Problem 8.4

Let the state be the number of busy servers. The state space
\{0, 1, 2, ..., c\}
The state transition rate diagram is

\[ \begin{array}{c}
0 \\
\mu \\
1 \\
2 \\
3 \mu \\
\vdots \\
\infty \\
\mu \\
\end{array} \]

From the diagram, we can write the balance equations
\[ \begin{align*}
\mathcal{P}_0 &= \mu \mathcal{P}_1 \\
\mathcal{P}_1 &= 2\mu \mathcal{P}_2 \\
\vdots \\
\mathcal{P}_i &= c \mu \mathcal{P}_c
\end{align*} \]

Where \( \mathcal{P}_i = \mathbb{P} [ \text{the number of busy servers at steady state} = i ] \)
\( (i = 0, 1, ..., c) \)

And
\( j = 40 \text{ customers/hour} \), \( \mu = 60 \text{ customers/hour} \)

Solving the balance equations, we have
\[ P_i = \frac{\frac{\mu^i}{i!}}{\mu^i} \mathcal{P}_0 = \left( \frac{2}{3} \right)^i \frac{1}{i!} \mathcal{P}_0 \quad (i = 0, 1, 2, ..., c) \]  \( (1) \)

Using the fact that \( \sum_{i=0}^{\infty} \mathcal{P}_i = 1 \), we have that
\[ \mathcal{P}_0 = \left[ \sum_{i=0}^{\infty} \left( \frac{2}{3} \right)^i \left( \frac{1}{i!} \right) \right]^{-1} \]  \( (2) \)

We know that
\( \mathcal{P}_c = \mathbb{P} [ \text{that all ATMs are busy} ] = \mathbb{P} [ \text{all arrivals find all ATMs busy} ] \)
Since, in one hour, there are 40 customers arriving, the loss per hour is 40p_c dollars.

We want the loss per hour to be less than 5 dollars. That is

$$40p_c \leq 5 \Rightarrow p_c \leq \frac{1}{8}$$

Using (1) and (2):

$$\frac{(\frac{1}{8})^c \frac{1}{c!}}{\sum_{c=0}^{\infty} (\frac{1}{8})^c \frac{1}{c!}} < \frac{1}{8} = 0.1250$$

When $c=1$, $p_c = 0.4$
$c=2$, $p_c = 0.1476 \leq 0.1250$

Hence
$c=2$ is recommended.
Problem 8.5

As suggested in the hint that was provided in the homework, this problem can be modeled as an $M/M/1//N$ queue and analyzed according to the results provided in Section 8.6.6 of your textbook.

In our problem, we have $p = 12$ requests per hour and $\lambda = \frac{1}{15}$ minutes, so $\lambda = \frac{1}{4}$ requests per hour. Thus, $p = \frac{2}{3} = \frac{4}{12} = \frac{1}{3}$.

Now, the average waiting time at the queue equals the average time spent at the station minus the average service time. Following the textbook notation:

$$E[W] = E[R] - \frac{1}{\mu}$$

and from (8.70)

$$E[R] = \frac{N}{\mu(1 - \pi_0)} - \frac{1}{\mu}$$

Also, from (8.67)

$$\pi_0 = \left[ \sum_{n=0}^{N} \frac{N!}{(N-n)!} \rho^n \right]^{-1}$$

Hence, the speed requirement can be expressed by:

$$E[W] = \frac{N}{12 \left( 1 - \frac{1}{1 - \sum_{n=0}^{N} \frac{N!}{(N-n)!} \rho^n} \right)} - \frac{1}{\lambda} - \frac{1}{12} < \frac{22}{60} \tag{2}$$

$$\leq \frac{N}{1 - \left[ \sum_{n=0}^{N} \frac{N!}{(N-n)!} \rho^n \right]} < \frac{22}{5} = 4.4 \tag{1}$$

The corresponding utilization of the workstation server is:

$$U(N) = 1 - \pi_0(N) \tag{2}$$
The required \( N \) can be identified from (1) through binary search and the corresponding utilization can be obtained from (2). Working like this, we find that the required number of fixtures is \( N = 2 \) and the resulting utilization is

\[ U(2) \approx 0.47. \]
For parts (i) and (ii), let \( \{X_t, t \geq 0\} \) denote the stochastic process that tracks the number of taxis waiting at the station. Then this process is a CTMC with the following structure:

