Self-Stabilizing Snapshot Objects for Asynchronous Fail-Prone Network Systems

(preliminary report)

Chryssis Georgiou † Oskar Lundström ‡ Elad M. Schiller §

1 Introduction

We propose self-stabilizing implementations of shared memory snapshot objects for asynchronous networked systems whose nodes may fail-stop.

A snapshot object simulates the behavior of an array of single-writer/multi-reader shared registers that can be read atomically. Delporte-Gallet et al. proposed two fault-tolerant algorithms for snapshot objects in asynchronous crash-prone message-passing systems. Their first algorithm is non-blocking; it allows snapshot operations to terminate once all write operations had ceased. It uses $O(n)$ messages of $O(n \cdot \nu)$ bits, where $n$ is the number of nodes and $\nu$ is the number of bits it takes to represent the object. Their second algorithm allows snapshot operations to always terminate independently of write operations. It incurs $O(n^2)$ messages.

The fault model of Delporte-Gallet et al. considers both node failures (crashes). We aim at the design of even more robust snapshot objects. We do so through the lenses of self-stabilization—a very strong notion of fault-tolerance. In addition to Delporte-Gallet et al.’s fault model, a self-stabilizing algorithm can recover after the occurrence of transient faults; these faults represent arbitrary violations of the assumptions according to which the system was designed to operate (as long as the code stays intact).

In particular, in this work, we propose self-stabilizing variations of Delporte-Gallet et al.’s non-blocking algorithm and always-terminating algorithm. Our algorithms have similar communication costs to the ones by Delporte-Gallet et al. and $O(1)$ recovery time (in terms of asynchronous cycles) from transient faults. The main differences are that our proposal considers repeated gossiping of $O(\nu)$ bits messages and deals with bounded space (which is a prerequisite for self-stabilization). Lastly, we explain how to extend the proposed solutions to reconfigurable ones.
Context and Motivation. Shared registers are fundamental objects that facilitate synchronization in distributed systems. In the context of networked systems, they provide a higher abstraction level than simple end-to-end communication, which provides persistent and consistent distributed storage that can simplify the design and analysis of dependable distributed systems. Snapshot objects extend shared registers. They provide a way to further make the design and analysis of algorithms that base their implementation on shared registers easier. Snapshot objects allow an algorithm to construct consistent global states of the shared storage in a way that does not disrupt the system computation. Their efficient and fault-tolerant implementation is a fundamental problem, as there are many examples of algorithms that are built on top of snapshot objects; see textbooks such as [32, 33], as well as recent reviews, such as [31].

Task description. Consider a fault-tolerant distributed system of \( n \) asynchronous nodes that are prone to failures. Their interaction is based on the emulation of Single-Writer/Multi-Reader (SWMR) shared registers over a message-passing communication system. Snapshot objects can read the entire array of system registers [2, 4]. The system lets each node update its own register via \texttt{write()} operations and retrieve the value of all shared registers via \texttt{snapshot()} operations. Note that these snapshot operations may occur concurrently with the write operations that individual nodes perform. We are particularly interested in the study of atomic snapshot objects that are linearizable [23]: the operations \texttt{write()} and \texttt{snapshot()} appear as if they have been executed instantaneously, one after the other (in other words, they appear to preserve real-time ordering).

Fault Model. We consider an asynchronous message-passing system that has no guarantees on the communication delay. Moreover, there is no notion of global (or universal) clocks and we do not assume that the algorithm can explicitly access the local clock (or timeout mechanisms). Our fault model includes (i) fail-stop failures of nodes, and (ii) communication failures, such as packet omission, duplication, and reordering. In addition, to the failures captured in our model, we also aim to recover from transient faults, i.e., any temporary violation of assumptions according to which the system and network were designed to behave, e.g., the corruption of control variables, such as the program counter and operation indices, which are responsible for the correct operation of the studied system, or operational assumptions, such as that at least half of the system nodes never fail. Since the occurrence of these failures can be combined, we assume that these transient faults can alter the system state in unpredictable ways. In particular, when modeling the system, we assume that these violations bring the system to an arbitrary state from which a self-stabilizing algorithm should recover the system. Therefore, starting from an arbitrary state, the correctness proof of self-stabilizing systems [13] has to demonstrate the return to a “correct behavior” within a bounded period, which brings the system to a legitimate state. The complexity measure of self-stabilizing systems is the length of the recovery period.

As transient faults can occur at any point in a system’s lifetime, self-stabilizing systems need to keep communicating its state structures for cleaning any potential corrupted (stale) information; to this respect, a self-stabilizing system cannot really terminate [14] Chapter 2.3. Specifically, the proposed solution repeatedly broadcasts \( \mathcal{O}(\nu) \)-size gossip messages that facilitate the system clean-up from stale information, where \( \nu \) is the number of bits it takes to represent the object. We note the trade-off between the cost related to these gossip messages and the recovery time. That is, one can balance this trade-off by, for example, reducing the rate of gossip messages, which prolongs the stabilization time. We clarify that the rate at which these repeated clean-up operations take place does not impact the execution time of the \texttt{write()} and \texttt{snapshot()} operations.

Related work. We follow the design criteria of self-stabilization, which was proposed by
Dijkstra [13] and detailed in [14]. We now overview existing work related to ours. Our review does not focus on algorithms for shared memory system; although there are examples for both non-self-stabilizing [25, 26] and self-stabilizing [1] solutions.

**Shared registers emulation in message-passing systems:** Attiya et al. [7] implemented SWMR atomic shared memory in an asynchronous networked system. They assume that the majority of the nodes do not crash or get disconnected. Their work builds on this assumption in the following manner: Any majority subset of the system nodes includes at least one non-failing node; thus, any two majority subsets of the system nodes have a non-empty intersection. They show that if a majority of the nodes acknowledge an update to the shared register, then that update can safely be considered visible to all non-failing nodes that retrieve the latest update from a majority of nodes. Attiya et al. also show that this assumption is essential for solvability. Their seminal work has many generalizations and applications [6]. The literature includes a large number of simulation of shared registers for networked systems, which differ in their fault tolerance properties, time complexity, storage costs, and system properties, e.g., [5, 20, 22, 28–30].

In the context of self-stabilization, the literature includes a practically-self-stabilizing variation for the work of Attiya et al. [7] by Alon et al. [3]. Their proposal guarantees wait-free recovery from transient faults. However, there is no bound on the recovery time. Dolev et al. [19] consider MWMR atomic storage that is wait-free in the absence of transient faults. They guarantee a bounded time recovery from transient faults in the presence of a fair scheduler. They demonstrate the algorithm’s ability to recover from transient faults using unbounded counters and in the presence of fair scheduling. Then they deal with the event of integer overflow via a consensus-based procedure. Since integer variables can have 64-bits, their algorithm seldom uses this non-wait-free procedure for dealing with integer overflows. In fact, they model integer overflow events as transient faults, which implies bounded recovery time from transient faults in the seldom presence of a fair scheduler (using bounded memory). They call these systems **self-stabilizing systems in the presence of seldom fairness**. Our work adopts these design criteria.

We also make use of their self-stabilizing quorum and gossip services [19, Section 13].

**Implementing a snapshot object on top of a message-passing system:** A straightforward way for implementing snapshot objects is to consider a layer of $n$ SWMR atomic registers emulated in a networked system. This way we can run on top of this layer any algorithm for implementing a snapshot object for a system with shared variables. Delporte-Gallet et al. [12] avoid this composition, obtaining, in this way, a more efficient implementation with respect to the communication costs. Specifically, they claim that when stacking the shared-memory atomic snapshot algorithm of [2] on the shared-memory emulation of [7] (with some improvements), the number of messages per snapshot operation is $8n$ and it takes four round trips. Their proposal, instead, takes $2n$ message per snapshot operation and just one round trip to complete. The algorithms we propose in the present work follow the non-stacking approach of Delporte-Gallet and they have the same communication costs for write and snapshot operations. Moreover, they tolerate any failure (in any communication or operation invocation pattern) that [12] can. Furthermore, our algorithms deal with transient faults by periodically removing stale information. To that end, the algorithms broadcast gossip message of $O(1)$ bits, where $\nu$ is the number of bits it takes to represent the object.

In the context of self-stabilization, there exist algorithms for the propagation of information with feedback, e.g., [11] that can facilitate the implementation of snapshot objects that can recover from transient faults, but not from node failures. For the sake of clarity, we note that “stacking” of self-stabilizing algorithms for asynchronous systems is not a straightforward process (since the existing “stacking” require schedule fairness, see [13, Section 2.7]). Moreover, we are unaware of an attempt in the literature to stack a self-stabilizing shared-memory atomic snapshot
algorithm (such as the weak snapshots algorithm of Abraham [11] that uses $O(n)$ register size) over a self-stabilizing shared-memory emulation, such as the one of Dolev et al. [15].

**Our Contributions.** We present an important module for dependable distributed systems: self-stabilizing algorithms for snapshot objects in networked systems. To the best of our knowledge, we are the first to provide a broad fault model that includes both node failures and transient faults. Specifically, we advance the state of the art as follows:

1. As a first contribution, we offer a self-stabilizing variation of the non-blocking algorithm presented by Delporte-Gallet et al. [12]. Their solution tolerates node failures as well as packet omission, duplication, and reordering. Each snapshot or write operation uses $O(n)$ messages of $O(\nu \cdot n)$ bits, where $n$ is the number of nodes and $\nu$ is the number of bits for encoding the object. The termination of a snapshot operation depends on the assumption that the invocation of all write operations cease eventually.

   Our solution broadens the set of failure types it can tolerate, since it can also recover after the occurrence of transient faults, which model any violation of the assumptions according to which the system was designed to operate (as long as the code stays intact). We increase the communication costs slightly by using $O(n^2)$ gossip messages of $O(\nu)$ bits, where $\nu$ is the number of bits it takes to represent the object.

2. Our second contribution offers a self-stabilizing all-operation always-terminating variation of the snapshot-only always-terminating algorithm presented by Delporte-Gallet et al. [12]. Our algorithm can: (i) recover from of transient faults, and (ii) both write and snapshot operations always terminate (regardless of the invocation patterns of any operation).

   We achieve (ii) by choosing to use safe registers for storing the result of recent snapshot operations, rather than a reliable broadcast mechanism, which often has higher communication costs. Moreover, instead of dealing with one snapshot task at a time, we take care of several at a time. We also consider an input parameter, $\delta$. For the case of $\delta = 0$, our self-stabilizing algorithm guarantees an always-termination behavior in a way that resembles the non-self-stabilizing algorithm by Delporte-Gallet et al. [12] that blocks all write operation upon the invocation of any snapshot operation at the cost of $O(n^2)$ messages. For the case of $\delta > 0$, our solution aims at using $O(n)$ messages per snapshot operation while monitoring the number of concurrent write operations. Once our algorithm notices that a snapshot operation runs concurrently with at least $\delta$ write operations, it blocks all write operations and uses $O(n^2)$ messages for completing the snapshot operations.

   Thus, the proposed algorithm can trade communication costs with an $O(\delta)$ bound on snapshot operation latency. Moreover, between any two consecutive periods in which snapshot operations block the system for write operations, the algorithm guarantees that at least $\delta$ write operations occur.

3. The two proposed algorithms presented in sections 4 and 5 consider unbounded counters. In Section 6, we explain how to bound these counters as well as how to extend our solutions to reconfigurable ones.

**Organization.** We state our system settings in Section 2. We review the non-self-stabilizing solutions by Delporte-Gallet et al. [12] in Section 3. Our self-stabilizing non-blocking and always-terminating algorithms are proposed in Sections 4 and 5, respectively; they consider unbounded counters. We explain how to bound the counters of the proposed self-stabilizing algorithms in Section 6. We conclude in Section 7.
2 System settings

We consider an asynchronous message-passing system that has no guarantees on the communication delay. Moreover, there is no notion of global (or universal) clocks and we do not assume that the algorithm can explicitly access the local clock (or timeout mechanisms). The system consists of \( n \) failure-prone nodes (or processors) with identifiers are unique and totally ordered in \( \mathcal{P} \).

2.1 Communication model

The network topology is of a fully-connected graph, \( K_n \), and any pair of nodes have access to a bidirectional communication channel that, at any time, has at most \( \text{capacity} \in \mathbb{N} \) packets. Every two nodes exchange (low-level messages called) packets to permit delivery of (high-level) messages. When node \( p_i \in \mathcal{P} \) sends a packet, \( m \), to node \( p_j \in \mathcal{P} \setminus \{p_i\} \), the operation send inserts a copy of \( m \) to channel \( i,j \), while respecting the upper bound \( \text{capacity} \) on the number of packets in the channel. In case channel \( i,j \) is full, i.e., \( |\text{channel}_{i,j}| = \text{capacity} \), the sending-side simply overwrites any message in channel \( i,j \). When \( p_j \) receives \( m \) from \( p_i \), the system removes \( m \) from channel \( i,j \). As long as \( m \in \text{channel}_{i,j} \), we say that \( m \)'s message is in transit from \( p_i \) to \( p_j \).

2.2 Execution model

Our analysis considers the interleaving model [14], in which the node’s program is a sequence of (atomic) steps. Each step starts with an internal computation and finishes with a single communication operation, i.e., message send or receive.

The state, \( s_i \), of node \( p_i \in \mathcal{P} \) includes all of \( p_i \)'s variables as well as the set of all incoming communication channels. Note that \( p_i \)'s step can change \( s_i \) as well as remove a message from channel \( j,i \) (upon message arrival) or add a message in channel \( i,j \) (when a message is sent). The term system state refers to a tuple of the form \( c = (s_1, s_2, \ldots, s_n) \) (system configuration), where each \( s_i \) is \( p_i \)'s state (including messages in transit to \( p_i \)). We define an execution (or run) \( R = c_0, a_0, c_1, a_1, \ldots \) as an alternating sequence of system states \( c_x \) and steps \( a_x \), such that each system state \( c_{x+1} \), except for the starting one, \( c_0 \), is obtained from the preceding system state \( c_x \) by the execution of step \( a_x \).

