 Minimal cost service rate in priority queuing models for emergency cases in hospitals

Nse. S. Udoh 1 *, Idorenyin A. Etukudo 2

1 Department of Mathematics and Statistics, University of Uyo, Nigeria
2 Department of Mathematics and Statistics, Akwa Ibom State University, Nigeria
*Corresponding author E-mail: nsesudoh@uniuyo.edu.ng

Abstract

Performance measures and waiting time cost for higher priority patients with severe cases over lower priority patients with stable cases using preemptive priority queuing model were obtained. Also, a total expected waiting time cost per unit time for service and the expected service cost per unit time for priority queuing models: M/M/2: ∞/NPP and M/M/2: ∞/PP were respectively formulated and optimized to obtain optimum cost service rate that minimizes the total cost. The results were applied to obtain optimum service rate that minimizes the total cost of providing and waiting for service at the emergency consulting unit of hospital.

Keywords: Priority Queuing Models; Preemptive Model; No Preemptive Model; Waiting Time Cost; Optimum Cost Service Rate; Emergency Case.

1. Introduction

Two priority queuing models; the preemptive priorities (PP) and the Nonpreemptive priorities (NPP) have been considered. In PP, a lower-priority patient with stable case being served can be ejected back into the queue or interrupted whenever a higher-priority patient with serious/critical case, where prompt treatment is vital enters the queuing system. In NPP, a patient being served cannot be ejected back into the queue even if a higher-priority customer enters the queuing system despite the severity of the case.

In order to solve this queuing problem, service facility must be organized so that an optimum balance is obtained between the cost of waiting time and the cost of idle equipment, Gupta and Hira (2009). The cost of waiting-patients generally includes either direct cost of idle doctors and other medical personnel or indirect cost of loss of goodwill due to dissatisfactory service. However, the cost of idle service facilities is the payment to be made for the period.

Hagen et al (2013) examined several queuing models for intensive care units and the effects on waiting times, utilization, return rates, mortalities and number of patients served. Then, Li, et al (2013) designed a hybridized-queuing model for effective packets scheduling in inter-vehicular communication with a view to obtaining the difference between delay-sensitivity and packet length of services. A study of five schemes proved that the non-preemptive short-packet-first scheme results in the minimal overall delay. The model was superior in terms of performance indices in packet delivery ratio and throughput. Also, Ke, Li and Ni (2012) used the priority queuing model to study the performance of various strategies based on delay sensitivity and packets length. The non-preemptive short-packets-first strategy was proved to result in the minimal overall delay with different strategies for delay-sensitive and non-delay sensitive services. Consequently, an optimal priority-queuing model for the scheduling of multiple internet services was designed based on the above conclusions. Siddarthan, Jones and Johnson (2006) investigated the increased waiting time costs imposed on society due to inappropriate use of the emergency department by patients seeking non-preemptive or primary care. It proposed a simple economic model to illustrate the effect of misuse at a public or not-for-profit hospital. The result showed that non-emergency patients contribute to lengthy delays in the emergency department for all classes of patients. It therefore proposed a priority queuing model to reduce average waiting times.

Several other works on the analysis of emergency waiting time and queuing systems with priority service discipline abound in the literature; see, for example, Udoh (2010), Bedford and Zephongsekkl (2003), Blake and Carter (1996), among others. The objective of this work is to formulate and optimize the total cost of waiting for service and service cost per unit time to determine the particular level of service which minimizes the total cost of providing service and waiting for that service. We also seek to obtain the associated waiting time costs for the higher priority patients and the lower priority patients.

