# Scaling ORAM for Secure Computation

Jack Doerner
Northeastern University
j@ckdoerner.net

abhi shelat
Northeastern University
abhi@neu.edu

## ABSTRACT

We design and implement a Distributed Oblivious Random Access Memory (DORAM) data structure that is optimized for use in two-party secure computation protocols. We improve upon the access time of previous constructions by a factor of up to ten, their memory overhead by a factor of one hundred or more, and their initialization time by a factor of thousands. We are able to instantiate ORAMs that hold $2^{34}$ bytes, and perform operations on them in seconds, which was not previously feasible with any implemented scheme.

Unlike prior ORAM constructions based on hierarchical hashing [19], permutation [19], or trees [39], our Distributed ORAM is derived from the new Function Secret Sharing scheme introduced by Boyle, Gilboa and Ishai [11, 12]. This significantly reduces the amount of secure computation required to implement an ORAM access, albeit at the cost of $O(n)$ efficient *local* memory operations.

We implement our construction and find that, despite its poor $O(n)$ asymptotic complexity, it still outperforms the fastest previously known constructions, Circuit ORAM [42] and Square-root ORAM [55], for datasets that are 32 KiB or larger, and outperforms prior work on applications such as *stable matching* [16] or *binary search* [23] by factors of two to ten.

## 1 INTRODUCTION

In spite of the substantial improvements to the efficiency of two-party secure computation protocols, they still encounter major obstacles when evaluating many types of functions. In particular, functions that make *data-dependent* accesses to memory remain difficult cases. A data-dependent memory access is an access to an element within an array, at an index $i$ that is computed from some secret input. A secure computation protocol must guarantee that no information about its inputs is leaked to either party, even via intermediate computations, and thus it must be able to execute such memory accesses without leaking any bits of $i$.

Data-dependent memory accesses are common even in textbook algorithms; they are required by, for example, binary search, most graph algorithms, sparse matrix methods, greedy algorithms, and dynamic programming algorithms. More generally, they are required by any program that is written in the RAM model of computation. Any attempt to evaluate such an algorithm in a secure context upon a large dataset certainly requires an efficient data-dependent memory access mechanism.

The simplest solution to this problem is the *linear scan* technique, which hides the index of an accessed element by touching every element in the memory and using multiplexers to ensure that only the desired element is actually read or written. This effectively ensures data-obliviousness, but it requires an expensive secure computation involving $O(n)$ gates for each individual memory access. With accesses incurring overhead linear in the size of the entire memory, scanning is impractical for all but the smallest amounts of data.

Another solution is *Oblivious Random Access Memory* (ORAM). Intuitively, ORAM is a technique to transform a memory access to a *secret* index $i$ into a sequence of memory accesses that can be revealed to an adversary, the indices of which appear independent of $i$. ORAM was first proposed by Goldreich and Ostrovsky in their seminal paper [19], which studied the general context of client-server memory outsourcing. In this setting, a client wishes to perform a computation on a database of size $n$, which is held by some untrusted server, but does not want the server to learn the semantic pattern of accesses to the database. Goldreich and Ostrovsky proposed two schemes to solve this problem, the second of which requires that the client perform $O(\text{polylog}\, n)$ accesses to the database for every access in the client's original program. In the subsequent two decades, ORAM techniques have been widely studied [7, 13, 14, 18, 20–22, 29, 33–35, 38, 40, 45–47] with the goals of reducing the communication overhead between the client and server, reducing the amount of memory required of the client, and reducing the server's overall memory overhead. State of the art approaches to ORAM design limit the overhead in all of these measures to $O(\log^c n)$ where $c \leq 3$.

ORAM can be applied to the domain of secure computation by implementing ORAM client operations as secure functions, while the mutually-untrusting computation parties share the role of the ORAM server. This arrangement was proposed by Ostrovsky and Shoup [34], who used it to show that secure computations need not take time linear in the size of their input. It was later taken up by Gordon *et al.* [23]. Subsequently, the development of secure-computation-specific ORAMs began.

Wang *et al.* [43] observed that memory and communication overhead, the metrics for which ORAM had traditionally been optimized, were inappropriate for the context of secure computation. They proposed that *circuit complexity* is a more relevant measure, and described a heuristic ORAM based on this idea. Subsequently, Wang *et al.* [42] proposed Circuit ORAM, which offers asymptotically strong parameters for a data-structure with small circuit complexity.

Zahur *et al.* [55] observed that by relaxing asymptotic bounds, it is possible to produce a scheme that has a smaller concrete circuit size. They described a modification of the original Goldreich-Ostrovsky Square-root ORAM that is asymptotically inferior to Circuit ORAM, but outperforms it for data sizes up to 4 MiB.


*CCS '17, October 30-November 3, 2017, Dallas, TX, USA*

ACM ISBN 978-1-4503-4946-8/17/10…$15.00
https://doi.org/10.1145/3133956.3133967

Although they represent a dramatic improvement over initial efforts, the ORAM constructions of Gordon *et al.*, Zahur *et al.*, and Wang *et al.* suffer drawbacks. For instance, they are all recursively structured. That is, accessing the top level ORAM data structure for $n$ elements requires recursively accessing another ORAM data structure of size $n/8$ elements, and so on, each layer adding a communication round. As a result, each semantic access requires accessing $O(\log n)$ different ORAM layers, incurring $O(\log n)$ rounds of communication and latency.

These constructions also have high concrete memory overhead, due in part to their recursive nature and to the fact that they store *wire labels* for each bit of their memory, each wire label being at least 80 times larger than the data it represents. All prior research efforts of which we are aware report on concrete experiments that involve at most $2^{20}$ elements. In our own experiments, we confirm that the constructions they describe cannot handle more elements in a reasonable amount of time and space.[1]

The last, and possibly most significant problem is initialization. In many cases, an ORAM must be filled with some initial data before it can be used. Circuit ORAM requires an individual write into each element, a process that is extremely expensive: we observed it to require more than 3000 seconds for a moderately-sized memory of $2^{15}$ elements.[2] Zahur *et al.*'s Square-root ORAM is asymptotically similar, but uses a permutation network [41] instead of individual writes to achieve a constant-factor improvement of roughly 100. Nevertheless, even for moderately-sized memories, initialization is a significant cost.

These bottlenecks limit the use of secure computation protocols mostly to data-independent algorithms (e.g. AES [36], edit distance [44], or linear regression [32]) or RAM programs that exploit specific algorithmic properties to restrict their access patterns (e.g. BFS [6], Dijkstra's algorithm [27], or stable matching [16]).

## 1.1 Contributions

We propose a new data structure that addresses the drawbacks discussed previously, and we demonstrate the first concrete secure computation memory implementation that is capable of hosting data at the scale of many gigabytes. Our scheme has faster access times than all prior constructions for memories that are larger than 32 KiB, and, as it does not have any recursive components, each access requires only three rounds in principle. Unlike prior ORAMs, our data structure supports read and write operations independently, and can perform read operations substantially faster. Instead of storing wire labels, we store either XOR-shares or encryptions of the data, and thereby reduce the memory overhead to a small constant. Additionally, we have a linear-time method to fill our structure with initial data that requires no secure computation. As a result, an instance with $2^{20}$ 4-byte elements can be initialized in 400 milliseconds, roughly 1000 times faster than the best prior initialization technique from Zahur *et al.*'s Square-root ORAM [55]. We show that our advantages hold not only in microbenchmarks, but also in previously-published application contexts such as binary search and stable matching.

In contrast to most prior secure computation ORAM research, we consider the *Distributed* ORAM model [31], and derive our scheme from two-server Private Information Retrieval (PIR) techniques. In PIR, a client wishes to retrieve an element $A^i$ at index $i$ in database $A$, copies of which are held by two servers. The client issues a query $q_1(i)$ to server 1 and query $q_2(i)$ to server 2, and the servers respond with short messages $m_1$ and $m_2$ respectively, which the client can use to reconstruct $A^i$. PIR schemes must satisfy two properties: the total communication between client and servers must be sub-linear in $n$, and the query $q_p(i)$ in isolation must reveal no information about $i$.

Gilboa and Ishai [17] and Boyle, Gilboa, and Ishai [11] recently presented a surprisingly efficient PIR construction that is based on the notion of a *function secret sharing* (FSS) scheme for a *distributed point function* (DPF). Their construction offers properties new to PIR which make it well-suited for use in an ORAM for secure computation. In particular, it produces a query message of size $O(\log n)$, as opposed to the size of $O(n^{1/3})$ required by many PIR schemes [50], and it requires only a cryptographic pseudo-random generator, whereas other PIR schemes with logarithmic query size require public key cryptography. We discuss the specifics of this primitive in Section 2. In our construction, the parties to the secure computation, Alice and Bob, also act as the two servers in the PIR scheme, and secure computation performs the role of the client. Owing to the efficiency of FSS, our ORAM requires a very small secure computation in comparison to prior ORAM designs (up to one hundred times smaller for the memory sizes that we explore).

The second novel property offered by Boyle *et al.*'s PIR scheme is support for "PIR-writing", which we use to implement ORAM write operations, in combination with a standard *stash* data structure that retains updated elements until they can be reintegrated into the ORAM's main memory. The secure computation needed to implement the stash has an amortized computation and communication complexity of $O(\sqrt{n})$ per access; however, as demonstrated by Zahur *et al.* [55], even schemes with a complexity of $O(\sqrt{n \log^3 n})$ can outperform poly-logarithmic schemes in practice. Our stash reintegration procedure is related to our initialization procedure, and similarly requires linear time with no secure computation.

The *theoretical* disadvantage of our PIR-derived ORAM stems from the fact that the servers in a PIR scheme (i.e., Alice and Bob, in our case) must perform $O(n)$ local computation. This is an unavoidable property of any PIR system. However, unlike the $O(n)$ secure computation required by a traditional linear scan, this computation is simple, highly parallelizable, and enjoys widespread hardware-acceleration support. In practice, secure computation protocols are typically bottlenecked by network or single-core CPU performance and utilize a very small portion of the total computational power and memory bandwidth available with modern hardware; thus, the approach of replacing secure computation with asymptotically-worse local computation can yield significant performance improvements. Despite the poor theoretical complexity of our scheme, we show via a concrete implementation that it outperforms all prior ORAMs, even for large datasets.

[1] Wang *et al.* [42] report on an instance of Circuit ORAM storing $2^{30}$ 4-byte elements using an older implementation of Circuit ORAM that stores its data as XOR-shares instead of wire labels, but they do not report concrete performance figures for that size. In this paper we evaluate the faster implementation reported by Zahur *et al.* [55]; with this implementation, an instance of Circuit ORAM larger than 64 MiB exhausts the 122 GiB of memory in each of our two test machines.

