Controlling the workload of $M/G/1$ queues via the q-policy

Val Andrei Fajardo[a], Steve Drekic[a,*]

[a] *Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, N2L 3G1, Canada*

Abstract

We consider a single-server queueing system with Poisson arrivals and generally distributed service times. To systematically control the workload of the queue, we define for each busy period an associated timer process, $\{R(t), t \geq 0\}$, where $R(t)$ represents the time remaining before the system is closed to potential arrivals. The process $\{R(t), t \geq 0\}$ is similar to the well-known workload process, in that it decreases at unit rate and consists of up-jumps at the arrival instants of admitted customers. However, if X represents the service requirement of an admitted customer, then the magnitude of the up-jump for the timer process occurring at the arrival instant of this customer is $(1-q)X$ for a fixed $q \in [0, 1]$. Consequently, there will be an instant in time within the busy period when the timer process hits level zero, at which point the system immediately closes and will remain closed until the end of the current busy period. We refer to this particular blocking policy as the q-policy. In this paper, we employ a level crossing analysis to derive the Laplace-Stieltjes transform (LST) of the steady-state waiting time distribution of serviceable customers. We conclude the paper with a numerical example which shows that controlling arrivals in this fashion can be beneficial.

Keywords:
Queueing; customer blocking; level crossing analysis; $M/G/1$ queue with accumulating priority.

1. Introduction

We study an $M/G/1$-type queueing model in which the arrival process is controlled by a system manager so as to decrease the lengths of the general busy period. In some applications, for example, a system manager may be more inclined to regularly decrease the overall length of the busy period

[*] Corresponding author. Tel.: 1 519 888 4567 ext. 35550
Email addresses: `andrei.fajardo@uwaterloo.ca` (Val Andrei Fajardo), `sdrekic@uwaterloo.ca` (Steve Drekic)

if it is the case that the server/machine becomes highly susceptible to expensive breakdowns after operating for extended periods of time. These breakdowns can be costly both in terms of the repair costs and the opportunity costs due to closures of the system. To alleviate the risk of incurring an expensive breakdown, a system manager may want to rest the server/machine during *closedown* periods on a regular basis. In addition, cost-effective *maintenance checks* can be performed during these rest periods to ensure the long-run functionality of the machine.

In this paper, we present one such policy which would allow a system manager to control the busy period lengths. Specifically, during each busy period, the control is exercised by closing the system to potential customers over a constant proportion of the overall busy period. The flexibility to disallow (or to block) customers from entering the system may be desirable if, for instance, a *holding cost* for customers during their sojourn in the system exists. The main focus of our research is to study the effect of the new policy, which we refer to as the *q-policy*, on various performance measures of interest such as the length of busy periods and the wait of serviceable customers.

The literature on the optimal design and control of queueing systems is quite extensive. In regards to the arrival control of queueing systems, the usual goal is to find the optimal policy which maximizes (or minimizes) a specific objective function. In the seminal paper by Naor [9], an $M/M/1$-type queueing system is studied where the arrival process is controlled by the administration of a toll charge for arriving customers. In particular, customers receive a fixed reward K upon successful service but also incur a holding cost h per unit time spent in the system. Naor studies the optimal policies from two perspectives, namely: (i) individual optimization, where the objective function is the individual expected net benefit rate function, and (ii) social optimization, where the objective function is the expected overall net benefit rate function. Naor assumes that the optimal policies for both problems is of the critical number form (i.e., customers are accepted for service if the number of customers currently occupying the system is less than the critical number), and this form of optimal policy can be validated through the use of Markov decision processes (see Stidham [14] and references therein). Under this framework, Naor establishes a key result which states that an individually optimal policy admits more customers than its counterpart, the socially optimal policy.

Naor's work inspired several other researchers to consider various generalizations for both the model and the net benefit structure. Rosenshine and Rue [11] considered Naor's model and studied the effect of the arrival rate on the parameters for both kinds of optimal policies. Yechiali [16] extended Naor's work by relaxing the assumption of the arrival process to be merely a renewal process. The $M/M/s$ variant was considered by Knudsen [8] where Naor's main result was shown to still hold true. Doshi [4] considered the continuous-time arrival control of an $M/G/1$ queueing system which operated under a policy that opened and closed the system to potential arrivals depending on the level of the workload. In Johansen and Stidham [5], the authors showed that Naor's main result actually holds true under a set of fairly general conditions (e.g., dependent arrivals, batch arrivals, and random rewards). For excellent surveys of the literature, we refer the interested reader to Stidham [13, 14]. To the best of our knowledge, the q-policy presented here has not been previously studied.

The optimal policies found by these researchers has usually resulted in the formulation of threshold-form policies (i.e., thresholds for the number of customers in the system or for the residual workload). We emphasize, however, that our focus is not one that searches for an optimal policy which maximizes a specific objective function, but instead analyzes the effects of a given policy which aims to lessen the workload of a system. Nonetheless, we do, in Section 6, formulate an optimization problem which illustrates that, in certain situations, the reduction of the busy cycle lengths via the q-policy can result in increased profits.

The rest of the paper is organized as follows. In Section 2, we introduce the queueing model and the q-policy. Section 3 is devoted to the study of the busy period as well as some fundamental steady-state probabilities associated with the system. The steady-state distribution of the waiting time of serviceable customers is analyzed in Section 4 by virtue of the level crossing methodology. In Section 5, we present a model which enables a system manager to block customers during busy periods similar to the q-policy, but has the property that it does not require knowledge of the service times upon arrival. Finally, following the numerical example in Section 6, we offer some concluding remarks and a discussion regarding potential future work in Section 7.

2. The model and the q-policy

We consider a queueing system which is of $M/G/1$-type. We assume that the Poisson arrival rate of customers to the system is $\lambda > 0$. If the system is *open* (i.e., accepting new customers) when a customer arrives, then this customer joins the queue. Otherwise, the customer is lost and unrecoverable. In Section 6 of this paper, we present an optimization problem which shows that in certain situations it may be desirable to block or prevent customers from entering the system.

Let $\{X_i, i = 1, 2, \ldots\}$ denote the sequence of independent and identically distributed (iid) customer service times having common mean $\mu = \mathbb{E}(X_i)$ and common second moment $\gamma = \mathbb{E}(X_i^2)$. Similar to the model studied by Johansen and Stidham [5], the customer service times are assumed to be known to the server (or system manager) immediately upon a customer's entry to the system. We denote the corresponding distribution function and Laplace-Stieltjes transform (LST) by

$$B(x) = \mathbb{P}(X_i \leq x) \quad \text{and} \quad \widetilde{B}(s) = \int_0^\infty e^{-sx} \mathrm{d}B(x), \tag{1}$$

respectively. Service is conducted by order of arrival (i.e., first-come-first-serve or FCFS for short). We denote the traffic intensity of the classical (i.e., unblocked) $M/G/1$ system, as usual, by $\rho = \lambda\mu$. Note that we reserve the notation $\bar{B}(x) = \mathbb{P}(X_i > x)$ for the complementary distribution function of $B(\cdot)$, and further that this style of notation will also be adopted for other complementary distribution functions throughout the paper.

Before we formally introduce the q-policy, we recall that for an arbitrary busy period of the classical (work-conserving) $M/G/1$ queue, any customer who arrives during this busy period will always be admitted for service (i.e., they will eventually be served in this busy period). However, suppose that a system manager would like to restrict (or control) the arrival process during a busy period, so that the system is not obligated to serve all customers who arrive during the busy period. In such a situation, a system manager could, for intervals of time within the busy period, close the system to potential arrivals. A blocking policy provides a set of guidelines which allows a system manager to administrate the openings and closures of the system. We denote such a policy in general by $\pi(t)$, where $\pi(t) = 1$ implies that the system is open at time t, and similarly $\pi(t) = 0$ implies that the system is closed at time t. An example of such a blocking policy is the q-policy, denoted by $\pi_q(\cdot)$, which we define next.

Definition 2.1 (The q-policy)**.** *Without loss of generality, assume that a customer arrives to an empty queue at time $\tau_1 = 0$, thereby initiating the start of a busy period. For all $t \geq 0$ during this busy period, we define the process $\{R(t), t \geq 0\}$, which is similar to the workload process. In particular, for $0 \leq q \leq 1$:*

1. $R(0) = (1 - q)X_1$, *where X_1 is initial customer's service time.*
2. $R(t)$ *decreases at unit rate unless the process is at level 0.*

3. *For the sequence of customer arrival epochs, $\{\tau_i, i = 2, 3, \ldots\}$, during this busy period,*

$$R(\tau_i) = \begin{cases} R(\tau_i^-) + (1-q)X_i & \textit{if } R(\tau_i^-) > 0, \\ 0 & \textit{if } R(\tau_i^-) = 0, \end{cases} \tag{2}$$

where $R(t^-) = \lim_{\epsilon \to 0} R(t - \epsilon)$.

Then, for all $t \geq 0$ during this busy period,

$$\pi_q(t) = \begin{cases} 1 & \textit{if } R(t) > 0, \\ 0 & \textit{if } R(t) = 0. \end{cases} \tag{3}$$

Remark 2.1. *The process $\{R(t), t \geq 0\}$ acts as a timer for the busy period. That is, $R(t)$ represents the time remaining, at time t, before the system is closed to potential arrivals.*

Figure 1 illustrates a busy period under the q-policy. Here, at some time during the servicing of the third customer, C_3, the timer becomes drained (i.e., $R(\cdot)$ hits level 0), at which point the system becomes closed for potential arrivals. Hence, both customers C_5 and C_6 are blocked from entering the system. It is important to note that, although the system is closed at this point, the server must still complete the servicing of C_3 and C_4. In other words, the busy period terminates when all admitted customers have been fully served. Moreover, the end of the busy period signals the reopening of the system and the commencement of the ensuing idle period which ends at the next customer arrival instant. The busy period and the subsequent idle period together form a busy cycle.

