Technical Report

Making Geo-Replicated Systems Fast as Possible, Consistent when Necessary

Cheng Li[†], *Daniel Porto*[*†], *Allen Clement*[†], *Johannes Gehrke*[‡], *Nuno Preguiça*[*], *Rodrigo Rodrigues*[*]
[†]*Max Planck Institute for Software Systems (MPI-SWS),* [*]*CITI / DI-FCT-Universidade Nova de Lisboa,* [‡]*Cornell University*

Abstract

Online services distribute and replicate state across geographically diverse data centers and direct user requests to the closest or least loaded site. While effectively ensuring low latency responses, this approach is at odds with maintaining cross-site consistency. We make three contributions to address this tension. First, we propose RedBlue consistency, which enables blue operations to be fast (and eventually consistent) while the remaining red operations are strongly consistent (and slow). Second, to make use of fast operation whenever possible and only resort to strong consistency when needed, we identify conditions delineating when operations can be blue and must be red. Third, we introduce a method that increases the space of potential blue operations by breaking them into separate generator and shadow phases. We built a coordination infrastructure called Gemini that offers RedBlue consistency, and we report on our experience modifying the TPC-W and RUBiS benchmarks and an online social network to use Gemini. Our experimental results show that RedBlue consistency provides substantial performance gains without sacrificing consistency.

1 Introduction

Scaling services over the Internet to meet the needs of an ever-growing user base is challenging. In order to improve user-perceived latency, which directly affects the quality of the user experience, services replicate system state across geographically diverse sites and direct users to the closest or least loaded site.

To avoid paying the performance penalty of synchronizing concurrent actions across data centers, some systems, such as Amazon's Dynamo [9], resort to weaker consistency semantics like eventual consistency where state can temporarily diverge. Others, such as Yahoo!'s PNUTS [8], avoid state divergence due to the undesirable sets of behaviors it allows, by requiring all operations that update the service state to be funneled through a primary site and thus incurring increased latency.

This paper addresses the inherent tension between performance and meaningful consistency. A first step towards addressing this tension is to allow multiple levels of consistency to coexist [19,33,34]: some operations can be executed optimistically, without synchronizing with concurrent actions at other sites, while others require a stronger consistency level and thus require cross-site synchronization. However, this places a high burden on the developer of the service, who must decide which operations to assign which consistency levels. This requires reasoning about the consistency semantics of the overall system to ensure that the behaviors that are allowed by the different consistency levels satisfy the specification of the system.

In this paper we propose a comprehensive and principled approach to this problem, aiming at enabling geo-replicated systems to be as fast as possible, while ensuring that they are consistent when necessary. We make the following three contributions:

1. We propose a novel consistency definition called RedBlue consistency. The intuition behind RedBlue consistency is that blue operations execute locally and are lazily replicated in an eventually consistent manner [9, 12, 24, 25, 32, 33, 37]. Red operations, in contrast, are serialized with respect to each other and require immediate cross-

Consistency level	Example systems	Immediate response	State convergence	Single value	General operations	Stable histories	Classification strategy
Strong	RSM [20, 30]	no	yes	yes	yes	yes	N/A
Timeline/snapshot	PNUTS [8], Megastore [3]	reads only	yes	yes	yes	yes	N/A
Fork	SUNDR [23]	all ops	no	yes	yes	yes	N/A
Eventual	Bayou [37], Depot [25]	all ops	yes	no	yes	yes	N/A
	Sporc [12], CRDT [32]	all ops	yes	yes	no	yes	N/A
	Zeno [33], COPS [24]	weak/all ops	yes	yes	yes	no	no / N/A
Multi	PSI [34]	cset	yes	yes	partial	yes	no
	lazy repl. [19], Horus [38]	immed./causal ops	yes	yes	yes	yes	no
RedBlue	Gemini	Red ops	yes	yes	yes	yes	yes

Table 1: Tradeoffs in geo-replicated systems and various consistency levels.

site coordination. RedBlue consistency preserves causality by ensuring that dependencies established when an operation is invoked at its primary site are preserved as the operation is incorporated at other sites.

2. We identify the conditions under which operations must be colored red and may be colored blue in order to ensure that application invariants are never violated and that all replicas converge on the same final state. Intuitively, operations that commute with all other operations and do not impact invariants may be blue.
3. We observe that the commutativity requirement limits the space of potentially blue operations. To address this, we decompose operations into two components: (1) a generator operation that identifies the changes the original operation should make, but has no side effects itself, and (2) a shadow operation that performs the identified changes and is replicated to all sites. Only shadow operations are colored red or blue. This allows for a fine-grained classification of operations and broadens the space of potentially blue operations.

We built a system called Gemini that coordinates RedBlue replication, and use it to extend three applications to be RedBlue consistent: the TPC-W and RUBiS benchmarks and the Quoddy social network. Our evaluation using microbenchmarks and the three applications shows that RedBlue consistency provides substantial latency and throughput benefits. Furthermore, our experience with modifying these applications indicates that shadow operations can be created with modest effort.

The rest of the paper is organized as follows: we position our work in comparison to existing proposals in §2. We define RedBlue consistency and introduce shadow operations along with a set of principles of how to use them in §4 and §5. We describe our prototype system in §6, and report on the experience transitioning three application benchmarks to be RedBlue consistent in §7. We analyze experimental results in §8 and conclude in §9.

2 Background and related work

Target end-to-end properties. To frame the discussion of existing systems that may be used for geo-replication, we start by informally stating some desirable properties that such solutions should support. The first property consists of ensuring a good user experience by providing **low latency** access to the service [31]. Providing low latency access implies that operations should proceed after contacting a small number of replicas, but this is at odds with other requirements that are often sacrificed by consistency models that privilege low latency. The first such requirement is preserving **causality**, both in terms of the monotonicity of user requests within a session and preserving causality across clients, which is key to enabling natural semantics [27]. Second, it is important for all operations executed at one replica to be propagated to all remaining replicas, a property we call **eventual propagation**. Third, it is important that all replicas that have executed the same set of operations are in the same state, i.e., that they exhibit **state convergence**. Fourth, we also want to avoid marked deviations from the conventional, single server semantics. In particular, operations should return a **single value**, precluding solutions that return a set of values corresponding to the outcome of multiple concurrent updates; the system should provide a set of **stable histories**, meaning that user actions cannot be undone; and it should provide support for **general operations**, not restricting the type of operations that can be executed. Finally, the behavior of the service must obey a service-dependent specification, which may be defined as a set of **invariants** that must be preserved.

Table 1 summarizes several proposals of consistency definitions, which strike different balances between the requirements mentioned above. While other consistency definitions exist, we focus on the ones most closely related to the

problem of offering fast and consistent responses in geo-replicated systems.

Strong vs. weak consistency. On the strong consistency side of the spectrum there are definitions like linearizability [17], where the replicated system behaves like a single server that serializes all operations. This, however, requires coordination among replicas to agree on the order in which operations are executed, with the corresponding overheads that are amplified in geo-replication scenarios. Somewhat more efficient are timeline consistency in PNUTS [8] and snapshot consistency in Megastore [3]. These systems ensure that there is a total order for updates to the service state, but give the option of reading a consistent but dated view of the service. Similarly, Facebook has a primary site that handles updates and a secondary site that acts as a read-only copy [22]. This allows for fast reads executed at the closest site but writes still pay a penalty for serialization. Fork consistency [23, 26] addresses the performance limitations of strong consistency by allowing users to observe distinct causal histories. The primary drawback of fork consistency is that once replicas have forked, they can never be reconciled. Such approach is useful when building secure systems but is not appropriate in the context of geo-replicating a single service.

Eventual consistency [37] is on the other end of the spectrum. Eventual consistency is a catch-all phrase that covers any system where replicas may diverge in the short term as long as the divergence is eventually repaired and may or may not include causality. (See Saito and Shapiro [29] for a survey.) In practice, as shown in Table 1, systems that embrace eventual consistency have limitations. Some systems waive the stable history property, either by rolling back operations and re-executing them in a different order at some of the replicas [33], or by resorting to a last writer wins strategy, which often results in loss of one of the concurrent updates [24]. Other systems expose multiple values from divergent branches in operations replies either directly to the client [9, 25] or to an application-specific conflict resolution procedure [37]. Finally, some systems restrict operations by assuming that all operations in the system commute [12, 32], which might require the programmer to rewrite or avoid using some operations.

Coexistence of multiple consistency levels. The solution we propose for addressing the tension between low latency and strongly consistent responses is to allow different operations to run with different consistency levels. Existing systems that used a similar approach include Horus [38], lazy replication [19], Zeno [33], and PSI [34]. However, none of these proposals guide the service developer in choosing between the available consistency levels. In particular, developers must reason about whether their choice leads to the desired service behavior, namely by ensuring that invariants are preserved and that replica state does not diverge. This can be challenging due to difficulties in identifying behaviors allowed by a specific consistency level and understanding the interplay between operations running at different levels. Our research addresses this challenge, namely by defining a set of conditions that precisely determine the appropriate consistency level for each operation.

Other related work. Consistency rationing [18] allows consistency guarantees to be associated with data instead of operations, and the consistency level to be automatically switched at runtime between weak consistency and serializability based on specified policies. TACT [40] consistency bounds the amount of inconsistency of data items in an application-specific manner, using the following metrics: numerical error, order error and staleness. In contrast to these models, the focus of our work is not on adapting the consistency levels of particular data items at runtime, but instead on systematically partitioning the space of operations according to their actions and the desired system semantics.

One of the central aspects of our work is the notion of shadow operations, which increase operation commutativity by decoupling the decision of the side effects from their application to the state. This enables applications to make more use of fast operations. Some prior work also aims at increasing operation commutativity: Weihl exploited commutativity-based concurrency control for abstract data types [39]; operational transformation [10, 12] extends non-commutative operations with a transformation that makes them commute; Conflict-free Replicated Data Types (CRDTs) [32] design operations that commute by construction; Gray [15] proposed an open nested transaction model that uses commutative compensating transactions to revert the effects of aborted transactions without rolling back the transactions that have seen their results and already committed; delta transactions [35] divide a transaction into smaller pieces that commute with each other to reduce the serializability requirements. Our proposal of shadow operations can be seen as an extension to these concepts, providing a different way of broadening the scope of potentially commutative operations. There exist other proposals that also decouple the execution into two parts, namely two-tier replication [16] and CRDT downstreams [32]. In contrast to these proposals, for each operation, we may generate different shadow operations based on the specifics of the execution. Also, shadow operations can run under different

consistency levels, which is important because commutativity is not always sufficient to ensure safe weakly consistent operation.

