

Declarative Programming over Eventually Consistent Data Stores

KC Sivaramakrishnan *

University of Cambridge, UK
sk826@cl.cam.ac.uk

Gowtham Kaki

Purdue University, USA
gkaki@cs.purdue.edu

Suresh Jagannathan

Purdue University, USA
suresh@cs.purdue.edu

Abstract

User-facing online services utilize geo-distributed data stores to minimize latency and tolerate partial failures, with the intention of providing a fast, always-on experience. However, geo-distribution does not come for free; application developers have to contend with weak consistency behaviors, and the lack of abstractions to compositably construct high-level replicated data types, necessitating the need for complex application logic and invariably exposing inconsistencies to the user. Some commercial distributed data stores and several academic proposals provide a lattice of consistency levels, with stronger consistency guarantees incurring increased latency and throughput costs. However, correctly assigning the right consistency level for an operation requires subtle reasoning and is often an error-prone task.

In this paper, we present QUELEA, a declarative programming model for eventually consistent data stores (ECDS), equipped with a *contract* language, capable of specifying fine-grained application-level consistency properties. A *contract enforcement system* analyses contracts, and *automatically* generates the appropriate consistency protocol for the method protected by the contract. We describe an implementation of QUELEA on top of an off-the-shelf ECDS that provides support for *coordination-free* transactions. Several benchmarks including two large web applications, illustrate the effectiveness of our approach.

Categories and Subject Descriptors D.1.3 [Concurrent Programming]: Distributed Programming; C.2.4 [Distributed Systems]: Distributed databases; D.3.2 [Language Classifications]: Applicative (Functional) Languages; F.3.1 [Logics and Meanings of Programs]: Specifying and Verifying and Reasoning about Programs

General Terms Languages, Performance

Keywords Eventual Consistency, Availability, CRDTs, Axiomatic Contracts, Contract Classification, Distributed Transactions, SMT solvers, Decidable Logic, Quelea, Cassandra, Haskell

* This work was done at Purdue University, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

PLDI'15, June 13–17, 2015, Portland, OR, USA.

Copyright © 2015 ACM 978-1-4503-3468-6/15/06...\$15.00.

<http://dx.doi.org/10.1145/2737924.2737981>

1. Introduction

Many real-world web services — such as those built and maintained by Amazon, Facebook, Google, Twitter, etc. — replicate application state and logic across multiple *replicas* within and across data centers. Replication is intended not only to improve application throughput and reduce user-perceived latency, but also to tolerate partial failures without compromising overall service availability. Traditionally programmers have relied on *strong consistency* guarantees such as linearizability [15] or serializability [21] in order to build correct applications. While strong consistency is an easily stated property, it masks the reality underlying large-scale distributed systems with respect to non-uniform latency, availability, and network partitions [8, 14]. Indeed, modern web services, which aim to provide an "always-on" experience, overwhelmingly favor availability and partition tolerance over strong consistency. To this end, several *weak consistency* models such as eventual consistency, causal consistency, session guarantees, and timeline consistency have been proposed.

Under weak consistency, the developer needs to be aware of concurrent conflicting updates, and has to pay careful attention to avoid unwanted inconsistencies (e.g., negative balances in a bank account, or having an item appear in a shopping cart after it has been removed [13]). Oftentimes, these inconsistencies leak from the application and are witnessed by the user. Ultimately, the developer must decide the consistency level appropriate for a particular operation; this is understandably an error-prone process requiring intricate knowledge of both the application as well as the semantics and implementation of the underlying data store, which typically have only informal descriptions. Nonetheless, picking the correct consistency level is critical not only for correctness but also for scalability of the application. While choosing a weaker consistency level than required may introduce program errors and anomalies, choosing a stronger one than necessary can negatively impact program scalability and performance.

Weak consistency also hinders compositional reasoning about programs. Although an application might be naturally expressed in terms of well-understood and expressive data types such as maps, trees, queues, or graphs, geo-distributed stores typically only provide a minimal set of data types with in-built conflict resolution strategies such as last-writer-wins (LWW) registers, counters, and sets [16, 25]. Furthermore, while traditional database systems enable composability through transactions, geo-distributed stores typically lack unrestricted serializable transactional access to the data. Working in this environment thus requires application state to be suitably coerced to function using only the capabilities of the store.

To address these issues, we describe QUELEA, a declarative programming model and implementation for ECDS. The key novelty of QUELEA is an expressive *contract language* to declare and

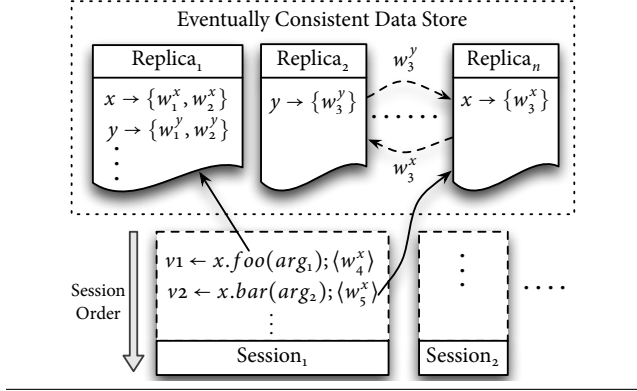


Figure 1: QUELEA system model.

verify fine-grained application-level consistency properties. The programmer uses the contract language to axiomatically specify the set of legal executions allowed over the replicated data type. Contracts are constructed using primitive consistency relations such as *visibility* and *session order* along with standard logical and relational operators. A *contract enforcement system* statically maps operations over the datatype to a particular consistency level available on the store, and provably validates the correctness of the mapping. The paper makes the following contributions:

- We introduce QUELEA, a shallow extension of Haskell that supports the description and validation of replicated data types found in an ECDS. Contracts are used to specify fine-grained application-level consistency properties, and are statically analyzed to assign the most efficient and sound store consistency level to the corresponding operation.
- QUELEA supports coordination-free transactions over arbitrary datatypes. We extend our contract language to express fine-grained transaction isolation guarantees, and utilize the contract enforcement system to automatically assign the correct isolation level for a transaction.
- We provide meta-theory that certifies the soundness of our contract enforcement system, and ensures that an operation is only executed if the required conditions on consistency are met.
- We describe an implementation of QUELEA as a transparent shim layer over Cassandra [16], a well-known general-purpose data store. Experimental evaluation over a set of real-world applications, including a Twitter-like micro-blogging site and an eBay-like auction site illustrates the practicality of our approach.

The rest of the paper is organized as follows. The next section describes the system model. We describe the challenges in programming under eventual consistency, and introduce QUELEA contracts as a proposed solution to overcome these issues in § 3. § 4 provides more details on the contract language, and its mapping to store consistency levels, along with meta-theory for certifying the correctness of the mapping. § 6 introduces transaction contracts and their classification. § 7 describes the implementation of QUELEA on top of Cassandra. § 8 discusses experimental evaluation. § 9 and 10 present related work and conclusions.

2. System Model

In this section, we describe the system model and introduce the primitive relations that our contract language is seeded with. Figure 1 presents a schematic diagram of our system model. The distributed store is composed of a collection of *replicas*, each of which stores a set of *objects* (x, y, \dots). We assume that every object is replicated

at every replica in the store. The state of an object at any replica is the set of all updates (*effects*) performed on the object. For example, the state of x at replica 1 is the set composed of effects w_1^x and w_2^x .

Each object is associated with a set of *operations*. The clients interact with the store by invoking operations on objects. The sequence of operations invoked by a particular client on the store is called a *session*. The data store is typically accessed by a large number of clients (and hence sessions) concurrently. Importantly, the clients are oblivious to which replica an operation is applied to; the data store may choose to route the operation to any replica in order to minimize latency, balance load, etc. For example, the operations *foo* and *bar* invoked by the same session on the same object, might end up being applied to different replicas because replica 1 (to which *foo* was applied) might be unreachable when the client invokes *bar*.

When *foo* is invoked on a object x with arguments arg_1 at replica 1, it simply *reduces* over the current set of effects at that replica on that object (w_1^x and w_2^x), produces a result $v1$ that is sent back to the client, and emits a *single* new effect w_4^x that is appended to the state of x at replica 1. Thus, every operation is evaluated over a *snapshot* of the state of the object on which it is invoked. In this case, the effects w_1^x and w_2^x are *visible* to w_4^x , written logically as $\text{vis}(w_1^x, w_4^x) \wedge \text{vis}(w_2^x, w_4^x)$, where *vis* is the visibility relation between effects. Visibility is an irreflexive and asymmetric relation, and only relates effects produced by operations on the same object. Executing a read-only operation is similar except that no new effects are produced. The effect added to a particular replica is asynchronously sent to other replicas, and eventually merged into all other replicas. Observe that this model does not assume a particular resolution strategy for concurrent conflicting updates, and instead preserves *every* update. Update conflicts are resolved when an operation reduces over the set of effects on an object at a particular replica.

Two effects w_4^x and w_5^x that arise from the same session are said to be in *session order* (written logically as $\text{so}(w_4^x, w_5^x)$). Session order is an irreflexive, transitive relation. The effects w_4^x and w_5^x arising from operations applied to the same object x are said to be under the *same object* relation, written $\text{sameobj}(w_4^x, w_5^x)$. Finally, we can associate every effect with the operation that generated the effect with the help of a relation *oper*. In the current example, $\text{oper}(w_4^x, \text{foo})$ and $\text{oper}(w_5^x, \text{bar})$ hold. For simplicity, we assume all operation names across all object are distinct.

This model admits all the inconsistencies associated with eventual consistency. The goal of this work is to identify the precise consistency level for each operation such that application-level constraints are not violated. In the next section, we will concretely describe the challenges associated with constructing a consistent bank account on top of an ECDS. Subsequently, we will illustrate how our contract and specification language, armed with the primitive relations *vis*, *so*, *sameobj* and *oper*, mitigates these challenges.

3. Motivation

Consider how we might implement a highly available bank account on top of an ECDS, with the *integrity* constraint that the balance must be non-negative. We begin by implementing a bank account replicated data type (RDT) in QUELEA, and then describe the mechanisms to obtain the desired correctness guarantees.

3.1 RDT Specification

A key novelty in QUELEA is that it allows the addition of new RDTs to the store, which obviates the need for coercing application logic to utilize store-provided data types. In addition, QUELEA treats the convergence semantics (i.e., *how* conflicting updates are resolved) of the data type separately from its consistency properties (i.e., *when* updates become visible). This separation of

```

data Acc = Deposit Int | Withdraw Int | GetBal

getBalance :: [Acc] → () → (Int, Maybe Acc)
getBalance hist _ =
  let res = sum [x | Deposit x ← hist] -
            sum [x | Withdraw x ← hist]
  in (res, Nothing)

deposit :: [Acc] → Int → ((), Maybe Acc)
deposit hist amt = ((), Just $ Deposit amt)

withdraw :: [Acc] → Int → (Bool, Maybe Acc)
withdraw hist v =
  if sell $ getBalance hist () ≥ v
  then (True, Just $ Withdraw v)
  else (False, Nothing)

```

Figure 2: Definition of a bank account expressed in Quelea.

concerns permits *operational* reasoning for conflict resolution, and *declarative* reasoning for consistency. The combination of these techniques enhances the programmability of the store.

Let us assume that the bank account object provides three operations: `deposit`, `withdraw` and `getBalance`, with the assumption that the `withdraw` fails if the account has insufficient balance. Every operation in QUELEA is of the following type, written in Haskell syntax:

```

type Operation e a r = [e] → a → (r, Maybe e)

```

An operation takes a list of effects (the *history* of updates to that object), and an input argument, and returns a result along with an optional effect (read-only operations return `Nothing`). The new effect (if emitted) is added to the state of the object at the current replica, and asynchronously sent to other replicas. The implementation of the bank account operations in QUELEA is given in Figure 2.

The datatype `Acc` represents the effect type for the bank account. The function `sum` returns the sum of elements in the list, and `sell` returns the first element of a tuple. For each operation, `hist` is a *snapshot* of the state of the object at some replica. In this sense, every operation on the RDT is atomic, and thus amenable to sequential reasoning. Here, `getBalance` is a read-only operation, `deposit` always emits an effect, and `withdraw` only emits an effect if there is sufficient balance in the account. We have implemented a large corpus of RDTs for realistic benchmarks including shopping carts, auction and micro-blogging sites, etc. in a few tens of lines of code, expressed in this style.

3.1.1 Summarization

Observe that the definition of `getBalance` reduces over the *entire history* of updates to an account. If we are to realize an efficient implementation of this bank account RDT, we need a *summary* of the account history. Intuitively, the current account balance summarizes the state of an account. A bank account with the history `[Deposit 10, Withdraw 5]` is *observably equivalent* to a bank account with a single deposit operation `[Deposit 5]`; we can replace the earlier history with the latter and a client of the store would not be able to tell the difference between the two.

This notion of observable equivalence can be generalized to other RDTs as well. For example, a last-writer-wins register with multiple updates is equivalent to a register with only the last write. Similarly, a set with a collection of add and remove operations is equivalent to a set with a series of additions of live elements from the original set. Since the notion of observable equivalence is specific to each RDT, programmers can provide a summarization function - of type `[e]`

`→ [e]` - as a part of the RDT specification. The summarization function for the bank account is:

```

summarize hist =
  [Deposit $ sell $ getBalance hist ()]

```

Given a bank account history `hist`, the `summarize` function returns a new history with a single deposit of the current account balance. Our implementation invokes the summarization function associated with an RDT to reduce the size of the effect sets maintained by replicas.

3.2 Anomalies under Eventual Consistency

Our goal is to choose the correct consistency level for each of the bank account operations such that (1) the balance remains non-negative and (2) the `getBalance` operation never incorrectly returns a negative balance.

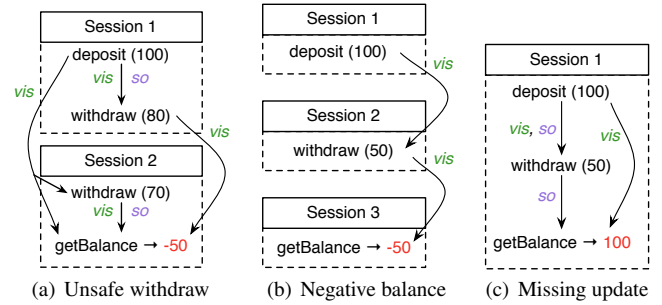


Figure 3: Anomalies possible under eventual consistency for the get balance operation.

