Obladi: Oblivious Serializable Transactions in the Cloud

Natacha Crooks*†  Matthew Burke‡  Ethan Cecchetti‡
Sitar Harel‡  Rachit Agarwal‡  Lorenzo Alvisi‡

*University of Texas at Austin  ‡Cornell University

Abstract
This paper presents the design and implementation of Obladi, the first system to provide ACID transactions while also hiding access patterns. Obladi uses as its building block oblivious RAM, but turns the demands of supporting transactions into a performance opportunity. By executing transactions within epochs and delaying commit decisions until an epoch ends, Obladi reduces the amortized bandwidth costs of oblivious storage and increases overall system throughput. These performance gains, combined with new oblivious mechanisms for concurrency control and recovery, allow Obladi to execute OLTP workloads with reasonable throughput: it comes within 5× to 12× of a non-oblivious baseline on the TPC-C, SmallBank, and FreeHealth applications. Latency overheads, however, are higher (70× on TPC-C).

1 Introduction
This paper presents Obladi, the first cloud-based key value store that supports transactions while hiding access patterns from cloud providers. Obladi aims to mitigate the fundamental tension between the convenience of offloading data to the cloud, and the significant privacy concerns that doing so creates. On the one hand, cloud services [3, 4, 48, 49, 62] offer clients scalable, reliable IT solutions and present application developers with feature-rich environments (transactional support, stronger consistency [23, 52], etc.). Medical practices, for instance, increasingly prefer to use cloud-based software to manage electronic health records (EHR) [17, 39]. On the other hand, many applications that could benefit from cloud services store personal data that can reveal sensitive information even when encrypted or anonymized [53, 54, 74, 83]. For example, charts accessed by oncologists can reveal not only the content of an object, but also when and how frequently it is accessed, is thus often desirable.

Responding to this challenge, the systems community has taken a fresh look at private data access. Recent solutions, whether based on private information retrieval [2, 31], Oblivious RAM [15, 44, 70], function sharing [83], or trusted hardware [5, 7, 25, 44, 81], show that it is possible to support complex SQL queries without revealing access patterns.

Obladi addresses a complementary issue: supporting ACID transactions while guaranteeing data access privacy. This combination raises unique challenges [3], as concurrency control mechanisms used to enforce isolation, and techniques used to enforce atomicity and durability, all make hiding access patterns more problematic (§3).

Obladi takes as its starting point Oblivious RAM, which provably hides all access patterns. Existing ORAM implementations, however, cannot support transactions. First, they are not fault-tolerant. For security and performance, they often store data in a client-side stash; durability requires the stash content to be recoverable after a failure, and preserving privacy demands hiding the stash’s size and contents, even during failure recovery. Second, ORAM provides limited or no support for concurrency [12, 70, 75, 86], while transactional systems are expected to sustain highly concurrent loads.

Obladi demonstrates that the demands of supporting transactions can not only be met, but also turned into a performance opportunity. Its key insight is that transactions actually afford more flexibility than the single-value operations supported by previous ORAMs. For example, serializability [61] requires that the effects of transactions be reflected consistently in the state of the database only after they commit. Obladi leverages this flexibility to delay committing transactions until the end of fixed-size epochs, buffering their execution at a trusted proxy and enforcing consistency and durability only at epoch boundaries. This delay improves ORAM throughput without weakening privacy.

The ethos of delayed visibility is the core that drives Obladi’s design. First, it allows Obladi to implement a multiversioned database atop a single-versioned ORAM, so that read operations proceed without blocking, as with other multiversioned
write-back phase of the ORAM algorithm is delayed until 2.

1. 728 13th USENIX Symposium on Operating Systems Design and Implementation

server, but choose to outsource the integrity and availability
doctors and nurses access medical records through an on-site
deployments similar to those of medical practices, where
Obladi’s threat and failure assumptions aim to model
operations did not affect the set of values that
transactions executing within the same epoch can observe.

Second, it allows Obladi to securely parallelize Ring
ORAM [69], the ORAM construction on which it builds.
Obladi pipelines conflicting ORAM operations rather than
processing them sequentially, as existing ORAM implementa-
tions do. This parallelization, however, is only secure if the
write-back phase of the ORAM algorithm is delayed until
pre-determined times, namely, epoch boundaries.

Finally, delaying visibility gives Obladi the ability to
abort entire epochs in case of failure. Obladi leverages this
flexibility, along with the near-deterministic write-back
algorithm used by Ring ORAM, to drastically reduce the
information that must be logged to guarantee durability and
privacy-preserving crash recovery.

The results of a prototype implementation of Obladi
are promising. On three applications (TPC-C [80], Small-
Bank [22], and FreeHealth [42], a real medical application)
Obladi is within 5×-12× of the throughput of non-private
baselines. Latency is higher (70×), but remains reasonable
(in the hundreds of milliseconds).

To summarize, this paper makes three contributions:
1. It presents the design, implementation, and evaluation of
the first ACID transactional system that also hides access
patterns.
2. It introduces an epoch-based design that leverages the
flexibility of transactional workloads to increase overall
system throughput and efficiently recover from failures.
3. It provides the first formal security definition of a trans-
actional, crash-prone, and private database. Obladi uses
the UC-security framework [14], ensuring that security
guarantees hold under concurrency and composition.

Obladi also has several limitations. First, like most
ORAMs that regulate the interactions of multiple clients,
it relies on a local centralized proxy, which introduces
issues of fault-tolerance and scalability. Second, Obladi
does not currently support range or complex SQL queries.
Addressing the consistency challenge of maintaining
oblivious indices [5, 25, 89] in the presence of transactions
is a promising avenue for future work.

2 Threat and Failure Model

Obladi’s threat and failure assumptions aim to model
deployments similar to those of medical practices, where
doctors and nurses access medical records through an on-site
server, but choose to outsource the integrity and availability
of those records to a cloud storage service [17, 39].

Threat Model. Obladi adopts a trusted proxy threat
model [70, 75, 86]: it assumes multiple mutually-trusting
client applications interacting with a single trusted proxy in a
single shared administrative domain. The applications issue
transactions and the proxy manages their execution, sending
read and write requests on their behalf over an asynchronous
and unreliable network to an untrusted storage server. This
server is controlled by an honest-but-curious adversary that
can observe and control the timing of communication to
and from the proxy, but not the on-site communication
between application clients and the proxy. We extend our
threat model to a fully malicious adversary in our technical
report [20]. We consider attacks that leak information
by exploiting timing channel vulnerabilities in modern
processors [13, 36, 43] to be out of scope. Obladi guarantees
that the adversary will learn no information about: (i) the
decision (commit/abort) of any ongoing transaction; (ii) the
number of operations in an ongoing transaction; (iii) the type
of requests issued to the server; and (iv) the actual data they
access. Obladi does not seek to hide the type of application
that is currently executing (ex: OLTP vs OLAP).

Failure Model. Obladi assumes cloud storage is reliable,
but, unlike previous ORAMs, explicitly considers that both
application clients and the proxy may fail. These failures
should invalidate neither Obladi’s privacy guarantees nor the
Durability and Atomicity of transactions.

3 Towards Private Transactions

Many distributed, disk-based commercial database sys-
tems [8, 19, 58] separate concurrency control logic from
storage management: SQL queries and transactional
requests are regulated in a concurrency control unit and
are subsequently converted to simple read-write accesses
to key-value/file system storage. As ORAMs expose a
read-write address space to users, a logical first attempt
at implementing oblivious transactions would simply
replace the database storage with an arbitrary ORAM. This
black-box approach, however, raises both security concerns
§3.1 and performance/functionality issues §3.2

Security guarantees can be compromised by simply enforc-
ing the ACID properties. Ensuring Atomicity, Isolation, and
Durability imposes additional structure on the order of in-
dividual reads and writes, introducing sources of information
leakage [5, 72] that do not exist in non-transactional ORAMs
§3.1. Performance and functionality, on the other hand, are
hampered by the inability of current ORAMs to efficiently
support highly concurrent loads and guarantee Durability.

