Amin Arab

Matt Roberts

Nathan Whitaker

Vineet Amonkar

# Metadata Server Project (Team C)

This project was to implement three different mechanisms for storing a metadata server. The metadata server is responsible for implementing a filesystem on top of an existing object store. This means that reading and writing to files is handled by the object storage engine but the part that our team implemented managed, creating directories, creating files in the meteadata server, removing files, removing directories, and listing flies. (The corresponding commands of LS, MKDIR, TOUCH, RM, RMDIR in a unix based filesystem)

This project was implemented in three ways, with a Centralized server, with Distributed Servers similar to Ceph MDS, and using a general-purpose distributed database as the backend (Cassandra &Redis)

The additional components implemented were a lockserver, a client, and a client simulator. The lockserver was responsible for maintaining locks for the metadata stored in the general purpose database, (a similar component was implemented in Ceph for Cross MDS locking). The client makes it possible for people to enter commands for the servers to run. The client simulator reads in a file (containing a large or small directory structure tested with files up to 4gb) and then randomly walks down the parsed directory structure, creating directories and files. Optionally the client simulator will then walk down the created tree and delete directories and files to test those operations.

# Centralized and Ceph Server

Both the Centralized and Ceph servers are HTTP based and respond to common Directory manipulation commands through the /command Resource. The Metadata is stored in memory in a Tree Structure and operations submitted to the server manipulate that tree structure. Each Node in the Tree has a ReadWrite lock that ensures that concurrent requests don’t interfere with eachother, leaving the server in an invalid state. The only difference between the Centralized and Ceph Server is that the Ceph Server has an additional feature to enable the Tree to be “Partitioned” this partitioning takes a single directory and moves it to a different process (on the current server or on a different one). This is for scaling if the metadata information or the request load becomes larger than a single server can handle.

The Ceph Server has an additional component that responds to a custom TCP protocol, this is to ensure that if a server hosing a subpartition of the tree needs to take out read locks on directories hosted on a different server. This custom tcp protocol is based on the lockserver implemented for the general purpose database solution.

Here is the UML Diagram that outlines the important classes in the Ceph and Centralized solution.



Simplified Class Diagram

# Unix Solution

Unix Solution is the name given to the method of using a general purpose database to store all the directory information, as this is similar to the unix methodology of “everything is a file”. The information is stored as a text blob in the database with the filename as the key.

Since the component that stores the information is a distributed database, all of the logic for interacting with the metadata had to be programmed into the client, additionally there’s nothing stopping anyone with access to the client from maliciously modifying the data in the database.

Additionally the lockserver could potentially become a bottleneck as it has to respond to all requests, with a single thread being taken out per request.   


Performance

Centralized Server

To measure the performance of the centralized server,bla bla

Two Clients

Depth 500

Times: 10000

Prob 0.25

Client1

LS: DoubleSummaryStatistics{count=74, sum=127.000000, min=1.000000, average=1.716216, max=5.000000}

MKDIR: DoubleSummaryStatistics{count=24099, sum=28597.000000, min=1.000000, average=1.186647, max=1207.000000}

TOUCH: DoubleSummaryStatistics{count=8208, sum=9124.000000, min=0.000000, average=1.111598, max=13.000000}

RM: DoubleSummaryStatistics{count=11, sum=16.000000, min=1.000000, average=1.454545, max=2.000000}

RMDIR: DoubleSummaryStatistics{count=63, sum=103.000000, min=1.000000, average=1.634921, max=4.000000}

Client 2:

LS: DoubleSummaryStatistics{count=73, sum=126.000000, min=1.000000, average=1.726027, max=2.000000}

MKDIR: DoubleSummaryStatistics{count=24003, sum=26671.000000, min=1.000000, average=1.111153, max=240.000000}

TOUCH: DoubleSummaryStatistics{count=8232, sum=10879.000000, min=0.000000, average=1.321550, max=1920.000000}

RM: DoubleSummaryStatistics{count=19, sum=35.000000, min=1.000000, average=1.842105, max=2.000000}

RMDIR: DoubleSummaryStatistics{count=54, sum=98.000000, min=1.000000, average=1.814815, max=2.000000}

Four Clients

Depth 500

Times: 10000

Prob 0.7

Client 1

LS: DoubleSummaryStatistics{count=51, sum=70.000000, min=1.000000, average=1.372549, max=3.000000}

MKDIR: DoubleSummaryStatistics{count=23963, sum=26771.000000, min=0.000000, average=1.117181, max=1590.000000}

TOUCH: DoubleSummaryStatistics{count=8240, sum=8469.000000, min=0.000000, average=1.027791, max=12.000000}

RM: DoubleSummaryStatistics{count=6, sum=6.000000, min=1.000000, average=1.000000, max=1.000000}

RMDIR: DoubleSummaryStatistics{count=45, sum=70.000000, min=1.000000, average=1.555556, max=7.000000}

Client 2

LS: DoubleSummaryStatistics{count=31, sum=62.000000, min=2.000000, average=2.000000, max=2.000000}

