f4: Facebook’s Warm BLOB Storage System [OSDI 2014]

Satadru Pan,

Facebook
BLOBs@FB

Immutable & Unstructured

Diverse

A LOT of them!!
Facebook PHOTO growth

Photos Stored

- 600 Million in 2006
- 240 Billion in 2013
- 400 Billion in 2014

Years:
- 2004
- 2006
- 2013
- 2014
Data cools off rapidly

HOT DATA

WARM DATA

Normalized Read Rates

< 1 Days 1 Day 1 Week 1 Month 1 Year

510X 590X 98X 30X 16X 14X 6X 7X 2X 1X 1X

Photo Video

3 Months
Handling failures

Replication: * 3 = 3.6
Handling load

Reduce space usage AND Not compromise reliability
Background: Data serving

- CDN protects storage
- Router abstracts storage
- Web tier adds business logic
Background: Haystack [OSDI2010]

- Volume is a series of BLOBs
- In-memory index
Introducing f4: Haystack on cells
Data splitting

Stripe 1

Reed Solomon Encoding

Stripe 2

10G Volume

4G parity

RS

BLOB 1
BLOB 2
BLOB 3

RS

BLOB 4
BLOB 5
BLOB 6
BLOB 7
BLOB 8
BLOB 9
BLOB 10
BLOB 11
Reed Solomon (10, 4) is used in practice (1.4X)
Tolerates 4 racks (→ 4 disk/host) failures
Reads

- 2-phase: Index read returns the exact physical location of the BLOB
Reads under cell-local failures

- Cell-Local failures (disks.getHosts/racks) handled locally
Reads under datacenter failures (2.8X)

2 * 1.4X = 2.8X
Cross datacenter XOR (1.5 * 1.4 = 2.1X)

67%

33%

Cell in Datacenter1

Cell in Datacenter2

Cell in Datacenter3

Cross –DC index copy
Reads with datacenter failures (2.1X)
<table>
<thead>
<tr>
<th></th>
<th>Haystack with 3 copies</th>
<th>f4 2.8</th>
<th>f4 2.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replication</td>
<td>3.6X</td>
<td>2.8X</td>
<td>2.1X</td>
</tr>
<tr>
<td>Irrecoverable Disk Failures</td>
<td>9</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Irrecoverable Host Failures</td>
<td>3</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Irrecoverable Rack failures</td>
<td>3</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Irrecoverable Datacenter failures</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Load split</td>
<td>3X</td>
<td>2X</td>
<td>1X</td>
</tr>
</tbody>
</table>
Evaluation

- What and how much data is “warm”? 

- Can f4 satisfy throughput and latency requirements?
Methodology

- CDN data: 1 day, 0.5% sampling
- BLOB store data: 2 week, 0.1%
- Random distribution of BLOBs assumed
- The worst case rates reported
Hot and warm divide

HOT DATA
< 3 months → Haystack

WARM DATA
> 3 months → f4

Reads/Sec per disk

Age
1 week 1 month 3 month 1 year

80 Reads/Sec
It is warm, not cold
Performance: Most loaded disk in cluster

Peak load on disk: 35 Reads/Sec
f4 Performance: Latency

- P80 = 30ms
- P99 = 80ms
Concluding Remarks

- Facebook’s BLOB storage is big and growing
- BLOBs cool down with age
  - ~100X drop in read requests in 60 days
- Haystack’s 3.6X replication over provisioning for old, warm data.
- f4 encodes data to lower replication to 2.1X