Dynamic Atomic Storage Without Consensus

Marcos K. Aguilera
Microsoft Research Silicon Valley
Mountain View, CA, USA
aguilera@microsoft.com

Idit Keidar
Dept. of Electrical Engineering
Technion, Haifa, Israel
idish@ee.technion.ac.il

Dahlia Malkhi
Microsoft Research Silicon Valley
Mountain View, CA, USA
dalia@microsoft.com

Alexander Shraer*
Dept. of Electrical Engineering
Technion, Haifa, Israel
shralex@tx.technion.ac.il

ABSTRACT

This paper deals with the emulation of atomic read/write (R/W) storage in *dynamic* asynchronous message passing systems. In static settings, it is well known that atomic R/W storage can be implemented in a fault-tolerant manner even if the system is completely asynchronous, whereas consensus is not solvable. In contrast, all existing emulations of atomic storage in dynamic systems rely on consensus or stronger primitives, leading to a popular belief that dynamic R/W storage is unattainable without consensus.

In this paper, we specify the problem of dynamic atomic R/W storage in terms of the interface available to the users of such storage. We discover that, perhaps surprisingly, dynamic R/W storage is solvable in a completely asynchronous system: we present DynaStore, an algorithm that solves this problem. Our result implies that atomic R/W storage is in fact easier than consensus, even in dynamic systems.

Categories and Subject Descriptors

B.3.2 [**Memory Structures**]: Design Styles—*shared memory*; C.2.4 [**Computer-Communication Networks**]: Distributed Systems—*Distributed applications*; D.4.2 [**Operating Systems**]: Storage Management—*secondary storage, distributed memories*; D.4.5 [**Operating Systems**]: Reliability—*fault-tolerance*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*distributed systems*

General Terms

Algorithms, Design, Reliability, Theory

Keywords

Shared-memory emulations, dynamic systems, atomic storage.

*Supported by Eshkol Fellowship from the Israeli Ministry of Science.

PODC'09, August 10–12, 2009, Calgary, Alberta, Canada.

1. INTRODUCTION

Distributed systems provide high availability by replicating the service state at multiple processes. A fault-tolerant distributed system may be designed to tolerate failures of a minority of its processes. However, this approach is inadequate for long-lived systems, because over a long period, the chances of losing more than a minority inevitably increase. Moreover, system administrators may wish to deploy new machines due to increased workloads, and replace old, slow machines with new, faster ones. Thus, real-world distributed systems need to be *dynamic*, i.e., adjust their membership over time. Such dynamism is realized by providing users with an interface to *reconfiguration* operations that *add* or *remove* processes.

Dynamism requires some care. First, if one allows arbitrary reconfiguration, one may lose liveness. For example, say that we build a fault tolerant solution using three processes, p_1, p_2, and p_3. Normally, the adversary may crash one process at any moment in time, and the up-to-date system state is stored at a majority of the current configuration. However, if a user initiates the removal of p_1 while p_1 and p_2 are the ones holding the up-to-date system state, then the adversary may not be allowed to crash p_2, for otherwise the remaining set cannot reconstruct the up-to-date state. Providing a general characterization of allowable failures under which liveness can be ensured is a challenging problem.

A second challenge dynamism poses is ensuring safety in the face of concurrent reconfigurations, i.e., when some user invokes a new reconfiguration request while another request (potentially initiated by another user) is under way. Early work on replication with dynamic membership could violate safety in such cases [8, 22, 10] (as shown in [27]). Many later works have rectified this problem by using a centralized sequencer or some variant of consensus to agree on the order of reconfigurations (see discussion of related work in Section 2).

Interestingly, consensus is not essential for implementing replicated storage. The ABD algorithm [3] shows that atomic *read/write (R/W)* shared memory objects can be implemented in a fault-tolerant manner even if the system is completely asynchronous. Nevertheless, to the best of our knowledge, all previous dynamic storage solutions rely on consensus or similar primitives, leading to a popular belief that dynamic storage is unattainable without consensus.

In this work, we address the two challenges mentioned above, and debunk the myth that consensus is needed for dynamic storage. We first provide a precise problem specification of a dynamic problem. To be concrete, we focus on atomic R/W storage, though we believe the approach we take for defining a dynamic problem can be carried to other problems. We then present *DynaStore*, a solution to this problem in an asynchronous system where processes may

undetectably crash, so that consensus is not solvable. We note that our solution is given as a possibility proof, rather than as a blueprint for a new storage system. Given our result that consensus-less solutions are possible, we expect future work to apply various practical optimizations to our general approach, in order to build real-world distributed services. We next elaborate on these two contributions.

Dynamic Problem Specification

In Section 3, we define the problem of an atomic R/W register in a dynamic system. Clearly, the progress of such service is conditioned on certain failure restrictions in the deployed system. It is well understood how to state a liveness condition of the static version of this problem: t-resilient R/W storage guarantees progress if fewer than t processes crash. For an n-process system, it is well known that t-resilient R/W storage exists when $t < n/2$, and does not exist when $t \geq n/2$ [3].

The liveness condition of a dynamic system needs to also capture changes introduced by the user. Suppose the system initially has four processes $\{p_1, p_2, p_3, p_4\}$ in its configuration (also called *view*). Initially, any one process may crash. Suppose that p_1 crashes. Then, additional crashes would lead to a loss of liveness. Now suppose the user requests to reconfigure the system to remove p_1. While the request is pending, no additional crashes can happen, because the system must transfer the up-to-date state from majority of the previous view to a majority of the new one. However, once the removal is completed, the system can tolerate an additional crash among the new view $\{p_2, p_3, p_4\}$. Overall, two processes may crash during the execution. Viewed as a simple threshold condition, this exceeds a minority threshold, which contradicts lower bounds. However, the liveness condition we will formulate will not be in the form of a simple threshold; rather, we require crashes to occur gradually, and adapt to reconfigurations.

A dynamic system also needs to support additions. Suppose the system starts with three processes $\{p_1, p_2, p_3\}$. In order to reconfigure the system to add a new process p_4, a majority of the view $\{p_1, p_2, p_3\}$ must be alive to effect the change. Additionally, a majority of the view $\{p_1, p_2, p_3, p_4\}$ must be alive to hold the state stored by the system. Again, the condition here is more involved than a simple threshold. That is, if a user requests to *add* p_4, then while the request is pending, a majority of both old and new views need to be alive. Once the reconfiguration is completed, the requirement weakens to a majority of the new view.