Consider the execution shown in Figure 3(a). Assume that all operations in the figure are on the same bank account object with the initial balance being zero. Session 1 performs a `deposit` of 100, followed by a `withdraw` of 80 in the same session. The `withdraw` operation witnesses the deposit and succeeds¹. Subsequently, session 2 performs a `withdraw` operation, but importantly, due to eventual consistency, only witnesses the deposit from session 1, but not the subsequent `withdraw`. Hence, this `withdraw` also *incorrectly* succeeds, violating the integrity constraint. A subsequent `getBalance` operation, that happens to witness all the previous operations, would report a negative balance.

It is easy to see that preventing concurrent `withdraw` operations eliminates this anomaly. This can be done by insisting that `withdraw` be executed as a strongly consistent operation. Despite this strengthening, the `getBalance` operation may still incorrectly report a negative balance to the user. Consider the execution shown in fig. 3(b), which consists of three concurrent sessions performing a `deposit`, a `withdraw`, and a `getBalance` operation, respectively, on the same bank account object. As the `vis` edge indicates, operation `withdraw(50)` in session 2 witnesses the effects of `deposit(100)` from session 1, concludes that there is sufficient balance, and completes successfully. However, the `getBalance` operation may only witness this successful `withdraw`, but not the causally preceding `deposit`, and reports the balance of negative 50 to the user.

Under eventual consistency, the users may also be exposed to other forms of inconsistencies. Figure 3(c) shows an execution where the `getBalance` operation in a session does not witness the effects of an earlier `withdraw` operation performed in the same session, possibly because it was served by a replica that has

¹ Although visibility and session order relations relate effects, we have abused the notation in these examples to relate operations, with the idea that the relations relate the effect emitted by those operations.

not yet merged the `withdraw` effect. This anomaly leads the user to incorrectly conclude that the `withdraw` operation failed to go through.

Although it is easy to understand the reasons behind the occurrence of the aforementioned anomalies, finding the appropriate fixes is not readily apparent. Making `getBalance` a strongly consistent operation is definitely sufficient to avert anomalies, but is it really necessary? Given the cost of enforcing strong consistency [25, 28], it is preferable to avoid imposing such stringent conditions unless there are no viable alternatives. Exploring the space of these alternatives requires understanding the subtle differences in semantics of various kinds of weak consistency alternatives.

3.3 Contracts

QUELEA helps facilitate the mapping of operations to appropriate consistency levels by letting the programmer declare application-level consistency constraints as *contracts*² (Figure 4) that axiomatically specify the set of allowed executions involving this operation. In the case of the bank account, any execution that does not exhibit the anomalies described in the previous section is a *well-formed* execution on the bank account object. By specifying the set of legal executions for each data type in terms of a trace of operation invocations on that type, QUELEA ensures that all executions over that type are well-formed.

In our running example, it is clear that in order to preserve the critical integrity constraint, the `withdraw` operation must be strongly consistent. That is, given two `withdraw` operations a and b , either a is visible to b or vice-versa. We express this application-level consistency requirement as a contract (ψ_w) over `withdraw`:

$$\begin{aligned} &\forall(a : \text{withdraw}). \\ &\text{sameobj}(a, \hat{\eta}) \Rightarrow a = \hat{\eta} \vee \text{vis}(a, \hat{\eta}) \vee \text{vis}(\hat{\eta}, a) \end{aligned}$$

Here, $\hat{\eta}$ stands for the effect emitted by the `withdraw` operation. The syntax $a : \text{withdraw}$ states that a is an effect emitted by a `withdraw` operation i.e., $\text{oper}(a, \text{withdraw})$ holds. The contract specifies that if the current operation emits an effect $\hat{\eta}$, then for any operation a which was emitted by a `withdraw` operation, it is the case that $a = \hat{\eta}$ or a is visible to $\hat{\eta}$, or vice versa. Any execution on a bank account object that preserves the above contract for a `withdraw` operation is said to be derived from a correct implementation of `withdraw`.

To prevent `getBalance` from ever showing a negative balance, it is necessary to prevent the scenario depicted in Figure 3(b). Let $\hat{\eta}$ stand for the effect emitted by the `getBalance` operation. If the effect (b) of a `withdraw` operation is visible to $\hat{\eta}$, and the effect (a) of a `deposit` operation is visible to the effect (b) of the `withdraw` operation, then it must be the case that a is also visible to $\hat{\eta}$. A contract (ψ_{gb}^1) for `getBalance` that precisely captures this application-level consistency requirement can be written thus:

$$\begin{aligned} &\forall(a : \text{deposit}), (b : \text{withdraw}). \\ &(\text{vis}(a, b) \wedge \text{vis}(b, \hat{\eta}) \Rightarrow \text{vis}(a, \hat{\eta})) \end{aligned}$$

To prevent the missing update anomaly described in Figure 3(c), it is necessary for a `getBalance` operation on a bank account to witness the effects of all previous `deposit` and `withdraw` operations performed on the same bank account in the same session. We can express an additional contract (ψ_{gb}^2) for `getBalance` that captures this consistency requirement:

$$\begin{aligned} &\forall(c : \text{deposit} \vee \text{withdraw}). \\ &((\text{so} \cap \text{sameobj})(c, \hat{\eta}) \Rightarrow \text{vis}(c, \hat{\eta})) \end{aligned}$$

Our contract language provides operators to compose relations. The syntax $(R_1 \cap R_2)(a, b)$ is equivalent to $R_1(a, b) \wedge R_2(a, b)$.

² QUELEA exposes the contract construction language as a Haskell library

	$x, y, \hat{\eta} \in \text{EffVar}$	$\text{Op} \in \text{OpName}$
ψ	$\in \text{Contract}$	$::= \forall(x : \tau). \psi \mid \forall x. \psi \mid \pi$
τ	$\in \text{EffType}$	$::= \text{Op} \mid \tau \vee \tau$
π	$\in \text{Prop}$	$::= \text{true} \mid R(x, y) \mid \pi \vee \pi$
		$\mid \pi \wedge \pi \mid \pi \Rightarrow \pi$
R	$\in \text{Relation}$	$::= \text{vis} \mid \text{so} \mid \text{sameobj} \mid =$
		$\mid R \cup R \mid R \cap R \mid R^+$

Figure 4: Contract language.

The above contract (ψ_{gb}^2) says that if a `deposit` or a `withdraw` operation precedes a `getBalance` operation in session order, and is applied on the same object as the `getBalance` operation, then it must be the case that the `getBalance` operation witnesses the effects of the preceding operations.

The final contract (ψ_{gb}) of the `getBalance` operation is merely a conjunction of the previous two versions (ψ_{gb}^1 and ψ_{gb}^2):

$$\begin{aligned} &\forall(a : \text{deposit}), (b : \text{withdraw}), (c : \text{deposit} \vee \text{withdraw}). \\ &(\text{vis}(a, b) \wedge \text{vis}(b, \hat{\eta}) \Rightarrow \text{vis}(a, \hat{\eta})) \\ &\wedge ((\text{so} \cap \text{sameobj})(c, \hat{\eta}) \Rightarrow \text{vis}(c, \hat{\eta})) \end{aligned}$$

Intuitively, this specification prohibits both the `getBalance` anomalies described in Figures. 3(b) and 3(c) from occurring.

Finally, since there are no restrictions on when or how a `deposit` operation can execute, its contract is simply true.

3.4 From Contracts to Implementation

Notice that the contracts for `withdraw` and `getBalance` only express application-level consistency requirements, and make no reference to the semantics of the underlying store. To write contracts, a programmer only needs to reason about the semantics of the application under the QUELEA system model. The mapping of application-level consistency requirements to appropriate store-level guarantees is done automatically behind-the-scenes. How might one go about ensuring that an execution adheres to a contract? The challenge is that a contract provides a declarative (axiomatic) specification of an execution, while what is required is an operational procedure for *enforcing* its implicit constraints.

One strategy would be to execute operations speculatively. Here, operations are tentatively applied as they are received from the client or other replicas. We can maintain a runtime manifestation of executions, and check well-formedness conditions at runtime, rolling back executions if they are ill-formed. However, the overhead of state maintenance and the complexity of user-defined contracts is likely to make this technique infeasible in practice.

We devise a static approach instead. Contracts are analyzed with the help of a theorem prover, and statically mapped to a particular store-level consistency property that the prover guarantees preserves contract semantics. We call this procedure *contract classification*. Given the variety and complexity of store level consistency properties, the idea is that the system implementer parameterizes the classification procedure by describing the store semantics in the *same* contract language as the one used to express the contract on the operations. In the next section, we describe the contract language in detail and describe the classification procedure for a particular store semantics.

4. Contract Language

4.1 Syntax

The syntax of our core contract language is shown in Figure 4. The language is based on first-order logic (FOL), and admits prenex universal quantification over typed and untyped effect variables. We use a special effect variable ($\hat{\eta}$) to denote the effect of *current*

$\eta \in \text{Effect}$	$\psi \in \text{Contract}$	$\bar{\eta} \in \text{Effect Set}$
A	$\in \text{EffSoup}$	$::= \bar{\eta}$
vis, so, sameobj	$\in \text{Relations}$	$::= A \times A$
E	$\in \text{ExecState}$	$::= (A, \text{vis}, \text{so}, \text{sameobj})$

Figure 5: Axiomatic execution.

operation - the operation for which a contract is being written. Notice that $\hat{\eta}$ occurs free in the contract. We fix its scope when classifying contracts (§ 4.4). The type of an effect is simply the name of the operation (eg: *withdraw*) that induced the effect. We admit disjunction in types to let an effect variable range over multiple operation names. The contract $\forall(a : \tau_1 \vee \tau_2). \psi$ is just syntactic sugar for $\forall a. (\text{oper}(a, \tau_1) \vee \text{oper}(a, \tau_2)) \Rightarrow \psi$. An untyped effect variable ranges over all operation names.

Quantifier-free propositions in our contract language are conjunctions, disjunctions and implications of predicates expressing relations between pairs of effect variables. The syntactic class of relations is seeded with primitive *vis*, *so*, and *sameobj* relations, and also admits derived relations that are expressible as union, intersection, or transitive closure³ of primitive relations. Commonly used derived relations are the *same object session order* ($\text{soo} = \text{so} \cap \text{sameobj}$), *happens-before order* ($\text{hb} = (\text{so} \cup \text{vis})^+$) and the *same object happens-before order* ($\text{hbo} = (\text{soo} \cup \text{vis})^+$).

4.2 Semantics

QUELEA contracts are constraints over axiomatic definitions of program executions. Figure 5 summarizes artifacts relevant to define an axiomatic execution. We formalize an axiomatic execution as a tuple $(A, \text{vis}, \text{so}, \text{sameobj})$, where A , called the *effect soup*, is the set of all effects generated during the program execution, and $\text{vis}, \text{so}, \text{sameobj} \subseteq A \times A$ are *visibility*, *session order*, and *same object* relations, respectively, witnessed over generated effects at run-time.

Note that the axiomatic definition of an execution (E) provides interpretations for primitive relations (eg: *vis*) that occur free in contract formulas, and also fixes the domain of quantification to set of all effects (A) observed during the program execution. As such, E is a potential model for any first-order formula (ψ) expressible in our contract language. If E is indeed a valid model for ψ (written as $E \models \psi$), we say that the execution E satisfied the contract ψ :

Definition 1. An axiomatic execution E satisfies a contract ψ if and only if $E \models \psi$.

4.3 Capturing Store Semantics

An important aspect of our contract language is its ability to capture store-level consistency guarantees, along with application-level consistency requirements. Similar to [10], we can rigorously define a wide variety of store semantics including those that combine any subset of session and causality guarantees, and multiple consistency levels. However, for our purposes, we identify three particular consistency levels – eventual, causal, and strong, commonly offered by many distributed stores with tunable consistency, with increasing overhead in terms of latency and availability.

- **Eventual consistency:** Eventually consistent operations can be satisfied as long as the client can reach at least one replica. In the bank account example, *deposit* is an eventually consistent operation. While an ECDS typically offers *basic* eventual consistency with all possible anomalies, we assume that our store

³ Strictly speaking, R^+ is not the transitive closure of R , as transitive closure is not expressible in FOL. Instead, R^+ in our language denotes a superset of transitive closure of R . Formally, R^+ is any relation R' such that for all x, y , and z , a) $R(x, y) \Rightarrow R'(x, y)$, and b) $R'(x, y) \wedge R'(y, z) \Rightarrow R'(x, z)$.

$\frac{\psi \leq \psi_{sc}}{\text{WellFormed}(\psi)}$	$\frac{\psi \leq \psi_{ec}}{\text{EventuallyConsistent}(\psi)}$
$\frac{\psi \not\leq \psi_{ec} \quad \psi \leq \psi_{cc}}{\text{CausallyConsistent}(\psi)}$	$\frac{\psi \not\leq \psi_{cc} \quad \psi \leq \psi_{sc}}{\text{StronglyConsistent}(\psi)}$

Figure 6: Contract classification.

provides stronger semantics that remain highly-available [2, 19]; the store always exposes a *causal cut* of the updates. This semantics can be formally captured in terms of the following contract definition:

$$\psi_{ec} = \forall a, b. \text{hbo}(a, b) \wedge \text{vis}(b, \hat{\eta}) \Rightarrow \text{vis}(a, \hat{\eta})$$

The above contract mandates that an effect a must be visible to the current effect $\hat{\eta}$ if some effect b that causally succeeds a is also visible to $\hat{\eta}$. Thus, if every store replica always maintains and exposes a causal cut of updates to the client, then such a store will satisfy this contract. In such a system, an eventually consistent operation, such as *deposit*, which requires weaker guarantees than offered by the store, can be satisfied as long as some replica is reachable.