3.1 Security for Isolation and Durability

The mechanisms used to guarantee Isolation, Atomicity, and
Durability introduce timing correlations that directly leak
information about the data accessed by ongoing transactions.

Concurrency Control. Pessimistic concurrency controls
like two-phase locking [26] delay operations that would vi-
olate serializability: a write operation from transaction \( T_1 \)
cannot execute concurrently with any operation to the same object from transaction $T_2$. Such blocking can potentially reveal sensitive information about the data, even when executing on top of a construction that hides access patterns: a sudden drop in throughput could reveal the presence of a deadlock, of a write-heavy transaction blocking the progress of read transactions, or of highly contended items accessed by many concurrent transactions. More aggressive concurrency control schemes like timestamp ordering or multiversioned concurrency control [1, 10, 34, 41, 66, 67, 87] allow transactions to observe the result of the writes of other ongoing transactions. These schemes improve performance in contended workloads, but introduce the potential for cascading aborts; if a transaction aborts, all transactions that observed its write must also abort. If a write-heavy transaction $T_{\text{heavy}}$ aborts, it may cause a large number of transactions to rollback, again revealing information about $T_{\text{heavy}}$ and, perhaps more problematically, about the set of objects that $T_{\text{heavy}}$ accessed.

**Failure Recovery.** When recovering from failure, Durability requires preserving the effects of committed transactions, while Atomicity demands removing any changes caused by partially-executed transactions. Most commercial systems [50, 58, 59] preserve these properties through variants of undo and redo logging. To guarantee Durability, write and commit operations are written to a redo log that is replayed after a failure. To guarantee Atomicity, writes performed by partially-executed transactions are undone via an undo log, restoring objects to their last committed state. Unfortunately, this undo process can leak information: the number of undo operations reveals the existence of ongoing transactions, their length, and the number of operations that they performed.

3.2 Performance/functionality limitations

Current ORAMs align poorly with the need of modern OLTP workloads, which must support large numbers of concurrent requests; in contrast, most ORAMs admit little to no concurrency [12, 70, 75, 86] (we benchmark the performance of sequential Ring ORAM in Figure 10a).

More problematically, ORAMs provide no support for fault-tolerance. Adding support for Durability presents two main challenges. First, most ORAMs require the use of a stash that temporarily buffers objects at the client and requires that these objects be written to server storage in very specific ways (as we describe further in §4). This process aligns poorly with guaranteeing Durability for transactions. Consider for example a transaction $T_1$ that reads the version of object $x$ written by $T_2$ and then writes object $y$. To recover the database to a consistent state, the update to $x$ should be flushed to cloud storage before the update to $y$. It may however not be possible to securely flush $x$ from the stash before $y$. Second, ORAMs store metadata at the client to ensure that cloud storage observes a request pattern that is independent of past and currently executing operations. As we show in §8, recovering this metadata after a failure can lead to duplicate accesses that leak information.

3.3 Introducing Obladi

These challenges motivate the need to co-design the transactional and recovery logic with the underlying ORAM data structure. The design should satisfy three goals: (i) security—the system should not leak access patterns; (ii) correctness—Obladi should guarantee that transactions are serializable; and (iii) performance—Obladi should scale with the number of clients. The principle of workload independence underpins Obladi’s security: the sequence of requests sent to cloud storage should remain independent of the type, number, and access set of the transactions being executed. Intuitively, we want Obladi’s sequence of accesses to cloud storage to be statistically indistinguishable from a sequence that can be generated by an Obladi simulator with no knowledge of the actual transactions being run by Obladi. If this condition holds, then observing Obladi’s accesses cannot reveal to the adversary any information about Obladi’s workload. We formalize this intuition in our security definition in §9.

Much of Obladi’s novelty lies not in developing new concurrency control or recovery mechanisms, but in identifying what standard database techniques can be leveraged to lower the costs of ORAM while retaining security, and what techniques instead subtly break obliviousness.

To preserve workload independence while guaranteeing good performance in the presence of concurrent requests, Obladi centers its design around the notion of delayed visibility. Delayed visibility leverages the observation that, on the one hand, ACID consistency and Durability apply only when transactions commit, and, on the other, commit operations can be delayed. Obladi leverages this flexibility to delay commit operations until the end of fixed-size epochs. This approach allows Obladi to (i) amortize the cost of accessing an ORAM over many concurrently executing requests; (ii) recover efficiently from failures; and (iii) preserve workload independence: the epochs’ deterministic structure allows Obladi to decouple its externally observable behavior from the specifics of the transactions being executed.

4 Background

Oblivious Remote Access Memory is a cryptographic protocol that allows clients to access data outsourced to an untrusted server without revealing what is being accessed [29]; it generates a sequence of accesses to the server that is completely independent of the operations issued by the client. We focus specifically on tree-based ORAMs, whose constructions are more efficiently implementable in real systems: to date, they have been implemented in hardware [27, 46] and as the basis for blockchain ledgers [15] with reasonable overheads. Most tree-based ORAMs follow a similar structure: objects (usually key-value pairs) are mapped to a random leaf (or path) in a binary tree and physically reside (encrypted) in
some tree node (or bucket) along that path. Objects are logically removed from the tree and remapped to a new random path when accessed. These objects are eventually flushed back to storage (according to their new path) as part of an eviction phase. Through careful scheduling, this write-back phase does not reveal the new location of the objects; objects that cannot be flushed are kept in a small client-side stash.

Ring ORAM. Obladi builds upon Ring ORAM [69], a tree-based ORAM with two appealing properties: a constant stash size and a fully deterministic eviction phase. Obladi leverages these features for efficient failure recovery.

As shown in Figure 1, server storage in Ring ORAM consists of a binary tree of buckets, each with a fixed number $Z + S$ of slots. Of these, $Z$ are reserved for storing actual encrypted data (real objects); the remaining $S$ exclusively store dummy objects. Dummy objects are blocks of encrypted but meaningless data that appear indistinguishable from real objects; their presence in each bucket prevent the server from learning how many real objects the bucket contains and which slots contains them. A random permutation (stored at the client) determines the location of dummy slots. In Figure 1, the root bucket contains a real slot followed by two dummy slots; the real slot contains the data object $a$; its left child bucket instead contains dummy slots in positions one and three, and an empty real slot in second position.

Client storage, on the other hand, is limited to (i) a constant sized stash, which temporarily buffers objects that have yet to be replaced into the tree and, unlike a simple cache, is essential to Ring ORAM’s security guarantees; (ii) the set of current permutations, which identify the role of each slot in each bucket and record which slot have already been accessed (and marked invalid); and (iii) a position map, which records the random leaf (or path) associated with every data object. In Ring ORAM, objects are mapped to individual leaves of the tree but can be placed in any one of the buckets along the path from the root to that leaf. For instance, object $a$ in Figure 1 is mapped to path 4 but stored in the root bucket, while object $b$ is mapped to path 2 and stored in the leaf bucket of this path.

Ring ORAM maintains two core invariants. First, each data object is mapped to a new leaf chosen uniformly at random after every access, and is stored either in the stash, or in a bucket on the path from the tree’s root to that leaf (path invariant). Second, the physical positions of the $Z + S$ dummy and real objects in each bucket are randomly permuted with respect to all past and future writes to that bucket (i.e., no slot can be accessed more than once between permutations) (bucket invariant). The server never learns whether the client accesses a real or a dummy object in the bucket, so the exact position of the object along that path is never revealed.

