Billion-Gate Secure Computation with Malicious Adversaries

Benjamin Kreuter
brk7bx@virginia.edu
University of Virginia

abhi shelat
abhi@virginia.edu
University of Virginia

Chih-hao Shen
cs6zb@virginia.edu
University of Virginia

Abstract

The goal of this paper is to assess the feasibility of two-party secure computation in the presence of a malicious adversary. Prior work has shown the feasibility of billion-gate circuits in the semi-honest model, but only the 35k-gate AES circuit in the malicious model, in part because security in the malicious model is much harder to achieve. We show that by incorporating the best known techniques and parallelizing almost all steps of the resulting protocol, evaluating billion-gate circuits is feasible in the malicious model. Our results are in the standard model (i.e., no common reference strings or PKIs) and, in contrast to prior work, we do not use the random oracle model which has well-established theoretical shortcomings.

1 Introduction

Protocols for secure computation allow two or more mutually distrustful parties to collaborate and compute some function on each other's inputs, with privacy and correctness guarantees. Andrew Yao showed that secure two-party protocols can be constructed for any computable function [33]. Yao's protocol involves representing the function as a boolean circuit and having one party (called the *generator*) encrypt the circuit in such a way that it can be selectively decrypted by the other party (called the *evaluator*) to compute the output, a process called *garbling*. In particular, oblivious transfers are used for the evaluator to obtain a subset of the decryption keys that are needed to compute the output of the function.

Yao's protocol is of great practical significance. In many real-world situations, the inputs to a function may be too valuable or sensitive to share. Huang et al. explored the use of secure computation for biometric identification [14] in national security applications, in which it is desirable for individual genetic data to be kept private but still checked against a classified list. In a similar security application, Osadchy et al. described how face recognition could be performed in a privacy-preserving manner [29]. The more general case of multiparty computation has already seen real-world use in computing market clearing prices in Denmark [2].

Yao's original protocol ensures the privacy of each party's input and the correctness of the output under the *semi-honest* model, in which both parties follow the protocol honestly. This model has been the basis for several scalable secure computation systems [4, 10, 12, 13, 17, 22, 26]. It is conceivable, however, that one of the parties may deviate from the protocol in an attempt to violate privacy or correctness. Bidders may attempt to manipulate the auction output in their favor; spies may attempt to obtain sensitive information; and a computer being used for secure computation may be infected with malware. Securing against *malicious* participants, who may deviate arbitrarily from pre-agreed instructions, in an efficient manner is of more practical importance.

There have been several attempts on practical systems with security against active, malicious adversaries. Lindell and Pinkas presented an approach based on garbled circuits that uses the cut-and-choose technique [23], with an implementation of this system having been given by Pinkas et al. [30]. Nielsen et al. presented the LEGO+ system [28], which uses efficient oblivious transfers and authenticated bits to enforce honest behaviors from participants. shelat and Shen proposed a hybrid approach that integrates sigma protocols into the cut-and-choose technique [32]. The protocol compiler presented by Ishai, Prabhakaran, and Sahai [16] also uses an approach based on oblivious transfer, and was implemented by Lindell, Oxman, and Pinkas [21]. In all these cases, AES was used as a benchmark for performance tests.

Protocols for general multiparty computation with security against a malicious *majority* have also been presented. Canetti et al. gave a construction of a *universally composable* protocol in the *common reference string* model [5]. The protocol compiler of Ishai et al., mentioned above, can be used to construct a multiparty protocol with security against a dishonest majority in the UC model [16]. Bendlin et al. showed a construction based on homomorphic encryption [1], which was improved upon by Damgård et al. [7]; these protocols were also proved secure in the UC model, and thus require additional setup assumptions. The protocol of Damgård et al. (dubbed "SPDZ" and pronounced "speedz") is based on a preprocessing model, which improves the amortized performance. Damgård et al. presented an implementation of their protocol, which could evaluate the function $(x \times y) + z$ in about 3 seconds with a 128 bit security level, but with an amortized time of a few milliseconds.

This paper presents a scalable two-party secure computation system which guarantees privacy and correctness in the presence of a malicious party. The system we present can handle circuits with hundreds of millions or even billions of gates, while requiring relatively modest computing resources. Our system follows the Fairplay framework, allowing general purpose secure computation starting from a high level description of a function. We present a system with numerous technical advantages over the Fairplay system, both in our compiler and in the secure computation protocol. Unlike previous work, we do not rely solely on AES circuits as our benchmark; our goal is to evaluate circuits that are orders of magnitude larger than AES in the malicious model, and we use AES only as a comparison with other work. We prove the security of our protocol assuming *circular 2-correlation robust* hash functions and the hardness of the elliptic curve discrete logarithm problem, and require neither additional setup assumptions nor the random oracle model.

2 Contributions

Our principal contribution is to build a high performance secure two-party computation system that integrates state-of-the-art techniques for dealing with *malicious* adversaries efficiently. Although some of these techniques have been reported individually, we are not aware of any attempt to incorporate them all into one system, while ensuring that a security proof can still be written for that system. Even though some of the techniques are claimed to be compatible, it is not until everything is put together and someone has gone through all the details can a system as a whole be said to be provably secure.

System Framework We start by using Yao's garbled circuit [33] protocol for securely computing functions in the presence of semi-honest adversaries, and shelat and Shen's cut-and-choose-based transformation [32] that converts Yao's garbled circuit protocol into one that is secure against malicious adversaries.

We then modify the above to use Ishai et al.'s oblivious transfer extension [15] that has efficient amortized computation time for oblivious transfers secure against malicious adversaries, and Lindell and Pinkas' random combination technique [23] that defends against selective failure attacks. We implement Kiraz's randomized circuit technique [18] that guarantees that the generator gets either no output or an authentic output, i.e., the generator cannot be tricked into accepting arbitrary output.

Optimization Techniques For garbled circuit generation and evaluation, we incorporate Kolesnikov and Schneider's free-XOR technique that minimizes the computation and communication cost for XOR gates in a circuit [20]. We also adopt Pinkas et al.'s garbled-row-reduction technique that reduces the communication cost for k-fan-in non-XOR gates by $1/2^k$ [30], which means at least a 25% communication saving in our system since we only have gates of 1-fan-in or 2-fan-in. Finally, we implement Goyal et al.'s technique for reducing communication as follows: during the cut-and-choose step, the check circuits are given to the evaluator by revealing the random seeds used to produce them rather than the check circuits themselves [11]. Combined with the 60%-40% check-evaluation ratio proposed by shelat and Shen [32], this technique provides a near 60% saving in communication. As far as we know, although these techniques exist individually, ours is the first system to incorporate all of these mutually compatible state-of-the-art techniques.

Circuit-Level Parallelism The most important new technique that we use is to exploit the embarrassingly parallel nature of shelat and Shen's protocol for achieving security in the malicious model. Exploiting this, however, requires careful engineering in order to achieve good performance while maintaining security. We parallelize all computation-intensive operations such as oblivious transfers or circuit construction by splitting the generator-evaluator pair into hundreds of slave pairs. Each of the pairs works on an independently generated copy of the circuit in a parallel but synchronized manner as synchronization is required for shelat and Shen's protocol [32] to be secure.

Computation Complexity For the computation time of a secure computation, there are two main contributing factors: the input processing time I (due to oblivious transfers) and the circuit processing time C (due to garbled circuit construction and evaluation). In the semi-honest model, the system's computation time is simply $I + C$. Security in the malicious model, however, requires several extra checks. In the first instantiation of our system, through heavy use of circuit-level parallelism, our system needs roughly $I+2C$ to compute hundreds of copies of the circuit. Thus when the circuit size is sufficiently larger than the input size, our system (secure in the malicious model) needs roughly twice as much computation time as that needed by the original Yao protocol (secure in the semi-honest model). This is a tremendous improvement over prior work [30,32] which needed 100x more time than the semi-honest Yao. In the second instantiation of our scheme, we are able to achieve $I+C$ computation time, albeit at the cost of moderately more communication overhead.

Large Circuits In the Fairplay system, a garbled circuit is fully constructed before being sent over a network for the other party to evaluate. This approach is particularly problematic when hundreds of copies of a garbled circuit are needed against malicious adversaries. Huang et al. [13] pointed out that keeping the whole garbled circuit in memory is unnecessary, and that instead, the generation and evaluation of garbled gates could be conducted in a "pipelined" manner. Consequently, not only do both parties spend less time idling, only a small number of garbled gates need to reside in memory at one time, even when dealing with large circuits. However, this pipelining idea does not work trivially with other optimization techniques for the following two reasons:

- The cut-and-choose technique requires the generator to finish constructing circuits before the coin flipping (which is used to determine check circuits and evaluation circuits), but the evaluator cannot start checking or evaluating before the coin flipping. A naive approach would ask the evaluator to hold the circuits and wait for the results of the coin flipping before she proceeds to do her jobs. When the circuit is of large size, keeping hundreds of copies of such a circuit in memory is undesirable.
- Similarly, the random seed checking technique [11] requires the generator to send the hash for each garbled circuit, and later on send the random seeds for check circuits so that the communication for check circuits is vastly reduced. Note that the hash for an evaluation circuit is given away before the garbled circuit itself. However, a hash is calculated only after the whole circuit is generated. So the generation-evaluation pipelining cannot be applied directly.