3 System model

We assume a distributed system with state fully replicated across k sites denoted $site_0 \ldots site_{k-1}$. We follow the traditional deterministic state machine model, where there is a set of possible states $\mathscr{S}$ and a set of possible operations $\mathscr{O}$, each replica holds a copy of the current system state, and upon applying an operation each replica deterministically transitions to the next state and possibly outputs a corresponding reply.

In our notation, $S \in \mathscr{S}$ denotes a system state, and $u, v \in \mathscr{O}$ denote operations. We assume there exists an initial state S_0. If operation u is applied against a system state S, it produces another system state S'; we will also denote this by $S' = S + u$. We say that a pair of operations u and v *commute* if $\forall S \in \mathscr{S}, S + u + v = S + v + u$. Given a total order $T(U, <)$ over a set of operations U, if we apply all operations U against a system state S according to $<$, we denote the final state by $S(T)$. $S(T) = S + u_0 + u_1 + ... + u_i + ... + u_{|U|-1}$, where $\forall i, 0 <= i < |U|$. The system maintains a set of application-specific invariants. We define the primitive *valid*(S) to be *true* if state S satisfies all these invariants and *false* otherwise. We say an operation u is *correct* if for any valid state S, $S + u$ is also valid. Each operation u is initially submitted at one site which we call u's *primary site* and denote *site*(u); the system then later replicates u to the other sites.

4 RedBlue consistency

In this section we introduce RedBlue consistency, a novel consistency model that allows replicated systems to be fast as possible and consistent when necessary. "Fast" is an easy concept to understand—it equates to providing low latency responses to user requests. "Consistent" is more nuanced—consistency models technically restrict the state that operations can observe, which can be translated to an order that operations can be applied to a system. Eventual consistency [12, 24, 25, 37], for example, permits operations to be partially ordered and enables fast systems—sites can process requests locally without coordinating with each other—but sacrifices the intuitive semantics of serializing updates. In contrast, linearizability [17] or serializability [5] provide strong consistency and allow for systems with intuitive semantics— in effect, all sites process operations in the same order —but require significant coordination between sites, precluding fast operation.

RedBlue consistency is designed to allow systems to support fast eventually consistent execution when possible and (slower) strongly consistent execution when necessary. It is based on an explicit division of operations into blue operations whose order of execution can vary from site to site, and red operations that must be executed in the same order at all sites.

4.1 Defining RedBlue consistency

The definition of RedBlue consistency has two components: (1) A RedBlue order, which defines a partial order of operations, and (2) a set of local causal serializations, which define site-specific total orders in which the operations are locally applied.

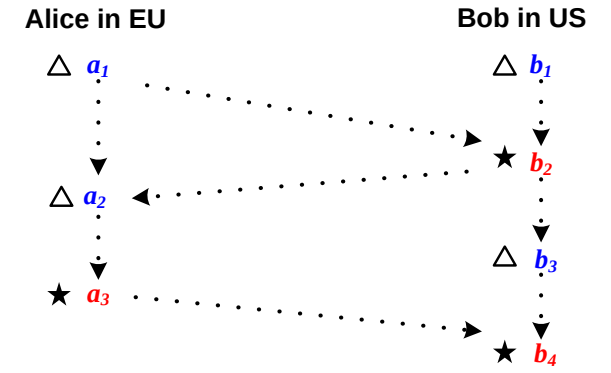
Definition 1 (RedBlue order) *Given a set of operations $U = R \cup B$, where $R \cap B = \emptyset$, a* RedBlue order *is a partial order $O = (U, \prec)$ with the restriction that $\forall u, v \in R$ such that $u \neq v$, $u \prec v$ or $v \prec u$ (i.e., red operations are totally ordered).*

Recall that each site is modeled as a deterministic state machine capable of processing a totally ordered sequence of operations. We define which serializations are allowed for a given RedBlue order as follows:

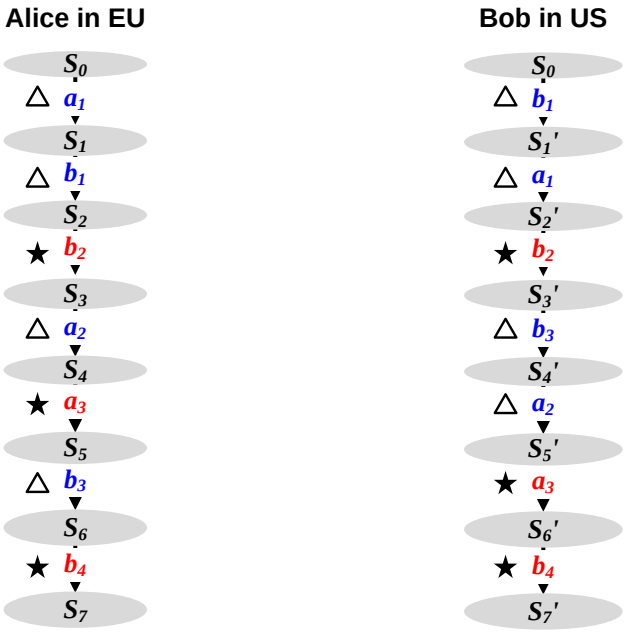
Definition 2 (Legal serialization) *$O' = (U, <)$ is a* legal serialization *of RedBlue order $O = (U, \prec)$ if*

- *O' is a linear extension of O; i.e., $<$ is a total order compatible with the partial order defined by $\prec$.*

This definition forces the serial order by which replicas execute operations to be compatible with the RedBlue order. However, it fails to enforce causality, meaning that if an operation v sees the effects of operation u at its primary site,



(a) RedBlue order O of operations



(b) Causal serializations of O

Figure 1: RedBlue order and causal serializations for a system spanning two sites. Operations marked with ★ are red; operations marked with △ are blue. Dotted arrows in (a) indicate dependencies between operations.

then any operation w that sees the effects of v must also see the effect of u at all sites in the system. Causality is an important component of a good user experience [36]. Thus we extend the above definition by saying that if operation v sees the effects of u at its primary site, $site(v)$, then u must be serialized before v at all sites.

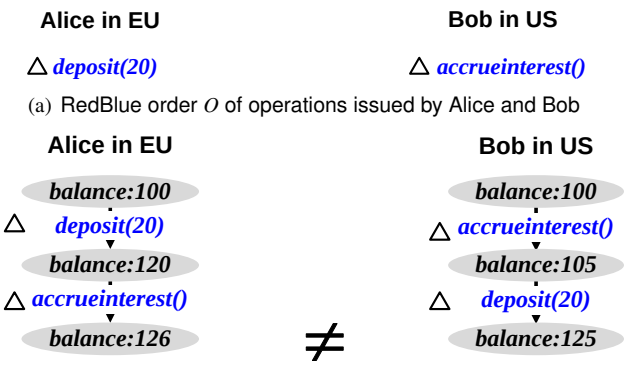
Definition 3 (Causal legal serialization) *Given a site i, $O_i = (U, <)$ is an* i-causal legal serialization *(or short, a* causal serialization*) of RedBlue order $O = (U, \prec)$ if (a) O_i is a legal serialization of O, and (b) for any two operations $u, v \in U$, if $site(v) = i$ and $u < v$ in O_i, then $u \prec v$.*

A replicated system with k sites is then RedBlue consistent if every site applies a causal serialization of the same global RedBlue order O.

Definition 4 (RedBlue consistency) *A replicated system is* O-RedBlue consistent *(or short, RedBlue consistent) if each site i applies operations according to an i-causal serialization of RedBlue order O.*

```
1  float balance, interest = 0.05;
2  func deposit( float money ):
3    balance = balance + money;
4  func withdraw ( float money ):
5    if ( balance - money >= 0 ) then:
6      balance = balance - money;
7    else print "failure";
8  func accrueinterest():
9    float delta = balance × interest;
10   balance = balance + delta;
```

Figure 2: Pseudocode for the bank example.



(a) RedBlue order O of operations issued by Alice and Bob

(b) Causal serializations of O leading to diverged state

Figure 3: A RedBlue consistent account with initial balance of $100.

Figure 1 shows a RedBlue order and a pair of causal serializations of that RedBlue order. In systems where every operation is labeled red, RedBlue consistency is equivalent to serializability [5]; in systems where every operation is labeled blue, RedBlue consistency allows the same set of behaviors as eventual consistency [24,25,37]. It is important to note that while RedBlue consistency constrains possible orderings of operations at each site and thus the states the system can reach, it does not ensure *a priori* that the system achieves all the end-to-end properties identified in §2, in particular, state convergence and invariant preservation, as discussed next.

4.2 State convergence and a RedBlue bank

In order to understand RedBlue consistency it is instructive to look at a concrete example. For this example, consider a simple bank with two users: Alice in the EU and Bob in the US. Alice and Bob share a single bank account where they can deposit or withdraw funds and where a local bank branch can accrue interest on the account (pseudocode for the operations can be found in Figure 2). Let the `deposit` and `accrueinterest` operations be blue. Figure 3 shows a RedBlue order of deposits and interest accruals made by Alice and Bob and causal serializations applied at both branches of the bank.

State convergence is important for replicated systems. Intuitively a pair of replicas is state convergent if, after processing the same set of operations, they are in the same state. In the context of RedBlue consistency we formalize state convergence as follows:

Definition 5 (State convergence) *A RedBlue consistent system is state convergent if all causal serializations of the underlying RedBlue order O reach the same state S.*

The bank example as described is not state convergent. The root cause is not surprising: RedBlue consistency allows sites to execute blue operations in different orders but two blue operations in the example correspond to non-commutative operations—addition (`deposit`) and multiplication (`accrueinterest`). A sufficient condition to guarantee state convergence in a RedBlue consistent system is that every blue operation is *globally commutative*, i.e., it commutes with all other operations, blue or red.

Theorem 1 *Given a RedBlue order O, if all blue operations are globally commutative, then any O-RedBlue consistent system is state convergent.*

In order to prove this theorem, we introduce the following lemmas and their proofs:

The first lemma asserts that, given a legal serialization, swapping two adjacent operations in the legal serialization that are not ordered by the RedBlue order results in another legal serialization.

