Homework 4

CS 547

Due Friday, Feb. 24, in class

Justification for the Poisson Process

This problem is designed to give you some insight into why the Poisson process is a reasonable model for arrivals in many real-world systems. To do this, we'll perform a simulation and show that the Poisson distribution reasonably approximates the results.

Suppose we perform a large number of independent and simultaneous Bernoulli trials and count the number of successes (i.e. imagine that we flip a large number coins and keep count of the number of heads). Theoretically, the distribution of the number of successes out of N Bernoulli trials with parameter p is described by the *binomial distribution*,

$$P[k \text{ successes out of } N \text{ trials}] = \binom{N}{k}(1-p)^{N-k}p^k$$

If N is large and p is small, the binomial distribution can be closely approximated by a Poisson distribution.

Here's a simulation procedure that will illustrate this fact:

1. Generate 1000 Bernoulli trials with $p = .01$ and count the number of successes that occur.
2. Repeat step 1 at least 5000 times, keeping track of the number of successes in each experiment.
3. Plot the total distribution of successes generated from all 5000 experiments. That is, for $k = 0, 1, 2, \ldots$, plot the fraction of experiments that resulted in k successes.

Now, on the same figure, plot the Poisson distribution with parameter $\lambda = Np = 1000 * .01$. Recall that the Poisson distribution is

$$P[k \text{ events}] = \frac{e^{-\lambda}\lambda^k}{k!}$$

Your plot should show that the two distributions have very similar shapes. As you increase the number of experiments and the number of trials per experiment, the approximation should become even more accurate. These results demonstrate that applications with reasonably independent customers and sufficiently heavy traffic tend to experience Poisson arrivals.

M/M/1 Calculation Practice

Given an M/M/1 queue with $\lambda = 10$ arrivals per second and $\overline{s} = 75$ ms, calculate the following quantities:

- the utilization, U
- the probability that an arriving customer finds the queue idle
- the average residence time, $\overline{R}$
- the average waiting time, $\overline{W}$

- the average number in the queue, $\overline{Q}$
- the average number in the queue at an arrival instant
- the average number waiting and not being served
- the average number waiting and not being served at an arrival instant

Suppose we keep λ fixed at 10 arrivals per second. What value of $\overline{s}$ is required to achieve an average residence time of .10 seconds?

What value of $\overline{s}$ is required to keep $\overline{Q} \leq 2$?

If $\overline{s}$ must stay fixed at 75 ms, what is the maximum arrival rate we can sustain while keeping $\overline{R} \leq 1$ second?

Power Management in Datacenters

There has been a large amount of recent research on reducing the economic and environmental impact caused by the energy demands of modern datacenters.

One of the most basic strategies for conserving energy is *dynamic voltage scaling* (DVS). By reducing the operating voltage of a processor, we can its reduce energy consumption in exchange for decreased performance.

Suppose we have a set of k servers, each operating as an M/M/1 queue capable of running at a maximum rate μ. The total arrival rate to the entire set of servers is λ. Assume the service rate of each server scales linearly with its power consumption[1]. Consider two basic operating strategies:

- Turn on all k servers, each running at a fraction $\frac{1}{k}$ of its maximum power. In this scenario, the arrival rate at each queue is $\frac{\lambda}{k}$ and the each queue's service rate is $\frac{\mu}{k}$.
- Turn on one server at full power. The server receives all arrivals at rate λ and has service rate μ.

Of these two strategies – multiple slow servers or one fast server – which one minimizes customer residence time? By how much? Can you provide an intuitive explanation for this result?

Stuff Managers Say

Imagine that you've (finally) finished this class, graduated, and gone on to your dream job with a prominent Internet company.

One day, during a conversation about your company's datacenter, your boss makes the following statement.

Our machines are too expensive to sit idle 25% of the time! I want to buy the smallest possible number of machines and run them all at 100% utilization!

Thinking back to your time in 547, how do you respond?

[1] This is fairly unrealistic. In most DVS models, performance is some sort of nonlinear function of power consumption. Also, there's typically a minimum level of power required to simply keep the server on. For this problem, ignore these complications.