P2P filesystem
Sam Prentice & Sudeep Pillai
Central Extent Server
Distributed Extents

- YFS
  - Local Extent

- Light-weight extent membership service

- Lock Server
Shared+Exclusive Locking

- Shared access: Multiple readers
  - acquire_read

- Exclusive access: Single writer
  - acquire_write

- Pitfalls
  - Writer Starvation: Lock contention where writer waits on same lock that several readers request
    - Lock granted to either single writer in queue head
    OR
    - Lock granted to group of readers in queue head
Optimizations

• Typical lock transition mechanism
  • acquire-revoke-release-retry cycle

• Lock upgrading/downgrading
  • Upgrading (R→W): Revoke read access for every other client
  • Downgrading (W→R): Clients with write access permitted to read until revocation
    - Optimize special cases: new read requests not preceded by write request granted immediately
Distributed Extents

- Extents reside distributed across clients
- Freshest extents are directly read by clients via P2P
- Avoid flushing data to central extent server
- Extent ID membership held centrally
  - Latest writer/owner is tracked
  - Peers can request for latest owner for fresh copies
- Allows distributing of hot-spots in filesystem
Load Balancing

- Load distributed evenly across clients
- Reduced bandwidth per client compared to central extent server which did the heavy-lifting

P2P Load balancing test with 10 clients
Benchmarking

• Handles well with hot-spots of file-system usage
Thanks!
Sequential consistency

- Readers and writers see the most recent extent copy
- Modifications only made by exclusive writer
- Difficulties
  - Ensure extent attribute synchronization
Benchmarking

- **RPC counts**

<table>
<thead>
<tr>
<th></th>
<th>Central</th>
<th>P2P</th>
</tr>
</thead>
<tbody>
<tr>
<td>test-lab-3-c</td>
<td>1400</td>
<td>400</td>
</tr>
</tbody>
</table>

- **Timing**

<table>
<thead>
<tr>
<th></th>
<th>Central</th>
<th>P2P</th>
</tr>
</thead>
<tbody>
<tr>
<td>test-lab-3-c</td>
<td>~1600 ms</td>
<td>~1250 ms</td>
</tr>
</tbody>
</table>

- **Bandwidth**

<table>
<thead>
<tr>
<th></th>
<th>Central</th>
<th>P2P</th>
</tr>
</thead>
<tbody>
<tr>
<td>test-lab-3-c</td>
<td>27.02 MB</td>
<td>14.31 MB</td>
</tr>
</tbody>
</table>