Computing with Reads and Writes in the Absence of Step Contention

(Extended Abstract)

Hagit Attiya[*] Rachid Guerraoui[†] Petr Kouznetsov[†] [‡]

[*]Department of Computer Science, Technion
[†]School of Computer and Communication Sciences, EPFL

Abstract

This paper studies implementations of concurrent objects that exploit the absence of *step contention*. These implementations use only reads and writes when a process is running solo. The other processes might be busy with other objects, swapped-out, failed, or simply delayed by a contention manager. We study in this paper two classes of such implementations, according to how they handle the case of step contention. The first kind, called *obstruction-free* implementations, are not required to terminate in that case. The second kind, called *solo-fast* implementations, terminate using powerful operations (e.g., C&S).

We present a generic obstruction-free object implementation that has a linear contention-free step complexity (number of reads and writes taken by a process running solo) and uses a linear number of read/write objects. We show that these complexities are asymptotically optimal, and hence generic obstruction-free implementations are inherently slow. We also prove that obstruction-free implementations cannot be *gracefully degrading*, namely, be nonblocking when the contention manager operates correctly, and remain (at least) obstruction-free when the contention manager misbehaves.

Finally, we show that any object has a *solo-fast* implementation, based on a solo-fast implementation of consensus. The implementation has linear contention-free step complexity, and we conjecture solo-fast implementations must have non-constant step complexity, i.e., they are also inherently slow.

Keywords: shared memory, contention, step contention, solo-fast, obstruction-free, lock-free.

[‡] **Contact author:** petr.kouznetsov@epfl.ch, EPFL IC IIF LPD, 313 (Bât. INR), Station 14, CH 1015, Switzerland, phone: +41 21 693 5274, fax: +41 21 6937570.

1 Introduction

At the heart of many distributed systems are *shared objects*—data structures that are concurrently accessed by many processes. Often, these objects are *implemented* in software, out of more elementary *base objects*. *Lock-free* implementations of such objects do not rely on mutual exclusion or locking, and thereby allow processes to overcome adverse operating systems affects. This includes both *wait-free* algorithms, in which *every* process completes its operations in a finite number of steps, and *nonblocking* algorithms, where *some* process completes an operation in every sufficiently long execution [16]. The safety property typically required from both nonblocking and wait-free implementations is *linearizability* [16, 19]; roughly, every operation on the object should appear instantaneous.

Although they provide very attractive guarantees, lock-free implementations were claimed to have limited usability. This is because nonblocking implementations of many objects are often impossible, e.g., when only read/write objects are available [11, 13, 24]. Even when the implementations are possible, which can be achieved under specific timing assumptions (e.g., encapsulated within failure detector abstractions), or using strong synchronization operations (like C&S), these implementations are typically complex and expensive [7, 10, 21]. The complexity and computability price paid by lock-free algorithms often originates in situations in which steps of concurrent processes are interleaved, i.e., when there is *step contention*.

In this paper, we study implementations that exploit the rarity of these situations: it is indeed often argued that, in practice, step contention is rare, or at least can be made so through operating system support. That is, only one process is typically performing visible (non local) steps within any object operation, whereas the rest of the processes are busy with other objects, swapped-out or failed. The absence of step contention does not preclude common scenarios where other processes have pending operations on the same implemented object but are not accessing the base objects. This is fundamentally different from alternative contention metrics: *point* contention [5] and *interval* contention [2]; both count also failed or swapped-out processes. (See the scenario presented in Figure 1.)

We first study *obstruction-free* implementations that guarantee termination only in the absence of step contention. This is formalized by the *solo termination* property [12]: a process that takes sufficiently many steps on its own returns a value. Clearly obstruction-free implementations cannot rely on mutual exclusion or locks, and hence, they are lock-free. Whereas all nonblocking implementations are obstruction-free; the converse is not necessarily true, however, since obstruction-free implementations may incur scenarios (when there is step contention) in which no process is able to

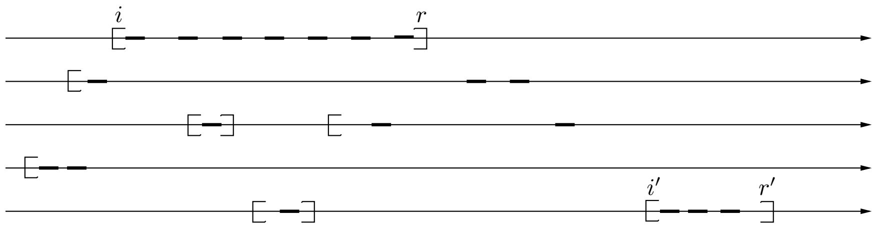

Figure 1: An example illustrating types of contention: Operation $[i, r]$ has interval contention 5, point contention 4, and step contention 3; operation $[i', r']$ has interval and point contention 4, and step contention 1 ($[i', r']$ is step contention-free). (Square brackets denote invocations and responses, while solid intervals denote steps on base objects.)

complete its operation in a finite number of steps.

An obstruction-free implementation has to provide a *legal* response if it returns at all, but termination is required only under very restricted conditions. One contribution of this paper is to disambiguate the behavior of an obstruction-free implementation when an operation cannot return a legal response. In the presence of step contention, an operation may return control to a higher-level entity, which we call the *client*. Ideally, the obstruction-free implementation should return a *fail* indication to the client, allowing it to either re-invoke the operation, or to invoke another operation. We show however that there is inherent uncertainty as to whether the operation could have had an effect on the object or not, by reduction to wait-free consensus. This implies that the implementation must sometimes return a special *pause* value, indicating that the client should re-invoke the same operation. We extend the notion of linearizability so as to accommodate failed operations and re-invocations of paused operations.

An obstruction-free implementation of any object is presented (Section 3), which exemplifies how pause and fail values are returned when a legal response is not possible. A natural way to evaluate obstruction-free implementations is by considering the *contention-free step complexity*, namely, the number of steps taken by a process running alone, until it returns a value. Our implementations have linear contention-free step complexity and use a linear number of read/write base objects. By reduction to the lower bound of Jayanti, Tan and Toueg [20], we show that obstruction-free implementations of many long-lived objects from historyless base objects must have $\Omega(n)$ contention-free step complexity and must use $\Omega(n)$ historyless objects.

In practice, the burden of providing termination of obstruction-free implementations is shifted to a system-supported *contention manager* that relies on low-level mechanisms such as timers, identifiers and interrupts [18, 26]. The contention manager instructs the clients if and when to invoke operations, trying to ensure that only a single process eventually accesses the concurrent object. To explore inherent characteristics of obstruction-free implementations, we consider a specific contention manager that can turn any obstruction-free implementation into a nonblocking one (none of those of [15,18,26,27] can do so). The contention manager indicates the client whether to continue or not (a binary indication), and should eventually indicate only to a single client to continue [9]. [1]

We show (Section 3) that there are no *gracefully degrading* consensus implementations, which are nonblocking when the contention manager operates correctly, but remain (at least) obstruction-free when the contention manager is unsuccessful.

