DECISION SCIENCES INSTITUTE

Service Delivery Delay in Multi-Class Service Processes

Xiaofeng Zhao
University of Mary Washington
Email:xzhao@umw.edu

ABSTRACT

This paper measures service delay time in multi-class service processes based on a non-preemptive priority queue model. It derives tractable expressions for Markov queues and uses the concept of isomorphism to approximate the mean and standard deviation of delay time for the general queue. The numerical results are reported with simulation experiments to demonstrate the accuracy of the new method. The approximation methods are mathematically tractable to conduct what-if analysis. This practice-based research can be efficiently and effectively used for managers to estimate the delay time for different customer classes.

KEYWORDS: queuing theory, priority queue, service delivery, stochastic processes

INTRODUCTION

The speed of service is increasingly becoming a critical service attribute (Larson 1987; Shang and Liu 2011) as waiting lines and delays have become commonplace in service operations. The waiting experiences are typically considered to have a negative influence on customers' overall satisfaction with the products and services (Kumar 1997, Davis 1998), so the importance of proper management of customer waiting time is of significant interest to most organizations. Managers have made efforts not only to reduce actual waiting times but also to improve customer satisfaction with a given waiting time. Many companies have used waiting time guarantee strategy to compete in service industries. The waiting time guarantee strategy requires managers to know both the average of waiting time and the variance around average waiting time.

This research develops a math attractable expression for the mean and variance of waiting time for the M/M/n queue and approximates the customer waiting time for multi-class, multi-server queueing system with non-preemptive priorities. The approximation also has the desirable properties of being exact for the specific case of Markov queue models. The models can be easily applied to service operations to calculate the variance of customer waiting time. In comparison with simulations, the method gives accurate approximations.

A significant feature of the approximation method is that it is mathematically tractable and can be implemented in a spreadsheet format. The results demonstrate the queuing model application in service operations. Another feature of our models is that managers can conduct what-if analysis and select appropriate capacity levels to commit themselves to a given waiting time guarantee. Hopefully, the approximation benefits practitioners in providing simple, quick, and practical answers to their multi-class, multi-server queuing systems.

The paper is organized as follows. Section 2 provides a review of the relevant literature. Section 3 develops an exact expression for the coefficient of variation of waiting time for Markov queues and presents an approximation for the variance of waiting time for the general queue. Section 4

develops priority queue algorithms for different customer classes. Section 5 reports numerical results. Section 6 concludes the paper.

LITERATURE REVIEW

Customer waiting time has been studied in the domain of queuing theory. Queues occur because of uncertainty in supply and demand. A waiting line forms whenever the demand for service exceeds the ability to provide service. A queue can be classified according to the number of servers and the distributions that characterize the arrival rates of customers and the service rates. Priorities are often applied in queueing systems as a mechanism for service or resource allocation. Selecting appropriate rules of priorities can greatly improve system performances, such as improvements in throughput, cycle time, utilization, and the balance of flow.

Queueing models with priority discipline have been studied extensively, and most literature on multi-server priority queues deals with complex numerical methods, and the results are intractable. For instance, Gail et al. (1988) studied the M/M/n non-preemptive queues with two classes of jobs. Kao and Narayanan (1990) used the matrix-geometric approach in conjunction with the state reduction method to develop a computationally efficient procedure for solving the model. Sleptchenko et al. (2003) derived a method in analyzing multi-class M/M/n priority queues with partial blocking. Each group with different priority can contain several classes of items with different arrival and service rates. The proposed method is based on the solution of the stationary state equations. Tabet-Aouel and Kouvatsos (1992) proposed an approximation to the mean response times of priority classes in a stable G/G/n queue under pre-emptive resume scheduling.

Takine (1996) considered a non-preemptive priority queue with two classes of customers. Customers in each priority class arrive at the system according to a Markov distribution. Since the Markov arrival is weakly dense in the class of stationary point processes, it is a fairly general arrival process. The service times of customers in each priority class are i.i.d.in a general distribution which may differ among two priority classes. He et al. (2012) discussed a general multi-class priority queueing system with customer priority upgrades. They assume that the customer arrival processes are marked Markovian arrival processes and service times are of general distribution. Recently, Jouini and Roubos (2014) analyzed Markovian multi-server queues with two types of impatient customers: high and low-priority ones.

