Queueing Models and
Service Management
Vol. 5, No. 1, page 1-25, 2022

QMSM

Waiting Times in an Exponential Queue with Hysteretic Service Rate Control: A Numerical Investigation

B. Madhu Rao*

Department of Business Systems and Analytics
School of Business Administration
Stetson University
DeLand, FL 32720, USA
(Received January 2021 ; accepted June 2021)

Abstract: Increasing the service rate when a queue is long is a common practice in many service systems. The control mechanism can be uni-level, where the service rate is increased whenever a preset threshold is exceeded and returns to normal when the number of customers drops below the same threshold. In a hysteretic control, the service rate remains at high level until the number in the queue is brought down to a value much lower than the threshold for increasing the service rate. When the server can operate at k levels, a k-level hysteretic control can be used. While explicit results are obtained for the equilibrium distribution of the number of customers in the system, there is no study dealing with the waiting time characteristics of the customers entering the system. The difficulty arises from the fact that service rates, and hence the waiting times, are affected by arrivals during a customer's sojourn through the system. In this paper algorithms are developed to compute the equilibrium probability distributions of customer waiting times and the times spent by the server at each of the service levels during each visit to that state. Numerical results are presented to describe the behavior of the system under different traffic and control structures.

Keywords: Algorithmic probability, control, design, hysteretic control, queues.

1. Introduction

Service systems are typically not designed to meet the peak demand. Customers, in general, tolerate a certain amount of waiting before frustration and dissatisfaction with the level of service sets in. Thus, it is prudent to provide service at a normal or lower rate, leading to a steadily increasing waiting line and increase the service rate when the potential waiting time approaches the limit of customers' tolerance. This is especially appropriate when arrivals are subject to periodic or random fluctuations. In the design of such systems, one needs to balance the need to avoid customer dissatisfaction with the need for high utilization of the servers and the cost of service rate changes. We refer the readers to [8] for the practical significance of this model. A simple and direct approach to managing queue size in such systems is to increase the service rate when the queue builds up *sufficiently*, and to bring the waiting line down to a *reasonable* level before reducing the service rate to the normal level. Gebhard [3] coined the term *hysteretic control*, to describe such two level control of

*Corresponding author
Email : bmrao@stetson.edu

service rates. Increasing the service rate for a short time typically requires the use of more expensive resources (e.g., a manager works as a checkout clerk, use of expensive gasoline power generators to meet the peak demand to supplement the base capacity provided by the larger and less expensive coal or nuclear generators) and should be used sparingly.

Early analysis of exponential queues with hysteretic control considered the equilibrium behavior of the number in the system [3, 5, 11]. Tijms [10] Federgruen and Tijms [2] among others [1] studied the determination of the optimal values of u (upper control limit) and l (lower control limit) by considering the cost of waiting customers and the costs of operating the server at the two service rates. Neuts and Rao [6] considered a system with phase type service time and finite waiting space. They introduced the cost of lost customers to the analysis and developed algorithmic methodology to study the effect of service time distribution.

For the basic exponential queuing system with two level hysteretic control, results for the number in system are obtained in closed form [3] but no analysis of the behavior of the system under various parameter values has been carried out. More importantly, the waiting time of customers has not been studied, mainly because the service rate can change several times during a customer's sojourn through the system [5]. Similarly, analysis of the time spent by the server at each service rate under equilibrium conditions has also not been studied. These characteristics provide valuable insight into the system behavior useful to the decision makers. This paper proposes an algorithmic methodology to compute these important system characteristics and present a detailed description of the system behavior under various parameter values. The subscripts n and h are used throughout the paper to identify parameters and system characteristics when the system is operating at the *normal* and *higher* service rates respectively. Properties of phase type distributions and related computational considerations for use in this paper are presented in the Appendix for readers' ready reference.

2. Mathematical Model

The arrival process of customers is characterized by a Poisson stream of rate λ. Service times are exponentially distributed with either the *normal* rate (μ_n) or the *higher* rate (μ_h). Service rate is increased from μ_n to μ_h when the number of customers exceeds the upper threshold (u), and returns to μ_n when the number drops below the lower threshold l, $(1 \leq l < u)$.

System state is described by the two tuple (i, k), where i denotes the number of customers waiting for or receiving service, and k denotes the state of the server, taking values 0 or 1 depending on whether the server is operating at the normal or higher service rate. For this system, the service rate will always be at the *normal* level (i.e., $k = 0$) when i is below l, and will always be at the *higher* rate (i.e., $k = 1$) when i is above u. When i is between l and u ($l \leq i \leq u$), there will be two states for each value of i, corresponding to the two service rates (i.e., one each for $k = 0$ and $k = 1$).

Arranging the states in lexicographic order, in increasing order of k and i, in that order, the system dynamics can be described by a continuous parameter Markov chain with infinitesimal generator Q as shown in Figure 1, where, $\theta_n = -\lambda - \mu_n$ and $\theta_h = -\lambda - \mu_h$. In the infinitesimal generator below, states corresponding to normal service rate are denoted in *italics* and states corresponding to higher service rate are denoted in **bold**.

$$
Q = \begin{array}{c|ccccc|ccccc|ccccc}
 & \mathit{0} & \mathit{1} & \mathit{2} & . & \mathit{l-1} & \mathit{l} & \mathit{l+1} & . & \mathit{u} & \mathbf{l} & \mathbf{l+1} & . & \mathbf{u} & \mathbf{u+1} & \mathbf{u+2} & . & . & . \\
\hline
\mathit{0} & -\lambda & \lambda & & & & & & & & & & & & & & & & \\
\mathit{1} & \mu_n & \theta_n & \lambda & & & & & & & & & & & & & & & \\
\mathit{2} & & \mu_n & \theta_n & \lambda & & & & & & & & & & & & & & \\
\cdots & & & . & . & . & & & & & & & & & & & & & \\
\mathit{l-1} & & & & \mu_n & \theta_n & \lambda & & & & & & & & & & & & \\
\hline
\mathit{l} & & & & & & \theta_n & \lambda & & & & & & & & & & & \\
\mathit{l+1} & & & & & & \mu_n & \theta_n & \lambda & & & & & & & & & & \\
\cdots & & & & & & & . & . & . & & & & & & & & & \\
\mathit{u} & & & & & & & & \mu_n & \theta_n & & & & & \lambda & & & & \\
\hline
\mathbf{l} & & & & & \mu_h & & & & & \theta_h & \lambda & & & & & & & \\
\mathbf{l+1} & & & & & & & & & & \mu_h & \theta_h & \lambda & & & & & & \\
\cdots & & & & & & & & & & & . & . & . & & & & & \\
\mathbf{u} & & & & & & & & & & & & \mu_h & \theta_h & \lambda & & & & \\
\hline
\mathbf{u+1} & & & & & & & & & & & & & \mu_h & \theta_h & \lambda & & & \\
\mathbf{u+2} & & & & & & & & & & & & & & \mu_h & \theta_h & \lambda & & \\
. & & & & & & & & & & & & & & & . & . & . & \\
. & & & & & & & & & & & & & & & & . & . &
\end{array}
$$

Figure 1. Infinitesimal Generator.

Let the row vector $\mathbf{z} = [\mathbf{x}, \mathbf{y}]$ denote the equilibrium probability vector of the generator Q shown in Figure 1, where, $\mathbf{x} = [x_0, x_1, x_2, \cdots x_{u-1}, x_u]$ and $\mathbf{y} = [y_l, y_{l+1}, y_{l+2}, \cdots]$ represent equilibrium probabilities of being in the states where the server is operating at the normal and higher service rates respectively. $\mathbf{x}$ is of dimension $u+1$, and the dimension of $\mathbf{y}$ is infinite. The following results are due to Gebhardt [3], adapted to the notation in this paper.

$$
\begin{aligned}
x_0 &= \left[\frac{1}{1-\rho_n} - \frac{(u-l+2)\,\rho_n^u\,(\rho_n - \rho_h)}{(1-\rho_n^{u-l+2})\,(1-\rho_h)}\right]^{-1}, \\
x_r &= x_0\,\rho_n^r, \quad \text{for } r = 1, 2, \cdots l-1, \\
x_r &= x_0\left(\frac{\rho_n^r - \rho_n^{u+1}}{1-\rho_n^{u-l+2}}\right), \quad \text{for } r = l, l+1, l+2, \cdots u,
\end{aligned}
$$

$$
\begin{aligned}
y_r &= x_0\, A \left(\frac{\rho_h - \rho_h^{r-l+2}}{1-\rho_h^{u-l+2}}\right), \quad \text{for } r = l, l+1, l+2, \cdots u, u+1, \\
\text{where, } A &= \rho_n^u \left(\frac{1-\rho_n}{1-\rho_h}\right) \left(\frac{1-\rho_h^{u-l+2}}{1-\rho_n^{u-l+2}}\right). \\
y_{u+2} &= x_0\, A\, \rho_h^2, \\
y_r &= y_{r-1}\rho_h, \text{ for } r = u+3, u+4, \cdots \infty.
\end{aligned}
$$

Equilibrium probability vector $\mathbf{z}$ is indeterminate at ρ_n and/or ρ_h is equal to 1. This can be resolved by the application of L'Hospital's rule. $\mathbf{z}$ can also be evaluated using iterative methods such as those used in [7], which are effective even when ρ_n and/or ρ_h are equal to 1. The following results for the moments of N, the number in the system, are due to Gebhard [3].

$$
\begin{aligned}
E(N) &= x_0 \left[\frac{\rho_n}{(1-\rho_n)^2} - \frac{(u-l+2)\rho_n^u(\rho_n-\rho_h)}{(1-\rho_n^{u-l+2})(1-\rho_h)} \left\{\frac{u+l-1}{2} + \frac{1-\rho_n\rho_h}{(1-\rho_n)(1-\rho_2)}\right\}\right] \\
E[N(N-1)] &= x_0 \left[\frac{2\rho_n^2}{(1-\rho_n)^3} - \frac{(u-l+2)\rho_n^u(\rho_n-\rho_h)}{(1-\rho_n^{u-l+2})(1-\rho_h)}\right. \\
&\quad \left\{\frac{3(l-1)(u-1)+(u-l+1)(u-l)}{3} + \frac{(u+l-1)(1-\rho_n\rho_h)}{(1-\rho_n)(1-\rho_h)} + \right. \\
&\quad \left.\left. \frac{2(\rho_n+\rho_h-3\rho_n\rho_h+\rho_n^2\rho_h^2)}{(1-\rho_n)^2(1-\rho_h)^2}\right\}\right]
\end{aligned}
$$

Variance and coefficient of variation of N can be computed from the above results. The following additional measures of system performance are defined to facilitate the discussion of the system behavior in Section 5.