Let \( R' \) and \( R'' \) be a prefix, and respectively, a suffix of \( R \), such that \( R' \) is a finite sequence, which starts with a system state and ends with a step \( a_x \in R' \), and \( R'' \) is an unbounded sequence, which starts in the system state that immediately follows step \( a_x \) in \( R \). In this case, we can use \( \circ \) as the operator to denote that \( R = R' \circ R'' \) concatenates \( R' \) with \( R'' \).

2.3 Fault model

We model a failure as a step that the environment takes rather than the algorithm. We consider failures that can and cannot cause the system to deviate from fulfilling its task (Figure 1). The set of legal executions (\( LE \)) refers to all the executions in which the requirements of the task \( T \) hold. In this work, \( T_{\text{snapshot}} \) denotes our studied task of snapshot object emulation and \( LE_{\text{snapshot}} \) denotes the set of executions in which the system fulfills \( T_{\text{snapshot}} \)'s requirements. We say that a system state \( c \) is legitimate when every execution \( R \) that starts from \( c \) is in \( LE \). When a failure cannot cause the system execution (that starts in a legitimate state) to leave the set \( LE \), we refer to that failure as a benign one. We refer to any temporary violation of the assumptions according to which the system was designed to operate (as long as program code remains intact) as transient faults. Self-stabilizing algorithms deals with benign failures (while
| Frequency          | Rare                                                                 | Not rare                                                                 |
|--------------------|----------------------------------------------------------------------|--------------------------------------------------------------------------|
| **Duration**       | Any violation of the assumptions according to which the system is assumed to operate (as long as the code stays intact). This can result in any state corruption. | Packet failures: omissions, duplications, reordering (assuming communication fairness holds). |
| **Transient**      |                                                                      |                                                                          |
| **Permanent**      | Fail-stop failures.                                                  |                                                                          |

| Prior to the system start, consider all faults | Execution's starting state | Recovery period | Legal execution (LE) | Consider only benign faults |

Figure 1: The table above details our fault model and the chart illustrates when each fault set is relevant. The chart’s gray shapes represent the system execution, and the white boxes specify the failures considered to be possible at different execution parts and recovery guarantees of the proposed self-stabilizing algorithm. The set of benign faults includes both packet failures and fail-stop failures.

fulfilling the task requirements) and they can also recover, within a bounded period, after the occurrence of transient faults.

2.3.1 Benign failures

The algorithmic solutions that we consider are oriented towards asynchronous message-passing systems and thus they are oblivious to the time in which the packets arrive and departure (and require no explicit access to clock-based mechanisms, which may or may not be used by the system underlying mechanisms, say, for congestion control at the end-to-end protocol).

Communication fairness. Recall that we assume that the communication channel handles packet failures, such as omission, duplication, reordering (Section 2.1). We consider standard terms for characterizing node failures [21]. A crash failure considers the case in which a node stops taking steps forever and there is no way to detect this failure. A fail-stop failure considers the case in which a node stops taking steps and there is a way to detect this failure, say, using unreliable failure detectors [10]. We say that a failing node resumes when it returns to take steps without restarting its program — the literature sometimes refer to this as an undetectable restart. The case of a detectable restart allows the node to restart all of its variables. We assume that if $p_i$ sends a message infinitely often to $p_j$, node $p_j$ receives that message infinitely often. We refer to the latter as the fair communication assumption. For example, the proposed algorithm sends infinitely often GOSSIP messages from any processor to any other. Despite the possible loss of messages, the communication fairness assumption implies that every processor receives infinitely often GOSSIP messages from any non-failing processor. Note that fair communication provides no bound on the channel communication delays. It merely says that a message is received within some finite time if its sender does not stop sending it (until that sender receives the acknowledgment for that message). We refer to the latter as the fair communication assumption. We note that without the communication fairness assumption, the communication channel between any two correct nodes eventually becomes non-functional.

Node failure. We assume that the failure of node $p_i \in \mathcal{P}$ implies that it stops sending and receiving messages (and it also stops executing any other step) without any warning. We assume
that the number of failing nodes is bounded by \( f \) and that \( 2f < n \) for the sake of guaranteeing correctness \[27\]. In the absence of transient faults, failing nodes can simply crash (or fail-stop and then resume at some arbitrary time), as in Delporte-Gallet et al. \[12\]. In the presence of transient faults, we assume that failing nodes resume within some unknown finite time. The latter assumption is needed only for recovering from transient faults; we bring more details in Section \[2.6\]. In Section \[6\] we discuss how to relax this assumption.

2.3.2 Transient faults

As already mentioned, we consider arbitrary violations of the assumptions according to which the system and the communication network were designed to operate. We refer to these violations and deviations as transient faults and assume that they can corrupt the system state arbitrarily (while keeping the program code intact). The occurrence of a transient fault is rare. Thus, we assume that transient faults occur before the system execution starts \[14\]. Moreover, it leaves the system to start in an arbitrary state.

2.4 The snapshot object task

The task of snapshot object emulation requires the fulfillment of two properties: termination and linearizability. The definition of these two terms is based on the term event, which we defined next before defining the term event histories that is needed for the definition of linearizability.

Events: Let \( \text{op} \) be a \text{write()} or \text{snapshot()} operation. The execution of an operation \( \text{op} \) by a processor \( p_i \) is modeled by two steps: the invocation step, denoted by \( \text{invoc}(\text{op}) \), which calls the \( \text{op} \) operation, and a response event, denoted \( \text{resp}(\text{op}) \) (termination), which occurs when \( p_i \) terminates (completes) the operation. For the sake of simple presentation, by \text{event} we refer to either an operation’s start step or an operation’s end step.

Effective operations: We say that a \text{snapshot()} operation is effective when the invoking processor does not fail during the operation’s execution. We say that a \text{write()} operation is effective when the invoking processor does not fail during its execution, or in case it does fail, the operation’s effect is returned by an effective snapshot operation.

Histories: a history is a sequence of operation start and end steps that are totally ordered. We consider histories to compare in an abstract way between two executions of the studied algorithms. Given any two events \( e \) and \( f \), \( e < f \) if \( e \) occurs before \( f \) in the corresponding history. A history is denoted by \( \hat{H} = (E, <) \), where \( E \) is the set of events. Given an infinite history \( \hat{H} \), we require that: (i) its first event is an invocation and (ii) each invocation is followed by its matching response event. If \( \hat{H} \) is finite, then \( \hat{H} \) might not contain the matching response event of the last invocation event.

Linearizable snapshot history: A snapshot-based history \( \hat{H} = (H, <) \) models a computation at the abstraction level at which the write and snapshot operations are invoked. It is linearizable if there is an equivalent sequential history \( \hat{H}_{\text{seq}} = (H, <_{\text{seq}}) \) in which the sequence of effective \text{write()} and \text{snapshot()} operations issued by the processes is such that:

1. Each effective operation appears as executed at a single point of the timeline between its invocation event and its response event, and

2. Each effective \text{snapshot()} operation returns an array \( \text{reg} \) such that: (i) \( \text{reg}[i] = (v, \bullet) \) if the operation \( \text{write}(v) \) by \( p_i \) appears previously in the sequence. (ii) Otherwise \( \text{reg}[i] = \bot \).
2.5 Dijkstra’s self-stabilization criterion

An algorithm is self-stabilizing with respect to the task of LE, when every (unbounded) execution \( R \) of the algorithm reaches within a bounded period a suffix \( R_{\text{legal}} \in LE \) that is legal. That is, Dijkstra \[13\] requires that \( \forall R : \exists R': R = R' \circ R_{\text{legal}} \land R_{\text{legal}} \in LE \), where the length of \( R' \) is the complexity measure, which we refer to as the recovery time (other calls it the stabilization time). We say that a system execution is fair when every step that is applicable infinitely often is executed infinitely often and fair communication is kept. Self-stabilizing algorithms often assume that \( R \) is a fair execution. Wait-free algorithms guarantee that non-failing operations always become (within a finite number of steps) complete even in the presence of benign failures. Note that fair executions do not consider fail-stop failures (that were not detected by the system whom then excluded these failing nodes from reentering the system, as in \[16\]). Therefore, we cannot demonstrate that an algorithm is wait-free by assuming that the system execution is always fair.

2.6 Self-stabilization in the presence of seldom fairness

As a variation of Dijkstra’s self-stabilization criterion, Dolev et al. \[19\] proposed design criteria in which (i) any execution \( R = R_{\text{recoveryPeriod}} \circ R' : R' \in LE \), which starts in an arbitrary system state and has a prefix (\( R_{\text{recoveryPeriod}} \)) that is fair, reaches a legitimate system state within a bounded prefix \( R_{\text{recoveryPeriod}} \). (Note that the legal suffix \( R' \) is not required to be fair.) Moreover, (ii) any execution \( R = R'' \circ R_{\text{globalReset}} \circ R''' \circ R_{\text{globalReset}} \circ \ldots : R''', R'''' \ldots \in LE \) in which the prefix of \( R \) is legal, and not necessarily fair but includes at most \( O(n \cdot z_{\text{max}}) \) write or snapshot operations, has a suffix, \( R_{\text{globalReset}} \circ R''' \circ R_{\text{globalReset}} \circ \ldots \), such that \( R_{\text{globalReset}} \) is required to be fair and bounded in length but might permit the violation of liveness requirements, i.e., a bounded number of operations might be aborted (as long as the safety requirement holds). Furthermore, \( R''' \) is legal and not necessarily fair but includes at least \( z_{\text{max}} \) write or snapshot operations before the system reaches another \( R_{\text{globalReset}} \). Since we can choose \( z_{\text{max}} \in \mathbb{Z}^+ \) to be a very large value, say \( 2^{64} \), and the occurrence of transient faults is rare, we refer to the proposed criteria as one for self-stabilizing systems that their executions fairness is unrequited except for seldom periods.

2.7 Complexity Measures

The main complexity measure of self-stabilizing systems is the time it takes the system to recover after the occurrence of a last transient fault. In detail, in the presence of seldom fairness this complexity measure considers the maximum of two values: (i) the maximum length of \( R_{\text{recoveryPeriod}} \), which is the period during which the system recovers after the occurrence of transient failures, and (ii) the maximum length of \( R_{\text{globalReset}} \). We consider systems that use bounded among of memory and thus as a secondary complexity measure we bound the memory that each node needs to have. However, the number of messages sent during an execution does not have immediate relevance in the context of self-stabilization, because self-stabilizing systems never stop sending messages \[14\] Chapter 3.3]. Next, we present the definitions, notations and assumptions related to the main complexity measure.

2.7.1 Message round-trips and iterations of self-stabilizing algorithms

The correctness proof depends on the nodes’ ability to exchange messages during the periods of recovery from transient faults. The proposed solution considers quorum-based communications
that follow the pattern of request-replay as well as gossip messages for which the algorithm
does not wait for any reply. The proof uses the notion of a message round-trip for the cases of
request-reply messages as well as the term of algorithm iteration.

We give a detailed definition of round-trips as follows. Let \( p_i \in \mathcal{P} \) be a node and \( p_j \in \mathcal{P} \setminus \{p_i\} \) be a network node. Suppose that immediately after state \( c \) node \( p_i \) sends a message \( m \) to \( p_j \), for which \( p_i \) awaits a reply. At state \( c' \), that follows state \( c \), node \( p_j \) receives message \( m \) and sends a reply message \( r_m \) to \( p_i \). Then, at state \( c'' \), that follows state \( c' \), node \( p_j \) receives \( p_i \)’s response, \( r_m \). In this case, we say that \( p_i \) has completed with \( p_j \) a round-trip of message \( m \).

Self-stabilizing algorithms cannot terminate their execution and stop sending messages [14].
Moreover, their code includes a do forever loop. Thus, we define a complete iteration of a self-stabilizing algorithm. Let \( N_i \) be the set of nodes with whom \( p_i \) completes a message round trip infinitely often in execution \( R \). Moreover, assume that node \( p_i \) sends a gossip message infinitely often to \( p_j \in \mathcal{P} \setminus \{p_i\} \) (regardless of the message payload). Suppose that immediately after the state \( c_{\text{begin}} \), node \( p_i \) takes a step that includes the execution of the first line of the do forever loop, and immediately after system state \( c_{\text{end}} \), it holds that: (i) \( p_i \) has completed the iteration it has started immediately after \( c_{\text{begin}} \) (regardless of whether it enters branches), (ii) every request-reply message \( m \) that \( p_i \) has sent to any node \( p_j \in \mathcal{P} \) during the iteration (that has started immediately after \( c_{\text{begin}} \)) has completed its round trip, and (iii) it includes the arrival of at least one gossip message from \( p_i \) to any non-failing \( p_j \in \mathcal{P} \setminus \{p_i\} \). In this case, we say that \( p_i \)’s iteration (with round-trips) starts at \( c_{\text{begin}} \) and ends at \( c_{\text{end}} \).

2.7.2 Asynchronous cycles

We measure the time between two system states in a fair execution by the number of (asynchronous) cycles between them. The definition of (asynchronous) cycles considers the term of complete iterations. The first (asynchronous) cycle (with round-trips) of a fair execution \( R = R'' \circ R''' \) is the shortest prefix \( R'' \) of \( R \), such that each non-failing node in the network executes at least one complete iteration in \( R'' \), where \( \circ \) is the concatenation operator (Section 2.2). The second cycle in execution \( R \) is the first cycle in execution \( R''' \), and so on.

Remark 2.1 For the sake of simple presentation of the correctness proof, we assume that any message that arrives in \( R \) without being transmitted in \( R \) does so within \( O(1) \) asynchronous cycles in \( R \).

