ARES: Adaptive, Reconfigurable, Erasure coded, atomic Storage

Nicolas Nicolaou[1], Viveck Cadambe[2], N. Prakash[3], Kishori M. Konwar[3], Muriel Medard[3] and Nancy Lynch[3]
[1]*KIOS Research and Innovation Center of Excellence, Univ. of Cyprus, Nicosia, Cyprus*
[2]*EE Department Penn. State University* University Park, PA, US
[3]*Department of EECS, Massachusetts Institute of Technology, Cambridge MA, USA*
nicolasn@cs.ucy.ac.cy, vxc12@engr.psu.edu, prakashn@mit.edu, kishori@csail.mit.edu, medard@mit.edu, lynch@csail.mit.edu

***Abstract*—Emulating a shared *atomic*, read/write storage system is a fundamental problem in distributed computing. Replicating atomic objects among a set of data hosts was the norm for traditional implementations (e.g., [6]) in order to guarantee the availability and accessibility of the data despite host failures. As replication is highly storage demanding, recent approaches suggested the use of erasure-codes to offer the same fault-tolerance while optimizing storage usage at the hosts. Initial works focused on a fix set of data hosts. To guarantee longevity and scalability, a storage service should be able to dynamically mask hosts failures by allowing new hosts to join, and failed host to be removed without service interruptions. This work presents the first erasure-code based atomic algorithm, called ARES, which allows the set of hosts to be modified in the course of an execution. ARES is composed of three main components: (i) a *reconfiguration protocol*, (ii) a *read/write protocol*, and (iii) a set of *data access primitives*. The design of ARES is modular and is such to accommodate the usage of various erasure-code parameters on a per-configuration basis. We provide bounds on the latency of read/write operations, and analyze the storage and communication costs of the ARES algorithm.**

I. INTRODUCTION

Distributed Storage Systems (DSS) store large amounts of data in an affordable manner. Cloud vendors deploy hundreds to thousands of commodity machines, networked together to act as a single giant storage system. Component failures of commodity devices, and network delays are the norm, therefore, ensuring consistent data-access and availability at the same time is challenging. Vendors often solve availability by replicating data across multiple servers. These services use carefully constructed algorithms that ensure that these copies are consistent, especially when they can be accessed concurrently by different operations. The problem of keeping copies consistent becomes even more challenging when failed servers need to be replaced or new servers are added, without interrupting the service. Any type of service interruption in a heavily used DSS usually translates to immense revenue loss.

The goal of this work is to provide an algorithm for implementing strongly consistent (i.e., atomic/linearizable), fault-tolerant distributed read/write storage, with low storage and communication footprint, and the ability to reconfigure the set of data hosts without service interruptions.

Replication-based Atomic Storage. A long stream of work used replication of data across multiple servers to implement atomic (linearizable) read/write objects in message-passing, asynchronous environments where servers (data hosts) may crash fail [5], [6], [14], [15], [16], [18], [19], [30]. A notable replication-based algorithm appears in the work by Attiya, Bar-Noy and Dolev [6] (we refer to as the ABD algorithm) which implemented non-blocking atomic read/write data storage via logical timestamps paired with values to order read/write operations. Replication based strategies, however, incur high storage and communication costs; for example, to store 1,000,000 objects each of size 1MB (a total size of 1TB) across a 3 server system, the ABD algorithm replicates the objects in all the 3 servers, which blows up the worst-case *storage cost* to 3TB. Additionally, every write or read operation may need to transmit up to 3MB of data (while retrieving an object value of size 1MB), incurring high *communication cost*.

Erasure Code-based Atomic Storage. Erasure Coded-based DSS are extremely beneficial to save storage and communication costs while maintaining similar fault-tolerance levels as in replication based DSS [11]. Mechanisms using an $[n, k]$ erasure code splits a value v of size, say 1 unit, into k elements, each of size $\frac{1}{k}$ units, creates n *coded elements* of the same size, and stores one coded element per server, for a total storage cost of $\frac{n}{k}$ units. So the $[n = 3, k = 2]$ code in the previous example will reduce the storage cost to 1.5TB and the communication cost to 1.5MB (improving also operation latency). Maximum Distance Separable (MDS) codes have the property that value v can be reconstructed from any k out of these n coded elements; note that replication is a special case of MDS codes with $k = 1$. In addition to the potential cost-savings, the suitability of erasure-codes for DSS is amplified with the emergence of highly optimized erasure coding libraries, that reduce encoding/decoding overheads [7], [1], [33]. In fact, an exciting recent body of systems and optimization works [41], [2], [33], [36], [38], [24], [37], [39] have demonstrated that for several data stores, the use of erasure coding results in lower latencies than replication based approaches. This is achieved by allowing the system to carefully tune erasure coding parameters, data placement strategies, and other system parameters that improve workload characteristics – such as load and spatial distribution. A complementary body of work has proposed novel non-blocking algorithms that use erasure coding to provide an atomic storage over asynchronous message passing models [8], [10], [13], [25], [26], [11], [40]. Since erasure code-based algorithms, unlike their replication-based counter parts, incur the additional burden of synchronizing the access of multiple pieces of coded-elements from the *same version* of the data object, these algorithms are quite complex.

Reconfigurable Atomic Storage. *Configuration* refers to the set of storage servers that are collectively used to host the data and implement the DSS. *Reconfiguration* is the process of adding or removing servers in a DSS. In practice, reconfigurations are often desirable by system administrators [4], for a wide range of purposes, especially during system maintenance. As the set of storage servers becomes older and unreliable they are replaced with new ones to ensure data-durability. Furthermore, to scale the storage service to increased or decreased load, larger (or smaller) configurations may be needed to be deployed. Therefore, in order to carry out such reconfiguration steps, in addition to the usual read and write operations, an operation called reconfig is invoked by reconfiguration clients. Performing reconfiguration of a system, without service interruption, is a very challenging task and an active area of research. RAMBO [29] and DynaStore [3] are two of the handful of algorithms [12], [17], [20], [23], [34], [35] that allows reconfiguration on live systems; all these algorithms are replication-based.

Despite the attractive prospects of creating strongly consistent DSS with low storage and communication costs, so far, no algorithmic framework for reconfigurable atomic DSS employed erasure coding for fault-tolerance, or provided any analysis of bandwidth and storage costs. Our paper fills this vital gap in algorithms literature, through the development of novel reconfigurable approach for atomic storage that use *erasure codes* for fault-tolerance. From a practical viewpoint, our work may be interpreted as a bridge between the systems optimization works [41], [2], [33], [36], [38], [24], [37], [39] and non-blocking erasure coded based consistent storage [8], [10], [13], [25], [26], [11], [40]. Specifically, the use of our *reconfigurable* algorithms would potentially enable a data storage service to dynamically shift between different erasure coding based parameters and placement strategies, as the demand characteristics (such as load and spatial distribution) change, without service interruption.

Our Contributions. We develop a *reconfigurable*, *erasure-coded*, *atomic* or *strongly consistent* [21], [28] read/write storage algorithm, called ARES. Motivated by many practical systems, ARES assumes clients and servers are separate processes * that communicate via logical point-to-point channels.

In contrast to the, replication-based reconfigurable algorithms [29], [3], [12], [17], [20], [23], [34], [35], where a configuration essentially corresponds to the set of servers that stores the data, the same concept for erasure coding need to be much more involved. In particular, in erasure coding, even if the same set of n servers are used, a change in the value of k defines a new configuration. Furthermore, several erasure coding based algorithms [10], [13] have additional parameters that tune how many older versions each server store, which in turn influences the concurrency level allowed. Tuning of such parameters can also fall under the purview of reconfiguration.

To accommodate these various reconfiguration requirements, ARES takes a modular approach. In particular, ARES uses a set of primitives, called *data-access primitives* (DAPs). A different implementation of the DAP primitives may be specified in each configuration. ARES uses DAPs as a "black box" to: (i) transfer the object state from one configuration to the next during reconfig operations, and (ii) invoke read/write operations on a single configuration. Given the DAP implementation for each configuration we show that ARES correctly implements a *reconfigurable*, *atomic* read/write storage.

The DAP primitives provide ARES a much broader view of the notion of a configuration as compared to replication-based algorithms. Specifically, the DAP primitives may be parameterized, following the parameters of protocols used for their implementation (e.g., erasure coding parameters, set of servers, quorum design, concurrency level, etc.). While transitioning from one configuration to another, our modular construction, allows ARES to reconfigure between different sets of servers, quorum configurations, and erasure coding parameters. In principle, ARES even allows to reconfigure between completely different protocols as long as they can be interpreted/expressed in terms of the primitives; though in this paper, we only present one implementation of the DAP primitives to keep the scope of the paper reasonable. From a technical point of view, our modular structure makes the atomicity proof of a complex algorithm (like ARES) easier.