[2] See Figure 8d

Due to the heavy influence of the FSS scheme and the fact that the computation parties make local linear scans of the memory for each operation, we call our ORAM construction Function-secret-sharing Linear ORAM, or *Floram*.

As with most prior ORAM research, our implementation is in the honest-but-curious adversarial setting. We conjecture that our scheme can be hardened more easily than others due to its simplicity, but we leave that question for future work.

*Organization.* The remainder of the paper is organized as follows: In Section 2, we review definitions of techniques we use, including ORAM and the recently developed technique of Function Secret Sharing. In Section 3 we construct simple single-function ORAMs based upon FSS, and analyze their properties, and in Section 4 we combine and extend these constructions to yield a fully functional ORAM. In Section 5 we present a technique for outsourcing the FSS computation that yields a significant practical speed increase over a naïve implementation. Finally, in Section 6, we describe an implementation of our scheme and evaluate its performance. In the full version of this document [25] we give formal definitions and security proofs.

## 2 BACKGROUND

*Secure Multi-party Computation.* The field of Secure Multi-Party Computation (MPC) studies mechanisms by which a group of individuals, each individual $i$ having some secret input $x_i$, can evaluate a function $y = f(x_1, x_2, \ldots)$ jointly, in such a way that no party $i$ learns anything other than what is revealed by the output $y$ and their private input $x_i$. Specifically, party $i$ must neither learn any $x_j$ for all $j \neq i$, nor any intermediate value derived from $x_j$ during the evaluation of $f$. A special case of MPC is Two-Party Computation (2PC), in which only two parties, Alice and Bob, participate. Though many variations of MPC have been developed in its thirty-plus year history, and it is likely possible to adapt our work to suit a significant subset of them, this paper focuses on Yao's Garbled Circuits [51, 52].

Yao's Garbled Circuits conforms to the *honest-but-curious* or *semi-honest* security model, in which Alice and Bob are trusted to follow the protocol instructions, but are curious adversaries who may attempt to learn each others' secrets by analyzing protocol transcripts. Outside observers may also analyze protocol transcripts, but must learn nothing in so doing. Selective security for Yao's Garbled Circuits in this model has been proven by Lindell and Pinkas [30], and adaptive security by Jafargholi and Wichs [26]. We provide a standard security definition in the full version of this document [25].

*Oblivious RAM.* ORAM [19] is a data structure that provides the familiar semantics of random access memory, but translates the logical access instructions it receives into sequences of physical accesses in such a way that no adversary can recover the logical accesses by observing the physical access patterns. An ORAM must support the functions $\mathsf{Read}(i)$ and $\mathsf{Write}(i, v)$, which perform semantic reads and writes to locations specified by a private index $i$. An ORAM may also support functions $\mathsf{Apply}(f, i, v)$, which applies some function privately to a single location, and $\mathsf{Init}(V)$, which fills the ORAM with data from the array $V$.

As traditionally defined, an ORAM must satisfy the security property that, for any two sequences of logical accesses of the same length, transcripts of the physical accesses produced must be indistinguishable. We concern ourselves with a variant, *Distributed* Oblivious RAM (DORAM) [31], which considers the context wherein the underlying memory is split among multiple parties, and which satisfies a slightly weaker security property: for any two sequences of logical accesses of the same length, transcripts of the physical accesses performed by any *single* party must be indistinguishable. Intuitively, no party may learn anything about the semantic memory by observing their own share of the physical memory. We provide formal definitions for DORAM in the full version of this document [25].

ORAMs are traditionally considered to have some manner of secure CPU that transforms semantic memory accesses into physical ones. In the setting of MPC, the CPU is typically implemented as a multiparty protocol. Thus, in some sense, all ORAMs become DORAMs when applied to MPC: the constructions as wholes can be only as secure as the MPC protocols that implement their CPUs, and no protocol can be secure when all participants are corrupt. For simplicity, we refer to our scheme as an ORAM, except where the distinction is important.

*Function Secret Sharing.* Secret Sharing [37] allows a *dealer* to divide a secret value into $m$ shares, one for each of $m$ parties, such that none of the parties can individually gain any insight into the secret value, yet all $m$ shares, as a group, contain enough information to reconstruct it. Recently, Gilboa and Ishai [17] observed that it is possible to secret-share a point function using shares with sizes sublinear in the size of the function's domain; they call this concept a *Distributed Point Function* (DPF). Boyle *et al.* [11, 12] subsequently improved upon this work and described how to construct a two-server PIR scheme using a DPF. We begin by formally defining a Function Secret Sharing Scheme for two parties.

*Definition 2.1 (Point Function).* A point function is a function $f_{\alpha,\beta} : [1, n] \rightarrow G$ such that

$$f_{\alpha,\beta}(x) = \begin{cases} \beta & \text{if } x = \alpha \\ 0 & \text{otherwise} \end{cases}$$

*Definition 2.2 (Function Secret Sharing Scheme for Point Functions [11, 17]).* A two-party function secret sharing scheme is a pair of Probabilistic Polynomial Time algorithms (Gen, Eval) of the following form

(1) $\mathsf{Gen}(1^\lambda, (\alpha, \beta))$ is a key generation algorithm, which on input $1^\lambda$ (a security parameter), and a description of a point function function $f_{\alpha,\beta}$, outputs a tuple of keys $(k_a^{\mathsf{FSS}}, k_b^{\mathsf{FSS}})$.
(2) $\mathsf{Eval}(k_p^{\mathsf{FSS}}, x)$ is an evaluation algorithm, which on input $k_p^{\mathsf{FSS}}$ (party key share for party $p \in \{a, b\}$), and evaluation point $x \in [1, n]$, outputs a group element $y_p^x \in G$ and a bit $t_p^x \in \{0, 1\}$ such that $y_p^x = f_p(x)$ (party $p$'s share of $f(x)$) and $t_p^x$ is a share of 0 if $f(x) = 0$, or a share of 1 otherwise.

*Definition 2.3 (Security for an FSS Scheme for Point Functions).* A two-party FSS for point functions is secure if

(1) (Correctness) For all point functions $f_{\alpha,\beta}$, and for every $x \in [1, n]$ in the domain of $f_{\alpha,\beta}$

$$(k_a^{\text{FSS}}, k_b^{\text{FSS}}) \leftarrow \text{Gen}(1^\lambda, (\alpha, \beta)) \implies \Pr\left[\text{Eval}(k_a^{\text{FSS}}, x) - \text{Eval}(k_b^{\text{FSS}}, x) = f(x)\right] = 1$$

(2) (Privacy) For every corrupted party $p$ (either $a$ or $b$), and every sequence of point function descriptions $f_1, f_2, \ldots$, there exists a simulator Sim such that:

$$\left\{(k_a^{\text{FSS}}, k_b^{\text{FSS}}) \leftarrow \text{Gen}(1^\lambda, f_\lambda) : k_p^{\text{FSS}}\right\}_{\lambda\in\mathbb{N}} \stackrel{c}{\equiv} \left\{\text{Sim}(p, 1^\lambda)\right\}_{\lambda\in\mathbb{N}}$$

In other words, the simulator can produce a share (without knowing the function) that is indistinguishable from the real share for the function. Thus, the function share leaks nothing about $f_{\alpha,\beta}$ other than its domain and the group that contains its range.

We summarize the FSS construction of a distributed point function $f_{\alpha,\beta}$ from Boyle *et al.* [11, 12] in Figure 1. The $\text{Gen}(1^\lambda, (\alpha, \beta))$ method produces shares $k_a^{\text{FSS}}, k_b^{\text{FSS}}$ of the point function $f_{\alpha,\beta}$. These shares consist of one private seed each ($s_a, t_a$ and $s_b, t_b$ respectively), and the rest of the information in the share is the same for both parties. The FSS scheme follows a tree-based PRF construction, wherein each node of the tree is associated with a seed, and a pseudo-random generator (PRG) is used to double the seed into two seeds, one for the left child, and one for the right. At each level $j$ of the tree, Alice and Bob will have exactly the same seed for all nodes except for the node along the path from the root to the leaf $\alpha$. At this node, Alice and Bob have different seeds, $s_a^{j,\alpha_j}$ and $s_b^{j,\alpha_j}$ respectively, and thus the expansion of their seeds result in different seeds for the children of this node at level $j+1$, $s_a^{j+1,0}, s_a^{j+1,1}$ and $s_b^{j+1,0}, s_b^{j+1,1}$. The scheme provides a correction word $\sigma^j$ and two advice bits, $\tau^{j,0}$ and $\tau^{j,1}$, for each level. $\sigma^j$ is conditionally applied to both child seeds of a node according to $t^j = \text{Lsb}(s_p^{j,\alpha_j}) \oplus t^{j-1} \cdot \tau^{j,\alpha_j}$. This modifies the child seeds such that afterward, Alice and Bob share the same seed for all nodes except for the node along the path to leaf $\alpha$. That is, of the two children of each node along the path to leaf $\alpha$, for which Alice and Bob's seeds differ, one is "deactivated" (i.e. Alice and Bob's seeds at that position are made identical), and the other is not. This correction is performed in such a way that neither party can determine which branch has been deactivated.

A Private Information Retrieval (PIR) system is a mechanism by which a client may retrieve an item from a database replicated among some number of servers, without revealing to any server which item was retrieved. Though similar to ORAMs, PIR systems are notably distinct: they typically do not concern themselves with writing or with hiding the contents of the memory from the servers, they do not require any initialization or allow reorganization of the database, and they do not incur memory overheads for the client or servers. On the other hand, PIR schemes take for granted that servers must perform $O(n)$ work for each access, whereas ORAM literature has hitherto focused on providing sublinear-in-$n$ computation complexity. When combined with memory encryption, a PIR scheme may be thought of as an Oblivious Read-only Memory (OROM), and we show how to construct such a primitive from FSS in Section 3.

