Transactions of the Nigerian Association of Mathematical Physics
***Vol. 6* (Jan., 2018), pp. 63– 70**

In honour of Prof. Ekhaguere at 70

On the analysis of priority scheduling and its applications in the $M/G/1$ queueing system

S. O. Agboola*

Department of Applied Mathematics with Statistics , Bells University of Technology, Ota, Nigeria

Abstract. This paper presents the analysis and application of preemptive resume and non-preemptive priorities scheduling for an $M/G/1$ queueing system. Our analysis is based on the fact that, in real life problems such as telecommunication system, supermarket operation, production system, e.t.c. It is usual to assign priorities according to the perceived importance of the customer. We consider the general case of $j \geq 2$ different classes of customers in which the different classes can have different service requirement. for both preemptive resume priority and non-preemptive priority scheduling, we develop efficient algorithms to calculate the time it takes to serve all customers of equal or higher priority that are present at the moment a class j customer arrives, we as well obtained the following measures for the two priorities scheduling, mean response or sojourn time of a class j customers, the number waitng in the queue, average number of class j customers present in the system and total waiting time for a class j customer.

Keywords: preemptive resume priority scheduling, non-preemptive priority scheduling, Little's law, PASTA property.

1. Introduction

In a queueing system in which customers are distinguished by class, it is usual to assign priorities according to the perceived importance of the customer. The most important class of customers is assigned priority 1; classes of customers of lesser importance are assigned priorities $2, 3, \cdots$. When the system contains customers of different classes, those with priority j are served before those of priority $j+1, \quad j = 1, 2, \cdots$. Customers within each class are served in first-come, first-served order. The question remains as to how a customer in service should be treated when a higher-priority customer arrives. This gives rise to two scheduling policies. The first policy is called preemptive priority and in this case, a lower-priority customer in service is ejected from service the moment a higher-priority customer arrives. The interrupted customer is allowed back into service once the queue contains no customer having a higher priority. The interruption of service may mean that all of the progress made toward satisfying the ejected customers service requirement has been lost, so that it becomes necessary to start this service from the beginning once again. This is called preempt-restart. Happily, in many cases, the work completed on the ejected customer up to the point of interruption is not lost so that when that customer is again taken into service, the service process can continue from where it left off. This is called preempt-resume.

The second policy is nonpreemptive. In this scheduling, the service of a low-priority customer will begin when there are no higher-priority customers present. Now, however, once service has been initiated on a low-priority customer, the server is obligated to serve this customer to completion, even if one or more higher-priority customers arrive during this service.

Nomenclature

λ: Arrival rate
j : Class of customer in service
$E[y]$: Average length of a busy period

*Corresponding author. Email: larrysoa_7519@yahoo.com

μ: Service rate
Y: The random variable that describes the length of a busy period
N_i: Random variable of the number of customers served during the i^th, $\quad i = 1, 2, \cdots, n$ busy period
X : Service time random variable when arrival occurs
x : duration of service time random variable X
R : Response time
$\Re$: Residual service time
$f_X(x)$: Density function of random variable X
$b(x)dx$: Frequency of occurrences of service interval having length x
$M(x)$: Number of customers served by time t
W_q: Expected time arriving customer must wait until its service begins. L_q : Expected number of customers waiting in the queue
$E[\Re]$: Mean residual service time
ρ : Workload intensity
W_j^q : the total time spent waiting by a class j customer
λ_1 : Arrival rate of class-1 customers
λ_2 : Arrival rate of class-2 customers