Our system, however, integrates this pipelining idea with the optimization techniques mentioned above, and is capable of handling circuits of billions of gates.

AES-NI Besides the improvements by the algorithmic means, we also incorporate the Intel Advanced Encryption Standard Instructions (AES-NI) in our system. While the encryption is previously suggested to be

$$\mathrm{Enc}_{X,Y}(Z) = H(X||Y) \oplus Z$$

in the literature [6, 20], where H is a 2-circular correlation robust function instantiated either with SHA-1 [13] or SHA-256 [30], we propose an alternative that

$$\mathrm{Enc}^{k}_{X,Y}(Z) = \text{AES-256}_{X||Y}(k) \oplus Z,$$

where k is the index of the garbled gate. With the help of the latest instruction set, an AES-256 operation could take as little as 30% of the time for SHA-256. Since this operation is heavily used in circuit operations, with the help of AES-NI instructions, we are able to reduce the circuit computation time C by at least 20%.

Performance To get a sense of our improvements, we list the experimental results of the benchmark function—AES—from the most recent literature and our system. The latest reported system in the semi-honest model was built by Huang et al. [13] and needs 1.3 seconds (where $I = 1.1$ and $C = 0.2$) to complete a block of secure AES computation. The fastest known system in the malicious model was proposed by Nielson et al. [28] and has an amortized performance 1.6 seconds per block (or more precisely, $I = 79$ and $C = 6$ for 54 blocks). Our system provides security in the malicious model and needs 1.4 ($= I+2C$, where $I = 1.0$ and $C = 0.2$) seconds per block. Note that both the prior systems require the full power of a random oracle, while ours requires a weaker cryptographic primitive, 2-circular correlation robust functions, which was recently shown to be sufficient to prove the security of the free-XOR technique. It should also be noted that our system benefits greatly from parallel computation, which was not tested for LEGO+.

Scalable Circuit Compiler One of the major bottlenecks that prevents large-scale secure computation is the need for a scalable compiler that generates a circuit description from a function written in a high-level programming language. Prior tools could barely handle circuits with $50,000$ gates, requiring significant computational resources to compile such circuits. While this is just enough for an AES circuit, it is not enough for the large circuits that we evaluate in this paper.

We present a scalable boolean circuit compiler that can be used to generate circuits with billions of gates, with moderate hardware requirements. This compiler performs some simple but highly effective optimizations, and tends to favor XOR gates. The toolchain is flexible, allowing for different levels of optimizations and can be parameterized to use more memory or more CPU time when building circuits.

As a first sign that our compiler advances the state of the art, we observe that it automatically generates a smaller boolean circuit for the AES cipher than the hand-optimized circuit reported by Pinkas et al. [30]. AES plays an important role in secure computation, and oblivious AES evaluation can be used as a building block in cryptographic protocols. Not only is it one of the most popular building blocks in cryptography and real life security, it is often used as a benchmark in secure computation. With the textbook algorithm, the well-known Fairplay compiler can generate an AES circuit that has 15,316 non-XOR gates. Pinkas et al. were able to develop an optimized AES circuit that has 11,286 non-XOR gates. By applying an efficient S-box circuit [3] and using our compiler, we were able to construct an AES circuit that has 9,100 non-XOR gates. As a result, our AES circuit only needs 59% and 81% of the communication needed by the other two, respectively.

Most importantly, with our system and the scalable compiler, we are able to run experiments on circuits with sizes in the range of billions of gates. To the best of our knowledge, secure computation with such large circuits has never been run in the malicious model before. These circuits include 256-bit RSA (266,150,119 gates) and 4095x4095-bit edit distance (5,901,194,475). As the circuit size grows, resource management becomes crucial. A circuit of billions of gates can easily result in several GB of data stored in memory or sent over the network. Special care is required to handle these difficulties.

Paper Organization The organization of this paper is as follows. A variety of security decisions and optimization techniques will be covered in Section 3 and Section 4, respectively. Then, our system, including a compiler, will be introduced in Section 5. Finally, the experimental results are presented in Section 6 followed by the conclusion and future work in Section 7.

3 Techniques Regarding Security

The Yao protocol, while efficient, assumes honest behavior from both parties. To achieve security in the malicious model, it is necessary to enforce honest behavior. The *cut-and-choose* technique is one of the most efficient methods in the literature and is used in our system. Its main idea is for the generator to prepare multiple copies of the garbled circuit with independent randomness, and the evaluator picks a random fraction of the received circuits, whose randomness is then revealed. If any of the chosen circuits (called *check circuits*) is not consistent with the revealed randomness, the evaluator aborts; otherwise, she evaluates the remaining circuits (called *evaluation circuits*) and takes the majority of the outputs, one from each evaluation circuit, as the final output.

The intuition is that to pass the check, a malicious generator can only sneak in a few faulty circuits, and the influence of these (supposedly minority) faulty circuits will be eliminated by the majority operation at the end. On the other hand, if a malicious generator wants to manipulate the final output, she needs to construct faulty majority among evaluation circuits, but then the chance that none of the faulty circuits is checked will be negligible. So with the help of the cut-and-choose method, a malicious generator either constructs many faulty circuits and gets caught with high probability, or constructs merely a few and has no influence on the final output.

However, the cut-and-choose technique is not a cure-all. Several subtle attacks have been reported and would be a problem if not properly handled. These attacks include the *generator's input inconsistency attack*, the *selective failure attack*, and the *generator's output authenticity attack*, which are discussed in the following sections. Note that in this section, n denotes the input size and s denotes the number of copies of the circuit.

Generator's Input Consistency Recall that in the cut-and-choose step, multiple copies of a circuit are constructed and then evaluated. A malicious generator is therefore capable of providing altered inputs to different evaluation circuits. It has been shown that for some functions, there are simple ways for the generator to extract information about the evaluator's input [23]. For example, suppose both parties agree to compute the inner-product of their input, that is, $f([a_2,a_1,a_0],[b_2,b_1,b_0]) \mapsto a_2b_2+a_1b_1+a_0b_0$ where a_i and b_i is the generator's and evaluator's i-th input bit, respectively. Instead of providing $[a_2,a_1,a_0]$ to all evaluation circuits, the generator could send $[1,0,0]$, $[0,1,0]$, and $[0,0,1]$ to different copies of the evaluation circuits. After the majority operation from the cut-and-choose technique, the generator learns major(b_2,b_1,b_0), the majority bit in the evaluator's input, which is not what the evaluator agreed to reveal in the first place.

There exist several approaches to deter this attack. Mohassel and Franklin [27] proposed the equality-checker that needs $O(ns^2)$ commitments to be computed and exchanged. Lindell and Pinkas [23] developed an approach that also requires $O(ns^2)$ commitments. Later, Lindell and Pinkas [24] proposed a pseudorandom synthesizer that relies on efficient zero-knowledge proofs for specific hardness assumptions and requires $O(ns)$ group operations. shelat and Shen [32] suggested the use of malleable claw-free collections, which also uses $O(ns)$ group operations, but they showed that witness-indistinguishability suffices, which is more efficient than zero-knowledge proofs by a constant factor.

In our system, we incorporate the malleable claw-free collection approach because of its efficiency. Although the commitment-based approaches can be implemented using lightweight primitives such as collision-resistant hash functions, they incur high communication overhead for the extra complexity factor s, that is, the number of copies of the circuit. On the other hand, the group-based approach could be more computationally intensive, but this discrepancy is compensated again due to the parameter s.[1] Hence, with similar computation cost, group-based approaches enjoy lower communication overhead.

Selective Failure A more subtle attack is *selective failure* [19, 27]. A malicious generator could use inconsistent keys to construct the garbled gate and OT so that the evaluator's input can be inferred from whether or not the protocol completes. In particular, a cheating generator could assign (K_0, K_1) to an input wire in the garbled circuit while using (K_0, K_1^*) instead in the corresponding OT, where $K_1 \neq K_1^*$. As a result, if the evaluator's input is 0, she learns K_0 from OT and completes the evaluation without complaints; otherwise, she learns K_1^* and gets stuck during the evaluation. If the protocol expects the evaluator to share the result with the generator at the end, the generator learns whether or not the evaluation failed, and therefore, the evaluator's input is leaked.

Lindell and Pinkas [23] proposed the random input replacement approach that involves replacing each of the evaluator's input bits with an XOR of s additional input bits, so that whether the evaluator aborts due to a selective failure attack is almost independent (up to a bias of 2^{1-s}) of her actual input value. Both Kiraz [18] and shelat and Shen [32] suggested a solution that exploits committing OTs so that the generator commits to her input for the OT, and the correctness of the OTs can later be checked by opening the commitments during the cut-and-choose. Lindell and Pinkas [24] also proposed a solution to this problem using cut-and-choose OT, which combines the OT and the cut-and-choose steps into one protocol to avoid this attack.