In the literature there are some high-quality approximations for a single class in the GI/G/n queue (Kimura 1986; Shore 1988; Whitt 1993). In Sphicas and Shimshak (1978), Whitt (1993) and Zhao et al. (2014) developed approximation models to estimate the variance of waiting time, but only for single customer class. Defraeye and Van Nieuwenhuyse (2013) developed a single-stage multi-server M (t)/G/s (t) +G queue with time-varying arrivals and customer impatience. There exists no mathematically tractable formula for approximating the standard deviation of waiting time for multi-classes in the GI/G/n queue.

For practitioners, it's imperative to estimate the delays for each class when the service systems with different capacity provided. Computer simulation can be an alternative to mathematical models. But simulation has significant limitations as well. It is required to obtain correct simulation with enough samples. It is expensive for managers to develop and run simulations. Analytical models are relatively inexpensive and are useful for quick initial estimates (Louw and Page 2004). To the best of our knowledge, no spreadsheet model has dealt with both the

average and the variance of waiting time for different customer classes in GI/G/n queue. Therefore, in this practice-based research, we concentrate on mathematically explicit expressions of the mean and the standard deviation of waiting time for different classes.

APPROXIMATION MODELS FOR SINGLE CUSTOMER CLASS

The mean waiting time

To approximate the variance of waiting time in the GI/G/n queue, we need to develop a formula to approximate the coefficient of variation of waiting time $c_q = \sigma_q / W_q$. In the literature, there are some approximations available for the mean waiting time (Whitt 1993). For instance, Sakasegawa (1977) developed an attractable expression for the mean waiting time with high utilization in the GI/G/n queue:

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

This formula offers several advantages (Whitt 1993). It does not need iteration and can be easily implemented on the computer. The method is used in our research when calculating average customer waiting time for GI/G/n queue.

This approximation is exact for M/M/n queue. It has been applied in several commercial software packages (Hopp and Spearman 2007). The approximation (1) is a two-moment approximation for mean waiting time: it depends only on the first two moments of inter-arrival times and service times. Two-moment approximations for queuing characteristics such as GI/G/n usually have sufficient accuracy for practical purpose, since extreme cases which need higher moments tend not to arise (Kimura 1986, 1991).

The variance of waiting time

The approximation used in the single class queue for the variances of the waiting times in GI/G/m queue is developed by Whitt (1993).By definition, the variance of the expected waiting time in the queue $Var(W_q)$ is determined by $Var(W_q) = W_q(GI/G/n)^2 \cdot CV_w^2$ and

$$CV_w^2 = \frac{CV_d^2 + 1 - P(W > 0)}{P(W > 0)} \qquad (2)$$

Where $CV_d^2 = 2\rho - 1 + 4(1-\rho)d_s^3 / 3(CV_s^2 + 1)^2$

$$d_s^3 = \begin{cases} 3CV_s^2(1+CV_s^2), & CV_s^2 \geq 1 \\ (2CV_s^2 + 1)(CV_s^2 + 1) & CV_s^2 < 1 \end{cases}$$

APPROXIMATION MODELS FOR MULTI-CUSTOMER CLASSES

Multi-class service process with priority

The above models have the property of a first come first served discipline. In priority schemes, customers with the highest priorities are selected for services ahead of those with lower

priorities. Priority queues are more difficult to model than non-priority situations. The determination of stationary probabilities in a non-preemptive Markov system is extremely difficult.

There are two further refinements in priority situations, namely, preemption and non-preemption. In preemptive cases, a customer with the highest priority is allowed to obtain service immediately even if another with lower priority is already accepted in service when the higher customer arrives. That means the lower priority customer in service is preempted to be resumed again after the higher priority customer is served. A priority discipline is defined to be non-preemptive if there is no interruption and the highest-priority customer just goes to the front of the waiting line. 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.