2. Materials and methods

2.1 *M/M/1: Priority queue with two customers classes*

2.1.1 M/M/1: Preemptive priority policy

To analyse the M/M/1 queue with two customer classes that operate under the preemptive priority policy. Let Customers of class 1 arrive according to a Poisson process with rate λ_1; those of class 2 arrive according to a second (independent) Poisson process having rate λ_2. Customers of both classes receive the same exponentially distributed service at rate μ. We shall let $\rho_1 = \frac{\lambda_1}{\mu}$, $\rho_2 = \frac{\lambda_2}{\mu}$, and assume that $\rho = \rho_1 + \rho_2 < 1$. Given that the service is exponentially distributed, it matters little whether we designate preempt-resume or preempt-restartthanks to the memoryless property of the exponential distribution. Furthermore, since we have conveniently chosen the service time of all customers to be the same, the total number of customers present, N, is independent of the scheduling policy chosen. It then follows from standard $M/M/1$ results that $E[N] = \frac{\rho}{1-\rho}$. From the perspective of a class 1 customer, class 2 customers do not exist, since service to customers of class 2 is immediately interrupted upon the arrival of a class 1 customer. To a class 1 customer, the system behaves exactly like an $M/M/1$ queue with arrival rate λ_1 and service rate μ. The mean number of class 1 customers present and the mean response time for such a customer are given, respectively, by $E[N_1] = \frac{\rho_1}{1-\rho_1}$ and $E[R_1] = \frac{1}{\mu(1-\rho_1)}$. Mean number of class 2 customers present is

$$E[N_2] = E[N] - E[N_1] = \frac{\rho}{1-\rho} - \frac{\rho_1}{1-\rho_1} = \frac{\rho_1+\rho_2}{1-\rho_1-\rho_2} - \frac{\rho_1}{1-\rho_1} = \frac{\rho_2}{(1-\rho_1)(1-\rho_1-\rho_2)}$$

By Little's law, the mean response time of class 2 customers is given as

$$E[R_2] = \frac{1/\mu}{(1-\rho_1)(1-\rho_1-\rho_2)}.$$

2.1.2 M/M/1: Nonpreemptive scheduling policy

For the nonpreemptive scheduling policy. This time an arriving class 1 customer finding a class 2 customer in service is forced to wait until that class 2 customer finishes its service. From PASTA, we know that an arriving class 1 customer will find, on average, E[N_1] class 1 customers already

present, each of which requires $\frac{1}{\mu}$ time units to complete its service. The arriving customer also has a mean service time of $\frac{1}{\mu}$, and if the arriving customer finds a class 2 customer in service, a further $\frac{1}{\mu}$ time units must be added into the total time the arriving class 1 customer spends in the system. The probability of an arriving customer finding a class 2 customer in service is equal to ρ_2, recall that the probability the system has at least one customer is $\rho = \rho_1 + \rho_2$. Summing these three time periods together, we compute the mean response time for a class 1 customer as

$$E[R_1] = \frac{E[N_1]}{\mu} + \frac{1}{\mu} + \frac{\rho_2}{\mu} \tag{1}$$

By Little's law, we have $E[N_1] = \lambda_1 E[R_1]$, which when substituted into Equation (1) gives

$$E[R_1] = \frac{\lambda_1 E[R_1]}{\mu} + \frac{1}{\mu} + \frac{\rho_2}{\mu}$$

Solving for $E[R_1]$ yields $E[R_1] = \dfrac{(1+\rho_2)/\mu}{1-\rho_1}$. Mean number of class 1 customers is obtained as $E[N_1] = \dfrac{(1+\rho_2)\rho_1}{1-\rho_1}$. Mean number of class 2 customers can be found from

$$E[N_2] = E[N] - E[N_1] = \frac{\rho_1+\rho_2}{1-\rho_1-\rho_2} - \frac{(1+\rho_2)\rho_1}{1-\rho_1} = \frac{\rho_2 - \rho_1\rho_2 + \rho_1^2\rho_2 + \rho_1{\rho_2}^2}{(1-\rho_1)(1-\rho_1-\rho_2)}$$

$$= \frac{\rho_2[1-\rho_1(1-\rho_1-\rho_2)]}{(1-\rho_1)(1-\rho_1-\rho_2)}$$

Finally, from Little's law, the mean response time for class 2 customers is

$$E[R_2] = \frac{E[N_2]}{\lambda_2} = \frac{[1-\rho_1(1-\rho_1-\rho_2)]/\mu}{(1-\rho_1)(1-\rho_1-\rho_2)}$$

