Paxos based directory updates for geo-replicated cloud storage

Modern geo-distributed cloud data stores require and rely upon significant amounts of data movement to optimize access latencies for their objects. This flexible, revisable scheme for placement of data with varying levels of consistency promotes the need for a system that can manage the migration process for data items in a consistent fashion. This system must also keep track of meta-data detailing each object’s current deployment and associated information in a strictly correct manner as this now serves as a directory/lookup table to track the location of the object to be read or written to. In this work, we present an open sourced implementation of such a system built around an open sourced implementation of the Paxos protocol. Our system entertains requests for migration of data elements through a CRUD-like API and guarantees correctness along the process of completion of the migration request across a cluster. We also present our studies and results of the instrumentation and behavior of the Paxos protocol in simulated WAN environments.

An emerging trend in cloud computing is cloud data stores which replicate data for availability, redundancy and optimized latencies across geographically distributed data centers. One important class of such systems is a directory based scheme which offers the advantage of flexible placement of data by maintaining a directory to track the location of the same. As workloads vary dynamically thus prompting a revision in data placement, a key challenge of maintaining and updating the directory state emerges. In this thesis we present an implementation to address the problem of correctly updating these directories. Built around an open-sourced implementation, JPaxos, of the Paxos consensus protocol, our system manages and completes requests to migrate data elements using a cluster of systems. Using a distributed approach makes our system failure tolerant by avoiding the single-point-of-failure problem of centralized coordinator based systems. We instrument and evaluate our implementation on a large scale research testbed, PRObE using DummyNet as a network simulator. Our results show that such a scheme is feasible and that the latencies are acceptable in simulated WAN environments. Our contributions include the performance data of a fully functional Paxos based system and learnings from practical design considerations of implementing the same.