Performance Analysis of Queue Length Monitoring of M/G/1 Systems

Nan Chen,[1] Yuan Yuan,[2] Shiyu Zhou[2]

[1] *Department of Industrial and Systems Engineering, National University of Singapore, Singapore*

[2] *Department of Industrial and Systems Engineering, University of Wisconsin, Madison, Wisconsin*

Received 17 August 2010; revised 20 August 2011; accepted 7 September 2011
DOI 10.1002/nav.20483
Published online 18 October 2011 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: This study investigates the statistical process control application for monitoring queue length data in M/G/1 systems. Specifically, we studied the average run length (ARL) characteristics of two different control charts for detecting changes in system utilization. First, the nL chart monitors the sums of successive queue length samples by subgrouping individual observations with sample size n. Next is the individual chart with a warning zone whose control scheme is specified by two pairs of parameters, (upper control limit, d_u) and (lower control limit, d_l), as proposed by Bhat and Rao (*Oper Res* 20 (1972) 955–966). We will present approaches to calculate ARL for the two types of control charts using the Markov chain formulation and also investigate the effects of parameters of the control charts to provide useful design guidelines for better performance. Extensive numerical results are included for illustration. © 2011 Wiley Periodicals, Inc. Naval Research Logistics 58: 782–794, 2011

Keywords: average run length; Markov chain; M/G/1 queueing models; queue length; statistical monitoring

1. INTRODUCTION

Queueing models and queueing theory have been widely used in recent years to investigate the behavior of different performance measures in many practical systems, including manufacturing and production systems [19], computer systems and networks [16], tele-traffic systems [10], and healthcare systems [11]. The success of queueing models has enabled us to study operational performance, such as response time, queue length, or system throughput, systematically, which has provided us with opportunities for productivity improvement.

Although extensive research exists on queueing systems, relatively little attention has been given to the statistical monitoring of operational performance, an area which may help achieve more efficient operations. For example, by keeping track of the cycle time each individual item experiences in a production system, it is possible to detect changes in service rates or identify irregular patterns of customer arrivals. As another example, statistical monitoring of customer waiting times at a service call center may help to detect abnormal operational conditions and potentially provide guidelines for resource allocation. Despite its importance, several challenges remain when applying statistical monitoring methodologies directly to the monitoring of the operational performance. Specifically, (i) the observations from queueing systems are often autocorrelated or autocorrelated [13] and (ii) the distribution of the observations is typically unknown and is often highly skewed [18, 20]. These properties render the underlying assumptions of many standard statistical monitoring methods invalid. To overcome these difficulties, some researchers have devised new control charts to monitor queueing systems.

Bhat and Rao [6] first proposed a control scheme for M/G/1 and GI/M/1 systems based on observations of queue length data. The control scheme is specified by two pairs of parameters [upper control limit (UCL), d_u] and [lower control limit (LCL), d_l]. When the queue length is larger than UCL for d_u consecutive observations or smaller than LCL for d_l consecutive observations, the system is considered out of control (OC). Different from traditional control charts, UCL and LCL can be considered as forming a warning zone (WZ) as values beyond the limits do not signal OC alarms immediately. It is expected that, with a carefully designed WZ, the control chart would be more sensitive to system changes. Bhat [3] later developed a control chart using a sequential probability ratio test with partially observed queue length data: only the number of periods with an empty queue is used. Whereas this chart is useful when only this data is available, a more efficient scheme may be used when full observations of queue

Correspondence to: Shiyu Zhou (szhou@engr.wisc.edu)

$$\cdots \underbrace{X_{(n-1)k}, X_{(n-1)k+1}, X_{(n-1)k+2} \cdots X_{nk}}_{S_k}, \underbrace{X_{nk+1}, X_{nk+2} \cdots X_{(n+1)k}}_{S_{k+1}}, \underbrace{X_{(n+1)k+1}, X_{(n+2)k+2} \cdots X_{(n+2)k}}_{S_{k+2}} \cdots$$

Figure 1. Samples and sample sums in the *nL* chart.

lengths are available. Recently, Shore [20] proposed a modified attribute control chart using the first three moments of the queue length distribution to handle the skewness. By including the third moment when computing the control limits, it provides roughly equal sensitivity in detecting changes in both directions. Shore [20] applied the method to monitor the queue length in a G/G/S system where satisfactory results have been reported. Recently, Ben-Gal et al. [1, 2] proposed a context-based statistical process control technique to monitor the buffer levels in a production system. Their method may be applied to a general variable-length-state-dependent process.

Although many measures have been adopted to evaluate and design control charts, average run length (ARL) is more widely used in practice. ARL is defined as the expected number of samples to be taken before an OC alarm is signaled. Unlike the false alarm rate (α error) or the miss detection rate (β error) whose definitions are less meaningful when the data are autocorrelated, ARL may be used to compare control charts for both correlated data and independent data. Consequently, ARL has been frequently used in practice (e.g., Refs. 7, 14, 17, 22, 24) as the main criterion to design and evaluate control charts. Thus, accurate evaluation of ARL becomes very important. In general, ARL can only be estimated through replicated simulations which are often time consuming and not very accurate. In this article, we have developed a Markov chain model to compute the ARL of typical charts used in the monitoring of queueing systems without the help of simulation. This method enables us to efficiently compare the performance of different charts and provides us with the opportunity to design the control charts with specified ARL performances.

In particular, we would like to investigate two types of control charts for queue length monitoring in M/G/1 systems. First is the *nL* chart, a simple extension of the chart proposed by Shore [20]. The chart by Shore monitors an individual sample, whereas the *nL* chart monitors the sums of successive samples. The second control chart is the individual chart with the WZ proposed by Bhat and Rao [6]. The observations used in both charts are the queue lengths counted immediately after each departure. According to Gross et al. [12], these observations form a Markov chain whose states are the possible values of the queue length. We denote the observed queue length immediately after the kth departure by X_k. For the *nL* chart, nonoverlapping sample sums with the sample size n

$$S_k = \sum_{i=n(k-1)+1}^{nk} X_i, \quad k = 1, 2, \ldots \tag{1}$$

are monitored, as illustrated in Fig. 1.

If S_k is larger than UCL or smaller than LCL, the system is considered OC.

Another chart, proposed by Bhat and Rao [6], directly monitors the individual queue length observations X_k with warning limit parameters UCL and LCL. If either of the two conditions in (2) is met, it will generate OC signals.

$$\begin{aligned} &\min\left\{m \in N^+; X_{k+m} < \text{UCL} \,|X_k \geq \text{UCL}\right\} > d_{\text{u}} \quad \text{OR} \\ &\quad \min\left\{m \in N^+; X_{k+m} > \text{LCL} \,|X_k \leq \text{LCL}\right\} > d_{\text{l}}, \quad \forall k \end{aligned} \tag{2}$$

where N^+ is the set of positive integers.

We call this chart the warning zone control chart or WZ chart for short.

Using the developed Markov chain models discussed in the following sections, we are able to study the ARL performance of these two charts analytically. Based on their ARL calculations, we can further investigate the effects of the sample size in the *nL* chart and the WZ parameters in the WZ chart to provide design guidelines. To compare the performances of the two control charts, we set the design parameters in such a way that both control charts have the same ARL_0 (ARL in normal operation conditions). Then, the chart with the shorter ARL_1 (ARL in abnormal conditions) is considered better; shorter ARL_1 often indicates a quicker response to process shifts under some nonrestrictive assumptions [22]. It needs to be pointed out that different charts may use different sample sizes, for example, n in the *nL* chart and 1 in the WZ chart. Thus, for fair comparison, we need to use the average number of observations to signal (ANOS), which is defined as the ARL multiplied by the sample size for the comparison. Because of the simple correspondence between ARL and ANOS, we will discuss the calculations of the ARL in the following sections instead.

