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 to provide 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 composably 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, and provide 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

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 is 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
The rest of the paper is organized as follows. The next section presents related work and conclusions. § 7 discusses experimental evaluation. § 8 and 9 cover more details on the contract language, and its mapping to the store. We describe the system model. We describe the challenges associated with constructing a consistent replicated data type (RDT) in QUELEA, and then describe the mechanisms to obtain the desired correctness guarantees.

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 \((x, y, \ldots)\). 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 \(\text{effects}\) 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 \(\text{foo}\) and \(\text{bar}\) invoked by the same session on the same object, might end up being applied to different replicas because replica 1 (to which \(\text{foo}\) was applied) might be unreachable when the client invokes \(\text{bar}\).

When \(\text{foo}\) is invoked on an object \(x\) with arguments \(\text{arg}_1\), 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 \(v_1\) that is sent back to the client, and emits a single new effect \(w^3_x\) 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 \(w^4_x\) and \(w^5_x\) are visible to \(w^3_x\), written logically as \(\text{vis}(w^4_x, w^5_x) \land \text{vis}(w^3_x, w^5_x)\), where \(\text{vis}\) 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 \(w^6_x\) and \(w^7_x\) arising from operations applied to the same object \(x\) are said to be under the same object relation, written \(\text{sameobj}(w^6_x, w^7_x)\). Finally, we can associate every effect with the operation that generated the effect with the help of a relation \(\text{op}\). In the example above, \(\text{op}(w^4_x, \text{foo})\) and \(\text{op}(w^5_x, \text{bar})\) 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 correctness guarantees 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 \(w^6_x\) and \(w^7_x\) arising from operations applied to the same object \(x\) are said to be under the same object relation, written \(\text{sameobj}(w^6_x, w^7_x)\). Finally, we can associate every effect with the operation that generated the effect with the help of a relation \(\text{op}\). In the example above, \(\text{op}(w^4_x, \text{foo})\) and \(\text{op}(w^5_x, \text{bar})\) hold. For simplicity, we assume all operation names across all object are distinct.

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 the application logic to utilize the 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 concerns permits operational reasoning for conflict resolution, and
data Acc = Deposit Int | Withdraw Int | GetBalance

getBalance :: [Acc] → () → (Int, Maybe Acc)
getBalance hist _ =
  let res = ∑[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.

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 opera-
tions: 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 ∑ 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
100, Withdraw 50] is observably equivalent to a bank account with
a single deposit operation [Deposit 100]: we can replace the earlier
history with the latter and a client of the store would not 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 - (summarize) 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.

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 succeeds. Subsequently, session
2 perform 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 suc-
cedes, 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 vS edge indicates, operation
withdraw(50) in session 2 witnesses the effects of deposit(100)
from session 1, concludes that there is sufficient balance, and com-
pletes 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
not yet merged the withdraw effect. This anomaly leads the user
to incorrectly conclude that the withdraw operation failed to go
through.

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
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 it 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 *contract* (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 a 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 integrity constraint, the withdraw operation must be strictly 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 \((\psi_w)\) over *withdraw*:

\[
\forall(a:\text{withdraw}), (\text{sameobj}(a,\eta)) \Rightarrow a = \eta \lor \text{vis}(a,\eta) \lor \text{vis}(\eta, a)\]

Here, \(\eta\) stands for the effect emitted by the withdraw operation.

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 operation* - the operation for which a contract is being written. Notice that \(\hat{\eta}\) occurs free in the contract. We will fix its scope when

\[
\begin{align*}
x, y, \hat{\eta} \in \text{EffVar} & \quad \text{Op} \in \text{OperName} \\
\psi \in \text{Contract} & : = \forall(x: \tau), \psi \lor \forall x, \psi \lor \pi \\
\tau \in \text{EffType} & : = \text{Op} \lor \pi \lor \tau \\
\pi \in \text{Prop} & : = \text{true} \lor R(x, y) \lor \pi \lor \pi \\
R \in \text{Relation} & : = \text{vis} \lor \text{so} \lor \text{sameobj} \lor \pi \\
\end{align*}
\]

*Figure 4: Contract language.*

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 \((\psi_{gb})\) of the getBalance operation is merely a conjunction of the previous two versions \((\psi^1_{gb} \text{ and } \psi^2_{gb})\):

\[
\forall(a: \text{deposit}), (b: \text{withdraw}, (c: \text{deposit} \lor \text{withdraw}) \\
\text{vis}(a, b) \land \text{vis}(b, \hat{\eta}) \Rightarrow \text{vis}(c, \hat{\eta}) \lor \text{vis}(\hat{\eta}, a))
\]

Intuitively, this prevents both the getBalance anomalies described in figs. 3(b) and 3(c) from ever 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-scene. 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 prove 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 operation* - the operation for which a contract is being written. Notice that \(\hat{\eta}\) occurs free in the contract. We will fix its scope when
An important aspect of our contract language is its ability to capture \( R^+ \) and transitive closure of \( R \) 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 \( z, a) R(x, y) \Rightarrow R'(x, y, z) \Rightarrow R'(x, z) \).

