A Two Stage Queuing System with delay time

S.Shyamala¹,*, K.Suresh², S.R. Anantha Lakshmi³

¹,² Department of Mathematics, Government Arts College, Tiruvannamalai 606603, India
³ Department of Mathematics, Easwari Engineering College, Chennai 600089, India
E-mail: subramaniyanshyamala@gmail.com

Abstract. This paper discusses about a two stage queueing system with shared server and the delay time between two end to end responses. This system is examined as almost the process of quasi-birth and death and the necessary and sufficient conditions for system stability are discussed. The ErnPS algorithm for VoIP Services has recently been investigated for average delay in queueing and it has been proven that ErnPS can allow more users than UGS. However, we need the distribution of probabilities of queue delays in order to include information necessary for QoS rather than the average queue delays. Some fascinating performance measures are achieved by matrix geometric method. The server wishes to determine the optimal point of differentiation and its optimal buffer size. Finally we find the delay time differentiation and the specific probabilistic system descriptors.

Keywords: Buffer size, Delay time, Shared server, Stability.

1. Introduction

In recent years, there was an improved interest in queuing structures and their applications. This is predominantly because of the reality that queuing models certainly rise up with inside the overall performance analysis of a huge range of systems in data distributed networks, telecommunications, traffic management on high-speed networks and manufacturing engineering. Further, new outcomes in queuing systems have frequently been stimulated with the aid of using new technological advances in computer systems and data communication networks. Recent queues were broadly used to model many issues in telephone switching systems, telecommunication networks and computer systems competing to benefit carrier from a central server.

The communication network provides a wide range of services such as data communication, telephone and interactive games. The development of high-speed networks has caused distributed applications in the field of computer communication networks. Distributed applications are divided into components that can communicate and run on distributed systems (for example, interconnected servers and databases). In the highly competitive information and telecommunications market, being able to provide applications with high performance and guaranteed reliability has become increasingly important to the users in an economical way. IEEE 802.16e is intended to support large-scale, high-data services and multimedia networks. IEEE 802.16e aims to bridge the difference between WLAN (Wireless Local Area Networks) high data speeds and high WANs (Wide Area Networks) mobility. Five traffic services support IEEE 802.16e systems for uplink planning algorithms, such as UGS, Real Time polling service, rtPS,
nrtPS, BE (best effort service), and ErtPS (Extended Real Time polling service) SEE 802.16e. IEEE 802.16e systems support 5 UGS traffic systems (IEEE 802.16-2004; IEEE 802.16a-2003). The UGS, rtPS and ErtPS are among these five algorithms designed to help service flows in real time. Voice over IP (VoIP) service in IEEE 802.16e may be provided as a form of Unsolicted Grant Service (UGS) or Real-Time Polling Service (rtPS) or Extended ERTPS. However, the UGS and the rtPS algorithm are considered to not be suitable for VoIP service scheduling as a result of wasting time (in the case of UGS), and MAC overhead and queue delays (in the case of rtPS) and ErtPS were also applied to the IEEE 802.16e-2005 Standards for VoIP Service.

For latest bibliographies about retrial queues, See Falin (1990) [6], Falin and Templeton (1997)[5], [3], [4], [7], [12], [13] and Artalejo (2001)[2], a good survey of results on sojourn times in queuing networks offered by Boxma and Daduna is available. There are several approximations for the expected value of the sojourn time available for the $M|G|C – FCFS$ queue. Nates and Rao (1990) [11] developed an attractive method to create an infinitesimal generator, which is basically a quasi birth and death process (QBD) with several boundary states. In the Queueuing Literature (see Neuts (1981)[10] and Latoche and Ramaswami (1999)[9], QBD procedures have been extensively studied to analyse the methods to describe the ergodicity condition and to compute the stationary probabilities. The $M/G/1$ retrial queue with feedback and starting failures is addressed by Krishnakumar et.al(2002)[8]. Recently Artalejo et al. Al (2000, 2001)[1], [2] studied numerical solutions. According to this solution, the number of customers seeking service by matrix-geometric technique does not have to do with repeated attempts.