MKDIR: DoubleSummaryStatistics{count=24047, sum=27038.000000, min=0.000000, average=1.124381, max=1908.000000}

TOUCH: DoubleSummaryStatistics{count=8226, sum=8524.000000, min=0.000000, average=1.036227, max=12.000000}

RM: DoubleSummaryStatistics{count=3, sum=6.000000, min=2.000000, average=2.000000, max=2.000000}

RMDIR: DoubleSummaryStatistics{count=28, sum=55.000000, min=1.000000, average=1.964286, max=2.000000}

Client 3

LS: DoubleSummaryStatistics{count=57, sum=83.000000, min=1.000000, average=1.456140, max=2.000000}

MKDIR: DoubleSummaryStatistics{count=24054, sum=26977.000000, min=0.000000, average=1.121518, max=1713.000000}

TOUCH: DoubleSummaryStatistics{count=8226, sum=8576.000000, min=0.000000, average=1.042548, max=11.000000}

RM: DoubleSummaryStatistics{count=10, sum=13.000000, min=1.000000, average=1.300000, max=2.000000}

RMDIR: DoubleSummaryStatistics{count=47, sum=70.000000, min=1.000000, average=1.489362, max=3.000000}

Client 4

LS: DoubleSummaryStatistics{count=57, sum=93.000000, min=1.000000, average=1.631579, max=2.000000}

MKDIR: DoubleSummaryStatistics{count=24054, sum=27322.000000, min=0.000000, average=1.135861, max=1816.000000}

TOUCH: DoubleSummaryStatistics{count=8192, sum=8577.000000, min=0.000000, average=1.046997, max=11.000000}

RM: DoubleSummaryStatistics{count=18, sum=33.000000, min=1.000000, average=1.833333, max=2.000000}

RMDIR: DoubleSummaryStatistics{count=39, sum=65.000000, min=1.000000, average=1.666667, max=2.000000}

Creative Clients Settings

Depth 500

Times:10000

Prob 0.9

Four separate clients were created to create the tree structure on one node.

Then data was partitioned and performance was measures using four new clients.

with a creative walk following with a destructive walk.

Ceph Server

Depth 1000

Times: 40000

Prob 0.9

Four clients

Four Partitions(The four most active directories)

Client 1

LS: DoubleSummaryStatistics{count=85, sum=89.000000, min=1.000000, average=1.047059, max=2.000000}

MKDIR: DoubleSummaryStatistics{count=95947, sum=111735.000000, min=0.000000, average=1.164549, max=205.000000}

TOUCH: DoubleSummaryStatistics{count=33078, sum=39359.000000, min=0.000000, average=1.189885, max=1447.000000}

RM: DoubleSummaryStatistics{count=18, sum=17.000000, min=0.000000, average=0.944444, max=1.000000}

RMDIR: DoubleSummaryStatistics{count=67, sum=71.000000, min=1.000000, average=1.059701, max=2.000000}

Client 2

LS: DoubleSummaryStatistics{count=57, sum=102.000000, min=1.000000, average=1.789474, max=2.000000}

MKDIR: DoubleSummaryStatistics{count=96047, sum=118196.000000, min=0.000000, average=1.230606, max=2302.000000}

TOUCH: DoubleSummaryStatistics{count=32988, sum=38912.000000, min=0.000000, average=1.179580, max=37.000000}

RM: DoubleSummaryStatistics{count=19, sum=34.000000, min=1.000000, average=1.789474, max=2.000000}

RMDIR: DoubleSummaryStatistics{count=38, sum=73.000000, min=1.000000, average=1.921053, max=3.000000}

Client 3

LS: DoubleSummaryStatistics{count=24, sum=45.000000, min=1.000000, average=1.875000, max=2.000000}

MKDIR: DoubleSummaryStatistics{count=96042, sum=116891.000000, min=0.000000, average=1.217082, max=209.000000}

TOUCH: DoubleSummaryStatistics{count=33047, sum=39219.000000, min=0.000000, average=1.186764, max=124.000000}

RM: DoubleSummaryStatistics{count=0, sum=0.000000, min=Infinity, average=0.000000, max=-Infinity}

RMDIR: DoubleSummaryStatistics{count=24, sum=49.000000, min=2.000000, average=2.041667, max=3.000000}

Cleint 4

LS: DoubleSummaryStatistics{count=42, sum=99.000000, min=2.000000, average=2.357143, max=4.000000}

MKDIR: DoubleSummaryStatistics{count=96419, sum=117978.000000, min=0.000000, average=1.223597, max=2136.000000}

TOUCH: DoubleSummaryStatistics{count=33002, sum=38547.000000, min=0.000000, average=1.168020, max=31.000000}

RM: DoubleSummaryStatistics{count=2, sum=3.000000, min=1.000000, average=1.500000, max=2.000000}

RMDIR: DoubleSummaryStatistics{count=40, sum=122.000000, min=2.000000, average=3.050000, max=13.000000}