Freezing Exabytes of Data at Facebook’s Cold Storage

Kestutis Patiejunas (kestutip@fb.com)
### 1990 vs. 2014

<table>
<thead>
<tr>
<th>Spec</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form</td>
<td>3.5&quot;</td>
</tr>
<tr>
<td>Platters</td>
<td>5</td>
</tr>
<tr>
<td>Heads</td>
<td>9</td>
</tr>
<tr>
<td>Capacity</td>
<td>300MB</td>
</tr>
<tr>
<td>Interface</td>
<td>SCSI</td>
</tr>
<tr>
<td>Seek time</td>
<td>17ms</td>
</tr>
<tr>
<td>Data transfer rate</td>
<td>1 MB/sec</td>
</tr>
</tbody>
</table>

**Seagate 94171-327 (300MB)**

**iPhone 5 16 GB**
# History of Hard Drive data transfer rates

<table>
<thead>
<tr>
<th>Manufacturer</th>
<th>Capacity</th>
<th>Transfer speed (MB/sec)</th>
<th>Time to read all data</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seagate</td>
<td>300MB</td>
<td>1</td>
<td>5 mins</td>
<td>1990</td>
</tr>
<tr>
<td>IBM</td>
<td>10GB</td>
<td>12</td>
<td>13 mins</td>
<td>1998</td>
</tr>
<tr>
<td>Seagate</td>
<td>750GB</td>
<td>72</td>
<td>3 hours</td>
<td>2006</td>
</tr>
<tr>
<td>Hitachi</td>
<td>1TB</td>
<td>85</td>
<td>3.2 hours</td>
<td>2007</td>
</tr>
<tr>
<td>WD/Seagate</td>
<td>4TB</td>
<td>100</td>
<td>11 hours</td>
<td>2012</td>
</tr>
<tr>
<td>Seagate</td>
<td>8TB</td>
<td>120</td>
<td>18 hours</td>
<td>2014</td>
</tr>
</tbody>
</table>
Tape Is Dead
Disk is Tape
Flash is Disk
RAM Locality is King

Tape Is Dead
Disk is Tape

- 1TB disks are available
- 10+ TB disks are predicted in 5 years
- Unit disk cost: ~$400 → ~$80

- But: ~5..15 hours to read (sequential)
- ~15..150 days to read (random)

- Need to treat most of disk as Cold-storage archive
Building Facebook
HDD Cold Storage

Distinct goals and principles
(otherwise we will get another HDFS)
Goals and non goals

1. **Durable**
2. **High efficiency**
3. **Reasonable availability**
4. **Scale**
5. **Support evolution**
6. **Gets better as it gets bigger**

1. Have low latency for write/read operations
2. Have high availability
3. Be efficient for small objects
4. Be efficient for the objects with short lifetime
Principles

#1. **Durability** comes from eliminating single points of failure and ability to recover full system out of the remaining portions.

#2. **High efficiency** is from batching and trading latency for the efficiency. We spend mostly on the storing the data and not the metadata.

#3. **Simplicity** leads to reliability. Trade complexity and features for simplicity, gain durability and reliability.

#4. **Handle failures** from the day one. Distributed systems fail even on the sunny day, we learn about the mistakes when we find that intended recovery doesn’t work.
Architecture from 36,000 feet
Facebook HDD Cold Storage – HW parts of the solution

1/3 The cost of conventional storage servers

1/5 The cost of conventional data centers
Software architecture when we started in 2013
Raid rebuild vs Distributed volume rebuild

1) Default (Current State): The rebuild cards are set at 30% rebuild rate priority, with this a typical 4TB drive take almost 20 days to rebuild.
2) Set the rebuild rate priority to 100%. By doing so, we can possibly increase the rebuild rate and get the volume in healthy state in ~10-12 days.
# Gets better as it gets bigger

<table>
<thead>
<tr>
<th>Number of racks (PB)</th>
<th>Capacity (PB)</th>
<th>Amount of data to read/write in 1h at 50%</th>
<th>PB in 24 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>52</td>
<td>0.1</td>
<td>1.2</td>
</tr>
<tr>
<td>90</td>
<td>156</td>
<td>0.2</td>
<td>3.7</td>
</tr>
<tr>
<td>200</td>
<td>346</td>
<td>0.3</td>
<td>8.2</td>
</tr>
<tr>
<td>500</td>
<td>865</td>
<td>0.9</td>
<td>20.6</td>
</tr>
<tr>
<td>1000</td>
<td>1730</td>
<td>1.7</td>
<td>41.2</td>
</tr>
<tr>
<td>2000</td>
<td>3460</td>
<td>3.4</td>
<td>82.4</td>
</tr>
</tbody>
</table>
The Real Software architecture after 9 months in production
Raw disk storage

Problem:
- Takes 12h to fill 4TB HDD
- XFS can be formatted/erased in ~1sec

Solution
- Custom Raw disk storage format
- Metadata stored in 3 places
- Metadata is distributed
- Have to do full disk overwrite to erase data
Is this good enough?

What if we had a simplest roof water leak?
Disgruntled employee?
Software bug?
Fire?
Storm?
Earthquake?
What if we use Reed-Solomon across datacenters?