Lemma 1 *Given a legal serialization $O_i = (U, <_i)$ of RedBlue order $O = (U, \prec)$ with operations $u, v \in U$ such that $u <_i v$ and $u \not\prec v$ and there exists no s such that $u <_i s <_i v$, and let $P = \{p | p \in U \wedge p <_i u\}$ and $Q = \{q | q \in U \wedge v <_i q\}$. The serialization $O_k = (U, <_k)$ where*

- $\forall p, q \in P \cup Q : p <_k q \iff p <_i q$,
- $\forall p \in P : p <_k v$,
- $v <_k u$,
- $\forall q \in Q : u <_k q$

is a legal serialization.

Proof: It suffices to show that $\forall r, s \in U : r <_k s$ is compatible with $\prec$.

Case 1: $r, s \in P \cup Q$. Since O_i is a legal serialization, each $r <_i s$ is compatible with $\prec$ by definition. By construction $\forall p, q \in P \cup Q : r <_k s \iff r <_i s$, so each $r <_k s$ is also compatible with $\prec$.

Case 2: $r \in P$, $s = v$. $r <_k s$ is compatible with $\prec$ by similar logic as above.

Case 3: $r = u$, $s \in Q$. $r <_k s$ is compatible with $\prec$ by similar logic as above.

Case 4: $v <_k u$. Since $u \not\prec v$, $v <_k u$ is compatible with $\prec$.

Case 5: $r \in P$, $s = u$. Since $v <_k u \wedge \forall p \in P : p <_k v \implies p <_k u$. By the construction of P, $\forall p \in P : p <_k u \iff p <_i u$. So each $r <_k s$ is also compatible with $\prec$.

Case 6: $r = v$, $s \in Q$. Since $v <_k u \wedge \forall q \in Q : v <_k q \implies v <_k q$. $r <_k s$ is compatible with $\prec$ by similar logic as above.

As $U = P \cup Q \cup \{u, v\}$, by all above cases, $\forall r, s \in U : r <_k s$ is compatible with $\prec$. ∎

The following lemma asserts that given a RedBlue order and its legal serialization, if there exists a pair of elements that are not ordered by the RedBlue order, then there exists an adjacent pair of elements between u and v in the legal serialization that are not ordered by the RedBlue order.

Lemma 2 *Given a legal serialization $O_i = (U, <_i)$ of RedBlue order $O = (U, \prec)$, if $\exists u, v \in U$ such that $u <_i v$ and $u \not\prec v$, let $U' = \{u, v\} \cup \{q | u <_i q \wedge q <_i v\}$, then $\exists r, s \in U'$ such that $r <_i s \wedge r \not\prec s \wedge \nexists p \in U' : r <_i p \wedge p <_i s$.*

Proof: We prove this by performing the following exhaustive analysis. The analysis terminates when the required pair of elements is found.

Let's start with u, v. Consider Q to be the sequence of elements strictly between u and v, i.e., $Q = \{q \in U | u <_i q \wedge q <_i v\}$. There are two cases we have to analyze:

Case 1: Q is empty. This implies that u and v are adjacent, so the analysis terminates.

Case 2: Q is not empty. This implies that u and v are not adjacent. Consider p to be the first element in U' according to $<_i$, i.e., $p \in U' : \forall q \in U' \setminus \{p\}, p <_i q$. There are two cases to consider:

Case 2a: $u \not\prec p$. It follows that p is the successor of u in O_i, then u, p is the adjacent pair that are not ordered by O. The analysis terminates.

Case 2b: $u \prec p$. It follows from the assertion that $u \not\prec v$ and the transitivity of $\prec$ that $p \not\prec v$. Then we run the analysis from the beginning with p, v. Since we are removing the first element of the sequence Q, the analysis will either eventually terminate with an empty sequence, or before that. ∎

The following lemma asserts that two legal serializations that differ in the order of exactly one pair of adjacent operations (one of which is blue) are state convergent.

Lemma 3 *Assume $O_i = (U, <_i)$ and $O_j = (U, <_j)$ are both legal serializations of RedBlue order $O = (U, \prec)$ that are identical except for two adjacent operations u and v such that $u <_i v$ and $v <_j u$ and that all operations $r \in B$ are globally commutative. Then $S(O_i) = S(O_j)$.*

Proof: Let P and Q be the greatest common prefix and suffix respectively of O_i and O_j. Further, let S be an initial state, let $S_P = S(P)$, $S_{uv} = S_P + u + v$, and $S_{vu} = S_P + v + u$.

It follows from the definition of a RedBlue order and a legal serialization that either $u \in B$ or $v \in B$. Without loss of generality, assume $u \in B$.

By assumption u commutes with all operations in U, therefore $S_{uv} = S_{vu}$. It then follows the definition of deterministic state machine that $S_{uv}(Q) = S_{vu}(Q)$. By the definition of legal serialization 2, the final state reached by sequentially executing operations in O_i against S according to $<$ is equal to the final state obtained by sequentially applying operations in Q against S_{uv} according to $<$, namely $S(O_i) = S_{uv}(Q)$. As the similar logic, we know $S(O_j) = S_{vu}(Q)$. Finally, we have $S(O_i) = S(O_j)$. ∎

With the above lemmas, we could prove Theorem 1 as follows:

Proof: To prove a RedBlue consistent system is state convergent, it is sufficient to show that any pair of legal serializations of their underlying RedBlue order O is state convergent. Let O_i and O_j be two legal serializations of O. There are two cases to consider:

Case 1: $O_i = O_j$. The underlying deterministic state machine ensures that $S(O_i) = S(O_j)$.

Case 2: $O_i \neq O_j$, in which case $\exists u, v \in U$ such that $u <_i v$ and $v <_j u$. Since both O_i and O_j are legal serializations of O, it follows that $u \not\prec v$ and $v \not\prec u$. It then follows Lemma 2 we can find an adjacent pair of operations r, s such that $r <_i s \wedge s <_j r \wedge r \not\prec s \wedge s \not\prec r$. We construct a new serialization O_{i+1} by duplicating O_i but swapping the order of r and s in O_{i+1}, i.e., $r <_i s \wedge s <_{i+1} r$. By Lemma 1, O_{i+1} is also a legal serialization of O.

If $O_{i+1} \neq O_j$, we continue the construction by finding an adjacent pair of elements whose order is different in O_{i+1}, O_j. By swapping the two operations, we obtain another legal serialization O_{i+2}. We can then continue to swap all such adjacent pairs until the last constructed serialization is equal to O_j. This is achievable since the number of pairs in the new serialization O'' whose orders are different in O_j becomes smaller than the number observed in the serialization O' that O'' was constructed from. At the end, the construction process results in a chain of legal serializations where the first one is O_i and the last is O_j, and any consecutive pair of legal serializations is identical except for the order of an adjacent pair of elements. It then follows Lemma 3 that every consecutive pair of serializations in the chain is state convergent. ∎

Theorem 1 highlights an important tension inherent to RedBlue consistency. On the one hand, low latency requires an abundance of blue operations. On the other hand, state convergence requires that blue operations commute with all other operations, blue or red. In the next section we introduce a method for addressing this tension by increasing commutativity.

5 Replicating side effects

In this section, we observe that while operations themselves may not be commutative, *we can often make the changes they induce on the system state commute.* Let us illustrate this issue within the context of the RedBlue bank from §4.2. We can make the `deposit` and `accrueinterest` operations commute by first computing the amount of interested accrued and then treating that value as a deposit.

5.1 Defining shadow operations

The key idea is to split each original application operation u into two components: a *generator operation* g_u with no side-effects, which is executed only at the primary site against some system state S and produces a *shadow operation* $h_u(S)$, which is executed at every site (including the primary site). The generator operation decides which state transitions should be made while the shadow operation applies the transitions in a state-independent manner.

The implementation of generator and shadow operations must obey some basic correctness requirements. Generator operations, as mentioned, must not have any side effects. Furthermore, shadow operations must produce the same

```
1  func deposit' ( float money ):
2    balance = balance + money;
3  func withdrawAck' ( float money ):
4    balance = balance - money;
5  func withdrawFail' ():
6    /* no-op */
7  func accrueinterest' ( float delta ):
8    balance = balance + delta;
```

Figure 4: Pseudocode for shadow bank operations.

effects as the corresponding original operation when executed against the original state S used as an argument in the creation of the shadow operation.

Definition 6 (Correct generator / shadow operations) *The decomposition of operation u into generator and shadow operations is correct if for all states S, the generator operation g_u has no effect and the generated shadow operation $h_u(S)$ has the same effect as u, i.e., for any state S: $S + g_u = S$ and $S + h_u(S) = S + u$.*

Note that a trivial decomposition of an original operation u into generator and shadow operations is to let g_u be a no-op and let $h_u(S) = u$ for all S.

In practice, as exemplified in §7, separating the decision of which transition to make from the act of applying the transition allows many objects and their associated usage in shadow operations to form an abelian group and thus dramatically increase the number of commutative (i.e., blue) operations in the system. Unlike previous approaches [16, 32], for a given original operation, our solution allows its generator operation to generate state-specific shadow operations with different properties, which can then be assigned different colors in the RedBlue consistency model.

5.2 Revisiting RedBlue consistency

Decomposing operations into generator and shadow components requires us to revisit the foundations of RedBlue consistency. In particular, only shadow operations are included in a RedBlue order while the causal serialization for site i additionally includes the generator operations initially executed at site i. The causal serialization must ensure that generator operations see the same state that is associated with the generated shadow operation and that shadow operations appropriately inherit all dependencies from their generator operation.

We capture these subtleties in the following revised definition of causal serializations. Let U be the set of shadow operations executed by the system and V_i be the generator operations executed at site i. Note that the definitions of legal serialization and RedBlue order remain fundamentally unchanged, once "operation" is replaced with "shadow operation."

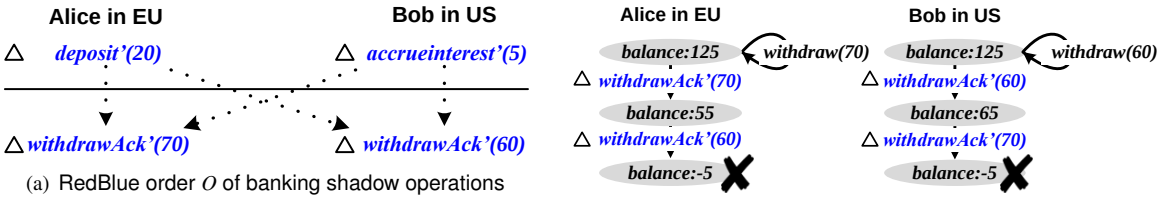
Definition 7 (Causal serialization–revised) *Given a site i, $O_i = (U \cup V_i, <)$ is an i-causal serialization of RedBlue order $O = (U, \prec)$ if*

- *O_i is a total order;*
- *$(U, <)$ is a linear extension of O;*
- *For any $h_v(S) \in U$ generated by $g_v \in V_i$, S is the state obtained after applying the sequence of shadow operations preceding g_v in O_i;*
- *For any $g_v \in V_i$ and $h_u(S) \in U$, $h_u(S) < g_v$ in O_i iff $h_u(S) \prec h_v(S')$ in O.*

Note that shadow operations appear in every causal serialization, while generator operations appear only in the causal serialization of the initially executing site.

