Siyuan Liu

Distributed Systems

Quick Book

June 4, 2016

Springer

Contents

1

Distributed File Systems

1.1 Problems and Goals

When you want to have a file syustem and manage files, you need to consider following problems:

- Naming: allow uses to find files with a human-friendly name.
- Accessing: create, delete, read, write, append
- Physical Allocation
- Security and Protection: ensure privacy
- Resource Administration: enforce quotas and implement priorities.

Besides the problems in mind when design a DFS (actually general file system), we have some goals for DFS. **The big difference isn't what it does but the environment in which it lives.** A distributed file system typically operates in an environment where the data may be spread out across many, many hosts on a network, and the users of the system may be equally distributed.

- Coordinate file systems on machines.
- Hide the existence of distributed file system from the user.

Why we need a DFS. Actually this problem can be generalized into a problem of distributed system.

- More storage
- More fault tolerance
- Users are distributed who need to access file system from many places.

1.2 Operations

1.2.1 Unit of Transmission

There is a lot of data movement across the network, **how much do we move one time?** There are two intuitive answers to this question: whole files and blocks.

- File: Only the users know how to use the data in one file. File systems do not need to get known how a file is organized.
- Block: Reduce the payload for one operation. When the first block of the file arrives, the user can start to operate it.

1.2.2 Implementation Idea

Caching

Cache is the same usage. Cache data and reduce access to the servers. There are two kinds of caching in distributed systems:

- Cache and Validate Approach: ask servers that if the data is newest. This approach is used in *NFS*.
- Callback: if servers have some modification for files, they inform the users. This approach is used in *AFS* and *Coda*.

1.3 Coda

1.3.1 Disconnected Operations: cache and write conflict

hoard daemon.

keep certain files in the client's cache, requesting them as necessary, just in case the client should later find itself unable to communicate with the server. Cache and validate. This ensure the users get access the data efficiently.

write conflict.

keep a version number. Before a client writes a file to the server, it checks the version of the file on the server.

- If that version number matches the version number of the file that the client read before the write, the client is safe and can send the new version of the file. The server can then increment the version number.
- If the version number has increased, the client missed a callback promise. Then users must take care of the conflicts.

1.3.2 Replication: Volume Storage Group

This part is discussed in the Replication chapter. There are some topics for this.

First, **how it requests a file?**

1. It asks all replicas for their version number.
2. It asks the replica with the greatest version number for the file.
3. If there is a conflict, the client can direct the servers to update or inform them of the conflict.

Second, **how it writes a file?**

1. The clients sends the file to all servers, along with the original CVV.
2. Each server increments its entry in the file's CVV and ACKS the client.
3. The client merges the entries from all the servers and sends the new CVV back to each server.

Third, **what if partition happens?** If one or more servers fail, the client cannot contact the servers. The collections of volume servers that the client can communicate with is known as *Available Volume Storage Group*.

If the network is partitioned, Coda will still work, but generate some conflicts and inconsistency. (As is said in Chapter Replication.)

If the partitioned or failed servers become accessible, the files need to be updated. The client needs to check VVV or CVV for conflicts. If there is a conflict, the client drops all callbacks in the volume, because the servers should have all the callbacks.

Conflicts

There are several kinds of conflicts appearing in this problem.

- read before write. When the client decide to write the updates to servers, it needs to check if the update reads before write. Otherwise, it may miss some callbacks. Situation could be that the client is broken down for a while and misses some callbacks.
- server consistency. After the updates written back to servers and the client getting the CVVs from servers, the client may find conflict in the CVVs. Situation could be that there is concurrent operation on the files.
- partitioned network recovered. Replication inconsistency conflict, users need to take actions.

1.3.3 Weakly Connected Mode

If a client finds itself with a limited connection to the servers, it will pick a server and send the update. This server will propagate the update to other servers.

1.4 File System Interface

There are some file systems that are designed at user level, rather than in the kernel.

1.4.1 MogileFS

Difference

Consider the file systems where files are not changed by users. Users are only allowed to upload and download the files.

Method

There are some key idea behind the implementation:

- RAID. Replication.
- Namespaces, rather than directory tree. In different applications. they can use a same MogileFS, but they can use different name.

Because there is little change happened, locks will be rarely used. It makes a lot of work efficient.

1.4.2 HDFS

Difference

HDFS supports the computation of Hadoop.

Idea

There are some key idea behind the implementation:

- The data needs to be very heavily distributed.
- The data needs to be local to each other, so the system requires location-awareness.
- The system needs to be implemented in a portable way at the user-level, to be operated at scale.

