Recitation 13: Priority Queues

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 $2, \ldots$; lowest to K.
- Assume FCFS within each priority class.
- Non preemptive first (Later, preemptive-resume).

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

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

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

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

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

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

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

   $\uparrow$

   substitute $E(W_q^1)$

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

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

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

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

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

The last equality can be derived via simple calculations.

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

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

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

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

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

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

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

$$\uparrow$$

$$j \leq k-1 \text{ preempts } k$$

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

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

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

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

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 $\dfrac{C_k\lambda_k}{\rho_k} = \dfrac{C_k}{m_k} = C_k\mu_k$.

Conservation Law for multi-class $M/G/1$

For any work-conserving, non-preemptive strategy,

$$\begin{aligned} \textstyle\sum_k \rho_k E(W_q^k) = & \ \tfrac{\rho}{1-\rho}E(R) & \rho < 1\ , \\ = & \ \infty & \rho \geq 1\ . \end{aligned}$$

Proof. Recall that the unfinished work is independent of strategy, therefore

$$E(U) = E(R) + \sum_k m_k E(L_q^k) = E(R) + \sum_k \rho_k E(W_q^k)$$

Set the policy when all customers are routed into a common queue and served by the single server on a First-Come-First-Serve basis, i.e., usual $M/G/1$.
Then it is known that $E(U) = E(W_q) = \dfrac{E(R)}{1-\rho}$ when $\rho < 1$.

$$\Rightarrow \qquad \boxed{\sum_{k=1}^K \rho_k E(W_q^k) = \frac{\rho}{1-\rho}\ \frac{1}{2}\ \sum_{k=1}^K \lambda_k E(S_k^2)}$$

When $\rho > 1$, at least one of the classes will have $EW_q^k \to \infty$.

Proof of $c\mu$-rule: Assume that the classes are labelled in a "usual" way: highest priorities to 1, then $2, \ldots$; lowest to K. By Little's formula, rewrite the cost as

$$J = \sum_k C_k E(L_q^k) = \sum_k C_k \lambda_k E(W_q^k) = \sum_k (C_k \mu_k) \rho_k E(W_q^k)$$

By the Conservation Law, the quantity $\sum_k \rho_k E(W_q^k)$ is constant. This will be a key to the proof. Recall

$$E(W_q^k) = \frac{E(R)}{(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)}. \tag{1}$$

Denote

$$w_k := E(W_q^k), \quad k = 1, 2..., K, \qquad \text{Then} \quad J = \sum_k (C_k \mu_k) \rho_k w_k.$$

Recall that by the priorities assignment we have

$$w_1 \leq w_2 \leq ...w_K. \tag{2}$$

Pick arbitrarily two adjacent classes i and $j = i+1$, and exchange the priorities among them. The resulting average waiting times will be denoted by $\widetilde{w}_k$ and the new cost $\widetilde{J} = \sum_k (C_k \mu_k) \rho_k \widetilde{w}_k$.

We will show that $\widetilde{J} \geq J$ and that will be enough for the proof. It is simple to derive from formula (1) that

$$\widetilde{w}_k = w_k, \quad \text{for} \quad k \neq i, j.$$

as well as

$$\widetilde{w}_i > w_i, \quad \widetilde{w}_j < w_j.$$

Therefore,

$$\widetilde{J} - J = C_i \mu_i (\rho_i \widetilde{w}_i - \rho_i w_i) + C_j \mu_j (\rho_j \widetilde{w}_j - \rho_j w_j) \tag{3}$$

From the Conservation Law we have $\sum_k \rho_k w_k = \sum_k \rho_k \widetilde{w}_k$, hence

$$\rho_i w_i + \rho_j w_j = \rho_i \widetilde{w}_i + \rho_j \widetilde{w}_j \quad \Rightarrow \quad \rho_j \widetilde{w}_j - \rho_j w_j = -(\rho_i \widetilde{w}_i - \rho_i w_i) \tag{4}$$

Combining (3) and (4), we have

$$\widetilde{J} - J = (\rho_i \widetilde{w}_i - \rho_i w_i)(C_i \mu_i - C_j \mu_j) \geq 0,$$

since $\widetilde{w}_i \geq w_i$ and $C_i \mu_i \geq C_j \mu_j$.

End of the proof.

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 type1.

