle counters are integers with a granularity of at best 1. Recall that CryptOpt can optimize functions of arbitrary size, ranging in our experiments from fewer than 100 to around 2000 instructions. That means that the accuracy of the fixed-granularity counter relative to the length-varying function changes. In other words, the same "stopwatch" is not appropriate to time an Olympic sprinter and a marathon runner.

The solution to that is to measure multiple executions of the same function instead of just one. We call multiple executions of the same function a *batch*. The batch size *bs* specifies how many executions of the same function are being measured with one measurement. It follows logically that small functions should have a bigger *bs* than long functions in order to get similar cycle accuracy.

Hence, we came up with the idea of the *cyclegoal*. That is, we want the measured cycles count for one batch across all functions to be in the order of $cyclegoal = 10\,000$ measured cycles. The reason for this particular number is that we then have four to five digits of accuracy. CryptOpt then scales the value for *bs* up or down by comparing the measured cycles against the *cyclegoal*.

This technique also mitigates another inherent problem that arises when measuring performance in cycles on hardware: different hardware platforms interpret "cycles" differently, and additional challenges arise due to different types of boosting. However, because CryptOpt adjusts *bs* in each

Algorithm 3: Execution Time Measurement

```
input  : A pointer to code A,
         B pointer to code B,
         bs size of a measurement batch,
         nob number of batches to measure
output : PA performance of code A in cycles,
         PB performance of code B in cycles

function measure(A, B, bs, nob)
begin
    /* initialize cycle lists                              */
    $cycles_A \leftarrow$ empty list
    $cycles_B \leftarrow$ empty list
    repeat
        /* f points to either A or B                       */
        $f \leftarrow \mathrm{randomSelect}(A,\ B)$

        /* run bs times and get elapsed cycles             */
        $cycles \leftarrow \mathrm{countCyclesForNRuns}(bs, f)$

        /* append elapsed cycles to list A/B resp.         */
        $\mathrm{append}(cycles_f, cycles)$
    until both have been measured nob times;

    $PA \leftarrow \mathrm{median}(cycles_A)$
    $PB \leftarrow \mathrm{median}(cycles_B)$

    return PA, PB
end
```

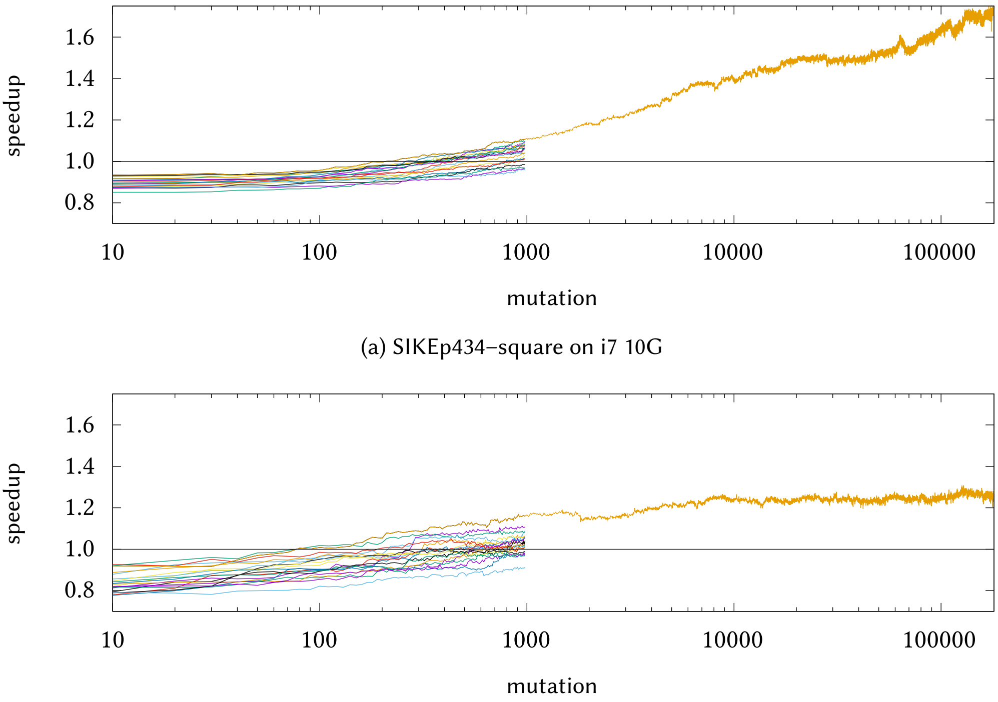
evaluation of the mutation (i.e. before each call to "measure") and it does so on each platform independently, we get comparable accuracy for all functions across platforms.

