# Estimating the time buffer size for different customer classes in lean supply chain operations

**Xiaofeng Zhao**
University of Mary Washington, Virginia,U.S.A.

**JianrongHou**
Shanghai Jiaotong University, Shanghai, China

**Key words**
Priority queue, buffer management, lean operations, supply chain management

**Abstract**
*This paper uses quantitative methods to estimate the sizes of the time buffer for different customer classes in lean supply chain operation. It discusses mathematically exact expressions for expected waiting time and the variation of waiting time for multi-classes. A queuing theory based method is applied to calculate the time buffer sizes. It provides a mathematically tractable expression for Markov queues and approximates the mean and variance of waiting time in general queue. The results are implemented in manual or spreadsheet calculations which can be used to conduct what-if analysis.*

## 1. Introduction

Supply chains have proven instrumental in improving efficiency within many industries, such as Wal-Mart in retail and Dell in computers .The link between supply chain and financial performance revealed that virtually all winning business strategies have, at their core, supply chain strategies that provide a competitive advantage. It becomes increasingly apparent that the battleground has been shifting from competition between enterprises to competition between supply chains. A survey by the Supply Chain Council found that, on average, enterprises spent about 11 percent of revenue on supply chain management, yet best-in-class enterprises go to the cost down to between 3 and 6 percent. Reducing logistics driven costs by 5 percent would save about $50 billion in US.

Lean supply chain strategies focus on waste reduction, helping firms eliminate non-value adding activities related to excess time, labor, equipment, space, and inventories across the supply chain (Ma 2011). Such strategies enable firms to improve quality, reduce costs, and improve service to customers as traditional batch and queue mass production and supply chain approaches are transformed. Lean supply chains adapt to changing customer needs and still deliver products quickly. As a result, firms that are part of a lean supply chain have lower costs than their competitors. Lean practices are becoming increasingly difficult to implement and sustain as supply chains increase in complexity and length.

Although the potential benefits are huge, building and managing a lean supply chain poses a challenge because supply chain activities are so highly interconnected. The present business environment is also significantly more challenging than the business environment of the production-centric era that prevailed for the greater part of 20th century. In today's customer centric world, production capacity exceeds customer demand in many industries. Prices are now determined by more competitive market forces than existed when capacity constrained sales volume. Consumers are demanding better products, and want them cheaper and faster. To stay competitive, firms are compelled to respond to these customer demands even as product life cycles are getting smaller.

The consumer centric era requires business to manage their supply chain in a radically different manner. To respond to increasingly demanding consumers, firms not only have to excel at producing the goods or service they sell; they must also excel at delivering products quickly and efficiently to the consumer. A firm that provides quick response to customers' needs is able to attract more clients and encourage brand loyalty, increasing its market share; it can even win a price premium for speed and punctuality of its deliveries. A certain degree of elasticity to delivery time characterizes not only consumers' demand, but intermediaries' too, due to lower risk for less anticipated replenishment orders.

Both empirical evidence and logic suggest that there is a strong negative correlation between waiting time and a customer's evaluation of the quality of a supply chain service (Davis and Heineke 1998). A waiting time guarantee is a firm's commitment to its customers that it will deliver the products within a specified period of time. A firm can enhance customer waiting experiences by providing assurances of products or services within the expected time as well as evidence for the progress that customers are making in the system. Firms commit themselves to a given waiting time guarantee by selecting appropriate capacity levels. To estimate the waiting time guarantee (maximum waiting time), managers need to know both the average of waiting time and the variance of waiting time. In many systems, the worst case-time buffer value of flow time is very relevant because it represents the turnaround time that can safely be promised to the customers. Thus, it is essential to recognize and understand how to estimate the buffer time in supply chain operations.

Research on customer waiting time has traditionally been the domain of queuing theory. Queues occur because of uncertainty in the environment; whenever the demand for service exceeds the ability to provide service, a queue forms. A major distinction classifies queues according to the number of servers and the distributions that characterize the arrival rates of customers (or their inter-arrival times) and the service times. From a statistical perspective, the random arrival process is not necessary described with the Poisson probability distribution. Similarly, the exponential probability distribution is inappropriate when a wide range of service times is possible (Hopp and Spearman 2000). Kendall notation A/B/n is widely accepted in queuing system. In this notation, the A, B, and n denote, respectively, the inter-arrival time distribution, the service time distribution and the number of servers. In other words, most service operation queuing problems are represented by a general GI/G/n system (G for general, I for independent arrivals).

