**MP3 Report from Haorong Sun (Haorong4) and Haoyu Zhang (haoyuz3)**

MP2 is used as the failure detector.

The heartbeat will also carry synchronization information.

The system is built as a virtual ring structure.

Distributed File System:

There will be a master for the entire system. The master will decide where the replicas of file is located. And every time a file’s status is updated, the master will share the file information though the ring structure. Every member who receives the update request will compare the new information’s timestamps to the local ones, and merge incoming list with the local list, choosing the data with the newest timestamp. Then pushing it to all its neighbors. If all the incoming info is older or equal to local info, block the message.

If the current master fails, the system will start a new election. It will check the membership list and choose the member whose IP address has the highest hash value as the new master.

In order to be tolerate to three failures, every file will have four replicas, so that even if three of its replicas fail, the remaining replica are still able to do the re-replication.

The master process will also check if there are needs for re-replication, we set the master to check through its failure list for every 5 seconds. If there are new failures, the master will go through the file information list, and choose one remaining replica to send file copy to new positions.

We use TCP connection to transfer files. We transfer files in blocks of 1024KB. Each node has two threads listening for incoming connections. One thread listens for connections that query master, since every node may be a master. The other thread listens for connections, which comes from other nodes, that put or get files.

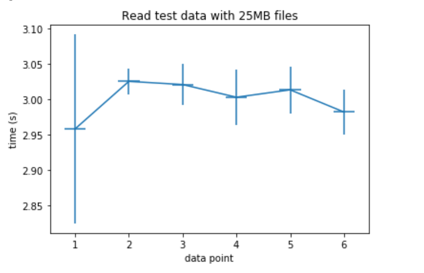
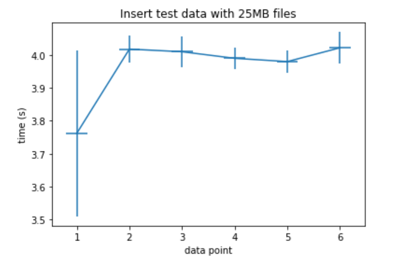
While debugging the system, MP1 is very helpful, since it could help us to gather all the file information from 10 vms to one terminal, saving lots of time and effert.

The following measurements are all measured using Bmon.

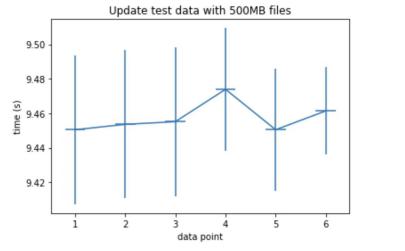
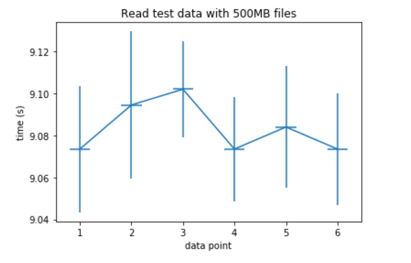
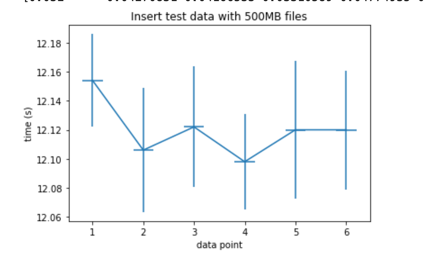
Measurement 1:

Upon a failure, the machine that receives the replica has a bandwidth of 28.68MBps. Since we check failure every 5 seconds, the time from failure to re-replication completion is around 6.5 s. However, the actual file transferring time is around 2 s.

Measurement 2:



the standard deviations for the first the data point are relatively larger than others, this might because the internet environment is not that stable during that measurement. Overall the insert time is around 4 seconds, the read and update is a little faster because we will use quorum. We will only visit 2 replicas for the read, and 3 replicas for update.



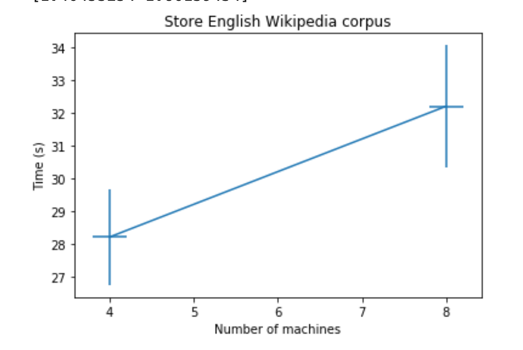
The overall behavior does not change much. The time for transferring files will increase while files size increases. However, the standard deviation doesn’t differ too much, since the network environment remains the same.

Measurement 3:

This is the plot for the time of storing the English Wikipedia corpus on 4-machine and 8-machine system:

4-machine Average: 28.22s Std: 1.465s

8-machine Average: 32.20s Std: 1.861s



The time increases a little for the 8-machine system is because the processing time on the communication between members is larger.