Clearly, under the q-policy, the resulting busy periods are stochastically smaller than those corresponding to a system not implementing any sort of blocking policy. It is also apparent that if we set $q = 0$, then $\{R(t), t \geq 0\}$ exactly becomes the workload process during a busy period in the classical $M/G/1$ queue. In fact, a blocking proportion equal to zero simply implies that no customers are blocked from service, and thus the resulting model is equivalent to the classical $M/G/1$ queue. Moreover, we obtain the $M/G/1/1$ queue as a special case when $q = 1$.

3. The busy period and steady-state probabilities

In this section, we first establish a functional equation for the LST corresponding to the distribution of the busy period duration operating under the q-policy. Let T be the length of such a busy period, whose distribution function and LST are denoted by $G(x)$ and $\widetilde{G}(s)$, respectively.

To derive the LST of T, we note that the order in which serviceable customers are served does not, in any way, affect the duration of the busy period. As in the classical case, this important observation leads to the derivation of a functional equation for $\widetilde{G}(s)$. We now introduce a new service discipline which we refer to as the *q-restricted last-come-first-serve (q-restricted LCFS for short)* discipline. First of all, recall that $\{R(t), t \geq 0\}$ consists of up-jumps at the arrival epochs of each serviceable customer, and further that the magnitude of the jump is equal to the service time of the customer multiplied by $(1-q)$. Let us refer to these entities simply as the unblocked portions of the service times. Now, the order of service determined by the q-restricted LCFS discipline is precisely the order of service obtained by applying the usual LCFS discipline to a system in which the unblocked portions are effectively considered as the actual service times (i.e., $(1-q)X_i$ instead of X_i).

Figure 2 demonstrates the q-restricted LCFS discipline in a typical busy period. Again, we determine the order of service under this discipline by effectively considering the unblocked portions as the actual service times. Specifically, in Figure 2, one can determine the order of service by projecting the arrival epochs to the axis a^* and applying the usual LCFS discipline. Moreover, under the q-restricted LCFS discipline, we see that the interval of time during which $R(t)$ is positive (i.e., the system is open to accepting new customers) can be decomposed into smaller, well-understood sub-intervals of time. Indeed, these sub-intervals are merely the acceptance periods of their corresponding sub-busy periods. For example, in Figure 2, C_4 generates a sub-busy period in which C_5 and C_6 both are serviced; the length of the acceptance period for this sub-busy period is equal to $(1-q) \times (X_4 + X_5 + X_6)$. It is clear that these sub-busy periods are identically distributed to the overall busy period (generated by C_1). However, we do note that in the intermediate sub-busy periods (i.e., sub-busy periods generated by C_4 and C_3 in Figure 2), customers who fail to arrive in their acceptance periods are not blocked from the system, but instead are serviced in the next sub-busy period.

Theorem 3.1. *If $\lambda^{(q)} = \lambda(1-q)$ and $\rho^{(q)} = \lambda^{(q)}\mu < 1$, then T has a proper (i.e., non-defective) distribution and its corresponding LST satisfies the functional equation*

$$\widetilde{G}(s) = \widetilde{B}(s + \lambda^{(q)}(1 - \widetilde{G}(s))). \tag{4}$$

Proof. Similar to the LST derivation of the busy period duration in the classical $M/G/1$ queue (e.g., see Kleinrock [7, Section 5.8]), we invoke the fact that T is independent of the service discipline, so long as it is a work-conserving one. Kleinrock's derivation involves the usual LCFS discipline, but here, we employ the q-restricted LCFS discipline. Define N to be the number of customers who arrive during the unblocked portion of the initial customer's service time. As discussed above, each of the N customers generates a sub-busy period of their own which is identically distributed to the overall busy period and, moreover, is mutually independent from the others.

Conditioning on both $N = n$ and the first service time $X_1 = x$, we obtain

$$\mathbb{E}(e^{-sT}|X_1 = x, N = n) = e^{-sx}\big(\widetilde{G}(s)\big)^n. \tag{5}$$

Given $X_1 = x$, N is Poisson distributed with rate $\lambda^{(q)}x$, and this leads to

$$\mathbb{E}(e^{-sT}|X_1 = x) = e^{-sx}e^{-\lambda^{(q)}x}\sum_{n=0}^{\infty}\frac{\big(\lambda^{(q)}x\widetilde{G}(s)\big)^n}{n!} = e^{-x(s+\lambda^{(q)}-\lambda^{(q)}\widetilde{G}(s))}. \tag{6}$$

Lastly, removing the condition on X_1 immediately yields

$$\widetilde{G}(s) = \mathbb{E}(e^{-sT}) = \widetilde{B}(s + \lambda^{(q)}(1 - \widetilde{G}(s))), \tag{7}$$

and the result is proven. □

As in the classical case, we are left with an implicit expression for the LST of T. Nonetheless, we are still able to obtain the moments of T through successive differentiation. In particular, the first two moments of T are:

$$\mathbb{E}(T) = \frac{\mu}{1-\rho^{(q)}}, \tag{8}$$

$$\mathbb{E}(T^2) = \frac{\gamma}{(1-\rho^{(q)})^3}. \tag{9}$$

Remark 3.2. *Theorem 3.1 implies that the busy period under the q-policy is distributed equivalently to the busy period of a classical $M/G/1$ queue with arrival rate $\lambda^{(q)}$ and service time distribution $B(\cdot)$. Furthermore, the busy period is also equivalently distributed to the busy period of the following $M/G/1$ system with a Bernoulli-type blocking policy:*

(i) *customers arrive according to a Poisson process with rate $\lambda > 0$;*

(ii) *at each customer arrival epoch, the server conducts a Bernoulli experiment, where with probability $(1-q)$ the customer is admitted for service, and with probability q the customer is blocked.*

A commonality of this model with the system under the q-policy is that during busy periods, the probability that an arriving customer is blocked from entering the system is precisely q.

We next establish the form of the probability generating function (pgf) for N_{bp}, the number of customers served in a busy period. We define $m(z) = \mathbb{E}(z^{N_{bp}})$ to be the pgf of N_{bp}. Like the duration of the busy period T, the number served in a busy period is unaffected by the order of service. Hence, by implementing the q-restricted LCFS discipline, we obtain

$$\mathbb{E}(z^{N_{bp}}|N = n) = \mathbb{E}(z^{1+M_1+M_2+\cdots+M_n}), \tag{10}$$

where N is the number of customers in the initial queue (i.e., those customers arriving during the unblocked portion of the initial customer's service time) and M_i denotes the number of customers served in the i-th customer's sub-busy period. By independence, we have

$$\mathbb{E}(z^{N_{bp}}|N = n) = z\big(m(z)\big)^n. \tag{11}$$

It immediately follows, by removing the condition on N, that

$$m(z) = z\widetilde{B}\left(\lambda^{(q)}(1 - m(z))\right), \tag{12}$$

from which the first moment of N_{bp} is clearly given by

$$\mathbb{E}(N_{bp}) = \frac{1}{1 - \rho^{(q)}}. \tag{13}$$

To conclude this section, we shift our focus to the derivation of some key steady-state probabilities of the system, namely:

$$\begin{aligned}
P_I &\equiv \text{steady-state probability the server is idle;}\\
P_B &\equiv \text{steady-state probability the server is busy;}\\
P_{B,0} &\equiv \text{steady-state probability the server is busy and the system is closed;}\\
P_{B,1} &\equiv \text{steady-state probability the server is busy and the system is open.}
\end{aligned}$$

To obtain these probabilities, we apply the theory of regenerative processes (e.g., see Kao [6, Section 3.6]). Define a busy cycle, D, to consist of a busy period T and the ensuing idle period I (i.e., $D = T + I$). Clearly, the set of regeneration points associated with D are the epochs defined by busy period commencements. Thus, from elementary renewal theory, we readily obtain:

$$P_I = \frac{\mathbb{E}(I)}{\mathbb{E}(D)} = \frac{1 - \rho^{(q)}}{1 + \rho q}, \tag{14}$$

$$P_B = \frac{\mathbb{E}(T)}{\mathbb{E}(D)} = \frac{\rho}{1 + \rho q}, \tag{15}$$

$$P_{B,0} = qP_B = \frac{\rho q}{1 + \rho q}, \tag{16}$$

$$P_{B,1} = (1 - q)P_B = \frac{\rho^{(q)}}{1 + \rho q}. \tag{17}$$

4. Steady-state wait of serviceable customers

4.1. The workload and virtual wait processes

The motivation for our study of the virtual wait process stems from the well-known fact that for $M/G/1$-type queues, the distributions of virtual wait and actual wait are equivalent in steady-state. In what follows, we denote the (unfinished) workload process under a q-policy by $\{U_q(t), t \geq 0\}$, whereas the virtual wait process is denoted by $\{W_q(t), t \geq 0\}$.

Obviously, $\{U_0(t), t \geq 0\}$ and $\{W_0(t), t \geq 0\}$ are the corresponding workload and virtual wait processes for the classical $M/G/1$ system. Now, for times $t > 0$ when the system is open (i.e., $\pi_q(t) = 1$), one notes that $U_q(t)$ behaves in the same manner as $U_0(t)$ in that:

(i) $U_q(t)$ decreases at unit rate, except during times of idleness;

(ii) $U_q(t)$ up-jumps at customer arrival epochs, with the magnitude of the jumps being equal to the arriving customer's service time.

