Problem 1 - A single fast server, or multiple slower servers?

In this problem we will compare two queues that are apparently very similar: a queue with a single server, and one with $c$ distinct servers that are each $c$ times slower than the single server. In the real world, we are interested to questions as: at the same price, would you buy a single Internet connection at 30 Mbps speed or 3 distinct slower connections each at 10 Mbps (assuming the existence of a box able to manage multiple connections)? The answer to this problem is the business idea at the basis of San Diego-based Mushroom Network Inc.

M/M/1 Queue

1) Consider the $M/M/1$ queue with arrival rate $\lambda$ and departure (or service) rate $\mu$. Let the state of the system be the number of packets in the system (queue and service), and let $s(t)$ denote the state at time $t$. The state transition diagram of this queue is represented in the figure below (where $\lambda = \lambda_1 = \lambda_2 = \cdots$ and $\mu = \mu_1 = \mu_2 = \cdots$). Remember that there is no upper bound to the queue size, and assume that $\lambda < \mu$. We want to compute the stationary distribution of the queue size, that tells us how likely it is to be in each state when the queue has reached a stationary condition (read “after a large number of steps”). To be precise the stationary distribution is a vector

$$\pi = (\pi_0, \pi_1, \pi_2, \ldots),$$

such that

$$\Pr(s(t) = k) = \pi_k$$
for all \( k \geq 0 \) and for all \( t \) suitably large. As the values in \( \pi \) constitute a probability distribution, we have that
\[
\sum_{k=0}^{\infty} \pi_k = 1.
\]

As the diagram above suggests, we have that
\[
\begin{align*}
\lambda \pi_0 &= \mu \pi_1 \\
\lambda \pi_1 &= \mu \pi_2 \\
&\vdots \\
\lambda \pi_i &= \mu \pi_{i+1} \\
&\vdots
\end{align*}
\]

1) Use the fact above to express \( \pi_k, k > 0 \), as a function of \( \pi_0 \).
2) Using \( \lambda < \mu \) and the fact that all \( \pi_k \)'s sum to 1, compute \( \pi_0 \) (as a function of \( \lambda \) and \( \mu \)).
3) Using the results above, compute the expected number of packets in the system at any given time. As you learnt in class, you should get \( \frac{\lambda}{\mu - \lambda} \). You may find it useful that \( \rho = \frac{\lambda}{\mu} < 1 \).
4) What is the expected time \( T_1 \) that a packet spends in the system (queue and service) if the arrival rate is \( \lambda \) and the departure rate is \( 3\mu \)?

**M/M/c Queue**

Let’s now consider the generalization of the M/M/1 queue, the M/M/c queue, whose only difference is that there are \( c > 0 \) servers each with the same service rate \( \mu \). In the case of a M/M/c queue, if a packet arrives in the system when there are up to \( c - 1 \) other packets, then it goes to service right away without waiting in the queue (as at least one server is free). The utilization of this system is defined as \( \rho_c = \frac{\lambda}{c\mu} \). Defining the stationary distribution in the same way as for the M/M/1 queue, you could show (but you don’t have to) that, for the M/M/c queue,
\[
\begin{align*}
\pi_i &= \frac{(c\rho_c)^k}{k!} \pi_0, \quad k = 1, 2, \ldots, c - 1, \\
\pi_i &= \frac{c^k \rho^k}{c!} \pi_0, \quad k \geq c.
\end{align*}
\]

5) What is \( \pi_0 \)?
The probability that all servers are occupied (and thus a new arrived packet has to wait in the queue) is
\[
\pi_{c+} = \sum_{k=c}^{\infty} \pi_k.
\]
This means that, with probability $\pi_{c+}$ the new packet has to wait in line, while with probability $1 - \pi_{c+}$ it is served right away.

6) Compute $\pi_{c+}$ as a function of $c, \rho_c, \pi_0$.

