A Control Chart Approach to Optimize Sojourn Times of Customers in K-policy Queues

R. Sivasamy*[1] W.M. Thupeng[1] K. Kebotsamang[1]

[1]*Department of Statistics, University of Botswana, Private Bag 0022, Gaborone Botswana*

Abstract

This article presents a control chart approach that optimizes the sojourn times of customers in a dual rate queuing model for serving a single queue with a K-policy. Each member of the initial space $D\_1 = \{0,1,2,\dots,K\}$ is served by a single server (S_1) with an average service rate of μ_1 per unit time. When the queue length increases to a higher space $D\_2 = \{K+1, K+2, \dots\}$, an additional server ($S_2$) is installed to join the main server (S_1) to provide service at a rate of $\mu 2 = \mu 1 + \mu$ ($\mu > 0$) per unit time. When each service is started at a starting point with a queue length of $j = 0$ or greater, no prioritization is allowed regardless of the amount of work available on the system. The arrival process is a Poisson process with an average rate of λ. Based on a predefined target level (i.e. maximum queue length or maximum latency), the K-policy is optimized using a modified Shewart-like control chart approach. By providing simulation time weights, each performance metric of interest is calculated and used in the proposed control chart. To find the optimal value for the queue length $K = K_0$ and install the second server, a number table is created numerically to plot this optimal value.

Mathematics Subject Classification (2020). xxx, yyy, zzz

Keywords. K- policy queue, control chart, queue length distribution, mean queue length

1. Introduction

One of the foremost concerns of any service provider is the time customers have to spend waiting for service. Queuing models are often used to predict queue length and job waiting time. These timeout predictions can help us determine the optimum resource allocation, such as the number of servers needed to maintain the pre-set timeouts for all jobs. For example, we may use the models to determine system capacity needs and install a new server to keep job downtime within a pre-set target level.

Numerous authors have proposed several infinite capacity queue models using probabilistic and theoretical metric representations of distributions such as queue length, waiting time and busy interval. However, we can only find finite capacity scenarios in most of the queuing systems in practice such as in banks, airport counters and hospitals.

In addition, if the maximum number of customers, n, per session is less than the maximum capacity of the system (N), then the chance that a queue length takes any value between n and N is zero. This calls for some novel methods that can be used to obtain a best finite capacity model to capture most of our real-life waiting line aspects.

The application of queueing models has been the subject of intensive research studies mainly because of their ability to approximately represent real life systems. For a basic understanding of queuing theory, researchers can refer to [1–4]. Shore [6, 7] uses conventional approximation to create Shewhart control charts for variable data and attributes. Montgomery [5] and Sivasamy and Jayanthi [8] describe important issues related to statistical process control (SPC) and product control theory. Statistical control charts for monitoring client latency in M/M/1 queues are described in Zhao and Gilbert [11]. For more information on single-queue applications and process control methods, see the references in Sivasamy [8–10].

This paper proposes a control chart approach to choosing the optimal value for a threshold queue length K based on some objective functions related to the M/M/1/K policy queues and the G/G/1 queue in order to determine which K-policy would be optimal. It is hoped that the introduction of the K-policy leads to a waiting time that is less than the waiting time for jobs processed by any other queue without a K-policy. Among the main observable characteristics in queues, the number of users in the system can be controlled to keep waiting times as minimal as possible. The design of efficient control charts is an attempt to monitor and control such systems. The control charts proposed here can be applied to monitor infinite queues with Markovian arrivals, exponential service times, and s identical parallel servers.

The rest of the paper is organized as follows. Section 2 deals with an M/M/1/K-policy queues where one of its applications to the waiting line of patients of hospitals is discussed. Section 3 extends the methodology to a few G/G/1/K-policy queues whereas section 4 discusses M/G/1/K-policy queues. In each case, an optimal value of queue length K is fixed from a table of values constructed for different combinations of input values and, simulated performance measures are used to construct Shewhart-like control charts for illustrative purposes. Section 5 presents concluding remarks and scope for future research.