Our system is based on the random input replacement approach due to its scalability. It is a fact that the committing OT or the cut-and-choose OT does not alter the circuit while the random input replacement approach inflates the circuit by $O(sn)$ additional gates. However, it has been shown that $\max(4n, 8s)$ additional gates suffice [30]. Moreover, both the committing OT and the cut-and-choose OT require $O(ns)$ group operations, while the random input replacement approach needs only $O(s)$ group operations. Furthermore, we observe that the random input replacement approach is in fact compatible with the OT extension technique. Therefore, we were able to build our system which has the group operation complexity independent of the evaluator's input size, and as a result, our system is particularly attractive when handling a circuit with a large evaluator input.

Generator's Output Authenticity It is not uncommon that *both* the generator and evaluator receive outputs from a secure computation, that is, the goal function is $f(x,y) = (f_1, f_2)$, where the generator with input x gets output f_1, and the evaluator with input y gets f_2.[2] In this case, the security requires that *both* the input and output are hidden from each other. In the semi-honest setting, the straightforward solution is to let the generator choose a random number c as an extra input, convert $f(x,y) = (f_1, f_2)$ into a new function $f^*((x,c),y) = (\lambda, (f_1 \oplus c, f_2))$, run the original Yao protocol for f^*, and instruct the evaluator to pass the encrypted output $f_1 \oplus c$ back to the generator, who can then retrieve her real output f_1 with the secret input c chosen in the first place. However, the situation gets complicated when either of the participants could potentially be malicious. In particular, a malicious evaluator might claim an arbitrary value to be the generator's output coming from the circuit evaluation. Note that the two-output protocols we consider are not *fair* since the evaluator always learns her own output and may refuse to send the generator's output. However, they can satisfy the notion that the evaluator cannot trick the generator into accepting arbitrary output.

Many approaches have been proposed to ensure the generator's output authenticity. Lindell and Pinkas [23] proposed a solution similar to the aforementioned solution in the semi-honest setting, where the goal function is modified to compute $f_1 \oplus c$ and its MAC so that the generator can verify the authenticity of her output. This approach incurs a cost of adding $O(n^2)$ gates to the circuit. Kiraz [18] presented a two-party computation protocol in which a zero knowledge proof of size $O(s)$ is conducted at the end. shelat and Shen [32] suggested a signature-based solution which, similar to Kiraz's, adds n gates to the circuit, and requires a proof of size $O(s+n)$ at the end. However, they observed that witness-indistinguishable proofs are sufficient.

Lindell and Pinkas' approach, albeit straightforward, might introduce greater communication overhead than the description function itself. We therefore employ the approach that takes the advantages of the remaining two solutions. In particular, we implement Kiraz's approach

[1] To give concrete numbers, with an Intel Core i5 processor and 4GB DDR3 memory, a SHA-256 operation (from OpenSSL) requires 1,746 cycles, while a group operation (160-bit elliptic curve from the PBC library with preprocessing) needs 322,332 cycles. It is worth-mentioning that s is at least 256 in order to achieve security level 2^{-80}. The gap between a symmetric operation and an asymmetric one becomes even smaller when modern libraries such as RELIC are used instead of PBC.

[2] Here f_1 and f_2 are short for $f_1(x,y)$ and $f_2(x,y)$ for simplicity.

(smaller proof size), but only a witness-indistinguishable proof is performed (weaker security property).

4 Techniques Regarding Performance

Yao's garbled circuit technique has been studied for decades. It has drawn significant attention for its simplicity, constant round complexity, and computational efficiency (since circuit evaluation only requires fast symmetric operations). The fact that it incurs high communication overhead has provoked interest that has led to the development of fruitful results.

In this section, we will first briefly present the Yao garbled circuit, and then discuss the optimization techniques that greatly reduce the communication cost while maintaining the security. These techniques include free-XOR, garbled row reduction, random seed checking, and large circuit pre-processing. In addition to these original ideas, practical concerns involving large circuits and parallelization will be addressed.

4.1 Baseline Yao's Garbled Circuit

Given a circuit that consists of 2-fan-in boolean gates, the generator constructs a garbled version as follows: for each wire w, the generator picks a random permutation bit $\pi_w \in \{0,1\}$ and two random keys $w_0, w_1 \in \{0,1\}^{k-1}$. Let $W_0 = w_0||\pi_w$ and $W_1 = w_1||(\pi_w \oplus 1)$, which are associated with bit value 0 and 1 of wire w, respectively. Next, for gate $g \in \{f | f : \{0,1\} \times \{0,1\} \mapsto \{0,1\}\}$ that has input wire x with (X_0, X_1, π_x), input wire y with (Y_0, Y_1, π_y), and output wire z with (Z_0, Z_1, π_z), the garbled truth table for g has four entries:

$$GTT_g \begin{cases} \text{Enc}(X_{0\oplus\pi_x}||Y_{0\oplus\pi_y},\ Z_{g(0\oplus\pi_x,0\oplus\pi_y)}) \\ \text{Enc}(X_{0\oplus\pi_x}||Y_{1\oplus\pi_y},\ Z_{g(0\oplus\pi_x,1\oplus\pi_y)}) \\ \text{Enc}(X_{1\oplus\pi_x}||Y_{0\oplus\pi_y},\ Z_{g(1\oplus\pi_x,0\oplus\pi_y)}) \\ \text{Enc}(X_{1\oplus\pi_x}||Y_{1\oplus\pi_y},\ Z_{g(1\oplus\pi_x,1\oplus\pi_y)}). \end{cases}$$

$\text{Enc}(K,m)$ denotes the encryption of message m under key K. Here the encryption key is a concatenation of two labels, and each label is a random key concatenated with its permutation bit. Intuitively, π_x and π_y permute the entries in GTT_g so that for $i_x, i_y \in \{0,1\}$, the $(2i_x + i_y)$-th entry represents the input pair $(i_x \oplus \pi_x, i_y \oplus \pi_y)$ for gate g, in which case the label associated with the output value $g(i_x \oplus \pi_x, i_y \oplus \pi_y)$ could be retrieved. More specifically, to evaluate the garbled gate GTT_g, suppose $X||b_x$ and $Y||b_y$ are the retrieved labels for input wire x and wire y, respectively, the evaluator will use $X||b_x||Y||b_y$ to decrypt the $(2b_x + b_y)$-th entry in GTT_g and retrieve label $Z||b_z$, which is then used to evaluate the gates at the next level. The introduction of the permutation bit helps to identify the correct entry in GTT_g, and thus, only one, rather than all, of the four entries will be decrypted.

4.2 Free-XOR

Kolesnikov and Schneider [20] proposed the free-XOR technique that aims for removing the communication cost and decreasing the computation cost for XOR gates.

The idea is that the generator first randomly picks a global key R, where $R = r||1$ and $r \in \{0,1\}^{k-1}$. This global key has to be hidden from the evaluator. Then for each wire w, instead of picking both W_0 and W_1 at random, only one is randomly chosen from $\{0,1\}^k$, and the other is determined by $W_b = W_{1\oplus b} \oplus R$. Note that π_w remains the rightmost bit of W_0. For an XOR gate having input wire x with $(X_0, X_0 \oplus R, \pi_x)$, input wire y with $(Y_0, Y_0 \oplus R, \pi_y)$, and output wire z, the generator lets $Z_0 = X_0 \oplus Y_0$ and $Z_1 = Z_0 \oplus R$. Observe that

$$X_0 \oplus Y_1 = X_1 \oplus Y_0 = X_0 \oplus Y_0 \oplus R = Z_0 \oplus R = Z_1$$
$$X_1 \oplus Y_1 = X_0 \oplus R \oplus Y_0 \oplus R = X_0 \oplus Y_0 = Z_0.$$

This means that while evaluating an XOR gate, XORing the labels for the two input wires will directly retrieve the label for the output wire. So no garbled truth table is needed, and the cost of evaluating an XOR gate is reduced from a decryption operation to a bitwise XOR.

This technique is only secure when the encryption scheme satisfies certain security properties. The solution provided by the authors is

$$\text{Enc}(X||Y, K) = H(X||Y) \oplus Z,$$

where $H : \{0,1\}^{2k} \mapsto \{0,1\}^k$ is a random oracle. Recently, Choi et al. [6] have further shown that it is sufficient to instantiate $H(\cdot)$ with a weaker cryptographic primitive, *2-circular correlation robust functions*. Our system instantiates this primitive with $H(X||Y) = \text{SHA-256}(X||Y)$. However, when AES-NI instructions are available, our system instantiates it with $H^k(X||Y) = \text{AES-256}(X||Y, k)$, where k is the gate index.

4.3 Garbled Row Reduction

The GRR (Garbled Row Reduction) technique suggested by Pinkas et al. [30] is used to reduce the communication overhead for non-XOR gates. In particular, it reduces the size of the garbled truth table for 2-fan-in gates by 25%.

Recall that in the baseline Yao's garbled circuit, both the 0-key and 1-key for each wire are randomly chosen. After the free-XOR technique is integrated, the 0-key and 1-key for an XOR gate's output wire depend on input key and R, but the 0-key for a non-XOR gate's output wire is still free. The GRR technique is to make a smart choice for this degree of freedom, and thus, reduce one entry in the garbled truth table to be communicated over network.

