An Efficient Adaptive Partial Snapshot Implementation

Benyamin Bashari
benyamin.bashari@ucalgary.ca
University of Calgary
Calgary, Canada

Philipp Woelfel
woelfel@ucalgary.ca
University of Calgary
Calgary, Canada

ABSTRACT
The standard single-writer snapshot type allows processes to obtain a consistent snapshot of an array of \( n \) memory locations, each of which can be updated by one of \( n \) processes. In almost all algorithms, a \( \text{Scan}() \) operation returns a linearizable snapshot of the entire array. Under realistic assumptions, where hardware registers do not have the capacity to store many array entries, this inherently leads to a step complexity of \( \Omega(n) \).

In this paper, we consider an alternative version of the snapshot type, where a \( \text{Scan}() \) operation stores a consistent snapshot of all \( n \) memory locations, but does not return anything. Instead, a process can later observe the value of any component of that snapshot using a separate \( \text{Observe}() \) operation. This allows us to implement the type from fetch-and-increment and compare-and-swap objects, such that \( \text{Scan}() \) operations have constant step complexity and \( \text{Update}() \) and \( \text{Observe}() \) operations have step complexity \( O(\log n) \).

CCS CONCEPTS
• Theory of computation → Concurrent algorithms;

KEYWORDS
Shared Memory; Concurrency; Snapshot Object; Predecessor Object; Asynchrony

1 INTRODUCTION
Taking linearizable snapshots of memory is one of the most fundamental and best studied problems in the area of concurrent shared memory algorithms. The problem can be described abstractly in terms of a snapshot type, which maintains an array of \( m \) memory locations, \( A[0 \ldots m-1] \). It supports the operations \( \text{Update}(i, x) \), which changes the value of the \( A[i] \) to \( x \), and \( \text{Scan}() \), which returns the vector \((A[0], \ldots, A[m-1])\). The first linearizable snapshot implementations have been known since the 1990’s [1, 4].

Most research focuses on single-writer snapshot implementations, where \( m \) is equal to the number of processes, \( n \), and each array index is associated with a unique process, which is the only one that is allowed to update the corresponding array entry. Attiya, Herlihy, and Rachman [9] showed that implementing single-writer snapshots reduces to solving lattice agreement. They presented a lattice agreement algorithm from registers that then yields a snapshot algorithm with subquadratic (in \( n \)) worst-case step complexity. By devising a faster lattice agreement algorithm, Inoue and Chen [14] obtained a single-writer snapshot algorithm with linear step complexity. It follows from a proof technique by Jayanti, Tan, and Toueg [17] that this is optimal. Aspnes, Attiya, Censor-Hillel, and Ellen [6] proved that it is possible to break through this barrier by restricting the number of operations. They implemented a deterministic algorithm that has worst-case step complexity of \( O(\log^3 n) \), as long as the number of operations on the object is polynomial in \( n \). Subsequently, Aspnes and Censor-Hillel, gave a randomized algorithm with poly-logarithmic expected step complexity [7], and Ahad Baig, Hendler, Milani, and Travers devised a deterministic algorithm with poly-logarithmic amortized worst-case step complexity [10].

Generally, snapshot implementations from registers are not practical: Almost all of them assume that a single hardware register can store the entire snapshot array. The only exceptions we are aware of have either exponential running time [4], or do not permit concurrent \( \text{Scan}() \) operations [18]. Therefore, snapshot implementations from stronger primitives have been considered: Using single-word compare-and-swap (CAS) and fetch-and-increment (FAI) objects, Riany, Shavit, and Touitou [19] devised a single-writer algorithm with step complexity \( O(1) \) for \( \text{Update}() \) and \( O(n) \) for \( \text{Scan}() \). Jayanti [15] generalizes this to the multi-writer case, and achieves \( O(m) \) step complexity for \( \text{Scan}() \) operations, even without relying on FAI.

It is not hard to see that the \( \text{Scan}() \) operation of any snapshot object implemented from single-word objects must have step complexity \( \Omega(m) \). Therefore, the standard snapshot specification leads to inherently inefficient \( \text{Scan}() \) operations for large values of \( m \), and in particular in the common case of single-writer snapshots, where \( m = n \). The specification does not allow for a faster performance, even if a process is only interested in the value of some but not all components of the snapshot object.

1.1 Result
To improve upon this inherent complexity barrier, we implement a natural extension of the single-writer snapshot: A \( \text{Scan}() \) operation takes a snapshot of the array \( A[0\ldots n-1] \), but does not return anything. Instead, after taking a snapshot, a process can
observe each individual component \( A[i] \) of that snapshot by calling \( \text{Observe}(i) \). (Hence, the entire snapshot can be obtained using \( n \) \( \text{Observe}(i) \) calls.) Using single-word registers, CAS, and FAI objects, we obtain an implementation, where \( \text{Scan()} \) has step-complexity \( O(1) \), and \( \text{Update()} \) and \( \text{Observe()} \) have step complexity \( O(\log n) \).

Since our algorithm uses sequence numbers, only up to \( 2^{W-1} - 1 \) operations can be executed, if each base object can store \( W \) bits. This is not a restriction that has any practical impact on current 64-, or even 32-bit architectures.

Note that our snapshot algorithm is more flexible than a standard snapshot: It can simulate “full” single-writer standard snapshots (with a logarithmic performance penalty over the best known algorithms [15, 19]), but can be much more efficient, if not all components of the snapshot need to be observed. Moreover, which components a process wants to observe can be decided adaptively after a process has taken a snapshot. For example, if a snapshot represents a data structure, the search path through the data structure can depend on the actual values found. The fact that \( \text{Scan()} \) operations have constant step complexity can be useful when a process would have to take a snapshot before it even knows for certain that it will need the snapshot later. This could, for example, be the case in applications that require backups or error recovery.

### 1.2 Related Solutions

Attiya, Guerraoui, and Ruppert [8] defined a partial snapshot type, where a process can choose to scan only some of the \( m \) array components. This can be considered an obvious version of our specification, because processes must decide at the beginning of a \( \text{Scan()} \) operation, which memory locations they are interested in. The authors provide a multi-writer algorithm from registers, CAS, and FAI objects, in which scanning \( r \) array components takes \( O(r^2) \) steps in the worst-case. While \( \text{Update()} \) operations are not bounded wait-free, their amortized step complexity is bounded by the maximum interval contention, as well as the maximum number of components accessed by \( \text{Scan()} \) operations.

A more flexible multi-writer snapshot specification was proposed by Wei, Ben-David, Biellioch, Fatourou, Ruppert, and Sun [21]. Their object allows a process to take a snapshot of multiple CAS objects, and returns a handle to that snapshot. Using that handle, a process can later determine the value of any of those CAS objects at the point in time the snapshot was taken. They gave an implementation of that type from CAS objects, where each CAS() and snapshot operation takes constant number of steps. However, the step complexity of reading a single snapshotted value of a memory location grows linearly with the number of updates that may have occurred on that location, since the corresponding snapshot was taken. Hence, their algorithm is not bounded wait-free.

The authors use a version list for each CAS object, that stores the complete history of updates performed on the object. Each update is associated with a global sequence number, which is also stored in the version list. The sequence number is incremented with each \( \text{Scan()} \) operation and is returned as the handle. It can then be used to identify the latest update that was applied to a CAS object, using the version list.

Instead of version lists, our algorithm uses single-writer predecessor objects, which we implement from sequential persistent red-black trees [12] This allows processes to find the latest value that was written to a component of the snapshot object in logarithmic time, with respect to the total number of nodes stored in the red-black tree. We then devise a method for pruning the data structure from outdated values, in order to keep the size polynomial in \( n \).

The idea of associating data structure modifications with timestamps has also been used in algorithms for software transactional memory [20] and multi-version databases [22]. These algorithms are either not wait-free or allow operations to fail. Moreover, using an FAI object as a global clock to linearize operations at the time of \( \text{FAI()} \) operations is used in [5] to achieve a general method to add range queries to data structures.

As far as we know, our algorithm is the only solution to the partial snapshot problem, where snapshots of \( o(n/\log n) \) components can be obtained in \( o(n) \) time. A comparison of our implementation with the partial snapshot objects of [8, 21] is shown in Table 1.

