

# Oh-RAM! One and a Half Round Atomic Memory

Theophanis Hadjistasi <sup>\*</sup>    Nicolas Nicolaou <sup>†</sup>    Alexander Schwarzmann<sup>\*</sup>

September 9, 2021

## Abstract

Implementing atomic read/write shared objects in a message-passing system is an important problem in distributed computing. Considering that communication is the most expensive resource, efficiency of read and write operations is assessed primarily in terms of the needed communication and the associated latency. The seminal result of Attiya, Bar-Noy, and Dolev established that two communication round-trip phases involving in total *four* message exchanges are sufficient to implement atomic operations when a majority of processors are correct. Subsequently it was shown by Dutta et al. that one round involving *two* communication exchanges is sufficient as long as the system adheres to certain constraints with respect to crashes on the number of readers and writers in the system. It was also observed that three message exchanges are sufficient in some settings.

This work explores the possibility of devising algorithms where operations are able to complete in *three* communication exchanges without imposing constraints on the number of participants, i.e., the aim is *One and half Round Atomic Memory*, hence the name OHRAM! A recent result by Hadjistasi et al. suggests that three-exchange implementations are *impossible* in the MWMR (multi-writer/multi-reader) setting. This paper shows that this is achievable in the SWMR (single-writer/multi-reader) setting and also achievable for read operations in the MWMR setting by “sacrificing” the performance of write operations. In particular, we present an atomic SWMR memory implementation, where reads complete in *three* and writes complete in *two* communication exchanges. Next, we provide an atomic MWMR memory implementation, where reads involve *three* and writes involve *four* communication exchanges. In light of the impossibility result these algorithms are optimal in terms of the number of communication exchanges. Both algorithms are then refined to allow some reads to complete in just *two* communication exchanges. To evaluate these algorithms we use the NS3 simulator and compare their performance in terms of operation latency. The algorithms are evaluated with different topologies and operation loads.

---

<sup>\*</sup>University of Connecticut, Storrs CT, USA. Email: [theo@uconn.edu](mailto:theo@uconn.edu), [aas@engr.uconn.edu](mailto:aas@engr.uconn.edu)

<sup>†</sup>IMDEA Networks Institute, Madrid, Spain. Email: [nicolas.nicolaou@imdea.org](mailto:nicolas.nicolaou@imdea.org)

# 1 Introduction

Emulating atomic [8] (or linearizable [7]) read/write objects in message-passing environments is an important problem in distributed computing. Atomicity is the most intuitive consistency semantic as it provides the illusion of a single-copy object that serializes all accesses such that each read operation returns the value of the latest preceding write operation. Solutions to this problem are complicated when the processors are failure-prone and when the environment is asynchronous. To cope with processor failures, distributed object implementations use *redundancy* by replicating the object at multiple network locations. Replication introduces the problem of consistency because operations may access different object replicas possibly containing obsolete values.

The seminal work of Attiya, Bar-Noy, and Dolev [2] provided an algorithm, colloquially referred to as ABD, that implements single-writer/multiple-reader (SWMR) atomic objects in message-passing crash-prone asynchronous environments. The operations are ordered with the help of logical *timestamps* associated with each value. Here each operation is guaranteed to terminate as long as some majority of replica servers do not crash. Each write operation takes one communication round-trip phase, or round, involving *two* communication exchanges and each read operation takes two rounds involving in total *four* communication exchanges. Subsequently, [10] showed how to implement multi-writer/multiple-reader (MWMR) atomic memory where both read and write operations involve two communication round trips involving in total four communication exchanges.

The work by Dutta et al. [3] introduced a SWMR implementation where both reads and writes involve a single round consisting of *two* communication exchanges. Such an implementation is called *fast*, and it was shown that this is possible only when the number of readers  $r$  is bounded with respect to the number of servers  $s$  and the number of server failures  $f$ , viz.  $r < \frac{s}{f} - 2$ . An observation made in [3] suggests that atomic memory may be implemented (using a max/min technique) so that each read and write operation complete in *three* communication exchanges. The authors did not elaborate on the inherent limitations that such a technique may impose on the distributed system.

Subsequent works, e.g., [4, 5], focused in relaxing the bound on the number of readers and writers in the service by proposing hybrid approaches where some operations complete in *one* and others in *two* rounds. Tight bounds were provided in [4] on the number of rounds that read and write operations require in the MWMR model.

A natural question arises whether one can devise implementations where all operations complete in at most *three* communication exchanges without imposing any restrictions on the numbers of participants in the service. A recent work by Hadjistasi, Nicolaou, and Schwarzmam [6] showed that such implementations are impossible in the MWMR setting. It is not known whether there is an SWMR implementation and whether there exists some trade off that allows operations to complete in three communication exchanges in the MWMR setting.

Model	Algorithm	Read Exchanges	Write Exchanges	Read Comm.	Write Comm.
SWMR	ABD	4	2	$4 \mathcal{S} $	$2 \mathcal{S} $
SWMR	OhSAM	3	2	$ \mathcal{S} ^2 + 2 \mathcal{S} $	$2 \mathcal{S} $
SWMR	OhSAM'	2 or 3	2	$ \mathcal{S} ^2 + 3 \mathcal{S} $	$2 \mathcal{S} $
MWMR	ABD	4	4	$4 \mathcal{S} $	$4 \mathcal{S} $
MWMR	OhMAM	3	4	$ \mathcal{S} ^2 + 2 \mathcal{S} $	$4 \mathcal{S} $
MWMR	OhMAM'	2 or 3	4	$ \mathcal{S} ^2 + 3 \mathcal{S} $	$4 \mathcal{S} $

Table 1: Summary of communication exchanges and communication complexities.

**Contributions.** We focus on the gap between one-round and two-round algorithms by presenting atomic memory algorithms where read operations can take “one and a half rounds,” i.e., complete in *three* communication exchanges. We also provide SWMR and MWMR algorithms where read operations complete in either *two* or *three* communication exchanges. We rigorously reason about the correctness of the algorithms. To assess the practicality of these implementations we simulate them and compare their performance. Additional details are as follows.

1. We present a new SWMR algorithm (OHSAM) for atomic objects in the asynchronous message-passing model with processor crashes. Write operations take *two* communication exchanges and are similar to the write operations of ABD. Read operations take *three* communication exchanges: (1) the reader sends a message to servers, (2) the servers share this information, and (3) once this is “sufficiently” done, servers reply to the reader. A key idea of the algorithm is that the reader returns the value that is associated with the *minimum* timestamp (cf. the observation in [3]). The read operations are optimal in terms of communication exchanges in light of [6]. (Section 3.)
2. We extend the SWMR algorithm to yield a MWMR algorithm (OHMAM). In the new algorithm the write operations are more complicated, taking *four* communication exchanges (cf. [10]). Read operations complete as before in *three* communication exchanges. (Section 4.)
3. We then present a revised SWMR algorithm (OHSAM’) and a revised MWMR algorithm (OHMAM’), where read operations complete in either *two* or *three* communication exchanges. The original and the revised versions of each algorithm are presented for pedagogical reasons: for ease of understanding and reasoning about the algorithms. (Section 5.)
4. We simulate our algorithms using the NS3 simulator and assess their performance under practical considerations. We note that the relative performance of our algorithms depends on the simulation topologies and object server placement; this is another reason for presenting both versions of each algorithm. (Section 6.)

The summary of complexity results in comparison with ABD [2] is in Table 1. Improvements in the latency (in terms of the number of exchanges) are obtained in a trade-off with communication complexity. We note that increases in the communication complexity need not necessarily have negative consequences in some practical settings, such as data centers, where servers communicate over high-bandwidth links.

## 2 Models and Definitions

The system consists of a collection of crash-prone, asynchronous processors with unique identifiers from a totally-ordered set  $\mathcal{I}$  partitioned into: set  $\mathcal{W}$  of writer identifiers, set  $\mathcal{R}$  of reader identifiers, and set  $\mathcal{S}$  of replica server identifiers with each *server* maintaining a copy of the object. Any subset of writers and readers, and up to  $f$  servers,  $f < \frac{|\mathcal{S}|}{2}$ , may crash at any time. Processors communicate by exchanging messages via asynchronous point-to-point reliable channels; messages may be reordered. For convenience we use the term *broadcast* as a shorthand denoting sending point-to-point messages to multiple destinations.

**Executions.** An algorithm  $A$  is a collection of processes, where process  $A_p$  is assigned to processor  $p \in \mathcal{I}$ . The *state* of processor  $p$  is determined over a set of state variables, and the state of  $A$  is

a vector that contains the state of each process. Algorithm  $A$  performs a *step*, when some process  $p$  (i) receives a message, (ii) performs local computation, (iii) sends a message. Each such action causes the state at  $p$  to change. An *execution* is an alternating sequence of states and actions of  $A$  starting with the initial state and ending in a state. A process  $p$  *crashes* in an execution if it stops taking steps; otherwise  $p$  is *correct*.

**Atomicity.** An implementation of a read or a write operation contains an *invocation* action (such as a call to a procedure) and a *response* action (such as a return from the procedure). An operation  $\pi$  is *complete* in an execution  $\xi$ , if  $\xi$  contains both the invocation and the *matching* response actions for  $\pi$ ; otherwise  $\pi$  is *incomplete*. An execution is *well formed* if any process invokes one operation at a time. We say that an operation  $\pi$  *precedes* an operation  $\pi'$  in an execution  $\xi$ , denoted by  $\pi \rightarrow \pi'$ , if the response step of  $\pi$  appears before the invocation step in  $\pi'$  in  $\xi$ . Two operations are *concurrent* if neither precedes the other. The correctness of an atomic read/write object implementation is defined in terms of *atomicity* (safety) and *termination* (liveness) properties. Termination requires that any operation invoked by a correct process eventually completes. Atomicity is defined following [9]. For any execution  $\xi$ , if all invoked read and write operations are complete, then the operations can be partially ordered by an ordering  $\prec$ , so that the following properties are satisfied:

- A1** The partial order  $\prec$  is consistent with the external order of invocation and responses, that is, there do not exist operations  $\pi$  and  $\pi'$ , such that  $\pi$  completes before  $\pi'$  starts, yet  $\pi' \prec \pi$ .
- A2** All write operations are totally ordered and every read operation is ordered with respect to all writes.
- A3** Every read operation returns the value of the last write preceding it in the partial order, and any read operation ordered before all writes returns the initial value of the object.

**Efficiency and Communication Exchanges.** Efficiency of implementations is assessed in terms of *operation latency* and *message complexity*. *Latency* of each operation is determined by the *computation time* and the *communication delays*. Computation time accounts for the computation steps that the algorithm performs in each operation. Communication delays are measured in terms of *communication exchanges*. The protocol implementing each operation involves a collection of sends (or broadcasts) of typed messages and the corresponding receives. *Communication exchange* within an execution of an operation is the set of sends and receives for the specific message type within the protocol. Note that using this definition, traditional implementations in the style of ABD are structured in terms of *rounds*, cf. [2, 5], where each round consists of two communication exchanges, the first, a broadcast, is initiated by the process executing an operation, and the second, a convergecast, consists of responses to the initiator. We refer to the  $i^{th}$  exchange using the notation  $E_i$ . The number of messages that a process expects during a convergecast depends on the implementation. *Message complexity* measures the worst-case total number of messages exchanged during an operation.

### 3 SWMR Algorithm OHSAM

We now present our SWMR algorithm OHSAM: *One and a Half round Single-writer Atomic Memory*. The write operations are fast, that is, they take *two* communication exchanges to complete (similarly to ABD [2]). We show that atomic operations do not need to involve complete communication round

trips between clients and servers. In particular, we allow *server-to-server* communication and we devise read operations that take *three* communication exchanges using the following communication pattern: exchange E1 the reader sends message to the participating servers, in exchange E2 each server that receives the request it then *relays* the request to all the servers, and once a server receives the relay for a particular read from a majority of servers, it replies to the requesting reader forming exchange E3. The read completes once the invoker collects a majority of acknowledgment replies. A key idea of the algorithm is that the reader returns the value that is associated with the *minimum* timestamp. In particular, while the replica servers update their local value to the associated with the *maximum* timestamp received, the reader returns the value associated with the *minimum* timestamp discovered in the set of the received acknowledgment messages. The code is given in Algorithm 1. We now give the details of the protocols; in referring to the numbered lines of code we use the prefix “L” to stand for “line”.

Counter variables *read\_op*, *operations* and *relays* are used to help processes identify “new” read and write operations, and distinguish “fresh” from “stale” messages (since messages can be reordered). The value of the object and its associated timestamp, as known by each process, are stored in variables *v* and *ts* respectively. Set *rAck*, at each reader *r<sub>i</sub>*, stores all the received acknowledgment messages. Variable *minTS* holds the minimum timestamp discovered in the set of the received acknowledgment messages *rAck*. Below we provide a brief description of the protocol of each participant of the service.

**Writer Protocol.** Writer *w* increments its local timestamp *ts* and broadcasts a `writeRequest` message to all the participating servers  $\mathcal{S}$  (L24-26). Once the writer receives replies from at least a majority of servers,  $|\mathcal{S}|/2 + 1$ , the operation completes (L27-27).

**Reader Protocol.** When a read process *r* invokes a read operation it first monotonically increases its local read operation counter *read\_op* and empties the set of the received acknowledgment messages, *rAck* (L8). Then, it creates a  $\langle \text{readRequest}, r, \text{read\_op} \rangle$  `readRequest` message in which it encloses its id and local read counter and it broadcasts this request message to all the participating servers  $\mathcal{S}$ , forming exchange E1 (L10). It then waits to collect at least  $|\mathcal{S}|/2 + 1$  messages from servers. While collecting `readAck` messages from exchange E3, reader *r* discards any delayed messages from previous operations due to asynchrony. When “fresh” messages are collected from a majority of servers, then the reader returns the value *v* associated with the *minimum* timestamp, *minTS*, among the set of the received acknowledgment messages, *rAck* (L12-14).

**Server Protocol.** Each server  $s \in \mathcal{S}$  expects three types of messages:

(1) Upon receiving a  $\langle \text{readRequest}, r, \text{read\_op} \rangle$  message the server creates a `readRelay` message, containing its *ts*, *v*, and its id *s*, and it broadcasts it to all the servers  $\mathcal{S}$  (L39-40).