In order to provide a protocol-independent specification, we must expose in the model the completion of reconfigurations. Our service interface therefore includes explicit *reconfig* operations that allow the user to add and remove processes. These operations return OK when they complete. Given these, we state the following requirement for liveness for dynamic R/W storage: At any moment in the execution, let the *current view* consist of the initial view with all completed reconfiguration operations (add/remove) applied to it. We require that the set of crashed processes and those whose removal is pending be a minority of the current view, and of any pending future views. Moreover, like previous reconfigurable storage algorithms [19, 12], we require that no new *reconfig* operations will be invoked for "sufficiently long" for the started operations to complete. This is formally captured by assuming that only a finite number of *reconfig* operations are invoked.

Note that a dynamic problem is harder than the static variant. In particular, a solution to dynamic R/W is a fortiori a solution to the static R/W problem. Indeed, the solution must serve read and write requests, and in addition, implement reconfiguration operations. If deployed in a system where the user invokes only read and write requests, and never makes use of the reconfiguration interface, it must solve the R/W problem with precisely the same liveness condition, namely, tolerating any minority of failures. Similarly, dynamic consensus is harder than static consensus, and is therefore a fortiori not solvable in an asynchronous setting with one crash failure allowed. As noted above, in this paper, we focus on dynamic R/W storage.

DynaStore: Dynamic Atomic R/W Storage

Our algorithm does not need consensus to implement the reconfiguration operations. Intuitively, previous protocols used consensus, virtual synchrony, or a sequencer, in order to provide processes with an agreed-upon sequence of configurations, so that the membership views of processes do not diverge. The key observation in our work is that it is sufficient that such a sequence of configurations exists, and there is no need for processes to know precisely which configurations belong to this sequence, as long as they have some assessment which includes these configurations, possibly in addition to others which are not in the sequence. In order to enable this property, in Section 4 we introduce *weak snapshots*, which are easily implementable in an asynchronous system. Roughly speaking, such objects support *update* and *scan* operations accessible by a given set of processes, such that *scan* returns a set of updates that is guaranteed to include the *first update* made to the object (but the object cannot identify which update that is).

In DynaStore, which we present in Section 5, each view w has a weak snapshot object $ws(w)$, which stores reconfiguration proposals for what the next view should be. Thus, we can define a unique global sequence of views, as the sequence that starts with some fixed initial view, and continues by following the first proposal stored in each view's *ws* object. Although it is impossible for processes to learn what this sequence is, they can learn a DAG of views that includes this sequence as a path. They do this by creating a vertex for the current view, querying the *ws* object, creating an edge to each view in the response, and recursing. Reading and writing from a chain of views is then done in a manner similar to previous protocols, e.g., [19, 12, 5, 23, 24].

Summary of Contributions

In summary, our work makes two contributions.

- We define a dynamic R/W storage problem that includes a clean and explicit liveness condition, which does not depend on a particular solution to the problem. The definition captures a dynamically changing resilience requirement, corresponding to reconfiguration operations invoked by users. The approach easily carries to other problems, such as consensus. As such, it gives a clean extension of existing static problems to the dynamic setting.

- We discover that dynamic R/W storage is solvable in a completely asynchronous system with failures, by presenting a solution to this problem. Along the way we define a new abstraction of weak snapshots, employed by our solution, which may be useful in its own right. Our result implies that the dynamic R/W is weaker than the (dynamic) consensus problem, which is not solvable in this setting. This was known before for static systems, but not for the dynamic version. The result counters the intuition that emanates from all previous dynamic systems, which used agreement to handle configuration changes.

2. RELATED WORK

Several existing solutions can be viewed in retrospect as solving a dynamic problem. Most closely related are works on reconfig-

urable R/W storage. RAMBO [19, 12] solves a similar problem to the one we have formulated above; other works [21, 23, 24] extend this concept for Byzantine fault tolerance. All of these works have processes agree upon a unique sequence of configuration changes. Some works use an auxiliary source (such as a single reconfigurer process or an external consensus algorithm) to determine configuration changes [18, 11, 19, 12, 21, 24], while others implement fault-tolerant consensus decisions on view changes [5, 23]. In contrast, our work implements reconfigurable R/W storage without any agreement on view changes.

Since the closest related work is on RAMBO, we further elaborate on the similarities and differences between RAMBO and DynaStore. In RAMBO, a new configuration can be proposed by any process, and once it is installed, it becomes the current configuration. In DynaStore, processes suggest changes and not configurations, and thus, the current configuration is determined by the set of all changes proposed by complete reconfigurations. For example, if a process suggests to add p_1 and to remove p_2, while another process concurrently suggests to add p_3, DynaStore will install a configuration including both p_1 and p_3 and without p_2, whereas in RAMBO there is no guarantee that any future configuration will reflect all three proposed changes, unless some process explicitly proposes such a configuration. In DynaStore, a quorum of a configuration is any majority of its members, whereas RAMBO allows for general quorum-systems, specified explicitly for each configuration by the proposing process. In both algorithms, a non-faulty quorum is required from the current configuration. A central idea in allowing dynamic changes is that a configuration can be replaced, after which a quorum of the old configuration can crash. In DynaStore, a majority of a current configuration C is allowed to crash as soon as C is no longer current. In RAMBO, two additional conditions are needed: C must be garbage-collected at every non-faulty process $p \in C$, and all *read* and *write* operations that began at p before C was garbage-collected must complete. Thus, whereas in DynaStore the conditions allowing a quorum of C to fail can be evaluated based on events visible to the application, in RAMBO these conditions are internal to the algorithm. Note that if some process $p \in C$ might fail, it might be impossible for other processes to learn whether p garbage-collected C or not. Assuming that all quorums required by RAMBO and DynaStore are responsive, both algorithms require additional assumptions for liveness. In both, the liveness of *read* and *write* operations is conditioned on the number of reconfigurations being finite. In addition, in both algorithms, the liveness of reconfigurations does not depend on concurrent *read* and *write* operations. However, whereas reconfigurations in RAMBO rely on additional synchrony or failure-detection assumptions required for consensus, reconfigurations in DynaStore, just like its *read* and *write* operations, only require the number of reconfigurations to be finite.

View-oriented group communication systems provide a membership service whose task is to maintain a dynamic view of active members. These systems solve a dynamic problem of maintaining agreement on a sequence of views, and additionally provide certain services within the members of a view, such as atomic multicast and others [6]. Maintaining agreement on group membership in itself is impossible in asynchronous systems [4]. However, perhaps surprisingly, we show that the dynamic R/W problem is solvable in asynchronous systems. This appears to contradict the impossibility but it does not: We do not implement group membership because our processes do not have to agree on and learn a unique sequence of view changes.

The State Machine Replication (SMR) approach [15, 25] provides a fault tolerant emulation of arbitrary data types by forming agreement on a sequence of operations applied to the data. Paxos [15] implements SMR, and allows one to dynamically reconfigure the system by keeping the configuration itself as part of the state stored by the state machine. Another approach for reconfigurable SMR is to utilize an auxiliary configuration-master to determine view changes, and incorporate directives from the master into the replication protocol. This approach is adopted in several practical systems, e.g., [17, 20, 26], and is formulated in [16]. Naturally, a reconfigurable SMR can support our dynamic R/W memory problem. However, our work solves it without using the full generality of SMR and without reliance on consensus.

