facebook

SYSTEM DESIGN

About Facebook

- Most used online social network in the world.
- 2.91 billion monthly users as of Feb 14, 2022. (Statista)
- Every minute: 317000 statuses are updated, 147000 photos are uploaded
- Facebook users generate 8 billion video views per day on average, 20% are live broadcast
- For perspective of the scale, Facebook's 6 hour downtime in Oct 2021 cost them 100 million US dollars (Fortune), 47.2 billion loss in market cap
- Main challenge is to keep the website online and functional
- Facebook operates 18 data centre campuses worldwide, 16 in United States, 1 in Ireland (datacenters.fb.com)

Facebook's Tech Stack overview

What technology Facebook uses to solve the problem of scalability

- Open Source Technologies
  - **Memcached** – distributed memory caching system, caching layer used between web servers and MySQL servers
  - **Scribe** – distributed queuing based logging system for handling logging at scale (several petabytes per hour)
  - **Varnish Cache** - HTTP accelerator, for load balancing and content caching
  - The **LAMP** Stack (Linux, Apache, MySQL, PHP)

Facebook's Tech Stack overview

What technology Facebook uses to solve the problem of scalability

- Personalized in-house developed systems
  - **Haystack** - highly scalable object store for storing billions of photos
  - **HipHop VM** – converts PHP code into C++ code for better performance
  - **BigPipe** – dynamic web page serving system to accelerate page rendering, divides web page into pagelets for optimal performance
  - **Thrift** – Cross-language framework, allows different languages to communicate, business logic exposed as services
  - **React** – front-end JavaScript library for building web and mobile user interfaces
  - **Gatekeeper** – software engineering system to get quick feature feedback and release feature to production for specific users, also involves "dark launches"
  - **XHProf** – Live performance monitoring system of PHP environment in production
  - **MyRocks** – Facebook first developed and then integrated it with MySQL storage engine, previously it was InnoDB
  - **TAO** (The Association and Objects)

Facebook's High-level Architecture

How Facebook solves the problem of scalability and reliability

- News feed server
- Photos and videos server
- Database server
    - Persistence layer: MySQL, Memcached, MyRocks

Scribe – Distributed queuing system

- Handles processing, storing and serving of logs
- Volume of logs – several petabytes per hour
- Low latency and high throughput

News Feed

- First thing users can see when they visit facebook from browser or mobile application
- Collection of posts, photos , comments of all the friends of the user and then rank it by relevancy.
- Over billions of users simultaneously visits their feeds.
- All this data is distributed across data centers
- Content is tailored to each user which nececciates dynamic data loading
- **How such distributed data is rendered at such a scale ?**