In particular, the generator picks (Z_0, Z_1, π_z) by letting $Z_{g(0\oplus\pi_x,0\oplus\pi_y)} = H(X_{0\oplus\pi_x}||Y_{0\oplus\pi_y})$, that is, either Z_0 or Z_1

is assigned to the encryption mask for the 0-th entry of the GTT_g, and the other one is computed by the equation $Z_b = Z_{1\oplus b} \oplus R$. Therefore, when the evaluator gets $(X_{0\oplus\pi_x}, Y_{0\oplus\pi_y})$, both $X_{0\oplus\pi_x}$ and $Y_{0\oplus\pi_y}$ have rightmost bit 0, indicating that the 0-th entry needs to be decrypted. However, with GRR technique, she is able to retrieve $Z_{g(0\oplus\pi_x, 0\oplus\pi_y)}$ by running $H(\cdot)$ without inquiring GTT_g.

Pinkas et al. claimed that this technique is compatible with the free-XOR technique [30]. For rigorousness purposes, we carefully went through the details and came up with a security proof for our protocol that confirms this compatibility. The proof will be included in the full version of this paper.

4.4 Random Seed Checking

Recall that the cut-and-choose approach requires the generator to construct multiple copies of the garbled circuit, and more than half of these garbled circuits will be fully revealed, including the randomness used to construct the circuit. Goyal, Mohassel, and Smith [11] therefore pointed out an insight that the evaluator could examine the correctness of those check circuits by receiving a hash of the garbled circuit first, acquiring the random seed, and reconstructing the circuit and hash by herself.

This technique results in the communication overhead for check circuits independent of the circuit size. This technique has two phases that straddle the coin-flipping protocol. Before the coin flipping, the generator constructs multiple copies of the circuit as instructed by the cut-and-choose procedure. Then the generator sends to the evaluator the hash of each garbled circuit, rather than the circuit itself. After the coin flipping, when the evaluation circuits and the check circuits are determined, the generator sends to the evaluator the full description of the evaluation circuits and the random seed for the check circuits. The evaluator then computes the evaluation circuits and tests the check circuits by reconstructing the circuit and comparing its hash with the one received earlier. As a result, even for large circuits, the communication cost for each check circuit is simply a hash value plus the random seed. Our system provides that 60% of the garbled circuits are check circuits. Thus, this optimization significantly reduces communication overhead.

4.5 Working with Large Circuits

A circuit for a reasonably complicated function can easily consist of billions of gates. For example, a 4095-bit edit distance circuit has 5.9 billion gates. When circuits grow to such a size, the task of achieving high performance secure computation becomes challenging.

An $(I+2C)$-time solution Our solution for handling large circuits is based on Huang et al.'s work [13], which is the only prior work capable of handling large circuits (of up to 1.2 billion non-XOR gates) in the semi-honest setting. Intuitively, the generator could work with the evaluator in a pipeline manner so that small chunks of gates are being processed at a time. The generator could start to work on the next chunk while the evaluator is still processing the current one. However, this technique does not work directly with the random seed checking technique described above in Section 4.4 because the generator has to finish circuit construction and hash calculation before the coin flipping, but the evaluator could start the evaluation only after the coin flipping. As a result, the generator needs a way to construct the circuit first, wait for the coin flipping, and send the evaluation circuits to the evaluator without keeping them in memory the whole time. We therefore propose that the generator constructs the evaluation circuits all over again after the coin flipping, with the same random seed used before and the same keys for input wires gotten from OT.

We stress that when fully parallelized, the second construction of an evaluation circuit does not incur overhead to the overall execution time. Although we suggest to construct an evaluation circuit twice, the fact is that according to the random seed checking, a check circuit is already being constructed twice—once before the coin flipping by the generator for hash computation and once after by the evaluator for correctness verification. As a result, when each generator-evaluator pair is working on a single copy of the garbled circuit, the constructing time for a evaluation circuit totally overlaps with that for a check circuit. We therefore achieve the overall computation time $I+2C$ mentioned earlier, where the first C is for the generator to calculate the circuit hash, and the other C is either for the evaluator to reconstruct a check circuit or for both parties to work on an evaluation circuit in a pipeline manner as suggested by Huang et al. [13].

Achieving an $(I+C)$-time solution We observe that there is a way to achieve $I+C$ computation time, which exactly matches the running time of Yao in the semi-honest setting. This idea, however, is not compatible with the random-seed technique, and therefore represents a trade-off between communication and computation. Recall that the generator has to finish circuit construction and hash evaluation before beginning coin flipping, whereas the evaluator can start evaluating only after receiving the coin flipping results. The idea is to run the coin flipping in the way that only the evaluator gets the result and does not reveal it to the generator until the circuit construction is completed. Since the generator is oblivious to the coin flipping result, she sends every garbled circuit to the evaluator, who could then either evaluate or check the received circuit. In order for the evaluator to get the generator's input keys for evaluation circuits and the random seed for the check circuits, they run an OT, where the evaluator uses the coin flipping result as input and the generator provides either the random seed (for the check circuit) or his input keys (for the evaluation circuit). After the generator completes circuit construction and reveals the circuit hash, the evaluator compares the hash with her own calculation, if the hashes match, she proceeds with the rest of the original protocol. Note that this approach comes at the cost of sacrificing the random seed checking technique and its 60% savings in communication.

Working Set Optimization Another problem encountered while dealing with large circuits is the *working set minimization problem*. Note that the *circuit value problem* is log-space complete for P. It is suspected that L≠P, that is, there exist some circuits that can be evaluated in polynomial time but require more than logarithmic space. This open problem captures the difficulty of handling large circuits during both the construction and evaluation, where at any moment there is a set of wires, called the *working set*, that are available and will be referenced in the future. For some circuits, the working set is inherently super-logarithmic. A naive approach is to keep the most recent D wires in the working set, where D is the upper bound of the input-output distance of all gates. However, there may be wires which are used as inputs to gates throughout the entire circuit, and so this technique could easily result in adding almost the whole circuit to the working set, which is especially problematic when there are hundreds of copies of a circuit of billions of gates. While reordering the circuit or adding identity gates to minimize D would mitigate this problem, doing so while maintaining the topological order of the circuit is known to be an NP-complete problem, the *graph bandwidth problem* [9].

Our solution to this difficulty is to pre-process the circuit so that each gate comes with a usage count. Our system has a compiler that converts a program in high-level language into a boolean circuit. Since the compiler is already using global optimization in order to reduce the circuit size, it is easy for the global optimizer to analyze the circuit and calculate the usage count for each gate. With this information, it is easy for the generator and evaluator to decrement the counter for each gate whenever it is being referenced and to toss away the gate whenever its counter becomes zero. In other words, we keep track of merely useful information and heuristically minimize the size of the working set, which is small compared with the original circuit size as shown in Table 1.

	AES	Dot_4^{64}	RSA-32	EDT-255
circuit size	49,912	460,018	1,750,787	15,540,196
wrk set size	323	711	235	2,829

Table 1: The size of the working set for various circuits (sizes include input gates)

5 Boolean Circuit Compiler

Although the Fairplay circuit compiler can generate circuits, it requires a very large amount of computational resources to generate even relatively small circuits. Even on a machine with 48 gigabytes of RAM, Fairplay terminates with an out-of-memory error after spending 20 minutes attempting to compile an AES circuit. This makes Fairplay impractical for even relatively small circuits, and infeasible for some of the circuits tested in this project. One goal of this project was to have a general purpose system for secure computation, and so writing application specific programs to generate circuits, a technique used by others [13], was not an option.

To address this problem, we have implemented a new compiler that generates a more efficient output format than Fairplay, and which requires far lower computational resources to compile circuits. We were able to generate the AES circuit in only a few seconds on a typical desktop computer with only 8GB of RAM, and were able to generate and test much larger non-trivial circuits. We used the well-known *flex* and *bison* tools to generate our compiler, and implemented an optimizer as a separate tool. We also use the results from [30] to reduce 3 arity gates to 2 arity gates.

As a design decision, we created an imperative, untyped language with static scoping. We allow code, variables, and input/output statements to exist in the global scope; this allows very simple programs to be written without too much extra syntax. Functions may be declared, but may not be recursive. Variables do not need to be declared before being used in an unconditional assignment; variables assigned within a function's body that are not declared in the global scope are considered to be local. Arrays are a language feature, but array indices must be constants or must be determined at compile time. If run-time determined indices are required for a function, a loop that selects the correct index may be used; this is necessary for oblivious evaluation. Variables may be arbitrarily concatenated, and bits or groups of bits may be selected from any variable and bits or ranges of bits may be assigned to; as with arrays, the index of a bit must be determined at compile time, or else a loop must be used. Note that loop variables may be used as such an index, since loops are always completely unrolled, and therefore the loop index can always be resolved at compile time. Additional language features are planned as future work.

We use some techniques from the Fairplay compiler in our own compiler. In particular we use the single assignment algorithm from Fairplay, which is required to deal with assignments that occur inside of *if* statements. Otherwise, our compiler has several distinguishing characteristics that make it more resource efficient than Fairplay. The front end of our compiler attempts to generate circuits as quickly as possible, using as little memory as possible and performing only rudimentary optimizations before emitting its output. This can be done with very modest computational resources, and the intermediate output can easily be translated into a circuit for evaluation. The main optimizations are performed by the back end of the compiler, which identifies gates that can be removed without affecting the output of the circuit as a whole.