(2) Upon receiving message  $\langle \text{readRelay}, ts', v', r, \text{read\_op} \rangle$  server *s* compares its local timestamp *ts* with *ts'* enclosed in the message. If  $ts < ts'$ , then *s* sets its local timestamp and value to those enclosed in the message (L46-47). In any other case, no updates are taking place. As a next step *s* checks if the received `readRelay` message marks a new read operation by *r*. This is achieved by checking if reader’s *r* operation counter is newer than the local one, i.e.,  $\text{read\_op} > \text{operations}[r]$  (L48). If this holds, then *s*: a) sets its local read operation counter for reader *r* to be equal to the received counter, i.e.,  $\text{operations}[r] = \text{read\_op}$ ; and b) re-initializes the relay counter for *r* to zero, i.e.,  $\text{relays}[r] = 0$  (L48-50). Server *s* also updates the number of collected `readRelay` messages regarding the read request created by reader *r* (L51-52). When *s* receives  $\langle \text{readRelay}, ts, v, \text{read\_op}, s_i \rangle$  messages from a majority of servers, it creates a  $\langle \text{readAck}, ts, v, \text{read\_op}, s \rangle$  message in which it encloses its local timestamp and value, its id, and the reader’s operation counter and sends it to the

---

**Algorithm 1** Reader, Writer, and Server Protocols for SWMR algorithm OHSAM

---

```
1: At each reader  $r$ 
2: Variables:
3:  $ts \in \mathbb{N}^+$ ,  $minTS \in \mathbb{N}^+$ ,  $v \in V$ 
4:  $read\_op \in \mathbb{N}^+$ ,  $rAck \subseteq \mathcal{S} \times M$ 
5: Initialization:
6:  $ts \leftarrow 0$ ,  $minTS \leftarrow 0$ ,  $v \leftarrow \perp$ ,  $read\_op \leftarrow 0$ 
7: function READ
8:    $read\_op \leftarrow read\_op + 1$ 
9:    $rAck \leftarrow \emptyset$ 
10:  broadcast ( $\langle readRequest, r, read\_op \rangle$ ) to  $\mathcal{S}$ 
11:  wait until ( $|rAck| = |\mathcal{S}|/2 + 1$ )
12:   $minTS \leftarrow \min\{m.ts' \mid m \in rAck\}$ 
13:   $v \leftarrow \{m.val \mid m \in rAck \wedge m.ts' = minTS\}$ 
14:  return( $v$ )

15: Upon receive  $m$  from  $s$ 
16: if  $m.read\_op = read\_op$  then
17:    $rAck \leftarrow rAck \cup \{(s, m)\}$ 

18: At writer  $w$ 
19: Variables:
20:  $ts \in \mathbb{N}^+$ ,  $v \in V$ ,  $wAck \subseteq \mathcal{S} \times M$ 
21: Initialization:
22:  $ts \leftarrow 0$ ,  $v \leftarrow \perp$ 
23: function WRITE( $val : input$ )
24:    $(ts, v) \leftarrow (ts + 1, val)$ 
25:    $wAck \leftarrow \emptyset$ 
26:   broadcast ( $\langle writeRequest, ts, v, w \rangle$ ) to  $\mathcal{S}$ 
27:   wait until ( $|wAck| = |\mathcal{S}|/2 + 1$ )
28:   return

29: Upon receive  $m$  from  $s$ 
30: if  $m.ts = ts$  then
31:    $wAck \leftarrow wAck \cup \{(s, m)\}$ 

32: At server  $s_i$ 
33: Variables:
34:  $ts \in \mathbb{N}^+$ ,  $v \in V$ 
35:  $operations[1..|\mathcal{R}|+1]$ ,  $relays[1..|\mathcal{R}|+1]$  : array of int
36: Initialization:
37:  $ts \leftarrow 0$ ,  $v \leftarrow \perp$ 
38:  $operations[i] \leftarrow 0$  for  $i \in \mathcal{R}$ ,  $relays[i] \leftarrow 0$  for  $i \in \mathcal{R}$ 

39: Upon receive( $\langle readRequest, r, read\_op \rangle$ )
40:   broadcast( $\langle readRelay, ts, v, r, read\_op, s_i \rangle$ ) to  $\mathcal{S}$ 

41: Upon receive( $\langle writeRequest, ts', v', w \rangle$ )
42: if ( $ts < ts'$ ) then
43:    $(ts, v) \leftarrow (ts', v')$ 
44: send ( $\langle writeAck, ts, v, s_i \rangle$ ) to  $w$ 

45: Upon receive( $\langle readRelay, ts', v', r, read\_op, s_i \rangle$ )
46: if ( $ts < ts'$ ) then
47:    $(ts, v) \leftarrow (ts', v')$ 
48: if ( $operations[r] < read\_op$ ) then
49:    $operations[r] \leftarrow read\_op$ 
50:    $relays[r] \leftarrow 0$ 
51: if ( $operations[r] = read\_op$ ) then
52:    $relays[r] \leftarrow relays[r] + 1$ 
53: if ( $relays[r] = |\mathcal{S}|/2 + 1$ ) then
54:   send ( $\langle readAck, ts, v, read\_op, s_i \rangle$ ) to  $r$ 
```

---

requesting reader  $r$  (L53-54).

(3) Upon receiving message  $\langle writeRequest, ts', v', w \rangle$  server  $s$  compares its local timestamp  $ts$  with the received one,  $ts'$ . If  $ts < ts'$ , then the server sets its local timestamp and value to be equal to those in the received message (L42-43). In any other case, no updates are taking place. Finally, the server always sends an acknowledgement,  $writeAck$ , message to the requesting writer (L44).

### 3.1 Correctness.

To prove correctness of algorithm OHSAM we reason about its *liveness* (termination) and *atomicity* (safety).

**Liveness.** Termination holds with respect to our failure model: up to  $f$  servers may fail, where  $f < |\mathcal{S}|/2$  and each operation waits for messages from some majority of servers. We now give more detail on how each operation satisfies *liveness*.

*Write Operation.* Per algorithm OHSAM, writer  $w$  creates a  $writeRequest$  message and then it

broadcasts it to all servers in exchange E1 (L26). Writer  $w$  then waits for `writeAck` messages from a majority of servers from E2 (L27-27). Since in our failure model up to  $f < \frac{|S|}{2}$  servers may crash, writer  $w$  collects `writeAck` messages from a majority of live servers during E2 and the write operation  $\omega$  terminates.

*Read Operation.* The reader  $r$  begins by broadcasting a `readRequest` message all the servers forming exchange E1. Since  $f < \frac{|S|}{2}$ , then at least a majority of servers receives the `readRequest` message sent in E1. Any server  $s$  that receives this message it then broadcasts a `readRelay` message to all the servers, forming E2, (L39-40), and no server ever discards any incoming `readRelay` messages. Any server, whether it is aware or not of the `readRequest`, always keeps a record of the incoming `readRelay` messages and takes action as if it is aware of the `readRequest`. The only difference between server  $s_i$  that received a `readRequest` message and server  $s_k$  that did not, is that  $s_i$  is able to broadcast a `readRelay` message, and  $s_k$  broadcasts a `readRelay` message when it receives the corresponding `readRequest` message (L39-40). Each non-failed server receives `readRelay` messages from a majority of servers during E2 and sends a `readAck` message to the requesting reader  $r$  at E3 (L51-52). Therefore, reader  $r$  collects `readAck` messages from a majority of servers during E3, and the read operation terminates (L12-14).

Based on the above, it is always the case that acknowledgment messages `readAck` and `writeAck` are collected from at least a majority of servers in any read and write operation, thus ensuring *liveness*.

**Atomicity.** To prove atomicity we order the operations with respect to timestamps written and returned. More precisely, for each execution  $\xi$  of the algorithm there must exist a partial order  $\prec$  on the operations in on the set of completed operations  $\Pi$  that satisfy conditions A1, A2, and A3 as given in Section 2. Let  $ts_\pi$  be the value of the timestamp at the completion of  $\pi$  when  $\pi$  is a write, and the timestamp computed as the *maximum*  $ts$  at the completion of a read operation  $\pi$ . With this, we denote the partial order on operations as follows. For two operations  $\pi_1$  and  $\pi_2$ , when  $\pi_1$  is any operation and  $\pi_2$  is a write, we let  $\pi_1 \prec \pi_2$  if  $ts_{\pi_1} < ts_{\pi_2}$ . For two operations  $\pi_1$  and  $\pi_2$ , when  $\pi_1$  is a write and  $\pi_2$  is a read we let  $\pi_1 \prec \pi_2$  if  $ts_{\pi_1} \leq ts_{\pi_2}$ . The rest of the order is established by transitivity and reads with the same timestamps are not ordered. We now state and prove the following lemmas.

It is easy to see that the  $ts$  variable in each server  $s$  is monotonically increasing. This leads to the following lemma.

**Lemma 1** *In any execution  $\xi$  of OHSAM, the variable  $ts$  maintained by any server  $s$  in the system is non-negative and monotonically increasing.*

**Proof.** When a server  $s$  receives a timestamp  $ts$  then  $s$  updates its local timestamp  $ts_s$  if and only if  $ts > ts_s$  (L42-43 and L46-47). Thus the local timestamp of the server monotonically increases and the lemma follows.  $\square$

Next, we show that if a read operation  $\rho_2$  succeeds read operation  $\rho_1$ , then  $\rho_2$  always returns a value at least as recent as the one returned by  $\rho_1$ .

**Lemma 2** *In any execution  $\xi$  of OHSAM, if  $\rho_1$  and  $\rho_2$  are two read operations such that  $\rho_1$  precedes  $\rho_2$ , i.e.,  $\rho_1 \rightarrow \rho_2$ , and  $\rho_1$  returns the value for timestamp  $ts_1$ , then  $\rho_2$  returns the value for timestamp  $ts_2 \geq ts_1$ .*

**Proof.** Let the two operations  $\rho_1$  and  $\rho_2$  be invoked by processes with identifiers  $r_1$  and  $r_2$  respectively (not necessarily different). Also, let  $RSet_1$  and  $RSet_2$  be the sets of servers that sent a `readAck` message to  $r_1$  and  $r_2$  during  $\rho_1$  and  $\rho_2$ .

Assume by contradiction that read operations  $\rho_1$  and  $\rho_2$  exist such that  $\rho_2$  succeeds  $\rho_1$ , i.e.,  $\rho_1 \rightarrow \rho_2$ , and the operation  $\rho_2$  returns a timestamp  $ts_2$  that is smaller than the  $ts_1$  returned by  $\rho_1$ , i.e.,  $ts_2 < ts_1$ . According to our algorithm,  $\rho_2$  returns a timestamp  $ts_2$  that is smaller than the minimum timestamp received by  $\rho_1$ , i.e.,  $ts_1$ , if  $\rho_2$  obtains  $ts_2$  and  $v$  in the `readAck` message of some server  $s_x \in RSet_2$ , and  $ts_2$  is the minimum timestamp received by  $\rho_2$ .

Let us examine if  $s_x$  replies with  $ts'$  and  $v'$  to  $\rho_1$ , i.e.,  $s_x \in RSet_1$ . By Lemma 1, and since  $\rho_1 \rightarrow \rho_2$ , then it must be the case that  $ts' \leq ts_2$ . According to our assumption  $ts_1 > ts_2$ , and since  $ts_1$  is the smallest timestamp sent to  $\rho_1$  by any server in  $RSet_1$ , then it follows that  $r_1$  does not receive the `readAck` message from  $s_x$ , and hence  $s_x \notin RSet_1$ .

Now let us examine the actions of the server  $s_x$ . From the algorithm, server  $s_x$  collects `readRelay` messages from a majority of servers in  $\mathcal{S}$  before sending a `readAck` message to  $\rho_2$  (L53-54). Let  $RRSet_{s_x}$  denote the set of servers that sent `readRelay` to  $s_x$ . Since, both  $RRSet_{s_x}$  and  $RSet_1$  contain some majority of the servers then it follows that  $RRSet_{s_x} \cap RSet_1 \neq \emptyset$ .

Thus there exists a server  $s_i \in RRSet_{s_x} \cap RSet_1$ , which sent (i) a `readAck` to  $r_1$  for  $\rho_1$ , and (ii) a `readRelay` to  $s_x$  during  $\rho_2$ . Note that  $s_i$  sends a `readRelay` for  $\rho_2$  only after it receives a read request from  $\rho_2$  (L39-40). Since  $\rho_1 \rightarrow \rho_2$ , then it follows that  $s_i$  sent the `readAck` to  $\rho_1$  before sending the `readRelay` to  $s_x$ . By Lemma 1, if  $s_i$  attaches a timestamp  $ts_{s_i}$  in the `readAck` to  $\rho_1$ , then  $s_i$  attaches a timestamp  $ts'_{s_i}$  in the `readRelay` message to  $s_x$ , such that  $ts'_{s_i} \geq ts_{s_i}$ . Since  $ts_1$  is the minimum timestamp received by  $\rho_1$ , then  $ts_{s_i} \geq ts_1$ , and hence  $ts'_{s_i} \geq ts_1$  as well. By Lemma 1, and since  $s_x$  receives the `readRelay` message from  $s_i$  before sending a `readAck` to  $\rho_2$ , it follows that  $s_x$  sends a timestamp  $ts_2 \geq ts'_{s_i}$ . Thus,  $ts_2 \geq ts_1$  and this contradicts our initial assumption.  $\square$

The next lemma shows that any read operation following a write operation receives `readAck` messages from servers where each included timestamp is at least as large as the one returned by the complete write operation.

**Lemma 3** *In any execution  $\xi$  of OHSAM, if a read operation  $\rho$  succeeds a write operation  $\omega$  that writes  $ts$  and  $v$ , i.e.,  $\omega \rightarrow \rho$ , and receives `readAck` messages from a majority of servers  $RSet$ , then each  $s \in RSet$  sends a `readAck` message to  $\rho$  with a timestamp  $ts_s$  s.t.  $ts_s \geq ts$ .*

**Proof.** Let  $WSet$  be the set of servers that send a `writeAck` message in  $\omega$  and let  $RRSet$  be the set of servers that send `readRelay` messages to server  $s$ .

By Lemma 1, if a server  $s$  receives timestamp  $ts$  from process  $p$ , then  $s$  includes timestamp  $ts'$  s.t.  $ts' \geq ts$  in any subsequent message. This, means that every server in  $WSet$ , sends a `writeAck` message to  $\omega$  with a timestamp greater or equal to  $ts$ . Hence, every server  $s_x \in WSet$  has timestamp  $ts_{s_x} \geq ts$ . Let us now examine timestamp  $ts_s$  that server  $s \in RSet$  sends in read operation  $\rho$ .