We finally explore *solo-fast* implementations [25]. These are wait-free linearizable object implementations that use only read/write base objects when there is no step contention, but may fall back on more powerful objects like compare&swap, when contention occurs. Luchangco, Moir and Shavit [25] call these implementations solo-fast because read and write operations are considered comparatively cheap. They presented a generic object implementation that uses only reads and writes when an operation runs in the absence of contention. However, in their implementation this also means lack of pending operations, namely, lack of *point* contention; moreover, a transient raise of point contention will cause a subsequent operation (that has no point contention) to invoke costly C&S operations. In light of this, it is challenging to design truly solo-fast implementations that do not invoke C&S operations in the more common case of no *step* contention.

Surprisingly, we show in this paper that any object has a solo-fast implementation by describing a solo-fast consensus implementation, and employing it within Herlihy's universal con-

[1]This specification style is inspired by the way *failure detectors* [9,10] abstract away (partial) synchrony assumptions. It highlights the intriguing connection between obstruction-free implementations and Paxos-style algorithms for consensus and state-machine replication [22].

struction [16] (Section 4). The implementation has linear contention-free step complexity. We conjecture that, just like obstruction-free implementations, solo-fast ones are also inherently slow: they must have (at least) non-constant step complexity.

2 Model

Processes, objects and implementations. A system contains a set Π of $n > 1$ *processes* $p_1, \ldots, p_n$ that communicate through shared *objects*.

Every object has a *type* that is defined by a triple (O, R, Δ), where O is a set of *invocations*, R is a set of *responses*, and Δ is a set of sequences of invocation-response pairs. The set Δ, known as the *sequential specification* of the type, contains all the sequences of invocations and responses allowed by the object.

As an example, consider the *compare&swap* object, which is accessed by a *C&S*$(r_1,\ r_2,\ m)$ operation; the operation compares that value in memory location m with the content of local variable r_1, and if equal, writes the value of r_2 to m. The operation returns the *old* value of m. The sequential specification of the *compare&swap* type includes all sequences of *C&S* operations that obey this rule.

Another important example is the *consensus* object, on which processes perform a *propose* operation with an argument in some set V. The sequential specification of consensus includes all sequences of *propose* operations that return the argument of the first operation in every sequence.

To implement a (high-level) object from a collection of *base* objects, processes follow an *algorithm* A, which is a collection of state machines $A_1, \ldots A_n$, one for each process.

When receiving an *invocation* (to the high-level object), process p_i takes steps according to A_i. In each step, p_i can either (a) invoke an operation on a base object, or (b) receive the response of its previous base operation, or (c) perform some local computation. After each step, p_i changes its local state according to A_i, and possibly returns a response on the pending high-level operation.

Executions and histories. We investigate implementations that work in environments where process speeds are highly-variable, and at the extreme case, a process may stop taking steps.

An *execution* e of an algorithm A is a sequence of interleaved *events*. Every execution induces a *history* that includes only the invocations and responses of the high-level operations. Each invocation or response is associated with a single process and a single object. A *local history* of process p_j in H, $H|j$, is the subsequence of H containing only events of p_j. Similarly, $H|x$ is the subsequence of H of operations on an object x.

A response *matches* an invocation if they are associated with the same process and the same object. A matching invocation-response pair $[i, r]$ is called a *complete operation*, and we say that i *returns* r. An invocation i without a matching response is called a *pending operation*; a *completion* of a pending operation, that is, an invocation, is the invocation together with with an appropriate response. The fragment of H (or e, its corresponding execution) between the invocation i and its matching response r (if it exists) is the operation's *interval*.

In an infinite execution, a process is *correct* if it takes an infinite number of steps or it has no pending operation; otherwise, it is *faulty*.

A local history is *well-formed* if it is a sequence of matching invocation-response pairs, except perhaps for the last invocation in a finite local history. A history H is *well-formed* if every local history in H is well-formed.

A history H is *sequential* if every invocation is immediately followed by its matching response. A sequential history H is *legal* if for every object x, $H|x$ is in the sequential specification of x.

Linearizability. Two different invocations i and i' on the same object x are *concurrent* in a history H, if i and i' are both pending in some finite prefix of H. This implies that their intervals overlap. We say that two operations $[i, r]$ and $[i', r']$ (or i' if i' is pending) are *non-concurrent* if their intervals are non-overlapping: Either r appears before i' in H, in which case we say that $[i, r]$ *precedes* $[i', r']$, or r' appears before i in H, in which case we say that $[i, r]$ *follows* $[i', r']$.

A well-formed history H satisfies *extended linearizability* [16] (see also [6, Chapter 10]) if there is a permutation H' containing all the complete operations and completions of a subset of the pending operations in H, such that (1) H' is legal, and (2) H' respects the order of non-concurrent operations in H.

Measures of contention. This paper explores the benefits induced by the scenarios in which contention is rare. Formally, we define the *step contention* of a fragment in execution e to be the number of processes that take steps in this fragment. An operation $[i, r]$ is *step contention-free in* e if step contention of its interval in e is 1. An operation is *eventually step contention-free in* e if its interval in e has a suffix with step contention 1.

Alternative ways to measure contention during an operation's interval were previously defined [2, 5]: The *interval contention* during $[i, r]$ is the number of processes whose operations are concurrent with $[i, r]$ in e. The *point contention* of $[i, r]$ is the maximum number of operations *simultaneously* concurrent with $[i, r]$ in e. Clearly, both are always equal to or higher than step contention. Note also when no operation overlaps $[i, r]$, then both the point contention and the interval contention are 1.

3 Obstruction-Free Implementations

This section considers obstruction-free implementations [17, 18], which guarantee progress only in the absence of step contention.

Definitions. Originally [17, 18], an implementation is called obstruction-free *"if it guarantees progress for every thread that eventually executes in isolation. Even though other threads may be in the midst of executing operations, ..."* [18, Page 522]. This requirement is also called *solo termination* [12], and it echoes the liveness correctness conditions stated for Paxos-style algorithms for state-machine replication [22].

Using our terminology, an implementation is *obstruction-free* if every operation that is eventually step contention-free eventually returns.

Obstruction-freedom is a very weak liveness condition, and it requires the operation to return only under very restricted conditions. In all other circumstances, we only require that an operation's response is legal, *if it returns a response at all.*

If an operation cannot return a legal response, it is useful to return control to a higher-level entity, which we call the *client.* The client may consult a system-specific mechanism called a *contention manager*, in order to expedite termination.

There are two ways in which an obstruction-free implementation returns control to the client, depending on whether the implementation is certain that the operation did not have any effect on the object or not. In the former case, a special *pause* value $\perp$ is returned, and the client must re-invoke the same operation until a non-$\perp$ response is received. In the latter case, a special *fail* value is returned, indicating that the operation was not applied, and the client is free to invoke any operation it wishes. (We come back to the need for the two indications in Section .)