2

Processor Allocation and Process Migration

2.1 Problems and Goals

Scheduling a System

The discussion involves

- pick a processor to run a process
- move a process from one processor to another processor

Process Migration

The discussion involves

- decide a process that should be migrated.
- select a new host for the process
- migrate the resource of the process

The migration here is talking about process. The threads of one process should be migrated together. We do not discuss that the threads are dispatched on different processors.

2.2 Processor Allocation

You need to be careful when you want to migrate a process. There is a difference between *remote execution* and *processor allocation*. People may not know that their local processes have been migrated to a remote processor and executed remotely.

2.2.1 Approaches

A Centralized Approach: Up-Down

There are CPU consumers and suppliers existing in the network. We give credits to the suppliers and take credits from the consumers. The hosts with more credits can have high probability to acquire a CPU.

Hierarchical Approach

Goal

Hierarchical Approach can reduce the communication and information across the network in order to balance the load.

Method

There are workers on the leaves, and managers on the inner nodes. If a worker gets too much work or too little work, it will inform the manager above.

- The manager will use the information to try a shift between workers of it.
- The manager meets the limit of its quotas, then tell the directors who try to balance the load among the managers.

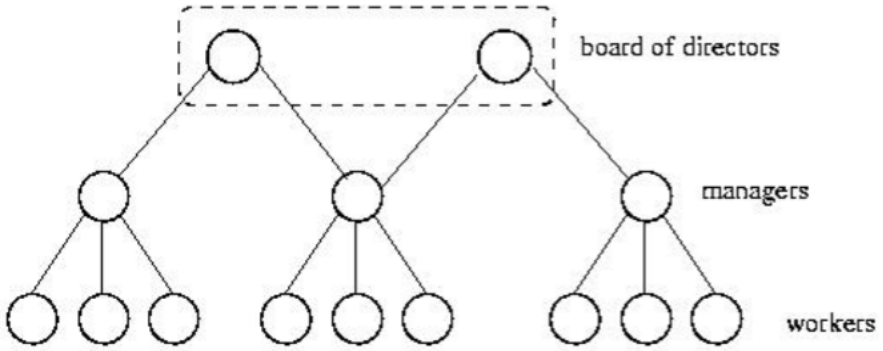

Fig. 2.1. Hierarchical Processor Management

Loads

We discussed the method for balance loads above that the workers gets too much work or too little work, it will inform managers. But how it communicate the information with managers and workers.

1. Yell out. When a worker meets the situation, it yells out to others. If it is idle, it may cause *thundering herds* that everyone wants to use it which makes it busy immediately.
2. Receiver Initiated. The idle processor asks around. But this approach leads to heavy communications.
3. Sender Initiated. The busy processor asks around to reduce load.
4. Hybrid Approach. These try to balance the costs of the above two approaches. They only "yell out" if they are substantially overworked or under worked.

2.3 Process Behavior and Scheduling

In general-purpose systems, recently started jobs are likely to be short lived, whereas long-running jobs are likely to keep running for a very long time.

2.4 Process Migration

Some ideas are here:

- Virtual Machine can make the migration easier, but it arises some network problems, such as IP address.
- If we want to migrate processes, we need to build a recoverable, portable communication layer (do not hack with TCP). This layer allows suspend, update and resume programs.
- There are some systems working for this scheduling: HTCondor, TORQUE.

3

Log and Checkpoint

3.1 Problems and Goals

As for **recovery**, there are three ways

- Replication. It is true that replication can make the system more secure, but it is expensive and also require a lot of synchronization work.
- Checkpoint. If the system make changes (progress), we can make checkpoints to store the system states. But the problem for checkpoint is that the systems need a **freeze** when backing up. However, we can try to find a consistent recovery line, which will be discussed later.
- Log. The classic technique is **Write Ahead Logging**. Events need to be written in the log first before executed.

For cooperation between checkpoint and log, we use logs between checkpoints. Recovery involves checkpointing and logging. Checkpoints store the state of process, logging involves recording the operations that produced the current state.

3.2 Sender-based Logging

Sender based logging is very important in those cases where receivers are thin or unreliable. In other words, we would want to log messages on the sender if the receiver does not have the resources to maintain the logs, or if the receiver is likely to fail.

Ensure that the order in both senders and receivers are correct.

Senders can play back the logging in the same order in which they were dispatched.It is difficult to order the messages from senders in the receiver side.

To ensure this, we follow the protocol:

- The sender logs a message and send it.
- The receiver gets the message and ACKs it with the time local to the receiver.
- The sender adds this timestamp in the ACK into that log entry.

Under this protocol, we can resend the message with the order recorded by the senders when dispatching. At the same time, when messages arrive the receivers, they can also order the messages with the receivers' time stamp.

