Alternative Reliability Models in Ceph

David Bigelow
University of California, Santa Cruz

SSRC Retreat
May 31, 2007
Introduction

Objective: To introduce RAID-like methods into Ceph to replace or supplement the current system of mirroring

◆ Primary Goals:
  • Reduce Hardware Overheads
  • Increase Reliability

◆ Secondary Goals:
  • Allow multiple reliability schemes in the same system, creating a hierarchy
  • Keep all changes transparent to the client if possible
    • However, this may not be possible in the more advanced variants
Proposed Ceph Model (very basic)

Each color is a different group, each block is one object
Application of RAID-Like Methods

- Have a choice of several RAID-Like Methods
  - RAID-4/5
  - Error-Detecting and Erasure Codes (such as Reed-Solomon)
  - Other parity based schemes (EVENODD, RDP, etc.)

- Differences from standard RAID:
  - Parity groups differ on a per-object basis
  - Recovery can happen immediately on an entirely new device, aside from the usual ‘hot spare’
  - I/O performance is not necessarily a top consideration anymore
Current Ceph Model

Client

Primary

Replica

Replica

Time

Write

Ack

Commit to disk

Commit
RAID Methods (Basic Parity)

RAID Within

\[
P = H \oplus I \oplus J \oplus K
\]

RAID Across

\[
H \oplus I \oplus J \oplus K \oplus P
\]
## Comparison of Methods

<table>
<thead>
<tr>
<th></th>
<th>RAID Across</th>
<th>RAID Within</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write Speed</td>
<td>Fast</td>
<td>Slow</td>
</tr>
<tr>
<td>Read Speed</td>
<td>Fast</td>
<td>Slow or Fast (depending on client)</td>
</tr>
<tr>
<td>Required Calculations</td>
<td>2x (single parity)</td>
<td>1x (single parity)</td>
</tr>
<tr>
<td>Amount of time spent in</td>
<td>Very Little</td>
<td>Slightly more than RAID Across</td>
</tr>
<tr>
<td>degraded mode</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance in</td>
<td>Poor</td>
<td>Good</td>
</tr>
<tr>
<td>degraded mode</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Current Direction

- Currently developing a simulator to model different schemes of encoding
- Integrate a basic RAID-like scheme into Ceph
- Extend Ceph to handle a hierarchical model
Questions?