We add to R, the set of responses of an object, a special *pause* value $\perp \notin R$ and a special *fail* value $\emptyset \notin R$. The definition of a *well-formed* local history is extended to require that if an invocation i is followed by the response $\perp$, then the subsequent event, if exists, is i.

The definition of extended linearizability is further extended so that invocations returning *fail* are removed from the linearized history, while a sequence of invocations returning *pause* are considered as one pending operation.

Formally, let H be any history, and H' be any well-formed local history of H. Let i be an invocation in H' on an object x. A fragment of the form i, x in H', where $x \in R$, is called an *occurrence of* i (returning x). Since an invocation occurrence might return $\perp$ and be re-invoked later, there might be a number of occurrences of i in a history. Consider the longest fragment of the form i or $i, \perp, i, \ldots, \perp, i$ in H'. If the fragment is followed by a matching response $r \notin \{\perp, \emptyset\}$, we call i, r or, resp., $i, \perp, i \ldots, \perp, i, r$ a *complete operation.* If the fragment is followed by a fail response $\emptyset$, we call $i, \emptyset$ or, resp., $i, \perp, i \ldots, \perp, i, \emptyset$ a *failed operation.* If the fragment is followed by no event or by $\perp$, we call i or $i, \perp$ or, resp., $i, \perp, i, \ldots, \perp, i$ or $i, \perp, i \ldots, \perp, i, \perp$ a *pending operation.* Since H' is well-formed, a pending operation is a suffix of H' ($\perp$ cannot be followed by an invocation other than i). The operation's interval is the shortest fragment of H that includes all events of that operation. If the fragment $i, \perp, i, \ldots, \perp, i$ is followed by no event in H' then the operation's *interval* is infinite.

As defined before, a well-formed history H is *linearizable* if there is a permutation H' containing all the complete operations in H and completions of a subset of the pending operations in H, such that H' is legal and it respects the order of non-concurrent operations in H. When taken in the context of the extended notions of complete and pending operations, this definition means that we order all non-failed operations, with paused operations "spanned" across their re-invocations.

An implementation is *live* if every invocation occurrence returns in a finite number of its own steps (although a value in $\{\perp, \emptyset\}$ can be returned). An implementation is *valid* if (1) an invocation occurrence returns $\perp$ (that is, *pause*) only when it is not step contention-free, and (2) an invocation occurrence i returns $\emptyset$ (that is, *fail*) only when the corresponding *operation* (the longest fragment of the form $i, \emptyset$ or $i, \perp, i, \ldots, \perp, i, \emptyset$ in the local history) is not step contention-free. It is immediate that any live and valid implementation is obstruction-free.

Ideally, the obstruction-free implementation should always return a *fail* indication to the client, allowing it to either re-invoke the operation, or to invoke another operation. We show that it is impossible to implement an obstruction-free consensus object that is only allowed to return *fail* in the case of step contention. Thus, the choice of returning $\perp$ in obstruction-free implementations is sometimes unavoidable.

Theorem 1 *There is no obstruction-free consensus implementation from registers that never returns $\perp$.*

Proof. By contradiction, consider an implementation of obstruction-free consensus that is allowed to returns only $\emptyset$ in the case of step contention. Then it is possible to solve consensus among two processes p_0 and p_1 using one such consensus object, denoted C, and one register R, contradicting [11, 24]. The algorithm is presented in Figure 2.

Validity of the algorithm follows from the fact that $\emptyset$ is returned only in case of step contention. If C returns $\emptyset$ at p_0, then p_1 can only decide its own value. Thus, Agreement is satisfied. Since p_1 eventually runs in the absence of contention, it eventually decides. Thus, Termination is also satisfied. □

Shared variables: *register* R, initially $\perp$, and *"only-fail" OF consensus object* C

```
Code for process $p_0$:
  upon $propose(v_0)$ do
    $d_0 \leftarrow C.propose(v_0)$
    if $d_0 = \emptyset$ then
      $d_0 \leftarrow R$
    return $d_0$
```

```
Code for process $p_1$:
  upon $propose(v_1)$ do
    $R \leftarrow v_1$
    repeat
      $d_1 \leftarrow C.propose(v_1)$
    until $d_1 \neq \emptyset$
    return $d_1$
```

Figure 2: Wait-free consensus from "only-fail" obstruction-free consensus

Obstruction-Free Generic Object Implementation. This section gives an algorithm that obstruction-free implements any object of type T, using only registers. Like previous universal implementations, it is built from consensus objects. (A simple obstruction-free consensus algorithm is detailed in Appendix A.)

The universal obstruction-free implementation relies on a sequential implementation of the object type T; it is live, valid and linearizable. Herlihy's universal nonblocking implementation [16] cannot be applied "off-the-shelf" since it does not handle re-invocations and failing. Instead, the algorithm builds on similar ideas, while making sure that *pause* or *fail* are returned only in the absence of step contention.

An object of type T is represented as a linked list; an element of the list represents an operation applied to the object. The list of operations clearly determines the list of corresponding responses.[2] A process makes an invocation by appending a new element to the end of the list. The algorithm assumes a function $response(invs, inv)$ that returns the response matching the invocation inv in a sequential execution of invocations from list $invs$ (under the condition that $inv \in invs$).

The algorithm (Figure 5 in Appendix B) uses the following shared variables:

– n atomic single-writer, multi-reader registers $L_1, \ldots, L_n$. Process p_i stores in L_i its last view of the object state in the form of a linked list of operations that p_i witnessed to be applied on the object.

– $C[\,]$ is an unbounded array of obstruction-free consensus objects. The array is used to agree on the order in which invocations are put into the linked list of operations.

Roughly, the algorithm works as follows. When a process p_i executes an invocation inv, it identifies the longest list L_j (let $k = |L_j|$). If inv is already in L_j, the response associated with inv in L_i is returned (line 5). This ensures that an operation takes effect at most once, even if repeated several times. If it is not the first instance of inv, and $k > |L_i|$ (i.e., inv was not decided in any OF Consensus to which it was proposed), p_i returns $\emptyset$ (line 9). Otherwise, p_i proposes inv to $C[k+1]$ (line 10). If $C[k+1]$ returns $\perp$ (step contention is detected), then p_i returns $\perp$ (line 13). If the propose operation fails, or returns a non-inv response while it is not the first instance of inv, then p_i returns $\emptyset$ (line 16). If $C[k+1]$ returns inv, then p_i returns the response associated with inv (line 20). Otherwise, the procedure is repeated, now at position $k+2$. Now if $C[k+2]$ returns a non-$\{inv, \perp\}$ response, then p_i returns $\emptyset$ (line 29). The second consensus operation ensures *validity* of the implementation, namely, that $\emptyset$ is never returned in line 29 if the corresponding operation is step contention-free.

This algorithm implies the next theorem (the correctness proof is sketched in Appendix B):

[2] A non-deterministic object can be implemented by deterministically restricting its sequential specification.