1. ϕ_n and ϕ_h, the proportions of times spent by the server at the normal (including idle time) and higher service levels are given by:

$$
\begin{aligned}
\phi_n &= \sum_{k=0}^{u} x_k, \text{ and} \\
\phi_h &= \sum_{k=l}^{\infty} y_k.
\end{aligned}
$$

2. η_n and η_h, the proportions of customers served while the server is operating at the normal and higher service rates are given by:

$$
\begin{aligned}
\eta_n &= \frac{(\phi_n - x_0)\mu_n}{(\phi_n - x_0)\mu_n + \phi_h\mu_h}, \text{ and} \\
\eta_h &= \frac{\phi_h\mu_h}{(\phi_n - x_0)\mu_n + \phi_h\mu_h}.
\end{aligned}
$$

3. μ_{eff}, the effective steady state rate at which the server operates is given by,

$$\mu_{eff} = \phi_n \mu_n + \phi_h \mu_h.$$

The server is idle with probability x_0 at the normal service rate.

4. μ_{eq}, the service rate in an equivalent $M/M/1$ queue with arrival rate λ and average number in the system equal to $E(N)$ can be obtained by considering the properties of $M/M/1$ queue as follows.

$$\mu_{eq} = \frac{1 + E(N)}{\lambda E(N)}$$

3. Time Spent by the Server at Each Service Rate

The server can be in one of two macro states, namely *normal* and *high* service rate. Let random variables t_n and t_h respectively denote the times spent by the server at *normal* and *high* service rates during each visit to the respective macro state, when the system is operating under steady state conditions.

System state alternates between the two intervals t_n (including the idle state) and t_h, with the start of $t_n(t_h)$ representing the end of $t_h(t_n)$. t_n is initiated when the number drops from $l+1$ to l by a service completion. t_n ends when the number increases from u to $u+1$ by an arrival. t_n and t_h represent alternative renewal processes.

3.1. Time spent at normal service rate

The random variable t_n can be described as the time to absorption in the infinitesimal generator R_n, where A_n is an absorbing state indicating the end of t_n, $\mathbf{e}$ is a column vector of 1's and $\mathbf{0}$ is a row vector of 0's.

$$R_n = \begin{array}{c|cccccccccc|c} & 0 & 1 & 2 & . & l & l{+}1 & l{+}2 & . & u & A_n \\ \hline 0 & -\lambda & \lambda & & & & & & & & \\ 1 & \mu_n & \theta_n & \lambda & & & & & & & \\ 2 & & \mu_n & \theta_n & \lambda & & & & & & \\ \cdots & & & . & . & . & & & & & \\ l & & & & \mu_n & \theta_n & \lambda & & & & \\ l{+}1 & & & & & \mu_n & \theta_n & \lambda & & & \\ l{+}2 & & & & & & \mu_n & \theta_n & \lambda & & \\ \cdots & & & & & & & . & . & . & \\ u & & & & & & & & \mu_n & \theta_n & \lambda \\ \hline A_n & & & & & & & & & & 0 \end{array} = \begin{bmatrix} T_n & -T_n\mathbf{e} \\ \mathbf{0} & 0 \end{bmatrix},$$

Figure 2. Interval t_n.

t_n has a phase type probability distribution [6], with $(u+1)$ phases and representation $[\boldsymbol{\alpha}_n, T_n]$ where $\boldsymbol{\alpha}_n$ is a $(u+1)$ row vector of initial probabilities. Since t_n always starts with the number in the system equal to $(l-1)$, $\boldsymbol{\alpha}_n = [0\ 0\ \cdots\ 1\ \cdots\ 0\ 0]$, has all 0's except for a 1 in the l^{th} position, corresponding to $l-1$ customers in the system. $E(t_n)$ and $Var(t_n)$ can be obtained from the properties of phase type distributions [6].

3.2. Time spent at higher service rate

Let the random variable t_h describe the time to absorption in the infinitesimal generator R_h , where A_h is an absorbing state denoting the end of t_h.

$$
R_h = \left\| \begin{array}{c|c|cccccccc}
 & A_h & \mathbf{l{+}1} & \mathbf{l{+}2} & . & \mathbf{u} & \mathbf{u{+}1} & \mathbf{u{+}2} & . & . \\
\hline
\mathbf{l{+}1} & \mu_h & \theta_n & \lambda & & & & & & \\
\mathbf{l{+}2} & & \mu_h & \theta_n & \lambda & & & & & \\
\cdots & & & . & . & . & & & & \\
\mathbf{u} & & & & \mu_h & \theta_n & \lambda & & & \\
\mathbf{u{+}1} & & & & & \mu_h & \theta_n & \lambda & & \\
\mathbf{u{+}2} & & & & & & \mu_h & \theta_n & \lambda & . \\
\cdots & & & & & & & . & . & . \\
\hline
A_h & 0 & & & & & & & &
\end{array} \right\| = \begin{bmatrix} T_h & -T_h\mathbf{e} \\ \mathbf{0} & 0 \end{bmatrix},
$$

Figure 3. Interval t_h.

t_h has a phase type probability distribution with representation $[\boldsymbol{\alpha}_h, T_h]$ where $\boldsymbol{\alpha}_h$ is a row vector of initial probabilities. Since t_h always starts with $(u+1)$ customers in the system, $\boldsymbol{\alpha}_h = [0\ 0\ \cdots\ 1\ \cdots\ 0\ 0]$, with 0's except for a 1 in the position corresponding to $u+1$ customers in the system. $E(t_h)$ and $Var(t_h)$ can be obtained from the properties of phase type distribution.

Unlike t_n, t_h has infinite number of phases. For computational purposes R_h (and T_h) need to be truncated of at a value that does not result in significant loss in accuracy. This issue is addressed in Section 4, in connection with the computation of the sojourn time density function. Details for the efficient computation of $E(t_n)$, $Var(t_n)$, $E(t_h)$ and $Var(t_h)$ are provided in the Appendix.

t_h can also be described in terms of the busy period of an M/M/1 queue, which is defined as the interval between the first arrival to the system when it is empty to the time when the system becomes empty again. Busy period of an $M/M/1$ queue with arrival rate λ and service rate mu_h is mathematically equivalent to the current system starting with $(u+1)$ customers, and reaching u customers for the first time. Since t_h always starts with $(u+1)$ customers in the system and ends with $(l-1)$ customers in the system, t_h is the sum of $(u-l+2)$ busy periods of an $M/M/1$ queue with arrival rate λ and service rate μ_h. The expected value and variance of the busy period are given by $\frac{1}{\mu_h-\lambda}$ and $\frac{\mu_h+\lambda}{(\mu_h-\lambda)^3}$ respectively. Since successive busy periods are statistically independent, the expected value and variance

of t_h are given by $\frac{u-l+2}{\mu_h-\lambda}$ and $\frac{(u-l+2)(\mu_h+\lambda)}{(\mu_h-\lambda)^3}$ respectively. The probability density function of the busy period of an $M/M/1$ queue can be expressed in terms of the modified Bessel function of the first kind of order one. In principle, t_h can be obtained as the $(u-l+2)$ fold convolution of this density function but it is difficult to implement.

While the results of Sections 3.1 and 3.2 can be used to compute the density functions for t_n and t_h, numerical results are presented only for the moments of t_n and t_h.

3.3. Direct computation of $E(T_n)$ and $E(T_h)$

When only the expected values of t_n and t_h are needed, they may be obtained directly by appealing to the properties of alternating renewal processes [4].

$$\text{P(system is operating at higher service rate)} = \sum_{k=l}^{\infty} y_k = \frac{E(t_h)}{E(t_h) + E(t_n)}$$

Computing $E(t_h)$ directly based on the discussion in Section 3, $E(t_n)$ can be obtained.

4. Sojourn Time of Customers

In systems with hysteretic control of service rates, sojourn (or waiting) time of a customer is not determined at the time a customer enters the system, because the service rate may change during the customer's sojourn from arrival to service completion, possibly more than once and even during service. This makes obtaining an analytical solution for the sojourn (or waiting) time distribution function very difficult. In this paper, we adopt an algorithmic approach, suggested by Neuts [6], by describing a virtual customer's sojourn from arrival to service completion, as the time to absorption in a continuous time Markov chain (CTMC). By supplementing the state description in the CTMC described in Figure 1 with additional information, it is possible to capture the effect of possible service rate changes during the sojourn of a customer through the system.