5.3 Shadow banking and invariants

Figure 4 shows the shadow operations for the banking example. Note that the `withdraw` operation maps to two distinct shadow operations that may be labeled as blue or red independently—`withdrawAck'` and `withdrawFail'`.



(a) RedBlue order O of banking shadow operations

(b) Invalid but convergent causal serializations of O

Figure 5: A RedBlue consistent bank with only blue operations. The starting balance of \$125 is the result of applying shadow operations above the solid line to an initial balance of \$100. Loops indicate generator operations.

Figure 5 illustrates that shadow operations make it possible for all operations to commute, provided that we can identify the underlying abelian group. This does not mean, however, that it is safe to label all operations blue.

In this example, such a labeling would allow Alice and Bob to successfully withdraw \$70 and \$60 at their local branches, thus ending up with a final balance of \$-5. This violates the fundamental invariant that a bank balance should never be negative.

To determine which operations can be safely labeled blue, we begin by defining that a shadow operation is invariant safe if, when applied to a valid state, it always transitions the system into another valid state.

Definition 8 (Invariant safe) *Shadow operation $h_u(S)$ is invariant safe if for all valid states S and S', the state $S' + h_u(S)$ is also valid.*

We also assume that the original applications without being RedBlue consistent replicated are correct, i.e., all their original operations always transition from a valid system state to another valid state. This is captured by the following trivial definition:

Definition 9 (Correct original operation) *Original operation t is correct if for all valid states S, $S+t$ is also valid.*

The following theorem states that in a RedBlue consistent system with appropriate labeling, each replica transitions only through valid states.

Theorem 2 *Given a RedBlue consistent system, if all its original operations and shadow operations are correct and all its blue shadow operations are invariant safe and globally commutative, then for any execution of that system that starts from a valid state, no site is ever in an invalid state.*

Proof. Let $O = (U, \prec)$ be a RedBlue order and $U = B \cup R$ such that $\forall u \in U$ is correct and $\forall v \in B$ is invariant safe and globally commutative. All original operations are correct. The initial state S_0 is valid.

Let L be a causal serialization of O, which is shown in Figure 6. Assume that L is in an invalid state. We prove this theorem by performing the following exhaustive analysis and showing the contradictions found.

Analysis: Let $P(U_P, <_P)$ be the shortest prefix of L that produces an invalid state. If P is empty, then $S_0(P) = S_0$, and L is in a valid state. This violates the assumption that L is in an invalid state. The theorem is proved.

If P is non-empty, then consider u to be the last shadow operation in P such that $P = P' + u$. Let t be the original operation of u. By the definition of shadow operation, we know $u = h_t(S')$, where S' is the state in which u was generated.

Case 1: u is blue. As the assertion that u is invariant safe, the state reached by P', $S_0(P')$, must be invalid as well. This contradicts the assumption that P is the shortest prefix that introduces an invalid state. The theorem is proved.

Case 2: u is red. S' has two possible values.

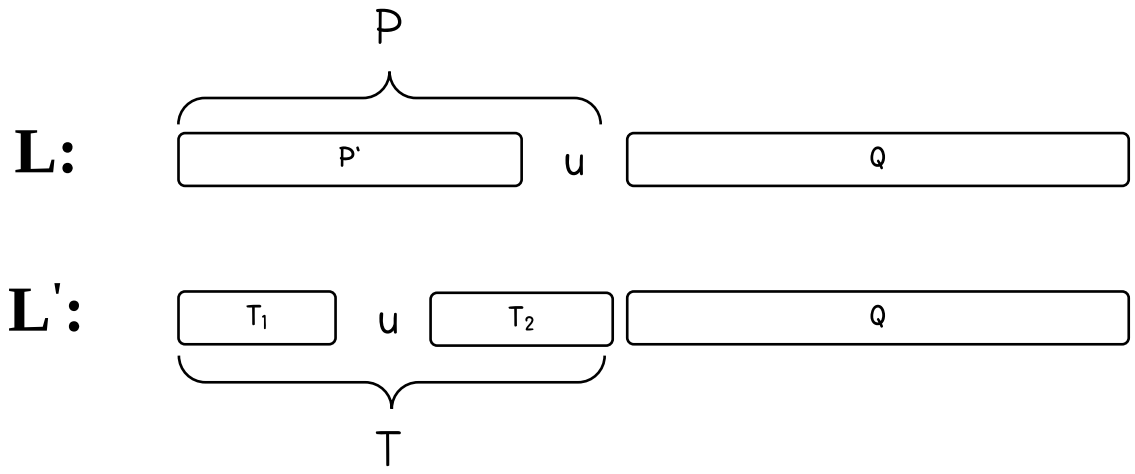


Figure 6: Two legal serializations L and L'

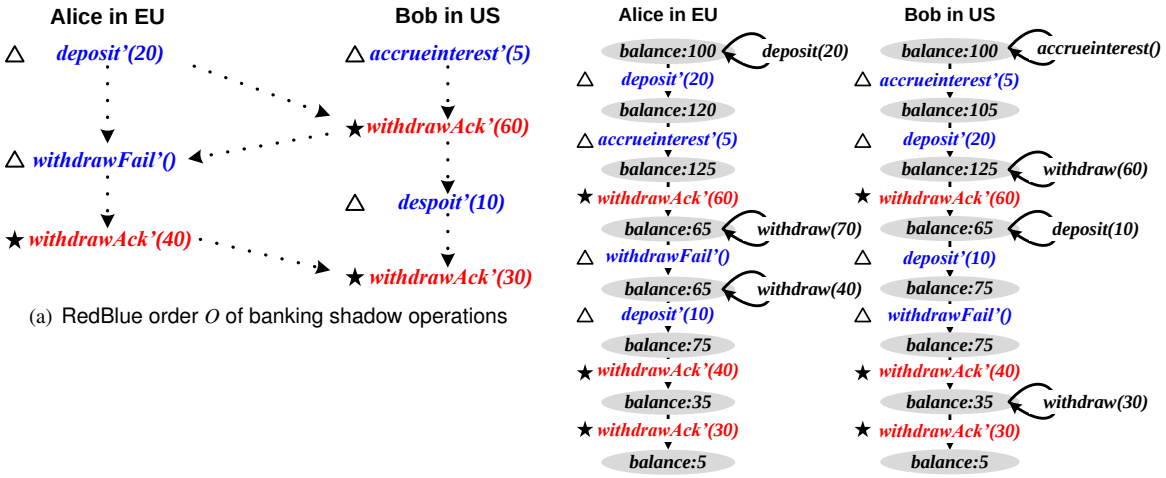
- **Case 2a:** $S' = S_0(P')$, i.e., the state that u was applied against is the same as the state that u was created from. By Definition 6, u is a correct shadow operation, so $S' + u = S' + t$. $S' + t$ is invalid. It follows the Definition 9 that t is a correct original operation. Thus, S' must be invalid as well. By the same logic in Case 1, we could find a shorter prefix P' other than P that produces an invalid state. By contradiction, the theorem is proved.
- **Case 2b:** $S' \neq S_0(P')$. It follows the definitions of a RedBlue order and a causal serialization that there exists some blue shadow operations v such that $v <_L u$ and $u \not\prec_O v \wedge v \not\prec_O u$. It then follows Lemmas 1 and 2 that we can construct a causal serialization L' of O by duplicating L and swapping the order between u and every v, so that u is bubbled up over every such v. The result is shown in Figure 6. The only difference between L and L' is as follows: $\forall i \in U_p : i <_L u \wedge i \not\prec_O u \implies u <_{L'} i$. By the state convergence Theorem 1 and the fact that every i is blue and globally commutative, so the prefix T and P are state convergent, i.e., $S_0(P) = S_0(T)$. As $S_0(P)$ is invalid, $S_0(T)$ is also invalid.

  By the deterministic state machine model, we know $S_0(T) = S_0(T_1) + u + T_2$, where as shown in Figure 6 T_1 and T_2 are prefix and suffix of T, respectively. By the construction of L', we know that T_2 is a sequence of blue shadow operations. As all blue shadow operations are invariant safe, we conclude that $S_0(T_1) + u$ must be invalid. By the definition of a causal serialization 7 and the construction of L', we know that $\forall i <'_L u \iff i \prec_O u$. Therefore, the state $S_0(T_1)$ is the state in which u was generated. Then, we have $S_0(T_1) + u = S_0(T_1) + t$. It follows that t is a correct original operation. Thus, $S_0(T_1)$ must be invalid as well. We proceed by starting again the analysis with the input a new causal serialization of O and a new shortest prefix that produces an invalid state, i.e., $P = T_1 \wedge L = L'$. ∎

What can be blue? What must be red? The combination of Theorems 1 and 2 leads to the following procedure for deciding which shadow operations can be blue or must be red if a RedBlue consistent system is to provide both state convergence and invariant preservation:

1. For any pair of non-commutative shadow operations u and v, label both u and v red.
2. For any shadow operation u that may result in an invariant being violated, label u red.
3. Label all non-red shadow operations blue.

Applying this decision process to the bank example leads to a labeling where `withdrawAck'` is red and the remaining shadow operations are blue. Figure 7 shows a RedBlue order with appropriately labeled shadow operations and causal serializations for the two sites.



(a) RedBlue order O of banking shadow operations

(b) Convergent and invariant preserving causal serializations of O

Figure 7: A RedBlue consistent bank with correctly labeled shadow operations and initial balance of $100.

5.4 Discussion

Shadow operations introduce some surprising anomalies to a user experience. Notably, while the effect of every user action is applied at every site, the final system state is not guaranteed to match the state resulting from a serial ordering of the original operations. The important thing to keep in mind is that the decisions made always make sense in the context of the *local* view of the system: when Alice accrues interest in the EU, the amount of interest accrued is based on the balance that Alice observes at that moment. If Bob concurrently makes a deposit in the US and subsequently observes that interest has been accrued, the amount of interest *will not* match the amount that Bob would accrue based on the balance as he currently observes it.