An important consideration in the design choice of ARES, is to ensure that we gain/retain the advantages that come with erasure codes – cost of data storage and communication is low – while having the flexibility to reconfigure the system. Towards this end, we present an erasure-coded implementation of DAPs which satisfy the necessary properties, and are used by ARES to yield the first reconfigurable, *erasure-coded*, read/write atomic storage implementation, where read and write operations complete in *two-rounds*. We provide the atomicity property and latency analysis for any operation in ARES, along with the storage and communication costs resulting from the erasure-coded DAP implementation. In particular, we specify lower and upper bounds on the communication latency between the service participants, and we provide the necessary conditions to guarantee the termination of each read/write operation while concurrent with reconfig operations.

Table I compares some characteristics of ARES with a few well-known erasure-coded atomic memory algorithms and also the static and reconfigurable replication-based algorithms. From the table we observe that ARES is the only algorithm to combine a dynamic behavior with the use of erasure codes, and also the only to allow adaptive change of the atomic algorithm per configuration, while reducing the storage and communcation costs associated with the read or write operations. Moreover, in ARES the number of rounds per write and read is at least as good as in any of the remaining algorithms.

Document Structure. In Section II, we present the model assumptions for our setting and in Section III, the DAP primitives. In Section IV, we present the implementation of the reconfiguration and read/write protocols in ARES using the DAPs. In Section V, we present an erasure-coded implementation of a set of DAPs, which can be used in every configuration of the ARES algorithm. Section VI provides operation latency and cost analysis, and Section VII the DAP flexibility. We conclude our work in Section VIII. Due to lack of space omitted proofs can be found in [31].

*Note that in practice, these processes can be on the same node or different nodes.

Algorithm	#rounds /write	#rounds /read	Reconfig.	Repl. or EC	Storage cost	read bandwidth	write bandwidth
CASGC [9]	3	2	No	EC	$(\delta+1)\frac{n}{k}$	$\frac{n}{k}$	$\frac{n}{k}$
SODA [25]	2	2	No	EC	$\frac{n}{k}$	$(\delta+1)\frac{n}{k}$	$\frac{n^2}{k}$
ORCAS-A [13]	3	≥ 2	No	EC	n	n	n
ORCAS-B [13]	3	3	No	EC	∞	∞	∞
ABD [6]	2	2	No	Repl.	n	$2n$	n
RAMBO [29]	2	2	Yes	Repl.	$\geq n$	$\geq n$	$\geq n$
DYNASTORE [3]	≥ 4	≥ 4	Yes	Repl.	$\geq n$	$\geq n$	$\geq n$
SMARTMERGE [23]	2	2	Yes	Repl.	$\geq n$	$\geq n$	$\geq n$
ARES (this paper)	2	2	Yes	EC	$(\delta+1)\frac{n}{k}$	$(\delta+1)\frac{n}{k}$	$\frac{n}{k}$

TABLE I: Comparison of ARES with previous algorithms emulating atomic Read/Write Memory for replication (Repl.) and erasure-code based (EC) algorithms. δ is the maximum number of concurrent writes with any read during the course of an execution of the algorithm. In practice, $\delta < 4$ [11].

II. MODEL AND DEFINITIONS

A shared atomic storage, consisting of any number of individual objects, can be emulated by composing individual atomic memory objects. Therefore, herein we aim in implementing a single atomic *read/write* memory object. A read/write object takes a value from a set $\mathcal{V}$. We assume a system consisting of four distinct sets of processes: a set $\mathcal{W}$ of writers, a set $\mathcal{R}$ of readers, a set $\mathcal{G}$ of reconfiguration clients, and a set $\mathcal{S}$ of servers. Let $\mathcal{I} = \mathcal{W} \cup \mathcal{R} \cup \mathcal{G}$ be the set of clients. Servers host data elements (replicas or encoded data fragments). Each writer is allowed to modify the value of a shared object, and each reader is allowed to obtain the value of that object. Reconfiguration clients attempts to introduce new configuration of servers to the system in order to mask transient errors and to ensure the longevity of the service. Processes communicate via *messages* through *asynchronous*, and *reliable* channels.

Configurations. A *configuration*, with a unique identifier from a set $\mathcal{C}$, is a data type that describes the finite set of servers that are used to implement the atomic storage service. In our setting, each configuration is also used to describe the way the servers are grouped into intersecting sets, called *quorums*, the consensus instance that is used as an external service to determine the next configuration, and a set of data access primitives that specify the interaction of the clients and servers in the configuration (see Section III). More formally, a configuration, $c \in \mathcal{C}$, consists of: (i) $c.Servers \subseteq \mathcal{S}$: a set of server identifiers; (ii) $c.Quorums$: the set of quorums on $c.Servers$, s.t. $\forall Q_1, Q_2 \in c.Quorums, Q_1, Q_2 \subseteq c.Servers$ and $Q_1 \cap Q_2 \neq \emptyset$; (iii) $DAP(c)$: the set of primitives (operations at level lower than reads or writes) that clients in $\mathcal{I}$ may invoke on $c.Servers$; and (iv) $c.Con$: a consensus instance with the values from $\mathcal{C}$, implemented and running on top of the servers in $c.Servers$. We refer to a server $s \in c.Servers$ as a *member* of configuration c. The consensus instance $c.Con$ in each configuration c is used as a service that returns the identifier of the configuration that follows c.

Executions. An algorithm A is a collection of processes, where process A_p is assigned to processor $p \in \mathcal{I} \cup \mathcal{S}$. The *state*, of a process A_p is determined over a set of state variables, and the state σ of A is a vector that contains the state of each process. Each process A_p implements a set of actions. When an action α occurs it causes the state of A_p to change, say from some state σ_p to some different state σ'_p. We call the triple $\langle \sigma_p, \alpha, \sigma'_p \rangle$ a *step* of A_p. Algorithm A performs a step, when some process A_p performs a step. An action α is *enabled* in a state σ if $\exists$ a step $\langle \sigma, \alpha, \sigma' \rangle$ to some state σ'. An *execution* is an alternating sequence of states and actions of A starting with the initial state and ending in a state. An execution ξ is *well-formed* if any process invokes one operation at a time and it is *fair* if enabled actions perform a step infinitely often. In the rest of the paper we consider executions that are fair and well-formed. A process p *crashes* in an execution if it stops taking steps; otherwise p is *correct* or *non-faulty*. We assume there exists a function $c.\mathcal{F}$ that describes the failure model of a configuration c.

Reconfigurable Atomic Read/Write Objects. A reconfigurable atomic object supports three operations: read(), write(v) and reconfig(c). A read() operation returns the value of the atomic object, write(v) attempts to modify the value of the object to $v \in \mathcal{V}$, and the reconfig(c) that attempts to install a new configuration $c \in \mathcal{C}$. We assume *well-formed* executions where each client may invoke a single operation (read(), write(v) or reconfig(c)) at a time.

An implementation of a read/write or a reconfig operation contains an *invocation* action (such as a call to a procedure) and a *response* action (such as a return from the procedure). An operation π is *complete* in an execution, if it contains both the invocation and the *matching* response actions for π; otherwise π is *incomplete*. We say that an operation π *precedes* an operation π' in an execution ξ, denoted by $\pi \rightarrow \pi'$, if the response step of π appears before the invocation step of π' in ξ. Two operations are *concurrent* if neither precedes the other. An implementation A of a read/write object satisfies the atomicity property if the following holds [28]. Let the set Π contain all complete operations in any well-formed execution of A. Then for operations in Π there exists an irreflexive partial ordering $\prec$ satisfying the following:

A1. For any operations π_1 and π_2 in Π, if $\pi_1 \rightarrow \pi_2$, then it cannot be the case that $\pi_2 \prec \pi_1$.

A2. If $\pi \in \Pi$ is a write operation and $\pi' \in \Pi$ is any operation,

then either $\pi \prec \pi'$ or $\pi' \prec \pi$.

A3. The value returned by a read operation is the value written by the last preceding write operation according to $\prec$ (or the initial value if there is no such write).

Storage and Communication Costs. We are interested in the *complexity* of each read and write operation. The complexity of each operation π invoked by a process p, is measured with respect to three metrics, during the interval between the invocation and the response of π: (i) *number of communication round*, accounting the number of messages exchanged during π, (ii) *storage efficiency* (storage cost), accounting the maximum storage requirements for any single object at the servers during π, and (iii) *message bit complexity* (communication cost) which measures the size of the messages used during π.

