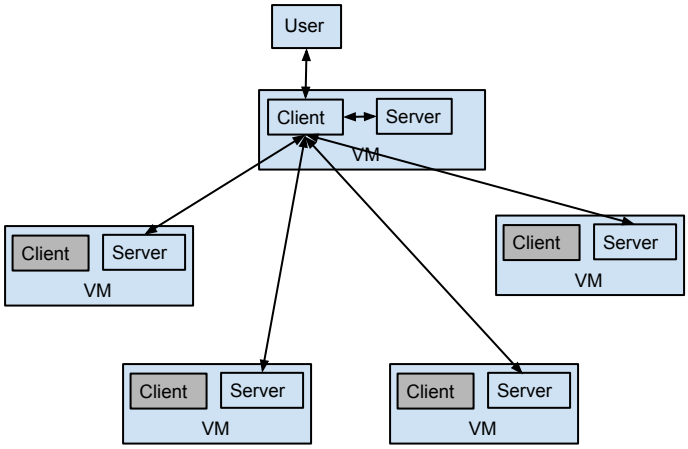
CS 425 Distributed Systems

MP 1 Report

Anish Shenoy (ashenoy3) & Omkar Lokhande (lokhand2)

For this MP, we had to write a program that allows you to query log files on a distributed system from any one of the machines. In our implementation, we setup servers on all machines which listen for incoming connections, and a client which runs on the querying machine, as explained in the following picture.



Since the data was distributed, it is faster to query the data at the source using the server and then send the results back to the client. The client uses a file that contains the server IP addresses and port numbers. Again, it is faster to perform the searching operation in parallel, so we have to connect to each server asynchronously. For this purpose, the client connects to individual servers on a separate thread, and then collects their output as it is received. On the server side, the server uses a wrapper class that runs the query using the system grep, and collects its output, which is relayed to the client using the socket connection. The client collects all the grep output and writes it to a user defined file.

Unit Tests

- Generates log files with frequent (70%), semifrequent (25%), and rare (5%) patterns.
- Checking wrapper class for nonexistent, frequent, semifrequent, and rare patterns - 4 tests
- Setup a single server - client, query frequent, semifrequent and rare patterns - 3 Tests
- Setup 4 servers and 1 client, query for frequent patterns with same logfile on all machines.
- Setup 4 servers and 1 client, query for a pattern which appears frequently, semifrequently, rarely or not at all in each server log.

Querying Latencies: Same 100 MB log on each machine, 4 machines in all

- Querying for 'anish' which appears in 70 % of the lines: 19.26 s
- Querying for 'omkar' which appears in 25 % of the lines: 7.19 s
- Querying for 'patrick' which appears in 5% of the lines: 1.80 s