Intuitively, the path invariant removes any correlation between two accesses to the same object (each access will access independent random paths), while the bucket invariant prevents the server from learning when an object was last accessed (the server cannot distinguish an access to a real slot from a dummy slot). Together, these invariants ensure that, regardless of the data or type of operation, all access patterns will look indistinguishable from a random set of leaves and slots in buckets.

Access Phase. The procedures for read and write requests is identical. To access an object $o$, the client first looks up $o$’s path in the position map, and then reads one object from each bucket along that path. It reads $o$ from the bucket in which it resides and a valid dummy object from each other bucket, identified using its local permutation map. Finally, $o$ is remapped to a new path, updated to a new value (if the request was a write), and added to the stash; importantly, $o$ is not immediately written back out to cloud storage.

Figure 1 illustrates the steps involved in reading an object $b$, initially mapped to path 2. The client reads a dummy object from the first two buckets in the path (at slots two and three respectively), and reads $b$ from the first slot of the bottom bucket. The three slots accessed by the client are then marked as invalid in their respective buckets, and $b$ is remapped to path 1. To write a new object $c$, the client would have to read three valid dummy objects from a random path, place $c$ in the stash, and remap it to a new path.

Access Security. Remapping objects to independent random paths prevents the server from detecting repeated accesses to data, while placing objects in the stash prevents the server from learning the new path. Marking read slots as invalid forces every bucket access to read from a distinct slot (each selected according to the random permutation). The server consequently observes uniformly distributed accesses (without repetition) independently of the contents of the bucket. This lack of correlation, combined with the inability to distinguish real slots from dummy slots, ensures that the server does not learn if or when a real object is accessed. Accessing dummy slots from buckets not containing the target object (rather than real slots), on the other hand, is necessary for efficiency: in combination with Ring ORAM’s eviction phase (discussed next) it lets the stash size remain constant by preventing multiple real objects from being added to the stash on a single access.

Eviction Phase and Reshuffling. The aforementioned protocol falls short in two ways. First, if objects are placed in the stash after each access, the stash will grow unbounded. Second, all slots will eventually be marked as invalid. Ring
ORAM sidesteps these issues through two complementary processes: eviction and bucket reshuffling. Every $A$ accesses, the evict path operation evicts objects from the client stash to cloud storage. It deterministically selects a target path, flushes as much data as possible, and permutes each bucket in the path, revalidating any invalid slots. Evict path consists of a read and write phase. In the read phase, it retrieves $Z$ objects from each bucket in the path: all remaining valid real objects, plus enough valid dummies to reach a total of $Z$ objects read. In the write phase, it places each stashed object—including those read by the read phase—to the deepest bucket on the target path that intersects with the object’s assigned path. Evict path then permutes the real and dummy values in each bucket along the target path, marking their slots as valid, and writes their contents to server storage. Figure 2 and 3 show the evict path procedure applied to path 4. In the read phase, evict path reads the unread object $a$ from the root node and dummies from other buckets on the path. In the write phase (Fig. 3), $a$ is flushed to leaf 4, as its path intersects completely with the target path. Finally, we note that randomness may cause a bucket to contain only invalid slots before its path is evicted, rendering it effectively unaccessible. When this happens, Ring ORAM restores access to the bucket by performing an early reschedule operation that executes the read phase and write phase of evict path only for the target bucket.

Eviction Security. The read phase leaks no information about the contents of a given bucket. It systematically reads exactly $Z$ valid objects from the bucket, selecting the valid real objects from the $Z$ real objects in the bucket, padding the remaining $Z-z$ required reads with a random subset of the $S$ dummy blocks. The random permutation and randomized encryption ensure that the server learns no information about how many real objects exist, and how many have been accessed. Similarly, the write phase hides the values and locations of objects written. At every bucket, the storage server observes only a newly encrypted and permuted set of objects, eliminating any correlation between past and future accesses to that bucket. Together, the read and write phases ensure that no slot is accessed more than once between reshuffles, guaranteeing the bucket invariant.

Similarly, the eviction process leaks no information about the paths of the newly evicted objects: since all paths intersect at the root and the server cannot infer the contents of any individual bucket, any object in the stash may be flushed during any evict path. Moreover, since all paths intersect at the root, any object in the stash may be flushed during any evict path.

5 System Architecture

Obladi, like most privacy-preserving systems [70, 76, 86] consists of a centralized trusted component, the proxy, that communicates with a fault-tolerant but untrusted entity, cloud storage (Figure 4). The proxy handles concurrency control, while the untrusted cloud storage stores the private data. Obladi ensures that requests made by the proxy to the cloud storage over the untrusted network do not leak information. We assume that the proxy can crash and that when it does so, its state is lost. This two-tier design allows applications to run a lightweight proxy locally and delegate the complexity of fault-tolerance to cloud storage.

The proxy has two components: (i) a concurrency control unit and (ii) a data manager comprised of a batch manager and an ORAM executor. The batch manager periodically schedules fixed-size batches of client operations that the ORAM executor then executes on a parallel version of Ring ORAM’s algorithm. The executor accesses one of two units located on server storage: the ORAM tree, which stores the actual data blocks of the ORAM; and the recovery unit, which logs all non-deterministic accesses to the ORAM to a write-ahead log [51] to enable secure failure recovery (§8).

6 Proxy Design

The proxy in Obladi has three goals: guarantee good performance, preserve correctness, and guarantee security. To meet these goals, Obladi designs the proxy around the concept of epochs. The proxy partitions time into a set of fixed-length, non-overlapping epochs. Epochs are the granularity at which Obladi guarantees durability and consistency. Each transaction, upon arriving at the proxy, is assigned to an epoch and clients are notified of whether a transaction has committed only when the epoch ends. Until then, Obladi buffers all updates at the proxy.

This flexibility boosts performance in two ways. First, it allows Obladi to implement a multiversioned concurrency control (MVCC) algorithm on top of a single versioned Ring ORAM. MVCC algorithms can significantly improve throughput by allowing read operations to proceed with
limited blocking. These performance gains are especially significant in the presence of long-running transactions or high storage access latency, as is often the case for cloud storage systems. Second, it reduces traffic to the ORAM, as only the database state at the end of the epoch needs to be written out to cloud storage.

Importantly, Obladi’s choice to enforce consistency and durability only at epoch boundaries does not affect correctness; transactions continue to observe a serializable and recoverable schedule (i.e., committed transactions do not see writes from aborted transactions).

For transactions executing concurrently within the same epoch, serializability is guaranteed by concurrency control; transactions from different epochs are naturally serialized by the order in which the proxy executes their epochs. No transaction can span multiple epochs; unfinished transactions at epoch boundaries are aborted, so that no transaction is ongoing during epoch changes.

Durability is instead achieved by enforcing epoch fate-sharing [82] during proxy or client crashes: Obladi guarantees that either all completed transactions (i.e., transactions for which a commit request has been received) in the epoch are made durable or all transactions abort. This way, no committed transaction can ever observe non-durable writes.

Finally, the deterministic pattern of execution that epochs impose drastically simplifies the task of guaranteeing workload independence: as we describe further below, the frequency and timing at which requests are sent to untrusted storage are fixed and consequently independent of the workload.

The proxy processes epochs with two modules: the concurrency control unit (CCU) ensures that execution remains serializable, while the data handler (DH) accesses the actual data objects. We describe each in turn.