We define the total storage cost as the size of the data stored across all servers, at any point during the execution of the algorithm. The communication cost associated with a read or write operation is the size of the total data that gets transmitted in the messages sent as part of the operation. We assume that metadata, such as version number, process ID, etc. used by various operations is of negligible size, and is hence ignored in the calculation of storage and communication cost. Further, we normalize both costs with respect to the size of the value v; in other words, we compute the costs under the assumption that v has size 1 unit.

Erasure Codes. We use an $[n, k]$ linear MDS code [22] over a finite field $\mathbb{F}_q$ to encode and store the value v among the n servers. An $[n, k]$ MDS code has the property that any k out of the n coded elements can be used to recover (decode) the value v. For encoding, v is divided into k elements $v_1, v_2, \ldots v_k$ with each element having size $\frac{1}{k}$ (assuming size of v is 1). The encoder takes the k elements as input and produces n coded elements $e_1, e_2, \ldots, e_n$ as output, i.e., $[e_1, \ldots, e_n] = \Phi([v_1, \ldots, v_k])$, where Φ denotes the encoder. For ease of notation, we simply write $\Phi(v)$ to mean $[e_1, \ldots, e_n]$. The vector $[e_1, \ldots, e_n]$ is referred to as the codeword corresponding to the value v. Each coded element c_i also has size $\frac{1}{k}$. In our scheme we store one coded element per server. We use Φ_i to denote the projection of Φ on to the i^{th} output component, i.e., $e_i = \Phi_i(v)$. Without loss of generality, we associate the coded element e_i with server i, $1 \leq i \leq n$.

Tags. We use logical tags to order operations. A tag τ is defined as a pair (z, w), where $z \in \mathbb{N}$ and $w \in \mathcal{W}$, an ID of a writer. Let $\mathcal{T}$ be the set of all tags. Notice that tags could be defined in any totally ordered domain and given that this domain is countably infinite, then there can be a direct mapping to the domain we assume. For any $\tau_1, \tau_2 \in \mathcal{T}$ we define $\tau_2 > \tau_1$ if (i) $\tau_2.z > \tau_1.z$ or (ii) $\tau_2.z = \tau_1.z$ and $\tau_2.w > \tau_1.w$.

III. Data Access Primitives

In this section we introduce a set of primitives, we refer to as *data access primitives (DAP)*, which are invoked by the clients during read/write/reconfig operations and are defined for any configuration c in ARES. The DAPs allow us: (i) to describe ARES in a *modular* manner, and (ii) a cleaner reasoning about the correctness of ARES.

We define three data access primitives for each $c \in \mathcal{C}$: $c.\mathsf{put\text{-}data}(\langle \tau, v \rangle)$, via which a client can ingest the tag value pair $\langle \tau, v \rangle$ in to the configuration c; (ii) $c.\mathsf{get\text{-}data}()$, used to retrieve the most up to date tag and vlaue pair stored in the configuration c; and (iii) $c.\mathsf{get\text{-}tag}()$, used to retrieve the most up to date tag for an object stored in a configuration c. More formally, assuming a tag τ from a set of totally ordered tags $\mathcal{T}$, a value v from a domain $\mathcal{V}$, and a configuration c from a set of identifiers $\mathcal{C}$, the three primitives are defined as follows:

Definition 1 (Data Access Primitives)**.** *Given a configuration identifier $c \in \mathcal{C}$, any non-faulty client process p may invoke the following data access primitives during an execution ξ, where c is added to specify the configuration specific implementation of the primitives:*

D1: $c.\mathsf{get\text{-}tag}()$ *that returns a tag $\tau \in \mathcal{T}$;*

D2: $c.\mathsf{get\text{-}data}()$ *that returns a tag-value pair $(\tau, v) \in \mathcal{T} \times \mathcal{V}$,*

D3: $c.\mathsf{put\text{-}data}(\langle \tau, v \rangle)$ *which accepts the tag-value pair $(\tau, v) \in \mathcal{T} \times \mathcal{V}$ as argument.*

In order for the DAPs to be useful in designing the ARES algorithm we further require the following consistency properties. As we see later in Section VI-B, the safety property of ARES holds, given that these properties hold for the DAPs in each configuration.

Property 1 (DAP Consistency Properties)**.** *In an execution ξ we say that a DAP operation in an execution ξ is complete if both the invocation and the matching response step appear in ξ. If Π is the set of complete DAP operations in execution ξ then for any $\phi, \pi \in \Pi$:*

C1 *If ϕ is $c.\mathsf{put\text{-}data}(\langle \tau_\phi, v_\phi \rangle)$, for $c \in \mathcal{C}$, $\langle \tau_\phi, v_\phi \rangle \in \mathcal{T} \times \mathcal{V}$, and π is $c.\mathsf{get\text{-}tag}()$ (or $c.\mathsf{get\text{-}data}()$) that returns $\tau_\pi \in \mathcal{T}$ (or $\langle \tau_\pi, v_\pi \rangle \in \mathcal{T} \times \mathcal{V}$) and ϕ completes before π is invoked in ξ, then $\tau_\pi \geq \tau_\phi$.*

C2 *If ϕ is a $c.\mathsf{get\text{-}data}()$ that returns $\langle \tau_\pi, v_\pi \rangle \in \mathcal{T} \times \mathcal{V}$, then there exists π such that π is $c.\mathsf{put\text{-}data}(\langle \tau_\pi, v_\pi \rangle)$ and ϕ did not complete before the invocation of π. If no such π exists in ξ, then (τ_π, v_π) is equal to (t_0, v_0).*

In Section V we show how to implement a set of DAPs, where erasure-codes are used to reduce storage and communication costs. Our DAP implementation satisfies Property 1.

As noted earlier, expressing ARES in terms of the DAPs allows one to achieve a modular design. Modularity enables the usage of different DAP implementation per configuration, during any execution of ARES, as long as the DAPs implemented in each configuration satisfy Property 1. For example, the DAPs in a configuration c may be implemented using replication, while the DAPs in the next configuration say c', may be implemented using erasure-codes. This in turn, allows the usage of different coding algorithms and code parameters in each individual configuration without affecting the correctness of ARES.

In Section VII, we show that the presented DAPs are not only suitable for algorithm ARES, but can also be used to implement a large family of atomic read/write storage implementations. By describing an algorithm A according to a simple algorithmic template (see Alg. 7), we show that A preserves safety (atomicity) if the used DAPs satisfy Property 1, and A preserves liveness (termination), if every invocation of the used DAPs terminate, under the failure model assumed.

IV. ARES PROTOCOL

In this section, we describe ARES. In the presentation of ARES algorithm we decouple the reconfiguration service from the shared memory emulation, by utilizing the DAPs presented in Section III. This allows ARES, to handle both the reorganization of the servers that host the data, as well as utilize a different atomic memory implementation per configuration. It is also important to note that ARES adopts a client-server architecture and separates the reader, writer and reconfiguration processes from the server processes that host the object data. More precisely, ARES algorithm comprises of three major components: (i) The reconfiguration protocol which consists of invoking, and subsequently installing new configuration via the reconfig operation by recon clients. (ii) The read/write protocol for executing the read and write operations invoked by readers and writers. (iii) The implementation of the DAPs for each installed configuration that respect Property 1 and which are used by the reconfig, read and write operations.

A. Implementation of the Reconfiguration Service.

In this section, we describe the reconfiguration service in ARES. The service relies on an underlying sequence of configurations (already proposed or installed by reconfig operations), in the from of a "distributed list", which we refer to as the *global configuration sequence (or list)* $\mathcal{G}_L$. Conceptually, $\mathcal{G}_L$ represents an ordered list of pairs $\langle c, status\rangle$, where c is a configuration identifier ($c \in \mathcal{C}$), and a binary state variable $status \in \{F, P\}$ that denotes whether c is *finalized* (F) or is still *pending* (P). Initially, $\mathcal{G}_L$ contains a single element, say $\langle c_0, F\rangle$, which is known to every participant in the service.

To facilitate the creation of $\mathcal{G}_L$, each server in $c.Servers$ maintains a local variable $nextC \in \{\mathcal{C} \cup \{\bot\}\} \times \{P, F\}$, which is used to point to the configuration that follows c in $\mathcal{G}_L$. Initially, at any server $nextC = \langle \bot, F\rangle$. Once $nextC$ it is set to a value it is never altered. As we show below, at any point in the execution of ARES and in any configuration c, the $nextC$ variables of the non-faulty servers in c that are not equal to $\bot$ agree, i.e., $\{s.nextC : s \in c.Servers \wedge s.nextC \neq \bot\}$ is either empty of has only one element.

