Abstract

MongoDB is a popular general-purpose, document-oriented, distributed NoSQL database. It supports transactions in three different deployments: single-document transactions utilizing the WiredTiger storage engine in a standalone node, multi-document transactions in a replica set which consists of a primary node and several secondary nodes, and distributed transactions in a sharded cluster which is a group of multiple replica sets, among which data is sharded. A natural and fundamental question about MongoDB transactions is: What transactional consistency guarantee do MongoDB transactions in each deployment provide? However, it lacks both concise pseudocode of MongoDB transactions in each deployment and formal specification of the consistency guarantees which MongoDB claimed to provide. In this work, we formally specify and verify the transactional consistency protocols of MongoDB. Specifically, we provide a concise pseudocode for the transactional consistency protocols in each MongoDB deployment, namely WiredTiger, ReplicaSet, and ShardedCluster, based on the official documents and source code. We then prove that WiredTiger, ReplicaSet, and ShardedCluster satisfy different variants of snapshot isolation, namely StrongSI, RealtimeSI, and SessionSI, respectively. We also propose and evaluate efficient white-box checking algorithms for MongoDB transaction protocols against their consistency guarantees, effectively circumventing the NP-hard obstacle in theory.

1 Introduction

MongoDB is a popular general-purpose, document-oriented, distributed NoSQL database. A MongoDB database consists of a set of collections, a collection is a set of documents,
MongoDB achieves scalability by partitioning data into shards and fault-tolerance by replicating each shard across a set of nodes. MongoDB deployment is a sharded cluster, replica set, or standalone; see Figure 1. A standalone is a storage node that represents a single instance of a data store. A replica set consists of a primary node and several secondary nodes. A sharded cluster is a group of multiple replica sets, among which data is shared.

MongoDB transactions have evolved in three stages so far (Figure 1): In version 3.2, MongoDB used the WiredTiger storage engine as the default storage engine. Utilizing the Multi-Version Concurrency Control (MVCC) architecture of WiredTiger storage engine, MongoDB was able to support single-document transactions (called WiredTiger) in the standalone deployment. In version 4.0, MongoDB supported multi-document transactions (called REPLICASET) in replica sets. In version 4.2, MongoDB further introduced distributed (multi-document) transactions (called SHARDEDCLUSTER) in sharded clusters. Each kind of MongoDB transactions has advantages and disadvantages. In particular, distributed transactions should not be a replacement for multi-document transactions or single-document transactions, since “in most cases, (they) incur a greater performance cost over single document writes.”

A natural and fundamental question about MongoDB transactions is: What transactional consistency guarantee do MongoDB transactions in each deployment provide? This question poses three main challenges, in terms of specification, protocols, and checking algorithms:

- There are many variants of snapshot isolation in the literature. Though it was officially claimed that MongoDB implements a so-called speculative snapshot isolation protocol, it is unclear which kind of MongoDB transactions in different deployments satisfies which specific variant of snapshot isolation.
- It lacks concise pseudocode of the transactional consistency protocols of MongoDB in

---

Footnotes:

3 Roughly speaking, a document is an analog to a row in a relational database, and a collection is to a table.

4 Snapshots and Checkpoints. [https://docs.mongodb.com/manual/core/wiredtiger/#snapshots-and-checkpoints](https://docs.mongodb.com/manual/core/wiredtiger/#snapshots-and-checkpoints)

5 Transactions and Atomicity. [https://docs.mongodb.com/manual/core/transactions/#transactions-and-atomicity](https://docs.mongodb.com/manual/core/transactions/#transactions-and-atomicity)
different deployments, let alone the rigorous correctness proofs for them. Recently Biswas et al. proved that the problem of checking whether a given history without the version order satisfies (Adya) snapshot isolation \cite{1} is NP-complete \cite{5}. Therefore, it is challenging to efficiently check whether MongoDB in production satisfies some variant of snapshot isolation \cite{16}.

To answer the question above, we formally specify and verify the transactional consistency protocols of MongoDB. Specifically,

- We formally specify several variants of snapshot isolation in the well-known \((\text{vis, ar})\) specification framework for transactional consistency models proposed by Cerone et al. \cite{7}.
- We provide a concise pseudocode for the transactional consistency protocols in each MongoDB deployment, based on the official documents and source code.
- We prove that \textsc{WiredTiger}, \textsc{ReplicaSet}, and \textsc{ShardedCluster} satisfy \textsc{StrongSI}, \textsc{RealtimeSI}, and \textsc{SessionSI}, respectively (Figure 1). In particular, \textsc{RealtimeSI} and \textsc{SessionSI} are natural variants of snapshot isolation that we introduce for characterizing \textsc{ReplicaSet} and \textsc{ShardedCluster}, respectively.
- We design white-box polynomial-time checking algorithms for the transactional protocols of MongoDB against \textsc{StrongSI}, \textsc{RealtimeSI}, and \textsc{SessionSI}. These checking algorithms make use of the properties of the transactional protocols to infer the version order of histories, effectively circumventing the NP-hard obstacle in theory \cite{5}. We then intensively test the transactional consistency protocols of MongoDB using Jepsen \cite{6}. The results show that our checking algorithms are effective and efficiency.

The rest of the paper is organized as follows. Section 2 formally specifies several variants of snapshot isolation in the \((\text{vis, ar})\) specification framework for transactional consistency models. Sections 3, 4, and 5 describe the transactional consistency protocols for MongoDB transactions in each deployment, respectively. We also prove the correctness of them in the \((\text{vis, ar})\) framework in Section B of the Appendix. Section 6 proposes and evaluates the efficient white-box checking algorithms for MongoDB transactions in each deployment. Section 7 discusses related work. Section 8 concludes the paper with possible future work.

2 Snapshot Isolation

We consider a MongoDB deployment \(d \in \text{Dep} \triangleq \{\text{wt}, \text{rs}, \text{sc}\}\) managing a set of keys \(\text{Key}\), ranged over by \(\text{key}\), which take on a set of values \(\text{Val}\), ranged over by \(\text{val}\). We denote by \(\text{Op}\) the set of possible read or write operations on keys: \(\text{Op} = \{\text{read}(\text{key}, \text{val}), \text{write}(\text{key}, \text{val}) \mid \text{key} \in \text{Key}, \text{val} \in \text{Val}\}\). We assume that each key \(\text{key}\) has an dedicated initial value \(\text{key}_0 \in \text{Val}\).

Each invocation of an operation is denoted by an event from a set \(\text{Event}\), ranged over by \(e\) and \(f\). A function \(\text{op} : \text{Event} \rightarrow \text{Op}\) determines the operation a given event denotes. Below we follow the \((\text{vis, ar})\) specification framework proposed by \cite{7,8}.

2.1 Relations and Orderings

A binary relation \(R\) over a given set \(A\) is a subset of \(A \times A\), i.e., \(R \subseteq A \times A\). For \(a, b \in A\), we use \((a, b) \in R\) and \(a \xrightarrow{R} b\) interchangeably. The inverse relation of \(R\) is denoted by \(R^{-1}\), i.e., \((a, b) \in R \iff (b, a) \in R^{-1}\). We use \(R^{-1}(b)\) to denote the set \(\{a \in A \mid (a, b) \in R\}\). For some subset \(A' \subseteq A\), the restriction of \(R\) to \(A'\) is \(R|_{A'} \triangleq R \cap (A' \times A)\). For a relation \(R\) and
23:4 Verifying Transactional Consistency of MongoDB

Table 1 Consistency axioms, constraining an abstract execution \((H, \text{Vis}, \text{Ar})\). (Adapted from [3])

\[
\begin{align*}
\forall (E, po) \in H. \forall e \in \text{Event}. \forall \text{key}, \text{val}. \exists \text{op}(c) = \text{read}((\text{key}, \text{val}) \land \{ f | (\text{op}(f) = \text{commit}((\text{key}, ...) \land f = c) \neq \emptyset \}) \quad \text{(Int)}
\end{align*}
\]

\[
\forall T \in H. \forall \text{key}, \text{val}. T \vdash \text{read}((\text{key}, \text{val}) \Rightarrow \text{max(vis) }\text{write}(T, \text{key}, \text{val})) \quad \text{(Ext)}
\]

\[
\begin{align*}
\text{SO} \subseteq \text{vis} \quad \text{(Session)} & \quad \text{AR} \cup \text{vis} \subseteq \text{vis} \quad \text{(Prefix)} & \quad \text{VIS}, T \in H. S \equiv T \Rightarrow (S \rightarrow_{\text{AR}} T \vee T \rightarrow_{\text{vis}} S) \quad \text{(NoConflict)}
\end{align*}
\]

\[
\begin{align*}
\text{RB} \subseteq \text{vis} \quad \text{(ReturnBefore)} & \quad \text{VIS} \subseteq \text{RB} \quad \text{(IRReturnBefore)} & \quad \text{VIS} = \text{RB} \quad \text{(RealTimeSnapshot)} & \quad \text{CH} \subseteq \text{AR} \quad \text{(CommitBefore)}
\end{align*}
\]

A deployment \(d \in \text{Dep}\), we use \(R_d\) to denote an instantiation of \(R\) specific to this deployment \(d\). Given two binary relations \(R\) and \(S\) over set \(A\), we define the composition of them as \(R ; S = \{ (a, c) \mid \exists b \in A : a \rightarrow R b \rightarrow S c \}\). A strict partial order is an irreflexive and transitive relation. A total order is a relation which is a partial order and total.

2.2 Histories and Abstract Executions

Definition 1 (Transactions). A transaction is a pair \((E, po)\), where \(E \subseteq \text{Event}\) is a finite, non-empty set of events and \(po \subseteq E \times E\) is a strict total order called program order.

For simplicity, we assume a dedicated transaction that writes initial values of all keys. We also assume the existence of a time oracle that assigns distinct real-time start and commit timestamps to each transaction \(T\), and access them by \(\text{start}(T)\) and \(\text{commit}(T)\), respectively. For an transaction \(T\) in deployment \(d \in \text{Dep}\), we use \(\text{start}_d(T)\) (resp. \(\text{commit}_d(T)\)) to denote the instantiation of \(\text{start}(T)\) (resp. \(\text{commit}(T)\)) specific to the deployment \(d\). We define two strict partial orders involving real time on transactions.

Definition 2 (Returns Before). A transaction \(S\) returns before \(T\) in real time, denoted \(S \rightarrow_{\text{RB}} T\), if \(\text{commit}(S) < \text{start}(T)\).

Definition 3 (Commits Before). A transaction \(S\) commits before \(T\) in real time, denoted \(S \rightarrow_{\text{CB}} T\), if \(\text{commit}(S) < \text{commit}(T)\).

Clients interact with MongoDB by issuing transactions via sessions. We use a history to record the client-visible results of such interactions.

Definition 4 (Histories). A history is a pair \(H = (T, \text{So})\), where \(T\) is a set of transactions with disjoint sets of events and the session order \(\text{So} \subseteq T \times T\) is a union of strict total orders defined on disjoint sets of \(T\), which correspond to transactions in different sessions.

To justify each transaction in a history, we need to know how these transactions are related to each other. This is captured declaratively by the visibility and arbitration relations.