Before server  $s$  sends a `readAck` message in  $\rho$ , it must receive `readRelay` messages from the majority of servers,  $RRSet$  (L53-54). Since both  $WSet$  and  $RRSet$  contain a majority of servers, then  $WSet \cap RRSet \neq \emptyset$ . By Lemma 1, any server  $s_x \in WSet \cap RRSet$  has a timestamp  $ts_{s_x}$  s.t.  $ts_{s_x} \geq ts$ . Since server  $s_x \in RRSet$  and from the algorithm, server's  $s$  timestamp is always updated to the highest timestamp it receives (L46-47), then when server  $s$  receives the message from  $s_x$ , it updates its timestamp  $ts_s$  s.t.  $ts_s \geq ts_{s_x}$ . Thus, by Lemma 1, each  $s \in RSet$  sends a `readAck`



(L53-54) in  $\rho$  with a timestamp  $ts_s$  s.t.  $ts_s \geq ts_{s_x} \geq ts$ . Therefore,  $ts_s \geq ts$  holds and the lemma follows.  $\square$

Next show that if a read operation succeeds a write operation, then it returns a value at least as recent as the one that was written.

**Lemma 4** *In any execution  $\xi$  of OHSAM, if a read  $\rho$  succeeds a write operation  $\omega$  that writes timestamp  $ts$ , i.e.  $\omega \rightarrow \rho$ , and returns a timestamp  $ts'$ , then  $ts' \geq ts$ .*

**Proof.** Suppose that read operation  $\rho$  receives `readAck` messages from a majority of servers  $RSet$ . By lines 12-14 of the algorithm, it follows that  $\rho$  decides on the minimum timestamp,  $ts' = ts\_min$ , among all the timestamps in the `readAck` messages of the servers in  $RSet$ . From Lemma 3,  $ts\_min \geq ts$  holds, where  $ts$  is the timestamp written by the last complete write operation  $\omega$ , then  $ts' = ts\_min \geq ts$  also holds. Therefore,  $ts' \geq ts$  holds and the lemma follows.  $\square$

**Theorem 5** *Algorithm OHSAM implements an atomic SWMR object.*

**Proof.** We now use the lemmas stated above and the operations order definition to reason about each of the three *atomicity* conditions A1, A2 and A3.

**A1** For any  $\pi_1, \pi_2 \in \Pi$  such that  $\pi_1 \rightarrow \pi_2$ , it cannot be that  $\pi_2 \prec \pi_1$ .

When the two operations  $\pi_1$  and  $\pi_2$  are reads and  $\pi_1 \rightarrow \pi_2$  holds, then from Lemma 2 it follows that the timestamp returned from  $\pi_2$  is always greater or equal to the one returned from  $\pi_1$ ,  $ts_{\pi_2} \geq ts_{\pi_1}$ . If  $ts_{\pi_2} > ts_{\pi_1}$  then by the ordering definition  $\pi_1 \prec \pi_2$  is satisfied. When  $ts_{\pi_2} = ts_{\pi_1}$  then the ordering is not defined, thus it cannot be the case that  $\pi_2 \prec \pi_1$ . If  $\pi_2$  is a write, the sole writer generates a new timestamp by incrementing the largest timestamp in the system. By well-formedness (see Section 2), any timestamp generated by the writer for any write operation that precedes  $\pi_2$  must be smaller than  $ts_{\pi_2}$ . Since  $\pi_1 \rightarrow \pi_2$ , then it holds that  $ts_{\pi_1} < ts_{\pi_2}$ . Hence, by the ordering definition it cannot be the case that  $\pi_2 \prec \pi_1$ . Lastly, when  $\pi_2$  is a read and  $\pi_1$  a write and  $\pi_1 \rightarrow \pi_2$  holds, then from Lemmas 3 and 4 it follows that  $ts_{\pi_2} \geq ts_{\pi_1}$ . By the ordering definition, it cannot hold that  $\pi_2 \prec \pi_1$  in this case either.

**A2** For any write  $\omega \in \Pi$  and any operation  $\pi \in \Pi$ , then either  $\omega \prec \pi$  or  $\pi \prec \omega$ .

If the timestamp returned from  $\omega$  is greater than the one returned from  $\pi$ , i.e.  $ts_\omega > ts_\pi$ , then  $\pi \prec \omega$  follows directly. Similarly, if  $ts_\omega < ts_\pi$  holds, then  $\omega \prec \pi$  follows. If  $ts_\omega = ts_\pi$ , then it must be that  $\pi$  is a read and  $\pi$  discovered  $ts_\omega$  as the *minimum* timestamp in at least a majority of servers. Thus,  $\omega \prec \pi$  follows.

**A3** Every read operation returns the value of the last write preceding it according to  $\prec$  (or the initial value if there is no such write).

Let  $\omega$  be the last write preceding read  $\rho$ . From our definition it follows that  $ts_\rho \geq ts_\omega$ . If  $ts_\rho = ts_\omega$ , then  $\rho$  returned the value written by  $\omega$  on a majority of servers. If  $ts_\rho > ts_\omega$ , then it means that  $\rho$  obtained a larger timestamp. However, the larger timestamp can only be originating from a write that succeeds  $\omega$ , thus  $\omega$  is not the preceding write and this cannot be the case. Lastly, if  $ts_\rho = 0$ , no preceding writes exist, and  $\rho$  returns the initial value.  $\square$

Having shown liveness and atomicity of algorithm OHSAM the result follows.

### 3.2 Performance.

We now assess the performance of OHSAM in terms of (i) latency of read and write operations as measured by the number of communication exchanges, and (ii) the message complexity of read and write operations.

In brief, for algorithm OHSAM write operations take 2 exchanges and read operations take 3 communication exchanges. The (worst case) message complexity of read operations is  $|\mathcal{S}|^2 + 2|\mathcal{S}|$  and the (worst case) message complexity of write operations is  $2|\mathcal{S}|$ . This follows directly from the structure of the algorithm. We now give additional details.

**Operation Latency.** *Write operation latency:* According to algorithm OHSAM, writer  $w$  sends a `writeRequest` message to all the servers in exchange E1, and, awaits for `writeAck` messages from at least a majority of servers in exchange E2. Once the `writeAck` messages are received, no further communication is required and the write operation terminates. Therefore, any write operation consists of 2 communication exchanges.

*Read operation latency:* A reader sends a `readRequest` message to all the servers in the first communication exchange E1. Once a server receives a `readRequest` message, it broadcasts a `readRelay` message to all the servers in exchange E2. Any active servers now await `readRelay` messages from at least a majority of servers, and then, the servers send a `readAck` message to the reader during communication exchange E3. We note that servers do not reply to any incoming `readRelay` messages. Thus, a read operation consists of 3 communication exchanges.

**Message Complexity.** We measure operation message complexity as the worst case number of exchanged messages in each read and write operation. The worst case number of messages corresponds to failure-free executions where all participants send messages to all destinations according to the protocols.

*Write operation:* A single write operation in algorithm OHSAM takes 2 communication exchanges. In the first exchange E1, the writer sends a `writeRequest` message to all the servers in  $\mathcal{S}$ . The second exchange E2, occurs when all servers in  $\mathcal{S}$  send a `writeAck` message to the writer. Thus, at most  $2|\mathcal{S}|$  messages are exchanged in a write operation.

*Read operation:* Read operations take 3 communication exchanges. Exchange E1 occurs when a reader sends a `readRequest` message to all the servers in  $\mathcal{S}$ . Exchange E2 occurs when servers in  $\mathcal{S}$  send a `readRelay` message to all the servers in  $\mathcal{S}$ . The last exchange, E3, occurs when servers in  $\mathcal{S}$  send a `readAck` message to the requesting reader. Therefore,  $|\mathcal{S}|^2 + 2|\mathcal{S}|$  messages are exchanged during a read operation.

## 4 MWMR Algorithm OHMAM

We seek a solution for the MWMR setting that involves *three* communications exchanges per read operation and *four* exchanges per write operation. We now present our MWMR algorithm OHMAM: *One and a half round Multi-writer Atomic Memory*. To impose an ordering on the values written by the writers we associate each value with a tag  $tg$  defined as the pair  $\langle ts, id \rangle$ , where  $ts$  is a timestamp and  $id$  is the identifier of a writer. Tags are ordered lexicographically (cf. [10]). The read protocol is identical to the one presented in section ?? for algorithm OHSAM (except that tags are used instead of timestamps). Thus, for algorithm OHMAM we briefly describe only the protocols that differ, and that is, the writer and server protocols.

---

**Algorithm 2** Reader, Writer and Server Protocols for MWMM algorithm OHMAM
 

---

```

55: At each reader r
56: Components:
57:  $tg \in \langle \mathbb{N}^+, \mathcal{I} \rangle$ ,  $minTAG \in \langle \mathbb{N}^+, \mathbb{N}^+ \rangle$ 
58:  $v \in V$ ,  $read\_op \in \mathbb{N}^+$ ,  $rAck \subseteq \mathcal{S} \times M$ 
59: Initialization:
60:  $tg \leftarrow \langle 0, r \rangle$ ,  $minTAG \leftarrow \langle 0, 0 \rangle$ 
61:  $v \leftarrow \perp$ ,  $read\_op \leftarrow 0$ 
62: function READ
63:    $read\_op \leftarrow read\_op + 1$ .
64:    $rAck \leftarrow \emptyset$ 
65:   broadcast ( $\langle readRequest, r, read\_op \rangle$ ) to  $\mathcal{S}$ 
66:   wait until ( $|rAck| = |\mathcal{S}|/2 + 1$ )
67:    $minTAG \leftarrow \min(\{m.tg' \mid m \in rAck\})$ 
68:    $v \leftarrow \{m.val \mid m \in rAck \wedge m.tg' = minTAG\}$ 
69:   return( $v$ )

70: Upon receive  $m$  from  $s$ 
71: if  $m.read\_op = read\_op$  then
72:    $rAck \leftarrow rAck \cup \{(s, m)\}$ 

73: At each writer w
74: Variables:
75:  $tg \in \langle \mathbb{N}^+, \mathcal{I} \rangle$ ,  $v \in V$ ,  $write\_op \in \mathbb{N}^+$ 
76:  $maxTS \in \mathbb{N}^+$ ,  $mAck \subseteq \mathcal{S} \times M$ 
77: Initialization:
78:  $tg \leftarrow \langle 0, w \rangle$ ,  $v \leftarrow \perp$ ,  $write\_op \leftarrow 0$ 
79:  $maxTS \leftarrow 0$ 
80: function WRITE( $val : input$ )
81:    $write\_op \leftarrow write\_op + 1$ 
82:    $mAck \leftarrow \emptyset$ 
83:   broadcast( $\langle discover, write\_op, w \rangle$ ) to  $\mathcal{S}$ 
84:   wait until ( $|mAck| = |\mathcal{S}|/2 + 1$ )
85:    $maxTS \leftarrow \max\{m.ts' \mid m \in mAck\}$ 
86:    $(tag, v) \leftarrow (\langle maxTS + 1, w \rangle, val)$ 
87:    $write\_op \leftarrow write\_op + 1$ 
88:    $mAck \leftarrow \emptyset$ 
89:   broadcast( $\langle writeRequest, \langle tg, v \rangle, write\_op, w \rangle$ ) to  $\mathcal{S}$ 
90:   wait until ( $|mAck| = |\mathcal{S}|/2 + 1$ )
91:   return

92: Upon receive  $m$  from  $s$ 
93: if  $m.write\_op = write\_op$  then
94:    $mAck \leftarrow mAck \cup \{(s, m)\}$ 

95: At each server  $s_i$ 
96: Variables:
97:  $tg \in \langle \mathbb{N}^+, \mathcal{I} \rangle$ ,  $v \in V$ ,
98:  $write\_ops[1 \dots |\mathcal{W}| + 1] : \text{array of int}$ 
99:  $operations[1 \dots |\mathcal{R}| + 1] : \text{array of int}$ 
100:  $relays[1 \dots |\mathcal{R}| + 1] : \text{array of int}$ 
101: Initialization:
102:  $tg \leftarrow \langle 0, s_i \rangle$ ,  $v \leftarrow \perp$ 
103:  $write\_ops[i] \leftarrow 0$  for  $i \in \mathcal{W}$ 
104:  $operations[i] \leftarrow 0$  for  $i \in \mathcal{R}$ 
105:  $relays[i] \leftarrow 0$  for  $i \in \mathcal{R}$ 

106: Upon receive( $\langle readRequest, r, read\_op \rangle$ )
107: broadcast ( $\langle readRelay, \langle tg, v \rangle, r, read\_op, s_i \rangle$ ) to  $\mathcal{S}$ 

108: Upon receive( $\langle readRelay, \langle tg', v' \rangle, r, read\_op, s_i \rangle$ )
109: if ( $tg < tg'$ ) then
110:    $\langle tg, v \rangle \leftarrow \langle tg', v' \rangle$ .
111: if ( $operations[r] < read\_op$ ) then
112:    $operations[r] \leftarrow read\_op$ .
113:    $relays[r] \leftarrow 0$ .
114: if ( $operations[r] = read\_op$ ) then
115:    $relays[r] \leftarrow relays[r] + 1$ .
116: if ( $relays[r] = |\mathcal{S}|/2 + 1$ ) then
117:   Send ( $\langle readAck, \langle tg, v \rangle, read\_op, s_i \rangle$ ) to  $r$ 

118: Upon receive( $\langle discover, write\_op, w \rangle$ )
119: Send ( $\langle discoverAck, \langle tg, v \rangle, write\_op, s_i \rangle$ ) to  $w$ 

120: Upon receive( $\langle writeRequest, tg', v', write\_op, w \rangle$ )
121: if ( $(tg < tg') \wedge (write\_op[w] < write\_op)$ ) then
122:    $\langle tg, v \rangle \leftarrow \langle tg', v' \rangle$ 
123:    $write\_ops[w] \leftarrow write\_op$ 
124: send ( $\langle writeAck, \langle tg, v \rangle, write\_op, s_i \rangle$ ) to  $w$ 

```

---

**Writer Protocol.** This protocol is similar to [10]. When a write operation is invoked, a writer  $w$  monotonically increases its local write operation counter  $write\_op$ , empties the set  $mAck$  that holds the received acknowledgment messages (L81 - 82), and it broadcasts a **discover** message to all servers  $s \in \mathcal{S}$  (L83). It then waits to collect **discoverAck** messages from a majority of servers,  $|\mathcal{S}|/2 + 1$ . While collecting **discoverAck** messages, writer  $w$  checks the  $write\_op$  variable that is included in the message  $m$  and discards any message where the value of  $write\_op < m.write\_op$  (L92 - 94). This, happens in order to avoid any delayed **discoverAck** messages sent during previous

write operations. Once the `discoverAck` messages are collected, writer  $w$  determines the maximum timestamp  $maxTS$  from the tags included in the received messages (L85) and creates a new local tag  $tg$ , in which it assigns its id and sets the timestamp to one higher than the maximum one,  $tg = \langle maxTS + 1, w \rangle$  (L86). The writer then broadcasts a `writeRequest` message, including the updated tag, the value to be written, its write operation counter and id,  $tg, v, write\_op$  and  $w$ , to all the participating servers (L89). It then waits to collect  $|\mathcal{S}|/2 + 1$  `writeAck` messages (L90) for completion.