Clients discover the configuration that follows a $\langle c, *\rangle$ in the sequence by contacting a subset of servers in $c.Servers$ and collecting their $nextC$ variables. Every client in $\mathcal{I}$ maintains a local variable $cseq$ that is expected to be some subsequence of $\mathcal{G}_L$. Initially, at every client the value of $cseq$ is $\langle c_0, F\rangle$. We use the notation $\widehat{x}$ (a caret over some name) to denote state variables that assumes values from the domain $\{\mathcal{C} \cup \{\bot\}\} \times \{P, F\}$.

Reconfiguration clients may introduce new configurations, each associated with a unique configuration identifier from $\mathcal{C}$. Multiple clients may concurrently attempt to introduce different configurations for same next link in $\mathcal{G}_L$. ARES uses consensus to resolve such conflicts. In particular, a subset of servers in $c.Servers$, in each configuration c, implements a distributed consensus service (such as Paxos [27], RAFT [32]) , denoted by $c.Con$.

The reconfiguration service consists of two major components: (i) *sequence traversal*, responsible of discovering a recent configuration in $\mathcal{G}_L$, and (ii) *reconfiguration operation* that installs new configurations in $\mathcal{G}_L$.

Sequence Traversal. Any read/write/reconfig operation π utilizes the sequence traversal mechanism to discover the latest state of the global configuration sequence, as well as to ensure that such a state is discoverable by any subsequent operation π'. See Fig. 1 for an example execution in the case of a reconfig operation. In a high level, a client starts by collecting the $nextC$ variables from a quorum of servers in a configuration c, such that $\langle c, F\rangle$ is the last finalized configuration in that client's local $cseq$ variable (or c_0 if no other finalized configuration exists). If any server s returns a $nextC$ variable such that $nextC.cfg \neq \bot$, then the client (i) adds $nextC$ in its local $cseq$, (ii) propagates $nextC$ in a quorum of servers in $c.Servers$, and (iii) repeats this process in the configuration $nextC.cfg$. The client terminates when all servers reply with $nextC.cfg = \bot$. More precisely, the sequence parsing consists of three actions (see Alg. 1):

get-next-config(c): The action get-next-config returns the configuration that follows c in $\mathcal{G}_L$. During get-next-config(c), a client sends READ-CONFIG messages to all the servers in $c.Servers$, and waits for replies containing $nextC$ from a quorum in $c.Quorums$. If there exists a reply with $nextC.cfg \neq \bot$ the action returns $nextC$; otherwise it returns $\bot$.

put-config(c, c'): The put-config(c, c') action propagates c' to a quorum of servers in $c.Servers$. During the action, the client sends (WRITE-CONFIG, c') messages, to the servers in $c.Servers$ and waits for each server s in some quorum $Q \in c.Quorums$ to respond.

read-config(seq): A read-config(seq) sequentially traverses the installed configurations in order to discover the latest state of the sequence $\mathcal{G}_L$. At invocation, the client starts with the last finalized configuration $\langle c, F\rangle$ in the given seq (Line A1:2), and enters a loop to traverse $\mathcal{G}_L$ by invoking get-next-config(c), which returns the next configuration, say $\widehat{c}'$. While $\widehat{c}' \neq \bot$, then: (a) $\widehat{c}'$ is appended at the end of the sequence seq; (b) a put-config($c, \widehat{c}'$) is invoked to inform a quorum of servers in $c.Servers$ to update the value of their $nextC$ variable to $\widehat{c}'$; and (c) variable c is set to $\widehat{c}'.cfg$. If $\widehat{c}' = \bot$ the loop terminates and the action read-config returns seq.

Reconfiguration operation. A reconfiguration operation reconfig(c), $c \in \mathcal{C}$, invoked by any reconfiguration client rec_i, attempts to append c to $\mathcal{G}_L$. The set of server processes in c are not a part of any other configuration different from c. In a high-level, rec_i first executes a sequence traversal to discover the latest state of $\mathcal{G}_L$. Then it attempts to add the new configuration c, at the end of the discovered sequence by proposing c in the consensus instance of the last configuration in the sequence. The client accepts and appends the decision of the consensus instance (that might be different than c). Then it attempts to transfer the latest value of the read/write object to the latest installed configuration. Once the information is transferred, rec_i finalizes the last configuration in its local sequence and propagates the finalized tuple to a quorum of servers in that configuration. The operation consists of four phases, executed consecutively by rec_i (see Alg. 2):

read-config(seq): The phase read-config(seq) at rec_i, reads the recent global configuration sequence as described in the sequence traversal.

add-config(seq, c): The add-config(seq, c) attempts to ap-

Algorithm 1 Sequence traversal at each process $p \in \mathcal{I}$ of algorithm ARES.

```
    procedure read-config(seq)
2:      μ = max({j : seq[j].status = F})
        ĉ ← seq[μ]
4:      while ĉ ≠ ⊥ do
            ĉ' ←get-next-config(ĉ.cfg)
6:          if ĉ' ≠ ⊥ then
                μ ← μ + 1
8:              seq[μ] ← ĉ'
                put-config(seq[μ − 1].cfg, seq[μ])
10:             ĉ ← seq[μ]
        end while
12:     return seq
    end procedure

14: procedure get-next-config(c)
        send (READ-CONFIG) to each s ∈ c.Servers
16:     until ∃Q, Q ∈ c.Quorums s.t. rec_i receives nextC_s from ∀s ∈ Q
        if ∃s ∈ Q s.t. nextC_s.status = F then
18:         return nextC_s
        else if ∃s ∈ Q s.t. nextC_s.status = P then
20:         return nextC_s
        else
22:         return ⊥
    end procedure

24: procedure put-config(c, nextC)
        send (WRITE-CONFIG, nextC) to each s ∈ c.Servers
26:     until ∃Q, Q ∈ c.Quorums s.t. rec_i receives ACK from ∀s ∈ Q
    end procedure
```

Algorithm 2 Reconfiguration protocol of algorithm ARES.

```
    at each reconfigurer rec_i
2:  State Variables:
    cseq[]s.t.cseq[j] ∈ C × {F, P} with members:
4:  Initialization:
    cseq[0] = ⟨c_0, F⟩

6:  operation reconfig(c)
        if c ≠ ⊥ then
8:          cseq ←read-config(cseq)
            cseq ← add-config(cseq, c)
10:         update-config(cseq)
            cseq ← finalize-config(cseq)
12: end operation

    procedure add-config(seq, c)
14:     ν ← |seq|
        c' ← seq[ν].cfg
16:     d ← c'.Con.propose(c)
        seq[ν + 1] ← ⟨d, P⟩
18:     put-config(c', ⟨d, P⟩)
        return seq
20: end procedure

    procedure update-config(seq)
22:     μ ← max({j : seq[j].status = F})
        ν ← |seq|
24:     M ← ∅
        for i = μ : ν do
26:         ⟨t, v⟩ ← seq[i].cfg.get-data()
            M ← M ∪ {⟨τ, v⟩}
28:     ⟨τ, v⟩ ← max_t{⟨t, v⟩ : ⟨t, v⟩ ∈ M}
        seq[ν].put-data(⟨τ, v⟩)
30: end procedure

    procedure finalize-config(seq)
32:     ν = |seq|
        seq[ν].status ← F
34:     put-config(seq[ν − 1].cfg, seq[ν])
        return seq
36: end procedure
```

Algorithm 3 Server protocol of algorithm ARES.

```
    at each server s_i in configuration c_k
2:  State Variables:
    τ ∈ ℕ × W, initially, ⟨0, ⊥⟩
4:  v ∈ V, intially, ⊥
    nextC ∈ C × {P, F}, initially ⟨⊥, P⟩

6:  Upon receive (READ-CONFIG)_{s_i, c_k} from q
        send nextC to q
8:  end receive

    Upon receive (WRITE-CONFIG, cfgT_in)_{s_i, c_k} from q
10:     if nextC.cfg = ⊥ ∨ nextC.status = P then
            nextC ← cfgT_in
12:     send ACK to q
    end receive
```

pend a new configuration c to the end of seq (client's view of $\mathcal{G}_L$). Suppose the last configuration in seq is c' (with status either F or P), then in order to decide the most recent configuration, rec_i invokes $c'.Con.propose(c)$, on the consensus object associated with configuration c'. Let $d \in \mathcal{C}$ be the configuration identifier decided by the consensus service. If $d \neq c$, this implies that another (possibly concurrent) reconfiguration operation, invoked by a reconfigurer $rec_j \neq rec_i$, proposed and succeeded d as the configuration to follow c'. In this case, rec_i adopts d as it own propose configuration, by adding $\langle d, P\rangle$ to the end of its local $cseq$ (entirely ignoring c), using the operation put-config$(c', \langle d, P\rangle)$, and returns the extended configuration seq.