The queuing model of interest consists of s identical servers serving N types of customers: type 1, type 2. . and type N customers. Type 1, 2…and N customers from queue 1, 2…and N, respectively. Type N customers have the highest service priority, type N-1 the second highest priority . . . and type 1 the lowest priority. When the service is available, a customer is chosen from the non-empty queue of the highest priority.

Type 1,2,… and N customers arrive according to independent Poisson processes with parameters λ_1, λ_2 …and λ_N , respectively. The service times of type 1, 2… and N customers are exponentially distributed with parameters μ_1, μ_2 …and μ_N , respectively. The arrival processes and service times are independent. Since the service time of a type j customer is exponentially distributed, there is no need to assume that its interrupted service, if it occurs, will be repeated. For the same reason, if a server is available to serve type j customers, it makes no difference (to system stability/instability) which waiting type j customer enters the server.

The number of priority classes can be any number greater than one. If there can be more than a single customer in any given priority class in the system simultaneously, then we must specify the discipline of selecting customers within the same priority class. 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 $GI/G/n$, 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.

The mean waiting time of class *i* customer

For non-preemptive Markov systems with many priorities, 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

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

The system is stationary for $\sigma_r = \rho = \sum_{k=1}^{r} \rho_k < 1$. we have $W_q^{(i)} = \frac{\sum_{k=1}^{r} (\rho_k / \mu_k)}{(1-\sigma_{i-1})(1-\sigma_i)}$. (Gross and Harris (2002) 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 $W_q = \sum_{i-1}^{r} \frac{\lambda_i}{\lambda} W_q^{(i)}$.

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

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

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

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

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

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

For multi-servers, $n \ne 1$

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

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

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*.

The variance of waiting time of class *i* customer

For each customer class, by definition, the variance of the mean waiting time is determined by $Var(W_{q_i}) = W_{qi}(GI/G/n)^2 \cdot CV_{w_i}^2$. Using the concept of isomorphism, we make the similar conjecture that (2) holds for estimating the standard deviation of waiting time for each customer class. To calculate the approximations for the coefficient of waiting time, we adopt the results used by Whitt (1993),

$$CV_{wi}^2 = \frac{CV_{di}^2 + 1 - P(W_i > 0)}{P(W_i > 0)}$$

$$CV_{di}^2 = 2\rho_i - 1 + 4(1-\rho_i)d_{si}^3 / 3(CV_{si}^2 + 1)^2$$

$$d_{si}^3 = \begin{cases} 3CV_{si}^2(1+CV_{si}^2), & CV_{si}^2 \geq 1 \\ (2CV_{si}^2 + 1)(CV_{si}^2 + 1) & CV_{si}^2 < 1 \end{cases}$$

$P(W_i > 0) \approx \min\{\pi, 1\}$ where

$$\pi = \left\{ \begin{array}{l} \pi_1, \text{ if } m_i \leq 6 \quad or \quad r_i \leq 0.5 \quad or \quad C_{ai}^2 \geq 1 \\ \pi_2, \text{if } m_i \geq 7 \quad and \quad r_i \geq 1 \quad and \quad C_{ai}^2 < 1 \\ \pi_3, \text{if } m_i > 7 \quad and \quad 0.5 \leq r_i \leq 1 \; and \; C_{ai}^2 < 1 \end{array} \right\}$$

$$r_i = (m_i - m_i\rho_i - 0.5) / \sqrt{m_i\rho_i z_i}$$

$$z_i = (CV_{ai}^2 + 1)/2$$

$$\pi_1 = \rho_i^2\pi_4 + (1-\rho_i^2)\pi_5$$

$$\pi_2 = CV_{ai}^2\pi_1 + (1-CV_{ai}^2)\pi_6$$

$$\pi_3 = 2(1-CV_{ai}^2)(r_i - 0.5)\pi_2 + (1 - \left[2(1-CV_{ai}^2)(r_i - 0.5)\pi_2\right])\pi_1$$

$$\pi_4 = \min\left\{1, \frac{1-\Phi((1+CV_{si}^2)(1-\rho_i)m_i^{1/2}/(CV_{ai}^2 + CV_{si}^2)}{1-\Phi(((1-\rho_i)m_i^{1/2})} P(W_i(M/M/n) > 0)\right\}$$