Theorem 2 *Every sequential type T has an obstruction-free linearizable implementation from registers.*

Remark. The algorithm satisfies one additional property. In any execution, every operation takes effect (if it does) before it stops taking steps in that execution. In other words, the implementation stays linearizable even if we restrict an operation's interval to the shortest fragment of the execution which contains all steps of that operation. As a result, an operation invoked by a faulty process takes effect (if it does) before the process fails, which makes our implementations *strictly linearizable* [3].

Obstruction-Free Implementations are Slow. The universal construction presented in Figure 5 is not very efficient: finding the longest list of invocations requires to collect information from all processes. The next theorem shows that this is inherent in obstruction-free universal implementations from read/write base objects, by proving a lower bound of $\Omega(n)$ on the number of steps and on the number of registers for implementing a compare&swap object.

Theorem 3 *Let A be any obstruction-free implementation of n-valued compare&swap from registers, then A has an execution in which a step contention-free operation takes $n-1$ or more steps and accesses $n-1$ or more different objects.*

Proof. Follows directly from the result of Jayanti, Tan and Toueg [20]. Indeed, it is shown in [20] that any implementation of n-valued compare&swap that satisfies the solo termination property has an execution in which a solo operation (i.e., an operation that does not observe step contention) takes $n-1$ or more steps and accesses at $n-1$ or more different objects. Since any obstruction-free implementation ensures the solo termination property, we immediately have the theorem. $\square$

Leveraging Obstruction-Free Objects. The next two subsections discuss how obstruction-free implementations can be turned into nonblocking ones using a contention manager. The contention manager we consider provides the client with a binary indication whether to continue or not. The contention manager works well when it indicates only to a single client to continue. Formally, in response to the client's query, the contention manager returns either 0 or 1, telling the client to abort or to continue (respectively); in the latter case, we say that the client is a *leader*. The *eventual* contention manager, denoted Ω, guarantees that eventually exactly one correct client is a leader; it is deliberately the same as the "sloppy leader" failure detector and can be implemented using partial synchrony assumptions [9].

A single obstruction-free consensus object and a weak contention manager, Ω, can implement *nonblocking* consensus using the following simple algorithm: A process queries the contention manager and, if it is a leader, the process makes a *propose* invocation on the underlying obstruction-free consensus object. If the response is neither $\perp$ nor $\emptyset$, it is returned; otherwise, the process repeats. This implies the following result:

Theorem 4 *Consensus has a nonblocking implementation from obstruction-free consensus and Ω.*

Graceful Degradation of Obstruction-Free Implementations. *Obstruction-free* consensus can be implemented from registers only [18]; on the other hand, *wait-free* consensus can be implemented from registers using Ω [23]. However, these two liveness properties cannot be combined in the same implementation, namely, there is no wait-free consensus implementation using registers and Ω which becomes (at least) obstruction-free when the contention manager fails to eventually elect a single correct leader. In fact, we prove the claim even for *nonblocking* consensus implementations.

Theorem 5 *There is no nonblocking consensus implementation using registers and Ω that ensures obstruction-freedom when the contention manager fails to eventually elect a single correct leader.*

Proof. Suppose, by contradiction, that an algorithm A provides such an implementation. We show that it is then possible to devise an algorithm A' that implements nonblocking consensus for two processes, p_1 and p_2 with registers only — a contradiction with [11, 24].

In A', processes take steps like in A, except that, instead of using Ω, processes assume that Ω always indicates p_1 as the only leader. In doing so, processes cyclically invoke *propose* operations until a non-$\{\perp, \emptyset\}$ value is returned. Note that A' cannot violate safety properties of consensus, since every finite execution of A' is also an execution of A. To establish a contradiction, it is thus sufficient to show that at least one correct process eventually terminates in A', i.e., obtains a non-$\{\perp, \emptyset\}$ value from the underlying algorithm A.

Every execution of A' belongs to one of the following classes:
(1) Executions in which p_1 is correct, i.e., the assumed output of the contention manager complies with the specification of Ω. Such an execution is indistinguishable to p_1 and p_2 from executions of A in which processes $p_3, \ldots, p_n$ are initially faulty, and p_1 is the only correct leader. Since A implements a nonblocking consensus using Ω, some correct process (p_1 or p_2) eventually obtains a non-$\{\perp, \emptyset\}$ value from A and decides.
(2) Executions in which p_1 is faulty, i.e., the assumed output of the contention manager does not comply with the specification of Ω. Assume that p_2 is correct in such an execution (if both p_1 and p_2 are faulty, consensus is trivially solved). Any finite prefix of our execution is indistinguishable to p_2 from an execution of A in which processes $p_3, \ldots, p_n$ are initially faulty, and the contention manager malfunctions. Since p_2 is eventually running in the absence of step contention, and A ensures obstruction-freedom even when the contention manager is incorrect, p_2 eventually obtains a non-$\{\perp, \emptyset\}$ value from A and decides.

In other words, A' guarantees that whenever there is at least one correct process, some correct process eventually decides — a contradiction. $\Box$

4 Solo-Fast Implementations

We say that a wait-free linearizable implementation of a sequential type T from registers and other objects (e.g., compare&swap) is *solo-fast* if only read and write operations are invoked by any step contention-free operation on it.

Figure 3 presents a solo-fast consensus implementation. In the algorithm, each process p_i starts from the smallest round in which a value can be *fixed*, i.e., returned in line 19 (we say that p_i *joins* in that round). In every round, p_i tries to fix its current estimate. It is ensured that if no other process tries to fix concurrently a different value in the current or higher round, then the estimate must be fixed. If p_i is not able to fix the estimate in the current round (we say that p_i *fails* in that round), which can only happen when there is step contention, it updates the estimate using a C&S operation and goes to the next round. The algorithm guarantees that whenever process p_i fails in round k, and no process joins in round $k+1$, then p_i fixes its estimate in round $k+1$ (C&S ensures that no two processes that fail in round k try to fix different values in round $k+1$).

