Priority Queues

- We have considered so far systems with only FCFS (or FIFO) queue discipline.
- Alternative queue disciplines include LCFS, selection in random order (SRO), priority, etc.
- In priority schemes, the higher priorities are selected for service ahead of those with lower priorities.
  - Two ways: With preemption and without preemption.

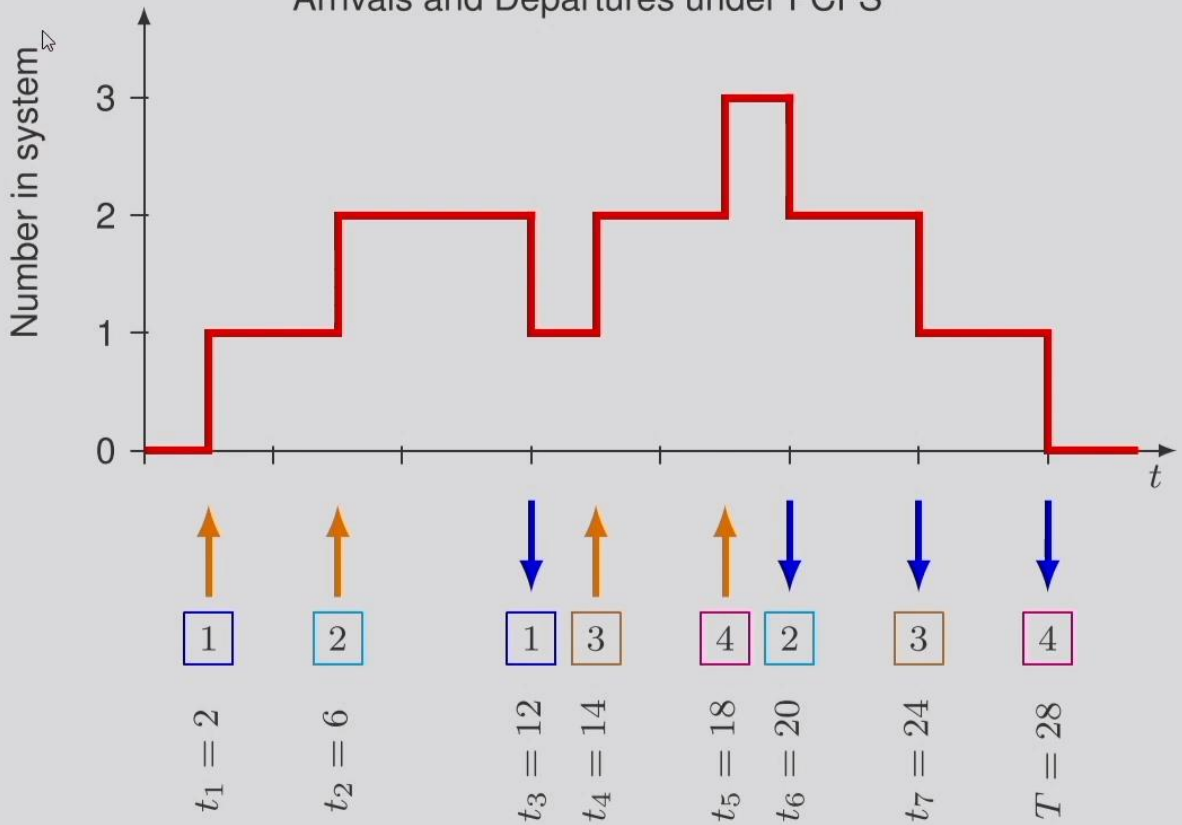
Priority Queues

- We have considered so far systems with only FCFS (or FIFO) queue discipline.
- Alternative queue disciplines include LCFS, selection in random order (SRO), priority, etc.
- In priority schemes, the higher priorities are selected for service ahead of those with lower priorities.
  ♦ Two ways: With preemption and without preemption.
- In preemptive case, the customer with the higher priority either gets into service immediately on arrival interrupting the service (if any) of the lower priority customer, or gets into the queue if the ongoing service is of the same or higher priority customer.
  ► When the preempted customer returns to the service, the service can be resumed from the point of interruption (Preemptive-resume) or can be started afresh (Preemptive-non-resume).
- In the non-preemptive case, the higher priority customer goes into the queue (ahead of any lower priority customers), but must wait for any customer (irrespecive of priority) in service to complete the service.
- We use 1 to denote highest priority customer and 2 to denote second highest, and so on (smaller the number higher the priority).
  [In case of two priorities, it is simply *higher* and *lower*, for convenience]