### 2 PRELIMINARIES

We consider the standard asynchronous shared memory model with \( n \) processes with IDs 0, ..., \( n-1 \), which communicate using atomic (or linearizable) shared memory operations on base objects. We assume that each process’s ID is stored in a process-local variable \( \text{myID} \). Invocation and response of an operation \( op \) are denoted \( \text{inv}(op) \) and \( \text{rsp}(op) \), respectively.

The following base objects are relevant for our work: A read-write register supports two operations, \( \text{Write}(v) \), which changes its value to \( v \) and returns nothing, and \( \text{Read()} \), which returns the value of the register. An LL/SC object provides two operations, \( \text{LL}(\cdot) \) and \( \text{SC}(\cdot) \). An \( \text{LL}(\cdot) \) operation returns the object’s value, and an \( \text{SC}(\cdot) \) operation called by process \( p \) updates the value to \( v \), if \( p \) has previously called \( \text{LL}(\cdot) \) and no successful \( \text{SC}(\cdot) \) operation has occurred since then. An \( \text{SC}(\cdot) \) operation returns true if it succeeds to update the object’s value, and returns false otherwise. A FAI object stores an integer, initially 1, and provides an operation \( \text{FAI()} \), which increments the object’s value by 1 and returns the value before the increment. A CAS object provides an operation \( \text{CAS}(\text{old}, \text{new}) \). If the value of the object is \( \text{old} \), this operation updates the value to \( \text{new} \) and returns true, otherwise, the object remains unchanged and the operation returns false.

Let \( D_{\text{val}} \) be the domain of values stored in a snapshot component. We assume that the system provides atomic registers, FAI, and CAS objects of word size \( W \geq \max[\log |D_{\text{val}}|, 3\log n + c] \) bits for a sufficiently large constant \( c \). In order to avoid having to deal with ABAs, we will use LL/SC objects instead of CAS objects. It is trivial to replace those LL/SC objects with CAS objects and sequence numbers. Alternatively, one can use the efficient construction of LL/SC from single-word CAS by Jayanti and Petrovic [16], which has \( O(1) \) step complexity. Due to [11] we may assume w.l.o.g. that LL/SC objects have arbitrary large word size \( O(W) \).

Generally, we measure efficiency of wait-free shared memory algorithms in terms of the number of shared memory steps executed. Thus, step complexity ignores local computation time, which does not include shared memory steps. Time complexity, on the other hand, measures local and shared memory steps of processes (assuming a standard word-RAM model).
As a building block for our main algorithm, we use the destination array of Blelloch and Wei [11]. A destination array stores a sequence of values, and supports the operations Read() and Copy(). Operation Read(i) takes as argument an integer i ∈ {1, ..., n}, and returns the value of the i-th component of the array. Operation Copy(R) takes as argument a reference R to a register, and if process p calls that operation, it changes the value of the p-th component of the array to the value of register R. Blelloch and Wei [11] show that a linearizable and wait-free destination array can be implemented from O(n^2) single-word CAS objects and registers in such a way that each Read() and each Copy() operation can be executed in a constant number of steps.

**Specification of the New Algorithms.** In this paper we present linearizable single-writer predecessor and adaptive partial snapshots objects. The former is used as a building block for the latter, but may be of independent interest. In the following we provide the sequential specifications of the underlying types.

The predecessor type maintains a set of pairs, each comprising a key and a value. The domain of keys must be totally ordered.

The predecessor type provides four operations. An Insert(k, v) operation inserts a pair with key k and value v, provided that the data structure does not contain a pair with key k. In that case, the Insert() operation succeeds and returns true, otherwise, it fails and returns false. Operations Remove(k), Pred(k), and Succ(k) take as a single argument a key k. Each of them fails and returns false if the data structure contains no pair with key k. Otherwise, Remove(k) removes the pair with key k from the data structure, and Pred(k) and Succ(k) return the pairs with greatest key smaller than k and smallest key larger or equal than k, respectively, if such pairs exist. If not, these operations fail and return false.

We call operations Insert() and Remove() update operations, and all other operations query operations.

A predecessor object is single-writer, if there is only one dedicated process that is allowed to perform updates. We will consider single-writer predecessor objects with bounded capacity Δ, which informally means that at most Δ elements can be stored in the data structure at any point. Since the object is single-writer, it is uniquely determined at the point of invocation of an update operation, whether that operation will be successful or not. (We call an incomplete update operation successful, if it must be successful in any extension of the execution in which it completes.)) Bounded capacity Δ formally means that at any point in time the number of invocations of successful Insert(t) operations minus the number of responses of successful Remove() operations is at most Δ.

The adaptive partial snapshots type stores an n-component array and supports three operations Update(v), Scan(), and Observe(k). An Update(v) operation called by process p changes the value of the p-th component to v, and returns nothing. Method Scan() does not return anything, and its behaviour is only defined in terms of method Observe(k). A process is only allowed to call an Observe(k) operation after it has performed at least one Scan() call, and an Observe(k) call by process p returns the value that the k-th component of the object had at the point of p’s latest preceding Scan() operation.

### Table 1: Comparison of Partial Snapshot Objects

| Algorithm | Step complexity | Update complexity | Base objects | Number of primitives |
|-----------|----------------|-------------------|--------------|---------------------|
|           | Reading r locations | Unbounded | n-word CAS and FAI | Unbounded |
| [8]       | O(r^2) | O(1) | single-word CAS | Unbounded |
| [21]      | Unbounded | Unbounded | single-word CAS and FAI | O(n log n) |
| This work | O(r log n) | O(log n) | single-word CAS and FAI | O(n^3 log n) |

All algorithms use sequence numbers that increment with update or scan operations, and store them in base objects. Therefore, these algorithms implicitly assume that the number of scan or update operations is bounded by 2^W, where W is the word size.\(^1\) It is safe to assume that this assumption will never be violated on current 64-bit architectures.

### 3 SINGLE-WRITER PREDECESSOR ALGORITHM

In this section we present our linearizable and wait-free single-writer implementation of the predecessor type from registers and CAS objects. First, we will show that the sequential balanced red-black tree of [12] can be used in a concurrent system (i.e., is linearizable), as long as there is only one process that performs update operations. This concurrent red-black tree almost immediately yields a single-writer predecessor object: it natively supports the operations Insert(t) and Remove(), and adding operations Pred() and Succ() is straightforward. However, in infinite executions, this algorithm may need an unbounded number of nodes. In Section 3.2 we will use memory reclamation to bound the space of our concurrent data structure.

#### 3.1 The Basic Algorithm

Driscoll, Sarnak, Sleator, and Tarjan [12], present a technique called node-copying to make linked data structures persistent. We first describe a basic version of this technique that can be applied to any binary search tree (BST) implementations, where each node store only pointers to its children (i.e., there are no parent pointers):

A dedicated variable R stores a pointer to the root r of the tree. An update operation does not modify any nodes of the data structure. Instead, it adds copies of all nodes that need to be modified, as well as a new root r’ and, finally changes the pointer R so that it points to r’ instead of r. To be more precise, suppose the set of nodes reachable from r form a conventional BST T_r. Let T_2 be the BST obtained by applying a conventional update operation to T_r (e.g., an insertion). Let S be the set of nodes in T_2, that are either added

---

\(^1\)We believe that in [8] unbounded sequence numbers can be avoided by using LL/SC instead of CAS.
or modified by this update operation, and \( S' \) the parent-closure of \( S \) (i.e., if \( v \) is in \( S' \), then the parent of \( v \) is also in \( S' \)). Instead of modifying the nodes in \( S \), we create a copy \( v' \) of each node \( v \) in \( S' \). Each field of \( v' \) has the same value as the corresponding field in \( v \), except that a pointer to a node \( u \) in \( S' \) is replaced with a pointer to the copy \( u' \) of that node. Since \( S' \) is parent-closed, the root \( r' \) of \( S' \) is also copied into a node \( r' \). It is easy to see that the nodes reachable from \( r' \) now form a red-black tree that is equivalent to \( T_2 \). Hence, to complete the update operation, it suffices to replace the value of \( R \) with a pointer to \( r' \).

