Machine Programming 3 – Simple Distributed File System

Yu Tao, Jie Yin
yutao2@illinois.edu; jiey3@illinois.edu

Design

Given the predefined fault tolerance is up to 2 VMs, the number of replica on SDFS is set to 3 (excluding the local file). The consistence level is Quorum to ensure that the latest value can be read. The Quorum is specified to be W=3 and R=1.

To write a file into SDFS, the hash ID of the file corresponding to its sdfsfilename is calculated. Like chord, the primary replica is stored on the first VM with node ID greater than the ID of the file. Two other replicas are stored on its two successors. So, once the "put" is issue, the client sends replicas to three nodes directly by TCP.

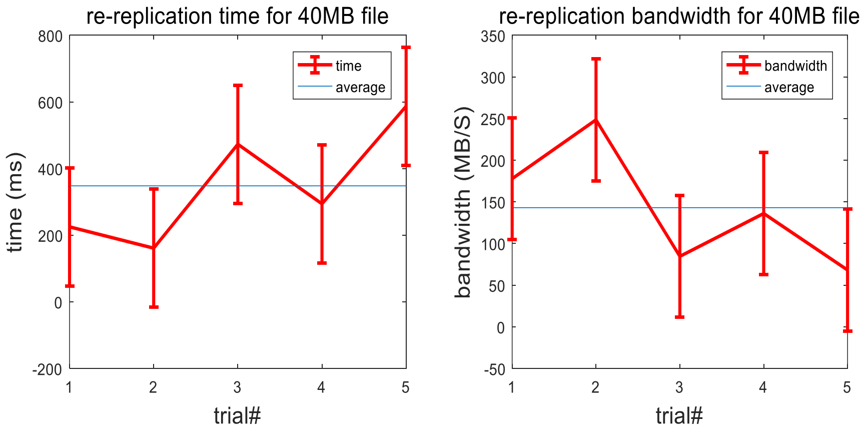
When the replica has been written to node, the node will multicast the acknowledgement with the sdfsfilename and timestamp so that every node in the group knows where the replica is stored. To read the file, the client fetch the replica from the node which stores the primary replica directly.

When some (up to 2) node(s) fails, the other nodes which store the same replica as the failed one can figure out the new node(s) to store the replica. They will send the request to that node, the node receives the replica if it does not have the replica. To delete a file, every node deletes the replica if exists.

For "put" command, the timestamp difference of the last update and current update is calculated. Confirmation is needed for back-to-back update within 60 seconds. Write will be rejected if the confirmation is not received within 30 seconds.