On the other hand, for times $t > 0$ when $\pi_q(t) = 0$, we have that $U_q(t)$ decreases at unit rate. In particular, if $t_* > 0$ is such that $\pi_q(t_*) = 0$ and $\pi_q(t_*^-) = 1$, then starting from time t_*, the workload depletes at unit rate until it hits level 0. Now, similar to how $\{U_0(t), t \geq 0\}$ and $\{W_0(t), t \geq 0\}$ are equivalent processes, during times t when the system is open, the processes $\{W_q(t), t \geq 0\}$ and $\{U_q(t), t \geq 0\}$ are also equivalent. However, the virtual wait process is further complicated by the fact that during a closure period for the system, the process is essentially undefined (i.e., does not exist).

Figure 3 depicts the sample paths of both processes for three consecutive busy periods of the system. The grey-shaded regions correspond to the times during which the system is closed (i.e.,

$\pi_q(t) = 0$), and thus, also represents the times when $W_q(t)$ is undefined. Customer arrival epochs are marked on the time axis with diamond symbols, and observe that both processes up-jump at arrivals occurring only during times when the system is open. As is also evident from Figure 3, the instant in time at which the system becomes closed during a busy period is exactly the same instant in time that $W_q(t)$ (or equivalently $U_q(t)$) hits level qT_i, where T_i is the duration of the i-th busy period. In what follows, we define $G_q(x) = \mathbb{P}(qT \leq x) = G(x/q)$ as well as $\widetilde{G}_q(s) = \mathbb{E}(e^{-s(qT)}) = \widetilde{G}(sq)$.

In order to study the wait of admitted customers, it is clear that we must analyze the virtual wait process only during times of its existence. Hence, we introduce the *censored* virtual wait process $\{\mathcal{W}_q(t), t \geq 0\}$, as illustrated in Figure 4. This process can be considered as $\{W_q(t), t \geq 0\}$ with the censorship (or removal) of the periods of non-existence. Indeed, by simply removing these periods, the resulting censored process will have a different time clock than the non-censored version. However, due to the memoryless property of the Poisson arrival process, the analysis of $\{W_q(t), t \geq 0\}$ during its times of existence must be equivalent to the analysis of $\{\mathcal{W}_q(t), t \geq 0\}$.

As is evident in Figure 4, the sample path never continuously hits level 0 (unless $q = 0$), but instead always down-jumps to level 0. Furthermore, the magnitude of these down-jumps have distribution $G_q(\cdot)$. This simple observation allows us to derive the steady-state integral equation for the probability density function (pdf) of the virtual wait (during times of its existence).

4.2. Steady-state integral equation for the pdf of the virtual wait

We characterize the transient distribution of the censored virtual wait by the functions

$$\left.\begin{aligned} F_t(x) &= \mathbb{P}(\mathcal{W}_q(t) \leq x), \quad x \geq 0,\ t \geq 0; \\ f_t(x) &= \tfrac{\partial}{\partial x} F_t(x), \quad x > 0,\ t \geq 0; \\ P_0(t) &= \mathbb{P}(\mathcal{W}_q(t) = 0), \quad t \geq 0. \end{aligned}\right\} \tag{18}$$

The steady-state distribution is obtained by letting $t \to \infty$ in the functions of Eq. (18), resulting in

$$F(x) = \lim_{t\to\infty} F_t(x), \quad f(x) = \lim_{t\to\infty} f_t(x), \quad \text{and} \quad P_0 = \lim_{t\to\infty} P_0(t). \tag{19}$$

When appropriate, we will use $f(x;q)$ equivalently as $f(x)$ to specify the value of q being used in the blocking policy. Also, in what follows, we extend the definition of $P_0(t)$ by defining $P_0(t) = 0$ for all $t < 0$.

If we consider the censored virtual wait process, let $\mathcal{U}_t(x)$ and $\mathcal{D}_t(x)$ denote the number of sample path up- and down-crossings of level x, respectively, during the time interval $(0,t)$. Moreover, let $\mathcal{D}_t^c(x)$ (and $\mathcal{D}_t^j(x)$) denote the number of continuous down-crossings (jump down-crossings) of level x in the time interval $(0,t)$. Clearly,

$$\mathcal{D}_t(x) = \mathcal{D}_t^c(x) + \mathcal{D}_t^j(x). \tag{20}$$

Correspondingly, we remark that $\mathcal{U}_t^j(x) = \mathcal{U}_t(x)$ for all $x \geq 0$. The ingenuity of the level crossing methodology lies in the principle of set balance (e.g., see Brill [2, Section 2.4.6]). That is, in steady-state, the up-crossing and down-crossing rates of level x are equal:

$$\lim_{t\to\infty} \frac{\mathbb{E}(\mathcal{D}_t(x))}{t} = \lim_{t\to\infty} \frac{\mathbb{E}(\mathcal{U}_t(x))}{t}, \tag{21}$$

$$\lim_{t\to\infty} \frac{\mathcal{D}_t(x)}{t} \overset{a.s.}{=} \lim_{t\to\infty} \frac{\mathcal{U}_t(x)}{t}, \tag{22}$$

where "*a.s.*" means almost surely, or with probability 1. Thus, to develop an integral equation for the steady-state pdf of the virtual wait (provided it exists), we must establish both the up- and down-crossing rates of level x. The next theorem provides the means to do so.

Theorem 4.1. *The up- and down-crossing rates of level x are given by*

$$\lim_{t\to\infty} \frac{\mathbb{E}(\mathcal{U}_t(x))}{t} = \lambda \bar{B}(x) P_0 + \lambda \int_{y=0}^{x} \bar{B}(x-y) f(y)\,\mathrm{d}y, \quad x > 0, \tag{23}$$

$$\lim_{t\to\infty} \frac{\mathbb{E}(\mathcal{D}_t^c(x))}{t} = f(x), \quad x > 0, \tag{24}$$

$$\lim_{t\to\infty} \frac{\mathbb{E}(\mathcal{D}_t^j(x))}{t} = \lambda P_0 \bar{G}_q(x), \quad x > 0. \tag{25}$$

Proof. The proof for both the up-crossing rate and the continuous down-crossing rate (i.e., Eq. (23) and Eq. (24)) can be derived in the exact same manner as for the classical M/G/1 virtual wait process (e.g., see Brill [2, Theorems 3.3 and 3.4]). Thus, we omit their proofs and only prove Eq. (25).

To establish Eq. (25), we consider $\mathbb{E}(\mathcal{D}_{t+h}^j(x) - \mathcal{D}_t^j(x))$ for very small h. Clearly, $\mathcal{D}_{t+h}^j(x) - \mathcal{D}_t^j(x)$ represents the number of jump down-crossings of level x in a small interval of size h. Thus, $\mathcal{D}_{t+h}^j(x) - \mathcal{D}_t^j(x)$ can take values in the set of non-negative integers. Concerning the expectation of this quantity, we can obviously omit the case of it being equal to 0. In addition, it is not difficult to see that $\mathbb{P}(\mathcal{D}_{t+h}^j(x) - \mathcal{D}_t^j(x) \geq 2) = o(h)$.

Therefore, the only event we must really consider is when $\mathcal{D}^j_{t+h}(x) - \mathcal{D}^j_t(x) = 1$. This event implies that a busy period initiates before time t, and also that sometime within the time interval $(t, t+h)$, the server finishes processing all but the last q-th proportion of the workload of this busy period (assume again that the system is empty at time 0). Conditioning on the length of this busy period leads to

$$\mathbb{P}(\mathcal{D}^j_{t+h}(x) - \mathcal{D}^j_t(x) = 1) = \int_{y=x/q}^{\infty} \lambda h P_0(t - (1-q)y)\, \mathrm{d}G(y) + o(h). \tag{26}$$

The above result is obtained by recalling that the sample path immediately jumps down to level 0 as soon as the censored virtual wait process hits level qy. In particular, a jump down-crossing of level x will occur only if the busy period duration y is such that $qy > x$. Thus,

$$\mathbb{E}(\mathcal{D}^j_{t+h}(x) - \mathcal{D}^j_t(x)) = \int_{y=x/q}^{\infty} \lambda h P_0(t - (1-q)y)\, \mathrm{d}G(y) + o(h). \tag{27}$$

Dividing the above equality by h and letting $h \to 0$, we subsequently obtain

$$\frac{\partial}{\partial t}\mathbb{E}(\mathcal{D}^j_t(x)) = \lambda \int_{y=x/q}^{\infty} P_0(t - (1-q)y)\, \mathrm{d}G(y). \tag{28}$$

It then follows (since $\mathbb{E}(\mathcal{D}^j_0(x)) = 0$) that

$$\mathbb{E}(\mathcal{D}^j_t(x)) = \lambda \int_{s=0}^{t} \int_{y=x/q}^{\infty} P_0(s - (1-q)y)\, \mathrm{d}G(y)\mathrm{d}s. \tag{29}$$

Finally, Eq. (25) follows because $\lim_{s\to\infty} \int_{y=x/q}^{\infty} P_0(s - (1-q)y)\, \mathrm{d}G(y) = P_0 \bar{G}_q(x)$ via the dominated convergence theorem (e.g., see Parzen [10, Section 6-10]). □

Corollary 4.2. *If $\rho^{(q)} < 1$, then*

$$\lim_{t\to\infty} \frac{\mathcal{D}^c_t(x)}{t} \overset{a.s.}{=} f(x), \quad x \geq 0 \quad \textit{and} \quad \lim_{t\to\infty} \frac{\mathcal{D}^j_t(x)}{t} \overset{a.s.}{=} \lambda P_0 \bar{G}_q(x), \quad x \geq 0. \tag{30}$$

Proof. By the memoryless property of Poisson arrivals, both $\{\mathcal{D}^j_t(x), t \geq 0\}$ and $\{\mathcal{D}^c_t(x), t \geq 0\}$ are (delayed) renewal processes. The desired result then follows from a well-known limiting theorem from renewal theory (e.g., see Parzen [10, Theorem 3A]). □