<table>
<thead>
<tr>
<th>Metric</th>
<th>2 replicas</th>
<th>10/15 Reed solomon (3 datacenters)</th>
<th>Savings (percentage)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Storage</td>
<td>2.8</td>
<td>1.5</td>
<td>187%</td>
</tr>
<tr>
<td>Network required per DC</td>
<td>100%</td>
<td>50%</td>
<td>200%</td>
</tr>
<tr>
<td>Availability</td>
<td>99.998910%</td>
<td>99.99674%</td>
<td>1%</td>
</tr>
<tr>
<td>Downtime per year (minutes)</td>
<td>5.7</td>
<td>17.1</td>
<td>33%</td>
</tr>
</tbody>
</table>
Conclusion: trade between storage, network and CPU

Like RAID systems do
Like HDFS and similar systems do

Just do this at the datacenter level
(can mix Cold and Regular datacenters)
So was Jim Gray 100% right about the future?

Tape is Dead
Disk is Tape
Flash is Disk
RAM Locality is King

Jim Gray
Microsoft
December 2006

Tape is Dead
Disk is Tape

- 1TB disks are available
- 10+ TB disks are predicted in 5 years
- Unit disk cost: ~$400 → ~$80

But: ~5..15 hours to read (sequential)
~15..150 days to read (random)

Need to treat most of disk as Cold-storage archive
Questions and possibilities for mass storage industry

Hard drives:

• hit density wall with PMR – 1TB/platter
• adding more platters – 4-8TB
• adding SMR (only 15-20% increase)
• waiting for HAMR!
• going back to 5” factor?

Optical:

• 100GB/disc is cheap
• 300GB within 18 months
• 500GB 2-3 years
• Sony and Panasonic has 1TB/disc on the roadmap
Questions and possibilities for mass storage industry

**Hard drives:**

- Less demand from IOPS intensive workload (either shifting to SSD, or can’t use all of the capacity)
- Small demand from consumers for large capacity

**Optical:**

- 4k or 8k movies will need lots of storage
Facebook Blu-Ray storage rack
Facebook Blu-Ray storage rack
Facebook Blu-Ray storage rack
BluRay rack software/hardware components

- BluRay Storage rack boundary
- BluRay Rack agent
  - receive/sent data to network
  - read/write to DVD
  - call to Robotics for load/unload
- Head node computer
  - software fully controlled by Facebook
- Robotics control unit
  - software/firmware provided by HIT
- 12 BluRay Burners Readers
- SATA
- Data to DVD write/read path
- 2 x 10 gbps public Ethernet
- Private network
  - Commands only
- Ethernet
- Control commands and data to read or write from Cold storage service

12 BluRay Burners Readers

BluRay Storage rack boundary

Control commands and data to read or write from Cold storage service

BluRay Rack agent
receive/sent data to network
read/write to DVD
call to Robotics for load/unload

Head node computer
software fully controlled by Facebook

Robotics control unit
software/firmware provided by HIT
Details data write path into BluRay racks

1. Public Thrift interface:
   - Migrate to Blu-Ray

2. Management calls (inventory, etc)

3. Blu-Ray Pool Manager
   - Blu-Ray racks
   - Empty/Full
   - Media
   - Volumes
   - Files
   - Blocks
   - Single CDB for one deployment
   - Sharded per user per bucket
   - Blu-Ray racks

4. Thrift interface:
   - Get portion of Migration Plan
   - FindVolumeForWrite

5. Calls to SN to read Chunk #1, #2, .. #N (20/28)

6. Calls to Blu-Ray rack to write Chunk #1, #2, .. #N (maybe 4/5)

7. Update DB with inventory

Thrift interface:
- Write chunk
- Read Chunk
- Verify Chunk
- Get inventory

Thrift interface:
- Write chunk
- Read Chunk
- Verify Chunk
- Get inventory

StorageNode #1

StorageNode #N

Cold Storage Knox nodes

Blu-Ray Data Migrator Worker #1

Data: RS decoded from Cold Storage - RS encoded for Blu-Ray nodes

Data Migration Plane

Get portion of Migration Plan
Work on moving data
Update with info about migration completed

Blu-Ray Data Migrator Worker #N

Blu-Ray Node #1

Blu-Ray Storage racks

Blu-Ray Pool Manager

Update DB with inventory

CreateFile
CreateBlock

Management calls (inventory, etc)

Blu-Ray racks
Empty/Full
Media
Volumes
Files
Blocks

Sharded per user per bucket

Single CDB for one deployment

Thrift interface:
- API Similar to HDD storage nodes

Thrift interface:
- API Similar to HDD storage nodes

Calls to Blu-Ray rack to write Chunk #1, #2, .. #N (maybe 4/5)

Calls to SN to read Chunk #1, #2, .. #N (20/28)

Get portion of Migration Plan, Work on moving data
Update with info about migration completed
Limits of BluRay storage, pros and cons vs HDD storage

1. **Load time**
   Time to load media for read/write ~30s – loadings should be amortized for reading/writing big data sets (full discs).

2. **IO speed**
   Current time to read/write 100GB disc is 1.3 hours (about 5-6x longer than on HDDs).

3. **Small fraction of active storage**
   BluRay rack has only 12 burners or **1.2TB of active storage**. HDD rack has 32 active HDDs or **128TB of active storage**. Write/read strategy is to cluster the data across the time and spread across multiple racks.

4. **Efficiency and longevity**
   Optical has big edge vs. HDD
Conclusions on Cold Storage

When data amounts are massive – efficiency is important

Specialization allows to achieve efficiency

If we are approaching the end of Moore’s and Kryder’s laws which of the storage media has more iterations left: silicon, magnetic or optical?

If we can’t see the future can we hedge our bets and how far we can push unexplored technologies to extract extra efficiency?