update-config(seq): Let us denote by μ the index of the last configuration in the local sequence $cseq$, at rec_i, such that its corresponding status is F; and ν denote the last index of $cseq$. Next rec_i invokes update-config$(cseq)$, which gathers the tag-value pair corresponding to the maximum tag in each of the configurations in $\widehat{cseq[i]}$ for $\mu \leq i \leq \nu$, and transfers that pair to the configuration that was added by the add-config action. The get-data and put-data DAPs are used to transfer the value of the object to the new configuration, and they are implemented

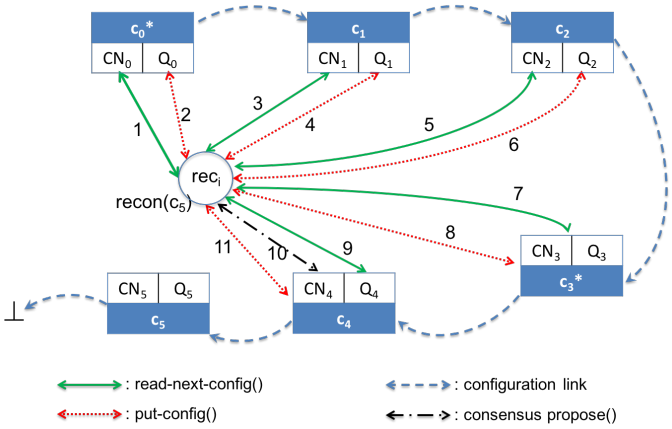

Fig. 1: Illustration of an execution of the reconfiguration steps.

with respect to the configuration that is accessed. Suppose $\langle t_{max}, v_{max}\rangle$ is the tag value pair corresponding to the highest tag among the responses from all the $\nu - \mu + 1$ configurations. Then, $\langle t_{max}, v_{max}\rangle$ is written to the configuration d via the invocation of $\widehat{cseq[\nu]}.cfg$.put-data($\langle \tau_{max}, v_{max}\rangle$).

finalize-config($cseq$): Once the tag-value pair is transferred, in the last phase of the reconfiguration operation, rec_i executes finalize-config($cseq$), to update the status of the last configuration in $cseq$, say $d = \widehat{cseq[\nu]}.cfg$, to F. The reconfigurer rec_i informs a quorum of servers in the previous configuration $c = \widehat{cseq[\nu-1]}.cfg$, i.e. in some $Q \in c.Quorums$, about the change of status, by executing the put-config($c, \langle d, F\rangle$) action.

Server Protocol. Each server responds to requests from clients (Alg. 3). A server waits for two types of messages: READ-CONFIG and WRITE-CONFIG. When a READ-CONFIG message is received for a particular configuration c_k, then the server returns $nextC$ variables of the servers in $c_k.Servers$. A WRITE-CONFIG message attempts to update the $nextC$ variable of the server with a particular tuple $cfgT_{in}$. A server changes the value of its local $nextC.cfg$ in two cases: (i) $nextC.cfg = \perp$, or (ii) $nextC.status = P$.

Fig. 1 illustrates an example execution of a reconfiguration operation recon(c_5). In this example, the reconfigurer rec_i goes through a number of configuration queries (read-next-config) before it reaches configuration c_4 in which a quorum of servers replies with $nextC.cfg = \perp$. There it proposes c_5 to the consensus object of c_4 ($c_4.Con.propose(c_5)$ on arrow 10), and once c_5 is decided, recon(c_5) completes after executing finalize-config(c_5).

B. Implementation of Read and Write operations.

The read and write operations in ARES are expressed in terms of the DAP primitives (see Section III). This provides the flexibility to ARES to use different implementation of DAP primitives in different configurations, without changing the basic structure of ARES. At a high-level, a write (or read) operation is executed where the client: (i) obtains the *latest configuration sequence* by using the read-config action of the reconfiguration service, (ii) queries the configurations, in $cseq$, starting from the last finalized configuration to the end of the discovered configuration sequence, for the latest $\langle tag, value\rangle$ pair with a help of get-tag (or get-data) operation as specified for each configuration, and (iii) repeatedly propagates a new $\langle tag', value'\rangle$ pair (the largest $\langle tag, value\rangle$ pair) with put-data in the last configuration of its local sequence, until no additional configuration is observed. In more detail, the algorithm of a read or write operation π is as follows (see Alg. 4):

A write (or read) operation is invoked at a client p when line Alg. 4:8 (resp. line Alg. 4:31) is executed. At first, p issues a read-config action to obtain the latest introduced configuration in $\mathcal{G}_L$, in both operations.

If π is a write p detects the last finalized entry in $cseq$, say μ, and performs a $cseq[j].conf$.get-tag() action, for $\mu \leq j \leq |cseq|$ (line Alg. 4:9). Then p discovers the *maximum tag* among all the returned tags (τ_{max}), and it increments the maximum tag discovered (by incrementing the integer part of τ_{max}), generating a new tag, say τ_{new}. It assigns $\langle \tau, v\rangle$ to $\langle \tau_{new}, val\rangle$, where val is the value he wants to write (Line Alg. 4:13).

if π is a read, p detects the last finalized entry in $cseq$, say μ, and performs a $cseq[j].conf$.get-data() action, for $\mu \leq j \leq |cseq|$ (line Alg. 4:32). Then p discovers the *maximum tag-value* pair ($\langle \tau_{max}, v_{max}\rangle$) among the replies, and assigns $\langle \tau, v\rangle$ to $\langle \tau_{max}, v_{max}\rangle$.

Once specifying the $\langle \tau, v\rangle$ to be propagated, both reads and writes execute the $cseq[\nu].cfg$.put-data($\langle \tau, v\rangle$) action, where $\nu = |cseq|$, followed by executing read-config action, to examine whether new configurations were introduced in $\mathcal{G}_L$. The repeat these steps until no new configuration is discovered (lines Alg. 4:15–21, or lines Alg. 4:37–43). Let $cseq'$ be the sequence returned by the read-config action. If $|cseq'| = |cseq|$ then no new configuration is introduced, and the read/write operation terminates; otherwise, p sets $cseq$ to $cseq'$ and repeats the two actions. Note, in an execution of ARES, two consecutive read-config operations that return $cseq'$ and $cseq''$ respectively must hold that $cseq'$ is a prefix of $cseq''$, and hence $|cseq'| = |cseq''|$ only if $cseq' = cseq''$. Finally, if π is a read operation the value with the highest tag discovered is returned to the client.

V. Implementation of the DAPs

In this section, we present an implementation of the DAPs, that satisfies the properties in Property 1, for a configuration c, with n servers using a $[n, k]$ MDS coding scheme for storage. We implement an instance of the algorithm in a configuration of n server processes. We store each coded element c_i, corresponding to an object at server s_i, where $i = 1, \cdots, n$. The implementations of DAP primitives used in ARES are shown in Alg. 5, and the servers' responses in Alg. 6.

Each server s_i stores one state variable, $List$, which is a set of up to $(\delta + 1)$ (tag, coded-element) pairs. Initially the set at s_i contains a single element, $List = \{(t_0, \Phi_i(v_0)\}$. Below we describe the implementation of the DAPs.

c.get-tag(): A client, during the execution of a c.get-tag() primitive, queries all the servers in $c.Servers$ for the highest tags in their $Lists$, and awaits responses from $\lceil \frac{n+k}{2} \rceil$ servers. A server upon receiving the GET-TAG request, responds to the client with the highest tag, as $\tau_{max} \equiv \max_{(t,c)\in List} t$. Once the client receives the tags from $\lceil \frac{n+k}{2} \rceil$ servers, it selects the highest tag t and returns it .

Algorithm 4 Write and Read protocols at the clients for ARES.

