CSC/ECE 579 – Introduction to Computer Performance Modeling

Simulation Project #3
Due Date: April 26, 2017

Description

For this project we will study a queueing system with multiple servers. Specifically, we will assume that the system has m identical servers. For the purposes of your simulation program, you will always assume that $m = 3$. The objective of the project is to: **(i)** evaluate multiple server systems, **(ii)** compare the FCFS and SJF service disciplines, and **(iii)** investigate how the results change when we consider a heavy tailed distribution for the service times.

As in the previous projects, you will make the following assumptions regarding the interarrival and service times:

- interarrival times are *exponentially distributed* with parameter λ
- service times are either *exponentially distributed* with parameter μ (M/M/m system), or follow a *bounded Pareto distribution*, defined below (M/G/m system).

Unlike previous projects, you will use $\mu = 1/3000$; this choice of value for μ is explained below. You will also assume that the size of the queue is **infinite**, therefore all arriving customers must be accommodated in the queue (no losses).

The Bounded Pareto Distribution.
Many recent measurements of computing systems have observed job size and file size distributions which are well-modeled by a bounded Pareto distribution. The PDF for the bounded Pareto distribution $B(k, p, \alpha)$ is defined as:

$$f_X(x) = \frac{\alpha k^\alpha}{1 - \left(\frac{k}{p}\right)^\alpha} \; x^{-\alpha-1}, \quad k \leq x \leq p, \quad 0 < \alpha < 2$$

In your simulation you will use $k = 332$, $p = 10^{10}$, $\alpha = 1.1$. For this choice of values, the mean value is fixed to 3000 (therefore, to make the results comparable when service times are exponentially distributed, μ must be equal to 1/3000).

From the definition of the bounded Pareto distribution, the maximum service time is p ($= 10^{10}$ in our case), and the minimum service time is k ($= 332$). One key property of heavy tailed distributions and (many) bounded Pareto distributions is that a tiny fraction ($< 1\%$) of the very largest jobs comprise over half of the total load; this is called the heavy-tailed property. The distribution $B(k = 332, p = 10^{10}, \alpha = 1.1)$ has a strong heavy tailed property, in that the largest .3% of the jobs comprise half the total load, and it has a finite variance. Note that while the bounded Pareto distribution has both the heavy tailed property and finite moments, in general, heavy tailed distributions have infinite variance, and sometimes even infinite mean.

Confidence Intervals.
For all the performance results you collect and plot, you must also provide confidence intervals, computed in the same manner as described in project #2.

Service Disciplines.
You will simulate the following two service disciplines:

1. *First-come, first-served* (FCFS).
2. *Shortest job first* (SJF).

Your simulation program will terminate once C customers have completed service, where C is an input parameter. For initial conditions assume that at time $t = 0$ the system is empty.

Submission and Deliverables
Submit your source code (*no object files*!) using `submit` by midnight on the day due. There are several weeks until the due date for you to work on this project, therefore **no** late submissions will be accepted. Name the file containing the source code `proj3`, with the appropriate extension (e.g., `proj3.c` if you code in C).

Input: Set up your simulation program to accept values for the following parameters *from the command line* (in this order):

1. the parameter λ of the distribution of interarrival times
2. the number C of customers serviced before the program terminates
3. an integer L such that: 0 – FCFS, 1 – SJF.
4. an integer M such that: 0 – M/M/3 system, 1 – M/G/3 system.

Upon completion, your program should print to the standard output the following:

- the value of the input parameter λ
- the value of the input parameter C
- the value of the master clock at the end of the simulation
- the average service time and corresponding confidence intervals (based on the C serviced customers only)
- the average waiting time and corresponding confidence intervals (based on the C serviced customers only)

Include a `makefile` with your submission. This is a large class, and we would like to ensure that the TA does not spend an inordinate amount of time compiling and running your programs. Therefore, if you fail to include a `makefile` we will **subtract 5 points** from your project grade.

In addition to submitting your source code, you must answer the following questions and include any necessary plots. Recall that you must always use $\mu = 1/3000$ in your simulation for exponentially distributed service times.

Task 1. Calculate the mean and variance of the bounded Pareto distribution for the parameters given above. Show all the math for the derivation.

Task 2. Read the paper [CL97] in the reading list, and answer the following questions:

1. How do the authors define "heavy-tailed" distributions?
2. What is an α-Stable distribution, and why are they introduced?
3. What is a "swamping observation" and what is its significance?
4. Under what conditions do the authors claim that a simulation may *always be in transient state*?
5. Overall, what are the conclusions that the authors draw regarding simulations with heavy-tailed workload?

Task 3. This task involves the *3-server* queueing system. Plot the average customer system time against the value of ρ, for $\rho = 0.1$ to $\rho = 0.9$, in increments of 0.1 (include confidence intervals). Note that $\rho = \frac{\lambda \bar{x}}{3}$ for this system. Use $C = 50,000$ customers. Compile two pairs of plots, one pair for each of the service disciplines (FCFS and SJF), where one plot of each pair corresponds to the M/M/3 system and one plot corresponds to the M/G/3 system. Discuss the relative performance of the plots; did you expect these results based on what we have covered in class?

Task 4. This task involves the *single-server* queueing system. You will assume that $\rho = 0.5$ and that $C = 100,000$. Let $T(x)$ denote the total system time for a customer whose service time is x, and $W(x) = T(x) - x$ its waiting time. We define the *slowdown* for a customer with service time x as $S(x) = \frac{W(x)}{x}$. The slowdown can be thought of as a measure of *fairness* [1]. The objective of this task is to see how the size of the service time affects the slowdown of the customer. For this, you need to compute the slowdown for each of the C customers, and for each of the two service disciplines (FCFS and SJF). Since there are too many values of $S(x)$, making it difficult to plot, you will plot the slowdown against the *percentile* of the service time distribution. Use $B = 100$ bins, and place each customer to the appropriate bin according to its service time (i.e., the 1% of customers with the smallest service times go into the first bin, the 1% of customers with the next smallest service times go to the second bin, etc). For each bin, take an average of the slowdown for the customers in that bin. Then, plot the slowdown against the number of bins; compile two plots, one for FCFS and one for SJF. What can you say about the fairness of each of these service disciplines? How do the results compare to the slowdown of the processor sharing discipline which we discussed in class?

Grading

Source code	50 Points
Four tasks @ 12.5 points each	50 Points
	100 Points

[1] Why?