Definition 5 (Abstract Executions). An abstract execution is a tuple \(A = (T, \text{So}, \text{Vis}, \text{Ar})\), where \((T, \text{So})\) is a history, visibility \(\text{Vis} \subseteq T \times T\) is a strict partial order, and arbitration \(\text{Ar} \subseteq T \times T\) is a strict total order such that \(\text{Vis} \subseteq \text{Ar}\).

For \(H = (T, \text{So})\), we often shorten \((T, \text{So}, \text{Vis}, \text{Ar})\) to \((H, \text{Vis}, \text{Ar})\).

2.3 Consistency Axioms and (Adya) Snapshot Isolation

A consistency model is a set \(\Phi\) of consistency axioms constraining abstract executions. The model allows those histories for which there exists an abstract execution that satisfies the axioms: \(\text{Hist}_\Phi = \{ H \mid \exists \text{vis}, \text{ar}. (H, \text{Vis}, \text{Ar}) \models \Phi \}\).
We first briefly explain the consistency axioms that are necessary for defining (Adya) snapshot isolation [1], namely \textit{Int}, \textit{Ext}, \textit{Prefix}, and \textit{NoConflict} [7, 8]. In Section 2.4, we will introduce a few new consistency axioms and formally define several variants of snapshot isolation using them. Table 1 summarizes all the consistency axioms used in this paper. For $T = (E, po)$, we let $T \vdash \text{write}(key, val)$ if $T$ writes to $key$ and the last value written is $val$, and $T \vdash \text{read}(key, val)$ if $T$ reads from $key$ before writing to it and $val$ is the value returned by the first such read. We also use $\text{WriteTx}_key = \{T \mid T \vdash \text{write}(key, \_))\}$. Two transactions $S$ and $T$ conflicts, denoted $S \bowtie T$, if they write on the same key.

The \textit{internal consistency axiom} \textit{Int} ensures that, within a transaction, a read from a key returns the same value as the last write to or read from this key in the transaction. The \textit{external consistency axiom} \textit{Ext} ensures that an external read in a transaction $T$ from a key returns the value written by the last transaction in $ar$ among all the transactions that proceed $T$ in terms of $vis$ and write this key. The \textit{prefix axiom} \textit{Prefix} ensures that if the snapshot taken by a transaction $T$ includes a transaction $S$, than this snapshot also include all transactions that committed before $S$ in terms of $ar$. The \textit{no-conflict axiom} \textit{NoConflict} prevents concurrent transactions from writing on the same key.

The classic (Adya) snapshot isolation is then defined as follows.

\begin{definition}[\textit{SI} [7]]\textit{SI} = \textit{Int} \land \textit{Ext} \land \textit{Prefix} \land \textit{NoConflict}.
\end{definition}

\subsection{2.4 Variants of Snapshot Isolation}

There are several variants of snapshot isolation [10], including ANSI-SI [4], generalized snapshot isolation (GSI) [11], strong snapshot isolation (\textit{StrongSI}) [11], strong session snapshot isolation (\textit{SSS}) [11], parallel snapshot isolation (\textit{PSI}) [24], write snapshot isolation [28], non-monotonic snapshot isolation (\textit{NMSI}) [2], and prefix-consistent snapshot isolation [11]. We now formally define these SI variants in the \textit{(vis, ar)} framework and illustrate their similarities and differences with examples. For brevity, we concentrate on three variants that are concerned with MongoDB deployments, namely SessionSI, \textit{RealtimeSI}, and \textit{StrongSI}. Particularly, the latter two variants are new. More variants are explained in Appendix A.

Session snapshot isolation, denoted \textit{SessionSI}, requires a transaction to observe all the transactions that precedes it in its session.

\begin{definition}[\textit{SessionSI} [8]]\textit{SessionSI} = \textit{SI} \land \textit{Session}.
\end{definition}

\begin{example}[\textit{SI} vs. \textit{SessionSI}]
Consider the history $\mathcal{H}$ in Figure 2. To show that $\mathcal{H} \models \textit{SI}$, we construct an abstract execution $\mathcal{A} = (\mathcal{H}, \textit{vis}, \textit{ar})$ such that $\textit{vis}$ is transitive, $T_c \xrightarrow{\text{vis}} T_b \xrightarrow{\text{vis}} T_d \xrightarrow{\text{vis}} T_c \xrightarrow{\text{vis}} T_a$ holds, and $\textit{ar} = \textit{vis}$. It is straightforward to justify that $\mathcal{A} \models \textit{SI}$.
\end{example}
In this section, we describe the protocol WiredTiger of snapshot isolation implemented in WiredTiger. Table 2 summarizes the types and variables used in WiredTiger. For
now the readers should ignore the underlined and highlighted lines, which are needed for REPLICASet and SHARDEDCluster protocols, respectively. We reference pseudocode lines using the format algorithm#:line#. For conciseness, we write something like $S \leftarrow @ \oplus T$ to denote $S \leftarrow S \oplus T$, where $\oplus$ is an operator.

3.1 Key Designs

3.1.1 Transactions and the Key-Value Store

We assume that each WiredTiger transaction $txn \in WT\_TXN$ is associated with a unique transaction identifier $txn.tid$ from set $TID = \mathbb{N} \cup \{-1, \bot\}$. When a transaction $txn$ starts, it initializes $txn.tid$ to 0. The actual (non-zero) $txn.tid$ is assigned when its first update operation is successfully executed. $tid$ tracks the next monotonically increasing transaction identifier to be allocated. A transaction $txn$ with $txn.tid \neq 0$ may be aborted due to a conflict caused by a later update. When this happens, we set $txn.tid = -1$. Note that a read-only transaction $txn$ always has $txn.tid = 0$. We use dummy $\bot$ to indicate that there is no such a transaction.

We model the key-value store, denoted $store$, as a function which maps each key $key \in Key$ to a (possibly empty) list of pairs of the form $\langle tid, val \rangle$ representing that the transaction with $tid$ has written value $val$ on $key$.

3.1.2 Sessions

Clients interact with WiredTiger via sessions. Each client is bind to a single session with a unique session identifier $wt\_sid \in WT\_SID$. At most one transaction is active on a session at any time. The mapping $wt\_session$ maintains the currently active transaction on each session and $wt\_global$ records which transaction has obtained its identifier on which session.

3.1.3 The Visibility Rule

To guarantee snapshot isolation, each transaction $txn$ needs to identify the set of transactions that are visible to it throughout its lifecycle when it starts. Intuitively, each transaction $txn$ is only aware of all the transactions that have already been committed before it starts. To this end, the transaction $txn$ maintains

- $txn.concur$: the set of identifiers of currently active transactions that have obtained their identifiers; and
- $txn.limit$: the next transaction identifier (i.e., $tid$) when $txn$ starts.

The procedure $visible$ states that a transaction with $tid$ is invisible to another (active) transaction $txn$ if (line 1–40)

- it is aborted (i.e., $tid = -1$), or
- it is concurrent with $txn$ (i.e., $tid \in txn.concur$), or
- it starts after $txn$ and thus has a larger transaction identifier than $txn.limit$ (i.e., $tid \geq txn.limit$). Note that when $visible$ is called, $txn$ may have been assigned an identifier larger than $txn.limit$. The second conjunction $tid \neq txn.id$ allows $txn$ to observe itself.

**Example 12 (Visibility).** Consider the scenario in Figure 4, where transaction $T_i$ is on session $wt\_sid_i$ and would obtain its identifier $tid_i$ if any. We assume that $\forall 1 \leq i \leq 6. \; tid_i < tid_{i+1}$. Suppose that when $T_6$ starts, (1) $T_1$ and $T_4$ have been committed; (2) $T_2$ and $T_3$ are active and have obtained their identifiers; (3) $T_3$ has been aborted; and (4) $T_7$ has not started. We
Verifying Transactional Consistency of MongoDB

Figure 4 Illustration of the visibility rule; see Example 12.

Table 3 Types and variables used in ReplicaSet.

| RS_SID = N | TS: the set of timestamps | ct ∈ TS | oplog ← ∅ ∈ oplog* | rs_wt ∈ [RS_SID → WT_SID] |
|------------|--------------------------|--------|-------------------|--------------------------|
| oplog = [ts : TS, ops : (Key × Val)*] ∪ [ts : TS, cts : TS] | txn_mods ∈ [RS_SID → (Key × Val)*] |

have \(T_6\).concur = \{tid_2, tid_5\} and \(T_6\).limit = tid_6. According to the visibility rule, only \(T_1\) and \(T_4\) are visible to \(T_6\). Note that even when \(T_5\) commits later, it is still invisible to \(T_6\).

3.2 Protocol

For simplicity, we assume that each handler in the protocols executes atomically; see Section 8 for discussions.

3.2.1 Start Transactions

A client starts a transaction on a session \(wt\_sid\) by calling \(wt\_start\), which creates and populates a transaction \(txn\) (lines 1:2–1:5). Particularly, it scans \(wt\_global\) to collect the concurrently active transactions on other sessions into \(txn\).concur.

3.2.2 Read and Update Operations

To read from a key \(key\), we iterate over the update list \(store[\text{key}]\) forward and returns the value written by the first visible transaction (line 1:10).

To update a key \(key\), we first check whether the transaction, denoted \(txn\), should be aborted due to conflicts (lines 1:14–1:17). To this end, we iterates over the update list \(store[\text{key}]\). If there are updates on \(key\) made by transactions that are invisible to \(txn\) and are not aborted, \(txn\) will be rolled back. If \(txn\) passes the conflict checking, it is assigned a unique transaction identifier, i.e., \(tid\), in case it has not yet been assigned one (line 1:19). Finally, the key-value pair \(⟨key, val⟩\) is added into the modification set \(txn.mod\)s and is inserted at the front of the update list \(store[\text{key}]\).