Write Operation:
at each writer w_i
State Variables:
$cseq[]s.t.cseq[j] \in \mathcal{C} \times \{F, P\}$ with members:
Initialization:
$cseq[0] = \langle c_0, F \rangle$

operation write(val), $val \in V$
  $cseq \leftarrow$read-config($cseq$)
  $\mu \leftarrow \max(\{i : cseq[i].status = F\})$
  $\nu \leftarrow |cseq|$
  for $i = \mu : \nu$ **do**
    $\tau_{max} \leftarrow \max(cseq[i].cfg.\text{get-tag}(), \tau_{max})$
  $\langle \tau, v \rangle \leftarrow \langle \langle \tau_{max}.ts + 1, \omega_i \rangle, val \rangle$
  $done \leftarrow false$
  while not $done$ **do**
    $cseq[\nu].cfg.\text{put-data}(\langle \tau, v \rangle)$
    $cseq \leftarrow$read-config($cseq$)
    if $|cseq| = \nu$ **then**
      $done \leftarrow true$
    else
      $\nu \leftarrow |cseq|$
  end while
end operation

Read Operation:
at each reader r_i
State Variables:
$cseq[]s.t.cseq[j] \in \mathcal{C} \times \{F, P\}$ with members:
Initialization:
$cseq[0] = \langle c_0, F \rangle$

operation read( )
  $cseq \leftarrow$read-config($cseq$)
  $\mu \leftarrow \max(\{j : cseq[j].status = F\})$
  $\nu \leftarrow |cseq|$
  for $i = \mu : \nu$ **do**
    $\langle \tau, v \rangle \leftarrow \max(cseq[i].cfg.\text{get-data}(), \langle \tau, v \rangle)$
  $done \leftarrow$ **false**
  while not $done$ **do**
    $cseq[\nu].cfg.\text{put-data}(\langle \tau, v \rangle)$
    $cseq \leftarrow$read-config($cseq$)
    if $|cseq| = \nu$ **then**
      $done \leftarrow true$
    else
      $\nu \leftarrow |cseq|$
  end while
  return v
end operation

Algorithm 5 DAP implementation for ARES.

at each process $p_i \in \mathcal{I}$

procedure c.get-tag()
  send (QUERY-TAG) to each $s \in c.Servers$
  until p_i receives $\langle t_s, e_s \rangle$ from $\lceil \frac{n+k}{2} \rceil$ servers in $c.Servers$
  $t_{max} \leftarrow \max(\{t_s : \text{ received } \langle t_s, v_s \rangle \text{ from } s\})$
  return t_{max}
end procedure

procedure c.get-data()
  send (QUERY-LIST) to each $s \in c.Servers$
  until p_i receives $List_s$ from each server $s \in \mathcal{S}_g$ s.t. $|\mathcal{S}_g| = \lceil \frac{n+k}{2} \rceil$ and $\mathcal{S}_g \subset c.Servers$
  $Tags_*^{\geq k}$ = set of tags that appears in k lists
  $Tags_{dec}^{\geq k}$ = set of tags that appears in k lists with values
  $t^*_{max} \leftarrow \max Tags_*^{\geq k}$
  $t^{dec}_{max} \leftarrow \max Tags_{dec}^{\geq k}$
  if $t^{dec}_{max} = t^*_{max}$ **then**
    $v \leftarrow$ decode value for t^{dec}_{max}
  return $\langle t^{dec}_{max}, v \rangle$
end procedure

procedure c.put-data($\langle \tau, v \rangle$))
  $code\text{-}elems = [(\tau, e_1), \ldots, (\tau, e_n)]$, $e_i = \Phi_i(v)$
  send (WRITE, $\langle \tau, e_i \rangle$) to each $s_i \in c.Servers$
  until p_i receives ACK from $\lceil \frac{n+k}{2} \rceil$ servers in $c.Servers$
end procedure

Algorithm 6 The response protocols at any server $s_i \in \mathcal{S}$ in ARES for client requests.

at each server $s_i \in \mathcal{S}$ in configuration c_k

State Variables:
$List \subseteq \mathcal{T} \times \mathcal{C}_s$, initially $\{(t_0, \Phi_i(v_0))\}$

Upon receive (QUERY-TAG) $_{s_i, c_k}$ **from** q
  $\tau_{max} = \max_{(t,c) \in List} t$
  Send τ_{max} to q
end receive

Upon receive (QUERY-LIST) $_{s_i, c_k}$ **from** q
  Send $List$ to q
end receive

Upon receive (PUT-DATA, $\langle \tau, e_i \rangle$) $_{s_i, c_k}$ **from** q
  $List \leftarrow List \cup \{\langle \tau, e_i \rangle\}$
  if $|List| > \delta + 1$ **then**
    $\tau_{min} \leftarrow \min\{t : \langle t, * \rangle \in List\}$
    /* remove the coded value and retain the tag */
    $List \leftarrow List \setminus \{\langle \tau, e \rangle : \tau = \tau_{min} \ \wedge \langle \tau, e \rangle \in List\}$
    $List \leftarrow List \cup \{(\tau_{min}, \perp)\}$
  Send ACK to q
end receive

c.put-data($\langle t_w, v \rangle$): During the execution of the primitive c.put-data($\langle t_w, v \rangle$), a client sends the pair $(t_w, \Phi_i(v))$ to each server $s_i \in c.Servers$. When a server s_i receives a message (PUT-DATA, t_w, c_i) , it adds the pair in its local $List$, trims the pairs with the smallest tags exceeding the length $(\delta + 1)$ of the $List$, and replies with an ack to the client. In particular, s_i replaces the coded-elements of the older tags with $\perp$, and maintains only the coded-elements associated with the $(\delta +$

1) highest tags in the $List$ (see Line Alg. 6:16). The client completes the primitive operation after getting acks from $\lceil \frac{n+k}{2} \rceil$ servers.

c.get-data(): A client, during the execution of a c.get-data() primitive, queries all the servers in $c.Servers$ for their local variable $List$, and awaits responses from $\lceil \frac{n+k}{2} \rceil$ servers. Once the client receives $Lists$ from $\lceil \frac{n+k}{2} \rceil$ servers, it selects the highest tag t, such that: (i) its corresponding value v is decodable from the coded elements in the lists; and (ii) t is the highest tag seen from the responses of at least k $Lists$ (see lines Alg. 5:11-14) and returns the pair (t, v). Note that in the case where anyone of the above conditions is not satisfied the corresponding read operation does not complete.

VI. Correctness, performance and latency of Ares

In this section, we prove the atomicity property of Ares. We also provide an analysis of its storage and communication costs, and the latency of read and write operations. The atomicity property of Ares hinges on the Property 1 of the DAP implementation in each individual configuration used in Ares. Therefore, we start by proving, in subsection VI-A, that Property 1 holds for the DAP implementation in Section V. Based on this, in subsection VI-B, we prove the atomicty of Ares. Next, in sub-section VI-C, we derive the storage and communication costs of read and write operations, and in sub-section VI-D, we derive the latency of reads and writes in terms of the minimum and maximum delays of any point-to-point messages of the underlying network. Due to lack of space proofs are omitted and can be found in the extended version of the paper [31].

A. Safety (Property 1) proof of the DAPs

Correctness. In this section we are concerned with only one configuration c, consisting of a set of servers $c.Servers$. We assume that at most $f \leq \frac{n-k}{2}$ servers from $c.Servers$ may crash. Lemma 2 states that the DAP implementation satisfies the consistency properties Property 1 which will be used to imply the atomicity of the Ares algorithm.

Theorem 2 (Safety)**.** *Let Π a set of complete DAP operations of Algorithm 5 in a configuration $c \in \mathcal{C}$,* c.get-tag*,* c.get-data *and* c.put-data*, of an execution ξ. Then, every pair of operations $\phi, \pi \in \Pi$ satisfy Property 1.*

Liveness. To reason about the liveness of the proposed DAPs, we define a concurrency parameter δ which captures all the put-data operations that overlap with the get-data, until the time the client has all data needed to attempt decoding a value. However, we ignore those put-data operations that might have started in the past, and never completed yet, if their tags are less than that of any put-data that completed before the get-data started. This allows us to ignore put-data operations due to failed clients, while counting concurrency, as long as the failed put-data operations are followed by a successful put-data that completed before the get-data started. In order to define the amount of concurrency in our specific implementation of the DAPs presented in this section the following definition captures the put-data operations that overlap with the get-data, until the client has all data required to decode the value.

Definition 3 (Valid get-data operations)**.** *A* get-data *operation π from a process p is valid if p does not crash until the reception of $\lceil \frac{n+k}{2} \rceil$ responses during the get-data phase.*

Definition 4 (put-data concurrent with a valid get-data)**.** *Consider a valid* get-data *operation π from a process p. Let T_1 denote the point of initiation of π. For π, let T_2 denote the earliest point of time during the execution when p receives all the $\lceil \frac{n+k}{2} \rceil$ responses. Consider the set $\Sigma = \{\phi : \phi$ is any* put-data *operation that completes before π is initiated$\}$, and let $\phi^* = \arg\max_{\phi \in \Sigma} tag(\phi)$. Next, consider the set $\Lambda = \{\lambda : \lambda$ is any* put-data *operation that starts before T_2 such that $tag(\lambda) > tag(\phi^*)\}$. We define the number of* put-data *concurrent with the valid* get-data *π to be the cardinality of the set Λ.*

Termination (and hence liveness) of the DAPs is guaranteed in an execution ξ, provided that a process no more than f servers in $c.Servers$ crash, and no more that δ put-data may be concurrent at any point in ξ. If the failure model is satisfied, then any operation invoked by a non-faulty client will collect the necessary replies independently of the progress of any other client process in the system. Preserving δ on the other hand, ensures that any operation will be able to decode a written value. These are captured in the following theorem:

