Homework 5

CS 547

Due Friday, Mar. 2, in class

Service Level Objectives

Suppose you manage a datacenter with 100 servers that can be modeled as M/M/1 queues. The stream of incoming requests to your system is a Poisson process with rate $\lambda = 5000$ requests per second, and all incoming requests are randomly divided across your 100 servers. What average service time is required to guarantee that 99.9% of requests exit the system with a residence time less than 100 ms?

Changes in 95th Percentile Latency

How does 95th percentile latency in the M/M/1 queue change as utilization increases?

First, rearrange the M/M/1 residence time CDF to calculate the 95th percentile latency as a function of utilization. You can assume $\overline{s} = 1$ to simplify the derivation.

Now plot both 95th percentile latency and average latency for values of U between 1% and 99%. What conclusions can you draw about the behavior of 95th percentile latency at high utilizations?

What implications do these results have for system planning and operation?

Queueing Simulator Implementation

Implement a simulator for an M/M/1 queue, as described in class. You can use any language you wish, but your implementation should record the arrival time, the service start time, and the completion time of each customer. From these measures, calculate each customer's waiting time and residence time. Use the inverse CDF method to generate exponentially distributed interarrival and service times.

Let N be the number of customers generated in each run of the simulation. Use your simulator to predict $\overline{R}$ when $\lambda = 1$, $\overline{s} = .75$, and $N = 10000$.

Simulation Error

Running the simulator one time gives a single estimate of the value of $\overline{R}$. Because the estimate is calculated from a finite number of data points, it contains some inherent randomness. Therefore, using only one simulated value as a prediction of $\overline{R}$ is unlikely to be accurate unless N is very large.

If you run the simulator multiple times with the same value of N, you'll generate a set of estimates for $\overline{R}$. These estimates all contain some random error that separates them from the real, true value of $\overline{R}$. It's possible to very clearly describe the behavior and statistical properties of this error.[1]

[1] See the notes on the website to learn more about error, the Central Limit Theorem, and confidence intervals.

Keep N fixed at 10000 and run your simulator 100 times, recording the value of $\overline{R}$ produced each time. Calculate the mean and the variance of the 100 $\overline{R}$ values. How does the mean of the means compare to the the actual expected residence time?

Simulating an $\mathbf{M/E_2/1}$ Queue

Modify your simulation to use service times drawn from a two-stage Erlang distribution. Set $\lambda = 1$ and set the Erlang distribution's rate parameter so that $\overline{s} = .75$. Perform 100 simulations with $N = 10000$ and average the 100 estimates of $\overline{R}$ to obtain a final value.

The true value is $\overline{R} = 2.4375$.