Let Q_s denote the modified CTMC. State of Q_s is defined by the three tuple $(i, j, k), i = 1, 2, \ldots,\ j = 1, 2, ..i,$ and $k = 0, 1$, where i denotes the number in the system, j the position of the customer in the waiting line, with $j = 1$ indicating that the customer is in service. As before, k = 0 or 1 depending on whether the server is operating at normal or higher service rate. It is known from Section 2 that for $i < l$, k is always 0, and for $i > u$, k is always 1. In the following discussion, whenever it is clear from the context, k will not be used in the state description. A_s denotes an absorbing state, indicating the end of the sojourn of the virtual customer. For stable systems (i.e., when $\rho_h < 1$) eventual absorption into the state A_s from any initial state is certain.

State changes that occur due to arrivals and service completions when the system is operating under steady state conditions are detailed below.

- A new arrival always increases i (number in the system) by 1 but does not affect j, the position of the customer already in the queue.

  - A new arrival encounters an empty system with probability x_0. This arrival enters the system in state $(1, 1)$, with the new arrival representing the only customer in the system and enters service immediately.
  - A new arrival encounters i customers in the system with probability x_i or y_i depending on whether server is operating at the normal of high service rate. This arrival enters the system in state $(i + 1, i + 1)$. (i.e., there will be $(i + 1)$ customers in the system with new arrival occupying the $(i + 1)^{th}$ position). The service rate remains the same except when $i = u$ and the server is operating at the normal service rate. In this case, the service rate increases from μ_n to μ_h.
  - A new arrival cannot encounter l customers in the system while the server is operating at the higher service rate because this state can only be reached by a service completion.

- A service completion decreases the number in the system and advances the position of the customers in the queue by 1.
  - After a service completion, the service rate remains the same except if it occurs when $i = l$. In this case, i drops to $l - 1$, j decreases by 1, and the service rate decreases from μ_h to μ_n.
  - Service completion from any state with $j = 1$, represents the end of the virtual customer's sojourn and leads to absorption in state A_s.

Q_s for a system with $u = 5$ and $l = 4$ is shown in Figure 4, where the steady state probabilities of an arrival entering the system in various states are shown in the column to the left of the state descriptions.

Let the random variable S denote the sojourn time of a customer who enters the system in steady state. S is the time to absorption in Q_s and has a phase type probability distribution. Partitioning the matrix Q_s by separating the last row and last column corresponding to the absorbing state A_s we have,

$$Q_s = \begin{bmatrix} T & \mathbf{t} \\ \mathbf{0} & 0 \end{bmatrix}.$$

Let $\boldsymbol{\alpha}_s$ describe the vector of initial probabilities for Q_s (i.e., the system state immediately following the arrival of a customer in steady state). The last element of $\boldsymbol{\alpha}_s$ indicates the probability of the system starting in the state A_s and will be 0. The probability distribution and density functions as well as the first two moments of X can be obtained from the properties of phase type distributions described in the Appendix.

The matrix Q_s is infinite and for computational purposes it must be truncated at a reasonable value such that its effect on the accuracy of quantities computed is within acceptable tolerances. This truncation point can be established by recognizing that at higher service rate, the system acts as an $M/M/1$ queue with arrival rate λ and service rate μ_h and $\rho_h = \frac{\lambda}{\mu_h}$.

Prob	State	(1,1)	(2,1) (2,2)	(3,1) (3,2) (3,3)	(4,1) (4,2) (4,3) (4,4)	(5,1) (5,2) (5,3) (5,4) (5,5)	**(4,1) (4,2) (4,3) (4,4)**	**(5,1) (5,2) (5,3) (5,4) (5,5)**	**(6,1) (6,2) (6,3) (6,4) (6,5) (6,6)**	**(7,1) (7,2)** $\cdots$	A_s
x_0	(1,1)	$*$	λ								μ_n
	(2,1)		$*$	λ							μ_n
x_1	(2,2)	μ_n	$*$	λ							
	(3,1)			$*$	λ						μ_n
	(3,2)		μ_n	$*$	λ						
x_2	(3,3)		μ_n	$*$	λ						
	(4,1)			$*$	λ						μ_n
	(4,2)			μ_n $*$	λ						
	(4,3)			μ_n $*$	λ						
x_3	(4,4)			μ_n $*$	λ						
	(5,1)				$*$				λ		μ_n
	(5,2)				μ_n $*$				λ		
	(5,3)				μ_n $*$				λ		
	(5,4)				μ_n $*$				λ		
x_4	(5,5)				μ_n $*$				λ		
	(4,1)					$*$		λ			μ_h
	(4,2)			μ_h		$*$		λ			
	(4,3)			μ_h		$*$		λ			
	(4,4)			μ_h		$*$		λ			
	(5,1)						$*$		λ		μ_h
	(5,2)						μ_h $*$		λ		
	(5,3)						μ_h $*$		λ		
	(5,4)						μ_h $*$		λ		
y_4	**(5,5)**						μ_h $*$		λ		
	(6,1)							$*$		λ	μ_h
	(6,2)							μ_h $*$		λ	
	(6,3)							μ_h $*$		$\cdots$	
	(6,4)							μ_h $*$			
	(6,5)							μ_h $*$			
x_5+y_5	**(6,6)**							μ_h $*$			
	(7,1)								$*$	$*$	μ_h
	(7,2)								μ_h $*$	$*$	
	(7,3)								μ_h $*$	$\cdots$	
	(7,4)								μ_h $*$		
	(7,5)								μ_h $*$		
	(7,6)								μ_h $*$		
y_6	**(7,7)**								μ_h $*$		
	$\cdots$									$\cdots$	$\cdots$
	A_s										0

Figure 4. Generator Q_s when $u = 5$ and $l = 4$.

For this $M/M/1$ queue, a truncation point can be set such that the probability mass lost due to truncation is limited to an arbitrarily small value ϵ as,

$$P(\text{Number in system} \geq m) = (\rho_h)^m < \epsilon, \text{ or}$$

$$m \geq \frac{log(\epsilon)}{log(\rho_h)}.$$

For the system under consideration, the higher service rate always starts with $u+1$ customers in the system and ends with $l-1$ customers in the system. Thus, a truncation value of $n = m + u - l + 2$, will ensure that no more than ϵ of probability mass is lost due to truncation.

The effect of truncation is to cutoff arrivals into the system when the number in system reaches n. A small improvement in accuracy can be achieved by considering the probability of an arrival encountering n customers in the truncated system as $\sum_{i=n}^{\infty} y_i$.

The dimension of the truncated matrix Q_s will be $n(n+1)/2$. This can be quite large, especially when ρ_h is close to 1. Thus, evaluation of $f_s(x)$, especially if required at a set of finely spaced values of S, is computationally very demanding. When the matrix P and vector $\mathbf{p}$ are sparse, as in this case, computation of vectors $\boldsymbol{\psi}(k)$ can be organized efficiently as described in the Appendix.

Waiting time (W) density function can be computed by modifying Q_s to obtain Q_w as follows.

- Remove all states with $j = 1$. States with $j = 2$ will lead to absorption in state A_w at the end of service.
- All initial probabilities will remain the same except for the deletion of x_0 corresponding to state $(1, 1)$.
- The initial probabilities sum to $1 - x_0$. Correspondingly, the waiting time density function will have an impulse function at $W = 0$ equal to x_0, corresponding to an arrival encountering an empty system and thus, having zero waiting time.
- To compute the conditional waiting time distribution, the vector of initial probabilities need to be normalized to sum to 1.

5. Numerical Results for System Behavior

This Section presents a summary of observations on the behavior of the system under various traffic intensities (ρ_n and ρ_h), and upper and lower control limits (u and l). This summary is based on an extensive set of computer runs for a wide range of values for ρ_n (0.7 to 1.3), ρ_h (0.4 to 0.8), u (5 to 40), and l (1 to u, with $l = 1$ meaning that the higher service rate is maintained until the system is empty). In the interests of brevity, detailed tables and graphs are presented only for two combinations of (ρ_n, ρ_h), namely, (0.9, 0.7)

and (1.2. 0.6) and a limited set of values for u and l. Without loss of generality, $\lambda = 1$ is used, implying that the time unit is set equal to the mean interarrival time.

It is useful to compare the performance measures of the system under consideration (henceforth referred to as the *hysteretic* system) with those of $M/M/1$ queues with the same arrival rate, and service rates μ_n and/or μ_h (referred to as $MM1n$ and $MM1h$). In the tables below, wherever appropriate, relevant information for the $MM1n$ and/or $MM1h$ queues is included for ready comparison with the *hysteretic* system. For example, in Table 1, which summarizes the values of x_0, values of the probabilities of empty system for $MM1n$ and/or $MM1h$ are included. No data is presented for $MM1n$ systems with $\rho_n \geq 1$. Subscript n and h have the same meaning as before.