- **Causal consistency:** Causally consistent operations are required to see a causally consistent snapshot of the object state, including the actions performed on the same session. The latter requirement implies that if two operations o_1 and o_2 from the same session are applied to two different replicas r_1 and r_2 , the second operation cannot be discharged until the effect of o_1 is included in r_2 . The *getBalance* operation requires causal consistency, as it requires operations from the same session to be visible, which cannot be guaranteed under eventual consistency. The corresponding store semantics is captured by the contract ψ_{cc} defined below:

$$\psi_{cc} = \forall a. \text{hbo}(a, \hat{\eta}) \Rightarrow \text{vis}(a, \hat{\eta})$$

- **Strong consistency:** Strongly consistent operations may block indefinitely under network partitions. An example is the total-order contract on *withdraw* operation. The corresponding store semantics is captured by the ψ_{sc} contract definition:

$$\psi_{sc} = \forall a. \text{sameobj}(a, \hat{\eta}) \Rightarrow \text{vis}(a, \hat{\eta}) \vee \text{vis}(\hat{\eta}, a) \vee a = \hat{\eta}$$

4.4 Contract Classification

Our goal is to map application-level consistency constraints on operations to appropriate store-level consistency guarantees capable of satisfying these constraints. The ability to express both these kinds of constraints as contracts in our contract language lets us compare and determine if contract (ψ_{op}) of an operation (op) is weak enough to be satisfied under a store consistency level identified by the contract ψ_{st} . Towards this end, we define a binary *weaker than* relation for our contract language as following:

Definition 2. A contract ψ_{op} is said to be weaker than ψ_{st} (written $\psi_{op} \leq \psi_{st}$) if and only if $\Delta \vdash \hat{\eta}. \psi_{st} \Rightarrow \psi_{op}$.

The quantifier in the sequent binds $\hat{\eta}$ that occurs free in ψ_{st} and ψ_{op} . The context (Δ) of the sequent is a conjunction of assumptions about the nature of primitive relations. A *well-formed* axiomatic execution (E) is expected to satisfy these assumptions (i.e., $E \models \Delta$).

Definition 3. An axiomatic execution $E = (A, \text{vis}, \text{so}, \text{sameobj})$ is *well-formed* if the following axioms (Δ) hold:

- *The happens-before relation is acyclic:* $\forall a. \neg \text{hbo}(a, a)$.
- *Visibility only relates actions on the same object:*

- $\forall a, b. \text{vis}(a, b) \Rightarrow \text{sameobj}(a, b)$.
- *Session order is a transitive relation:*
 - $\forall a, b, c. \text{so}(a, b) \wedge \text{so}(b, c) \Rightarrow \text{so}(a, c)$.
- *Same object is an equivalence relation:*
 - $\forall a. \text{sameobj}(a, a)$.
 - $\forall a, b. \text{sameobj}(a, b) \Rightarrow \text{sameobj}(b, a)$.
 - $\forall a, b, c. \text{sameobj}(a, b) \wedge \text{sameobj}(b, c) \Rightarrow \text{sameobj}(a, c)$.

If the contract (ψ_{op}) of an operation (op) is *weaker than* a store contract (ψ_{st}), then constraints expressed by the former are implied by guarantees provided by the latter. The completeness of first-order logic allows us to assert that any well-formed execution (E) that satisfies ψ_{st} (i.e., $E \models \psi_{st}$) also satisfies ψ_{op} (i.e., $E \models \psi_{op}$). Consequently, it is safe to execute operation op under a store consistency level captured by ψ_{st} .

Observe that the contracts ψ_{sc} , ψ_{cc} and ψ_{ec} are themselves totally ordered with respect to the \leq relation: $\psi_{ec} \leq \psi_{cc} \leq \psi_{sc}$. This concurs with the intuition that any contract satisfiable under ψ_{ec} or ψ_{cc} is satisfiable under ψ_{sc} , and any contract that is satisfiable under ψ_{ec} is satisfiable under ψ_{cc} . We are interested in the *weakest* guarantee (among ψ_{ec} , ψ_{cc} , and ψ_{sc}) required to satisfy the contract. We define the corresponding consistency level as the *consistency class* of the contract.

The classification scheme, presented formally in Figure 6, defines rules to judge the consistency class of a contract. For example, the scheme classifies the `getBalance` contract (ψ_{gb}) from § 3 as a CausallyConsistent contract, because the sequent $\Delta \vdash \forall \hat{\eta}. \psi_{cc} \Rightarrow \psi_{gb}$ is valid in first-order logic (therefore, $\psi_{gb} \leq \psi_{cc}$), whereas the sequent $\Delta \vdash \forall \hat{\eta}. \psi_{ec} \Rightarrow \psi_{gb}$ is invalid (therefore, $\psi_{gb} \not\leq \psi_{ec}$). Since we confine our contract language to a decidable subset of the logic, validity of such sequents can be decided mechanically allowing us to automate the classification scheme in QUELEA.

Along with three straightforward rules that classify contracts into consistency classes, the classification scheme also presents a rule that judges well-formedness of a contract. A contract is well-formed if and only if it is satisfiable under ψ_{sc} - the strongest possible consistency guarantee that the store can provide. Otherwise, it is considered ill-formed, and rejected statically.

4.5 Generality of Contracts

It is important to note that our contract language provides a generic way to capture application-level consistency properties and is not tied to a particular store semantics. In particular, the same application-level contracts can easily be mapped to a different store with a varied consistency lattice. To illustrate this, let us consider the consistency lattice proposed by Terry et al. [27] based on session guarantees. Terry et al. propose the following four incomparable session guarantees, whose semantics is captured in the contracts below:

Read Your Writes (RYW)	::= $\forall a. \text{soo}(a, \hat{\eta}) \Rightarrow \text{vis}(a, \hat{\eta})$
Monotonic Reads (MR)	::= $\forall a, b. \text{vis}(a, b) \wedge \text{soo}(b, \hat{\eta}) \Rightarrow \text{vis}(a, \hat{\eta})$
Monotonic Writes (MW)	::= $\forall a, b. \text{soo}(a, b) \wedge \text{vis}(b, \hat{\eta}) \Rightarrow \text{vis}(a, \hat{\eta})$
Writes Follow Reads (WFR)	::= $\forall a, b, c. \text{vis}(a, b) \wedge \text{vis}(c, \hat{\eta}) \wedge (\text{soo} \cup =)(b, c) \Rightarrow \text{vis}(a, \hat{\eta})$

In this scheme, the consistency level of an operation is any combination of the above guarantees, which form a partially ordered consistency lattice show in Figure 7. Each element in this lattice corresponds to a store-consistency level, and is represented by its contract. An edge from an upper level element to a lower level element corresponds to a weaker-than relation between the corresponding contracts. Classifying a contract under this scheme is a directed search in the lattice, starting from the bottom, and

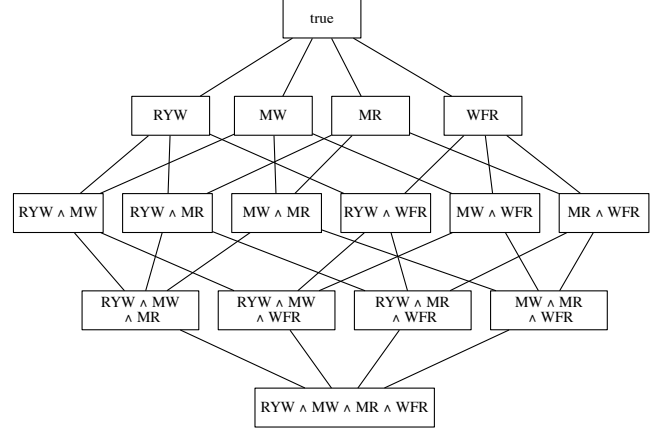


Figure 7: Lattice of consistency levels under session guarantees.

determining the weakest consistency level under which the contract can be satisfied. Under this scheme, `deposit` operations do not need any guarantees, `getBalance` needs RYW and WFR ($\psi_{gb} \leq \text{RYW} \wedge \text{WFR}$), and `withdraw` cannot be satisfied ($\psi_w \not\leq \text{RYW} \wedge \text{MW} \wedge \text{MR} \wedge \text{WFR}$).

4.6 Soundness of Contract Classification

We now present a meta-theoretic result that certifies the soundness of classification-based contract enforcement. To help us state the result, we define an operational semantics of the system described informally in § 2:

op	\in	Operation
τ	\in	ConsistencyClass ::= ec, cc, sc
σ	\in	Session ::= $\cdot \mid \langle op, \tau \rangle; \sigma$
Σ	\in	Session Soup ::= $\sigma \parallel \Sigma \mid \emptyset$
		Config ::= E, Σ

We model the system as a tuple E, Σ , where the axiomatic execution E captures the data store's current state, and session soup Σ is the set of concurrent client sessions interacting with the store. A session σ is a sequence of pairs composed of replicated data type operations op , tagged with the consistency class τ of their contracts (as determined by the contract classification scheme). We assume a reduction relation of form:

$$E, \langle op, \tau \rangle; \sigma \parallel \Sigma \xrightarrow{\eta} E', \sigma \parallel \Sigma$$

on the system state. The relation captures the progress of the execution (from E to E') due to the successful completion of a client operation op from one of the sessions in Σ , generating a new effect η . If the resultant execution E' satisfies the store contract ψ_τ (i.e., $E' \models \psi_\tau$), then we say that the store has *enforced* the contract ψ_τ in the execution E' . With help of the operational semantics, we now state the soundness of contract enforcement as follows:

Theorem 4 (Soundness of Contract Enforcement). *Let ψ be a well-formed contract of a replicated data type operation op , and let τ denote the consistency class of ψ as determined by the contract classification scheme. For all well-formed execution states E, E' such that $E, \langle op, \tau \rangle; \sigma \parallel \Sigma \xrightarrow{\eta} E', \sigma \parallel \Sigma$, if $E' \models \psi_\tau[\eta/\hat{\eta}]$, then $E' \models \psi[\eta/\hat{\eta}]$.*

The theorem states that if a data store correctly enforces ψ_{sc} , ψ_{cc} , and ψ_{ec} contracts in all well-formed executions, then the same store, extended with the classification scheme shown in Figure 6, can enforce all well-formed QUELEA contracts. The proof of the theorem is given below:

Proof. Hypothesis:

$$\begin{array}{l} E, \langle op, \tau \rangle; \sigma \parallel \Sigma \xrightarrow{\eta} E', \sigma \parallel \Sigma \quad H0 \\ E' \models \psi_\tau[\eta/\hat{\eta}] \quad H1 \end{array}$$

Since τ is the contract class of ψ , by inversion, we have $\psi \leq \psi_\tau$. By the definition of \leq relation:

$$\Delta \vdash \forall \hat{\eta}. \psi_\tau \Rightarrow \psi \quad H2$$

Since η denotes new effect, it is a fresh variable that does not occur free in Δ . From $H3$, after instantiating bound $\hat{\eta}$ with η , we have:

$$\Delta \vdash \psi_\tau[\eta/\hat{\eta}] \Rightarrow \psi[\eta/\hat{\eta}] \quad H3$$

Due to the soundness of natural deduction for first-order logic, $H4$ implies that for all models \mathcal{M} such that $\mathcal{M} \models \Delta$, if $\mathcal{M} \models \psi_\tau[\eta/\hat{\eta}]$ then $\mathcal{M} \models \psi[\eta/\hat{\eta}]$. Since E' is well-formed, we have:

$$E' \models \Delta \quad H4$$

Proof follows from $H1$, $H4$, and $H3$. \blacksquare

It is important to note that Theorem 4 does not ascribe any semantics to the reduction relation (\rightarrow). As such, it makes no assumptions about how the store actually implements ec, cc and sc guarantees. The specific implementation strategy is determined by the operational semantics of the store, which *defines* the reduction relation for that particular store. The following section describes operational semantics of the store used by the QUELEA implementation.

5. Operational Semantics

We now describe operational semantics of a data store that implements strong, causal and eventual consistency guarantees. The semantics also serves as a high-level description of our implementation of the store underlying QUELEA.

Figure 8 presents operational semantics as rules defining the reduction relation (\rightarrow) over the execution state. Since we now have a concrete store, we extend our system model with Θ , a representation of the store as a map from replicas to their local states. The local state of a replica r (i.e., $\Theta(r)$) is a set of effects that are currently visible at r . An operation op performed at replica r can only witness the set of effects ($\Theta(r)$) visible at r . To avoid too many parentheses in our formal presentation, we represent operation op , whose contract is in consistency class τ , as op_τ instead of the usual $\langle op, \tau \rangle$. For the sake of clarity, we only consider a single replicated object of well-defined type (for eg: a replicated object of type `BankAccount`) in our formalization. Our semantics are parametric over the specification of this replicated data type. Figure 8 formalizes replicated data type (RDT) specification as tuple (δ, Λ, Ψ) , where δ is the data type, Λ maps labels (op) of operations on δ to their definitions, while Ψ maps them to their consistency contracts (ψ). The definition of an operation is expected to be a lambda expression, although we do not enforce this in our formalization. For technical reasons, we tag each session with a session identifier (s) and the sequence number (i) of the next operation in the session.

The state of an operational execution (E) is a tuple $(A, \text{vis}, \text{so}, \text{sameobj})$, where A is a set of effects, and vis , so , $\text{sameobj} \subseteq A \times A$ are *visibility*, *session order*, and *same object* relations over effects, respectively. We define an effect (η) as a tuple (s, i, op, v) , which records the fact that i^{th} action in session with `SessID` s , which is an operation op on the replicated object, has been successfully executed on some replica yielding a return value v . Note that the combination of s and i uniquely identifies the effect. Session order relation (so) relates effects generated by the same session. An effect $\eta = (s, i, op, v)$ is said to precede another effect $\eta' = (s', i', op', v')$ in session order if and only if $s' = s$ and $i' \geq i$. Since we only consider one replicated object in our formalization, the sameobj relation relates every pair of effects in the effect soup (A). An effect generated at a replica becomes visible at rest of the replicas eventually. If we denote the effect generated by the

operation op as η_{op} , then $\Theta(r) \times \{\eta_{op}\} \subseteq \text{vis}$. Often, in our formalization, we use vis and so binary relations to obtain a set of effects visible to a given effect η , or set of effects that precede a given effect η in the session order. As a syntactic convenience, whenever R is a binary relation, we write $R(\eta)$ to denote the set of all η' such that $(\eta, \eta') \in R$. Conversely, we write $R^{-1}(\eta)$ to denote the set of all η' such that $(\eta', \eta) \in R$.