2. Methodology

In this section, we derive conditional distributions of waiting time for customers of a single server Poisson facility, i.e. M/M/1/K-policy queue, that serves its customers according to a state dependent K-policy. The exponential service rate is set to a slow rate of $\mu_1 > 0$ for all clients waiting conditional on whether the system size, the length of the queue (L_s), is less than or equal to some integer threshold K. If the system size, L_s is larger than K, the service rate changes to a faster rate of $\mu_2 = (\mu_1 + \mu)$, where $\mu > 0$. These constraints introduce the K-policy in the system to allow customers to experience shorter waiting times than the job waiting times offered by other M/M/1 queues.

Let the traffic intensity be $\rho_1 = \lambda/\mu_1$ and $\rho = \lambda/(\mu_1 + \mu)$, where λ denotes the mean inter arrival times of an exponential distribution. Further, let the probability that the queue length equals n be $q_n = P(L_s = n)$. Then by adopting the works of Sivasamy [9, 10], the distribution, $\{q_n\}$, of the queue length variable L_s is given by the following equation

$$q_j = \begin{cases} \frac{(1-\rho)(1-\rho_1)}{1-\rho-\rho_1^{K+1}}, & j = 0 \\ q_0\rho_i^j, & j = 1,2,\ldots,K \\ q_{K+1}\rho^{j-K+1}, & j = K+1, K+2, \ldots. \end{cases} \quad (2.1)$$

We simulate the performance measures to assess the finite sample performance of the proposed control chart approach to optimizing sojourn times of customers in a K-policy queue. The same metrics are then used to construct a Shewhart-like control chart and a table which records queue length and waiting times for customer arrival number n for $n = 1, 2, \ldots, N_1$, where N_1 is the observed maximum queue length.

For simplicity, we partition the queue length Ls into two variables V_1 and V defined on the initial space and the higher space $D\_1 = \{0,1,2, \ldots, K\}$ and $D\_2 = \{K+1, K+2, \ldots\}$, respectively. Now let's assume that all customers in D_1 are served by an M/M/1/K queue with arrival rate λ and service rate, μ1 ($\rho_1 = \lambda/mu_1$). Define a probability Q_K:

$$Q_K = \sum_{i=0}^{n} q_j = \frac{(1-\rho)(1-\rho_1^{K+1})}{(1-\rho)(1-\rho_1^{K+1})+(1-\rho_1)\rho_1^{K+1}} \quad (2.2)$$

If we let V_1 be the system size and $V_1(n) = P(V_1 = n)$ be the steady state probability that there are n customers in the underlying M/M/1/K queue, then

$$V_1(n) = \frac{1-\rho_1}{1-\rho_1^{K+1}}\rho_1^n; n = 0,1,2, \ldots, K \quad (2.3)$$

Furthermore, every customer in the D_2 subspace is served by another M/M/1 queue with arrival rate, λ and service rate, $\mu_1+\mu$, with traffic intensity of $\rho = \lambda/(\mu_1+\mu)$ and probability $1-Q_K$. Thus, if $V(n) = P(V = n)$ is the steady state probability of the underlying M/M/1 queue, then

$$V(n) = (1-\rho)\rho^n; \quad n = 0, 1, 2, \ldots \infty \quad (2.4)$$

Thus under the K-policy, the joint distribution of the system size $G = L_s$, say $g(n) = P(G = n)$, is now re-organized as a convex combination of two proper geometric distributions $V_1(n)$ and $V(n)$, with probabilities Q_K and $1-Q_K$ and ρ_1, respectively:

$$g(n) = (1-Q_K)V(n) + Q_K V_1(n) \quad (2.5)$$

Consider any random variable, X, having a probability mass function $f(n) = P(X = n)$ and and an r^{th} raw moment, $E(X^r) = \sum_{n=0}^{\infty} n^r f(n)$. The latter formula can be used to compute both the raw and central moments of $V \sim V(n)$, $V_1 \sim V_1(n)$ and G~ $g(n)$. That is, the moments of G ~ $g(n)$ are :

$$\mu_G = (1-Q_K)\mu_V + Q_K\mu_{V_1} \quad (2.6)$$

$$\sigma_G = (1-Q_K)\sigma_V + Q_K\sigma_{V_1} \quad (2.7)$$

$$\nu_G = (1-Q_K)\nu_V + Q_K\nu_{V_1} \quad (2.8)$$

where μ_G, σ_G and ν_G are the mean, standard deviation and the skewness measure of the queue size G, respectively.