From Theorem 4.1, we can obtain an integral equation for the steady-state pdf of the virtual wait (provided it exists). Specifically, by using Eq. (23) through Eq. (25) along with the balance rate equation given by Eq. (21), we end up with

$$f(x) + \lambda P_0 \bar{G}_q(x) = \lambda \bar{B}(x) P_0 + \lambda \int_{y=0}^{x} \bar{B}(x-y) f(y)\mathrm{d}y. \tag{31}$$

Remark 4.3. *An attractive feature of the level crossing technique is that we are able to intuitively explain each of the individual algebraic components of the resulting integral equation. We note that Eq.* (31) *is almost identical to that for the classical $M/G/1$ virtual wait, with only the addition of the second term on the left-hand side of the equality sign. This term (the jump down-crossing rate of level x) can be explained as follows: the rate that a busy period initiates is λP_0, where the proportion of these busy periods that result in a jump down-crossing of level x is $\bar{G}_q(x) = \mathbb{P}(qT > x)$. The other terms are interpreted in the same manner as for the classical $M/G/1$ virtual wait.*

Remark 4.4. *Letting $x \to 0$ in Eq.* (31) *results in $f(0^+) = 0$ where, in general, $f(z^+) = \lim_{\epsilon \to 0} f(z + \epsilon)$. This result is as expected since $f(x)$ represents the continuous down-crossing rate of level x, and under the q-policy, any sample path of $\{\mathcal{W}_q(t), t \geq 0\}$ never down-crosses level* 0 *continuously – it always jumps down to level 0.*

To find P_0, we use the normalizing condition $\int_0^\infty f(x)\,\mathrm{d}x + P_0 = 1$. Now,

$$\int_0^\infty f(x)\,\mathrm{d}x = \lambda P_0(\mu - \mathbb{E}(qT)) + \lambda \int_{y=0}^\infty \int_{x=y}^\infty \bar{B}(x-y) f(y)\,\mathrm{d}x\,\mathrm{d}y, \tag{32}$$

which implies that $\int_0^\infty f(x)\mathrm{d}x(1 - \lambda\mu) = \lambda P_0(\mu - q\mathbb{E}(T))$. Using Eq. (8), we get

$$\begin{aligned} \int_0^\infty f(x)\,\mathrm{d}x &= P_0 \frac{\rho(1 - \rho^{(q)} - q)}{(1-\rho)(1-\rho^{(q)})} \\ &= P_0 \frac{\rho(1-q)(1-\rho)}{(1-\rho)(1-\rho^{(q)})} \\ &= P_0 \frac{\rho^{(q)}}{1-\rho^{(q)}}. \end{aligned} \tag{33}$$

Therefore, $P_0 = 1 - \rho^{(q)}$. This result too is as expected, since P_0 represents the long-run proportion of time that the server is idle conditional on the system being open for arrivals (i.e., conditional on the existence of the virtual wait process). From Eq. (14) and Eq. (17), the long-run fraction of time the system accepts new customers is $P_I + P_{B,1} = (1 + \rho q)^{-1}$. Thus, $P_0 = P_I/(1 + \rho q)^{-1}$.

From Eq. (31), we can readily obtain the LST of the steady-state actual wait of serviceable customers.

Theorem 4.5. *The LST of the steady-state waiting time of serviceable customers is*

$$\widetilde{W}(s) = \frac{(1 - \rho^{(q)})(s - \lambda + \lambda \widetilde{G}(qs))}{s - \lambda + \lambda \widetilde{B}(s)}. \tag{34}$$

Proof. Clearly, $\widetilde{W}(s) \equiv \int_0^\infty e^{-sx} \mathrm{d}F(x) = P_0 + \int_0^\infty e^{-sx} f(x)\mathrm{d}x$. Thus, the desired result is readily obtained by first multiplying both sides of Eq. (31) by e^{-sx} and then integrating over $x \in (0, \infty)$. $\square$

Alternatively, we can express the above LST as follows:

$$\widetilde{W}(s) = (1 - \rho^{(q)}) + \rho^{(q)} \widetilde{W}_+(s), \tag{35}$$

where W_+ represents the stationary waiting time for those customers who are admitted for service upon their arrival but incur a positive wait time prior to entering service. We refer to W_+ as the *delayed* waiting time whose LST $\widetilde{W}_+(s)$ is given by

$$\widetilde{W}_+(s) = \frac{(1 - \rho^{(q)})(\widetilde{G}(qs) - \widetilde{B}(s))}{\mu(1 - q)(s - \lambda + \lambda \widetilde{B}(s))}. \tag{36}$$

One can obtain the first moment of waiting time as usual by differentiating $\widetilde{W}(s)$ and twice applying L'Hôpital's rule. After some algebra, we acquire the following illuminating form of the mean waiting time:

$$\mathbb{E}(W) = \frac{\lambda^{(q)} \gamma}{2(1 - \rho^{(q)})} \times (1 + \sigma(q)), \tag{37}$$

where $\sigma(q) = q/(1 - \rho^{(q)})$. We observe that the first term is equal to the average waiting time in the classical $M/G/1$ queue with arrival rate $\lambda^{(q)}$ and service time distribution $B(\cdot)$. Clearly, $\sigma(q) \geq 0$ since $0 \leq q \leq 1$, which implies that a system under the q-policy has a greater average waiting time than a classical $M/G/1$ system with the aforementioned parameters.

In addition, the first moment of waiting time can be rewritten as

$$\mathbb{E}(W) = \frac{\lambda \gamma}{2} \times \kappa(q), \quad 0 \leq q \leq 1, \tag{38}$$

where

$$\kappa(q) = \frac{1 - q}{1 - \rho^{(q)}}(1 + \sigma(q)) = \frac{(1 - q)(1 - \rho^{(q)} + q)}{(1 - \rho^{(q)})^2}. \tag{39}$$

Differentiating $\kappa(q)$ (with respect to q) yields

$$\kappa'(q) = -\frac{2q}{(1 - \rho^{(q)})^3}. \tag{40}$$

Therefore, for $\rho^{(q)} < 1$, $\mathbb{E}(W)$ is a decreasing function of q. Considering $\mathbb{E}(W)$ at the extreme values of q, we see that for $q = 0$, $\mathbb{E}(W) = \lambda\gamma(1 - \rho)^{-1}/2$ which is the classical $M/G/1$ average waiting time without a blocking policy, and for $q = 1$, $\kappa(1) = 0$ so that $\mathbb{E}(W) = 0$. The latter result is due to the fact that during busy periods, the system is closed to all potential arrivals, and above that, only customers who arrive to an idle server will be served (and these customers experience zero wait).

Finally, we close this analysis by considering the first moment of delayed waiting time, namely:

$$\mathbb{E}(W_+) = \frac{\mathbb{E}(W)}{\rho^{(q)}} = \frac{\gamma}{2\mu} \times \frac{1-\rho^{(q)}+q}{(1-\rho^{(q)})^2}, \quad 0 \leq q \leq 1. \tag{41}$$

It is indeed true that for $q = 1$, there is zero probability that an arbitrary customer will experience positive wait; however, as $q \to 1$, we see that $\mathbb{E}(W_+)$ becomes

$$\mathbb{E}(W_+)\big|_{q=1} = \frac{\gamma}{\mu}. \tag{42}$$

We recognize Eq. (42) as the mean of the limiting *total-life* random variable of a renewal process with $B(\cdot)$ serving as the interarrival time distribution function (e.g., see Kao [6, Section 3.3]).

4.3. $M/G/1$ queue under a q-policy with closedown periods

We now consider a slight variant of the $M/G/1$ queue operating under the q-policy. Specifically, we incorporate a closedown period, S, after each busy period. It is assumed that the sequence of successive closedown periods are iid with distribution function $A(x) = \mathbb{P}(S \leq x)$. The facility is closed to all potential arrivals during a closedown period. Thus, the incorporation of a closedown period will increase the proportion of customers that are blocked from the system. In addition, it is obvious that the closedown periods do not affect the waiting time distributions for serviceable customers, and so our analysis of waiting time in the previous subsections is still applicable.

We view the total idle period as the durations of time when the server is not busy. Hence, similar to the partitioning of the steady-state probability of the system being busy, we define the following:

$P_{I,0} \equiv$ steady-state probability the server is idle and the system is closed;

$P_{I,1} \equiv$ steady-state probability the server is idle and the system is open.

In this variation, the busy cycle remains $D = T + I$ (note though that the closedown period is contained in I). Again, applying elementary renewal theory arguments, we obtain:

$$P_I = \frac{\mathbb{E}(I)}{\mathbb{E}(D)} = \frac{(1-\rho^{(q)})(1+\lambda\mathbb{E}(S))}{(1+\lambda\mathbb{E}(S))(1+\rho q)-\lambda\mathbb{E}(S)\rho}, \tag{43}$$

$$P_{I,0} = \frac{\mathbb{E}(S)}{\mathbb{E}(I)}P_I = \frac{(1-\rho^{(q)})\lambda\mathbb{E}(S)}{(1+\lambda\mathbb{E}(S))(1+\rho q)-\lambda\mathbb{E}(S)\rho}, \tag{44}$$

$$P_{I,1} = \frac{\lambda^{-1}}{\mathbb{E}(I)}P_I = \frac{1-\rho^{(q)}}{(1+\lambda\mathbb{E}(S))(1+\rho q)-\lambda\mathbb{E}(S)\rho}, \tag{45}$$

$$P_B = \frac{\mathbb{E}(T)}{\mathbb{E}(D)} = \frac{\rho}{(1+\lambda\mathbb{E}(S))(1+\rho q)-\lambda\mathbb{E}(S)\rho}, \tag{46}$$

$$P_{B,0} = qP_B = \frac{\rho q}{(1+\lambda\mathbb{E}(S))(1+\rho q)-\lambda\mathbb{E}(S)\rho}, \tag{47}$$

$$P_{B,1} = (1-q)P_B = \frac{\rho^{(q)}}{(1+\lambda\mathbb{E}(S))(1+\rho q)-\lambda\mathbb{E}(S)\rho}. \tag{48}$$

Thus, the long-run fraction of time the system is accepting of new customers is $P_{I,1} + P_{B,1} = [(1+\lambda\mathbb{E}(S))(1+\rho q)-\lambda\mathbb{E}(S)\rho]^{-1}$.