Basic guarantee provided by the store is causal visibility, which is captured by the rule [EFFVIS] as a condition for an effect to be visible at a replica. The rule makes an effect (η) visible at a replica r only after all the effects that causally precede η are made visible at r . It is important to note that that enforcing causal visibility does not require any inter-replica coordination. Any eventually consistent store can provide causal visibility while being eventually consistent. Therefore, we do not lose any generality by assuming that the store provides causal visibility.

Rule [OPER] is an auxiliary reduction of the form:

$$\Theta \vdash (E, \langle s, i, op \rangle) \xrightarrow{r} (E', \eta)$$

Under the store configuration Θ , the rule captures the progress in execution (from E to E') due to the application of operation op to replica r resulting in a new effect η . The rule first constructs a *context* for the application from the local state ($\Theta(r)$) of the replica, by projecting⁴ relevant information from effects in $\Theta(r)$. It then substitutes the definition ($\Lambda(op)$) of the operation for its label (op), and relies on the reduction relation (\rightarrow) of the server-side language to reduce the application $\Lambda(op)(\text{ctx})$ to a value v' . Subsequently, the attributes of execution state, namely A , vis , so , and sameobj are extended to incorporate the new effect (η).

If the operation op is `EventuallyConsistent`, we simply apply the operation to any replica r . Since the store provides causal visibility, eventually consistent operations are satisfiable under any replica. If the operation is `CausallyConsistent`, the operation can be applied to a replica r only if it already contains the effects of all the previous operations from the same session. This guarantee can be satisfied by applying all operations from the same session to the same logical copy of the database. If such a logical copy is untenable, then the operation might block. Since the store is assumed to converge eventually, the blocked causally consistent operation guaranteed to unblock eventually.

A `StronglyConsistent` operation expects sequential consistency. That is, universe of all effects (A) in an execution (E) must be partitionable into a set of effects that *happened before* η and another set that *happened after* η , where η is the effect generated by an strongly consistent operation. The rule [SC] enforces this sequencing in two steps; firstly, it insists that the the strongly consistent operation (op) witness effects of all operations executed so far by requiring the global set of effects A to be a subset of local state ($\Theta(r)$) of the replica (r) executing op . Secondly, the rule requires the effect (η) generated by op to be added to the local state of every other replica in the store, so that further operations on these replicas can witness the effect of op . Since both these steps require global coordination among replicas, the store is *unavailable* during the time it is executing op .

5.1 Soundness of Operational Semantics

We now prove a meta-theoretic property that establishes the soundness of our operational semantics in enforcing ψ_{ec} , ψ_{cc} , and ψ_{sc} consistency guarantees at every reduction step. As a corollary of this result, and Theorem 4, we have the assurance that QUELEA correctly enforces all well-formed consistency contracts.

First, we prove a useful lemma:

⁴ ctx^* is auxiliary function ctx extended straightforwardly to set of effects

RDT Specification Language

δ	\in	ReplicatedDatatype	Λ	\in	OperationDef.	$::=$	$op \mapsto e$
v	\in	Value	ψ	\in	Contract		
e	\in	Expression	Ψ	\in	OperationContract	$::=$	$op \mapsto \psi$
op	\in	Operation			RDTSpecification	$::=$	(δ, Λ, Ψ)

System Model

$s \in \text{SessID}$	$i \in \text{SeqNo}$	$r \in \text{ReplID}$	Θ	\in	Store	$::=$	$r \mapsto \bar{\eta}$
η	\in	Effect	τ	\in	ConsistencyClass	$::=$	$\text{ec}, \text{cc}, \text{sc}$
A	\in	EffSoup	σ	\in	Session	$::=$	$\cdot \mid op_\tau :: \sigma$
$\text{vis}, \text{so}, \text{sameobj}$	\in	Relations	Σ	\in	Session Soup	$::=$	$\langle s, i, \sigma \rangle \parallel \Sigma \mid \emptyset$
E	\in	ExecState			Config	$::=$	(E, Θ, Σ)

Auxiliary Definitions

$$\text{oper}(s, i, op, v) = op \quad \text{ctxt}(s, i, op, v) = (op, v)$$

Auxiliary Reductions

$$\Theta \vdash (E, \langle s, i, op \rangle) \xrightarrow{r} (E', \eta)$$

[OPER]

$$\frac{r \in \text{dom}(\Theta) \quad \text{ctxt} = \text{ctxt}^*(\Theta(r)) \quad \Lambda(op)(\text{ctxt}) \rightsquigarrow^* v \quad \eta = (s, i, op, v) \quad \{\eta'\} = A_{(\text{SessID}=s, \text{SeqNo}=i-1)} \quad A' = \{\eta\} \cup A \quad \text{vis}' = \Theta(r) \times \{\eta\} \cup \text{vis} \quad \text{so}' = (\text{so}^{-1}(\eta') \cup \eta') \times \{\eta\} \cup \text{so} \quad \text{sameobj}' = A' \times A'}{\Theta \vdash ((A, \text{vis}, \text{so}, \text{sameobj}), \langle s, i, op \rangle) \xrightarrow{r} ((A', \text{vis}', \text{so}', \text{sameobj}'), \eta)}$$

Operational Semantics

$$(E, \Theta, \Sigma) \xrightarrow{\eta} (E', \Theta', \Sigma')$$

[EFFVIS]

$$\frac{\eta \in A \quad \Theta' = \Theta \cup [r \mapsto \{\eta\} \cup \Theta(r)] \quad \eta \notin \Theta(r) \quad E.\text{vis}^{-1}(\eta) \cup E.\text{so}^{-1}(\eta) \subseteq \Theta(r)}{(E, \Theta, \Sigma) \xrightarrow{\eta} (E', \Theta', \Sigma')}$$

[EC]

$$\frac{\tau = \text{EventuallyConsistent} \quad \Theta \vdash (E, \langle s, i, op \rangle) \xrightarrow{r} (E', \eta)}{(E, \Theta, \langle s, i, op_\tau :: \sigma \rangle \parallel \Sigma) \xrightarrow{\eta} (E', \Theta, \langle s, i+1, \sigma \rangle \parallel \Sigma)}$$

[CC]

$$\frac{\tau = \text{CausallyConsistent} \quad \Theta \vdash (E, \langle s, i, op \rangle) \xrightarrow{r} (E', \eta) \quad E'.\text{so}^{-1}(\eta) \subseteq \Theta(r)}{(E, \Theta, \langle s, i, op_\tau :: \sigma \rangle \parallel \Sigma) \xrightarrow{\eta} (E', \Theta, \langle s, i+1, \sigma \rangle \parallel \Sigma)}$$

[SC]

$$\frac{\tau = \text{StronglyConsistent} \quad \Theta \vdash (E, \langle s, i, op \rangle) \xrightarrow{r} (E', \eta) \quad E.A \subseteq \Theta(r) \quad \text{dom}(\Theta') = \text{dom}(\Theta) \quad \forall r' \in \text{dom}(\Theta'). \Theta'(r') = \Theta(r') \cup \{\eta\}}{(E, \Theta, \langle s, i, op_\tau :: \sigma \rangle \parallel \Sigma) \xrightarrow{\eta} (E, \Theta', \langle s, i+1, \sigma \rangle \parallel \Sigma)}$$

Figure 8: Operational semantics of a replicated data store.

Lemma 5 (Auxiliary Reduction Preserves Well-Formedness). *For every well-formed execution state E , if $\Theta \vdash (E, \langle s, i, op \rangle) \xrightarrow{r} (E', \eta)$, then E' is well-formed.*

Proof. Let us denote $E.A$, $E.\text{vis}$, $E.\text{so}$, and $E.\text{sameobj}$ as A , vis , so , and sameobj respectively. Likewise, let us denote $E'.A$, $E'.\text{vis}$, $E'.\text{so}$, and $E'.\text{sameobj}$ as A' , vis' , so' , and $\text{sameobj}'$, respectively. By inversion on $\Theta \vdash (E, \langle s, i, op \rangle) \xrightarrow{r} (E', \eta)$, we have following hypotheses:

$$\begin{array}{ll} r \in \text{dom}(\Theta) & H0 \\ \text{ctxt} = \text{ctxt}^*(\Theta(r)) & H1 \\ \eta = (s, i, op, v) & H2 \\ \{\eta'\} = A_{(\text{SessID}=s, \text{SeqNo}=i-1)} & H3 \\ A' = \{\eta\} \cup A & H4 \\ \text{vis}' = \Theta(r) \times \{\eta\} \cup \text{vis} & H5 \\ \text{so}' = (\text{so}^{-1}(\eta') \cup \eta') \times \{\eta\} \cup \text{so} & H6 \\ \text{sameobj}' = A' \times A' & H7 \end{array}$$

Since E is well-formed, from the definition of well-formedness and *models* relation, we have the following:

$$\begin{array}{ll} \forall (a \in A). \neg \text{hb}(a, a) & H8 \\ \forall (a, b \in A). \text{vis}(a, b) \Rightarrow \text{sameobj}(a, b) & H9 \\ \forall (a, b, c \in A). \text{so}(a, b) \wedge \text{so}(b, c) \Rightarrow \text{so}(a, c) & H10 \\ \forall (a \in A). \text{sameobj}(a, a) & H11 \\ \forall (a, b \in A). \text{sameobj}(a, b) \Rightarrow \text{sameobj}(b, a) & H12 \\ \forall (a, b, c \in A). \text{so}(a, b) \wedge \text{so}(b, c) \Rightarrow \text{so}(a, c) & H13 \end{array}$$

Since $\text{hb} = (\text{vis} \cup \text{so})^+$, $H8$ is equivalent to conjunction of following assertions:

$$\begin{array}{ll} \forall (a \in A). \neg \text{vis}(a, a) & H14 \\ \forall (a \in A). \neg \text{so}(a, a) & H15 \\ \forall (a, b \in A). \neg (\text{vis}(a, b) \wedge \text{so}(b, a)) & H16 \end{array}$$

Since η is fresh, $\eta \notin \Theta(r)$. From $H4$, $H5$, and $H14$, we have the acyclicity property for vis' :

$$\forall (a \in A'). \neg \text{vis}'(a, a) \quad H17$$

Also, $\eta \notin A$, and from $H4$, $H6$ and $H15$, we have acyclicity for so' :

$$\forall(a \in A'). \neg \text{so}'(a, a) \quad H18$$

Similarly, from the uniqueness of η , and $H5$, $H6$, and $H16$, we have the following:

$$\forall(a, b \in A'). \neg(\text{vis}'(a, b) \wedge \text{so}'(b, a)) \quad H19$$

From $H17 - 19$, we prove the acyclicity of hb' :

$$\forall(a \in A'). \neg \text{hb}'(a, a) \quad G0$$

The $\text{sameobj}'$ relation is simply the cross product $A' \times A'$. Hence following trivially hold:

$$\begin{aligned} \forall(a, b \in A'). \text{vis}'(a, b) &\Rightarrow \text{sameobj}'(a, b) & G1 \\ \forall(a \in A'). \text{sameobj}'(a, a) & & G2 \\ \forall(a, b \in A'). \text{sameobj}'(a, b) &\Rightarrow \text{sameobj}'(b, a) & G3 \end{aligned}$$

Finally, from $H3$, $H4$, $H6$ and $H10$, we have transitivity for so' :

$$\forall(a, b, c \in A'). \text{so}'(a, b) \wedge \text{so}'(b, c) \Rightarrow \text{so}'(a, c) \quad G4$$

Well-formedness of E' follows from $G0 - 4$. ■

We now define causal consistency property of the store formally:

Definition 6. A store Θ is said to be causally consistent under an execution $E = (A, \text{vis}, \text{so}, \text{sameobj})$ if and only if:

$$\begin{aligned} \forall(r \in \text{dom}(\Theta)). \forall(\eta \in \Theta(r)). \\ \forall(a \in A). \text{hbo}(a, \eta) \Rightarrow a \in \Theta(r) \quad H20 \end{aligned}$$

Where, $\text{hbo} = (\text{vis} \cup \text{soo})^+$

The following theorem proves that our operational semantics correctly enforce ψ_{ec} , ψ_{cc} , and ψ_{sc} guarantees:

Theorem 7 (Soundness Modulo Classification). For every well-formed execution state E , for every store Θ that is causally consistent under E , and for every contract class $\tau \in \{\text{ec}, \text{cc}, \text{sc}\}$, if:

$$(E, \Theta, \langle s, i, \text{op}_\tau :: \sigma \rangle \parallel \Sigma) \xrightarrow{\eta} (E', \Theta', \langle s, i+1, \sigma \rangle \parallel \Sigma)$$

then (i) E' is well-formed, and (ii) $E' \models \psi_\tau[\eta/\hat{\eta}]$

Proof. By case analysis on the derivation:

$$(E, \Theta, \langle s, i, \text{op}_\tau :: \sigma \rangle \parallel \Sigma) \xrightarrow{\eta} (E', \Theta', \langle s, i+1, \sigma \rangle \parallel \Sigma)$$

Cases:

• Case [EC]: Hypotheses:

$$\begin{aligned} \tau &= \text{EventuallyConsistent} & H0 \\ \Theta \vdash (E, \langle s, i, \text{op} \rangle) &\xrightarrow{\tau} (E', \eta) & H1 \end{aligned}$$

Goal (i) follows from $H1$ and lemma 5. Goal (ii) is the following:

$$E' \models \forall a, b. \text{hbo}(a, b) \wedge \text{vis}(b, \eta) \Rightarrow \text{vis}(a, \eta) \quad G0$$

Let $A' = E'.A$, $\text{vis}' = E'.\text{vis}$, $\text{so}' = E'.\text{so}$, and $\text{hbo}' = (\text{so}' \cup \text{vis}')^+$. By inversion on $H1$, we get the following hypotheses:

$$\begin{aligned} A' &= A \cup \{\eta\} & H2 \\ \text{vis}' &= \Theta(r) \times \eta \cup \text{vis} & H3 \\ r &\in \text{dom}(\Theta) & H4 \\ \text{sameobj}' &= A' \times A' & H5 \end{aligned}$$

Since E' defines A' as the universe of values. Therefore, the goal can be rewritten:

$$\forall(a, b \in A'). E' \models \text{hbo}(a, b) \wedge \text{vis}(b, \eta) \Rightarrow \text{vis}(a, \eta) \quad G1$$

New hypotheses after *intros*:

$$\begin{aligned} a &\in A' & H6 \\ b &\in A' & H7 \end{aligned}$$

And new goal:

$$E' \models \text{hbo}(a, b) \wedge \text{vis}(b, \eta) \Rightarrow \text{vis}(a, \eta) \quad G2$$

Since $(\mathcal{M} \models A \Rightarrow B) \Leftrightarrow (\mathcal{M} \models A \Rightarrow \mathcal{M} \models B)$, we prove $G0$ by proving:

$$(E' \models \text{hbo}(a, b) \wedge \text{vis}(b, \eta)) \Rightarrow (E' \models \text{vis}(a, \eta)) \quad G3$$

After *intros*:

$$E' \models \text{hbo}(a, b) \wedge \text{vis}(b, \eta) \quad H8$$

Since E' defines hbo' and vis' as interpretations for hbo and vis respectively, we have:

$$\text{hbo}'(a, b) \wedge \text{vis}'(b, \eta) \quad H9$$

And the goal is:

$$\text{vis}'(a, \eta) \quad G4$$

Inversion on $H9$:

$$\begin{aligned} \text{hbo}'(a, b) & \quad H10 \\ \text{vis}'(b, \eta) & \quad H11 \end{aligned}$$

Since η is unique, from $H3$ and $H11$ we have the following:

$$b \in \Theta(r) \quad H12$$

Since $a, b \neq \eta$, we have that $\text{hbo}'(a, b) \Rightarrow \text{hbo}(a, b)$. Since Θ is causally consistent under E , using $H10$ and $H12$ we derive the following:

$$a \in \Theta(r) \quad H13$$

Now, from $H3$ and $H13$, we deduce:

$$(a, \eta) \in \text{vis}'$$

which is what needs to be proven ($G4$).