To perform a query operation on the persistent data structure, a process simply reads the pointer \( R \) to obtain a root \( r \), and then performs the same operations as it would in the conventional BST algorithm, using \( r \) as a root.

For the purpose of concreteness, we will now consider a red-black tree [13]. Driscoll, Sarnak, Sleator, and Tarjan have applied the node copying technique described above to that data structure to obtain a persistent red-black tree, where each of the operations \( \text{Insert}(x) \), \( \text{Remove}(x) \) and \( \text{Find}(x) \) take \( O(\log m) \) steps [12], where \( m \) is the number of elements stored in the tree. It is straightforward to augment the data structure with query operations \( \text{Pred}(x) \) and \( \text{Succ}(x) \) so that all operations have time complexity \( O(\log m) \). Thus, we obtain an implementation of a persistent sequential predecessor type with the same asymptotic time complexity.

We can now use that persistent predecessor implementation in a shared memory system, by storing \( R \) and each node of the data structure in an atomic register. We will allow only one process, \( p_w \), to perform update operations, but all processes are allowed to execute query operations. Observe that if at some point \( t \) pointer \( R \) points to a root \( r \) in the data structure, all nodes reachable from \( r \) form the BST that was obtained as a result of the update operation that wrote the pointer to \( r \) into \( R \). None of these nodes can change after point \( t \). Hence, if a query operation reads the pointer to \( r \) from \( R \), then it will visit exactly the same nodes that were visited in the sequential case. Similarly, an update operation by process \( p_w \) does initially not make changes to any reachable nodes, and all tree modifications will become visible to other processes only when \( p_w \) changes the root pointer, \( R \), to point to the new root copy \( p_w \) created. Hence, it is easy to see that each update operation can linearize with the write to \( R \), and each query operation can linearize with the read of \( R \). It follows that this object is linearizable, provided that only one process can perform update operations (in fact, it is linearizable as long as no two update operations are concurrent).

### 3.2 Recycling Outdated Nodes

During each update operation on the concurrent red-black tree, described in the previous section, \( p_w \) makes copies of up to \( \Theta(\log m) \) nodes [12], and needs to allocate space for the registers storing them. Thus, in unbounded executions, an infinite number of registers is needed. In the following, we apply a memory reclamation technique to bound the space. We add a \( \text{Recycle}() \) method to the object, whose purpose is to remove the nodes from the tree that cannot be accessed anymore by any process. The registers storing these nodes’ information can then be reused for future nodes.

We will only consider single-writer predecessor objects with bounded capacity \( \Delta \). In our snapshot application, \( \Delta \leq 3n \).

For the ease of discussion and readability of the pseudocode (see Algorithm 1), we assume that \( p_w \) has access to a method \( \text{Allocate}(x) \), which allocates a new node and returns a reference to it, and a method \( \text{Deallocate}(x) \), which deallocates a node \( x \) that \( p_w \) previously allocated. We will show that at any point in time, there are at most \( \lambda = O(nA \log \Delta) \) nodes that \( p_w \) has allocated but not deallocated. Thus, implementing methods \( \text{Allocate}(x) \) and \( \text{Deallocate}(x) \) with bounded memory is straightforward, by having \( p_w \) maintain a local pool of \( \lambda \) registers, one for each node. Our algorithm guarantees that when a process accesses a node \( v \), then at that point \( v \) has been allocated but not yet deallocated.

Consider an execution. Let \( r_0 \) be the initial root of the red-black tree pointed to by \( R \), and let \( r_1 \) be the root pointed to by \( R \) after the \( i \)-th update operation. Let \( \text{reachable}(r_i) \) denote the set of nodes reachable from \( r_i \). A node \( v \) is \textit{outdated} at point \( t \), if at that point \( R \) points to a root \( r_k \), and there exists \( i \in \{0, \ldots, k - 1\} \), such that \( v \in \text{reachable}(r_i) \setminus \text{reachable}(r_k) \). Outdated nodes are candidates for deallocation. However, it is possible that an outdated node \( v \) may still be accessed by a process \( q \), if \( v \) is reachable from a root \( r \), and \( R \) pointed to \( r \) when \( q \) read that pointer at the beginning of its query operation. To prevent \( v \) from being deallocated in such a situation, process \( q \) will initially protect a root \( r \), when it reads the pointer to \( r \) from \( R \) at the beginning of its query operation. It does so by storing the address of \( r \) in a register \( \text{protectedRoot}[q] \), where \( \text{protectedRoot} \) is an array with one entry for each process. We say a node is protected, if it is reachable from a root stored in some array entry \( \text{protectedRoot}[q] \), \( j \in \{0, \ldots, n - 1\} \). The \( \text{Recycle}() \) method will then deallocate only outdated, but not protected nodes.

Using only registers, it is difficult to protect nodes this way: Process \( q \) may fall asleep immediately after reading the address of the current root \( r \) from \( R \), and then only wake up again when \( r \) has already been deallocated. It is then too late to protect \( r \) by writing it into \( \text{protectedRoot}[q] \). Sophisticated techniques to deal with that in constant time and using only registers have been described in [2, 3]. An easier way is to use a destination array for \( \text{protectedRoot} \). (The main motivation for the definition of the destination array in [11] has been to solve the same type of “protection” problem.) At the beginning of its query operation, process \( q \) calls \( \text{protectedRoot}[q].\text{Copy}(R) \). This copies the address of the root \( r \) pointed to by \( R \) into \( \text{protectedRoot}[q] \), and thus protects all nodes reachable from \( r \). As mentioned above, in a \( \text{Recycle}() \) call, \( p_w \) now only needs to find all outdated nodes (i.e., those that are not reachable from the current root pointed to by \( R \), and remove those that are not protected.

It remains to show that this can be done without decreasing the worst-case time complexity of update and query operations. For query operations this is trivial, as only one \( \text{Copy}() \) operation is added.

We now show how to modify update operations, and how to implement the \( \text{Recycle}() \) method. The updater, process \( p_w \), maintains a (local) list \( \text{Outdated} \) of outdated nodes that have not yet been deallocated. Initially, \( \text{Outdated} \) is an empty list. Suppose that at the end of an update operation by process \( p_w \), the root pointed to by \( R \) changes from \( r \) to \( r' \). Then \( p_w \) computes \( \text{reachable}(r) \setminus \text{reachable}(r') \) and adds all nodes in that set to \( \text{Outdated} \). Note that each node in \( \text{reachable}(r) \setminus \text{reachable}(r') \) is in the parent closure of the set of
nodes that get modified by a standard red-black tree update operation. Since that red-black tree contains at most $\Delta$ nodes, it is easy to see that

$$|\text{reachable}(r) \setminus \text{reachable}(r')| \leq c \cdot \log \Delta \quad (1)$$

for some constant $c$. Moreover, the set $\text{reachable}(r) \setminus \text{reachable}(r')$ can be computed in time $O(\log \Delta)$.

Every $n\Delta$ update operations, process $p_w$ begins a new Recycle() operation. The total number of steps required to complete such a Recycle() call will be $O(n\Delta \log \Delta)$, and with each update operation, $p_w$ contributes $O(\log \Delta)$ steps to an ongoing Recycle() call. In a Recycle() call, $p_w$ first renames outdated to outdated', and then sets outdated to an empty list. (This way, future update operations will fill the set outdated again, while the Recycle() call can process outdated'. And if, while that Recycle() call is ongoing, a process $p$ starts protecting a new node $v$ by copying its root to $\text{protectedRoot}[p]$, then $v$ will not be in outdated', and thus will not get deallocated.) Then, $p_w$ initializes a local empty list protected, and reads all roots stored in $\text{protectedRoot}[i]$ for each $i \in \{0, \ldots, n-1\}$, and for each such root $r$, adds reachable($r$) to protected. Finally, $p_w$ deallocates all nodes in outdated' that are not in protected, and append list protected to outdated', so that the protected nodes can be reconsidered in the next Recycle() operations.