- Priority queues are difficult to analyze in general (compared to nonpriority ones).
- Recall that, for the $M/M/1$ queue (and some other models too), the derivation of steady state system size probabilities did not depend on the queue discipline.
- It can be shown that as long as selection of customers for service is independent of the relative size of the service time, $\{p_n\}$ are independent of queue discipline.
- In such cases, Little's formula remains unchanged and hence the average waiting time remains the same.
  - ► But there will be changes in the waiting time distributions (again recall our $M/M/1$ analysis).

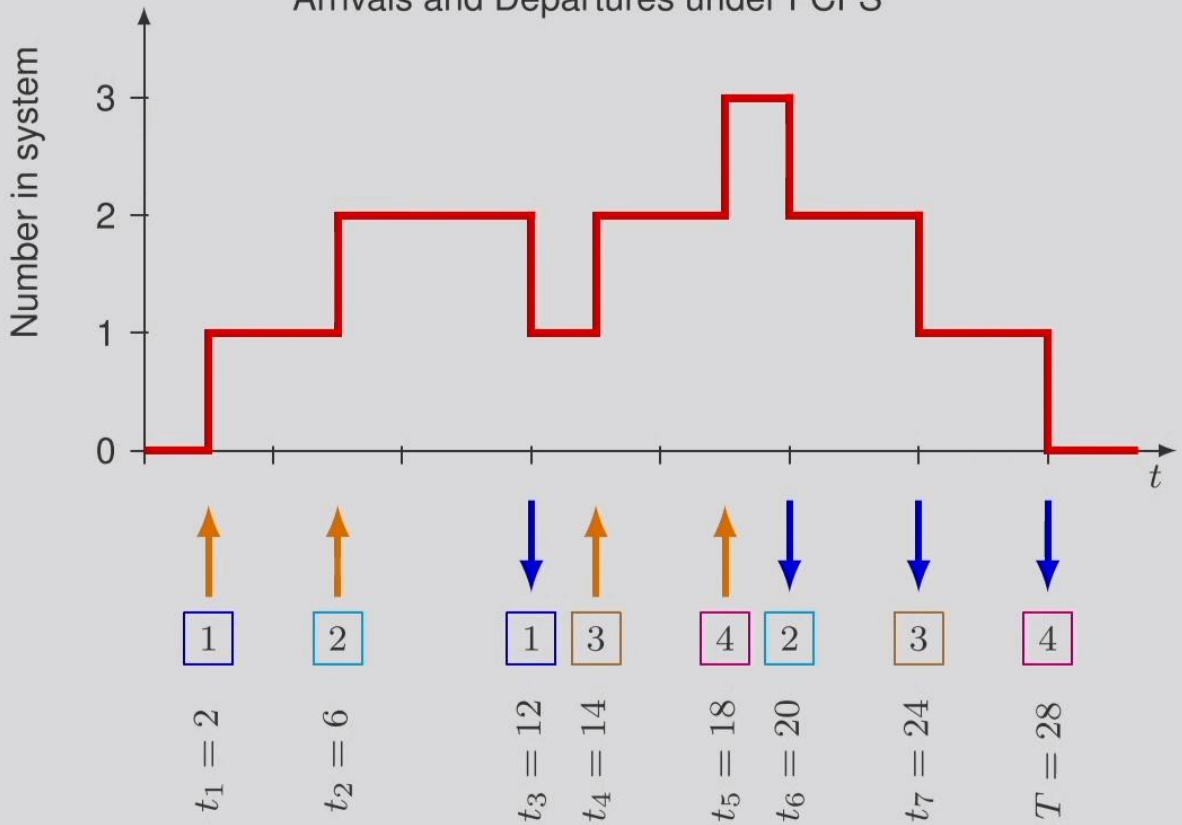
- Priority queues are difficult to analyze in general (compared to nonpriority ones).
- Recall that, for the $M/M/1$ queue (and some other models too), the derivation of steady state system size probabilities did not depend on the queue discipline.
- It can be shown that as long as selection of customers for service is independent of the relative size of the service time, $\{p_n\}$ are independent of queue discipline.
- In such cases, Little's formula remains unchanged and hence the average waiting time remains the same.
  ► But there will be changes in the waiting time distributions (again recall our $M/M/1$ analysis).
- Claim: Waiting times are stochastically smallest under FCFS (with all other things being equal), i.e., introduction of any scheme of priorities not depending on service times makes higher order moments worse than under FCFS.
- With a lower variance, FCFS can be viewed as the most fair scheme (lower variance means greater equality among waiting times).
- Though it would be difficult to prove the above at this point, we can get an idea about the happening in a single-server queue context with an illustration.