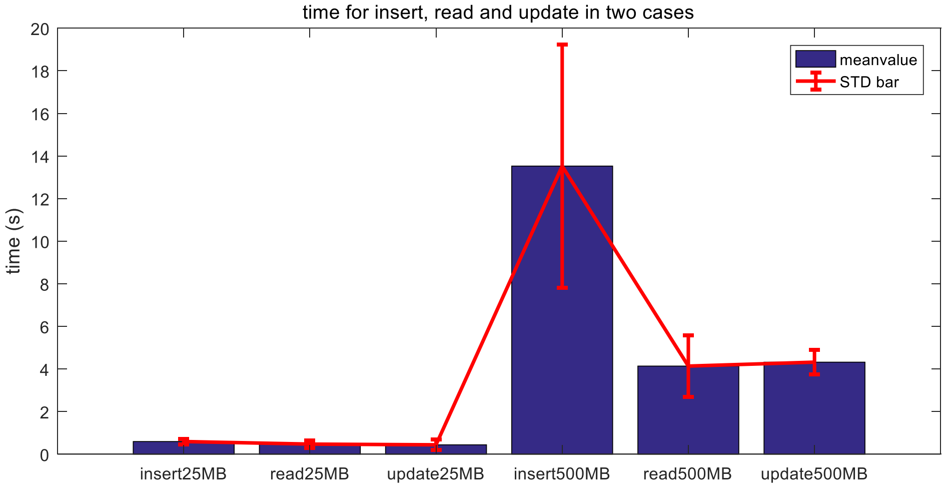
Measurement

a. **Re-replication time for a 40MB file upon one failure**



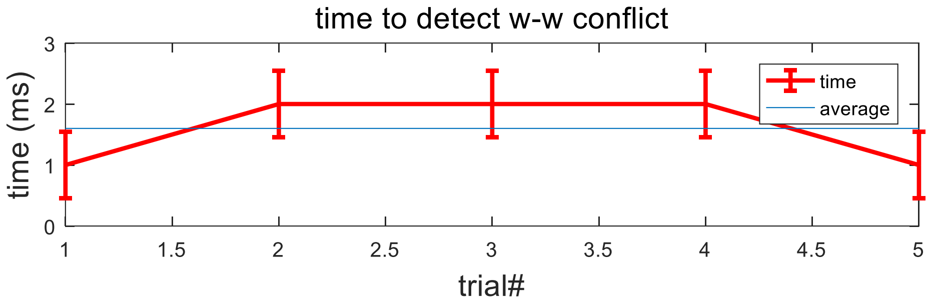
From the graph, it takes an average value of around 300ms to complete re-replication for a 40MB file when there is one failure. However, the value varies significantly due to the network condition, so the standard deviation is large.

b. times to insert, read, and update, file of size 25 MB, 500 MB



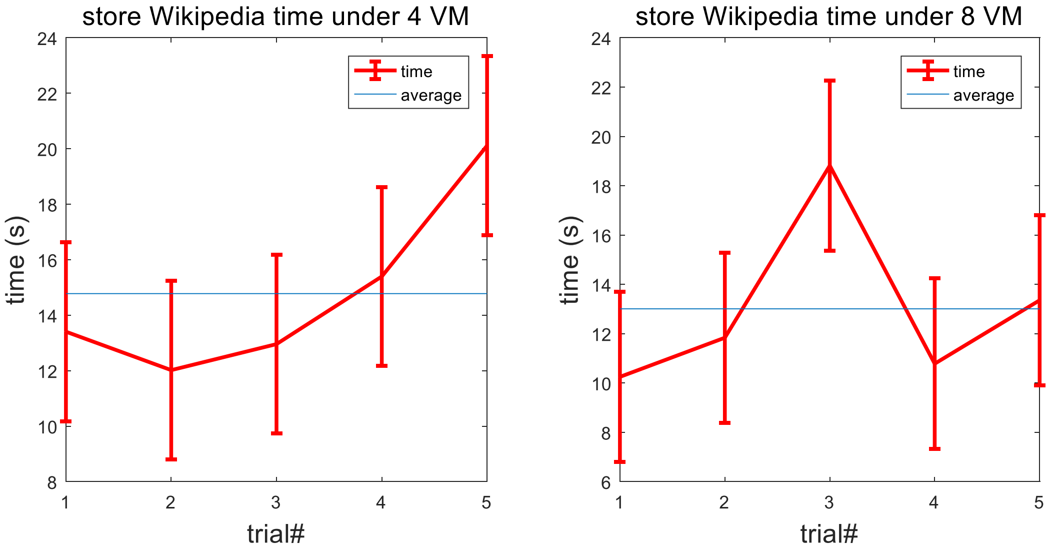
From the graph, manipulation on 25MB file is much faster than on 500MB file in SDFS, which is straightforward. When the file size is large, the insert performance varies significantly, which is reflected by its large standard deviation.

c. Time to detect write-write conflicts for two consecutive writes within 1 minute to the same file



The time to detect write-write conflicts within 1 minute is very fast (around an average of 1.5ms). This is because the detection is done locally, i.e. checking its own timestamp difference.

d. time to store the entire English Wikipedia corpus into SDFS with 4 machines and 8 machines



From the graph, storing a large file (1.13GB for Wikipedia) takes around 14 seconds. The number of VMs in the system does not influence the performance significantly. However, the performance is not stable due to the network condition, which is reflected by its large standard deviation.