M/M/N with priorities

- K customer classes, indexed by $k = 1, \ldots, K$.
- Highest priorities to 1, then $2, \ldots$; lowest to K.
- FCFS within priority class.
- Non Preemptive first (Later, preemptive-resume).

Class k: Poisson arrivals, at rate λ_k
Exponential service time: $m_k \equiv 1/\mu$ equal for all classes.
(Note this is a restriction, relative to the $M/G/1$ model analyzed previously.)

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

Non Preemptive (Kella & Yechiali 1985)

Let $E_{2,N}$ be the probability of delay in a single class $M/M/N$ system as given by the Erlang-C formula:.

$$E_{2,N} = \frac{(N\rho)^N}{N!(1-\rho)} \left[\sum_{k=0}^{N-1} \frac{(N\rho)^k}{k!} + \frac{(N\rho)^N}{N!(1-\rho)} \right]^{-1} . \tag{5}$$

Then the average waiting time of the k^{th} class is:

$$E\left(W_q^k\right) = \frac{1}{N\mu} \frac{E_{2,N}}{(1-\bar{\rho}_k)(1-\bar{\rho}_{k-1})} \tag{6}$$

where, as before, $\bar{\rho}_k = \sum_{i=1}^{k} \rho_i$, $\bar{\rho}_0 = 0$.

Proof

We will show that

1. $P\{W_q^k > 0\} \equiv E_{2,N}, \quad \forall k = 1, \ldots, K$;
2. $E\left(W_q^k | W_q^k > 0\right) = \frac{1}{N\mu} \frac{1}{(1-\bar{\rho}_k)(1-\bar{\rho}_{k-1})}$.

Thus,

$$E\left(W_q^k\right) = E\left(W_q^k | W_q^k > 0\right) P\{W_q^k > 0\} = \frac{1}{N\mu} \frac{E_{2,N}}{(1-\bar{\rho}_k)(1-\bar{\rho}_{k-1})} .$$

Step 1: A customer of class k is delayed if and only if upon its arrival all servers are busy. The total number of customers in system and the number of busy servers are independent of the policy as long as it is work conserving.

Step 2: Let us look at the system when all servers are busy. In that case we have a single server system with service rate $N\mu$.

As long as all servers are busy, queue of class k customers behaves like an $M/G/1$, where G represents the busy period of an $M/M/1$ queue with arrival rate equal to $\sum_{i=1}^{k-1} \lambda_i$ and service rate $N\mu$.

Denote by S_k this busy period. The first two moments of S_k are given by (Kleinrock I, p. 215):

$$\begin{aligned} E(S_k) &= \frac{1}{N\mu(1-\bar{\rho}_{k-1})} \\ E(S_k^2) &= \frac{2}{(N\mu)^2(1-\bar{\rho}_{k-1})^3} \end{aligned}$$

Hence, $\dfrac{1+C^2(S_k)}{2} = \dfrac{1}{1-\bar{\rho}_{k-1}}$.

Recalling that $\rho^{M/G/1} = \lambda_k E(S_k)$, applying Khinchine-Pollatcheck and performing straightforward calculations we get:

$$\boxed{E\left(W_q^k | W_q^k > 0\right) = E(S_k) \cdot \frac{1}{1-\rho^{M/G/1}} \quad \frac{1+C^2(S_k)}{2} = \frac{1}{N\mu} \frac{1}{(1-\bar{\rho}_k)(1-\bar{\rho}_{k-1})}}$$

Classical Application Suppose cost C_k for one unit wait of class k and we wish to minimize $\sum_k C_k \lambda_k E(W_k)$

Optimal (Federgruen & Groenvelt 1988) Highest priority to largest C_k.

Alternative Proof for Non-Preemptive Case can be provided via the the same algorithm as for M/G/1 with priorities.

Waiting time (given wait) of class k customer can be divided into three components:

- Residual service time, which is $\exp(n\mu)$ distributed.
- Wait due to service of classes $1-k$ that were in queue on arrival of a customer.
- Wait due to service of customers from classes $1-(k-1)$ that arrived during customer's wait.