The remaining sections of this article are organized as follows. The second section gives a detailed description of the Markov chain approach used to calculate the ARL of the two types of control charts and shows possible extensions to other

charts as well. The third section compares ARL performances and investigates the effects of different design parameters. Finally, we conclude the article with a discussion on future research.

2. MARKOV CHAIN APPROACH FOR ARL CALCULATION

This section describes how to calculate the ARL of the *nL* chart and WZ chart for monitoring queue length data observed at each departure epoch in M/G/1 systems. For many control charts monitoring independent observations, the ARL has a simple correspondence with the α error and the β error. However, for complex control chart schemes (e.g., cumulative sum (CUSUM) charts and exponentially-weighted moving average (EWMA) charts) or autocorrelated observations, exact calculation of the ARL is much more involved. In the rest of Section 2, we will give a brief introduction to the general Markov chain approach that is commonly used to calculate the ARL in complicated situations. Following the introduction, we will demonstrate that this approach can also be tailored to the ARL calculations for the two control charts discussed in this article. For simplicity, we only consider one-sided control charts with the LCL set to zero. However, it should be noted that the methodology presented in this article can be easily extended to two-sided charts.

2.1. Introduction to the Markov Chain Approach

The Markov chain approach was first presented by Brook and Evans [8] to analyze the performance of conventional CUSUM charts. It has wide applications in calculating run length distributions or the ARL of a variety of control charts. For instance, Champ and Woodall [9] studied the run length properties of Shewhart charts with supplementary run rules using this method; Lucas and Saccucci [17] discussed the performance of EWMA charts and presented the ARL and distribution of run length for CUSUM schemes with a fast initial response feature; Shu and Jiang [21] extended the Markov chain approach to the adaptive CUSUM charts.

To use the Markov chain approach, first we need to formulate the in-control conditions and the OC conditions as states of a Markov chain. The OC conditions are often combined as an absorbing state. Next, the expected number of transitions from each in-control state to the first hit of the OC state is computed using Markov chain theory. The expected number of transitions from state i corresponds to the expected run length of the control chart with initial observation i. To illustrate, we consider a simple upper-sided control chart for individual queue length observations in an M/G/1 queue. Let X_k represent the queue length observed at the kth departure epoch, which forms a Markov chain according to Gross et al. [12]. When $X_k \leq \text{UCL}$, we say the system is in-control. Thus, we can formulate a Markov chain with state space defined as $\Omega = \{0, 1, \ldots \text{UCL}, \text{OC}\}$, where state i means the queue length is i when observed and OC denotes the OC state aggregated from all the OC queue lengths. By the operational convention of control charts, once the OC alarm is signaled, the current monitoring cycle stops and the process is considered to remain in the OC state until correction. As a result, from the point of view of the control chart operation, the OC state in the defined Markov chain is an absorbing state. Successive observations of the queue length X_k correspond to the transitions in the state space and the one-step transition matrix takes the form of Eq. (3), where $p_{ij} = \Pr\{X_{k+1} = j | X_k = i\}$:

$$
\mathbf{T} = \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2 & & \text{UCL} & \text{OC} & \text{States} \end{array} \\ \left[\begin{array}{ccccccc} p_{00} & p_{01} & p_{02} & \cdots & p_{0,\text{UCL}} & 1-\sum_{j=0}^{\text{UCL}} p_{0j} & 0 \\ p_{10} & p_{11} & p_{12} & \cdots & p_{1,\text{UCL}} & 1-\sum_{j=0}^{\text{UCL}} p_{1j} & 1 \\ p_{20} & p_{21} & p_{22} & \cdots & p_{3,\text{UCL}} & 1-\sum_{j=0}^{\text{UCL}} p_{2j} & 2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ p_{\text{UCL},0} & p_{\text{UCL},1} & p_{\text{UCL},2} & \cdots & p_{\text{UCL,UCL}} & 1-\sum_{j=0}^{\text{UCL}} p_{\text{UCL},j} & \text{UCL} \\ 0 & 0 & 0 & \cdots & 0 & 1 & \text{OC} \end{array}\right] \end{array} . \tag{3}
$$

Using r_i to denote the expected number of transitions to the OC state starting from state i, Brook and Evans [8] show that the linear equations

$$r_i = 1 + \sum_{j=0}^{\mathrm{UCL}} T_{ij} \cdot r_j \qquad \forall i = 0, 1, \ldots, \mathrm{UCL}, \tag{4}$$

are valid and can be solved efficiently. When considering the expected run length across all possible initial states, the ARL to the absorbing state OC can be expressed as:

$$\mathrm{ARL} = \sum_{i=0}^{\mathrm{UCL}} \pi_i \cdot r_i + 1, \tag{5}$$

where π_i refers to the steady state probability of being in state i. As the first sample is counted in the run length but not in the number of transitions, we need to add one to the expected transitions to get the correct ARL, as shown in Eq. (5). We would like to point out that Eq. (4) is equivalent to the approach using the "fundamental matrix" as illustrated in Refs. 5, 15, 21. In fact, solving Eq. (4) is essentially the same as obtaining the fundamental matrix $(\mathbf{I} - \mathbf{T}')^{-1}$, where $\mathbf{T}'$ is the upper $(\mathrm{UCL} + 1) \times (\mathrm{UCL} + 1)$ submatrix of $\mathbf{T}$.

As shown by the above description, two basic elements are essential to this approach: Markov chain formulation and its transition matrix $\mathbf{T}$, as well as the steady state probability π_i of each in-control state. We would like to emphasize that the ARL defined in Eq. (5) may be viewed as the weighted average of r_i, $i = 0, 1, \ldots, \mathrm{UCL}$, where r_i is the ARL when monitoring the system from a specific queue length i. Therefore, if we are interested in the ARL starting from certain initial states (e.g., r_0 as in Ref. 6), it is not necessary to obtain the steady state distribution π_i. In the next section, the general Markov chain approach will be tailored to compute the ARL of the *nL* chart and the WZ chart for the monitoring of M/G/1 systems. First, we will briefly introduce some relevant properties of M/G/1 systems.

Assume that a M/G/1 queue is observed at departure epochs $t_1, t_2, \ldots$, and $X_k = X(t_k)$ denotes the number of customers in the system at time t_k. Therefore, the maximum decline from X_k to X_{k+1} is one, whereas the increment could be more than one if there are multiple arrivals between two successive departures. Let A be the random variable representing the number of customers arriving during a service period, and define

$$\begin{aligned} a_i &= \Pr\{i \text{ arrivals during a service time}\} = \Pr\{A = i\} \\ &= \int_0^\infty \frac{e^{-\lambda t}(\lambda t)^i}{i!} dG(t), \end{aligned} \tag{6}$$

where $G(t)$ is the cumulative distribution function of the service time and λ is the arrival rate. Denote the transition probability from state i to j by p_{ij}. Then, according to the relationship between transition probabilities and a_i, p_{ij} can be expressed as:

$$\begin{aligned} p_{ij} &= \Pr\{X_{n+1} = j | X_n = i\} \\ &= \begin{cases} \Pr\{A = j - i + 1\} = a_{j-i+1} & (i \geq 1) \\ \Pr\{A = j\} = a_j & (i = 0) \end{cases}. \end{aligned} \tag{7}$$