An alternative way to break the minority barrier in R/W emulation is by strengthening the model using a failure detector. Delporte et al. [9] identify the weakest failure detector for solving R/W memory with arbitrary failure thresholds. Their motivation is similar to ours– solving R/W memory with increased resilience threshold. Unlike our approach, they tackle more than a minority of failures right from the outset. They identify the *quorums failure detector* as the weakest detector required for strengthening the asynchronous model, in order to break the minority resilience threshold. Our approach is incomparable to theirs, i.e., our model is neither weaker nor stronger. On the one hand, we do not require a failure detector, and on the other, we allow the number to failures to exceed a minority only after certain actions are taken. Moreover, their model does not allow for additions as ours does. Indeed, our goal differs from [9], namely, to model dynamic reconfiguration in which resilience is adaptive to actions by the processes.

3. DYNAMIC PROBLEM DEFINITION

The goal of our work is to implement a read/write service with atomicity guarantees. The storage service is deployed on a collection of processes that interact using asynchronous message passing. We assume an unknown, unbounded and possibly infinite universe of processes Π. Communication channels between all processes are reliable. Below we revisit the definition of reliable links in a dynamic setting.

The service stores a value v from a domain $\mathcal{V}$ and offers an interface for invoking *read* and *write* operations and obtaining their result. Initially, the service holds a special value $\perp \notin \mathcal{V}$. The sequential specification for this service is as follows: In a sequence of operations, a read returns the latest written value or $\perp$ if none was written. Atomicity [14] (also called linearizability [13]) requires that for every execution, there exist a corresponding sequential execution, which conforms with the operation precedence relation, and which satisfies the sequential specification.

In addition to the above API, the service exposes an interface for invoking reconfigurations. We define $\mathit{Changes} \stackrel{def}{=} \{\mathit{Remove, Add}\} \times \Pi$. We informally call any subset of Changes a *set of changes*. A *view* is a set of changes. A *reconfig* operation takes as parameter a set of changes c and returns OK. We say that a change ω is *proposed* in the execution if a *reconfig*(c) operation is invoked s.t. $\omega \in c$. A process p_i is *active* if p_i does not crash, some process invokes a *reconfig* operation to add p_i, and no process invokes a *reconfig* operation to remove p_i. We do not require all processes in Π to start taking steps from the beginning of the execution, but instead we assume that if p_i is active then p_i takes infinitely many steps (if p_i is not active then it may stop taking steps).

For a set of changes w, the *removal-set of* w, denoted $w.\mathit{remove}$, is the set $\{i \mid \mathit{(Remove, i)} \in w\}$. The *join set of* w, denoted $w.\mathit{join}$, is the set $\{i \mid \mathit{(Add, i)} \in w\}$. Finally, the *membership of* w, denoted $w.\mathit{members}$, is the set $w.\mathit{join} \backslash w.\mathit{remove}$.

At any time t in the execution, we define $V(t)$ to be the union of all sets c s.t. a *reconfig(c)* completes by time t. Note that removals

are permanent, that is, a process that is removed will never again be in members. More precisely, if a reconfiguration removing p_i from the system completes at time t_0, then p_i is excluded from $V(t)$.*members*, for every $t \geq t_0$[1]. We assume a non-empty view $V(0)$ which is initially known to every process in the system and we say, by convention, that a *reconfig*$(V(0))$ completes by time 0. Let $P(t)$ be the set of *pending changes* at time t, i.e., for each element $\omega \in P(t)$ some process invokes a *reconfig*(c) operation s.t. $\omega \in c$ by time t, and no process completes such a *reconfig* operation by time t. Denote by $F(t)$ the set of processes which have crashed by time t.

Intuitively, only processes that are members of the current system configuration should be allowed to initiate actions. To capture this restriction, *read*, *write* and *reconfig* operations at a process p_i are initially disabled, until *enable operations* occurs at p_i. Intuitively, any pending future view should have a majority of processes that did not crash and were not proposed for removal; we specify a simple condition sufficient to ensure this. A dynamic R/W service guarantees the following liveness properties:

Definition 1. [Dynamic Service Liveness]
If at every time t in the execution, fewer than $|V(t).members|/2$ processes out of $V(t).members \cup P(t).join$ are in $F(t) \cup P(t).remove$, and the number of different changes proposed in the execution is finite, then the following holds:

1. Eventually, the *enable operations* event occurs at every active process that was added by a complete *reconfig* operation.
2. Every operation invoked at an active process eventually completes.

A common definition of reliable links states that if processes p_i and p_j are "correct", then every message sent by p_i is eventually received by p_j. We adapt this definition to a dynamic setting as follows: for a message sent at time t, we require eventual delivery if both processes are active and $j \in V(t).join \cup P(t).join$, i.e., a *reconfig*$(c)$ operation was invoked by time t s.t. $(Add, j) \in c$.

4. THE WEAK SNAPSHOT ABSTRACTION

A weak snapshot object S accessible by a set P of processes supports two operations, $update_i(c)$ and $scan_i()$, for a process $p_i \in P$. The $update_i(c)$ operation gets a value c and returns OK, whereas $scan_i()$ returns a set of values. Note that the set P of processes is fixed (i.e., *static*). We require the following semantics from *scan* and *update* operations:

NV1 Let o be a *scan*$_i()$ operation that returns C. Then for each $c \in C$, an *update*(c) operation is invoked by some process prior to the completion of o.

NV2 Let o be a *scan*$_i()$ operation that is invoked after the completion of an *update*$_j(c)$ operation, and that returns C. Then $C \neq \emptyset$.

NV3 Let o be a *scan*$_i()$ operation that returns C and let o' be a *scan*$_j()$ operation that returns C' and is invoked after the completion of o. Then $C \subseteq C'$.

NV4 There exists c such that for every *scan*$()$ operation that returns $C \neq \emptyset$, it holds that $c \in C$.

NV5 If some majority M of processes in P keep taking steps then every *scan*$_i()$ and *update*$_i(c)$ invoked by every process $p_i \in M$ eventually completes.

[1]In practice, one can add back a process by changing its id.