### 6.1 Concurrency Control

Obladi, like many existing commercial databases [57, 65], uses multiversioned concurrency control [10]. Obladi specifically chooses multiversioned timestamp ordering (MVTSO) [10, 68] because it allows uncommitted writes to be immediately visible to concurrently executing transactions. To ensure serializability, transactions log the set of transactions whose uncommitted values they have observed (their write-read dependencies) and abort if any of their dependencies fail to commit. This optimistic approach is critical to Obladi’s performance: it allows transactions within the same epoch to see each other’s effects even as Obladi delays commits until the epoch ends. In contrast, a pessimistic protocol like two-phase locking [26], which precludes transactions from observing uncommitted writes, would artificially increase contention by holding exclusive write-locks for the duration of an epoch. When a transaction starts, MVTSO assigns it a unique timestamp that determines its serialization order. A write operation creates a new object version marked with its transaction’s timestamp and inserts it in the version chain associated with that object. A read operation returns the object’s latest version with a timestamp smaller than its transaction’s timestamp. Read operations further update a read marker on the object’s version chain with their transaction’s timestamp. Any write operation with a smaller timestamp that subsequently tries to write to this object is aborted, ensuring that no read operation ever fails to observe a write from a transaction that should have preceded it in the serialization order.

Consider for example the set of transactions executing in Figure 5. Transaction $t_1$’s update to object $a$ ($w(a_1)$) is immediately observed by transaction $t_2$ ($r_2(a_1)$). $t_2$ becomes dependent on $t_1$ and can only commit once $t_1$ also commits. In contrast, $t_2$’s write to object $d$ causes $t_2$ to abort: a transaction with a higher timestamp ($t_2$) had already read version $d_0$, setting the version’s read marker to 3.

### 6.2 Data Handler

Once a version is selected for reading or writing, the DH becomes responsible for accessing or modifying the actual object. Whereas it suffices to guarantee durability and consistency only at epoch boundaries, security must hold at all times, posing two key challenges. First, the number of requests executed in parallel can leak information, e.g., data dependencies within the same transaction [11, 70]. Second, transactions may abort (§6.1), requiring their effects to be rolled back without revealing the existence of contended objects [5, 72]. To decouple the demands of these workloads from the timing and set of requests that it forwards to cloud storage, Obladi leverages the following observation: transactions can always be re-organized so that all reads from cloud storage execute before all writes [19, 38, 47, 88]. Indeed, while operations within a transaction may depend on the data
returned by a read from cloud storage, no operation depends on the execution of a write. Accordingly, Obladi organizes the DH into a read phase and a write phase: it first reads all necessary objects from cloud storage, before applying all writes.

**Read Phase.** Obladi splits each epoch’s read phase into a fixed set of $R$ fixed-sized read batches ($b_{\text{read}}$) that are forwarded to the ORAM executor at fixed intervals ($\Delta_{\text{epoch}}$). This deterministic structure allows Obladi to execute dependent read operations without revealing the internal control flow of the epoch’s transactions. Read operations are assigned to the epoch’s next unfilled read batch. If no such batch exists, the transaction is aborted. Conversely, before a batch is forwarded to the ORAM executor, all remaining empty slots are padded with dummy requests. Obladi further deduplicates read operations that access the same key. As we describe in §7, this step is necessary for security since parallelized batches may leak information unless requests all access distinct keys [12, 86]. Deduplicating requests also benefits performance by increasing the number of operations that can be served within a fixed-size batch.

**Write Phase.** While transactions execute, Obladi buffers their write operations into a version cache that maintains all object versions created by transactions in the epoch. At the end of an epoch, transactions that have yet to finish executing (recall that epochs terminate at fixed intervals) are aborted and their operations are removed. The latest versions of each object in the version cache according to the version chain are then aggregated in a fixed-size write batch ($b_{\text{write}}$) that is forwarded to the ORAM executor, with additional padding if necessary.

This entire process, including write buffering and deduplication, must not violate serializability. The DH guarantees that write buffering respects serializability by directly serving reads from the version cache for objects modified in the current epoch. It guarantees serializability in the presence of duplicate requests by only including the last write of the version chain in a write batch. Since Obladi’s epoch-based design guarantees that transactions from a later epoch are serialized after all transactions from an earlier epoch, intermediate object versions can be safely discarded. In this context, MVTSO’s requirement that transactions observe the latest committed write in the serialization order reduces to transactions reading the tail of the previous epoch’s version chain.

In the presence of failures, Obladi guarantees serializability and recoverability by enforcing epoch fate sharing; either all transactions in an epoch are made durable or none are. If a failure arises during epoch $e_i$, the system simply recovers to epoch $e_{i-1}$, aborting all transactions in epoch $e_i$. Once again, this flexibility arises from Obladi delaying commit notifications until epoch boundaries.

**Example Execution.** We illustrate the batching logic once again with the help of Figure 5. Transactions $t_1$, $t_2$, $t_3$ first execute read operations. These operations are aggregated into the first read batch of epoch $i$. The values returned by these reads are then cached into the version cache. $t_2$ then executes a write operation, which Obladi also buffers into the version cache. When executing $r_2(d_3)$, $t_3$ reads object $d$ directly from the version cache (we discuss the security of this step in the next section). Similarly, $r_1(a_1)$ reads the buffered uncommitted version of $a$. In contrast, Obladi schedules $r_1(b_0)$ to execute as part of the next read batch as $b_0$ is not present in the version cache. The read batch is then padded to its fixed $b_{\text{read}}$ size and executed. $t_4$ contains no read operations: its write operations are simply executed and buffered at the version cache. Obladi then finalizes the epoch by aborting all transactions (and their dependencies) that have not yet finished executing; $t_4$ is consequently aborted. Finally, Obladi aggregates the last version of every update into the write batch (skipping version $c_1$ of object $c$ for instance, instead only writing $c_2$), before notifying clients of the commit decision.

### 6.3 Reducing Work

Obladi reduces work in two additional ways: it caches reads within an epoch and allows Ring ORAM to execute write operations without also executing dummy queries. While these optimizations may appear straightforward, ensuring that they maintain workload independence requires care.

**Caching Reads.** Ring ORAM maintains a client-side stash (§4) that stores ORAM blocks until their eviction to cloud storage. Importantly, a request for a block present in the stash still triggers a dummy request: a dummy object is still retrieved from each bucket along its path. While this access may appear redundant at first, it is in fact necessary to preserve workload independence: removing it removes the guarantee that the set of paths that Obladi requests from cloud storage is uniformly distributed. In particular, blocks present in the stash are more likely to be mapped to paths farther away from the one visited by the last evict path, as they correspond to paths that could not be flushed: buckets have limited space for real blocks and blocks mapped to paths that only intersect near the top of the tree are less likely to find a free slot to which they can be flushed. The degree to which this effect skews the distribution leaks information about the stash size, and, consequently, about the workload. To illustrate, consider the execution in Figure 6. Objects mapped to paths 1 and 2 ($a$, $b$, and $f$) were not flushed from the stash in the previous eviction of path 4. When these objects are subsequently accessed, naively reading them from the stash without performing dummy reads skews the set of paths accessed toward the right subtree (paths 3 and 4).

Obladi securely mitigates some of this work by drawing a novel distinction between objects that are in the stash as a result of a logical access and those present because they could not be evicted. The former can be safely accessed without performing a dummy read, while the latter cannot. Objects present in the stash following a logical access are mapped to independently uniformly distributed paths. Ring ORAM’s path invariant ensures that, without caching, the
set of accessed paths is uniformly distributed. Removing an independent uniform subset of those paths (namely, the dummy requests) will consequently not change the distribution. Thus, caching these objects, and filling out a read batch with other real or dummy requests, preserves the uniform distribution of paths and leaks no information. Obladi consequently allows all read objects to be placed in the version cache for the duration of the epoch. Objects \(a, b, d\) are, for instance, placed in the version cache in Figure 5, allowing read \(r_2(d_0)\) to read \(d\) directly from the cache. In contrast, objects present in the stash because they could not be evicted are mapped to paths that skew away from the latest evict path. Caching these objects would consequently skew the distribution of requests sent to the storage away from a uniform distribution, as illustrated in Figure 6.

