MP1 Report Amaan Khan (amaanmk2) & Samaah Khan (skhan330)

Design & Algorithm

Every machine acts as a server and a client by having both run in parallel. Every machine is connected to every other machine using a TCP socket.

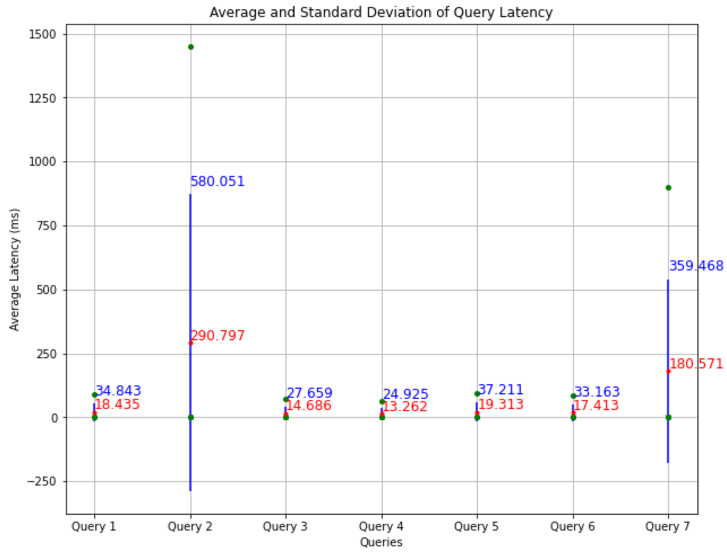
When a grep query was written by the user, the client on that machine sends the query to all actively connected peer machines (i.e., it is not sent to machines who's connection was disconnected). The server on each peer machine would read the query, check if the query is present in its local cache, retrieve the output from the cache if its present otherwise it executes the grep query locally, and then it sends back the grep output as a serialized struct containing the filename, output, number of lines, and execution latency. In the meantime, the client waits to receive the outputs from its connected peer machines, and once it does receive everything it prints it all out, including the execution latency of the time between the query was written to the time all the outputs were printed out to stdout.

Since frequent grep queries are slow and take time to execute, every machine caches the outputs of a grep query it executes to save time for subsequent calls to the same query. Every machine has an in-memory LRU cache where the key is the grep query and the value is the grep output struct. Also, we executed grep queries by having it execute the existing grep shell command.

Unit Tests (brief)

We wrote unit tests on the grep functionality on a single machine, as well as testing the grep functionality when ran on multiple VMs by running the program under TEST mode (see README for more details). Also tested the cache by checking that the speed with the cache must faster on subsequent runs of the same query. Also wrote unit tests on checking serialization/deserialization, and other helper functions.

Plot



Legend:
- Blue numbers = standard deviation for query i
- Red numbers = average latency for query i
- Green dots = latency of a trial for query i

This is the plot of 7 queries that we ran using 4 VMs and on 60mb data log files (the log files given to us for demoing). Query 2 and Query 7 were complex regular expressions (exact queries can be seen under `scripts/plot_data.ipynb`), which is why we believe the standard deviation was so high. The first time it was ran the latency was quite large, but in subsequent runs the latency was consistently small since it was pulling the output from the caches, explaining why the standard deviation was so high on those. The other queries, however, didn't have large standard deviations which we believe is because despite the result being cached for subsequent runs, the bottleneck for the latency was more so for sending the data across the network rather than the grep run or the cache retrieval itself. We know this because we displayed the latency for individual grep queries from each machine.