Furthermore, according to the Poisson-arrivals-see-time-averages property of the queues with Poisson arrivals [10], the steady state distribution of queue length observed at departure epochs equals the steady state distribution observed at any time. Particularly in M/G/1 queues, a general formula can be expressed as [4]:

$$\begin{aligned} \pi_0 &= 1 - \rho \\ \pi_j &= (1 - \rho) \int_0^\infty e^{-\lambda t} \sum_{l=0}^{\infty} \left[\frac{(\lambda t)^{l+j-1}}{(l + j - 1)!} - \frac{(\lambda t)^{l+j}}{(l + j)!} \right] dG_l(t), \end{aligned} \tag{8}$$

where $G_l(t)$ is the l-fold convolution of the service time distribution $G(t)$. When $G_l(t)$ is known, the corresponding steady state distribution can be calculated. Specifically, for M/M/1 queues with utilization rate ρ, the steady state distribution is:

$$\pi_j = (1 - \rho)\rho^j, \quad j = 0, 1, 2, \ldots. \tag{9}$$

Except for some special cases, the computation of l-fold convolution is generally expensive. Readers may either refer to existing works (e.g., Refs. [15–17]) for efficient computation algorithms or use Monte Carlo simulation methods to obtain an approximation of π_j. In Appendix, we have included the method to compute π_j for queues with Erlang-k distributed service times that is commonly used in practice. It should be pointed out that if we only consider the ARL starting from an empty system or any particular initial states, then we do not need to compute the steady state distribution π_j. Instead, r_j in Eq. (5) is sufficient to be used as the performance measure in such circumstances.

2.2. Calculation of the ARL of *nL* Charts

The *nL* chart monitors sample sums by grouping successive individual observations into samples of size n (see Fig. 1). When S_k is larger than UCL, the process is signaled to be out of control. Regrettably, $\{S_k\}$ itself is not a Markov process and this violates the basic conditions for using the Markov chain approach for ARL calculation. To solve this problem, we redefine the states as (X_{kn}, Q_k) where X_{kn} is the last element in the kth sample and Q_k is a variable that checks sample sums:

$$Q_k = \begin{cases} 1 & \text{if } S_k \leq \mathrm{UCL} \\ 0 & \text{if } S_k > \mathrm{UCL} \end{cases} \tag{10}$$

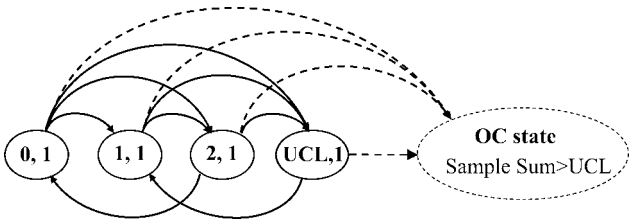

Figure 2. Transitions between states in the *nL* chart (transition probabilities are not shown).

It is shown that (X_{kn}, Q_k) forms a Markov chain with $(X_{kn}, 0)$ as the absorbing state indicating the OC condition. The Markovian property of (X_{kn}, Q_k) is proved in Appendix. The state space becomes $\Omega = \{(0,1),(1,1),(2,1),\ldots(\mathrm{UCL},1),\mathrm{OC}\}$, in which OC incorporates all states where $Q_k = 0$. Figure 2 illustrates the transitions among states in Ω, where states are represented by ellipses with the values of X_{kn} and Q_k in the center. In the figure, solid arrows are the transitions between in-control states and the dashed arrows represent the transitions from in-control states to the OC state. Once the process jumps into the OC state, it will remain in the OC state.

From the formulation, we find that the transition probability depends on the choice of sample size and UCL. We denote $V_{ij}(n, UCL)$ as the transition probability among in-control states when sample size equals n and the upper control limit is UCL. We then have

$$\begin{aligned}
&V_{ij}(n,\mathrm{UCL})\\
&\quad=\Pr\{X_{(k+1)n}=j, Q_{k+1}=1|X_{kn}=i, Q_k=1\}\\
&\quad=\Pr\{X_{(k+1)n}=j, S_{k+1}\le \mathrm{UCL}|X_{kn}=i, S_k\le \mathrm{UCL}\}\\
&\quad=\Pr\left\{X_{(k+1)n}=j, S_{k+1}\right.\\
&\qquad\left.\le \mathrm{UCL}|X_{kn}=i, \sum_{l=(k-1)n}^{kn-1} X_l \le \mathrm{UCL}-i\right\}\\
&\quad=\Pr\{X_{(k+1)n}=j, S_{k+1}\le \mathrm{UCL}|X_{kn}=i\}, \qquad (11)
\end{aligned}$$

where the last step follows the Markov property of $\{X_k\}$.

However, when the sample size is large, it is difficult to evaluate the transition probability directly through Eq. (11). For efficiency, we explore a recursive relationship to reduce the computational load as follows:

$$\begin{aligned}
&V_{ij}(n,\mathrm{UCL})\\
&\quad=\Pr\{X_{(k+1)n}=j, S_{k+1}\le \mathrm{UCL}|X_{kn}=i\}\\
&\quad=\Pr\left\{X_{(k+1)n}=j, \sum_{l=nk+1}^{(k+1)n} X_l \le \mathrm{UCL}|X_{kn}=i\right\}\\
&\quad=\Pr\left\{X_{(k+1)n}=j, \sum_{l=kn+1}^{kn+n-1} X_l \le \mathrm{UCL}-j|X_{kn}=i\right\}\\
&\quad=\sum_{q=0}^{\mathrm{UCL}-j}\Pr\left\{X_{(k+1)n}=j, X_{kn+n-1}=q, \sum_{l=kn+1}^{kn+n-1} X_l\right.\\
&\qquad\left.\le \mathrm{UCL}-j|X_{kn}=i\right\}\\
&\quad=\sum_{q=0}^{\mathrm{UCL}-j}\Pr\{X_{(k+1)n}=j|X_{kn+n-1}=q\}\\
&\qquad\times\Pr\left\{X_{kn+n-1}=q, \sum_{l=kn+1}^{kn+n-1} X_l \le \mathrm{UCL}-j|X_{kn}=i\right\}\\
&\quad=\sum_{q=0}^{\mathrm{UCL}-j} p_{qj}\cdot V_{iq}(n-1,\mathrm{UCL}-j). \qquad (12)
\end{aligned}$$

when $n = 1$ the probability $V_{ij}(1, \mathrm{UCL})$ can be expressed as

$$\begin{aligned}
V_{ij}(1,\mathrm{UCL}) &= \Pr\{X_{k+1}=j, X_{k+1}\le UCL|X_k=i\}\\
&= \begin{cases} p_{ij} & \text{if } j\le \mathrm{UCL}\\ 0 & \text{otherwise}\end{cases}, \qquad (13)
\end{aligned}$$

where $p_{ij} = \Pr\{X_{k+1} = j|X_k = i\}$ as defined earlier. For the convenience of the readers, an algorithm sketch for calculating V_{ij} is included in Appendix. Based on the recursive relationship in Eq. (12) and the initial conditions in Eq. (13), we can easily get the transition matrix of the Markov chain $\{(X_{kn}, Q_k)\}$. Next, the expected run length from each initial state can be calculated using the methodology introduced in Section 2.1. Suppose X_0 is the initial observation according to the queue length data we observed, then the next queue length observation should be at least $\max\{0, X_0 - 1\}$. To ensure the first sample S_1 with n observations has the possibility to be in control, X_0 should satisfy

$$\sum_{j=1}^{n}\max\{0, X_0 - j\} \le \mathrm{UCL}. \qquad (14)$$

Denote the largest value of X_0 satisfying Eq. (14) as $N_{\max}$, then the ARL can be calculated as:

$$\mathrm{ARL} = \sum_{i=0}^{N_{\max}} \pi_i \cdot \left[\sum_{j=0}^{\mathrm{UCL}} V_{ij}(n,\mathrm{UCL})\cdot r_j\right] + 1, \qquad (15)$$

where the term in brackets represents the expected number of transitions to the OC state starting from the initial observation $X_0 = i$.