Quantifier-free propositions in our contract language are con-
jectured to have a wide variety of store semantics including those that combine any store-level consistency guarantees, along with application-level consistency levels – eventual, causal, and strong, commonly offered over a wide variety of store semantics including those that combine any store-level consistency guarantees, along with application-level consistency levels – eventual, causal, and strong, commonly offered.

**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: \( \text{vis} \)) 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 \((\psi)\) expressible in our contract language. If \( E \) is indeed a valid model for \( \psi \) (written as \( E \models \psi \)), we say that the execution \( E \) satisfied the contract \( \psi \).

**Definition 1.** An axiomatic execution \( E \) satisfies a contract \( \psi \) 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 offer basic eventual consistency with all possible anomalies, we assume that our store provides stronger semantics that remain highly-available.

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 \( z, a) R(x, y) \Rightarrow R'(x, y, z) \Rightarrow R'(x, z) \), 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.

\[ \psi \leq \psi_{\text{ec}} \]

\[ \psi \leq \psi_{\text{ec}} \]

\[ \text{WellFormed}(\psi) \]

\[ \text{EventuallyConsistent}(\psi) \]

\[ \psi \leq \psi_{\text{ec}} \]

\[ \psi \leq \psi_{\text{ec}} \]

\[ \text{CausallyConsistent}(\psi) \]

\[ \text{StronglyConsistent}(\psi) \]

\[ \psi_{\text{ec}} = \forall a, b. \ \text{hbo}(a, b) \land \text{vis}(b, \hat{\eta}) \Rightarrow \text{vis}(a, \hat{\eta}) \]

- **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 the operations from the same session to be visible, which cannot be guaranteed under eventual consistency. The corresponding store semantics is captured by the contract \( \psi_{\text{ec}} \) defined below:

\[ \psi_{\text{ec}} = \forall a. \ \text{vis}(a, \hat{\eta}) \land \text{vis}(\hat{\eta}, a) \]

- **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 contract \( \psi_{\text{ec}} \) defined following:

\[ \psi_{\text{ec}} = \forall a. \ \text{sameobj}(a, \hat{\eta}) \Rightarrow \text{vis}(a, \hat{\eta}) \land \text{vis}(\hat{\eta}, a) \land a = \hat{\eta} \]

**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 \((\psi_{\text{op}})\) of an operation \((op)\) is weak enough to be satisfied under a store consistency level identified by the contract \(\psi_{\text{st}}\). Towards this end, we define a binary weaker than relation for our contract language as following:

**Definition 2.** A contract \(\psi_{\text{op}}\) is said to be weaker than \(\psi_{\text{st}}\) (written \(\psi_{\text{op}} \leq \psi_{\text{st}}\)) if and only if \(\text{If } \Delta \models \forall \eta, \psi_{\text{st}} \models \psi_{\text{op}}\).

The quantifier in the sequent binds \(\hat{\eta}\) that occurs free in \(\psi_{\text{st}}\) and \(\psi_{\text{op}}\). Context \((\Delta)\) 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 \((\Delta)\) 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**: \(\forall a, b, \text{so}(a, b) \land \text{so}(b, c) \Rightarrow \text{so}(a, c)\).
- **Same object is an equivalence relation**: \(\forall a. \ \text{sameobj}(a, a)\).
- \(\forall a, b. \ \text{sameobj}(a, b) \Rightarrow \text{sameobj}(b, a)\).
- \(\forall a, b, c. \ \text{sameobj}(a, b) \land \text{sameobj}(b, c) \Rightarrow \text{sameobj}(a, c)\).
If the contract $(\psi_{op})$ of an operation $(op)$ is weaker than a store contract $(\psi_{st})$, 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 $\psi_{op}$ (i.e., $E \models \psi_{op}$) also satisfies $\psi_{op}$ (i.e., $E \models \psi_{op}$). Consequently, it is safe to execute operation $(op)$ under a store consistency level captured by $\psi_{st}$.