1 **function** $\text{Gen}(1^\lambda, \alpha = \alpha_m \ldots \alpha_2\alpha_1, \beta)$:
2 $s'^0_a, s'^0_b \stackrel{\$}{\leftarrow} \{0,1\}^\lambda$ // pick random seeds
3 $t^0_a, t^0_b \leftarrow$ a random xor−share of 1
4 **for** $j \in [1, m]$:
5 $\left\{\left(s_p^{j,0} \middle\| s_p^{j,1}\right)\right\}_{p\in\{a,b\}} \leftarrow \left\{\text{Prg}\left(s_p'^{j-1}\right)\right\}_{p\in\{a,b\}}$
6 $\sigma^j \leftarrow s_a^{j,\overline{\alpha_j}} \oplus s_b^{j,\overline{\alpha_j}}$ // xor off-path children
7 $\tau^{j,0} \leftarrow \text{Lsb}\left(s_a^{j,0}\right) \oplus \text{Lsb}\left(s_b^{j,0}\right) \oplus \alpha_j \oplus 1$
8 $\tau^{j,1} \leftarrow \text{Lsb}\left(s_a^{j,1}\right) \oplus \text{Lsb}\left(s_b^{j,1}\right) \oplus \alpha_j$
9 $\left\{s_p'^j\right\}_{p\in\{a,b\}} \leftarrow \left\{s_p^{j,\alpha_j} \oplus t_p^{j-1} \cdot \sigma^j\right\}_{p\in\{a,b\}}$
10 $\left\{t_p^j\right\}_{p\in\{a,b\}} \leftarrow \left\{\text{Lsb}\left(s_p^{\alpha_j}\right) \oplus t_p^{j-1} \cdot \tau^{j,\alpha_j}\right\}_{p\in\{a,b\}}$
11 $\gamma \leftarrow s_a^m \oplus s_b^m \oplus \beta$
12 $k_a^{\text{FSS}} \leftarrow \left(s'^0_a, t^0_a, \{\sigma^j, \tau^{j,0}, \tau^{j,1}\}_{j\in[1,m]}, \gamma\right)$
13 $k_b^{\text{FSS}} \leftarrow \left(s'^0_b, t^0_b, \{\sigma^j, \tau^{j,0}, \tau^{j,1}\}_{j\in[1,m]}, \gamma\right)$
14 **return** $k_a^{\text{FSS}}, k_b^{\text{FSS}}$
15
16 **function** $\text{Eval}(k_p^{\text{FSS}}, x = x_m \ldots x_2x_1)$
17 // Parse key $k_p^{\text{FSS}}$ as $(s_p^0, t_p^0, \{\sigma^j, \tau^{j,0}, \tau^{j,1}\}_{j\in[1,m]}, \gamma)$
18 **for** $j \in [1, m]$:
19 $\left(s^{j,0} \middle\| s^{j,1}\right) \leftarrow \text{Prg}\left(s'^{j-1}\right)$
20 $s'^j \leftarrow s^{j,x_j} \oplus t^{j-1} \cdot \sigma^j$
21 $t^j \leftarrow \text{Lsb}\left(s^{j,x_j}\right) \oplus t^{j-1} \cdot \tau^{j,x_j}$
22 $y \leftarrow s'^m \oplus t^m \cdot \gamma$
23 **return** $y, t^m$

**Figure 1: Pseudocode for the Function Secret Sharing scheme. Our design follows Boyle *et al.* [11, 12].**

## 3 SINGLE-FUNCTION MEMORY

We begin by explaining how to construct write-only and read-only random access memories from the FSS scheme described in Section 2. The constructions presented here may be independently useful in scenarios wherein simultaneous read and write capabilities are not needed; we combine them into a full ORAM in Section 4.

*Oblivious Write-Only Memory.* We first construct an Oblivious Write-Only Memory (OWOM), based on the folkloric technique of *PIR-writing*. Both parties hold a local XOR-share of each memory location; in order to write to a location $i$ (this index being given as private data within the MPC protocol), the secure computation must determine the difference, $v^\Delta$, between the value already stored there and the value to be written. It must then use the FSS scheme to construct a distributed point function that evaluates to 0 everywhere except location $i$, whereat the DPF evaluates to $v^\Delta$. Alice and Bob individually evaluate their shares of the DPF, and add these shares into the memory-shares that they hold. Because they are adding shares of zero at all locations other than $i$, those values remain unchanged. At index $i$, they add shares of the difference between the old and new values to shares of the old value, producing shares of the value that was to be written.

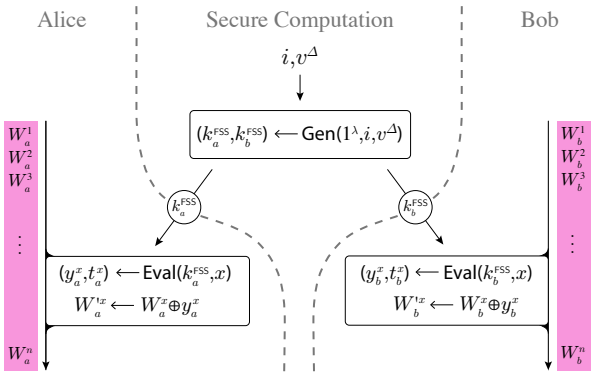



**Figure 2: Diagram of Oblivious Write-only Memory. To perform a write, the secure computation generates shares of a DPF, $k_a^{\text{FSS}}$ and $k_b^{\text{FSS}}$, which are distributed to Alice and Bob. Alice and Bob each evaluate the DPF at every value $x \in [1, n]$ and XOR the result into their respective corresponding shares of the OWOM memory.**

More precisely, we represent the value at memory location $i$ as $W^i$, and party $p$'s share as $W_p^i$, where $W^i = W_a^i \oplus W_b^i$. To write value $W'^i$ into the memory, the secure computation calculates $v^{\Delta} = W^i \oplus W'^i$ and then $(k_a^{\text{FSS}}, k_b^{\text{FSS}}) \leftarrow \text{Gen}(1^{\lambda}, (i, v^{\Delta}))$, delivering $k_a^{\text{FSS}}$ to Alice and $k_b^{\text{FSS}}$ to Bob, who use these keys to derive $(y_p^x, t_p^x) \leftarrow \text{Eval}(k_p^{\text{FSS}}, x)$ for all $x \in [1, n]$. For the purpose of writing, the parties will ignore $t_p^x$ and use the main DPF output $y_p^x$, which they XOR into the underlying memory to perform the write, $W_p'^x \leftarrow W_p^x \oplus y_p^x$.

Because write operations are performed by *cumulatively* XORing adjustment values with each $W^i$, it is necessary to write the difference between the old and new values, rather than writing the new value directly. In absence of any mechanism for reading (or otherwise determining which values are currently stored), this limits our OWOM to use only in write-only, write-once situations. However, it will become a building block for a full ORAM in the next section. We depict this scheme in Figure 2.

*Oblivious Read-Only Memory.* We implement read-only memory in a manner similar to classic PIR constructions. Alice and Bob, in their roles as the PIR servers, each hold identical copies of the memory, masked by the output of a pseudo-random function (PRF) using a key $k^{\text{PRF}}$ that is known to the secure computation, but not to Alice or Bob individually. To read an element $R^i$ from the memory at a private index $i$ (again, this index is given as private data within the protocol), Alice and Bob engage in a secure computation protocol to calculate $(k_a^{\text{FSS}}, k_b^{\text{FSS}}) \leftarrow \text{Gen}(1^{\lambda}, (i, \beta))$. Each party receives a $k_p^{\text{FSS}}$ and uses it to calculate $(y_p^x, t_p^x) \leftarrow \text{Eval}(k_p^{\text{FSS}}, x)$ for all $x \in [1, n]$. Although the DPF $y_p^x$ may have an arbitrary range $\beta$, for the purpose of reading, it is necessary that they hold a DPF of magnitude 1. Thus, the parties will use the final advice bits, $t_p^x$, which essentially represent the same DPF normalized to $\{0, 1\}$. Both parties compute $v_p = \bigoplus_x t_p^x \cdot R^x$. According to the properties of our FSS scheme, since $t_a^x = t_b^x$ for all $x \neq i$, it follows that $v_a \oplus v_b = R^i$. Finally, Alice and Bob use a secure computation to evaluate $\text{Prf}_{k^{\text{PRF}}}(i) \oplus R^i$, effectively importing the semantic value of interest into the secure computation. We depict this scheme in Figure 3.

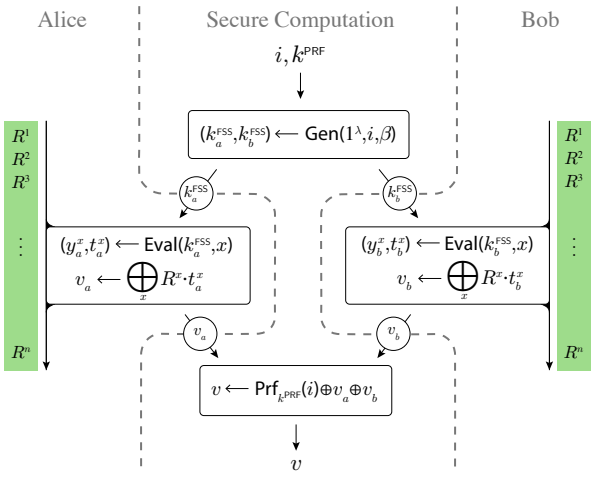



**Figure 3: Diagram of Oblivious Read-Only Memory. To perform a read, the secure computation generates shares of a DPF, $k_a^{\text{FSS}}$ and $k_b^{\text{FSS}}$, which are distributed to Alice and Bob. Alice and Bob each evaluate a normalized version of the DPF at every value $x \in [1, n]$, calculate the dot product of the normalized DPF with their respective copies of the OROM memory, and feed the result back into the secure computation to compute the value $v$ at location $i$.**

Though this scheme permits an unlimited number of reads, it cannot be written. Each party stores a PRF-masked copy (i.e. an encryption) of the data rather than a secret share: were any single memory location to be changed by a write, the access pattern would be revealed; on the other hand, if *all* memory locations were changed during a write, the semantic values of those not being updated must be destroyed.

*Complexity Analysis.* For both schemes, the secure FSS component (which forms the bulk of the secure computation) is identical. The computation of $\text{Gen}(1^{\lambda}, (\alpha, \beta))$ requires $4 \log_2(n)$ evaluations of the PRG function, along with some basic boolean operations. It must be seeded with random data of length $O(\lambda)$, and it produces an output of size $O(\lambda \log n)$ where $\lambda$ is the security parameter. This output can be revealed to the computation parties all at once, or incrementally, in $\log n$ chunks of $\lambda$ bits, one for each layer of the FSS scheme. In the former case, the secure component incurs a memory complexity of $O(\lambda \log n)$ and $O(1)$ communication rounds. In the latter case, the secure component incurs a memory complexity of $O(\lambda)$, and no additional rounds, as the secure computation does not need to wait for replies. In either case, the communication and computation complexities are $O(\lambda \log n)$.