Then

1. $$E(W_q^1|W_q^1>0)=\frac{1}{n\mu}+\frac{1}{n\mu}E(L_q^1|W_q^1>0)=\frac{1}{n\mu}+\rho_1 E(W_q^1|W_q^1>0)$$

$$\Rightarrow E(W_q^1|W_q^1>0)=\frac{1}{n\mu}\cdot\frac{1}{1-\rho_1}.$$

2. $$E(W_q^2|W_q^2>0)=\frac{1}{n\mu}+\underbrace{\frac{1}{n\mu}\cdot\Big(E(L_q^1|W_q^2>0)+E(L_q^2|W_q^2>0)\Big)}_{\text{wait due to class 1 \& 2 in queue}}+\underbrace{\frac{\lambda_1}{n\mu}\cdot E(W_q^2|W_q^2>0)}_{\substack{\text{wait due to class 1,}\\ \text{arriving during wait of 2}}}$$

$$\Rightarrow E(W_q^2|W_q^2>0)=\frac{1}{n\mu}+\rho_1 E(W_q^1|W_q^1>0)+\rho_2 E(W_q^2|W_q^2>0)+\rho_1 E(W_q^2|W_q^2>0)$$

$$\Rightarrow E(W_q^2|W_q^2>0)=\left[\frac{1}{n\mu}+\rho_1 E(W_q^1|W_q^1>0)\right]/(1-\rho_1-\rho_2)=$$

$$=\underset{\substack{\uparrow\\ \text{substitute } E(W_q^1|W_q^1>0)}}{\frac{1}{n\mu}}/[(1-\rho_1)(1-\rho_1-\rho_2)]$$

Then the formula:

$$E\left(W_q^k|W_q^k>0\right)=\frac{1}{N\mu}\frac{1}{(1-\bar{\rho}_k)(1-\bar{\rho}_{k-1})}$$

can be derived by induction similarly to M/G/1 case.

Preemptive Resume

In the case of preemptive resume we have the following recursive relation:

$$E\left(W_q^k\right) = \left[\Lambda_k \bar{W_q}^{(1\to k)} - \sum_{i=1}^{k-1} \lambda_i E(W_q^i)\right] / \lambda_k \ , \ k = 1, 2, \ldots, K,$$

where $\Lambda_k = \sum_{i=1}^{k} \lambda_i$ and $\bar{W_q}^{(1\to k)}$ is the average waiting time in a single class $M/M/N$ FCFS system with arrival rate Λ_k (i.e. ignoring arrivals from the lower classes $i > k$).

Proof

Let L_q^i be the average number class i customers in the queue.

Let $L_q^{(1\to k)}$ be the average number of customers in a single class $M/M/N$ FIFO queue with arrival rate Λ_k (i.e. ignoring arrivals from lower classes $i > k$).

Note that the total number of customers in queue is independent of the policy chosen and hence

$$L_q^{(1\to k)} = \sum_{i=1}^{k} L_q^i$$

Calculation of $E(W_q^k)$ = average wait of class k.

1. $E(W_q^1) = \bar{W_q}^{(1\to 1)}$
2. $L_q^{(1\to 2)} = L_q^1 + L_q^2$

   $\Rightarrow (\lambda_1 + \lambda_2)\bar{W_q}^{(1\to 2)} = \lambda_2 E(W_q^2) + \lambda_1 E(W_q^1)$

   $\Rightarrow E(W_q^2) = \left[\Lambda_2 \bar{W_q}^{(1\to 2)} - \lambda_1 E(W_q^1)\right] / \lambda_2$

   .
   .
   .

k. $\boxed{E\left(W_q^k\right) = \left[\Lambda_k \bar{W_q}^{(1\to k)} - \sum_{i=1}^{k-1} \lambda_i E(W_q^i)\right] / \lambda_k}$

A Numerical Example

Non Preemptive

Assume we have a system with 10 servers, $\mu = 1$ (average handling time of one minute) and two customers classes such that $\lambda_1 = 4$ and $\lambda_2 = 3$.

We calculate $E_{2,N} = 0.222$ using 4CC to obtain:

$$E\left(W_q^1\right) = \frac{1}{10}\frac{0.22}{(1-0.4)} = 2.2 \text{ seconds}$$
$$E\left(W_q^2\right) = \frac{1}{10}\frac{0.22}{(1-0.4)(1-0.7)} = 7.3 \text{ seconds}$$

Preemptive

Using 4CC we calculate:

$W_q^{(1\to 2)} = 4.4$ seconds

$E(W_q^1) = 0.09$ seconds

$E(W_q^2) = (0.5133 - 0.006)/3 = 0.1691$ minutes $= 10.146$ seconds.