Observe that the contracts $\psi_{sc}$, $\psi_{cc}$, and $\psi_{ec}$ are themselves totally ordered with respect to the $\leq$ relation: $\psi_{sc} \leq \psi_{cc} \leq \psi_{ec}$. This concurs with the intuition that any contract satisfiable under $\psi_{sc}$ or $\psi_{cc}$ is satisfiable under $\psi_{ec}$, and any contract that is satisfiable under $\psi_{ec}$ is satisfiable under $\psi_{cc}$. We are interested in the weakest guarantee (among $\psi_{sc}$, $\psi_{cc}$, and $\psi_{ec}$) 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 $(\psi_{gb})$ from § 3 as a Causally-Consistent contract, because the sequent $\Delta \vdash \psi_{gb} \Rightarrow \psi_{gb}$ is valid in first-order logic (therefore, $\psi_{gb} \leq \psi_{gb}$), whereas the sequent $\Delta \vdash \psi_{gb} \Rightarrow \psi_{gb}$ is invalid (therefore, $\psi_{gb} \not\leq \psi_{gb}$). 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 $\psi_{cc}$ - 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, b. so0(a, b) \Rightarrow vis(a, b)
Monotonic Reads (MR) := \forall a, b, c. vis(a, b) \land so0(b, c) \Rightarrow vis(a, c)
Monotonic Writes (MW) := \forall a, b, c. so0(a, b) \land vis(b, c) \Rightarrow vis(a, c)
Writes Follow Reads (WFR) := \forall a, b, c, d. vis(a, b) \land so0(b, c) \land so0(d, a) \Rightarrow vis(a, d)

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 weaker than relation between their corresponding contracts. Classifying a contract under this scheme is a directed search in the lattice, starting from the bottom, and determining the weakest consistency level under which the contract can be satisfied. Under this scheme, deposit operations does not need any guarantees, getBalance needs RYW and WFR $(\Delta \vdash \text{RYW} \land \text{WFR} \Rightarrow \psi_{gb})$, and withdraw cannot be satisfied $(\Delta \vdash \text{RYW} \land \text{MW} \land \text{MR} \land \text{WFR} \not\Rightarrow \psi_{st})$.

Figure 7: Lattice of consistency levels under session guarantees.

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:

\[ \text{Op} \in \text{Operation} \]
\[ \tau \in \text{ConsistencyClass} ::= \text{ec}, \text{cc}, \text{sc} \]
\[ \sigma \in \text{Session} ::= \cdots | (\text{op}, \tau); \sigma \]
\[ \Sigma \in \text{Session Soup} ::= \sigma \parallel \Sigma | \emptyset \]
\[ \text{Config} ::= \Sigma, \Sigma \]

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

$E,(\text{op}, \tau); \sigma \parallel \Sigma \xrightarrow{\Delta} E', \sigma \parallel \Sigma$

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 $\Sigma$, generating a new effect $\eta$. If the resultant execution $E'$ satisfies the store contract $\psi_r$ (i.e., $E \models \psi_r$), then we say that the store has enforced the contract $\psi_r$ 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 $\psi$ be a well-formed contract of a replicated data type operation $op$, and let $\tau$ denote the consistency class of $\psi$ as determined by the contract classification scheme. For all well-formed execution states $E, E'$ such that $E,(\text{op}, \tau); \sigma \parallel \Sigma \xrightarrow{\Delta} E', \sigma \parallel \Sigma$, if $E' \models \psi_r[\eta/\hat{\eta}]$, then $E' \models \psi[\eta/\hat{\eta}]$.

The theorem states that if a data store correctly enforces $\psi_{sc}$, $\psi_{cc}$, and $\psi_{ec}$ 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 in the supplementary material.
The supplementary material also provides the concrete reduction rules for enforcing the consistency classes.
5. 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 consistency semantics of individual operations.

5.1 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 its consistency specification. 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 \( \text{sametxn}(a, b) \) which holds for effects from the same transaction. \( \text{txn}(a, b) \{c, d\} \) is just syntactic sugar for \( \text{sametxn}(a, b) \land \text{sametxn}(c, d) \land \neg\text{sametxn}(a, c) \), 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 contracts provide atomicity. Hence, we include the following axiom in \( \Delta \):

\[ \forall a, b, c. \text{txn}(a)\{b, c\} \land \text{sameobj}(b, c) \land \text{vis}(b, a) \Rightarrow \text{vis}(c, a) \]

The semantics of this contract is illustrated in Figure 8(a).

