Malleable Flow for Time-Bounded Replica Consistency Control

Yuqing Zhu #1, Jianmin Wang #2, Philip S. Yu #3
#Tsinghua University, School of Software
*University of Illinois at Chicago, Dept. of Computer Science
1 zhuyuqing@gmail.com, 2 jimwang@tsinghua.edu.cn, 3 ppsyu@uic.edu

THSS-ISE Lab

Big Picture

A system with malleable flow (M-flow) supports latency-bounded operations that maximize replica consistency within the given time. A better consistency status indicates a shorter latency for a following consistent read, or a more recent value returned by a following read within the same latency.

The M-flow replication process starts from when the request is received till when the response is sent. It is decomposed into six stages, i.e. reception, transmission, coordination, execution, compaction and acquisition.

M-Flow Decomposition

1. Reception: Receive requests
2. Transmission: Send requests to remote data centers
3. Coordination: Coordinate updates across data centers
4. Execution: Process requests at the data centers
5. Compaction: Compact replicas across data centers
6. Acquisition: Retrieve consistent data

Motivation

- Replication is important for availability, scalability and reliability in large-scale systems
- Trade-off between replica consistency and availability, latency exists
- Guaranteeing best replica consistency within a given latency
  - Not provided yet
  - But highly DESIRABLE

Key Idea

- To decompose the replication process into stoppable stages and then into a directed graph of ordered minute steps
- To reform an execution flow by finding a path of steps in the graph that
  - Meets the time requirement
  - Maximizes replica consistency status
  - Guarantees durability and fault tolerance

The deomposed replication process enables this flexibility and the control of replica consistency (and latency) by reforming a suitable execution process with carefully selected stages and writes.

Bounded Time & Traded Consistency

Consistency versus latency under different cross-DC bandwidths (the number of returned values versus the given read latency constraints):
- A larger latency bound for instantaneous reads following writes leads to a larger number of returned values.

Implementation over Cassandra

The storage architecture guarantees the durability of writes, and enables the execution flow to stop at the end of any stage.

Implementation over Cassandra

The decomposed steps and their execution orders form a directed graph, consisting of six partitions corresponding to the six stages.

The Malleable Flow (M-FLOW)

Algorithm: (CON Algorithm finding the legal path with maximized consistency)

* This work is supported by the National Science and Technology Major Special Projects of China under Grant No. 2010ZX01042-002-002-01, and the National Natural Science Foundation of China under Grant No. 61073705.