![Diagram of CTMC structure with states 0, 1, 2, 3, 4, ... and transition rates \( \lambda \) and \( \mu \).]

where the transition rates \( \lambda \) and \( \mu \) are defined as follows:

* \( \lambda = \) taxi arrival rate = \( \lambda / \text{min} \)
* \( \mu = \) customer service rate = \( \mu / \text{min} \)

In fact, the above CTMC is a birth-death process with identical structure to the birth-death process that models the dynamics of an \( N/M/1 \) queue with arrival rate \( \lambda \) and processing rate \( \mu \).

Hence, working as in that case, we can infer the following:

The above CTMC is ergodic (i.e., it reaches steady-state) iff \( \lambda / \mu < 1 \). Then, the steady-state distribution is given by

\[
\pi_i = (1-p)^i p, \quad i = 0, 1, 2, \ldots
\]

Furthermore, the average number of taxis waiting is given by

\[
\sum_{i=0}^{\infty} \pi_i i = \frac{\mu}{\mu - \lambda}.
\]

Finally, an arriving customer will get a taxi as long as the considered CTMC is in a state \( X_t = i \geq 1 \). But the total probability for this class of states is

\[
1 - \pi_0 = 1 - (1-p) = p.
\]

For the provided values for \( \lambda \) and \( \mu \), we can see that \( \lambda / \mu = 1/2 \) and the considered system is ergodic. The average number of taxis waiting is equal to \( \frac{1}{2} \). Also, an arriving customer will be served with prob. 0.5.
To answer part (iii), we need also to trace the evolution of the customer queue in our definition of the system state. Recognizing, however, that only one of the two queues (i.e., taxi or customer) can be non-empty at any time point, the necessary extension of the CTMC structure can be performed as follows:

In the above state space, a negative state models accumulation of customers and a positive state models accumulation of taxis. Furthermore, under an appropriate relabeling of the states, the process retains the structure of the CTMC that models an M/M/1 queue with respective arrival and service rates \( \lambda \) and \( \mu \). Hence, under the assumption that \( \rho = \frac{\lambda}{\mu} < 1 \), the steady-state distribution of this process is characterized as follows:

\[
\pi_i = (1 - \rho)^i \rho^{N+1}, \quad i = -N, -N+1, -N+2, \ldots
\]

But then,

\[
\text{Avg # of taxis waiting} = \sum_{i=1}^{\infty} i (1-\rho)^i \rho^{N+1} = \rho^{N+1} \sum_{i=1}^{\infty} i \rho^i = \rho^{N+1} \frac{\rho}{1-\rho} = \frac{\rho^{N+1}}{1-\rho}.
\]

Also, the probability an arriving customer will get a taxi is \( 1 - \pi_{-N} = 1 - (1 - \rho) = \rho \).
The average waiting time $W$ for a customer who joined the queue can be expressed as follows:

$$W = \sum_{i=N}^{i} a_i \tau_i$$

where

* $a_i = \text{prob. that the customer found the system in state } i$,
* $\tau_i = \text{expected waiting time given that the customer found the system in state } i$.

From PASTA we have that:

$$a_i = \begin{cases} \emptyset & , \quad i = 0 \\ \frac{\pi_i}{1 - \pi_N} & , \quad i = -N, -N+1, \ldots \end{cases}$$

Also, from the memoryless property of the exp. distribution that characterizes the taxi inter-arrival times, we get that:

$$\tau_i = \begin{cases} (-i+1)/\lambda & , \quad i = -N+1, -N+2, \ldots, 0 \\ 0 & , \quad \text{otherwise} \end{cases}$$