We would like to point out that during the calculation of the ARL with sample size n, the ARL of charts with sample size $m \leq n$ are readily available as byproducts. Therefore, we can specify the largest reasonable sample size $n_{\max}$ we want to use and get the results for all sample sizes ranging from 1 to $n_{\max}$ simultaneously.

2.3. Calculation of the ARL of WZ Charts

Unlike the nL chart, individual observations are used as the monitoring statistics in the WZ chart. When the queue length of the system falls and stays beyond the UCL for d_{u} consecutive observations, the process is considered to be OC.

To formulate a Markov chain model for this chart, we need to cover all possible in-control situations in the state space. Additionally, when the queue length is larger than UCL, we also need to keep track of how long it remains beyond UCL. In this article, we use the bivariate state $(\mathrm{UCL}+k, i)$, where $k > 0$ and $i < d_{\mathrm{u}}$, to represent the case when the current queue length is $\mathrm{UCL}+k$, and the $(i-1)$ most recent observations are also larger than UCL. For simplicity, we use (X_k, R_k) to denote the state variables. When $X_k > \mathrm{UCL}$, R_k denotes the successive number of observations beyond UCL up to now; when $X_k \leq \mathrm{UCL}$, R_k has no practical meaning and is set to 0 by default. As $\{X_k\}$ itself forms a Markov chain according to Ref. 12, it can be easily shown that $\{(X_k, R_k)\}$ also has the Markov property needed for ARL calculation. Among all the bivariate states, many of them can indicate the OC situation immediately, even before meeting the defined OC condition. Specifically, when $X_k > \mathrm{UCL} + d_{\mathrm{u}} - R_k + 1$, there is no chance of returning to UCL within d_{u} observations because each departure can only cause the queue length to decrease by one at most. Thus, the OC condition would be inevitably met. We aggregate all such states to form an OC state. To better illustrate, take $d_{\mathrm{u}} = 2$ and the state space $\Omega = \{(0,0), (1,0), \ldots (\mathrm{UCL}-1, 0), (\mathrm{UCL}, 0), (\mathrm{UCL}+1, 1), (\mathrm{UCL}+2, 1), (\mathrm{UCL}+1, 2), \mathrm{OC}\}$, where states $(\mathrm{UCL}+3, 1)$ and $(\mathrm{UCL}+2, 2)$ are included in the OC state, as shown in Fig. 3.

As there are two types of in-control states in the Markov chain, the transition matrix is much more complicated. When $R_k = 0$ (i.e., $X_k \leq \mathrm{UCL}$), we have

$$\Pr\{X_{k+1} = j, R_{k+1} = v | X_k = i, R_k = 0\} = \begin{cases} p_{ij}, & j \leq \mathrm{UCL} \\ p_{ij}, & j > \mathrm{UCL} \text{ and } v = 1 \\ 0, & \text{otherwise} \end{cases}, \quad (16)$$

where $p_{ij} = \Pr\{X_{k+1} = j | X_k = i\}$ as in Eq. (7). On the other hand, when $0 < R_k \leq d_{\mathrm{u}}$ (i.e., $X_k > \mathrm{UCL}$), we have

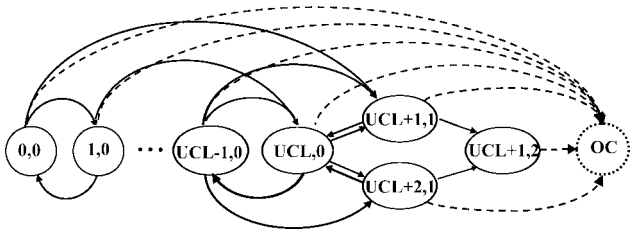

Figure 3. Transitions between states in the WZ chart (transition probabilities are not shown).

$$\Pr\{X_{k+1} = j, R_{k+1} = v | X_k = i, R_k = u\} = \begin{cases} p_{ij}, & j \leq \mathrm{UCL} \text{ and } v = 0 \\ p_{ij}, & j > \mathrm{UCL} \text{ and } v = u + 1 \\ 0, & \text{otherwise} \end{cases}. \quad (17)$$

The transition probability to the OC state is the sum of transition probabilities to different OC states, or

$$\Pr\{\mathrm{OC} | X_k = i, R_k = u\} = 1 - \sum_{v \leq d_{\mathrm{u}}} \Pr\{X_{k+1} = j, R_{k+1} = v | X_k = i, R_k = u\}. \quad (18)$$

From the complete transition matrix, we can obtain the expected run length r_i from each state $(i, 0)$ when $i \leq \mathrm{UCL}$ or $(i, 1)$ when $i > \mathrm{UCL}$ using Eq. (4), and consequently calculate the ARL for the WZ chart as:

$$\mathrm{ARL} = \sum_{i=0}^{\mathrm{UCL}+d_{\mathrm{u}}+1} \pi_i r_i + 1, \quad (19)$$

where π_i is again the steady state probability of queue length with value i.

We would like to point out that the proposed methods may be extended to other control charts as well. For example, we can combine the nL chart and the WZ chart by monitoring the sample sums with WZ parameters, and the methods presented in the last two subsections can be integrated to calculate the ARL of the combined chart.

3. COMPARISON OF CONTROL CHART PERFORMANCE

The methodology presented in Section 2 provides us with an efficient way to evaluate the performance of the two types of control charts. It also enables us to investigate the effects of different design parameters and design the control charts with the desired performance. In this section, we compare the nL and WZ charts through numerical results and provide relevant discussion regarding their performance. Please note that the calculated ARL has been converted to the corresponding ANOS in this section for fair comparison between these two charts.

Table 1. $ANOS_0$ of *nL* charts for M/M/1 queues.

	UCL								
Sample size	1	2	3	4	5	6	7	8	9
$\rho_0 = 0.3$									
1	18.09	69.15	243.67	829.70	2787.47	9317.72	31089.55	$1.04E+05$	$3.46E+05$
5	15.84	23.08	35.08	50.68	68.61	101.02	135.36	178.41	233.57
10	21.97	26.48	32.36	40.60	51.00	64.79	82.83	104.49	131.17
20	38.75	41.15	43.51	46.62	50.83	56.33	63.37	72.19	83.07
	1	4	7	10	13	16	19	22	25
$\rho_0 = 0.7$									
1	4.01	32.19	142.43	494.82	1554.23	4675.28	$1.38E+04$	$4.05E+04$	$1.18E+05$
5	8.25	13.13	18.62	25.38	33.65	44.90	58.95	77.61	101.72
10	15.23	19.63	23.11	26.87	31.27	36.39	42.22	48.84	56.21
20	30.20	36.72	39.33	41.01	42.77	44.93	47.56	50.68	54.28
	1	6	11	16	21	26	31	36	41
$\rho_0 = 0.9$									
1	1.85	19.33	85.19	245.11	571.49	1183.99	2283.54	4209.40	7535.49
5	6.10	9.28	12.74	16.88	22.18	29.02	37.75	48.72	62.27
10	11.92	15.72	18.84	21.66	24.37	27.14	30.06	33.23	36.76
20	23.80	30.46	34.55	37.32	39.42	41.22	42.92	44.62	46.39

3.1. In-control Performance $ANOS_0$

In control chart design, a critical issue is whether it is possible and straightforward to find design parameters that ensure the specified in-control performance (a specified $ANOS_0$). The problem becomes even more prominent when the data are discrete. Therefore, we calculate the ANOS using different control chart parameters in a given system to see if the specified $ANOS_0$ can be easily achieved. Table 1 lists the $ANOS_0$ of *nL* charts for M/M/1 systems at different utilization rates. Four different sample sizes ($n = 1, 5, 10, 20$) are compared in the table. From Table 1, we find that, as the sample size increases, the increments of $ANOS_0$ caused by the increase of UCL by the same amount become smaller and smaller. In other words, it is relatively easier to find parameters giving the specified $ANOS_0$ when the sample size increases.