• Case [CC]: Hypotheses:

$$\begin{aligned} \tau &= \text{CausallyConsistent} & H14 \\ \Theta \vdash (E, \langle s, i, \text{op} \rangle) &\xrightarrow{\tau} (E', \eta) & H15 \\ E'.\text{so}(\eta) &\subseteq \Theta(r) & H16 \end{aligned}$$

Goal (i) follows from $H15$ and lemma 5. We now prove Goal (ii). Expanding the definition of ψ_{cc} , goal is the following:

$$E' \models \forall a. \text{hbo}(a, \eta) \Rightarrow \text{vis}(a, \eta) \quad G5$$

Let $A' = E'.A$, $\text{vis}' = E'.\text{vis}$, $\text{so}' = E'.\text{so}$, and $\text{hbo}' = (\text{so}' \cup \text{vis}')^+$. By inversion on $H15$, we get the following:

$$\begin{aligned} A' &= A \cup \{\eta\} & H17 \\ \text{vis}' &= \Theta(r) \times \eta \cup \text{vis} & H18 \\ \{\eta'\} &= A_{(\text{SessID}=s, \text{SeqNo}=i-1)} & \\ \text{so}' &= (\text{so}^{-1}(\eta') \cup \eta') \times \{\eta\} \cup \text{so} & H19 \\ \text{sameobj}' &= A' \times A' & H20 \end{aligned}$$

Expanding \models followed by *intros* on $G5$ yields a context with following hypotheses:

$$\begin{aligned} a &\in A' & H21 \\ \text{hbo}'(a, \eta) & & H22 \end{aligned}$$

And the goal is the following:

$$\text{vis}'(a, \eta) \quad G6$$

Since *happens-before* is transitive, by inversion on $H22$, we get two cases⁵:

• SCase 1: Hypotheses:

$$(\text{so}' \cup \text{vis}')(a, \eta) \quad H23$$

Note that we Inversion on $H23$ leads to two subcases. In one case, we assume $\text{vis}'(a, \eta)$ and try to prove the goal $G6$. The proof for this case mimics the proof for Case [EC]. Alternatively, in second case, we assume:

$$\text{so}'(a, \eta) \quad H24$$

and prove $G6$. From $H24$ and $H16$, we infer:

$$a \in \Theta(r) \quad H25$$

Now, from $H25$ and $H18$ we know:

$$(a, \eta) \in \text{vis}'$$

which is the goal ($G6$). 26

⁵ Recall that we only consider single replicated object in our formalization. Accordingly, for any execution $E = (A, \text{vis}, \text{so}, \text{sameobj})$, we have $\text{sameobj} = A \times A$. Since, $\text{soo} = \text{so} \cap \text{sameobj}$ and $\text{so} \subseteq A \times A$, we use so and soo interchangeably in proofs.

- SCase 2: Hypotheses (after abbreviating the occurrence of $(\text{so}' \cup \text{vis}')^+$ as hbo'):

$$\exists(c \in A'). \text{hbo}'(a, c) \wedge (\text{so}' \cup \text{vis}')(c, \eta) \quad H27$$

Inverting $H27$, followed by expanding $\text{so}' \cup \text{vis}'$:

$$\begin{aligned} c \in A' & \quad H28 \\ \text{hbo}'(a, c) \wedge (\text{so}'(c, \eta) \vee \text{vis}'(c, \eta)) & \quad H29 \end{aligned}$$

Inverting the disjunction in $H29$, we get two cases:

- SSCase R: Hypothesis is

$$\text{hbo}'(a, c) \wedge \text{vis}'(c, \eta) \quad H30$$

Observe that hypothesis $H30$ and current goal ($G6$) are same as hypothesis $H9$ and goal ($G4$) in Case [EC]. The proof for this SSCase is also the same.

- SSCase L: Hypothesis is

$$\text{hbo}'(a, c) \wedge \text{so}'(c, \eta) \quad H31$$

Inverting $H31$:

$$\begin{aligned} \text{hbo}'(a, c) & \quad H32 \\ \text{so}'(c, \eta) & \quad H33 \end{aligned}$$

From $H33$ and $H16$, we infer:

$$c \in \Theta(r) \quad H34$$

Since $a, c \neq \eta$, we know that $\text{hbo}'(a, c) \Rightarrow \text{hbo}(a, c)$. Since Θ is causally consistent under E , using $H32$ and $H34$ we derive the following:

$$a \in \Theta(r) \quad H35$$

Proof follows from $H35$ and $H18$.

- Case [SC]: Hypotheses:

$$\begin{aligned} \tau &= \text{StronglyConsistent} & H36 \\ \Theta \vdash (E, \langle s, i, \text{op} \rangle) &\xrightarrow{\tau} (E', \eta) & H37 \\ E.A \subseteq \Theta(r) & & H38 \\ \text{dom}(\Theta') = \text{dom}(\Theta) & & H39 \\ \forall r' \in \text{dom}(\Theta'). \Theta'(r') = \Theta(r') \cup \{\eta\} & & H40 \end{aligned}$$

Let $A' = E'.A$, $\text{vis}' = E'.\text{vis}$, $\text{so}' = E'.\text{so}$, and $\text{hbo}' = (\text{so}' \cup \text{vis}')^+$. Inversion on $H37$ gives:

$$\begin{aligned} A' &= A \cup \{\eta\} & H41 \\ \text{vis}' &= \Theta(r) \times \eta \cup \text{vis} & H42 \\ \{\eta'\} &= A_{(\text{SessID}=s, \text{SeqNo}=i-1)} & \\ \text{so}' &= (\text{so}^{-1}(\eta') \cup \eta') \times \{\eta\} \cup \text{so} & H43 \\ \text{sameobj}' &= A' \times A' & H44 \end{aligned}$$

Goal (i) follows from $H37$ and Lemma 5. Expanding the definition of ψ_{sc} , followed by expanding \models relation, and then doing *intros*, we get the following context:

$$a \in A' \quad H45$$

And the goals are following:

$$\begin{aligned} \text{sameobj}'(a, \eta) &\Rightarrow \text{hbo}'(a, \eta) \vee \text{hbo}'(\eta, a) \vee a = \eta & G7 \\ \text{hbo}'(a, \eta) &\Rightarrow \text{vis}'(a, \eta) & G8 \\ \text{hbo}'(\eta, a) &\Rightarrow \text{vis}'(\eta, a) & G9 \end{aligned}$$

From $H41$ and $H45$, we know that either $a = \eta$ or $a \in A$. When $a = \eta$:

- $G7$ follows trivially
- From lemma 5, we know that hbo' is acyclic. Hence $\text{hbo}'(\eta, \eta) = \text{false}$. Therefore, $G8 - 9$ are valid vacuously.

When $a \in A$:

- Intros on $G7$ gives following hypothesis:

$$\text{sameobj}'(a, \eta)$$

From $H38$ we know that $a \in \Theta(r)$. Using $H42$, we derive:

$$\text{vis}'(a, \eta) \quad H46$$

Introducing disjunction:

$$(\text{vis}' \cup \text{so}')(a, \eta) \quad H47$$

Now, since $\text{hbo}' = (\text{vis}' \cup \text{so}')^+$, proof follows from last hypothesis.

- Intros on $G8$ gives following hypothesis:

$$\text{hbo}'(a, \eta)$$

From $H38$ we know that $a \in \Theta(r)$. Using $H42$, we derive:

$$\text{vis}'(a, \eta) \quad H48$$

Which proves $G8$.

- Intros on $G9$ gives:

$$\text{hbo}'(\eta, a) \quad H49$$

From lemma 5, we know that E' is well-formed. Hence:

$$\neg \text{hbo}'(a, a) \quad H50$$

Since $\text{sameobj}' = A' \times A'$, we have:

$$\text{sameobj}'(a, \eta) \quad H51$$

Using previous hypothesis, we can reuse the proof for $G7$ to derive:

$$\text{hbo}'(a, \eta) \quad H52$$

Since hbo' is a transitive relation, from $H49$ and $H52$, we derive:

$$\text{hbo}'(a, a) \quad H53$$

$H53$ and $H50$ are contradicting hypothesis. Proof follows. \blacksquare

We now show that every configuration of the store that is reachable via the reduction relation (\rightarrow) is causally consistent.

Theorem 8 (Causal Consistency Preservation). *For every well-formed execution state E , and a store Θ that is causally consistent under E , if:*

$$(E, \Theta, \langle s, i, \text{op}_\tau :: \sigma \rangle \parallel \Sigma) \xrightarrow{\eta} (E', \Theta', \langle s, i+1, \sigma \rangle \parallel \Sigma)$$

then Θ' is causally consistent under E' .

Proof. By case analysis on on:

$$(E, \Theta, \langle s, i, \text{op}_\tau :: \sigma \rangle \parallel \Sigma) \xrightarrow{\eta} (E', \Theta', \langle s, i+1, \sigma \rangle \parallel \Sigma)$$

Cases:

- Case [EFFVIS]: Final execution state is same as the initial (i.e., $E' = E$). Therefore, we need to prove that new store configuration Θ is causally consistent under $E = (A, \text{vis}, \text{so}, \text{sameobj})$. Hypotheses:

$$\begin{aligned} r &\in \text{ReplID} & H0 \\ \eta &\in A & H1 \\ \eta &\notin \Theta(r) & H2 \\ E.\text{vis}^{-1}(\eta) \cup E.\text{so}^{-1}(\eta) &\subseteq \Theta(r) & H3 \\ \Theta' &= \Theta \cup [r \mapsto \{\eta\} \cup \bar{\Theta}(r)] & H4 \\ \forall (r \in \text{dom}(\Theta)). \forall (\eta \in \Theta(r)). & & \\ \forall (a \in A). \text{hbo}(a, \eta) &\Rightarrow a \in \Theta(r) & H5 \end{aligned}$$

From $H4$ and $H5$, it suffices to prove:

$$\forall (a \in A). \text{hbo}(a, \eta) \Rightarrow a \in \Theta'(r) \quad G0$$

After *intros*, hypotheses:

$$\begin{aligned} a &\in A & H6 \\ \text{hbo}(a, \eta) & & H7 \end{aligned}$$

Goal:

$$a \in \Theta'(r) \quad G1$$

Inversion on $H7$ leads to two cases:

- SCase a directly precedes η : Hypothesis:

$$(\text{vis} \cup \text{so})(a, \eta) \quad H8$$

From $H3$ and $H8$, we conclude that $a \in \Theta'(r)$.

- SCase a transitively precedes η : Hypothesis:

$$\exists(c \in A). \text{hbo}(a, c) \wedge (\text{vis} \cup \text{so})(c, \eta) \quad H9$$

Inverting $H9$:

$$\begin{array}{ll} c \in A & H10 \\ \text{hbo}(a, c) & H11 \\ (\text{vis} \cup \text{so})(c, \eta) & H12 \end{array}$$

From $H3$ and $H12$, we have:

$$c \in \Theta'(r) \quad H13$$

From $H5$, $H13$ and $H11$, we conclude that $a \in \Theta'(r)$.

- Cases [EC] and [CC]: Store configuration (Θ) remains unchanged. Further, no new *happens before* order is added either among existing effects, or *from* the newly generated effect *to* existing effects. Consequently, proof is trivial.
- Case [SC]: Let $E = (A, \text{vis}, \text{so}, \text{sameobj})$ and $E' = (A', \text{vis}', \text{so}', \text{sameobj}')$. Hypotheses:

$$\begin{array}{ll} r \in \text{RepIID} & H14 \\ \tau = \text{StronglyConsistent}(\psi) & H15 \\ \Theta \vdash (E, \langle s, i, op \rangle) \xrightarrow{\tau} (E', \eta) & H16 \\ A \subseteq \Theta(r) & H17 \\ \text{dom}(\Theta') = \text{dom}(\Theta) & H18 \\ \forall r' \in \text{dom}(\Theta'). \Theta'(r') = \Theta(r') \cup \{\eta\} & H19 \\ \forall(r' \in \text{dom}(\Theta)). \forall(\eta' \in \Theta(r')). & \\ \quad \forall(a \in A). \text{hbo}(a, \eta') \Rightarrow a \in \Theta(r') & H20 \end{array}$$

The goal:

$$\begin{array}{l} \forall(r' \in \text{dom}(\Theta')). \forall(\eta' \in \Theta(r')). \\ \quad \forall(a \in A'). \text{hbo}'(a, \eta') \Rightarrow a \in \Theta'(r') \end{array} \quad G2$$

Inverting $H16$:

$$\begin{array}{ll} A' = A \cup \{\eta\} & H21 \\ \text{vis}' = \Theta(r) \times \eta \cup \text{vis} & H22 \\ \{\eta'\} = A_{(\text{SessID}=s, \text{SeqNo}=i-1)} & \\ \text{so}' = (\text{so}^{-1}(\eta') \cup \eta') \times \{\eta\} \cup \text{so} & H23 \\ \text{sameobj}' = A' \times A' & H24 \end{array}$$

Using $H19$ and $H20$, we can reduce the goal ($G2$) to:

$$\forall(r' \in \text{dom}(\Theta')). \forall(a \in A'). \text{hbo}'(a, \eta) \Rightarrow a \in \Theta'(r') \quad G3$$

After *intros*, hypotheses:

$$\begin{array}{ll} r' \in \text{dom}(\Theta') & H25 \\ a \in A' & H26 \\ \text{hbo}'(a, \eta) & H27 \end{array}$$

Goal:

$$a \in \Theta'(r') \quad G4$$

From $H26$ and $H21$, we know that either $a \in A$ or $a = \eta$.