5. $M/G/1$ queue with accumulating priority

In order to implement the q-policy, a system manager must know the service times of the customers upon their arrival to the system. However, such knowledge may not always be available. In this section, we introduce another $M/G/1$-type queueing model which enables a system manager to reduce the length of busy periods, in a similar fashion as the q-policy, without the knowledge of service times upon arrival. This system is a variant of the $M/G/1$ queue with accumulating priority, which was recently studied by Stanford, Taylor, and Ziedins [12].

The first key aspect of the $M/G/1$ queue with accumulating priority has to do with how priority is accumulated for customers. Specifically, customers arrive to the system with zero initial priority and, throughout their sojourn in the system, earn priority linearly at rate $\xi_1 > 0$. At service completion epochs, the customer with the greatest accumulated priority is serviced next. The second key feature of this model lies in the concept of an accreditation threshold, which increases linearly at rate ξ_2 where $0 \leq \xi_2 \leq \xi_1$. In fact, the accreditation threshold is a stochastic process which we denote as $\{\Theta(t), t \geq 0\}$. It is important to note that the accreditation threshold and its implementation does not, in any way, affect the order of service for customers. Hence, the way in which the $M/G/1$ queue with accumulating priority operates is actually equivalent to the classical $M/G/1$ queue under the FCFS discipline. However, the incorporation of the accreditation threshold does shed new light on the structuralization of the general busy period, providing a useful classification of those customers who arrive during busy periods.

The above basic model was introduced by Stanford et al. [12] in their analysis of a particular multi-class non-preemptive priority system. In order to analyze the $M/G/1$ queue with accumulating priority, these authors defined a new stochastic process which they called the *maximal priority process*. To incorporate a blocking policy into this system, we require a slight modification to their definition of the maximal priority process. We then establish the relation between our modified maximal priority process and the censored virtual wait process of the previous section. We exploit this relation to obtain the steady-state integral equation of the *accumulated priority* of serviceable customers.

5.1. The maximal priority process

Upon arrival to the system, customers begin to accumulate priority at a linear rate. During busy periods, a customer will be admitted for service only if its priority overtakes (i.e., becomes greater than) the accreditation threshold, governed by $\{\Theta(t), t \geq 0\}$. At a service completion instant, if there are any admitted customers present in the system, the one with the greatest accumulated priority is selected next for service. The busy period ends at a service completion instant which leaves no more admitted customers in the system. Note that the busy period may end while there are still customers present in the system. In this situation, these customers depart the system without ever entering into service.

Let τ_k denote the arrival epoch of customer C_k, so that we may define $\Phi_k(t)$ to be this customer's priority function (i.e., the amount of accumulated priority C_k has at time t), namely:

$$\Phi_k(t) = \xi_1(t - \tau_k), \quad t > \tau_k. \tag{49}$$

Furthermore, let $n(k)$ denote the arrival position of the k-th customer to be serviced. The definition of the maximal priority process now follows.

Definition 5.1. *The maximal priority process is a two-dimensional stochastic process* $\mathcal{M}(t) = \{(M(t), \Theta(t)), t \geq 0\}$*, satisfying the following conditions:*

1. $\mathcal{M}(t) = (0, 0)$ *for all* t *corresponding to idle periods.*

2. *For all t not corresponding to service commencement/completion instants, we have*

$$\frac{\mathrm{d}M(t)}{\mathrm{d}t} = \xi_1 \quad \textit{and} \quad \frac{\mathrm{d}\Theta(t)}{\mathrm{d}t} = \xi_2, \tag{50}$$

where $0 \leq \xi_2 \leq \xi_1$.

3. *At the sequence of service completion times $\{\delta_k, k = 1, 2, \ldots\}$,*

$$M(\delta_k) = 1\big\{\Phi^{\vee}(\delta_k^-) > \Theta(\delta_k^-)\big\} \cdot \Phi^{\vee}(\delta_k^-), \tag{51}$$

$$\Theta(\delta_k^+) = \min\{M(\delta_k), \Theta(\delta_k^-)\}, \tag{52}$$

where

$$\Phi^{\vee}(\delta_k^-) = \max_{m \in \{n(k)+1, n(k)+2, \ldots\}} \Phi_m(\delta_k^-) \tag{53}$$

and $1\{A\}$ is the indicator function of the event A.

The above definition shows that $\{M(t), t \geq 0\}$ is closely related to the well-known age process (i.e., when $\xi_1 = 1$, $M(t)$ represents the age of the oldest admitted customer at time t). Furthermore, the accreditation threshold process increases linearly at rate ξ_2 during busy periods. Stanford et al. [12] referred to those customers who arrive during busy periods and whose priority overtakes the accreditation threshold as *accredited customers.*

With this definition in place, we can now introduce the blocking scheme for our modified $M/G/1$ queue with accumulating priority. In particular, serviceable customers consist of accredited customers and customers who arrive during idle times. On the other hand, those customers whose priority fails to overtake the accreditation threshold during a busy period are blocked, thereby departing the system without ever entering into service. We refer to such customers as non-accredited customers.

Figure 5 depicts a typical sample path of $\{\mathcal{M}(t), t \geq 0\}$. Note that customers C_4, C_5, and C_9 are of the non-accredited type and thus end up being blocked from service. Moreover, a notable difference between the current model and the one considered in Section 4 is that with the current system, blocked customers experience some wait before being forced to depart the system.

Suppose now at the end of an arbitrary busy period, we wish to find the latest time by which a customer would have to arrive in order to be admitted for service. This can be done by simply dividing the height of the accreditation threshold at time t_* (i.e., the time at which the busy period completes) by ξ_1 and subsequently subtracting this quantity from t_*. For a sample path such as the one shown in Figure 5, this is equivalent to determining the t-intercept of a line with slope ξ_1 which crosses the point $(t_*, \Theta(t_*^-))$.

For each busy period, we define the accreditation interval as the duration of time within which customers must arrive in order to be admitted for service. An important observation is that the ratio of the accreditation interval to the busy period is always $(1 - \xi_2/\xi_1)$. Therefore, this model is similar to the one of Section 4 in that admitted customers must arrive within the first $(1-q)$-th proportion of the busy period with $q = \xi_2/\xi_1$. In fact, it can be shown that the LST of the busy period is the solution to Eq. (4) with $q = \xi_2/\xi_1$ (see Stanford et al. [12] and their discussion on accredited busy periods). In addition, using the same argument as in Brill [1], we can show that the steady-state distribution of $\{M(t), t \geq 0\}$ when $\xi_1 = 1$ is equivalent to the steady-state distribution of the workload process $\{U_{\xi_2}(t), t \geq 0\}$ of Section 4.

5.2. The distribution of accumulated priority for serviceable customers

Let ${}^{\xi_1}\phi_n$ be the accumulated priority, immediately prior to entering service, of the n-th customer to be serviced where ξ_1 is the priority accumulation rate. Trivially, we have

$$ {}^{\xi_1}\phi_n = \xi_1 \cdot W_n, \tag{54}$$

where W_n is the waiting time of the n-th serviceable customer. Furthermore, ${}^{1}\phi_n \equiv \phi_n = W_n$ represents the age of the customer prior to entering service. Let $g_{\xi_1}(x)$ denote the steady-state pdf of the accumulated priority (immediately prior to entering service) for the serviceable customers.

Since waiting times for the serviceable customers in the current model are equivalently characterized by the waiting times for serviceable customers in the model of Section 4 with $q = \xi_2/\xi_1$, we have $g_1(x) \equiv g(x) = f(x; \xi_2)$ for all $x > 0$, from which it immediately follows that

$$ g_{\xi_1}(x) = \frac{g(x/\xi_1)}{\xi_1}, \quad x > 0. \tag{55}$$

From Eq. (31), we get

$$ g_{\xi_1}(x) = \frac{\lambda \bar{B}(x/\xi_1)P_0 - \lambda P_0 \bar{G}_{\xi_2/\xi_1}(x/\xi_1)}{\xi_1} + \frac{\lambda}{\xi_1}\int_{y=0}^{x/\xi_1} \bar{B}(x/\xi_1 - y)g(y)\mathrm{d}y. \tag{56}$$

Thus, multiplying both sides of Eq. (56) by e^{-sx} and integrating over $x \in (0, \infty)$, we obtain the LSTs of the steady-state accumulated priority for serviceable customers and accredited-type customers as

$$ {}^{\xi_1}\widetilde{\phi}(s) = \frac{(1 - \rho^{(\xi_2/\xi_1)})(\xi_1 s - \lambda + \lambda \widetilde{G}(\xi_2 s))}{\xi_1 s - \lambda + \lambda \widetilde{B}(\xi_1 s)}, \tag{57}$$

and

$$
{}^{\xi_1}\widetilde{\phi}_+(s) = \frac{(1-\rho^{(\xi_2/\xi_1)})(\widetilde{G}(\xi_2 s) - \widetilde{B}(\xi_1 s))}{\mu(1-\xi_2/\xi_1)(\xi_1 s - \lambda + \lambda\widetilde{B}(\xi_1 s))}, \tag{58}
$$

respectively. Alternatively, Eq. (57) and Eq. (58) can be obtained by substituting $s = \xi_1 s$ and $q = \xi_2/\xi_1$ into Eq. (34) and Eq. (36), respectively.