Analysis. At any point in time the predecessor object stores a set of size at most $\Delta$. Hence, from any root stored in $\text{protectedRoot}[i]$, at most $\Delta$ nodes can be reached, and thus

$$|\text{protected}| \leq \Delta \cdot n, \quad (2)$$

and protected can be computed in time $O(|\text{protected}|) = O(\Delta n)$.

We will now argue that

$$|\text{outdated}| \leq (c \log \Delta + 1)n\Delta \quad (3)$$

at all times, where $c$ is the constant from (1). Clearly, this is true initially and at the beginning of each Recycle() call, when outdated is set to an empty list. In the interval during which a Recycle() call completes (or until the first Recycle() call is invoked), $n\Delta$ update operations are executed. Hence, by (1), at most $cn\Delta \log \Delta$ nodes become outdated. In addition, by (2), at the end of the Recycle() call at most $\Delta n$ nodes are added from protected to outdated. Thus, once the Recycle() call terminates, $|\text{outdated}| \leq \Delta n + cn\Delta \log \Delta$, and thus (3) is true.

By definition, each node in the data structure is either outdated or reachable from the root pointed to by $R$. Since at most $\Delta$ nodes are reachable from that node, it follows from (3) that at any point $p_w$ needs to have only $\lambda = O(n\Delta \log \Delta)$ nodes allocated. Thus, a pool $P_j$ of $\lambda$ registers suffices to store all nodes.

Since $p_w$ needs only $\lambda$ nodes in its entire pool (unallocated and allocated nodes), it is easy to see that $p_w$ can compute the set difference $S_1 \setminus S_2$ of two sets $S_1, S_2 \subseteq P_j$ (given as linked lists) in time $O(\lambda) = O(n\Delta \log \Delta)$ using the standard lookup-table technique. In particular, at the end of its Recycle() method, process $p_w$ can compute all nodes that are in outdated' but not in protected in time $O(n\Delta \log \Delta)$.

To summarize, a complete execution of the Recycle() method takes $O(n\Delta \log \Delta)$ time. In each update operation, process $p_w$ computes the set of $O(\log \Delta)$ nodes that have become unreachable (see (1)) in $O(\log \Delta)$ time. Then it contributes sufficiently many steps towards a (new or ongoing) Recycle() operation, so that the Recycle() method completes during $n\Delta$ update operations by $p_w$. Since the Recycle() method takes $O(n\Delta \log \Delta)$ time, there is a constant $\lambda$, such that $p_w$ needs to contribute at most $\lambda \cdot \log \Delta$ steps to the Recycle() method during each update. Hence, each update operation takes time $O(\log \Delta)$, and thus it also comprises only $O(\log \Delta)$ shared memory steps.

As discussed, the algorithm needs to store $\lambda = O(n\Delta \log \Delta)$ nodes in registers. In addition it requires a destination array of size $n$, which can be implemented from $O(n^2)$ registers and CAS objects (see Section 2).

**Theorem 1.** A wait-free linearizable single-writer predecessor object with bounded capacity $\Delta$ can be implemented from $O(n\Delta \log \Delta + n^2)$ single-word CAS objects and registers, such that each update and query operation has time and step complexity $O(\log \Delta)$.

A full proof is omitted due to space constraints.

**Algorithm 1:** Single-Writer Predecessor Algorithm

```plaintext
Shared:
   Register R = \bot
   Destination Array protectedRoot
Local for Writer:
   List outdated

Function update()
   Let r be the root pointed to by R
   Perform the corresponding update operation on R as in the sequential implementation in [12], allocating new nodes using method Allocate().
   Let r' be the new root pointed to by R
   Add all nodes in reachable(r) \ reachable(r') to outdated.
   Every n\Delta Update() calls, start a new Recycle() method
   Contribute $\lambda \cdot \log \Delta$ steps towards an ongoing Recycle() call, where $\lambda$ is a sufficiently large constant.

Function Query()
   protectedRoot.Copy(R)
   Perform the corresponding query operation on protectedRoot.Read(myID) as in the sequential implementation in [12]

Function Recycle()
   Rename outdated to outdated'
   Let outdated and protected be new empty lists
   for i \in \{0, \ldots, n-1\} do
      if r \not= \bot then
         protected ← protected \cup reachable(r)
      For each node x in outdated', such that x \not= protected, perform Deallocate(x).
   Append protected to outdated
```

4 ADAPTIVE PARTIAL SNAPSHOT ALGORITHM

4.1 The Basic Algorithm

In this section we present a basic linearizable implementation of the adaptive partial snapshot type, see Algorithm 2. The algorithm uses, for each process $p$, a predecessor object versions[$p$]. Each
Update() by process $p$ adds an element to $\text{versions}[p]$, and the time complexity of Update() and Observe() operations is dominated by operations on the predecessor objects. This basic algorithm never removes elements from the predecessor objects. In order to obtain our desired space and step complexity bounds, we will later show how we can prune outdated elements from the predecessor object. This will allow us to use a predecessor object of bounded capacity $\Delta = 3n$.

Algorithm 2: Snapshot Algorithm without Pruning

AlGORITHM 2 uses a shared FAI object $clk$ in addition to the predecessor objects $\text{versions}[p]$ for each process $p$. The value of $clk$ is incremented at least once for each Update() and Scan() operation. Each Update() and Scan() operation is associated with exactly one such increment (even though multiple ones can happen during such an operation execution), and the operation linearizes at the point of that increment. I.e., each Update() and Scan() operation is associated with a unique value $x$, where $clk$ is incremented from $x$ to $x + 1$ at the linearization point of that operation.

We will first discuss a very simple high-level idea, which leads to an incorrect algorithm, and then show how to fix it in order to arrive at Algorithm 2. In an Update$(val)$ operation, the calling process, $p$, increments the FAI object $clk$ from $x$ to $x + 1$, and then inserts the pair $(x, val)$ into $\text{versions}[p]$. In a Scan() operation, a process $q$ determines the current value of $clk$ (also by executing an FAI$(\cdot)$), and stores it in $lastScan[q]$. (For this basic algorithm, $lastScan[q]$ can be a local variable. But for the advanced algorithm we will need it to be an LL/SC object.) In a later Observe$(\cdot)$ operation, $q$ can then determine the last value $val$ that was written to component $p$ of the array, by calling $\text{versions}[p].\text{Pred}(sTime)$, where $sTime$ is the value of $lastScan[q]$. This call returns the pair $(x, val)$ stored in the predecessor object, such that $x$ is the largest key less than or equal to $sTime$.

However, this simple approach is not linearizable: Suppose that during its Update$(val)$, process $p$ executes FAI$(\cdot)$, but then before it inserts the pair $(x, val)$ into its predecessor object $\text{versions}[p]$, some process $q$ performs a Scan$(\cdot)$ operation, during which it fetches $x' > x$ from $clk$. After that, $q$ calls Observe$(\cdot)$, and determines the predecessor of $x'$ in $\text{versions}[p]$. Whether or not $q$ observes the value $val$ depends on when $p$ inserts the pair $(x, val)$ into $\text{versions}[p]$. Thus, the relative order of $p$'s Update$(val)$ and $q$'s Scan$(\cdot)$ would now have to be determined by the execution of $q$'s Observe$(\cdot)$ call relative to $p$'s insert of $(x, val)$ into the predecessor object. In particular, $q$ may perform two subsequent Observe$(\cdot)$ calls that return different values, which is incorrect.

To deal with that problem, we let any process calling Observe$(\cdot)$ help process $p$ linearize a possible pending Update$(\cdot)$ operation. This is facilitated by method HelpUpdate$(\cdot)$ and an auxiliary LL/SC object $lastUpdate[p]$ for each process $p$. To indicate that process $p$ has started performing Update$(val)$, it writes the pair $\langle \bot, val \rangle$ into that variable (using a pair of $LL(\cdot)$ and $SC(\cdot)$ operations). A process $q$ that wants to observe the value of $p$'s array component following a Scan$(\cdot)$, can retrieve that value either from the predecessor object $\text{versions}[p]$, or, if $\text{versions}[p]$ is outdated, from $lastUpdate[p]$. (How a process determines if the value is outdated, will be explained in the low level description.)