3.2.3 Commit or Rollback Operations

To commit the transaction on session \(wt\_sid\), we simply resets \(wt\_global[\text{wt\_sid}]\) to \(⊥_{\text{tid}}\), indicating that there is currently no active transaction on this session (line 1:32). To roll back a transaction \(txn\), we additionally reset \(txn.tid\) in \(store\) to \(-1\) (line 1:38). Note that read-only transactions can always commit successfully.
Algorithm 1 WiredTiger: the snapshot isolation protocol in WiredTiger

1: procedure wt_start(wt_sid)
2:   txn ← (0, 0, 0, \perp, wt_sid)
3: for (_ → (tid, _)) ∈ wt_global ∧ tid ≠ \perp
4:   txn.commit ← ⋈ \cup \{tid\}
5:   txn.limit ← tid
6:   wt_session[wt_sid] ← txn

7: procedure wt_read(wt_sid, key)
8:   txn ← wt_session[wt_sid]
9: for (_, val, ts, phase) ∈ store[key]
10:   if visible(txn, tid, ts)
11:      return ⟨val, phase⟩

12: procedure wt_update(wt_sid, key, val)
13:   txn ← wt_session[wt_sid]
14: for (_, val, ts, phase) ∈ store[key]
15:   if ¬visible(txn, tid, ts) ∧ tid ≠ −1
16:      wt rollback(wt_sid)
17: return rollback
18: if txn.tid = 0
19:   txn.tid ← tid
20: tid ← ⋈ + 1
21: wt_global[wt_sid] ← ⟨txn.tid, wt⟩
22: txn.mods ← ⋈ \cup \{⟨key, val⟩\}
23: wt_session[wt_sid] ← txn
24: store[key] ← ⟨txn.tid, val, wt⟩ o ⋈
25: return ok

26: procedure wt_commit(wt_sid)
27:   txn ← wt_session[wt_sid]
28: for (_, _) ∈ txn.mods
29:   for (_, val, phase) ∈ store[key]
30:   if tid = txn.tid
31:      store[key] ← ⟨tid, val, txn.commit_ts, phase⟩
32:   end
33: wt_global[wt_sid] ← ⟨⊥, ⊥⟩

33: procedure wt_rollback(wt_sid)
34:   txn ← wt_session[wt_sid]
35: for (_, _) ∈ txn.mods
36:   for (_, val, phase) ∈ store[key]
37:   if tid = txn.tid
38:      store[key] ← ⟨−1, val, phase⟩
39: wt_global[wt_sid] ← ⟨⊥, ⊥⟩

40: procedure visible(txn, tid, ts)
41: return ¬(tid = −1 ∨ tid ∈ \perp ∧ (tid ≥ txn.limit ∧ tid ≠ txn.tid)) ∧ (ts ≠ ⊥ ∧ ts ≤ txn.read_ts)

42: procedure all_committed()
43: return the largest timestamp smaller than max_commit_ts and min(commit_ts | \langlecommit_ts, wt_global ∧ commit Ts ≠ \perp\rangle)}

44: procedure wt_set_read_ts(wt_sid, read_ts)
45:   wt_session[wt_sid].read_ts ← read_ts

46: procedure wt_set_commit_ts(wt_sid)
47:   txn ← wt_session[wt_sid]
48:   commit_ts ← txn.commit_ts
49: max_commit_ts ← max\{0, commit_ts\}
50: wt_global[wt_sid] ← ⟨txn.tid, commit_ts⟩
51: return ok

52: procedure wt_prepare(wt_sid, prepare_ts)
53:   txn ← wt_session[wt_sid]
54:   prepare_ts ← prepare_ts
55: for (_, _) ∈ txn.mods
56:   for (_, val, _) ∈ store[key]
57:   if tid = txn.tid
58:      store[key] ← ⟨tid, val, txn.prepare_ts, PREPARED⟩
59: wt_global[wt_sid] ← ⟨⊥, ⊥⟩
60: return ok

61: procedure wt_commit_prepare_ts(wt_sid, commit_ts)
62:   txn ← wt_session[wt_sid]
63:   commit_ts ← commit_ts
64:   wt_global[wt_sid] ← ⟨txn.tid, txn.commit_ts⟩
65: procedure wt_commit_prepare(wt_sid)
66:   txn ← wt_session[wt_sid]
67: for (_, _) ∈ txn.mods
68:   for (_, val, _) ∈ store[key]
69:   if tid = txn.tid
70:      store[key] ← ⟨tid, val, txn.commit_ts, COMMITTED⟩
71: wt_global[wt_sid] ← ⟨⊥, ⊥⟩
72: max_commit_ts ← max\{0, txn.commit_ts\}
4 Replica Set Transactions

We now describe the protocol ReplicaSet of snapshot isolation implemented in replica sets. Table 3 summarizes the types and variables used in ReplicaSet. For now the readers should ignore the highlighted lines, which are needed for the ShardedCluster protocol.

4.1 Key Designs

A replica set consists of a single primary node and several secondary nodes. All transactional operations, i.e., start, read, update, and commit, are first performed on the primary. Committed transactions are replicated to the secondaries via a leader-based consensus protocol similar to Raft [20, 30].

4.1.1 Hybrid Logical Clocks

ReplicaSet uses hybrid logical clocks (HLC) [17] as the read and commit timestamps of transactions. Without going into the details, we assume that HLCs are compared lexicographically and are thus totally ordered, and HLCs can be incremented via tick.

All nodes and clients maintain a cluster time ct, which is also an HLC [27]. They distribute their latest cluster time when sending any messages and update it when receiving a larger one in incoming messages. The cluster time is incremented (“ticks”) only when a ReplicaSet transaction is committed on the primary.

4.1.2 Speculative Snapshot Isolation

ReplicaSet implements a so-called speculative snapshot isolation protocol with readConcern = “snapshot” and writeConcern = “majority” [22]. It guarantees that each read obtains data that was majority committed in the replica set, and requires all updates be majority committed before the transaction completes. To reduce aborts due to conflicts in back-to-back transactions, ReplicaSet adopts an innovative strategy called “speculative majority” [22]. In this strategy, transactions read the latest data, instead of reading at a timestamp at or earlier than the majority commit point in WiredTiger. At commit time, they wait for all the data they read to become majority committed.

In implementation, it is unnecessary for update transactions to explicitly wait at commit time for the data read to become majority committed [22]. This is because we must wait for the updates in those transactions to be majority committed, which, due to the replication mechanism (Section 4.2.3), implies that the data read was also majority committed. Read-only transactions, however, need to issue a special “noop” operation at commit time and wait for it to be majority committed.

4.1.3 Read and Commit Timestamps

The primary node maintains an oplog of transactions, where each entry is assigned a unique commit timestamp. These commit timestamps determine the (logical) commit order of ReplicaSet transactions, no matter when they are encapsulated into WiredTiger transactions and are committed in WiredTiger.

---

7 For brevity, we omit the distribution of cluster time in pseudocode.
When a transaction starts, it is assigned a read timestamp on the primary such that all transactions with smaller commit timestamps have been committed in WiredTiger. That is, the read timestamp is the maximum point at which the oplog of the primary has no gaps. Specifically, in all_committed (line 112), the read timestamp of a transaction \( txn \) is computed as the largest timestamp smaller than the minimum of max_commit_ts and the set of commit timestamps of transactions concurrent with \( txn \), where max_commit_ts is the maximum commit timestamp that WiredTiger knows.

4.2 Protocol

Clients interacts with \textsc{ReplicaSet} via sessions. Each client is bind to a single session with a unique session identifier \( rs\_sid \in RS\_SID \), and at most one transaction is active on a session at any time. Each active \textsc{ReplicaSet} transaction on a session \( rs\_sid \) is encapsulated into a WiredTiger transaction on a new session \( wt\_sid \), as recorded in \( rs\_wt \).

4.2.1 Read and Update Operations

When the primary receives the first operation of an transaction (lines 24 and 211), it calls open_wt_session to open a new session \( wt\_sid \) to WiredTiger, start a new WiredTiger transaction on \( wt\_sid \), and more importantly set the transaction’s read timestamp.

The primary delegates the read/update operations to WiredTiger (lines 217 and 214). If an update succeeds, the \( \langle key, val \rangle \) pair is recorded in \( \text{txn\_mods}[rs\_sid] \) (line 216).

4.2.2 Commit Operations

To commit a transaction, the primary first atomically increments its cluster time \( ct \) via \textsc{tick}, takes it as the transaction’s commit timestamp (line 223), uses it to update max_commit_ts, and records it in \( wt\_global \) (lines 224 and 146).
If this is a read-only transaction, the primary appends a noop entry to its oplog (line 2:27, Section 4.1.2). Otherwise, it appends an entry containing the updates of the transaction. Each oplog entry is associated with the commit timestamp of the transaction. Then, the primary asks WiredTiger to locally commit this transaction in \texttt{wt_commit} (line 2:30), which associates the updated key-value pairs in store with the commit timestamp (line 1:31). Note that \texttt{wt_commit} needs not to be atomically executed with \texttt{tick} and \texttt{wt_set_commit_ts}.

Finally, the primary waits for all updates of the transaction to be majority committed (line 2:31). Specifically, it waits for \texttt{last\_majority\_committed} \(\geq ct\), where \texttt{last\_majority\_committed} is the timestamp of the last oplog entry that has been majority committed (discussed shortly).

### 4.2.3 Replication

Each secondary node periodically pulls the oplog entries with larger commit timestamps than \texttt{last\_pulled} from the primary (line 3:2). It appends the retrieved entries to its own oplog, updates \texttt{last\_pulled} accordingly, and sends a \texttt{replicate\_ack} to the primary. The primary maintains the timestamp of the last pulled oplog entry for each secondary \(s\) in \texttt{last\_pulled\_ack[\(s\)]}. Whenever it receives a \texttt{replicate\_ack} from a secondary, the primary updates \texttt{last\_majority\_committed} accordingly (line 3:12).

### 5 Sharded Cluster Transactions

This section describes the protocol \texttt{ShardedCluster} of snapshot isolation implemented in sharded clusters. Table 4 summarizes the types and variables used in \texttt{ShardedCluster}.

### 5.1 Key Designs

A client issues distributed transactions via a session connected to a mongos. The mongos, as a transaction router, uses its cluster time as the read timestamp of the transaction and forwards the transactional operations to corresponding shards. The shard which receives the first read/update operation of a transaction is designated as the transaction coordinator.
5.1.1 Two Phase Commit among Shards

If a transaction has not been aborted due to write conflicts in sc_update, the mongos can proceed to commit it. If this transaction is read-only, the mongos instructs each of the participants to directly commit locally via rs_commit; otherwise, the mongos instructs the transaction coordinator to perform a variant of two-phase commit (2PC) that always commits among all participants (line 4:9). Specifically, the coordinator sends a PREPARE message to all participants. After receiving the PREPARE message, a participant computes a local prepare timestamp and returns it to the coordinator in a PREPARE_ACK message. When the coordinator receives PREPARE_ACK messages from all participants, it calculates the transaction’s commit timestamp by taking the maximum of all prepare timestamps (line 4:14), and sends a COMMIT message to all participants. After receiving DEC_ACK messages from all participants, the coordinator replies to the mongos (line 4:18).

5.1.2 Replication within Replia Sets

ShardedCluster uses state-machine replication [21, 9] to achieve fault tolerance in replica sets. On the one hand, the transaction coordinator persists the participant information within its replica set (line 4:10) before sending the PREPARE messages, and the transaction’s commit timestamp (line 4:15) before sending the COMMIT messages. On the other hand, the primary of a shard waits for a quorum of secondary nodes to persist its oplog entry (line 4:34) before sending the PREPARE_ACK message, and the final decision (lines 4:42 and 4:46) before sending the DEC_ACK message.

5.1.3 Consistent Snapshots and Read Timestamps

ShardedCluster uses HLCs which are loosely synchronized to assign read and commit timestamps to transactions. Due to clock skew or pending commit, a transaction may receive a read timestamp from a mongos, but the corresponding snapshot is not yet fully available at transaction participants [13]. Therefore, ShardedCluster will delay the read/update operations until the snapshot becomes available. There are four cases [13].

1. (Case-Clock-Skew) When a transaction participant receives a read/update operation forwarded by the mongos and finds that its cluster time is behind the read timestamp, it first increments its cluster time to catch up. ShardedCluster achieves this by issuing a noop write with the read timestamp (line 4:51).

2. (Case-Holes) ShardedCluster transactions on a primary are not necessarily committed in the increasing order of their commit timestamps in WiredTiger. To guarantee snapshot isolation, we need to ensure that there are no “holes” in the oplog before the read timestamp. Therefore, the primary waits until ALLCOMMITTED() is larger than the read timestamp (line 4:52).