We remark that Eq. (58) was first presented by Stanford et al. [12]. However, their result was obtained under a different setting, as they studied a particular multi-class non-preemptive priority system and obtained the steady-state marginal waiting time distributions of each class. We emphasize that in their model, there is no concept of customer blocking. The authors obtained their result for a random variable which they called the *additional accumulated priority.* We direct readers to their paper for more details. Moreover, the authors' method of analysis differs from ours in that their proof of Eq. (58) is inspired by the Conway et al. [3, Chapter 8-4] derivation of the flow time LST in a classical FCFS $M/G/1$ system.

In summary, our level crossing analysis provides an alternate proof of Stanford et al.'s [12] main result (i.e., Eq. (58)) and also yields the steady-state integral equation for the pdf of accumulated priority in Eq. (56). Furthermore, our model provides an alternate interpretation of the wait of customers in their non-preemptive priority system. Specifically, the additional wait that high priority customers serviced in an accredited busy period experience is identical to the wait experienced by delayed customers in an $M/G/1$ system under the q-policy.

5.3. The overall distribution of wait

We next establish the distribution of the overall waiting time random variable. Clearly, by design of the model, customers who are blocked from service will experience a (steady-state) waiting time (or total time in the system) which follows the limiting distribution of the forward recurrence time of qT. Defining W_0 to be the wait of such non-serviced customers, it readily follows that

$$
\widetilde{W_0}(s) = \frac{1-\widetilde{G}_{\xi_2/\xi_1}(s)}{\mathbb{E}(T)s\xi_2/\xi_1}. \tag{59}
$$

Since priority is accumulated linearly at rate ξ_1, the LST of the waiting time for serviceable customers is obtained from Eq. (57) as

$$
\widetilde{W_1}(s) = {}^{\xi_1}\widetilde{\phi}(s/\xi_1) = \frac{(1-\rho^{(\xi_2/\xi_1)})(s-\lambda+\lambda\widetilde{G}(\xi_2 s/\xi_1))}{s-\lambda+\lambda\widetilde{B}(s)}. \tag{60}
$$

Using the steady-state probabilities given by Eq. (14) through Eq. (17), we derive the overall LST of waiting time as

$$\widetilde{W}(s) = \frac{1}{1+\rho\xi_2/\xi_1}\widetilde{W}_1(s) + \frac{\rho\xi_2/\xi_1}{1+\rho\xi_2/\xi_1}\widetilde{W}_0(s). \tag{61}$$

After some elementary algebra, we obtain

$$\widetilde{W}(s) = \left(\frac{1-\rho^{(\xi_2/\xi_1)}}{1+\rho(\xi_2/\xi_1)}\right) \times \left(\frac{s-\lambda+\lambda\widetilde{G}(s\xi_2/\xi_1)}{s-\lambda+\lambda\widetilde{B}(s)} + \frac{\lambda(1-\widetilde{G}(s\xi_2/\xi_1))}{s}\right). \tag{62}$$

6. An optimization problem

In this section, we formulate a numerical study to demonstrate a potential usage of the q-policy. We remark that the inspiration for this study originates from a similar study considered by Kao [6, Example 3.6.4]. In what follows, we consider a queueing system with closedown periods as described in Section 4.3. For this system, suppose we have the following monetary parameters:

K ≡ the cost of each closedown period;
h ≡ the cost of holding one customer per unit time;
R ≡ the toll fee paid by each serviced customer.

The objective function which we seek to optimize is the long-run expected profit per unit time. Clearly, the instants of busy period commencements define a set of regeneration points. Thus, our objective function is

$$P(q) = \frac{R\cdot\mathbb{E}(N_{bp}) - K - \mathbb{E}(C_{bp})}{\mathbb{E}(D)}, \tag{63}$$

where $\mathbb{E}(C_{bp})$ is the expected holding cost incurred during a busy period. We remark that $\mathbb{E}(N_{bp})$ is given by Eq. (13) and $\mathbb{E}(D) = \mathbb{E}(T) + \mathbb{E}(S) + \lambda^{-1}$. Moreover, it can be shown, following a similar line of reasoning to Kao [6, pp.139-140], that for all work-conserving service disciplines (e.g., both FCFS and the q-restricted LCFS disciplines),

$$\mathbb{E}(C_{bp}) = h\mathbb{E}(N_{bp})(\mu + \mathbb{E}(W)). \tag{64}$$

Note that the quantity $\mu + \mathbb{E}(W)$ represents the long-run average flow time.

By recalling the form of $\mathbb{E}(W)$ in Eq. (37), it is immediately clear that $\mathbb{E}(C_{bp})$ depends only on the first two moments of the service time distribution. Consequently, the expected profit function $P(q)$ is also affected by the variability of the service time distribution. We use the coefficient of variation of the service time distribution, denoted by $CV = \sqrt{\gamma - \mu^2}/\mu$, to assess the effect of the variability of the service time distribution on the profit function. In particular, we present five numerical examples of nearly identical models, differing only in their respective coefficients of variation of the service time distribution. In Examples 1 through 5, we consider five service time distributions with common mean $\mu = 1$, but with coefficients of variation 0, 0.5, 1, 1.5, and 2, respectively.

Figure 6 displays the profit functions corresponding to the five examples. With the exception of the profit functions for Examples 1 and 2, we observe that the expected profit per unit time can be maximized by implementing the q-policy. Letting q^* denote the optimal blocking proportion which maximizes $P(q)$, we find q^* (to 4 decimal places of accuracy) for Examples 1 through 5 to be 0, 0, 0.1000, 0.1710, and 0.2538, respectively. In Table 1, we calculate the expected profit function and several other quantities of interest corresponding to various values of the blocking proportion q for Examples 1 through 5. We note that since $\mu = 1$, Eq. (8) and Eq. (13) together imply that $\mathbb{E}(T) = \mathbb{E}(N_{bp})$ for all values of q.

Although it is indeed true that the maximum long-run expected profit per unit time is obtained without the usage of a q-policy (i.e., $q^* = 0$) for both Examples 1 and 2, there are other viable reasons for the implementation of a q-policy. In regard to Example 2, let us define q_r^* to be the relative maxima of $P(q)$. By using standard calculus-based methods, we find that $q_r^* = 0.0406$. From Table 1 (and the rows corresponding to Example 2), we see that the resulting expected profits with $q = q^* = 0$ and $q = q_r^*$ differ only by a small amount. However, the advantage of implementing a q-policy still lies in the fact that both the cycle and busy period lengths are smaller when compared to the system without a q-policy in place. Ultimately, with $q = q_r^*$, the system is essentially earning the same expected profit as for the case with $q = 0$, but at the same time allowing for more frequent maintenance checks on the server/machine. Similar remarks can be made for Example 1.

In these numerical examples, we showed that by reducing the cycle lengths, a system manager can significantly decrease the incurred costs and thus capture the potential profit (or, as in both Examples 1 and 2, obtain nearly maximal expected profit). It is also apparent that as CV increases, so too does the optimal blocking proportion q^*, as evidenced in Figure 7. It is interesting to note the presence of a discontinuity point in Figure 7, which occurs for a certain value of CV residing in the interval $(0.6014, 0.6015)$. This particular value of CV corresponds to the first instance in which a non-zero blocking proportion yields a higher expected profit.

7. Concluding Remarks

In conclusion, we have presented a queueing model which enables a system manager, through choice of which blocking proportion q to use, to effectively reduce the duration of a busy period. We have studied several performance measures of interest, including the wait of serviceable customers and the busy period duration. We have also shown that, in certain situations, the reduction of busy period lengths can be accompanied with an increase in profit. A possible avenue for future research lies in the area of server vacation models (e.g., see Takagi [15, Chapter 2]). In particular, rather than closing the system to potential arrivals, one could consider a situation where the server goes on vacation, during which customers are still allowed to join the queue.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada. The authors are very grateful to the anonymous referees for their careful reading of the paper and their helpful suggestions that have considerably improved the presentation of the paper.

References

[1] Brill, P.H. (1988). Single-server queues with delay dependent arrival streams. *Probability in the Engineering and Informational Sciences* 2(2): 231-247.

[2] Brill, P.H. (2008). *Level Crossing Methods in Stochastic Models.* New York: Springer.

[3] Conway, R.W., Maxwell, W.L., & Miller, L.W. (1967). *Theory of Scheduling.* Reading, MA: Addison-Wesley.

[4] Doshi, B.T. (1977). Continuous time control of the arrival process in an $M/G/1$ queue. *Stochastic Processes and their Applications* 5(3): 265-284.

[5] Johansen, S.G. & Stidham, S. (1980). Control of arrivals to a stochastic input-output system. *Advances in Applied Probability* 12(4): 972-999.

[6] Kao, E. (1996). *An Introduction to Stochastic Processes.* Belmont, CA: Duxbury Press.

[7] Kleinrock, L. (1975). *Queueing Systems, Volume I, Theory.* New York: John Wiley & Sons.

[8] Knudsen, N.C. (1972). Individual and social optimization in a multiserver queue with a general cost-benefit structure. *Econometrica* 40(3): 515-528.

[9] Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica* 37(1): 15-24.

[10] Parzen, E. (1962). *Stochastic Processes.* Oakland, CA: Holden-Day.

[11] Rosenshine, M. & Rue, R.C. (1981). Some properties of optimal control policies for entry to an $M/M/1$ queue. *Naval Research Logistics Quarterly* 28(4): 525-532.

[12] Stanford, D.A., Taylor, P., & Ziedins, I. (2014). Waiting time distributions in the accumulating priority queue. *Queueing Systems*, 77(3): 297-330.

[13] Stidham, S. (1985). Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control* 30(8): 705-713.