Recent years have witnessed a growing volume of good quality approximations for the GI/G/n queue (Kimura 1986, Shore 1988, Whitt 1993, 2004, Holland and Griffiths1999, Atkinson 2008). While the accuracy of these approximations is usually satisfactory, they often result in algebraically intractable expressions. This hinders attempts to derive closed-form solutions to the decision variables incorporated in optimization models. It often leads to the use of complex numerical methods or to recursive schemes of calculation. Furthermore, actual application of many of these approximations is often obstructed due to the thorough specification that is needed of inter-arrival or service time distribution. No general theoretical formula exists that provides a platform to calculate control limits for GI/G/n.

In addition, all literature focuses on the probability of customer waiting and the average waiting time. The analysis of the variance of waiting time remains unsolved due to its inherent complexity. There exists no mathematically tractable general formula for approximating the standard deviation of waiting time $\sigma_q$ in the GI/G/n queue. Only bounds or approximations of waiting time have been found in the literature. When these

bounds are used as approximations, they appear to be rather crude (Whitt 2004, Metters and Pullman 2003). Simple and easy to use formulae are available for determining the average (e.g. Sakasegawa 1977), but not the standard deviation. For operations practitioners, this requires spreadsheet formulae for determining not only the average waiting time but also the standard deviation of the waiting time in the GI/G/n queue. Authors are not aware of any other spreadsheet model that is specifically designed to analyze the variance of waiting time in GI/G/n queue.

This paper develops an easy to use spreadsheet model to estimate the mean and standard deviation of waiting time for different customer classes. The approximation requires only the mean and standard deviation or the coefficient of variation of the inter-arrival and service time distributions, and the number of servers. It is simple enough to be implemented in spreadsheet calculations. The rest of this paper is organized as follows. In section 2, we derive exact expression for the coefficient of variation of waiting time for G/M/n, M/G/1 queues and develop interpolation approximation for variance of waiting time for the general queue GI/G/n. In section 4, numerical results show that the approximations are accurate enough to be applied to practical service operations. Section 5 delivers concluding remarks.

## 2. **Analytical Models**

To develop the approximation of the standard deviation of waiting time in the GI/G/n queue, we have studied the equivalent problem of finding a mathematically tractable formula to approximate the coefficient of variation of waiting time $c_q = \sigma_q / W_q$, where $W_q$ and $\sigma_q$ are respectively the average and standard deviation of waiting time. There exist some good approximations for the average waiting time (Kimura 1986, Whitt 1993). For instance, Sakasegawa (1977) presented the following closed-form expression for the average waiting time in the GI/G/n queue:

$$W_q(GI/G/n) = \left(\frac{c_a^2 + c_s^2}{2}\right)\left(\frac{\rho^{\sqrt{2(n+1)}-1}}{n(1-\rho)}\right)\left(\frac{1}{\mu}\right) \qquad (1)$$

$c_a$ Is the coefficient of variation of inter-arrival time and $c_s$ the coefficient of variation of service time.

This formula offers several advantages (Whitt 1993). Although it may appear complex, it does not require any type of iterative algorithm to solve and therefore can be easily implemented into a spreadsheet program. This also makes it possible to couple the single-station approximation with the multiple-server to create a spreadsheet tool for analyzing the performance of a series of queues. The formula is used in our research when calculating average customer waiting time for GI/G/n queue.

We first present a general expression for $c_q$ which is applicable to G/M/n and M/G/1 queues. Then we form a conjecture that the expression provides a good approximation for GI/G/n queues and test this conjecture via Mont Carlo simulations. In the following, λ is the arrival rate, μ is the service rate, and $\rho$ is utilization.