Table 1. Probability of an empty system (x_0)

$\rho_n = 0.9, \rho_h = 0.7, \lambda = 1$ ($x_{0n} = 0.1, x_{0h} = 0.3$)						
u/l	1	5	10	20	30	40
5	0.202	0.171				
10	0.159	0.145	0.132			
20	0.124	0.120	0.116	0.109		
30	0.110	0.109	0.107	0.105	0.103	
40	0.104	0.104	0.103	0.103	0.102	0.101
$\rho_n = 1.2, \rho_h = 0.6, \lambda = 1$ ($x_{0h} = 0.4$)						
5	0.159	0.084				
10	0.092	0.050	0.027			
20	0.046	0.024	0.012	0.004		
30	0.030	0.015	0.007	0.002	0.001	
40	0.022	0.011	0.005	0.001	0.000	0.000

Table 2. Mean Number in the system ($E(N)$)

$\rho_n = 0.9, \rho_h = 0.7, \lambda = 1$ ($E(N_n) = 9.0, E(N_h) = 2.33$)						
u/l	1	5	10	20	30	40
5	3.070	3.457				
10	4.050	4.316	4.843			
20	5.850	5.962	6.204	6.870		
30	7.163	7.208	7.309	7.619	7.993	
40	8.004	8.021	8.061	8.191	8.366	8.551
$\rho_n = 1.2, \rho_h = 0.6, \lambda = 1$ ($E(N_h = 1.5$)						
5	2.983	3.925				
10	4.944	5.855	7.612			
20	9.363	10.262	12.034	16.429		
30	14.082	14.989	16.789	21.253	26.099	
40	18.931	19.846	21.668	26.180	31.055	36.021

Tables 1 and 2 summarize the values of x_0 and $E(N)$. Since λ is to 1, values in Table 2 also represent the values of $E(S)$. Increasing u, or increasing l for a given u (i.e., decreasing $(u - l)$) increased $E(N)$, as a result of the system spending increasing amount of time at the normal service rate, thus serving customers at a slower rate. When $\rho_n < 1$, the value of x_0 and $E(N)$ fell between the corresponding values for $MM1n$ and $MM1h$ systems. Increasing u, moved these values away from the values for $MM1h$ and closer to the values for $MM1n$. When $\rho_n > 1$, the value of $E(N)$ were always greater than the corresponding values for $MM1h$ system and, except for very small values of u, x_0 is very close to zero.

Table 3. Effective Service rate (μ_{eq})

$\rho_n = 0.9, \rho_h = 0.7, \lambda = 1$ ($\mu_n = 1.11, \mu_h = 1.429$)						
u/l	1	5	10	20	30	40
5	1.224	1.190				
10	1.177	1.162	1.147			
20	1.137	1.133	1.129	1.122		
30	1.122	1.121	1.119	1.117	1.115	
40	1.116	1.115	1.115	1.114	1.113	1.112
$\rho_n = 1.2, \rho_h = 0.6, \lambda = 1$ ($\mu_n = 0.833, \mu_h = 1.667$)						
5	1.133	1.070				
10	1.077	1.041	1.022			
20	1.038	1.020	1.010	1.003		
30	1.025	1.013	1.006	1.002	1.001	
40	1.018	1.009	1.004	1.001	1.000	1.000

Table 4. Service rate in an equivalent $M/M/1$ system (μ_{eff})

$\rho_n = 0.9, \rho_h = 0.7, \lambda = 1$ ($\mu_n = 1.11, \mu_h = 1.429$)						
u/l	1	5	10	20	30	40
5	1.326	1.289				
10	1.247	1.232	1.207			
20	1.171	1.168	1.161	1.146		
30	1.140	1.139	1.137	1.131	1.125	
40	1.125	1.125	1.124	1.122	1.120	1.117
$\rho_n = 1.2, \rho_h = 0.6, \lambda = 1$ ($\mu_n = 0.833, \mu_h = 1.667$)						
5	1.335	1.255				
10	1.202	1.171	1.131			
20	1.107	1.097	1.083	1.061		
30	1.071	1.067	1.060	1.047	1.038	
40	1.053	1.050	1.046	1.038	1.032	1.028

It is interesting to compare μ_{eq} and μ_{eff} (defined in Section 2) summarized in Tables 3 and 4. It is important to note that for stable systems (i.e., $\rho_h < 1$), μ_{eq} will always be greater than λ (in this case > 1). μ_{eq} is always found to be greater than μ_{eff}, indicating that

a *hysteretic* system is more efficient than an equivalent $M/M/1$ system (defined as one with the same arrival rate and same expected number in the system). This is further supported by noting that all values of x_0 from Table 1, are smaller than probabilities of emptiness for the equivalent system obtained as $(\mu_{eff} - 1)$ (since $\lambda = 1$). Hysteretic systems, however, have the additional cost of service rate changes.

Values of μ_{eq} and μ_{eff} always fell between the service rates for $MM1n$ and $MM1h$ systems. As u increased, μ_{eff} gradually approached μ_n when $\rho_n < 1$, and approached λ $(= 1$, the minimum equivalent service rate required for system stability), when $\rho_n > 1$. Effect of l is similar but moderate.

Table 5. Percent of time spent at high service rate (ϕ_h)

$\rho_n = 0.9, \rho_h = 0.7, \lambda = 1$						
u/l	1	5	10	20	30	40
5	35.58%	24.77%				
10	20.79%	15.86%	11.34%			
20	8.27%	6.93%	5.49%	3.26%		
30	3.51%	3.09%	2.59%	1.74%	1.07%	
40	1.49%	1.35%	1.18%	0.86%	0.59%	0.37%
$\rho_n = 1.2, \rho_h = 0.6, \lambda = 1$						
5	35.93%	28.36%				
10	29.18%	24.96%	22.69%			
20	24.58%	22.43%	21.22%	20.39%		
30	22.95%	21.53%	20.72%	20.19%	20.06%	
40	22.16%	21.10%	20.50%	20.11%	20.03%	20.01%

Table 6. Percent of customers served at high service rate(η_h)

$\rho_n = 0.9, \rho_h = 0.7, \lambda = 1$						
u/l	1	5	10	20	30	40
5	50.82%	35.38%				
10	29.70%	22.66%	16.20%			
20	11.81%	9.91%	7.84%	4.66%		
30	5.01%	4.41%	3.70%	2.48%	1.53%	
40	2.13%	1.93%	1.69%	1.23%	0.84%	0.52%
$\rho_n = 1.2, \rho_h = 0.6, \lambda = 1$						
5	59.89%	47.26%				
10	48.64%	41.60%	37.81%			
20	40.97%	37.38%	35.36%	33.98%		
30	38.25%	35.88%	34.53%	33.65%	33.44%	
40	36.93%	35.16%	34.16%	33.52%	33.38%	33.35%

Tables 5 and 6 summarize ϕ_h and η_h, the percent of time spent by the system at the higher service rate, and the percent of customers served at the higher service rate. As u and

l increased, ϕ_h and η_h decreased, with the effect of u being more pronounced than l. η_h is always higher than ϕ_h because more customers are served per unit time at the higher service rate.

Table 7. Mean value of t_n $(E(t_n))$

$\rho_n = 0.9, \rho_h = 0.7, \lambda = 1$						
u/l	1	5	10	20	30	40
5	25.35	14.18				
10	97.80	86.62	36.49			
20	543.53	532.35	482.22	138.28		
30	1990.00	1978.80	1928.70	1584.70	430.21	
40	6306.50	6295.30	6245.20	5901.30	4746.70	1267.40
$\rho_n = 1.2, \rho_h = 0.6, \lambda = 1$						
5	16.05	7.58				
10	40.04	31.57	10.22			
20	96.65	88.18	66.84	11.71		
30	156.11	147.64	126.29	71.17	11.95	
40	216.02	207.55	186.20	131.08	71.87	11.99

Table 8. Mean value of t_h $(E(t_h))$

$\rho_n = 0.9, \rho_h = 0.7, \lambda = 1$						
u/l	1	5	10	20	30	40
5	14.00	4.67				
10	25.67	16.33	4.67			
20	49.00	39.67	28.00	4.67		
30	72.33	63.00	51.33	28.00	4.67	
40	95.67	86.33	74.67	51.33	28.00	4.67
$\rho_n = 1.2, \rho_h = 0.6, \lambda = 1$						
5	9.00	3.00				
10	16.50	10.50	3.00			
20	31.50	25.50	18.00	3.00		
30	46.50	40.50	33.00	18.00	3.00	
40	61.50	55.50	48.00	33.00	18.00	3.00

Tables 7 and 8 summarize the values of $E(t_n)$ and $E(t_h)$. Increasing u results in fewer increases in service rate so that the server spends more time at the normal service rate, giving larger $E(t_n)$. $E(t_h)$ clearly increases with increasing $(u - l)$ (i.e., decreasing l for a given value of u). It follows that increasing u for a given l will result in increases in both $E(t_n)$ and $E(t_h)$.

When $\rho_n < 1$, the server enters the higher rate less frequently and spends relatively more time at the normal service rate compared to when $\rho_n > 1$. For example, when u=20 and l=10, the ratio of $E(t_n)$ to $E(t_h)$ was 17.22 when $\rho_n = 0.9, \rho_h = 0.7$. The same ratio

was 3.71 when $\rho_n = 1.2, \rho_h = 0.6$. These results can be explained by considering t_{nx}, the time taken by the system, when operating at normal service rate, to visit any state $(r+1)$ for the first time, starting from the state r. Since t_n starts with $(l-1)$ in the system and ends with $(u+1)$ in the system, t_n is the sum of $(u-l+2)$ independent intervals of t_{nx}. When $\rho_n < 1$, there will be a drift towards smaller number in the system so that the interval t_{nx}, and correspondingly t_n, can be very large. When $\rho_n > 1$, the drift will be towards larger levels, so that the interval t_{nx}, and correspondingly t_n, will be relatively small. In both cases, t_h depends very strongly on the value of $(u-l)$.