2.2 *M/G/1: Priority queue with two customers classes*

2.2.1 M/G/1: NonPreemptive priority scheduling

Let us consider the general case of $J \geq 2$ different classes of customer in which the different classes can have different service requirements. We assume that the arrival process of class j customers, $j = 1, 2, \cdots, J$, is Poisson with parameter λ_j and that the service time distribution of this class of customers is general, having a probability density function denoted by $b_j(x), \quad x \geq 0$, and expectation $\bar{x_j} = \dfrac{1}{\mu_j}$. We shall let $\rho_j = \dfrac{\lambda_j}{\mu_j}$ and assume that $\rho = \sum_{j=1}^{J} \rho_j < 1$. Let L_j and L_j^q be the mean number of class j customers in the system and waiting in the queue respectively; we shall let $E[R_j]$ and W_j^q be the mean response time and mean time spent waiting respectively; and we shall let $E[\Re_j]$ be the expected residual service time of a class j customer. We consider first the non-preemptive priority scheduling case. Here the time an arriving class j customer, to whom we refer to as the tagged customer, spends waiting in the queue is the sum of the following three time periods.
1. The residual service time of the customer in service.
2. The sum of the service time of all customers of classes 1 through j that are already present the moment the tagged customer arrives.
3. The sum of the service time of all higher-priority customers who arrive during the time the tagged customer spends waiting in the queue.

The response time of the tagged customer is found by adding its service requirement to this total. Two features allow us to treat the first of these three periods relatively simply. First, an arriving customer, whatever the class, must permit the customer in service to finish its service: the probability that the customer in service is of class j is given by ρ_j (observe that $\rho = \rho_1 + \rho_2 + \cdots + \rho_J$ is the probability that the server is busy). Second, we know that the expected remaining service time of any customer in service as seen by an arriving customer whose arrival process is Poisson is $E[\Re_j]$ if the customer in service is of class j. Thus the expected residual service time as experienced by the tagged customer is $E[\Re] = \sum_{i=1}^{J} E[\Re_i]$. The second time period is found by using the PASTA property: the tagged customer finds the queueing system at steady state, and hence observes the stationary distribution of all classes of customer already present. Given that, at equilibrium, the mean number of class i customers waiting in the queue is L_i^q, the mean duration to serve all customers of equal or higher priority found by the arriving tagged customer is $\sum_{i=1}^{j} \bar{x_i}$. The mean length of time such customers spend waiting in the queue is $W_1^q = L_1^q \bar{x_1} + \sum_{i=1}^{J} \rho_i E[\Re_i]$. By applying Little's law, $L_1^q = \lambda_1 W_1^q$ and $W_1^q = \lambda_1 W_1^q \bar{x_1} + \sum_{i=1}^{J} \rho_i E[\Re_i] = \rho_1 W_1^q + \sum_{i=1}^{J} \rho_i E[\Re_i]$ leads to

$$W_1^q = \frac{\sum_{i=1}^{J} \rho_i E[\Re_i]}{1 - \rho_1} \tag{2}$$

From this result, the response time of class 1 customers can be computed and then, using Littles law, the mean number in the system and the mean number waiting in the queue. Equation (2) will serve as the basis clause of a recurrence relation involving customers of lower priority classes. For customers of class 2 or greater, we need to compute the third time period, the time spent waiting for the service completion of all higher-priority customers who arrive while the tagged customer is waiting. Given that we have defined W_j^q to be the total time spent waiting by a class j customer, the time spent serving higher-priority customers who arrive during this wait is $\sum_{i=1}^{J-1} \lambda_i W_j^q \bar{x_i} = W_j^q \sum_{i=1}^{J-1} \rho_i$. Thus the total time spent waiting by a class j customer is

$$W_j^q = \sum_{i=1}^{J} \rho_i E[\Re_i] + \sum_{i=1}^{j} L_1^q \bar{x_1} + W_j^q \sum_{i=1}^{J-1} \rho_i.$$

$$W_j^q (1 - \sum_{i=1}^{J-1} \rho_i) = \sum_{i=1}^{J} \rho_i E[\Re_i] + \sum_{i=1}^{j} \lambda_i W_i^q \bar{x_i}$$

$$= \sum_{i=1}^{j-1} \rho_i E[\Re_i] + \sum_{i=1}^{J-1} \rho_i W_i^q + \rho_j W_j^q \tag{3}$$

which leads to

$$W_j^q (1 - \sum_{i=1}^{J-1} \rho_i) = \sum_{i=1}^{j} \rho_i E[\Re_i] + \sum_{i=1}^{J-1} \rho_i W_i^q$$

