07_Priority.pdf

Model (B): Unequal Service Rates

- Assume that the service rates of the two classes are not necessarily equal and μ_1 is the service rate for priority-1 customer & μ_2 for priority-2 customers. Define

$$\rho_1 = \frac{\lambda_1}{\mu_1}, \quad \rho_2 = \frac{\lambda_2}{\mu_2} \text{ and } \rho = \rho_1 + \rho_2.$$

- A similar analysis one be done for this model too. One gets finally

$$\boxed{L_q^{(1)} = \frac{\lambda_1(\frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2})}{(1-\rho_1)}, \quad L_q^{(2)} = \frac{\lambda_2(\frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2})}{(1-\rho_1)(1-\rho)}, \quad L_q = L_q^{(1)} + L_q^{(2)}.}$$

- Extra (again for interested, refer Miller (1981)): The probabilities for priority-1 customers are

$$p_{n_1} = (1-\rho)\left(\frac{\lambda_1}{\mu_1}\right)^{n_1} + \frac{\lambda_2}{\lambda_1 + \mu_2 - \mu_1}\left[\left(\frac{\lambda_1}{\mu_1}\right)^{n_1} - \frac{\mu_1 \lambda_1^{n_1}}{(\lambda_1 + \mu_2)^{(n_1+1)}}\right] \qquad (n_1 \geq 0).$$

Create account Login 202%

Model (C): Two-Class FCFS

- There are two customer classes with respective arrival rates λ_1 and λ_2 and with respective service rates μ_1 and μ_2.
- Service times are exponential and customers are served on an FCFS basis. There are no priorities.
- This two-class FCFS model can be viewed as single-class $M/H_2/1$ queue, where customers are grouped into a single arrival stream and the service distribution is a mixture of two exponential distribution.
- One can obtain (perhaps you can do later, after having seen the analysis of $M/G/1$)

$$\boxed{L_q^{(1)} = \frac{\lambda_1(\frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2})}{1-\rho}, \quad L_q^{(2)} = \frac{\lambda_2(\frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2})}{1-\rho}, \quad L_q = \frac{\lambda(\frac{\rho_1}{\mu_1} + \frac{\rho_2}{\mu_2})}{1-\rho}.}$$

Note: The L_q above is always greater than that of the standard $M/M/1$ model with mean service time equal to the weighted average of the respective means, namely, $\frac{1}{\mu} = \frac{(\frac{\lambda_1}{\lambda})}{\mu_1} + \frac{(\frac{\lambda_2}{\lambda})}{\mu_2}$ (due to the higher variability in the service times).

07_Priority.pdf

Create account Login

Comparison of Models (A), (B) and (C)

- (B) Vs. (C) : Priority queues (unequal service rates) with the nonpriority queue.
  - ▶ Imposition of priorities decreases the mean number of priority-1 customers $(L_q^{(1)})$ and increases the mean number of priority-2 customers $(L_q^{(2)})$. This result is quite intuitive.
- Comparison of average overall number in the queue L_q between the two models $(B\ \&\ C)$.
  - ▶ They (and W_q's) differ by a factor $\frac{\lambda-\lambda_1\rho}{\lambda-\lambda\rho_1}$ as evident from the below mentioned equations:

$$L_q\ of\ B = \left(\frac{\frac{\rho_1}{\mu_1}+\frac{\rho_2}{\mu_2}}{1-\rho}\right)\frac{\lambda-\lambda_1\rho}{1-\rho_1}$$

$$L_q\ of\ C = \left(\frac{\frac{\rho_1}{\mu_1}+\frac{\rho_2}{\mu_2}}{1-\rho}\right)\lambda.$$

▶ Thus there are fewer customers waiting in the priority queue of B when

$$\frac{\lambda-\lambda_1\rho}{\lambda-\lambda\rho_1} < 1 \iff \lambda_1\rho > \lambda\rho_1 \iff \lambda_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}\right) > (\lambda_1+\lambda_2)\frac{\lambda_1}{\mu_1} \iff \mu_2 < \mu_1.$$