Algorithm 1 Weak snapshot - code for process p_i.

```
1:  operation update_i(c)
2:      if collect() = ∅ then
3:          Mem[i].Write(c)
4:      return OK

5:  operation scan_i()
6:      C ← collect()
7:      if C = ∅ then return ∅
8:      C ← collect()
9:      return C

10: procedure collect()
11:     C ← ∅;
12:     for each p_k ∈ P
13:         c ← Mem[k].Read()
14:         if c ≠ ⊥ then C ← C ∪ {c}
15:     return C
```

Although these properties bear resemblance to the properties of atomic snapshot objects [1], NV1-NV5 define a weaker abstraction: we do not require that *all update*s are ordered as in atomic snapshot objects, and even in a sequential execution, the set returned by a *scan* does not have to include the value of the most recently completed *update* that precedes it. Intuitively, these properties only require that the "first" *update* is seen by all *scan*s that see any *update*s. As we shall see below, this allows for a simpler implementation than of a snapshot object.

DynaStore will use multiple weak snapshot objects, one of each view w. The weak snapshot of view w, denoted *ws*(w), is accessible by the processes in $w.members$. To simplify notation, we denote by *update*$_i(w, c)$ and *scan*$_i(w)$ the $update$ and $scan$ operation, respectively, of process p_i of the weak snapshot object *ws*(w). Intuitively, DynaStore uses weak snapshots as follows: in order to propose a set of changes c to the view w, a process p_i invokes *update*$_i(w, c)$; p_i can then learn proposals of other processes by invoking *scan*$_i(w)$, which returns a set of sets of changes.

Implementation.

Our implementation of *scan* and *update* is shown in Algorithm 1. It uses an array *Mem* of $|P|$ single-writer multi-reader (SWMR) atomic registers, where all registers are initialized to $\perp$. Such registers support *Read*$()$ and *Write*(c) operations s.t. only process $p_i \in P$ invokes *Mem*$[i]$.*Write*(c) and any process $p_j \in P$ can invoke *Mem*$[i]$.*Read*$()$. The implementation of such registers in message-passing systems is described in the literature [3].

A *scan*$_i()$ reads from all registers in *Mem* by invoking *collect*, which returns the set C of values found in all registers. After invoking *collect* once, *scan*$_i()$ checks whether the returned C is empty. If so, it returns $\emptyset$, and otherwise invokes *collect* one more time. An *update*$_i(c)$ invokes *collect*, and in case $\emptyset$ is returned, writes c to *Mem*$[i]$. Intuitively, if *collect*$()$ returns a non-empty set then another *update* is already the "first" and there is no need to perform a *Write* since future *scan* operations would not be obligated to observe it. In DynaStore, this happens when some process has already proposed changes to the view, and thus, the weak snapshot does not correspond to the most up-to-date view in the system and there is no need to propose additional changes to this view.

Standard emulation protocols for atomic SWMR registers [3] guarantee integrity (property NV1) and liveness (property NV5). We next explain why Algorithm 1 preserves properties NV2-NV4; the formal proof of correctness appears in the full paper [2]. First, notice that at most one *Mem*$[i]$.*Write* operation can be invoked in the execution, since after the first *Mem*$[i]$.*Write* operation completes, any *collect* invoked by p_i (the only writer of this register)

will return a non-empty set and p_i will never invoke another *Write*. This together with atomicity of all registers in *Mem* implies properties NV2-NV3. Property NV4 stems from the fact that every *scan*() operation that returns $C \neq \emptyset$ executes *collect* twice. Observe such operation o that is the first to complete one *collect*. Any other *scan*() operation o' begins its second *collect* only after o completes its first *collect*. Atomicity of the registers in *Mem* along with the fact that each register is written at-most once, guarantees that any value returned by a *Read* during the first *collect* of o will be read during the second *collect* of o'.

5. DYNASTORE

This section describes DynaStore, an algorithm for multi-writer multi-reader (MWMR) atomic storage in a dynamic system, which is presented in Algorithm 2. A key component of our algorithm is a procedure *ContactQ* (lines 31-41) for reading and writing from/to a quorum of members in a given view, used similarly to the *communicate* procedure in ABD [3]. When there are no reconfigurations, *ContactQ* is invoked twice by the *read* and *write* operations – once in a read-phase and once in a write-phase. More specifically, both *read* and *write* operations first execute a read-phase, where they invoke *ContactQ* to query a quorum of the processes for the latest value and timestamp, after which both operations execute a write-phase as follows: a *read* operation invokes *ContactQ* again to write-back the value and timestamp obtained in the read-phase, whereas a *write* operation invokes *ContactQ* with a higher and unique timestamp and the desired value.

To allow reconfiguration, the members of a view also store information about the current view. They can change the view by modifying this information at a quorum of the current view. We allow the reconfiguration to occur concurrently with any *read* and *write* operations. Furthermore, once reconfiguration is done, we allow future reads and writes to use (only) the new view, so that processes can be expired and removed from the system. Hence, the key challenge is to make sure that no reads linger behind in the old view, while updates are made to the new view. Atomicity is preserved using the following strategy.

- The read-phase is modified so as to first read information on reconfiguration, and then read the value and its timestamp. If a new view is discovered, the read-phase repeats with the new view.
- The write-phase, which works in the last view found by the read-phase, is modified as well. First, it writes the value and timestamp to a quorum of the view, and then, it reads the reconfiguration information. If a new view is discovered, the protocol goes back to the read-phase in the new view (the write-phase begins again when the read-phase ends).
- The *reconfig* operation has a preliminary phase, writing information about the new view to the quorum of the old one. It then continues by executing the phases described above, starting in the old view.

The core of a read-phase is procedure *ReadInView*, which reads the configuration information (line 67) and then invokes *ContactQ* to read the value and timestamp from a quorum of the view (line 68). It returns a non-empty set if a new view was discovered in line 67. Similarly, procedure *WriteInView* implements the basic functionality of the write-phase, first writing (or writing-back) the value and timestamp by invoking *ContactQ* in line 73, and then reading configuration information in line 74 (we shall explain lines 71-72 in Section 5.3).

First, for illustration purposes, consider a simple case where only one reconfiguration request is ever invoked, from w_1 to w_2. We shall refer to this reconfiguration operation as *RC*. The main insight into why the above regime preserves read/write atomicity is the following. Say that a *write* operation performs a write-phase W writing in w_1 the value v with timestamp ts. Then there are two possible cases with respect to *RC*. One is that *RC*'s read-phase observes W. Hence, *RC*'s write-phase writes-back a value into w_2, whose timestamp is at least as high as ts. Otherwise, *RC*'s read-phase does not observe W. This means that W's execution of *ContactQ* writing a quorum of w_1 did not complete before *RC* invoked *ContactQ* during its read-phase, and so W starts to read w_1's configuration information after *RC*'s preliminary phase has completed, updating this information. Hence, W observes w_2 and the *write* operation continues in w_2 (notice that if a value v' with timestamp higher than ts is found in w_2 then the *write* will no longer send v, and instead simply writes back v' to a quorum of w_2).