Using the results in equations 2.6 – 2.8, we are now able to develop a control chart for the mean queue size, μ_G, by employing the methodology of Shore [6, 7] for the attributes data with 3-σ control limits:

$$UCL = \mu_G + 3\sigma_G + 1.324\sigma_G\nu_G - 0.5 \quad (2.9)$$

$$CL = \mu_G \quad (2.10)$$

$$LCL = \mu_G - 3\sigma_G + 1.324\sigma_G\nu_G + 0.5 \quad (2.11)$$

It is noticed that when the skewness measure takes the value zero, i.e., skew(G)=0, the proposed control chart automatically reduces to the traditional Shewhart chart. Hence, the modified chart limits for the number of customers in the system may be viewed as a natural extension of the traditional Shewhart control chart by adding a measure of skewness.

We use Little's law to relate the capacity of a queuing system, the average time spent in the system, and the average arrival rate into the system without knowing any other features of the queue. Since the mean waiting time, $w_n = n/\lambda$ for each queue length, $L_s = n$ according to the Little's law, we claim that $P(L_s = n) = P(L_s/\lambda = n/\lambda)$. Hence, the empirical distribution of the waiting time statistic, $W = \{w_n = n/\lambda;\ n = 0, 1, 2, \ldots\}$, has the same distribution as the queue size L_s. Therefore, to monitor the waiting statistic, W, we have the following approximate 3-σ control limits:

$$UCL = \mu_W + 3\sigma_W + 1.324\sigma_W \nu_W - 0.5 \quad (2.12)$$

$$CL = \mu_W \quad (2.13)$$

$$LCL = \mu_W - 3\sigma_W + 1.324\sigma_W \nu_W + 0.5 \quad (2.14)$$

To actually apply the control limits in equations 2.12 – 2.14, we should observe the arrival and service time of an appropriate number of customers, the average inter-arrival time ($1/\hat{\lambda}$) and the average service time of an appropriate number of customers ($1/\hat{\mu}$) for both the first and second servers. Each is a simulated estimate with $\hat{\lambda} = (M-1)/\sum_{n=1}^{M-1} A_n$ and $\hat{\mu} = K/\sum n = 1^K B_n$.

The variance test of Chi-square, χ^2, can be applied to the null hypothesis that a set of independent observations $S_n: n = 1,2, \ldots, N$ is drawn from an exponential distribution with parameters μ and variance σ^2. That is, $H_0: \sigma^2 = 1/\mu^2$. Under H_0, the test statistic follows Chi-square distribution with (N-1) degrees of freedom:

$$\sum_{n=1}^{N} \frac{(S_n - 1/\hat{\mu})^2}{(1/\hat{\mu})^2}$$

We can therefore monitor the queue length or sojourn period of the customers in the system base based on the accepted values. As a result, we can make the best decisions regarding the K-policy and its optimization using the proposed chart control process. Thus, the simulation helps us locate the arrival and departure epochs for any number of prefixed arrivals, for example 150 used in the simulation in the next section, and to use the accepted values by the variance test of Chi-square.

3. Simulation

Suppose we are interested in the ATM queue. We can graphically describe the operation of this system by plotting the number of customers present at the ATM; system status.

Every time a customer arrives, the chart increases by one unit while the customer leaves decreases the graph by one unit. This graph (called a sampling line), can be obtained from observing an actual survey, but can also be constructed artificially. The artificial construction and analysis of the resulting sample path (or multiple sample paths in more complex cases) is simulation.

The above sample path consisted of only horizontal and vertical lines, as customer arrivals and departures occurred at distinct points of time, what we refer to as events. Between two consecutive events, nothing happens - the graph is horizontal. If the number of possible events is finite, we call the simulation a "discrete event." So simulation is generally about pretending you're working with a real thing while actually working with an imitation, so discrete event simulation can serve as an approximation.