Inherent with measuring time on hardware is measurement noise, which we as developers can hardly control. This noise can be due to the OS's interrupt event handlers or process scheduling. It can also be due to temperature rises the hardware limits itself. The technique above mitigates the problems having a bad "stopwatch" but at the same time increases the captured noise. Similar to the R3-validation, we mitigate this problem by simply measuring each batch multiple times, called *numbers of batches* or *nob* for short. That means that we have *nob* measurements for the same function. We take the median as the performance measurement for a function to drop outlier measurements. (Using the minimum rather than median did not have statistically significant effects – neither positive nor negative.) Empirically, setting $nob = 31$ works well on our experimental platforms.

Complex processors nowadays can cause biases to one or another function by their speculative execution, prediction behavior and caching, just to name a few reasons. The R3-validation also mitigates this issue by using a random sequence of the functions to measure. We do this by randomly selecting either function to measure a batch of.

A.3 AssemblyLine

CryptOpt uses AssemblyLine [0xADE1A1DE 2022] to assemble the generated initial and mutated code into memory positions *A* and *B*. AssemblyLine is a lightweight in-memory assembler for



(a) SIKEp434−square on i7 10G

(b) secp256k1−square on i7 11G

Fig. 9. CryptOpt optimization progress of different method on two platforms. Each line traces an optimization run over time (X-axis, log-scale), showing the relative performance gain over Clang. Bet-and-Run used for 20 runs with a budget of 1 000 mutations each. The best-performing candidate continues for another 180 000 mutations. Progress over time is roughly logarithmic in the number of mutations, achieving a performance gain of 52–86%. We generate one data point every ten mutations.

x86-64 assembly instructions. It takes an input string of instructions, assembles them into machine code and returns a pointer to the executable code. We use AssemblyLine to eliminate the overhead of writing to the file system and invoking a typical compiler tool chain for every evaluation. Additionally, CryptOpt uses AssemblyLine to assemble the code to the start of a memory page. This reduces noise in measurements because it always assembles aligned code for all functions, which, in turn, reduces memory-biased performance impacts [Mytkowicz et al. 2009]. Once assembled, CryptOpt calls the measure function with pointers to that code and values for batch size *bs* and number of batches *nob*.

B OPTIMIZATION PROGRESS

The first evaluation question we answer is how the optimization progresses over time. Figure 9 shows examples of optimization runs on i7 10G and i7 11G, targeting the field-multiplication operation corresponding to the prime field for SIKEp434 and secp256k1 respectively.

The figures show performance over time, where performance is measured as gain over the baseline Clang-compiled code, while optimization time is measured in mutations (shown in log scale).

The figure shows the Bet-and-Run strategy in action. The optimization starts with 20 runs, each with a budget of 1 000 mutations. CryptOpt then selects the best-performing and continues optimizing it for a further 180 000 mutations. As the figure shows, in all runs, optimization progress is roughly logarithmic in the number of mutations.

A notable difference between the two optimizations is the overall performance gain. On the i7 11G platform the optimization converges rather quickly, whereas on i7 10G we see monotonic improvement until the very end. Finally, we note that these variations and performance gain are dependent on field, method and machine; here we see a gain 52–86%.

C PLATFORM-SPECIFIC OPTIMIZATION

CryptOpt optimizes the execution time of the function on the platform it executes on. Because different platforms have different hardware components, the fastest code on one platform is not necessarily the fastest on another. Thus, in this section we ask whether CryptOpt overfits to the platform it executes on, as opposed to universally optimizing for all platforms. That is, we test how *native optimization*, where the code executes on the same platform it was optimized on, compares with *cross optimization*, where code is optimized on one platform and executes on another.

To test this, we first optimize each of the functions on each of the platforms. We then measure the performance of the produced code from all platforms on all platforms (as per evaluation described in Appendix A). Table 6 shows the results for the operations multiply and square on all nine fields we tested. Rows correspond to the platform the code was optimized on and columns to the platform the code is executed on. The value in a cell shows the ratio between the execution time of the cross-optimized code and that of the native code. An example value of 1.10 indicates that the cross-platform code (from the "row" platform) needs 1.10 times as many cycles as the native code (from the "column" platform). In other words, a blue cell means that code from the column architecture outperforms code from the row architecture, and the number indicates by how much.