Hence,

$$W = \frac{1}{1 - \pi_N} \left[ \frac{\pi_{-N}}{\lambda} \frac{N}{2} + \frac{\pi_{-N+1}}{\lambda} \frac{N-1}{2} + \cdots + \frac{\pi_N}{2} \frac{1}{2} + \pi_0 \frac{1}{2} \right] =$$

$$= \frac{1-p}{p^2} \left[ N p + (N-1) p^2 + \cdots + 2 p^{N-1} + p^N \right] =$$

$$= \frac{1-p}{p^2} \sum_{i=0}^{N-1} (N-i) p^i = \frac{1-p}{p^2} \left[ N \sum_{i=0}^{N-1} p^i - \sum_{i=0}^{N-1} i p^i \right]$$
Since we need \( p < 1 \) for stability (ergodicity), we proceed as follows:
\[
\sum_{i=0}^{N-1} p^i = \frac{1-p^N}{1-p}
\]
\[
\sum_{i=0}^{N-1} i p^i = p \sum_{i=1}^{N-1} i p^{i-1} = p \frac{d}{dp} \left( \sum_{i=1}^{N-1} p^i \right) = p \frac{d}{dp} \left( \frac{1-p^N}{1-p} \right) =
\]
\[
= p \frac{d}{dp} \left( \frac{1-p^N}{1-p} \right) = \frac{p}{(1-p)^2} \left[ (1-p)(-Np^{N-1}) - (1-p^N)(-1) \right] =
\]
\[
= \frac{p}{(1-p)^2} \left[ (N-1)p^N - Np^{N-1} + 1 \right]
\]
Hence,
\[
W = \frac{1}{2} \left\{ N(1-p^N) - \frac{p}{1-p} \left[ (N-1)p^N - Np^{N-1} + 1 \right] \right\} =
\]
\[
= \frac{1}{2(1-p)} \left[ Np^N - Np + Np^{N+1} - (N-1)p^N + Np - p \right] =
\]
\[
= \frac{1}{2(1-p)} \left[ N - (N+1)p + p^{N+1} \right]
\]
The above expression is pretty compact and therefore useful for computation. On the other hand, it can also be processed further as follows:
\[
W = \frac{1}{2} \left\{ N(1-p) - p(1-p^N) \right\} = \frac{1}{2} \left[ N - p \sum_{i=0}^{N-1} p^i \right] =
\]
\[
= \frac{1}{2} \sum_{i=1}^{N} (1-p^i)
\]
What is an intuitive interpretation of the last expression?
\[ W = \frac{1}{2} \sum_{i=1}^{N} (1 - p^i) = \frac{1}{2} \times \text{Prob (joining customer waits for his own cab)} + \]
\[ + \frac{1}{2} \times \text{Prob (joining cab of another customer)} + \]
\[ + \frac{1}{2} \times \text{Prob (joining customer also waits for the cab of a third customer)} + \ldots + \]
\[ + \frac{1}{2} \times \text{Prob (joining customer also waits for the cab of the (N-1)st customer in the queue)} \]

\[ \text{Prob (joining customer waits for his own cab)} = \sum_{i=N+1}^{\infty} a_i = 1 - \sum_{i=1}^{\infty} a_i = \]
\[ = 1 - \sum_{i=1}^{\infty} \frac{p^i}{1-p} = 1 - \frac{1}{p} \sum_{i=1}^{\infty} (1-p)^i = 1 - (1-p)^N \sum_{i=0}^{\infty} p^i = \]
\[ = 1 - p^N \]

Similarly,
\[ \text{Prob (joining customer also waits for the cab of the (N-1)st customer)} = \sum_{i=N+1}^{\infty} a_i = 1 - \sum_{i=1}^{\infty} a_i = 1 - \frac{1}{p} \sum_{i=N+2}^{\infty} (1-p)^i = \]
\[ = 1 - \frac{(1-p)^2}{p} \sum_{i=0}^{\infty} p^i = 1 - p \]

The probabilities for the other terms in the above sum can be computed in the same manner.
An alternative computation of \( \lambda \) can be developed through Little's law. Let \( \lambda \) denote the average number of customers waiting for a taxi. Then, from Little's law,

\[
W = \frac{\lambda}{\rho \mu} = \frac{\lambda}{\eta}
\]

where we have taken into consideration that from the stream of the arriving customers, only a percentage \( p \) of them joins the queue.