[14] Stidham, S. (2002). Analysis, design, and control of queueing systems. *Operations Research* 50(1): 197-216.

[15] Takagi, H. (1991). *Queueing Analysis, Volume I: Vacation and Priority Systems, Part 1.* Amsterdam: North-Holland.

[16] Yechiali, U. (1971). On optimal balking rules and toll charges in the $GI/M/1$ queueing process. *Operations Research* 19(2): 349-370.

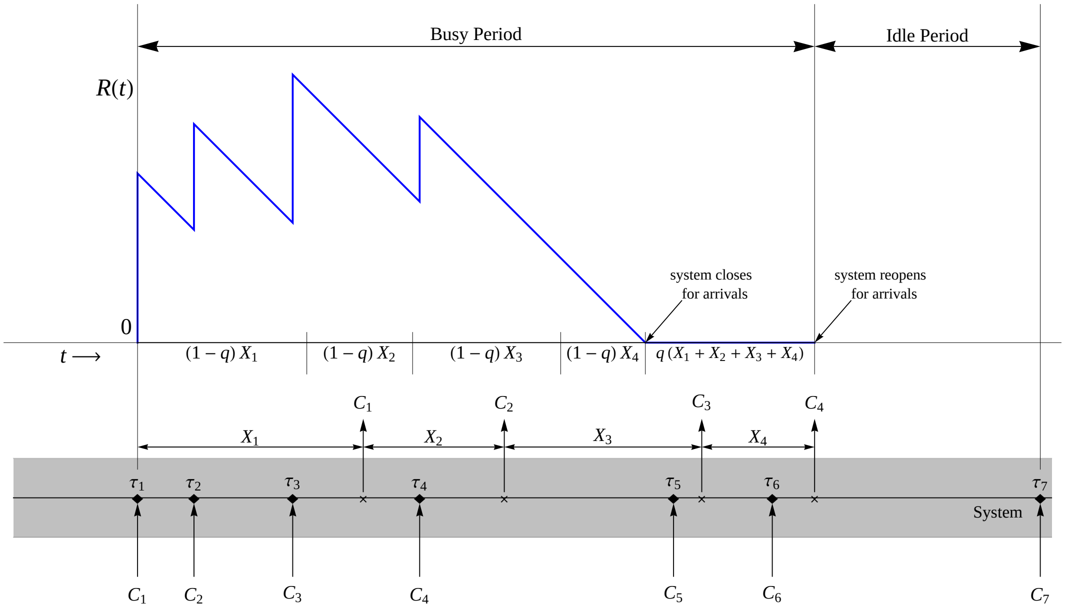


Figure 1: A typical busy period under the q-policy

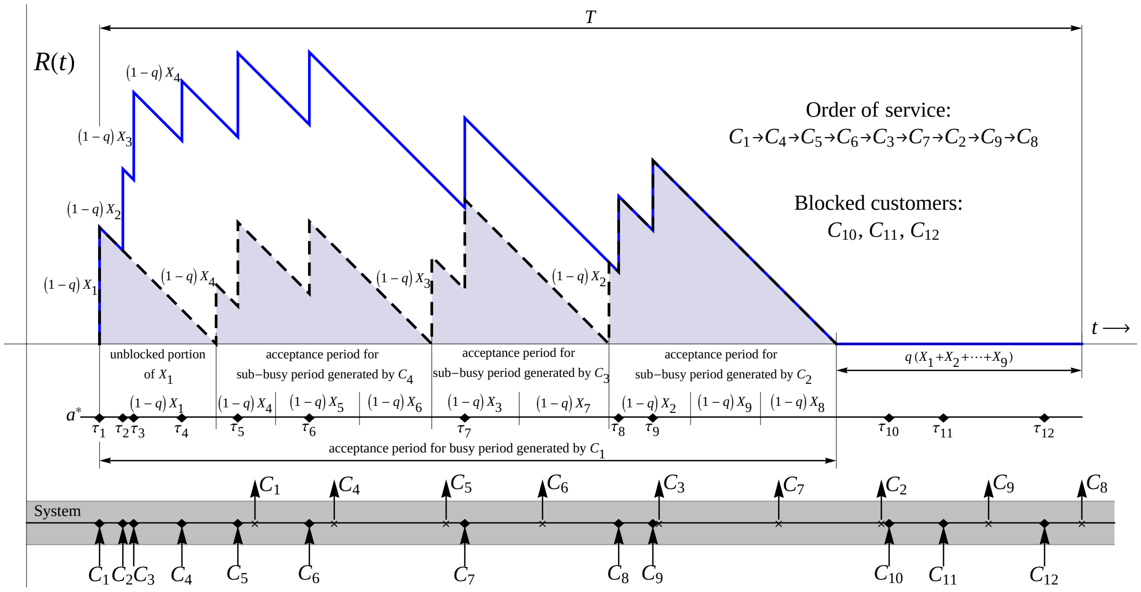
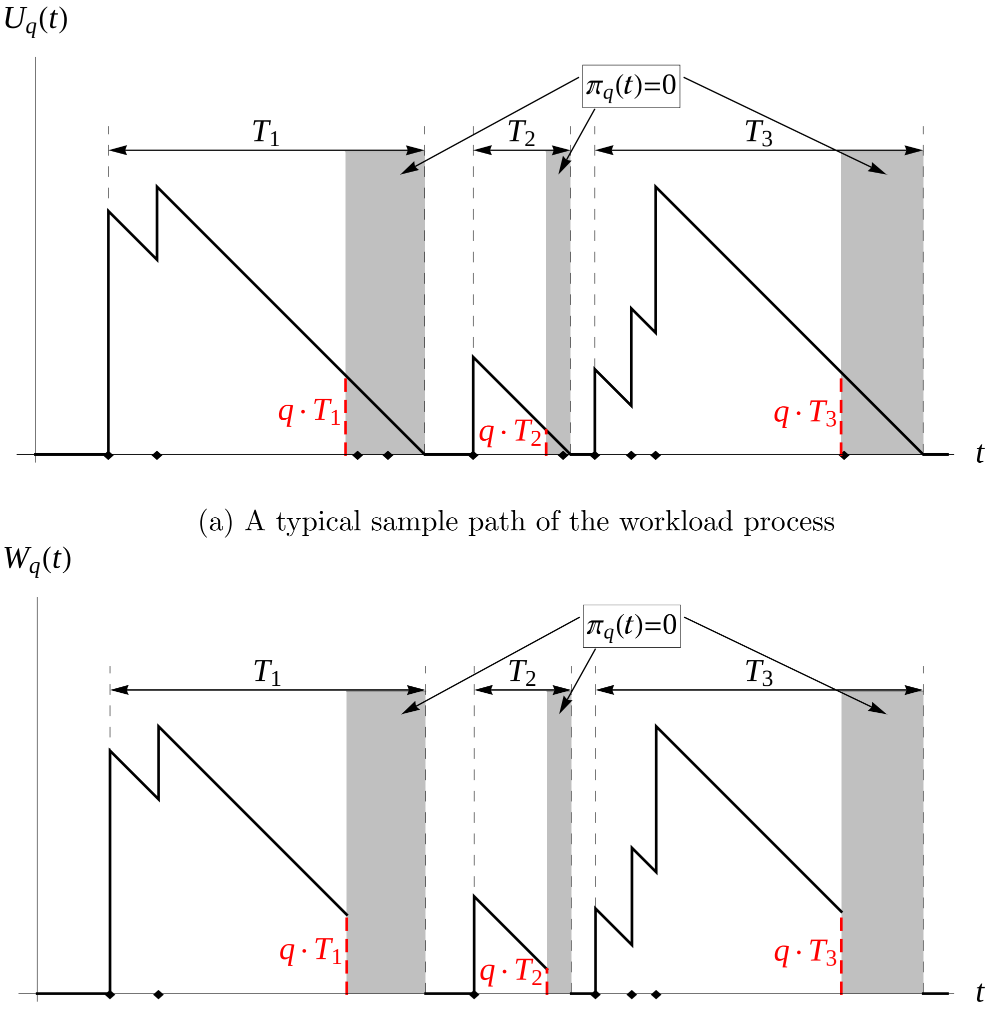


Figure 2: A busy period under the q-restricted LCFS discipline



(a) A typical sample path of the workload process

(b) Corresponding sample path of the virtual wait process

Figure 3: Typical sample paths of the processes $\{U_q(t), t \geq 0\}$ and $\{W_q(t), t \geq 0\}$

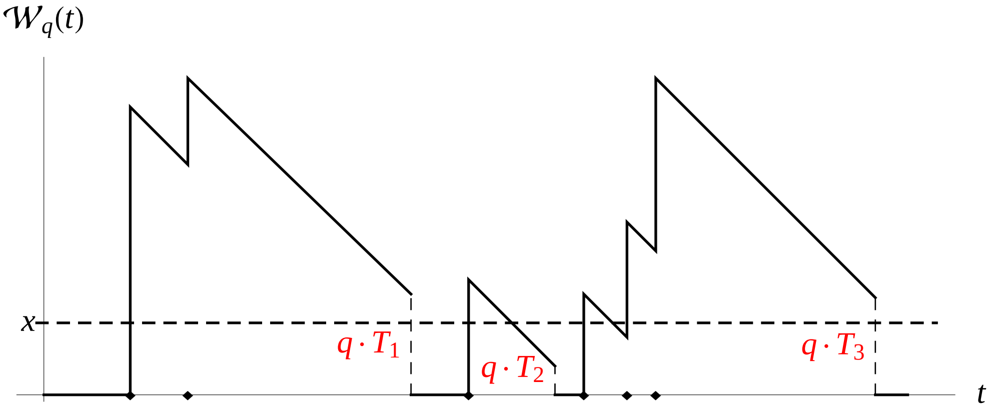


Figure 4: Sample path up- and down-crossings of level x for $\{\mathcal{W}_q(t), t \geq 0\}$