- If $a \in A$, then from $H17$, we know that $a \in \Theta(r')$. However, from $H19$ we know that $\Theta(r') \subset \Theta'(r')$, which lets us conclude that $a \in \Theta'(r')$.
- If $a = \eta$, then $H27$ is $\text{hbo}'(\eta, \eta)$. However, from lemma 5 we know that E' is well-formed, which means that hbo' is acyclic. Hence, a contradiction. *ex_falso_quodlibet*. ■

Corollary 9 (Soundness). *For every well-formed execution state E , for every store Θ that is causally consistent under E , for every contract class $\tau \in \{\text{ec}, \text{cc}, \text{sc}\}$, and for every consistency contract ψ in the contract class τ , If:*

$$(E, \Theta, \langle s, i, op_\tau :: \sigma \rangle \parallel \Sigma) \xrightarrow{\eta} (E', \Theta', \langle s, i+1, \sigma \rangle \parallel \Sigma)$$

then (i) E' is well-formed, (ii) Θ' is causally consistent under E' , and (iii) $E' \models \psi[\eta/\hat{\eta}]$

Proof. Follows from Theorems 4, 8, and 7. ■

6. Transaction Contracts

While contracts on individual operations offer the programmer object-level declarative reasoning, real-world scenarios often involve operations that span multiple objects. In order to address this problem, several recent systems [2, 9, 26] have proposed eventually consistent transactions in order to compose operations on multiple objects. However, given that classical transaction models such as serializability and snapshot isolation require inter-replica coordination, these systems espouse *coordination-free transactions* that remain available under network partitions, but only provide weaker isolation guarantees. Coordination-free transactions have intricate consistency semantics and widely varying runtime overheads. As with operation-level consistency, the onus is on the programmer to pick the correct transaction kind. This choice is further complicated by the consistency semantics of individual operations.

6.1 Syntax and Semantics Extensions

QUELEA automates the choice of assigning the correct and most efficient transaction isolation level. Similar to contracts on individual operations, the programmer associates contracts with transactions, declaratively expressing consistency specifications. We extend the contract language with a new term under quantifier-free propositions - $\text{txn } S_1 S_2$, where S_1 and S_2 are sets of effects, and introduce a new primitive equivalence relation sametxn that holds for effects from the same transaction. $\text{txn}\{a, b\}\{c, d\}$ is just syntactic sugar for $\text{sametxn}(a, b) \wedge \text{sametxn}(c, d) \wedge \neg \text{sametxn}(a, c)$, where a and b considered to belong to the *current* transaction.

We assume that operations not part of any transaction belong to their own unique transaction. While transactions may have varying isolation guarantees, we make the standard assumption that all transactions provide atomicity. Hence, we include the following axiom in Δ :

$$\forall a, b, c. \text{txn}\{a\}\{b, c\} \wedge \text{sameobj}(b, c) \wedge \text{vis}(b, a) \Rightarrow \text{vis}(c, a)$$

The semantics of this contract is illustrated in Figure 9(a).

6.2 Transactional Bank Account

In order to illustrate the utility of declarative reasoning for transactions, consider an extension of our running example to use two accounts (objects) – current (c) and savings (s). Each account provides operations `withdraw`, `deposit` and `getBalance`, with the same contracts as defined previously. We consider two transactions – `save(amt)`, which transfers `amt` from current to savings, and `totalBalance`, which returns the sum of the balances of individual accounts. Our goal is to ensure that `totalBalance` returns the result obtained from a consistent snapshot of the object states. The QUELEA code for these transactions is given below:

```

save amt =                                     totalBalance =
  x ← $(classify  $\psi_{sv}$ )                         x ← $(classify  $\psi_{tb}$ )
  atomically x $ do                               atomically x $ do
    b ← withdraw c amt                             b1 ← getBalance c
    when b $ deposit s amt                         b2 ← getBalance s
                                                    return b1 + b2

```

ψ_{sv} and ψ_{tb} are the contracts on the corresponding transactions. The function `classify` assigns the contracts *statically* to one of the transaction isolation levels offered by the store; $\$(\cdot)$ is meta-programming syntax for splicing the result into the program. The `atomically` construct invokes the enclosing operations at the given isolation level x , ensuring that the effects of the operations are made visible atomically.

While making both transactions serializable would ensure correctness, distributed stores rarely offer serializable transactions given their availability requirements and implications for scalabil-

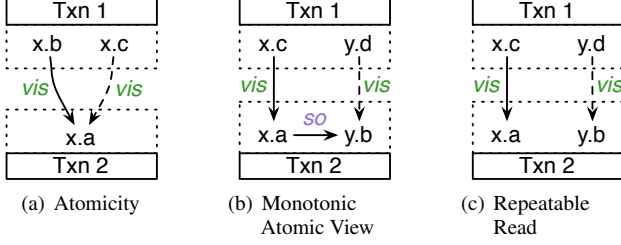


Figure 9: Semantics of transaction contracts. x and y are distinct objects. The dotted line represents the visibility requested by the contracts.

ity [2]. Fortunately, these transactions can be satisfied with much weaker isolation guarantees. However, despite the atomicity offered by the transaction, anomalies are still possible. For example, the two `getBalance` operations in a `totalBalance` transaction might be served by different replicas with a distinct set of committed `save` transactions. If the first(second) `getBalance` operation witnesses a `save` transaction that is not witnessed by the second(first) `getBalance` operation, then the balance returned will be less(greater) than the actual balance. It is not immediately apparent how to choose the weakest isolation guarantee that would be sufficient to prevent the anomaly.

Instead, QUELEA requires the programmer to simply state the consistency requirement as a contract. Since we would like both the `getBalance` operations to witness the same set of `save` transactions, we define the constraint on the `totalBalance` transaction ψ_{tb} as:

$$\begin{aligned} \psi_{tb} = & \forall a : \text{getBalance}, b : \text{getBalance}, \\ & (c : \text{withdraw} \vee \text{deposit}), (d : \text{withdraw} \vee \text{deposit}). \\ & \text{txn}\{a, b\}\{c, d\} \wedge \text{vis}(c, a) \wedge \text{sameobj}(d, b) \Rightarrow \text{vis}(d, b) \end{aligned}$$

The key idea in the above definition is that the `txn` primitive allows us to relate operations on different objects.

The `save` transaction only needs to ensure that the two writes it performs are made visible atomically. Since this is ensured by combining them in a transaction, `save` does not require any additional constraints, and ψ_{sv} is simply true.

6.3 Coordination-free Transactions

In order to illustrate the utility of transaction contract classification, we identify three well-understood coordination-free transaction semantics – Read Committed (RC) [7], Monotonic Atomic View (MAV) [2] and Repeatable Read (RR) [7], and illustrate the classification strategy. Our technique can indeed be applied to a different isolation-level lattice.

A transaction with ANSI RC semantics only witnesses committed operations. Let us assume that a replica will buffer transactional updates until all the updates from the same transaction are available at that replica. Once all the updates from a transaction are available, the buffered updates are made visible to subsequent client requests. This ensures atomicity of transactions. Importantly, RC does not entail any other guarantees. As a result, a store implementing RC does not require inter-replica coordination. We can express RC as follows:

$$\begin{aligned} \psi_{rc} = & \forall a, b, c. \text{txn}\{a\}\{b, c\} \wedge \text{sameobj}(b, c) \\ & \wedge \text{vis}(b, a) \Rightarrow \text{vis}(c, a) \end{aligned}$$

Notice that the above definition is the same as the atomicity guarantee of transaction described in § 6.1. The `save` operation is an example of an RC transaction.

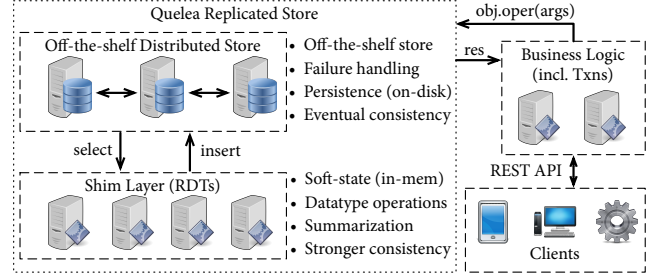


Figure 10: Implementation Model.

MAV semantics ensures that if some operation in a transaction T_1 witnesses the effects of another transaction T_2 , then subsequent operations in T_1 will also witness the effects of T_2 . MAV semantics is useful for maintaining the integrity of foreign key constraints, materialized views and secondary updates [2]. In order to implement MAV, a store only needs to keep track of the set of transactions S_t witnessed by the running transaction, and before performing an operation at some replica, ensure that the replica includes all the transactions in S_t . Hence, MAV is coordination-free. MAV semantics is captured with the following contract:

$$\begin{aligned} \psi_{\text{mav}} = & \forall a, b, c, d. \text{txn}\{a, b\}\{c, d\} \wedge \text{so}(a, b) \wedge \text{vis}(c, a) \\ & \wedge \text{sameobj}(d, b) \Rightarrow \text{vis}(d, b) \end{aligned}$$

whose semantics is illustrated in the Figure 9(b).

ANSI RR semantics requires that the transaction witness a snapshot of the data store state. Importantly, this snapshot can be obtained from any replica, and hence RR is also coordination-free. An example for such a transaction is the `totalBalance` transaction. The semantics of RR is captured by the following contract:

$$\begin{aligned} \psi_{rr} = & \forall a, b, c, d. \text{txn}\{a, b\}\{c, d\} \wedge \text{vis}(c, a) \\ & \wedge \text{sameobj}(d, b) \Rightarrow \text{vis}(d, b) \end{aligned}$$

whose semantics is illustrated in the Figure 9(c).

6.4 Classification

Similar to operation-level contracts, with respect to \leq relation, the coordination-free transaction semantics described here form a total order: $\psi_{rc} \leq \psi_{\text{mav}} \leq \psi_{rr}$. The transaction classification is also similar to the operation-level contract classification presented in Figure 6; given a contract ψ on a transaction, we start from the weakest transaction contract ψ_{rc} , and progressively compare its strength to the known transaction contracts until we find a isolation level under which ψ can be safely discharged. Otherwise, we report a type error.

7. Implementation

QUELEA is implemented as a shallow extension of GHC Haskell and runs on top of Cassandra, an off-the-shelf eventually consistent distributed data (or backing) store responsible for all data management issues (i.e., replication, fault tolerance, availability, and convergence). Template Haskell is used to implement static contract classification, and proof obligations are discharged with the help of the Z3 [30] SMT solver. Figure 10 illustrates the overall system architecture.

Replicated data types and various consistency semantics are implemented and enforced in the *shim layer*. Our implementation supports eventual, causal, and strong consistency for data type operations, and RC, MAV, and RR semantics for transactions. This functionality is implemented entirely on top of the standard interface exposed by Cassandra. From an engineering perspective, leveraging

```

1: atomically {
2:   o1.oper1(v1); //Emits effect e1
3:   o2.oper2(v2); //Emits effect e2
4:   o3.oper3(v3); //Emits effect e3
5: }

```

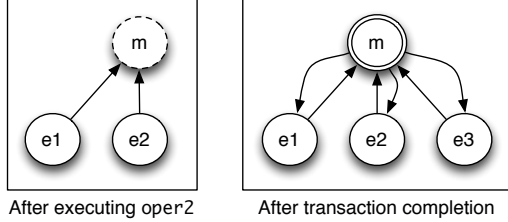


Figure 11: Implementing atomicity semantics.

an off-the-shelf data store enables an implementation comprising roughly only 2500 lines of Haskell code, which is packaged as a library.

7.1 Operation Consistency

The shim layer maintains a causally consistent in-memory snapshot of a subset of objects in the backing store, by explicitly tracking dependencies introduced between effects due to visibility, session and same transaction relations. Dependence tracking is similar to the techniques presented in [3] and [20]. Because Cassandra provides durability, convergence, and fault tolerance, each shim layer node simply acts as a soft-state cache, with no inter-node communication, and can safely be terminated at any point. Similarly, new shim layer nodes can be spawned on demand.

Each effect generated as a result of an effectful operation on an object inserts a new row ($o, e, txn, val, deps$) into the backing store, where o and e are object and *unique* effect identifiers, txn is an optional transaction identifier, and val is the value associated with the effect (eg: `Withdraw 50`). $deps$ is the set of identifiers of *dependencies* of this operation and is defined as $deps(e) = \{e_1 \mid vis(e_1, e) \wedge \neg(\exists e_2. vis(e_1, e_2) \wedge vis(e_2, e))\}$. At any shim layer node, an effect is included only if all of its dependencies are also included in that node. This ensures that the state maintained by the shim layer node is causally consistent. Hence, our dependence tracking strategy ensures that QUELEA does not track every effect as the number of writes in the system grows.

The shim layer nodes periodically fetch updates from the backing store for eventually consistent operations, and on-demand for causally consistent and strongly consistent operations. Strongly consistent operations are performed after obtaining exclusive leases on objects. The lease mechanism is implemented with the help of Cassandra’s support for conditional updates and expiring columns.

7.2 Transactions

Cassandra does not provide general-purpose transactions. Since the transaction guarantees provided by QUELEA are coordination-free [2], we realize efficient implementations by explicitly tracking dependencies between operations and transactions. Importantly, the weaker isolation semantics of transactions in QUELEA permit transactions to be discharged if at least one shim layer node is reachable.