2.8 External Building Blocks: Gossip and Quorum Services

We utilize the gossip service from [19], which guarantees the following: (a) every gossip message that the receiver delivers to its upper layer was indeed sent by the sender, and (b) such deliveries occur according to the communication fairness guarantees (Section 2.3.1). I.e., this gossip service does not guarantee reliability.

We consider a system in which the nodes behave according to the following terms of service. We assume that at any time, any node runs only at most one operation (that is, either a write() or a snapshot(); one at a time). These operations access the quorum sequentially, i.e., send one request at a time, by sending messages to all other nodes via a broadcast interface. The receivers of this message reply. For this request-reply behavior, the quorum-based communication functionality guarantees the following: (a) at least a quorum of nodes receive, deliver and acknowledge every message, (b) a (non-failing) sending node receives at least a majority of these replies or fulfill another return condition, e.g., arrival of a special message, and (c) immediately before returning from the quorum access, the sending-side of this service clears its state from information related
Algorithm 1: The non-self-stabilizing and non-blocking algorithm by Delporte-Gallet et al. [12] that emulates snapshot object; code for $p_i$

\begin{verbatim}
Definitions of $\leq$: For integers $t$ and $t'$; $\bullet, t \leq (\bullet, t') \iff t \leq t'$; For arrays $tab$ and $tab'$ of $\bullet, integer$:
\[ tab \leq tab' \iff \forall p \in P: tab[p] \leq tab'[p] \] Also, $a < b \equiv a \leq b \land a \neq b$;

local variables initialization (optional in the context of self-stabilization):
\begin{itemize}
    \item $ssn := 0$; $ts := 0$; /* snapshot, resp., write operation indices */
    \item $reg := \lfloor 0, \ldots, 1 \rfloor$; /* shared registers (⊥ is smaller than any possibly written value) */
\end{itemize}

macro $merge(Rec)$ for $p_k \in P$ do $reg[k] \leftarrow \max\{reg[k] \cup \{v[k], r \in Rec\})$;

operation write($v$) begin
\begin{itemize}
    \item $ts \leftarrow ts + 1$; $reg[i] \leftarrow (v, ts)$; let $lReg := reg$;
    \item repeat broadcast $WRITE(IReg)$; until $WRITEack(reg.J \geq lReg)$ received from a majority;
\end{itemize}

operation snapshot() begin
\begin{itemize}
    \item repeat broadcast $SNAPSHOT(reg, ssn)$; until $SNAPSHOTack(reg, ssn.J = ss)$ received from a majority;
    \item $merge(Rec)$ where $Rec$ is the set of $reg$ arrays received at line 14;
\end{itemize}

upon message $WRITE(regJ)$ arrival from $p_j$ begin
\begin{itemize}
    \item for $p_k \in P$ do $reg[k] \leftarrow \max\{reg[k], regJ[k]\}$;
    \item send $WRITEack(reg)$ to $p_j$;
\end{itemize}

upon message $SNAPSHOT(regJ, ssn)$ arrival from $p_j$ begin
\begin{itemize}
    \item for $p_k \in P$ do $reg[k] \leftarrow \max\{reg[k], regJ[k]\}$;
    \item send $SNAPSHOTack(reg, ssn)$ to $p_j$;
\end{itemize}
\end{verbatim}

to this quorum request. We use the above requirements in Corollary 2.1, its correctness proof can be found in [19].

Corollary 2.1 (Self-stabilizing gossip and quorum-based communications) Let $R$ be an (unbounded) execution of the algorithm that appears in [19, Algorithm 3] and satisfies the terms of service of the quorum-based communication functionalities. Suppose that $R$ is fair and its starting system state is arbitrary. Within $O(1)$ asynchronous cycles, $R$ reaches a suffix $R'$ in which (1) the gossip, and (2) the quorum-based communication functionalities are correct. (3) During $R'$, the gossip and quorum-based communication complete their operations correctly within $O(1)$ asynchronous cycles.

3 Background

For the sake of completeness, we review the solutions of Delporte-Gallet et al. [12].

3.1 The non-blocking algorithm by Delporte-Gallet et al.

The non-blocking solution to snapshot object emulation by Delporte-Gallet et al. [12, Algorithm 1] allows all write operations to terminate regardless of the invocation patterns of the other write or snapshot operations (as long as the invoking processors do not fail during the operation). However, for the case of snapshot operations, termination is guaranteed only if eventually the system execution reaches a period in which there are no concurrent write operations. Algorithm 1 presents Delporte-Gallet et al. [12, Algorithm 1]. That is, we have changed some of the notation
of Delporte-Gallet to fit the presentation style of this paper. Moreover, we use the broadcast primitive according to its definition in Section 2.8.

Local variables. The node state appears in lines 2 to 4 and automatic variables (which are allocated and deallocated automatically when program flow enters and leaves the variable’s scope) are defined using the let keyword, e.g., the variable prev (line 13). Also, when a message arrives, we use the parameter name xJ to refer to the arriving value for the message field x.

Processor \( p_i \) stores an array \( \text{reg} \) of \(|\mathcal{P}|\) elements (line 4), such that the \( k \)-th entry stores the most recent information about processor \( p_k \)’s object and \( \text{reg}[i] \) stores \( p_i \)’s actual object value. Every entry is a pair of the form \((v, ts)\), where the field \( v \) is a \( \nu \)-bits object value and \( ts \) is an unbounded integer that stores the object timestamp. The values of \( ts \) serve as the index of \( p_i \)’s write operations. Similarly, \( p_i \) maintains an index for the snapshot operations, \( \text{ssn} \) (sequence number). Algorithm 1 defines also the relation \( \preceq \) that compares \((v, ts)\) and \((v', ts')\) according to the write operation indices (line 1).

The write(\( v \)) operation. Algorithm 1’s write(\( v \)) operation appears in lines 6 to 10 (client-side) and lines 11 to 17 (server-side). The client-side operation write(\( v \)) stores the pair \((v, ts)\) in \( \text{reg}[i] \) (line 7), where \( p_i \) is the calling processor and \( ts \) is a unique operation index. The primitive broadcast sends to all the processors in \( \mathcal{P} \) the message WRITE about \( p_i \)’s local perception of \( \text{reg} \)’s value.

Upon the arrival of a WRITE message to \( p_i \) from \( p_j \) (line 18), the server-side code is run. Processor \( p_i \) updates \( \text{reg} \) according to the timestamps of the arriving values (line 19). Then, \( p_i \) replies to \( p_j \) with the message WRITEack (line 23), which includes \( p_i \)’s local perception of the system shared registers.

Getting back to the client-side, \( p_i \) repeatedly broadcasts the message WRITE to all processors in \( \mathcal{P} \) until it receives replies from a majority of them (line 8). Once that happens, it uses the arriving values for keeping \( \text{reg} \) up-to-date (line 9).

The snapshot(\( v \)) operation. Algorithm 1’s snapshot() operation appears in lines 11 to 17 (client-side) and lines 21 to 23 (server-side). Recall that Delporte-Gallet et al. [12, Algorithm 1] is non-blocking with respect to the snapshot operations as long as are no concurrent write operations. Thus, the client-side is written in the form of a repeat-until loop. Processor \( p_i \) tries to query the system for the most recent value of the shared registers. The success of such attempts depends on the above assumption. Therefore, before each such broadcast, \( p_i \) copies \( \text{reg} \)’s value to \( \text{prev} \) (line 13) and exits the repeat-until loop only when the updated value of \( \text{reg} \) indicates that there are no concurrent write operations.

Figure 2 depicts two examples of Algorithm 1’s execution. The upper drawing illustrates a write operation that is followed by a snapshot operation and then a second write operation. We use this example when comparing algorithms 2, 3 and 4. The lower drawing illustrates a case of an unbounded sequence of write operations that disrupts a snapshot operation, which does not terminate for an unbounded period.

3.2 The always-terminating algorithm by Delporte-Gallet et al.

Delporte-Gallet et al. [12, Algorithm 2] guarantee termination for any invocation pattern of write and snapshot operations, as long as the invoking processors do not fail during these operations. Its advantage over Delporte-Gallet et al. [12, Algorithm 1] is that it can deal with an infinite number of concurrent write operations. This is because it guarantees the non-blocking progress criterion for the snapshot operations. We present [12, Algorithm 2] in Algorithm 2 using the presentation style of this paper. We review Algorithm 2 while pointing out some key challenges that exist when considering the context of self-stabilization.
Figure 2: Examples of Algorithm 1’s executions. The upper drawing illustrates a case of a terminating snapshot operation (dashed line arrows) that occurs between two write operations (solid line arrows). The acknowledgments of these messages are arrows that start with circles and squares, respectively. The lower drawing illustrates a case in which every execution of line 77 occurs concurrently with write operations (regardless of whether the algorithm is self-stabilizing or not). Thus, snapshot operations cannot terminate.

**High-level overview.** Delporte-Gallet et al. [12, Algorithm 2] use a job-stealing scheme for allowing rapid termination of snapshot operations. Processor \( p_i \in P \) starts its `snapshot` operation by queueing this new task at all processors \( p_j \in P \). Once \( p_j \) receives \( p_i \)’s new task and when that task reaches the queue front, \( p_j \) starts the `baseSnapshot(s,t)` procedure, which is similar to Algorithm 1’s `snapshot()` operation. This joint participation in all snapshot operations makes sure that all processors are aware of all on-going snapshot operations.

This joint awareness allows the system processors to make sure that no write operation can stand in the way of on-going snapshot operations. To that end, the processors wait until the oldest snapshot operation terminates before proceeding with later operations. Specifically, they defer write operations that run concurrently with snapshot operations. This guarantees termination of snapshot operations via the interleaving and synchronization of snapshot and write operations.

**Detailed description.** Algorithm 2 extends Algorithm 1 in the sense that it uses all of Algorithm 1’s variables and two additional ones, which is a second operation index, \( sns \), and an array `repSnap`, which `snapshot()` operations use. The entry `repSnap[x,y]` holds the outcome of \( p_x \)’s \( y \)-th snapshot operation, where no explicit bound on the number of invocations of snapshot operations is given.

In the context of self-stabilization, the use of such unbounded variables is not possible. The reasons are that real-world systems have bounded size memory as well as the fact that a single transient fault can bring any counter to its near overflow value and fill up any finite capacity buffer. We discuss the way around this challenge in Section 5.

*The `write()` operation and the `baseWrite()` function.* Since `write(v)` operations are preemptible, \( p_i \) cannot always start immediately to write. Instead, \( p_i \) stores \( v \) in `writePending_i` together with a unique operation index (line 35). The algorithm then runs the write operation as a background task (line 29) using the `baseWrite()` function (lines 39 to 42).

*The `snapshot()` operation.* A call to `snapshot()` (line 37) causes \( p_i \) to reliably broadcast, via
Algorithm 2: The non-self-stabilizing and always-terminating algorithm by Delporte-Gallet et al. [12] that emulates snapshot object; code for $p_k$

```plaintext
local variables initialization: ssn := 0; ts := 0; /* snapshot, resp., write operation indices */
reg := [$\bot$, $\ldots$, $\bot$]; /* shared registers ($\bot$ is smaller than any possibly written value) */
foreach $k, s : repSnap[k, s] := \bot; /* stores $p_k$'s snapshot task result for index $s$ */
macro merge(Rec) for $p_k \in \mathcal{P}$ do $\text{reg}[k] \leftarrow \max(\{\text{reg}[k]\} \cup \{r[k] \mid r \in \text{Rec}\});$
do forever begin
  if (writePending $\neq \bot$) then $\text{baseWrite(writePending)}$; writePending $\leftarrow \bot$;
  if (there are messages SNAP() received and not yet processed) then
    let SNAP(source, sn) be the oldest of these messages;
  baseSnapshot(source, sn);
  wait until ($\text{repSnap}[source, sn] \neq \bot$);
operation write(v) begin
  ssns $\leftarrow$ ssns + 1; \text{reliableBroadcast} SNAP(i, ssns);
  wait until ($\text{repSnap}[i, ssns] \neq \bot$); return($\text{repSnap}[i, ssns]$);
function baseWrite(v) begin
  ts $\leftarrow$ ts + 1; reg[i] $\leftarrow$ (ts, v); let lReg := reg;
  \text{repeat} broadcast WRITE(lReg); until WRITEack(lReg) \geq lReg received from a majority;
  merge(Rec) \text{ where } Rec is the set of reg arrays received at line 41
function baseSnapshot(s, t) begin
  while $\text{repSnap}[s, t] = \bot$ do
    let prev := reg; ssns $\leftarrow$ ssns + 1;
    repeat
      broadcast SNAPSHOT(s, t, reg, ssns);
      until ($sJ = s, tJ = t, \star, ssnsJ = ssns$) received from a majority;
      merge(Rec) \text{ where } Rec is the set of reg arrays received at line 47
    if prev = reg then \text{reliableBroadcast} END(source, sn, prev);
upon message WRITE(regJ) arrival from $p_i$ begin
  for $p_k \in \mathcal{P}$ do $\text{reg}[k] \leftarrow \max_{\text{sn}} (\text{reg}[k], \text{reg}[k]);$
  send WRITEack(regJ) to $p_j$;
upon message SNAPSHOT(s, t, regJ, ssnsJ) arrival from $p_j$ begin
  for $p_k \in \mathcal{P}$ do $\text{reg}[k] \leftarrow \max_{\text{sn}} (\text{reg}[k], \text{reg}[k]);$
  send SNAPSHOTack(s, t, regJ, ssnsJ) to $p_j$;
upon message END(s, t, val) arrival from $p_j$ do $\text{repSnap}[s, t] \leftarrow$ val;
```

the primitive \text{reliableBroadcast}, a new \text{ssn} index in a \text{SNAP} to all processors in $\mathcal{P}$. Processor $p_i$ then places it as a background task (line 38). We note that for our proposed solutions we do not assume access to a reliable broadcast mechanism such as \text{reliableBroadcast}; see Section 5 for details and an alternative approach that uses safe registers instead of the \text{reliableBroadcast} primitive, which often has higher communication costs.