The additional column G.M. shows the geometric mean of the numbers in the row. This gives a measure of "how platform-specific" one optimization is: A value larger than one (blue) in column G.M. in the row for platform X indicates that on average, code from other platforms outperform the code from X, if ran on platforms other than X. In turn, a value smaller than one means that the code from platform X tends to outperform platform-specific code on other platforms. Please note that a "large number" can arise for two reasons: (a) The platform-specific code is very bad (thus easy to outperform), or (b) the generated code is very generic and can thus compete with the platform-specific ones. To see which is applicable per platform, one needs to see how well the code compares to off-the-shelf compilers. Note that this is also not a universal measure, as off-the-shelf compilers generate different code per platform. Alternatively, running the same code on all platforms would also not give a fair baseline, as compilers would not be enabled to generate platform-specific code, whereas CryptOpt would be.

As can be seen in Table 6 for Curve25519, optimizing and running on the same architecture tends to outperform cross-architecture optimizations (which would mean the 8-by-8 is mostly blue, and the G.M. column is fully blue, too). However, there are cases where cross-architectural optimizations are better. In particular it appears that optimization of SIKEp434 functions on AMD is not ideal, and the results underperform code optimized on Intel processors. (Still, faster than compiling with off-the-shelf compilers).

Another effect that the table highlights is that optimization overfits for processor families, for some families more than others. In particular, Intel 6th and 10th generations show little differences in the respective inter-platform performance. For example, the related 3x3 sub-matrices for Curve25519 in Table 6 have values in the small interval [0.99, 1.05] (mul), [0.90, 1.05] (square);

Table 6. Optimization results. We show the relative improvements in % for the multiplication (top) and squaring (bottom) operations; time savings are marked in blue. First, to observe hardware-specific optimization, the 8-by-8 matrix shows the performance the optimized operation that have been optimized on one machine and then run on another. The subsequent two rows (Clang/GCC) then show the time savings of our optimized operations over off-the-shelf-compilers. Lastly, "Final" shows the time savings of our best-performing implementation over the best-performing compiler-generated version.

Curve25519

run on	1900X	5800X	5950X	i7 6G	i7 10G	i9 10G	i7 11G	i9 12G	G.M.
opt on									
1900X		1.09	1.07	1.12	1.13	1.24	1.19	1.21	1.13
5800X	1.11		0.98	1.14	1.15	1.15	1.15	1.20	1.11
5950X	1.12	1.01		1.12	1.16	1.18	1.14	1.22	1.12
i7 6G	1.06	1.16	1.13		1.01	1.00	1.05	1.15	1.07
i7 10G	1.07	1.12	1.10	1.05		1.00	1.09	1.18	1.07
i9 10G	1.07	1.13	1.17	0.99	1.00		1.14	1.29	1.09
i7 11G	1.08	1.11	1.09	1.10	1.20	1.12		1.20	1.11
i9 12G	1.03	1.03	1.01	1.04	1.05	1.12	1.04		1.04
Clang	1.14	1.16	1.13	1.19	1.20	1.19	1.22	1.27	1.19
GCC	1.04	1.12	1.10	1.13	1.14	1.15	1.18	1.28	1.14
Final	1.04	1.12	1.12	1.15	1.14	1.15	1.18	1.27	1.14
1900X		1.09	1.11	1.15	1.04	1.20	1.20	1.31	1.14
5800X	1.10		1.00	1.13	1.04	1.21	1.18	1.15	1.10
5950X	1.16	1.00		1.13	1.03	1.15	1.21	1.18	1.10
i7 6G	1.25	1.17	1.18		0.91	1.02	1.19	1.30	1.12
i7 10G	1.23	1.17	1.18	1.04		1.05	1.15	1.37	1.14
i9 10G	1.20	1.17	1.15	0.99	0.90		1.11	2.39	1.18
i7 11G	1.18	1.10	1.09	1.06	0.97	1.10		1.40	1.11
i9 12G	1.06	1.05	1.04	1.03	0.94	1.05	1.09		1.03
Clang	1.18	1.12	1.11	1.15	1.04	1.16	1.14	1.27	1.14
GCC	1.15	1.14	1.13	1.18	1.08	1.21	1.26	1.31	1.18
Final	1.15	1.12	1.11	1.16	1.16	1.16	1.14	1.27	1.16