**Server Protocol.** Servers react to messages from the readers exactly as in Algorithm OHSAM. Here we describe server actions for `discover` and `writeRequest` messages.

(1) Upon receiving message  $\langle \text{discover}, write\_op, w \rangle$ , server  $s$  attaches its local tag and local value in a new `discoverAck` message that it sends back to the requesting writer  $w$  (L118-119).

(2) Upon receiving  $\langle \text{writeRequest}, \langle tg', v' \rangle, write\_op, w \rangle$  message server compares its local tag  $tg$  with the received tag  $tg'$ . If the message is not stale and server's local tag is older,  $tg < tg'$ , it updates its local timestamp and local value to those received (L121-123). Otherwise, no update takes place. Server  $s$  acknowledges the requesting writer  $w$  by creating and sending it a `writeAck` message (L124).

#### 4.1 Correctness.

To show correctness of Algorithm 2 we prove that it satisfies the *termination* and *atomicity* properties.

**Liveness.** Similarly to OHSAM, termination holds with respect to our failure model: up to  $f$  servers may fail, where  $f < |\mathcal{S}|/2$  and each operation waits for messages from some majority of servers. We now give additional details.

*Write Operation.* Writer  $w$  finds the maximum tag by broadcasting a `discover` message to all servers forming exchange E1, and waiting to collect `discoverAck` replies from a majority of servers during exchange E2 (L83-85 and L118-119). Since we tolerate  $f < \frac{|\mathcal{S}|}{2}$  crashes, then at least a majority of live servers will collect the `discover` messages from E1 and reply to writer  $w$  in E2. Once the maximum timestamp is determined, then writer  $w$  updates its local tag and broadcasts a `writeRequest` message to all the servers forming E3. Writer  $w$  then waits to collect `writeAck` replies from a majority of servers before completion. Again, at least a majority of servers collects the `writeRequest` message during E3, and acknowledges to the requesting writer  $w$  in E4.

*Read Operation.* A read operation of algorithm OHMAM differs from OHSAM only by using tags instead of timestamps in order to impose an ordering on the values written. The structure of the read protocol is identical to OHSAM, thus *liveness* is ensured as reasoned in section 3.1.

Based on the above, any read or write operation collects a sufficient number of messages to terminate, guaranteeing *liveness*.

**Atomicity.** MWMR setting we use tags instead of timestamps, and here we show how algorithm OHMAM (algorithm 2) satisfies *atomicity* using *tags*. We now state and prove the following lemmas.

It is easy to see that the  $tg$  variable in each server  $s$  is monotonically increasing. This leads to the following lemma.

**Lemma 6** *In any execution  $\xi$  of OHMAM, the variable  $tg$  maintained by any server  $s$  in the system is non-negative and monotonically increasing.*

Proof When server  $s$  receives a tag  $tg$  then  $s$  updates its local tag  $tg_s$  iff  $tg > tg_s$  (L109-110 and L121-123). Thus the local tag of the server monotonically increases and the lemma follows.

Next, we show that if a read operation  $\rho_2$  succeeds read operation  $\rho_1$ , then  $\rho_2$  always returns a value at least as recent as the one returned by  $\rho_1$ .

**Lemma 7** *In any execution  $\xi$  of OHMAM, If  $\rho_1$  and  $\rho_2$  are two read operations such that  $\rho_1$  precedes  $\rho_2$ , i.e.,  $\rho_1 \rightarrow \rho_2$ , and  $\rho_1$  returns a tag  $tg_1$ , then  $\rho_2$  returns a tag  $tg_2 \geq tg_1$ .*

Proof Let the operations  $\rho_1$  and  $\rho_2$  be invoked by processes  $r_1$  and  $r_2$  respectively (not necessarily different). Let  $RSet_1$  and  $RSet_2$  be the sets of servers that reply to  $r_1$  and  $r_2$  during  $\rho_1$  and  $\rho_2$  respectively.

Suppose, for purposes of contradiction, that read operations  $\rho_1$  and  $\rho_2$  exist such that  $\rho_2$  succeeds  $\rho_1$ , i.e.,  $\rho_1 \rightarrow \rho_2$ , and the operation  $\rho_2$  returns a tag  $tg_2$  which is smaller than the  $tg_1$  returned by  $\rho_1$ , i.e.,  $tg_2 < tg_1$ .

According to our algorithm,  $\rho_2$  returns a tag  $tg_2$  which is smaller than the minimum tag received by  $\rho_1$ , i.e.,  $tg_1$ , if  $\rho_2$  discovers tag  $tg_2$  and value  $v$  in the `readAck` message of some server  $s_x \in RSet_2$ , and  $tg_2$  is the minimum tag received by  $\rho_2$ .

Assume that server  $s_x$  replies with tag  $tg'$  and value  $v'$  to read operation  $\rho_1$ , i.e.,  $s_x \in RSet_1$ . By monotonicity of the timestamp at the servers (Lemma 6), and since  $\rho_1 \rightarrow \rho_2$ , then it must be the case that  $tg' \leq tg_2$ . According to our assumption  $tg_1 > tg_2$ , and since  $tg_1$  is the smallest tag sent to  $\rho_1$  by any server in  $RSet_1$ , then it follows that  $r_1$  does not receive the `readAck` message from  $s_x$ , and hence  $s_x \notin RSet_1$ .

Now examine the actions of server  $s_x$ . From the algorithm, server  $s_x$  collects `readRelay` messages from a majority of servers in  $\mathcal{S}$  before sending `readAck` message to  $\rho_2$  (L116-117). Let  $RRSet_{s_x}$  denote the set of servers that send `readRelay` to  $s_x$ . Since, both  $RRSet_{s_x}$  and  $RSet_1$  contain some majority of the servers then it follows that  $RRSet_{s_x} \cap RSet_1 \neq \emptyset$ .

This above means that there exists a server  $s_i \in RRSet_{s_x} \cap RSet_1$  that sends (i) a `readAck` message to  $r_1$  for  $\rho_1$ , and (ii) a `readRelay` message to  $s_x$  during  $\rho_2$ . Note that  $s_i$  sends a `readRelay` message for  $\rho_2$  only after it receives a read request from  $\rho_2$  (L106-107). Since  $\rho_1 \rightarrow \rho_2$ , then it follows that  $s_i$  sends the `readAck` message to  $\rho_1$  before sending the `readRelay` message to  $s_x$ . Thus, by Lemma 6, if  $s_i$  attaches a tag  $tg_{s_i}$  in the `readAck` to  $\rho_1$ , then  $s_i$  attaches a tag  $tg'_{s_i}$  in the `readRelay` message to  $s_x$ , such that  $tg'_{s_i} \geq tg_{s_i}$ . Since  $tg_1$  is the minimum tag received by  $\rho_1$ , then  $tg_{s_i} \geq tg_1$ , then  $tg'_{s_i} \geq tg_1$  as well. By Lemma 6, and since  $s_x$  receives the `readRelay` message from  $s_i$  before sending a `readAck` to  $\rho_2$ , it follows that  $s_x$  sends a tag  $tg_2 \geq tg'_{s_i}$ . Therefore,  $tg_2 \geq tg_1$  and this contradicts our initial assumption and completes our proof.

Next, we reason that if a write operation  $\omega_2$  succeeds write operation  $\omega_1$ , then  $\omega_2$  writes a value associated with a tag strictly higher than  $\omega_1$ .

**Lemma 8** *In any execution  $\xi$  of OHMAM, if a write operation  $\omega_1$  writes a value with tag  $tg_1$  then for any succeeding write operation  $\omega_2$  that writes a value with tag  $tg_2$  we have  $tg_2 > tg_1$ .*

Proof Let  $WSet_1$  be the set of servers that send a `writeAck` message within write operation  $\omega_1$ . Let  $Disc_2$  be the set of servers that send a `discoverAck` message within write operation  $\omega_2$ .

Based on the assumption, write operation  $\omega_1$  is complete. By Lemma 6, we know that if a server  $s$  receives a tag  $tg$  from a process  $p$ , then  $s$  includes tag  $tg'$  s.t.  $tg' \geq tg$  in any subsequently message. Thus the servers in  $WSet_1$  send a `writeAck` message within  $\omega_1$  with tag at least tag  $tg_1$ . Hence, every server  $s_x \in WSet$  obtains tag  $tg_{s_x} \geq tg_1$ .

When write operation  $\omega_2$  is invoked, it obtains the maximum tag,  $max\_tag$ , from the tags stored in at least a majority of servers. This is achieved by sending **discover** messages to all servers and collecting **discoverAck** replies from a majority of servers forming set  $Disc_2$  (L83-85 and L118-119).

Sets  $WSet_1$  and  $Disc_2$  contain a majority of servers, and so  $WSet_1 \cap Disc_2 \neq \emptyset$ . Thus, by Lemma 6, any server  $s_k \in WSet \cap Disc_2$  has a tag  $tg_{s_k}$  s.t.  $tg_{s_k} \geq tg_{s_x} \geq tg_1$ . Furthermore, the invoker of  $\omega_2$  discovers a  $max\_tag$  s.t.  $max\_tag \geq tg_{s_k} \geq tg_{s_x} \geq tg_1$ . The invoker updates its local tag by increasing the maximum tag it discovered, i.e.  $tg_2 = \langle max\_tag + 1, v \rangle$  (L86), and associating  $tg_2$  with the value to be written. We know that,  $tg_2 > max\_tag \geq tg_1$ , hence  $local\_tag > tg_1$ .

Now the invoker of  $\omega_2$  includes its tag  $local\_tag$  with **writeRequest** message to all servers, and terminates upon receiving **writeAck** messages from a majority of servers. By Lemma 6,  $\omega_2$  receives **writeAck** messages with a tag  $tg_2$  s.t.  $tg_2 \geq local\_tag > tg_1$  hence  $tg_2 > tg_1$ , and the lemma follows.

At this point we have to show that any read operation which succeeds a write operation, will receive **readAck** messages from the servers where each included timestamp will be greater or equal to the one that the complete write operation returned.

**Lemma 9** *In any execution  $\xi$  of OHMAM, if a read operation  $\rho$  succeeds a write operation  $\omega$  that writes value  $v$  with tag  $tg$ , i.e.,  $\omega \rightarrow \rho$ , and receives **readAck** messages from a majority of servers  $RSet$ , then each  $s \in RSet$  sends a **readAck** message to  $\rho$  with a tag  $tg_s$  s.t.  $tg_s \geq tg$ .*

*Proof* Let  $WSet$  be the set of servers that send a **writeAck** message to the write operation  $\omega$  and let  $RRSet$  be the set of servers that sent **readRelay** messages to server  $s$ .

It is given that write operation  $\omega$  is complete. By Lemma 6, we know that if server  $s$  receives a tag  $tg$  from process  $p$ , then  $s$  includes a tag  $tg'$  s.t.  $tg' \geq tg$  in any subsequent message. Thus a majority of servers, forming  $WSet$ , send a **writeAck** message in  $\omega$  with tag greater or equal to tag  $tg$ . Hence, every server  $s_x \in WSet$  has a tag  $tg_{s_x} \geq tg$ . Let us now examine tag  $tg_s$  that server  $s$  sends to read operation  $\rho$ .

Before server  $s$  sends a **readAck** message to  $\rho$ , it must receive **readRelay** messages for the majority of servers,  $RRSet$  (L116-117). Since both  $WSet$  and  $RRSet$  contain a majority of servers, then it follows that  $WSet \cap RRSet \neq \emptyset$ . Thus, by Lemma 6, any server  $s_x \in WSet \cap RRSet$  has a tag  $tg_x$  s.t.  $tg_x \geq tg$ .

Since server  $s_x \in RRSet$  and by the algorithm, server's  $s$  tag is always updated to the highest tag it observes (L109-110), then when server  $s$  receives the message from  $s_x$ , it updates its tag  $tg_s$  s.t.  $tg_s \geq tg_x$ .

Furthermore, server  $s$  creates a **readAck** message where it includes its local tag  $tg_s$  and its local value  $v_s$ , and sends this **readAck** message within the read operation  $\rho$  (L116-117). Each  $s \in RSet$  sends a **readAck** to  $\rho$  with a tag  $tg_s$  s.t.  $tg_s \geq tg_x \geq tg$ . Therefore,  $tg_s \geq tg$  and the lemma follows.

Next we show that if a read operation succeeds a write operation, then it returns a value at least as recent as the one written.

**Lemma 10** *In any execution  $\xi$  of OHMAM, if read operation  $\rho$  succeeds write operation  $\omega$ , i.e.,  $\omega \rightarrow \rho$ , that writes value  $v$  associated with tag  $tg$ , and returns tag  $tg'$ , then  $tg' \geq tg$ .*

*Proof* Suppose that read operation  $\rho$  receives **readAck** messages from a majority of servers  $RSet$  and decides on a tag  $tg'$  associated with value  $v$  and terminates.

In this case, by Algorithm (L67-69) it follows that read operation  $\rho$  decides on a tag  $tg'$  that belongs to a **readAck** message among the messages from servers in  $RSet$ ; and it is the minimum tag among all the tags that are included in messages of servers  $RSet$ , hence  $tg' = min\_tag$ .

Furthermore, since  $tg' = \text{min\_tag}$  holds and from Lemma 9,  $\text{min\_tag} \geq tg$  holds, where  $tg$  is the tag returned from the last complete write operation  $\omega$ , then  $tg' = \text{min\_tag} \geq tg$  also holds. Therefore,  $tg' \geq tg$  holds and the lemma follows.

**Lemma 11** *In any execution  $\xi$  of OHMAM, if a write  $\omega$  succeeds a read operation  $\rho$  that reads tag  $tg$ , i.e.  $\rho \rightarrow \omega$ , and returns a tag  $tg'$ , then  $tg' > tg$ .*

*Proof* Let  $RR$  be the set of servers that sent **readRelay** messages to  $\rho$ . Let  $dAck$  be the set of servers that sent **discoverAck** messages to  $\omega$ . Let  $wAck$  be the set of servers that sent **writeAck** messages to  $\omega$  and let  $RA$  be the set of servers that sent **readAck** messages to  $\rho$ . It is not necessary that  $a \neq b \neq c$  holds.