But $p$'s Update$(val)$ may not yet have been linearized, if $p$ has not yet incremented $clk$. Therefore, during its Observe$\langle \cdot \rangle$ operation, $q$ increments $clk$ itself from $x$ to $x + 1$ and tries to communicate that to other processes by trying to store the pair $(x, val)$ into $lastUpdate[p]$, using an SC$(\cdot)$ operation. If successful, then $p$'s Update$(val)$ operation can linearize with $q$'s increment of $clk$. And otherwise, some other process has already helped $p$'s Update$(\cdot)$ linearize.

4.2 Pruning Predecessor Objects

In Algorithm 2, the number of pairs stored in the predecessor objects increases with the number of Update$(\cdot)$ operations, and thus needs predecessor objects with unbounded capacity.

In order to be able to use a predecessor objects with bounded capacity, we remove unnecessary elements. The corresponding pseudo-code is show in Algorithm 3. We will need an auxiliary method $\text{HelpScan}(\cdot)$. Its implementation guarantees that if at the invocation of a $\text{HelpScan}(\cdot)$ call a Scan$(\cdot)$ call by some process $p$ starts, then by the time the $\text{HelpScan}(\cdot)$ call responds, the value of $clk$ that is associated with the Scan$(\cdot)$ operation will have been written to $lastScan[q]$. This helping mechanism is implemented in essentially the same way as that of $\text{HelpUpdate}(\cdot)$.

We now describe how a process $p$ can remove pairs from its predecessor object, $\text{versions[p]}$, which are not needed anymore. To that end, $p$ computes a set of required keys, which it will not remove from $\text{versions[p]}$. Consider an arbitrary point $t$ in the execution. A
pair \((k, v)\) in \(versions[p]\) can be safely removed after point \(t\), if no \(\text{Observe}(p)\) operation that responds after \(t\) needs to return \((k, v)\).

Conversely, a key is required, if it may at any point after \(t\) be the predecessor of some value that is then stored in \(\text{lastScan}(q)\) of some process \(q\).

Process \(p\) can compute the set of required keys as follows: First, at some point \(t_0\) it determines the largest key, \(\text{maxKey}\), stored in \(versions[p]\). Observe that then \(\text{clk} > \text{maxKey}\) at that point, and also at any later point. Next, process \(p\) calls \(\text{HelpScan}(q)\) for each process \(q\), and then reads a value \(v_q\) from \(\text{lastScan}(q)\). Let \(V\) be the set of values \(v_q\neq\perp\), for each process \(q\). It is not hard to see that at any point after \(p\) has read \(v_q\) from \(\text{lastScan}(q)\), the value of that object is either still \(v_q\), or \(\perp\) or an integer of value at least \(\text{clk}\).

Since \(\text{clk} > \text{maxKey}\) at any point after \(t_0\), it suffices if the set of required keys contains at least all predecessors of values larger than \(\text{maxKey}\), and of the keys in \(V\). Let \(S\) be the set of all predecessors of values in \(V\) that are in \(versions[p]\). Since \(\text{maxKey}\) is in \(versions[p]\), the predecessors of values larger than \(\text{maxKey}\) have a value of at least \(\text{maxKey}\). Hence, \(p\) determines as the set of required keys as the set of keys \(k\), where \(k \in S\) or \(k \geq \text{maxKey}\).

To prune its predecessor object, a process \(p\) can call method \(\text{Prune}(\cdot)\). In lines 62–67 of that method, \(p\) computes the value \(\text{maxKey}\) and the set \(S\) exactly as described above. Then, as indicated in the last line of the method, \(p\) removes all keys from \(versions[p]\) that are not in \(S\) and that are smaller than \(\text{maxKey}\).

Clearly, the step complexity of method \(\text{Prune}(\cdot)\) is super-linear in \(n\). But we can distribute the total work of a \(\text{Prune}(\cdot)\) call over \(n\) \(\text{Update}(\cdot)\) calls. We maintain the invariant that \(p\)'s predecessor object contains at most \(3n\) pairs at any point in time, and thus we can use a predecessor object with bounded capacity \(\Delta = 3n\). Once the predecessor object contains \(2n\) elements, process \(p\) begins distributing the total work of a single \(\text{Prune}(\cdot)\) call over its next \(n\) \(\text{Update}(\cdot)\) operations. This way, \(p\) can remove at least \(n\) elements from \(versions[p]\), while \(n\) new elements are added.

We will now argue that the total amount of work of each \(\text{Prune}(\cdot)\) call is \(O(n \log n)\). Thus, to preserve the \(O(\log n)\) step complexity for all snapshot operations, it suffices if during each \(\text{Update}(\cdot)\), \(p\) devotes \(O(\log n)\) steps to this recycling method.

Let \(K\) be the set of keys with value at most \(\text{maxKey}\) that are in \(versions[p]\) at any point after \(t_0\), which is when \(p\) determines \(\text{maxKey}\). Clearly, no such keys are added to this set after \(t_0\), so \(|K| = O(n)|\). Process \(p\) can find all keys in \(K\) by first determining the successor of \(-1\), which is the smallest such key, and then following the chain of successors, until \(\text{maxKey}\) is reached. Since the step complexity of computing a successor in \(versions[p]\) is \(O(\log n)\), the total work complexity of \(O(n \log n)\) for \(\text{Prune}(\cdot)\) follows.

### 4.3 Linearizability Proof

In the following we will prove that any execution \(E\) is linearizable. Since our algorithm is wait-free, we assume w.l.o.g. that in \(E\) all operations complete. (If they don’t we can construct a completion of the execution, by letting processes with pending operations take steps in an arbitrary order, until their operations have completed.)

In Section 4.3.1, we define a mapping \(\text{lin}\) that maps each operation \(op\) in \(E\) to a point in time. Then, in Section 4.3.2, we show that for every \(\text{Update}(\cdot)\), \(\text{Scan}(\cdot)\) and \(\text{Observe}(\cdot)\) operation, \(\text{lin}(op) = \text{lastScan}(op)\). Similarly, if \(\text{lastScan}(op)\) changes in \(\text{lin}(op)\), \(\text{lin}(op) \leq \text{lastScan}(op)\). In Sections 4.3.3 and 4.3.4 we prove that if we order all operations by \(\text{lin}\), the resulting sequential history is valid (w.r.t. the specification of the adaptive partial snapshot type). Let \(p\) be a process, \(up\) an \(\text{Update}(\cdot)\) operation by \(p\) and \(sc\) a \(\text{Scan}(\cdot)\) operation by \(p\). Let \(t(up)\) be the point in time at which process \(p\) performs \(\text{lastUpdate}[p]\). \(\text{Scan}(\cdot)\) in line 22 of \(up\). Similarly, let \(t(sc)\) be the point in time at which process \(p\) performs \(\text{lastScan}[p]\). \(\text{Scan}(\cdot)\) in line 53 of \(sc\). If \(\text{lastUpdate}[p]\) changes in \((t(up), \infty)\), then we let \(t'(up)\) be the first point when this happens in that interval, and otherwise we let \(t'(up) = \infty\). Similarly, if \(\text{lastScan}[p]\) changes in \((t(sc), \infty)\), \(t'(sc)\) denotes the first point when this happens in that interval, and otherwise \(t'(sc) = \infty\).

#### 4.3.1 Linearization Points

Below we will prove two lemmas that facilitate the definition of mapping \(\text{lin}\). Lemma 2 shows that during each \(\text{Update}(\cdot)\) operation by process \(p\), \(\text{lastUpdate}[p]\) changes exactly once from \(\perp\) to a positive integer fetched from \(\text{clk}\). It is possible that \(p\) changes \(\text{lastUpdate}[p]\) when it calls \(\text{HelpUpdate}(\cdot)\) during its own \(\text{Update}(\cdot)\), or some other process may do that during a \(\text{HelpUpdate}(\cdot)\) call.) Lemma 3 proves a symmetric statement for \(\text{lastScan}[p]\). Due to space constraints, proofs of these two lemmas are omitted in this extended abstract.
LEMMA 2. Let p be a process that executes first an Update(y) operation up and later an Update(y') operation up', and assume p executes no Update() in-between up and up'. Then
(a) at point t(up) process p changes lastUpdate[p].key to 1, (b) t'(up) is before the point at which process p reads lastUpdate[p] in line 24 of up.
(c) at point t'(up) some process performs a successful lastUpdate[p].SC(x,y) operation in line 30 of a HelpUpdate(p) call, for some x ∈ N, and (d) lastUpdate[p] remains unchanged throughout (t'(up), t(up'),)
LEMMA 3. Let p be a process that executes first a Scan() operation sc and later a Scan() operation sc', and no Scan() operation between sc and sc'.
(a) Process p changes lastScan[p] to ⊥ at t(sc), (b) t'(sc) < rsp(sc).
(c) at point t'(sc) some process performs a successful lastScan[p].SC(x) operation in line 59 of a HelpScan(p) call, where x ∈ N, and (d) lastScan[p] remains unchanged throughout (t'(sc), t(sc')).