3.1. Waiting time monitoring and M/M/1 queues under an optimum K-policies

We randomly fix $n = 150$, $\lambda = 0.5$, $\mu = 0.03$, $\mu_1 = 0.27$, and $K = 7, 10, 12, 15, 17, 20, 23$ and 27, in order to assess the finite sample performance of the proposed control chart. We simulate the arrival times, service start and end times, and departure epochs for each customer based on the fixed parameters above. Then the number of customers present and the waiting time were calculated for each arrival and departure from the simulated data. Furthermore, the mean inter-arrival times, and mean service times were computed.

We also used the chi-square statistic tested the hypothesis of no difference between the estimated and the given values of the arrival and service rates $\lambda = 0.5$, $\mu = 0.27$, $\mu_1 = 0.03$ is performed by the chi-square test. Then, for those values validated by the chi-square test, we calculate the numerical values of the theoretical expressions through, the system size L_s, the sojourn time W, control limits and the timeout value for $n = 150$ customers managed by the M/M/1/N/K-policy system.

The results of the simulation in Table 1 show that the maximum queue length reached at the time of the migration is 17 out of 150 arrivals. Also recorded in Table Table 1 are some other important simulation outputs for different values of K. As K varies from 17 to 27, all clients are processed by the M/M /1/N (> 17)/K-policy queue with $\lambda = 0.5$ and $\mu = 0.27$. The corresponding control graphs observed at the arrival and departure points are plotted in Figures 1a and 1b to monitor waiting times against transition numbers 1–300 together with the control lines.

Table 1. The control limits for M/M/1/N(>17) /K-policy queues when $n = 150$, $\lambda = 0.25$, $\mu_1 = 0.27$, $\mu = 0.03$ and for $K = 7, 10, 12, 15, 17, 20, 23$ and 27.

K	Max L	UCL	CL	LCL	Max W	UCLw	CLw	LCLw
7.00	**11.00**	**11.20**	**4.10**	**0.10**	**44.00**	**46.32**	**16.41**	**1.11**
10.00	14.00	14.62	5.87	-0.72	56.00	59.97	23.47	-4.37
12.00	17.00	16.75	7.80	-1.23	68.00	68.49	31.21	-6.43
15.00	17.00	15.70	8.11	-0.83	68.00	64.32	32.42	-4.82
17.00	17.00	15.37	8.30	-0.73	68.00	62.97	33.21	-4.42
20.00	17.00	15.37	8.30	-0.73	68.00	62.97	33.21	-4.42
23.00	17.00	15.37	8.30	-0.73	68.00	62.97	33.21	-4.42
27.00	17.00	15.37	8.30	-0.73	68.00	62.97	33.21	-4.42

Important decisions can be made from the results in Table 1. For example, suppose the intensive care unit (ICU) of a medical facility wants to be advised on reducing the waiting times of their customers. The administrator's goal of treatment may be to ensure that the patient is not delayed by more than 45 units of time. From Table 1, only a maximum of 11 patients will wait in the ICU and no patient will wait for more than 45 units of time when K is 7. Also, the queue length is between (K =) 7 and 11, so the average wait time per patient is 16.41 minutes if a second server is installed. This case is represented in Figure 1a with control limits and Figure 1b for K=7 and maximum wait = 45 units of time.