P-224

run on	1900X	5800X	5950X	i7 6G	i7 10G	i9 10G	i7 11G	i9 12G	G.M.
opt on									
1900X		1.01	1.02	1.12	1.12	1.06	1.06	1.09	1.06
5800X	1.06		1.01	1.16	1.16	1.10	1.11	1.20	1.10
5950X	1.08	1.00		1.22	1.19	1.12	1.20	1.25	1.13
i7 6G	1.06	1.04	1.03		1.01	0.95	1.04	1.05	1.02
i7 10G	1.03	1.01	1.00	1.00		0.95	1.04	1.04	1.01
i9 10G	1.10	1.04	1.04	1.07	1.06		1.34	1.84	1.16
i7 11G	1.03	1.01	1.01	1.06	1.09	1.01		1.05	1.03
i9 12G	1.04	1.02	1.02	1.05	1.08	1.00	1.04		1.03
Clang	1.28	1.26	1.26	1.35	1.35	1.28	1.26	1.43	1.31
GCC	1.71	1.56	1.56	2.06	2.06	1.96	1.91	2.22	1.87
Final	1.28	1.27	1.26	1.36	1.35	1.35	1.26	1.43	1.32
1900X		1.02	1.08	1.03	1.09	1.11	1.16	1.20	1.08
5800X	1.15		1.04	1.06	1.10	1.14	1.18	1.25	1.11
5950X	1.10	0.96		1.04	1.11	1.12	1.14	1.78	1.14
i7 6G	1.13	1.08	1.14		1.02	1.12	1.13	1.12	1.09
i7 10G	1.11	1.03	1.09	0.97		1.04	1.11	1.08	1.05
i9 10G	1.09	1.04	1.09	0.93	0.96		1.08	1.05	1.03
i7 11G	1.09	1.00	1.06	1.01	1.04	1.08		1.06	1.04
i9 12G	1.07	1.03	1.08	0.99	1.02	1.08	1.06		1.04
Clang	1.23	1.15	1.21	1.20	1.24	1.30	1.22	1.37	1.24
GCC	1.77	1.53	1.61	1.89	1.95	2.02	1.95	2.08	1.84
Final	1.23	1.20	1.21	1.29	1.29	1.30	1.22	1.37	1.26

P-256

run on	1900X	5800X	5950X	i7 6G	i7 10G	i9 10G	i7 11G	i9 12G	G.M.
opt on									
1900X		1.02	0.99	1.11	1.11	1.14	1.13	0.67	1.01
5800X	1.07		0.97	1.15	1.12	1.19	1.16	0.70	1.03
5950X	1.08	1.01		1.15	1.12	1.18	1.15	0.68	1.03
i7 6G	1.05	1.01	0.99		0.98	1.03	1.05	1.03	1.02
i7 10G	1.06	1.00	0.98	1.03		1.08	1.07	0.64	0.97
i9 10G	1.04	0.98	0.96	0.97	0.95		1.04	0.61	0.93
i7 11G	1.04	0.97	0.95	1.06	1.03	1.13		0.61	0.96
i9 12G	1.03	1.01	0.98	1.04	1.02	1.08	1.09		1.03
Clang	1.32	1.30	1.27	1.38	1.34	1.43	1.33	0.89	1.27
GCC	1.72	1.61	1.56	2.08	2.02	2.14	2.01	1.36	1.79
Final	1.32	1.34	1.34	1.42	1.42	1.43	1.33	1.47	1.38
1900X		1.08	1.06	1.02	1.12	1.08	1.15	1.11	1.08
5800X	1.06		1.00	1.00	1.12	1.06	1.10	1.80	1.12
5950X	1.06	1.01		1.05	1.11	1.08	1.12	1.09	1.06
i7 6G	1.09	1.08	1.05		1.04	1.02	1.10	1.06	1.06
i7 10G	1.04	1.04	1.02	0.91		0.97	1.07	1.04	1.01
i9 10G	1.11	1.06	1.03	0.96	1.06		1.15	1.06	1.05
i7 11G	1.06	1.04	1.02	0.97	1.08	1.03		1.76	1.10
i9 12G	1.06	1.09	1.06	1.00	1.09	1.06	1.07		1.05
Clang	1.28	1.31	1.28	1.23	1.35	1.32	1.26	1.43	1.30
GCC	1.75	1.65	1.61	1.87	2.04	1.98	1.98	1.98	1.85
Final	1.28	1.31	1.28	1.35	1.35	1.36	1.26	1.43	1.32