Arrivals and Departures under FCFS



- Under FCFS, the waiting times are $10, 14, 10$ and 10, with mean as 11 and unbiased variance as 4.
- Under LCFS, the waiting times are $10, 14, 6$ and 14, with mean as 11 and unbiased variance as $44/3 = 14.67$. [customer-4 departs at 24 and customer-3 at 28]
- Variance higher under LCFS.

Arrivals and Departures under FCFS



- Under FCFS, the waiting times are $10, 14, 10$ and 10, with mean as 11 and unbiased variance as 4.
- Under LCFS, the waiting times are $10, 14, 6$ and 14, with mean as 11 and unbiased variance as $44/3 = 14.67$. [customer-4 departs at 24 and customer-3 at 28]
- Variance higher under LCFS.

- Furthermore, the remaining total service or work required for a single server at any point during an arbitrary busy period is independent of the order of service as long as the system is conservative (i.e., no service needs are created or destroyed with the system).

Nonpreemptive Priority Systems with Two Classes

- Customers arrive as a Poisson process to a single exponential channel.
- A customer, upon arrival, is assigned to one of the two priority classes.
- First or higher priority customers arrive as PP(λ_1) and the second or lower priority customers arrive as PP(λ_2).
  - ▶ The total arrival rate is $\lambda = \lambda_1 + \lambda_2$.
- Assume that there is no preemption.
- The system can be modelled by a CTMC with state space $S = \{0\} \cup \{(m, n, r) : m, n = 0, 1, 2, \ldots, \max\{m, n\} > 0, r = 1, 2\}$ and the corresponding steady state probabilities (where m & n not both 0) are denoted by

$$p_{mnr} = P\{m \text{ priority-1 customers in the system,}$$
$$n \text{ priority-2 customers in the system, and}$$
$$\text{the customer in service is of priority } r = 1 \text{ or } 2\}$$
$$p_0 = P\{\text{the system is empty}\}$$

♦ Let $L^{(i)}, L_q^{(i)}, W_q^{(i)}, W^{(i)}$ denote the measures of effectiveness for class-i customers.