Shadow operations always provide for a coherent sequence of state transitions that reflects the effects demanded by user activity; while this sequence of state transitions is coherent (and convergent), the state transitions are chosen based on the locally observable state when/where the user activity initiated and not the system state when they are applied.

6 Gemini design & implementation

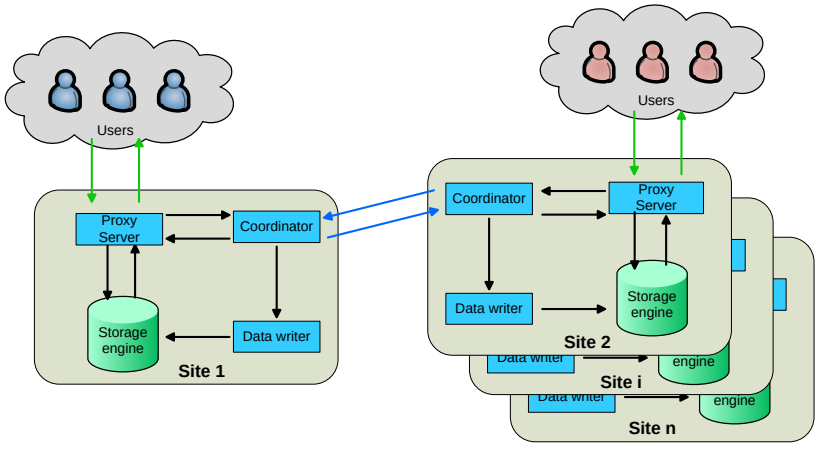


Figure 8: Gemini system architecture.

We implemented the Gemini storage system to provide RedBlue consistency. The prototype consists of 10K lines of java code and uses MySQL as its storage backend. As shown in Figure 8, each Gemini site consists of four components: a storage engine, a proxy server, a concurrency coordinator, and a data writer. A multi-site deployment is constructed by replicating the single data center components across multiple sites.

The basic flow of user requests through the system is straightforward. A user issues requests to a *proxy server* located at the closest site. The proxy server processes a request by executing an appropriate application transaction, which is implemented as a single Gemini operation, comprising multiple data accesses; individual data accesses within a generator operation execute in a temporary private scratchpad, providing a virtual private copy of the service state. The original data lies in a *storage engine*, which provides a standard storage interface. In our implementation, the storage engine is a relational database, and scratchpad operations are executed against a temporary table.Upon completion of the generator operation, the proxy server sends the produced shadow operation on to the *concurrency coordinator* to admit or reject this operation according to RedBlue consistency. The concurrency coordinator notifies the proxy server if the operation is accepted or rejected. Additionally, accepted shadow operations are appended to the end of the local causal serialization and propagated to remote sites and to the local *data writer* for execution against the storage engine. When a shadow operation is rejected, the proxy server re-executes the generator operation and restarts the process.

6.1 Optimistic concurrency control

Gemini relies on optimistic concurrency control (OCC) [5] to run generator operations without blocking.

Gemini uses timestamps to determine if operations can complete successfully. Timestamps are logical clocks [20] of the form $\langle\langle b_0, b_1, \ldots, b_{k-1}\rangle, r\rangle$, where b_i is the local count of shadow operations initially executed by site i and r is the global count of red shadow operations. To ensure that different sites do not choose the same red sequence number (i.e., all red operations are totally-ordered) we use a simple token passing scheme: only the coordinator in possession of a unique red token is allowed to increase the counter r and approve red operations. In the current prototype, a coordinator holds onto the red token for up to 1 second before passing it along.

When a generator operation completes, the coordinator must determine if the operation (a) reads a coherent system snapshot and (b) obeys the ordering constraints of a causal serialization, as described in §5. To do this, the coordinator checks the timestamps of the data items read and written by the completing operation, and compares them to the timestamps associated with operations completing concurrently and the remote shadow operations that were being applied simultaneously at that site.

Upon successful completion of a generator operation the coordinator assigns the corresponding shadow operation a timestamp that is component-wise equal to the latest operation that was incorporated at its site, and increments its blue and, if this shadow operations is red, the red component of the logical timestamp. This timestamp determines the position of the shadow operation in the RedBlue order, with the normal rules that determine that two operations are partially ordered if one is equal to or dominates the other in all components. It also allows sites to know when it is safe to incorporate remote shadow transactions: they must wait until all shadow operations with smaller timestamps have already been incorporated in the local state of the site. When a remote shadow operation is applied at a site, it is assigned a new timestamp that is the entry-wise max of the timestamp assigned to the shadow operation in the initial site and the local timestamps of accessed data objects. This captures dependencies that span local and remote operations.

Read-only shadow operations. As a performance optimization, blue shadow operations can be marked as read-only. Read-only shadow operations receive special treatment from the coordinator: once the generator operation passes the coherence and causality checks, the proxy is notified that the shadow operation has been accepted but the shadow operation is *not* incorporated into the local serialization or global RedBlue order.

6.2 Failure handling

The current Gemini prototype is designed to demonstrate the performance potential of RedBlue consistency in geo-replicated environments and as such is not implemented to tolerate faults of either a local (i.e., within a site) or catastrophic (i.e., of an entire site) nature. Addressing these concerns is orthogonal to the primary contributions of this paper, nonetheless we briefly sketch mechanisms that could be employed to handle faults.

Application	Original					RedBlue consistent extension				
	user requests	transactions			LOC	shadow operations				LOC changed
		total	read-only	update		blue no-op	blue update	red	LOC	
TPC-W	14	20	13	7	9k	13	14	2	2.8k	429
RUBiS	26	16	11	5	9.4k	11	7	2	1k	180
Quoddy	13	15	11	4	15.5k	11	4	0	495	251

Table 2: Original applications and the changes needed to make them RedBlue consistent.

Isolated component failure. The Gemini architecture consists of four main components at each site, each representing a single point of failure. Standard state machine replication techniques [20, 30] can be employed to make each component robust to failures.

Site failure. Our Gemini prototype relies on a simple ring-exchange for coordinating red epochs. To avoid halting the system upon a site failure, a fault tolerant consensus protocol like Paxos [21] can regulate red tokens.

Operation propagation. Gemini relies on each site to propagate its own local operations to all remote sites. A pairwise network outage or failure of a site following the replication of a operation to some but not all of the sites could prevent sites from exchanging operations that depend on the partially replicated operation. This can be addressed using standard techniques for exchanging causal logs [2, 25, 27, 37] or reliable multicast [13].

Cross-session monotonicity. The proxy that each user connects to enforces the monotonicity of user requests within a session [36]. However, a failure of that proxy, or the user connecting to a different site may result in a subset of that user's operations not carrying over. This can be addressed by allowing the user to specify a "last-read" version when starting a new session or requiring the user to cache all relevant requests [25] in order to replay them when connecting to a new site.

7 Case studies

In this section we report on our experience in modifying three existing applications—the TPC-W shopping cart benchmark [7], the RUBiS auction benchmark [11], and the Quoddy social networking application [14]—to work with RedBlue consistency. The two main tasks to fulfill this goal are (1) decomposing the application into generator and shadow operations and (2) labeling the shadow operations appropriately.

Writing generator and shadow operations. Each of the three case study applications executes MySQL database transactions as part of processing user requests, generally one transaction per request. We map these application level transactions to the original operations and they also serve as a starting point for the generator operations. For shadow operations, we turn each execution path in the original operation into a distinct shadow operation; an execution path that does not modify system state is explicitly encoded as a no-op shadow operation. When the shadow operations are in place, the generator operation is augmented to invoke the appropriate shadow operation at each path.

Labeling shadow operations. Table 2 reports the number of transactions in the TPC-W, RUBiS, and Quoddy, the number of blue and red shadow operations we identified using the labeling rules in §5.3, and the application changes measured in lines of code. Note that read-only transactions always map to blue no-op shadow operations. In the rest of this section we expand on the lessons learned from making applications RedBlue consistent.

7.1 TPC-W

TPC-W [7] models an online bookstore. The application server handles 14 different user requests such as browsing, searching, adding products to a shopping cart, or placing an order. Each user request generates between one and four transactions that access state stored across eight different tables. We extend an open source implementation of the benchmark [28] to allow a shopping cart to be shared by multiple users across multiple sessions.

Writing TPC-W generator and shadow operations. Of the twenty TPC-W transactions, thirteen are read-only and admit no-op shadow operations. The remaining seven update transactions translate to one or more shadow operations according to the number of distinct execution paths in the original operation .

```
1  doBuyConfirm(cartId){
2   beginTxn();
3   cart = exec(SELECT * FROM cartTb WHERE cId=cartId);
4   cost = computeCost(cart);
5   orderId = getUniqueId();
6   exec(INSERT INTO orderTb VALUES(orderId, cart.item.id,
        cart.item.qty, cost));
7   item =exec(SELECT * FROM itemTb WHERE id=cart.item.id);
8   if item.stock − cart.item.qty < 10 then:
9      delta = item.stock − cart.item.qty + 21;
10     if delta > 0 then:
11        exec(UPDATE itemTb SET item.stock+ = delta);
12     else rollback();
13  else exec(UPDATE itemTb SET item.stock− = cart.item.qty);
14  exec(DELETE FROM cartContentTb WHERE cId=cartId AND
        id=cart.item.id);
15  commit();}
```

(a) Original transaction that commits changes to database.

```
1  doBuyConfirmGenerator(cartId){
2   sp = getScratchpad();
3   sp.beginTxn();
4   cart = sp.exec(SELECT * FROM cartTb WHERE cId=cartId);
5   cost = computeCost(cart);
6   orderId = getUniqueId();
7   sp.exec(INSERT INTO orderTb VALUES (orderId, cart.item.id,
        cart.item.qty, cost));
8   item = sp.exec(SELECT * FROM itemTb WHERE id=cart.item.id);
9   if item.stock − cart.item.qty < 10 then:
10     delta = item.stock − cart.item.qty + 21;
11     if delta > 0 sp.exec(UPDATE itemTb SET item.stock+ = delta)
           ;
12     else sp.discard(); return;
13  else sp.exec(UPDATE itemTb SET item.stock− = cart.item.qty);
14  sp.exec(DELETE FROM cartTb WHERE cId=cartId AND id=
        cart.item.id);
15  L_TS = getCommitOrder();
16  sp.discard();
17  if replenished return (doBuyConfirmIncre'(orderId, cartId,
        cart.item.id, cart.item.qty, cost, delta, L_TS));
18  else return (doBuyConfirmDecre'(orderId, cartId, cart.item.Id,
        cart.item.qty, cost, L_TS));}
```

(b) Generator operation that manipulates data via a private *scratchpad*.

```
1  doBuyConfirmIncre'(orderId, cartId, itId, qty, cost, delta, L_TS){
2   exec(INSERT INTO orderTb VALUES(orderId, itId, qty, cost,
        L_TS));
3   exec(UPDATE itemTb SET item.stock+ = delta);
4   exec(UPDATE itemTb SET item.l_ts = L_TS WHERE
        item.l_ts < L_TS);
5   exec(UPDATE cartContentTb SET flag = TRUE WHERE
        id = itId AND cid = cartId AND l_ts <= L_TS);}
```

(c) Shadow doBuyConfirmIncre (Blue) that replenishes the stock value.

```
1  doBuyConfirmDecre'(orderId, cartId, itId, qty, cost, L_TS){
2   exec(INSERT INTO orderTb VALUES(orderId, itId, qty, cost,
        L_TS));
3   exec(UPDATE itemTb SET item.stock− = qty);
4   exec(UPDATE itemTb SET item.l_ts = L_TS WHERE
        item.l_ts < L_TS);
5   exec(UPDATE cartContentTb SET flag = TRUE WHERE
        id = itId AND cid = cartId AND l_ts <= L_TS);}
```

(d) Shadow doBuyConfirmDecre (Red) that decrements the stock value.