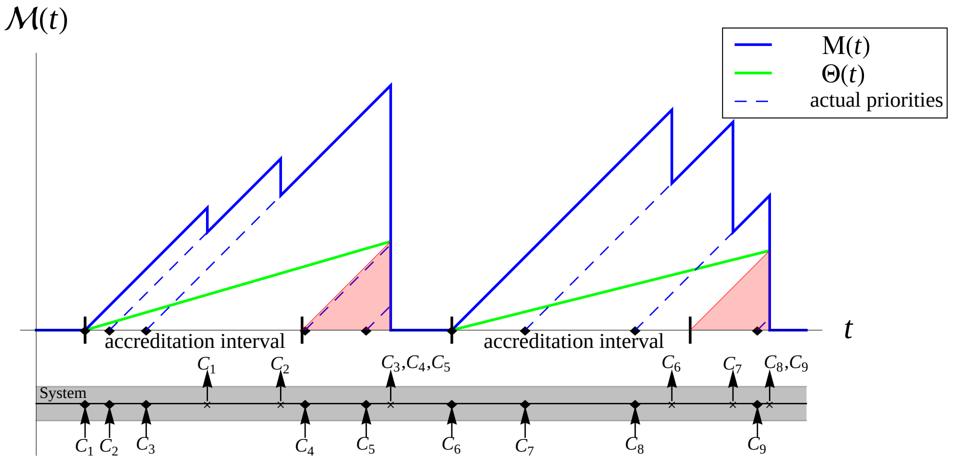


Figure 5: A typical sample path of $\{M(t), t \geq 0\}$

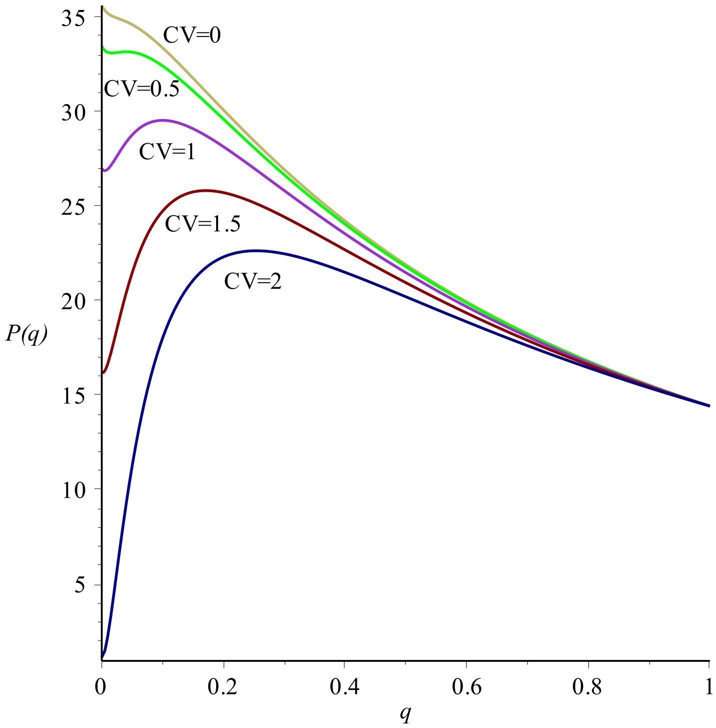


Figure 6: Expected profit per unit time for Examples 1 through 5

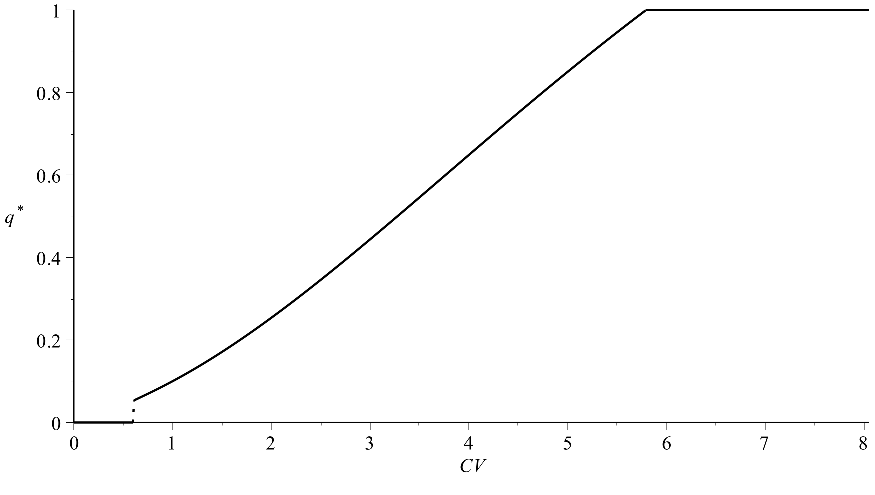


Figure 7: Behaviour of the optimal blocking proportion q^* as a function of CV

Table 1: Expected profit per unit time and other quantities of interest against various q-values for Examples 1 through 5

$M/G/1$ queue with $\lambda = 0.95$; $\mu = 1$; $\mathbb{E}(S) = 1$; $h = 1$; $K = 5$; $R = 50$								
Example 1: $CV = 0$; $q^* = 0.00$								
Quantity	———	$q = 0.00$	$q = 0.05$	$q = 0.10$	$q = 0.20$	$q = 0.40$	$q = 0.55$	$q = 0.85$
$P(q)$	———	35.5967	34.5886	33.3634	30.0792	24.2058	20.8748	16.1395
$\mathbb{E}(D)$	———	22.0526	12.3090	8.9492	6.2193	4.3782	3.7994	3.2188
$\mathbb{E}(C_{bp})$	———	210.0000	82.0681	41.2522	16.2616	5.3008	3.0254	1.3591
$\mathbb{E}(N_{bp})$	———	20.0000	10.2564	6.8966	4.1667	2.3256	1.7467	1.1662
$\mathbb{E}(W)$	———	9.5000	7.0016	4.9816	2.9028	1.2793	0.7321	0.1655
Example 2: $CV = 0.5$; $q^* = 0.00$; $q_r^* = 0.0406$								
Quantity	$q = q_r^*$	$q = 0.00$	$q = 0.05$	$q = 0.10$	$q = 0.20$	$q = 0.40$	$q = 0.55$	$q = 0.85$
$P(q)$	33.1506	33.4427	33.1301	32.4037	29.5931	24.0359	20.7907	16.1245
$\mathbb{E}(D)$	13.3439	22.0526	12.3090	8.9492	6.2193	4.3782	3.7994	3.2188
$\mathbb{E}(C_{bp})$	117.2050	257.5000	100.0211	49.8411	19.2853	6.0446	3.3451	1.4074
$\mathbb{E}(N_{bp})$	11.2912	20.0000	10.2564	6.8966	4.1667	2.3256	1.7467	1.1662
$\mathbb{E}(W)$	9.3802	11.8750	8.7521	6.2270	3.6285	1.5992	0.9151	0.2068
Example 3: $CV = 1$; $q^* = 0.1000$								
Quantity	$q = q^*$	$q = 0.00$	$q = 0.05$	$q = 0.10$	$q = 0.20$	$q = 0.40$	$q = 0.55$	$q = 0.85$
$P(q)$	29.5245	26.9809	28.7545	29.5245	28.1345	23.5263	20.5383	16.0795
$\mathbb{E}(D)$	8.9491	22.0526	12.3090	8.9492	6.2193	4.3782	3.7994	3.2188
$\mathbb{E}(C_{bp})$	75.6071	400.0000	153.8799	75.6079	28.3565	8.276	4.3041	1.5521
$\mathbb{E}(N_{bp})$	6.8965	20.0000	10.2564	6.8966	4.1667	2.3256	1.7467	1.1662
$\mathbb{E}(W)$	9.9631	19.0000	14.0033	9.9631	5.8056	2.5587	1.4641	0.3309
Example 4: $CV = 1.5$; $q^* = 0.1710$								
Quantity	$q = q^*$	$q = 0.00$	$q = 0.05$	$q = 0.10$	$q = 0.20$	$q = 0.40$	$q = 0.55$	$q = 0.85$
$P(q)$	25.8100	16.2112	21.4619	24.7257	25.7036	22.6768	20.1176	16.0046
$\mathbb{E}(D)$	6.7606	22.0526	12.3090	8.9492	6.2193	4.3782	3.7994	3.2188
$\mathbb{E}(C_{bp})$	55.9079	637.5000	243.6445	118.5524	43.4751	11.995	5.9025	1.7933
$\mathbb{E}(N_{bp})$	4.7080	20.0000	10.2564	6.8966	4.1667	2.3256	1.7467	1.1662
$\mathbb{E}(W)$	10.8751	30.8750	22.7553	16.1901	9.4340	4.1579	2.3792	0.5377
Example 5: $CV = 2$; $q^* = 0.2538$								
Quantity	$q = q^*$	$q = 0.00$	$q = 0.05$	$q = 0.10$	$q = 0.20$	$q = 0.40$	$q = 0.55$	$q = 0.85$
$P(q)$	22.6279	1.1337	11.2523	18.0075	22.3003	21.4876	19.5286	15.8997
$\mathbb{E}(D)$	5.4881	22.0526	12.3090	8.9492	6.2193	4.3782	3.7994	3.2188
$\mathbb{E}(C_{bp})$	42.5880	970.0000	369.3151	178.6748	64.6412	17.2016	8.1402	2.1309
$\mathbb{E}(N_{bp})$	3.4354	20.0000	10.2564	6.8966	4.1667	2.3256	1.7467	1.1662
$\mathbb{E}(W)$	11.3967	47.5000	35.0082	24.9078	14.5139	6.3967	3.6603	0.8273