In our example above, additional measures are needed to preserve atomicity if several processes concurrently propose changes to w_1. Thus, the rest of our algorithm is dedicated to the complexity that arises due to multiple contending reconfiguration requests. Our description is organized as follows: Section 5.1 introduces the pseudo-code of DynaStore, and clarifies its notations and atomicity assumptions. Section 5.2 presents the DAG of views, and shows how every operation in DynaStore can be seen as a traversal on that graph. Section 5.3 discusses *reconfig* operations. Finally, Section 5.4 presents the notion of established views, which is central to the analysis of DynaStore. Proofs are deferred to the full paper [2].

5.1 DynaStore Basics

DynaStore uses *operations*, *upon clauses*, and *procedures*. Operations are invoked by the user, whereas upon-clauses are event handlers – they are actions that may be triggered whenever their condition is satisfied. Procedures are called from an operation. In the face of concurrency, operations and upon clauses act like separate monitors: at most one of each kind can be executed at a time. Note that an operation and an upon-clause might execute concurrently. In addition, all accesses to local variables are atomic (even if accessed by an operation and an upon-clause concurrently), and when multiple local variables are assigned as a tuple (e.g., line 72), the entire assignment is atomic. Operations and local variables at process p_i are denoted with subscript i.

Operations and upon-clauses access different variables for storing the value and timestamp: v_i and ts_i are accessed in upon-clauses, whereas operations (and procedures) manipulate v_i^{max} and ts_i^{max}. Procedure *ContactQ* sends a write-request including v_i^{max} and ts_i^{max} (line 35) when writing a quorum, and a read-request (line 36) when reading a quorum ($msgNum_i$, a local sequence number, is also included in such messages). When p_i receives a write-request, it updates v_i and ts_i if the received timestamp is bigger than ts_i, and sends back a REPLY message containing the sequence number of the request (line 45). When a read-request is received, p_i replies with v_i, ts_i, and the received sequence number (line 46).

Every process p_i executing Algorithm 2 maintains a local estimation of the latest view, $curView_i$ (line 9), initialized to $V(0)$ when the process starts. Although p_i is able to execute all event-handlers immediately when it starts, recall that invocations of *read*, *write* or *reconfig* operations at p_i are only allowed once they are enabled for the first time; this occurs in line 11 (for processes in $V(0)$) or in line 81 (for processes added later). If p_i discovers that it is being removed from the system, it simply halts (line 53). In this section, we denote changes of the form (Add, i) by $(+, i)$ and changes of the form $(Remove, i)$ by $(-, i)$.

Algorithm 2 Code for process p_i.

```
1: state
2:    v_i ∈ V ∪ {⊥}, initially ⊥                                  // latest value received in a WRITE message
3:    ts_i ∈ ℕ_0 × (Π ∪ {⊥}), initially (0, ⊥)                    // timestamp corresponding to v_i (timestamps have selectors num and pid)
4:    v_i^max ∈ V ∪ {⊥}, initially ⊥                              // latest value observed in Traverse
5:    ts_i^max ∈ ℕ_0 × (Π ∪ {⊥}), initially (0, ⊥)                // timestamp corresponding to v_i^max
6:    pickNewTS_i ∈ {FALSE, TRUE}, initially FALSE                // whether Traverse should pick a new timestamp
7:    M_i: set of messages, initially ∅
8:    msgNum_i ∈ ℕ_0, initially 0                                 // counter for sent messages
9:    curView_i ∈ Views, initially V(0)                           // latest view

10: initially:
11:    if (i ∈ V(0).join) then enable operations

12: operation read_i():
13:    pickNewTS_i ← FALSE
14:    newView ← Traverse(∅, ⊥)
15:    NotifyQ(newView)
16:    return v_i^max

17: operation write_i(v):
18:    pickNewTS_i ← TRUE
19:    newView ← Traverse(∅, v)
20:    NotifyQ(newView)
21:    return OK

22: operation reconfig_i(cng):
23:    pickNewTS_i ← FALSE
24:    newView ← Traverse(cng, ⊥)
25:    NotifyQ(newView)
26:    return OK

27: procedure NotifyQ(w)
28:    if did not receive ⟨NOTIFY, w⟩ then
29:       send ⟨NOTIFY, w⟩ to w.members
30:    wait for ⟨NOTIFY, w⟩ from majority of w.members

31: procedure ContactQ(msgType, D)
32:    M_i ← ∅
33:    msgNum_i ← msgNum_i + 1;
34:    if msgType = W then
35:       send ⟨REQ, W, msgNum_i, v_i^max, ts_i^max⟩ to D
36:    else send ⟨REQ, R, msgNum_i⟩ to D
37:    wait until M_i contains a ⟨REPLY, msgNum_i, ···⟩
          from a majority of D
38:    if msgType = R then
39:       tm ← max{t:⟨REPLY, msgNum_i, v, t⟩ is in M_i}
40:       vm ← value corresponding to tm
41:       if tm > ts_i^max then (v_i^max, ts_i^max)←(vm, tm)

42: upon receiving ⟨REQ, msgType, num, v, ts⟩ from p_j:
43:    if msgType = W then
44:       if (ts > ts_i) then (v_i, ts_i) ← (v, ts)
45:       send ⟨REPLY, num⟩ to p_j
46:    else send message ⟨REPLY, num, v_i, ts_i⟩ to p_j

47: procedure Traverse(cng, v)
48:    desiredView ← curView_i ∪ cng
49:    Front ← {curView_i}
50:    do
51:       s ← min{|ℓ| : ℓ ∈ Front}
52:       w ← any ℓ ∈ Front s.t. |ℓ| = s
53:       if (i ∉ w.members) then halt
54:       if w ≠ desiredView then
55:          update_i(w, desiredView\w)
56:       ChangeSets ← ReadInView(w)
57:       if ChangeSets ≠ ∅ then
58:          Front ← Front \ {w}
59:          for each c ∈ ChangeSets
60:             desiredView ← desiredView ∪ c
61:             Front ← Front ∪ {w ∪ c}
62:       else ChangeSets ← WriteInView(w, v)
63:    while ChangeSets ≠ ∅
64:    curView_i ← desiredView
65:    return desiredView

66: procedure ReadInView(w)
67:    ChangeSets ← scan_i(w)
68:    ContactQ(R, w.members)
69:    return ChangeSets

70: procedure WriteInView(w, v)
71:    if pickNewTS_i then
72:       (pickNewTS_i, v_i^max, ts_i^max) ←
             (FALSE, v, (ts_i^max.num + 1, i))
73:    ContactQ(W, w.members)
74:    ChangeSets ← scan_i(w)
75:    return ChangeSets

76: upon receiving ⟨NOTIFY, w⟩ for the first time:
77:    send ⟨NOTIFY, w⟩ to w.members
78:    if (curView_i ⊂ w) then
79:       pause any ongoing Traverse
80:       curView_i ← w
81:       if (i ∈ w.join) then enable operations
82:       if paused in line 79, restart Traverse from line 48

83: upon receiving ⟨REPLY, ···⟩:
84:    add the message and its sender-id to M_i
```