You could also (but, again, you don’t have to) show that the expected time $T_c$ that a packet spends in this system (queue and service) is

$$T_c = \frac{\rho_c}{\lambda(1 - \rho_c)} \pi_{c+}.$$

7) What is the expected time $T_3$ that a packet spends in a M/M/3 queue with arrival rate $\lambda$ and three servers each with service rate $\mu$? (Do not write the complete expression for $\pi_{3+}$)

8) Compare $T_3$ with $T_1$ (computed above). For the same price, would you rather buy one outgoing link with rate 30Mbps or three outgoing links each with rate 10Mbps (assuming that you have a box that distributes your outgoing flow between them)?

**Problem 2: balancing the traffic.**

In this problem we consider the notion of fairness, according to which we want to give the same treatment to different services (in this case, packets flows). We first consider two independent queues, each receiving a certain flow of packets. We then consider an additional flow of packets, and we allow for the possibility to split it among the two queues in such a way that the average packet delay on each queue is the same. This will require to split the new packet flow unequally between the queues, in a measure depending on their capacity and their current utilization factor.

Consider the following M/M/1 queues.

**Queue 1:** There is an incoming flow of packets of constant length $\ell = 1kbyte$.

The service time for a packet is $T_s^{(1)} = 0.2ms$ (note that this value is a constant, as all packets have the same length), and the average time that a packet spends in the queue (waiting time in the buffer plus service time) is $T^{(1)} = 1.6ms$.

1) Compute the departure rate $\mu^{(1)}$.
2) Compute the arrival rate $\lambda^{(1)}$.
3) Compute the utilization factor $\rho^{(1)}$.

**Queue 2:** There is an incoming flow of packets of constant length $\ell = 1kbyte$, of rate $R^{(2)} = 10Mbps$. The outgoing link is used at 75% of its capacity.

4) What is the utilization factor $\rho^{(2)}$?
5) Compute the arrival rate $\lambda^{(2)}$.
6) Compute the departure rate $\mu^{(2)}$.
7) Compute the service time $T_s^{(2)}$.
8) Compute the average time that a packet spends in the system $T^{(2)}$.

Consider an additional flow of packets, each of length $\ell = 1kbyte$, with arrival rate $\lambda_{new} = 400$ packets per second. We want to split this flow among the two queue in a fair way, that is, in such a way that the two queues experience the
same average packet delay. Writing
\[ \lambda_{new} = \lambda^{(1)}_{new} + \lambda^{(2)}_{new}, \]
we mean that the new flow is split into two smaller flows, one with arrival rate \( \lambda^{(1)}_{new} \) added to Queue 1, and the other with arrival rate \( \lambda^{(2)}_{new} \) added to Queue 2, resulting in a total flow to Queue 1 with arrival rate
\[ \lambda^{(1)}_{tot} = \lambda^{(1)} + \lambda^{(1)}_{new}, \]
and a total flow to Queue 2 with arrival rate
\[ \lambda^{(2)}_{tot} = \lambda^{(2)} + \lambda^{(2)}_{new}. \]

9) What are the packet service time in Queue 1 and Queue 2 once the new flows are added? Did they change?

10) What is the average time \( T^{(1)}_{tot} \) a packet spends in Queue 1 once the new flow is added? Write its expression as a function of \( \mu^{(1)}, \lambda^{(1)}, \lambda^{(1)}_{new} \).

11) What is the average time \( T^{(2)}_{tot} \) a packet spends in Queue 2 once the new flow is added? Write its expression as a function of \( \mu^{(2)}, \lambda^{(2)}, \lambda_{new}, \lambda^{(1)}_{new} \) (not \( \lambda^{(2)}_{new} \)).

12) To reach our goal of fairness, we want \( T^{(1)}_{tot} = T^{(2)}_{tot} \). Equate the two expressions obtained in parts 10) and 11), and write an expression for \( \lambda^{(1)}_{new} \) in terms of \( \mu^{(1)}, \lambda^{(1)}, \mu^{(2)}, \lambda^{(2)}, \lambda_{new} \).

13) Compute \( \lambda^{(1)}_{new}, \lambda^{(2)}_{new} \).

14) Compute \( T^{(1)}_{tot} \) and \( T^{(2)}_{tot} \) and check they are equal.

15) Compute the utilization factors of the queues (after the new flows are added).