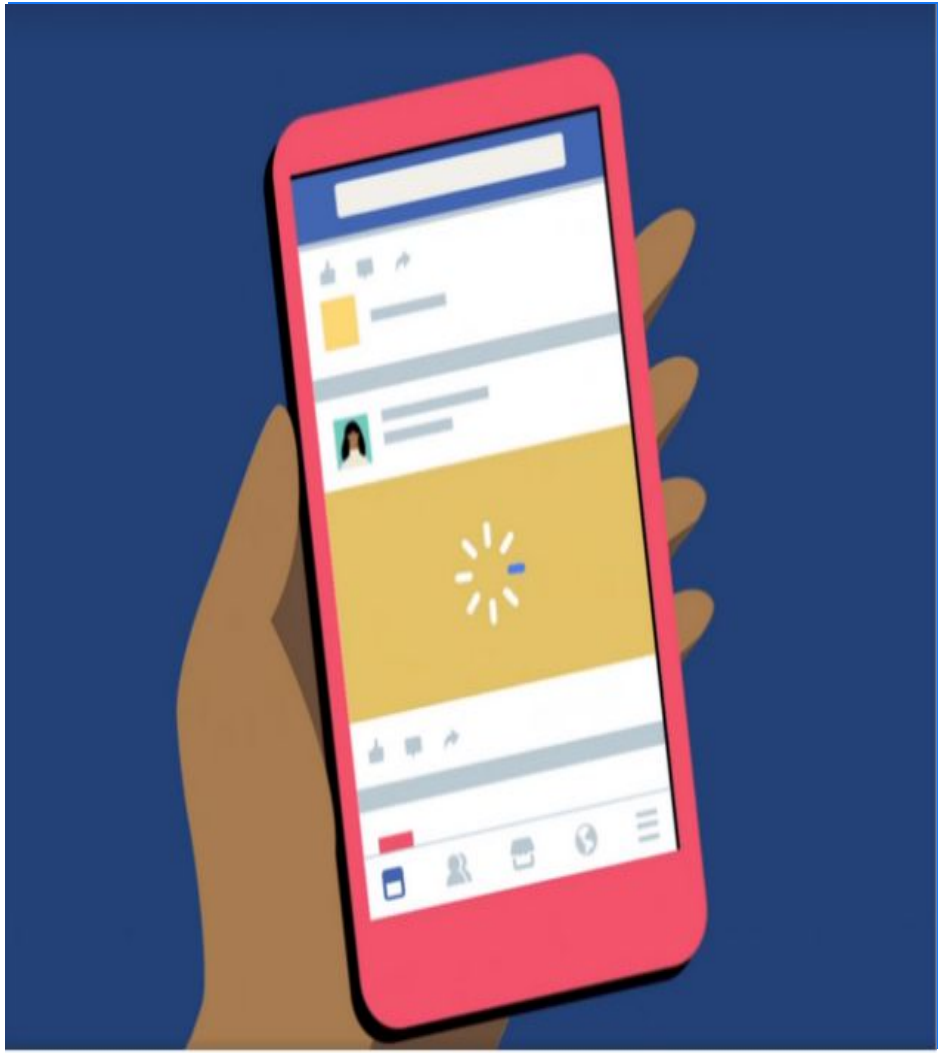

Photo Storage

Facebook classifies photos into three categories, 'hot', 'warm' and 'cold' photos, and uses different mechanisms to process these images:

- **Hot: Popular, a lot of views (approx. 90% of views)**

  CDN (Content Distribution Network)

- **Warm: Somewhat popular, but still a lot of views in aggregate**

  Haystack (Facebook has designed its own storage called Haystack)

- **Cold: Unpopular, occasional views**

  f4 (It is an “archival” storage designed by Facebook)

CDN

- **What is CDN**

  1. CDN is a content delivery network;
  2. A CDN is a cache, not a permanent store;
  3. Content providers are CDN customers;

- **Pros and cons of CDN**

  - Pros: Very good performance
  - Cons: no reliability guarantee

- **How does the CDN work?**

  - CDN company (e.g., Akamai) installs thousands of servers throughout Internet (In large datacenters close to users)
  - CDN replicates customers' content
  - When provider updates content, CDN updates servers

Haystack

- **What is Haystack**
  - Designed for performance and reliability
  - "Default" photo storage
- **Haystack Directory**
  - Helps the URL construction for an image
  - Logical & physical volumes
- **Pros and cons of Haystack**
  - Pros: Good throughput and reliability;
  - Cons: Somewhat inefficient use of storage space (mainly due to replication).
- **Haystack Cache & Store**
  - Haystack cache

f4

- **f4's Replication**
  - (n, k) Reed-Solomon code
  - Parity example: XOR
- **f4: Single Datacenter**
  - Within a single data center, (14, 10) Reed-Solomon code
  - Distribute blocks across different racks
- **f4: Cross-Datacenter**
  - Additional parity block
  - Overall average space usage per block: 2.1X
  - With 2.1X space usage,

Content Storage

- **Architecture**
- **To served by the follwing steps**
  1. Dedicated webservers,
  2. Scribe–Hadoop Clusters
  3. Hive–Hadoop
  4. Mysql

- **Distributed systems components**
  1. Two main components - Hadoop
     - Map–Reduce
     - Hadoop Distributed File System (HDFS)
  2. Master nodes - Hadoop

     Hadoop consists of multiple master nodes to avoid single point of failure in any environment.
  3. The elements of master node
     - Job Tracker
     - Task tracker
     - Name node (NN)
     - Data Node (DN)
     - Worker Nodes

- **Map Reduce (M-R)**
  1. Index any data comes from HDFS and being divided into blocks.
  2. Submit the M-R Job and its details to the Job tracker.
  3. Mapper process data blocks and generates a list of key value pairs.
  4. M-R merge list of key value pairs to generate final results.

- **HDFS**
  5. Run on low-cost hardware
  6. Highly fault-tolerance (as it supports block replication)
  7. Store very large data sets
  8. Reliability
  9. High bandwidth
  10. Ability to dynamically scale

Hadoop and Hive

In Facebook Hive is a data warehouse infrastructure built on top of Hadoop technology.

Role

1. **Easy data summarization;**
2. **Heavily reporting;**
3. **Adhoc querying;**
4. **Analysis of large datasets data stored;**
5. **HiveQL.**

   A simple query language called HiveQL which is based on SQL and which enables users familiar with SQL to query this data.

Apache HBase

Facebook messaging system by the support of Apache HBase which is a database-like layer built on Hadoop designed to support billions of messages per day.

Memcahed servers

Facebook, let Hadoop performing a random access workloads that provides low latency access to HDFS, by using a combination of large clusters of MySQL databases and caching tiers built using memcached ,that will be support a better in performance while all results from Hadoop are directed to MySQL or memcached for consumption by the web tier side.