It is also worth noting that as ρ_0 increases the change of the $ANOS_0$ becomes smaller for the same amount of UCL changes. This is because when ρ_0 is small, the distribution of the queue length is more concentrated. Therefore, a small change of UCL can result in a large shift in the OC probability. We can use a relatively large sample size to compensate for this effect when monitoring systems at low utilization.

In addition to the *nL* chart, we also investigated how the parameters UCL and d_u in WZ charts may influence the in-control performance. Tables 2 and 3 illustrate the $ANOS_0$ of different parameters of the WZ charts for M/M/1 systems at utilization rates 0.3 and 0.9, respectively.

From the two tables, we observe that both UCL and d_u can influence the in-control performance significantly, especially when the utilization rate is low. As a result, and similar to the *nL* chart, it is difficult to design a WZ chart with specified $ANOS_0$ when the utilization rate is low. The effects of d_u also have strong interactions with the effects of UCL. This means, for larger UCL, the increase of d_u can result in a much more significant increase of $ANOS_0$. Hence, in such cases, keeping UCL small and adjusting d_u may help find the parameters that achieve the specified in-control performance.

Table 2. $ANOS_0$ of the WZ charts for M/M/1 queue with $\rho_0 = 0.3$.

d_u	0	1	2	3	4	5	6	7	8
UCL									
1	17.18	45.98	98.81	190.73	345.65	601.00	1014.96	1677.20	2725.32
2	68.18	166.80	344.63	652.29	1169.64	2021.63	3402.13	5610.18	9104.37
3	242.67	573.90	1168.38	2195.14	3920.63	6761.37	$1.14E+04$	$1.87E04$	$3.04E+04$
4	828.70	1935.27	3918.52	7342.30	$1.31E+04$	$2.26E+04$	$3.79E+04$	$6.24E+04$	$1.01E+05$
5	2786.47	6477.48	$1.31E+04$	$2.45E+04$	$4.37E+04$	$7.52E+04$	$1.26E+05$	$2.07E+05$	$3.32E+05$

Table 3. $ANOS_0$ of the WZ charts for M/M/1 queue with $\rho_0 = 0.9$.

d_u	0	1	2	3	4	5	6	7	8
UCL									
1	1.66	2.41	3.34	4.43	5.64	6.99	8.44	10.00	11.66
2	2.85	4.19	5.66	7.27	9.00	10.84	12.79	14.83	16.97
3	4.93	6.99	9.14	11.38	13.72	16.15	18.68	21.29	23.99
4	8.11	11.07	14.01	17.00	20.06	23.19	26.39	29.67	33.01
5	12.65	16.67	20.56	24.43	28.32	32.25	36.22	40.26	44.35

3.2. Comparison of the Detection Power

In this subsection, we will compare how fast the control charts can detect changes of different magnitudes in terms of $ANOS_1$. We will present numerical results for M/G/1 queues with different distributions of the service time. Four in-control utilization rates, $\rho_0 = 0.3$, 0.5, 0.7, and 0.9, are considered in M/M/1 queues. In each case, we selected sample sizes 5, 10, and 20 for the *nL* chart (when $\rho_0 = 0.9$, $n = 1$ is also considered). For each sample size, we found the corresponding UCL that leads to the $ANOS_0$ closest to 370. For comparison, two groups of parameters of the WZ charts with similar $ANOS_0$ are also located. To ensure the comparisons are fair and meaningful, we also tried to adjust the design parameters of different charts to make their $ANOS_0$ as close as possible. To evaluate their OC performance, the system utilization gradually increases from ρ_0 to 0.995 and the corresponding $ANOS_1$ are computed and compared. To detect the same amount of utilization shifts, a smaller value of $ANOS_1$ often suggests a quicker response to the changes. Similar experiments are also conducted on general M/G/1 systems and selected results are presented.

3.2.1. *Performance for Systems Starting Empty*

As a special case, we first compare how long it takes the control charts to signal alarms if the monitored systems start from an empty queue. As this is common practice and an empty queue is a regeneration point in M/G/1 systems, the assessment of performance in systems initially empty has its own merits. This performance measure is manifested by the value r_0 in Section 2, that is, ARL $= r_0$. We use an M/M/1 system as an example. The utilization rate is set to 0.7 and three sample sizes of *nL* charts and two sets of parameters (UCL, d_u) of WZ charts are compared. Please note that the control limits of the WZ charts are different from that reported in Ref. 6 because we used ANOS as the design criterion rather than the α risk defined in Ref. 6. When the monitoring statistics are correlated, ANOS or ARL is the preferred design criterion as stated in Ref. 23. In Table 4, the normal utilization rate is $\rho_0 = 0.7$ and the control limits are determined such that ANOS is around 370 (the exact values are listed in the first row in Table 4 and are named $ANOS_0$). When the utilization rate shifts from $\rho_0 = 0.7$ upward, ANOS also changes. The corresponding ANOS for different shifts in the utilization (termed $ANOS_1$) are summarized in the rest of Table 4.

From the table, we find there is no evident difference between these charts in their ability to detect small shifts in the utilization when starting from an empty queue. However, the *nL* chart with a large sample size ($n = 20$) starts to underperform as the shift becomes larger, which is consistent with observations in many other types of control charts. Additionally, the WZ charts perform slightly worse at detecting large shifts compared with the *nL* charts using a reasonable sample size ($n = 5$ and 10) but a slightly better performance compared with the *nL* chart when $n = 20$. This can be explained as follows: for *nL* charts with $n = 5$ or 10, as the shift becomes larger, the decrease of $ANOS_1$ from $ANOS_0$ gets slightly larger than in the WZ chart. However, for the *nL* chart with $n = 20$, the WZ chart has a larger reduction in $ANOS_1$ instead. Similar conclusions can be drawn from other systems (e.g., M/M/1 systems with different utilization rates, or M/G/1 systems with different service distributions).

Table 4. Comparisons of $ANOS_1$ starting from an empty queue in M/M/1 systems.

	Empty M/M/1 ($\rho_0 = 0.7$)				
	nL chart			WZ chart	
ρ	$n = 5$, UCL = 40	$n = 10$, UCL = 70	$n = 20$, UCL = 109	$d_u = 14$, UCL = 4	$d_u = 4$, UCL = 7
0.7	367.5	371.3	366.8	362.3	362.6
0.75	239.6	241.9	238.5	233.9	235.9
0.8	166.0	168.1	166.7	161.7	163.2
0.85	121.1	123.4	123.9	118.0	118.5
0.9	92.4	94.9	96.7	90.2	89.8
0.95	73.2	75.9	78.7	71.7	70.8
0.96	70.1	72.9	75.8	68.7	67.6
0.97	67.3	70.1	73.2	66.0	64.8
0.98	64.6	67.4	70.7	63.4	62.2
0.99	62.1	65.0	68.4	61.0	59.7
0.995	60.9	63.8	67.3	59.9	58.5

Table 5. $ANOS_1$ of different utilization shifts in M/M/1 systems with $\rho_0 = 0.3$.

