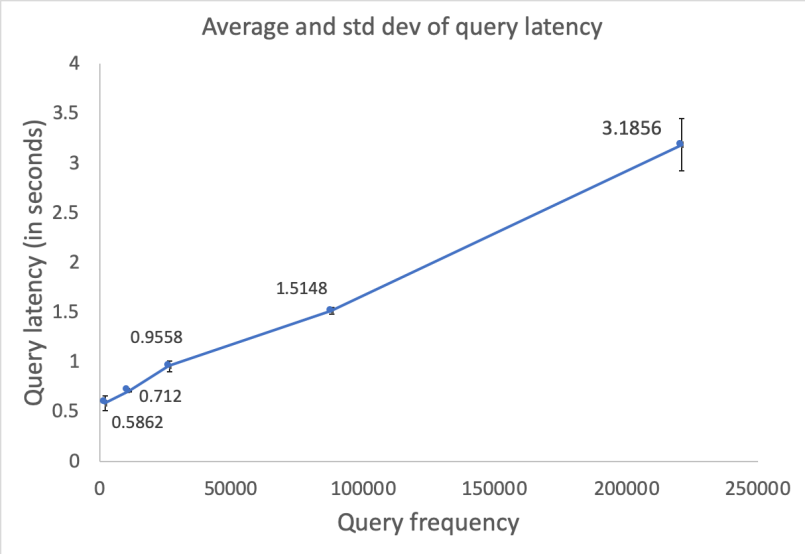
CS425 MP1 in Golang

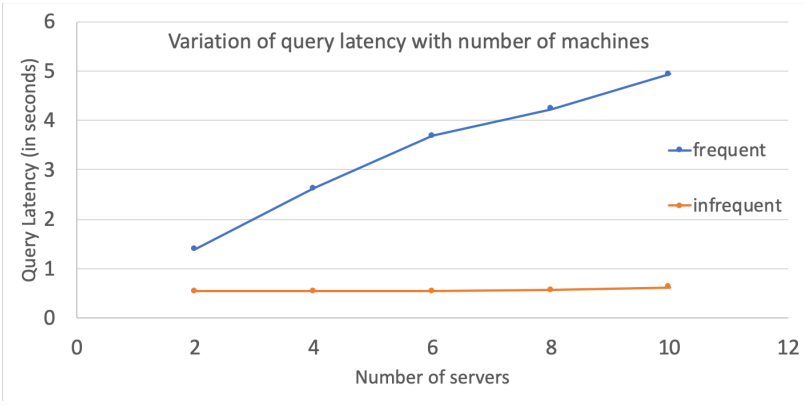
Bhavana Jain (bmjain2), Dipayan Mukherjee (dipayan2)

We have used TCP client and servers to implement the distributed file grep. The client launches multiple goroutines, one for each server. Each goroutine starts a TCP connection with the assigned server and sends the filename and pattern. The server searches the file for the required pattern. It iterates over the lines and in case of a match, sends it back immediately. The client goroutine receives this matched line and outputs it to a buffer specific to the server. Once the server exhausts the file contents, it sends the total number of matches and a special token to indicate the end. We have also added a -visual flag, when set, it prints the annotated output to os.Stdout and highlights all the matching patterns. Additionally, we handle errors like invalid regexp pattern, server failures and invalid files.

To perform **unit testing**, we have launch test servers and a test client. The test client generates files with known frequent, moderate and infrequent patterns. It also generates an expected output file with the filename, linenum and line content recorded. The client sends the generated files to test servers and queries the various patterns. The generated results from each query are recorded in a file. Finally, we check if all lines from expected output files exactly match at least one line in the generated output file.



The average latency of a frequent query (GET) that occurs ~221K times is 3.1856s whereas the average latency of an infrequent query (3/Mar) that occurs ~2K is 0.5862s. The cost per query for infrequent pattern is 3e-4 whereas for frequent pattern is 1.4e-5. This stems from the fact that frequent patterns have a very high density in all the files and hence get matched and returned quickly. We observe a higher standard deviation for frequent patterns, since the volume of packets (matched lines) is high for frequent patterns, a slight variation in network i/o leads to a large deviation in latency values.



The plot on the left shows that average query latency for frequent patterns in sublinear for number of machines . However the query time for infrequent pattern stays the same as the latency is dominated by fixed network overhead (setup).