Unlike the Fairplay compiler, we avoided the use of hash tables in our compiler, using more memory-efficient storage. Our system can use one of three storage strategies: memory-mapped files, flat files without any mapping, and Berkeley DB. In our tests, we found that memory mapped files always resulted in the highest performance, but that Berkeley DB is only sometimes better than direct access without any mapping.

In the following sections, we describe these contributions in more detail, and provide experimental results.

5.1 Circuit Optimizations

The front-end of our compiler tends to generate inefficient circuits, with large numbers of unnecessary gates. As an example, for some operations the compiler generates large numbers of identity gates i.e. gates whose outputs follow one of their inputs. It is therefore essential to optimize the circuits emitted by the front end, particularly to meet our system's overall goal of practicality.

Our compiler uses several stages of optimization, most of which are global. As a first step, a local optimization removes redundant gates, i.e. gates that have the same truth table and input wires. This first step operates on a fixed-size chunk of the circuit, but we have found that there are diminishing improvements as the size of this window is increased. We also remove constant gates, identity gates, and inverters, which are generated by the compiler and which may be inadvertently generated during the optimization process. Finally, we remove gates that do not influence the output, which can be thought of as dead code elimination. The effectiveness of each optimization on different circuits is shown in Figure 1. The circuit that was least optimizable was the edit distance circuit, being reduced to only 82% of its size from the front end, whereas the RSA signing and the dot product circuits were the most optimizable, being reduced to roughly half of the gates emitted by the front end.

Gate Removal The front-end of the compiler emits gates in topological order, and similar to Fairplay, our compiler assigns explicit identifiers to each emitted gate. To remove gates efficiently, we store a table that maps the identifiers of gates that were removed to the previously emitted gates, and for each gate that is scanned the inputs are rewritten according to this table. The table itself is then emitted, so that the identifiers of non-removed gates can be corrected. This mapping process can be done in linear time and space using an appropriate key-value store.

Removing Redundant Gates Some of the gates generated by the front end of our compiler have the same truth table and input wires as previously generated gates; such gates are redundant and can be removed. This removal process has the highest memory requirement of any other optimization step, since a description of every non-redundant gate must be stored. However, we found during our experiments that this optimization can be performed on discrete chunks of the circuit with results that are very close to performing the optimization on the full circuit, and that there are diminishing improvements in effectiveness as the size of the chunks is increased. Therefore, we perform this optimization using chunks, and can use hash tables to improve the speed of this step.

Removing Identity Gates and Inverters The front end may generate identity gates or inverters, which are not necessary. This may happen inadvertently, such as when a variable is incremented by a constant, or as part of the generation of a particular logic expression. While removing identity gates is straightforward, the removal of inverters requires more work, as gates which have inverted input wires must have their truth tables rewritten. There is a cascading effect in this process; the removal of some identity gates or inverters may transform later gates into identity gates or inverters. This step also removes gates with constant outputs, such as an XOR gate with two identical inputs. Constant propagation and folding occur as a side effect of this optimization.

Removing Unused Gates Finally, some gates in the circuit may not affect the output value at all. For this step, we scan the circuit backwards, and store a table of live gates; we then re-emit the live gates in the circuit and skip the dead gates. Immediately following this step, the circuit is prepared for the garbled circuit generator, which includes generating a usage count for each gate.

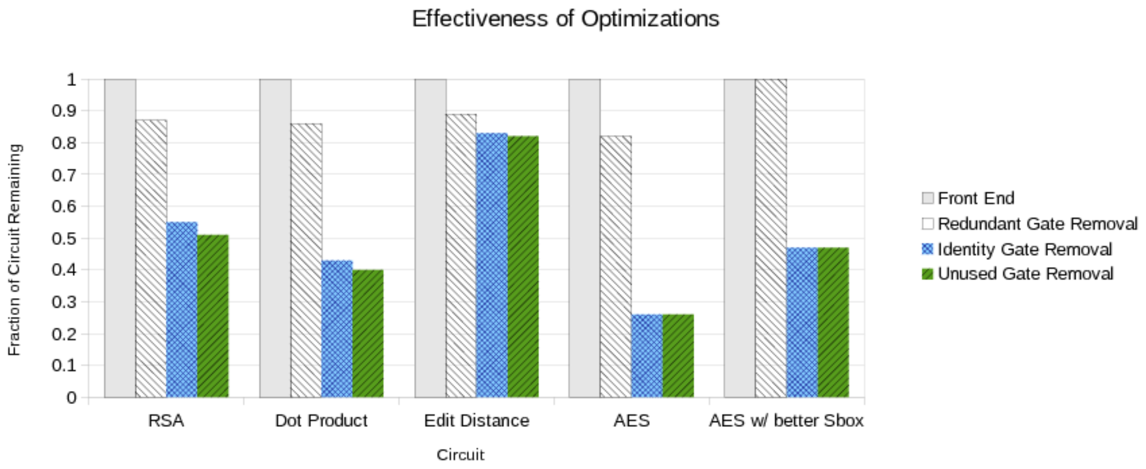

Figure 1: Average fraction of circuits remaining after each optimization is applied in sequence. We see that the *relative change* in circuit sizes after each optimization is dependent on the circuit itself, with some circuits being optimized more than others.

Circuit	DB (s)	mmap (s)	flat (s)
	7200RPM Spinning Disk (ext4-fs)		
AES	4.3 ±0.5%	1.05 ± 1%	3.48 ±0.3%
RSA-32	103 ±0.3%	24.6 ±0.2%	78.4 ±0.3%
Dot$_4^{64}$	32.56 ±0.1%	7.1 ±0.3%	28.37 ±0.1%
EDT-255	975 ±0.1%	240 ± 1%	700 ±0.9%
	Solid-State Drive		
AES	3.62 ±0.3%	0.86 ± 1%	3.17 ±0.6%
RSA-32	96.5 ±0.2%	21.6 ±0.4%	68.3 ±0.3%
Dot$_4^{64}$	30.5 ±0.5%	6.27 ± 1%	25.9 ±0.2%
EDT-255	907 ±0.1%	200 ±0.4%	590 ± 1%
	Amazon EC2		
AES	5.56 ± 4%	1.12 ± 0%	7.11 ±0.3%
RSA-32	208 ±0.4%	45.7 ± 3%	240 ±0.1%
Dot$_4^{64}$	46.3 ±0.1%	9.2 ±0.2%	60.7 ±0.2%
EDT-255	2500 ± 1%	405 ±0.2%	2050 ±0.2%
Circuit Sizes			
AES	RSA-32	Dot$_4^{64}$	EDT-255
49,912	1,750,787	460,018	15,540,196

Table 2: Compile times for different storage systems for small circuits (sizes include input gates), using different storage media. Results are averaged over 30 experiments, with 95% confidence intervals. On EC2, a high-memory quadruple extra large instance was used.

Key-Value Stores Unfortunately, even though our compiler is more resource efficient than Fairplay, it still requires space that is linear in the size of the circuit. For very large circuits, circuits with billions of gates or more, this may exceed the amount of RAM that is available. Our compiler can make use of a computer's hard drive to store intermediate representations of circuits and information about how to remove gates from the circuit. We used memory-mapped I/O to reduce the impact this has on performance; however, our use of *mmap* and *ftruncate* is not portable, and so our system also supports using an unmapped file or Berkeley DB. Our tests revealed that, as expected, memory-mapped I/O achieves the highest performance, but that Berkeley DB is sometimes better than unmapped files on high-latency filesystems. A summary of the performance of each method on a variety of storage systems is shown in Table 2.

Using the hard drive in this manner, we were able to compile our largest circuits. The performance impact of writing to disk should not be understated; a several-billion-gate edit distance 4095x4095 circuit required more than 3 days to compile on an Amazon EC2 high-memory image, with 68 GB of RAM, one third of which was spent waiting on I/O. Note, however, that this is a one-time cost; a compiled circuit can be used in unlimited evaluations of a secure computation protocol.

5.2 Compiler Testing Methodology

We tested the performance of our compiler using five circuits. The first was AES, to compare our compiler with the Fairplay system. We also used AES with the compact S-Box description given by Boyar and Parelta [3], which results in a smaller AES circuit. We used an RSA

RSA Size	Circuit Size	Compile Time (s)	**Gates/s**	Edit-Dist Size	Circuit Size	Compile Time (s)	**Gates/s**
16	208,499	2.6 ± 7%	**80,000**	31x31	144,277	1.70 ±0.7%	**84,900**
32	1,750,787	21.6 ±0.4%	**81,100**	63x63	717,233	8.56 ±0.7%	**83,800**
64	14,341,667	189 ±0.3%	**75,900**	127x127	3,389,812	41.7 ±0.5%	**81,300**
128	116,083,983	1810 ±0.3%	**64,100**	255x255	15,540,196	200 ±0.4%	**77,700**

Table 3: Time required to compile and optimize RSA and edit distance circuits on a workstation with an Intel Xeon 5506 CPU, 8GB of RAM and a 160GB SSD, using the textbook modular exponentiation algorithm. Note that the throughput for edit distance is higher even for comparably sized circuits; this is because the front end generates a more efficient circuit without any optimization. Compile times are averaged over 30 experiments, with 95% confidence intervals reported.

signing circuit with various toy key sizes, up to 128 bits, to test our compiler's handling of large circuits; RSA circuits have cubic size complexity, allowing us to generate very large circuits with small inputs. We also used an edit distance circuit, which was the largest test case used by Huang et al. [13]; unlike the other test circuits, there is no multiplication routine in the inner loop of this function. Finally we used a dot product with error, a basic sampling function for the LWE problem, which is similar to RSA in creating large circuits, but also demonstrates our system's ability to handle large input sizes.