P-384

run on	1900X	5800X	5950X	i7 6G	i7 10G	i9 10G	i7 11G	i9 12G	G.M.
opt on									
1900X		0.96	0.96	1.59	1.60	1.64	1.02	0.97	1.18
5800X	1.25		1.00	1.82	1.86	1.88	1.10	1.08	1.32
5950X	1.16	1.00		1.85	1.83	1.87	1.12	1.06	1.31
i7 6G	0.76	0.71	0.71		0.98	1.03	0.67	0.60	0.79
i7 10G	0.76	0.71	0.70	0.99		1.02	0.67	0.60	0.79
i9 10G	0.78	0.71	0.71	0.97	0.98		0.67	0.61	0.79
i7 11G	1.04	0.99	0.98	1.68	1.68	1.74		0.99	1.22
i9 12G	1.13	1.05	1.05	1.79	1.82	1.84	1.05		1.29
Clang	0.98	0.94	0.94	1.52	1.53	1.56	0.85	0.90	1.12
GCC	1.33	1.29	1.28	2.28	2.30	2.35	1.56	1.34	1.66
Final	1.29	1.34	1.35	1.56	1.56	1.56	1.27	1.49	1.42
1900X		1.02	1.00	1.76	1.77	1.67	1.08	0.98	1.24
5800X	1.21		0.98	1.74	1.70	1.63	1.11	0.97	1.26
5950X	1.05	1.01		1.78	1.75	1.68	1.10	0.98	1.25
i7 6G	0.74	0.71	0.70		0.99	0.94	0.71	0.57	0.78
i7 10G	0.74	0.73	0.71	0.99		0.94	0.70	0.56	0.78
i9 10G	0.77	0.73	0.72	1.06	1.05		0.73	0.60	0.82
i7 11G	1.02	1.00	0.98	1.62	1.62	1.53		0.92	1.18
i9 12G	1.15	1.10	1.08	1.92	1.92	1.83	1.15		1.34
Clang	0.98	0.95	0.93	1.51	1.49	1.42	0.84	0.81	1.08
GCC	1.36	1.16	1.14	2.38	2.36	2.25	1.62	1.19	1.60
Final	1.32	1.33	1.34	1.53	1.52	1.52	1.21	1.44	1.40

SIKEp434

run on	1900X	5800X	5950X	i7 6G	i7 10G	i9 10G	i7 11G	i9 12G	G.M.
opt on									
1900X		0.99	0.95	1.95	1.83	1.71	1.27	1.01	1.28
5800X	1.32		0.96	2.02	1.91	1.79	1.37	1.04	1.37
5950X	1.39	1.03		2.04	1.92	1.81	1.40	1.05	1.40
i7 6G	0.70	0.66	0.63		0.95	0.89	0.80	0.60	0.77
i7 10G	0.72	0.67	0.64	1.04		0.94	0.80	0.60	0.79
i9 10G	0.72	0.69	0.66	1.12	1.06		0.88	0.94	0.87
i7 11G	0.84	0.83	0.80	1.42	1.36	1.27		0.77	1.00
i9 12G	1.36	1.06	1.02	1.99	1.91	1.78	1.37		1.39
Clang	1.15	1.11	1.06	1.84	1.75	1.65	1.01	1.12	1.30
GCC	1.40	1.16	1.12	2.55	2.43	2.28	2.08	1.32	1.70
Final	1.64	1.68	1.68	1.84	1.84	1.85	1.27	1.87	1.70
1900X		0.99	0.98	1.88	1.79	1.93	1.36	1.02	1.31
5800X	1.17		0.98	1.88	1.79	1.93	1.39	1.00	1.34
5950X	1.21	1.01		1.89	1.80	1.96	1.46	1.06	1.37
i7 6G	0.67	0.66	0.65		0.95	1.01	0.82	0.57	0.77
i7 10G	0.70	0.70	0.68	1.06		1.09	0.89	0.62	0.82
i9 10G	0.68	0.68	0.66	0.98	0.93		0.83	0.59	0.78
i7 11G	0.85	0.80	0.79	1.32	1.26	1.35		0.77	0.99
i9 12G	1.30	1.03	1.02	1.93	1.82	1.96	1.47		1.39
Clang	1.10	1.10	1.08	1.76	1.68	1.82	1.04	1.06	1.29
GCC	1.38	1.45	1.43	2.49	2.37	2.56	2.25	1.28	1.83
Final	1.65	1.67	1.67	1.80	1.81	1.82	1.28	1.85	1.68

