Management and Content Delivery for Smart Networks: Algorithms and Modeling

Lab. 1: Simulation of Simple Queuing Models

The objective of this laboratory is to practice the basic steps for simulating queuing systems. You will simulate the most basic queue models, investigating the effects of model parameters, e.g., arrival and service rates. This will help to understand the workings of a network simulator.

Exercise 1 - A Simplified Web Server

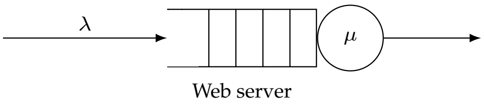

Figure 1: M/M/1 model.

Assume a server handles HTTP requests for a popular web site. The server is modeled as a simple queue station as in Figure 1. The server satisfies requests sequentially, one at a time. If a request arrives while the server is busy, it is put in a queue – first-come, first-served (FCFS). You will simulate two scenarios:

(a) $M/M/1$ queue: requests are satisfied according to a Poisson process with service rate μ; HTTP requests arrive following a Poisson process with arrival rate λ; the server has an infinite memory buffer to hold pending requests while it is busy.

(b) $M^X/M/1/B$ queue: Web servers have limited buffer capacity in practice. Moreover, client browsers fire many HTTP requests simultaneously to render a web page (e.g., to retrieve images and HTML files in parallel). Thus, HTTP requests arrive in batches to servers. In this scenario, assume requests are satisfied according to a Poisson process with service rate μ; the server can hold up to B HTTP requests in total; batches of HTTP requests arrive following a Poisson process with arrival rate λ; and the size of each batch is uniformly distributed in the interval $[a, b]$.

For both scenarios, study how the *response time* per HTTP request and the *buffer occupancy* vary according to λ and μ. For the second scenario, study how the number of HTTP requests rejected due to lack of buffer space is affected by parameters of the model.

Discuss the problem of simulation warm-up and derive confidence intervals for at least one of the considered performance metrics.

Exercise 2 - HTTP Accelerator + Web Server

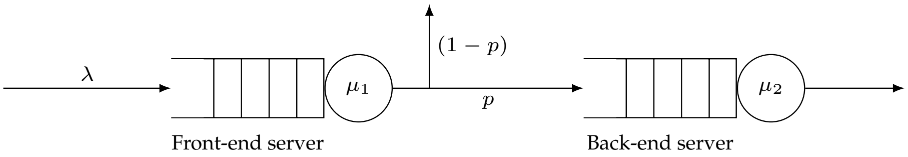

Figure 2: Deployment with a cache to accelerate the delivery of pages and static objects.

Assume the system administrator has decided to invest on a new web accelerator solution to improve performance. Figure 2 depicts a model for this scenario: a front-end server is deployed to satisfy HTTP requests. This server operates as a cache: It has large amounts of memory to satisfy popular requests with higher service rate μ_1. Some requests cannot be satisfied by the front-end server and must go to the back-end server with lower service rate μ_2 (i.e., $\mu_1 >> \mu_2$). The fraction of requests that are passed to the back-end server is modeled by the parameter p.

As in the previous exercise, assume that services follow Poisson processes with parameters μ_1 and μ_2; HTTP requests arrive at the front-end server in batches with similar characteristics as before; and both servers have limited buffer capacity B_1 and B_2.

Again, study the overall *response time* as well as *buffer occupancy* in both servers. Considering different values for μ_1 and μ_2, what would be a reasonable value for p to justify the deployment of the web accelerator in terms of response time?

Groups and Final Reporting

You are expected to work on groups of up to three students. Each group is required to prepare a short report describing results obtained during all labs in the course. This report must not exceed 10 pages.

You have as starting point the skeleton of a network simulator written in Python. You are however free to pick other programming languages you are familiar with to code your solution, provided that all members of the group are able to code and explain me the solution.

You need to delivery both the written report and your source code by June 16.

References

[1] SimPy in 10 Minutes. `https://simpy.readthedocs.io/en/latest/simpy_intro/`
[2] matplotlib. `http://matplotlib.org/`