For M/G/1queue, the variance of waiting time is $\sigma_q{}^2 = W_q{}^2 + \frac{\lambda E[s^3]}{3(1-\rho)}$ and the average waiting time is $W_q = \frac{\lambda E[s^2]}{2(1-\rho)}$ (Kleinrock 1976), where $E[s^2]$, $E[s^3]$ are the second and the third

moments of the service time distribution. For $M/G/1$, we know $P(T_q = 0) = 1 - P(T_q > 0) = 1 - \rho$. Therefore,

$$c_q = \frac{\sigma_q}{W_q} = \sqrt{1 + \frac{\lambda E[s^3]}{3(1-\rho)W_q^2}} = \sqrt{1 + \frac{4E[s^3]}{3\lambda}\frac{P(T_q = 0))}{(E[s^2])^2}} \qquad (2)$$

We conjecture that formula (2) can be used as an approximation for the $GI/G/n$ queue since it applies to both $G/M/n$ and $M/G/1$. Whitt (1993) conjectured that the exact formula for the distribution of waiting times of $M/G/1$ can be used as an approximation for the $M/G/n$ model. Seelan and Tijms (1984) provided additional support for this approximation.

## 3. Priority Queues

In the above models, all the models considered have the property of a first come first served discipline. This is not the only manner of service, and there are many alternatives, such as last come, first served, selection in random order, and selection by priority. A very considerable portion of real life queuing situations contain priority considerations. Some companies may decide to divide their customers into different priority queue classes to receive services according to their required service time and the price they are willing to pay. For example, customers in a high priority queue class would receive their goods or services immediately while a lower priority queue class may accept a delay in return for a discounted price. It is very important to analyze how to manage this system as it is a complex task to provide goods and services to customers with different delivery or service waiting times and in different priority classes.

In priority schemes customers with the highest priorities are selected for services ahead of those with lower priorities, independent of their time of arrival into the system. Priority queues are generally more difficult to model than non-priority situations. The determination of stationary probabilities in a non-preemptive Markov system is an extremely difficult matter.

There are two further refinements possible in priority situations, namely, preemption and non-preemption. In preemptive cases, a customer with the highest priority is allowed to enter service immediately even if another with lower priority is already present in service when the higher customer arrives. That is the lower priority customer in service is preempted, his service stopped, to be resumed again after the higher priority customer is served. In addition, a decision has to be made whether to continue the preempted customer's service from the point of preemption when resumed or to start anew. On the other hand, a priority discipline is defined to be non-preemptive if there is no interruption and the highest-priority customer just goes to the head of the queue to wait its turn. The customer can't get into service until the customer presently in services is completed, even though this customer has a lower priority. The non-preemptive approach is preferred for most call center priority applications because it best preserves the invisibility aspect of the prioritization process. It is also preferred in most applications where immediate service is not the sole reason for the priority and where the line behavior is observed by all customers because it is perceived as being fairer than the preemptive approach.

Queuing systems with customer priorities and queuing systems with customer transfers have wide applications in manufacturing, computer networks, telecommunication systems, and vehicle traffic control. The study of such queuing systems is extensive. Existing works address issues related to system stability, optimal scheduling, routing, and

performance analysis. For example, some of the existing works focus on system stability conditions, some on the stationary analysis of the queue length(s) and waiting times, and some on customer transfer strategies. The queuing model of interest is also related to, but not included in, stochastic transfer networks. In addition, the literatures focus on a product-form solution, rather than stability conditions.

The queuing model of interest consists of s identical servers serving N types of customers: type 1, type 2. . .and type Customers. Type 1, 2…and N customers form queue 1, 2…and N, respectively. Type N customers have the highest service priority, type N-1 the second highest service priority . . . and type 1 the lowest service priority. When a server is available, it chooses a customer from the non-empty queue of the highest priority and begins to serve it. If some servers are serving type j customers when a type k customer arrives, for j < k, there is no idle server, and type j customers are the lowest priority customers in service, then one of the type j customers in service is pushed back to queue j and the server begins to serve the type k customer. The type j customer will resume or repeat its service if a server is available to serve type j customers.

Type 1, 2, and N customers arrive according to independent Poisson processes with parameters $\lambda_1, \lambda_2$ …and $\lambda_N$ , respectively. The service times of type 1, 2… and N customers are exponentially distributed with parameters $\mu_1, \mu_2$ …and $\mu_N$ ,respectively. The arrival processes and service times are independent. Since the service time of a type j customer is exponentially distributed, it does not make a difference to assume that its interrupted service, if it occurs, will be repeated or resumed. For the same reason, if a server is available to serve type j customers, it does not matter (to system stability/instability) which waiting type customer enters the server to receive service?

The number of priority classes can be any number greater than one, and if there can be more than a single customer in any given priority class in the system simultaneously, then the discipline of selecting customers within the same priority class must also be specified.