Curve448

run on	1900X	5800X	5950X	i7 6G	i7 10G	i9 10G	i7 11G	i9 12G	G.M.
opt on									
1900X		1.14	1.12	1.40	1.34	1.44	1.25	1.04	1.21
5800X	0.85		1.00	1.24	1.19	1.23	1.03	1.07	1.07
5950X	0.86	1.01		1.20	1.16	1.26	1.02	1.05	1.06
i7 6G	0.78	0.97	0.95		0.97	1.03	0.94	0.89	0.94
i7 10G	0.81	0.96	0.94	1.03		1.09	0.94	0.95	0.96
i9 10G	0.78	0.94	0.92	0.99	0.95		0.89	0.91	0.92
i7 11G	0.90	1.08	1.06	1.25	1.18	1.25		1.02	1.08
i9 12G	0.91	1.17	1.15	1.44	1.36	1.46	1.18		1.19
Clang	0.80	0.92	0.91	1.22	1.18	1.25	1.03	0.98	1.02
GCC	0.74	0.90	0.88	1.10	1.07	1.13	0.92	0.91	0.95
Final	0.95	0.96	0.96	1.12	1.13	1.13	1.03	1.02	1.04
1900X		1.05	1.02	1.24	1.19	1.22	1.07	1.00	1.10
5800X	1.00		0.97	1.29	1.26	1.26	1.10	1.06	1.11
5950X	1.05	1.02		1.35	1.27	1.31	1.22	1.09	1.16
i7 6G	0.95	1.05	1.01		0.97	0.99	1.07	1.50	1.06
i7 10G	0.96	1.01	0.99	1.04		1.00	1.00	0.97	1.00
i9 10G	0.98	1.04	1.00	1.06	0.98		1.02	0.98	1.01
i7 11G	1.01	1.05	1.02	1.23	1.18	1.23		1.63	1.15
i9 12G	1.05	1.12	1.09	1.33	1.29	1.32	1.24		1.17
Clang	0.88	0.95	0.92	1.10	1.07	1.09	1.04	0.93	1.00
GCC	0.90	0.96	0.93	1.11	1.07	1.10	0.99	0.91	0.99
Final	0.93	0.95	0.95	1.10	1.10	1.11	0.99	0.94	1.01

P-521

run on	1900X	5800X	5950X	i7 6G	i7 10G	i9 10G	i7 11G	i9 12G	G.M.
opt on									
1900X		1.05	1.07	1.32	1.31	1.31	1.15	1.08	1.15
5800X	0.94		1.01	1.26	1.26	1.25	1.15	1.06	1.11
5950X	0.94	0.99		1.29	1.27	1.31	1.14	1.59	1.18
i7 6G	0.86	0.96	0.99		0.99	0.99	0.93	1.00	0.97
i7 10G	0.87	0.97	0.97	1.01		1.03	0.91	0.98	0.97
i9 10G	0.87	0.97	0.99	1.00	1.04		0.91	1.51	1.02
i7 11G	0.93	1.02	1.04	1.26	1.20	1.21		0.98	1.07
i9 12G	0.98	1.08	1.10	1.38	1.37	1.36	1.22		1.18
Clang	0.98	1.07	1.08	1.38	1.37	1.37	1.20	1.19	1.20
GCC	0.89	1.03	1.04	1.18	1.18	1.18	1.03	1.03	1.06
Final	1.04	1.06	1.08	1.18	1.18	1.19	1.13	1.05	1.11
1900X		1.02	1.05	1.18	1.12	1.13	1.07	1.16	1.09
5800X	1.10		1.03	1.28	1.20	1.23	1.10	1.15	1.13
5950X	1.07	0.98		1.26	1.17	1.19	1.08	1.12	1.10
i7 6G	1.03	1.01	1.04		0.96	0.97	1.02	1.89	1.09
i7 10G	1.09	1.05	1.11	1.04		1.21	1.06	1.23	1.10
i9 10G	1.08	1.07	1.10	0.98	0.94		1.05	1.16	1.05
i7 11G	1.08	1.06	1.08	1.17	1.16	1.13		1.14	1.10
i9 12G	1.10	1.04	1.05	1.19	1.14	1.15	1.10		1.09
Clang	1.10	1.13	1.15	1.38	1.32	1.34	1.39	1.21	1.25
GCC	1.10	1.04	1.07	1.15	1.10	1.12	1.06	1.21	1.11
Final	1.10	1.07	1.07	1.17	1.17	1.15	1.06	1.21	1.12

Poly1305

run on	1900X	5800X	5950X	i7 6G	i7 10G	i9 10G	i7 11G	i9 12G	G.M.
opt on									
1900X		1.09	1.14	1.12	1.11	1.10	1.10	1.15	1.10
5800X	1.23		1.10	1.11	1.10	1.09	1.12	1.17	1.11
5950X	1.14	0.96		1.13	1.18	1.10	1.09	1.08	1.08
i7 6G	1.26	1.15	1.24		1.00	0.99	1.14	1.35	1.13
i7 10G	1.25	1.09	1.16	1.01		0.99	1.05	1.37	1.11
i9 10G	1.34	1.12	1.18	1.01	1.01		1.06	1.29	1.12
i7 11G	1.24	1.09	1.15	1.07	1.06	1.23		1.35	1.14
i9 12G	1.25	1.09	1.15	1.13	1.06	1.05	1.09		1.10
Clang	1.18	1.05	1.09	1.10	1.09	1.07	1.07	1.20	1.10
GCC	1.19	1.09	1.14	1.15	1.15	1.14	1.12	1.24	1.15
Final	1.18	1.09	1.09	1.10	1.09	1.09	1.07	1.20	1.11
1900X		1.19	1.17	1.14	1.14	1.14	1.10	1.14	1.13
5800X	1.13		1.00	1.11	1.11	1.11	1.08	1.08	1.08
5950X	1.15	1.00		1.11	1.11	1.11	1.08	1.14	1.09
i7 6G	1.30	1.17	1.14		1.00	1.00	1.04	1.40	1.12
i7 10G	1.23	1.16	1.16	1.00		1.00	1.17	1.41	1.13
i9 10G	1.26	1.18	1.16	1.00	1.00		1.06	2.51	1.21
i7 11G	1.24	1.13	1.11	1.09	1.09	1.09		1.35	1.13
i9 12G	1.20	1.11	1.10	1.07	1.07	1.06	1.06		1.08
Clang	1.12	1.08	1.07	1.06	1.07	1.07	1.03	1.21	1.09
GCC	1.17	1.15	1.14	1.15	1.15	1.15	1.13	1.23	1.16
Final	1.12	1.08	1.07	1.06	1.07	1.07	1.03	1.21	1.09

