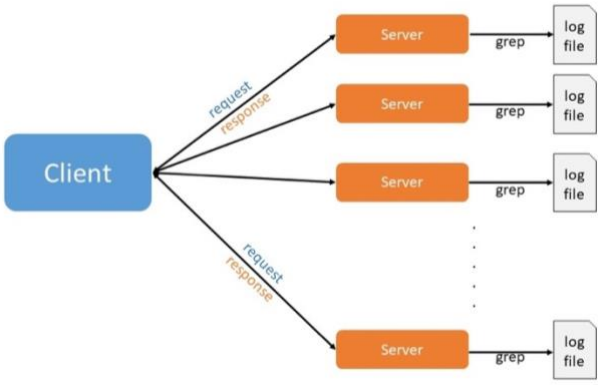
Distributed System MP1: Distributed Log Querier

Group 41 Cheng-Hsuan Huang(chhuang5), Yen-Chen Yeh(yy63)

Design

In our C++ based implementation of a distributed log querier, we adopted a client-server socket communication architecture for interaction. When a query machine is selected by the user, the machine acts as the client and sends simultaneous requests to all connected servers, initiating the query process. Servers process queries locally using two distinct methods: string matching uses the C++ 'find' function for exact matches and regular expression pattern searches uses the Linux 'grep' command. After processing, servers aggregate the results by summing matching line counts and send the outcome back to the client.



Unit test

Our unit test function generates log files locally, which contain known lines common to all machines and random lines specific to each machine. The function then executes several 'grep' functions with different query patterns of varying frequencies and receives the results from all machines, subsequently checking them for accuracy.

average query latency (4 machines each store 60 MB log files)

Result

We did our testing with 4 distributed machines, and tried three string matches and three Regex patterns, each including frequent, somewhat infrequent, and rare tokens. We try each token 10 times and plot their respective average time and standard deviation.

Total lines: 1089189				
	Token	Line Sum	Avg. Time(s)	Std. Dev
Frequent	“com”	676074	0.148744	0.021266
S/W Infrequent	“log”	126341	0.117714	0.023769
Rare	“428”	264	0.09632	0.017473
Frequent	“*.com”	676074	0.287157	0.007152
S/W infrequent	“428**”	136132	0.269425	0.012167
Rare	“edi*”	15546	0.091886	0.005589