202%

- This gives rise to an optimal design rule called "the shortest processing time (SPT) rule".
  ▶ The priority queue results in less overall waiting (compared with corresponding FCFS) when the first-priority customers have a faster service rate (or shorter service times).
  ▶ Conversely, the priority queue results in more overall waiting when the priority customers have longer service rates.
  ▶ If you want reduction in the total number waiting (or equivalently the overall mean delay), give priority to the class with the faster server rate.
- (B) Vs. (A) : Priority $two - rate$ model of B with the priority $one - rate$ model of A.
  ▶ We must make some choice for the value of the single rate μ. For example, choose μ so that

$$\frac{1}{\mu} = \left(\frac{\lambda_1}{\lambda}\right)\frac{1}{\mu_1} + \left(\frac{\lambda_2}{\lambda}\right)\frac{1}{\mu_2}.$$

  or one can choose μ to lie somewhere between μ_1 & μ_2.
  ▶ If $\mu = max\{\mu_1, \mu_2\}$ then $L_q^{(1)}, L_q^{(2)}, and\ L_q$ of A are less than that of B.
  ▶ If $\mu = min\{\mu_1, \mu_2\}$ then the reverse happens.
  ▶ If μ is strictly between μ_1 & μ_2 then the comparison would depend on the parameter values.

Extensions in Nonpreemptive Systems

- Though the idea of priority classes can be extended in principle, it is nearly impossible to determine the stationary probabilities (mainly because of its multi-dimensional nature) in case of more than two priorities.
- A direct expected-value procedure can be used to determine the mean-value measures L_q and W_q (not in our scheme of things).
- Continuous priority classes (based on actual service times – assumed to be known – leading to 'shortest job first' rule) and multi-server cases are some other extensions.

System with Preemptive Priorities

- Consider the same Markovian two-class model considered earlier, but with preemption now.
- Units of higher priority preempt units of lower priority in service.
- Lower priority units that are ejected from service cannot reenter service until the system is free of all higher priority units.
  - ▶ Ejected units must start over thereby losing the partial work already completed.
  - ▶ Ejected units resume service from the point of interruption.
- No difference in this model between preempt-resume & preempt-non-resume as service times are exponential (otherwise, one needs to worry about this).
- The state space for this preemptive priority two-class system is $S\{(m,n) : m,n \geq 0\}$ with their steady state system size probability given by

$$p_{mn} = P\{m \text{ units of priority-1 } \& \ n \text{ units of priority-2 in the system in steady-state}\}$$

♦ (λ_1, μ_1) and (λ_2, μ_2) are the corresponding arrival and service rates (of the two classes).

♦ $\lambda = \lambda_1 + \lambda_2, \qquad \rho = \dfrac{\lambda_1}{\mu_1} + \dfrac{\lambda_2}{\mu_2} < 1$ (assume)

- As earlier, one can proceed to write the balance equations (there will be $2^2 = 4$ sets of equations) as

$$\begin{aligned}
\lambda p_{00} &= \mu_1 p_{10} + \mu_2 p_{01} \\
(\lambda + \mu_1) p_{m0} &= \lambda_1 p_{m-1,0} + \mu_1 p_{m+1,0}, \quad m \geq 1 \\
(\lambda + \mu_2) p_{0n} &= \mu_1 p_{1,n} + \lambda_2 p_{0,n-1} + \mu_2 p_{0,n+1}, \quad n \geq 1 \\
(\lambda + \mu_1) p_{mn} &= \lambda_1 p_{m-1,n} + \lambda_2 p_{m,n-1} + \mu_1 p_{m+1,n}, \quad m, n \geq 1
\end{aligned}$$

- After deriving various partial generating functions, one can obtain the moments of the number of units in the system. This gives us

$$L^{(1)} = \frac{\rho_1}{1 - \rho_1}$$

$$L^{(2)} = \frac{\rho_2 - \rho_1 \rho_2 + \rho_1 \rho_2 (\frac{\mu_2}{\mu_1})}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}$$

Here, $L^{(i)}$ is the average number of class-i customers in the system in steady state.

- Class-1 customers are not affected by the presence of the class-2 customers. Thus the class-1 customers are effectively operating as if they were in an $M/M/1$.