5.2 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 - \( \text{save}(amt) \), which transfers \( amt \) from current to savings, and \( \text{totalBalance} \), which returns the sum of the balances of individual accounts. Our goal is to ensure that \( \text{totalBalance} \) returns the result obtained from a consistent snapshot of the object states. The QUELEA code for the transactions is given below:

\[
\begin{align*}
\text{save } \text{amt} = & \\
\text{totalBalance} = & \\
\text{x } & \leftarrow \$\text{(classify } \psi_{\text{save}}\text{)} \quad \text{x } & \leftarrow \$\text{(classify } \psi_{\text{totalBalance}}\text{)} \\
\text{atomically } x & \text{ do} \quad \text{atomically } x & \text{ do} \\
\text{b } & \leftarrow \text{withdraw } c \text{ amt} \quad \text{b1 } & \leftarrow \text{getBalance } c \\
\text{when } b & \text{ do} \text{ deposit } s \text{ amt} \quad \text{b2 } & \leftarrow \text{getBalance } s \\
\text{return } b1 & + b2
\end{align*}
\]

\( \psi_{\text{save}} \) and \( \psi_{\text{totalBalance}} \) 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; \( \$() \) is meta-programming syntax for splicing the result into the program. The \text{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 since it is unavailable and hinders scalability [2]. As we will see, these transactions can be satisfied with much weaker isolation guarantees. Despite the atomicity offered by the transaction, anomalies are still possible. For example, the two getBalance operations in \( \text{totalBalance} \) transactions might be served by different replicas with distinct set of committed \( \text{save} \) transactions. If the first(second) getBalance operation witness a \( \text{save} \) 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 which weakest isolation guarantee will 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 \( \text{save} \) transactions, we define the constraint on \( \text{totalBalance} \) transaction \( \psi_{tB} \) as:

\[ \psi_{tB} = \forall a : \text{getBalance}, b : \text{getBalance}. \]

\[ (c : \text{withdraw } \lor \text{deposit}), (d : \text{withdraw } \lor \text{deposit}) \]

\[ \text{txn}(a, b)\{c, d\} \land \text{vis}(c, a) \land \text{sameobj}(b, d) \Rightarrow \text{vis}(d, b) \]

The key idea in the above definition is that the \( \text{txn} \) primitive allows us to relate operations on different objects.

The \text{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, \text{save} does not require any additional constraints, and \( \psi_1 \) is simply true.

5.3 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 ensure 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:

\[ \psi_{\text{rc}} = \forall a, b, c. \text{txn}(a)\{b, c\} \land \text{sameobj}(b, c) \land \text{vis}(b, a) \Rightarrow \text{vis}(c, a) \]

Notice that the above definition is the same as the atomicity guarantee of transaction described in § 5.1. The \text{save} is an example for RC transaction.

MAV semantics ensures that if some operation in a transaction \( T_1 \) witnesses the effects of another transaction \( T_2 \), then subsequent
operations in $T_2$ 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:

$$\psi_{mv} = \forall a, b, c, d. \exists \text{txn}(a, b) \{c, d\} \land \text{so}(a, b) \land \text{vis}(c, a) \land \text{sameobj}(d, b) \Rightarrow \text{vis}(d, b)$$

whose semantics is illustrated in the Figure 8(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 coordination-free. An example for such a transaction is the totalBalance transaction. The semantics of RR is captured by the following contract:

$$\psi_r = \forall a, b, c, d. \exists \text{txn}(a, b) \{c, d\} \land \text{vis}(c, a) \land \text{sameobj}(d, b) \Rightarrow \text{vis}(d, b)$$

whose semantics is illustrated in the Figure 8(c).

5.4 Classification

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

6. 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 9 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 an off-the-shelf data store enables an implementation comprising roughly only 2500 lines of Haskell code, which is packaged as a library.

6.1 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 availability, session and same transaction relations. Dependence tracking is similar to the techniques presented in [13] 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 $(o, e, t_xn, val, deps)$ into the backing store, where $o$ and $e$ are object and unique effect identifiers, $t_xn$ is an optional transaction identifier, and $val$ is the value associated with the effect (e.g., Withdraw 50). $deps$ is the set of identifiers of dependencies of this operation and is defined as $deps(e) = \{e_1 \mid \text{vis}(e_1, e) \land \neg (\exists e_2. \text{vis}(e_1, e_2) \land \text{vis}(e_2, e))\}$. At any shim layer node, an effect is included only if all its dependencies are also included in that node. This ensures that the state maintained by the shim layer node is causally consistent. 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.