Figure 9: Pseudocode for the product purchase transaction in TPC-W. For simplicity the pseudocode assumes that the corresponding shopping cart only contains a single item.

We now give an example transaction, `doBuyConfirm`, which completes a user purchase. The pseudocode for the original transaction is shown in Figure 9(a).

The `doBuyConfirm` transaction removes all items from a shopping cart, computes the total cost of the purchase, and updates the stock value for the purchased items. If the stock would drop below a minimum threshold, then the transaction also replenishes the stock. The key challenge in implementing shadow operations for `doBuyConfirm` is that the original transaction does not commute with itself or any transaction that modifies the contents of a shopping cart. Naively treating the original transaction as a shadow operation would force every shadow operation to be red.

Figure 9(b) shows the generator operation of `doBuyConfirm`, and Figures 9(c) and 9(d) depict the corresponding pair of shadow operations: `doBuyConfirmIncre'` and `doBuyConfirmDecre'`. The former shadow operation is generated when the stock falls below the minimum threshold and must be replenished; the latter is generated when the purchase does not drive the stock below the minimum threshold and consequently does not trigger the replenishment path. In both cases, the generator operation is used to determine the number of items purchased and total cost as well the shadow operation that corresponds to the initial execution. At the end of the execution of the generator operation these parameters and the chosen shadow operation are then propagated to other replicas.

Labeling TPC-W shadow operations. For 29 shadow operations in TPC-W, we find that 27 can be blue and only two must be red. To label shadow operations, we identified two key invariants that the system must maintain. First, the number of in-stock items can never fall below zero. Second, the identifiers generated by the system (e.g., for items or shopping carts) must be unique.

The first invariant is easy to maintain by labeling `doBuyConfirmDecre'` (Figure 9(d)) and its close variant `doBuyConfirmAddrDecre'` red. We observe that they are the only shadow operations in the system that de-

crease the stock value, and as such are the only shadow operations that can possibly invalidate the first invariant. Note that the companion shadow operation `doBuyConfirmIncre'` (Figure 9(c)) *increases* the stock level, and can never drive the stock count below zero, so it can be blue.

The second invariant is more subtle. TPC-W generates IDs for objects (e.g., shopping carts, items, etc.) as they are created by the system. These IDs are used as keys for item lookups and consequently must themselves be unique. To preserve this invariant, we have to label many shadow operations red. This problem is well-known in database replication [6] and was circumvented by modifying the ID generation code, so that IDs become a pair $\langle$*appproxy_id*, *seqnumber*$\rangle$, which makes these operations trivially blue.

7.2 RUBiS

RUBiS [11] emulates an online auction website modeled after eBay [1]. RUBiS defines a set of 26 requests that users can issue ranging from selling, browsing for, bidding on, or buying items directly, to consulting a personal profile that lists outstanding auctions and bids. These 26 user requests are backed by a set of 16 transactions that access the storage backend.

Of these 16 transactions, 11 are read-only, and therefore trivially commutative. For the remaining 5 update transactions, we construct shadow operations to make them commute, similarly to TPC-W. Each of these transactions leads to between 1 and 3 shadow operations. The effort to write the shadow operations was nominal and mechanically very similar to our efforts with TPC-W.

Through an analysis of the application logic, we determined three invariants. First, that identifiers assigned by the system are unique. Second, that nicknames chosen by users are unique. Third, that item stock cannot fall below zero. Again, we preserve the first invariant using the global id generation strategy described in §7.1. The second and third invariants require both `RegisterUser'`, checking if a name submitted by a user was already chosen, and `storeBuyNow'`, which decreases stock, to be labeled as red.

7.3 Quoddy

Quoddy [14] is an open source Facebook-like social networking site. Despite being under development, Quoddy already implements the most important features of a social networking site, such as searching for a user, browsing user profiles, adding friends, posting a message, etc. These main features define 13 user requests corresponding to 15 different transactions. Of these 15 transactions, 11 are read-only transactions, thus requiring trivial no-op shadow operations.

Writing and labeling shadow operations for the 4 remaining transactions in Quoddy was straightforward. Besides reusing the recipe for unique identifiers, we only had to handle an automatic conversion of dates to the local timezone (performed by default by the database) by storing dates in UTC in all sites. In the social network we did not find system invariants to speak of; we found that all shadow operations could be labeled blue.

7.4 Experience and discussion

Our experience showed that writing shadow operations is easy; it took us about one week to understand the code, and implement and label shadow operations for all applications. We also found that the strategy of generating a different shadow operation for each distinct execution path is beneficial for two reasons. First, it leads to a simple logic for shadow operations that can be based on operations that are intrinsically commutative, e.g., *increment/decrement*, *insertion/removal*. Second, it leads to a fine-grained classification of operations, with more execution paths leading to blue operations. Finally, we found that it was useful in more than one application to make use of a standard last-writer-wins strategy to make operations that overwrite part of the state commute.

8 Evaluation

We evaluate Gemini and RedBlue consistency using microbenchmarks and our three case study applications. The primary goal of our evaluation is to determine if RedBlue consistency can improve latency and throughput in geo-replicated systems.

	UE	UW	IE	BR	SG
UE	0.4 ms 994 Mbps	85 ms 164 Mbps	92 ms 242 Mbps	150 ms 53 Mbps	252 ms 86 Mbps
UW		0.3 ms 975 Mbps	155 ms 84 Mbps	207 ms 35 Mbps	181 ms 126 Mbps
IE			0.4 ms 996 Mbps	235 ms 54 Mbps	350 ms 52 Mbps
BR				0.3 ms 993 Mbps	380 ms 65 Mbps
SG					0.3 ms 993 Mbps

Table 3: Average round trip latency and bandwidth between Amazon datacenters.

8.1 Experimental setup

We run experiments on Amazon EC2 using extra large virtual machine instances located in five sites: US east (UE), US west (UW), Ireland (IE), Brazil (BR), and Singapore (SG). Table 3 shows the average round trip latency and observed bandwidth between every pair of sites. For experiments with fewer than 5 sites, new sites are added in the following order: UE, UW, IE, BR, SG. Unless otherwise noted, users are evenly distributed across all sites. Each VM has 8 virtual cores and 15GB of RAM. VMs run Debian 6 (Squeeze) 64 bit, MySQL 5.5.18, Tomcat 6.0.35, and Sun Java SDK 1.6. Each experimental run lasts for 10 minutes.

8.2 Microbenchmark

We begin the evaluation with a simple microbenchmark designed to stress the costs and benefits of partitioning operations into red and blue sets. Each user issues requests accessing a random record from a MySQL database. Each request maps to a single shadow operation; we say a request is blue if it maps to a blue shadow operation and red otherwise. The offered workload is varied by adjusting the number of outstanding requests per user and the ratio of red and blue requests.

We run the microbenchmark experiments with a dataset consisting of 10 tables each initialized with 1,000,000 records; each record has 1 text and 4 integer attributes. The total size of the dataset is 1.0 GB.

8.2.1 User observed latency

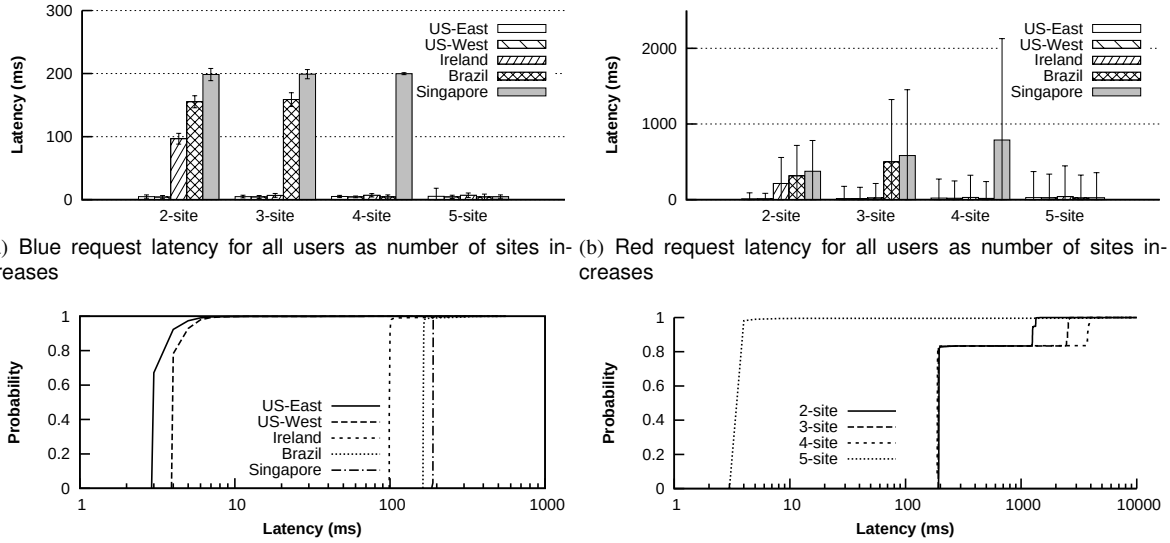
The primary benefit of using Gemini multiple sites is the decrease in latency from avoiding the intercontinental round-trips as much as possible. As a result, we first explore the impact of RedBlue consistency on user experienced latency. In the following experiments each user issues single outstanding request at one time.