Model (A): Equal Service Rates

- Assume that the service rates of both the classes are equal to μ. Define

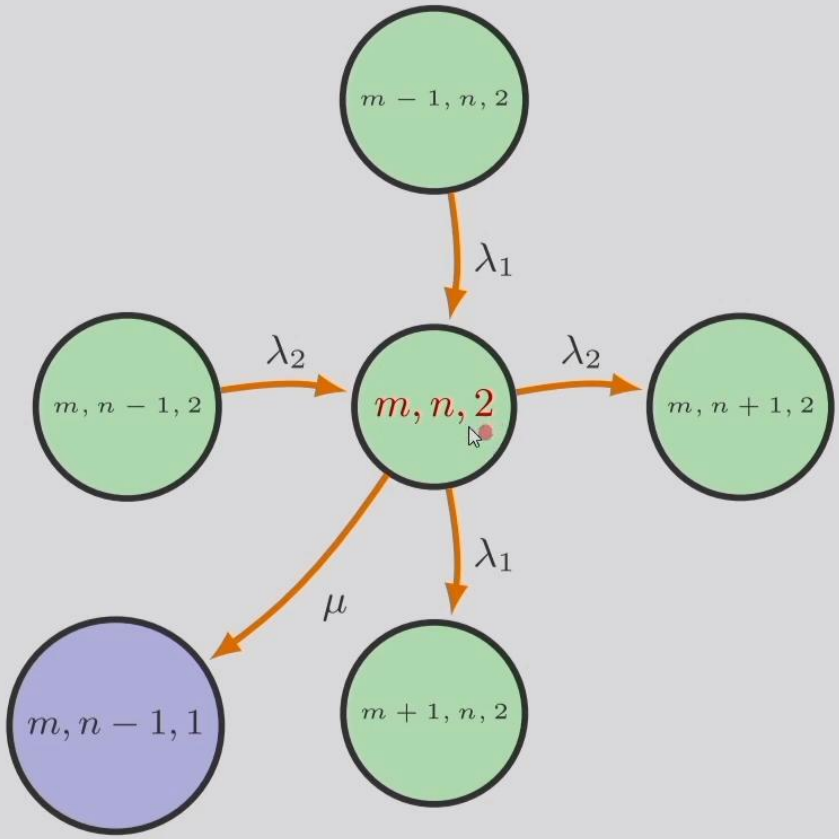
$$\rho_1 = \frac{\lambda_1}{\mu}, \quad \rho_2 = \frac{\lambda_2}{\mu}, \quad \rho = \rho_1 + \rho_2 = \frac{\lambda}{\mu}$$

  We assume that $\rho < 1$.

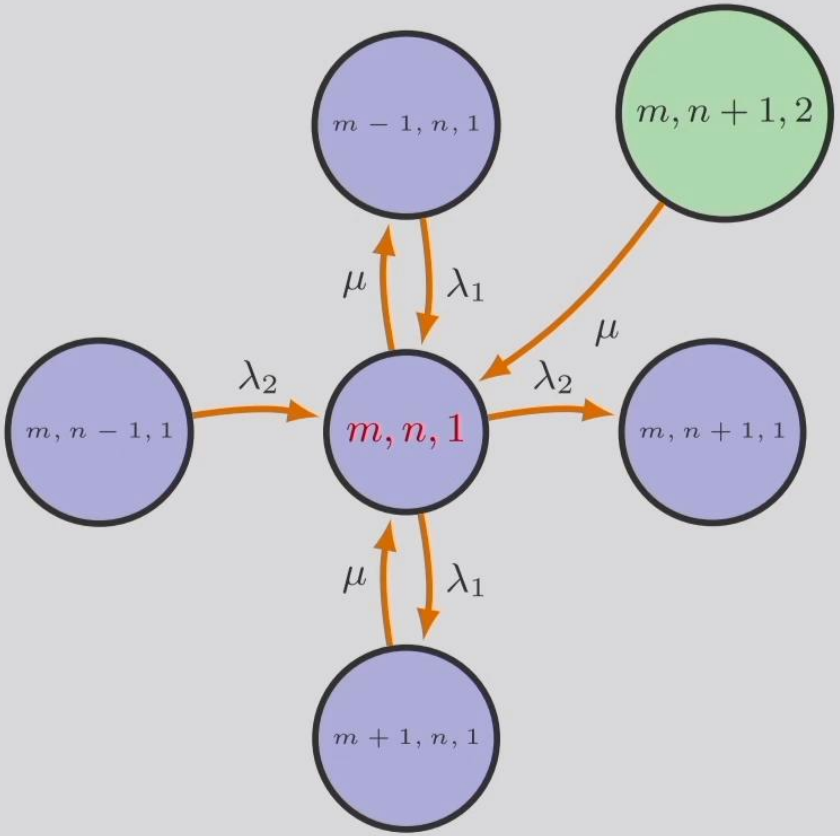
- The balance equations are:

