Alex Shen
Tiancheng Wu

CS 425 MP1 Report

Design

The distributed log querier consists of a server running on each machine and a client on one machine. The client connects to each server and sends a grep command for each server to execute. The server responds first with a packet containing the size of the command output. Then, the server sends the entire output and follows that with the line count. The client uses the size data to resize the data buffer accordingly, and writes each result to a file.

Implementation

The design is implemented in C++11 with Boost.Asio for socket programming and compiled with CMake. A vector of client pointers is used to connect to each server. Asynchronous I/O is performed on the socket to read/write data between the server and the client. The size and line counts are encoded as 8-byte packets and read appropriately by the client. The line count is retrieved by running the grep command on the server, with the "-c" flag appended. If a server fails, the client for that server will not be able to connect, so it indicates that the output is 0.

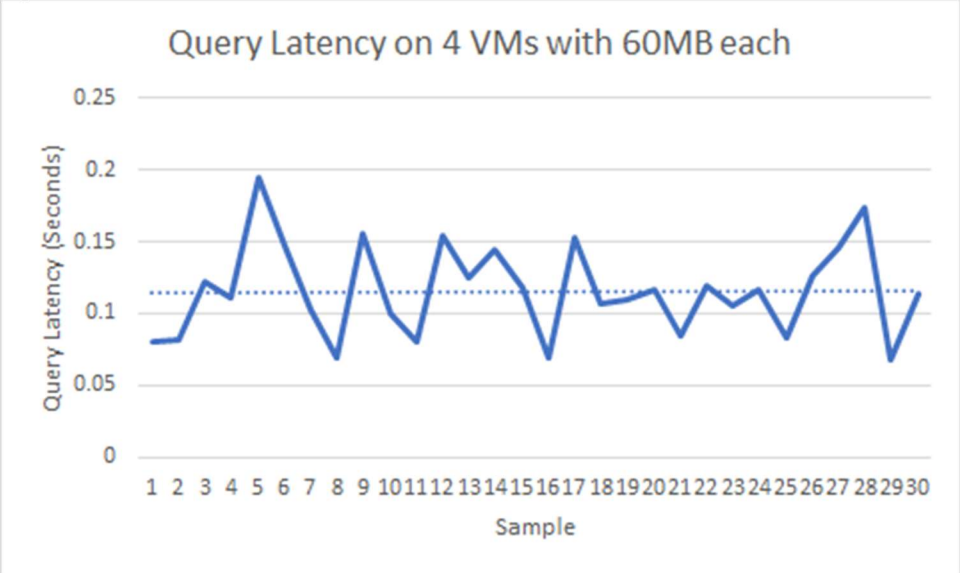
Evaluation

Correctness:

Unit tests determine correctness by executing in a distributed manner from the viewpoint of the client. The output of a hardcoded command is compared to the expected results to determine if the command was successfully executed and the correct data were transferred to/from the servers. The tests cover rare, somewhat, or frequent patterns on one, some, or all machines.

Performance:

The average query latency on 4 VMs with 60MB logs in each machine is used to determine performance. The command "grep NASA benchmark*.log" is used since it is a somewhat frequent pattern among all the logs. The sample size is 30 and each sample is plotted in the figure below.



Average: 0.11584 seconds
Standard deviation: 0.031662 seconds