Figure 10(a) shows that the average latency for blue requests is dominated by the latency between the user and the closest site; as expected, average latency decreases as additional sites appear close to the user. Figure 10(b) shows that this trend also holds for red requests. The average latency and standard deviation, however, are higher for red requests than for blue requests.

Figures 10(c) and 10(d) show the CDFs of observed latencies for blue and red requests, respectively, from the perspective of users located in Singapore. The observed latency for blue requests tracks closely with the round-trip latency to the closest site. In the $k = 2$ through $k = 4$ site configurations, four red requests from a user in Singapore are processed at the closest site during the one second in which the closest site holds the red token; every fifth request must wait $k-1$ seconds for the token to return. In the 5 site configuration, the local site also becomes a replica of the service and therefore a much larger number of requests (more than 300) can be processed while the local site holds the red token. This changes the format of the curve, even though the request issued immediately after the red token is released also needs to wait four seconds for the token to return.

8.2.2 Peak throughput

We now shift our attention to the throughput implications of RedBlue consistency. Figure 11 shows a throughput-latency graph for a 2 site configuration and three workloads: 100% blue, 100% red, and a 70% blue/30% red mix.



(a) Blue request latency for all users as number of sites increases

(b) Red request latency for all users as number of sites increases

(c) Blue latency CDF for Singapore users as number of sites increases

(d) Red latency CDF for Singapore users as number of sites increases

Figure 10: (a) and (b) show the average latency and standard deviation for blue and red requests issued by users in different locales as the number of sites is increased, respectively. (c) and (d) show the CDF of latencies for blue and red requests issued by users in Singapore as the number of sites is increased, respectively.

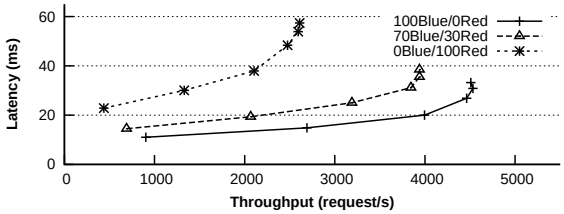


Figure 11: Throughput versus latency graph for a 2 site configuration with varying red-blue workload mixes.

The different points in each curve are obtained by increasing the offered workload, which is achieved by increasing the number of outstanding requests per user. For the mixed workload, users are partitioned into blue and red sets responsible for issuing requests of the specified color and the ratio is a result of this configuration.

The results in Figure 11 show that increasing the ratio of red requests degrades both latency and throughput. In particular, the two-fold increase in throughput for the all blue workload in comparison to the all red workload is a direct consequence of the coordination (not) required to process red (blue) requests: while red requests can only be executed by the site holding the red token to process, every site may independently process blue requests. The peak throughput of the mixed workload is proportionally situated between the two pure workloads.

8.3 Case studies: TPC-W and RUBiS

Our microbenchmark experiments indicate that RedBlue consistency instantiated with Gemini offers latency and throughput benefits in geo-replicated systems with sufficient blue shadow operations. Next, we evaluate Gemini using TPC-W and RUBiS.

8.3.1 Configuration and workloads

In all case studies experiments a single site configuration corresponds to the original unmodified code with users distributed amongst all five sites. Two through five site configurations correspond to the modified RedBlue consis-

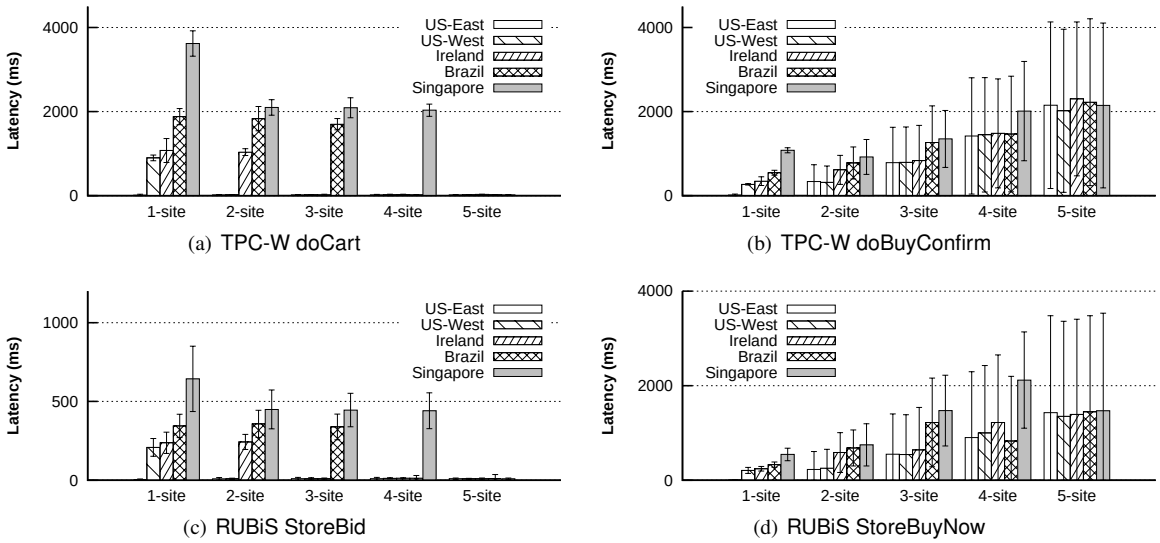


Figure 12: Average latency for selected TPC-W and RUBiS user interactions. Shadow operations for doCart and StoreBid are always blue; for doBuyConfirm and StoreBuyNow they are red 98% and 99% of the time respectively.

	Blue	Red	read-only	update
TPC-W shop	99.2	0.8	85	15
TPC-W browse	99.5	0.5	96	4
TPC-W order	93.6	6.4	63	37
RUBiS bid	97.4	2.6	85	15

Table 4: Proportion of blue and red shadow operations and read-only and update requests in TPC-W and RUBiS workloads at runtime.

tent systems running on top of Gemini. When necessary, we modified the provided user emulators so that each user maintains k outstanding requests and issues the next request as soon as a response is received.

TPC-W. TPC-W [7] defines three workload mixes differentiated by the percentage of client requests related to making purchases: browsing (5%), shopping (20%), ordering (50%). The dataset is generated with the following TPC-W parameters: 50 EBS and $10,000$ items.

RUBiS. RUBiS defines two workload mixes: browsing, exclusively comprised of read-only interactions, and bidding, where 15% of user interactions are updates. We evaluate only the bidding mix. The RUBiS database contains $33,000$ items for sale, 1 million users, $500,000$ old items and is 2.1 GB in total.

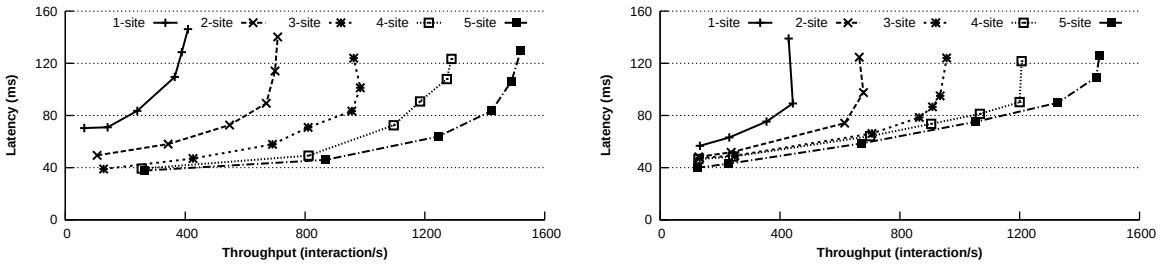
8.3.2 Prevalence of blue and red shadow operations

Table 4 shows the distribution of blue and red shadow operations during execution of the TPC-W and RUBiS workloads. The results show that TPC-W and RUBiS exhibit sufficient blue shadow operations for it to be likely that we can exploit the potential of RedBlue consistency.

8.3.3 User observed latency

We first explore the per request latency for a set of exemplar blue and red requests from TPC-W and RUBiS. For this round of experiments, each site hosts a single user issuing one outstanding request to the closest site.

From TPC-W we select doBuyConfirm (discussed in detail in §7.1) as an exemplar for red requests and doCart (responsible for adding/removing items to/from a shopping cart) as an exemplar for blue requests; from RUBiS we identify



(a) TPC-W shopping mix

(b) RUBiS bidding mix

Figure 13: Throughput versus latency for the TPC-W shopping mix and RUBiS bidding mix. The 1-dc line corresponds to the original code; the 2/3/4/5-dc lines correspond to the RedBlue consistent system variants.

StoreBuyNow (responsible for purchasing an item at the buyout price) as an exemplar for red requests and StoreBid (responsible for placing a bid on an item) as an exemplar for blue requests. Note that doBuyConfirm and StoreBid can produce either red or blue shadow operations; in our experience they produce red shadow operations 98% and 99% of the time respectively.

Figures 12(a) and 12(c) show that the latency trends for blue shadow operations are consistent with the results from the microbenchmark—observed latency is directly proportional to the latency to the closest site. The raw latency values are higher than the round-trip time from the user to the nearest site because processing each request involves sending one or more images to the user.

For red requests, Figures 12(b) and 12(d) show that latency and standard deviation both increase with the number of sites. The increase in standard deviation is an expected side effect of the simple scheme that Gemini uses to exchange the red token and is consistent with the microbenchmark results. Similarly, the increase in average latency is due to the fact that the time for a token rotation increases, together with the fact that red requests are not frequent enough that several cannot be slipped in during the same token holding interval. We note that the token passing scheme used by Gemini is simple and additional work is needed to identify an optimal strategy for regulating red shadow operations.

8.3.4 Peak throughput

We now shift our attention to the throughput afforded by our RedBlue consistent versions of TPC-W and RUBiS, and how it scales with the number of sites. For these experiments we vary the workload by increasing the number of outstanding requests maintained by each user. Throughput is measured according to interactions per second, a metric defined by TPC-W to correspond to user requests per second.

Figure 13 shows throughput and latency for the TPC-W shopping mix and RUBiS bidding mix as we vary the number of sites. In both systems, increasing the number of sites increases peak throughput and decreases average latency. The decreased latency results from situating users closer to the site processing their requests. The increase in throughput is due to processing blue and read-only operations at multiple sites, given that processing their side effects is relatively inexpensive. The speedup for a 5 site Gemini deployment of TPC-W is 3.7x against the original code for the shopping mix; the 5 site Gemini deployment of RUBiS shows a speedup of 2.3x.

