f4: Facebook’s Warm BLOB Storage System


*Facebook Inc., ψUniversity of Southern California, *Princeton University
BLOBs@FB

Profile Photo

Immutable & Unstructured

Diverse

A LOT of them!!
Data cools off rapidly
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

Stripe1 => BLOB2

RS

Stripe2 => 10G Volume

RS

4G parity

BLOB1

BLOB2

BLOB3

BLOB4

BLOB5

BLOB6

BLOB7

BLOB8

BLOB9

BLOB10

BLOB11
Data placement

- Reed Solomon (10, 4) is used in practice (1.4X)
- Tolerates 4 racks (→ 4 disk/host) failures
2-phase: Index read returns the exact physical location of the BLOB
Reads under cell-local failures

- Cell-Local failures (diskshostsracks) handled locally
Reads under datacenter failures (2.8X)

$$2 \times 1.4X = 2.8X$$
Cross datacenter XOR ($1.5 \times 1.4 = 2.1X$)
Reads with datacenter failures (2.1X)
### Haystack v/s f4 2.8 v/s f4 2.1

<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?

- How much space does f4 save

- f4 failure resilience
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**
  - 1 week: 350
  - 1 month: 150
  - 3 months: 80
  - 1 year: 0

- **Photo**
  - 80 Reads/Sec
It is warm, not cold

CDF

Day 2 4 Week 2 Month 2 Year 4 8

Haystack (50%)

F4 (50%)

HOT DATA

WARM DATA
f4 Performance: Most loaded disk in cluster

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

CDF of Read Responses

Latency (ms)

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