Priority Queues

Service Engineering - Recitation 13

- Review: M/G/1 without Priorities (p.2)
- M/G/1 with Priorities (p.3-7)
- A Numerical Example (p.8)
- $c\mu$-Rule (p.9-11)

January 16, 2014

$M/G/1$ without Priorities

Recall the Khintchine-Pollaczek Formula:

$$E(W_q) = E(S) \cdot \frac{\rho}{1-\rho} \cdot \frac{1+C^2(S)}{2}$$

Expected residual service time (where ρ = Prob. of arriving to a busy server (PASTA+Little))?

$$E(R) = (1-\rho)\cdot 0 + \rho \cdot E(S) \cdot \frac{1+C^2(S)}{2}$$

$M/G/1$ with Priorities

- K customer classes, indexed by $k = 1, \ldots, K$.
- **Class k arrivals:** Poisson, rate λ_k.
- **Class k service times:** S_k - generally distributed, with $m_k = E(S_k)$ and $E(S_k^2)$ both finite.
- **Setting the priorities:** Set highest priorities to 1, then 2, ...; lowest to K.
- Assume FCFS within each priority class.
- Non preemptive first (Later, preemptive-resume).

$M/G/1$ with Priorities (2)

Steady state $\Leftrightarrow \rho \stackrel{\Delta}{=} \rho_1 + \cdots + \rho_K < 1$, where $\rho_k = \lambda_k m_k$.

Convenient notation: $\bar{\rho}_k = \rho_1 + \cdots + \rho_k,\ 1 \leq k \leq K$.

Note: $\rho_k =$ fraction of time allocated by server to class k.

$1 - \rho =$ idleness/availability.

- $E(W_q^k)$ - expected waiting time of class k customer.
- $E(L_q^k)$ - expected number of waiting class k customers.
- $E(U)$ - expected unfinished work in the system.
- $E(R)$ - expected residual service time.

Non-preemptive regime

1. $E(W_q^1) = E(R) + m_1 E(L_q^1) = E(R) + \rho_1 E(W_q^1)$

$\Rightarrow E(W_q^1) = E(R)/(1-\rho_1)$, as before $(K=1)$.

2. $E(W_q^2) = E(R) + \underbrace{m_1 E(L_q^1) + m_2 E(L_q^2)}_{\text{wait due to class 1 \& 2 in queue}} + \underbrace{m_1 \lambda_1 E(W_q^2)}_{\text{wait due to class 1, arriving during wait of 2.}}$

$\Rightarrow E(W_q^2) = E(R) + \rho_1 E(W_q^1) + \rho_2 E(W_q^2) + \rho_1 E(W_q^2)$

$\Rightarrow E(W_q^2) = [E(R) + \rho_1 E(W_q^1)]/(1-\rho_1-\rho_2) =$

$= E(R)/[(1-\rho_1)(1-\rho_1-\rho_2)]$

$\uparrow$

substitute $E(W_q^1)$

k. $EW_q^k = ER + m_1 \cdot EL_q^1 + \cdots + m_k \cdot EL_q^k + \lambda_1 m_1 EW_q^k + \cdots + \lambda_{k-1} m_{k-1} EW_q^k$

$\Rightarrow \quad = ER + \rho_1 EW_q^1 + \cdots + \rho_{k-1} EW_q^{k-1} + (\rho_1 + \cdots + \rho_k) EW_q^k$

$$E(W_q^k) = \frac{E(R) + \rho_1 E(W_q^1) + \cdots + \rho_{k-1} E(W_q^{k-1})}{(1-\rho_1-\rho_2-\cdots-\rho_k)} = \frac{E(R)}{(1-\bar{\rho}_{k-1})(1-\bar{\rho}_k)}$$

Non-preemptive regime (2)

where

$$E(R) = (1-\rho)\cdot 0 + \sum_k \rho_k \cdot m_k \cdot \frac{1+C_k^2(S)}{2} = \frac{1}{2}\sum_k \lambda_k E(S_k^2)$$

$$\Rightarrow E(W_q^k) = \frac{\frac{1}{2}\sum_{j=1}^{K} \lambda_j E(S_j^2)}{(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)} \ , 1 \leq k \leq K \ .$$

Preemptive-resume regime

Now, Class k does not *see* classes $k+1,\ldots,K$.