**Dummiless Writes.** Ring ORAM must hide whether requests correspond to read or write operations, as the specific pattern in which these operations are interleaved can leak information [89]; that is why Ring ORAM executes a read operation on the ORAM for every access. In contrast, since transactions can always perform all reads before all writes, no information is leaked by informing the storage server that each epoch consists of a fixed-size sequence of potentially dummy reads followed by a fixed-size sequence of potentially dummy writes. Obladi thus modifies Ring ORAM’s algorithm to directly place the new version of an object in the stash, without executing the corresponding read. Note, though, that Obladi continues to increment the evict path count on write operations, a necessary step to preserve the bounds on the stash size, which is important for durability (§8).

### 6.4 Configuring Obladi

Obladi’s good performance hinges on appropriately configuring the size/frequency of batches and ORAM tree for a target application. Table 1 summarizes the parameter space.

**Ring ORAM.** Configuring Ring ORAM first requires choosing an appropriate \(Z\) parameter. Larger values of \(Z\) reduce the total size of the ORAM on cloud-storage by decreasing the required height of the ORAM tree and decrease eviction frequency (reducing network/CPU overhead). In contrast, this increase the maximum stash size. Traditional ORAMs thus choose the largest value of \(Z\) for which the stash size fits on the proxy. Obladi adds an additional consideration: for durability (as we describe in §8), the stash must be synchronously written out every epoch. One must thus take into account the throughput loss associated with the stash writeback time. Given an appropriate value of \(Z\), Obladi then chooses \(L\), \(S\), and \(A\) according to the analytical model proposed in [69].

**Epochs and batching.** Identifying the appropriate size and number of batches hinges on several considerations. First, Obladi must provision sufficiently many read batches \((R)\) to handle control flow dependencies within transactions. A transaction that executes in sequence five dependent read operations, will for instance require five read batches to execute (it will otherwise repeatedly abort). Second, the ratio of reads \((R+b_{\text{read}})\) to writes \((w_{\text{write}})\) must closely approximate the application’s read/write ratio. An overly large write batch will waste resources as it will be padded with many dummy requests. A write batch that is too small will lead to frequent aborts caused by the batch filling up. Third, the size of a read or write batch (respectively \(b_{\text{read}}\) and \(b_{\text{write}}\)) defines the degree of parallelism that can be extracted. The desired batch size is thus a function of the concurrent load of the system, but also of hardware considerations, as increasing parallelism beyond an I/O or CPU bottleneck serves no purpose. Finally, the number and frequency of read batches within an epoch increases overall latency, but reduces amortized resource costs through caching and operation pipelining (introduced in §7). Latency-sensitive applications may favor smaller batch sizes, while others may prefer longer epochs, but lower overheads.

**Security Considerations.** Obladi does not attempt to hide the size and frequency of batches from the storage server (we formalize this leakage in §9). Carefully tuning the size and frequency of batches to best match a given application may thus leak information about the application itself. An OLTP application, for instance, will likely have larger batch sizes \((b_{\text{read}})\), but fewer read batches \((R)\), as OLTP applications sustain a high concurrent load of fairly short transactions. OLAP applications will prefer small or non-existent write batches \((b_{\text{write}})\), as they are predominantly read-only, but require many read batches to support the complex joins/aggregates that they implement. Obladi does not attempt to hide the type of application that is being run. It does, however, continue to hide what data is being accessed and what transactions are currently being run at any given point in time. While Obladi’s configuration parameters may, for instance, suggest

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N)</td>
<td>Number of real objects</td>
</tr>
<tr>
<td>(Z)</td>
<td>Number of real slots</td>
</tr>
<tr>
<td>(L)</td>
<td>Number of read batches</td>
</tr>
<tr>
<td>(A)</td>
<td>Frequency of evict path</td>
</tr>
<tr>
<td>(S)</td>
<td>Number of dummy slots</td>
</tr>
<tr>
<td>(B_{\text{read}})</td>
<td>Size of a read batch</td>
</tr>
<tr>
<td>(B_{\text{write}})</td>
<td>Size of a write batch</td>
</tr>
<tr>
<td>(\Delta)</td>
<td>Batch frequency</td>
</tr>
</tbody>
</table>

Table 1: Obladi’s configuration parameters
that a medical application like FreeHealth is being run, they
do not in any way leak information about how, when, or
which patient records are being accessed.

7 Parallelizing the ORAM

Existing ORAM constructions make limited use of paral-
elism. Some allow requests to execute concurrently between
eviction or shuffle phases [12, 70, 86], while others target
intra-request parallelism to speed up execution of a single re-
quest [44]. Obladi explicitly targets both forms of parallelism.
Parallelizing Ring ORAM presents three challenges: (i) pre-
serving the correct abstraction of a sequential datastore, (ii)
enforcing security by concealing the position of real blocks
in the ORAM (thereby maintaining workload independence),
and (iii) preserving existing bounds on the stash size. While
these issues also arise in prior work [70], the idiosyncrasies
of Ring ORAM add new dimensions to these challenges.

Correctness. Obladi makes two observations. First, while
all operations conflict at the Ring ORAM tree’s root, they
can be split into suboperations that access mostly disjoint
buckets (§4). Second, conflicting bucket operations can be
further parallelized by distinguishing accesses to the bucket’s
metadata from those to its physical data blocks.

Obladi draws from the theory of multilevel serializabil-
ity [84], which guarantees that an execution is serializable if
the system enforces level-by-level serializability: if operation
o is ordered before o’ at level i, all suboperations of o must
precede conflicting suboperations of o’. Thus, if Obladi
orders conflicting operations at a level i, it enforces the same
order at level i + 1 for all their conflicting suboperations;
conversely, if two operations do not conflict at level i, Obladi
executes their suboperations in parallel. To this end, Obladi
simply tracks dependencies across operations and orders
conflicting suboperations accordingly. Obladi extracts further
parallelism in two ways. First, since in Ring ORAM reads to
the same bucket between consecutive eviction or reshuffling
operations always target different physical data blocks
(even when bucket operations conflict on metadata access),
Obladi executes them in parallel. Second, Obladi’s own
batching logic ensures that requests within a batch touch
different objects, preventing read and write methods from
ever conflicting. Together, these techniques allow Obladi to
execute most requests and evictions in parallel.

We illustrate the dependency tracking logic in Figure 7.
The read operation to path 1 conflicts with the evict path for
path 2, but only at the root (bucket 1). Thus, reads to buckets
2 and 3 can proceed concurrently, even though accesses to
the root’s metadata must be serialized, as both operations
update the bucket access counter and valid/invalid map (§4).

Security. For security, Obladi’s parallel evict path operation
must flush the same blocks flushed by a sequential imple-
mentation. Reproducing this behavior without sacrificing
parallelism is challenging. It requires that all real objects
brought in during the last A accesses be present in the
stash when data is flushed, which may introduce data dependencies. Unlike dependencies that arise between
operations that access the same physical location in cloud
storage, these dependencies are not a deterministic function
of an epoch’s operations already known to the adversary.

Consider, for instance, block b in Figure 7. In a sequential
implementation, b would enter the stash as a result of reading
path 1 and be flushed to bucket 3 by the following evict path.
Thus, evict path would have to wait until b is placed in the
stash. Honoring these dependencies opens a timing channel:
delay in flushing certain blocks can reveal object placement.
As blocks holding real objects can exist anywhere in the
tree and be remapped to any path, it follows that it is never
secure to execute an eviction operation until all previous
access operations have terminated.

