**RAIDP: ReplicAtion with Intra-Disk Parity**

Eitan Rosenfeld, Aviad Zuck, Nadav Amit, Michael Factor, Dan Tsafrir

### Triplication vs. Erasure Coding

**Background:** Modern cloud storage systems use redundancy to withstand simultaneous disk failures.

<table>
<thead>
<tr>
<th>Disk 1</th>
<th>Disk 2</th>
<th>Disk 3</th>
<th>Disk 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>a</td>
<td>b</td>
<td>a</td>
</tr>
<tr>
<td>c</td>
<td>b</td>
<td>c</td>
<td>c</td>
</tr>
</tbody>
</table>

#### Three-Way Replication

- **Erasure Coding**: 2a+b, 2c+d
- **Data Chunks**: a, b, c, d

#### Triplication

- **Data Chunks**: a, b, c, d

### Recovery in RAIDP

**0. Two disks fail simultaneously**

- **Disk 1**: d, c
- **Disk 2**: b, c
- **Disk 3**: a, d
- **Disk 4**: a, b

**Replacement Drives**

- d
- b
- a
- a

### RAIDP Hybrid Layout

**Problem:** Triplication is expensive and used typically for warm data only.

**Idea:** Two independent NVRAM disk add-ons are cheaper than a third replica.

**Proposed Solution:**
- Disks partitioned to superchunks (e.g., 4GB) with two replicas only.
- Bitwise mirror superchunk writes.
- Disks share < 1 superchunk.
- Disk add-ons use NVRAM to store local chunk parity.
- Add-ons fail separately from disk.

<table>
<thead>
<tr>
<th>Disk 1</th>
<th>Disk 2</th>
<th>Disk 3</th>
<th>Disk 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td>c</td>
<td>b</td>
<td>a</td>
</tr>
</tbody>
</table>

**Super Chunks**

- d, c
- b, c
- a, d
- a, b

**Local NVRAM Add-Ons**

- d, c
- b, c
- a, d
- a, b

### Evaluation

- PoC implementation in Hadoop 1.0.4
- Cluster of 16 Intel Xeon E3-1220 machines (3.10GHz)
  - Ubuntu 14.04 (kernel 3.13)
- 6GB superchunks, add-ons simulated in memory.
- Append-only baseline

**HDFS w/triplication vs. RAIDP**

<table>
<thead>
<tr>
<th></th>
<th>HDFS</th>
<th>RAIDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Terasort</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Wordcount</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>Read</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

**Recovery Time:** 14-15x faster than RAID-6 following double disk failure.