In this research, we focus on the non-preemptive $GI/G/n$ system with many priorities. Within each priority class the FIFO discipline holds. Due to the difficulty of the determination of stationary priorities of $GI/G/n$, and the difficulty of handling multi-index generating functions when there are more than two priority classes, we use the similar approximation method analogous to the $M/M/n$ priority queue.

For non-preemptive Markov systems with many priorities, we use the results of Gross and Harris (2002) to derive the formula we used in our spreadsheet. Suppose that the items of the $kth$ priority (the smaller the number, the higher the priority) arrive before a single channel according to a Poisson distribution with parameter $\lambda_k (k = 1,2,...r)$ and that these customers wait on a FIFO basis within their respective priorities. Let the service distribution for the $kth$ priority be exponential with mean $1/\mu_k$ .Whatever the priority of a unit in service, it completes its service before another item is admitted. We begin by defining

$$\rho_k = \frac{\lambda_k}{\mu_k} \qquad (1 \le k \le r) \text{ and } \sigma_k = \sum_{i=1}^{k} \rho_i \qquad (\sigma_0 \equiv 0, \sigma_r \equiv \rho)$$

The system is stationary for $\sigma_r = \rho = \sum_{k=1}^{r} \rho_k < 1$. we have $W_q^{(i)} = \dfrac{\sum_{k=1}^{r} (\rho_k / \mu_k)}{(1-\sigma_{i-1})(1-\sigma_i)}$.

The analysis for the multiple-channel case is very similar to that of the proceeding model except that it must now be assumed that service is governed by identical exponential distributions for each priority at each of s channels. For multiple channels we must assume no service time distinction between priorities or else the mathematics becomes quite intractable.

Define $\rho_k = \frac{\lambda_k}{s\mu_k}$ $(1 \le k \le r)$ and $\sigma_k = \sum_{i=1}^{k} \rho_i$ $(\sigma_r \equiv \rho = \lambda / c\mu)$

Again the system is completely stationary for $\sigma_r = \rho = \sum_{k=1}^{r} \rho_k < 1$.

$$W_q^{(i)} = \frac{E[S_0]}{(1-\sigma_{i-1})(1-\sigma_i)} = \frac{\left[ s!(1-\rho)(s\mu)\sum_{n=0}^{s-1}(s\rho)^{n-s}/n! + s\mu \right]^{-1}}{(1-\sigma_{i-1})(1-\sigma_i)}$$

and the expected waiting time taken over all priorities is thus $W_q = \sum_{i-1}^{r} \frac{\lambda_i}{\lambda} W_q^{(i)}$.

$$Fract1 = \frac{\lambda_1}{\lambda}, \quad Fract2 = \frac{\lambda_2}{\lambda}, \quad Fract3 = \frac{\lambda_3}{\lambda}, \quad Fract4 = \frac{\lambda_4}{\lambda}$$

$\lambda = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$

In spreadsheet, we use

$$L_{q1} = \frac{L_q \cdot Fract1 \cdot (1-\rho)}{(1 - Fract1 \cdot \rho)}$$

$$L_{q2} = \frac{L_q \cdot Fract2 \cdot (1-\rho)}{(1 - Fract1 \cdot \rho) \cdot (1 - Fract1 \cdot \rho - Fract2 \cdot \rho)}$$

$$L_{q3} = \frac{L_q \cdot Fract3 \cdot (1-\rho)}{(1 - Fract1 \cdot \rho - Fract2 \cdot \rho) \cdot (1 - Fract1 \cdot \rho - Fract2 \cdot \rho - Fract3 \cdot \rho)}$$

$$L_{q4} = \frac{L_q \cdot Fract4 \cdot (1-\rho)}{(1 - Fract1 \cdot \rho - Fract2 \cdot \rho - Fract3 \cdot \rho) \cdot (1 - Fract1 \cdot \rho - Fract2 \cdot \rho - Fract3 \cdot \rho - Fract4 \cdot \rho)}.$$

For multi-servers, $n \neq 1$

$A = s!(\frac{s\mu - \lambda}{\rho^s})\sum_{j=0}^{s-1} \frac{\rho^j}{j!} + s\mu$ (Note here $\rho = \frac{\lambda}{\mu}$)

$B_0 = 1;\ B_k = 1 - \frac{\sum_{i=1}^{k} \lambda_i}{s\mu}.$ *for* $k = 1, 2, \ldots\ldots N$