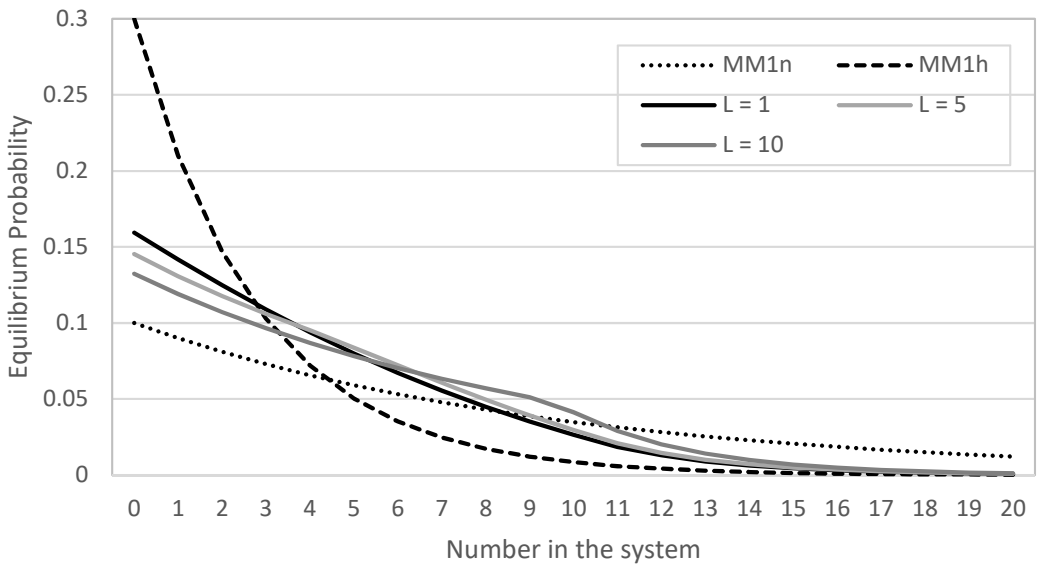

Figure 5. Distribution of Number in System (ρ_n=0.9, ρ_h=0.7, u=10)

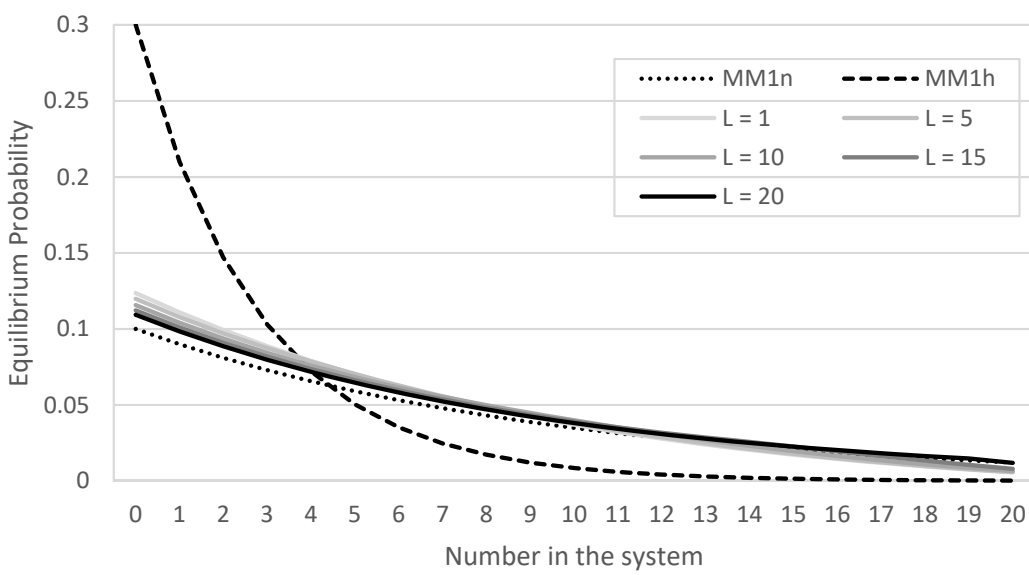

Figure 6. Distribution of Number in System (ρ_n=0.9, ρ_h=0.7, u=20)

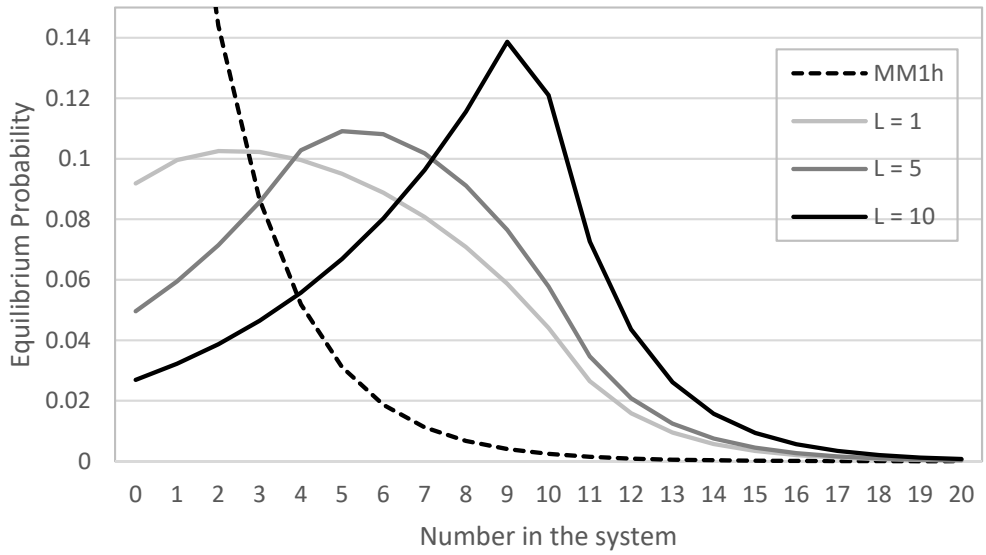

Figure 7. Distribution of Number in System (ρ_n=1.2, ρ_h=0.6, u=10)

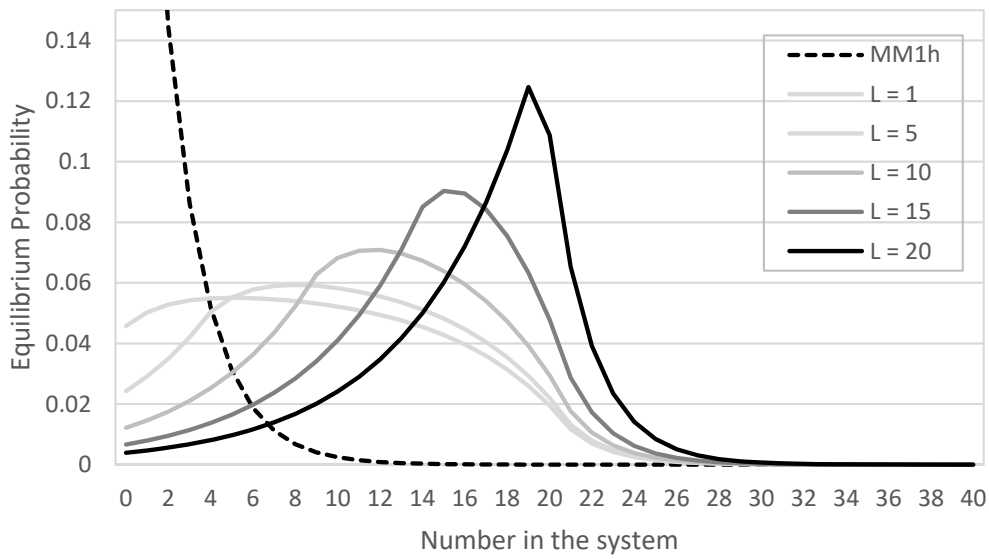

Figure 8. Distribution of Number in System (ρ_n=1.2, ρ_h=0.6, u=20)

Figures 5 to 8 display the distributions of number in the system. These distributions are dramatically different for $\rho_n < 1$ and $\rho_n > 1$. For systems with $\rho_n < 1$ (Figures 5 and 6), they are close to a geometric distribution, with $MM1n$ and $MM1h$ systems serving as the upper and lower bounds. For systems with $\rho_n > 1$ (Figures 7 and 8), the probability distributions have a distinct mode and the distributions got *"peekier"* (positive and increasing kurtosis) as l is increased (for a given u), with the mode occurring near l.