In this paper, we will study the two stage tandem queue with a shared server. Here, we consider the delay time between two end-to-end responses. End-to-end response does not depend on the service time distribution. The two layered tandem queuing model is a simplified model, which considers the actual queuing network in the communication network. We create relations among distributions pertaining to the various subsystems to determine the distributions of the service times for the two servers on each sub system. The UGS, rtPS and ErtPS are among the five algorithms to support real-time service flows. This article provides an overview of the Quasi-Birth-and-Death chain (QBD) of stage dependence. We analyze the level-dependent nested QBD chain, and give a complete characterization of its fundamental matrix based on the minimum non-negative solution of a series of quadratic matrix equations. If the off-diagonal elements of the matrix $Q$ are non-negative and the diagonal elements are non positive and the sum of the rows is less than or equal to zero, it is called a generator. If the sum of rows of a generator $Q$ are all zero, $Qe = 0$, where $e$ is a column vector of appropriate size with all 1 and then $Q$ is called a conservative generator.

Delay time differentiation is an approach in which an arrival receives a service from station 1 and then receives a service from station 2 to complete the service. The time delay is triggered by a delay where the dependency between the first and second stations is decreased. The Delay time is an approach whereby an arrival is partly served by station 1 and then completed by station 2 only after an arrival enters into the system. The Delay time principle is to delay the time when components partially served are assigned to a given finished service. When dealing with uncertainty in the system, delay can increase the flexibility of the server. Part exchange can be accomplished. The model is described in this paper in section 2. Section 3 deals with the performance measures and section 4 discusses the stationary distribution.

2. Model description

There is a tandem of two stations in the queuing network. These service locations represent locations where the server can process requests. We assume a customer does not need to wait
in Station 1, we assume that a customer \( \theta (0 < \theta < 1) \) gets 100% service from Station 1. Then at station 2 serve the remaining \((1 - \theta)\) customers. Note if \( \theta = 0 \), the customer doesn’t have to wait for station 2 service. Instead, if \( \theta = 1 \), then only in station 1 the customer provides the service and in station 2 it does not. These two cases we don’t consider, so we have \( 0 < \theta < 1 \).

In fact, it could be feasible for Station 2 to have a small number of values of \( q \). We have decided to model \( \theta \) as a continuous variable to give us a better idea of the relationship between \( \theta \) and the system performance. The assumption makes our estimation analysis simpler as well. Therefore, we present findings as if station 2 may enforce a \( \theta \) value. In the event that this is not the case, our model will quickly identify a variety of feasible alternatives for the best \( \theta \) option.

Station 1 takes the time to service an item, where the parameter \( \frac{\delta \mu}{\theta} \) distribution is exponential, and where \( \theta > 0 \) is used as a scaling factor in the association of station 1 with station2 service rates. Station 1 sends the customer to station 2 after service, but as soon as the customer number reaches \( K \) in station 2. Station 1 stops serving until the waiting number is lower than \( k \). Station 2 takes the time to complete the processing on the item with parameter \( \frac{\mu}{1-\theta} \) as an exponential distribution. Changes in the time needed for station 2 to fulfill various customer needs are assumed to take account of in the time distribution of processing.

3. Performance measures
At a Poisson rate of \( \lambda \), customers arrive at these stations and are served by FCFS. We assume that customers have infinite space to wait. Customers may not be able to get full service at station 1, especially if the main service is needed at the station 2. The time delay may require the use of common components or functions within the product in order to make it suitable for different end products.

Let \( \phi \) be the probability that a semi-served arrival does not receive service and leaves the system and \( \phi \) will be increased by \( q \) on a monotonous basis. \((1 - \phi)\) is the probability to receive service at Station 2 for the semi served arrival. Station 1 now has \( \frac{\delta \mu(1-\phi)}{\theta} \) as its service rate. The \( \phi \) value can be considered as the characteristic of the arrival. Assume \( \phi = b\theta^n \), \( n \geq 1 \), \( 0 < b < 1 \), Assume \( n = 1 \). i.e, \( \phi = b\theta \).

