Load Balancing In Ceph:

Load balancing with pseudorandom placement

Esteban Molina-Estolano, Carlos Maltzahn, Scott Brandt

University of California, Santa Cruz

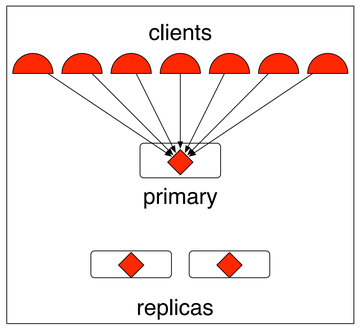
Background

• Pseudorandom placement in distributed storage systems offers scalability benefits

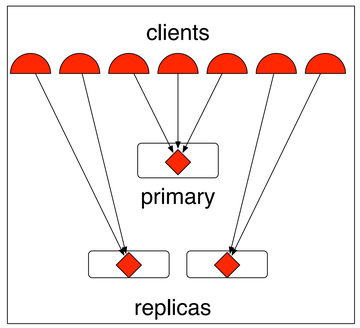
• Pseudorandom placement makes load balancing harder; new techniques are required

• We explore different load balancing techniques using Ceph, an object-based storage system developed at UCSC

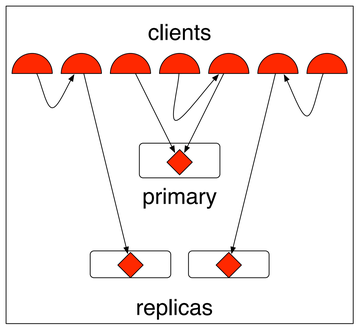
Read Shedding



read flash crowd: many clients read the same object concurrently

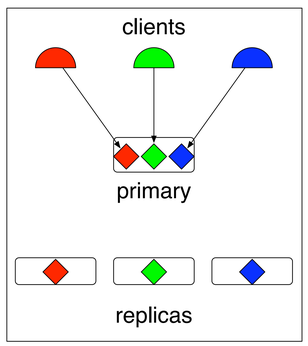


read shedding: redirect some clients to read the object from replicas

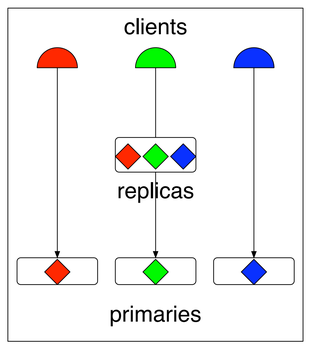


extended read shedding: redirect some clients to read the object from other clients

Primary Switching

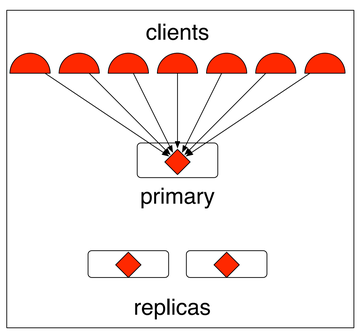


coincidental overload: one node holds primary copies for concurrent requests for different objects

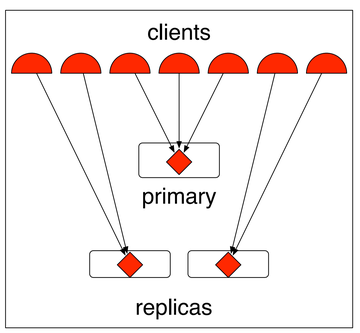


primary switching: swap the primary and replica roles for each object to distribute load

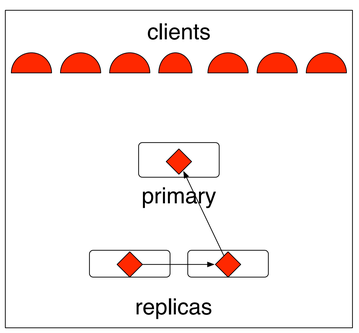
Write Shedding



write flash crowd: many clients write to the same object concurrently



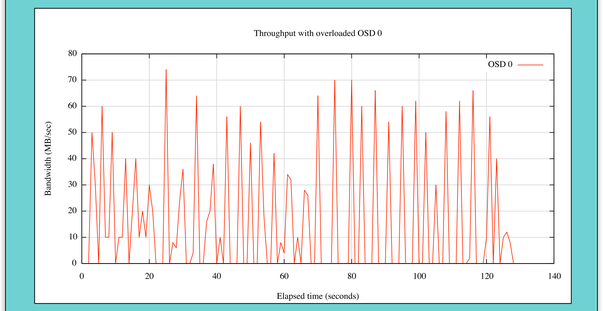
write shedding: redirect some clients to write to replicas

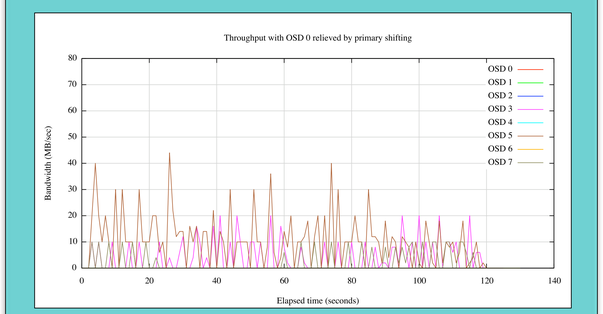


write consolidation: easy with lazy POSIX I/O extensions, complicated without them

Preliminary Experiments

• Primary switching implemented in Ceph and tested for a write workload





These graphs show per-node bandwidth with respect to time for an unbalanced workload writing to a single node. In the top graph, primary switching is turned off; in the bottom graph, primary switching distributes the load among several nodes.

Status and Results

• Preliminary experiments show that primary switching is feasible

• Simulation in development to test techniques for large cluster sizes