The \text{baseSnapshot()} function. This function essentially follows the \text{snapshot()} operation of Algorithm 1. That is, Algorithm 1’s snapshot repeat-until loops iterates until the retrieved \text{reg} vector equals to the one that was known prior to the last repeat-until iteration. Algorithm 1’s \text{baseSnapshot()} procedure returns after at least one snapshot process has terminated. In detail, processor $p_i$ stores in $\text{repSnap}[s, t]$, via a reliable broadcast of the \text{END} message, the result of the snapshot process (line 50 and 57).

Synchronization between the \text{baseWrite()} and \text{baseSnapshot()} functions. Algorithm 2 interleaves the background tasks in a do forever loop (lines 29 to 33). As long as there is an awaiting write task, processor $p_i$ runs the \text{baseWrite()} function (line 29). Also, if there is an awaiting snapshot task, processor $p_i$ selects the oldest task, $(\text{source}, \text{sn})$, and uses the \text{baseSnapshot(source, sn)}
Figure 3: Algorithm 2’s execution for the case depicted by the upper drawing of Figure 2. The drawing illustrates a case of a terminating snapshot operation (dashed line arrows) that occurs between two write operations (solid line arrows). The acknowledgments of these messages are arrows that start with circles and squares, respectively.

4 An Unbounded Self-stabilizing Non-blocking Algorithm

We propose Algorithm 3 as an elegant extension of Delporte-Gallet et al. [12, Algorithm 1]; we have simply added the boxed code lines to Algorithm 3. Algorithms 1 and 3 differ in their ability to deal with stale information that can appear in the system when starting in an arbitrary state. Note that we model the appearance of stale information as the result of transient faults and assume that they occur only before the system starts running.

4.1 Algorithm description

Our description refers to the values of variable $X$ at node $p_i$ as $X_i$, i.e., the variable name with a subscript that indicates the node identifier. Algorithm 3 considers the case in which any of $p_i$’s operation indices, $ssn_i$ and $ts_i$, is smaller than some other $ssn$ or $ts$ value, say, $ssn_m reg_i[t].ts$ or $reg_j[t].ts$, where $X_m$ appears in the $X$ field of some in-transit message.

For the case of corrupted $ssn$ values, $p_i$’s client-side simply ignores arriving message with $ssn$ values that do not match $ssn_i$ (line 77). For the sake of clarity of our proposal, we also remove periodically any stored snapshot replies that their $ssn$ fields are not equal to $ssn_i$.

For the case of corrupted $ts$ values, $p_i$’s do forever loop makes sure that $ts_i$ is not smaller than $reg_[t].ts$ (line 67) before gossiping to every processor $p_j \in \mathcal{P}$ its local copy of $p_j$’s shared register (line 68). Also, upon the arrival of such gossip messages, Algorithm 3 merges the arriving information with the local one (line 82). Moreover, when replies from write or snapshot messages arrive to $p_i$, it merges the arriving $ts$ value with the one in $ts_i$ (line 53).

On the presentation side, we clarify that the code lines 8 to 9 and lines 14 to 15 are equivalent to lines 3 to 5 and lines 10 and 12 of [12 Algorithm 1], respectively, because it is merely a more
Figure 4: Algorithm 3’s execution for the case depicted in the upper drawing of Figure 2. The drawing illustrates a case of a terminating snapshot operation (dashed line arrows) that occurs between two write operations (solid line arrows). The acknowledgments of these messages are arrows that start with circles and squares, respectively.

detailed description of the code described in [12, Algorithm 1].

Figure 4 depicts an example of Algorithm 3’s execution in which a write operation is followed by a snapshot operation. Note that gossip messages do not interfere with write and snapshot operations.

4.2 Correctness

Although the extension performed to Algorithm 1 for obtaining Algorithm 3 includes only few changes, proving convergence and closure for Algorithm 3 is not straightforward. We proceed with the details.

Notation and definitions Definition 4.1 refers to $p_i$’s timestamps and snapshot sequence numbers, where $p_i \in P$. The set of $p_i$’s timestamps includes $ts_i$, $reg_i[ts_i]$, $reg_j[ts_j]$ in the payload of any message $m$ that is in transit in the system. The set of $p_i$’s snapshot sequence numbers includes $ssn_i$ and the value of $ssn_m$ in the payload of any message $m$ that is in transit in the system.

Definition 4.1 (Algorithm 3’s consistent operation indices) (i) Let $c$ be a system state in which $ts_i$ is greater than or equal to any $p_i$’s timestamp values in the variables and fields related to $ts$. We say that the $ts$’ timestamps are consistent in $c$. (ii) Let $c$ be a system state in which $ssn_i$ is greater than or equal to any $p_i$’s snapshot sequence numbers in the variables and fields related to $ssn$. We say that the $ssn$’s snapshot sequence numbers are consistent in $c$.

Theorems 4.1 and 4.11 show the properties required by the self-stabilization design criteria.

Theorem 4.1 (Algorithm 3’s convergence) Let $R$ be a fair and unbounded execution of Algorithm 3. Within $O(1)$ asynchronous cycles in $R$, the system reaches a state $c \in R$ in which $ts$’ timestamps and $ssn$’s snapshot sequence numbers are consistent in $c$.

Proof. The proof of the theorem follows by Lemmas 4.2 and 4.7.

Lemma 4.2 (Timestamp convergence) Let $R$ be an unbounded fair execution of Algorithm 3. Within $O(1)$ asynchronous cycles in $R$, the system reaches a state $c \in R$ in which the value of $ts_i$ is greater than or equal to any $p_i$’s timestamp value. Moreover, suppose that node $p_i$ takes a step immediately after $c$ that includes the execution of line 70. Then in $c$, it holds that $ts_i = reg_i[ts_i]$, $ts = reg_j[ts_j]$ as well as for every messages $m \in channel_{i,j}$, channel_{j,i} that is in transit from $p_i$ to $p_j$ or $p_j$ to $p_i$ it holds that $m.reg[ts_i] = ts_i$. 

15
**Algorithm 3:** Self-stabilizing algorithm for non-blocking snapshot object; code for $p_i$.

