

Recitation 13: Priority Queues

M/G/1 with priorities

- K customer classes, indexed by $k = 1, \dots, K$.
- **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).

Steady state $\Leftrightarrow \rho \triangleq \rho_1 + \dots + \rho_K < 1$, where $\rho_k = \lambda_k m_k$.
Convenient notation: $\bar{\rho}_k = \rho_1 + \dots + \rho_k$, $1 \leq k \leq K$.

Note: ρ_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.

Calculation of $E(W_q^k)$. Non-preemptive regime

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

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

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

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

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

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

↑

substitute $E(W_q^1)$

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

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

$$E(W_q^k) = \frac{E(R) + \rho_1 E(W_q^1) + \dots + \rho_{k-1} E(W_q^{k-1})}{(1 - \rho_1 - \rho_2 - \dots - \rho_k)}, \quad k \geq 1$$

$$= (\text{Induction}) \frac{E(R) \cdot \left[1 + \frac{\rho_1}{1 - \rho_1} + \frac{\rho_2}{(1 - \rho_1)(1 - \rho_1 - \rho_2)} + \frac{\rho_{k-1}}{(1 - \bar{\rho}_{k-2})(1 - \bar{\rho}_{k-1})} \right]}{1 - \bar{\rho}_k}$$

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

The last equality can be derived via simple calculations.

$$\text{We now show} \quad E(R) = \frac{1}{2} \sum_{k=1}^K \lambda_k E(S_k^2)$$

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

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

Calculation of $E(W_q^k)$. Preemptive regime

Now, Class k does not “see” classes $k + 1, \dots, K$.

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

$$E(W_q^k) = \frac{E(R^k)}{1 - (\rho_1 + \dots + \rho_k)} + \sum_{j=1}^{k-1} \lambda_j m_j [E(W_q^k) + m_k]$$

\uparrow
 $j \leq k - 1$ preempts k

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

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

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

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

A Numerical Example

Non-Preemptive

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

When no priorities are applied we have that

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

When non-preemptive priorities are applied we have

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

Preemptive

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

$c\mu$ -Rule

CLASSICAL APPLICATION 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 .

Remark: The cost J is derived from the "actual" cost, that is $\sum_k \int_0^t C_k L_q^k(t) dt$.

Some intuition: Equal m 's \Rightarrow costliest first

Equal C 's \Rightarrow shortest processing time - first.

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 $C_1 = 3$, $C_2 = 5$ dolar/minute.

Calculating the $C\mu$ rule we have $C_1\mu_1 = 10 \cdot 3 = 30$, and $C_2\mu_2 = 5 \cdot 5 = 25$. Therefore we should give priority to customer type 1.

Adding abandonments...