1. Finite Birth-Death Chain. Consider a queue at a supermarket. To make life easier let us assume that time is slotted. At each time slot, only one customer can arrive to the line and this happens with probability $p$. The head of line customer leaves the line with probability $q$. The queue can be modeled with a birth-death Markov chain.

(a) What is the probability that there are $n$ customers in line?

(b) What is the average number of customers in queue? What happens if $q < p$?

Sooner or later, the number of customers in line will reach any finite $n$. Assume that arrivals who find that there are already $l$ customers in line, depart immediately without receiving service. In other words the number of customers in queue is limited by $l$.

(c) What is the probability that there are $n$ customers in queue now, for $n \leq l$?

(d) What is the average number of customers in queue? What happens in case of $q < p$ or $q = p$?

(e) Compute the limit of answers of part (c) and (d) when $l$ goes to infinity. Are the limits the same as part (a) and (b)?

2. Doubly Stochastic Matrix. (From Wolff: Problem 3-16) A square matrix $P = (p_{ij})$ of nonnegative elements is called stochastic if for all rows $i$, $\sum_j p_{ij} = 1$. It is called doubly stochastic if it also has the property: for all columns $j$, $\sum_i p_{ij} = 1$. For a finite Markov chain (i.e., with finite states) which is also irreducible and aperiodic, with a doubly stochastic transition probability matrix, “guess” the form of stationary
probability vector $\pi$. Verify that $\pi = \pi P$.

3. Constraint Sets. The constraint set method was studied in the class. Using the pigeon-hole principle, we saw that an $N \times N$ switch and FIFO service policy requires $3N - 1$ memory blocks (where each can perform one memory operation per time slot).

Now consider a switch with high priority packets. Whenever a packet from the high priority class arrives, it pushes in to the head of line. The arrival rate of high priority packets is very small; when a high priority packet arrives, all the packets that were affected by the previous high priority packet have departed. Using constraint sets, determine how many memory blocks are needed.

4. Comparison of scheduling policies. Consider the 2-queue, 1-server system shown in Figure 1 below. Arrivals occur at each buffer according to independent rate $\lambda$ Poisson processes. The service times are independent and exponential of rate $\mu$. Suppose for stability that $2\lambda < \mu$.

\[
\begin{align*}
Poisson(\lambda) & \quad Poisson(\lambda) \\
\downarrow & \quad \downarrow \\
\text{exp}(\mu) & \\
\end{align*}
\]

Figure 1:

a. Consider 3 service policies: (1) serve the longest queue, (2) serve the shortest nonempty queue, and (3) serve a nonempty queue at random. When ties occur under (1) or (2), they are broken at random (independently and fairly). Let $Q^i(t) = Q_1^i(t) + Q_2^i(t)$ be the total number of packets in both queues under policy (i). Assuming that the system is in equilibrium, for each $i = 1, 2, 3$, what is $P(Q^i(t) = k)$ as $k$ ranges over $\{0, 1, 2, \ldots\}$?

b. Now consider the system shown in Figure 2. In this system arrivals occur according to a rate $2\lambda$ Poisson process. Each arrival is assigned to one of the two queues equally likely. Assuming that the server serves the longest queue (breaking ties at random), what is the equilibrium distribution of $Q(t)$, where $Q(t) = Q_1(t) + Q_2(t)$?
Figure 2:

\[
Poisson(2\lambda)
\]

\[
\exp(\mu)
\]

\[
\text{Figure 2:}
\]

c. Again consider the system in Figure 2. With everything as described in part (b), now suppose that each arriving packet is assigned to the shortest queue (ties broken at random). What is the equilibrium distribution of \( Q(t) = Q_1(t) + Q_2(t) \)?

Consider \( R(t) = |Q_1(t) - Q_2(t)| \). What is the equilibrium distribution of \( R(t) \)?

d. For the system in Figure 2, with everything being as in part (c), now assume that there are \( n \) buffers and that \( n\lambda < \mu \). Let \( R(t) = \max_i \{Q_i(t)\} - \min_i \{Q_i(t)\} \). What values does \( R(t) \) take? What is its distribution?

Conclude that the policies “join the shortest queue” and “serve the longest queue” balance the load in the system.

5. Work conserving and frame conserving. Let’s consider two systems each with a single server and infinite buffer space. Thus, no packets are ever dropped. Time is slotted, and each server can serve only one packet per time slot. Packets depart at the beginning of a time slot and new packets arrive at the end of a time slot. The queue occupancy is examined at the end of a time slot (after arrivals). System 1 has a work conserving server. Whenever the buffer is non-empty, one packet can leave the system in every time slot. The buffer in system 2 is divided into \( N \) queues. An incoming packet is put into one of the queues according to some criteria (such as final destination). The server is “sticky”, i.e., once it starts serving a queue, it serves it for \( b \) consecutive time slots, and up to \( b \) packets from this queue can leave the system. Therefore, it decided not to do any work unless there are queues with at least \( b \) packets. When there are such queues, and the server is free to choose a new queue to serve, it picks one of them, and serves it for \( b \) time slots. If we define a frame to be \( b \) packets in the same queue, then this system is “frame conserving”.

In other words, it is work conserving if there are full frames. Let the two systems be empty at time 0, and have the same cumulative arrival process $A(t)$, for $t \geq 0$. Let $Q_i(t)$ denote the total number of packets in queue at time $t$ for system $i$, $i = 1, 2$.

(a) If you are only told the total number of packets $Q_2(t)$ in system 2, what’s the minimum value $Q_2(t)$ has to be for you to be sure that server 2 is going to serve packets in the next time slot?

(b) What is the maximum difference between $Q_1(t)$ and $Q_2(t)$? Please provide a traffic pattern which achieves this difference and prove that it is the maximum.