	Steady state M/M/1 ($\rho_0 = 0.3$)				
	nL chart			WZ chart	
ρ	$n = 5$, UCL = 11	$n = 10$, UCL = 14	$n = 20$, UCL = 19	$d_u = 4$, UCL = 1	$d_u = 2$, UCL = 2
0.3	402.8	367.5	383.4	345.6	344.6
0.35	224.1	199.3	196.3	190.5	195.8
0.4	138.5	122.0	117.3	120.4	125.8
0.45	92.6	81.6	78.5	80.0	84.3
0.55	48.5	43.9	44.6	42.0	44.4
0.65	28.9	27.7	31.2	24.3	25.5
0.75	18.3	19.3	25.0	14.7	15.2
0.85	11.7	14.4	22.0	8.3	8.4
0.95	6.9	11.2	20.4	3.2	3.2
0.96	6.5	10.9	20.3	2.8	2.8
0.98	5.8	10.4	20.2	1.8	1.8
0.995	5.2	10.1	20.0	1.2	1.2

3.2.2. *Performance in Steady State Systems*

When the systems operate in the steady state, the ARL calculated in Section 2 indicates the average performance without knowing the initial queue length when the monitoring starts. To measure the steady state performance, we first compare the *nL* chart with the WZ chart for M/M/1 systems at four different in-control utilization rates. The $ANOS_1$ for different shift magnitudes are summarized in Tables 5–8. By comparing these tables we find that, for systems with light traffic intensity, the *nL* chart with a larger sample size results in faster detection of small shifts in the utilization. On the other hand, it also results in slower detection of large shifts. However, if the systems operate at medium to high utilization (larger than 0.7) in normal conditions, the advantage of the large sample size diminishes even for small shifts.

Table 6. $ANOS_1$ of different utilization shifts in M/M/1 systems with $\rho_0 = 0.5$.

	Steady State M/M/1 ($\rho_0 = 0.5$)				
	nL chart			WZ chart	
ρ	$n = 5$, UCL = 23	$n = 10$, UCL = 35	$n = 20$, UCL = 48	$d_u = 3$, UCL = 4	$d_u = 1$, UCL = 5
0.5	408.9	401.9	405.6	396.5	386.8
0.55	251.6	243.6	239.3	242.8	238.8
0.6	164.6	158.4	153.9	161.7	159.7
0.65	112.7	108.6	105.9	109.5	108.4
0.75	57.5	56.7	57.9	55.3	54.7
0.85	29.6	31.3	35.9	26.6	26.2
0.95	12.0	15.9	24.0	8.6	8.4
0.96	10.6	14.6	23.1	7.2	7.0
0.97	9.1	13.4	22.3	5.3	5.2
0.98	7.7	12.3	21.5	4.0	3.9
0.99	6.4	11.1	20.7	2.6	2.6
0.995	5.7	10.6	20.4	1.7	1.7

Table 7. $ANOS_1$ of different utilization shifts in M/M/1 systems with $\rho_0 = 0.7$.

	Steady state M/M/1 ($\rho_0 = 0.7$)				
	nL chart			WZ chart	
ρ	$n = 5$, UCL = 41	$n = 10$, UCL = 72	$n = 20$, UCL = 113	$d_u = 15$, UCL = 4	$d_u = 11$, UCL = 5
0.7	376.5	380.7	376.4	380.8	378.9
0.75	232.5	234.6	231.5	231.0	230.5
0.8	148.1	150.1	149.5	150.0	149.3
0.85	94.5	96.9	98.8	95.5	94.4
0.9	57.3	60.3	64.7	57.7	56.3
0.95	28.9	32.8	39.8	29.3	28.1
0.96	23.9	27.9	35.5	22.6	21.6
0.97	19.0	23.3	31.4	17.7	16.8
0.98	14.2	18.7	27.5	12.7	12.0
0.99	9.6	14.3	23.7	6.1	5.7
0.995	7.3	12.1	21.8	4.4	4.2

Likewise, in systems at a low utilization, larger d_u (or smaller UCL) has a slightly better detection performance for small shifts. However, the benefits are barely visible in systems with medium to high utilization. However, in contrast to the *nL* chart where different sample sizes have different performances at detecting large shifts, different groups of parameters in the WZ chart provide the same detection power for large shifts. This phenomenon indicates that larger d_u is always preferable in control chart design as its performance is consistently better or not worse than charts with smaller d_u.

Besides the M/M/1 systems, we also investigated the performance of the *nL* and WZ charts in M/G/1 systems with general service distributions. Similar conclusions can be drawn from the numerical results in M/G/1 systems with different nominal utilization rates. For example, Tables 9 and 10 compare $ANOS_1$ in an M/G/1 system with Erlang-*k* distributed service time with parameter $k = 4$. The monitoring performance in the system at two utilization rates is compared. A larger sample size in the *nL* chart and larger d_u in the WZ chart still exhibit some benefits at detecting changes of small magnitude. To limit the length of this article, we do not list all the results of M/G/1 systems (we have considered uniform, Erlang-*k* distributed service times at different utilizations). However, we would like to emphasize that the insights and conclusions derived from M/M/1 systems are still valid in those settings.

3.2.3. *Performance of Charts with Small* ARL_0

Considering the applications in some service systems, we also provide some comparisons when $ANOS_0$ is small. The ANOS in M/E4/1 systems at two nominal utilization rates are

Table 8. $ANOS_1$ of different utilization shifts in M/M/1 systems with $\rho_0 = 0.9$.

	Steady State M/M/1 ($\rho_0 = 0.9$)					
	nL chart				WZ chart	
ρ	$n = 1$, UCL = 18	$n = 5$, UCL = 85	$n = 10$, UCL = 160	$n = 20$, UCL = 290	$du = 15$, UCL = 13	$du = 11$, UCL = 14
0.9	349.3	357.6	358.3	359.6	366.2	367.2
0.91	299.4	306.9	308.0	310.1	319.3	319.9
0.92	254.7	261.6	263.1	266.1	263.7	263.8
0.93	214.4	220.8	222.7	226.4	226.3	226.1
0.94	177.6	183.6	185.9	190.4	191.9	191.5
0.95	143.8	149.4	152.1	157.5	149.7	149.1
0.96	112.3	117.6	120.8	126.9	120.2	119.5
0.97	82.5	87.7	91.3	98.3	92.1	91.4
0.98	54.2	59.3	63.2	71.1	56.1	55.6
0.99	26.8	31.8	36.3	45.2	29.9	29.5
0.995	13.3	18.3	23.1	32.5	12.6	12.4

provided in Tables 11 and 12. From the table, it is not surprising to observe that charts with larger sample sizes perform worse. Further, there is no practical difference between the WZ charts and the *nL* charts with sample size 1.

To summarize, the calculation of ARL (ANOS) does provide us with very useful guidelines for designing the *nL* and WZ charts. In general, larger sample size in *nL* charts and larger d_u in WZ charts are preferable because: (i) it is easier to find the parameters that lead to the specified in-control performance ($ANOS_0$); (ii) they often offer faster detection of small shifts in the system utilization. However, when $ANOS_0$ is small, the larger sample size in the *nL* chart loses its advantage. It is also worth noting that while a larger sample size in the *nL* chart has disadvantages at detecting large shifts, larger d_u in the WZ chart has no evident side effects. Although no uniform conclusion can be made by comparing the *nL* chart with the WZ chart, we found the *nL* chart tends to perform slightly better at detecting small shifts, whereas the WZ chart stands out at detecting large shifts.

4. CONCLUSIONS

This article studies the performance of two types of control charts used to monitor queue length data in M/G/1 systems. Based on the general Markov chain models, we analyzed the ARL (ANOS) of the two control charts without extensive and time consuming simulations. Numerical examples using these methods also revealed some guidelines for selecting control chart parameters. These guidelines can help design control charts with better performance at monitoring operations in production/service systems. It should be pointed out that the proposed method can also be applied to G/M/1 systems by taking the queue length observations right before each arrival. Also, the method for the *nL* chart with sample size *n* may possibly be extended to CUSUM and EWMA charts when monitoring M/G/1 systems.