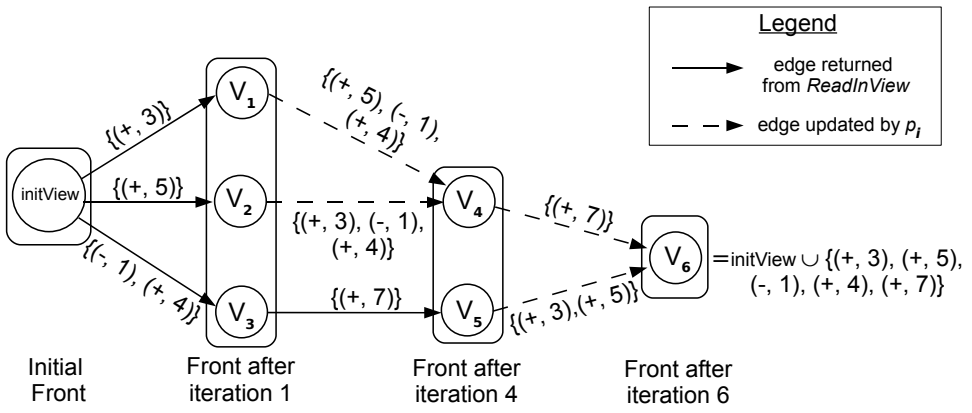

Figure 1: Example DAG of views.

5.2 Traversing the Graph of Views

Weak snapshots organize all views into a DAG, where views are the vertices and there is an edge from a view w to a view w' whenever an $update_j(w, c)$ has been made in the execution by some process $j \in w.members$, updating $ws(w)$ to include the change $c \neq \emptyset$ s.t. $w' = w \cup c$; $|c|$ can be viewed as the weight of the edge – the distance between w' and w in changes. Our algorithm maintains the invariant that $c \cap w = \emptyset$, and thus w' always contains more changes than w, i.e., $w \subset w'$. Hence, the graph of views is acyclic.

The main logic of Algorithm 2 lies in procedure *Traverse*, which is invoked by all operations. This procedure traverses the DAG of views, and transfers the state of the emulated register from view to view along the way. *Traverse* starts from the view $curView_i$. Then, the DAG is traversed in an effort to find all membership changes in the system; these are collected in the set *desiredView*. After finding all changes, *desiredView* is added to the DAG by updating the appropriate *ws* object, so that other processes can find it in future traversals.

The traversal resembles the well-known Dijkstra algorithm for finding shortest paths from some single source [7], with the important difference that our traversal modifies the graph. A set of views, *Front*, contains the vertices reached by the traversal and whose outgoing edges were not yet inspected. Initially, $Front = \{curView_i\}$ (line 49). Each iteration processes the vertex w in *Front* closest to $curView_i$ (lines 51 and 52).

During an iteration of the loop in lines 50–63, it might be that p_i already knows that w does not contain all proposed membership changes. This is the case when *desiredView*, the set of changes found in the traversal, is different from w. In this case, p_i installs an edge from w to *desiredView* using $update_i$ (line 55). As explained in Section 4, in case another update to $ws(w)$ has already completed, *update* does not install an additional edge from w; the only case when multiple outgoing edges exist is if they were installed concurrently by different processes.

Next, p_i invokes *ReadInView*(w) (line 56), which reads the state and configuration information in this view, returning all edges outgoing from w found when scanning $ws(w)$ in line 67. By property NV2, if p_i or another process had already installed an edge from w, a non-empty set of edges is returned from *ReadInView*. If one or more outgoing edges are found, w is removed from *Front*, the next views are added to *Front*, and the changes are added to *desiredView* (lines 59–61). If p_i does not find outgoing edges from w, it invokes *WriteInView*(w) (line 62), which writes the latest known value to this view and again scans $ws(w)$ in line 74, returning any outgoing edges that are found. If here too no edges are found, the traversal completes.

Notice that *desiredView* is chosen in line 52 only when there are no other views in *Front*, since it contains the union of all views observed during the traversal, and thus any other view in *Front* must be of smaller size (i.e., contain fewer changes). Moreover, when $w \neq$ *desiredView* is processed, the condition in line 54 evaluates to true, and *ReadInView* returns a non-empty set of changes (outgoing edges) by property NV2. Thus, *WriteInView*$(w, *)$ is invoked only when *desiredView* is the only view in *Front*, i.e., $w =$ *desiredView* (this transfers the state found during the traversal to *desiredView*, the latest-known view). For the same reason, when the traversal completes, $Front = \{$*desiredView*$\}$. Then, *desiredView* is assigned to $curView_i$ in line 64 and returned from Traverse.

To illustrate such traversals, consider the example in Figure 1. Process p_i invokes *Traverse* and let *initView* be the value of $curView_i$ when *Traverse* is invoked. Assume that *initView.members* includes at least p_1 and p_i, and that $cng = \emptyset$ (this parameter of *Traverse* will be explained in Section 5.3). Initially, its *Front*, marked by a rectangle in Figure 1, includes only *initView*, and *desiredView* = *initView*. Then, the condition in line 54 evaluates to false and p_i invokes *ReadInView*(*initView*), which returns $\{\{(+,3)\}, \{(+,5)\}, \{(-,1),(+,4)\}\}$. Next, p_i removes *initView* from *Front* and adds vertices V_1, V_2 and V_3 to *Front* as shown in Figure 1. For example, V_3 results from adding the changes in $\{(-,1),(+,4)\}$ to *initView*. At this point, *desiredView* = *initView*$\cup\{(+,3), (+,5), (-,1), (+,4)\}$. In the next iteration of the loop in lines 50–63, one of the smallest views in *Front* is processed. In our scenario, V_1 is chosen. Since $V_1 \neq$ *desiredView*, p_i installs an edge from V_1 to *desiredView*. Suppose that no other updates were made to $ws(V_1)$ before p_i's update completes. Then, *ReadInView*(V_1) returns $\{\{(+,5),(-,1),(+,4)\}\}$ (properties NV1 and NV2). Then, V_1 is removed from *Front* and $V_4 = V_1 \cup \{(+,5),(-,1),(+,4)\}$ is added to *Front*. In the next iteration, an edge is installed from V_2 to V_4 and V_2 is removed from *Front*.