Subsequently, a local computation is required to construct the DPF, $(y_p^x, t_p^x) \leftarrow \text{Eval}(k_p^{\text{FSS}}, x)$ for all $x \in [1, n]$. If all $n$ FSS evaluations are combined into a single operation, then the FSS tree can be constructed in its entirety only once, requiring $O(n)$ PRG calls. In the case of a write, each of the $n$ elements in the output DPF's domain must be XORed into the corresponding memory location; in the case of a read, the dot product of the DPF and the memory must be taken instead. In either case, this incurs $O(n)$ memory accesses. All of the operations performed by the local FSS evaluation and the application of the output DPF are highly parallelizable. We make

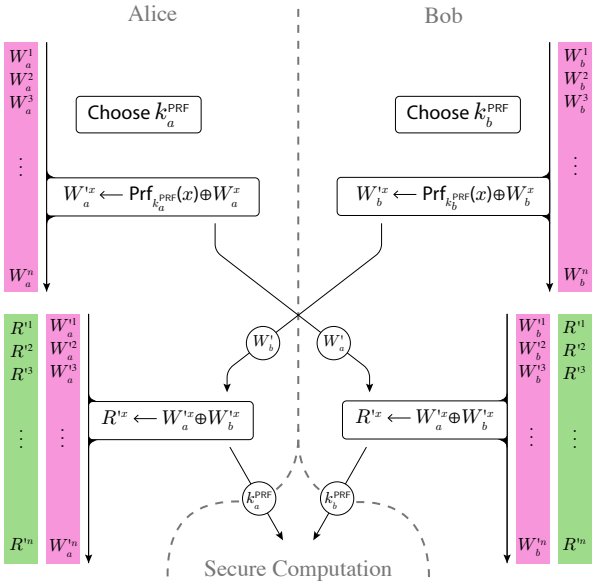



**Figure 4: Diagram of the Floram Refresh method. In addition to the operations illustrated here, the secure computation must clear the stash.**

extensive use of this fact in our concrete implementation, and in Section 6 we show experimentally that the local component does not become a significant burden until the amount of data stored is on the order of hundreds of megabytes.

## 4 READING AND WRITING

We now combine the OWOM and OROM from Section 3 into an ORAM construction. We need a few building blocks in order to make this combination possible, and conjecture that these building blocks are sufficient for the combination of any PIR and PIR-writing schemes into an ORAM, assuming that the schemes themselves are suitable (that is, their access patterns and underlying memory formats are secure).

At a high level, the construction works as follows: we initialize both an OROM and an OWOM with the same data, and create a linear-scan *stash* that stores elements while they are waiting to be returned to the main memory. *Read* operations are performed by inspecting both the stash and the OROM, and returning the most recent data. *Write* operations are performed by first reading the current value at the specified index, using it to calculate the difference necessary to correctly update the OWOM, and finally writing the new value into both the OWOM and the stash. When the stash fills, we perform a refresh operation to convert the OWOM memory into OROM memory, and then clear the stash. The cost of this refresh can be amortized over the refresh period of the construction. Because we use this stash-and-refresh technique, our amortized secure computation complexity becomes $O(\sqrt{n})$.

*Refresh Procedure.* To refresh our ORAM construction, we need to convert the underlying memory of an OWOM into the underlying memory of an OROM. The former stores its data as XOR-shares, while the latter uses a masked copy of the data as the underlying format. We can avoid incurring any secure computation overhead at all if, instead of masking the OROM memory only once, using a key known only to the secure computation, we mask it first with a key known only to Alice, and then with a key known only to Bob. To convert the OWOM into an OROM, Alice and Bob mask their local OWOM memory shares using two PRFs with individual secret keys, $k_a^{\mathsf{PRF}}$ and $k_b^{\mathsf{PRF}}$.

$$W'_p \leftarrow \left\{ W'^x_p \leftarrow \mathsf{Prf}_{k_p^{\mathsf{PRF}}}(x) \oplus W^x_p \right\}_{x \in [1,n]}$$

They each transmit their masked OWOM memory share to the other party, and both parties calculate

$$R' \leftarrow \left\{ R'^x \leftarrow W'^x_a \oplus W'^x_b \right\}_{x \in [1,n]}$$

Finally, each party feeds their key $k_p^{\mathsf{PRF}}$ into the secure computation, so that the OROM memory can be unmasked via $v \leftarrow \mathsf{Prf}_{k_a^{\mathsf{PRF}}}(x) \oplus \mathsf{Prf}_{k_b^{\mathsf{PRF}}}(x) \oplus R^x$. This refresh procedure is illustrated in Figure 4. Unlike previous Square-root ORAM constructions [19, 55], our refresh procedure does not require access to the stash. Instead, we simply clear it. Our stash serves only the purpose of allowing updated elements to be accessed multiple times between refreshes.

*Semi-private Access.* It may be the case that some algorithms call for both private (i.e. data-dependent) and data independent accesses to the same memory. Ostrovsky and Shoup refer to the latter type of accesses as *semi-private* [34]. To our knowledge, it has heretofore been necessary to implement *all* accesses as fully private accesses in such a scenario, or to perform costly import and export operations upon the entire ORAM. Floram, however, allows for a secondary, semi-private access mechanism, which has a significantly reduced asymptotic and practical cost. Unlike all other ORAMs of which we are aware, Floram stores each memory element at the physical address corresponding to its semantic index. Thus, to read the element at the publicly known semantic index $i$, the two parties feed their OWOM memory shares $W_a^i$ and $W_b^i$ into the secure computation, which computes the value $W^i$ in $O(1)$ complexity (and potentially using only free gates [28]). Semi-private writes must additionally append to the stash.

*Private Read Access.* Read operations that are *publicly known* to be read operations can also be performed without invoking the full-access mechanism: neither a write to the stash nor a write to the OWOM is required. Because no write to the stash is required, ORAM reads do not contribute to the refresh period.

*Full Private Access.* A full private access accepts some arbitrary oblivious function $f$ and applies it to a single element within the ORAM. $f$ takes an ORAM element and some auxiliary input $v^f$, and produces a new element and some auxiliary output $y^f$. We use this general-purpose mechanism to implement ORAM writes via simple $f_{\mathsf{write}}$ that returns $v^f$ as the output element. To perform a full access, our scheme first retrieves the desired element from the OROM, then scans the stash to determine whether a newer version of the same element exists. $f$ is then applied to it. Finally, the result is stored using an OWOM operation and appended to the stash. Because the OROM and OWOM access the same element, they can share a single FSS evaluation. This process is illustrated in Figure 5.

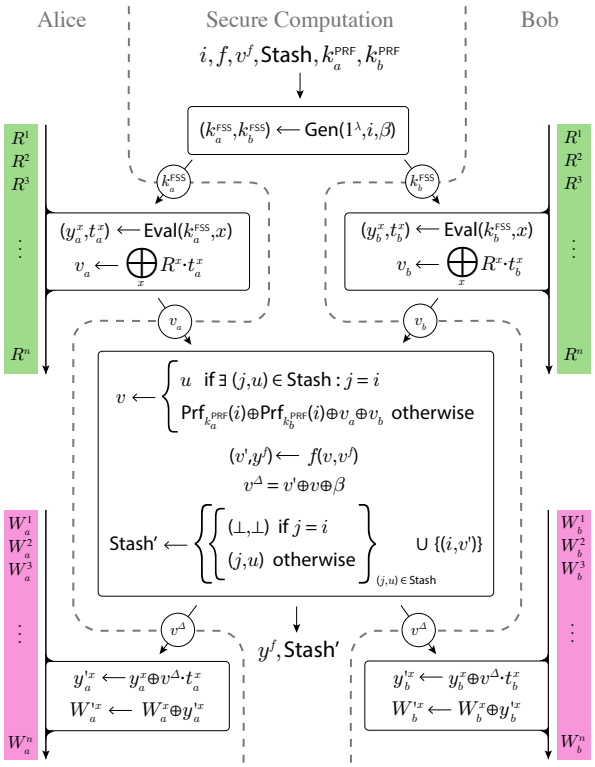



**Figure 5: Diagram of the Floram Access method. Note that $\beta$ is randomly chosen on each access.**

*Initialization.* The initialization of our ORAM can be performed efficiently using the mechanism for refreshing that we described earlier. That is, assuming that the parties begin with some secret sharing of the data values with which the ORAM is to be filled, they may initialize it by copying those shares into the OWOM's memory and performing a refresh. If the ORAM is hosted by a Yao's Garbled Circuits protocol, then the point-and-permute technique of Beaver *et al.* [4] can be used to encode XOR shares of the data within the protocol's wire labels, effectively making the generation of shares a free action. Furthermore, because this technique encodes the XOR sharing of each data bit only in the final bit of a much larger wire-label, it is actually a significant constant factor *faster* to initialize our ORAM than it is to perform a single linear scan on the same data. To our knowledge, this property is unique among all known ORAMs.

*Complexity Analysis.* If we briefly set aside the stash, the complexities of our scheme for full access to private indices closely follow the complexities of the individual components described in Section 3. That is, each access requires a single FSS Gen execution within the secure context, incurring $O(\log n)$ communication and secure computation, followed by the evaluation of the DPF at all points in its domain, incurring $O(n)$ local computation by both parties. This is in turn followed by a memory scan for the ROM component, adding a further $O(n)$ local computation, an unmasking within the secure computation context, which accounts for $O(1)$ communication and secure computation complexity, and a local memory scan for the WOM component, which incurs a further $O(n)$ local computation. Thus, still ignoring the stash, a standard access operation incurs $O(\log n)$ secure computation and communication overall, as well as $O(n)$ local computation.

The stash must be traversed on each access, and its length depends upon the refresh period of the ORAM. The refresh operation requires a simple masking (i.e. encryption), transmission, and element-wise XOR of $n$ memory elements by each of the two parties, without any secure computation. Thus the total cost of a refresh is $O(n)$ in terms of local computation and communication. This is optimally amortized over $O(\sqrt{n})$ accesses, and thus the cost of each access must include the cost of scanning $O(\sqrt{n})$ elements in the stash. The optimal constant can be determined by the relative costs of secure and local scans. Our concrete implementation uses a stash of size $\sqrt{n}/8$. A summary of these costs, along with comparisons to other ORAM schemes, is provided in Table 1.

The asymptotic complexity of our initialization procedure is $O(n)$ in terms of local computation, memory, and communication. Like the refresh procedure on which it is based, it requires no secure computation at all. This is optimal, at least from a complexity standpoint. Furthermore, as we shall see in Section 6, the practical costs of our initialization procedure are so low that it is actually faster in practice than a simple `memcpy` over the same data.