3. (Case-Pending-Commit-Read) Consider a read operation of a transaction \( t_{rn} \). By the visibility rule in WiredTiger, the read may observe an update of another transaction \( t_{rn'} \) which is prepared but not yet committed. To guarantee snapshot isolation, the read cannot be applied until \( t_{rn'} \) commits or aborts. To this end, ShardedCluster keeps trying the read until it returns a value without the PREPARED flag (line 2:25).

4. (Case-Pending-Commit-Update) Similarly, an update operation of a transaction may observe an update of another transaction which is prepared but not yet committed. ShardedCluster delays this update by first performing a read operation on the same key in the way described in Case-Pending-Commit-Read (line 2:12).
Verifying Transactional Consistency of MongoDB

Algorithm 4 ShardedCluster: the snapshot isolation protocol in sharded cluster (on the primary nodes)

```plaintext
1: when received sc_read(sc_sid, key)
2: sc_start(sc_sid)
3: val ← rs_read(sc_rs[sc_sid], key)
4: return val
5: when received sc_update(sc_sid, key, val)
6: sc_start(sc_sid)
7: status ← rs_update(sc_rs[sc_sid], key, val)
8: return status
9: when received 2pc(sc_sid) from m
10: wait until (sc_sid, shards[sc_sid]) majority committed in collection config.transaction_coords
11: P ← PRIMARY_OP(shards[sc_sid])
12: send PREPARE(sc_sid) to P
13: when receive PREPARE_ACK(prepare.ts_p) from p ∈ P
14: commit_ts ← max_p_p prepare.ts_p
15: send COMMIT(sc_sid, commit_ts) to P
16: when receive DEC_ACK(sc_sid) from P
17: send 2PC_ACK() to m
18: procedure sc_start(sc_sid)
19: if this is the first operation of the transaction received by the primary
20: read_ts ← read_ts[sc_sid]
21: wt_sid ← rs_wt[sc_rs[sc_sid]]
22: wt_set_read_ts(wt_sid, read_ts)
23: wait_for_read_concern(sc_sid)
24: procedure wait_for_read_concern(sc_sid)
25: when received PREPARE(sc_sid) from p
26: rs_sid ← sc_rs[sc_sid]
27: ct ← TICK()
28: wt_prepare(rs_wt[rs_sid], ct)
29: ops ←txn_mods[rs_sid]
30: if ops = ∅
31: oplog ← @ 0 (ct, noop)
32: else
33: oplog ← @ 0 (ct, ops)
34: wait until(ops)committed ≥ ct
35: send PREPARE_ACK(ct) to p
36: when received commit(sc_sid, commit_ts) from p
37: ct ← TICK()
38: wt_sid ← rs_wt[sc_rs[sc_sid]]
39: wt_prepare(sc_contact.ts, commit.ts)
40: send COMMIT(sc_sid, commit_ts) to P
41: when received abort(sc_sid) from m
42: wait until last_majority_committed ≥ ct
43: send DEC_ACK(sc_sid) to m
44: procedure case_start(sc_sid)
45: if this is the first operation of the transaction received by the primary
46: read_ts ← read_ts[sc_sid]
47: wt_sid ← rs_wt[sc_rs[sc_sid]]
48: wt_set_read_ts(wt_sid, read_ts)
49: wait_for_read_concern(sc_sid)
50: procedure wait_for_read_concern(sc_sid)
51: if ct < read_ts
52: oplog ← @ 0 (read_ts, noop)
53: wait until all_committed() ≥ read_ts
```

5.2 Protocols

For brevity, we focus on the behavior of the primary nodes in the sharded cluster. Consider a session sc_sid connected to a mongos. We use read_ts[sc_sid] to denote the read timestamp, assigned by the mongos, of the currently active transaction on the session.

5.2.1 Read and Update Operations

If this is the first operation the primary receives, it calls sc_start to set the transaction’s read timestamp in WiredTiger (line 23). In sc_start, it also calls wait_for_read_concern to handle Case-Clock-Skew and Case-Holes (line 24).

The primary then delegates the operation to ReplicaSet (lines 11 and 13). To handle Case-Pending-Commit-Read, rs_read has been modified to keep trying reading from WiredTiger until it returns a value updated by a committed transaction (line 28). To handle Case-Pending-Commit-Update, rs_update first performs an sc_read on the same key (line 21). Moreover, if the update fails due to write conflicts, the mongos will send an abort message to the primary nodes of all other participants, without entering 2PC.
Table 5 Cloud virtual machines used in our experiments.

| VMs           | Configuration                                                                 | OS       | Region       |
|---------------|-------------------------------------------------------------------------------|----------|--------------|
| VM-a[0-6]     | 3.10GHz Intel(R) Xeon(R) Platinum 8269CY CPU with 2 virtual cores and 4GB of RAM | Ubuntu 20.04 | Chengdu-A     |
| VM-b[0-2]     | 2.50GHz Intel(R) Xeon(R) Platinum 8269CY CPU with 1 virtual core and 1GB of RAM | Ubuntu 20.04 | Chengdu-B     |

Table 6 Configurations of MongoDB deployments.

| Deployment    | Version | Configuration                                                                 |
|---------------|---------|-------------------------------------------------------------------------------|
| Standalone    | WiredTiger 3.3.0 | A standalone WiredTiger storage engine in VM-a0.                             |
| Replica Set   | MongoDB 4.4.5 | A replica set of 5 nodes with VM-a1 as the primary and VM-a(2-5) as secondaries. |
| Sharded Cluster | MongoDB 4.4.5 | A cluster consisting of 1 config server and 2 shards. The config server, shard1, and shard2 are all replica sets, and deployed in VM-a(1-3), VM-a(4-6), and VM-b(0-2), respectively. |

5.2.2 Commit Operations

In 2PC, the transaction coordinator behaves as described in Sections 5.1.1 and 5.1.2 for atomic commitment and fault tolerance, respectively. We now explain how the participants handle the PREPARE and COMMIT messages in more detail.

After receiving a PREPARE message, the participant advances its cluster time and takes it as the prepare timestamp (lines [4.27] 4.28 1.54 and [1.58]). Note that the transaction’s tid in \( wt_{global} \) is reset to \( \_tid \) (line [1.59]). Thus, according to the visibility rule, this transaction is visible to other transactions that starts later in WiredTiger. Next, the participant creates an oplog entry containing the updates executed locally or a noop oplog entry for the “speculative majority” strategy (Section 4.1.2). Then, it waits until the oplog entry has been majority committed (line [4.34]).

When a participant receives a COMMIT message, it ticks its cluster time. After setting the transaction’s commit timestamp (line [4.39]), it asks WiredTiger to commit the transaction locally (line [4.40]). Note that the status of the transaction is changed to COMMITTED (line [1.70]). Thus, this transaction is now visible to other waiting transactions (line [2.23]). Then, the participant generates an oplog entry containing the commit timestamp and waits for it to be majority committed.

6 Checking Snapshot Isolation

In this section, we design and evaluate white-box polynomial-time checking algorithms for the transactional protocols of MongoDB against STRONGSI, REALTIME_SI, and SESSION_SI. The project can be found at [https://github.com/Tsunaou/MongoDB-SI-Checker](https://github.com/Tsunaou/MongoDB-SI-Checker).

6.1 White-box Checking Algorithms

6.1.1 Basic Ideas

The three checking algorithms work in the same manner. For example, in the checking algorithm for the WiredTiger protocol against STRONGSI, we first extract the \( \text{vis}_{wt} \) and \( \text{ar}_{wt} \) relations from a given history \( H \) according to Definitions 14 and 22. Then, we check whether the abstract execution \( A \equiv (H, \text{vis}_{wt}, \text{ar}_{wt}) \) satisfies all the axioms required by STRONGSI according to Definition 10. Since the total order \( \text{ar}_{wt} \) offers the version order \( \ll \), this checking, particularly for \( \text{EXT} \), can be easily done in polynomial time.
Table 7 Transaction generation parameters (supported by Jepsen).

| Parameters        | Default | Range          | Description                                      |
|-------------------|---------|----------------|--------------------------------------------------|
| #txn-num          | 3000    | {1000, 2000, 3000, 4000, 5000} | The total number of transactions. |
| #concurrency      | 9       | {3, 6, 9, 12, 15} | The number of clients.                           |
| #max-txn-len      | 12      | {4, 8, 12, 16, 20} | The maximum number of operations in each transaction. |

Fixed Parameters

| Parameters        | Value  | Description                                      |
|-------------------|--------|--------------------------------------------------|
| #key-count        | 10     | There are 10 distinct keys at any point for generation. |
| #max-writes-per-key| 128    | There are at most 128 updates on each key.       |
| #key-dist         | exponential | Probability distribution for keys.               |
| #read:update ratio| 1 : 1  | The default (and fixed) read update ratio in Jepsen. |
| #timeout          | 5s/10s/30s | The timeout for WiredTiger/ReplicaSet/ShardedCluster transactions. |

6.1.2 Implementation Considerations

We now explain how to obtain the additional information necessary for extracting the appropriate Vis and Ar relations for each protocol.

For each WiredTiger transaction txn, we record the real time when it starts and commits. They are taken as $\text{start}_\text{wt}(txn)$ and $\text{commit}_\text{wt}(txn)$, respectively. However, this poses a technical challenge involving real time: Due to the potential non-accurate records of physical time, it is possible that a transaction $txn'$ reads data written by another transaction $txn$, but $txn'$ starts before $txn$ commits, violating the InReturnBefore axiom. Therefore, we measure the degree of inaccuracy: we enumerate all pairs of transactions like $txn'$ and $txn$ above and take the maximum over all $\text{commit}_\text{wt}(txn) - \text{start}_\text{wt}(txn')$. We call it the “real time error” and report it in the experiments. To further reduce the impacts caused by the real time issue, we can also utilize some other information to fix Vis$_{WT}$ and Ar$_{WT}$. For example, Lemma 26 says that for any two conflicting transactions, the Ar$_{WT}$ relation between them can be defined by their transaction identifiers. Therefore, in the experiments, we also obtain the transaction identifiers of update transactions from the write-ahead log of WiredTiger, and omitted the histories which violate Lemma 26.

For ReplicaSet, we obtain the read timestamps of transactions from the mongod.log file, and commit timestamps from the oplog (stored in the collection oplog.rs) on the primary.

For ShardedCluster, we obtain the transactions’ read timestamps from the mongod.log file. For update transactions that involve only a single shard, their commit timestamps are recorded in the oplog of the primary of the shard. For a distributed transaction across multiple shards that enters 2PC, its commit timestamp is stored in the oplog entry on the primary of any shard involved.

6.2 Evaluations

6.2.1 Experimental Setup

To demonstrate the effectiveness and the efficiency of the white-box checking algorithms, we implement them and use them to test MongoDB in different deployments.

We utilize the Jepsen testing framework to schedule the execution of transactions in each deployment. Each MongoDB deployment is considered as a key-value store as described in Section 2. A group of clients concurrently issue random transactions to MongoDB. Table 7
summarizes both the tunable and fixed parameters for executions. We consider 10 histories for each combination of parameter values. Note that we use relatively long timeout values to reduce possible rollbacks triggered by client timeout, since we do not consider failures in the work.

Our experiments evaluate WiredTiger 3.3.0 and MongoDB 4.4.5. Tables 5 and 6 shows the configurations of the machines and the deployments. The average RTT in Region-A, Region-B, and between regions are 0.14ms, 0.19ms, and 0.86ms, respectively.