After compiling these circuits, we tested the correctness by first performing a direct, offline evaluation of the circuit, and comparing the output to a non-circuit implementation. We then compared the output of an online evaluation to the offline evaluation. Additionally, for the AES circuit, we compared the output of the circuit generated by our compiler to the output of a circuit generated using Fairplay. We tested all three key-value stores on a variety of file systems, including a fast SSD, a spinning disk, and an Amazon EC2 instance store, checking for correctness as described above in each case.

5.3 Summary of Compiler Performance

Our compiler is able to emit and optimize large circuits in relatively short periods of time, less than an hour for circuits with tens of millions of gates on an inexpensive workstation. In Figure 1 we summarize the effectiveness of the various optimization stages on different circuits; in circuits that involve multiplication in finite fields or modulo an integer, the identity gate removal step is the most important, removing more than half of the gates emitted by the front-end. The edit distance circuit is the best-case for our front end, as less than $1/5$ of the gates that are emitted can be removed by the optimizer. The throughput of our compiler is dependent on the circuit being compiled, with circuits which are more efficiently generated by the front-end being compiled faster; in Table 3 we compare the generation of RSA circuits to edit distance circuits.

6 Experimental Results

In this section, we give a detailed description of our system, upon which we have implemented various real world secure computation applications. The experimental environment is the Ranger cluster in the Texas Advanced Computing Center. Ranger is a blade-based system, where each node is a SunBlade x6240 blade running a Linux kernel and has four AMD Opteron quad-core 64-bit processors, as an SMP unit. Each node in the Ranger system has 2.3 GHz core frequency and 32 GB of memory, and the point-to-point bandwidth is 1 GB/sec. Although Ranger is a high-end machine, we use only a small fraction of its power for our system, only 512 out of 62,976 cores. Note that we use the PBC (Pairing-Based Cryptography) library [25] to implement the underlying cryptographic protocols such as oblivious transfers, witness-indistinguishable proofs, and so forth. However, moving to more modern libraries such as RELIC [31] is likely to give even better results, especially to those circuits with large input and output size.

System Setup In our system, both the generator and the evaluator run an equal number of processes, including a root process and many slave processes. A root process is responsible for coordinating its own slave processes and the other root process, while the slave processes work together on repeated and independent tasks. There are three pieces of code in our system: the generator, the evaluator, and the IP exchanger. Both the generator's and evaluator's program are implemented with Message Passing Interface (MPI) library. The reason for the IP exchanger is that it is common to run jobs on a cluster with dynamic working node assignment. However, when the nodes are dynamically assigned, the generator running on one cluster and the evaluator running on another might have a hard time locating each other. Therefore, a fixed location IP exchanger helps the match-up process as described in Figure 2. Our system provides two modes—the user mode and the simulation mode. The former works as mentioned above, and the latter simply spawns an even number of processes, half for the generator and the other half for the evaluator. The network match-up process is omitted in the latter mode to simplify the testing of this system.

To achieve a security level of 2^{-80}, meaning that a malicious player cannot successfully cheat with probability better than 2^{-80}, requires at least 250 copies of the garbled circuit [32]. For simplicity, we used 256 copies in our experiments, that is, security parameters $k = 80$ and $s = 256$. Each experiment was run 30 times (unless stated otherwise), and in the following sections we report the average runtime of our experiments.

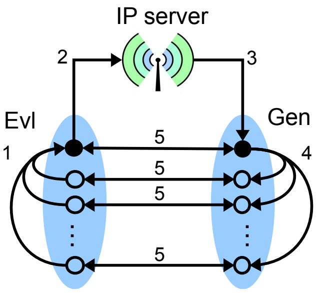

Figure 2: Both the generator and evaluator consist of a root process (solid dot) and a number of slave processes (hollow dots). The match-up works as follows: the slave evaluator processes send their IPs to the root evaluator process (Step 1), who then forwards them to the IP exchanger (Step 2). Next, the root generator process comes to acquire these IPs (Step 3) and dispatch them to its slaves (Step 4), who then proceed to pair up with one of the slave evaluator processes (Step 5) and start the main protocol. The arrows show the message flow.

Timing methodology When there is more than one process on each side, care must be taken in measuring the timings of the system. The timings reported in this section are the time required by the root process at each stage of the system. This was chosen because the root process will always be the *longest* running process, as it must wait for each slave process to run to completion. Moreover, in addition to doing all the work that the slaves do, the root processes also perform the input consistency check and the coin tossing protocol.

Impacts of the Performance Optimization Techniques We have presented several performance optimization techniques in Section 4 with theoretical analyses, and here we demonstrate their empirical effectiveness in Table 4. As we have anticipated, the Random Seed Checking reduces the communication cost for the garbled circuits by 60%, and the Garbled Row Reduction further reduces by another 25%. In the RS and GRR columns, the small deviation from the theoretical fraction 40% and 30%, respectively, is due to certain implementation needs. Our compiler is designed to reduce the number of non-XOR gates. In these four circuits, the ratio of non-XOR gates is less than 43%. So after further applying the Free-XOR technique, the final communication is less than 13% of that in the baseline approach.

	non-XOR (%)	Baseline (MB)	RS (%)	GRR (%)	FX (%)
AES	30.81	509	39.97	30.03	9.09
Dot_4^{64}	29.55	4,707	39.86	29.91	8.88
RSA-32	34.44	17,928	39.84	29.88	10.29
EDT-255	41.36	159,129	39.84	29.87	12.36

Table 4: The impact of various optimization techniques: The Baseline shows the communication cost for 256 copies of the original Yao garbled circuit when $k = 80$; RS shows the remaining fraction after Random Seed technique is applied; GRR shows when Garbled Row Reduction is further applied; and FX shows when the previous two techniques and the Free-XOR are applied. (The communication costs here only include those in the generation and evaluation stages.)

Performance Gain by AES-NI On a machine with 2.53 GHz Intel Core i5 processor and 4GB 1067 MHz DDR3 memory, it takes 784 clock cycles to run a single SHA-256 (with OpenSSL 1.0.0g), while it needs only 225 cycles for AES-256 (with AES-NI). To measure the benefits of AES-NI, we use two instantiations to construct various circuits, listed in Table 5, and observe a consistent 20% saving in circuit construction.[3]

	size (gate)	AES-NI (sec)	SHA-256 (sec)	Ratio (%)
AES	49,912	0.12± 1%	0.15± 1%	78.04
Dot_4^{64}	460,018	1.11±0.4%	1.41±0.5%	78.58
RSA-32	1,750,787	4.53±0.5%	5.9±0.8%	76.78
EDT-255	15,540,196	42.0±0.5%	57.6± 1%	72.92

Table 5: Circuit generation time (for a single copy) with different instantiations (AES-NI vs SHA-256) of the 2-circular correlation robust function.

AES We used AES as a benchmark to compare our compiler to the Fairplay compiler, and as a test circuit

[3] The reason that saving 500+ cycles does not lead to more improvements is that this encryption operation is merely one of the contributing factors to generating a garbled gate. Other factors, for example, include GNU `hash_map` table insertion (∼1,200 cycles) and erase (∼600 cycles).

for our system. We tested the full AES circuit, as specified in FIPS-197 [8]. In the semi-honest model, it is possible to reduce the number of gates in an AES circuit by computing the key schedule offline; e.g. this is one of the optimizations employed by Huang et al. [13]. In the malicious model, however, such an optimization is not possible; the party holding the key could attempt to reduce the security level of the cipher by computing a malicious key schedule. So in our experiments we compute the entire function, including the key schedule, online.

In this experiment, two parties collaboratively compute the function $f : (x, y) \mapsto (\perp, \mathrm{AES}_x(y))$, i.e., the circuit generator holds the encryption key x, while the evaluator has the message y to be encrypted. At the end, the generator will not receive any output, whereas the evaluator will receive the ciphertext $\mathrm{AES}_x(y)$.