2. Problem formulation

The waiting time and its associated losses can be decreased by increasing the investment in service facilities such as Doctors and other medical facilities. It is desirable to obtain the minimum sum of these two cost: cost of investment and operation, and cost due to waiting by patients for service. The optimum balance of costs can be obtained by scheduling the flow of patients requiring service and/or providing proper number of doctors and facilities. Both service facilities and the flow of the patients can be controlled by effective schedule of consulting time as well as provide necessary number of doctors and facilities to minimize the overall cost. According to Taha (2007), the estimated waiting time and idle time cost can be obtained as follows:

Let $C_w = \text{expected waiting cost/patient/unit time}$

$L_s = \text{expected number of patients in the system}$
C_t = cost of treating one patient/unit time (that is, cost per unit time
Of having a doctor available
\( \mu = \) service rate or average number of patients treated/unit time
\( \lambda = \) average number of patients’ arriving/unit time in the queue

The expected waiting cost per unit time;

\[
W_c = C_s L_s = \frac{\lambda}{\mu - \lambda}
\]  
(1)

The expected service cost per unit time;

\[
W_s = C_s \mu \Sigma c
\]  
(2)

In addition, total cost

\[
c = c_s \frac{\lambda}{\mu - \lambda} + C_s \mu
\]  
(3)

The minimization condition is \( \frac{\partial c}{\partial \mu} = 0 \)

\[
\hat{\mu} = \lambda \pm \sqrt{\frac{2C_s}{\mu - \lambda}}
\]  
(4)

where \( \hat{\mu} \) is the minimum cost service rate.

3. Performance measures and estimated waiting time cost for M/M/2: \( \infty/FCFS \) (in the highest priority class)

Notations

Let \( P_n = \) probability of exactly \( n \) customers in the queuing system

\( L = \) expected number of customers in queuing system

\( L_q = \) expected queue length (excluding customers being served)

\( W = \) waiting time in the system (includes service time) for each individual customer; \( W = E(w) \)

\( W_q = \) waiting time in queue (excludes service time) for each individual customer; \( W_q = E(w) \)

\( N_t = \) number of servers in the queuing system at time \( t \) \((t \geq 0)\)

\( P_n(\lambda) = \) probability of exactly \( n \) customers in queuing system at time \( t \)

\( S (\geq 2) = \) number of servers (parallel service channels) in the queuing system

\( \lambda_s = \) mean arrival rate (expected number of arrival per unit time of new customers completing service per unit time when \( n \) customers are in system)

\( \rho = \frac{\lambda}{\mu} = \) traffic intensity (utilization factor for the service facility)

\( C_s = \) expected waiting cost in queue

\( C_i = \) expected waiting cost in the system

Assume \( \hat{\lambda} = \lambda s, \) then in a steady-state queuing process \( L_s = \lambda W_s \)

From little’s law, Hillier and Lieberman (1995);

\[
L_q = \lambda W_q \, \text{and} \, W_q = W - \frac{1}{\mu}
\]  
(5)

\[
\therefore \, W_i = E(w) = \frac{\lambda}{\mu(\mu - \lambda)}
\]  
(6)

When \( s > 1 \)

The utilization factor is

\[
U_s = \left[ \frac{1}{\left( \frac{\lambda}{\mu} \right)^n} \right]^{1/n} \left( \frac{1}{\left( \frac{\lambda}{\mu} \right)^n} \right)^{1/n} \left( \frac{1}{\left( \frac{\lambda}{\mu} \right)^n} \right)^{1/n}
\]  
(7)

If \( N \rightarrow \infty \) and \( \lambda < \frac{\lambda}{\mu} \), so that \( \rho = \frac{\lambda}{\mu} < 1 \), then \( P_0 = \)

\[
1 \cdot \sum_{n=1}^{\infty} \left[ \left( \frac{\lambda}{\mu} \right)^n \right]^{1/n} \left( \frac{1}{\left( \frac{\lambda}{\mu} \right)^n} \right)^{1/n} \left( \frac{1}{\left( \frac{\lambda}{\mu} \right)^n} \right)^{1/n}
\]  
(8)

and \( P_n = \)

\[
\begin{pmatrix}
\left( \frac{\lambda}{\mu} \right)^n & P_n \quad \text{if} \quad 0 < n < s \\
\left( \frac{\lambda}{\mu} \right)^n & P_n \quad \text{if} \quad n \geq s
\end{pmatrix}
\]  
(9)