### 6.2.2 Experimental Results

First of all, the experimental results confirm our theoretical analysis: the transactional protocols on the replica set deployment and the sharded cluster deployment satisfy RealtimeSI and SessionSI, respectively. The transactional protocol in WiredTiger satisfies StrongSI, if we can tolerate the real time error within about 25ms, which is crucial for validating the ReturnBefore axiom.

Figure 5 shows the performance of our white-box checking algorithms. All experiments are performed on VM-a0. The three algorithms are fast to check the transactions for all deployments, since each consistency axiom can be efficiently checked. For example, it takes less than 2s to check against a replica set or sharded cluster history consisting of 5000 transactions. Therefore, they are quite efficient for checking the satisfaction of large-scale MongoDB transactional workloads against variants of snapshot isolation.

### 7 Related Work

**Consistency Models in MongoDB.** Schultz et al. [22] discussed the tunable consistency models in MongoDB, which allow users to select the trade-off between consistency and latency at a per operation level by choosing different readConcern, writeConcern, and readPreference parameters. Tyulenev et al. [27] discussed the design and implementation of causal consistency introduced by MongoDB 3.6, which provides session guarantees including read-your-writes, monotonic-reads, monotonic-writes, and writes-follow-reads [26, 6]. In this work, we are concerned about the more challenging transactional consistency of MongoDB.

**Specification and Verification of MongoDB.** The MongoDB Inc. has been actively working on the formal specification and verification of MongoDB. Zhou et al. [30] presented the design and implementation of MongoDB Raft, the pull-based fault-tolerant replication protocol in MongoDB. Schultz et al. [23] presented a novel logless dynamic reconfiguration protocol for the MongoDB replication system. Both protocols have been formally specified using TLA+ [19] and verified using the TLC model checker [29]. In this work, we formally specify
Verifying Transactional Consistency of MongoDB

and verify the transactional consistency protocols of MongoDB.

Transactional Consistency Checking. Biswas et al. [5] have studied the complexity issues of checking whether a given history of a database satisfies some transactional consistency model. Particularly, they proved that it is NP-complete to check the satisfaction of (Adya) snapshot isolation for a history without the version order. Gan et al. [15] proposed IsoDiff, a tool for debugging anomalies caused by weak isolation for an application. IsoDiff finds a representative subset of anomalies. Kingsbury et al. [16] presented Elle, a checker which infers an Adya-style dependency graph among client-observed transactions. By carefully choosing appropriate data types and operations, Elle is able to infer the version order of a history. Tan et al. [25] proposed Cobra, a black-box checking system of serializability. By leveraging the SMT solver MonoSAT [3] and other optimizations, Cobra can handle real-world online transactional processing workloads. In this work, we design white-box polynomial-time checking algorithms for the transactional protocols of MongoDB against StrongSI, RealtimeSI, and SessionSI. These checking algorithms make use of the properties of the transactional protocols to infer the version order of histories, effectively circumventing the NP-hard obstacle in theory [5].

8 Conclusion and Future Work

In this paper, we have formally specified and verified the transactional consistency protocols in each MongoDB deployment, namely WiredTiger, ReplicaSet, and ShardedCluster. We proved that they satisfy different variants of snapshot isolation, namely StrongSI, RealtimeSI, and SessionSI, respectively. We have also proposed and evaluated efficient white-box checking algorithms for MongoDB transaction protocols against their consistency guarantees.

Our work is a step towards formally verifying MongoDB, and creates plenty of opportunities for future research. We list some possible future work.

- First, for simplicity, we have assumed that each procedure executes atomically. However, the implementation of MongoDB is highly concurrent with intricate locking mechanisms [12], and needs to be modelled and verified more carefully.
- Second, we did not consider failures in this paper. MongoDB employs a pull-based Raft protocol for data replication and tolerates any minority of node failures [30]. We plan to verify and evaluate MongoDB Raft in the future work. We will also explore the fault-tolerance and recovery mechanisms in distributed transactions.
- Third, there is a gap between our (simplified) pseudocode and the real implementation of MongoDB. We plan to use model-based testing [12] to bridge such a gap.
- Fourth, MongoDB also supports non-transactional consistency, including tunable consistency [22] and causal consistency [27]. It is unclear what the consistency model is like when non-transactional operations are involved.
- Finally, it is interesting to explore the transactional consistency checking algorithms utilizing SMT solvers [5, 25].

References

1. Atul Adya and Barbara H. Liskov. Weak Consistency: A Generalized Theory and Optimistic Implementations for Distributed Transactions. PhD thesis, USA, 1999. AAI0800775.
2. Masoud Saeida Ardekani, Pierre Sutra, and Marc Shapiro. Non-monotonic snapshot isolation: Scalable and strong consistency for geo-replicated transactional systems. In Proceedings of the
3 Sam Bayless, Noah Bayless, Holger H. Hoos, and Alan J. Hu. Sat modulo monotonic theories. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI’15, page 3702–3709. AAAI Press, 2015.

4 Hal Berenson, Phil Bernstein, Jim Gray, Jim Melton, Elizabeth O’Neil, and Patrick O’Neil. A critique of ansi sql isolation levels. SIGMOD Rec., 24(2):1–10, May 1995. doi:10.1145/568271.223785

5 Ranadeep Biswas and Constantin Enea. On the complexity of checking transactional consistency. Proc. ACM Program. Lang., 3(OOPSLA), October 2019. doi:10.1145/3360591.

6 J. Brzezinski, C. Sobaniec, and D. Wawrzyniak. From session causality to causal consistency. In Proceedings of 12th Euromicro Conference on Parallel, Distributed and Network-Based Processing, pages 152–158, 2004. doi:10.1109/EMPDP.2004.1271440.

7 Andrea Cerone, Giovanni Bernardi, and Alexey Gotsman. A Framework for Transactional Consistency Models with Atomic Visibility. In Luca Aceto and David de Frutos Escrig, editors, 26th International Conference on Concurrency Theory (CONCUR 2015), volume 42 of Leibniz International Proceedings in Informatics (LIPIcs), pages 58–71, Dagstuhl, Germany, 2015. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. URL: http://drops.dagstuhl.de/opus/volltexte/2015/5375, doi:10.4230/LIPIcs.CONCUR.2015.58.

8 Andrea Cerone and Alexey Gotsman. Analysing snapshot isolation. J. ACM, 65(2), January 2018. doi:10.1145/3152396.

9 Gregory V. Chockler and Alexey Gotsman. Multi-shot distributed transaction commit. Distributed Comput., 34(4):301–318, 2021. doi:10.1007/s00446-021-00389-4.

10 Natacha Crooks, Youer Pu, Lorenzo Alvisi, and Allen Clement. Seeing is believing: A client-centric specification of database isolation. In Proceedings of the ACM Symposium on Principles of Distributed Computing, PODC ’17, page 73–82, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3087801.3087802.

11 Khuzaima Daudjee and Kenneth Salem. Lazy database replication with snapshot isolation. In Proceedings of the 32nd International Conference on Very Large Data Bases, VLDB ’06, page 715–726. VLDB Endowment, 2006.

12 A. Jesse Jiryu Davis, Max Hirschhorn, and Judah Schwimer. Extreme modelling in practice. Proc. VLDB Endow., 13(9):1346–1358, May 2020. doi:10.14778/3397230.3397233.

13 Jiaqing Du, Sameh Elnikety, and Willy Zwaenepoel. Clock-si: Snapshot isolation for partitioned data stores using loosely synchronized clocks. In Proceedings of the 2013 IEEE 32nd International Symposium on Reliable Distributed Systems, SRDS ’13, page 173–184, USA, 2013. IEEE Computer Society. doi:10.1109/SRDS.2013.26.

14 Sameh Elnikety, Willy Zwaenepoel, and Fernando Pedone. Database replication using generalized snapshot isolation. SRDS ’05, page 73–84, USA, 2005. IEEE Computer Society. doi:10.1109/RELDIS.2005.14.

15 Yifan Gan, Xuexuan Ren, Drew Ripberger, Spyros Blanas, and Yang Wang. Isodiff: Debugging anomalies caused by weak isolation. Proc. VLDB Endow., 13(12):2773–2786, July 2020. doi:10.14778/3407790.3407860.

16 Kyle Kingsbury and Peter Alvaro. Elle: Inferring isolation anomalies from experimental observations. Proc. VLDB Endow., 14(3):268–280, November 2020.

17 Sandeep S Kulkarni, Murat Demirbas, Deepak Madappa, Bharadwaj Avva, and Marcelo Leone. Logical physical clocks. In International Conference on Principles of Distributed Systems (OPODIS’2014), pages 17–32, 2014.

18 Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. Commun. ACM, 21(7):558–565, jul 1978. doi:10.1145/359545.359563.

19 Leslie Lamport. Specifying Systems: The TLA+ Language and Tools for Hardware and Software Engineers. Addison-Wesley Longman Publishing Co., Inc., USA, 2002.
Generalized snapshot isolation (denoted GSI), which is shown equivalent to ANSI-SI in [10], limits a transaction $T$ to read only from snapshots that do not include transactions that committed in real time after $T$ starts (i.e., InReturnBefore), and requires the total order respect the commit order (i.e., CommitBefore).

**Definition 13.**

$$GSI = SI \land \text{InReturnBefore} \land \text{CommitBefore}.$$
B Correctness Proofs

B.1 Correctness of WiredTiger

Consider a history \( \mathcal{H} = (WT\_TXN, so_{WT}) \) of WiredTiger, where we restrict WT\_TXN to be set of committed transactions. Denote by WT\_TXN\_UPDATE \( \subseteq WT\_TXN \) be the set of update transactions in WT\_TXN. We prove that \( \mathcal{H} \) satisfies StrongSI by constructing an abstract execution \( \mathcal{A} = (\mathcal{H}, \text{vis}_{WT}, \text{AR}_{WT}) \) (Theorem 29).

B.1.1 The Visibility Relation

\textbf{Definition 14} \((\text{vis}_{WT})\). \( \text{vis}_{WT} \triangleq \text{RB}_{WT} \).

\textbf{Lemma 15.} \( A \models \text{RealTimeSnapshot} \).

\textbf{Lemma 16.} \( \forall \text{txn}, \text{txn}' \in \text{WT\_TXN\_UPDATE}. \text{txn} \xrightarrow{\text{vis}_{WT}} \text{txn}' \implies \text{txn}.\text{tid} < \text{txn}'.\text{tid} \).

\textbf{Proof.} Consider any two transactions \( \text{txn}, \text{txn}' \in \text{WT\_TXN} \) that have obtained non-zero transaction identifiers such that \( \text{txn} \xrightarrow{\text{vis}_{WT}} \text{txn}' \). By Definition 14 of \( \text{vis}_{WT} \), \( \text{txn} \xrightarrow{\text{RB}_{WT}} \text{txn}' \). By lines 119 and 120, \( \text{txn}.\text{tid} < \text{txn}'.\text{tid} \). \( \blacksquare \)

\textbf{Definition 17} \((\text{wt}_\text{vis})\). For a transaction \( \text{txn} \in \text{WT\_TXN} \), we define \( \text{wt}_\text{vis}(\text{txn}) \subseteq \text{WT\_TXN\_UPDATE} \) to be the set of update transactions (excluding \( \text{txn} \) itself) that are visible to \( \text{txn} \) according to the visibility rule in \text{VISIBLE}.