Based on the read protocol, the read operation  $\rho$  terminates when it receives **readAck** messages from a majority of servers. It follows that  $\rho$  decides on the minimum tag,  $tg = \text{minTG}$ , among all the tags in the **readAck** messages of the set  $RA$  and terminates. Writer  $\omega$ , initially it broadcasts a **discover** message to all servers, and it then awaits for “fresh” **discoverAck** messages from a majority of servers, that is, set  $dAck$ . Each of  $RA$  and  $dAck$  sets are from majorities of servers, and so  $RA \cap dAck \neq \emptyset$ . By Lemma 6, any server  $s_k \in RA \cap dAck$  has a tag  $tg_{s_k}$  s.t.  $tg_{s_k} \geq tg$ . Since  $\omega$  generates a new local tag-value  $(tg', v)$  pair in which it assigns the timestamp to be one higher than the one discovered in the *maximum* tag from set  $dAck$ , it follows that  $tg' > tg$ . Furthermore,  $\omega$  broadcasts the value to be written associated with  $tg'$  in a **writeRequest** message to all servers and it awaits for **writeAck** messages from a majority of servers before completion, set  $wAck$ . Observe that each of  $dAck$  and  $wAck$  sets are from majority of servers, and so  $dAck \cap wAck \neq \emptyset$ . By Lemma 6, any server  $s_k \in dAck \cap wAck$  has a tag  $tg_{s_k}$  s.t.  $tg_{s_k} \geq tg' > tg$  and the result follows.

**Theorem 12** *Algorithm OHMAM implements an atomic MWMMR object.*

*Proof* We use the above lemmas and the operations order definition to reason about each of the three *atomicity* conditions A1, A2 and A3.

**A1** For any  $\pi_1, \pi_2 \in \Pi$  such that  $\pi_1 \rightarrow \pi_2$ , it cannot be that  $\pi_2 \prec \pi_1$ .

If both  $\pi_1$  and  $\pi_2$  are writes and  $\pi_1 \rightarrow \pi_2$  holds, then from Lemma 8 it follows that  $tg_{\pi_2} > tg_{\pi_1}$ . By the ordering definition  $\pi_1 \prec \pi_2$  is satisfied. When  $\pi_1$  is a write,  $\pi_2$  a read and  $\pi_1 \rightarrow \pi_2$  holds, then from Lemmas 9 and 10 it follows that  $tg_{\pi_2} \geq tg_{\pi_1}$ . By definition  $\pi_1 \prec \pi_2$  is satisfied. If  $\pi_1$  is a read,  $\pi_2$  a write and  $\pi_1 \rightarrow \pi_2$  holds, then from Lemma 11 it follows that  $\pi_2$  always returns a tag  $tg_{\pi_2}$  s.t.  $tg_{\pi_2} > tg_{\pi_1}$ . By the ordering definition  $\pi_1 \prec \pi_2$  is satisfied. If both  $\pi_1$  and  $\pi_2$  are reads and  $\pi_1 \rightarrow \pi_2$  holds, then from Lemma 7 it follows that the tag returned from  $\pi_2$  is always greater or equal to the one returned from  $\pi_1$ .  $tg_{\pi_2} \geq tg_{\pi_1}$ . If  $tg_{\pi_2} > tg_{\pi_1}$ , then by the ordering definition  $\pi_1 \prec \pi_2$  holds. When  $tg_{\pi_2} = tg_{\pi_1}$  then the ordering is not defined but it cannot be that  $\pi_2 \prec \pi_1$ .

**A2** For any write  $\omega \in \Pi$  and any operation  $\pi \in \Pi$ , then either  $\omega \prec \pi$  or  $\pi \prec \omega$ .

If  $tg_\omega > tg_\pi$ , then  $\pi \prec \omega$  follows directly. Similarly, if  $tg_\omega < tg_\pi$  holds, then it follows that  $\omega \prec \pi$ . When  $ts_\omega = ts_\pi$  holds, then the uniqueness of each tag that a writer creates ensures that  $\pi$  is a read. In particular, remember that each tag is a  $\langle ts, id \rangle$  pair, where  $ts$  is a natural number and  $id$  a writer identifier. Tags are ordered lexicographically, first with respect to the timestamp and then with respect to the writer id. Since the writer ids are unique, then even if two writes use the same timestamp  $ts$  in the tag pairs they generate, the two tags cannot be equal as they will differ on the writer id. Furthermore, if the two tags are generated by the same writer, then by well-formedness it must be the case that the latter write will contain a timestamp larger than any timestamp used

by that writer before. Since  $\pi$  is a read operation that receives `readAck` messages from a majority of servers, then the intersection properties of majorities ensure that  $\omega \prec \pi$ .

**A3** Every read operation returns the value of the last write preceding it according to  $\prec$  (or the initial value if there is no such write).

Let  $\omega$  be the last write preceding read  $\rho$ . From our definition it follows that  $tg_\rho \geq tg_\omega$ . If  $tg_\rho = tg_\omega$ , then  $\rho$  returned a value written by  $\omega$  in some servers. Read  $\rho$  waited for `readAck` messages from a majority of servers and the intersection properties of majorities ensure that  $\omega$  was the last complete write operation. If  $tg_\rho > tg_\omega$  holds, it must be the case that there is a write  $\omega'$  that wrote  $tg_\rho$  and by definition it must hold that  $\omega \prec \omega' \prec \rho$ . Thus,  $\omega$  is not the preceding write and this cannot be the case. Lastly, if  $tg_\rho = 0$ , no preceding writes exist, and  $\rho$  returns the initial value.

Having shown liveness and atomicity of algorithm OHMAM the result follows.

## 4.2 Performance.

Briefly, for algorithm OHMAM write operations take 4 communication exchanges and read operations take 3 exchanges. The (worst case) message complexity of read operations is  $|\mathcal{S}|^2 + 2|\mathcal{S}|$  and the (worst case) message complexity of write operations is  $4|\mathcal{S}|$ . This follows directly from the structure of the algorithm. We now give additional details.

**Operation Latency.** *Write operation latency:* According to algorithm OHMAM, writer  $w$  broadcasts a `discover` message to all the servers during exchange E1, and, awaits for `discoverAck` messages from a majority of servers during E2. Once the `discoverAck` messages are received, then writer  $w$  broadcasts a `writeRequest` message to all servers in exchange E3. Lastly, it waits for `writeAck` messages from a majority of servers in E4. No further communication is required and the write operation terminates. Thus any write operation consists of 4 communication exchanges.

*Read operation latency:* The structure of the read protocol of OHMAM is identical to OHSAM, thus a read operation consists of 3 communication exchanges as reasoned in Section 3.2.

**Message Complexity.** Similarly as in Section 3.2, we measure operation message complexity as the worst case number of exchanged messages in each read and write operation.

*Write operation:* A write operation in algorithm OHMAM takes 4 communication exchanges. The first and the third exchanges, E1 and E3, occur when a writer sends `discover` and `writeRequest` messages respectively to all servers in  $\mathcal{S}$ . The second and fourth exchanges, E2 and E4, occur when servers in  $\mathcal{S}$  send `discoverAck` and `writeAck` messages back to the writer. Thus, in a write operation there are  $4|\mathcal{S}|$  messages exchanged.

*Read operation:* The structure of the read protocol of OHMAM is identical to OHSAM thus, as reasoned in Section 3.2, during a read operation,  $|\mathcal{S}|^2 + 2|\mathcal{S}|$  messages are exchanged.

## 5 Reducing the Latency of Read Operations

In this section we revise the protocol implementing read operations of algorithms OHSAM and OHMAM to yield protocols that implement read operations that terminate in either *two* or *three* communication exchanges. The key idea here is to let the reader determine “quickly” that a majority of servers holds the same timestamp (or tag) and return its associated value. This is achieved by having the servers send relay messages to each other as well as to the requesting reader. While the reader collects the relays and the read acknowledgment messages, if it observes in the set of the received relay messages that a majority of servers holds the same timestamp (or tag),



then it safely returns the associated value and the read operation terminates in *two* communication exchanges. If that was not the case, then the reader proceeds similarly to algorithm OHSAM and terminates in *three* communication exchanges. We name the revised algorithms as OHSAM' and OHMAM'.

### 5.1 Algorithm OHSAM' for the SWMR setting

The code for OHSAM' that presents the revised read protocol is given in Algorithm 3. In addition, for the servers protocol we show only the changes from algorithm OHSAM. We now give additional details.

---

#### Algorithm 3 Read Protocol Changes for SWMR algorithm OHSAM'

---

<pre> 119: <i>At each reader r</i> 120: <b>Variables:</b> 121: <math>ts \in \mathbb{N}^+, minTS \in \mathbb{N}^+, v \in V</math> 122: <math>read\_op \in \mathbb{N}^+, rRelay, rAck \subseteq \mathcal{S} \times M</math> 123: <b>Initialization:</b> 124: <math>ts \leftarrow 0, minTS \leftarrow 0, v \leftarrow \perp, read\_op \leftarrow 0</math> 125: <b>function</b> READ 126:   <math>read\_op \leftarrow read\_op + 1.</math> 127:   <math>rAck, rRelay \leftarrow (\emptyset, \emptyset)</math> 128:   <b>bdcst</b> (readRequest, <math>r, read\_op</math>) to <math>\mathcal{S}</math> 129:   <b>wait until</b> (<math> rAck  =  \mathcal{S} /2 + 1</math>) <b>OR</b> 130:     (<math>\exists Z \subseteq rRelay : ( Z  \geq  \mathcal{S} /2 + 1) \wedge</math> 131:       (<math>\forall (m', s'), (m'', s'') \in Z : m'.ts = m''.ts</math>)) 132:   <b>if</b> (<math>rAck =  \mathcal{S} /2 + 1</math>) <b>then</b> 133:     <math>minTS \leftarrow \min\{m.ts' \mid m \in rAck\}</math> 134:     <math>v \leftarrow \{m.val \mid m \in rAck \wedge m.ts' = minTS\}</math> 135:     <b>return</b>(<math>v</math>) 136:   <b>else</b> 137:     <math>v \leftarrow \{m.val \mid m \in rRelay\}</math> 138:     <b>return</b>(<math>v</math>) </pre>	<pre> 139: <b>Upon receive</b> <math>m</math> from <math>s</math> 140: <b>if</b> <math>m.read\_op = read\_op</math> <b>then</b> 141:   <b>if</b> <math>m.type = readAck</math> <b>then</b> 142:     <math>rAck \leftarrow rAck \cup \{(s, m)\}</math> 143:   <b>else</b> 144:     <math>rRelay \leftarrow rRelay \cup \{(s, m)\}</math>  145: <i>At each server <math>s_i</math></i> 146: <b>Upon receive</b>(readRequest, <math>r, read\_op</math>) 147:   <b>bdcst</b>(readRelay, <math>ts, v, r, read\_op, s_i</math>) to <math>\mathcal{S} \cup \{r\}</math> </pre>
--	---

---

In order to construct algorithm OHSAM' we modify readers and servers protocol of algorithm OHSAM. In particular, we update the read protocol to wait for both readRelay and readAck messages (L129-131). Next, in order for the servers protocol to support the broadcast of a readRelay message to all the servers and the reader process we replace lines 39-40 of OHSAM with lines 146-147 of OHSAM', as shown in algorithm 3. The combination of those changes yields algorithm OHSAM'.

**Revised Protocol for the Server.** The server sends a readRelay message to all the servers *and* to the requesting reader (L147).

**Revised Protocol for the Reader.** Here we let the reader await for either (i) readAck messages or (ii) readRelay messages from a majority of servers (L129-131). Notice that in latter, all the  $|\mathcal{S}|/2 + 1$  readRelay messages must include the same timestamp to ensure that at least a majority of servers is informed regarding the last complete write operation (L130-131). In addition, since at least a majority of servers is informed regarding the last complete operation, then it is safe for the reader to return the value  $v$  associated with the timestamp  $ts$  found in the readRelay messages collected from the majority of servers. Otherwise, when case (i) holds, then the read protocol proceeds as in OHSAM.

### 5.1.1 Correctness.

*Liveness* and *atomicity* of the revised algorithm OHSAM' is shown similarly to algorithm OHSAM as reasoned in Section 3.1.

**Liveness.** Termination of Algorithm OHSAM' is guaranteed with respect to our failure model: up to  $f$  servers may fail, where  $f < |\mathcal{S}|/2$ , and any type of operation awaits for messages from a majority of servers before completion. We now provide additional details.

*Read Operation.* A read operation of algorithm OHSAM' terminates when the client either (i) collects `readRelay` messages from at least a majority of serves and all of them include the same timestamp; or (ii) collects `readAck` messages from a majority of servers. When case (i) occurs then the operation terminates immediately (and faster). Otherwise, case (ii) holds, the read operation proceeds identically to algorithm OHSAM and its termination is ensured as reasoned in Section 3.1.

*Write Operation.* A write operation of algorithm OHSAM' is identical to that of algorithm OHSAM, thus *liveness* is guaranteed as reasoned in Section 3.1.

**Atomicity.** Next, we show how algorithm OHSAM' satisfies atomicity. Due to the similarity of the writer and server protocols of algorithm OHSAM' to those in OHSAM, we state the lemmas and we omit some of the proofs. Lemma 13 shows that the timestamp variable  $ts$  maintained by each server  $s$  in the system is monotonically non-decreasing.

**Lemma 13** *In any execution  $\xi$  of OHSAM', the variable  $ts$  maintained by any server  $s$  in the system is non-negative and monotonically increasing.*

**Proof.** Lemma 1 for algorithm OHSAM also applies to algorithm OHSAM' because the modification in line 147 does not affect the update of the timestamp  $ts$  at the server protocol.  $\square$

Next, we show that if a read operation  $\rho_2$  succeeds read operation  $\rho_1$ , then  $\rho_2$  always returns a value at least as recent as the one returned by  $\rho_1$ .

**Lemma 14** *In any execution  $\xi$  of OHSAM', if  $\rho_1$  and  $\rho_2$  are two read operations such that  $\rho_1$  precedes  $\rho_2$ , i.e.,  $\rho_1 \rightarrow \rho_2$ , and  $\rho_1$  returns the value for timestamp  $ts_1$ , then  $\rho_2$  returns the value for timestamp  $ts_2 \geq ts_1$ .*

**Proof.** Since read operations in algorithm OHSAM terminate in 3 communication exchanges then from Lemma 2 we know that any two non-concurrent 3-exchange read operations satisfy this. Thus we have to show that the lemma holds for the cases where (i) a 2-exchange read operation  $\rho_1$  precedes a 2-exchange read operation  $\rho_2$ ; (ii) a 2-exchange read operation  $\rho_1$  precedes a 3-exchange read operation  $\rho_2$ ; and (iii) a 3-exchange read operation  $\rho_1$  precedes a 2-exchange read operation  $\rho_2$ . Let the two operations  $\rho_1$  and  $\rho_2$  be invoked by processes with identifiers  $r_1$  and  $r_2$  respectively (not necessarily different).