From its definition, \( \lambda \) itself can be computed as follows:

\[
\lambda = \sum_{i=-N}^{-1} (-i) \pi_i = -\sum_{i=-N}^{-1} i (1-p)^{N+i} = (1-p) \sum_{i=1}^{N} i p^{i-1} = (1-p)^N \sum_{i=1}^{N-1} i p^{i-1} = (1-p)^{N-1} \sum_{i=1}^{N-1} i u^{i-1}
\]

where \( u = \frac{1}{\rho} \)

But \( \sum_{i=1}^{N} i u^{i-1} = \frac{d}{du} \sum_{i=0}^{N-1} u^i = \frac{d}{du} \frac{1-u^{N+1}}{1-u} = \frac{-N u^{N+1} - (N+1) u^N + 1 - u^{N+1}}{(1-u)^2} = \frac{N u^{N+1} - (N+1) u^N + 1 - u^{N+1}}{(1-u)^2} = \)

\[
N \left( \frac{u^{N+1}}{\rho^{N+1}} - \frac{u^{N+1}}{\rho^{N+1}} \right) = \frac{N - (N+1) p + \rho^{N+1}}{\rho^{N+1} (1-p)^2}
\]

Substituting back to the expression for \( \lambda \) we get:

\[
\lambda = (1-p) p^{N-1} \frac{N - (N+1) p + \rho^{N+1}}{\rho^{N-1} (1-p)^2} = \frac{N - (N+1) p + \rho^{N+1}}{1-p}
\]

And finally,

\[
W = \frac{1}{2(1-p)} \left[ N - (N+1) p + \rho^{N+1} \right]
\]
For part (iv), notice that when \( A \) increases to 2 taxis per minute, \( p = 1 \), and therefore, the considered birth-death process becomes unstable. In this regime, the \# of taxis waiting will grow to infinity and therefore arriving customers will always get a taxi immediately.
Consider the M/M/1 queuing system operating in steady state and let:
- \( W \) denote the random time spent in system by an arriving customer,
- \( N \) denote the random number of customers in the system that are encountered by the aforementioned customer upon his arrival,
- \( F_{W}(t) \) denote the cdf for \( W \).

Then,
\[
F_{W}(t) = P(W \leq t) = \sum_{n=0}^{\infty} P(W \leq t | N=n) P(N=n) = \sum_{n=0}^{\infty} F_{\text{Er} \{n+1\}}(t) \cdot a_n
\]
where the last inequality results by:
- setting \( a_n = P(N=n) \), and
- taking into consideration that in an M/M/1 queue
  \* customers are processed in a FCFS basis, and
  \* their processing times are iid rv's distributed according to an exponential distribution with rate \( \rho \).

Furthermore, from the Markovian nature of the arrival process and the MTA, we have:
\[
a_n = \frac{a_{n-1}}{1 - p} = (1 - p)^n \quad \text{where} \quad p = \frac{\lambda}{\mu} < 1 \quad (\text{from stability})
\]

Then,
\[ F_n(t) = (1 - \rho) \sum_{n=0}^{\infty} \rho^n \left(1 - e^{-tk} \sum_{j=0}^{\infty} \frac{(kt)^j}{j!}\right) = \]

\[ = (1 - \rho) \sum_{n=0}^{\infty} \rho^n - (1 - \rho) e^{-tk} \sum_{n=0}^{\infty} \rho^n \sum_{j=0}^{\infty} \frac{(kt)^j}{j!} = * \]

\[ = 1 - (1 - \rho) e^{-tk} \sum_{n=0}^{\infty} \frac{(\rho t)^n}{n!} \sum_{j=0}^{\infty} \frac{(kt)^j}{j!} = \]

\[ = 1 - (1 - \rho) e^{-tk} \sum_{n=0}^{\infty} \frac{1}{1-\rho} \frac{(\rho t)^n}{n!} \frac{(kt)^j}{j!} = \]

\[ = 1 - e^{-kt} e^{\rho t} = 1 - e^{-(\rho - 1)t} \]

But this is the cdf of an exponential distribution with rate \( \rho - 1 \).