Obladi mitigates this restriction by again leveraging delayed
visibility and the idea to separate read and write operations
within an epoch—but with an important difference. In §6.2
the proxy created separate batches for logical read and write
operations; to improve parallelism, Obladi, expanding on an
idea used by Shroud [44], assigns to separate phases within
an epoch the physical read and write operations that underlie
each of those logical operations. The read phase computes
all necessary metadata and executes the set of physical read
operations for all logical read path, early reschedule, and evict
path operations. This set is workload independent, so its
operations need not be delayed. Physical writes, however,
are only flushed at the end of an epoch. The proxy can again
apply write deduplication: if a bucket is repeatedly modified
during an epoch, only the last version must be written back.
Reads that should have read an intermediate write are served
locally from the buffered buckets.

The adversary thus always observes a set of reads to random
paths followed by a deterministic set of writes independent
of the contents of the ORAM and, consequently, of the work-
load. Data dependencies between read and evict operations
no longer create a timing channel. Meanwhile parallelism re-
mains high, as the physical blocks accessed in each phase are
guaranteed to be distinct—Ring ORAM directly guarantees
this for reads, while bucket deduplication does it for writes.
8 Durability

Obladi guarantees durability at the granularity of epochs: after a crash, it recovers to the state of the last failure-free epoch. Obladi adds two demands to the need of recovering to a consistent state: recovery should leak no information about past or future transactions, and it should be efficient, accessing minimal data from cloud storage. Obladi guarantees the former by ensuring that recovery logic and data logged for recovery maintain workload independence (§3). It strives towards the latter by leveraging the determinism of Ring ORAM.

Consistency. Obladi recovery logic relies on two well-known techniques: write-ahead logging [51] and shadow paging [30]. Obladi mandates that transactions be durable only at the end of an epoch; thus, on a proxy failure, all ongoing transactions can be aborted, and the system reverted to the previous epoch. To make this possible, Obladi must (i) recover the proxy metadata lost during the proxy crash, and (ii) ensure that the ORAM does not contain any of the aborted transactions’ updates. To recover the metadata, Obladi logs three data structures before declaring the epoch committed: the position map, the permutation map, and the stash. The position map and the permutation map identify the position of real objects in the ORAM tree (respectively, in a path and in a bucket); logging them prevents the recovery logic from having to scan the full ORAM to recover the position of buckets. Logging the stash is necessary for correctness. As eviction may be unable to flush the entire stash, some newly written buckets may be present only in the stash, even at epoch boundaries. Failing to log the stash could thus lead to data loss.

To undo partially executed transactions, Obladi adapts the traditional copy-on-write technique of shadow paging [30]: rather than updating buckets in place, it creates new versions of each bucket on every write. Obladi then leverages the inherent determinism of Ring ORAM to reconstruct a consistent snapshot of the ORAM at a given epoch. In Ring ORAM, the current version of a bucket (i.e., the number of times a bucket has been written) is a deterministic function of the number of prior evict paths. The number of evict paths per epoch is similarly fixed (evict paths happen every A accesses, and epochs are of fixed size). Obladi can then trivially revert the ORAM on failures by setting the evict path counter to its value at the end of the last committed epoch. This counter determines the number of evict paths that have occurred, and consequently the object versions of the corresponding epoch.

Security. Obladi ensures that (i) the information logged for durability remains independent of data accesses, and (ii) that the interactions between the failed epoch, the recovery logic, and the next epoch preserve workload independence.

Obladi addresses the first issue by encrypting the position map and the contents of the permutations table. It similarly encrypts the stash, but also pads it to its maximum size, as determined in canonical Ring ORAM [69], to prevent it from indicating skew (if a small number of objects are accessed frequently, the stash will tend to be smaller).

The second concern requires more care: workload independence must hold before, during, and after failures. Ring ORAM guarantees workload independence through two invariants: the bucket invariant and the path invariant (§4). Preserving bucket slots from being read twice between evictions is straightforward. Obladi simply logs the invalid/valid map to track which slots have already been read and recovers it during recovery; there is no need for encryption, as the set of slots read is public information. Ensuring that the ORAM continues to observe a uniformly distributed set of paths is instead more challenging. Specifically, read requests from partially executed transactions can potentially leak information, even when recovering to the previous epoch. Traditionally, databases simply undo partially executed transactions, mark them as aborted, and proceed as if they had never existed. From a security standpoint, however, these transactions were still observed by the adversary, and thus may leak information. Consider a transaction accessing object a (mapped to path 1) that aborts because of a proxy failure. Upon recovery, it is likely that a client will attempt to access a again. As the recovery logic restores the position map of the previous epoch, that new operation on a will result in another access to path 1, revealing that the initial access to path 1 was likely real (rather than padded), as the probability of collisions between two uniformly chosen paths is low. To mitigate this concern while allowing clients to request the same objects after failure, Obladi durably logs the list of paths and slot indices that it accesses, before executing the actual requests, and replays those paths during recovery (remapping any real blocks). While this process is similar to traditional database redo logging [51], the goal is different. Obladi does not try to reapply transactions (they have all aborted), but instead forces the recovery logic to be deterministic: the adversary always sees the paths from the aborted epoch repeated after a failure.

Optimizations. To minimize the overhead of checkpointing, Obladi checkpoints deltas of the position, permutation, and valid/invalid map, and only periodically checkpoints the full data structures. While the number of changes to the permutation and valid/invalid maps directly follows from the set of physical requests made to cloud storage, the size of the delta for the position map reveals how many real requests were included in an epoch—padded requests do not lead to position map updates. Obladi thus pads the map delta to the maximum number of entries that could have changed in an epoch (i.e., the read batch size times the number of read batches, plus the size of the single write batch).

9 System Security

We now outline Obladi’s security guarantees, deferring a formal treatment to the associated technical report [20]. To the best of our knowledge, we are the first to formalize the notion of crashes in the context of oblivious RAM.

Model We express our security proof within the Universal Composability (UC) framework [14], as it aligns well with
the needs of modern distributed systems: a UC-secure system remains UC-secure under concurrency or if composed with other UC-secure systems. Intuitively, proving security in the UC model proceeds as follows. First, we specify an ideal functionality $F$ that defines the expected functionality of the protocol for both correctness and security. For instance, Obladi requires that the execution be serializable, and that only the frequency of read and write batches be learned. We must ensure that the real protocol provides the same functionality to honest parties while leaking no more information than $F$ would. To establish this, we consider two different worlds: one where the real protocol interacts with an adversary $A$, and one where $F$ interacts with $S_A$, our best attempt at simulating $A$’s transcript—including its inputs, outputs, and randomness—and $S_A$’s output are given to an environment $E$, which can also observe all communications within each world. $E$’s goal is to determine which world contains the real protocol. To prompt the worlds to diverge, $E$ can delay and reorder messages, and even control external inputs (potentially causing failures). Intuitively, $E$ represents anything external to the protocol, such as concurrently executing systems. We say that the real protocol is secure if, for any adversary $A$, we can construct $S_A$ such that $E$ can never distinguish between the worlds.

### Assumptions

The security of Obladi relies on four assumptions. (i) Canonical Ring ORAM is linearizable (ii) MVTSO generates serializable executions. (iii) The network will retransmit dropped packets. The adversary learns of the rertransmissions, but nothing more.

### Ideal Functionality

To define the ideal functionality $F_{Ob}$, recall that the proxy is considered trusted while interactions with the cloud storage are not. This allows $F_{Ob}$ to replace the proxy and intermediate between clients and the storage server, performing the same functions as the proxy (we do not try to hide the concurrency/batching logic). We must, however, define $F_{Ob}$ to obliviously hide data values and access patterns. To this end, when the proxy logic finalizes a batch, $F_{Ob}$ simply informs the storage server that it is executing a read or write batch. Since $F_{Ob}$ is a theoretical ideal, we allow it to manage all storage internally, so it then updates its local storage and furnishes the appropriate response to each client.