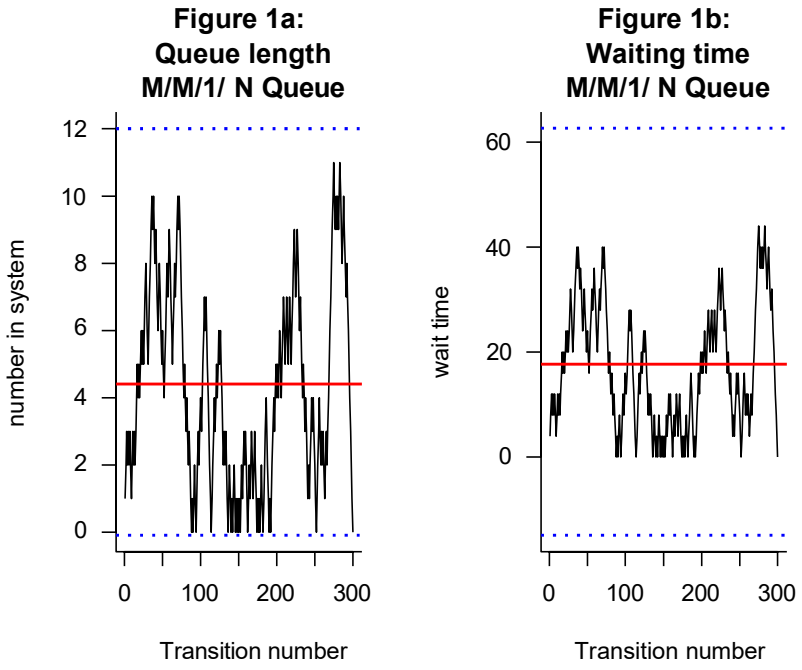

Figure 1. Control charts for monitoring queue lengths based on a simulated data set observed from 150 arrivals leading to 300 transitions

Consider another case; Suppose some health department wants to set its own goals and set a best K-policy, with no reference to control limits. In the case the goal of the administrator of the health Department is to choose a value of K that guarantees a maximum waiting time of about 56 time units, i.e. UCL_W = 59.97, in this case the targeted 56 time units is lower than 59.97. Hence, the better value for K is 10 (see Table 1). Further, the maximum expected queue length is 15 and average waiting time for queue is 23.47 time units.

Similarly, if the maximum waiting time set by the administrator is between 59.97 and 68.49 time units (for example, 62 or 65 time units), the optimal policy is K = 12. That indicates that the second server should be installed whenever the queue length reaches the value 12 and at the same time whenever the queue length falls below 12, the second server should be removed.

3.2. Waiting Time monitoring and G/G/1 Queues under an optimum K-policy

As an extension of the methodology applied in the preceding sections for applications of M/M/1 queues, an analysis of G/G/1 queues operating under a best K policy is now investigated. Here, we extend the above methodology to G/G/1/K-policy queue models with a general inter-arrival distribution, a general service time distributions and one server. The inter-arrival times and service times are generated from the respective distributions given in Tables 2, Table 3, and Table 4. The service times of the second server are generated from a general distribution which is reported in Table 4. This second server happens as a slow speed server as compared with the service rate of the first server reported in the Table 3.

Table 2. Inter arrival (A) distribution, λ=0.25

Time units	2	3	9
Probabilities	0.5	0.25	0.25
Mean $(1/\lambda = 4)$	SD = 2.92	CV(A) = 0.73	

Table 3. Service (S_1) time distribution, μ=0.294

Time units	2	4	5
Probabilities	0.4	0.4	0.2
Mean $(1/\mu = 3.4)$	SD = 1.2	CV(S) = 0.35	

Table 4. Service (S_2) time distribution, μ_1 = 0.04

Time units	20	25	30
Probabilities	0.2	0.6	0.2
Mean $(1/\mu_1 = 25)$			

The steady state characteristics of G/G/1/K-policy queues are then simulated as before for 45 arrivals. Replacing the input data and the exponential parameters by new input values $n = 45$, $\lambda = 0.25$, $\mu = 0.294$, $\mu_1 = 0.04$ and $K \in \{4, 6, 9, 10, 12, 15, 17\}$ in the simulation algorithm used for the M/M/1/K-policy queues, and running the algorithm, we obtained results presented in Table 5.

Table 5. The control limits for G/G/1/N(>14)/K-policy queues when n=45 λ=0.25, μ=0.2941, μ_1=0.04 and K $\in$ {4,6,9,10,12,15, 17}

K	Max L	UCL	CL	LCL	Max W	UCLw	CLw	LCLw
4.00	8.00	6.87	2.85	-0.44	32.00	28.97	11.41	-3.28
6.00	10.00	8.70	4.07	-0.84	40.00	36.31	16.28	-4.85
9.00	11.00	10.83	6.42	-1.37	44.00	44.82	25.68	-6.98
10.00	13.00	11.79	7.00	-1.03	52.00	48.66	28.02	-5.61
12.00	14.00	12.36	7.94	-1.22	56.00	50.92	31.77	-6.40
15.00	14.00	12.39	8.04	-1.28	56.00	51.06	32.16	-6.63
17.00	14.00	12.39	8.04	-1.28	56.00	51.06	32.16	-6.63