Comparing the right-hand side of this equation with the right-hand side of Equation (3), $W_j^q (1 - \sum_{i=1}^{j} \rho_i) = W_{j\,j-1}^q (1 - \sum_{i=1}^{j-2} \rho_i)$ and multiplying both sides with $(1 - \sum_{i=1}^{j-1} \rho_i)$,

$$W_j^q (1 - \sum_{i=1}^{j} \rho_i)(1 - \sum_{i=1}^{j-1} \rho_i) = W_{j\,j-1}^q (1 - \sum_{i=1}^{j-2} \rho_i)(1 - \sum_{i=1}^{j-1} \rho_i)$$

For the highest-priority customers. Repeated application of this recurrence leads to

$$W_j^q(1-\sum_{i=1}^{j}\rho_i)(1-\sum_{i=1}^{j-1}\rho_i) = W_1^q(1-\rho_1)$$

By substituting for W_1^q using equation (2),

$$W_j^q = \frac{\sum_{i=1}^{J}\rho_i E[\Re_i]}{(1-\sum_{i=1}^{j}\rho_i)(1-\sum_{i=1}^{j-1}\rho_i)}, \quad j=1,2,\cdots,J \tag{4}$$

The mean response time of class j customers can now be found by adding $\bar{x_j}$ to this equation, and then Littles law can be used to determine the mean number of class k customers in the system and waiting in the queue.

2.2.2 M/G/1: preempt-resume priority scheduling

This is when the scheduling policy is such that a low-priority customer in service is interrupted to allow an arriving customer of a higher priority to begin service immediately. The interrupted customer is later scheduled to continue its service from the point at which it was interrupted. With this policy, customers of class $j+1, j+2, \cdots, J$ do not affect the progress of class j customers; customers with lower priorities are essentially invisible to higher-priority customers. In light of this we may set $\lambda_k = 0$ for $k = j+1, j+2, \cdots, J$ when analyzing the performance of class j customers. Therefore T_1^A is equal to the sum of the residual service time of the customer in service and the time required to serve all waiting customers: i.e., $T_1^A = \sum_{i=1}^{j}\rho_i E[\Re_i] + \sum_{i=1}^{j}\rho_i T_1^A$ which leads to $T_1^A = \dfrac{\sum_{i=1}^{j}\rho_i E[\Re_i]}{1-\sum_{i=1}^{j}\rho_i}$.

Also

$$T_2^A = \sum_{i=1}^{j-1}\rho_i E[\mathbb{R}_j] = (W_j^q + 1/\mu_j)\sum_{i=1}^{j-1}\rho_i.$$

The total waiting time for a class j customer is then equal to

$$W_j^q = T_1^A + T_2^A = \frac{\sum_{i=1}^{j}\rho_i E[\Re_i]}{1-\sum_{i=1}^{j}\rho_i} + (W_j^q + 1/\mu_j)\sum_{i=1}^{j-1}\rho_i.$$

Therefore,

$$W_j^q = \frac{\sum_{i=1}^{J}\rho_i E[\Re_i]}{(1-\sum_{i=1}^{j}\rho_i)(1-\sum_{i=1}^{j-1}\rho_i)} + \frac{1/\mu_j\sum_{i=1}^{j-1}\rho_i}{1-\sum_{i=1}^{j}\rho_i} \tag{5}$$

The mean response or sojourn time of a class j customer is

$$E[R_j] = W_j^q + \frac{1}{\mu_j} = \frac{\sum_{i=1}^{J}\rho_i E[\Re_i]}{(1-\sum_{i=1}^{j}\rho_i)(1-\sum_{i=1}^{j-1}\rho_i)} + \frac{1/\mu_j}{1-\sum_{i=1}^{j}\rho_i}$$

It is also possible to solve for the mean response time directly from

$$E[R_j] = T_1^A + T_2^A + \frac{1}{\mu_j}$$

$$E[R_j] = \frac{1}{(1-\sum_{i=1}^{j}\rho_i)}[\frac{\sum_{i=1}^{J}\rho_i E[\Re_i]}{(1-\sum_{i=1}^{j}\rho_i)} + \frac{1}{\mu_j}]$$