Furthermore, \( L_q = \sum_{n=1}^{\infty} n P_n \)

\[
= \sum_{j=0}^{\infty} \sum_{j=0}^{\infty} \left( \frac{\lambda}{\mu} \right)^n P_n = P_0 \left( \frac{\lambda}{\mu} \right)^n \frac{\rho}{\mu} \left( \sum_{j=0}^{\infty} \rho^j \right)
\]  
(10)

\[
= P_0 \left( \frac{\lambda}{\mu} \right)^n \frac{\rho}{(1-\rho)^2} = P_0 \left( \frac{\lambda}{\mu} \right)^n \frac{\rho}{s(1-\rho)}
\]  
(11)

\[
W_q = \frac{L_q}{S}
\]  
(12)
\( C_q = W_q C_w \) (10)

\( W_s = W_q + 1/\lambda = \frac{L_q}{\lambda} + 1/\mu \) (11)

\( C_s = W_s C_w \) (12)

\( L_s = \lambda (W_q + 1/\mu) = L_q + \frac{\lambda}{\mu} \) (13)

\[ P(\omega > t) = e^{-\mu t} \left[ 1 + \frac{p(\omega)}{s(1-\rho)} \left( 1 - e^{-(s-1)\mu t} \right) \right] \] and \[ P(\omega_q > t) = \left( 1 - P(\omega_q = 0) \right) e^{-\mu t(1-\rho)t}, \] where \( P(\omega_q = 0) = \sum_{n=0}^{\infty} P_n \)

3.1. Waiting time measure and waiting time cost for M/M/2: \( \infty/NPP \)

Let \( S (\geq 2) \) = number of servers, \( \mu \) = mean service rate per busy server, \( k \) = the number of priority classes;

\[ \lambda = \sum_{i=1}^{k} \lambda_i \] (14)

And

\[ r = \frac{\lambda}{\mu} \]

This result assumes that \( \sum \lambda_i < \mu \); see Hillier and Lieberman (1995).

Let \( w_{sk} \) be the steady state expected waiting time in the system (including service time) for a member of priority class \( k \).

Then, \( w_s = \frac{1}{AB_i B_k} \cdot \frac{1}{\mu} \), \( k = 1, 2, \ldots, N \) (15)

where

\[ A = s\frac{\mu - \lambda}{\rho} \sum_{j=1}^{k} \frac{1}{\mu j} + s\mu, j = 0, 1, 2, \ldots, s-1 \] (16)

\[ B_0 = 1 \]

\[ B_i = \frac{1}{s\mu} \sum_{j=0}^{i} \lambda_j \] for \( k = 1, 2, \ldots, N \) (17)

\[ L_s = \lambda w_s = \lambda \left( \frac{1}{AB_i B_k} + \frac{1}{\mu} \right) \]

And the estimated waiting time cost for this class is given by

\( W_s = W_q C_w \) (18)

Where

\( W_q = \frac{1}{AB_i B_k} + \frac{1}{\mu} \) and \( L_q = \lambda w_q = \lambda \frac{1}{AB_i B_k} \)

3.1.1. Minimum cost service rate (\( \mu^* \)) for the NPP queuing model

The waiting time cost

\[ W_c = C w = C w \]

\[ \lambda w_s + \frac{1}{\mu} \]

And the expected service cost \( S_i \) per unit time is; \( S_i = C \mu \).

The total cost;

\[ C_{np} = W_s + S = C \left[ \lambda \left( \frac{1}{AB_i B_k} + \frac{1}{\mu} \right) + C \mu \right] \] (19)

Let \( C_{s} = A, s = A; s^2 = B, s^2 = D; s^2 = E, s^2 = F \)

\[ C_{s} = A \left[ \frac{1}{BD} \frac{\mu - \lambda}{\rho^2} + \frac{1}{\mu} \right] + C \mu \]

\[ C_{np} = A \left[ \frac{1}{BD\mu} + \frac{1}{\mu} \right] + C \mu \]