We use the same reasoning: $L_q = [\frac{(\lambda/\mu)^s \lambda\mu}{(s-1)!(s\mu - \lambda)^2}]P_0$

$$P_0 = \left[ \sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \frac{s\mu}{(s\mu - \lambda)} \right]^{-1}$$

By Little's rule, we can have $W_{q1}, W_{q2}, W_{q3}$, *and* $W_{q4}$ *etc.*

For the $GI/G/n$ priority queue, we use the similar approximate method analogous to the $M/M/n$ priority queue. We conjecture that the mean waiting time for each priority class has the similar relations of those of the $M/M/n$ priority queue. The above formulas are used in our spreadsheet to calculate average flow times in non-preemptive priority queues.

## 4. Simulation and Numerical Comparisons

To evaluate the accuracy of the approximations, we conduct simulation experiments using the ExtendSim simulation program. The testing of our approximations has been based on extensive simulation experiments. In this simulation research, we performed independent replications using 54000 minutes of simulation time and estimated 95 % confidence intervals. Both Weibull and Gamma distributions are used as general distribution. For Gamma distribution, when shape parameter k is positive integer, Gamma is reduced to Erlang. When k=1, it is exponential. When k$\rightarrow\infty$, it is deterministic.

Simulation experiments confirm that the approximations perform remarkably well across a wide range of cases. In most of these cases the standard deviation of the time in the system obtained with the spreadsheet was within 10% of that obtained in the simulation. The limitation is that our result is under the assumption that the coefficients of variation of the inter-arrival times and the service times are between 0 and 1.25, which is usual in practice. When coefficients of variation are greater than1.5, the performance of the queue itself becomes very unstable. As noted by Whitt (1993), greater variability means less reliable approximation, because such descriptions evidently depend more critically on the missing information.

We present a representative set of tables comparing the approximations with exact (simulation) values. There are two standard ways to measure the quality of queuing approximations: absolute difference and relative percentage error (Whitt 1993). We contend that neither procedure alone is usually suitable over the entire range of values. We can obtain satisfactory results if either the absolute difference is below a critical threshold or the relative percentage error is below another critical threshold. Thus, a final adjusted measure of error (AME) might be:

$$Error = \min\{|exact - approx.|,\ 100(|exact - approx.|)/exact.\}.$$

Either the relative percentage error or the absolute difference should be small. Here we have simulation results corresponding to different experiments. These tables display expected mean and standard deviation of cycle time in specific queuing systems. The difference and relative error analysis are displayed in a separate spreadsheet. For those cases with both $c_a, c_s \le 1.25$, the approximations appear to be remarkably accurate.

Simulation experiments confirm that the approximations perform remarkably well across a wide range of cases. In most of these cases the standard deviation of the time in the system obtained with the spreadsheet was within 10% of that obtained in the simulation. The limitation is that our result is under the assumption that the coefficients of variation of the inter-arrival times and the service times are between 0 and 1.25, which is usual in practice. When coefficients of variation are greater than1.5, the performance of the queue itself becomes very unstable. As noted by Whitt (1993), greater variability means less reliable approximation, because such descriptions evidently depend more critically on the missing information.

The tests show that the standard deviation does not change dramatically when Weibull distributions are used instead of gamma distribution. This suggests that the standard

deviation tends to be insensitive to "reasonable" changes in the distribution assumptions, and hence our approximation will work well for these different distributional assumptions. The approximation tends to be closer to the simulation results obtained with gamma distributions than with Weibull.

## 4. Conclusions

This research provides a mathematically exact expression for the standard deviation of waiting time for Markov queues. It then applies this expression to give a two-moment approximation to the standard deviation of waiting time for a general queue with infinite waiting capacity. With quantitative results, this paper has presented an analytical approach to estimate the sizes of the time buffers in lean supply chain operations. The measurement requires only the mean and standard deviation or the coefficient of variation of the inter-arrival and service time distributions, and the number of servers. The quality of the approximations is not the same for all cases, but in comparisons to Monte Carlo simulations has proven to give good approximations. A significant feature of the approximation methods is that it is mathematically intractable and can be implemented in a spreadsheet format.

It is clear that the scheduling goal is to minimize the value for each measure. Because priority rules do not all affect performance measure to the same degree or the same manner, a manager should select a rule that best address the performance measure that is most important for his business. If all jobs must go through the same sequence of steps or operations, the queuing analysis models discussed above can be used to determine which scheduling priority rule provides the lowest performance measure values for a particular business.