Now consider a process p that executes an Update() operation up, or a Scan() operation sc. If a process p executes an Update() operation up, then by Lemma 2 (c), lastScan[p].key changes due to a shared memory operation executed at some point t'(up) in a HelpUpdate(p) operation. If p executes a Scan() operation sc, then by Lemma 3 (c), lastScan[p] changes at t'(sc) during a HelpUpdate(p). We call these HelpUpdate(p) and HelpScan(p) operations helpup and helpsc, respectively.

Let op be an Observe(), Scan() or Update() operation. Recall that we consider only complete executions. If op is an Observe() operation, then we define lin(op) = rset(op). If it is an Update() or Scan(), then lin(op) is the point at which clk.FAI() is executed in helpop (in line 29 or line 58). We say that operation op linearizes at point lin(op).

4.3.2 Correctness of Mapping lin.

CLAIM 4. Let op an Update() or Scan() operation executed by process p. Then, t(op) < lin(op) < t'(op).

Proof. We prove this claim for the case that op is an Update() operation. A symmetric proof applies, when op is a Scan(). Let q be the process calling helpop. By definition, op linearizes when q performs clk.FAI() in line 29 of helpop. By Lemma 2 (c), at point t'(op) process q executes line 30 of helpop. So lin(op) < t'(op). Thus, it remains to show that t(op) < lin(op).

Let t1 be the point at which q performs lastUpdate[p].LL() in line 27 of helpop. Since q's lastUpdate[p].SC at t'(op) in line 30 of helpop succeeds, there is no successful lastUpdate[p].SC() throughout [t1, t'(op)]. Also, based on Lemma 2 (a), there is a successful lastUpdate[p].SC() at t(op), and by definition t(op) < t'(op). Therefore, t(op) < t1 and since t1 < lin(op), it follows that t(op) < lin(op).

Claim 5. Let op be an Update(), Observe(), or Scan() operation. Then, inv(op) ≤ lin(op) ≤ rset(op).

Proof. If op is an Observe() operation, then lin(op) = rset(op).
If it is an Update() or Scan() operation, then t(op) < lin(op) < t'(op) by Claim 4. Moreover, inv(op) < t(op) by definition, and t'(op) < rset(op) by Lemmas 2 (b) and 3 (b), respectively. Therefore, inv(op) ≤ lin(op) ≤ rset(op).

Assume that process p executes a Scan() and later completes an Observe(q) operation ob that returns a value v. (Recall that p is not allowed to call Observe() until it completed a Scan().) Then the sequential specification dictates that either no Update(⊥) by process q linearizes before ob and v = ⊥, or the latest Update() by process q that linearizes before ob uses parameter v. In fact, since neither Update() nor Scan() returns anything, execution E is linearizable, if this is true for every Observe() operation. We will show in Section 4.3.4 that this is the case. To facilitate that proof, we will first show in Section 4.3.3, that if a process q executes an Update(u) operation, then once it has added u to versions[q], it won’t remove u from that predecessor object until all Observe() operations that might have to return u (for linearizability), have responded.

4.3.3 Correctness of the Memory Reclamation. Recall that for each Update() or Scan() operation op by process p, at t'(op) some integer is written to lastUpdate[p] or lastScan[p]. Let val(op) denote that integer. Claim 6 below states that the linearization order of Update() and Scan() operations op matches the order of integers val(op).

CLAIM 6. Let each of op1 and op2 be an Update() or Scan() operation. Then, lin(op1) < lin(op2) if and only if val(op1) < val(op2).

Proof. The definition of lin together with Lemmas 2 (c) and 3 (c), imply that val(op1) is the value of clk just before lin(op1) and val(op2) is the value of clk just before lin(op2). There is a clk.FAI() operation at lin(op1) and lin(op2). Therefore, by that fact that the value of clk is increasing, lin(op1) < lin(op2) if and only if val(op1) < val(op2).

We now show that during an Update(x) operation op by process p, the pair (val(op), x) gets inserted into versions[p].

CLAIM 7. If process p completes an Update(x) operation up, then in line 25 of up it inserts (val(up), x) into versions[p].

Proof. Let t1 be the point in time at which process p reads lastUpdate[p] in line 24 of up. By Lemma 2 (b), t'(up) < t1. Also, by Lemma 2 (b), lastUpdate[p].key does not change throughout (t'(up), rset(up)]. Therefore, the value assigned to uTime in line 24 of up is val(up). So process p inserts (val(up), x) into versions[p].

The next claim shows that if a Scan() operation sc by p linearizes before the invocation of a HelpScan(p) operation hs by q, then, val(sc) is written to lastScan[p] before rset(hs). Therefore, if q reads lastScan[p] after it performed hs, then it either reads val(sc) or a value that is written to lastScan[p] by p during a Scan() operation that linearizes after inv(hs). This shows that the HelpScan() method prevents processes from reading outdated values from lastScan[θ...n−1].

CLAIM 8. Suppose process p executes a Scan() operation sc, and process q executes a HelpScan(p) operation hs, such that lin(sc) < inv(hs). Then, t'(sc) < rset(hs).
Proof. For the purpose of a contradiction, assume $t'(sc) \geq rs(p, hs)$. By definition and Lemma 3(a), lastScan[p] = ⊥ throughout $(t(sc), t'(sc))$. Also, by Claim 4, $(t(sc) < lim(sc) < t'(sc))$. Therefore, lastScan[p] = ⊥ throughout the execution of hs. So process p reads ⊥ from lastScan[p] in line 56 of hs, and changes lastScan[p] in line 59 of hs to a positive integer fetched from clk. This contradicts the fact that lastScan[p] = ⊥ throughout the execution of hs.

Finally, we use the above to prove that elements added to versions[0...n-1] are not removed prematurely, i.e., our memory reclamation works correctly. Suppose a process p adds a pair $(key, v)$ to versions[p] during an Update(v) operation. The memory reclamation scheme ensures that the pair remains in the predecessor object until every Observe operation that may need to return v has completed. Note that an Observe operation by process q needs to return v if q's latest preceding Scan[p] linearizes after p's Update(v) and no Update() by p linearizes in between.

Observation 9. Suppose process p executes an Update() operation up during which it inserts a pair $(key, v)$ into versions[p]. Let sc be a Scan() operation by some process q, such that lin(up) < lim(sc), and no Update() operation by p linearizes between lin(up) and lim(sc). If at point $t^*$ > rs(p, hs) the pair $(key, v)$ is in versions[p], and some process executes versions[p].Pred(val(sc)), then that operation returns $(key, v)$.

Proof. By Claim 7, $key = val(up)$. Therefore, by the assumption of this observation, $(val(up), v)$ is in versions[p] at $t^*$. Also, by Claim 6, since lin(up) < lim(sc), we have val(up) < val(sc). Hence, if there is no pair E in versions[p] at $t^*$, such that val(up) < E.key < val(sc), then, versions[p].Pred(val(sc)) operation returns $(key, v)$. Suppose at point $t^*$ there is a pair E in versions[p], such that val(up) < E.key < val(sc). By Claim 7, this pair is inserted into versions[p] during an Update() operation up' by p, such that $E'.key = val(up')$. Therefore, val(up) < val(up') < val(sc). By Claim 6, lin(up) < lin(up') < lim(sc). This contradicts the fact that no Update() operation by p linearizes between lin(up) and lim(sc).