The boxed code lines mark the added code to Algorithm 1.

```plaintext
Definitions of $\preceq$: For integers $t, t'$: $(t, t') \preceq (t', t')$ $\iff$ $t \leq t'$; For arrays $\overline{tab}$ and $\overline{tab'}$ of $(\star, \text{integer})$: $\overline{tab} \preceq \overline{tab'} \iff \forall k \in P: \overline{tab}[k] \preceq \overline{tab'}[k]$; Also, $a < b \iff a \preceq b \land a \neq b$.

local variables initialization (optional in the context of self-stabilization):

- $ssn := 0$; $ts := 0$; $reg := \{1, \ldots, \|\|$ /* snapshot, resp., write operation indices */

**Proof of claim.**

Claim 4.3: The sequences $ts_{i, \ell}, reg_{i, \ell}[i], ts, reg_{i, \ell}[i], ts, reg_{i, \ell}[i]$ and $reg_{i, \ell}[i]$ are non-decreasing.

**Proof of claim.** We note that Algorithm 3 does only the following actions on $ts$ and $reg$ fields:
- increment (line 70) and merge using the max function (lines 63, 64, 67, 78, 82, 84 and 87). That is, there are no assignments. Thus, the claim is true, because the value of these fields is never decremented during $R$.

Claim 4.4: Within $O(1)$ asynchronous cycles, $ts_i \geq reg_i[i].ts$.

**Proof of claim.** Since $R$ is unbounded, it holds that node $p_j \in P$ calls line 68 for an unbounded number of times during $R$. Recall the line numbers that may change the value of $ts_i$ and $reg_i[i].ts$, cf. the proof of Claim 4.3. Note that only line 70 change the value of $ts_i$, via an increment (thus we do not have a simple equality) whereas lines 63, 67 and 82 update $ts_i$ and $reg_i[i].ts$ by taking
the maximum value of $ts_{i}$ and $reg_{i}[i].ts$. The rest of the proof is implied by Claim 4.3, and the
fact that $p_{i}$ executes line 67 at least once in every $O(1)$ asynchronous cycles. □

Algorithm 3 sends GOSSIP messages in line 68 requests messages in lines 71 and 77 as well as replies in lines 85 and 88. Claim 4.5’s proof considers lines 71 and 77 in which $p_{i}$ sends a request message to $p_{j}$, whereas Claim 4.6’s proof considers lines 68, 85 and 88 in which $p_{j}$ replies or gossips to $p_{i}$.

Claim 4.5 Let $m \in \text{channel}_{i,j}$ be a message on transit from $p_{i}$ to $p_{j}$ (during the first asynchronous cycles of $R$) and $reg_{m}$ the value of the reg filed in $m$, where $p_{i}, p_{j} \in \mathcal{P}$ are non-failing nodes. Within $O(1)$ asynchronous cycles, $reg_{i}[i].ts \geq reg_{m}[i].ts$ and $reg_{i}[i].ts \geq reg_{j}[i].ts$ whenever $p_{j}$ raises the events GOSSIP($reg_{J}$), WRITE($reg_{J}$) or SNAPSHOT($reg_{J}, \bullet$).

Proof of claim. Suppose during the first asynchronous cycles of $R$, node $p_{i}$ indeed sends message $m$, i.e., $m$ does not appear in $R$’s starting system state. Let $a_{k} \in R$ be the first step in $R$ in which $p_{i}$ calls line 68 or 71 and for which there is a step $a_{\text{depart},k} \in R$, which appears in $R$ after $a_{k}$ and in which message $m$ is sent (in a packet by the end-to-end or quorum protocol). Note that the value of $reg_{m}[i].ts$ in the message payload is defined by the value of $reg_{i}[i].ts$ in the system state that immediately precedes $a_{k}$. The rest of the proof relies on the fact that until $m$ arrives to $p_{j}$, the invariant $reg_{i}[i].ts \geq reg_{m}[i].ts$ holds (due to Claim 4.3).

Let $a_{\text{arrival},k} \in R$ be the first step that appears after $a_{\text{depart},k} \in R$, if there is any such step, in which the node at $p_{j}$ delivers the packet (token) that $a_{\text{depart},k}$ transmits the message $m$ (if there are several such packets, consider the last to arrive). By the correctness of the end-to-end [15, 17] or quorum service (Corollary 2.1), step $a_{\text{arrival},k}$ appears in $R$ within $O(1)$ asynchronous cycles. During $a_{\text{arrival},k}$, node $p_{j}$ raises the message delivery event GOSSIP($reg_{J}$) (when $a_{k}$ considers line 82), WRITE($reg_{J}$) (when $a_{k}$ considers line 83) or SNAPSHOT($reg_{J}, ssn$) (when $a_{k}$ considers line 84), such that $reg_{i}[i].ts \geq reg_{m}[i].ts = reg_{J}[i].ts$.

Suppose that step $a_{k}$ does not appear in $R$, i.e., $m$ appears in $R$’s starting system state. By the definition of asynchronous rounds with round-trips (Remark 2.1), within $O(1)$ asynchronous cycles, all messages in transit to $p_{j}$ arrive (or leave the communication channel). Immediately after that, the system starts an execution in which this claim holds trivially. □

Claim 4.6 Let $m’ \in \text{channel}_{i,k}$ be a message on transit from $p_{i}$ to $p_{k}$ (during the first asynchronous cycles of $R$) and $reg_{m}$ the value of the reg filed in $m'$, where $p_{i}, p_{j}, p_{k} \in \mathcal{P}$ are non-failing nodes and $i = k$ may or may not hold. Within $O(1)$ asynchronous cycles, $reg_{i}[i].ts \geq reg_{m’}[i].ts$ and $reg_{i}[i].ts \geq reg_{j}[i].ts$ whenever $p_{k}$ raises the events GOSSIP($reg_{J}$), WRITEACK($reg_{J}$) or SNAPSHOTack($reg_{J}, \bullet$).

Proof of claim. Suppose during the first asynchronous cycles of $R$, node $p_{k}$ indeed sends message $m’$, i.e., $m’$ does not appear in $R$’s starting system state. Let $a_{k} \in R$ be the first step in $R$ in which $p_{k}$ calls line 68, 85 or 88 and for which there is a step $a_{\text{depart},k} \in R$, which appears in $R$ after $a_{k}$. Note that the value of $reg_{m’}[i].ts$ in the message payload is defined by the value of $reg_{i}[i].ts$ in the system state that immediately precedes $a_{k}$. The rest of the proof relies on the fact that until $m’$ arrives to $p_{j}$, the invariant $reg_{i}[i].ts \geq reg_{m’}[i].ts$ holds (due to Claim 4.3).

Let $a_{\text{arrival},k} \in R$ be the first step that appears after $a_{\text{depart},k} \in R$, if there is any such step, in which the node at $p_{j}$ delivers the packet (token) that $a_{\text{depart},k}$ transmits the message $m$ (if there are several such packets, consider the last to arrive). By the correctness of the gossip and quorum services (Corollary 2.1), step $a_{\text{arrival},k}$ appears in $R$ within $O(1)$ asynchronous cycles. During $a_{\text{arrival},k}$, node $p_{j}$ raises the message delivery event GOSSIP($reg_{J}$) (when $a_{k}$ considers
Lemma 4.7 (Sequence number convergence) Let $R$ be a fair and unbounded execution of Algorithm 3. Within $O(1)$ asynchronous cycles in $R$, the system reaches a state $c_x \in R$ in which the value of $ssn_i$ is greater than or equal to any $p_i$’s snapshot sequence number.

Proof. Claims 4.8, 4.9 and 4.10 prove the lemma.

Claim 4.8 The sequence $ssn_{i,\ell}$ is non-decreasing.

Proof of claim. Algorithm 3 only increments (line 76), and assigns (lines 77 and 88) $ssn$ values. Thus, the claim is true, because the value of this field is never decremented during $R$. □

The proofs of Claims 4.9 and 4.10 are followed by similar arguments to the ones that appear in the proofs of Claims 4.5 and 4.6.

Claim 4.9 Let $m \in \text{channel}_{i,j}$ be a $\text{SNAPSHOT}$ message on transit from $p_i$ to $p_j$ (during the first asynchronous cycles of $R$) that includes the field $ssn$ with the value of $ssn_m$. Within $O(1)$ asynchronous cycles, $ssn_i \geq ssn_{m'}$ including when $p_j$ raises the event $\text{SNAPSHOT}(regJ, ssn_{m'})$.

Claim 4.10 Let $m' \in \text{channel}_{j,i}$ be a $\text{SNAPSHOTack}$ message on transit from $p_j$ to $p_i$ (during the first asynchronous cycles of $R$) and $ssn_{m'}$ the value of the reg field in $m'$. Within $O(1)$ asynchronous cycles, $ssn_i \geq ssn_m$ including when $p_j$ raises the event $\text{SNAPSHOTack}(regJ, ssn_{m'})$.

This completes the proof of the lemma, which completes the proof of the theorem. ■ ■

Theorem 4.11 (Algorithm 3’s termination and linearization) Let $R$ be an execution of Algorithm 3 that starts in system state $c$, in which the timestamps and snapshot sequence numbers are consistent (Definition 4.1). Execution $R$ is legal with respect to the task of emulating snapshot objects.

Proof. We start the proof by observing the differences between Algorithms 1 and 3. Note that Algorithms 1 and 3 use the same variables. Any message that Algorithm 1 sends, also Algorithm 3 sends. The only exception are gossip messages: Algorithm 3 sends gossip messages, while Algorithm 1 does not. The two algorithms differ in line 63, lines 66 to 68 and line 82.

The next step in the proof is to show that during $R$, any step that includes the execution of line 82 does not change the state of the calling processor. This is due to the fact the every timestamp uniquely couples an object value (line 67) and that timestamps are consistent in every system state throughout $R$ (Lemma 4.2).

The rest of the proof considers $\text{Alg}_{\text{noGOSSIP}}$ that is obtained from the code of Algorithm 3 by the removal of lines 68 and 82 in which the gossip messages are sent and received, respectively. We use this definition to show that $\text{Alg}_{\text{noGOSSIP}}$ simulates Algorithm 1. This means that from the perspective of its external behavior (i.e., its requests, replies and failure events), any trace of $\text{Alg}_{\text{noGOSSIP}}$ has a trace of Algorithm 1 (as long as indeed the starting system state, $c$, encodes consistent timestamps and snapshot sequence numbers). Since Algorithm 1 satisfies the task of emulating snapshot objects, it holds that $\text{Alg}_{\text{noGOSSIP}}$ also satisfies the task. This implies that Algorithm 3 satisfies the task as well.
Recall the fact that every timestamp uniquely couples an object value (line 67) as well as that timestamps and snapshot sequence numbers are consistent in every system state throughout $R$ (Lemma 4.2). These facts imply that also line 63 and lines 66 to 67 do not change the state of the calling node.

5 An Unbounded Self-stabilizing Always Terminating Algorithm

We propose Algorithm 4 as a variation of Delporte-Gallet et al. [12, Algorithm 2]. Algorithms 2 and 4 differ mainly in their ability to recover from transient faults. This implies some constraints. For example, Algorithm 4 must have a clear bound on the number of pending snapshot tasks as well as on the number of stored results from snapshot tasks that have already terminated (see Section 3.2 for details). For lack of simple presentation, Algorithm 4 assumes that the system needs, for each processor, to cater for at most one pending snapshot task. It turns out that this assumption allows us to avoid the use of a self-stabilizing mechanism for reliable broadcast, as an extension of the non-self-stabilizing reliable broadcast that Delporte-Gallet et al. [12, Algorithm 2] use. Instead, Algorithm 4 uses a simpler mechanism for safe registers.

The above opens up another opportunity: Algorithm 4 can defer pending snapshot tasks until either (i) at least one processor was able to observe at least $\delta$ concurrent write operations, where $\delta$ is an input parameter, or (ii) no $\delta$ concurrent write operations were observed, i.e., $\Delta = \emptyset$ (line 92). Our intention here is to have $\delta$ as a tunable parameter that balances the latency (with respect to snapshot operations) vs. communication costs. That is, for the case of $\delta$ being a very high (finite) value, Algorithm 4 guarantees termination in a way that resembles [12, Algorithm 1], which uses $O(n)$ messages per snapshot operation, and for the case of $\delta = 0$, Algorithm 4 behaves in a way that resembles [12, Algorithm 2], which uses $O(n^2)$ messages per snapshot operation.

5.1 High-level description

Algorithms 2 uses reliable broadcasts for informing all non-failing processors about new snapshot tasks (line 37) as well as the results of snapshot tasks that have terminated (line 50). Since we assume that each processor can have at most one pending snapshot task, we can avoid the need of using a self-stabilizing mechanism for reliable broadcast. Indeed, Algorithm 4 simply lets every processor disseminate its (at most one) pending snapshot task and use a safe register for facilitating the delivery of the task result to its initiator. That is, once a processor finishes a snapshot task, it broadcasts the result to all processors and waits for replies from a majority of processors, which may possibly include the initiator of the snapshot task (using the macro $\text{safeReg}(\cdot)$, line 93). This way, if processor $p_j$ notices that it has the result of an ongoing snapshot task, it sends that result to the requesting processor.

5.2 Algorithm details

We review Algorithm 4’s do forever loop (lines 96 to 102), the baseSnapshot() function together with the dealing of message SNAPSHOT (lines 124 to 125), as well as the macro $\text{safeReg}(s,r)$ (line 93) together with the dealing of message SAVE (lines 116 to 118).

The do forever loop. Algorithm 4’s do forever loop (lines 96 to 102), includes a number of lines for cleaning stale information, such as out-of-synch SNAPSHOTack messages (line 96), out-dated operation indices (line 97), illogical vector-clocks (line 98) or corrupted $pndTsk$ entries (line 99). The gossiping of operation indices (lines 100 and 119) also helps to remove stale information (as in Algorithm 4 but only with the addition of sns values).
The synchronization between write and snapshot operations (lines 101 and 102) starts with a write, if there is any such pending task (line 101), before running its own snapshot task, if there is any such pending, as well as any snapshot task (initiated by others) for which \( p_i \) observed that at least \( \delta \) write operations occur concurrently with it (line 102).

**The write() operation and the baseWrite() function.** As in Algorithm 2, \( p_i \) does not start immediately a write operation. Node \( p_i \) permits concurrent write operations by storing \( v \) and a unique index in \( writePending_i \) (line 103). The algorithm then runs the write operation as a background task (line 101) using the baseWrite() function (line 106).

**The baseSnapshot() function and the SNAPSHOT message.** Algorithm 4 maintains the state of every snapshot task in the array \( pndTsk \). The entry \( pndTsk_i[k] = (sns, vc, fnl) \) includes: (i) the index \( sns \) of the most recent snapshot operation that \( p_k \in \mathcal{P} \) has initiated and \( p_i \) is aware of, (ii) the vector clock representation of \( reg_k \) (i.e., just the timestamps of \( reg_k \), cf. line 91) and (iii) the final result \( fnl \) of the snapshot operation (or \( \bot \), in case it is still running).

The baseSnapshot() function includes an outer loop part (lines 109 and 115), an inner loop part (lines 109 to 112), and a result update part (lines 113 to 114). The outer loop increments the snapshot index, \( ssn \) (line 109), so that it can consider a new query attempt by the inner loop. The outer loop ends when (i) there are no more pending snapshot tasks that this call to baseSnapshot() needs to handle, or (ii) the only pending snapshot task for the current invocation of baseSnapshot() is the one of \( p_i \) and \( p_i \) has not observed at least \( \delta \) concurrent writes. The inner loop broadcasts SNAPSHOT messages, which includes all the pending \( (S \cap \Delta) \) that are relevant to this call to baseSnapshot() together with the local current value of \( reg \) and the snapshot query index \( ssn \). The inner loop ends when acknowledgments are received from a majority of processors and the received values are merged (line 112). The results are updated by writing to an emulated safe shared register (line 113) whenever \( prev = reg \). In case the results do not allow \( p_i \) to terminate its snapshot task (line 114). Algorithm 4 uses the query results for storing the timestamps in the field \( vs \). This allows to balance a trade-off between snapshot operation latency and communication costs, as we explain next.