Consider a Markov chain with continuous time \( \{(i,j), i \geq 0, 0 \leq j \leq k\} \), where \( i \) is the number of arrivals waiting for service in the system at station 1. \( j \) is the number of semi-served arrivals at station 2 in the system. This chain’s generator matrix \( Q \) is given by

\[
Q = \begin{pmatrix}
B & A_0 & \cdot & \cdot & \cdot \\
A_2 & A_1 & A_0 & \cdot & \cdot \\
\cdot & A_2 & A_1 & A_0 & \cdot \\
\cdot & A_2 & A_1 & A_0 & \cdot 
\end{pmatrix}
\]

Where,

\[
B = \begin{pmatrix}
-(\lambda + \frac{\delta \mu(1-\phi)}{\theta}) & \frac{\delta \mu(1-\phi)}{\theta} & \cdot & \cdot & \cdot \\
\cdot & -(\lambda + \frac{\delta \mu(1-\phi)}{\theta}) & \frac{\delta \mu(1-\phi)}{\theta} & \cdot & \cdot \\
\cdot & \cdot & -(\lambda + \frac{\delta \mu(1-\phi)}{\theta}) & \cdot & \cdot \\
\cdot & \cdot & \cdot & -(\lambda + \frac{\mu}{1-\theta}) & \cdot \\
\cdot & \cdot & \cdot & \cdot & -\left(\lambda + \frac{\mu}{1-\theta}\right) 
\end{pmatrix}
\]

\[
A_1 = \begin{pmatrix}
-\lambda - \frac{\delta \mu(1-\phi)}{\theta} & \cdot & \cdot & \cdot & \cdot \\
\cdot & -\left(\lambda + \frac{\delta \mu(1-\phi)}{\theta} + \frac{\mu}{1-\theta}\right) & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & -\left(\lambda + \frac{\mu}{1-\theta}\right) & \cdot \\
\cdot & \cdot & \cdot & \cdot & -\left(\lambda + \frac{\mu}{1-\theta}\right) 
\end{pmatrix}
\]
\[ A_2 = \begin{bmatrix} 0 & 0 \\ I_{1-\theta} & 0 \end{bmatrix}, \]

\[ A_0 = I \lambda \]

\[ A_0, A_1, A_2 \] are \((k+1)\times(k+1)\) matrices which show a rate that increases by one the number of arrivals in the system, decreases by one level or remains on the same level. \((\text{for level exceeding 0})\)

\(B\) is a matrix rate where the system is transfer at level one from level zero.

Let \(A = A_0 + A_1 + A_2\)

\[ A = \begin{bmatrix} \lambda - \lambda - \frac{\delta \mu (1-\phi)}{\theta} + \theta & \frac{\delta \mu (1-\phi)}{\theta} + \frac{\mu}{1-\theta} & \frac{\delta \mu (1-\phi)}{\theta} & \ldots \\ \frac{\mu}{1-\theta} & -\frac{\mu}{1-\theta} \end{bmatrix} \]

\(A\) is a generator matrix with stationary distribution \(\pi = [\pi_0, \pi_1, \pi_2, \ldots, \pi_k]\), given as a solution to \(\pi A = 0, \pi_1 = 1\).

Let \(\alpha = \frac{\delta (1-\phi)(1-\theta)}{\theta}\) be the ratio of service rate of station 1 to that of rate of station 2.

\[ \pi A_2 e > \pi A_0 e \] (Neuts 1981 Thm 3.1.1 pp82 [10])

\[ \pi A_0 e < \pi A_2 e \] is the necessary and sufficient condition for stability.

\[ \lambda (1 - \alpha^{k+1}) < \frac{\mu}{1-\theta} \alpha (1 - \alpha^k) \]

\[ \lambda < \frac{\delta \mu (1 - \phi) (1 - \alpha^k)}{\theta (1 - \alpha^{k+1})} \]

\[ \frac{(1-\alpha^{k+1})}{(1-\alpha^{k+1})} \] is the problem that the system has fewer than \(K\) items waiting for service. The term \(\frac{\delta \mu (1 - \phi) (1 - \alpha^k)}{\theta (1 - \alpha^{k+1})}\) is the effective rate at which the station 1 is able to send items to station 2.