*Case (i).* Let  $RRSet_1$  and  $RRSet_2$  be the sets of servers that send a `readRelay` message to  $r_1$  and  $r_2$  during  $\rho_1$  and  $\rho_2$ . Assume by contradiction that 2-exchange read operations  $\rho_1$  and  $\rho_2$  exist such that  $\rho_2$  succeeds  $\rho_1$ , i.e.,  $\rho_1 \rightarrow \rho_2$ , and the operation  $\rho_2$  returns a timestamp  $ts_2$  that is smaller than the  $ts_1$  returned by  $\rho_1$ , i.e.,  $ts_2 < ts_1$ . According to our algorithm,  $\rho_2$  returns a timestamp  $ts_2$  that is smaller than the timestamp that  $\rho_1$  returned i.e.,  $ts_1$ , if  $\rho_2$  received  $|\mathcal{S}|/2 + 1$  `readRelay` messages that all included the same timestamp  $ts_2$  and  $ts_2$  is smaller than  $ts_1$ , which is included in  $|\mathcal{S}|/2 + 1$

readRelay messages received by  $\rho_1$ . Since, both  $RRSet_1$  and  $RRSet_2$  contain some majority of servers then it follows that  $RRSet_1 \cap RRSet_2 \neq \emptyset$ . And since by Lemma 13 the timestamp variable  $ts$  maintained by servers is monotonically increasing, then it is impossible that  $\rho_2$  received  $|\mathcal{S}|/2 + 1$  readRelay messages that all included the same timestamp  $ts_2$  and  $ts_2 < ts_1$ . In particular, since at least a majority of servers have a timestamp at least as  $ts_1$  then  $\rho_2$  can receive only  $|\mathcal{S}|/2$  readRelay messages with a timestamp  $ts_2$  s.t.  $ts_2 < ts_1$ . Therefore, this contradicts our assumption.

*Case (ii).* Since  $\rho_1$  is a 2-exchange operation, then  $r_1$  receives at least  $|\mathcal{S}|/2 + 1$  readRelay messages that includes the same timestamp  $ts_1$ . Thus after the completion of  $\rho_1$  at least a majority of servers hold a timestamp at least as  $ts_1$ . In addition, we know that the servers relay to each other and wait for readRelay messages from a majority of servers before they send a readAck message to the reader  $r_2$ . By Lemma 13 the timestamp variable  $ts$  maintained by servers is monotonically increasing then each server  $s_i$  that sends a readAck message to  $r_2$  must include a timestamp  $ts_{s_i}$  s.t.  $ts_{s_i} \geq ts_1$ . Therefore, the minimum timestamp  $ts_2$  that  $r_2$  can observe in each readAck message received from  $s_i$  must be  $ts_2 \geq ts_{s_i} \geq ts_1$ . Since a 3-exchange read operation decides on the minimum timestamp observed in the readAck responses, then reader  $r_2$  will decide on a timestamp  $ts_2$  s.t.  $ts_2 \geq ts_1$ .

*Case (iii).* Since  $\rho_1$  is a 3-exchange operation, then  $r_1$  receives at least  $|\mathcal{S}|/2 + 1$  readAck messages that include the minimum timestamp  $ts_1$ . Servers relay to each other before they send a readAck message to  $\rho_1$  and timestamps in servers are monotone (Lemma 13), thus after the completion of  $\rho_1$  at least a majority of servers,  $s_i \in RSet$ , hold a timestamp no smaller than  $ts_1$ . Let  $RRSet$  be the set of servers that send a readRelay message to  $r_2$  during  $\rho_2$ . In order for  $\rho_2$  to terminate, based on the read protocol the size of  $RRSet$  must be at least  $|\mathcal{S}|/2 + 1$ . Let  $ts_{s_i}$  be a timestamp received from a server  $s_i \in RRSet$ . Since  $RSet \cap RRSet \neq \emptyset$  then the 2-exchange operation  $\rho_2$  that succeeds  $\rho_1$  can receive at most  $|\mathcal{S}|/2$  (minority) readRelay messages with a timestamp  $ts_{s_i}$  s.t.  $ts_{s_i} < ts_1$ . Thus, when  $\rho_2$  terminates it must return a timestamp  $ts_2$  s.t.  $ts_2 \geq ts_1$  and the lemma follows.  $\square$

Now we show that if a read operation succeeds a write operation, then it returns a value at least as recent as the one written.

**Lemma 15** *In any execution  $\xi$  of the algorithm, if a read  $\rho$  succeeds a write operation  $\omega$  that writes timestamp  $ts$ , i.e.  $\omega \rightarrow \rho$ , and returns a timestamp  $ts'$ , then  $ts' \geq ts$ .*

**Proof.** From Lemma 4 for algorithm OHSAM we know that lemma holds if a 3-exchange read operation succeeds a write operation. We now show that the same holds in case where the read operation terminates in 2 exchanges.

Assume by contradiction that a 2-exchange read operation  $\rho$  and a write operation  $\omega$  exist such that  $\rho$  succeeds  $\omega$ , i.e.  $\omega \rightarrow \rho$ , and  $\rho$  returns a timestamp  $ts'$  that is smaller than  $ts$  that  $\omega$  wrote,  $ts' < ts$ . From our algorithm, in order for this to happen,  $\rho$  receives  $|\mathcal{S}|/2 + 1$  readRelay messages that all include the same timestamp  $ts'$  and  $ts' < ts$ . Since  $\omega$  is complete it means that at least a majority of servers hold a timestamp  $ts_s$  s.t.  $ts_s \geq ts$ . Since any two majorities have a non empty intersection, this contradicts the assumption that  $\rho$  received  $|\mathcal{S}|/2 + 1$  readRelay messages that all included the same timestamp  $ts'$  where  $ts' < ts$  and the lemma follows.  $\square$

**Theorem 16** *Algorithm OHSAM' implements an atomic SWMR object.*

**Proof.** We now use the lemmas stated above and the operations order definition to reason about each of the three *atomicity* conditions A1, A2 and A3.

**A1** For any  $\pi_1, \pi_2 \in \Pi$  such that  $\pi_1 \rightarrow \pi_2$ , it cannot be that  $\pi_2 \prec \pi_1$ .

When the two operations  $\pi_1$  and  $\pi_2$  are reads and  $\pi_1 \rightarrow \pi_2$  holds, then from Lemma 14 it follows that the timestamp returned from  $\pi_2$  is always greater or equal to the one returned from  $\pi_1$ ,  $ts_{\pi_2} \geq ts_{\pi_1}$ . If  $ts_{\pi_2} > ts_{\pi_1}$  then by the ordering definition  $\pi_1 \prec \pi_2$  is satisfied. When  $ts_{\pi_2} = ts_{\pi_1}$  then the ordering is not defined, thus it cannot be the case that  $\pi_2 \prec \pi_1$ . If  $\pi_2$  is a write, the sole writer generates a new timestamp by incrementing the largest timestamp in the system. By well-formedness (see Section 2), any timestamp generated by the writer for any write operation that precedes  $\pi_2$  must be smaller than  $ts_{\pi_2}$ . Since  $\pi_1 \rightarrow \pi_2$ , then it holds that  $ts_{\pi_1} < ts_{\pi_2}$ . Hence, by the ordering definition it cannot be the case that  $\pi_2 \prec \pi_1$ . Lastly, when  $\pi_2$  is a read and  $\pi_1$  a write and  $\pi_1 \rightarrow \pi_2$  holds, then from Lemma 15 it follows that  $ts_{\pi_2} \geq ts_{\pi_1}$ . By the ordering definition, it cannot hold that  $\pi_2 \prec \pi_1$  in this case either.

**A2** For any write  $\omega \in \Pi$  and any operation  $\pi \in \Pi$ , then either  $\omega \prec \pi$  or  $\pi \prec \omega$ .

If the timestamp returned from  $\omega$  is greater than the one returned from  $\pi$ , i.e.  $ts_{\omega} > ts_{\pi}$ , then  $\pi \prec \omega$  follows directly. Similarly, if  $ts_{\omega} < ts_{\pi}$  holds, then  $\omega \prec \pi$  follows. If  $ts_{\omega} = ts_{\pi}$ , then it must be that  $\pi$  is a read and  $\pi$  either discovered  $ts_{\omega}$  as the *minimum* timestamp in at least a majority of servers or returned fast  $ts_{\omega}$  because it was noticed in at least a majority of servers. Thus,  $\omega \prec \pi$  follows.

**A3** Every read operation returns the value of the last write preceding it according to  $\prec$  (or the initial value if there is no such write).

Let  $\omega$  be the last write preceding read  $\rho$ . From our definition it follows that  $ts_{\rho} \geq ts_{\omega}$ . If  $ts_{\rho} = ts_{\omega}$ , then  $\rho$  returned the value written by  $\omega$  on a majority of servers. If  $ts_{\rho} > ts_{\omega}$ , then it means that  $\rho$  obtained a larger timestamp. However, the larger timestamp can only be originating from a write that succeeds  $\omega$ , thus  $\omega$  is not the preceding write and this cannot be the case. Lastly, if  $ts_{\rho} = 0$ , no preceding writes exist, and  $\rho$  returns the initial value. □

Having shown liveness and atomicity of algorithm OHSAM' the result follows.

### 5.1.2 Performance.

In algorithm OHSAM' write operations take 2 exchanges and read operations take 2 or 3 exchanges. The (worst case) message complexity of read operations is  $|\mathcal{S}|^2 + 3|\mathcal{S}|$  and the (worst case) message complexity of write operations is  $2|\mathcal{S}|$ . These results follows directly from the structure of the algorithm. We now provide additional details.

**Operation Latency.** *Write operation latency:* The structure of the write protocol of OHSAM' is identical to OHSAM, thus a write operation consists of 2 communication exchanges as reasond in Section 3.2.

*Read operation latency:* A reader sends a `readRequest` message to all servers in the first communication exchange E1. Once the servers receive the `readRequest` message they broadcast a `readRelay` message to all the servers *and* to the requesting reader in the exchange E2. The reader can terminate at the end of the second exchange, E2, if it can be “fast” and complete. If not, then the operation waits for exchange E3 as in algorithm OHSAM before completion. Thus, a read operation terminate either in 2 or 3 communication exchanges.

**Message Complexity.** *Write operation:* The structure of the write protocol of OHSAM' is identical to OHSAM, thus, as reasond in Section 3.2,  $4|\mathcal{S}|$  messages are exchanged during a write operation.

*Read operation:* Read operations in the worst case take 3 communication exchanges. Exchange E1 occurs when a reader sends a `readRequest` message to all servers in  $\mathcal{S}$ . The second exchange E2 occurs when servers in  $\mathcal{S}$  send `readRelay` messages to all servers in  $\mathcal{S}$  and to the requesting reader. The final exchange E3 occurs when servers in  $\mathcal{S}$  send a `readAck` message to the reader. Summing together  $|\mathcal{S}| + (|\mathcal{S}|^2 + |\mathcal{S}|) + |\mathcal{S}|$ , shows that in the worst case,  $|\mathcal{S}|^2 + 3|\mathcal{S}|$  messages are exchanged during a read operation.

## 5.2 Algorithm OHMAM' for the MWMR setting

Algorithm OHMAM' is obtained similarly to OHSAM' by (i) using tags instead of timestamps in the revised read protol of OHSAM' and (ii) using the write protocol of OHMAM without any modifications. Next, we reason about OHMAM' correctness.

### 5.2.1 Correctness.

*Liveness* and *atomicity* of the revised algorithm OHMAM' is shown similarly to algorithm OHMAM as reasoned in Section 4.1.

**Liveness.** Termination of Algorithm OHMAM' is guaranteed with respect to our failure model: up to  $f$  servers may fail, where  $f < |\mathcal{S}|/2$ , and operations wait for messages only from a majority of servers. We now give additional details.

*Read Operation.* A read operation of OHMAM' differs from OHSAM' by using tags instead of timestamps in order to impose an ordering on the values written. The structure of the read protocol is identical to OHSAM', thus *liveness* is ensured as reasoned in section 3.1.

*Write Operation.* Since the write protocol of algorithm OHMAM' is identical to the one that algorithm OHMAM uses, *liveness* is guaranteed as discussed in Section 4.1.

**Atomicity.** Next, we show how algorithm OHMAM' satisfies atomicity. Due to the similarity of the writer and server protocols of algorithm OHMAM' to those in OHMAM, we state the lemmas and we omit some of the proofs. Lemma 17 shows that the timestamp variable  $ts$  maintained by each server  $s$  in the system is monotonically non-decreasing.

**Lemma 17** *In any execution  $\xi$  of OHMAM', the variable  $tg$  maintained by any server  $s$  in the system is non-negative and monotonically increasing.*

**Proof.** Lemma 6 for algorithm OHMAM also applies to algorithm OHMAM' because the modification in line 147 does not affect the update of the tag  $tg$  at the server protocol.  $\square$

Next, we show that if a read operation  $\rho_2$  succeeds read operation  $\rho_1$ , then  $\rho_2$  always returns a value at least as recent as the one returned by  $\rho_1$ .

**Lemma 18** *In any execution  $\xi$  of OHMAM', If  $\rho_1$  and  $\rho_2$  are two read operations such that  $\rho_1$  precedes  $\rho_2$ , i.e.,  $\rho_1 \rightarrow \rho_2$ , and  $\rho_1$  returns a tag  $tg_1$ , then  $\rho_2$  returns a tag  $tg_2 \geq tg_1$ .*

**Proof.** Since read operations in algorithm OHMAM terminate in 3 communication exchanges then from Lemma 7 we know that the any two non-concurrent 3-exchange satisfy this. Thus we now have to show that the lemma holds for the cases where (i) a 2-exchange read operation  $\rho_1$  precedes

a 2-exchange read operation  $\rho_2$ ; (ii) a 2-exchange read operation  $\rho_1$  precedes a 3-exchange read operation  $\rho_2$ ; and (iii) a 3-exchange read operation  $\rho_1$  precedes a 2-exchange read operation  $\rho_2$ . Let the two operations  $\rho_1$  and  $\rho_2$  be invoked by processes with identifiers  $r_1$  and  $r_2$  respectively (not necessarily different).

*Case (i).* Assume by contradiction that 2-exchange read operations  $\rho_1$  and  $\rho_2$  exist such that  $\rho_2$  succeeds  $\rho_1$ , i.e.,  $\rho_1 \rightarrow \rho_2$ , and operation  $\rho_2$  returns a tag  $tg_2$  that is smaller than tag  $tg_1$  returned by  $\rho_1$ , i.e.,  $tg_2 < tg_1$ . Since both operations  $\rho_1$  and  $\rho_2$  complete in 2 exchanges, they both have to collect  $|\mathcal{S}|/2 + 1$  **readRelay** messages with the same tag  $tg_1$  and  $tg_2$  respectively. We know that after the completion of  $\rho_1$  at least  $|\mathcal{S}|/2 + 1$  servers have a tag at least as  $tg_1$ . By monotonicity of the tag at the servers (Lemma 17) and the fact that  $\rho_1$  is completed it follows that it is impossible for  $\rho_2$  to collect  $|\mathcal{S}|/2 + 1$  **readRelay** messages with the same tag  $tg_2$  s.t.  $tg_2 < tg_1$ . In particular,  $\rho_2$  can receive only  $|\mathcal{S}|/2$  **readRelay** messages with a timestamp  $tg_2$  s.t.  $tg_2 < tg_1$ . Therefore, this contradicts our assumption.