**The use of the input parameter \( \delta \) for balancing the trade-off between snapshot operation latency and communication costs.** For the case of \( \delta = 0 \), the set \( \Delta \) (line 92) includes all the nodes for which there is no stored result, i.e., \( pndTsk[k].fnl = \bot \). Thus, no snapshot tasks are ever deferred, as in Delporte-Gallet et al. [12, Algorithm 2]. The case of \( \delta > 0 \) uses the fact that Algorithm 11 samples the vector clock value of \( reg \), and stores it in \( pndTsk[i].vc \) (line 114) once it had completed at least one iteration of the repeat-until loop (line 111 and 112). This way, we can be sure that the sampling of the vector clock is an event that occurred not before the start of \( p_i \)’s snapshot operation that has the index of \( pndTsk[i].sns \).

**Many-jobs-stealing scheme for reduced blocking periods.** We note that \( p_k \)’s task is considered active as long as \( pndTsk[k].fnl \neq \bot \). For helping all currently actives snapshot tasks, \( p_i \) samples the set of currently pending task \( (S \cap \Delta_i) \) (line 109) before starting the inner repeat-until loop (lines 109 to 112). Processor \( p_i \) broadcasts from the client-side the SNAPSHOT message, which includes the most recent snapshot task information, to all processors. The reception of this SNAPSHOT message on the server-side (lines 121 to 125) updates the local information (line 126) and prepares the response information (line 127) before sending the reply to the client-side (line 128). Note that if the receiver notices that it has the result of an ongoing snapshot task, it sends that result to the requesting processor (line 128).

**The safeReg() function and the SAVE message.** The safeReg() function considers a snapshot task that was initiated by processor \( p_k \in \mathcal{P} \). This function is responsible for storing the result \( r \) of this snapshot task in a safe register. It does so by broadcasting the client-side
Figure 5: The upper drawing depicts an example of Algorithm 4’s execution for a case that is equivalent to the one depicted in the upper drawing of Figure 3, i.e., only one snapshot operation. The lower drawing illustrates the case of concurrent invocations of snapshot operations by all nodes.

message SAVE to all processors in the system (line 93). Upon the arrival of the SAVE message to the server-side, the receiver stores the arriving information, as long as the arriving information is more recent than the local one. Then, the server-side replies with a SAVEack message to the client-side, who is waiting for a majority of such replies (line 93).

Figure 5 depicts two examples of Algorithm 4’s execution. In the upper drawing, a write operation is followed by a snapshot operation. Note that fewer messages are considered when comparing to Figure 3’s example. The lower drawing illustrates the case of concurrent invocations of snapshot operations by all nodes. Observe the potential improvement with respect to number of messages (in the upper drawing) and throughput (in the lower drawing) since Algorithm 2 uses $O(n^2)$ messages for each snapshot task and handles only one snapshot task at a time.

5.3 Correctness

We now prove the convergence (recovery), termination and linearization of Algorithm 4.

Definition 5.1 (Algorithm 4’s consistent system states and executions) (i) Let $c$ be a system state in which $ts_i$ is greater than or equal to any $p_i$’s timestamp values in the variables and fields related to $ts$. We say that the $ts$’ timestamps are consistent in $c$. (ii) Let $c$ be a system state in which $ssn_i$ is greater than or equal to any $p_i$’s snapshot sequence numbers in the variables and fields related to $ssn$. We say that the $ssn$’s snapshot sequence numbers are consistent in $c$. (iii) Let $c$ be a system state in which $sns_i$ is greater than or equal to any $p_i$’s snapshot operation index in the variables and fields related to $sns$. Moreover, $\forall p_i \in \mathcal{P} : sns_i = \text{pndTsk}_i[i].sns$ and $\forall p_i, p_j \in \mathcal{P} : \text{pndTsk}_i[i].sns \leq \text{pndTsk}_j[i].sns$. We say that the $sns$’s snapshot sequence numbers are consistent in $c$. (iv) Let $c$ be a system state in which $\forall p_i, p_k \in \mathcal{P} : \text{pndTsk}_i[k].vc \preceq VC_i$ holds, where $VC_i$ is the returned value from a macro defined in line 91 when executed by processor $p_i$. We say that the vector clock values are consistent in $c$. We say that system state $c$ is consistent if it is consistent with respect to invariants (i) to (iv). Let $R$ be an execution of Algorithm 4 that all of its system states are consistent and $R'$ be a suffix of $R$. We say that execution $R'$ is
Algorithm 4: Self-stabilizing always-terminating snapshot objects; code for $p_i$

Input: $\delta$ a number of observed concurrent writes after which write operations block temporarily;

Variables: $ts := 0$ is $p_i$’s write operation index; $ssn, sns := 0$ are $p_i$’s snapshot operation indices;

$\text{reg}[n] := \ldots \ldots \ldots \ldots$ buffers all shared registers; $\text{pndTsk}[n] := \ldots \ldots \ldots \ldots$ control variables of snapshot operations; each entry form is $(sns, vc, fnl)$, where $sns$ is an index, $vc$ is a vector clock that timestamps the snapshot operation $sns$, and $fnl$ is the operation’s returned value;

Macro $\text{VC} := (tsk)_n \in P$ where $tsk := 0$ when $\text{reg} = 1$ otherwise $\text{reg} = (\ast, tsk)$;

Macro $\Delta := \{(k, \text{pndTsk}[k], sns, \text{pndTsk}[k].vc)p \in P \land \text{pndTsk}[k].fnl = \bot \land ((\delta = 0 \land \text{pndTsk}[k], sns > 0) \lor (\text{pndTsk}[k].vc = 0 \land \Delta \leq \sum_{i \in \{1, \ldots, n\}} \text{VC}(i) - \text{pndTsk}[k].vc(i))) \lor \{(i, \text{pndTsk}[i], sns, \text{pndTsk}[i].vc) : \text{pndTsk}[i], sns > 0 \land \text{pndTsk}[i].fnl = \bot\}$;

Macro $\text{safeReg}(A)$ repeat broadcast $\text{SAVE}(A)$ until majority of $\text{SAVE}each(A) = \{(k, s) : (k, s, \ast) \in A\}$ arrived;

Macro $\text{merge}(\text{Rec}) \{ts \leftarrow \max(\text{ts}, \text{reg}[i].ts) \cup \{r[i]_r : r \in \text{Rec}\}; \text{for } p \in P \text{ do}\}$

redefine $\text{reg}[k] \leftarrow \max(\text{reg}[k] \cup \{r[k]_r : r \in \text{Rec}\});$

do forever begin

foreach $ssn \neq ssn$ do delete $\text{SNAPSHOT}(\ast, ssn);$;

$(ts, sns) \leftarrow (\max(ts, reg[i].ts), \max(sns, pndTsk[i].sns));$

for $k \in \{1, \ldots, n\}$ : $\text{pndTsk}[k].vc \notin \text{VC}, \text{where line 55 defines the relation} \prec \text{ do } \text{pndTsk}[k].vc \leftarrow \bot;$

if $\text{sns} \neq \text{pndTsk}[i].sns$ then $\text{pndTsk}[i] \leftarrow (\text{sns}, \bot);$

for $p \in \text{P}; k \neq i$ do send $\text{GOSSIP}(\text{reg}[k], \text{pndTsk}[k].sns)$ to $p$;

if $\text{writePending} \neq \bot$ then (baseWrite(\text{writePending}); writePending $\leftarrow \bot$);

if $\Delta \neq 0$ then baseSnapshot($\Delta$);

operation write(v) \{writePending $\leftarrow v$; wait until (writePending $\leftarrow \bot$); return $\bot$;\}

operation snapshot() begin

$(ssn, \text{pndTsk}[i]) \leftarrow (ssn + 1, (ssn, \bot, \bot));$ wait until (pndTsk[i].fnl $\neq \bot$); return(pndTsk[i].fnl);

function baseWrite(v) $(ts \leftarrow ts + 1; \text{reg}[i] \leftarrow (ts, v);$ let $\text{LC} := \text{reg};$ repeat broadcast $\text{WRITE}(\text{LC});$

merge(\text{Rec}) where $\text{Rec}$ is the received $\text{reg}$ arrays) until $\text{WRITE}each(\text{reg} \geq \text{LC})$ received from a majority;

function baseSnapshot($\Delta$) begin

repeat

$\text{ssn} \leftarrow \text{ssn} + 1;$ let $\text{prev} := \text{reg};$ repeat

broadcast $\text{SNAPSHOT}((S \Delta \Delta), \text{reg}[ssn]);$

until $(S \Delta \Delta) = \emptyset$ or majority of (SNAPSHOT)($\ast, ssn. ssn = ssn$) arrived;

merge(Rec) where $\text{Rec}$ is the set of $\text{reg}$ arrays received at line 101;

if $\text{prev} = \text{reg} \land (S \Delta \Delta) \neq \emptyset$ then $\text{safeReg}((k, \text{pndTsk}[k].sns, \text{prev} : (k, s, \ast) \in S));$

else if $(t, \ast) \in (S \Delta \Delta) \land (pndTsk[i].vc = \bot)$ then $\text{pndTsk}[i].vc \leftarrow \text{VC} =$

until $(S \Delta \Delta) = \emptyset$ or $(S \Delta \Delta) = (t, \ast) \land \text{pndTsk}[k], sns > 0 \land \text{pndTsk}[k].fnl = \bot \land \Delta \leq \sum_{i \in \{1, \ldots, n\}} (\text{VC}(i) - \text{pndTsk}[k].vc(i)))$;

upon message $\text{SAVE}(A)$ arrival from $p_j$ begin

foreach $(k, s, r) \in S : \text{pndTsk}[k], sns < s \lor \text{pndTsk}[k] = (s, \ast, \bot) \text{ do}$

$(\text{pndTsk}[k].sns, \text{pndTsk}[k].fnl) \leftarrow (s, r);$ wait until $\text{Reg}each(\{(k, s) : (k, s, \ast) \in A\})$ to $p_j$;

upon message $\text{GOSSIP}(\text{reg}, \text{sns})$ arrival from $p_j$ begin

$\text{reg}[i] \leftarrow \max(\text{reg}[i], \text{reg}[j]);$(ts, sns) $\leftarrow (\max(ts, \text{reg}[i].ts), \max(sns, \text{sns}));$

upon message $\text{WRITE}(\text{reg})$ arrival from $p_j$ begin

for $p \in P$ do $\text{reg}[k] \leftarrow \max_{\ast < s} (\text{reg}[k], \text{reg}[k]);$

send $\text{WRITE}each(\text{reg})$ to $p_j$;

upon message $\text{SNAPSHOT}(S_j, \text{reg}, \text{sns})$ arrival from $p_j$ begin

for $p \in P$ do $\text{reg}[k] \leftarrow \max_{\ast < s} (\text{reg}[k], \text{reg}[k]);$

foreach $(s, sns, vc) \in S_j : \text{pndTsk}[s], sns < sn \lor \text{pndTsk}[s] = (sn, \bot, \bot) \text{ do}$

$\text{pndTsk}[s] \leftarrow (sn, vc, \bot);$ let $A := \{(k, \text{pndTsk}[k], sns, \text{pndTsk}[k].fnl) : (k, \ast) \in S \land \text{pndTsk}[k].fnl \neq \bot\};$

send $\text{SNAPSHOT}(\text{reg}, \text{sns});$ to $p_j$; if $A \neq \emptyset$ then $\text{SAFE}(A)$ to $p_j$ (* piggyback messages *);

consistent (with respect to $R$) if any message arriving in $R'$ was indeed sent in $R$ and any reply arriving in $R'$ has a matching request in $R$.

Theorem 5.1 (Algorithm 4’s convergence) Let $R$ be a fair and unbounded execution of Algorithm 3. Within $O(1)$ asynchronous cycles in $R$, the system reaches a consistent state $c \in R$ (Definition 5.1). Within $O(1)$ asynchronous cycles after $c$, the system starts a consistent execution $R'$.

Proof. Note that Lemmas 4.2 and 4.7 imply invariants (i), and respectively, (ii) of Definition 5.1.
also for the case of Algorithm 4 because they use the similar code lines for asserting these invariants.

We now consider the proof of invariant (iii) of Definition 5.1. Note that the variables and fields of sns and the data structure pndTsk in Algorithm 4 follow the same patterns of information as the variables and fields of ts and the data structure reg in Algorithm 3. Moreover, within one asynchronous cycle, every processor \( p_i \in P \) executes line 99 at least once. Therefore, the proof of invariant (iii) can follow similar arguments to the ones appearing in the proof of Lemma 4.2. Specifically, \( \forall p_i, p_j \in P : pndTsk_i[i].sns \leq pndTsk_i[i].sns \) holds due to arguments that appear in the proof of Claim 4.3 with respect to the variables and the fields of ts and the structure reg.

The proof of invariant (iv) is implied by the fact that within one asynchronous cycle, every processor \( p_i \in P \) executes line 108 at least once and the fact that VC_i is assigned to pndTsk_i[k].vc in line 114. Note that these are the only lines of code that assign values to pndTsk_i[k].vc and that value of every entry in VC_i is not decreasing (Claim 4.3).

By the definition of asynchronous cycles (Section 2.7.2), within one asynchronous cycle, \( R \) reaches a suffix \( R' \), such that every received message during \( R' \) was sent during \( R \). By repeating the previous argument, it holds that within \( O(1) \) asynchronous cycles, \( R \) reaches a suffix \( R' \) in which for every received reply message, we have that its associated request message was sent during \( R \). Thus, \( R' \) is consistent.

The proof of Theorem 5.2 considers both complete and not complete snapshot() operations. We say that a snapshot() operation is complete if it starts due to a step \( a_i \) in which \( p_i \) calls the snapshot() operation (line 104) and its operation index, \( s \), is greater than any of \( p_i \)'s snapshot indices in the system state that appears immediately before \( a_i \). Otherwise, was say that it is not complete.

**Theorem 5.2 (Algorithm 4's termination and linearization)** Let \( R \) be a consistent execution (as defined by Definition 5.1) with respect to some execution of Algorithm 4. Suppose that there exists \( p_i \in P \), such that in \( R \)'s second system state (which immediately follows \( R \)'s first step that may include a call to the snapshot() operation in line 104) it holds that pndTsk_i[i] = (s, •, ⊥) and \( s > 0 \). Within \( O(\delta) \) asynchronous cycles, the system reaches a state \( c \in R \) in which pndTsk_i[i] = (s, •, x) : x ̸= ⊥.

**Proof.** Lemmas 5.3, 5.7, and 5.10 prove the theorem. These lemmas use the function \( S_i() \) that we define next. Whenever \( p_i \)'s program counter is outside of the function baseSnapshot(), the \( S_i() \) function returns the value of \( \Delta_i \). Otherwise, the function returns the value of \( (S_i \cap \Delta_i) \).

**Lemma 5.3 (Algorithm 4's termination — part 1)** Let \( R \) be a consistent execution (Definition 5.1) with respect to some execution of Algorithm 4. Suppose that there exists \( p_i \in P \), such that in \( R \)'s second system state (which immediately follows \( R \)'s first step that may include a call to the snapshot() operation in line 104) it holds that pndTsk_i[i] = (s, •, ⊥) and \( s > 0 \). Within \( O(\delta) \) asynchronous cycles, the system reaches a state \( c \in R \) in which either: (i) for any non-failing processor \( p_j \in P \) it holds that \( (i, •) \in S_j() \) (line 92) and pndTsk_j[i] = (s, •, ⊥), (ii) any majority \( M \subseteq P : |M| > |P|/2 \) include at least one \( p_j \in M \), such that pndTsk_j[i] = (s, •, x) : x ̸= ⊥ or (iii) pndTsk_i[i] = (s, •, x) : x ̸= ⊥.

**Proof.** Towards a proof in the way of contradiction, suppose that the lemma is false. That is, \( R \) has a prefix \( R' \) that includes \( O(\delta) \) asynchronous cycles, such that none of the lemma invariants hold during \( R' \). The proof uses claims 5.4 and 5.5 for demonstrating a contradiction with the above assumption in Claim 5.6.
Claim 5.4 \( R' \) does not include a step in which processor \( p_i \) evaluates the if-statement condition in line 113 to be true (or at least one of the lemma invariants holds).

