Solution to COMP9334 Revision problems - Week 3A

Question 1

This aim of this question is to compare 3 different ways to upgrade an existing system. Let us first calculate the response time of the existing system.

Existing system

The assumptions imply that the queue is an M/M/1 queue with arrival rate $\lambda = 9$ and mean service time (= $1/\mu$) = 0.1. This gives an utilisation $\rho = \lambda / \mu = 9 * 0.1 = 0.9$. Plugging these numbers into the mean response time T for M/M/1 = $1/(\mu (1- \rho))$, we have T = 1s.

Alternative 1

We upgrade the CPU to one twice as fast and since the service time is inversely proportional to the speed, the new mean service time = 0.05. The system is an M/M/1 queue with $\lambda = 9$ and $1/\mu$ = 0.05. The utilisation $\rho = \lambda / \mu = 9 * 0.05 = 0.45$. Using the M/M/1 mean response time formula gives T = 0.0909s.
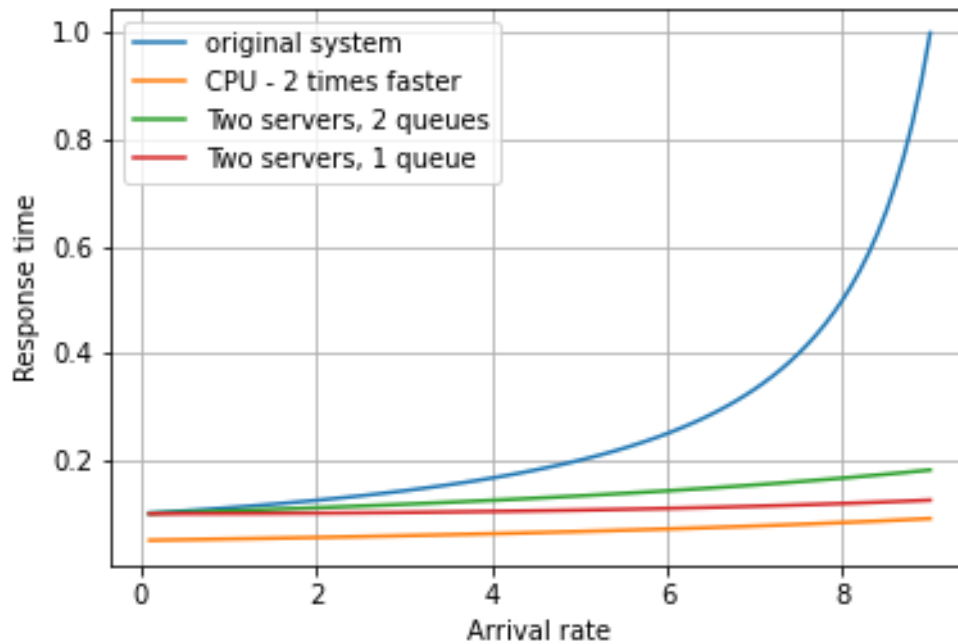
Alternative 2

We use 2 existing systems and each system maintains its own queue. The arrival rate at each queue is half of the system arrival rate. (This is known as Poisson splitting, see Section 11.7 of Harchol-Balter.) Effectively, we have two M/M/1 queues with with $\lambda = 4.5$ and $1/\mu$ = 0.1. The utilisation $\rho = \lambda / \mu = 4.5 * 0.1 = 0.45$. Using the M/M/1 mean response time formula gives T = 0.1818.

Alternative 3

This alternative is effectively an M/M/2 queue with $\lambda = 9$ and $1/\mu$ = 0.1. The utilisation $\rho = \lambda / \mu / 2 = 9 * 0.1 / 2 = 0.45$. Using the M/M/2 mean response time formula gives = 0.1254s.

Part (b)

We use a number of different arrival rates and we plot a curve of how the response time changes with the arrival rate. See the graph below:

Observations:

Note all the three alternatives have the same utilisation but have different response times. All three alternatives improve the response time of the current system but to different extend.

Alternative 1 gives the best response time but it means throwing away the existing hardware and buy an entirely new system. This can be costly.

Both Alternatives 2 and 3 make use of the existing hardware but Alternative 3 gives a better response time. It may be cheaper to scale up the system by using Alternatives 2 and 3 in the long term because it allows incremental expansion of the system.

Note: I used Python to create the above graph. I've posted the files on the course web site.

Question 2

Part (a)

There is no queueing at all in this case. The mean waiting time is zero second.

Part (b)

The mean arrival rate is 1 customer / s. The mean service rate is 2 customers/s. By the M/M/1 formula, we know the mean response time is $1/(\mu\,(1-\rho)) = 1$s. The mean waiting time = mean response time – mean service time = $1 - 0.5 = 0.5$s.

You may draw a number of conclusions here:
- Queueing is a result of the randomness in arrival and service patterns. In Part (b), the mean arrival rate is 1 customer/s, which means that the instantaneous arrival rate can sometimes be greater than 1 customer/s, the high instantaneous arrival rate can create a backlog in the system and create queueing.
- Waiting time is not just a function of the mean service time and mean arrival rate. It depends on the arrival and service time distributions too.

Question 3

This is effectively a M/M/4/4 system with arrival rate $\lambda = 3$ and mean service rate $\mu = 1/1.5$. The probability that a connection request is blocked is the same as the probability that there are 4 requests in the system. This is given by the formula

$$P_m = \frac{\dfrac{\rho^m}{m!}}{\displaystyle\sum_{k=0}^{m} \dfrac{\rho^k}{k!}} \qquad \textbf{where} \qquad \rho = \frac{\lambda}{\mu}$$

With m = 4.