\textbf{Lemma 18.} For any transaction \( \text{txn} \in \text{WT\_TXN} \), \( \text{wt}_\text{vis}(\text{txn}) \) is well-defined. That is, \( \text{wt}_\text{vis}(\text{txn}) \) does not change over the lifecycle of \( \text{txn} \).

\textbf{Proof.} Consider a transaction \( \text{txn} \in \text{WT\_TXN} \) which starts at time \( \tau \). Denote by \( \text{wt}_\text{vis}_\tau(\text{txn}) \) the set of transactions that are visible to \( \text{txn} \) at time \( \tau \). Let \( \text{txn}' \neq \text{txn} \in \text{WT\_TXN\_UPDATE} \) be an update transaction on session \( wt\_sid \). Suppose that \( \text{txn}' \) obtains its transaction identifier \( \text{txn}'.\text{tid} \) at time \( \tau' \) and finishes at time \( \tau'' \). We distinguish between two cases.

\text{CASE I:} \( \tau < \tau'' \). There are two cases. If \( \tau' < \tau \), we have \( \text{wt}_\text{global}[wt\_sid] \neq \perp_{\text{tid}} \) at time \( \tau \). Therefore, \( \text{txn}'.\text{tid} \in \text{txn}.\text{concur} \). Since \( \text{txn}.\text{concur} \) does not change, \( \text{txn}' \) is visible to \( \text{txn} \) over the lifecycle of \( \text{txn} \). If \( \tau < \tau' \), then \( \text{txn}'.\text{tid} \geq \text{txn}.\text{limit} \). Since \( \text{txn}.\text{limit} \) does not change, \( \text{txn}' \) is invisible to \( \text{txn} \) over the lifecycle of \( \text{txn} \).

\text{CASE II:} \( \tau'' < \tau \). Therefore, \( \text{txn}'.\text{tid} \notin \text{txn}.\text{concur} \land \text{txn}'.\text{tid} < \text{txn}.\text{limit} \). Since either \( \text{txn}.\text{concur} \) or \( \text{txn}.\text{limit} \) does not change, \( \text{txn}' \) is visible to \( \text{txn} \) over the lifecycle of \( \text{txn} \). \( \blacksquare \)

By Lemma 13, we can extend the definition of \( \text{wt}_\text{vis} \) to events.

\textbf{Definition 19} \((\text{wt}_\text{vis} \text{ for Events})\). Let \( e \) be an event of transaction \( \text{txn} \). We abuse the notation \( \text{wt}_\text{vis}(e) \) to denote the set of update transactions that are visible to \( e \) when it occurs. We have \( \text{wt}_\text{vis}(e) = \text{wt}_\text{vis}(\text{txn}) \).

\textbf{Lemma 20.} \( \forall \text{txn} \in \text{WT\_TXN}. \text{vis}_{\text{WT}^{-1}}(\text{txn}) \cap \text{WT\_TXN\_UPDATE} = \text{wt}_\text{vis}(\text{txn}) \).

\textbf{Proof.} We need to show \( \text{RB}_{\text{WT}^{-1}}(\text{txn}) \cap \text{WT\_TXN\_UPDATE} = \text{wt}_\text{vis}(\text{txn}) \).

We first show that \( \text{wt}_\text{vis}(\text{txn}) \subseteq \text{RB}_{\text{WT}^{-1}}(\text{txn}) \cap \text{WT\_TXN\_UPDATE} \). Consider an update transaction \( \text{txn}' \in \text{wt}_\text{vis}(\text{txn}) \). We need to show that \( \text{txn}' \xrightarrow{\text{RB}_{\text{WT}^{-1}}} \text{txn} \). Suppose by contradiction that \( \text{txn} \) starts before \( \text{txn}' \) commits. By the same argument in (CASE I) of the proof of Lemma 18, \( \text{txn}' \notin \text{wt}_\text{vis}(\text{txn}) \). Contradiction.

Next we show that \( \text{RB}_{\text{WT}^{-1}}(\text{txn}) \cap \text{WT\_TXN\_UPDATE} \subseteq \text{wt}_\text{vis}(\text{txn}) \). Consider an update transaction \( \text{txn}' \in \text{RB}_{\text{WT}^{-1}}(\text{txn}) \). By the same argument in (CASE II) of the proof of Lemma 18, \( \text{txn}' \in \text{wt}_\text{vis}(\text{txn}) \). \( \blacksquare \)
Lemma 21. \( A \models \text{NoConflict} \).

Proof. Consider any two update transactions \( txn, txn' \in \text{WT}_\text{TXN}_\text{UPDATE} \) such that \( txn \lessdot txn' \). Without loss of generality, assume that they both update key \( key \), and that \( txn \) updates \( key \) before \( txn' \) does. By Definition 14 of \( \text{vis}_{\text{WT}} \), \( \neg (txn' \xrightarrow{\text{vis}_{\text{WT}}} txn) \). We show that \( txn \xrightarrow{\text{vis}_{\text{WT}}} txn' \). Suppose by contradiction that \( \neg (txn \xrightarrow{\text{vis}_{\text{WT}}} txn') \). By Lemma 20 \( txn \notin \text{wt}_{\text{vis}}(txn') \). Therefore, \( txn' \) would abort when it updates \( key \).

B.1.2 The Arbitrary Relation

Definition 22 (\( AR_{\text{WT}} \)). \( AR_{\text{WT}} \triangleq CB_{\text{WT}} \).

Lemma 23. \( A \models \text{CommitBefore} \).

Lemma 24. \( \text{vis}_{\text{WT}} \subseteq AR_{\text{WT}} \).

Proof. By Definition 14 of \( \text{vis}_{\text{WT}} \) and Definition 22 of \( AR_{\text{WT}} \), \( \text{vis}_{\text{WT}} = \text{RB}_{\text{WT}} \subseteq CB_{\text{WT}} = AR_{\text{WT}} \).

Lemma 25. \( A \models \text{Prefix} \).

Proof. By Definition 14 of \( \text{vis}_{\text{WT}} \), Definition 22 of \( AR_{\text{WT}} \), Definition 2 of \( \text{RB} \), and Definition 3 of \( \text{CB} \), \( AR_{\text{WT}} = \text{vis}_{\text{WT}} = \text{RB}_{\text{WT}} \subseteq CB_{\text{WT}} = \text{vis}_{\text{WT}} \).

Lemma 26. \( \forall txn, txn' \in \text{WT}_\text{TXN}. \, txn \lessdot txn' \implies (txn \xrightarrow{AR_{\text{WT}}} txn' \iff txn.\text{tid} < txn'.\text{tid}) \).

Proof. Consider any two transactions \( txn, txn' \in \text{WT}_\text{TXN} \) such that \( txn \lessdot txn' \). By Lemma 21 \( txn \xrightarrow{\text{vis}_{\text{WT}}} txn' \lor txn' \xrightarrow{\text{vis}_{\text{WT}}} txn \). In the following, we proceed in two directions.

Suppose that \( txn \xrightarrow{AR_{\text{WT}}} txn' \). By Lemma 24 \( \neg (txn' \xrightarrow{\text{vis}_{\text{WT}}} txn) \). Therefore, \( txn \xrightarrow{\text{vis}_{\text{WT}}} txn' \).

Suppose that \( txn.\text{tid} < txn'.\text{tid} \). By Lemma 16 it cannot be that \( \neg (txn' \xrightarrow{\text{vis}_{\text{WT}}} txn) \). Therefore, \( txn \xrightarrow{\text{vis}_{\text{WT}}} txn' \).

Lemma 27. \( A \models \text{Int} \).

Proof. Consider a transaction \( txn \in \text{WT}_\text{TXN} \). Let \( e \) be an event such that \( \text{op}(e) = \text{read}(key, val) \) is an \text{internal} \( \text{wt}_\text{READ} \) operation in \( txn \). Let \( e' \triangleq \max_{\text{po}} \{f \mid \text{op}(f) = \_ (key, \_) \land f \xrightarrow{\text{po}} e\} \). We need to show that \( e' = \_ (key, val) \).

By Lemma 15 \( \text{wt}_{\text{vis}}(e') = \text{wt}_{\text{vis}}(e) \). If \( \text{op}(e') = \text{read}(key, \_) \), \( e \) obtains the same value as \( e' \). Therefore, \( \text{op}(e') = \text{read}(key, val) \). If \( \text{op}(e') = \text{write}(key, \_) \), by the visibility rule, \( e' \) is visible when \( e \) occurs. Thus, \( e \) reads from \( e' \). Therefore, \( \text{op}(e') = \text{write}(key, val) \).

Lemma 28. \( A \models \text{Ext} \).

Proof. Consider a transaction \( txn \in \text{WT}_\text{TXN} \). Let \( e \) be an event such that \( \text{op}(e) = \text{read}(key, val) \) is an \text{external} \( \text{wt}_\text{READ} \) operation in \( txn \). Let \( W \triangleq \text{vis}_{\text{WT}}^{-1}(txn) \cap \text{Write}_{\text{key}} \) be the set of transactions that update key \( key \) and are visible to \( txn \). If \( W = \emptyset \), obviously \( e \) obtains the initial value of \( key \). Now suppose that \( W \neq \emptyset \). By Lemma 26 \( W = \text{wt}_{\text{vis}}(txn) \cap \text{Write}_{\text{key}} \). By Lemma 29 the \( AR_{\text{WT}} \mid W \) order is consistent with the increasing \( \text{tid} \) order of the transactions in \( W \), which is also the list order at line 110. Therefore, \( e \) reads from \( \max_{AR_{\text{WT}}} W \). Thus, \( \max_{AR_{\text{WT}}} W \vdash \text{write}(key, val) \).

Theorem 29. \( \text{WiredTiger} \models \text{StrongSI} \).
Proof. For any history $H$ of WiredTiger, we construct an abstract execution $A = (H, \text{vis}_{\text{wt}}, \text{ar}_{\text{wt}})$, where $\text{vis}_{\text{wt}}$ and $\text{ar}_{\text{wt}}$ are given in Definitions 14 and 22 respectively. By Lemmas 13, 21, 23, 25, 27, and 28, $A \models \text{STRONGSI}$. Since $H$ is arbitrary, WiredTiger $\models \text{STRONGSI}$.

B.2 Correctness of ReplicaSet

Consider a history $H = (\text{RS\_TXN}, \text{soc}_\text{rs})$ of ReplicaSet, where we restrict RS\_TXN to be set of committed transactions. Denote by $\text{RS\_TXN\_UPDATE} \subseteq \text{RS\_TXN}$ be the set of update transactions in RS\_TXN. We prove that $H$ satisfies RealtimeSI by constructing an abstract execution $A = (H, \text{vis}_\text{rs}, \text{ar}_\text{rs})$ (Theorem 48).

Lemma 30. Transactions are majority committed in the increasing order of their commit timestamps.

Proof. By the replication mechanism (Algorithm 3).

Definition 31 (vis$_\text{rs}$). $\forall \text{txn}, \text{txn}' \in \text{RS\_TXN}. \text{txn} \xrightarrow{\text{vis}_\text{rs}} \text{txn}' \iff \text{txn.commit}_\text{ts} \leq \text{txn}'\text{read}_\text{ts}$

Lemma 32. vis$_\text{rs}$ is acyclic.