QUELEA implements atomic visibility by exploiting shim layer causality guarantees – an effect is included only if all the effects it depends on are also included. Consider the example given in Figure 11. In the figure, the graphs represent the state of the store, where the circles represent effects and the edges represent the

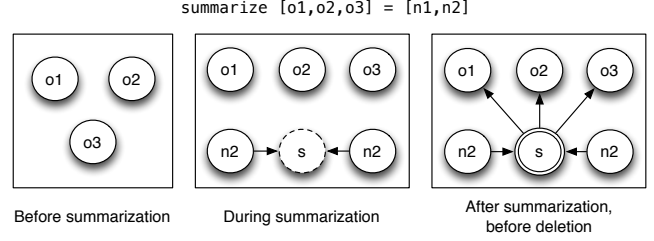


Figure 12: Summarization in the backing store.

dependence between effects. The dotted circle represents effects that are not yet inserted into the store. The graph on the left shows that state of the store after executing `oper2`. For every transaction in QUELEA, we instantiate a special transaction marker effect m that is importantly not inserted into the backing store. Marker m is included as a dependence to every effect generated in the transaction. Since the causally preceding effect m has not yet been written to the store, no operation will witness $e1$ and $e2$ while the transaction is in progress. After the transaction has finished execution, we insert m into the backing store, marking all the effects from the transactions as a dependence for m , as shown in the graph on the right. Now, any replica which includes one of the effects from the transaction must include m , and transitively must include every effect from the transaction. This ensures atomicity and satisfies the RC requirement.

The above scheme prevents a transaction from witnessing its own effects. This might conflict with causality requirements on the operations. Hence, transactions piggy-back the previous effects from the same transaction for each request. MAV semantics is implemented by keeping track of the set of transaction markers M witnessed by the transaction, and before performing an operation at some replica, ensuring that M is a subset of the transaction markers included at that replica. If not, the missing effects are synchronously fetched. RR semantics is realized by capturing an optimized snapshot of the state of some replica; each operation from an RR transaction is applied to this snapshot state. Any generated effects are added to this snapshot.

7.3 Summarization

We utilize the `summarize` function (§ 3.1.1) to summarize the object state both in the shim layer node and the backing store, typically when the number of effects on an object crosses a tunable threshold. Shim layer summarization is straight-forward; a summarization thread takes the local lock on the cached object, and replaces its state with the summarized state. The shim layer node only remains unavailable for that particular object during summarization (usually a few milliseconds).

Performing summarization in the backing store is more complicated since the whole process needs to be atomic from a client’s perspective, but Cassandra does not provide multi-row transactions. Summarization in the backing store involves deleting previously inserted rows and inserting new rows, where each row corresponds to an effect. It is essential that concurrent client operations are permitted, but are not allowed to witness the intermediate state of the summarization process.

To this end, we adopt a novel summarization strategy that builds on the causality property of the store. Figure 12 illustrates the summarization strategy. Suppose the original set of effects on an object are $o1, o2$ and $o3$. When summarized, the new effects yielded are $n1$ and $n2$. We first instantiate a summarization marker s , and similar to transaction marker, we do not insert it into the store immediately. We insert the new effects $n1$ and $n2$, with strong consistency, including s as a dependence. Since s is not yet in the store, the new effects are not made visible to the clients. Then we

Table 1: The distribution of classified contracts. #T refers to the number of tables in the application. The columns 4-6 (7-9) represent operations (transactions) assigned to this consistency (isolation) level.

Benchmark	LOC	#T	EC	CC	SC	RC	MAV	RR
LWW Reg	108	1	2	2	2	0	0	0
DynamoDB	126	1	3	1	2	0	0	0
Bank Account	155	1	1	1	1	1	0	1
Shopping List	140	1	2	1	1	0	0	0
Online store	340	4	9	1	0	2	0	1
RUBiS	640	6	14	2	1	4	2	0
Microblog	659	5	13	6	1	6	3	1

insert s with strong consistency, including the original effects $o1$, $o2$ and $o3$ as dependence. Strongly consistent insertions ensure that a shim layer node witnessing s on some object must also witness $n1$ and $n2$ on the same object. A shim layer node which witnesses all the effects removes the original effects from its cache since they are superseded by the new effects. Finally, the old effects are deleted from the backing store. This process ensures that clients either witness the old or the new effects, but not both; the summarization process appears to be atomic from the clients perspective.

8. Evaluation

We present an evaluation study of our implementation, report contract profiles of benchmark programs, and illustrate the performance benefits of fine-grained consistency classification on operations and transactions. We also evaluate the impact of the summarization. We have implemented the following applications, which includes individual RDTs as well as larger applications composed of several RDTs:

- **LWW register:** A last-write-wins register that provides read and write operations, where the read returns the value of the latest write.
- **DynamoDB register:** An integer register that allows eventual and strong puts and gets, conditional puts, increment and decrement operations.
- **Bank account:** Our running example.
- **Shopping list:** A collaborative shopping list that allows concurrent addition and deletion of items.
- **Online store:** An online store with shopping cart functionality and dynamically changing item prices. The checkout process verifies that the customer only pays the accepted price.
- **RUBiS:** An eBay-like auction site [23]. The application allows users to browse items, bid for items on sale, and pay for items from a wallet modeled after a bank account.
- **Microblog:** A twitter-like microblogging site, modeled after Twissandra [29]. The application allows adding new users, adding and replying to tweets, following, unfollowing and blocking users, and fetching a user’s timeline, userline, followers and following.

The distribution of contracts in these applications is given in Table 1. We see that majority of the operations and transactions are classified as eventually consistent and RC, respectively. Operation contracts are used to enforce integrity and visibility constraints on individual fields in the tables. Transactions are mainly used to consistently modify and access related fields across tables. In QUELEA, the contract classification process is completely performed at com-

pile time and incurs no runtime overhead. The proof obligations associated with contract classification is discharged through the Z3 SMT Solver. Across our benchmarks, classifying a contract took 11.5 milliseconds on average.

For our performance evaluation, we deploy QUELEA applications in *clusters*, where each cluster is composed of five fully replicated Cassandra replicas within the same datacenter. We instantiate one shim layer node co-located on the same VM as a Cassandra replica. Clients are instantiated within the same data center as the store, and run transactions. We deploy each cluster and client node on a `c3.4xlarge` Amazon EC2 instance. We call this a `1DC` configuration. For our geo-distributed experiments (`2DC`), we instantiate 2 clusters, each with five nodes, and place the clusters on US-east (Virginia) and US-west (Oregon) locations. The average inter-region latency was 85ms.

Figure 13 shows throughput vs. latency of operations in the bank account example as we increase the number of clients in a `1DC` configuration. Our client workload was generated using the YCSB benchmark [12]. The benchmark uniformly chooses from 100,000 keys, where the operation spread was 25% withdraw, 25% deposit and 50% getBalance, which corresponds to the default 50:50 read:write mix in YCSB. We increased the number of clients from 128 to 1024, and each experiment ran for 180 seconds.

The lines marked EC and CC correspond to all operations (including `withdraw`) being assigned EC and CC consistency levels. These levels compromise correctness as `withdraw` has to be an SC operation. The SC line corresponds to a configuration where all operations are strongly consistent; this ensures application correctness, at the cost of performance. QUELEA corresponds to our implementation, which classifies operations based on their contract specifications. With 512 clients, the QUELEA implementation was within 41% of the latency and 18% of the throughput of EC, whereas SC operations had 162% higher latency and 52% lower throughput than EC operations. Observe that there is a point in each case after which the latency increases while the throughput decreases; these correspond to points where the store becomes saturated with client requests. In a `2DC` configuration (not shown here), the average latency of SC operations with 512 clients increased by $9.4\times$ due to the cost of geo-distributed coordination, whereas QUELEA operations were only $2.2\times$ slower, mainly due to the increased cost of `withdraw` operations. Importantly, the latency of `getBalance` and `deposit` remained almost the same, illustrating the benefit of fine-grained contract classification.

We compare the performance of different transaction isolation level choices in Figure 14 using the LWW register. The numbers were obtained under a `1DC` configuration. The YCSB workload was modified to issue 10 operations per transaction, with a default 50:50 read:write mix. Each operation is assumed to be eventually consistent. NoTxn corresponds to a configuration that does not use transactions. Compared to this, RC is only 12% slower in terms of latency with 512 clients, whereas RR is $2.3\times$ slower. The difference between RC and NoTxn is due to the meta-data overhead of recording transaction information in the object state. For RR transactions, the cost of capturing and maintaining a snapshot is the biggest source of overhead.

We also compared (not shown) the performance of EC LWW operations directly against Cassandra (our backing store), which uses last-writer-wins as the only convergence semantics. While Cassandra provides no stronger-than-eventual consistency properties, QUELEA was within 30%(20%) of latency(throughput) of Cassandra with 512 clients, illustrating that the programmers only have to pay a minimal overhead for the expressive and stronger QUELEA programming model.

Figure 15 compares the QUELEA implementation of RUBiS in `1DC` configuration against a single replica (NoRep) and strongly

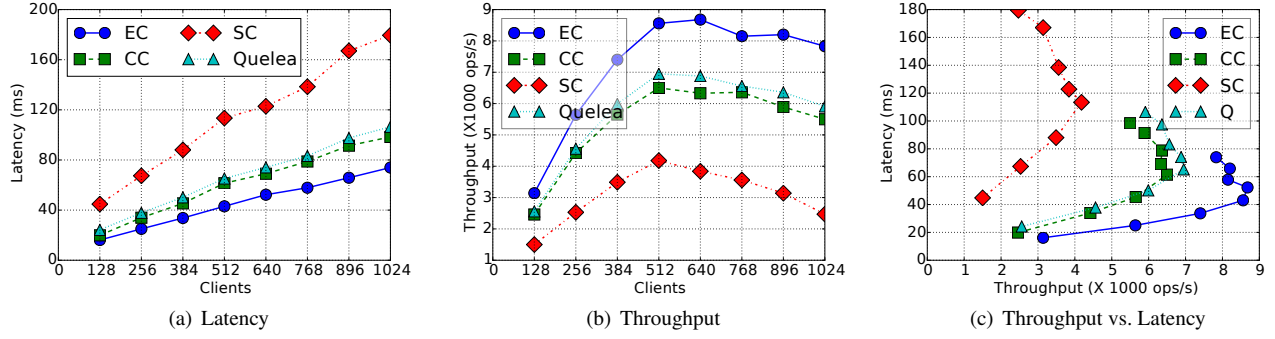


Figure 13: Bank account performance.

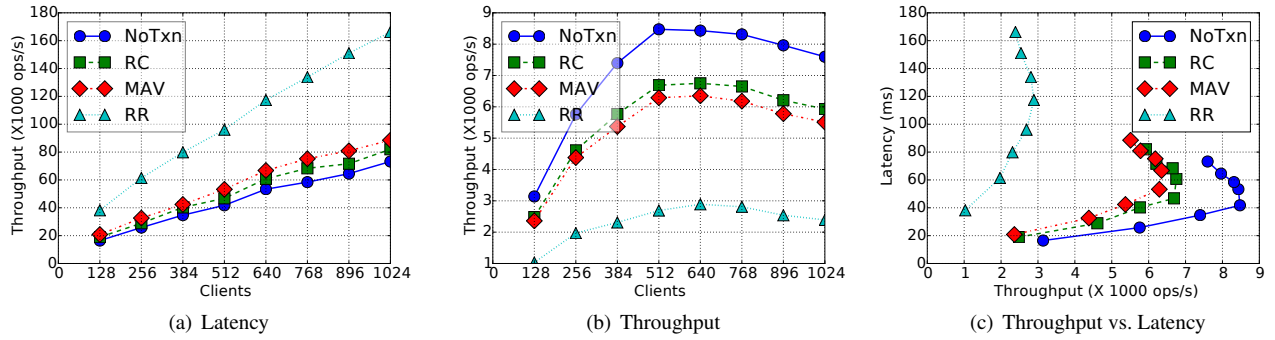


Figure 14: LWW register transaction performance.

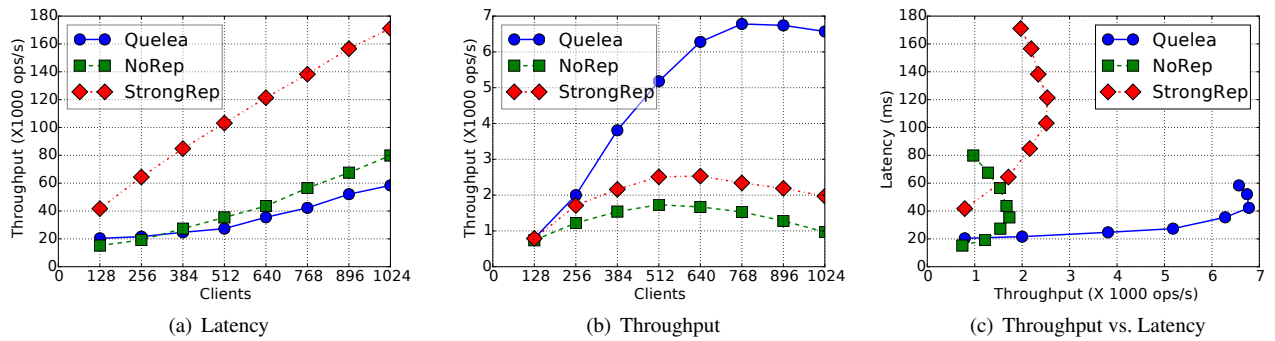


Figure 15: Rubis bidding mix performance.

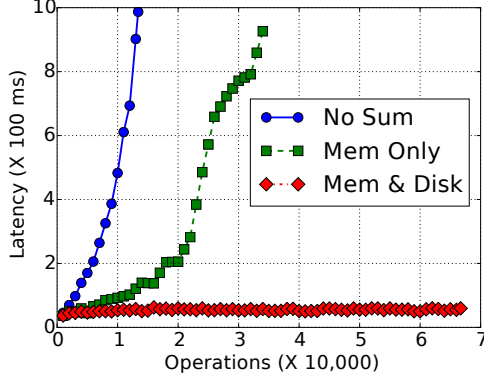


Figure 16: Impact of summarization.

replicated (StrongRep) ¹DC deployment. The benchmark was RUBiS bidding mix, which has 15% read-write interactions, which is representative of the auction workload. Without replication, NoRep trivially provides strong consistency. However, this deployment does not scale beyond 1750 operations per second. Strong replication offers better throughput at the cost of greater latency due to inter-replica coordination. QUELEA deployment offers the benefit of replication, while only paying the cost of coordination when necessary.