$$\begin{aligned}
(\lambda+\mu)p_{mn2} &= \lambda_1 p_{m-1,n,2} + \lambda_2 p_{m,n-1,2} && (m \geq 1, \quad n \geq 2) \quad \text{(Eq.A)} \\
(\lambda+\mu)p_{mn1} &= \lambda_1 p_{m-1,n,1} + \lambda_2 p_{m,n-1,1} + \mu(p_{m+1,n,1} + p_{m,n+1,2}) && (m \geq 2, \quad n \geq 1) \quad \text{(Eq.B)} \\
(\lambda+\mu)p_{m12} &= \lambda_1 p_{m-1,1,2} && (m \geq 1) \\
(\lambda+\mu)p_{1n1} &= \lambda_2 p_{1,n-1,1} + \mu(p_{2n1} + p_{1,n+1,2}) && (n \geq 1) \\
(\lambda+\mu)p_{0n2} &= \lambda_2 p_{0,n-1,2} + \mu(p_{1n1} + p_{0,n+1,2}) && (n \geq 2) \\
(\lambda+\mu)p_{m01} &= \lambda_1 p_{m-1,0,1} + \mu(p_{m+1,0,1} + p_{m12}) && (m \geq 2) \\
(\lambda+\mu)p_{012} &= \lambda_2 p_0 + \mu(p_{111} + p_{022}) \\
(\lambda+\mu)p_{101} &= \lambda_1 p_0 + \mu(p_{201} + p_{112}) \\
\lambda p_0 &= \mu(p_{101} + p_{012})
\end{aligned}$$

State Transitions for $(m, n, 2)$ (Eq.A above)



State Transitions for $(m, n, 1)$ (Eq.B above)



- By Little's law (to the server), ρ is the fraction of time the server is busy, or equivalently $p_0 = 1 - \rho$.
- Similarly, the fraction of time the server is busy with a priority-r customer is ρ_r. Thus,

$$\sum_{m=1}^{\infty}\sum_{n=0}^{\infty} p_{mn1} = \rho_1 \quad \text{and} \quad \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} p_{mn2} = \rho_2.$$

- And, since the service times are $Exp(\mu)$ for both the priority classes, the system size steady state distribution for the number of customers is

$$p_n = \sum_{m=0}^{n-1} (p_{n-m,m,1} + p_{m,n-m,2}) = (1-\rho)\rho^n, \quad n > 0.$$

- Obtaining a complete analytical solution is very difficult here. We will try to obtain the expected value measures.

- Define two-dimensional generating functions as:

$$P_{m1}(z) = \sum_{n=0}^{\infty} p_{mn1} z^n \quad (m \geq 1), \qquad P_{m2}(z) = \sum_{n=1}^{\infty} z^n p_{mn2} \qquad (m \geq 0),$$

$$H_1(y,z) = \sum_{m=1}^{\infty} y^m P_{m1}(z) \quad (\text{with } H_1(1,1) = \rho_1),$$

$$H_2(y,z) = \sum_{m=0}^{\infty} y^m P_{m2}(z) \quad (\text{with } H_2(1,1) = \rho_2),$$

and
$$\begin{aligned} H(y,z) &= H_1(y,z) + H_2(y,z) + p_0 \\ &= \sum_{m=1}^{\infty}\sum_{n=0}^{\infty} y^m z^n p_{mn1} + \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} y^m z^n p_{mn2} + p_0 \\ &= \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} y^m z^n (p_{mn1} + p_{mn2}) + \sum_{m=1}^{\infty} y^m p_{m01} + \sum_{n=1}^{\infty} z^n p_{0n2} + p_0, \end{aligned}$$

where $H(z,y)$ is the joint generating function for the two classes, regardless of which type is in service.

- Note that $H(y,y) = \dfrac{p_0}{(1-\rho y)}$, with $H(1,1) = 1$, since $H(y,z)$ collapses to the generating function of an $M/M/1$ queue when $z = y$ and thus no priority distinction is made. Also,

$$\left.\frac{\partial H(y,z)}{\partial y}\right|_{y=z=1} = L^{(1)} = L_q^{(1)} + \rho_1 = \lambda_1 W^{(1)},$$

$$\left.\frac{\partial H(y,z)}{\partial z}\right|_{y=z=1} = L^{(2)} = L_q^{(2)} + \rho_2 = \lambda_2 W^{(2)}$$