To understand better the step denoted by \((\ast)\) in the above derivation, consider the following figure:

![Diagram](image)

The double sums are computed for the pairs indicated by the dotted matrix above, and in the first expression the summation is computed on a column by column basis, while in the second expression on a row by row basis.
An insightful interpretation of this result is as follows:

From the dynamics of the M/M/1 queue we have that

\[ \text{Prob (an arriving customer leaves at the n-th processing completed after his arrival)} = \]

\[ = \text{Prob (arriving customer encountered n-1 customers when he joined the queue)} = \]

\[ = T_{n-1} = p^{n-1} (1-p), \quad n = 2, 3, \ldots \]

The next-to-last step in the above derivation results from PASTA. Furthermore, the above result implies a geometric distribution with "success" prob. equal to 1-p; each "trial" in the interpretation of this distribution corresponds to a job completion, and the "successful trial" is the job completion that corresponds to the "tagged" customer. While the customer is in the system, the server must be continuously busy, and therefore, during this time job completion take place with rate \( p \). The above discussion further implies that any of these completions will be the completion of our tagged customer with prob. \( 1-p \), and therefore, by the properties of the Poisson process, the completion of our tagged customer occurs with rate \( p(1-p) = p - 2p \). Since this rate is constant, the distribution that characterizes the "time-to-completion" (i.e., otherwise, the sojourn time) in our customer is exponential with the same rate.
For parts (1-3) of this problem, you can see the pages 7 and 9 from the excerpt by Chen and Yao on "Birth-Death processes and Jackson queuing networks" that has been posted at the library academic reserve. Part (4) follows directly from the result of part (3).

An alternative derivation of part (4) had revealed the connection of this result to some other results presented in [ref], and also to part (c) of this homework, as follows:

Let \( X(t_a) = \) # of customers that are encountered by an arrival at steady state

Then, \( S = \) the sojourn time of this arrival

\[
P(X(t_a) = n | S = t) = \lim_{\Delta t \to 0} \frac{P(X(t_a) = n \land S \in [t, t+\Delta t])}{P(S \in [t, t+\Delta t])}
\]

\[
= \lim_{\Delta t \to 0} \frac{P(X(t_a) = n) \cdot P(S \in [t, t+\Delta t])}{P(S \in [t, t+\Delta t])} / \Delta t
\]

\[
= \frac{pdf[\text{Geometric}(n; \lambda)] \cdot pdf[\text{Erlang}(t; n+1, \mu)]}{pdf[\text{Exponential}(t; \mu)]}
\]

\[
= \frac{\left[ (\frac{2}{\mu})^n (1-\frac{2}{\mu}) \right] \cdot \left[ \mu^{n+1} \cdot e^{-\mu t} / n! \right]}{(\mu-2) e^{-(\mu-2) t}}
\]

\[
= \frac{\left[ (\frac{2}{\mu})^n (1-\frac{2}{\mu}) \right] \cdot \left[ \mu^{n+1} \cdot e^{-\mu t} / n! \right]}{(\mu-2) e^{-(\mu-2) t}}
\]

\[
= \text{pdf}[\text{Poisson}(n; 2t)].
\]
Consider the $i$-th arrival. Due to the preemptive nature of the applied LIFO policy, this job goes immediately in service and it will finish in $Z_i$ time units if $Z_i \leq Y_{i+2}$.

If, on the other hand, $Z_i > Y_{i+1}$, then this job will be preempted by the next arrival, and it will finish in $Z_i + Z_{i+1}$ time units if $Z_i + Z_{i+1} \leq Y_{i+1} + Y_{i+2}$.

More generally, let $j = \min \{ j = 0, 1, 2, \ldots \text{s.t. } \sum_{k=i}^{i+j} Z_k \leq \sum_{k=i+1}^{i+j+1} Y_k \}$

Then, reasoning as above, we can see that the sojourn time in job $i$ is

$$S_i = \sum_{k=i}^{i+j} Z_k$$
Consider the $n$-th arrival at this queue. Assuming that the first arrival took place at time $t = \lambda$, the $n$-th arrival takes place at time $t = (n-1)\tau$. Since service times are exponentially distributed with rate $\mu$, the $i$-th arrival, for $i = 1, \ldots, n-1$, will be still in service upon the arrival of the $n$-th job with prob. $P_i = e^{-\mu \tau i}$. Hence, setting $A_n = \#$ of jobs encountered by the $n$-th arrival, we have that

$$E[a_n] = E \left[ \sum_{i=1}^{n-1} I_{\{i\text{-th job still in service upon n-th arrival}\}} \right] =$$

$$= \sum_{i=1}^{n-1} E \left[ I_{\{i\text{-th job still in service}\}} \right] = \sum_{i=1}^{n-1} P_i = \sum_{i=1}^{n-1} e^{-\mu \tau i} =$$

$$= \sum_{j=1}^{n-1} e^{-\mu \tau j}, \text{ where in the last step we have set } j = n-i.$$