Theorem 5 (Liveness)**.** *Let ξ be well-formed and fair execution of DAPs, with an $[n, k]$ MDS code, where n is the number of servers out of which no more than $\frac{n-k}{2}$ may crash, and δ be the maximum number of* put-data *operations concurrent with any valid* get-data *operation. Then any* get-data *and* put-data *operation π invoked by a process p terminates in ξ if p does not crash between the invocation and response steps of π.*

B. Atomicity Property of Ares

Ares **Correctness.** The correctness of Ares highly depends on the way the configuration sequence is constructed at each client process. Let $\mathbf{c}_\sigma^p$ denote the configuration sequence $cseq$ at process p in a state σ and $\mu(\mathbf{c}_\sigma^p)$ the index of the last finalized configuration in $\mathbf{c}_\sigma^p$. Then the following properties are preserved by the reconfiguration service:

Theorem 6. *Let π_1 and π_2 two completed* read-config *actions invoked by processes $p_1, p_2 \in \mathcal{I}$ respectively, such that $\pi_1 \rightarrow \pi_2$ in an execution ξ. Let σ_1 be the state after the response step of π_1 and σ_2 the state after the response step of π_2. Then the following properties hold:*

- (a) ***Configuration Consistency**: $\mathbf{c}_{\sigma_2}^{p_2}[i].cfg = \mathbf{c}_{\sigma_1}^{p_1}[i].cfg$, for $1 \leq i \leq |\mathbf{c}_{\sigma_1}^{p_1}|$,*
- (b) ***Sequence Prefix**: $\mathbf{c}_{\sigma_1}^{p_1} \preceq_p \mathbf{c}_{\sigma_2}^{p_2}$, and*
- (c) ***Sequence Progress**: $\mu(\mathbf{c}_{\sigma_1}^{p_1}) \leq \mu(\mathbf{c}_{\sigma_2}^{p_2})$*

Given the properties satisfied by the reconfiguration algorithm of Ares and assuming that the DAP used satisfy Property 1, as presented in Section III, then we have the following result.

Theorem 7 (Atomicity)**.** *In any execution ξ of* Ares*, if in every configuration $c \in \mathcal{G}_L$, c.*get-data()*, c.*put-data()*, and c.*get-tag() *satisfy Property 1, then* Ares *satisfy atomicity.*

Remark 8. *Algorithm* ARES *satisfies atomicity even when the implementaton of the DAPs in the two different configurations c_1 and c_2 are not the same, given that the c_i.get-tag, c_i.get-data, and the c_i.put-data primitives used in each c_i satisfy Property 1.*

C. Storage and Communication Costs for ARES.

We now briefly present the storage and communication costs associated with the presented DAPs. Recall that by our assumption, the storage cost counts the size (in bits) of the coded elements stored in variable $List$ at each server. We ignore the storage cost due to meta-data. For communication cost we measure the bits sent on the wire between the nodes.

Theorem 9. *The* ARES *algorithm has: (i) storage cost $(\delta+1)\frac{n}{k}$, (ii) communication cost for each write at most to $\frac{n}{k}$, and (iii) communication cost for each read at most $(\delta+2)\frac{n}{k}$.*

D. Latency Analysis for read and writes in ARES

Liveness properties cannot be specified for ARES, without restricting asynchrony or the rate of arrival of reconfig operations, or if the consensus protocol never terminates. Here, we provide some conditional performance analysis of the operation, based on latency bounds on the message delivery. We assume that local computations take negligible time and the latency of an operation is due to the delays in the messages exchanged during the execution. We measure delays in *time units* of some global clock, which is visible only to an external viewer. No process has access to the clock. Let d and D be the minimum and maximum durations taken by messages, sent during an execution of ARES, to reach their destinations. Also, let $T(\pi)$ denote the duration of an operation (or action) π. In the statements that follow, we consider any execution ξ of ARES, which contains k reconfig operations. For any configuration c in an execution of ARES, we assume that any $c.Con.$propose operation, takes at least $T_{min}(CN)$ time units.

The following lemma shows the maximum latency of a read or a write operation, invoked by any non-faulty client. From ARES algorithm, the latency of a read/write operation depends on the delays of the DAPs operations. For our analysis we assume that all get-data, get-tag and put-data primitives use two phases of communication. Each phase consists of a communication between the client and the servers.

Lemma 10. *Suppose π, ϕ and ψ are operations of the type* put-data, get-tag *and* get-data, *respectively, invoked by some non-faulty reconfiguration clients, then the latency of these operations are bounded as follows: (i) $2d \leq T(\pi) \leq 2D$; (ii) $2d \leq T(\phi) \leq 2D$; and (iii) $2d \leq T(\psi) \leq 2D$.*

In the following lemma, we estimate the time taken for a read or a write operation to complete, when it discovers k configurations between its invocation and response steps.

Lemma 11. *Consider any execution of* ARES *where at most k reconfiguration operations are invoked. Let σ_s and σ_e be the states before the invocation and after the completion step of a read/write operation π, in some fair execution ξ of* ARES. *Then we have $T(\pi) \leq 6D\,(k+2)$ to complete.*

It remains now to examine the conditions under which a read/write operation may catch up with an infinite number of reconfiguration operations. For the sake of a worst case analysis we will assume that reconfiguration operations suffer the minimum delay d, whereas read and write operations suffer the maximum delay D in each message exchange. Also, we assume that any consensus operation takes the least amount of time to complete $T_{min}(CN)$. The following theorem is the main result of this section, in which we define the relation between $T_{min}(CN)$, d and D so to guarantee that any read or write issued by a non-faulty client always terminates.

Theorem 12. *Suppose $T_{min}(CN) \geq 3(6D-d)$, then any read or write operation π completes in any execution ξ of* ARES *for any number of reconfiguration operations in ξ.*

VII. FLEXIBILITY OF DAPS

In this section, we argue that various implementations of DAPs also be used in ARES. In fact, via reconfig operations, one can implement a highly adaptive strongly consistent DSS: replication-based can be transformed into erasure-code based DSS; increase or decrease the number of storage servers; study the performance of the DSS under various code parameters, etc. The insight to implementing various DAPs comes from the observation that the simple algorithmic template A (see Alg. 7) for reads and writes protocol combined with any implementation of DAPs, satisfying Property 1 gives rise to a MWMR atomic memory service. Moreover, the read and writes operations terminate as long as the implemented DAPs complete.

Algorithm 7 Template A for the client-side read/write steps.

```
     operation read()
2:     ⟨t, v⟩ ← c.get-data()
       c.put-data(⟨t, v⟩)
4:     return ⟨t, v⟩
     end operation

6:   operation write(v)
       t ← c.get-tag()
8:     t_w ← inc(t)
       c.put-data(⟨t_w, v⟩)
10:  end operation
```

A read operation in A performs c.get-data() to retrieve a tag-value pair, $\langle \tau, v \rangle$ from a configuration c, and then it performs a c.put-data($\langle \tau, v \rangle$) to propagate that pair to the configuration c. A write operation is similar to the read but before performing the put-data action it generates a new tag which associates with the value to be written. The following result shows that A is atomic and live, if the DAPs satisfy Property 1 and live.

Theorem 13 (Atomicity of template A)**.** *Suppose the DAP implementation satisfies the consistency properties $C1$ and $C2$ of Property 1 for a configuration $c \in \mathcal{C}$. Then any execution ξ of algorithm A in configuration c is atomic and live if each DAP invocation terminates in ξ under the failure model $c.\mathcal{F}$.*

A number of known tag-based algorithms that implement atomic read/write objects (e.g., ABD [6], FAST[14] – see [31]), can be expressed in terms of DAP.

VIII. CONCLUSIONS

We presented an algorithmic framework suitable for reconfigurable, erasure code-based atomic memory service in asynchronous, message-passing environments. Future work will involve adding efficient repair and reconfiguration using regenerating codes.

REFERENCES

[1] Intel storage acceleration library (open source version). https://goo.gl/zkVl4N.

[2] ABEBE, M., DAUDJEE, K., GLASBERGEN, B., AND TIAN, Y. Ec-store: Bridging the gap between storage and latency in distributed erasure coded systems. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)* (July 2018), pp. 255–266.

[3] AGUILERA, M. K., KEIDAR, I., MALKHI, D., AND SHRAER, A. Dynamic atomic storage without consensus. In *Proceedings of the 28th ACM symposium on Principles of distributed computing (PODC '09)* (New York, NY, USA, 2009), ACM, pp. 17–25.

[4] AGUILERA, M. K., KEIDARY, I., MALKHI, D., MARTIN, J.-P., AND SHRAERY, A. Reconfiguring replicated atomic storage: A tutorial. *Bulletin of the EATCS 102* (2010), 84–081.

