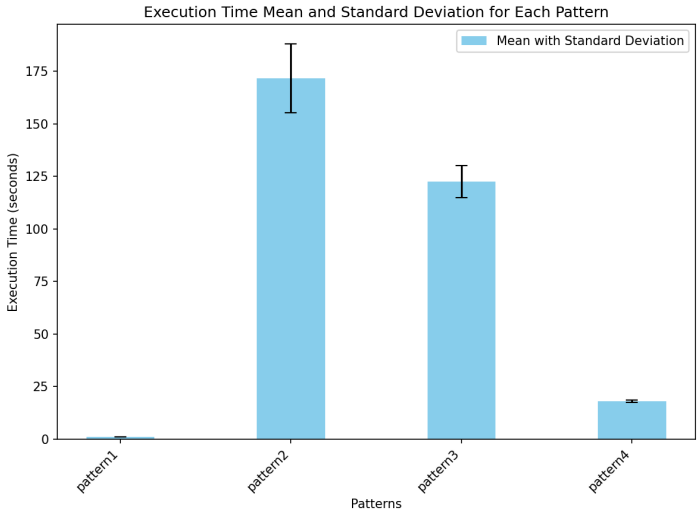
MP1 Report
Group No. 69 (sthakar3 and chaskar2)

We have created a distributed logger on a cluster of 10 VMs. Each of the 10 machines has a server running listening for socket connections. A client socket can be created on any of the 10 machines. The client will parallelly open connections to the server and run grep command on each machine. We have used Thread Pooling to enable parallel execution. The grep output will then be passed on to the client where it is displayed on the terminal.

Some implemented unit tests:

- test_rare_pattern_on_one_machine -> rare pattern log has 1 known line and 99 random lines. This rare log is on 1 machine other 9 machines have blank or random log files
- test_somewhatfreq_on_5_machine -> somewhat freq pattern log has 50 known lines and 50 random lines. This log is on 5 machines and other 5 machines have blank or random log files
- test_freq_pattern_on_all_machine -> freq pattern log has 100 known lines and 0 random lines. This log is on all machines

Observation Statistics:



Pattern 1: -r 'http://reynolds.com/' /home/chaskar2/log/
Pattern 2:-r '"GET' /home/chaskar2/log/
Pattern 3:-r -oP '"Mozilla[^"]+"' /home/chaskar2/log/
Pattern 4: -r -P '"GET /wp-content[^"]*" (200|404)' /home/chaskar2/log/

Discussing the observations:

-r 'http://reynolds.com/' is a simple string match, and it's the **fastest** with consistent times (~0.12 seconds) (434 matches)
-r '"GET' is a simpler pattern, but due to the **large number of lines matched** (654,694), it takes significantly longer (145 to 194 seconds)
-r -oP '"Mozilla[^"]+" uses a **character class** which makes it slightly slower due to the **complexity** of the pattern. It takes between 116 to 136 seconds and scans a large number of lines (1,036,161)
-r -P '"GET /wp-content[^"]*" (200|404)' uses alternation ((200|404)), but it still completes fairly quickly due to a **smaller number of matches** (77,098) and the fact that it doesn't need to scan as many lines as -r '"GET'
We see that as the average time increases the standard deviation increases as well, which is expected.