secp256k1

run on	1900X	5800X	5950X	i7 6G	i7 10G	i9 10G	i7 11G	i9 12G	G.M.
opt on									
1900X		1.01	0.99	1.11	1.10	1.11	1.11	1.15	1.07
5800X	1.13		0.98	1.22	1.22	1.25	1.18	1.24	1.15
5950X	1.17	1.12		1.27	1.27	1.30	1.26	1.34	1.21
i7 6G	1.04	0.95	0.94		1.00	1.01	1.01	1.04	1.00
i7 10G	1.07	0.99	0.96	1.00		1.01	1.04	1.05	1.01
i9 10G	1.05	1.00	0.98	1.03	1.04		1.04	1.04	1.02
i7 11G	1.08	0.98	0.97	1.07	1.09	1.07		1.03	1.03
i9 12G	1.05	1.00	0.98	1.08	1.08	1.12	1.02		1.04
Clang	1.35	1.27	1.24	1.42	1.42	1.43	1.19	1.43	1.34
GCC	1.61	1.39	1.36	1.94	1.94	1.96	1.77	2.05	1.73
Final	1.35	1.33	1.33	1.42	1.42	1.43	1.19	1.43	1.36
1900X		1.05	1.04	1.06	1.13	1.10	1.13	1.13	1.08
5800X	1.11		0.98	1.11	1.15	1.16	1.15	1.18	1.10
5950X	1.16	1.02		1.13	1.22	1.18	1.21	1.23	1.14
i7 6G	1.14	1.04	1.02		1.00	1.00	1.07	1.08	1.04
i7 10G	1.07	1.05	1.03	0.96		1.00	1.10	1.06	1.03
i9 10G	1.05	1.02	1.00	1.00	1.02		1.07	1.05	1.03
i7 11G	1.11	1.04	1.02	1.03	1.07	1.07		1.09	1.05
i9 12G	1.07	1.02	1.01	1.02	1.07	1.07	1.06		1.04
Clang	1.29	1.27	1.24	1.34	1.39	1.39	1.21	1.46	1.32
GCC	1.62	1.50	1.47	1.86	1.93	1.93	1.84	1.86	1.74
Final	1.29	1.27	1.26	1.39	1.39	1.39	1.21	1.46	1.33

Table 7. Cost of scalar multiplication (in cycles) of different implementations benchmarking on different machines.