Type	Fairplay	Ours-A	Pinkas et al.	Ours-B
non-XOR	15,316	15,300	11,286	9,100
XOR	35,084	34,228	22,594	21,628

Table 6: The components of the AES circuits from different sources. Ours-A comes from the textbook AES algorithm, and Ours-B uses an optimized S-box circuit from [3]. (Sizes do not include input or output wires)

First of all, we demonstrate the performance of our compiler in Table 6. We have shown in Section 5 that our compiler is capable of large circuit generation. We also found in our experiments that our compiler produces smaller AES circuit than Fairplay. Given the same high-level description of AES encryption (textbook AES), our compiler produces a circuit with a smaller gate count and even fewer non-XOR gates. When applying the compact S-Box description proposed by Boyar and Parelta [3] to the high-level description as input to our compiler, a smaller AES circuit than the hand-optimized one from Pinkas et al. is generated with less effort.

In Table 7, both the computational and communication costs for each main stage are listed under the traditional setting, where there is only one process on each side. These main stages include oblivious transfer, garbled circuit construction, the generator's input consistency check, and the circuit evaluation. Each row includes both the computation and communication time used. Note that network conditions could vary from setting to setting. Our experiments run in a local area network, and the data can only give a rough idea on how fast the system could be in an ideal environment. However, the precise amount of data being exchanged is reported.

We notice in Table 7 that the evaluator spends an unreasonable amount of time on communication with respect to the amount of data to be transmitted in both the oblivious transfer and circuit construction stages.

		Gen (sec)	Eval (sec)	Comm (KB)
OT	comp	45.8±0.09%	34.0±0.2%	5,516
	comm	0.1± 1%	11.9±0.6%	
Gen.	comp	35.6± 0.5%	–	3
	comm	–	35.6±0.5%	
Inp. Chk	comp	–	1.75±0.2%	266
	comm	–	–	
Evl.	comp	14.9± 0.6%	32.4±0.4%	28,781
	comm	18.2± 1%	3.2±0.8%	
Total	comp	96.3± 0.3%	68.0±0.2%	34,566
	comm	18.3± 1%	50.8±0.4%	

Table 7: The 95% two-sided confidence intervals of the computation and communication time for each stage in the experiment $(x, y) \mapsto (\perp, \mathrm{AES}_x(y))$.

This is because the evaluator spends that time waiting for the generator to finish computation-intensive tasks. The same reasoning explains why in the circuit evaluation stage the generator spends more time in communication than the evaluator. This waiting results from the fact that both parties need to run the protocol in a synchronized manner. A generator-evaluator pair cannot start next communication round while any other pair has not finished the current one. This synchronization is crucial since our protocol's security is guaranteed only when each communication round is performed sequentially. While the parallelization of the program introduces high performance execution, it does not and should not change this essential property. A stronger notion of security such as universal security will be required if asynchronous communication is allowed. By using TCP sockets in "blocking" mode, we enforce this communication round synchronization.

Note that the low communication during the circuit construction stage is due to the random seed checking technique. Also, the fact that the generator spends more time in the evaluation stage than she traditionally does comes from the second construction for evaluation circuits. Recall that only the evaluation circuits need to be sent to the evaluator. Since only 40% of the garbled circuits (102 out of 256) are evaluation-circuits, the ratio of the generator's computation time in the generation and evaluation stage is 35.63:14:92 $\simeq$ 5:2.

We were unfortunately unable to find a cluster of hundreds of nodes that all support AES-NI. Our experimental results, therefore, do not show the full potential of all the optimization techniques we have proposed. However, recall that for certain circuits the running time in the semi-honest setting is roughly half of that in the

node #	4		16		64		256	
	Gen	Evl	Gen	Evl	Gen	Evl	Gen	Evl
OT	12.56±0.1%	8.41±0.1%	4.06±0.1%	2.13±0.2%	1.96±0.1%	0.58±0.2%	0.64±0.1%	0.19±0.2%
Gen.	8.18±0.4%	–	1.92±0.7%	–	0.49±0.4%	–	0.14± 1%	–
Inp. Chk	–	0.42± 4%	–	0.10± 10%	–	–	–	–
Evl.	3.3± 4%	7.08± 1%	0.80± 10%	1.58± 4%	0.23± 17%	0.37± 7%	0.12±0.5%	0.05±0.6%
Inter-com	4± 5%	13.2±0.3%	0.93± 10%	4.08±0.8%	0.31± 20%	1.98± 1%	0.11± 40%	0.72±0.2%
Intra-com	0.17± 30%	0.23± 20%	0.18± 8%	0.25± 6%	0.45± 20%	0.48± 15%	0.34± 30%	0.34± 30%
Total time	28.3±0.3%	29.4±0.3%	7.90±0.5%	8.17±0.4%	3.45± 2%	3.44± 2%	1.4± 10%	1.3± 9%

Table 8: The average and error interval of the times (seconds) running AES circuit. The number of nodes represents the degree of parallelism on each side. "–" means that the time is smaller than 0.05 seconds. Inter-com refers to the communication between the two parties, and intra-com refers to communication between nodes for a single party.

malicious setting. We estimate a 20% improvement in the performance of garbled circuit generation when the AES-NI instruction set becomes ubiquitous, based on the preliminary results presented above in Table 5.

Table 8 shows that the Yao protocol really benefits from the circuit-level parallelization. Starting from Table 7, where each side only has one process, all the way to when each side has 256 processes, as the degree of parallelism is multiplied by four, the total time reduces into a quarter. Note that the communication costs between the generator and evaluator remain the same, as shown in Table 7. It may seem odd that the communication costs are *reduced* as the number of processes increase. The real interpretation of this data is that as the number of processes increases, the "waiting time" decreases.

Notice that as the number of processes increases, the ratio of the time the generator spends in the construction and evaluation stage decreases from 5:2 to 1:1. The reason is that the number of garbled circuit each process handles is getting smaller and smaller. Eventually, we reach the limit of the benefits that the circuit-level parallelism could possibly bring. In this case, each process is dealing with merely a single copy of the garbled circuit, and the time spent in both the generation and evaluation stages is the time to construct a garbled circuit.

To the best of our knowledge, completing an execution of secure AES in the malicious model within 1.4 seconds is the best result that has ever been reported. The next best result from Nielsen et al. [28] is 1.6 seconds, and it is an amortized result (85 seconds for 54 blocks of AES encryption in parallel) in the random oracle model. This is only a crude comparison, however; our experimental setup uses a cluster computer while Nielsen et al. used only two desktops. A better comparison would be possible given a parallel implementation of Nielsen et al.'s system, and we are interested in seeing how much of an improvement such an implementation could achieve.

Large Circuits In this experiment, we run the 4095-bit edit distance circuit, that is, $(x,y) \mapsto (\perp, \mathrm{EDT}(x,y))$, where $x,y \in \{0,1\}^{4095}$. In particular, we use the $I+C$ approach, where the computation time could be roughly a half of that of the $I+2C$ approach with the price of not getting to use the random-seed technique. Recall that in the $I+C$ approach, the generator and the evaluator conduct the cut-and-choose in a way that the generator does not know the check circuits until she finishes transferring all the garbled circuits. Next, both the parties run the circuit generation and evaluation in a pipeline manner, where one party is generating and giving away garbled gates on one end, and the other party is evaluating and checking the received gates at the other end at the same time. The results are shown in Table 9.

	Gen (sec)	Eval (sec)	Comm (Byte)
OT	19.73±0.5% 1.1± 6%	5.26±0.4% 15.6±0.6%	1.7×10^8
Cut-& Choose	1.1±0.8% –	– 1.5± 2%	6.5×10^7
Gen./Evl.	24,400± 1% 4,900± 1%	14,600± 3% 14,700± 2%	1.8×10^{13}
Inp. Chk	0.6± 20% 0.4± 40%	– 0.60± 20%	8.5×10^6
Total	24,400± 1% 4,900± 1%	14,600± 3% 14,700± 2%	1.8×10^{13}

Table 9: The result of $(x,y) \mapsto (\perp, \text{EDT-4095}(x,y))$. Each party is comprised of 256 cores in a cluster. This table comes from 6 invocations of the system. Similarly, the upper row in each stage is the computation time, while the lower is the communication time.

This circuit generated by our compiler has 5.9 billion gates, and 2.4 billion of those are non-XOR. It is worth mentioning that, without the random-seed technique, the communication cost shown in Table 9 can also be estimated by $256 \times 2.4 \times 10^9 \times 3 \times 10 = 1.8 \times 10^{13}$, since 256 copies of the garbled circuits need to be transferred, each copy has 2.4 billion non-free gates, each non-free gate has three entries, and each entry has $k = 80$ bits.

In additional to showing that our system is capable of handling the largest circuits ever reported, we also have shown a speed in the malicious setting that is comparable to those in the semi-honest setting. In particular, we were able to complete an single execution of 4095-bit edit distance circuit in less than 8.2 hours with a rate of 82,000 (non-XOR) gates per second. Note that Huang et al.'s system is the only one, to the best of our knowledge, that is capable of handling such large circuits [13]; they reported a rate of over 96,000 (non-XOR) gates per second for an edit-distance circuit in the semi-honest setting.

7 Conclusion

We have presented a general purpose secure two party computation system which offers security against malicious adversaries and which can efficiently evaluate circuits with hundreds of millions and even billions of gates on affordable hardware. Our compiler can generate large circuits using fewer computational resources than similar compilers, and offers improved flexibility to users of the system. Our evaluator can take advantage of parallel computing resources, which are becoming increasingly common and affordable. As future work, we plan further improvements to our compiler and language, as well as experiments on systems other than Ranger.