*Comparison to other ORAM schemes.* Our ORAM scheme stands in contrast to those that have preceded it in a number of respects, as summarized in Table 1. Here we discuss their implications. We focus primarily on the secure component of our scheme (which cannot be parallelized), and explore the practical consequences of the local component in Section 6. Although our ORAM uses a simple stash that incurs square-root overhead, it does not use recursive position maps or permutations required by Zahur *et al.*'s construction [55], nor does it need the sorting and binary searching required by the classic Goldreich and Ostrovsky construction [19]. Consequently, its optimal stash size is much smaller. Moreover, our scheme can be refreshed more efficiently than that of Zahur *et al.*, and much more efficiently than classic Square-root ORAM, which requires $O(n)$ encryptions within the secure context as well as an oblivious sort for each refresh operation. In previous Square-root ORAM constructions, stash scan and amortized refresh operations accounted for the vast majority of per-access cost; in having provided asymptotic improvements to both (as well as significant constant cost improvements), we have made our new ORAM far more suitable than its predecessors for handling large data sizes. On the other hand, our ORAM requires $O(\log n)$ calls to a PRG within the secure context for each access. Because these PRG calls are expensive, our ORAM is less suitable than that of Zahur *et al.* for small data sizes. In Section 5, we describe a method for reducing the number of secure PRG calls to $O(1)$ at the cost of incurring $O(\log n)$ communication rounds. This significantly improves our performance for small values of $n$, but for *very* small values, the construction of Zahur *et al.* remains more efficient in practice.

A comparison to Circuit ORAM (and other tree-based ORAMs) is somewhat less straightforward. Our ORAM enjoys an initialization procedure many orders of magnitude more efficient; however, in terms of access complexity, Circuit ORAM remains ahead. Nonetheless, as we shall discuss in Section 6, reduction in constant costs

| | Access | | | | Initialization | | | |
|---|---|---|---|---|---|---|---|---|
| | Floram | Florom | Square-root | Circuit | Floram | Florom | Square-root | Circuit |
| Secure Comp. | $O(\sqrt{n})$ | $O(\log n)$ | $O(\sqrt{n \log^3 n})$ | $O(\log^3 n)$ | – | – | $O(n \log^2 n)$ | $O(n \log^3 n)$ |
| Local Comp. | $O(n)$ | $O(n)$ | $O(\sqrt{n \log n})$ | $O(1)$ | $O(n)$ | $O(n)$ | $O(n \log n)$ | $O(1)$ |
| Communication | $O(\sqrt{n})$ | $O(\log n)$ | $O(\sqrt{n \log^3 n})$ | $O(\log^3 n)$ | $O(n)$ | $O(n)$ | $O(n \log^2 n)$ | $O(n \log^3 n)$ |
| Rounds | $O(1)$ | $O(1)$ | $O(\log n)$ | $O(\log n)$ | $O(1)$ | $O(1)$ | $O(\log n)$ | $O(n \log n)$ |

**Table 1: Access and Initialization Complexities. Complexities include amortized refresh operations where relevant. *Florom* refers an instantiation of Floram with a stash size of zero (i.e. one which has recently been refreshed); due to the fact that only writes increase the stash size, refreshes can be forced before long sequences of reads to achieve these complexities.**

renders our scheme far more efficient in practice. Boyle *et al.* [10] propose a parallelization method for tree-based ORAMs, from which it is possible to derive an initialization procedure that uses permutations in place of individual writes. With this mechanism, Circuit ORAMs could achieve initialization performance similar to that of Zahur *et al.*'s construction, at best.[3] Although the local component of our ORAM is highly parallelizable, no equivalent parallelization scheme for our secure component is possible.

Finally, it is worthwhile to acknowledge the distinctions between our scheme and the recent work of Abraham *et al.* [2], which also combined ORAM with PIR. Like Floram, their scheme is properly a *Distributed* ORAM, but in contrast, their scheme uses PIR to retrieve single elements along the branches of a larger recursive tree ORAM. Consequently, it shares more with Circuit ORAM and Onion ORAM [15] than it does with our scheme. They optimize for communication overhead, and their scheme achieves a communication complexity of $O(\log n)$ per access, which we can match only when no writes are performed. Furthermore, it is likely that PIR-server computation is significantly less burdensome in their scheme, since their PIR requires no PRG and is evaluated over only $O(\log n)$ elements. On the other hand, they primarily consider the outsourcing model, and do not account for costs in an MPC context. We find it likely[4] that these would be similar to Circuit ORAM.

*Security Analysis.* To argue that our scheme is semi-honest secure, we must present a simulator that produces a party's view of an ORAM operation (without receiving any information about other parties' private inputs) that is indistinguishable from the same party's view of the real ORAM operation. Simulators for access and initialization, along with proofs of computational indistinguishability, are presented in the full version of this document [25]. Informally, the security of our scheme follows from the security properties of the MPC technique chosen to host the construction and the security of the FSS scheme, which guarantees that the neither the FSS key share nor the output leaks any information about the associated point function, other than its domain and range. The underlying memory itself reveals nothing about its contents due to its mechanism of representation: each party views an OROM memory that is masked by the output of a PRF for which they key is not known, as well as an information-theoretically secure secret-share of an OWOM memory

*PRG and PRF.* Among several options for the PRG, we have chosen AES-128 [1]. Significant research effort has been put toward optimizing the boolean-circuit representation of AES [8, 49], and these optimizations have naturally been adapted for the context of secure computation [24]. Specifically, we use the AES S-box circuit of Boyar and Peralta [9], which requires less than 5000 non-free gates per block, and we accelerate local AES evaluations using Intel's AES-NI instruction set. In order to avoid the cost of repeated key expansion, we assume that AES satisfies the ideal cipher property and use the Davies-Meyer construction [48], with independent keys for left and right expansions in the FSS tree. We use AES in counter mode as the PRF that masks the OROM.

## 5 CONSTANT SECURE PRG EVALUATIONS

The costliest single component of our scheme is the repeated evaluation of the PRG function within the secure computation of the FSS Gen algorithm. In this section, we present an optimization that can be used to achieve a significant constant-factor speed improvement relative to a naïve implementation by *outsourcing* the evaluations of the PRG in the FSS Gen algorithm to Alice and Bob. That is, instead of Alice and Bob performing a *single* secure computation which uses $O(\log n)$ PRG expansions to compute their shares of the FSS key (line 5 in Figure 1), we instead divide Gen into $m = \log_2 n$ iterative computations that compute the FSS key one part at a time. Surprisingly, we can divide the computation in a manner that requires no PRG evaluations inside the secure computation, and that also maintains the security properties of the original.[5] Specifically, we devise an equivalent method of computing the value $\sigma^j$ (line 6 in Figure 1) that does not require the PRG to be evaluated in a secure computation. Hereafter, we refer to this as the *Constant PRG* or CPRG optimization.

Thus far, our FSS notation has only identified seeds $s_p^{j,\alpha_j}$ that are on the path from the root to the leaf $\alpha$ in the FSS evaluation tree. We now introduce notation to identify *all* of the nodes in the evaluation tree. Let $S_p^{j,\ell}$ denote the $\ell^{\text{th}}$ node from the left at level $j$ of player $p$'s FSS evaluation tree, where $p \in \{a, b\}$, $j \in [1, m]$, and

[3] This mechanism has not yet been implemented, so we cannot currently provide concrete data to support this claim.

[4] As we have no implementation of their scheme (MPC-oriented or otherwise), we cannot perform a practical evaluation.

[5] i.e., we will still be able to simulate the view of Alice or Bob given only the output of the function. Notice that we would not be able to simulate the view if our protocol simply asked Alice and Bob to evaluate line 5 in Figure 1.

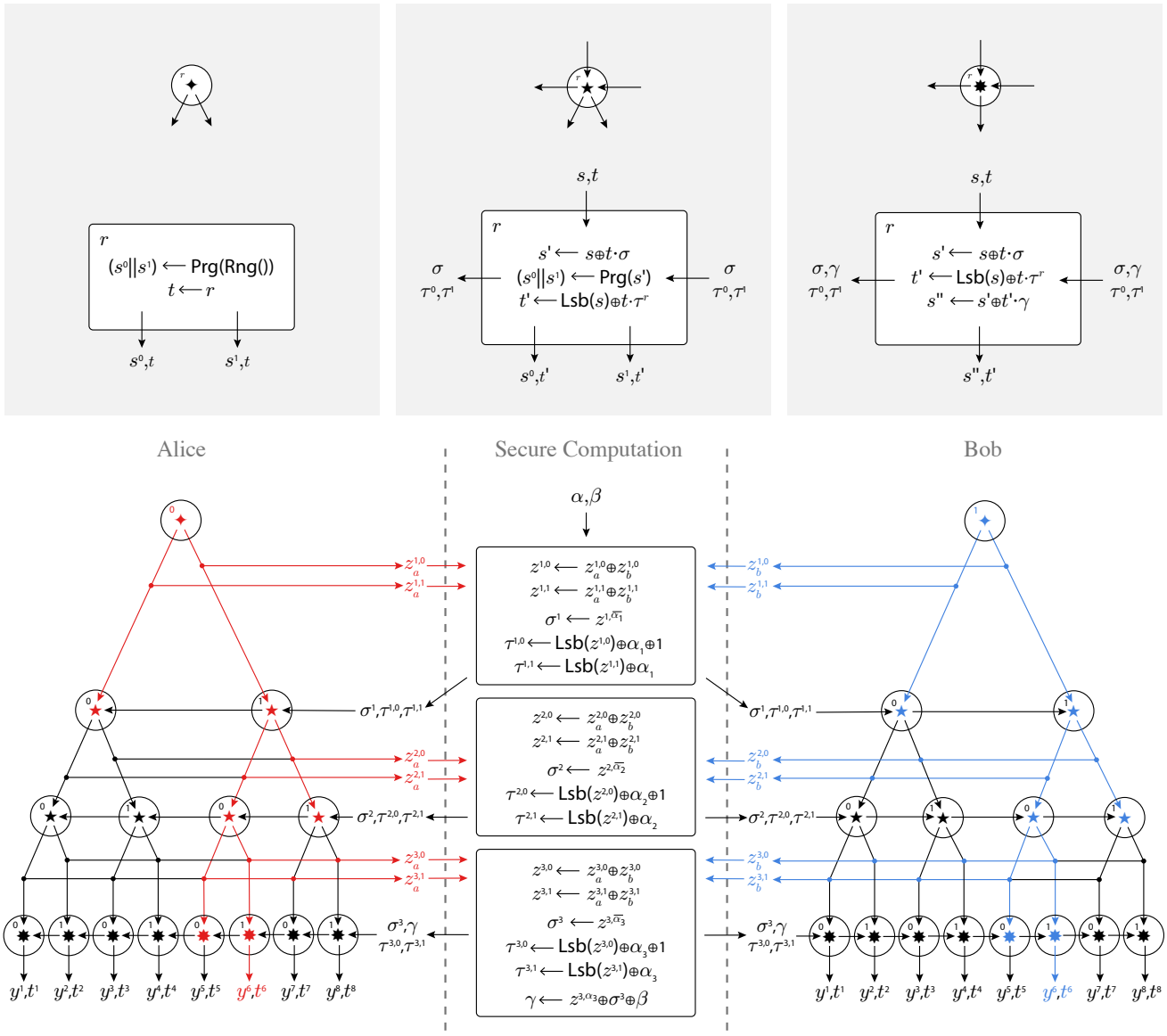



**Figure 6: Diagram of the modified Gen/Eval algorithm used by the CPRG optimization. Variables and processes for which Alice and Bob's views are identical are rendered in black. Variables and processes for which Alice and Bob's views differ are rendered in red for Alice and blue for Bob. In this example, $n = 8$, $m = 3$, and $\alpha$ is a three-bit number with value 6.**

$\ell \in [0, 2^j)$. Thus, seed $s_a^{j,\alpha_j}$ can also be identified as node $S_a^{j,\alpha_j^*}$ where $\alpha_j^*$ is the integer with the binary representation $\alpha_j \ldots \alpha_2\alpha_1$.