The following observations can be easily made from Table 5:

(1) If K varies between 15 and 17, all clients are processed by the G/G /1/K-policy queue with $\lambda = 0.25$ and $\mu = 0.2941$. Corresponding control maps covering all 90 (=45+45) arrivals. Comparing the UCL, Max L with the current value of K=15 or 17 in Table 5. Since a few sample points can fall above the UCL due to the fact max L=14, we conclude that the current K =15 (or 17) value is to be changed to a value less than 15. This means the system is busy throughout the observational period.

(2) Assuming that the experimental unit is a patient, assume that the administrator does not aim to delay the patient for more than 45 minutes. This decision can be reached from Table 5. With K = 9, there are only up to 11 clients waiting in the system, with no patients having to wait for more than 44 minutes. In addition, the average waiting time per patient is 26 minutes, and all clients are handled by the policy G/G /1 /K (= 9) on both servers $\lambda = 0.25$, $\mu = 0.2941$ and $\mu_1 = 0.04$. If, instead, the administrator chooses a K value with a maximum timeout of around 36 minutes, then the better value for K is 6, the maximum expected queue length is 10, and the average wait time is 16.28 minutes

4. Conclusion

Using simulated Models for M/M/1/K-policy queues and G/G/1/K-policy queues, we have developed numerically tractable control limits for monitoring the waiting times of customers and provided numerical illustrations. The modified control chart proposed in this investigation ensures that queueing models require relatively small amount of data and simple algorithms.

We now conclude with the following recommendation: In order to deal with management of resources in any kind of service facilities like healthcare facilities, the proposed control chart for monitoring the waiting time of customers/patients is strongly recommended over other tools. This type of chart can be used as a practical tool to quickly assess job latency and compare different options to provide the best QoS. Overall, medical service resources are limited, but demand is often high.

Future studies will develop smart, easy-to-implement policies for asset management. Further, each simulation model presented in this article can be used in inventory, web services, call centers, insurance management, reliability, and many other service operations.

Acknowledgment.

References

[1] A.O. Allen, Probability, statistics, and queueing theory, with computer science applications. New York, Academic Press.1978

[2] D. Gross and Harris, S.M. Fundamentals of Queueing Theory, 3rd Edition. Wiley, New York, NY,1998

[3] J.J. Hunter, *Filtering of Markov Renewal Queues, I: Feedback Queues* Adv. Appl. Prob. Vol. 15, No. 2, pp. 349-375, 1983

[4] J.J. Hunter, *Filtering of Markov Renewal Queues, II: Feedback Queues* Adv. Appl. Prob. Vol. 15, No. 2, pp. 376-393, 1983

[5] D.C. Montgomery, Introduction to Statistical Quality Control 5th Edition. Wiley, New York, NY, 2004

[6] H. Shore, *General control charts for attributes*, IIE Transactions, 32, 1149–1160, 2000

[7] H. Shore, *Control charts for queue lengths in a G/G/s system*, IIE Transactions, 38, 1117–1130, 2006

[8] R. Sivasamy, and S. Jayanthi, *Charts for Detecting Small Shift*, Economic Quality Control, Vol 8, No.1, pp 1-8, 2003

[9] R. Sivasamy, N.Thillaigovindan, G. Paulraj, and N. Paranjothi, *Quasi birth and death processes of two-server queues with stalling* , OPSEARCH, 56:739-756.,2019

[10] R. Sivasamy,*An M/M(μ_1)+M(μ_2)/1 Queue with a State Dependent N-Policy*, Journal of Xi'an University of Architecture & Technology, (ISSN No:1006-7930), Volume XIII, Issue I , pp 111-124, 2021

[11] X. Zhao, and K. Gilbert, *A statistical control chart for monitoring customer waiting time* Int. J. Data Analysis Techniques and Strategies, Vol.7 No.3, pp.301 – 321, 2015