Adapting RAID Methods for Use in Object Storage Systems

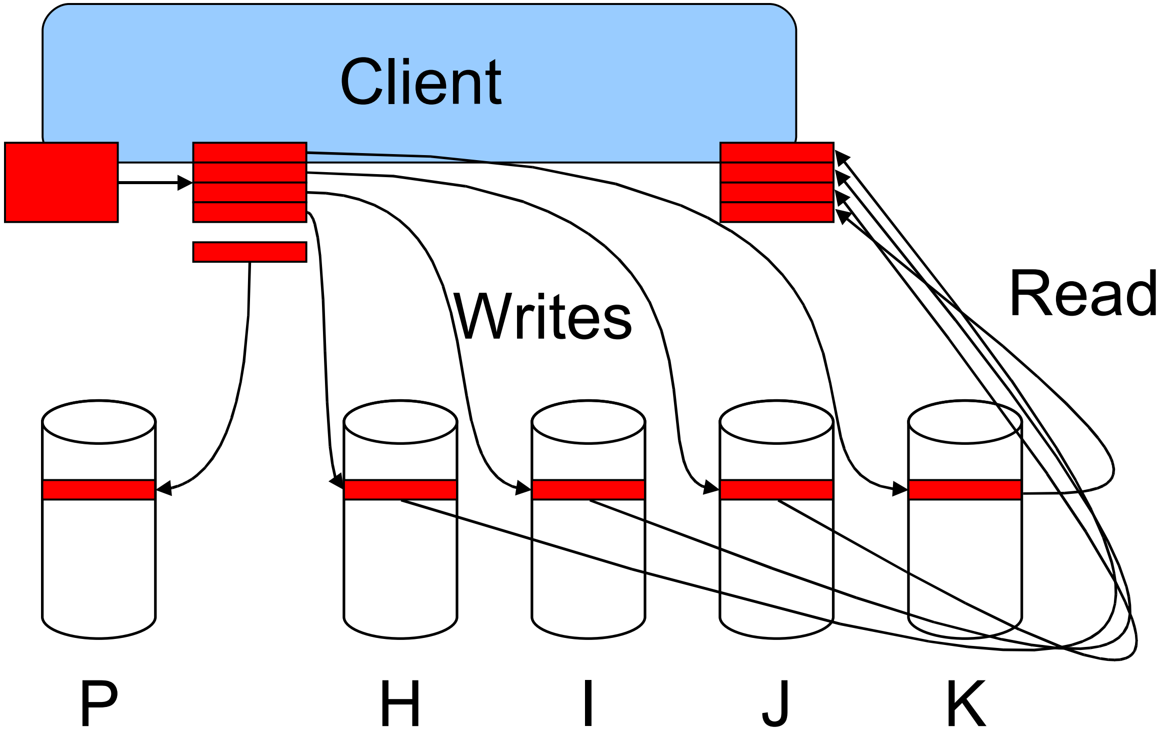
David Bigelow, Scott A. Brandt, Carlos Maltzahn, Sage Weil
University of California, Santa Cruz
{dbigelow, scott, carlosm, sage}@cs.ucsc.edu

FAST 2007 Work-In-Progress
February 27, 2008

Motivation: OSD Reliability

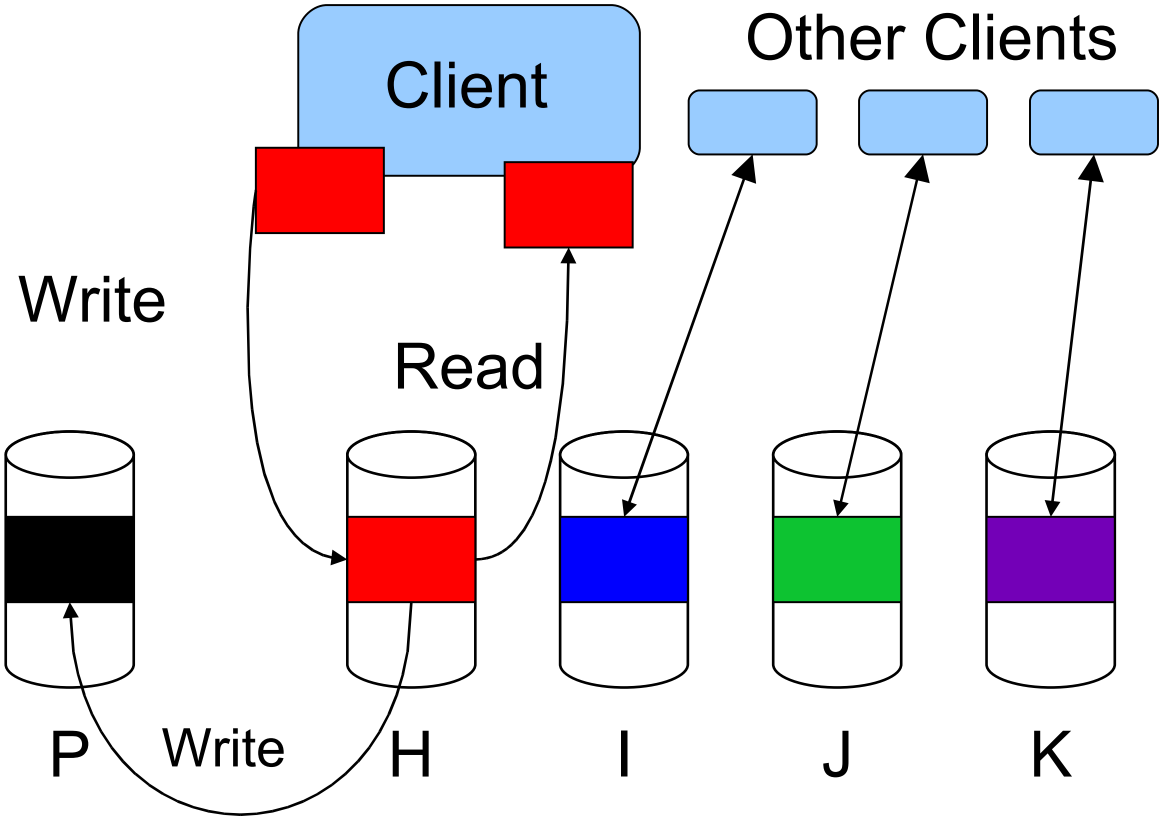
- Mirroring is Expensive
  - System may have petabytes of data in thousands of devices
  - For two-way mirroring alone, the system cost doubles
  - Linear scaling of system cost for each additional degree of protection
- RAID (and other error-correction codes)
  - Simple RAID codes can reduce overhead to (N + 1)/N
  - More advanced error-correction codes (like Reed-Solomon) are available
  - How to adapt these methods for use in object-based storage?
- High-Performance Storage
  - Typical systems will have very high performance requirements
  - Can RAID maintain the necessary performance level?