We observed that using priorities increases the variability of waiting times: the higher the percentages of customers getting preferential treatment, the higher the variability. Because variability adds uncertainty to business outcomes, using priority rules in processing waiting line customers should be carefully considered. If used, it should be limited to only a small percentage of the arrival population. Some models have been developed to determine the increased variability in average waiting time when using both non-preemptive and preemptive priorities. (Hillier and Lieberman 2010, Haussman 1970). These models also aid in determining the degree of reduction in the average waiting time for higher priority customers and the concomitant increase in waiting time for lower priority customers.

## References

Allon, G. and Federgruen, A. ,2008. Service competition with general queuing facilities. *Operations Research*, 56(4), 827-849.

Alotaibi, Y. and Liu, F., 2013. Average waiting time of customers in a new queue system with
different classes.*Business Process Management Journal,* 19 (1), 146-168.

Atkinson, J.B. ,2009. Two new heuristics for the GI/G/nqueuing loss system with examples based on the two-phase distribution.*Journal of Operations Research Society*, 60(6), 818-830.

Babad, Y. Dada ,M.,and Saharia, A. ,1996. An appointment-based service center with guaranteed service.*European Journal of Operational Research*, 89(2),246-258.

Bertsimas,D. ,1990. An analytic approach to a general class of G/G/s queuing systems. *Operations Research*, 38(1),139-149.

Davis, M. and Heineke, J., 1998.How disconfirmation, perception and actual waiting times

impact customer satisfaction.*International Journal of Service Industry Management,* 9(1),64-73.

Gross, D. and Harris, C. M. ,2002.*Fundamentals of Queuing Theory*, John Wiley &Sons, New York.

He, Q.M, Xie, J, and Zhao, X.B. ,2012. Priority Queue with Customer Upgrades.*Naval Research Logistics*, 59(5),362–375.

Hopp, W.J. and Spearman, M.L. ,2000.*Factory Physics*, Irwin/McGraw-Hill, New York.

Jaiswell, N.K., 1968. *Priority Queues*, Academic Press, New York.

Kimura, T. ,1986. A two-moment approximation for the mean waiting time in the GI/G/s Queue. *Management Science*, 32(6),751-763.

Kleinrock, L. ,1976.*Queuing Systems, Volume I & II: Theory,* John Wiley & Sons, New York.

Kumar, P.,Kalwani, M. and Dada, M. ,1997. The impact of waiting time guarantees on Customers' waiting experiences.*Marketing Science*, 16(4), 295-314.

Ma, J. , Wang, K.& Xu, L., 2011.Modelling and analysis of workflow for lean supply chains,*Enterprise Information Systems*, 5(4),423–447.

Peter, J. and Peppiatt, E. ,1996. Managing perceptions of waiting times in service queues. *International Journal of Service Industry Management*, 7(5).47-61.

Sakasegawa, H.,1977. An approximate formula Lq = αβ ρ / (1 – ρ).*Annals of the Institute of Statistical Mathematics,*29(a), 67-75.

Seelen, L.P. and TIJMS, H.C., van Hoorn, M.H.,1984. Approximations for the conditional waiting times in GI/G/c queue.*Operations Research Letters,*3,183- 190.

Shore, H. (1988), "Simple approximations for the GI/G/c queue-I: The steady-state Probabilities", *The Journal of the Operational Research Society*, 39(3),279- 284.

So, K.C. and Song, J.S. ,1998. Price, delivery time guarantees and capacity selection. *European Journal of Operational Research,* 111(1), 28-49.

Suresh ,S. and Whitt, W. ,1990. Arranging queues in series: simulation experiments. *Management Science,* 36(9), 1080-1091.

Taylor, S., 1994. Waiting for service: the relationship between delays and the evaluation of Service, *Journal of Marketing*,58(2),56-69.

Whitt, W.,1993. Approximations for the GI/G/m queue.*Production and Operations Management*, 2(2), 114-161.

Whitt, W.,2004. A diffusion approximation for the GI/G/n/m queue.*Operations Research*, 52 (6), 922-941.

Zhang, H., Shi, D.,2010. Explicit Solution for M/M/1 Preemptive Priority Queue. *International Journal of Information and Management Sciences*, 21,197-208.