Lemma 10. Suppose process p executes an Update() operation up and inserts a pair $(key, v)$ into versions[q] in line 25. Let q be a process and sc a Scan() operation by q, such that lin(up) < lim(sc), and no Update() operation by p linearizes between lin(up) and lim(sc). Let $t^* > max(lin(sc), rs(p, hs))$, such that q invokes no Scan() in $[rs(p, hs), t^*)$. Then $(key, v)$ is in versions[p] at $t^*$.

Proof. Let $t_1$ be the point in time at which q inserts $(key, v)$ into versions[q]. We need to prove that this pair is not removed from versions[q] throughout $(t_1, t^*)$. For the purpose of contradiction, assume p removes $(key, v)$ from versions[q] in line 72 a Prune() operation pr at $t_{mv}$ < $t^*$. Let $t_2$ be the point when p executes its versions[p].Pred(∞) operation in line 62 of its Prune() call pr, and let $key'$ be the key that this operation returns, and which it assigns to maxKey. Process p removes pair $(key, v)$ from versions[p] in line 72, and thus $key < key'$. Let $t_2 < t_1$ be the point when p inserts the pair with key $key'$ into versions[p]. Then p executes an versions[p].Insert() operation in line 25 at point $t_2$, during an Update() operation up'. Since $key < key'$ it follows from Claims 6 and 7 that p executes up before up'. By Lemma 2(b), inv(up') < lin(up') < $t_2 < t_1$. Since up' is p's last update that linearizes before sc, we conclude lin(sc) < inv(up') < $t_1$.

Let hs be the HelpScan(q) operation that p executes in line 64 of pr. Recall that $t_1$ is the point when p executes line 62 of pr, and thus $t_2 < inv(hs)$. Since lin(sc) < $t_1$ we obtain lin(sc) < inv(hs).

Therefore, by Claim 8, $t'(sc) < rs(p, hs)$. By definition, lastScan[p] = val(sc) at point $t'(sc)$. Since there is no Scan(q) by q throughout $(rs(p, hs), t^*)$, by Lemma 3(d), lastScan[p] = val(sc) throughout $(t'(sc), t^*)$. By the assumption that $t_{mv} \leq t^*$, and, as shown above, $t'(sc) < rs(p, hs)$, lastScan[p] is val(sc) throughout $(rs(p, hs), t_{mv})$. Note that p calls hs in line 64 of pr, and at point $t_{mv}$ it executes line 72 of pr (where it removes pair (key, v) from versions[p]). Hence, when p executes line 65 during $(rs(p, hs), t_{mv})$, it reads val(sc) from lastScan[p], and in line 67 it determines a key $y = versions[p].Pred(val(sc))$, which it adds to S. Since $(key, v)$ is stored in versions[p].Pred(val(sc)), when it adds it to S, it has the lastUpdate() preceding sc, by Observation 9, $y = key$. Thus, when p executes line 72 of pr, $key \in S$, and so p does not remove key from versions[p] in that line. This is a contradiction.

4.3.4 Validity. We will now show that the return values of Observe(q) operations are correct. Since no other operations return anything, this is sufficient for validity. The observation below follows from Lemma 3(c) and (d).

Observation 11. Let ob be an Observe() operation executed by process p, and let sc be p’s latest Scan() operation preceding ob. Then, lastScan[p] = val(sc) throughout $[inv(ob), rs(ob)]$.

Claim 12. Suppose process p executes a Scan() operation sc and later an Observe(q) operation ob, and no other Scan() between sc and ob. Let $t_1$ be the point at which p performs lastUpdate[q].LL() in line 33 during ob. Let up be the last Update() operation by q, such that $t(up) < t_1$. If lin(sc) < lin(up), then p does not execute line 36 throughout ob.

Proof. In line 33 of ob, process p reads a value x from lastUpdate[q].key at $t_1$, and in line 34 it reads a value y from lastScan[p]. For the purpose of contradiction, assume p executes line 36 of ob. Therefore, the if-statement in line 35 evaluates to true. Then $x = y$.

By Observation 11, $y = val(sc)$. We will now show that $x = val(up)$, and then arrive at a contradiction.

By Lemma 2, the value of lastUpdate[q].key is ⊥ throughout $(t(up), t'(up))$, and changes to val(up) at $t'(up)$. Since $x = y$, $t_1 \notin (t(up), t'(up))$. Because $t(up) < t_1$, it follows that $t'(up) < t_1$. Also, by Lemma 2(d), lastUpdate[q].key remains val(up) until process q performs lastUpdate[q].SC(LL) in line 22 of a later Update() operation up' at point $t'(up')$. Then, because up is the last Update() by q satisfying $t(up) < t_1$, we obtain $x = val(up)$. Since lin(sc) < lin(up) it follows from Claim 6 that $y = val(sc) < val(up) = x$. This is a contradiction.

For the proof of linearizability, we will use the following claim. The proof is omitted due to space restrictions.

Claim 13. Suppose process p executes a Scan() operation sc and later an Observe(q) operation ob, and no other Scan() between sc and ob. Let $t_1$ be the point at which p performs lastUpdate[q].LL() in
Let \( t^* \) be the point in time at which \( p \) reads \( \text{maxKey} \) in line 62 of \( pr \). By Claim 6 and Claim 7, the pairs that are inserted into \( \text{versions}[p] \) after \( t^* \) have keys greater than \( \text{maxKey} \). Let \( U \) be the set of pairs in \( \text{versions}[p] \) at \( t^* \). Then in line 72, all pairs in \( U \setminus S \), except for the one with largest key, are removed from \( \text{versions}[p] \). Since \( |S| \leq n \), at most \( n \) elements from \( U \) are not removed. Moreover, at most \( n \) elements are inserted into \( \text{versions}[p] \) throughout \( pr \). Hence, the number of elements in \( \text{versions}[p] \) at \( \text{rs}t(\text{pr}) \) is at most \( 2n \).

We follow that it can use our single-writer predecessor object from Section 3 with bounded capacity \( \Delta = 3n \).

Our algorithm also uses multi-word LL/SC objects. We will use the algorithm of Blelloch and Wei [11], which implements a collection of \( M \)-word LL/SC objects from \( O((M + n^2)L) \) registers and CAS objects with step complexity of \( O(L) \).

**Lemma 16.** The time complexity of a \( \text{Prune()} \) operation is \( O(n \log n) \).

**Proof.** Let \( p \) be a process and \( pr \) a \( \text{Prune()} \) operation by \( p \). By Lemma 15, the number of elements in \( \text{versions}[p] \) is at most \( \Delta = 3n \) at any point in time. Therefore, by Theorem 1, performing a \( \text{Pred()} \) operation on \( \text{versions}[p] \) takes time \( O(\log n) \).

We use a sequential set, implemented by a balanced binary search tree for the local set \( S \) that is initialized in line 61 of \( pr \). During each iteration of the for-loop starting in line 63, at most one element is inserted into \( S \), so \( |S| \leq n \). Therefore, inserting and finding a pair in \( S \) takes time \( O(\log n) \).

Each \( \text{HealScan()} \) operation in line 64 and each \( \text{LL()} \) operation in line 65 takes constant time. The \( \text{Pred()} \) operation on \( \text{versions}[p] \) in line 67 takes \( O(\log n) \) time, and so does adding the predecessor to \( S \). Hence, the for-loop comprising lines 63-67, takes \( O(n \log n) \) time.

By Claim 6 and Claim 7, all pairs that are inserted into \( \text{versions}[p] \) after \( p \) reads \( \text{maxKey} \) in line 62 of \( pr \) have keys greater than \( \text{maxKey} \). Moreover, process \( p \) can remove all outdated pairs from \( \text{versions}[p] \) in the while-loop in line 69 of \( pr \) by traversing through all possible pairs with key less than \( \text{maxKey} \). By Lemma 15, the number of pairs in \( \text{versions}[p] \) less than \( \text{maxKey} \) is at most \( 3n \). Checking if a key is in \( S \) takes \( O(\log n) \) time complexity, and traversing through elements can be done by calling successor of elements starting from the element with the smallest key. (\( \text{versions}[p].\text{Succ}(\infty) \)). Therefore, line 72 of \( pr \) takes \( O(n \log n) \) steps. So the total amount of work throughout the execution of \( pr \) is \( O(n \log n) \).

