DV1567 - Project 1
Performance and scalability analysis based on queuing network theory

Solve in your favourite programming language.

Consider a large software company that uses an internal Web server to allow its programmers, testers, and documentation personnel to download files containing documentation, manuals, software code and executables. The Web server has one CPU and four identical disks.

By collecting and analyzing http logs and OS performance logs we obtain the average service demand for the CPU and the four disks: D_{CPU}=39.4 msec, D_{disk1}=77.1msec, D_{disk2}=123.8msec, D_{disk3}=80.4msec, D_{disk4}=235msec.

The main question of interest for the system administrator and the management are:

Q1. What is the maximum number of concurrent file downloads that can be in progress while the web server is in steady state conditions? λ/mμ < 1, μ= service rate

Q2. What is the maximum number of concurrent file downloads that can be in progress in order to satisfy a certain pre-specified Service Level Agreement (SLA)? The SLA is question is SLA={download time <= 20sec}

Q3. In case the load on the disks is balanced, what would be the maximum number of concurrent file downloads that can be in progress in order to satisfy a SLA={download time <= 20sec}.

Q4. How the Web server capacity could be improved in order to guarantee a throughput of 10 tps (i.e. 10 download per second)? What is the maximum throughput of the new system configuration? What is the maximum number of concurrent file downloads that can be in progress in order to satisfy a download time less that 20sec? [*Compare at least 2 different solutions that could be applied to improve the web server performance*]

Q5. Suppose the company would start using the TLS protocol to enable secure download from the web server. Results of performance test on the system with 4 disks and the load balanced among the disks shows that the use of encrypted communication impact only the CPU service demand. The average value measured for D_{CPU} is 137msec. What is, in that case, the maximum number of concurrent file downloads that can be in progress in order to satisfy the SLA={download time <= 20sec}?

Q6. How the Web server capacity can be improved to guarantee a maximum download time of 20 seconds with at least 150 concurrent users?

The students are requested to define the more appropriate queuing network model for the system and to use the QNM to answer the questions *Q1 - Q6*. When answering the questions *Q1 - Q6* students should comment on the system scalability and on the system bottleneck.

Upload the report and the code.