Proof. This holds due to the property that $\forall \text{txn} \in \text{RS\_TXN}. \text{txn.read}_\text{ts} < \text{txn.commit}_\text{ts}$ (line 21).

Lemma 33. $A \models \text{RETURNBEFORE}$.

Proof. Consider any two transactions $\text{txn}, \text{txn}' \in \text{RS\_TXN}$ such that $\text{txn} \xrightarrow{\text{vis}_\text{rs}} \text{txn}'$. By line 23 when $\text{txn}'$ starts, $\text{txn}$ has been majority committed. By Lemma 30, all transactions with commit timestamps $\leq \text{txn.commit}_\text{ts}$ have been majority committed and thus locally committed on the primary. Therefore, when $\text{txn}'\text{read}_\text{ts}$ is computed (line 237), all $\text{commit}_\text{ts} \neq \bot_\text{rs}$ in wt.global are larger than $\text{txn.commit}_\text{ts}$ (line 143). Thus, $\text{txn.commit}_\text{ts} \leq \text{txn}'\text{read}_\text{ts}$. By Definition 31 of vis$_\text{rs}$, $\text{txn} \xrightarrow{\text{vis}_\text{rs}} \text{txn}'$.

Since ReplicaSet transactions are encapsulated into WiredTiger transactions, we can extend vis$_\text{wt}$ over ReplicaSet transactions. Generally speaking, the following lemma shows that the timestamps of ReplicaSet overrides the transaction identifiers of WiredTiger.

Lemma 34. vis$_\text{rs} \subseteq$ vis$_\text{wt}$.

Proof. Consider two transactions $\text{txn}$ and $\text{txn}'$ in RS\_TXN such that $\text{txn} \xrightarrow{\text{vis}_\text{rs}} \text{txn}'$. By Definition 31 of vis$_\text{rs}$, $\text{txn.commit}_\text{ts} \leq \text{txn}'\text{read}_\text{ts}$. We need to show that $\text{txn} \xrightarrow{\text{vis}_\text{wt}} \text{txn}'$, which, by Definition 14 of vis$_\text{wt}$, is $\text{commit}_\text{wt}(\text{txn}) < \text{start}_\text{wt}(\text{txn}')$.

Consider the following three time points: $\tau_1$ when $\text{txn}$ obtained its commit timestamp $\text{txn.commit}_\text{ts}$ (line 223) and atomically set $\text{txn.commit}_\text{ts}$ in WiredTiger (line 150), $\tau_2 \triangleq \text{commit}_\text{wt}(\text{txn})$ when $\text{txn}$ was locally committed in WiredTiger on the primary (line 126), and $\tau_3 \triangleq \text{start}_\text{wt}(\text{txn}')$ when $\text{txn}'$ computed its read timestamp $\text{txn}'\text{read}_\text{ts}$ (lines 237 and 142).

We need to show that $\tau_2 < \tau_3$. Note that $\tau_1 < \tau_2$.

We first show that $\tau_1 < \tau_3$. Suppose by contradiction that $\tau_3 < \tau_1$. Since the commit timestamps of ReplicaSet transactions are strictly monotonically increasing, by the way $\text{txn}'\text{read}_\text{ts}$ is computed, $\text{txn}'\text{read}_\text{ts} < \text{txn.commit}_\text{ts}$. Contradiction.

Next we show that $\tau_2 < \tau_1$. Suppose by contradiction that $\tau_3 < \tau_2$. Since $\tau_1 < \tau_3$, $\tau_1 < \tau_3 < \tau_2$. That is, at time $\tau_3$, $\text{txn}$ has obtained its commit timestamp, but has not been
locally committed. By the way $txn'.read_ts$ is computed, $txn'.read_ts < txn.commit_ts$. Contradiction.

Definition 35. For a transaction $txn \in RS\_TXN$, we define $rs\_vis txn$ to be the set of update transactions in $RS\_TXN\_UPDATE$ that are visible to $txn$ according to the visibility rule in VISIBLE.

Lemma 36. For any transaction $txn \in RS\_TXN$, $rs\_vis txn$ is well-defined. That is, $rs\_vis txn$ does not change over the lifecycle of $txn$.

Proof. By Lemma 41 and the fact that the read timestamp and commit timestamp of a transaction do not change over its lifecycle.

As with $wt\_vis$, by Lemma 36 we can extend the definition of $rs\_vis$ to events.

Definition 37 (rs\_vis for Events). Let be an event of transaction $txn$. We abuse the notation $rs\_vis\langle e \rangle$ to denote the set of update transactions that are visible to e when it occurs. We have $rs\_vis\langle e \rangle = rs\_vis\langle txn \rangle$.

Lemma 38. $\forall txn \in RS\_TXN. \ vis\_rs\_vis\langle \langle texn \rangle \cap RS\_TXN\_UPDATE = rs\_vis\langle txn \rangle$.

Proof. Consider a transaction $txn \in RS\_TXN$. By Definition 31 of $vis\_rs\_vis\langle \langle texn \rangle \cap RS\_TXN\_UPDATE = \{txn' \in RS\_TXN\_UPDATE \mid txn'.commit_ts \leq txn.read_ts\}$. By the visibility rule (line 140) and Lemmas 34 and 36 $rs\_vis\langle txn \rangle = \{txn' \in RS\_TXN\_UPDATE \mid txn'.commit_ts \leq txn.read_ts\}$. Thus, $vis\_rs\_vis\langle \langle texn \rangle \cap RS\_TXN\_UPDATE = rs\_vis\langle txn \rangle$.

Lemma 39. $\mathcal{A} \models No\text{CONFLICT}$.

Proof. Consider any two transactions $txn, txn' \in RS\_TXN$ such that $txn \triangleright\triangleright txn'$. Suppose that they both update key $key$. Suppose by contradiction that $\neg((txn \triangleright\triangleright txn' \triangleright\triangleright txn)$. By Definition 31 of $vis\_rs\_vis\langle \langle texn \rangle \cap RS\_TXN\_UPDATE = \{txn' \in RS\_TXN\_UPDATE \mid txn'.commit_ts \leq txn.read_ts\}$. By the visibility rule (line 140) and Lemmas 34 and 36 $rs\_vis\langle txn \rangle = \{txn' \in RS\_TXN\_UPDATE \mid txn'.commit_ts \leq txn.read_ts\}$. Thus, $vis\_rs\_vis\langle \langle texn \rangle \cap RS\_TXN\_UPDATE = rs\_vis\langle txn \rangle$.

The Arbitrary Relation

Definition 40 (AR\_rs). $\forall txn1, txn2 \in RS\_TXN. txn1 \xrightarrow{AR\_rs} txn2 \iff\txn1.commit_ts < txn2.commit_ts$.

Lemma 41. $AR\_rs$ is a total order.

Proof. By line 221 the commit timestamps of all RS\_TXN transactions are unique.

Lemma 42. $vis\_rs \subseteq AR\_rs$.

Proof. Consider any two transactions $txn, txn' \in RS\_TXN$ such that $txn \xrightarrow{vis\_rs} txn'$. By Definition 31 of $vis\_rs\langle \langle texn \rangle \cap RS\_TXN\_UPDATE = \{txn' \in RS\_TXN\_UPDATE \mid txn'.commit_ts \leq txn.read_ts\}$. Since $txn'.read_ts < txn'.commit_ts$, we have $txn.commit_ts < txn'.commit_ts$. By Definition 40 of $AR\_rs\langle \langle texn \rangle \cap RS\_TXN\_UPDATE = \{txn' \in RS\_TXN\_UPDATE \mid txn'.commit_ts \leq txn.read_ts\}$. Thus, $vis\_rs\langle \langle texn \rangle \cap RS\_TXN\_UPDATE = rs\_vis\langle txn \rangle$.

Lemma 43. $\mathcal{A} \models Prefix$. 
Proof. Consider transactions \(txn_1, txn_2, txn_3 \in RS_{\text{TXN}}\) such that \(txn_1 \xrightarrow{\text{commit}} txn_2 \xrightarrow{\text{vis}} txn_3\). By Definition 40 of \(AR_{\text{vis}}\) and Definition 31 of \(\text{vis}_{\text{rs}}\), \(txn_1.\text{commit}\_\text{ts} < txn_2.\text{commit}\_\text{ts} \leq txn_3.\text{read}\_\text{ts}\). By Definition 31 of \(\text{vis}_{\text{rs}}\), \(txn_1 \xrightarrow{\text{vis}_{\text{rs}}} txn_3\).

► Lemma 44. \(A \models \text{CommitBefore}\).

Proof. Consider two transactions \(txn, \: txn' \in RS_{\text{TXN}}\) such that \(txn \xrightarrow{\text{commit}} txn'\). By line 2 in \(\text{commit}\_\text{ts} < txn'.\text{commit}\_\text{ts}\). By Definition 40 of \(AR_{\text{vis}}\), \(txn \xrightarrow{\text{vis}_{\text{rs}}} txn'\).

► Lemma 45. \(\forall txn, \: txn' \in RS_{\text{TXN}}. \: txn \triangleright txn' \implies (txn \xrightarrow{\text{vis}_{\text{rs}}} txn' \iff txn.\text{tid} < txn'.\text{tid})\).

Proof. Consider any two transactions \(txn, \: txn' \in RS_{\text{TXN}}\) such that \(txn \triangleright txn'\). By Lemma 39, \(txn \xrightarrow{\text{vis}_{\text{rs}}} txn' \lor txn' \xrightarrow{\text{vis}_{\text{rs}}} txn\). In the following, we proceed in two directions.

Suppose that \(txn \xrightarrow{\text{vis}_{\text{rs}}} txn'\). By Lemma 42, \(\neg (txn' \xrightarrow{\text{vis}_{\text{rs}}} txn)\). Therefore, \(txn \xrightarrow{\text{vis}_{\text{rs}}} txn'\). By Lemma 16, \(txn.\text{tid} < txn'.\text{tid}\).

Suppose that \(txn.\text{tid} < txn'.\text{tid}\). By Lemma 42, \(\neg (txn' \xrightarrow{\text{vis}_{\text{rs}}} txn)\). By Lemma 42, \(txn \xrightarrow{\text{vis}_{\text{rs}}} txn'\).

► Lemma 46. \(A \models \text{Int}\).

Proof. Consider a transaction \(txn \in RS_{\text{TXN}}\). Let \(e\) be an event such that \(\text{op}(e) = \text{read}(\text{key}, \text{val})\) is an \emph{internal rs_read} operation in \(txn\). Let \(e' = \max_{\text{ps}} \{f \mid \text{op}(f) = \_\text{read}(\_\text{key}, \_\text{val}) \land f.\text{op} = e\}\). We need to show that \(e' = \_\text{read}(\_\text{key}, \_\text{val})\).

By Lemma 36, \(\text{rs\_vis}(e') = \text{rs\_vis}(e)\). If \(\text{op}(e') = \_\text{read}(\_\text{key}, \_\text{val})\), \(e\) obtains the same value as \(e'\). Therefore, \(\text{op}(e') = \_\text{read}(\_\text{key}, \_\text{val})\). If \(\text{op}(e') = \_\text{write}(\_\text{key}, \_\text{val})\), by the visibility rule, \(e'\) is visible when \(e\) occurs. Thus, \(e\) reads from \(e'\). Therefore, \(\text{op}(e') = \_\text{write}(\_\text{key}, \_\text{val})\).