Figures 9 to 12 present the sojourn time density functions for the same set of parameter values as in Figures 5 to 8. For systems with $\rho_n < 1$ (Figures 9 and 10), sojourn time density

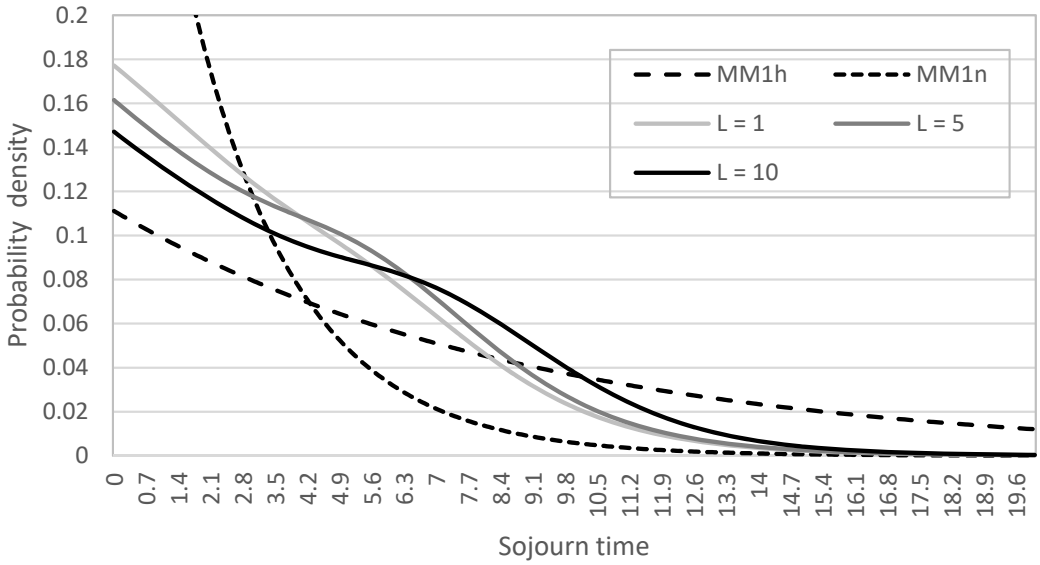

Figure 9. Sojourn Time Density (ρ_n=0.9, ρ_h=0.7, u=10)

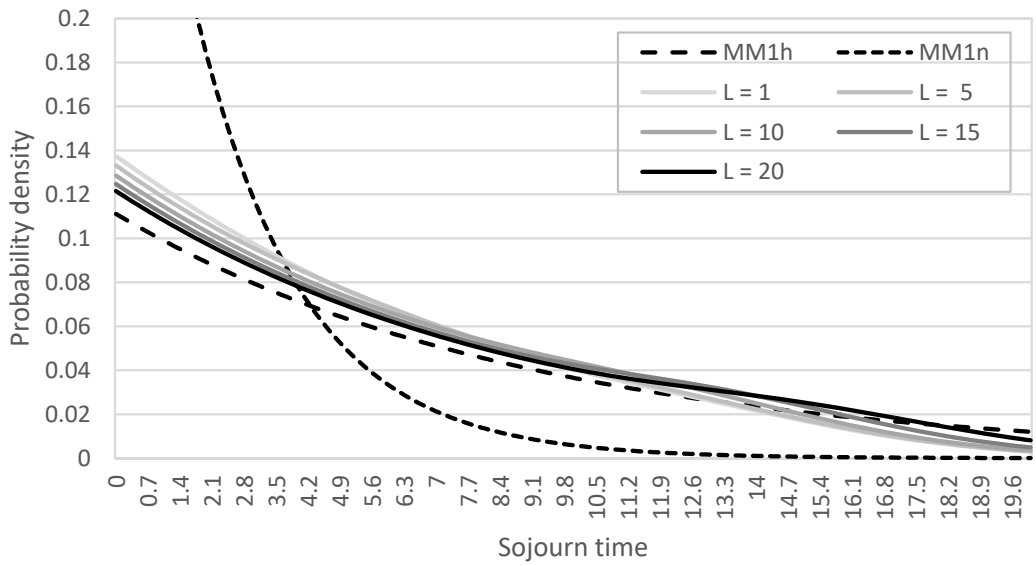

Figure 10. Sojourn Time Density (ρ_n=0.9, ρ_h=0.7, u=20)

functions are similar in shape to an exponential density function, falling between the sojourn times for $M/M/1n$ and $M/M/1h$ queues. The effect of l is relatively small and got less pronounced as u increased. For systems with $\rho_n > 1$ (Figures 11 and 12), as l increased (for a given u), the probability mass shifted to lower values of sojourn time. For most parameter values, the density functions appeared similar to a normal distribution.

Tables 9 and 10 summarize $s(N)$ and $s(S)$, the standard deviations of N and S. They also include the standard deviations of number in the system ($= \frac{\sqrt{\rho}}{(1-\rho)}$) and sojourn time (= $E(S)$ because $\lambda = 1$) for $MM1n$ and $MM1h$ systems. $s(N)$ and $s(S)$ are always less than

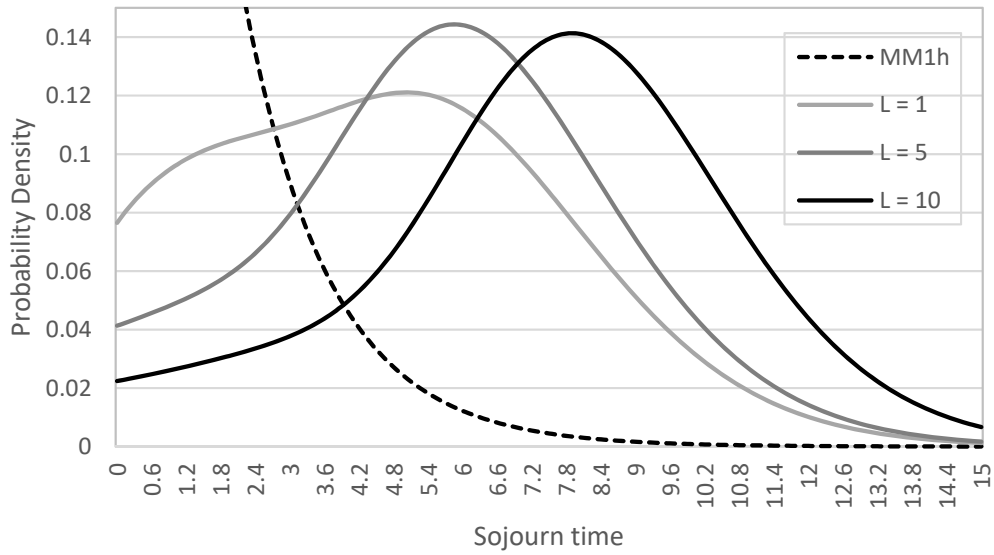

Figure 11. Sojourn Time Density (ρ_n=1.2, ρ_h=0.6, u=10)

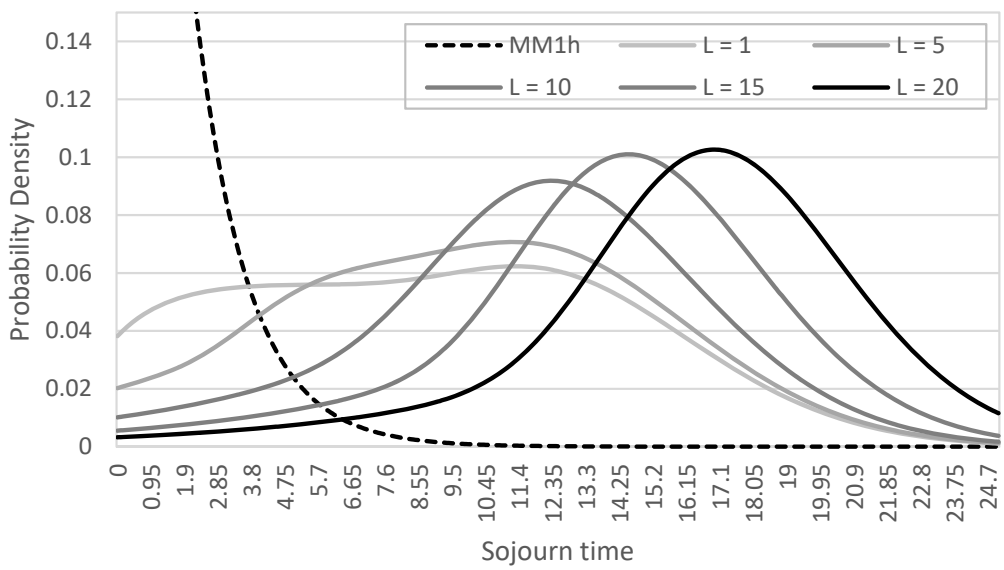

Figure 12. Sojourn Time Density (ρ_n=1.2, ρ_h=0.6, u=20)

the corresponding values for $E(N)$ (=$E(S)$) in Table 2, indicating that $CV(N)$ and $CV(S)$ are always less than 1. $s(N)$ and $s(S)$ increased with increasing u and l. Together, Tables 2, 9 and 10 indicate that the overall behavior of the *hysteretic* system is less variable than both $MM1n$ and $MM1h$ system due to the moderating effect of the service rate changes. $s(S)$ is always smaller than $s(N)$ because when the server is operating at higher service rate, customers are served at a faster rate resulting sojourn time being less variable than the number in the system.