	Implementation	Lang.	1900X	5800X	5950X	i7 6G	i7 10G	i9 10G	i7 11G	i9 12G	G.M.
Curve25519	sandy2x [Chou 2015]	asm-v	603k (1.06x)	**427k (1.00x)**	**426k (1.00x)**	521k (1.15x)	521k (1.15x)	521k (1.15x)	492k (1.16x)	480k (1.31x)	496k (1.08x)
	amd64-64 [Chen et al. 2014]	asm	622k (1.09x)	573k (1.34x)	573k (1.34x)	569k (1.26x)	571k (1.26x)	572k (1.27x)	533k (1.26x)	453k (1.24x)	556k (1.21x)
	amd64-51 [Chen et al. 2014]	asm	771k (1.35x)	584k (1.37x)	586k (1.37x)	568k (1.25x)	567k (1.25x)	567k (1.26x)	496k (1.17x)	424k (1.16x)	563k (1.22x)
	donna [Langley 2022]	asm-v	1092k (1.92x)	970k (2.27x)	970k (2.27x)	1013k (2.24x)	1015k (2.25x)	1015k (2.25x)	955k (2.26x)	887k (2.43x)	988k (2.14x)
	donna-c64 [Langley 2022]	C	806k (1.42x)	607k (1.42x)	602k (1.41x)	581k (1.28x)	582k (1.29x)	583k (1.29x)	543k (1.28x)	452k (1.24x)	588k (1.27x)
	OSSL ots [OpenSSL 2022]	C	700k (1.23x)	538k (1.26x)	540k (1.27x)	569k (1.26x)	567k (1.26x)	568k (1.26x)	492k (1.16x)	420k (1.15x)	544k (1.18x)
	OSSL fe-51 ots [OpenSSL 2022]	asm	695k (1.22x)	540k (1.27x)	539k (1.26x)	568k (1.25x)	567k (1.26x)	568k (1.26x)	491k (1.16x)	419k (1.15x)	543k (1.18x)
	OSSL fe-51+**CryptOpt**	asm	715k (1.26x)	531k (1.24x)	537k (1.26x)	561k (1.24x)	558k (1.24x)	559k (1.24x)	497k (1.17x)	402k (1.10x)	539k (1.17x)
	OSSL fe-64 ots [OpenSSL 2022]	asm	**570k (1.00x)**	444k (1.04x)	444k (1.04x)	462k (1.02x)	460k (1.02x)	464k (1.03x)	461k (1.09x)	412k (1.13x)	463k (1.00x)
	OSSL fe-64+**CryptOpt**	asm	590k (1.04x)	459k (1.08x)	464k (1.09x)	494k (1.09x)	494k (1.09x)	497k (1.10x)	451k (1.06x)	394k (1.08x)	478k (1.04x)
	HACL* fe-64 [HACL 2022]	asm	588k (1.03x)	494k (1.16x)	494k (1.16x)	**452k (1.00x)**	**452k (1.00x)**	**451k (1.00x)**	**423k (1.00x)**	**365k (1.00x)**	**461k (1.00x)**
secp256k1	libsecp256k1 [Bitcoin Core 2022]	asm	718k (1.06x)	568k (1.06x)	566k (1.05x)	564k (1.03x)	566k (1.05x)	564k (1.04x)	540k (1.08x)	437k (1.04x)	561k (1.04x)
	libsecp256k1 [Bitcoin Core 2022]	C	**676k (1.00x)**	543k (1.01x)	540k (1.00x)	563k (1.03x)	563k (1.05x)	562k (1.04x)	503k (1.00x)	438k (1.04x)	545k (1.01x)
	libsecp256k1+**CryptOpt**	asm	727k (1.08x)	547k (1.02x)	546k (1.01x)	552k (1.01x)	561k (1.04x)	551k (1.02x)	**501k (1.00x)**	**420k (1.00x)**	545k (1.01x)
	libsecp256k1+**CryptOpt**(CS2)	asm	715k (1.06x)	**535k (1.00x)**	**539k (1.00x)**	**548k (1.00x)**	**538k (1.00x)**	**541k (1.00x)**	506k (1.01x)	420k (1.00x)	**538k (1.00x)**

as do AMD Zen 3 processors in the 2x2 sub-matrices have values in the small interval $[0.98, 1.01]$ (mul), and equal for square.

This specialization to platforms is a double-edged sword. On the one hand, it allows CryptOpt to produce high-performing code, as previously shown. On the other hand, it also means that the code is potentially less generic, and it might be less performant if executed on platforms other than originally intended, as observed in our a-posteriori analysis.

D DETAILED PERFORMANCE INFORMATION

Table 7 shows the detailed performance information of the scalar multiplication experiments. The table is divided in two sections, one for Curve25519 and one for secp256k1. In each of those sections we compare different implementations of the scalar multiplication for respective curve against each other. The language, in which the core operations are implemented, is written in the column *Lang*. The following columns show the elapsed cycles for the scalar multiplication. The fastest implementation per section and per platform is highlighted in (**bold** text) and notes a (1.00x). All other implementations are then compared against this fastest in the form of the ratio, which is written in parenthesis. E.g. 1.05x means that it takes 1.05 times as many cycles than the fastest. The last column, G.M., is the geometric mean of the cycles across all platforms, the ratio is then recalculated on that G.M. as well.

For secp256k1, we base all evaluation on the implementation of `libsecp256k1`. We compare its asm and C implementations (first two rows) against our optimized version, when we plug in verified and optimized field operations (third row) and when we optimize its own implementation (in our Case Study 2).

Note: “ots” stands for off-the-shelf; “asm” means assembly; “-v” indicates the use of vector instructions. The HACL* implementation [HACL 2022] uses parallelized field arithmetic. For Curve25519, “-51“ and “-64“ indicates whether the representation for the field elements is unsaturated or saturated.