6.2 Transactions

Cassandra does not provide general-purpose transactions. Since the transaction guarantees provided by QUELEA are coordination-free, 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 guarantee – an effect is included only if all the effects it depends on are also included. Consider the example given in Figure 10. For every transaction in QUELEA, we instantiate a special transaction marker effect $m$. But importantly, do not insert into the backing store. $m$ is included as a dependence to every effect generated in the transaction. In the figure, the graph on the left shows the state of the store in the middle of a transaction. Each circle represents an effect. The dotted circle indicates that the effect has been instantiated, but has not yet been inserted into the store.
We utilize the \( M \) when the number of effects on an object crosses a tunable threshold. The effects removes the original effects from its cache since they are not yet in the backing store, making all the effects from the transactions as a dependence for \( M \). 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 the causality requirement 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 a 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.

### 6.3 Summarization

We utilize the `summarize` function (§3.1.4) 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 11 illustrates the summarization strategy. Suppose the original set of effects on an object are \( o_1, o_2 \) and \( o_3 \). When summarized, the new effects yielded are \( n_1 \) and \( n_2 \). 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 \( n_1 \) and \( n_2 \), 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 insert \( s \) with strong consistency, including the original effects \( o_1, o_2 \) and \( o_3 \) as dependence. Strongly consistent inserts ensure that a shim layer node witnessing \( s \) on some object must also witness \( n_1 \) and \( n_2 \) 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 arises to be atomic from the clients perspective.

---

**Figure 11:** Summarization in the backing store. Dotted circle represents effects not yet inserted into the backing store.

Since the causally preceding effect \( m \) has not yet been written to the store, no operation will witness \( e_1 \) and \( e_2 \) while the transaction 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 \( m \). 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.

### 7. 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, username, 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 compile time and has no overheads at runtime. 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 nodes.

### 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.

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>LOC</th>
<th>#T</th>
<th>EC</th>
<th>CC</th>
<th>SC</th>
<th>RC</th>
<th>MAV</th>
<th>RR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LWW Reg</td>
<td>156</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DynamoDB</td>
<td>126</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bank Account</td>
<td>155</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Shopping List</td>
<td>140</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Online store</td>
<td>340</td>
<td>4</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RUBIS</td>
<td>640</td>
<td>6</td>
<td>14</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Microblog</td>
<td>659</td>
<td>5</td>
<td>13</td>
<td>6</td>
<td>1</td>
<td>6</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>
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 an e3.4xlarge 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 12(a) 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 9.4x due to the cost of geo-distributed coordination, whereas QUELEA operations were only 2.2x 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 12(b) 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 2.3x 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, 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, supporting our thesis that programmers only have to incur relatively low overhead for a more expressive programming model which provides stronger provable consistency guarantees.

Figure 12(c) compares the QUELEA implementation of RUBiS in a 1DC configuration against a single replica (NoRep) and a strongly replicated (StrongRep) 1DC deployment. The benchmark uses the default 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. The 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 12(d). 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.

8. 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 RDTs stronger consistency models 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 on 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, 25] 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 expects contracts expressing declarative visibility requirements.

Our contract language and system model is inspired by the axiomatic description of RDT semantics proposed by [10]. While they use axioms for formal verification of correctness of an RDT implementation, we utilize them as a means for the user to express the desired consistency guarantees in the application. Operational semantics of QUELEA (described in the accompanying tech report [2]) describe 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 differ in the way they aim to empower the application programmer. In [11], they expose their operational model to the application programmer, along with primitives, such as push, pull and fetch, that make it easy to implement the required consistency and isolation guarantees. In contrast, QUELEA completely abstracts its operational model, and instead relies on contract classification to 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 [25].

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 [13] define invariant confluence and invariant safety, respectively, as conditions for safely executing an operation without coordination. [13] 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 operation, if programmer chooses invariant-violation avoidance over violation repair, the system employs various techniques, such as escrow reservation, to ensure that the offending set of operations are effectively serialized. Unlike QUELEA, these works focus on a coarse-grained classification of consistency as eventual or strong, and do not focus on finer-grained transaction isolation levels. However, the analyses that some of them propose relieve programmers of the burden to tag operations with consistency levels. Indeed, we do consider automatic inference of consistency contracts from application-specific integrity constraints as the next step for QUELEA.

9. 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 to the 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 off-the-shelf distributed store, Cassandra, and illustrate the benefit of fine-grained contract classification by implementing and evaluating several scalable applications.

References