Ensure the sender get the timestamped ACKs from receivers (make sure logs are complete)

Require the sender to send a ACK-ACK to the receiver before send a following message.

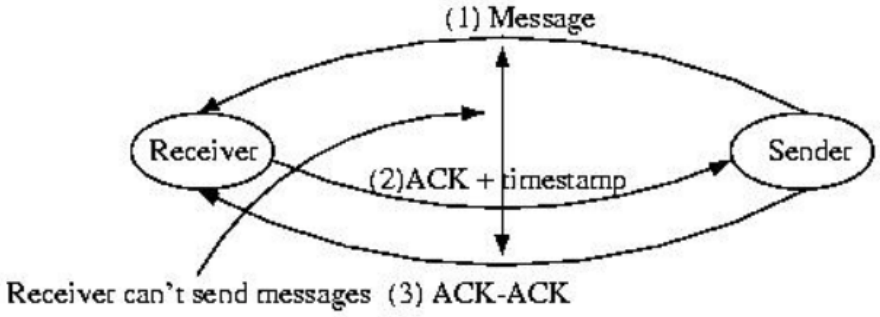

Fig. 3.1. ACK-ACK log.

Before the receiver gets the ACK-ACK for the first message, it cannot send messages to the sender. It make sure that we get the correct order in the receiver side.

3.3 Recovery from Failure

Recovery involves sending lots of history messages. **Duplicate messages** are the messages sent to other normal systems. **Orphan messages** are that after a rollback some systems may receive the messages that the recovering system does not remember sending. Rollback to a fail system may causes another system to rollback, which is known as **cascading rollbacks**. Eventually the systems will reach a state where they can move forward together, which is known as **recovery line**. After a rollback, a system may duplicate output, or request the input again, which is called **studdering**.

3.3.1 Incarnation Numbers

Incarnation is the period between checkpoints. Rebooting a system or restarting a cooperating process results in a new incarnation. We can number these incarnations. This number can be used to eliminate duplicate messages.

When a system is reincarnated, it sends a message to the cooperating systems informing them of the new incarnation number. The incarnation number is also send out with all messages. Therefore, the receivers can determine whether or not it is a duplicate message.

How the receiver handles the incarnation numbers?

- If the incarnation number of the message is less than the expected number, the message is a duplicate, so it should be discarded.
- If the incarnation number is in the message is greater than the expected number, the sender is recovering, so block accepting messages, until it informs us about its new incarnation number.
- If they are the same, accept the message.

3.3.2 Checkpoint: consistency

The checkpoints of all the servers are unnecessary consistent. Therefore, we need to find a maximum recovery line.

Interval Dependency Graph (IDG)

The graph is constructed by creating a node for each interval, and then connecting subsequent intervals on the same processor by constructing an edge from a predecessor to its successor. Then an edge is draw from each interval during which one or more messages were received to the interval or intervals during which the message(s) was or were sent.

Where to store the graph? Each processor keeps the nodes and edges that are associated with it.

How to find the recovery line? If there are some lost intervals, the other servers rollback to a interval that is independent of the lost intervals. And this rollback continues until there is a recovery line established.

Coordniated Checkpoints

We can decrease the rollbacks happened in the whole system, by coordinating checkpoints. Actually we only discussed how to use IDG, but not how to implicitly form a IDG. Here is the discussion.

There are two methods:

- Record message sequence number (the sender information). If the receiver gets a message, it send back a message to tell the senders to check if they checkpointed since the last time they sent the message. If not, the senders need to checkpoint in order to satisfy the dependency.

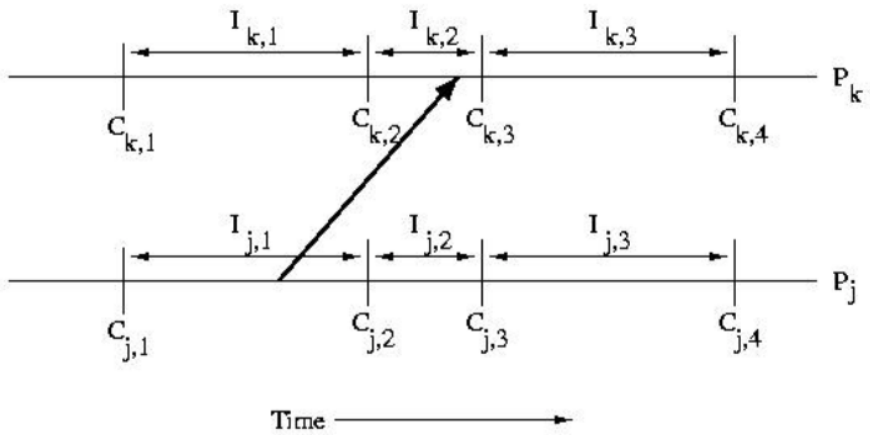

Fig. 3.2. IDG

- Synchronized clock. Each processor creates a checkpoint every T units of time.