4. Stationary Distribution

Under the stability condition, the stationary problem vector \(X\) exits. The stationary problem vector \(X\), partitioned as \(X = (X_0, X_1, X_2, \ldots)\) is given by

\[ XQ = 0, \quad X_1 = 1 \]

\[ X_0(B_0 + RA_2) = 0 \]

\[ X_n = X_0 R^n, n \geq 1 \]

\[ X_0 (I - R)^{-1} e = 1 \]

\(I\) is the unit matrix, while \(R\) is the least non negative solution for \(R^2 A_2 + RA_1 + A_0 = 0\) matrix equation, of which less than one is spectral radius and \(X_n, n \geq 0\) are \((k+1)\) dimensional row vector. Formula (1) is a closed form expression but contains all the parameters of non-trivial
relations. Therefore, when defining the remaining parameters, we wish to examine the impact of system parameters in the main performance measures.

To compute $R$

$$R(n+1) = -(A_0 + R_0^2 A_2) A_1^{-1} = -A_0 A_1^{-1} - R_0^2 A_2 A_1^{-1}, n \geq 0$$

Define

$$P_{ij} = (X_i)_{j}$$

$$Max \{(R_n)_{i,j} - (R_{n-1})_{i,j}\} \leq \epsilon = 10^{-8}. \text{ With } R_0(0) = 0$$

(i) The average number of customers (including the individual being served) is calculated as follows $E(N) = X_1(1 - R)^{-2}I$.

(ii) Mean number of customers delay $E(W) = E(L)$ (Using Littles formula)

(iii) Mean number of units received service from station 1 and waiting for service in station 2 is given by

$$E(S) = X_0(1 - R)^{-1} V$$

$$V = [0, 1, 2, \cdots K]^{T}$$

(iv) The mean number of units received service from station 1 and does not receive service from station 2 per unit time is given by

$$E(U) = (1 - P_{r(j=k)})V_{0j}$$

where $j$ is number of items received service from station 1.

5. Conclusion

The features of a two stage queueing system with shared server and delay time between two end to end response is highlighted. Explicit expressions for the probabilities of the system are obtained. The performance measures such as mean number of customers, customer delay and units received for service are carried out. Therefore it is emphasized that the queueing system is not of theoretical interest but may attract practitioners based on the easiness of the implementation.

6. References

[1] J.R. Artalejo, A Gomez-Corral and M.F. Neuts, 2000, Numerical analysis of multiserver retrial queues operating under full access policy, (Advances in Algorithmic methods of Stochastic Models, Eds. G. Latrobe and Peter Taylor, Notable Publications Inc., New Jersey, USA,) pp. 1-19.

[2] J.R. Artalejo, A. Gomez-Corral and M.F. Neuts, 2001, Analysis of multiserver queues with constant retrial rate, (European Journal of operations Research vol 135), 569-581.

[3] B.D. Choi, Y.W. Shin and W.C., Ahn, 1992, Retrial queues with collision arising from unslotted CSMA/CD protocol, (queueing systems, vol 11), 335-336.

[4] J.W. Cohen, 1957, Basic problems of telephone traffic theory and the influence of repeated calls, (Phillips Telecommunication Review, 18), 49-100.

[5] G.I. Falin and J.C.G. Templeton, 1997, Retrial Queues, (Chapman and Hall, London.).

[6] G.I. Falin, 1990, A survey of retrial queues, (Queueing Systems, vol 7), 127-168.

[7] G. Fayolle, 1986, A simple telephone exchange with delayed feedbacks, In Teletraffic analysis and computer performance evaluation, Edited by O.J. Boxma, J.W. Cohen, and H.C. Tijms, (Elsevier, Amsterdam, ) pp. 245-253.

[8] B. Krishna Kumar, S. Pivai Madheswari and A. Vijayakumar, 2002, The M/G/1 retrial queue with feedback and starting failures, (Applied Mathematical Modelling, 26), 1057-1075.

[9] G. Latouche and V. Ramaswami, 1999, Introduction to Matrix Analytic methods in Stochastic Modeling, (ASA-SIAM, Philadelphia.).

[10] M.F. Neuts, 1981, Matrix-Geometric solutions in Stochastic Models An Algorithmic Approach, (John Hopkins University Press, Baltimore, MD).
[11] M.F. Neuts and B.M. Rao, 1990, *Numerical investigation of a multi server retrial model*, (Queuing Systems, 7), 169-190.

[12] Neuts, M. F., 1989, *Structured Stochastic matrices of M/G/1 type and their applications*, (New York: Marcel Dekker).

[13] H. Takagi, 1993, *Queueing Analysis. Amsterdam, Netherlands: North-Holland.*