The time spent waiting by a class j customer prior to entering service for the first time, is given by

$$T_1^B = \frac{\sum_{i=1}^{J}\rho_i E[\Re_i]}{(1-\sum_{i=1}^{j}\rho_i)(1-\sum_{i=1}^{j-1}\rho_i)}$$

The remaining time spent in service and in interrupted periods caused by higher priority customers who arrive after the customer first starts its service, is given by

$$T_2^B = \frac{1/\mu_j \sum_{i=1}^{j-1}\rho_i}{1-\sum_{i=1}^{j}\rho_i}$$

which is the second term in Equaion 7.

3. Numerical examples

Example 1 Consider a queueing system which caters to three different classes of customers whose arrival processes are all Poisson. The most important customers require $\bar{x_1} = 1$ time unit of service and have a mean interarrival period of $\frac{1}{\lambda_1} = 4$ time units. The corresponding values for classes 2 and 3 are $\bar{x_2} = 5, \quad \frac{1}{\lambda_2} = 20$, and $\bar{x_3} = 20, \quad \frac{1}{\lambda_2} = 50$, respectively. Thus $\rho_1 = 1/4, \quad \rho_2 = 5/20, \quad \rho_3 = 20/50$, and $\rho = \rho_1 + \rho_2 + \rho_2 = .9 < 1$. To facilitate the computation of the residual service times, we shall assume that all service time distributions are deterministic. Thus $\Re_1 = .5, \quad \Re_2 = 2.5$, and $\Re_3 = 10.0$. With the nonpreemptive priority policy, the times spent waiting in the queue by a customer of each of the three classes are as follows:

$$W_1^q = \frac{\rho_1\Re_1 + \rho_2\Re_2 + \rho_3\Re_3}{1-\rho_1} = \frac{4.75}{0.75} = 6.333$$

$$W_2^q = \frac{\rho_1\Re_1 + \rho_2\Re_2 + \rho_3\Re_3}{(1-\rho_1)(1-\rho_1-\rho_2)} = \frac{4.75}{0.50 \times 0.75} = 12.667$$

$$W_3^q = \frac{\rho_1\Re_1 + \rho_2\Re_2 + \rho_3\Re_3}{(1-\rho_1-\rho_2)(1-\rho_1-\rho_2-\rho_3)} = \frac{4.75}{0.10 \times 0.5} = 95$$

For the preemptive- resume scheduling, the corresponding waiting times are

$$W_1^q = \frac{\rho_1\Re_1}{1-\rho_1} = \frac{0.125}{0.75} = 0.1667$$

$$W_2^q = \frac{\rho_1\Re_1 + \rho_2\Re_2}{(1-\rho_1)(1-\rho_1-\rho_2)} + \frac{\rho_1/\mu_2}{1-\rho_1} = \frac{0.75}{0.50 \times 0.75} + \frac{1.25}{0.75} = 3.667$$

$$W_3^q = \frac{\rho_1\Re_1 + \rho_2\Re_2 + \rho_3\Re_3}{(1-\rho_1-\rho_2)(1-\rho_1-\rho_2-\rho_3)} + \frac{(\rho_1+\rho_1)/\mu_3}{1-\rho_1-\rho_2} = \frac{4.75}{0.10 \times 0.5} + \frac{10}{0.5} = 115.0$$

Example 2 Consider a nonpreemptive priority M/G/1 queue with three classes of customer. The first two moments of the highest priority customers are given by $E[S_1] = 5$ and $E[S_1^2] = 28$, respectively. The corresponding values for classes 2 and 3 are $E[S_2] = 4; \quad E[S_2^2] = 24$ and $E[S_3] = 22; \quad E[S_3^2] = 1,184$, respectively. Arrival rates for the three classes are such that $\rho_1 = 1/3, \quad \rho_2 = 11/30$, and $\rho_3 = 11/60$, respectively, and so $\rho = \rho_1 + \rho_2 + \rho_3 = 53/60$. The mean residual service times are