Let also \( H = \left( F + E + B D \lambda \right), G = \left( \frac{BD}{\rho^2} \right) \) and \( P = \left( EF \right) \)

\[ C_{s} = \left[ \frac{1}{G \mu + s \mu + \frac{1}{\mu} \rho H} \right] + C \mu \]

\[ C_{np} = \left[ \frac{1}{G \mu + s \mu + \frac{1}{\mu} \rho H} \right] + C \mu \]

Theorem 1: Suppose \( \tau(\mu) \) is the total expected waiting cost per unit time in a queue, we assumed that the expected service cost per unit time in a queue is continuous in the interval \( x_1 \leq \mu \leq x_2 \) and is differentiable with relative minimum cost at \( \mu^* \). Then \( x_1 \leq \mu^* \leq x_2 \) if \( \tau(\mu^*) = 0 \).

Proof: consider any interval \( x_1 \) and \( x_2 \) in the queue and k an increment, we have

\[ \tau(\mu^*) = \lim_{k \to 0} \left( \frac{\tau(\mu^* + k) - \tau(\mu^*)}{k} \right) \]

If (21) exists as a definite number, \( \tau(\mu^*) \leq \tau(\mu^* + k) \) for all k sufficiently close to zero.

Then,

\[ \tau(\mu^* + k) - \tau(\mu^*) \geq 0 \] if \( k > 0 \) and \( \tau(\mu^* + k) - \tau(\mu^*) \leq 0 \) if \( k < 0 \)

Thus \( \tau(\mu^*) \) in (21) gives the limit as \( k \to 0 \) through positive values as \( \tau(\mu^*) \geq 0 \) and through negative values as \( \tau(\mu^*) \leq 0 \)

Therefore, to satisfy both inequalities in (23), we have \( \tau(\mu^*) = 0 \)

From theorem 1:
\[
\frac{\partial C_{\mu_2}}{\partial \mu} = \frac{P_{A_2} - Q_{A_2} \mu \lambda^2}{Q_{\lambda}(\mu_2^2 + \lambda^2 + \mu^2)} + C_{A_2} = 0
\]

\[
(P_{A_2} - Q_{A_2} \lambda \mu^2 - \lambda^2(Q_{\lambda}^2 \mu^2 + \lambda^2 + \mu^2)) + C_{A_2}((\lambda^2 \mu^2 + \lambda^2 + \mu^2)) \mu^2 = 0
\]

\[
(P_{A_2} - Q_{A_2} \lambda \mu^2 - \lambda^2(Q_{\lambda}^2 \mu^2 + \lambda^2 + \mu^2)) + C_{A_2}((\lambda^2 \mu^2 + \lambda^2 + \mu^2)) \mu^2 = 0
\]

The resulting polynomial in (25) is a function of \(\mu\) which can be solved for optimum service rate \(\mu^*\) that minimizes the total cost of providing service and waiting for that service in an NPP queuing model by substituting the values of the constants C, A, Q, P, s, and C for a given queuing process.

3.2. Waiting time measure and waiting time cost for preemptive priority model: MM/2: \(\infty/PP\)

If preemption is allowed under the conditions in NPP, the total expected waiting time in the system (including the total service time) will change to:

\[
w_{sk} = w_s = \frac{1}{\mu_{sk}}
\]

\[
L_s = L_s w_{sk}
\]

When \(s > 1\), \(w_s\) can be calculated by an iterative procedure.

To determine the expected waiting time in the queue (excluding service time) for priority class \(k\), we have:

\[
w_{sk} = w_{sk} - \frac{1}{\mu_{sk}}
\]

Also, if \(s > 1\), \(w_s\) can be calculated by an iterative procedure. The associated expected waiting time cost is; \(w_s C_{w_s}\).

3.2.1. Minimum cost service rate (\(\mu^*\)) for PP queuing model

The waiting time cost is:

\[
w_s = C_s L_s = C_s \left( \frac{\mu_{sk}}{\mu_{sk} - 1} (1 + \frac{1}{\mu}) \right)
\]

where \(w_{sk} = \frac{1}{\mu_{sk}}\).

Note that preemption does not affect the service process in any way because of lack of memory property of exponential distribution.