- Multiplying the balance equations by the appropriate powers of y and z and summing, we have

$$(1+\rho-\rho_1 y-\rho_2 z-\frac{1}{y})H_1(y,z)=\frac{H_2(y,z)}{z}+\rho_1 y p_0-P_{11}(z)-\frac{P_{02}(z)}{z},$$

$$(1+\rho-\rho_1 y-\rho_2 z)H_2(y,z)=P_{11}(z)+\frac{P_{02}(z)}{z}-(\rho-\rho_2 z)p_0$$

- In order to determine H_1 and H_2 completely, we need the values of $P_{11}(z), P_{02}(z)$and p_0. An equation giving the relationship between these quantities can be found by summing z^n $(n=2,3,\dots)$ times the balance equations involving p_{0n2}, and then using the final three balance equations. This gives

$$P_{11}(z)=\left(1+\rho-\rho_2 z-\frac{1}{z}\right)P_{02}(z)+(\rho-\rho_2 z)p_0.$$

- Substituting the above equation for $P_{11}(z)$ into the previous equations gives H_1 and H_2 as functions of p_0 and $P_{02}(z)$.

- We can thus obtain $H(y,z)$ also in terms of p_0 and $P_{02}(z)$ as

$$H(y,z) = H_1(y,z) + H_2(y,z) + p_0.$$
$$= \frac{(1-y)p_0}{1-y-\rho y+\rho_1 y^2+\rho_2 yz} + \frac{(1+\rho-\rho z+\rho_1 z)(z-y)P_{02}(z)}{z(1+\rho-\rho_1 y-\rho_2 z)(1-y-\rho y+\rho_1 y^2+\rho_2 yz)}.$$

- From $H(1,1)=1$, we get $P_{02}(1)=\frac{\rho_2}{1+\rho_1}$. We can then determine $L^{(1)}$ from $L^{(1)} = \left.\frac{\partial H(y,z)}{\partial y}\right|_{y=z=1}$ (the partial derivative cannot be evaluated directly at $(1,1)$ so a limit must be taken). Fortunately, in these steps, only $P_{02}(1)$ is required, and not the function $P_{02}(z)$.

- Since the total number of customers in the system is the same as that of the $M/M/1$ system, we have that $L^{(1)} + L^{(2)} = \dfrac{\rho}{1-\rho}$ and hence $L^{(2)} = \dfrac{\rho}{1-\rho} - L^{(1)}$.
- The other measures can be obtained from

$$L_q^{(i)} = L^{(i)} - \rho_i, \quad L_q^{(i)} = \lambda_i W_q^{(i)}, \quad L^{(i)} = \lambda_i W^{(i)}$$

- The final results for $L_q^{(i)}$ are (recall that $\rho = \lambda_1/\mu + \lambda_2/\mu$)

$$\boxed{L_q^{(1)} = \frac{\lambda_1 \rho}{\mu - \lambda_1}, \qquad L_q^{(2)} = \frac{\lambda_2 \rho}{(\mu - \lambda_1)(1-\rho)}, \qquad L_q = \frac{\rho^2}{1-\rho}}$$

- Extra (for interested, refer Miller (1981)): The actual probabilities for priority-1 customers can be shown to be given by

$$p_{n_1} = (1-\rho)\left(\frac{\lambda_1}{\mu}\right)^{n_1} + \frac{\lambda_2}{\lambda_1}\left(\frac{\lambda_1}{\mu}\right)^{n_1}\left[1 - \left(\frac{\mu}{\lambda_1 + \mu}\right)^{n_1+1}\right] \quad (n_1 \geq 0).$$

Some Observations on the Mean-Value Results

1. Lower priority customers always wait in queue longer (on average) than the higher priority customers. This can be seen as follows:

$$W_q^{(2)} = \frac{\rho}{(\mu - \lambda_1)(1-\rho)} = \frac{\left(\frac{\rho}{\mu - \lambda_1}\right)}{1-\rho} = \frac{W_q^{(1)}}{1-\rho} > W_q^{(1)} \quad \text{(when } \rho < 1\text{)}.$$

However, it is not always the case that $L_q^{(2)} > L_q^{(1)}$.