Theorem 6 *There is a solo-fast consensus implementation from registers and C&S objects, takes takes $O(n)$ steps in the solo path.*

Proof. Validity follows immediately from the algorithm.

```
Shared variables:
    Registers $\{A_j\}, \{B_j\}, j \in \{1, 2, \ldots, n\}$, initially $\bot$
    C&S $C_1, \ldots C_{n-1}$, initially $\bot$

 1: upon $propose(input_i)$ do
 2:   $V \leftarrow$ collect $A$                                   { $\bot$'s are ignored in each collect }
 3:   $k_i \leftarrow \min\{k \geq 1 \mid \forall(k', v') \in V : k' \leq k \wedge \forall(k, v'), (k, v'') \in V : v' = v''\}$
 4:   if $\exists(k, v) \in V$ then
 5:     $v_i \leftarrow v$
 6:   else
 7:     $V' \leftarrow$ collect $B$
 8:     if $V' \neq \emptyset$ then
 9:       $v_i \leftarrow$ the highest timestamped value in $V'$
10:     else
11:       $v_i \leftarrow input_i$
12:   while (true) do
13:     $A_i \leftarrow (k_i, v_i)$
14:     $V \leftarrow$ collect $A$
15:     if $\forall(k', v') \in V : k' < k_i \vee (k' = k_i \wedge v' = v_i)$ then
16:       $B_i \leftarrow (k_i, v_i)$
17:       $V \leftarrow$ collect $A$
18:       if $\forall(k', v') \in V : k' < k_i \vee (k' = k_i \wedge v' = v_i)$ then
19:         return $v_i$
20:     $V' \leftarrow$ collect $B$
21:     if $V' \neq \emptyset$ then
22:       $v_i \leftarrow$ the highest timestamped value in $V'$
23:     $v' \leftarrow C_{k_i}.CS(\bot, v_i)$
24:     if $v' \neq \bot$ then $v_i \leftarrow v'$
25:     $k_i \leftarrow k_i + 1$
```

Figure 3: An n-process solo-fast consensus: code for process p_i

We say that a process p_i *reaches round* k in an execution of the algorithm if it reaches line 13 with $k_i = k$ in that execution. We say that a process p_i *joins in round* k if k is the first round p_i reaches. We say that p_i *participates* in an execution if it reaches round $k \geq 1$ in that execution. We establish the correctness of our algorithm through Claims 1–3 (the proofs of the claims are given in Appendix C).

Claim 1 *For all $k \geq 2$, if no process joins in round k or later, then registers A contain no (k', v') such that $k' > k$ and no two (k, v') and (k, v'') such that $v' \neq v''$.*

Claim 2 *Let $k \geq 2$ be the number of processes that participate in the algorithm. Then no process joins in round k or later.*

Since there are at most n participants, by Claim 2, no process joins in round n or later. By Claim 1, in any execution, registers A contain no (k', v') such that $k' > k$ and no two (k, v') and (k, v'') such that $v' \neq v''$. Thus, any process that reaches round n will pass the "if" clauses in lines 15 and 18 and return in line 19. Thus, every process decides in round n or earlier — the Termination property of consensus is ensured.

Claim 3 *Let process p_i return v in round k. Then registers A contain no (v', k') such that $v' \neq v$ and $k' \geq k + 1$ and registers B contain no (v', k') such that $v' \neq v$ and $k' \geq k$.*

Let $1 \leq k \leq n$ be the first round such that some process p_i decides some value v in round k. By the algorithm, if a process p_j decides v' in round k', then it previously wrote (k', v') in B_j (line 16).

By Claim 3, no process p_j writes (k', v') such that $v' \neq v$ and $k' \geq k$. Thus, no process decides $v' \neq v$, which implies the Agreement property of consensus.
Solo-fast. Assume that process p_i joins in round k. The only reason for p_i to fail in that round is to observe a value timestamped with $k' > k$ or two different values timestamped with k in registers A (line 15 or 18). But p_i previously observed the opposite in line 3 and adopted the value timestamped with k found in registers A (if any) in line 5. Thus, p_i can fail in round k only when some other process p_j concurrently writes (k', v') in A_j such that $k' > k \vee k' = k \wedge v' \neq v$, i.e., when there is step contention. Clearly, if p_i does not fail in round k, then it takes a linear number of read and write operations. $\Box$

From Theorem 6 and Herlihy's universal construction [16], we immediately obtain:

Corollary 7 *Every sequential type T has a solo-fast implementation from registers and C&S objects.*

5 Discussion

This paper studies the notion of step contention, which inherently does not charge for processes stalled, e.g., due to failures or swap-outs, and is, in this sense, fundamentally different from point or interval contention. We show that registers are powerful enough to ensure liveness in the absence of step contention (which leads to a wider set of executions than when looking at other forms of contention). However, we suggest that such implementations are inherently expensive and of limited benefit.

We believe our paper opens up several interesting avenues for further research:

Complexity of obstruction-free consensus. We have shown tight bounds on the cost of *generic* obstruction-free implementations. However, there might be more efficient obstruction-free solutions for specific problems. For obstruction-free consensus, for example, an $\Omega(\sqrt{n})$ lower bound on the number of registers (or historyless objects) can be derived from the lower bound of Fich, Herlihy and Shavit [12]. This bound is not tight (the upper bound is $O(n)$) and moreover, it does not bound the *contention-free* step complexity of obstruction-free consensus.

Complexity of solo-fast implementations. Our solo-fast implementation performs $O(n)$ steps, even in the absence of step contention; by employing adaptive collect [1, 5], the step complexity can be made to depend only on the point contention; by employing adaptive collect for unbounded concurrency [14], it can be made independent of the number of processes.

By a simple variation on the proof of [20], the contention-free step complexity of any generic solo-fast implementation using *non-readable* compare&swap objects can be shown to be at least linear in the point contention. (This matches the contention-free step complexity of our implementation.) We conjecture that a non-constant lower bound holds even if compare&swap objects are readable, making solo-fast implementations rather inefficient. On the other hand, it is possible that the step and space complexities of solo-fast consensus can be made constant if objects *slightly* more powerful than read/write registers, e.g., counters or queues, are used on a fast path.

Better mechanisms for contention management. The contention manager we considered is fundamentally different from those considered in [15, 18, 26, 27]. It is easy to see that none of those can transform any obstruction-free implementation into a nonblocking one. Those contention managers do not provide any worst case nonblocking deterministic guarantees (with the exception of [15] in the absence of failures), and were actually rather designed to provide a high throughput in the average case. Devising a contention manager that would provide deterministic worst case guarantees with acceptable average case throughput is an interesting research direction.

Acknowledgments: We would like to thank Partha Dutta, Ron Levy and Eric Ruppert for important discussions on the topic of this paper.

References

[1] Y. Afek, G. Stupp, and D. Touitou. Long-lived adaptive collect with applications. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 262–272, 1999.

[2] Y. Afek, G. Stupp, and D. Touitou. Long-lived adaptive splitter and applications. *Distributed Computing*, 15(2):67–86, 2002.

[3] M. K. Aguilera and S. Frølund. Strict linearizability and the power of aborting. Technical report, HP Laboratories Palo Alto, 2003.

[4] J. Aspnes and M. Herlihy. Fast randomized consensus using shared memory. *J. Algorithms*, 11(3):441–461, 1990.

[5] H. Attiya and A. Fouren. Algorithms adapting to point contention. *J. ACM*, 50(4):444–468, 2003.

[6] H. Attiya and J. L. Welch. *Distributed Computing: Fundamentals, Simulations and Advanced Topics (2nd edition).* Wiley, 2004.

[7] B. N. Bershad. Practical considerations for non-blocking concurrent objects. In *Proceedings of the 14th IEEE International Conference on Distributed Computing Systems (ICDCS'93)*, pages 264–273, 1993.

[8] R. Boichat, P. Dutta, S. Frølund, and R. Guerraoui. Deconstructing Paxos. *ACM SIGACT News Distributed Computing Column*, 34(1):47 – 67, March 2003. Revised version of EPFL Technical Report 200106, January 2001.