Finally, we study the impact of summarization in Figure 16. We use 128 clients and a single QUELEA replica, with all clients operating on the *same* LWW register to stress test the summarization mechanism. The shim layer cache (memory) is summarized every 64 updates, while the updates in the backing store (disk) are summarized every 4096 updates. Each point in the graph represents the average latency of the previous 1000 operations. Each experiment is run for one minute. Without summarization, the average latency of operations increases exponentially to almost one second, and only 13K operations were performed in a minute. Since every operation has to reduce over the set of all previous operations, operations take increasingly more time to complete since they must contend with an ever growing set. With summarization only in memory, performance still degrades due to the cost of fetching all previous updates from the backing store into the shim layer. Fetching the latest updates from the backing store is essential for SC operations. With summarization enabled on both disk and memory, latency does not increase over time, and the implementation realizes throughput of 67K operations/minute.

9. Related Work

Operation-based RDTs have been widely studied in terms of their algorithmic properties [10, 24], and several systems utilize this model to construct distributed data structures [5, 16, 22]. These systems typically propose to implement the datatypes directly over a cluster of nodes, and only focus on basic eventual consistency. Hence, these systems implement custom solutions for durability and fault-tolerance. QUELEA realizes stronger consistency models for RDTs on top of off-the-shelf eventually consistent distributed stores. In this respect, QUELEA is similar to [3] where causal consistency is achieved through a shim layer layered on top of Cassandra, which explicitly tracks and enforces dependencies between updates. However, [3] does not support user-defined RDTs, automatic contract classification and transactions.

Since eventual consistency alone is insufficient to build correct applications, several systems [17, 22, 28] propose a lattice of stronger consistency levels. Similarly, traditional database processing systems [7] and their replicated variants [2] propose weaker isolation levels for performance. In these systems, the onus is on the

developer to choose the correct consistency(isolation) level for operations(transactions). QUELEA relieves the developer of this burden, and instead only requires contracts to express declarative visibility requirements.

Our contract language and system model is inspired by the axiomatic description of RDT semantics proposed in [10]. While [10] uses axioms for formal verification of an RDT implementation, we utilize them as a means for the user to express desired consistency guarantees in the application. The operational semantics of QUELEA describes an operational manifestation of our system model. [11] also presents an operational model of a replicated data store that is based on the abstract system model presented in [10]. Their claims about the expressivity and practicality of the system model vindicate ours. However, both approaches differ in the way they aim to empower the application programmer. [11] exposes the operational model to the programmer, along with primitives, such as `push`, `pull` and `fetch`, that make it easy to *implement* required consistency and isolation guarantees. In contrast, QUELEA completely abstracts its operational model, and instead relies on contract classification to automatically *choose* appropriate consistency and isolation levels. Similar to [10] and [11], our system model does not incorporate real (i.e., wall-clock) time. Hence, our contract language cannot describe store semantics such as recency or bounded-staleness guarantees offered by certain stores [28].

Several conditions have been proposed to judge whether an operation on a replicated data object needs coordination or not. [1] defines *logical monotonicity* as a sufficient condition for coordination freedom, and proposes a consistency analysis that marks code regions performing non-monotonic reasoning (eg: aggregations, such as `COUNT`) as potential coordination points. [4] and [18] define *invariant confluence* and *invariant safety*, respectively, as conditions for safely executing an operation without coordination. [18] also proposes a program analysis that conservatively marks certain operations as *blue* (coordination not required), while marking the remaining as *red* (coordination required). [6] requires programmers to declare application semantics, and the desired application-specific invariants as formulas in first-order logic. It performs static analysis on these formulas to determine *I*-offender sets - sets of operations, which, when performed concurrently, result in violation of one or more of the stated invariants. For each offending set of operations, if the programmer chooses invariant-violation avoidance over violation repair, the system employs various techniques, such as escrow reservation, to ensure that the offending set is effectively serialized. Unlike QUELEA, these approaches focus on a coarse-grained classification of consistency as eventual or strong, and do not consider finer-grained transaction isolation levels. However, these analyses do relieve programmers of the burden of tagging operations with consistency levels. We consider automatic inference of consistency contracts from application-specific integrity constraints as an obvious next step for QUELEA.

10. Conclusions

This paper presents QUELEA, a shallow Haskell extension for declarative programming over ECDS. The key idea underlying QUELEA’s design is the automatic classification of fine-grained consistency contracts on operations and distributed transactions with respect various consistency and isolation levels offered by the store. Our contract language is carefully crafted from a decidable subset of first-order logic, enabling the use of automated verification tools to discharge the proof obligations associated with contract classification. We realize an instantiation of QUELEA on top of an off-the-shelf distributed store, Cassandra, and illustrate the benefit of fine-grained contract classification by implementing and evaluating several scalable applications.

Acknowledgements

We thank the anonymous reviewers for their detailed comments. This work is supported in part by the National Science Foundation under grants CCF-1216613 and CCF-1318227.

References

- [1] P. Alvaro, N. Conway, J. Hellerstein, and W. R. Marczak. Consistency Analysis in Bloom: a CALM and Collected Approach. In *CIDR 2011, Fifth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 9-12, 2011, Online Proceedings*, pages 249–260, 2011. URL http://www.cidrdb.org/cidr2011/Papers/CIDR11_Paper35.pdf.
- [2] P. Bailis, A. Davidson, A. Fekete, A. Ghodsi, J. M. Hellerstein, and I. Stoica. Highly Available Transactions: Virtues and Limitations. *PVLDB*, 7(3):181–192, 2013.
- [3] P. Bailis, A. Ghodsi, J. M. Hellerstein, and I. Stoica. Bolt-on Causal Consistency. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, SIGMOD '13*, pages 761–772, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2037-5. doi: 10.1145/2463676.2465279.
- [4] P. Bailis, A. Fekete, M. J. Franklin, A. Ghodsi, J. M. Hellerstein, and I. Stoica. Coordination-Avoiding Database Systems. *CoRR*, abs/1402.2237, 2014. URL <http://arxiv.org/abs/1402.2237>.
- [5] M. Balakrishnan, D. Malkhi, T. Wobber, M. Wu, V. Prabhakaran, M. Wei, J. D. Davis, S. Rao, T. Zou, and A. Zuck. Tango: Distributed Data Structures over a Shared Log. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles, SOSP '13*, pages 325–340, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2388-8. doi: 10.1145/2517349.2522732.
- [6] V. Balesgas, N. Preguiça, R. Rodrigues, S. Duarte, C. Ferreira, M. Najafzadeh, and M. Shapiro. Putting the Consistency back into Eventual Consistency. In *Proceedings of the Tenth European Conference on Computer System, EuroSys '15*, Bordeaux, France, 2015. URL <http://lip6.fr/Marc.Shapiro/papers/putting-consistency-back-EuroSys-2015.pdf>.
- [7] H. Berenson, P. Bernstein, J. Gray, J. Melton, E. O'Neil, and P. O'Neil. A Critique of ANSI SQL Isolation Levels. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, SIGMOD '95*, pages 1–10, New York, NY, USA, 1995. ACM. ISBN 0-89791-731-6. doi: 10.1145/223784.223785.
- [8] E. Brewer. Towards Robust Distributed Systems (Invited Talk), 2000.
- [9] S. Burckhardt, D. Leijen, M. Fähndrich, and M. Sagiv. Eventually Consistent Transactions. In *Proceedings of the 21st European Conference on Programming Languages and Systems, ESOP '12*, pages 67–86, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-28868-5. doi: 10.1007/978-3-642-28869-2_4.
- [10] S. Burckhardt, A. Gotsman, H. Yang, and M. Zawirski. Replicated Data Types: Specification, Verification, Optimality. In *Proceedings of the 41st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, POPL '14*, pages 271–284, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2544-8. doi: 10.1145/2535838.2535848.
- [11] S. Burckhardt, D. Leijen, J. Protzenko, and M. Fähndrich. Global Sequence Protocol: A Robust Abstraction for Replicated Shared State. In *Proceedings of the 29th European Conference on Object-Oriented Programming, ECOOP '15*, Prague, Czech Republic, 2015. URL <http://research.microsoft.com/pubs/240462/gsp-tr-2015-2.pdf>.
- [12] B. F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears. Benchmarking Cloud Serving Systems with YCSB. In *Proceedings of the 1st ACM Symposium on Cloud Computing, SoCC '10*, pages 143–154, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0036-0. doi: 10.1145/1807128.1807152.
- [13] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Voshall, and W. Vogels. Dynamo: Amazon's Highly Available Key-value Store. In *Proceedings of Twenty-first ACM SIGOPS Symposium on Operating Systems Principles, SOSP '07*, pages 205–220, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-591-5. doi: 10.1145/1294261.1294281.
- [14] S. Gilbert and N. Lynch. Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-tolerant Web Services. *SIGACT News*, 33(2):51–59, June 2002. ISSN 0163-5700. doi: 10.1145/564585.564601.
- [15] M. P. Herlihy and J. M. Wing. Linearizability: A Correctness Condition for Concurrent Objects. *ACM Transactions on Programming Languages and Systems*, 12(3):463–492, July 1990. ISSN 0164-0925. doi: 10.1145/78969.78972.
- [16] A. Lakshman and P. Malik. Cassandra: A Decentralized Structured Storage System. *SIGOPS Operating Systems Review*, 44(2):35–40, Apr. 2010. ISSN 0163-5980. doi: 10.1145/1773912.1773922.
- [17] C. Li, D. Porto, A. Clement, J. Gehrke, N. Preguiça, and R. Rodrigues. Making Geo-replicated Systems Fast As Possible, Consistent when Necessary. In *Proceedings of the 10th USENIX Conference on Operating Systems Design and Implementation, OSDI'12*, pages 265–278, Berkeley, CA, USA, 2012. USENIX Association. ISBN 978-1-931971-96-6. URL <http://dl.acm.org/citation.cfm?id=2387880.2387906>.
- [18] C. Li, J. a. Leitão, A. Clement, N. Preguiça, R. Rodrigues, and V. Vafeiadis. Automating the Choice of Consistency Levels in Replicated Systems. In *Proceedings of the 2014 USENIX Conference on USENIX Annual Technical Conference, USENIX ATC'14*, pages 281–292, Berkeley, CA, USA, 2014. USENIX Association. ISBN 978-1-931971-10-2. URL <http://dl.acm.org/citation.cfm?id=2643634.2643664>.
- [19] W. Lloyd, M. J. Freedman, M. Kaminsky, and D. G. Andersen. Don't Settle for Eventual: Scalable Causal Consistency for Wide-area Storage with COPS. In *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles, SOSP '11*, pages 401–416, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0977-6. doi: 10.1145/2043556.2043593.
- [20] W. Lloyd, M. J. Freedman, M. Kaminsky, and D. G. Andersen. Stronger Semantics for Low-latency Geo-replicated Storage. In *Proceedings of the 10th USENIX Conference on Networked Systems Design and Implementation, nsdi'13*, pages 313–328, Berkeley, CA, USA, 2013. USENIX Association. URL <http://dl.acm.org/citation.cfm?id=2482626.2482657>.
- [21] C. H. Papadimitriou. The Serializability of Concurrent Database Updates. *Journal of the ACM*, 26(4):631–653, Oct. 1979. ISSN 0004-5411. doi: 10.1145/322154.322158.
- [22] K. Petersen, M. J. Spreitzer, D. B. Terry, M. M. Theimer, and A. J. Demers. Flexible Update Propagation for Weakly Consistent Replication. In *Proceedings of the Sixteenth ACM Symposium on Operating Systems Principles, SOSP '97*, pages 288–301, New York, NY, USA, 1997. ACM. ISBN 0-89791-916-5. doi: 10.1145/268998.266711.
- [23] RUBiS. Rice University Bidding System, 2014. URL <http://rubis.ow2.org/>. Accessed: 2014-11-4 13:21:00.
- [24] M. Shapiro, N. Preguiça, C. Baquero, and M. Zawirski. Conflict-Free Replicated Data Types. In X. Défago, F. Petit, and V. Villain, editors, *Stabilization, Safety, and Security of Distributed Systems*, volume 6976 of *Lecture Notes in Computer Science*, pages 386–400. Springer Berlin Heidelberg, 2011. ISBN 978-3-642-24549-7. doi: 10.1007/978-3-642-24550-3_29.
- [25] S. Sivasubramanian. Amazon dynamoDB: A Seamlessly Scalable Non-relational Database Service. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, SIGMOD '12*, pages 729–730, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1247-9. doi: 10.1145/2213836.2213945.
- [26] Y. Sovran, R. Power, M. K. Aguilera, and J. Li. Transactional Storage for Geo-replicated Systems. In *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles, SOSP '11*, pages 385–400, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0977-6. doi: 10.1145/2043556.2043592.
- [27] D. B. Terry, A. J. Demers, K. Petersen, M. Spreitzer, M. Theimer, and B. W. Welch. Session Guarantees for Weakly Consistent Replicated

Data. In *Proceedings of the Third International Conference on Parallel and Distributed Information Systems*, PDIS '94, pages 140–149, Washington, DC, USA, 1994. IEEE Computer Society. ISBN 0-8186-6400-2. URL <http://dl.acm.org/citation.cfm?id=645792.668302>.

- [28] D. B. Terry, V. Prabhakaran, R. Kotla, M. Balakrishnan, M. K. Aguilera, and H. Abu-Libdeh. Consistency-based Service Level Agreements for Cloud Storage. In *Proceedings of the Twenty-Fourth ACM Symposium*

on Operating Systems Principles, SOSP '13, pages 309–324, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2388-8. doi: 10.1145/2517349.2522731.

- [29] Twissandra. Twitter clone on Cassandra, 2014. URL <http://twissandra.com/>. Accessed: 2014-11-4 13:21:00.

- [30] Z3. High-performance Theorem Prover, 2014. URL <http://z3.codeplex.com/>. Accessed: 2014-11-4 13:21:00.