Some Observations on the Mean-Value Results

1. Lower priority customers always wait in queue longer (on average) than the higher priority customers. This can be seen as follows:

$$W_q^{(2)} = \frac{\rho}{(\mu - \lambda_1)(1-\rho)} = \frac{\left(\frac{\rho}{\mu - \lambda_1}\right)}{1-\rho} = \frac{W_q^{(1)}}{1-\rho} > W_q^{(1)} \quad \text{(when } \rho < 1\text{)}.$$

   However, it is not always the case that $L_q^{(2)} > L_q^{(1)}$.

2. As $\rho \to 1, L_q^{(2)} \to \infty$ (and so do $W_q^{(2)}, W^{(2)}$, and $L^{(2)}$). However, $L_q^{(1)}$ approaches a finite limit, if $\frac{\lambda_1}{\mu} < 1$ is held constant. The first-priority means go to ∞ only when $\frac{\lambda_1}{\mu} \to 1$.
   ▶ Possible that higher priority customers do not accumulate, even when an overall steady state does not exist.

Some Observations on the Mean-Value Results

1. Lower priority customers always wait in queue longer (on average) than the higher priority customers. This can be seen as follows:

$$W_q^{(2)} = \frac{\rho}{(\mu-\lambda_1)(1-\rho)} = \frac{\left(\frac{\rho}{\mu-\lambda_1}\right)}{1-\rho} = \frac{W_q^{(1)}}{1-\rho} > W_q^{(1)} \quad \text{(when } \rho < 1\text{)}.$$

   However, it is not always the case that $L_q^{(2)} > L_q^{(1)}$.

2. As $\rho \to 1, L_q^{(2)} \to \infty$ (and so do $W_q^{(2)}, W^{(2)}$, and $L^{(2)}$). However, $L_q^{(1)}$ approaches a finite limit, if $\frac{\lambda_1}{\mu} < 1$ is held constant. The first-priority means go to ∞ only when $\frac{\lambda_1}{\mu} \to 1$.
   ▶ Possible that higher priority customers do not accumulate, even when an overall steady state does not exist.

3. The presence of class-2 customers still creates delays for class-1 customers (because of nonpreemptiveness). In particular,

$$\{L_q^{(1)} \text{ when } \lambda_2 = 0\} < \{L_q^{(1)} \text{ when} \lambda_2 > 0\}.$$

   However, if the class-1 customers have the power of preemption, then the class-2 customers do not effect the class-1 customers.

Some Observations on the Mean-Value Results

1. Lower priority customers always wait in queue longer (on average) than the higher priority customers. This can be seen as follows:

$$W_q^{(2)} = \frac{\rho}{(\mu-\lambda_1)(1-\rho)} = \frac{\left(\frac{\rho}{\mu-\lambda_1}\right)}{1-\rho} = \frac{W_q^{(1)}}{1-\rho} > W_q^{(1)} \quad \text{(when } \rho < 1\text{)}.$$

   However, it is not always the case that $L_q^{(2)} > L_q^{(1)}$.
2. As $\rho \to 1, L_q^{(2)} \to \infty$ (and so do $W_q^{(2)}, W^{(2)}$, and $L^{(2)}$). However, $L_q^{(1)}$ approaches a finite limit, if $\frac{\lambda_1}{\mu} < 1$ is held constant. The first-priority means go to ∞ only when $\frac{\lambda_1}{\mu} \to 1$.
   ▶ Possible that higher priority customers do not accumulate, even when an overall steady state does not exist.
3. The presence of class-2 customers still creates delays for class-1 customers (because of nonpreemptiveness). In particular,

$$\{L_q^{(1)} \text{ when } \lambda_2 = 0\} < \{L_q^{(1)} \text{ when} \lambda_2 > 0\}.$$

   However, if the class-1 customers have the power of preemption, then the class-2 customers do not effect the class-1 customers.
4. The average number in queue is the same as an $M/M/1$ queue. Similarly, the unconditional average wait, $W_q = \left(\frac{\lambda_1}{\lambda}\right) W_q^{(1)} + \left(\frac{\lambda_2}{\lambda}\right) W_q^{(2)}$ is the same as an $M/M/1$ queue.