Figure 14 shows the throughput and latency graph for a two site configuration running the TPC-W browsing, shopping, and ordering mixes. As expected, the browsing mix, which has the highest percentage of blue and read-only requests, exhibits the highest peak throughput, and the ordering mix, with the lowest percentage of blue and read-only requests, exhibits the lowest peak throughput.

8.4 Case study: Quoddy

Quoddy differs from TPC-W and RUBiS in one crucial way: it has no red shadow operations. We use Quoddy to show the full power of RedBlue geo-replication.

Quoddy does not define a benchmark workload for testing purposes. Thus we design a social networking workload generator based on the measurement study of Benevenuto et al. [4]. In this workload, 85% of the interactions are read-

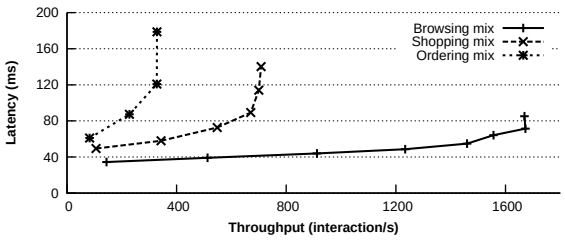


Figure 14: TPC-W: Throughput vs. latency graph for TPC-W with Gemini spanning two sites when running the three workload mixes.

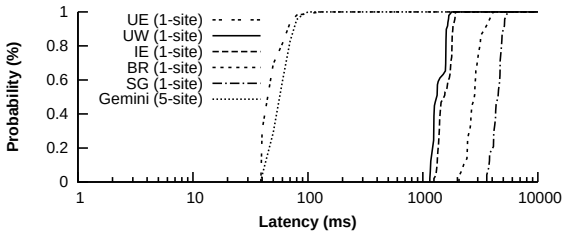


Figure 15: User latencies CDF for the addFriend request in single site Quoddy and 5-site Gemini deployments.

only page loads and 15% of the interactions include updates, e.g., request friendship, confirm friendship, or update status. Our test database contains 200,000 users and is 2.6 GB in total size.

In a departure from previous experiments, we run only two configurations. The first is the original Quoddy code in a single site. The second is our Gemini based RedBlue consistent version replicated across 5 sites. In both configurations, users are distributed in all 5 regions.

Figure 15 shows the CDF of user experienced latencies for the addFriend operation. All Gemini users experience latency comparable to the local users in the original Quoddy deployment; a dramatic improvement for users not based in the US East region. The significantly higher latencies for remote regions are associated with the images and javascripts that Quoddy distributes as part of processing the addFriend request.

8.5 Gemini overheads

Gemini is a middleware layer that interposes between the applications that leverage RedBlue consistency and a set of database systems where data is stored. We evaluate the performance overhead imposed by our prototype by comparing the performance of a single site Gemini deployment with the unmodified TPC-W and RUBiS systems directly accessing a database. For this experiment we locate all users in the same site as the service.

Table 5 presents the peak throughput and average latency for the TPC-W shopping and RUBiS bidding mixes. The peak throughput of a single site Gemini deployment is between 82% and 94% of the original and Gemini increases latency by 1ms per request.

9 Conclusion

In this paper, we presented a principled approach to building geo-replicated systems that are fast as possible and consistent when needed. Our approach is based on our novel notion of RedBlue consistency allowing both strongly consistent (red) operations and eventually consistent (blue) operations to coexist, a concept of shadow operation enabling the maximum usage of blue operations, and a labeling methodology for precisely determining which operations to be assigned which consistency level. Experimental results from running benchmarks with our system Gemini show that RedBlue consistency significantly improves the performance of geo-replicated systems.

	TPC-W shopping		RUBiS biding	
	Original	Gemini	Original	Gemini
Thput. (inter/s)	409	386	450	370
Avg. latency	14 ms	15 ms	6 ms	7 ms

Table 5: Performance comparison between the original code and the Gemini version for both TPC-W and RUBiS within a single site.

Acknowledgments

We sincerely thank Edmund Wong, Rose Hoberman, Lorenzo Alvisi, our shepherd Jinyang Li, and the anonymous reviewers for their insightful comments and suggestions. The research of R. Rodrigues has received funding from the European Research Council under an ERC starting grant. J. Gehrke was supported by the National Science Foundation under Grant IIS-1012593, the iAd Project funded by the Research Council of Norway, a gift from amazon.com, and a Humboldt Research Award from the Alexander von Humboldt Foundation. N. Preguiça is supported by FCT/MCT projects PEst-OE/EEI/UI0527/2011 and PTDC/EIA-EIA/108963/2008.

References

[1] Ebay website. `http://http://www.ebay.com/`, 2012.

[2] M. Ahamad, G. Neiger, J. E. Burns, P. Kohli, and P. Hutto. Causal memory: Definitions, implementation and programming. Technical report, Georgia Institute of Technology, 1994.

[3] J. Baker, C. Bond, J. C. Corbett, J. Furman, A. Khorlin, J. Larson, J.-M. Leon, Y. Li, A. Lloyd, and V. Yushprakh. Megastore: Providing scalable, highly available storage for interactive services. In *CIDR*, 2011.

[4] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In *IMC*, 2009.

[5] P. A. Bernstein, V. Hadzilacos, and N. Goodman. *Concurrency control and recovery in database systems*. 1987.

[6] E. Cecchet, G. Candea, and A. Ailamaki. Middleware-based database replication: the gaps between theory and practice. In *SIGMOD*, 2008.

[7] T. consortium. Tpc benchmark-w specification v. 1.8. `http://www.tpc.org/tpcw/spec/tpcw_v1.8.pdf`, 2002.

[8] B. F. Cooper, R. Ramakrishnan, U. Srivastava, A. Silberstein, P. Bohannon, H.-A. Jacobsen, N. Puz, D. Weaver, and R. Yerneni. Pnuts: Yahoo!'s hosted data serving platform. In *VLDB*, 2008.

[9] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels. Dynamo: amazon's highly available key-value store. In *SOSP*, 2007.

[10] C. A. Ellis and S. J. Gibbs. Concurrency control in groupware systems. In *SIGMOD*, 1989.

[11] C. Emmanuel and M. Julie. Rubis: Rice university bidding system. `http://rubis.ow2.org/`, 2009.

[12] A. J. Feldman, W. P. Zeller, M. J. Freedman, and E. W. Felten. Sporc: group collaboration using untrusted cloud resources. In *OSDI*, 2010.

[13] S. Floyd, V. Jacobson, C.-G. Liu, S. McCanne, and L. Zhang. A reliable multicast framework for light-weight sessions and application level framing. *IEEE/ACM Trans. Netw.*, 1997.

[14] Fogbeam Labs. Quoddy code repository, 2012. `"http://code.google.com/p/quoddy/"`.

[15] J. Gray. The transaction concept: Virtues and limitations. In *VLDB*, 1981.

[16] J. Gray, P. Helland, P. O'Neil, and D. Shasha. The dangers of replication and a solution. In *SIGMOD*, 1996.

[17] M. P. Herlihy and J. M. Wing. Linearizability: a correctness condition for concurrent objects. *ACM Trans. Program. Lang. Syst.*, 1990.

[18] T. Kraska, M. Hentschel, G. Alonso, and D. Kossmann. Consistency rationing in the cloud: pay only when it matters. In *VLDB*, 2009.

[19] R. Ladin, B. Liskov, L. Shrira, and S. Ghemawat. Providing high availability using lazy replication. *ACM Trans. Comput. Syst.*, 1992.

[20] L. Lamport. Time, clocks, and the ordering of events in a distributed system. *Commun. ACM*, 1978.

[21] L. Lamport. The part-time parliament. *ACM Trans. Comput. Syst.*, 1998.

[22] H. Li. Practical consistency tradeoffs. In *PODC*, 2012.

[23] J. Li, M. Krohn, D. Mazières, and D. Shasha. Secure untrusted data repository (sundr). In *OSDI*, 2004.

[24] 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 *SOSP*, 2011.

[25] P. Mahajan, S. Setty, S. Lee, A. Clement, L. Alvisi, M. Dahlin, and M. Walfish. Depot: cloud storage with minimal trust. In *OSDI*, 2010.

[26] D. Mazières and D. Shasha. Building secure file systems out of byzantine storage. In *PODC*, 2002.

[27] K. Petersen, M. J. Spreitzer, D. B. Terry, M. M. Theimer, and A. J. Demers. Flexible update propagation for weakly consistent replication. In *SOSP*, 1997.

[28] A. Rito da Silva et al. Project fenix applications and information systems of instituto superior tcnico. `https://fenix-cvs.ist.utl.pt`, 2012.

[29] Y. Saito and M. Shapiro. Optimistic replication. *ACM Comput. Surv.*, 2005.

[30] F. B. Schneider. Implementing fault-tolerant services using the state machine approach: a tutorial. *ACM Comput. Surv.*, 1990.

[31] E. Schurman and J. Brutlag. Performance relatedchanges and their user impact. Presented at velocity web performance and operations conference, 2009.

[32] M. Shapiro, N. Preguiça, C. Baquero, and M. Zawirski. Conflict-free replicated data types. In *SSS*, 2011.

[33] A. Singh, P. Fonseca, P. Kuznetsov, R. Rodrigues, and P. Maniatis. Zeno: eventually consistent byzantine-fault tolerance. In *NSDI*, 2009.

[34] Y. Sovran, R. Power, M. K. Aguilera, and J. Li. Transactional storage for geo-replicated systems. In *SOSP*, 2011.

[35] D. Stocker. Delta transactions. `http://collectiveweb.wordpress.com/2010/03/01/delta-transactions/`, 2010.

[36] D. B. Terry, A. J. Demers, K. Petersen, M. Spreitzer, M. Theimer, and B. W. Welch. Session guarantees for weakly consistent replicated data. In *PDIS*, 1994.

[37] D. B. Terry, M. M. Theimer, K. Petersen, A. J. Demers, M. J. Spreitzer, and C. H. Hauser. Managing update conflicts in Bayou, a weakly connected replicated storage system. In *SOSP*, 1995.

[38] R. van Renesse, K. P. Birman, and S. Maffeis. Horus: A flexible group communication system. *Commun. ACM*, 1996.

[39] W. E. Weihl. Commutativity-based concurrency control for abstract data types. *IEEE Trans. Comput.*, 1988.

[40] H. Yu and A. Vahdat. Design and evaluation of a continuous consistency model for replicated services. In *OSDI*, 2000.