**Proof of claim.** Arguments (1), (2) and (3) show that during \( a_i \in R' \) processor \( p_i \) calls the function \( \text{safeReg} \{ \{ (k, \text{pndTsk})[k], \text{sns}, \text{prev} \} : (k, s, \bullet) \in S \} \). Argument (4) shows that this implies that invariant (ii) holds. Thus, we reached a contradiction with the assumption in the lemma proof.

Argument (1): a call to \( \text{baseWrite()} \) ends within \( \mathcal{O}(1) \) asynchronous cycles.

A call to \( \text{baseWrite}(v) \) starts with \( p_i \) incrementing \( ts_i \) and sorting it in \( \text{reg}_i[i] \) (line 106) to a value that is unique (in the system state that immediately follows) with respect to \( ts \)'s variables and fields that are associated with \( p_i \) (Theorem 4.2). We note that the repeat-until loop in line 109 terminates due to the correctness of the quorum service (Corollary 2.1) and the uniqueness of \( ts_i \)'s value.

Argument (2): the repeat-until loop in lines 109 to 112 ends within \( \mathcal{O}(1) \) asynchronous cycles.

The call to \( \text{baseSnapshot}(S_i) : (i, \bullet) \in S_i \) starts with \( p_i \) incrementing \( ssn_i \) (line 109) to a value that is unique (in the system state that immediately follows) with respect to \( ssn \)'s variables and fields that are associated with \( p_i \) (Theorem 4.2). We note that the repeat-until loop in line 111 terminates due to the correctness of the quorum service (Corollary 2.1) and the uniqueness of \( ssn_i \)'s value (or the fact that \( S_i() = \emptyset \), which implies the lemma since then invariant (iii) holds).

Argument (3) consider a call that \( p_i \) performs to \( \text{baseSnapshot()} \) with the parameter \( S_i \).

Argument (3): showing that within \( \mathcal{O}(1) \) asynchronous cycles, \( p_i \in \mathcal{P} \) executes \( \text{baseSnapshot}_i(S_i) : (i, \bullet) \in S_i \), where it takes a step \( a_i \) that includes the execution of the if-statement in line 113.

The assumption that invariant (iii) does not hold in \( R' \) implies that \( (i, \bullet) \in S_i \) whenever processor \( p_i \) takes a step that includes the execution of \( \text{baseSnapshot}_i(S_i) \) or line 102, which is part of Algorithm 4’s do forever loop. The latter occurs within \( \mathcal{O}(1) \) asynchronous cycles (due to Argument (1) of this claim) and it includes the call to \( \text{baseSnapshot}_i() \) (line 102). Thus, the execution of line 113 is implied by the fact that the repeat-until loop in lines 109 to 112 eventually ends due to Argument (2) of this claim.

Argument (4): showing that invariant (ii) holds.

The function \( \text{safeReg}_i() \), which \( p_i \) calls in line 93 repeatedly sends to all processors the message \( \text{SAVE} \) until \( p_i \) receives matching \( \text{SAVEack} \) messages from a majority of processors. Theorem 5.1 and the assumption that \( R' \) is consistent imply that every received \( \text{SAVEack} \) message can be associated with a matching \( \text{SAVE} \) message that was indeed sent during \( R \). Thus, the rest of the proof shows that the existence of this majority of acknowledgments from processors \( p_j \in \mathcal{P} \) implies that invariant (ii) holds (due to the intersection property of majority groups). According to lines 116 to 118 the arrival of the message \( \text{SAVE} \) to \( p_j \in \mathcal{P} \) assures that \( \text{pndTsk}_{j[i]}[i].\text{fnl} \neq \bot \) before sending the message \( \text{SAVEack} \) back to \( p_i \). This is due to Theorem 5.1 and the assumption that \( R \) is consistent.


Claim 5.5 Within \( \mathcal{O}(1) \) asynchronous cycles, the system reaches a state \( c' \in R' \) in which for any non-faulty processor \( p_j \in \mathcal{P} \) it holds that \( \text{pndTsk}_{j[i]}[i] = (s, y, \bullet) : y \neq \bot \) (or at least one of the lemma invariants holds).

**Proof of claim.** We first consider the case of \( j = i \) before considering the case of \( j \neq i \).

The \( j = i \) case. Within \( \mathcal{O}(1) \) asynchronous cycles, \( p_i \) calls \( \text{baseSnapshot}(S_i) : (i, \bullet) \in S_i \) (line 102) due to Argument (3) in Claim 5.4. This, Argument (2) in Claim 5.4 and Claim 5.4 imply the execution of line 114 in every call for \( \text{baseSnapshot}(S_i) \). Hence, the claim for the case of \( j = i \).
The \( j \neq i \) case. By the arguments of the case of \( j = i \), within two asynchronous cycles, processor \( p_i \) executes lines 109 and 110 in which \( p_i \) broadcasts the record \((i, \text{pndTsk}_i[i], \text{sns}, \text{pndTsk}_i[i], \text{vc}) \in S'\) to all processors in the system via \text{SNAPSHOT}(S', \bullet) messages. Note that \( \text{pndTsk}_i[i], \text{vc} \neq \perp \) holds by the above case of \( j = i \). Moreover, once processor \( p_j \) receives this \text{SNAPSHOT} message, \( \text{pndTsk}_j[i], \text{vc} \neq \perp \) holds (line 126). The above arguments for the case of \( j \neq i \) can be repeated as long as invariant (iii) does not hold. Thus, the arrival of such a \text{SNAPSHOT} message to all \( p_j \in P \) occurs within \( O(1) \) asynchronous cycles (or one of the lemma invariants holds).

Claim 5.6 Let \( c' \in R' \) be a system state in which for any non-faulty processor \( p_j \in P \) it holds that \( \text{pndTsk}_j[i] = (s,y,\bullet) : y \neq \perp \) (as Claim 5.5 showed existence). Let \( x \) be the (finite or infinite) number of iterations of Algorithm 4’s outer loop in \text{baseSnapshot()} function (lines 109 and 115) that processor \( p_i \) takes between \( c' \) and \( c'' \in R' \), where \( c'' \) is a system state after which it takes at most \( O(\delta) \) asynchronous cycles until the system reach the state \( c''' \) in which at least one of the lemma invariants holds. The value of \( x \) is actually finite and \( x \leq \delta \).

Proof of claim. Arguments (1) to (3) show that \( x \leq \delta \). Moreover, between \( c'' \) and \( c''' \) there are \( O(\delta) \) asynchronous cycles.

Argument (1): as long as none of the proof invariants hold, whenever processor \( p_i \) iterates over the outer loop in \text{baseSnapshot()} function (lines 109 and 115), \( p_i \) takes a step in which it tests the if-statement condition at line 113 and that condition does not hold.

Within \( O(1) \) asynchronous cycle, \( p_i \) takes a step that includes a call to \text{baseSnapshot}(S_i) \: (i,\bullet) \in S_i \) (line 102) at least once (Argument (3) in Claim 5.4). By Claim 5.4, that call includes the execution of line 113 in which the if-statement condition does not hold (because then Argument (4) in Claim 5.4 implies that invariant (ii) holds).

Argument (2): suppose that there are at least \( x \) consecutive and complete iterations of \( p_i \)'s outer loop in the \text{baseSnapshot()} function (lines 109 and 113) between \( c' \) and \( c'' \) in which the if-statement condition at line 113 does not hold. There are at least \( x \) write operations that run concurrently with the \text{SNAPSHOT} operation that has the index of \( s \).

The only way that the if-statement condition in line 113 does not hold in a repeated manner is by repeated changes of \( ts \) field values in \text{regi}, during the different executions of lines 109 to 112. Such changes can only happen due to increments of \( ts_j : p_j \in P \) (line 103) at the start of \text{write()} operations.

Argument (3): there exists \( x' \leq \delta \) for which \((i,\bullet) \in S_i() \) (or at least one of the lemma invariants hold), where \( x' \) is the number of consecutive and complete iterations of the outer loop in the \text{baseSnapshot()} function (lines 109 and 113) between \( c' \) and \( c'' \) in which the if-statement condition at line 113 does not hold.

Argument (2) implies that the number of iterations continues to grow (as long as none of the lemma invariants holds). The proof of Argument (2) and Claim 4.3 imply that during every such iteration there are increments of at least one of the summation \( \sum_{\ell \in \{1,...,n\}} \text{VC}_i[\ell] - \text{pndTsk}_i[i].\text{vc}[\ell] \) until that summation is at least \( \delta \). Recall that \( \text{pndTsk}_i[i].\text{vc} \neq \perp \) (Claim 5.5) and \( \text{pndTsk}_i[i].\text{fnd} = \perp \) (the assumption that none of the lemma invariants hold in \( R' \)). Thus, \((i,\bullet) \in S_i() \) holds (line 92 for the case of \( k = i \)).

Argument (4): suppose that \( p_i \) has taken at least \( x' \) iterations of the outer loop in \text{baseSnapshot()} function (lines 109 and 115) after system state \( c' \) (which is defined in Claim 5.3). After those \( x' \) iterations, suppose that the system has reached a state \( c''' \) in which \((i,\bullet) \in S_i() \), as in Argument (3). Within \( O(1) \) asynchronous cycles after \( c'' \), the system reaches the state \( c''' \) in
which \((i, \bullet) \in S_i()\) holds for any non-failing processor \(p_j \in \mathcal{P}\) (or at least one of the two other lemma invariants holds).

Claim 5.5 states that in \(c'\) it holds for any non-faulty processor \(p_i \in \mathcal{P}\) that \(pndTsk_j[i] = (s, y, \bullet) : y \neq \bot\) (or at least one of the lemma invariants holds) and \(y \leq VC_j\) (otherwise, \(R\) is not a consistent execution). Within \(O(1)\) asynchronous cycles after \(c'\) (which Argument (3) defines), it holds that \(reg_j's ts\) fields are not smaller than the ones of \(reg_i's ts\) fields in \(c'\). This is because in every iteration of the outer loop in baseSnapshot() function (lines 109 and 115), processor \(p_i\) broadcasts \(reg_i\) to all processors (line 110). These SNAPSHOT messages arrive within one asynchronous cycle to all non-faulty processors \(p_j \in \mathcal{P}\) and upon their arrival \(p_j\) updates \(reg_j\) (lines 124 to 125). The rest of the proof shows that \((i, \bullet) \in S_i()\) holds (line 92 for the case of \(k = i\)); the reasons for that are similar to the ones that appear in the proof of Argument (3). □

This completes the proof of the lemma.

We note that invariants (i) and (i) of lemmas 5.3 and 5.7 match and invariant (iii) of Lemma 5.3 implies that the theorem holds.

**Lemma 5.7 (Algorithm 4’s termination — part II)** Let \(R\) be a consistent execution of Algorithm 4 (Definition 5.7) and \(p_i \in \mathcal{P}\). Moreover, suppose that either (i) in any system state of \(R\), it holds that \(pndTsk_j[i] = (s, \bullet, \bot)\), \(s > 0\) as well as for any non-failing processor \(p_i \in \mathcal{P}\) it holds that \((i, \bullet, \bot) \in S_i()\) (line 92) and \(pndTsk_j[i] = (s, \bullet, \bot)\), or (ii) in any system state of \(R\), it holds that \(pndTsk_j[i] = (s, \bullet, \bot)\), \(s > 0\) as well as for any majority \(M \subseteq \mathcal{P} : |M| > |\mathcal{P}|/2\) include at least one \(p_j \in M\), such that \(pndTsk_j[i] = (s, \bullet, x) : x \neq \bot\). Within \(O(1)\) asynchronous cycles, the system reaches a state \(c \in R\) in which \(pndTsk_j[i] = (s, \bullet, x) : x \neq \bot\).

**Proof.** The proof is implied by Claims 5.8 and 5.9.

Claim 5.8 Suppose that \(pndTsk_j[i].sns > 0\) holds in any system state of \(R\) and that for any majority \(M \subseteq \mathcal{P} : |M| > |\mathcal{P}|/2\) includes at least one \(p_j \in M\), such that \(pndTsk_j[i] = (s, \bullet, x) : x \neq \bot\). Within \(O(1)\) asynchronous cycles, the system reaches a state \(c \in R\) in which \(pndTsk_j[i] = (s, \bullet, x) : x \neq \bot\).

**Proof of claim.** Towards a proof in the way of contradiction, suppose that the lemma is false. That is, \(R\) has a prefix \(R'\) that includes at least \(O(1)\) asynchronous cycles, such that \(pndTsk[i] = (s, \bullet, x) : x = \bot\) holds in any system state in \(R'\). Arguments (1) to (2) show the needed contradiction. Recall that by Argument (3) in Claim 5.4, it holds that every iteration of the do forever loop during \(R'\) includes a call to baseSnapshot() at line 102.

Argument (1): within \(O(1)\) asynchronous cycles, a majority of nodes acknowledge the message baseSnapshot(\(\bullet, reg_i, ssn_i\)), such that for at least one baseSnapshot\(\bullet, reg_i, ssn_i\) acknowledgment, it holds that \(ssn.J = ssn_i\), \((s, pndTsk_j[i].sns, \bullet, x) \in AJ\) and \(\bot \neq x = pndTsk_j[i].fnl\).

We show that within \(O(1)\) asynchronous cycles, for at least one baseSnapshot\(\bullet, reg_i, ssn_i\) message, say the one from \(p_j\), it holds that \(ssn.J = ssn_i\), \((s, pndTsk_j[i].sns, \bullet, x) \in AJ\) and \(\bot \neq x = pndTsk_j[i].fnl\). This is followed from the fact that line 110 broadcasts repeatedly the SNAPSHOT\(\bullet, ssn_i\) message until at least a majority receives it and acknowledges it. By Argument (2) in Claim 5.4, the repeat-unti loop in lines 109 to 112 ends within \(O(1)\) asynchronous cycles. Moreover, by the proof of Argument (2) in Claim 5.4, the received acknowledgments indeed refer to these messages, and at least one of these acknowledgments includes \((s, pndTsk_j[i].sns, \bullet, x) \in AJ : \bot \neq x = pndTsk_j[i].fnl\) due to the claim assumption about \(M\).

Argument (2): within \(O(1)\) asynchronous cycles, \(pndTsk_j[i].fnl \neq \bot\) holds.
By the proof of Argument (1), a majority $M$ of processors successfully acknowledge the message SNAPSHOT (line 128). By the claim assumption and the intersection property of majority sets, for at least one of the acknowledging node, say, $p_k$, it holds that $\text{pndTsk}_k[i] = (s, \bullet, x) : x \neq \bot$. Since line 128 piggyback the SNAPSHOTack($\text{reg}, \text{ssn.J}$) and SAVE($A$) messages, the message SAVE($A$) successfully arrives from $p_k$ to $p_i$. The proof is done by the fact that once the message SAVE arrives to $p_i$, line 117 updates $\text{pndTsk}_i[i] = (s, \bullet, x) : x \neq \bot$.

\[\square\]

**Claim 5.9** Suppose that, during the first $O(1)$ asynchronous cycles of $R$, for any non-failing processor $p_j \in \mathcal{P}$ it holds that $(i, \bullet) \in S_j(\ell)$ (line 92) and $\text{pndTsk}_j[i] = (s, \bullet, \bot)$. Within $O(1)$ asynchronous cycles, the system reaches a state $c \in R$ in which $\text{pndTsk}_i[i] = (s, \bullet, x) : x \neq \bot$.

**Proof of claim.** The proof is by a sequence of statements, i.e., arguments (1) to (3).

Argument (1): *within* $O(1)$ asynchronous cycles, there are no active write operations.