Client-Based RAID



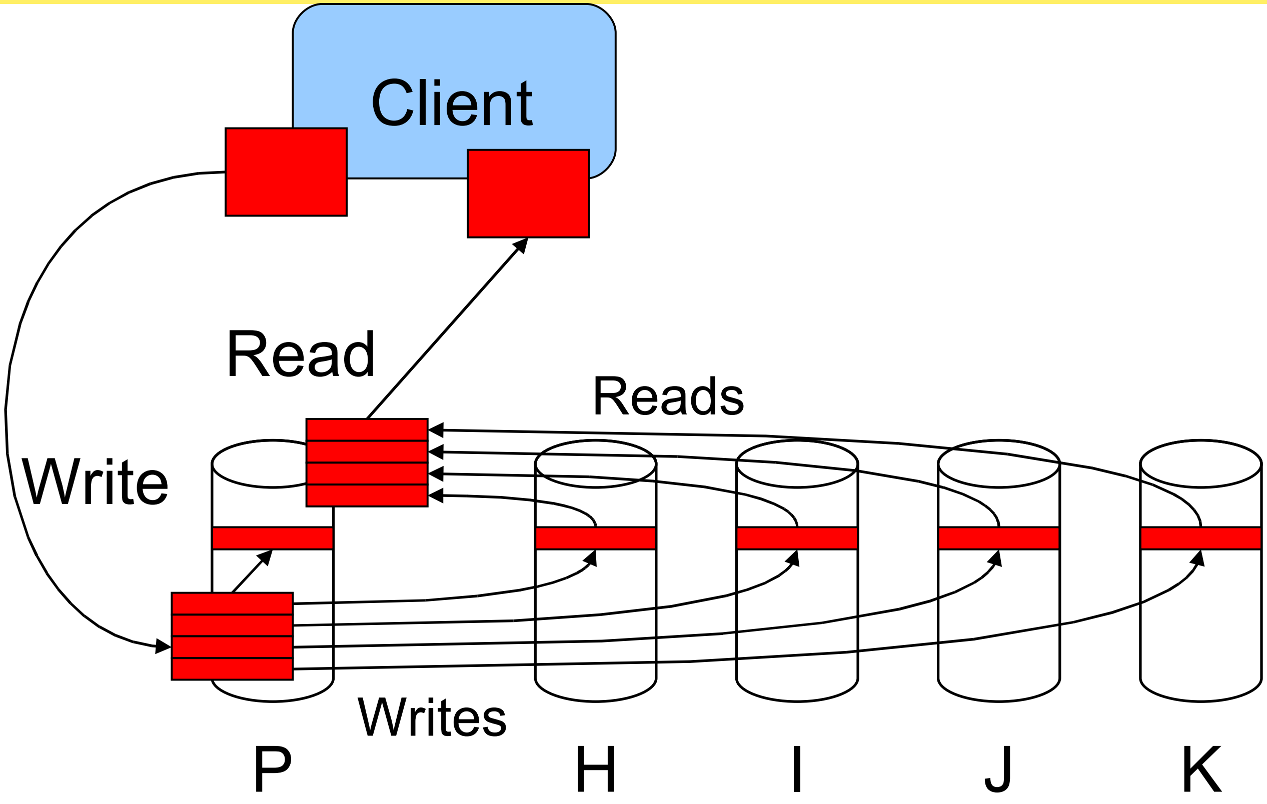
- The client alone determines how its data will be stored
- Storage system only responsible for storing and returning objects

RAID Across Objects



- No overhead to client -- storage system maintains own records
- Device failure can lead to large reconstruction times
- Very jagged performance curve in degraded mode

RAID Within Objects



- Always additional delay to the client for both reading and writing
- Device failure has smaller reconstruction times
- Smoother performance curve in degraded mode

Current Status

- Simulation
  - Measuring of relative performance
- Implementation
  - Applying techniques to Ceph Object Storage System
  - Initial approach of parity based RAID
- Continuing Work
  - More complex schemes to tolerate multiple failures
  - Hierarchical model to allow multiple reliability schemes