There are still some open questions in the current research. Although the two types of control charts have been proposed and draw increasing attention, there is no proof that they are optimal in terms of detection power. Therefore, developing an optimal or near optimal control chart based on queue length observations could improve the performance of the monitoring. Additionally, under the current monitoring scheme, the time intervals between successive departures are not used. It would be interesting to investigate whether we can improve the performance of the monitoring by incorporating additional information. If so, how much improvement we can make to justify the additional cost of the new data is also worth investigation.

Table 9. $ANOS_1$ of different utilization shifts in M/E4/1 systems with $\rho_0 = 0.5$.

	Steady state M/E4/1 ($\rho_0 = 0.5$)				
	nL chart			WZ chart	
ρ	$n = 5$, UCL = 17	$n = 10$, UCL = 25	$n = 20$, UCL = 34	$d_u = 5$, UCL = 2	$d_u = 2$, UCL = 3
0.5	356.4	383.2	377.4	366.0	331.9
0.55	220.3	231.2	225.2	221.5	206.8
0.6	144.8	150.8	149.2	142.7	135.7
0.65	100.0	104.8	107.6	96.4	92.6
0.7	71.7	76.7	83.2	67.4	65.0
0.75	52.9	58.5	67.8	48.1	46.3
0.8	39.6	45.9	57.7	34.5	33.0
0.85	29.6	36.8	50.7	24.1	22.8
0.9	21.7	29.8	45.8	15.7	14.6
0.95	14.3	23.1	41.1	8.2	7.5
0.96	12.7	21.5	39.6	6.8	6.2
0.97	11.0	19.6	37.4	5.3	4.9
0.98	9.2	17.2	34.1	3.9	3.6
0.99	7.2	14.1	28.7	2.5	2.3
0.995	6.2	12.2	24.9	1.7	1.6

Table 10. $ANOS_1$ of different utilization shifts in M/E4/1 systems with $\rho_0 = 0.9$.

	Steady state M/E4/1 ($\rho_0 = 0.9$)				
	nL chart			WZ chart	
ρ	$n = 5$, UCL = 64	$n = 10$, UCL = 118	$n = 20$, UCL = 210	$d_u = 16$, UCL = 9	$d_u = 3$, UCL = 12
0.9	381.5	377.5	382.7	372.9	377.6
0.91	324.0	322.2	329.2	315.6	320.5
0.92	273.8	273.9	282.6	265.5	270.7
0.93	229.6	231.3	241.5	221.4	226.7
0.94	190.2	193.3	205.0	182.0	187.4
0.95	154.7	159.0	172.0	146.4	151.7
0.96	122.1	127.7	142.0	113.7	118.8
0.97	91.8	98.6	114.3	83.4	87.9
0.98	63.0	71.1	88.4	54.6	58.3
0.99	34.6	43.8	63.0	27.0	29.2
0.995	20.2	28.7	48.1	13.5	14.7

Table 11. $ANOS_1$ of different utilization shifts in M/E4/1 systems with $\rho_0 = 0.7$ (small $ANOS_0$).

	Steady state M/E4/1 ($\rho_0 = 0.7$)					
	nL chart			WZ chart		
ρ	$n = 1$, UCL = 4	$n = 5$, UCL = 10	$n = 10$, UCL = 11	$n = 20$, UCL = 2	$d_u = 2$, UCL = 2	$d_u = 5$, UCL = 1
0.7	29.9	28.5	28.9	31.5	30.8	33.0
0.75	22.4	23.5	26.2	30.0	23.1	24.9
0.8	16.6	19.5	24.0	28.3	17.2	18.8
0.85	11.9	16.2	22.1	26.5	12.4	13.7
0.9	7.8	13.0	19.8	24.5	8.3	9.3
0.95	4.3	9.6	16.4	22.3	4.5	5.2
0.96	3.6	8.8	15.5	21.9	3.8	4.4
0.97	2.9	8.0	14.4	21.4	3.1	3.5
0.98	2.3	7.0	13.1	21.0	2.4	2.7
0.99	1.6	6.1	11.7	20.5	1.7	1.9
0.995	1.3	5.5	10.9	20.2	1.3	1.4

Table 12. $ANOS_1$ of different utilization shifts in M/E4/1 systems with $\rho_0 = 0.9$ (small $ANOS_0$).

	Steady state M/E4/1 ($\rho_0 = 0.9$)					
	nL chart			WZ chart		
ρ	$n = 1$, UCL = 6	$n = 5$, UCL = 18	$n = 10$, UCL = 16	$n = 20$, UCL = 2	$d_u = 2$, UCL = 4	$d_u = 11$, UCL = 1
0.9	23.7	23.3	23.1	24.5	24.1	22.5
0.91	21.1	21.6	22.4	24.1	21.5	20.4
0.92	18.6	20.0	21.7	23.7	19.0	18.4
0.93	16.2	18.4	20.9	23.2	16.6	16.3
0.94	13.9	16.8	19.9	22.8	14.2	14.3
0.95	11.6	15.1	18.9	22.3	11.9	12.2
0.96	9.4	13.4	17.7	21.9	9.7	10.1
0.97	7.3	11.6	16.3	21.4	7.5	8.0
0.98	5.2	9.6	14.6	21.0	5.3	5.7
0.99	3.1	7.4	12.5	20.5	3.1	3.4
0.995	2.0	6.3	11.3	20.2	2.1	2.2

APPENDIX

Proof of the Markovian Property of (X_{kn}, Q_k)

According to the operation of control charts, once $Q_k = 0$ (i.e., the process is OC), we have $Q_n = 0, n > k$ to form an absorbing state for ARL calculation regardless of previous observations. Therefore,

$$\begin{aligned}\Pr\{X_{(k+1)n}, Q_{k+1} = 0 | X_{kn}, Q_k = 0, X_{(k-1)n}, Q_{k-1}, \ldots, X_n, Q_1\} \\ = \Pr\{X_{(k+1)n}, Q_{k+1} = 0 | X_{kn}, Q_k = 0\} = 1. \end{aligned} \quad \text{(A1)}$$

On the other hand, when $Q_k = 1$, we have

$$\begin{aligned}\Pr\{X_{(k+1)n}, Q_{k+1} | X_{kn}, Q_k = 1\} &= \frac{\Pr\{X_{(k+1)n}, Q_{k+1}, X_{kn}, Q_k = 1\}}{\Pr\{X_{kn}, Q_k = 1\}} = \frac{\sum_{X_{(k-1)n+1}+\cdots+X_{kn} \le \text{UCL}} \Pr\{X_{(k+1)n}, Q_{k+1}, X_{kn}, X_{kn-1}, \ldots, X_{(k-1)n+1}\}}{\sum_{X_{(k-1)n+1}+\cdots+X_{kn} \le \text{UCL}} \Pr\{X_{kn}, X_{kn-1}, \ldots, X_{(k-1)n+1}\}} \\ &= \frac{\sum_{X_{(k-1)n+1}+\cdots+X_{kn} \le \text{UCL}} \Pr\{X_{(k+1)n}, Q_{k+1} | X_{kn}, X_{kn-1}, \ldots, X_{(k-1)n+1}\} \cdot \Pr\{X_{kn}, X_{kn-1}, \ldots, X_{(k-1)n+1}\}}{\sum_{X_{(k-1)n+1}+\cdots+X_{kn} \le \text{UCL}} \Pr\{X_{kn}, X_{kn-1}, \ldots, X_{(k-1)n+1}\}} \\ &= \frac{\Pr\{X_{(k+1)n}, Q_{k+1} | X_{kn}\} \cdot \sum_{X_{(k-1)n+1}+\cdots+X_{kn} \le \text{UCL}} \Pr\{X_{kn}, X_{kn-1}, \ldots, X_{(k-1)n+1}\}}{\sum_{X_{(k-1)n+1}+\cdots+X_{kn} \le \text{UCL}} \Pr\{X_{kn}, X_{kn-1}, \ldots, X_{(k-1)n+1}\}} = \Pr\{X_{(k+1)n}, Q_{k+1} | X_{kn}\}. \end{aligned} \quad \text{(A2)}$$