[9] T. D. Chandra, V. Hadzilacos, and S. Toueg. The weakest failure detector for solving consensus. *Journal of the ACM*, 43(4):685–722, July 1996.

[10] T. D. Chandra and S. Toueg. Unreliable failure detectors for reliable distributed systems. *Journal of the ACM*, 43(2):225–267, March 1996.

[11] D. Dolev, C. Dwork, and L. J. Stockmeyer. On the minimal synchronism needed for distributed consensus. *Journal of the ACM*, 34(1):77–97, January 1987.

[12] F. Fich, M. Herlihy, and N. Shavit. On the space complexity of randomized synchronization. *J. ACM*, 45(5):843–862, 1998.

[13] M. J. Fischer, N. A. Lynch, and M. S. Paterson. Impossibility of distributed consensus with one faulty process. *Journal of the ACM*, 32(3):374–382, April 1985.

[14] E. Gafni, M. Merritt, and G. Taubenfeld. The concurrency hierarchy, and algorithms for unbounded concurrency. In *Proceedings of the 20th Annual ACM Symposium on Principles of Distributed Computing (PODC)*, pages 161–169, 2001.

[15] R. Guerraoui, M. Herlihy, and B. Pochon. Toward a theory of transactional contention managers. In *Proceedings of the 24th Annual ACM Symposium on Principles of Distributed Computing (PODC)*, 2005.

[16] M. Herlihy. Wait-free synchronization. *ACM Transactions on Programming Languages and Systems*, 13(1):124–149, January 1991.

[17] M. Herlihy, V. Luchangco, M. Moir, and W. N. Scherer III. Software transactional memory for dynamic-sized data structures. In *Proceedings of the 22nd Annual ACM Symposium on Principles of Distributed Computing (PODC)*, pages 92–101, 2003.

[18] M. Herlihy, V. Luchango, and M. Moir. Obstruction-free synchronization: Double-ended queues as an example. In *Proceedings of the 23rd IEEE International Conference on Distributed Computing Systems (ICDCS'03)*, pages 522–529, 2003.

[19] M. Herlihy and J. M. Wing. Linearizability: a correctness condition for concurrent objects. *ACM Transactions on Programming Languages and Systems*, 12(3):463–492, June 1990.

[20] P. Jayanti, K. Tan, and S. Toueg. Time and space lower bounds for nonblocking implementations. *SIAM Journal on Computing*, 30(2):438–456, 2000.

[21] A. LaMarca. A performance evaluation of lock-free synchronization protocols. In *Proceedings of the 13th Annual ACM Symposium on Principles of Distributed Computing (PODC)*, pages 130–140, 1994.

[22] L. Lamport. The part-time parliament. *ACM Transactions on Computer Systems*, 16(2):133–169, May 1998.

[23] W.-K. Lo and V. Hadzilacos. Using failure detectors to solve consensus in asynchronous shared-memory systems. In *Proceedings of the 8th International Workshop on Distributed Algorithms (WDAG'94)*, volume 857 of *LNCS*, pages 280–295. Springer Verlag, 1994.

[24] M. C. Loui and H. H. Abu-Amara. Memory requirements for agreement among unreliable asynchronous processes. *Advances in Computing Research*, pages 163–183, 1987.

[25] V. Luchango, M. Moir, and N. Shavit. On the uncontended complexity of consensus. In *Proceedings of the 17th International Symposium on Distributed Computing (DISC'03)*, pages 45–59, 2003.

[26] M. L. Scott and W. N. Scherer III. Contention management in dynamic software transactional memory. In *PODC Workshop on Concurrency and Synchronization in Java Programs*, July 2004.

[27] M. L. Scott and W. N. Scherer III. Advanced contention management for dynamic software transactional memory. In *Proceedings of the 24th Annual ACM Symposium on Principles of Distributed Computing (PODC)*, 2005.

```
Shared variables:
    $Register\ R_1, \ldots, R_n \leftarrow (0,0,\perp), \ldots, (0,0,\perp)$
Local variables:
    $r_i \leftarrow 0;\ w_i \leftarrow 0;\ v_i \leftarrow \perp;\ check \leftarrow false$

 1: upon propose($v$) do
 2:   $regSet \leftarrow \{R_1, \ldots, R_n\}$
 3:   $mr \leftarrow \max\{r : (r,*,*) \in regSet\}$                    { Adopt the highest round number }
 4:   $r_i \leftarrow$ the smallest integer s.t. $((r_i\ mod\ n = i)$ and $(r_i > mr))$
 5:   $R_i \leftarrow (r_i, w_i, v_i)$                                   { Register the round number }
 6:   $regSet \leftarrow \{R_1, \ldots, R_n\}$
 7:   choose $v'$ s.t. $(*, mw, v') \in regSet$ and $mw = \max\{w : (*,w,*) \in regSet\}$   { Choose the "highest" announced value }
 8:   if $(v' \neq \perp)$ then $v_i \leftarrow v'$ else $v_i \leftarrow v$
 9:   $w_i \leftarrow r_i$
10:   $R_i \leftarrow (r_i, w_i, v_i)$                                   { Announce the current estimate }
11:   $check \leftarrow true$
12:   $regSet \leftarrow \{R_1, \ldots, R_n\}$
13:   if $(\exists (r,*,*) \in regSet$ s.t. $r > r_i)$ then
14:     if $(\exists (*, w, val) \in regSet$ s.t. $w > r_i$ and $val \neq \perp$ and $val \neq v_i)$ then
15:       $check \leftarrow false$
16:       return $\emptyset$                          { Fail if a different value is announced in a higher round }
17:     else
18:       $check \leftarrow true$
19:       return $\perp$                                                  { Pause }
20:   else
21:     $check \leftarrow false$
22:     return $v_i$                                                      { Decide on $v_i$ }
```

Figure 4: Obstruction-free consensus implementation: code for process p_i

A Obstruction-Free Consensus

As suggested in [18], an obstruction-free consensus algorithm can be derived by "de-randomizing" randomized consensus algorithms [4]. Here we show how "fail" and "pause" values are returned by a simple single-shot "Paxos-style" algorithm for obstruction-free consensus. Our algorithm (Figure 4) translates the $\Diamond$Register implementation [8], from message-passing to read-write shared memory.

Every process p_i maintains a current estimate of the decision value of the consensus, denoted by v_i and initialized to $\perp$, and two counters r_i and w_i, both initialized to 0. The counter r_i denotes the round number adopted by the current operation, and w_i denotes the last round number in which p_i "announced" its estimate v_i. Each process p_i is designated a single-writer multi-reader register R_i, initialized to $(0, 0, \perp)$ that is written by p_i and read by all processes. A boolean *check*, initialized to *false*, indicates whether p_i has been paused.