Next, we observe that the FSS construction guarantees that at any level $j$, $S_a^{j,\ell} = S_b^{j,\ell}$ for all $\ell \neq \alpha^*$ (that is, for all nodes *except* the one along the path to leaf $\alpha$), and $S_a^{j,\alpha_j^*} \neq S_b^{j,\alpha_j^*}$. It follows that all of the PRG expansions of the nodes at level $j$, i.e., the *uncorrected children* at level $j+1$, are equal except for the two children of the node along the path to $\alpha$. Finally, consider the sum of the PRG expansions of $S_p^{j,\ell}$ for $\ell \in [0, 2^j)$:

$$\left(z_p^{j+1,0} \middle\| z_p^{j+1,1}\right) = \bigoplus_{\ell \in [0,2^j)} \mathrm{Prg}\left(S_p^{j,\ell}\right)$$

From the above, we have:

$$z_a^{j,0} \oplus z_b^{j,0} = s_a^{j,0} \oplus s_b^{j,0}$$
$$z_a^{j,1} \oplus z_b^{j,1} = s_a^{j,1} \oplus s_b^{j,1}$$
$$\sigma^j = z_a^{j,\overline{\alpha_j}} \oplus z_b^{j,\overline{\alpha_j}}$$

Thus, we instruct Alice and Bob to locally compute $z_p^{j,0}$ and $z_p^{j,1}$ by accumulating the XOR of all left children and all right children at each level. These two values are submitted to a secure computation, which selects the correct sum using bit $\overline{\alpha_j}$, computes the next advice words $(\sigma^j, \tau^{j,0}, \tau^{j,1})$ and returns them to both parties. Both parties can then apply these values (per lines 9–10 in Figure 1) to

**function** $\mathsf{Gen}(1^\lambda, \alpha = \alpha_m \ldots \alpha_2\alpha_1, \beta)$:

$S'^{0,0}_a, S'^{0,0}_b \xleftarrow{\$} \{0,1\}^\lambda$ // pick random seeds

$t^{0,0}_a, t^{0,0}_b \leftarrow$ a random xor−share of 1

**for** $j \in [1, m]$:

**for** $p \in \{a, b\}$: // local computations

$$\left\{\left(S^{j,2\ell}_p \middle\| S^{j,2\ell+1}_p\right)\right\}_{\ell\in[0,2^{j-1})} \leftarrow \left\{\mathsf{Prg}\left(S'^{j-1,\ell}_p\right)\right\}_{\ell\in[0,2^{j-1})}$$

$$z^{j,0}_p \leftarrow \left(\bigoplus_{\ell\in[0,2^{j-1})} S^{j,2\ell}_p\right)$$

$$z^{j,1}_p \leftarrow \left(\bigoplus_{\ell\in[0,2^{j-1})} S^{j,2\ell+1}_p\right)$$

$\sigma^j \leftarrow z^{j,\overline{\alpha_j}}_a \oplus z^{j,\overline{\alpha_j}}_b$ // xor off-path children

$\tau^{j,0} \leftarrow \mathsf{Lsb}\left(z^{j,0}_a\right) \oplus \mathsf{Lsb}\left(z^{j,0}_b\right) \oplus \alpha_j \oplus 1$

$\tau^{j,1} \leftarrow \mathsf{Lsb}\left(z^{j,1}_a\right) \oplus \mathsf{Lsb}\left(z^{j,1}_b\right) \oplus \alpha_j$

**for** $p \in \{a, b\}$: // local computations

$$\left\{S'^{j,\ell}_p\right\}_{\ell\in[0,2^j)} \leftarrow \left\{S^{j,\ell}_p \oplus t^{j-1,\lfloor\ell/2\rfloor}_p \cdot \sigma^j\right\}_{\ell\in[0,2^j)}$$

$$\left\{t^{j,\ell}_p\right\}_{\ell\in[0,2^j)} \leftarrow \left\{\mathsf{Lsb}\left(S^{j,\ell}_p\right) \oplus t^{j-1,\lfloor\ell/2\rfloor}_p \cdot \tau^{j,\mathsf{Lsb}(\ell)}\right\}_{\ell\in[0,2^j)}$$

$\gamma \leftarrow z^{m,\alpha_m}_a \oplus z^{m,\alpha_m}_b \oplus \sigma^m \oplus \beta$

$k^{\mathsf{FSS}}_a \leftarrow \left(S'^{0,0}_a, t^{0,0}_a, \{\sigma^j, \tau^{j,0}, \tau^{j,1}\}_{j\in[1,m]}, \gamma\right)$

$k^{\mathsf{FSS}}_b \leftarrow \left(S'^{0,0}_b, t^{0,0}_b, \{\sigma^j, \tau^{j,0}, \tau^{j,1}\}_{j\in[1,m]}, \gamma\right)$

**return** $k^{\mathsf{FSS}}_a, k^{\mathsf{FSS}}_b$

**Figure 7: Pseudocode for the Constant PRG optimization applied to the FSS** Gen **method. This optimization is discussed in Section 5.**

generate the corrected seeds for all nodes at the next level, and then continue the process until level $m$. Revised pseudocode is presented in Figure 7. Although we model this function as returning a pair of key values $(k^{\mathsf{FSS}}_a, k^{\mathsf{FSS}}_b)$, note that most components of each party's key are revealed to them over the course of the function, and furthermore, that both parties will have had to perform most of the work of evaluating $\mathsf{Eval}(k^{\mathsf{FSS}}_p, x)$ for all $x \in [1, n]$ in order to calculate $(z^{j,0}_p, z^{j,1}_p)$. Consequently, in practice, the CPRG-optimized Gen algorithm returns only those key components that have not already been revealed, and Alice and Bob evaluate Eval simultaneously with the evaluation of Gen. This process is illustrated in Figure 6.

*Security Analysis.* Relative to the original Gen algorithm, nothing additional is revealed to either party, i.e., the output of the CPRG-optimized Gen is exactly the same, and the view of each party can be easily simulated with the final key. The only difference is that the advice strings included in the output key are revealed one by one. In the honest-but-curious setting that we consider here, the adversary has no additional power when receiving outputs in this manner.

*Efficiency Analysis.* The CPRG optimization requires no calls to the PRG function within the secure evaluation of Gen, and only two calls to the PRF to unmask the value retrieved from the OROM. We still perform $O(\log n)$ differencing and advice bit generation steps, but these require only a handful of gates each. On the other hand, our local stage now requires a reduction to be performed over all of the blocks in each layer of the FSS Eval algorithm. Consequently, this variant is significantly more efficient for small and medium sized memories, where secure computation dominates total runtime, but slightly less efficient for memories on the scale of gigabytes, as shown by our evaluations in Section 6.

# 6 EVALUATION

*Experimental Setup.* We implemented and benchmarked Floram, using Obliv-C [53], a C derivate that compiles and executes Yao's Garbled Circuits protocols [51] with many protocol-level optimizations [4, 5, 24, 28, 54]. Additionally, we made use of Obliv-C-based Square-root and Circuit ORAM implementations that were provided by the original authors of those works and are identical to the ones reported on previously by Zahur *et al.* [55].

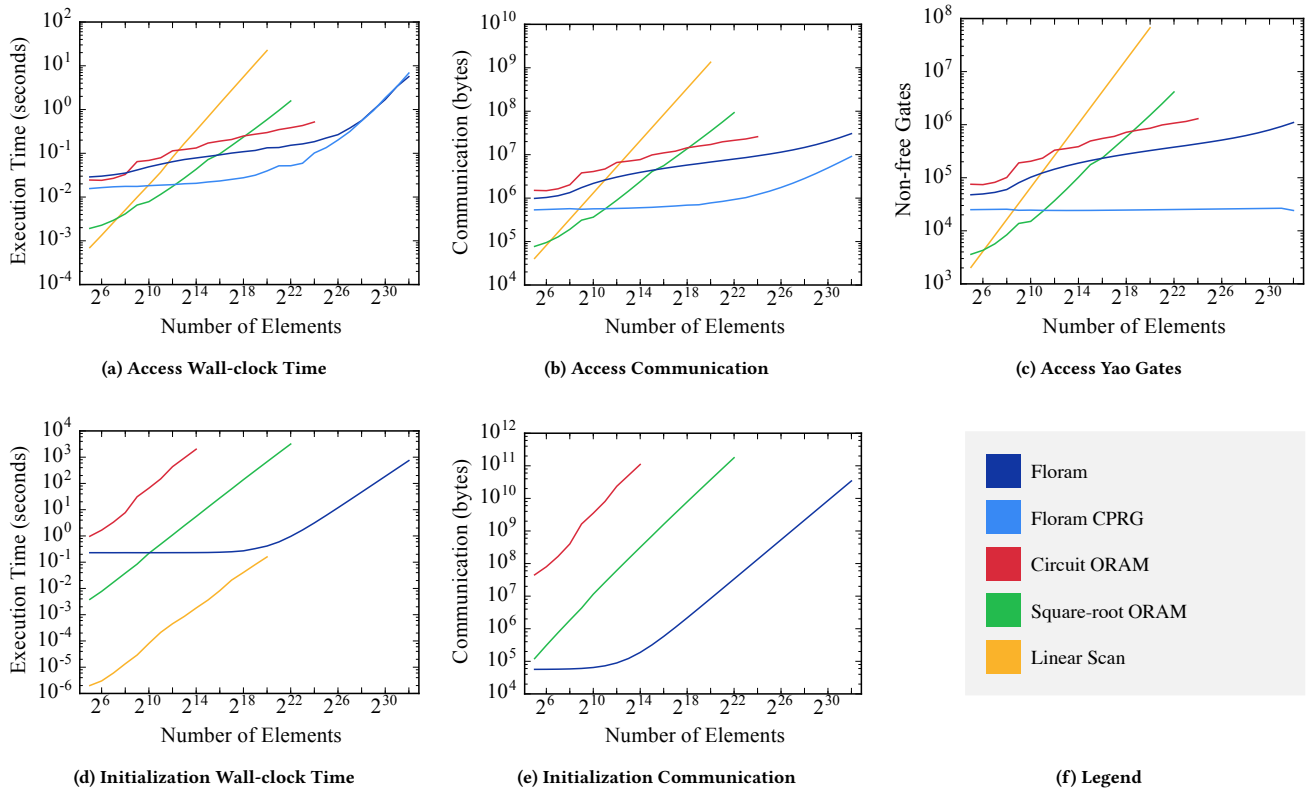
We created two variants of our ORAM, one using the basic construction described in Section 4, and the other using the CPRG method from Section 5. Both variants have optimized scheduling, as described in the full version of this document [25]. Our concrete implementation uses a 128 bit block size, this being the block size of AES-128, our chosen PRG function. For ORAMs with element sizes smaller than 128 bits, we pack multiple elements into a single block and linearly scan them. For ORAMS with element sizes greater than 128 bits, we perform an additional expansion and correction stage after the last layer of the FSS in order to enlarge the blocks to the correct length.

