MP1 Report

Harsh Agarwal - *harsha4*
Maya Moy - *mjmoy2*

Design Algorithm:
The log file query system employs a client-server architecture. The client (Client.java) initiates queries, establishing socket connections with multiple server machines for parallel processing. Servers (Server.java) receive incoming queries and execute `grep` commands on their log files. The client maintains a list of server addresses, marking them as active or inactive. It selectively connects to active servers. Parallelism is achieved by deploying individual threads for each server, enhancing query execution efficiency and reducing latency.

Unit Tests:
In ServerTest.java, unit tests ensure the server effectively listens on a designated port and responds accurately to client requests. In QueryTestVM1234.java, client tests validate the system's correctness by verifying the proper identification of active and inactive servers, while also testing query functionality by evaluating pattern-based queries for accurate line accumulation. For query testing, VMs 1-4 run servers, and each log file is 60MB.

Average Query Latency:

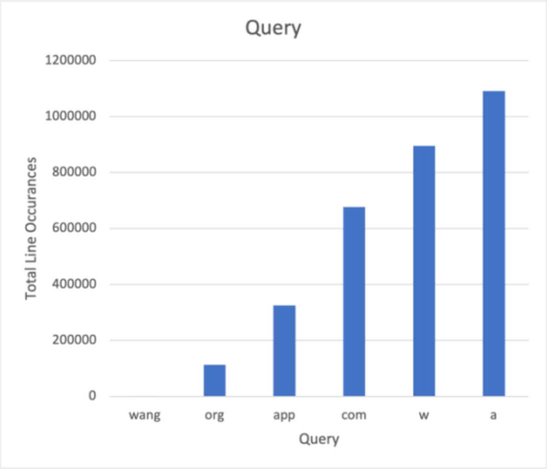

Figure 1 – Query input phrase vs line occurrences found

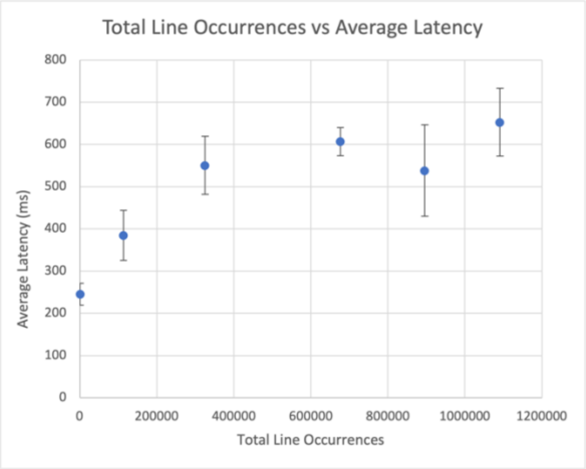

Figure 2 – Line occurrences vs average query latency

Figure 1 shows the queries run during QueryTestVM1234.java, and Figure 2 shows the results. Each data point in Figure 2 represents the average latency of 5 trials run with the same query. The error bars represent the standard deviation across the 5 trials. There appears to be a weak correlation between higher occurrences and higher latency. We expected a stronger correlation, however the standard deviation for the data point with around 900,000 occurrences is quite large, which could account for this data point not fitting the overall trend. In Figure 1, we see shorter query phrase correlates with higher total line occurrences. The Boyer-Moore algorithm used to run the grep command executes faster when given longer query phrases, and this likely plays a large role in determining latency. In this way, the higher latency seen with short phrases and higher total line occurrences makes sense.