In this setup, modeling proxy crashes is straightforward. Crashes can occur at any time and cause the proxy to lose all state. So, on an external input to crash, $F_{Ob}$ simply clears its state. Since we accept that $A$ may learn of proxy crashes, $F_{Ob}$ also sends a message to the storage server that it has crashed.

### Proof Sketch

The correctness of the system is straightforward, as $F_{Ob}$ behaves much the same as the proxy.

To prove security, we must demonstrate that, for any algorithm $A$ defining the behavior of the storage server, we can accurately simulate $A$’s behavior using only the information provided by $F_{Ob}$. Note that the simulator $S_A$ can run $A$ internally, as $A$ is simply an algorithm. Thus we can define $S_A$ to operate as follows. When $S_A$ receives notification of a batch, it constructs a parallel ORAM batch from uniformly random accesses of the correct type. It provides these accesses to $A$ and produces $A$’s response.

The security of this simulation hinges on two key properties: (i) the caching and deduplication logic do not affect the distribution of physical accesses, and (ii) the physical access pattern of a parallelized batch is entirely determined by the physical accesses prescribed by sequential Ring ORAM for the same batch. The first follows from Ring ORAM’s guarantee that each access will be an independent uniformly random path—removing an independently-sampled element does not change the distribution of the remaining set. The second follows from the parallelization procedure simply aggregating all accesses and performing all reads followed by all writes.

These properties ensure that the random access pattern produced by $S_A$ is identical to the access pattern produced by the proxy when operating on real data. Thus the simulated $A$ must behave exactly as it would when provided with real data, and produce indistinguishable output.

### 10 Implementation

Our prototype consists of 41,000 lines of Java code. We use the Netty library for network communication (v4.1.20), Google protobufs for serialization (v3.5.1), the Bouncy Castle library (v1.59) for encryption, and the Java MapDB library (v3) for persistence. We additionally implement a non-private baseline (NoPriv). NoPriv shares the same concurrency control logic (TSO), but replaces the proxy data handler with non-private remote storage. NoPriv neither batches nor delays operations; it buffers writes at the local proxy until commit, and serves writes locally whenever possible.

### 11 Evaluation

Obladi leverages the flexibility of transactional commits to mitigate the overheads of ORAM. To quantify the benefits and limitations of this approach, we ask:

1. How much does Obladi pay for privacy? (§11.1)
2. How do epochs affect these overheads? (§11.2)
3. Can Obladi recover efficiently from failures? (§11.3)

#### Experimental Setup

The proxy runs on a c5.xlarge Amazon EC2 instance (16 vCPUs, 32GB RAM), and the storage on an m5.4xlarge instance (16 vCPUs, 64GB RAM). The ORAM tree is configured with $Z = 100$ and optimal values of $S$ and $A$ (respectively, 196 and 168) [69]. We report the average of three 90 seconds runs (30 seconds ramp-up/down).
Benchmarks We evaluate the performance of our system using three applications: TPC-C [22, 80], SmallBank [22], and FreeHealth [28, 42]. Our microbenchmarks use the YCSB [18] workload generator. TPC-C, the de facto standard for OLTP workloads, simulates the business logic of e-commerce suppliers. We configure TPC-C to run with 10 warehouses [87]. In line with prior transactional key-value stores [79], we use a separate table as a secondary index on the order table to locate a customer’s latest order in the order status transaction, and on the customer table to look up customers by their last names (order status and payment). SmallBank [22] models a simple banking application supporting money transfers, withdrawals, and deposits. We configure it to run with one million accounts. Finally, we port FreeHealth [28, 42], an actively-used cloud EHR system (Figure 8). FreeHealth supports the business logic of medical practices and hospitals. It consists of 21 transaction types that doctors use to create patients and look up medical history, prescriptions, and drug interactions.

11.1 End-to-end Performance

Figure 9 summarizes the results from running the three end-to-end applications in two setups: a local setup in which the latency between proxy and server is low (0.3ms) (Obladi, NoPriv), and a more realistic WAN setup with 10ms latency (ObladiW, NoPrivW). We additionally compare those results with a local MySQL setup. MySQL, unlike NoPriv, cannot buffer writes. We consequently do not evaluate MySQL in the WAN setting.

TPC-C Obladi comes within 8x of NoPriv’s throughput, as NoPriv is contention-bottlenecked on the high rate of conflicts between the new-order and payment transactions on the district table. NoPriv’s performance is itself slightly higher than MySQL as the use of MVTSO allows for the new-order and payment transactions to be pipelined. In contrast, MySQL acquires exclusive locks for the duration of the transactions. Latency, however, spikes to 70x over NoPriv because of the inflexible execution pattern Obladi needs for security. Transactions in TPC-C vary heavily in size. Epochs must be large enough to accommodate all transactions, and hence artificially increase the latency of short instances. Moreover, write operations must be applied atomically during epoch changes. For a write batch size of 2,000, this process takes on average 340ms, further increasing latency for individual transactions. The write-back process also limits throughput, even preventing non-conflicting operations from making progress (in contrast, NoPriv can benefit from writes never blocking reads in MVTSO). Epoch changes also introduce additional aborts for transactions that straddle epochs. The additional 10ms latency of the WAN setting has comparatively little effect, as the large write batch size of TPC-C is the primary bottleneck: throughput remains within 9x of NoPrivW. Also NoPrivW’s performance does not degrade: since MVTSO exposes uncommitted writes immediately, increasing commit latency does not increase contention.

SmallBank Transactions in Smallbank are more homogeneous (between three and six operations); thus, the length of an epoch can be set to more closely approximate most transactions, reducing latency overheads (17x NoPriv). NoPriv is CPU bottlenecked for Smallbank; the relative throughput drop for Obladi is higher (12x) because of the overhead of changing epochs and the blocking that it introduces. Transaction dependency tracking becomes a bottleneck in NoPriv, resulting in a 15% throughput loss over MySQL. Increasing latency between proxy and storage causes both systems’ throughput to drop. ObladiW’s 35% drop is due to the increased duration of epoch changes (during which no other transactions can execute) while NoPrivW’s 30% drop stems from the larger dependency chains that arise from the relatively long commit phase.

FreeHealth Like SmallBank, FreeHealth consists of fairly short transactions and can thus choose a fairly small epoch (five read batches), reducing the impact on latency (20x NoPriv). Unlike Smallbank, however, FreeHealth consists primarily of read operations, and so it can choose a much smaller write batch (200), minimizing the cost of epoch changes and maximizing throughput (only a 4x drop over NoPriv and a 5.5x over NoPrivW for ObladiW). Both NoPriv and Obladi are contention-bottlenecked on the creation of episodes, the core units of EHR systems that encapsulate prescriptions, medical history, and patient interaction.