Finally, Figures 13, 14 and 15 display the probability sub-vectors $\mathbf{x}$ and $\mathbf{y}$ on the same graph for comparison. In these figures, for $i < l$ and $i > u$, the total probability of i in

Table 9. Standard Deviation of N ($s(N)$)

$\rho_n = 0.9, \rho_h = 0.7, \lambda = 1$ ($s(N_n) = 9.487, s(N_h) = 2.789$)						
u/l	1	5	10	20	30	40
5	3.063	3.190				
10	3.700	3.785	4.094			
20	5.251	5.281	5.406	6.007		
30	6.644	6.656	6.705	6.978	7.462	
40	7.711	7.715	7.735	7.849	8.080	8.395
$\rho_n = 1.2, \rho_h = 0.6, \lambda = 1$ ($s(N_h) = 1.936$)						
5	2.489	2.577				
10	3.517	3.428	3.549			
20	5.975	5.698	5.288	4.945		
30	8.617	8.252	7.590	6.240	5.552	
40	11.344	10.929	10.135	8.274	6.606	5.755

Table 10. Standard deviation of S ($s(S)$)

$\rho_n = 0.9, \rho_h = 0.7, \lambda = 1$ ($s(S_n) = 9.0, s(S_h) = 2.33$)						
u/l	1	5	10	20	30	40
5	2.543	2.672				
10	3.149	3.225	3.560			
20	4.696	4.721	4.846	5.480		
30	6.104	6.114	6.161	6.443	6.948	
40	7.186	7.190	7.209	7.324	7.565	7.891
$\rho_n = 1.2, \rho_h = 0.6, \lambda = 1$ ($s(S_h) = 1.5$)						
5	1.951	2.095				
10	2.962	2.851	3.110			
20	5.432	5.095	4.674	4.641		
30	8.094	7.680	6.949	5.786	5.469	
40	10.842	10.387	9.526	7.636	6.459	5.989

the system is the same as the individual probabilities in vectors $\mathbf{x}$ or $\mathbf{y}$. For $l \leq i \leq u$, the total probability is the sum of the corresponding probabilities from $\mathbf{x}$ and $\mathbf{y}$. These graphs provide a sense of the relative values of probabilities of the server being in normal or higher service rate when the number in the system is between l and u for systems with $\rho_n < 1$ and $\rho_n > 1$.

In summary, the upper control limit u has the more dominant effect on the system characteristics than the lower control limit l. The difference $(u - l)$ has a strong impact on the frequency of service rate changes, and the mean and variance of the times spent at the two service levels. For systems with $\rho_n < 1$, the sojourn time density functions appear to be closer in shape to an exponential density function and for systems with $\rho_n > 1$, the sojourn time density functions appear to be closer in shape to a normal density function. Many observations are consistent with intuitive expectations, and the tools developed in this paper

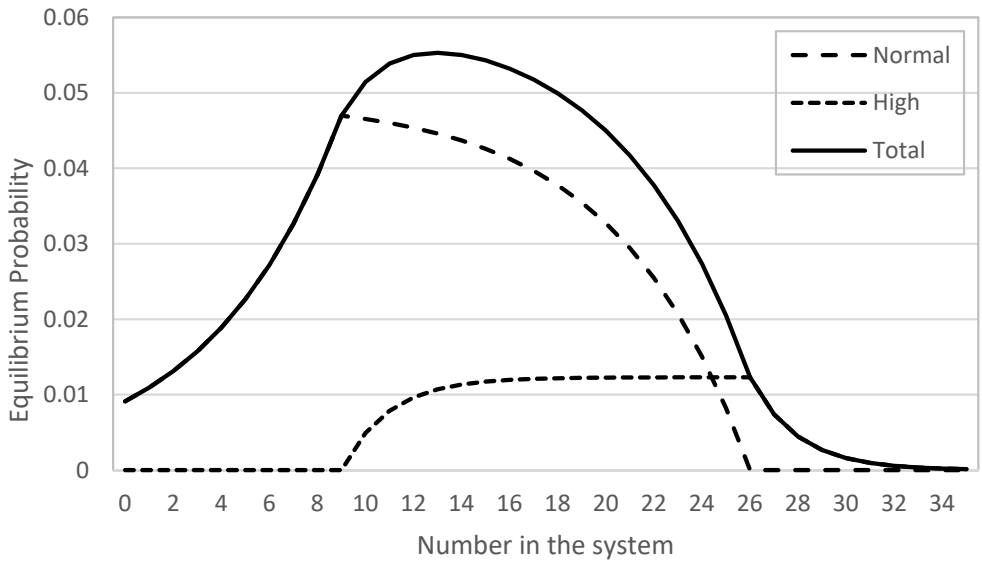

Figure 13. Distribution of Number in System (ρ_n=1.2, ρ_h=0.6, u=25, l=10)

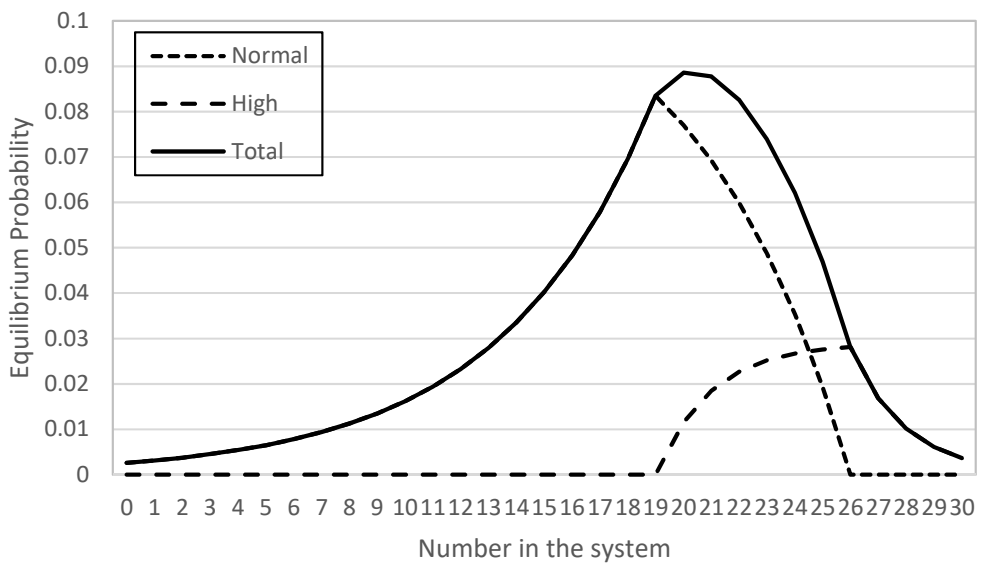

Figure 14. Distribution of Number in System (ρ_n=1.2, ρ_h=0.6, u=25, l=20)

provide the means of quantifying the intuitive expectations. Some observations, especially those dealing with the sojourn time density functions present new insights into the system behavior.

6. Concluding Remarks

Let the costs per unit time of operating the server at the normal and higher service rates be given by c_n and c_h respectively. Let the fixed cost of increasing [decreasing] the service rate be denoted by c_i [c_d]. Let c_s denote the cost per unit time, associated with the time

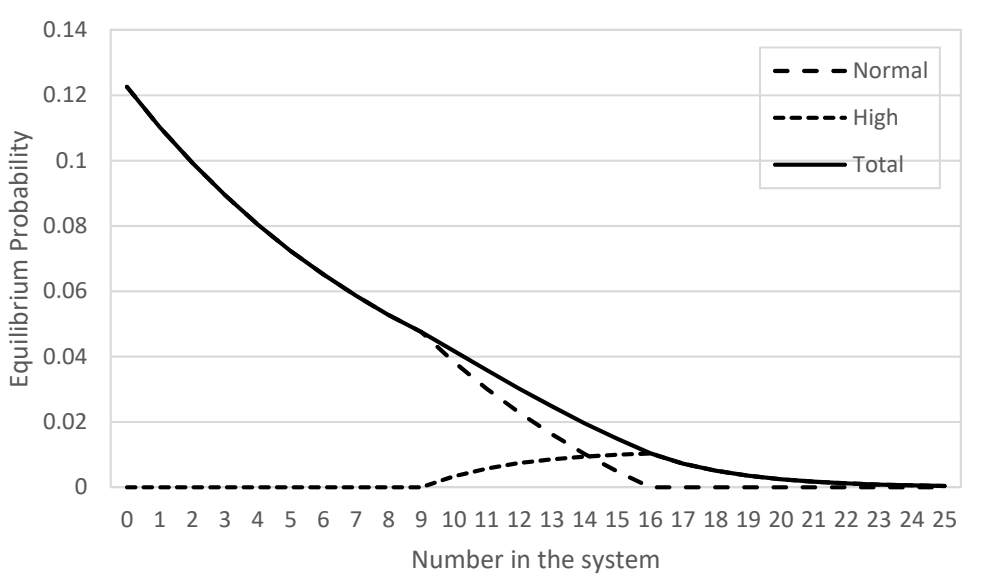

Figure 15. Distribution of Number in System (ρ_n=0.9, ρ_h=0.7, u=15, l=10)

spent by a customer in the system. C_1, C_2, and C_3, the expected costs per unit time of (i) operating the server, (ii) service rate changes, and (iii) waiting customers under equilibrium condition are given by,

$$\begin{aligned} C_1 &= \left[\sum_{i=0}^{u} x_i\right] c_n + \left[\sum_{i=l}^{\infty} y_i\right] c_h, \\ C_2 &= \left[\lambda x_u c_i + y_{u+1} c_d\right], \\ C_3 &= \left[E(N)\right] c_s. \end{aligned}$$

The algorithms developed in this paper can be used interactively to perform an efficient search for the optimal combination of trigger points for given values of ρ_n and ρ_h. One can start with a reasonable set of values for l and u and fine tune them until the desired performance measure is optimized. Strategies for the search algorithms can be based on the heuristic understanding of system behavior discussed in Section 5 and any other available information about the system behavior. Such a search procedure permits the analyst to incorporate intuitive understanding of the system behavior into the search process and will likely lead to the optimal solution with reduced effort.

As studied in Neuts and Rao [7], systems with finite waiting space presents interesting questions. In such systems, C_1, C_2, and C_3, defined above must be balanced with the cost of lost customers when the waiting space is full. The methodology presented in this paper can be used as is to study such systems by simply replacing the truncation level n discussed in Section 4, by the size of the waiting space. The proposed algorithmic methods can also be readily adopted to systems with k-level hysteretic control.