Our benchmarks were performed under Ubuntu 16.04 with Linux kernel 4.4.0 64-bit, running on a pair of identical Amazon EC2 `R4.4xlarge` instances. All code was compiled using `gcc` version 5.4.0, with the `-O3` flag enabled, OpenMP was used to manage multithreading and SIMD operations, and local AES computations were implemented using Intel's AES-NI instructions. Each machine had 122GB of DDR4 memory and eight physical cores partitioned from an Intel Xeon E5-2686 v4 CPU clocked at 2.3 GHz, each core being capable of executing two threads. We measured the bandwidth between our two instances to be roughly four gigabits per second. In order to ensure that the secure computation would be bandwidth-bound, as we would expect it to be in real-world conditions, we artificially restricted the bandwidth to 500 megabits per second, using the linux tool `tc`.

*Multithreading.* Our two Floram implementations make extensive use of multithreading for their local components, but we have not attempted to multithread their secure components, nor have we multithreaded the other ORAMs against which we make comparisons. Multithreading a secure computation does not reduce the total communication between parties, and thus in bandwidth-bound environments provides no advantage. Neither Square-root nor Circuit ORAM performs significant local computation, and so they cannot benefit significantly from local parallelism.

## 6.1 Full ORAM Microbenchmarks

*Full Access.* We performed single-access microbenchmarks for Floram, as well as Floram with the CPRG optimization discussed in Section 5. For the purpose of comparison, we also performed benchmarks for the Square-root ORAM of Zahur *et al.* [55], Circuit




(a) Access Wall-clock Time (b) Access Communication (c) Access Yao Gates

(d) Initialization Wall-clock Time (e) Initialization Communication (f) Legend

**Figure 8: Microbenchmark Results. Access figures are averages from at least 100 samples; for refreshing ORAMs, the sample count was a multiple of the refresh period. Initialization figures are averages from 30 samples. For all benchmarks, elements were 4 bytes in size.**

ORAM [42], and linear scan. For all ORAMs, we used an element sizes of 4 bytes. For linear scan, we varied the number of ORAM elements between $2^5$ and $2^{20}$, and for Square-root ORAM, between $2^5$ and $2^{22}$. In both cases, this is far past the range in which those schemes are competitive. For Circuit ORAM, we performed benchmarks with up to $2^{24}$ 4-byte elements, corresponding to 64 MiB of data; beyond this the ORAM's physical size was so large that it could not be instantiated on our machine. We benchmarked Floram with sizes up to $2^{32}$ 4-byte elements, corresponding to 16 GiB of data; these were the largest instances that our machine could handle. We recorded the wall-clock times for both parties, the number of bytes transmitted, and the number of non-free Yao gates executed. Our results are reported in Figures 8a, 8b, and 8c, respectively.

As we expected, the wall-clock time of our scheme exhibits a piecewise behavior. Up to roughly $2^{25}$ 4-byte elements, secure computation (specifically, the FSS Gen algorithm) dominates the total access time, and thus the time grows with $O(\log n)$—noticeably more slowly than any other ORAM. In this region, as expected, the CPRG optimization leads to a significant concrete performance gain, amounting to roughly a four-fold improvement. Beyond $2^{25}$ elements, local computation becomes the dominant factor, and thus the wall-clock time grows with $O(n)$ and the standard FSS scheme becomes more efficient. We estimate that the break-even point with Circuit ORAM lies at $2^{30}$ elements.

*Initialization.* We also performed initialization benchmarks. That is, beginning with an array of data, we evaluated each construction's native mechanism for importing that data into a fresh ORAM instance. As before, we varied the number of elements for linear scan between $2^5$ and $2^{20}$, and for Square-root ORAM between $2^5$ and $2^{22}$. Circuit ORAM has the slowest initialization process by several orders of magnitude, and so we benchmarked only up to $2^{14}$ elements, after which continuing was impractical. Both variants of Floram share the same initialization procedure, and we tested instances up to the largest size that our machines supported: $2^{32}$ 4-byte elements, or 16 GiB of data in total. Results for wall-clock time and total communication are reported in Figures 8d and 8e respectively; gate counts are not reported, as our ORAM requires no gates to initialize.

As we expected, our ORAM has a clear asymptotic advantage over other schemes in terms of initialization. Moreover, at $2^{22}$ elements, it has a 3000-fold concrete performance advantage over Square-root ORAM, the fastest previously known construction in this respect. In fact, in the context of garbled circuits, our construction even initializes somewhat faster than a linear scan, which requires only a simple `memcpy` by each party. Thus, so long as a single access in our scheme is faster than a single linear scan, the efficiency break-even point between the two is exactly *one* access.

This is far better than other schemes, which require $\Omega(\log n)$ accesses in order to reach their break-even points. We note that the constant component of the Floram performance curve is due to the cost of setting up OT extensions for a second thread, and it could be optimized away.

## 6.2 Applications

In order to assess the performance of our ORAM construction in realistic scenarios, we implemented two secure applications, and benchmarked them with each of the ORAMs considered previously.

*Binary Search.* In order to highlight the ways in which the novel properties of our ORAM differentiate it from previous ORAM constructions, we begin with a simple binary search benchmark. The use of ORAM for performing binary searches was first considered by Gordon *et al.* [23], who reported that searching a database of $2^{20}$ 64-byte elements required roughly 1000 seconds.[6] Our ORAM benchmark procedure is derived from that used by Square-root ORAM [55]: first, the data is loaded from secure computation into an ORAM, and then a number of searches are performed (each requiring $\log_2 n$ semantic accesses to complete). In this context, linear scan has a special advantage: because it touches each element in the memory, it requires only a single semantic access to perform a search. As a consequence of this property, ORAM has thus far yielded little improvement over the trivial solution for the problem of searching.

We executed instances of this benchmark upon databases of $2^{15}$ and $2^{20}$ 16-byte elements, with 1, $2^5$, and $2^{10}$ searches being performed. In addition, we benchmarked single searches of databases of $2^{25}$ elements under Floram (due to exhaustion of memory, it was not possible to instantiate Square-root or Circuit ORAMs of this size). We do not include in our benchmark the cost of sorting the data, which is unnecessary for the linear scan solution. Sorting can be performed with a Batcher Mergesort [3] in $O(n \log^2 n)$, with practical costs being lower than the that of instantiating any of the tested ORAMs, other than Floram. Results are reported in Table 2.

Floram has the fastest access and initialization procedures at these sizes, and so, not surprisingly, it is the fastest among the ORAMs regardless of the number of searches performed. What is surprising, however, is that it is significantly faster than linear scan, *even when only a single search is performed.* To our knowledge, such a thing is not possible under any other ORAM scheme, at any data size. Our scheme achieves this due to the fact that, considering initialization and a single access, only two full scans of XOR shares are required, whereas in the context of Yao's Garbled Circuits a linear scan requires iterating over wire labels that are at least eighty times larger than the equivalent secret-shared representation.

*Stable Matching.* Many previous research efforts have sought to optimize the secure evaluation of the Gale-Shapley algorithm for stable matching. Recently, Doerner *et al.* [16] developed algorithmic improvements which yielded a significant increase in asymptotic and concrete performance, allowing them to execute a secure stable matching using the related Roth-Peranson algorithm on the scale of the stable matching performed annually by the National Resident Matching Program (NRMP) to match graduating doctors to medical residencies in the United States. This algorithm requires $O(nr)$ ORAM accesses in $n$, the number of doctors, and $r$, the number of hospitals for which the doctors are allowed to submit rankings, to a comparatively small ORAM of size $O(m)$ in $m$, the number of hospitals (in practice, around 5000 for NRMP-scale matchings). Nonetheless, in terms of gates, the NRMP matching is one of the largest secure computations ever reported. In other words, this is a benchmark for which Floram's initialization advantage matters very little. The parameters of the benchmark were derived by Doerner *et al.* from the 2016 NRMP Statistical Report; specifically: 35,476 residents submitting up to 15 rankings each, and 4836 hospitals submitting up to 120 rankings each, and having at most 12 open positions. Individual preferences were generated at random. We collected one sample each for Square-root ORAM and Floram CPRG, and, following Doerner *et al.*, we did not collect any data for Circuit ORAM or linear scan, which would not be competitive. The results are shown in Table 3, and demonstrate a factor of 1.83 improvement over prior work for a very small ORAM used in a real application.

| $n$ | $s$ | Linear | Circuit | Square-root | Floram | CPRG |
|---|---|---|---|---|---|---|
| $2^{15}$ | 1 | 2.80 | 5192.4 | 12.87 | 1.85 | **0.64** |
| | $2^5$ | 89.75 | 5284.2 | 37.24 | 51.28 | **11.97** |
| | $2^{10}$ | 2872.1 | 8126.8 | 1210.0 | 1625.3 | **383.7** |
| $2^{20}$ | 1 | 89.52 | – | 690.99 | 3.79 | **2.20** |
| | $2^5$ | 2864.5 | – | 800.23 | 95.97 | **44.02** |
| | $2^{10}$ | 91,663. | – | 12,736. | 3023.5 | **1386.2** |
| $2^{25}$ | 1 | 2864.5 | – | – | 26.75 | **24.79** |

**Table 2: Binary Search Benchmark Results. We measured the wall-clock time required for $s$ searches through $n$ 16-byte data elements, including initialization. Figures are averages in seconds from 30 samples for databases of $2^{15}$ elements, or 3 samples for larger databases. Linear scan figures are estimated from results in Section 6.1.**

| | Square-root | Floram CPRG |
|---|---|---|
| Wall-clock Time (Hours) | 28.98 | **15.78** |
| Billions of Non-free Gates | 226.87 | **143.29** |

**Table 3: Roth-Peranson Benchmark Results. Our wall-clock time result for Square-root ORAM differs from that presented by Doerner *et al.* [16]; this is due to differences in benchmarking environments used.**

## ACKNOWLEDGMENT

The authors would like to thank the authors of the Square-root ORAM paper [55], and especially Samee Zahur, for his insight and technical expertise.