$$E[\Re_1] = \frac{E[S_1^2]}{2E[S_1]} = \frac{28}{10} = 2.8,$$

$$E[\Re_2] = \frac{E[S_2^2]}{2E[S_2]} = \frac{24}{8} = 3,$$

$$E[\Re_3] = \frac{E[S_3^2]}{2E[S_3]} = \frac{1184}{44} = 26.9091.$$

The mean time spent waiting by customers of each class given that

$$\sum_{j=1}^{3} \rho_j E[\Re_j] = 6.9667,$$

$$W_1^q = \frac{\sum_{j=1}^{3} \rho_j E[\Re_j]}{1-\rho_1} = \frac{6.9667}{2/3} = 10.45,$$

$$W_2^q = \frac{\sum_{j=1}^{3} \rho_j E[\Re_j]}{(1-\rho_1-\rho_2)(1-\rho_1)} = \frac{6.9667}{3/10 \times 2/3} = 34.8333,$$

$$W_3^q = \frac{\sum_{j=1}^{3} \rho_j E[\Re_j]}{(1-\rho_1-\rho_2-\rho_3)(1-\rho_1-\rho_2)} = \frac{6.9667}{7/60 \times 3/10} = 199.0476.$$

4. Conclusion

This study has surveyed the questions of priorities scheduling for $M/G/1$ queueing model that arise in queueing theory and its application such as Preempt-Resume Priority Scheduling and Nonpreemptive priority policy. we were able to developed efficient algorithms to calculate the time it takes to serve all customers of equal or higher priority that are present at the moment a class j customer arrives, we as well obtained the following measures for the two priorities scheduling, mean response or sojourn time of a class j customers, the number waitng in the queue, average number of class j customers present in the system and total waiting time for a class j customer while the numerical examples were presented to reflect its applicatiation to real life problems.

References

[1] Agboola S. O. (2016): Repairman Problem with Multiple Batch Deterministic Repair. Unpublished Ph.D. Mathematics (Statistics Option) Thesis submitted to Department of Mathematics, Obafemi Awolowo University, Ile-Ife.

[2] Agboola, S.O. (2007): On the Analysis of M/G/1 Queues. Unpublished M.Sc. (Mathematics) Thesis submitted to Department of Mathematics, University of Ibadan.

[3] Agboola, S.O. (2011): The Analysis of Markov Inter-arrival Queues Model with K - Server under Various Service Points. Unpublished M.Sc. (Statistics) Thesis submitted to Department of Mathematics, Obafemi Awolowo University, Ile-Ife.
[4] Bolch, G. Greiner, S. and Trivedi, K. S. (1998): Queueing Networks and Markov Chains. Wiley Interscience, New York.
[5] Jain, M. (2013): Transient analysis of machining system with service interruption, mixed standbys and priority. International Journal of Mathematics in Operations Research, vol.5, no. 5, pp. 604-625 (indersience).
[6] Jain, M. and Singh M., (2013): Bi-level control of degraded machining system with warm standby setup and vacation, Applied Mathematical Modelling, 28, 1015 - 1026.
[7] Kleinrock, L. (1975): Queueing systems, New York,pp. 60 - 158.
[8] Law, A. M. and Kelton, W. D. (2000): Simulation Modeling and Analysis. Third Edition, McGraw-Hill, New York.
[9] Lucantoni, D.M. (1993): The BMAP/G/1 Queue: Models and Techniques for performance evaluation of computer and communications systems, Springer.
[10] Medhi, J. (1980): Stochastic application of queueing theory. Guahandi University, Guwahata, Indian. pp. 23 - 120.
[11] Meini, B. (1997): New Convergence Results on Functional Techniques for the Numerical Solution of $M/G/1$ Type Markov Chains. Numer. Math., Vol. 78, pp. 39-58.
[12] Ross, S. M. (1997): Simulation. Second edition, Academic Press, New York.
[13] Saad, Y. (2003): Iterative Methods for Sparse Linear Systems. Second edition, SIAM Publications, Philadelphia.
[14] William Stewart (2009): probability, Markov chains, Queues and Simulation, Princeton University press, Princeton and oxford.