The source code for this system can be downloaded from the authors' website (http://crypto.cs.virginia.edu/), along with example functions, including those describe in this paper.

8 Acknowledgements

We would like to thank Benny Pinkas, Thomas Schneider, Nigel Smart and Stephen Williams for providing us with a copy of their optimized AES circuit. We would also like to thank Gabriel Robins for his advice on minimizing circuits in VLSI systems. We are particularly grateful to Ian Goldberg for his very helpful comments.

This work is supported by Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under contract FA8750-11-2-0211. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US government.

References

[1] Bendlin, R., Damgård, I., Orlandi, C., and Zakarias, S. Semi-homomorphic encryption and multiparty computation. In *Proceedings of the 30th Annual international conference on Theory and applications of cryptographic techniques: advances in cryptology* (Berlin, Heidelberg, 2011), EUROCRYPT'11, Springer-Verlag, pp. 169–188.

[2] Bogetoft, P., Christensen, D. L., Damgård, I., Geisler, M., Jakobsen, T. P., Krøigaard, M., Nielsen, J. D., Nielsen, J. B., Nielsen, K., Pagter, J., Schwartzbach, M. I., and Toft, T. Secure Multiparty Computation Goes Live. In *Financial Cryptography* (2009), pp. 325–343.

[3] Boyar, J., and Peralta, R. A new combinational logic minimization technique with applications to cryptology. In *Proceedings of the 9th international conference on Experimental Algorithms* (Berlin, Heidelberg, 2010), SEA'10, Springer-Verlag, pp. 178–189.

[4] Brickell, J., and Shmatikov, V. Privacy-preserving graph algorithms in the semi-honest model. In *Proceedings of the 11th international conference on Theory and Application of Cryptology and Information Security* (Berlin, Heidelberg, 2005), ASIACRYPT'05, Springer-Verlag, pp. 236–252.

[5] Canetti, R., Lindell, Y., Ostrovsky, R., and Sahai, A. Universally composable two-party and multi-party secure computation. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing* (New York, NY, USA, 2002), STOC '02, ACM, pp. 494–503.

[6] Choi, S. G., Katz, J., Kumaresan, R., and Zhou, H.-S. On the security of the "free-xor" technique. In *Proceedings of the 9th international conference on Theory of Cryptography* (Berlin, Heidelberg, 2012), TCC'12, Springer-Verlag, pp. 39–53.

[7] Damgard, I., Pastro, V., Smart, N., and Zakarias, S. Multiparty Computation from Somewhat Homomorphic Encryption. In *Proceedings of the 32th Annual International Cryptology Conference on Advances in Cryptology* (2012), CRYPTO '12. http://eprint.iacr.org/2011/535.

[8] FIPS. *Advanced Encryption Standard (AES)*, 2001.

[9] Garey, M., Graham, R., Johnson, D., and Knuth, D. Complexity results for bandwidth minimization. *SIAM Journal on Applied Mathematics 34*, 3 (1978), 477–495.

[10] Gentry, C., Halevi, S., and Smart, N. P. Homomorphic Evaluation of the AES Circuit. In *Proceedings of the 32th Annual International Cryptology Conference on Advances in Cryptology* (2012), CRYPTO '12. http://eprint.iacr.org/2012/099.

[11] Goyal, V., Mohassel, P., and Smith, A. Efficient two party and multi party computation against covert adversaries. In *Proceedings of the theory and applications of cryptographic techniques 27th annual international conference on Advances in cryptology* (Berlin, Heidelberg, 2008), EUROCRYPT'08, Springer-Verlag, pp. 289–306.

[12] Henecka, W., K ögl, S., Sadeghi, A.-R., Schneider, T., and Wehrenberg, I. Tasty: tool for automating secure two-party computations. In *Proceedings of the 17th ACM conference on Computer and communications security* (New York, NY, USA, 2010), CCS '10, ACM, pp. 451–462.

[13] Huang, Y., Evans, D., Katz, J., and Malka, L. Faster secure two-party computation using garbled circuits. In *Proceedings of the 20th USENIX conference on Security* (Berkeley, CA, USA, 2011), SEC'11, USENIX Association, pp. 35–35.

[14] Huang, Y., Malka, L., Evans, D., and Katz, J. Efficient Privacy-Preserving Biometric Identification. In *NDSS'11* (2011).

[15] ISHAI, Y., KILIAN, J., NISSIM, K., AND PETRANK, E. Extending Oblivious Transfers Efficiently. In *CRYPTO'03*, vol. 2729 of *LNCS*. Springer Berlin / Heidelberg, 2003, pp. 145–161.

[16] ISHAI, Y., PRABHAKARAN, M., AND SAHAI, A. Founding cryptography on oblivious transfer — efficiently. In *Proceedings of the 28th Annual conference on Cryptology: Advances in Cryptology* (Berlin, Heidelberg, 2008), CRYPTO 2008, Springer-Verlag, pp. 572–591.

[17] JHA, S., KRUGER, L., AND SHMATIKOV, V. Towards practical privacy for genomic computation. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy* (Washington, DC, USA, 2008), SP '08, IEEE Computer Society, pp. 216–230.

[18] KIRAZ, M. *Secure and Fair Two-Party Computation*. PhD thesis, Technische Universiteit Eindhoven, 2008.

[19] KIRAZ, M., AND SCHOENMAKERS, B. A Protocol Issue for The Malicious Case of Yao's Garbled Circuit Construction. In *27th Symposium on Information Theory in the Benelux* (2006).

[20] KOLESNIKOV, V., AND SCHNEIDER, T. Improved garbled circuit: Free xor gates and applications. In *Proceedings of the 35th international colloquium on Automata, Languages and Programming, Part II* (Berlin, Heidelberg, 2008), ICALP '08, Springer-Verlag, pp. 486–498.

[21] LINDELL, Y., OXMAN, E., AND PINKAS, B. The IPS Compiler: Optimizations, Variants and Concrete Efficiency. In *CRYPTO'11* (2011), pp. 259–276.

[22] LINDELL, Y., AND PINKAS, B. Privacy preserving data mining. In *Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology* (London, UK, UK, 2000), CRYPTO '00, Springer-Verlag, pp. 36–54.

[23] LINDELL, Y., AND PINKAS, B. An efficient protocol for secure two-party computation in the presence of malicious adversaries. In *Proceedings of the 26th annual international conference on Advances in Cryptology* (Berlin, Heidelberg, 2007), EUROCRYPT '07, Springer-Verlag, pp. 52–78.

[24] LINDELL, Y., AND PINKAS, B. Secure two-party computation via cut-and-choose oblivious transfer. In *Proceedings of the 8th conference on Theory of cryptography* (Berlin, Heidelberg, 2011), TCC'11, Springer-Verlag, pp. 329–346.

[25] LYNN, B. Pairing-Based Cryptography Library, 2006. http://crypto.stanford.edu/pbc/.

[26] MALKA, L. Vmcrypt: modular software architecture for scalable secure computation. In *Proceedings of the 18th ACM conference on Computer and communications security* (New York, NY, USA, 2011), CCS '11, ACM, pp. 715–724.

[27] MOHASSEL, P., AND FRANKLIN, M. Efficiency tradeoffs for malicious two-party computation. In *Proceedings of the 9th international conference on Theory and Practice of Public-Key Cryptography* (Berlin, Heidelberg, 2006), PKC'06, Springer-Verlag, pp. 458–473.

[28] NIELSEN, J. B., NORDHOLT, P. S., ORLANDI, C., AND BURRA, S. S. A New Approach to Practical Active-Secure Two-Party Computation. In *Proceedings of the 32th Annual International Cryptology Conference on Advances in Cryptology* (2012), CRYPTO '12. http://eprint.iacr.org/2011/091.

[29] OSADCHY, M., PINKAS, B., JARROUS, A., AND MOSKOVICH, B. Scifi - a system for secure face identification. In *Proceedings of the 2010 IEEE Symposium on Security and Privacy* (Washington, DC, USA, 2010), SP '10, IEEE Computer Society, pp. 239–254.

[30] PINKAS, B., SCHNEIDER, T., SMART, N. P., AND WILLIAMS, S. C. Secure two-party computation is practical. In *Proceedings of the 15th International Conference on the Theory and Application of Cryptology and Information Security: Advances in Cryptology* (Berlin, Heidelberg, 2009), ASIACRYPT '09, Springer-Verlag, pp. 250–267.

[31] RELIC. http://code.google.com/p/relic-toolkit/.

[32] SHELAT, A., AND SHEN, C.-H. Two-output secure computation with malicious adversaries. In *Proceedings of the 30th Annual international conference on Theory and applications of cryptographic techniques: advances in cryptology* (Berlin, Heidelberg, 2011), EUROCRYPT'11, Springer-Verlag, pp. 386–405.

[33] YAO, A. C. Protocols for secure computations. In *Proceedings of the 23rd Annual Symposium on Foundations of Computer Science* (Washington, DC, USA, 1982), SFCS '82, IEEE Computer Society, pp. 160–164.