[5] ANTA, A. F., NICOLAOU, N., AND POPA, A. Making "fast" atomic operations computationally tractable. In *International Conference on Principles Of Distributed Systems* (2015), OPODIS'15.

[6] ATTIYA, H., BAR-NOY, A., AND DOLEV, D. Sharing memory robustly in message passing systems. *Journal of the ACM 42(1)* (1996), 124–142.

[7] BURIHABWA, D., FELBER, P., MERCIER, H., AND SCHIAVONI, V. A performance evaluation of erasure coding libraries for cloud-based data stores. In *Distributed Applications and Interoperable Systems* (2016), Springer, pp. 160–173.

[8] CACHIN, C., AND TESSARO, S. Optimal resilience for erasure-coded byzantine distributed storage. IEEE Computer Society, pp. 115–124.

[9] CADAMBE, V. R., LYNCH, N., MÉDARD, M., AND MUSIAL, P. A coded shared atomic memory algorithm for message passing architectures. In *Network Computing and Applications (NCA), 2014 IEEE 13th International Symposium on* (Aug 2014), pp. 253–260.

[10] CADAMBE, V. R., LYNCH, N. A., MÉDARD, M., AND MUSIAL, P. M. A coded shared atomic memory algorithm for message passing architectures. *Distributed Computing 30*, 1 (2017), 49–73.

[11] CHEN, Y. L. C., MU, S., AND LI, J. Giza: Erasure coding objects across global data centers. In *Proceedings of the 2017 USENIX Annual Technical Conference (USENIX ATC 17)* (2017), pp. 539–551.

[12] CHOCKLER, G., GILBERT, S., GRAMOLI, V., MUSIAL, P. M., AND SHVARTSMAN, A. A. Reconfigurable distributed storage for dynamic networks. *Journal of Parallel and Distributed Computing 69*, 1 (2009), 100–116.

[13] DUTTA, P., GUERRAOUI, R., AND LEVY, R. R. Optimistic erasure-coded distributed storage. In *DISC '08: Proceedings of the 22nd international symposium on Distributed Computing* (Berlin, Heidelberg, 2008), Springer-Verlag, pp. 182–196.

[14] DUTTA, P., GUERRAOUI, R., LEVY, R. R., AND CHAKRABORTY, A. How fast can a distributed atomic read be? In *Proceedings of the 23rd ACM symposium on Principles of Distributed Computing (PODC)* (2004), pp. 236–245.

[15] FAN, R., AND LYNCH, N. Efficient replication of large data objects. In *Distributed algorithms* (2003), F. E. Fich, Ed., vol. 2848 of *Lecture Notes in Computer Science*, pp. 75–91.

[16] FERNÁNDEZ ANTA, A., HADJISTASI, T., AND NICOLAOU, N. Computationally light "multi-speed" atomic memory. In *International Conference on Principles Of Distributed Systems* (2016), OPODIS'16.

[17] GAFNI, E., AND MALKHI, D. Elastic Configuration Maintenance via a Parsimonious Speculating Snapshot Solution. In *International Symposium on Distributed Computing* (2015), Springer, pp. 140–153.

[18] GEORGIOU, C., NICOLAOU, N. C., AND SHVARTSMAN, A. A. On the robustness of (semi) fast quorum-based implementations of atomic shared memory. In *DISC '08: Proceedings of the 22nd international symposium on Distributed Computing* (Berlin, Heidelberg, 2008), Springer-Verlag, pp. 289–304.

[19] GEORGIOU, C., NICOLAOU, N. C., AND SHVARTSMAN, A. A. Fault-tolerant semifast implementations of atomic read/write registers. *Journal of Parallel and Distributed Computing 69*, 1 (2009), 62–79.

[20] GILBERT, S. RAMBO II: Rapidly reconfigurable atomic memory for dynamic networks. Master's thesis, MIT, August 2003.

[21] HERLIHY, M. P., AND WING, J. M. Linearizability: a correctness condition for concurrent objects. *ACM Transactions on Programming Languages and Systems 12*, 3 (1990), 463–492.

[22] HUFFMAN, W. C., AND PLESS, V. *Fundamentals of error-correcting codes*. Cambridge university press, 2003.

[23] JEHL, L., VITENBERG, R., AND MELING, H. Smartmerge: A new approach to reconfiguration for atomic storage. In *International Symposium on Distributed Computing* (2015), Springer, pp. 154–169.

[24] JOSHI, G., SOLJANIN, E., AND WORNELL, G. Efficient redundancy techniques for latency reduction in cloud systems. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS) 2*, 2 (2017), 12.

[25] KONWAR, K. M., PRAKASH, N., KANTOR, E., LYNCH, N., MÉDARD, M., AND SCHWARZMANN, A. A. Storage-optimized data-atomic algorithms for handling erasures and errors in distributed storage systems. In *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (May 2016), pp. 720–729.

[26] KONWAR, K. M., PRAKASH, N., LYNCH, N., AND MÉDARD, M. Radon: Repairable atomic data object in networks. In *The International Conference on Distributed Systems (OPODIS)* (2016).

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

[28] LYNCH, N. *Distributed Algorithms*. Morgan Kaufmann Publishers, 1996.

[29] LYNCH, N., AND SHVARTSMAN, A. RAMBO: A reconfigurable atomic memory service for dynamic networks. In *Proceedings of 16th International Symposium on Distributed Computing (DISC)* (2002), pp. 173–190.

[30] LYNCH, N. A., AND SHVARTSMAN, A. A. Robust emulation of shared memory using dynamic quorum-acknowledged broadcasts. In *Proceedings of Symposium on Fault-Tolerant Computing* (1997), pp. 272–281.

[31] NICOLAOU, N., CADAMBE, V., KONWAR, K., PRAKASH, N., LYNCH, N., AND MÉDARD, M. Ares: Adaptive, reconfigurable, erasure coded, atomic storage. *CoRR abs/1805.03727* (2018).

[32] ONGARO, D., AND OUSTERHOUT, J. In search of an understandable consensus algorithm. In *Proceedings of the 2014 USENIX Conference on USENIX Annual Technical Conference* (Berkeley, CA, USA, 2014), USENIX ATC'14, USENIX Association, pp. 305–320.

[33] RASHMI, K., CHOWDHURY, M., KOSAIAN, J., STOICA, I., AND RAMCHANDRAN, K. Ec-cache: Load-balanced, low-latency cluster caching with online erasure coding. In *OSDI* (2016), pp. 401–417.

[34] SHRAER, A., MARTIN, J.-P., MALKHI, D., AND KEIDAR, I. Data-centric reconfiguration with network-attached disks. In *Proceedings of the 4th Int'l Workshop on Large Scale Distributed Sys. and Middleware (LADIS 10)* (2010), p. 2226.

[35] SPIEGELMAN, A., KEIDAR, I., AND MALKHI, D. Dynamic Reconfiguration: Abstraction and Optimal Asynchronous Solution. In *31st International Symposium on Distributed Computing (DISC 2017)* (2017), vol. 91, pp. 40:1–40:15.

[36] WANG, S., HUANG, J., QIN, X., CAO, Q., AND XIE, C. Wps: A workload-aware placement scheme for erasure-coded in-memory stores. In *Networking, Architecture, and Storage (NAS), 2017 International Conference on* (2017), IEEE, pp. 1–10.

[37] XIANG, Y., LAN, T., AGGARWAL, V., AND CHEN, Y.-F. R. Multi-tenant latency optimization in erasure-coded storage with differentiated services. In *2015 IEEE 35th International Conference on Distributed Computing Systems (ICDCS)* (2015), IEEE, pp. 790–791.

[38] XIANG, Y., LAN, T., AGGARWAL, V., CHEN, Y.-F. R., XIANG, Y., LAN, T., AGGARWAL, V., AND CHEN, Y.-F. R. Joint latency and cost optimization for erasure-coded data center storage. *IEEE/ACM Transactions on Networking (TON) 24*, 4 (2016), 2443–2457.

[39] YU, Y., HUANG, R., WANG, W., ZHANG, J., AND LETAIEF, K. B. Sp-cache: load-balanced, redundancy-free cluster caching with selective partition. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis* (2018), IEEE Press, p. 1.

[40] ZHANG, H., DONG, M., AND CHEN, H. Efficient and available in-memory kv-store with hybrid erasure coding and replication. In *14th USENIX Conference on File and Storage Technologies (FAST 16)* (Santa Clara, CA, 2016), USENIX Association, pp. 167–180.

[41] ZHOU, P., HUANG, J., QIN, X., AND XIE, C. Pars: A popularity-aware redundancy scheme for in-memory stores. *IEEE Transactions on Computers* (2018), 1–1.