Note that during $R$, any processor that executes the write() function, returns from this call to write() within $O(1)$ asynchronous cycles (Argument (1) in Claim 5.4). Thus, $p_i$ calls baseSnapshot($S_i$) with the parameter $S_i$, such that $(i, \bullet) \in S_i$ follows (Argument (3) in Claim 5.3). By this claim assumption that $\forall p_j \in \mathcal{P} : (i, \bullet) \in S_j(\ell)$, it holds that the repeat-until loop (lines 109 to 115) does not end during the first $O(1)$ asynchronous cycles of $R$ (due to the fact that the end condition cannot hold). By the same assumptions and similar arguments as above, the same holds for any $p_j \in \mathcal{P}$ within $O(1)$ asynchronous cycles. Thus, within $O(1)$ asynchronous cycles the system execution reaches a suffix, $R'$, during which there are no active write operations. This completes the proof of Argument (1.)

The rest of the proof shows that $p_i$’s snapshot operation terminates within the first $O(1)$ asynchronous cycles of $R'$. Towards a proof in the way of contradiction, suppose that the statement is false. That is, let $R''$ be a prefix of $R'$ (that includes at least $O(1)$ asynchronous cycles), in which $p_i$’s snapshot operation does not terminate. In other words, in every system state of $R''$, it holds that $\text{pndTsk}_i[i] = (s, \bullet, \bot)$.

By Argument (2) of Claim 5.4, the repeat-until loop in lines 110 to 112 terminates. By line 111 this happens only when (i) $(S \cap \Delta) \neq \emptyset$ or (ii) majority of matching SNAPSHOTack messages arrived. The former case implies that the proof is done, because the only way in which $(i, \bullet)$ leaves the set $(S \cap \Delta)$ is by having $\text{pndTsk}_k[i] = (s, \bullet, x) : x \neq \bot$ (due to the execution of line 117). For the latter case, we note that Argument (1) implies that the if-statement condition in line 113 holds. We complete the proof with Argument (2), which show the needed contradiction.

Argument (2): a call to safeReg($\cdot$) during $R''$ implies the claim.

Suppose that the if-statement condition in line 113 holds after the execution of lines 110 to 112. The call to safeReg ($\{(\bullet, r_i) : r_i \neq \bot\}$) causes $p_i$ to send the message SAVE ($\{(\bullet, r_i) : r_i \neq \bot\}$) to itself and the reception of this message assigns $r_i \neq \bot$ to $\text{pndTsk}_k[i], \text{fnl}$ (line 117). (Since there is no need to actually send this message, this is done within an atomic step.)

This completes the proof of the lemma.

\[\square\]

**Lemma 5.10 (Algorithm 4’s linearization)** Algorithm 4 respects the sequential specification of the snapshot object.

**Proof.** We note that the baseWrite() functions in Algorithms 2 and 4 are identical. Moreover, Algorithm 2’s lines 15 to 47 are similar to Algorithm 4’s lines 109 to 112 but differ in the following manner: (i) the dissemination of the operation tasks is done outside of Algorithm 2’s
lines 45 to 47 but inside of Algorithm 4’s lines 109 and (ii) Algorithm 2 considers one snapshot operation at a time whereas Algorithm 4 considers many snapshot operations.

The proof is based on observing that the definition of linearizability (Section 2.4) allows concurrent snapshot operations to have the same result (as long as they each individually respect all the other constraints that appear in the definition of linearizability). Moreover, by the same definition, the linearizability property does not depend on the way in which the snapshot tasks (and their results) are disseminated. (Indeed, the linearizability proof of Delporte-Gallet et al. [12, Lemma 7] does not consider the way in which the snapshot tasks, and their results, are disseminated when selecting linearization points. These linearization points are selected according to some partition, defined in [12, Lemma 7]. The proof there explicitly allows the same partition to include more than one snapshot result.)

This completes the proof of the theorem.

6 Bounded Variations on Algorithms 3 and 4

In this section, we discuss how we can obtain bounded variations of our two unbounded self-stabilization algorithms. Dolev et al. [19, Section 10] present a solution to a similar transformation: They show how to take a self-stabilizing atomic MWMR register algorithm for message passing systems that uses unbounded operation indices and transform it to an algorithm that uses bounded indices. We review the techniques that Dolev et al. use and explain how a similar transformation can also be used for Algorithms 3 and 4 with respect to their different operation indices.

The procedure by Dolev et al. [19, Section 10] considers operation indices, which they call tags, whereas the proposed algorithms refer to operation indices as (i) ts values in the variables and message fields, as well as (ii) ssn and sns values in the variables and message fields. That is, Dolev et al. consider just one type of operation index whereas we consider several types. For lack of simple presentation, when describing next the procedure by Dolev et al., we refer to the case of many types of operation indices.

1. Once node \( p_i \in \mathcal{P} \) stores an operation index that is at least MAXINT, node \( p_i \) disables the invocation of all operations (of all types) while allowing the completion of the existing ones (until all nodes agree on the highest index for each type of operation, cf. item 2), where \( \text{MAXINT} \in \mathbb{Z}^+ \) is a very large constant, say, \( \text{MAXINT} = 2^{64} - 1 \).

2. While the invocation of new operations (of all types) is disabled (by item 1), the gossip procedure keeps on propagating the maximal operation indices (and merge the arriving information with the local one). Eventually, all nodes share the same operation indices (for all types). At that point in time, the procedure for dealing with integer overflow events uses a consensus-based global reset procedure for replacing, per operation type, the highest operation index with its initial value 0, while keeping the values of all shared registers unchanged.

Self-stabilizing global reset procedure. The implementation of the self-stabilizing procedure for global reset can be based on existing mechanisms, such as the one by Awerbuch et al. [8]. We note that the system settings of Awerbuch et al. [8] assume execution fairness. This assumption is allowed by our system settings (Section 2.6). This is because we assume that reaching MAXINT can only occur due to a transient fault. Thus, execution fairness, which implies all nodes are eventually alive, is seldom required (only for recovering from transient faults).
An extension: quorum reconfiguration. We consider an extension of our system settings that include quorum reconfiguration. The advantage here is two folded: (i) systems that can reconfigure the set $\mathcal{P}$ are more durable since they can replace failing nodes with new ones, and (ii) they allow us to relax the assumption that failing node eventually restart (Section 2).

As an alternative approach for implementing the self-stabilizing procedure for global reset, we propose to base the reset procedure on a self-stabilizing consensus algorithm, e.g., [9], and quorum reconfiguration [16]. Note that the system settings of [9, 16] assume the availability of failure detector mechanisms, and the relevant liveness conditions for implementing these mechanisms. Moreover, quorum reconfiguration requires the use of state transfer procedure after every reconfiguration. In the context of the proposed solutions, the array $\text{reg}$ should be adjusted by adding new entries for every joining node and removing entries associated with nodes that are no longer part of the quorum configuration. The reconfiguration-based reset procedure is similar to the above procedure described in items 1 and 2. The only difference is that between these two steps, a quorum reconfiguration needs to be imposed to assure that all nodes in $\mathcal{P}$ are up and connected.

7 Conclusions

We showed how to transform the two non-self-stabilizing algorithms of Delporte-Gallet et al. [12] into ones that can recover after the occurrence of transient faults. This requires some non-trivial considerations that are imperative for self-stabilizing systems, such as the explicit use of bounded memory and the reoccurring clean-up of stale information. Interestingly, these considerations are not restrictive for the case of Delporte-Gallet et al. [12]. For our self-stabilizing atomic snapshot algorithm that always terminates, we chose to use safe registers for storing the results of recent snapshot operations, rather than a mechanism for reliable broadcast, which is more expensive to implement. Moreover, instead of dealing with one snapshot operation at a time, we deal with several at a time. In addition, we consider a tunable input parameter, $\delta$, for allowing the system to balance a trade-off between the latency of snapshot operations and communications costs, which range from $O(n)$ to $O(n^2)$ messages per snapshot operation.

One future direction emanating from this work is to consider the $O(n)$ gap in the number of messages when designing future applications. For example, one might prefer the use of repeated snapshots (using the proposed solution) over a replicated state machine, which always costs $O(n^2)$ messages per state transition. Another future direction is to consider the techniques presented here for providing self-stabilizing versions of more advanced snapshot algorithms, such as the one by Imbs et al. [24].

References

[1] Uri Abraham. Self stabilizing weak snapshots and timestamps. Technical report, Department of Computer Science, Ben-Gurion University of the Negev, September 2007. Accessible via https://www.cs.bgu.ac.il/~abraham/papers/cs/EfficientStableTS30.pdf

[2] Yehuda Afek, Hagit Attiya, Danny Dolev, Eli Gafni, Michael Merritt, and Nir Shavit. Atomic snapshots of shared memory. J. ACM, 40(4):873–890, 1993.

[3] Noga Alon, Hagit Attiya, Shlomi Dolev, Swan Dubois, Maria Potop-Butucaru, and Sébastien Tixeuil. Practically stabilizing SWMR atomic memory in message-passing systems. J. Comput. Syst. Sci., 81(4):692–701, 2015.
[4] James H. Anderson. Multi-writer composite registers. *Distributed Computing*, 7(4):175–195, 1994.

[5] Hagit Attiya. Efficient and robust sharing of memory in message-passing systems. *J. Algorithms*, 34(1):109–127, 2000.

[6] Hagit Attiya. Robust simulation of shared memory: 20 years after. *Bulletin of the EATCS*, 100:99–113, 2010.

[7] Hagit Attiya, Amotz Bar-Noy, and Danny Dolev. Sharing memory robustly in message-passing systems. *J. ACM*, 42(1):124–142, 1995.

[8] Baruch Awerbuch, Boaz Patt-Shamir, George Varghese, and Shlomi Dolev. Self-stabilization by local checking and global reset. In *WDAG*, volume 857 of *LNCS*, pages 326–339. Springer, 1994.

[9] Peva Blanchard, Shlomi Dolev, Joffroy Beauquier, and Sylvie Delaët. Practically self-stabilizing Paxos replicated state-machine. In *NETYS*, volume 8593 of *LNCS*, pages 99–121. Springer, 2014.

[10] Tushar Deepak Chandra and Sam Toueg. Unreliable failure detectors for reliable distributed systems. *J. ACM*, 43(2):225–267, 1996.

[11] Sylvie Delaët, Stéphane Devismes, Mikhail Nesterenko, and Sébastien Tixeuil. Snap-stabilization in message-passing systems. *J. Parallel Distrib. Comput.*, 70(12):1220–1230, 2010.

[12] Carole Delporte-Gallet, Hugues Fauconnier, Sergio Rajsbaum, and Michel Raynal. Implementing snapshot objects on top of crash-prone asynchronous message-passing systems. *IEEE Trans. Parallel Distrib. Syst.*, 29(9):2033–2045, 2018.

[13] Edsger W. Dijkstra. Self-stabilizing systems in spite of distributed control. *Commun. ACM*, 17(11):643–644, 1974.

[14] Shlomi Dolev. *Self-Stabilization*. MIT Press, 2000.

[15] Shlomi Dolev, Swan Dubois, Maria Potop-Butucaru, and Sébastien Tixeuil. Stabilizing datalink over non-fifo channels with optimal fault-resilience. *Inf. Process. Lett.*, 111(18):912–920, 2011.

[16] Shlomi Dolev, Chryssis Georgiou, Ioannis Marcoullis, and Elad Michael Schiller. Self-stabilizing reconfiguration. In *NETYS*, volume 10299 of *LNCS*, pages 51–68, 2017.

[17] Shlomi Dolev, Ariel Hanemann, Elad Michael Schiller, and Shantanu Sharma. Self-stabilizing end-to-end communication in (bounded capacity, omitting, duplicating and non-fifo) dynamic networks - (extended abstract). In *SSS*, volume 7596 of *LNCS*, pages 133–147. Springer, 2012.

[18] Shlomi Dolev, Amos Israeli, and Shlomo Moran. Resource bounds for self-stabilizing message-driven protocols. *SIAM J. Comput.*, 26(1):273–290, 1997.

[19] Shlomi Dolev, Thomas Petig, and Elad Michael Schiller. Self-stabilizing and private distributed shared atomic memory in seldomly fair message passing networks. *CoRR*, abs/1806.03498, 2018.
[20] Partha Dutta, Rachid Guerraoui, Ron R. Levy, and Marko Vukolic. Fast access to distributed atomic memory. *SIAM J. Comput.*, 39(8):3752–3783, 2010.

[21] Chryssis Georgiou and Alexander A. Shvartsman. *Cooperative Task-Oriented Computing: Algorithms and Complexity*. Synthesis Lectures on Distributed Computing Theory. Morgan & Claypool Publishers, 2011.

[22] Theophanis Hadjistasi, Nicolas C. Nicolaou, and Alexander A. Schwarzmann. Oh-ram! one and a half round atomic memory. In *NETYS*, volume 10299 of *LNCS*, pages 117–132, 2017.

[23] Maurice Herlihy and Jeannette M. Wing. Linearizability: A correctness condition for concurrent objects. *ACM Trans. Program. Lang. Syst.*, 12(3):463–492, 1990.

[24] Damien Imbs, Achour Mostéfaoui, Matthieu Perrin, and Michel Raynal. Set-constrained delivery broadcast: Definition, abstraction power, and computability limits. In 19th *Distributed Computing and Networking, ICDCN*, pages 7:1–7:10. ACM, 2018.

[25] Lefteris M. Kirousis, Paul G. Spirakis, and Philippas Tsigas. Reading many variables in one atomic operation: Solutions with linear or sublinear complexity. *IEEE Trans. Parallel Distrib. Syst.*, 5(7):688–696, 1994.

[26] Lefteris M. Kirousis, Paul G. Spirakis, and Philippas Tsigas. Simple atomic snapshots: A linear complexity solution with unbounded time-stamps. *Inf. Process. Lett.*, 58(1):47–53, 1996.

[27] Nancy A. Lynch. *Distributed Algorithms*. Morgan Kaufmann, 1996.

[28] Achour Mostéfaoui and Michel Raynal. Two-bit messages are sufficient to implement atomic read/write registers in crash-prone systems. In *PODC*, pages 381–389. ACM, 2016.

[29] Michel Raynal. *Communication and Agreement Abstractions for Fault-Tolerant Asynchronous Distributed Systems*. Synthesis Lectures on Distributed Computing Theory. Morgan & Claypool Publishers, 2010.

[30] Michel Raynal. *Distributed Algorithms for Message-Passing Systems*. Springer, 2013.

[31] Eric Ruppert. Implementing shared registers in asynchronous message-passing systems. In *Encyclopedia of Algorithms*, pages 954–958. Springer, 2016.

[32] Gadi Taubenfeld. *Distributed Computing Pearls*. Synthesis Lectures on Distributed Computing Theory. Morgan & Claypool Publishers, 2018.

[33] Jennifer L. Welch and Jennifer E. Walter. *Link Reversal Algorithms*. Synthesis Lectures on Distributed Computing Theory. Morgan & Claypool Publishers, 2011.