Let $N_{nk}$ = number in queue for priority $k$
$W_k$ = queueing (waiting) time for priority $k$
$R$ = residual service time
$P_k = \frac{\lambda_k}{\mu} = \text{utilization for priority } k$
$N = \text{number of customers in the system}$

*First, we evaluate $E[R]$. Consider the two following cases:

$N < m$

1. Arriving customer goes immediately to service i.e. $R = 0$

$N \geq m$

1. In this case, the residual service time is the time until one server empties
2. Let $T_i$, $i=1, \ldots, m$, be the service time of customer at server $i$
   $T_i \sim \text{exponential, mean} = \frac{1}{\mu}$
   $R = \min\{T_1, \ldots, T_m\}$

$P(R \leq t) = P(\min\{T_i\} \leq t)$
   $= 1 - P(\min\{T_i\} > t)$
   $= 1 - \prod_{i=1}^{m} P(T_i > t)$
   $= 1 - (e^{-\mu t})^m$
   $= 1 - e^{-\mu t}$

$E[R | N < m] = 0$

$E[R | N \geq m] = \frac{1}{\mu}$
\[ E[R] = \frac{P(N<m)}{1-P_0} \cdot E[R|N<m] + \frac{P(N>m)}{P_0} \cdot \frac{1}{m} \cdot E[R|N>m] \]

\[ E[R] = \frac{P_0}{m} \]

*Next, we evaluate \( E[W_i] \)

Priority class 1:

\[ E[W_1] = E[R] + \frac{1}{m} \cdot E[N_{ai}] \]

Intuitively speaking,

- avg waiting = avg residual time + aug time it takes to serve the customers of priority 1 in the queue

\[ = \frac{\text{avg # of customers}}{m} \times \text{avg service time for each} \]

(we divide by \( m \) because the servers work in parallel)

From Little's law, we have

\[ E[N_{ai}] = \lambda_1 \cdot E[W_1] \]

Therefore,

\[ E[W_1] = \frac{E[R]}{1-P_1} \quad \Rightarrow \quad P_1 = \frac{\lambda_1}{m_1} \]

Priority class 2:

\[ E[W_2] = E[R] + \frac{1}{m_2} \cdot E[N_{ai}] + \frac{1}{m_2} \cdot E[N_{a2}] + \frac{1}{m_2} \cdot \lambda_1 \cdot E[W_2] \]

Using Little's theorem, we have \( E[N_{ai}] = \lambda_1 \cdot E[W_1] \) and \( E[N_{a2}] = \lambda_2 \cdot E[W_2] \)

\[ E[W_2] = E[R] + P_1 \cdot E[W_1] + P_2 \cdot E[W_2] + P_1 \cdot E[W_2] \]

\[ E[W_2] = \frac{E[R]}{1-P_1} \quad \Rightarrow \quad P_1 = \frac{\lambda_1}{m_1} \text{ and } P_2 = \frac{\lambda_2}{m_2} \]

\[ E[W_2] = \frac{E[R]}{(1-P_1)} \]

Priority class \( k \):

\[ E[W_k] = \frac{E[R]}{(1-P_1)(1-P_2) \cdots (1-P_{k-1})(1-P_k)} \]

where \( \frac{\lambda_k}{m_k} \)

and \( E[R] = \frac{P_0}{m} \) (\( P_0 \) as defined in sec. 3.4.1, with \( P = P_1 + P_2 + \ldots + P_n \))
Check: substitute $m=1$, $\ldots$, $m$

\[ p_k = \frac{\lambda_k}{\mu} \]

\[ E[W_k] = \frac{E[R]}{(1-p_1\cdots-p_{k-1})(1-p_1\cdots-p_k)} \]

\[ p_0 = 1-p_1 = 1-(1-p) = p = p_1 + p_2 + \cdots + p_n = \sum_{i=1}^{n} \frac{\lambda_i}{\mu} \]

\[ E[R] = \frac{p_0}{\sum_{i=1}^{n} \lambda_i} = \frac{\sum_{i=1}^{n} \lambda_i}{\mu} \]

Compare this with what we get by considering exponential service time in the M/G/1 expressions

\[ E[W_k] = \frac{E[R]}{(1-p_1\cdots-p_{k-1})(1-p_1\cdots-p_k)} \]

\[ E[R] = \frac{1}{2} \sum_{i=1}^{n} \lambda_i \bar{X}^2 \]

$X \sim \text{exponential with mean} = \frac{1}{\mu}$

\[ \Rightarrow E[X^2] = \frac{2}{\mu^2} \]

\[ \therefore E[R] = \frac{\sum_{i=1}^{n} \lambda_i}{\mu^2} \quad \checkmark \quad (\text{expressions match}) \]
Define \( W(k) = \) average time in queue averaged over the first \( k \) priorities.

Obviously, \( W(1) = W_1 \)

Note consider two types of traffic: type a and type b

\[
\begin{align*}
N_{Qa} &= \lambda_a W_a, & N_{Qb} &= \lambda_b W_b \\
N_Q &= (\lambda_a+\lambda_b) W
\end{align*}
\]

\[
\begin{align*}
N_Q &= N_{Qa} + N_{Qb} \\
(\lambda_a+\lambda_b) W &= \lambda_a W_a + \lambda_b W_b \quad \text{(Recall Prob 39)}
\end{align*}
\]

We will use Little's theorem in a similar manner.

We have \( W(k) = \) average waiting time of an \( M/M/m \) with

\[
W_k = \frac{1}{\lambda_k} \left[ \left( \sum_{i=1}^{k-1} \lambda_i \right) W(k-1) - \left( \sum_{i=1}^{k-1} \lambda_i \right) W(k-1) \right]
\]

\( W_k \) is the average waiting time of an \( M/M/m \) with

arrival rate \( \lambda = \sum_{i=1}^{k} \lambda_i \) and mean service time \( \frac{1}{\mu} \)