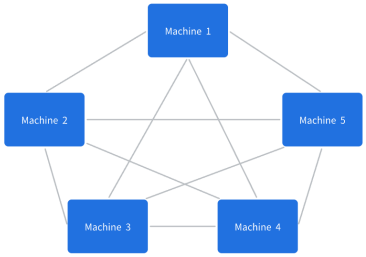
Jiachen Yuan
Sizheng Zhang
09/06/23

CS-425 MP1 Report

Design and Algorithm

We took a leaderless architecture for the distributed log querier. Every node in the cluster acts as both a server and a client, running two separate coroutines/processes. User inputs are read in and sent out by the client process to all other VMs. Each recipient VM parallelly executes it and replies with the results. The client then assembles the results and statistics and then prints them to stdout. Since there is no coordinator in the system, our design is highly fault tolerant, as any failed nodes would not impair the overall functioning and performance of the log querying commands. As shown in the graph, each node can send requests to and receive responses from every other node in the cluster.



Testing

We divided our testing into two parts, unit tests and end-to-end tests. Unit tests verify whether the output of core functions matches expectations. For instance, one such test makes sure a function gives errors if the user inputs an invalid "grep" command. For the end-to-end tests, we automatically generate 10 chunks of log files and distribute each to corresponding machines. Each log file has some known and some other random names. The querying program runs multiple "greps" on frequent patterns, infrequent patterns, and patterns that exist in one/all/some logs and compare the results with our expectations.

Evaluation

Setup - We ran 4 patterns in the form of `**grep -c {Pattern} vm*.log**` in a cluster of 4 machines each with about 60 MB logs and locally with the same 4 log files. The details of the queries are shown in the following table:

Pattern to Match	Type of pattern
"GET"	Frequent
"4992"	Infrequent
"http://www.powell.com/faq/"	Appears in some logs
"\[[0-9][0-9]/[A-Za-z][A-Za-z][A-Za-z]/[0-9][0-9][0-9][0-9]:[0-9][0-9]:[0-9][0-9]:[0-9][0-9][0-9]-[0-9][0-9][0-9][0-9]\]"	Regex for date

Result - The results are shown in the graph. In all cases we test, the distributed grep on VMs have significantly less latency than locally executed grep. This is expected because our system executes "grep" commands in parallel among all VMs on their part of the "global" log file, effectively eliminating the waiting time in local serial execution. On the other hand, local results are more stable, as indicated by the smaller standard deviation. This stability is likely due to the lack of network fluctuation, unlike in a distributed context.