11.2 Impact of Epochs

Though epochs create blocking and cause aborts, they are key to reducing the cost of accessing ORAM, as they allow to (i) securely parallelize the ORAM and (ii) delay and buffer bucket writes. To quantify epochs’ impact on performance as a function of their size and the underlying storage properties, we instantiate an ORAM with 100K objects and choose three different storage backends: a local dummy (storing no real data) that responds to all reads with a static value and ignores writes (dummy); a remote server backend with an in-memory hashmap (server, ping time 0.3ms) and a remote WAN server backend with an in-memory hashmap (server WAN, ping time 10ms); and DynamoDB (dynamo, provisioned for 80K req/s, read ping 1ms, write 3ms).
Parallelization We first focus on the performance impact of parallelizing Ring ORAM (ignoring other optimizations). Graph 10a shows that, unsurprisingly, the benefits of parallelism increase with the latency of individual requests. Parallelizing the ORAM for dummy, for instance, yields no performance gain; in fact, it results in a 3x slowdown (from 72K req/s to 24K req/s). Sequential Ring ORAM on dummy is CPU-bound on metadata computation (remapping paths, shuffling buckets, etc.), so adding coordination mechanisms to guarantee multi-level serializability only increases the cost of accessing a bucket. As storage access latency increases and the ORAM becomes I/O-bound, the benefits of parallelism become more salient. For a batch size of 500, throughput increases by 12x for server, as much as 51x for dynamo, and 510x for WAN server. The available parallelism is a function of both the size/fan-out of the tree and the underlying resource bottlenecks of the proxy. Graph 10b captures the parallelization speedup for both intra- and inter-request parallelism, while Graph 10b quantifies the latency impact of batching. The parallelization speedup achieved for a batch size of one captures intra-request parallelism: the eleven levels of the ORAM can be accessed concurrently, yielding an 11x speedup. As batch sizes increase, Obladi can leverage inter-request parallelism to process non-conflicting physical operations in parallel, with little to no impact on latency. Dynamo peaks early (at 1750 req/s) because its client API uses blocking HTTP calls, and dummy’s storage eventually bottlenecks on encryption, but server and WAN server are more interesting. Their throughput is limited by the physical and data dependencies on the upper levels of the tree (recall that paths always conflict at the root (§7)).

Work Reduction To amortize ORAM overheads across a large number of operations, Obladi relies on delayed visibility to buffer writes to the end of an epoch, when they can be executed in parallel, discarding intermediate writes. Reads to those buckets are directly served from the proxy, reducing network communication and CPU work (as encryption is not needed). Graph 10d shows that enabling this optimization for an epoch of eight batches (a setup suitable for FreeHealth and TPC-C) yields a 1.5x speedup on both dynamo and the server, a 1.6x speedup on the WAN server, but only minimal gains for dummy (1.1x). When using a small number of batches, throughput gains come primarily from combining duplicate operations in buckets near the top of the tree. For example, the root bucket is written 27 times in an epoch of eight (once per eviction, every 168 requests). As these operations conflict, they must be executed sequentially and quickly become the bottleneck (other buckets have fewer operations to execute). Our optimization lets Obladi write the root bucket only once, significantly reducing latency and thus increasing throughput. As epochs grow in size, increasingly many buckets are buffered locally until the end of the epoch (§7), allowing reads to be served locally and further reducing I/O with the storage. Consider Graph 10c: throughput increases almost logarithmically; metadata computation eventually becomes a bottleneck for dummy, while server and WAN eventually run out of memory from storing most of the tree (our AWS account did not allow us to provision dynamo adequately for larger batches). Larger epochs reduce the raw amount of work per operation: with one batch, Obladi requires 41 physical requests per logical access. For real transactional workloads, however, epochs are not a silver bullet. Graph 10f suggests that applications are very sensitive to identifying the right epoch duration: too short and transactions cannot make progress, repeatedly aborting; too long and the system will remain unnecessarily idle.

11.3 Durability

Table 11b quantifies the efficiency of failure recovery and the cost it imposes on normal execution for ORAMs of different sizes (we show space results for only the WAN server as Dynamo follows a similar trend). During normal execution, durability imposes a moderate throughput drop (from 0.83x for 10K to 0.89x for 1M). This slowdown is due to the need to checkpoint client metadata and to synchronously log read paths to durable storage before reading. As seen in Graph 11a, computing diffs mitigates the
impact of checkpointing. Recovery time similarly increases as the ORAM grows, from 1.5s to 6.1s (Table 11b, RecTime).

The costs of decrypting the position and permutation maps (Pos and Perm) are low for small datasets, but grow linearly with the number of keys. Read path logging (Paths) instead starts much larger, but grows only with the depth of the tree.

12 Related Work

Batching Obladi amortizes ORAM costs by grouping operations into epochs and committing at epoch boundaries. Batching can mitigate expensive security primitives, e.g., it reduces server-side computation in private information retrieval (PIR) schemes [9, 31, 33, 45], amortizes the cost of shuffling networks in Atom [40] and the cost of verifying integrity in Concerto [6]. Changing when operations output commit is a popular performance-boosting technique: it yields significant gains for state-machine replication [35, 37, 64], file systems [55], and transactional databases [21, 47, 82].

ORAM parallelism Obladi extends recent work on parallel ORAM constructions [11, 44, 86] to extract parallelism both within and across requests. Shroud [44] targets intra-request parallelism by concurrently accessing different levels of tree-based ORAMs. Chung et al [12] and PrivateFS [86] instead target inter-request parallelism, respectively in tree-based [73] and hierarchical [85] ORAMs. Both works execute requests to distinct logical keys concurrently between resuffles or evictions and deduplicate concurrent requests for the same key to increase parallelism. Obladi leverages delayed visibility to separate batches into read and write phases, extracting concurrency both within requests and across evictions. Furthermore, Obladi parallelizes across requests by deduplicating requests at the trusted proxy.

ObliviStore [77] and Taostore [70] instead approach parallelization by focusing on asynchrony. ObliviStore [77] formalizes the security challenges of scheduling requests asynchronously; the oblivious scheduling mechanism that it presents for that model however is computationally expensive and requires a large stash, making ObliviStore unsuitable for implementing ACID transactions. Like ObliviStore, Taostore leverages asynchrony to parallelize Path ORAM [78], a tree-based construction from which Ring ORAM descends. Taostore, however, targets a different threat model: it assumes both that requests must be processed immediately, and that the timing of responses is visible to the adversary. Request latencies thus necessarily increase linearly with the number of clients [86].

Hiding access patterns for non-transactional systems

Many systems seek to provide access pattern protections for analytical queries: Opaque [89] and Cipherbase [5] support oblivious operators for queries that scan or shuffle full tables. Both rely on hardware enclaves for efficiency: Opaque runs a query optimizer in SGX [32], while Cipherbase leverages secure co-processors to evaluate predicates more efficiently. Others seek to hide the parameters of the query rather than the query itself: Olumofin et al. [56] do it via multiple rounds of keyword-based PIR operations [16]; Splinter [83] reduces the number of round-trips necessary by mapping these database queries to function secret sharing primitives. Finally, ObliDB [25] adds support for point queries and efficient updates by designing an oblivious B-tree for indexing. The concurrency control and recovery mechanisms of all these approaches introduce timing channels and structure writes in ways that leak access patterns [5].

Encryption

Many commercial systems offer the possibility to store encrypted data [24, 71]. Efficiently executing data-dependent queries like joins, filters, or aggregations without knowledge of the plaintext is challenging: systems like CryptDB [63], Monomi [81], and Seabed [60] tailor encryption schemes to allow executing certain queries directly on encrypted data. Others leverage trusted hardware [7]. In contrast, executing transactions on encrypted data is straightforward: neither concurrency control nor recovery requires knowledge of the plaintext data.

13 Conclusion

This paper presents Obladi, a system that, for the first time, considers the security challenges of providing ACID transactions without revealing access patterns. Obladi guarantees security and durability at moderate cost through a simple observation: transactional guarantees are only required to hold for committed transactions. By delaying commits until the end of epochs, Obladi inches closer to providing practical oblivious ACID transactions.

Acknowledgements We thank our shepherd, Jay Lorch, for his commitment to excellence, and the anonymous reviewers for their helpful comments. We are grateful to Sebastian Angel, Soumya Basu, Vijay Chidambaram, Trinabah Gupta, Paul Grubbs, Malte Schwarzkopf, Yunhao Zhang, and the MIT PDOS reading group for their feedback. This work was supported by NSF grants CSR-1409555 and CNS-1704742, and an AWS EC2 Education Research grant.

References


Cloud, C. 5 advantages of a cloud-based EHR.


[77] Stefanov, E., Shi, E., and Song, D. Towards Practical Oblivious RAM.