Now, the size of V_3 is smallest in *Front*, and suppose that another process p_j has already completed $update_j(V_3, \{(+,7)\})$. p_i executes *update* (line 55), however since an outgoing edge already exists, a new edge is not installed. Then, *ReadInView*(V_3) is invoked and returns $\{\{(+,7)\}\}$. Next, V_3 is removed from *Front*, $V_5 = V_3 \cup \{(+,7)\}$ is added to *Front*, and $(+,7)$ is added to *desiredView*. Now, $Front = \{V_4, V_5\}$, and we denote the new *desiredView* by V_6. When V_4 and V_5 are processed, p_i installs edges from V_4 and V_5 to V_6. Suppose that *ReadInView* of V_4 and V_5 in line 56 return only the edge installed in the preceding line. Thus, V_4

and V_5 are removed from *Front*, and V_6 is added to *Front*, resulting in *Front* $= \{V_6\}$. During the next iteration *ReadInView*(V_6) and *WriteInView*(V_6) execute and both return $\emptyset$ in our execution. The condition in line 63 terminates the loop, V_6 is assigned to *curView*$_i$ and Traverse completes returning V_6.

5.3 Reconfigurations (Liveness)

A *reconfig(cng)* operation is similar to a *read*, with the only difference that *desiredView* initially contains the changes in *cng* in addition to those in *curView*$_i$ (line 48). Since *desiredView* only grows during a traversal, this ensures that the view returned from *Traverse* includes the changes in *cng*. As explained earlier, *Front* $=$ {*desiredView*} when *Traverse* completes, which means that *desiredView* appears in the DAG of views.

When a process p_i completes *WriteInView* in line 62 of *Traverse*, the latest state of the register has been transfered to *desiredView*, and thus it is no longer necessary for other processes to start traversals from earlier views. Thus, after Traverse completes returning *desiredView*, p_i invokes *NotifyQ* with this view as its parameter (lines 15, 20 and 25), to let other processes know about the new view. *NotifyQ*(w) sends a NOTIFY message (line 29) to w.*members*. A process receiving such a message for the first time forwards it to all processes in w.*members* (line 77), and after a NOTIFY message containing the same view was received from a majority of w.*members*, *NotifyQ* returns. In addition to forwarding the message, a process p_j receiving $\langle$NOTIFY, $w\rangle$ checks whether *curView*$_j \subset w$ (i.e., w is more up-to-date than *curView*$_j$), and if so it pauses any ongoing *Traverse*, assigns w to *curView*$_j$, and restarts *Traverse* from line 48. Restarting *Traverse* is necessary when p_j waits for responses from a majority of some view w' where less than a majority of members are active. Intuitively, Definition 1 implies that w' must be an old view, i.e., some *reconfig* operation completes proposing new changes to system membership. We prove in the full paper [2] that in this case p_j will receive a $\langle$NOTIFY, $w\rangle$ message s.t. *curView*$_j \subset w$ and restart its traversal.

Note that when a process p_i restarts *Traverse*, p_i may have an outstanding *scan*$_i$ or *update*$_i$ operation on a weak snapshot *ws*(w) for some view w, in which case p_i restarts *Traverse* without completing the operation. Later, it is possible that p_i needs to invoke another operation on *ws*(w). In that case, we require that p_i first terminates previous outstanding operations on *ws*(w) before it invokes the new operation. The mechanism to achieve this is a simple queue, and it is not illustrated in the code.

Restarts of *Traverse* introduce an additional potential complication for *write* operations: suppose that during its execution of *write*(v), p_i sends a WRITE message with v and a timestamp ts. It is important that if *Traverse* is restarted, v is not sent with a different timestamp (unless it belongs to some other write operation). Before the first message with v is sent, we set the *pickNewTS*$_i$ flag to *false* (line 72). The condition in line 71 prevents *Traverse* from re-assigning v to v_i^{max} or incorrect ts_i^{max}, even if a restart occurs.

In the full paper [2] we prove that DynaStore preserves Dynamic Service Liveness (Definition 1). Thus, liveness is conditioned on the number of different changes proposed in the execution being finite. In reality, only the number of such changes proposed concurrently with every operation has to be finite. Then, the number of times Traverse can be restarted during that operation is finite and so is the number of views encountered during the traversal, implying termination.

5.4 Sequence of Established Views (Safety)

Our traversal algorithm performs a *scan*(w) to discover outgoing edges from w. However, different processes might invoke *update*(w) concurrently, and different *scan*s might see different sets of outgoing edges. In such cases, it is necessary to prevent processes from working with views on different branches of the DAG. Specifically, we would like to ensure an intersection between views accessed in reads and writes. Fortunately, property NV4 guarantees that all *scan*(w) operations that return non-empty sets (i.e., return some outgoing edges from w), have at least one element (edge) in common. Note that a process cannot distinguish such an edge from others and therefore traverses all returned edges. This property of the algorithm enables us to define a totally ordered subset of the views, which we call *established*, as follows:

Definition 2. [Sequence of Established Views] The unique sequence of established views $\mathcal{E}$ is constructed as follows:

- the first view in $\mathcal{E}$ is the initial view $V(0)$;
- if w is in $\mathcal{E}$, then the next view after w in $\mathcal{E}$ is $w' = w \cup c$, where c is an element chosen arbitrarily from the intersection of all sets $C \neq \emptyset$ returned by some *scan*(w) operation in the execution.

Note that each element in the intersection mentioned in Definition 2 is a set of changes, and that property NV4 guarantees a non-empty intersection. In order to find such a set of changes c in the intersection, one can take an arbitrary element from the set C returned by the first *collect*(w) that returns a non-empty set in the execution. This unique sequence $\mathcal{E}$ allows us to define a total order relation on established views. For two established views w and w' we write $w \dot{\leq} w'$ if w appears in $\mathcal{E}$ no later than w'; if in addition $w \neq w'$ then $w \dot{<} w'$. Notice that for two established views w and w', $w \dot{<} w'$ if an only if $w \subset w'$.

Notice that the first graph traversal in the system starts from *curView*$_i = V(0)$, which is established by definition. When *Traverse* is invoked with an established view *curView*$_i$, every time a vertex w is removed from *Front* and its children are added, one of the children is an established view, by definition. Thus, *Front* always includes at least one established view, and since it ultimately contains only one view, *desiredView*, we conclude that *desiredView* assigned to *curView*$_i$ in line 64 and returned from *Traverse* is also established. Thus, all views sent in NOTIFY messages or stored in *curView*$_i$ are established. Note that while a process encounters all established views in its traversal, it only recognizes a subset of established views as such (whenever *Front* contains a single view, that view must be in $\mathcal{E}$).

It is easy to see that each traversal performs a *ReadInView* on every established view in $\mathcal{E}$ between *curView*$_i$ and the returned view *desiredView*. Notice that *WriteInView* (line 62) is always performed in an established view. Thus, intuitively, by reading each view encountered in a traversal, we are guaranteed to intersect any write completed on some established view in the traversed segment of $\mathcal{E}$. Then, performing the *scan* before *ContactQ* in *ReadInView* and after the *ContactQ* in *WriteInView* guarantees that in this intersection, indeed the state is transferred correctly, as explained in the beginning of this section. A formal correctness proof of our protocol appears in the full paper [2].