Recall: for M/G/1-like queues, $E(U) = \dfrac{E(R)}{1-\rho} = E(W_q)$

$$\begin{aligned} E(W_q^k) &= \frac{E(R^k)}{1-(\rho_1+\cdots+\rho_k)} + \sum_{j=1}^{k-1} \lambda_j m_j [E(W_q^k) + m_k] \\ &= \frac{E(R^k)}{1-\bar{\rho}_k} + \bar{\rho}_{k-1}[E(W_q^k) + m_k] \end{aligned}$$

$$E(W_q^k) = \frac{E(R^k)}{(1-\bar{\rho}_k)(1-\bar{\rho}_{k-1})} + \frac{\bar{\rho}_{k-1}}{1-\bar{\rho}_{k-1}}\ m_k$$

where $E(R^k) = \sum_{j=1}^{k} \rho_j \cdot m_j \cdot \frac{1+C^2(S_j)}{2} = \frac{1}{2}\sum_{j=1}^{k} \lambda_j E(S_j^2)$

$$E(W_q^k) = \frac{\frac{1}{2}\sum_1^k \lambda_j E(S_j^2)}{(1-\bar{\rho}_{k-1})(1-\bar{\rho}_k)} + \frac{\bar{\rho}_{k-1}}{1-\bar{\rho}_{k-1}}\ E(S_k)$$

A Numerical Example

Assume we have two classes $k = 1, 2$, **exponential** service with rates $\mu_1 = \mu_2 = 10$ customers/minute, $\lambda_1 = 4$, $\lambda_2 = 3$

- No priorities:

  $$E(W_q^1) = E(W_q^2) = E(W) = \frac{\rho}{\mu(1-\rho)} = 14 \text{ seconds}$$

- Non-Preemptive priorities:

  $$E(W_q^1) = \frac{\rho}{\mu(1-\rho_1)} = 7 \text{ seconds}$$
  $$E(W_q^2) = \frac{\rho}{\mu(1-\rho_1)(1-\rho_1-\rho_2)} = 23.32 \text{ seconds}$$

- Preemptive-resume priorities:

  $$E(W_q^1) = \frac{\rho_1}{\mu(1-\rho_1)} = 4 \text{ seconds}$$
  $$E(W_q^2) = \frac{\rho}{\mu(1-\rho_1)(1-\rho_1-\rho_2)} + \frac{\rho_1}{1-\rho_1}\frac{1}{\mu} = 23.32 + 4 = 27.32 \text{ seconds}$$

$c\mu$-Rule

Suppose that there is a cost C_k per unit time for each class-k customer, that waits in queue. Consider the "steady-state" cost

$$J = \sum_k C_k E(L_q^k).$$

Find a non-preemptive policy that minimizes J, i.e., assign the priorities to classes so that to minimize J.

- Simple Case 1: If $m_1 = m_2 = ... = m_k$, order by cost.
- Simple Case 2: If $c_1 = c_2 = ... = c_k$, shortest process time first.

$c\mu$-Rule (2)

Optimal priorities assignment: Highest priority to largest

$$\frac{C_k \lambda_k}{\rho_k} = \frac{C_k}{m_k} = C_k \mu_k$$

A Numerical Example:

- Assume we have two customer types $k = 1, 2$, exponential service with rates $\mu_1 = 10,\ \mu_2 = 5$ customers/minute, $\lambda_1 = 4$, $\lambda_2 = 3$, and $C_1 = 3,\ C_2 = 5$ dollar/minute.
- We have $C_1 \mu_1 = 10 \cdot 3 = 30$, and $C_2 \mu_2 = 5 \cdot 5 = 25$.
  $\rightarrow$ Give priority to type 1.

$c\mu$-Rule (3)

Add **abandonments**? $\frac{C_k' \mu_k}{\theta_k}$ - longest patience first

Idea:

- Before we minimized $J = \sum_k C_k E(L_q^k) = \sum_k C_k \lambda_k E[W^k]$.
- Now we minimize $J' = \sum_k [C_k^1 \lambda_k E[W^k] + C_k^2 \lambda_k P_k(Ab)]$.
- Using $P_k(Ab) = \theta_k E[W^k]$,

$$\begin{aligned} C_k^1 \lambda_k E[W^k] + C_k^2 \lambda_k P_k(Ab) &= C_k^1 \lambda_k E[W^k] + C_k^2 \lambda_k \theta_k E[W^k] \\ &= (C_k^1 + \theta_k C_k^2) E(L_q^k) = C_k' E(L_q^k) \end{aligned}$$