References

[1] Crabill, T. B., Gross, D. & Magazine, M. J. (1977). A classified bibliography of research on optimal design and control of queues, *Operations Research*, 25, 219-232.

[2] Federgruen, A. & Tijms, H. C. (1980). Computation of the stationary distribution of the queue size in an $M/G/1$ queueing system with variable service rate, *Journal of Applied Probability*, 17, 515-522.

[3] Gebhard, R. F. (1967). A queueing process with bilevel hysteretic service-rate control, *Naval Research logistics Quarterly*, 14, 55-67.

[4] Kulkarni, V. G. (2011). Modeling and Analysis of Stochastic Systems, 2nd Edition, Springer, New York, New York, USA.

[5] Matthews, D. E. (1975). *Controlling Single Server Queues*, Ph. D. Thesis, University of London.

[6] Neuts, M. F. (1981). *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, Johns Hopkins University Press, Baltimore, MD.

[7] Neuts, M. F. & Rao, B.M. (1992). On the Design of a Finite Capacity Queue With Phase Type Service Times and Hysteretic Control, *European Journal of Operations Research*, 62, 221-240.

[8] Sias, M., Noble, G. & Singh, P. (2017). Operations Research at Kroger, ORMS Today, December 2017.

[9] Latouche, G., & Ramaswami, V. (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling*, ASA-SIAM Series on Statistics and Applied Mathematics.

[10] Tijms, H. C. (1976). Optimal control of the workload in an $M/G/1$ queueing system with removable server, *Mathematische Operationsforschung und Statistik*, 7, 933-944.

[11] Yadin, M. & Naor, P. (1967). On queueing systems with variable service capacities, *Naval Research logistics Quarterly*, 14, 43-53.

Appendix
Phase Type Probability Distributions

A probability distribution is said to be of *phase type* or *PH*, if it can be described as the probability distribution of the time until absorption in a finite Markov chain [6]. Consider an $m+1$ state continuous time Markov chain with infinitesimal generator

$$Q = \begin{bmatrix} T & \mathbf{t} \\ \mathbf{0} & 0 \end{bmatrix},$$

and initial probability row vector $(\boldsymbol{\alpha}, \alpha_{m+1})$ where, $\mathbf{t} = -T\mathbf{e}$ and $\mathbf{e}$ is a column vector of 1s. If T is non-singular, eventual absorption into the state $(m+1)$ from any initial state is certain. α_{m+1} represents the probability that absorption occurs instantaneously at the start, and represents an impulse function at $t = 0$. *PH* distributions are generalizations of the Erlang and hyperexponential distributions and remain analytically and computationally tractable under a variety of operations in the analysis of stochastic models. They have been used successfully in developing efficient and computationally stable algorithms for system characteristics for a wide variety of queueing models [6]. We refer the reader to Neuts [7] for a complete discussion of the properties of *PH* distributions.

Let X be the random variable denoting the time till absorption in Q. X is said to have a *phase type* distribution with representation $(\boldsymbol{\alpha}, T)$. The probability distribution and density functions are given by,

$$\begin{aligned} F(x) &= 1 - \boldsymbol{\alpha}\, e^{Tx}\, \mathbf{e}, \quad x > 0, \\ f(x) &= \boldsymbol{\alpha}\, e^{Tx}\, \mathbf{t}, \quad x > 0. \end{aligned}$$

The first two moments of X are given by,

$$\begin{aligned} E(X) &= -\boldsymbol{\alpha} T^{-1} \mathbf{e}, \\ Var(X) &= 2\boldsymbol{\alpha} T^{-2} \mathbf{e} - (\boldsymbol{\alpha} T^{-1} \mathbf{e})^2. \end{aligned}$$

Computation of the moments

When the number of phases is large, matrix inversion in implementing the above equations may be avoided by recognizing that one only needs the vectors $T^{-1}\ \mathbf{e}$ and $T^{-2}\ \mathbf{e}$. Rewriting $E(X)$ and $Var(X)$ as

$$\begin{aligned} E(X) &= -\boldsymbol{\alpha} T^{-1} \mathbf{e} \ = \ \boldsymbol{\alpha}\boldsymbol{\phi}, \\ Var(X) &= 2\boldsymbol{\alpha} T^{-2} \mathbf{e} - (E(X))^2 \ = \ 2\boldsymbol{\alpha}\boldsymbol{\omega} - (E(X))^2 \end{aligned}$$

$\boldsymbol{\phi}(= -T^{-1}\ \mathbf{e})$ and $\boldsymbol{\omega}(= -T^{-2}\ \mathbf{e})$ can be evaluated by solving the following two systems of equations sequentially.

$$T\ \boldsymbol{\phi} = -\mathbf{e},$$

$$T\,\boldsymbol{\omega} = -\boldsymbol{\phi}.$$

In many cases, matrix T is very sparse, making it highly suitable to use an iterative approach. The basic principle calls for the infinitesimal generator T to be split as $T = T_1 - T_2$ where T_1 is a non-singular matrix. The equation $T\boldsymbol{\phi} = \mathbf{e}$ can then be rewritten as $T_1\boldsymbol{\phi} = T_2\boldsymbol{\phi} + \mathbf{e}$ which suggests the iterative scheme given by,

$$T_1\boldsymbol{\phi}(t+1) = T_2\boldsymbol{\phi}(t) + \mathbf{e}\ ,$$

where $\boldsymbol{\phi}(t)$ is the estimate of $\boldsymbol{\phi}$ obtained at iteration t. Under fairly general conditions, convergence to the correct value of $\boldsymbol{\phi}$ is guaranteed. Many schemes are described in the literature for the choice of T_1 and T_2 and perhaps the simplest of them is the point Gauss-Seidel scheme, in which $T_2 = T_{ut} + T_{lt}$ and $T_1 = -T_d$, where T_d, T_{ut} and T_{lt} are the diagonal, strictly upper triangular and strictly lower triangular portions of T. Higher moments can be computed sequentially by solving additional systems of equations.

Computation of the density function

Computing the density or distribution functions require the evaluation of e^{Tx}. The presence of negative diagonal elements in T makes the direct computation of e^{Tx} numerically hazardous and is not recommended. For computational stability and error control, the recommended method is *uniformization* [9], where computations are performed in terms of a corresponding discrete time Markov chain, embedded in a Poisson process with rate τ equal to the absolute value of the most negative diagonal element in T [9].

Let K denote a substochastic matrix defined as follows.

$$K = \frac{1}{\tau}Q + I = \begin{bmatrix} P & \mathbf{p} \\ \mathbf{0} & 1 \end{bmatrix}.$$

We then have,

$$e^{Tx} \;=\; \sum_{k=0}^{\infty} e^{-\tau x}\frac{(\tau x)^k}{k!}P^k.$$

The distribution and density functions can be expressed as,

$$F(x) \;=\; 1 - \boldsymbol{\alpha}\, e^{Tx}\, \mathbf{e} = 1 - \boldsymbol{\alpha}\, \sum_{k=0}^{\infty} e^{-\tau x}\frac{(\tau x)^k}{k!}P^k\mathbf{e}, \quad x > 0,$$

$$f(x) \;=\; \boldsymbol{\alpha}\, e^{Tx}\, \mathbf{t} = \boldsymbol{\alpha}\, \sum_{k=0}^{\infty} e^{-\tau x}\frac{(\tau x)^k}{k!}P^k\, \mathbf{t}, \quad x > 0.$$

These equations can be expressed as follows for efficient algorithmic implementation.

$$F_s(x) \;=\; 1 - e^{-\tau x}\left[\sum_{k=0}^{\infty} \boldsymbol{\psi}(k)\mathbf{e}\right],$$

$$f_s(x) \;=\; e^{-\tau x}\left[\sum_{k=0}^{\infty}\boldsymbol{\psi}(k)\mathbf{t}\right].$$

where,

$$\begin{aligned}\boldsymbol{\psi}(0) &= \boldsymbol{\alpha}\\ \boldsymbol{\psi}(k) &= (\frac{\tau x}{k})\boldsymbol{\psi}(k-1)P, \;\text{ for } k=1,2,\cdots\end{aligned}$$

Only two vectors $\boldsymbol{\psi}(k)$ need to be stored as they are calculated recursively as the scalar values of $F_s(x)$ and $f_s(x)$ are accumulated. In many cases, as in in the present case, the matrix P and vector $\mathbf{p}$ are very sparse so that the computations can be organized efficiently without actually generating and storing P and $\mathbf{p}$.

For extremely large values of τx (> 250 [4]), the value of $e^{-\tau x}$ may be very small and lead to underflow problems. The method suggested in [9] should be followed under those circumstances.

Evaluation of $f_s(x)$ would require the truncation of an infinite series. [9] suggests a method to determine the cutoff value for k, which requires the inversion of matrix $(I-P)$. When the dimension of the matrix $(I-P)$ is large, we may simply terminate the infinite series when the individual term in the summation $\sum_{k=0}^{\infty}\boldsymbol{\psi}(k)\mathbf{p}$ drops below a preset arbitrarily small value. In implementing this method, it is important to remember that individual terms in the summation increase initially when $\tau x < k$, but eventually start to decrease as k increases.

In many applications, as in the present case, matrix P is infinite and needs to be truncated at a reasonable value. The methods to be used for these two truncations depend very much on the specific application. Truncation of the matrix T for the present system are discussed in the main body of the paper.