Roughly, the algorithm can be decomposed into two phases. In the first phase, process p_i chooses the highest unique round number, "registers" it (writes in R_i, line 5), and collects the shared memory. In the second phase, p_i adopts the value announced in the highest round (or its own proposal if no value is announced so far) as v_i, "announces" it with the current round number, and collects the shared memory. If p_i observes that some process registered a higher round (which can occur only if step contention is > 1) and a value different from v_i is announced in a higher round, then p_i returns $\emptyset$ (line 16). If step contention is detected and v_i is still the highest announced value, then p_i returns $\perp$ (line 19). Otherwise, p_i returns v_i. These two phases (register and announce) ensure that once p_i returns v_i, no process will ever announce a value different from v_i in a higher round. As a result no two processes can return different non-$\{\perp, \emptyset\}$ values, and $\emptyset$ is never returned if the corresponding operation took effect.

Theorem 8 *There exists an obstruction-free linearizable consensus implementation from registers.*

Proof. The liveness property of the implementation in Figure 4 is straightforward: there are no cycles or wait statements in the code.

To prove linearizability of the implementation, it is sufficient to show that (i) every returned non-$\{\perp, \emptyset\}$ value is a previously proposed value, and (ii) no two processes return different non-$\{\perp, \emptyset\}$ values.

We observe that (i) follows directly from the algorithm.

To prove (ii), we first show that if a process p_i returns a non-$\{\perp, \emptyset\}$ value v in a round r, then no process can announce a value different from v in a higher round. Indeed, let p_j be the first process to announce a value v' in a round $r' > r$. We immediately observe that p_j registered r' (line 5) *after* p_i announced v in round r (otherwise, p_i would see that a round higher than r is registered in line 13 and return $\perp$ or $\emptyset$). Thus, when p_j collects the shared memory in round r' (line 6), v is the value announced in the highest round number. Thus, $v = v'$. Since before returning a non-$\{\perp, \emptyset\}$ value, every process announces the value, we have (ii).

Now we prove validity of the implementation. Responses in $\{\perp, \emptyset\}$ are returned only if step contention is detected (line 13). Moreover, if $\emptyset$ is returned (line 16), then p_i made sure that its current estimate is not the value announced in the highest round (line 14), so the estimate cannot be decided. That is, the $\mathsf{propose}(v)$ operation of p_i did not take effect in the execution.

Finally, we obtain an obstruction-free linearizable consensus implementation from registers. □

B Proof of Theorem 2

We prove that the implementation in Figure 5 is live, valid and linearizable.

We assume that invocations corresponding to different operations are uniquely identifiable (e.g., by associating a unique tag with every operation). We say that an invocation *inv* is *fixed at k* in *e*, and we write *fixed(e,inv,k)*, if some propose operation on $C[k]$ returns *inv* in *e* (in line 10 or 21). Clearly, at most one invocation can be fixed at every index.

We say that index k is *decided in e* if there is an invocation *inv* such that *fixed(e,inv,k)*. By inspecting the code in Figure 5, we observe that if index $k > 1$ is decided in e, then all indexes $k' < k$ are also decided. Moreover, if an invocation *inv* is not fixed at any index k at which it was proposed to OF Consensus in an execution *e*,then *inv* did not take effect in *e*.

The liveness follows from the fact that the algorithm has no loops and wait statements, and the underlying obstruction-free consensus objects are live.

Now we prove validity of the implementation. We observe that boolean *check* is true if and only if the current instance of invocation *inv* does not represent a new operation and all previous instances of *inv* returned $\perp$.

Assume that an operation returns a value in $\{\perp, \emptyset\}$. The following cases are only possible:

(1) $\emptyset$ is returned in line 9. This can only happen when all previous instances of *inv* returned $\perp$, and *inv* was not decided in any OF consensus object to which it was proposed. Thus, *inv* did not take effect, and the corresponding operation was not step contention-free.

Note that if p_i reaches line 10 while *check* is true, then *inv* is not fixed at any $k' \leq k$ and the last $\mathsf{propose}(inv)$ invoked by p_i on $C[k+1]$ returned $\perp$.

(2) $\perp$ is returned in line 13 or line 24. Thus, a propose operation on $C[k+1]$ or $C[k+2]$ returned $\perp$. This can only happen if the current operation is not step contention-free.

```
Shared variables:
    Register $L_1, \ldots, L_n \leftarrow \emptyset, \ldots, \emptyset$
    OF-Consensus $C[\,]$
Local variables:
    $check \leftarrow false$; $dec \leftarrow \bot$

1:  upon Invoking $inv$ do
2:    $invs \leftarrow longest(\{L_1, \ldots, L_n\})$          { Select the longest invocation list }
3:    if $inv \in invs$ then
4:      $check \leftarrow false$
5:      return $response(invs, inv)$          { Return if inv is already completed }
6:    $k \leftarrow |invs|$
7:    if $(k > |L_i|)$ and $check$ then
8:      $check \leftarrow false$
9:      return $\emptyset$          { Fail the operation }
10:   $dec \leftarrow C[k+1].propose(inv)$          { The 1st consensus operation }
11:   if $dec = \bot$ then
12:     $check \leftarrow true$
13:     return $\bot$
14:   if $(dec = \emptyset)$ or $(dec \neq inv$ and $check)$ then
15:     $check \leftarrow false$
16:     return $\emptyset$          { Fail the operation }
17:   $invs \leftarrow invs \cdot dec$; $L_i \leftarrow invs$          { Update $L_i$ }
18:   if $dec = inv$ then
19:     $check \leftarrow false$
20:     return $response(invs, inv)$          { Return if inv is decided }
21:   $dec \leftarrow C[k+2].propose(inv)$          { The 2nd consensus operation }
22:   if $dec = \bot$ then
23:     $check \leftarrow true$
24:     return $\bot$
25:   if $dec \neq \emptyset$ then $invs \leftarrow invs \cdot dec$; $L_i \leftarrow invs$
26:   if $dec = inv$ then
27:     $check \leftarrow false$
28:     return $response(invs, inv)$          { Return if inv is decided }
29:   return $\emptyset$          { Fail if inv is ignored twice }
```

Figure 5: An obstruction-free implementation of T: code for process p_i

(3) $\emptyset$ is returned in line 16. That is, either $C[k+1]$ returned $\emptyset$, or *inv* was not fixed at any $k' \leq k+1$. Thus, *inv* did not take effect, and the corresponding operation was not step contention-free.

(4) $\emptyset$ is returned in line 29. That is, $C[k+2]$ returned a value $dec \notin \{\bot, \emptyset, inv\}$. Let p_j be the process that previously proposed *dec* to $C[k+2]$. By the algorithm, before proposing, p_j has made sure that for some $p_l \in \Pi$, $|L_l| = k+1$ (lines 2, 6 and 17). But the longest list in $\{L_1, \ldots, L_n\}$ seen by p_i in the beginning of its operation had length k (line 6). Thus, p_l took steps in the interval of the current instance of *inv*, i.e., the operation is not step contention-free. Moreover, *inv* was not fixed at any $k' \leq k+2$, and thus did not take effect.