*Case (ii).* We know that since  $\rho_1$  is a 2-exchange operation then  $r_1$  receives at least  $|\mathcal{S}|/2 + 1$  **readRelay** messages that include the same tag  $tg_1$ . Thus after the completion of  $\rho_1$  at least a majority of servers hold a timestamp at least as  $tg_1$ . Servers relay to each other and wait for **readRelay** messages from a majority of servers before they send a **readAck** message to the reader  $r_2$ . By Lemma 17 since the tag variable  $tg_s$  maintained by servers is monotonically increasing then each server  $s_i$  that sends a **readAck** message to  $r_2$  must include a tag  $tg_{s_i}$  s.t.  $tg_{s_i} \geq tg_1$ . Therefore, the minimum tag  $tg_2$  that  $r_2$  can observe in each **readAck** message received from  $s_i$  must be  $tg_2 \geq tg_{s_i} \geq tg_1$ . Since a 3-exchange read operation decides on the minimum tag observed in the **readAck** responses, reader  $r_2$  decides on a timestamp  $tg_2$  s.t.  $tg_2 \geq tg_1$ .

*Case (iii).* Since  $\rho_1$  is a 3-exchange operation,  $r_1$  receives at least  $|\mathcal{S}|/2 + 1$  **readAck** messages that include the minimum tag  $tg_1$ . Servers relay to each other before they send a **readAck** message to  $\rho_1$ , then, by the monotonicity of tags at servers (Lemma 17), after the completion of  $\rho_1$  at least a majority of servers  $s_i \in RSet$  hold a tag at least as  $tg_1$ . Let  $RRSet$  be the set of servers that send a **readRelay** message to  $r_2$  during  $\rho_2$ . In order for  $\rho_2$  to terminate the size of  $RRSet$  must be at least  $|\mathcal{S}|/2 + 1$  based on the read protocol. Let  $tg_{s_i}$  be a tag received from a server  $s_i \in RRSet$ . Since  $RSet_1 \cap RRSet \neq \emptyset$  then the 2-exchange operation  $\rho_2$  that succeeds  $\rho_1$  can receive at most  $|\mathcal{S}|/2$  (minority) **readRelay** messages with a tag  $tg_{s_i}$  s.t.  $tg_{s_i} < tg_1$ . Thus, when  $\rho_2$  terminates it must return a tag  $tg_2$  s.t.  $tg_2 \geq tg_1$  and the lemma follows.  $\square$

Now we show that if a read operation succeeds a write operation, then it returns a value at least as recent as the one written.

**Lemma 19** *In any execution  $\xi$  of OHMAM', if read operation  $\rho$  succeeds write operation  $\omega$  (i.e.,  $\omega \rightarrow \rho$ ) that writes value  $v$  associated with tag  $tg$  and returns tag  $tg'$ , then  $tg' \geq tg$ .*

**Proof.** From Lemma 10 for algorithm OHMAM we know that the lemma holds if a 3-exchange read operation succeeds a write operation. We now show that the same holds for 2-exchange read operations.

Assume by contradiction that a 2-exchange read operation  $\rho$  and a write operation  $\omega$  exist such that  $\rho$  succeeds  $\omega$ , i.e.  $\omega \rightarrow \rho$ , and  $\rho$  returns a tag  $tg'$  that is smaller than the tag  $tg$  that  $\omega$  wrote,  $tg' < tg$ . From the algorithm, in order for this to happen,  $\rho$  receives  $|\mathcal{S}|/2 + 1$  **readRelay** messages that all include the same tag  $tg'$  and  $tg' < tg$ . Since  $\omega$  is complete it means that at least a majority of servers hold tag  $tg_s$  s.t.  $tg_s \geq tg$ . Since any two majorities intersect, this contradicts

the assumption that  $\rho$  receives  $|S|/2 + 1$  `readRelay` messages that all include the same timestamp  $tg'$  where  $tg' < tg$  and the lemma follows.  $\square$

Next, we reason that if a write operation  $\omega_2$  succeeds write operation  $\omega_1$ , then  $\omega_2$  writes a value associated with a tag strictly higher than  $\omega_1$ .

**Lemma 20** *In any execution  $\xi$  of OHMAM', if a write operation  $\omega_1$  writes a value with tag  $tg_1$  then for any succeeding write operation  $\omega_2$  that writes a value with tag  $tg_2$  we have  $tg_2 > tg_1$ .*

**Proof.** The modifications of OHMAM' do not have an impact on the write operations thus this lemma follows directly from lemma 8 of OHMAM.  $\square$

**Lemma 21** *In any execution  $\xi$  of OHMAM', if a write  $\omega$  succeeds a read operation  $\rho$  that reads tag  $tg$ , i.e.  $\rho \rightarrow \omega$ , and returns a tag  $tg'$ , then  $tg' > tg$ .*

**Proof.** The case where the read operation takes three communication exchanges to terminate is identical as in lemma 11 of algorithm OHMAM. Thus, we are interested to examine the case where the read terminates in two communication exchanges.

Let  $RR$  be the set of servers that sent `readRelay` messages to  $\rho$ . Let  $dAck$  be the set of servers that sent `discoverAck` messages to  $\omega$ . Let  $wAck$  be the set of servers that sent `writeAck` messages to  $\omega$  and let  $RA$  be the set of servers that sent `readAck` messages to  $\rho$ . It is not necessary that  $a \neq b \neq c$  holds.

In the case we examine, the read operation  $\rho$  terminates when it receives `readRelay` messages from a majority of servers and  $\rho$  decides on a tag that all servers attached in the set  $RA$  and lastly it terminates. Writer  $\omega$ , initially it broadcasts a `discover` message to all servers, and it then awaits for “fresh” `discoverAck` messages from a majority of servers, that is, set  $dAck$ . Each of  $RA$  and  $dAck$  sets are from majorities of servers, and so  $RA \cap dAck \neq \emptyset$ . By Lemma 17, any server  $s_k \in RA \cap dAck$  has a tag  $tg_{s_k}$  s.t.  $tg_{s_k} \geq tg$ . Since  $\omega$  generates a new local tag-value  $(tg', v)$  pair in which it assigns the timestamp to be one higher than the one discovered in the *maximum* tag from set  $dAck$ , it follows that  $tg' > tg$ . Furthermore,  $\omega$  broadcasts the value to be written associated with  $tg'$  in a `writeRequest` message to all servers and it awaits for `writeAck` messages from a majority of servers before completion, set  $wAck$ . Observe that each of  $dAck$  and  $wAck$  sets are from majority of servers, and so  $dAck \cap wAck \neq \emptyset$ . By Lemma 6, any server  $s_k \in dAck \cap wAck$  has a tag  $tg_{s_k}$  s.t.  $tg_{s_k} \geq tg' > tg$  and the result follows.  $\square$

Similarly to Theorem 16 we show the following for algorithm OHMAM'.

**Theorem 22** *Algorithm OHMAM' implements an atomic MWMR object.*

**Proof.** We use the above lemmas and the operations order definition (using tags instead of timestamps) to reason about each of the three *atomicity* conditions A1, A2 and A3.

**A1** For any  $\pi_1, \pi_2 \in \Pi$  such that  $\pi_1 \rightarrow \pi_2$ , it cannot be that  $\pi_2 \prec \pi_1$ .

If both  $\pi_1$  and  $\pi_2$  are writes and  $\pi_1 \rightarrow \pi_2$  holds, then from Lemma 21 it follows that  $tg_{\pi_2} > tg_{\pi_1}$ . By the ordering definition  $\pi_1 \prec \pi_2$  is satisfied. When  $\pi_1$  is a write,  $\pi_2$  a read and  $\pi_1 \rightarrow \pi_2$  holds, then from Lemma 19 it follows that  $tg_{\pi_2} \geq tg_{\pi_1}$ . By definition  $\pi_1 \prec \pi_2$  is satisfied. If  $\pi_1$  is a read,  $\pi_2$  a write and  $\pi_1 \rightarrow \pi_2$  holds, then from Lemma 21 it follows that  $\pi_2$  always returns a tag  $tg_{\pi_2}$

s.t.  $tg_{\pi_2} > tg_{\pi_1}$ . By the ordering definition  $\pi_1 \prec \pi_2$  is satisfied. If both  $\pi_1$  and  $\pi_2$  are reads and  $\pi_1 \rightarrow \pi_2$  holds, then from Lemma 18 it follows that the tag returned from  $\pi_2$  is always greater or equal to the one returned from  $\pi_1$ .  $tg_{\pi_2} \geq tg_{\pi_1}$ . If  $tg_{\pi_2} > tg_{\pi_1}$ , then by the ordering definition  $\pi_1 \prec \pi_2$  holds. When  $tg_{\pi_2} = tg_{\pi_1}$  then the ordering is not defined but it cannot be that  $\pi_2 \prec \pi_1$ .

**A2** For any write  $\omega \in \Pi$  and any operation  $\pi \in \Pi$ , then either  $\omega \prec \pi$  or  $\pi \prec \omega$ .

If  $tg_{\omega} > tg_{\pi}$ , then  $\pi \prec \omega$  follows directly. Similarly, if  $tg_{\omega} < tg_{\pi}$  holds, then it follows that  $\omega \prec \pi$ . When  $ts_{\omega} = ts_{\pi}$  holds, then the uniqueness of each tag that a writer creates ensures that  $\pi$  is a read. In particular, remember that each tag is a  $\langle ts, id \rangle$  pair, where  $ts$  is a natural number and  $id$  a writer identifier. Tags are ordered lexicographically, first with respect to the timestamp and then with respect to the writer id. Since the writer ids are unique, then even if two writes use the same timestamp  $ts$  in the tag pairs they generate, the two tags cannot be equal as they will differ on the writer id. Furthermore, if the two tags are generated by the same writer, then by well-formedness it must be the case that the latter write will contain a timestamp larger than any timestamp used by that writer before. Since  $\pi$  is a read operation that receives either (i) **readAck** messages from a majority of servers, or (ii) **readRelay** messages from a majority of servers with the same  $tg$ , then the intersection properties of majorities ensure that  $\omega \prec \pi$ .

**A3** Every read operation returns the value of the last write preceding it according to  $\prec$  (or the initial value if there is no such write).

Let  $\omega$  be the last write preceding read  $\rho$ . From our definition it follows that  $tg_{\rho} \geq tg_{\omega}$ . If  $tg_{\rho} = tg_{\omega}$ , then  $\rho$  returned a value written by  $\omega$  in some servers. Read  $\rho$  waited either (i) for **readAck** messages from a majority of servers, or (ii) **readRelay** messages from a majority of servers with the same  $tg$ . Thus the intersection properties of majorities ensure that  $\omega$  was the last complete write operation. If  $tg_{\rho} > tg_{\omega}$  holds, it must be the case that there is a write  $\omega'$  that wrote  $tg_{\rho}$  and by definition it must hold that  $\omega \prec \omega' \prec \rho$ . Thus,  $\omega$  is not the preceding write and this cannot be the case. Lastly, if  $tg_{\rho} = 0$ , no preceding writes exist, and  $\rho$  returns the initial value.  $\square$

Having shown liveness and atomicity of algorithm OHMAM' the result follows.

### 5.2.2 Performance.

In algorithm OHMAM' write operations take 4 exchanges and read operations take 2 or 3 exchanges. The (worst case) message complexity of read operations is  $|\mathcal{S}|^2 + 3|\mathcal{S}|$  and the (worst case) message complexity of write operations is  $4|\mathcal{S}|$ . We now provide additional details.

**Operation Latency.** *Write operation latency:* The structure of the write protocol of OHMAM' is identical to OHMAM, thus a write operation consists of 4 communication exchanges as reasoned in Section 4.2.

*Read operation latency:* The structure of the read protocol of OHMAM' is identical to OHSAM', thus a read operation consists of *at most* 3 communication exchanges as reasoned in Section 5.1.2.

**Message Complexity.** *Write operation:* The structure of the write protocol of OHMAM' is identical to OHMAM, thus, as reasoned in Section 4.2,  $4|\mathcal{S}|$  messages are exchanged in a write operation.

*Read operation:* The structure of the read protocol of OHMAM' is identical to OHSAM', thus, as reasoned in Section 5.1.2,  $|\mathcal{S}|^2 + 3|\mathcal{S}|$  messages are exchanged during a read operation.



## 6 Empirical Evaluations

Here we present a comparative study of our algorithms by simulating them using the NS3 discrete event simulator [1]. We implemented the following three SWMR algorithms: ABD [2], OHSAM, and OHSAM'. We also implemented the corresponding three MWMR algorithms: ABD-MW (following the multi-writer extension [10]), OHMAM, and OHMAM'. For comparison we also implemented a benchmark, called LB, that mimics the minimum message requirements for the SWMR and MWMR settings but without performing any computation or ensuring consistency. In particular, LB performs two communication exchanges for read and write operations thus providing a lower bound on performance in simulated scenarios. Note that LB, does not serve the properties of Atomicity and its use is strictly serving comparison purposes.

**Experimentation Setup.** The experimental configuration consists of a single (SWMR) or multiple (MWMR) writers, a set of readers, and a set of servers. We assume that at most one server may fail. This is done to subject the system to a high communication burden. Communication among the nodes is established via point-to-point bidirectional links implemented with a DropTail queue.

For our evaluation, we use simulations representing two different topologies, *Series* and *Star*, that include the same array of routers but differ in the deployment of server nodes. In both topologies clients are connected to routers over 5Mbps links with 2ms delay, the routers are connected over 10Mbps links with 4ms delay. In the *Series* topology in Figure 1(a), a server is connected to each router over 10Mbps bandwidth with 2ms delay. This topology models a network where servers are separated and appear to be in different networks. In the *Star* topology in Figure 1(b) all servers are connected to a single router over 50Mbps links with 2ms delay, modeling a network where servers are in a close proximity and are well-connected, e.g., as in a datacenter. In both topologies readers and writer(s) are located uniformly with respect to the routers. We ran NS3 on a Mac OS X with 2.5Ghz Intel Core i7 processor. The results are compiled as averages over five samples per each scenario.

**Performance.** We assess algorithms in terms of *operation latency* that depends on communication delays and local computation delays. NS3 supports simulated time events, but does not measure delays due to local computation. In order to measure operation latency we combine two clocks: the simulation clock to measure communication delays, and a real time clock to measure computation delays. The sum of the two times yields operation latency.