In steady state, we have:

$$E[A] = E \left[ \lim_{n \to \infty} A_n \right] = E \left[ \lim_{n \to \infty} \sum_{i=1}^{n-1} I_{\{i\text{-th job still in service}\}} \right] =$$

$$= \lim_{n \to \infty} E \left[ \sum_{i=1}^{n-1} I_{\{i\text{-th job still in service}\}} \right] = \lim_{n \to \infty} \sum_{j=1}^{n-1} e^{-\mu \tau j} =$$

$$= \sum_{j=1}^{\infty} (e^{-\mu \tau})^j, \text{ where the interchange of the lim and the expectation operation is justified by monotone convergence. Furthermore, since } \mu \tau > 0, \text{ we have:}$$

$$E[A] = \sum_{j=0}^{\infty} (e^{-\mu \tau})^j - 1 = \frac{1}{1 - e^{-\mu \tau}} - 1 = \frac{1}{e^{\mu \tau} - 1}.$$
(ii) The difference between parts (i) and (ii) is that in the second case, the system is observed at some arbitrary time point $t$. Suppose that $t \in [(n-1)\tau, n\tau)$, i.e., $t$ is the time point of the $n$-th arrival at a point between this arrival and the next one, for some arbitrary $n$. Hence, $T = (n-1)\tau + T$, where $T$ is a r.v. uniformly distributed in $[0, \tau)$. Let $N(t)$ denote the number of jobs in service at time $t$.

Then, writing as in part (i) above, we can see that:

$$E[X] = E\left[\lim_{t \to \infty} X(t)\right] = E_T\left[E\left[\lim_{t \to \infty} X(t+T)\right]\right] =$$

$$= E_T\left[E\left[\lim_{j \to \infty} \sum_{j=0}^{\infty} I_{\text{arrival still in service at time } t+j\tau}\right]\right] =$$

$$= E_T\left[\sum_{j=0}^{\infty} e^{-\lambda(j+1)\tau}\right]$$,

where again we have interchanged the internal expectation and the limit thanks to the monotone convergence theorem. Proceeding with the above calculation, we have:

$$E[X] = E_T\left[\sum_{j=0}^{\infty} e^{-\lambda j \tau} (e^{-\lambda \tau})^j\right] = E_T\left[e^{-\lambda \tau} \frac{1}{1-e^{-\lambda \tau}}\right] =$$

$$= \frac{1}{1-e^{-\lambda \tau}} \int_0^\tau e^{-\lambda y} \frac{d\tau}{y} = \frac{1}{1-e^{-\lambda \tau}} \frac{1}{\tau} \left(\frac{1}{\lambda}\right) \left[ e^{-\lambda \tau} - 1 \right] =$$

$$= \frac{1}{\tau} \frac{1}{\lambda},$$

where we set $\lambda = \frac{1}{\tau}$, i.e., the arrival rate.

Notice that the obtained result admits the standard interpretation of the traffic intensity $\rho = \frac{\lambda}{\mu}$, i.e., the expected number of busy servers (which in this case coincides with the expected number of jobs in the system) is equal to the expected workload that is received each time unit.
(iii) For this part, notice that \( e^{\lambda x} - 1 = e^{\lambda x} - (1 + \lambda x) = \sum_{i=2}^{\infty} \frac{(\lambda x)^i}{i!} > 0. \)

Hence, \( E[A] = \frac{1}{e^{\lambda x} - 1} < \frac{1}{\lambda x} = E[X]. \)

This last result further implies that the two distributions that define the above means are different. Hence, the considered system does not present any PASTA-type effect.

(iv) To answer this part, first notice that by the problem statement, arrivals occur one at a time. Also, the exponential nature of the service times, combined with the independence of the service times, imply that the departure process is (non-homogeneous) Poisson and therefore, the probability of having more than one departure at the same time is zero. But then, from the relevant theory discussed in class, we can infer that the sought expectation is equal to \( E[A]. \)