3.4 Logging

3.4.1 Kinds of Logging

- Synchronous Logging. Logging before execution. It is expensive and slow.
- Asychronous Logging. Occasionally write the logs. Some messages may be logged, while others may not be logged.

3.4.2 One Approach For Asychronous Logging: GDM

First, the system maintains a **Global Dependency Matrix (GDM)**, each processor has a vector recording the interval numbers of processors it knows.

We can check if the matrix is consistent. The method is similar as before. You can check if the number on the processor's self is the greatest.

The method to find a new recovery line:

1. Get the previous recovery line and new updates after the recovery line.
2. Check for each update, if it can hold a consistent state for the system. If so, update the recovery line and go on checking.

3.4.3 Adaptive Logging

We do not need to log every message. It only needs to log those messages that have originated from processors that have taken checkpoints more recently than it has. Before there are new updates existing on the senders. If the receiver is ahead, it won't worry about it and do not need to checkpoint.

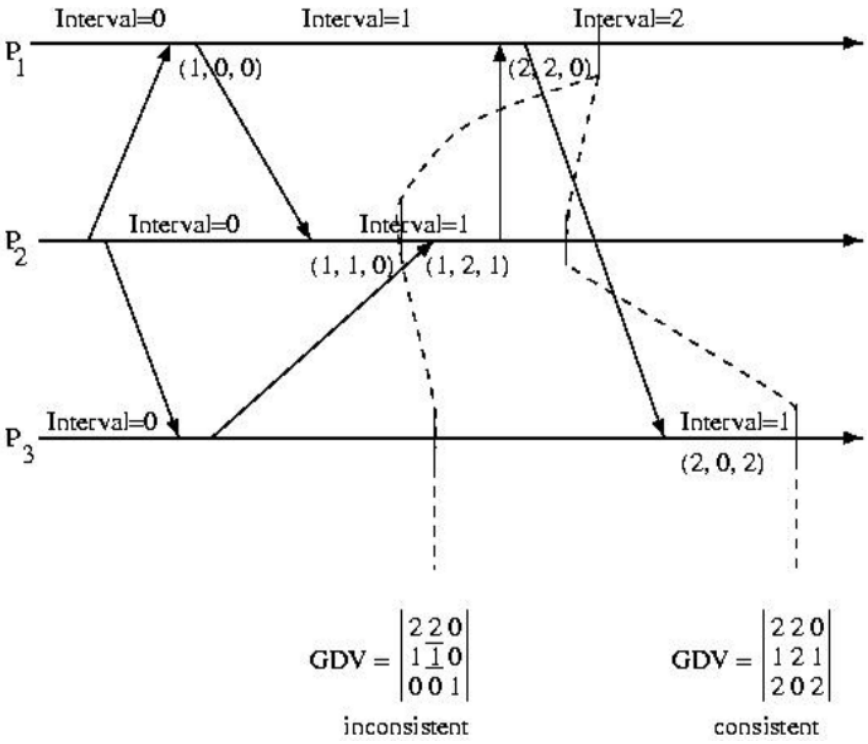
Interval=0
Interval=1
Interval=2
P1
(1, 0, 0)
(2, 2, 0)
Interval=0
Interval=1
P2
(1, 1, 0)
(1, 2, 1)
Interval=0
Interval=1
P3
(2, 0, 2)
GDV = 2 2 0 / 1 1 0 / 0 0 1
inconsistent
GDV = 2 2 0 / 1 2 1 / 2 0 2
consistent

Fig. 3.3. GDM

4

MapReduce and Hadoop

4.1 Problems and Goals

- How many maps
- How many reduces
- Locality, replication
- Combining Maps and Reduces: combiner only reduce the size of results from Maps, and the results are written in an intermediate file.
- Worker failure. When a Map worker dies, it needs to be re-executed from scratch. The reason for this is the results are stored on the Worker's local disk and are now inaccessible to Reduces. But, should a Reduce Worker fail, its results remain available in the global file system.
- Master failure. The master isn't scaled up. It is just one central Master. Like your desktop. Failures are years apart. And, checkpointing things will waste tons of time. Instead, if a computation times out, the program can just restart the computation a new, perhaps after checking the status of and with the Master.
- How many Map-Reduce phases is optimal? You can say it is one. But in reality, it is not possible and also not necessary.