**Theorem 17.** A wait-free linearizable single-writer adaptive partial snapshot object can be implemented from \( O(n^3 \log n) \) single-word \( \text{CAS} \) objects, \( \text{FAI} \) objects, and registers, such that \( \text{Scan()} \) has constant time (and step) complexity, and \( \text{Update()} \) and \( \text{Observe()} \) have \( O(\log n) \) time (and step) complexity.

**Proof of Theorem 17.** Linearity follows immediately from Lemma 14.

By Lemma 16, the total amount of work devoted to each \( \text{Prune()} \) operation is \( O(n \log n) \). Moreover, by Lemma 15 and Theorem 1, each operation on \( \text{versions}[p] \) takes time \( O(\log n) \). The time complexity of a \( \text{Scan()} \) operation is \( O(1) \). The time complexity of \( \text{Update()} \) and \( \text{Observe()} \) is dominated by the operations on a predecessor object, and thus is \( O(\log n) \).
Our algorithm uses \( n \) single-writer predecessor objects of bounded capacity \( \Delta = 3n \). By Theorem 1, these can be obtained from a total of \( O(n^3 \log n) \) base objects. In addition, we use \( 2n \) 1-word LL/SC objects, which can be implemented from \( O(n^3) \) single-word CAS objects and registers [11]. Finally, we need a single FAI object. Thus, in total \( O(n^3 \log n) \) base objects suffice.

5 CONCLUSION

In this paper we present a powerful variant of the single-writer snapshot type, which allows a process to adaptively read a consistent view \( k \) memory components in \( O(\ell \log n) \) steps. Contrary to most fast snapshot solutions, our algorithm does not make unrealistic assumptions about the size of base objects. Instead, it employs powerful synchronization primitives, which are readily available in most common hardware architectures. To achieve our result we implement as a building block a bounded memory single-writer predecessor object, which may be of independent interest.

Our algorithm uses unbounded sequence numbers, which has no practical impact on modern 64-bit architectures. However, from a theoretical point of view, this is not satisfying. But we believe that our unbounded sequence numbers can be replaced with a bounded timestamp system. Unfortunately, no practical bounded timestamp systems are known that could replace the unbounded FAI in our algorithm, without significantly reducing efficiency.

An important open problem is to generalize our algorithm to a multi-writer version, or to allow snapshots of stronger primitives, such as CAS objects (similar to [21]).

ACKNOWLEDGMENTS

We thank Trevor Brown and Wojciech Golab for their insightful discussions, and the anonymous PODC reviewers for many useful comments that helped improve our paper.

Support is gratefully acknowledged from the Natural Science and Engineering Research Council of Canada (NSERC) under Discovery Grant RGPIN/2019-04852, and the Canada Research Chairs program.

REFERENCES

[1] Y. Afek, H. Attiya, D. Dolev, E. Gafni, M. Merritt, and N. Shavit. 1993. Atomic Snapshots of Shared Memory. J. of the ACM 40, 4 (1993), 873–890.
[2] Zahra Aghazadeh, Wojciech Golab, and Philipp Woelflé. 2014. Making Objects Writable. (2014). Under review.
[3] Zahra Aghazadeh and Philipp Woelflé. 2016. Upper Bounds for Boundless Tagging with Boundless Objects. In Proc. of 30th DISC. 442–457. https://doi.org/10.1007/978-3-662-53426-7_32
[4] James H. Anderson. 1990. Composite Registers. In Proc. of 9th PODC, Cynthia Dwork (Ed.). ACM, 15–29. https://doi.org/10.1145/93385.93396
[5] Maya Achel-Raviv and Trevor Brown. 2018. Harnessing epoch-based reclamation for efficient range queries. ACM SIGPLAN Notices 53, 1 (2018), 14–27.
[6] James Aspnes, Hagit Attiya, Keren Censor-Hillel, and Faith Ellen. 2015. Limited-use atomic snapshots with polylogarithmic step complexity. Journal of the ACM (JACM) 62, 1 (2015), 1–22.
[7] James Aspnes and Keren Censor-Hillel. 2013. Atomic snapshots in \( O(\log 3 n) \) steps using randomized helping. In International Symposium on Distributed Computing. Springer, 254–268.
[8] Hagit Attiya, Rachid Guerraoui, and Eric Ruppert. 2008. Partial snapshot objects. In Proc. of 20th SPAA, Friedhelm Meyer auf der Heide and Nir Shavit (Eds.). ACM, 336–343. https://doi.org/10.1145/1378533.1378591
[9] Hagit Attiya, Maurice Herlihy, and Ophir Rachman. 1992. Efficient atomic snapshots using lattice agreement. In International Workshop on Distributed Algorithms. Springer, 35–53.
[10] Mirza Ahad Baig, Danny Hendler, Alessia Milani, and Corentin Travers. 2020. Long-Lived Snapshots with Polylogarithmic Amortized Step Complexity. In Proc. of the 2020 ACM PODC, Yuval Emek and Christian Cachin (Eds.). ACM, 31–40. https://doi.org/10.1145/3382734.3406005
[11] Guy E. Blelloch and Yuanhao Wei. 2019. LL/SC and atomic copy: Constant time, space efficient implementations using only pointer-width CAS. arXiv preprint arXiv:1911.09671 (2019).
[12] James R Driscoll, Neil Sarnak, Daniel D Sleator, and Robert E Tarjan. 1989. Making data structures persistent. Journal of computer and system sciences 38, 1 (1989), 86–124.
[13] Leo J Guibas and Robert Sedgewick. 1978. A dichromatic framework for balanced trees. In 19th Annual Symposium on Foundations of Computer Science (sfcs 1978). IEEE, 8–21.
[14] Michiko Inoue and Wei Chen. 1994. Linear-Time Snapshot Using Multi-writer Multi-reader Registers. In Proc. of 8th WDAG (LNCS), Gerard Tel and Paul M. B. Vitanyi (Eds.), Vol. 857. Springer, 130–140. https://doi.org/10.1007/BFb0020429
[15] Prasad Jayanti, King Tan, and Sam Toueg. 2000. Time and Space Lower Bounds Many Variables in One Atomic Operation: Solutions with Linear or Sublinear Complexity. J. of the ACM 47, 5 (2000), 438–456. https://doi.org/10.1145/3360590.1060697
[16] Prasad Jayanti and Srjdan Petrovic. 2003. Efficient and practical constructions of LL/SC variables. In Proceedings of the twenty-second annual symposium on Principles of distributed computing. 285–294.
[17] Prasad Jayanti, King Tan, and Sam Toueg. 2000. Time and Space Lower Bounds for Nonblocking Implementations. SIAM J. on Comp. 30, 2 (2000), 438–456. https://doi.org/10.1137/S0097539797317299
[18] Lefteris M. Kirousis, Paul G. Spirakis, and Philippas Tsigas. 1994. Reading Many Variables in One Atomic Operation: Solutions with Linear or Sublinear Complexity. IEEE Trans. Parallel Distrib. Syst. 5, 7 (1994), 688–696. https://doi.org/10.1109/71.296315
[19] Yaron Rinyan, Nir Shavit, and Dan Touitou. 2001. Towards a practical snapshot algorithm. Theor. Comp. Sci. 269, 1–2 (2001), 163–201. https://doi.org/10.1016/S0304-3975(00)00412-6
[20] Torvald Riegel, Pascal Felber, and Christof Fetzer. 2006. A lazy snapshot algorithm with eager validation. In International Symposium on Distributed Computing. Springer, 284–298.
[21] Yuanhao Wei, Naama Ben-David, Guy E. Blelloch, Panagiota Fatourou, Eric Ruppert, and Yihan Sun. 2020. Constant-Time Snapshots with Applications to Concurrent Data Structures. (2020). arXiv:cs.DC/2007.02372 (v3)
[22] Yingjun Wu, Joy Arulraj, Jieyi Lin, Ran Xian, and Andrew Pavlo. 2017. An empirical evaluation of in-memory multi-version concurrency control. Proceedings of the VLDB Endowment 10, 7 (2017), 781–792.