**Scenarios.** To measure performance we define several scenarios. The scenarios are designed to test (i) the scalability of the algorithms as the number of readers, writers, and servers increases; (ii) the contention effect on efficiency, by running different concurrency scenarios; and (iii) the effects of chosen topologies on the efficiency. For scalability we test with the number of readers  $|\mathcal{R}|$  from the set  $\{10, 20, 40, 80, 100\}$  and the number of servers  $|\mathcal{S}|$  from the set  $\{10, 15, 20, 25, 30\}$ .

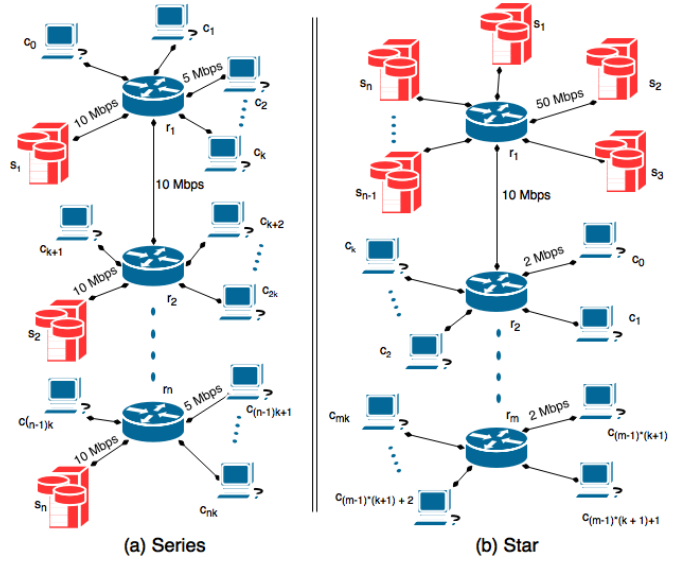


Figure 1: Simulated topologies.

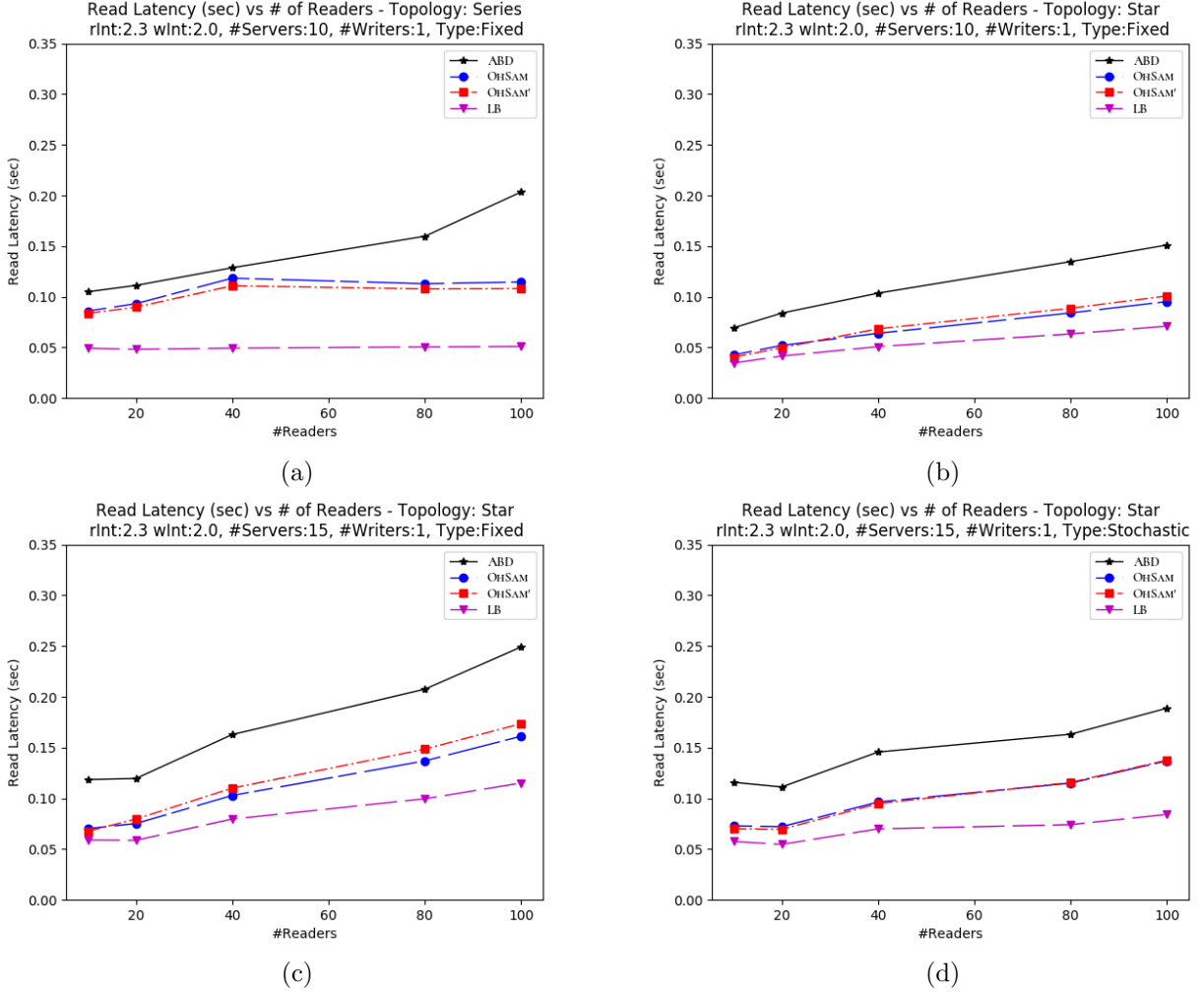
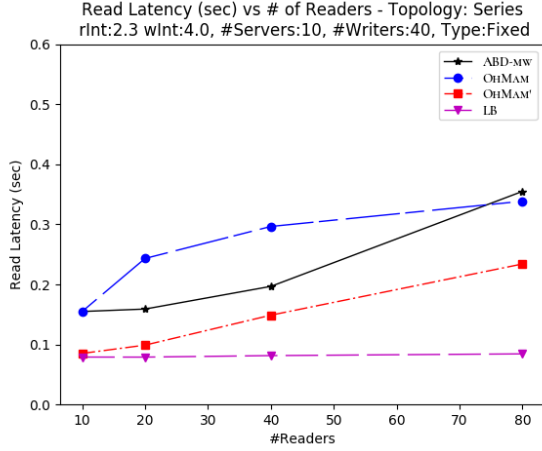


Figure 2: SWMR Simulation Results.

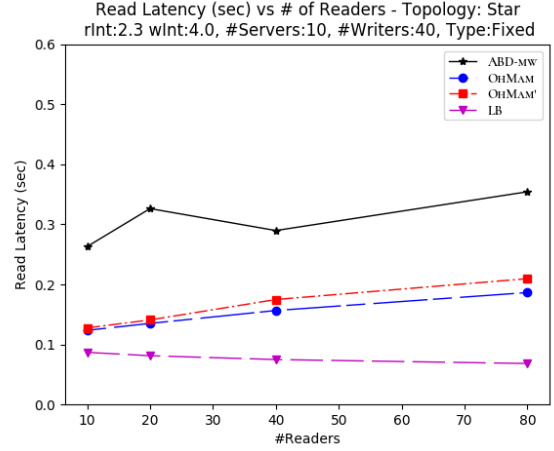
For the MWMR setting we use at most 80 readers and we range the number of writers  $|\mathcal{W}|$  over the set  $\{10, 20, 40\}$ . To test contention we set the frequency of each read and write operation to be constant and we define two different invocation schemes. We issue reads every  $rInt = 2.3$  seconds and write operations every  $wInt = 4$  seconds. We define two invocation schemes: *fixed* and *stochastic*. In the *fixed* scheme all operations are scheduled periodically at a constant interval. In the *stochastic* scheme read and write operations are scheduled randomly from the intervals  $[1, rInt]$  and  $[1, wInt]$  respectively. To test the effects of topology we run our simulations using the *Series* and *Star* topologies.

**Results.** We generally observe that the proposed algorithms outperform algorithms ABD and ABD-MW in most scenarios by a factor of 2. A closer examination yields the following observations.

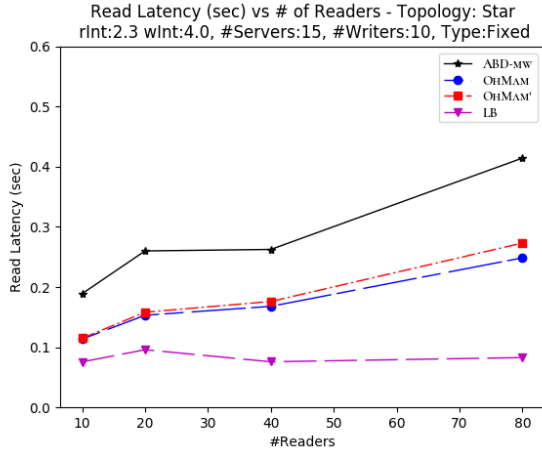
*Scalability:* As seen in Figures 2(b) and 2(c), increasing the number of readers and servers increases latency in the SWMR algorithms. The same observation holds for the MWMR algorithms. When the number of the participating readers and writers is reduced then not surprisingly the latency improves, but the relative performance of the algorithms remains the same.



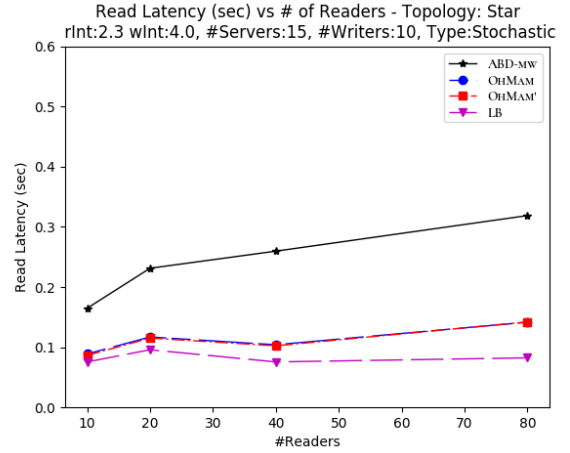
(e)



(f)



(g)



(h)

Figure 3: MWMR Simulation Results.

*Contention:* We examine the efficiency of our algorithms under different concurrency schemes. We set the operation frequency to be constant and we observe that in the *stochastic* scheme read operations complete faster than in the *fixed* scheme; see Figures 2(c) and 2(d) for the SWMR setting, and Figures 3(g) and 3(h) for the MWMR setting. This is expected as the *fixed* scheme causes congestion. For the *stochastic* scheme the invocation time intervals are distributed uniformly, this reduces congestion and improves latency.

*Topology:* Figures 2(a) and 2(b) for the SWMR setting, and Figures 3(e) and 3(f) for the MWMR setting show that topology substantially impacts performance. For both the SWMR and MWMR settings our algorithms outperform algorithms ABD and ABD-MW by a factor of at least 2 in *Star* topology where servers are well-connected. Our SWMR algorithms perform much better than ABD also in the *Series* topology. For the MWMR setting and *Series* topology, we note that ABD-MW generally outperforms algorithm OHMAM, however the revised algorithm OHMAM' noticeably outperforms ABD-MW.

Lastly we compare the performance of algorithms OHSAM and OHMAM with revised versions

OHSAM' and OHMAM'. We note that OHSAM' and OHMAM' outperform all other algorithms in *Series* topologies. However, and perhaps not surprisingly, OHSAM and OHMAM outperform OHSAM' and OHMAM' in *Star* topology. This is explained as follows. In *Star* topology *readRelay* and *readAck* messages are exchanged quickly at the servers and thus delivered quickly to the clients. On the other hand, the bookkeeping mechanism used in the revised algorithms incur additional computational latency, resulting in worse latency.

An important observation is that while algorithms OHSAM' and OHMAM' improve the latencies of some operations (allowing some reads to complete in two exchanges), their performance relative to algorithms OHSAM and OHMAM depends on the deployment setting. Simulations show that OHSAM and OHMAM are more suitable for datacenter-like deployment, while in the “looser” settings algorithms OHSAM' and OHMAM' perform better.

## 7 Conclusions

We focused on the problem of emulating atomic read/write shared objects in message-passing settings with the goal of using three communication exchanges (to the extent allowed by the impossibility result [6]). We presented algorithms for the SWMR and MWMR models. We then revised the algorithms to speed up some read operations. We rigorously reasoned about the correctness of our algorithms. The algorithms do not impose any constraints on the number of readers (SWMR and MWMR) and on the number of the writers for the MWMR model. Finally we performed an empirical study of the performance of algorithms using simulations.

## References

- [1] NS3 network simulator. <https://www.nsnam.org/>.
- [2] ATTIYA, H., BAR-NOY, A., AND DOLEV, D. Sharing memory robustly in message passing systems. *Journal of the ACM* 42(1) (1996), 124–142.
- [3] DUTTA, P., GUERRAUI, R., LEVY, R. R., AND CHAKRABORTY, A. How fast can a distributed atomic read be? In *Proceedings of the 23rd ACM symposium on Principles of Distributed Computing (PODC)* (2004), pp. 236–245.
- [4] ENGLERT, B., GEORGIOU, C., MUSIAL, P. M., NICOLAOU, N., AND SHVARTSMAN, A. A. On the efficiency of atomic multi-reader, multi-writer distributed memory. In *Proceedings 13th International Conference On Principle Of Distributed Systems (OPODIS 09)* (2009), pp. 240–254.
- [5] GEORGIOU, C., NICOLAOU, N. C., AND SHVARTSMAN, A. A. Fault-tolerant semifast implementations of atomic read/write registers. *Journal of Parallel and Distributed Computing* 69, 1 (2009), 62–79.
- [6] HADJISTASI, T., NICOLAOU, N., AND SCHWARZMANN, A. A. On the impossibility of one-and-a-half round atomic memory. [www.arxiv.com](http://www.arxiv.com), 2016.
- [7] HERLIHY, M. P., AND WING, J. M. Linearizability: a correctness condition for concurrent objects. *ACM Transactions on Programming Languages and Systems (TOPLAS)* 12, 3 (1990), 463–492.

- [8] LAMPORT, L. How to make a multiprocessor computer that correctly executes multiprocess program. *IEEE Trans. Comput.* 28, 9 (1979), 690–691.
- [9] LYNCH, N. *Distributed Algorithms*. Morgan Kaufmann Publishers, 1996.
- [10] LYNCH, N. A., AND SHVARTSMAN, A. A. Robust emulation of shared memory using dynamic quorum-acknowledged broadcasts. In *Proceedings of Symposium on Fault-Tolerant Computing* (1997), pp. 272–281.