Hence, the expected total waiting time for any customer is still \(1/\mu\), and the expected service cost \(S_s\) per unit time is; \(S_s = C_s \mu\).

The total cost

\[
\frac{\partial C_{\mu_2}}{\partial \mu} = C_s \left[ \frac{\lambda}{\mu^2 - \mu_2^2} \right] + C_{A_2} = 0
\]

\[
C_{A_2} = C_{A_2} \left[ \frac{\lambda}{\mu_{sk} - \mu} \right] + C_{A_2} \left[ \frac{\lambda}{\mu^2 - \mu_2} \right] + \lambda \mu \frac{\lambda^2}{\mu^2 - \mu_2} C_{A_2} = 0
\]

Similarly, from theorem 1:

\[
C_{\mu} = C_s \left[ \frac{\lambda}{\mu^2 - \mu_2^2} \right] + C_{A_2} \left[ \frac{\lambda}{\mu_{sk} - \mu} \right] + C_{A_2} \left[ \frac{\lambda}{\mu^2 - \mu_2} \right] + \lambda \mu \frac{\lambda^2}{\mu^2 - \mu_2} C_{A_2} = 0
\]

\[
C_{\mu^*} = C_s \left[ \frac{(\mu^2 - \mu_2^2)}{\mu^2 - \mu_2} \right] + C_{A_2} \left[ \frac{\lambda}{\mu_{sk} - \mu} \right] + C_{A_2} \left[ \frac{\lambda}{\mu^2 - \mu_2} \right] = 0
\]

Also, the resulting polynomial equation in (31) can be solved for real values of \(\mu^*\) as the optimum service rate which minimizes the total cost of providing service and waiting for service in PP queuing model by substituting the constants \(C_s\) and \(\lambda\).

4. Application

We consider queuing data on emergency consulting unit of University of Uyo Teaching Hospital in Nigeria. The following parameters were obtained there: \(\lambda = 2.76; s = 2, C_1 = 17.36; r = 1.38\) and \(C_s = 4.37\). The associated cost component was obtained as the mean of the salaries of patients/patients’ benefactors. By substituting these parameters in the results in (25) and (31) we obtained the respective service rate equations of the non-preemptive and preemptive priority queuing models as follows;

\[
Y_p = 3.4025 \lambda^2 - 4.3525 \lambda^2 - 6.1554 \lambda^2 + 2.7843 \lambda^2 - 1.1525 \lambda^2 + 4.459 \lambda - 39.95 + 0
\]

\[
P_p = 17.36 \lambda_{p2} - 1.376 \lambda_{p2} - 0.261 \lambda_{p2} - 0.437 \lambda = 0
\]

The solution of (32) yields the following rational values as roots of \(\mu^*\) for the Nonpreemptive priority queuing model; 1. 4447, -0.8607, 0.7011 and 0.1029. The optimum cost service rate which yield the minimum service rate function that minimizes the waiting time cost of patients and the service time cost of doctors is \(\mu^* = 0.7011\).

Similarly, the solution of (33) yields a rational value of \(\mu^* = -0.2786\) as the optimum cost service rate that yields a minimum service rate function for the Preemptive priority queuing model that minimizes the waiting time cost of patients and the service time cost of doctors.

5. Conclusion

The total expected waiting time cost per unit time and the expected service time cost per unit time for priority queuing models: MM/2; \(\infty/\infty\) and MM/2: \(\infty/PP\) were formulated and optimized to obtain optimum service rate that both minimizes the expected waiting time cost and the expected service time cost per unit time for the models. The resulting polynomial equations were solved to obtain real values of \(\mu^*\) as the optimum cost service rate (\(\mu^*\)) that minimizes the total cost of time of providing service by doctors and waiting for service by patients at the emergency unit in hospital. The optimal value of \(\mu^* = 0.7011\) provide the optimum service time that guarantees minimum waiting time cost and service time cost for the non-preemptive priority model while \(\mu^* = 0.2786\) provides an optimum service time that guarantees a minimum waiting time cost and service time cost using the Preemptive priority model. This, however suggest an overstretch of available medical facilities.

References