Following similar derivations, it is not difficult to get

$$\begin{aligned}\Pr\{X_{(k+1)n}, Q_{k+1} | X_{kn}, Q_k = 1, X_{(k-1)n}, Q_{k-1} = 1, \ldots, X_n, Q_1 = 1\} \\ = \Pr\{X_{(k+1)n}, Q_{k+1} | X_{kn}\}. \end{aligned} \quad \text{(A3)}$$

Please note that there is a probability of zero to obtain $Q_k = 1$ while $Q_i = 0$ for some $i < k$ following Eq. (A1). Therefore, the Markov property of $\{(X_{kn}, Q_k)\}$ proved as $(X_{(k+1)n}, Q_{k+1})$ only depends on X_{kn} and Q_k.

Steady State Distribution of M/E$_k$/1 Systems

This appendix presents the algorithm for computing the steady state distribution of queue length in M/E$_k$/1 systems. The general formula for other service distributions has been provided in Eq. (8) of Section 2.

Erlang-k service distribution may be viewed as a series of k successive independent service steps where each step takes an exponentially distributed amount of time. Each job passes through these services step-by-step. Assuming that the average service time for each phase is $1/\mu$ and the arrival rate is λ, then the overall Erlang-k distribution has mean service time r/μ.

Consider the number of uncompleted phases of service left in the system. There exists a one-to-one correspondence between the number of customers in the system as well as the total uncompleted phases of work. Let q_i be the steady state probability of i uncompleted phases of work in the system. Then,

$$\pi_0 = q_0, \quad \pi_i = \sum_{j=(i-1)r+1}^{ir} q_j \forall i \ge 1. \quad \text{(A4)}$$

It can be shown that q_i is a function of the r distinct roots $x_1, x_2, \ldots, x_r$ with $|x| < 1$ of the polynomial equation

$$(\lambda + \mu)x^r = \lambda + \mu \cdot x^{r+1}. \quad \text{(A5)}$$

In fact, we have

$$q_j = \sum_{i=1}^{r} c_i x_i^j, \quad \text{where } c_i = \frac{1-\rho}{\Pi_{v \ne i}(1 - x_v/x_i)}. \quad \text{(A6)}$$

By substituting Eq. (A6) into Eq. (A4), the steady state distribution of queue length in M/E$_k$/1 queues can be easily obtained.

Algorithm Sketch for Computing V_{ij}

In this article, Eq. (12) provides a recursive relationship to compute the matrix V_{ij} for different sample sizes and control limits. In the following, we present an algorithm sketch to implement the procedure. Please note that, at iteration n and N, the matrices $V_{ij}(n', N')$ have already been computed for all the $n' < n$ and $N' < N$cases in the previous iterations.

Initialize the iteration number $n = 1$;
Set the initial values $V_{ij}(1, N) = p_{ij}$ if $j \le N$,
  and $V_{ij}(1, N) = 0$ otherwise for all $N \le$ UCL
repeat $n = 2$ to UCL
  repeat $N = 1$ to UCL
    repeat i, $j = 0$ to UCL

$$V_{ij}(n, N) = \sum_{q=0}^{N-j} p_{qj} \cdot V_{iq}(n-1, N-j)$$

ACKNOWLEDGMENTS

The financial support of this work is provided by NSF Awards CMMI-0545600 and CMMI-0757683. The authors appreciate the editor's and referees' valuable comments and suggestions.

REFERENCES

[1] I. Ben-Gal, G. Morag, and A. Shmilovici, Context-based statistical process control, Technometrics 45 (2003), 293–311.
[2] I. Ben-Gal and G. Singer, Statistical process control via context modeling of finite-state processes: an application to production monitoring, IIE Trans 36 (2004), 401–415.
[3] U.N. Bhat, A sequential technique for the control of traffic intensity in Markovian queues, Ann Oper Res 8 (1987), 151–164.
[4] U.N. Bhat, An introduction to queueing theory: Modeling and analysis in applications, Birkhäuser, Boston, 2008.

[5] U.N. Bhat and G.K. Miller, Elements of applied stochastic processes (2002).
[6] U.N. Bhat and S.S. Rao, A statistical technique for the control of traffic intensity in the queuing systems M/G/1 and GI/M/1, Oper Res 20 (1972), 955–966.
[7] C.M. Borror, C.W. Champ, and S.E. Rigdon, Poisson EWMA control charts, J Qual Technol 30 (1998), 352–361.
[8] D. Brook and D.A. Evans, An approach to the probability distribution of CUSUM run length, Biometrika 59 (1972), 539–549.
[9] C.W. Champ and W.H. Woodall, Exact results for shewhart control charts with supplementary runs rules, Technometrics 29 (1987), 393–399.
[10] R.B. Cooper, Introduction to queueing theory, North Holland, New York, 1981.
[11] S. Fomundam and J. Herrmann, A survey of queuing theory applications in healthcare, The Institue for System Research, University of Maryland, College Park, MD, 2007.
[12] D. Gross, J.F. Shortle, J.M. Thompson, and C.M. Harris, Fundamentals of queueing theory, Wiley, Hoboken, NJ, 2008.
[13] K.B. Hendricks and J.O. McClain, The output process of serial production lines of general machines with finite buffers, Manage Sci 39 (1993), 1194–1201.
[14] L. Huwang, A.B. Yeh, and C.-W. Wu, Monitoring multivariate process variability for individual observations, J Qual Technol 39 (2007), 258–278.
[15] J.G. Kemeny and J.L. Snell, Finite Markov chains, Springer-Verlag, Princeton, NJ, 1976.
[16] L. Kleinrock, Queueing systems, Volume 1: Theory, Wiley, New York, 1975.
[17] J.M. Lucas and M.S. Saccucci, Exponentially weighted moving average control schemes: Properties and enhancements, Technometrics 32 (1990), 1–12.
[18] J.E. Mcneill, J.W. Fowler, G.T. Mackulak, and B.L. Nelson, "Cycle time quantile estimation in manufacturing systems employing dispatching rules," Proceedings of the 2005 Winter Simulation Conference, Orlando, FL, USA, 2005, pp. 749–755.
[19] H.T. Papadopoulos and C. Heavey, Queueing theory in manufacturing systems analysis and design: A classification of models for production and transfer lines, Eur J Oper Res 92 (1996), 1–27.
[20] H. Shore, Control charts for the queue length in a G/G/S system, IIE Trans 38 (2006), 1117–1130.
[21] L. Shu and W. Jiang, A Markov chain model for the adaptive CUSUM control chart, J Qual Technol 38 (2006), 135–147.
[22] D.G. Wardell, H. Moskowitz, and R.D. Plante, Control charts in the presence of data correlation, Manage Sci 38 (1992), 1084–1105.
[23] W.H. Woodall, The use of control charts in health-care and public-health surveillance, J Qual Technol 38 (2006), 89–104.
[24] P.F. Zantek, Design of cumulative sum schemes for start-up processes and short runs, J Qual Technol 38 (2006), 365–375.