C8DB: Geo-Replicated, Conflict-Free Document Database

Durga Gokina, Chetan Venkatesh, Christopher S. Meiklejohn
Macrometa

Modern applications are becoming increasingly concerned with providing low-latency, global operation at scale. To provide low-latency operations, applications typically need to be deployed in multiple datacenters (DCs), and user data replicated to each of these DCs, allowing user requests to be serviced by the nearest geographic DC. However, while reducing user perceived latency through the replication of data, application developers are now faced with an additional, even more difficult challenge: managing the consistency of multiple replicas.

First generation Content Delivery Networks (CDNs) serve to reduce user perceived latency when performing read operations against shared data. Each data item in the system is designated with a primary site where all data modifications occur and read-only replicas are maintained at every other site. These replicas are periodically refreshed on demand and cache eviction messages are used to expire the data stored at remote replicas. Updates are totally ordered by the primary site: every replica in the system sees the same updates in the same order. Therefore, users observe *eventual consistency*: writes are performed at the designated primary replica and reads at other DCs will eventually return the result of the most recent write. While CDNs work well for a read dominated workload, they fail to assist the developer in situations where the workload may be write dominated.

To solve the issue of low-latency, write heavy workloads, every replica needs to be able to accept and process writes on behalf of the user. However, handling writes at each DC raises a number of difficult challenges. First, if all replicas in the system are expected to observe the same events in the same order, the system must use some form of coordination between replicas to obtain this order. In the case of geo-replication, this can be costly as (i.) the geographic distance between replicas communicating with one another will slow down the system to the speed of the slowest link and (ii.) remote replicas may be unavailable due to network partitions between two remote DCs. Second, if we forego the requirement on totally ordering all writes, concurrent writes for the same data item that occur at different DCs may *conflict* and the system will have to make a decision on how to resolve these conflicting writes.

Operation-based Conflict-Free Replicated Data Types (CRDTs) are one way of dealing with automatic conflict resolution. Operation-based CRDTs rely on two properties: (i.) causal delivery of updates, ensuring that updates are delivered to remote replicas observing the causal order of events; and (ii.) the commutativity of concurrent operations, ensuring that any updates that these operations result in the same outcome regardless of delivery order. However, since some operations cannot be made commutative (e.g., bank transfer while ensuring a non-negative balance invariant) the system must also support strong consistency *when necessary* in order to provide a comprehensive solution.

We present C8DB, a geo-replicated, conflict-free replicated document database. C8DB is a *multi-master* database where the system provides session guarantees and transactions with snapshot isolation. Documents in C8DB are JSON: these documents can contain graphs, registers, sequences, and references to other documents. Modifications are performed using a SQL-like language called C8QL, that provides the user the ability to perform updates to single or multiple documents using atomic transactions. When strong consistency is required, C8DB will operate using a *single-master* for those documents: thereby, ensuring serializabile transactions and strong consistency.

In C8DB, operation-based CRDTs are used to represent modifications to documents. Modifications in C8DB are represented as operations performed on single attributes for each document and aggregated in an append-only log of updates. Reliable causal broadcast is used to deliver these updates to other replicas in the system. When conflicting updates are encountered, they are resolved on a per-attribute basis using data type specific conflict resolution. C8DB is currently deployed in 25 DC's globally today, and is expanding to 50 additional sites by the end of the year.

C8DB is novel in a few aspects. First, it is the only industrial database to take advantage of the lower cost of operation-based CRDTs: previous designs have relied on state-based CRDTs that rely on transmitting the entire state of objects in the replication process. Second, operation-based CRDTs typically rely on storage of the entire log of operations for materliazation during read: C8DB avoids this cost by implementing a novel garbage collection algorithm that reduces the cost of this operation by only storing a single operation for each attribute of a given object.