$$\pi_5 = \min\left\{1, \frac{1-\Phi(2(1-\rho_i)m_i^{1/2}/(1+CV_{ai}^2)}{1-\Phi(((1-\rho_i)m_i^{1/2})} P(W_i(M/M/n) > 0)\right\}$$

$$\pi_6 = 1 - \Phi((m_i - m_i\rho_i - 0.5)/\sqrt{m_i\rho_i z_i}$$

Φ is the standard normal PDF. The above approximations of the mean and the standard deviation of waiting time reduce to exact M/M/n priority formulas when the coefficients of variance are 1.

NUMERICAL ANALYSIS AND DISCUSSIONS

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 and estimated 95 % confidence intervals. For Gamma distribution, when shape parameter k is a

positive integer, Gamma is reduced to Erlang. When k=1, it is exponential. When k→∞, 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 for the entire range of values. We can obtain satisfactory results if either the absolute difference is below a defined threshold or the relative percentage error is below another specific threshold. Thus, a final adjusted measure of error (AME) is:

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

Either the relative percentage error or the absolute difference should be within the accepted threshold. 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 present accurate prediction.

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 input measurement requires the mean and standard deviation 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 with simulation results, has proven to give a good estimation. 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 management goal is to minimize the waiting times. 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 addresses the performance measure that is most important for his business. If all customers must go through the same sequence of service 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 analyzed. If applied, it should be limited to only a small percentage of the arrival populations. Some models have been developed to determine the increased variability in average waiting time when using both non-preemptive and preemptive priorities. 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 the lower priority customers.

REFERENECS

Allon, G and Federgruen, A (2008). Service competition with general queuing facilities. *Operations Research,* 56(4), 827-849.
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.
Gail, H.R., Hantler, S.L and Taylor, B.A (1988). Analysis of a non-preemptive priority multi-server queue. *Advanced in Applied Probability,* 20(4), 852-879.
Gross, D and Harris, C.M. (2002). *Fundamentals of Queuing Theory*. NY, John Wiley &Sons.
Hanbali, A.L, Alvarez, E.M and van der Heijden, M.C(2015). Approximations for the waiting-time distribution in an M/PH/c priority queue. *OR Spectrum,* 37(2), 529-552.
Hillier, F.S and Lieberman, G.J (1986). *Introduction to Operations Research*. Oakland, CA.
Hopp, W.J and Spearman, M.L (2000). *Factory Physics*. Irwin/McGraw-Hill: New York.
Jouini, O., & Roubos, A. (2013). On multiple priority multi-server queues with impatience. *Journal of the Operational Research Society,* 65(5), 616-632.
Kimura, T (1991). Approximating the mean waiting time in the GI/G/s queue. *Journal of the Operational Research Society,* 42(1), 959-970.
Kleinrock, L (1976). *Queuing Systems, Volume I & II: Theory.* New York: John Wiley & Sons.
Kumar, P, Kalwani, M and Dada, M (1997).The impact of waiting time guarantees on customers' waiting experiences. *Marketing Science,* 16(4), 295-314.
Louw, L and Page, D.C (2004). Queuing network analysis approach for estimating the sizes of the time buffers in Theory of Constraints-controlled production systems. *International Journal of Production Research,* 42(6), 1207-1226.
Sakasegawa, H (1977). An approximate formula Lq = αβρ/(1 – ρ). *Annals of the Institute of Statistical Mathematic,* 29(a), 67-75.
So, K.C (2000). Price and time competition for service delivery. *Manufacturing & Service Operations Management,* 2(4), 392-409.
Sleptchenko, A, van Harten, A and van der Heijden, A (2003). An exact analysis of the multi-class M/M/k priority queue with partial blocking. *Stochastic Models,* 19(4), 527-548.
Whitt, W (2004). A diffusion approximation for the GI/G/n/m queue. *Operations Research,* 52(6), 922- 941.
Zhao, X, Hou, J and Gilbert, K. (2014). Measuring the variance of customer waiting time in service operations. *Management Decision, 52*(2), 296-312.