Thus, the implementation is obstruction-free. In particular, every step contention-free invocation representing a new operation returns a non-$\{\bot, \emptyset\}$ response.

Let H be the finite well-formed history generated by an execution e of our implementation from which we remove all failed operations, all repeated invocations and $\bot$ responses. To prove that the implementation is linearizable, it is sufficient to show that there is a permutation H' containing all the complete operations in H and completions of a subset of the pending operations in H, such that H' is legal and it respects the order of non-concurrent operations in H.

Let k^* be the highest decided index in e. We construct H' as the sequence $inv_1, r_1, inv_2, r_2, \ldots, inv_{k^*}, r_{k^*}$, where, for every $k \in \{1, \ldots, k^*\}$, $fixed(e, inv_k, k)$ and r_k is the response that corresponds (w.r.t. sequential specification of T) to inv_k in H. By construction, H' is legal with respect to T. By the algorithm, a non-$\{\bot, \emptyset\}$ response returned only if the corresponding invocation is fixed. That is, each response event in H occurs in H', i.e., H' contains all complete operations in H. Further, *fixed(e,inv,k)* implies that an invocation of *inv* was made by some process p_i in H before *inv* got fixed, i.e., H' contains only complete operations in H and completions of some pending operations in H.

Assume now that H contains operations $[inv, r]$ and $[inv', r']$ (or a pending operation *inv'*) such that r precedes *inv'* in H. Let p_i be the process that invoked *inv*, and let k be such that *fixed(e,inv,k)*. By the algorithm, before returning r, p_i makes sure that some process p_l previously updated L_l so that $inv \in L_l$ and $|L_l| \geq k$ (lines 3, 17 and 25). Thus, the invocation *inv'* that follows r will observe $L_l \geq k$, i.e., *inv'* can only be fixed at $k' \geq k+1$. Hence, H' preserves the order of non-concurrent operations in H.

Finally, we obtain an obstruction-free linearizable implementation of T.

C Proofs of Claims 1–3

Claim 1 *For all $k \geq 2$, if no process joins in round k or later, then registers A contain no (k', v') such that $k' > k$ and no two (k, v') and (k, v'') such that $v' \neq v''$.*

Proof. Indeed, if no process joins in round k or later ($k \geq 2$), then every process that reaches round k, previously completed round $k-1$ and invoked $CS(\bot, v_i)$ to update its current estimate (line 23). Since $CS(\bot, v_i)$ returns no two different values, all processes that reach round k have the same estimate. Thus, a process can fail in round k only if it reads (k', v') such that $k' > k$. Since no process joins in round k or later, this value can be written only by a process that previously failed in round k. Thus, no process can fail in round k and the claim follows. □

Claim 2 *Let $k \geq 2$ be the number of processes that participate in the algorithm. Then no process joins in round k or later.*

Proof. By induction on k.

If there is only one participant, then it trivially joins and decides in round 1. Assume that there are two participants, p_i and p_j. Consider a prefix e of our execution in which p_i just completed executing line 3. Since p_i does not write in e, in line 2, p_i finds at most one non-$\bot$ value in registers A. Note that if p_i finds a non-$\bot$ value (k', v') in A, then $k' = 1$ (otherwise, p_j reaches round 2 being the only participant — a contradiction). Thus, in line 3, p_i either finds no non-$\bot$ values, or finds exactly one value $(1, v')$ in registers A. In both cases, p_i joins in round 1.

Assume now that the claim holds for all $2 \leq k' \leq k$ and consider any execution with $k+1$ participants. Let p_i be the *last* process that joins in that execution. Consider a prefix e of our execution in which p_i just completed executing line 3. Let V be the result of collect taken by p_i in line 2. Since p_i does not write in e, the other (at most k) participants cannot distinguish e from an execution with at most k participants. By the induction hypothesis, no process joined in round k or later in e. By Claim 1, V contains no (k', v') such that $k' > k$ and no two (k, v') and (k, v'') such that $v' \neq v''$. By the algorithm (line 3), p_i must join in round k or earlier. The claim follows. □

Claim 3 *Let process p_i return v in round k. Then registers A contain no (v', k') such that $v' \neq v$ and $k' \geq k+1$ and registers B contain no (v', k') such that $v' \neq v$ and $k' \geq k$.*

Proof. By induction on the number m of processes that reach round $k+1$.

Let $m=0$, i.e., no process reaches round $k+1$. Thus, no process reaches any round $k' > k+1$. Trivially, no value with $k' > k$ can be written in registers A and B.

Assume, by contradiction that some process p_j writes (k, v') with $v' \neq v$ in B_j. Thus, p_j previously wrote (k, v') in A_j, and then p_j did not read (k, v) in A_i (otherwise, p_j would not pass the "if" clause in line 15). Hence, p_i has written (k, v) in A_i *after* p_j has written (k', v') in A_j. But then p_i will necessarily read (k, v') (or a value with a higher timestamp) in A_j and will not pass the"if" clause in line 15 — a contradiction.

Now assume that the claim holds when m processes reach round $k+1$ and consider an execution in which $m+1$ processes reach round $k+1$. Let p_j be *the last* process to reach round $k+1$ in that execution. The following two cases are possible.

(1) p_i joins in round $k+1$.

Then previously p_i first collected registers A in line 2 and then collected registers B in line 7. Let V and V' be the results of these two collects.

By the hypothesis, V contains no value $v' \neq v$ timestamped with $k' \geq k+1$ and V' contains no value $v' \neq v$ timestamped with $k' \geq k$. Since p_j joins in round $k+1$ and V contains no values timestamped with $k+1$, p_j does not pass the "if" clause in line 4 and adopts the highest timestamped value in V' (if any) in line 9.

(2) p_j participated in round k.

Thus, in round k, p_j read a value timestamped with $k' > k$ or two different values timestamped with k in registers A (line 14 or line 17) and then collected registers B in line 20. Let V and V' be the results of these two collects. Again, by the hypothesis, V contains no value $v' \neq v$ timestamped with $k' \geq k+1$ and V' contains no value $v' \neq v$ timestamped with $k' \geq k$. By the algorithm, p_i adopts the highest timestamped value in V' (if any) in line 22.

In both cases, (a) V contains a value timestamped with $k' > k$ or two different values timestamped with k, and (b) before reaching round $k+1$, p_j tries to adopt the highest timestamped value in V'.

Assume that v is *not* the highest timestamped value in V'. Since V' contains no $v' \neq v$ timestamped with $k' \geq k$, it follows that V' does not contain (k, v). That is, p_i wrote (k, v) in B_i *after* p_j read B_i. But p_j read B_i *after* it found a value timestamped with $k' > k$ or two different values timestamped with k in registers A. Hence, p_i will necessarily find a value timestamped with $k' > k$ or two different values timestamped with k in registers A. As a result, p_i cannot pass the "if" clause in line 15 in round k and decide, which is a contradiction. □