**CS425, Fall 2018 MP 1 Report– Distributed Log Querier**

Group 42
Netid: yidanli2 siqi5

## Design

· Distributed Service Architecture

A server-client architecture adopting RPC over TCP has been built to implement distributed log querier. All machines serve as servers with different logs. When any one of machines executes a grep command, it functions as a client and ask servers to run grep command and return the results to the client, which displays matched lines in all log files across all servers.

· Concurrency and Fault Tolerance

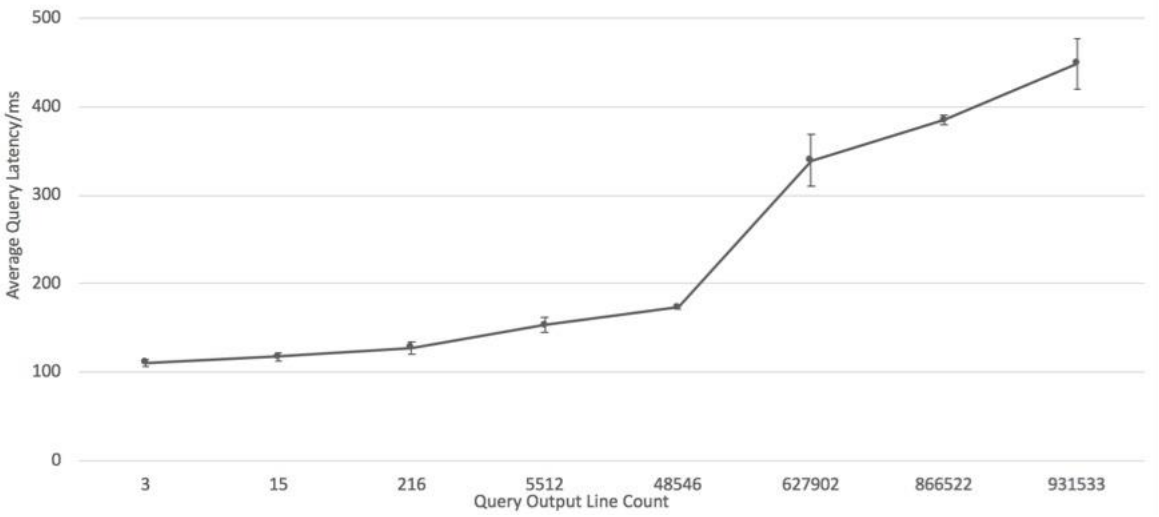
For every server, the client creates a goroutine to call grep commands on servers using RPC connection, which ensures concurrency. The servers will return the results to the client using channels. As for fault tolerance, since the connection between the client and servers is on different goroutines, when a server crashes, it will just return a message "could not connect to server #id" and have no impact on other goroutines.

## Unit Testing

In unit tests, we designed several patterns as input, run greps on distributed log querier program and compared the number of matched lines with expected line counts. We tested a comprehensive set of patterns including frequent, somewhat frequent and rare patterns, regular expressions, and those only occur in one/ part of/ all the log files. client_test.go verifies automatically the correctness of the results.

## Performance

With 4 machines each storing 100 MB log files, we run 9 grep queries, each 5 trials, and got the query output line count & average query latency plot.




In the plot, patterns with high frequency have higher query latency and the less frequent patterns have lower query latency. Query latency is composed of data transfer latency and pattern matching latency. In general, pattern matching time is irrelevant of pattern frequency because every line in all log files will be "scanned". However, the larger matched line count is, the longer time it will take for servers to return the output to the querying machine, thus the query latency is highly correlated with the output line count. The analysis is consistent with what the plot demonstrates.