[6] Though we show significant improvement upon this number, our construction is not directly comparable to theirs, due to differences in the underlying protocol and benchmarking hardware.

## CODE AVAILABILITY

Complete reference implementations of the constructions described in this paper along with implementations of Square-root and Circuit ORAM sharing a common interface are available under the 3-clause BSD license from https://gitlab.com/neucrypt/floram.

## REFERENCES


[1] 2001. Advanced Encryption Standard. (2001).
[2] Ittai Abraham, Christopher W. Fletcher, Kartik Nayak, Benny Pinkas, and Ling Ren. 2017. Asymptotically Tight Bounds for Composing ORAM with PIR. In *PKC*.
[3] Ken Batcher. 1968. Sorting Networks and Their Applications. In *Spring Joint Computer Conference*.
[4] Donald Beaver, Silvio Micali, and Phillip Rogaway. 1990. The Round Complexity of Secure Protocols. In *ACM STOC*.
[5] Mihir Bellare, Viet Tung Hoang, Sriram Keelveedhi, and Phillip Rogaway. 2013. Efficient Garbling from a Fixed-Key Blockcipher. In *IEEE S&P*.
[6] Marina Blanton, Aaron Steele, and Mehrdad Alisagari. 2013. Data-oblivious Graph Algorithms for Secure Computation and Outsourcing. In *ACM Asia CCS*.
[7] Dan Boneh, David Mazieres, and Raluca Ada Popa. 2011. Remote Oblivious Storage: Making Oblivious RAM practical. http://dspace.mit.edu/bitstream/handle/1721.1/62006/MIT-CSAIL-TR-2011-018.pdf. (2011).
[8] Joan Boyar and René Peralta. 2010. A New Combinational Logic Minimization Technique with Applications to Cryptology. In *Lecture Notes in Computer Science*.
[9] Joan Boyar and René Peralta. 2012. *A Small Depth-16 Circuit for the AES S-Box*.
[10] Elette Boyle, Kai-Min Chung, and Rafael Pass. 2016. Oblivious Parallel RAM and Applications. In *TCC*.
[11] Elette Boyle, Niv Gilboa, and Yuval Ishai. 2015. Function Secret Sharing. In *EUROCRYPT*.
[12] Elette Boyle, Niv Gilboa, and Yuval Ishai. 2016. Function Secret Sharing: Improvements and Extensions. In *ACM CCS*.
[13] Kai-Min Chung, Zhenming Liu, and Rafael Pass. 2013. Statistically-secure ORAM with $\tilde{O}(\log^2 n)$ Overhead. *arXiv preprint arXiv:1307.3699* (2013).
[14] Ivan Damgård, Sigurd Meldgaard, and Jesper Buus Nielsen. 2011. Perfectly Secure Oblivious RAM without Random Oracles. In *TCC*.
[15] Srinivas Devadas, Marten van Dijk, Christopher W. Fletcher, Ling Ren, Elaine Shi, and Daniel Wichs. 2016. Onion ORAM: A Constant Bandwidth Blowup Oblivious RAM. In *TCC*.
[16] Jack Doerner, David Evans, and abhi shelat. 2016. Secure Stable Matching at Scale. In *ACM CCS*.
[17] Niv Gilboa and Yuval Ishai. 2014. *Distributed Point Functions and Their Applications*.
[18] Oded Goldreich. 1987. Towards a theory of software protection and simulation by oblivious RAMs. In *ACM STOC*.
[19] Oded Goldreich and Rafail Ostrovsky. 1996. Software Protection and Simulation on Oblivious RAMs. *Journal of the ACM* 43, 3 (1996).
[20] Michael T. Goodrich and Michael Mitzenmacher. 2011. Privacy-Preserving Access of Outsourced Data via Oblivious RAM Simulation. In *ICALP*.
[21] Michael T. Goodrich, Michael Mitzenmacher, Olga Ohrimenko, and Roberto Tamassia. 2011. Oblivious RAM Simulation with Efficient Worst-Case Access Overhead. In *ACM CCSW*.
[22] Michael T. Goodrich, Michael Mitzenmacher, Olga Ohrimenko, and Roberto Tamassia. 2012. Privacy-preserving group data access via stateless oblivious RAM simulation. In *ACM-SIAM SODA*.
[23] Dov Gordon, Jonathan Katz, Vladimir Kolesnikov, Fernando Krell, Tal Malkin, Mariana Raykova, and Yevgeniy Vahlis. 2012. Secure Two-party Computation in Sublinear (Amortized) Time. In *ACM CCS*.
[24] Yan Huang, David Evans, Jonathan Katz, and Lior Malka. 2011. Faster Secure Two-party Computation Using Garbled Circuits. In *USENIX Security Symposium*.
[25] Jack Doerner and abhi shelat. 2017. Scaling ORAM for Secure Computation. Cryptology ePrint Archive, Report 2017/827. https://eprint.iacr.org/2017/827.pdf. (2017).
[26] Zahra Jafargholi and Daniel Wichs. 2016. *Adaptive Security of Yao's Garbled Circuits*.
[27] Marcel Keller and Peter Scholl. 2014. *Efficient, Oblivious Data Structures for MPC*.
[28] Vladimir Kolesnikov and Thomas Schneider. 2008. Improved Garbled Circuit: Free XOR Gates and Applications. In *ICALP*.
[29] Eyal Kushilevitz, Steve Lu, and Rafail Ostrovsky. 2012. On the (In)security of Hash-based Oblivious RAM and a New Balancing Scheme. In *ACM-SIAM SODA*.
[30] Yehuda Lindell and Benny Pinkas. 2009. A Proof of Security of Yao's Protocol for Two-Party Computation. *Journal of Cryptology* 22, 2 (2009).
[31] Steve Lu and Rafail Ostrovsky. 2013. *Distributed Oblivious RAM for Secure Two-Party Computation*.
[32] Valeria Nikolaenko, Udi Weinsberg, Stratis Ioannidis, Marc Joye, Dan Boneh, and Nina Taft. 2013. Privacy-Preserving Ridge Regression on Hundreds of Millions of Records. In *IEEE S&P*.
[33] Rafail Ostrovsky. 1990. Efficient computation on oblivious RAMs. In *ACM STOC*.
[34] Rafail Ostrovsky and Victor Shoup. 1997. Private Information Storage (Extended Abstract). In *ACM STOC*.
[35] Benny Pinkas and Tzachy Reinman. 2010. Oblivious RAM revisited. In *CRYPTO*.
[36] Benny Pinkas, Thomas Schneider, Nigel P. Smart, and Stephen C. Williams. 2009. Secure Two-Party Computation Is Practical. In *ASIACRYPT*.
[37] Adi Shamir. 1979. How to Share a Secret. *Commun. ACM* 22, 11 (Nov. 1979).
[38] Elaine Shi, T.-H. Hubert Chan, Emil Stefanov, and Mingfei Li. 2011. Oblivious RAM with $O((\log N)^3)$ Worst-Case Cost. In *ASIACRYPT*.
[39] Emil Stefanov, Marten Van Dijk, Elaine Shi, Christopher Fletcher, Ling Ren, Xiangyao Yu, and Srinivas Devadas. 2013. Path ORAM: an Extremely Simple Oblivious RAM Protocol. In *ACM CCS*.
[40] Emil Stefanov, Marten van Dijk, Elaine Shi, Christopher Fletcher, Ling Ren, Xiangyao Yu, and Srinivas Devadas. 2013. Path ORAM: an Extremely Simple Oblivious RAM Protocol. In *ACM CCS*.
[41] Abraham Waksman. 1968. A Permutation Network. *Journal of the ACM* 15, 1 (Jan. 1968).
[42] Xiao Wang, Hubert Chan, and Elaine Shi. 2015. Circuit ORAM: On Tightness of the Goldreich-Ostrovsky Lower Bound. In *ACM CCS*.
[43] Xiao Wang, Yan Huang, Hubert Chan, Abhi Shelat, and Elaine Shi. 2014. SCORAM: Oblivious RAM for Secure Computation. In *ACM CCS*.
[44] Xiao Wang, Yan Huang, Yongan Zhao, Haixu Tang, XiaoFeng Wang, and Diyue Bu. 2015. Efficient Genome-Wide, Privacy-Preserving Similar Patient Query Based on Private Edit Distance. In *ACM CCS*.
[45] Peter Williams and Radu Sion. 2008. Usable PIR. In *NDSS*.
[46] Peter Williams and Radu Sion. 2012. Round-Optimal Access Privacy on Outsourced Storage. In *ACM CCS*.
[47] Peter Williams, Radu Sion, and Bogdan Carbunar. 2008. Building castles out of mud: Practical access pattern privacy and correctness on untrusted storage. In *ACM CCS*.
[48] R. S. Winternitz. 1984. A Secure One-Way Hash Function Built from DES. In *IEEE S&P*.
[49] Johannes Wolkerstorfer, Elisabeth Oswald, and Mario Lamberger. 2002. An ASIC Implementation of the AES SBoxes. In *RSA Conference on Topics in Cryptology*.
[50] David Woodruff and Sergey Yekhanin. 2005. A Geometric Approach to Information-Theoretic Private Information Retrieval. In *Proceedings of the 20th Annual IEEE Conference on Computational Complexity*.
[51] Andrew Chi-Chih Yao. 1982. Protocols for Secure Computations. In *IEEE FOCS*.
[52] Andrew Chi-Chih Yao. 1986. How to Generate and Exchange Secrets (Extended Abstract). In *IEEE FOCS*.
[53] Samee Zahur and David Evans. 2015. Obliv-C: A Lightweight Compiler for Data-Oblivious Computation. Cryptology ePrint Archive, Report 2015/1153. http://oblivc.org. (2015).
[54] Samee Zahur, Mike Rosulek, and David Evans. 2015. Two Halves Make a Whole: Reducing Data Transfer in Garbled Circuits Using Half Gates. In *EUROCRYPT*.
[55] Samee Zahur, Xiao Wang, Mariana Raykova, Adrià Gascón, Jack Doerner, David Evans, and Jonathan Katz. 2016. Revisiting Square Root ORAM: Efficient Random Access in Multi-Party Computation. In *IEEE S&P*.