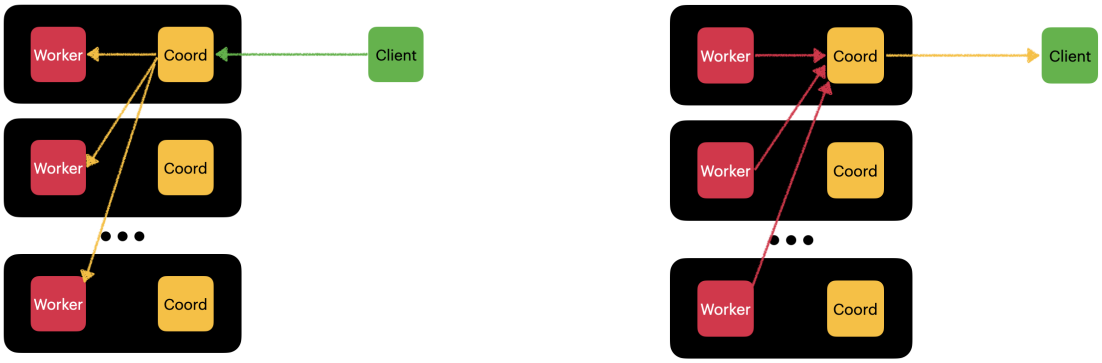
# CS425 MP1 – Distributed Log Querier

Group 04: Hsin-Yu Huang (hyhuang3), Yun-Liang Huang (ylh2)

## Design and Algorithms




We implement the distributed log querier in C++ with client-server architecture. We use socket to enable client-server communication. There are two services on the server side: Worker service and Coordinator service. Any Coordinator service in the cluster may be queried by a client program, and the Coordinator will then simultaneously query every Worker service in the cluster. After receiving response from all active Workers, the Coordinator collects logs and sums up line counts and sends them back to the Client. We design an abstraction layer above the socket to (de)serialize messages for each communication in order to send and receive massive amounts of data.

For executing queries on workers, we implement **string search**, **case insensitive search**, and **regular expression**. String search, case insensitive search are implemented using string find function in C++, while regular expression is executed by calling a linux command due to performance issues.

## Unit Test

We designed test cases for querying frequent, infrequent, and rare logs, and tested the regular expression grep feature. We check the cluster query result with local grep result by comparing the match line count.

## Average Query Latency (4 machines each store 60 MB log files with total line count: 1091212)

| Query type | Grep command | Match line | Avg time (s) | Standard deviation |
|---|---|---|---|---|
| Frequent | ./client grep "Intel" ".*.log" | 163628 | 1.174 | 0.075 |
| Somewhat frequent | ./client grep "allen" ".*.log" | 6110 | 0.115 | 0.011 |
| Rare | ./client grep "17/Aug/2022:19:09:09" ".*.log" | 1 | 0.102 | 0.002 |
| Frequent regex | ./client grep -regex "Mac OS.*Firefox" ".*.log" | 182057 | 1.286 | 0.053 |
| Somewhat frequent regex | ./client grep -regex "^124.*" ".*.log" | 4257 | 0.232 | 0.01 |
| Rare regex | ./client grep -regex "GET.*5023.*Mozilla.*Mac OS X 10_5_7.*Chrome" ".*.log" | 17 | 0.14 | 0.007 |

## Analysis

We experimented with six grep commands with querying frequent, somewhat frequent, and rare logs both with and without regex enabled. We anticipate that frequent logs should take longer time to process than rare ones and that regex queries should generally run slower. The outcome confirms our expectations. The standard deviation is lower when there are fewer match lines. We think that it might be because more match lines will take more socket packets to send the result. The communication between numerous sockets might cause the time difference for each time using the same grep command.