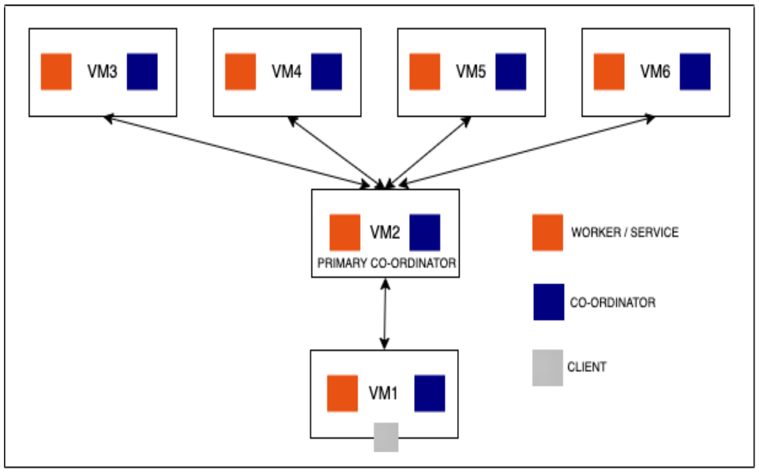
# CS425 - MP1 - Distributed Log Querier

Lavanya Ramkumar (lr15)
Shahid Ikram (shahidi3)

## Design:

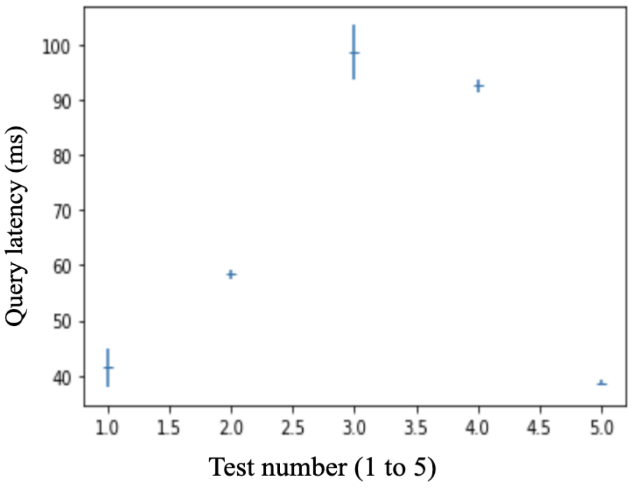


The client process, oblivious to the system, queries for logs from a set of machines. The client process first chooses a coordinator process to which the query request is to be forwarded to. Each machine in the cluster runs the coordinator and worker/service processes. If the connection to a chosen coordinator fails, the client randomly chooses another coordinator process until a connection succeeds. This ensures that even if there is one coordinator process running the client's request will be processed. The coordinator process then forwards the query request to the service processes who fetch logs pertaining to the query on the chunk of log file that resides on them and returns the results to the coordinator. The coordinator aggregates the results from all the service processes and returns the response back to the client.

## Unit tests:

| Test # | Query | # of matches | Exec Time (ms) |
|---|---|---|---|
| 1 | privacy | 154387 | 41.05 |
| 2 | http://www.burke.com/homepage.html | 3 | 61.31 |
| 3 | http:/* | 2709263 | 100.56 |
| 4 | \[(0?[1-9]\|[12][0-9]\|3[01])/Aug/([0-9]+(:[0-9]+)+) -[0-9]+] | 236668 | 103.01 |
| 5 | this query doesnt exist | 0 | 41.79 |

## Analysis:



Firstly, we observed that the latencies of a distributed grep was significantly better than running a grep over all the files on a single machine. Moreover, over the multiple runs of a query (10 runs) the variance seen in the latency numbers were quite low except for the test number 3 which might be due to presence of network congestion as in some cases we used multiple clients to issue queries simultaneously. There also may have been a failure in connecting to the coordinator node in the first attempt. The average latency was seen to be higher in the case of regex patterns which is expected. The average latency was quite less (~45ms) for queries that had abundant matches. The queries with infrequent matches expectedly had higher latency than a frequent occurring pattern as seen in Test number 2 vs Test number 1.