Google File System

Architecture

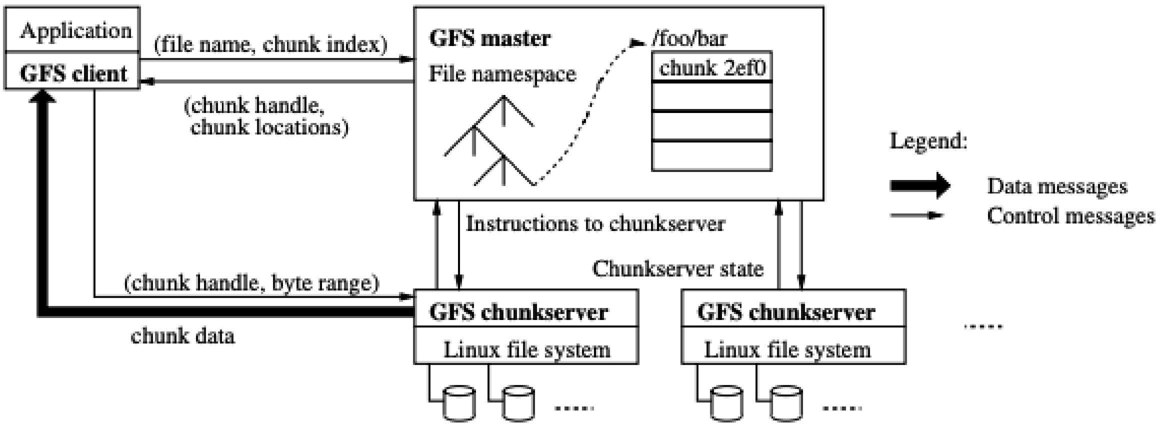

Figure 1: GFS Architecture

A GFS cluster consists of a single master and multiple chunkservers and is accessed by multiple clients.

The 4 steps illustrate a simple read operation. In step 1, the client asks the master for metadata: where are the replicas and locations where my data is stored? The metadata is returned in step 2, and then in step 3, the client asks for one of the replicas (usually the closest one) for the actual data, which is transferred in step 4.

Every client has to interact with the master. If the master started forwarding data to clients, bandwidth would be used up quickly, and the system couldn't serve many requests. By returning metadata, interactions with the master are quick, while interactions with chunkservers are longer (which is fine, since clients are probably talking to different chunkservers).

Features

- Namespace management and locking.
- Fault tolerance.
- Reduced client and master interaction because of large chunk server size.
- High availability.
- Critical data replication.
- Automatic and efficient data recovery.
- High aggregate throughput.

Data Flow

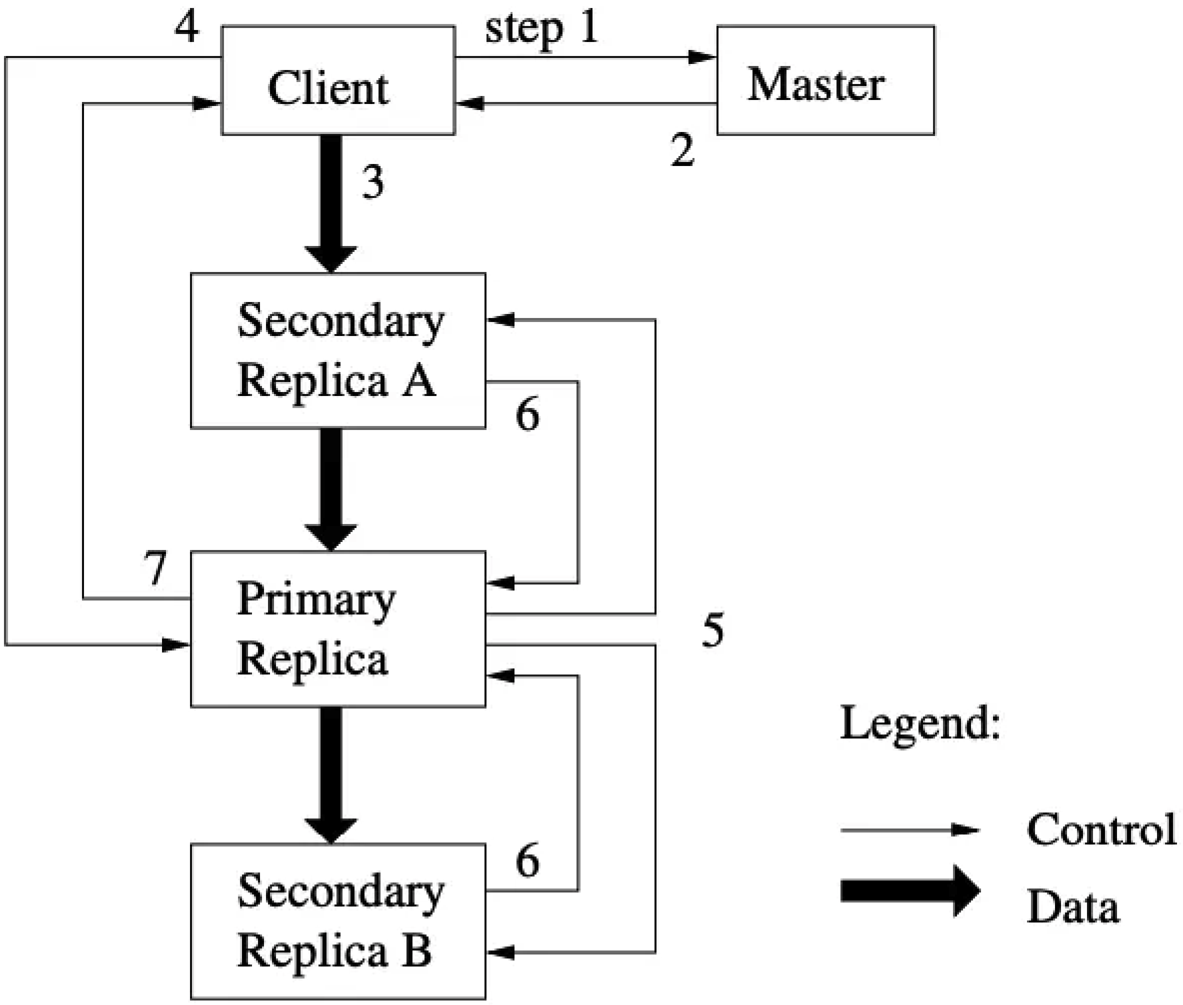

Figure 2: Write Control and Data Flow

Each piece of data must be replicated a configurable number of times (by default, 3) for availability. The diagram below shows what happens when a client writes data.

1. The client asks the master to write data.
2. The master responds with replica locations where the client can write.
3. The client finds the closest replica and starts forwarding data. (When a replica starts receiving data, it immediately forwards it to the next closest replica to speed up the process. This continues until all replicas have the data.)
4. The client asks the primary replica to commit the data.
5. The primary commits and asks the secondaries to do the same. If multiple changes happen concurrently, it also sends the order to commit changes.
6. The secondary replicas respond to the primary.
7. The primary responds to the client and reports errors.