6. CONCLUSIONS

We defined a dynamic R/W storage problem, including an explicit liveness condition stated in terms of user interface and independent of a particular solution. The definition captures a dynamically changing resilience requirement, corresponding to reconfiguration operations invoked by users. Our approach easily carries to other problems, and allows for cleanly extending static problems to the dynamic setting.

We presented DynaStore, which is the first algorithm we are aware of to solve the atomic R/W storage problem in a dynamic setting without consensus or stronger primitives. In fact, we assumed a completely asynchronous model where fault-tolerant consensus is impossible even if no reconfigurations occur. This implies that atomic R/W storage is weaker than consensus, not only in static settings as was previously known, but also in dynamic ones. Our result thus refutes a common belief, manifested in the design of all previous dynamic storage systems, which used agreement to handle configuration changes. Our main goal in this paper was to prove feasibility; future work may study the performance tradeoffs between consensus-based solutions and consensus-free ones.

Acknowledgments

We thank Ittai Abraham, Eli Gafni, Leslie Lamport and Lidong Zhou for early discussions of this work.

7. REFERENCES

[1] Y. Afek, H. Attiya, D. Dolev, E. Gafni, M. Merritt, and N. Shavit. Atomic snapshots of shared memory. *J. ACM*, 40(4):873–890, 1993.

[2] M. K. Aguilera, I. Keidar, D. Malkhi, and A. Shraer. Dynamic atomic storage without consensus. Technical Report CCIT 731, Department of Electrical Engineering, Technion, 2009.

[3] H. Attiya, A. Bar-Noy, and D. Dolev. Sharing memory robustly in message-passing systems. *J. ACM*, 42(1):124–142, 1995.

[4] T. D. Chandra, V. Hadzilacos, S. Toueg, and B. Charron-Bost. On the impossibility of group membership. In *Proceedings of the 15th Annual ACM Symposium on Principles of Distributed Computing (PODC'96)*, pages 322–330, 1996.

[5] G. Chockler, S. Gilbert, V. C. Gramoli, P. M. . Musial, and A. A. Shvartsman. Reconfigurable distributed storage for dynamic networks. In *9th International Conference on Principles of Distributed Systems (OPODIS)*, 2005.

[6] G. Chockler, I. Keidar, and R. Vitenberg. Group communication specifications: A comprehensive study. *ACM Computing Surveys*, 33(4):1–43, 2001.

[7] T. T. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to algorithms*. MIT Press, Cambridge, MA, USA, 1990.

[8] D. Davcev and W. Burkhard. Consistency and recovery control for replicated files. In *10th ACM SIGOPS Symposium on Operating Systems Principles (SOSP)*, pages 87–96, 1985.

[9] C. Delporte, H. Fauconnier, R. Guerraoui, V. Hadzilacos, P. Kouznetsov, and S. Toueg. The weakest failure detectors to solve certain fundamental problems in distributed computing. In *Proceedings of the ACM Symposium on Principles of Distributed Computing (PODC 2004)*, pages 338–346, 2004.

[10] A. El Abbadi and S. Dani. A dynamic accessibility protocol for replicated databases. *Data and Knowledge Engineering*, 6:319–332, 1991.

[11] B. Englert and A. A. Shvartsman. Graceful quorum reconfiguration in a robust emulation of shared memory. In *ICDCS '00: Proceedings of the The 20th International Conference on Distributed Computing Systems (ICDCS 2000)*, page 454, Washington, DC, USA, 2000. IEEE Computer Society.

[12] S. Gilbert, N. Lynch, and A. Shvartsman. Rambo ii: Rapidly reconfigurable atomic memory for dynamic networks. In *Proceedings of the 17th Intl. Symp. on Distributed Computing (DISC)*, pages 259–268, June 2003.

[13] M. P. Herlihy and J. M. Wing. Linearizability: a correctness condition for concurrent objects. *ACM Trans. Program. Lang. Syst.*, 12(3):463–492, 1990.

[14] L. Lamport. On interprocess communication – part ii: Algorithms. *Distributed Computing*, 1(2):86–101, 1986.

[15] L. Lamport. The part-time parliament. *ACM Trans. Comput. Syst.*, 16(2):133–169, May 1998.

[16] L. Lamport, D. Malkhi, and L. Zhou. Brief announcement: Vertical paxos and primary-backup replication. In 28th ACM Symposium on Principles of Distributed Computing (PODC), August 2009. Full version appears as Microsoft Technical Report MSR-TR-2009-63, May 2009.

[17] E. K. Lee and C. A. Thekkath. Petal: Distributed virtual disks. In *Proceedings of the Seventh International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 84–92, Cambridge, MA, 1996.

[18] N. Lynch and A. Shvartsman. Robust emulation of shared memory using dynamic quorum-acknowledged broadcasts. In *In Symposium on Fault-Tolerant Computing*, pages 272–281. IEEE, 1997.

[19] N. A. Lynch and A. A. Shvartsman. RAMBO: A reconfigurable atomic memory service for dynamic networks. In *5th International Symposium on Distributed Computing (DISC)*, 2002.

[20] J. MacCormick, N. Murphy, M. Najork, C. A. Thekkath, and L. Zhou. Boxwood: Abstractions as the foundation for storage infrastructure. In *OSDI*, pages 105–120, 2004.

[21] J.-P. Martin and L. Alvisi. A framework for dynamic byzantine storage. In *Proceedings of the International Conference on Dependable Systems and Networks*, 2004.

[22] J. Paris and D. Long. Efficient dynamic voting algorithms. In *13th International Conference on Very Large Data Bases (VLDB)*, pages 268–275, 1988.

[23] R. Rodrigues and B. Liskov. Rosebud: A scalable byzantine-fault-tolerant storage architecture. Technical Report TR/932, MIT LCS, 2003.

[24] R. Rodrigues and B. Liskov. Reconfigurable byzantine-fault-tolerant atomic memory. In *Twenty-Third Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC)*, St. John's, Newfoundland, Canada, July 2004. Brief Announcement.

[25] F. B. Schneider. Implementing fault-tolerant services using the state machine approach: a tutorial. *ACM Comput. Surv.*, 22(4):299–319, 1990.

[26] R. van Renesse and F. B. Schneider. Chain replication for supporting high throughput and availability. In *Sixth Symposium on Operating Systems Design and Implementation (OSDI 04)*, December 2004.

[27] E. Yeger Lotem, I. Keidar, and D. Dolev. Dynamic voting for consistent primary components. In *16th ACM Symposium on Principles of Distributed Computing (PODC)*, pages 63–71, August 1997.