► Lemma 47. \(A \models \text{Ext}\).

Proof. Consider a transaction \(txn \in RS_{\text{TXN}}\). Let \(e\) be an event such that \(\text{op}(e) = \_\text{read}(\_\text{key}, \_\text{val})\) is an \emph{external rs_read} operation in \(txn\). Let \(W = \text{vis}_{\text{rs}}^{-1}(txn) \cap \text{WriteTx}_{\text{key}}\) be the set of transactions that update key \(\text{key}\) and are visible to \(txn\). If \(W = \emptyset\), obviously \(e\) obtains the initial value of \(\text{key}\). Now suppose that \(W \neq \emptyset\). By Lemma 38, \(W = \text{rs\_vis}(txn) \cap \text{WriteTx}_{\text{key}}\). By Lemma 45, the \(AR_{\text{vis}}\) order is consistent with the increasing \(\text{tid}\) order of the transactions in \(W\), which is also the list order at line 11. Therefore, \(e\) reads from \(\max_{\text{vis}_{\text{rs}}} W\). Thus, \(\max_{\text{vis}_{\text{rs}}} W \vdash \_\text{write}(\_\text{key}, \_\text{val})\).

Consider a transaction \(txn \in RS_{\text{TXN}}\). Let \(e = \_\text{read}(k, v)\) be an \emph{external rs_read} operation in \(txn\). Consider the set \(W = \text{vis}_{\text{rs}}^{-1}(txn) \cap \{txn \mid txn \vdash \_\text{write}(k, \_\text{val})\}\) of transactions that update key \(k\) and are visible to \(txn\). If \(W\) is empty, obviously \(e\) obtains the initial value of \(k\). Now suppose that \(W \neq \emptyset\). By Lemma 38, it is also the set of transactions that pass the check of \(\text{visible}(txn, \_\text{key})\) for key \(k\) (line 11).

► Theorem 48. \(\text{ReplicaSet} \models \text{RealtimeSI}\).

Proof. For any history \(H\) of \(\text{ReplicaSet}\), we construct an abstract execution \(A = (H, \text{vis}_{\text{rs}}, AR_{\text{rs}})\), where \(\text{vis}_{\text{rs}}\) and \(AR_{\text{rs}}\) are given in Definitions 31 and 40 respectively. By Lemmas 32, 33, 39, 41, 42, 43, 44, 46 and 47, \(A \models \text{RealtimeSI}\). Since \(H\) is arbitrary, \(\text{ReplicaSet} \models \text{RealtimeSI}\).

Additionally, we illustrate by counterexample that the abstract execution \(A\) constructed in the proof of Theorem 48 does not satisfy \(\text{INReturnBefore}\).
Consider a transaction $txn \in RS_{TXN}$ that has committed locally, but has not been majority committed. Let $txn' \neq txn$ be a transaction that starts before $txn$ finishes. That is, $\neg (txn \xrightarrow{RS_{TXN}} txn')$. Due to the “speculative majority” strategy, it is possible for $txn'$ to read from $txn$ so that $txn \xrightarrow{vis_{RS}} txn'$. Therefore, $A \not\models \text{InReturnBefore}$.

### B.3 Correctness of ShardedCluster

Consider a history $H = (SC_{TXN}, so_{sc})$ of SHARDEDCLUSTER, where we restrict $SC_{TXN}$ to be set of committed transactions. Denote by $SC_{UPDATE_{TXN}} \subseteq SC_{TXN}$ and $SC_{RO_{TXN}} \subseteq SC_{TXN}$ be the set of update transactions and read-only transactions in $SC_{TXN}$, respectively. We prove that $H$ satisfies SESSIONS by constructing an abstract execution $A = (H, vis_{sc}, ar_{sc})$ (Theorem 65).

In SHARDEDCLUSTER, all transactions have a read timestamp. However, only update transactions will be assigned a commit timestamp in 2PC. For the proof, we need to assign commit timestamps to read-only transactions as well.

**Definition 50** (Commit Timestamps for Read-only SHARDEDCLUSTER Transactions). Let $txn \in SC_{RO_{TXN}}$ be a read-only transaction. We define $txn.commit_{ts} = txn.read_{ts}$.

We also use Lamport clocks [18] of transactions as usual to break ties when necessary. Let $txn \in RS_{TXN}$ be a transaction. We denote by $lc(txn)$ its Lamport clock.

#### B.3.1 The Visibility Relation

**Definition 51** ($vis_{sc}$).

\[
\forall txn, txn' \in SC_{TXN}. \ xrightarrow{vis_{sc}} \iff (txn.commit_{ts} < txn'.read_{ts}) \lor (txn.commit_{ts} = txn'.read_{ts} \land lc(txn) < lc(txn')).
\]

**Lemma 52.** $vis_{sc}$ is acyclic.

**Proof.** By the property that $\forall txn \in SC_{UPDATE_{TXN}}. \ xrightarrow{vis_{sc}} < xrightarrow{vis_{sc}}$ and that Lamport clocks are totally ordered.

**Lemma 53.** $A \models \text{SESSION}$.

**Proof.** This holds due to the way how nodes and clients maintain their cluster time (namely, they distribute their latest cluster time when sending any messages and update it when receiving a larger one in incoming messages; see Section 4.1.1) and the properties of Lamport clocks.

**Definition 54.** For a transaction $txn \in SC_{TXN}$, we define $sc_{vis}(txn) \subseteq SC_{UPDATE_{TXN}}$ to be the set of update transactions in $SC_{TXN}$ that are visible to $txn$ according to the visibility rule in $visible$ (line 140).

**Lemma 55.** For any transaction $txn \in SC_{TXN}$, $sc_{vis}(txn)$ is well-defined. That is, $sc_{vis}(txn)$ does not change over the lifecycle of $txn$.

**Lemma 56.** $\forall txn \in SC_{TXN}. \ sc_{vis}^{-1}(txn) \cap SC_{UPDATE_{TXN}} = sc_{vis}(txn)$.

**Lemma 57.** $A \models \text{NoConflict}$. 
Proof. Consider any two transactions \(txn, txn' \in SC\_TXN\) such that \(txn \triangleright txn'\). Suppose that they both update key \(k\) on the primary of some shard. Suppose by contradiction that \(\neg (txn \overset{\text{vis}}{\triangleright} txn' \lor txn' \overset{\text{vis}}{\triangleright} txn)\). By Definition 51 of \(\text{vis}_{sc}\), that is \((txn.\text{read}\_ts < txn'.\text{commit}\_ts) \land (txn'.\text{read}\_ts < txn.\text{commit}\_ts)\).

By Lemma 21, \(txn\)'s \text{wt\_commit} finishes before \(txn'\)'s \text{wt\_start} starts or \(txn'\)'s \text{wt\_commit} finishes before \(txn\)'s \text{wt\_start} starts. These two cases are symmetric. In the following, we consider the first case which implies that \(txn\) updates key before \(txn'\) does. When \(txn'\) updates key, \(\langle txn.\text{tid}, _, txn.\text{commit}\_ts \rangle \in \text{store}[k]\) (line 111). However, the check \(\text{visible}(txn', _, txn.\text{commit}\_ts)\) (line 115) fails because \(txn'.\text{read}\_ts < txn.\text{commit}\_ts\). Therefore, \(txn'\) would abort (line 116).

B.3.2 The Arbitrary Relation

\text{Definition 58 (AR}_{sc}).

\[\forall txn, txn' \in SC\_TXN. \ tx{\text{nx}} \overset{\text{AR}_{sc}}{\triangleright} tx{\text{n}}' \iff (txn.\text{commit}\_ts < txn'.\text{commit}\_ts) \lor (txn.\text{commit}\_ts = txn'.\text{commit}\_ts \land \text{lc}(txn) < \text{lc}(txn'))\]

\text{Lemma 59.} \(AR_{sc}\) is a total order.

\text{Proof.} It is easy to show that \(AR_{sc}\) is irreflexive, transitive, and total. 

\text{Lemma 60.} \(\text{vis}_{sc}\) \(\subseteq\) \(AR_{sc}\).

\text{Proof.} By a case analysis based on Definition 51 of \(\text{vis}_{sc}\) and Definition 58 of \(AR_{sc}\).

\text{Lemma 61.} \(A \models \text{PREFIX}\).

\text{Proof.} Consider transactions \(txn_1, txn_2, txn_3 \in SC\_TXN\) such that \(txn_1 \overset{\text{AR}_{sc}}{\triangleright} txn_2 \overset{\text{vis}_{sc}}{\triangleright} txn_3\). It is easy to show that \(txn_1 \overset{\text{vis}_{sc}}{\triangleright} txn_3\) by a case analysis based on Definition 58 of \(AR_{sc}\) and Definition 51 of \(vis_{sc}\).

\text{Lemma 62.} \(\forall txn, txn' \in SC\_TXN. \ tx{\text{n}} \triangleright tx{\text{n}}' \implies (txn \overset{\text{AR}_{sc}}{\triangleright} txn' \iff txn.\text{tid} < txn'.\text{tid})\).

\text{Lemma 63.} \(A \models \text{INT}\).

\text{Proof.} Consider a transaction \(txn \in SC\_TXN\). Let \(e = (\_, \text{read}(k, v))\) be an internal \(\text{RS\_READ}\) operation in \(txn\). Moreover, \(po^{-1}(e) \cap \text{HEvent}_{\text{e}} \neq \emptyset\). Let \(e' \triangleq \max_{po}(po^{-1}(e) \cap \text{HEvent}_{\text{e}})\). By Lemma 55 compared with \(e'\), no additional update operations on \(k\) (from other transactions) are visible to \(e\) at line 110. Thus, \(e' = (\_, -(k, v))\).

\text{Lemma 64.} \(A \models \text{EXT}\).

\text{Proof.} Consider a transaction \(txn \in SC\_TXN\). Let \(e = (\_, \text{read}(k, v))\) be an external \(\text{SC\_READ}\) operation in \(txn\). Consider the set \(W = \text{vis}_{sc}^{-1}(txn) \cap \{txn | \ txn \vdash \text{Write} k : \_\}\) of transactions that update key \(k\) and are visible to \(txn\). If \(W\) is empty, obviously \(e\) obtains the initial value of \(key\). Now suppose that \(W \neq \emptyset\). By Lemma 56, it is also the set of transactions that pass the check of \(\text{visible}(txn, \_\) for key \(k\) (line 110). By Lemma 62 the \(AR_{sc}|_W\) order is consistent with the increasing tid order of the transactions in \(W\), which is the list order at line 110. Therefore, \(\max_{AR_{sc}} W \vdash \text{Write} k : v\).

\text{Theorem 65.} \(\text{SHARDEDCluster} \models \text{SESSIONSI}\).
Proof. For any history $H$ of ShardedCluster, we construct an abstract execution $A = (H, \text{vis}_{sc}, \text{ar}_{sc})$, where $\text{vis}_{sc}$ and $\text{ar}_{sc}$ are given in Definitions 51 and 58 respectively. By Lemmas 52, 53, 57, 59, 60, 61, 63, and 64, $A \models \text{SessionSI}$. Since $H$ is arbitrary, $\text{ShardedCluster} \models \text{SessionSI}$. ◀