# CS 425 MP1: Distributed Grep

Abhishek Verma (averma11) & Dominic Le (ddle2)

## OVERVIEW:

The distributed grep command can get a grep output from a distributed systems by running a grep command on log files which have been distributed amongst different machines.
The command uses C - socket communication and sends a grep command to multiple machines and gets the output from them, then concatenates them to give a final resultant grep output. The grep command can take any number of hosts to read log files from, and this is provided in the form of a host file. The command is parallelized and uses a multithreaded model to communicate with different servers at one time to reduce latency. Servers are also parallelized using threads to ensure efficiency. The grep command is run local to the log file to improve performance.

The distributed grep design includes three steps. One, the client makes a socket connection with the server, sending the grep command to the server to execute. Two, the server executes the grep command on a local file to produce an output. Three, the server sends back the grep output to the client via the client/server socket connection. The client in turn receives all output from each machine and displays a concatenated output. The client program takes a flag to determine if the output should be written into separate files for each VM or in a concatenated version in stdout. Servers must be running on target machines before the client calls distributed grep. Otherwise, no connection can be made and no output will be received. A general call to the client looks as follows:

```
./client FILE_FLAG COMMAND QUERY FILENAME
./client 1 grep hello machine.i.log   \\flag = 1 for output in files, 0 for stdout output
```

## UNIT TESTS:

Unit tests were constructed to ensure the robustness of our distributed grep implementation. Unique log files were constructed and placed on each VM and the client computer. Each test calls distributed grep with grep commands created to test frequently used expressions, odd characters, regular expressions, grep with flags, and other edge case phrases. For the unit tests, the client queries all ten virtual machines. If a server is not running or crashes during a test, the client is notified but the test is still considered to have passed. The client runs grep locally and diffs the local output to the output received from distributed grep.

## QUERY LATENCY:

Query latency varies greatly depending on the frequency at which the phrase appears in each log file. The average query time for a file which was 60 MB [600,000 lines] when:

Query for common phrase: 45.2 Seconds
Query for uncommon phrase: .4 Seconds
Common Occurrence: 599983 lines for 600000 line file
Uncommon Occurrence: 4008 lines for 600000 line file